HELMHOLTZ GEMEINSCHAFT

Technische Universität München

Supporting the evidence for human trait-associated genetic variants by computational biology methods and multi-level data integration

Matthias Arnold

2016

Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Supporting the evidence for human trait-associated genetic variants by computational biology methods and multi-level data integration

Matthias Arnold

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. J. J. Hauner
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. H.-W. Mewes
	2. Univ.-Prof. Dr. Dr. F. J. Theis
	3. Univ.-Prof. Dr. F. Kronenberg Medizinische Universität Innsbruck Österreich

Die Dissertation wurde am 10.03.2016 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 18.07.2016 angenommen.

Danksagung

Das größte Dankeschön eines Doktoranden gilt nicht nur traditionell seinem Doktorvater, und somit möchte auch ich mich zuerst bei (Hans-)Werner Mewes für die Ermöglichung dieses Projektes, sein Vertrauen in meine Fähigkeiten, sowie die fast uneingeschränkten Freiheiten, die er mir während meiner Arbeit gestattet hat, herzlich bedanken.

Allerdings waren meine Doktorandenkollegen bei dieser Dissertation wohl ebenso beteiligt wie meine Betreuer und auch ich selbst. Für die fachlichen wie auch privaten Diskussionen, die regelmäßigen Motivationsschübe und die unvergesslichen "Betriebsausflüge" nach Schwabing möchte ich mich deswegen an zweiter Stelle bei Jörn Leonhardt, Johannes Raffler, Florian Büttner und Daniel Ellwanger bedanken.

Besonderer Dank gilt auch Gabi Kastenmüller, die meine direkte Betreuung übernommen hat, als mein ehemaliger Gruppenleiter das Institut verließ. Arne Pfeufer danke ich für die vielen Diskussionen und Ideen. Ohne die Unterstützung dieser beiden wären die meisten Projekte im Rahmen dieser Doktorarbeit nicht möglich gewesen, daher fällt ein Gutteil der Qualität der Arbeit auf sie zurück. Meinem Zweitprüfer Fabian Theis möchte ich für die positiven, motivierenden und anregenden Kommentare zu meiner Arbeit danken.

Den zahlreichen internen und externen Kooperationspartnern, die meine Projekte unterstützt haben oder die mich in Ihre Projekte involviert haben, muss ich ebenfalls meinen Dank aussprechen: Jan Krumsiek vom ICB, Karsten Suhre vom IBIS (und aus Katar), Matthias Wjst vom ILBD und Christian Gieger vom IGE am Helmholtz Zentrum München; Elke Rodriguez, Stephan Weidinger und Hansjörg Baurecht vom Uniklinikum Kiel; Nicole Soranzo vom Wellcome Trust Sanger Institute; So-Youn Shin von der University of Bristol; Stefan Herms von der Universität Basel; und Thilo Dörk-Bousset von der Medizinischen Hochschule Hannover.

Auch möchte ich "meinen" Studenten Kinga Balázs, Christoph Schramm, David Fuggersberger, Quirin Heiß, Niklas de Andrade Krätzig und Nick Lehner für die angenehme Zusammenarbeit, sowie allen nicht namentlich genannten Kollegen von IBIS und ICB für das kollegiale Umfeld danken.

Dank an Familie und Freunde möchte ich mir hier sparen, da eine Nennung auf einem Stück Papier nicht ausdrücken kann, was gesagt werden müsste. An meine Eltern möchte ich aber doch noch ein Zitat richten, da sie lange genug darauf gewartet haben:

"Ich habe fertig." (G. Trapattoni)

Abstract

Genome-wide association studies (GWAS) are an effective tool to map genetic regions contributing to multifactorial human traits and diseases and yielded a catalog of thousands of robust associations. The major recurring point of criticism with regards to the GWAS approach is that the obtained loci are of only limited value because in most cases the associations can neither be linked to a plausible causal gene nor provide information on the molecular background involved in trait development and progression. This thesis provides a detailed investigation of the challenges arising from this issue and proposes various evidenced-based and integrative computational approaches as well as a novel bioinformatics tool that enable comprehensive functional characterization of GWAS loci and thereby facilitate the elucidation of potential mechanisms underlying genotype-trait associations.

The first part of the thesis describes three GWA studies that have been conducted during this work to identify and characterize specific challenges in the interpretation of GWAS results. The first study investigates the influence of common genetic variants and rare copy number variants (CNVs) on sudden infant death syndrome (SIDS). While the results showed only indicative evidence for weak additive effects of common variants on SIDS risk, analysis of CNVs revealed rare deletion syndromes as likely causes of sudden infant death for a substantial number (12 of 301) of the cases. Two further GWAS focused on common genetic variants influencing the concentration of metabolites in human blood and urine samples. Here, we identified and replicated more than 150 genetic loci, thus providing a large compendium of genomic regions implicated in the genetic control of human metabolic homeostasis. In addition to the central study results, I illustrate the challenges associated with the GWAS approach by showing the complexity of interpreting weak genetic influences on extreme disease endpoints such as SIDS. In the GWAS on blood metabolic traits, I then emphasize the utility of thorough manual annotation of genetic associations to identify the most plausible causal gene and to suggest a testable effect hypothesis for each identified locus. Finally, in the urine metabolomics GWAS, I propose a method to automate the identification of predicted causal genes using a straightforward evidence-based gene prioritization metric.

To enable and facilitate automated causal gene prediction, in the second part of this thesis I developed an extensive data integration resource. This resource, representing the first genetic variant-based genome browser, allows for comprehensive annotation of the impact of genetic variation using evidence-based variant effect predictions. In the development process, I integrated, harmonized, and consolidated genome-wide annotation data from various sources

comprising genes, transcripts, proteins, genetic variants, regulatory elements including microRNA binding sites, enhancers and promoters, a set of genome- and exome-wide conservation and deleteriousness scores, as well as a large collection of trait annotations and associations for genes and genetic variants. The browser is extended by modules for the analysis, aggregation, and visualization of genomic annotations linked to genetic variants on a genome-wide scale. The resource thus provides both interfaces to the collected data and semantic categorization of the available variant-linked evidences in logical sections, which enables direct hypothesis generation using the modules' output.

In the third and final part of the thesis, I demonstrate the value of integrative bioinformatics approaches by utilizing the data incorporated in this resource to shed light on the potential molecular consequences of genetic variants identified by GWAS from three perspectives. In the first study, I present the concept of biological networks by integrating genetic variants and their previously collected associated diseases in a directed bipartite network. Analysis of this network showed that identical genetic loci frequently influence several different complex diseases both in agonistic and antagonistic effect directions. It is a yet unsolved question if such loci are to be considered *pleiotropic* featuring conditionally distinct functions, or if they pinpoint the same nodes in a cellular pathway that, in dependency of further genetic and environmental influences, lead to diverging phenotypic endpoints. The shared association signal observed for melanoma and vitiligo located in the *tyrosinase* gene, which has a central function in skin pigmentation, serves as an example for the former hypothesis. Here, the allelic effects suggest inverse trait-specific antigenicity of the encoded *TYR* protein that results in skin pigmentation being either elevated (as in skin cancer) or depleted (as in vitiligo) depending on allelically determined active or inactive targeting of *TYR* antigens via immune surveillance. The second study investigates the collected target sites of microRNAs, a special class of small non-coding RNAs involved in post-transcriptional gene regulation, for interrelations with trait-associated genetic variants. I demonstrate that trait-associated variants are significantly enriched in the 3'-untranslated region of human transcripts, which presents the major targeting region of microRNAs. Using the results of the blood metabolomics GWAS, I show that for a large fraction (>10%) of genetic loci linked to metabolic traits there is evidence for the involvement of genetically influenced microRNA regulation in metabolic control. The very specific mechanism described for genetic alteration of *lipoprotein lipase*-controlled lipid homeostasis by modulating its functioning potential via allele-dependent targeting of its transcripts by *miR-410* underlines the value of this approach. The third study explores regulatory effects of genetic

variants affecting the promoter and enhancer elements contained in the developed variant annotation resource. For the purpose of characterizing allele-specific effects on gene regulation, I used a novel clustering of cross-tissue regulatory element annotations. It is shown that the information aggregated within clusters can reveal direct interactions between enhancer elements, specific transcription factors, and the expression of more distal genes. The utility of the derived clusters in predicting allele-specific modifications of gene regulation is exemplified by a genetic locus from our blood metabolomics GWAS that is associated with *alpha-hydroxyisovalerate* levels. The associated haplotype is predicted to alter the binding motif of the *Myc/Max* transcription factor complex in a distal promoter-associated enhancer, leading to experimentally validated allele-specific changes of *lactate dehydrogenase A* expression. Combination with additional metabolic and enzymatic evidences further indicates a potential *pleiotropic* role of the encoded dehydrogenase in aerobic branched-chain amino acid and anaerobic lactate metabolism.

In summary, this thesis provides a detailed motivation for the application of large-scale integrative approaches in human genetic studies, illustrated using the findings of three GWA studies. With the implementation of a free-to-use, extensible, updatable, and programmatically accessible data integration resource, I introduce a novel bioinformatics platform that meets the requirements of integrative methods for causal gene prediction in the GWAS context in a comprehensive, yet user-friendly, way. In three studies covering different aspects of the molecular consequences introduced by genetic variation, I finally demonstrate that integrative methods based on this resource successfully mark novel, specific, as well as testable hypotheses for further investigation.

Zusammenfassung

Genomweite Assoziationsstudien (GWAS) stellen eine effektive Methode zur Verknüpfung von genetischen Regionen mit multifaktoriellen menschlichen Merkmalen und Erkrankungen dar und haben eine Sammlung von mehreren tausend statistisch robusten Assoziationen geschaffen. Der zentrale, immer wieder aufgegriffene Kritikpunkt an der GWAS-Methode ist, dass die identifizierten Regionen nur von begrenztem Nutzen sind, da die Assoziationen in vielen Fällen weder mit einem kausalen Gen verbunden werden können, noch Rückschlüsse auf die molekularen Vorgänge, die der Merkmalsentwicklung und -progression zugrunde liegen, zulassen. Die vorgelegte Arbeit schildert die Herausforderungen, die sich aus dieser Problematik ergeben, im Detail und führt mehrere unterschiedliche evidenzbasierte und integrative computergestützte Verfahren sowie eine webbasierte bioinformatische Umgebung an, die die umfangreiche funktionelle Charakterisierung von GWAS-identifizierten Regionen erlauben und damit die Aufdeckung potentieller Mechanismen, die den Genotyp-Merkmal-Assoziationen zugrunde liegen, unterstützen.

Der erste Teil der vorgelegten Arbeit beschreibt drei, im Verlauf dieses Projektes durchgeführte GWA-Studien, um die spezifischen Herausforderungen bei der Interpretation von GWAS-Ergebnissen zu identifizieren und zu erläutern. Die erste Studie untersucht den Einfluss häufiger genetischer Varianten und seltener Kopienzahlvariationen auf den plötzlichen Kindstod. Während die Ergebnisse nur andeutungsweise Evidenzen für schwache additive Effekte häufiger Varianten auf das Risiko für den plötzlichen Kindstod zeigten, ergab die Analyse der Kopienzahlvariationen, dass ein durchaus substantieller Anteil der Kindstodsfälle (12 von 301) wahrscheinlich von seltenen Deletionssyndromen verursacht wurde. Die zweite und dritte GWAS wendeten sich der Untersuchung von häufigen genetischen Varianten zu, die die Konzentrationen von Stoffwechselprodukten im menschlichen Blut und Urin beeinflussen. Mit über 150 genetischen Bereichen, die hier identifiziert und repliziert wurden, stellen diese beiden Studien eine große Sammlung von Bereichen des Genoms, die in der Steuerung des menschlichen Stoffwechsels involviert sind, zur Verfügung. Zusätzlich zu den zentralen Studienergebnissen werden in diesem Teil die Schwierigkeiten, die mit der GWAS-Methode assoziiert werden, anhand der Komplexität der Interpretation von schwachen genetischen Einflüssen auf extreme Krankheitsbilder wie dem plötzlichen Kindstod beleuchtet. In der GWAS zu Blutstoffwechselmerkmalen wird daraufhin die Bedeutung intensiver manueller Annotation genetischer Assoziationen für die Identifikation des naheliegensten kausalen Gens sowie für die Formulierung einer testbaren Effekthypothese für jeden der assoziierten Bereiche

herausgestellt. Schließlich wird im Kontext der GWAS zu Stoffwechselprodukten im Urin erklärt, wie die Vorhersage des kausalen Gens für einen Bereich mit einem geradlinigen evidenzbasierten Maß für die Gen-Priorisierung automatisiert werden kann.

Um diese automatisierte Vorhersagemethode kausaler Gene einfach zugänglich zu machen, wurde im zweiten Teil dieser Arbeit eine umfangreiche Datenintegrationsplattform angelegt. Diese Plattform, die den ersten variantenbasierten Genom-Browser darstellt, ermöglicht die umfassende Annotation der Auswirkungen genetischer Variation durch die evidenzbasierte Vorhersage von Varianteneffekten. Im Entwicklungsprozess wurden dabei genomweite Annotationsdatensätze zu Genen, Transkripten, Proteinen, genetischen Varianten, regulatorischen Elementen einschließlich microRNA-Bindestellen, Enhancern und Promotoren, mehreren genom- und exomweiten Konservierungs- und Schädlichkeitsscores, sowie Merkmalsassoziationen und -annotationen für Gene und genetische Varianten aus unterschiedlichsten Quellen integriert und einheitlich zusammengeführt. Der Browser wird durch Module, die die Analyse, Aggregation und Visualisierung von mit genetischen Varianten verknüpften Annotationen auf genomweiter Ebene ermöglichen, komplettiert. Die Plattform stellt dadurch sowohl Schnittstellen zur Datensammlung als auch eine semantische Kategorisierung der für die Varianten verfügbaren Evidenzen in logische Abschnitte zur Verfügung, wodurch die direkte Generierung von Hypothesen aus der Ausgabe der einzelnen Module ermöglicht wird.

Im dritten und letzten Teil der Arbeit wird der Nutzen integrativer bioinformatischer Ansätze unter Benutzung dieser Datensammlung anhand dreier Anwendungen verdeutlicht, die verschiedene Aspekte potentieller Auswirkungen genetischer Varianten auf molekularer Ebene im Kontext GWAS-identifizierter genetischer Assoziationen beleuchten. Die erste Untersuchung führt dazu das Konzept biologischer Netzwerke ein, wobei genetische Varianten und deren gesammelte Krankheitsassoziationen in einem gerichteten, bipartiten Netzwerk verknüpft werden. Die Analyse dieses Netzwerkes ergab, dass dieselben genetischen Bereiche häufig mehrere unterschiedliche komplexe Erkrankungen beeinflussen, und zwar sowohl in gleichgerichteten als auch in entgegengesetzten Effektrichtungen. Bis jetzt konnte die Frage nicht global geklärt werden, ob solche Loci als *pleiotrop* anzusehen sind, also konditional unterschiedliche Funktionen ausüben, oder ob hier derselbe Knoten eines zellulären Pfades betroffen ist, der in Abhängigkeit anderer genetischer und umweltbedingter Einflüsse zu unterschiedlichen phänotypischen Endpoints führt. Die überlappenden Assoziationssignale für die Bildung von Melanomen und Vitiligo am *Tyrosinase*-Gen, das eine zentrale Funktion in

der Hautpigmentierung ausübt, dient hier als Beispiel für die erste Hypothese. Die Alleleffekte deuten in diesem Fall eine gegensätzliche, von der jeweiligen Krankheit abhängige Antigenizität des kodierten *TYR*-Proteins an, die dazu führt, dass die Pigmentierung der Haut abhängig von der aktivierten oder deaktivierten Immunantwort auf *TYR*-Antigene entweder erhöht (bei Hautkrebs) oder verringert (bei Vitiligo) wird. Die zweite Studie untersucht die gesammelten Bindestellen von microRNAs, einer speziellen Klasse von post-transkriptional aktiven Molekülen, auf Beeinflussung durch merkmalsassoziierte genetische Varianten. Hier wird nachgewiesen, dass merkmalsassoziierte Marker in der 3'-untranslatierten Region menschlicher Transkripte, welche die zentral von microRNAs anvisierten Bereiche darstellen, statistisch signifikant angereichert auftreten. Unter Benutzung der Ergebnisse der GWAS zu Blutstoffwechselmerkmalen wird gezeigt, dass es für einen großen Teil (>10%) der genetischen Bereiche, die mit Konzentrationsänderungen von Stoffwechselprodukten assoziiert sind, Hinweise auf genetisch bedingte Veränderungen der microRNA-Regulation gibt, die in der Steuerung des Stoffwechsels involviert sind. Mit dem sehr spezifischen Mechanismus, der für die genetische Veränderung der von der *Lipoprotein Lipase* gesteuerten Lipid-Homöostase beschrieben wird, nämlich ein modifiziertes Funktionspotential, das durch die allelspezifische Regulation der Transkripte durch *miR-410* kontrolliert wird, unterstreicht die Aussagekraft des vorgestellten Ansatzes. Die dritte Studie beschäftigt sich mit regulatorischen Effekten genetischer Varianten, die die in der vorher beschriebenen Plattform enthaltenen Promoter- und Enhancer-Elemente beeinträchtigen. Dazu wird ein neuer Clustering-Ansatz zur Gruppierung von gewebeübergreifenden Annotationen regulatorischer Elemente für die Charakterisierung Allel-spezifischer Effekte auf die Genregulation verwendet. Wie belegt wird, können die in den Clustern zusammengefassten Informationen direkte Verbindungen zwischen Enhancer-Elementen, spezifischen Transkriptionsfaktoren und der Expression entfernterer Gene aufdecken. Die Anwendbarkeit der Cluster in der Vorhersage von allelspezifischen Modifikationen der Genregulation wird mit einer Region, die in der GWAS zu Blutstoffwechselmerkmalen mit der Konzentration von *alpha-Hydroxyisovalerat* assoziiert ist, belegt. Laut Vorhersage beeinträchtigt der assoziierte Haplotyp das Bindemotif des *Myc/Max*-Transkriptionsfaktorkomplexes in einem distalen, Promotor-assoziierten Enhancer, was zu einer Allel-spezifischen, experimentell belegten Änderung des Expressionsniveaus der *Laktatdehydrogenase A* führt. Das Heranziehen weiterer Evidenzen zu Stoffwechselprodukten und Enzymen deutet darüber hinaus eine potentiell *pleiotrope* Funktion für diese

Dehydrogenase im aeroben Abbau der verzweigtkettigen Aminosäuren im Kontrast zur ihrer klassischen Funktion im anaeroben Laktatstoffwechsel an.

Zusammengefasst liefert die vorgelegte Arbeit eine detaillierte Motivation für die Anwendung von umfangreichen integrativen Analyseansätzen in humangenetischen Studien, die Anhand der Ergebnisse dreier GWA-Studien veranschaulicht wird. Durch die Implementierung einer kostenlosen, erweiter- und aktualisierbaren, sowie programmatisch ansteuerbaren Datenintegrationsplattform wird eine neue Bioinformatik-Schnittstelle bereitgestellt, die die Voraussetzungen zur Entwicklung integrativer Methoden zur Vorhersage von kausalen Genen im GWAS-Kontext umfassend und dennoch anwenderfreundlich zur Verfügung stellt. In drei Studien zu verschiedenen Aspekten der molekularen Auswirkungen genetischer Varianz wird abschließend gezeigt, dass sich mit integrativen Methoden, die auf dieser Plattform aufsetzen, erfolgreich neue sowohl spezifische als auch überprüfbare Hypothesen zur weiteren Erforschung kennzeichnen lassen.

Scientific Publications

The following list specifies peer-reviewed publications written in the course of this thesis:

* Fard D, Läer K, Rothämel T, Schürmann P, Arnold M, Cohen M, Vennemann M, Pfeiffer H, Bajanowski T, Pfeufer A, Dörk T, Klintschar M. **Candidate gene variants of the immune system and Sudden Infant Death Syndrome**. *International Journal of Legal Medicine*, 130(4):1025-33, 2016.
* Raffler J, Friedrich N, Arnold M, Kacprowski T, Rueedi R, Altmaier E, Bergmann S, Budde K, Gieger C, Homuth G, Pietzner M, Römisch-Margl W, Strauch K, Völzke H, Waldenberger M, Wallaschofski H, Nauck M, Völker U, Kastenmüller G, Suhre K. **Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality**. *PLoS Genetics*, 11 (9):e1005487, 2015.
* Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. ***SNiPA*: an interactive, genetic variant-centered annotation browser**. *Bioinformatics*, 31 (8):1334-6, 2015.
* Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, Arnold M, Erte I, Forgetta V, Yang TP, Walter K, Menni C, Chen L, Vasquez L, Valdes AM, Hyde CL, Wang V, Ziemek D, Roberts P, Xi L, Grundberg E, The MuTHER Consortium, Waldenberger M, Richards JB, Mohney RP, Milburn MV, John SL, Trimmer J, Theis FJ, Overington JP, Suhre K, Brosnan MJ, Gieger C, Kastenmüller G, Spector TD, Soranzo N. **An atlas of genetic influences on human blood metabolites**. *Nature genetics*, 46 (6):543-550, 2014.
* Wjst M, Sargurupremraj M, Arnold M. **Genome-wide association studies in asthma: what they really told us about pathogenesis**. *Current Opinion in Allergy and Clinical Immunology*, 13 (1):112-118, 2013.
* Arnold M, Hartsperger ML, Baurecht H, Rodríguez E, Wachinger B, Franke A, Kabesch M, Winkelmann J, Pfeufer A, Romanos M, Illig T, Mewes HW, Stümpflen V, Weidinger S. **Network-based SNP meta-analysis identifies joint and disjoint genetic features across common human diseases**. *BMC Genomics*, 13 (1):490, 2012.
* Arnold M, Ellwanger DC, Hartsperger ML, Pfeufer A, Stümpflen V. ***Cis*-Acting Polymorphisms Affect Complex Traits through Modifications of MicroRNA Regulation Pathways**. *PLoS ONE*, 7 (5):e36694, 2012.

Contents

1 Introduction

In the 21st century, applied genetics or genetic engineering can be found everywhere. The flowers in our backyard, the dairy cows at the neighboring farmstead, the detergents for the washer, or agricultural products at the supermarket – all of them are the results of breeding *(artificial selection)* or targeted genetic modifications. This goes so far that, in the context of the Transatlantic Trade and Investment Partnership between the European Union, the United States of America, and other countries, the labeling obligation of genetically modified maize and soybeans as such is a matter of public dispute. However, while the breeding of plants and animals had its beginning with the sedentariness of humans (at least 15,000 years before present [1]), it was not before the late 18th century that the first scientists laid the cornerstones for our current understanding of genetics. Until then, the prevailing view of the origin of life in the Western world was dictated by creationists postulating the work of a divine creator. In his 1809 paper, JEAN-BAPTISTE LAMARCK was the first to suggest that living things had evolved over time:

> *"EVERYTHING IN TIME UNDERGOES VARIOUS MUTATIONS, MORE OR LESS RAPID ACCORDING TO THE NATURE OF THE OBJECTS AND THE CONDITIONS; (…) IN SHORT, EVERYTHING ON THE SURFACE OF THE EARTH CHANGES ITS SITUATION, SHAPE, NATURE AND APPEARANCE, AND EVEN CLIMATES ARE NOT MORE STABLE."*
>
> **Jean-Baptiste Lamarck, 1809 [2]**

LAMARCK not only stated that evolution takes place via mutations, but also that it is influenced by several factors, including environmental conditions [2]. Although his conclusions were widely accepted by academics and scientists, the Christian churches as well as most Western states interpreted them as a threat to the social establishment. The access to existing evidences for evolution was very limited and this led to what is known as the creation-evolution controversy, a heated debate about the origin of species. This dispute was put to an end by CHARLES ROBERT DARWIN, when he laid out his interpretation of the many evidences he brought back to Europe from the voyage of H.M.S Beagle during the years 1832-1836. In his "Origin of Species" (1859), DARWIN took LAMARCK'S realizations to the next level by hypothesizing that all life may have developed from one common ancestor (for a timeline of the following discoveries, see Figure 1). Using his enormous collection of specimens on animals and plants, DARWIN deduced an evolution theory by means of *natural selection*: he concluded that there are differences between individual organisms induced by random mutations and that only the ones increasing fitness to environmental challenges get fixed in an organism's population pool (they are *selected* and *inherited* to following generations), while disadvantageous variants are lost by genetic drift [3]. However, DARWIN was not able to deduce a hypothesis regarding the mechanisms of inheritance from his natural selection theory:

> *„I HAVE HITHERTO SOMETIMES SPOKEN AS IF THE VARIATIONS – SO COMMON AND MULTIFORM IN ORGANIC BEINGS (…) – HAD BEEN DUE TO CHANCE. THIS, OF COURSE, IS A WHOLLY INCORRECT EXPRESSION, BUT IT SERVES TO ACKNOWLEDGE PLAINLY OUR IGNORANCE OF THE CAUSE OF EACH PARTICULAR VARIATION."*
>
> **Charles R. Darwin, 1859 [3]**

1.1 The beginnings of genetics

It was the research of the Austrian monk JOHANN GREGOR MENDEL in the 1860's that illustrated the mechanisms of heredity [4]. In thousands of cross-breeding experiments with pea plants MENDEL recorded the phenotypic outcomes. From his observations he concluded that there is always a pair of units of inheritance *(alleles)* for each trait. In gamete (germ cell)

formation, the two alleles segregate (are separated) so that only one allele per trait is left (*principle of segregation*). This happens for each trait independently (*principle of independent assortment*). During reproduction, the two parental alleles are recombined such that each parent transmits one allele for each trait to the offspring. There are dominant (A) and recessive (a) alleles, with a total of four possible allele combinations for each trait: AA, Aa, aA, and aa (as the phenotypic outcomes of the heterozygotes Aa and aA are identical, the two combinations are often summarized as 2Aa). On the phenotype level, the dominant trait occurs with a ratio of 3:1, meaning that the dominant allele has full phenotypic penetrance (outmatches the recessive allele if present; *principle of dominance*). The significance of MENDEL'S findings was not recognized in his lifetime. It was not before the year 1900 that his work was rediscovered and highlighted independently by three researchers: HUGO DE VRIES [5], CARL CORRENS [6], and ERICH VON TSCHERMAK [7]. In addition to emphasizing the importance of MENDEL's conclusions, CORRENS also claimed that the matter of inheritance has to be located in the cell nucleus and that segregation of alleles takes place during a *"NUCLEAR DIVISION"* [6].

That the nucleus of cells contains special molecules (discoverer FRIEDRICH MIESCHER called them *nuclein*) which were later found to consist of deoxyribonucleic acid (DNA) was known since 1871 [8], however, neither the identity and structure of the molecules nor that macromolecules consisting of DNA encode genetic information was yet discovered. Around 1880, WALTHER FLEMMING found that *nuclein* was organized in densely packed threadlike structures during cell division [9]. As he was – in contrast to MIESCHER – not persuaded that the substance was specific to the nucleus, he called it *chromatin* because of its stainability. FLEMMING also observed that the threadlike structures (later called *chromosomes* by WILHELM VON WALDEYER-HARTZ [10]) were separated during cell division. As he was not familiar with the findings of MENDEL, he did not recognize the implications of his findings to inheritance. It were WALTER SUTTON and THEODOR BOVERI – the latter had already shown that chromosomes persist as organized and individual structures during cell division as well as that paternal and maternal gametes contribute an equal number of chromosomes to the embryonic cells [11, 12] – who finally connected the Mendelian principles of inheritance with chromosomes as carrier substance of hereditary factors (later called *genes* by WILHELM JOHANSSEN [13]) around 1902 [14–16]. Soon afterwards, the first human disorders were classified as Mendelian traits: in 1902, ARCHIBALD GARROD laid out his theory that biochemical disorders (he later termed such disorders *inborn errors of metabolism* [17]) such as alkaptonuria (MIM: 203500) may be recessively inherited following the Mendelian pattern [18]

and, three years later, WILLIAM FARABEE deduced dominant inheritance for brachydactyly (shortness of fingers and toes; MIM: 112500) [19].

In spite of the growing number of examples where MENDEL'S theories fitted evidence, still several principles and findings – both theoretical and observational – stood in conflict with the Mendelian model of heredity.

On the observational side, WILLIAM BATESON (maybe the most passionate proponent of MENDEL'S ideas) together with EDITH SAUNDERS and REGINALD PUNNETT found in 1905 that not all pea phenotypes follow MENDEL'S principle of independent assortment but that there exists a *"CORRELATION OR 'COUPLING'"* [20] of their occurrence (now called *genetic linkage*). What BATESON and colleagues failed to recognize in their experiments was that linked genetic factors together are again inherited in accordance to the Mendelian scheme. It was THOMAS H. MORGAN who achieved this result in 1910 when he investigated crossing results of a male fruit fly (*D. melanogaster*) with white instead of red eye color as seen in the wild-type [21]. Previously, NETTIE STEVENS [22] and EDMUND WILSON described a pair of chromosomes that, while paired normally in females, in males consisted of one normal-sized chromosome (the *X chromosome*) paired with one smaller *"'ACCESSORY' OR HETEROTROPIC ONE"* [23] (the *Y chromosome*). Both STEVENS and WILSON correctly concluded that this chromosome pair determines the sex of an organism, in humans with the combination XX for females and XY for males. When MORGAN crossed the white-eyed male fly, the offspring (F_1) was purely red-eyed. When inbreeding F_1 flies, the second generation (F_2) showed flies with both eye colors in the Mendelian ratio 3 (red) : 1 (white) for dominant traits. However: all white-eyed flies were males and, thus, the heredity factor of white eyes was shown to be linked to the sex-determining chromosomes. In the next experiment, MORGAN crossed the original white-eyed mutant with its F_1 daughters, and this cross produced males and females with both red and white eyes in the ratio 1 (female/red) : 1 (female/white) : 1 (male/red) : 1 (male/white) [21]. Thus, MORGAN not only proved the existence of a recessively inherited X chromosome-linked trait, but also established Mendelian inheritance for linked phenotypes. In the following years, MORGAN and his co-workers (prominently his student ALFRED H. STURTEVANT) also identified the source of genetic linkage: recombination of parts of chromosomes via crossover [24]. The theory states that heredity factors (i.e. genes) are in fact physical objects linearly positioned on the chromosomes and the smaller the distance between two genes the more often they are inherited together. As the structure and the code encrypted in the chromosomes was still not resolved (a *physical distance* was not available), STURTEVANT defined the *genetic distance* between two

genes as the frequency of recombination events that separated the genes per 100 gamete formations, a measure that was later termed *centiMorgan (cM)* by JOHN HALDANE [25, 26].

On the theoretical side, it was mainly a dispute of two schools with different evolutionary theories: *Gradualism / the biometric school* (evolution happens through gradual changes based on the Darwinian theory of a common ancestor and natural selection; model of inheritance – the *law of ancestral heredity* – created by FRANCIS GALTON [27] and mathematically elaborated by KARL PEARSON [28]) and *Saltationism / the Mendelian school* (intra-species evolution/adaption takes place gradually, but new species are the result of sudden, crucial alterations, not natural selection; Mendelian model of inheritance). G. UNDY YULE, a statistician and former co-worker of PEARSON, expressed substantial points of criticism of the biometric school in his 1902 paper. He postulated that: firstly, if MENDEL'S principle of dominance should apply, the dominant alleles should at some point supersede their recessive counterparts if evolutionary forces are absent; and secondly, human traits as complex as body height that are determined by *"NOT LESS THAN 50"* [29] genes could only be explained by the Mendelian laws if they are, due to the large number of possible combinations, changed gradually, not saltatory. YULE concluded that Mendelian inheritance (mainly because it was propagated by saltationists), while holding good for some traits (in his opinion limited to some hybridization experiments), cannot be readily transferred to inheritance patterns in general populations.

REGINALD PUNNETT was opposed to YULE'S first hypothesis [30], however, he was not able to derive a mathematical formula to prove YULE'S reckoning wrong. In 1908, the mathematician GODFREY H. HARDY and the physician WILHELM WEINBERG independently presented a mathematical equilibrium of allele frequencies in the absence of evolutionary forces (now called the *Hardy-Weinberg equilibrium* or HWE). While WEINBERG observed the equilibrium when he was studying the accumulation of twin-births in families [31], HARDY had read PUNNETT'S paper and, thus familiar with YULE'S calculations, found it *"NOT DIFFICULT TO PROVE (…) THAT SUCH AN EXPECTATION WOULD BE QUITE GROUNDLESS"* [32]. Based on the assumption of an ideal or random population (a large population consisting of individuals of equal fitness where mating is fully random and neither mutations, gene flow, genetic drift nor natural selection appear), HARDY deduced the model as follows:

The numbers of homozygotes of the dominant allele (AA), heterozygotes (Aa), and homozygotes of the recessive allele (aa) are considered as $p : 2q : r$. Under random mating, the numbers in the next generation are

$$(p+q)^2 : 2(p+q)(q+r) : (q+r)^2$$

or $p_1 : 2q_1 : r_1$. HARDY simply states that the condition for this distribution to be stable is $q^2 = pr$. And as, independently of the explicit values of $p, q,$ and r, $q_1^2 = p_1 r_1$, "*THE DISTRIBUTION WILL IN ANY CASE CONTINUE UNCHANGED AFTER THE SECOND GENERATION*" [32]. While the latter conclusion is obviously correct

$$\underbrace{((p+q)(q+r))^2}_{q_1^2} = \underbrace{(p+q)^2}_{p_1}\underbrace{(q+r)^2}_{r_1}$$

the condition for the equilibrium is not that straightforward to see. However, transferring the numbers of individuals to the frequencies of the alleles, like WEINBERG did, shows that the concept is rather intuitive: the distribution of $p : 2q : r$ is then expressed as $P[A]^2 : 2P[A]P[a] : P[a]^2$ with $P[A] + P[a] = 1$ (here, obviously $q^2 = pr$). In the next generation (under random mating and no selection) the relative frequencies are $P[A]^2(P[A] + P[a])^2 : 2P[A](P[A] + P[a])^2 P[a] : (P[A] + P[a])^2 P[a]^2$ and therefore $((P[A] + P[a])^2 = 1^2 = 1)$ again $P[A]^2 : 2P[A]P[a] : P[a]^2$.

While the formulation of the Hardy–Weinberg equilibrium was substantial as it proved that, in a large, interbreeding population, allelic frequencies remain stable without evolutionary forces active, its reversion is equally interesting, because deviations from the HWE are evidence for either inbreeding or the presence of evolutionary pressure on alleles – something that, until then, was not possible.

1.2 The foundation of population genetics

The second point of criticism stated by YULE was highly controversial, mainly due to the hardened fronts between biometricians and Mendelians. It was the groundbreaking 1918 paper of RONALD A. FISHER that not only mathematically integrated Mendelism and biometry by modeling the genetic basis of quantitative traits as being determined by many Mendelian factors but also laid the ground for many of the principles of population genetics in place [33]. Although the basis for FISHER'S findings originated from many other researchers (including the aforementioned BATESON, MORGAN, YULE, GALTON, PEARSON, HARDY, MENDEL, and

DARWIN), he was the first to combine all the evidences and theories into one generalized mathematical framework.

FISHER starts with the simplifying assumption that genetic factors for a complex trait (a trait that is determined by many genetic factors) act together additively, that is, they are not interacting and thus independent in the sense of MENDEL, and that the environment plays no role in trait variance. The statistical benefit of additivity is that, for measurements of many independent random variables (here: the genetic factors as causes of trait variability) with finite means and variance, the central limit theorem applies, stating that the arithmetic mean of such measurements is distributed approximately normal. Thus, the complete variance of the trait due to genetic factors (the environment is neglected) in the population, σ^2, can be calculated by summing up the variances α_i^2 contributed by each individual genetic factor i as $\sigma^2 = \sum_{i=1}^{n} \alpha_i^2$. The effect of each factor in its three forms (AA, Aa, and aa) is then calculated as the deviation from the *"SOMATIC"* [33] (phenotypic) midpoint of the two homozygous phenotypes (AA and aa) denoted as $+a$ and $-a$, respectively. To account for potential dominant effects, the heterozygote can differ from the exact midpoint by a quantity d ($d = 0$: no dominance; $0 < d < a$: partial dominance; $d = a$: full dominance). Assuming the HWE (random mating, no selection), genotype frequencies are $p, 2q, r$ with $p + 2q + r = 1$. The population mean for one single genetic factor is then $m = pa + 2qd - ra$ and its variance is the quadratic form $\alpha^2 = p(a - m)^2 + 2q(d - m)^2 + r(a + m)^2$. Next, substituting $+a$, d, and $-a$ with linear quantities, FISHER – now accounting for dominance and environmental influences leading to imperfect additivity of effects – splits additive genetic effects (β^2) from non-additive effects (δ^2) via least squares, obtaining $\sigma^2 = \tau^2 + \epsilon^2$ with $\tau^2 = \sum_{i=1}^{n} \beta_i^2$ and $\epsilon^2 = \sum_{i=1}^{n} \delta_i^2$. In the non-additive effects, deviations from linearity resulting from genetic interactions *(epistasis)* are also included due to their *"SIMILAR STATISTICAL EFFECTS TO DOMINANCE"* [33]. Going even more into complexity, the paper also derives formulae to account for assortative mating, introducing factors $f_{n,m}$ that depict deviations from HWE-defined stable allele frequencies (and the correlations between relatives, respectively) due to inbreeding. The same symbol is later used by SEWALL WRIGHT to formulate the *inbreeding coefficient* that, in dependency of the extent of assortative mating, in its simplest form gives the allele frequencies as

$$\begin{aligned} AA &: P[A]^2(1-f) + P[A]f \\ Aa &: 2P[A]P[a](1-f) \\ aa &: P[a]^2(1-f) + P[a]f \end{aligned}$$

with $f = 0$ for random mating, $f = 0.5$ for full siblings in a random population, and $f \to 1$ (loss of heterozygosity) for individuals in closely inbred populations [34].

To summarize, based on the numbers for genetic correlations between individuals in randomly mating populations (Table 1), FISHER establishes the analysis of variance (ANOVA) which is still used to disassemble the factors involved in the variability of heritable traits. Using the derived formulae, FISHER shows that his basic hypothesis that complex traits *"ARE DETERMINED BY A LARGE NUMBER OF MENDELIAN FACTORS, AND THAT THE LARGE VARIANCE AMONG CHILDREN OF THE SAME PARENTS IS DUE TO THE SEGREGATION OF THOSE FACTORS IN RESPECT TO WHICH THE PARENTS ARE HETEROZYGOUS"* [33] complies surprisingly well with population-based anthropometric data. Thus, he created a theoretical framework that allowed for the assessment of genetic effects on complex trait variation in the general population, not only in nuclear families or larger pedigrees.

Generations	**Ancestral Line**	**1st degree relatives**	**2nd degree relatives**	**3rd degree relatives**	**4th degree relatives**
F_0	1	$1/2$	$1/4$	$1/8$	$1/16$
F_1	$1/2$	$1/4$	$1/8$	$1/16$	$1/32$
F_2	$1/4$	$1/8$	$1/16$	$1/32$	$1/64$
F_3	$1/8$	$1/16$	$1/32$	$1/64$	$1/128$
F_4	$1/16$	$1/32$	$1/64$	$1/128$	$1/256$

Table 1: **Genetic correlations between individuals in random mating populations**.

The current notation of the different summands for the analysis of variance has changed slightly since FISHER and the term *heritability* has been introduced for the proportion of the phenotypic variance of complex traits contributed by the variance of genetic factors, but the statistical idea has remained the same. To define heritability (and, later on, models to estimate the explained variance), we will at this point shortly introduce the current notations for the partitioned variances.

The observed phenotype (P) is the result of unknown genetic (G) as well as environmental factors (E) acting together: $P = G + E$, or in terms of variances:

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2.$$

As suggested by FISHER, the variance of both genetic and environmental factors can be separated further. The genetic variance is partitioned into the variance of additive genetic effects σ_A^2, the variance of dominance σ_D^2, and the variance of epistasis/genetic interactions σ_I^2:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2.$$

Similarly, environmental effects can be partitioned into environmental variance shared between individuals such as siblings, σ_C^2, and the residual (individual) environmental variance including measurement error, σ_{RE}^2:

$$\sigma_E^2 = \sigma_C^2 + \sigma_{RE}^2.$$

Because it is often not possible to determine specific environmental factors contributing to the environmental variance, the symbols for environmental factors and individual environmental factors (E and RE) are used interchangeably (like FISHER did). For instance, in the twin-based ACE model that is used to estimate trait heritability, the total variance is partitioned into additive genetic variance (A), shared environmental variance (C) and individual environmental variance (E) [35]. Further disassembly of variances to account for genotype-by-environment interactions or to include more complex genetic backgrounds are possible [36] but may be neglected here for simplification.

The term heritability is defined for two ratios of variances: the *broad-sense heritability,* H^2*,* and the *narrow-sense heritability,* h^2. The former refers to the proportion of the phenotypic variance contributed by the variance of all genetic factors irrespective of their mode of action (including dominance and epistasis):

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}.$$

The latter refers to the proportion of the phenotypic variance due to the variance introduced by additive genetic effects only:

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}.$$

Again it was FISHER who claimed in his *Fundamental Theorem of Natural Selection* that the evolutionary response to natural selection in fitness equals the additive genetic variance in fitness of an organism [37] (actually, FISHER only speaks of *"GENETIC VARIANCE"*; that he meant *additive* genetic variance has been pointed out later by his co-worker EDWARDS [38] and is now generally accepted [36]). In other words, due to the segregation of alleles, the phenotypic resemblance between relatives is mainly the result of additive genetic effects as non-additive genetic effects do not predictably change allele frequencies in the next generation (dominance as well as epistasis are not trivially predictable as both depend on two or more alleles). Therefore, if not stated otherwise, the term heritability most commonly refers to h^2.

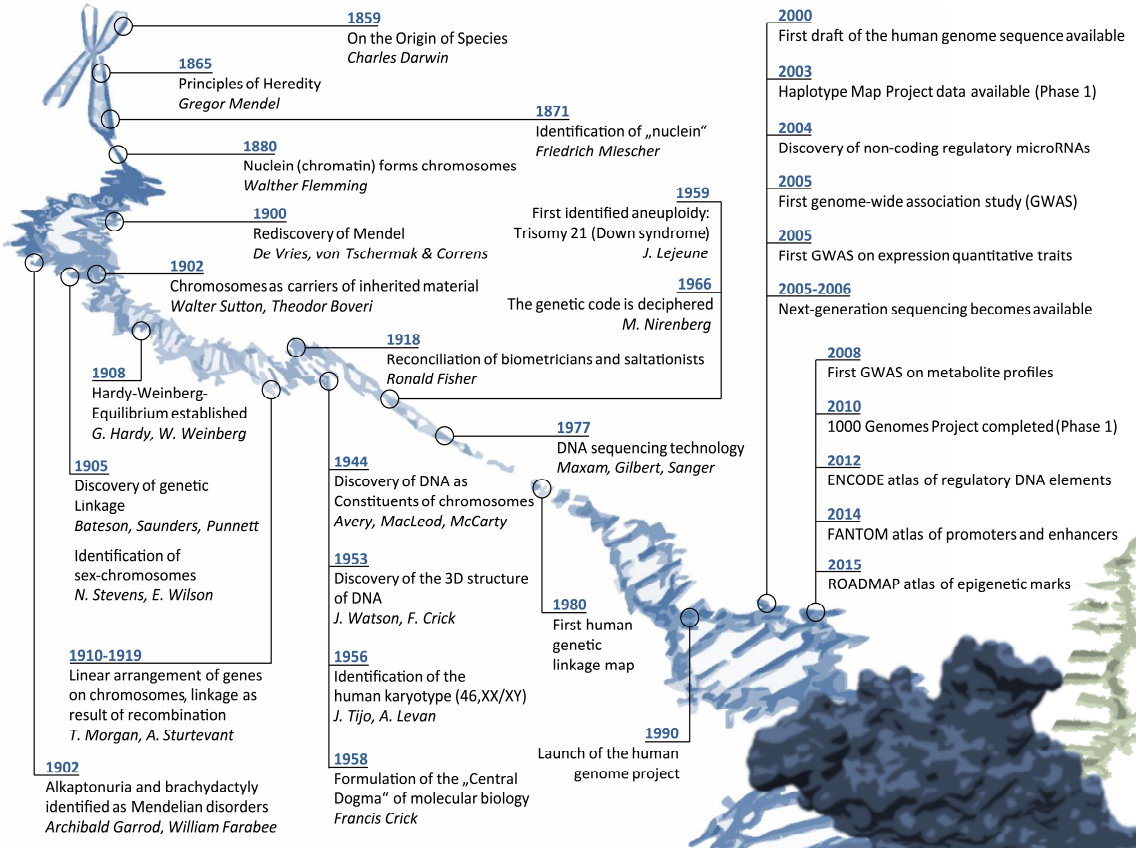


Figure 1: **Overview of milestones in human genetics**.

1.3 The molecularization of genetics

The theoretical approaches introduced by FISHER disentangle sources of variance of complex human traits, enabling a comparison of the proportional contributions to trait variance attributable to additive and non-additive genetic as well as environmental factors. Of equal importance was his insight that inheritance patterns of quantitative/multifactorial human traits are not contradictory to the Mendelian principles of inheritance, thus reconciling the two competing camps of geneticists. After this breakthrough, new discoveries followed at a rapid pace, both in the field of single-gene disorders as well as in the newly founded area of population and quantitative genetics. Nevertheless, there are still many black spots left regarding our knowledge on the heredity of multifactorial human traits. And in spite of the available sophisticated and complex mathematical and statistical frameworks, we are still not yet able to

give answers to many questions concerning heritability and the extent to which it contributes to phenotype prevalence in human populations (see also Box 2).

However, across the 20th and the beginning 21st century, theory has been substantiated by revolutionizing advances in molecular biology. In 1944, the *Avery–MacLeod–McCarty experiment* identified DNA as the substance encoding the genetic information [39]. Composite of the four nucleobases adenine, cytosine, guanine, and thymine (described by ALBRECHT KOSSEL [40–43]) linked to a backbone of phosphate and deoxyribose sugar (resolved by PHOEBUS LEVENE [44, 45]) and organized in a stable double-helical structure (determined by JAMES WATSON and FRANCIS CRICK [46]), DNA was found to be bound by special proteins (*histones;* also discovered by KOSSEL [47]) condensing it to chromatin. In his experiments, LEVENE also found a second class of nucleic acids with ribose sugar instead of deoxyribose in its backbone [44]: ribonucleic acid (RNA). RNA contains the same nucleobases as DNA except for thymine which is replaced by uracil (discovered by ASCOLI, OSBORNE, and HARRIS [48, 49]). The interrelation between DNA, RNA, and proteins (termed by BERZELIUS [50]) was unknown, although the latter were already known to consist of amino acids, to categorize into several functional groups, as well as to be the main class of functionally active molecules in the cell [51]. It was again FRANCIS CRICK who formulated the *Central Dogma* of molecular biology [52]: genes encoded in the DNA are transcribed into RNA and the nucleic acid four letter code is then translated into the 20 letter amino acid code of proteins (Figure 2). The *genetic code* underlying this translation was found to work *codon*-based (three nucleobases = one codon = the code for one amino acid) [53] and only five years later, in 1966, the complete code was deciphered [54]. With that, it was also discovered that the genetic code is buffered against some mutations (single nucleotide variants or SNVs) as for all amino acids except tryptophan and methionine there is more than one codon available (this is called *codon degeneracy:* there are 61 codons encoding 20 amino acids). For this reason, SNVs that alter a gene's DNA code but do not change the amino acid sequence of the encoded protein are called *synonymous* variants. Accordingly, SNVs altering both the DNA sequence and the amino acid sequence of a protein are called *non-synonymous* or *missense* variants. The first non-synonymous variant that causes a Mendelian trait was described for sickle-cell anemia (MIM:603903), an autosomal-recessive disorder which, in its most common form, is caused by a single SNV *(rs334)* in the *HBB* gene entailing a single amino acid exchange in the *hemoglobin subunit beta* protein [55–57].

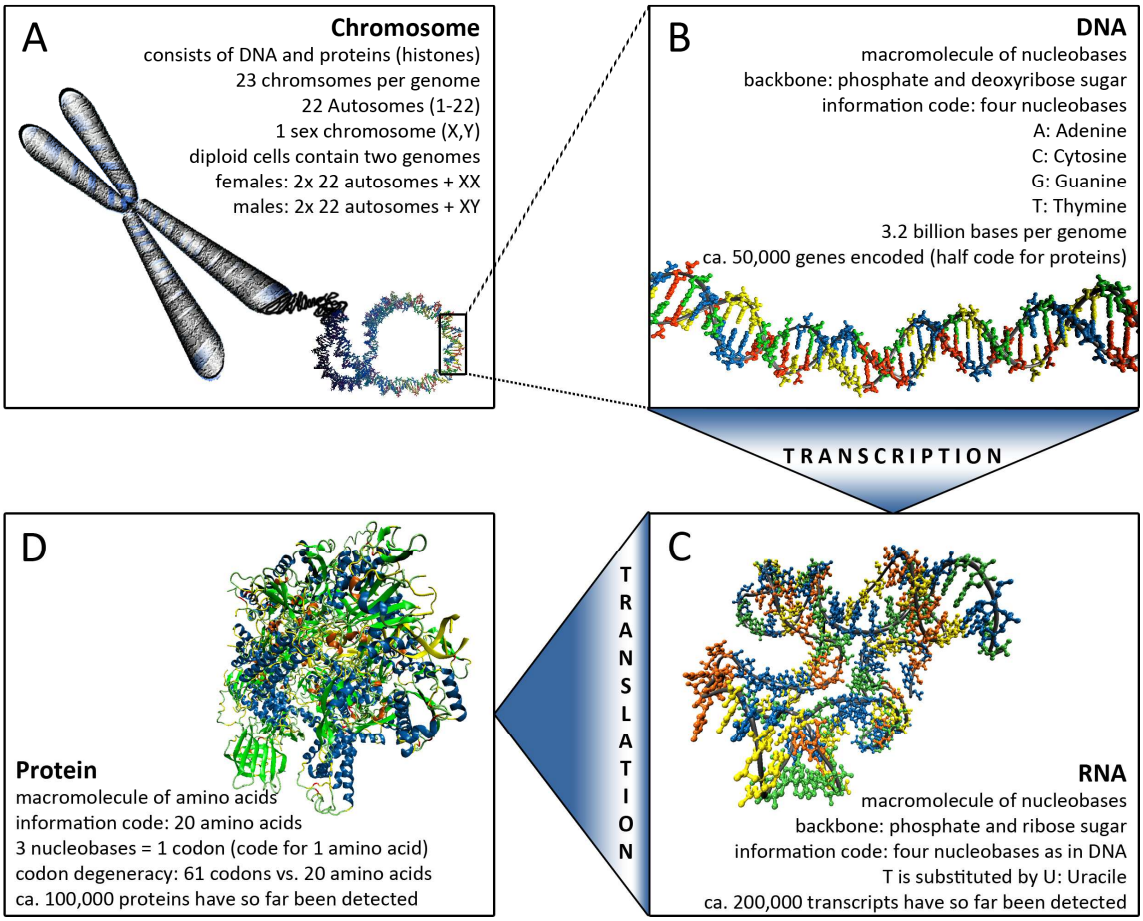


Figure 2: **Central Dogma of molecular biology**. DNA, by histones condensed to chromosomes, contains the genetic information. The genes encoded in the DNA are transcribed into RNA sequences that then serve as template for the translation into proteins via the genetic code. Numbers include all known forms of the entities (isoforms, transcribed pseudogenes, etc.) and are specific to humans (taken from [58]).

Shortly afterwards, it was recognized that the processes of transcription and translation from DNA to RNA to protein work not as linearly as expected. Although there exist *transcripts* (the RNA products of transcription) that are translated into proteins without modifications (*colinear* transcripts), it was found that this is not the prevailing mode of operation in humans and many other organisms. Instead, transcripts are modified in a process called *splicing* where sequence parts of the RNA *(introns)* are cut out before the remaining sequence parts *(exons)* are translated [59, 60]. Further complexity is added as two transcripts of the same gene can differ based on differential removal of introns or skipping of exons *(alternative splicing)* [59, 61]. Other RNA molecules were, in contrast to protein-coding *messenger* RNA (mRNA), found to remain untranslated (*non-coding* RNAs) featuring a variety of functions: transfer RNA (tRNA) and ribosomal RNA (rRNA) that are involved in protein synthesis, microRNA (miRNA), small interfering RNA (siRNA), and PIWI-interacting RNA (piRNA) that are involved in post-

transcriptional regulation of mRNA translation, and several others such as long non-coding RNA (lncRNA) with an extensive palette of additional functions (reviewed in [62]). Another layer of information was added when the proteins involved in the regulation of *gene expression* (active transcription of a gene) were identified. Besides the primary enzymes transcribing DNA into RNA – the RNA polymerases – a plethora of other proteins called *transcription factors* (TFs) were discovered to affect gene expression and its efficacy by binding to specific DNA sequences (*promoters* close to the transcription start site and more distal *enhancers/repressors*) [63]. These also include for instance the histones that are responsible for unfolding chromatin to make the DNA accessible for the transcription machinery.

The breakthrough of central importance to genetics was the development of experimental methods to determine DNA sequences – *DNA sequencing* – by synthesizing DNA using the target sequence as template. The two first approaches applicable for larger DNA segments were the Maxam-Gilbert method and Sanger sequencing, each named after their inventors [64, 65]. Both applications involved radioactive labelling of the synthesized DNA sequences, single runs for each nucleotide, and manual "reading" of the sequence from X-ray films of the combined per-nucleotide experiments. The Sanger method could be further refined using fluorescent labelling with different dyes for each nucleotide which enabled sequencing of all four nucleotides in a single run and thus allowed for automatization of the sequencing process [66]. With this technology in hand, in 1990 the International Human Genome Sequencing Consortium was initiated and the complete human genetic material – the *genome* consisting of about 3.2 billion bases (gigabases) – was sequenced [67, 68]. Simultaneously, high-resolution genetic linkage maps of human chromosomes were developed for the study of human single-gene disorders [69-71]. When the sequence of the human genome was available, the International Haplotype Map Project (HapMap) [72-75] was launched to obtain genetic maps of even higher granularity for 11 global populations using common SNVs with a frequency greater than five percent (referred to as *single nucleotide polymorphisms* or SNPs) to enable the study of complex multifactorial traits in genome-wide association studies (GWAS). Rapid advances in DNA sequencing techniques and the advent of next-generation sequencing (NGS) platforms facilitated the sequencing of human genomes enabling large-scale sequencing consortia such as the thousand genomes project (1000 genomes) to build again denser maps of human genetic variation [76, 77]. Other large consortia such as ENCODE (ENCyclopedia Of Dna Elements) and FANTOM (Functional ANnoTation Of the Mammalian genome) used the genome sequence to assemble catalogs of the functional elements within the human genome

from genes over non-coding RNAs to regulatory factors and their interactions. The compiled results of the two consortia published just recently showed that the majority of the 3.2 gigabases (Gb) of the genome are actually functional, further increasing the complexity of the functional interpretation of genetic and genomic studies [78-81].

Nowadays, there are high-throughput methods to assess almost any of the identified biological entities in place at large scale. In this context, the suffix –ome has been coined that refers to the total of a certain kind of biological entity such as the genome (the complete genetic material), the transcriptome (all expressed transcripts), the proteome (all present proteins), and the metabolome (the full set of metabolites) of an organism, a tissue, or a cell. The related suffix –omics is used for analyses that use high-throughput data on one of the –omes at a coverage as comprehensive as the available experimental platforms allow for. There are also more multidisciplinary fields of study that include several different assay methods. For instance, epigenomics (epigenetic modifiers influence gene expression) refers to the analysis of DNA methylation, histone modification, chromatin accessibility, and some other processes that are measured by very different means. These different measurements are then integrated and clustered to obtain a regulatory landscape on genome-wide scale. The largest available collection of such data has been published only recently by the NIH Roadmap epigenomics consortium [82].

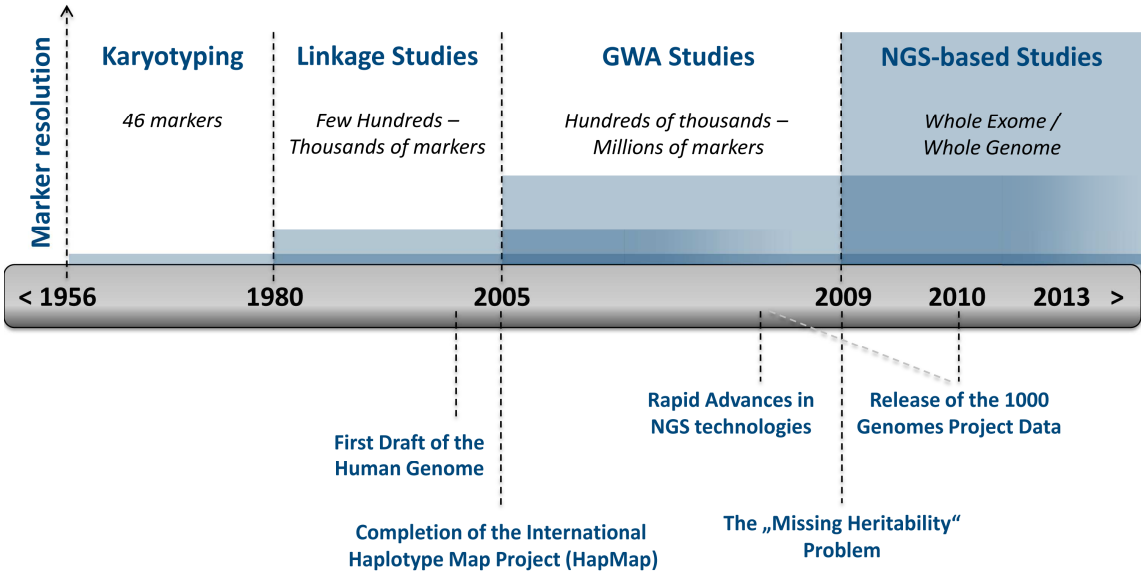


Figure 3: Timeline and marker resolution of analyses of genetic variation.

1.4 Analyzing genetic variation

For the analysis of genetic variation, several applications have been developed to provide experimental platforms for the analysis of the genetics of human traits and diseases. The resolution at which genetic markers could be explored steadily increased over time with modern studies measuring genetic differences on a genome-wide single nucleotide resolution (Figure 3). However, the early approaches still co-exist and will also be briefly introduced.

KARYOTYPING – This was the first experimental test for genetic aberrations based on visual inspection of the number of chromosomes and their structure – the *karyotype*. The first human karyotype was published in 1952, however, it contained the wrong number of chromosomes (48 instead of 46) [83]. The correct normal human karyotype consisting of two copies of the 22 autosomes and two sex chromosomes and denoted by 46,XX for females and 46,XY for males (Figure 4) was determined in 1956 by JOE TIJO and ALBERT LEVAN [85]. Notable deviations from the normal karyotype – *aneuploidy* (abnormal number of chromosomes), large structural rearrangements such as translocations and inversions, or large copy number variants (CNVs) such as deletions or duplications – can be identified by karyotyping combined with special chromosome staining methods [86]. Shortly after the publication of an experimental protocol to retrieve human chromosomes from cell cultures, the first genetic disorders were linked to aneuploidies: Down syndrome, that is caused by an excess copy of chromosome 21 (trisomy 21; karyotype: 47,XX/XY,+21; MIM: 190685) [87, 88], and Klinefelter syndrome, that manifests due to an additional copy of the X chromosome in males (karyotype: 47,XXY) [89].

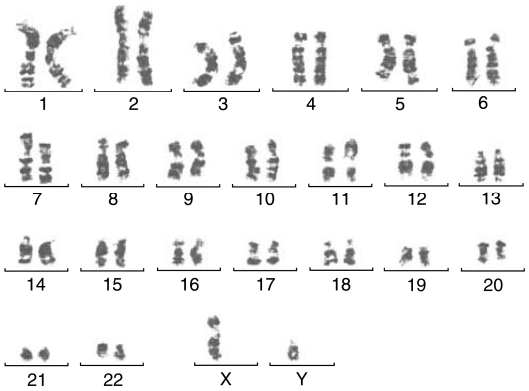


Figure 4: **Normal male human karyotype**. Adapted by permission from Macmillan Publishers Ltd: Leukemia [84], copyright 2003.

LINKAGE ANALYSIS – This study type makes use of linkage maps that contain a catalog of genetic markers with known chromosomal locations. The genetic markers (typically microsatellites or restriction fragment length polymorphisms, Figure 5) are investigated in pedigrees for co-segregation with the phenotype under study. To achieve that, the parental origin of genomic regions is traced by marker-based recombination event mapping. Regions

that display identical marker expression exclusively in individuals presenting the phenotype are said to be linked to the gene(s) causing the phenotype. Using a test statistic such as the logarithm of the odds (LOD) score [90], markers informative for the phenotype can be identified and ranked. Dependent on the size of the pedigree and the marker density in the linkage map, the identified regions can vary crucially in length [91].

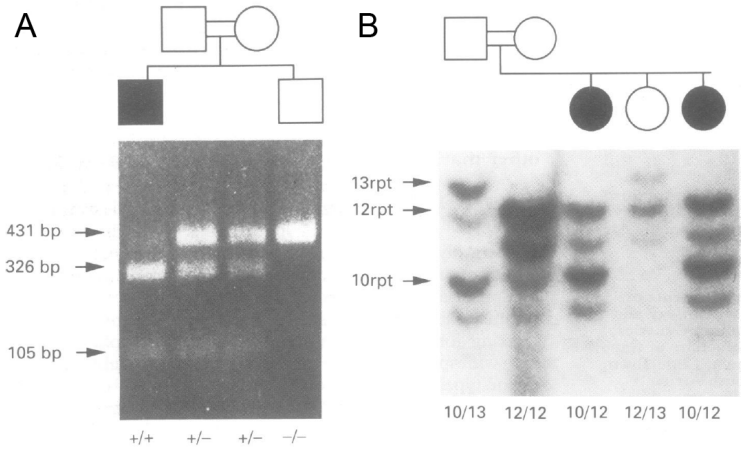
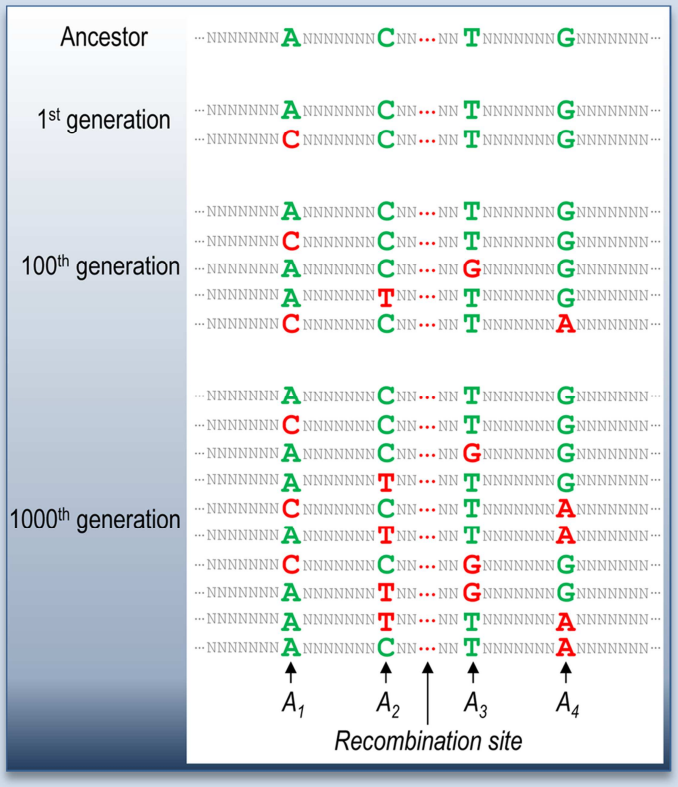


Figure 5: **Analysis of genetic markers in linkage analyses**. **A**: Gel showing the results of the analysis of a restriction fragment length polymorphism. The marker is informative as it shows a clearly distinguishable pattern for the diseased person (filled rectangle in the pedigree). **B**: Gel showing the results of the repeat count analysis of a microsatellite. This marker, too, is informative: the pattern is identical in cases (filled circles) and differs from controls (open circles and rectangle). Adapted by permission BMJ Publishing Group Ltd: Journal of Medical Genetics [92], copyright 1994.

GWAS – The basis for GWAS was the common disease/common variant (CDCV) hypothesis stating that common traits featuring a genetic component are likely to be caused by common variants [93–95]. The whole concept of GWAS thus presumes a direct link between trait prevalence and the frequency of occurrence of the underlying genetic causes. By definition, the term "common variant" implies a certain minimal frequency of the alleles in a population, and, therefore, GWAS relies on large population-based (case/control) cohorts to achieve the statistical power necessary to detect significant signals. By utilizing the data provided by HapMap, commercial experimental platforms (SNP arrays) were developed to genotype more than half a million SNPs at once. With these genotypes available, variants correlating via *linkage disequilibrium* (LD; see Box 1) can be imputed, that is, their genotype can be estimated using the HapMap or 1000 genomes haplotype map specific for the population under study. By this process, genotyping of only a fraction of the total catalogued common variants (ca. 3.1 mio in HapMap phase 2) can achieve almost genome-wide coverage of common markers [74]. In a statistical process (most commonly linear regression for quantitative and logistic regression for binary traits), the allele frequencies are compared with respect to case/control status or

Box 1: ***Linkage disequilibrium and genotype imputation***



Linkage disequilibrium or LD describes the non-random distribution of alleles in a locus. The phenomenon is based on co-segregation of genetic variants that is due to recombination.

In the figure on the left, the ancestral allelic setting is shown on the top. Along several generations, random mutations occur at certain positions (A_1-A_4, depicted in red) and are passed on to following generations. Random recombination events between markers A_2 and A_3 lead to many combinations of the mutated alleles. Some alleles, however, are not separated by recombination (A_1/A_2 and A_3/A_4, respectively) – they are in linkage and form a haplotype. This leads to a correlation between these alleles that deviates from the hypothetical random formation of the allelic structure.

There are several measures that quantify the extent of LD. The most commonly used measure of LD is called r^2, the squared correlation coefficient (coefficient of determination) of the alleles of two variants:

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}$$

where p_{AB} is the frequency of the occurrence of the major allele (the more frequent allele) at both variants and p_A and p_B is the frequency of the major allele of the first and the second variant, respectively. In the above example, for A_3/A_4 (1000^{th} gen.) this value would be $r^2 = 0.286$, a weak correlation resulting from mutations occurring at both sites. LD is considered to be strong at an threshold of $r^2 \geq 0.8$.

The GWAS approach utilizes this correlation as it allows to predict the genotype of complete haplotypes by genotyping only a few variants per haplotype. Modern imputation algorithms use more complex approaches than simply imputing alleles using LD measures: Haplotypes are first phased (the haploid allele sequences as in the above example are predicted) and stored in weighted graphs. These weights are as simple as the count of occurrence of the allele sequence in a population study such as the 1000 genomes project. Haplotypes that are statistically very unlikely to occur are then removed from this graph in a process called pruning. This is done to minimize the chances for wrong haplotype predictions. Afterwards, the genotyped alleles are entered in the graph and the best-fitting haplotype (selection is based on optimization of a summary statistic of the weights) is obtained [96, 97].

A secondary but equally important consequence of the haplotype structure of the human genome is the reduction of the number of tests for trait association necessary to yield genome-wide coverage of common markers. The estimated number of independent common haplotypes within the genome amounts to 1 million, defining the threshold for genome-wide significance at a Bonferroni-corrected association P_{value} to be $5.0 \cdot 10^{-8}$.

quantitative measures. Significant differences are then obtained by filtering for associations exceeding a P_{value} threshold adjusted for multiple testing. The large amount of both samples and measured genetic markers bears a high potential for false positive statistical associations. Therefore, scrutiny in quality control as well as in the association analysis and the interpretation of the results of GWAS is indispensable. Nonetheless, since the first successful GWAS published in 2005 [98] hundreds of studies followed yielding thousands of genetic loci linked to human traits, making this research method the probably most broadly applied study type in biomedical research to date [99, 100].

NGS-BASED STUDIES – To identify rare variants contributing to human traits it is necessary to get at the single nucleotide level. Although different enhancements have made it feasible to analyze variants with frequencies as low as one percent in the population using GWAS, for even rarer variant the only detection method at hand is sequencing. Therefore, as sequencing costs are decreasing, NGS-based studies have become increasingly popular over the past few years. And this holds true not only for rare disorders but also for common multifactorial diseases. Using different methods of association testing for common variants (handled similarly to GWAS), rare variants (for instance burden tests) or for the combination of both rare and common variants such as sequence kernel association tests (SKATs), sequencing studies can be used for the full spectrum of genetic analyses from single-individual to population-based genetics [101]. There are three major study subtypes: whole-genome sequencing (WGS), whole-exome sequencing (WES), and gene panel sequencing, with the latter being mostly used in clinical settings. WES differs from WGS in the way that only exons are sequenced. As the human exome makes up only about one percent of the genome, WES is much less expensive than WGS and therefore often preferred, although sequencing quality and coverage is generally higher with WGS [102]. While NGS-based studies were heralded as the solution to all shortcomings associated with the GWAS approach (mainly, its ignorance of the effects of very rare variants and the "missing heritability" problem [103, 104], see also Box 2), the complexity introduced by screening for variants exome- or genome-wide has its pitfalls: sequencing, in the best case, reveals all existing variation, including non-functional variants, that is, variants that are under neutral selection and are not affecting fitness or the phenotype. This results in the most challenging task of ranking or filtering variants for their likeliness of being effective in the trait under study – a process still undergoing development given that mostly there is no background information on this matter available [105] – even before their functional characterization.

1.5 Genetics of Mendelian diseases

Monogenic or Mendelian disorders are caused by altered functions of single genes that are grouped into several classes according to the location of the disease genes and their mode of inheritance (dominant/recessive) in which they segregate in families (Table 2) [105]. Due to the severity of monogenic diseases and the linked reduced reproductive fitness of affected individuals, these disorders individually are generally rare affecting fewer than 1 in 2,000 people in Europe. The exact number of Mendelian diseases is hard to report as the entry statistics of databases collecting Mendelian or orphan disorders such as the Online Mendelian Inheritance in Man (OMIM) catalogue [106] and OrphaNet [107] diverge significantly. Current estimates amount to 6,000–7,000 known rare genetic diseases with a comparable estimated number of additional, unknown phenotypes of the kind, collectively affecting millions of individuals worldwide [105–109].

Mode of inheritance	Example	Disease gene
Autosomal dominant	Huntington disease (MIM:143100)	Huntingtin (*HTT*)
Autosomal recessive	Sickle-cell anemia (MIM:603903)	Beta hemoglobin (*HBB*)
X-linked dominant	Rett syndrome (MIM:312750)	Methyl-CpG-binding protein 2 (*MECP2*)
X-linked recessive	Hemophilia A (MIM:306700)	Coagualtion factor 8 (*F8*)
Y-linked	SRY-related sex reversal (MIM:400044)	Sex-determining region Y (*SRY*)

Table 2: **Examples of single-gene disorders**, **their modes of inheritance**, **and the disease-causing genes**.

The etiology of more than half of the defined Mendelian diseases could already be identified with OMIM listing 4,322 phenotypes with known molecular basis (as of January 12th, 2015). Disease genes have primarily been discovered through linkage mapping, however, the application of NGS led to an accelerating pace of discoveries with more than 130 novel disease genes reported in 2012 alone [105, 110]. As rare diseases are caused by loss-of-function (LOF) or gain-of-function mutations in single genes, the primary targets for therapeutic intervention are also known. Nonetheless, there are still far less drugs or therapies available than targets known [105]. This is not only due to the challenges that are *per se* linked to drug development (such as drug specificity and side-effects). Studies have also shown that, for some disorders that are classified as monogenic diseases, so-called modifier genes exist that, based on their allelic structure, affect disease presentation and severity. One example for such an oligogenic trait is cystic fibrosis (MIM:219700), an autosomal recessive disorder that is primarily caused by

mutations in the *cystic fibrosis transmembrane conductance regulator (CTFR)* gene. Although defects in *CTFR* almost always cause cystic fibrosis, the individual phenotypic outcome depends on a number of alleles in up to seven additional genes [111].

1.6 Genetics of complex traits

As YULE correctly stated more than hundred years ago, multifactorial (or complex or common) traits such as height, type 2 diabetes, and coronary artery disease differ from mono- and oligogenic traits in that they have yet no clearly identified cause. Instead, they are the result of a complicated interplay of genetic, behavioral, and environmental factors. However, the estimated amount of genetic contributions to the development of complex human traits and diseases as derived from pedigree and twin studies is generally high (often >30% [112]). This led to massive investments in genetic scans using the GWAS approach to identify the hereditary factors predisposing to complex traits. Applied on hundreds of traits in study cohorts of

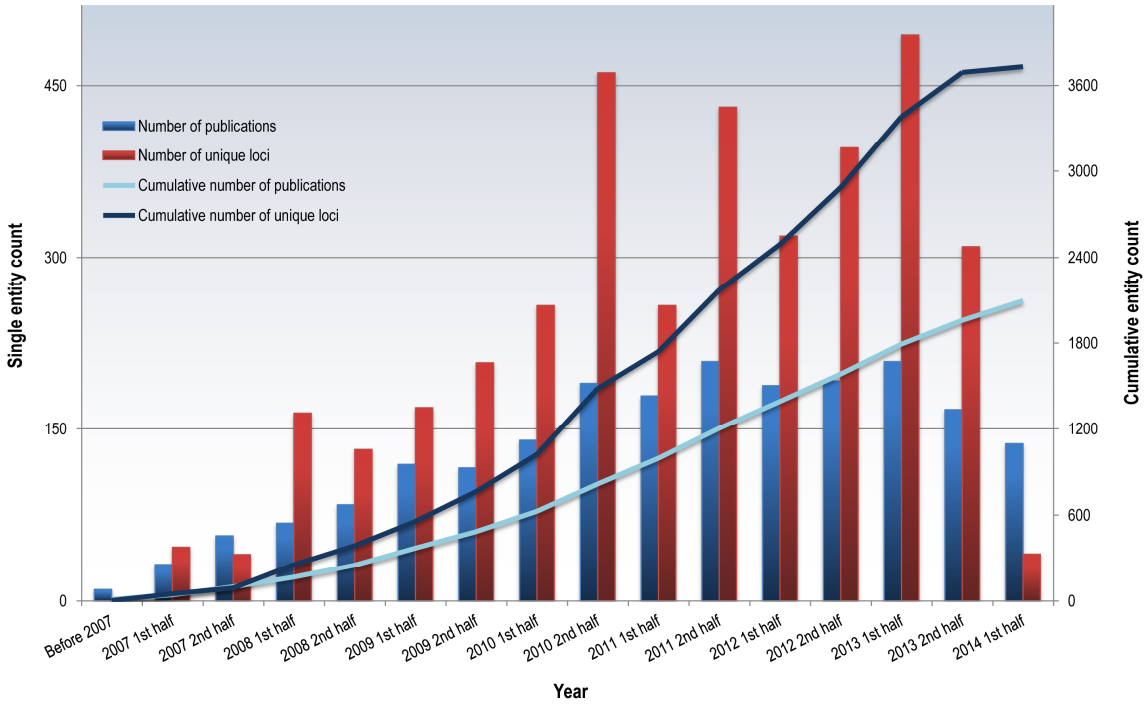


Figure 6: **Timeline of GWAS discoveries**. Shown are the number of GWAS publications and the number of genetic loci identified to predispose to complex traits. Data obtained from the GWAS Catalog [100]. Loci were determined as variants in strong LD ($r^2 \geq 0.8$) with an association $P_{value} \leq 5.0 \cdot 10^{-8}$.

(hundreds of) thousands of individuals, GWA studies – capturing most of the common (and, more recently, also a fair portion of the rarer) variation of the human genome – have identified several thousand genomic loci that add to the susceptibility to complex traits (Figure 6) [100]. In their 2012 review, VISSCHER and colleagues quantified the success of GWAS as a 16-fold increase in the number of independent loci identified to contribute to eleven complex autoimmune and metabolic diseases before ($N_{loci} = 24$) and after ($N_{loci} = 384$) the introduction of the GWAS approach [112].

Box 2: ***Missing heritability***

There has been great controversy about the fraction of heritability explained by GWAS-identified variants. Explained trait heritability is defined as the ratio of known additive heritability (the variance explained by all significant GWAS hits combined) divided by the complete (narrow-sense) heritability of a trait:

$$\pi_{explained} = \frac{h^2_{known}}{h^2_{complete}}.$$

The explained heritability differs substantially between complex traits ranging from high (70% for type 1 diabetes [113]) over mediocre (20% for multiple sclerosis [112]) to low (<3% for schizophrenia [114]). Interestingly, for intermediate complex traits such as gene expression levels and metabolite concentrations, additive genetic effects are generally strong and explain a large part of the genetic variance contributing to these traits [115, 116]. This is thought to be a result of the more direct (that is, less confounded) effects of genetic variance on such primary traits.

Different models have been proposed for the estimation of $\pi_{explained}$ [114], all including some background hypotheses such as the genetic model underlying the studied trait [36]. For the majority of complex traits, with these models only a minor proportion of heritability could be explained, leaving a substantial fraction of heritability missing. It was claimed that it may be unfeasible to identify all genetic variants contributing to trait predisposition, be it because of their small effect sizes or their low frequency in the population or both [103, 117]. Conversely, it was suggested that the models used to estimate $h^2_{complete}$ may be insufficient. For instance, an estimator was proposed including genetic interactions between loci contributing to disease risk, which plausibly showed that, if genetic interactions are present, $h^2_{complete}$ could be overestimated to a significant extent, showing unexplained heritability where in fact there is none (this phenomenon was termed "phantom heritability" [118]). Additionally, it was suggested that individual gene-environment interactions are likely to be underestimated in twin- or sibling-based measures, stating that estimates of $h^2_{complete}$ based on close relatives may be inflated [35]. Another source of error of current estimators is the transformation of observed heritability estimates to a liability scale (the heritability of liability [114]) that is assumed to be normally distributed. In most settings, this transformation leads to inflation of $h^2_{complete}$ estimates if disease prevalence is lower than 25% [35].

Indeed, recently evidence is growing that non-additive effects such as genotype-by-genotype (GxG) and genotype-by-environment (GxE) interactions have been substantially underestimated while they are simultaneously captured in narrow-sense heritability estimates (which is contradictory) [36, 115, 119, 120].

Despite the search for genetic loci has obviously been very productive, investigation of these loci and their effects on trait development and disease risk revealed even more intricacies than expected. Effect estimates of the genetic associations showed that the single loci explain only small proportions of the genetic variance of the associated trait and, even when combining the additive effects of all loci identified for a phenotype, the explained variance is substantially lower than the estimated heritability of the trait. Thus, a significant ratio of trait heritability remains missing (see Box 2) [103]. Moreover, the great majority (>90%) of GWAS signals is located in non-coding regions, not within protein-coding exons [100, 121]. It is assumed that non-coding variants affect the linked phenotypes via regulatory mechanisms, however, apart from a few exceptions, most of these loci still await functional characterization.

1.7 Bioinformatics and computational genetics

The advances in molecular biology over the past 60 years led to a large catalog of experimental approaches for the investigation of human traits and diseases. Many of those have been automatized and enhanced to high-throughput screenings, leading to an exponential growth of data that needs to be processed, stored, and interpreted. The research field of bioinformatics (also known as computational biology) was founded in the 1970s to explore biological information processes, that is, to accumulate the wealth of experimental data and utilize it to search for patterns, interactions, and dependencies, to model those as accurately as possible in mathematical frameworks, and, finally, to use these models to understand how cells and organisms function and to predict differential outcomes due to external or internal changes to the systems [122]. The subfield of computational genetics and genomics applies bioinformatics tools to model the impact of variations of the genetic material within and between organisms and cells. To get an idea of the complexity of this task, the amplitude of the various genomic entities described in the previous sections has to be considered. The human genome sequence in its latest release (GRCh38) contains 3,212,670,709 letters (bases). To this genomic sequence more than 50,000 genes have been aligned that are transcribed to over 200,000 transcripts encoding close to 100,000 different proteins and isoforms. The combined output of ENCODE and FANTOM5 lists additional 1.1 million regulatory DNA elements (all

numbers from [58]). All these entities have to be linked and augmented with annotation and interaction data which significantly increases the data load.

While studies with a limited set of genetic markers like karyotyping and candidate-gene linkage studies were still feasible without performant algorithms, GWAS and NGS-based studies cannot be realized without the power of modern computers due to the large sample sizes and the huge amount of data created. Recent GWAS combine genotype information of >2.5 million genetic markers for more than 200,000 individuals each, totaling to a collection of more than 1.5 trillion data points that have to be analyzed [123, 124]. NGS-based studies, although usually encompassing a considerably lower number of individuals, are even more complex. For instance the study of the Cohorts for Heart Aging Research in Genetic Epidemiology that performed WGS for 962 persons at 6-fold average coverage producing a total of 1.85 trillion

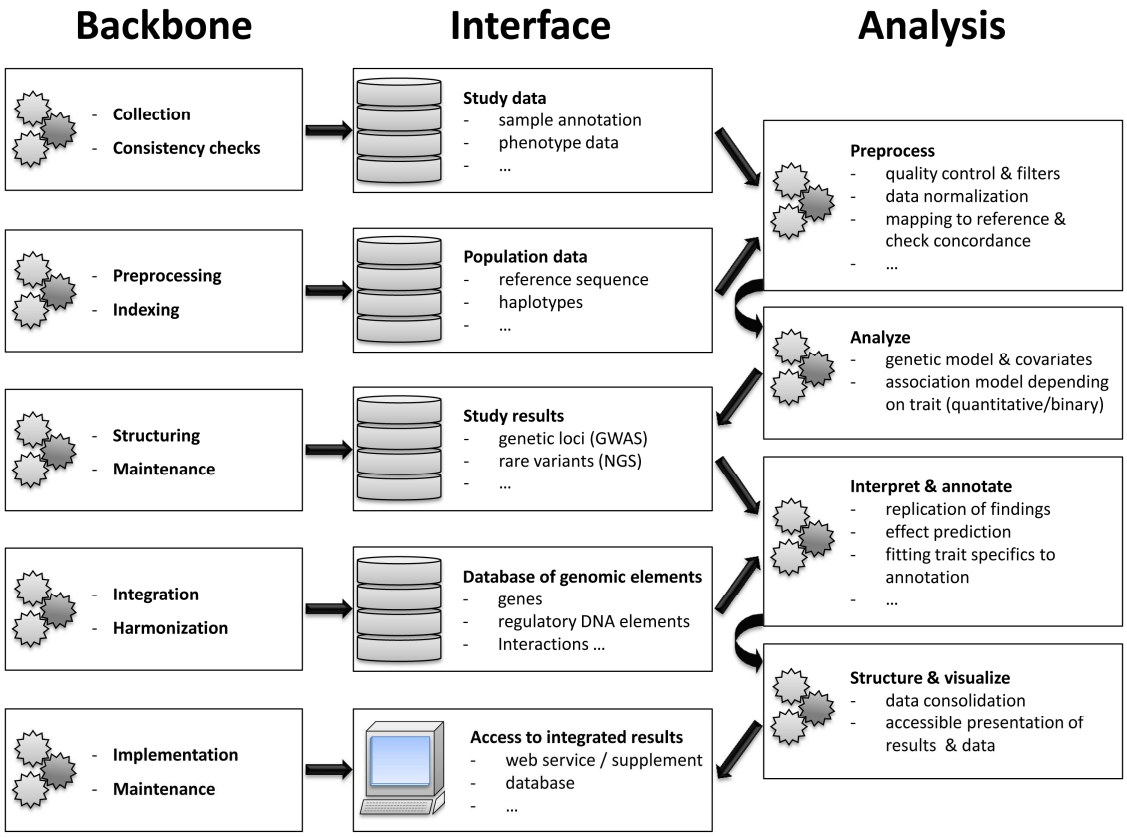


Figure 7: **Outline of bioinformatics work packages involved in modern genetic analyses**. High-throughput screens for genetic variation create a huge amount of data. Reference data from the population as well as genome annotations – which both are large data collections – have to be included in the analysis process. For efficient handling of study data, three logically separated but interlinked layers of information processing are needed: first, the programming backbone that provides the informatics framework to harmonize and consistently store the different data sets. Second, the interface layer that provides efficient access to the data without disclosure of the software routines in the backbone. And, third, the analytic layer where, depending on the study type, specific calculations are performed for data processing and normalization, quality control, association testing, and annotation of the identified genetic loci.

bases to be assembled, analyzed for variation, and then investigated for links to the trait [125]. After primary data analysis and association testing, the genetic association results have to be merged with the information backbone to retrieve hypotheses on the functional mechanisms linking the genetic loci to the trait or disease, which can then be tested experimentally in cell lines or model organisms. In the last step, it has become common practice to consolidate study results and put them at the scientific community's disposal either through custom web services or through large deposition sites such as the database of genotypes and phenotypes (dbGaP) or the European genome-phenome archive (EGA). Bioinformatics provides the complex frameworks, interfaces, and algorithms to achieve these tasks (Figure 7) and many of them will be introduced throughout this work.

1.8 Supporting the evidence for trait-associated variants

For multifactorial traits, clear causal genetic and molecular impact factors are still largely absent. The inability to identify the major genetic drivers of complex traits was – by some researchers – perceived as a failure of GWAS [104, 112, 117, 121]. By now, GWAS has been increasingly replaced by large sequencing studies intended to fill the gaps revealed by association studies. Nonetheless, the missing heritability paradox (Box 2) has not been solved yet, as well as most of the essential questions that arose from the findings of GWAS still persist.

GENETIC MODEL – One outcome of the discussion about missing heritability was the concept of "synthetic associations" (Box 3) [104]. The theory states that the signals detected by GWAS may be due to nearby rare, high penetrance variants that mediate the actual (dominant) effects on trait predisposition. While plausibly formulated, it was shown mathematically that synthetic associations are not to be considered the general rule [117]. Nonetheless, the theory about synthetic associations causing GWAS signals led to a controversial dispute about the genetic model underlying the heredity of complex traits, namely, if they are caused by *(i)* many common and rare variants with small effects (the infinitesimal model described by FISHER), *(ii)* several moderately highly penetrant rare variants, or *(iii)* a mixture of both of these models including broad-sense genetic factors (GxG and GxE interactions) as well as epigenetic effects such as DNA methylation patterns (the broad-sense heritability model) [117].

Box 3: ***Synthetic associations***

The concept of "synthetic associations" describes the phenomenon of a common variant that weakly mirrors the stronger effects of proximal rare causal variants and is thus detected by GWAS screens. While the concept is plausible to underlie a fraction of GWAS-identified signals [126], there are several reasons – besides parsimony – that contradict synthetic associations of being the major source of GWAS results. To be tagged by a common SNP, the rare variants have to correlate at least weakly with the identified variant. This correlation, naturally limited by the frequency of the rare variants, can be used to estimate the effect size of the assumed causal variants. The effects contributed by such rare variants would be significantly higher (with decreasing frequency of the assumed causal variants the effect increases exponentially) than that of the tagged variant (Figure 8) [127]. But:

(i) Variants with such strong effects would most probably have been identified in linkage studies.

(ii) When speaking of genetic variants, rare should be translatable into population-specific. Contrary to that, many GWAS hits from European studies could be replicated in other populations.

(iii) There is a substantial difference between the frequencies of the assumed rare variants and the prevalence of common diseases. The independent rare variants may all affect the same loci. But: these would again have been detected by linkage analysis.

(iv) The assumed rare variants would all have to be occurring in the same common haplotype within the locus to be tagged by SNPs. As GWAS markers have an average MAF <20%, this seems unlikely.

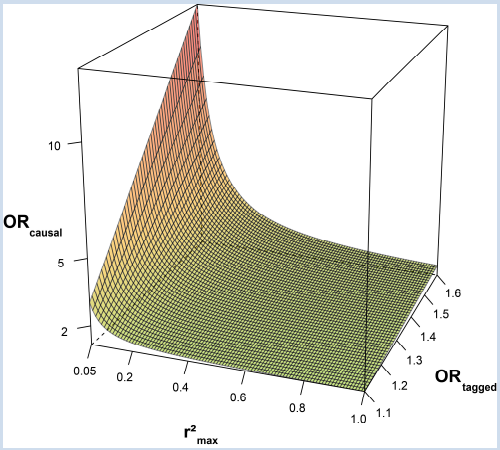


Figure 8: **Approximated effect size of a rare causal variant causing a synthetic association**. The effect of a rare variant (odds ratio, OR_{causal}) is approximated by a function of the OR of the tagged common variant (OR_{tagged}) and r^2_{max}, the maximum possible LD between the two variants. Assuming that the rare variant has an allele frequency far below 1% and the mean frequency of GWAS-identified variants is between 20-30%, r^2_{max} has to be <0.05. Thus, OR_{causal} should exceed OR_{tagged} at least by factor 5-10.

The strongest argument against synthetic associations is, however, of a different nature. It is given by the fact that, although many NGS-based studies (also NGS-based studies of parents-child trios) have been published, there has been no breakthrough on this subject. On the contrary, the more data becomes available, the more the hypothesis of rare high-penetrance variants causing synthetic associations falters. Just recently it has been shown for common autoimmune diseases that rare coding-region variants show *"NEGLIGIBLE IMPACT (…) ON MISSING HERITABILITY"* [128]. In another recent WGS-based study, the rationale is even more explicit: *"WE ESTIMATE THAT COMMON VARIATION CONTRIBUTES MORE TO HERITABILITY (…) THAN RARE VARIATION"* [125].While these findings may be specific to the studied traits, it is a validation of the theoretical refutation of synthetic associations being ubiquitous. Indeed, after several years of NGS-discoveries, there is only one report to be found in the literature of low-frequency variants with moderate effect sizes causing a synthetic association [129].

Although ample evidence on genetic associations is available, a conclusive synthesis on this topic could not been derived, yet. This is due to the obstacles hiding beneath each of the assumptions:

(i) Supposing the infinitesimal model, it is close to impossible to detect all variants contributing to the genetic variance of a complex trait. This is due to the small effect size of the variants assumed in this theory. Small effects are only barely measurable when summarized in the global genetic variance and, additionally, they complicate any proof of causality as their influence on the molecular level is hardly distinguishable from cellular noise such as temporary effects induced by cell cycle status or environmental stimuli. Variants that are frequent enough and exert measurable effects can be detected by GWAS or NGS-based studies using SKAT analysis. For rare variants with the same effect sizes, however, even the largest sample sizes will not suffice to push association statistics into significant ranges.

(ii) Assuming the rare variant model, penetrance of variants exerting moderately high effects has to be considerably less than 100%. Because the individual variants are very rare, there must be many such variants that are also present in apparently healthy controls to sustain the high heritability on population scale that is generally seen for complex traits. It follows that affected individuals need to carry several disease-causing variants in specific configurations to exceed the liability threshold to trait development. Thus, because the variants have to be both rare and present in the general population, their identification via sequencing is highly challenging, as typically rare variants present in controls are deliberately excluded from further analyses.

(iii) Finally, the broad-sense heritability model requires the identification not only of the causal genetic markers across the whole frequency spectrum, but also of the genetic and environmental interactions. In addition to the decreasing statistical power that results from the exponential increase in the number of tests necessary to detect interactions, especially environmental interactions are highly prone to individualized confounders such as diet or smoking behavior. To achieve sufficient power as well as to exclude the effect of confounders, huge study populations would be needed that live under very controlled conditions for a substantial amount of time – a study setting that for several reasons besides funding and ethics seems rather unrealistic. Recent evidence indicates that broad-sense heritability factors may substantially contribute to missing heritability [36, 115, 119, 120]. However, for the mentioned reasons, neither GxG nor GxE interactions have been systematically investigated on a global scale, yet.

With that, it becomes clear that the question of how complex traits are inherited cannot be readily answered. Independent of the genetic model, however, these considerations lead to another issue: how can we use the many known genetic associations to develop new diagnostic and therapeutic tools, if we are unable to identify the full set of causal, genetically predisposing variants?

> *"ULTIMATELY, THE MOST IMPORTANT GOAL FOR BIOMEDICAL RESEARCH IS NOT EXPLAINING HERITABILITY – THAT IS, PREDICTING PERSONALIZED PATIENT RISK – BUT UNDERSTANDING PATHWAYS UNDERLYING DISEASE AND USING THAT KNOWLEDGE TO DEVELOP STRATEGIES FOR THERAPY AND PREVENTION."*
>
> **Zuk et al., 2012 [118]**

CAUSALITY – The above quotation almost perfectly mirrors the contradictions that are involved in the study of genetic associations for complex traits. It correctly states that in the end it is not relevant for *therapy* whether missing heritability exists or which genetic model underlies the inheritance of complex traits, if the molecular determinants for trait development and progression can be identified using the available set of associations and simultaneously can be targeted by (novel) therapeutics. On the other hand, one may argue that *prevention* of trait development without knowledge of the personalized patient risk is hard to achieve. And personalized risk prediction should require the identification of the causal risk variants, if genetic predisposition plays only the smallest role in trait development.

However, as mentioned above, the molecular complexity of multifactorial traits as well as our incomplete understanding of the genomic landscape of genetic predisposition has severe implications for the identification of the causal variants. Because the direct, intermediate phenotypic variations introduced by most of the individual genetic components are unknown, the experimental validation of causality is almost impossible. Furthermore, because GWAS-identified alleles feature only minor effects on the organismal phenotype, it is likely that there are several allelic configurations leading to the same trait. This means that, theoretically, the absence of a risk allele in an affected individual tells nothing about its potential involvement in trait development – as well as the presence of a predisposing allele in a healthy person's genome does not necessarily allow predicting a trait endpoint. This has been impressively shown on the example of type 1 diabetes (T1D), a disease of high heritability with an estimated sibling recurrence risk of about 10-15: a genetic test using all variants associated with T1D that in

combination explain more than 90% of the genetic variance in the population would misclassify as many as 26% of T1D patients [130].

This does not mean that the variants incorporated in this predictor are not causal. On the contrary, the associations detected by GWAS are statistically robust and many of them could be replicated across different ethnic populations and even across species, which strongly suggests a true-positive causative link between the associated loci and the studied traits [131, 132]. It means that for complex traits (and even for some Mendelian disorders) causality of a variant does not translate into full penetrance on the organismal phenotypic level. Therefore, the statement of ZUK and colleagues is well-founded, although it may initially seem contradictory. Only the projection of all genetic and biochemical evidence available for each trait onto causally disturbed cellular pathways can enable therapy as well as prevention. This leads to the central question: is it possible to identify the causal molecular determinants for complex trait development using the available association data without knowledge of the actual causal variants?

FROM ASSOCIATIONS TO BIOLOGY – In several cases GWAS findings actively furthered our understanding of molecular disease mechanisms and there are several cases where discoveries from GWAS were successfully translated into clinical trials [112, 121, 133, 134]. An impressive example is the therapeutic targeting of the interleukin-23 subunit alpha protein *(IL-23-A)* in psoriasis, a common, highly heritable autoimmune disease of the skin. *IL-23-A* is active in the *IL-23* complex by binding another protein, *IL-12B*, that, when mutated, was identified to be involved in psoriasis predisposition. In addition to *IL-23-A*, *IL-12B* also binds a second cytokine, *IL-12A*, to form the *IL-12* complex. As the *IL-23* pathway was not well known at the time of the identification of *IL-12B* as predisposing factor, it was hypothesized that *IL-12* (the function of which was already determined) is causally involved in the development of psoriasis. And indeed, therapeutic targeting of *IL-12B* showed improvements of psoriasis severity in a phase I clinical trial, albeit the response rate was not as convincing as hoped for [135]. Therefore, the genetic predisposition landscape of psoriasis was more closely investigated in an association study and it was found that not only *IL-12B* is associated with psoriasis risk, but also *IL-23-A*, as well as the *IL-23* receptor, *IL-23R* [136]. This led to the closer exploration of the *IL-23* pathway that showed that this interleukin activates very different immune cells than *IL-12*. In addition, screening of expression data showed an up-regulation of the expression of *IL-12B* and *IL-23* mRNA in psoriatic lesions, but no change in *IL-12A* expression [137]. This knowledge motivated the development of an antibody specific to *IL-23-*

A (tildrakizumab). The recent report of the phase I clinical trial of this antibody showed a success rate of 92.8% [138].

However, in view of the thousands of robust association results produced by GWA studies, such examples are still rather the exception than the rule. The main challenge in translating common trait-associated markers into molecular pathways results from the haplotype structure of the human genome [121]. High LD between frequent markers complicates the selection of the most plausible candidate for inference of causality out of the set of correlating variants. As the variants' alleles occur together in almost all individuals, it is even possible that instead of a single causal variant the whole haplotype consisting of several correlating variants may affect trait predisposition by several separate mechanisms, for instance by affecting the expression of distinct genes by altering several regulatory elements. This again impedes the projection of variant effects on candidate genes that may be subjected to functional studies if the LD-block spans across multiple or no genes. And even if there is only one single gene, the possibility remains that the effector gene underlying the association may be located in a completely different genomic region if the variants affect *cis*- or *trans*-acting regulatory elements instead of the co-located gene.

For rare variants, on the other hand, the main challenge is to show that the individual variants are not neutral but indeed affect the trait. This is mostly done by predicting variant effects using a catalogue of genomic elements. There are several tools that provide predictions of functional consequences such as Annovar, the Ensemble Variant Effect Predictor, or SNPEff [139-141]. The most straightforward way to establish effectiveness of a variant was long considered to be the proof that a rare allele alters the amino acid sequence of a protein in a damaging way, leading to LOF or reduced activity either by functional amino acid substitutions, insertions of premature stop-codons or altered splicing. Therefore, the abovementioned tools all rely on algorithms like SIFT [142] or PolyPhen [143] that predict the extent of damage introduced by coding variants.

As mentioned before, this emphasis on variants located in protein-coding genes ignores the majority of association results. Yet, the ability to assess regulatory effects exerted by non-coding variants on a large scale has been established only recently. Although expression quantitative trait loci (eQTLs), that is, associations of genetic markers to changed levels of transcript expression, have been used to infer causality almost since the beginnings of GWAS, genome-wide datasets on the genomic localization of regulatory elements are released only since 2010. Due to the short timeframe since this shift of focus has taken place, there exist only few

examples where such regulatory variants have been thoroughly studied. Nonetheless, these studies show very promising results. For instance, a recent study on *cis*-regulatory variants affecting 21 autoimmune diseases was able to show strong enrichment of such variants in enhancer-like transcription factor binding sites (TFBSs) [144]. A related approach to underpin genetic loci with molecular mechanisms via an intermediate phenotype is the combination of GWAS with metabolomics screens (mGWAS). In inborn errors of metabolism, core enzymes of the human metabolism are impaired by rare mutations leading to extreme and eventually toxic metabolite concentrations. However, more common, less severe genetically influenced variations of metabolite levels can still provide valuable insights into the links between genetic variation and altered cellular functioning as well as global metabolic homeostasis. The first mGWAS was published in 2008 by GIEGER and colleagues, investigating associations between common SNPs and >350 metabolite concentrations [145]. Although the study had access to a limited number of samples ($n = 284$), they could still identify four significant associations between genetic loci and metabolite levels (metabolite quantitative trait loci, mQTLs). Several other mGWAS with larger sample sizes followed that could show that mQTLs can be replicated across cohorts, that they explain a significant fraction of the genetic variance of metabolic traits, and, intriguingly, that many of the identified loci harbor an enzyme that is functionally connected to the associated metabolic trait(s) [146]. For both eQTLs and mQTLs an enrichment of loci associated to clinical phenotypes became apparent, showing that the combination of data on different –omics can lead to further evidences usable for generation of hypotheses on the molecular mechanisms leading to human diseases.

However, considering all genes and their protein products, regulatory elements, eQTLs and their targets, the tissues and cell types where they are active, as well as additional intermediate traits such as mQTLs, and using all this data to annotate variant sets on a genome-wide scale is complicated by the complexity of comprehensive integrative analyses. Therefore, most studies limit their scope with regards to the tissue of interest, a single trait or a class of traits (such as the mentioned autoimmune diseases), or the set of genomic data that is incorporated into the variant annotation. The result is that, in order to be able to reuse a successful approach for other traits (or even if the setting is the same but additional association data become available), similar data integration approaches have to be applied before the actual analyses can be performed. Figure 6 shows the increasing pace at which GWAS results have been published for the last years, and with the advent of the broad use of NGS technologies the number of variants that need evidence-based annotation increases even faster. Thus, the redundancy in the application

of integrative approaches – that is still partly done manually – is a rate-limiting step in biomedical discovery and its translation to clinical use.

1.9 Objectives of this thesis

The central goal of this thesis is to address this bottleneck by demonstrating the benefits offered by automatized, re-usable integrative approaches to predict molecular effects of genetic markers linked to human traits and diseases. Beginning with the primary task of identifying genetic associations to disease and intermediate traits and ending with the generation of testable hypotheses regarding the putative molecular mechanisms underlying the association signals, I describe computational data processing pipelines, ready-to-use data integration resources, as well as integrative analysis approaches that facilitate the study of genetic variants and their potential impact.

The introductory **Chapter 2** illustrates the fundamentals of data integration, its complexity, and the associated obstacles with focus on applications in genomic and genetic studies. The technical descriptions of the major data integration frameworks and basic methods for integrating, harmonizing, and consolidating biological data are thus dissolved from the studies reported in **chapters 5 and 6** where I actually used the concepts. This is done in order to increase the accessibility of these more results-oriented chapters.

Chapter 3 lists the data, methods, and software tools used throughout this work. It introduces the cohorts and datasets that are used in the GWA studies reported in **chapter 4**. Further, it describes the main features of the developed modular workflows (or pipelines) to automatically perform complete GWAS and CNV analyses in concordance with best-practice guidelines. The chapter also includes the mathematical concepts, the bioinformatics tools, and the software utilized to study and annotate genetic variants as well as a comprehensive list of the datasets and resources incorporated in the integrative analyses described in **chapters 5 and 6**.

The results of three GWA studies are reported in **chapter 4**. The first GWAS investigates the sudden infant death syndrome (SIDS). Here, the process of performing a GWAS is introduced in detail, as is the protocol for calling CNVs using genotyping array data. The results of the GWAS show that SIDS is almost certainly not caused by strong complex genetic factors.

In an attempt to emphasize the complexity of linking genetic associations with disease traits to molecular mechanisms, the suggestive significant loci are annotated with the available evidences and we derive some plausible, albeit still speculative, hypotheses. The CNV analysis, on the other hand, provides evidence that for several of the included cases large deletions may have caused monogenic disorders resembling SIDS by leading to sudden unexpected death in infancy (SUDI) with inconclusive autopsy outcomes. Therefore, we suggest including cytogenetic screens into the standard protocol for autopsies if SIDS is considered as cause of death. The second and third GWAS explore the genetics of intermediate quantitative traits by searching for genetically influenced metabolite concentrations in human blood and urine samples. At the time of their publication, the studies were the largest in their fields, in combination identifying more than 160 genetic loci linked to metabolic traits. For both studies, it is adumbrated how additional data from genomic resources and text-mining can be included to support the process of linking biological information to genetic associations to enable the selection of the most plausible predicted causal genes. In the study on blood metabolites, this process was performed completely manually which was highly labor-intensive. In the investigation of genetic influences on urinary metabolic traits, we therefore developed a pipeline that performs most of the necessary annotation steps automatically.

Chapter 5 presents the data integration resource *SNiPA* that was used to develop the pipeline described in **chapter 4**. It contains a plethora of genomic annotations that are used for evidence-based characterization of genetic variants. The resource comprises genome-wide prediction of variant effects for the complete 1000 genomes variant set. Variant annotations are provided in a user-friendly webserver featuring the first genetic variant-based genome browser, as well as several other access modules that provide entry points to the analysis steps commonly applied in the interpretation of results from genetic association screenings. In this chapter, we also give a very detailed description of the development process of such a genomic resource, thus substantiating the introduction to data integration given in **chapter 2**.

In **Chapter 6**, I describe three studies that provide examples of how such an information basis can be used to advance our knowledge on the mechanisms underlying the genetics of complex human traits. In the first study, the concept of biological networks is introduced and applied to trait-associated variants to explore the specificity of genetic loci linked to human disorders. It is demonstrated that there is a substantial overlap of GWAS signals linked to distinct human diseases. By close investigation of the allelic structure of such overlapping genetic loci it is shown that there is evidence for both common pathways linking the etiology of similar

disorders and pleiotropic effects that, depending on the present alleles, predispose to one disease while protecting against another. The second study then describes the application of integrative analyses of the effects of genetic risk factors on specific regulatory entities by intersecting SNPs with miRNA target sites. It is shown that there is significant evidence for the interrelations of trait-linked variants and the regulatory level, indicating the complex mechanisms underlying the etiology of complex disorders. To investigate regulatory effects of non-coding genetic variants further, the third study compiles a novel clustering of gene-associated promoter and enhancer elements. Using chromatin immunoprecipitation DNA-sequencing (ChIP-seq) data for the annotation of active regulatory sites, we show that the annotations from distinct datasets on regulatory elements conform quite well to each other. Using eQTL data for assessing the applicability of these regulatory clusters in the assignment of the target genes of putative regulatory-acting genetic variants, we demonstrate that, to a large proportion, the correct genes can be predicted using the compiled set of regulatory elements.

The scientific contributions of this thesis are then summarized, embedded in the context of the field, and augmented with potential future research directions in the final **Chapter 7**.

2 Data integration

In the study of genetic variation and its effects on the phenotype, consideration of all biological entities possibly involved in the translation from genotype to phenotype is the key to identify the particular changes that lead to human disorders and that simultaneously may be targeted by therapeutic intervention. The collection, harmonization, consolidation, and provision of all data in a machine-accessible way is the task of a discipline called data integration that, when applied on biological data, is a subfield of computational biology. The complexity of the interrelations of biological elements has been adumbrated in the previous chapter and is detailed and illustrated further. The following also describes the process of integrating complex biological data without loss of information for the available knowledge to be usable by automated methods searching for patterns of genetic influences and how they may affect the development of complex traits. The introduced concepts are used in chapters 5 and 6, where the implementation of the single steps is described in detail.

2.1 Data integration frameworks

When investigating interrelations between biological entities originating from differing sources, it is important to ensure comparability of these sources to provide a unified view on the

complete data on which the global analyses are performed. In general, there are two different theoretical frameworks to achieve this: local-as-view (LAV) and global-as-view (GAV) [147]. Both frameworks consist of a global (unified) schema, source schemata (the primary data), and a mapping that describes the formulation of queries to the global schema by means of queries submitted to the source schemata. In LAV frameworks *(database federations)*, the sources are integrated in the global schema by expressing each source as a view on the global schema. This way, the global view is logically detached of the particular sources enabling efficient extension by additional sources, especially if these sources are of similar format. The efficacy of querying the source data, however, is affected as the results provided by each source view have to be merged and cleansed in accordance to the assertions established via the mapping to the global view. In GAV frameworks *(data warehouses)*, on the other hand, the global schema is expressed as views over the sources, is thus directly linked to the source schemata, and can be queried efficiently. Including additional sources, however, is complex as the definition of the global view is affected by each newly integrated source, making GAV frameworks rather inflexible when it comes to the integration of complex and dynamic data.

Integration of genomic data is often solved by a hybrid of the two frameworks *(database federations with mediated schemas)* [148]. This is due to the nature of genomic data. The backbone of all genomic data is provided by the reference genome sequence. To this sequence, all other physical entities such as genes, transcripts, variants, and regulatory regions are mapped via a coordinate system. Secondary data such as specific annotations for, interactions between, and cell type specificity of these entities, are then linked to the entities either by physical position or via entity identifiers. While the schema of the genomic sequence and the mapping of entities to it is very stable (unless a new entity type is discovered that cannot be expressed as one of the existing entities), this data can be conveniently stored in a GAV framework. This has the advantage that these mappings, containing hundreds of millions of data points, can be queried efficiently. In addition, upgrades to the GAV are only required if there are major updates to the underlying data sources – such as the release of a new genome assembly. Patches of the reference genome or mapping updates of contained entities can be easily applied. The more dynamic – as accumulating with every additional experiment – secondary data can be handled by a LAV framework. To be extendible, the LAV framework only needs to provide generalized mappings for each included experiment type. As data formats reporting the results of experimental protocols get more and more standardized, considerably less adjustment is needed for the LAV to include new data sets. Using a third mapping layer (the mediated schema), the

LAV and GAV frameworks are then linked into an analysis platform where the integrated and unified data can be investigated while data sources remain adjustable, updatable, and extendible. For the resource described in chapter 5, I used this hybrid to separate the more dynamic data from the genomic backbone, in order to enable several updatable layers.

2.2 Data harmonization

In the previous section, the data integration frameworks suitable for genomic data have been introduced. This section now describes how the data has to be harmonized for the annotation of genetic variants. As mentioned, genomic entities are usually aligned to the human reference genome sequence. This implies that there is a universally used reference genome which, in practice, is not the case. The Genome Reference Consortium (GRC) continues to decipher the complete sequence of the human genome and, during this process, patches of the reference sequence as well as fully-fledged new assemblies are released. This complicates the unification of genomic data sources as they may be mapped to different assemblies (or patches of assemblies) leading to inconsistencies of the mapping coordinates.

Therefore, the first task in integrating genomic data is to obtain a mapping of the coordinates of all contained entities to the same version of the reference sequence. Large resources like Ensembl [149], the National Center for Biotechnology Information (NCBI) [150], or the University of California Santa Cruz (UCSC) genome browser [151] integrate many sources conform to the same assembly version. For custom data sets, however, this task resides at the researcher. There are tools available at each of these resources that map genomic coordinates between assemblies, but there remain some obstacles. For instance, the current Ensembl gene build (GENCODE 21 [152]; contains genes, transcripts, and protein sequences) is only available for the current genome assembly GRCh38 while many other genomic data is only annotated for the older GRCh37 assembly. Global mapping of coordinates between these assemblies is not trivial as some sequence parts present in the older assembly are missing in the new one and vice versa. Also, in the realignment for the GRCh38 assembly, some sequences from the old assembly have been mapped to different chromosomes with some break points splitting entities mapped to one location in the old assembly into two or more pieces. Resolving

such discrepancies is very complex as it is often only achievable by merging the data available for two (or more) assembly versions. This was one of the major obstacles in designing the data integration resource described in chapter 5, where I propose an approach that solves this problem. A similar issue is the mapping of locally aligned genomic elements to the global coordinate system. For instance, binding sites of RNA-binding proteins (RBPs) are often mapped to the processed (exonic) transcript RNA sequences. In order to project coordinates relative to transcript positions back to the genomic coordinate system, RNA processing has to be reversed without losing the information of a binding site overlapping a splice site leading to a pair of genomic ranges instead of a single coordinate location. Individually, these steps do not rely on complex algorithms, however, to be performed on genome-wide scale, each step has to be automated and therefore all eventualities have to be taken into account before processing the data. For instance, in the study on interrelations between trait-associated SNPs and miRNA regulation pathways (chapter 6), we used the exonic sequence of human transcripts for prediction of miRNA target sites which were then mapped back to genomic coordinates to search for genetic variants affecting the obtained targeting intervals.

The second task is to unitize all sources containing data on genes, transcripts, and proteins to a single gene build. The three abovementioned resources all feature own gene builds each differing from the others and many further resources on genes, transcripts, and proteins are available. Secondary data sources on these elements are created using the gene build of the researcher's choice, leading to great heterogeneity of information. This problem has been recognized and led to the initiation of projects such as the consensus coding sequence (CCDS) project [153] that aims at building a common consensus set of protein-coding genes. However, CCDS identifiers (and others of the kind) are still used rather unfrequently. Mapping of genes, transcripts, and proteins between different gene builds therefore remains a major challenge in large-scale genomic data integration approaches.

The third task is the deduplication of data. Entity types such as SNVs or transcription factor binding motifs are, too, referred to using several distinct identifiers across resources or even within the same resource. To avoid ambiguity of the integrated data as well as to provide quick access to each unique entity, identifiers have to be merged to a single entry and, equally important, kept up-to-date to be able to link to the active entries of each source. To be able to integrate older data sources that may contain "retired" identifiers later on, building of an identifier mapping history is a further essential step.

2.3 Data consolidation

There are different ways of how to consolidate, to merge, and to annotate integrated data. Depending on the data representation, the implementation of (inter)relations between entities requires differing amounts of programming and different means of data optimization. Here, the most important data representation frameworks and their strengths and weaknesses will be shortly introduced.

RELATIONAL SCHEMAS – Maybe the most established data representation is the relational schema implemented in database management systems (DBMSs) like SQL or Oracle. All large data warehouses (Ensembl, UCSC, dbSNP, and others) use MySQL databases for data representation. In general, to store data in a relational schema, the data has to be normalized and transformed in order to be representable by relations. This bears the greatest benefit of relational databases: the data has to fulfil quality properties before it can be inserted in the schema. Relational storage engines display data as a set of transactions that all have to provide the ACID (Atomicity, Consistency, Isolation, Durability) properties as a minimal prerequisite [154, 155]. The most criticized weakness of relational databases is the query efficiency: to conform to the ACID principles, the data is normalized and split into several relations (tables) such that redundancies are minimized. Normalization requires that all entries in one relation contain comparable data, meaning that data fields have to be of identical format (numbers, characters, etc.). To get information on entries, the data distributed across the relations has to be rejoined during the query process. This is done using keys that are overlaid with index structures for quick access. The query performance of such indices is high if the keys are simple, such as unique names (identifiers) for entities. Coordinate-based queries as for genomic data, on the other hand, are not very efficient as the keys consist of three values (chromosome, start position, and end position) which complicates index construction resulting in very large (and, thus, inefficient) index structures. To optimize relational databases for genomic coordinate-based queries, chromosomes are split into smaller bins to reduce the size of the index tree. The disadvantage of this approach is, of course, that the inversed query, that is retrieval of the genomic coordinates of a specific entity, is slower because the bin structure has to be resolved beforehand. However, as there are by far fewer bins per chromosome than possible coordinates per whole chromosome, this optimization is very efficient. In summary, to use relational schemas for data representation, the highest workload has to be put into input data

normalization to obtain complete and structured data. Access to the data via queries as well as compliance checks with ACID properties is provided by the DBMS. Therefore, implementation of consistency measures or access interfaces is unnecessary. Efficiency optimization is query-based and can be adjusted to the use-case.

SEMI-STRUCTURED DATA – A major drawback associated with the relational schema is its perceived inflexibility regarding the inclusion of unstructured data types. Although basically relational schemas can represent any kind of data, the rigid technical structure of the database schemata complicates the illustration of data in an intuitive way. Semi-structured data representation formats such as XML (extensible markup language) or derived formats like HTML (hypertext markup language; used for displaying websites) provide a backbone to structure *per se* unstructured data. This is achieved by using markup tags that encapsulate the data and make entities identifiable. Using metadata style sheets, tags can be bound to data types, but this is optional and, contrary to relational representations, violations against the style specifications are tolerated. Additionally, tags can be self-defined, leading to unlimited possibilities of designating data. However, for these reasons, checks for consistency, completeness, and structure of the data as well as the provision of a query language to obtain data has to be implemented for each representation individually. The efficiency of accessing the data is highly dependent on this implementation.

One bioinformatics application of XML-like representations (e.g. GRAPHML) is the storage of network structures. A network is defined as a graph $G = (V, E)$ consisting of a set of vertices/nodes V and a set of edges E connecting the vertices. If there are several distinct sets of nodes $V = V_1 \cup \cdots \cup V_k$, the resulting graph is called k-partite. In biological networks, a node is defined as any biological entity like a genetic variant, a gene, or a phenotype. Edges linking vertices therefore define some biological connection, such as interactions, correlations, genomic co-location, or associations. To represent such diverse data, relational schemas would require relations for each node and edge type. In GRAPHML, the node and edge tags can contain any sub-tags defining various properties without the necessity of converting the data into equal formats. By using or ignoring tag classes of nodes and edges, entities can be included or excluded in global or local analyses, making semi-structured data representations prime candidates for network analysis. In our study on genetic correlations between complex traits (chapter 6), we utilized these properties both to store complex and diverse data and to closely investigate network properties by selecting sub-networks by different tags.

In summary, semi-structured data representations allow for fast, individualized and intuitive description of information. The highest workload has to be put into developing interfaces to access the data via queries, add new entries, as well as check for data consistency. Efficiency of accessing the data can range from very high to very low depending on the nesting depth and complexity of the data structure and the implementation of information traversing.

STRUCTURED DATA/ONTOLOGIES – As mentioned above, semi-structured data representations can be bound to data types using style sheets. Via extension using an ontology, that is including object-like attribute assignments to tags and metadata assertions that prohibit violations of data types and attribute specifications, semi-structured data can be converted into structured data representations [156]. Examples include OWL (web ontology language) or NoSQL (not only SQL) DBMS, however, structured data can be obtained by any assertion-based programmatic approach. As assertions are, in contrast to the implementation in relational schemas, detached from the data and can be defined for any subset of the data, structured data representation frameworks are less rigid than relational representations. This, however, again leads to the need for implementing checks for consistency, completeness, and structure of the data as well as the provision of a query language. In the age of big data, structured data representations are becoming increasingly popular because of their seemingly unlimited flexibility. In the process of integrating mGWAS results for reporting genetic associations with metabolic traits described in chapter 4, we incorporated a metabolite ontology in order to be able to categorize metabolites into pathways and use this categories for downstream analyses. However, as the performance of data management, structuring, and access relies heavily on the implementation of the structured framework (and also because query efficiency is often achieved by introducing data duplicates, and thus ACID violations, on purpose), there is a movement towards more efficient relational schemata (so-called NewSQL systems) [157]. Ontologies that are very standardized by assertions and thus avoid the mentioned shortcomings are very useful and valuable, but due to predefined structures as inflexible as relational schemata.

2.4 Data integration in human genetics

The complexity of interpreting the molecular effects of genetic variants linked to human traits lies in the non-linear interplay between the different regulatory layers that in concert

consummate the translation from the genotype to the phenotype (Figure 9). Understanding the processing of information between these different layers is the key to enable genotype-based prediction of phenotypic outcomes. However, despite the growing amount of data generated by the different –omics fields, current studies mostly investigate interrelations between only two of the layers. This results from the *curse of dimensionality* (termed by RICHARD BELLMAN [158]) – the lack of statistical power to detect significant interactions between the layers due to the exponential growth of data dimensions when combining –omics data of different types. A notable exception that is becoming more and more popular is posed by analyses of Mendelian randomization [159] (based on MENDEL'S principle of independent assortment), where it is tried to identify the effect direction exerted by a variant that is associated both with a trait with effect $\beta_{SNP \rightarrow Trait}$ and a potential intermediate phenotype (e.g. expression levels of transcripts) with effect $\beta_{SNP \rightarrow intermediate}$. In this setting, the two effect estimates are combined, for instance using the Wald ratio method [160], to obtain an estimate if the variant effects only the intermediate phenotype that then influences trait susceptibility or if the genetic effects on the two phenotypes are independent. However, until now there are only few cohorts available that have been subjected to analyses of more than two –omics types in sufficient sample sizes to provide the necessary statistical power for multi-dimensional analyses, and therefore different means of integrating the available data on the distinct layers are needed.

A commonly used approach is to collect the available evidence on the variants found to be associated with the trait under study as well as on the potentially affected genes from published data to formulate hypotheses on the molecular effects exerted by those variants. For instance, in 2010 the GABRIEL consortium published a genome-wide association meta-analysis investigating genetic predisposition to asthma [161]. One of the findings of the study is the replication of a previous association between SNPs within the chromosome 17q21 locus and childhood-onset asthma. In their rationale, the authors refer to two eQTL analyses that detected significant associations between the same SNPs and the expression of two genes within 17q21, *ORMDL3* and *GSDMB*. As a third study showed that changed expression levels of the yeast *ORMDL3* homolog leads to disturbances in sphingolipid metabolism and a fourth study observed the involvement of sphingolipids in inflammatory processes, the authors conclude that the associated locus is linked to asthma via *"MODULATION OF AIR WAY INFLAMMATION"* [161] due to an intermediate phenotype, i.e. the altered metabolism of sphingolipids. Thus, based on the outcome of a study on two layers (variance of the genome and asthma as part of the

phenome), the authors use published data to include two further layers, namely the transcriptome and the metabolome, in their hypothesis on the predisposing molecular effect.

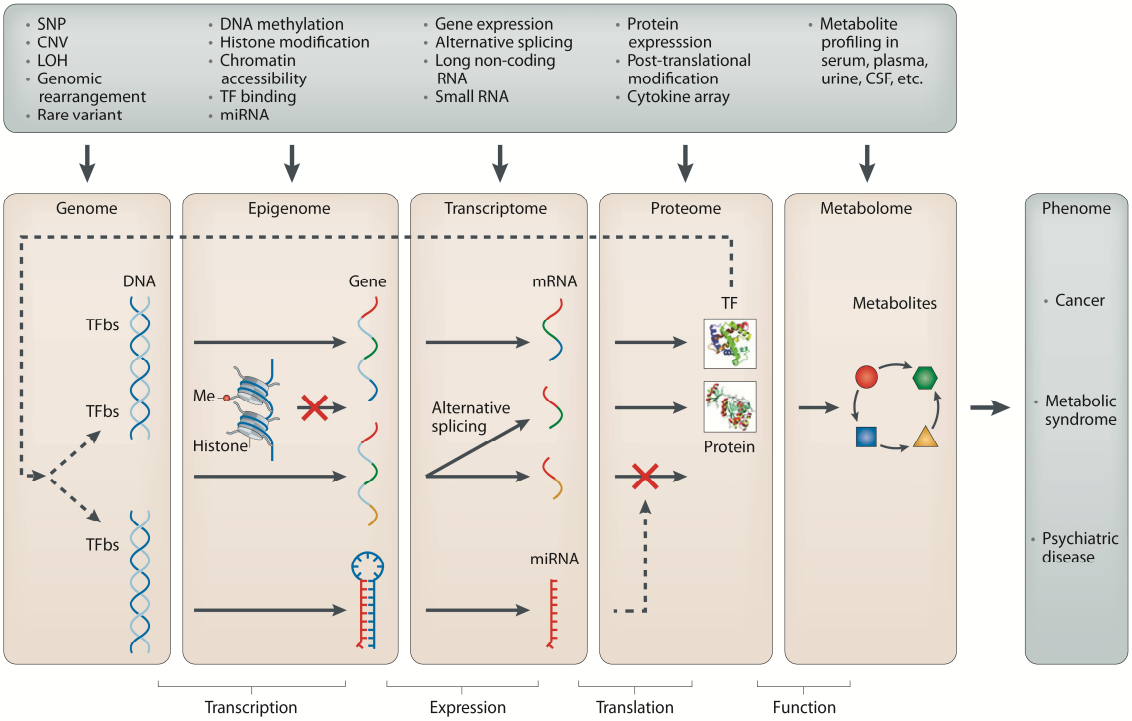


Figure 9: **Schematic view of the different –omics layers from genome, epigenome, transcriptome, proteome to the phenome**. Examples for biological entities and mechanisms are given. LOH: loss of heterozygosity; TFbs: transcription factor binding site; Me: methylation; CSF: cerebrospinal fluid; Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [162], copyright 2015.

The process of integrating the available data into study-specific context can be referred to as *results integration*. With the advances in –omics measurement techniques, nowadays there are many further possibilities to corroborate hypotheses with molecular evidence. Modelling of the effects of genetic variation across all available layers using results integration or directly combining measurements in multi-dimensional analyses is also referred to as *systems genomics*. In their recent review, RICHIE et al. describe the promises of systems genomics as follows:

> *"THE REDUCTIONIST PARADIGM OF LOOKING FOR THE 'LOW-HANGING FRUIT' (THE SINGLE VARIABLES THAT EXPLAIN SOME PORTION OF TRAIT VARIABILITY) IS SLOWLY BECOMING LESS PREVALENT. NOVEL QUESTIONS WILL BE ASKED ABOUT THE COMPLEX INTERPLAY OF DIFFERENT TYPES OF OMIC DATA USING NEW STATISTICAL AND MACHINE-LEARNING APPROACHES AS MORE*

researchers think 'outside the box'. These emerging systems genomics approaches yield more informative results, and the pace of development will accelerate. As the tools become more readily available and affordable, such systems genomics approaches will prevail as the dominant type of study design and analytical strategy — the days of studying molecular data variability in isolation are slowly coming to an end."

Ritchie et al., 2015 [162]

3 Materials and methods

This chapter introduces the cohorts, the genotyping arrays, applied quality control measures, and the data sets and software tools that were used for genotype imputation, GWAS analyses, and CNV calling and filtering. Further, it lists the data sources incorporated into integrative resources utilized for variant annotation, the used software, as well as a summary of the most relevant mathematical and statistical concepts used in the different studies. The materials and methods described in this section have been partly published in our papers [58, 116, 163–165].

3.1 Description of cohort data

3.1.1 German study on sudden infant death (GeSID)

The German study on sudden infant death (GeSID) [166] recruited 455 infants succumbed to sudden infant death syndrome (SIDS) in 18 study centers between 1998 and 2001. Cases were examined using a standardized autopsy protocol including morphological, histological, toxicological, and neuropathological parameters as well as microbiology and virology screens. For 373 cases, parents consented both to fill out a comprehensive questionnaire and to provide a thorough family history. Only infants suffering from sudden unexplained death where no clear

causes of death could be determined postmortem were classified as SIDS ($n = 331$). For 317 of those, genome-wide genotyping was performed using Illumina HumanHap660W-Quad BeadChips (657,366 markers).

3.1.2 Sheffield Children's Hospital SIDS Cohort (SCHC)

The Sheffield Children's Hospital recruited 121 cases of sudden unexpected death in infancy (SUDI) from the UK in a three-year period from 2004 to 2007 [167, 168]. SUDI cases underwent complete autopsy following a comparable protocol as in GeSID. Thereon, 51 infants were classified as succumbed to SIDS as no clear cause of death could be identified postmortem. Blood samples were available for 48 of these cases and were used for genome-wide genotyping using Illumina HumanHap660W-Quad BeadChips (657,366 markers).

3.1.3 Kooperative Gesundheitsforschung in der Region Augsburg (KORA)

The Kooperative Gesundheitsforschung in der Region Augsburg (KORA) [169] unites a set of epidemiological surveys and follow-up studies of participants from the general population in the region of Augsburg in southern Germany. The analyses described in this thesis include subsets of the data from the follow-up study KORA F4 (2006–2008) of the KORA S4 survey (1999/2000). As population-based controls for the study on genetics of SIDS, we used 823 healthy individuals (425 males and 398 females) which were genotyped on the Illumina HumanHap550-Quad+ BeadChip (539,741 markers). Genotypes were called using Bead Studio. As age matching is not feasible in SIDS, we performed sex-adjusted association analysis as well as right-censored Cox hazards regression to include age at death as a covariate. Further, for the characterization of genetic influences on metabolite concentrations in blood samples, we included 1,768 subjects (858 males and 910 female) from the F4 study which were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0. Genotypes were called with Birdseed 2. In this study, age, sex, and body mass index (BMI) were included as covariates [116]. And finally, in the analysis of genetic influences on urinary metabolite concentrations, we included 1,691 individuals (826 males and 865 females) from the F4 study. Those are a subset of the above 1,768 subjects genotyped on the Affymetrix 6.0 array. Here, age and sex were included as covariates [165].

3.1.4 UK Adult Twin Registry (TwinsUK)

In the UK Adult Twin Registry (TwinsUK), twins from the general population in the UK were recruited through national media campaigns [170]. Of the 6,056 samples used in the study on genetically influenced metabolite concentrations in human blood, 93% were female in the

age range of 17 to 85 years. Genotyping was performed with a combination of Illumina arrays (HumanHap300, HumanHap610-Quad, Human1M-Duo and Human1.2M-Duo 1M) and calling was performed with Bead Studio. As for KORA F4, age, sex, and BMI were included as covariates [116].

3.1.5 Study of Health In Pomerania (SHIP)

The Study of Health In Pomerania (SHIP) [171] was planned as a cross-sectional *"CLASSIC PREVALENCE STUDY"* [172] and later expanded into a longitudinal population study comprising three measurement time points: the initial survey (SHIP-0; 1997-2001), a five-year follow-up (SHIP-1; 2002-2006), and a twelve-year follow-up (SHIP-2; started in 2008). For the study of genetically influenced urinary metabolite concentrations, we used 3,861 (1,901 males and 1,960 females) individuals from the 4,308 SHIP-0 subjects for whom both genotype data (obtained using the Affymetrix Genome-Wide Human SNP Array 6.0 and called using Birdseed 2) and urine samples were available. As for KORA F4, age and sex were included as covariates [165].

3.1.6 Control cohort from the PopGen biobank

The PopGen biobank [173] was launched in 2003. Recruitment between 2005 and 2007 resulted in 1,317 study participants from the general population in Schleswig-Holstein, Germany. For 678 (257 males, 421 females) of those, genotype data was obtained using the Illumina HumanHap550-Quad+ BeadChip (539,741 markers) [174].

3.1.7 Metabolic profiling of KORA F4 and TwinsUK blood samples

We used the mass spectrometry (MS)-based platform of Metabolon Inc., Durham, USA, for profiling of 529 plasma and serum metabolites. Of those, 333 are of known identity and fall into a wide range of biochemical classes: amino acids, acylcarnitines, sphingomyelins, glycerophospholipids, carbohydrates, vitamins, lipids, nucleotides, peptides, xenobiotics and steroids. The remaining metabolites are "unknowns", meaning that they can be reproducibly measured but their chemical identity has not been identified yet [116].

3.1.8 Metabolic profiling of KORA F4 and SHIP-0 urine samples

Overnight-fasting urine samples for KORA and non-fasting, spontaneous urine samples for SHIP-0 were collected and stored at -80°C until analysis. Nuclear magnetic resonance (NMR) spectra were acquired at the University of Greifswald, Germany, using a Bruker DRX-400 spectrometer (Bruker BioSpin GmbH, Rheinstetten, Germany) [165]. Baseline-corrected, Fourier-transformed spectra were manually annotated (spectral pattern matching, Chenomx

Worksuite 7.0, Chenomx Inc., Edmonton, Canada) to retrieve 60 compound concentrations (including creatinine) for targeted analysis and automated spectral alignment and feature extraction implemented in the FOCUS software [175] to obtain individual NMR peaks for non-targeted analysis. To obtain absolute concentrations, signal intensities of NMR peaks were normalized using trimethylsilylphosphate of which 0.5mM was added to the samples before measurement. To account for dilution effects, signal intensities were additionally normalized using creatinine concentrations [165] (it has been shown that, provided kidney function is normal, in absence of dilution factors creatinine concentrations in human urine are close to constant to a value of 1g creatinine per 20kg muscle mass per day [176]).

3.1.9 Ethics

All 19 centers providing SIDS case data obtained the approval of their local medical ethics committees. All participants in SHIP-0, TwinsUK, PopGen, and KORA have given written informed consent and the respective local ethics committees (SHIP: ethics committee of the University of Greifswald; TwinsUK: Guy's and St. Thomas' Hospital Ethics Committee; PopGen: ethics committee of the Christian-Albrechts-University in Kiel; KORA: ethics committee of the Bavarian Chamber of Physicians in Munich) approved the studies.

3.2 GWAS analysis pipeline

During the GWAS on SIDS, I developed this pipeline consisting of a modular collection of BASH, PERL, and R scripts covering the steps of data preprocessing, quality control (QC), association analysis, imputation, and GWAS meta-analysis. In the Bachelor thesis of CHRISTOF SCHRAMM, the modular structure of the pipeline has been used to develop an automated GWAS analysis workflow system utilizing the graphical user interface of the KNIME (Konstanz Information Miner) platform [177].

3.2.1 Data preprocessing

There are several prerequisites that have to be met for variant data to be subjected to merging of data sets as well as for imputation. For instance, to impute genotypes using reference panels, all alleles have to be designated by the forward or plus strand allele of the same reference

genome. The same holds true for merging processes of several genotype datasets to avoid erroneous allele counts if allele codes are ambiguous. The first module of the pipeline therefore performs strand mapping and simultaneously adjusts the genomic coordinates of markers (e.g. for results plots) using a backbone of manufacturer manifest files of different genotyping platforms (Illumina and Affymetrix, obtained at the manufacturers' websites) as well as up-to-date variant coordinate mappings obtained at dbSNP [178]. I implemented the module also to provide the functionality of merging data sets correctly. This is performed in three steps: first, the core data set is updated with correct variant positions and alleles are flipped to the plus strand where necessary. Second, the marker set of the data set that shall be merged is compared to the core set and discordant missing / monomorphic markers are identified and cached. Alleles are then remapped and flipped if necessary and the data sets are merged. Merging is performed using the PLINK software [179]. Afterwards, the variants marked for removal in the previous step are filtered from the merged files.

3.2.2 Quality control

The large study populations and the huge count of included markers in GWAS bear a large potential for false positives. Therefore, stringent QC is indispensable and, therefore, best practice guidelines for QC in GWA studies have been established [180] and implemented in the module. These include:

1. Comparison of the sex predicted via the genotype data (estimated homozygosity of X-chromosomal markers) to the assigned sex for all individuals and exclusion of individuals were the two values mismatch
2. Computation of the kinship coefficient ($\hat{\pi}$; this is basically the correlation between related individuals described by FISHER) and iterative exclusion of one individual per pair exceeding a certain threshold for this coefficient (default is $\hat{\pi} > 5\%$) while minimizing the number of excluded individuals
3. Calculation of principal components or multidimensional scaling (MDS) dimensions for inclusion as covariates to account for population stratification. These can also be used to identify outliers by cluster analysis
4. Removal of markers with low overall call rate (default is exclusion at call rate $< 98\%$)
5. Removal of individuals with low overall call rate (default is exclusion at individual call rate$< 98\%$)

6. Removal of markers with minor allele frequencies (MAFs) which fall short of the range for which the study is statistically powered to detect associations (default is MAF$< 5\%$)
7. Removal of markers significantly violating the HWE. This value is, however, not to be set too strictly as markers associating with a trait also violate the HWE. Therefore, if this filter is applied on the whole data set (cases and controls) the P_{value} threshold for the test of compliance with HWE is to be set lower than if only including controls for HWE filtering (default is $5 \cdot 10^{-5}$ vs. $5 \cdot 10^{-3}$ for controls only)

As these filter steps are all implemented in PLINK, this module wraps and concatenates the consecutive PLINK commands. In addition, I wrote customized plotting scripts that are executed for steps 2 and 3 and PLINK log and statistics files are parsed and summarized in a single report.

3.2.3 Association tests

This module automatically checks if the studied trait is qualitative or quantitative and choses the right regression model accordingly. Covariates such as sex and MDS dimensions are automatically included. These can be selectively excluded as well as additional covariates can be included. Output P_{value} are adjusted for multiple testing and data files for plotting of the association results are generated. Except for survival analysis, this is done again by wrapping PLINK runs. For survival analysis, I wrote custom R scripts fitting a Cox proportional hazards regression model for the common genetic models (additive, dominant, recessive, over-dominant, and genotypic model) with and without covariates that can be called using PLINK via the Rserve connector [181, 182]. P_{values} for Cox regression are calculated using likelihood ratio tests (LRT) with one degree of freedom for additive, dominant, recessive, and over-dominant models and two degrees of freedom for the genotypic model. Association analysis is always followed by customized R scripts producing plots of the results (QQ-plot, Manhattan plot). For significant hits, regional association plots are automatically generated.

3.2.4 Genotype imputation

This module performs pre-phasing of the study genotypes using SHAPEIT2 [183] followed by genotype imputation using IMPUTE2 [97]. For this, first markers are split into chromosomes and then pre-phased chromosome-wise. This can be parallelized by a built-in SHAPEIT2 option. In the second step, genome-wide genotype imputation is performed on 5Mb chunks of the chromosomes with double-sided 250Kb overlaps. This can be either performed locally exploiting the count of CPU cores set by the user or completely parallelized

by batch-submission to a grid engine. Afterwards, imputation results are merged and filtered for the imputation quality (IMPUTE2 info score, default is info ≥ 0.8). IMPUTE2 output is then converted back to PLINK format (this is done either with GTOOL [184] or PLINK v1.9 [179, 185] using maximum likelihood genotypes) for further analysis.

3.2.5 GWAS meta-analysis

This is a custom R package for different approaches to GWAS meta-analysis. It implements the inversely weighted Z-score combination method [186] and the inversely weighted fixed and random effects models [187]. For more details, see section 3.6. Meta-analysis was used for the calculation of association statistics for X-chromosomal markers in the study of the genetics of SIDS (chapter 4). For this, association statistics for males and females have been computed separately and afterwards combined using this module. As the results showed no significant associations, this was omitted in the discussion of the results, in order to evade the unsolved, yet highly controversial, discussion about the male/female ratio seen in SIDS cases.

3.3 Copy number variant analysis pipeline

Copy number variant analysis requires the raw intensity data from the genotyping experiments. For Illumina BeadChips, intensity data comes in two IDAT-files per individual which each contains the intensity (red/green) measured for one of the two alleles of all markers.

3.3.1 Preprocessing and intensity-based marker QC

The pipeline performs conversion of IDAT-files, intensity quantile normalization, and removal of batch effects and array-to-array variability. For this, I used the Corrected Robust Linear Model with Maximum Likelihood Classification (CRLMM) R package [188] embedded in Bioconductor [189, 190]. Markers are filtered for the genotyping confidence score (>0.9) provided by CRLMM and for outliers from the normalized intensity distribution. Individuals are then filtered by individual call rate <95% and markers with genotyping call rate <95% are excluded. Furthermore, the overall signal-to-noise ratio (SNR) of all individuals are inspected and samples with $SNR < \mu(SNR) - 2\sigma(SNR)$ are excluded. A two-sample Kolmogorov-Smirnov-Test against a normal distribution with $\mu = \mu(SNR)$ and $\sigma = \sigma(SNR)$ is performed. If

the test detects a deviation of the normal distribution, the outlier removal is repeated on the filtered set of SNR values until the SNR distribution is approximately normal.

3.3.2 CNV calling

As CNV analysis software is prone to produce false-positive calls, I use two different tools (PennCNV [191] and the VanillaICE R package [192]) for CNV calling. PennCNV uses the log R ratio (LRR), a normalized representation of the total intensity for both alleles of a marker, and the B allele frequency (BAF), the normalized allelic intensity ratio of the two alleles, for CNV calling. These summary statistics are generated using an inbuilt CRLMM function and stored in PennCNV input format for each marker for each individual. For data input to VanillaICE, CRLMM estimates the raw, allele-specific copy number for each marker per individual. These estimates are centered to mean 2.0 ("normal" copy number for diploidy) for autosomes and X-chromosomes of females, while X-chromosomal markers of males are centered to 1.0. Centered estimates are then passed to VanillaICE which uses a hidden Markov model for smoothing and segmentation of copy numbers.

3.3.3 Quality control

I use the output generated by PennCNV for determining QC thresholds for outlier removal by means of number of CNVs and $\sigma(LRR)$. The threshold for numbers of CNVs is determined as the 90% quantile of the total distribution, while samples with $\sigma(LRR) > 0.25$ are excluded as recommended by PennCNV. The remaining CNV calls are intersected with VanillaICE calls and only CNVs that show at least 80% overlap between the two sets with the same copy number are retained for further analysis. Afterwards, CNVs overlapping with centromeric and immunoglobulin regions are removed. In a last step, I mark common CNVs by comparison with public control data from the Children's Hospital of Philadelphia CNV database [193] and the Database of Genomic Variants [194]. For computation of CNV burden and additional filtering in the SIDS study, I had access to control CNVs from a cohort of control samples (HYPERGENES Consortium [195], $n = 2{,}764$) that were analyzed for CNVs using the same methods as described above and genotyping arrays (Illumina Human Omni1M) with very similar marker content (A. MACÉ and Z. KUTALIK, *personal communication*). CNVs are further scored by a scoring method developed by A. MACÉ (*personal communication*).

3.4 Genomic resources

3.4.1 Genomic annotations and conservation/deleteriousness scores

ENSEMBL – The Ensembl project is a resource jointly developed by the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI). Its database includes a wide range of genome-level datasets [149] and provides an established backbone of annotations for the human genome. For genome annotation we downloaded GENCODE gene data (including OMIM [106] and DECIPHER [196] disease annotations) as well as associated transcripts and proteins, regulatory feature clusters and transcription factor binding motif data as well as linked information from the public MySQL database. We also used many of the allele-based annotations that are provided with the Variant Effect Predictor (VEP) [140] annotation, such as SIFT [142] and PolyPhen [143] predictions on the deleteriousness of non-synonymous nucleotide exchanges.

UCSC GENOME BROWSER – The UCSC table browser [197] is a data retrieval tool that allows access to the genomic annotations contained in the UCSC genome browser [151]. From its design, the UCSC genome browser is similar to the Ensembl database. A significant difference is that, while Ensembl for a long time exclusively used its own gene model and only integrated external gene models like NCBI reference sequence (RefSeq) transcripts [198] that could be mapped to one of their own transcripts (this has changed since Ensembl version 80), the UCSC genome browser integrated different gene models and did the mapping between different sources and its own gene model afterwards. Therefore, we used the UCSC table browser for retrieval of the original RefSeq gene, transcript, and protein annotations.

CONSERVATION SCORES: PHYLOP, PHASTCONS, AND GERP++ – Sequence conservation across species is an important indicator of the structural and functional importance of a nucleotide or sequence region. By now, many different scores have been developed that use different measures of conservation. In this work, three frequently used scores are utilized. Descriptions and interpretation guidelines of the scores are listed in Table 3.

Positional phyloP- as well as phastCons-100way-alignment PHAST conservation scores [199] were retrieved from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP100way/hg19.100way.phyloP100way.bw and http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons100way/hg19.100way.phastCons.bw. Further information on assemblies used in the

100way alignment can be obtained at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP100way/. GERP++ positional RS ("rejected substitutions") scores [200] were downloaded at http://hgdownload.cse.ucsc.edu/gbdb/hg19/bbi/All_hg19_RS.bw.

COMBINED ANNOTATION DEPENDENT DEPLETION (CADD) – KIRCHER et al. provide an annotation-aided score for genotype pathogenicity called CADD [201]. Genome-wide CADD-Scores were obtained from http://cadd.gs.washington.edu/download and mapped to 1000 genomes genotypes using allele-matching. We used the PHRED-like transformation of the C-score for variant annotation. Score description and interpretation is given in Table 3.

POLYA DB – The PolyA DB provides information on and the location of polyadenylation sites (polyA sites) for more than 25,000 human genes [202, 203]. The corresponding most abundant polyA signal variations are listed in [204].

Score	Description	Interpretation
phyloP	phyloP is a conservation score represented as -log(P) of a test for neutral evolution of a nucleotide.	Positive score: The position is predicted to be rather conserved. Negative score: The position is predicted to be rather fast-evolving.
phastCons	phastCons is a conservation score represented by the probability (i.e., range is 0 to 1) for a nucleotide to belong to a conserved element.	High score (max. 1): The position is predicted to be rather conserved. Low score (min. 0): The position is predicted to be rather fast-evolving.
GERP++	GERP++ is a conservation score quantified in terms of "rejected substitutions" per nucleotide, defined as number of substitutions expected under neutrality minus number of substitutions observed.	Positive score: The position shows a substitution deficit (it is conserved). Negative score: The position shows a substitution surplus (it is fast-evolving).
CADD	CADD integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations. The scaled (PHRED-like) C-scores range from 1 to 99.	A score ≥10 indicates that this is predicted to be one of the 10% most deleterious substitutions that you can do to the human genome, a score ≥20 indicates the 1% most deleterious and so on.

Table 3: **Description and interpretation of nucleotide-based scoring methods**. Described are three conservation scores (phyloP, phastCons, and GERP++) as well as one annotation-based, simulation-derived deleteriousness score (CADD).

MIRBASE – miRBase is a database developed by the WTSI that provides miRNA sequence data and annotation [205-207]. We downloaded the full set of miRNA annotation of release

18, including miRNA genes, miRNA hairpins and extracted all canonical seed sequences for each of the miRNAs according to [208].

3.4.2 Population-based haplotype panels for genotype imputation

We used the haplotype imputation panels of the HapMap Phase 2 release 22 containing 270 individuals and the 1000 genomes phase 1 version 3 release containing 1,092 individuals for genotype imputation [74, 77].

3.4.3 Regulatory element annotations

PROMOTERS & DISTAL ENHANCERS/REPRESSORS (DNaseI SCREEN) – In essence, THURMAN et al. [209] used DNaseI hypersensitive sites (DHSs) and mapped them to transcription start sites (TSSs) of human transcripts. Accessible DHSs in proximity to the TSSs are classified as promoters. The accessibility patterns of more distal DHSs have been correlated with the accessibility patterns of promoters across the analyzed cell types and are thus linked to the genes thought to be regulated by DHSs proximal to a TSS. After data processing, we obtained 412,798 distal elements (enhancers) and 23,749 promoters.

EXPRESSED PROMOTERS & ENHANCERS/REPRESSORS (FANTOM5) – Two papers of the FANTOM5 consortium [79, 80] describe the properties, location and transcript associations of expressed regulatory elements (promoters and enhancers). These datasets are provided at http://fantom.gsc.riken.jp/data/ and http://enhancer.binf.ku.dk/, respectively. After data processing, we included 435,881 expressed promoters and 43,002 expressed enhancers and their links to human transcripts in *SNiPA*.

STARBASE V2.0: MIRNA TARGET SITES – miRNA target sites located in RBP binding sites were obtained at the starBase v2.0 database (http://starbase.sysu.edu.cn/, released 09/2013, accessed 16/01/2014) [210]. We included target predictions from five prediction tools (provided by starBase) at positions that are located in experimentally identified regions bound by RBPs ($n = 606{,}408$).

3.4.4 eQTL associations

THE GTEX CONSORTIUM – The GTEx consortium collected 1,641 samples from 43 tissues of 175 donors and investigated gene-based *cis*-associations of RNA-sequencing-determined expression traits with about 6.8 million common variants [211]. As a prerequisite, a sample count of greater than 60 was chosen as lower bound for calculation of associations. In release 6, eQTL data was available for 44 tissues: adrenal gland, anterior cingulate cortex, aorta,

atrial appendage, blood, breast, caudate basal ganglia, cerebellar hemisphere, cerebellum, coronary artery, cortex, EBV lymphocytes, esophagus mucosa, frontal cortex, gastroesophageal junction, hippocampus, hypothalamus, left ventricle, liver, lung, muscularis mucosae, nucleus accumbens, ovary, pancreas, pituitary, prostate, putamen, sigmoid colon, skeletal muscle, spleen, stomach, subcutaneous adipocytes, sun exposed skin, terminal ileum, testis, thyroid, tibial artery, tibial nerve, transformed fibroblasts, transverse colon, unexposed skin, uterus, vagina, and visceral adipocytes. The release comprises 19,103,582 significant variant/gene expression *cis*-associations.

ZELLER ET AL. – MONOCYTES – ZELLER et al. investigated *cis*- and *trans*- associations of expression traits with >675,000 SNPs (Affymetrix SNP Array 6.0) in human monocytes from 1,490 unrelated individuals using the Illumina Human HT-12 v3 BeadChip. A SQLite database dump containing the association results is provided by the authors at http://genecanvas.ecgene.net/uploads/ForReview/ghs_probe_express030510.zip. This data-base comprises imputed association data on >2 Mio. SNPs. Following the protocol in [212] associations were filtered for genome-wide significance ($P_{value} > 5.78 \cdot 10^{-12}$). This filtered set was intersected with Kruskall-Wallis test (KWT) results and filtered to feature a KWT $P < 10^{-10}$ as described by ZELLER et al. [212]. SNPs were then split into *cis*-/*trans*-associations via distance to their associated expression target (up to 1MB apart: *cis*, else: *trans*).

MULTIPLE TISSUE HUMAN EXPRESSION RESOURCE (MUTHER) CONSORTIUM – LCL, ADIPOSE AND SKIN TISSUE – The MuTHER consortium collected samples from 856 female twins of the TwinsUK resource in three tissues (lymphoblastoid cell lines or LCL, adipose tissue, skin tissue) [213]. *cis*-eQTL associations comprising >2 Mio. SNPs were calculated using the Illumina Human HT-12 v3 BeadChip. Results files were retrieved from http://www.muther.ac.uk/Data.html and subjected to P_{value} filters as described in [213] ($P_{lcl} < 7.8 \cdot 10^{-5}$, $P_{adipose} < 5 \cdot 10^{-5}$, $P_{skin} < 3.8 \cdot 10^{-5}$) corresponding to a per-tissue false discovery rate (FDR) of 1%.

WESTRA ET AL. – PERIPHERAL BLOOD – WESTRA et al. performed a meta-analysis of eQTL associations in peripheral blood samples from 5,311 individuals [214]. Genotype data was imputed to HapMap2 CEU genotypes (>2 Mio. SNPs), expression data from different Illumina platforms (Human HT-12 v3, HT-12 v4, and H8 v2 BeadChips) were harmonized by mapping probe sequences to Human HT-12 v3 identifiers. Association data was obtained at http://genenetwork.nl/bloodeqtlbrowser/. Probes specified by Illumina array address IDs were

mapped to Illumina probe IDs using the developer manifest file (http://www.illumina.com). *Cis*- and *trans*-associations were filtered to have $P < 1.31 \cdot 10^{-4}$ and $P < 5.12 \cdot 10^{-7}$, respectively, corresponding to an FDR of 5%. In this study, eQTLs located less than 250 KB away from the probe midpoint are defined as *cis* while eQTLs more than 5 MB apart from the probe are defined as *trans* [214].

FAIRFAX ET AL. – B-CELLS AND MONOCYTES – FAIRFAX et al. investigated genotype associations with expression data from B-cells and monocytes from 288 individuals. For $> 600{,}000$ SNPs *cis*- (≤ 2.5 MB away from the probe) and *trans*-associations were determined at permutation ($n = 1{,}000$) $P < 1 \cdot 10^{-3}$ and Bonferroni-corrected $P < 1 \cdot 10^{-11}$, respectively. All significant associations from the online supplement [215] were mapped to Illumina Human HT-12 v4 probes using the genomic coordinates provided in the supplemental files to obtain an up-to-date mapping to the corresponding genes. For this, hg18/NCBI36 coordinates had to be converted to hg19/GRCh37 coordinates using the UCSC liftOver tool [216]. Probe mapping data was retrieved from the Ensembl public SQL database [149].

SEEQTL DATABASE – LCL AND BRAIN – The seeQTL database [217] contains several eQTL association datasets. Most of these are based on samples from individuals contained in the HapMap populations. On the data website of the seeQTL browser (http://www.bios.unc.edu/research/genomic_software/seeQTL/data_source), XIA et al. provide a meta-analysis association set on all HapMap-based studies which were included in *SNiPA*. In addition, association data from an eQTL study on human brain samples (MYERS et al. [218]) in the same file format is available and was also included.

DIXON ET AL. – LCL – DIXON et al. investigated genotype associations with expression data (using Affymetrix HG-U133 Plus 2.0 chip) from LCLs of 400 individuals [219]. The threshold for genome-wide significance was set to be a LOD score > 6.076 (equivalent to an FDR of 5%). Significant associations were extracted from the online supplement [219]. Associations with probes mapping to multiple locations in the genomes where removed ($n = 3{,}309$). Associations were defined as *trans* if SNPs are located more than 1 MB apart from the probe center, and *cis* else.

INNOCENTI ET AL. – HEPATOCYTES – INNOCENTI et al. investigated genotype associations with expression data (using Agilent 4x44K arrays) from liver tissue of 266 individuals [220]. The threshold for genome-wide significance was described to be a Bayes factor of > 5. We downloaded significant *cis*-associations from the online supplement [220].

SNiPA reports the P_{values} provided with the associations that, thus, may not always seem to be significant on a genome-wide level.

3.4.5 Variant-phenotype associations and annotations

DRUGBANK 4.0 – SNP–DRUG INTERACTIONS – DrugBank [221] is a bioinformatics resource collecting data on drug interactions. We downloaded the lists on "SNP Mediated Adverse Drug Reactions" as well as "SNP Mediated Pharmacological Effects" from the GenoBrowse interface.

GWAS CATALOG – The NHGRI-EBI GWAS catalog (original title: a catalog of published genome-wide association studies) [100, 222] is a text-mining resource formerly provided by the National Human Genome Research Institute (NHGRI) that collects the results of published GWAS. Before inclusion into the catalog, results are filtered and revised manually to ascertain several quality metrics. Although several other resources such as the HuGE Navigator [223] have been published, this is still the primary source for GWAS results. We downloaded the Catalog in tab-delimited format and retrieved trait annotations, association P_{values} as well as the source publications. Since the catalog's move to the EBI, we retrieve the data from Ensembl.

VARIANT-TRAIT ASSOCIATIONS FROM ENSEMBL – Ensembl includes variant-trait annotations and associations from several important resources. Data from OMIM [106], HGMD [224], UniProt [225], dbGaP [226], and ClinVar [227] were fetched from the public MySQL database. Association numbers are given in Table 4.

ASTHMA ASSOCIATIONS – GABRIEL CONSORTIUM – The GABRIEL (A Multidisciplinary Study to Identify the Genetic and Environmental Causes of Asthma in the European Community) Consortium performed an international genome-wide association meta-analysis of asthma and IgE levels in 10,365 cases and 16,110 controls recruited from 23 studies [161]. We downloaded the association results for 567,589 markers for each single study and itemized by age (adult/child) from www.cng.fr/gabriel.

Source	N (unique)
HGMD	53,420 (48,305)
dbGaP	40,254 (28,767)
ClinVar	156,160 (139,160)
OMIM variation	19,878 (18,442)
UniProt	3,484 (3,219)
GWAS Catalog	19,950 (18,769)
DrugBank 4.0	179 (169)

Table 4: Disease associations and annotations of genetic variants contained in *SNiPA* v3.

3.4.6 Gene-phenotype associations and annotations

ORPHANET – DISORDERS AND ASSOCIATED DISEASE GENES – OrphaNet [107] is a resource that collects information on rare diseases, associated disease genes, and orphan drugs. At its data repository (http://www.orphadata.org), it also offers mappings of diseases and disease genes to many other resources, including Ensembl. We downloaded the "Disorders with their associated genes"-XML file and used a PERL XML-library (XML::LiBXML) to extract the relevant data. Gene-trait association counts are given in Table 5.

OMIM – DISORDERS AND ASSOCIATED DISEASE GENES – The Online Mendelian Inheritance in Man database [106] collects data on Mendelian disorders and the associated disease genes as well as the known disease causing genetic variants. More recently, also strong associations from GWAS are collected in OMIM in a gene-centered manner. Entries in OMIM are manually curated and are thus very valuable. We downloaded links from genes to OMIM disease entries from the Ensembl database [149]. Gene-trait association counts are given in Table 5.

Source	N (unique)
DECIPHER	1,829 (1,829)
OMIM gene	4,886 (4,882)
OrphaNet	5,684 (5,684)

Table 5: Disease associations and annotations of human genes contained in *SNiPA* v3.

DECIPHER (DATABASE OF GENOMIC VARIATION AND PHENOTYPE IN HUMANS USING ENSEMBL RESOURCES) – DECIPHER [196] is an integrative resource of gene-disorder associations obtained by in-depth study of patients' genomes that is accessible through Ensembl. We downloaded DECIPHER gene-disease annotations from the Ensembl database [149]. Gene-trait association counts are given in Table 5.

GENETIC ASSOCIATION DATABASE – The genetic association database (GAD) was one of the first repositories for the standardized collection of genetic association results [228].

MIRNA-PHENOTYPE ASSOCIATIONS – Altered expression patterns of miRNAs are associated with several disease phenotypes. Two of the major resources for such miRNA-phenotype associations are the PhenomiR [229] and the miR2Disease [230] databases. We downloaded the full datasets comprising several hundred miRNAs associated with more than hundred human disorders.

3.4.7 Webservers, variant databases, and ontologies

HAPMAP – The HapMap project catalogues common sequence variants for all major human populations. We used variant and LD data from HapMap phase 2 [74] and phase 3 [75] panels.

DBSNP – The dbSNP database at NCBI is the primary reference resource for small sequence variants [178]. Consequently, its variant naming nomenclature (RefSNP- or rs-IDs) is ubiquitously used.

1000 GENOMES PROJECT – The international 1000 genomes project is currently the primary resource for population-based WGS haplotype data. While in phase 1 of the project, the consortium sequenced the promised 1000 human genomes, as of phase 3 more than 2,500 individual genomes are contained in the repository [77, 231].

SNAP – The SNAP (SNP Annotation and Proxy Search) resource is a user-friendly webserver for retrieval of LD data, basic variant annotations, variant/gene associations, as well as plotting of LD and association data [232].

DAVID – DAVID (Database for Annotation, Visualization and Integrated Discovery) is a bioinformatics resource for the functional annotation of gene sets [233, 234]. It provides several analysis modules, including gene function enrichment analysis as well as function-based gene clustering approaches.

GSEA – GSEA (Gene set enrichment analysis) is a web portal for functional as well as trait-association enrichment analysis of gene sets mainly based on altered gene expression analysis experiments [235]. It also provides enrichment analysis of gene ontology (GO) terms [236].

MESH – MeSH (Medical Subject Headings) is an ontology of organisms, traits, diseases, and other entities [237]. It uses a standardized vocabulary and manually curated categories. We used the category C (diseases) to map disease terms on the MeSH ontology tree.

PROTEIN-PROTEIN INTERACTION DATABASES – Protein-protein interaction (PPI) and pathway data were retrieved from IntAct [238], CORUM [239], and the Kyoto Encyclopedia of Genes and Genomes, KEGG [240].

CHEMICAL ANNOTATION AND PATHWAY DATABASES – Chemical pathway data was obtained at the Edinburgh Human Metabolic Network [241], KEGG [240], Recon 2 [242], and Reactome [243]. Additional metabolite and enzyme annotations were retrieved from BRENDA [244] (enzyme-metabolite relationships), the Human Metabolome Database (HMDB) [245], the Chemical Abstracts Service (CAS), ChEMBL [246] (drug/compound-gene

associations and approved drugs), and Citeline Pharmaprojects Pipeline (accessed on July 1, 2013, http://www.citeline.com/products/pharmaprojects/).

3.5 Software and tools

GENEVAR – Genevar (Gene Expression Variation) is a java applet connected to eQTL databases developed at the WTSI [247]. Before MuTHER consortium project was finished and the data was available for download, Genevar was the only interface to retrieve eQTL associations from the project. Furthermore, Genevar enabled access to eQTL data obtained by association studies in HapMap samples used in section 6.2.

CPMA – CPMA (cross-phenotype meta-analysis statistic) is an algorithm developed by the Cotsapas lab that allows for the inspection of an enrichment of significant association p-values for single variants across two or more phenotypes [248]. I used CPMA for validation of our network analysis approach described in section 6.1.

LIFTOVER – liftOver is the UCSC utility to map genomic coordinates between genome assemblies [216]. It was incorporated in the gene and regulatory build process used in chapter 5.

VARIANT EFFECT PREDICTOR – The Ensembl variant effect predictor is a tool to predict variant consequences on genomic entities [140]. It provides the baseline variant annotations used in the genomic resource described in chapter 5.

NNSPLICE – NNSplice is an algorithm from the Berkeley Drosophila Genome project that can be used for predicting variant-based changes to existing or creation of new splice sites [249]. We used NNSplice to estimate the effect of allele-specific splice variants on miRNA targeting as described in section 6.2.

VIENNA RNA PACKAGE – The Vienna RNA package is a toolbox that enables 2D-folding prediction of RNA sequences [250]. We used this software to calculate allele-induced folding changes to human mRNAs in order to estimate allele-specificity of RBP annealing potential described in section 6.2.

GENOMEGRAPHS – The GenomeGraphs R package is a toolbox for visualization of genomic data [251]. I used this software to provide the Association Maps module described in chapter 5.

NETWORK VISUALIZATION AND ANALYSIS TOOLS – For network visualization, we used the yWorks yEd graph editor (yFiles software, Tubingen, Germany) for displaying GRAPHML files. For the application of network analysis measures, we used the Cytoscape framework [252]. Both platforms were used in the analysis described in section 6.1.

PLINK – PLINK is a whole genome association analysis toolset that covers a multitude of analysis steps required for GWAS [179, 185]. It was utilized in several modules in the GWAS analysis pipeline described previously.

SHAPEIT2 – SHAPEIT2 is a linearly scaling method for fast but accurate genotype phasing [183]. I used it for haplotype phasing in the study on the genetics of SIDS.

IMPUTE2 – IMPUTE2 is a software toolbox for genotype imputation [97]. It is compatible with phased haplotypes as produced by SHAPEIT2.

QUANTO – Quanto is a tool for performing power calculations for genetic studies [253]. I used it in section 4.1 to estimate statistical power in dependency of the MAF.

VCFTOOLS – VCFtools is a software package for handling and analyzing files in variant call format (VCF) [254]. The VCF format has been developed by the 1000 genomes consortium and is now one of the standard formats to store variant data. We used it for LD calculations in chapter 5.

TABIX – Tabix is a software for positional indexing of block-compressed flat files containing genomic elements [255]. It is optimized for indexing genomic coordinates and is very fast even with extremely large files. As it also allows for remote file access, we used it as programmatic interface to the resource described in chapter 5.

JAVASCRIPT LIBRARIES – For development of the interactive webserver described in chapter 5, we relied on the following JavaScript libraries: jQuery and jQueryUI (The jQuery Foundation, 2014. www.jquery.org), Highcharts (Highcharts JS: Interactive JavaScript charts for your web projects. Highsoft AS, Vik i Sogn, Norway. www.highcharts.com), DataTables (DataTables: table plug-in for jQuery. SpryMedia. www.datatables.net), jQuery Chained (jquery_chained: chained selects for jQuery and Zepto. Mika Tuupola.

www.appelsiini.net/projects/chained), and Modernizr (Modernizr: the feature detection library for HTML5/CSS3. www.modernizr.com).

3.6 Mathematical and statistical concepts

3.6.1 Network analysis

The network concepts (density, centralization, and heterogeneity) which we used to compare network properties are defined as given in [256]:

1. $\text{Density} = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} = \frac{mean(k)}{n-1}$

where $a_{ij} = 1$ if nodes i and j are connected and 0 otherwise. $mean(k)$ denotes the mean connectivity which for a node i is defined as $k_i = \sum_{j \neq i} a_{ij}$.

2. $\text{Centralization} = \frac{n}{n-2}\left(\frac{\max(k)}{n-1} - Density\right)$.

3. $\text{Heterogeneity} = \frac{\sqrt{variance(k)}}{mean(k)}$.

To automatically distinguish the two node sets in bipartite networks, we used directed edges. Direction is always from one type (source) to the other (target). The distinct node degree distributions thus are identical to the indegree distribution and the outdegree distribution. The topological coefficient as a measure of modularity [257] T_i of a node i is defined as:

$$T_i = \begin{cases} 0, \text{ if } N_i < 2 \\ \text{avg}\left(\frac{S(i,j)}{N_i}\right), \text{ else} \end{cases}$$

where N_i is the number of neighbors of i and $S(i,j)$ is the number of shared neighbors of nodes i and j (undefined if i and j do not share a neighbor) plus one if j is a neighbor of i.

Power-law functions of the form $y = e^a x^b$ were fitted by least squares where coefficients are defined as $b = \frac{n\sum_{i=1}^n (\ln x_i \ln y_i) - \sum_{i=1}^n (\ln x_i) \sum_{i=1}^n (\ln y_i)}{n\sum_i^n (\ln x_i)^2 - \left(\sum_{i=1}^n \ln x_i\right)^2}$ and $a = \frac{\sum_{i=1}^n (\ln y_i) - b\sum_{i=1}^n (\ln x_i)}{n}$. As goodness-of-fit measure, we give the coefficient of determination

$$R^2 = \left(\frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n\sum_{i=1}^n {x_i}^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n\sum_{i=1}^n {y_i}^2 - \left(\sum_{i=1}^n y_i\right)^2}} \right)^2$$

for the linear transformation of the power-law functions, i.e. $\ln y = a + b \ln x$.

3.6.2 Survival analysis

For survival analysis with right-censored data, I use Cox proportional hazards using regression terms of the form $\lambda(time, status) = \lambda_0 \cdot e^{\beta_1 \cdot genotype + \beta_2 \cdot sex + \beta_3 \cdot MDS_1 + \cdots + \beta_{x+2} \cdot MDS_x}$. Here, $time$ is the vector of follow-up dates (or times at event), $status$ contains codes for event (e.g. death) occurred, sex is the vector of the gender of the individuals, and MDS_i is the i^{th} MDS dimension. The genotype vector is recoded beforehand to obtain effect estimates for the specified genetic model (Table 6).

Significance of association of the genotype with outcome under the selected genetic model is calculated using LRTs using the log-likelihood of the regression model including the genotype (LL_α) and that of the model excluding the genotype vector(s) (LL_0). The test statistic is then determined as $D = 2 \cdot (LL_\alpha - LL_0)$ which follows a χ^2-distribution with one degree of freedom for additive, dominant, recessive, and over-dominance models and two degrees of freedom for the genotypic model.

	Additive	Dominant	Recessive	Over-dominant	Genotypic
AA	0	0	0	0	0 0
Aa	1	1	0	1	1 0
aa	2	1	1	0	0 1

Table 6: **Genotype coding for the different genetic models**. Columns specify the genetic model, genotype conformations and their coding are denoted in the rows. As for the genotypic model, the genotype vector is splitted and recoded as two vectors (and thus, two regression variables), the test for significance has two degrees of freedom.

3.6.3 GWAS meta-analysis

Combined Z-scores for one variable (e.g. a variant) are computed as

$$Z_{meta} = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}, \text{ with } w_i = \sqrt{N_i} \text{ and } Z_i = \Phi^{-1}\left(\frac{P_i}{2}\right) \cdot \text{sign}(\beta_i)$$

where N_i is the sample size of study i, P_i is the two-tailed P_{value} of association for this variable, and β_i is the estimated effect (the estimate retrieved by linear regression or the logarithm of the odds ratio yielded by logistic regression). The two-tailed P_{value} for the meta-statistic is then calculated as $P_{meta} = 2 \cdot \Phi(-|Z_{meta}|)$ or, in case of continuous traits where Z is in fact the t-statistic and where only a limited number of samples is available, as $P_{meta} = 2 \cdot f_t(-|Z_{meta}|, \text{df} = n)$ where n is the number of samples minus the number of regressors and f_t is the distribution function of the t-distribution.

The DerSimonian and Laird random effects model is calculated using the standard error of the effects as weights and therefore needs more input. First, the between-study variance τ^2 needs to be calculated. For this, the Cochran's Q statistic [258] is calculated as

$$Q = \sum_{i=1}^{k} w_i \beta_i^2 - \frac{\left(\sum_{i=1}^{k} w_i \beta_i\right)^2}{\sum_{i=1}^{k} w_i} \text{, with } w_i = \frac{1}{SE(\beta_i)^2}$$

where k is the number of studies, the degrees of freedom are obtained as $\mathrm{df} = k - 1$, and a scaling factor C is introduced as

$$C = \sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}.$$

The between-study variance is then determined as

$$\tau^2 = \begin{cases} \dfrac{Q - \mathrm{df}}{C} & \text{if } Q > \mathrm{df} \\ 0 & \text{if } Q \leq \mathrm{df} \end{cases}$$

and included in w_i's as $w_i^* = \frac{1}{SE(\beta_i)^2 + \tau^2}$. Meta-statistics are then derived as

$$\beta_{meta} = \frac{\sum_{i=1}^{k} w_i^* \beta_i}{\sum_{i=1}^{k} w_i^*}, SE(\beta_{meta}) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}}, \text{ and } Z_{meta} = \frac{\beta_{meta}}{SE(\beta_{meta})}.$$

For binary traits and studies with a sufficient number of samples available, the two-tailed p-value is again calculated as $P_{meta} = 2 \cdot \Phi(-|Z_{meta}|)$ and log-scaled 95% confidence intervals (CI) can be obtained by $CI_{95} = [\beta_{meta} \pm 1.96 \cdot SE(\beta_{meta})]$. For continuous traits where only a limited number of samples is available, the formulas are as $P_{meta} = 2 \cdot f_t(-|Z_{meta}|, \mathrm{df} = n)$ and $CI_{95} = [\beta_{meta} \pm Q_t(0.975, \mathrm{df} = n) \cdot SE(\beta_{meta})]$ where n is the number of samples minus the number of regressors and Q_t is the quantile function of the t-distribution. Meta-effects and CI can be transferred to the odds ratio scale using the exponential function. Further, the Q-statistic is used to calculate heterogeneity statistics such as the probability of heterogeneity P_{het} and the I^2 and H statistics [259]. The random effects model can be directly transferred into a fixed effects model by forcibly setting $\tau^2 = 0$ which assumes that studies are comparable. Heterogeneity statistics are calculated in any case and results that are potentially caused by significant between-study variance are marked, suggesting the use of the random effects model.

4 Genetic association studies

The deciphering of the human genome sequence and the following successes of genetic studies and, in particular, of GWAS have raised hope that sooner or later medicine will be able to treat a patient on a personalized and not on a symptom-based level. Through the broader application of NGS in the clinic and the technological advances in other –omics fields, the idea of personalized medicine seems to have become even closer at hand. However, if we look below the questionable breakthroughs propagated by the media and the pharmaceutics sector, the truth is that we are far away from reaching this ambitious goal [260]. The complexity of the human organism as well as of the molecular, genetic, and environmental disturbances that lead to human traits and diseases is still poorly understood. While it is true that we are able to measure the genetic (or metabolomic or transcriptomic or proteomic) differences between diseased and (nominally – as based on the classical classification system of non-personalized medicine) healthy individuals, the translation of these differences into clinical use is progressing only slowly.

The first section of this chapter is intended to introduce the methodologies for the detection of trait-associated genetic variants (SNVs as well as CNVs) exemplified using a genetic analysis of the fatal disease syndrome SIDS. Using the outcomes of this study, the challenges represented in the functional, biological, and medical interpretation of the identified genetic loci will be outlined. In the second and the third section, I will focus on the genetics of intermediate traits that are more closely linked to cellular functions such as gene expression levels and metabolite

concentrations. Here, I want to highlight the value of considering genetic influences on intermediate traits in linking genomic regions as directly as possible to biochemical readouts.

Many parts of this chapter are based either on previously published data or are part of manuscripts in preparation. The references are given in the respective sections.

4.1 Genetics of the sudden infant death syndrome (SIDS)

Sudden infant death syndrome or SIDS is defined as the "*SUDDEN DEATH OF AN INFANT UNDER ONE YEAR OF AGE WHICH REMAINS UNEXPLAINED AFTER A THOROUGH CASE INVESTIGATION, INCLUDING PERFORMANCE OF A COMPLETE AUTOPSY, EXAMINATION OF THE DEATH SCENE, AND REVIEW OF THE CLINICAL HISTORY*" [261], a diagnosis *per exclusionem* as cases of sudden unexpected death in infancy of infants less than one year of age where no definite cause of death can be demonstrated [262]. It logically follows that SIDS etiology, contrary to diagnoses *per definitionem*, is very vaguely understood and is only defined by symptoms that have been observed in autopsies of SIDS victims. Nevertheless and in spite of these unclear conceptualizations, SIDS is still a major cause of infant death in industrialized countries (6.8% of all infant deaths in Germany in 2013, source: German Federal Statistical Office). In an attempt to formulate a generalized characterization of SIDS etiology, the triple-risk model has been derived that, in its best known form as formulated by FILIANO and KINNEY, consists of "*(1) A VULNERABLE INFANT; (2) A CRITICAL DEVELOPMENTAL PERIOD IN HOMEOSTATIC CONTROL, AND (3) AN EXOGENOUS STRESSOR(S). AN INFANT WILL DIE OF SIDS ONLY IF HE/SHE POSSESSES ALL THREE FACTORS; THE INFANT'S VULNERABILITY LIES LATENT UNTIL HE/SHE ENTERS THE CRITICAL PERIOD AND IS SUBJECT TO AN EXOGENOUS STRESSOR*" [263]. The cause of the vulnerability of an infant to sudden unexpected death remains controversial. However, the available data shows that there are several potential causes that frequently include inherited predisposition [264, 265]. While the contribution of monogenic risk factors such as long-QT-syndrome and medium-chain acyl-CoA dehydrogenase deficiency (MIM: 201450) to SIDS is established and generally accepted, the importance of complex genetic predispositions is less clear. Hypotheses regarding the involved pathways include cardiac channelopathies, inflammatory processes, impaired serotonergic

signaling, bacterial and viral infections, dysfunctional immune response, decreased energy production, abnormalities of the brain or the central nervous system, asphyxia, inborn errors of metabolism, and chronic hypoxia [168, 261-270]. Many of these hypotheses have been tested in genetic screens and the literature lists 42 genes that are described to confer risk for SIDS (reviewed in [264, 270]). However, the associations could only rarely be replicated and some have even been shown to have no effect on SIDS risk [271]. I therefore conducted a first-stage GWAS of SIDS in Europeans (318 cases and 1,493 controls). Additionally, I investigated CNVs for potential implication in SIDS etiology. As CNV studies using genotyping arrays have a large potential for detection of false positive CNV calls, I used another 2,764 European control samples for CNV analysis.

Most of the content of this section is still unpublished but is part of a manuscript in preparation and another one that has just been accepted at the International Journal of Legal Medicine (Fard et al., *"Candidate gene variants of the immune system and Sudden Infant Death Syndrome"*, Int J Legal Med, 2016) [272].

4.1.1 Methods summary

PATIENTS & CONTROLS – The study population consisted of 365 SIDS cases with both autopsy information and biosamples available. 317 of those originated from GeSID, the other 48 cases were recruited and autopsied in the UK at the Sheffield Children's Hospital. Both GeSID and SCHC used a standardized autopsy protocol and only infants suffering from SUDI where no causes of death could be determined postmortem were classified as SIDS. As controls for the GWAS, I used 823 adult individuals from the KORA F4 study, as well as 678 adults from the PopGen study. For filtering of CNVs, I had access to control CNVs obtained in a subset of the HYPERGENES cohort containing 2,764 individuals.

GWAS QUALITY CONTROL METRICS – Due to the different genotyping platforms used for cases and controls, I performed step-wise merging of the genotype data. GeSID and SCHC data, both obtained with the Illumina HumanHap660W-Quad BeadChips, could directly be merged. Exclusion of monomorphic and discordant missing markers resulted in 561,490 remaining markers. Analogously, KORA and PopGen, both genotyped with Illumina HumanHap550-Quad+ BeadChips, could also be directly merged. The final merge of case and control genotypes using the same filtering steps yielded a set of 505,759 markers available in all four cohorts. Based on these variants, global quality control was applied: *i)* individuals were filtered for an individual call rate (ICR) $< 95\%$ ($n = 48$) and one sample per duplicate or pair of

related individuals ($\hat{\pi} \geq 12.5\%$) was removed using the ICR as filter criterion ($n = 9$); *ii)* subjects where the annotated sex differed from the genotypic sex (determined calculating the average heterozygosity of X-chromosomal markers) were excluded ($n = 6$); and *iii)* markers were filtered at a genotype call rate $< 95\%$, HWE $P_{value} > 1 \cdot 10^{-5}$ (controls only), and minor allele frequency $< 5\%$. I performed all these steps using the pipeline described in section 3.2.

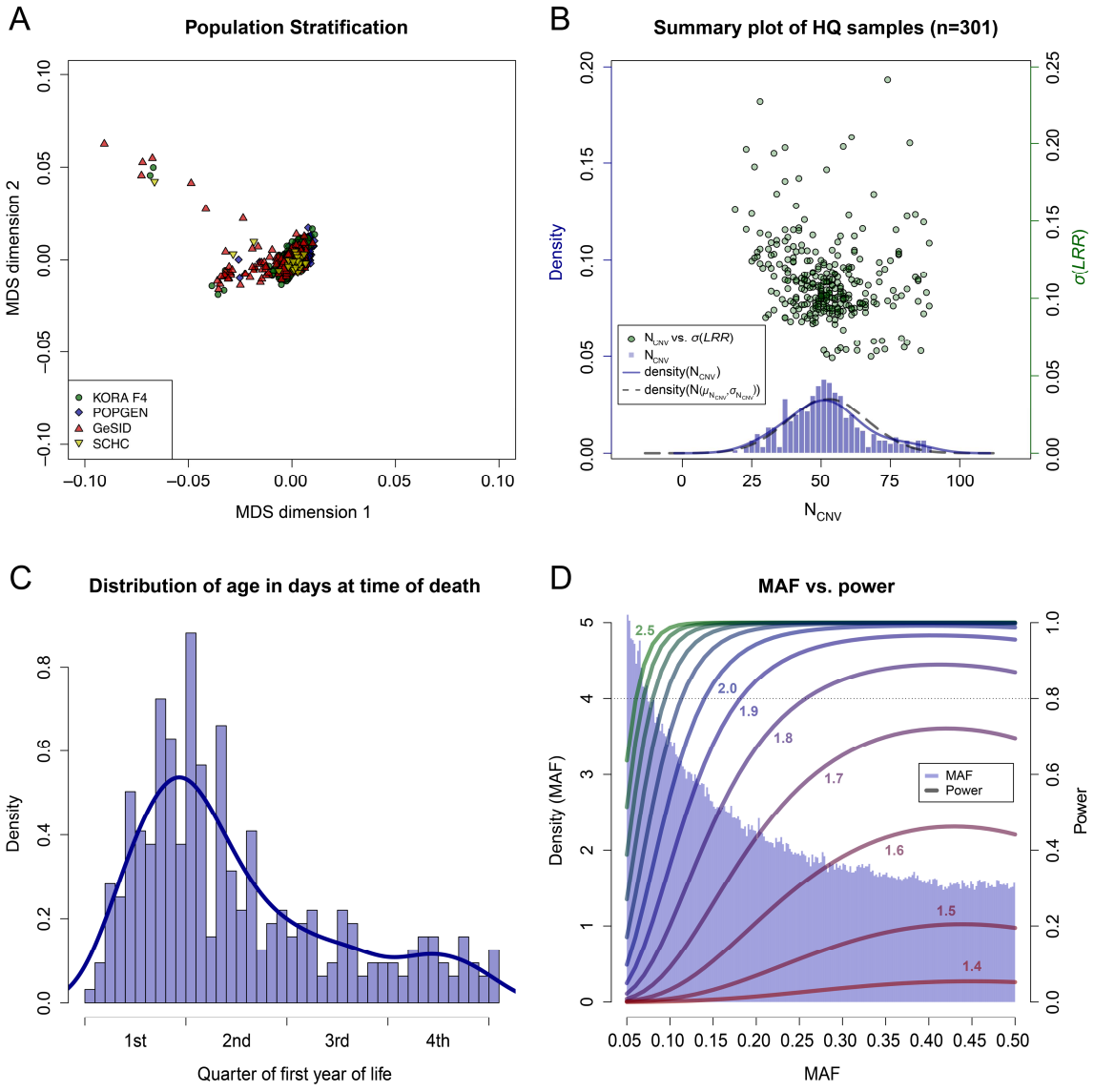


Figure 10: **Study characteristics**. **A**. **Population stratification** – Plot of the first two MDS dimensions. Although most variance is captured by the first two MDS dimensions, ten MDS vectors were included as covariates in the GWAS. **B**. **CNV QC summary** – Summary plot of intensity statistics for CNV calling after QC. Clustering of data points and normal distribution of CNV counts are identifiable. **C**. **Age distribution of SIDS cases** – It is known that SIDS risk is highest around the third month of life. This is also seen in our data. **D**. **MAF and power** – Density distribution of the MAF (bar plot) and statistical power of our study in dependency thereof (lines) for effect sizes between 1.4 and 2.5 (0.1 steps). The 80% power-threshold is denoted by the dotted line.

IMPUTATION AND POPULATION DATA – Genotype imputation for cases and controls was performed in a two-step process using SHAPEIT2 and IMPUTE2 as described in section 3.2. I used the haplotype imputation panel of the 1000 genomes project phase 1 version 3 release containing 1,092 individuals for genotype imputation [77].

CALCULATION OF ASSOCIATION STATISTICS – In addition to sex, I calculated the first ten multidimensional scaling dimensions for inclusion as covariates in all association statistics to account for population stratification (Figure 10A). For analyses excluding the SCHC set of cases (German-only analysis), I recalculated MDS vectors accordingly. I then used logistic regression analysis under the additive model. To include age in days at time of death, I additionally performed survival analysis using right-censored Cox-regression (proportional hazards).

	Cases			Controls		
	GeSID	**SCHC**	**total**	**KORA F4**	**PopGen**	**total**
N by gender (m/f)	283 (171/112)	35 (26/9)	318 (197/121)	822 (424/398)	671 (255/416)	1,493 (679/814)
male/female ratio	1.53	2.89	1.63	1.07	0.61	0.83

Table 7: **Sample statistics and covariates**. SIDS exhibits an approximate 2:1 male/female ratio. As in our control cohorts, more females than males are contained, I included sex as a covariate in all regression models.

CNV QUALITY CONTROL METRICS – In total, 605,701 probes had CRLMM genotype confidence scores (GCS) to pass QC. The filter for individuals violating the threshold for the signal-to-noise ratio removed 24 samples. The remaining set of SNR-values were normally distributed with $\mu_{SNR} = 40.53$ and $\sigma_{SNR} = 2.20$ (test for normality was performed using the two-sample Kolmogorov-Smirnov-Test). I then calculated for each remaining sample the log R ratio, the B allele frequency, and marker-based copy number estimates for input to PennCNV and VanillaICE. Markers with GCS< 0.9 were excluded before CNV calling for each individual. After CNV calling, I retained only CNVs that were called by both algorithms with at least 80% overlap and contained more than five markers. The resulting set of CNVs per person ($median = 53,\ min = 19,\ max = 297$) were used for determination of sample-based quality criteria (see section 3.3). 37 samples were filtered accordingly ($N_{CNV} > 90$ and/or $\sigma(LRR) > 0.25$) to retrieve the set of high-quality data (Figure 10B). Before their exclusion, duplicated samples ($n = 3$) were used as internal validation of CNV calls (concordance was $> 99\%$). Common CNVs were then marked using the CNV calls from Children's Hospital of Philadelphia CNV database [193] as well as the Database of Genomic Variants [194]. As calls of duplications based on SNP array data can show poor quality [273], I limited the analysis to CNVs with copy numbers less than two. Larger deletions >50Kb were compared to CNV calls

in the HYPERGENES controls and, where applicable, smaller common CNVs were tested for enrichment using Fisher's test. For this analysis, I coded CNV alleles as 0 = no deletion, 1 = hemizygous deletion, 2 = homozygous deletion.

4.1.2 No strong complex genetic background in SIDS

After extensive quality control, I performed genome-wide screening for associations between 4,778,167 genotyped and imputed common variants and SIDS using 318 cases and 1,493 controls. Depending on the allele frequency, our study was well powered to detect associations of moderate effect sizes above an odds ratio of 2.5 (I generated power calculations with Quanto; Figure 10D). At the threshold for genome-wide significance of $P < 5.0 \cdot 10^{-8}$, I was not able to detect any significant associations using logistic regression under the log-additive model. As the age at which infants succumb to SIDS is not randomly distributed across the first year of life but shows an accumulation between months two and four (Figure 10C) [274], I additionally performed right-censored (for controls) Cox proportional hazards estimation to include age in days at time of death in the regression model. This analysis yielded slightly different results, however, significant associations could not be identified either.

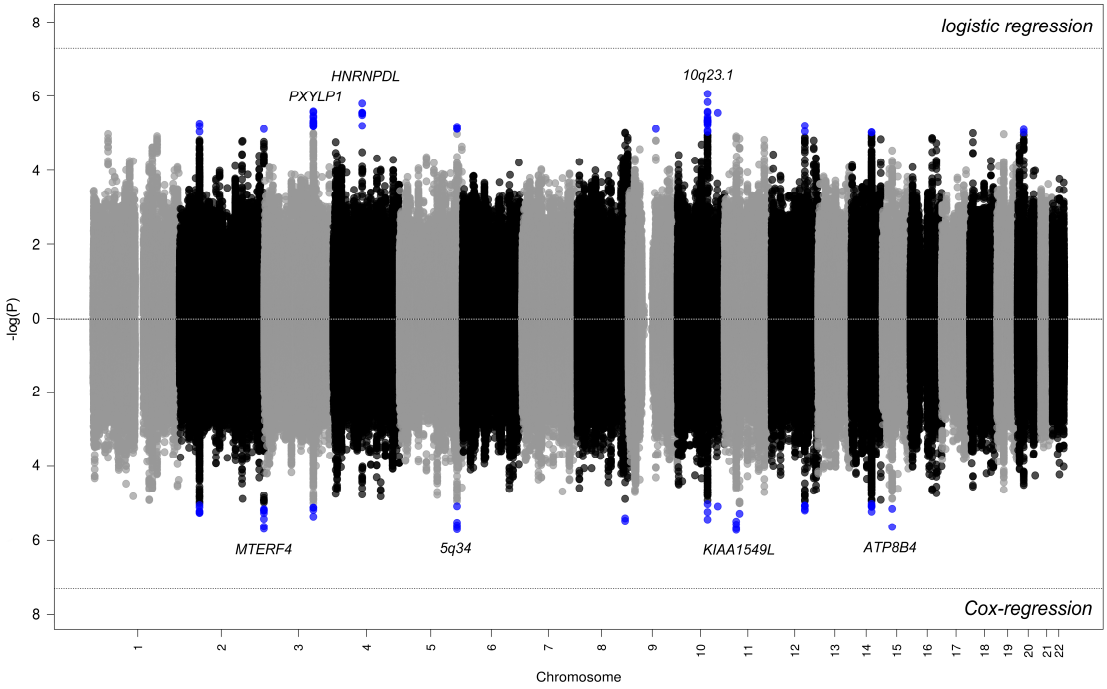


Figure 11: **Manhattan plot of association study results**. Results are shown for association analysis using logistic regression (top) and Cox-regression (bottom). For logistic regression, the top three loci are marked, while for Cox-regression loci are highlighted that show lower $\boldsymbol{P_{values}}$ as compared to the logistic regression results. Markers in blue have $\boldsymbol{P_{values} < 1.0 \cdot 10^{-5}}$. Overall, no variant reached genome-wide significance (threshold indicated by the dotted lines).

The most significant signal obtained by logistic regression analysis was for *rs11201232* in chromosome *10q23.1* ($OR = 1.63,\ P = 8.9 \cdot 10^{-7}$), followed by rs13952 ($OR = 1.61,\ P = 1.6 \cdot 10^{-6}$) in the *HNRNPDL* (*heterogeneous nuclear ribonucleoprotein D-like*) gene and *rs9880393* ($OR = 1.59,\ P = 2.7 \cdot 10^{-6}$) located in the promoter-containing region upstream of *2-phosphoxylose phosphatase 1* (*PXYLP1*). Cox-regression showed attenuated association statistics for these three loci, however, it revealed four loci with smaller P_{values} than in the logistic regression, namely *rs7929102* (hazard ratio (HR) $= 1.55,\ P = 1.9 \cdot 10^{-6}$) in the *KIAA1549L* (*KIAA1549-like*) gene, *rs12153272* ($HR = 1.69,\ P = 2.0 \cdot 10^{-6}$) in *5q34*, *rs6735450* ($HR = 2.01,\ P = 2.1 \cdot 10^{-6}$) located proximal to *MTERF4* (*mitochondrial transcription termination factor 4*), and *rs74012838* ($HR = 2.38,\ P = 2.3 \cdot 10^{-6}$) in the *ATP8B4* (*ATPase, class I, type 8B, member 4*) gene (Figure 11).

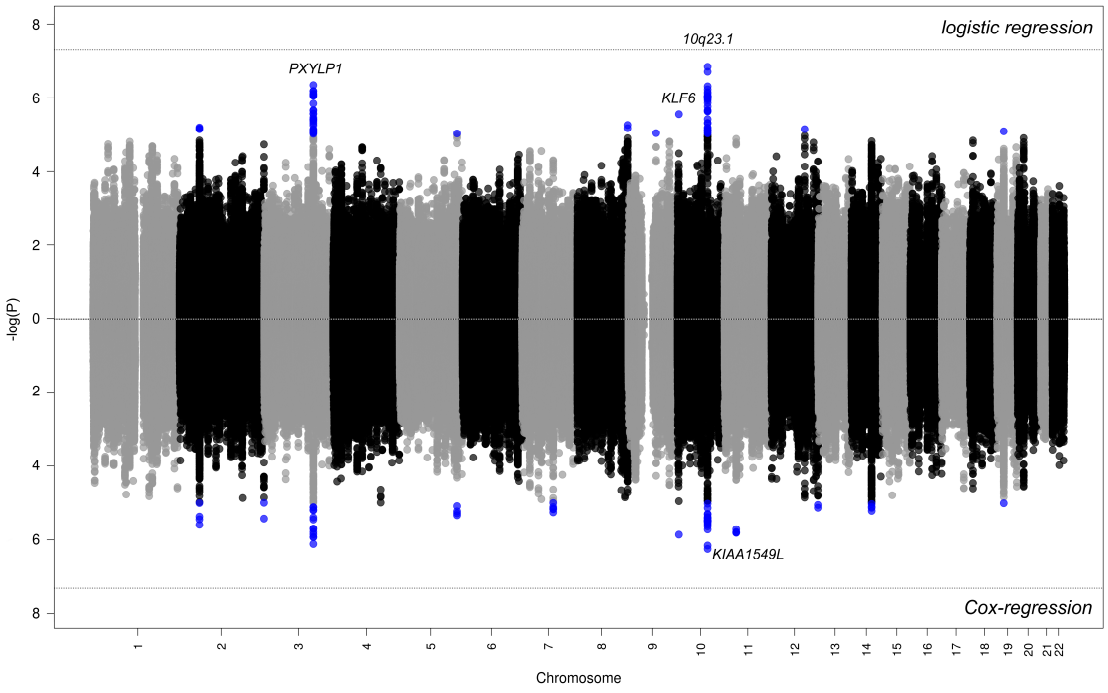


Figure 12: **Manhattan plot of the German-only association study results**. Results are shown for association analysis using logistic regression (top) and Cox-regression (bottom). For logistic regression, the top three loci are marked. The *KIAA1549L* locus was again only identified using Cox-regression. Markers in blue have $\boldsymbol{P_{values} < 1.0 \cdot 10^{-5}}$. As for the complete GWAS, no variant reached genome-wide significance (threshold indicated by the dotted lines).

To exclude the possibility that, due to potentially different country-specific case ascertainment criteria, associations get masked by non-SIDS cases, I performed a GWAS using only German cases (for an UK-specific analysis the case cohort was too small). Interestingly, despite the lower power (n_{cases}=283) the German-only GWAS yielded lower P_{values} for the top

associated SNPs at *10q23.1* ($OR = 1.73,\ P = 1.4 \cdot 10^{-7}$) and *PXYLP1* ($OR = 1.68,\ P = 4.6 \cdot 10^{-7}$). The third-best association signal was a locus that did not show up in the complete GWAS, *rs11252055* ($OR = 1.61,\ P = 2.8 \cdot 10^{-6}$) downstream of the *Kruppel-like factor 6 (KLF6)* gene (Figure 12). Cox-regression showed weaker attenuation of the signals at both *10q23.1* ($HR = 1.56,\ P = 5.7 \cdot 10^{-7}$) and *PXYLP1* ($HR = 1.57,\ P = 7.7 \cdot 10^{-7}$), as well as a stronger association at the *KIAA1549L* locus ($HR = 1.60,\ P = 1.5 \cdot 10^{-6}$).

4.1.3 Interpretation of suggestive significant GWAS loci

Before I attempt here to interpret the loci that might hold genome-wide significance in studies with larger sample sizes than ours, I want to point out that the central finding of this analysis is that SIDS cannot be attributed to strong complex genetic risk factors (i.e. common markers detectable by the GWAS approach at effect sizes our study is well powered to detect). However, in order to provide insights into the general approach that is used to annotated GWAS-identified genetic loci, I want to elucidate the loci emphasized in the Manhattan plots (Figures 11 and 12). First, it is noteworthy that the strongest association signal at chromosome *10q23.1* is located in a gene desert (as is the locus at chromosome *5q34*), which does not allow for sound hypotheses regarding potential molecular mechanisms that might affect the risk contribution to SIDS.

The *PXYLP1*-locus on the other hand may indeed be a promising target for further investigation. *PXYLP1* encodes a protein that functions as a phosphatase, dephosphorylating xylose residues in the linker region for glycosaminoglycan chains, enabling the polymerization of these molecules on the surface of proteoglycans [275]. This is a rate-limiting step in the post-translational modification and, thus, activation of this kind of proteins. Proteoglycans are a family of proteins with a large portfolio of cellular functions, including the control of the bioactivity of several inflammatory mediators as well as response to bacterial toxins and tissue repair [276]. Altered expression levels of *PXYLP1* due to risk variants in its promoter region may thus increase infant vulnerability to external stressors such as bacterial infections, which would be in line with the triple-risk model for SIDS.

The second best hit in the complete GWAS was for the variant located in *HNRNPDL*. This gene has been implicated in a neuromuscular disease that is inherited following the autosomal dominant pattern (limb-girdle muscular dystrophy (LGMD), type 1G; MIM: 609115) [277]. Equally to SIDS, sleep apnea and failed arousal from sleep are thought to be one of the predominantly contributing factors to mortality in the disease and seem to be independent from disease severity [278-280]. Because dominant forms of LGMDs are very rare

and usually show a late-onset of disease symptoms, a relationship between the association found in our study and the accumulation of hypoxia in SIDS victims is highly speculative. Nevertheless, as therapy may be as simple as bi-level positive airway pressure administered using a nasal mask during sleep hours [281], this hypothesis, if verified, could be implemented in clinical practice.

As to the loci showing stronger association statistics in the Cox-regression analysis, *MTERF4* seems to be the most promising target for further studies. The protein product of the gene directs the biogenesis of mitochondrial ribosomes and thus represents an important factor for translational activity. In mice, complete loss of *MTERF4* was shown to be embryonically lethal, while its targeted knockout in cardiac tissue significantly shortened life span [282]. Less severe effects on the activity of the gene via regulatory variants tagging the genetic signal outlined in my findings may still implicate affected energy homeostasis through reduced functioning of the mitochondrial translation machinery, a process that has been hypothesized to contribute to the susceptibility to SIDS. *KIAA1549L* and *ATP8B4* are both only poorly annotated regarding their molecular functions, although it should be mentioned that the genetic locus containing *KIAA1549L* (which is also linked via eQTL associations in blood and adipocytes) has been associated with heart rate traits (dbGaP analysis pha003053) and body height (dbGaP analyses pha003010 and pha003011). The *KLF6* locus, that only showed suggestive significance in the German-only GWAS, is also associated with heart rate traits (more specifically, with recovery of the heart rate after exercise) as well as with HDL-cholesterol levels (dbGaP analyses pha001678, pha000515, and pha000517). These findings support the general hypothesis of heart traits being linked to SIDS, but, based on the non-significant association statistics in our study, stay very hypothetical.

4.1.4 Genetic deletion syndromes may contribute to SIDS numbers

The analysis of rare CNVs detected large (>1Mb) hemizygous deletions in 10 of the 301 cases (3.3%) that passed quality control criteria (Table 8). Of those, at least three are pathogenic causing genetic disorders. The first is a 2.6Mb deletion at chromosome 22q11.21 comprising the DiGeorge/velocardiofacial syndrome (DGS/VCFS) interval (MIMs: 188400, 192430). In more than 70% of the cases, this deletion causes cardiovascular defects and immune deficiency leading to recurrent infections in an autosomal dominant pattern of inheritance, while almost one third of cases show no visible abnormalities (see [283], table 1; I detected the proximal A-D deletion). The second and third pathogenic CNVs are identical ~1.4Mb deletions on chromosome 7q11.23 in two of our cases, stretching across the region deleted in 95% of patients

with Williams-Beuren syndrome (WBS; MIM: 194050). WBS is an autosomal dominant multisystem disorder that includes cardiovascular, endocrine, and neurodevelopmental abnormalities and shows increased occurrence of sleep dysregulation and recurrent infections, symptoms that are also attributed to SIDS [284]. The prevalence of WBS has been estimated to 1 in 7,500 [285], whereas I find the typical microdeletion in 2 out of 301 cases.

coordinates	length	band	N_{SNP}	N_{genes}	phenotype candidate
chr3:2,856,134-4,168,500	1,312,366	3p26.3-.2	501	5	-
chr4:189,745,232-191,146,121	1,400,889	4q35.2	268	6	-
chr5:4,429,717-7,950,191	3,520,474	5p15.32-31	988	>10	Cri-du-chat syndrome (MIM: 123450)
chr5:10,585,639-11,923,554	1,337,915	5p15.2	472	3	Cri-du-chat syndrome (MIM: 123450)
chr5:158,185,817-163,691,759	5,505,942	5q33.3-q34	976	>20	GEFS (MIM: 611277)
chr7:72,360,917-73,777,987	1,417,070	7q11.23	79	>20	Williams-Beuren syndrome (MIM: 194050)
chr7:72,360,917-73,777,987	1,417,070	7q11.23	91	>20	Williams-Beuren syndrome (MIM: 194050)
chr10:45,938,621-51,406,960	5,468,339	10q11.21-23	688	>20	HHT (MIM: 615506)
chr18:2,246,904-4,459,918	2,213,014	18p11.32-31	570	>10	Holoprosencephaly (MIM: 142946)
chr22:17,175,037-19,796,478	2,621,441	22q11.21	572	>20	DGS/VCFS (MIMs: 188400, 192430)

Table 8: **The ten rare large hemizygous deletions found in SIDS cases**. Given are the 10 deletions with sizes greater than 1Mb and candidate phenotype where available. Genomic coordinates are given with respect to genome assembly NCBI36/hg18. N_{SNP}: number of markers within call; N_{genes}: number of genes contained in the deletion. GEFS: Generalized epilepsy with febrile seizures; HHT: Hereditary hemorrhagic telangiectasia; DGS/VCFS: DiGeorge/velocardiofacial syndrome.

Of the remaining seven large hemizygous deletions, another three regions are implicated in severe disease phenotypes. The largest (~5.5Mb) is located on chromosome 5q33.3-34 and contains more than 20 genes, including the known disease genes *IL12B*, *GABRA1*, *GABRG2*, and *HMMR*. Autosomal dominant disorders linked to these genes are generalized epilepsy with febrile seizures including familial febrile seizures (MIM: 611277), as well as susceptibility to other forms of epilepsy (MIMs: 611136, 607681). Febrile seizures can occur very early in life and in most cases progress without life-threatening course [286]. However, the case history describing the infant with this deletion as cyanotic with foam at mouth and nose when found by the mother and autopsy detecting gliosis in the thalamus (which can result from seizures [287]) may be indicative for recurrent febrile epileptic seizures leading to hypoxia, heart failure, and death. Another large (~5.47Mb) deletion was found to be located on chromosome 10q11.21-23 that contains more than 20 genes and includes six known disease genes. One of them, *GDF2*, is causative for hereditary hemorrhagic telangiectasia (MIM: 615506). The disease often manifests with arteriovenous malformations in lungs and brain that can lead to fatal hemorrhage and stroke [288]. Also, in a patient with Bohring-Opitz syndrome showing recurrent infections in the chest as well as generalized seizures, the same region was found to be deleted on the

maternally inherited chromosome [289]. The third likely pathogenic deletion (~2.2Mb) is located on chromosome 18p11.32-31 and comprises more than 10 genes including the known disease genes *SMCHD1*, *LPIN2*, and *TGIF1*. *TGIF1* causes autosomal dominant holoprosencephaly (MIM: 142946) that can lead to medical problems including abnormal heart and respiration rates and impaired organ growth even if brain structure seems normal [290]. The infant with this deletion showed edematous brain swelling and pericardial abnormalities that were attributed to the resuscitation attempts, but may be in line with active disease. However, this remains speculative as the brain of the infant was not investigated in more detail in the autopsy. The remaining four large deletions cannot be directly linked to a disease phenotype, however, due to their size it seems likely that they may have contributed to the infants' vulnerability to a life-threatening event causing sudden death.

To estimate the contribution of smaller CNVs to SIDS risk, I filtered CNV calls using additional criteria and excluded CNVs if they *i)* were called based on less than 10 markers, *ii)* had a CNV QC score less than 90%, *iii)* did not change the amino acid sequence of a known disease gene, *iv)* showed copy number changes not fitting the mode of inheritance of the disease linked to the affected protein (if available), and *v)* were found in equal proportions in the 2,764 control samples (Fisher's $P > 0.05$). This resulted in only two hemizygous deletions (260Kb and 81.5Kb in size) in two cases (one per case), both located in the *catenin (cadherin-associated protein), alpha 3 (CTNNA3)* gene that, when mutated, is suspected to be causally linked to autosomal dominant arrhythmogenic right ventricular dysplasia (ARVD; MIM: 615616). ARVD is a developmental disorder of the right ventricle of the heart and as such is a major cause of juvenile sudden death [291]. The significance of *CTNNA3* mutations in ARVD has yet to be conclusively resolved. However, both deletions detected in our study erase two complete exons which – provided that a causal relationship between mutated *CTNNA3* and ARVD can be proven – may explain the sudden death of these two respective cases.

4.1.5 Concluding remarks

The most important finding of this study is that there are no strong common genetic risk factors influencing the susceptibility to SIDS. This study is powered to detect associations with global effect sizes greater than 2.5 (Figure 10D) at the complete allele frequency spectrum addressed and thus is underpowered to detect associations with effect sizes typically seen in GWASs of complex traits. However, SIDS is a fatal syndrome which, if there is a common genetic predisposition, should show larger effect sizes than genetic influences on heterogeneous trait endpoints. And even if this assumption should be incorrect, there are many examples of

GWAS-identified loci with effect sizes larger than our detection limit, for complex traits as diverse as autoimmune diseases, Acne vulgaris, and coronary heart disease [222]. The fact that I fail to detect any significant associations permits only two settings. Either vulnerability to SIDS is solely caused by environmental circumstances/triggers. Or it is caused by rare *de novo* variants with moderate to large effect sizes that are unlikely to be detected using the GWAS approach. The identified rare deletions in twelve SIDS cases (almost 4%) support this hypothesis, which could also be the reason why I was not able to replicate any of the genetic loci previously implicated in SIDS pathogenesis. It may well be that I have identified even more copy number changes that are involved in SIDS predisposition. However, CNV analysis performed with genotyping arrays results in numerous calls and their pathogenicity is hard to estimate, especially if CNV status of the cases' parents is unknown (as in the presented study). The interpretation of smaller CNVs is a difficult task and should be handled with caution. Therefore, I used very strict measures to exclude spurious calls in the analysis.

To conclude, I was able to rule out a strong common genetic background of SIDS. While my results may indicate that specific mechanisms including heart function, energy homeostasis, as well as efficacy of the immune system may influence predisposition to SIDS, this can only be conclusively elucidated using whole-genome sequencing. And as genetic deletion syndromes often show severe developmental and skeletal abnormalities that must be detected in the autopsy, it is not to be expected that the numbers of SIDS cases due to such syndromes are significantly higher than in our case cohort. Nevertheless, as the death of a child is a traumatic event in the life of the parents, I suggest that cytogenetic screening for large deletions should be included in the standard autopsy protocol for infants suffering from sudden unexpected and unexplained death.

4.2 Genetic influences on human blood metabolites

As adumbrated in section 1.8, metabolic homeostasis is a crucial prerequisite for human health. The great success of GWAS has provided us with hundreds of associations between genetic loci and complex human traits. However, the functional interpretation of these associations and their mode of action by which they affect trait predisposition, development,

and progress on the molecular level is still very limited for the great majority of GWAS results. To address this issue, studies on genetic influences on intermediate phenotypes from –omics– levels closer (in terms of regulatory layers) linked to the genome are performed and correlations between genetic markers and epigenetic marks, transcript levels, protein levels, as well as metabolite concentrations gain more and more attention. In this section, I describe the key results of our mGWAS on human blood metabolite levels published in *Nature genetics* [116]. In addition to the report of 145 significant associations, we provide an extensive set of annotations on the biological relationship between the genetic loci and the linked metabolic phenotypes (metabotypes). Moreover, to support downstream analyses, I created a set of web resources that contain all information on the genetically influenced metabotypes (GIMs) collected in our study.

4.2.1 Methods summary

SAMPLES AND GENOTYPES – For the characterization of genetic influences on metabolite concentrations in blood samples, we included 7,824 individuals [116]. 1,768 of those originate from the KORA F4 study [169], the remaining 6,056 individuals were recruited by the UK Adult Twin Registry [170]. Further information on included individuals and covariates is listed in Table 9. After quality control and imputation, about 2.1 million SNPs were available for both KORA F4 and TwinsUK.

Study	N by gender (m/f)	Age [years] (mean±SD)	BMI [kg/m²] (mean±SD)
TwinsUK	6,056 (433/5,623)	53.4 ± 14.0	26.1 ± 4.9
KORA F4	1,768 (858/910)	60.8 ± 8.8	28.2 ± 4.8
total	**7,824 (1,291/6,533)**	**55.1 ± 12.9**	**26.6 ± 4.9**

Table 9: **Sample statistics and covariates**.

QUALITY CONTROL METRICS – SNPs were filtered based on HWE and call rate. Metabolomics raw data was normalized to the run-day median and proportional adjustment of each data point to account for variation due to instrument tuning differences. Metabolite concentrations were then log-transformed with base 10 and data points deviating more than four standard deviations from the mean were excluded. More details are given in [116].

ADDITIONAL DATASETS – The databases and resources used for the manual annotation of the identified loci are described in section 3.4.

DATA PROVISION – In addition to the extensive supplementary material provided at the publisher's website, I integrated all association and annotation data in two distinct web resources to enable convenient access to the results of this study.

First, an online supplement website (www.gwas.eu/si) that comprises all genome-wide significant genetic loci, the obtained metabolite-locus network, as well as all additional annotations including regional association plots for most significant metabotype. Entry point is a graphical genome atlas of the genetic loci (Figure 13).

Second, I developed a web resource for variant-based retrieval of variant-metabolite associations, the Metabolomics GWAS Server (www.gwas.eu). The database underlying the resource contains the full set of HapMap2-determined variants linked via LD data [74] as well as gene annotations from GENCODE v.14 [152] and metabolite associations from this and one additional publication [292]. Metabolites were linked to three external resources (The Human Metabolome Database, HMDB [245]; Kyoto Encyclopedia of Genes and Genomes, KEGG [240]; Chemical Abstracts Service, CAS) to provide further information on their properties.

4.2.2 Eighty-one newly discovered genetically influenced blood metabotypes

Genome-wide association analysis was performed for 529 metabolite profiles and the pairwise ratios thereof (for the ratio concept, see [145]). We identified a total of 299 significant mQTLs grouping into 145 statistically independent loci. Of those, 81 have not been reported previous to the publication of our study while the other 64 were observed before [145, 292-296]. Several of these known loci have been identified using tissues or fluids other than blood and/or different metabolite profiling platforms, suggesting that many GIMs are robust across platforms and tissues. In cases of associations with metabolite ratios, we were able to replicate the findings of previous studies [145, 292] that ratios can indicate changes in reactional activity (11 loci), substrate or product selectivity (five loci), or have normalizing statistical effects (seven loci).

4.2.3 Allelic architecture of metabolic loci

As described in Box 2, it has been observed that genetic variants associated with molecular traits such as metabolite concentrations on average explain more of the trait variance as compared to markers linked to complex trait endpoints [294]. Using the twin structure of the TwinsUK cohort, we applied the classical ACE twin model to estimate the heritability (see section 1.2) of metabolite concentrations. The variance explained by the mQTLs identified in our study was generally high (1-62% with a median of 6.9%). For 10% of the metabolites, the

explained heritability exceeded 20% and for four metabolites the identified associations accounted for more than half of the additive genetic variance (Figure 14). This is further proof that molecular (endo)phenotypes are more directly affected by variation of the genome sequence, leading to a greater statistical power to detect such associations.

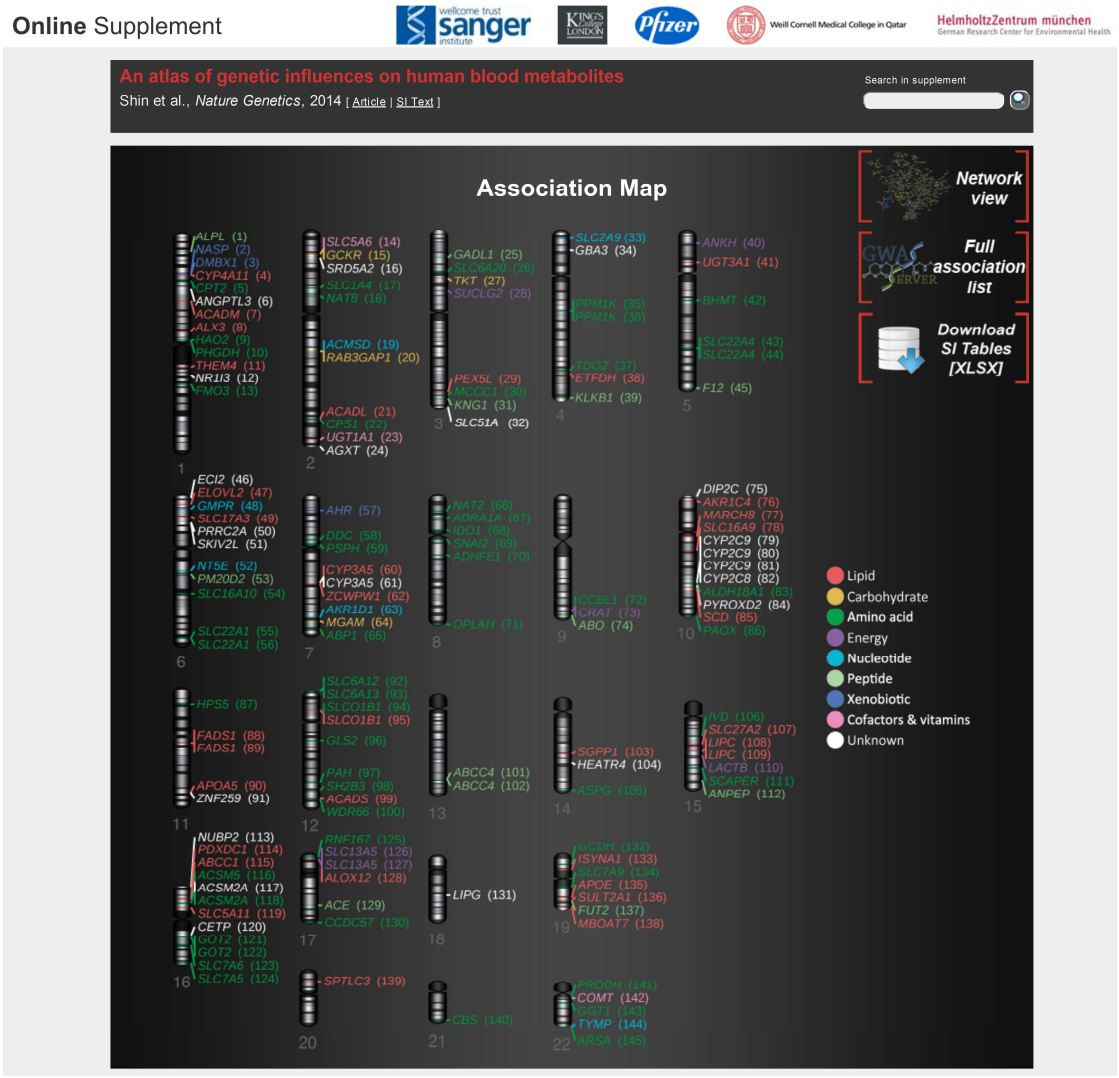


Figure 13: **Entry site for the Online Supplement of the study**. The association map at www.gwas.eu/si shows an ideogram augmented with the location of the 145 GIMs identified in our study. The color of a GIM shows the pathway of the associated metabolite (legend on the right). The figure is an image map, clicking on one locus leads to a detailed view of study results and biological annotations for this locus.

I have already discussed the potential of genetic interactions to inflate heritability estimates (Box 2, section 1.2). In this study, we conducted analysis of epistasis between all pairs of SNPs ($n = 106$) that were found to be significantly associated with the same metabolites ($n = 51$). We

detected only one significant interaction between the *NAT8* and the *PYROXD2* loci for the unknown metabolite X-12093 after Bonferroni-correction for 106 tests. As expected, including the interaction into the estimate of explained variance showed a slight increase compared to the purely additive model (15.6% versus 14.4% in TwinsUK and 27.7% versus 24.2% in KORA, respectively).

Figure 14: **Polar plot showing estimates for heritability and explained variance for metabolite concentrations**. Variances are partitioned into narrow-sense heritability and common and individual environmental factors as derived by the twin-based ACE model. For the numeric values, check the online supplement or the supplementary website at www.gwas.eu/si. Figure taken from [116].

In order to include even more molecular information into our study, we performed Mendelian randomization analysis (see section 2.4) of mQTLs and eQTLs obtained for a subset of 484 individuals of the TwinsUK cohort. We could identify two loci, *THEM4* and *CYP3A5*, where, depending on the present allele, increased metabolite levels were significantly associated

with decreased expression levels of the genes and vice versa (see Figure 4 in [116]). This shows the great value of including several –omics-levels into one analysis.

4.2.4 Relevance to biology, pharmacology, and disease

One of the central problems in GWAS interpretation is the projection of genetic variants to the causal gene underlying the disturbed mechanism which leads to the association. One benefit of mGWAS is that the molecular phenotype can be used to compare gene function to the observed outcome (provided that both the metabolite and the gene function are known). In this study, we used a multi-step approach to predict the genes that are most likely to be causal in the context of the observed associations. For this, we first extracted genes in a 500kb distance to the lead SNP of each association. The genes were then checked in pathway databases (KEGG, EHMN; for a description of utilized resources see section 3.4) for potential interrelations with the production or uptake of the associated metabolite. Co-occurrences of pairs of genes and metabolites in PubMed abstracts were mined to further establish the connection between. Based on the thus obtained evidences and a review using the enzyme resource BRENDA, the most functionally plausible gene was selected as predicted causal gene. This process was successful for almost two-thirds (94 of 145) of the identified loci. For another seven loci, we could assign biochemical functions not obviously connected to the associated metabolite (mostly, this is due to the metabolite being an unknown substance). These results are of great importance, as they present a validation of the functional annotation of the predicted causal genes as well as of the available metabolic reaction pathways.

It has been shown that genetic associations with complex traits often collocate with Mendelian disease genes [297], which led to the speculation that complex phenotypes may be milder forms of monogenic diseases caused by variants in the same genes but with smaller pathogenic effects. The same has been suggested for complex genetic influences on metabolic homeostasis [298], with inborn errors of metabolism being the monogenic counterparts of complex genetic factors modestly influencing metabolite concentrations. Therefore, we checked the 94 predicted causal genes for overlaps with the OrphaNet database (see section 3.4) which *inter alia* contains the known causal genes for inborn errors of metabolism. In 26 cases, our predicted causal genes are also causative for monogenic metabolic diseases, including *CPS1* (carbamoylphosphate synthetase I deficiency, MIM: 237300), *UGT1A1* (several recessive disorders linked to bilirubin levels, MIMs: 143500, 218800, 237900, 606785), *CBS* (homocystinuria due to cystathionine beta-synthase deficiency, MIM: 236200), and others. In addition to these 26 monogenic disease loci, I also annotated the 145 lead SNPs for overlaps

with genetic markers for complex traits as listed in the GWAS Catalog (see section 3.4). For almost one third of all reported loci ($n = 41$), we found associations to complex diseases or drug response endpoints. For instance, SNPs which are significantly associated with the concentration of several fatty acids in our study at both the *FADS1* and the *SLCO1B1* locus have been previously linked to response to therapy with lipid-lowering drugs (statins): the former with the efficacy of statins [299], the latter as risk factor for statin-induced myopathy [300]. In order to elucidate the relevance of GIMs to pharmacogenomics further, we collected data on genes that are involved in pharmacological targeting. To this end, we found evidence that more than 40% of our predicted causal genes are either targets ($n = 24$), metabolizing enzymes ($n = 11$), or transporters ($n = 3$) of approved drugs. An additional 11 genes are targets of drugs in preclinical or clinical trials.

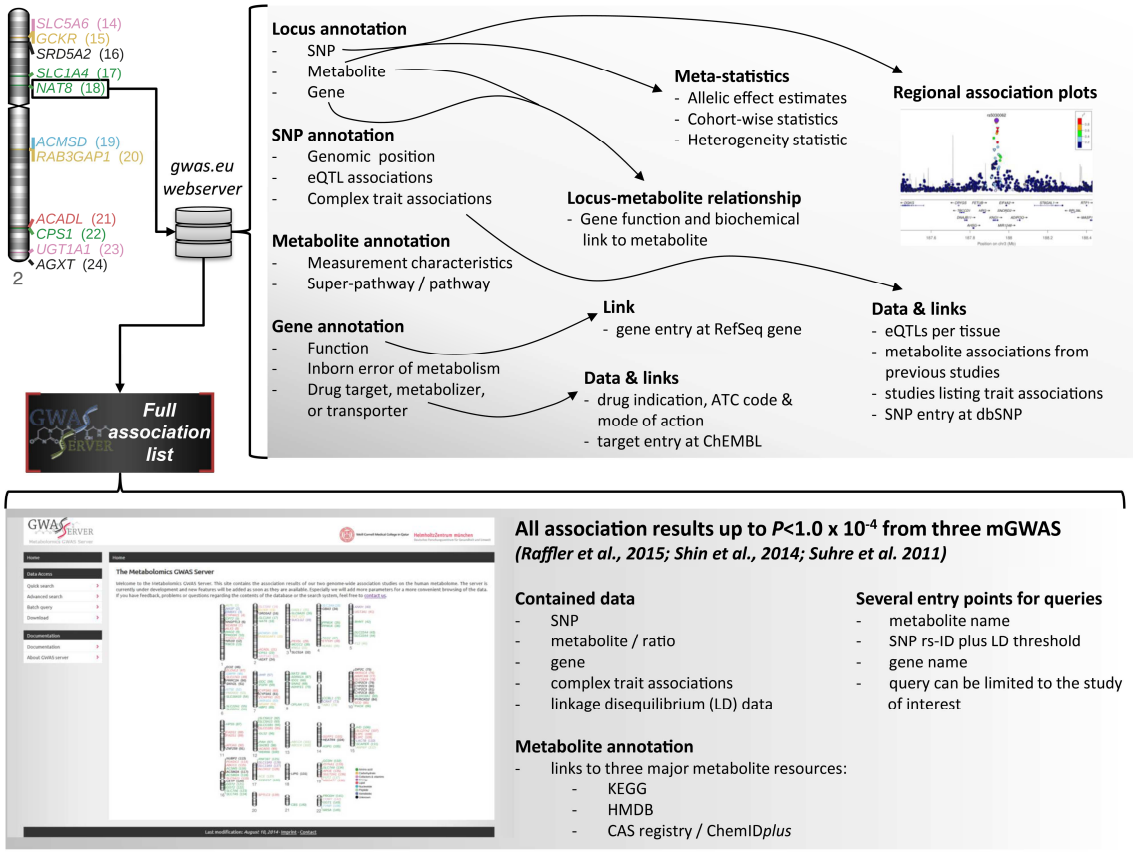


Figure 15: **Supplementary online resources for convenient access to the study data**. The upper part of the figure shows the content of the locus pages accessible through the online supplement at www.gwas.eu/si. Clicking on a GIM, e.g. the *NAT8*-locus, in the ideogram shown in Figure 13 directs the user to a detailed description of the locus comprising the information listed on the upper right part of the figure. Access to the full set of associations up to a P_{value}<0.0001 is provided through the Metabolomics GWAS server available at www.gwas.eu (lower part of the figure). The database also contains additional annotations including a mapping to GENCODE genes, URLs to locus views at dbSNP and Ensembl, metabolite annotations at three different resources, as well as links to PubMed articles that describe phenotype associations for the query SNP where available.

Combining all these results, we were able to obtain gene sets defined by GIMs that are either *i)* associated with complex traits or diseases or drug response endpoints, *ii)* causative for inborn errors of metabolism, or *iii)* targets, metabolizers, or transporters of drugs that are approved or under development. Intriguingly, these gene sets are not distinct but overlap (see Figure 5 in [116]), with genes that are listed in two ($n_{i) + ii)}$=4, $n_{i) + iii)}$=6, $n_{ii) + iii)}$=4) or even all three categories (n=3; *CPS1*, *SLC7A9*, and *UGT1A1*). These newly derived connections can not only be used to advance studies on the molecular background of complex diseases, they also identify potential targets for drug repositioning, development, and adjusted indications.

4.2.5 An online atlas of genetic influences on human blood metabolites

The scope of scientific publications on results obtained by large genome-wide screens of phenotype-SNP associations in print journals is space-limited to the description of only a few central findings. To allow for full exploitation of GWAS results (and especially of mGWAS results where many molecular traits (here >500) are tested for genetic associations), it has become common practice to provide access to the full association data sets through data deposition servers such as GEO, EGA and dbGaP. However, in order to enable access to the wealth of annotations for metabolite loci collected in our study, we decided to provide two supplementary webresources for convenient browsing of our study's results.

First, I reformatted many of the supplementary tables deposited at the online version of our article to be accessible through an online supplemental website (www.gwas.eu/si). Here, the entry point is an ideogram showing all autosomes annotated with the detected GIMs (Figure 13). Clicking on a GIM directs the user to a detailed description of the locus, augmented with the annotations for SNPs and metabolites defining the GIM as well as for the predicted causal gene. These locus pages are grouped into summary statistics of the meta-association analysis, biological annotations, pharmacological information, and locus information/regional association plots. Implementation of the resource is based on a PHP-framework nested with JavaScript elements for access to details such as heritability estimates that are hidden from the baseline information to preserve the clear and comparable structure of locus pages. I implemented the search interface on top of the website as exact full-text search of all included information obtained by indexing the complete data that shows a response time for any query in the range of 5ms.

Second, I developed a webresource for accessing mQTL data including associations not reaching genome-wide significance. As of now, the database lists association results from this

study as well as of two further mGWAS [165, 292]. Association data can be accessed through one of three query forms:

i) The "Quick search": a very simplistic query form where all associations contained in the database can be searched by entering a gene symbol, a metabolite name, or a SNP rs-number. When querying a SNP, a user-specified LD threshold can be used to expand the search to correlating markers (minimum correlation is $r^2 = 0.5$).

ii) The "Advanced search": as for some loci the quick search can produce a very long list of results, the advanced search can be used to limit the association results. In addition to selecting only associations from a single study or one of the included cohorts (where applicable), queries on associations with metabolites and ratios can be separated.

iii) The "Batch query": The batch query interface allows for batch annotation of large lists of SNP identifiers with mQTL results. For performance reasons, the interface is built similar to the advanced search and only specific datasets can be queried. Output is formatted as tab-delimited text, ordered by ascending P_{value}, and contains additional information such as metabolite resource entry identifiers.

Additionally, I provide bulk download for filtered associations ($P_{value} < 1.0 \cdot 10^{-4}$) as well as for the complete meta-analysis output.

When using the quick or advanced search, query results are listed in tabular format containing basic information including SNP position, alleles, allele frequency, effect allele, the association type (metabolite or ratio), the association source, and the minimal association P_{value}. If LD-expansion is used, the results for the query SNP and its proxies are listed in separate tables. For each SNP, the details for associations can be displayed in so-called SNP reports. These contain additional data on the SNP, a mapping to GENCODE genes, all metabolite/ratio associations from the selected study, and phenotype associations as contained in the GWAS catalog, including URLs to dbSNP, Ensembl, metabolite entries at HMDB, KEGG, and ChemID*plus* / the CAS registry, and PubMed articles describing the SNP-phenotype associations.

4.2.6 Concluding remarks

Here, I give only a short report of our findings. However, as we put much work into the provision of all data and its interpretation, I would like to point the attention of the reader to the original publication and the wealth of additional information provided in its supplement and the online resources.

When published, this study comprised the most comprehensive scan for genetically influenced concentrations of human blood metabolites. The multiple downstream analyses that we conducted in order to shed light on the relationships between the observed 145 GIMs, the genetics of other complex human phenotypes, and possible therapeutic interventions using drugs both approved and under development, as well as the systematic investigation of the allelic architecture of GIMs furthered our knowledge on several levels of biomedical research. To enable the convenient use of our study results, I developed two webservers that allow for browsing of all annotations that were collected and manually curated in our study. This combination of descriptive interpretations of metabolite loci with free full access to the complete association data is already widely used by the scientific community and facilitates the deeper functional investigation of the new hypotheses generated by our analysis.

4.3 Genetic influences on human urinary metabolites

In the last section, I described GIMs identified using plasma and serum metabolite levels. As mentioned, previous studies showed that genetic influences on metabolites seem to be robust across tissues, fluids, and measurement platforms. To investigate this matter further, we performed another large-scale discovery mGWAS in participants of the SHIP-0 cohort, but this time we used targeted and non-targeted proton nuclear magnetic resonance spectroscopy (^{1}H NMR) analysis of urine samples. Replication was performed using subjects from the KORA F4 cohort. In this section, I describe the key results of this study [165] which comprises the largest mGWAS on urinary metabolic traits to date, reporting 15 new urinary GIMs and replicating another seven previously identified loci. Intriguingly, 14 of all 22 identified GIMs (64%) also show associations with blood metabolite levels, enabling the study of the regulatory relationship between urinary excretion of metabolites and metabolic homeostasis in blood.

4.3.1 Methods summary

SAMPLES AND GENOTYPES – For the analysis of genetic influences on metabolite concentrations in urine, we performed a two-stage mGWAS including a discovery sample of 3,861 individuals from SHIP-0 and a replication sample of 1,691 subjects. Further information

on included individuals and covariates is listed in Table 10. Genotyping followed by pre-phasing and imputation yielded a final set of 15.9 million high-quality genotypes.

QUALITY CONTROL METRICS – SNPs were filtered based on HWE ($P > 1.0 \cdot 10^{-6}$), MAF ≥ 5%, and genotype call rate ≥ 95%. Before genotype imputation, 620,456 and 593,830 autosomal SNPs passed QC filters for SHIP-0 and KORA F4, respectively. Imputed genotypes were filtered for imputation quality score (IMPUTE info-score) ≥ 0.8 and MAF/HWE as for genotyped variants. Normalized metabolic traits were limited to compounds with at least 300 data points available for SHIP-0 and 100 data points available for KORA F4, respectively. Metabolite concentrations in the non-targeted analysis were log-transformed with base 10 and data points more than four standard deviations from the mean were removed, while in the targeted analysis outliers surpassing three standard deviations from the mean were excluded. The final dataset of urinary metabolic traits comprised 1,518 entries in the targeted analysis (55 metabolites and 1,463 ratios) and 13,861 entries in the non-targeted analysis (166 NMR peaks and 13,695 ratios).

Study	N by gender (m/f)	Age [years] (mean±SD)
SHIP-0 (discovery sample)	3,861 (1,901/1,960)	49.5 ± 16.2
KORA F4 (replication sample)	1,691 (826/865)	60.8 ± 8.8
total	**5,552 (2,727/2,825)**	**53.1 ± 15.2**

Table 10: Sample statistics and covariates.

REPLICATION OF ASSOCIATIONS – Replication in KORA F4 was performed for each locus significantly associated with metabolic traits in SHIP-0 in a stepwise manner, first trying to replicate the SNP/trait pair with the lowest association P_{value}. If this pair could not be replicated, we used the second best hit and so on.

METABOMATCHING – In order to obtain hints at the potential identity of non-targeted NMR peaks, we applied the metabomatching [301] annotation method. As reference set, we used the urine metabolome database [302], a subset of HMDB (see section 3.4). Final candidates as listed by metabomatching were manually curated.

4.3.2 Fifteen newly discovered genetically influenced urinary metabotypes

In order to reduce the computational burden, we first analyzed genotyped autosomal SNPs (n=620,456 after QC) for associations with metabolic traits in SHIP-0 at the classical threshold for genome-wide significance of $5.0 \cdot 10^{-8}$. In this step, we identified 499 mQTLs at 54 chromosomal regions (defined as 2Mb intervals centered to the lead SNPs). We then used all

genotyped and imputed variants within these regions for association analysis with all metabolic traits (targeted and non-targeted) below the significance level Bonferroni-corrected for the number of traits ($P_{value} < 3.25 \cdot 10^{-12}$). 2,882 variants within 23 distinct genetic loci were found to be significantly linked to one or more metabolic traits. For ratios, we additionally required a P-gain defined as $\min(P(M_1)/P(M_1 / M_2), P(M_2)/P(M_1 / M_2))$ ten times the number of tested metabolic traits (15,180 in the targeted and 138,610 in the non-targeted analysis, respectively) for genome-wide significance [303]. Replication in the KORA F4 samples was successful for 22 of the loci, 15 of which (*HIBCH*, *CPS1*, *AGXT*, *XYLB*, *TKT*, *ETNPPL*, *SLC6A19*, *DMGDH*, *SLC36A2*, *GLDC*, *SLC6A13*, *ACSM3*, *SLC5A11*, *PNMT*, and *SLC13A3*) have not been reported as associated to urinary metabolite concentrations before. The remaining seven loci were described in previous mGWAS in urine [296, 301, 304] and are replicated in our study with respect to both the associated metabolic traits and the direction of allele-specific effect estimates (for all loci and the collected annotations, see Table 3 in [165]).

4.3.3 From urinary GIMs to functional hypotheses

Consistent with the results described in section 4.2, we were again able to establish plausible biochemical relationships between the associated metabolic traits and the predicted causal genes for the majority of loci (15 of 22; 68%). To elucidate the functional relationship between metabolites as intermediate phenotypes and complex trait endpoints further, we again annotated all 22 GIMs with a large set of genotype-phenotype association and annotation databases as contained in the *SNiPA* webserver [58]: the GWAS catalog, OMIM variation, ClinVar, HGMD, and dbGaP (all described in section 3.4). Further annotations for the predicted target genes were retrieved from specialized databases and manual text mining. This integration of additional data led us to several plausible hypotheses that again show the benefit of including multiple omics-levels into analysis.

For instance, we identified a significant association of ethanolamine with variants upstream of the ethanolaminephosphate-phospholyase (*ETNPPL*) gene. It has been speculated that dysregulated homeostasis of its substrate, ethanolaminephosphate, which is the phosphorylated form of ethanolamine (EC 2.7.1.28) and which is degraded by *ETNPPL*, may contribute to psychiatric disorders such as schizophrenia [305]. This hypothesis is strengthened by differential gene expression analysis which yielded evidence for altered gene expression levels of *ETNPPL* in schizophrenia patients compared to control samples [306]. Proteomics data available at the Human Protein Atlas [307] shows that the protein product of *ETNPPL* is found in high levels

in the cerebral cortex and the kidneys, which labels these two tissues potential hotspots of ethanolaminephosphate degradation. The association found in our study indicates that accumulation of ethanolamine in urine may point towards a genetically reduced enzymatic velocity of ethanolaminephosphate degradation which could serve as a marker for predisposition to brain disorders.

4.3.4 Integration of mGWAS results in urine with blood GIMs

In order to compare our results to those of published blood mGWAS, we included mQTLs and their proxies in strong LD ($r^2 \geq 0.8$) from the metabolomics GWAS server (see section 4.2) and studies included in the GWAS catalog with an association $P_{value} < 5.0 \cdot 10^{-8}$. Eight of the urinary metabolite loci reported in our study seem to be urine-specific. However, the other 14 loci also comprise mQTLs in blood. Six of those (*CPS1*, *AGXT2*, *DMGDH*, *SLC6A13*, *HPD*, and *SLC5A11*) affect the same metabolite concentrations in both fluids with the same direction of effects in all but one (*SLC5A11*) of the six cases. In another five cases, the associated metabolic traits in blood and urine are biochemically linked *i)* via an enzymatic reaction (e.g. trimethylamine $\xleftrightarrow[\text{EC 1.5.8.2}]{}$ dimethylamine), *ii)* the enzymatic function of the predicted causal gene (e.g. *NAT8* which encodes an protein similar to N-acetyltransferases and is associated with different N-acetylated compounds in blood and urine), or *iii)* belong to the same molecular class (e.g. gluconate and erythronate are aldonates). For the remaining three loci, the metabolic traits in both media have no obvious connection.

The *SLC5A11* locus, which in this study is linked to increased urinary *myo*-inositol concentrations per copy of the T-allele of SNP *rs17702912*, was associated with lowered levels of the same metabolite in blood with the same allele in the mGWAS described in section 4.2. The s*olute carrier family 5 (sodium/inositol cotransporter), member 11 (SLC5A11)* transports inositol in concert with sodium [308]. The reversed genetic effects on *myo*-inositol concentrations seen in blood and urine suggest that *SLC5A11* may be implicated in the re-absorption of *myo*-inositol in the proximal tubule of the kidney as has been previously hypothesized [309]. To follow up on this theory, we calculated the association of *rs17702912* and the ratio of blood and urinary *myo*-inositol levels for the KORA F4 subjects from the replication sample of this study using the blood concentration measurements (normalized to circulating creatinine) of the same individuals contained in the study described in section 4.2. The association shows an increase in strength (by means of the P_{value}) of seven orders of magnitude ($P_{urine} < 1.95 \cdot 10^{-24}, P_{blood} < 1.50 \cdot 10^{-4}, P_{ratio} < 2.43 \cdot 10^{-31}$) as compared to the association in urine alone. This may indicate a direct relationship between the opposite genetic

effects on *myo*-inositol concentrations in the two fluids caused by altered function of *SLC5A11* that indeed would be in line with a reduced re-absorption rate in carriers of the effect allele.

4.3.5 An automated approach for assigning predicted causal genes

In the study on genetic influences on human blood metabolites (section 4.2), we put major effort into the assignment of the predicted causal gene for each locus, using manifold means of manual annotation. The objective, data-driven projection of the effects of genetic markers onto candidate genes, however, is one of the major challenges in current genetic analyses. In this study, I have developed such an evidence-based method for the detection of the most plausible predicted causal genes.

As the first step in candidate gene selection, we assigned the significantly associated SNPs to distinct loci using a physical distance threshold of 1Mb. Assigning variants within a locus to one of the covered genes based only on proximity or plausibility ignores haploblock structure and existing regulatory information for the SNPs such as eQTL associations. To take such information into account and to achieve an unbiased selection of candidate genes, I collected evidence for each significantly associated SNP and its proxies in strong LD ($r^2 \geq 0.8$). I received a list of candidate genes that are linked to any of these variants (including LD-proxies) via the following criteria to identify candidate genes: *i)* Genomic proximity: genes that harbor or are in close proximity (<5Kb) to any of the variants. *ii)* eQTL association (eQTL datasets are listed in section 3.4.3): genes where altered expression levels have been discovered to associate with any of the variants. *iii)* Regulatory elements (ENCODE and FANTOM5, see section 3.4.2): potentially regulated genes that are associated with a promoter/enhancer/repressor element containing one of the variants. Further evidence for potential involvement of a gene was assumed if *iv)* the variants contain a missense variant for a protein product of this gene and *v)* if an intragenic variant is annotated as pathogenic in one of the available phenotype databases (see section 3.4.4). For each gene, I counted how many of the aforementioned criteria are met. I then finally assigned the locus to the gene with the strongest functional evidence (i.e., the gene showing the highest number of different types of evidences (max. 5) among the candidate genes). In case of ambiguous assignments, the gene with the most plausible biological function as determined by manual text mining was chosen.

As an example, one locus on chromosome 2p13.1 contains a high number of SNPs with strong associations with non-targeted traits corresponding to N-acetylated compounds. These SNPs are distributed over 12 different genes. The gene covered by the highest number of SNPs is *ALMS1*. However, there are 3 more genes in this locus with the same functional evidence

count as *ALMS1*. One of these genes is *NAT8*, which encodes an N-acetyltransferase. Since there is a biologically meaningful link between the function of the *NAT8* gene product and the associated metabolic traits, I annotated this locus with *NAT8* as the most likely candidate gene. Interestingly, in the Shin et al. study, we manually selected the same gene for the overlapping blood GIM, a decision also based on the identity of the associated metabolic trait (which again was an N-acetylated compound) [116]. For the whole list of evidences collected for all loci, see supplementary table S3 in [165].

4.3.6 Concluding remarks

As I have shown in the previous section, the identification of genetic effects on intermediate phenotypes such as metabolite concentrations is a valuable tool to gain insights on the cellular disturbances possibly predisposing to human trait endpoints. However, it is not obvious if the findings obtained in one tissue or body fluid can mirror the effects taking place in the primary affected tissue. Here, I shift the focus from GIMs detected in blood to genetic variants modulating urinary metabolite concentrations. Although we also find some GIMs that seem to be specific to one fluid, we were able to show that, to the larger fraction, GIMs in both media are co-located. With the example of *SLC5A11*, we show that the combination of the genetic effects across different biological samples can even yield new hypotheses regarding the functional mechanisms underlying a genetic signal. For further downstream analyses, I have again made all association data available at the metabolomics GWAS server (www.gwas.eu; see section 4.2).

4.4 Summary

In this chapter, I described the process of how genetic associations with complex human traits can be investigated. In the first section on the genetics of the sudden infant death syndrome, I gave a detailed outline on the methodological challenges of GWAS and SNP array-based CNV analysis, discussed how best-practice quality control measures are to be applied to the primary data, and exemplified how genetic loci that show association to a trait can be interpreted functionally. In the following two sections, this matter was elucidated further in the context of metabolomics GWAS within and across the respective studied body fluid. Here, I

put emphasis on more specialized approaches (evidence-based locus to gene assignment) and outcomes (explained variance). Additionally, I underlined the importance to provide access to the results of biomedical studies in order to allow for further downstream analyses. In the next chapter, I will show how these results can be integrated and combined with the results from other studies, enabling the investigation of new relationships between genetic information and complex human traits.

5 Annotating the variome

The biological interpretation of genetic loci (as defined by variants in LD) and their effect on associated traits and diseases relies on the annotations available for the respective genetic region and the respective variants, as well as on the knowledge about the molecular pathways involved in the expression of the trait or disease phenotype. If there is no other information available, trait associations for a locus can *per se* be used to more closely characterize the biological mechanisms linked to both the genetic region and the trait. In the last chapter, I outlined the workflow of how genetic associations to human traits are obtained and gave an introduction to the methods of how these associations can be projected on candidate genes. Specifically, I described an automated approach to this task that uses a simple evidence-based metric to assign a target gene to a GWAS-identified genetic locus. This approach is based on a large, integrative data collection accessible through a user-friendly webserver called *SNiPA* that we developed in order to support scientists in the inspection of the potential functional implications underlying a genetic association signal. This chapter presents this resource, the integrative approaches we applied to obtain its data basis, and the data harmonization and consolidation methods we used, thus substantiating the introductory chapter 2 on data integration with a practical implementation. The following is based on and extends our application note on the resource published in *Bioinformatics* [58]. As I designed the resource to be both automatically updatable and easily extendible, the presented methods as well as the contained data sources for the current *SNiPA* version differ in part from what we have described in the original publication.

5.1 Methodological aspects

Development of large integrative genomic resources is an interdisciplinary task. First, the desired content, access interfaces, as well as the potential downstream applications and use-cases have to be determined. In the second step, datasets that fit the purpose of the specified focus of the resource have to be collected and integrated. In the third step, the integrated data has to be consolidated in an aggregated or mediated view that is optimized to be queried from the intended interfaces. Finally, the interfaces have to be related to data access points implemented in a technical platform that allows for interaction with requests by the end-user. This section will describe these process layers as we have designed them for the *SNiPA* web server.

5.1.1 Specification of requirements

The basic aim of the resource was to facilitate the annotation of genetic loci defined by the contained elements such as genes and regulatory elements with evidences for functional effects possibly exerted by genetic variation. We therefore conceptualized the standard approach for mechanistically annotating genetic variants:

1. Definition of the genetic variation. For SNVs, this includes the extraction of correlating variants via LD. For CNVs, this means the exact determination of the break points.
2. Annotation of the genetic variation with the available catalog of variant-linked information such as trait associations or conservation or deleteriousness scores.
3. Retrieval of potentially affected genes via a weighted combination of genomic proximity to the genetic variation, eQTL associations, and gene-associated regulatory elements (this is basically the approach described in the previous chapter).
4. Collection of further evidence that corroborate the relation between the genetic variation and the candidate genes (amino acid exchange, pathogenicity).
5. Basic annotation of the affected genes (disease associated, monogenic disease genes).

As further annotation of candidate genes is often context-specific (protein-protein interactions, metabolic or signaling pathway, information from knockout models, etc.), we decided to exclude more specialized gene annotations from the database. Also, we decided to put our focus on single nucleotide variants that are the most frequent (>99.9%) form of genetic variation in the human genome [231]. Also, for larger CNVs there are very good annotation tools and databases available (including the aforementioned Database of Genomic Variants [194]). This led to the definition of the desired content:

1. A background set of SNVs, including their frequency in the general population, their genomic location, and their correlations in terms of LD.
2. Data on variant-linked information, including conservation and deleteriousness scores, pathogenicity, eQTL associations, trait associations, as well as effects on gene, transcript, and protein sequences.
3. A basic genomic backbone consisting of up-to-date gene, transcript, and protein annotations and regulatory elements. This includes a mapping between the oligonucleotide probes used for eQTL analyses and the gene/transcript annotation.
4. Annotation of genes with trait associations and monogenic disorders.

For access to the data, we wanted to provide query interfaces to all tasks involving genetic variants. These we defined as:

1. An interface to define the genetic variants which, as mentioned before, for SNVs means retrieval of correlating variants.
2. An interface to retrieve variant-linked data as well as their evidence-based effect annotation.
3. An interface to retrieve combined annotations for a set of variants, defined either based on the results of a genetic association study, their correlation via LD, or their genomic position.
4. A graphical user interface to inspect LD-based genetic loci.
5. A graphical user interface to visualize results from genetic association studies
    a. in the locus context.
    b. across the whole genome.
6. A graphical user interface to screen the variant content in a genomic region.

5.1.2 Data integration and harmonization workflow

The first decisions in the design of the data integration workflow regarded the source for genome annotations and the background variant set. For the former, I chose the Ensembl database [149] as backbone for genomic data including gene, transcript, and protein data as well as basic regulatory element annotations, because the Ensembl development team closely collaborates with the ENCODE project consortium [78, 81] which is the major source of functional annotations for the human genome. For the variant set, I relied on the SNV data released by the 1000 genomes project [77, 231] which is the largest international sequencing project of population-based reference genomes. Although it has to be mentioned that, in non-coding regions, only low-coverage sequencing was used to generate the 1000 genomes data, the

consortium applied a whole set of quality control measures (including validation via deep sequencing in 10% of the individuals) that make this variant set unique in its form. To filter out further potential artifacts resulting from low sequencing coverage, I installed additional quality criteria (only bi-allelic variants; no short insertions and deletions; unique mapping by dbSNP [178] to two genome assembly versions). When I started collecting the data basis for the resource in autumn 2013, the current genome assembly was GRCh37 to which all of annotations from Ensembl and other data sources were aligned. The SNV data from 1000 genomes was also mapped against GRCh37 and, therefore, we could directly annotate variants with functional evidences using the respective genomic coordinates. This, however, changed in 2014 when the Ensembl database was converted to include the latest genome assembly, GRCh38. Afterwards, the data from Ensembl was conform to the new assembly, while all other annotation sources as well as 1000 genomes still used GRCh37 coordinates. To resolve this conflict, I had to apply several mappings between the two assemblies. For gene, transcript, and protein sequences as well as ENCODE regulatory feature clusters, I obtained the set of entities that could be unambiguously projected between the two assemblies using the UCSC liftOver tool [216]. For SNVs, I used the dbSNP database [178] that realigns all submitted variants (including those from the 1000 genomes project) to every genome assembly to retrieve the set of variants that can be mapped to both assemblies. These sets were used as whitelists in all further annotation steps, while entities not contained in the lists were excluded during the SNV annotation process. As up-to-date sequence mappings were only available for GRCh38, I further had to separate the annotation of effects on genes, transcripts, proteins, and regulatory feature clusters from the annotations aligned to GRCh37 (Table 11). Integration of the datasets is also differently handled for dynamic (i.e. regularly updated) and static datasets.

The dynamic data comprises primarily the gene, transcript, and protein alignments, as well as regulatory elements, mappings of gene-associated promoters and enhancers, eQTL probes, and miRNA target sites to the gene set, deleteriousness scores for amino acid exchanges, and trait associations for variants and genes. Datasets that are not as frequently updated, but still require the implementation of update routines include the variant background set from 1000 genomes (which can, except for LD calculations, be performed incrementally and, thus, is comparably trivial), realignments of the variants from dbSNP (which may affect the set of variants that can be mapped to both genome assemblies), dbSNP identifier history (contained in the dbSNP RsMergeArch table [178]), and genome-wide conservation scores (in case that additional mammalian genomes are included in the multiple sequence alignment).

As gene-, transcript-, and protein-linked annotations are dependent on the gene build, gene information (stable gene identifier (ID), gene symbol, genomic coordinates, size, strand, accession numbers, synonyms, biotype and full gene name) is gathered as first integration step via querying the latest release of the Ensembl MySQL database on the fly. The obtained gene set is annotated with transcript and protein information as well as trait associations and annotations from DECIPHER [196], OMIM [106], and OrphaNet [107]. Then, the bidirectional mapping between GRCh37 and GRCh38 is performed to obtain the final set of genes. This is followed by fetching the Ensembl regulatory build from the Ensembl MySQL database and again the final set of elements is derived using the bidirectional mapping between the assemblies. Next, the mapping of microarray probe sets (for eQTL mapping) and of gene-associated promoters and enhancers to the genes in the final list is obtained. Finally, variant-trait associations are downloaded from the different sources (DrugBank [221], dbGaP [226], ClinVar [227], GWAS Catalog [100, 222], HGMD [224], OMIM variation [106], UniProt [225], and the metabolomics GWAS server; see chapter 3), and deposited in standardized data files. As these data sources are more frequently updated than the Ensembl database, variant-trait associations can also be added to the existing annotation compendium without the need to complete the whole integration cycle.

Data type	$N_{Entities}$	Assembly	$N_{Sources}$	Mapping strategy
Genes	59,413	GRCh38	1	position-based
Regulatory feature clusters	406,822	GRCh38	1	position-based
AAE deleteriousness scores	e.w.	GRCh38	2	position-based
Bi-allelic single nucleotide variants	78.2 Mio.	both	1	position-based
Gene-associated promoters and enhancers	1,064,990	GRCh37	3	position-based
miRNA target sites	606,059	GRCh37	5	position-based
Conservation scores	g.w.	GRCh37	3	position-based
Deleteriousness scores	g.w.	GRCh37	3	position-based
eQTL associations	20,855,118	-	9	ID-based
Variant-trait associations	408,325	-	9	ID-based
Gene-trait annotations/associations	11,658	-	3	ID-based

Table 11: **Data integrated in the *SNiPA* resource**. The primary datasets and sources are listed in chapter 3. The annotation sets mapped to the older GRCh37 assembly clearly outnumber the annotations aligned to GRCh38. Therefore, I used the GRCh37 assembly as backbone. ID-based mappings are updated regularly to ensure full compatibility of ID-lists. Abbreviations: AAE – amino acid exchange; e.w. – exome-wide; g.w. – genome-wide.

SNV data is obtained from the 1000 genomes and filtered for bi-allelic variants in order to provide globally valid, allele-specific variant annotations. Variants are then mapped to the current dbSNP release and aligned to both genome assembly versions. Simultaneously, the

dbSNP ID-alias mapping is updated. Variants in the final SNV mapping are then stored in a database. LD calculations are performed on the 1000 genomes genotype data in full parallelization, meaning that the genome is split into 1Mb chunks with pairwise overlaps of 250Kb and LD-values are computed in 500Kb windows centered to the respective variants down to a threshold of $r^2 \geq 0.1$. After completion, the chunks are rejoined chromosome-wise.

Genome-wide conservation scores (GERP++ [200], PhastCons, and phyloP [199]) are only available in compressed and indexed bigWig format. Despite the index structure, access to the data in these files was not very performant which is due to the large file sizes leading to inefficient index structures. I therefore split the data chromosome-wise and used Tabix [255] for indexing which performed better than bigWig indexing.

The static datasets contain stable information from published studies. As of now, this applies only to eQTL association data, which are preprocessed (for details see chapter 3) and then stored in standardized formats, and the genome-wide CADD scores [201] (which *per se* needs not to be updated but only to be filtered to comply with the alleles of the latest variant mapping). As for the conservation scores, for performance reasons CADD scores were also split into single chromosomes to enable faster data retrieval.

5.1.3 Data consolidation

After all data sources have been integrated and harmonized, ID-mappings and the variant set as well as its mapping to both assemblies are available, the actual annotation of variants is performed (Figure 16). This is done in a slimmed version for the GRCh38 assembly to retrieve gene, transcript, and protein effects and associated deleteriousness scores as well as the Ensembl regulatory build comprising Encode ChIP-Seq clusters and TF binding motifs using the Ensembl VEP tool [140]. The same is done for the GRCh37 assembly, but here also the additional position-based datasets conform to the assembly (miRNA target sites, gene-associated promoters and enhancers, and the GERP++, phyloP, phastCons and CADD scores) are included in VEP annotation. For subsequent processing of the data and further annotation of variants, genes, transcripts, and regulatory elements that is not contained in VEP output, I have written a modular Perl program that combines the results from VEP with the additional data collected, as well as conflates the information to a consistent collection expressed and confined to the *SNiPA* variant, gene, and regulatory builds. The annotation software is controlled by a configuration file that can be adjusted to fit the collected data basis at time of build. This enables also partial updates to the variant annotations, for instance via adding entries for new eQTL or trait association data sets. The configuration almost completely determines the annotation

procedure, and is only adjusted by setting a special annotation mode (*--slim* for the slimmed annotation of VEP GRCh38 results and *--merge* for combining the slimmed annotation results with VEP GRCh37 results and the additional data specified in the configuration) at program execution. When executed, the program loads the Filehandle module that contains functions for handling all data formats as we have specified them in the data integration workflow and reads in the data sets listed in the configuration file. Afterwards, the user-specified annotation process module is loaded and performs variant annotation.

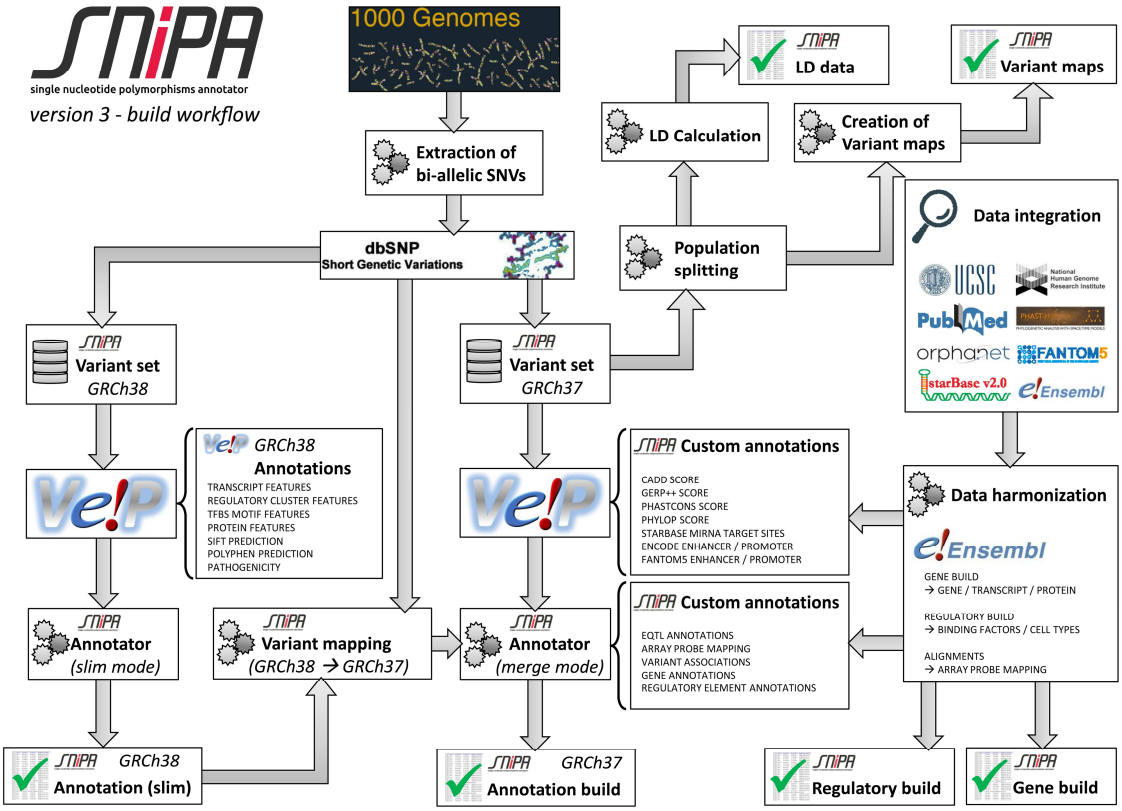


Figure 16: **Updatable data integration and consolidation workflow of the *SNiPA* resource**. The data basis of the resource is separated on five independent distributed systems: The gene build, the regulatory build, the variant annotation build, the variant maps, and LD data. Although stored in a distributed way, the production process of the single builds is interdependent. The pipeline starts simultaneously at two points: integration of 1000 genomes SNV data (top) and the Data integration node (right). The variant set is obtained by bidirectional mapping of variants over the two genome assemblies GRCh37 and GRCh38 using probe alignments and ID-alias mapping from dbSNP. Consecutively, LD calculations as well as the variant maps can be independently calculated across the whole genome. Data integration starts with producing the gene and regulatory builds also applying a bidirectional mapping between genome assemblies using the UCSC liftOver tool. Afterwards, the annotation data sets for variant effect prediction are harmonized to the gene and regulatory builds. As up-to-date gene mappings from Ensembl are only available for GRCh38 but the majority of annotations used by *SNiPA* are conform to GRCh37 coordinates, effect predictions on gene, transcript, and protein sequences as well as annotation with the corresponding deleteriousness scores is performed separately on GRCh38 (slim annotation mode). This annotation is then merged with the annotation data conform to GRCh37 using the variant mapping between the assemblies. Annotations that are in conflict with the gene or regulatory build are filtered out.

In default mode, the annotation is performed linearly, using the VEP output as input, performing additional data annotation, and creating output files that are formatted such that they can be used by the *SNiPA* web interface, which includes encoding of data arrays, bgzip compression, and Tabix-indexing of the results files. As VEP prints out one line per entity altered by one allele of the variant (which means that for the same variant there can be several lines), the input is read consecutively and checked for the variant of the current line being equal to the variant in the line before. If this is the case, the information contained in the current line is further annotated and added to the complete annotation of the variant. If not, the variant annotation of the previous variant is printed to the output file and cleared from the memory and a new structure for the current variant is created. Annotation is completely ID-based, as the VEP includes only very basic information in its output (e.g. the ID of the regulatory element hit by the variant, but no information on the activity of the regulatory element across cell types or its associated genes). Therefore, the whole gene and regulatory builds obtained before are loaded into hashes held in the memory, as are eQTL probe mappings, trait and eQTL associations, and the updated ID-mappings. For performance reasons as well as to enable the parallelized computation of variant annotations, all datasets are split into single chromosomes. This excludes eQTL probe mappings and the gene build as *trans*-eQTL datasets can contain associations to genes on other chromosomes. The cashed hashes are formatted such that the contained information can be directly used in the serialized output which makes the annotation *per se* very efficient as it only comprises obtaining of the information for the ID and its aliases from the hash. Complete processing of chromosome 1 containing approximately six million SNVs takes about 75 minutes. The output of the program is a hybrid of compressed annotation data and fields that are used for display of the data in a web interface. It contains the chromosome and position of the SNV (which are used for Tabix-indexing), the variant's current dbSNP ID, integer codes for plotting of the most severe effect annotation, disease association, and the information if there are several deviating annotations available for the variant, the HTML-code for the tooltip containing condensed annotation data (Figure 18), the encoded annotation array, and a comma-separated list of the effect categories annotated for this variant. As the variants subjected to VEP annotation are sorted with respect to their genomic position, VEP output and thus the output of the annotation program is also sorted, which allows for direct compression of the output using bgzip followed by position-based indexing of the output.

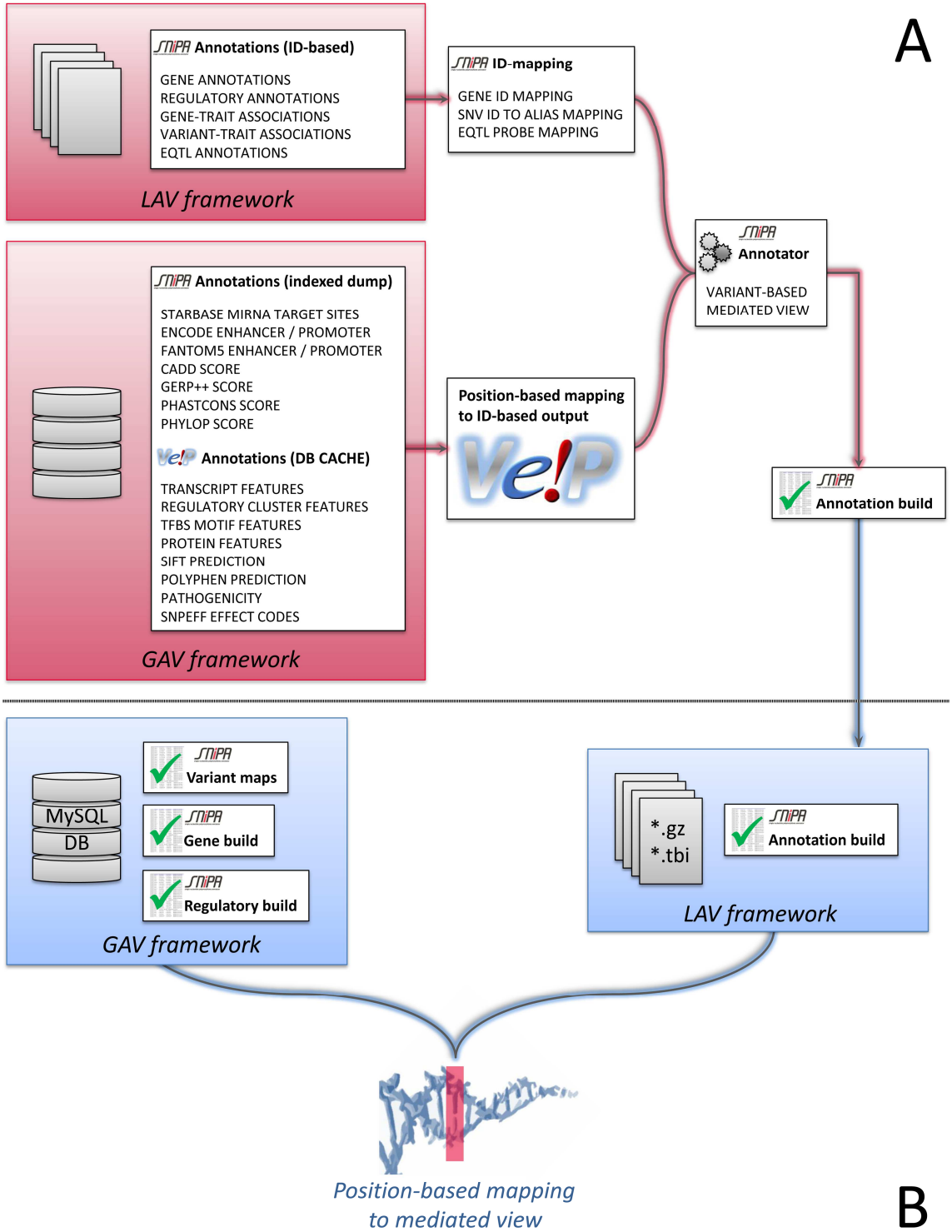


Figure 17: **Data integration frameworks for production database (A) and stable but updatable releases (B).** In order to be easily updatable and extendible, ID-based annotation datasets are stored in a LAV framework that is linked to the annotator software via up-to-date ID mappings. Data that is less frequently updated is stored in a data warehouse. On update of this GAV-stored data, position-based data is converted to ID-based data using the VEP tool and consolidated in a mediated view using a projection of evidences to variant effect predictions. The thus obtained variant annotation build is deployed to the stable data release server and combined with the position-based data of the backbone of genomic annotations.

In slim mode, the input from VEP is only reformatted to be more efficiently accessible in the downstream merge mode annotation. This is performed using compression of variant annotations in JavaScript object notation (JSON) arrays that can be conveniently converted into Perl hashes, leading to constant accession times as soon as the data is loaded to the memory. In addition, the mapping between the assemblies is applied here. Thus, the slim annotation data is saved with GRCh37 coordinates. As variant annotation is performed separately for each chromosome, this mapping step resolves the issue of genome locations that are aligned to a different chromosome in the GRCh38 assembly.

In merge mode, in addition to the full set of (chromosome-wise) annotations the slimmed annotation data obtained for GRCh38 is loaded into the memory. For the larger chromosomes, this leads to a quite large consumption of memory: the slimmed annotation for chromosome 2 has a file size of 6.2 GB of data, translating to more than 31GB of memory consumption. However, when using a machine with enough working memory, this results in equal run times for the merge mode and the linear default annotation mode as the preprocessing of variant effects on genes, transcripts, and proteins compensates for the quite extensive process of reading the slimmed annotation into memory.

5.1.4 Integration framework and data representation

As mentioned before, the data basis for the *SNiPA* resource consists of different types of information with some data being stable, sometimes updated, or regularly updated. To be able to account for this different nature of the data sources, I split the integration of the sources into different frameworks, as well as we separated the deployed stable releases from the production database. Although it is of course favorable to provide the latest data basis to the user, for large databases this is not feasible. Therefore, almost all large genomic resources (e.g. Ensembl, dbSNP, NCBI, and UCSC) use stable releases in data warehouse frameworks. Although in general we also follow this approach, we wanted to be able to both include new datasets at any time and update information from resources with very frequent update cycles dissolved from the whole annotation build cycle.

As I use the Ensembl database as backbone for genomic annotations, the major datasets in our resource follow the update cycle of Ensembl which is released quarterly. The information from Ensembl is integrated within a GAV framework in the production database. It consists of a cached and compressed flat file version of the VEP annotation database and is combined with our custom additional datasets with stable positional data (Figure 17). This additional data is stored in compressed and Tabix-indexed flat files. Although it would of course be possible to

dump the whole data into a database, the aforementioned performance reasons (see chapter 2) as well as the necessity to provide an additional interface between the database and the VEP tool disposed us to use this solution as VEP has an inbuilt interface for compressed and indexed flat files. Also, this allowed us to reduce the required disk space by a factor of almost 7 to 273GB.

ID-based data (e.g. eQTL or trait associations, and gene annotations) on the other hand was integrated in an LAV framework, using standardized data formats implemented in the Filehandle module of our annotator program. Thus, additional data sources can be integrated in the annotation process by simply reformatting the results files from a given study to fit my internal file type conventions. Mapping to the mediated view that, in the production database, consists of the variant-projected conversion of annotation data to variant effect predictions, is then automatically performed by the annotator software after the configuration has been adjusted. This also enables the dissolved update of phenotype association data that is more frequently updated than the Ensembl database. An important issue in this case is the update of ID mapping tables to fit the Ensembl release in use. This is, therefore, realized in independent processes incorporated in the finalization procedure of the gene, variant, and regulatory build cycle which are only executed if a new Ensembl release becomes available.

The deployed stable releases are also split into two frameworks. Similar to the above, this is mainly done because variant annotation data is more variable than the annotations contained in gene, regulatory, and variant builds. Therefore, the latter are stored in a GAV framework implemented in a MySQL-based data warehouse. However, after data integration and calculation of variant-based statistics, the disk space requirements exceeded ranges that can be conveniently handled in MySQL systems (uncompressed size ca. 4.1TB) without extensive normalization and optimization. Variant annotations and population-specific variant statistics (such as allele frequencies and LD data) were therefore stored in compressed and Tabix-indexed files (reducing the complete size of annotations plus MySQL warehouse to about 635GB) that are accessed via the variant build stored in the MySQL database. Thus, we ascertain ACID requirements and nevertheless escape the rigidity and the performance limitations of the relational schema. Using position-based queries, the variant annotation data is again consolidated with genomic datasets (i.e. the gene and regulatory builds) in a genomic coordinate-based mediated view.

5.2 Accessing variant annotations

In order to provide access to the extensive set of genome-wide annotations I have collected, we implemented user-friendly starting points for all the interfaces described in the previous section: *i)* a variant-centered implementation of a genome browser ("Variant Browser"); *ii)* "Association Maps" for browsing through GWAS results; *iii)* an interface for batch retrieval of variant annotations via ID-list, gene ID, or genomic coordinates ("Variant Annotation"); *iv)* a combined listing of annotations across a set of variants within LD blocks or chromosomal regions ("Block Annotation"); *v)* "Regional Association Plot" and "Linkage Disequilibrium Plot" that combine publication-ready plotting of association results and LD values, respectively, with the interactive interface of the "Variant Browser"; *vi)* "Proxy Search" and "Pairwise LD" that allow querying pre-calculated LD values augmented with variant annotations. The following gives a more detailed description of the individual interfaces.

5.2.1 The variant browser

The *SNiPA* Variant Browser is our version of a genome browser with a variant-centered point of view (Figure 18A). Our main focus was to enable the user to visually assess how well the variants in a locus are characterized by evidences. To achieve that, variants are plotted according to their highest effect category meaning that the higher a variant is located in the plot, the more evidence exists for it to feature strong effects. Variants that are assigned to more than one effect category are highlighted in green, while variants that have trait annotations available are highlighted in blue. Here, the symbols used for the variants and their location in the plot are redundant information. This is because the two other interactive plotting modules of *SNiPA* (LD plot and regional association plot) implement the interface of the browser and use other means of variant positioning, and there the used symbol is the only visual hint at the assigned effect categories. An additional feature of the browser (and of all visualizations implementing the browser's interface) is that the display can be exported as vector image, PDF, or PNG. Also, the browser is fully interactive, meaning that hovering over an element in the browser display will show a tooltip with further information on the selected element. This was implemented using the plotting library Highcharts and extension of interactivity by adding jQuery and the jQueryUI JavaScript environments (see chapter 3). Clicking on genes or regulatory feature clusters link to the elements' entry in the Ensembl database, promoters are linked to FANTOM5 promoter annotations, and enhancers as well as ENCODE promoter elements link

to the Ensembl genome browser for further inspection of the genomic context. Variants, on the other hand, are linked to the full set of annotations presented as "*SNiPA* cards" grouping information into semantic sections. All annotations therein are linked to their primary sources and to the Ensembl genome browser.

5.2.2 Association maps

To inspect variants or sets of variants that are associated with a specific trait (or a set of traits), I have implemented this module that allows for access to the data in *SNiPA* for variants with published associations (Figure 18C). "*SNiPA* cards" of the variants can be directly accessed from the karyogram. Furthermore, variants can be added to the Variant clipboard (which is similar to a shopping cart) and then be pasted from the clipboard into other modules such as the linkage disequilibrium plot for an LD-based locus inspection, the LD-based block annotation to get a summary of annotations for all correlating variants, the proxy search to retrieve a table of these variants with or without dense annotations, or the variant browser for further inspection of flanking regions of the locus. As it is a nice feature to plot genome-wide association results in a karyogram, I also provide the possibility to create own association maps via uploading ID-lists of variants.

5.2.3 Variant and block annotation

The Variant Annotation module provides direct access to variant annotations contained in *SNiPA*. Given a user-specified list of dbSNP identifiers, *SNiPA* returns a list of "*SNiPA* cards". The Block Annotation module enables retrieval of merged annotations of a set of variants that can be specified by four different ways: a list of dbSNP identifiers, one dbSNP identifier that is first used to obtain a list of correlating variants via a user-specified LD-threshold, a gene identifier, or a chromosomal region. Currently, only variants located on the same chromosome can be processed by block annotation. The merged annotation can be used to characterize a whole locus which can for instance be used to obtain evidence-based variant-to-gene projections as described in section 4.3.

5.2.4 Regional association plot

This is the classical plot for visualizing locus-based association results in a regional Manhattan plot. Input is a user-specified list of variant/association p-value pairs. Variants are plotted by their position on the x-axis and $-\log_{10}(P_{value})$ on the y-axis. In addition, variants are colored by their correlation with the sentinel variant (by default, this is the variant with the lowest P_{value} but optionally it can also be specified by the user). This plot implements the

interface of the variant browser, meaning that all functionalities of the variant browser are provided except for navigating to other loci.

5.2.5 Linkage disequilibrium plot

This plot is very useful for instance to inspect a published GWAS hit. It is common practice to select a single variant (e.g. the one with the lowest P_{value}) as published representative for an association signal. LD data can be used to reproduce the reported locus albeit there always will be differences as the study populations will not be perfectly resembled by 1000 genomes individuals. Input is a single dbSNP identifier. Variants are plotted by their position on the x-axis and their correlation (r^2) to the specified variant on the y-axis. This plot implements the interface of the variant browser, meaning that all functionalities of the variant browser are provided except for navigating to other loci. Instead, the plot can be updated by selecting any contained variant as locus representative.

An additional feature is that variants of special interest can be highlighted. The respective variants have to be entered in addition to the sentinel variant and the visualization type can then be specified. There are two options: either, the full plot is generated containing all variants showing any correlation with the sentinel, but variants that are not listed in the highlighting list are faded out (Figure 18B); or, the plot is restricted to the variants in the highlighting list. This plot basically presents a visual representation of the Pairwise LD module limited to LD correlations to the sentinel variant.

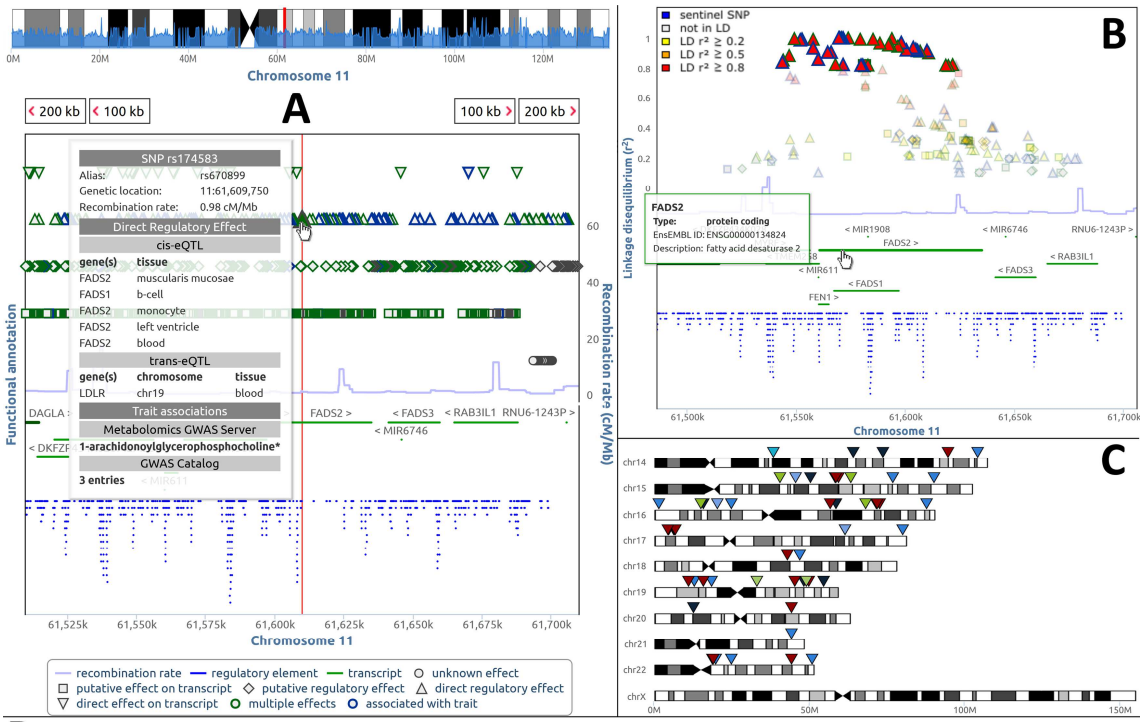
5.2.6 Proxy search and pairwise LD

The Proxy Search module allows for tabular retrieval of variants in LD with input variants (Figure 18D). In addition to the list of correlating variants, dense annotation of the resulting variant set is possible. Another common challenge of association studies is to find out if one locus contains more than one association signal. One possible (albeit not the optimal) approach to do so is to check the pairwise LD statistics of the variants contained in the locus which can be done using the Pairwise LD module. Input is a list of variants that are output with their LD statistics. Again, compressed annotation of the variants is provided as an optional feature.

5.2.7 Data downloads and programmatic access

Results tables from the Block Annotation, Proxy Search and Pairwise LD modules can be downloaded as is, thus providing batch retrieval of LD statistics as well as condensed top-level annotations. These include the set of genes linked to the variants via genomic proximity, eQTL associations, and gene-associated promoter and enhancer elements. For detailed variant

annotations, I have implemented a PDF-conversion of the "*SNiPA* cards" such that the user can collect annotations of variants of interest. The PDFs are interactive, meaning that the links to the data sources are active in the documents and can be used from within the PDF without having to take the indirect route over our server.



D

Proxy	Chr.	Proxy Pos.	LD r²	Proxy annotation	Associated with trait(s)	cis-eQTL	trans-eQTL	Genes hit or close-by	Potentially regulated genes	eQTL genes
rs99780	11	61,596,633	0.94	Direct Regulatory Effect, Putative Regulatory Effect, Putative Transcript Effect, Variation Proximal to Gene	no	yes	no	FADS1, FADS2	-	FADS2
rs1535	11	61,597,972	0.97	Direct Regulatory Effect, Putative Regulatory Effect, Trait Association, Putative Transcript Effect, Variation Proximal to Gene	yes	yes	yes	FADS1, FADS2	SCGB1D1, SCGB1D2, SCGB2A1, FADS2	LDLR, TMEM258, FADS2, FADS1

Figure 18: **Screenshots of *SNiPA* modules**. **A**: The *SNiPA* Variant Browser (query SNP *rs174583* on chromosome 11) shows variants (top), genes (center in green), and regulatory regions (bottom in blue). Variant symbols reflect the most severe effect determined by *SNiPA*. Top-level information is available in mouse-over tooltips for all plot elements as shown here for the query SNP. The example highlights the value of variant-centered accumulation of annotations: *rs174583* is associated with the concentration of a lipid metabolite as well as with the expression levels of two genes encoding enzymes involved in lipid metabolism (*FADS1/2*) and the gene coding for *LDL receptor*, a major regulator of cholesterol homeostasis. Furthermore, the variant has been linked to the response to lipid lowering drugs (statins), levels of *trans*-fatty acids, and the QT interval (information retrieved from the "*SNiPA* card" of the variant). Statins target *HMG-CoA reductase* that, among others, is again regulated by the *LDL receptor*. **B**: Linkage disequilibrium plot for *rs174583* with variants in strong LD ($\mathbf{r^2 \geq 0.8}$) highlighted. As illustrated for *FADS2*, basic annotations for genes and regulatory elements are also accessible by hovering over the elements. **C**: Association maps visualize published GWAS results for one or several traits. Variants can be selected here and further inspected using all other features of *SNiPA* while "*SNiPA* cards" can be directly accessed. **D**: Condensed tabular information for variants in this locus can be retrieved via the Proxy Search and other modules. The tables can be sorted, searched, and filtered online or downloaded for further offline processing.

We also decided to make all data used by the *SNiPA* webserver available for download. Because some datasets in our resource are very large (e.g. compressed LD-data for the latest 1000 genomes release sizes to 613GB), full downloads may be rather unfavorable. However, as we use the free Tabix software to retrieve variant annotations from the compressed and indexed chromosome files and as Tabix enables remote access to such files without the need to download the whole data but only the index files (which are generally very small), this deposition of our data simultaneously provides a programmatic access interface via Tabix.

5.3 Summary

SNiPA combines a comprehensive set of genomic and experimental data to simplify the task of comprehensive variant annotation. I include molecular evidences and mappings from 81 different datasets for annotation of each autosome (chromosome X is annotated to a lesser extent), summing up to more than 1,400 distinct annotation files. Of a total of 78,471,927 SNVs contained in the resource (Table 12), more than two thirds (68%) are annotated in at least one of our five effect categories *direct effect on transcript* (change of protein sequences), *direct effect on transcript regulation* (eQTLs), *putative effect on transcript regulation* (variants in regulatory regions, and TF binding sites), *putative effect on transcript* (change of transcript or gene sequences), and very close (±5Kb) *proximity to a transcript*. Of the annotated variants, more than ten million SNVs (19%) are located in intergenic regions which emphasizes the profits offered by my approach to include an extensive catalog of regulatory information in addition to effect annotations based only on gene, transcript or protein sequences. For more than half of the annotated variants and a substantial fraction of non-coding transcript-associated variants, I was able to predict regulatory effects.

Large genomic resources (e.g. Ensembl, UCSC, NCBI, etc.) aim at providing genome-wide genome-level annotation tracks from an extensive set of resources. This makes retrieving statistical and functional annotation relevant at the single nucleotide level remains difficult. For instance, common genome browsers often display SNVs as thin bars that trail away in the wealth of other annotation tracks and are even less prepared to display statistics such as LD relationships between variants. This limits visual distinction of relevant variants from those

without relevant annotations and leaves the complex task of aggregating position-based data to the researcher. Variant-centered resources, on the other hand, typically concentrate on specific types of data such as amino acid changes (e.g. SIFT [142] or PolyPhen [143]), eQTLs (e.g. seeQTL database [217] or GTEx Browser [211]), or regulatory effect predictions (e.g. RegulomeDB [310]). Moreover, these annotations are often presented in resource-specific data structures. And those variant-centered resources that indeed contain several distinct datatypes mostly lack a graphical representation of the data (e.g. GRASP [311] or GwasDB [312]). For individual inspection of single variants, both resource types are extremely valuable. However, for simultaneous processing of larger variant sets, collection and examination of annotations from different data sources quickly becomes cumbersome. This presents a major bottleneck in genome-wide scans of genetic influences on human traits since the collection of such evidences is the key to understanding the effects of phenotype-linked genetic variants. *SNiPA* on the other hand combines a large collection of evidences with data retrieval interfaces that allow for convenient access to single variant annotations, variant-centered genome browsing, and interactive visualization tools tailored for easy inspection of many variants in their locus context.

All variants	**count**	**percentage**
bi-alleleic variant set (GRCh37)	78,471,927	100.00%
integrated GRCh37/38 variant set	78,181,370	99.63%
intergenic variants	35,089,795	44.88%*
Variants with annotation	52,993,017	67.78%*
Annotated variants		
transcript variants	43,091,575	80.87%
regulatory variants	27,441,207	51.50%
intergenic variants	10,191,999	19.13%
regulatory transcript variants†	8,279,232	15.54%
eQTLs	1,889,879	3.55%
eQTLs in regulatory regions	958,368	1.80%
missense variants	634,460	1.19%
trait-associated variants	181,442	0.34%
trait-associated eQTLs	58,090	0.11%
trait-associated missense variants	16,506	0.03%

Table 12: **Statistics of variant annotations for the *SNiPA* release v3.1**. Given are the variant counts and percentages to the respective reference set for different variant stratification sets. Of the GRCh37 set of bi-allelic 1000 genomes SNVs, 99.6% could be mapped between genome assemblies. Of those, almost 70% could be annotated with an evidence-based effect prediction. The number for intergenic variants, which is lower than for transcript-associated variants, originates from the experimental design of the 1000 genomes project (low-coverage WGS vs. deep WES). eQTL and trait associations are reported as contained in the primary datasets. LD extension of associations leads to a large increase of these numbers (not shown here). * – w.r.t. the integrated variant set; † – excluding missense variants.

To conclude, in this chapter I addressed many of the challenges that are associated with the evidence-based functional annotation of trait-linked genetic variants. I showed how data integration can be used to provide a bioinformatics resource that simplifies the practical challenges posed by the biological interpretation of the results of genetic association studies. To sustain its value, I implemented *SNiPA* to be both automatically updatable and easily extendable by additional datasets. Accordingly, since its launch we have performed several substantial updates of *SNiPA's* data basis both on genome- and variome-wide level and integrated further datasets, such as the latest release of the GTEx portal (V6) covering >17 million new eQTL associations obtained from 44 tissues. With more and more data concerning other –omics layers becoming available, at some point this might enable tracking the translation of genetic variation from layer to layer, following the course of effects leading to phenotypic variance, trait endpoints, and, eventually, personalized medicine.

6 From evidence to biology

In the previous chapters, I have described the process of detecting associations between genetic variants and quantitative as well as binary human traits, how these associations are generally projected to the underlying predicted causal gene, and described a data integration resource that is intended to support this process on a genome-wide scale. However, although we put major efforts into designing *SNiPA* as an intuitively usable web server, it may not be immediately apparent how the genome-wide collection of more and more evidences for variant annotation can be utilized to gain new insights. Even worse, in some cases this accumulation of data leads to another challenge: while missing evidences for annotating genetic variants with potentially affected molecular pathways make it generally impossible to use genetic information for prognosis, diagnosis, and therapy, the opposite, i.e. too much available data, raises the difficulty of being forced to separate relevant from potentially irrelevant information. Furthermore, the large fraction of variants that are annotated as being active regulatory and not directly affecting gene, transcript, or protein sequences contradicts the lessons learned from the field of monogenic diseases.

The purpose of this chapter is therefore threefold: the first part describes the potential of a rather simple integrative analysis of genetic association signals that are shared between complex diseases to define and illuminate the presence and extent of pleiotropy in common disorders. This section is based on our 2012 *BMC Genomics* paper [164]. In the second part, I report a study correlating association signals with a very specific class of regulatory elements, namely miRNA target sites. This section is based on our 2012 *PLoS ONE* publication [163]. In the

final part, which is also focusing on the investigation of regulatory genetic variation, I want to illustrate an example of a genetic association were the available data are so manifold that evidence-based interpretation seems impossible. Using an integrative approach on the large collection of regulatory element data contained in *SNiPA*, I will illustrate the process of deriving a functional hypothesis using conclusive combination of the available evidences. This section is part of a paper in preparation.

6.1 Shared genetic features across complex diseases

As mentioned before, GWAS have provided a large set of genetic loci influencing the risk for many common diseases. Intriguingly, although published individual GWA studiess are usually carried out for one trait at a time, a significant overlap in the associations of several complex diseases becomes apparent [313]. Besides effects on a specific phenotype, loci and single SNPs thus may also exert pleiotropic effects by contributing to a variety of traits. While it is not surprising that susceptibility genes for closely related traits should be shared, multi-functionality of a gene in phenotype presentation, i.e. pleiotropy, *sensu stricto* refers to seemingly unrelated and distinct traits [314]. Loci or variants affecting several traits might have small effects on each specific trait, but may be of major biological interest while indicating shared or branching etiological mechanisms. In principle, the influence of such loci can be agonistic or antagonistic, i.e. involve concurrent similar or opposite effects of the same variant for different traits. So far, few studies attempted to study such loci in a systemic fashion and rather focused on shared risk variants in closely related traits like autoimmune diseases [248, 315, 316], heart diseases [317], or cancer [318]. In order to identify shared or branching pathways of related as well as diverse (i.e. medically and phenotypically distinct) diseases, we performed a systematic comparative analysis of genetic commonalities and differences across traditionally defined traits using the available repository of GWAS results that complements previous work on the topic such as the "Diseasome" concept introduced by GOH and colleagues [319]. For this variant-based approach we manually curated a high-quality data set to construct a network extending the knowledge on genetic overlaps between diseases as provided by GWA studies.

6.1.1 Methods summary

I obtained the core list of candidate sentinel SNPs from the GWAS Catalog. Additional associations were retrieved from HuGE Navigator and manually tested on compliance with the inclusion criteria of the GWAS Catalog before insertion in the candidate list. We then semi-automatically translated trait descriptions to the official terms in the Medical Subject Headings (MeSH). In this process, associations with quantitative and non-disease traits were eliminated.

For the construction of the locus-based data representation, we defined an associated locus as the whole genomic region captured by SNPs in strong LD, $r^2 \geq 0.8$, with the marker originally reported in a GWAS contained in our data set. The locus is then characterized as all genes located within this genomic region (referred to as "gene locus"). If the region contains no genes, the locus is assigned to its chromosomal location (referred to as "intergenic locus"). LD data and gene information were obtained with the SNAP tool. After locus assignment, our final data set consisted of 111 different traits linked via 1,120 SNPs to 508 gene loci and 226 intergenic loci. Removal of isolated traits, i.e. traits which share no associated locus with another trait ($n = 27$), and cutting out loci which are associated with only one trait ($n = 577$), I retrieved a bipartite network ("shared locus network" or SLN) of diseases and loci potentially linked via genetic correlations.

To obtain a variant-based representation of the data, I performed network generation on marker scale by utilizing the set of variants associated with more than one distinct trait. For this, I used the LD data to mutually assign the associated traits of sentinel SNPs in pairwise LD if not already present. In other words, each variant is, in addition to its own associated traits, assigned the traits associated with all correlated SNPs. This set consists of 175 SNPs located in 94 loci and associated with 55 diseases. In the resulting bipartite network (later referred to as "shared variant network"), a trait and a locus are linked if the locus contains a variant which comprises associations with this and at least one other trait. Here, the allele information was included in the graph visualization by coloring of the edges (Figure 19).

DETERMINATION OF AGONISTIC AND ANTAGONISTIC EFFECTS – For all variants associated with more than one trait, I manually extracted the risk alleles (OR > 1, independently of major or minor allele status) and odds ratios from the reporting studies. The alleles were mapped to the forward DNA strand according to dbSNP. The same procedure was applied to markers which were indirectly associated with a trait over LD. If for all traits the same associated risk allele (and correlating allele, respectively) was reported, the SNP was classified as

agonistic. If the risk alleles of a SNP were opposed in the associated diseases, the variant was classified as antagonistic.

GENETIC CLUSTERING – I applied complete-linkage hierarchical clustering to identify groups of traits genetically overlapping with respect to agonistic signals. Normalization was performed using the linear Pearson correlation coefficient (PCC) defined as $\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$ where the input are the vectors of the variant-based agonistic overlap of two distinct diseases X and Y to all other diseases. Thus, disorders which are clustered together show a homogeneous association overlap pattern to all other diseases, while diseases which are not clearly assigned to a cluster present a more heterogeneous pattern relatively unique in the SNP data. For cluster definition, I used a Euclidian distance threshold of 1.71. This threshold was determined as the maximal distance at which the six traits not or only weakly correlating with other diseases (Figure 20, bottom right) remain non-clustered.

CALCULATION OF THE CPMA STATISTIC FOR AUTOIMMUNE LOCI – I downloaded the dataset S1 from [248] and extracted the information on autoimmune-linked SNPs contained in the SVN. I used the Z-scores given in the file to compute two-sided association P_{value} for all seven GWAS. Using the CPMA (the cross-phenotype meta-analysis statistic) code provided on http://www.cotsapaslab.info/index.php/software/cpma/ I calculated the CPMA P_{values} as described by Cotsapas and colleagues.

6.1.2 The shared variant network

To provide a comprehensive base for the analysis of potentially multi-functional loci and variants, respectively, I compiled a network representation of the information made available by GWA studies which we called "shared variant network" (SVN, Figure 19A). Its degree distribution attributes the SVN a scale-free network, i.e. it approximates a power-law ($P(k) \sim k^{-\gamma}$; $\gamma = 1.32$; $R^2 = 0.69$). Interestingly, also when considering the two node types separately, disease nodes ($\gamma = 0.97$; $R^2 = 0.71$) as well as locus nodes ($\gamma = 2.98$; $R^2 = 0.93$) show scale-free degree distributions. The scale-free property classifies the network (and its two sets of node types, respectively) as structured, i.e. non-random [320]. It has to be considered that the limited size of the SVN leads to inaccuracies in distribution fitting and thus reduces the explanatory value of this observation. However, as clinically related diseases (i.e. diseases which present similar symptoms) should present a higher genetic overlap than unrelated disorders, this finding meets expectations.

We then tried to replicate the content of the SVN by comparing it to other variant-based approaches assessing the genetic overlap between traits. Recently, a statistic to identify SNPs with effects across phenotypes (CPMA) was proposed by Cotsapas et al. [248]. It compares the distribution of association P_{values} of a SNP across seven GWAS on distinct autoimmune diseases to the exponential distribution (e^{-x}, i.e. $\lambda = 1$) representing the expected decay rate of association P_{values}. As in our approach we use pre-filtered associations, this method cannot be readily employed on our data. However, using the data provided by Cotsapas and colleagues on autoimmune loci in the SVN, I retrieved CPMA P_{values} on 30 SNPs ($\sim 17\%$) corresponding to 28 loci ($\sim 30\%$) in our data. The CPMA classified all SNPs as significantly effective across diseases ($P < 0.05$). Thus, I was able to validate nearly one third of the loci contained in the SVN by an independent approach, which extrinsically validates at least in part the generation process of the network.

The SVN also shows no artificial character with regards to its topology. Both locus and disease node sets comprise hubs, here defined as nodes with a degree >3, which form the central elements in the network. As in each GWAS multiple markers are associated with a single disease, one would expect hubs to be constituted mostly of disease nodes. In line with that, 74% of the hubs in our network are disease nodes. The remaining 26% are loci hubs (seven gene loci and three intergenic loci). Several of these loci have been previously identified as influencing susceptibility to multiple diseases like the HLA region on chromosome 6 [321], a cancer locus at chromosome 8q24 [318], and a coronary artery disease locus at chromosome 9p21 [317]. Further hub loci are *PTPN22*, a known player across several autoimmunity disorders [322], and *IL23R*, which has been shown to direct inflammatory processes [323]. In addition, we observed hubs which have not yet been described as predisposing to a whole group of diseases, such as *TNPO3* which appears to predispose to various autoimmune diseases like systemic lupus erythematosus, systemic scleroderma, and rheumatoid arthritis [324-326], or *TNFSF15*, which shows associations with several inflammatory diseases [327-330]. As expected, in the majority of cases the traits linked to one hub can be assigned to the same disease group and, further, diseases which are not obviously related to other disorders linked to the respective hub are mostly associated with antagonistic signals. For instance, in a four-gene locus at chromosome 17q12 (*GSDML|IKZF3|ORMDL3|ZPBP2*), four autoimmune diseases are associated with the same risk allele that in turn has opposite effects on asthma [161, 326, 327, 331]. Thus, our results indicate that loci associated with several diseases have an effect specific to a certain disease group

rather than effects on unrelated diseases, and that, if there is an effect on an unrelated disease, it can often be distinguished by the direction of the effect.

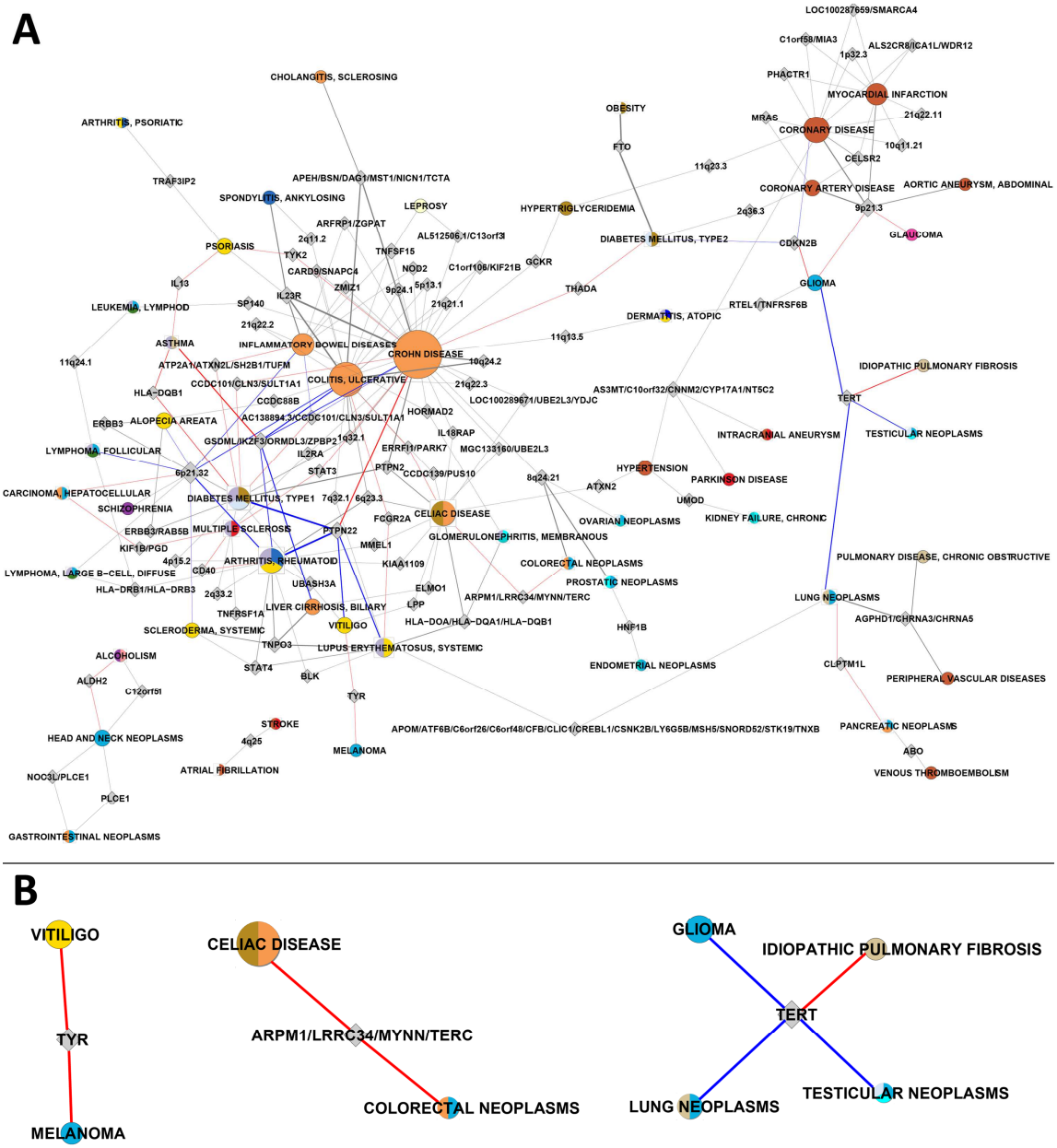


Figure 19: **A**: **The shared variant network** (**SVN**). A trait and a locus are linked if the locus contains a variant showing association with this and at least one other trait. The network consists of 175 SNPs located in 94 loci that are associated with 55 diseases. The colors of the disease nodes correspond to disease classes according to the MeSH ontology, multi-colored nodes indicate an association with different disease classes. Loci are depicted as transparent, diamond-shaped nodes. The node size reflects the number of loci a disease is associated with. The edge color reflects the allelic information: gray indicates agonistic variant(s), red corresponds to antagonistic variant(s), and blue marks both agonistic and antagonistic signals. **B**: **Examples of antagonistic loci**. For clearer accessibility, examples discussed in the text have been extracted from the whole network. Figure and caption adapted from [164].

6.1.3 Estimation of errors induced by naïve variant-to-gene projections

In the previous chapters, I have already indicated the difficulties in assigning predicted causal genes to trait-associated genetic variants, especially when only few evidences are available. Using established network property measures on the SLN and the SVN, I tried to quantify the amount of information loss and of misleading variant/gene assignments. Despite the size difference, the SLN shows greater network heterogeneity ($SLN = 1.30$, $SVN = 1.17$) and lower centralization ($SLN = 0.175$, $SVN = 0.205$) and density ($SLN = 0.014$, $SVN = 0.021$) values than the SVN. Furthermore, the intersection between the SVN and the SLN lacks not only 5% of the nodes but also 10% of the edges of the SVN. These numbers imply that the process of translating LD data into locus information is at least partly inconsistent. Analysis of the structure of the assigned LD blocks showed two error sources in shared locus analysis. First, variants in two independent LD blocks are assigned the same locus but are not in LD. Thus, shared loci are found that are not reflected in the variant based data. Second, if two SNPs are in strong LD but the individual LD patterns of the SNPs diverge (e.g. the LD block of one SNP covers a greater area at the given r^2-threshold), a second type of assignment error occurs. In this case the two SNPs are assigned to different loci (in the example above, this is due to the different sizes of their LD blocks which may contain distinct gene sets) and their LD connection is lost. These observations suggest that *i)* the SLN contains loci which overlap between traits but the associated markers are not in strong LD, *ii)* there are several traits which are connected to the SLN via a single, potentially misleading link (as not mirrored in the variant-based data), and *iii)* a locus assignment approach using only LD data is unable to identify all shared associations ($n = 25$ or 27% of unidentified loci, based on the second type of assignment error). This limited sensitivity and specificity in detecting LD-based correlations between the reported markers on locus scale shows the uncertainty in naïve automated variant/gene projections, which in this study prompted us to use the variant-based SVN for all further analyses.

6.1.4 Genetic correlations identify prevalence of frequent comorbidities

To identify shared and branching mechanisms I split the SNP association data into agonistic and antagonistic variants. Since in most cases there is no solid and comprehensive basis of experimental data that would allow for a more sensitive classification, we suggest that the best available indication of distinct effects of a variant on two diseases is the signal itself being different. Therefore, we define a SNP to be agonistic if all disorders are associated with the same risk allele of the SNP. Conversely, we consider a SNP antagonistic if the associated risk alleles

differ between diseases. Accordingly, in the analysis of genetic overlaps and correlations as a measure of trait similarity, only agonistic variants were included.

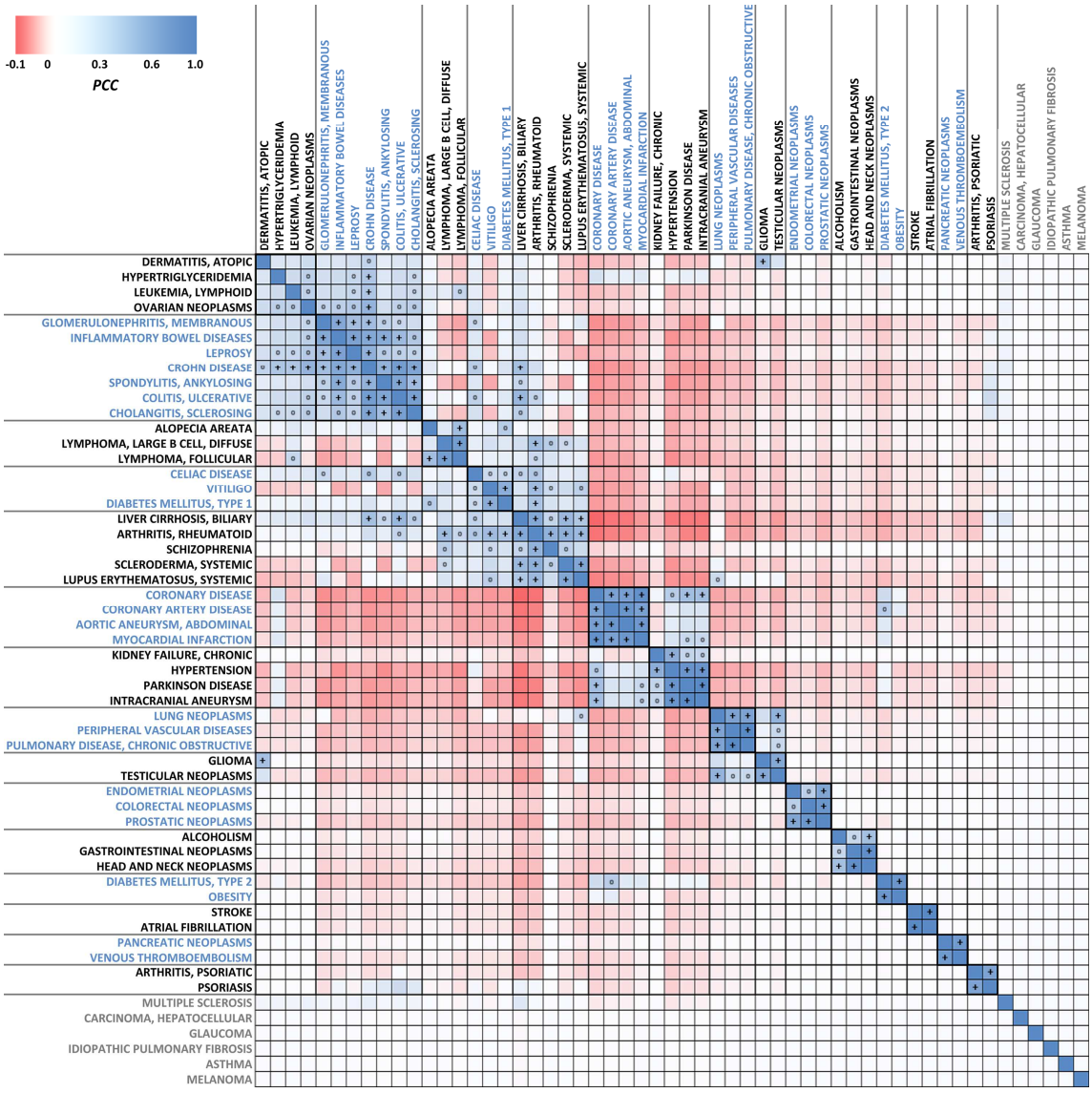


Figure 20: **Clustering of diseases based on genetic correlations**. We applied complete-linkage hierarchical clustering to identify groups of diseases that show homogeneous patterns of genetic overlap to other disorders. Disease clusters are illustrated by boxes and alternating font colors. The distance threshold for cluster definition was defined as the maximum distance at which the six diseases showing no or only very weak genetic correlations to any other disease remain unclustered (disease names in grey). Positive correlations that are significant are marked by plus signes for Bonferroni-corrected significance and by open circles for significance at an uncorrected $\boldsymbol{P_{value} < 0.05}$.

As similar diseases are more likely to share associations than diseases in distinct classes, we expected the SVN to be organized in a modular fashion. This was confirmed by the decrease of the degree distribution of the topological coefficient with the number of links per node. To retrieve these modules, we applied a hierarchical clustering approach. The SVN contains two

node types (loci and traits). As we wanted to directly assess variant-based disease relatedness, I performed a disease-centric projection of the adjacency matrix of the SVN for hierarchical correlation clustering. Genetic correlations were obtained by calculating the PCC for all pairs of diseases based on their genetic agonistic overlap with all other diseases. The clustering returned 15 disease clusters (Figure 20) and six diseases which show no or only weak correlation with any other disease. With the exception of one heterogeneous cluster (hypertriglyceridemia, ovarian neoplasms, lymphoid leukemia, and atopic dermatitis), the clusters mostly contain related diseases. Although some clusters also contain single traits unrelated to the other phenotypes, such as schizophrenia which is clustered together with four autoimmune diseases, this indicates that clinical disease classifications appear in general to be reflected on the genetic level. Notably, several small clusters contain diseases which are also linked through common environmental risk factors – like smoking for lung neoplasms, peripheral vascular diseases, and chronic obstructive pulmonary disease – or present high frequencies of comorbidity, e.g. type 2 diabetes and obesity. To get an insight into the overall extent of reported comorbidities of the diseases within the 15 clusters, I used publicly available resources [332, 333] and literature mining. The within-cluster fraction of disease co-occurrence ranged from 75% to 100% ($\mu = 95.89\%$, $\sigma = 8.66\%$) which provides empirical evidence of the epidemiological interrelation of diseases clustered together by genetic information. Clusters containing diseases that present high ratios of comorbidity may thus indicate potential artifacts due to "contaminated" disease cohorts including a substantial number of comorbid cases. The unbiased search for associations of genetic markers to a disease phenotype as performed in GWAS does not distinguish between markers for a primary or related secondary (comorbid) disease. For instance, in our approach pancreatic neoplasms and venous thromboembolism (VTE) are clustered together. While there is no direct molecular connection known between cancer and VTE, with incidence reports ranging from 17% to 57% pancreatic cancer is one of the cancer types most strongly associated with VTE complications [334-336]. Although the presence of independently shared etiological mechanisms can naturally not be ruled out in general, these results suggest that the potential of frequent comorbidities leading to spurious associations may have been underestimated in some studies.

6.1.5 Antagonistic markers suggest pleiotropic effects

We next searched for evidence that antagonistic signals represent genetic indicators of branching points in the etiologies of two diseases or disease groups. For the assessment of potentially multifunctional variants we therefore focused on markers with inverse effects. I

identified 44 such variants, which represent almost 4% of the original association data analyzed and about 25% of the SNPs associated with more than one disease. Of those 44 variants, about one fifth ($n = 9$) are located in the HLA region. SNP-markers in that region are known to differ in their ability to capture the classical HLA-alleles [337] and therefore were not considered further in the present analysis.

For cases where the function of the harboring genes is known, I was able to identify conclusive models (Figure 19B). For instance, *rs2736100* in the *telomerase reverse transcriptase* (*TERT*) gene was reported to exert antagonistic effects in idiopathic pulmonary fibrosis (IPF) and testicular germ cell tumor (TGCT) and two other cancer traits [338-344]. Whereas telomerase activity is generally upregulated in tumors sustaining proliferation and potentiating mutagenesis and transformation of cancer cells [345], in IPF limited cell division due to decreased telomerase activity is thought to contribute to the phenomenon of high percentages of apoptotic cells in fibroblasts [346]. Consistent with that observation, disturbed telomerase activity in TGCT is believed to form a distinct mechanism of cancerogenesis in this tumor type [344]. This distinction from other cancer traits is believed to be based on the fact that testicular germ cells are the only adult cell type with high telomerase expression [347]. Another example is the *telomerase RNA component* (*TERC*), which is essential for *TERT* functioning. Opposite alleles of SNP *rs10936599* are associated with celiac disease (CeD) and colorectal cancer (CRC) [348, 349]. Jones et al. showed that *rs2293607*, a variant tagged by *rs10936599* ($r^2 = 0.99$), alone is sufficient to modulate *TERC* expression [350]. While in CRC this leads to *TERC* overexpression and longer telomeres, the opposite might apply to CeD, which exhibits telomere reduction and genomic instability [350, 351]. The observation that both constituents of the telomerase complex contain independent antagonistic variants is an intriguing finding. It suggests parallel, autonomous evolution of two functionally interacting loci gone to fixation at a trade-off between early cell senescence or increased apoptosis rates (as in IPF and CeD) and oncogenesis.

A further example is *rs1393350* in the *tyrosinase* (*TYR*) gene where the opposite alleles are linked to vitiligo and melanoma [352, 353], potentially mirroring the inverse correlation observed for the two traits. The phenomenon is based on the presentation of *TYR* (self-) antigens on the cell surface of melanocytes. It is hypothesized that in vitiligo the immune system is hypersensitive towards *TYR* antigens, which are overexpressed in melanoma cells [354]. A possible explanation is that opposite alleles differentially influence the antigenicity of the *TYR* protein, possibly via the strongly ($r^2 = 0.95$) correlated *TYR* missense variant *rs1126809*,

thereby conferring protection from melanoma but susceptibility to vitiligo through immune surveillance and vice versa.

6.1.6 Concluding remarks

In this study, we identified overlapping genetic associations and their corresponding loci with analogous or contrasting effects on different diseases. Associations formally implicate genomic regions which are captured via tagging SNPs representing haplotype blocks. By using the population-specific LD-based haplotype data provided by the HapMap project or the 1000 genomes project, SNP arrays are constructed aiming at a high coverage of the total genome variation, but without considering biologically functional aspects. The advantage of GWAS as a method is its unbiased approach to identify genomic regions compromised in a disease; a major drawback is that the association of markers without knowledge of the causal variants and their effects does not allow for a straightforward biological interpretation.

As I show, the reliability of an automated assignment of LD-based loci to the trait-associated variants is strongly context-dependent. Especially in cases of high gene density or, conversely, in intergenic regions/gene deserts, assigning predicted causal genes to GWAS signals is not possible without further evidence. Simplifications such as even more basic locus assignment approaches which neglect the LD structure of the genome (e.g. classifying a SNP as affecting only the most proximal gene) may seem more intuitive, might facilitate analyses and could be useful to identify causal disease-gene associations. These correct associations of genes which are detected through significant enrichment of a harbored tagging variant in a patient cohort may not be discovered when incorporating LD data in cases where the LD block of the respective variant spans across several genes. However, such approaches disregard a basic principle defining the GWAS paradigm, namely the use of LD information in the design of genotyping arrays to achieve the genome-wide coverage of common variants.

Accordingly, we decided to use variant-based methods and concentrated on strong gene candidates identified via the gene function of single-gene loci whenever suggesting potential biological effects of the considered variants. In the analysis of genetic overlaps we followed the hypothesis that the effects of variants shared across several diseases correspond to the reported risk alleles. If the risk allele is the same in all associated diseases, we assume the effect to be the same, i.e. that there is a common underlying etiology. For closely related diseases a positive correlation is not surprising. For instance, a GWAS on psoriatic arthritis (PSA) will also detect agonistic variants such as *rs33980500* that are also associated with psoriasis (PS) [355, 356], leading to a highly significant ($\varrho = 0.765,\ P_{value} = 1.1 \cdot 10^{-11}$) genetic correlation between the

two diseases in our data. Indeed, the vast majority of agonistic variants in our data set links groups of related diseases and thus may mark interesting target regions for closer investigation. However, I also found a few agonistic signals connecting apparently unrelated diseases, e.g. *rs6010620* which exerts susceptibility for both glioma and atopic dermatitis [342, 343, 357, 358]. If our hypothesis is correct, an endophenotype influencing both diseases may be present which has yet to be identified. For antagonistic SNPs, on the other hand, we describe plausible mechanisms that may render variants protective against one trait and predisposing to another, labeling the affected genes/loci as pleiotropic. If pleiotropic effects are as frequent as evolutionary modelers postulate [359, 360] and these effects can be identified by analyses based on GWAS, this could point out interesting implications for the development and use of therapeutics because it would enable avoidance of potential side effects when targeting such loci. Already, there are several genotype/drug interactions known for which therapeutic dosing recommendations are available [361].

IBD	2nd disease	ϱ	P_{value}
Crohn's disease	sclerosing cholangitis	0.75	5.1E-11
Crohn's disease	leprosy	0.65	9.1E-08
Crohn's disease	ankylosing spondylitis	0.64	1.3E-07
Crohn's disease	ovarian neoplasms	0.63	3.2E-07
Crohn's disease	membranous glomerulonephritis	0.54	2.1E-05
Crohn's disease	biliary liver cirrhosis	0.51	6.6E-05
Crohn's disease	hypertriglyceridemia	0.46	3.8E-04
Crohn's disease	lymphoid leukemia	0.46	3.8E-04
inflammatory bowel diseases	leprosy	0.77	5.3E-12
inflammatory bowel diseases	membranous glomerulonephritis	0.73	3.4E-10
inflammatory bowel diseases	ankylosing spondylitis	0.70	3.2E-09
ulcerative colitis	sclerosing cholangitis	0.80	2.6E-13
ulcerative colitis	ankylosing spondylitis	0.75	5.6E-11
ulcerative colitis	biliary liver cirrhosis	0.59	2.0E-06
Crohn's disease	*ulcerative colitis*	*0.83*	*3.1E-15*
Crohn's disease	*inflammatory bowel diseases*	*0.72*	*4.5E-10*
ulcerative colitis	*inflammatory bowel diseases*	*0.57*	*7.0E-06*

Table 13: **Genetic correlations of IBDs to other diseases**. Shown are IBDs and diseases linked via significant *(P<0.0009)* genetic correlations in our analysis. On the bottom in italic, the genetic correlations between IBD types contained in our study are given. P_{values} are unadjusted. ϱ – correlation coefficient.

The observation that a surprisingly high fraction (>15%) of the SNPs considered in our study are associated both agonistically and antagonistically with related as well as unrelated disorders indicates that the molecular mechanisms influencing causes and progress of human diseases may in part be interrelated. Genetic overlaps between two diseases also suggest the

importance of the affected entities in the specific pathogenic pathways. Although these may be secondary, such as genes involved in inflammatory responses related to T2D as well as cancer [362, 363], they should nonetheless be investigated further. The findings presented however also demonstrate the need to clarify the relations of phenotypes linked to agonistically associated markers. For directly interrelated diseases such as PS and PSA, often PS patients without present arthritis or arthritis in the past are used as additional control group. Associations are then interpreted as PSA-specific if not as strongly associated with PS [364, 365]. Comparable procedures or direct adjustment of regression models in association testing may proof useful in frequently co-occurring diseases genetically linked by agonistic variants. For instance, inflammatory bowel diseases (IBDs) are associated with several frequently occurring extra-intestinal manifestations [366, 367]. Consistent with the epidemiological data, I found highly significant genetic correlations between IBDs and seven other diseases as well as with leprosy which is caused by infections with *Mycobacterium leprae* or *M. lepromatosis* (Table 13). The connections between the diseases, although in part assumed to be influenced by shared genetic factors, remain largely unknown [367], while infections with mycobacteria have been proposed as a cause of IBD [368]. Using our results for careful case stratification and inclusion of suitable covariates in the genetic analyses may provide deeper insights into these relationships.

Pleiotropic genetic effects, on the other hand, that are harbored in the same locus may trigger different mechanisms interfering with the genetic or environmental background. The detailed examination of antagonistically associated loci may thus lead to first insights into the mechanism of the various types of pleiotropy implicated in complex human diseases.

6.2 *Cis*-acting polymorphisms: miRNAs as disease mediators

In recent years, more and more evidence is emerging that microRNAs, a class of small non-coding RNAs, play an important role in the development of human traits. Databases collecting information on miRNAs mediating human disease such as miR2Disease or PhenomiR list several hundred miRNAs with established roles in way above 100 human diseases. miRNAs are key posttranscriptional regulators of most known cellular processes and have been associated with cell fate decision, development, and stress response. Additionally, miRNAs have been

identified to be usable as biomarkers for human diseases [369–372]. With growing knowledge on their targets, which are believed to make up more than 60% of all protein-coding genes [373], new regulatory and disease-mediating gene networks were discovered [374–377]. Because single miRNAs are able to regulate not only one but up to several hundred genes, they depict promising therapeutic targets for disease pathways involving multiple genes. With the advances of the crosslinking immunoprecipitation (CLIP) technology, it has become feasible to experimentally determine miRNA-target interactions and the exact binding sites of the RNA-binding proteins (RBPs) on transcriptome scale [378–381].

In order to investigate the role of miRNA functioning in human health more closely, approaches with the objective to identify potential interrelations of miRNA dysregulation and genetic variation were attempted. In these earlier studies, however, neither the data on trait-associated polymorphisms nor experimentally verified miRNA targeting information provided a sufficient basis for genome-wide integrative analyses of both entities. Genetic variants located in 3'-untranslated regions (3'-UTR) of human transcripts, the major target of miRNA-mediated regulation, were rarely reported in rationales of genetic association studies, as functional interpretation of such non-coding variants was challenging without available evidence for functional implications (such as miRNA target sites). Large-scale high-quality experimental data validating miRNA target site predictions, however, became available only in 2010. Consequently, at the time this study was conducted, only few particular examples of SNPs affecting miRNA regulation pathways had been identified [382, 383], and studies were mostly limited to effects on predicted miRNA target sites [383, 384].

It is important to mention that the 3'-UTR harbors several other functional elements besides miRNA target sites which may, if affected, also mediate disruption of miRNA regulation pathways. It has been assumed, for instance, that the loss of a polyadenylation (polyA) signal can cause genetic diseases by non-specific degradation of the mRNA [385]. Recent experiments suggest that this effect may be based on a functional connection between polyA signal efficiency and miRNA-mediated translational repression [386]. Further, the structural accessibility of an RNA region is an important feature for the binding affinity of RNA-induced silencing complex (RISC) target sites [387]. It has been shown that mutations in RNAs have large local as well as global structural effects [388] and that altered target accessibility can reduce miRNA-mediated posttranscriptional repression to a scale comparable to that of mutations disrupting miRNA recognition element (MRE) sequence complementarity [389]. Finally, polymorphisms affecting splice sites can lead to radical sequence changes increasing susceptibility

to diseases, an effect which is suspected to be partly due to altered translation efficiency of the affected mRNA [390] - which is characteristic for miRNA functioning.

The success of GWAS in identifying genetic loci involved in the susceptibility to common diseases has been repeatedly discussed in this thesis. I have also introduced the challenges associated with the prediction of causal genes for GWAS loci as well as with deriving hypotheses regarding the functional mechanisms underlying observed association signals. In this study, we concentrate on the influence of GWAS-identified variants on miRNA-mediated *cis*-acting regulatory effects as an example of predicted regulatory genetic variation. More specifically, we assess potential posttranscriptional effects exerted by trait-associated variants by systematically investigating SNPs located within the 3'-UTR of human transcripts for interference with polyA signals, 3'-UTR splicing, 3'-UTR secondary structure changes and MREs.

6.2.1 Methods summary

SNP DATA SETS – The core list of sentinel trait-associated SNPs (referred to as GWAS-SNPs) was again obtained from the GWAS Catalog, which at the time of this study included information about 5,101 unique SNP-trait associations with a $P_{value} < 1.0 \cdot 10^{-5}$. The GWAS-SNP set was extended by highly correlating SNPs in strong LD $r^2 \geq 0.8$ in the HapMap3 CEU panel. This set, further referred to as extended GWAS-SNPs, contains 18,884 variants retrieved using the SNAP tool. As background distribution for localization enrichment I used the complete set of 2.79 million SNPs from the CEU panel of the joint HapMap Phases I, II and III (release 27, referred to as HapMap-SNPs) for which genotype information was available. For background distributions of variant properties (see below), I randomly selected 5,101 SNPs from the HapMap-SNP set a 1,000 times and extended these set, analogous as for the GWAS-SNPs, with SNPs in strong LD.

ANNOTATION OF GENOMIC VARIANT PROPERTIES – I mapped all HapMap-SNPs on genomic locations of protein-coding genes and miRNA hairpin sequences from miRBase (release 18). The localization of SNPs was then categorized into five classes: intergenic, intronic, 5'-UTR, coding sequence (CDS), and 3'-UTR. For stratification of variants according to MAF and LD (r^2), SNPs were binned in 10% intervals for MAF and 5% intervals for LD. Genome-wide LD-based SNP binning was performed using an all-vs.-all $r^2 \geq 0.8$. The localization of the LD bins was defined as the localizations of the SNPs contained in the respective bin. Variant conservation was determined using phastCons (see Table 3) and considered as conserved sites at values greater than 0.57 or not conserved else [391].

ANNOTATION OF 3'-UTR VARIANTS – We included RISC target sites for the RBPs Argonaute and TNRC6 as provided by the starBase database that mapped to protein-coding genes (139,254 target sites in 24,442 transcripts; 48% of sites were located within a 3'-UTR). MREs were defined as complementary sites for canonical miRNA seed sequences [208] within 21 nucleotides to the center of a RISC target site in 3'-UTRs for both alleles of 3'-UTR SNPs. Additional filtering for miRNAs with MREs significantly overrepresented in RISC target sites ($LOR > 0$, $P_{\chi^2} < 0.05$) retrieved the final set of 258 and 324 miRNAs with affected and enhanced mRNA:miRNA hybrid formation, respectively. For examination of polyA signals, I integrated the PolyA DB for mRNA polyadenylation sites. The position of polyA sites is described to be located 10-30 bases downstream of the polyA signals [203, 204]. Therefore, I determined SNPs within this range, extracted 11 nucleotides long mRNA sequences centered around 3'-UTR SNPs, and examined the sequences for the most abundant polyA signal variants according to [204]. A SNP was classified as affecting a polyA signal if its non-reference allele either created a new polyA signal sequence or eliminated an existing signal. Synonymous variants (i.e. substitution of one polyA signal by another) were not considered as damaging. Changes to the splice site structure of mRNA products (loss/gain and increase/decrease of likelihood at cut-off 0.5) were predicted using the NNSplice algorithm [249]. RNA structural changes caused by SNPs were predicted with the RNAfold algorithm from the Vienna RNA Package. For this, using the 1,000 randomly created SNP sets we empirically determined a correlation coefficient of 0.55 between the structure predictions of reference to non-reference alleles as having a probability of less than 5% for a type I error and only assumed an effect for SNPs showing values above this threshold.

6.2.2 Trait-associated variants are significantly enriched in 3'-UTRs

I compared the amount of trait-associated variants within the predefined five localization categories of SNPs (intergenic, intronic, 5'-UTR, CDS, and 3'-UTR) to examine a potential location trend of these markers. Of 18,884 SNPs contained in the extended GWAS-SNP set, I found 436 to be located in the 3'-UTR of 326 human genes ($OR = 2.331$, $P < 10^{-52}$; enrichment of trait-associated SNPs located in human 3'UTRs vs. the background of all variants contained in HapMap). This is a higher enrichment than for sentinel SNPs only ($OR = 2.059$, $P < 10^{-10}$). Using 1,000 random subsets of HapMap-SNPs of comparable size and properties confirmed significance of the enrichment ($P < 1.1 \cdot 10^{-7}$). I further characterized dependencies between this enrichment and the MAF and the LD-based marker extension by stratification of SNP sets according to both measures. Stratification for r^2 in the extension of

GWAS-SNPs showed that the distribution of ORs locally stabilizes around a threshold of $r^2 = 0.8$ (Figure 21G). In order to identify a potential bias induced by correlating SNPs in the same 3'-UTRs (SNP-gene ratio was ~1.34) I binned the complete HapMap-SNPs into blocks with an all-vs.-all $r^2 \geq 0.8$. More than one million HapMap SNPs were binned together in about 371,000 blocks containing more than two SNPs. The remaining SNPs only showed pairwise or no LD at the chosen threshold. When I included all SNPs after binning, the OR for 3'-UTR enrichment was even more significant than without binning ($OR = 4.27$, $CI_{95\%} = [3.84 - 4.74]$, $p < 10^{-114}$). This eliminates LD-based extension of association signals as a source of bias. An interesting finding is that the size of LD blocks depends on the location of the SNPs. While intronic and intergenic SNPs are reduced to less than 35% (block-SNP ratio) by binning, SNPs in exonic regions present less extensive LD patterns (reduction only to about 81%). In total, binning reduced the HapMap SNP set to about 2.04 million tag regions translating to a genome-wide significance threshold for association testing of $2.45 \cdot 10^{-8}$. However, as an all-vs.-all $r^2 \geq 0.8$ is a very strict binning criterion, this is an implicit validation of the globally used threshold of $5.0 \cdot 10^{-8}$ for genome-wide significance.

When stratifying variants according to their MAF, I found the extended GWAS-SNPs to hold a commonly higher MAF than HapMap-SNPs in general, regardless of their assigned localization category. However, 3'-UTR SNP enrichment values remain highly significant for all MAF intervals (at least for the limited MAF spectrum accessible using HapMap and GWAS data) with comparable ORs, indicating that the MAF has negligible impact on the presented results (Figure 21H).

6.2.3 Evidence for impact on miRNA-mediated regulation

The efficacy of a miRNA to control target mRNA translation relies, among others, on three sequence-based features: correct mRNA processing, presence of a functional MRE, and accessibility of the RISC binding site. To find out to which extent trait-associated SNPs in 3'-UTRs affect miRNA functioning, we examined four mechanisms potentially compromising these features (Figure 21). This analysis was limited to transcripts featuring both 3'-UTR SNPs and validated RISC target sites. The according data set contained 288 SNPs on 409 transcripts and 219 genes, respectively.

First, I investigated allele-specific effects of SNPs on mRNA processing by interfering with polyA signals which yielded four SNPs affecting hexamers with a sequence characteristic for polyA signals. However, none of these hexamers were located near a validated polyA site and, thus, a functional effect of those variants on mRNA processing seems unlikely.

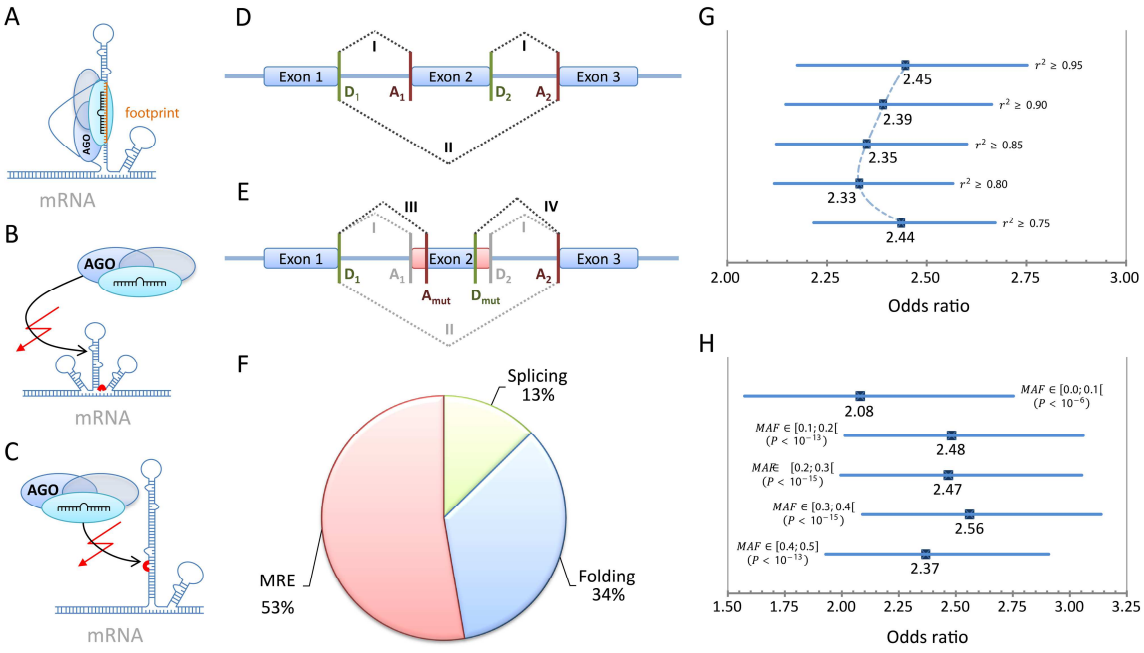


Figure 21: **Mechanisms of 3'-UTR SNPs affecting miRNA targeting, and impact of stratification analysis on 3'-UTR enrichment**. **A**: A microRNA-directed RISC regularly binds to the mRNA. The RISC binding region within the mRNA, the so-called footprint, is colored orange. **B**: Binding of the RISC and, thus, miRNA-mediated silencing is inhibited by a change in RNA secondary structure. **C**: A mutation within the MRE seed site disrupts the ability of a certain miRNA to target a transcript. Here, the opposite effect can occur, i.e. a new MRE seed site is formed by a polymorphism which enables targeting by a miRNA usually not controlling the respective transcript. **D** and **E**: Altered splicing by acceptor or donor splice site gain. The normally existing splice variants (I and II) are extended by additional splice variants introduced by a variant: (III) A present acceptor site (A_1) is substituted by a new acceptor site (A_{mut}), and (IV) a naturally occurring donor site (D_2) is replaced by a new donor site (D_{mut}). Both effects may lead to a considerable loss of exon sequence (displayed in red) and, thus, RISC binding sites. **F**: The percentages of classified SNPs mediating the single mechanisms. The greatest amount of functionally annotated 3'-UTR SNPs directly affect MRE sequences, followed by SNPs changing the RNA secondary structure and SNPs with an predicted effect on 3'-UTR splicing. **G**: SNP enrichment in the 3'-UTR in dependency of different LD thresholds. Displayed are the ORs and confidence intervals for five cut-offs. Accumulative 3'-UTR SNP sets were calculated. The fitted distribution (dashed line) points out the stabilization of the OR around a threshold of 0.8. **H**: SNP enrichment in the 3'-UTR in dependency of the minor allele frequency. Displayed are the ORs and confidence intervals for the 5 different MAF bins. SNP counts were compared within the respective bins. Figure and caption adapted from [163].

Second, we analyzed the predicted impact of 3'-UTR variants on mRNA splicing and identified seven SNPs (~2.4%) predicted to interfere with RNA splice sites (Figure 21D and E). Six of those are predicted to create new acceptor sites and one to create a new donor site. The probability to observe such an effect by chance was $P = 1.78 \cdot 10^{-2}$ for acceptor sites and $P = 1.41 \cdot 10^{-2}$ for donor sites. In all seven cases, the predicted gain of splice sites results in exon shortening, leading to a noticeable loss (46% on average) of RISC binding sites on the respective

transcript sequences. SNPs interfering with splice sites located at an exon/intron or intron/exon border were not found.

Third, we searched for SNPs which may affect the secondary structure of the 3'-UTR proximal to a validated RISC binding site causing an altered accessibility of the region. This resulted in 14 SNPs (~4.9%) predicted to affect the binding affinity of the RISC through changed secondary structure of the 3'-UTR (Figure 21B).

Finally, we examined direct effects of SNPs on MREs located in validated RISC target sites. We found 22 SNPs (~7.6%) disrupting MREs, and 28 SNPs (~9.7%) creating new MREs (Figure 21C). The overlap between the SNP sets creating and disrupting MREs, i.e. SNPs substituting the MRE of one miRNA by a MRE of another miRNA, amounts to 13 variants. Accordingly, a total of 37 unique SNPs (~12.8%) directly affect MREs. The probability of obtaining these amounts of SNPs affecting MREs randomly was $P = 1.27 \cdot 10^{-2}$ (disruption) and $P = 8.76 \cdot 10^{-4}$ (creation). Additionally, we found that only 11% of SNPs enhancing (i.e. extending an already existing seed match) or creating a MRE were conserved across mammals which was a lower fraction than for SNPs causing one of the other effects (folding=29%, splicing=29%, MRE disruption=27%).

6.2.4 Replication of results using 1000 genomes variants as background

The results described above have been compiled on the variant set of HapMap and in the meantime, the much denser backbone provided by the 1000 genomes project became available. It can be hypothesized that the limited variant background set as given in HapMap may, in contrast to the MAF and the LD-extension approach which have been excluded as sources of potential biases, have large effects on the study outcome. As mentioned before, the *SNiPA* resource (here I used version 3.1) also contains an updated set of StarBase miRNA target sites investigated in the initial study and, therefore, I attempted to replicate our results on the larger 1000 genomes project background set. As a complete revision of the results would include computationally very expensive calculations, this analysis was limited to very basic analyses which, nonetheless, underline the validity of the presented findings. The analysis includes enrichment analysis of trait-associated variants from the GWAS Catalog in the 3'-UTR and in miRNA target sites, extension to all variant-trait associations contained in *SNiPA* (including non-significant findings) as well as enrichment of eQTL markers in miRNA target sites. LD-extension of GWAS sentinel SNPs or eQTL associations was omitted here, as was LD-based genome-wide binning of variants.

The analysis yielded almost identical results to the approach using HapMap data. Enrichment of GWAS sentinels in the 3'-UTR was again highly significant ($OR = 2.15, CI_{95\%} = [1.96 - 2.35], P = 2.3 \cdot 10^{-50}$) and variants within miRNA target sites were two times as likely to be associated with a trait in the GWAS catalog than other variants ($OR = 2.06, CI_{95\%} = [1.67 - 2.50], P = 9.7 \cdot 10^{-11}$). Even when including non-significant trait associations (e.g. from dbGaP and the metabolomics GWAS server), the enrichment for trait-associated variants affecting miRNA target sites holds significance ($OR = 1.41, CI_{95\%} = [1.31 - 1.51], P = 7.2 \cdot 10^{-19}$). When intersecting the large catalog of eQTL associations contained in *SNiPA* with variants located in miRNA target sites, again a highly significant overrepresentation became apparent ($OR = 2.16, CI_{95\%} = [2.08 - 2.16], P < 4.9 \cdot 10^{-324}$). As RISC targeting is thought to affect mRNA expression levels via miRNA-mediated transcript degradation, this finding confirms expectations. In summary, this shows that the findings obtained using the HapMap variant set can be transferred to the 1000 genomes variant set which further emphasizes the importance of the reported findings: the incorporated 1000 genomes release contains >99% of all SNPs with MAFs >1% for five super-populations [231].

6.2.5 Models of allele-specific miRNA-mediated metabolic control

Analysis of *cis*-acting regulatory genetic variants is a difficult task. For instance, as useful as eQTL associations are in the interpretation of potential effects exerted by genetic variants, the information provided are again only associations. This means that, although it is known that the expression of a transcript is linked to a genetic variant, the mechanism behind this association, such as affected TF binding sites, have to be obtained in an additional data integration step (this is further discussed in section 6.3). Combining eQTL data with affected miRNA target sites, on the other hand, directly gives a hint on the underlying mechanism by which the transcript's expression levels may be affected.

In the original study, overall I found lipid concentration traits to be enriched in 3'-UTR SNPs ($P < 1.3 \cdot 10^{-3}$) which agrees with the established involvement of miRNAs in the regulation of lipid metabolism [392]. In line with that, the 1000 genomes analysis also yielded several cases of putative miRNA involvement in lipid homeostasis. For instance, the non-reference C-allele of the 3'-UTR variant *rs13702*, located in the *lipoprotein lipase (LPL)* gene and associated with levels of HDL cholesterol and triglycerides [393-397], was found to disrupt a RISC binding site containing MREs for the miRNAs *miR-495* and *miR-410*. Checking the *SNiPA* variant annotation of *rs13702* additionally showed that the variant is associated with *LPL* expression levels in human blood and monocytes [214, 215], as well as that the *LPL* gene is

associated with hyperlipidemia and *lipoprotein lipase* deficiency (MIMs: 144250, 238600; OrphaNet: 70470, 309015). Combination of these evidences suggests that *rs13702* may be the functional variant within its LD-block that, by disrupting the MREs for the two miRNAs, alters gene expression of *LPL*. According to UniProt (entry P06858), the *"PRIMARY FUNCTION OF THIS LIPASE IS THE HYDROLYSIS OF TRIGLYCERIDES OF CIRCULATING CHYLOMICRONS AND VERY LOW DENSITY LIPOPROTEINS"* [225], which thus has a direct enzymatic link to the phenotypes associated with *rs13702*. Interestingly, in a functional experiment, RICHARDSON and colleagues showed that indeed *miR-410* is actively downregulating *LPL* transcript levels, an effect that is completely abolished in the presence of the minor C-allele of *rs13702* [398].

gene	SNP	metabolic trait	N_{eQTL}	targeting miRNA(s)	metabolic function
ACADM	rs8763	hexanoylcarnitine	9	miR-203a	-
THEM4	rs13320	5-dodecenoate	20	miR-101	-
SLC5A6	rs7081	mannose	-	*miR-128*	energy homeostasis [399]
RAB3GAP1	rs4954221	1,5-anhydroglucitol	1	*miR-22*, miR-490	gluconeogenesis [400]
CPS1	rs715	glycine	-	miR-432, miR-496	-
NT5E	rs6922	inosine	1	miR-518f, miR-218, miR-134	-
SLC16A10	rs14399	tyrosine	-	*miR-214*, miR-503	lipid homeostasis [401], gluconeogenesis [402]
CCBL1	rs10988134	indolelactate	2	*miR-30c*, miR-30d, miR-30a, miR-30e, miR-30b	lipid homeostasis [403]
ALDH18A1	rs4037	citrulline	6	miR-376b, miR-377, miR-376a, miR-381, miR-300	-
ABCC4	rs3742106	indoleacetate	-	*miR-320a*	response to glucose [404]
IVD	rs7207	isovalerylcarnitine	1	*miR-23a*, *miR-23b*	glutamine metabolism [405]
LACTB	rs8468	succinylcarnitine	9	miR-181c, miR-544a, miR-181a, *miR-217*, miR-98, *miR-33b*, *miR-33a*	lipid homeostasis [406, 407]
PDXDC1	rs6498540	dihomo-linolenate	2	*miR-34a*	lipid, cholesterol, and energy homeostasis [407]
SLC7A5	rs1060253	kynurenine	-	*miR-301a*[*], *miR-130b*[*,†], *miR-130a*[*,†]	* energy homeostasis [399] † lipid homeostasis [408]
GCDH	rs8012	glutaroyl carnitine	22	miR-873	-

Table 14: **Fifteen blood GIMs potentially influenced by genotype-dependent miRNA regulation**. Shown are predicted causal genes from our blood mGWAS (see section 4.2) that harbour a SNP within a RISC binding site. Further, the metabolic traits associated with the variants, the count of tissues with identified eQTL associations (N_{eQTL}) of the variant or a LD proxy ($r^2 \geq 0.8$) and the respective gene, as well as the miRNAs with MREs in the RISC binding site are listed. miRNAs with implications in metabolic processes, as given in the last column, are highlighted in italic.

To investigate the impact of genetically influenced miRNA regulation pathways in human metabolism further, I used the results from our blood mGWAS study described before (section 4.2), as this data is already contained in *SNiPA*'s annotations. Interestingly, for 15 of our 145 loci (10.3%), I found an mQTL to be located in a miRNA target site within the predicted causal

gene. Of the 37 miRNAs predicted to target these genes, at least 14 (37.8%) have been previously described to affect metabolic homeostasis (Table 14). In the following investigation of the metabolic traits associated with the SNPs located in miRNA target sites for compliance with the described metabolic effects of the respective miRNAs, a functional relationship to the genotypic effects became apparent. For instance, *rs4954221*, located in a binding site of *miR-22-3p* in the *RAB3GAP1* gene, is significantly associated with 1,5-anhydroglucitol, which is a long-established marker for glycemic control [409-411]. In a functional assay, KAUR and colleagues showed that the same miRNA is a critical regulator of hepatic gluconeogenesis [400], which is directly linked to the glycemic status. Another example is *rs6498540* located within a RISC binding site on the *PDXDC1* gene containing the MRE of *miR-34a*. The variant is significantly associated with a fatty acid (dihomo-linolenate), while the miRNA has been described to affect cholesterol, lipid, and energy homeostasis via a complex regulatory network [407]. If our underlying hypothesis is correct that the allele-specific alterations of blood metabolite concentrations is due to disturbed miRNA targeting of the predicted causal genes, this would indicate a widespread functioning of genetic variance in human metabolism through miRNA-mediated regulatory pathways.

Intriguingly, our approach can also be used to identify plausible causal genes for mQTL associations not reaching genome-wide significance. For instance, I found *rs7942396* that is suggestively associated with lathosterol levels ($P_{value} = 7.68 \cdot 10^{-8}$) to be located in a RISC binding site within the 3'-UTR of the *sterol-C5-desaturase (SC5D)* gene. The encoded protein, lathosterol oxidase, catalyzes the conversion of lathosterol to 7-dehydrocholesterol (EC 1.14.19.20) [412]. In rare cases, both copies of *SC5D* are mutated and dysfunctional, leading to lathosterol accumulation, or lathosterolosis, which is an inborn error of metabolism (MIM: 607330). When *SC5D* is functional, lathosterol levels can be used as marker for cholesterol synthesis and have been shown to correlate with BMI, blood pressure, atherosclerosis, aortic stenosis, and coronary artery disease severity [413, 414]. The RISC binding site affected by *rs7942396* is targeted by two miRNAs. One of them, *miR-499*, has been shown to feature cardioprotective properties [415]. Following our hypothesis, in addition to the previously proposed mechanism of preventing cardiomyocyte apoptosis, this may result from confined cholesterol synthesis by downregulation of *SC5D* expression by *miR-499*. The allele-specific efficiency of miRNA-mediated silencing of *SC5D* expression may thus explain the association of *rs7942396* with lathosterol levels, although it was not genome-wide significant in the mGWAS due to the multiple testing burden.

6.2.6 Concluding remarks

In this study, we investigated if trait-associated variants in the 3'-UTR may exert regulatory effects that, to differing extent, affect trait development by interfering with miRNA targeting pathways. There is no evidence for considerable direct mutational disturbance of miRNA processing: only one SNP *(rs2168518)* is located in the hairpin sequence of *mir-4513*. However, our results uncover several lines of evidence on miRNA involvement in genetically influenced posttranscriptional trait emergence. The observed highly significant enrichment of trait-associated SNPs in the 3'-UTR strongly suggests a functional coherence between genetic variants and miRNA regulation pathways in *cis*.

In this context, the investigation of specific variant effects on functional elements in the 3'-UTR revealed several potential mechanisms of allele-specific miRNA-mediated regulation. The smallest fraction of predicted functional 3'-UTR SNPs affects 3'-UTR splicing. These variants are predicted to mediate miRNA target site loss, mostly through the gain of acceptor splice sites, resulting in shortened 3'-UTR sequences. The strong impact of alternative splice variants on miRNA targeting, manifesting in a high fraction of target site loss (46% on average) in the affected transcripts, could explain that altered splicing is rarely caused by common variants. The second most common genetic effect we observed is the SNP-mediated alteration of RNA secondary structure of a RISC binding region. The impact of RNA folding on the binding affinity of RBPs has already been described [387, 389]. However, the extent to which this phenomenon translates into miRNAs mediating human trait development is unknown. With our results, we provide a first data basis on RNA structural changes leading to phenotypic variability which may serve as a starting point to investigate this matter further. The most abundant mechanism in our study is the direct alteration of MRE sequences. We find not only that GWAS-identified markers in the 3'-UTR show a significant enrichment within MREs, but also identify a novel scenario of how miRNA dysregulation may take effect: the substitution of the recognition element of one miRNA by that of another miRNA. While a disruption (or creation, respectively) of a MRE enables a rather straightforward rationale, that is the tissue-specific repression (or enhancement, respectively) of miRNA regulation, this scenario makes interpretation rather complex. Such a substitution may imply concurrent but simultaneously diverging effects in different tissues, depending on the respective expression patterns of the two miRNAs, possibly leading to systemic disturbances of several cell types. The overlap between the two sets of SNPs which disrupt and create MREs amounts to 13 polymorphisms and constitutes more than one third of the set of variants affecting MREs - which is a surprisingly

high number. We believe that the transcripts affected by a SNP mediating this effect may present quite interesting targets for further studies. Moreover, a large fraction (39%) of 3'-UTR SNPs predicted to be effective shows a low conservation indicating that the creation of a MRE may be an abundant process of functional SNPs.

One shortcoming of our initial study was the limited variant set provided by the HapMap consortium. The final release of the 1000 genomes project as contained in *SNiPA* holds data on more than 78 million genetic variants, while the joint data of the HapMap release I, II, and III only lists 2.8 million variants. In order to find out if our results still hold true given this large increase in genetic data, I performed a reduced re-analysis of the original approach on the 1000 genomes variant set. Intriguingly, although this analysis was performed on 28 times the number of variants, enrichment statistics only changed marginally with respect to both enrichment and significance. This is especially interesting because 1000 genomes data is believed to contain virtually all existing genetic variants with MAF>1% in the large human populations, emphasizing the applicability and transferability of bioinformatics approaches in genetic analyses.

In the next part, I followed up on the finding in our original study that 3'-UTR SNPs are significantly enriched for influencing lipid homeostasis, which I replicated on 1000 genomes data and exemplified with the hypothesis on allele-specific regulation of *LPL*. I then used the large catalog of mGWAS results contained in *SNiPA* to annotate mQTLs with potential involvement in miRNA-mediated regulation of human metabolic homeostasis. This revealed a significant overlap of variants located in RISC binding sites and mQTL signals. More specifically, I was able to re-identify more than 10% of the predicted causal genes of our blood mGWAS using this approach. Further, I found conclusive evidence for mQTL interactions with miRNA targeting above the traditional threshold for genome-wide significance. In case of *SC5D*, the P_{value} is only narrowly falling short of the classical threshold of $5.0 \cdot 10^{-8}$. However, I also found other examples where insignificant mQTL associations have a large biological support. For instance, *rs1683787* is an mQTL for 1-palmitoylglycerophosphocholine (lysoPC a C16:0) levels with $P = 4.12 \cdot 10^{-5}$. It is located in a binding site of *miR-216a* in the *acyl-CoA dehydrogenase family, member 9 (ACAD9)* gene. *ACAD9* is highly expressed in the liver and, besides mediating the assembly of the mitochondrial respiratory chain Complex I, has dehydrogenase activity for palmitoyl-coenzyme A (C16:0) [416, 417]. *ACAD9* deficiency is an inborn error of metabolism (MIM: 611126, OrphaNet: 99901) that manifests with failure to thrive, cardiomyopathy, exercise intolerance, liver disease and mild to severe neurological dysfunction. Expression levels of *miR-216a* have been demonstrated to correlate positively

with LDL cholesterol levels and negatively with left ventricular ejection fraction and are strongly upregulated in ischemic heart failure [418, 419]. In summary, the metabolite associated with *rs1683787* is a close match to the substrate of *ACAD9*, which I predict to be targeted in allele-dependent efficacy by *miR-216a*, a miRNA that is implicated in symptoms and disease phenotypes connected to *ACAD9* deficiency.

This and the other proposed hypotheses clearly show the benefits posed by including more evidences (in this case, miRNA targeting information) into the molecular, functional interpretation of GWAS findings. The examples provide not only a superficial connection between molecular traits and genes based on genetic proximity of variants and genes, but include clearly accessible mechanisms by which genetic variation can contribute to trait variability. Of course, these hypotheses are still limited to co-located genes and do not allow for hypotheses regarding intergenic variants.

6.3 Predicting eQTLs via cross-tissue regulatory clusters

The major challenge in studying genetic associations with human traits is the identification of the causal genes that are responsible for the observed association signals. As the majority of identified trait-linked genetic variants are located in non-coding segments of DNA, investigating regulatory disturbances as possible molecular causes of complex diseases is a still a central bottleneck in genetic studies. In this context, I developed an automated approach for assigning predicted causal genes to the loci identified by our mGWAS on urinary metabolic traits (section 4.3). Interestingly, while our paper was under review, SPAIN and BARRET proposed a workflow for automated target gene identification that was very similar to my approach (Figure 22) [420]. In their workflow, markers are prioritized using fine-mapping and scoring methods to select the likely causal variants. These variants are then split into coding and non-coding variants. For coding variants, the selection of the target genes is naturally very straightforward. Additional data for downstream annotations can be retrieved using scoring systems such as PolyPhen2 and SIFT that estimate the deleteriousness of the sequence alteration to protein funcioning, while protein databases such as UniProt can be inspected for functional and domain-specific information. For non-coding variants, the authors recommend the

integration of various additional datasets. The classical approach of selecting target genes via genetic proximity (that in the beginnings of GWAS was without alternative due to the unavailability of regulatory element annotations) is then substituted by linking target genes via eQTL data or affected regulatory elements as provided by ENCODE, FANTOM5, or the NIH Roadmap project.

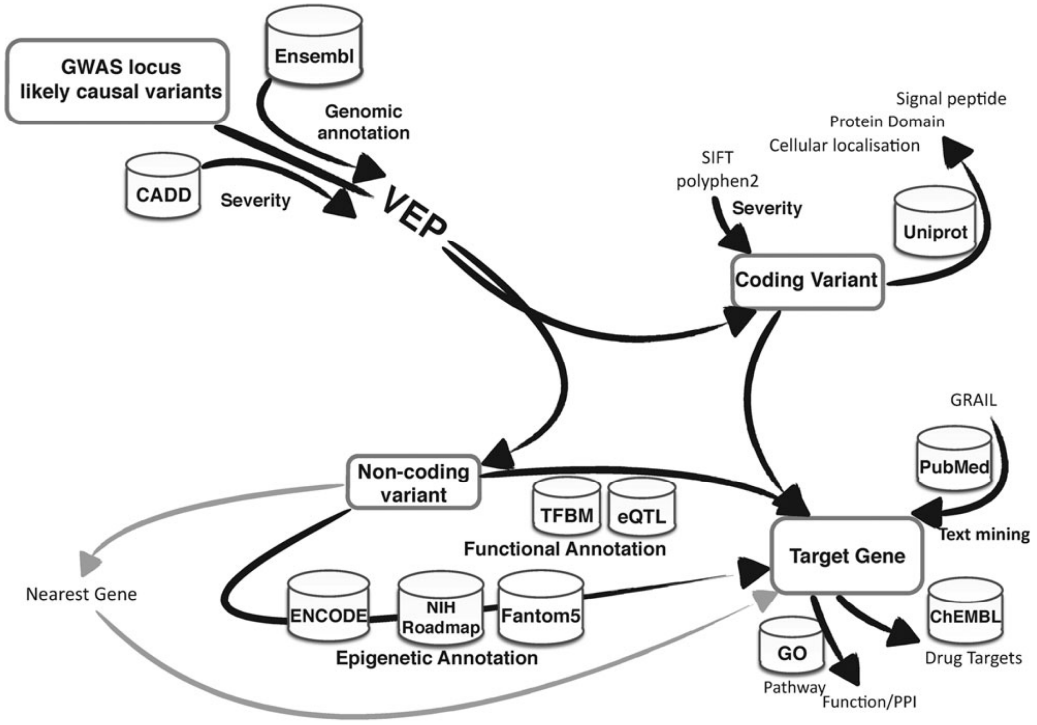


Figure 22: Workflow proposed by Spain and Barret illustrating the annotations to be included into the assignment of genes to genetic variants. Our *SNiPA* resource contains almost any of the listed data sources as well as hyperlinks to the databases suggested for post-processing (such as UniProt and PubMed). TFBM: transcription factor binding motif. Reprinted by permission from Oxford University Press: Human Molecular Genetics [420], copyright 2015.

Variants linked to multifactorial disorders have been found to substantially overlap with eQTLs. However, eQTL data, that can give direct hints on genes affected by a variant through regulatory mechanisms, have until very recently been only available for a limited set of tissues and cell types. And although the V6 release of the GTEx project [211] boosted the number of tissue types covered by eQTL studies, *SNiPA* (integrating GTEx V6 and several other eQTL datasets) contains eQTL data for only 50 tissues and cell types. eQTLs for tissues that are hard to access (e.g. brain) or rapidly disintegrate (e.g. kidney) are still underrepresented or even unavailable. Nonetheless, enrichment analysis performed on *SNiPA*'s annotations shows that a variant contained in the GWAS catalog is almost 30 times as likely to be an eQTL as variants contained in 1000 genomes without a trait association ($OR = 27.56$, $CI_{95\%} = [26.73 -$

$28.47], P < 4.9 \cdot 10^{-324}$). Including associations that are not genome-wide significant (covering dbGaP and the metabolomics GWAS server) decreases the OR, but still shows almost 20-fold enrichment for eQTLs ($OR = 19.58,\ CI_{95\%} = [19.39 - 19.78], P < 4.9 \cdot 10^{-324}$).

Despite these impressive numbers, looking at the total variant counts, only 40.5% of GWAS catalog variants and 32.0% of all trait-associated variants contained in *SNiPA*, respectively, can be linked to candidate genes via eQTL data. As both eQTL analyses and GWA studies use the same background set of genotypes, obtained by commonly used genotyping arrays and imputation to the same haplotype maps (HapMap2 or 1000 genomes), this is noteworthy. It has been frequently described (also by us [163]) that the frequency of GWAS markers with significant associations is skewed towards higher MAFs (>10%), which at least partially originates from the higher statistical power to detect associations using more frequently occurring alleles (see Figure 10D). Therefore, if GWAS-identified variants mainly affect regulatory mechanisms, the coverage by eQTL data should be in ranges equal to the proportion of non-coding variants, which, according to *SNiPA* v3.1, is >90%. However, most eQTL studies limit analyses to *cis*-effects at an arbitrary distance threshold (usually +-1Mb) in order to restrict the multiple testing problem. This leads to a limited number of eQTLs identified to affect the expression of genes located outside these fixed intervals. To find those as well as weaker *cis*-eQTLs, larger study populations would be needed. Especially for tissues that are not easily accessible for genetic and transcriptomic analyses, this holds not only major logistical and analytical challenges, but also a large cost burden.

The aforementioned consortia, however, have provided a great amount of data on regulatory regions across hundreds of cell types. And although mapping genetic variants to such regulatory elements as well as identifying the genes affected by changes to these elements remains difficult, predicting variant-gene associations utilizing these resources would abolish the issues linked to large-scale screens of cross-tissue eQTL analyses. In this study, I therefore used the data on regulatory DNA segments from these sources to build a new catalog of functional regulatory elements augmented with protein-DNA interaction data and target gene information. To this end, we combined ChIP-seq data clusters from ENCODE with ENCODE DNase I hypersensitive sites classified as promoters and enhancers and promoter upstream antisense RNA (uaRNA) and enhancer RNA (eRNA) mappings from FANTOM5 across several hundred tissues to retrieve a set of clustered enhancer and promoter regions. Using gene associations of elements clustered together, we assessed the compatibility of ENCODE and FANTOM5 with respect to promoter and enhancer annotations. Finally, we

used eQTLs located within the obtained clusters as benchmark for predicting regulatory variant-gene associations. The benchmark statistics given here have been mostly calculated by Nick Lehner in his Bachelor thesis [421].

6.3.1 Methods summary

DATA – For this analysis, I incorporated the data contained in version 2 of the SNiPA database, comprising 1,127,033 regulatory elements (defined as either DNAse accessible, transcribed regulatory, or TF-bound DNA) and 522,298 eQTL associations. eQTL data were extended to proxies in strong LD ($r^2 \geq 0.8$) using phase 3 version 5 data of the 1000 genomes project (contained in *SNiPA*), leading to a final set of 1,856,015 variants with eQTL associations. The datasets and entity numbers are given in Table 15.

Dataset	$N_{entries}$
ENCODE ChIP-Seq clusters	406,631
ENCODE DNase1 hypersensitive promoters	56,493
ENCODE DNase1 hypersensitive enhancers	538,515
FANTOM5 expressed promoters	82,419
FANTOM5 expressed enhancers	42,975
eQTL associations (from 8 tissues)	522,298
eQTL associations (LD extended $r^2 \geq 0.8$)	1,856,015

Table 15: Datasets and entry statistics for sources included in cross-tissue regulatory cluster generation.

CLUSTERING OF CROSS-TISSUE REGULATORY ELEMENTS – Regulatory feature clusters from ENCODE as contained in the Ensembl database are split into several different categories. More specifically, this means that ChIP-seq clusters for predicted enhancers, promoters, promoter flanking regions, etc. can overlap but are separate entities. The same applies for promoter and enhancer annotations for different tissues/cell types in FANTOM5. As we wanted to assess tissue-general effects this data representation was undesirable, albeit biologically meaningful. I therefore first determined the genetic ranges of regulatory segments using a dynamic programming approach, extending single elements up- and downstream according to the overlapping elements from all sources, simultaneously sustaining all annotation data of the single elements for later use. The final clusters are then classified as promoter, enhancer, or both based on the annotation of the included elements. After intersection of the clusters with the extended set of eQTLs, the associated annotations comprise tissue-specific ChIP-seq binding data, gene associations (as reported by ENCODE and FANTOM5) separated by enhancers and promoters as well as by data source, and eQTL-

associated genes including tissue information. Where available, I also overlaid the clusters with TF binding motifs affected by a variant ($n_{total} = 124{,}142$).

6.3.2 Construction of cross-tissue regulatory clusters (CTRCs)

When investigating variant effects in *SNiPA*'s variant browser, a stack-like accumulation of regulatory elements originating from different sources becomes evident (Figure 23). Visual investigation of these stacks often shows concordant annotations between ENCODE and FANTOM5. As variant effect predictions typically require a positional overlap between variant and an affected genomic entity, the information from the different sources is not readily available for the individual variants. In order to provide support for the interrelations of eQTLs, trait-associations, and regulatory elements, I assessed if there is enrichment for overlaps of these three entities. The results showed that trait-associated variants have a tendency to hit regulatory elements compared to all other 1000 genomes variants ($OR = 1.38,\ CI_{95\%} = [1.37 - 1.40], P < 4.9 \cdot 10^{-324}$). Intriguingly, the enrichment of reported-only eQTLs (i.e. without LD proxies) in regulatory sequences was almost identical to that of trait-linked markers ($OR = 1.37,\ CI_{95\%} = [1.36 - 1.37], P < 1.0 \cdot 10^{-324}$). As the ENCODE ChIP-seq feature clusters are generally larger than enhancers and promoters, we additionally checked for differences in the enrichment values for FANTOM5 elements. Surprisingly, although FANTOM5 annotations contain fewer elements that, additionally, are significantly smaller than ChIP-seq clusters, GWAS catalog variants showed a higher enrichment ($OR = 2.30,\ CI_{95\%} = [1.84 - 2.85], P = 1.2 \cdot 10^{-11}$). When using all trait associations contained in *SNiPA*, the enrichment was even stronger ($OR = 5.35,\ CI_{95\%} = [5.12 - 5.58], P < 4.9 \cdot 10^{-324}$). These results indicate that trait-linked genetic variants are more frequently affecting actively expressed regulatory sequences rather than ChIP-seq-identified signals. This may seem contradictory, because expressed sequences should also be detected by ChIP-seq screens (for instance for polymerase activity). However, as both uaRNAs and eRNAs are very short and unstable, the transcription complex quickly detaches from the DNA and may thus be missed in ChIP-seq settings [79, 80].

As simple enrichment analysis already supported the hypothesis that clustering of these datasets may provide further insights, I performed a position-based clustering using a dynamic programming approach. The 1,127,033 input elements could be clustered in 167,985 cross-tissue regulatory clusters (CTRCs), with 138,060 clusters (82%) containing ChIP-seq and promoter/enhancer annotations from ENCODE or FANTOM5 and 29,925 clusters (18%) with annotations from all three data sources. Our CTRCs covered 51.8% of ChIP-seq feature clusters, 51.1% and 37.5% of ENCODE promoters and enhancers, respectively, and 63.3% and

61.5% of FANTOM5 promoters and enhancers, respectively. Intersecting CTRCs with LD-extended eQTL data, I obtained 68,190 clusters (41%) for which variant-linked expression changes where available.

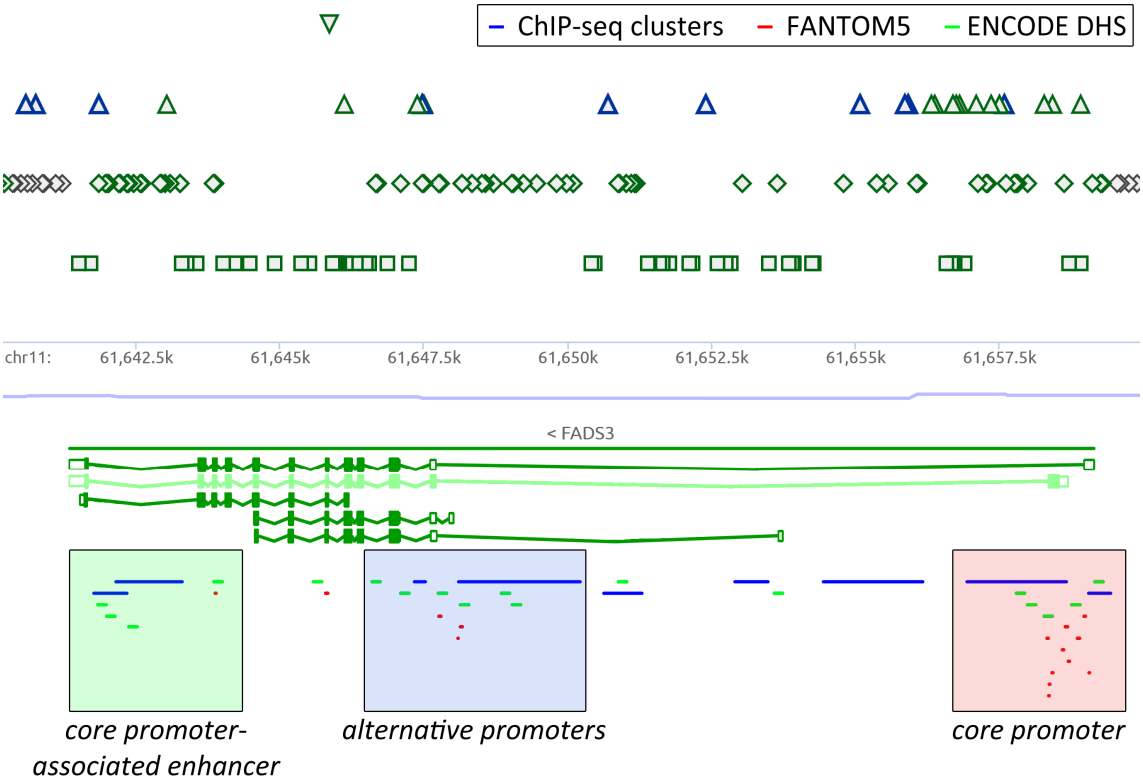


Figure 23: **Illustration of the stack-like accumulation of unclustered regulatory element annotations in *SNiPA* for the *FADS3* locus**. The sources of regulatory element annotations are color-coded. ENCODE ChIP-seq feature clusters are shown in blue, FANTOM5 promoters and enhancers in red, and ENCODE DHSs calssified as promoters and enhancers are in green. Manual clustering of the single elements into functional CTRCs would retrieve a core promoter cluster (red box), alternative promoters (blue box), and an enhancer element that interacts with the core promoter (green box). To demonstrate that alternative promoters actually control the expression of different isoforms of one gene, five transcript models from GENCODE are displayed. The canonical transcript with CCDS entry (ENST00000278829) is highlighted in light green. The screenshot shows *SNiPA* version 3 annotations. DHS: DNase hypersensitive site. FADS3: fatty acid desaturase 3.

6.3.3 Evaluation of compliance of within-CTRC annotations

It is common practice to include as many datasets as possible into effect prediction of non-coding variants to yield maximal coverage of the human genome sequence. However, it has never been empirically shown that datasets on regulatory elements are comparable, neither is known to which extent the results of different experiment types conform to each other. In this study, we used annotation data from four different experiment types: ChIP-seq signal peaks, DNase1 hypersensitivity screens, CAGE-based sequencing of uaRNAs and eRNAs, and eQTL

associations obtained by different gene expression microarrays. An important aspect of such integrative approaches is to determine the validity of the data consolidation approach. However, as a functional relationship between regulatory elements and the predicted target genes can only be verified in very targeted and controlled experiments, the catalog of such validated interactions is rather small. As for the proof of allele-specific effects of a genetic variant located within such a regulatory site has to be again experimentally tested, there is currently no possibility of deriving a large-scale reference set of true positive disturbances of gene expression via genetic variation of regulatory sequences.

Since there is no complete gold-standard available, we used benchmark statistics that are utilized in the assessment of the performance of machine learning applications. Simply put, we used one set of gene-associated elements (ENCODE enhancers and promoters; FANTOM5 enhancers and promoters; eQTL associations) as Criterion Standard and tested the performance of the other sets within each CTRC against it. Summing up the individual numbers for true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), we then derived global performance scores for the respective tested set. As commonly occurring in benchmarks applied on genome-wide (or other large-scale) annotations, we also observed extremely large numbers of TNs, which in our case are the genes that are neither predicted (in the test set) nor annotated as being regulated (in the Criterion Standard) by a CTRC. Consequently, we could not use performance measures that make use of TN counts, which unfortunately also includes the specificity. We therefore decided to use the F_1-measure that calculates the harmonic mean of the sensitivity and the precision and thus quantifies the effectiveness of a predictor (or, in our case, the respective tested set) in a single value [422].

When applied to the two promoter sets from ENCODE and FANTOM5, the F_1-measure showed excellent performance ($F_1 = 90.3\%$ for ENCODE and $F_1 = 90.4\%$ for FANTOM5). This can be considered a proof of concept for the application of the F_1-measure, as the annotation of promoters is methodologically biased in the ENCODE set, but not the FANTOM5 set. For ENCODE, DHSs in close proximity to transcription start sites (TSSs) are classified as promoter-associated. From a biological point of view, this approach is intuitive. However, such a basic hypothesis cannot be expected to hold true for all segments proximal to a TSS that are accessible to DNase1. FANTOM5 promoters on the other hand are measured as uaRNAs strongly correlating with the simultaneous expression of the target transcript, which is a methodologically sound approach. Therefore, the missing 10% in means of F_1 shows that the estimator well captures the differences between the two sets, while it also implies that promoters

are in general active across tissues. Using ChIP-seq feature clusters as classifier for actually active promoters (that is, using only CTRC with ChIP-seq binding annotations available), this has virtually no effect on the F_1-measure for FANTOM5 promoters. For ENCODE, however, it shows a significant increase to $F_1 = 95.1\%$ which can be interpreted as evidence that such rather basic approaches can be refined using additional annotations.

The benchmark of enhancer annotations produced only mediocre statistics: $F_1 = 25.3\%$ and $F_1 = 31.8\%$ for ENCODE and FANTOM5 enhancers, respectively. However, considering the experimental uncertainties in enhancer determination, this outcome is plausible. For ENCODE, cross-tissue correlations between the accessibility patterns of DHSs are used to link promoter-associated DHSs with distal DHSs. In case of FANTOM5, the same approach is used correlating co-expression of eRNAs and gene transcripts. In both studies the simultaneous activity of independently regulated genes cannot be separated from those that are actually linked to the regulatory site, naturally inflating FP counts while decreasing TP numbers. Interestingly, when again using ChIP-seq peaks as a prerequisite for CTRC classification as active enhancers, F_1-measures are more than doubled for both sets ($F_1 = 54.6\%$ for ENCODE and $F_1 = 69.0\%$ for FANTOM5). This is direct evidence that, independent of the experimental setting, the combination of data on both accessible/expressed DNA and protein binding strongly increases the agreement of diverse annotations.

6.3.4 Performance of CTRCs in regulatory variant effect prediction

The central aim of this study was to find out if it is possible to predict the correct target genes of putative regulatory-acting genetic variants for which no eQTL associations are available using only regulatory element annotations. For this, we actually had a concrete Criterion Standard available, namely the large catalog of eQTL associations contained in *SNiPA*. The highly significant enrichment of eQTLs (including LD proxies) within CTRCs ($OR = 1.53,\ CI_{95\%} = [1.52 - 1.54], P < 4.9 \cdot 10^{-324}$) shows a clear connection between CTRC regions and genetic control of gene expression via allele-specific alteration of functional regulatory elements. As annotation agreement was highest for CTRCs combining gene-associated elements with ChIP-seq peaks, for this analysis we pruned the CTRC set accordingly. The global benchmark (tested set: ChIP-Seq peaks plus enhancers and promoters from both ENCODE and FANTOM5) resulted in an $F_1 = 26.0\%$, meaning that without distinction between enhancer and promoter CTRCs, the predicted positive target gene for a putative regulatory variant would be correct in more than one out of four cases. The corresponding values of the sensitivity (43%) and of the precision (19%) further reveal that for

almost half of the genetic variants the correct target genes are contained in CTRC annotations. The lower precision indicates that CTRCs list more genes via gene-associated promoters and enhancers than are available in the eQTL association data. As in this analysis, we included eQTLs from only 8 tissues while enhancers and promoters from ENCODE and FANTOM5 have been obtained from several hundred cell types, it can be expected that the actual precision (and F_1-measure) is higher.

We next assessed the performance of the predictions separately by CTRC type, which showed that promoters performed significantly better than enhancers ($F_1 = 36.0\%$ vs. $F_1 = 22.8\%$). While the corresponding sensitivity values were almost equal at 46.2% and 41.1%, the precision of promoters (29.6%) was almost double that of enhancers (15.8%). When we dissected element annotations further, we observed that ChIP-seq paired FANTOM5 annotations globally outperform ChIP-seq paired ENCODE annotations (Table 16). We hypothesize that this is a result of the more sophisticated experimental design in the FANTOM5 study, as here active expression is measured and correlated with active expression of regulatory regions. Active expression may thus be a better predictor of regulatory element activity than DNA accessibility only. Globally, using ChIP-seq paired FANTOM5 annotations result in 41.3% correct positive target genes, which shows that this approach presents a highly valuable addition to traditional variant effect predictions. With close to 11.4 million genetic variants in *SNiPA* located in regulatory elements compared to only about 1.9 million eQTLs, this has also a large practical value.

tested set	measure	global	promoters	enhancers
FANTOM5	F_1	41.3%	49.4%	30.0%
	sensitivity	43.6%	47.2%	37.1%
	precision	39.2%	51.8%	25.2%
ENCODE	F_1	22.2%	31.7%	21.1%
	sensitivity	34.4%	24.6%	37.0%
	precision	16.4%	44.8%	14.8%

Table 16: **Benchmark statistics of gene–associated regulatory elements located in ChIP–seq peak clusters in the prediction of variant–gene associations**. Given are the values for the F_1-measure, sensitivity, and precision for the performance of FANTOM5 and ENCODE gene-associated elements in eQTL prediction globally and splitted by type of element, i.e. promoters and enhancers. FANTOM5 outperforms the ENCODE set in all instances which is probably due to its more sophisticated experimental and analytical design.

6.3.5 From evidence to biology: a case-study

As mentioned before, in our blood mGWAS (section 4.2), we performed manual annotation of loci influencing metabolic traits. This was very successful for some loci, however,

although we included many databases and annotations in this search for genetic links between the association signals and the predicted causal genes, for some loci this process did not result in a convincing hypothesis including a plausible target gene. For instance, on chromosome 11, we found an association signal clearly defined by recombination events (chr11:18.27–18.39 Mb) with alpha-hydroxyisovalerate (lead SNP *rs2403254*, $P = 1.26 \cdot 10^{-30}$; Figure 24A). Neglecting non-coding genes, the locus contains only three candidates: *serum amyloid A1 (SAA1)*, *Hermansky-Pudlak syndrome 5 (HPS5)*, and the *general transcription factor IIH subunit 1 (GTF2H1)*. Alpha-hydroxyisovalerate is a product of energy production by branched chain amino acid (BCAA) degradation, a process that is located to several tissues, including skeletal muscle, liver, kidney, heart, and brain [423]. More specifically, alpha-hydroxyisovalerate is produced by hydrogenation of alpha-oxoisovalerate [424], the first product of valine degradation. As we could not link any of the three genes in this locus to a molecular function interfering with alpha-hydroxyisovalerate levels, we selected *HPS5* as candidate as it contained the (intronic, i.e. non-coding) lead SNP. After we had finished the development of the *SNiPA* resource, I revisited the locus annotations for our blood mGWAS using *SNiPA*'s Block Annotation feature similar to the automated approach in our urine mGWAS. In case of the *HPS5* locus, it revealed eQTL evidence implicating not only the three genes within the locus but one additional gene outside of it: *lactate dehydrogenase A (LDHA)*. Trait annotations showed implication in Mendelian disorders for three of the four genes (*SAA1, HPS5,* and *LDHA)*, as well as an additional GWAS trait association with amyloid A serum levels [425]. Furthermore, several regulatory elements were displayed as being affected by variants significantly associated with alpha-hydroxyisovalerate levels. Using the data stored in *SNiPA*, I then characterized the locus further in order to retrieve the most likely candidate gene for causing our association signal. Inspection of the genetic locus linked to serum amyloid A levels using *SNiPA*'s LD Plot showed a clear distinction from our association signal (Figure 24B) with the signal being located in the promoter-containing region of *SAA1*. This gene is implicated in a rare recessive disorder, secondary amyloidosis (OrphaNet: 85445, MIM: 104750), and is primarily expressed and translated in adipocytes [307]. This agrees with eQTL association data showing significant associations only in adipocytes. According to OrphaNet, secondary or inflammatory amyloidosis manifests with accumulation of amyloid fibrils consisting of serum amyloid A protein, which is consistent with the GWAS association. It has been shown that expression changes of monogenic disease genes can mimic some of the symptoms even of recessive disorders [297], a phenomenon which seems to be true also for *SAA1*. As the most

strongly associated variants for this trait, however, were not significantly associated with our metabolite, I excluded *SAA1* as candidate gene.

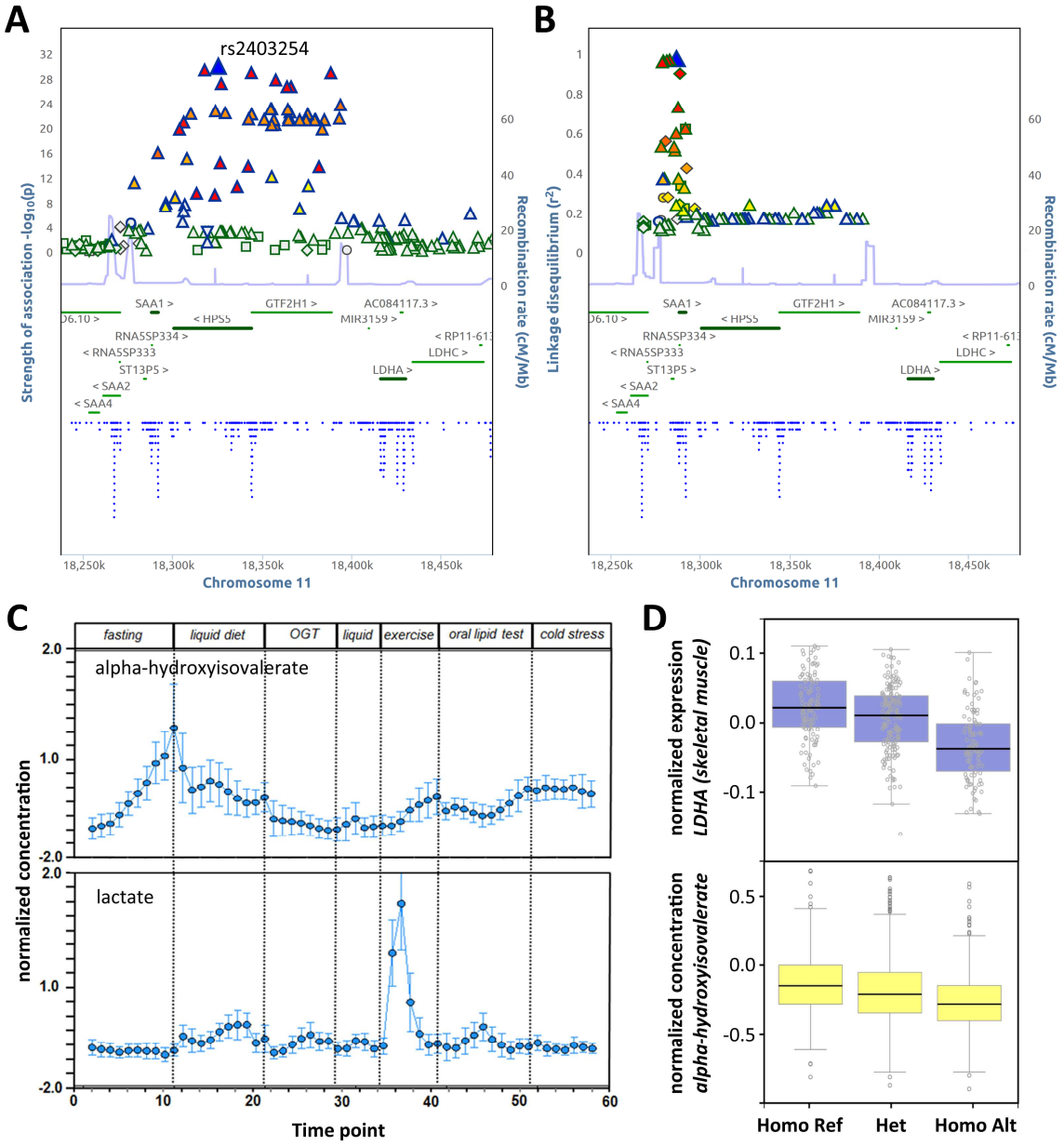


Figure 24: **Collection of evidences that support the identification of the predicted causal gene for the *HPS5* locus on chromosome 11 detected in our blood mGWAS**. **A**: *SNiPA* regional association plot of the association with alpha-hydroxyisovalerate. The lead SNP *rs2403254* is highlighted in blue. **B**: *SNiPA* linkage disequilibrium plot for one of the lead SNPs of the association with serum amyloid A levels. It clearly differs from our association signal. **C**: Blood concentration distribution of lactate and alpha-hydroxyisovalerate from the HuMet study across seven challenges and 56 time points, respectively. The course of concentration changes indicates antagonistic pathways for both metabolites. **D**: Allele-specific effects of rs2403254 in the eQTL (top in blue) and mQTL (bottom in yellow) data. The variant shows additive effects for both intermediate phenotypes. OGT: oral glucose test. Liquid: liquid diet. Homo Ref: homozygotes for reference allele of *rs2403254*. Het: heterozygotes. Homo Alt: homozygotes for the alternative allele.

Looking into the annotation for *HPS5* showed that the Mendelian syndrome (HPS5; MIMs: 614074, 203300; OrphaNet: 231512) linked to it is a very rare recessive disorder with often mild symptoms that is associated with impaired platelet formation, decreased immune function, and lung involvement. The symptoms again coincide with eQTL data listing significant associations in blood, monocytes, and lung tissue. However, Hermansky-Pudlak syndrome has never been linked to dysfunction of energy metabolism and, therefore, *HPS5* seems to be an unlikely candidate for causing the association with alpha-hydroxyisovalerate.

The third gene, *GTF2H1* is highly expressed in every measured tissue. Its protein product is involved in the general transcription complexes of RNA polymerases I and II and acts both as helicase in the transcription initiation complex and as enzyme in nucleotide excision repair [426, 427]. While for some mGWAS loci we detected a range of metabolic trait associations, the locus on chromosome 11 is specifically associated only with alpha-hydroxyisovalerate (ignoring ratios with alpha-hydroxyisovalerate). Such a specific association, however, is unlikely to be caused by a gene with such essential, vital functions.

Lactate dehydrogenase A, on the other hand, is a gene that by its function is directly involved in energy metabolism. In a recessive phenotype, it causes an inborn error of metabolism (glycogen storage disease; MIM: 612933; OrphaNet: 284426) associated with disturbed energy homeostasis and manifesting with increased levels of its primary pro- and educts, namely lactate and pyruvate, that are involved in anaerobic energy supply. Further investigation showed that the gene is only highly expressed in skeletal muscle [307] which is consistent with eQTL data. However, because of the aforementioned clear definition of the locus via recombination events and as *LDHA* lies outside of it, the genetic link between our association signal and *LDHA* seemed not very solid. I therefore examined the affected regulatory elements which listed an ENCODE enhancer element linked to the core promoter region of *LDHA*. Inspection of the annotation of the corresponding CTRC showed that it is located in the peak region of our association signal. Further, it contains a transcription factor binding motif of the Myc/Max complex which is altered by the minor allele of *rs3825025* that is significantly associated with decreased alpha-hydroxyisovalerate. Querying pathway databases revealed that the only gene within the whole locus that is actively regulated by this complex is *LDHA* [428]. To identify a potential link between alpha-hydroxyisovalerate and *LDHA*, I next investigated the allelic effects on *LDHA* expression and alpha-hydroxyisovalerate levels which showed that the effect directions are commutated (Figure 24D). This led me to the hypothesis that lactate dehydrogenase A might be able to produce alpha-hydroxyvalerate in the aerobic

valine degradation pathway if it is not occupied by lactate in its function in anaerobic energy supply.

When I checked the literature for this hypothesis, I found that HEEMSKERK and colleagues actually validated this hypothesis in an in vitro assay [429]. In contrast to my hypothesis, the authors speculated that *LDHA* produces alpha-hydroxyisovalerate during anaerobic energy consumption, which seems contradictory to their results where they show a specificity constant of *LDHA* for the conversion of alpha-hydroxyisovalerate 3000 times lower than for that of lactate [429]. To resolve this issue, I checked the results of the HuMet study where 15 healthy adults were metabolically screened across several challenges making up almost 60 measurement points [430]. Following the concentrations of the two metabolites over time indicated evidence for my hypothesis, namely opposite functions for *LDHA* in aerobic and anaerobic energy production, depicted by antagonistic patterns of concentrations of lactate and alpha-hydroxyisovalerate (Figure 24C). This alpha-hydroxycarboxylic acid could therefore be a marker of excess aerobic degradation of BCAAs. The example of maple syrup urine disease (MSUD; MIM: 248600), an inborn error of metabolism with loss of function of the branched-chain alpha-keto acid dehydrogenase complex that catalyzes the aerobic catabolism of alpha-keto acids coming from BCAA degradation (including alpha-oxoisovalerate), which HEEMSKERK et al. use to support their theory, is also rather evidence for my hypothesis, because the aggregation of alpha-hydroxyisovalerate in urine of MSUD patients is independent of physical activity [431]. In summary, taking into consideration all the available evidence and dissecting the information identifies *LDHA* as the most plausible candidate causing the association with alpha-hydroxyisovalerate. It also shows that the evidences point towards a pleiotropic function of lactate dehydrogenase which was verified experimentally and is supported by time-course data.

6.3.6 Concluding remarks

The central finding of this study is that, using CTRCs for variant effect prediction, we can obtain the correct target gene(s) in half of all cases for promoters and in one third of all cases for enhancers when comparing the predictions with the available eQTL associations. In case of enhancers, the sensitivity with close to 40% is much larger than the precision of the predictions, meaning that the correct target gene is listed along with several false positives in the annotations in four out of ten cases. Additionally, with the example on *LDHA,* I show how CTRC annotations can be used to dissect the evidences available for a genetic association signal leading to a sound hypothesis regarding the putative causal gene. This example also shows that data

integration can indeed support the geneticist in identifying the molecular mechanism underlying a genetic association signal and illustrates the whole process of categorizing the different evidence types into information relevant to the trait under study, deriving a functional hypothesis based on the combination of intermediate phenotypes (eQTL and mQTL data), and supporting it with additional evidences from independent studies. Of course, in the characterization of our association with alpha-hydroxyisovalerate, I neglect the evidence pointing towards further pleiotropy of the locus which is indicated by the strong eQTL associations linking our signal to the *HPS5* gene. However, if we look at associations with metabolite ratios, we see that the strongest signal for our lead SNP is that for the ratio of alpha-hydroxyisovalerate and 3-(4-hydroxyphenyl)lactate, a degradation product of tyrosine. One major symptom of Hermansky-Pudlak syndrome 5 is oculocutaneous albinism which is another recessive disorder that, in its most frequent form, is caused by homozygous mutations in the *tyrosinase* gene (MIM: 203100). *Tyrosinase* encodes a protein that is essential to produce the skin pigment melanin by catalyzing the conversion of tyrosine to L-Dopa. Degradation of tyrosine to either L-Dopa or 3-(4-hydroxyphenyl)lactate are distinct, mutually exclusive pathways. In my *LDHA* hypothesis, I assume that the primary enzymatic function of the locus is determined by *LDHA*'s potential to produce alpha-hydroxyisovalerate. When adjusting the regression model for this function by including alpha-hydroxyisovalerate as denominator in the ratio, the insignificant association of *rs2403254* with 3-(4-hydroxyphenyl)lactate ($P = 0.44$) is boosted to $P = 4.0 \cdot 10^{-39}$ with a highly significant P-gain of $3.15 \cdot 10^{8}$. This may indicate a metabolic link between *HPS5*, *LDHA*, and disturbed tyrosine metabolism. Interestingly, 3-(4-hydroxyphenyl)lactate is produced from 3-(4-hydroxyphenyl)pyruvate, a direct degradation product of tyrosine, by the same enzymatic reaction that converts pyruvate to lactate. And this latter reaction is catalyzed by *LDHA*. Of course, hypotheses on such effects in *trans* are very speculative, and it may well be that the fitting reaction type is sheer coincidence. With their large hydroxyphenyl groups, 3-(4-hydroxyphenyl)pyruvate and 3-(4-hydroxyphenyl)lactate are unlikely substrates of lactate dehydrogenase. On the other hand, a dehydrogenase specific for hydroxyphenyllactate was not yet found in human.

In conclusion, currently, a major focus in computational genetics is the development of variant effect prediction algorithms that are able to functionally characterize non-coding variants. In this context many datasets are included to yield maximal coverage of the human genome. However, the accordance of different sources of regulatory element annotations has never been investigated on a large scale. In this study, we combine regulatory evidences from

several experimental assays in cross-tissue regulatory clusters and demonstrate that the concordance between these datasets is surprisingly high. In fact, benchmark statistics for promoter and enhancer annotations are higher in CTRCs containing ChIP-seq data as compared to clusters without TF binding information. The validation of the clustering of regulatory elements via eQTL associations provides further evidence that our approach is highly valuable, while the use of eQTLs for validation is again justified by the strong enrichment of eQTLs within CTRCs. Additionally, our results underline the validity of using eQTL associations for candidate gene selection by supporting eQTL associations with regulatory element annotations. Our CTRCs thus enable further possibilities to select and prioritize candidate genes for trait-associated, non-coding, and potentially regulatory-acting genetic variants.

6.4 Summary

The major recurring point of criticism with regards to the GWAS approach is that the loci obtained by screenings for genetic trait associations are of only limited value because in most cases the associations can neither be linked to a plausible causal genetic variant nor provide information on the molecular mechanisms involved in trait development and progression. In this chapter, I describe three different approaches that are intended to support the process of prioritizing likely causal genes as well as of translating GWAS signals into relevant biochemical pathways.

The first study uses a comparably simple approach to select putative pathological variants using overlapping genetic signals for different traits. Using both agonistic and antagonistic signals, I investigate the genetic correlation between complex diseases as well as potentially pleiotropic mechanisms separating the development process of diseases on the molecular level. The genetic correlations clearly show that traits related on the symptom level also show overlaps in their genetic predisposition landscape. However, I also note that it is likely that a substantial part of these correlations may be due to biased covariate sampling or misclassification of (comorbid) cases. In a recent analysis, BULIK-SULLIVAN and colleagues proposed a novel approach for the identification of genetic correlations [432]. While they limit their rationale to

epidemiological aspects, they suggest that it would be an interesting research direction to characterize genetic correlation networks in a way similar to Mendelian randomization analyses where the objective is to identify the direction of causation for genetically correlated diseases. This would actually be a valuable addition to the available data, because it would allow for pruning genetic overlaps that are due to confounding by comorbidities. In our study, we used only summary statistics of published GWAS. Nevertheless, by inspection of their effect size we found evidence for SNPs appearing to be primarily associated with one disease which in turn represents a risk factor for another associated disorder. For instance, *rs2200733* on chromosome 4q25 is linked to atrial fibrillation (AF) with a higher effect size (OR=1.72) than to stroke (OR=1.26), for which AF is a major risk factor [433–435]. Interestingly, five variants in perfect LD with *rs2200733* are affecting an enhancer CTRC that is linked to *paired-like homeodomain 2 (PITX2)*, a gene that, in case of homozygous LOF, is causing familial AF (OrphaNet: 334). Another example is *rs964184* which is located proximal to the apolipoprotein gene cluster on chromosome 11q23 which is associated with hypertriglyceridemia with a markedly higher OR (OR=3.28) than to coronary heart disease (OR=1.13) [436, 437], which are again two diseases with a known connection. *rs964184* is in moderate LD ($r^2 = 0.37$) with *rs3135506*, a missense variant in the *apolipoprotein A-V (APOA5)* gene that has been previously linked to hypertriglyceridemia (MIM: 606368). The lower effects of the markers on the hypothesized "secondary sequels" may be explained by the fact that these are caused by the primary diseases, but with less than 100% penetrance. Intriguingly, in neither of the two examples above the risk factors (AF and hypertriglyceridemia) were included as covariates. Thus, it is also possible that the associations to coronary heart disease and stroke would vanish in the adjusted regression models, similar to what has been observed for the association of type 2 diabetes and the *FTO* locus when correcting for BMI [363]. On the side of putative pleiotropic genetic overlaps, we identify several loci that are interesting candidates for further studies. In this context, the interacting *TERC* and *TERT* loci may be of special interest. When studying the genetic overlaps between Mendelian and complex diseases, BLAIR and colleagues also report evidence for potential multi-functionality of the *TERT* locus [297]. *TERC* may thus also be a valuable target for more detailed exploration, especially as recent research on a wide range of non-coding RNAs such as *TERC* has demonstrated their potential to affect human diseases [438].

This topic is also the focus of the second study described in this chapter, where we investigate genetic influences on non-coding RNA functioning in the context of complex human phenotypes. Using a large catalog of experimentally supported miRNA binding sites, we

map trait-associated variants onto regulatory pathways to identify functional relationships between the genetic predisposition to human traits and the molecular level. Focusing on genetic influences on human blood metabolites, I detect a significant overlap between miRNA-mediated regulation and human metabolism. That miRNAs are involved in metabolic homeostasis has been shown previously [399, 400, 406, 407]. With our study, we demonstrate that there is ample evidence for genetic effects on miRNA-controlled metabolic pathways. Additionally, our approach provides another method to prioritize candidate genes for GWAS-identified association signals that can be used to combine evidences obtained in independent experiments. As we show for *LPL*-controlled lipid homeostasis, such a hypothesis-free approach can reveal molecular mechanisms for genetically influenced miRNA regulation pathways by connecting evidences from GWAS, miRNA screens, and enzymatic annotations of human genes. With examples on suggestive significant associations such as that with lathosterol from our blood mGWAS (section 4.2) that can be linked to genetically controlled, miRNA-mediated downregulation of lathosterol oxidase protein levels which furthermore may have pathophysiologically relevant downstream effects, we proof that our approach is not only very sensitive but also provides evidence-based support for genetic associations that fall short of the conservative significance threshold applied in GWAS.

In the third study, I extend this investigation of genetic influences on regulatory mechanisms to the classical elements involved in gene regulation, namely promoters and enhancers. Using over one million regions from several genome-wide datasets on regulatory elements, we were able to derive two important findings. First, we provide the first empirical analysis on the comparability of regulatory element annotations derived by different experimental assays. And while we find that more sophisticated experimental methods perform better (which could be expected), we also find that using ChIP-seq data as prerequisite for annotating active CTRCs removes a lot of putative false-positives, thus significantly increasing performance measures. Second, using available eQTL data, we demonstrate that CTRCs can be efficiently used to predict candidate genes linked to genetic variants. And although the numbers for the F_1-measure seem not too impressive, this is a highly valuable addition to the catalog of genomic annotations if used to predict the effects of variants for which no other annotations are available. Furthermore, the example on *LDHA* also shows that CTRCs can be very useful to dissect the evidences of highly annotated loci to find the most plausible predicted target gene for a GWAS-identified trait association. The aggregated regulatory annotations revealed a clear

connection between the well-defined association peak for alpha-hydroxyisovalerate and the more distal *LDHA* gene, a link that we missed in the manual annotation of the locus.

7 Discussion and outlook

Across the last two decades, the study of the genetics of human phenotypes has undergone unprecedented advances. The human genome sequence has been deciphered, ever-denser haplotype maps of the major populations became available, and GWA and sequencing studies have provided a catalog of several thousands of variants that influence trait predisposition and development [100, 222]. Population-based sequencing consortia have provided us with genomic and exomic sequences of ten thousands of individuals [231, 439]. By 2020, it is estimated that the genomes of about five million individuals will have been sequenced worldwide [440]. Other consortia obtained epigenetics and regulatory data for hundreds of human cell types [78, 82]. And with novel genome-editing techniques it became feasible to perform knock-out studies on human cell lines to identify sets of essential genes, something that previously was only possible in model systems like yeast [441, 442].

It seems paradoxical that, in spite of these impressive achievements and the rapidly growing compendium of data, current cost estimates for the development of new drugs are at 2.6 billion US dollars [443] – a number that more than triples the estimates from the beginning of the 21st century [444] – with an estimated success rate of less than 10%. However, it is an established fact that the translation of the information on disease-linked genetic variation into clinical use still follows at a much slower pace than new genetic data is produced. And although phenomena like the missing heritability problem are recognized and discussed, there are very few methodological advances developed to address these issues convincingly. Interestingly, this is not a new trend. During the Human Genome Project in the mid-1990s, the first large-scale

DNA sequences were accumulating and led some to generally question the classical theories in human genetics. But, in the words of KENNETH M. WEISS, *"IN MANY WAYS THESE ISSUES FADED WITHOUT RESOLUTION. IT WAS EASIER TO GENERATE EVEN MORE NEW DATA"* [445].

In this context, the demand to globally collect genotype–phenotype data in order to enhance exploitation of the available findings in a more expedient way receives increasing attention. GWAS, for instance, have provided a plethora of genetic associations with complex traits, however, most of the identified loci are still awaiting mechanistic characterization. For traits that are due to low-frequency or even rare genetic causes, sequencing can be used to obtain candidate variants for further investigation. However, this again holds several challenges, as *"EVEN AFTER EXTENSIVE BIOCOMPUTING AND FILTERING, THE MAIN RESULT IS TYPICALLY A LONG LIST OF VARIANTS OF UNKNOWN SIGNIFICANCE"* [446].

In this thesis, these obstacles are investigated from a bioinformatics perspective. We begin with the description of the standard GWAS approach utilized to detect genetic associations with the sudden infant death syndrome. Complemented by analysis of rare CNVs, we show the difficulties in the interpretation of the results of genetic association studies on trait endpoints. The focus is then shifted to the identification of genetic influences on intermediate phenotypes. Using the results of two large GWA studies on metabolic traits in human blood and urine, we demonstrate the great benefit posed by the investigation of molecular traits for the interpretation of genetic links to complex phenotypes. In this context, we further elaborate the utility of an integrative annotation of the identified genetic loci and exemplify how the identification of the predicted causal genes can be automated using a simple evidence-based gene prioritization metric. In order to provide this method to the scientific community, we developed the first genetic variant-centric annotation browser that, in addition to simple data retrieval, features several convenient tools for the analysis, aggregation, and visualization of genomic annotations linked to genetic variants on a genome-wide scale. To demonstrate the value of integrative analyses in the context of GWAS-identified genetic trait associations further, we then shed light on different aspects of the potential molecular consequences mediated by trait-linked genetic variants. First, we investigate the nature of genetic loci that show overlapping effects in a large set of common diseases and estimate the extent of pleiotropy for complex trait loci. Second, we examine trait-linked genetic influences on miRNA targeting and, using the previously obtained associations with metabolic traits, show that there is ample evidence for allele-specific modification of gene-regulatory mechanisms influencing energy homeostasis. To investigate this matter further, we thirdly inspect the value of genome-wide datasets on regulatory element

annotations in the prediction of regulatory effects exerted by genetic variants and provide the first empirical evidence on the performance of such predictions. In the following, these scientific contributions are summarized and potential future directions as well as links to the current emphases in the study of human genetics are discussed.

- The genetic analysis of the sudden infant death syndrome revealed that, at least in our cases and assuming the standard GWAS log-additive mode of inheritance, there are no major complex genetic risk factors determining susceptibility to SIDS. While we identified several suggestive significant loci, the resulting effect sizes are too small to be likely to lead to a significantly increased vulnerability of an infant to succumb to an external stressor. However, CNV analysis provided evidence that, for a quite substantial amount of our cases (12 of 301), rare monogenic disorders may have caused sudden death. To verify this finding, cytogenetic validation of the detected deletions is needed. Based on the very strict quality control measures applied to our data, we are confident that at least some of the CNVs called in our study are true positives. We therefore suggest including cytogenetic screens for chromosomal aberrations in the standard autopsy protocol of cases where SIDS is suspected as cause of death. Taking this outcome into consideration, examination of common recessive genetic causes of SIDS will be a valuable next step. Due to screening for recessive risk factors having lower detection power, we are currently extending our case cohort by additional cases from the Hannover SIDS cohort in collaboration with Thilo Dörk-Bousset. One promising target that already shows suggestive significant results for recessive effects ($OR = 4.40$, $P = 8.45 \cdot 10^{-8}$) in our cohort is the *acyl-CoA dehydrogenase, short-chain* (*ACADS*) gene locus. *ACADS*, like *ACADM* (that has been previously linked to SIDS pathogenesis), causes an inborn error of metabolism that can show infantile onset of acidosis and muscle weakness (MIM: 201470). With further candidates missing, thorough genetic characterization of SIDS will only be possible using sequencing. However, based on our study (which is, according to the literature, the first GWAS on SIDS) and the information available in the literature, prominently the success of the Back-to-sleep campaign, SIDS seems to be a highly heterogeneous phenotype that is predominantly caused by environmental factors.
- With our mGWAS on metabolic traits in human blood, we performed the largest mGWAS to date, detecting and replicating 145 blood GIMs. In addition to extending the catalog of variant-metabotype associations reported in earlier studies, we provide a

collection of downstream analyses that, in this depth, was a novelty. We give estimates on the heritability, explained variance, and environmental influence on 310 metabolites and examine epistasis between loci significantly associated with the same metabolic trait. We performed large-scale Mendelian randomization analyses of the direction of causation comparing the primary effects on gene expression versus metabolite concentrations identifying two loci where metabolic control is genetically modified by altered gene regulation. We provide a data-driven reconstruction of a metabolic network, augmented with the genetic loci that affect the pathways. Further, we annotated all loci integrating several genomic, phenotypic, chemical, and pharmacological databases, showing a substantial association between our GIMs and genes that are either drug targets, involved in drug uptake or metabolism, implicated in complex diseases, or causing inborn errors of metabolism. Finally, we compiled two comprehensive web resources to access the study data: first, a searchable supplemental website that enables browsing of all major results of our analyses and of the annotation of the detected GIMs, embedded in the reconstructed metabolic network; and, second, the metabolomics GWAS server that contains the complete set of variants and association statistics provided via a set of access and search interfaces and interlinked with several other specialized resources to enable further downstream analyses using our data.

The mGWAS on urinary metabolic traits, which is also the largest of its kind, complements the blood mGWA study on a second body fluid. Here, we established several methodological advances. First, we use non-targeted NMR spectra as phenotypic traits as proxy for unidentified metabolite profiles on a large scale that, coupled with the Metabomatching approach, can be used to more comprehensively exploit NMR data. And second, I developed a metric for automated evidence-based prioritization of predicted causal genes that, as this step is one of the major bottlenecks in large genetic association screens, constitutes a significant speedup in the annotation and interpretation of GWAS results. Furthermore, we integrate blood and urine metabolic traits in hypothesis generation which can provide deeper insights into allele-specific effects on metabolic homeostasis, as we show on the example of *SLC5A11*. Following the approach of the blood mGWAS, we again made all results publicly available by integrating and interlinking the association results genome-wide into the metabolomics GWAS server.

The finding that mGWAS-identified loci often contain enzymes that are functionally linked to the associated metabolic trait is an important one, as it shows that for GWAS of intermediate traits locus annotation and interpretation can in some cases be quite straightforward. However, a link between a gene and a metabolite in one of the metabolic pathway databases is not a proof of causality, and, therefore, causality has to be established in experiments. Nonetheless, the derived hypotheses are often biologically plausible and provide a valuable addition to the compendium of genetic annotations. To further elucidate genetically influenced metabolic control across tissues, in collaboration with the University Medicine Department in Greifswald we are currently performing an mGWAS study across human blood, urine, and saliva samples in the SHIP cohort. Another important step to characterize the genotype-metabotype relationship in more detail is the integration of as many available mGWAS datasets as possible. The current version of the metabolomics GWAS server was not intended to hold such amounts of data and we are therefore working on a relaunch of the resource that includes not only better-suited data access options but also data visualization modules that can cope with the multidimensionality of metabolomics data. We also plan on integrating the metabolomics GWAS server into the *SNiPA* resource to enable access to *SNiPA*'s large catalog of annotations while inspecting mQTLs. Additional extensions via integrative approaches, such as the investigation of genetically disturbed, miRNA-mediated metabolic control and genetic effects of GIMs on more distal enzymes, will be discussed in more detail below. Finally, NGS-based analyses of metabolic traits can provide further insights. It has already been shown for some examples that rare missense or LOF variants can push the metabolic profile of an individual to the extremes of the concentration distributions [125]. Identifying such variants on a large scale will help to identify key enzymes involved in metabolic homeostasis.

- With *SNiPA*, we developed a resource that, both from the integrated data and the provided data access and visualization modules, constitutes a universal starting point in the task of annotating genetic loci defined by single nucleotide variants. It contains LD information and functional annotations for almost all variants of the latest 1000 genomes project release and is extensible to include both larger variant sets and additional annotation data. Its current catalog contains, to the best of our knowledge, the largest available compilation of eQTL datasets, regulatory elements including miRNA target sites from CLIP-seq experiments, associations to complex, metabolic, as well as

pharmacogenomics traits from resources such as dbGaP, DrugBank, and the NHGRI-EBI GWAS catalog, annotations of missense variants from ClinVar, OMIM, and UniProt, monogenic disease gene annotations, and several genome-wide conservation (phastCons, phyloP, GERP++) and deleteriousness (PolyPhen, SIFT, CADD) scores. The block annotation module of the resource that enables an aggregated annotation of a complete association signal is currently unique in its form. And its variant-centered visualization methods provided a novelty that has already inspired others to follow our approach [447]. To maintain *SNiPA*'s value, we regularly update its data basis (current version is v3.1) with the latest releases of the included resources. For each major update, we sustain freezes of the annotations in order to enable citing a specific version of the database. Since January 2015, access statistics show that we have about 3,000 regular users that have accessed or downloaded more than 13 TB of data through almost 300,000 requests, which shows that the resource is broadly used by the scientific community. Future directions for the resource will include the incorporation of additional –omics layers that are now only scarcely represented in the data. For instance, protein quantitative trait loci (pQTLs) have until now been available in only small amounts. However, combining eQTL associations and predicted regulatory variants with this additional layer will enable to trace genetic effects from the genome across the transcriptome to the proteome. Additionally, further regulatory data sets, such as the epigenetic classifications from the NIH Roadmap Epigenomics Project, will enable categorization of non-coding genetic variants that are currently lacking annotation. Integrating the metabolomics GWAS server completely into *SNiPA* will add a further –omics layer. Finally, phenotypic ontologies as well as disease-symptom mappings can be used to extend the current trait annotation compendium of *SNiPA*. Listing all these data in the structured *SNiPA* cards will then enable to follow the effect of a genetic variant across the different –omics layers to the symptom-resolved phenotype.

- The integrative analyses in chapter 6 demonstrate the value of integrating different datasets by combining several layers of information in order to characterize the landscape of genetic trait predispositions more closely.

  The study on genetic overlaps between complex diseases applies a very straightforward approach and, based on the catalog of genetic associations selects those variants that are linked to more than one disease. Nevertheless, as we show, linking phenotypes over genetic associations can reveal interesting insights. On the one hand, we show that

almost 25% of all variants that are associated with two or more traits show pleiotropic signals. Investigating those loci, we generate several plausible hypotheses of branching disease etiologies. On the other hand, we show that agonistic associations across complex diseases not only mirror symptom-based trait similarity, but also identify potential comorbidities within the study cohorts. Using only the effect size of these associations without information on the individual phenotype or genotype, we are able to derive plausible hypotheses on the direction of causation if confounders have not inflated the association statistics and else to identify those confounders. In the Bachelor thesis of Niklas de Andrade Krätzig, we extended this approach by integrating GWAS-identified disease loci with proteinaceous biomarkers and detected that GWAS variants are twice as likely to be located in a gene encoding a biomarker ($OR = 1.97,\ CI_{95\%} = [1.74 - 2.24], P = 1.1 \cdot 10^{-24}$) [448]. Intriguingly, NELSON et al. recently found out that the same is true for drug targets [449]. Taking into consideration that GWAS hits are also enriched in Mendelian disease genes [297], it would be highly interesting to include all these datasets into one network representation to identify potential clusters defining disease pathways that may be druggable with existing substances not indicated in this context or present promising targets for the targeted development of new drugs. Using only their data, NELSON et al. already estimate a significant simplification and speed up of drug development that could be furthered by the combination with additional information.

Using CLIP-seq supported miRNA target sites co-localized with trait-associated variants, we then turn to investigate the influence of genetic variants on regulatory mechanisms. We establish a significant link between genetic variation and miRNA-mediated gene regulation by investigating the major functional elements within the 3'-UTR of human genes. While there is also evidence for disturbance of regulatory processes via affected splicing and changed RNA folding, we demonstrate that the predominant effect of trait-associated variants is the direct alteration of miRNA recognition elements. We also show that the efficacy of the control of metabolic homeostasis by miRNAs seems to be frequently dependent on the individual allelic configuration, which further emphasizes the importance of our findings. Conclusive examples such as that of *LPL* could be promising targets for siRNA-based drug development and should be examined further in functional assays. In order to extend these hypotheses by additional evidences, it would be a highly interesting approach to

include genotypes and miRNA and gene expression data from the same time points and individuals into one analysis. If our hypothesis is correct and efficacy of miRNA targeting is indeed individually affected by the allelic configuration this should result in patterns that can be identified using longitudinal data. Another future application of our approach will be enabled via including pQTL datasets into *SNiPA*. miRNAs are regulating gene expression on the posttranscriptional level via different mechanisms [450]. In addition to miRNA-mediated degradation of mRNA transcripts, transcripts can also be withheld from translation at different rates. Therefore, analyses on disturbed miRNA regulation pathways including proteomic data may show even further aspects that are missed using only gene expression data.

Further characterization of allele-specific effects on gene regulation is then obtained using a novel clustering of cross-tissue regulatory element annotations. Aside from showing that the individual annotation datasets originating from the different experimental assays conform well to each other, we demonstrate that using TF binding data as criterion for active regulatory elements increases the performance of the genomic annotations. Using the available eQTL data, we then examine the utility of our clusters in predicting allele-specific modifications of gene regulation and show that the combined cluster annotations are highly valuable in characterizing the molecular effects of non-coding variants. The aggregated information can reveal direct interactions between enhancer elements, specific transcription factors, and the expression of more distal genes, which we exemplify in detail by connecting our blood mQTL with alpha-hydroxyisovalerate to the *LDHA* gene. Other examples include *PITX2* that we identified in the re-analysis of our network of loci overlapping across complex diseases. For this analysis, there are several future directions. First, inclusion of the NIH Roadmap Epigenomics data will extend the ChIP-seq profiles for which previously only ENCODE data was available. Roadmap also provides genome-wide classification of DNA states based on dozens of ChIP-seq experiments for more than 100 tissue types which can be incorporated to refine our clustering approach. Using the additional 20 million eQTL associations provided by the GTEx consortium, we can then recalculate the performance measures for predicting eQTLs in a step-wise manner to identify the most valuable datasets for CTRC generation. The extension of cross-tissue clusters with a weighting scheme for the tissues where the cluster is most probably active would be a

further improvement. To obtain this, tissue and cell type ontologies could be included to dissect globally active elements from the cell type-specific activity states.

Conclusion

In this thesis, I investigated the genetics of human phenotypes on several distinct layers. In three genome-wide association studies we provide further insights on genetic loci involved in the presentation of human phenotypes with a focus on genetically controlled metabolic homeostasis. To address the challenge of interpreting genetic association signals on the molecular level, we developed a freely accessible web-based data integration resource that enables the thorough investigation of potential effects of the almost complete set of human genetic variation. Using the aggregated output of the block annotation module of this resource, we developed an automated approach to prioritize and weight candidate genes for association signals in order to facilitate the prediction of the causal genes. In three additional studies, we then used integrative analyses demonstrating the value of combining genomic and genetic data under different aspects, each generating new hypotheses that can be followed up experimentally to advance our understanding of the complex mechanisms that translate genetic variation into phenotypic variability. We expect that the future extension of these approaches with additional data, especially on genetic interactions and the modulating effects of environmental and behavioral factors, will further enhance the capabilities of bioinformatics approaches in supporting the interpretation of genetic association data.

List of abbreviations

1000 genomes	–	The thousand genomes project
ACID	–	Atomicity, Consistency, Isolation, Durability
AF	–	Artrial fibrillation
ANOVA	–	Analysis of variance
BAF	–	B allele frequency
BMI	–	Body mass index
CADD	–	Combined annotation dependent depletion
CAS	–	Chemical Abstracts Service
CCDS	–	Consensus Coding Sequence
CDCV	–	Common disease/common variant
CeD	–	Celiac disease
ChIP-seq	–	Chromatin immunoprecipitation DNA-sequencing
CLIP	–	Crosslinking immunoprecipitation
CNV	–	Copy number variant
CPMA	–	Cross-phenotype meta-analysis statistic
CRC	–	Colorectal cancer
CRLMM	–	Corrected Robust Linear Model with Maximum Likelihood Classification
CTRC	–	Cross-tissue regulatory clusters
dbGaP	–	Database of genotypes and phenotypes
DBMS	–	Database management system
DHS	–	DNaseI hypersensitive site
DNA	–	Deoxyribonucleic acid
EBI	–	European Bioinformatics Institute
EGA	–	European genome-phenome archive
ENCODE	–	ENCyclopedia Of Dna Elements
eQTL	–	Expression quantitative trait locus
eRNA	–	Enhancer RNA
FANTOM	–	Functional ANnoTation Of the Mammalian genome
FDR	–	False discovery rate
FN	–	False negative

FP	–	False positive
GAD	–	Genetic association database
GAV	–	Global-as-view
GCS	–	Genotype confidence Score
GeSID	–	German study on sudden infant death
GIM	–	Genetically influenced metabotype
GO	–	Gene ontology
GRC	–	Genome Reference Consortium
GSEA	–	Gene set enrichment analysis
GWAS	–	Genome-wide association study
GxE	–	Gene-by-environment
GxG	–	Gene-by-gene
HapMap	–	International Haplotype Map consortium
HTML	–	Hypertext markup language
HWE	–	Hardy-Weinberg equilibrium
IBD	–	Inflammatory bowel disease
ICR	–	Individual call rate
ID	–	Identifier
IPF	–	Idiopathic pulmonary fibrosis
JSON	–	JavaScript object notation
Kb, Mb, Gb	–	kilobases, megabases, gigabases; corresponds to 10^3, 10^6, and 10^9 nucleotides respectively
KORA	–	Kooperative Gesundheitsforschung in der Region Augsburg
KWT	–	Kruskal-Wallis test
LAV	–	Local-as-view
LCL	–	Lymphoblastoid cell lines
LD	–	Linkage disequilibrium
lncRNA	–	long non-coding RNA
LOD	–	Logarithm of the odds
LOF	–	Loss of function
LRR	–	Log R ration
LRT	–	Likelihood ratio test
MAF	–	Minor allele frequency

MDS – Multidimensional scaling
MeSH – Medical subject headings
MeSH – Medical Subject Headings
mGWAS – GWAS with metabolic traits
miRNA – microRNA
mQTL – Metabolite quantitative trait locus
MRE – miRNA recognition element
mRNA – Messenger RNA
MS – Mass spectrometry
MSUD – Maple syrup urine disease
MuTHER – Multiple tissue human expression resource
NCBI – National Center of Biotechnology Information
NGS – Next-generation sequencing
NHGRI – National Human Genome Research Institute
NMR – Nuclear magnetic resonance
OMIM – Online Mendelian Inheritance in Man
OR – Odds ratio
OWL – Web ontology language
piRNA – PIWI-interacting RNA
polyA – Polyadenylation
polyA sites – Polyadenylation sites
PPC – Pearson correlation coefficient
PPI – Protein-protein interactions
pQTL – Protein quantitative trait locus
PS – Psoriasis
PSA – Psoriatic arthritis
QC – Quality control
RBP – RNA-binding proteins
RBP – RNA-binding protein
RefSeq – NCBI reference sequence
RISC – RNA-induced silencing complex
RNA – Ribonucleic acid
rRNA – Ribosomal RNA

RS – Rejected substitutions
SCHC – Sheffield Children's Hospital SIDS Cohort
SHIP – Study of health in Pomerania
SIDS – Sudden infant death syndrome
siRNA – Small interfering RNA
SKAT – Sequence kernel association test
SLN – Shared locus network
SNP – Single nucleotide polymorphism
SNR – Signal-to-noise ratio
SNV – Single nucleotide variant
SUDI – Sudden unexpected death in infancy
SVN – Shared variant network
TF – Transcription factor
TFBS – Transcription factor binding site
TGCT – Testicular germ cell tumor
TN – True negative
TP – True positive
tRNA – Transfer RNA
TSS – Transcription start site
TSS – Transcription start site
TwinsUK – UK adult twin registry
uaRNA – Upstream antisense RNA
UCSC – University of California Santa Cruz
UTR – Untranslated region
VEP – Variant effect predictor
VTE – Venous thromboembolism
WES – whole-exome sequencing
WGS – whole-genome sequencing
WTSI – Wellcome Trust Sanger Institute
XML – Extensible markup language

References

1. Savolainen P, Zhang YP, Luo J, Lundeberg J, Leitner T: **Genetic evidence for an East Asian origin of domestic dogs**. *Science* 2002, **298**:1610–1613.
2. Lamarck J-B: *Zoological Philosophy*. 1809. Translated by Hugh Elliot. London: MacMillan and Co., Ltd.; 1914.
3. Darwin CR: *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life*. London: John Murray, Albemarle Street; 1859.
4. Mendel G: **Versuche über Pflanzen–Hybriden**. *Verhandlungen des Naturforschenden Vereins zu Brünn* 1866, **4**:3–47.
5. de Vries H: **Sur la loi de disjunction des hybrides**. *Comptes Rendus de l'Academie des Sciences* 1900, **130**:845–847.
6. Correns CE: **G**. **Mendel's Regel über das Verhalten der Nachkommenschaft der Rassenbastarde**. *Berichte der deutschen botanischen Gesellschaft* 1900, **18**:158–168.
7. von Tschermak E: **Ueber künstliche Kreuzung bei Pisum sativum**. *Berichte der Deutschen Botanischen Gesellschaft* 1900, **18**:232–249.
8. Miescher F: **Über die chemische Zusammensetzung der Eiterzellen**. *Medicinisch-chemische Untersuchungen* 1871, **4**:441–460.
9. Flemming W: **Beiträge zur Kenntniss der Zelle und Ihrer Lebenserscheinungen**. *Archiv für mikroskopische Anatomie* 1880, **18**:151–259.
10. von Waldeyer W: **Ueber Karyokinese und ihre Beziehungen zu den Befruchtungsvorgängen**. *Archiv für mikroskopische Anatomie* 1888, **32**:1–122.
11. Boveri T: **Zellenstudien II**. **Die Befruchtung und Teilung des Eies von Ascaris megalocephala**. *Jena Zeitschr Naturwiss* 1888, **22**:685–882.
12. Boveri T: **Zellenstudien III**: **Über das Verhalten der chromatischen Kernsubstanz bei der Bildung der Richtungskörper und bei der Befruchtung**. *Jena Zeitschr Naturwiss* 1890, **24**:314–401.
13. Johannsen W: *Elemente der exakten Erblichkeitslehre*. Jena: Verlag von Gustav Fischer; 1909.
14. Boveri T: **Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns**. *Verhandlungen der physikalisch-medizinischen Gesellschaft Würzburg* 1902, **35**:67–90.
15. Boveri T: *Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns*. Jena: Verlag von Gustav Fischer; 1904.
16. Sutton WS: **The chromosomes in heredity**. *The Biological Bulletin* 1903, **4**:231–250.
17. Garrod AE: *Inborn errors of metabolism*. London: Henry Frowde and Hodder & Stoughton; 1909.
18. Garrod AE: **The incidence of alkaptonuria**: **a study in chemical individuality**. *The Lancet* 1902, **160**:1616–1620.
19. Farabee WC: **Inheritance of digital malformations in man**. *Papers of the Peabody Museum of the American Archaeology and Ethnology* 1905, **3**:65–78.
20. Bateson W, Saunders ER, Punnett RC: **Experimental studies in the physiology of heredity**. *Reports to the Evolution Committee of the Royal Society* 1905, **2**:1–55, 80–99.
21. Morgan TH: **Sex–limited inheritance in Drosophila**. *Science* 1910, **32**:120–122.
22. Stevens NM: **Studies in spermatogenesis with especial reference to the "accessory chromosome"**. *Carnegie Institution of Washington Publication* 1905, **36**:1–33.
23. Wilson EB: **The chromosomes in relation to the determination of sex in insects**. *Science* 1905, **22**:500–502.
24. Morgan TH: **Random segregation versus coupling in Mendelian inheritance**. *Science* 1911, **34**:384.
25. Sturtevant AH: **The linear arrangement of six sex–linked factors in Drosophila**, **as shown by their mode of association**. *Journal of Experimental Zoology* 1913, **14**:43–59.
26. Haldane JBS: **The combination of linkage values**, **and the calculation of distance between the loci of linked factors**. *J Genet* 1919, **8**:299–309.
27. Galton F: **The average contribution of each several ancestor to the total heritage of the offspring**. *Proceedings of the Royal Society* 1897, **61**:401–413.

28. Pearson K: **Mathematical contributions to the theory of evolution**. **On the law of ancestral heredity**. *Proceedings of the Royal Society* 1898, **62**:386–412.
29. Yule GU: **Mendel's laws and their probable relations to intra–racial heredity**. *New Phytologist* 1902, **1**:193–207, 222–238.
30. Punnett RC: **Mendelism in Relation to Disease**. *Proceedings of the Royal Society of Medicine* 1908, **1**:135–168.
31. Weinberg W: **Über den Nachweis der Vererbung beim Menschen**. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg* 1908, **64**:368–382.
32. Hardy GH: **Mendelian proportions in a mixed population**. *Science* 1908, **28**:49–50.
33. Fisher RA: **The correlation between relatives on the supposition of Mendelian inheritance**. *Transactions of the Royal Society of Edinburgh* 1918, **52**:399–433.
34. Wright S: **Coefficients of inbreeding and relationship**. *The American Naturalist* 1922, **56**:330–338.
35. Tenesa A, Haley CS: **The heritability of human disease**: **estimation**, **uses and abuses**. *Nat Rev Genet* 2013, **14**:139–149.
36. Visscher PM, Hill WG, Wray NR: **Heritability in the genomics era––concepts and misconceptions**. *Nat Rev Genet* 2008, **9**:255–266.
37. Fisher RA: *The genetical theory of natural selection.* Oxford: Clarendon; 1930.
38. Edwards AWF: **The fundamental theorem of natural selection**. *Biological Reviews* 1994, **69**:443–474.
39. Avery OT, Macleod CM, McCarty M: **Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii**. *J Exp Med* 1944, **79**:137–158.
40. Kossel A: **Ueber eine neue Base aus dem Thierkörper**. *Ber Dtsch Chem Ges* 1885, **18**:79.
41. Kossel A: **Ueber das Thymin**, **ein Spaltungsproduct der Nucleinsäure**. *Ber Dtsch Chem Ges* 1893, **26**:2753–2756.
42. Kossel A: **Darstellung und Spaltungsprodukte der Nucleinsäure**. *Ber Dtsch Chem Ges* 1894, **27**:2215–2222.
43. Kossel A: **Ueber Guanin**. *Zeitschrift für Physiologische Chemie* 1883–84, **8**:404–410.
44. Levene PA, Bass L: *Nucleic Acids.* New York: The Chemical Catalog Co.; 1931.
45. Levene PA: **On the Chemistry of the Chromatin Substance of the Nerve Cell**. *J Med Res* 1903, **10**:204–211.
46. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid**. *Nature* 1953, **171**:737–738.
47. Kossel A: **Ueber einen peptonartigen Bestandtheil des Zellkerns**. *Zeitschrift für Physiologische Chemie* 1883–84, **8**:511–515.
48. Osborne TB, Harris IF: **Die Nucleinsäure des Weizenembryos**. *Zeitschrift für Physiologische Chemie* 1902, **36**:85–133.
49. Ascoli A: **Ueber ein neues Spaltungsprodukt des hefenucleins**. *Zeitschrift für Physiologische Chemie* 1900–01, **31**:161–164.
50. Hartley H: **Origin of the Word `Protein'**. *Nature* 1951, **168**:244–244.
51. Chittenden RH, Folin O, Gies WJ, Koch W, Osborne TB, Osborne TB, Levene PA, Mandel JA, Mathews AP, Mendel LB: **Joint Recommendations of the Physiological and Biochemical Committees on Protein Nomenclature**. *Science* 1908, **27**:554–556.
52. Crick FH: **On Protein Synthesis**. *Symposia of the Society for Experimental Biology* 1958, **12**:138–163.
53. Crick FH, Barnett L, Brenner S, Watts-Tobin RJ: **General nature of the genetic code for proteins**. *Nature* 1961, **192**:1227–1232.
54. Nirenberg M: **Historical review**: **Deciphering the genetic code––a personal account**. *Trends Biochem Sci* 2004, **29**:46–54.
55. Neel JV: **The Inheritance of Sickle Cell Anemia**. *Science* 1949, **110**:64–66.
56. Ingram VM: **Gene mutations in human haemoglobin**: **the chemical difference between normal and sickle cell haemoglobin**. *Nature* 1957, **180**:326–328.
57. Taliaferro WH, Huck JG: **The Inheritance of Sickle–Cell Anaemia in Man**. *Genetics* 1923, **8**:594–598.
58. Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmuller G: **SNiPA**: **an interactive**, **genetic variant–centered annotation browser**. *Bioinformatics* 2015, **31**:1334–1336.

59. Berget SM, Moore C, Sharp PA: **Spliced segments at the 5' terminus of adenovirus 2 late mRNA**. *Proc Natl Acad Sci U S A* 1977, **74**:3171-3175.
60. Chow LT, Gelinas RE, Broker TR, Roberts RJ: **An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA**. *Cell* 1977, **12**:1-8.
61. Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L: **Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways**. *Cell* 1980, **20**:313-319.
62. Morris KV, Mattick JS: **The rise of regulatory RNA**. *Nat Rev Genet* 2014, **15**:423-437.
63. Mitchell PJ, Tjian R: **Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins**. *Science* 1989, **245**:371-378.
64. Maxam AM, Gilbert W: **A new method for sequencing DNA**. *Proc Natl Acad Sci U S A* 1977, **74**:560-564.
65. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci U S A* 1977, **74**:5463-5467.
66. Edwards A, Voss H, Rice P, Civitello A, Stegemann J, Schwager C, Zimmermann J, Erfle H, Caskey CT, Ansorge W: **Automated DNA sequencing of the human HPRT locus**. *Genomics* 1990, **6**:593-608.
67. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.
68. International Human Genome Sequencing C: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**:931-945.
69. Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, et al.: **A genetic linkage map of the human genome**. *Cell* 1987, **51**:319-337.
70. NIH/CEPH Collaborative Mapping Group: **A comprehensive genetic linkage map of the human genome**. *Science* 1992, **258**:67-86.
71. Matise TC, Perlin M, Chakravarti A: **Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map**. *Nat Genet* 1994, **6**:384-390.
72. International HapMap Consortium: **The International HapMap Project**. *Nature* 2003, **426**:789-796.
73. International HapMap Consortium: **A haplotype map of the human genome**. *Nature* 2005, **437**:1299-1320.
74. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, et al: **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2007, **449**:851-861.
75. International HapMap Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, et al: **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, **467**:52-58.
76. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat Biotechnol* 2008, **26**:1135-1145.
77. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061-1073.
78. Encode Project Consortium: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**:57-74.
79. Fantom Consortium and the Riken PMI and CLST (DGT), Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Lassmann T, Itoh M, Summers KM, Suzuki H, Daub CO, Kawai J, Heutink P, Hide W, Freeman TC, Lenhard B, Bajic VB, Taylor MS, Makeev VJ, Sandelin A, Hume DA, Carninci P, Hayashizaki Y: **A promoter-level mammalian expression atlas**. *Nature* 2014, **507**:462-470.
80. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jorgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, et al: **An atlas of active enhancers across human cell types and tissues**. *Nature* 2014, **507**:455-461.

81. Encode Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, et al: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**:799-816.
82. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, et al: **Integrative analysis of 111 reference human epigenomes**. *Nature* 2015, **518**:317-330.
83. Hsu TC: **Mammalian chromosomes in vitro I. The karyotype of man**. *J Hered* 1952, **43**:167-172.
84. Strupp C, Hildebrandt B, Germing U, Haas R, Gattermann N: **Cytogenetic response to thalidomide treatment in three patients with myelodysplastic syndrome**. *Leukemia* 2003, **17**:1200-1202.
85. Tjio JH, Levan A: **The chromosome number of man**. *Hereditas* 1956, **42**:1-6.
86. Bickmore WA: **Karyotype Analysis and Chromosome Banding**. In *eLS*. John Wiley & Sons, Ltd; 2001
87. Lejeune J, Gauthier M, Turpin R: **Les chromosomes humains en culture de tissus**. *C R Hebd Seances Acad Sci* 1959, **248**:602-603.
88. Jacobs PA, Baikie AG, Court Brown WM, Strong JA: **The somatic chromosomes in mongolism**. *Lancet* 1959, **1**:710.
89. Jacobs PA, Strong JA: **A Case of Human Intersexuality Having a Possible XXY Sex-Determining Mechanism**. *Nature* 1959, **183**:302-303.
90. Morton NE: **Sequential tests for the detection of linkage**. *Am J Hum Genet* 1955, **7**:277-318.
91. Sklar P: **Linkage analysis in psychiatric disorders: the emerging picture**. *Annu Rev Genomics Hum Genet* 2002, **3**:371-413.
92. Dunnill MG, Richards AJ, Milana G, Mollica F, Atherton D, Winship I, Farrall M, al-Imara L, Eady RA, Pope FM: **Genetic linkage to the type VII collagen gene (COL7A1) in 26 families with generalised recessive dystrophic epidermolysis bullosa and anchoring fibril abnormalities**. *J Med Genet* 1994, **31**:745-748.
93. Risch N, Merikangas K: **The future of genetic studies of complex human diseases**. *Science* 1996, **273**:1516-1517.
94. Lander ES: **The new genomics: global views of biology**. *Science* 1996, **274**:536-539.
95. Chakravarti A: **Population genetics--making sense out of sequence**. *Nat Genet* 1999, **21**:56-60.
96. Delaneau O, Marchini J, Zagury JF: **A linear complexity phasing method for thousands of genomes**. *Nat Methods* 2012, **9**:179-181.
97. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR: **Fast and accurate genotype imputation in genome-wide association studies through pre-phasing**. *Nat Genet* 2012, **44**:955-959.
98. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: **Complement factor H polymorphism in age-related macular degeneration**. *Science* 2005, **308**:385-389.
99. Wjst M, Sargurupremraj M, Arnold M: **Genome-wide association studies in asthma: what they really told us about pathogenesis**. *Curr Opin Allergy Clin Immunol* 2013, **13**:112-118.
100. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *Proc Natl Acad Sci U S A* 2009, **106**:9362-9367.
101. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X: **Sequence kernel association tests for the combined effect of rare and common variants**. *Am J Hum Genet* 2013, **92**:841-853.
102. Linderman MD, Brandt T, Edelmann L, Jabado O, Kasai Y, Kornreich R, Mahajan M, Shah H, Kasarskis A, Schadt EE: **Analytical validation of whole exome and whole genome sequencing for clinical applications**. *BMC Med Genomics* 2014, **7**:20.
103. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore

AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**:747-753.

104. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: **Rare variants create synthetic genome-wide associations**. *PLoS Biol* 2010, **8**:e1000294.
105. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE: **Rare-disease genetics in the era of next-generation sequencing: discovery to translation**. *Nat Rev Genet* 2013, **14**:681-691.
106. Online Mendelian Inheritance in Man (OMIM®) [http://omim.org/ - accessed: 02/27/2014]. Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University; 1966-2014.
107. Orphanet encyclopedia: **Orphanet encyclopedia**. 03/2014 edition. http://orpha.net/).
108. Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, Chuzhanova N, Krawczak M, Kehrer-Sawatzki H, Stenson PD: **Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics**. *Hum Mutat* 2010, **31**:631-655.
109. Baird PA, Anderson TW, Newcombe HB, Lowry RB: **Genetic disorders in children and young adults: a population study**. *Am J Hum Genet* 1988, **42**:677-693.
110. Peltonen L, McKusick VA: **Genomics and medicine. Dissecting human disease in the postgenomic era**. *Science* 2001, **291**:1224-1229.
111. Badano JL, Katsanis N: **Beyond Mendel: an evolving view of human genetic disease transmission**. *Nat Rev Genet* 2002, **3**:779-789.
112. Visscher PM, Brown MA, McCarthy MI, Yang J: **Five years of GWAS discovery**. *Am J Hum Genet* 2012, **90**:7-24.
113. Polychronakos C, Li Q: **Understanding type 1 diabetes through genetics: advances and prospects**. *Nat Rev Genet* 2011, **12**:781-792.
114. Witte JS, Visscher PM, Wray NR: **The contribution of genetic variants to disease depends on the ruler**. *Nat Rev Genet* 2014, **15**:765-776.
115. Buil A, Brown AA, Lappalainen T, Viñuela A, Davies MN, Zheng HF, Richards JB, Glass D, Small KS, Durbin R, Spector TD, Dermitzakis ET: **Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins**. *Nat Genet* 2015, **47**:88-91.
116. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, Arnold M, Erte I, Forgetta V, Yang TP, Walter K, Menni C, Chen L, Vasquez L, Valdes AM, Hyde CL, Wang V, Ziemek D, Roberts P, Xi L, Grundberg E, Multiple Tissue Human Expression Resource C, Waldenberger M, Richards JB, Mohney RP, Milburn MV, John SL, Trimmer J, Theis FJ, Overington JP, et al: **An atlas of genetic influences on human blood metabolites**. *Nat Genet* 2014, **46**:543-550.
117. Gibson G: **Rare and common variants: twenty arguments**. *Nat Rev Genet* 2011, **13**:135-145.
118. Zuk O, Hechter E, Sunyaev SR, Lander ES: **The mystery of missing heritability: Genetic interactions create phantom heritability**. *Proc Natl Acad Sci U S A* 2012, **109**:1193-1198.
119. Hemani G, Knott S, Haley C: **An evolutionary perspective on epistasis and the missing heritability**. *PLoS Genet* 2013, **9**:e1003295.
120. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, McRae AF, Yang J, Gibson G, Martin NG, Metspalu A, Franke L, Montgomery GW, Visscher PM, Powell JE: **Detection and replication of epistasis influencing transcription in humans**. *Nature* 2014, **508**:249-253.
121. Manolio TA: **Bringing genome-wide association findings into clinical use**. *Nat Rev Genet* 2013, **14**:549-558.
122. Hogeweg P: **The roots of bioinformatics in theoretical biology**. *PLoS Comput Biol* 2011, **7**:e1002021.
123. Randall JC, Winkler TW, Kutalik Z, Berndt SI, Jackson AU, Monda KL, Kilpeläinen TO, Esko T, Mägi R, Li S, Workalemahu T, Feitosa MF, Croteau-Chonka DC, Day FR, Fall T, Ferreira T, Gustafsson S, Locke AE, Mathieson I, Scherag A, Vedantam S, Wood AR, Liang L, Steinthorsdottir V, Thorleifsson G, Dermitzakis ET, Dimas AS, Karpe F, Min JL, Nicholson G, et al: **Sex-stratified Genome-wide Association Studies Including 270,000 Individuals Show Sexual Dimorphism in Genetic Loci for Anthropometric Traits**. *PLoS Genet* 2013, **9**:e1003500.
124. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan Ja, Magi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, Weyant RJ, Segre AV, Estrada K, Liang L, Nemesh J, Park J-H,

Gustafsson S, Kilpelainen TO, et al: **Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index**. *Nat Genet* 2010, **42**:937-948.

125. Cohorts for Heart Aging Research in Genetic Epidemiology Consortium: **Whole-genome sequence-based analysis of high-density lipoprotein cholesterol**. *Nat Genet* 2013, **45**:899-901.
126. Goldstein DB: **The importance of synthetic associations will only be resolved empirically**. *PLoS Biol* 2011, **9**:e1001008.
127. Wray NR, Purcell SM, Visscher PM: **Synthetic associations created by rare variants do not explain most GWAS results**. *PLoS Biol* 2011, **9**:e1000579.
128. Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, Barker JN, Barrett JC, Blackburn H, Brand O, Burren O, Capon F, Compston A, Gough SC, Jostins L, Kong Y, Lee JC, Lek M, MacArthur DG, Mansfield JC, Mathew CG, Mein CA, Mirza M, Nutland S, Onengut-Gumuscu S, Papouli E, Parkes M, Rich SS, Sawcer S, Satsangi J, Simmonds MJ, et al: **Negligible impact of rare autoimmune-locus coding-region variants on missing heritability**. *Nature* 2013, **498**:232-235.
129. Saunders EJ, Dadaev T, Leongamornlert DA, Jugurnauth-Little S, Tymrakiewicz M, Wiklund F, Al Olama AA, Benlloch S, Neal DE, Hamdy FC, Donovan JL, Giles GG, Severi G, Gronberg H, Aly M, Haiman CA, Schumacher F, Henderson BE, Lindstrom S, Kraft P, Hunter DJ, Gapstur S, Chanock S, Berndt SI, Albanes D, Andriole G, Schleutker J, Weischer M, Nordestgaard BG, Canzian F, et al: **Fine-mapping the HOXB region detects common variants tagging a rare coding allele: evidence for synthetic association in prostate cancer**. *PLoS Genet* 2014, **10**:e1004129.
130. Clayton DG: **Prediction and interaction in complex disease genetics: experience in type 1 diabetes**. *PLoS Genet* 2009, **5**:e1000540.
131. Pryce JE, Hayes BJ, Bolormaa S, Goddard ME: **Polymorphic regions affecting human height also control stature in cattle**. *Genetics* 2011, **187**:981-984.
132. Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, Pistis G, Serbanovic-Canic J, Elling U, Goodall AH, Labrune Y, Lopez LM, Magi R, Meacham S, Okada Y, Pirastu N, Sorice R, Teumer A, Voss K, Zhang W, Ramirez-Solis R, Bis JC, Ellinghaus D, Gogele M, Hottenga JJ, Langenberg C, Kovacs P, O'Reilly PF, Shin SY, Esko T, Hartiala J, et al: **New gene functions in megakaryopoiesis and platelet formation**. *Nature* 2011, **480**:201-208.
133. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, Mooser V: **Use of genome-wide association studies for drug repositioning**. *Nat Biotechnol* 2012, **30**:317-320.
134. Daly AK: **Genome-wide association studies in pharmacogenomics**. *Nat Rev Genet* 2010, **11**:241-246.
135. Papp KA, Langley RG, Lebwohl M, Krueger GG, Szapary P, Yeilding N, Guzzo C, Hsu MC, Wang Y, Li S, Dooley LT, Reich K, investigators Ps: **Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 52-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 2)**. *Lancet* 2008, **371**:1675-1684.
136. Cargill M, Schrodi SJ, Chang M, Garcia VE, Brandon R, Callis KP, Matsunami N, Ardlie KG, Civello D, Catanese JJ, Leong DU, Panko JM, McAllister LB, Hansen CB, Papenfuss J, Prescott SM, White TJ, Leppert MF, Krueger GG, Begovich AB: **A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes**. *Am J Hum Genet* 2007, **80**:273-290.
137. Kauffman CL, Aria N, Toichi E, McCormick TS, Cooper KD, Gottlieb AB, Everitt DE, Frederick B, Zhu Y, Graham MA, Pendley CE, Mascelli MA: **A phase I study evaluating the safety, pharmacokinetics, and clinical response of a human IL-12 p40 antibody in subjects with plaque psoriasis**. *J Invest Dermatol* 2004, **123**:1037-1044.
138. Kopp T, Riedl E, Bangert C, Bowman EP, Greisenegger E, Horowitz A, Kittler H, Blumenschein WM, McClanahan TK, Marbury T, Zachariae C, Xu D, Hou XS, Mehta A, Zandvliet AS, Montgomery D, van Aarle F, Khalilieh S: **Clinical improvement in psoriasis with specific targeting of interleukin-23**. *Nature* 2015.
139. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic Acids Res* 2010, **38**:e164.
140. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor**. *Bioinformatics* 2010, **26**:2069-2070.

141. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3**. *Fly (Austin)* 2012, **6**:80-92.
142. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**:3812-3814.
143. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**:248-249.
144. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA, Hatan M, Carrasco-Alfonso MJ, Mayer D, Luckey CJ, Patsopoulos NA, De Jager PL, Kuchroo VK, Epstein CB, Daly MJ, Hafler DA, Bernstein BE: **Genetic and epigenetic fine mapping of causal autoimmune disease variants**. *Nature* 2015, **518**:337-343.
145. Gieger C, Geistlinger L, Altmaier E, Hrabe de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J, Illig T, Suhre K: **Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum**. *PLoS Genet* 2008, **4**:e1000282.
146. Kastenmuller G, Raffler J, Gieger C, Suhre K: **Genetics of human metabolism: an update**. *Hum Mol Genet* 2015.
147. Lenzerini M: **Data integration: a theoretical perspective**. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. pp. 233-246. Madison, Wisconsin: ACM; 2002:233-246.
148. Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P: **Data integration and genomic medicine**. *J Biomed Inform* 2007, **40**:5-16.
149. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kahari AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, et al: **Ensembl 2014**. *Nucleic Acids Res* 2014, **42**:D749-755.
150. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P, Maglott DR, Murphy TD, Ostell JM: **RefSeq: an update on mammalian reference sequences**. *Nucleic Acids Res* 2014, **42**:D756-763.
151. Kuhn RM, Haussler D, Kent WJ: **The UCSC genome browser and associated tools**. *Brief Bioinform* 2013, **14**:144-161.
152. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, et al: **GENCODE: the reference human genome annotation for The ENCODE Project**. *Genome Res* 2012, **22**:1760-1774.
153. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, et al: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes**. *Genome Res* 2009, **19**:1316-1323.
154. Gray J: **The Transaction Concept: Virtues and Limitations**. *Proceedings of Seventh International Conference on Very Large Databases* 1981.
155. Härder T, Reuter A: **Principles of transaction-oriented database recovery**. *ACM Comput Surv* 1983, **15**:287-317.
156. Gruber TR: **A translation approach to portable ontology specifications**. *Knowl Acquis* 1993, **5**:199-220.
157. Cattell R: **Scalable SQL and NoSQL data stores**. *SIGMOD Rec* 2011, **39**:12-27.
158. Bellman R: *Adaptive control processes - A guided tour*. Princeton University Press; 1961.
159. Smith GD, Ebrahim S: **'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?** *Int J Epidemiol* 2003, **32**:1-22.

160. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G: **Mendelian randomization: using genes as instruments for making causal inferences in epidemiology**. *Stat Med* 2008, **27**:1133-1163.
161. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WO, Consortium G: **A large-scale, consortium-based genomewide association study of asthma**. *N Engl J Med* 2010, **363**:1211-1221.
162. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D: **Methods of integrating data to uncover genotype-phenotype interactions**. *Nat Rev Genet* 2015, **16**:85-97.
163. Arnold M, Ellwanger DC, Hartsperger ML, Pfeufer A, Stumpflen V: **Cis-acting polymorphisms affect complex traits through modifications of microRNA regulation pathways**. *PLoS One* 2012, **7**:e36694.
164. Arnold M, Hartsperger ML, Baurecht H, Rodriguez E, Wachinger B, Franke A, Kabesch M, Winkelmann J, Pfeufer A, Romanos M, Illig T, Mewes HW, Stumpflen V, Weidinger S: **Network-based SNP meta-analysis identifies joint and disjoint genetic features across common human diseases**. *BMC Genomics* 2012, **13**:490.
165. Raffler J, Friedrich N, Arnold M, Kacprowski T, Rueedi R, Altmaier E, Bergmann S, Budde K, Gieger C, Homuth G, Pietzner M, Romisch-Margl W, Strauch K, Volzke H, Waldenberger M, Wallaschofski H, Nauck M, Volker U, Kastenmuller G, Suhre K: **Genome-Wide Association Study with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality**. *PLoS Genet* 2015, **11**:e1005487.
166. Findeisen M, Vennemann M, Brinkmann B, Ortmann C, Rose I, Kopcke W, Jorch G, Bajanowski T: **German study on sudden infant death (GeSID): design, epidemiological and pathological profile**. *Int J Legal Med* 2004, **118**:163-169.
167. Prtak L, Al-Adnani M, Fenton P, Kudesia G, Cohen MC: **Contribution of bacteriology and virology in sudden unexpected death in infancy**. *Arch Dis Child* 2010, **95**:371-376.
168. Scheimberg I, Ashal H, Kotiloglu-Karaa E, French P, Kay P, Cohen MC: **Weight charts of infants dying of sudden infant death in England**. *Pediatr Dev Pathol* 2014, **17**:271-277.
169. Holle R, Happich M, Lowel H, Wichmann HE, MONICA/KORA Study Group: **KORA--a research platform for population based health research**. *Gesundheitswesen* 2005, **67 Suppl 1**:S19-25.
170. Moayyeri A, Hammond CJ, Hart DJ, Spector TD: **The UK Adult Twin Registry (TwinsUK Resource)**. *Twin Res Hum Genet* 2013, **16**:144-149.
171. John U, Greiner B, Hensel E, Ludemann J, Piek M, Sauer S, Adam C, Born G, Alte D, Greiser E, Haertel U, Hense HW, Haerting J, Willich S, Kessler C: **Study of Health In Pomerania (SHIP): a health examination survey in an east German region: objectives and design**. *Soz Praventivmed* 2001, **46**:186-194.
172. Volzke H, Alte D, Schmidt CO, Radke D, Lorbeer R, Friedrich N, Aumann N, Lau K, Piontek M, Born G, Havemann C, Ittermann T, Schipf S, Haring R, Baumeister SE, Wallaschofski H, Nauck M, Frick S, Arnold A, Junger M, Mayerle J, Kraft M, Lerch MM, Dorr M, Reffelmann T, Empen K, Felix SB, Obst A, Koch B, Glaser S, et al: **Cohort profile: the study of health in Pomerania**. *Int J Epidemiol* 2011, **40**:294-307.
173. Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S: **PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships**. *Community Genet* 2006, **9**:55-61.
174. Nothlings U, Krawczak M: **PopGen. A population-based biobank with prospective follow-up of a control group**. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2012, **55**:831-835.
175. Alonso A, Rodriguez MA, Vinaixa M, Tortosa R, Correig X, Julia A, Marsal S: **Focus: a robust workflow for one-dimensional NMR spectral analysis**. *Anal Chem* 2014, **86**:1160-1169.
176. Wyss M, Kaddurah-Daouk R: **Creatine and creatinine metabolism**. *Physiol Rev* 2000, **80**:1107-1213.
177. Schramm CF: **A KNIME-based tool for automatizing genome-wide association analysis workflows**. Bachelor thesis; Munich; 2013.
178. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res* 2001, **29**:308-311.
179. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet* 2007, **81**:559-575.

180. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, Jarvik G, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto M, McCarty CA, McDavid AN, Mirel DB, Paschall JE, Pugh EW, Rasmussen LV, Wilke RA, Zuvich RL, Ritchie MD: **Quality control procedures for genome-wide association studies**. *Curr Protoc Hum Genet* 2011, **Chapter 1**:Unit1 19.
181. Therneau TM, Grambsch PM: **Modeling Survival Data: Extending the Cox Model**. *Stat Med* 2001, **20**:2053-2054.
182. Urbanek S: **Rserve: Binary R server**. 2013.
183. Delaneau O, Zagury JF, Marchini J: **Improved whole-chromosome phasing for disease and population genetic studies**. *Nat Methods* 2013, **10**:5-6.
184. Freeman C, Marchini J: **GTOOL**. http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html 2007.
185. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets**. *Gigascience* 2015, **4**:7.
186. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams Jr. RM: *Adjustment during Army Life*. Princeton, NJ: Princeton University Press; 1949.
187. DerSimonian R, Laird N: **Meta-analysis in clinical trials**. *Control Clin Trials* 1986, **7**:177-188.
188. Scharpf RB, Irizarry RA, Ritchie ME, Carvalho B, Ruczinski I: **Using the R Package crlmm for Genotyping and Copy Number Estimation**. *J Stat Softw* 2011, **40**:1-32.
189. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**:R80.
190. R core team: **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing; 2013.
191. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M: **PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data**. *Genome Res* 2007, **17**:1665-1674.
192. Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I: **Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays**. *Ann Appl Stat* 2008, **2**:687-713.
193. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, O'Hara R, Casalunovo T, Conlin LK, D'Arcy M, Frackelton EC, Geiger EA, Haldeman-Englert C, Imielinski M, Kim CE, Medne L, Annaiah K, Bradfield JP, Dabaghyan E, Eckert A, Onyiah CC, Ostapenko S, Otieno FG, Santa E, Shaner JL, Skraban R, Smith RM, Elia J, Goldmuntz E, Spinner NB, et al: **High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications**. *Genome Res* 2009, **19**:1682-1690.
194. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW: **The Database of Genomic Variants: a curated collection of structural variation in the human genome**. *Nucleic Acids Res* 2014, **42**:D986-992.
195. Salvi E, Kutalik Z, Glorioso N, Benaglio P, Frau F, Kuznetsova T, Arima H, Hoggart C, Tichet J, Nikitin YP, Conti C, Seidlerova J, Tikhonoff V, Stolarz-Skrzypek K, Johnson T, Devos N, Zagato L, Guarrera S, Zaninello R, Calabria A, Stancanelli B, Troffa C, Thijs L, Rizzi F, Simonova G, Lupoli S, Argiolas G, Braga D, D'Alessio MC, Ortu MF, et al: **Genomewide association study using a high-density single nucleotide polymorphism array and case-control design identifies a novel essential hypertension susceptibility locus in the promoter region of endothelial NO synthase**. *Hypertension* 2012, **59**:248-255.
196. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP: **DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources**. *Am J Hum Genet* 2009, **84**:524-533.
197. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool**. *Nucleic Acids Res* 2004, **32**:D493-D496.
198. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, **35**:D61-D65.

199. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**:1034–1050.
200. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: **Identifying a high fraction of the human genome to be under selective constraint using GERP++**. *PLoS Comput Biol* 2010, **6**:e1001025.
201. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nat Genet* 2014, **46**:310–315.
202. Lee JY, Yeh I, Park JY, Tian B: **PolyADB 2: mRNA polyadenylation sites in vertebrate genes**. *Nucleic Acids Res* 2007, **35**:D165–D168.
203. Zhang H, Hu J, Recce M, Tian B: **PolyADB: a database for mammalian mRNA polyadenylation**. *Nucleic Acids Res* 2005, **33**:D116–D120.
204. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D: **Patterns of variant polyadenylation signal usage in human genes**. *Genome Res* 2000, **10**:1001–1010.
205. Griffiths-Jones S: **The microRNA Registry**. *Nucleic Acids Res* 2004, **32**:D109–D111.
206. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature**. *Nucleic Acids Res* 2006, **34**:D140–D144.
207. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics**. *Nucleic Acids Res* 2008, **36**:D154–D158.
208. Ellwanger DC, Buttner FA, Mewes HW, Stumpflen V: **The sufficient minimal set of miRNA seed types**. *Bioinformatics* 2011, **27**:1346–1350.
209. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, et al: **The accessible chromatin landscape of the human genome**. *Nature* 2012, **489**:75–82.
210. Li JH, Liu S, Zhou H, Qu LH, Yang JH: **starBase v2.0: decoding miRNA–ceRNA, miRNA–ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data**. *Nucleic Acids Res* 2014, **42**:D92–97.
211. GTEx Consortium: **Human genomics. The Genotype–Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans**. *Science* 2015, **348**:648–660.
212. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, Eleftheriadis M, Sinning CR, Schnabel RB, Lubos E, Mennerich D, Rust W, Perret C, Proust C, Nicaud V, Loscalzo J, Hubner N, Tregouet D, Munzel T, Ziegler A, Tiret L, Blankenberg S, Cambien F: **Genetics and beyond--the transcriptome of human monocytes and disease susceptibility**. *PLoS One* 2010, **5**:e10693.
213. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A, Nisbett J, Sekowska M, Wilk A, Shin SY, Glass D, Travers M, Min JL, Ring S, Ho K, Thorleifsson G, Kong A, Thorsteindottir U, Ainali C, Dimas AS, Hassanali N, Ingle C, Knowles D, Krestyaninova M, Lowe CE, Di Meglio P, et al: **Mapping cis- and trans-regulatory effects across multiple tissues in twins**. *Nat Genet* 2012, **44**:1084–1089.
214. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JE, Zhernakova A, Zhernakova DV, Veldink JH, Van den Berg LH, Karjalainen J, Withoff S, Uitterlinden AG, Hofman A, Rivadeneira F, t Hoen PA, Reinmaa E, Fischer K, Nelis M, Milani L, Melzer D, Ferrucci L, Singleton AB, Hernandez DG, Nalls MA, Homuth G, et al: **Systematic identification of trans eQTLs as putative drivers of known disease associations**. *Nat Genet* 2013, **45**:1238–1243.
215. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, Ellis P, Langford C, Vannberg FO, Knight JC: **Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles**. *Nat Genet* 2012, **44**:502–510.
216. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2010**. *Nucleic Acids Res* 2010, **38**:D613–619.
217. Xia K, Shabalin AA, Huang S, Madar V, Zhou YH, Wang W, Zou F, Sun W, Sullivan PF, Wright FA: **seeQTL: a searchable database for human eQTLs**. *Bioinformatics* 2012, **28**:451–452.

218. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, Holmans P, Heward CB, Reiman EM, Stephan D, Hardy J: **A survey of genetic human cortical gene expression**. *Nat Genet* 2007, **39**:1494-1499.
219. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO: **A genome-wide association study of global gene expression**. *Nat Genet* 2007, **39**:1202-1207.
220. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, Ramirez J, Liu W, Lin YS, Moloney C, Aldred SF, Trinklein ND, Schuetz E, Nickerson DA, Thummel KE, Rieder MJ, Rettie AE, Ratain MJ, Cox NJ, Brown CD: **Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue**. *PLoS Genet* 2011, **7**:e1002078.
221. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS: **DrugBank 4.0: shedding new light on drug metabolism**. *Nucleic Acids Res* 2014, **42**:D1091-1097.
222. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations**. *Nucleic Acids Res* 2014, **42**:D1001-1006.
223. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: **A navigator for human genome epidemiology**. *Nat Genet* 2008, **40**:124-125.
224. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN: **The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine**. *Hum Genet* 2014, **133**:1-9.
225. UniProt Consortium: **Activities at the Universal Protein Resource (UniProt)**. *Nucleic Acids Res* 2014, **42**:D191-198.
226. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST: **The NCBI dbGaP database of genotypes and phenotypes**. *Nat Genet* 2007, **39**:1181-1186.
227. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR: **ClinVar: public archive of relationships among sequence variation and human phenotype**. *Nucleic Acids Res* 2014, **42**:D980-985.
228. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database**. *Nat Genet* 2004, **36**:431-432.
229. Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ: **PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes**. *Genome Biol* 2010, **11**:R6.
230. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y: **miR2Disease: a manually curated database for microRNA deregulation in human disease**. *Nucleic Acids Res* 2009, **37**:D98-104.
231. The 1000 Genomes Project Consortium: **A global reference for human genetic variation**. *Nature* 2015, **526**:68-74.
232. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW: **SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap**. *Bioinformatics* 2008, **24**:2938-2939.
233. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**:44-57.
234. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 2009, **37**:1-13.
235. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**:15545-15550.
236. Gene Ontology Consortium: **The Gene Ontology in 2010: extensions and refinements**. *Nucleic Acids Res* 2010, **38**:D331-335.
237. Lipscomb CE: **Medical Subject Headings (MeSH)**. *Bull Med Libr Assoc* 2000, **88**:265-266.

238. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K, Hermjakob H: **The IntAct molecular interaction database in 2010**. *Nucleic Acids Res* 2010, **38**:D525-531.

239. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW: **CORUM: the comprehensive resource of mammalian protein complexes--2009**. *Nucleic Acids Res* 2010, **38**:D497-501.

240. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 1999, **27**:29-34.

241. Hao T, Ma HW, Zhao XM, Goryanin I: **Compartmentalization of the Edinburgh Human Metabolic Network**. *BMC Bioinformatics* 2010, **11**:393.

242. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, Thorleifsson SG, Agren R, Bolling C, Bordel S, Chavali AK, Dobson P, Dunn WB, Endler L, Hala D, Hucka M, Hull D, Jameson D, Jamshidi N, Jonsson JJ, Juty N, Keating S, Nookaew I, Le Novere N, Malys N, Mazein A, et al: **A community-driven global reconstruction of human metabolism**. *Nat Biotechnol* 2013, **31**:419-425.

243. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P: **The Reactome pathway knowledgebase**. *Nucleic Acids Res* 2014, **42**:D472-477.

244. Schomburg I, Chang A, Placzek S, Sohngen C, Rother M, Lang M, Munaretto C, Ulas S, Stelzer M, Grote A, Scheer M, Schomburg D: **BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA**. *Nucleic Acids Res* 2013, **41**:D764-772.

245. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, et al: **HMDB: the Human Metabolome Database**. *Nucleic Acids Res* 2007, **35**:D521-526.

246. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP: **ChEMBL: a large-scale bioactivity database for drug discovery**. *Nucleic Acids Res* 2012, **40**:D1100-1107.

247. Yang TP, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE, Deloukas P, Dermitzakis ET: **Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies**. *Bioinformatics* 2010, **26**:2474-2476.

248. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, Abecasis GR, Barrett JC, Behrens T, Cho J, De Jager PL, Elder JT, Graham RR, Gregersen P, Klareskog L, Siminovitch KA, van Heel DA, Wijmenga C, Worthington J, Todd JA, Hafler DA, Rich SS, Daly MJ, Consortia FONo: **Pervasive sharing of genetic effects in autoimmune disease**. *PLoS Genet* 2011, **7**:e1002254.

249. Reese MG, Eeckman FH, Kulp D, Haussler D: **Improved splice site detection in Genie**. *J Comput Biol* 1997, **4**:311-323.

250. Hofacker IL: **Vienna RNA secondary structure server**. *Nucleic Acids Res* 2003, **31**:3429-3431.

251. Durinck S, Bullard J, Spellman PT, Dudoit S: **GenomeGraphs: integrated genomic data visualization with R**. *BMC Bioinformatics* 2009, **10**:2.

252. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, **27**:431-432.

253. John Morrison WJG: **Quanto**. http://biostats.usc.edu/Quanto.html; 2009.

254. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis G: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**:2156-2158.

255. Li H: **Tabix: fast retrieval of sequence features from generic TAB-delimited files**. *Bioinformatics* 2011, **27**:718-719.

256. Dong J, Horvath S: **Understanding network concepts in modules**. *BMC Syst Biol* 2007, **1**:24.

257. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE: **A human protein-protein interaction network: a resource for annotating the proteome**. *Cell* 2005, **122**:957-968.
258. Cochran WG: **The comparison of percentages in matched samples**. *Biometrika* 1950, **37**:256-266.
259. Higgins JP, Thompson SG: **Quantifying heterogeneity in a meta-analysis**. *Stat Med* 2002, **21**:1539-1558.
260. Browman G, Hebert PC, Coutts J, Stanbrook MB, Flegel K, MacDonald NE: **Personalized medicine: a windfall for science, but what about patients?** *CMAJ* 2011, **183**:E1277.
261. Willinger M, James LS, Catz C: **Defining the sudden infant death syndrome (SIDS): deliberations of an expert panel convened by the National Institute of Child Health and Human Development**. *Pediatr Pathol* 1991, **11**:677-684.
262. Krous HF, Beckwith JB, Byard RW, Rognum TO, Bajanowski T, Corey T, Cutz E, Hanzlick R, Keens TG, Mitchell EA: **Sudden infant death syndrome and unclassified sudden infant deaths: a definitional and diagnostic approach**. *Pediatrics* 2004, **114**:234-238.
263. Filiano JJ, Kinney HC: **A perspective on neuropathologic findings in victims of the sudden infant death syndrome: the triple-risk model**. *Biol Neonate* 1994, **65**:194-197.
264. Courts C, Madea B: **Genetics of the sudden infant death syndrome**. *Forensic Sci Int* 2010, **203**:25-33.
265. Weese-Mayer DE, Ackerman MJ, Marazita ML, Berry-Kravis EM: **Sudden Infant Death Syndrome: review of implicated genetic factors**. *Am J Med Genet A* 2007, **143A**:771-788.
266. Ferrante L, Opdal SH: **Sudden infant death syndrome and the genetics of inflammation**. *Front Immunol* 2015, **6**:63.
267. Guntheroth WG, Spiers PS: **The triple risk hypotheses in sudden infant death syndrome**. *Pediatrics* 2002, **110**:e64.
268. Thach BT: **Potential Central Nervous System Involvement in Sudden Unexpected Infant Deaths and the Sudden Infant Death Syndrome**. *Compr Physiol* 2015, **5**:1061-1068.
269. la Grange H, Verster J, Dempers JJ, de Beer C: **Review of immunological and virological aspects as contributory factors in Sudden Unexpected Death in Infancy (SUDI)**. *Forensic Sci Int* 2014, **245C**:12-16.
270. Van Norstrand DW, Ackerman MJ: **Genomic risk factors in sudden infant death syndrome**. *Genome Med* 2010, **2**:86.
271. Poetsch M, Todt R, Vennemann M, Bajanowski T: **That's not it, either-neither polymorphisms in PHOX2B nor in MIF are involved in sudden infant death syndrome (SIDS)**. *Int J Legal Med* 2015.
272. Fard D, Laer K, Rothamel T, Schurmann P, Arnold M, Cohen M, Vennemann M, Pfeiffer H, Bajanowski T, Pfeufer A, Dork T, Klintschar M: **Candidate gene variants of the immune system and sudden infant death syndrome**. *Int J Legal Med* 2016, **130**:1025-1033.
273. Zhang D, Qian Y, Akula N, Alliey-Rodriguez N, Tang J, Bipolar Genome S, Gershon ES, Liu C: **Accuracy of CNV Detection from GWAS Data**. *PLoS One* 2011, **6**:e14511.
274. Sullivan FM, Barlow SM: **Review of risk factors for sudden infant death syndrome**. *Paediatr Perinat Epidemiol* 2001, **15**:144-200.
275. Koike T, Izumikawa T, Sato B, Kitagawa H: **Identification of phosphatase that dephosphorylates xylose in the glycosaminoglycan-protein linkage region of proteoglycans**. *J Biol Chem* 2014, **289**:6695-6708.
276. Iozzo RV, Schaefer L: **Proteoglycan form and function: A comprehensive nomenclature of proteoglycans**. *Matrix Biol* 2015, **42**:11-55.
277. Vieira NM, Naslavsky MS, Licinio L, Kok F, Schlesinger D, Vainzof M, Sanchez N, Kitajima JP, Gal L, Cavacana N, Serafini PR, Chuartzman S, Vasquez C, Mimbacas A, Nigro V, Pavanello RC, Schuldiner M, Kunkel LM, Zatz M: **A defect in the RNA-processing protein HNRPDL causes limb-girdle muscular dystrophy 1G (LGMD1G)**. *Hum Mol Genet* 2014, **23**:4103-4110.
278. Polat M, Sakinci O, Ersoy B, Sezer RG, Yilmaz H: **Assessment of sleep-related breathing disorders in patients with duchenne muscular dystrophy**. *J Clin Med Res* 2012, **4**:332-337.

279. Della Marca G, Frusciante R, Dittoni S, Vollono C, Buccarella C, Iannaccone E, Rossi M, Scarano E, Pirronti T, Cianfoni A, Mazza S, Tonali PA, Ricci E: **Sleep disordered breathing in facioscapulohumeral muscular dystrophy**. *J Neurol Sci* 2009, **285**:54-58.

280. Shahrizaila N, Kinnear WJ, Wills AJ: **Respiratory involvement in inherited primary muscle conditions**. *J Neurol Neurosurg Psychiatry* 2006, **77**:1108-1115.

281. Robertson PL, Roloff DW: **Chronic respiratory failure in limb-girdle muscular dystrophy: successful long-term therapy with nasal bilevel positive airway pressure**. *Pediatr Neurol* 1994, **10**:328-331.

282. Camara Y, Asin-Cayuela J, Park CB, Metodiev MD, Shi Y, Ruzzenente B, Kukat C, Habermann B, Wibom R, Hultenby K, Franz T, Erdjument-Bromage H, Tempst P, Hallberg BM, Gustafsson CM, Larsson NG: **MTERF4 regulates translation by targeting the methyltransferase NSUN4 to the mammalian mitochondrial ribosome**. *Cell Metab* 2011, **13**:527-539.

283. Burnside RD: **22q11.21 Deletion Syndromes: A Review of Proximal, Central, and Distal Deletions and Their Associated Features**. *Cytogenet Genome Res* 2015, **146**:89-99.

284. Pober BR: **Williams-Beuren syndrome**. *N Engl J Med* 2010, **362**:239-252.

285. Stromme P, Bjornstad PG, Ramstad K: **Prevalence estimation of Williams syndrome**. *J Child Neurol* 2002, **17**:269-271.

286. Singh R, Scheffer IE, Crossland K, Berkovic SF: **Generalized epilepsy with febrile seizures plus: a common childhood-onset genetic epilepsy syndrome**. *Ann Neurol* 1999, **45**:75-81.

287. van Baalen A, Hausler M, Boor R, Rohr A, Sperner J, Kurlemann G, Panzer A, Stephani U, Kluger G: **Febrile infection-related epilepsy syndrome (FIRES): a nonencephalitic encephalopathy in childhood**. *Epilepsia* 2010, **51**:1323-1328.

288. Govani FS, Shovlin CL: **Hereditary haemorrhagic telangiectasia: a clinical and scientific review**. *Eur J Hum Genet* 2009, **17**:860-871.

289. Hastings R, Cobben JM, Gillessen-Kaesbach G, Goodship J, Hove H, Kjaergaard S, Kemp H, Kingston H, Lunt P, Mansour S, McGowan R, Metcalfe K, Murdoch-Davis C, Ray M, Rio M, Smithson S, Tolmie J, Turnpenny P, van Bon B, Wieczorek D, Newbury-Ecob R: **Bohring-Opitz (Oberklaid-Danks) syndrome: clinical study, review of the literature, and discussion of possible pathogenesis**. *Eur J Hum Genet* 2011, **19**:513-519.

290. Dubourg C, Bendavid C, Pasquier L, Henry C, Odent S, David V: **Holoprosencephaly**. *Orphanet J Rare Dis* 2007, **2**:8.

291. van Hengel J, Calore M, Bauce B, Dazzo E, Mazzotti E, De Bortoli M, Lorenzon A, Li Mura IE, Beffagna G, Rigato I, Vleeschouwers M, Tyberghein K, Hulpiau P, van Hamme E, Zaglia T, Corrado D, Basso C, Thiene G, Daliento L, Nava A, van Roy F, Rampazzo A: **Mutations in the area composita protein alphaT-catenin are associated with arrhythmogenic right ventricular cardiomyopathy**. *Eur Heart J* 2013, **34**:201-210.

292. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wagele B, Altmaier E, CardioGram, Deloukas P, Erdmann J, Grundberg E, Hammond CJ, de Angelis MH, Kastenmuller G, Kottgen A, Kronenberg F, Mangino M, Meisinger C, Meitinger T, Mewes HW, Milburn MV, Prehn C, Raffler J, Ried JS, Romisch-Margl W, Samani NJ, Small KS, Wichmann HE, Zhai G, Illig T, et al: **Human metabolic individuality in biomedical and pharmaceutical research**. *Nature* 2011, **477**:54-60.

293. Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmüller G, Kato BS, Mewes H-W, Meitinger T, Hrabé de Angelis M, Kronenberg F, Soranzo N, Wichmann H-E, Spector TD, Adamski J, Suhre K: **A genome-wide perspective of genetic variation in human metabolism**. *Nature Genetics* 2010, **42**:137-141.

294. Kettunen J, Tukiainen T, Sarin A-P, Ortega-Alonso A, Tikkanen E, Lyytikäinen L-P, Kangas AJ, Soininen P, Würtz P, Silander K, Dick DM, Rose RJ, Savolainen MJ, Viikari J, Kähönen M, Lehtimäki T, Pietiläinen KH, Inouye M, McCarthy MI, Jula A, Eriksson J, Raitakari OT, Salomaa V, Kaprio J, Jarvelin M-R, Peltonen L, Perola M, Freimer NB, Ala-Korpela M, Palotie A, et al: **Genome-wide association study identifies multiple loci influencing human serum metabolite levels**. *Nature Genetics* 2012, **44**:269-276.

295. Krumsiek J, Suhre K, Evans AM, Mitchell MW, Mohney RP, Milburn MV, Wagele B, Romisch-Margl W, Illig T, Adamski J, Gieger C, Theis FJ, Kastenmuller G: **Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information**. *PLoS Genet* 2012, **8**:e1003005.

296. Suhre K, Wallaschofski H, Raffler J, Friedrich N, Haring R, Michael K, Wasner C, Krebs A, Kronenberg F, Chang D, Meisinger C, Wichmann HE, Hoffmann W, Völzke H, Völker U, Teumer A, Biffar R, Kocher T, Felix SB, Illig T, Kroemer HK, Gieger C, Romisch-Margl W, Nauck M: **A genome-wide association study of metabolic traits in human urine**. *Nature Genetics* 2011, **43**:565-569.

297. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, Melamed R, Rabadan R, Bernstam EV, Brunak S, Jensen LJ, Nicolae D, Shah NH, Grossman RL, Cox NJ, White KP, Rzhetsky A: **A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk**. *Cell* 2013, **155**:70-80.

298. Mootha VK, Hirschhorn JN: **Inborn variation in metabolism**. *Nat Genet* 2010, **42**:97-98.

299. Barber MJ, Mangravite LM, Hyde CL, Chasman DI, Smith JD, McCarty CA, Li X, Wilke RA, Rieder MJ, Williams PT, Ridker PM, Chatterjee A, Rotter JI, Nickerson DA, Stephens M, Krauss RM: **Genome-wide association of lipid-lowering response to statins in combined study populations**. *PLoS One* 2010, **5**:e9763.

300. Search Collaborative Group, Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, Gut I, Lathrop M, Collins R: **SLCO1B1 variants and statin-induced myopathy--a genomewide study**. *N Engl J Med* 2008, **359**:789-799.

301. Rueedi R, Ledda M, Nicholls AW, Salek RM, Marques-Vidal P, Morya E, Sameshima K, Montoliu I, Da Silva L, Collino S, Martin FP, Rezzi S, Steinbeck C, Waterworth DM, Waeber G, Vollenweider P, Beckmann JS, Le Coutre J, Mooser V, Bergmann S, Genick UK, Kutalik Z: **Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links**. *PLoS Genet* 2014, **10**:e1004132.

302. Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, Bjorndahl TC, Krishnamurthy R, Saleem F, Liu P, Dame ZT, Poelzer J, Huynh J, Yallou FS, Psychogios N, Dong E, Bogumil R, Roehring C, Wishart DS: **The human urine metabolome**. *PLoS One* 2013, **8**:e73076.

303. Petersen AK, Krumsiek J, Wagele B, Theis FJ, Wichmann HE, Gieger C, Suhre K: **On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies**. *BMC Bioinformatics* 2012, **13**:120.

304. Nicholson G, Rantalainen M, Li JV, Maher AD, Malmodin D, Ahmadi KR, Faber JH, Barrett A, Min JL, Rayner NW, Toft H, Krestyaninova M, Viksna J, Neogi SG, Dumas ME, Sarkans U, Mol PC, Donnelly P, Illig T, Adamski J, Suhre K, Allen M, Zondervan KT, Spector TD, Nicholson JK, Lindon JC, Baunsgaard D, Holmes E, McCarthy MI, Holmes CC: **A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection**. *PLoS Genet* 2011, **7**:e1002270.

305. Veiga-da-Cunha M, Hadi F, Balligand T, Stroobant V, Van Schaftingen E: **Molecular identification of hydroxylysine kinase and of ammoniophospholyases acting on 5-phosphohydroxy-L-lysine and phosphoethanolamine**. *J Biol Chem* 2012, **287**:7246-7255.

306. Shao L, Vawter MP: **Shared gene expression alterations in schizophrenia and bipolar disorder**. *Biol Psychiatry* 2008, **64**:89-97.

307. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, et al: **Proteomics. Tissue-based map of the human proteome**. *Science* 2015, **347**:1260419.

308. Roll P, Massacrier A, Pereira S, Robaglia-Schlupp A, Cau P, Szepetowski P: **New human sodium/glucose cotransporter gene (KST1): identification, characterization, and mutation analysis in ICCA (infantile convulsions and choreoathetosis) and BFIC (benign familial infantile convulsions) families**. *Gene* 2002, **285**:141-148.

309. Lahjouji K, Aouameur R, Bissonnette P, Coady MJ, Bichet DG, Lapointe JY: **Expression and functionality of the Na+/myo-inositol cotransporter SMIT2 in rabbit kidney**. *Biochim Biophys Acta* 2007, **1768**:1154-1159.

310. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M: **Annotation of functional variation in personal genomes using RegulomeDB**. *Genome Res* 2012, **22**:1790-1797.

311. Eicher JD, Landowski C, Stackhouse B, Sloan A, Chen W, Jensen N, Lien JP, Leslie R, Johnson AD: **GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes**. *Nucleic Acids Res* 2015, **43**:D799-804.

312. Li MJ, Liu Z, Wang P, Wong MP, Nelson MR, Kocher JP, Yeager M, Sham PC, Chanock SJ, Xia Z, Wang J: **GWASdb v2: an update database for human genetic variants identified by genome-wide association studies**. *Nucleic Acids Res* 2016, **44**:D869-876.
313. Frazer KA, Murray SS, Schork NJ, Topol EJ: **Human genetic variation and its contribution to complex traits**. *Nat Rev Genet* 2009, **10**:241-251.
314. Wagner GP, Zhang J: **The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms**. *Nat Rev Genet* 2011, **12**:204-213.
315. Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ: **Autoimmune disease classification by inverse association with SNP alleles**. *PLoS Genet* 2009, **5**:e1000792.
316. Zhernakova A, van Diemen CC, Wijmenga C: **Detecting shared pathogenesis from the shared genetics of immune-related diseases**. *Nat Rev Genet* 2009, **10**:43-55.
317. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu X-D, Topol EJ, Rosenfeld MG, Frazer KA: **9p21 DNA variants associated with coronary artery disease impair interferon-? signalling response**. *Nature* 2011, **470**:264-268.
318. Meyer KB, Maia A-T, O'Reilly M, Ghoussaini M, Prathalingam R, Porter-Gill P, Ambs S, Prokunina-Olsson L, Carroll J, J BA: **A functional variant at a prostate cancer predisposition locus at 8q24 is associated with PVT1 expression**. *PLoS Genet* 2011, **7**:e1002165.
319. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: **The human disease network**. *Proc Natl Acad Sci U S A* 2007, **104**:8685-8690.
320. Barabási A-L, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease**. *Nat Rev Genet* 2011, **12**:56-68.
321. Klein J, Sato A: **The HLA system. First of two parts**. *N Engl J Med* 2000, **343**:702-709.
322. Gregersen PK: **Gaining insight into PTPN22 and autoimmunity**. *Nat Genet* 2005, **37**:1300-1302.
323. Di Meglio P, Di Cesare A, Laggner U, Chu CC, Napolitano L, Villanova F, Tosi I, Capon F, Trembath RC, Peris K, Nestle FO: **The IL23R R381Q gene variant protects against immune-mediated diseases by impairing IL-23-induced Th17 effector response in humans**. *PLoS One* 2011, **6**:e17160.
324. Chung SA, Taylor KE, Graham RR, Nititham J, Lee AT, Ortmann WA, Jacob CO, Alarcon-Riquelme ME, Tsao BP, Harley JB, Gaffney PM, Moser KL, Petri M, Demirci FY, Kamboh MI, Manzi S, Gregersen PK, Langefeld CD, Behrens TW, Criswell LA: **Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production**. *PLoS Genet* 2011, **7**:e1001323.
325. Radstake TR, Gorlova O, Rueda B, Martin JE, Alizadeh BZ, Palomino-Morales R, Coenen MJ, Vonk MC, Voskuyl AE, Schuerwegh AJ, Broen JC, van Riel PL, van 't Slot R, Italiaander A, Ophoff RA, Riemekasten G, Hunzelmann N, Simeon CP, Ortego-Centeno N, Gonzalez-Gay MA, Gonzalez-Escribano MF, Spanish Scleroderma G, Airo P, van Laar J, Herrick A, Worthington J, Hesselstrand R, Smith V, de Keyser F, Houssiau F, et al: **Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus**. *Nat Genet* 2010, **42**:426-429.
326. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A, Guiducci C, Chen R, Alfredsson L, Amos CI, Ardlie KG, Consortium B, Barton A, Bowes J, Brouwer E, Burtt NP, Catanese JJ, Coblyn J, Coenen MJ, Costenbader KH, Criswell LA, Crusius JB, Cui J, de Bakker PI, De Jager PL, Ding B, et al: **Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci**. *Nat Genet* 2010, **42**:508-514.
327. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, Lee JC, Goyette P, Imielinski M, Latiano A, Lagace C, Scott R, Amininejad L, Bumpstead S, Baidoo L, Baldassano RN, Barclay M, Bayless TM, Brand S, Buning C, Colombel JF, Denson LA, De Vos M, Dubinsky M, Edwards C, Ellinghaus D, Fehrmann RS, Floyd JA, Florin T, Franchimont D, et al: **Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47**. *Nat Genet* 2011, **43**:246-252.
328. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Consortium NIG, Libioulle C, Sandor C, Lathrop M, Belaiche J,

Dewit O, Gut I, et al: **Genome–wide association defines more than 30 distinct susceptibility loci for Crohn's disease**. *Nat Genet* 2008, **40**:955-962.

329. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, et al: **Genome–wide meta–analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci**. *Nat Genet* 2010, **42**:1118-1125.
330. Kugathasan S, Baldassano RN, Bradfield JP, Sleiman PM, Imielinski M, Guthery SL, Cucchiara S, Kim CE, Frackelton EC, Annaiah K, Glessner JT, Santa E, Willson T, Eckert AW, Bonkowski E, Shaner JL, Smith RM, Otieno FG, Peterson N, Abrams DJ, Chiavacci RM, Grundmeier R, Mamula P, Tomer G, Piccoli DA, Monos DS, Annese V, Denson LA, Grant SF, Hakonarson H: **Loci on 20q13 and 21q22 are associated with pediatric–onset inflammatory bowel disease**. *Nat Genet* 2008, **40**:1211-1215.
331. Liu X, Invernizzi P, Lu Y, Kosoy R, Lu Y, Bianchi I, Podda M, Xu C, Xie G, Macciardi F, Selmi C, Lupoli S, Shigeta R, Ransom M, Lleo A, Lee AT, Mason AL, Myers RP, Peltekian KM, Ghent CN, Bernuzzi F, Zuin M, Rosina F, Borghesio E, Floreani A, Lazzari R, Niro G, Andriulli A, Muratori L, Muratori P, et al: **Genome–wide meta–analyses identify three loci associated with primary biliary cirrhosis**. *Nat Genet* 2010, **42**:658-660.
332. Hidalgo CA, Blumm N, Barabasi AL, Christakis NA: **A dynamic network approach for the study of human phenotypes**. *PLoS Comput Biol* 2009, **5**:e1000353.
333. Park J, Lee DS, Christakis NA, Barabasi AL: **The impact of cellular networks on disease comorbidity**. *Mol Syst Biol* 2009, **5**:262.
334. Lee AY, Levine MN: **Venous thromboembolism and cancer: risks and outcomes**. *Circulation* 2003, **107**:I17-21.
335. Khorana AA, Fine RL: **Pancreatic cancer and thromboembolic disease**. *Lancet Oncol* 2004, **5**:655-663.
336. Castelli R, Porro F, Tarsia P: **The heparins and cancer: review of clinical trials and biological properties**. *Vasc Med* 2004, **9**:205-213.
337. de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, Morrison J, Richardson A, Walsh EC, Gao X, Galver L, Hart J, Hafler DA, Pericak-Vance M, Todd JA, Daly MJ, Trowsdale J, Wijmenga C, Vyse TJ, Beck S, Murray SS, Carrington M, Gregory S, Deloukas P, Rioux JD: **A high–resolution HLA and SNP haplotype map for disease association studies in the extended human MHC**. *Nat Genet* 2006, **38**:1166-1172.
338. Hsiung CA, Lan Q, Hong YC, Chen CJ, Hosgood HD, Chang IS, Chatterjee N, Brennan P, Wu C, Zheng W, Chang GC, Wu TC, Park JY, Hsiao CF, Kim YH, Shen HB, Seow A, Yeager M, Tsai YH, Kim YT, Chow WH, Guo HA, Wang WC, Sung SW, Hu ZB, Chen KY, Kim JH, Chen Y, Huang LM, Lee KM, et al: **The 5p15.33 Locus Is Associated with Risk of Lung Adenocarcinoma in Never–Smoking Females in Asia**. *Plos Genetics* 2010, **6**:e1001051.
339. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, Bergen AW, Li Q, Consonni D, Pesatori AC, Wacholder S, Thun M, Diver R, Oken M, Virtamo J, Albanes D, Wang Z, Burdette L, Doheny KF, Pugh EW, Laurie C, Brennan P, Hung R, Gaborieau V, McKay JD, Lathrop M, et al: **A genome–wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma**. *Am J Hum Genet* 2009, **85**:679-691.
340. Miki D, Kubo M, Takahashi A, Yoon KA, Kim J, Lee GK, Zo JI, Lee JS, Hosono N, Morizono T, Tsunoda T, Kamatani N, Chayama K, Takahashi T, Inazawa J, Nakamura Y, Daigo Y: **Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations**. *Nat Genet* 2010, **42**:893-+.
341. Mushiroda T, Wattanapokayakit S, Takahashi A, Nukiwa T, Kudoh S, Ogura T, Taniguchi H, Kubo M, Kamatani N, Nakamura Y, Pirfenidone Clinical Study G: **A genome–wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis**. *J Med Genet* 2008, **45**:654-656.
342. Sanson M, Hosking FJ, Shete S, Zelenika D, Dobbins SE, Ma Y, Enciso-Mora V, Idbaih A, Delattre JY, Hoang-Xuan K, Marie Y, Boisselier B, Carpentier C, Wang XW, Di Stefano AL, Labussiere M, Gousias K, Schramm J, Boland A, Lechner D, Gut I, Armstrong G, Liu Y, Yu R, Lau C, Di Bernardo MC, Robertson LB, Muir K, Hepworth S,

Swerdlow A, et al: **Chromosome 7p11.2 (EGFR) variation influences glioma risk**. *Hum Mol Genet* 2011, **20**:2897-2904.

343. Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, Simon M, Marie Y, Boisselier B, Delattre JY, Hoang-Xuan K, El Hallani S, Idbaih A, Zelenika D, Andersson U, Henriksson R, Bergenheim AT, Feychting M, Lonn S, Ahlbom A, Schramm J, Linnebank M, Hemminki K, Kumar R, Hepworth SJ, Price A, Armstrong G, Liu Y, Gu X, Yu R, et al: **Genome-wide association study identifies five susceptibility loci for glioma**. *Nat Genet* 2009, **41**:899-904.
344. Turnbull C, Rapley EA, Seal S, Pernet D, Renwick A, Hughes D, Ricketts M, Linger R, Nsengimana J, Deloukas P, Huddart RA, Bishop DT, Easton DF, Stratton MR, Rahman N, Collaboration UKTC: **Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer**. *Nat Genet* 2010, **42**:604-607.
345. Xu Y, He K, Goldkorn A: **Telomerase targeted therapy in cancer and cancer stem cells**. *Clin Adv Hematol Oncol* 2011, **9**:442-455.
346. Ramos C, Montano M, Garcia-Alvarez J, Ruiz V, Uhal BD, Selman M, Pardo A: **Fibroblasts from idiopathic pulmonary fibrosis and normal lungs differ in growth rate, apoptosis, and tissue inhibitor of metalloproteinases expression**. *Am J Respir Cell Mol Biol* 2001, **24**:591-598.
347. Schrader M, Burger AM, Muller M, Krause H, Straub B, Smith GL, Newlands ES, Miller K: **Quantification of human telomerase reverse transcriptase mRNA in testicular germ cell tumors by quantitative fluorescence real-time RT-PCR**. *Oncol Rep* 2002, **9**:1097-1105.
348. Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, Spain SL, Broderick P, Domingo E, Farrington S, Prendergast JGD, Pittman AM, Theodoratou E, Smith CG, Olver B, Walther A, Barnetson RA, Churchman M, Jaeger EEM, Penegar S, Barclay E, Martin L, Gorman M, Mager R, Johnstone E, Midgley R, Niittymaki I, Tuupanen S, Colley J, Idziaszczyk S, et al: **Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33**. *Nat Genet* 2010, **42**:973-U989.
349. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adany R, Aromaa A, Bardella MT, van den Berg LH, Bockett NA, de la Concha EG, Dema B, Fehrmann RS, Fernandez-Arquero M, Fiatal S, Grandone E, Green PM, Groen HJ, Gwilliam R, Houwen RH, Hunt SE, Kaukinen K, Kelleher D, Korponay-Szabo I, Kurppa K, MacMathuna P, Maki M, et al: **Multiple common variants for celiac disease influencing immune gene expression**. *Nat Genet* 2010, **42**:295-302.
350. Jones AM, Beggs AD, Carvajal-Carmona L, Farrington S, Tenesa A, Walker M, Howarth K, Ballereau S, Hodgson SV, Zauber A, Bertagnolli M, Midgley R, Campbell H, Kerr D, Dunlop MG, Tomlinson IP: **TERC polymorphisms are associated both with susceptibility to colorectal cancer and with longer telomeres**. *Gut* 2012, **61**:248-254.
351. Cottliar A, Palumbo M, La Motta G, de Barrio S, Crivelli A, Viola M, Gomez JC, Slavutsky I: **Telomere length study in celiac disease**. *Am J Gastroenterol* 2003, **98**:2727-2731.
352. Bishop DT, Demenais F, Iles MM, Harland M, Taylor JC, Corda E, Randerson-Moor J, Aitken JF, Avril MF, Azizi E, Bakker B, Bianchi-Scarra G, Bressac-de Paillerets B, Calista D, Cannon-Albright LA, Chin AWT, Debniak T, Galore-Haskel G, Ghiorzo P, Gut I, Hansson J, Hocevar M, Hoiom V, Hopper JL, Ingvar C, Kanetsky PA, Kefford RF, Landi MT, Lang J, Lubinski J, et al: **Genome-wide association study identifies three loci associated with melanoma risk**. *Nat Genet* 2009, **41**:920-925.
353. Jin Y, Birlea SA, Fain PR, Gowan K, Riccardi SL, Holland PJ, Mailloux CM, Sufit AJ, Hutton SM, Amadi-Myers A, Bennett DC, Wallace MR, McCormack WT, Kemp EH, Gawkrodger DJ, Weetman AP, Picardo M, Leone G, Taieb A, Jouary T, Ezzedine K, van Geel N, Lambert J, Overbeck A, Spritz RA: **Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo**. *N Engl J Med* 2010, **362**:1686-1697.
354. Spritz RA: **The genetics of generalized vitiligo: autoimmune pathways and an inverse relationship with malignant melanoma**. *Genome Med* 2010, **2**:78.
355. Ellinghaus E, Ellinghaus D, Stuart PE, Nair RP, Debrus S, Raelson JV, Belouchi M, Fournier H, Reinhard C, Ding J, Li Y, Tejasvi T, Gudjonsson J, Stoll SW, Voorhees JJ, Lambert S, Weidinger S, Eberlein B, Kunz M, Rahman P, Gladman DD, Gieger C, Wichmann HE, Karlsen TH, Mayr G, Albrecht M, Kabelitz D, Mrowietz U, Abecasis GR, Elder JT, et al: **Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2**. *Nat Genet* 2010, **42**:991-995.

356. Huffmeier U, Uebe S, Ekici AB, Bowes J, Giardina E, Korendowych E, Juneblad K, Apel M, McManus R, Ho P, Bruce IN, Ryan AW, Behrens F, Lascorz J, Bohm B, Traupe H, Lohmann J, Gieger C, Wichmann HE, Herold C, Steffens M, Klareskog L, Wienker TF, Fitzgerald O, Alenius GM, McHugh NJ, Novelli G, Burkhardt H, Barton A, Reis A: **Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis**. *Nat Genet* 2010, **42**:996-999.

357. Sun LD, Xiao FL, Li Y, Zhou WM, Tang HY, Tang XF, Zhang H, Schaarschmidt H, Zuo XB, Foelster-Holst R, He SM, Shi M, Liu Q, Lv YM, Chen XL, Zhu KJ, Guo YF, Hu DY, Li M, Li M, Zhang YH, Zhang X, Tang JP, Guo BR, Wang H, Liu Y, Zou XY, Zhou FS, Liu XY, Chen G, et al: **Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population**. *Nat Genet* 2011, **43**:690-694.

358. Wrensch M, Jenkins RB, Chang JS, Yeh RF, Xiao Y, Decker PA, Ballman KV, Berger M, Buckner JC, Chang S, Giannini C, Halder C, Kollmeyer TM, Kosel ML, LaChance DH, McCoy L, O'Neill BP, Patoka J, Pico AR, Prados M, Quesenberry C, Rice T, Rynearson AL, Smirnov I, Tihan T, Wiemels J, Yang P, Wiencke JK: **Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility**. *Nat Genet* 2009, **41**:905-908.

359. Stranger BE, Stahl EA, Raj T: **Progress and promise of genome-wide association studies for human complex trait genetics**. *Genetics* 2011, **187**:367-383.

360. Caspari E: **A synopsis of contemporary evolutionary thinking**. *Evolution* 1949, **3**:377.

361. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, Liu Y, Liu S, Oliver DE, Rubin DL, Shafa F, Stuart JM, Altman RB: **Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base**. *Pharmacogenomics J* 2001, **1**:167-170.

362. Basen-Engquist K, Chang M: **Obesity and cancer risk: recent review and evidence**. *Curr Oncol Rep* 2011, **13**:71-76.

363. Renstrom F, Payne F, Nordstrom A, Brito EC, Rolandsson O, Hallmans G, Barroso I, Nordstrom P, Franks PW, Consortium G: **Replication and extension of genome-wide association study results for obesity in 4923 adults from northern Sweden**. *Hum Mol Genet* 2009, **18**:1489-1496.

364. Bowes J, Barton A: **The genetics of psoriatic arthritis: lessons from genome-wide association studies**. *Discov Med* 2010, **10**:177-183.

365. Liu Y, Helms C, Liao W, Zaba LC, Duan S, Gardner J, Wise C, Miner A, Malloy MJ, Pullinger CR, Kane JP, Saccone S, Worthington J, Bruce I, Kwok PY, Menter A, Krueger J, Barton A, Saccone NL, Bowcock AM: **A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci**. *PLoS Genet* 2008, **4**:e1000041.

366. Yarur AJ, Czul F, Levy C: **Hepatobiliary manifestations of inflammatory bowel disease**. *Inflamm Bowel Dis* 2014, **20**:1655-1667.

367. Vavricka SR, Schoepfer A, Scharl M, Lakatos PL, Navarini A, Rogler G: **Extraintestinal Manifestations of Inflammatory Bowel Disease**. *Inflamm Bowel Dis* 2015, **21**:1982-1992.

368. Greenstein RJ: **Is Crohn's disease caused by a mycobacterium? Comparisons with leprosy, tuberculosis, and Johne's disease**. *Lancet Infect Dis* 2003, **3**:507-514.

369. Eisenberg I, Eran A, Nishino I, Moggio M, Lamperti C, Amato AA, Lidov HG, Kang PB, North KN, Mitrani-Rosenbaum S, Flanigan KM, Neely LA, Whitney D, Beggs AH, Kohane IS, Kunkel LM: **Distinctive patterns of microRNA expression in primary muscular disorders**. *Proc Natl Acad Sci U S A* 2007, **104**:17016-17021.

370. Gupta SK, Bang C, Thum T: **Circulating MicroRNAs as Biomarkers and Potential Paracrine Mediators of Cardiovascular Disease**. *Circulation-Cardiovascular Genetics* 2010, **3**:484-488.

371. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: **MicroRNA expression profiles classify human cancers**. *Nature* 2005, **435**:834-838.

372. Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, Benjamin H, Shabes N, Tabak S, Levy A, Lebanony D, Goren Y, Silberschein E, Targan N, Ben-Ari A, Gilad S, Sion-Vardy N, Tobar A, Feinmesser M, Kharenko O, Nativ O, Nass D, Perelman M, Yosepovich A, Shalmon B, Polak-Charcon S, Fridman E, Avniel A, Bentwich I, Bentwich Z, et al: **MicroRNAs accurately identify cancer tissue origin**. *Nat Biotechnol* 2008, **26**:462-469.

373. Friedman RC, Farh KKH, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome Res* 2009, **19**:92-105.

374. Hebert SS, De Strooper B: **Alterations of the microRNA network cause neurodegenerative disease**. *Trends Neurosci* 2009, **32**:199-206.

375. Johnston RJ, Jr., Chang S, Etchberger JF, Ortiz CO, Hobert O: **MicroRNAs acting in a double-negative feedback loop to control a neuronal cell fate decision**. *Proc Natl Acad Sci U S A* 2005, **102**:12449-12454.

376. Re A, Cora D, Taverna D, Caselle M: **Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human**. *Mol Biosyst* 2009, **5**:854-867.

377. Wilbert ML, Yeo GW: **Genome-wide approaches in the study of microRNA biology**. *Wiley Interdiscip Rev Syst Biol Med* 2011, **3**:491-512.

378. Chi SW, Zang JB, Mele A, Darnell RB: **Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps**. *Nature* 2009, **460**:479-486.

379. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr., Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T: **Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP**. *Cell* 2010, **141**:129-141.

380. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang XN, Darnell JC, Darnell RB: **HITS-CLIP yields genome-wide insights into brain alternative RNA processing**. *Nature* 2008, **456**:464-U422.

381. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB: **CLIP identifies Nova-regulated RNA networks in the brain**. *Science* 2003, **302**:1212-1215.

382. de la Chapelle A: **Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci**. *Oncogene* 2009, **28**:3345-3348.

383. Meola N, Gennarino VA, Banfi S: **microRNAs and genetic diseases**. *Pathogenetics* 2009, **2**:7.

384. Richardson K, Lai CQ, Parnell LD, Lee YC, Ordovas JM: **A genome-wide survey for SNPs altering microRNA seed sites identifies functional candidates in GWAS**. *BMC Genomics* 2011, **12**:504.

385. Bennett CL, Brunkow ME, Ramsdell F, O'Briant KC, Zhu Q, Fuleihan RL, Shigeoka AO, Ochs HD, Chance PF: **A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA -> AAUGAA) leads to the IPEX syndrome**. *Immunogenetics* 2001, **53**:435-439.

386. Walters RW, Bradrick SS, Gromeier M: **Poly(A)-binding protein modulates mRNA susceptibility to cap-dependent miRNA-mediated repression**. *RNA* 2010, **16**:239-250.

387. Li X, Quon G, Lipshitz HD, Morris Q: **Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure**. *RNA* 2010, **16**:1096-1107.

388. Waldispuhl J, Clote P: **Computing the partition function and sampling for saturated secondary structures of RNA, with respect to the Turner energy model**. *J Comput Biol* 2007, **14**:190-215.

389. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E: **The role of site accessibility in microRNA target recognition**. *Nat Genet* 2007, **39**:1278-1284.

390. Yang JO, Kim WY, Bhak J: **ssSNPTarget: genome-wide splice-site Single Nucleotide Polymorphism database**. *Hum Mutat* 2009, **30**:E1010-1020.

391. Betel D, Wilson M, Gabow A, Marks DS, Sander C: **The microRNA.org resource: targets and expression**. *Nucleic Acids Res* 2008, **36**:D149-153.

392. Flowers E, Froelicher ES, Aouizerat BE: **MicroRNA regulation of lipid metabolism**. *Metabolism* 2013, **62**:12-20.

393. Kraja AT, Vaidya D, Pankow JS, Goodarzi MO, Assimes TL, Kullo IJ, Sovio U, Mathias RA, Sun YV, Franceschini N, Absher D, Li G, Zhang Q, Feitosa MF, Glazer NL, Haritunians T, Hartikainen AL, Knowles JW, North KE, Iribarren C, Kral B, Yanek L, O'Reilly PF, McCarthy MI, Jaquish C, Couper DJ, Chakravarti A, Psaty BM, Becker LC, Province MA, et al: **A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium**. *Diabetes* 2011, **60**:1329-1339.

394. Kooner JS, Chambers JC, Aguilar-Salinas CA, Hinds DA, Hyde CL, Warnes GR, Gomez Perez FJ, Frazer KA, Elliott P, Scott J, Milos PM, Cox DR, Thompson JF: **Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides**. *Nat Genet* 2008, **40**:149-151.

395. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, Wahlstrand B, Hedner T, Corella D, Tai ES, Ordovas JM, Berglund G, Vartiainen E, Jousilahti P, Hedblad B, Taskinen MR, Newton-Cheh C, Salomaa V, Peltonen L, Groop L, Altshuler DM, Orho-Melander M: **Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans**. *Nat Genet* 2008, **40**:189-197.

396. Chasman DI, Pare G, Mora S, Hopewell JC, Peloso G, Clarke R, Cupples LA, Hamsten A, Kathiresan S, Malarstig A, Ordovas JM, Ripatti S, Parker AN, Miletich JP, Ridker PM: **Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis**. *PLoS Genet* 2009, **5**:e1000730.

397. Coram MA, Duan Q, Hoffmann TJ, Thornton T, Knowles JW, Johnson NA, Ochs-Balcom HM, Donlon TA, Martin LW, Eaton CB, Robinson JG, Risch NJ, Zhu X, Kooperberg C, Li Y, Reiner AP, Tang H: **Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations**. *Am J Hum Genet* 2013, **92**:904-916.

398. Richardson K, Nettleton JA, Rotllan N, Tanaka T, Smith CE, Lai CQ, Parnell LD, Lee YC, Lahti J, Lemaitre RN, Manichaikul A, Keller M, Mikkila V, Ngwa J, van Rooij FJ, Ballentyne CM, Borecki IB, Cupples LA, Garcia M, Hofman A, Ferrucci L, Mozaffarian D, Perala MM, Raitakari O, Tracy RP, Arnett DK, Bandinelli S, Boerwinkle E, Eriksson JG, Franco OH, et al: **Gain-of-function lipoprotein lipase variant rs13702 modulates lipid traits through disruption of a microRNA-410 seed site**. *Am J Hum Genet* 2013, **92**:5-14.

399. Siengdee P, Trakooljul N, Murani E, Schwerin M, Wimmers K, Ponsuksili S: **MicroRNAs Regulate Cellular ATP Levels by Targeting Mitochondrial Energy Metabolism Genes during C2C12 Myoblast Differentiation**. *PLoS One* 2015, **10**:e0127850.

400. Kaur K, Vig S, Srivastava R, Mishra A, Singh VP, Srivastava AK, Datta M: **Elevated Hepatic miR-22-3p Expression Impairs Gluconeogenesis by Silencing the Wnt-Responsive Transcription Factor Tcf7**. *Diabetes* 2015, **64**:3659-3669.

401. el Azzouzi H, Leptidis S, Dirkx E, Hoeks J, van Bree B, Brand K, McClellan EA, Poels E, Sluimer JC, van den Hoogenhof MM, Armand AS, Yin X, Langley S, Bourajjaj M, Olieslagers S, Krishnan J, Vooijs M, Kurihara H, Stubbs A, Pinto YM, Krek W, Mayr M, da Costa Martins PA, Schrauwen P, De Windt LJ: **The hypoxia-inducible microRNA cluster miR-199a approximately 214 targets myocardial PPARdelta and impairs mitochondrial fatty acid oxidation**. *Cell Metab* 2013, **18**:341-354.

402. Li K, Zhang J, Yu J, Liu B, Guo Y, Deng J, Chen S, Wang C, Guo F: **MicroRNA-214 suppresses gluconeogenesis by targeting activating transcriptional factor 4**. *J Biol Chem* 2015, **290**:8185-8195.

403. Soh J, Iqbal J, Queiroz J, Fernandez-Hernando C, Hussain MM: **MicroRNA-30c reduces hyperlipidemia and atherosclerosis in mice by decreasing lipid synthesis and lipoprotein secretion**. *Nat Med* 2013, **19**:892-900.

404. Feng B, Chakrabarti S: **miR-320 Regulates Glucose-Induced Gene Expression in Diabetes**. *ISRN Endocrinol* 2012, **2012**:549875.

405. Gao P, Tchernyshyov I, Chang TC, Lee YS, Kita K, Ochi T, Zeller KI, De Marzo AM, Van Eyk JE, Mendell JT, Dang CV: **c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism**. *Nature* 2009, **458**:762-765.

406. Yin H, Hu M, Zhang R, Shen Z, Flatow L, You M: **MicroRNA-217 promotes ethanol-induced fat accumulation in hepatocytes by down-regulating SIRT1**. *J Biol Chem* 2012, **287**:9817-9826.

407. Rottiers V, Naar AM: **MicroRNAs in metabolism and metabolic disorders**. *Nat Rev Mol Cell Biol* 2012, **13**:239-250.

408. Romao JM, Jin W, Dodson MV, Hausman GJ, Moore SS, Guan LL: **MicroRNA regulation in mammalian adipogenesis**. *Exp Biol Med (Maywood)* 2011, **236**:997-1004.

409. Yamanouchi T, Akanuma Y: **Serum 1,5-anhydroglucitol (1,5 AG): new clinical marker for glycemic control**. *Diabetes Res Clin Pract* 1994, **24 Suppl**:S261-268.

410. Pal A, Farmer AJ, Dudley C, Selwood MP, Barrow BA, Klyne R, Grew JP, McCarthy MI, Gloyn AL, Owen KR: **Evaluation of serum 1,5 anhydroglucitol levels as a clinical test to differentiate subtypes of diabetes**. *Diabetes Care* 2010, **33**:252-257.

411. Skupien J, Gorczynska-Kosiorz S, Klupa T, Wanic K, Button EA, Sieradzki J, Malecki MT: **Clinical application of 1,5-anhydroglucitol measurements in patients with hepatocyte nuclear factor-1alpha maturity-onset diabetes of the young**. *Diabetes Care* 2008, **31**:1496-1501.

412. Krakowiak PA, Wassif CA, Kratz L, Cozma D, Kovarova M, Harris G, Grinberg A, Yang Y, Hunter AG, Tsokos M, Kelley RI, Porter FD: **Lathosterolosis: an inborn error of human and murine cholesterol synthesis due to lathosterol 5-desaturase deficiency**. *Hum Mol Genet* 2003, **12**:1631-1641.

413. Weingartner O, Weingartner N, Scheller B, Lutjohann D, Graber S, Schafers HJ, Bohm M, Laufs U: **Alterations in cholesterol homeostasis are associated with coronary heart disease in patients with aortic stenosis**. *Coron Artery Dis* 2009, **20**:376-382.

414. Weingartner O, Pinsdorf T, Rogacev KS, Blomer L, Grenner Y, Graber S, Ulrich C, Girndt M, Bohm M, Fliser D, Laufs U, Lutjohann D, Heine GH: **The relationships of markers of cholesterol homeostasis with carotid intima-media thickness**. *PLoS One* 2010, **5**:e13467.

415. Vickers KC, Rye KA, Tabet F: **MicroRNAs in the onset and development of cardiovascular disease**. *Clin Sci (Lond)* 2014, **126**:183-194.

416. Zhang J, Zhang W, Zou D, Chen G, Wan T, Zhang M, Cao X: **Cloning and functional characterization of ACAD-9, a novel member of human acyl-CoA dehydrogenase family**. *Biochem Biophys Res Commun* 2002, **297**:1033-1042.

417. Schiff M, Haberberger B, Xia C, Mohsen AW, Goetzman ES, Wang Y, Uppala R, Zhang Y, Karunanidhi A, Prabhu D, Alharbi H, Prochownik EV, Haack T, Haberle J, Munnich A, Rotig A, Taylor RW, Nicholls RD, Kim JJ, Prokisch H, Vockley J: **Complex I assembly function and fatty acid oxidation enzyme activity of ACAD9 both contribute to disease severity in ACAD9 deficiency**. *Hum Mol Genet* 2015, **24**:3238-3247.

418. Menghini R, Casagrande V, Marino A, Marchetti V, Cardellini M, Stoehr R, Rizza S, Martelli E, Greco S, Mauriello A, Ippoliti A, Martelli F, Lauro R, Federici M: **MiR-216a: a link between endothelial dysfunction and autophagy**. *Cell Death Dis* 2014, **5**:e1029.

419. Greco S, Fasanaro P, Castelvecchio S, D'Alessandra Y, Arcelli D, Di Donato M, Malavazos A, Capogrossi MC, Menicanti L, Martelli F: **MicroRNA dysregulation in diabetic ischemic heart failure patients**. *Diabetes* 2012, **61**:1633-1641.

420. Spain SL, Barrett JC: **Strategies for fine-mapping complex traits**. *Hum Mol Genet* 2015, **24**:R111-119.

421. Lehner NJ: **Evaluation of the applicability of current promoter and enhancer annotations for the prediction of regulatory genetic variant effects using intrinsic and extrinsic classification testing**. Bachelor thesis; Munich; 2015.

422. D.M.W. P: **Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation**. *Journal of Machine Learning Technologies* 2011, **2**:37.

423. Wanders RJ, Duran M, Loupatty FJ: **Enzymology of the branched-chain amino acid oxidation disorders: the valine pathway**. *J Inherit Metab Dis* 2012, **35**:5-12.

424. Liebich HM, Forst C: **Hydroxycarboxylic and oxocarboxylic acids in urine: products from branched-chain amino acid degradation and from ketogenesis**. *J Chromatogr* 1984, **309**:225-242.

425. Marzi C, Albrecht E, Hysi PG, Lagou V, Waldenberger M, Tonjes A, Prokopenko I, Heim K, Blackburn H, Ried JS, Kleber ME, Mangino M, Thorand B, Peters A, Hammond CJ, Grallert H, Boehm BO, Kovacs P, Geistlinger L, Prokisch H, Winkelmann BR, Spector TD, Wichmann HE, Stumvoll M, Soranzo N, Marz W, Koenig W, Illig T, Gieger C: **Genome-wide association study identifies two novel regions at 11p15.5-p13 and 1p31 with major impact on acute-phase serum amyloid A**. *PLoS Genet* 2010, **6**:e1001213.

426. Drapkin R, Reinberg D: **The multifunctional TFIIH complex and transcriptional control**. *Trends Biochem Sci* 1994, **19**:504-508.

427. Hoogstraten D, Nigg AL, Heath H, Mullenders LH, van Driel R, Hoeijmakers JH, Vermeulen W, Houtsmuller AB: **Rapid switching of TFIIH between RNA polymerase I and II transcription and DNA repair in vivo**. *Mol Cell* 2002, **10**:1163-1174.

428. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database**. *Nucleic Acids Res* 2009, **37**:D674-679.

429. Heemskerk MM, van Harmelen VJ, van Dijk KW, van Klinken JB: **Reanalysis of mGWAS results and in vitro validation show that lactate dehydrogenase interacts with branched-chain amino acid metabolism**. *Eur J Hum Genet* 2016, **24**:142-145.

430. Krug S, Kastenmuller G, Stuckler F, Rist MJ, Skurk T, Sailer M, Raffler J, Romisch-Margl W, Adamski J, Prehn C, Frank T, Engel KH, Hofmann T, Luy B, Zimmermann R, Moritz F, Schmitt-Kopplin P, Krumsiek J, Kremer W, Huber F, Oeh U, Theis FJ, Szymczak W, Hauner H, Suhre K, Daniel H: **The dynamic range of the human metabolome revealed by challenges**. *FASEB J* 2012, **26**:2607-2619.

431. Chuang DT, Shih VE: **Maple syrup urine disease (branched-chain ketoaciduria)**. In *The Metabolic and Molecular Bases of Inherited Disease. Volume* II. 8th edition. Edited by Scriver CR, Beaudet AL, Sly WS, Valle D. New York: McGraw-Hill; 2001: Pp. 1971-2005

432. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, ReproGen C, Psychiatric Genomics C, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control C, Duncan L, Perry JR, Patterson N, Robinson EB, Daly MJ, Price AL, Neale BM: **An atlas of genetic correlations across human diseases and traits**. *Nat Genet* 2015, **47**:1236-1241.

433. Ellinor PT, Lunetta KL, Albert CM, Glazer NL, Ritchie MD, Smith AV, Arking DE, Muller-Nurasyid M, Krijthe BP, Lubitz SA, Bis JC, Chung MK, Dorr M, Ozaki K, Roberts JD, Smith JG, Pfeufer A, Sinner MF, Lohman K, Ding J, Smith NL, Smith JD, Rienstra M, Rice KM, Van Wagoner DR, Magnani JW, Wakili R, Clauss S, Rotter JI, Steinbeck G, et al: **Meta-analysis identifies six new susceptibility loci for atrial fibrillation**. *Nat Genet* 2012, **44**:670-675.

434. Gudbjartsson DF, Holm H, Gretarsdottir S, Thorleifsson G, Walters GB, Thorgeirsson G, Gulcher J, Mathiesen EB, Njolstad I, Nyrnes A, Wilsgaard T, Hald EM, Hveem K, Stoltenberg C, Kucera G, Stubblefield T, Carter S, Roden D, Ng MC, Baum L, So WY, Wong KS, Chan JC, Gieger C, Wichmann HE, Gschwendtner A, Dichgans M, Kuhlenbaumer G, Berger K, Ringelstein EB, et al: **A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke**. *Nat Genet* 2009, **41**:876-878.

435. Gretarsdottir S, Thorleifsson G, Manolescu A, Styrkarsdottir U, Helgadottir A, Gschwendtner A, Kostulas K, Kuhlenbaumer G, Bevan S, Jonsdottir T, Bjarnason H, Saemundsdottir J, Palsson S, Arnar DO, Holm H, Thorgeirsson G, Valdimarsson EM, Sveinbjornsdottir S, Gieger C, Berger K, Wichmann HE, Hillert J, Markus H, Gulcher JR, Ringelstein EB, Kong A, Dichgans M, Gudbjartsson DF, Thorsteinsdottir U, Stefansson K: **Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke**. *Ann Neurol* 2008, **64**:402-409.

436. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, Martins RA, Kennedy BA, Hassell RG, Visser ME, Schwartz SM, Voight BF, Elosua R, Salomaa V, O'Donnell CJ, Dallinga-Thie GM, Anand SS, Yusuf S, Huff MW, Kathiresan S, Hegele RA: **Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia**. *Nat Genet* 2010, **42**:684-687.

437. Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, Preuss M, Stewart AF, Barbalic M, Gieger C, Absher D, Aherrahrou Z, Allayee H, Altshuler D, Anand SS, Andersen K, Anderson JL, Ardissino D, Ball SG, Balmforth AJ, Barnes TA, Becker DM, Becker LC, Berger K, Bis JC, Boekholdt SM, Boerwinkle E, Braund PS, Brown MJ, Burnett MS, et al: **Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease**. *Nat Genet* 2011, **43**:333-338.

438. Esteller M: **Non-coding RNAs in human disease**. *Nat Rev Genet* 2011, **12**:861-874.

439. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, Hendricks AE, Danecek P, Li R, Floyd J, Wain LV, Barroso I, Humphries SE, Hurles ME, Zeggini E, Barrett JC, Plagnol V, Richards JB, Greenwood CM, Timpson NJ, Durbin R, Soranzo N: **The UK10K project identifies rare variants in health and disease**. *Nature* 2015, **526**:82-90.

440. Topol EJ, Steinhubl SR, Torkamani A: **Digital medical tools and sensors**. *JAMA* 2015, **313**:353-354.

441. Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van Diemen FR, Olk N, Stukalov A, Marceau C, Janssen H, Carette JE, Bennett KL, Colinge J, Superti-Furga G, Brummelkamp TR: **Gene essentiality and synthetic lethality in haploid human cells**. *Science* 2015, **350**:1092-1096.

442. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM: **Identification and characterization of essential genes in the human genome**. *Science* 2015, **350**:1096-1101.

443. Mullard A: **New drugs cost US$2.6 billion to develop**. *Nat Rev Drug Discov* 2014, **13**:877-877.

444. DiMasi JA, Hansen RW, Grabowski HG: **The price of innovation: new estimates of drug development costs**. *J Health Econ* 2003, **22**:151-185.

445. Weiss KM: **Is there a paradigm shift in genetics? Lessons from the study of human diseases**. *Mol Phylogenet Evol* 1996, **5**:259-265.

446. Topol EJ: **The big medical data miss: challenges in establishing an open medical resource**. *Nat Rev Genet* 2015, **16**:253-254.

447. Geihs M, Yan Y, Walter K, Huang J, Memari Y, Min JL, Mead D, Consortium UK, Hubbard TJ, Timpson NJ, Down TA, Soranzo N: **An interactive genome browser of association results from the UK10K cohorts project**. *Bioinformatics* 2015, **31**:4029-4031.

448. de Andrade Krätzig NT: **Mapping of established biomarkers in routine clinical use to complex trait associated genetic loci**. Bachelor thesis; Munich; 2015.

449. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P: **The support of human genetic evidence for approved drug indications**. *Nat Genet* 2015, **47**:856-860.

450. Jonas S, Izaurralde E: **Towards a molecular understanding of microRNA-mediated gene silencing**. *Nat Rev Genet* 2015, **16**:421-433.