



## Data-driven Approaches to Explore Precision Medicine

Garcia, Sara

*Publication date:*  
2021

*Document Version*  
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

*Citation (APA):*  
Garcia, S. (2021). *Data-driven Approaches to Explore Precision Medicine*. DTU Health Technology.

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

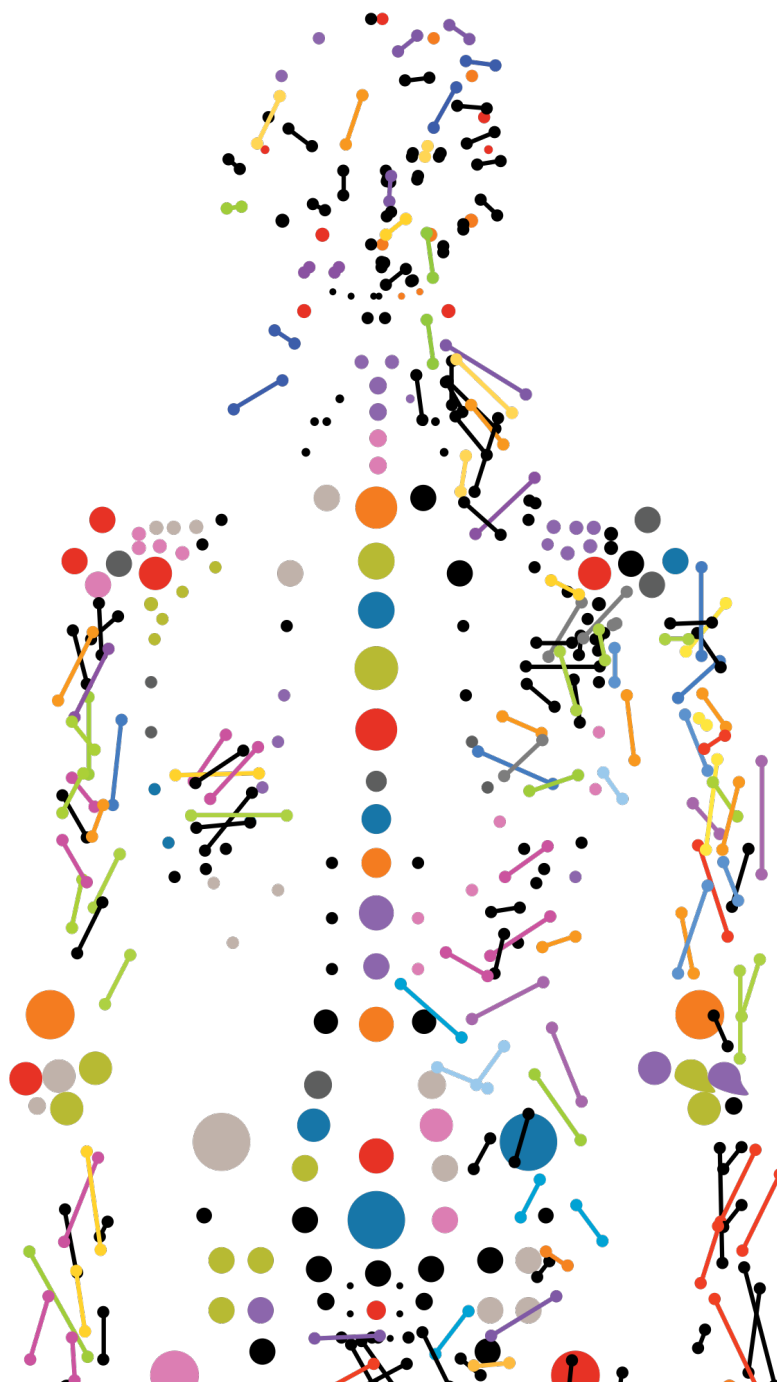
- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Data-driven Approaches to Explore Precision Medicine

PhD Thesis

Sara Garcia  
April, 2021



# Data-driven Approaches to Explore Precision Medicine

Sara Garcia

Kongens Lyngby, 2021



# Contents

<b>Preface</b>	<b>iii</b>
<b>Publications</b>	<b>iv</b>
<b>Abstract</b>	<b>viii</b>
<b>Danske Resumé</b>	<b>x</b>
<b>Acknowledgements</b>	<b>xi</b>
 <b>I Introduction</b>	 <b>1</b>
<b>1 Approaches to medicine and patient data</b>	<b>2</b>
1.1 Traditional and precision medicine – are they really that different? . . . . .	2
1.1.1 Ayurveda, Indian traditional medicine . . . . .	2
1.1.2 Conventional medicine . . . . .	4
1.1.3 Precision medicine . . . . .	5
1.1.4 Intersection between types of medicine . . . . .	5
1.2 Clinical and genomics data: challenges and opportunities . . . . .	6
1.2.1 Clinical data . . . . .	7
1.2.2 Genetic biomarkers . . . . .	8
 <b>2 Tools and approaches with genetic data</b>	 <b>10</b>
2.1 Genome-wide association studies . . . . .	10
2.1.1 SNP microarray . . . . .	10
2.1.2 Quality control . . . . .	11
2.1.3 Genome imputation . . . . .	13
2.1.4 Disease and single-nucleotide polymorphism association . . . . .	14
2.2 Next-generation sequencing . . . . .	15
2.2.1 Template preparation . . . . .	17
2.2.2 DNA sequencing . . . . .	17
2.2.3 Data analysis . . . . .	17



2.3	Polygenic risk score . . . . .	19
2.3.1	Calculate polygenic risk scores . . . . .	19
2.3.2	PRS clinical application . . . . .	21
2.4	Machine learning . . . . .	22
2.4.1	Data encoding . . . . .	22
2.4.2	Feature selection . . . . .	23
2.4.3	Cross-validation and hyperparameters tuning . . . . .	23
2.4.4	Model selection and training . . . . .	24
2.4.5	Model evaluation . . . . .	30
<b>3</b>	<b>Disease risk and risk management</b>	<b>32</b>
3.1	Disease predisposition . . . . .	32
3.2	Treatment response . . . . .	33
3.3	Late-side effects . . . . .	33
3.4	Translating cancer genomics into precision medicine . . . . .	34
<b>II</b>	<b>Papers</b>	<b>36</b>
<b>4</b>	<b>Paper I: Ayurveda GWAS</b>	<b>38</b>
<b>5</b>	<b>Paper II: Genetics scores in childhood cancer</b>	<b>57</b>
<b>6</b>	<b>Paper III: Hearing loss prediction</b>	<b>75</b>
<b>7</b>	<b>Paper IV: Nephrotoxicity prediction</b>	<b>99</b>
<b>III</b>	<b>Other projects</b>	<b>113</b>
<b>8</b>	<b>Application note: Fluctuation measures</b>	<b>114</b>
<b>9</b>	<b>Remote external stay: Dasatinib resistance in T-ALL</b>	<b>123</b>
<b>IV</b>	<b>Epilogue</b>	<b>130</b>
<b>V</b>	<b>Appendix: Research efforts in India, and thoughts around Ayurveda</b>	<b>134</b>
	<b>Bibliography</b>	<b>137</b>

# Preface

The PhD project was carried out at the department of Health Technology at the Technical University of Denmark between January 2018 and April 2021 to fulfil the requirements for acquiring a PhD degree. The project was funded by Fondation Idella.

This thesis consists of a general introduction followed by four research articles: one published, two submitted, and one in preparation. Additionally, one application note in preparation is included in chapter 8, as well as a description of the work done during my remote external stay at St. Jude Children's Research hospital in chapter 9.

The projects were carried out under the main supervision of associate professor Ramneek Gupta and co-supervision of professor B.K. Thelma, senior researcher Marlene Danner Dalgaard, and associate professor Elena Papaleo.

Kongens Lyngby, April 2021

Sara Garcia

A handwritten signature in dark ink, appearing to read 'Sara Garcia', with a horizontal line underneath.

# Publications

## Papers included in the thesis

### **Stratification of Rheumatoid Arthritis Cohort Using Ayurveda Based Deep Phenotyping Approach Identifies Novel Genes in a GWAS**

Garima Juyal, Anuj Pandey\*, Sara L Garcia\*, Sapna Negi, Ramneek Gupta, Uma Kumar, Bheema Bhat, Ramesh C Juyal, Thelma B K

Manuscript submitted to Journal of Traditional and Complementary Medicine

### **Evaluation of adult cancer polygenic risk scores for stratified disease prevention in childhood cancer**

Sara L Garcia, Marianne Helenius, Jonas Vestergaard, Adrian O. Laspior, Thomas van Overeem Hansen, Ulrik Soltze, Kjeld Schmiegelow, Ramneek Gupta, Rikke L. Nielsen, Karin Wadt

Manuscript in preparation

### **Predicting hearing loss after cisplatin chemotherapy in testicular cancer patients**

Sara L Garcia\*, Jakob Lauritsen\*, Bernadette K. Christiansen, Ida F. Hansen, Mikkil Bandak, Marlene D. Dalgaard, Gedske Dagaard, Ramneek Gupta

Manuscript submitted to JAMA Oncology

### **Prediction of Nephrotoxicity Associated With Cisplatin-Based Chemotherapy in Testicular Cancer Patients**

Sara L Garcia\*, Jakob Lauritsen\*, Zeyu Zhang\*, Mikkil Bandak, Marlene D Dalgaard, Rikke L Nielsen, Gedske Dagaard, Ramneek Gupta

*JNCI Cancer Spectrum, Volume 4, Issue 3 (June 2020)*

\* Equally contributing authors

## Application note included in the thesis

### **FLUCbio: a python package for fluctuation modelling on postprandial biological data**

Sara L Garcia, Cecilia B. Jensen, Rikke Linnemann Nielsen, Ramneek Gupta

Application note in preparation to Oxford Bioinformatics

## Papers not included in the thesis

### **Data integration for prediction of weight loss in randomized controlled dietary trials**

Rikke Linnemann Nielsen\*, Marianne Helenius\*, Sara L. Garcia, Henrik M. Roager, Derya Aytan-Aktug, Lea Benedicte Skov Hansen, Mads Vendelbo Lind, Josef K. Vogt, Marlene Danner Dalgaard, Martin I. Bahl, Cecilia Bang Jensen, Rasa Muktupavela, Christina Warinner, Vincent Aaskov, Rikke Gøbel, Mette Kristensen, Hanne Frøkiær, Morten H. Sparholt, Anders F. Christensen, Henrik Vestergaard, Torben Hansen, Karsten Kristiansen, Susanne Brix, Thomas Nordahl Petersen, Lotte Lauritzen, Tine Rask Licht, Oluf Pedersen, Ramneek Gupta

*Sci Rep* 10, 20103 (2020)

### **Association between brown eye colour in rs12913832:GG individuals and SNPs in *TYR*, *TYRP1*, and *SLC24A4***

Olivia S. Meyer , Maja M. B. Lunn, Sara L. Garcia, Anne B. Kjærbye, Niels Morling, Claus Børsting, Jeppe D. Andersen

*PloS one* 15, no. 9 (2020)

### **Whole genome and whole exome sequencing in patients with early onset colorectal cancer and familial colorectal cancer**

Djursby M, Hansen TVO, Asplund AM, Bak M, Garcia SL, Jane H. Frederiksen, Dunø M1, Risom L, Madsen MB, Nielsen FC, Melchior L, Willemoe GL, Nilbert M, Therkildsen C, Lone Sunde, Charlotte Lautrup, Karina Rønlund, Okkels H, Wikman F, Gerdes AM, Wadt K

Manuscript in preparation

\* Equally contributing authors

# List of Abbreviations

<b>AI</b>	artificial intelligence
<b>ALL</b>	acute lymphoblastic leukemia
<b>ANN</b>	artificial neural network
<b>AUC</b>	area under the curve
<b>BMI</b>	body mass index
<b>BQSR</b>	base quality score recalibration
<b>BWA-MEM</b>	burrows-wheeler alignment - maximal exact matches
<b>ddNTP</b>	dideoxynucleotide triphosphates
<b>DNA</b>	deoxyribonucleic acid
<b>dNTP</b>	deoxynucleotide triphosphates
<b>FDR</b>	false discovery rate
<b>FN</b>	false negative
<b>FP</b>	false positive
<b>GATK</b>	genome analysis toolkit
<b>GRCh37</b>	Genome Reference Consortium Human Build 37
<b>GWAS</b>	genome-wide association study
<b>HWE</b>	hardy–weinberg equilibrium
<b>IBD</b>	identity-by-descent
<b>IBS</b>	identity-by-state
<b>Indels</b>	insertions and deletions
<b>LASSO</b>	least absolute shrinkage and selection operator

<b>LD</b>	linkage disequilibrium
<b>MAF</b>	minor allele frequency
<b>MCMC</b>	markov chain monte carlo
<b>MDS</b>	multidimensional scaling
<b>ML</b>	machine learning
<b>NetBID</b>	network-based bayesian inference of drivers
<b>NGS</b>	next-generation sequencing
<b>NPV</b>	negative predictive value
<b>OR</b>	odds ratio
<b>PCA</b>	principal component analysis
<b>PCR</b>	polymerase chain reaction
<b>PPV</b>	positive predictive value
<b>PRS</b>	polygenic risk score
<b>ROC</b>	receiver operating characteristic
<b>RSS</b>	residual sum of squares
<b>SD</b>	standard deviation
<b>SNP</b>	single-nucleotide polymorphism
<b>TERT</b>	telomerase reverse transcriptase
<b>TLR</b>	toll-like receptors
<b>TN</b>	true negative
<b>TNF</b>	tumor necrosis factor
<b>TP</b>	true positive
<b>WGS</b>	whole-genome sequencing

# Abstract

Precision medicine in a contemporary context implies customising healthcare based on individual biomarkers, such as genetic variants or lifestyle factors. The purpose is to prevent, diagnose or find the most effective disease treatment approaches customised for the individual or subgroups of patients instead of a one-size-fits-all approach. As a systemised approach, this concept has come into focus in recent years in modern translational science; however, it should be noted that ancient medicine systems such as Ayurveda, a traditional medicine in India, has over centuries of history looking into patient stratification in relation to disease development and treatment, and a fairly layered system to describe it that incorporates elements of lifestyle, behaviour, diet and proxy biomarkers for underlying genetics.

In this PhD thesis, I have 1) explored precision medicine concepts from different perspectives; 2) used different approaches to analyse patient clinical (application note, chapter 8) and genomics data; 3) utilised genetics from genome-wide association studies and next-generation sequencing analysis; 4) developed stratification-based models, such as Ayurveda-based deep phenotyping, polygenic risk scores, and machine learning models; and 5) discussed how these models could be applied in a clinical setting for prediction of phenotypes, treatment response and late-side effects.

The first paper, presented in chapter 4, explores the use of Ayurveda medicine for patient stratification to help identify novel disease genetic variants that predispose towards rheumatoid arthritis. The second paper, chapter 5, uses two developed and validated adult cancer polygenic risk scores to explore risk stratification for different phenotypes in childhood cancer. The third and fourth papers, chapters 6 and 7, respectively, focus on the development of machine learning models to predict treatment late-side effects, specifically, cisplatin-induced hearing loss and nephrotoxicity, respectively, in testicular cancer patients, using clinical and genomics data. In chapter 9, it is presented a model that predicts dasatinib treatment response in T-cell acute lymphoblastic leukaemia. This work was developed at St. Jude Children's Research hospital during my external stay.

These stratification-based models may help leverage heterogeneous clinical data and find disease-associated genomic markers. Furthermore, implementing these models in a clinical

---

context, together with medical expertise, may allow for earlier disease diagnosis, personalised prevention, and treatment strategies for groups of people based on their genomics and clinical profiles. Ultimately, this will enable a better balance between treatment efficacy and patient's quality of life.



# Danske Resumé

Præcisionsmedicin betyder individualiseret medicin/behandling baseret på individuelle biomarkører, såsom genetiske varianter eller livsstilsfaktorer, for at forbygge, diagnosticere eller bestemme den mest effektive sygdomsbehandling til patienter i stedet for en one-size-fits-all-tilgang. Dette koncept er kommet i fokus de seneste år. På den anden side skal det dog bemærkes, at traditionel medicin såsom Ayurveda har over århundreder lang historie fra Indien. Her undersøges patientens stratificering i forhold til sygdomsudvikling og behandling. I denne PhD afhandling har jeg 1) udforsket konceptet med præcisionsmedicin fra forskellige perspektiver; 2) brugt forskellige tilgange til at analysere kliniske (ansøgningsnote, kapitel 8) og genomisk data, og øge genomisk forståelse ved at bruge "Genome-Wide Association Study" (GWAS) og næste generations sekventeringsanalyse (NGS); 3) udviklet modeller til forudsigelse af risiko, såsom Ayurveda-baseret dyb fænotypebestemmelse, polygenetiske risikoscorer og maskinlæringsmodeller; og 4) undersøgt, hvordan disse modeller kan anvendes i klinikken til forudsigelse af fænotype, behandlingsrespons og sene bivirkninger.

Den første artikel, kapitel 4, præsenterer en oversigt over Ayurveda-medicin, og hvordan dennes patientstratificeringsmetode kan hjælpe med at identificere nye sygdomsvarianter, specielt i leddegigt. Det andet kapitel, kapitel 5, undersøger anvendelse af to validerede polygene risikoscorer baseret på kolon- og bryst-kræft i voksne til at stratificere risiko-fænotyper i børn diagnosticeret med kræft. Arbejdet udviklet på St. Jude Children's Research hospital, kapitel 9, præsenterer en model, der forudsiger behandlingsrespons, specifikt dasatinib-respons i T-celle akut lymfoblastisk leukæmi. Artikler tre og fire i kapitlerne 6 og 7, fokuserer på udvikling af maskinlæringsmodeller til forudsigelse af senfølger ved behandling, specifikt cisplatin-induceret høretab og nefrotoksicitet hos testikelkræftpatienter ved hjælp af kliniske og genomiske data.

Implementeringen af disse modeller i en klinisk sammenhæng kan sammen med medicinsk ekspertise hjælpe med at løse kliniske behandlingsudfordringer, hvilket muliggør tidligere diagnose, personlig forebyggelse og behandlingsstrategier for grupper af mennesker baseret på deres genomiske og kliniske profiler.

# Acknowledgements

I want to thank my main supervisor Ramneek Gupta for the opportunity of doing a PhD in the DDI group. When I finished my master's, I decided to learn bioinformatics and, while very challenging, this PhD allowed me to learn exponentially. Thanks for this and all the exciting discussions, collaboration opportunities, and fun social events.

Furthermore, I would like to thank my co-supervisors. BK Thelma, for being so great in our visit to India in the Ayurveda collaboration, for all the insights on the Ayurveda field, and for opening up our minds about traditional medicine. Marlene Dalgaard, for checking up on me. Elena Papaleo, even though we did not have the opportunity to work much together, for the feedback on my last month.

During my PhD, I had the opportunity of doing a remote external stay with Jiyang Yu and Jun Yang at St.Jude Children's Research Hospital, Memphis, USA. It's been a pleasure to be part of Jiyang Yu group for approximately five months and getting to know the excellent research they do there. Unfortunately, COVID-19 hit when I was ready to go, but I hope we will have the opportunity to meet in person another time.

I also would like to thank the collaborators across India and at Rigshospitalet, Denmark, who always had a tremendous clinical insight that improved a lot of the thesis work and helped me understand how this could be implemented in a clinical setting.

Learning programming from scratch (thanks Peter Wad!), I was absorbing as much information as possible when I started. I thank everyone at the section of Bioinformatics and the DDI group, present and former members, for numerous discussions and joyful pre-COVID lunch breaks.

I want to thank my amazing friends in Portugal, Denmark, and around the globe. I hope I will be able to visit and travel with you very soon.

To my parents, my brother, my aunt, my grandparents, and anya, you were and are my greatest support that kept me going throughout tough times. Thank you for believing in me, for the good advice, and for the every day phone calls. Thanks mom, for always making me see the cup "half-full". Lastly, I want to thank Balint, my everyday partner and friend, for listening to me, for calming me down and for cheering me up. I am very lucky to have you all.

# Part I

## Introduction

# Chapter 1

## Approaches to medicine and patient data

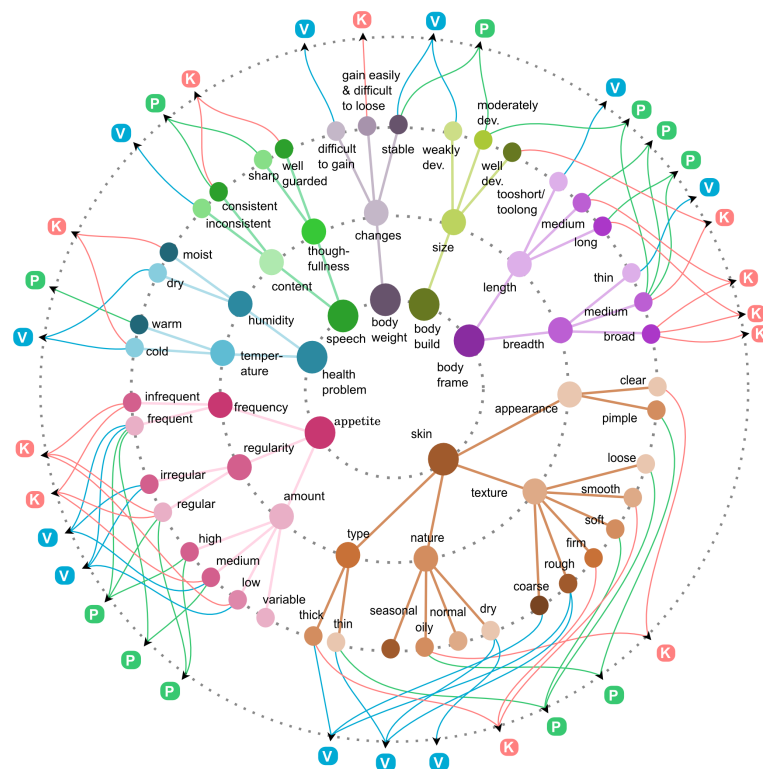
### 1.1 Traditional and precision medicine – are they really that different?

Precision medicine refers to individually tailored health care for disease prevention or better disease treatments. It is based on individual characteristics, such as one person's genes, lifestyle, and environmental factors. Currently, its use in the clinic is limited, even though a more systematic implementation is promising due to advances in genetics and the increase of patient data [1]. There is no doubt that "precision medicine" has become very popular recently [2]. However, physicians have tailored therapeutic recommendations to patients' specific characteristics for a long time. Pre-modern medical knowledge was often personalized, and one of the first examples in medical history is Ayurveda, one of the Indian traditional medicine documented and practised for centuries [3].

#### 1.1.1 Ayurveda, Indian traditional medicine

Ayurveda is a traditional medicine in India and one of the oldest in the world. It is a form of alternative medicine and likely the earliest example of predictive, preventive, personalized, and participatory (P4) medicine [4]. According to Ayurveda, we can define a person with a specific basic constitution at the time of birth. This basic constitution is known as Prakriti in the Ayurveda lexicon. Prakriti classification takes into account a person's physical, physiological and psychological constitution, and it will define, to a great extent, a person's predisposition to diseases and response to the environment, diet, and drug treatments [5] as can be observed in Figure 1.1. There are seven contrasting phenotypic categories. Three are the so called extreme phenotypes, named vata, pitta or

kapha. The other four categories, vata-pitta, pitta-kapha, vata-kapha, vata-pita-kapha, are non-extreme categories [6][7].



**Figure 1.1** | Illustration with examples of Prakriti characteristics. Figure from [6].

Prakriti is accessed by Ayurveda physicians and available software, such as the AyuSoft questionnaire. AyusSoft is based on 85 questions related to individuals' anatomy, physiology, and psychology. It gives a higher weight to anatomy-related questions than the others since those are more stable throughout life [8].

These ideas on aetiology, treatment and responses have been based on theoretical knowledge. Ayurveda uses a holistic approach emphasizing health improvement, disease prevention, early diagnosis and personalized treatment based on multiple aspects from herbal medicine to advice on diet, physical exercise, and lifestyle both to improve health or treat a disease. The 'Ayurveda pharmacopoeia of India' documents close to a thousand Ayurveda sources [9]. These Prakriti specific prescriptions are a form of personalized medicine [10]. One study that compared Ayurveda, methotrexate, and their combination for the treatment of rheumatoid arthritis, reported that the three treatments have similar efficacy, with the Ayurveda having fewer side effects. In this study, the patients were randomly assigned to different groups, and it also included a placebo group. This study was also double-blinded, meaning that neither the Ayurvedic nor the allopathic physicians had information about the treatment received by the patient [11].

In another randomized, double-blind study, two Ayurvedic treatment were compared with

glucosamine and celecoxib for knee osteoarthritis. It was observed that Ayurvedic drugs reduced knee pain considerably and improved knee function. Additionally, the Ayurvedic medicines had similar efficacy when compared to glucosamine and celecoxib, widely used for cases of knee osteoarthritis [12].

Integration of Ayurveda stratified approach in the modern world of genomics, i.e. Ayurgenomics could be a great complement to precision medicine [13], as there is a lot of knowledge we could use from our ancestors. Indeed, literature bridging Ayurveda to western medical contexts, with proof of principle, is growing [14][15][16][17]. However, applications on specific diseases and with controlled patient cohorts, have been few.

## Ayurgenomics

Ayurgenomics, a relatively recent concept, hypothesizes that integration of Prakriti with genomics and modern biology can validate Ayurveda concepts and help discover genetic markers important for disease predisposition and response to treatment [18]. There are two main objectives of Ayurgenomics: 1) provide scientific validation of Ayurveda system of medicine, and 2) obtain homogeneous disease cohorts for genetic analysis of common complex traits using Prakriti-based subgrouping [19].

One of the first research studies on this area demonstrated that individuals from the three different extreme Prakriti groups have evident differences when looking at biochemical parameters. For example, kapha was seen to have higher levels of triglycerides or total cholesterol, which confer a higher risk for cardiovascular diseases. Additionally, multiple differentially expressed genes were also found in the different groups [20]. Apart from genomics, recent manuscripts are starting to appear relating different Prakriti with different metagenomics signatures [21] and metabolomics pathways [22].

Ayurgenomics is currently a separate unit within the Institute of Genomics and Integrative Biology, New Delhi, India, collaborating with other institutions such as the All India Institute of Ayurveda, New Delhi, India.

### 1.1.2 Conventional medicine

Conventional medicine, also referred to as western in this thesis, refers to medicine as we know it today in most European and Northern American countries.

Diving a little into history, it was on the 19<sup>th</sup> century during the Industrial Revolution that many scientific discoveries in the field of biology were made. These led to medicine as we know it today. We can take the example of Gregor Mendel and his principles of inheritance, describing the transmission of genetic traits (1856-1863) before anyone knew what a gene was. This knowledge was expanded, and we know today that most diseases are complex or multifactorial, and complex methods are needed to find the multiple genes

associated with them. These will be discussed in more detail in chapter 2.

By the end of 20<sup>th</sup> century, the Human Genome Project was initiated. After 13 years, in 2003, its finished version was completed. This significant milestone led to a greater understanding of medicine and the discovery of single-nucleotide polymorphisms (SNPs) and genes associated with several phenotypes [23].

At the moment, it is much faster (1-2 days) and cheaper (2001: \$100,000,000; now: \$1,000) to sequence the human genome [24]. This changed the field of genomics. Now we have the fundamental knowledge and technologies available that allow us to read out a patient's genome routinely. We are gradually figuring out how to look at patient's genome differences and make medical decisions based on them, opening doors for more personalized medicine.

### 1.1.3 Precision medicine

Precision medicine implies customized healthcare to a subgroup of patients. By screening patient data and biomarkers, such as genetic biomarkers, and identifying which ones are specific to individual patients, we can point to better clinical decisions and resource allocation; instead of using the "one-size-fits-all" approach. The "one-size-fits-all" system aims to find treatments that work for the average patient. However, each of us is unique, and we respond in different ways to treatments and develop different early or late side effects depending on our genetic makeup.

It should be noted that some use the terms precision, personalized, P4, or stratified medicine interchangeably [25][2], while others point out some differences between them [26] [27].

In this PhD thesis, these terms were used interchangeably as I believe they have a common goal: develop and implement tailored healthcare for the diagnosis, prevention, and treatment of diseases.

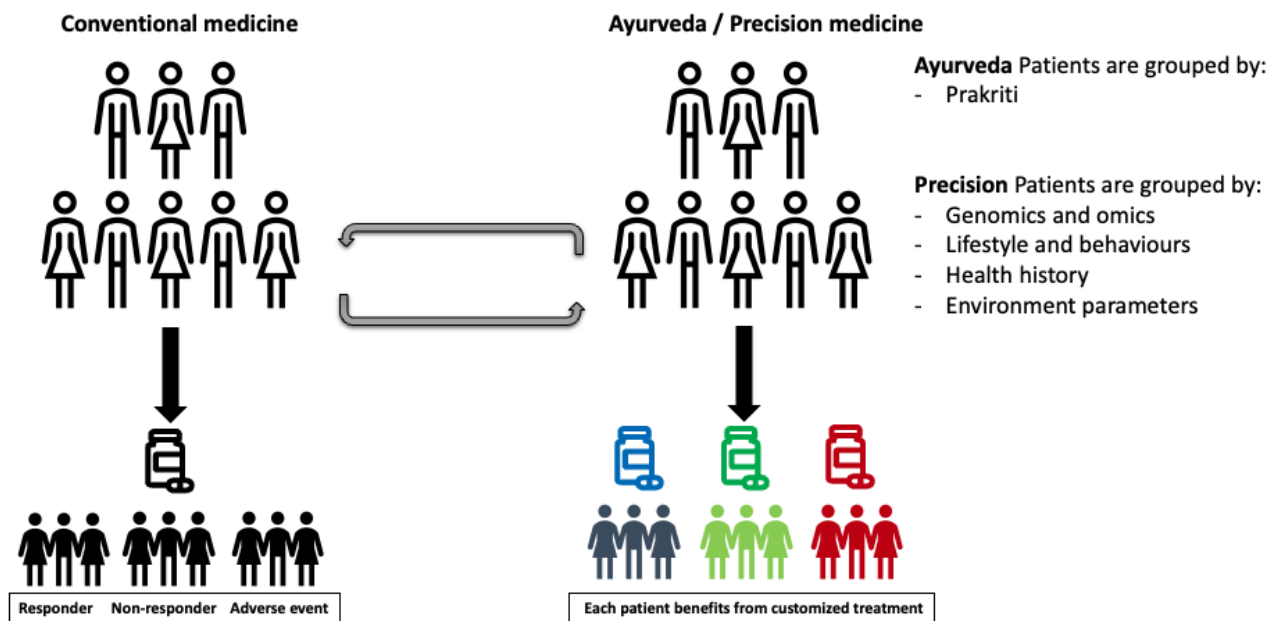
### 1.1.4 Intersection between types of medicine

Even though there is scepticism in the western culture concerning Ayurveda medicine, Ayurveda adopted precision medicine for centuries. Ayurveda and precision medicine share many aspects, as discussed in the current chapter. The first medical interventions were indeed personalized; however, as there was a lack of understanding of disease biology, these treatments were often ineffective.

The three types of medicine described here, Ayurveda, conventional and precision, can complement each other in some ways (Figure 1.2).

Traditional medicine was used when there was no technology available; thus, it is not data-driven but rather based on hundreds of years of traditional knowledge passed down from generation to generation. Today, we realize that, even though one drug may work

very well to treat some patients, the same drug can trigger early or late side effects in others, affecting the patients quality of life significantly. Thus, we need to be aware of the individual characteristics of each one of us and try to foresee how will a person respond to a specific treatment and if this person is prone to develop side effects that may persist throughout life. This viewpoint is shifting the concept of modern medicine, adopting the same ideas defended by our ancestors, and coming back to our roots.



**Figure 1.2** | Comparison between types of medicine: conventional, Ayurveda and precision. Ayurveda and precision medicine adopt a stratified-based approach based on different individual characteristics. Figure adapted from [28].

## 1.2 Clinical and genomics data: challenges and opportunities

Most complex disease development as well as individual treatment responses are the result of the interaction between hundreds of genome variants, defined at birth, and lifestyle behaviours, that change along the life course of the patients. Patient data is certainly very important but also very challenging to work with it. Normally there is also not a huge number of records, and that makes it more difficult to find patterns to better understand disease and disease sub-groups in such heterogeneous populations. To help with this, it is important to install a good healthcare data management system, and healthcare professionals are open to make recording and monitoring of these files better [29].

During my PhD, I have been working with clinical data and, in the OMICS field, genomics (mostly) and metabolomics ('Papers not included in the thesis', [30]), however, the work



on metabolomics will not be developed in this thesis, thus I will keep the focus on clinical and genomics data only.

### 1.2.1 Clinical data

Clinical data refers to basic information and/or health-related patient status that is part of the regular patient care such as age, gender, clinical biomarkers, or lifestyle behaviours like drinking or smoking habits.

In order to find relevant information, it is important to clean the data between different data formats, doctor's notes, and diagnostic codes. It is quite common to find duplicates due to an error of passing the information between paper records to electronic records, or due to multiple name variations given to a patient. We should also make sure the data does not contain inaccuracies and odd values. For instance, if body mass index (BMI) takes the value of 1 or 100, we know that is most probably inaccurate. In this case, we need to find the source of the error, which may be only a misplaced comma.

#### Missing data

It is very common to have sparse or incomplete clinical data; thus, extra attention is required. We do not want to take the risk of having biased data and unreliable and misleading results.

There may be different causes for the missing data. Missing at random holds the assumption that missingness is random within observed background characteristics in subgroups of the population, while missing not at random depends on unobserved variables [31][32]. It is hard to know for sure if missing values are missing at random or not. We need to compare observed with unobserved values, and the last are unknown; however, imputation should only be done if we are quite confident that values are missing at random. For example, if blood samples biomarkers or anthropometric measurements are missing due to an incident in the laboratory or lost files. Suppose there is any reason to assume that the values are not missing randomly. In that case, imputation should not be performed since we are not aware of the true nature of the missingness, and we may end up over or underestimating our results [33]. For example, if data is missing because the patients were not eligible for the study.

Missing data is avoided by doing good data collection or getting back to the patients to fill up the lacking information. This is not always possible; thus, several methods can be used. Simple methods include single imputation, where we use the variable observed values mean or median to replace the unobserved values. Another commonly used approaches are complete case analysis, where we remove samples with missing data and multivariate imputation [34]. Multivariate imputation differs from single imputation as missing data

is filled in many iterations with different possible values estimated for each missing value. This quantifies uncertainty by looking into the different values estimated at each iteration. There are multiple decisions that one needs to make, such as, what should be the first imputed value (for example, the most frequent value of the specific variable), or how many iterations on the dataset there should be. At each iteration, a variable is labelled as output  $y$  and the other variables input  $x$ . A linear or logistic regression is then fitted on  $x, y$  to then predict  $y$  [35].

These approaches help deal with the complexity of missing data, but none is a gold-standard solution. When stating one study's results, the extent of the missing data and the limitations should be reported clearly.

In this PhD thesis, if the phenotype or outcome was missing for one sample, this sample was removed as we did not want to risk modelling a false outcome. If other clinical variables were missing, the following methods were used: 1) multivariate imputation, modelling each variable containing missing values as a function of other variables in a round-robin fashion way (paper III, chapter 6); or 2) remove missing data, using a complete case analysis (paper IV, chapter 7).

### Single-time point vs longitudinal biological data

Clinical data can be either a single-time point measured at the time of visit, i.e., age or gender, or if a quantifiable biological parameter, measured over time. When dealing with a single-time point, this can be characterised by the value itself. In the case of longitudinal biological data, such as glucose postprandial responses, several methods have been developed to profile different glycemic patterns [36][37][38]. Various studies have also used the well-known area under the curve (AUC) [39][40][41]; however, with AUC there is a challenge capturing the fluctuation patterns of these temporal curves.

In this PhD thesis, different measurements were explored to model the dynamics of postprandial glucose responses. These fluctuation measurements are presented and discussed in more detail in the application note, chapter 8. A project GitHub page was also prepared; however, it will only be publicly available once the application note is published.

### 1.2.2 Genetic biomarkers

We find that 99.9% of the human genomes are the same, and it is the rest  $\approx 0.1\%$  that makes us all unique. This can also vary in some wildly divergent loci. The human leukocyte antigen region can reach over 10% variation across human genomes [42]. These genomic variations may impact diseases development. Still, some are simply associated with phenotypic characteristics, as skin or eye colour.

The genetic variations we find the most are SNPs. However, variants that cause disease

usually involve more than one single base-pair, such as insertions and deletions (indels), inversions and translocations [43].

Currently, several well-known genetic biomarkers are indeed used to diagnose and manage multiple diseases. For example, the well-known mutations in the BRCA genes that are known to confer a higher risk of breast and ovarian cancer [43].

Genomic tools such as DNA microarrays and next-generation sequencing (NGS) (discussed in more detail in chapter 2) allowed us to discover multiple variations associated with cancer and other diseases risk, evolution, and response to treatment. However, there is still a gap between the reported variants associated with diseases and the ones used in a clinical context. Some explanations for this gap are the use of inappropriate controls, the early disclose of rare variants with a not clear functional consequence and without functional validation, and the lack of replication in additional patients and laboratories [44][45].

## Chapter 2

# Tools and approaches with genetic data

In the field of cancer genomics, more and more data is being generated as new high throughput technologies become available, and sequencing becomes faster and cheaper. This raises the importance of data science and bioinformatics tools to analyse and make sense of all the information.

In this PhD thesis, GWASs and NGS analysis were performed to find variants associated with the disease of study. Further, Ayurveda-based phenotyping, polygenic risk scores (PRSs), and machine learning (ML) prediction models, integrating clinical and genomics data, were explored as primary approaches to translate these findings clinically.

GWASs, NGS, PRSs and ML are further explained in this chapter.

## 2.1 Genome-wide association studies

GWASs look into the genome to find associations between genetic variations and particular traits. Until now, they have shown multiple genetic influences on different physical characteristics [46][47], and multiple diseases and cancers [48].

### 2.1.1 SNP microarray

A SNP microarray is a technology used for SNP detection via hybridisation of single-stranded DNA sequences with unique oligonucleotides called probes bounded to the microarray. There are several alternative methods that have been developed by different companies [49][50][51].

In Illumina arrays, oligonucleotide probes targeting a specific locus in the genome are synthesised and attached to the array surface. DNA in each of the probes is oriented with the 5' end attached. The sample genomic DNA is fragmented and hybridised to the complementary sequence probe. The oligo on the probe is extended, where one of the four hapten-labelled dideoxynucleotide triphosphates (ddNTP) is added at a time. These haptens are detected by staining with fluorescently labelled proteins that bind each hapten

[49]. This method allows to simultaneously analyse hundreds of thousands of variants at locations in the genome that are known or suspected to correlate with disease.

There are many pre-made commercial available DNA arrays. Experts have selected various variants and locations spread in the genome to analyse their correlation with several traits.

### 2.1.2 Quality control

GenomeStudio software is used to process the array from raw intensity data to PLINK format. PLINK [52] is a program that allows quality control, an important step previous to the association analysis, to dismiss low-quality data and decrease the chance of false-positive associations [53].

Below, I will describe some of the standard steps of quality control, which were also performed on the research papers included in this PhD thesis.

**Genotyping data preparation** In this step, duplicated SNPs and those with ambiguous genome position, strand, and alleles (compared to a reference genome) are removed. The reference genome used in the research papers was Genome Reference Consortium Human Build 37 (GRCh37).

**Removal of individuals and SNPs with low call rate** A sample with a very low SNP call rate may indicate a poor quality DNA sample, and these samples should be removed from the analysis. SNP genotype failing in multiple DNA samples may point to systematic errors of array reaction or genotype-calling algorithms and SNPs in regions with a high number of copy number variations [54]. Even though this can slightly change between studies, a recommended threshold for SNPs call rate is 95%. For SNPs with a low minor allele frequency ( $MAF < 5\%$ ), this threshold should be stricter, i.e. 99%. For the samples, it is recommended to remove if more than 98% of the genotype is missing, but this can also depend on factors like the type of the study, genotype platform used, and quality of the DNA. We should determine a goal and find a balance between (minimising) the number of samples to remove and (maximising) genotyping efficiency [55][56].

**Removal of individuals with discordant sex information** Few times we find that the sex reported in the patient's clinical file and the sex found by the sex chromosome differ, and these samples are also considered unreliable and removed. In PLINK, a sex check can be made by X chromosome homozygosity estimate (F statistic). By default, if F estimate  $< 0.2$ , the sample is considered a female; otherwise, it is considered a male.

**Removal of individuals with excessive heterozygosity rate** Heterozygosity rate refers to the percentage of heterozygous genotypes for a specific individual. First, we need to calculate the heterozygosity rate for all samples and the mean heterozygosity

rate. Samples who deviate more than, for example, 3-4 standard deviations (SD) from the samples mean are excluded. Deviation from the samples mean may indicate DNA sample contamination, if high heterozygosity rate, or inbreeding, if low heterozygosity rate [56].

**Ancestry check** GWASs can be highly confounded due to population stratification yielding many false positives [57], which may be associated with the ancestry of cases versus controls and not with the trait under study. Principal component analysis (PCA) or multi-dimensional scaling (MDS) are two methods widely used to observe and correct population stratification. Both compare the population genetic diversity between the studied population and a reference genome, such as HapMap or 1000 Genomes. The HapMap Project, developed by an international consortium, consists of a map of shared patterns of DNA variations in the human genome, and it includes population from Africa (Kenia, Nigeria), Asia (China, Japan), Europe (Italia), and America [58]. The 1000 Genomes Project was another effort to include more samples diversity and perform a deeper characterisation of the human genome sequence [59].

**Removal of related individuals** The presence of high genetic similarity between individuals independent of the trait under study presents a source of potential bias in association tests in population-based studies, with the risk of yielding sub-population association instead of phenotype association. These samples should either be removed [60] or family relatedness should be taken into account in the case of family-based genetic association studies, for example, using variance components to account for family structure [61][62]. Relatedness can be accessed using identity-by-descent (IBD) and identity-by-state (IBS). If we look at a given locus in any pair of samples, IBS can be either 1) IBS0, if there are two different alleles (i.e., AA and BB); 2) IBS1, if there is one allele in common (i.e., AA and AB); or 3) IBS2, if the two alleles are the same (i.e., AA and AA). Two samples that share one or two alleles IBS at a given locus may have inherited those alleles from a common ancestor; thus, these alleles are identical-by-descent [63]. In PLINK, IBD is inferred from the observed IBS states using a hidden Markov model with a small hidden state space [64]. Subsequently, for each pair of samples, a proportion IBD ( $PI_{HAT}$ ) is calculated as defined in Equation 2.1, where  $P$  is probability.

$$PI_{HAT} = P(IBD = 2) + 0.5 \times P(IBD = 1) \quad (2.1)$$

Usually, a  $PI_{HAT}$  higher than 0.25 indicates close relatives [65], but an individual is removed from analysis if the  $PI_{HAT}$  is higher than 0.1875. A  $PI_{HAT}$  higher than 0.5 indicates first-degree relatives, and a  $PI_{HAT}$  of 1 shows duplicates or monozygotic twins [56].

**Removal of population outliers** MDS on an IBS matrix allows checking for population

outliers. PLINK uses raw Hamming distances to calculate pairwise IBS distance between individuals. A different threshold can be used to remove outlier samples from the analysis. One standard approach consists of removing samples with an IBS genetic distance from the sample mean of more than 3 SD on one or more clusters [66].

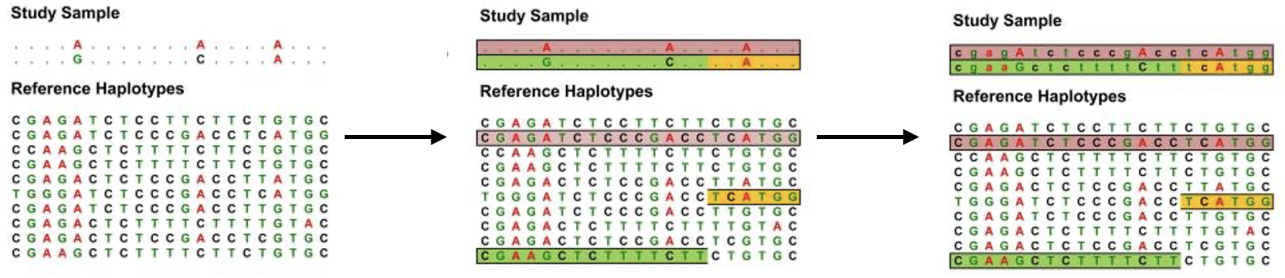
**Removal of rare and non-Hardy–Weinberg equilibrium SNPs** Even though low-frequency SNPs may represent a significant and understudied component, they are hard to detect as they are often specific to individual populations or families [67]. Rare or low-frequency SNPs are removed from the analysis. Usually, there are not many low-frequency alleles for traditional statistical tests, and they lead to a higher probability of type I errors [68]. Usually, MAF thresholds of 0.01 and 0.05 are used to exclude rare SNPs [61]. Additionally, to avoid type I errors, the Hardy-Weinberg equilibrium (HWE) is tested. HWE is a population genetic principle that assumes that genotype frequencies remain constant throughout generations in a random mating population [69]. If we take any dataset of a random mating population, it should not deviate from the HWE. Otherwise, this may be caused by genotyping errors [70]. Common used thresholds to exclude variants vary between HWE p-value  $< 1 \times 10^{-10}$  (cases) or HWE p-value  $< 1 \times 10^{-6}$  (controls). Stringent thresholds may cause the removal of phenotype-associated SNPs, as the deviation from HWE can also be the results of a true genetic association. Therefore, less strict thresholds are normally used for cases than for controls [61].

### 2.1.3 Genome imputation

Genome imputation refers to predicting genotypes that were not directly genotyped, as illustrated in Figure 2.1. This has been used extensively in GWASs to enhance analysis power for fine-mapping or to help in combining and comparing studies using meta-analysis. Genome imputation can be done across the whole genome or in a specific region we are interested [71].

There are multiple methods for imputing genotypes, such as IMPUTE (v1 and v2), BEAGLE, and MACH. I will not go through an exhaustive explanation of each method, but rather the one used in this PhD thesis: SHAPEIT2-IMPUTE2 (paper II, chapter 5). The human reference genome used for imputation was 1000 Genomes Project phase 3.

**SHAPEIT for haplotype estimation** Haplotype refers to a set of SNPs along a chromosome that tends to be inherited together due to the short distance between them. These SNPs on the same haplotype block are in linkage disequilibrium (LD), and if one of the SNPs carry a specific allele, we can often predict the alleles carried on the SNPs in the same block [73][74]. This is known as haplotype estimation or phasing. It can be used for other purposes since some diseases are associated with a specific haplotype, i.e., an exact allele arrangement found on each copy of homologous chromosomes can influence



**Figure 2.1** | Genotype imputation workflow in two samples, where regions shared between the study samples and samples in the reference genome are identified and this shared information is combined to fill in the missing genotype of the study samples. Figure adapted from [72].

the gene expression of a disease-associated gene. In standard genetic studies, this will not be detected, since phasing usually is not done in these cases, perhaps because it will add complexity in the analysis [75].

In this PhD thesis, SHAPEIT v2 (r790) (Segment HAPlotype Estimation and Imputation Tool) was used for haplotype estimation. Before phasing, the dataset was split by chromosome using PLINK2. In SHAPEIT2, haplotypes are inferred using a Gibbs sampling approach. At the core of this approach is a Hidden Markov model used to linearly model the conditional distributions of the Gibbs sampler, where an individual's haplotypes are updated iteratively based on the haplotype estimates of all other samples [76][77].

**IMPUTE2 for imputation** Once phasing is done, we can infer missing genotypes by performing imputation. IMPUTE2 also performs phasing; however, IMPUTE2 authors recommend using SHAPEIT2 for it, followed by IMPUTE2 for imputation [78]. To reduce computation demand, each chromosome is split into several segments or chunks of 5000 kilobases, which can be merged in the end once imputation is performed.

IMPUTE2 uses a Markov chain Monte Carlo (MCMC) method to integrate all possible haplotypes from the phasing step and predict the alleles of missing SNPs. As a standard procedure, 30 MCMC iterations are performed in 500 reference haplotypes [79]. These probabilities are then averaged across iterations and produce a marginal posterior genotype probability at each imputed SNP. Further, IMPUTE2 reports an imputation quality score, known as INFO score, for each SNP based on posterior genotype probabilities. This INFO score ranges between 0 and 1, where 1 means the highest certainty. The INFO score is used to filter SNPs with low imputation accuracy [80].

#### 2.1.4 Disease and single-nucleotide polymorphism association

Once quality control (and sometimes genotype imputation) is performed, all genotyped variants are tested from association with the phenotype under study. The test to be used



depends on what we want to investigate, and sometimes we do not know at the beginning what are we looking for. Instead of trying one specific test, we rather do an exploratory analysis using several tests.

In this PhD thesis, PLINK was used to perform the association tests. The most basic one is the allele test chi-square (binary traits) or Wald (quantitative traits), which compares the allele frequency or counts between cases and controls. PLINK also uses other alternative association analysis, such as Fisher's exact test, genetic models (dominant, recessive, and genotypic), stratified analysis when clusters have been specified, and logistic and linear models with the possibility of adding possible covariates. As thousands of SNPs are tested, this will cause the inflation of type I error; thus, adjustment for multiple testing, using Bonferroni, Sidak or false discovery rate (FDR), is also an option [52].

A "diagnostic plot" widely used is the quantile-quantile (Q-Q) plot of the observed vs expected p-values on a log 10 scale. This plot indicates if there is a deviation from HWE, i.e. if the study has generated more significant results than expected by chance. A large inflation factor, or lambda, will be obtained in this case. This can happen due to population stratification and relatedness between samples [81]. To facilitate the visualisation of the analysis, a Manhattan plot is usually used to check the results from a single-locus association analysis. In a Manhattan plot, one can easier visualise which SNPs passed the defined thresholds for genome-wide significance.

## 2.2 Next-generation sequencing

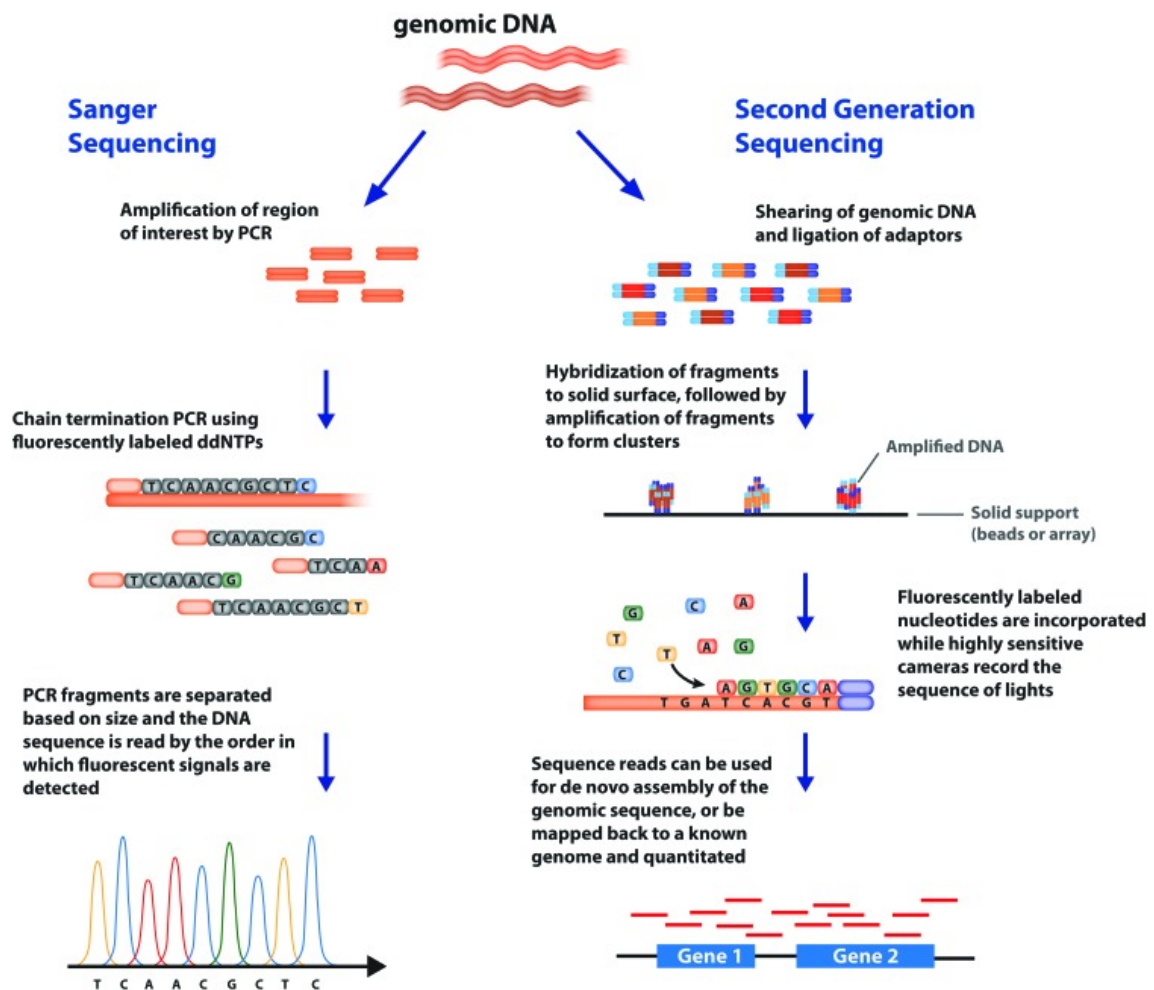
NGS revolutionised the field of genomics. It describes high-throughput DNA sequencing technologies that now dominate the DNA sequencing field, taking the place of the previous gold-standard Sanger sequencing.

NGS allows whole-genome sequencing (WGS), or parts of the genome, i.e., whole-exome or targeted sequencing [82]. Sanger is considered the first-generation; thus, NGS is known as second-generation sequencing. NGS is also known as massive parallel sequencing due to its advantage of analysing millions of DNA strands in parallel and producing large volumes of data [83][84].

**The past:** Sanger sequencing is a targeted technique that uses oligonucleotides primers to seek specific DNA regions. After DNA amplification, the double-stranded DNA is denatured using heat, and the primers bind to the 5' end of the single-stranded DNA. Next, this primed DNA is dispersed in four reaction vessels and DNA polymerase, four deoxynucleotide triphosphates (dNTPs): adenine, cytosine, thymine, and guanine, and chain-terminating ddNTPs for each nucleotide is added in each vessel. The single-stranded DNA is elongated by the dNTPs until a ddNTP binds since ddNTP lacks a hydroxyl group at the 3' carbon. Each ddNTP contains a unique fluorescent label, so a laser recognises

this fluorescent signal on the automated machine, which detects the fluorescent intensity, translated into a “peak”. As dNTPs and ddNTPs have the same probability of binding to the sequence, the sequences will have different lengths. In the end, polyacrylamide gel electrophoresis is used to get the complementary sequence of the DNA sample. Sanger sequencing is nevertheless costly compared to NGS, and it can only sequence short regions of DNA each time [85].

In Figure 2.2, we can visualise some of the main differences between these sequencing technologies.



**Figure 2.2** | Illustration comparing Sanger sequencing and second-generation sequencing. Figure from [86].

In this PhD thesis, Illumina technology has been used for NGS in paper II, chapter 5 and in additional work for paper IV, chapter 7. The following sub-sections are focused on the three main steps of NGS analysis: 1) template preparation; 2) DNA sequencing; and 3) data analysis.

### 2.2.1 Template preparation

The first step for NGS is template preparation which consists of library preparation. Libraries are created by the fragmentation of DNA in smaller fragments, i.e., between 150 to 800 base pairs, depending on the platform used [87]. Following DNA fragmentation, DNA fragments are end-repaired and A-tailed. The A-tail allows adapters to bind. Before sequencing, DNA fragments are amplified through bridge amplification.

### 2.2.2 DNA sequencing

NGS Illumina uses sequencing by synthesis approach. In sequencing by synthesis, there is the extension of the sequencing primer. A fluorescent-label nucleotide competes to be added to the growing strand in each sequencing cycle. Once a labelled dNTP is added to the nucleic acid chain, a fluorescent signal is emitted. The emission wavelength along with the signal intensity determines the base call. The number of sequencing cycles determines the length of the read [88].

### 2.2.3 Data analysis

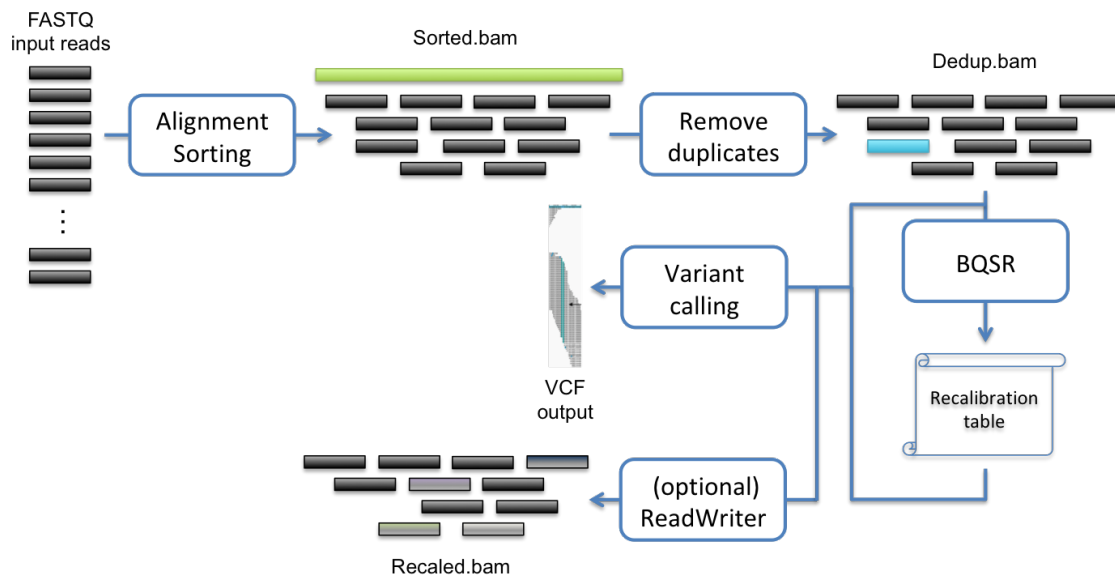
Once the samples are run through the sequencer machine, data is stored in FASTQ format files. This is a standard file format used for sequencing data containing both the sequence and the corresponding per base quality score or Phred quality score encoded as ASCII characters (human-readable) [89].

The raw sequence data needs to go through several steps until the final output is generated. A standard data analysis pipeline for NGS includes remove adapter sequences and low-quality reads, align the data to a reference sequence or construct a genome from multiple DNA fragments via de novo assembly, and lastly, analyse the compiled sequences. Genome Analysis Toolkit (GATK) from the Broad Institute [90][91] is widely used for analysing NGS data. Still, other alternatives provide faster variant calling, such as Sentieon DNASEq (pipeline illustrated in Figure 2.3).

In this PhD thesis, Sentieon version 201808.03 was used for germline variant calling in WGS (paper II, chapter 5) and targeted sequencing (addition to paper IV, chapter 7). Pipeline steps are described below. GRCh37 was used as the reference genome.

**Short read quality assessment using FastQC (v0.11.2)** FastQC is used to assess the overall quality of the sequencing run and adapter contamination [92].

**Adapter removal using AdapterRemoval (v2.1.3)** If adapter contamination is found, adapters are removed as they can interfere with the correct mapping of the read to a reference genome and influence downstream analysis. AdapterRemoval is a widely used



**Figure 2.3** | Sentieon DNaseq pipeline. Figure from Sentieon manual support.

tool for the task and it can pre-process single and paired-end data in FASTQ format. The adapter sequence is specified (`--adapter1` and `--adapter2`) or the default Illumina TruSeq adapters are used [93].

The following described steps were executed using Sentieon version 201808.03.

**Mapping reads with BWA-MEM** Burrows-Wheeler Alignment uses the maximal exact matches (BWA-MEM) algorithm to align reads against the reference genome. It finds at each query position the longest exact match covering this position [94].

**Access metrics** Multiple quality control metrics are available to access the number of reads with low mapping quality after alignment to the reference genome. Here, the following algorithms and respective plots were used: `MeanQualityByCycle`, `QualDistribution`, `GCBias`, `AlignmentStat`, and `InsertSizeMetricAlgo` (description at Sentieon version 201808.03 support webpage).

**Remove duplicate reads** PCR duplicates may occur when the same DNA fragment is sequenced two or more times. These duplicates usually are removed as they may be counted as additional evidence and lead to false-positive variant calls [95]. Here, `LocusCollector` algorithm was used to collect read information and generate a score file. `Dedup` algorithm was used to remove duplicate reads or simply mark these duplicates without removing them (if `--rmdup` is not set).

**Indel realignment** During alignment of each read to the reference genome, alignments artefacts can arise. It is well known that indels close to the end of the reads are difficult to align to the reference genome [96]. Local realignment of these problematic regions

against a known reference set of indels helps mitigate false discoveries related to indel regions. Here, given a VCF file containing known indels, the Realigner algorithm was used to perform indel realignment.

**Base Quality Score Recalibration (BQSR)** Base quality scores are per-base estimated of error emitted by the sequencing machines. These scores are subject to various sources of errors. BQSR adjusts these quality scores of reads using ML algorithms. Here, the QualCal algorithm was used to calculate the recalibration table, using a database of known indels and known SNPs.

**Apply the results of BQSR (optional)** ReadWriter algorithm outputs the results of applying the BQSR to a file. This step is optional as variant callers can also perform the recalibration using the recalibrated bam and the recalibration table [97].

**Variant caller** Lastly, the Haplotype algorithm is used for variant calling. In this step, only variants that pass a specified threshold are added to the VCF file. The flags used are `--call_conf`, which determines the threshold to call a variant, and `--emit_conf`, which determines the threshold to emit a variant. VCF stands for variant call format; thus, these files only contain variant sites as the name indicates. If we are interested in getting all sites, the GVCFTyper algorithm can be used.

## 2.3 Polygenic risk score

As mentioned in the previous chapters, several complex diseases are highly polygenic. This means that hundreds or thousands of genetic variants have a cumulative effect on the disease risk and help understand the biological pathway(s) related to the phenotype. These genetic variants may have enormous clinical utility and be used to predict disease risk when combined into a PRS [98][99].

### 2.3.1 Calculate polygenic risk scores

A PRS is one value estimated from an individual's genetic predisposition to a phenotype. Standard PRSs are calculated as a weighted sum of genome-wide genotypes as shown in Equation 2.2.

$$PRS = \sum_i^n \chi_i \times \beta_i \quad (2.2)$$

In Equation 2.2,  $\chi_i$  is the allele dosage for SNP  $i$  where  $i \in 0, 1, 2$ , for 0, 1 and 2 alleles, respectively; and  $\beta_i$  is the effect size of SNP  $i$  estimated from the relevant GWAS data

[99]. When calculating PRSs on a binary trait, the effect sizes are normally reported as log odds ratios ( $\log(\text{ORs})$ ); and when calculating in a continuous trait, standardized mean differences normally are used [100].

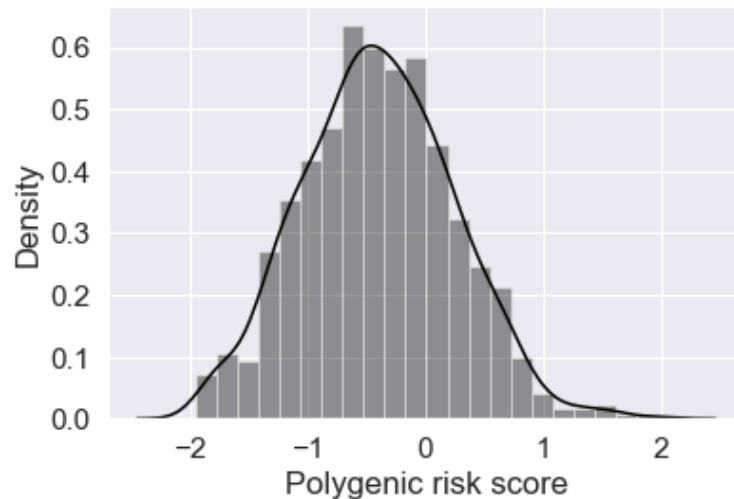
One challenge of building a PRS is to decide which SNPs to include, as not all of them influence the phenotype under study [101]. A widely used approach to calculating these standard PRSs is LD-clumping and p-value thresholding. In LD-clumping, SNPs are LD-pruned before building the PRSs to avoid redundant correlated effects between SNPs. In p-value thresholding, only SNPs that passed a pre-determined threshold are retained. The predictive performance of a PRS can be tested at different p-value thresholds, or one can define a single “hard-threshold” to retain the SNPs. LD-clumping can be performed before or after p-value thresholding. It is common to perform LD-clumping after p-value thresholding, so if there are two correlated SNPs (for example,  $r^2$  threshold of 0.25), the one with the lowest p-value for association is kept [102][103].

In the end, PRSs can be standardised to facilitate its interpretation and conversion of an individual’s PRS to quantiles [104].

As a sum of SNPs with identical distributions, PRSs should look like a normal Gaussian distribution (Figure 2.4). The majority of people will find their scores to be in the middle, while others will be on the left (lower quantile) or right (upper quantile) tail ends, which indicates a low or high risk, respectively, of developing the phenotype under study. When studying cancer, people in the right tail may benefit from discussing preventive treatments with their clinicians. It is important to remember that PRSs only provide a relative risk by comparing a person’s risk based on the genetic constitution and not providing information on the disease progression. On the other hand, absolute risk shows the likelihood of a disease occurring without any comparison to any groups of people. Giving a concrete example, if we have two people with the same PRS, and one is 20 years old while the other is 90 years old, they will most probably have different lifetimes risks.

In paper II, chapter 5, I have calculated PRSs in two childhood cancer cohorts. These PRSs were based on two recently published GWASs, for adult breast [105] and colon cancers [106]. Due to pleiotropic effects and the presence of better established and validated PRSs on the genetic predisposition of adult cancers, as opposed to childhood cancers, we have evaluated these for stratified disease prevention in childhood cancer.

In paper IV, chapter 7, PRSice software [107] was used to calculate the PRSs. Effect sizes were estimated from the performed GWAS. SNPs that passed a Bonferroni corrected threshold of  $8.02 \times 10^{-8}$ , and a set of SNPs in the same gene that passed a suggestive threshold of  $1 \times 10^{-5}$ , were included.



**Figure 2.4** | Illustration representing normal distribution of polygenic risk scores.

### 2.3.2 PRS clinical application

There is a growing interest in the clinical implementation of PRSs to measure disease predisposition, improve diseases diagnosis, and select the best treatment for each patient [108].

As an example, if we develop a PRS for coronary artery disease risk and if we find a group of patients in the upper quantile, a preventive strategy can be adopted, i.e., these patients can adopt a healthy diet and physical exercise at an early stage [98].

At the time of writing this PhD thesis, there were 26 studies found for “polygenic risk score” at ClinicalTrials.gov, either complete or active. These aim to validate several PRSs in multiple diseases, such as breast cancer, coronary artery disease, type 2 diabetes, schizophrenia or ovarian cancer [accessed on 05-02-2021].

PRS is indeed a relatively cheap and non-invasive “procedure”, as it can be calculated from a saliva sample using genotyping technologies that are becoming cheaper [99]. As research in the field shows that PRSs may have a potential benefit in patient care and help on clinical decision making, there is also a rising interest in incorporating genomic data into electronic health records. Several groups have been working together on this, such as Electronic Medical Records and Genomics (eMERGE) Network, funded in 2007, and the Clinical Genome Resource (ClinGen) project [109].

There are already PRS commercially available, for example, for breast cancer risk (Myriad Genetics) or for type 2 diabetes risk (23andMe) [110]. While some argue that PRSs are still not ready to be implemented in a clinical setup [111], and others defend that we should start its implementation [112][113], there are still some challenges that need to be addressed. There is a need of having a more standardised protocol on how to present genomics information and how to translate genomic variants into relevant phenotypes that

medical doctors can understand and use for patient care [109]. Additionally, more research is needed in other non-European populations. The majority of genomic studies done until now have included individuals of European ancestry, bringing a limitation for its use worldwide [114].

## 2.4 Machine learning

ML is “an artificial intelligence (AI) technique that can be used to design and train software algorithms to learn from and act on data” (FDA definition). ML and AI are terms used interchangeable many times, and while ML methods are AI, not all AI is ML. AI is a broader field and it is described as “the science and engineering of making intelligent machines” [115][116]. One example of AI without ML is the chatbots, which can answer a limited number of questions that a human previously fed.

In this PhD thesis, I will be only focusing on ML algorithms.

While previously described PRSs focus on genetic variants only, ML allows integrating different data types handling multidimensional data [117]. ML algorithms have been previously applied in a few GWASs [118][119][120][121], as they offer an opportunity to find complex relationships between various genetic factors, which GWASs alone cannot uncover. This is essential to make biological sense of the data and understand highly complex biological systems and disease mechanisms to improve diagnosis and treatments [122]. I have worked with ML to integrate clinical and genomics data, either independent SNPs or combined into a PRS.

### 2.4.1 Data encoding

In most ML models, we need to encode the different features as numerical variables. For clinical data, if the feature is continuous, we can use the absolute value. If the feature is categorical without an inherent order associated with it, we can use a one-hot encoding, generating one binary variable for each category. Furthermore, in regression and artificial neural networks (ANN), there is a need for feature scaling if we have features with different ranges; otherwise, features with larger values will be treated as more important while training the model. As for random forest, each feature is evaluated independently, there is no need for feature scaling.

For the genomic biomarkers, if we assume an additive genetic effect, we can encode each variant as 0, 1, or 2, for homozygous for the reference allele, heterozygous, or homozygous for the alternative allele, respectively; or use one-hot encoding, where we create three different variables for each of the categories.



Biological data have high inherited complexity, being highly heterogeneous and consequently very noisy [123]. Also, we usually are faced with the “curse of dimensionality”, which means that the number of variables is much higher when comparing with the number of samples in the study, leading to data sparsity, multicollinearity, multiple testing, and overfitting [124]. Thus, it is crucial to perform feature selection, hyperparameters tuning, have a cross-validation setup, perform randomisation, and when possible, have a new completely independent cohort where we deploy our model in the end. Ideally, and if possible, an external dataset. These concepts are discussed below.

### 2.4.2 Feature selection

Feature selection allows filtering for non-relevant and correlated features in the dataset. This is important not only to speed up learning but also to avoid overfitting and low-performance models [125][126].

Three main feature selection methods are filter, wrapper, and embedded methods [127]. Filter methods look at each feature and use univariate or multivariate analysis to remove irrelevant or high correlated features. Some examples are t-test, Relief-based algorithms and correlation-based feature selection (Pearson or Spearman correlation). Embedded methods select optimal feature subsets to build a suitable classification model. Some examples are least absolute shrinkage and selection operator (Lasso) and Ridge regression. Wrapper methods use predictive models to evaluate selected features in a training-hold-out set. These are more computationally expensive than the filter and embedded methods, and some examples are forward or backwards feature elimination [128][129].

### 2.4.3 Cross-validation and hyperparameters tuning

In the models developed in papers III, chapter 6 and IV, chapter 7, a nested cross-validation setup was used, so feature selection and hyperparameters tuning could be performed in the internal cross-validation to avoid overfitting to the test set.

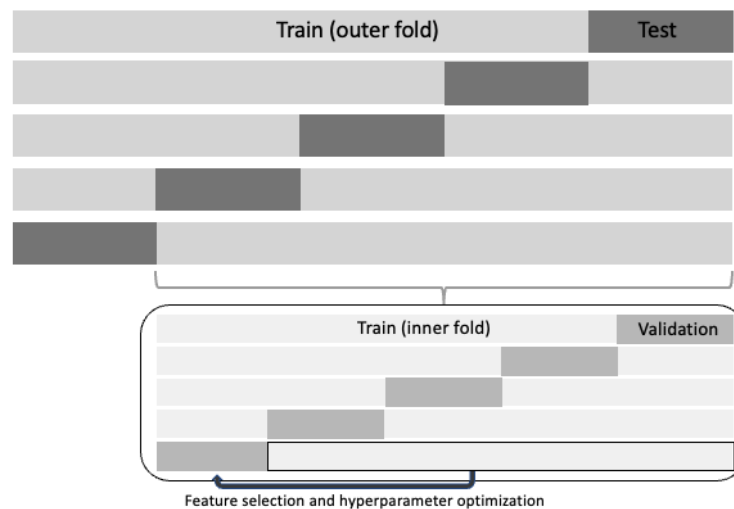
In a standard K-fold cross-validation, the dataset is split into K folds. Each K fold is used as a test set, and the other folds ( $K - 1$ ) are used as the training set to build the model. A total of K models are fit and evaluated in each test set, and the mean performance metrics calculated on these test sets are reported [130]. The number of K's should be a balance between the number of folds we choose to perform with the size of the dataset we have available. If we expect 100 samples to be enough to test our model (with a balanced number of cases and controls), we can choose a 5-outer fold if we have 500 samples available or a 10-outer fold if we have 1000 samples available.

Each model includes one or more hyperparameters tuned to maximise the models' predictive performance. There is no perfect way to configure these hyperparameters, and

usually, different configurations are tested while training a ML model, using a grid-search or random search strategies. Some examples of hyperparameters that are optimised are the number of trees or estimators in the random forest or the number of hidden layers in an ANN [131].

If we use the standard K-fold cross-validation to tune and select a model, we will most probably overfit to the test set. One way of overcoming this is to apply nested cross-validation. Nested cross-validation allows performing hyperparameter tuning and feature selection to train an optimal prediction model. Nested cross-validation is more computationally expensive than a standard K-fold cross-validation, as it increases the number of model evaluations to be performed [132].

In the nested cross-validation, the dataset is split in  $K$  outer folds, and each  $K - 1$  inner-fold is further divided into inner-training set and validation set. Hyperparameter tuning and feature selection are fully made on the inner-fold, using the training set and the validation set to evaluate the different combinations of hyperparameters and features. In the end, a completely unbiased model is deployed in the test set, avoiding any data leakage while training the model and thus avoiding overfitting, as shown in Figure 2.5.



**Figure 2.5** | Example of 5,5-outer, inner nested cross-validation setup.

#### 2.4.4 Model selection and training

ML models can be either supervised or unsupervised. In supervised learning, the labels or output is known, while in unsupervised learning, we don't have the labels or output available. For the supervised model, the training data is analysed and the function produced can be used to classify new samples. For the unsupervised, the idea is to unravel hidden signals within the data by detecting clusters. If a new sample is available, it can be assigned to the closest cluster. Apart from the two main groups, there is also semi-supervised learning, which combines both labelled and unlabelled data. Semi-supervised

learning algorithms learn from partially labelled data and are mainly used if it is expensive and very time-consuming to label all data available [133].

In this PhD thesis, only supervised learning models were used.

### Linear regression

A linear regression model assumes a consequent one-unit change in the outcome for each one-unit change in the variable. Linear regression assumes linearity, normality, and homoscedasticity. Violation of these assumptions lead to type I and type II errors.

Linear regression makes a prediction  $\hat{y}$  estimating a weighted ( $\theta$ ) sum of  $n$  input features ( $x$ ), and adding a constant called the bias or the intercept term (Equation 2.3).

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n \quad (2.3)$$

The values of  $\theta$  are estimated so that the model best fits the training set, i.e., find a value of  $\theta$  that minimises the residual sum of squares (RSS) (Equation 2.4).

$$RSS = \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (2.4)$$

### Logistic regression

Regression algorithms can also be used for classification by assuming a linear dependence between the variable (independent variables) and the logit of the outcome (dependent variable). Like linear regression, logistic regression also estimates a weighted sum of the input features. Instead of outputting the result directly, the output is transformed using the logistic sigmoid function (Equation 2.5).

$$\sigma(\hat{y}) = \frac{1}{1 + e^{-(\theta_0 + \sum_{i=1}^n \theta_i x_i)}} \quad (2.5)$$

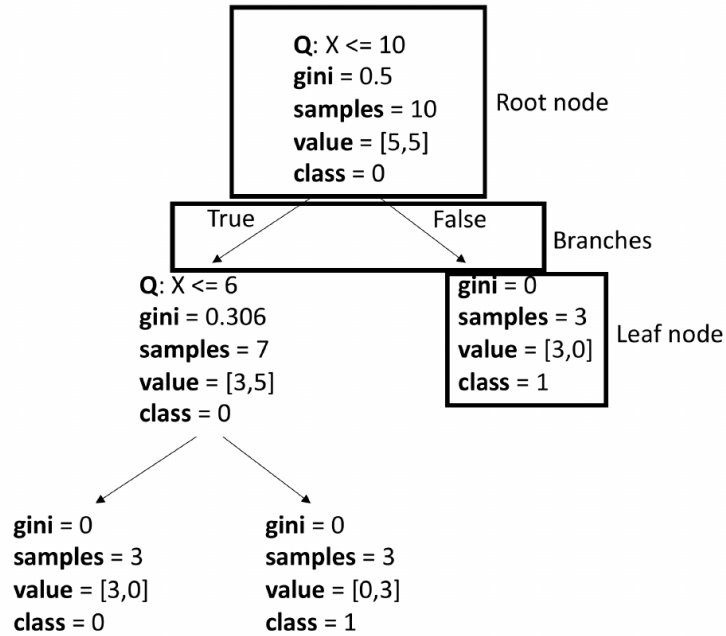
The final prediction  $\hat{y}$  can be made using Equation 2.6.

$$\hat{y} = \begin{cases} 0 & \text{if } \sigma(\hat{y}) < 0.5; \\ 1 & \text{if } \sigma(\hat{y}) \geq 0.5. \end{cases} \quad (2.6)$$

## Random forest

Random forests are an “ensemble learning” algorithm, i.e., a combination of decision tree predictors. Decision trees are made up of decision nodes, branches and leaf nodes. For each node, a question is asked and the data is divided into smaller subsets until it reaches the leaf nodes with no further divisions as shown in Figure 2.6 [134].

The two most used criterion to measure the quality of the split are Gini (Equation 2.7) and entropy (Equation 2.8) coefficients which are based on the  $C$  total number of classes and  $p$  the proportion of a class in the node. These are also used to measure the relative importance of each feature, as it estimates how much impurity was reduced in the tree nodes that used a specific feature. For each feature, its importance is calculated as the normalised total reduction of the criterion brought by that feature [135].



**Figure 2.6** | Representative example of a decision tree. Q refers to the question asked. Gini measures the quality of the split (entropy is another possibility). Samples refer to the total number of samples in the dataset. Value refers to the number of samples in each category. The class value refers to the final prediction.

$$Gini = 1 - \sum_{i=1}^C (p_i)^2 \quad (2.7)$$

$$Entropy = \sum_{i=1}^C p_i \log_2(p_i) \quad (2.8)$$

Each tree outputs a prediction for each sample, and the final prediction is the classes mode or the mean predictions, for classification and regression problems, respectively, of all individual trees. To avoid overfitting in the dataset, it is ensured that each decision tree is independent of each other using bootstrapping and feature randomness. Bootstrapping consists of random sampling with replacement from the dataset, which results in different decision trees. Feature randomness consists of considering every possible feature from a random set of all features to split a node and choose the one that results in the best separation between the different classes [136][137].

### Hyperparameters

There are multiple hyperparameters one can optimise for in the random forest to increase the model's predictive power. I will indicate some of the most common optimised hyperparameters [135] and indicate the default values for these hyperparameters in the sklearn's built-in random forest function [138] in parenthesis.

1. Number of decision trees: this should be large enough to allow for each feature to be selected by the model, or until predictive performance reaches a plateau, but not too large to not slow down computation ( $n\_estimators=100$ ).
2. Maximum number of features: this refers to the number of features randomly considered to split a node, and while a low value gives a chance for features with small effects to be selected, a high value reduces the risk of having too many non-informative candidate features ( $max\_features=\sqrt{totalnumberof\ features}$ ).
3. Minimum number of leaf nodes: the smaller the value, the larger decision trees we will end up with ( $min\_sample\_leaf=1$ ).

### **Artificial neural network**

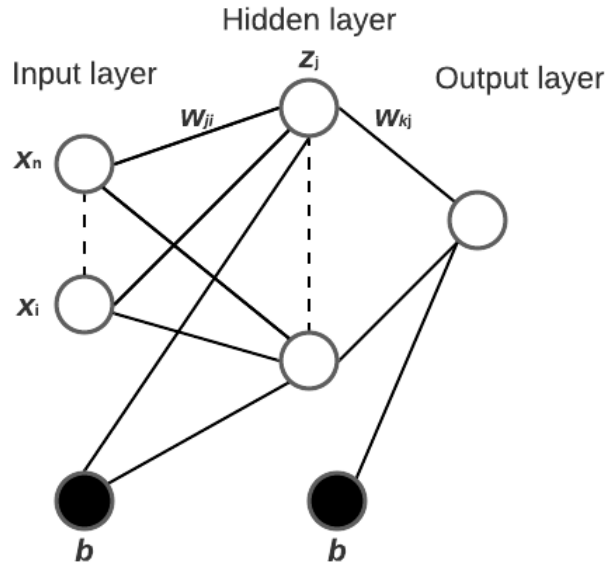
ANNs were inspired by the brain's architecture. Our brain is composed of billions of neurons that receive electrical impulses from other neurons via synapses. A neuron will release its signal, or be excited, if it gets enough signals from other neurons. The biological neural networks architecture is still the subject of much active research, but from previous studies, it seems that biological neurons are organised in consecutive layers [139]. The smallest units of ANNs are artificial neurons.

The feedforward neural network is one of the simplest ANN types and consists of one input layer, one or more hidden layers and the output layer. If an ANN contains more than two hidden layers, this is referred to as deep learning. There are other more complex ANNs, such as convolutional neural networks primarily used in image recognition; and recurrent neural networks used primarily for time series data analysis.

In this PhD thesis, feedforward neural networks were used and described further. The mathematical notations are from Bishop textbook [140] and [141].

### Feedforward neural networks

Feedforward neural networks are, as referred above, the earliest and simplest form of a neural network, where the data is fed forward from one layer to the next and finally to the output layer, computing a function  $f$  on input data  $x$ , where  $f(x) \approx y$ . The neurons are arranged in a directed and fully connected acyclic graph. The bias neuron is added to each layer in the neural network. A constant term is added to the calculation of the hidden and output neurons values, which allows the network to set an individual activation threshold for each neuron (Figure 2.7).



**Figure 2.7** | Illustration of ANN with one input layer, hidden layer and output layer.

### Training

For each training instance there is a series of functions transformation.  $M$  linear combinations of the input variables  $x_i, \dots, x_n$  are constructed, where  $j = 1, \dots, M$ , and the superscript (1) stands for the first training example of the network. Additionally  $w_{ji}^{(1)}$  stands for the weights and  $b^{(1)}$  for the biases (Equation 2.9).

$$a_j = \sum_{i=1}^n w_{ji}^{(1)} x_i + b^{(1)} \quad (2.9)$$

In Equation 2.9,  $a_j$  are known as activations. These are transformed using a differentiable and nonlinear activation function  $h$  such that:  $z_j = h(a_j)$ . These nonlinear functions  $h$  can be sigmoidal (example, logistic or tanh function). Other very common used activation

function is the rectified linear unit (ReLU). The output  $o$ , for a 1-hidden-layer ANN, is defined in Equation 2.10.

$$o = \sum_{j=1}^M w_{kj}^{(2)} z_j + b^{(2)} \quad (2.10)$$

In the end, the output unit activations are transformed to give an output  $\hat{y}$ . We can use the identity function for regression tasks, so  $\hat{y} = o$  or the logistic sigmoid function for binary tasks, as described in Equation 2.5.

### Backpropagation

The backpropagation algorithm goal is to update the weights based on the loss or error function  $E$ , allowing the information from the error function to flow backwards through the network. The error function considers the difference between the output or predicted values  $\hat{y}$  and the target value  $y$ , where  $n$  is the number of data points. Commonly used measures are cross-entropy (binary traits, Equation 2.11) and mean squared error (quantitative traits, Equation 2.12).

$$E = - \sum_i^n (y_i \log(\hat{y}_i)) \quad (2.11)$$

$$E = \frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2 \quad (2.12)$$

The error function's gradient in regards to a weight  $w$  is given by the derivative of the error function  $\frac{\partial E}{\partial w}$ . Here, the most commonly used algorithms are gradient descent, stochastic gradient descent, or adaptive moment estimation.

Finally, the weights are adjusted to minimise the error by a fixed-size amount (Equation 2.13). Here, a learning rate  $\gamma$  controls how much the weights should be adjusted with respect to the error gradient.

$$\Delta w = -\gamma \frac{\partial E}{\partial w} \quad (2.13)$$

### Hyperparameters

The flexibility of the neural network is also one of its main drawbacks, with the high amount of hyperparameters one can choose from [139]. I will give examples of some of the

most important and indicate the default values for these hyperparameters in the sklearn's built-in MLPClassifier function [138] in parenthesis.

1. Hidden layers size and neurons per hidden layer: this will be dependent on the size and complexity of the input data (*hidden\_layer\_sizes=100*).
2. Activation function: this has a big impact on the prediction and performance of the model, and while the sigmoid function performs well for classification problems, the ReLU function avoids the vanishing gradient problem, meaning that the weights will not be prevented from update its value (*activation=relu*).
3. Learning rate: if too small, the gradient descent can be slow, while if too large, it can exceed the minimum and it may fail to converge (*learning\_rate\_init=0.001*).

### 2.4.5 Model evaluation

Model evaluation is an essential part of building an effective ML model and there are several metrics used.

For continuous labels, the model can be evaluated by calculating the difference between the predicted values  $\hat{y}$ , and the observed values  $y$ . One standard measure used to access this correlation is the correlation coefficient,  $R^2$ , between observed and predicted values, where a value of 1 would mean a perfect prediction (Equation 2.14).

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y} - y)^2}{\sum_{i=1}^n (y - \bar{y})^2} \quad (2.14)$$

For binary labels, the outcome of the prediction model can be either: true positive (TP), false negative (FN), false positive (FP), or true negative (TN). We can get these numbers from the confusion matrix and use them to calculate multiple different error measures (Table 2.1).

**Table 2.1** | Confusion matrix and performance measures for categorical outcome.

		Predicted outcome		
		Predicted cases	Predicted controls	
Actual outcome	Cases	True positive (TP)	False positive (FP)	Positive predictive value = TP/(TP+FP)
	Controls	False negative (FN)	True negative (TN)	Negative predictive value = TN/(TN+FN)
		Sensitivity = TP/(TP+FN)	Specificity = (TN/TN+FP)	

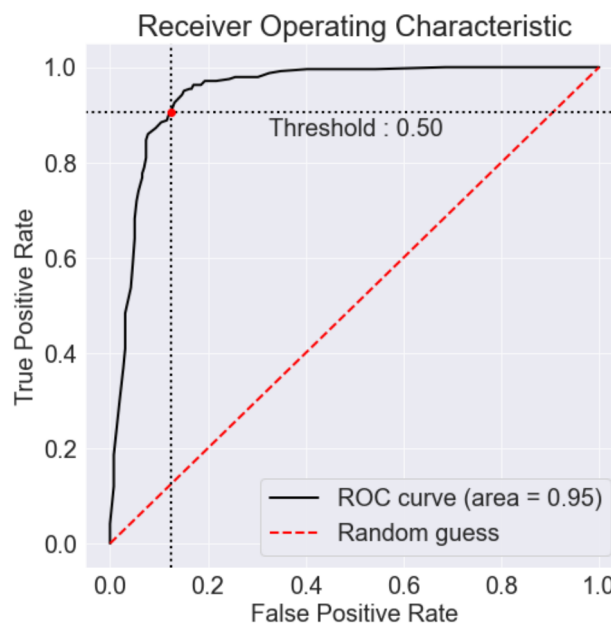


The definition ensures that all error measures in Table 2.1 have a range between 0 and 1, where for all of them, a value of 1 implies an error-free classification. These measurements look into different information from the model; thus, each is important to infer the model's utility.

Sensitivity looks only at cases and from those, which ones were correctly predicted as cases. Specificity looks only at the controls and from those, which were correctly predicted as controls. On the other hand, we have positive predictive value (PPV) and negative predictive value (NPV), which look into the predicted outcome. PPV looks into the predicted cases and from those, which ones are actual cases. NPV looks into the predicted controls and from those, which ones are real controls.

The relationship between sensitivity and specificity can be observed with a Receiver Operating Characteristic (ROC) curve, where we have the sensitivity in the y-axis and the false positive rate (1-specificity) in the x-axis (Figure 2.8). The relationship between sensitivity and specificity is visualised for different classification cutoffs; thus one efficient way of evaluating a specific model for different cutoffs is by calculating the area under the ROC curve (ROC-AUC). The ROC-AUC ranges between 0 and 1, where 1 is the best possible classifier and 0.5 means a random classification [142].

The default classification cutoff to calculate the performance measures referred to in the confusion matrix is 0.5, i.e., if the model prediction score is below 0.5, the sample will be classified as a control. Otherwise it will be classified as a case. However, other cutoffs can be accessed and selected to find a better trade-off between a high true positive rate and a low false-positive rate.



**Figure 2.8** | Illustration of ROC curve and highlight of 0.5 threshold (red dot). A model with random performance yields ROC-AUC of 0.5 (red dashed diagonal line).

## Chapter 3

# Disease risk and risk management

In this chapter, I will briefly mention how the use of the bioinformatic analysis described in chapter 2 can help evaluate phenotype predisposition or predict treatment response in both short (resistant versus sensitive) or long-term (late side effects from treatment). This is discussed in more detail in the respective papers.

### 3.1 Disease predisposition

Genetics may increase the likelihood of developing a particular disease due to DNA variations. Identifying these DNA variations and genes associated with disease has been possible due to the fast expansion of multiple bioinformatics and statistical tools. This is important to understand the mechanisms of pathogenesis and detect any risk groups early to adopt preventive strategies.

We are well aware by now that most diseases are very complex, meaning that multiple genes may be associated with disease appearance and progress. Recently, there has been a growing interest in studying genetic correlations of multiple cancers together since they, to a certain extent, have the same underlying biology. The aim is to uncover new shared genetic variants and better understand the complex biological pathways that lead to cancer. For example, a recent study had identified a link between the following cancers 1) lung and head/neck; 2) colorectal and lung; 3) breast and lung; and 4) breast and colorectal. A shared genetic basis was also found between breast and ovarian cancer [143]. Another recent study also investigated pan-cancer pleiotropy in well-defined populations and found multiple cancer pairs that showed either positive or negative genetic correlation, as well as novel pleiotropic risk variants. Many of them were enriched for regulatory elements and influenced cross-tissue gene expression [144].

PRSs allow us to put together information from multiple genetic variants, providing an individual genetic predisposition profile that may influence disease prediction and stratification in different risk groups. In a recent phenome-wide association study, it was found

that PRSs for common cancers, such as breast, prostate or melanoma, were also associated with other phenotypes [145].

The cancer burden is more challenging to quantify in children than adults; thus, in paper II, chapter 5, the underlying biology between adult breast and colon cancer and childhood cancer was investigated. We have calculated PRSs in two childhood cancer cohorts using GWAS summary statistics from previously developed breast and colon cancer PRSs. The goal was to evaluate genetic predisposition's risk on subgroups that may reflect childhood cancer aetiology. Furthermore, this could contribute to developing further downstream treatment stratification in children.

## 3.2 Treatment response

Pharmacogenetics, which studies how a person's genes influence their response to drug therapy, is a field of great interest in medicine. Multiple genes have shown to influence response to treatment in diverse diseases. While it is true that other factors are involved in drug response, such as environmental factors, diet, lifestyle, and age, it is believed that the genetic makeup of a person is the strongest indicator of drug response [146].

I have explored dasatinib resistance in T-cell acute lymphoblastic leukaemia (T-ALL). This project took place during my remote external stay at St. Jude Children's Research Hospital in Memphis, USA, under the supervision of Jiyang Ju from Computation Biology department and Jun J. Yang from Pharmaceutical Sciences department. My project was inspired by the research paper previously developed at St. Jude Children's Research Hospital [147]. This work is described in chapter 9.

## 3.3 Late-side effects

Chemotherapy is the most common treatment for cancer, together with radiation therapy. While quite effective, this is also a very aggressive treatment. After chemotherapy treatment, some patients may continue living with no problems. In contrast, others end up developing long-term side effects, which can happen months or years after treatment. These late effects vary depending on the cancer type, treatment used and amount of chemotherapy given, and other characteristics of the patient, such as age, gender or lifestyle factors. They can also vary from mild to severe.

Methods to identify patients at high risk of developing chemotherapy-based toxic effects are of great interest as they allow the development of targeted therapies, which can reduce patients' distress and costs of treatment-related hospitalisations [148]. In recent years, researchers have been working on ML-based models to predict treatment toxicity. However,

the application of these models in the clinic is still far from being implemented, partly due to their low interpretability and the lack of communication between data analysts and clinicians [149].

In paper III, chapter 6 and paper IV, chapter 7, I will describe two prediction models to classify testicular cancer survivors at high risk of developing hearing loss and nephrotoxicity, respectively, after cisplatin-based chemotherapy. This work has been done in close collaboration with clinicians at Rigshospitalet, Copenhagen, to better understand the needs in the clinic.

### 3.4 Translating cancer genomics into precision medicine

With genome sequencing came the promise that we would understand disease biology much better. Indeed, tremendous progress has been made in mapping genes to their function and relating the molecular pathology of monogenic diseases to the respective phenotype. However, we still do not understand all gene functions. Many disease markers are still uncovered, as most clinical phenotypes are complex, meaning that multiple genes and environmental factors contribute to it. Other reasons are: 1) an inadequate description of the clinical phenotype; 2) the high biological complexity; and 3) not enough data available, the well-known “large  $p$ , small  $n$ ” scenario in most patient-data, i.e., the total number of predictors  $p$  is usually much larger than the sample size  $n$  [150].

We now have available several methods that help deal with some of the described challenges. GWASs made it possible to identify several risk genetic variants for complex diseases in specific populations. NGS allowed the sequencing of the entire genome, thus, increasing the probability of finding the gene(s) associated with a disease.

To increase our genomic understanding and develop personalised medical healthcare using disease risk prediction models, PRSs and ML are the primary selected methods [151][152]. They aim to improve clinical decision-making for each patient using a risk stratification-based approach; thus, communication with clinicians is an essential part of the process to understand what is needed.

These prediction models are far from perfect, and it is essential to state the model’s limitations and uncertainty concerning patient classification in risk groups. Additionally, there are other few challenges. These include 1) the use of performance metrics that do not demonstrate the model’s clinical relevance and clinicians are not familiar with, such as ROC-AUC - while ROC-AUC is a good overall performance measure, it is essential to state other performance metrics that can capture different properties of a model, such as sensitivity, specificity, PPV and NPV; 2) difficulty in comparing algorithms as each study reports different methods applied on populations with different characteristics; 3) logistics related to the implementation of these models in the clinic and difficulty in combining

data in different formats such as personal doctor's notes or electronic health records; and 4) workforce to keep improving and updating models throughout time [153][154][155]. So, though many of these prediction models end up not being implemented in the clinic, we are moving that way. Being aware of these challenges is the first step to enable a model's safe clinical deployment, making a better and more responsible decision.

In the next part of the thesis, I will be presenting the different work developed throughout the PhD. Each paper uses a stratification-based approach, and its potential clinical applicability is discussed, including limitations and opportunities.

## Part II

## Papers

In this PhD thesis, I explored different sides of precision medicine and its potential application in a clinical setting. I had close contact with 1) Ayurveda researchers and Ayurveda clinician (paper I); 2) as well as clinicians practising evidence-based medicine in Denmark (paper II, III, and IV).

### **Paper I: under review at Journal of Traditional and Complementary Medicine**

In paper I, I present a genome-wide case-control study of rheumatoid arthritis in Indian individuals. The concept of Ayurveda medicine is presented. The goal was to identify new or constitution-specific rheumatoid arthritis biomarkers using Ayurveda-based deep phenotyping into vata, pitta, and kapha predominant groups, and at the same time, scientifically validate the principles of Ayurveda.

### **Paper II: in preparation**

In paper II, I have calculated genetic PRSs in two childhood cancer cohorts. These PRSs were based on published GWAS summary statistics previously used to estimate PRSs in adults of European ancestry with colon and breast cancer. There have been various studies showing shared biological mechanisms between multiple cancers. Furthermore, as adult cancers have better established PRSs than children, the goal was to identify risk groups that reflect downstream treatment stratification and prognosis in childhood cancer using the adults-based PRSs.

This study summarises current results. This is part of an ongoing project and could not be submitted before handing in this thesis. My main contribution consisted on data analysis, results interpretation and manuscript draft. Parts of data analysis previously done and not performed by me consisted of 1) quality control (NOPHO cohort) by MH and RLN; and 2) Sentieon analysis (STAGING cohort) by AOL and JV. In 2), I have run Sentieon pipeline for gVCF calling (last step).

### **Paper III: submitted to JAMA Oncology**

In paper III, I have used a cohort of Danish testicular cancer survivors and have developed a logistic regression model to classify patients into high or low risk of developing hearing loss as part of ototoxicity, a common side-effect of cisplatin-based chemotherapy. Hearing loss, as reported here, is not an objective medical measurement but rather a self-reported measurement. Other more complex machine learning models, i.e., random forests and artificial neural networks were tried, and logistic regression shown to be the most effective.

### **Paper IV: published in JNCI Cancer Spectrum**

In paper IV, using the same patient data as in paper III, I have explored nephrotoxicity development in testicular cancer survivors, another common late side-effect from cisplatin treatment. I have also discussed the key genes likely to affect the development of kidney damage.

## Chapter 4

### Paper I: Ayurveda GWAS

Garima Juyal, Anuj Pandey, Sara L Garcia, Sapna Negi,  
Ramneek Gupta, Uma Kumar, Bheema Bhat, Ramesh C Juyal,  
Thelma B K

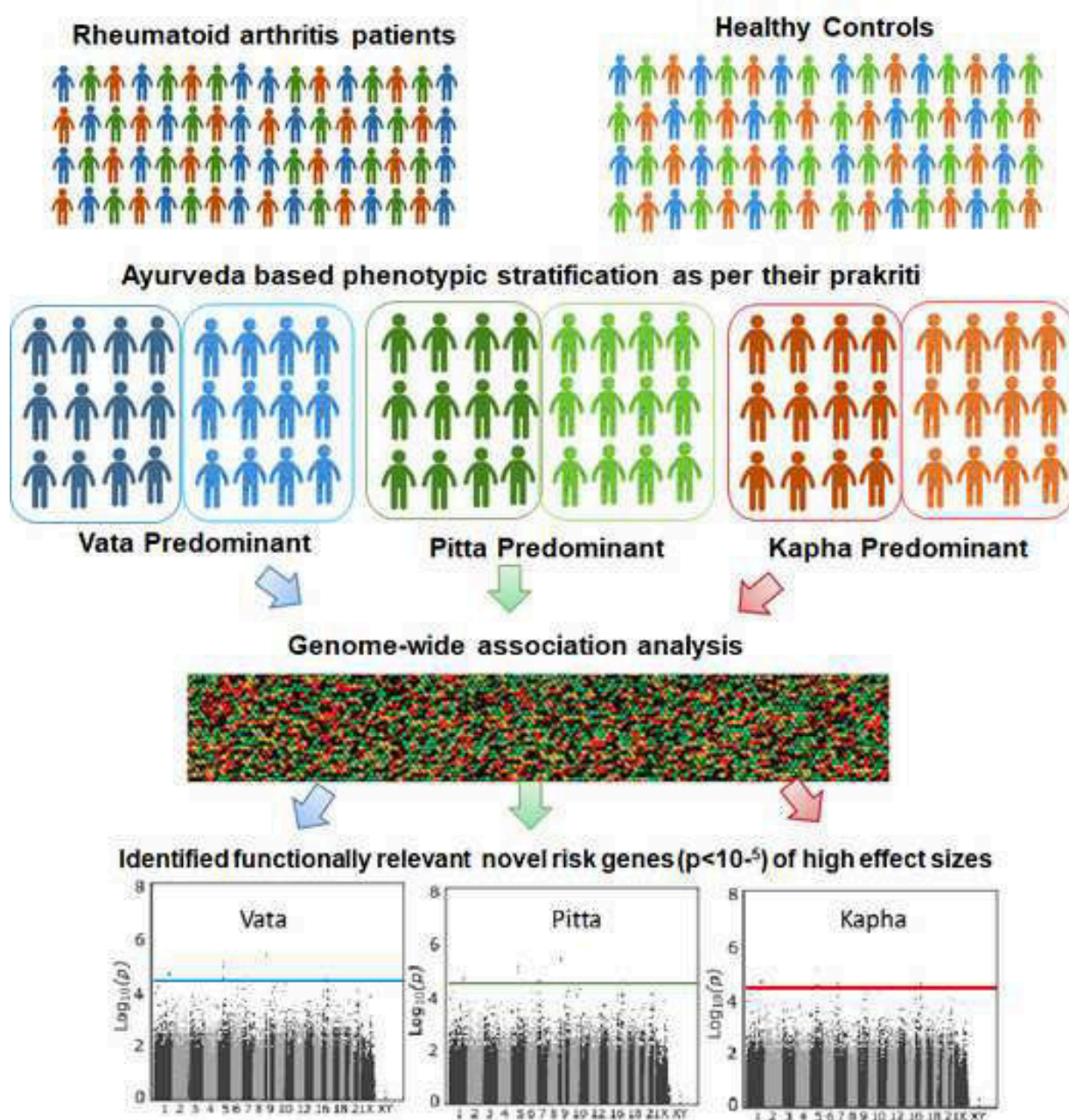
**Stratification of Rheumatoid Arthritis Cohort Using  
Ayurveda Based Deep Phenotyping Approach Identifies  
Novel Genes in a GWAS.**



Graphical Abstract

[Click here to access/download;Graphical Abstract;Graphical abstract final.tif](#)

Pasted Layer



## **Stratification of Rheumatoid Arthritis Cohort Using Ayurveda Based Deep Phenotyping Approach Identifies Novel Genes in a GWAS**

**Authors:** Garima Juyal<sup>1\*</sup>, Anuj Pandey<sup>2§</sup>, Sara L Garcia<sup>3§</sup>, Sapna Negi<sup>4</sup>, Ramneek Gupta<sup>3</sup>, Uma Kumar<sup>5</sup>, Bheema Bhat<sup>6</sup>, Ramesh C Juyal<sup>7\*\*</sup>, Thelma B K<sup>2\*</sup>

### **Author Affiliations**

1. School of Biotechnology, Jawaharlal Nehru University, New Delhi-110067, India
2. Department of Genetics, University of Delhi South Campus, New Delhi-110021, India
3. Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark
4. National Institute of Pathology, Safdarjung Hospital Campus, New Delhi-110029, India
5. Department of Rheumatology, All India Institute of Medical Sciences, New Delhi-110029, India
6. Department of Ayurveda, Holy Family Hospital, New Delhi-110025, India
7. National Institute of Immunology, New Delhi-110067, India

### **\*Co-corresponding authors**

Dr. Garima Juyal, School of Biotechnology, Jawaharlal Nehru University, New Delhi-110067, India;  
garimajuyal@gmail.com

Prof. B K Thelma, Department of Genetics, University of Delhi South Campus, Benito Juarez Road,  
New Delhi-110021, India; thelmabk@south.du.ac.in

\*\*Affiliation at the time of study

§These authors contributed equally

### **Footnotes**

None

### **Conflict of Interest statement**

The authors declare that they have no conflict of interest

**List of Abbreviations**

Genome-Wide Association Study (GWAS), Rheumatoid arthritis (RA), Predictive, Preventive, Personalized and Participatory medicine (P4 medicine), Single Nucleotide Polymorphisms (SNPs); Thioredoxin (Trx); Cullin E3 ubiquitin ligase (Cul3); long intergenic non-coding RNA (lncRNA)

**Keywords**

Rheumatoid arthritis, Ayurgenomics, Ayurveda, Genome-wide association study, P4 medicine, Prakriti

**Highlights of the findings and novelties**

- i) Ayurveda based deep phenotyping of study cohort for prakriti determination
- ii) First genome-wide association study using prakriti matched cases and controls
- iii) >80% power for detecting true associations in all three subgroups
- iv) Identification of novel risk genes of high effect sizes in different RA sub-groups
- v) Ayurgenomics may propel momentum in personalized medicine

**Type of Article**

Original Research article

**Length of the Manuscript**

**Title:** 109 characters; **Abstract:** 248 words; **Text:** 5953 words; **References:** 50; **Figures and Tables:** 2 and 1; **Supplementary files:** 2

**Section:** History, Philosophy and Social-Cultural aspects of Traditional Medicine

**Taxonomy (classification by EVISE):** the methodology.

## Abstract

**Background and Aim:** Genome wide association studies have scaled up both in terms of sample size and range of complex disorders investigated, but these have explained relatively little phenotypic variance. Of the several reasons, phenotypic heterogeneity seems to be a likely contributor for missing out genetic associations of large effects. Ayurveda, the traditional Indian system of medicine is one such tool which adopts a holistic deep phenotyping approach and classifies individuals based on their body constitution/prakriti. We hypothesized that Ayurveda based phenotypic stratification of healthy and diseased individuals will allow us to achieve much desired homogeneous cohorts which would facilitate detection of genetic association of large effects. In this proof of concept study, we performed a genome wide association testing of clinically diagnosed rheumatoid arthritis patients and healthy controls, who were re-phenotyped into Vata, Pitta and Kapha predominant prakriti sub-groups.

**Experimental Procedure:** Genotypes of rheumatoid arthritis cases (Vata=49; Pitta=117; Kapha=78) and controls (Vata=33; Pitta=175; Kapha=85) were retrieved from the total genotype data, used in a recent genome-wide association study performed in our laboratory. A total of 528461 SNPs were included after quality control. Prakriti-wise genome-wide association analysis was employed.

**Results and Conclusion:** This study identified (i) prakriti-specific novel disease risk genes of high effect sizes; (ii) putative candidates of novel therapeutic potential; and (iii) a good correlation between genetic findings and clinical knowledge in Ayurveda. Adopting Ayurveda based deep phenotyping may facilitate explaining hitherto undiscovered heritability in complex traits and may propel much needed progress in personalized medicine.

## Background

The inherent goal of Predictive, Preventive, Personalized and Participatory (P4) Medicine is to shift the paradigm in medicine from reactive and generalized to proactive and personalized and hence from disease to wellness. This transformation in healthcare can be achieved by (i) predicting an individual's predisposition to a disease; (ii) stratifying patients to facilitate potential personalized nutritional and drug treatment strategies; (iii) reducing adverse drug reactions; (iv) identifying new druggable targets and their development; and (v) reducing the time, cost, and also failure rate of clinical trials for new therapies. Past decade witnessed explosion of genome-wide association studies (GWASs) of clinically defined cases against non-compromised individuals as controls with a hope that discovering risk genes in large cohorts would provide insights into patient stratification and further aid towards achieving P4 medicine goal. Despite the apparent success of this approach in complex traits, this technology-driven GWAS strategy which primarily relied on large sample sizes has witnessed serious limitations such as non-identification of disease associated genes of large effects which contribute to missing heritability and also non-replication of observed associations. This limitation has been largely due to the inherent heterogeneity owing to endogenous and exogenous factors involved in complex disorders, and by increasing sample sizes, led to inadvertent scaling up of heterogeneity in parallel and hence reduced the statistical power to detect real associations. Different strategies are now being used to overcome the GWAS related limitations and these include turning the emphasis on to rare variants, epigenetic modifications, miRNA etc<sup>1</sup>. Phenotype resolution however seems a likely major determinant of the success or failure of GWAS to date. The importance of accurate phenotyping over increasing sample size to detect true associations has been addressed in a recent study. The authors used both simulated and GWAS data for Type I and Type II diabetes and demonstrated that statistical power to detect real association was reduced when the study cohort was heterogeneous. In another words, if GWAS were carried out in more homogeneous sample sets the magnitude of risk conferred by the marginally/modestly significant risk variants would have been larger<sup>2</sup>.

Phenotype definitions in modern medicine largely depend on quantifiable parameters and ignores the underlying heterogeneity in disease pathogenesis. Therefore, we believe it is time to revisit and adopt newer non-conventional phenotyping approaches which may be able to capture molecular variability underlying the disease. Like personalised medicine, Ayurveda (the Indian system of medicine and one of the oldest in the world) is not a 'one-size-fit-all' approach but on the contrary addresses inter-individual variability effectively. It adopts a holistic approach towards healthy living on the basis of the concepts of Tridosha and prakriti. According to Ayurveda, all matter is comprised of the five basic elements or building blocks of nature: Earth, Air, Water, Fire and Space. Varying combination of these elements form the three basic humours/forces of human body namely Vata dosha, Pitta dosha and Kapha dosha collectively called as Tridoshas<sup>3</sup>. Each of these doshas have distinct properties and when there is balance between the constituents, they work in harmony through the body to maintain homeostasis. In other words, according to Ayurveda, maintenance of this balance is health and imbalance is disease. The innate proportion of doshas that a zygote acquires at the time of conception determines its prakriti, and this represents a summed-up phenotype or basic constitution type of an individual<sup>3</sup>. Prakriti defines physical, physiological, and psychological traits of an individual and is the template for individualized diet, lifestyle counselling and disease treatment. To this extent one's prakriti may be considered as the Ayurvedic equivalent of describing the unique genetic constitution (genome) of each individual in modern biology. However, ayurveda doctrines go further and according to Tridosha theory, depending upon the individual or combinatorial proportions of Tridosha in each person, there are seven possible prakriti types namely Vata, Pitta, Kapha, Vata-Pitta, Pitta-Kapha, Vata-Kapha and Vata-Pitta-Kapha contributing to wide phenotypic diversity highlighting their practice of deep phenotyping. Furthermore, according to its doctrines and practice, each of these prakriti types is the determinant of its own characteristic features such as metabolic profiles, disease predisposition, and natural history in individuals with respective prakriti<sup>4,5,6</sup>. This dosha type is also postulated to be

responsible for disease characteristics such as severity, therapeutic recommendations, and treatment outcome in individuals.

An empirical validation of this concept has been provided by our previous pilot study on candidate gene associations with rheumatoid arthritis (RA; termed Amavata in Ayurveda), among three subgroups that were based on Ayurveda based phenotyping<sup>7</sup>. This study revealed association of inflammatory genes with RA among Vata predominant prakriti subjects whereas oxidative stress genes were associated with Pitta subgroup. Further, disease severity was significantly higher in Vata compared to Pitta and Kapha predominant subgroups which is in agreement with what is known for Amavata in Ayurveda literature.

Significant correlations have also been established between prakriti and single nucleotide polymorphisms (SNPs) in genes such as *HLA* in healthy individuals and *EGLN1* in patients suffering from high-altitude pulmonary edema<sup>8,9</sup>. Apropos to the above evidence, differences in genome wide expression and biochemical profiles have been observed between the three extreme prakriti types<sup>10</sup>. Interestingly, a genome-wide analysis has revealed 52 prakriti differentiating SNPs in healthy individuals across the three predominant prakriti groups<sup>11</sup>. Of note, another recent study utilized modelling methods using the phenotyping data and was able to classify healthy individuals into three distinct clusters, which matched with the extreme prakriti groups as classified by clinicians<sup>12</sup>.

With this background, we hypothesize that GWAS carried out on homogeneous but small cohorts obtained by deep phenotyping based on Ayurveda doctrines will be more insightful since Ayurveda has comprehensive criteria to stratify not only controls but also cases. This approach would facilitate identification of genetic associations of large effect sizes which may enable filling the present knowledge gap in complex disease biology. In the present study, an Ayurgenomics approach was adopted wherein we carried out a GWAS on RA patients and healthy controls who were re-phenotyped for their prakriti type. Genome-wide analysis of prakriti matched RA cases and controls revealed potential prakriti-specific genetic associations. We believe this Ayurgenomics approach offers itself as a novel tool to perform prakriti based deep phenotyping of study cohorts prior to their inclusion in contemporary GWASs to obtain homogeneous cohorts and consequently identify true genetic associations with high effect sizes.

## Materials and methods

### Study Cohort

In the present study, 244 RA cases [49 Vata; 117 Pitta; 78 Kapha] and 293 controls [33 Vata; 175 Pitta; 85 Kapha] sub-grouped based on their predominant prakriti type (briefly described below) were recruited from Department of Ayurveda, Holy family hospital, New Delhi. Both cases and controls were matched for age, gender and ethnicity. DNA from venous blood drawn from study subjects was extracted according to routine phenol-chloroform protocol. This cohort has been used in two previous studies<sup>7,13</sup>.

### Ayurveda based phenotyping

The prakriti of each subject was assessed independently by two Ayurveda physicians using a validated questionnaire based on physical, physiological and psychological characteristics recommended by the Central Council for Research in Ayurveda and Siddha, Department of AYUSH, Ministry of health and family welfare, Government of India, New Delhi (<http://www.ccras.nic.in/>). Physique, skin texture, hunger, thirst, digestive capacity, temperament and memory are some of the major attributes evaluated to determine an individual's prakriti. Predominant prakriti was allotted if  $\geq 70\%$  dominance of a single dosha score was obtained. Only individuals with predominance of either Vata, Pitta or Kapha doshas were included in the study as described previously<sup>7</sup>. Objective parameters like height, weight, body mass index, blood pressure, swelling, blood/serum examination, X-rays and magnetic resonance imaging were used in the clinical assessment of cases. In addition, visual analogue scale was used for

most of the subjective features like pain, swelling, burning sensation, heaviness etc. Blood was drawn and used for analysis of haemoglobin%, erythrocyte sedimentation rate, rheumatoid factor and anti-cyclic citrullinated peptide antibody levels as described elsewhere<sup>7</sup>. RA diagnosis of all the patients thus enrolled were independently confirmed by an orthopaedic surgeon.

### Genotyping and Quality Control

Genotype information of above mentioned cohort was retrieved from the total genotype data generated on Illumina Human660W Quad BeadChip v1.C (655 216 markers) genotyping platform, as described before<sup>13</sup>. Genomic data was filtered using standard quality control steps on PLINK (v1.9beta3)<sup>14</sup>. Individuals and SNPs with a call rate of < 90%, individuals with discordant/ambiguous sex, putative inbreeding/contaminated samples (heterozygosity rate >4±standard deviation), ethnic outliers, duplicates and first-degree relatives (PIHAT >0.25) and SNPs with minor allele frequency <0.01 and under Hardy-Weinberg disequilibrium (p-value>5 E-06) were excluded.

### Statistical Analysis

Allele frequencies between cases and controls belonging to the same constitution type/prakriti were compared using Fisher's exact test on PLINK. A p-value threshold of 1 E-05 and a Bonferroni-corrected p-value threshold of 9.46 E-08 were considered suggestive and genome-wide significant, respectively. To identify potential link between the novel genetic variants identified with RA subgroups in this study and RA based on literature evidence, p-values were retrieved from Open Targets platform<sup>15</sup>. Biological interactions between genes were inferred using GeneMANIA prediction server which provides interactive functional association network<sup>16</sup>.

### Results

Power analysis performed with Quanto software showed >80% power in Vata, Pitta and Kapha subgroups.

### Association findings

A total of 444 individuals [229 cases (45 Vata, 113 Pitta, 71 Kapha) and 215 controls (24 Vata, 131 Pitta, 60 Kapha)] and 528461 SNPs remained for downstream analysis after quality control (**Supplementary figure 1**). GWA analysis was performed for the total study cohort as well as for the three predominant Prakriti groups separately. Novel SNPs with suggestive p-value 1 E-05 [**Table 1**] were found to be associated with the three RA sub-groups. However, none of the SNPs surpassed Bonferroni-corrected p-value. Manhattan and Q-Q plots are shown in **figure 1**. It was notable that all the associated markers/genes were unique to each of the groups. Broad function of the gene/nearest gene identified in prakriti-wise analysis are also described below.

#### a) Total cases vs controls

*TMEM179*, *TMEM18*, *NUAK1*, *TBC1D8*, and *LEF1-AS1*, together with *MZT1* (downstream of rs340575) and *TCERG1* (downstream of rs10056189) were identified when analysing total RA cases and controls. Most of these were found to be associated with suggestive significance (GWAS, p-value 1 E-05) in the three RA sub-groups [**Table 1**]. All six genes, except for the RNA gene *LEF1-AS1*, were associated with musculoskeletal system disease (Open Targets, p-value 0.02), which includes conditions that affect joints, such as osteoarthritis, rheumatoid arthritis, psoriatic arthritis, gout, ankylosing spondylitis among others. Of note, we also found that all these six genes physically and genetically interact with each other and are co-expressed (**Figure 2**). Furthermore, *LEF1-AS1*, *NUAK1*, and *TBC1D8* were seen to be associated with systemic juvenile idiopathic arthritis (Open Targets, p-value 0.002), and a genomic marker located at *C14orf180* (overlapped with *TMEM179*), rs4264325, was also seen to be significant for RA susceptibility.

#### b) Vata

Test of association between RA cases and controls categorised under Vata predominant prakriti identified one SNP (rs1953175) in *RP11-536O18.1* and two SNPs (rs4352629 and rs7448716) in *CTC-*

498M16.4 [Table 1]. None of these three Vata-specific SNPs showed association in the total group, or in Pitta and Kapha cohorts [Table 1]. Although no direct link was found between RA and the two genes, *CTC-498M16.4* has been shown to be associated with attention deficit/hyperactivity disorder (ADHD), rs4916723 (GWAS, p-value 2.67 E-05) being the lead SNP conditioning the gene<sup>17</sup>. This is important considering that though ADHD is currently conceptualized as a neurodevelopmental disorder, recent findings<sup>18</sup> have shown genetic connection between ADHD and immune alterations/autoimmune disorders. Further, the infective component in RA etiology indicates that there could be shared risk pathways between RA and ADHD with pleiotropic genetic effects.

### c) Pitta

In this group, we found six genes namely *TXNDC16* (rs11625685; rs11623917), *PCDH8* (rs9527038), *KLHL25* (rs4620912) *NTF3* (rs10849264), *RP11-93I21.3* (rs1390079), and *SERTM1* (rs7323558) significantly associated [Table 1] and all of which were functionally relevant. Additionally, physical interaction, co-expression or their presence in common pathways were found among these genes (Figure 2).

*TXNDC16* which encodes for Thioredoxin (Trx) Domain Containing 16, is an endoplasmic reticulum-associated glycoprotein and is believed to have putative redox activity<sup>19</sup>. A substantial number of studies have demonstrated the role of oxidative stress in RA pathogenesis<sup>20</sup>. It has been reported that cytosolic Trx system has a role in RA and Trx1 has shown to be significantly increased in the synovial fluid of RA patients<sup>19</sup>.

*PCDH8* is a protocadherin involved in neural development and function and it is shown to be dysregulated in several types of cancers and playing a critical role in tumor progression. Notably, a global gene expression profiling of chondrogenic tissues during *in vivo* development in mice showed involvement of *Pcdh8* in chondrogenesis<sup>21</sup>, a process by which cartilage is formed.

*KLHL25* belongs to Kelch family of protein that function as substrate-specific adaptors for Cullin E3 ubiquitin ligase (Cul3), a core component of the ubiquitin-proteasome system to regulate the protein turnover. It is important to mention here that our earlier study has shown an association between *CUL1* haplotype and methotrexate response in a north Indian population<sup>22</sup>. Similar findings were also observed in a RA cohort of Japanese origin<sup>23</sup>. Our findings lend further support to the role of ubiquitin pathway in autoimmunity and inflammation. Recent studies have identified mutations of several Kelch proteins in skeletal muscle disorders<sup>24</sup>. Though, no direct role of *KLHL25* has been implicated in RA, a recent study has proposed that increase in inflammatory processes and reactive oxygen species production leads to skeletal muscle deterioration<sup>25</sup> which in turn contributes to a vicious cycle of disease activity, muscle inflammatory signalling and disrupted remodelling, physical inactivity, and disability in patients with RA<sup>26</sup>.

Protein (NT-3) encoded by *NTF3* is a member of the neurotrophin family which are essential for the development and maintenance of the vertebrate nervous system. Neurotrophins and their receptors are shown to be expressed in the non-neuronal cells<sup>27</sup> supporting the role of neurotrophins beyond neurogenesis. A study has shown that LPS-treated mouse macrophages resulted in up-regulation of NT-3 leading to overproduction of nitric oxide, suggesting that NT-3 may play important roles in the function of macrophages during inflammatory responses and in tissue repair<sup>28</sup>. The role of NT-3 in RA has been empirically demonstrated by recent studies wherein over-expression of NT-3 in serum of RA patients<sup>29</sup> and high expression of *NTF-3* in RA synovial fibroblasts compared with healthy synovial fibroblasts under normoxic conditions has been observed<sup>30</sup>. In a recent study, NT-3 and its high affinity receptor TrkC were found to be highly induced at the injury site and endogenous NT-3 was found to promote bone repair<sup>31</sup>. In addition, NT-3 has also been implicated in neuropathic pain which is often poorly alleviated by first- and second-line medications due to lack of efficacy and/or dose-limiting side-effects<sup>32</sup>. Notably, neuropathic pain in substantial number of RA patients has been associated with



vitamin D deficiency<sup>33</sup> or with high disease activity and weight<sup>34</sup>. These observations are of clinical relevance since a better understanding of neuropathic pain mechanisms will provide a more targeted approach to pain treatment in RA.

RP11-93I21.3 is a long intergenic non-coding RNA (lncRNA). Though there is no report suggesting direct functional involvement of RP11-93I21.3 in RA pathogenesis, several lncRNAs are shown to be dysregulated in RA and are correlated with disease activity<sup>35</sup>.

*SERTM1*, encoding for serine rich and transmembrane domain containing 1 gene and has shown to be downregulated in psoriasis patients, thereby likely to play an important role in inflammation associated with RA patients<sup>36</sup>.

#### d) Kapha

Three genes namely *ZBTB34* (rs3120029), *ITGB8* (rs11762117), and *GPR12* (rs9512378) were found to be significantly associated in Kapha group but not in Vata and Pitta [Table 1]. All three genes were found to be physically interacting or were part of common pathways [Figure 2]. Genomic markers in all three genes were associated with scoliosis (Open Targets, p-value 0.002), fat body mass (Open Targets, p-value 0.01), and also rheumatic disease (Open Targets, p-value 0.01).

*ZBTB34*, a nuclear protein, is a new member of the BTB/POZ zinc finger protein family. Although exact role of *ZBTB34* is not known, some of the proteins of this family critically regulate development of specific lineages in the immune system, promote oncogenesis and maintain stem cells. It has also been suggested that *ZBTB34* might function as a transcriptional repressor<sup>37</sup>. In zebrafish, *ZBTB* is predicted to be involved in negative regulation of transcription by RNA polymerase II; regulation of cytokine production; and regulation of immune system process<sup>38</sup>. *ZBTB34* has been shown to be overexpressed in the whole blood of axial spondylarthritis/ankylosing spondylitis patients compared to healthy controls<sup>39</sup>. Of note, our previous GWAS has shown significant association of a different SNP (rs561041) closest to *ZBTB34* with RA<sup>13</sup>.

*GPR12* is classified as an orphan G protein-coupled receptor. Disruption of *Gpr12* gene in mice has shown to provoke changes in both lipid and carbohydrate metabolism resulting in dyslipidemia and obesity<sup>40</sup> and therefore considered to be involved in regulating energy expenditure and important for future drugs that target this receptor. Of note, *GPR12* has been identified as a novel target of Cannabidiol, which is shown to have therapeutic potential for arthritis pain-related behaviors and inflammation without evident side-effects<sup>41</sup>.

*ITGB8* is a member of the integrin beta chain family and has been involved in angiogenesis deregulation in systemic sclerosis, a chronic autoimmune rheumatic disorder<sup>42</sup>. Furthermore, importance of *ITGB8* in chondrogenesis has been previously established<sup>43</sup>, suggesting its direct involvement in RA pathology, as progressive loss of cartilage due to inflammatory response is one of the disease characteristics. This derives further support from a gene expression study wherein *ITGB8* was shown to be highly expressed in RA synovial fibroblasts compared with healthy synovial fibroblasts under normoxic conditions suggesting its role in chronic synovitis<sup>30</sup>.

#### Discussion

RA is a chronic inflammatory joint disease affecting synovial tissue in multiple joints but with poorly uncovered etiology. It is a clinically and biologically heterogeneous disease with respect to both disease course and treatment outcome suggesting distinct molecular mechanisms contributing to RA in different patients. For instance, differences in the activation of the STAT1 pathway between rheumatoid tissues confirms etiological heterogeneity<sup>44</sup>. Continuous efforts are being made to subclassify clinically diagnosed RA on the basis of molecular criteria/signatures using OMICS or more recently using phenome wide association study approach<sup>45</sup>. While we still await their deliverables, exploring non-conventional phenotyping approaches and providing scientific validation for their utility

as an adjunct could accelerate the progress in this endeavour. In this proof-of-concept study, we performed a GWAS of RA patients and healthy controls who were phenotypically sub-classified into three prakriti groups namely Vata, Pitta and Kapha predominant by employing the ancient deep phenotyping principles practiced in Ayurveda system<sup>4</sup>. Despite small numbers in each of the three subgroups, this novel Ayurgenomics approach identified prakriti-specific genes of high effect sizes [Table 1], most of them not hitherto identified for RA across multiple GWASs performed in large cohorts or even in meta-analyses data (www.ebi.ac.uk/gwas). Our results also highlight the striking difference in the genetic associations identified in the total group versus those in each of the three prakriti based subgroups (Table 1). These findings imply that even small sample size of tightly defined cases and controls or just precise phenotyping may have led to comparatively more genetically homogeneous groups which were sufficient to maximize the detection of common alleles conferring high risk and minimize statistical noise. This derives support from a recent report wherein a locus with genome-wide significance was identified near the gene encoding parathyroid hormone-like hormone in a GWAS performed in a cohort of only 40 patients with peripartum cardiomyopathy<sup>46</sup>. In addition, we also found lack of HLA markers [Supplementary table 1] which may suggest that p-values of genes of minor/moderate effect are largely driven by sample size. This is also witnessed in our study wherein GWAS p-values for HLA markers are more significant in the total cohort (n=229 cases and 215 controls) and Pitta subgroup (n=117 cases and n=175 controls) compared to Vata and Kapha [Supplementary table 1].

Taken together, the novel study findings lend credence that association studies conducted on homogeneous subgroups enable identification of disease specific genes of major effect size. Of the genes identified in the different subgroups in our study, *NTF3*, *KLH25*, *TXNDC16*, *PCDH8*, *ITGB8*, and *GPR12* [Table 1] look promising and may provide a new perspective and prompt us to explore their active involvement and therapeutic potential in chronic inflammatory arthritis. At the moment we lack clarity on the correlation between function of these genes in the different Ayurveda subgroups and prakriti-specific disease etiology/mechanism, yet Ayurveda wisdom (explained briefly below) supports our findings to some extent.

#### Insights into RA biology from Ayurveda

Stratifying healthy individuals into seven constitution types or prakriti for predicting prakriti-specific disease susceptibilities and clinical outcomes such as treatment response forms the basis of Ayurveda medical practice and also explains inter-individual variability. To elaborate this further, individuals with Vata prakriti are more predisposed to RA and are the most difficult group to treat compared to the Pitta subgroup who are less prone, manifest mild to moderate symptoms and are also easier to treat with better outcome<sup>47</sup>. Furthermore, disease severity is more pronounced in RA patients with Vata prakriti, who suffer severe throbbing pain, which worsens in cold weather; Pitta patients experience burning sensation, redness, swelling, and inflammation, which worsens with hot weather; and Kapha patients show loss of movement, itching, joint swelling and edema (without inflammation), with other symptoms including dullness, heaviness and aches<sup>48</sup>.

As for treatment, Ayurveda believes that RA (amavata) is a problem of the gut, or in other words, a metabolic disorder, and therefore improving the digestive capacity which varies according to individual's prakriti is the primary focus of its treatment regime. This is in line with the emerging role of gut microbiome dysbiosis in RA. According to Ayurveda, hypo functioning of Agni/digestive power (corresponding to enzymes, chemicals, hormones, neurotransmitters and cytokines known to modern medicine) results in impaired digestion and absorption of food, which leads to the formation of immunologic and toxic substances called "Ama"<sup>49</sup>. This ama when circulates in the body lodges in the joints and leads to inflammation. Of note, Agni which is responsible for metabolism, absorption, etc is believed to be prakriti specific with best/strong digestive/metabolic power in Pitta (Pachaka Pitta) followed by Kapha (Kledaka Kapha) and then Vata (Samana Vata) sub-groups<sup>50</sup>. Therefore, the

treatment of Amavata focuses primarily on improving the digestive capacity, and removal of Ama (which in other words is treating the cause of the disease). Reducing the pain/ inflammation is a secondary treatment based on the disease symptoms and as mentioned above, this treatment is also prakriti specific.

### Conclusions

Identification of novel prakriti specific and more importantly, functionally relevant susceptibility genes (as shown to be supported by other functional studies) of intermediate/high effect size for RA, suggest that Ayurveda based deep phenotyping could be an effective approach to achieve the highly desirable sample homogeneity in complex trait genetics. This may propel **i)** a better development of multi-omics signature based prognostic and diagnostic markers and **ii)** allow prakriti specific nutritional and therapeutic intervention strategies. Further, such homogeneous cohorts catalyse rare variant identification as the focus of genetic studies turns from common to rare variants. We strongly believe that using non-conventional phenotyping approaches practiced in complementary systems of medicine such as Ayurveda, Unani, and Chinese traditional medicine along with modern medicine diagnostic/therapeutic knowledge will broaden our horizon of disease biology and provide insights into disease genetics, which remains an urgent unmet need to break ground in complex traits and fulfil the P4 medicine goal. However, these novel findings endorse replication in independent cohorts.

### Ethical approval

Institutional ethics committee clearance was obtained from Ayurveda Department, Holy Family Hospital and Department of Genetics, University of Delhi South Campus, New Delhi, prior to initiation of this study and the methods were carried out in accordance with the approved guidelines.

### Informed consent

Written informed consent was obtained from all subjects participating in this study.

### Consent for publications

Not applicable

### Availability of data and materials

The datasets used and /or analysed during the current study are available from the corresponding authors on reasonable request

### Authors' contributions

BKT, RCJ and BB conceived the idea; Study was designed by GJ, RCJ and BKT; BB, UK, RCJ and BKT generated the funds; Primary and secondary data analyses were performed by GJ, AP, SLG, SN and RG; GJ and BKT wrote the manuscript; all authors reviewed and approved the final manuscript

### Legends

**Figure 1. Manhattan plots depicting SNP associations with the (a) total RA cohort and (b-d) prakriti specific RA sub-groups in the north Indian population, and respective Q-Q plots.** On the Manhattan plot, all SNPs are plotted according to their position on each chromosome on x-axis, against their association ( $-\log_{10}(\text{p-value})$ ) on y-axis. The red line represents the Bonferroni-corrected threshold ( $\text{p-value}=9.46 \text{ E-}08$ ), while the blue line represents the suggestive association threshold ( $\text{p}=1\text{E-}05$ ). The inset Q-Q plots show the observed (y-axis) against the expected (x-axis) distribution of GWAS p-values under the null hypothesis for the total RA cohort and prakriti sub-groups

**Figure 2. GeneMANIA network showing potential interactions with novel RA genes identified in total study cohort and in prakriti-specific RA sub-groups.** The genes of interest are represented in the middle with striped circles. Pink lines represent physical Interactions; green lines represent genetic interactions; purple lines represent co-expression; blue lines represent pathways

**Table 1. Genome-wide suggestive significant (p-value < 1 E-05) variants associated with RA in (a) total study cohort and (b-d) prakriti specific RA groups.** Positions refer to assembly GRCh37

**Supplementary Table 1: List of SNPs in and around HLA region which surpassed genome-wide significance ( $p < 10^{-5}$ ) in our previous RA GWAS study<sup>13</sup> and their significance status in the three Ayurveda sub-groups and total cohort**

**Supplementary Figure 1. Flow diagram of the quality control steps performed before the GWAS for individuals and SNPs**

### Acknowledgments

Gratefully acknowledge Dr. Preeti Wakhode, Ayurveda physician, for independent assessment of the study cohort for their prakriti using the questionnaire; and Apoorva Anand, Anuroop Venkateswaran and Neha Rana for their help with initial data collation.

### Funding

The financial assistance for this work was provided by Central Council for Research in Ayurvedic Sciences, Ministry of AYUSH, New Delhi, vide F.No. Z.31018/18/2006-R&P; and #BT/01/COE/07/UDSC/2008 (Phase I) from Department of Biotechnology, New Delhi, India. Senior research fellowship from Council of Scientific and Industrial Research, New Delhi to AP; Financial support from Idella Foundation to SLG; Infrastructure support provided to the Department of Genetics, University of Delhi South Campus, by the University Grants Commission, New Delhi under the Special Assistance Programme and Department of Science and Technology, New Delhi under FIST and DST PURSE programmes are gratefully acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### References:

1. Seyhan AA, Carini C. Are innovation and new technologies in precision medicine paving a new era in patients centric care? *Open Access Journal of Translational Medicine. J Transl Med.* 2019;17:114. doi:10.1186/s12967-019-1864-9
2. Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, Alda M. The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases. *PLoS One.* 2013;8(10). doi:10.1371/journal.pone.0076295
3. Chakraborty R. Genesis of Personalized Medicine: Relevance of Ayurveda in the Present Millennium. Published online 2017. doi:10.4172/1747-0862.1000285
4. Charak. *Charaka Samhita Vimana Sthana* 8.
5. *Sushruta Samhita Sharira Sthana* 4.
6. *Ashtanga Hridaya Sharira Sthana* 3.
7. Juyal RC, Negi S, Wakhode P, Bhat S, Bhat B, Thelma BK. Potential of Ayurgenomics Approach in Complex Trait Research: Leads from a Pilot Study on Rheumatoid Arthritis. *PLoS One.* 2012;7(9). doi:10.1371/journal.pone.0045752
8. Bhushan P, Kalpana J, Arvind C. Classification of human population based on HLA gene polymorphism and the concept of Prakriti in Ayurveda. *J Altern Complement Med.* 2005;11(2):349-353. doi:10.1089/acm.2005.11.349
9. Aggarwal S, Negi S, Jha P, et al. EGLN1 involvement in high-altitude adaptation revealed through genetic analysis of extreme constitution types defined in Ayurveda. *Proc Natl Acad Sci U S A.* 2010;107(44):18961-18966. doi:10.1073/pnas.1006108107
10. Prasher B, Negi S, Aggarwal S, et al. Whole genome expression and biochemical correlates of extreme constitutional types defined in Ayurveda. *J Transl Med.* 2008;6(1):48. doi:10.1186/1479-5876-6-48

11. Govindaraj P, Nizamuddin S, Sharath A, et al. Genome-wide analysis correlates Ayurveda Prakriti. *Sci Rep*. 2015;5:15786. doi:10.1038/srep15786
12. Tiwari P, Kutum R, Sethi T, et al. Recapitulation of Ayurveda constitution types by machine learning of phenotypic traits. Chaubey G, ed. *PLoS One*. 2017;12(10):e0185380. doi:10.1371/journal.pone.0185380
13. Negi S, Juyal G, Senapati S, et al. A genome-wide association study reveals ARL15, a novel non-HLA susceptibility gene for rheumatoid arthritis in North Indians. *Arthritis Rheum*. 2013;65(12):3026-3035. doi:10.1002/art.38110
14. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795
15. Carvalho-Silva D, Pierleoni A, Pignatelli M, et al. Open Targets Platform: New developments and updates two years on. *Nucleic Acids Res*. 2019;47(D1):D1056-D1065. doi:10.1093/nar/gky1133
16. Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*. 2010;38(SUPPL. 2). doi:10.1093/nar/gkq537
17. Liao C, Laporte AD, Spiegelman D, et al. Transcriptome-wide association study of attention deficit hyperactivity disorder identifies associated genes and phenotypes. *Nat Commun*. 2019;10(1). doi:10.1038/s41467-019-12450-9
18. Tylee DS, Sun J, Hess JL, et al. Genetic correlations among psychiatric and immune-related phenotypes based on genome-wide association data. *Am J Med Genet Part B Neuropsychiatr Genet*. 2018;177(7):641-657. doi:10.1002/ajmg.b.32652
19. Hanschmann EM, Godoy JR, Berndt C, Hudemann C, Lillig CH. Thioredoxins, glutaredoxins, and peroxiredoxins-molecular mechanisms and health significance: From cofactors to antioxidants to redox signaling. *Antioxidants Redox Signal*. 2013;19(13):1539-1605. doi:10.1089/ars.2012.4599
20. Fonseca LJS Da, Nunes-Souza V, Goulart MOF, Rabelo LA. Oxidative Stress in Rheumatoid Arthritis: What the Future Might Hold regarding Novel Biomarkers and Add-On Therapies. *Oxid Med Cell Longev*. 2019;2019. doi:10.1155/2019/7536805
21. Cameron TL, Belluoccio D, Farlie PG, Brachvogel B, Bateman JF. Global comparative transcriptome analysis of cartilage formation in vivo. *BMC Dev Biol*. 2009;9(1). doi:10.1186/1471-213X-9-20
22. Negi S, Kumar A, Thelma BK, Juyal RC. Association of Cullin1 haplotype variants with rheumatoid arthritis and response to methotrexate. *Pharmacogenet Genomics*. 2011;21(9):590-593. doi:10.1097/FPC.0b013e3283492af7
23. Kawaida R, Yamada R, Kobayashi K, et al. CUL1, a component of E3 ubiquitin ligase, alters lymphocyte signal transduction with possible effect on rheumatoid arthritis. *Genes Immun*. 2005;6(3):194-202. doi:10.1038/sj.gene.6364177
24. Gupta VA, Beggs AH. Kelch proteins: Emerging roles in skeletal muscle development and diseases. *Skelet Muscle*. 2014;4(1):11. doi:10.1186/2044-5040-4-11
25. Oyenihni AB, Ollewagen T, Myburgh KH, Powrie YSL, Smith C. Redox status and muscle pathology in rheumatoid arthritis: Insights from various rat hindlimb muscles. *Oxid Med Cell Longev*. 2019;2019. doi:10.1155/2019/2484678
26. Huffman KM, Jessee R, Andonian B, et al. Molecular alterations in skeletal muscle in rheumatoid arthritis are related to disease activity, physical inactivity, and disability.

- Arthritis Res Ther.* 2017;19(1):12. doi:10.1186/s13075-016-1215-7
27. Nockher WA, Renz H. Neurotrophins in clinical diagnostics: pathophysiology and laboratory investigation. *Clin Chim Acta.* 2005;352(1-2):49-74. doi:10.1016/j.cccn.2004.10.002
  28. Barouch R, Appel E, Kazimirsky G, Brodie C. Macrophages express neurotrophins and neurotrophin receptors: Regulation of nitric oxide production by NT-3. *J Neuroimmunol.* 2001;112(1-2):72-77. doi:10.1016/S0165-5728(00)00408-2
  29. Panezai J, Ali A, Ghaffar A, et al. Upregulation of circulating inflammatory biomarkers under the influence of periodontal disease in rheumatoid arthritis patients. *Cytokine.* 2020;131. doi:10.1016/j.cyto.2020.155117
  30. Del Rey MJ, Izquierdo E, Usategui A, et al. The transcriptional response of normal and rheumatoid arthritis synovial fibroblasts to hypoxia. *Arthritis Care Res.* 2010;62(12):3584-3594. doi:10.1002/art.27750
  31. Su YW, Chung R, Ruan CS, et al. Neurotrophin-3 Induces BMP-2 and VEGF Activities and Promotes the Bony Repair of Injured Growth Plate Cartilage and Bone in Rats. *J Bone Miner Res.* 2016;31(6):1258-1274. doi:10.1002/jbmr.2786
  32. Khan N, Smith MT. Neurotrophins and neuropathic pain: Role in pathobiology. *Molecules.* 2015;20(6):10657-10688. doi:10.3390/molecules200610657
  33. Yesil H, Sungur U, Akdeniz S, Gurer G, Yalcin B, Dundar U. Association between serum vitamin D levels and neuropathic pain in rheumatoid arthritis patients: A cross-sectional study. *Int J Rheum Dis.* 2018;21(2):431-439. doi:10.1111/1756-185X.13160
  34. Ito S, Kobayashi D, Murasawa A, Narita I, Nakazono K. An analysis of the neuropathic pain components in rheumatoid arthritis patients. *Intern Med.* 2018;57(4):479-485. doi:10.2169/internalmedicine.9235-17
  35. Lao MX, Xu HS, Guo LS. Involvement of long non-coding RNAs in the pathogenesis of rheumatoid arthritis. *Chin Med J (Engl).* 2020;133(8):941-950. doi:10.1097/CM9.0000000000000755
  36. Ahn R, Yan D, Chang HW, et al. RNA-seq and flow-cytometry of conventional, scalp, and palmoplantar psoriasis reveal shared and distinct molecular pathways. *Sci Rep.* 2018;8(1):11368-11368. doi:10.1038/s41598-018-29472-w
  37. Qi J, Zhang X, Zhang HK, Yang HM, Zhou YB, Han ZG. ZBTB34, a novel human BTB/POZ zinc finger protein, is a potential transcriptional repressor. *Mol Cell Biochem.* 2006;290(1-2):159-167. doi:10.1007/s11010-006-9183-x
  38. ZFIN The Zebrafish Information Network. Accessed October 25, 2020. <https://zfin.org/>
  39. Park R, Kim T-H, Ji JD. Gene Expression Profile in Patients with Axial Spondyloarthritis: Meta-analysis of Publicly Accessible Microarray Datasets. *J Rheum Dis.* 2016;23(6):363. doi:10.4078/jrd.2016.23.6.363
  40. Bjursell M, Gerdin AK, Jönsson M, et al. G protein-coupled receptor 12 deficiency results in dyslipidemia and obesity in mice. *Biochem Biophys Res Commun.* 2006;348(2):359-366. doi:10.1016/j.bbrc.2006.07.090
  41. Hammell DC, Zhang LP, Ma F, et al. Transdermal cannabidiol reduces inflammation and pain-related behaviours in a rat model of arthritis. *Eur J Pain (United Kingdom).* 2016;20(6):936-948. doi:10.1002/ejp.818
  42. Giusti B, Margheri F, Rossi L, et al. Correction: Desmoglein-2-integrin Beta-8 interaction regulates actin assembly in endothelial cells: Dereulation in Systemic sclerosis (PLoS ONE (2013) 8, 7, (e68117) doi: 10.1371/journal.pone.0068117). *PLoS One.* 2013;8(7).

- doi:10.1371/annotation/b41766f2-c23d-455e-8d6e-e4bce5ae1d80
43. LaPointe VLS, Verpoorte A, Stevens MM. The Changing Integrin Expression and a Role for Integrin  $\beta 8$  in the Chondrogenic Differentiation of Mesenchymal Stem Cells. Connon CJ, ed. *PLoS One*. 2013;8(11):e82035. doi:10.1371/journal.pone.0082035
44. Van der Pouw Kraan TCTM, Van Gaalen FA, Kasperkovitz P V., et al. Rheumatoid arthritis is a heterogeneous disease: Evidence for differences in the activation of the STAT-1 pathway between rheumatoid tissues. *Arthritis Rheum*. 2003;48(8):2132-2145. doi:10.1002/art.11096
45. Liao KP, Sparks JA, Hejblum BP, et al. Phenome-Wide Association Study of Autoantibodies to Citrullinated and Noncitrullinated Epitopes in Rheumatoid Arthritis. *Arthritis Rheumatol*. 2017;69(4):742-749. doi:10.1002/art.39974
46. Horne BD, Rasmusson KD, Alharethi R, et al. Genome-Wide Significance and Replication of the Chromosome 12p11.22 Locus Near the *PTHLH* Gene for Peripartum Cardiomyopathy. *Circ Cardiovasc Genet*. 2011;4(4):359-366. doi:10.1161/CIRCGENETICS.110.959205
47. *Ashtanga Hridaya Sutra Sthana 13/25*.
48. Shrikrishnadass K. *Madhava MNidhana, Chapter 25 Verse 11.*; 2007.
49. *Madhava Nidana 25/1*.
50. *Charaka Samhita Sutra Sthana 20*.

<b>Table 1</b>											
<b>SNP</b>	<b>Chr</b>	<b>Position</b>	<b>Location</b>	<b>Gene/Nearest gene</b>	<b>Risk allele</b>	<b>P<sub>GWAS</sub></b>	<b>OR (95% CI)*</b>	<b>P<sub>GWAS</sub>Total</b>	<b>P<sub>GWAS</sub>Vata</b>	<b>P<sub>GWAS</sub>Pitta</b>	<b>P<sub>GWAS</sub>Kapha</b>
<b>Total study cohort</b>											
rs340575 (C>A)	13	72876358	intergenic	MZT1	C	9.7E-08	6.52 (2.91-14.62)		0.21	0.00002255	0.003
rs10056189 (C>T)	5	145804118	intergenic	TCERG1	C	1.4E-06	3.41 (2.02-5.78)		0.05	0.0008	0.01
rs4956041 (C>T)	4	109113914	intron	LEF1-AS1	T	2.1E-06	1.92 (1.47-2.51)		0.01	0.01	0.006
rs4983599 (G>A)	14	105011436	intron	TMEM179	A	2.6E-06	2.13 (1.55-2.93)		0.05	0.0005	0.06
rs2867116 (C>A)	2	682363	upstream	TMEM18	C	4.7E-06	4.56 (2.26-9.23)		0.0001	0.02	0.02
rs7556762 (T>G)	2	101765922	intron	TBC1D8	T	6.7E-06	2.41 (1.63-3.55)		0.01	0.002	0.06
rs4548807 (G>A)	14	105040988	intron	TMEM179	A	7.6E-06	2.12 (1.53-2.96)		0.003	0.007	0.06
rs3782690 (T>G)	12	106464063	intron	NUAK1	T	8.9E-06	1.84 (1.41-2.42)		0.2	0.005	0.0008
<b>Vata</b>											
rs1953175 (G>T)	9	13505851	intergenic	RP11-536O18.1	T	3.7E-06	6.41 (2.87-14.34)	0.2		0.2	0.3
rs4352629 (C>T)	5	87756821	intron	CTC-498M16.4	T	7E-06	5.64 (2.63-12.08)	0.3		0.7	0.1
rs7448716 (A>G)	5	87752695	intron	CTC-498M16.4	G	7E-06	5.64 (2.63-12.08)	0.2		0.7	0.1
<b>Pitta</b>											
rs11625685 (T>C)	14	52931884	intron	TXNDC16	C	1.4E-06	4.50 (2.35-8.61)	0.00005	0.4		0.2
rs11623917 (A>G)	14	52921896	intron	TXNDC16	G	2.4E-06	4.37 (2.28-8.38)	0.0001	0.5		0.3
rs9527038 (A>G)	13	53503775	intergenic	PCDH8	G	3.9E-06	2.48 (1.69-3.64)	0.003	0.5		0.8
rs4620912 (C>T)	15	86389516	intergenic	KLHL25	T	4.5E-06	2.80 (1.8-4.37)	0.00004	0.5		0.5
rs10849264 (A>G)	12	5531307	intergenic	NTF3	G	4.8E-06	2.35 (1.62-3.39)	0.0003	0.3		0.7
rs1390079 (T>C)	4	125523042	intergenic	RP11-93I21.3	C	5.2E-06	13.32 (3.09-57.45)	0.007	0.5		0.3
rs7323558 (C>T)	13	37250048	intron	SERTM1	T	9.3E-06	2.305 (1.60-3.33)	0.002	0.9		0.7
<b>Kapha</b>											
rs3120029 (G>A)	9	129649356	downstream	ZBTB34	A	6.2E-06	4.243 (2.21-8.16)	0.2	0.4	0.2	
rs9512378 (A>G)	13	27363688	intergenic	GPR12	A	7.4E-06	4.60 (3.31-9.15)	0.01	0.8	0.7	
rs11762117 (C>A)	7	20391383	intron	ITGB8	C	8.8E-06	8.74 (2.93-26.03)	0.02	0.3	0.6	



Fig. 1

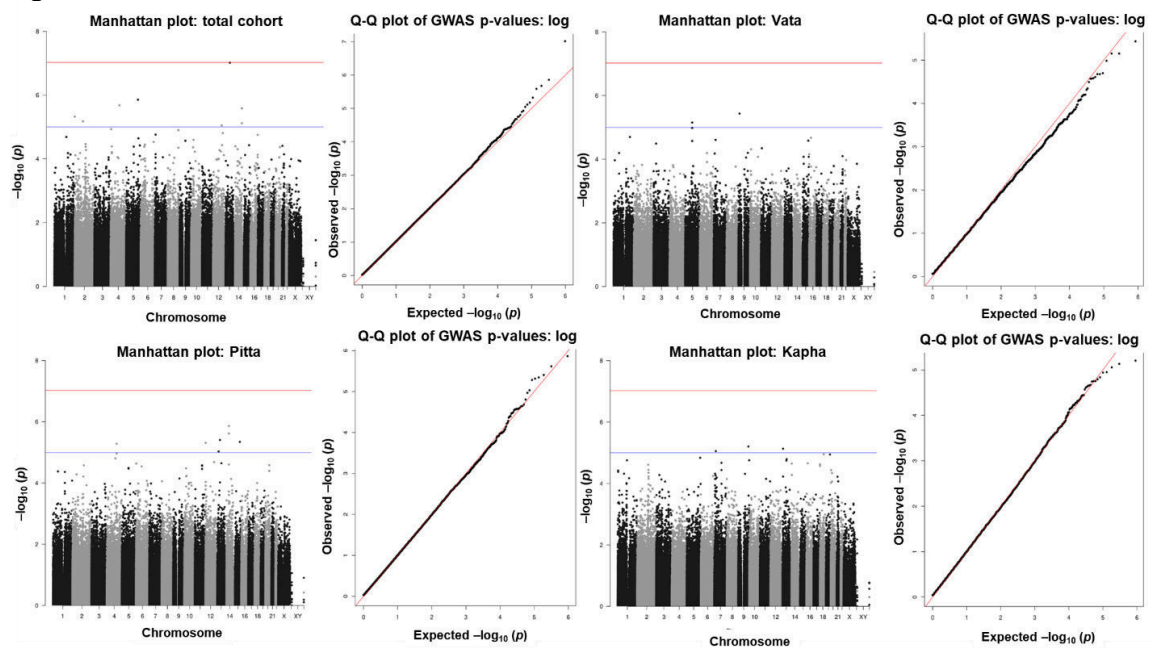
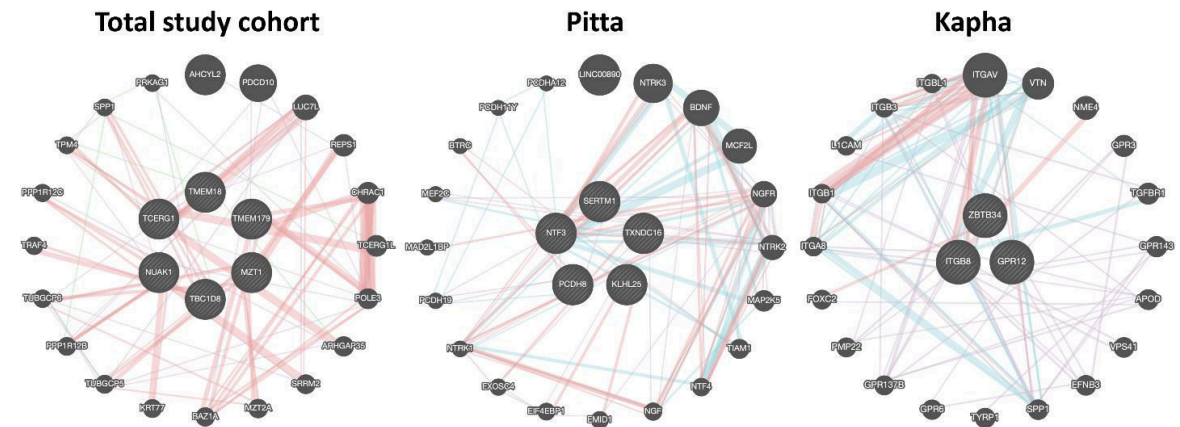


Fig. 2



Supplementary Figure 1

[Click here to access/download;Figure;Supplementary Figure-Final.tif](#)

Supplementary Figure-Final.pdf

	Number of patients	Number of markers
	537	655 216
Step 0. Remove poorly mapped markers and those w/o agreement between v1.0 and v1.1 and Will Rayner and Illumina	537	573 502
Step 1. Removal of individuals/SNPs with low call rate ↓ Keep individuals with call rate $\geq 90\%$ and SNPs $\geq 90\%$	478	549 466
Step 2. Removal of individuals with discordant sex information ↓ Keep individuals with chromosome X homozygosity rate $\leq 20\%$	475	549 466
Step 3. Removal of individuals with excessive heterozygosity rate ↓ Keep individuals with heterozygosity rate $\leq 4 \times \text{S.D.}$	472	549 466
Step 4. Removal of related individuals ↓ Keep individuals with identity-by-descent $\leq 0.25$	444	549 466
Step 5. Removal of population outliers ↓ Keep individuals within $4 \times \text{S.D.}$ from the population structure cluster center mean	444	549 466
Step 6. Removal of rare and non-HWE markers Keep SNPs with MAF $\geq 1\%$ and follow Hardy-Weinberg Equilibrium ( $p\text{-value} < 5 \times 10^{-6}$ )	444	528 461

## Chapter 5

# Paper II: Genetics scores in childhood cancer

Sara L Garcia, Marianne Helenius, Jonas Vestergaard, Adrian O. Laspior, Thomas van Overeem Hansen, Ulrik Soltze, Kjeld Schmiegelow, Ramneek Gupta, Rikke L. Nielsen, Karin Wadt

**Evaluation of adult cancer polygenic risk scores for stratified disease prevention in childhood cancer.**

## Evaluation of adult cancer polygenic risk scores for stratified disease prevention in childhood cancer

**Tentative author list:** Sara L. Garcia<sup>1</sup>, Marianne Helenius<sup>1</sup>, Jonas Vestergaard<sup>1</sup>, Adrian O. Laspior<sup>1</sup>, Thomas van Overeem Hansen<sup>2,3</sup>, Ulrik Stoltze<sup>3</sup>, Kjeld Schmiegelow<sup>3,4</sup>, Ramneek Gupta<sup>1,5</sup>, Rikke Linnemann Nielsen<sup>1,3\*</sup>, Karin Wadt<sup>2,4\*</sup>

<sup>1</sup>Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

<sup>2</sup>Department of Clinical Genetics, Rigshospitalet, University Hospital of Copenhagen, Denmark

<sup>3</sup>Department of Pediatrics and Adolescent Medicine, University Hospital Rigshospitalet, Copenhagen, Denmark

<sup>4</sup>Institute of Clinical Medicine, Faculty of Medicine, University of Copenhagen, Copenhagen, Denmark

<sup>5</sup>Novo Nordisk Research Centre Oxford, Oxford, OX3 7FZ, UK

\*Co-corresponding authors.

**Keywords: childhood cancer, polygenic risk score, breast cancer, colorectal cancer, phenotype stratification**

Genome-wide association studies (GWAS) have identified several single nucleotide polymorphisms (SNPs) across the genome that may contribute to a person's risk of developing diseases, although the individual contribution is very low with hazard ratio (HR) being <3, and mostly <2. This has provided understanding of which pathways influence complex conditions. Many of these variants showed association with multiple cancer types indicating pleiotropic effects and shared biological and etiologic mechanisms.

Polygenic risk score (PRS) provides an overall summary estimate of the genetic propensity to a trait at the individual level, which may provide stronger and clinically applicable risk scores than individual SNPs. Thus, PRSs are useful to identify patients with a substantially increased genetic risk of a disease.

Here, we hypothesized that adult cancer PRSs may provide useful risk scores for childhood cancers, as we observe 1) pleiotropic effects and the presence of well-established and validated PRSs on genetic predisposition of several adult cancers, and 2) that germline mutations predisposing to adult cancers (e.g. BRCA genes) are frequent in children with cancer. Using acute lymphoblastic leukaemia (ALL), the most common childhood cancer as a prototype, we compared these PRSs between different cancer subgroups (B or T-cell, cytogenetic alterations, and others) within a childhood ALL case cohort (NOPHO, N=1952), and differences between patients with solid tumours, haematological cancer, or CNS tumours in a Danish childhood cancer cohort (STAGING, N=425), to evaluate genetic predisposition's risk on subgroups that reflect aetiology, but also in time could contribute to further develop a downstream treatment stratification or knowledge with prognostic implications.

### INTRODUCTION

Cancer burden quantification is more challenging in children than in adults, since the first is rarer and often presents non-specific symptoms often confused with infections and/or nutritional conditions (Spector, Pankratz, and Marcotte 2015)(Johnston et al. 2021).

While few environmental risk factors for childhood cancer have been established, recent studies have identified germline predisposition in established childhood cancer genes, surprisingly also enriched in known adult onset-cancer predisposition genes (Spector, Pankratz, and Marcotte 2015). In multiple types of adult-onset cancers, 5-10% of affected

**Cross-cancer adult-child PRS**

patients carry a germline disposition in a high penetrant cancer gene (Lu et al. 2014), and several studies have recently confirmed a similar frequency of adult cancer predisposition syndromes in childhood-onset cancers (Zhang et al. 2015) (Parsons et al. 2016)(Byrjalsen et al. 2020), which underscores that a link could be present. Twin studies of adult-onset cancers have suggested a much higher heritability fraction (Lichtenstein et al. 2000)(Mucci et al. 2016), which partially can be explained by accumulation of single nucleotide polymorphism (SNPs), that individually only contributes with a slight increase in cancer risk. Genome-wide association studies (GWAS) have contributed to understand better genetic predisposition of cancer. These results have been used to create polygenic risk scores (PRS), which give an individual approximation of genetic propensity to develop cancer as it accounts for the estimated effect of several genetic variants (Kachuri et al. 2020). Some studies have found genetic correlations between different cancer types (Sampson et al. 2015)(Lindström et al. 2017)(Jiang et al. 2019). Furthermore, several PRSs have been evaluated in multiple different cancer types and it was possible to identify several positive correlations, indicating universal traits affecting carcinogenesis, telomere maintenance genes, haematopoiesis, inflammation, and cell migration, as well as non-cancer traits, such as obesity-induced chronic inflammation (Graff et al. 2021). Here, we hypothesized that two recently developed and validated PRSs may additionally be associated with leukaemia subgrouping of patients and/or prognosis of disease. This strategy is motivated by recent literature on genetic correlations between different cancer types and may help identify new paediatric cancer genes in rare and complex childhood cancer subgroups. These PRSs were previously associated with two of the most frequent adult onset-cancer types: breast- (Mavaddat et al. 2019) and colorectal (Huyghe et al. 2019) cancer, and were independently validated. These were also utilised in cross-cancer PRS evaluation studies for breast (Kachuri et al. 2020)(Jia et al. 2020) and colorectal (Graff et al. 2021) cancer. For this study, two childhood cancer cohorts were used: the Nordic Society of Pediatric Hematology and Oncology (NOPHO) cohort, consisting of children with several sub-types of leukemia etiology and the Sequencing Tumor and Germline DNA – Implications for National Guidelines (STAGING) cohort, consisting of children with multiple cancer types, such as hematological, solid, and central nervous system (CNS).

**MATERIALS AND METHODS****Study populations****NOPHO cohort**

For the NOPHO cohort, patients between one and 45 years old, diagnosed with ALL and treated according to the NOPHO ALL2008 protocol in Sweden, Denmark, Norway, Lithuania, Finland, Iceland, and Estonia between 2008 and 2018, were registered in the NOPHO ALL2008 database. In Finland and Iceland, only children were included. This study and patient collection have been described in previous studies (Rank et al. 2018)(Toft et al. 2013)(Frandsen et al. 2014)(Toft et al. 2016). Only children of the same age range as in the other used cohort (STAGING, described next) were included (age below 19 years old). From 1952 patients kept after genotype quality control (Table 1), 203 participants were > 19 years, thus 1749 were kept. Additionally, patients with ALL predisposition Down syndrome, not following the NOPHO-ALL 2008 treatment protocol, non-European, or without given consent to be part of the study were excluded (Figure 1). Finally, 1437 participants were included in this study.

This is a provisional file, not the final typeset article

**Cross-cancer adult-child PRS**

Information was available for age at the time of diagnosis; sex; body-mass index (BMI); white blood cell count (WBC); DNA index (DI); immunophenotype; central nervous system (CNS) status [CNS1: no cerebrospinal fluid (CSF) blasts; CNS2:  $<5$  leukocytes/ $\mu$ l CSF with blasts; CNS3:  $\geq 5$  leukocytes/ $\mu$ l with blasts or signs of CNS involvement]; CNS involvement type [only cells in CSF; cranial nerve palsy; clinical signs of CNS leukemia/mass on magnetic resonance imaging (MRI); combination of all]; cytogenetic information [dic(9;20); iAMP21; *MLL* 11q23; t(1;19); t(12;21); hyper- ( $>50$  chromosomes) or hypodiploidy ( $<44$  chromosomes); minimal residual disease (MRD) measured by PCR and flow cytometry as well as stratification in standard, intermediate or high-risk therapy after genetic analysis, day 15 (early response), day 29 (intermediate response) and day 79 (final stratification); and relapse leukaemia cells (bone marrow, testis, or CNS).

**STAGING cohort**

The STAGING cohort, included paediatric cancer patients diagnosed in Denmark. Details on patient collection in this Danish, prospective, nationwide study were previously described in a recent study (Byrjalsen et al. 2018). Here, 425 patients from the STAGING study at Rigshospitalet (Copenhagen University Hospital, Denmark) were included. Information was available for age at the time of diagnosis, sex, cancer type [solid, CNS, hematological], and cancer predisposition syndromes (CPSs) assessed according to Jongmans criteria (Jongmans et al. 2016).

**Duplicated samples**

A genotype comparison was made between 52 duplicate samples, i.e., present in both STAGING (whole-genome sequencing) and NOPHO (genotype imputation) cohort, using the 4227 genomic positions genotype included in the PRSs. For all the duplicates tested, the similarity percentage was  $> 95\%$  (Figure 2).

**NOPHO: Quality Control, Genotyping, and Genotype Imputation****Quality control**

The genotype data was genotyped using Illumina InfiniumOmni 2.5 Exome chips, but as the samples were received in batches over time, not all used the same version of the SNP chip. Therefore, the batches were strand aligned and the quality control (QC) was done first individually, followed by a few extra QC steps on the merged dataset. The preprocessing of the genotype data was performed using PLINK version 1.90beta3 (Chang et al. 2015). More details in Supplementary note 1 and Supplementary table 1.

**Imputation**

The genotypes were imputed after QC using SHAPEIT v2 (Delaneau, Marchini, and Zagury 2012) for phasing and IMPUTE v2.3.2 (Howie, Donnelly, and Marchini 2009) for imputation with default parameters. As a reference panel, 1000 Genomes Project (Phase 3) ("A Global Reference for Human Genetic Variation" 2015) was used. Post-imputation quality was based on the imputation accuracy that is provided by IMPUTE v2.3.2 (Howie, Donnelly, and Marchini 2009) info score. This metric takes values between 0 and 1, where higher values indicate higher imputation certainty and 1 implying perfect imputation. The calling of SNPs for the imputed data was done using an info score of 0.60. Afterwards, the genotype probabilities output by IMPUTE v.2.3.2 were converted into hard genotype calls using the

**Cross-cancer adult-child PRS**

genotype with the highest likelihood. This was done in PLINK version 1.90beta3(Chang et al. 2015) using the command line “--hard-call-threshold 0.49”.

**STAGING: whole-genome sequencing analysis****DNA extraction, library preparation, and whole-genome sequencing**

Genomic DNA was isolated from peripheral blood samples. Whole-genome sequencing (WGS) was performed by the Beijing Genomics Institute (Hong Kong, China) using the HiSeqX platform (Illumina, San Diego, CA, USA) for 309 participants and by the Center for Genomic Medicine at Rigshospitalet (Copenhagen, Denmark) using the MiSeq Illumina platform (Illumina, San Diego, CA, USA) for 116 participants (N=425). After excluding 57 non-European participants, 368 participants were included.

**Variant calling and filtering**

For each sample, mapping to the human reference genome hg19 using BWA algorithm, removal of read duplicates, realignment around insertions and deletions, and base-score recalibration and variant calling were done using Sentieon DNaseq software (Sentieon version 201808.03). Sentieon Haplotyper algorithm with option --emit\_mode gvcf was used to generate a variant call format (VCF) file per sample, or in this case, a gVCF file, including non-variant positions. Afterwards, a joint variant calling was performed using Sentieon GVCFTyper algorithm. Only bases above Q30 were kept in the VCF files. For individual-level SNPs, a depth of 10 was applied using vcftools software (version 0.1.16) (Danecek et al. 2011).

VCF files were filtered to include only variants referred in the PRSs using vcftools software (version 0.1.16) (Danecek et al. 2011).

**Geographic variation**

For both STAGING and NOPHO cohorts, multidimensional scaling was used to identify ethnic outliers using PLINK version 1.90beta3(Chang et al. 2015). Datasets were filtered for minor allele frequency (MAF) > 0.01, human leukocyte antigen (HLA) region was removed and only autosomes were kept. Patients were defined as being of non-European ancestry when deviating more than 4 standard deviations (SDs) from the EU panel mean value in any of the first four genomic components (Figure 3). 1000Genomes data was used as reference for checking population ethnicities. A total of 251 participants were removed from further analysis. From these, 199 were from the NOPHO cohort, while 57 were from the STAGING cohort. Five patients were duplicated in NOPHO and STAGING.

**Polygenic risk score**

We have performed PRS analysis on two childhood cancer cohorts, NOPHO and STAGING, based on two large GWAS of adult breast (Mavaddat et al. 2019) and colorectal (Huyghe et al. 2019) cancer. Both of these studies included samples of European ancestry. Effect size metrics were extracted from the respective manuscripts (Mavaddat et al. 2019) (Huyghe et al. 2019).

In both studies described below, the ultimate goal was to construct a PRS as a weighted sum of the allele dosage (Mavaddat et al. 2019) or the number of risk alleles (Huyghe et al. 2019) carried by an individual, using the per-allele log OR for each variant as weights:  $PRS = \sum_{i=1}^n \beta_i \chi_i$ , where  $\beta_i$  is the per-allele log odds ratio (OR) associated with SNP  $i$ ,

This is a provisional file, not the final typeset article

## Cross-cancer adult-child PRS

$\chi_k$  is the allele dosage (Mavaddat et al. 2019) or the number of risk alleles (Huyghe et al. 2019) for SNP  $k$ , and  $n$  is the total number of SNPs included in the PRS.

**Breast cancer**

For the breast cancer study, the authors applied two different approaches to develop a PRS using a large cohort of 94,075 cases and 75,017 controls of European ancestry from 69 studies. The PRSs were further validated in an independent set of 11,428 cases and 18,323 controls (10 prospective studies) and 190,040 women (UK Biobank) (Mavaddat et al. 2019).

Here, we have used the best performing PRS consisting of 313 SNPs, further referred to in this study as PRS313. Briefly, this was obtained using a “hard-thresholding” approach, based on a series of stepwise regression analysis that retained SNPs significantly associated with overall breast cancer (cases vs controls) or ER-negative disease (cases only) (305 SNPs) using a p-value  $< 10^{-5}$  in the largest available genome-wide association dataset. The SNP effect sizes were estimated in a single logistic regression model. SNPs associated with ER-position (p-value  $< 10^{-6}$ ) but not with overall breast cancer (p-value  $< 10^{-5}$ ) were added (6 SNPs). Two rarer variants (*BRCA2* p.Lys3326X and *CHEK2* p.Ile157Tyr) which are established to confer a moderate risk of breast cancer were also added.

Another approach for the PRS development was penalized regression using least absolute shrinkage and selection operator (LASSO). Here, the authors pre-selected for inclusion SNPs with p-value  $< 0.001$  in overall breast cancer or ER-negative disease in the training set, and *BRCA2* p.Lys3326X and *CHEK2* p.Ile157Thr were added. For overall breast cancer, variable selection and parameter estimation was carried out selecting the best penalty parameter (lambda) in the validation set. In this approach the PRS consisted of 3820 SNPs, further referred to in this study as PRS3820.

Both approaches are described in more detail at (Mavaddat et al. 2019).

**Colorectal cancer**

For the colorectal cancer (CRC), the authors performed the largest and most comprehensive WGS study and GWAS meta-analysis for CRC so far combining data from three consortia: the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), the Colorectal Cancer Transdisciplinary Study (CORECT) and the Colon Cancer Family Registry (CCFR), and including participants from European and East Asian ancestry. The resulting stage 1 meta-analysis informed the design of a custom genotyping array which was used to genotype 12,007 CRC cases and 12,000 controls. They combined this with additional new or existing GWAS data imputed to the Haplotype Reference Consortium panel, resulting in a stage 2 meta-analysis of up to 23,262 CRC cases and 38,296 controls. They report new association signals discovered through their two-stage custom genotyping experiment and replicating at the Bonferroni significance threshold ( $P < 7.8 \times 10^{-6}$ ). This list of 8 new loci includes the first rare variant association for sporadic CRC. Next, the authors performed a combined (stage 1 + stage 2) meta-analysis of up to 125,478 CRC samples (58,131 cases and 67,347 controls) and reported all distinct association signals passing the genome-wide significance threshold of p-value  $< 5 \times 10^{-8}$  in the combined meta-analysis. This list comprises 30 new loci, including all eight loci discovered through the custom genotyping experiment, and 10 additional signals discovered through conditional meta-analysis. 55 previously described autosomal risk variants that showed evidence for colorectal were added (Huyghe et al. 2019). A total of 95 SNPs were included to build this PRS, further referred to in this study as PRS95.



## Cross-cancer adult-child PRS

**Statistical tests**

The utility of the PRSs for patient stratification was assessed using the different available clinical features. Samples were stratified in accordance with the PRS percentile: bottom (0-20), medium (20-90), and top (90-100). First, statistical significance of mean PRS between different groups in each clinical feature was accessed using independent two-sided t-test. For the continuous variables, correlation between PRS and variable was accessed using Pearson (normal distributed) or Spearman (non-normal distribution) correlations. The OR for the different PRS-based stratified groups were calculated using Fisher's exact test (categorical variables) and logistic regression (continuous variables).

**RESULTS****PRS distribution in NOPHO and STAGING cohorts**

PRS was calculated for each patient and the overall cohort distribution approximated a normal distribution. PRSs distributions were very similar between NOPHO and STAGING cohorts (Figure 4). Some of the SNPs included in the PRS were not present in our dataset after imputation (due to info score  $< 0.60$ ), or WGS (due to quality  $< 30$  or alleles present being different from the ones referred in the (Mavaddat et al. 2019) and (Huyghe et al. 2019) (Table 2). Missing data at individual level (due to "--hard-call-threshold for NOPHO, and depth for STAGING) was removed based on samples, so all possible relevant SNPs could be kept. Only for PRS3820 in the NOPHO cohort, missing data was removed based on SNPs, otherwise only 40 samples with complete genomic data would have remained. Plots for all other variables (statistically non-significant results) can be found in Supplementary material online.

**Risk score for NOPHO cohort**

PRS313/PRS3820 and age showed to have a very weak correlation (Figure 5A-B). However, when we stratified the population according to PRS percentiles, we found significant results, as described below.

**Breast cancer PRS313**

*Age at the time of diagnosis:* The OR of finding younger patients in the top percentile was 1.105 (95% CI 1.066-1.146, p-value  $7.222 \times 10^{-8}$ ) and 1.418 (95% CI 1.359-1.480, p-value  $5.363 \times 10^{-58}$ ) compared with the bottom or the other percentiles, respectively.

**Breast cancer PRS3820**

*Age at the time of diagnosis:* The OR of finding younger patients in the top percentile was 1.09 (95% CI 1.059-1.122, p-value  $3.871 \times 10^{-9}$ ) and 1.449 (95% CI 1.354-1.449, p-value  $4.422 \times 10^{-84}$ ) compared with the bottom or the other percentiles, respectively.

*Risk-group stratification at day 15 (early response):* Even though mean PRS was not significantly different between participants in the two different groups (Figure 6A); the OR of finding a patient of the high-risk group in the top percentile was 1.507 (95% CI 0.904-2.496, p-value 0.1027) and 1.444 (95% CI 0.941-2.179, p-value 0.078) compared with the bottom or the other percentiles, respectively (Figure 6D).

*MRD measured by PCR at days 29:* Mean PRS was not significantly different between participants in the two different groups (Figure 6B); however the OR of finding a patient with a higher MRD measured by PCR on day 29 in the top

## Cross-cancer adult-child PRS

percentile was 2.252 (95% CI 0.954-5.350, p-value 0.04) and 2.055 (95% CI 1.007-4.126, p-value 0.03) compared with the bottom and the others percentiles, respectively (Figure 6E).

*MRD measured by PCR at days 79:* Mean PRS was significantly different between participants in the two different groups (p-value 0.04) (Figure 6C). The OR of finding a patient with a higher MRD measured by PCR on day 79 in the top percentile was 2.522 (95% CI 0.428-10.368, p-value 0.163) compared with the bottom percentile (Figure 6F).

**Risk score for STAGING cohort**

All PRSs and age showed to have a very weak correlation (Figure 5C-E). However, when we stratified the population according to PRS percentiles, we found significant results, as described below.

**Breast cancer PRS313**

*Age at the time of diagnosis:* The OR of finding younger patients in the top percentile was 1.111 (95% CI 1.021-1.111, p-value 0.003) and 1.257 (95% CI 1.198-1.318, p-value  $5 \times 10^{-21}$ ) compared with the bottom or the other percentiles, respectively.

*Diagnosis:* Mean PRS was not significantly different between participants in the two different groups (Figure 7A); however the OR of finding patients with CNS tumour in the top percentile was 4.831 (95% CI 1.201-22.388, p-value 0.0152) and 2.702 (95% CI 0.817-10.412, p-value 0.112) (top vs bottom and top vs others percentile, respectively) compared with the bottom or the other percentiles, respectively (Figure 7B).

**Breast cancer PRS3820**

*Age at the time of diagnosis:* The OR of finding younger patients in the top percentile was 1.103 (95% CI 1.049-1.103, p-value 0.0001) and 1.337 (95% CI 1.258-1.421, p-value  $7.24 \times 10^{-21}$ ) compared with the bottom or the other percentiles, respectively.

**Colorectal cancer PRS95**

*Age at the time of diagnosis:* The OR of finding younger patients in the top percentile was 1.063 (95% CI 1.018-1.111, p-value 0.006) and 1.261 (95% CI 1.202-1.324, p-value  $4.678 \times 10^{-21}$ ) compared with the bottom or the other percentiles, respectively.

**DISCUSSION**

In this study, we found few significant PRS313, PRS3820, and PRS95-based stratification on the two childhood cancer cohorts. Those with higher genetic risk were also the younger patients, except for the PRS95 in the NOPHO cohort, where no difference was found. Additionally, using a PRS3820-based stratification on the NOPHO cohort, we found that higher PRS was also associated with high-risk group stratification at day 15 and higher MRD levels following induction (day 29) and consolidation (day 79). For the STAGING cohort, a significant stratification was found based on the PRS313, where patients with CNS tumours had a higher genetic risk score than patients with solid tumours. However, for the majority of the variables, no significant stratification was found.

Genetic basis and cancer aetiology in children is still largely unknown and new strategies to identify paediatric cancer genes are much needed. These findings, being further validated, could contribute to a better understanding of disease

**Cross-cancer adult-child PRS**

aetiology trajectory in childhood cancer and for the increasing hope that PRS may contribute to precision medicine in enhancing risk assessment.

Furthermore, this would indicate some shared genetic basis of breast, colorectal and childhood cancer, however, further work would need to be done to understand the exact mechanism shared between these cancers in order to understand better disease progression.

**Limitations and further work**

In this study interactions between features were not accounted, when considering the risk of later outcomes. Giving a specific example, relapse may be very much dependent on both PRS and treatment, and here we simply considered PRS. Additionally, we have not accounted for any cofounders. We know for example that WBC is extremely heterogeneous and associated with age (immune system maturation), karyotype, and immunophenotype (Vaitkeviciene et al. 2011)(Vrooman et al. 2018)(Toft et al. 2013).

It would have been interesting to compare cancer prognosis in specific subtypes, for example, hyperdiploidy group stratified by immunophenotype (NOPHO cohort), haematological vs solid and CNS tumours (STAGING cohort) or perform this analysis in different age groups to understand if PRS-based stratification would work better in certain groups. For the age groups, other studies have stated that PRS may be a better predictor at different ages (Isgut et al. 2021)(Damask et al. 2020).

It is also worth noting that we expected to observe a higher risk score associated with the worst prognosis in childhood cancer, however, it could also go the other way around, where being predisposed to leukaemia makes one less predisposed to other phenotypes. Indeed, a study has shown that survivors of breast cancer who did not develop breast cancer in the future had higher probability of developing other late side effects and diseases, such as leukaemia and CNS tumours (Wang et al. 2018).

The PRS-based stratification was only performed on subjects of European ancestry, thus further studies of ancestry-specific genetic architectures are needed to understand if this could generalise across population substructures. Here, we only had large numbers of patients of European ancestry, but it would be interesting to explore the PRS-based stratification on the non-European samples removed from the initial analysis (n=251).

As an additional study, it would be interesting to compare the PRS distribution of the childhood cancer cohort with a healthy cohort to investigate if PRS distribution is different, and if a higher PRS is observed in the childhood cancer cohort. As a more distant future study, one could also follow-up on the development of breast and colorectal cancer in the childhood cancer survivors of the NOPHO and STAGING cohort and determine if the patients in the higher percentile were indeed the ones who developed breast and colorectal cancer later in life.

**Cross-cancer adult-child PRS****DATA AVAILABILITY STATEMENT**

Datasets in the current study can be made available on reasonable request to the corresponding author, in accordance with patient approvals and GDPR.

**ETHICS STATEMENT**

Verbal and written consent was obtained. The database containing phenotype data was approved by the regional ethical review board of The Capital Region of Denmark (H-2-2010-022), the Danish Data Protection Authorities (j.nr.: 2012-58-0004), and by relevant regulatory authorities in all participating countries (NOPHO cohort). Ethical approval was obtained through the regional scientific ethical committee (the Ethical Scientific Committees for the Capital Region, H-15016782) and the Danish Data Protection Agency (RH-2016-219, I-Suite no: 04804). All parents/guardians and patients 15 years or older gave formal written consent to participation (STAGING cohort).

**CONFLICT OF INTEREST**

RG is employed with Novo Nordisk Research Centre Oxford since February 2020. The other authors have no conflicts of interest to disclose.

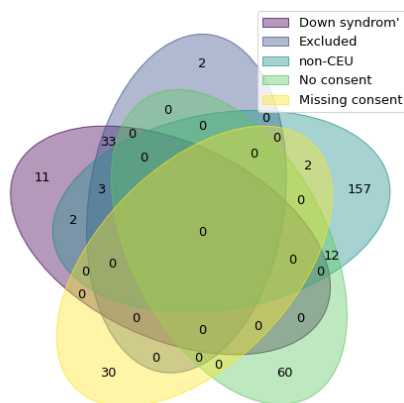
**FUNDING**

This study was financially supported by the Danish Childhood Cancer Foundation, University Hospital Rigshospitalet, Danish Cancer Society, Otto Christensen's Foundation, Swedish Childhood Cancer Foundation, Nordic Cancer Union, Novo Nordisk Foundation, European Union's Interregional Öresund–Kattegat–Skagerrak and Idella Foundation. No funding sources played a role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

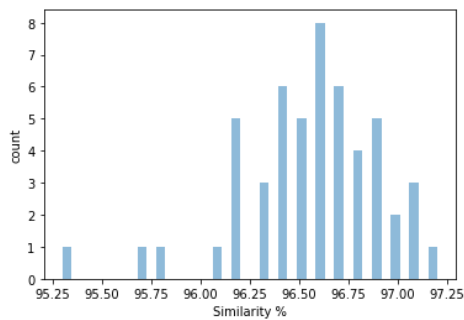
**ACKNOWLEDGMENTS**

We thank all the researchers who scrutinized patient files and completed phenotype questionnaires, and also the organizational support from the research staff at Bonkolab, at the University Hospital Rigshospitalet.

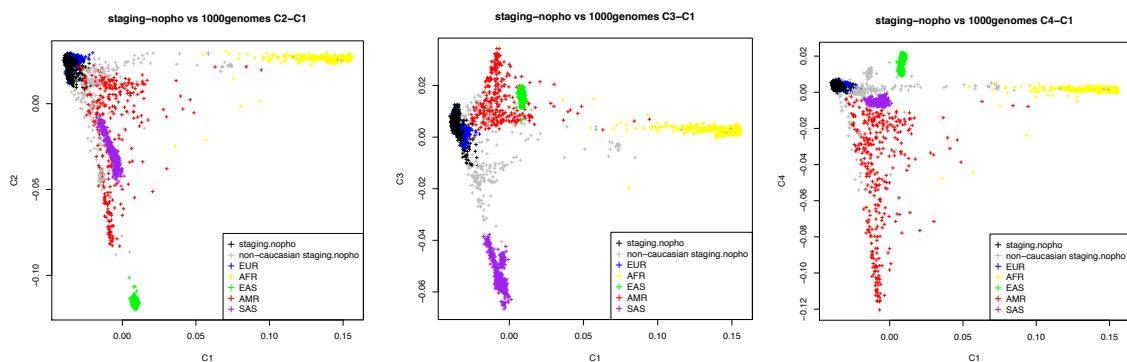
### Cross-cancer adult-child PRS



**Figure 1:** Venn diagram representing the number of participants in each excluded group.

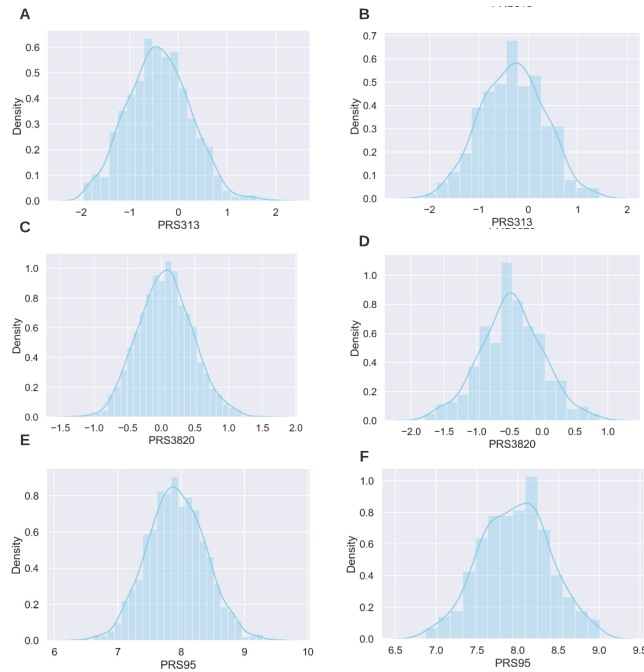


**Figure 2:** Percentage similarity between pairs of duplicated samples.

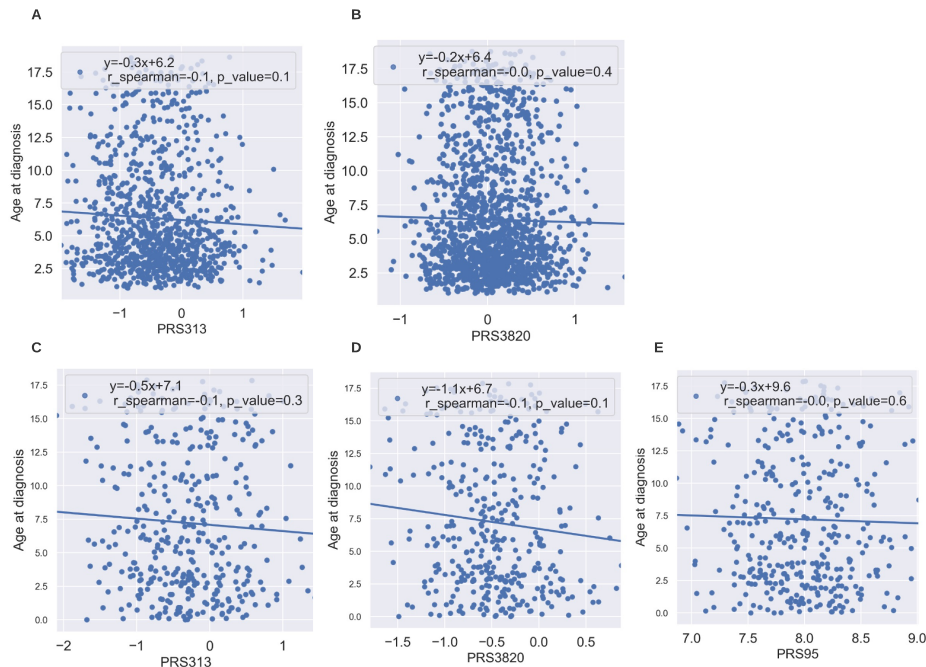


**Figure 3:** Multidimensional scaling performed in both NOPHO and STAGING cohorts. A total of 185, 29, 27, and 10 participants were removed in components 1, 2, 3 and 4, respectively.

## Cross-cancer adult-child PRS

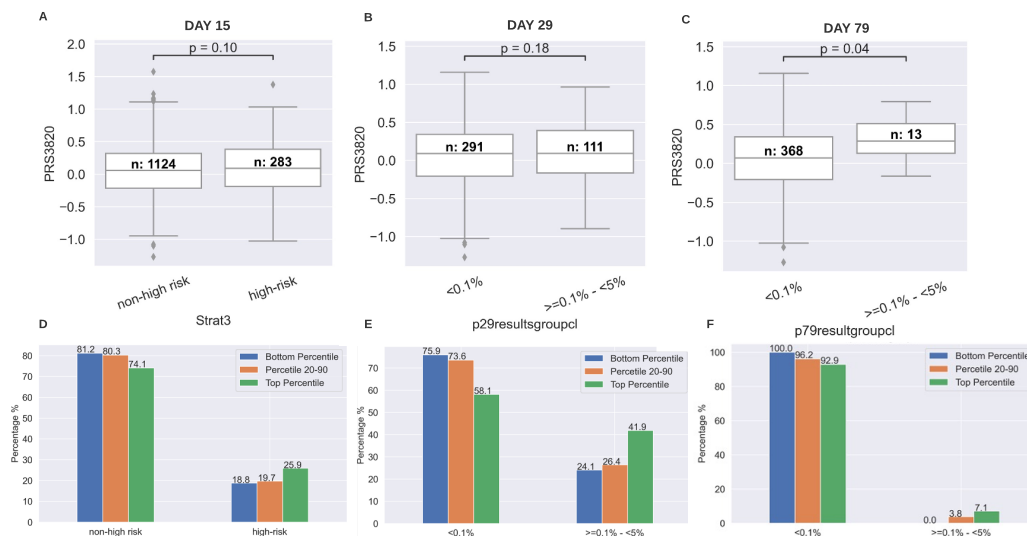


**Figure 4:** Polygenic risk score distribution for NOPHO (A: 952 samples, 313 SNPs; C: 1437 samples, 2663 SNPs; E: 1351 samples, 93 SNPs) and STAGING (B: 368 samples, 307 SNPs; D: 368 samples, 3763 SNPs; F: 368 samples, 94 SNPs) for the PRS313 (A,B), PRS3820 (C,D), and PRS95 (E,F).

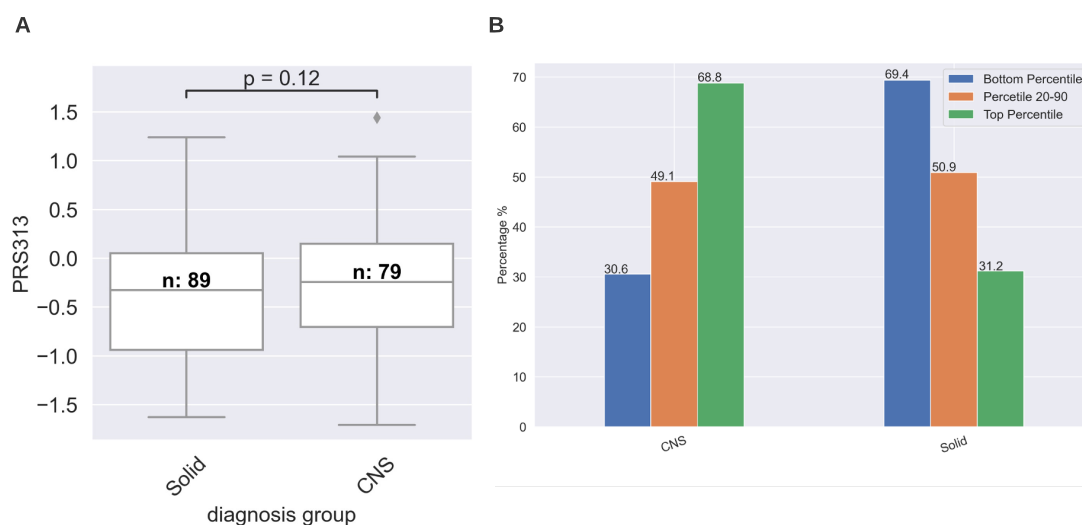


**Figure 5:** Correlation between age at the time of diagnosis (in years) and PRS. A-B: NOPHO cohort, PRS313 and PRS3820, respectively; C-E: STAGING cohort, PRS313, PRS3820 and PRS95, respectively.

## Cross-cancer adult-child PRS



**Figure 6:** Boxplots of PRS3820 values by risk-stratification at day 15 (A), MRD values at day 29 (B), and MRD values at day 79 (C) for the NOPHO cohort. The respective stratification in top, middle and bottom percentile is represented below in the bar plots where we can compare the percentage of samples in each percentile within groups (D-F).



**Figure 7:** Boxplots of PRS313 values by diagnosis group (A) for the STAGING cohort and the respective stratification in top, middle and bottom percentile in the bar plot in the side, where we can compare the percentage of samples in each percentile within groups (B).

## Cross-cancer adult-child PRS

**Table 1** Preprocessing of genotype data batches. This table depicts the number of patients and SNPs in each batch of genotype data during preprocessing.

	Patients	SNPs
<b>Remove duplicate samples and SNPs</b>	1,966	2,534,414
<b>QC step: SNP and sample missingness (2%)</b>	1,966	2,146,366
<b>QC step: Relatedness (keep identity-by-descent <math>\leq 0.1875</math>)</b>	1,961	2,146,366
<b>QC step: Remove individuals with excess homozygosity</b>	1,952	2,146,366

QC: quality control, SNP: single nucleotide polymorphism

**Table 2** Markers removed after imputation or WGS due to low quality or non-relevant alleles present.

		NOPHO (SNP array)	STAGING (WGS)
<b>PRS<sub>313</sub> SNPs</b>	Below info score 0.60	-	-
	Below Q30	-	3
	Different alleles	0	3
<b>PRS<sub>3820</sub> SNPs</b>	Below info score 0.60	2	-
	Below Q30	-	15
	Different alleles	0	42
<b>PRS<sub>95</sub> SNPs</b>	Below info score 0.60	2	-
	Below Q30	-	1
	Different alleles	0	0

SNP: single nucleotide polymorphism, WGS: whole-genome sequencing



## Cross-cancer adult-child PRS

## SUPPLEMENTARY MATERIAL

## Supplementary note 1 Quality control description.

For the individual batches, strand alignment to the PLUS strand was performed using the William Rayner (WR) strand files as reference (<https://www.well.ox.ac.uk/~wrayner/strand/index.html>) and human genome build 37. A check was performed to ensure that the original files were TOP strand oriented and that the alleles matched the TOP strand reference file. SNPs that were not present in the reference strand file and the SNPs found mapped to multiple locations according to the WR .multiple file, were excluded.

*Sample and SNP duplicates check:* Each batch was checked for duplicate samples, where the sample with lower SNP missingness was kept. A check for duplicate SNPs was performed based on SNP location and alleles, where SNP identifiers were merged and remaining duplicates excluded.

*Quality control:* The batches were subject to individual QC steps, as listed below:

- i) Missingness, where SNPs and samples with more than 2% missing data were removed.
- ii) Sex check, where samples with mismatch between genetic and clinical sex were removed. Genetic sex was determined by the inbreeding coefficient, where  $F < 0.2$  are female and  $F > 0.8$  are male.
- iii) Check for excess heterozygosity and homozygosity, where samples are removed based on the inbreeding coefficient as a measure of excess heterozygosity and homozygosity (more than 4 standard deviations away from the mean of  $F$ ). This removes possibly contaminated samples and population substructure.
- iv) Relatedness check, where any related samples or potentially remaining duplicate samples are removed.

Following the individual batch pre-processing, the genotype batches were merged and checked for duplicate SNPs and samples across batches. The merged set was then QC'ed by the steps listed below:

- i) Missingness, where SNPs with more than 2% missing data were excluded, before excluding samples with more than 2% missing SNPs.
- ii) Relatedness check, where any related samples or potentially remaining duplicate samples are removed.
- iii) Check for excess homozygosity, in order to remove potential population substructure in the merged dataset.

**Supplementary table 1** Overview of pre-processing and quality control (QC) on individually genotyped batches. The batches were genotyped using different chips, why these were initially quality controlled separately.

Batch	A		B		C		D	
Genotyping chip	HumanOmni2-5Exome-8-v1-1-A		InfiniumOmni2-5Exome-8v1-3 A1		InfiniumOmni2-5Exome-8v1-3 A1		InfiniumOmni2-5Exome-8v1-4 A1	
	Patients	SNPs	Patients	SNPs	Patients	SNPs	Patients	SNPs
NOPHO genotypes	1,519	2,546,527	334	2,612,357	362	2,612,357	135	2,617,655
Strand alignment	1,519	2,498,653	334	2,560,920	362	2,560,920	135	2,566,350
Remove duplicate samples and SNPs	1,452	2,453,541	320	2,514,171	362	2,515,787	135	2,523,930
QC step 1: SNP and sample missingness	1,355	2,383,820	241	2,314,622	345	2,434,130	134	2,462,363
QC step 2: Sex mismatch	1,343	2,383,820	232	2,314,622	340	2,434,130	134	2,462,363
QC step 3: Remove individuals with excess heterozygosity and homozygosity	1,321	2,383,820	224	2,314,622	334	2,434,130	133	2,462,363
QC step 4: Relatedness (keep identity-by-descent $\leq 0.1875$ )	1,313	2,383,820	224	2,314,622	331	2,434,130	133	2,462,363

QC: quality control, SNP: single nucleotide polymorphism, NOPHO: Nordic Society for Pediatric Hematology and Oncology.

This is a provisional file, not the final typeset article

## Cross-cancer adult-child PRS

## References

- “A Global Reference for Human Genetic Variation.” 2015. *Nature* 526 (7571): 68–74.  
<https://doi.org/10.1038/nature15393>.
- Byrjalsen, Anna, Thomas V O Hansen, Ulrik K Stoltze, Mana M Mehrjouy, Nanna Moeller Barnkob, Lisa L Hjalgrim, René Mathiasen, et al. 2020. “Nationwide Germline Whole Genome Sequencing of 198 Consecutive Pediatric Cancer Patients Reveals a High Incidence of Cancer Prone Syndromes.” *PLoS Genetics* 16 (12): e1009231.  
<https://doi.org/10.1371/journal.pgen.1009231>.
- Byrjalsen, Anna, Ulrik Stoltze, Karin Wadt, Lisa Lyngsie Hjalgrim, Anne-Marie Gerdes, Kjeld Schmiegelow, and Ayo Wahlberg. 2018. “Pediatric Cancer Families’ Participation in Whole-Genome Sequencing Research in Denmark: Parent Perspectives.” *European Journal of Cancer Care* 27 (6): e12877. <https://doi.org/10.1111/ecc.12877>.
- Chang, Christopher C., Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4 (1). <https://doi.org/10.1186/s13742-015-0047-8>.
- Damask, Amy, P. Gabriel Steg, Gregory G. Schwartz, Michael Szarek, Emil Hagström, Lina Badimon, M. John Chapman, et al. 2020. “Patients With High Genome-Wide Polygenic Risk Scores for Coronary Artery Disease May Receive Greater Clinical Benefit From Alirocumab Treatment in the ODYSSEY OUTCOMES Trial.” *Circulation* 141 (8): 624–36. <https://doi.org/10.1161/CIRCULATIONAHA.119.044434>.
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. <https://doi.org/10.1093/bioinformatics/btr330>.
- Delaneau, Olivier, Jonathan Marchini, and Jean-François Zagury. 2012. “A Linear Complexity Phasing Method for Thousands of Genomes.” *Nature Methods* 9 (2): 179–81. <https://doi.org/10.1038/nmeth.1785>.
- Frandsen, Thomas Leth, Mats Heyman, Jonas Abrahamsson, Kim Vettenranta, Ann Åsberg, Goda Vaitkeviciene, Kaie Pruunsild, et al. 2014. “Complying with the European Clinical Trials Directive While Surviving the Administrative Pressure – An Alternative Approach to Toxicity Registration in a Cancer Trial.” *European Journal of Cancer* 50 (2): 251–59. <https://doi.org/10.1016/j.ejca.2013.09.027>.
- Graff, Rebecca E., Taylor B. Cavazos, Khanh K. Thai, Linda Kachuri, Sara R. Rashkin, Joshua D. Hoffman, Stacey E. Alexeeff, et al. 2021. “Cross-Cancer Evaluation of Polygenic Risk Scores for 16 Cancer Types in Two Large Cohorts.” *Nature Communications* 12 (1): 970. <https://doi.org/10.1038/s41467-021-21288-z>.
- Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. “A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies.” Edited by Nicholas J. Schork. *PLoS Genetics* 5 (6): e1000529. <https://doi.org/10.1371/journal.pgen.1000529>.
- Huyghe, Jeroen R., Stephanie A. Bien, Tabitha A. Harrison, Hyun Min Kang, Sai Chen, Stephanie L. Schmit, David V. Conti, et al. 2019. “Discovery of Common and Rare Genetic Risk Variants for Colorectal Cancer.” *Nature Genetics* 51 (1): 76–87. <https://doi.org/10.1038/s41588-018-0286-6>.
- Isgut, Monica, Jimeng Sun, Arshed A. Quyyumi, and Greg Gibson. 2021. “Highly Elevated Polygenic Risk Scores Are Better Predictors of Myocardial Infarction Risk Early in Life than Later.” *Genome Medicine* 13 (1): 13. <https://doi.org/10.1186/s13073-021-00828-8>.
- Jia, Guochong, Yingchang Lu, Wanqing Wen, Jirong Long, Ying Liu, Ran Tao, Bingshan Li, Joshua C Denny, Xiao-Ou Shu, and Wei Zheng. 2020. “Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals

## Cross-cancer adult-child PRS

- for Eight Common Cancers.” *JNCI Cancer Spectrum* 4 (3). <https://doi.org/10.1093/jncics/pkaa021>.
- Jiang, Xia, Hilary K. Finucane, Fredrick R. Schumacher, Stephanie L. Schmit, Jonathan P. Tyrer, Younghun Han, Kyriaki Michailidou, et al. 2019. “Shared Heritability and Functional Enrichment across Six Solid Cancers.” *Nature Communications* 10 (1): 431. <https://doi.org/10.1038/s41467-018-08054-4>.
- Johnston, W.T., Friederike Erdmann, Robert Newton, Eva Steliarova-Foucher, Joachim Schüz, and Eve Roman. 2021. “Childhood Cancer: Estimating Regional and Global Incidence.” *Cancer Epidemiology* 71 (April): 101662. <https://doi.org/10.1016/j.canep.2019.101662>.
- Jongmans, Marjolijn C.J., Jan L.C.M. Loeffen, Esmé Waanders, Peter M. Hoogerbrugge, Marjolijn J.L. Ligtenberg, Roland P. Kuiper, and Nicoline Hoogerbrugge. 2016. “Recognition of Genetic Predisposition in Pediatric Cancer Patients: An Easy-to-Use Selection Tool.” *European Journal of Medical Genetics* 59 (3): 116–25. <https://doi.org/10.1016/j.ejmg.2016.01.008>.
- Kachuri, Linda, Rebecca E. Graff, Karl Smith-Byrne, Travis J. Meyers, Sara R. Rashkin, Elad Ziv, John S. Witte, and Mattias Johansson. 2020. “Pan-Cancer Analysis Demonstrates That Integrating Polygenic Risk Scores with Modifiable Risk Factors Improves Risk Prediction.” *Nature Communications* 11 (1): 6084. <https://doi.org/10.1038/s41467-020-19600-4>.
- Lichtenstein, Paul, Niels V. Holm, Pia K. Verkasalo, Anastasia Iliadou, Jaakko Kaprio, Markku Koskenvuo, Eero Pukkala, Axel Skytthe, and Kari Hemminki. 2000. “Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland.” *New England Journal of Medicine* 343 (2): 78–85. <https://doi.org/10.1056/NEJM200007133430201>.
- Lindström, Sara, Hilary Finucane, Brendan Bulik-Sullivan, Fredrick R. Schumacher, Christopher I. Amos, Rayjean J. Hung, Kristin Rand, et al. 2017. “Quantifying the Genetic Correlation between Multiple Cancer Types.” *Cancer Epidemiology Biomarkers & Prevention* 26 (9): 1427–35. <https://doi.org/10.1158/1055-9965.EPI-17-0211>.
- Lu, Karen H., Marie E. Wood, Molly Daniels, Cathy Burke, James Ford, Noah D. Kauff, Wendy Kohlmann, et al. 2014. “American Society of Clinical Oncology Expert Statement: Collection and Use of a Cancer Family History for Oncology Providers.” *Journal of Clinical Oncology* 32 (8): 833–40. <https://doi.org/10.1200/JCO.2013.50.9257>.
- Mavaddat, Nasim, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P. Tyrer, et al. 2019. “Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes.” *The American Journal of Human Genetics* 104 (1): 21–34. <https://doi.org/10.1016/j.ajhg.2018.11.002>.
- Mucci, Lorelei A., Jacob B. Hjelmborg, Jennifer R. Harris, Kamila Czene, David J. Havelick, Thomas Scheike, Rebecca E. Graff, et al. 2016. “Familial Risk and Heritability of Cancer Among Twins in Nordic Countries.” *JAMA* 315 (1): 68. <https://doi.org/10.1001/jama.2015.17703>.
- Parsons, D Williams, Angshumoy Roy, Yaping Yang, Tao Wang, Sarah Scollon, Katie Bergstrom, Robin A Kerstein, et al. 2016. “Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors.” *JAMA Oncology* 2 (5): 616–24. <https://doi.org/10.1001/jamaoncol.2015.5699>.
- Rank, Cecilie Utke, Nina Toft, Ruta Tuckuviene, Kathrine Grell, Ove Juul Nielsen, Thomas Leth Frandsen, Hanne Vibeke Hansen Marquart, et al. 2018. “Thromboembolism in Acute Lymphoblastic Leukemia: Results of NOPHO ALL2008 Protocol Treatment in Patients Aged 1 to 45 Years.” *Blood* 131 (22): 2475–84. <https://doi.org/10.1182/blood-2018-01-827949>.
- Sampson, Joshua N., William A. Wheeler, Meredith Yeager, Orestis Panagiotou, Zhaoming Wang, Sonja I. Berndt, Qing

## Cross-cancer adult-child PRS

- Lan, et al. 2015. "Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types." *Journal of the National Cancer Institute* 107 (12): djv279. <https://doi.org/10.1093/jnci/djv279>.
- Spector, Logan G., Nathan Pankratz, and Erin L. Marcotte. 2015. "Genetic and Nongenetic Risk Factors for Childhood Cancer." *Pediatric Clinics of North America* 62 (1): 11–25. <https://doi.org/10.1016/j.pcl.2014.09.013>.
- Toft, Nina, Henrik Birgens, Jonas Abrahamsson, Per Bernell, Laimonas Griškevičius, Helene Hallböök, Mats Heyman, et al. 2013. "Risk Group Assignment Differs for Children and Adults 1-45 Yr with Acute Lymphoblastic Leukemia Treated by the NOPHO ALL-2008 Protocol." *European Journal of Haematology* 90 (5): 404–12. <https://doi.org/10.1111/ejh.12097>.
- Toft, Nina, Henrik Birgens, Jonas Abrahamsson, Laimonas Griškevičius, Helene Hallböök, Mats Heyman, Tobias Wrenfeldt Klausen, et al. 2016. "Toxicity Profile and Treatment Delays in NOPHO ALL2008-Comparing Adults and Children with Philadelphia Chromosome-Negative Acute Lymphoblastic Leukemia." *European Journal of Haematology* 96 (2): 160–69. <https://doi.org/10.1111/ejh.12562>.
- Vaitkevičienė, Goda, Erik Forestier, Marit Hellebostad, Mats Heyman, Olafur G. Jonsson, Päivi M. Lähteenmäki, Susanne Rosthøj, Stefan Söderhäll, and Kjeld Schmiegelow. 2011. "High White Blood Cell Count at Diagnosis of Childhood Acute Lymphoblastic Leukaemia: Biological Background and Prognostic Impact. Results from the NOPHO ALL-92 and ALL-2000 Studies." *European Journal of Haematology* 86 (1): 38–46. <https://doi.org/10.1111/j.1600-0609.2010.01522.x>.
- Vrooman, Lynda M., Traci M. Blonquist, Marian H. Harris, Kristen E. Stevenson, Andrew E. Place, Sarah K. Hunt, Jane E. O'Brien, et al. 2018. "Refining Risk Classification in Childhood B Acute Lymphoblastic Leukemia: Results of DFCI ALL Consortium Protocol 05-001." *Blood Advances* 2 (12): 1449–58. <https://doi.org/10.1182/bloodadvances.2018016584>.
- Wang, Zhaoming, Qi Liu, Carmen L. Wilson, John Easton, Heather Mulder, Ti-Cheng Chang, Michael C. Rusch, et al. 2018. "Polygenic Determinants for Subsequent Breast Cancer Risk in Survivors of Childhood Cancer: The St Jude Lifetime Cohort Study (SJLIFE)." *Clinical Cancer Research* 24 (24): 6230–35. <https://doi.org/10.1158/1078-0432.CCR-18-1775>.
- Zhang, Jinghui, Michael F Walsh, Gang Wu, Michael N Edmonson, Tanja A Gruber, John Easton, Dale Hedges, et al. 2015. "Germline Mutations in Predisposition Genes in Pediatric Cancer." *The New England Journal of Medicine* 373 (24): 2336–46. <https://doi.org/10.1056/NEJMoa1508054>.

## Chapter 6

### Paper III: Hearing loss prediction

Sara L Garcia, Jakob Lauritsen, Bernadette K. Christiansen,  
Ida F. Hansen, Mikkel Bandak, Marlene D. Dalgaard, Gedske  
Daugaard, Ramneek Gupta

**Predicting hearing loss after cisplatin chemotherapy in  
testicular cancer patients.**

## Predicting hearing loss after cisplatin chemotherapy in testicular cancer patients

Sara L. Garcia\*, MSc; Jakob Lauritsen\*, MD; Bernadette K. Christiansen, BSc; Ida F. Hansen, BSc; Mikkel Bandak, MD; Marlene D. Dalgaard, PhD; Gedske Daugaard, DMSc; Ramneek Gupta, PhD

\*Sara L. Garcia and Jakob Lauritsen contributed equally to this work.

**Manuscript word count: body: 3155; abstract: 364**

### Article information

**Corresponding Author:** Jakob Lauritsen, MD, Department of Oncology, Copenhagen University Hospital, Copenhagen, Denmark ([Jakob.Lauritsen@regionh.dk](mailto:Jakob.Lauritsen@regionh.dk))

**Author Affiliations:** Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark (Garcia, Christiansen, Hansen, Dalgaard, Gupta); Department of Oncology, Copenhagen University Hospital, Copenhagen, Denmark (Lauritsen, Bandak, Daugaard); Department of Computational Biology, Novo Nordisk Research Centre Oxford, Oxford, United Kingdom (Gupta)

#### Author Contributions:

*Concept and design:* Lauritsen, Daugaard, Gupta.

*Acquisition, analysis, or interpretation of data:* Garcia, Lauritsen, Christiansen, Hansen, Bandak.

*Drafting of the manuscript:* Garcia, Lauritsen.

*Critical revision of the manuscript for important intellectual content:* All authors.

*Statistical Analysis:* Garcia, Christiansen, Hansen.

*Study supervision:* Daugaard, Gupta.

**Conflict of interest:** RG is employed with Novo Nordisk Research Centre Oxford since February 2020. The other authors have no conflicts of interest to disclose.

**Funding/support:** This study was supported by the Danish cancer society. Garcia was supported by Idella foundation.

**Role of the Funder/Sponsor:** The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit.

### Key Points

**Question:** What is the risk of developing hearing loss after cisplatin-based chemotherapy in testicular cancer patients?

**Findings:** In this case-control study of 433 testicular cancer patients, a predictive model that identified patients at high or low risk of developing hearing loss was built. Furthermore, a possible biological mechanism that may contribute to hearing loss is proposed.

**Meaning:** Identification of patients with a higher risk of developing hearing loss will allow to explore other treatment strategies to mitigate or prevent this common late-side effect.

**Abstract**

**Importance:** In industrialized countries, testicular cancer is the most common solid tumor in men between 20-40 years old. Even though this is a curable cancer, the long-term side effects of chemotherapy are worrisome as they may have severe impact on quality of life. Here in this research study, we focus on hearing loss.

**Objective:** To identify testicular cancer patients at higher risk of hearing loss after cisplatin-based chemotherapy.

**Design:** Case-control study comprising testicular cancer survivors treated with cisplatin-based chemotherapy from 1984 until 2007.

**Setting:** Data was collected from the Danish Testicular Cancer-Late cohort in October 2014.

**Participants:** Clinical patient data on 433 individuals was collected from hospital files, and saliva samples were used for genotyping.

**Exposure:** The standard chemotherapy treatment for disseminated testicular cancer is three to four cycles of bleomycin-etoposide-cisplatin. The larger part of the patients received standard-dose cisplatin  $20 \text{ mg/m}^2 \times 5 \text{ q3w}$ , etoposide  $100 \text{ mg/m}^2 \times 5 \text{ q3w}$ , and bleomycin  $15.000 \text{ IU/m}^2 \text{ q1w}$ , and 25 patients received double-dose cisplatin and etoposide: cisplatin  $40 \text{ mg/m}^2 \times 5 \text{ q3w}$ , etoposide  $200 \text{ mg/m}^2 \times 5 \text{ q3w}$ , and bleomycin  $15.000 \text{ IU/m}^2 \text{ q1w}$ .

**Main Outcome(s) and Measure(s):** Hearing loss was classified according to the FACT/GOG-Ntx-11 version 4 self-reported NTX6 question that measures difficulty in hearing. A logistic regression with inner-GWAS was developed to identify patients at high risk of developing hearing loss.

**Results:** Of a total of 433 patients, 424 answered NTX6 and 34.4% scored 2 to 4, phenotypical ototoxicity. These patients had a median age at diagnosis (interquartile range (IQR)) of 34 (27-41) years while the non-affected patients had a median age (IQR) of 29 (26-36) years. The prediction model comprising clinical and genomics data was able to identify 67% of the patients with hearing loss, however, with a false discovery rate of 49%. For the non-affected patients, the model identified 66% of the patients with a false omission rate of 19%. An area under the receiver operating characteristic curve (ROC-AUC) of 0.73 (95% CI, 0.71-0.74) was obtained, while a ROC-AUC of 0.66 (95% CI, 0.65-0.66) was obtained for the model with only clinical data.

**Conclusions and Relevance:** A prediction model and a discussion concerning possible biological mechanism for hearing loss development are presented. This prediction model may be used in the clinic to allow for earlier detection and prevention of hearing loss. These findings need to be confirmed in larger studies before applying to clinical practice.

## Introduction

Testicular cancer is the most common cancer in men below 40 years of age in developed countries with a continuously rising incidence in many countries<sup>1</sup>. It is a highly curable tumor with a 5-year survival of more than 90% disregarding initial stage, which results in an increasing population of long-term testicular cancer survivors<sup>2,3</sup>. Treatment for patients with disseminated disease includes a multi-modality approach with initial surgery – orchiectomy, and either radiotherapy or chemotherapy followed by possible secondary surgery. The majority of patients receive chemotherapy in the form of bleomycin-etoposide-cisplatin, which has been standard of care since early 1980's<sup>4</sup>. Although testicular cancer is associated with a high curability, treatment with chemotherapy is hampered by late effects, such as ototoxicity, neurotoxicity, nephrotoxicity, cardiovascular disease, and psychosocial problems<sup>5,6</sup>. Prevalence of ototoxicity is associated with cumulative cisplatin doses and age at diagnosis and possibly genetic factors<sup>7-9</sup>. Yet, there is a need for further identification of risk factors to identify patients at risk of ototoxicity and possibly initiate preventive measures. In this study we aimed at identifying risk factors for ototoxicity in testicular cancer survivors (TCS) via the usage of a prediction logistic regression model to address the burden of cisplatin-induced hearing loss. Using clinical and genomics data integration, we have identified a subgroup of individuals who were more predisposed to develop hearing loss. Additionally, a new biological mechanism is proposed.

## Material & Methods

### Source of the data

Long-term TCS were identified in the Danish Testicular Cancer (DaTeCa)-Late cohort<sup>10</sup> which houses patients initially treated for testicular cancer from 1984 through 2007. In October 2014, TCS were invited to fill in a questionnaire including the Functional Assessment of Cancer Therapy/Gynecologic Oncology Group-Neurotoxicity (FACT/GOG-Ntx)-11 version 4, with 11 questions related to overall neurotoxicity, sensory neuropathy, ototoxicity, motor difficulty, and dysfunction neurotoxicity. Patients with renal function measurements before and after chemotherapy were asked to deliver a saliva sample for genotyping, as previously published,  $n = 433$ <sup>11</sup>. Clinical features were originally extracted from hospital files as registered in the DaTeCa database<sup>12</sup>.

Patients gave informed consent to participate in this study, and the study was approved by the regional ethical committee (File number, H-2-2012-044), as well as the National Board of Data Protection (File number, 2012-41-0751).

### Treatment and clinical information

All patients received bleomycin-etoposide-cisplatin for disseminated testicular cancer, three cycles or more as previously described<sup>11</sup>.

Clinical information consisted of age at diagnosis, body mass index (BMI), normal dose vs double-dose bleomycin-etoposide-cisplatin, glomerular filtration rate before treatment, cumulative cisplatin dose per square meter of body surface area (BSA), number of bleomycin-etoposide-cisplatin (BEP) treatment cycles, histology (seminoma vs nonseminoma), prognostic classification as per IGCCCG<sup>13</sup>, and information about alcohol consumption in number of units per week, and smoking habits (never; former; or current). BMI, alcohol and smoking information were collected at the time of the questionnaire in October 2014. Age at the time of questionnaire was highly correlated with age at diagnosis (Pearson



correlation 0.76). Normal dose vs double-dose BEP was not included in the analysis as its information is covered by cumulative cisplatin dose per square meter of BSA, and number of treatment cycles.

### Assessment of hearing loss

Self-perceived hearing loss was assessed with the Ntx subscale of FACT/GOG-Ntx-11, version 4. The FACT/GOG-Ntx is a self-reported questionnaire in the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System<sup>14</sup> that evaluates the severity and impact of neuropathy. The questionnaire consists of 11 items rated from 0 (“not at all”) to 4 (“very much”). The scale can be divided into four subscales: sensory neuropathy, motor neuropathy, auditory neuropathy, and dysfunctional problems<sup>14</sup>. Auditory neuropathy comprises two different questions, where NTX6 measures difficulty hearing, and NTX7 measures tinnitus.

Here, only NTX6 is further explored. NTX7 will be evaluated in a separate paper as the genotypes involved may vary between the two toxicities<sup>15,7</sup>. For NTX6, to ensure clinical relevance, the outcome was dichotomized. Two different cut-offs for the binarization of the risk group were considered: (1) low (score from 0-1) and high-risk group (score from 2-4), and (2) low (score from 0-2) and high-risk group (score from 3-4). Further results and discussion refer to (1), where better results were achieved.

Additionally, the patients were asked if they recalled experiencing worse hearing during treatment (hearing change question 1 (HC Q1)) and whether it returned to normal afterwards (hearing change question 2 (HC Q2)).

### DNA preparation and quality control

DNA samples were prepared at DTU Multi-Assay Core (Lyngby, Denmark), and genotyped at AROS Applied Biotechnology A/S company (Aarhus, Denmark) using Illumina® HumanOmniExpressExome-8-v1-2-B-b37 chip (around 1 million markers).

Genotyping data were converted into pedigree format using GenomeStudio® (v2011.1) with PLINK Input Report Plug-in (v2.1.3). Quality control for both SNPs and patient samples is described in Supplementary Figure 1.

### SNPs data feature selection

SNPs were selected via (1) inner-GWAS (described in “Statistical Analysis”) and (2) gene literature search. A systematic literature search was conducted on databases Uniprot<sup>16</sup>, DrugBank<sup>17</sup>, KEGG<sup>18,19,20</sup>, and BioCyc<sup>21</sup>. In the databases, literature search was done on genes associated with cisplatin metabolism and genes related to ototoxicity (Supplementary Table 1 and Supplementary Note 2). To extract the markers available in our dataset located at the literature search genes, the SNPs were annotated with Ensembl Variant Effect Predictor (VEP)<sup>22</sup>, and all SNPs from a specific gene or with a specific consequence were easily extracted. The literature search resulted in a large number of genetic features, thus SNPs were further filtering using Ensembl VEP for “IMPACT is HIGH” OR “CLIN\_SIG” is drug\_response, leaving a total of 19 SNPs included from literature: *CYP2J2* rs11572279, *MGST3* rs9333378, *ABCA12* rs10498027, *ABCC5* rs939336, *WFS1* rs1801206, *SLC44A4* rs494620, *NOX3* rs12195525, *CEP78* rs17787781, *CYP2C9* rs4917639, *CYP2C8* rs2071426, *SYCE1* rs2149616, *ABCC8* rs2074308, *DUSP6* rs808820, *DMXL2* rs2414105, *ABCA10* rs10491178, *ABCA7* rs3752229, *CYP2B6* rs2279345, *ERCC1* rs3212986, *MCM8* rs3761873. Additionally, seven SNPs reported to be associated with cisplatin ototoxicity in a recent

systematic review paper<sup>23</sup> were also included (two SNPs were not available in our dataset): *LRP2* rs2075252, *LRP2* rs4668123, *TPMT* rs1800460, *SOD2* rs4880, *GSTP1* rs1695, *COMT* rs4646316, *COMT* rs9332377.

## Statistical analysis

### Missing data

All patients had full phenotypical data. In patients with missing values in predictors ( $n = 2$  for BMI and smoking) a multiple imputation method with 10 iterations was used. Imputation was performed separately in the training and test sets to avoid leak of information between these two sets.

### Logistic regression with inner-GWAS

A nested cross-validation (CV) was implemented. First, the data was split into training and test sets in an outer 5-fold CV loop. Second, the training set was further split into a training and validation subset. Forward feature selection and parameter optimization were done in the training-validation subsets, and the model was deployed on the independent test set. An inner-GWAS was performed on the training set of the inner-fold to select SNPs for model training. Each genetic variant was tested for its association with hearing loss using a logistic regression model after adjusting for potential confounding effects: age at the time of questionnaire and cisplatin per body surface area. A P value threshold of  $1 \times 10^{-5}$  was used to select SNPs for model training; however, none or very few SNPs passed this P value threshold. Thus, a less strict P value threshold of  $1 \times 10^{-4}$  was used (Supplementary Figure 2).

Initially, only clinical data was included. The area under the receiver operating characteristic curve (ROC-AUC) was used to evaluate the model's prediction ability. Clinical features selected until the ROC-AUC reached a plateau were selected. The genomic data, which consisted of the SNPs filtered by literature search and inner-GWAS, were then added to the model with the selected clinical data. For the model with clinical and genomics data, the same forward feature selection approach was followed. The variables included until the model reached a plateau ROC-AUC were accessed. SHapley Additive exPlanations (SHAP) values<sup>24</sup> were accessed to interpret the impact of each feature in the model.

The dataset was randomly split 30 different times in training (inner and outer fold), validation, and test set, to ensure model reproducibility and robustness.

For more information on model hyperparameters, encoding of variables and feature normalization consult Supplementary Note 1.

Randomization tests were also tried to make sure the model was not fitting random noise. This was achieved by eliminating any association between the features and the outcome. For the model with clinical and genomics data, this was achieved by adding random SNPs.

All statistics were calculated using SciKit-learn<sup>25</sup> (v0.23.2) and PLINK<sup>26</sup> (v1.9) in Python (v3.6.10).

## Results

### Population characteristics and quality control

Out of all 433 patients, 424 filled in the NTX6. Of those, 146 (34.4%) scored 2 to 4 in NTX6, phenotypical ototoxicity. These affected patients had a median age at diagnosis (interquartile range [IQR]) of 34 (27-41) years while the non-affected ( $n = 278$ ) patients had a median age (IQR) of 29 (26-36) years. Demographical features are presented in Table 1.

After genotype quality control, 393 patients and 611129 SNP remained (Supplementary Figure 1).

### Prediction model – integration of clinical and genomic features

First, the prediction power of the clinical features alone was assessed. One feature was added at a time, following a forward feature selection approach, until all nine features (previously referred in methods) were included in the model. The ROC-AUC reached a plateau when two features were added to the model with a mean ROC-AUC of 0.66 (95% CI, 0.65-0.66) (Figure 1A,C). These two features were age at diagnosis (selected 30 times out of 30 models: one model for each random data split) and the number of treatment cycles (selected 17 times out of 30 models).

### Genomic feature selection (inner-GWAS) as part of the model

Genomics data was then added to the model with age at diagnosis and number of treatment cycles. Again, a forward feature selection approach was followed. The ROC-AUC reached a plateau when 8 features were added into the model with a mean ROC-AUC of 0.73 (95% CI, 0.71-0.74) (Figure 1B,D and Figure 2A). Two out of the eight features were the clinical data. For the remaining six, the SNPs selected were: *SOD2* rs4880, *MGST3* rs9333378, intergenic rs4389005, *ABCA10* rs10491178, *ABCA12* rs10498027, *MCM8* rs3761873 (Table 2). Out of 30 models, these SNPs were selected 15, 9, 7, 6, 5, and 4 times, respectively. Genotypes *SOD2* rs4880:AA and *MGST3* rs9333378:AA (two most selected SNPs) were found in 47% of patients who replied NTX6 = 0/1, 63% of patients who replied NTX6 = 2 and 76% of patients who replied NTX6 = 3/4.

For each sample, the prediction scores ranged between 0 and 1, where a value closer to 1 means higher probability of hearing loss. Using a default cut-off of 0.50, a sensitivity of 67% was reached (identification of patients with hearing loss), equivalent to a positive predictive value (PPV) of 51%. Correspondingly this resulted in a specificity of 66% and a negative predictive value (NPV) of 80% (Figure 2B). The model performed best on patients with the highest toxicity (score of 4 “very much”), Figure 2C.

The most important feature for the prediction was age at diagnosis, followed by the number of treatment cycles and the SNPs in the same order of times they were selected in the model (Figure 3).

For most patients, adding genomics data helped with the prediction (320 out of 393). For 18 out of 393 patients, the addition of genomics data worsened the prediction, even though 11 out of 18 were still correctly classified. For 42 out of 393 patients, genomics helped with the prediction, however this was still not enough to correctly classify these patients as either affected or non-affected. For 55 out of 393 patients, neither clinical nor genomics data seemed to help the classification (Supplementary Figure 3).

### Hearing loss at the time of treatment Vs. at the time of questionnaire

NTX6 from the validated FACT/GOG-Ntx questionnaire correlated with the hearing loss reported from the two additional questions concerning self-perceived changes during treatment (Spearman’s rank correlation coefficient 0.56 for HC Q1 and 0.76 for HC Q2).

## Randomization tests

Two randomization tests were applied to assess model robustness. For the models with only clinical data, variables permutation was done and a mean ROC-AUC close to 0.50 was obtained (Supplementary Figure 4A). The mean ROC-AUC for the random model with 2 features (benchmark against the non-random model) was 0.50 (95% CI, 0.49-0.51) (Supplementary Figure 4B). When adding random genomic variants, ROC-AUC decreased as non-informative SNPs were being added (Supplementary Figure 4C). Mean ROC-AUC was 0.67 (95% CI, 0.66-0.68) for the random model with 8 features (Supplementary Figure 4D).

## Discussion

In this study, we present a model for prediction of hearing loss after cisplatin-based chemotherapy based on a combination of clinical and genetic features. We observed an improvement of the prediction when including both clinical and genomics data compared to only clinical data, yet still observed misclassifications. Age at diagnosis and cisplatin dose were the most important clinical predictors, as previously reported in other studies<sup>8,9,27</sup>.

### SNPs and its biological importance

SNPs rs4880 *SOD2* and rs9333378 *MGST3* were the two most selected SNPs by the model, and it is hypothesized here that they may have a combined effect on ototoxicity.

The functional rs4880 SNP is located on codon 16 exon 2 of *SOD2*, which codes for superoxide dismutase 2, a mitochondrial protein<sup>28</sup>. SNP rs4880 is the most studied *SOD2* SNP<sup>29</sup>, however there is no agreement regarding the risk allele<sup>30,31,32</sup>.

The SNP rs9333378 is located in *MGST3*, that codes for the microsomal glutathione S-transferase 3<sup>28</sup>. Between the microsomal glutathione S-transferases, MGST1, MGST2, and MGST3 have been reported to be important in the detoxification process<sup>33</sup>.

SNP rs4880 (G > A) produces an amino acid change from alanine to valine (Ala16Val), which may change the structure of the SOD2 mitochondrial targeting sequencing<sup>34</sup>. Computer models have predicted a beta-sheet structure if a valine is present, and a partial alpha-helix structure if an alanine is present<sup>34</sup>. Due to partial arrest of the beta-sheet structure during transport across the inner mitochondrial membrane, this will likely inhibit efficient mitochondrial import of SOD2 precursors<sup>32,30</sup>.

When platinum enters the cells, it is metabolized by the mitochondria, which will lead to the production of reactive oxygen species (ROS), such as superoxide. SOD2 will then degrade superoxide into hydrogen peroxide until complete superoxide anion degradation. However, if SOD2 is retained from entering the mitochondria due to partial arrest of beta-helix, this will lead to an accumulation of ROS. ROS cause lipid peroxidation, activation of pro-inflammatory factors, and cell death by apoptosis, including hair cells<sup>35,36</sup>. Indeed, we observed the A-allele with a higher frequency in patients who reported hearing loss (odds ratio = 1.55, 95% CI: 1.13-2.13). Furthermore, glutathione's, including glutathione S-transferase, are known to help with complete superoxide anion degradation<sup>32</sup>. In the brain, MGST3 expression levels seem to be lower if rs9333378:A, as reported in Genotype-Tissue Expression (GTEx)<sup>37</sup>. Thus, if there is decreasing MGST3 activity, this will enhance the accumulation of cisplatin.

Additionally, potential novel variants associated with cisplatin-induced hearing loss were found. SNP rs4389005 located in an intergenic region was found in the inner-GWAS. The closest gene is GPR12 (64415 base pairs to canonical transcription start site). The other variants selected in the model were found via gene literature search. SNP *ABCA10* rs10491178 and

*ABCA12* rs10498027 both code for ATP-binding cassette (ABC) transporters and overexpression of ABC transporters have been associated with multidrug resistance, including cisplatin, in multiple tumors<sup>38</sup>. SNP *ABCA10* rs10491178 seems to affect expression of protein ABCA6, being the genotype *ABCA10* rs10491178:GG associated with lower expression of ABCA6<sup>37</sup>, which can lead to higher sensitivity to cisplatin and higher toxicity<sup>39</sup>. The last SNP selected by the model, rs3761873 is located in *MCM8* that codes for the mini-chromosome maintenance 8 homologous recombinant repair factor protein (MCM8), and in a recent study, it was shown that in mice, inhibition of MCM8 (and MCM9) hypersensitized cells to cisplatin<sup>40</sup>.

### Patients with self-reported higher levels of hearing loss

Even though we observed a false discovery rate of 49% using a 0.50 cut-off, it is encouraging to see that only four out of 23 patients with the highest score (NTX6=4) were misclassified. Three of them had a prediction score very close to the 0.50 cut-off (2 patients with 0.48 prediction score and one with 0.49 prediction score). For the other misclassified patient, the prediction score was 0.31 and this patient was also the youngest of the 23. Furthermore, this patient received one of the lowest amounts of cisplatin (300 mg/m<sup>2</sup> and three treatment cycles) and he was heterozygous for all SNPs, except *ABCA10* rs10491178 and *MCM8* rs3761873 where he was homozygous for the reference allele. This may point to other relevant genetic predispositions that might be underrepresented in this dataset and hence may not have been detected. The cutoff choice will always have an impact on the tradeoff of positive and negative errors, for example, by increasing it, the number of false positives decreases, but so does the number of true positives.

### Limitations and strengths

The diagnosis of ototoxicity is very challenging to perform and ototoxicity definition is still far from being entirely defined<sup>41</sup>. Here, several potential factors for hearing loss were not explored, such as noise, infection, or vascular problems and the toxicity was assessed several years after exposure. However, long-term toxicity also has the highest impact on quality of life and may be most important to predict.<sup>9</sup>

The models were trained on labels that derive from the FACT/GOG-Ntx questionnaire, which are not objectively measured. Other measurements such as pure-tone audiometry or other hearing tests could have been done to improve precision<sup>41</sup>. On the other hand, using quality of life measures ensures that the focus is on the patient<sup>42</sup>, as objective measurement might detect the same level of toxicity between two individuals, however only one may be affected by the symptoms.

As a potential limitation, body mass index as well as information about alcohol consumption and smoking habits were retrieved in 2014 when the questionnaire was done. These clinical features were used as a proxy at the time of treatment, but they may not represent the true values. While those features were not selected in the final model, we are unaware if the real values at the time of treatment could have added relevant information to the model.

### Conclusions

Cisplatin is essential in the treatment of several neoplasms, however, the inability to predict accurately how patients will react to chemotherapy represents a major challenge. Ototoxicity is one of the most common late-side effects of cisplatin-based chemotherapy. In this study, we present a logistic regression prediction model based on a combination of genetic and clinical features able to classify patients at high or low risk of hearing loss after cisplatin-based treatment. We also propose a new

mechanism of hearing loss development involving the SNPs *SOD2* rs4880:AA and *MGST3* rs9333378:AA. These SNPs have not yielded significant results when single associations between SNPs and outcome have been performed. Before application to clinical practice, confirmation in a prospective clinical setting and replication in larger studies are required. This model could be used as a complement to support the clinical decision and help on reducing hearing loss cases by adjusting treatment for patients in the high-risk group.

## References

1. Chia VM, Quraishi SM, Devesa SS, Purdue MP, Cook MB, McGlynn KA. International trends in the incidence of testicular cancer, 1973-2002. *Cancer Epidemiol Biomarkers Prev.* 2010;19(5):1151-1159. doi:10.1158/1055-9965.EPI-10-0031
2. Shanmugalingam T, Soultati A, Chowdhury S, Rudman S, Van Hemelrijck M. Global incidence and outcome of testicular cancer. *Clin Epidemiol.* 2013;5:417-427. doi:10.2147/CLEP.S34430
3. Kier MG, Lauritsen J, Mortensen MS, et al. Prognostic Factors and Treatment Results After Bleomycin, Etoposide, and Cisplatin in Germ Cell Cancer: A Population-based Study. *Eur Urol.* 2017;71(2):290-298. doi:10.1016/j.eururo.2016.09.015
4. Einhorn LH, Donohue J. Cis-diamminedichloroplatinum, vinblastine, and bleomycin combination chemotherapy in disseminated testicular cancer. *Ann Intern Med.* 1977. doi:10.7326/0003-4819-87-3-293
5. Chovanec M, Lauritsen J, Bandak M, et al. Late adverse effects and quality of life in survivors of testicular germ cell tumour. *Nat Rev Urol.* 2021;18(4):227-245. doi:10.1038/s41585-021-00440-w
6. Lauritsen J, Mortensen MS, Kier MGG, et al. Renal impairment and late toxicity in germ-cell cancer survivors. *Ann Oncol.* 2014;26(1):173-178. doi:10.1093/annonc/mdl506
7. El Charif O, Mapes B, Trendowski MR, et al. Clinical and genome-wide analysis of cisplatin-induced tinnitus implicates novel ototoxic mechanisms. *Clin Cancer Res.* 2019;25(13):4104-4116. doi:10.1158/1078-0432.CCR-18-3179
8. Frisina RD, Wheeler HE, Fossa SD, et al. Comprehensive audiometric analysis of hearing impairment and tinnitus after cisplatin-based chemotherapy in survivors of adult-onset cancer. *J Clin Oncol.* 2016;34(23):2712-2720. doi:10.1200/JCO.2016.66.8822
9. Lauritsen J, Bandak M, Kreiberg M, et al. Long-term neurotoxicity and quality of life in testicular cancer survivors—a nationwide cohort study. *J Cancer Surviv.* 2020. doi:10.1007/s11764-020-00944-1
10. Kreiberg M, Bandak M, Lauritsen J, et al. Cohort Profile: The Danish Testicular Cancer Late Treatment Effects Cohort (DaTeCa-LATE). *Front Oncol.* 2018;8:37. doi:10.3389/fonc.2018.00037
11. Garcia SL, Lauritsen J, Zhang Z, et al. Prediction of Nephrotoxicity Associated With Cisplatin-Based Chemotherapy in Testicular Cancer Patients. *JNCI Cancer Spectr.* 2020;4(3):1-8. doi:10.1093/jncics/pkaa032
12. Daugaard G, Kier MGG, Bandak M, et al. The Danish testicular cancer database. *Clin Epidemiol.* 2016;8:703-707. doi:10.2147/CLEP.S99493
13. International Germ Cell Cancer Collaborative Group. Germ Cell Consensus Classification: a prognostic factor-based staging system for metastatic germ cell cancers. International Germ Cell Cancer Collaborative Group. *J Clin Oncol.*

- 1997;15(2):594-603.
14. Huang HQ, Brady MF, Cella D, Fleming G. Validation and reduction of FACT/GOG-Ntx subscale for platinum/paclitaxel-induced neurologic symptoms: a gynecologic oncology group study. *Int J Gynecol cancer Off J Int Gynecol Cancer Soc.* 2007;17(2):387-393. doi:10.1111/j.1525-1438.2007.00794.x
  15. Henry JA, Roberts LE, Caspary DM, Theodoroff SM, Salvi RJ. Underlying Mechanisms of Tinnitus: Review and Clinical Implications. *J Am Acad Audiol.* 2014;25(01):005-022. doi:10.3766/jaaa.25.1.2
  16. Consortium TU. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2018;47(D1):D506-D515. doi:10.1093/nar/gky1049
  17. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 2018;46(D1):D1074-D1082. doi:10.1093/nar/gkx1037
  18. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* 2019;28(11):1947-1951. doi:10.1002/pro.3715
  19. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* 2019;47(D1):D590-D595. doi:10.1093/nar/gky962
  20. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 2000;28(1):27-30. doi:10.1093/nar/28.1.27
  21. Caspi R, Billington R, Ferrer L, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 2016;44(D1):D471-D480. doi:10.1093/nar/gkv1164
  22. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 2016;17(1):122. doi:10.1186/s13059-016-0974-4
  23. Tserga E, Nandwani T, Edvall NK, et al. The genetic vulnerability to cisplatin ototoxicity: a systematic review. *Sci Rep.* 2019;9(1):3455. doi:10.1038/s41598-019-40138-z
  24. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. May 2017. <http://arxiv.org/abs/1705.07874>. Accessed April 2, 2021.
  25. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. January 2012. <http://arxiv.org/abs/1201.0490>. Accessed July 16, 2019.
  26. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet.* 2007;81(3):559-575. doi:https://doi.org/10.1086/519795
  27. Haugnes HS, Stenklev NC, Brydøy M, et al. Hearing loss before and after cisplatin-based chemotherapy in testicular cancer survivors: a longitudinal study. *Acta Oncol (Madr).* 2018;57(8):1075-1083. doi:10.1080/0284186X.2018.1433323
  28. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinforma.* 2016;54(1). doi:10.1002/cpbi.5
  29. Pourvali K, Abbasi M, Mottaghi A. Role of Superoxide Dismutase 2 Gene Ala16Val Polymorphism and Total Antioxidant Capacity in Diabetes and its Complications. *Avicenna J Med Biotechnol.* 8(2):48-56. <http://www.ncbi.nlm.nih.gov/pubmed/27141263>.
  30. Sutton A, Imbert A, Igoudjil A, et al. The manganese superoxide dismutase Ala16Val dimorphism modulates both mitochondrial import and mRNA stability. *Pharmacogenet Genomics.* 2005;15(5):311-319. doi:10.1097/01213011-200505000-00006



- 369 31. Bastaki M, Huen K, Manzanillo P, et al. Genotype-activity relationship for Mn-superoxide dismutase, glutathione  
370 peroxidase 1 and catalase in humans. *Pharmacogenet Genomics*. 2006;16(4):279-286.  
371 doi:10.1097/01.fpc.0000199498.08725.9c
- 372 32. Brown AL, Lupo PJ, Okcu MF, Lau CC, Rednam S, Scheurer ME. SOD2 genetic variant associated with treatment-  
373 related ototoxicity in cisplatin-treated pediatric medulloblastoma. *Cancer Med*. 2015;4(11):1679-1686.  
374 doi:10.1002/cam4.516
- 375 33. Uno Y, Murayama N, Kunori M, Yamazaki H. Characterization of Microsomal Glutathione S -Transferases MGST1,  
376 MGST2, and MGST3 in Cynomolgus Macaque. *Drug Metab Dispos*. 2013;41(9):1621-1625.  
377 doi:10.1124/dmd.113.052977
- 378 34. Sutton A, Khoury H, Prip-Buus C, Cepanec C, Pessayre D, Degoul F. The Ala16Val genetic dimorphism modulates  
379 the import of human manganese superoxide dismutase into rat liver mitochondria. *Pharmacogenetics*.  
380 2003;13(3):145-157. doi:10.1097/01.fpc.0000054067.64000.8f
- 381 35. Romano A, Capozza MA, Mastrangelo S, et al. Assessment and Management of Platinum-Related Ototoxicity in  
382 Children Treated for Cancer. *Cancers (Basel)*. 2020;12(5):1266. doi:10.3390/cancers12051266
- 383 36. Paciello F, Fetoni AR, Mezzogori D, et al. The dual role of curcumin and ferulic acid in counteracting  
384 chemoresistance and cisplatin-induced ototoxicity. *Sci Rep*. 2020;10(1):1063. doi:10.1038/s41598-020-57965-0
- 385 37. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45:580.  
386 <https://doi.org/10.1038/ng.2653>.
- 387 38. Pasello M, Giudice AM, Scotlandi K. The ABC subfamily A transporters: Multifaceted players with incipient  
388 potentialities in cancer. *Semin Cancer Biol*. 2020;60:57-71. doi:10.1016/j.semcancer.2019.10.004
- 389 39. Robey RW, Pluchino KM, Hall MD, Fojo AT, Bates SE, Gottesman MM. Revisiting the role of ABC transporters in  
390 multidrug-resistant cancer. *Nat Rev Cancer*. 2018;18(7):452-464. doi:10.1038/s41568-018-0005-8
- 391 40. Morii I, Iwabuchi Y, Mori S, et al. Inhibiting the MCM8-9 complex selectively sensitizes cancer cells to cisplatin and  
392 olaparib. *Cancer Sci*. 2019;110(3):1044-1053. doi:10.1111/cas.13941
- 393 41. Ganesan P, Schmiedge J, Manchaiah V, Swapna S, Dhandayutham S, Kothandaraman PP. Ototoxicity: A Challenge  
394 in Diagnosis and Treatment. *J Audiol Otol*. 2018;22(2):59-68. doi:10.7874/jao.2017.00360
- 395 42. Higginson IJ. Measuring quality of life: Using quality of life measures in the clinical setting. *BMJ*.  
396 2001;322(7297):1297-1300. doi:10.1136/bmj.322.7297.1297
- 397 43. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
- 398 44. HannahVMeyer. meyer-lab-cshl/plinkQC: plinkQC version 0.3.3. February 2021. doi:10.5281/ZENODO.4521453



**Table 1. Comparison of baseline characteristics between affected and non-affected patients (NTX6 FACT/GOG-Ntx).**

Out of 433, nine patients had not replied on NTX6, thus only 424 patients are represented in the table. Values show the median and interquartile range (IQR; 25%–75%) or number of patients and percentages (%).

*IQR = interquartile range; BMI = body mass index; BEP = bleomycin-etoposide-cisplatin; GFR = glomerular filtration rate; IQR = interquartile range.*

		Affected, No. (%)	Non-affected, No. (%)	P values <sup>a</sup>
<b>Number of patients</b>		146 (34.4)	278 (65.6)	-
<b>Age at diagnosis, median (IQR)</b>		34 (27-41)	29 (26-36)	0.002
<b>BMI, median (IQR)</b> Unknown: 8 Affected ; 10 Non-affected		21 (19-27)	22 (19-26)	0.38
<b>BEP regimen</b>	Normal dose	113 (78.5)	260 (95.6)	6x10 <sup>-27</sup>
	Double dose	31 (21.5)	12 (4.4)	
	Unknown	2	6	
<b>GFR before treatment, median (IQR), mL/min/1.73m<sup>2</sup></b> Unknown: 2 Non-affected		122 (111-135)	121 (110-133)	0.68
<b>Cisplatin, median (IQR), mg/m<sup>2</sup></b>		400 (385-403)	400 (300-400)	P < .001
<b>Treatment cycles</b>	3	30 (20.5)	86 (30.9)	1x10 <sup>-5</sup>
	4	85 (58.2)	180 (64.7)	
	5 or more	9 (6.2)	10 (3.6)	
	High-dose	22 (15.1)	2 (0.7)	
<b>Histology</b>	Seminoma	34 (23.3)	54 (19.4)	0.42
	Non-Seminoma	112 (76.7)	224 (80.6)	
<b>Prognostic group</b>	Good	103 (70.5)	239 (86)	P < .001
	Intermediate	32 (21.9)	30 (10.8)	
	Poor	11 (7.5)	9 (3.2)	
<b>Alcohol consumption in number of units per week</b>		5 (1-10)	5 (2-10)	0.30
<b>Smoking</b>	Never	61 (41.8)	128 (46.4)	0.40
	Former	55 (37.7)	88 (31.9)	
	Current	30 (20.5)	60 (21.7)	
	Unknown	-	2	

<sup>a</sup>P values were calculated by 2-sided Mann-Whitney *U* test for continuous or ordinal characteristics. For “histology,” *P* value was calculated by  $\chi^2$  test. All tests are appropriate for unpaired data and in the case of continuous variables, non-normal distributed data. Distribution of continuous variables was accessed through Shapiro-Wilk normality test.

**Table 2 Single-nucleotide polymorphisms (SNPs) selected on the final prediction model.** SNPs are ordered by genomic position and not by the number of times selected in the model.

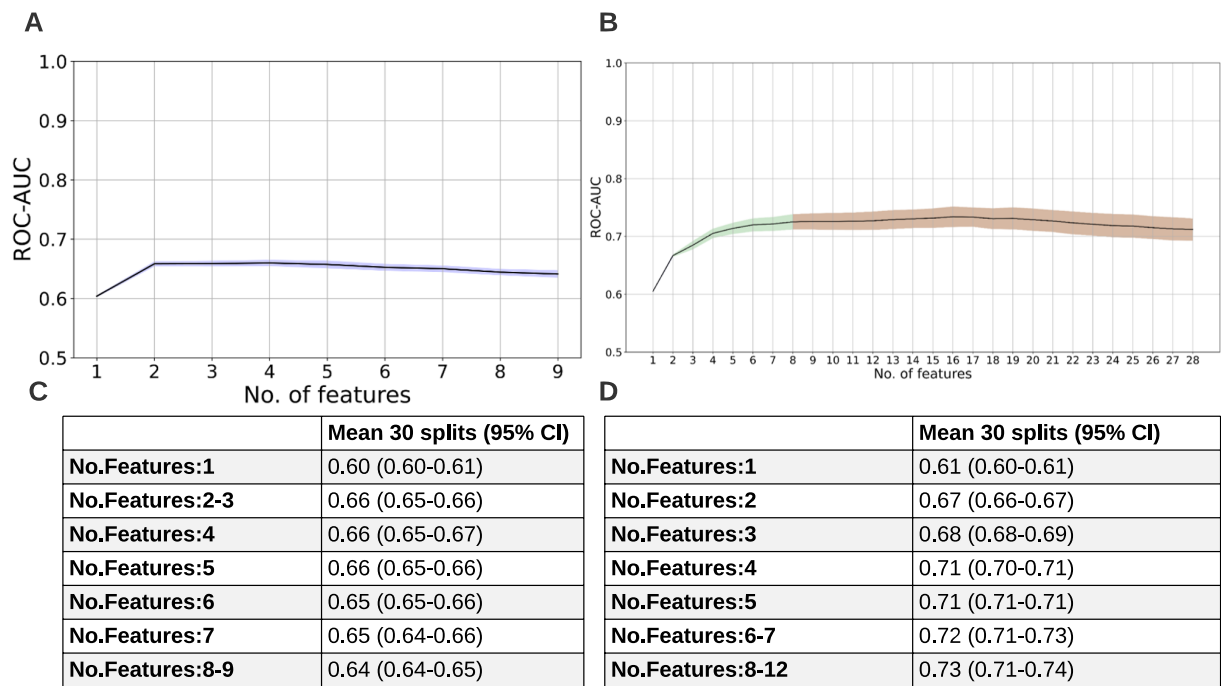
*Chr.* = Chromosome; *SNP* = single nucleotide polymorphism; *MAF* = minor allele frequency; *CEU* = European; *glomerular filtration rate*; *OR* = odds ratio; *CI* = confidence interval.

Chr.	SNP	Genomic position <sup>a</sup>	Gene	Reference allele	Alternative allele	Risk allele	MAF (CEU)	Effect	OR (95% CI) <sup>b</sup>	P <sup>c</sup>
1	rs9333378	165601466	MGST3	G	A	A	G: 0.35	Splice acceptor	1.37 (1-1.86)	0.0441
2	rs10498027	215820013	ABCA12	G	A	G	A: 0.35	Stop gained	1.11 (0.81-1.51)	0.5158
6	rs4880	160113872	SOD2	A	G	A	G: 0.41	Missense	1.55 (1.13-2.13)	0.007183
13	rs4389005	27399338	GPRR12 (closest gene)	A	G	A	G: 0.31	Intergenic	2.09 (1.56-2.89)	7×10 <sup>-6</sup>
17	rs10491178	67149973	ABCA10	G	A	G	A: 0.09	Stop gained	1.84 (0.80-4.22)	0.1525
20	rs3761873	5939214	MCM8	A	C	A	C: 0.12	Stop gained	1.35 (0.66-2.75)	0.4148

<sup>a</sup>Genomic position based on NCBI Human Genome Build 37 coordinates.

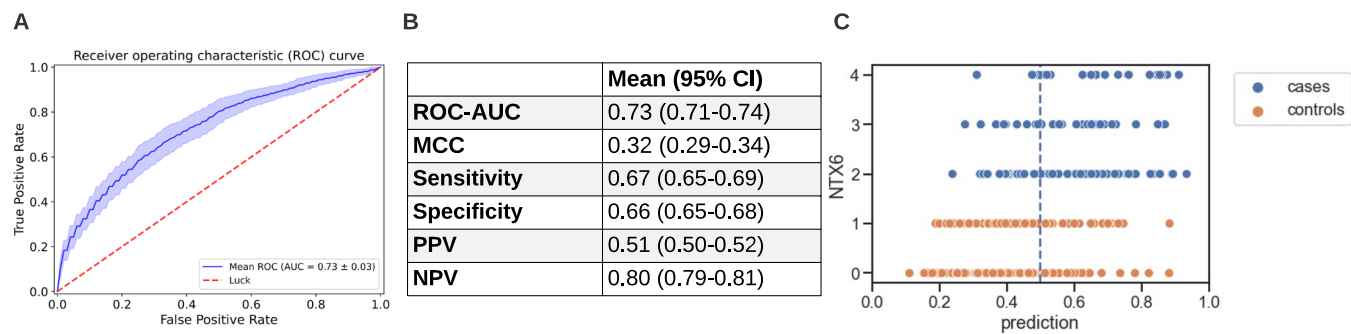
<sup>b</sup>Odds ratio with 95% Confidence Interval for the risk allele.

<sup>c</sup>A logistic model was adjusted for cisplatin dosage and age at the questionnaire and *P* values represent how likely the variant association was by random chance.

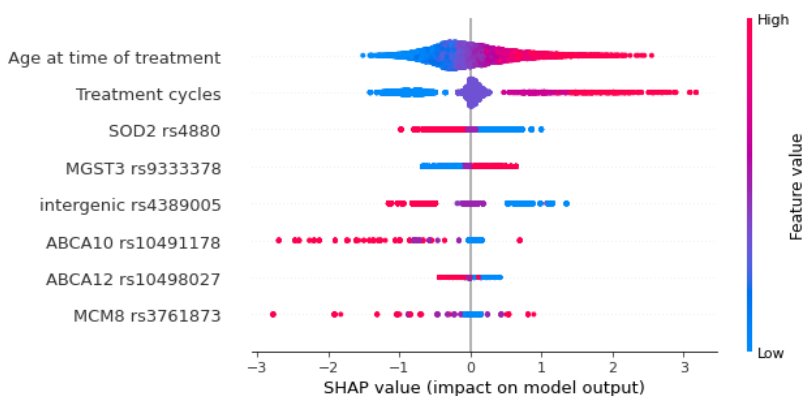


**Figure 1. ROC-AUC mean (30 random data splits) performances in each step of the forward feature selection. A:** Model with clinical data with forward feature selection up until nine features. Shaded blue area indicates 95% CI. Exact ROC-AUC mean and 95% CI in **C**. **B:** Model with clinical and genomics data with forward feature selection up until 28 features. Shaded areas indicate 95% CI and blue color indicates that only clinical data was added, green color that clinical and genomics data were added, and red color that ROC-AUC reached a plateau. Exact ROC-AUC mean and 95% CI in **D**. For illustration purposes, exact ROC-AUC mean and 95% CI are not indicated in **D** from 13 features. From 13-28 features, ROC-AUC mean (95% CI) was 0.73 (0.71-0.75) (13-15 features); 0.73 (0.71-0.75) (14-15 features); 0.73 (0.72-0.75) (16-17 features); 0.73 (0.71-0.75) (18-21 features); 0.72 (0.70-0.74) (22-25 features); 0.72 (0.70-0.73) (26 features); and 0.71 (0.69-0.73) (27-28 features).

ROC-AUC = area under the receiver operating characteristic curve; No. = number; CI = confidence interval.



**Figure 2. Final model performance measures and prediction scores.** **A:** Model performance shown as ROC-AUC curve. Solid blue line and shaded area indicate the mean and standard deviation across 30 data splits. Dashed red line indicates a random classifier. **B:** ROC-AUC and other performance measures, i.e., MCC, sensitivity, specificity, PPV and NPV using a cut-off of 0.50. **C:** Final prediction scores (x-axis) for each patient, represented by a dot. Orange dots represent controls or non-affected patients (NTX6 score 0-1), while blue dots represent cases or affected patients (NTX6 score 2-4). Dashed vertical line represents a cut-off of 0.50, where patients with a prediction score of 0.50 or higher are considered cases. ROC-AUC = area under the receiver operating characteristic curve; MCC = Matthews correlation coefficient; PPV = positive predictive value; NPV = negative predictive value.



**Figure 3. SHAP value feature importance.** Individual features are ranked by importance, where age at diagnosis is the most important feature. The color represents the feature value (red: high; blue: low). Negative SHAP values (x-axis) contribute towards a negative model outcome (control or non-affected), while positive SHAP values contribute towards a positive model outcome (case or affected).

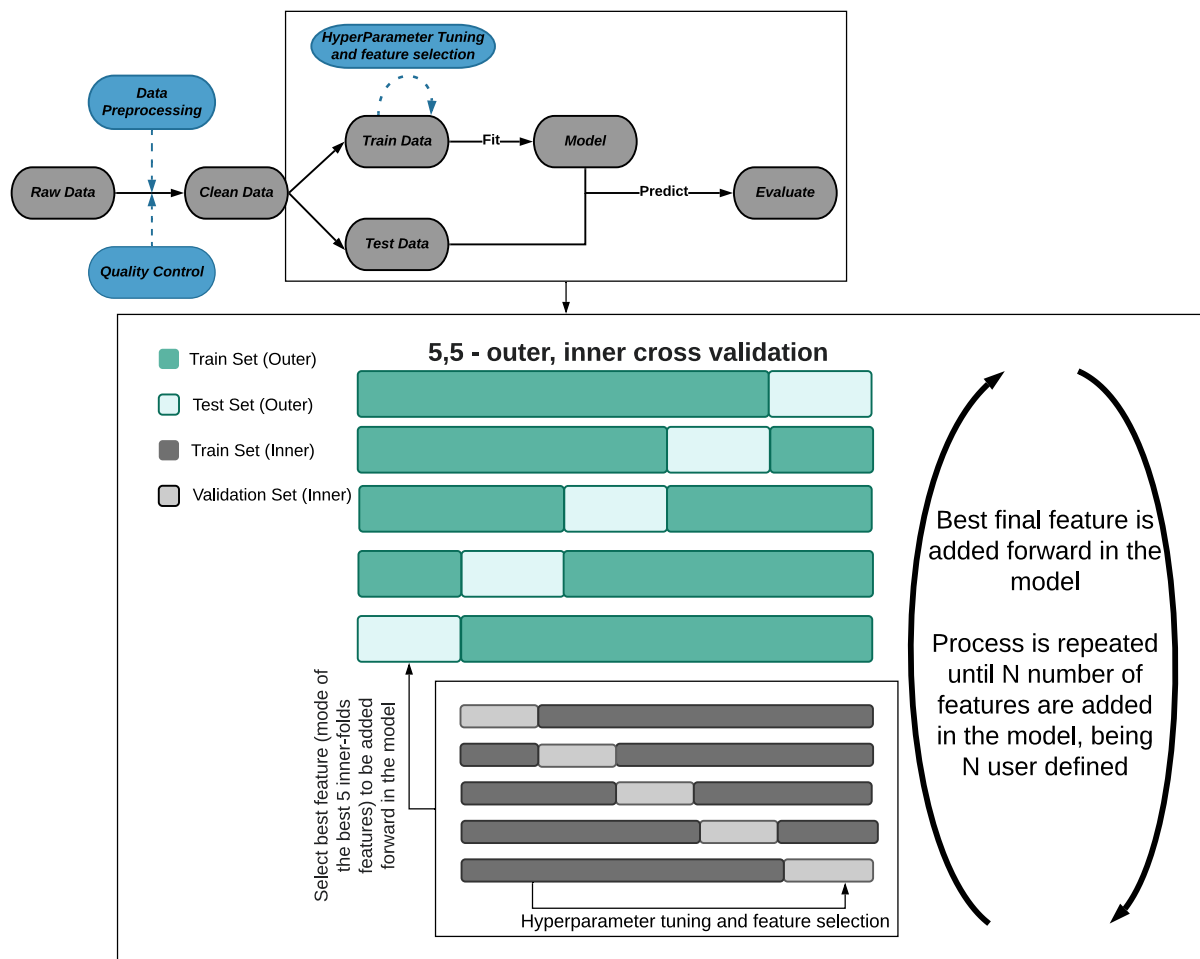
**SUPPLEMENTARY MATERIAL**

**Supplementary Table 1** Overview of the gene literature search identifying genetic markers associated with cisplatin metabolism and ototoxicity.

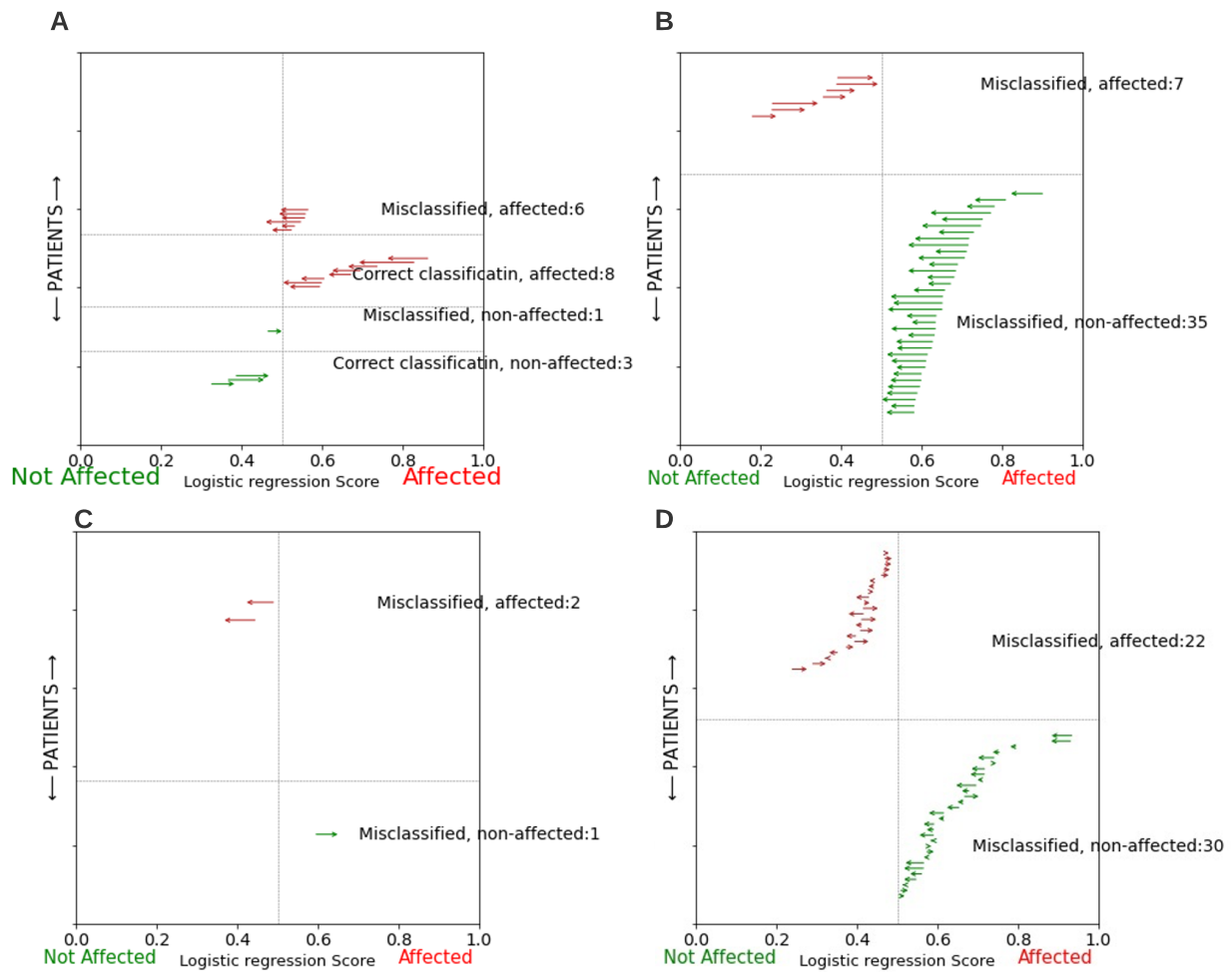
		Description	Number of genes
Cisplatin metabolism	Resistance Pathway (KEGG Pathways)	Overview of genes and interactions resulting in platinum-based drugs resistance.	46
	Detoxification Pathway (BioCyc Pathway)	Cisplatin is degraded via the glutathione-mediated detoxification pathway.	9
	Glutathione Transferases Cytochrome P450 Enzymes ABC Transporters (Uniprot)	The three protein groups may be associated with cisplatin introduced neurotoxicity, since they affect the uptake and disposition. Genes associated with the groups were identified with Uniprot.	26 61 49
	Cisplatin (Uniprot)	Systematic search identifying cisplatin-related genes conducted with Uniprot.	22
	Cisplatin (DrugBank)	The DrugBank database contains information on pharmaceutical drugs including cisplatin.	31
Ototoxicity	Sensorineural Hearing Loss	Sensorineural hearing loss-related genes conducted with Uniprot.	155
	Ototoxicity	Ototoxicity-related genes conducted with Uniprot.	2

Step	Number of patients	Number of SNPs
Raw data	478	964,193
Prepared input	478	921,861
Low call rates	455	873,835
Gender check	455	873,835
Excess hetero- and homozygosity	454	873,835
Non CEU population	452	873,835
Relatedness	450	873,835
Population outliers	450	873,835
Non HWE and low MAF	450	611,129
Missing Questionnaire Information	401	611,129
Missing Label (NTX6)	393	611,129

**Supplementary Figure 1. Step-by-step demonstration of genomic data quality control and information on patients where questionnaire information was missing.** Single-nucleotide polymorphism (SNP) quality filtering included removal of duplicated SNPs and those with ambiguous genome position, strand, and alleles; call rate ( $<98\%$ ); extreme deviation from Hardy–Weinberg equilibrium ( $P$  value  $<5 \times 10^{-6}$ ); and MAF ( $<1\%$ ). Quality controls applied on the patient samples were based on genotype (chromosome X homozygosity rate  $<20\%$  for females and  $>80\%$  for males) and phenotype sex discordance; extreme heterozygosity or homozygosity ( $\pm 4$  SD from sample's hetero-/homozygosity rate mean); outliers from the European descent using 1000 Genomes<sup>43</sup> as reference samples; cryptic relatedness (IBD  $>18.75\%$ ); and population outliers ( $\pm 4$  SD from cluster centroid mean). European outliers were detected by 1) doing principal component analysis (PCA) to find the center of the European reference samples, and 2) remove samples whose Euclidean distance from the center  $> 1.5 \times$  maximum Euclidean distance of the European reference samples<sup>44</sup>. Patients with missing questionnaire information consisted of 45 patients who received more than one line of treatment and therefore were not relevant for the present study and were not invited for the questionnaire in 2014. These were still included for the purpose of quality control only.

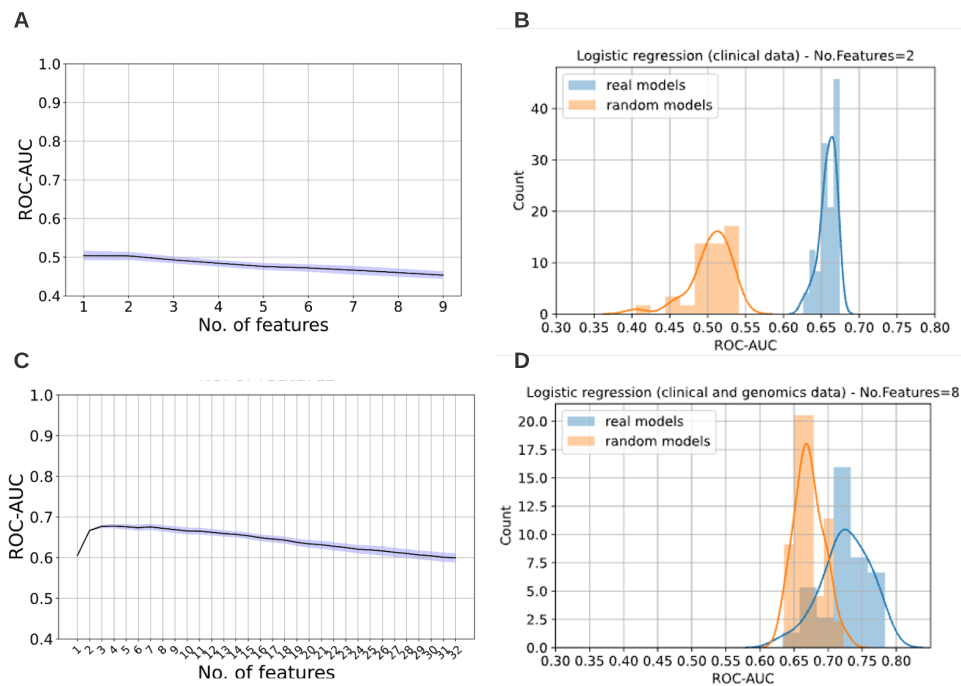


**Supplementary Figure 2. Illustration of logistic regression model used in this study.** Model was run at Computerome 2.0 (<https://www.computerome.dk>). The 30 random data splits were run in parallel to reduce running time, thus 32 nodes were used (30 allocated for each random split and 2 for other initializations). Each node contains 2 CPUs with 20 cores/CPU. 192 GB is the memory distributed through all cores.



**Supplementary Figure 3 Misclassified patients and/or patients where genomics “pushed” final classification in the wrong direction.** Arrow starts at prediction score of model with only clinical data (model 1) and ends at prediction score of model with clinical and genomics data (model 2). **A:** Patients where genomic data “pushed” the classification in the wrong direction, even though some of them were correctly classified; **B:** Inclusion of genomics data helped but not enough to correctly classify these patients; **C, D:** Neither clinical nor genomics data helped on these patients classification (in **D**, score difference between model 2 and 1 was below 0.05). All other patients not represented here were correctly classified and genomic data “pushed” the classification in the right direction (or if in the wrong direction, score difference between model 2 and model 1 was below 0.05).





**Supplementary Figure 4 ROC-AUC mean (30 random splits) performances for the random and real models.** **A:** Model with random clinical data with forward feature selection up until nine features. Shaded blue area indicates 95% CI. **B:** Comparison between real model (mean ROC-AUC 0.66 (95% CI, 0.65-0.66, blue histogram) and random models (mean ROC-AUC 0.50 (95% CI, 0.49-0.51, orange histogram); **C:** Model with clinical and non-informative genomics data with forward feature selection up until 32 features. Shaded areas indicate 95% CI. **D:** Comparison between real model (mean ROC-AUC of 0.73 (95% CI, 0.71-0.74) and random models (mean ROC-AUC was 0.67 (95% CI, 0.66-0.68). In **B** and **C**, count (y-axis) sums up to 150 as the model consists of 5 outer folds and 30 data splits were done (5x30).

ROC-AUC = area under the receiver operating characteristic curve; No. = number; CI = confidence interval.

**Supplementary Note 1. Model hyperparameters, encoding, and normalization**

On the logistic regression model, hyperparameter optimization was done for the inverse of regularization strength. Values between  $1 \times 10^{-4}$  and  $1 \times 10^4$  were tried (20 values in total spaced evenly on a log scale). Smaller values specify stronger regularization.

Ordinal categorical variables were encoded as one-column vectors, i.e.  $(\{1,2,3,4\})$  for treatment cycles 3, 4, 5 or more, and high dose (double dose of chemotherapy, unspecified number of cycles) BEP cycles, respectively;  $(\{1,2,3\})$  for prognosis good, intermediate, and poor, respectively; and  $(\{0,1,2\})$  for smoking habits never, former, and current, respectively. The nominal variable histology was encoded in one column,  $(\{1,2\})$ , for non-seminoma, or seminoma, respectively. Continuous variables were represented in absolute values (age at diagnosis, body mass index, glomerular filtration rate before treatment, cumulative cisplatin dose per square meter of BSA, and alcohol consumption in number of units per week. Single-nucleotide polymorphism data was encoded as one column vector with counts of minor alleles  $(\{0,1,2\})$ .

Features were scaled down to values between 0 and 1 using Sklearns MinMaxScaler. Rescaling of features was done separately in the training and test set to avoid leakage of the test set information.

## Supplementary Note 2. Genes obtained from literature search.

Resistance Pathway (KEGG Pathways\*) (with aliases):

GST, gst, PIK3CA, APAF1, BAD, BAX, BCL2, CASP3, REV3L, POLZ, FADD, PIK3R1, TOP2, POLH, ERK, MAPK1, TNFSF6, FASL, CD178, TNFRSF6, FAS, CD95, CASP8, CASP9, MAP3K5, ASK1, TP53, P53, AKT, BCL2L1, bcl-xL, XIAP, BIRC4, BID, ATM, TEL1, ERBB2, HER2, CD340, ABCC2, PDPK1, CDKN2A, P16, INK4A, CDKN1A, P21, CIP1, MDM2, BIRC5, MLH1, MSH2, MSH3, MSH6, CYC, PMAIP1, NOXA, BBC3, PUMA, BRCA1, XPA, ERCC1, BAK, BAK1, SLC31A1, CTR1, BIRC2, copA, ctpA, ATP7, GSTP, BIRC3, MAPK3, PIK3R3, PIK3R2, PIK3CB, PIK3CD

\*KEGG entry name: "Platinum drug resistance"

Detoxification Pathway (BioCyc Pathway)

GSTZ1, GSTA1, GSTA2, ABCC1, GGT1, GGT5, CCBL1, CCBL2, NAT8

\*Name of pathway in BioCyc: "glutathione-mediated detoxification I"

Glutathione Transferases (Uniprot\*) (with aliases)

PTGES, MGST1L1, MPGES1, PGES, PIG12, MGST3, MGST2, GST2, MGST1, GST12, MGST, LANCL1, GPR69A, HPGDS, GSTS, PGDS, PTGDS2, GSTZ1, MAA1, GSTT4, GSTTP1, GSTT2B, GSTT2, GSTT2, GSTT1, GSTP1, FAEES3, GST3, GSTO2, GSTO1, GSTTLP28, GSTM5, GSTM4, GSTM3, GST5, GSTM2, GST4, GSTM1, GST1, GSTK1, HDCMD47P, GSTCD, GSTA5, GSTA4, GSTA3, GSTA2, GST2, GSTA1

\*search string: name:glutathione name:transferase AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]"

Cytochrome P450 enzymes (Uniprot\*) (with aliases)

CYP2C9, CYP2C10, CYP21A2, CYP21, CYP21B, CYP3A4, CYP3A3, CYP2C8, CYP27B1, CYP1ALPHA, CYP27B, CYP26B1, CYP26A2, P450RAI2, CYP1A1, CYP11B2, CYP4A11, CYP4A2, CYP51A1, CYP51, CYP26A1, CYP26, P450RAI1, CYP2C19, CYP17A1, CYP17, S17AH, CYP1A2, CYP11B1, S11BH, CYP2A6, CYP2A3, CYP2C18, CYP27A1, CYP27, CYP3A5, CYP2B6, CYP2D6, CYP2DL1, CYP11A1, CYP11A, CYP4F2, CYP2E1, CYP2E, CYP4B1, CYP3A7, CYP4F3, LTB4H, CYP24A1, CYP24, CYP19A1, ARO1, CYAR, CYP19, CYP2A7, CYP7A1, CYP7, CYP4A22, CYP2F1, CYP4F12, UNQ568, PRO1129, CYP39A1, CYP7B1, CYP2A13, CYP26C1, CYP46A1, CYP46, CYP2U1, POR, CYPOR, CYP4F11, CYP4F22, CYP4V2, CYP2W1, CYP2S1, UNQ891, PRO1906, CYP2J2, CYP2R1, CYP4F8, CYP1B1, CYP3A43, CYP2D7, CYP27C1, TBXAS1, CYP5, CYP5A1, CYP8B1, CYP12, CYP4X1, UNQ1929, PRO4404, CYP4Z1, UNQ3060, PRO9882, CYP2G1P, CYP2GP1, CYP20A1, UNQ667, PRO1301, CYP4Z2P, CYP4F30P, C2orf14

\*search string: name:cytochrome name:p450 AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]"

ATP binding cassette (Uniprot\*) (with aliases)

TAP2, ABCB3, PSF2, RING11, Y1, TAP1, ABCB2, PSF1, RING4, Y3, CFTR, ABCC7, ABCG8, ABCG5, ABCG4, WHITE2, ABCG2, ABCP, BCRP, BCRP1, MXR, ABCG1, ABC8, WHT1, ABCD4, PXMP1L, ABCD3, PMP70, PXMP1, ABCD2, ALD1, ALDL1, ALDR, ALDRP, ABCD1, ALD, ABCC9, SUR2, ABCC8, HRINS, SUR, SUR1, ABCC6, ARA, MRP6, ABCC5, MRP5, ABCC4, MRP4, ABCC3, CMOAT2, MLP2, MRP3, ABCC2, CMOAT, CMOAT1, CMRP, MRP2, ABCC12, MRP9, ABCC11, MRP8, ABCC10, MRP7, SIMRP7, ABCC1, MRP, MRP1, ABCB9, KIAA1520, ABCB8, MABC1, MITOSUR, ABCB7, ABC7, ABCB6, MTABC3, PRP, UMAT, ABCB5, ABCB4, MDR3, PGY3, ABCB11, BSEP, ABCB10, ABCB1, MDR1, PGY1, ABCA9, ABCA8, KIAA0822, ABCA7, ABCA6, ABCA5, KIAA1888, ABCA4, ABCR, ABCA3, ABC3, ABCA2, ABC2, KIAA1062, ABCA13, ABCA12, ABC12, ABCA10, ABCA1, ABC1, CERP

\*search string: name:atp name:binding name:cassette AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]"

Cisplatin (Uniprot\*) (with aliases)

ATP11B, ATP1F, ATP1R, KIAA0956, LRRC8D, LRRC5, UNQ213, PRO239, SSRP1, FACT80, RAD23B, SIVA1, SIVA, PRIMPOL, CCDC111, LRRC8A, KIAA1437, LRRC8, SWELL1, UNQ221, PRO247, MCM8, C20orf154, ABCC2, CMOAT, CMOAT1, CMRP, MRP2, POLH, RAD30, RAD30A, XPV, SLC22A2, OCT2, SRSF2, SFRS2, DCLRE1A, KIAA0086, SNM1, SNM1A, FAM168A, KIAA0280, TCRP1, RDM1, RAD52B, XPC, XPCC, YBX1, NSE1, YB1, NOX3, MOX2, ADIRF, AFRO, APM2, C10orf116, DNAJC15, DNAJD1, GIG22, HSD18, TMEM205, UNQ501, PRO1018, CLPTMIL, CRR9

\*search string: (annotation:(type:function cisplatin) OR annotation:(type:"activity regulation" cisplatin) OR annotation:(type:disease cisplatin) OR annotation:(type:pharmaceutical cisplatin) OR annotation:(type:mutagen cisplatin)) AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]"

Cisplatin (Drugbank\*)

MPG, A2M, TF, ATOX1, MPO, XDH, CYP4A11, PTGS2, nat, CYP2C9, CYP2B6, BCHE, GSTT1, MT1A, MT2A, SOD1, GSTP1, NQO1, GSTM1, ALB, ABCC3, ABCC5, ABCC2, SLC22A2, SLC31A1, SLC31A2, ABCC6, ABCB1, ATP7B, ATP7A, ABCG2

\*Name of drug in Drugbank: "cisplatin"

ABHD12, ACS4, ACSL4, ACTG, ACTG1, ADMLX, AFG2, AIE75, AIGF, ALR, AMMECR1, AMMECR2, ANOS1, AP19, AP1S1, APNH1, ARSG, ATP1A3, ATP6B1, ATP6B2, ATP6N1B, ATP6N2, ATP6V0A4, ATP6V1B1, ATP6V1B2, AXOR12, BCS1, BCS1L, BFGFR, BHLHE32, BM28, BOM, BRWD2, BV8, C14orf10, C19orf64, C1orf7, C20orf22, C20orf54, C21orf29, C6orf125, C6orf29, C6orf32, C9orf75, C9orf81, CCNL1, CD164, CDC14A, CDCL1, CDH23, CEACAM16, CEAL2, CEK, CEP2, CEP250, CEP78, CGI-47, CHD7, CIAS1, CIB2, CLAPS1, CLDN14, CLLD7, CNAP1, COI, COL11A1, COL11A2, COL2A1, COL4A6, COL9A2, COLL6, COMT2, COXI, CRYM, CTL4, DCDC2, DFNA5, DFNB31, DFNB36, DIABLO, DIAP1, DIAP3, DIAPH1, DIAPH3, DIFF48, DLX5, DMXL2, DUSP6, E4.5, ECHOS1, EDG5, EIF3F, EIF3S5, ELMOD3, ENT3, EPS8L2, EPS8R2, ESPN, EXOSC2, EYA4, FACL4, FAM65B, FER1L2, FEZ, FEZF1, FGF17, FGF8, FGFR, FGFR1, FGFR3, FKHL7, FLG, FLRT3, FLT2, FOXC1, FP17425, FREAC3, G5PR, GAS3, GFER, GIPC3, GJB2, GJB6, GMPBB, GNRH, GNRH1, GNRHR, GPR54, GPR73L1, GRAP, GRH, GRHL2, GRHR, GSDME, HBGFR, HCCS4, HERV1, HGF, HOMER2, HPO, HPTA, HRIHFB2122, HS6ST, HS6ST1, HXB, IARS2, ICERE1, IL17RD, IL17RLM, ILDR1, IRX2A, IRX5, IRXB2, JTK4, KAL, KAL1, KALIG1, KCNE1L, KCNE5, KCNJ10, KCNQ4, KIAA0030, KIAA0386, KIAA0389, KIAA0567, KIAA0772, KIAA0856, KIAA1001, KIAA1154, KIAA1171, KIAA1351, KIAA1416, KIAA1469, KIAA1526, KIAA1662, KIAA1774, KIAA1812, KIAA1897, KIAA2034, KIP2, KISS1, KISS1R, KRML, LACS4, LHFPL5, LHRH, LP2654, LRP2, LRTOMT, MAFB, MAP3K20, MARS2, MARVELD2, MCM2, MET, MITF, MKP3, MKS3, MLTK, MNF1, MT-CO1, MTCO1, MYH14, MYH9, MYO15, MYO15A, MYO1F, MYO6, MYO7A, NALP3, NELF, NG22, NHE1, NK3R, NKNB, NLRP3, NRSF, NSMF, NTRKR1, OPA1, ORP2, OSBPL2, OTOA, OTOF, P2RX2, P2X2, PAF1, PAF3, PCDH15, PCNA, PDS, PEX1, PEX10, PEX12, PEX13, PEX2, PEX26, PEX5, PEX6, PI6, PKR2, PL48, PMP22, PMP3, PMP35, PP13181, PP4068, PP5098, PP7517, PPP2R3C, PRES, PRO1155, PRO1380, PRO1571, PRO1777, PRO1865, PRO187, PRO20026, PRO382, PRO4340, PRO874, PROK2, PROKR2, PRPS1, PTI, PTPRQ, PUS7, PXAAA1, PXMP3, PXR1, PYPAF1, PYST1, RAB40AL, RBED1, RBM29, RCBTB1, RDX, REST, RFT2, RFTV3, RIPOR2, RLGP, RNF69, RNF72, ROR1, RRP4, RU2, S1PR2, SANS, SEF, SEMA3A, SEMAD, SERPINB6, SLC17A8, SLC26A4, SLC26A5, SLC29A3, SLC44A4, SLC52A3, SLC9A1, SLITRK6, SMAC, SPAF, SPATA5, SPRY4, STRC, TAC3, TAC3R, TACR3, TADG12, TARA, TBC1D24, TBL1, TBL1Y, TECTA, TFCP2L3, THBP, TMC1, TMEM132E, TMEM67, TMHS, TMIE, TMPRSS3, TNC, TOMT, TPPT1, TPRN, TRIC, TRIOBP, TRNT1, TSPEAR, TUBB2C, TUBB4B, UNQ161, UNQ1894, UNQ323, UNQ441, UNQ585, UNQ6115, UNQ717, UNQ777, UNQ839, UNQ856, UQCC2, USH1B, USH1C, USH1F, USH1G, USH2A, VATB, VGLUT3, VPP3, VPP3, WBP2, WDR11, WDR15, WFS1, WHRN, XBR, XPNPEP3, ZAK, ZNF312B

\*search string: reviewed:yes AND organism:"Homo sapiens (Human) [9606]" AND (annotation:(type:disease \*Search string: "ototoxicity") OR annotation:(type:"disruption phenotype" "ototoxicity") OR annotation:(type:mutagen "ototoxicity") OR annotation:(type:function "ototoxicity") OR annotation:(type:pathway "ototoxicity"))

## Chapter 7

### Paper IV: Nephrotoxicity prediction

Sara L Garcia, Jakob Lauritsen, Zeyu Zhang, Mikkel Bandak,  
Marlene D Dalgaard, Rikke L Nielsen, Gedske Daugaard,  
Ramneek Gupta

**Prediction of Nephrotoxicity Associated With  
Cisplatin-Based Chemotherapy in Testicular Cancer  
Patients.**

*JNCI Cancer Spectrum, Volume 4, Issue 3 (June 2020)*









OXFORD

JNCI Cancer Spectrum (2020) 4(3): pkaa032

doi: 10.1093/jncics/pkaa032

Article

## Prediction of Nephrotoxicity Associated With Cisplatin-Based Chemotherapy in Testicular Cancer Patients

Sara L. Garcia , MSc,<sup>1,†</sup> Jakob Lauritsen , MD,<sup>2,\*‡</sup> Zeyu Zhang , MSc,<sup>1,3,‡</sup> Mikkel Bandak , MD,<sup>2</sup> Marlene D. Dalgaard , PhD,<sup>1</sup> Rikke L. Nielsen , MSc,<sup>1,4</sup> Gedske Daugaard , DMSc,<sup>2</sup> Ramneek Gupta , PhD<sup>1</sup>

<sup>1</sup>Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark; <sup>2</sup>Department of Oncology, Copenhagen University Hospital, Copenhagen, Denmark; <sup>3</sup>Key Laboratory of Genetic Network Biology, Institute of Genetics and Developmental Biology, University of Chinese Academy of Sciences, Beijing, China; and <sup>4</sup>Sino-Danish Center for Education and Research, Eastern Yanqihu campus, University of Chinese Academy of Sciences, Beijing, China

\*Correspondence to: Jakob Lauritsen, MD, Department of Oncology, Copenhagen University Hospital, Copenhagen, Blegdamsvej 9, Copenhagen 2100, Denmark (e-mail: jakob.lauritsen@regionh.dk).

†Authors contributed equally to this work.

### Abstract

**Background:** Cisplatin-based chemotherapy may induce nephrotoxicity. This study presents a random forest predictive model that identifies testicular cancer patients at risk of nephrotoxicity before treatment. **Methods:** Clinical data and DNA from saliva samples were collected for 433 patients. These were genotyped on Illumina HumanOmniExpressExome-8 v1.2 (964 193 markers). Clinical and genomics-based random forest models generated a risk score for each individual to develop nephrotoxicity defined as a 20% drop in isotopic glomerular filtration rate during chemotherapy. The area under the receiver operating characteristic curve was the primary measure to evaluate models. Sensitivity, specificity, and positive and negative predictive values were used to discuss model clinical utility. **Results:** Of 433 patients assessed in this study, 26.8% developed nephrotoxicity after bleomycin-etoposide-cisplatin treatment. Genomic markers found to be associated with nephrotoxicity were located at NAT1, NAT2, and the intergenic region of CNTN6 and CNTN4. These, in addition to previously associated markers located at ERCC1, ERCC2, and SLC22A2, were found to improve predictions in a clinical feature-trained random forest model. Using only clinical data for training the model, an area under the receiver operating characteristic curve of 0.635 (95% confidence interval [CI] = 0.629 to 0.640) was obtained. Retraining the classifier by adding genomics markers increased performance to 0.731 (95% CI = 0.726 to 0.736) and 0.692 (95% CI = 0.688 to 0.696) on the holdout set. **Conclusions:** A clinical and genomics-based machine learning algorithm improved the ability to identify patients at risk of nephrotoxicity compared with using clinical variables alone. Novel genetics associations with cisplatin-induced nephrotoxicity were found for NAT1, NAT2, CNTN6, and CNTN4 that require replication in larger studies before application to clinical practice.

Standard treatment in patients with disseminated testicular cancer is chemotherapy consisting of bleomycin-etoposide-cisplatin (BEP). Cisplatin is also central in the treatment of many other solid tumors such as bladder, ovarian, and lung cancer (1). Treatment containing cisplatin has a wide range of side effects, one of which is nephrotoxicity (2,3).

Cisplatin is excreted by the kidneys and may induce nephrotoxicity resulting in glomerular filtration rate (GFR) decline (4). Maintenance of sufficient renal function during treatment with chemotherapy is vital, and identification of patients at risk for developing nephrotoxicity could influence the treatment of choice if alternatives exist. Additionally, impaired renal function has been associated with increased risk of cardiovascular

disease (5), which may pose a problem in long-term cancer survivors.

Previous studies have improved the understanding of molecular mechanisms of cisplatin-induced nephrotoxicity (6), and several candidate gene studies have identified single-nucleotide polymorphisms (SNPs) associated with cisplatin-induced nephrotoxicity (7–9). However, these studies were conducted with surrogate measures of GFR (creatinine clearance or estimated GFR) rather than measured GFR as outcome.

The scope of this study was 2-fold: first, to conduct a genome-wide association study (GWAS) using a linear model controlling for cisplatin dosage (high or normal) to identify new genetic variants associated with cisplatin-induced

nephrotoxicity; and second, to investigate the utility of germline genetic markers together with clinical prognostic factors to predict nephrotoxicity using a random forest-recursive feature elimination algorithm. Patients treated for disseminated testicular cancer were chosen for this study because this patient group does not normally have comorbidity, which could influence renal function.

## Methods

### Patients

Patients were identified in the Danish Testicular Cancer-Late cohort (10), which includes 2572 Danish patients treated for testicular cancer from 1984 through 2007. Clinical features from 433 patients were originally extracted from hospital files as registered in the Danish Testicular Cancer database (Table 1). In 2014, all patients with measurements of renal function before and after treatment with BEP were invited to deliver a saliva sample for DNA analysis (Supplementary Figure 1, available online). Patients provided informed consent, and the study was approved by the regional ethical committee (H-2-2012-044) and the National Board of Data Protection (2012-41-0751).

### Treatment and Renal Measurement

All 433 patients received 3 cycles or more of BEP. The majority received normal-dose cisplatin  $20\text{ mg/m}^2 \times 5$  q3w, etoposide

$100\text{ mg/m}^2 \times 5$  q3w, and bleomycin  $15\text{ IU/m}^2$  q1w, and 25 patients received double-dose cisplatin and etoposide: cisplatin  $40\text{ mg/m}^2 \times 5$  q3w, etoposide  $200\text{ mg/m}^2 \times 5$  q3w, and bleomycin  $15\text{ IU/m}^2$  q1w. Hydration remained uniform over time with 2L isotonic saline before cisplatin and an additional 1-2L after. Diuretics were administered only in special cases, and no magnesium was added to hydration. There was no predefined cutoff of renal function where patients would not receive cisplatin-based triplets; however, to ensure toxicity was related to treatment, only patients with a GFR greater than  $90\text{ mL/min/1.73m}^2$  before chemotherapy were included.

GFR was measured by the 1-sample  $^{51}\text{Cr}$ -ethylenediamine-tetra acetic acid clearance technique using 2 samples 200 minutes after tracer injection and normalized to a body surface area (BSA) of  $1.73\text{ m}^2$ .

### Genomic Information

Genomic DNA was collected and purified using GeneFiX Saliva DNA Midi Kit from Isohelix (Harrietsham, UK). DNA samples were prepared at DTU Multi-Assay Core (Lyngby, Denmark) and genotyped at AROS Applied Biotechnology A/S (Aarhus, Denmark) using Illumina HumanOmniExpressExome-8 v1.2 chip (964 193 markers).

Genomic data were filtered using standard quality control steps (Supplementary Figure 2, available online). GWAS testing for single SNP association was conducted using PLINK (11)

**Table 1.** Comparison of baseline characteristics between affected (GFR high-drop) and nonaffected patients <sup>a</sup>

Characteristics	Affected, No. (%)	Nonaffected, No. (%)	P <sup>b</sup>
No. of patients	116 (26.8)	317 (73.2)	
Clinical characteristics			
Age, median (IQR)	34 (27-43)	30 (26-37)	.001
BEP regimen			
Normal dose	92 (79.3)	295 (93.4)	<.001
Double dose	24 (20.7)	21 (6.6)	
Unknown	—	1	
GFR before treatment, median (IQR), mL/min/1.73 m <sup>2</sup>	128 (115-139)	119 (110-131)	.001
GFR after treatment, median (IQR), mL/min/1.73 m <sup>2</sup>	88 (75-99)	109 (100-119)	<.001
Cisplatin, median (IQR), mg/m <sup>2</sup>	400 (391-410)	400 (300-400)	<.001
Treatment cycles			
3	20 (17.2)	97 (30.6)	<.001
4	72 (62.1)	199 (62.8)	
5 or more	6 (5.2)	14 (4.4)	
High dose	18 (15.5)	7 (2.2)	
Histology			
Seminoma	23 (19.8)	68 (21.5)	.78
Nonseminoma	93 (80.2)	249 (78.5)	
Prognostic group			
Good	71 (61.2)	277 (87.4)	<.001
Intermediate	30 (25.9)	35 (11.0)	
Poor	15 (12.9)	5 (1.6)	
Stage			
Extragenital	15 (12.9)	15 (4.7)	.87
Stage Ia	7 (6.0)	30 (9.6)	
Stage IIa	22 (19.1)	80 (25.5)	
Stage IIb	21 (18.1)	77 (24.5)	
Stage IIc	23 (19.8)	42 (13.4)	
Stage III	28 (24.1)	70 (22.3)	
Unknown	—	3	

<sup>a</sup>BEP = bleomycin-etoposide-cisplatin; GFR = glomerular filtration rate; IQR = interquartile range.

<sup>b</sup>P values were calculated by 2-sided Mann-Whitney U test for continuous or ordinal characteristics. For "histology," P value was calculated by  $\chi^2$  test.

(v1.9beta3), with the GFR decline after chemotherapy as the measure of toxicity and discretized cisplatin dosage as covariate with double-dose and normal-dose groups. The cutoff of 5 cycles was made to differentiate between normal and historically higher doses of cisplatin.

SNPs were annotated by ANNOVAR (v2015-06-17) (12) against the human reference genome hg19. Gene expression profiles were retrieved from GTExPortal (13).

We used a suggestive  $P$  value threshold of  $1 \times 10^{-5}$  (14) and a stringent threshold of  $8.02 \times 10^{-8}$  [Bonferroni corrected (15)].

In addition to the GWAS hits, 4 SNPs, rs11615 and rs3212986 (ERCC1), rs13181 (ERCC2), and rs316019 (SLC22A2), found in previous literature to be associated with cisplatin-induced nephrotoxicity (9), were added to the input feature search space in the machine learning modeling.

### Clinical Information

The clinical features used as input feature variables in the machine learning model were age at time of treatment, GFR before treatment, cumulative cisplatin dose per square meter of BSA, normal dose vs double-dose BEP, number of treatment cycles, histology (seminoma vs nonseminoma), prognostic classification as per IGCCCG (16) and stage of the disease as surrogate for size of retroperitoneal tumor size, which was represented as 3 features in the model (details on [Supplementary Methods](#), available online).

### Statistical Analysis and Model Development

A random forest model (17), which identified different risk subgroups of GFR drop, was developed using SciKit-learn (18) in Python (v3.7.1). A GFR decline of more than 20% after chemotherapy was chosen as outcome to indicate a clinically significant change and to avoid selection of cases due to random variation. A 20% decline has been associated with, for example, cognitive deterioration (19) and risk of cardiovascular and all-cause mortality compared with those with stable GFR (20).

As a first stage, the predictive power of a model driven by clinical features only was established. In a second stage, genomic markers were added to the model.

From all 433 individuals, about 20% (78 individuals: 20 nephrotoxicity affected) of the data, with no missing values, was randomly separated ahead of time to be used as a holdout set. Therefore, for machine model training, we omitted those 78 individuals present on the holdout set and excluded individuals with missing data in either clinical or genomic data ([Supplementary Figure 1](#), available online). Patients' baseline characteristics in each of these sets are available in [Supplementary Table 2](#) (available online).

Training and testing of the algorithm was performed with a 5 outer, 2 inner fold nested cross-validation (21,22) ([Supplementary Figure 3](#), available online).

The sample-splitting process for training and testing cohorts was random and repeated 100 times. Area under the receiver operating characteristic curve (ROC-AUC) was used as the primary performance measure for model optimization.

A recursive backwards feature elimination approach was used for feature selection initiated with 10 clinical features and then reduced (23). To identify when the algorithm should stop removing features, a paired  $t$  test (level of statistical significance,  $P < .05$ ) was calculated for each round of feature elimination on mean ROC-AUCs ([Figure 1, A and B](#)). A

statistically significant AUC drop ( $P < .05$ ) was indicative of an important feature being eliminated. All statistical tests were 2-sided. Details on model optimization and variable importance are described in the [Supplementary Methods](#) (available online).

The top-ranked clinical features constituted the baseline for adding prioritized SNPs from GWAS (17 SNPs) and the literature (4 SNPs), and feature selection was done using recursive backwards feature elimination approach.

### Polygenic Risk Score (PRS)-Derived Models

We also calculated PRS-derived models weighted by effect sizes estimated by the GWAS using the R-Package PRSice (24). These were tested in the random forest models in place of individual SNPs. Two different approaches were used: the risks associated with all the 21 SNPs were combined to determine a PRS, and a PRS per gene was estimated.

### Model Performances and Risk Groups

The primary reported performance was assessed with a 0.50 cutoff on the random forest model scores. In addition, to determine clinical applicability, we assessed different cutoffs on the random forest scores with a goal of 10% false discovery or omission rate (positive or negative predictive values  $> 90\%$ ).

For the SNPs and clinical-based models from the best round, the split that had a representative ROC-AUC close to the mean was used to assess different cutoffs (25) ([Supplementary Figure 4](#), available online).

Based on this, specific cutoffs for detection of 3 risk groups were used on the holdout set: a high-risk group for developing nephrotoxicity; a low-risk group for developing nephrotoxicity; and an intermediate group, which refers to individuals whose prediction is not adequately compelling to change the clinical decision.

## Results

### Study Population

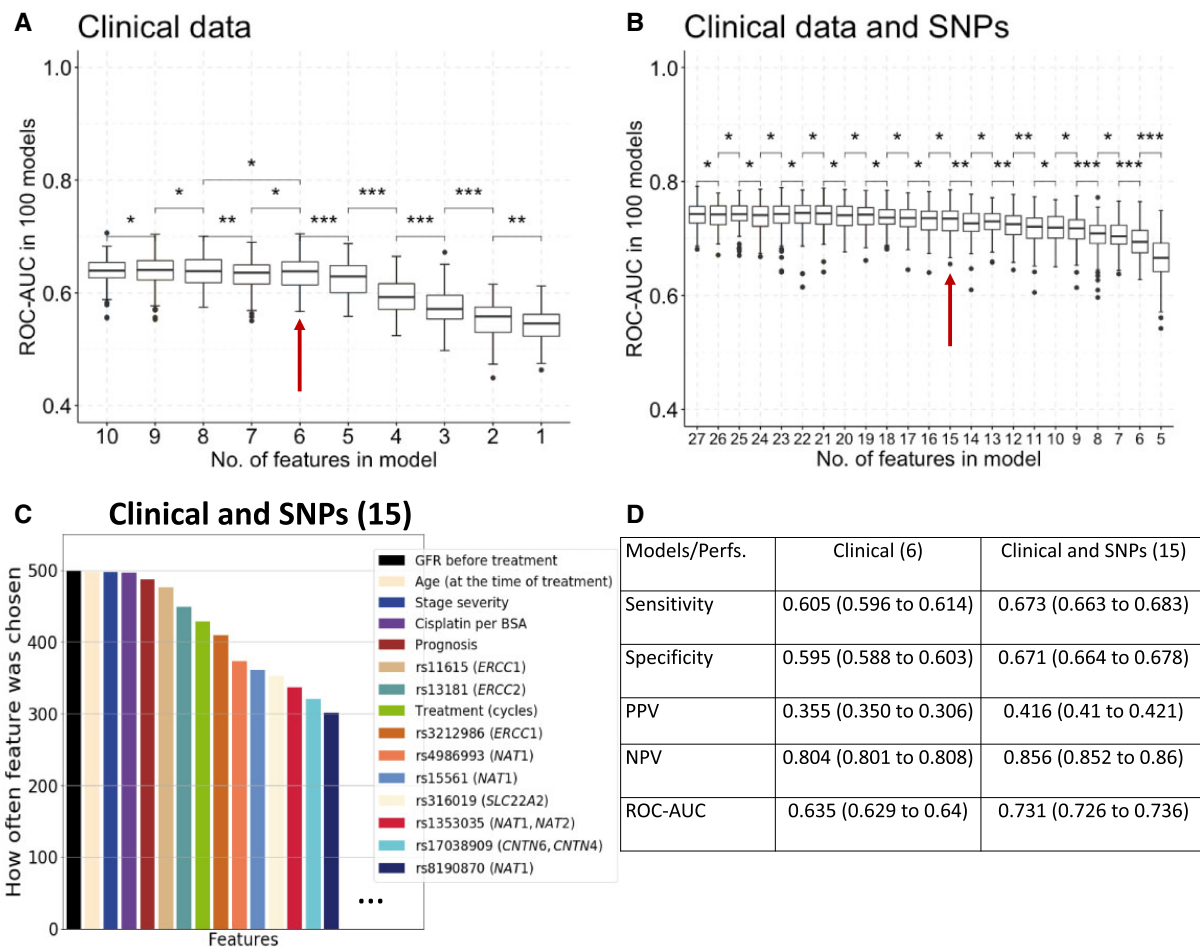
Overall, 433 individuals (26.8% nephrotoxicity affected) were assessed in this study, with a median (interquartile range [IQR]) age of 34 (27-43) years for affected patients ( $N = 116$ ) and 30 years (26-37) for nonaffected patients ( $N = 317$ ). The majority received 3 or 4 cycles of BEP. Before treatment, the median (IQR) GFR (mL/min/1.73 m<sup>2</sup>) was 128 (115-139) for affected and 119 (110-131) for nonaffected, and after treatment it decreased to 88 (75-99) for affected and 109 (100-119) for nonaffected ([Table 1](#)).

### Genome-Wide Association Study

Of 433 saliva samples received, 8 failed to yield high-quality genetic data. After quality control filtering, a total of 411 patients and 623 289 SNPs were eligible for GWAS ([Supplementary Figures 1 and 2](#), available online).

There was no indication of population stratification or inflation in the quantile-quantile plot of observed vs expected  $-\log_{10}$  ( $P$  values) ([Supplementary Figure 5](#), available online). GWAS controlling for cisplatin-based chemotherapy dosage identified 17 SNPs associated with GFR decline. Seven SNPs located





**Figure 1.** Feature selection using random forest-recursive feature elimination algorithm and diagnostic performances. **A** and **B**) Boxplots with different number of features, -10 to 1 and 27 to 5, for clinical and clinical plus genomics, respectively, and respective area under the receiver operating characteristic curve (ROC-AUC) throughout 100 different replications for data shuffling. Asterisks between boxplots represent P values (paired t test) of  $>.05$  (\*),  $\leq .05$  (\*\*), and  $\leq .01$  (\*\*\*). All tests were 2-sided. The red arrow represents the block chosen for further analysis. **C**) The features chosen the most on the 15-features clinical and SNP-based models. **D**) Performances obtained (mean and 95% confidence intervals) on the clinical models (6 features) and on the clinical and SNP-based models (15 features) using 0.50 cutoff for classification for sensitivity, specificity, positive predictive value, and negative predictive value. NPV = negative predictive value; Perfs. = performances; PPV = positive predictive value; ROC-AUC = area under the receiver operating characteristic curve; SNP = single-nucleotide polymorphism.

contiguous on chromosome 14 within the intergenic region between *LINC00645* and *FOXG1* passed a genome-wide statistical significance threshold of  $P = 8.02 \times 10^{-8}$  (Figure 2; Table 2). Nine additional SNPs located on chromosome 8, cytoband p22, passed a suggestive threshold of  $P = 1 \times 10^{-5}$  and were located in the intron and 3' untranslated region of *NAT1* or the intergenic region between *NAT1* and *NAT2*. SNP rs17038909 ( $P = 6.70 \times 10^{-8}$ ), located in the intergenic region between *CNTN6* and *CNTN4*, passed the genome-wide statistical significance threshold.

These 17 SNPs were included in input feature space of the machine learning models.

### Risk Prediction Model

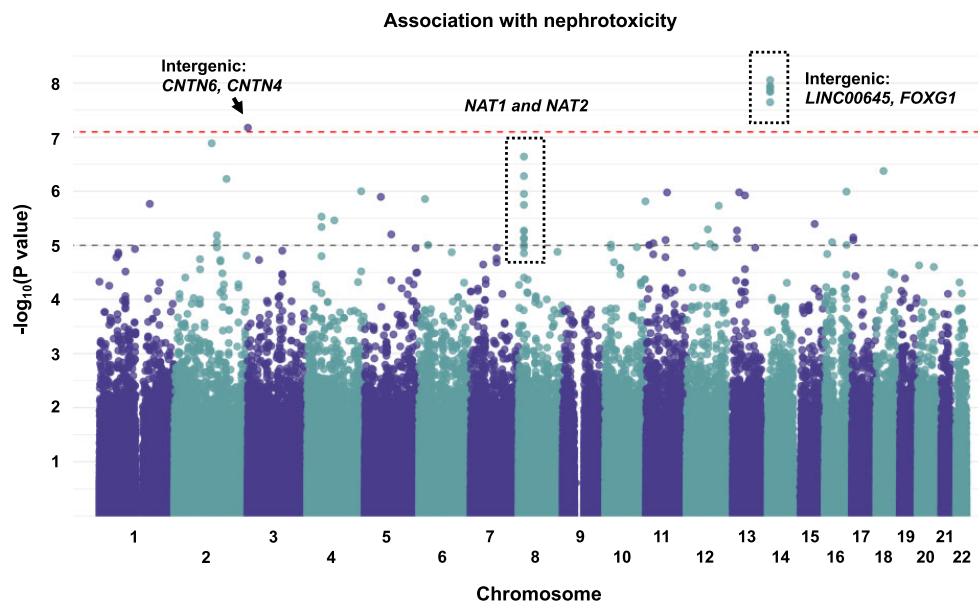
A baseline predictive model with only clinical features was trained using random forests. Of the initial 10 clinical features, 6 features were prioritized through recursive backwards elimination (Figure 1A): age at time of treatment, GFR before treatment, cumulative cisplatin-dose per square meter of BSA,

number of treatment cycles, prognostic classification as per IGCCCG (1)<sup>2</sup> (16), and stage of the disease, excluding group and histology. Univariate analysis also highlighted features selected in the random forest model (Table 1).

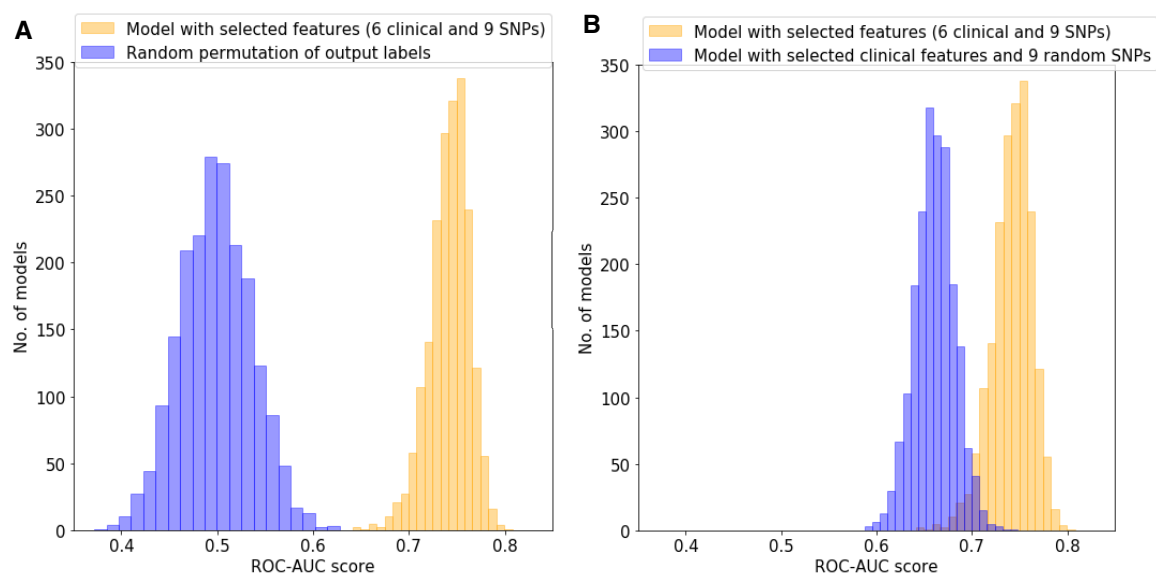
### SNPs and Clinical-Based Model

A selection of genomic markers was added to the baseline clinical prediction model: 17 SNPs from the GWAS and 4 additional SNPs from prior literature. Through recursive backwards elimination, 15 features were prioritized (6 clinical and 9 SNPs). The selected SNPs were rs11615 and rs3212986 (*ERCC1*), rs13181 (*ERCC2*), rs4986993, rs15561, rs8190870 (*NAT1*), rs1353035 (*NAT1/NAT2*), rs316019 (*SLC22A2*), and rs17038909 (*CNTN6/CNTN4*) (Figure 1, B and C). None of the SNPs located within the intergenic region between *LINC00645* and *FOXG1* were selected.

By adding genomic markers, ROC-AUC increased from 0.635 (95% confidence interval [CI] = 0.629 to 0.640) to 0.731 (95% CI = 0.726 to 0.736) (Figure 1D for additional performance metrics).



**Figure 2.** Genome-wide association study. Manhattan plot for association of 623 289 single-nucleotide polymorphisms with glomerular filtration rate decline. Linear model adjusted for cisplatin dosage was performed. The **black dashed line** represents a suggestive threshold:  $1 \times 10^{-5}$ , and the **red dashed line** represents a stringent Bonferroni corrected threshold:  $8.02 \times 10^{-8}$ . Markers in a contiguous pattern that pass the suggestive threshold are marked with a dotted box.



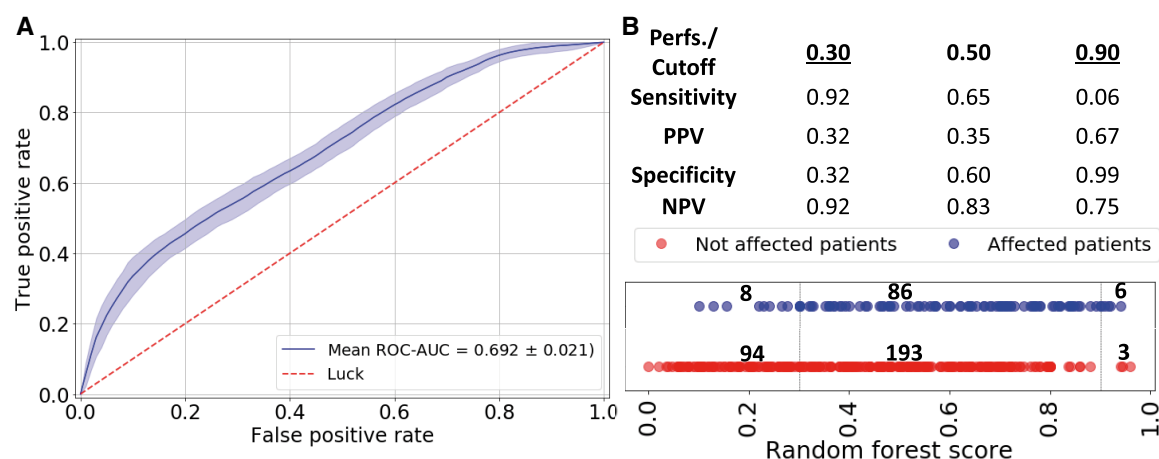
**Figure 3.** Benchmarking of the models. A) Test for random outcome simulated by permuting the labels 2000 times. B) Test for random single-nucleotide polymorphisms selection by combining 9 random markers, instead of the 9 selected markers, with the selected clinical traits. ROC-AUC = area under the receiver operating characteristic curve; SNP = single nucleotide polymorphism.

Additionally, 2 PRS were added independently to the baseline clinical model but did not outperform the individual SNPs (Supplementary Table 1, available online).

### Model Robustness

As a further validation, we tested for random outcome, simulated by permuting the labels 2000 times. This generated random performance for the model based on the clinical traits in

combination with the 9 SNPs previously reported, with a ROC-AUC mean of 0.498 (95% CI = 0.497 to 0.500). Furthermore, to assess if the SNP selection was meaningful, the performance of 9 random GWAS SNPs instead of the previously described 9 selected SNPs was tested when combined with the selected clinical traits; this process was repeated 2000 times. This performed very similarly to clinical traits alone, with a ROC-AUC mean of 0.661 (95% CI = 0.660 to 0.661) against the model scores with a ROC-AUC mean of 0.742 (95% CI = 0.741 to 0.743) (Figure 3).



**Figure 4.** Final model evaluation (clinical and genomic markers) on the holdout set. **A)** Area under (AUC) the receiver operating characteristic curve (ROC; mean and 95% confidence interval) analysis of clinical risk factors and genetic variables for prediction of cisplatin-based nephrotoxicity in testicular cancer patients using the holdout dataset. **B)** Diagnostic performances obtained with 3 prediction cutoffs and independent evaluation (random forest score) for each individual: 78 individuals ( $\times 5$  cross-validated models) (blue: affected; red: nonaffected). One validation external set was used. The 3 groups are represented: low-risk group (8% false negatives), undetermined zone, and high-risk group (33% false positives). Perfs. = performances; PPV = positive predictive value; NPV = negative predictive value; FN = false negatives; FP = false positives.

### Replication Dataset

The holdout set (78 individuals: 20 nephrotoxicity affected) was used for replication of the random forest models with clinical and genetic features. A ROC-AUC of 0.692 (95% CI = 0.688 to 0.696) was obtained on the final evaluation (Figure 4A).

A prediction cutoff of 0.90 and 0.30 for high risk and low risk, respectively, of developing nephrotoxicity was chosen for further analysis on 1 validation external set to discuss the model clinical utility. A random forest score between 0.30 and 0.90 was not enough to make a clinical decision. In the high-risk group, we had a positive predictive value of 0.67 (33% false discovery rate) and specificity of 0.99 while capturing 6% of all nephrotoxicity, whereas in the low-risk group we had a sensitivity of 0.92 and negative predictive value of 0.92 (8% false omission rate), which captured 32% of all nonaffected patients (Figure 4B).

### Discussion

In this study, we were able to predict patients at risk of developing nephrotoxicity after BEP chemotherapy based on clinical and genetic features with a machine learning algorithm. Clinical features selected on the random forests-driven baseline clinical model were known risk factors of renal toxicity (2) and were statistically significant in univariate analysis. The aim of the baseline model was to mimic and codify clinical intuition, which relies on the available clinical information at the time of treatment.

When genomic markers were added to the baseline model, prediction power substantially improved. We believe that genomic information, although not being predictive on its own, improves a baseline clinical model for identification of patients at risk for nephrotoxicity.

PRS did not perform as well as independent SNPs when added to the model, suggesting that nonlinear correlations between SNPs drove the increase in performance opposed to the linear combination that PRS offer, as has also been suggested elsewhere (26).

SNPs located in the *LINC00645* and *FOXG1* intergenic regions, although strongly associated in the GWAS ( $P = 5 \times 10^{-8}$ ), were not selected in the machine learning model because of either limited contribution or low minor allele frequencies (Table 2) that made it harder to detect in cross-validated setups.

SNPs rs4986993, rs15561, and rs8190870 (*NAT1*), rs1353035 (*NAT1/NAT2*), and rs17038909 (*CNTN6/CNTN4*) were newly discovered in the present GWAS to be associated with nephrotoxicity and added performance to the machine learning model.

*NAT1* and *NAT2* encode for arylamine *N*-acetyltransferases that take part in metabolizing drugs and chemical compounds in humans with a role in folate metabolism (27). These 2 genes encode similar protein sequences [identity = 81.03%, Clustal-Omega, Uniprot (28)], yet differ on expression profiles (13). *NAT1* is ubiquitously expressed in the central nervous system, and *NAT2* is specifically expressed in the liver, colon, and small intestine (Supplementary Figure 6, available online). It has been reported that cisplatin can impair *NAT1* by blocking its transferase activity in human breast cancer cells and impair murine *Nat2* activity in cultured mouse tissues (liver and kidney) (29), which on one hand contributes to the therapeutic effects of cisplatin, but on the other hand may lead to accumulation of cisplatin in the kidneys.

*CNTN6* and *CNTN4* encode for contacting proteins, which mediate cell surface interactions during nervous system development and have been suggested to be associated with neurodevelopmental disorders (30–32), though the association with nephrotoxicity needs to be further explored. SNPs found previously to be associated with nephrotoxicity were incorporated in this model. These SNPs were located at *ERCC1*, *ERCC2*, and *SLC22A2*.

*ERCC1* and *ERCC2* encode for excision repair proteins, and polymorphisms in *ERCC1/2* have been reported to alter *ERCC1/2* DNA repair function (33–35), which may affect nephron repair capacity after cisplatin exposure during chemotherapy (36–39). If not adequately repaired, cisplatin-induced DNA damage can induce cell death (40,41).

**Table 2.** Top GWAS hits and literature SNP hits for cisplatin-based nephrotoxicity in testicular cancer patients<sup>a</sup>

SNP	Gene	CHR	Position	Region/Consequence	Alleles (ref/alt)	MAF (all)	MAF (EUR)	P <sup>b</sup>
Top GWAS								
rs17038909	CNTN6, CNTN4	3	1467145	Intergenic	A/G	G: 0.10	G: 0.08	$6.70 \times 10^{-8}$
rs8190845	NAT1	8	18078628	Intronic	G/A	A: 0.20	A: 0.15	$1.79 \times 10^{-6}$
rs15561	NAT1	8	18080651	3 UTR	A/C	A: 0.44	A: 0.28	$2.29 \times 10^{-7}$
rs4986993	NAT1	8	18080747	3 UTR	T/G	T: 0.44	T: 0.28	$5.25 \times 10^{-7}$
rs8190870	NAT1	8	18081272	Downstream	C/T	T: 0.14	T: 0.15	$1.12 \times 10^{-6}$
rs13270034	NAT1, NAT2	8	18082354	Intergenic	G/A	A: 0.08	A: 0.13	$7.64 \times 10^{-6}$
rs13277177	NAT1, NAT2	8	18086096	Intergenic	A/G	G: 0.06	G: 0.10	$9.72 \times 10^{-6}$
rs13277481	NAT1, NAT2	8	18086217	Intergenic	A/G	G: 0.08	G: 0.13	$5.47 \times 10^{-6}$
rs13270961	NAT1, NAT2	8	18139163	Intergenic	T/C	C: 0.08	C: 0.11	$7.31 \times 10^{-6}$
rs1353035	NAT1, NAT2	8	18140633	Intergenic	C/T	C: 0.15	C: 0.17	$5.35 \times 10^{-6}$
rs17095485	LINC00645, FOXG1	14	28500775	Intergenic	C/T	T: 0.07	T: 0.06	$1.13 \times 10^{-8}$
rs17382424	LINC00645, FOXG1	14	28529219	Intergenic	C/T	T: 0.02	T: 0.06	$1.29 \times 10^{-8}$
rs4551947	LINC00645, FOXG1	14	28584430	Intergenic	C/A	A: 0.05	A: 0.06	$2.26 \times 10^{-8}$
rs8020589	LINC00645, FOXG1	14	28604708	Intergenic	C/T	T: 0.07	T: 0.06	$1.44 \times 10^{-8}$
rs10131751	LINC00645, FOXG1	14	28681216	Intergenic	C/A	A: 0.07	A: 0.07	$1.45 \times 10^{-8}$
rs9671720	LINC00645, FOXG1	14	28714229	Intergenic	C/T	T: 0.05	T: 0.04	$8.81 \times 10^{-9}$
rs12323487	LINC00645, FOXG1	14	28837771	Intergenic	C/A/T	A: 0.09	A: 0.05	$1.19 \times 10^{-8}$
Literature								
rs316019	SLC22A2	6	160670282	Missense	A/C	A: 0.14	A: 0.11	0.21
rs13181	ERCC2	19	45854919	Stop gained	T/A/G	G: 0.24	G: 0.36	0.03
rs3212986	ERCC1	19	45912736	Stop gained	C/A/G/T	A: 0.30	A: 0.25	0.11
rs11615	ERCC1	19	45923653	Synonymous	A/G	A: 0.33	G: 0.38	0.004

<sup>a</sup>Positions refer to assembly GRCh37. alt = alternative(s); CHR = chromosome; EUR = Europe; GWAS = genome-wide association study; MAF = minor allele frequency; ref = reference; ; SNP = single-nucleotide polymorphism; UTR = untranslated region.

<sup>b</sup>A linear model was adjusted for cisplatin dosage and scored by P values representing how likely the variant association was by random chance.

SLC22A2 encodes for organic cation transporter 2 (OCT2) protein, which is expressed in the proximal tubule epithelial cells of the kidney and involved in the absorption and excretion of xenobiotics and metabolites (42). OCT2 efficiently mediates cisplatin cellular uptake, leading to high cisplatin accumulation in renal proximal tubule cells (43) where cisplatin-induced nephrotoxicity typically occurs (44). OCT2 may be a key regulator in the renal accumulation of cisplatin, affecting drug handling and inducing nephrotoxicity (42,45).

During primary treatment of disseminated testicular cancer, about one-third of the patients develop cisplatin-induced nephrotoxicity (46,47).

This clinical and genomics-based model could be used as an early assessment for nephrotoxicity risk, assisting in identifying patients at high and low nephrotoxicity risk and influencing decisions on cisplatin chemotherapy cycles.

Using a 0.50 cutoff on the random forest model scores, we were able to achieve a sensitivity of 0.65, positive predictive value of 0.35, specificity of 0.60, and negative predictive value of 0.83. Differential thresholding of the nephrotoxicity model classified patients into high, low, and intermediate risk. For the high-risk group, the model correctly classified 67% of the patients who developed nephrotoxicity, yet only a small fraction of affected individuals was captured (0.06 sensitivity). On the other hand, for the low-risk group, the model correctly classified 92% of the patients who did not develop nephrotoxicity and captured 32% of the nonaffected population (Figure 4B).

Even though the model shows utility in the ability to predict toxicity throughout the score range, extreme cutoffs to identify the highest and lowest risk patients could point at the least disruptive implementation of such a model within current practice.

A strength of this study is the large dataset with a good representation of patients who developed nephrotoxicity after cisplatin-based chemotherapy, using exact renal measurements, and the first application, to our knowledge, of artificial intelligence on predicting such a phenotype.

The machine learning models appeared to be robust with stable performance across 100 random cross-validation splits of the training data, demonstrating performance of 0.731 mean ROC-AUC in cross-validation and 0.692 (95% CI = 0.688 to 0.696) ROC-AUC in the holdout set. Yet, as a limitation, the machine learning setups use some of the association results from the GWAS on the same cohort; therefore, replication on another cohort from an external dataset would be of substantial interest. NAT1 and NAT2 appear as interesting genetic targets to prioritize for assaying in future nephrotoxicity studies and would benefit from functional validation.

The ability to develop machine learning models for patient stratification in different nephrotoxicity risk groups has the potential to balance aggressive treatment against predicted toxicity risk.

In the future, toxicity may play a larger role in guiding treatment across several complex diseases, where data-driven prediction models may aid in decision making. Some of the clinical features used in this model, such as age at the time of treatment and GFR before chemotherapy as well as some of the identified genomics markers, could be applicable to other tumors types. Cisplatin is one of the most compelling drugs used in cancer treatment, and nephrotoxicity is a well-known side effect from its use. Our model could be applicable to ovarian, bladder, and lung cancer, where more elderly patients are at risk of nephrotoxicity and early identification of toxicity risks (or lack thereof) may influence treatment aggression or increase monitoring for selected patients.



## Funding

This work was supported by the Danish cancer society (R40-A2119). SLG was supported by Idella Foundation. ZZ and RLN were supported by Sino-Danish Center for Education and Research.

## Notes

**Role of the funder:** The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit.

**Conflicts of interest:** RG is employed with Novo Nordisk Research Centre Oxford since February 2020. The other authors have no conflicts of interest to disclose.

**Author contributions:** JL, GD, RG: Study concept and design. SLG, JL, ZZ, MB, RLN, RG: Acquisition, analysis, or interpretation of data. SLG, JL, ZZ: Drafting of the manuscript. SLG, JL, ZZ, MB, MDD, RLN, GD, RG: Critical revision of the manuscript for important intellectual content. SLG, ZZ: Statistical Analysis. MDD, GD, RG: Study supervision.

## References

- Dilruba S, Kalayda GV. Platinum-based drugs: past, present and future. *Cancer Chemother Pharmacol*. 2016;77(6):1103–1124.
- Lauritsen J, Mortensen MS, Kier MCG, et al. Renal impairment and late toxicity in germ-cell cancer survivors. *Ann Oncol*. 2015;26(1):173–178.
- Fung C, Fossa SD, Williams A, Travis LB. Long-term morbidity of testicular cancer treatment. *Urol Clin North Am*. 2015;42(3):393–408.
- Dasari S, Tchounwou PB. Cisplatin in cancer therapy: molecular mechanisms of action. *Eur J Pharmacol*. 2014;740:364–378.
- Astor BC, Hallan SI, Miller ER, Yeung E, Coresh J. Glomerular filtration rate, albuminuria, and risk of cardiovascular and all-cause mortality in the US population. *Am J Epidemiol*. 2008;167(10):1226–1234.
- Karasawa T, Steyger PS. An integrated view of cisplatin-induced nephrotoxicity and ototoxicity. *Toxicol Lett*. 2015;237(3):219–227.
- Nematbakhsh M, Pezeshki Z, Eshraghi Jazi F, et al. Cisplatin-induced nephrotoxicity; protective supplements and gender differences. *Asian Pac J Cancer Prev*. 2017;18(2):295–314.
- Achkar IW, Abdulrahman N, Al-Sulaiti H, Joseph JM, Uddin S, Mraiche F. Cisplatin based therapy: the role of the mitogen activated protein kinase signaling pathway. *J Transl Med*. 2018;16(1):96.
- Zazuli Z, Vijverberg S, Slob E, et al. Genetic variations and cisplatin nephrotoxicity: a systematic review. *Front Pharmacol*. 2018;9:1111.
- Kreiberg M, Bandak M, Lauritsen J, et al. Cohort profile: The Danish Testicular Cancer late treatment effects cohort (DaTeCa-LATE). *Front Oncol*. 2018;8:37.
- Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–575.
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
- Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580–585.
- MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*. 2017;45(D1):D896–D901. doi: 10.1093/nar/gkw1133
- Bland JM, Altman DG. Multiple significance tests: The Bonferroni method. *BMJ*. 1995;310(6973):170–170.
- International Germ Cell Cancer Collaborative Group. Germ cell consensus classification: a prognostic factor-based staging system for metastatic germ cell cancers. International Germ Cell Cancer Collaborative Group. *J Clin Oncol*. 1997;15(2):594–603.
- Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
- Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
- Chen Y-C, Weng S-C, Liu J-S, Chuang H-L, Hsu C-C, Tarn D-C. Severe decline of estimated glomerular filtration rate associates with progressive cognitive deterioration in the elderly: a community-based cohort study. *Sci Rep*. 2017;7(1):42690.
- Cheng T-Y, Wen S-F, Astor BC, Tao XG, Samet JM, Wen CP. Mortality risks for all causes and cardiovascular diseases and reduced GFR in a middle-aged working population in Taiwan. *Am J Kidney Dis*. 2008;52(6):1051–1060.
- Picard RR, Berk KN. Data splitting. *Am Stat*. 1990;44(2):140–147.
- Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7(1):91.
- Lai C, Guo S, Cheng L, Wang W. A comparative study of feature selection methods for the discriminative analysis of temporal lobe epilepsy. *Front Neurol*. 2017;8:633.
- Euesden J, Lewis CM, O'Reilly P. PRSice: Polygenic Risk Score software. *Bioinformatics*. 2015;31(9):1466–1468.
- Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*. 2008;56(1):45–50.
- Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. *Front Genet*. 2019;10:267.
- Sim E, Abuhammad A, Ryan A. Arylamine N-acetyltransferases: from drug metabolism and pharmacogenetics to drug discovery. *Br J Pharmacol*. 2014;171(11):2705–2725.
- Consortium TU. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2018;47(D1):D506–D515.
- Ragunathan N, Dairou J, Pluvineau B, et al. Identification of the xenobiotic-metabolizing enzyme arylamine N-acetyltransferase 1 as a new target of cisplatin in breast cancer cells: molecular and cellular mechanisms of inhibition. *Mol Pharmacol*. 2008;73(6):1761–1768.
- Hu J, Liao J, Sathanoori M, et al. CNTN6 copy number variations in 14 patients: a possible candidate gene for neurodevelopmental and neuropsychiatric disorders. *J Neurodev Disord*. 2015;7(1):26.
- Mercati O, Huguet G, Danckaert A, et al. CNTN6 mutations are risk factors for abnormal auditory sensory perception in autism spectrum disorders. *Mol Psychiatry*. 2017;22(4):625–633.
- Tassano E, Uccella S, Giacomini T, et al. Clinical and molecular characterization of two patients with CNTN6 copy number variations. *Cytogenet Genome Res*. 2018;156(3):144–149.
- Ni M, Zhang W, Qiu J, et al. Association of ERCC1 and ERCC2 polymorphisms with colorectal cancer risk in a Chinese population. *Sci Rep*. 2015;4(1):4112.
- Yang L, Ritchie A-M, Melton DW. Disruption of DNA repair in cancer cells by ubiquitination of a destabilizing dimerization domain of nucleotide excision repair protein ERCC1. *Oncotarget*. 2017;8(33):55246–55264.
- Basu A, Krishnamurthy S. Cellular responses to cisplatin-induced DNA damage. *J Nucleic Acids*. 2010;2010:1–16.
- Khrunin A V, Moiseev A, Gorbunova V, Limborska S. Genetic polymorphisms and the efficacy and toxicity of cisplatin-based chemotherapy in ovarian cancer patients. *Pharmacogenomics J*. 2010;10(1):54–61.
- Tzvetkov MV, Behrens G, O'Brien VP, Hohloch K, Brockmüller J, Benöhr P. Pharmacogenetic analyses of cisplatin-induced nephrotoxicity indicate a renoprotective effect of ERCC1 polymorphisms. *Pharmacogenomics*. 2011;12(10):1417–1427.
- Benhamou S, Sarasin A. ERCC2/XPD gene polymorphisms and cancer risk. *Mutagenesis*. 2002;17(6):463–469.
- Windsor RE, Strauss SJ, Kallis C, Wood NE, Whelan JS. Germline genetic polymorphisms may influence chemotherapy response and disease outcome in osteosarcoma: a pilot study. *Cancer*. 2012;118(7):1856–1867.
- Zamble DB, Lippard SJ. Cisplatin and DNA repair in cancer chemotherapy. *Trends Biochem Sci*. 1995;20(10):435–439.
- Rocha CRR, Silva MM, Quinet A, Cabral-Neto JB, Menck C. DNA repair pathways and cisplatin resistance: an intimate relationship. *Clinics (Sao Paulo)*. 2018;73(suppl 1):e478s.
- Nigam SK, Wu W, Bush KT, Hoenig MP, Blantz RC, Bhatnagar V. Handling of drugs, metabolites, and uremic toxins by kidney proximal tubule drug transporters. *Clin J Am Soc Nephrol*. 2015;10(11):2039–2049.
- Ciarimboli G, Deuster D, Knief A, et al. Organic cation transporter 2 mediates cisplatin-induced oto- and nephrotoxicity and is a target for protective interventions. *Am J Pathol*. 2010;176(3):1169–1180.
- Leibbrandt ME, Wolfgang GH, Metz AL, Ozobia AA, Haskins JR. Critical subcellular targets of cisplatin and related platinum analogs in rat renal proximal tubule cells. *Kidney Int*. 1995;48(3):761–770.
- Filipski KK, Loos WJ, Verweij J, Sparreboom A. Interaction of cisplatin with the human organic cation transporter 2. *Clin Cancer Res*. 2008;14(12):3875–3880.
- Prasaja Y, Sutandyo N, Andrajati R. Incidence of cisplatin-induced nephrotoxicity and associated factors among cancer patients in Indonesia. *Asian Pac J Cancer Prev*. 2015;16(3):1117–1122.
- Kidera Y, Kawakami H, Sakiyama T, et al. Risk factors for cisplatin-induced nephrotoxicity and potential of magnesium supplementation for renal protection. *PLoS One*. 2014;9(7):e101902.

## Additional work: Further investigation of *NAT1* and *NAT2* variants in the development of cisplatin-induced nephrotoxicity in testicular cancer patients

As a continuation of paper IV, and to explore further the genomic regions comprising *NAT1* and *NAT2*, a targeted NGS analysis was performed by the Beijing Genomics Institute (Hong Kong, China) using the HiSeqX platform (Illumina, San Diego, CA, USA) on the same patients. DNA probes have been previously designed to bind to several sequences of interest in multiple genes, and I was not part of this process. Here, due to time restrictions, only *NAT1* (chromosome 8:18027986-18081198) and *NAT2* (chromosome 8:18248755-18258728) were explored in detail.

Sentieon DNaseq software (Sentieon version 201808.03) pipeline was used. For each sample, the following steps were performed: 1) removal of duplicates, 2) mapping to the human reference genome hg19 using BWA algorithm, 3) realignment around insertions and deletions, 4) base-score recalibration, and 5) variant calling. Sentieon Haplotyper algorithm with option `--emit_mode gvcf` was used to generate a gVCF file per sample. Afterwards, a combined variant calling was performed using Sentieon GVCFTyper algorithm. Only bases above Q10 were kept in the gVCF files. After the intersection with the BED file, only markers located at chromosome 8:18027000-18259000 were included (873 variants).

There were few challenges with this dataset:

- Low coverage (below 10) for some of the samples. There are many potential reasons for low coverage, such as low sample quality, guanine-cytosine content, and sequences with many homologous, hypervariable and low complexity regions. Additionally, poorly designed probes could have been an issue. Here, we observed a pattern of every ten samples with either low or high coverage (Figure 7.1). During the experimental setup, samples had been pooled together in blocks of 10 samples per sequencing lane in the flow cell. Thus, the issue seems to point to the flow cell sequencing lanes. To ensure that poorly covered variants were not included in the analysis, the VCF file was further filtered using a depth of 10.
- Due to some lost translation between patients IDs and samples, I have written an in-house script to compare genomic markers available in both microarray and targeted sequencing, to match the samples with patient IDs.

Once data was "ready" to use, genomic variants present in samples with a glomerular filtration rate drop of more than 5% after chemotherapy were retrieved (265 variants).

This not so strict "threshold" was used to ensure no significant variants were lost in the process. In Figure 7.2, SNPs are represented by order of frequency in the dataset, and only SNPs that appear more than four times are present (44 variants).

Ensembl Variant Effect Predictor was used to determine the effect of each variant. Three variants had a moderate impact: rs4987076, rs4986783 and rs56172717, meaning that they may change protein effectiveness.

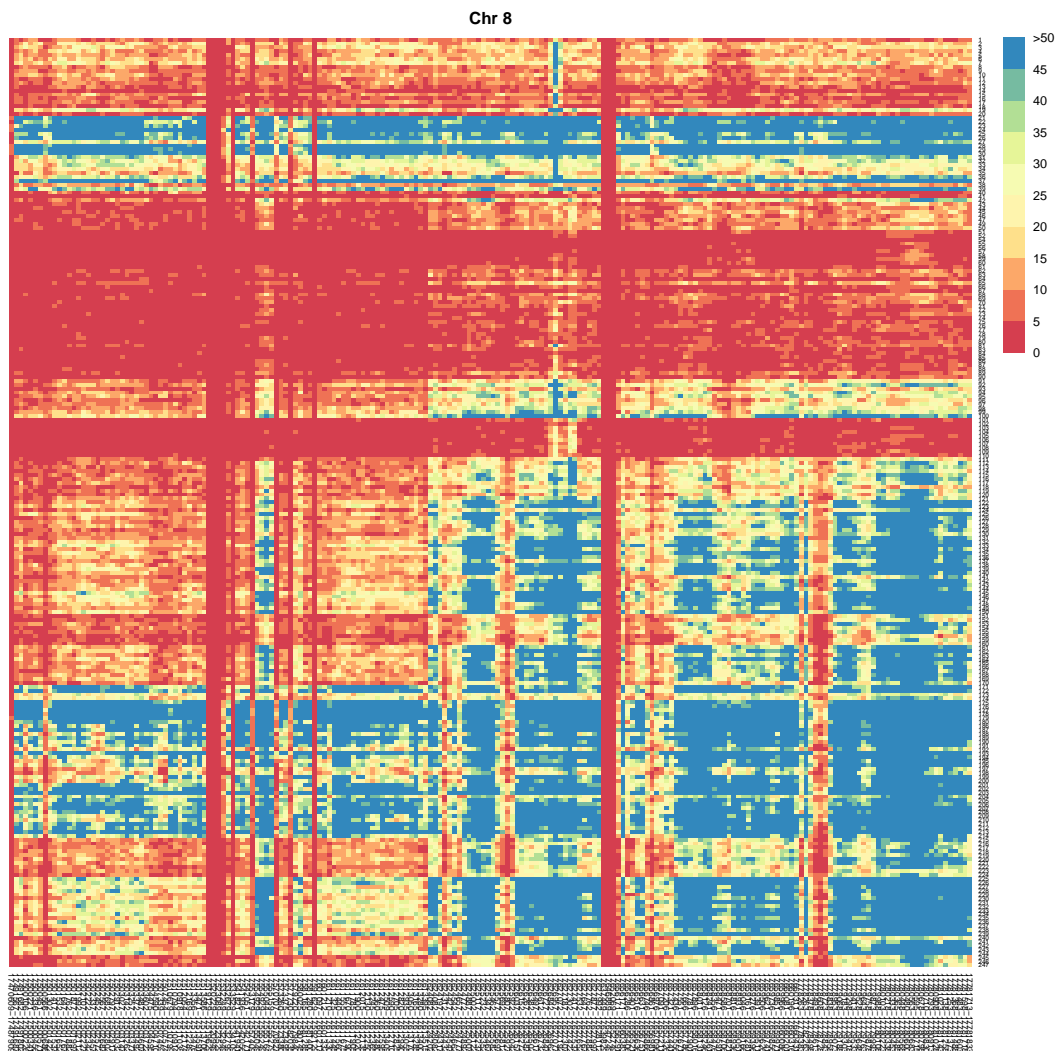
SNPs rs4987076 and rs4986783 also appeared with higher frequency in the dataset. Furthermore, these SNPs are in high linkage disequilibrium [156] with the SNPs found in the GWAS in paper IV (Figure 7.3). These two SNPs are present in the same seven samples with glomerular filtration rate decline of 37, 23, 19, 18, 7 (x2) and 6%. SNP rs4987076 is present in one additional sample with a glomerular filtration rate decline of 33%.

For SNP rs56172717, it was classified as malignant in both SIFT and Polyphen. This SNP was present in two samples with glomerular filtration rate decline of 22 and 14%.

Other variants from this 44 SNP list would be worth exploring (Table 7.1). From those 44 variants, six were located in regulatory regions, and two were located in transcription factor binding sites. None of these SNPs in Table 7.1 were considered on the GWAS in paper IV.

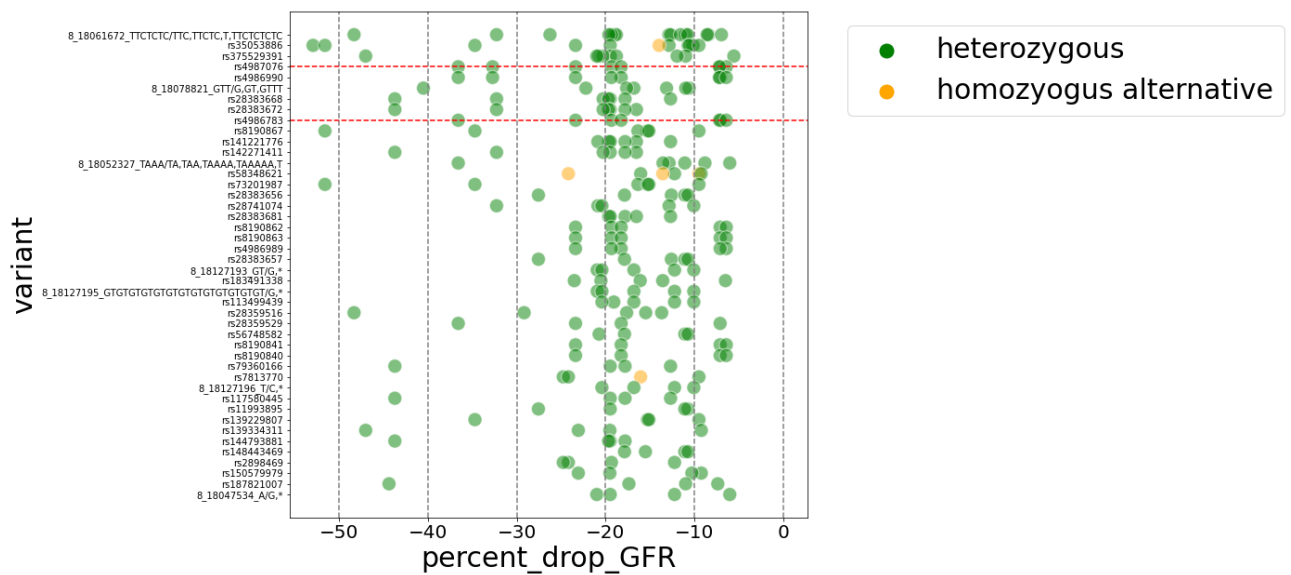
**Table 7.1** | Potential SNPs worth to explore in the future as for the consequence.

Uploaded variant	Location	MAF	Consequence	Impact	Symbol	SIFT	PolyPhen
rs117733044	8:18037479	<0.01 (A)	regulatory_region_variant	-	<i>NAT1</i>	-	-
rs73666897	8:18053812	0.01 (C)	regulatory_region_variant	-	<i>NAT1</i>	-	-
rs28383681	8:18054584	0.01 (G)	regulatory_region_variant	-	<i>NAT1</i>	-	-
rs28383686	8:18054844	0.12 (C)	regulatory_region_variant	-	<i>NAT1</i>	-	-
rs28359484	8:18067876	0.05 (T)	TF_binding_site_variant	-	<i>NAT1</i>	-	-
rs28359489	8:18068149	0.01 (A)	TF_binding_site_variant	-	<i>NAT1</i>	-	-
rs4987076	8:18080001	0.02 (A)	missense_variant	MODERATE	<i>NAT1</i>	1	0.003
rs4986783	8:18080196	0.02 (G)	missense_variant	MODERATE	<i>NAT1</i>	0.8	0.001
rs56172717	8:18080308	<0.01 (T)	missense_variant	MODERATE	<i>NAT1</i>	0	1
rs7834402	8:18145281	0.14 (C)	regulatory_region_variant	-	-	-	-
rs7818916	8:18145405	0.13 (G)	regulatory_region_variant	-	-	-	-

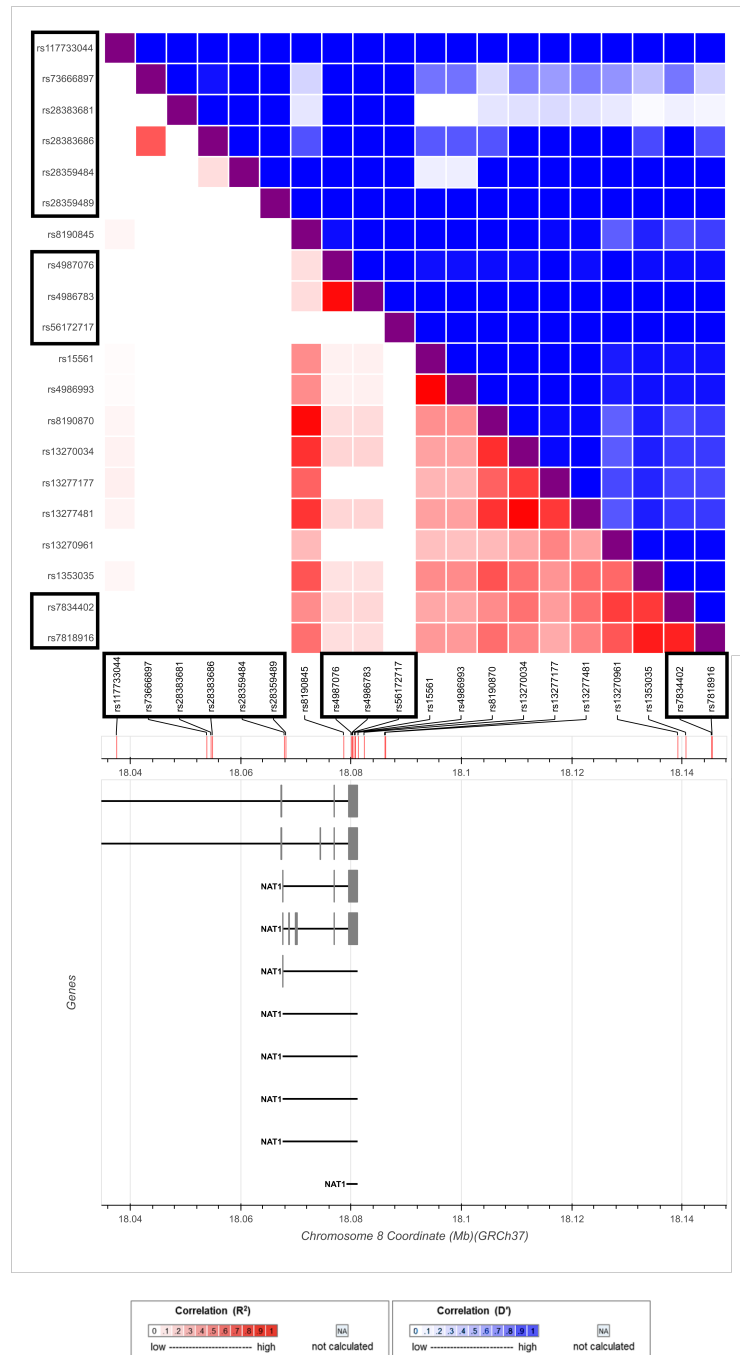


**Figure 7.1** | Representative illustration of coverage throughout targeted sequences. Here, only 247 samples (y-axis) and few regions on genes *NAT1* and *NAT2* (x-axis) are included. This pattern was observed for other regions as well. Figure generated by Freja Dahl Hede.





**Figure 7.2** | Genomic variants (y-axis) present in patients with more than 5% drop in glomerular filtration rate (x-axis). Each dot represents a samples and colors stand for heterozygous (green) or homozygous for the alternative allele (yellow). Red dashed lines stand out variants rs4987076 and rs4986783 with moderate impact.



**Figure 7.3** | Linkage disequilibrium between 11 genomic variants present in table 7.1 (black box) and genomic variants described in paper IV [157]. Plots generated from LDlink [156].

## Part III

### Other projects

## Chapter 8

# Application note: Fluctuation measures

Sara L Garcia, Cecilia B. Jensen, Rikke Linnemann Nielsen,  
Ramneek Gupta

**FLUCbio: a python package for fluctuation modelling  
on postprandial biological data**

---

# FLUCbio: a python package for fluctuation modelling on postprandial biological data

Sara L. Garcia<sup>1</sup>, Cecilia B. Jensen<sup>1</sup>, Rikke Linnemann Nielsen<sup>1,2</sup>, and Ramneek Gupta<sup>1,3\*</sup>

<sup>1</sup>Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark

<sup>2</sup>Department of Paediatrics and Adolescent Medicine, Rigshospitalet, University Hospital of Copenhagen, Denmark

<sup>3</sup>Department of Computational Biology, Novo Nordisk Research Centre Oxford, Oxford, United Kingdom

\*To whom correspondence should be addressed.

## Abstract

**Motivation:** Glucose and haemoglobin A1C levels over time are useful to predict clinical outcomes, such as pre-diabetes, type 2 diabetes and cardiovascular diseases, however, there is a challenge on how to capture the fluctuation patterns of these temporal curves as cardiometabolic disease progress. Area under the curve is still a preferred method to profile these curves, however, some information is lost in this translational process. Here, we present FLUCbio, a package that outputs different fluctuation measurements.

**Results:** Our described methods captured other characteristics from postprandial curves which were not captured by the area under the curve approach, thus this could be used as complementary methods.

**Availability:** FLUCbio source code is freely available at GitHub\*\*. All code was implemented in Python.

**Contact:** rmgp@novonordisk.com

\*\*Code will be made available once application note is published.

---

## 1 Introduction

In biological research many variables contain temporal patterns, which makes them time-dependent. In many of these biological processes it is important to understand the pattern and if the output is something that we want to control. For example, for pre-diabetes and type 2 diabetes (T2D), postprandial hyperglycaemia is a serious threat and one of the earliest signs of glucose homeostasis associated with T2D (American Diabetes Association 2001). Understanding the individual variation of glycemic measurements are important markers for early metabolic diagnosis that can help target timely lifestyle interventions before cardiometabolic disease progress (Brezina, Orekhova, and Weiss 1997)(Kavakiotis et al. 2017).

Due to the huge variability of glucose responses across people, several methods (Hulman, Vistisen, et al. 2018); (Hulman, Witte, et al. 2018); (Hall et al. 2018); (Schüssler-Fiorenza Rose et al. 2019) have been developed for categorisation of these postprandial curves, and hence, classification of patients diabetes status. Several studies have also used the well-known area under the curve (AUC) measured by continuous glucose monitoring to profile glycemic patterns (Freckmann et al.

2007); (Jackson et al. 2010); (Zeevi et al. 2015); (Søndertoft et al. 2020). However, the same AUC can hide very different individual postprandial responses, therefore using it as the only measure may not be the most optimal approach. Specifically, the fluctuation of the response variables convey information on stability of glucose control, which is not captured by AUC.

In this study, we developed three simple heuristic measures of curve volatility to capture the fluctuation of postprandial curves and we demonstrate that these measures add complementary information to AUC.

### Temporal curve: variation and fluctuation

Variation and fluctuation are two terms which many times are used interchangeable (Hajime et al. 2018), however, they represent different characteristics of a postprandial curve (Figure 1), which are import to identify inter- and intra-individuals responses. AUC captures the curve variation, and does a great job if we only observe steady changes over time; however, when observing more irregular changes over time, the AUC may not be the most optimal measure to use, as we will show in the next section.

## 2 Methods

The methods described in this application note can be used as complementary methods to the AUC, in order to capture as much information as possible from a temporal curve. All implementations can be found at GitHub.

### Fluctuation measures

#### Fluctuation measure 1

On fluctuation measure 1, we calculate the difference of the first two consecutive time points ( $\Delta_1 = \text{time point 2 } (y_2) - \text{time point 1 } (y_1)$ ).  $\Delta_1$  is positive or negative, in case of a positive or negative slope, respectively. Next, we calculate the difference between time point 3 and time point 2 ( $\Delta_2$ ). Finally, we calculate the absolute difference of  $\Delta_1$  and  $\Delta_2$  ( $\alpha_1$ ). If  $\Delta_1$  and  $\Delta_2$  are both positive or negative, the absolute difference of those will be smaller than if  $\Delta_1$  and  $\Delta_2$  have opposite signs (Equation 1, Figure 2). A curve which changes direction constantly will therefore have a higher fluctuation measure.

$$\begin{aligned} fluc(y) &= \sum_{i=2}^{len(x)} abs((y_i - y_{i-1}) - (y_{i+1} - y_i)) = abs(\Delta_1 - \Delta_2) + abs(\Delta_2 - \Delta_3) + abs(\Delta_3 - \Delta_4) + abs(\Delta_4 - \Delta_5) \\ &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \end{aligned}$$

**Equation 1:** Fluctuation measure equation.

#### Fluctuation measure 2

On fluctuation measure 2, we use a grid analysis approach where we plotted the non-normalized temporal profile for each patient. The image of the temporal profile is segmented into a grid of  $n \times n$  size,  $n$  being user-defined.

All individuals should have the same number of postprandial measurements and these should be evenly-spaced over time.

If missing data is present, interpolation is performed with function `interp1d` from SciPy (version 1.2.1) Python package.

The presence of the temporal profile line in the grid boxes is translated into 0s and 1s for non-presence and presence respectively. In the end, an image vector is created from the grid reading vertically from lower to upper boundary, concatenating the columns into one vector (Figure 3). The measurement obtained is the sum of the 1s in the vector. Curves with higher glycaemic fluctuation will have a higher sum.

### Fluctuation measure 3

Fluctuation measure 3 is obtained by filtering the sum, and only summing up two or more consecutive 1s in the vector obtained in fluctuation measure 2.

The aim of using this fluctuation measure is to remove spurious measurements, as we are taking a continuous shift in the amplitude.

### Other measures

Using FLUCbio it is also possible to calculate the AUC or a simple variation measure. For the AUC we have used the trapezoidal rule performed by the function `numpy.trapz` (v.1.16.4). To calculate the variation of the curve, we have used Equation 2.

$$var(y) = \frac{\sum_{i=2}^{len(y)} abs(y_i - y_{i-1})}{len(y)}$$

**Equation 2:** Variation measure equation.

## 3 Example analysis

### Example 1: methods applicability

Fluctuation measures 2 and 3 were used in our previous paper to predict gain weight loss on individuals following a whole grain, low gluten or refined grain diet during 8 weeks randomized clinical trials using machine learning models (Nielsen et al. 2020). Briefly, the cohort consisted of generally healthy Danish adults at a cardiometabolic risk whose postprandial biomarkers was reported including plasma glucose concentrations measured after a standardized breakfast at five time points (0,30,60,120,and 180min). Data was imputed using linear imputation (available in FLUCbio package) so the measurements were evenly-spaced over time (in this case, every 30 minutes). Ten examples of comparisons between samples whose postprandial glucose curves have the same AUC, but it shows very different fluctuation responses is represented in Figure 4. These differences were captured using the three fluctuations measures described in “methods”.

### Example 2: correlation with biomarkers

To show further utility of the presented methods, we have run FLUCbio in an online available dataset which consisted of glucose concentrations (in mg/dL) values from 30 minutes before a standardized meal and 2.5 hours after, drawn every 5 minutes, thus a total of 37 time points on a total of 30 samples were available (Supporting Information, S6 from (Hall et al. 2018)). We have only included participants who were put on the standardized meal consisting of peanut butter sandwich. Linear imputation was done and 100 time points were used. The correlation between several biomarkers, glucose AUC and glucose fluctuation measures were calculated using Spearman’s rank or Pearson’s correlation

---

depending on the normal distribution of the data. For the total cholesterol, high-density lipoprotein (HDL) cholesterol, and low-density lipoprotein (LDL) cholesterol, we could observe higher correlation using the glucose fluctuation measures than the glucose AUC, (Figure 5) emphasizing the need of having other measurements which can add extra information to the AUC.

## 4 Conclusion

The interest in the inter- and intra-individual differences in postprandial glucose responses has been gaining major interest. Understanding and knowing how postprandial biological data can be modelled is important in multiple fields, such as diet management, diabetes research and care, or to gain insights into the transition of human metabolic mechanism from normal to impaired glucose tolerance.

The presented three methods shown to capture other characteristics of postprandial curves that were not captured by the AUC alone, thus can be used as complementary methods.

These methods output a number which have a biological meaning, however the output depends on the units used, in this case, we have used glucose values. This needs to be taken into account if we want to compare between studies with different units.

These methods were only tried on postprandial data, and no other types of longitudinal data. Furthermore, the examples used in this application note were only used for glucose values, however, this can be extended to other variables such as free fatty acid, insulin, HbA1c, glucagon-like peptide-1 receptor agonists.

## Funding

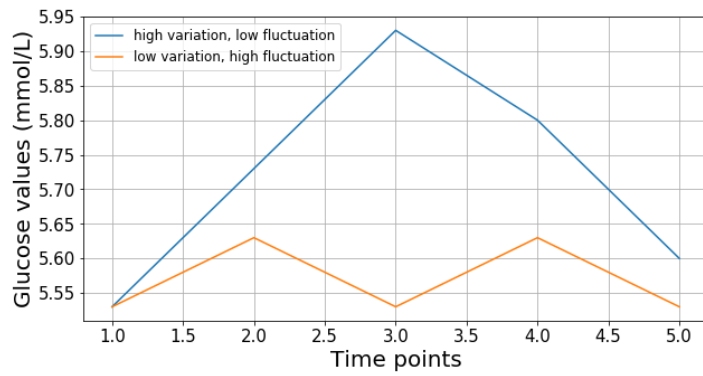
This project was executed at Technical University of Denmark. SLG is funded by Fondation Idella.

*Conflict of Interest:* none declared.

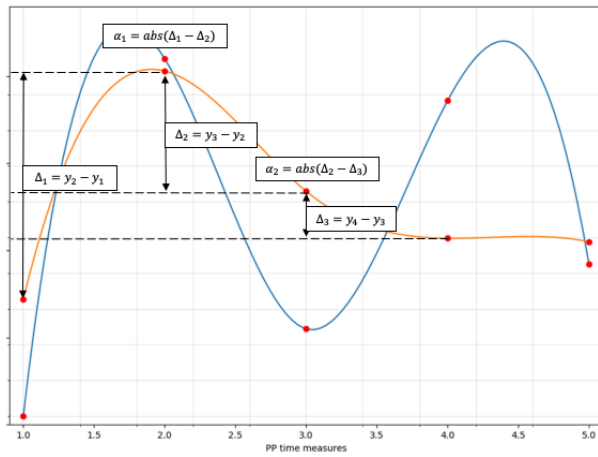


## References

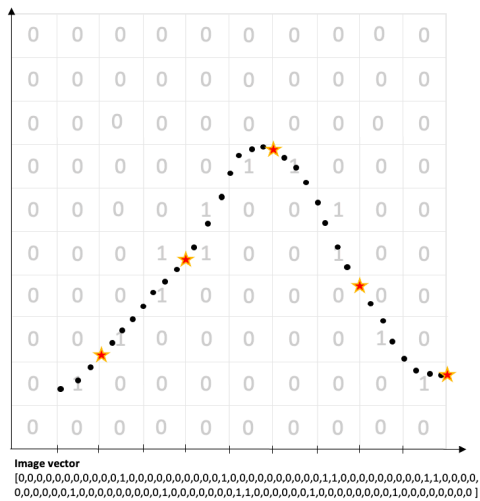
- American Diabetes Association. 2001. "Postprandial Blood Glucose." *Diabetes Care* 24(4): 775 LP – 778. <http://care.diabetesjournals.org/content/24/4/775.abstract>.
- Brezina, Vladimir, Irina V Orekhova, and Klaudiusz R Weiss. 1997. "Control of Time-Dependent Biological Processes by Temporally Patterned Input." *Proceedings of the National Academy of Sciences* 94(19): 10444 LP – 10449. <http://www.pnas.org/content/94/19/10444.abstract>.
- Freckmann, Guido et al. 2007. "Continuous Glucose Profiles in Healthy Subjects under Everyday Life Conditions and after Different Meals." *Journal of diabetes science and technology* 1(5): 695–703.
- Hajime, Maiko et al. 2018. "Twenty-Four-Hour Variations in Blood Glucose Level in Japanese Type 2 Diabetes Patients Based on Continuous Glucose Monitoring." *Journal of diabetes investigation* 9(1): 75–82.
- Hall, Heather et al. 2018. "Glucotypes Reveal New Patterns of Glucose Dysregulation." *PLoS biology* 16(7): e2005143.
- Hulman, Adam, Dorte Vistisen, et al. 2018. "Glucose Patterns during an Oral Glucose Tolerance Test and Associations with Future Diabetes, Cardiovascular Disease and All-Cause Mortality Rate." *Diabetologia* 61(1): 101–7.
- Hulman, Adam, Daniel R Witte, et al. 2018. "Pathophysiological Characteristics Underlying Different Glucose Response Curves: A Latent Class Trajectory Analysis From the Prospective EGIR-RISC Study." *Diabetes care* 41(8): 1740–48.
- Jackson, Kim G et al. 2010. "Introduction to the DISRUPT Postprandial Database: Subjects, Studies and Methodologies." *Genes & nutrition* 5(1): 39–48.
- Kavakiotis, Ioannis et al. 2017. "Machine Learning and Data Mining Methods in Diabetes Research." *Computational and structural biotechnology journal* 15: 104–16.
- Nielsen, Rikke Linnemann et al. 2020. "Data Integration for Prediction of Weight Loss in Randomized Controlled Dietary Trials." *Scientific Reports* 10(1): 20103. <http://www.nature.com/articles/s41598-020-76097-z>.
- Schüssler-Fiorenza Rose, Sophia Miryam et al. 2019. "A Longitudinal Big Data Approach for Precision Health." *Nature Medicine* 25(5): 792–804. <https://doi.org/10.1038/s41591-019-0414-6>.
- Søndertoft, Nadja B. et al. 2020. "The Intestinal Microbiome Is a Co-Determinant of the Postprandial Plasma Glucose Response" ed. Erwin G Zoetendal. *PLOS ONE* 15(9): e0238648. <https://dx.plos.org/10.1371/journal.pone.0238648> (October 29, 2020).
- Zeevi, David et al. 2015. "Personalized Nutrition by Prediction of Glycemic Responses." *Cell* 163(5): 1079–94.



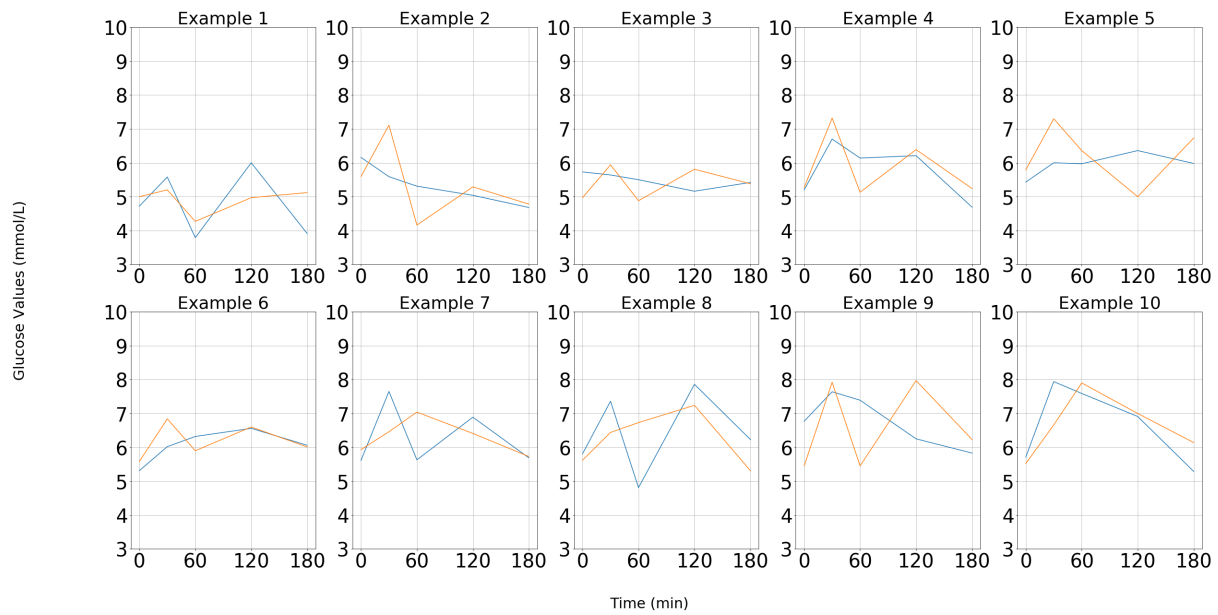
**Figure 1:** Example illustration of variation versus fluctuation.



**Figure 2:** Example of fluctuation calculation (refers to Equation 2).

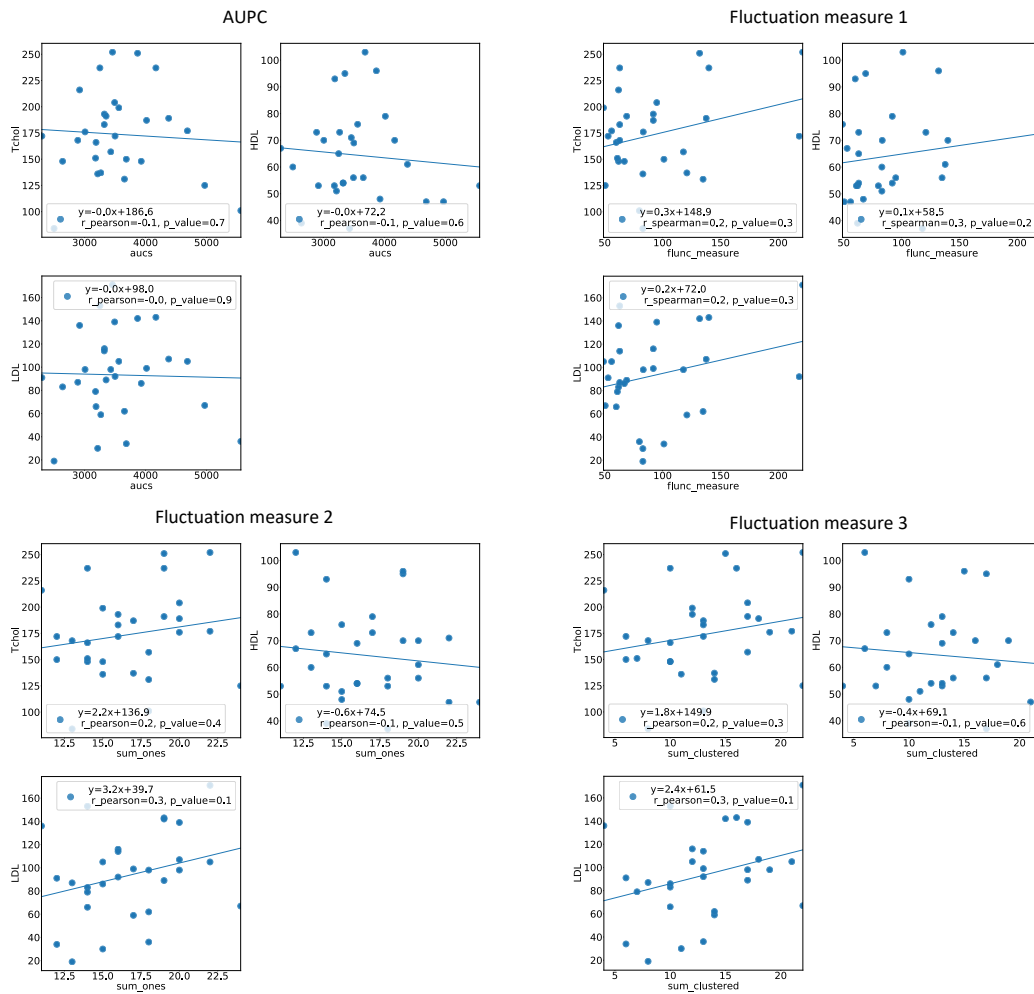


**Figure 3:** Grid analysis approach used in fluctuation measure 2 and 3. In this example, sum of 1's = 12 and sum of clustered 1's = 6.



	AUC	Fluc measure 1	Var_measure	Fluc measure 2	Fluc measure 3
Example 1	29	[7.7,2.69]	[0.99,0.28]	[31,26]	[31,25]
Example 2	31	[0.48,8.81]	[0.21,0.87]	[19,25]	[16,22]
Example 3	32	[0.38,4.24]	[0.12,0.48]	[21,34]	[20,34]
Example 4	35	[3.45,8.26]	[0.52,0.95]	[25,32]	[22,32]
Example 5	36	[1.21,4.25]	[0.2,0.79]	[25,30]	[24,30]
Example 6	37	[0.96,4.13]	[0.25,0.5]	[22,31]	[20,31]
Example 7	38	[7.94,0.97]	[0.93,0.35]	[32,24]	[32,23]
Example 8	39	[10.52,1.79]	[1.25,0.51]	[30,22]	[30,19]
Example 9	40	[1.8,10.8]	[0.38,1.31]	[22,35]	[21,35]
Example 10	41	[3.06,1.81]	[0.7,0.59]	[27,24]	[26,23]

**Figure 4:** Ten examples showing pairs of samples with the same AUC but distinct curves (top) and different fluctuation measures (bottom). In the table, for the fluctuation and variation measures, values are listed in square brackets for the blue curve and the orange curve, respectively.



**Figure 5:** Correlations between glucose fluctuation measures and AUC with total cholesterol, HDL cholesterol, and LDL cholesterol. AUC = area under the postprandial curve.

## Chapter 9

# Remote external stay: Dasatinib resistance in T-ALL

### Dasatinib resistance in T-cell acute lymphoblastic leukemia

**BRIEF INTRODUCTION** ALL is a cancer that affects white blood cells. It is defined as a malignant alteration and proliferation of lymphoid progenitor cells in the bone marrow, blood and extramedullary sites, such as liver and spleen. Although rare, ALL is the most common leukaemia in children (80% of cases) [158].

ALL develops relatively quick; thus, treatment begins a few days after diagnosis, and it consists of typically three main stages. Stage 1 is remission induction, and the aim is to destroy the leukaemia cells in the bone marrow and blood. Stage 2 is consolidation and the aim is to eliminate any remaining leukaemia cells to avoid relapse. Stage 3 aims to stop leukaemia from coming back, and usually, lower dosage of drugs are used than in the two first stages.

Different treatment drugs are used depending on the lineage affected (B- or T-cell). In B-ALL, drugs such as imatinib and dasatinib are very efficient if fusions involving B-cell receptor (BCR) and ABL class kinases are found, such as Philadelphia chromosome-positive (BCR-ABL1 fusion gene). These drugs are strong ABL inhibitors. T-ALL, rarer than B-ALL (12-20% of cases), is associated with a more aggressive haematological malignancy and treatment options are more limited. T-ALL is characterized by chromosomal rearrangements and enhancer mutations involving transcription factor genes such as TAL1, TAL2, HOXA, TXL1, TXL3, LMO1, and LMO2 [147].

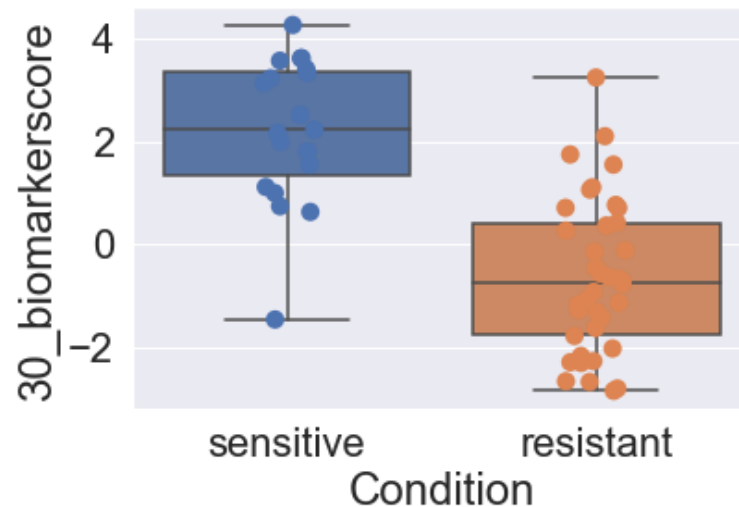
**PREVIOUS WORK[147]** One recent study at St. Jude Children's Research Hospital has profiled leukaemia drug responses ex-vivo in ALL. Surprisingly, they found that 44.4% of childhood T-ALL were sensitive to dasatinib, even though they did not have the BCR-ABL1 fusion gene. Leukaemia blasts were obtained from either bone marrow or peripheral blood collected during remission and used as germline samples as described in [147].

In B-ALL, they found that both children and adults, who presented BCR-ABL1 fusion gene or fusions involving ABL class genes, consistently exhibited high sensitivity to dasatinib. In T-ALL, some cases were classified as dasatinib-sensitive, and surprisingly, none of the dasatinib-sensitive T-ALLs harboured ABL class fusion genes.

Testing three other ABL inhibitors in a subset of these primary T-ALL samples, they observed that dasatinib-sensitive cases were universally resistant to ABL-specific inhibitors imatinib and nilotinib, but responded to ponatinib, which shares non-ABL targets with dasatinib. These results firmly pointed to an ABL-independent mechanism driving dasatinib sensitivity in a significant proportion of T-ALL.

Network-based Bayesian Inference of Drivers (NetBID)[159], a data-driven system biology approach, was used to identify drivers, either transcription factors or signalling factors, from RNA-seq-derived expression profiles. NetBID algorithm measures these drivers activity based on the expression level of its downstream targets and infers genes activity.

A biomarker of 30 drug-sensitivity driver genes (further referred to as "30 biomarker") were identified. These 30 drivers were obtained by first filtering the top 461 driver genes identified by NetBID analysis against preTCR pathway genes and dasatinib targets. Summing the weighed NetBID-inferred activity of these 30 genes, a dasatinib sensitivity score for each T-ALL case was estimated, as represented in Figure 9.1 [147].



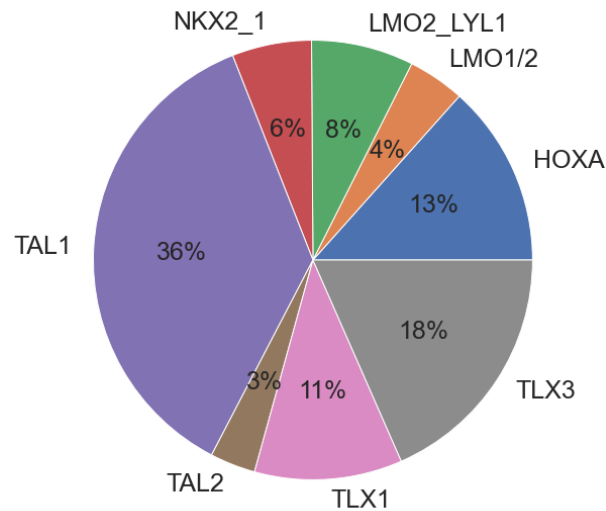
**Figure 9.1** | Biomarker score estimated from NetBID-inferred 30 genes activity for 57 T-ALL sensitive and resistant cases.

**EXTERNAL STAY PROJECT** In my project, I have tried to further optimize the 30 biomarker by building a model to predict dasatinib sensitivity in T-ALL with fewer driver genes and without compromising accuracy.

The main cohort consisted of 57 T-ALL cases (19 sensitive and 38 resistant). Additionally, a dataset of 239 T-ALL cases was available with information on molecular subtypes (Figure

9.2) rather than dasatinib sensitivity. This cohort is part of the previously published TARGET T-ALL cohort [160].

The network analysis on NetBID had been previously done, and I had direct access to NetBID inferred genes activity.



**Figure 9.2** | T-ALL subtypes in the 239 childhood leukemia cases from the TARGET T-ALL cohort

Pre-feature selection As in most clinical data scenarios, the number of patients was much lower than the number of genes (57 patients and 7689 genes). The pre-feature selection was performed by checking the correlation between all pairs of genes. If highly correlated (absolute Pearson correlation coefficient  $> 0.90$ ), the gene higher correlated with the phenotype was kept. Additionally, genes weakly associated with the phenotype were removed from the analysis (absolute point biserial correlation coefficient  $< 0.50$ ). Out of 7689 genes, 230 genes remained included in the logistic regression model.

Model development A logistic regression model using with L-BFGS optimization algorithm was developed on SciKit-learn in Python (v3.7.6). For logistic regression optimization, an inverse of regularization strength search space between  $1 \times 10^{-4}$  and  $1 \times 10^4$  was used in the inner-fold. Default values were used for the other hyperparameters, and can be assessed at SciKit-learn library v0.23.2 – LogisticRegression.

A 2 outer, 2 inner fold nested cross-validation was used on a total of 57 samples and the allocation of samples in the training or testing set was random and repeated five times.

A forward feature selection was performed with the 230 genes, adding one gene at a time in the model. ROC-AUC computed on the testing set was used as the primary performance measure. The mean ROC-AUC was then obtained from the two replication test sets. The performance kept increasing until seven genes were added. These seven genes were: *TLR9*,

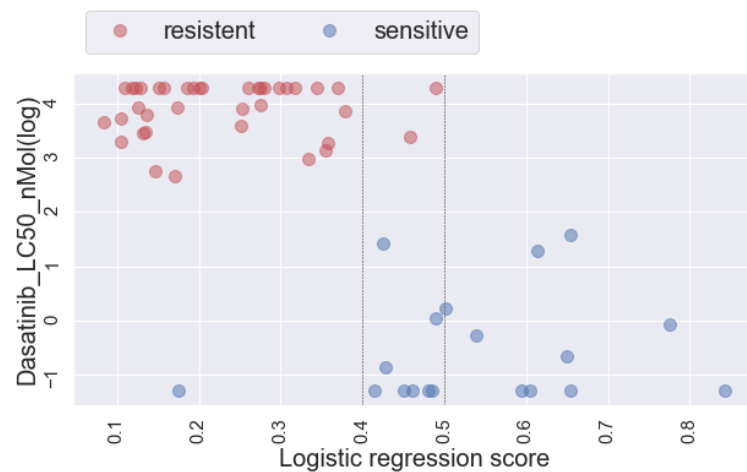
*TRAF3*, *TERT*, *ANP32B*, *IRF9*, *NFIB*, and *NAP1L1*.

For the model with seven genes, a ROC-AUC of 0.954 (std.dev. 0.006) was obtained. A logistic regression prediction score between 0 and 1 was given for each sample, where a score closer to 1 means a higher probability of being sensitive to dasatinib. Further performance metrics using a prediction cutoff of 0.50 (default) and 0.40 for classification can be accessed in Table 9.1. A higher number of true positives was obtained when a cutoff of 0.40 was used instead of the default 0.50 (Figure 9.3).

**Table 9.1** | Performance measures obtained for the model with seven genes.

	Sensitivity (0.50/0.40 cutoff)	Specificity (0.50/0.40 cutoff)	PPV (0.50/0.40 cutoff)	NPV (0.50/0.40 cutoff)	ROC-AUC test	ROC-AUC train
Mean	[0.537; 0.842]	[0.984; 0.942]	[0.957; 0.882]	[0.812; 0.924]	0.9543	0.9659
Std.dev.	[0.131; 0.074]	[0.021; 0.026]	[0.057; 0.042]	[0.04; 0.033]	0.0058	0.010

std.dev.=standard deviation; PPV=positive predictive value; NPV=negative predictive value; ROC-AUC=area under the receiver operating characteristic curve.



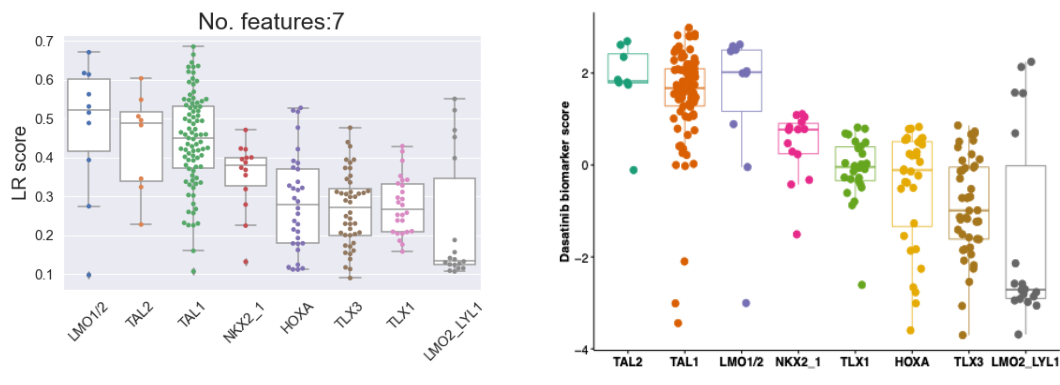
**Figure 9.3** | Comparison between dasatinib LC50 (log scale) in T-ALL and logistic regression score. Red: dasatinib resistant; blue: dasatinib sensitive.

Model evaluation The logistic regression model was deployed on the TARGET T-ALL cohort, and a variation in the predicted dasatinib sensitivity was observed between the different molecular subtypes. These results were not so different from those obtained with the 30 biomarker (Figure 9.4). As in the 30 biomarker, a higher likelihood of responding to dasatinib in T-ALL with overexpression of *TAL1*, *TAL2* or *LMO1/2* genes was observed. Additionally, the total number of dasatinib responders was similar in both models. Using a cutoff of 0.40 for classification in the logistic regression model, we ended up with a total of 37% predicted dasatinib responders (Table 9.2), while this number was 44.4% using the 30 biomarker as referred in the paper [147].



**Table 9.2** | Number (and percentage) of predicted resistant and sensitive cases in the total cohort and in the different molecular subtypes groups using 0.40 and 0.50 cutoffs for classification in the logistic regression model.

Prediction cutoff = 0.50	Prediction cutoff = 0.40
<b>TOTAL TARGET No. resistant = 195 ( 81.59 %)</b>	<b>TOTAL TARGET No. resistant = 150 ( 62.76 %)</b>
<b>TOTAL TARGET No. sensitive = 44 ( 18.41 %)</b>	<b>TOTAL TARGET No. sensitive = 89 ( 37.24 %)</b>
TLX1 : No. resistant = 26 ( 100.0 %)	TLX1 : No. resistant = 24 ( 92.31 %)
TLX1 : No. sensitive = 0 ( 0.0 %)	TLX1 : No. sensitive = 2 ( 7.69 %)
TAL2 : No. resistant = 5 ( 62.5 %)	TAL2 : No. resistant = 3 ( 37.5 %)
TAL2 : No. sensitive = 3 ( 37.5 %)	TAL2 : No. sensitive = 5 ( 62.5 %)
HOXA : No. resistant = 29 ( 90.62 %)	HOXA : No. resistant = 27 ( 84.38 %)
HOXA : No. sensitive = 3 ( 9.38 %)	HOXA : No. sensitive = 5 ( 15.62 %)
TAL1 : No. resistant = 57 ( 65.52 %)	TAL1 : No. resistant = 28 ( 32.18 %)
TAL1 : No. sensitive = 30 ( 34.48 %)	TAL1 : No. sensitive = 59 ( 67.82 %)
NKX2_1 : No. resistant = 14 ( 100.0 %)	NKX2_1 : No. resistant = 10 ( 71.43 %)
NKX2_1 : No. sensitive = 0 ( 0.0 %)	NKX2_1 : No. sensitive = 4 ( 28.57 %)
TLX3 : No. resistant = 44 ( 100 %)	TLX3 : No. resistant = 41 ( 93.18 %)
TLX3 : No. sensitive = 0 ( 0.0 %)	TLX3 : No. sensitive = 3 ( 6.82 %)
LMO2_LYL1 : No. resistant = 16 ( 88.89 %)	LMO2_LYL1 : No. resistant = 14 ( 77.78 %)
LMO2_LYL1 : No. sensitive = 2 ( 11.11 %)	LMO2_LYL1 : No. sensitive = 4 ( 22.22 %)
LMO1/2 : No. resistant = 4 ( 40.0 %)	LMO1/2 : No. resistant = 3 ( 30.0 %)
LMO1/2 : No. sensitive = 6 ( 60.0 %)	LMO1/2 : No. sensitive = 7 ( 70.0 %)



(a) Logistic regression (LR) score in the TARGET T-AL cohort (n=239).

(b) Dasatinib 30 biomarker score in the TARGET T-AL cohort (n=239).

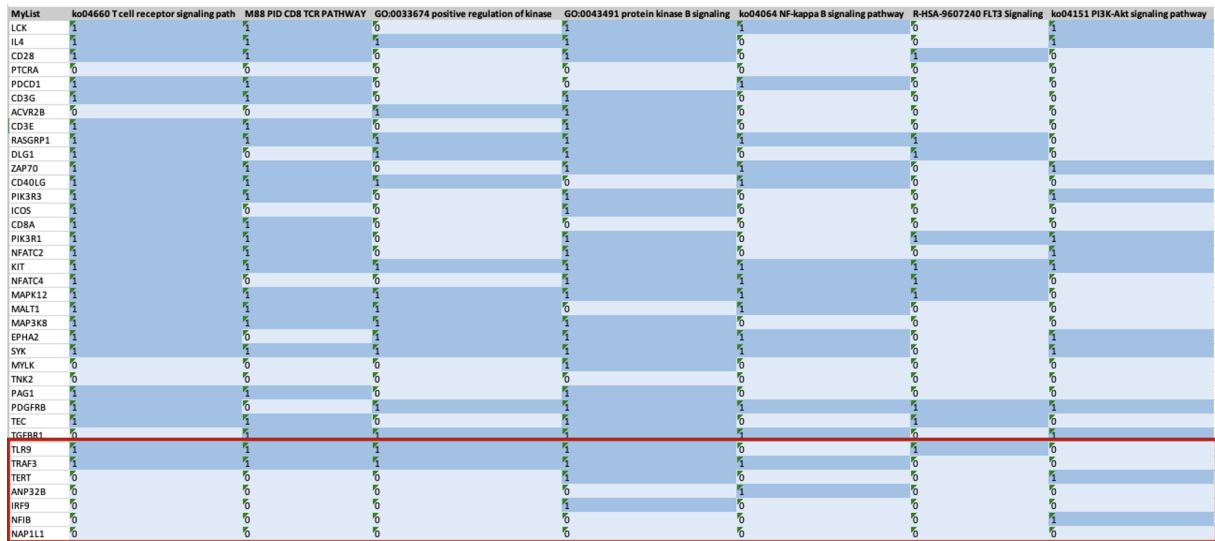
**Figure 9.4** | Comparison between dasatinib sensitivity prediction in the logistic regression (LR) model versus 30 biomarker score.

Biology interpretation The model selected a total of seven genes. Gene ontology analysis was done in Metascape [161] to understand the possible biological functions of these genes. This was also done for the 30 genes included in the 30 biomarker to understand shared biological pathways. Seven shared relevant pathways were found. The first two genes selected by the model, *TLR9* and *TRAF3*, were enriched in five of them, such as the T-cell receptor signalling pathway, positive regulation of kinases or protein kinase B signalling (Figure 9.5).

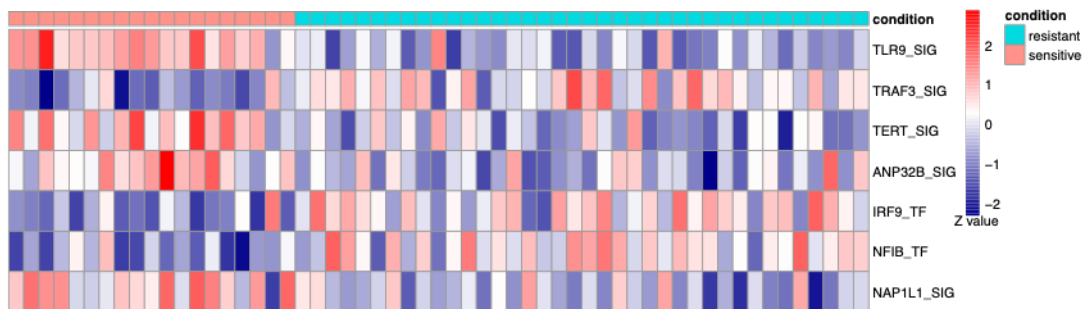
The involvement of TLR9 in dasatinib response was further supported by the literature. Toll-like receptors (TLRs) are co-stimulatory receptors involved in T-cell and cytokine production, complementing TCR-induced signals [162]. The engagement of different TLRs, including TLR9, on helper CD4 T-cells leads to an increase of IL-2 and consequently increases proliferation [163]. Dasatinib was seen to inhibit the secretion of tumour necrosis factor (TNF)- $\alpha$  after TLR stimulation. After dasatinib treatment, TNF- $\alpha$  blood samples levels were shown to decrease significantly in a multiple sclerosis mouse model [164]. Thus, if it happens the same in ALL, this could explain the higher TLR9 activity in dasatinib responders (Figure 9.6). Higher activity of TLR9 will lead to higher production of TNF- $\alpha$ , and TNF- $\alpha$  is a dasatinib target.

For the third gene selected by the model, *TERT*, even though it was only present in two out of seven metascape pathways, literature that may support the involvement of this gene in dasatinib response was found. Telomerase reverse transcriptase (TERT) is responsible for the transcription and translation of the enzyme telomerase. Telomerase maintains the telomeres, which are composed of repeated segments of DNA found at the end of chromosomes. Numerous drugs, including dasatinib, have been identified with off-target effects on telomerase activity. These include drugs that act via downregulation of

hTERT gene transcription [165], which could help explaining the higher activity of TERT in sensitive samples (Figure 9.6).



**Figure 9.5** | Metascape [161] gene ontology analysis for the 30 genes in the 30 biomarker and the seven genes included in my model (red box).



**Figure 9.6** | Heatmap of NetBID-inferred activity of seven driver genes included in the logistic regression model in the 57 patients cohort.

Before claiming clinical applicability in predicting dasatinib response, the seven genes found in this study need to be further validated in a larger and external cohort. Furthermore, it would be beneficial to develop functional studies to identify the causal genes and biological mechanisms underlying. This was not explored here. It was seen, for example, that TLR9 activity was highly correlated (Pearson's  $r=0.90$ ) with other 22 genes, including LCK, the main driver identified in the paper [147].

## Part IV

## Epilogue

Many therapies applied in western medicine are developed with a focus on the "average" patient, or with statistically demonstrated safety and efficacy across a wide population. At the individual level, systematic guidance for choice or application of therapy regimes is rather limited. An individual's response to therapy or possible short and long-term side-effects varies greatly and is influenced by several factors, including genetic predisposition and exposures later in life. We have seen increasing progress with high efficacy rates in treatment strategies against certain cancers. Thus, we have now the luxury and data-driven ability to begin a larger focus on the side effects these therapies may bring and how they can affect these patients' long-term quality of life. Are we now sitting on enough data that allows us to focus on long-term quality of life, and to what extent would we dare to adjust treatment to improve quality of life?

Since the completion of the Human Genome Project in 2003, many genetic variants have been found to be associated with multiple phenotypes. This allowed the development of prognostic markers, hopefully leading to more tailored interventions.

At the moment, clinical genomics analysis is only done on specific types of cancers or rare diseases. Soon, this may become a routine procedure in clinical care and may start to be integrated with electronic health records, as we see its high potential in multiple common diseases for guiding prevention, diagnosis and treatment. Routine genomic sequencing will also result in more data that can be used for research.

In this PhD thesis, I have summarised different types of medicine, starting from our roots (traditional medicine) until now. I have described ways of analysing patient genomics and clinical data. For clinical data, I have focused on single-time point measurements and longitudinal data, which are touched upon in the application note, chapter 8, where I have explored different measures that can capture fluctuation information from postprandial temporal curves. For genomics data, I have used GWASs and NGS analysis to identify genetic variants associated with several phenotypes.

Finally, I have explored different stratification-based approaches such as Ayurveda-based deep phenotyping, PRSs, and ML models and discussed how these could be applied in a clinical setting. In paper I, chapter 4, I have used an Ayurveda-based deep phenotyping to stratify individuals in more homogeneous subgroups and facilitate the identification of genetic associations of large effect sizes in rheumatoid arthritis. In paper II, chapter 5, I have investigated two PRSs developed and validated in adult cancers to subgrouping patients and disease prognosis in two childhood cancer cohorts. This was done to understand disease aetiology trajectory in childhood cancer better. In papers III, chapter 6, and paper IV, chapter 7, I have developed two prediction models for hearing loss and nephrotoxicity, respectively, after cisplatin-based chemotherapy in testicular cancer patients. I have integrated both clinical and genomics data in the last two papers. Putting patients

in different risk groups for developing these late side effects is useful for identifying toxicity risks, which will influence treatment intensity or monitorisation for selected patients. Some additional work was done for paper IV, chapter 7, where I further explored *NAT1* and *NAT2* involvement in nephrotoxicity development.

Additionally, during my external stay in St.Jude, chapter 9, I have developed a prediction model of dasatinib response in T-ALL. By identifying which patients are dasatinib resistant, the long-term goal is to reduce adverse events and cut treatment costs.

All this work was done in close collaboration with clinicians, which I consider very meaningful and has been super valuable for me to understand the considerations, opportunities and challenges in a clinical setting. Through the different PhD projects, some of the challenges and opportunities with patient data analyses are pointed out.

### **Limitations**

The limitations of each project are described in more detail in each paper chapter; however, some main and common limitations across all projects, which I also believe applies in many other precision medicine research papers, are: 1) further validation is needed in external datasets; 2) functional studies would have been a great addition to the projects, to facilitate the validation and mechanistic understanding of causal variants and genes; and 3) most studies presented here, except for paper I, chapter 4, are done in European populations; thus its applicability in other populations may be limited.

### **Future directions**

New therapy regimes and guidance are much needed to change the paradigm of the current "one-size fits all" approach. If no better treatments are developed, clinicians may continue using some that may have differential benefits on individual patients just because it is the best available therapy at the moment across a wide population. This impacts individual care as well as the overall burden of healthcare costs.

There are still multiple challenges that one needs to be aware of regarding the application of data models to implement precision medicine in the clinic. These include 1) ethical regulation and how we can use and release the patient information appropriately to address privacy and security concerns; 2) how to better approach patients to be part of the studies - clear communication is essential so physicians and patients can understand the benefits and trust the process; and 3) how to make the most of the limited, but very valuable clinical datasets. Furthermore, an effort is needed to bring this to all individuals in all parts of the world; otherwise, we risk having high racial disparities in the future regarding health systems around the globe.

The ultimate goal of stratification-based models is to help and support clinicians in deciding which treatment would be better for each patient based on their clinical characteristics and biomarkers. We are now scratching the surface, but I believe that in the next decade,

with close collaboration between people from different fields, such as clinicians, bioinformaticians, data scientists, biologists and statisticians, we will see the continued growth and application of these models into clinical care. This will lead to improvements in how one defines and classifies diseases based on a more precise understanding of what is occurring at the molecular and cellular level, as well as across more holistically for an individual.

## Part V

### Appendix: Research efforts in India, and thoughts around Ayurveda



Ayurveda is one of the oldest disciplines in the world, with quite a lot of information available but not compiled in scientific journals.

One of the challenges I had while working on this area was the scepticism around it. I think that either scepticism or fanaticism, in either direction, is dangerous for the development of science, especially if we are to be encouraged out of the box for the development of patient stratification.

Precision medicine is promising, but success has been fewer than expected, even though there are high hopes that there will be a significant development in the next decade. One thing is certain; we need to have open-minded thinking to explore opportunities for patient stratification that will eventually lead to treatment improvements and diseases prevention. As my knowledge was limited in regards to Ayurveda medicine, and to better understand the concept and current state of the art, different visits and studies in India were organised to meet and find partners and collaborators. We had the opportunity to work with 1) professor BK Thelma, from the Department of Genetics at the University of Delhi, who had worked with Ayurveda before and published several papers on the topic; 2) Dr Bheema Bhatta, Head of Ayurveda Department at Holy Family Hospital, who practices Ayurveda for many years; and 3) Dr Uma Kumar, Head of Rheumatology at All India Institute of Medical Sciences (AIIMS) who uses conventional medicine to treat rheumatoid arthritis (RA) patients but is also very interested about Ayurveda and its patient stratification principle. Dr Uma is also helping with patient recruitment.

We have submitted a project description (briefly described below in "Aim 1") to the Institute Ethics Committee of AIIMS. This was approved in February 2020; thus, there are still ongoing efforts in India concerning projects we had in mind to complete. There were several delays due to the inability to travel to India in 2020 when COVID-19 hit.

**Aim 1 (in progress) Understanding and predicting response in first-line therapy patients with RA and impact of Ayurvedic stratification across the two treatment arms: conventional therapy and Ayurveda.**

Recruitment of RA patients had begun in early 2020 at the AIIMS from regular patient visits. The study for ayurvedic examination was approved by the Ethics board at AIIMS. The goals in this effort are 1) to understand the differential patient response to standardised therapy on RA combining Ayurveda based patient Prakriti information; and 2) to identify Prakriti specific multi-omics markers of treatment response/non-response.

a) Target group 1: 600 patients (300 responders; 300 non-responders): Response/non-response will be measured by collecting multi-omics (genomics, metagenomics, metabolomics) data before and after initiation of therapy, with an outcome measurement at 3 and 6 months.

b) Target group 2: 60 newly diagnosed and drug-naïve RA patients to identify biomarkers that might be associated with response/non-response by multi-omics approaches.

**Aim 2. Gut microbiome characterisation of 30 RA patients who followed ayurvedic treatment or methotrexate (pre and post-treatment).**

Blood samples and frozen stool samples have been collected. Negotiation on costs and logistics for microbiome characterisation is underway. As we expect high inter-individual variability in microbiome composition, this project has a lower likelihood of generating insights across the vata, pita and kapha ayurvedic subgroups. However, it will be the first such characterisation and thus would be the first stepping stone in this direction.

# Bibliography

- [1] R. Hodson. “Precision medicine”. In: *Nature* 537.7619 (Sept. 2016), S49–S49 (cit. on p. 2).
- [2] H. Fröhlich et al. “From hype to reality: data science enabling personalized medicine”. In: *BMC Medicine* 16.1 (Dec. 2018), p. 150 (cit. on pp. 2, 5).
- [3] A. Dance. “Medical histories”. In: *Nature* 537.7619 (Sept. 2016), S52–S53 (cit. on p. 2).
- [4] B. PRASHER, G. GIBSON, and M. MUKERJI. “Genomic insights into ayurvedic and western approaches to personalized medicine”. In: *Journal of Genetics* 95.1 (Mar. 2016), pp. 209–228 (cit. on p. 2).
- [5] R. Sharma and P. K. Prajapati. “Predictive, Preventive and Personalized Medicine: Leads From Ayurvedic Concept of Prakriti (Human Constitution)”. In: *Current Pharmacology Reports* 6.6 (Dec. 2020), pp. 441–450 (cit. on p. 2).
- [6] P. Tiwari et al. “Recapitulation of Ayurveda constitution types by machine learning of phenotypic traits”. In: *PLOS ONE* 12.10 (Oct. 2017). Ed. by G. Chaubey, e0185380 (cit. on p. 3).
- [7] T. P. Sethi, B. Prasher, and M. Mukerji. “Ayurgenomics: a new way of threading molecular variability for stratified medicine.” In: *ACS chemical biology* 6.9 (Sept. 2011), pp. 875–80 (cit. on p. 3).
- [8] H. Rotti et al. “Determinants of prakriti, the human constitution types of Indian traditional medicine and its correlation with contemporary science.” In: *Journal of Ayurveda and integrative medicine* 5.3 (July 2014), pp. 167–75 (cit. on p. 3).
- [9] V. K. Joshi, A. Joshi, and K. S. Dhiman. “The Ayurvedic Pharmacopoeia of India, development and perspectives”. In: *Journal of Ethnopharmacology* 197 (Feb. 2017), pp. 32–38 (cit. on p. 3).
- [10] B. Patwardhan. “Bridging Ayurveda with evidence-based scientific approaches in medicine”. In: *EPMA Journal* 5.1 (Dec. 2014), p. 19 (cit. on p. 3).

- [11] D. E. Furst et al. “Double-Blind, Randomized, Controlled, Pilot Study Comparing Classic Ayurvedic Medicine, Methotrexate, and Their Combination in Rheumatoid Arthritis”. In: *Journal of Clinical Rheumatology* 17.Suppl 1 (June 2011), pp. 185–192 (cit. on p. 3).
- [12] A. Chopra et al. “Ayurvedic medicine offers a good alternative to glucosamine and celecoxib in the treatment of symptomatic knee osteoarthritis: a randomized, double-blind, controlled equivalence drug trial”. In: *Rheumatology* 52.8 (Aug. 2013), pp. 1408–1417 (cit. on p. 4).
- [13] P. Gupta. “Pharmacogenetics, pharmacogenomics and ayurgenomics for personalized medicine: A paradigm shift”. In: *Indian Journal of Pharmaceutical Sciences* 77.2 (2015), p. 135 (cit. on p. 4).
- [14] S. Aggarwal, A. Gheware, A. Agrawal, S. Ghosh, B. Prasher, and M. Mukerji. “Combined genetic effects of EGLN1 and VWF modulate thrombotic outcome in hypoxia revealed by Ayurgenomics approach”. In: *Journal of Translational Medicine* 13.1 (Dec. 2015), p. 184 (cit. on p. 4).
- [15] P. Bhushan, J. Kalpana, and C. Arvind. “Classification of Human Population Based on HLA Gene Polymorphism and the Concept of Prakriti in Ayurveda”. In: *The Journal of Alternative and Complementary Medicine* 11.2 (Apr. 2005), pp. 349–353 (cit. on p. 4).
- [16] S. Aggarwal et al. “EGLN1 involvement in high-altitude adaptation revealed through genetic analysis of extreme constitution types defined in Ayurveda”. In: *Proceedings of the National Academy of Sciences* 107.44 (Nov. 2010), pp. 18961–18966 (cit. on p. 4).
- [17] P. Govindaraj et al. “Genome-wide analysis correlates Ayurveda Prakriti”. In: *Scientific Reports* 5.1 (Dec. 2015), p. 15786 (cit. on p. 4).
- [18] B. Chatterjee and J. Pancholi. “Prakriti-based medicine: A step towards personalized medicine.” In: *Ayu* 32.2 (Apr. 2011), pp. 141–6 (cit. on p. 4).
- [19] R. C. Juyal, S. Negi, P. Wakhode, S. Bhat, B. Bhat, and B. K. Thelma. “Potential of Ayurgenomics Approach in Complex Trait Research: Leads from a Pilot Study on Rheumatoid Arthritis”. In: *PLoS ONE* 7.9 (Sept. 2012). Ed. by G. Novelli, e45752 (cit. on p. 4).
- [20] B. Prasher et al. “Whole genome expression and biochemical correlates of extreme constitutional types defined in Ayurveda”. In: *Journal of Translational Medicine* 6.1 (2008), p. 48 (cit. on p. 4).

- [21] N. S. Chauhan et al. "Western Indian Rural Gut Microbial Diversity in Extreme Prakriti Endo-Phenotypes Reveals Signature Microbes". In: *Frontiers in Microbiology* 9 (Feb. 2018) (cit. on p. 4).
- [22] A. Shirolkar, S. Chakraborty, T. Mandal, and R. Dabur. "Plasma metabolomics reveal the correlation of metabolic pathways and Prakritis of humans". In: *Journal of Ayurveda and Integrative Medicine* 9.2 (Apr. 2018), pp. 113–122 (cit. on p. 4).
- [23] E. D. Green and M. S. Guyer. "Charting a course for genomic medicine from base pairs to bedside". In: *Nature* 470.7333 (Feb. 2011), pp. 204–213 (cit. on p. 5).
- [24] Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)* (cit. on p. 5).
- [25] L. H. Goetz and N. J. Schork. "Personalized medicine: motivation, challenges, and progress." In: *Fertility and sterility* 109.6 (2018), pp. 952–963 (cit. on p. 5).
- [26] G. S. Ginsburg and K. A. Phillips. "Precision Medicine: From Science To Value." In: *Health affairs (Project Hope)* 37.5 (2018), pp. 694–701 (cit. on p. 5).
- [27] S. Erikainen and S. Chan. "Contested futures: envisioning "Personalized," "Stratified," and "Precision" medicine." In: *New genetics and society* 38.3 (2019), pp. 308–330 (cit. on p. 5).
- [28] Y. Wang, S. Sun, Z. Zhang, and D. Shi. "Nanomaterials for Cancer Precision Medicine". In: *Advanced Materials* 30.17 (Apr. 2018), p. 1705660 (cit. on p. 6).
- [29] K. Adane, M. Gizachew, and S. Kendie. "The role of medical data in efficient patient care delivery: a review". In: *Risk Management and Healthcare Policy* Volume 12 (Apr. 2019), pp. 67–73 (cit. on p. 6).
- [30] R. L. Nielsen et al. "Data integration for prediction of weight loss in randomized controlled dietary trials". In: *Scientific Reports* 10.1 (Dec. 2020), p. 20103 (cit. on p. 6).
- [31] I. Cornelisz, P. Cuijpers, T. Donker, and C. van Klaveren. "Addressing missing data in randomized clinical trials: A causal inference perspective". In: *PLOS ONE* 15.7 (July 2020). Ed. by V. Berger, e0234349 (cit. on p. 7).
- [32] D. Mavridis and I. R. White. "Dealing with missing outcome data in meta-analysis." In: *Research synthesis methods* 11.1 (Jan. 2020), pp. 2–13 (cit. on p. 7).
- [33] S. Fielding, P. M. Fayers, A. McDonald, G. McPherson, and M. K. Campbell. "Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data". In: *Health and Quality of Life Outcomes* 6.1 (2008), p. 57 (cit. on p. 7).

- [34] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel. “When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts”. In: *BMC Medical Research Methodology* 17.1 (Dec. 2017), p. 162 (cit. on p. 7).
- [35] S. v. Buuren and K. Groothuis-Oudshoorn. “mice : Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3 (2011) (cit. on p. 8).
- [36] A. Hulman, D. R. Witte, D. Vistisen, B. Balkau, J. M. Dekker, C. Herder, M. Hatunic, T. Konrad, K. Færch, and M. Manco. “Pathophysiological Characteristics Underlying Different Glucose Response Curves: A Latent Class Trajectory Analysis From the Prospective EGIR-RISC Study”. In: *Diabetes Care* 41.8 (Aug. 2018), pp. 1740–1748 (cit. on p. 8).
- [37] A. Hulman, D. Vistisen, C. Glümer, M. Bergman, D. R. Witte, and K. Færch. “Glucose patterns during an oral glucose tolerance test and associations with future diabetes, cardiovascular disease and all-cause mortality rate”. In: *Diabetologia* 61.1 (Jan. 2018), pp. 101–107 (cit. on p. 8).
- [38] H. Hall, D. Perelman, A. Breschi, P. Limcaoco, R. Kellogg, T. McLaughlin, and M. Snyder. “Glucotypes reveal new patterns of glucose dysregulation”. In: *PLOS Biology* 16.7 (July 2018). Ed. by J. Locasale, e2005143 (cit. on p. 8).
- [39] G. Freckmann, S. Hagenlocher, A. Baumstark, N. Jendrike, R. C. Gillen, K. Rössner, and C. Haug. “Continuous Glucose Profiles in Healthy Subjects under Everyday Life Conditions and after Different Meals”. In: *Journal of Diabetes Science and Technology* 1.5 (Sept. 2007), pp. 695–703 (cit. on p. 8).
- [40] D. Zeevi et al. “Personalized Nutrition by Prediction of Glycemic Responses”. In: *Cell* 163.5 (Nov. 2015), pp. 1079–1094 (cit. on p. 8).
- [41] N. B. Søndertoft et al. “The intestinal microbiome is a co-determinant of the postprandial plasma glucose response”. In: *PLOS ONE* 15.9 (Sept. 2020). Ed. by E. G. Zoetendal, e0238648 (cit. on p. 8).
- [42] S. Tian, H. Yan, C. Neuhauser, and S. L. Slager. “An analytical workflow for accurate variant discovery in highly divergent regions”. In: *BMC Genomics* 17.1 (Dec. 2016), p. 703 (cit. on p. 8).
- [43] B. Roig, M. Rodríguez-Balada, S. Samino, E. W.-F. Lam, S. Guaita-Esteruelas, A. R. Gomes, X. Correig, J. Borràs, O. Yanes, and J. Gumà. “Metabolomics reveals novel blood plasma biomarkers associated to the BRCA1-mutated phenotype of human breast cancer”. In: *Scientific Reports* 7.1 (Dec. 2017), p. 17831 (cit. on p. 9).

- [44] R. Simon. “Genomic biomarkers in predictive medicine: an interim analysis.” In: *EMBO molecular medicine* 3.8 (Aug. 2011), pp. 429–35 (cit. on p. 9).
- [45] V. M. Lauschke and M. Ingelman-Sundberg. “Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation”. In: *npj Genomic Medicine* 5.1 (Dec. 2020), p. 9 (cit. on p. 9).
- [46] S. I. Candille et al. “Genome-Wide Association Studies of Quantitatively Measured Skin, Hair, and Eye Pigmentation in Four European Populations”. In: *PLoS ONE* 7.10 (Oct. 2012). Ed. by N. J. Timpson, e48294 (cit. on p. 10).
- [47] O. S. Meyer, M. M. B. Lunn, S. L. Garcia, A. B. Kjærbye, N. Morling, C. Børsting, and J. D. Andersen. “Association between brown eye colour in rs12913832:GG individuals and SNPs in TYR, TYRP1, and SLC24A4”. In: *PLOS ONE* 15.9 (Sept. 2020). Ed. by N. R. Parine, e0239131 (cit. on p. 10).
- [48] A. Xue et al. “Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes”. In: *Nature Communications* 9.1 (Dec. 2018), p. 2941 (cit. on p. 10).
- [49] R. Bumgarner. “Overview of DNA Microarrays: Types, Applications, and Their Future”. In: *Current Protocols in Molecular Biology*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Jan. 2013 (cit. on pp. 10, 11).
- [50] V. Trevino, F. Falciani, and H. A. Barrera-Saldaña. “DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research”. In: *Molecular Medicine* 13.9-10 (Sept. 2007), pp. 527–541 (cit. on p. 10).
- [51] S. Maouche, O. Poirier, T. Godefroy, R. Olaso, I. Gut, J.-P. Collet, G. Montalescot, and F. Cambien. “Performance comparison of two microarray platforms to assess differential gene expression in human monocyte and macrophage cells”. In: *BMC Genomics* 9.1 (2008), p. 302 (cit. on p. 10).
- [52] S. Purcell et al. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses”. In: *The American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575 (cit. on pp. 11, 15).
- [53] S. Zhao, W. Jing, D. C. Samuels, Q. Sheng, Y. Shyr, and Y. Guo. “Strategies for processing and quality control of Illumina genotyping arrays”. In: *Briefings in Bioinformatics* 19.5 (Sept. 2018), pp. 765–775 (cit. on p. 11).
- [54] T. Miyagawa et al. “Appropriate data cleaning methods for genome-wide association study”. In: *Journal of Human Genetics* 53.10 (Oct. 2008), pp. 886–893 (cit. on p. 11).
- [55] S. Turner et al. “Quality Control Procedures for Genome-Wide Association Studies”. In: *Current Protocols in Human Genetics* 68.1 (Jan. 2011) (cit. on p. 11).

- [56] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. “Data quality control in genetic case-control association studies”. In: *Nature Protocols* 5.9 (Sept. 2010), pp. 1564–1573 (cit. on pp. 11, 12).
- [57] J. N. Hellwege, J. M. Keaton, A. Giri, X. Gao, D. R. Velez Edwards, and T. L. Edwards. “Population Stratification in Genetic Association Studies.” In: *Current protocols in human genetics* 95 (2017), pp. 1–1 (cit. on p. 12).
- [58] “The International HapMap Project”. In: *Nature* 426.6968 (Dec. 2003), pp. 789–796 (cit. on p. 12).
- [59] “A map of human genome variation from population-scale sequencing”. In: *Nature* 467.7319 (Oct. 2010), pp. 1061–1073 (cit. on p. 12).
- [60] J. R. I. Coleman, J. Euesden, H. Patel, A. A. Folarin, S. Newhouse, and G. Breen. “Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray”. In: *Briefings in Functional Genomics* 15.4 (July 2016), pp. 298–304 (cit. on p. 12).
- [61] A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks. “A tutorial on conducting genome-wide association studies: Quality control and statistical analysis.” In: *International journal of methods in psychiatric research* 27.2 (2018), e1608 (cit. on pp. 12, 13).
- [62] P. McArdle, J. O’Connell, T. Pollin, M. Baumgarten, A. Shuldiner, P. Peyser, and B. Mitchell. “Accounting for Relatedness in Family Based Genetic Association Studies”. In: *Human Heredity* 64.4 (2007), pp. 234–242 (cit. on p. 12).
- [63] E. L. Stevens, G. Heckenberg, E. D. O. Roberson, J. D. Baugher, T. J. Downey, and J. Pevsner. “Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State”. In: *PLoS Genetics* 7.9 (Sept. 2011). Ed. by D. B. Allison, e1002287 (cit. on p. 12).
- [64] S. Bercovici, C. Meek, Y. Wexler, and D. Geiger. “Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping”. In: *Bioinformatics* 26.12 (June 2010), pp. i175–i182 (cit. on p. 12).
- [65] Y. Guo, J. He, S. Zhao, H. Wu, X. Zhong, Q. Sheng, D. C. Samuels, Y. Shyr, and J. Long. “Illumina human exome genotyping array clustering and quality control”. In: *Nature Protocols* 9.11 (Nov. 2014), pp. 2643–2662 (cit. on p. 12).
- [66] M. A. Simonson, A. G. Wills, M. C. Keller, and M. B. McQueen. “Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk”. In: *BMC Medical Genetics* 12.1 (Dec. 2011), p. 146 (cit. on p. 13).



- [67] J. Höglund, N. Rafati, M. Rask-Andersen, S. Enroth, T. Karlsson, W. E. Ek, and Å. Johansson. “Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers”. In: *Scientific Reports* 9.1 (Dec. 2019), p. 16844 (cit. on p. 13).
- [68] T.-H. Schwantes-An, H. Sung, J. A. Sabourin, C. M. Justice, A. J. M. Sorant, and A. F. Wilson. “Type I error rates of rare single nucleotide variants are inflated in tests of association with non-normally distributed traits using simple linear regression methods”. In: *BMC Proceedings* 10.S7 (Oct. 2016), p. 62 (cit. on p. 13).
- [69] J. Graffelman, D. Jain, and B. Weir. “A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data.” In: *Human genetics* 136.6 (2017), pp. 727–741 (cit. on p. 13).
- [70] S. Rodriguez, T. R. Gaunt, and I. N. M. Day. “Hardy-Weinberg equilibrium testing of biological ascertainment for Mendelian randomization studies.” In: *American journal of epidemiology* 169.4 (Feb. 2009), pp. 505–14 (cit. on p. 13).
- [71] J. Marchini and B. Howie. “Genotype imputation for genome-wide association studies”. In: *Nature Reviews Genetics* 11.7 (July 2010), pp. 499–511 (cit. on p. 13).
- [72] Y. Li, C. Willer, S. Sanna, and G. Abecasis. “Genotype imputation.” In: *Annual review of genomics and human genetics* 10 (2009), pp. 387–406 (cit. on p. 14).
- [73] International HapMap Consortium. “A haplotype map of the human genome.” In: *Nature* 437.7063 (Oct. 2005), pp. 1299–320 (cit. on p. 13).
- [74] O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, and J. Marchini. “Haplotype estimation using sequencing reads.” In: *American journal of human genetics* 93.4 (Oct. 2013), pp. 687–96 (cit. on p. 13).
- [75] Z. Al Bkhetan, J. Zobel, A. Kowalczyk, K. Verspoor, and B. Goudey. “Exploring effective approaches for haplotype block phasing”. In: *BMC Bioinformatics* 20.1 (Dec. 2019), p. 540 (cit. on p. 14).
- [76] O. Delaneau, J.-F. Zagury, and J. Marchini. “Improved whole-chromosome phasing for disease and population genetic studies”. In: *Nature Methods* 10.1 (Jan. 2013), pp. 5–6 (cit. on p. 14).
- [77] O. Delaneau and J. Marchini. “Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel”. In: *Nature Communications* 5.1 (Sept. 2014), p. 3934 (cit. on p. 14).
- [78] E. M. van Leeuwen et al. “Population-specific genotype imputations using minimac or IMPUTE2”. In: *Nature Protocols* 10.9 (Sept. 2015), pp. 1285–1296 (cit. on p. 14).

- [79] B. Howie, J. Marchini, and M. Stephens. “Genotype imputation with thousands of genomes.” In: *G3 (Bethesda, Md.)* 1.6 (Nov. 2011), pp. 457–70 (cit. on p. 14).
- [80] H.-F. Zheng, J.-J. Rong, M. Liu, F. Han, X.-W. Zhang, J. B. Richards, and L. Wang. “Performance of genotype imputation for low frequency and rare variants from the 1000 genomes.” In: *PloS one* 10.1 (2015), e0116487 (cit. on p. 14).
- [81] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn. “Genome-wide association studies for complex traits: consensus, uncertainty and challenges”. In: *Nature Reviews Genetics* 9.5 (May 2008), pp. 356–369 (cit. on p. 15).
- [82] K. Schwarze, J. Buchanan, J. C. Taylor, and S. Wordsworth. “Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature”. In: *Genetics in Medicine* 20.10 (Oct. 2018), pp. 1122–1130 (cit. on p. 15).
- [83] M. L. Metzker. “Sequencing technologies — the next generation”. In: *Nature Reviews Genetics* 11.1 (Jan. 2010), pp. 31–46 (cit. on p. 15).
- [84] Y.-H. Rogers and J. C. Venter. “Massively parallel sequencing”. In: *Nature* 437.7057 (Sept. 2005), pp. 326–327 (cit. on p. 15).
- [85] A. Gomes and B. Korf. “Genetic Testing Techniques”. In: *Pediatric Cancer Genetics*. Elsevier, 2018, pp. 47–64 (cit. on p. 16).
- [86] E. M. Bunnik and K. G. Le Roch. “An Introduction to Functional Genomics and Systems Biology”. In: *Advances in Wound Care* 2.9 (Nov. 2013), pp. 490–498 (cit. on p. 16).
- [87] S. Ambardar, R. Gupta, D. Trakroo, R. Lal, and J. Vakhlu. “High Throughput Sequencing: An Overview of Sequencing Chemistry”. In: *Indian Journal of Microbiology* 56.4 (Dec. 2016), pp. 394–404 (cit. on p. 17).
- [88] B. E. Slatko, A. F. Gardner, and F. M. Ausubel. “Overview of Next-Generation Sequencing Technologies”. In: *Current Protocols in Molecular Biology* 122.1 (Apr. 2018) (cit. on p. 17).
- [89] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.” In: *Nucleic acids research* 38.6 (Apr. 2010), pp. 1767–71 (cit. on p. 17).
- [90] *Best Practices Workflows* (cit. on p. 17).
- [91] *GATK workflows* (cit. on p. 17).

- [92] U. H. Trivedi, T. CÃ©zard, S. Bridgett, A. Montazam, J. Nichols, M. Blaxter, and K. Gharbi. “Quality control of next-generation sequencing data without a reference”. In: *Frontiers in Genetics* 5 (May 2014) (cit. on p. 17).
- [93] S. Lindgreen. “AdapterRemoval: Easy Cleaning of Next Generation Sequencing Reads”. In: *BMC Research Notes* 5.1 (2012), p. 337 (cit. on p. 18).
- [94] H. Li. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”. In: (Mar. 2013) (cit. on p. 18).
- [95] M. T. W. Ebbert, M. E. Wadsworth, L. A. Staley, K. L. Hoyt, B. Pickett, J. Miller, J. Duce, J. S. K. Kauwe, and P. G. Ridge. “Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches”. In: *BMC Bioinformatics* 17.S7 (July 2016), p. 239 (cit. on p. 18).
- [96] N. Homer and S. F. Nelson. “Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA”. In: *Genome Biology* 11.10 (Oct. 2010), R99 (cit. on p. 18).
- [97] K. I. Kendig et al. “Sentieon DNaseq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy”. In: *Frontiers in Genetics* 10 (Aug. 2019) (cit. on p. 19).
- [98] C. M. Lewis and E. Vassos. “Polygenic risk scores: from research tools to clinical instruments”. In: *Genome Medicine* 12.1 (Dec. 2020), p. 44 (cit. on pp. 19, 21).
- [99] S. W. Choi, T. S.-H. Mak, and P. F. O’Reilly. “Tutorial: a guide to performing polygenic risk score analyses”. In: *Nature Protocols* 15.9 (Sept. 2020), pp. 2759–2772 (cit. on pp. 19–21).
- [100] J. A. Durlak. “How to Select, Calculate, and Interpret Effect Sizes”. In: *Journal of Pediatric Psychology* 34.9 (Oct. 2009), pp. 917–928 (cit. on p. 20).
- [101] Y. Ruan, S. W. Choi, and P. O’Reilly. “INVESTIGATING SHRINKAGE METHODS TO IMPROVE ACCURACY OF GWAS AND PRS EFFECT SIZE ESTIMATES”. In: *European Neuropsychopharmacology* 29 (2019), S896–S897 (cit. on p. 20).
- [102] H.-C. So and P. C. Sham. “Improving polygenic risk prediction from summary statistics by an empirical Bayes approach”. In: *Scientific Reports* 7.1 (Mar. 2017), p. 41262 (cit. on p. 20).
- [103] F. Privé, B. J. Vilhjálmsson, H. Aschard, and M. G. Blum. “Making the Most of Clumping and Thresholding for Polygenic Scores”. In: *The American Journal of Human Genetics* 105.6 (Dec. 2019), pp. 1213–1221 (cit. on p. 20).

- [104] C. M. Lewis and E. Vassos. “Prospects for using risk scores in polygenic medicine.” In: *Genome medicine* 9.1 (Nov. 2017), p. 96 (cit. on p. 20).
- [105] N. Mavaddat et al. “Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes”. In: *The American Journal of Human Genetics* 104.1 (Jan. 2019), pp. 21–34 (cit. on p. 20).
- [106] J. R. Huyghe et al. “Discovery of common and rare genetic risk variants for colorectal cancer”. In: *Nature Genetics* 51.1 (Jan. 2019), pp. 76–87 (cit. on p. 20).
- [107] J. Euesden, C. M. Lewis, and P. F. O’Reilly. “PRSice: Polygenic Risk Score software”. In: *Bioinformatics* 31.9 (May 2015), pp. 1466–1468 (cit. on p. 20).
- [108] A. Torkamani, N. E. Wineinger, and E. J. Topol. “The personal and clinical utility of polygenic risk scores”. In: *Nature Reviews Genetics* 19.9 (Sept. 2018), pp. 581–590 (cit. on p. 21).
- [109] M. S. Williams et al. “Genomic Information for Clinicians in the Electronic Health Record: Lessons Learned From the Clinical Genome Resource Project and the Electronic Medical Records and Genomics Network”. In: *Frontiers in Genetics* 10 (Oct. 2019) (cit. on pp. 21, 22).
- [110] A. C. F. Lewis and R. C. Green. “Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues”. In: *Genome Medicine* 13.1 (Dec. 2021), p. 14 (cit. on p. 21).
- [111] N. J. Wald and R. Old. “The illusion of polygenic disease risk prediction”. In: *Genetics in Medicine* 21.8 (Aug. 2019), pp. 1705–1707 (cit. on p. 21).
- [112] M. Warren. “The approach to predictive medicine that is taking genomics research by storm”. In: *Nature* 562.7726 (Oct. 2018), pp. 181–183 (cit. on p. 21).
- [113] A. V. Khera et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”. In: *Nature Genetics* 50.9 (Sept. 2018), pp. 1219–1224 (cit. on p. 21).
- [114] G. Sirugo, S. M. Williams, and S. A. Tishkoff. “The Missing Diversity in Human Genetic Studies”. In: *Cell* 177.1 (Mar. 2019), pp. 26–31 (cit. on p. 22).
- [115] *Artificial Intelligence and Machine Learning in Software as a Medical Device US food and drug administration*. 2019 (cit. on p. 22).
- [116] T. S. Toh, F. Dondelinger, and D. Wang. “Looking beyond the hype: Applied AI and machine learning in translational medicine”. In: *EBioMedicine* 47 (Sept. 2019), pp. 607–615 (cit. on p. 22).
- [117] C. Xu and S. A. Jackson. “Machine learning and complex biological data”. In: *Genome Biology* 20.1 (Dec. 2019), p. 76 (cit. on p. 22).

- [118] B. Mieth et al. “Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies”. In: *Scientific Reports* 6.1 (Dec. 2016), p. 36671 (cit. on p. 22).
- [119] S. Uppu, A. Krishna, and R. P. Gopalan. “A Deep Learning Approach to Detect SNP Interactions”. In: *Journal of Software* 11.10 (Oct. 2016), pp. 965–975 (cit. on p. 22).
- [120] S. Lee et al. “Machine learning on genome-wide association studies to predict the risk of radiation-associated contralateral breast cancer in the WECARE Study”. In: *PLOS ONE* 15.2 (Feb. 2020). Ed. by Y. Zheng, e0226157 (cit. on p. 22).
- [121] M. Arabnejad, C. G. Montgomery, P. M. Gaffney, and B. A. McKinney. “Nearest-Neighbor Projected Distance Regression for Epistasis Detection in GWAS With Population Structure Correction”. In: *Frontiers in Genetics* 11 (July 2020) (cit. on p. 22).
- [122] R. S. Olson, W. L. Cava, Z. Mustahsan, A. Varik, and J. H. Moore. “Data-driven advice for applying machine learning to bioinformatics problems.” In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 23 (2018), pp. 192–203 (cit. on p. 22).
- [123] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman. “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities”. In: *Information Fusion* 50 (Oct. 2019), pp. 71–91 (cit. on p. 23).
- [124] N. Altman and M. Krzywinski. “The curse(s) of dimensionality”. In: *Nature Methods* 15.6 (June 2018), pp. 399–400 (cit. on p. 23).
- [125] K. Kira and L. A. Rendell. “A Practical Approach to Feature Selection”. In: *Machine Learning Proceedings 1992*. Elsevier, 1992, pp. 249–256 (cit. on p. 23).
- [126] N. Mahendran, P. M. Durai Raj Vincent, K. Srinivasan, and C.-Y. Chang. “Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues, and Future Research Directions”. In: *Frontiers in Genetics* 11 (Dec. 2020) (cit. on p. 23).
- [127] Y. Saeys, I. Inza, and P. Larranaga. “A review of feature selection techniques in bioinformatics”. In: *Bioinformatics* 23.19 (Oct. 2007), pp. 2507–2517 (cit. on p. 23).
- [128] Y. Mao and Y. Yang. “A Wrapper Feature Subset Selection Method Based on Randomized Search and Multilayer Structure”. In: *BioMed Research International* 2019 (Nov. 2019), pp. 1–9 (cit. on p. 23).
- [129] D. Panda, R. Ray, A. A. Abdullah, and S. R. Dash. “Predictive Systems: Role of Feature Selection in Prediction of Heart Disease”. In: *Journal of Physics: Conference Series* 1372 (Nov. 2019), p. 012074 (cit. on p. 23).

- [130] F. J. W. M. Dankers, A. Traverso, L. Wee, and S. M. J. van Kuijk. “Prediction Modeling Methodology”. In: *Fundamentals of Clinical Data Science*. Cham: Springer International Publishing, 2019, pp. 101–120 (cit. on p. 23).
- [131] D. Preuveneers, I. Tsingenopoulos, and W. Joosen. “Resource Usage and Performance Trade-offs for Machine Learning Models in Smart Environments”. In: *Sensors* 20.4 (Feb. 2020), p. 1176 (cit. on p. 24).
- [132] S. Parvande, H.-W. Yeh, M. P. Paulus, and B. A. McKinney. “Consensus features nested cross-validation”. In: *Bioinformatics* 36.10 (May 2020). Ed. by A. Valencia, pp. 3093–3098 (cit. on p. 24).
- [133] X.-w. Chen and J. X. Gao. “Big Data Bioinformatics”. In: *Methods* 111 (Dec. 2016), pp. 1–2 (cit. on p. 25).
- [134] Z. Zhang. “Decision tree modeling using R”. In: *Annals of Translational Medicine* 4.15 (Aug. 2016), pp. 275–275 (cit. on p. 26).
- [135] R. Couronné, P. Probst, and A.-L. Boulesteix. “Random forest versus logistic regression: a large-scale benchmark experiment”. In: *BMC Bioinformatics* 19.1 (Dec. 2018), p. 270 (cit. on pp. 26, 27).
- [136] L. Breiman. “Random forests”. In: *Mach Learn* 45 (2001) (cit. on p. 27).
- [137] D. Denisko and M. M. Hoffman. “Classification and interaction in random forests”. In: *Proceedings of the National Academy of Sciences* 115.8 (Feb. 2018), pp. 1690–1692 (cit. on p. 27).
- [138] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: (Jan. 2012) (cit. on pp. 27, 30).
- [139] Aurlien Gron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O’Reilly Media, Inc., p. 568 (cit. on pp. 27, 29).
- [140] G. A. Bishop. “TRAF3 as a powerful and multitalented regulator of lymphocyte functions”. In: *Journal of Leukocyte Biology* 100.5 (Nov. 2016), pp. 919–926 (cit. on p. 28).
- [141] S. Shi. “Facial Keypoints Detection”. In: (Oct. 2017) (cit. on p. 28).
- [142] F. Emmert-Streib, S. Moutari, and M. Dehmer. “A comprehensive survey of error measures for evaluating binary decision making in data science”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.5 (Sept. 2019) (cit. on p. 31).
- [143] X. Jiang et al. “Shared heritability and functional enrichment across six solid cancers”. In: *Nature Communications* 10.1 (Dec. 2019), p. 431 (cit. on p. 32).

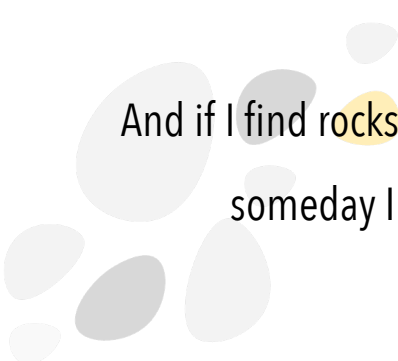
- [144] S. R. Rashkin et al. “Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts”. In: *Nature Communications* 11.1 (Dec. 2020), p. 4423 (cit. on p. 32).
- [145] L. G. Fritsche et al. “Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative”. In: *The American Journal of Human Genetics* 102.6 (June 2018), pp. 1048–1061 (cit. on p. 33).
- [146] A. T P. “Pharmacogenomics: The Right Drug to the Right Person”. In: *Journal of Clinical Medicine Research* (2009) (cit. on p. 33).
- [147] Y. Gocho et al. “Network-based systems pharmacology reveals heterogeneity in LCK and BCL2 signaling and therapeutic sensitivity of T-cell acute lymphoblastic leukemia”. In: *Nature Cancer* (Jan. 2021) (cit. on pp. 33, 123, 124, 126, 129).
- [148] G. A. Brooks, A. J. Kansagra, S. R. Rao, J. I. Weitzman, E. A. Linden, and J. O. Jacobson. “A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy”. In: *JAMA Oncology* 1.4 (July 2015), p. 441 (cit. on p. 33).
- [149] L. J. Isaksson et al. “Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy”. In: *Frontiers in Oncology* 10 (June 2020) (cit. on p. 34).
- [150] P. N. Benfey and T. Mitchell-Olds. “From Genotype to Phenotype: Systems Biology Meets Natural Variation”. In: *Science* 320.5875 (Apr. 2008), pp. 495–497 (cit. on p. 34).
- [151] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O’Sullivan. “Machine Learning SNP Based Prediction for Precision Medicine”. In: *Frontiers in Genetics* 10 (Mar. 2019) (cit. on p. 34).
- [152] N. Mena Mamani. “Machine Learning techniques and Polygenic Risk Score application to prediction genetic diseases”. In: *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal* 9.1 (Jan. 2020), pp. 5–14 (cit. on p. 34).
- [153] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. “Key challenges for delivering clinical impact with artificial intelligence”. In: *BMC Medicine* 17.1 (Dec. 2019), p. 195 (cit. on p. 35).
- [154] B. Kompa, J. Snoek, and A. L. Beam. “Second opinion needed: communicating uncertainty in medical machine learning”. In: *npj Digital Medicine* 4.1 (Dec. 2021), p. 4 (cit. on p. 35).

- [155] S. Kakarmath, A. Esteva, R. Arnaout, H. Harvey, S. Kumar, E. Muse, F. Dong, L. Wedlund, and J. Kvedar. “Best practices for authors of healthcare-related artificial intelligence manuscripts”. In: *npj Digital Medicine* 3.1 (Dec. 2020), p. 134 (cit. on p. 35).
- [156] M. J. Machiela and S. J. Chanock. “LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants: Fig. 1.” In: *Bioinformatics* 31.21 (Nov. 2015), pp. 3555–3557 (cit. on pp. 109, 112).
- [157] S. L. Garcia, J. Lauritsen, Z. Zhang, M. Bandak, M. D. Dalgaard, R. L. Nielsen, G. Daugaard, and R. Gupta. “Prediction of Nephrotoxicity Associated With Cisplatin-Based Chemotherapy in Testicular Cancer Patients”. In: *JNCI Cancer Spectrum* 4.3 (June 2020) (cit. on p. 112).
- [158] E. Vadillo, E. Dorantes-Acosta, R. Pelayo, and M. Schnoor. “T cell acute lymphoblastic leukemia (T-ALL): New insights into the cellular origins and infiltration mechanisms common and unique among hematologic malignancies”. In: *Blood Reviews* 32.1 (Jan. 2018), pp. 36–51 (cit. on p. 123).
- [159] X. Du et al. “Hippo/Mst signalling couples metabolic state and immune function of CD8 $\alpha$ + dendritic cells”. In: *Nature* 558.7708 (June 2018), pp. 141–145 (cit. on p. 124).
- [160] Y. Liu et al. “The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia”. In: *Nature Genetics* 49.8 (Aug. 2017), pp. 1211–1218 (cit. on p. 125).
- [161] Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, and S. K. Chanda. “Metascape provides a biologist-oriented resource for the analysis of systems-level datasets”. In: *Nature Communications* 10.1 (Dec. 2019), p. 1523 (cit. on pp. 128, 129).
- [162] A. H. Rahman, D. K. Taylor, and L. A. Turka. “The contribution of direct TLR signaling to T cell responses”. In: *Immunologic Research* 45.1 (Oct. 2009), pp. 25–36 (cit. on p. 128).
- [163] C. Morrison, M. R. Baer, D. P. Zandberg, A. Kimball, and E. Davila. “Effects of Toll-like receptor signals in T-cell neoplasms”. In: *Future Oncology* 7.2 (Feb. 2011), pp. 309–320 (cit. on p. 128).
- [164] C. K. Fraser, E. L. Lousberg, R. Kumar, T. P. Hughes, K. R. Diener, and J. D. Hayball. “Dasatinib inhibits the secretion of TNF- $\alpha$  following TLR stimulation in vitro and in vivo”. In: *Experimental Hematology* 37.12 (Dec. 2009), pp. 1435–1444 (cit. on p. 128).



- [165] K. Jäger and M. Walter. “Therapeutic Targeting of Telomerase”. In: *Genes* 7.7 (July 2016), p. 39 (cit. on p. 129).

**Cover image** by mcmurryjulie from Pixabay



And if I find rocks on my way I shall keep them all ...  
someday I will build up my own castle.

Fernando Pessoa

**Technical University of Denmark**  
DTU Health Tech  
Department of Health Technology

Kemitorvet, Building 204  
2800 Kgs. Lyngby

[www.healthtech.dtu.dk](http://www.healthtech.dtu.dk)

April, 2021