Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Medicine 2055

Genomic Analysis of Adverse Drug Reactions

JOEL ÅS

ACTA UNIVERSITATIS
UPSALIENSIS
2024

ISSN 1651-6206
ISBN 978-91-513-2139-4
urn:nbn:se:uu:diva-527102

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Rosénsalen, Akademiska sjukhuset, ing. 95/96, Uppsala, Thursday, 13 June 2024 at 13:00 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Professor Volker Lauschke (Department of Physiology and Pharmacology, Karolinska Institutet, and Deputy Head of the Margarete Fischer-Bosch Institute of Clinical Pharmacology in Stuttgart, Germany).

Abstract
Ås, J. 2024. Genomic Analysis of Adverse Drug Reactions. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 2055. 67 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2139-4.

Adverse drug reactions (ADRs) pose a significant global challenge, leading to substantial costs, suffering, and even loss of life. Genetic factors can play a role in determining a patient's response to the drug treatments and predicting ADRs. While many genetic associations with ADRs have been identified, there are still numerous ADRs suspected to have genetic components.

In Paper I, the collection and curation strategies for ADR cases in the Swedegene biobank are established, presenting a cohort of 2,550 ADR-cases. Paper II presents the association between genetic variations in human leukocyte antigen (HLA) genes and the development of pancreatitis as a response to azathioprine treatment in patients with Crohn's disease. Paper III reports on an international collaboration to investigate the genetic aetiology of atypical femur fractures (AFF) during bisphosphonate treatment. The study found that previously identified genetic variants did not replicate, and --- as the cohort is the largest of its kind --- provides valuable insights into common genetic factors of AFF. Paper IV examines the genetic associations with central nervous system (CNS) toxicity as an ADR to antimicrobial drugs, identifying correlations with three genes linked to suicide and schizophrenia, although the biological connection remains unclear. Finally, Paper V presents a methodology for the experimental design of ADR studies by analysing the known protein interactions of drugs and proteins associated with ADRs. This approach aims to mitigate the impact of competing genetic correlations by identifying common protein interactions to validate the inclusion of drugs and ADRs in the study. These interactions are then ranked based on importance to the selected drugs and ADRs and used to propose genetic targets of interest.

Overall, the findings of these studies contribute to the understanding of genetic predispositions to ADRs and provide a novel approach for data-driven experimental design for phenotype and genetic target selection.

Keywords: Adverse drug reactions, Genetic association, Network biology

Joel Ås, Department of Medical Sciences, Clinical pharmacogenomics and osteoporosis, Akademiska sjukhuset, Uppsala University, SE-75185 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-513-2139-4
URN urn:nbn:se:uu:diva-527102 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-527102)

Dedicated to all you keeping me
curious and creative

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Hallberg, P., Yue, Q., Eliasson, E., Melhus, H., **Ås J**., Wadelius, M., SWEDEGENE: a Swedish nation-wide DNA sample collection for pharmacogenomic studies of serious adverse drug reactions. The Pharmacogenomics Journal, ISSN 1470-269X, E-ISSN 1473-1150, Vol. 20, no 4, p. 579-585 (2020)

II **Ås, J**., Bertulyte, I., Eriksson, N., Wadelius, M., Hallberg P. HLA variants associated with azathioprine-induced pancreatitis in patients with Crohn's disease. Clinical and Translational Science, ISSN 1752-8054, E-ISSN 1752-8062, Vol. 15, no 5, p. 1249-1256 (2022)

III Zhou, W., **Ås, J**., Shore-Lorenti, C., Nguyen H., van de Laarschot D., Sztal-Mazer, S., Grill, V., Girgis, C., Stricker, B., van der Eerden, B., Thakker, R., Appelman-Dijkstra, N., Wadelius, M., Clifton-Bligh, R., Hallberg, P., Verkerk, A., van Rooij, J., Ebeling, P., Zillikens, C. Gene-based association analysis of a large patient cohort identifies potential gene candidates for atypical femur fractures. Journal of Bone and Mineral Research (manuscript 2023)

IV **Ås, J**., Bertulyte, I., Norgren, N., Johansson, A., Eriksson, N., Green, H., Wadelius, M., Hallberg, P. Whole genome case-control study of central nervous system (CNS) toxicity due to antimicrobial drugs. PLOS ONE, https://doi.org/10.1371/journal.pone.0299075 (2024)

V **Ås, J**., Eriksson, N., Hallberg, P. Wadelius, M. Network-Based Analysis of Protein Interactions among Drugs and Adverse Reactions: Identifying Phenotype-Groupings and Key Genes. BMC Methods (manuscript 2024)

Contents

Abbreviations

ADME	Absorption, Distribution, Metabolism, and Excretion
ADR	Adverse Drug Reaction
AFF	Atypical Femur Fracture
BAM	Binary Sequence Alignment Map
BLT	Bead-Linked Transposomes
BMI	Body Mass Index
BQSR	Base Quality Score Recalibration
CADD	Combined Annotation Dependent Depletion
CNS	Central Nervous System
dNTP	Deoxyribose Nucleoside Triphosphate
ddNTP	Dideoxynucleotides Triphosphate
GATK	Genome Analysis Tool Kit
GRCh37	Genome Reference Consortium Human Build 37
GVCF	Genomic Variant Calling File
HLA	Human leukocyte Antigen
HPO	Human Phenotype Ontology
HWE	Hardy-Weinberg Equilibrium
IBD	Inflammatory Bowel Disease
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCC	Large Connected Component
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
NSAID	Nonsteroidal Anti-Inflammatory Drug
PC	Principal Component
PCA	Principal Component Analysis
PPI	Protein-Protein Interaction
QC	Quality Control
SJS/TEN	Stevens-Johnson Syndrome / Toxic Epidermal Necrolysis
SKAT	Sequence Kernel Association Test
SKAT-O	Sequence Kernel Association Test with Optimal weight
SNP	Single Nucleotide Polymorphism
SV	Structural Variation
VCF	Variant Calling File
VF	Variant Frequency
VQSR	Variant Quality Score Recalibration
WGS	Whole Genome sequencing

1. Introduction

This thesis will discuss genetic research aiming to uncover predisposing genetic variation for adverse drug reactions (ADRs) and the approaches to do so. ADRs pose a significant problem in healthcare and society and research aims to understand and develop the ability to predict ADRs.

There are known genetic predispositions to adverse reactions to certain drugs. The predisposing genetic variation is tested in healthcare proactively to avoid the ADR but those cases where the genetic predisposition is known are few. Therefore, there is potential to alleviate cost, suffering, and death by uncovering genetic predispositions of ADRs.

Severe ADRs are rare as drugs with such common side effects would typically not be approved by regulatory authorities. The consequence of this is that recorded cases and samples from patients are also rare. To tackle the scarcity of samples, the Swedegene project has collected cases over almost two decades while periodically conducting genomic studies on samples in their biobank. In **Paper I**, the recruitment strategies and data curation employed by Swedegene are presented as well as ADR diagnosis of the first 2550 cases.

Paper II is one such study where the samples' genomic variants were genotyped using single nucleotide polymorphism (SNP) microarrays, showing a correlation between drug-induced pancreatitis and genetic variation. However, as SNP microarrays only genotype predetermined genetic loci, a large cohort of whole genomes from Swedegene samples was sequenced, capturing the majority of the samples' genetic variation. In **Paper III** and **Paper IV** genetic variation of samples from this cohort is investigated for genetic predisposition to the ADRs atypical femur fracture and central nervous system (CNS) toxicity.

When investigating genetic predisposition statistical tests are conducted on genetic variation to find differences between cases and controls. The results of such tests have an inherent risk of falsely identifying significant differences.

The convention is that this risk must be below 5% for a single test. If multiple tests are conducted, the per-test acceptable risk decreases to ensure that the overall risk stays at acceptable levels. In genome-wide association studies, hundreds of thousands of genetic variants are tested. Consequently, the adjusted per-sample acceptable risk is very low, resulting in only the strongest correlations being detected.

Often, a collection of genetic loci is chosen in beforehand to test to detect weaker correlations. These loci are usually chosen based on the working hypothesis of the study, but such information is not always available. **Paper V**

presents a methodology to suggest what drugs and ADRs could be included in a study while suggesting genetic loci to test based on public gene-drug and gene-ADR association data.

In conclusion, this thesis will present sample collection strategies, genomic association studies on genetic predisposition of ADRs, and a novel method for phenotype and genetic target selection.

2. Background

2.1 Drug reactions in general

The absolute majority of drugs have ADRs. The definition of these reactions is broad, being defined as *any reaction except the intended therapeutic effect*. By this definition, both a lack of therapeutic effect and a too strong therapeutic effect is considered an ADR. Consequently, many ADRs are dependent on the dosage. ADRs are classified into two categories based on this property; intrinsic (type A) or idiosyncratic (type B)[1]. Intrinsic ADRs are dependent on dosage, such as the dose-dependent toxicity-reactions to the cytotoxic agent 5-fluorouracil[2]. Idiosyncratic ADRs do not depend on dosage and are harder to predict. The development of narcolepsy among Swedish youth during the vaccination campaign against the swine flu in 2009-2010[3] is an example of an idiosyncratic reaction.

2.2 Adverse drug reactions — the problem in healthcare

Before any drug is approved for routine use it must undergo rigorous clinical trials, where any observed serious ADR can prevent approval by regulatory authorities. Consequently, common ADRs are in general mild or moderate. There are exceptions to this where the risk of ADRs is out-weighed by the potential benefit of treatment, such as drugs for the treatment of cancer.

When a drug is undergoing a clinical trial, the frequency and severity of ADRs are studied among the participants. Since clinical trials study a finite number of participants, rare ADRs will seldom be observed. Therefore, rare ADRs have a higher probability of being severe once they manifest in routine use, as the observation of such an ADR could have hindered the drug from being approved. Additionally, the ancestry of participants in clinical trials could influence the outcome as some ADRs to the same drug are more common in some ethnicities than others[4]. Due to this, continued monitoring of ADRs is essential. Some international and national efforts dedicated to this task are the WHO Programme for International Drug Monitoring and the Pharmacovigilance Unit at the Swedish Medical Products Agency.

Despite this, ADRs still cause considerable morbidity[5] in Swedish healthcare where 12 % of Hospital admissions[6] have been shown to be attributed to ADRs during the 2000s and 9.5 % of all costs could be attributed to ADRs in 2008[7]. ADRs are also estimated to represent 3 % of fatalities in Sweden[8] and were estimated to be the fourth leading cause of death in the USA between 1966 and 1998[9].

2.3 Causes and Prevention

The majority of ADRs can be attributed to interactions of three factors: pharmacological, environmental, and biological.

- Pharmacological: Many drugs may interact with other drugs resulting in increased or decreased drug effects. Pharmacokinetic drug interactions occur when one drug affects the metabolism of another drug. An example of a pharmacodynamic drug interaction is an enhanced risk of opioid-induced decrease in the respiratory drive when combining morphine and benzodiazepines[10].
- Environmental: Similar to the potential interactions between drugs there are small molecules in our environment that can cause comparable effects. Additionally, exposure to infectious agents can change the drug response by modifying our microbiome[11] and immune system[12].
- Biological: Decreased kidney or liver function can lead to accumulation of the drug and predispose the patient to an ADR. The drug response is also dictated by the genetic predisposition where a specific genetic variation influences the metabolism of the drug or prompts adverse immunological responses.

In some cases, an ADR can be prevented if the causal factors are known. There are knowledge-based systems for pharmacological interactions such as the SFINX drug-drug interaction database[13] supporting clinical decisions. Environmental factors can be both simple and complex, as drugs may have known interactions with small molecules present in herbal medications (such as St John's wort[14]) or foodstuffs such as grapefruit[15]. However, the clinician cannot take the patient's complete infection history into account or have control over all small molecules that the patient is exposed to. As for biological interactions, the state of the patient is evaluated by a physician, and the known potential dangers of a drug are weighed against the potential benefit. Genetically mediated ADRs can be prevented by proactively genotyping known causal genetic variations and adjusting treatment to suit the genetic makeup. The patient's genotype will remain the same during the patient's lifetime and future treatments can be tailored to it. For instance, specific genetic variants have been found to decrease or inactivate the activity of the enzyme dihydropyrimidine dehydrogenase (DPYD) which places carriers of these variants at risk of toxicity during treatment with 5-fluorouracil[2]. The decreased enzyme activity results in a slower breakdown of 5-fluorouracil and prolongs drug exposure, resulting in toxicity. The HLA-type HLA-DQB1*06:02 was correlated with an increased risk of narcolepsy among youths following the swine flu vaccination[3] in Sweden 2009-2010. Had this knowledge been available prior to vaccination, some cases of narcolepsy could have been averted. However, while nearly 30% of the Swedish population possesses this specific HLA-type, only a small fraction of them developed narcolepsy. Still,

knowledge of this risk could have given Swedish youth carrying the HLA-type an informed choice.

2.4 Scarcity of data when studying rare ADRs

There is thus a need to study genetic associations between drugs and their ADRs. However, numerous obstacles must be overcome when designing such studies. Even though, collectively, serious ADRs are not a rare occurrence, the individual cases of a specific reaction to a specific drug are rare. As previously mentioned, a drug commonly associated with severe ADRs would never be approved by a regulatory agency. An additional crux when researching ADRs in healthcare is that the causality of a reaction can be hard to attribute to the drug. Complicating factors such as co-morbidity and multiple simultaneous medical treatments confound the correlation between a drug and a reaction. Restarting a discontinued treatment due to an ADR to see if the reaction reoccurs, known as rechallenge, can establish causality but it requires that alternative treatments are insufficient and that the ADR is not too severe. Of course, this also assumes that the ADR is not chronic, such as narcolepsy mentioned earlier. Still, this is only the medical information about the patient, while environmental factors such as food habits, lifestyle habits, and possible self-treatments are still possible confounding factors. After checking all of these, and still finding the reaction to be a possible ADR, the presence of a genetic predisposition to that ADRs is still not guaranteed.

3. Aims

As evident from Section 2.2, genetic factors of ADRs should be studied as it has potential to alleviate suffering and death. Therefore the overarching aim of this thesis has been to explore genetic predispositions of ADRs. However, there are specific question relating to this aim.

These Questions are:

- Does ADR data be collected and curated the way the Swedegene Project does further the understanding of genetic factors underpinning ADRs?
- Can previous findings of azathioprine-induced pancreatitis be replicated?
- Can novel genetic associations be found for ADRs and categories of ADRs using whole genome sequencing?
- Can shared protein-protein interaction be used to formulate genetic hypotheses of genetic predispositions for drugs relating to ADRs?

4. Material and Methods

This chapter presents the different computational, bioinformatical, and statistical tools used during the project as well as the samples used for association studies.

4.1 Swedegene WGS Cohort

The Swedegene WGS (whole genome sequencing) cohort consists of whole genomes from 978 individuals who had ADRs. Among these cases there are 16 categories of ADRs to be investigated for genetic predisposition. Each ADR category has a set of suspected drug categories, listed in table 4.1. Note that the sum of patients from all categories exceeds the number of samples in the cohort, since some patients are included in multiple categories. Originally, the patients in the cohort were to be compared to individuals from the SweGen project[16] exposed to the same drug types. However, the medical phenotype data from these individuals was not obtained and, unfortunately, a batch effect between the Swedegene and SweGen samples gave rise to spurious genetic correlations. Due to this, the other patients in the cohort were used as population controls going forward.

Table 4.1. *ADR type drug-categories included in each grouping in the Swedegene cohort*

ADR-type	Drug-type	Number of cases
Hypersensitivity inclusive	Antibiotics, Anti-epileptics Immunosuppressants NSAIDs	372
Hypersensitivity narrow	Antibiotics Immunosuppressants NSAIDs	210
Phototoxicity	Antibiotics, NSAIDs	45
SJS/TEN	Antibiotics, Anti-epileptics Immunosuppressants NSAIDs	47
Cytopenias	Antibiotics, Anti-epileptics Antimycotics Immunosuppressants	98
Leukopenia	Antibiotics, Anti-epileptics Antimycotics Immunosuppressants	80
Thrombocytopenia	Antibiotics, Anti-epileptics Antimycotics Immunosuppressants Antimalarias	31
Haemorrhage	NOACs	50
CNS-toxicity	Antibiotics, Antivirals Antimycotics, Interferone Antimalarias	76
Hyponatremia/SIADH	Antidepressants Anti-epileptics	32
Liver toxicity	Antidepressants Anti-epileptics Immunosuppressants Statins NSAIDs	96
Metabolic symptoms	Antidepressants Neuroleptics	17
Renal toxicity	Immunosuppressants	33
Pancreatitis	Antibiotics, Immunosuppressants	40
Narcolepsy	Swine flu vaccine Pandemrix	66
Atypical fracture	Bisphosphonates	54

4.2 SNP Microarray Genotyping

SNP microarrays mainly genotype SNPs but are generally cheaper and require less data pre-processing — that is processing steps needed to prepare genomic data for analysis — than whole genome, exome, or targeted sequencing. The drawback is that they only genotype nucleotides (*base*) at predetermined positions in the genome. SNP genotyping works by hybridisation of the sample DNA and designed complementary probe sequences affixed to silica beads[17]. These beads are located in wells on a microarray panel where each well contains beads of unique probes. When single-strand DNA fragments of a sample are washed over the panel, the fragments hybridise to the complementary sequence on the probes. These probes are designed in such a way that the hybridisation will end a single base short of the position of interest. After this, a chain-terminating nucleotide (ddNTP) coupled with fluorescent probes with unique fluorescent spectra per nucleotide-type is introduced. The ddNTP compliments the nucleotide of interest while terminating the probe and preventing further nucleotide additions. Fluorescence emissions are measured when each well is excited via laser. The mixed emission spectra correspond to the fluorescence of the probes present. Homozygote variants produce a single spectrum, while a heterozygote results in a mix. The positions of what well contains which probe is known, allowing interpretation of the genotype at each genomic position based on the emitted spectra.

4.2.1 Imputation of Genetic Variants

Though the SNP microarrays analyse many hundreds of thousands of positions, they will only cover a fraction of all known human SNPs. However, many other variants can be *imputed* from them. Genetic variation is not inherited independently but in *haplotypes*. Haplotypes refer to sets of genetic variations that tend to be inherited together, generally due to parts of chromosomes being less likely to be split by chromosomal crossover during meiosis. With enough samples, genetic variation can be inferred via the probabilities of simultaneous inheritance. Using the obtained genetic variation, variants not targeted in the SNP microarray can be imputed to the data set[18].

4.3 Whole Genome Sequencing

While SNP genotyping can provide substantial amounts of genetic information, it does not yield any novel genetic variations, as array-based genotyping will only convey the genetic information it is designed to obtain. On the other hand, whole genome sequencing results in genetic information from the whole genome without selecting its targets. There are multitudes of sequencing methods and all revolve around obtaining different sub-sequences from

the full nucleotide sequence (e.g. a chromosome). By obtaining numerous sub-sequences, called *reads*, the full sequence can be reconstructed by aling the reads either to a reference or to each other. There is a subdivision in sequencing methodologies divided by the number of nucleotides obtained in each read. These methods are categorised as short-read sequencing (hundreds of bases per read) and long-read sequencing (tenths of thousands of bases per read).

4.3.1 Library Preparation and Sequencing

Illumina HighSeq X is a short-read sequencing method that employs flowcells with billions of nanowells. Each of these wells has two surface-bound complementary oligonucleotides, also known as the *forward* and *reverse* adapters. These adapters' purpose is to capture and amplify fragments of sample DNA. However the sample DNA needs to be prepared correctly, in a process called library preparation.

The DNA is extracted from the patient-samples, for example blood or saliva. The DNA is then introduced to silica beads with bound transposons (BLT) to their surface[19]. The transposons bind the sample DNA and fragment it into predictable lengths and then these transposons add the complementary adapters to the ends of the fragments, Illumina calls this process tagmentation. The expected insert size (the distance between adapters) can be controlled by altering concentrations of DNA and reagents. In this cohort the insert length was set to 350 bp. A subsequent purification using solid-phase reversible immobilisation (SPRI) beads can remove outlying, unwanted fragment lengths[20].

A solution containing the adapter-ligated fragments is washed over the flowcell. The double-stranded fragments are denatured and then renatured. Some single-stranded fragments will pair with the oligonucleotides in the nanowells in the flowcell surface instead of re-hybridising with their complementary strand. The enzyme DNA-polymerase is introduced — together with free DNA nucleotides (dNTPs) — and rebuilds the complementary strands, from the adapter-hybridised primers resulting in copies of the fragments adhered to the flowcell. The fragments are denatured again and any free oligonucleotides are washed away, leaving only the adhered single-stranded copies.

Through the bridge amplification process the fragments are multiplied, creating distinct clusters of copies in the nanowells. These clusters are created when the complementary primer at the end of the fragment hybridises with a complementary adhered adapter nearby. DNA-polymerase and dNTPs are reintroduced, resulting in a doubling of the fragments. This process is repeated until there are clusters of copies of different fragments across the flowcell. The fragments of the reverse strand are cleaved and washed away, leaving only the forward strands. A free reverse primer is introduced, which hybridises with

the forward strands. Subsequently, ddNTP coulpled to fluorescent probes, and DNA-polymerase are added. The DNA-Polymerase adds the ddNTP corresponding to the complement of the next base after the hybridised reverse primer. The fluorescence emission spectra of each cluster are measured and the first base of the corresponding fragment is determined. After recording the emission and the positions, the probe blocking further complementary ddNTPs is cleaved away. This allows the next base to be added and its emission to be recorded. In a process called sequencing by synthesis, the cycle is repeated, sequencing one base per cluster and repetition.The cycle is repeated until the desired read-length of the fragments is obtained.

In a process called base-calling, these emissions are categorised as belonging to a specific nucleotide. A quality measurement, known as the sequencing quality, is then estimated based on the properties of the cluster and empirical data models studied by Illumina. Many sequencing methods quality scores are known to have certain biases but these can be re-calibrated through a process called base quality score recalibration. As many sites of genetic variation are known, a model of covariation is estimated based on the sequencing errors at the locations of known variation. The model of covariation can then re-calibrate the quality scores accordingly.

4.3.2 Short-Read Whole Genome Sequence data Pre-processing

Before any testing of differences in genetic variation can be done, genetic loci that differ between individuals need to be identified and filtered for systematic errors. This is the pre-processing of the sequencing data.

4.3.3 Alignment to a reference genome

First, to find the genetic differences between individuals, the fragments obtained through sequencing needs to be mapped to the correct position in relation to each other. In a process called *alignment*, each fragment is assigned a position at its most likely position along a reference genome.

Reference genomes are standardised genomic sequences unique to a species, in our case humans. These reference genomes serve as a foundation for studying genetic variations between individuals as comparing the sample genome to the reference genome, the differences between the two are obtained, defining genomic variation as deviation from the reference. With the increase in human genomes sequenced and studied, the human reference genome has become more complete and well-suited for its purpose. The reference genome used in our studies was the Genome Assembly *GRCh37*, or Genome Reference Consortium Human Build 37[21].

The human genome consists of 6.27 Gigabase pairs and with each base being aligned to 30 fragments, the alignment procedure is computationally

heavy, requiring both efficient algorithms and access to computational clusters.

The BWA-MEM[22] algorithm is the most common method used for short-read sequencing data.

4.3.4 Variant calling and structural variation

The process of determining the most probable base at each position in the alignment is performed using the BAM-files. If 30 reads all show a high-quality base at the same position, the sample is assigned a homozygous base for that position through *variant calling*. Any position that differs from the reference genome is considered a variant, and a variant quality measurement is assigned based on the sequencing quality of the reads. Variants can take the form of single nucleotide polymorphisms (SNPs), shorter insertions or deletions of bases known as *indels*, or larger duplications, insertions, deletions or re-arrangements of genomic segments (referred to as structural variations). The likelihood and variant quality are partially determined by previously observed genetic variation, such as those documented in the 1000 Genome Project, as already observed variants has lower probabiliy of being sequencing errors.

In larger cohorts, additional information can be obtained from other samples in the cohort through joint genotyping. This process takes into account that if sequence variations are observed in one sample, it increases the probability of variation at the same position in another sample, while reducing the likelihood of variation being solely due to sequencing errors or missing data.

In general, this is how variant calling works. However, the most popular methods are more sophisticated in their methodology. A widely used variant caller is GATK HaplotypeCaller[23] which first identifies regions with significant variation and after that realigns that region in a De Bruijn-like graph approach. Traversing the graphs, all haplotypes within the region can be estimated. These haplotypes can then be used as the hidden states in a hidden Markov model and the probability of each variant belonging to the most likely haplotype can be derived as a variant score.

The other, more recent, approach is DeepVariant[24] from Google. DeepVariant is a deep learning approach to variant calling using convolutional neural networks to discern correctly sequenced variations from sequencing errors, misalignments, and other sources of errors.

After genotyping calls are made and variant quality is assigned, re-calibration of these qualities takes place. The Variant Quality Score Recalibration method (VQSR) from GATK utilises Gaussian mixture models to estimate clusters of variants, initially using previously known variants (from databases such as dbSNP[25]) observed in the sample. By employing this technique, the model establishes profiles for high-quality variants, which can be used to com-

pare and assign re-calibrated quality measurements to new variations detected within the sample, based on their similarity to these established profiles.

Structural variation (SV) necessitates distinct methodologies compared to simple variants like SNPs and indels. Generally, three types of data are utilised for SV calling: *read-pairs*, *read-depth*, and *split-reads*. Read pairs are paired sequences acquired from the same fragment, but with opposite sequencing directions (a forward read and a reverse read). By analysing the positional distance and orientation of these mapped paired reads, larger variations can be inferred by considering the approximated insert length of the fragments. Additionally, the read-depth data provides valuable information about SVs. For instance, if a fragment has been duplicated we would expect a deviation from the expected read-depth of 30, as more fragments would be sequenced. Conversely, larger deletions would result in fewer fragments being mapped to the region. Split-reads, on the other hand, exhibit a split in their mapping. In the case of insertions, the inserted fragment would not map to the genome without the inserted piece, as it would not be present in the reference genome. In a deletion the mapped read would be split, leaving an empty region in the middle.

4.3.5 Quality control (QC) and filtering

A high percentage of the variants identified are true at this point, but even a half a percent of erroneous calls would result in 25000 wrongly typed variants, as the average human carries $\sim$5 million genetic variations compared to the reference genome[26]. To lower the number of erroneous calls, variants are filitered on cross-sample statistics such as variant genotyping rate (successful versus failed calls), testing of HWE among controls, or other sequencing statistics.

The genotyping rate of a variant is the fraction of samples where sequencing was successful in calling a variant or wildtype. The minimum genotyping rate allowed is commonly set to 5% missing calls among samples. A higher number of missing data would indicate that the variant loci is difficult to sequence or is prone to incorrect mapping of fragments. Moreover, variants that did not have viable sequence coverage of the loci in more than 95% of the samples are commonly excluded. Variants only observed in single samples within a cohort are called *singletons* and are usually excluded as well, as they have a higher risk of being erroneous. If the sum of singletons per sample is larger when compared to other samples in the cohort, this could indicate that the sample has poor quality.

In a large cohort, samples are sequenced in batches since all samples cannot be sequenced with sufficient coverage simultaneously. However, the batching of samples risks introducing systematic batch-specific variations, resulting from differences in sequencing protocol execution or encountering variations

due to batch-specific differences in DNA-sample quality. Therefore it is important to spread the cases and controls evenly across the different batches so potential batch effects do not cofound case-control differences.

In the end, the process of QC and filtering determines which genomic variation and which samples are of high enough quality to be tested, without underlying sequencing bias.

4.3.6 Genome Annotation

When the processing of the genetic information is complete, the information from the samples is obtained. In a process called *annotation*, the variation obtained can be analysed to predict its impact and previous knowledge can be annotated to the variation.

When estimating impact of genetic variation, tools such as SnpEff[27] and VEP[28] can estimate the direct variant effect, such as effects on protein transcripts, amino acid alterations, and the presence of regulatory regions. These tools can also annotate variations with secondary information such as previously reported correlating phenotype and the frequency observed of variants in population studies.

4.4 HLA imputation

The Human Leukocyte Antigen (HLA) complex is a set of genes that encodes immune-related cell-surface proteins involved in immune response and suppression. Genetic variation in HLA has been linked to both disease and drug-response. The HLA region is highly polymorphic due to the Red Queen dynamic[29] between humans and parasites. However, these polymorphisms come in haplotypes, also called *HLA-types*.

The nomenclature of these HLA-types is updated by an international committee and the classification is based on loci (referring to a specific position on a chromosome) and serology(antibody tests for variants of antigens)[30]. The HLA-type HLA-A*01:02:03 conveys that the type is for the gene HLA-A, where 01 represents the serological type, 02 describes the general haplotype and 03 is the class of additional genetic variation. Genomic variants that do not provide information about the phase of the variant need additional analysis to infer HLA-types. Without the phase — what strand the variant is located on — of a variant, the HLA haplotypes are not available. Assigning a type to an HLA gene using variation from both alleles intermixed would risk incorrect typing. However, the most probable HLA-type can be inferred by referencing frequencies of HLA-types within the population and the haplotypes previously observed. Such analysis is available in methods such as SNP2HLA[31].

4.5 Hardy-Weinberg equilibrium

It is common practice to exclude variants among controls deviating from the Hardy-Weinberg equilibrium[32] (*HWE*) as it can indicate underlying errors in the data[33]. HWE states the ratio of expected homozygous, heterozygous, and wildtype alleles in a population under random mating. Given that an allele has a frequency of $p \in [0,1]$ in a population, the frequency of the wildtype would be $q = 1 - p$. HWE states that the expected ratio of homozygous, heterozygous and wildtype individuals in the next generation should be p^2, $2pq$ and q^2 respectively. Divergence of the observed number of heterozygous individuals from the expected number, given the variant frequency (VF), under HWE can indicate that a variant is prone to genotyping error or affected by underlying population stratification. However, HWE should only be evaluated among controls in case-control studies as it would otherwise risk excluding correlating variants. The variants that correlate with the phenotype would be overrepresented among cases meaning that there would be a case-specific VF and a control-specific VF. If HWE is evaluated simultaneously among all samples the causal variant risks exclusion since the variant is very unlikley to have the exact proportions of zygosity to follow HWE.

4.6 Logistic regression on genetic variation

Perhaps the most straight forward way to test for genetic correlation of a binary phenotype is through logistic regression. For one genetic variant, this is defined in equation 4.1

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{4.1}$$

The outcome is represented by $Y \in [0,1]$. This will be the absence or presence of an ADR, e.g. a patient with narcolepsy from Pandemrix is 1 and without is 0. This means that cases are represented by 1 and controls by 0. X represents the genotype value of the genetic variant being studied. The numerical interpretation of the genotype depends on the chosen genetic model where the model can be additive, recessive, or dominant. Depending on the model, the genotype value will be encoded as per table 4.2.

Table 4.2. *The value representation of variant zygosity depending on genetic model*

	Wildtype	heterozygous	homozygous
Additive	0	1	2
Recessive	0	0	1
Dominant	0	1	1

In the earlier examples, the risk of toxicity during 5-fluorouracil treatment should be analysed using an additive model. A patient heterozygous for a

decreased or non-functioning variant of DPYD will metabolise 5-fluorouracil faster than a patient homozygous for the same variant. Conversely, the autoimmune reaction due to Pandemrix is correlated with the presence of HLA-DQB1*06:02 and not the zygosity, meaning that the model of choice should be dominant.

Fitting the data to the model yields a regression coefficient for each analysed variant. The regression coefficient for the genetic variation is recalculated as a standard score, or *Z-score*. The Z-score value obtained will reflect the number of standard deviations the observed value deviates from no effect, which is a regression coefficient of zero. Assuming that the probability follows a normal distribution, the probability of obtaining a value *at least* as extreme as the observed can be calculated. This probability is also known as the *p-value*.

However, this value can be misleading as it can covary with variables not included in the model. Common covariates are age, sex, BMI, etc. but, arguably, the most important covariate when comparing genetic variables is genetic variation due to population stratification.

4.7 Population stratification and stratification mitigation

Population stratification arises when there exists systematic differences in allele frequencies depending on genetic ancestry among individuals. These systematic differences are evident between populations and arise from the isolation and migrations of those populations. Humans inherit the absolute majority of their genetic makeup, except from mutations or retrovirus infection, and humans have populated even the most remote places on earth. Every time humans migrated to a new region population bottlenecks occured influencing which genetic variances that were more or less likely to be inherited by the next generation. Through successive generations and migrations, the frequencies of alleles and genetic variation have diverged between separate populations. Therefore, if a genomic case-control study does not account for population structure among the samples, any genetic correlation between cases and controls could be confounded by the population structure differing between them[34].

Fortunately, there are many ways to control for this and one such method is using principal component analysis, *PCA*, on the genotype covariance matrix[34]. PCA decomposes the genotype covariance matrix along the axis of maximum total variance decreasingly[35]. The genotype-covariance matrix is the sample covariance matrix calculated on the normalised genotype values. This means that the first principal component (PC) is the axis spanning an N-dimensional — one dimension for each sample — space that maximises the covariance explained by the linear combination. Once that PC (also called PC1) has been established, the variance explained by it is removed from the genotype matrix

by subtracting the linear combination PC1 times the samples projected. The process is repeated, resulting in the linear combinations from PC1 to PCN.

When including principal components as covariates in logistic regression (such as equation 4.1) the values used are each sample's projected value along each PC. This results in equation 4.2, where X_{PCj} are the projected values along the jth PC.

$$P(Y=1) = (1+e^{-(\beta_0+\beta_1*X+\beta_{PC1}X_{PC1}+...+\beta_{PCj}*X_{PCj})})^{-1} \tag{4.2}$$

This means that most of the systematic variation in the data can be explained by the regression coefficients of the principal components, instead of the covariance being explained by the regression-coefficient of the genetic marker.

4.8 Multiple hypothesis testing

The classical significance threshold is a p-value of 0.05, which means that given the test results, there is a 5% probability of generating a test statistic at least as extreme as the obtained value. So when testing a single variant the risk of incorrectly rejecting the null hypothesis is 5%. This means that when performing 100 tests, five are expected to falsely reject the null hypotheses. When testing millions of variants across the genome, the multiple-test problem needs to be addressed.

One common way of adjusting the significance threshold is Bonferroni correction[36]. This correction states that the threshold should be $p < \frac{\alpha}{m}$ where $\alpha = 0.05$ is the classical significance threshold and m is the number of hypotheses tested.

A widely used significance threshold is estimated based on Bonferroni correction, the *genome-wide significance threshold* [37]. The genome-wide significance threshold, usually defined as $p < 5*10^{-8}$, is the Bonferroni correction of the number of approximated independent genetic variants in humans with an allele frequency greater than 0.05.

Another approach is to control for the positive false discovery rate. This estimates the expected ratio of false positives obtained at a certain p-value. When plotting the probability distribution of all p-values obtained for multiple tests, the distribution could be bimodal. This bimodal distribution is obtained from two groups of tests, one where the null hypothesis is false and one where it is true. The *p-value distribution*, that is the distribution obtained when the null hypothesis is true, is per definition uniform. For tests where the null hypothesis is false, p-values are more probable to be closer to zero than one. This way, the contribution of the uniform p-value distribution can be estimated in the obtained bimodal distribution. Using this distribution the expected false positive rate at observed p-values can be obtained, this rate is denoted as the *q-value*[38].

4.9 Gene based variant analysis with SKAT-O

SKAT-O (sequence kernel association test with optimal weights)[39] is a gene-based test developed for uncommon variants, generally meaning variants with VF below 0.05. In contrast to the common variant testing, SKAT-O tests sets of variants reducing the number of tests conducted compared to testing each variant individually. These variant sets usually contain variants located in the same gene coupled with selection and weighting criteria. These criteria are commonly: only include protein-altering variants and weighting on the resulting predicted severity protein function alteration or the variants MAF. SKAT-O consists of two separate tests, a burden test and the SKAT (sequence kernel association test)[40] non-burden test and the output test statistic is a weighted average of both.

The burden test performs a multivariable sum regression — multivariate sum logistic regression for binary phenotypes — on an additive genetic model of the selected genetic variance and covariates, such as principal components. The summed genotype values of all selected variations are modeled against the phenotype. This means that the model assumes that all variants are assumed to have the same effect, resulting in a single regression coefficient for all selected variants. The genetic component of the equation becomes, $\beta \sum_{j=1}^{m} w_j g_{ij}$ where w_j is the weight of the j:th variant and g_{ij} is the number representation of the variants in the ith sample, depending on the genetic model. The null hypothesis, that there is no correlation between the phenotype and genetics, becomes $\beta = 0$. The drawback of this approach is that it assumes that all variants have the same *direction*, i.e. that all included variants in a gene are only either causal or protective. Without prior studies this cannot be assumed, therefore the burden test is presumed to be conservative.

SKAT, on the other hand, views each genetic variant's regression coefficient as an independent variable with a mean of 0 and a variance of $w_j^2\tau$. This means that the null hypothesis $\beta = 0$ is equivalent to $\tau = 0$. Given a logistical model, the estimation can be expressed as the weighted sum of a single variant test statistics $Q_s = \sum_{j=1}^{m} w_j^2 S_j^2$ where S_j is the test statistic for the single variant $B_j = 0$ formulated as $S_j = \sum_{i=1}^{n} g_{ij}(y_i - \hat{\pi}_i)$. Here, $\hat{\pi}_i$ is the null model i.e. the phenotype is only explained by the covariates.

Burden test can be more powerful than SKAT when testing regions with variants of the same effect but suffers if there are both protective and causal variants present[39]. The SKAT-O method performs both tests and instead reports the test statistics as the weighted sum that minimises the p-value.

4.10 Causality of Predisposition and Enrichment Testing

The fact is that it is not the single genetic variations that cause the phenotype but rather the phenotype is a consequence of the organism's environment and the emergent properties of its cellular interactions and mechanisms. In turn, these cellular processes emerge from the interplay between all molecules present, from protein-protein interactions, transcriptomic and epigenomic regulation, metabolic chains, and much more. This is where the single genetic variations can result in a change within a pattern (e.g. expression of mRNA), and that change cascades from molecule to function, from function to pathway, and finally from pathway to organism. If the phenotype is an ADR, the obvious exposure to a drug is required, but also the biological predisposition of the patient. For example, if looking at the *ADME*(absorption, distribution, metabolism, and excretion) networks of the immunosuppressant prodrug azathioprine[41], changes in this network can lead to myelosuppression as a result of drug exposure. However, there are known variations in different genes in this network predispositioning the ADR[42]. This implies that the ADR occurs in response to any singular or polygenic variation that causes comparable changes in a biological process. The resulting conclusion becomes, that when selecting tests, or genetic targets for testing, they should be selected to prioritise shared biological interactions and processes.

To evaluate the common ground of, for example, 10 correlating genetic variations in a case-control study, one must first determine what the common ground is. Suppose each of these genetic variations all map to different genes. This set of 10 genes can be tested for overrepresentation, or *enrichment*, among known categorisation of genes such as KEGG pathways [43], disease ontologies[44], or other gene ontologies[45, 46]. The enrichment results could convey the underlying process that predisposes the ADR. Fortunately, these tests are easily performed using analysis packages such as ClusterProfiler[47].

4.11 The Interactome

Another way to detect shared processes, due to the modularity of biological processes among genes, is testing for an overrepresentation of *interactions* between genes. Generally, these interactions are the molecular interactions within a cell or an organism, and the complete list of the interactions would be called its *interactome*. The interactome consists of interactions between many different types but the most commonly studied are protein-protein interactions (PPIs)[48]. As these PPIs create networks where each node is a protein and every edge indicates interaction between proteins, the resulting network (also called a graph) can be analysed. The previously identified correlated genes can be studied if they are situated within subgraphs, where a subgraph is sets of

nodes and edges that are connected in a smaller graph which is a subset of the PPI network. The size of the subgraph can then be compared to the expected size when randomly selecting a comparable set of genes. The shared interactions among a set of proteins can therefore be compared to random chance where divergence would indicate a shared impact on biological processes.

4.12 Modularity-Based Distance

In terms of genetic predisposition, it has been shown that most diseases and disease outcomes are not governed by variation in single genes but rather are due to the polygenic variation leading to alteration of biological processes[49]. As such, it has been shown that disease-associated genes cluster in the PPI network[50, 51]. With this in mind, the set of genes associated with a drug can be regarded as a whole rather than a singular gene. For example, the distance — a metric based on the shortest paths between protein-coding genes in a PPI network — from the targeted genes of a drug to the disease-causing genes can be viewed as a measurement of efficacy[52], but also suggest potential repurposing of drugs[53].

4.13 Drug, Symptom, and Protein-Protein Interaction databases

Importantly, this approach requires previous knowledge such as drug-gene association data, ADR-gene association data, and protein-protein interaction data. As the amount of information and research is ever-growing, being up to date on all findings is a challenging task. Fortunately, there are databases dedicated to gathering and curating knowledge relevant to interactome analysis. The Drugbank database[54] collects information on known and proven drug-gene interactions, chemistry, and genetic interaction of drugs. The Human Phenotype Ontology (HPO) database[55] collects data on genetic and disease associations of phenotypes, such as symptoms. The STRING database[56] contains known and predicted, direct and functional, protein-protein interactions (PPI). STRING assigns each protein-protein interaction with a score, ranging from 0-1000, estimating the probability of the interaction given the evidence available. In 2006 the completeness of the human PPI networks was estimated to be around 10% [57].

5. Results

5.1 Processing and Testing of WGS data

The samples in the Swedegene WGS cohort was sequenced using the Illumina HighSeq X platform. The libray preperation (Section 4.3.1) was conducted by National Genomics Infrastructure — Uppsala (NGI-U). The insert length was set to 350 bp and the sequencing were perfomred to achive an average coverage 30X with a read length of 150 bp.

The specific target of 30 fragments per base was set as it was estimated to have a sufficient accuracy over a majority of SNPs in the human genome [58].

The sequened reads were aligned (Section 4.3.1) to the reference genome GRCh37 using the BWA-MEM[22] algorithm, resulting in *Binary Sequence Alignment Maps* (BAM) files. The sample BAM files from the Swedegene WGS cohort was delivered by NGI-U. From the BAM files, re-calibrated quality tables were estimated using GATK[59] 3.8 BaseRecalibrator. Preliminary variant calling of each sample was conducted using the GATK HaplotypeCaller[23] in GVCF (Genomic Variant Calling File) mode to prepare each sample for joint genotyping (Section 4.3.4). The individual GVCFs were split along predetermined genomic intervals and joined using GATK CombineGVCFs, resuting in multi-sample GVCFs containing the 978 Swedegene samples. These multi-sample GVCF were jointly genotyped using GATK GenotypeGVCFs. The resulting segmented multi-sampled VCF (Variant Calling File) from GenotypeGVCFs were concatenated into a whole-genome, multi-sample VCF. The variants in the full VCF were recalibrated using GATK VQSR as described in Section 4.3.4. VQSR was executed twice sequentially, once for SNPs and once for indels.

Multiallelic SNPs and indels were split using BCFtools. BCFtools was used to filter out variants that failed quality criteria, such as variants failed by VQSR. Variant-based missingness was calculated using KING autoQC[60] and any variant with over 5 % missing calls was filtered out as described in Section 4.3.5. HWE was calculated on the controls using PLINK2.0[61] and any variant with a p-value below 10^{-8} was filtered out. This resulted in a multi-sample VCF containing $\sim$45 million variants.

The finished VCF was annotated using ANNOVAR[62] and VEP[28] while structural variation was called using the software Manta[63].

5.1.1 Implementation

For the many steps needed in this analysis the usage of a workflow management system was paramount. For this task, the Python-based workflow man-

agement system Snakemake[64] was selected. In Snakemake each step in the chain of procedures constituting the full analysis is implemented as its subroutine, called a rule. These rules expect an input file to exist and will expect an output file to be created once the code defined in the rule is executed. If an input cannot be found for a rule, the remaining rules are searched for an output matching the expected input. This way, if a single rule fails it will only be a small part of the whole tree of procedures contained in the analysis failing and the parts that need to be reanalysed are kept to a minimum. Despite of this, the full analysis required roughly 1.6 million core hours on Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) secure computational cluster Bianca. However, half of this computational time was spent rerunning the analysis trying to eliminate the batch effect between Swedegene and SweGen (elaborated on more in Section 5.1.2)

Pre-processing Implementation
The full implementation of the WGS pre-processing pipeline and analysis can be found at `https://github.com/JoelAAs/WGS_pipeline`.

The Snakemake implementation of pre-processing are detailed in the following three steps:

1. main/ pipeline_part1 / SnakeFile_batching: Defines the processes for base quality recalibration and preparation of joint genotyping

2. main/ pipeline_part1 / SnakeFile_rest : Contains joint genotyping, variant quality recalibration, failing variant filtration, collection of QC metrics, and Annotation. This requires the first workflow to be completed.

3. main/ pipeline_part2 / SnakeFile_filter : Contains the filtering steps based on genotyping rates, HWE failing variants as well as filtering out variation failing cohort-specific variation based on the QC metrics defined by the previous steps.

Genetic testing implementation
The main genetic tests are implemented in main/ pipeline_part3 / SnakeFile_assoc which outlines the single variant testing and allele frequency calculations. The SKAT-O testing was done on a gene-basis and the workflow can be found at main/ pipeline_part3 / src / skato . smk.

Additional analysis and visualisation for paper IV can be found at `https://github.com/JoelAAs/WGS_testing`.

5.1.2 Batch-effects between Cases and Population Controls

The batch-effect was the reason the SweGen population controls were not used in **Paper IV**. All Swedegene cases were sequenced in batches in close

succession and processed in the same lab as the SweGen samples, but were sequenced a couple of years later. Figure 5.1 shows tests of common variants between Pandemrix-induced narcolepsy and SweGen controls. The spike at HLA-DQB1*06:02 on chromosome 6 shows the variants associated with Pandemrix-induced narcolepsy. The other significant variants were due to batch-effects between Swedegene and SweGen. The genetic correlations observed when testing any hypothesis would be the sum of the correlations due to the batch-effect and the correlations between cases and controls. These contributions cannot be separated unless the cases existed in both Swedegene and SweGen. Therefore the SweGen controls could not be used in the analysis.

So even though the sequencing was performed on the same machines, at the same center using the same protocols, the root of the batch effect could not be determined. However, there had been some updates in the Illumina sequencing chemistry in the years between the cohorts and additional minor differences in sample handling that could have led to this systematic difference.

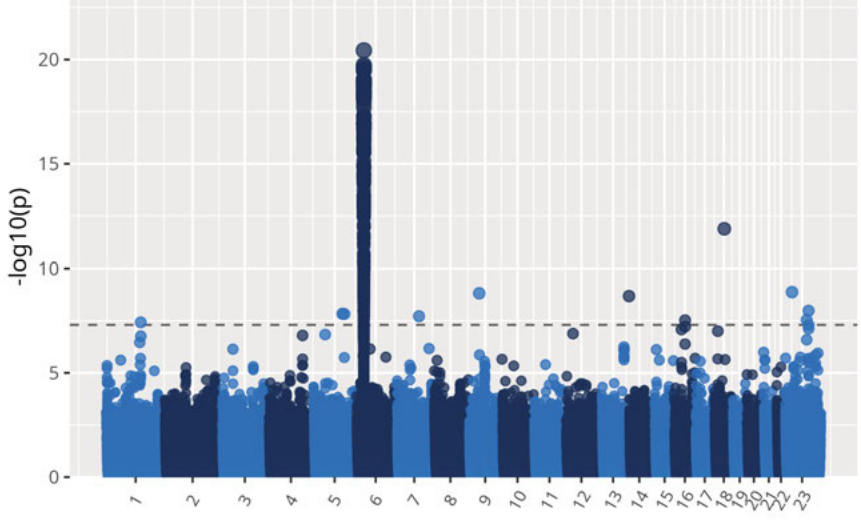

Figure 5.1. Logistic regression test of variants with VF above 0.1 for Swedegene narcolepsy cases and SweGen controls. The dotted line is the genomewide significance threshold $p < 5 * 10^{-8}$. The spike at chromosome 6 shows the variants correlating with HLA-DQB1*06:02, which is associated with Pandemrix-induced narcolepsy.

5.2 Phenotype Clustering and Genetic Target Suggestion Utilising Network Biology

One of the main differences between the studies **papers II-IV** is how they handle the interplay between multiple hypothesis problems and the scarcity of

data, each with their pros and cons. **Paper II** presents genetic associations with azathioprine-induced pancreatitis in patients with Crohn's disease. The genetic targets studied were chosen as they had been previously reported as associated. The fact is that these genetic variants had such a high correlation that they could be viewed as a single test therefore eliminating the need for any correction for multiple hypotheses. The obvious drawback to this approach is that it will never discover any novel correlations.

Paper III and **Paper IV** follow suit but are studying a larger set of *candidate genes* that have previously been associated with, or have been theorised to be involved in, the phenotype. None of these studies found any associated genetic correlation among the candidate genes. As a secondary aim, exploratory testing was performed to study any genetic variation that passed basic filtration criteria (such as major allele frequency thresholds, variants resulting in protein alteration, exon-proximity, etc.). The exploratory analysis drastically increased the number of hypotheses, demanding a much more stringent correction of test results, thereby leaving both studies underpowered.

The way **Paper III** and **Paper IV** included samples was also quite different. By collaborating internationally,**Paper III** managed to recruit 248 cases (53 from Swedegene) with the same ADR due to a collection of five drugs (four of which have high chemical similarity). The cases from **Paper IV**, on the other hand, were collected with the candidate genes in mind. The initial hypothesis was that genetic variation in blood-brain barrier transport proteins lead to drugs, or metabolites of drugs, entering the brain and causing CNS-toxic reactions. This hypothesis led to a heterogeneous cohort of 66 cases with 29 different CNS-related ADRs due to 33 different anti-microbial drugs. This heterogeneity among the samples made to the secondary aim, the exploratory analysis, challenging to interpret.

This problem was the motivation behind **Paper V**, which presents data-driven methodology to suggest or validate selections of drugs and ADR in studies. Moreover, the methodology allows for suggestions as to what genes study and which biological processes that might be of interest, based on the drugs and ADRs included.

5.2.1 Implementation

The full analysis pipeline is implemented as a Snakemake workflow and can be obtained at `https://github.com/JoelAAs/phenotype_mapping`.

The results in **Paper V** can be replicated by following instructions outlined in the *replication* branch.

5.3 Summary Paper I

SWEDEGENE—a Swedish nation-wide DNA sample collection for pharmacogenomic studies of serious adverse drug reactions. The Swedegene initiative started in 2008 and has established a biobank of adjudicated cases of predominantly rare ADRs to investigate possible genetic predisposition. Swedegene aims to collect enough cases of ADRs to reach sufficient sample sizes. The cases are mainly recruited through the Swedish Medical Products Agency's registry of spontaneously reported ADRs, but also through direct recruitment from healthcare facilities and advertising. Each participant, whose blood or saliva samples are stored within the biobank, has answered a questionnaire relating to general health, lifestyle (such as smoking, alcohol consumption, etc.), herbal medications, allergies and a separate questionnaire specific to the type of ADR in question. These answers are then compared to and complemented by the medical records of the patient to ensure the causality of a drug-ADR relationship is deemed possible.

This paper outlines the goal and methodology of recruiting, interviewing, and curating patient- and ADR-data in the Swedegene project. The Swedegene project is a Swedish nation-wide biobank established to facilitate studies of genetic risk factors of ADRs. At the time of publication, the project had recruited, approved, and received DNA and medical information from about 2550 adults and 580 drug-exposed controls.

5.4 Summary Paper II

HLA variants associated with azathioprine-induced pancreatitis in patients with Crohn's disease. In this paper, we investigated and replicated the association between genetic factors of azathioprine-induced pancreatitis in patients with from Crohn's disease or ulcerative colitis, which are both diagnoses of inflammatory bowel disease (IBD). This study comprised three batches of SNP microarray data containing 19 cases where 11 suffered from Crohn's disease and 8 from ulcerative colitis. They were compared with 81 matched controls of which 39 were diagnosed with Crohn's disease and 48 with ulcerative colitis as well as an additional 4891 population controls. All variants failing HWE $p < 5 * 10^{-8}$, quality control (QC) checks or with a minor allele frequency (MAF) < 0.005 were excluded. The variants were used for genetic imputation using the Sanger imputation server[65] with the haplotype reference consortium panel (V1.1)[65]. Imputed variants with an impute2 quality measurement < 0.7 or a MAF < 0.0001 after merging the batches were excluded. HLA alleles were imputed with first and second field resolution using SNP2HLA[31] and the T1DGC European HLA reference panel.

Previous studies have shown that the genetic variant rs2647087 correlated to the HLA haplotype HLA-DQA1*02:01 - HLA-DRB1*07:01, which has been associated with azathioprine-induced pancreatitis[66]. However, as the

rs2647087 variant was not present in the post-imputation dataset, the high linkage disequilibrium (LD) variant rs2647085 (LD: 0.98 r², D' 0.99 vs rs2647087) was selected as a proxy. The outcome was modelled as an interaction between disease and genotype as per equation 5.1. Utilising contrast analysis we could show that the association between azathioprine-induced pancreatitis and IBD was only present in patients with Crohn's disease. Odds ratios and confidence intervals are shown in Figure 5.2.

$$\begin{aligned} P(Y=1) &= (1+e^{-\bar{\beta}\bar{X}})^{-1} \\ \bar{\beta}\bar{X} &= \beta_0 + \beta_{variant}X_{genetic} + \beta_{IBD-type}X_{IBD-type} + \\ &\quad \beta_{interaction}X_{genetic}X_{IBD-type} + \\ &\quad \bar{\beta}_{PCA}\bar{X}_{PCA} \end{aligned} \tag{5.1}$$

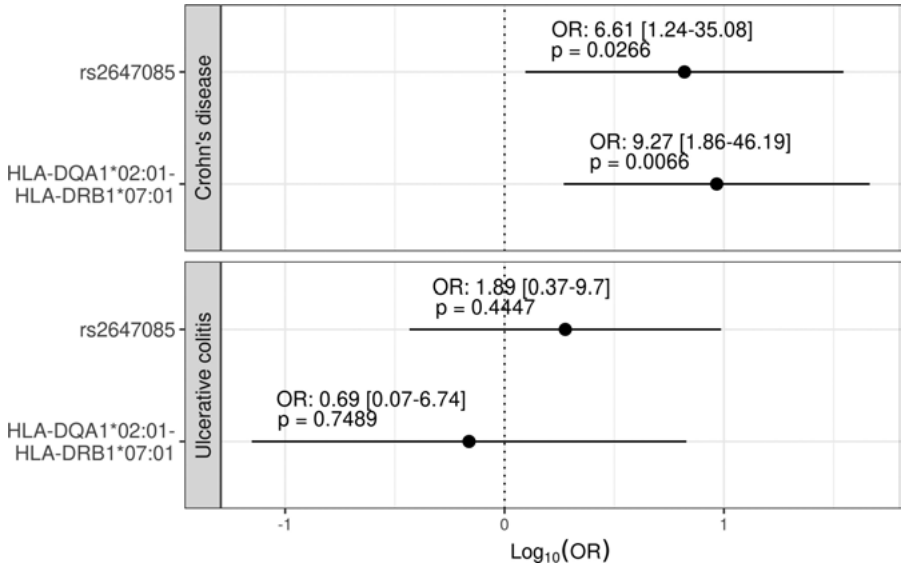

Figure 5.2. Odds ratio (OR) for azathioprine-induced pancreatitis from inflammatory bowel disease (IBD) type specific contrast tests. The term being contrasted is the genetic factor, while the IBD subtype is held constant for each type. Each test was performed for HLA-DQA1*02:01-HLA-DRB1*07:01 and the SNP rs2647085. Odds ratio, p-value and 95% confidence interval (CI) can be found above each line.

5.5 Summary Paper III

Gene-based association analysis of a large patient cohort identifies potential gene candidates for atypical femur fractures. This paper presents an exploratory study of genetic predisposition to the adverse reaction of atypical

femur fracture (AFF) during bisphosphonate use. In this study, gene-based analysis of 238 cases of AFF was performed using a discovery cohort of 185 cases recruited in the Netherlands and Australia and a replication cohort contaning 53 cases of AFF from the Swedegene cohort.

This study was conducted in two steps. First, predicted function-altering genetic variation in candidate genes was studied. These genes had previously been shown to be associated with monogenic bone disorders or mineralization and bisphosphate metabolism. Any variation with a high enough predicted deleteriousness, estimated by CADD[67] annotation, was tested using SKAT-O.

The second step was an exploratory analysis. Any protein-altering variants with a sufficiently low allele frequency (VF < 0.01) were selected, resulting in 43,101 variants in 13,198 genes. These variants were analysed using SKAT-O on a gene basis. This resulted in 57 associations with nominal p-values ($<$ 0.01) that were not significant after multiple hypothesis corrections. These results were corrected for sex and four principal components which resulted in 14 genes with p-values below 0.01.

Four of the 14 gene associations was discovered to be driven by an indel. This indel was deemed to be a sequencing error resulting in the exclusion of the four genes. The remaining ten genes was tested among the 53 cases in the replication cohort but resulted in no significant correlations.

5.6 Summary Paper IV

Whole genome case-control study of central nervous system (CNS) toxicity due to antimicrobial drugs This study investigated genetic associations between central nervous system (CNS) toxicity and antimicrobial drugs. The primary aim was to look for genetic association with candidate genes. These genes are known drug transporters in the CNS that are theorised to have an impact on CNS-toxic reactions by allowing the drug or metabolites of the drugs to pass the blood-brain barrier[68]. The secondary aim was to conduct an exploratory genome-wide study for any association.

Several genetic variants and tests were conducted. Single variation was first divided into two sets, those with a VF higher than 0.123 or lower. This VF threshold was set as it is the VF where a single homozygous individual is expected to exist among the 66 cases. The common variants were tested using logistic regression with four principal components as covariates. The less frequent variants were tested with SKAT-O where the weights were set as a sigmoidal curve based on the distance from an exon, meaning that all less common variations within exons were included. This was established to capture variation in the intron close to the exons to include possible splicing affecting variation. Additionally, a carrier/non-carrier correlation test of struc-

tural variation on a gene basis and a four-digit dominant HLA association test were conducted.

Among the candidate genes, no significant association was found. However, in the exploratory analysis, three genes: LCP1(q = 0.013), RETSAT(q=0.013) and SFMBT2(q=0.035) were significantly associated in the SKAT-O tests as seen in Figure 5.3.

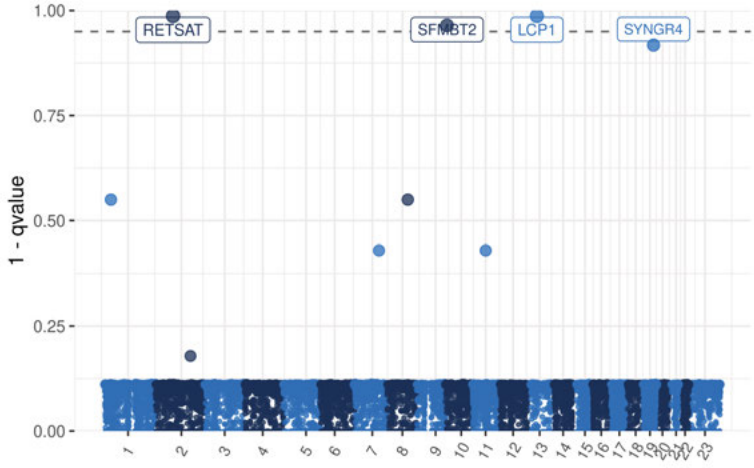

Figure 5.3. Gene association tested using SKAT-O in cases with CNS toxicity (n=66) vs controls (n=833). The number of genes tested was 19250. Uncommon variants (frequency < 0.123) within 140 bases from the closest exon/3'-UTR or 5'-UTR were included. The dotted line represents q-value 0.05.

The variation within the genes was selected and tested with a Bonferroni corrected significance threshold based on the 19250 tested genes added to the 106 single variants tested in the three genes ($p = \frac{0.05}{19356} = 2.58x10^{-6}$). Test results are shown in Figure 5.4.

In LCP1, the two variants rs6561297 and rs10492451 were individually associated with CNS toxicity ($p = 1.15 * 10^{-6}$, VF cases = 0.121, VF controls = 0.032, OR: 4.60 [95% CI: 2.51 –8.46], and $p = 1.15x10^{-6}$, VF cases = 0.121, VF controls = 0.032, OR: 4.60 [95% CI: 2.51 –8.46], respectively). Both are intronic variants, with rs10492451 being positioned in the distal enhancer-like signature EH38E1673884 (listed in SCREEN V2[69]) linked to LCP1.

L-Plastin (expressed by LCP1) is related to the transport of T-cell activation modules to the cell surface, and has been shown to be differentially expressed in suicide victims[70], compared to those with an accidental death. Additionally, a rare missense variant in LCP1 has been associated with schizophrenia[71]. RETSAT encodes retinol saturase which is involved in the metabolism of vitamin A, but the biological function of the enzyme product is unknown[72]. There is evidence of dysregulation of vitamin A metabolism in the aetiology of schizophrenia[73]. Nevertheless, due to the heterogeneity of antimicrobial

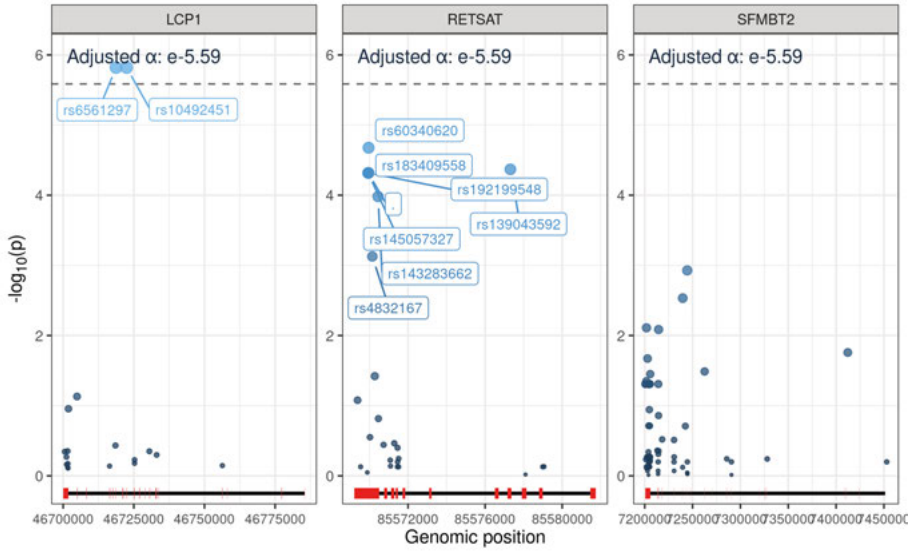

Figure 5.4. Variants associated in gene based tests using SKAT-O in cases with CNS toxicity (n=66) vs controls (n=833). Each variant was tested using logistic regression with principal components one to four, calculated on all genetic variation, as covariates. The dotted line represents a Bonferroni corrected significance threshold. Exons are marked in red and introns in black at the bottom of each graph.

drugs and CNS toxic reactions, providing a rationale of these results is challenging.

5.7 Summary Paper V

Network-Based Analysis of Protein Interactions among Drugs and Adverse Reactions: Identifying Phenotype-Groupings and Key Genes. This methodology paper presents a novel way to abstract the distance between drugs or symptoms while utilising the PPI network. There are successful methods for doing this when studying drugs versus diseases[52, 53]. It has been shown that diseases associated with genes form disease-specific clusters in PPI networks[50]. By comparing the distances from the relatively few genes targeted for drugs the similarity of action of the drugs can be compared. However, severe ADRs are rare by necessity resulting in fewer studies and fewer associated proteins and processes. Additionally, the genetic predisposition to an ADR is not guaranteed to be interacting, directly or indirectly, with the drug target but could be an effect of any gene it interacts with. This means that any shared PPIs or common PPI network topology could be of interest when comparing drugs or ADRs. To calculate the proximity between two sets of genes associated with drugs we calculate the probability of this question: *"If I pick*

a random gene from one set, and a random gene from the other set, what is the probability of picking the first gene in the interaction neighbourhood of the second?". The answer represents the pair of proteins as random variables whose sample space is defined by the drug or symptom.

We define the *neighbourhood* as the shortest possible path - where a path is a chain of interactions starting from one protein and ending with another and the path length is the number of proteins contained in that chain - from a starting gene to any other gene of path length D. As the probability of the genes interacting in the same processes decreases with each step, a maximum D_{max} was chosen in such a way that the likelihood of proteins sharing functionality remains high. Additionally, D_{max} should be sufficiently small to ensure that possible interactions remain meaningful and do not exceed the boundaries of the graph. The PPI from StringDB[56] consists of approximately 16800 proteins with approximately 30 interactions per protein at a *combined threshold score* of 700. A shortest path longer that 4 would risk reaching irrelevant interactions as well as the boundaries of the network. Therefore, we chose the longest possible shortest path $D_{max} = 4$.

With this definition and the earlier question in mind, we can represent the model as a Bayesian network as in figure 5.5. Here g_n and g_t are randomly selected proteins, where the subscript n represents the neighbourhood and t the target. The random variable p is the interaction path taken and y is a random gene chosen in the interaction path. Finally, the boolean operator for selecting the same protein within the target gene set and the interaction path gene y is defined as $m = \delta(y = g_t)$. The terms for drugs or symptoms is represented by T where the specific sample spaces defined by the gene sets are $T_n \in \{g_{n1}, ..., g_{nn}\}$ and $T_t \in \{g_{t1}, ..., g_{tm}\}$.

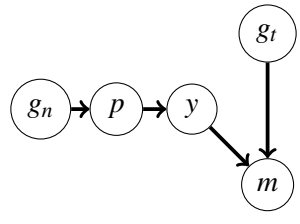

Figure 5.5. Bayesian network over the probability of finding matching proteins in the interaction neighbourhood of proteins from one protein set to another

The probability of selecting the same gene from the target gene set and the neighbourhood gene set, expressed as $P(m|T_n, T_t)$ can be calculated. This probability serves as a proxy for the functional similarity between the genes within T_t and those within the neighbourhood of T_n. The neighbourhood is defined in equation 5.2.

$$P(y|T_n) = \sum_{g_n \in T_n} \sum_{p} P(y|p)P(p|g_n)P(g_n) \tag{5.2}$$

The probability of selecting a random protein from the target gene set, $P(g_t)$, and picking the same protein, m, in the neighbourhood of the other gene set, $P(y|T_n)$ is expressed as Equation 5.3

$$P(m|T_n, T_t) = \sum_{g_t \in T_t} \sum_{y} \delta(y = g_t) P(g_t) P(y|T_n) \tag{5.3}$$

Using this probability as a measurement of similarity from drug A to drug B we obtain a directed graph that can be clustered via spectral clustering. The results of this clustering can be seen in Figure 5.6.

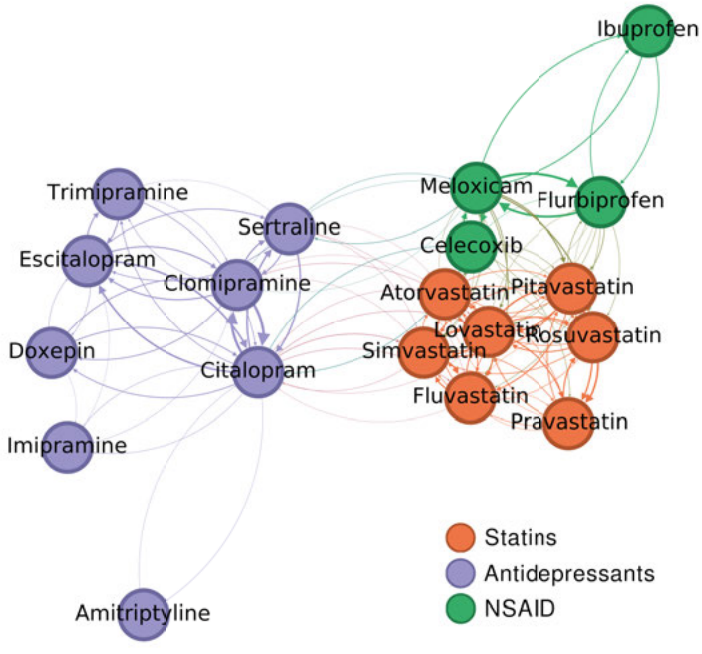

Figure 5.6. Clustering of drugs using spectral clustering on graph of drugs nodes with edge-weights calculated as in Figure 5.5

To analyse what genes are important in a cluster Equation 5.2 can be expanded to answer the question *"What is the probability of picking a specific gene in the neighbourhood of genes within gene sets, belonging to a specific cluster"*. This is formulated in equation 5.4 where C represents the set of all nodes within the cluster. It is assumed that the probability of selecting a node, T, within a cluster is uniform.

$$P(y|C) = \sum_{p} \sum_{g_n \in T} \sum_{T \in C} P(y|p) P(p|g_n) P(g_n|T) P(T) \tag{5.4}$$

However, these probabilities are not only influenced by the input genes of the gene sets but also by the picked genes position and connectivity in the PPI. A fitting metaphor is that the central station in a generic subway network has a higher probability of being included in the shortest path between two random stations than other less connected stations. So in order to mitigate this positional bias, random comparable sets for each cluster are simulated and a probability density function for picking each gene is estimated. Using the probability density function, the cumulative probability density value of the observed value can be obtained resulting in a metric where the degree bias is mitigated. Still, there is also a path-specific bias as a consequence of the degree bias. Following the subway metaphor, if we have a higher probability of the central station being among shortest routes, then the stops close to the central station will also have an increased probability, even if the stop itself does not have high numbers of connections. This is partly mitigated by estimating the probability density function, but when testing a cluster of terms with associated genes of relative proximity some of the paths through these hub nodes will be favoured. Even if the higher probability in the hub node is mitigated, its influence persists in some connected genes. The solution to this problem was to only select genes with a high cumulative density value that forms Large Connected Components (LCCs), such as the one shown in figure 5.7. The LCC selection ensures that all high-scoring nodes are connected to other high-scoring nodes, while excluding nodes under hub-influence.

To estimate the validity of the gene selection, enrichment tesing of the LCCs among the top 500 genes for the three drug clusters (figure 5.6) were performed. This resulted in expected top enriched pathways for the drug-network clusters. General drug metabolism, such as via cytochrome P450, was highly enriched for all three groups as expected. In addition, the test identified anticipated group-specific results such as Steroid Hormone synthesis[74] or Linoleic acid metabolism[75] for statins, Serotonergic synapse or Neuroactive ligand-receptor interaction for antidepressants, and Arachidonic acid metabolism[76] (COX and cytochrome P450) for NSAID clusters.

The clustering method was compared against the z-score transformation of the mean shortest distance[52] and the closest network seperation of drugs[77]. Both these measurement had to be adapted to provide a proximity value instead of a distance value as spectral clustering requires an affinity matrix. The method performed well when classifying drugs and ADRs reaching a sensitivity of 94 % and 86 % respectively, outcompeting the other metrics.

Unfortunately, enrichment testing for ADRs could not be performed while excluding input-genes. The HPO terms matched for ADRs report any gene associated with the phenotype (symptom), meaning that genes included are mainly associated with diseases that cause the symptom. Moreover, the number of associated genes tends to be numerous and diverse between HPO-terms. Consequently, the pool of unique genes used as input for candidate gene suggestion is large and most top genes suggested can be found in at least one

Large Connected Component among top Cumulative Probability values

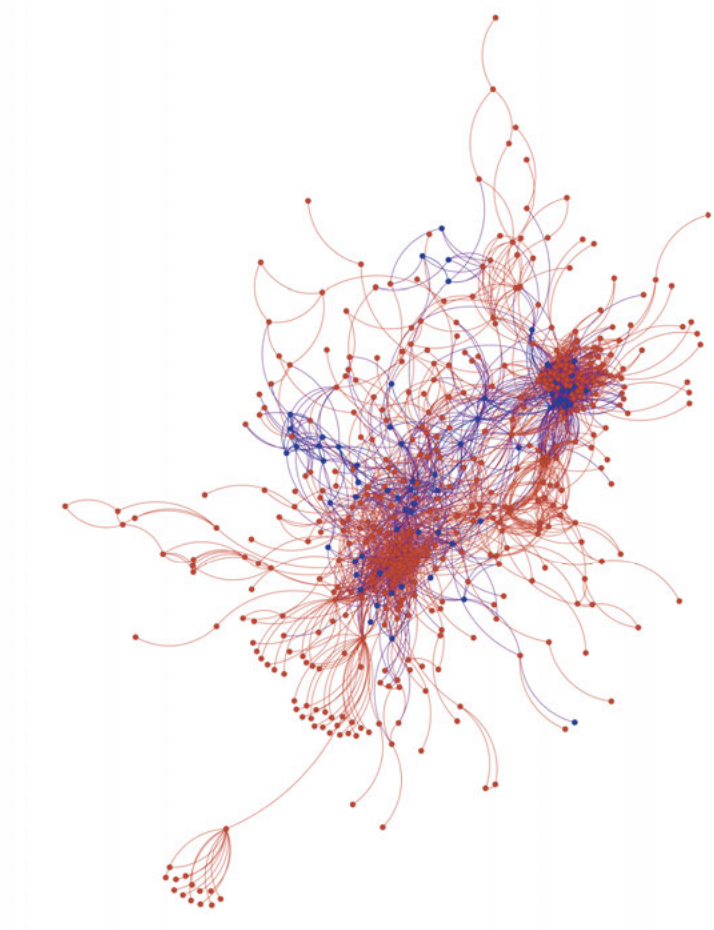

Figure 5.7. Large Connected Component (LCC) of top 700 cumulative probability values obtained from the antidepressant cluster, observed in figure 5.6. Blue nodes represent genes present in input drug-gene sets.

input ADR gene set. The method shows promise, but further development is possible and warrants experimental validation.

6. Conclusions

To revisit the milestones from aims:

- **Does ADR data be collected and curated the way the Swedegene Project does further the understanding of genetic factors underpinning ADRs?**:
  This question had been answered before this thesis and my PhD project. The Swedegene project has already produced studies from the material collected. However, the methodology it employs had not been published. My contribution has been modernising and securing the database and the interface from which new data is added.
- **Can previous findings of azathioprine-induced pancreatitis in patients with IBD be replicated in patients from Sweden?**:
  The association between HLA-DQA1*02:01-HLA-DRB1*07:01 and azathioprine-induced pancreatitis was confirmed in Swedish patients with Crohn's disease, but not in ulcerative colitis. Under assumptions of HLA-DQA1*02:01-HLA-DRB1*07:01 frequency in the population and the probability of developing azathioprine-induced pancreatitis, a conservative risk was estimated to be 7.3 % in carriers while being 2.2 % in non-carriers. Interestingly, the one-sidedness of the genetic association with IBD (only patients with Crohn's disease) observed is contrary to previous studies[66].
- **Can novel genetic associations be found for ADRs and categories of ADRs using whole genome sequencing?**:
  In the largest cohort of atypical femur fracture (AFF) during bisphosphonate use to date, the Swedegene 53 WGS AFF samples were tested together with 185 exon sequencing samples from The Netherlands and Australia. In this study, 203 genes previously found associated or theorised to be associated with AFF were tested for gene-association using SKAT-O, selecting for rare deleterious variants. None of these tests were significant, nor were any tests significant in the secondary exploratory analysis. However, suggestive evidence was found for four nominally significant genes that could be interesting as candidate genes in future studies. However, my conclusion of these results it that there is no strong simple genetic predisposition to AFF during bisphosphonate use.
  Patients from the Swedegene WGS cohort who had CNS-toxic reactions to anti-microbial drugs were tested for association with genetic variation in transporter proteins in the blood-brain barrier. No correlation between CNS-toxicity and these genes was significant or nominally significant. In the exploratory analysis, gene-association tests

were significant for the genes LCP1(q = 0.013), RETSAT(q=0.013), and SFMBT2(q=0.035). These genes can relate to CNS-related symptoms but are also related to psychiatric and neurological diseases such as schizophrenia, suicide, Parkinson's disease, and dementia. Due to the heterogeneity of exposure and response among the 66 cases, the interpretation of these results is challenging.

To answer the question: Yes, genetic associations of ADRs and ADR categories can be identified in WGS data. However, the increased number of possible hypotheses requires clear biological reasoning when selecting which hypotheses to test.

- **Can shared protein-protein interaction be used to formulate genetic hypotheses of genetic predispositions for drugs relating to ADRs?**:
  Yes, the method presented classifies drugs and ADRs with high sensitivity and outperforms compared methods on test data. Using enrichment analysis of KEGG pathways on the suggested candidate genes, during evaluation of drug clusters, reveals that the method identifies genes relevant to biological processes shared among the drugs. The method needs experimental validation but results from testing data suggest that it could be used to formulate genetic hypotheses by selecting drugs or ADRs to be included and what genes to test.

I believe the overarching aim of exploring genetic predispositions of ADRs has been reached as my PhD project has resulted in providing nuance and new findings while previous associations have been rejected. I have also contributed with a methodology for how hypothesis formulation of case-control genetic st§udies of ADRs can be performed.

7. Discussion

The study of pharmacogenomics is closely tied to the research into personalised medicine[78], where the core belief is that treatment should adapt to the unique amalgamation of molecular, genetic, physiological, environmental, and behavioural factors of the patient. Today, there are already many drug treatments that can, and are, personalised for genetic factors. Some examples are the genotyping of variants in TPMT, NUDT15 for the cytotoxic agent azathioprine[42], DPYD for the cytotoxic agent 5-fluorouracil[2], CYP2D6 for a multitude of drugs[79], different genetic variants causing ADRs from antiviral treatment of HIV[80] and many more. Today, there are many sources keeping track of genetic factors predisposing ADRs while providing clinical recommendations. Two important such sources are CPIC (Clinical Pharmacogenetics Implementation Consortium)[81] and DPWG (Dutch Pharmacogenetics Working Group)[82]. But before any clinical recommendations can be made, the predispositioning genetic factors must first be detected and proven causal or predictive.

It is important to note, when testing genetic correlation between cases and controls, that any significant results are correlated and correlation does not equal causality. This is evident in papers III and IV. In paper III, the three genes correlated were LCP1, RETSAT, and SFMBT2. Where L-Plastin (expressed by LCP1) is related to the transport of T-cell activation modules to the cell surface, and has been shown to be differentially expressed in suicide victims[70], compared to those with an accidental death. Additionally, a rare missense variant in LCP1 has been associated with schizophrenia[71]. RETSAT encodes retinol saturase which is involved in the metabolism of vitamin A, but the biological function of the enzyme product is unknown[72]. And, there is evidence of dysregulation of vitamin A metabolism in the aetiology of schizophrenia[73]. As evident, these genes have previously been reported as associated with psychiatric and neurological disorders. Some of the CNS-toxic reactions included are symptoms, such as hallucinations, of such disorders but none of the cases in the study had such a diagnosis. Nevertheless, it poses an alternative explanation for the correlation. Still, the associations found are still valuable as they hone the hypothesis of future studies and enrich the understanding of the aetiology of the ADR, not only by what is correlated but also by what is not. Paper II is the result of such studies as paper IV, where the hypothesis was honed down to the testing of a single genetic variable by previous studies. The result of that paper does not provide causality but it does provide additional evidence for the predictive capabilities of the genetic variation. Interestingly, Paper III provides information of both previous

studies and sugeests new genetic targets. The the candidate genes consisted of previous findings[83, 84, 85, 86, 87] and genes theorised to be involved with the phenotype. While none of these previously reported associations could be replicated, ten genes were suggested as candidate genes for future studies. An important thing to note is that this is the study with the largest sample size of its kind, implying that the effect-sizes of previously implicated genetic variation cannot be very large. In turn, this means that the variance explained — the number of cases that can be explained by the genetic variants — cannot be very large. However, how would genetic targets to study be selected when there is limited information from previous studies relating to the ADRs?

When investigating genetic associations with ADRs it is crucial to have a clear hypothesis, since is the hypothesis that will prioritise what to test. Ideally, the hypothesis should frame the ADR from a biological perspective. In the most direct of such cases, it has been shown that when drugs interact with genes related to Mendelian diseases there is a high risk of causing an ADR similar to the symptoms of the disease[88]. Similarly, there is evidence that drugs that target genes associated with symptoms have higher odds of resulting in an ADR similar to the symptom[89]. This means that the drug affects underlying biological processes and the alteration gives rise to the reaction. The scope and specificity of the aforementioned hypothesis will influence which processes are implicated as affected. In paper II the hypothesis was that a specific configuration of genetic variations (a haplotype of HLA-types) was associated with a specific ADR (azathioprine-induced pancreatitis) while the recipients of the drug all had the same disease (IBD). If previous findings had not associated the specific HLA-types, the hypothesis might have been broadened to suspect any HLA-type. If this were the case, the multiple hypothesis correction would have rendered the study underpowered as, using Bonferroni correction, the observed p-value of $p = 0.0275$ falls above the significance threshold after the correction of two hypotheses. This highlights the balancing act between the number of hypotheses, sample size, and the discoverable magnitude of correlations. Effectively, as the number of hypotheses increases, the p-value significance threshold decreases. This implies that significant test statistics need either an increased mean or a decreased variation to remain significant after correction. Therefore, testing multiple hypotheses with a limited sample size will only yield significant test results on the most extreme differences. These heightened demands on test statistics are also reflected in the selection criteria of the cases in the study. If cases are selected in a manner that introduces competing correlations of the response, (e.g. different correlating variants in two different genetic loci) the increased variation and decrease of the mean will jeopardise the significance of both associations.

As mentioned earlier, severe ADR cases are rare, and reported cases are even rarer. This means that genetic studies of ADRs that have multiple hypotheses or hypothesis-generating exploratory studies are faced with underpower issues[90]. In order to get samples, Swedegene has collected cases of

different ADRs nationwide since 2008, but acquiring a sufficient number of samples is challenging. In such scenarios, there are two alternatives: either increase the sample size through collaboration, as in paper III, or broaden the inclusion criteria for phenotypes, as in paper IV. Each of these approaches carries their separate challenges and advantages. Pooling of samples internationally is a sure way to increase sample size while keeping narrow hypotheses but could introduce batch-effects (such as between Swedegene and SweGen) due to differences in data acquisition, population stratification, and risks confounding results due to centre-specific inclusion criteria. On the other hand, increasing the sample size by broadening the case inclusion criteria risks discordant subsets of cases if the drugs or ADRs included do not share sufficient attributes — chemical similarity, implicated biological pathways, etc.— relating to the hypotheses. Additionally, the broadening of the inclusion criteria would require reformulation of the research question, which in turn can expand the scope and therefore lead to additional genetic hypotheses.

In some cases, underlying biological processes can be implicated by the p-values obtained in a GWAS even if no individual tests are significant[91]. These processes or patterns can be observed via the detection of non-random distributions of p-values from variants tested located in genes. Where the analysis of these genes can detect non-random interconnected clusters in a PPI network. When designing studies, a similar approach can be applied in order to create or validate the selection of drugs and ADRs by analysing what genes they are previously reported associated with. Combining public data on what proteins the drugs interact with and what proteins are associated with the ADRs, models for abstracting similarity between drugs or ADRs in a PPI network can be constructed. Utilising these similarities, the validity of including or excluding a drug or an ADRs in a study can be estimated. This type of rationale has been applied to estimate *in silico* drug efficacy screening[52], suggest drug repurposing candidates[53], and predict synergistic drug combinations[77]. In these methods, the drug targets' position in a PPI network is compared to interconnected clusters of disease-associated genes (a.k.a. disease modules). However, when studying ADRs though the same approach would not produce the same results. The two main differences are: there are considerably fewer gene associations to ADRs generally than those associated with disease, so the associations needed to construct meaningful ADR-clusters are limited. The second reason is that the ADR can be the result form different categories of interactions with the drug, such as off-target proteins[88], on-target proteins[89], metabolising enzymes[89] or other proteins. If several drugs are considered in a ADR-related study the causality in the drug-ADR relationship might be maintained via different classes of interactions (e.g. the off-target for one drug is the on-target for others). Paper V presents a methodology that is designed with this in mind as the similarities estimated between drugs are based on the shared shortest chains of interactions between any of the proteins associated with the drugs in a PPI network.

Moreover, this similarity measurement is extended to include multiple drugs allowing ranking of the interaction paths of all included drugs. Subsequently, the highest ranking of these interaction chains can be investigated for what biological processes they are a part of, which can be used as the basis of a working hypothesis.

8. Acknowledgements

The path of a PhD project is seldom straight. The fact is that the aims and the opportunities available can often change and not always without friction. However, if the outcome already is known it cant be called research and without that uncertainty there is no room for creativity and curiosity. Nonetheless, such a journey is not without strife and there are many individuals I have to thank for allowing me to complete it. I'd like to extend my gratitude to:

My main supervisor **Mia Wadelius** and co-supervisor **Pär Hallberg** for providing the possibility for this thesis in the first place. It is through your diligent work and initiative that the Swedegene project exists and without it, none of these studies could exist.

My co-supervisor **Niclas Eriksson** for your support and feedback on implementation, analysis, and holistic reviewing of both papers and this thesis.

My former co-supervisor **Adam Ameur** for the feedback and plans made for incorporating the SweGen data into analysis. It was unfortunate that those plans fell through, but I am looking forward to other projects in the future.

My colleges working with the Swedegene project: **Eva Prado**, **Sara Ask**, **Ulrica Ramqvist**, **Hugo Kohnke**, and **Soffia Attelind**.

My colleague **Marco Cavalli** for teaching the ropes of regulatory element analysis.

The long-term support from NBIS, **Nina Norgren** and **Anna Johansson** for your input, expertise, and your practical and theoretical support during the development of the WGS pipeline.

My dear cousin **Sofia Forslund** for methodology feedback, reviewing papers and this thesis, but most of all introducing me to the wonderful combination of hard science and fiction ever since you made me the booklet "Mathematics of the Orcs" when I was a child. Our discussions have been invaluable to me.

My friends **Björn Forsberg**, **Stephanie Herman**, **Åland Sharyari**, and **Olle Nordesjö** for providing feedback on papers, the half-time summary and/or this thesis. Conversations with you are always meaningful and provide perspectives on science, analysis, and life.

My colleges present and former with whom I've shared room, lunch and interesting discussions with at Akademiska sjukhuset: **Asma Al-Grety**, **Christina Zjukovskaja**, **Shibu Krishnan**, **Sandy Abujrais**, **Henrik Carlsson**, **Payam Emami Khoonsari**, **Yassine Noui**, **Akshai Parakkal Sreenivasan**, **Jenny Jakobsson**, **Aina Vaivade**, **Ida Erngren**, and **Kim Kultima**.

My other colleges at entrance 61, floors three and four.

The bioinformatics group of clinical genomics: **Claes Ladenvall**, **Patrik Smeds**, **Malin Melin**, **Arielle Munters**, **Jonas Almlöf** and **John Morrsion** for the input and feedback during the GMS pharmacogenomics project.

My opponent **Volker Lauschke** and committee members **Jessica Nordlund**, **Mika Gustafsson**, and **Arne Reimers**.

My parents **Anna** and **Stefan Ås** for providing me with a wonderful life, curiosity, meaning and reviewing my thesis.

My sisters **Runa** and **Tove Ås** for always being able to provide me perspective and security.

My partner **Ramona Reldin** for the laughter, your companionship and love.

My **friends and relatives**, too numerous to name but you know who you are. Thank you for providing my life with joy and wonder.

And finally, even if you can not read, my cat **Felix** for reminding me of work-life balance by standing on the keyboard and screaming in my face ever so often.

Thank you all.

References

[1] Munir Pirmohamed, Alasdair M Breckenridge, Neil R Kitteringham, and B Kevin Park. Adverse drug reactions. 316(7140):1295–1298. ISSN 0959-8138. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1113033/.

[2] Steven M. Offer, Natalie J. Wegner, Croix Fossum, Kangsheng Wang, and Robert B. Diasio. Phenotypic profiling of DPYD variations relevant to 5-fluorouracil sensitivity using real-time cellular analysis and in vitro measurement of enzyme activity. 73(6):1958–1968. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-12-3858. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3602211/.

[3] N. Feltelius, I. Persson, J. Ahlqvist-Rastad, M. Andersson, L. Arnheim-Dahlström, P. Bergman, F. Granath, C. Adori, T. Hökfelt, S. Kühlmann-Berenzon, P. Liljeström, M. Maeurer, T. Olsson, Å Örtqvist, M. Partinen, T. Salmonson, and B. Zethelius. A coordinated cross-disciplinary research initiative to address an increased incidence of narcolepsy following the 2009-2010 pandemrix vaccination programme in sweden. 278(4):335–353. ISSN 1365-2796. doi: 10.1111/joim.12391.

[4] F. Yang, J. Xuan, J. Chen, H. Zhong, H. Luo, P. Zhou, X. Sun, L. He, S. Chen, Z. Cao, X. Luo, and Q. Xing. HLA-b*59:01: a marker for stevens–johnson syndrome/toxic epidermal necrolysis caused by methazolamide in han chinese. 16(1):83–87. ISSN 1473-1150. doi: 10.1038/tpj.2015.25. URL https://www.nature.com/articles/tpj201525. Number: 1 Publisher: Nature Publishing Group.

[5] Rostam Osanlou, Lauren Walker, Dyfrig A. Hughes, Girvan Burnside, and Munir Pirmohamed. Adverse drug reactions, multimorbidity and polypharmacy: a prospective analysis of 1 month of medical admissions. *BMJ open*, 12(7): e055551, July 2022. ISSN 2044-6055. doi: 10.1136/bmjopen-2021-055551.

[6] Ingegerd Odar-Cederlöf, Per Oskarsson, Gunnar Ohlén, Yohannes Tesfa, Annica Bergendal, Anders Helldén, and Ulf Bergman. [adverse drug effect as cause of hospital admission. common drugs are the major part according to the cross-sectional study]. 105(12):890–893. ISSN 0023-7205.

[7] Hanna Gyllensten, Katja M. Hakkarainen, Staffan Hägg, Anders Carlsten, Max Petzold, Clas Rehnberg, and Anna K. Jönsson. Economic impact of adverse drug events – a retrospective population-based cohort study of 4970 adults. 9 (3):e92061. ISSN 1932-6203. doi: 10.1371/journal.pone.0092061. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0092061. Publisher: Public Library of Science.

[8] Karin Wester, Anna K. Jönsson, Olav Spigset, Henrik Druid, and Staffan Hägg. Incidence of fatal adverse drug reactions: a population based study. *British Journal of Clinical Pharmacology*, 65(4):573–579, April 2008. ISSN 1365-2125. doi: 10.1111/j.1365-2125.2007.03064.x.

[9] J. Lazarou, B. H. Pomeranz, and P. N. Corey. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*, 279(15):1200–1205, April 1998. ISSN 0098-7484. doi: 10.1001/jama.279.15.1200.

[10] Deborah Dowell. CDC clinical practice guideline for prescribing opioids for pain — united states, 2022. 71. ISSN 1057-59871545-8601. doi: 10.15585/mmwr.rr7103a1. URL https://www.cdc.gov/mmwr/volumes/71/rr/rr7103a1.htm.

[11] Qing Zhao, Yao Chen, Weihua Huang, Honghao Zhou, and Wei Zhang. Drug-microbiota interactions: an emerging priority for precision medicine. *Signal Transduction and Targeted Therapy*, 8(1):1–27, October 2023. ISSN 2059-3635. doi: 10.1038/s41392-023-01619-w. URL https://www.nature.com/articles/s41392-023-01619-w. Publisher: Nature Publishing Group.

[12] Werner J. Pichler and Marie-Charlotte Brüggen. Viral infections and drug hypersensitivity. *Allergy*, 78(1):60–70, 2023. ISSN 1398-9995. doi: 10.1111/all.15558. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/all.15558. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/all.15558.

[13] Ylva Böttiger, Kari Laine, Marine L. Andersson, Tuomas Korhonen, Björn Molin, Marie-Louise Ovesjö, Tuire Tirkkonen, Anders Rane, Lars L. Gustafsson, and Birgit Eiermann. SFINX-a drug-drug interaction database designed for clinical decision support systems. 65(6):627–633. ISSN 1432-1041. doi: 10.1007/s00228-008-0612-5.

[14] L. Henderson, Q. Y. Yue, C. Bergquist, B. Gerden, and P. Arlett. St john's wort (hypericum perforatum): drug interactions and clinical outcomes. 54(4): 349–356. ISSN 0306-5251. doi: 10.1046/j.1365-2125.2002.01683.x.

[15] David G. Bailey, George Dresser, and J. Malcolm O. Arnold. Grapefruit–medication interactions: Forbidden fruit or avoidable consequences? 185(4):309–316. ISSN 0820-3946. doi: 10.1503/cmaj.120951. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3589309/.

[16] Adam Ameur, Johan Dahlberg, Pall Olason, Francesco Vezzi, Robert Karlsson, Marcel Martin, Johan Viklund, Andreas Kusalananda Kähäri, Pär Lundin, Huiwen Che, Jessada Thutkawkorapin, Jesper Eisfeldt, Samuel Lampa, Mats Dahlberg, Jonas Hagberg, Niclas Jareborg, Ulrika Liljedahl, Inger Jonasson, Åsa Johansson, Lars Feuk, Joakim Lundeberg, Ann-Christine Syvänen, Sverker Lundin, Daniel Nilsson, Björn Nystedt, Patrik KE Magnusson, and Ulf Gyllensten. SweGen: a whole-genome data resource of genetic variability in a cross-section of the swedish population. 25(11):1253–1260. ISSN 1476-5438. doi: 10.1038/ejhg.2017.130. URL https://www.nature.com/articles/ejhg2017130. Number: 11 Publisher: Nature Publishing Group.

[17] David G. Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, Jessica Spencer, Leonid Kruglyak, Lincoln Stein, Linda Hsie, Thodoros Topaloglou, Earl Hubbell, Elizabeth Robinson, Michael Mittmann, Macdonald S. Morris, Naiping Shen, Dan Kilburn, John Rioux, Chad Nusbaum, Steve Rozen,

Thomas J. Hudson, Robert Lipshutz, Mark Chee, and Eric S. Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. 280(5366):1077–1082, . doi: 10.1126/science.280.5366.1077. URL https://www.science.org/doi/10.1126/science.280.5366.1077. Publisher: American Association for the Advancement of Science.

[18] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. 10:387–406. ISSN 1527-8204. doi: 10.1146/annurev.genom.9.081307.164242. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2925172/.

[19] Rachel Marine, Shawn W. Polson, Jacques Ravel, Graham Hatfull, Daniel Russell, Matthew Sullivan, Fraz Syed, Michael Dumas, and K. Eric Wommack. Evaluation of a Transposase Protocol for Rapid Generation of Shotgun High-Throughput Sequencing Libraries from Nanogram Quantities of DNA. *Applied and Environmental Microbiology*, 77(22):8071–8079, November 2011. ISSN 0099-2240. doi: 10.1128/AEM.05610-11. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3209006/.

[20] Margaret M. DeAngelis, David G. Wang, and Trevor L. Hawkins. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Research*, 23(22):4742–4743, November 1995. ISSN 0305-1048. doi: 10.1093/nar/23.22.4742. URL https://doi.org/10.1093/nar/23.22.4742.

[21] Deanna M. Church, Valerie A. Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, Richa Agarwala, William M. McLaren, Graham R. S. Ritchie, Derek Albracht, Milinn Kremitzki, Susan Rock, Holland Kotkiewicz, Colin Kremitzki, Aye Wollam, Lee Trani, Lucinda Fulton, Robert Fulton, Lucy Matthews, Siobhan Whitehead, Will Chow, James Torrance, Matthew Dunn, Glenn Harden, Glen Threadgold, Jonathan Wood, Joanna Collins, Paul Heath, Guy Griffiths, Sarah Pelan, Darren Grafham, Evan E. Eichler, George Weinstock, Elaine R. Mardis, Richard K. Wilson, Kerstin Howe, Paul Flicek, and Tim Hubbard. Modernizing Reference Genome Assemblies. *PLOS Biology*, 9(7):e1001091, July 2011. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001091. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001091. Publisher: Public Library of Science.

[22] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. 25(14):1754–1760. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324. URL https://doi.org/10.1093/bioinformatics/btp324.

[23] Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples, . URL https://www.biorxiv.org/content/10.1101/201178v3. Pages: 201178 Section: New Results.

[24] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T. Afshar, Sam S. Gross, Lizzie Dorfman, Cory Y. McLean, and Mark A. DePristo. A universal SNP and small-indel variant caller using deep neural networks. 36(10):983–987, . ISSN 1546-1696. doi: 10.1038/nbt.4235. URL https://www.nature.com/articles/nbt.4235. Number: 10 Publisher: Nature Publishing Group.

[25] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. 29(1): 308–311. ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC29783/.

[26] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Jun Wang, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Paul Flicek, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R. Mardis, Richard K. Wilson, Lucinda Fulton, Robert Fulton, Stephen T. Sherry, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Adam Auton, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Richard A. Gibbs, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Donna Muzny, Aniko Sabo, Zhuoyi Huang, Jun Wang, Lachlan J. M. Coin, Lin Fang, Xiaosen Guo, Xin Jin,

Guoqing Li, Qibin Li, Yingrui Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, David M. Altshuler, Eric Banks, Gaurav Bhatia, Guillermo del Angel, Stacey B. Gabriel, Giulio Genovese, Namrata Gupta, Heng Li, Seva Kashin, Eric S. Lander, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Andrew G. Clark, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Jan O. Korbel, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Paul Flicek, Kathryn Beal, Laura Clarke, Avik Datta, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Richard E. Smith, Daniel Zerbino, Xiangqun Zheng-Bradley, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, David R. Bentley, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Sudbrak, Vyacheslav S. Amstislavskiy, Ralf Herwig, Elaine R. Mardis, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Bioinformatics Institute European Molecular Biology Laboratory, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, and NIH National Eye Institute. A global reference for human genetic variation. 526(7571):68–74. ISSN 1476-4687. doi: 10.1038/nature15393. URL `https://www.nature.com/articles/nature15393`. Number: 7571 Publisher: Nature Publishing Group.

[27] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. 6(2):80–92. ISSN 1933-6934. doi: 10.4161/fly.19695. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3679285/`.

[28] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. 17(1):122. ISSN 1474-760X. doi: 10.1186/s13059-016-0974-4. URL `https://doi.org/10.1186/s13059-016-0974-4`.

[29] Ricard Solé. Revisiting leigh van valen's "a new evolutionary law" (1973). 17 (2):120–125. ISSN 1555-5550. doi: 10.1007/s13752-021-00391-w. URL https://doi.org/10.1007/s13752-021-00391-w.

[30] Carolyn Katovich Hurley. Naming HLA diversity: A review of HLA nomenclature. 82(7):457–465. ISSN 0198-8859. doi: 10.1016/j.humimm.2020.03.005. URL https://www.sciencedirect.com/science/article/pii/S0198885920300653.

[31] Xiaoming Jia, Buhm Han, Suna Onengut-Gumuscu, Wei-Min Chen, Patrick J. Concannon, Stephen S. Rich, Soumya Raychaudhuri, and Paul I. W. de Bakker. Imputing amino acid polymorphisms in human leukocyte antigens. 8(6): e64683. ISSN 1932-6203. doi: 10.1371/journal.pone.0064683. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064683. Publisher: Public Library of Science.

[32] Oliver Mayo. A century of hardy–weinberg equilibrium. 11(3):249–256. ISSN 1839-2628, 1832-4274. doi: 10.1375/twin.11.3.249. URL https://www.cambridge.org/core/journals/twin-research-and-human-genetics/article/century-of-hardyweinberg-equilibrium/7290655079251A83AF3D89D176DA3AFF. Publisher: Cambridge University Press.

[33] Stephen Turner, Loren L. Armstrong, Yuki Bradford, Christopher S. Carlson, Dana C. Crawford, Andrew T. Crenshaw, Mariza de Andrade, Kimberly F. Doheny, Jonathan L. Haines, Geoffrey Hayes, Gail Jarvik, Lan Jiang, Iftikhar J. Kullo, Rongling Li, Hua Ling, Teri A. Manolio, Martha Matsumoto, Catherine A. McCarty, Andrew N. McDavid, Daniel B. Mirel, Justin E. Paschall, Elizabeth W. Pugh, Luke V. Rasmussen, Russell A. Wilke, Rebecca L. Zuvich, and Marylyn D. Ritchie. Quality control procedures for genome-wide association studies. 68(1):1.19.1–1.19.18. ISSN 1934-8258. doi: 10.1002/0471142905.hg0119s68. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0471142905.hg0119s68. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471142905.hg0119s68.

[34] Jacklyn Hellwege, Jacob Keaton, Ayush Giri, Xiaoyi Gao, Digna R. Velez Edwards, and Todd L. Edwards. Population stratification in genetic association studies. 95:1.22.1–1.22.23. ISSN 1934-8266. doi: 10.1002/cphg.48. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6007879/.

[35] Nick Patterson, Alkes L. Price, and David Reich. Population structure and eigenanalysis. 2(12):e190. ISSN 1553-7404. doi: 10.1371/journal.pgen.0020190. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020190. Publisher: Public Library of Science.

[36] Winston Haynes. Bonferroni correction. In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, *Encyclopedia of Systems Biology*, pages 154–154. Springer. ISBN 978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_1213. URL https://doi.org/10.1007/978-1-4419-9863-7_1213.

[37] Frank Dudbridge and Arief Gusnanto. Estimation of significance thresholds for genomewide association scans. 32(3):227–234. ISSN 0741-0395. doi:

10.1002/gepi.20297. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2573032/.

[38] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. 100(16):9440–9445. doi: 10.1073/pnas.1530509100. URL https://www.pnas.org/doi/10.1073/pnas.1530509100. Publisher: Proceedings of the National Academy of Sciences.

[39] Seunggeun Lee, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, David C. Christiani, Mark M. Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. 91(2):224–237. ISSN 0002-9297. doi: 10.1016/j.ajhg.2012.06.007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3415556/.

[40] Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. 89(1):82–93, . ISSN 0002-9297. doi: 10.1016/j.ajhg.2011.05.029. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3135811/.

[41] B Jassal. Azathioprine ADME. URL http://reactome.org/content/detail/R-HSA-9748787.1. Institution: Reactome - a curated knowledgebase of biological pathways.

[42] Laura Dean. Azathioprine Therapy and TPMT and NUDT15 Genotype. In Victoria M. Pratt, Stuart A. Scott, Munir Pirmohamed, Bernard Esquivel, Brandi L. Kattman, and Adriana J. Malheiro, editors, *Medical Genetics Summaries*. National Center for Biotechnology Information (US), Bethesda (MD), 2012. URL http://www.ncbi.nlm.nih.gov/books/NBK100661/.

[43] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. 28(1):27–30. ISSN 0305-1048. doi: 10.1093/nar/28.1.27.

[44] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. 40:D940–D946. ISSN 0305-1048. doi: 10.1093/nar/gkr972. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245088/.

[45] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556. URL https://www.nature.com/articles/ng0500_25. Publisher: Nature Publishing Group.

[46] The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina

Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1): iyad031, May 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031. URL `https://doi.org/10.1093/genetics/iyad031`.

[47] Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. 2(3):100141, . ISSN 2666-6758. doi: 10.1016/j.xinn.2021.100141.

[48] Javier De Las Rivas and Celia Fontanillo. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLOS Computational Biology*, 6(6):e1000807, June 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000807. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000807`. Publisher: Public Library of Science.

[49] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. 104(21):8685–8690.

ISSN 0027-8424. doi: 10.1073/pnas.0701361104. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1885563/.

[50] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLOS Computational Biology*, 11(4):e1004120, April 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004120. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004120. Publisher: Public Library of Science.

[51] Hendrik A de Weerd, Tejaswi V S Badam, David Martínez-Enguita, Julia Åkesson, Daniel Muthas, Mika Gustafsson, and Zelmina Lubovac-Pilav. MODifieR: an ensemble r package for inference of disease modules from transcriptomics networks. 36(12):3918–3919. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa235. URL https://doi.org/10.1093/bioinformatics/btaa235.

[52] Emre Guney, Jörg Menche, Marc Vidal, and Albert-László Barábasi. Network-based in silico drug efficacy screening. *Nature Communications*, 7(1): 10331, February 2016. ISSN 2041-1723. doi: 10.1038/ncomms10331. URL https://www.nature.com/articles/ncomms10331. Publisher: Nature Publishing Group.

[53] Deisy Morselli Gysi, Ítalo do Valle, Marinka Zitnik, Asher Ameli, Xiao Gan, Onur Varol, Susan Dina Ghiassian, J. J. Patten, Robert A. Davey, Joseph Loscalzo, and Albert-László Barabási. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences*, 118(19):e2025581118, May 2021. doi: 10.1073/pnas.2025581118. URL https://www.pnas.org/doi/10.1073/pnas.2025581118. Publisher: Proceedings of the National Academy of Sciences.

[54] David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. 46:D1074–D1082. ISSN 1362-4962. doi: 10.1093/nar/gkx1037.

[55] Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M. Brower, Tiffany J. Callahan, Christopher G. Chute, Johanna L. Est, Peter D. Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L. Harris, Michael J. Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A. McMurry, Jillian A. Miller, Monica C. Munoz-Torres, Rebecca L. Peters, Christina K. Rapp, Ana M. Rath, Shahmir A. Rind, Avi Z. Rosenberg, Michael M. Segal, Markus G. Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A. Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J. Mungall, Melissa A. Haendel, and Peter N. Robinson. The human phenotype ontology in 2021. 49:D1207–D1217. ISSN 1362-4962.

doi: 10.1093/nar/gkaa1043.

[56] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. 47:D607–D613. ISSN 0305-1048. doi: 10.1093/nar/gky1131. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323986/.

[57] G. Traver Hart, Arun K. Ramani, and Edward M. Marcotte. How complete are current yeast and human protein-interaction networks? 7(11):120. ISSN 1474-760X. doi: 10.1186/gb-2006-7-11-120. URL https://doi.org/10.1186/gb-2006-7-11-120.

[58] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M. J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M. D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung-Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R.

Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W. Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie vandeVondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, November 2008. ISSN 1476-4687. doi: 10.1038/nature07517. URL `https://www.nature.com/articles/nature07517`. Publisher: Nature Publishing Group.

[59] Genomics in the cloud [book]. URL `https://www.oreilly.com/library/view/genomics-in-the/9781491975183/`. ISBN: 9781491975190.

[60] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. 26(22):2867–2873. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq559. URL `https://doi.org/10.1093/bioinformatics/btq559`.

[61] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. 4(1):s13742–015–0047–8. ISSN 2047-217X. doi: 10.1186/s13742-015-0047-8. URL `https://doi.org/10.1186/s13742-015-0047-8`.

[62] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. 38(16): e164, . ISSN 0305-1048. doi: 10.1093/nar/gkq603. URL `https://doi.org/10.1093/nar/gkq603`.

[63] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. 32(8):1220–1222. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv710.

[64] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with snakemake. URL `https://f1000research.com/articles/10-33`. Type: article.

[65] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R. Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, Yang Luo, Carlo Sidore, Alan Kwong, Nicholas Timpson, Seppo Koskinen, Scott Vrieze, Laura J. Scott, He Zhang, Anubha

Mahajan, Jan Veldink, Ulrike Peters, Carlos Pato, Cornelia M. van Duijn, Christopher E. Gillies, Ilaria Gandin, Massimo Mezzavilla, Arthur Gilly, Massimiliano Cocca, Michela Traglia, Andrea Angius, Jeffrey C. Barrett, Dorrett Boomsma, Kari Branham, Gerome Breen, Chad M. Brummett, Fabio Busonero, Harry Campbell, Andrew Chan, Sai Chen, Emily Chew, Francis S. Collins, Laura J. Corbin, George Davey Smith, George Dedoussis, Marcus Dorr, Aliki-Eleni Farmaki, Luigi Ferrucci, Lukas Forer, Ross M. Fraser, Stacey Gabriel, Shawn Levy, Leif Groop, Tabitha Harrison, Andrew Hattersley, Oddgeir L. Holmen, Kristian Hveem, Matthias Kretzler, James C. Lee, Matt McGue, Thomas Meitinger, David Melzer, Josine L. Min, Karen L. Mohlke, John B. Vincent, Matthias Nauck, Deborah Nickerson, Aarno Palotie, Michele Pato, Nicola Pirastu, Melvin McInnis, J. Brent Richards, Cinzia Sala, Veikko Salomaa, David Schlessinger, Sebastian Schoenherr, P. Eline Slagboom, Kerrin Small, Timothy Spector, Dwight Stambolian, Marcus Tuke, Jaakko Tuomilehto, Leonard H. Van den Berg, Wouter Van Rheenen, Uwe Volker, Cisca Wijmenga, Daniela Toniolo, Eleftheria Zeggini, Paolo Gasparini, Matthew G. Sampson, James F. Wilson, Timothy Frayling, Paul I. W. de Bakker, Morris A. Swertz, Steven McCarroll, Charles Kooperberg, Annelot Dekker, David Altshuler, Cristen Willer, William Iacono, Samuli Ripatti, Nicole Soranzo, Klaudia Walter, Anand Swaroop, Francesco Cucca, Carl A. Anderson, Richard M. Myers, Michael Boehnke, Mark I. McCarthy, Richard Durbin, and Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. 48(10):1279–1283. ISSN 1546-1718. doi: 10.1038/ng.3643.

[66] Graham A. Heap, Michael N. Weedon, Claire M. Bewshea, Abhey Singh, Mian Chen, Jack B. Satchwell, Julian P. Vivian, Kenji So, Patrick C. Dubois, Jane M. Andrews, Vito Annese, Peter Bampton, Martin Barnardo, Sally Bell, Andy Cole, Susan J. Connor, Tom Creed, Fraser R. Cummings, Mauro D'Amato, Tawfique K. Daneshmend, Richard N. Fedorak, Timothy H. Florin, Daniel R. Gaya, Emma Greig, Jonas Halfvarson, Alisa Hart, Peter M. Irving, Gareth Jones, Amir Karban, Ian C. Lawrance, James C. Lee, Charlie Lees, Raffi Lev-Tzion, James O. Lindsay, John Mansfield, Joel Mawdsley, Zia Mazhar, Miles Parkes, Kirstie Parnell, Timothy R. Orchard, Graham Radford-Smith, Richard K. Russell, David Reffitt, Jack Satsangi, Mark S. Silverberg, Giacomo C. Sturniolo, Mark Tremelling, Epameinondas V. Tsianos, David A. van Heel, Alissa Walsh, Gill Watermeyer, Rinse K. Weersma, Sebastian Zeissig, Jamie Rossjohn, Arthur L. Holden, and Tariq Ahmad. HLA-DQA1–HLA-DRB1 variants confer susceptibility to pancreatitis induced by thiopurine immunosuppressants. 46(10):1131–1134. ISSN 1546-1718. doi: 10.1038/ng.3093. URL `https://www.nature.com/articles/ng.3093`. Number: 10 Publisher: Nature Publishing Group.

[67] Max Schubach, Thorben Maass, Lusiné Nazaretyan, Sebastian Röner, and Martin Kircher. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. 52:D1143–D1154. ISSN 1362-4962. doi: 10.1093/nar/gkad989.

[68] Ethan G. Geier, Eugene C. Chen, Amy Webb, Audrey C. Papp, Sook Wah Yee, Wolfgang Sadee, and Kathleen M. Giacomini. Profiling Solute Carrier Transporters in the Human Blood-Brain Barrier. *Clinical pharmacology and*

therapeutics, 94(6):636–639, December 2013. ISSN 0009-9236. doi: 10.1038/clpt.2013.175. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3906042/.

[69] Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J. Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, and Zhiping Weng. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, July 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2493-4. URL https://www.nature.com/articles/s41586-020-2493-4. Publisher: Nature Publishing Group.

[70] Daniela Glavan, Victor Gheorman, Andrei Gresita, Dirk M. Hermann, Ion Udristoiu, and Aurel Popa-Wagner. Identification of transcriptome alterations in the prefrontal cortex, hippocampus, amygdala and hippocampus of suicide victims. *Scientific Reports*, 11(1):18853, September 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-98210-6.

[71] Jon M. McClellan and Mary-Claire King. A tipping point in neuropsychiatric genetics. *Neuron*, 109(9):1411–1413, May 2021. ISSN 1097-4199. doi: 10.1016/j.neuron.2021.04.002.

[72] Michael Schupp, Martina I. Lefterova, Jürgen Janke, Kirstin Leitner, Ana G. Cristancho, Shannon E. Mullican, Mohammed Qatanani, Nava Szwergold, David J. Steger, Joshua C. Curtin, Roy J. Kim, Moo-Jin Suh, Martin R. Albert, Stefan Engeli, Lorraine J. Gudas, and Mitchell A. Lazar. Retinol saturase promotes adipogenesis and is downregulated in obesity. *Proceedings of the National Academy of Sciences of the United States of America*, 106(4): 1105–1110, January 2009. ISSN 1091-6490. doi: 10.1073/pnas.0812065106.

[73] William R. Reay and Murray J. Cairns. The role of the retinoids in schizophrenia: genomic and clinical perspectives. *Molecular Psychiatry*, 25(4): 706–718, April 2020. ISSN 1476-5578. doi: 10.1038/s41380-019-0566-2. URL https://www.nature.com/articles/s41380-019-0566-2. Number: 4 Publisher: Nature Publishing Group.

[74] Edra London, Christina Tatsi, Steven J. Soldin, Christopher A. Wassif, Peter Backlund, David Ng, Leslie G. Biesecker, and Constantine A. Stratakis. Acute Statin Administration Reduces Levels of Steroid Hormone Precursors. *Hormone and metabolic research = Hormon- und Stoffwechselforschung = Hormones et metabolisme*, 52(10):742–746, October 2020. ISSN 0018-5043.

doi: 10.1055/a-1099-9556. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7495505/.

[75] Patrizia Risé, Silvia Ghezzi, and Claudio Galli. Relative potencies of statins in reducing cholesterol synthesis and enhancing linoleic acid metabolism. *European Journal of Pharmacology*, 467(1):73–75, April 2003. ISSN 0014-2999. doi: 10.1016/S0014-2999(03)01594-2. URL https://www.sciencedirect.com/science/article/pii/S0014299903015942.

[76] Yiran Zhang, Yingxiang Liu, Jin Sun, Wei Zhang, Zheng Guo, and Qiong Ma. Arachidonic acid metabolism in health and disease. *MedComm*, 4(5):e363, September 2023. ISSN 2688-2663. doi: 10.1002/mco2.363. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10511835/.

[77] Feixiong Cheng, István A. Kovács, and Albert-László Barabási. Network-based prediction of drug combinations. 10(1):1197. ISSN 2041-1723. doi: 10.1038/s41467-019-09186-x. URL https://www.nature.com/articles/s41467-019-09186-x. Publisher: Nature Publishing Group.

[78] Laura H. Goetz and Nicholas J. Schork. Personalized medicine: Motivation, challenges and progress. 109(6):952–963. ISSN 0015-0282. doi: 10.1016/j.fertnstert.2018.05.006. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6366451/.

[79] Christopher Taylor, Ian Crosby, Vincent Yip, Peter Maguire, Munir Pirmohamed, and Richard M. Turner. A review of the important role of CYP2d6 in pharmacogenomics. 11(11):1295. ISSN 2073-4425. doi: 10.3390/genes11111295. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7692531/.

[80] David W. Haas, Daniel R. Kuritzkes, Marylyn D. Ritchie, Shashi Amur, Brian F. Gage, Gary Maartens, Dan Masys, Jacques Fellay, Elizabeth Phillips, Heather J. Ribaudo, Kenneth A. Freedberg, Christos Petropoulos, Teri A. Manolio, Roy M. Gulick, Richard Haubrich, Peter Kim, Marjorie Dehlinger, Rahel Abebe, and Amalio Telenti. Pharmacogenomics of HIV therapy: Summary of a workshop sponsored by the national institute of allergy and infectious diseases. 12(5): 277–285. ISSN 1528-4336. doi: 10.1310/hct1205-277. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322423/.

[81] M. V. Relling and T. E. Klein. CPIC: Clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. 89(3): 464–467. ISSN 1532-6535. doi: 10.1038/clpt.2010.279.

[82] J. J. Swen, M. Nijenhuis, A. de Boer, L. Grandia, A. H. Maitland-van der Zee, H. Mulder, G. a. P. J. M. Rongen, R. H. N. van Schaik, T. Schalekamp, D. J. Touw, J. van der Weide, B. Wilffert, V. H. M. Deneer, and H.-J. Guchelaar. Pharmacogenetics: from bench to byte–an update of guidelines. 89(5):662–673. ISSN 1532-6535. doi: 10.1038/clpt.2011.34.

[83] Isabel Pérez-Núñez, José L. Pérez-Castrillón, María T. Zarrabeitia, Carmen García-Ibarbia, Laura Martínez-Calvo, José M. Olmos, Laisa S. Briongos, Javier Riancho, Victoria Camarero, Josep M. Muñoz Vives, Raquel Cruz, and José A. Riancho. Exon array analysis reveals genetic heterogeneity in atypical femoral fractures. a pilot study. 409(1):45–50. ISSN 1573-4919. doi: 10.1007/s11010-015-2510-3. URL

https://doi.org/10.1007/s11010-015-2510-3.

[84] Roca-Ayats Neus, Balcells Susana, Garcia-Giralt Natàlia, Falcó-Mascaró Maite, Martínez-Gil Núria, Abril Josep F., Urreizti Roser, Dopazo Joaquín, Quesada-Gómez José M., Nogués Xavier, Mellibovsky Leonardo, Prieto-Alhambra Daniel, Dunford James E., Javaid Muhammad K., Russell R. Graham, Grinberg Daniel, and Díez-Pérez Adolfo. GGPS1 mutation and atypical femoral fractures with bisphosphonates. 376(18):1794–1795. doi: 10.1056/NEJMc1612804. URL https://www.nejm.org/doi/full/10.1056/NEJMc1612804. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMc1612804.

[85] Lauren E. Surface, Damon T. Burrow, Jinmei Li, Jiwoong Park, Sandeep Kumar, Cheng Lyu, Niki Song, Zhou Yu, Abbhirami Rajagopal, Yangjin Bae, Brendan H. Lee, Steven Mumm, Charles C. Gu, Jonathan C. Baker, Mahshid Mohseni, Melissa Sum, Margaret Huskey, Shenghui Duan, Vinieth N. Bijanki, Roberto Civitelli, Michael J. Gardner, Chris M. McAndrew, William M. Ricci, Christina A. Gurnett, Kathryn Diemer, Fei Wan, Christina L. Costantino, Kristen M. Shannon, Noopur Raje, Thomas B. Dodson, Daniel A. Haber, Jan E. Carette, Malini Varadarajan, Thijn R. Brummelkamp, Kivanc Birsoy, David M. Sabatini, Gabe Haller, and Timothy R. Peterson. ATRAID regulates the action of nitrogen-containing bisphosphonates on bone. 12(544):eaav9166. ISSN 1946-6242. doi: 10.1126/scitranslmed.aav9166.

[86] Francesca Marini, Francesca Giusti, Elena Marasco, Luciano Xumerle, Katarzyna Malgorzata Kwiatkowska, Paolo Garagnani, Emmanuel Biver, Serge Ferrari, Giovanni Iolascon, Teresa Iantomasi, and Maria Luisa Brandi. High frequency of heterozygous rare variants of the SLC34a1 and SLC9a3r1 genes in patients with atypical femur fracture. 188(1):lvad001. ISSN 1479-683X. doi: 10.1093/ejendo/lvad001.

[87] Hiroshi Furukawa, Shomi Oka, Naoki Kondo, Yasuaki Nakagawa, Naofumi Shiota, Kenji Kumagai, Keiji Ando, Tsutao Takeshita, Takenori Oda, Yoshinori Takahashi, Kazutaka Izawa, Yoichi Iwasaki, Kazuhiro Hasegawa, Hiroshi Arino, Takeshi Minamizaki, Norie Yoshikawa, Shinjiro Takata, Yasuo Yoshihara, and Shigeto Tohma. The contribution of deleterious rare alleles in ENPP1 and osteomalacia causative genes to atypical femoral fracture. 107(5): e1890–e1898. ISSN 1945-7197. doi: 10.1210/clinem/dgac022.

[88] Aimee M. Deaton, Fan Fan, Wei Zhang, Phuong A. Nguyen, Lucas D. Ward, and Paul Nioi. Rationalizing secondary pharmacology screening using human genetic and pharmacological evidence. 167(2):593–603. ISSN 1096-0929. doi: 10.1093/toxsci/kfy265.

[89] Eric Vallabh Minikel and Matthew R. Nelson. Human genetic evidence enriched for side effects of approved drugs. URL http://medrxiv.org/lookup/doi/10.1101/2023.12.12.23299869.

[90] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with GWAS: a review. 9(1):29. ISSN 1746-4811. doi: 10.1186/1746-4811-9-29. URL https://doi.org/10.1186/1746-4811-9-29.

[91] Niclas Björn, Tejaswi Venkata Satya Badam, Rapolas Spalinskas, Eva Brandén, Hirsh Koyi, Rolf Lewensohn, Luigi De Petris, Zelmina Lubovac-Pilav, Pelin Sahlén, Joakim Lundeberg, Mika Gustafsson, and Henrik Gréen. Whole-genome sequencing and gene network modules predict gemcitabine/carboplatin-induced myelosuppression in non-small cell lung cancer patients. 6(1):25. ISSN 2056-7189. doi: 10.1038/s41540-020-00146-6.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 2055

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-527102

ACTA UNIVERSITATIS
UPSALIENSIS
2024