Development of computational approaches for whole-genome sequence variation and deep phenotyping

Matthias Haimel

Supervisors: Dr. Stefan Gräf
Prof. Nicholas Morrell

Department of Medicine
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Gonville and Caius College September 2018

Dedicated to my family.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Matthias Haimel
September 2018

Acknowledgements

First and foremost I would like to thank my supervisors Stefan Gräf and Nicholas Morrell for giving me the opportunity to carry out this project and for all their invaluable advice, support and encouragement. I would like to express my appreciation to all the patients and relatives that participated in the National Institute for Health Research (NIHR) BioResource – Rare (BR-RD) diseases study. Many thanks also to all the people involved in the NIHR BR-RD for their hard work, support and advice. I thank the NIHR for funding my PhD.

I extend my gratitude to Marta Bleda for her patience, advice and snacks during the course of the project. Special thanks to Charaka Hadinnapola for bringing the clinical perspective to the project. I am also grateful to Marta Bleda, Charaka Hadinnapola, Jennifer Martin, Carmen Treacy and all research nurses dedicated to the pulmonary arterial hypertension study for their tireless effort of collecting, validating and correcting phenotype information. I would also like to thank all the PIs from the participating NHS trusts and international collaborating centres for their dedication.

Many thanks to the high performance computing team at the university of Cambridge that provided me with an excellent service. I would like to express my gratitude to Stuart Rankin to dedicate a complete Hadoop cluster to the project and for resolving issues in the middle of the night. I am also grateful to the OpenCB development team around Nacho Medina, for their endless conversations about distributed computing technologies and their support in integrating the analysis platform into the OpenCGA project.

Nicholas Morrell, Stefan Gräf and Marta Bleda kindly read the draft of this thesis and provided me with valuable feedback. Thank you.

On a personal note, I want to thank my wife Annette Haimel for supporting me during my endeavour. Special thanks to my children Maria and Zara for the many chats, pictures, presents and smiles you brought to me, when I was in the home office.

Summary

The rare disease pulmonary arterial hypertension (PAH) results in high blood pressure in the lung caused by narrowing of lung arteries. Genes causative in PAH were discovered through family studies and very often harbour rare variants. However, the genetic cause in heritable (31%) and idiopathic (79%) PAH cases is not yet known but are speculated to be caused by rare variants. Advances in high-throughput sequencing (HTS) technologies made it possible to detect variants in 98% of the human genome. A drop in sequencing costs made it feasible to sequence 10,000 individuals including 1,250 subjects diagnosed with PAH and relatives as part of the NIHR Bioresource – Rare (BR-RD) disease study. This large cohort allows the genome-wide identification of rare variants to discover novel causative genes associated with PAH in a case-control study to advance our understanding of the underlying aetiology.

In the first part of my thesis, I establish a phenotype capture system that allows research nurses to record clinical measurements and other patient related information of PAH patients recruited to the NIHR BR-RD study. The implemented extensions provide a programmatic data transfer and an automated data release pipeline for analysis ready data.

The second part is dedicated to the discovery of novel disease genes in PAH. I focus on one well characterised PAH disease gene to establish variant filter strategies to enrich for rare disease causing variants. I apply these filter strategies to all known PAH disease genes and describe the phenotypic differences based on clinically relevant values. Genome-wide results from different filter strategies are tested for association with PAH. I describe the findings of the rare variant association tests and provide a detailed interrogation of two novel disease genes.

The last part describes the data characteristics of variant information, available non SQL (NoSQL) implementations and evaluates the suitability and scalability of distributed compute frameworks to store and analyse population scale variation data. Based on the evaluation, I implement a variant analysis platform that incrementally merges samples, annotates variants and enables the analysis of 10,000 individuals in minutes. An incremental design for variant merging and annotation has not been described before. Using the framework, I develop a quality score to reduce technical variation and other biases. The result from the rare variant association test is compared with traditional methods.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

1kG	1000 Genome Project
AC	Allele Count
AC_Hemi	Allele Count Hemizygous
AC_Het	Allele Count heterozygous
AC_Hom	Allele Count Homozygous
ACMG	American College of Medical Genetics and Genomics
ACVRL1	Activin-receptor–like kinase 1
AF	Allele Frequency
AFR	African
AN	Allele Number
API	application programming interface
BAM	Binary Sequence Alignment/Map
BCF	Binary Variant Call Format
BCL	binary base call
BMP	Bone Morphogenetic Protein
BMPR2	Bone Morphogenetic Protein Receptor Type 2
bp	base pair

CADD	Combined Annotation Dependent Depletion
CATGO	Cambridge Translational GenOmics
CAV1	Caveolin-1
CDISC	Clinical Data Interchange Standards Consortium
ClinGen	Clinical Genome
CMR	Cardiac Magnetic Resonance
CNV	Copy Number Variation
CPIC	Clinical Pharmacogenetics Implementation Consortium
CPIC	Pharmacogenomics Research Network
CPU	central processing unit
CR	Call Rate
CSCS	Clinical School Computing Service
dbSNP	Database of Single Nucleotide Polymorphism
EAS	East Asians
EBI	EMBL – European Bioinformatics Institute
EC	Endothelial Cell
eCRF	electronic Case Report Form
EIF2AK4	Eukaryotic Translation Initiation Factor 2 Alpha Kinase 4
EMBL	European Molecular Biology Laboratory
eMERGE	Electronic Medical Records and Genomics
EMR	Electronic Medical Record
ENG	Endoglin
EUR	European
ExAC	Exome Aggregation Consortium

FASTQ	text based sequence file format
FTP	File Transfer Protocol
GA4GH	Global Alliance for Genomics and Health
GATK	Genome Analysis Toolkit
GB	Gigabyte
GFF	Gene Feature Format
GiaB	Genome in a Bottle
gnomAD	Genome Aggregation Database
GRC	Genome Reference Consortium
GRCh	Genome Reference Consortium human genome
GTF	Gene Transfer Format
gVCF	genome VCF
GWAS	Genome-Wide Association Studies
HapMap	Haplotype Map
HDFS	Hadoop File System
HET	heterozygous
het/hom	heterozygous / homozygous
HGMD	Human Gene Mutation Database
HGP	Human Genome Project
HHT	Hereditary Haemorrhagic Telangiectasia
HI	Haploinsufficiency
HOM	homozygous
HPCS	High Performance Computing Service
HPO	Human Phenotype Ontology

HTS	High-Throughput Sequencing
HUVEC	Human Umbilical Vein Endothelial Cells
HWE	Hardy-Weinberg Equilibrium
IBD	Identity-By-Descent
IGV	integrative genomics viewer
INDEL	Insertions or Deletions
IQR	Interquartile Range
IRDiRC	International Rare Diseases Research Consortium
JAX-WS	Java API for XML Web Services
JDK	Java Development Kit
JSON	JavaScript Object Notation
Kb	Kilo-bases
KCNK3	potassium channel, subfamily K, member 3
KCO	transfer coefficient for carbon monoxide and vasoreactivity
KVM	Kernel-based Virtual Machine
LD	Linkage Disequilibrium
LoF	Loss-of-Function
LVEDP	Left Ventricular End-Diastolic Pressure
MAF	Minor Allele Frequency
Mb	Mega-bases
minOPR	minimum OPR
MNV	Multi Nucleotide Variants
mPAP	mean Pulmonary Arterial Pressure
mPAWP	mean Pulmonary Arterial Wedge Pressure

NGS	Next Generation Sequencing
NIH	National Institutes of Health
NIHR BR-RD	NIHR BioResource – Rare Diseases
NIHR	National Institute for Health Research
NIST	National Institute of Standards and Technology
NMD	nonsense-mediated decay
NoSQL	Not Only SQL
OC	OpenClinica
ODM	Operational Data Model
OID	Object Identifier
OPR	Overall Pass Rate
PAEC	Pulmonary Artery Endothelial Cell
PAHAFF	affected PAH adults
PAHIDX	affected PAH adult index cases
PAH	Pulmonary Arterial Hypertension
PASMC	Pulmonary Artery Smooth Muscle Cell
PCA	Principal Component Analysis
PCH	Pulmonary Capillary Haemangiomatosis
PDF	Portable Document Format
PF	Pass Frequency
PH	Pulmonary Hypertension
PTV	Protein Truncating Variant
PVOD	Pulmonary Veno-Occlusive Disease
Q1	First quartile

QR	Quick Response
RDBMS	Relational Database Management System
SAS	South Asian
SD	Standard Deviation
Sift	Sort intolerant from tolerant
SMAD	the mothers against decapentaplegic
SNV	Single Nucleotide Variation
SOAP	Simple Object Access Protocol
SOP	Standard Operating Procedures
SPS	Small Patella Syndrome
ssh	Secure Shell
SV	Structural Variation
TASK-1	twik-related acid sensitive K+
TB	Terabyte
TBX4	T-box 4
TSS	Transcription Start Site
Ts/Tv	Transition/ Transversion
UBRG_EUR	Unrelated European NIHR BR-RD subjects
UBRG	Unrelated NIHR BR-RD subjects
UPAHC	unrelated non-PAH control subjects
UWGS10K	unrelated WGS10K
VCF	Variant Call Format
VILMAA	Variant Infrastructure for Loading, Merging, Annotating and Analysing human genomes

WGS	Whole Genome Sequencing
WSDL	Web Service Description Language
XML	Extensible Markup Language
YCSB	Yahoo! Cloud Serving Benchmark

Chapter 1

Introduction

1.1 Genome variation in health and disease

1.1.1 Human reference genome

The first human draft reference genome was published in January 2001 by the Human Genome Project (HGP) after more than 10 years in making (International Human Genome Sequencing Consortium, 2001). The success of the collaborative achievement marked the beginning of human genomics. In 2004, the international collaboration provided the finished human genome with high accuracy and approximately 99% coverage to the research community as reference Build 35 (International Human Genome Sequencing Consortium, 2004). The published reference was assembled from multiple individuals and represent a haploid consensus of diploid human genomes. Further assessment of the genome provided an insight into single nucleotide variation (SNV) and described the complex architecture of so far uncharacterised structural variation (SV) or copy number variation (CNV) (Bailey et al., 2001). Classification of SNVs in gene coding regions separates non-synonymous (alters amino acid in the protein) from synonymous (change does not cause amino acid alteration) variants. The terminology for further variation types is described in Fig. 1.1. Identified variants were collected in the single nucleotide polymorphism database (dbSNP), which provides a public archive of SNVs and multi-base insertions or deletions (INDELs) (www.ncbi.nlm.nih.gov/snp/). Disease causing variants or variants associated with inherited diseases were collected in databases and include the online mendelian inheritance in man (OMIM) (McKusick, 2007), ClinVar (Landrum et al., 2016) as well as the human gene mutation database (HGMD) (Stenson et al., 2003) among others. The frequency of common SNVs was determined in samples with African, Asian and European ancestry and explored for common patterns across the genome. Such patterns included the identification of SNV

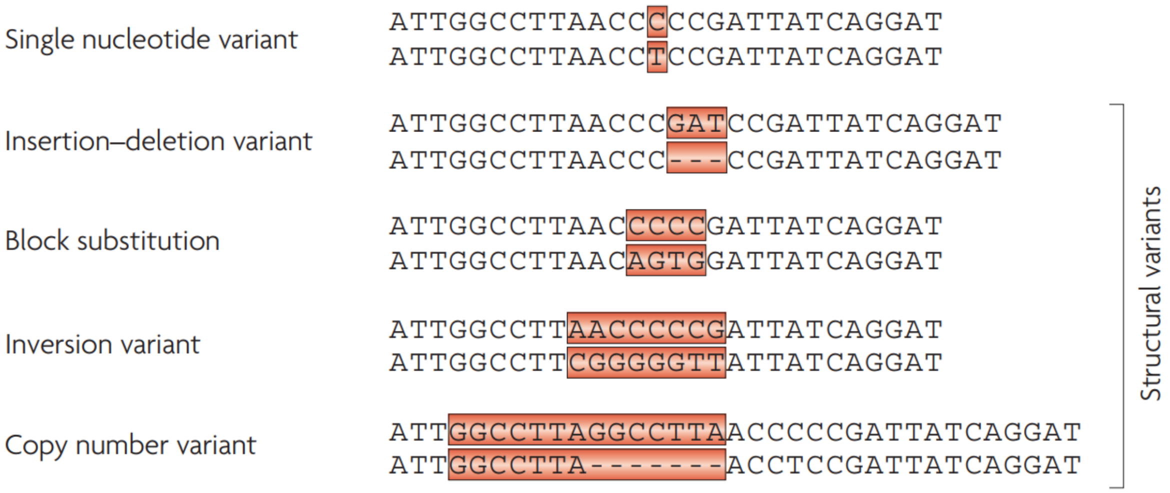

Single nucleotide variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCTCCGATTATCAGGAT
Insertion–deletion variant
ATTGGCCTTAACCCGATCCGATTATCAGGAT
ATTGGCCTTAACCC---CCGATTATCAGGAT
Block substitution
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACAGTGGATTATCAGGAT
Inversion variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTCGGGGGTTATTATCAGGAT
Copy number variant
ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTA-------ACCTCCGATTATCAGGAT
Structural variants

Figure 1.1 Description of variant types with their respective term. Figure redrawn from Frazer et al. (2009)

with a strong correlation with their neighbours as blocks of linkage disequilibrium, also called haplotype blocks (see Fig. 1.2). A genome wide analysis described the diversity of these haplotype blocks as part of the haplotype map (HapMap) project (International HapMap Consortium, 2003, 2005, 2007). The genome reference consortium (GRC) (Church et al., 2011) continued working on the human assembly to correct erroneous bases, update the tiling path in variable regions and increase the coverage of the genome by closing sequencing gaps. Initially, the concept of "a golden path" guided the collaboration to reduce the human genome to one non-redundant haploid representation (Kent and Haussler, 2001) that allows the characterisation of SNVs. Recent discoveries of the large diversity of SV (Feuk et al., 2006; Mefford and Eichler, 2009) required a rethink to model the structural complexity. The latest release of the human reference genome assembly GRCh38 was published in 2013 (Genome Reference Consortium, 2013) and provides mechanisms to describe alternate representations of a locus as well as unlocalised and unplaced sequences. Additional improvements included extra sequence content, removal of redundant or falsely duplicated sequences and correction of clinically relevant regions. After the availability of the reference genome, raw sequence information were processed and annotated with the latest available reference annotations by the integrated data platform Ensembl (Hubbard et al., 2002) to offer free access to a complete data set. These annotations included gene model, comparative genomics, variant and regulatory region information for genomic research (Cunningham et al., 2015). In practice, the previous genome release GRCh37 (Flicek et al., 2010) is still in use. The GRCh37 release was released in 2009 and coincided with the introduction of next generation sequencing (NGS), which accelerated the generation of sequence data and identification of variation information. While the reference genome was a mosaic of different individuals, the NGS technology put the personal genome in reach to be used in diagnosis and treatment guidance of genetic diseases.

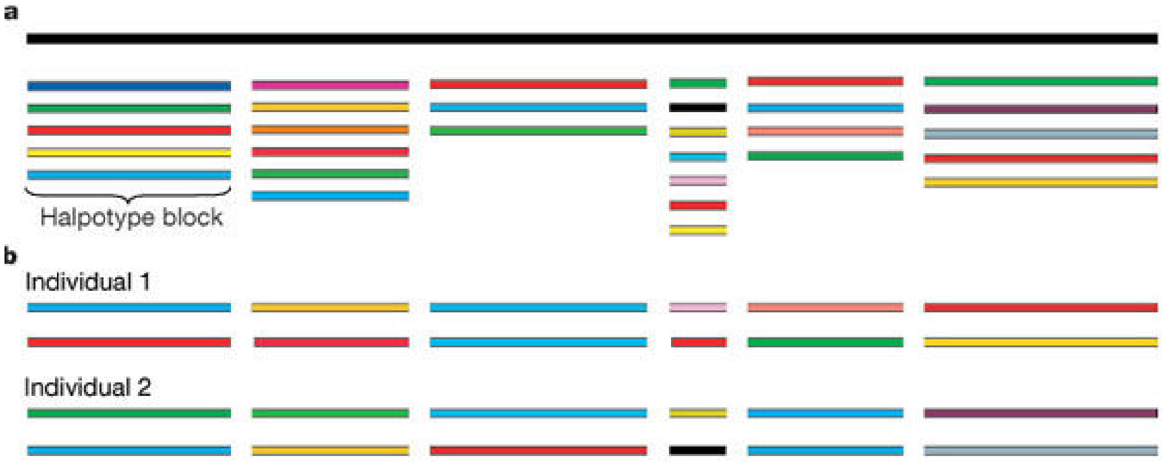


Figure 1.2 Haplotype blocks in a population. (a) Human chromosome are constructed of different haplotype blocks. Such haplotype blocks are indicated as bars of different colours. (b) Example chromosomes for two individuals are made up by two copies of each haplotype block. By chance, one individual can have the two copies of the same haplotype block. Figure redrawn from Pääbo (2003)

1.1.2 Personal human genome

In 2007, whole-genome shotgun in combination with conventional Sanger dideoxy sequencing produced the genome sequence of one individual (Levy et al., 2007). The personal genome sequence of J. Craig Venter was assembled from scaffolds and resulted in 2.8 Gb of continuous sequence with every given region being covered approximately 7.5-fold. The genome provided single base resolution. The analysis of the genome identified 4.1 million variants compared to the human reference assembly and 1.3 million of these variants were novel (not in dbSNP). Levy et al. (2007) showed that SNVs accounted for 78% of changes, non-SNVs variants (from small INDELs up to large CNVs) cover 74% of all changed variant bases.

The Venter genome was followed by the personal genome of James D. Watson (Wheeler et al., 2008), which shared 1.68 million of the SNVs and identified 7,648 different protein coding changes compared to the Venter genome. Non-synonymous variants were compared against HGMD and revealed 32 matching entries previously reported as disease causing.

Further analysis identified the subject as a carrier of 10 highly penetrant genetic disease loci (Wheeler et al., 2008), which is in line with the expected number of HGMD disease-causing mutations in a healthy individuals (Gambin et al., 2015; Xue et al., 2012). Many pathogenic variants in HGMD were also shown to be too frequent in a human population (Lek et al., 2016) to be consistent with the prevalence and penetrance of the disease (Blekhman et al., 2008).

The genome from an African (Bentley et al., 2008; McKernan et al., 2009) and an Asian (Wang et al., 2008) individual each were sequenced and compared against the Venter and Watson genome. The number of non-synonymous SNVs was found to be similar between the Asian (7,062), Venter (6,889) and Watson (7,319), but showed an increase in the African (9,902) genome. Personal genomes provided a first glimpse of the genomic diversity and private variation on an individual level of different ancestry. The genomic diversity was further explored by sequencing a 100 kilo-bases (Kb) region in 692 individuals of diverse populations as part of the HapMap project (International HapMap 3 Consortium et al., 2010). The unearthed common and rare genetic variation in the region contained 77% of novel variants (not in dbSNP) of which 99% had a minor allele frequency (MAF) <5%. The scale of new and low frequency variants highlighted the importance of a reference resource including a large cohort of multiple populations to assess personal genomes for rare disease causing variants.

1.1.3 Human population genomics for medical interpretation

The 1000 Genomes (1kG) project aimed to characterise and provide accurate haplotype information for all types of human variation in different populations, which is a progression of the International HapMap Project. Advances in high-throughput sequencing (HTS) technologies enabled to access more than 95% of the human genome compared to the 100 Kb region analysed as part of the HapMap project. The analysis of the 1kG project was conducted in 3 phases and covered allele frequencies of 1% or higher in present populations. The first phase tested different sequencing approaches from genome wide low depth (2-6x), high coverage (40x) in trios and high depth (>50x) exon targeted sequencing for a total of 1,092 samples. The number of samples were extended to around 1,700 in phase 2 to develop new methods in order to handle and analyse the complexity of present variant information. Phase 3 represented 2,504 samples from 26 populations in Africa, East Asia, Europe, South Asia, and the Americas. The described variants were the first comprehensive assessment of common (MAF >5%) and low-frequency (MAF 1-5%) genetic variants across diverse populations (1000 Genomes Project Consortium et al., 2015).

The project raised questions about the true clinical significance and consistency of variants reported in the public archives ClinVar (Landrum et al., 2016) and HGMD (Stenson et al., 2014). Variants in ClinVar showed an equal distribution per super population for different frequency ranges while HGMD saw an enrichment of low-frequency variants in Europeans that are more likely to have a pathogenic effect. Loss-of function (LoF) variants (stop gained, splice acceptor, splice donor and frameshift) were found to be under expected purifying selection and biased towards low-frequency variants. Low-frequency LoF variants were private to individuals or sub-populations and singleton heterozygous variants were very similarly represented across different populations.

Whole genome sequencing also enabled further insight into the extent of structural variation (SV) (1000 Genomes Project Consortium et al., 2010). Structural variation (SV), their importance and association with disease and phenotypic variation was first recognised by Sebat et al. (2004) using microarrays. Previously regarded as rare events, the study reported the extent and size of large scale ($\geq$ 100Kb) gains and losses of copy number variation (CNV) in 20 healthy individuals from a variety of geographic backgrounds. These individuals revealed a total of 221 CNVs representing 76 unique CNVs with an average length of 465 Kb and overlapped 70 different genes associated with disease or involved in functional and regulatory processes. Whole genome sequencing enabled further insight into the extent of eight different classes of SV described by the 1kG project in 26 human populations (Sudmant et al., 2015) . The constructed integrated SV map included naturally occurring homozygous gene knock-outs with detailed breakpoint information, variant class specific size distribution and detailed description of populations.

In 2015, the UK10K project characterised 24 million novel sequence variants, assessed the contribution of rare genetic variation (MAF <1%) to human traits, which are insufficiently represented in genome-wide association studies (GWAS) (UK10K Consortium et al., 2015). Samples from extensively phenotyped cohorts were whole exome (high read depth, 80x) or whole genome (low read depth, 7x) sequenced. From a cohort of nearly 10,000 samples, disease-causing variants were reported in 2.3% of the UK10K cases (42 out of 1,805 individuals), which demonstrated the possibility of clinical application. Complex trait-associated loci harboured novel rare variants that suggested a genetic association. The UK10K showed the value of the large-scale sequencing data for complex traits and provides whole-genome variation information as a reference resource to increase accuracy and coverage of rare and low-frequency variants in addition to the 1kG panel.

In 2016, the exome aggregation consortium (ExAC) provided the allele frequencies of 60,706 humans to study protein-coding genetic variation following the UK10K as a reference resource (Lek et al., 2016). One year later, the genome aggregation database (gnomAD) made a reference resource available, that consists of 15,496 whole-genome and 123,136 exome sequences. The whole-genome part of the gnomAD release contained 241 million variants and showed the scale of rare variation in a population. The availability of such extensive variant catalogues allowed the reassessment of pathogenic genes in Mendelian diseases and found the under- and overestimation of clinical importance of some genes (Walsh et al., 2017). The ExAC MAF of the reported disease-causing variants was disproportionate high compared to the reported penetrance and the paper showed that genes with a low disease association were often initially discovered through candidate gene studies. In contrast to disease causing variants, the discovery of protective genetic variants was aided by large reference resources for severe Mendelian conditions and opened the prospect of therapeutic strategies (Chen et al., 2016a).

The driving force behind these research results was the advance in HTS technology, which enabled an increase in sequencing output and a reduction in price. The main used methods were whole exome (WES) and whole genome (WGS) sequencing using Illumina technology and are described later.

1.1.4 High-throughput sequencing workflow

The first human genome was sequenced using Sanger-based (dideoxy chain termination sequencing) methods at an estimated cost of 2.7 billion dollars and took 13 years to complete (Gyles, 2008). In 2007, the HTS technology 454-sequencing was used to sequence the Watson genome at a price of 2 million dollars and in a matter of 2 months. Advances in HTS technologies reduced the cost per human genome further to a price of 1,245 dollars in October 2015 (Fig. 1.3) and the market is currently dominated by Illumina machines.

Illumina's sequencing technology is based on the reversible terminator chemistry concept (Bentley et al., 2008). For whole genome sequencing (WGS), the preparation of DNA samples involves the fragmentation of the DNA (Kozarewa et al., 2009). These genomic DNA fragments attach to a glass surface and clonal amplification creates 'clusters' of the same fragment. DNA clusters are then sequenced by repeated cycles of single base extension using a set of four reversible terminators. The terminators are labelled with different fluorophores and laser-induced excitation allows the identification of the specific base for each cluster by imaging technology. The number of cycles determines the number of called bases and a quality value is provided for each base call. Lower quality scores assigned by the base-calling algorithm can be down to a cluster of mixed fragments, a failed cycle (phasing) or multiple bases being synthesized in one cycle (pre-phasing) (Ledergerber and Dessimoz, 2011). For each cycle, base-calling information are stored in binary base call (BCL) format for the image coordinate of the cluster.

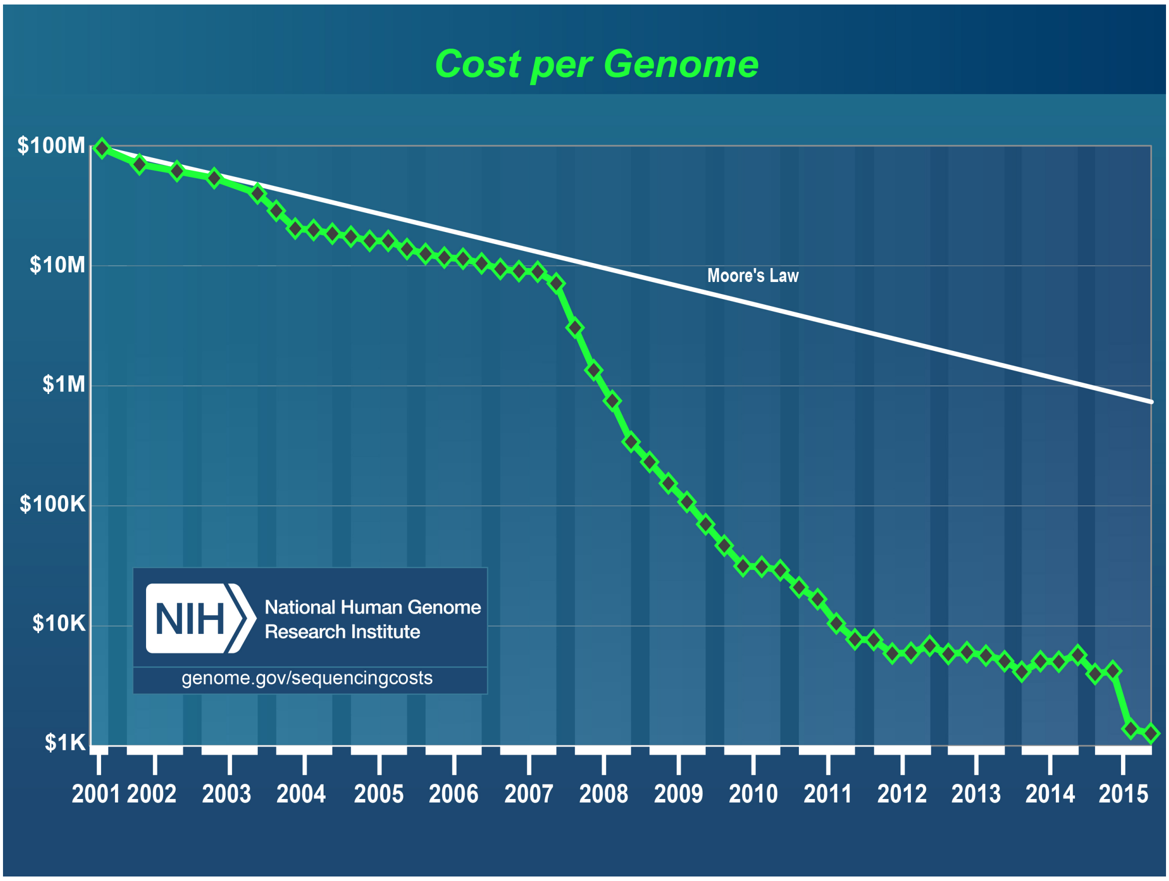


Figure 1.3 Sequencing cost per human genome interrupted by high-throughput sequencing technology in 2007. Figure redrawn from Wetterstrand, KA (2015)

A DNA fragment, also called read, is the reconstructed base-calling information for one cluster from the glass surface across the images from all cycles. After the last base-calling cycle, fragments are extracted from the BCL file and can be stored in a standard text based sequence file format (FASTQ) (Picard, 2017). The extracted fragments represent the information of the sequenced individual in millions of small pieces. Different computational approaches have been developed to reconstruct the genomic information of the individual from these small pieces. These approaches include the *de novo* assembly of the individual genome from fragments (Gnerre et al., 2011; Iqbal et al., 2012; Li et al., 2010; Simpson et al., 2009; Zerbino and Birney, 2008) and the alignment of the fragments (Langmead et al., 2009b; Li and Durbin, 2009; Li et al., 2009b; Raczy et al., 2013) to the reference genome. The *de novo* assembled genome can be stored in a graph structure and compared against a reference genome graph to identify simple as well as structural variation (Iqbal et al., 2012). In contrast, fragments aligned against the reference genome can be stored in the standardised sequence alignment / map (SAM) or in binary SAM (BAM) format and the differences in the alignment are utilised to identify single base variants or insertions / deletions of up to tens of bases (DePristo et al., 2011; Li et al., 2009a; Raczy et al., 2013; Rimmer et al., 2014; Wei et al., 2011). Larger copy number variations (CNV) or structural variation (SV) can be detected by extracting read-depth coverage (Campbell et al., 2008; Chiang et al., 2009; Roller et al., 2016), split-read (Chen et al., 2016b; Wang et al., 2011; Ye et al., 2009) or both information (Abyzov and Gerstein, 2011; Rausch et al., 2012) from the BAM file. Resulting simple CNVs and complex SVs can also be stored in variant call format (VCF) (Danecek et al., 2011).

The extensive list of developed tools to process FASTQ, BAM or VCF files resulted in method specific variants. A best practices workflow by the gene analysis toolkit (GATK) provided guidance to retrieve high-quality variants starting with the FASTQ files and protocols for variant filtering and annotation (Van der Auwera et al., 2013). Further development of alignment, variant calling and variant filtering algorithms prompted a comparison of different tools and pipelines. Variants of one subject from the 1kG project were experimentally validated, which resulted in the comprehensive variant dataset provided by the national institute of standards and technology led by the genome in a bottle (NIST-GiaB) consortium (Zook et al., 2014). The results of a comparison of established pipelines against the GiaB reference data set provided an insight into the performance differences (Cornish and Guda, 2015). Difference in variant representation, specifically for INDELs, were highlighted and standardised by a formal definition of a normalisation process (Tan et al., 2015).

1.1.5 Sample variant aggregation

Since the first personal genome in 2007, samples are sequenced and processed on an individual level resulting in a single sample VCF or genome VCF (gVCF) file. Case-control and trio based study designs focus on groups of individuals and analysis tools require merged or aggregated variant information as a multi-sample VCF file for the analysis. First approaches bridged the gap between a single and multi-sample VCF file by aggregating VCF files (Li et al., 2009a), but missed reference or no-call information for positions in the merged file, that were not present in the single sample VCF. In addition, the performance of aggregating individuals deteriorated with increasing number of individuals. The gVCF files were introduced for a complete variant, reference and no-call representation for each position of the whole genome, but were found to be very large in size. Illumina collapsed regions of the same type to reduce the space and developed an aggregation tool (Illumina, Inc., 2015) surrounding the collapsed gVCF format, which scales beyond 3,000 samples. A different approach was followed by FreeBayes (Garrison and Marth, 2012), which provided aggregated variant calling based on the alignment data. The scalability was limited due to the quantity of information required for an aggregated variant call, but allowed the selection of the most-confident genotype across a population. The genome analysis toolkit (GATK) applied a similar strategy by creating single sample gVCF files with additional information to perform an aggregated variant call (McKenna et al., 2010). Large sequencing projects like ExAC (Lek et al., 2016) provided intermediate releases with increasing number of samples until the final public release (ftp://ftp.broadinstitute.org/pub/ExAC_release). The time between releases ranged from months to years and highlights the challenge in aggregating, preparing and releasing variants for a large number of samples. Current approaches from GATK and FreeBayes require a recall of the whole cohort. The computational time can be as high as 100,000 central processing unit (CPU) days for analysing a population of 28,075 whole genome samples (Eggertsson et al., 2017), of which variant calling is one part of. Calling variants is a huge computational burden for rapidly growing cohorts.

1.1.6 Prioritisation of genome variation

The analysis of genome variation data involves sifting through millions of variants across the whole genome with some falling within protein coding regions. Computational approaches integrate genome annotation, functional genomic data, allele frequency and effect prediction information to assess and annotate each variant. For Mendelian diseases, prioritisation tools dissect these annotations to separate likely disease causing variants from benign and private changes and identify the one or more responsible variants. Variant effect predictor (VEP) provides a rich toolset to interrogate variants genome wide by providing a comprehensive collection of deleteriousness, conservation and allele frequency annotations (McLaren et al., 2010). In addition to a predefined set of annotations, the custom annotation module allows to add additional public available or private information to a variant annotation set. A set of annotation is created for each transcript or regulatory feature specific to present a comprehensive view for cases where one variant overlaps more than one transcript or regulatory region. Based on the transcript and regulatory information, the consequence type(s) of a variant can be predicted and is (are) reported as a standardised sequence ontology (SO) term(s) illustrated in Fig. 1.4. Other tools were developed to use different layers of information to infer the functional importance of variants (Gnad et al., 2013; Thusberg et al., 2011). The conservation of sequence is assessed at different levels to predict the importance. Genomic conservation measurements are provided by the genomic evolutionary rate profiling (GERP), PhastCons and PhyloP (Cooper et al., 2005; Davydov et al., 2010; Margulies et al., 2003). Protein sequence conservation is used to judge the importance of an amino-acid position for the functioning of a protein and to infer the deleteriousness of amino-acid substitution. The sort intolerant from tolerant (Sift) (Ng and Henikoff, 2003) and MutationAssessor (Reva et al., 2011) implement methods to calculate the protein conservation, but the resulting scores are sensitive to the chosen input alignment (Hicks et al., 2011). The structure of the protein is an additional layer integrated with other sequence features by PolyPhen-2 (Adzhubei et al., 2013) to predict the functional consequence. The combined annotation dependent depletion (CADD) score comprises 63 separate annotations for a genome wide assessment and includes protein effect prediction, conservation, regulatory information as well as predictions from other tools (Kircher et al., 2014). The accuracy of the predictions was assessed by selecting variants with known consequences and found a high number of false positives in all assessed

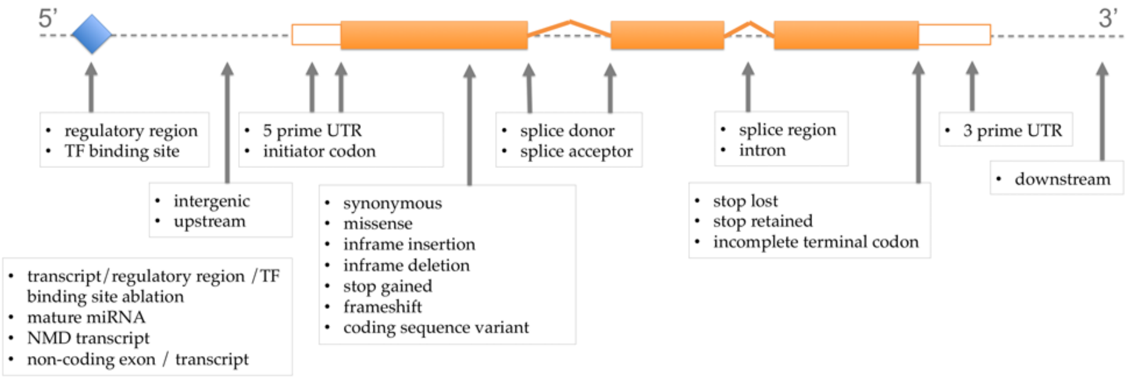


Figure 1.4 Variant consequence display terms relative to the transcript structure. Figure redrawn from Ensembl (2015)

tools, which highlights the complexity of predicting the affect of a variant (Miosge et al., 2015; Wang and Wei, 2016).

After the annotation process, the enriched dataset is available for filtering to identify causal variants. Filter tools provide a computational approach to select predicted deleterious variants and enables to filter on one value or combine information into a complex query (Cingolani et al., 2012). In addition to variant annotations, pedigree information can be included into the analysis to identify autosomal dominant, autosomal recessive or *de novo* mutations in a family (Paila et al., 2013). After applying filter strategies, the result requires cautious examination in classifying a variant as deleteriousness due to the highlighted limitation of computational approaches.

The american college of medical genetics and genomics (ACMG) developed guidelines to standardise the terminology, assessment and classification of variants in genes that cause mendelian disorders (Richards et al., 2015). The guidelines recommend to classify variants into five categories using the terms “pathogenic”, “likely pathogenic”, “uncertain significance”, “likely benign”, and “benign”. Evidence items for the classification process include population data, computational data, functional data, segregation data and allelic data. Public variant archives and knowledge bases like ClinVar (Landrum et al., 2016) or clinical genome (ClinGen) (Rehm et al., 2015) adopted the terminology for Mendelian diseases. Further automation and standardisation was provided by computational framework, which translated the ACMG guidelines into a Bayesian classifier (Tavtigian et al., 2018). The standardised approach aims to reduce the number of variants being reported as “causative” and to improve the genomic variant interpretation.

1.1.7 Rare variant study design

Genome-wide association studies (GWASs) identified robust associations between thousands of common variants and complex traits and diseases. Low-frequency ($0.5\% \leq$ MAF <5%) and rare (MAF <0.5%) variants are not well represented by genetic markers in a GWAS. The missing heritability of disorders is attributed to these variants unrepresented in GWAS (Cirulli and Goldstein, 2010; Maher, 2008; Manolio et al., 2009). Highly penetrant rare variants are known to cause many mendelian disorders or rare forms of common diseases (Gibson, 2012). The availability of whole genome sequencing (WGS) allowed the identification of rare variants of individual genotypes in large cohorts and evaluate the contribution of rare genetic variation to disorders. The 1000 Genomes Project sequenced >2,500 individuals at low depth and enabled to identify 95% of variants that have an allele frequency of 1% (1 in 100) or higher. Larger studies focused on sequencing the 2% of the genome that codes for protein (Bamshad et al., 2011) at a higher depth for accuracy. The exome sequence project (ESP) comprised 6,515 individuals as part of the National Heart, Lung, and Blood Institute (NHLBI) and sequenced the exome to identify rare genetic variations (Fu et al., 2013). Recently, the exome aggregation consortium (ExAC) studied 60,706 samples and enables to filter for very rare variants with a MAF <0.01% and identify single variants with a frequency below 1 in 120,000 (Lek et al., 2016). The genome aggregation database (gnomAD) analysed 15,496 whole genomes and 123,136 exomes, and provided the allele frequencies across different populations.

The high resolution of genomes variation and reference allele frequencies empowered the study of rare and undiagnosed diseases and interpretation of variants (Bahcall, 2016). For a robust identification of associations, the design of a whole genome sequencing (WGS) study involves multiple processing and analysis steps (see Fig. 1.5). Planing the rare variant analysis starts with the selection of inclusion and exclusion criteria for patients as well as control subjects. The selection criteria rely on an agreed clinical classification of the disease or require the collection of phenotype information to group patients during the analysis (Grateau et al., 2013; Pathak et al., 2017; Westbury et al., 2015). Following the study design, the study is submitted for ethical approval to the ethics committee to judge the ethical acceptability for

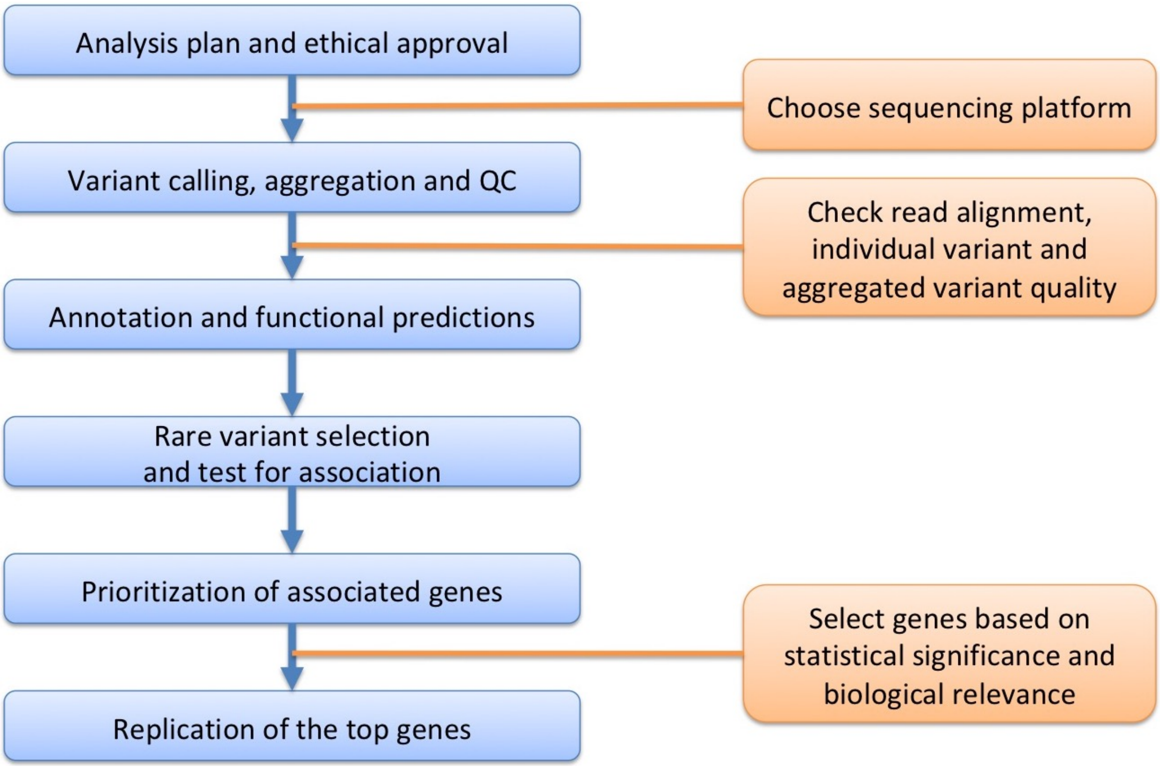

Analysis plan and ethical approval
Choose sequencing platform
Variant calling, aggregation and QC
Check read alignment, individual variant and aggregated variant quality
Annotation and functional predictions
Rare variant selection and test for association
Prioritization of associated genes
Select genes based on statistical significance and biological relevance
Replication of the top genes

Figure 1.5 Data and analysis workflow for rare variant association studies applying whole genome sequencing. Figure redrawn from Lee et al. (2014)

involving human participants (Fletcher, 2015). Quality control processes are required at every step to ensure the accurate identification, handling and standardised processing of samples. Data quality issues can occur at various analysis levels and the careful investigation of possible issues is important. Sequencing artefacts and errors can be introduced by including different sequencing platforms or changing sequencing protocols (Ross et al., 2013). To discover and quantify biases, a number of measurements are available to assess sequence alignment, individual and aggregated variants (Van der Auwera et al., 2013; Wang et al., 2015). Variants are evaluated by bioinformatic methods to predict the likely impact or consequence and to annotate variants with public available information including allele frequencies from large sequencing projects (see chapter 1.1.6 on page 9). These variant information can be used in the association analysis and help with the interpretation, discovery and prioritization of genes. Replication of rare-variant associations is a challenge for rare diseases and the strategy of follow-up studies depends on multiple factors, including disease rarity, characteristics of the gene discovery and study budget. Follow-up studies could include functionally analysing the discovered gene by studying model organisms or cell lines (Edwards et al., 2013).

1.2 Aims

The overarching objective of the research presented in this Thesis is to advance our understanding of the genetic architecture of PAH by the application of whole-genome sequencing technology.

Specifically my aims were as follows:

1. To collect and curate multi-dimensional clinical phenotype data from a large cohort of affected PAH patients and validate the measurements for correctness and completeness.

2. To undertake an analysis of whole-genome sequence variation in a case-control study to identify an enrichment of deleterious variants in protein-coding genes in affected PAH patients. Genes demonstrating enrichment will be assessed for clinical patterns by integrating the phenotype data.

3. To facilitate this effort by developing a scalable analysis platform to incorporate variation, variant annotation and phenotype information for enhanced genome-wide integrative data analysis.

Chapter 2

Phenotype capture

2.1 Introduction

The general introduction (see chapter 1.1.3 on page 4) provided an overview of the application of whole genome sequencing (WGS) to characterise human variation in different ethnic as well as disease populations. I have discussed the contribution of rare genetic variation to human traits and the importance of deep phenotyped cohorts to discover trait-associated loci. This chapter focuses on patients with the rare disease idiopathic and heritable pulmonary arterial hypertension, the collection of phenotype data and management of biological samples.

2.1.1 Pulmonary arterial hypertension

Pulmonary arterial hypertension (PAH) is characterised by elevated blood pressure in the lung. PAH is a rare disease, with an estimated prevalence in the range from 10 to 52 cases per million per year (Escribano-Subias et al., 2012; Frost et al., 2011; Humbert et al., 2006; Peacock et al., 2007). Shortness of breath, fatigue or weakness, angina, syncope, peripheral edema and abdominal distension are some of the non-specific symptoms a PAH patient presents with a mean age of 38 (± 11) (Barst et al., 2004; Elliott et al., 2006). Females have a higher prevalence compared to males at a ratio of 2.3:1 (Larkin et al., 2012). PAH is part of the disease group pulmonary hypertension (PH) and the underlying disease pathogenesis divides PAH into four subgroups. The subgroups include idiopathic PAH (IPAH), heritable PAH (HPAH) caused by gene mutations, drug/toxin induced PAH and PAH with associated diseases (Simonneau et al., 2013). Prognosis and management for each subcategory is different, which highlights the importance of a correct diagnosis (Benza et al., 2012; Escribano-Subias et al., 2012). The diagnosis is based on haemodynamic measurements performed during right heart catheterisation (see Fig. 2.1). PH is defined by a mean pulmonary artery pressure $\geq$ 25 mm Hg at rest (Hoeper et al., 2013). PAH is further characterised by the presence of pre-capillary PH, defined by a normal pulmonary artery wedge pressure (PAWP) of $\leq$ 15 mm Hg at rest. The recommendation suggest the catheterisation of the left heart to measure the left ventricular end-diastolic pressure (LVEDP), in case the PAWP is unreliable (Barst et al., 2004; Galie et al., 2015). The normal range for LVEDP in PAH is $\leq$ 12 mm Hg. Elevated LVEDP is an indication of left heart disease, while an elevated PAWP characterises postcapillary PH (Galiè et al., 2004).

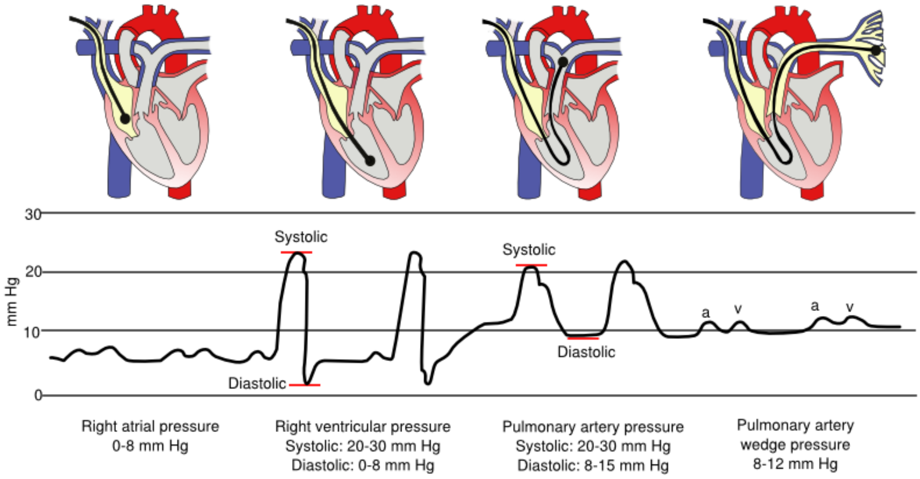


Figure 2.1 Pressure and pressure waveform description at different locations during pulmonary artery catheterisation. Pulmonary artery pressure (PAP) and pulmonary artery wedge pressure (PAWP) are the last two points of measure (left ventricular end-diastolic pressure not provided). Reference values are shown for each locations and measurement. Figure adapted from Wierda and Hoftijzer (2016)

For my project I focus on the idiopathic and heritable forms, which represent 38.9% and 4.2% respectively of PAH prevalent cases (Humbert et al., 2006). This amounts to around 1 in 170,000 and 1 in 1.6 million people for IPAH and HPAH respectively. The median survival for adults with IPAH and without treatment is 2.8 years from the point of diagnosis (D'Alonzo et al., 1991). Therapies are available, but Charalampopoulos et al. (2014) have shown differences in response to treatment based on risk factors, which emphasises the need to further elucidate the underlying mechanisms. For patients on treatment, Humbert et al. (2010) observed a three year survival of 54.9% for idiopathic, heritable or anorexigen-induced PAH cases.

PAH is caused by remodelled pulmonary arteries leading to a raised pulmonary vascular resistance. The remodelling is characterised by increased proliferation and apoptosis resistance of pulmonary artery cells (Budhiraja et al., 2004). Arteries are blood vessels carrying blood away from the heart. An artery (see Fig. 2.2) consists of endothelial cells (ECs) and smooth muscle cells (SMCs), which build up functional parts of the vessel wall. The endothelial cells (ECs) form a mono-cellular layer, the endothelium, that lines the blood vessel wall separating the tissue (abluminal side) from the circulating blood (luminal side). This selective barrier traffics molecules, proteins and nutrients to surrounding cells. Intima is the layer beneath the endothelium and largely consists of elastic fibers, collagens and amorphous proteoglycans. Smooth muscle cells are part of the media layer and control the blood pressure by changing the diameter of the vessel lumen. The interface between the vessel and surrounding tissue is the externa, a layer of connective tissue and fibroblasts. Pulmonary and umbilical arteries carry deoxygenated blood from the heart to the lung, while

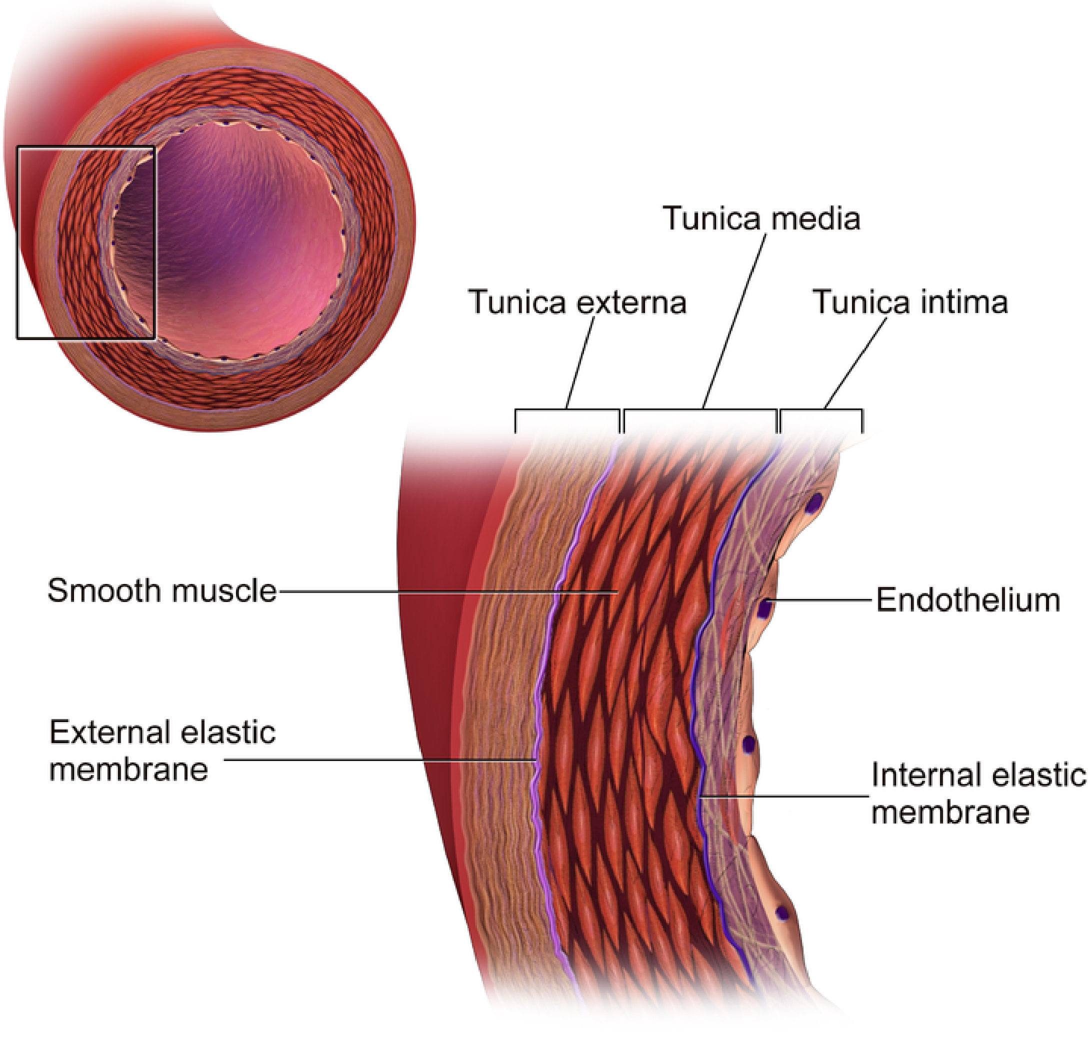


Figure 2.2 Structure of an artery wall. Extracted from Wikiversity (2014)

the remaining arteries carry oxygenated blood. The ECs and SMCs part of the pulmonary arteries are implicated in PAH and called pulmonary artery endothelial cells (PAEC) and pulmonary artery smooth muscle cells (PASMCs).

The gene bone morphogenetic protein receptor type 2 (*BMPR2*) binds members of the transforming growth factor beta (TGF-β) family of signaling molecules. Variants in *BMPR2* have been found to be the main cause of HPAH (69%) and also explains 21% of cases in IPAH (Aldred et al., 2006). The majority of IPAH cases are still unexplained, but other genes mainly part of the TGF-β pathway are implicated in the disease and discussed later.

2.2 Methods

2.2.1 Study design

Subjects were recruited for two separate studies, the national cohort study of idiopathic and heritable PAH and the national institute for health research (NIHR) BioResource - Rare Diseases (BR-RD) study PAH cohort. The purpose of the national cohort study was to set up a national cohort and biorepository of heritable PAH cases and their relatives, and idiopathic PAH. Participants were followed to initiate longitudinal clinical evaluation and sampling of HPAH family members and to elucidate the underlying genetic architecture of idiopathic and heritable PAH. The main objective for the NIHR BR-RD was the establishment of a comprehensive BioResource of participants with rare diseases to identify the cause of disease in those individuals using Next Generation Sequencing Techniques (NGST). The study aimed to improve the diagnostic rate through NGST for future use as diagnostic test and perform subsequent studies to validate possible novel treatments. The outcome of the studies should be used for the health services research to develop funding schemes for use in the NHS.

2.2.2 Ethical approval

The national cohort study of idiopathic and heritable PAH recruited cases and their relatives from the UK National Pulmonary Hypertension Centres, Universite Sud Paris (France), the VU University Medical Center Amsterdam (The Netherlands), the Universities of Gießen and Marburg (Germany), San Matteo Hospital, Pavia (Italy), and Medical University of Graz (Austria). All participants or their parents provided written informed consent (UK Research Ethics Committee: 13/EE/0203). During their routine 6 months visit, PAH patients had additional blood and urine samples taken for research purposes. Relatives were followed up at annual intervals.

Cases were recruited for the national institute for health research (NIHR) BioResource – Rare Diseases (BR-RD) study from the same centres as the national cohort study of idiopathic and heritable PAH. All cases had a clinical diagnosis of idiopathic PAH, heritable PAH, drug- and toxin-associated PAH, or PVOD/PCH established by their expert centre. The non-PAH cohort for the case-control comparison were unrelated subjects recruited to the NIHR BR-RD study. All PAH and non- PAH patients or their parents provided written informed consent (UK Research Ethics Committee: 13/EE/0325), or local forms consenting to genetic testing in deceased patients and non-UK cases.

Blood and saliva samples were collected under written informed consent of the participants or their parents for use in gene identification studies (UK Research Ethics Committee: 08/H0802/32). Patient's notes were only viewed by the direct care team and named researchers. Subjects were given a unique study number and all data were linked anonymised (also known as pseudonymised) at study entry. Each centre had a master list which matched study number with the participant, only local study members had access to this. Authorized researchers had only access to pseudonymised (non-identifiable) data and biological samples. Linking anonymised data was necessary to provide the clinical care team with pertinent information that they could feed back to the participant. The local study team could request participant's contact information to enable participants to be contacted for future ethically approved studies. Contact information were held securely at the participant's local recruiting centre.

2.2.3 Informatics infrastructure and data security

Phenotype information were electronically collected for two separate studies, the national cohort study and the NIHR BR-RD study PAH cohort. Both studies collected the same information at the time of diagnosis and recruited mainly the same patients at separate occasions. Due to the overlap of information, the same infrastructure was used by both PAH projects to securely collect and retrieve phenotypes.

The Clinical School Computing Service (CSCS) was chosen to provide the support and computational resources to host a virtual server in a secure environment. Primary software installed on the server was Ubuntu, Apache HTTP, Apache Tomcat and PostgreSQL with default security settings only allowing HTTPS and secure shell (SSH) access. The server provided 4 CPUs, 8 GB of memory and 50 GB of expandable disk space. Direct public access to the virtual server was not possible and only enabled through the internal network using SSH. The CSCS firewall filtered and accepted web requests before internally forwarding the web traffic to the virtual server. All web traffic was protected by the encrypted hypertext transfer protocol secure (HTTPS) protocol implementing the transport layer security (TLS) 1.2 standard.

The phenotype information were captured by the web-based OpenClinica (OC) system. The open source package was deployed on a Apache Tomcat webserver with a PostgreSQL relational database management system (RDBMS) on the secure virtual machine. Information entered through web services were stored in the PostgreSQL database on the server, which required authentication. Researchers required approval from the data access committee before user accounts were issued and access was granted to the required data.

Approved research nurses, study coordinators and researchers were assigned OpenClinica accounts for user authentication. Data access and permissions of users were limited based on the assigned role and study site in OpenClinica.

2.2.4 Electronic phenotype collection

The established infrastructure collected phenotype data and catalogued biological sample from individuals, which participated in the PAH cohort, NIHR BR-RD or both studies. The open source software OpenClinica (OC) was chosen after evaluating the NHS computer restrictions and different electronic phenotype data capture systems. The OC web interface captured data based on electronic case report form (eCRF) definition specified in an Excel spreadsheet. The eCRF definition file controls the web form appearance, description, unit of the requested value and the validation of the patient data. Each row in the Excel eCRF describes one value field, also called item, in OC. Amongst other values, the descriptive text, type of value and range checks can be specified per item. Definitions are specific for each project and for PAH we developed 24 different eCRFs. Each eCRF captures a different category of data and were based on specifications defined by the PAH consortium.

2.2.5 Software release information

The data were collected and analyses were performed using the software products and versions listed in Tab. 2.1 unless stated otherwise.

2.2.6 Phenotype definition

The PAH consortium specified the clinical measurements in a Microsoft Word document and provided ranges for selected numerical values. The clinical values were grouped into 20 categories listed in Tab. 2.2. The definitions of each category were translated into an OC

Name	Version	Info
Apache HTTP	2.2.22	
Apache Tomcat	6.0.32	
JAVA	8u66	Java Development Kit (JDK)
KVM	1.0	QEMU emulator version
MySQL	5.1.73	
OpenClinica	3.3	OpenClinica community edition
PostgreSQL	8.4	
Python	3.4.1	
R	3.2.4	
Ubuntu	12.04.2 LTS	

Table 2.1 Software release versions used

eCRF Microsoft Excel definition file. Numerical range checks were defined as rules and the conditional displaying of items was used to show only required fields.

Categories	Description
Demographics	Basic characteristics
Body system	General measurements
Family history	Description of possible PAH cases in the family
Risk Factors	Drugs and toxin ingestion
Drug Treatment History (PAH)	Record of PAH related drug treatment
Drug Treatment History (other)	Record of drug treatments not related to PAH
Associated diseases	Record of diseases associated with PAH
Clinical feature by history	Clinical symptoms before diagnosis
Clinical feature by examination	Clinical symptoms at diagnosis
Functional class	New York Heart Association (NYHA) classification
Haemodynamics	Blood movement related measurements
Clinical blood tests	Measurements of the blood
Arterial blood gases	Arterial blood measurements
Exercise performance	Fitness measurement
Lung function	Assessment of the lung function
Echocardiography	Heart measurements from ultrasound
Electrocardiogram	Electrical activity of the heart
Imaging investigation	Cardiac magnetic resonance (CMR) data
Imaging Report	Report of the CMR imaging
Survival	Record of study participation

Table 2.2 Clinical categories specified by the PAH consortium.

2.2.7 OpenClinica SOAP API access

The programming language JAVA was used for the interaction with OC. The wsimport tool part of the JAVA development kit (JDK) was applied to generate JAVA API for XML web services (JAX-WS) portable artifacts. These artifacts were generated from the web service description language (WSDL) schema definition files provided by the OC simple object access protocol (SOAP) application programming interface (API). Separate artifacts were generated for the production and test OC instance.

2.2.8 Sample submission

Samples were sent to the NIHR Cambridge BioRepository for storage and required a sample manifest listing the expected sample and tube identifiers for verification. The JAVA application BioAPP was developed to extract the OC ID and visit number from the 'tracking log'. The OC SOAP API was used to download the sample log. The identifiers contained in the 'sample log' were validated before producing the 'sample manifest'.

2.2.9 Automated data import

The automated import JAVA application was developed to process clinical information stored in a predefined Excel file. The extracted values were submitted to OC using the SOAP API and a report was produced to identify failed entries.

2.2.10 Data export, version mapping, normalisation and validation

An automated file exports was scheduled using the OC DataSet export feature. All data items were selected for export and data were exported separately per event. The automated pipeline transferred the files to a release folder on an internal server and executed the load process. The load process was written in Python and loaded each file separately into the MySQL database. After loading the data, the version mapping was performed as a python script using the mapping instructions. The statistical computing language R extracted data from MySQL, converted units based on mapping instructions and stored the resulting values as R object. The standard deviation (SD) was calculated for numerical values and entries outside 2.5 SD were recorded to be manually validated (see 2.3.8 on page 32). For validation, the specific recording (patient, event, field, value) was sent to the respective centre to be confirmed based on the patient records or corrected. The values in the normalised R data object were validated based on an Excel file definition with the specific records. This Excel file contained previously detected outliers and either removed the entry for pending requests, replaced the entry with the corrected value or included confirmed values. There were 755 outliers entries for 415 subjects recorded in the Excel file affecting 104 different items, 133 of these entries have been checked by the centre and changed if required. Sample identifiers were validated for consistency and the cleaned data was stored as an `R` object ready for analysis.

2.3 Results

The national cohort study of idiopathic and heritable pulmonary arterial hypertension (PAH) also called PAH Cohort Study recruited affected PAH patient, relatives and unrelated controls across 10 participating NHS and and 4 international collaborating PAH specialist centres. Detailed clinical information were captured from patients at the time of diagnosis, during recruitment and additional measurements and samples were collected during each follow-up visit. Participating PAH centres also recruited affected PAH patient and relatives for the PAH cohort of the national institute for health research (NIHR) BioResource – Rare Diseases (BR-RD) study. Clinical information were collected from patient at the date of diagnosis as part of the NIHR BR-RD study, the same as for the PAH Cohort Study.

The NHS computer restrictions were evaluated for the integration with different electronic phenotype data capture systems. After the evaluation, I established the open source software OpenClinica (OC) to capture clinical data of PAH patients for both studies and developed an automated data release pipeline for analysis. The OC web interface captures data based on an electronic case report form (eCRF) definition specified in an Excel spreadsheet. Each eCRF was designed based on specifications defined by the PAH consortium and one eCRF was created for each data category. Clinical information were entered by research nurses using the web-browser based forms after authentication. In addition, OC provides an encrypted application programming interface (API) with authentication to enter and extract data electronically.

The initial test phase started in October 2013 and the production system went live in February 2014 to collect data. Until December 2016, I lead the development of 109 different versions of 24 eCRFs that contain 3,845 items and changed the capture behaviour to reduce the workload on the nurses. These changes reflect the dynamic nature of the project and OC provided the required flexibility to capture these information. The overview in Fig. 2.3 shows the different data entry workflows, which followed our standard operating procedures (SOP) described later.

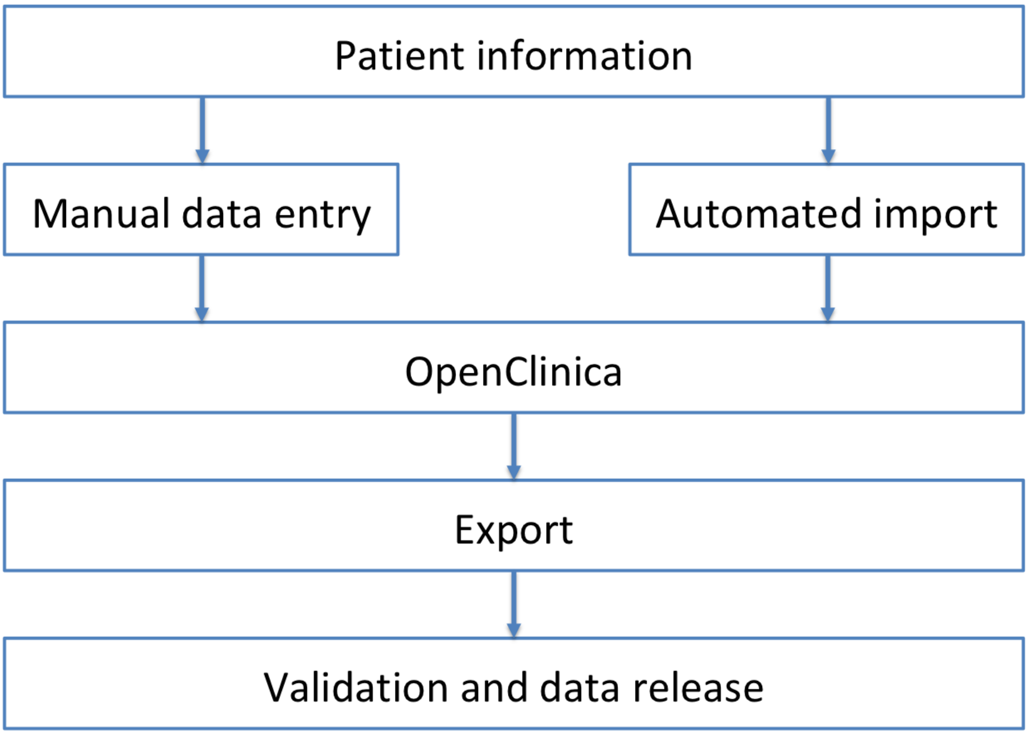
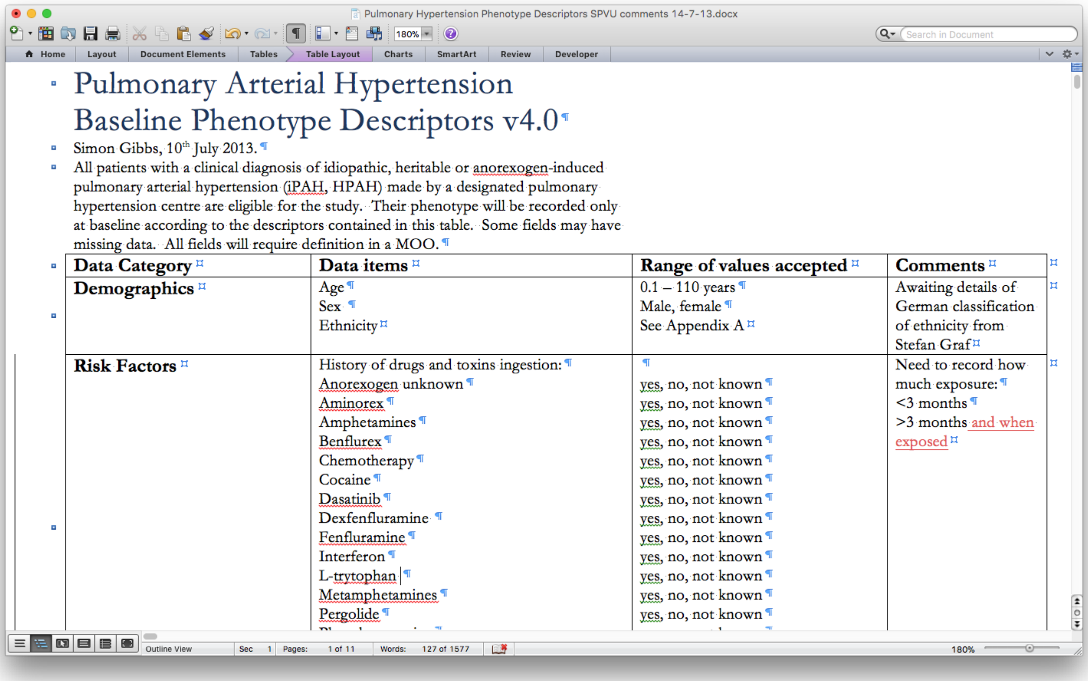


Figure 2.3 Patient information capture and automated validation workflow. OpenClinica supports manual as well as electronic data import. Data are automatically exported, validated and prepared for analysis.

2.3.1 Phenotype definition

We aimed to capture the baseline phenotype of PAH patients in OpenClinica as comprehensively as possible. The PAH consortium provided an itemised description of the baseline phenotype items including ranges and comments. Fig. 2.4 shows an extract of the specification document, which was captured in OC using eCRFs. I contributed to the manual translation of the descriptors to eCRF, which were organised based on 20 categories listed in methods. Study identifiers and sample tracking information were captured as additional eCRFs. An eCRF is specified in an Excel document and enables to group items into pages, sections and subsections with specific. The definitions are interpreted by OC and dynamically creates web forms for data entry. Entered values are validated during the submission process and unexpected records can be blocked, produce warnings or trigger predefined actions to document discrepancies. The same validation rules apply to electronically transferred data.



Pulmonary Arterial Hypertension
Baseline Phenotype Descriptors v4.0

Simon Gibbs, 10th July 2013.

All patients with a clinical diagnosis of idiopathic, heritable or anorexogen-induced pulmonary arterial hypertension (iPAH, HPAH) made by a designated pulmonary hypertension centre are eligible for the study. Their phenotype will be recorded only at baseline according to the descriptors contained in this table. Some fields may have missing data. All fields will require definition in a MOO.

Data Category	Data items	Range of values accepted	Comments
Demographics	Age Sex Ethnicity	0.1 – 110 years Male, female See Appendix A	Awaiting details of German classification of ethnicity from Stefan Graf
Risk Factors	History of drugs and toxins ingestion: Anorexogen unknown Aminorex Amphetamines Benflurex Chemotherapy Cocaine Dasatinib Dexfenfluramine Fenfluramine Interferon L-trytophan Metamphetamines Pergolide	 yes, no, not known yes, no, not known yes, no, not known yes, no, not known yes, no, not known yes, no, not known yes, no, not known yes, no, not known yes, no, not known yes, no, not known yes, no, not known yes, no, not known yes, no, not known	Need to record how much exposure: <3 months >3 months and when exposed

Figure 2.4 Screenshot of agreed phenotype definition document partly showing two categories with data items, value ranges / options and additional comments.

2.3.2 Event definition

Clinical measurements were defined as eCRFs and different types of information are collected in separate CRFs. Event provide the possibility to group CRFs, that collected during the same visit or for different groups of patients. The event definition defines the default CRF version included in the event. The same CRFs are used to capture patient, relative and unrelated control information in different events and allows the transition from a relative to become a patient (see Fig. 2.5). Each CRF in an event is marked as completed after entering the data. The exception is the Continuous data event that collects continuously updated values. Such updated values include prescribed drugs with their amount, start and end date, if it applies. Tab. 2.3 provides a complete list of CRFs and their use in different events.

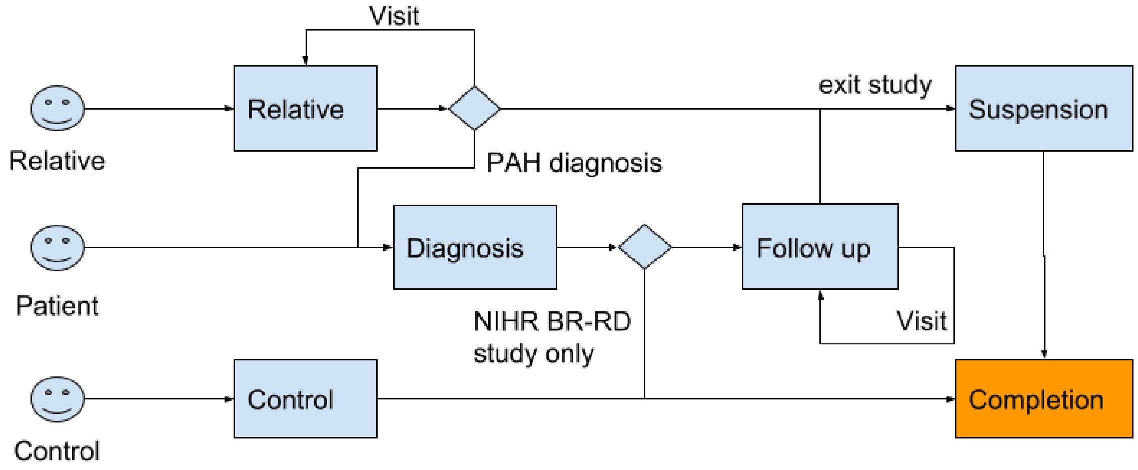


Figure 2.5 Subject specific workflows. A subject can participate as a ‘Relative’, ‘Patient’ or ‘Control’. Data from a ‘Relative’ are entered into the `Relative` event for every visit until the subject exits the study through `Suspension` or through the transition to a ‘Patient’ due to a diagnosis of PAH. For a ‘Patient’, `Diagnosis` information are captured at the date of diagnosis. After the `Diagnosis` entry, a NIHR BR-RD study ‘Patient’ completes the data entry. Data from a ‘Patient’ recruited to the PAH cohort study continue to be entered in the `Follow up` event by 6-monthly visits until the exit from the study through `Suspension`. ‘Control’ subjects date are entered for the `Control` event only.

2.3.3 Subject and sample identification

Both the PAH cohort study and NIHR BR-RD PAH cohort used OC to record phenotype information and identifiers. In order to upload information, each subject in the study received an OC subject study ID (OC ID). The OC ID was study independent and bridged between different studies and visits. In addition, subject were provided with study specific identifiers. These identifiers were registered in OC in order to track subjects. A subject taking part in both studies had an OC ID, PAH cohort ID and an NIHR BR-RD ID assigned. All these identifiers are printed as barcode labels and attached to the physical patient record at the

CRF	Diagnosis	Cohort study entry / Follow up	Continuous data	Suspension	Relative	Control
ID capture	X	X			X	
Demographics	X	X			X	X
Clinical features by history	X	X				
Associated Diseases with PAH			X			
Family history			X			
Cohort relatives			X			
Functional class	X	X			X	
Comorbidities						X
Clinical features by history			X			
Clinical features by examination	X	X			X	
Body system	X	X			X	
Haemodynamics	X	X			X	
Exercise performance	X	X			X	
Lung function	X	X			X	
Arterial blood gases	X	X				
Echocardiography	X	X			X	
Electrocardiogram	X	X			X	
Imaging Investigation	X	X				
Imaging Report	X	X				
Clinical blood tests	X	X			X	
Drug treatment history (PAH)			X			X
Drug treatment history (other)			X			X
Risk Factors			X			
Suspension				X	X	
Sample collection / Sample log		X			X	
Relative Symptoms					X	

Table 2.3 Description of CRF components for each event. Events are composed of a combination of CRFs and a CRF can be included in multiple events. This matrix lists the events on the top, the available CRFs on the left and the **X**s at the crossing points between them highlight the inclusion of CRFs in studies.

relevant NHS trust. For the PAH cohort study, the research nurse collected samples from subjects at each cohort visit. These samples were processed and stored in barcoded tubes. The tube types were sample specific. All barcodes were scanned and recorded in a sample log with the OC ID and PAH cohort ID. The completed document was uploaded to the eCRF in OpenClinica. I contributed to the tracking of subject identifiers, tubes and capture of phenotype information. Study and tube identifiers were printed as barcodes or had pre-printed quick response (QR) codes, which should be scanned with the provided barcode scanners. The generated barcodes had a character appended at the end to check the consistency of the identifier. This consistency check was independent of other resources and my sample submission application described below validated the generated identifiers for consistency during submission. Identifiers failing the consistency checks were reported. Incorrect identifiers in sample log files required correction by the relevant centres and resubmission to OpenClinica.

2.3.4 Sample submission

All participating centres followed a standard operating procedure, which we developed for the sample submission to the NIHR Cambridge BioRepository. I developed an application (BioAPP) that automates the validation and transformation to the agreed exchange format. Research nurses completed a tracking log for a delivery, which contains the OC ID and visit number of a subject. The BioAPP read in the 'tracking log', extracted the corresponding 'sample log' and tube id for each visit from OpenClinica. The extracted identifiers were validated and produced the 'sample manifest'. The 'sample manifest' contained the PAH cohort id, tube id, sample type, volume, volume unit, collection date, centre, visit and if the subject was a patient, relative or unrelated control. This 'sample manifest' was sent to the NIHR Cambridge BioRepository, imported into their computer system and used for validation on arrival of the samples.

2.3.5 Automated data import

Imperial College and Hammersmith Hospital provided the largest collection of historical PAH samples for the NIHR BR-RD study. The hospital has electronic health records for the enrolled individuals. The records closest to the date of diagnosis was exported. Although OpenClinica provided SOAP API for automated data submission, the submitted file required mapping to OC specific item identifiers, also called unique OID.

We agreed on an Excel file based data exchange format to allow for human intervention and correction. I wrote the software tool and defined the structure of the Excel file. The file structure allowed the inclusion of multiple eCRFs in one Excel file and further described in Tab. 2.4. The Excel template was generated from the eCRF information stored in OC.

				Section name
Number	Openclinica ID	Bridge Id	Cohort Id	Description
OC	oc_study_subject_id			Variable
1	OC001A			
2	OC002B			

Table 2.4 Excel based OC data exchange format. Each sheet in the excel file represents one eCRF identified by the sheet name. In each sheet, individuals are stored in rows and each column represents a captured measurement. The first two rows are ignored and used for description. The third row starts with 'OC' to comply with the exchange format followed item OIDs for each column to identify the measurement. In addition, the key word 'oc_study_subject_id' is used to specify the column with the OC ID. From the fourth row onwards, the clinical measurements are recorded.

To ensure consistent data validation, the import tool used the web-service interface provided by OpenClinica. The same checks were applied for data submitted through the web-page or the web-service. The only difference was that no history is stored about changes to item values during the automated import. For this reason, the Excel file was the main source of information and an import overwrites existing values in OpenClinica.

I generated Excel templates for the required eCRFs and these were used to manually integrate and curate data. The developed tool imported the majority of available baseline data for Imperial College and Hammersmith Hospital. The developed tool was also vital for restructuring eCRFs and moving eCRFs to different events. In these cases, phenotype data from the MySQL DB described below were exported into Excel on the latest eCRF version and reimported into a different event.

2.3.6 Data export

OpenClinica is focused on capturing phenotype information and does not provide an open source mechanism to analyse the captured information. The export facility provided by OC created a file in extensible markup language (XML) format compliant with the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM) standard Version 1.3 (see http://www.cdisc.org/odm). In addition, OC extended their schema to include annotations about discrepancy notes. Each XML file provides a complete description of the field definitions and the entered data from the specified time frame.

I developed the MySQL database schema shown in Fig. 2.6 and a command line tool to automate the import of the XML structure in a secure environment. Access to the data

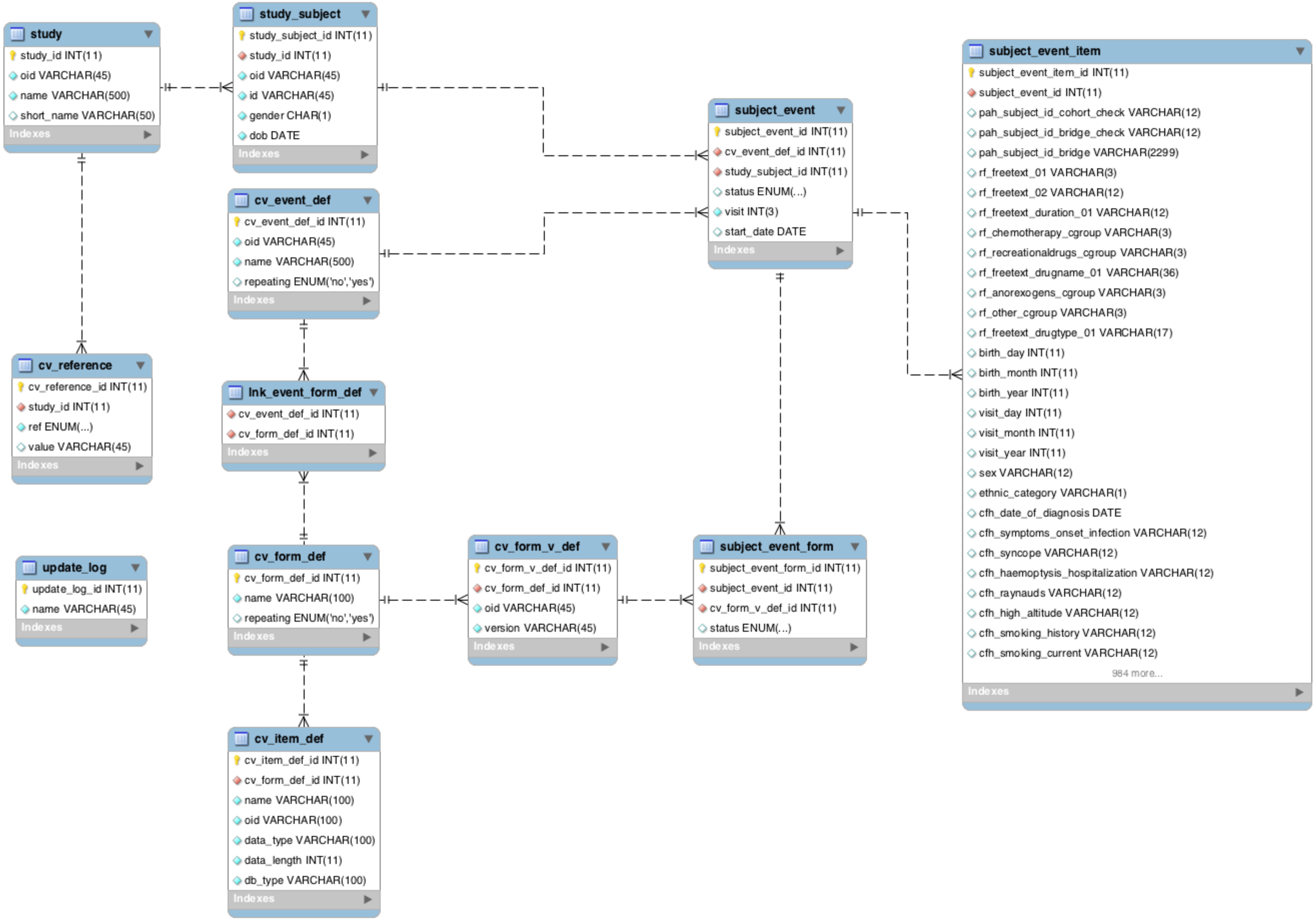


Figure 2.6 Entity relationship model of the phenotype export database. The database schema is designed to store, clean and query the exported data from OpenClinica. Each box in the schema represents a table in the database and the lines between the boxes describe their relationship.

stored in the secure environment required a successful application granted by the data access committee. In the database schema, a study or centre (*study*) has a study subject (*study_subject*) which can attend several subject events (*subject_event*). The captured measurements for a subject event are stored as items in *subject_event_item* and the status of each eCRF is stored in *subject_event_form*. A list of possible events is defined in *cv_event_def*. The form definition (*cv_form_def*) represent the eCRFs and has a list of items *cv_item_def* and form versions (*cv_form_v_def*) defined. The link table *lnk_event_form_def* describes the relationship between eCRFs and events since one eCRF can be used in different events. In addition the schema stores generated identifiers (*cv_reference*) for each study site as well as time stamps of the last schema update (*update_log*).

The tool created the table schema and loaded the following meta-information from the XML file:

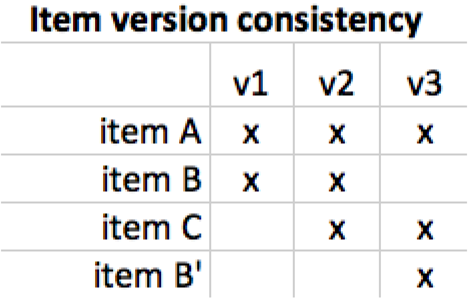
- Study definitions
- Case Report Form definitions
- Item definition
- Event definition
- Study Subject (identifiers and labels)
- Subject Event (an ‘Event’ created for a ‘Subject’)
- Subject Event Form (a ‘Form’ filled or started to be filled in for a ‘Subject Event’)

After loading the meta-information, the values for the items were loaded.

2.3.7 Version mapping

As described in Phenotype definition, specified descriptors were translated into eCRFs for OC. After the test and first production release, we reviewed the feedback from OC users. The collected feedback was incorporated during several iterations of eCRF changes and event restructuring. We focused to improve the usability, reduce the number of mouse clicks, remove ambiguity and add conditional displaying of fields. The process of corrections involved the removal as well as addition of units, fields and reduction of data range checks.

During the implementation process we realised that capturing phenotype data was a dynamic process. The continuous evolution of eCRFs required a robust way to map captured data forward to the latest version. The consistency of items in one eCRF across versions is demonstrated in Fig. 2.7. It also highlights that columns used in older versions can be missing in the latest version of the same eCRF for various reasons. The version processor module was added to the automated processing pipeline. The process mapped otherwise obsolete columns to columns renamed in the latest version. The module processes a tab delimited file



Item version consistency

	v1	v2	v3
item A	x	x	x
item B	x	x	
item C		x	x
item B'			x

Figure 2.7 Example of item consistency and changes across versions in OC. Item A and C is consistent across all and across the last two versions respectively, while item B is renamed in v3 to item B'.

of instructions and acts on the operations listed in Fig. 2.8(a). The file consists of a header line followed by one item for each line. The tool allows for the following columns:

- **Operation**: The operation to perform
- **Name**: The item name to perform the operation on
- **Target**: The item name to map a field to (if different from target)
- **Table**: The table name for the target (for list items) described later

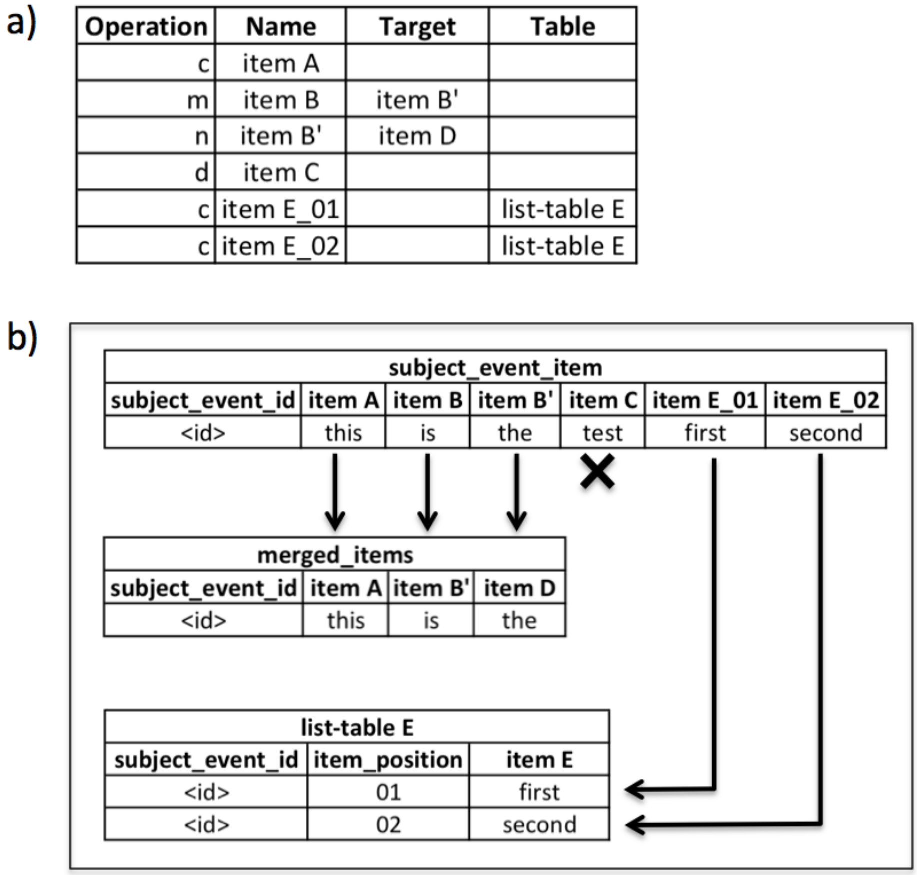
The permitted operations are:

- **copy** (c): No version change
- **map** (m): Update the item name to the latest item version name.
- **new** (n): No compatible version available and store as different name.
- **delete** (d): Removed in the latest version

A special case was required for the last two items shown in Fig. 2.8(a) and were called list items. These items are used to store a collection of values of the same type e.g. list of prescribed drugs in OC. The list item names are constructed as <prefix>_<suffix> using the same prefix with different suffixes (Fig. 2.8(a)). This naming convention allowed to extract list values into a separate table as shown in Fig. 2.8(b). Before the version mapping 1,014 columns (column without values are not created) were represented in the exported data set. After the version mapping it was reduced to 393 columns and 11 list tables with all items from the last eCRF versions represented. These tables represent a comprehensive collection of the data in OpenClinica.

2.3.8 Data normalisation, correction and cleaning

Research nurses in participating centres collected and entered phenotype information into OpenClinica, which were then exported and mapped to the same version as described above.



Operation	Name	Target	Table
c	item A		
m	item B	item B'	
n	item B'	item D	
d	item C		
c	item E_01		list-table E
c	item E_02		list-table E

subject_event_item

subject_event_id	item A	item B	item B'	item C	item E_01	item E_02
<id>	this	is	the	test	first	second

merged_items

subject_event_id	item A	item B'	item D
<id>	this	is	the

list-table E

subject_event_id	item_position	item E
<id>	01	first
<id>	02	second

Figure 2.8 Version mapping process. The (a) version mapping format consists of a header line followed by one item per line. The table lists the **copy** (c), **map** (m), **new** (n) and **delete** (d) operations in the Operation column. The ‘Name’ column contains the item name with ‘Target’ and ‘Table’ column filled depending on the operation. The mapping is applied to the table (b) *subject_event* item and two new tables *merged_items* and *list-table E* are created with the mapped values.

The recorded data per item were then available, but the values for one item are in different units and contain outliers, invalid and inconsistent values. These issues were addressed with the normalisation, correction, cleaning and validation before a cohort wide comparison can be done.

The normalisation process converted the measurements of each item into one unit per item. The developed `R` script identified and converted values with different units to the same unit. An item mapping file specified the converted functions between units. After normalisation, the script calculated the distribution of each numeric item and recorded entries with 2 and 2.5 standard deviation (SD) from the mean value to assess potential outliers. We assessed such outliers and recorded likely incorrect values in a centrally stored verification Excel file. Outliers were assessed by a clinician for their feasibility and for their consistence with other values of the overall disease status. Incorrect entered values included a diastolic pulmonary artery pressure (dPAP) of 82 mmHg above a mean pulmonary arterial pressure (mPAP) measure of 67 mmHg or a physiological near impossible mean right atrial pressure (mRAP) of 54 mmHg. The 2.5 SD ranges were 5.7 - 62.6 and -12.1 - 31.2 mmHg for dPAP and mRAP respectively. After checking with the local centres, research nurses corrected the 35 and 10 mmHg for dPAP and mRAP respectively. These entries were sent to the recruitment centre and the value either was confirmed or corrected by the research nurse. Until the confirmation or correction, the verification file allowed to correct or remove the identified values during a verification step. The verification also removes individuals with invalid, duplicate or inconsistent identifier (NIHR BR-RD, PAH cohort or OC ID). The resulting phenotype information were a high confident data set for further in-depth analysis.

Specific OpenClinica eCRF data were extracted from the MySQL database, normalised, validated and stored as an `R` object. All these steps from OpenClinica to the validated `R` object were automated and produced on a weekly bases. Email alerts were distributed to coordinators to flag entries for validation. On 1st of January 2017, the data release contained 351 fields for 1,023 individuals that passed the validation. In addition, phenotype information for 128 subjects were available from collaborators as Excel documents, but not yet transferred into OpenClinica. These additional measurements were combined with the data release to increase the total number of available samples. The combined release resulted in 1,151 individuals and 351 items.

2.3.9 Assessment of clinical data

Capturing phenotype data is a continuous process and the incompleteness of data has to be monitored. For the ability to assess missing data, we developed a visualisation method to inspect and detect pattern of completion. Fig. 2.9 and 2.10 show the status of entered and missing values from the diagnoses event for all patients in an unbiased way. Based on the visualised matrix, we were able to identify items consistently filled or empty for subjects across all centres. For example, the Imperial and Hammersmith Hospital completed data for all patients per eCRF, which is indicated by large gray or blue areas for whole eCRFs across subjects. The matrix was an efficient way to give feedback to centres regarding the data entry progress.

The validated phenotype information contained 1,151 subjects and 351 items, which included clinical measurement used to diagnose PAH patients. In the opinion of the clinicians at each specialist centre the main clinical diagnoses was pulmonary arterial hypertension. By definition, diagnosed PAH patients have a mean pulmonary artery pressure (mPAP) of $\geq$ 25 mmHG at rest, pulmonary artery wedge pressure (PAWP) of $\leq$ 15 mmHG and left ventricular end-diastolic pressure (LVEDP) of $\leq$ 12 mmHG (see chapter 2.1). These values were captured in the Haemodynamics eCRF and we used them to confirm the diagnoses of the subjects (see Fig. 2.11, 2.12 and 2.13). For 1,151, the completeness was 99.5% (n=1145), 92.6% (n=1066), 73.1% (n=842) and 19.7% (n=227) for gender, mPAP, PAWP and LVEDP respectively. Subjects outside the defined diagnostic threshold were 0.9% (n=11), 3.5% (n=40) and 11.2% (n=129) for mPAP, PAWP and LVEDP respectively. It is possible for the PAWP and LVEDP to be elevated in severe PAH with fluid overload from right heart failure and compression of the left ventricular (LV) by an overloaded right ventricular (RV). Subsequent right heart catheter after diuresis may have shown normal PAWP and LVEDP despite the elevated measurement at the time of diagnosis. All elevated PAWPs and LVEDPs were queried with the centres to confirm the diagnosis and rule out incorrect values. Later data releases included updated values (corrected where needed) from the currently displayed elevated entries. The PAWP values $\geq$ 40 mmHg from Sheffield were outliers and were corrected by the centres. In cases of severe PAH, LVEDP values are possible to vary. The observed gender ratio was 2.1 (779 female : 366 male), which is in line with the published literature.

A visual comparison of the mPAP measurements identified centre specific trends, most noticeable “Great Ormond Street” and “Lincoln” with the lowest median mPAP values compared to other centres 2.11. All recruited cases in both of these centres were below the age of 16 at diagnosis with a mean age of 6 compared to 50 for the remaining centres. In order to elucidate inter-centre differences, I applied a linear regression approach to model the relationship between (**A**) mPAP and centre, (**B**) mPAP and age at diagnosis as well as (**C**) mPAP and centre with age at diagnosis. Analysis of variance (ANOVA) separately compared the models **A**, **B** and **C**. The calculated p-values comparing the similarity of the groups **A** with **C**, **B** with **C** and **A** with **B** were 2.2E-16, 4.9E-12 and 0.9988 respectively. The result

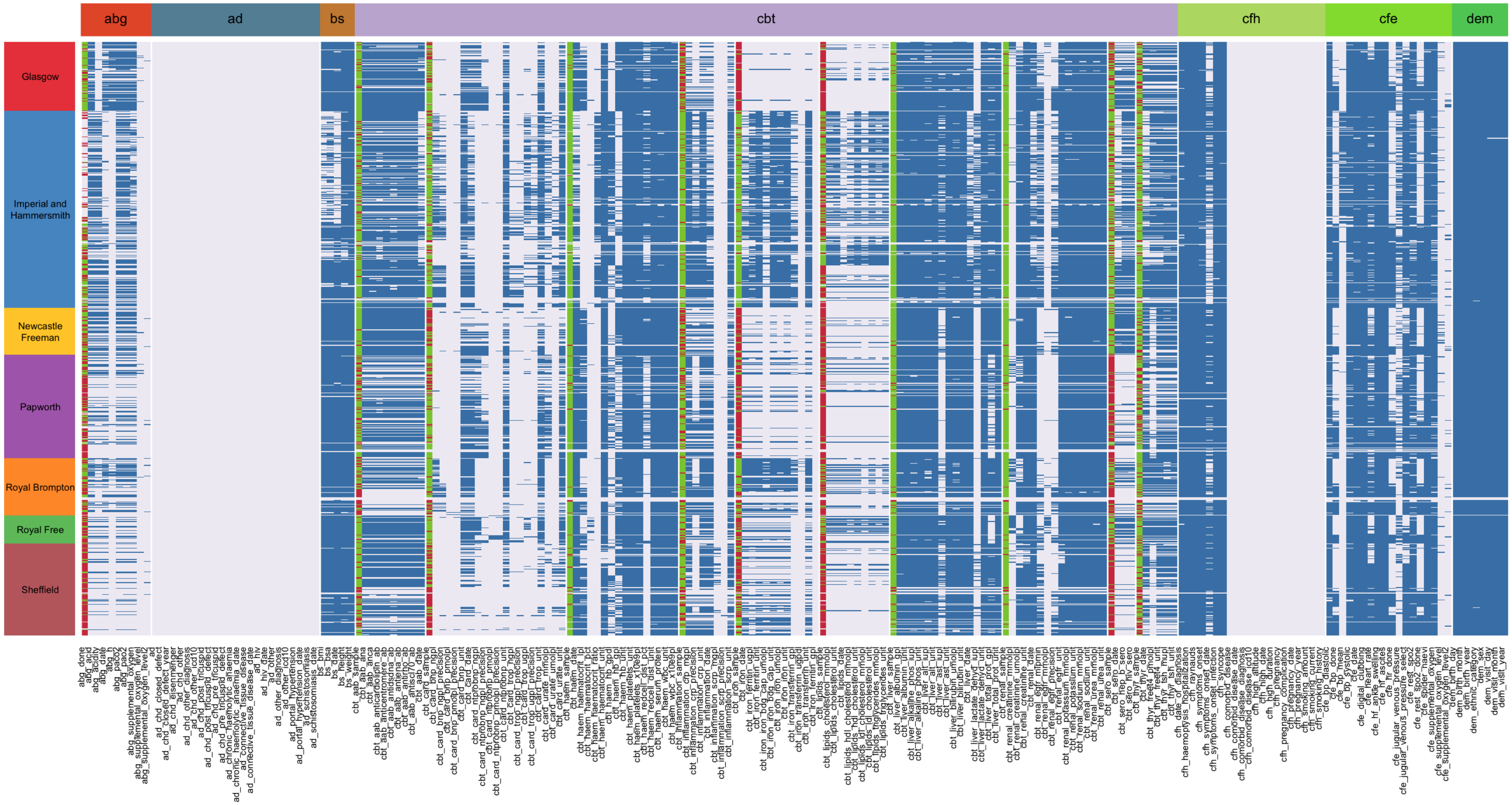


Figure 2.9 Completeness matrix (part 1) of diagnoses data per centre [Release 10th January 1018] Rows represent subjects grouped by centre and columns are items grouped by eCRF. Empty items are coloured as grey and entered values are colour in red, green and blue. The value of some items control if further items require completion. The colour of these control items indicate require (green) or not required (red) fields. The eCRF groups stand for: abg (Arterial blood gases); ad (Associated diseases); bs (Body System); cbt (Clinical blood tests); cfh (Clinical features by history); cfe (Clinical features by examination); dem (Demographics);

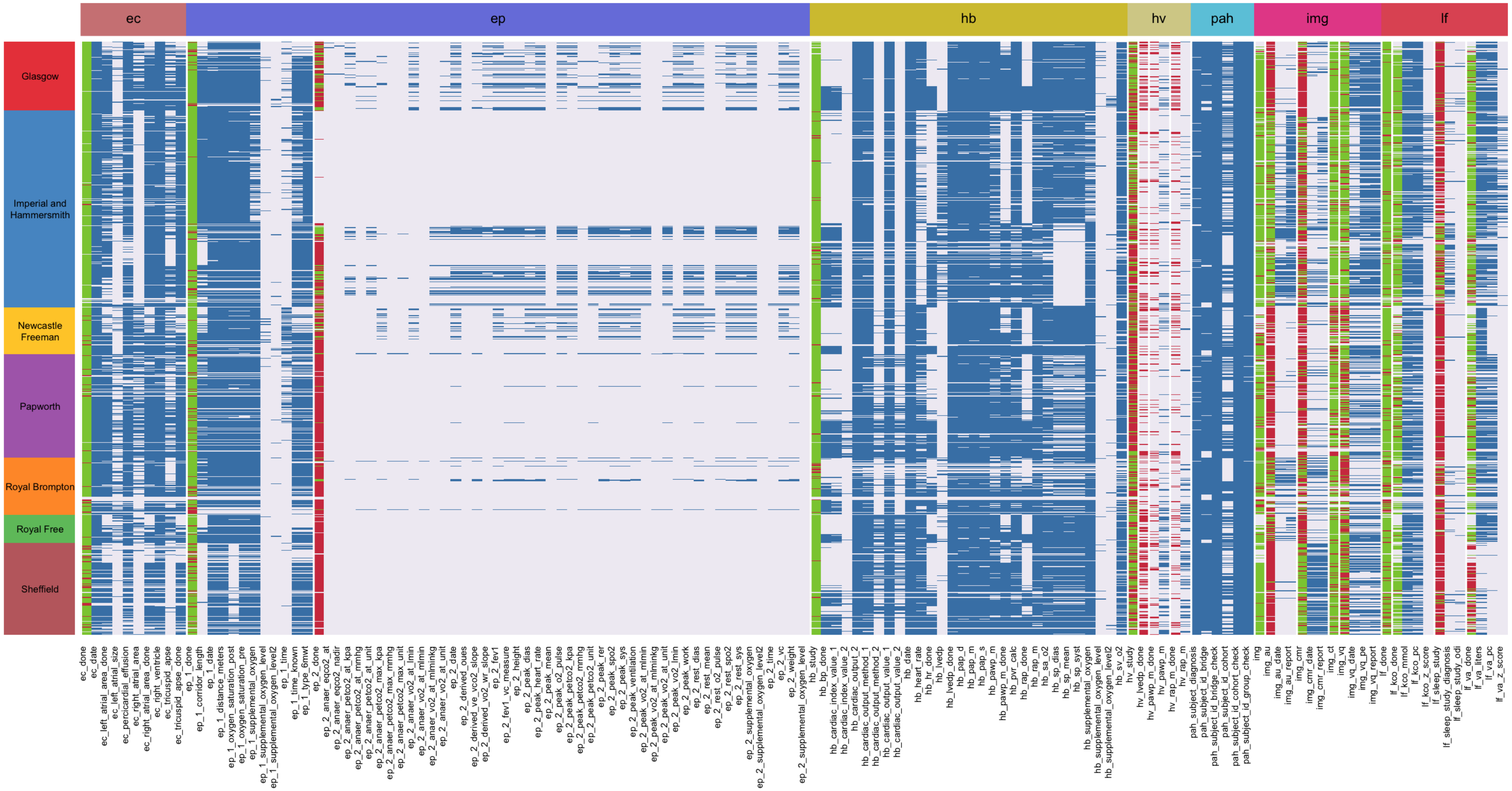


Figure 2.10 Completeness matrix (part 2) of diagnoses data per centre [Release 10th January 1018] Rows represent subjects grouped by centre and columns are items grouped by eCRF. Empty items are coloured as grey and entered values are colour in red, green and blue. The value of some items control if further items require completion. The colour of these control items indicate require (green) or not required (red) fields. The eCRF groups stand for: ec (Electrocardiogram); ep (Exercise performance); hb (Haemodynamics); hv (Haemodynamics vasodilator study); pah (Study identifiers and status); img (Imaging); lf (Lung function);

suggests, that adding age as a variable to centre is a more significant change than adding the variable centre to age. There was no significant difference between the regression models using age or centre as a variable, which can be explained by centres specialised in pediatric cases. After excluding pediatric cases (age at diagnosis <16), I rerun the analyses. The recalculated p-values were 2.2E-16 and 0.003 comparing **A** with **C** and **B** with **C** respectively. No p-value was retrievable comparing **A** with **B**. Adding age in addition to centre as a variable to the model was still similarly significant for adults, while adding the centre in addition to age was not as significant compared to the previous analysis including children. An exploration of the relationship between (**A**) mPAP and centre with age in adult cases found a difference (p-value of 0.014) for the centre "University of Giessen". The mean age at diagnosis in "University of Giessen" was 52 and elevated compared to 49 in the other centres for adult cases only.

Differences in mPAP value distribution were found between centres and explained by the age at diagnosis due the recruitment of only pediatric cases by two centres. The remaining difference for "University of Giessen" in the adult only dataset could not be fully explained, but age remains an important variable due to the elevated age at diagnosis compared to other centres.

2.4 Discussion

We captured clinical and related phenotypic information from subjects participating in the PAH national cohort study, NIHR BR-RD study or both. The longitudinal PAH national cohort study recruited patients and relatives for 6 month follow-up visits, while the NIHR BR-RD PAH project focused on capturing information at the time of diagnosis for affected PAH patients. Research nurses record phenotype data from participating subjects in the web-based OpenClinica phenotype capture system and an automated pipeline produces weekly data releases. While the defined number of fields is very extensive, captured information is sparse and a challenge for further analyses.

2.4.1 Data completeness

The hundreds of defined fields includes a core set of measurement for cohort characterisation, classification and diagnosis of PAH (see 3.3.3 on page 87). Completeness for this core set of information was mainly below 90% and in case of LVEDP was 19.7% (see 2.3.9 on page 34). Some values can be explained by historical samples, which did not record or collect these information at the time or patients unable to undergo the procedure. Visualising the

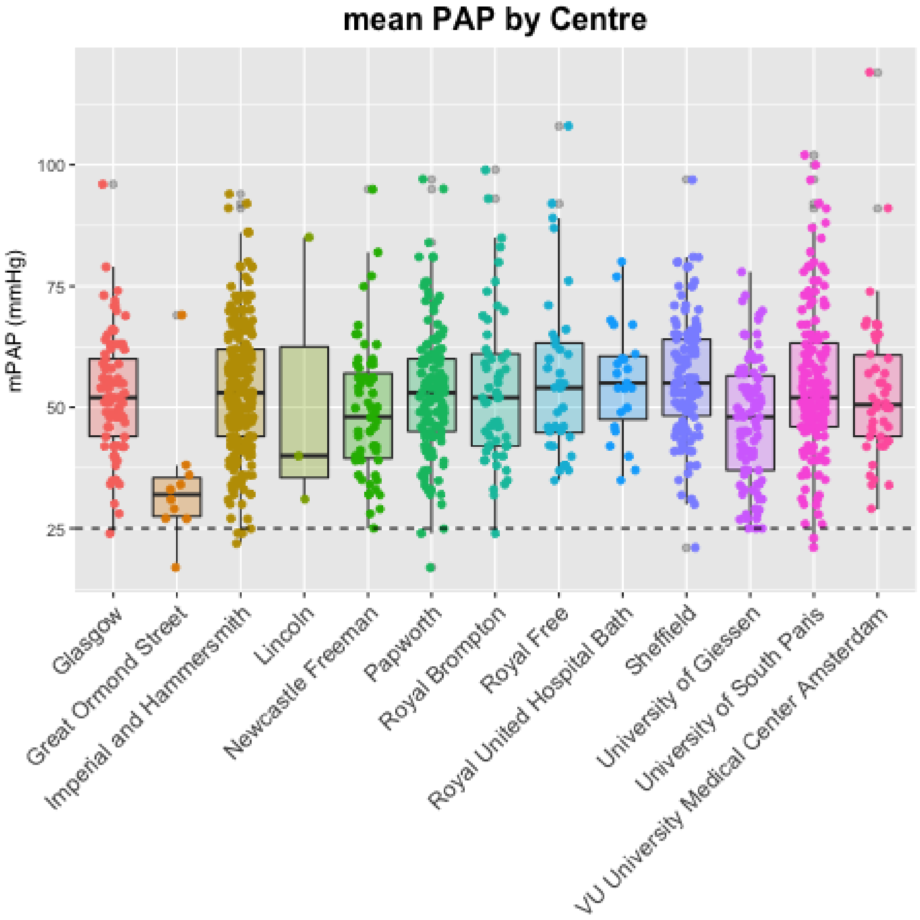


Figure 2.11 Distribution of mPAP values per centre. PAH patients should be above the horizontal line, which indicates 25 mm Hg. Differences in distribution for “Great Ormond Street” and “Lincoln” were explained by the age at diagnosis due to the large proportion of pediatric (<16) cases. Elevated age at diagnosis were found for “University of Giessen”, but does not fully explain the differences of mPAP distribution compared to other centres.

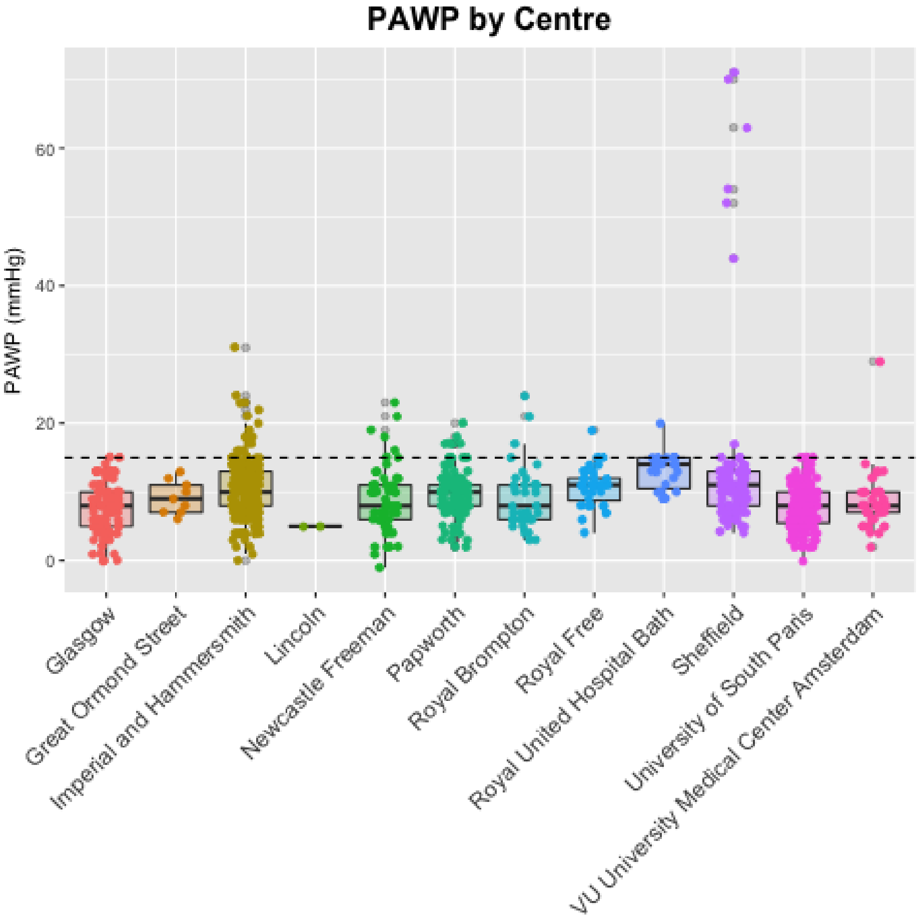


Figure 2.12 Distribution of PAWP measurements per centre. PAH patients should be below the horizontal line of 15 mm Hg, but elevated values in severe PAH are possible. PAH values $\geq$ 40 mmHg from Sheffield were found to be outliers and were corrected by the centre for later data releases.

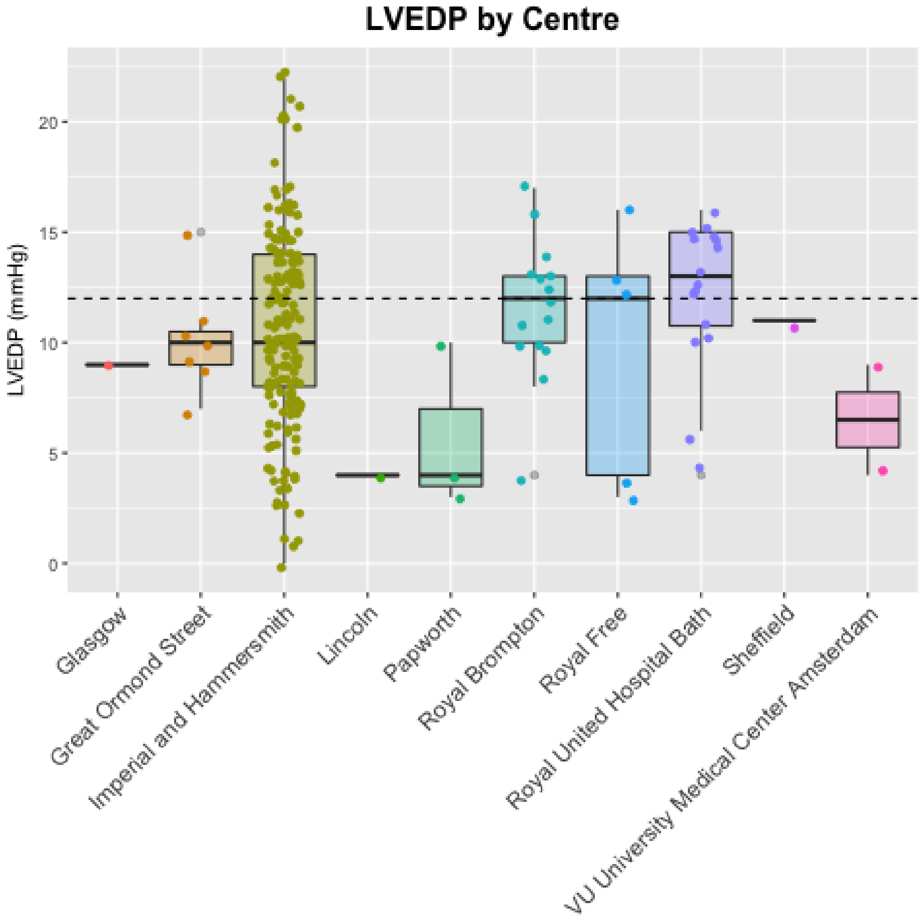


Figure 2.13 Distribution of LVEDP measurements per centre. PAH patients should be below the horizontal line of 12 mm Hg, but variable values in severe PAH are possible. Elevated values were queried with the individual centres for validation of the diagnosis and to rule out incorrect values.

completed values (see Fig. 2.9 on page 36) revealed distinctive patterns, where some values are collected by all or only by a few centres. Even the list of measured values captured during a blood test differ between centres. A standardised full blood count would benefit the analysis of the cohort, where the samples are tested centrally, capturing the same measurements and the data are transferred electronically. Central sample processing would ensure consistent and complete measurements from the same machine for all samples encoded in the same unit. Computational data transfer of these measurements would free-up time of local research nurses and reduce the risk of typos or entering the results of the wrong subject. However, the current dataset is the largest phenotype collection of PAH patients so far, which needs further work to utilise the data to its full potential.

2.4.2 Personal trends

The PAH cohort study is an observational study, which collects phenotype data during routine visits at 6 monthly intervals. Collected clinical measurements collected during the visit are dependent on the condition of a patient on that day and are not able to take short-term fluctuations into account. The clinical environment and the travel to the clinic could further skew results. Introducing mobile technology, including smart devices, would allow to continuously monitor patients in their own environment. The increased frequency of these measurements would enable to create an accurate personalised profile to follow the disease progression. Current methods focus on cohort wide analysis at specific time points, but modern technology could provide further insight and establish further subgroups or disease stages of PAH for precision medicine.

Chapter 3

Discovery of novel disease genes in PAH

3.1 Introduction

The general introduction (see chapter 1.1.3 on page 4) provided an overview of the application of whole genome sequencing (WGS) and discussed the contribution of rare genetic variation to human traits. The rare diseases pulmonary arterial hypertension (PAH) was described in chapter 2 on page 15 and defined the characteristics of the cardiovascular system. In this chapter, I discuss the currently known genetic causes of the PAH subgroups idiopathic PAH (IPAH) and heritable PAH (HPAH) before focusing on the genetic comparison of PAH patients to non-PAH controls using next generation sequencing (NGS) technology.

3.1.1 Genetics of Pulmonary Arterial Hypertension

The rare disease PAH is a form of high blood pressure in the lung and with an estimated prevalence of 10 to 52 cases per million per year (Escribano-Subias et al., 2012; Frost et al., 2011; Humbert et al., 2006; Peacock et al., 2007). PAH is part of the pulmonary hypertension (PH) disease group and can be further divided into the four subgroups idiopathic PAH (IPAH), heritable PAH (HPAH), drug/toxin induced PAH and PAH with associated diseases (Simonneau et al., 2013). HPAH was previously known as familial PAH (FPAH) and defines patients with a known family history of PAH. Collaborative effort on HPAH, the rarer form of PAH, identified chromosome 2q31-q32 as important locus through linkage analysis of genotype data across the autosomes (Morse et al., 1997; Nichols et al., 1997). Further investigation of the locus by targeted sequencing of exons revealed mutations predicted to cause premature termination of *BMPR2* (Deng et al., 2000; International PPH Consortium et al., 2000). Although *BMPR2* is inherited as an autosomal dominant trait, family studies found unaffected parents with a variant in *BMPR2*, which suggested incomplete penetrance (Thomson et al., 2000). Following the revelation of *BMPR2* in the heritable form, other forms of pulmonary hypertension (PH) were investigated. Targeted sequencing of the *BMPR2* gene identified variants in IPAH cases that were predicted to alter the protein function or cause premature termination (Cogan et al., 2006; Koehler et al., 2004; Sztrymf et al., 2008). Disease causing *BMPR2* variants were also described in patients with pulmonary veno-occlusive disease (PVOD), a rare form of PH (Montani et al., 2008; Runo et al., 2003). Identified cases of *BMPR2* variants explained 69% of HPAH and 21% of IPAH cases and the latest report collated 384 distinct variants across the gene (Aldred et al., 2006; Machado et al., 2015). In contrast, HPAH and IPAH represent 4.2% HPAH and 38.9% of PAH prevalent cases (Humbert et al., 2006). This highlights that IPAH is more frequent with less cases explained by *BMPR2* compared to HPAH.

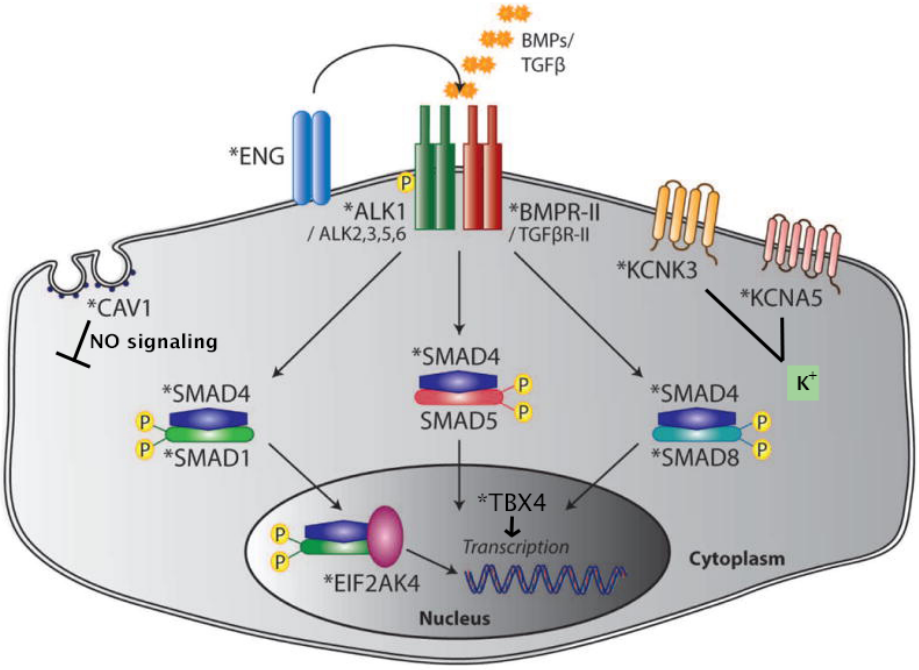


Figure 3.1 Genes and pathways implicated in PAH pathogenesis. Implicated genes involved in TGF-β/bone morphogenetic protein (BMP) signalling pathway are *BMPR2*, *ALK1*, *END*, *SMAD1*, *SMAD4*, *SMAD8* and *TBX4*. Other genes and pathways contributing to PAH: *EIF2AK4* part of angiogenesis-regulating gene in response to cellular stress; *CAV1* as part of the NO signalling and oxidative stress pathway; *KCNK3* and *KCNA5* are potassium channel-related genes; (*) indicate implicated genes. *SMAD8* and *ALK1* are now known as *SMAD9* and *ACVRL1* respectively. Figure adapted from Machado et al. (2015)

Here, I discuss four different potential pathways to causation (TGF-β/bone morphogenetic protein (BMP), cellular stress response, nitric oxide (NO) and potassium channel) for 11 different genes, the analysis strategy and their supporting evidence. The seven TGF-

β/BMP genes implicated in PAH are *BMPR2*, activin-receptor–like kinase 1 (*ACVRL1* or *ALK1*), endoglin (*ENG*), the mothers against decapentaplegic 1 (*SMAD1*), *SMAD4*, *SMAD9* and T-box 4 (*TBX4*) (Attisano et al., 1993). The disease causing gene *BMPR2* was first discovered through linkage analysis followed by targeted sequencing (Deng et al., 2000; International PPH Consortium et al., 2000; Morse et al., 1997; Nichols et al., 1997). The same strategy identified *ACVRL1* and *ENG* mutations in patients with HPAH and hereditary haemorrhagic telangiectasia (HHT), a PAH associated disease (Chaouat et al., 2004; Johnson et al., 1996; Trembath et al., 2001). Follow-up studies identified in total 66 mutations, 57 in *ACVRL1* and 9 in *ENG* (Abdalla et al., 2004; Best et al., 2011; Chaouat et al., 2004; Chen et al., 2013; Chida et al., 2012; Eyries et al., 2012; Fujiwara et al., 2008; Girerd et al., 2010; Harrison et al., 2003; Ishiwata et al., 2014; Jones et al., 2014; Machado et al., 2009, 2015; Mache et al., 2008; Pfarr et al., 2013; Smoot et al., 2009; Trembath et al., 2001). *ACVRL1* and *ENG* encode receptors of the TGF-β family and molecular defects are known to cause the vascular disorder HHT characterized by the presence of mucocutaneous telangiectasia and visceral arteriovenous malformations (Johnson et al., 1996; Massagué, 1998; McAllister et al., 1994; Scharpfenecker et al., 2007). Early-onset PAH patients present without HHT, but could develop the condition later in live (Fujiwara et al., 2008; Harrison et al., 2003). The genes *BMPR2*, *ACVRL1* and *ENG* were implicated in PAH pathogenesis and highlighted the role of the TGF-β/BMP pathway in the disease. Candidate gene analyses were undertaken to screen further members of the BMP pathway for potential deleterious variants, more specifically the SMAD family (Nasim et al., 2011; Shintani et al., 2009). Three separate studies run targeted sequencing screens and identified *SMAD1* (n=1), *SMAD4* (n=2) and *SMAD9* (n=3) variants in 348 PAH cases (Drake et al., 2011; Nasim et al., 2011; Shintani et al., 2009). Functional analyses demonstrated a significant reduction of a downstream BMP target gene *Id2*, but found an unclear impact on the canonical pathway for *SMAD1* and *SMAD4* (Drake et al., 2011; Nasim et al., 2011; Shintani et al., 2009). Finally, the PAH association of *TBX4* was suggested by a study of childhood-onset PAH cases (Kerstjens-Frederikse et al., 2013). Patients were found to have overlapping deletions and a candidate gene analysis of genes contained in the overlapping region resulted in the discovery of *TBX4* (Kerstjens-Frederikse et al., 2013). The study identified 6 *TBX4* variants (mutations, n = 3; deletions, n = 3) in 6 out of 20 children and one mutation in one out of 49 adult patients with PAH (Kerstjens-Frederikse et al., 2013). Mutations in *TBX4* are known to cause small patella syndrome (SPS) and previously unrecognised features of SPS were identified in all PAH patients with *TBX4* mutations (Bongers et al., 2004; Kerstjens-Frederikse et al., 2013). *TBX4* is a required regulator of the embryonic development, part of the T-box gene family (Naiche and Papaioannou, 2003; Sakiyama et al., 2003) and T-box gene mutations have been associated with several developmental disorders (Bamshad et al., 1997; Basson et al., 1997; Kirk et al., 2007; Packham and Brook, 2003; Yagi et al., 2003). Mutations in T-box transcription factor-encoding genes have also been found to lead to congenital heart defects (Hoogaars et al., 2007; Stennard and Harvey, 2005).

Genes not directly related to the TGF-β signalling pathway were associated with PAH and exploring potential novel disease causing pathways. Cellular stress response was implicated in PAH by the discovery of variants in eukaryotic translation initiation factor 2 alpha kinase 4 (*EIF2AK4*). Whole exome sequencing (WES) identified biallelic recessive *EIF2AK4* variants as disease-causing changes by two separate groups in the PAH related diseases pulmonary veno-occlusive disease or pulmonary capillary haemangiomatosis (PVOD/PCH) (Best et al., 2014; Eyries et al., 2014). In total, the studies analysed 14 families followed by 31 unrelated cases and identified 22 subjects with biallelic *EIF2AK4* variants (Best et al., 2014; Eyries et al., 2014). The identified gene *EIF2AK4* is found to regulate cellular stress related angiogenesis, including oxidative stress that is important in pulmonary hypertension development (Anthony et al., 2004; Chaveroux et al., 2011; Donnelly et al., 2013; Fessel et al., 2013).

Nitric oxide (NO) signalling pathway was implicated by the revelation of the novel PAH associated gene caveolin-1 (*CAV1*) (Austin et al., 2012). The study applied WES to analyse a 3-generation family with multiple cases of PAH and an unrelated child (Austin et al., 2012). Two frameshift variants were identified in the study, one in the family and one in the child. Both variants were in highly conserved regions in *CAV1*, in close proximity of each other and adjacent to a cysteine palmitoylation site (Austin et al., 2012). Caveolin-1 is a main component of the caveolae plasma membrane, required for the anchoring process and important for the receptor signalling cascades relevant to PAH, including the nitric oxide pathway (Cohen et al., 2004; Engelman et al., 1997, 1998; Mathew et al., 2004; Patel et al., 2007). Mutations in *CAV1* have been described before adjacent to another palmitoylation site and could explain the reduced caveolin-1 staining in PAH patients (Austin et al., 2012; Hayashi et al., 2001).

The last pathway includes two separate genes from the potassium channel gene family, which were associated with PAH (Ma et al., 2013; Wang et al., 2014). A larger WES PAH study identified variants in the potassium channel, subfamily K, member 3 (*KCNK3*) also called twik-related acid sensitive K+ (*TASK-1*) as disease-causing (Ma et al., 2013). In this study, 13 patients with six separate *KCNK3* heterozygous missense variants were identified in three multi-member family with HPAH, 92 unrelated HPAH and and 230 unrelated IPAH patients (Ma et al., 2013). One family member with an identified *KCNK3* variant developed the disease after recruitment and two unaffected family members in separate families were also carriers of a variant, suggesting incomplete penetrance or of late-onset disease (Ma et al., 2013). The gene *KCNK3* encodes a pH sensitive potassium channel, which controls the resting membrane potential in PASMCs (Czirják and Enyedi, 2002; Hartness et al., 2001; Olschewski et al., 2006). The second potassium channel gene is the voltage gated shaker-related subfamily A, member 5 (*KCNA5*) identified by targeted sequencing of known PAH causing genes as well as further genes of the potassium channel gene family (Wang et al., 2014). The study identified a so-called "second hit" in *KCNA5* in addition to missense mutations in *BMPR2* in one early onset patient with severe PAH (Wang et al., 2014). Replication of a rare digenic genotype has not been observed and the significance of the *KCNA5* report requires caution in the absent of a comprehensive functional analysis.

Genes associated with PAH were predominantly discovered in small cohorts by family based studies, but the interpretation of such results requires caution. Following novel discoveries, additional functional work was performed for the majority of the discovered genes and replication in independent cohorts provided sufficient evidence to confirm causality except *KCNA5*. Earlier studies focused on candidate genes and genes in specific region, but the availability of WGS allowed recent studies to screen for variants in all genes. The majority of disease causing variants are novel, identified in unrelated cases and highlights the role of rare variant in PAH.

3.2 Methods

The national institute for health research (NIHR) BioResource - Rare Diseases(BR-RD) study recruited 9,224 subjects for 16 participating projects (see Tab. 3.1) and used whole genome sequencing (WGS) to analyse their genomes. Individual WGS data were quality controlled, resulting variants aggregated from 9,110 individuals and released for analysis to all participating projects. The PAH project from the NIHR BR-RD study aimed to elucidate the complete genetic basis of the rare disease PAH.

3.2.1 Study design

Subjects were recruited for the NIHR BR-RD study with the objective to identify the cause of rare diseases using next generation sequencing techniques (see chapter 2.2.1 on page 18). The PAH project was part of the NIHR BR-RD study and recruited mainly patients with a diagnosis of PAH (see chapter 2.2.2 on page 18) for diagnosis and the discovery of novel disease-gene associations. Whole genome sequencing (WGS) strategy was deployed at high depth ($\geq$ 15x coverage of at least 95% of the genome with an average depth of 30x) for increased genotype accuracy (Bentley et al., 2008). The PAH disease cohort was compared against the remaining 15 non-PAH participating projects. Subjects for the PAH disease and non-PAH control cohorts were selected based on calculated relatedness information as well as collected phenotype data from the PAH project (see Fig. 3.2). The variant filtering strategy was developed by characterising the known disease gene *BMPR2* and aimed to enrich for disease causing variants. All previously reported PAH disease genes were then assessed for variants explaining PAH by applying the developed filter strategy. Following the assessment, a case-control study design compared the unexplained PAH index cases with unrelated non-PAH controls to determine a disease association with genes previously not described in PAH. Significant associated genes were then assessed in more detail.

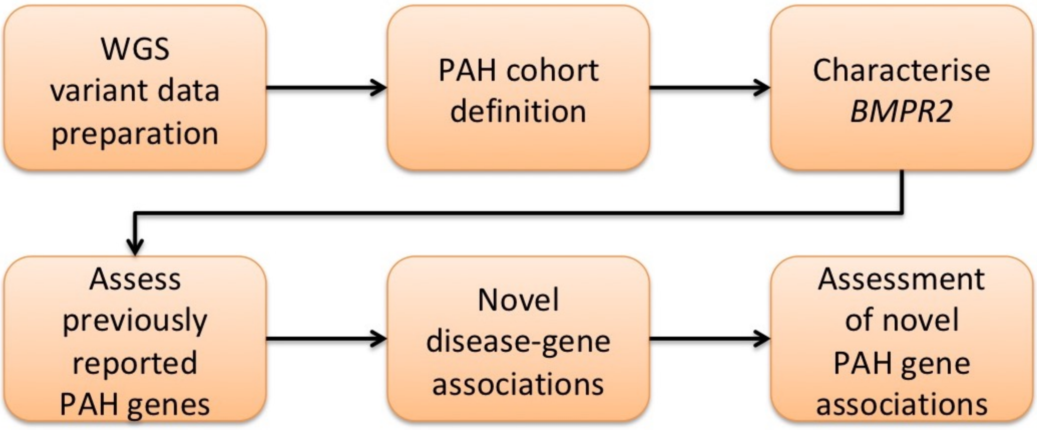

WGS variant data preparation
PAH cohort definition
Characterise *BMPR2*
Assess previously reported PAH genes
Novel disease-gene associations
Assessment of novel PAH gene associations

Figure 3.2 Overview of the analysis steps

3.2.2 Ethical approval

Cases were recruited by 16 projects listed in Tab. 3.1 for the NIHR BR-RD study. The PAH project recruited patients from the same centres as the national cohort study of idiopathic and heritable PAH (see 2.2.2 on page 18). For the PAH project, all cases had a clinical diagnosis of idiopathic PAH, heritable PAH, drug- and toxin-associated PAH, or PVOD/PCH established by their expert centre. Patients with known underlying cause of PAH (chronic thromboembolic disease, congenital heart disease, connective tissue disease, HIV, liver cirrhosis, left heart disease, chronic lung disease) were excluded. All PAH and non-PAH patients or their parents provided written informed consent (UK Research Ethics Committee: 13/EE/0325), or local forms consenting to genetic testing in deceased patients and non-UK cases. Blood and saliva samples were collected under written informed consent of the participants or their parents for use in gene identification studies (UK Research Ethics Committee: 08/H0802/32).

Cardiovascular	
Acronym	Project
BPD	Bleeding, Thrombotic and Platelet Diseases
PAH	Pulmonary Arterial Hypertension
HCM	Myofilament-gene negative Hypertrophic Cardiomyopathy
SMD	Stem Cell and Myeloid Disorders
EDS	Ehlers Danlos Syndrome
Infection & immunity	
PID	Primary Immune Disorders
SRNS	Steroid Resistant Nephrotic Syndrome
PMG	Primary Membranoproliferative Glomerulonephritis
Neuroscience	
SPEED	Retinal Dystrophies / Paediatric Neurology / Metabolic Disease
CSVD	Cerebral Small Vessel Disease
NPD	Neuropathic Pain Disorders
Other rare diseases (including rare cancers)	
MPMT	Multiple Primary Malignant Tumours
ICP	Intrahepatic Cholestasis of Pregnancy
LHON	Leber Hereditary Optic Neuropathy
Other	
GEL	Genomic England
CNTRL	Control samples

Table 3.1 Recruiting projects and sample sizes

3.2.3 Sample collection and whole genome sequencing

Blood samples were collected from rare disease patients recruited to NIHR BR-RD study by participating NHS trusts and collaborating hospitals. The samples were sent to the Cambridge translational genomics (CATGO) laboratory for DNA extraction and quality control before being plated and submitted to Illumina for whole-genome sequencing (WGS).

Sequencing technology and protocols

DNA extracted from venous blood underwent whole-genome sequencing using the Illumina TruSeq DNA PCR-Free Sample Preparation kit. The first 377 and 3,138 samples were sequenced by Illumina on the HiSeq 2000 generating reads of 100 and 125 base pair (bp) length respectively. The remaining samples were processed using the HiSeq X sequencer, generating reads of 150 bp length. The agreed measurements for the sequence data were as follows: $\leq 5\%$ of insert sizes below two times the read length, at least 95% of non-N bases on the autosome covered at $\geq$ 15x and at least 95% of bases along exons covered at $\geq$ 15x.

Software and data release information

The analyses were based on the reference data sets and software versions listed in Tab. 3.2 and 3.3 respectively unless stated otherwise.

Data security

The storage and computational infrastructure was provided by the High Performance Computing service (HPCS) of the University of Cambridge. Access to the HPCS infrastructure is provided through the cryptographic network protocol secure shell (ssh) and open only to a specified list of internet protocol (IP) addresses. Data for the NIHR BR-RD were accessible on a dedicated storage location. Access to the data required approval by the data access committee. Researches were granted access after a successful applications and with the necessary computational ability to analyse genomic data.

Sequence read alignment and variant calling

Illumina processed the sequence reads using the `Isaac` aligner and variant caller using the version specified in Tab. 3.3 (Raczy et al., 2013). The genome reference consortium human genome (GRCh) version 37 was used as the reference to align short reads and to call variants. The reference sequence contained chromosomes 1 to 22, X, Y and mitochondrial sequence. `Manta` and `Canvas` were deployed to call copy number variation (CNV) and

Name	Version	Description
CADD	v1.3	CADD score whole genome SNV and INDELs
CAGE peaks	phase1and2	CAGE combined peak BED file downloaded from the FANTOM 5 archive
Ensembl 37way GERP	75	Conserved regions in humans based on eutherian mammals
Ensembl gene annotations	GRCh37.75	Gene annotations in gene transfer format (GTF)
Ensembl Regulatory build	87	Gene regulatory build 20161117 gene feature format (GFF) downloaded from FTP including HUVEC specific dataset
Ensembl VEP	84	Variant effect prediction for GRCh37
ExAC	r0.3	Whole Exome frequencies
GERP	hg19	Downloaded BigWig file from UCSC FTP
Human GRCh37 reference	75	Human autosomes, X and Y downloaded from Ensembl FTP
Human GRCh38 reference	GCA_000001405.15	Human primary assembly, EBV and decay contigs downloaded from NCBR FTP
NIHR BR-RD	20170104-A	Variant release of the NIHR BioResource – Rare Diseases
Phenotype data	2017-03-05	PAH phenotype data release
PhyloP	hg19	Downloaded 100way PhyloP BigWig file from UCSC FTP
PhastCons	hg19	Downloaded 100way PhastCons BigWig file from UCSC FTP
UK10K	20130411	Exome and whole genome frequencies

Table 3.2 Reference data release versions used throughout the project.

Name	Version	Info
agg	v0.3.3.dev-31-gaa44755	
BCFtools	1.3.1	including git commit bdb01d8
BEAGLE	4.1	executable beagle.22Feb16.8ef.jar downloaded
Canvas	1.1.0.5	
ensemblVEP	1.10.3	R package
GENISIS	2.2.7	
ggplot2	2.1.0	
Isaac	iSAAC-SAAC00776.15.01.27	
Manta	0.23.1	
MUMmer	3.23	
perl	5.20.0	
PLINK	v1.90b21	
Primus	1.8.0	
Python	3.4.1	
PyVCF	0.6.8	Python package
R	3.2.4	
Samtools	1.3.1	

Table 3.3 Software release versions used throughout the project

structural variation (SV) respectively (Chen et al., 2016b; Roller et al., 2016). Canvas uses read coverage information, while Manta uses paired-end and partial read mapping information to determine the breakpoints for SV and CNV and optimised for medium-sized insertions or deletions (INDELs) up to 10 Kb. The resulting files were securely transferred from Illumina to the dedicated storage on the HPCS. The WGS and genotype data were transferred in binary sequence alignment/map (BAM) file format and variant call format (VCF) respectively (Danecek et al., 2011; Li et al., 2009a). The single nucleotide variants (SNV), multi nucleotide variant (MNV) and INDELs were delivered as VCF and genome VCF (gVCF) files. A separate VCF files was provided for CNVs and SVs. Alignment and variant summary statistics were provided in tab-delimited format and as portable document format (PDF) files.

Quality control analysis

The coverage information from the BAM file was used to infer the gender and was compare against the recorded gender. Summary statistics provided by Illumina were collected per sample and analysed for outliers. The BAM summary included the number of reads (total, aligned, duplicated paired, percent of bases with a base phred quality score greater or equal to 30), mean coverage and fragment length (median, min, max, standard deviation). Variant measurements contained the counts of total, pass, ratio of heterozygous / non-reference homozygous (Het/Hom) ratio, transition/transversion (Ts/Tv) ratio, in the database of single nucleotide polymorphisms (dbSNP), in genes / exons / coding regions, stop lost, synonymous, non-synonymous for SNV, insertions and deletions.

Sequence data backup and availability

The risk of loss of data was reduced by mirroring the alignment and variant files to an off-site data centre. Alignment files were also submitted to the European Genome-phenome Archive (EGA) at the EMBL – European Bioinformatics Institute (EBI) for public availability through controlled access and as an additional backup.

File based variant normalisation and aggregation

The gVCF aggregation tool (agg) was used to normalise and aggregate SNV, INDELs, MNV, no-call regions and reference call regions (Illumina, Inc., 2015). The normalization by agg was based on the `BCFtools norm` implementation and included left-alignment / trimming of INDELs and the decomposition of MNVs into SNVs part of the ingest1 command for consistent representation (Danecek et al., 2011). The ingest1 step stored the normalised VCF in binary variant call format (BCF) per sample, which were used by the ingest2 command to pre-merge the BCF files in chunks of 200 samples. Selected pre-merged chunks were then merged into a multi-sample VCF using the genotype command.

Detection of sample duplication

The detection of duplicated samples was based on a representative set of SNVs and the deployment of the `PLINK` package to perform an identity-by-descent (IBD) analysis (Purcell et al., 2007). The 20,000 SNVs were selected at the beginning of the project and were present in 3,000 NIHR BR-RD samples and 2,504 samples of the 1,000 Genomes (1kG) project. Selected SNVs had a minor allele frequency (MAF) > 0.05 in NIHR BR-RD, were retained with PASS filter in all samples and existed in the 1kG project. Following the SNV selection, the representative SNVs were extracted from the single sample VCFs, merged into a multi-sample VCF file using `BCFtools merge` and LD pruning was performed using `PLINK` with a window size of 50 bp, window shift of 5 and a variance inflation factor threshold of 0.2. The remaining 14,721 autosomal, linkage disequilibrium (LD) pruned SNVs were used to perform the IBD calculation. Pairs of samples with a relatedness measure (pi-hat) value > 0.9 were flagged as possible duplicated samples. Flagged pairs of samples were checked with the submitting centres and either one sample was removed as a confirmed duplication, both samples were removed as mislabeled or were both passed as a confirmed case of twins.

Estimation of population and family structure

Ethnicities and relatedness were estimated based on a representative set of SNVs and the deployment of the GENISIS package to perform PC-Air and PC-Relate respectively (Conomos et al., 2015, 2016). The selected 35,114 autosomal SNVs were present on Illumina genotyping arrays (HumanCoreExome-12v1.1, HumanCoreExome-24v1.0, HumanOmni2.5-8v1.1), did not overlap regions excluded following quality control or multiallelic sites in the 1kG Phase 3 dataset (Sudmant et al., 2015). In addition, these SNVs did not have any missing genotypes in NIHR BR-RD, had a MAF of 0.3 or above and LD pruning was performed using PLINK with a window size of 50 bp, window shift of 5bp and a variance inflation factor threshold of 2. The 2,110 samples from the 1kG project including the European (EUR), African (AFR), South Asian (SAS) and East Asians (EAS) populations (excluding the admixed American population) were merged with the NIHR BR-RD samples, filtered on the selected SNVs and used to perform a principal component analysis (PCA) using PC-Air considering the 1kG samples as an independent set. The scores of the leading five principal components (PC) were modelled as data generated by a population specific multivariate Gaussian distribution and the corresponding mean and covariance parameters were estimated. The likelihood was computed for the genotypes of every NIHR BR-RD sample that it belonged to each subpopulation under a mixture of multivariate Gaussians model after projecting the loadings for the leading five PC from the 1kG PCA. The population with the highest likelihood was assigned to each sample, unless the highest likelihood was similar to values from other populations, as expected for unrepresented populations or admixed ancestry, and labeled as ‘other’. PC-Relate was used to identify related individuals in NIHR BR-RD. We used the first 20 PCs from PC-Air to adjust for relatedness and extracted the pairwise IBS and kinship values. The pairwise information was used by Primus to infer family networks and calculate the maximum set of unrelated samples (Staples et al., 2014).

Data freeze for analysis

The latest available NIHR BR-RD variant data (see section 3.3.1 on page 64 for details) and PAH phenotype data release (see chapter 2 on page 15 for a detailed description) were used for the following analyses (see Tab. 3.2 for release dates).

3.2.4 Definition of PAH and control cohort

Diagnosis and phenotype information were extracted for PAH project subjects included in the NIHR BR-RD release. From these PAH subjects, patients were included in the affected PAH adults (PAHAFF) cohort with an age ≥ 16 at date of diagnosis and a clinical diagnosis of IPAH, HPAH or PVOD/PCH. Unrelated affected PAH adults or subjects with the higher sequence identifier of related pairs were also included in the PAH index cohort (PAHIDX). The maximum unrelated set of subjects was retrieved from the provided NIHR BR-RD release and defined as the unrelated WGS cohort (UWGS10K) (see section 3.2.3 on page 54). The unrelated non-PAH control group (UPAHC) was defined as the UWGS10K cohort excluding all subjects part of the PAH project. In addition, the female, European and female European subjects were identified as sub cohorts for each of UPAHC, PAHAFF and PAHIDX using the genetically identified gender and population information. Files with the identifiers of the defined three cohorts and nine sub cohorts were created and used for downstream analyses.

Variant file annotation

The aggregated variants were annotated using Ensembl's Variant Effect Predictor (VEP) version 84 (McLaren et al., 2010). The VEP annotation included consequence type prediction, gene annotation, SIFT (Ng and Henikoff, 2003), PolyPhen-2 (Adzhubei et al., 2013) and allele frequencies in 1kG (1000 Genomes Project Consortium et al., 2015). VEP defined the annotated fields in the VCF header and stored the annotations as structured text as one 'ANN' entry in the info field. The structured text encoded a list of feature annotations separated by ',' and each feature contained a list of annotation values for the defined fields separated by '|'. Custom annotation was added for UK10K (UK10K Consortium et al., 2015), ExAC (Song et al., 2016), GERP (Cooper et al., 2005), 100 way PhyloP (Pollard et al., 2010), 100 way PhastCons (Siepel et al., 2005) and CADD (Kircher et al., 2014). The allele count (AC), allele number (AN), hemizygous (AC_Hemi), homozygous (AC_Hom), heterozygous states (AC_Het), allele frequency (AF) and minor allele frequency (MAF) were calculated for the three defined cohorts (UPAHC, PAHAFF, PAHIDX) and nine sub cohorts using the `fill-tags` plugin. I extended the fill-tags plugin for improved efficiency [1], which is part of `BCFtools` (see Tab. 3.3).

Merging of copy number variants

Copy number variation (CNV) and structural variation (SV) were called in each samples by applying `Isaac` copy number variant caller (`Canvas`, Illumina) and Isaac structural

[1] https://github.com/samtools/bcftools/pull/503

variant caller (Manta, Illumina), which use different algorithms (Chen et al., 2016b; Roller et al., 2016). The identified CNV and SV calls were first grouped into the seven reported events (Canvas: deletion, duplication; Manta: deletion, tandem duplication, translocation, insertion, inversion) for each sample using the library pyVCF. An R script was used to separate autosomes from allosomes and stored them as BED files (R Core Team, 2016). Deletions called by Canvas and Manta for a sample were combined with a reciprocal overlap of $\geq 20\%$. The combined deletions required at least one PASS call and the support from both Canvas and Manta to be selected. These selected deletions were compared against the Zarrei database containing previously published deletions (Zarrei et al., 2015). Deletions overlapping $\geq 50\%$ with database entries were removed using bedtools and R scripts (Quinlan and Hall, 2010). The remaining deletions were assessed for reciprocal overlap of $\geq 50\%$ across samples and the number of samples with overlapping deletions recorded. Gene annotations from protein-coding canonical transcripts were added to deletions, which overlapped within a 10bp window of an exon. The annotation was extracted from Ensembl gene annotations (see Tab. 3.2).

3.2.5 Rare variant selection

The merged VCF file was filtered for PASS in the FILTER column to retrieve variants with an OPR greater or equal to 0.8 using the filter command of BCFtools. Autosomal variants with a MAF greater or equal to 1 in 10,000 in UPAHC, UK10K, ExAC and 1kG were excluded by applying the filter command of BCFtools and stored as a rare variant set. The X chromosome was filtered separately to reflect that females contain two alleles compared to one allele in males. The frequency was adjusted to 1 in 8,000 in order to take into account the haploidy in males and to retain variants with an AC of 1 in UPAHC. The rare variants were further filtered using a regular expression query as part of BCFtools filter to select variants with consequence types annotated as protein truncating variant (PTV) (splice_acceptor_variant, splice_donor_variant, stop_gained, stop_lost, frameshift_variant, start_lost, transcript_amplification, transcript_ablation) or missense variants (missense_variant, inframe_insertion, inframe_deletion, protein_altering_variant).

Protein coding variation filtering in *BMPR2*

Filtered and annotated variants from the *BMPR2* loci were extracted using the filter command of BCFtools and further processed using python, PyVCF package and custom code to parse VEP annotation. The following fields were extracted for the assessment: canonical (yes/no), consequence type, gene name, UPAHC MAF, combined annotation dependent depletion

(CADD) PHRED score, SIFT score, PolyPhen score, UK10K WES AF, UK10K WGS AF, EXAC AF, 1kG MAF and subject IDs with genotypes.

Non-coding region analysis upstream of *BMPR2*

In order to highlight interesting locations of the 5KB upstream regions of *BMPR2*, the regulatory features including promoter position information were extracted from the GRCh37 Ensembl regulatory build. The human CAGE peaks BED file was used to define transcription start site (TSS) locations and conserved regions were extracted from the file provided by Ensembl (see Tab. 3.2). Extracted region information was translated into BED format and `BCFtools` view was used to extract the corresponding variation and genotype information for the analyses. The region and variant information were visualised using the integrative genomics viewer (IGV) (Robinson et al., 2011a).

Copy number variation analysis of *BMPR2*

The merged `Manta` and `Canvas` deletions were extracted for the *BMPR2* locus. Deletions were removed, if the entry had more than, or equal to, 1 in 1,000 overlapping entries recorded in WGS10K subjects, or no confident support was provided by `Manta` or `Canvas`. The filtering was performed in `R`. Coverage plots were created by extracting regions with `SAMtools`, calculating the coverage using `BEDtools` and the mean was calculated over small regions using `Python`.

Identification of distantly related cases

The NIHR BR-RD study prepared aggregated variant releases during various points of the study and used the date of the data freeze as release identifier. The analysis of shared segments of the genome was only performed for one release identified as 20160212-A. The NIHR BR-RD 20160212-A release consisted of 8,066 subjects and included 5,707 unrelated non-PAH controls, 864 affected PAH adults and 856 affected PAH adult index cases. For the analysis, SNVs were selected with an AF greater than 0.05 in 1kG, an AF greater than 0 in NIHR BR-RD and with a filter flag of PASS by using the `filter` command from `BCFtools`. BEAGLE was used to phase the genotypes of subjects and to identify shared segments between pairs of subjects (Browning and Browning, 2009). Only pairs of subjects were considered, where both subjects were part of the PAH cohort. Shared segments overlapping the start of the *BMPR2* gene (position 2:203241659) were selected for the *BMPR2* loci analysis. Filtering was performed using `R`.

3.2.6 Filter and analyse variants

The variant filtering strategy was developed based on the *BMPR2* gene to enrich for deleterious variants (see section 3.3.3 on page 71). The same filter strategy was then applied to identify deleterious variants in previously reported PAH disease genes (see 3.3.4 on page 89). Finally, all variants were filtered using the same strategy. For the developed filter strategy, variants were removed from the annotated, rare variant set if the variant was not present in the cases or control cohort (AC > 0 required for PAHIDX or UPAHC). PTV, missense and combined filtering strategies were applied for the remaining variants and regarded only annotations from the canonical transcript of protein coding genes (see Fig. 3.3). First, variants were selected for the PTV filter with a consequence type of frameshift, splice donor / acceptor, start lost, stop lost / gained or transcript ablation / amplification. Second, the missense filter included consequence types of missense, inframe insertion / deletion or protein altering variant. Variants were excluded with a CADD score of less than 15, or both SIFT and PolyPhen-2 prediction of tolerated and benign respectively. Third, the combined filter included variants passing either the PTV or missense filter. The variant sets retrieved by the PTV, missense or combined filter were each analysed separately and was performed using the ensemblVEP package (McLaren et al., 2016). For each protein coding gene, we calculated the number of subjects carrying filtered variants in cases and in controls. If a subject carried more than one variant for the same gene, the subject was counted only once. The total number of carriers in each cohort was then used to test for over-representation of variants in cases for that gene. A one-tailed (greater) Fisher's exact test was applied with Bonferroni post hoc correction for multiple testing to determine the p-values for genome-wide significance. Subjects with likely deleterious variants identified in previously reported genes were removed from the PAHIDX cohort to increase the power in detecting signals possibly masked by these individuals in novel genes. The remaining PAHIDX cases were extracted from the recorded subjects per gene and the same association tests were applied as described above.

Rare loss of copy number variation analysis

Copy number variation (CNV) and structural variation (SV) were called from the WGS data (see chapter 3.2.3 on page 50). With each data release a CNV file was provided, that included merged `Canvas` and `Manta` calls described in chapter 3.2.4 on page 55. These CNV events were analysed to identify an over-representation of deletions in the PAH cohort compared to controls (see chapter 3.3.5 on page 100). `Canvas` uses read coverage information, while `Manta` uses paired-end and partial read mapping information to determine the breakpoints for SV and CNV and optimised for medium-sized INDELs up to 10 Kb. Each deletion call had a

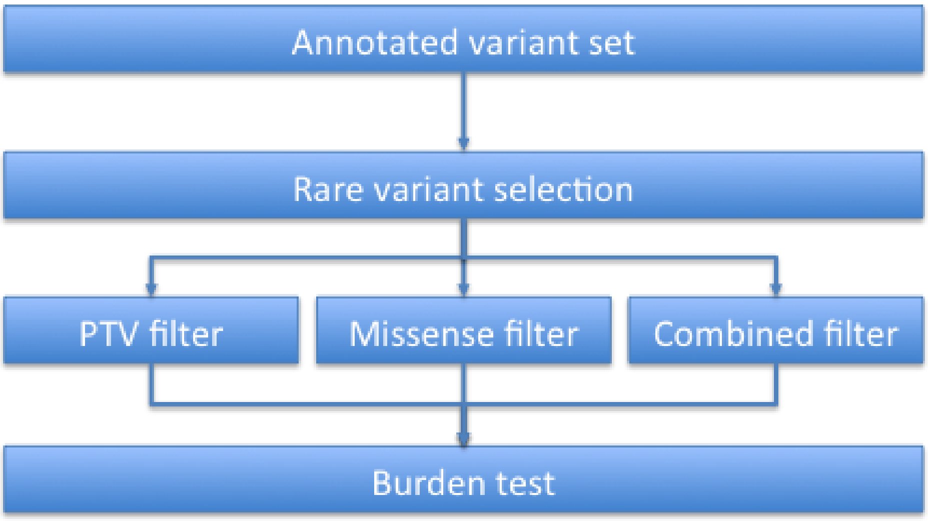


Figure 3.3 Data flow diagram showing different filtering strategies evaluated by a common burden test.

quality score provided and a filter flag to highlight low as well as high confident ('PASS') calls. In addition, deletions shorter than 10Kb were flagged as 'CLT10kb' in `Canvas`, deletions longer than 10Kb were flagged as 'MGE10kb' by `Manta`. We used `R` to select deletions with less than 1 in 1,000 overlapping entries recorded in 9,110 WGS10K subjects and present in either PAHIDX or UPAHC. The assessment of deletions required independent support from both `Manta` and `Canvas` with at least one confident ('PASS' filter flag) call. Calls supported by both algorithms without a 'PASS' flag or by only one method were removed. False positive deletion calls were reduced by requiring support from two algorithms assessing different type of evidence. The stringent setting also removed valid deletion calls (see section 3.3.3 on page 79), which reduced the chance of finding novel gene associations. The Ensembl api and the Ensembl gene annotation (see Tab. 3.2) were used to define exonic regions for canonical transcripts, which are protein coding. The defined exonic regions were extended by 10bp for each start and end position to account for splice region deletions and imprecise breakpoint positions. The `GenomicRanges` package was used to select deletions overlapping these extended exonic positions. The selected variants were tested for association as described above.

Confirmation of CNVs using read depth and homozygosity

The gene `GDF2` harboured three individuals with deletion called by *Canvas*, while *Manta* provided inversion calls for the same region. These calls (deletion and inversion) were labelled as high confidence ('PASS') by both callers. A visual inspection was not conclusive due to the size of the deletion. Alternative methods were developed to assess the correctness of the deletion calls due to the discordance of calls from *Canvas* and *Manta*. These included the comparison of the exonic coverage between genes and the heterozygous / homozygous (het/hom) ratio. For the coverage analysis, the exonic start and end positions for the canonical transcripts of *GDF2* and *BMPR2* were selected from the Ensembl gene annotation file. The read depth for each exonic position was extracted using `SAMtools` depth and the average depth calculated for each gene per sample. Deletions were validated in *GDF2* with a drop of coverage <50% relative to *BMPR2*. For visual inspection, outliers were highlighted with the standard cut-off of the first quartile (Q1) - 1.5x interquartile range (IQR) for *BMPR2* and `GDF2`. The het/hom ration was calculated by selecting the variant calls from each sample for the genomic region 10:48400000-48600000, which overlaps *GDF2*. The variants were extracted from the single sample VCF file with a minimum allele count of 1 and a ‘PASS’ filter flag. The number of heterozygous (HET) and homozygous (HOM) alternate variant calls was counted for the selected region and the ratio (HET divided by HOM) was calculated using `R`. The statistical significance for the coverage and het/hom ratio were determined using a one-sided (less), unequal-variance Student’s t-test, comparing samples with a *Canvas* call with the remaining PAH samples.

Reference genome assembly comparison

The *Manta* algorithm called an inversion overlapping the *GDF2* region in 90% (n=6704) of subjects using paired-end and split-read information. Such high number of inversion calls indicated a possible alternative, incomplete or misrepresentation of the GRCh37 reference. To assess these possibilities, the reference genome GRCh37 and GRCh38 of the wider *GDF2* loci were compared for structural rearrangements. The fasta sequence for the region 10:45000000-50000000 was extracted for both reference assemblies using `SAMtools faidx`. The `NUCmer` tool from the `MUMmer` package was used to perform an all versus all comparison of the extracted reference region. A sequence alignment was performed with `nucmer` to detect minimum clusters of 1,000 bp with a maximum gap of 500bp gap between adjacent matches in the cluster. The alignment was run on both strands and filtered the longest and consistent alignments found for the reference query sequence. The graphics were generated from the filtered information by `mummerplot`.

3.3 Results

3.3.1 Whole genome sequencing

The latest data release (20170104-A) of the NIHR BR-RD study comprises 9,110 samples from 15 different rare disease cohorts and included multiple quality control steps described later (see Fig. 3.4). Each data release provides sample metadata, where whole blood samples were collected, sequenced, assessed and selected over a four year time period. The recruiting projects and their sample sizes are listed in Tab. 3.4 grouped by research interest. The largest five projects (GEL, SPEED, PID, BPD and PAH) account for 76% (n=6,953) of the samples. The whole genome sequencing (WGS) and accompanying sample data produced by Illumina comprised on average 70 gigabyte (GB) per sample. The sample data amounted to 630 terabytes (TB), which was transferred to the high performance computing service (HPCS) for automated validation and quality assessment. Sequence data arrived on average in batches of 131 samples and the quality measurements of the sequence data stored in binary sequence alignment/map (BAM) files were collected for each batch. Fig. 3.5 visualises the effect of protocol changes during the course of the project. There was a shift in (a) fragment size between 100 and 125 bp and subtle quality differences between 100 and 125 bp protocols in respect of duplication (b) and read length (c,d). A significant increase was observed in duplication rate (c) and a drop in the second read quality (d) for the 150bp protocol. The changes between the fragment sizes were within the defined quality specifications (see chapter 3.2.3). Protocol differences were further assessed after variant calling (see chapter 3.3.1).

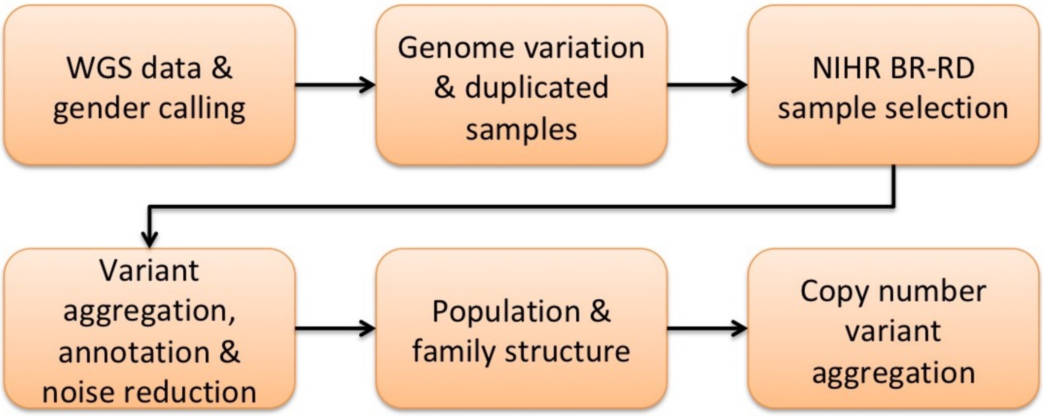

WGS data & gender calling
Genome variation & duplicated samples
NIHR BR-RD sample selection
Variant aggregation, annotation & noise reduction
Population & family structure
Copy number variant aggregation

Figure 3.4 NIHR BR-RD analysis and quality control steps

Cardiovascular

# Samples	Acronym	Project
1,167	BPD	Bleeding, Thrombotic and Platelet Diseases
1,131	PAH	Pulmonary Arterial Hypertension
241	HCM	Myofilament-gene negative Hypertrophic Cardiomyopathy
221	SMD	Stem Cell and Myeloid Disorders
15	EDS	Ehlers Danlos Syndrome

Infection & immunity

# Samples	Acronym	Project
1,308	PID	Primary Immune Disorders
249	SRNS	Steroid Resistant Nephrotic Syndrome
151	PMG	Primary Membranoproliferative Glomerulonephritis

Neuroscience

# Samples	Acronym	Project
1,384	SPEED	Retinal Dystrophies / Paediatric Neurology / Metabolic Disease
244	CSVD	Cerebral Small Vessel Disease
168	NPD	Neuropathic Pain Disorders

Other rare diseases (including rare cancers)

# Samples	Acronym	Project
521	MPMT	Multiple Primary Malignant Tumours
261	ICP	Intrahepatic Cholestasis of Pregnancy
71	LHON	Leber Hereditary Optic Neuropathy

Other

# Samples	Acronym	Project
1,963	GEL	Genomic England
15	CNTRL	Control samples

Table 3.4 Samples included per project

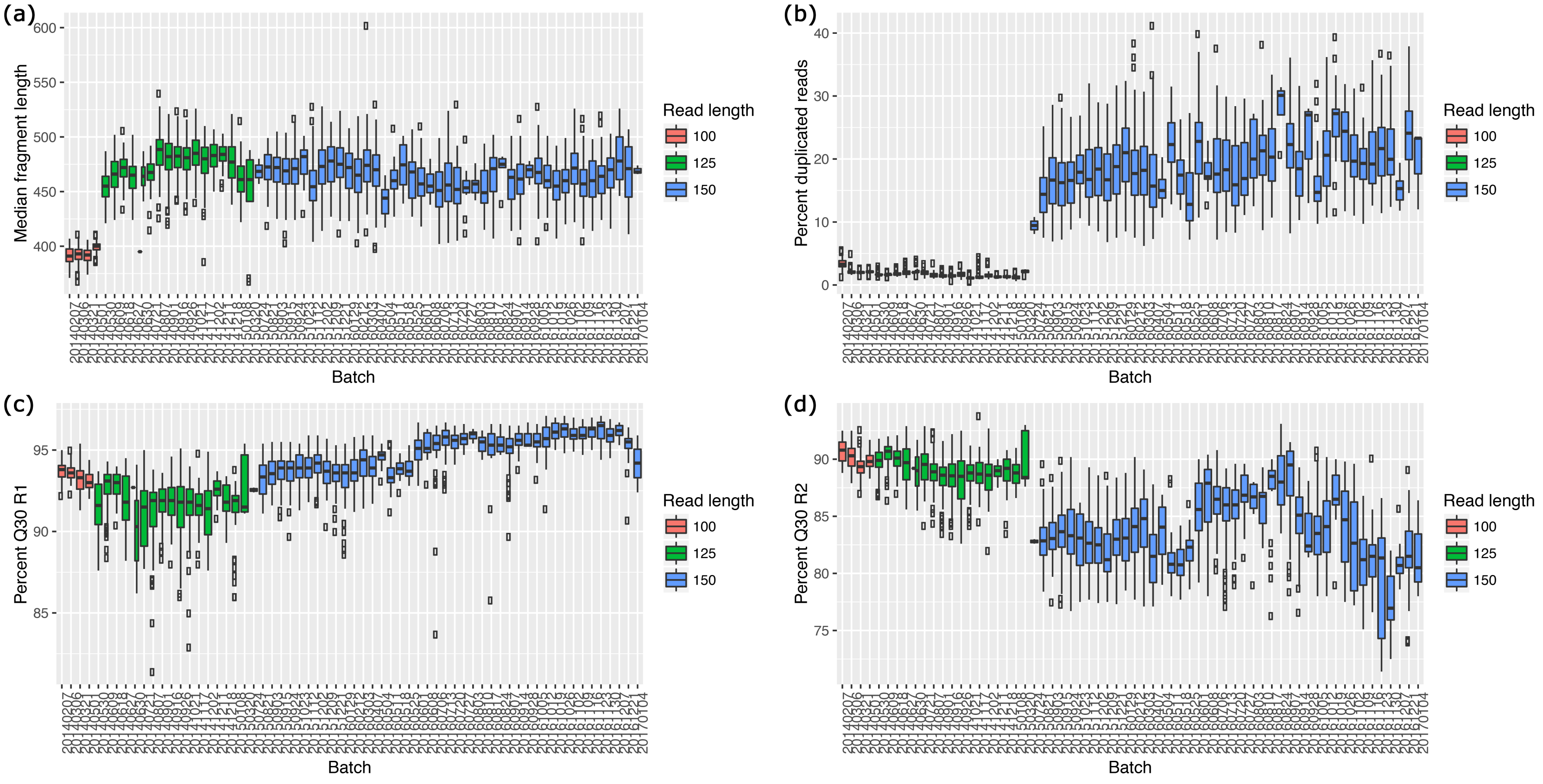


Figure 3.5 Effect of protocol changes during the course of the project. The change in protocol from 100 to 125 bp and 125 to 150 bp can be observed by the change in the fragment size (a) and the increase of duplicate reads (b) respectively. The quality of the first read (c) is lower for the 125bp and we observed a drop in the read quality of the second read (d) for the 150bp.

Sequence based gender calling

The gender was detected of delivered samples using the difference in read coverage information between the autosomes and the allosomes. The sequence-based gender was compared to the self reported gender and checked for discrepancies. For 99.8% (n=9,095) of the samples a reported gender was available, comprising 3,907 (43%) males and 5,188 (57%) females.

Whole genome variation

Variant quality and summary statistics were collected and assessed across batches to detect outliers and to estimate the impact of protocol changes to the analysis. The batch effect of read length changes is highlighted in Fig. 3.6 for (b) transition/ transversion (Ts/Tv), (c) number of SNVs and (d) number of INDELs while there was no differences in (a) the number of synonymous / non-synonymous variants. The increase of read length reduced the Ts/Tv and one sample was found to have an unexpected low value of 2.05. The number of (c) SNVs and (d) INDELs show a significant increase of variants in 150 bp read length compared to the rest. Outliers mainly present with increased number of variants. The increase is not read length specific and suggests to reflect genetic diversity within the NIHR BR-RD cohort.

Sample duplication detection

We screened incoming samples for duplicated submissions, so that one of the samples from the same subject could be removed from the analysis. Duplications occurred due to the same individual being recruited twice in different hospitals to the same project, or the same individual being recruited by different projects, or the same sample being sent twice for sequencing. The relatedness of pairs of individuals was calculated using the identity-by-descent (IBD) method providing the pi-hat score to identify duplicated samples. Identified duplicates were manually checked to avoid removing genuine twins. Fig. 3.7 shows the clear separation of the duplicated pair of samples or twins with a pi-hat score close to 1 compared to the rest. From the 9,224 sequenced samples, there were 1.2% (n=104) samples removed due to duplication.

Sample selection

The variation information from 9,224 samples were available and the quality control assessments excluded 114 samples. The exclusions are due to missing contract specifications (n=1), mislabelling (n=9) and duplication (n=104). The remaining 9,110 subjects assemble the 20170104-A release. After the variant quality assessment and sample selection, we normalised, aggregated and annotated the single sample data sets as described below.

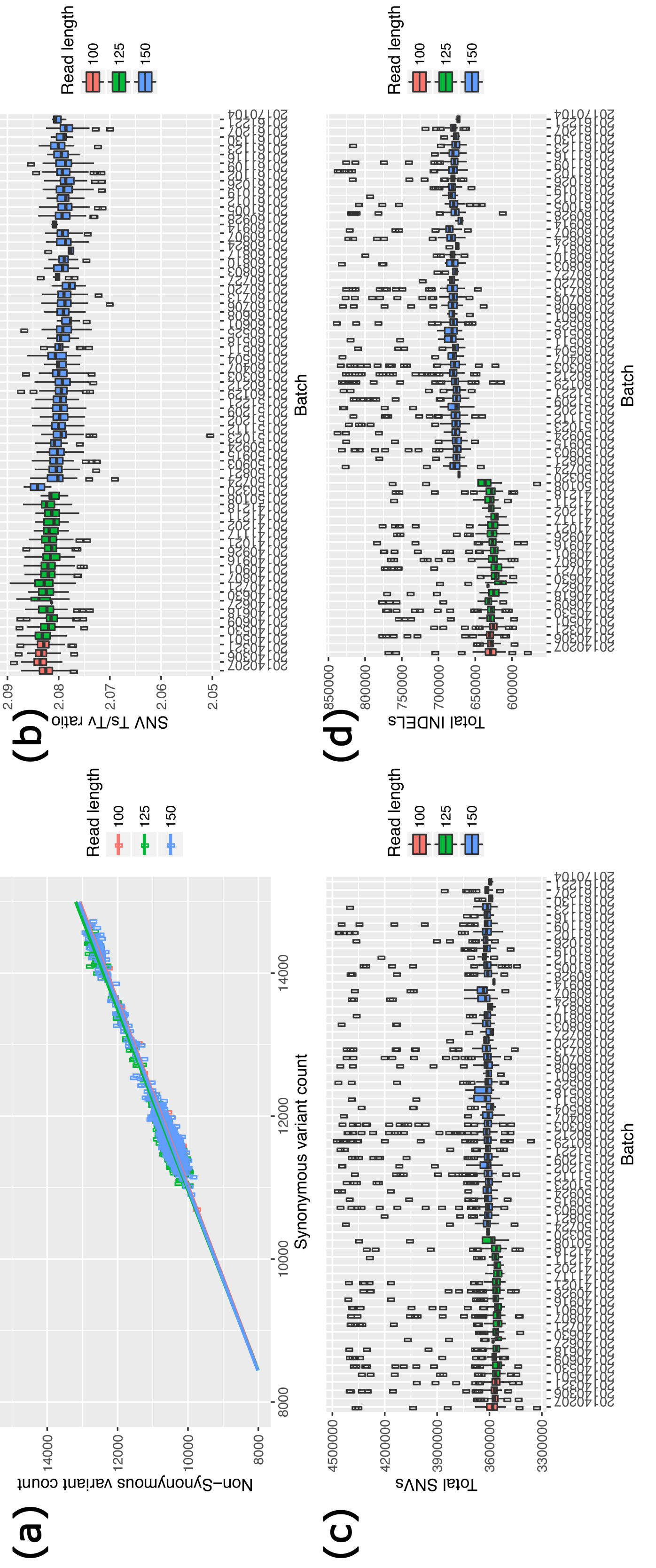


Figure 3.6 Consistency measurements to assess impact of read lengths. (a) Correlation of synonymous / non-synonymous variants, (b) Ts/Tv ratio for SNVs, and total number of (c) SNVs and (d) INDELs per batch respectively. While the Ts/Tv ratio for (b) SNVs is mainly stable across different read lengths, the total count of SNVs (c) and INDELs (d) shows an overall increase for the 150 bp read length library.

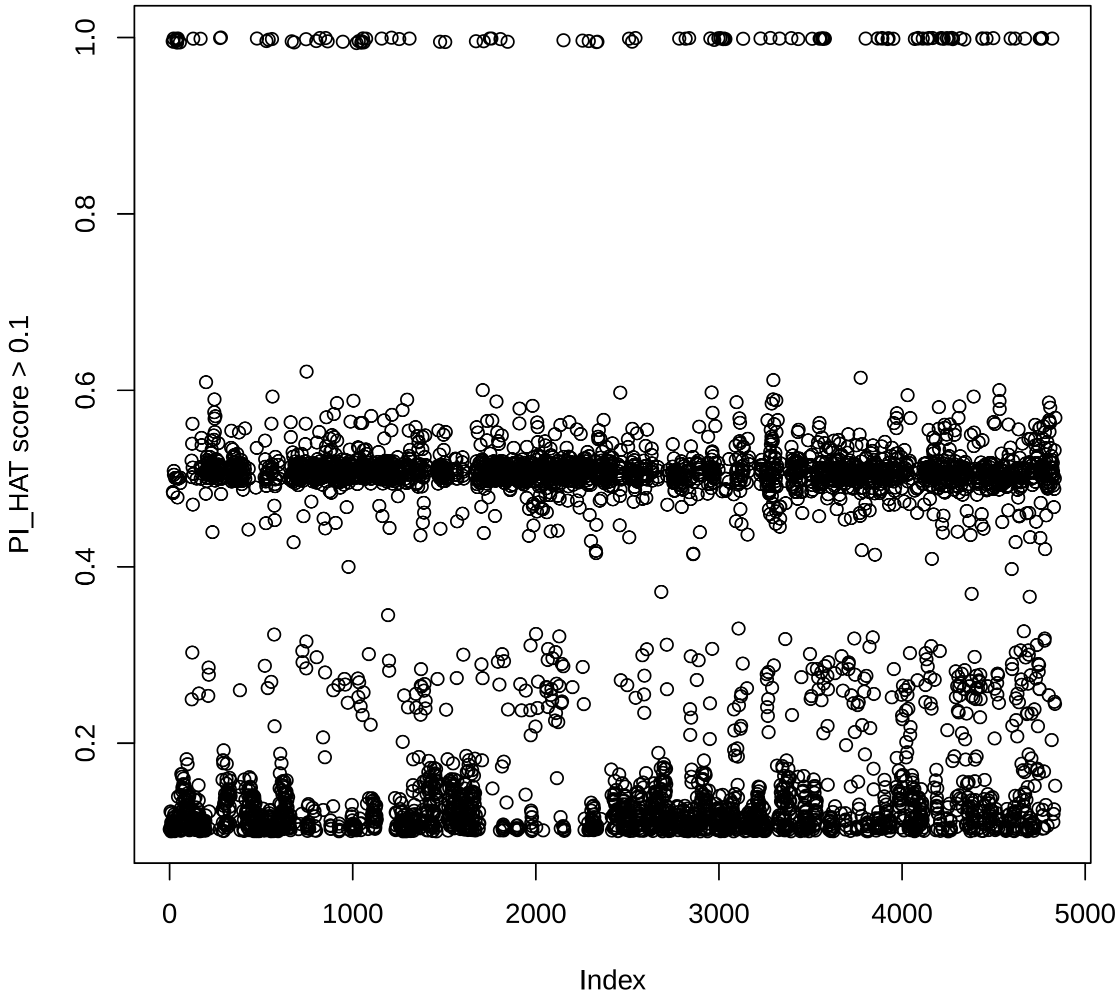


Figure 3.7 Detection of duplicated samples. The pi-hat score is a measure of relatedness between pairs of samples. Pairs of samples with a pi-hat score greater than 0.9 were regarded as duplicated or twins. One samples of each duplicated pair was removed.

Variant normalisation and aggregation

The genome variant call format (gVCF) files for the 9,110 subjects were used for the aggregation process. As a first step, we normalised the variants from the gVCF files for a consistent representation of SNVs and INDELs. On average each individual presents with 4.3 million variants. The normalised variants from the single sample gVCF files were aggregated into one multi-sample VCF file containing 291 million variants (84% SNVs, 15% INDELs, 1% others) for 9,110 subjects.

Variant noise reduction

The normalised samples were aggregated into one multi-sample VCF file as described above, where each variant has a genotype and associated values for each sample. The collated information allows the identification of common and rare variants, but does not distinguish between biological and technical events. The AGG tool provides aggregated annotation regarding the call rate (CR) and pass frequency (PF), that describes the proportion of non-missing genotypes and proportion of PASS calls of alternate genotypes respectively. To distinguish between biological and technical variants, I made use of the CR, which measures the ability to call alleles for a genomic position, and pass frequency (PF), which is indicating the consistency of confident calls for a given position. The combined overall pass rate (OPR = CR x PF) accounts for challenging regions that are difficult to call. The OPR distribution of variants was compared with the minor allele frequency (MAF) in NIHR BR-RD (see Fig. 3.8). For a random sample of 1% of variants, The comparison shows an enrichment of rare variants (MAF $<0.5\%$) on both OPR extremes (0 and 1), while low-frequency ($0.5\% \leq$ MAF $<5\%$) and common variants (MAF $\geq 5\%$) are located closer to OPR 1. This suggests that variants called in more samples contain less technical noise and were called with more confidence. To determine the OPR cutoff, I selected low-frequency and common (MAF $\geq 0.5\%$) SNVs, calculated the first quartile (Q1) - 1.5*IQR and the resulting value of 0.807684 was rounded to one decimal place (0.8 OPR). The inclusive OPR of 0.8 was then used to reduce technical artefacts and enrich for true biologically events. The filtering retained 56% of variants with an OPR ≥ 0.8 (see Tab. 3.5).

	SNV	INDELs	others	Out of total
All variants	84% 246M	15% (42M)	1% (3M)	100% (291M)
$\geq$ 0.8 OPR	89% (145M)	10% (17M)	1% (2M)	56% (164M)

Table 3.5 Count of variants per type. The table shows the breakdown of different variant types for all variants and filtered for high confident ($\geq$ 0.8 OPR) variants only.

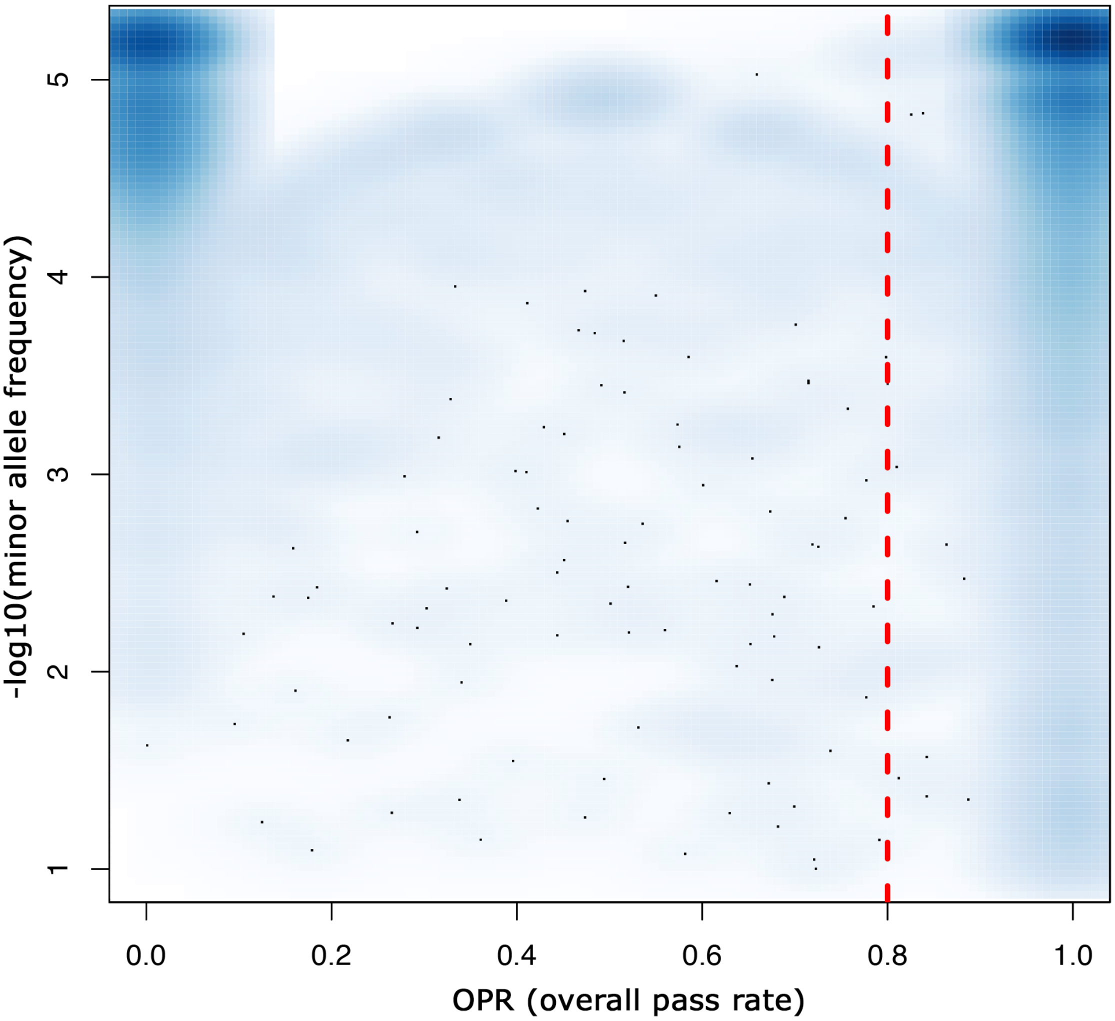


Figure 3.8 Variant exclusion based on OPR distribution. The calculated OPR is compared to the MAF of variants. The darker areas indicate an enrichment of variants close to OPR 0 for rare and close to 1 for rare and common variants. The red line indicates the selected cut-off of 0.8. Variants with an OPR greater or equals to 0.8 were labelled as PASS.

Population and family structure

We assessed the family structure and geographical ancestry of individuals using 1kG and performing principal component analysis (PCA). The NIHR BR-RD subjects with their assigned 1kG population are highlighted in Fig. 3.9. We identified 80.2% European (n=7,307), 9.2% Other (n=844), 7.2% South-Asian (n=649), 2.3% African (n=213), 0.08% East-Asian (n=78) and 0.02% Finnish-European (n=19) in NIHR BR-RD. There were 36.1% related individuals (n=3,293) part of 1,178 families (with two or more individuals) in NIHR BR-RD. The maximum set of unrelated individuals comprises 7,493 (82.2%). In addition, the total number of variants varies between population and African subjects have on average an increase of 20% compared to the rest shown in Fig. 3.10. The 'Other' population were uncategorised samples and shows the largest distribution of values.

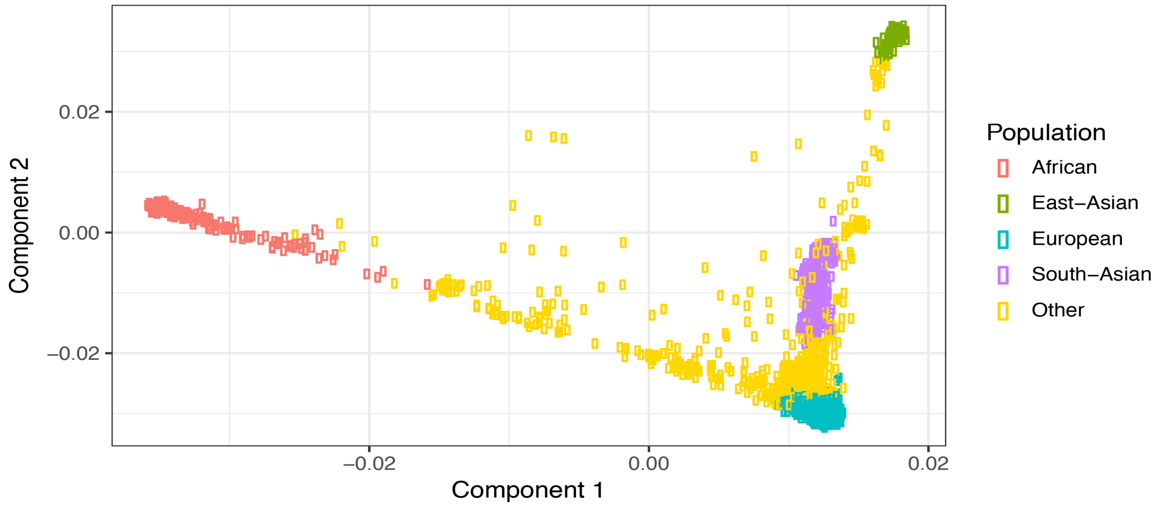


Figure 3.9 Principal component analysis (PCA) of subjects with the assigned ancestry based on the 1kG data. The NIHR BR-RD samples are indicated as points and the colour represents the assigned population.

Variant annotation

The identified population and relatedness information were used to calculate the MAF for unrelated (UBRG) and unrelated European subjects (UBRG_EUR) in NIHR BR-RD. The 301M normalised variants were annotated with the calculated MAF, MAF from 1kG, ExAC and UK10K and with the OPR, consequence type, deleteriousness and conservation scores as defined in methods.

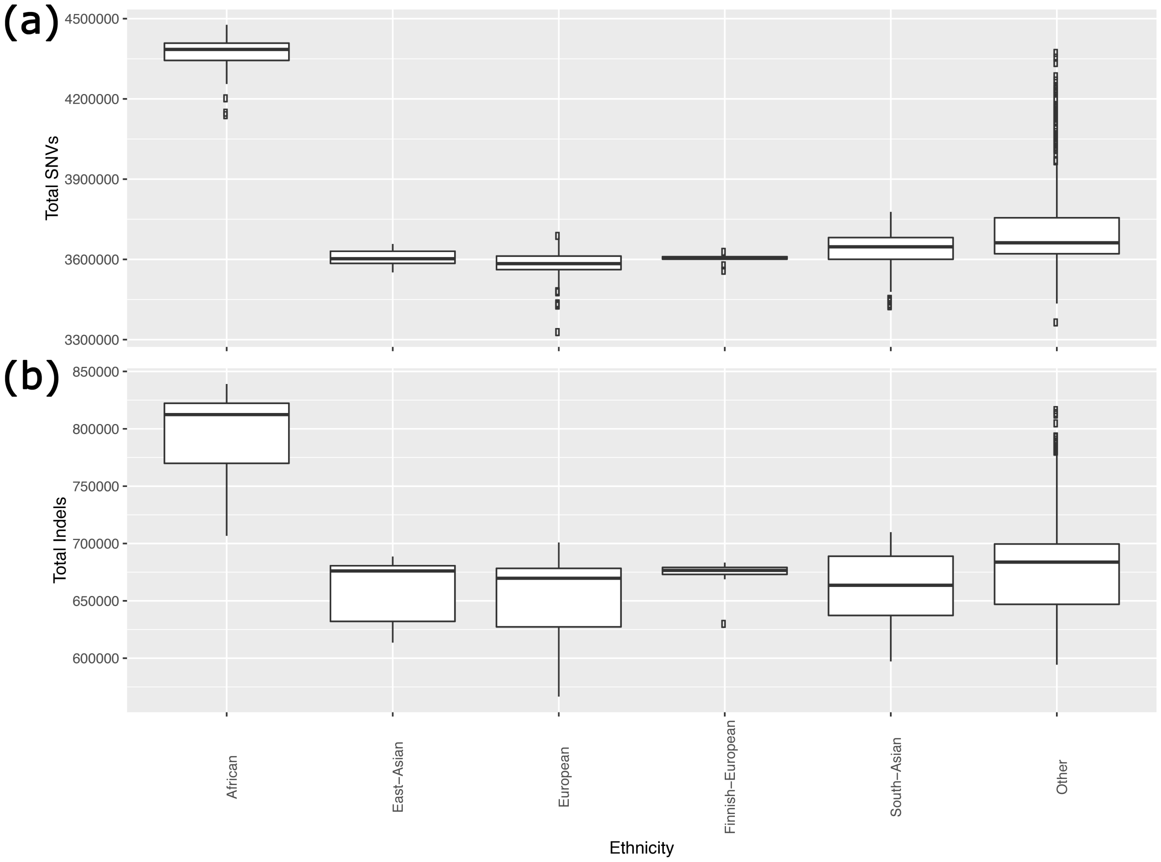


Figure 3.10 Difference in number of variants based on ethnicity. The number of (a) SNVs and (b) INDELs is counted in samples and shown per ethnicity. The African population show an increased number of variants compared to the rest. Unassigned samples in the ‘Other’ population are more diverse.

Copy number variant aggregation and annotation

The deletions called by `Canvas` (n=4.7M) and `Manta` (n=61.7M) were collected for 9,110 NIHR BR-RD samples. We merged 2.8M deletions with sufficient overlap between `Canvas` and `Manta`, of which 2.6M were supported by at least one `PASS` call. These deletions were filtered for a frequency of less than 1 in 1,000 in NIHR BR-RD, which retained 66K and 15K for the whole genome and overlapping protein coding exons respectively.

3.3.2 Definition of PAH and control cohort

The NIHR BR-RD consists of 15 rare disease projects (see Tab. 3.4 on page 62) of which one is PAH. The PAH cohort recruited mainly unrelated idiopathic or heritable PAH patients, while some of the other disease cohorts focussed on the recruitment of related subjects. For the genetic analysis, 1,131 individuals from the PAH cohort were matched with phenotypic information from OpenClinica (OC) to exclude unaffected subjects (n=22), subjects with an age of diagnosis <16 (n=22) due to likely different genetic etiology in children (Ma and Chung, 2017) and subjects with diagnosis other than idiopathic/heritable PAH or PVOD/PCH (n=39). This resulted in 1,048 affected PAH adults (PAHAFF) including 10 affected related individuals, 1,038 unrelated PAH index cases (PAHIDX) and 6,385 unrelated non-PAH controls (shown in Fig. 3.11) to be included in the following analyses.

The suitability of the control cohort was assessed in terms of ethnicity by comparing the proportion of samples belonging to each ancestry in cases and controls. The cohorts were further separated by gender (shown in Fig. 3.12) to allow for the increased female ratio in PAH. We found a comparable representation of ancestries in the gender specific cohorts. The largest ethnic group was Europeans and comprising 84.5% (n=878) and 79% (n=5,089) of the PAHIDX and UPAHC groups, respectively. Ancestral diversity represents a challenge to the identification of causal variants due to population specific variation and a matched control cohort helps to control for possible ancestral differences in rare variant frequencies.

3.3.3 Variant characterisation in *BMPR2*

Previous studies have reported disease causing variants in *BMPR2* and determined the frequency in familial and idiopathic cases based on the aggregation of small studies (Evans et al., 2016). We used the well characterised gene *BMPR2* to develop filter strategies, assess prediction tools, describe the mixture of protein truncating variants (PTV) (i.e. frameshift, start lost, stop gained, stop lost, splice donor and acceptor site) and missense (including inframe insertions and deletions) variants and to establish the frequency of rare coding variants in this large cohort. The first filter strategy was to remove common variants in

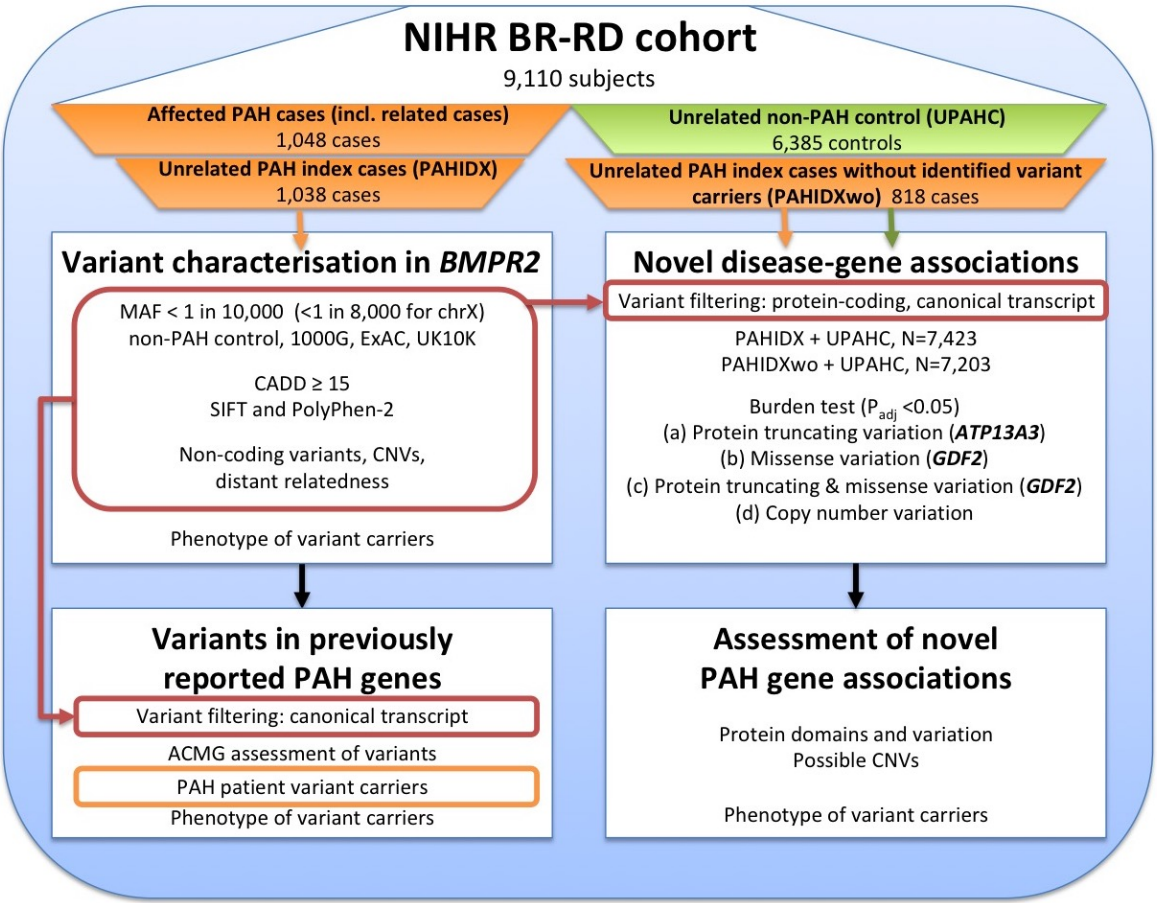

NIHR BR-RD cohort
9,110 subjects
Affected PAH cases (incl. related cases)
1,048 cases
Unrelated PAH index cases (PAHIDX)
1,038 cases
Unrelated non-PAH control (UPAHC)
6,385 controls
Unrelated PAH index cases without identified variant carriers (PAHIDXwo) 818 cases
Variant characterisation in *BMPR2*
MAF < 1 in 10,000 (<1 in 8,000 for chrX)
non-PAH control, 1000G, ExAC, UK10K
CADD ≥ 15
SIFT and PolyPhen-2
Non-coding variants, CNVs,
distant relatedness
Phenotype of variant carriers
Novel disease-gene associations
Variant filtering: protein-coding, canonical transcript
PAHIDX + UPAHC, N=7,423
PAHIDXwo + UPAHC, N=7,203
Burden test (P_{adj} <0.05)
(a) Protein truncating variation (***ATP13A3***)
(b) Missense variation (***GDF2***)
(c) Protein truncating & missense variation (***GDF2***)
(d) Copy number variation
Variants in previously reported PAH genes
Variant filtering: canonical transcript
ACMG assessment of variants
PAH patient variant carriers
Phenotype of variant carriers
Assessment of novel PAH gene associations
Protein domains and variation
Possible CNVs
Phenotype of variant carriers

Figure 3.11 Diagram summarising the analyses. The number of subjects part of the NIHR BR-RD release is listed on the top and each white box below represents a separate analysis. Cases and control cohorts included in these analyses are listed in the orange and green boxes respectively. The developed filter strategy (red border) is applied by multiple analyses and highlighted by the red arrows. Identified PAH variant carries (orange border) are removed from the unrelated PAH index cases (PAHIDX) to create a cohort without identified variant carriers (PAHIDXwo). The applied burden test identifies novel disease-gene associations with a $\mathbf{P}_{adj}$ <0.05 highlighted in bold.

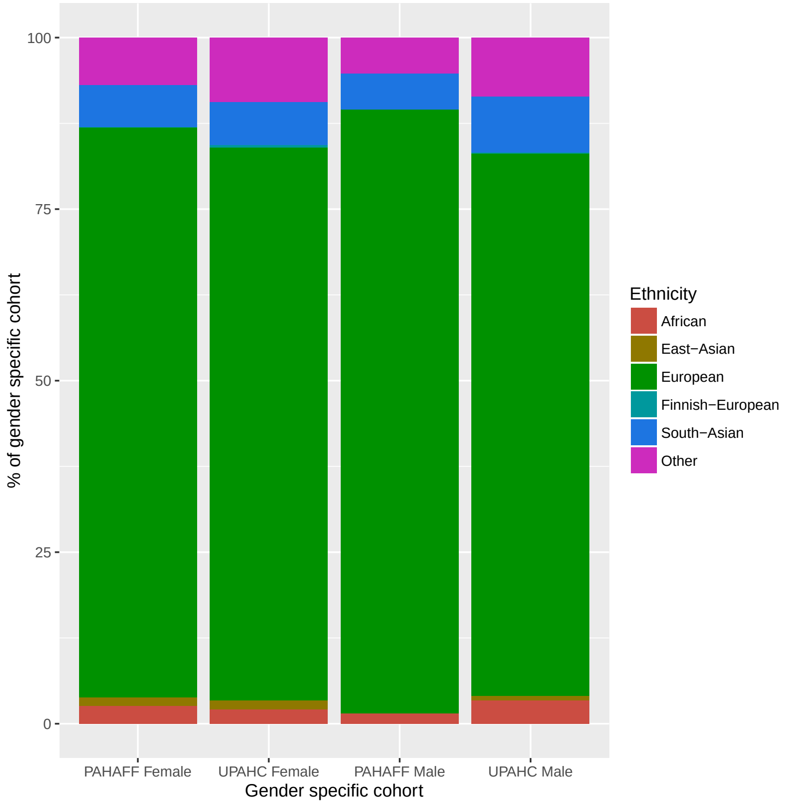


Figure 3.12 Gender specific ethnicity ratio of subjects for PAHIDX and UPAHC. Ethnicities represented in cases were equally represented in the control cohort. The biggest ethnic group in both cohorts were Europeans.

the PAH cohort by filtering on a minor allele frequency of less than 1 in 10,000 in the above defined unrelated non-PAH controls, 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015), UK10K (UK10K Consortium et al., 2015) and ExAC (Lek et al., 2016). This retained variants with an AC of 1 in UPAHC, and was adjusted to 1 in 8,000 for the X chromosome. The adjustment was based on a male ratio of 44% and an expected X chromosome AN of 9,938 for UPAHC. The effect of this filtering strategy on unrelated non-PAH control and affected PAH cases is shown in Fig. 3.13a and 3.13b, and reduced the number of missense variants down to 75% (n=58) listed in Tab. 3.6. One missense variant was found to be shared between cases and controls, but the number of PTVs remained the same. This analysis supported the hypothesis that PTVs are highly disruptive and rare in disease causing genes compared to missense variants. Missense variants alter or reduce the function based on their protein position and often require additional functional and structural prediction to assess their impact. Premature termination codons created by PTVs causes the degradation of mRNAs by nonsense-mediated decay (NMD), but some truncated proteins can still be partially or fully functional depending on the position of the variant (Holbrook et al., 2004). The lack of variants at the end of *BMPR2* in PAHAFF (see Fig. 3.13) is an indication, that missense variants and PTVs in this region might not effect the function of the protein.

Our second filtering strategy was to remove likely benign variants based on deleterious predictions, which included the combined annotation dependent depletion (CADD) deleteriousness score (Kircher et al., 2014), SIFT (Ng and Henikoff, 2003) and PolyPhen-2 (Adzhubei et al., 2013). The CADD score combines 63 annotations including DNA conservation and protein function and a score between 10 and 20 is generally accepted as deleterious. In order to determine a CADD cut-off, I selected the values from missense variants from the PAHAFF cohort, calculated the Q11.5x IQR and the resulting value of 15.22719 was rounded to the integer 15. The selected cut-off value of 15 was then used to remove the most unlikely causative variants (shown in Fig. 3.14) and reduced the missense variants further down to 70% (n=54). The previously found variant shared between cases and controls remained. Further analysis of the remaining variants revealed SIFT and PolyPhen-2 predictions of tolerated and benign, respectively, which we regarded as ambiguous. Excluding these ambiguous variants reduced missense variants to 62% (n=48), which removed variants shared between cases and controls and had no effect on the frequency of PTVs. The combined exclusion of variants based on allele frequency and functional predictions retrieved the most likely disease causing candidates. Filtering on a single value did not achieve the same results even though the CADD score includes protein prediction scores. The measurement was based on the ability to remove variants shared between cases and

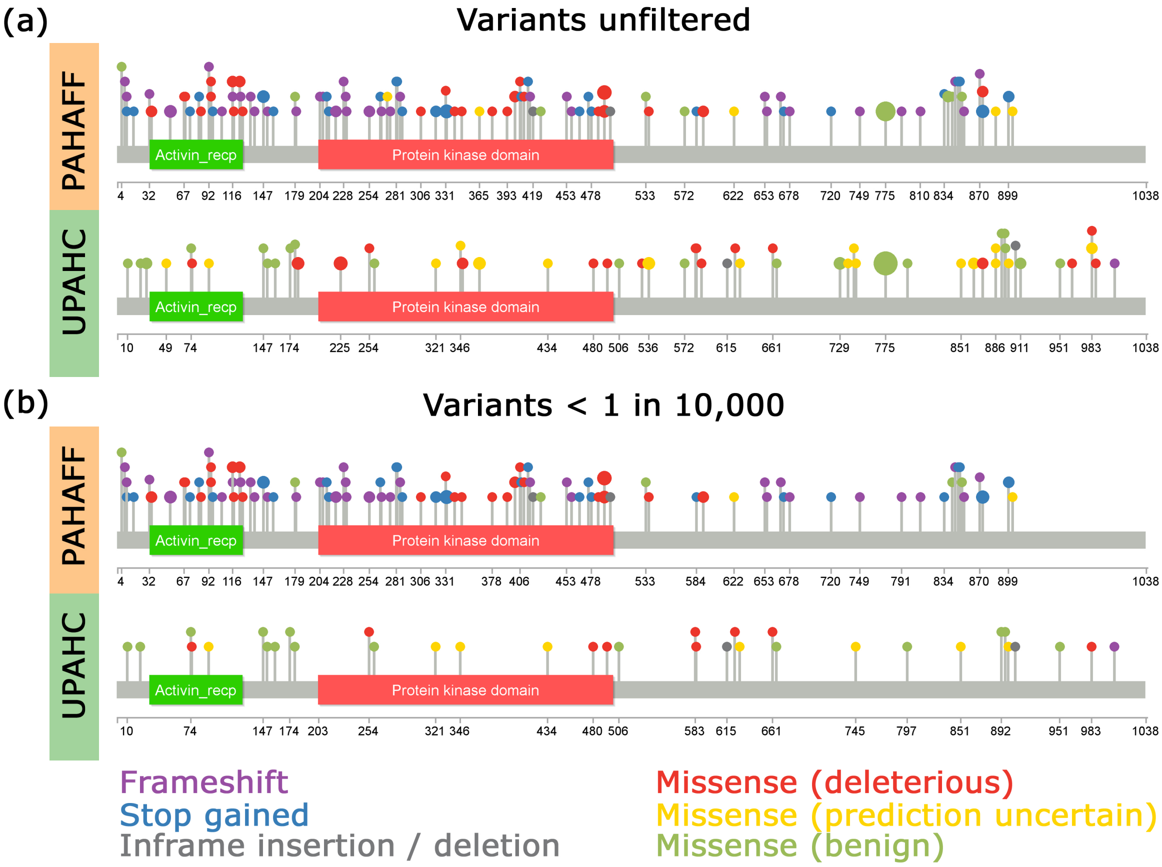


Figure 3.13 Visualisation of the effect of the filter strategy comparing cases and non-PAH controls using lollipop plots. Variants (a) unfiltered and (b) filtered for 1 in 10,000 in UPAHC are displayed for affected PAH cases and unrelated non-PAH controls. Variant types are separated by colour. Missense variants are further separated based on SIFT and PolyPhen-2 predictions of deleterious and damaging (deleterious), tolerated and benign (benign) respectively or labeled as prediction uncertain for the remaining. The increase in lollipop diameter relates to an increase in number of samples with the same variant.

controls. The stringently filtered missense or PTVs in *BMPR2* explained 12.7% (n=134)

MAF <1 in 10,000	CADD $\geq$ 15	SIFT and PolyPhen-2	Variants (non-/missense)	Cases	Control	Fisher's exact p-value
no	no	no	155 (78 / 77)	141 (13.5%)	52 (0.8%)	9.9E-80
yes	no	no	136 (78 /58)	138 (13.2%)	26 (0.4%)	4.2E-93
yes	yes	no	132 (78 /54)	136 (12.9%)	25 (0.4%)	3.4E-92
yes	yes	yes	126 (78 / 48)	134 (12.7%)	20 (0.3%)	2.0E-94

Table 3.6 Summary of filtering strategy. The rows of the table represent different filter settings and the resulting variant counts and number of subjects part of the affected PAH cases (PAHAFF) or unrelated non-PAH control (UPAHC) cohort (no double counting). The MAF filter used to the unrelated PAH control cohort, ExAC, UK10K and 1kG. The recalculated CADD scores provided by scoring service were used for variants with no CADD score annotation part of variant effect predictor.

of affected PAH cases, of which 29% (n=39) were stop gained, 29% (n=39) frameshift, 27.7% (n=37) missense, 9.7% (n=13) splice donor, 3.7% (n=5) splice acceptor and 1% (n=2) inframe insertion / deletion. A closer inspection of the consequence types revealed two splice region variants within 3 cases (one variant found in two subjects) with a CADD score $\geq$ 15. Splice region variants are regarded to have a lower impact and were not part of the analysis.

The developed filtering strategy aimed at removing variants shared between affected PAH and non-PAH control cases to enrich for likely disease causing variants. Remaining variation in *BMPR2* included subjects from both cohorts, affected PAH and non-PAH control. The position of the variant in the genome was not taking into account, but showed a depletion of variants at the end of *BMPR2* in affected PAH cases. The lack of variation suggests that variants near the 3‘ end do not effect the function of *BMPR2* and non-PAH controls could wrongly be assigned a diagnosis of PAH. At the same time, the filtering strategy is also likely to remove variants from PAH patients, which cause the disease due to the stringent filtering criteria and the incomplete penetrance. The filtering strategy does show an enrichment of affected PAH cases compared to non-PAH controls, but would need to be reconsidered for diagnostic purposes.

Non-coding region upstream of *BMPR2*

The foregoing analysis showed the effectiveness of the chosen filter strategy to enrich for likely deleterious variants in index cases. Therefore, the same MAF and CADD score filters were used to explore non-coding regions upstream of *BMPR2*. The considered region extended 5 Kb upstream from the *BMPR2* translation start site and overlapped with the

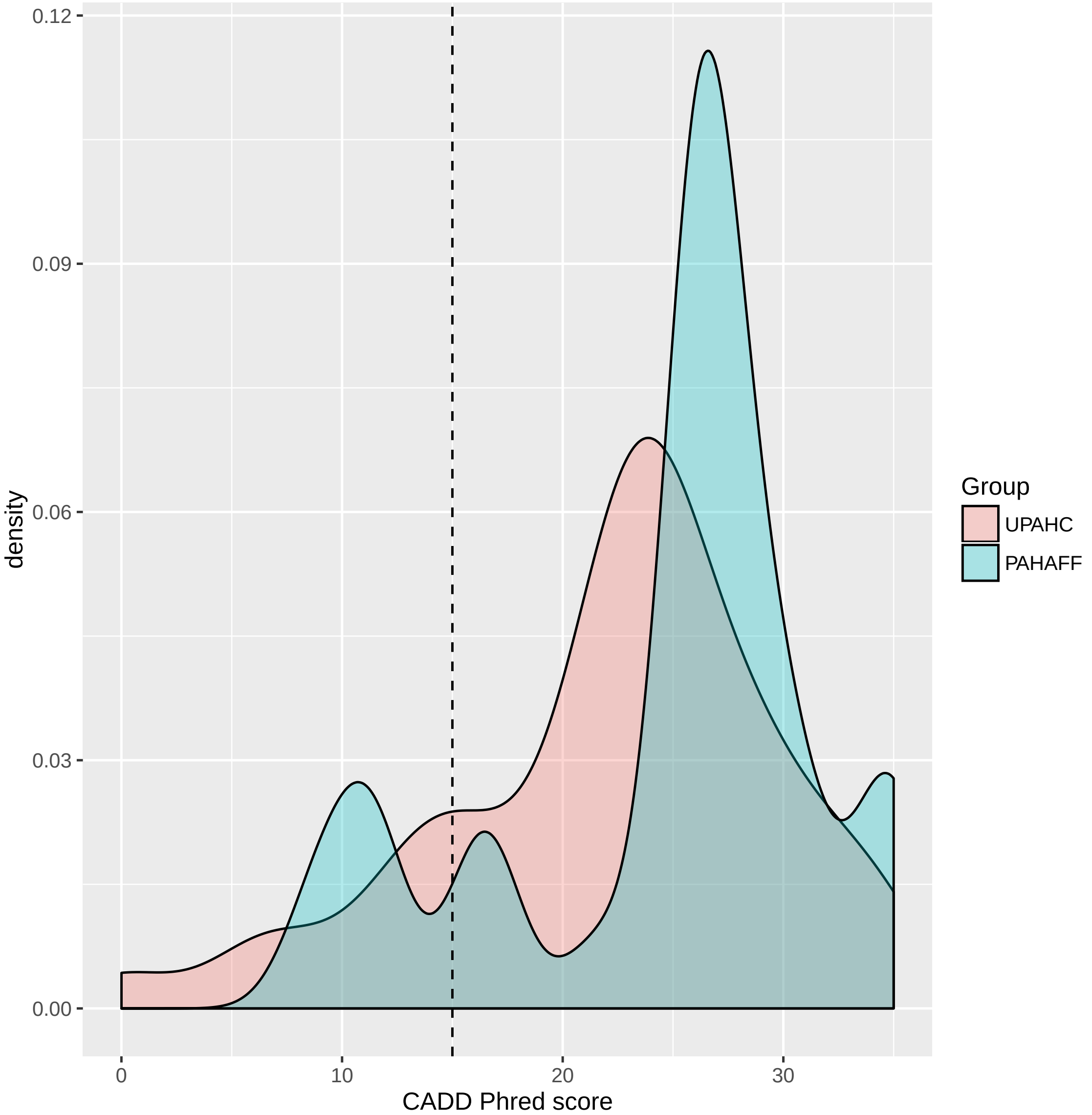


Figure 3.14 Density plot comparing the CADD Phred score distribution for *BMPR2* missense variants in affected PAH (PAHAFF) and unrelated non-PAH subjects (UPAHC). The line indicates the chosen filter value of 15, aimed to reject likely benign variants and lies between the community accepted values of 10 and 20.

annotation of cap analysis of gene expression (CAGE) for transcription start sites (TSS), genomic evolutionary rate profiling (GERP) elements for conserved regions and ENSEMBL regulatory build for regions involved in gene regulation (see Fig. 3.15).

The ENSEMBL regulatory build (see Tab. 3.2 on page 51) included many different cell types. For a more focused analysis, the human umbilical vein endothelial cells (HUVEC) specific dataset was downloaded to represent a PAH relevant cell type and to retrieve a cell type specific feature status. The feature status describes the activity levels of a region with epigenetic signature for ACTIVE (active epigenetic signature), POISED (potential to be activated), REPRESSED (repressed), INACTIVE (no epigenetic modifications) and NA (no available data). Open chromatin and promoter feature (see Fig. 3.15) were annotated as INACTIVE and ACTIVE respectively in HUVEC. On closer inspection of the features, the identifiers in the release did not match with the displayed features from the ENSEMBL website. The source for the HUVEC experiments were listed on the ENSEMBL website and were provided by the encyclopedia of DNA elements (ENCODE) project (Consortium, 2012). The DNase-seq HUVEC experiment in ENCODE specifically supported the open chromatin feature in ENSEMBL while the broader histone modification marker H3K27ac covered both features (see Fig. 3.15). A narrower region compared to H3K27ac would be beneficial to locate smaller functional regions in order to locate likely disease causing variants. Additional supporting experiments from ENCODE for the region is not shown. Chromatin state segmentation analysis for HUVEC based on ENCODE aggregated all experiments and annotated the same regions as ‘strong enhancer’ and ‘active promoter’ respectively (Ernst et al., 2011). The HUVEC specific activity status for the open chromatin feature was discordant between ENSEMBL and ENCODE. Further work is required to establish an unambiguous dataset of a HUVEC or other relevant cell types that includes the latest available data for an in-depth analysis.

The Fig. 3.15 also shows rare (filtered using MAF) and CADD (filtered by MAF and CADD score) variants in cases and controls. These variants were close to regions with TSS and conserved regions. The number of variants for PAH affected cases and unrelated PAH controls are listed in Tab. 3.7 and does not identify an over-representation of variants. All variants filtered by CADD overlap conserved regions and highlights the composition of the CADD score, which includes GERP scores amongst other conservation scores. For comprehensive CADD annotation, the 211 rare variants (45 variants had no annotated CADD score) were submitted to the CADD scoring service for reannotation. The updated CADD scores identified one additional variant with a score $\geq$ 15 for a control subject. In total, the 5 Kb upstream region contained 211 rare variants of which 27 variants had a CADD score $\geq$

15 and comprised 5 PAHAFF cases and 21 UPAHC (one subjects found with two variants) controls.

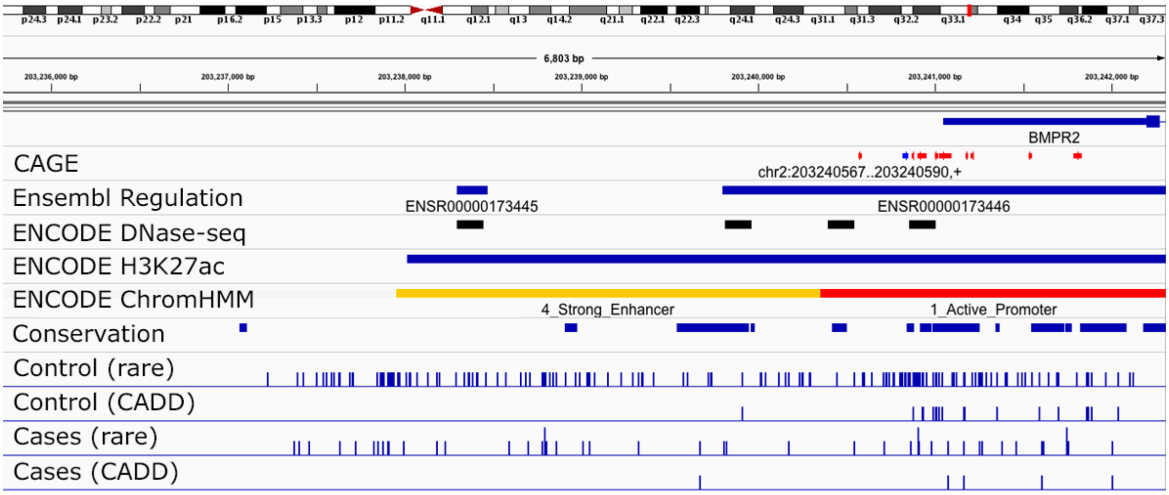


Figure 3.15 Rare deleterious variants in region upstream of *BMPR2* for cases and controls. Variants were filtered by 1 in 10,000 in unrelated PAH controls (rare) and retained variants with a CADD score $\geq$ 15 (CADD). The CAGE annotation identifies confident (red) and less confident (blue) TSS regions. Ensembl regulation annotation shows a open chromatin region followed by a promoter region overlapping with the 5' UTR region of the *BMPR2* gene model. HUVEC specific ENCODE data for DNase-seq (open chromatin), activation (H3K27ac) and chromatin state segmentation provide support for the Ensembl regulatory features.

Region	#PAH affected	#unrelated Controls	Fisher's exact p-value
Promoter	4	20	0.7665
Regulatory features	4	20	0.7665
Conservation	5	20	0.3853
TSS	1	6	1
3 Kb upstream	5	20	0.3853
5 Kb upstream	5	20	0.3853

Table 3.7 Number of identified cases and controls of rare variants filtered by CADD score. The rows provide the count for the number of unique subjects containing a variant in different regions. No significant enrichment was found.

BMPR2 copy number variation

Structural variation at the *BMPR2* locus has been described in previous studies (Machado et al., 2015). The copy number variation (CNV) analysis assessed deletions of exons encoding the *BMPR2* gene in the NIHR BR-RD cohort. The complementary tools `Canvas` and `Manta` were applied that use the read coverage and read alignment information, respectively, to call deletions. Confident deletion calls were identified with a 'PASS' flag in the VCF file, while low quality calls receive other tool and context specific flags. Our conservative approach considered deletions called by both tools, filtered for deletions with a NIHR BR-RD frequency of less than 1 in 1000 (n=9.11), and removed previous published deletions. Using this approach, I identified deletions in 23 PAH cases and no controls that at least partially overlap one or more *BMPR2* exons. Every exon of *BMPR2* was deleted in at least one subject and we identified 5 subjects carrying deletions that covered the entire *BMPR2* gene region shown in Fig. 3.16. For the *BMPR2* locus, the same deletion called by `Canvas` and `Manta` overlapped on average 93% and 95% respectively with each other and all breakpoints were located in the intronic space, except for one `Canvas` start position.

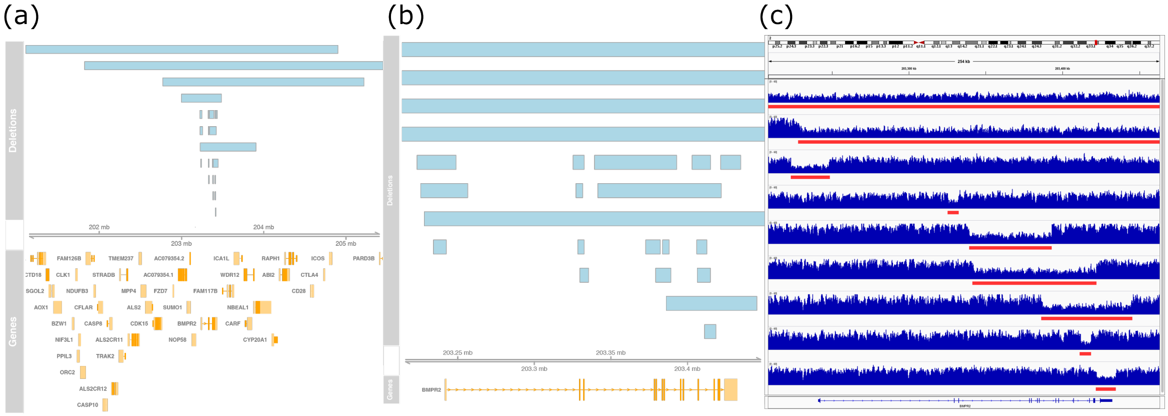


Figure 3.16 Deletions overlapping *BMPR2* protein coding exons. (a) Identified 23 subjects with deletions overlapping *BMPR2* exon. The size of the deletions ranged from deleting a single *BMPR2* exon (4.1 Kb deletion) to deleting 37 protein coding genes (3.7 MB deletion). (b) *BMPR2* focused view shows deletions affecting one or more exons. (c) IGV screenshot highlighting the change in read coverage for selected samples with deletions.

After this conservative analysis, further investigation revealed low quality and tool specific deletions for `Canvas` (cases=7, control=2) and Manta (cases=41, control=189) with an average size of 12 Kb and 125 MB respectively. I decided to focus first on the `Canvas` only calls due to the average length. Seven additional PAH subjects presented deletions called by Canvas only that overlapped with protein coding regions. One control subject covered part of the 5' UTR and another control subject overlapped the end of the 3' UTR shown in Fig. 3.17. The last *BMPR2* exon was covered by three deletions which had more than 97% overlap between each other and near identical start and end positions. These individuals had no declared relatedness and were genetically not closely related (3rd degree or more distant) to each other. However, I identified an affected PAH sibling for one of these subjects and both siblings had deletions of the last *BMPR2* exon with 94% overlap. Only the deletion of one sibling was part of the strict deletion set, while the deletion of the other sibling was filtered out. The inspection of the coverage and alignment information in Fig. 3.18 suggested the correctness of both deletions and a co-inheritance. The four individuals with a deletion of the last exon (1 strict, 3 `Canvas`) were sequenced at different times on different chemistries, but recruited by the same centre. This suggests the possibility of distant relationships between these cases, rather than technical artifacts related to read length.

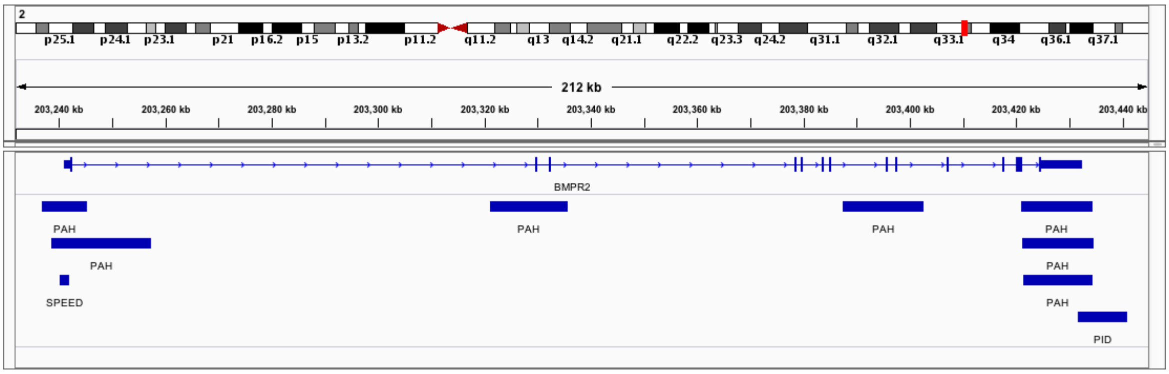


Figure 3.17 Low quality deletions called by `Canvas` overlapping the *BMPR2* loci. Three PAH cases had similar start and end positions covering the last exon of the gene. Control subjects did not overlap with protein coding regions of *BMPR2*.

The discovery of 7 additional cases using `Canvas` prompted a further analysis of the `Manta` data set. `Manta` also called 4 of the 7 additional `Canvas` cases with low confidence. The 'MGE10kb' filter flag of these calls indicates low quality due to a length greater than 10 Kb, because the `Manta` algorithm is optimised for SV and INDELs of up to 10 Kb. The precision decreases beyond 10 Kb length and explains the filter flag (Chen et al., 2016b). However, the 3 additional `Canvas` cases were of similar size and below the 10 Kb limitation. While `Canvas` uses read depth information to detect deletions and duplications (see Fig. 3.19(a)), `Manta` relies on changes in paired-end distance and split reads (partially mapping the same read on both ends of a deletion), illustrated in Fig. 3.19(b). The read alignment showed a drop in coverage for both subjects while the split reads on both ends of the deletion were specific to 3.19(b) in the same region. The subject without a `Manta` call was sequenced using 125bp chemistry rather than the 150bp chemistry, but only the 150bp chemistry resulted in confident `Manta` calls.

This analysis of inconsistencies between two variant callers highlights the deficiencies of the algorithms, the dependency on the alignment tool and ultimately the read length and mappability of reads surrounding the breakpoints of a deletion. The lack of confident variant

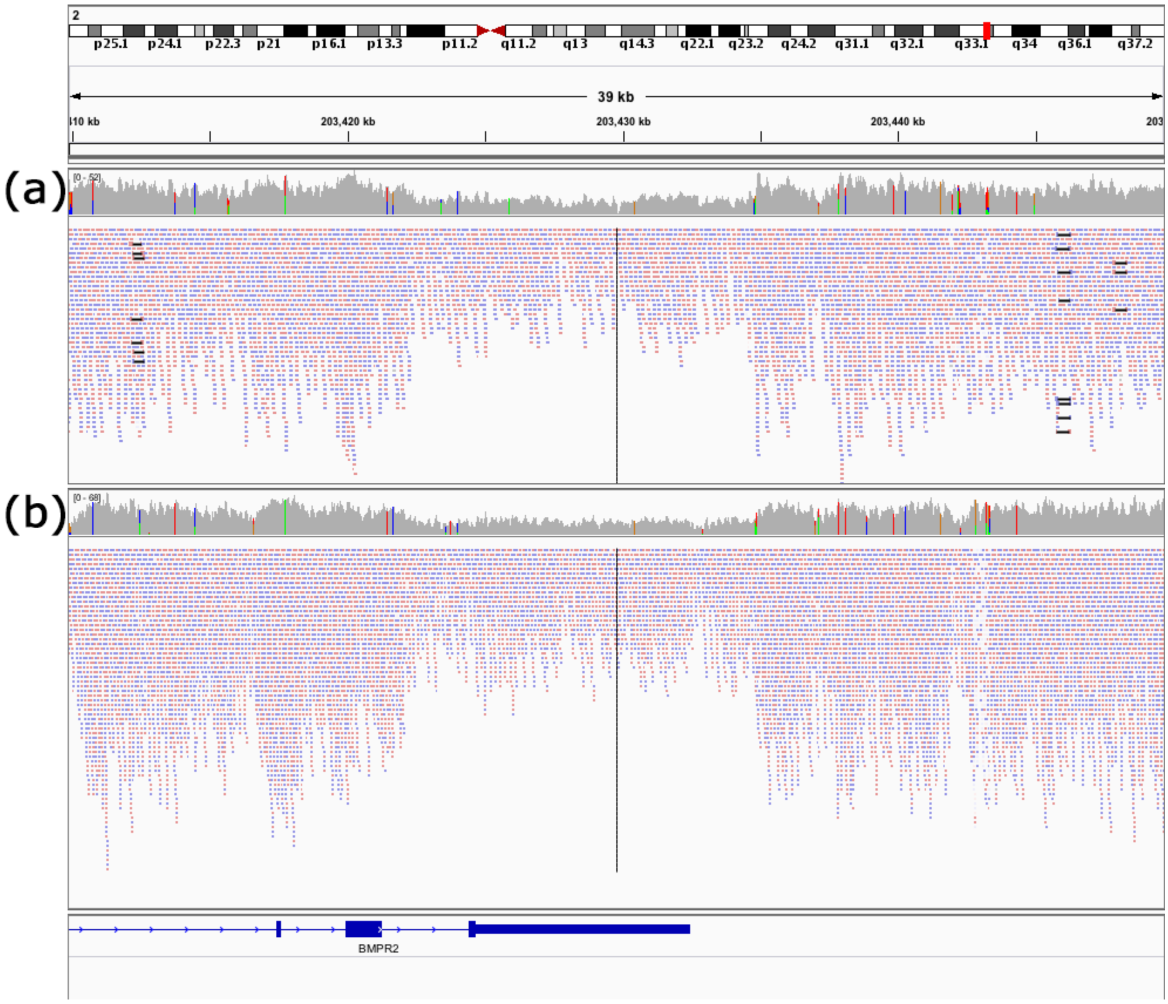


Figure 3.18 Whole genome read alignment with a read coverage plot of siblings covering the last three exons of *BMPR2*. Subject (b) was called by `Manta` and `Canvas` while subject (a) was only called by `Canvas` with low confidence. Both coverage and alignment showed the same pattern and suggested a deletion in both subjects. Split read information are missing in both cases to determine the exact breakpoints of the deletion.

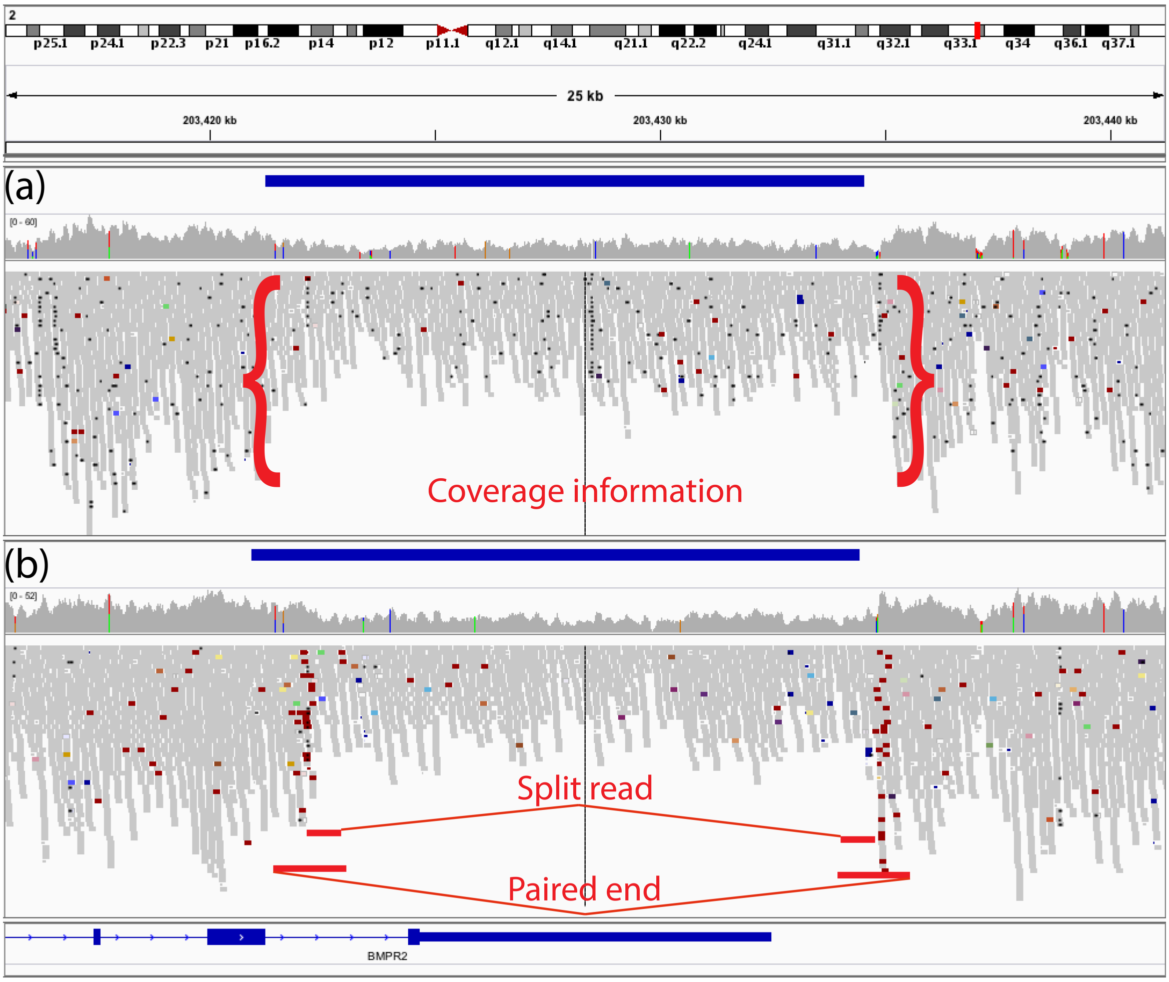


Figure 3.19 Chemistry dependent deletion calls of the last *BMPR2* exon in two subjects. The deletion in (a) is only called by Canvas while (b) is called by Canvas and Manta. Coverage information was used by Canvas highlighted at (a) while split read and paired end information were used by Manta to detect deletions highlighted in (b), which is more reliant on accurate read alignment information. The read length used for (a) was 125bp compared to (b) 150bp and could explain the absence of the Manta call for (a).

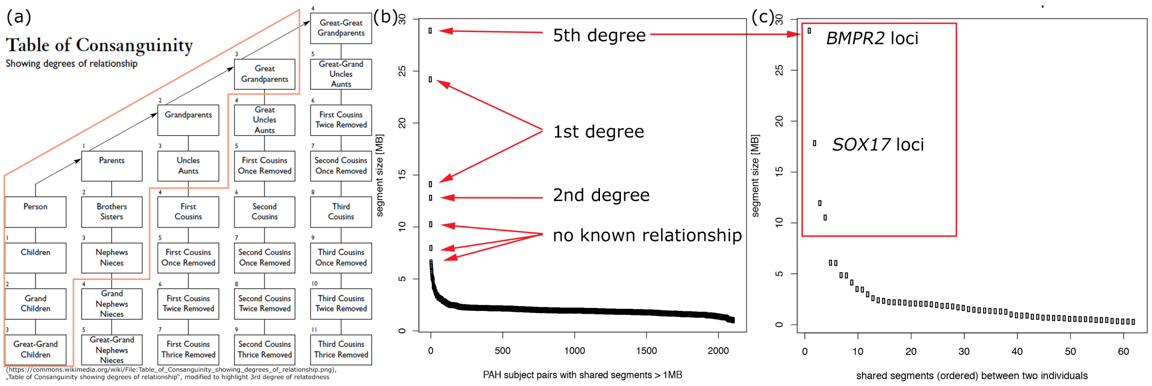
calls by the two algorithms raises further questions about the possibility of missed diagnoses and the ability to distinguish false positive from true positive calls for low quality calls. The conservative approach reduced the number of false CNV deletions and was used for the genome-wide analysis.

Identify distant relationships using shared regions of the *BMPR2* locus

The above discovery of near identical deletions of the *BMPR2* locus raises questions about possible unknown and distant relatedness in PAH families. The identification of relationships was currently based on the analysis of SNPs across the whole genome and limited to 3rd degree relationships shown in Fig. 3.20(a). To identify subjects with 4th degree relatedness and beyond, I analysed the shared segments of the genome in pairs of subjects using identity-by-descent (IBD). The analysis was performed on 6,224 subjects part of the NIHR BR-RD variant release identified as 20160212-A and selected 2.4M SNVs with a 'PASS' flag and an AF greater than 0.05 in 1kG. We applied `BEAGLE` to phase the genotypes of the subjects and identify shared segments. The release contained two of the four subjects with a near identical deletion and these subjects were not known to be related. The length of the shared segments were reported between all pairs of individuals shown in Fig. 3.20(b) focusing on the *BMPR2* locus. The accumulation of long shared segments indicated possible relatedness. Fig. 3.20(b) shows the lengths of shared haplotypes around the *BMPR2* locus and subjects with first and second degree relatedness were labelled as reference points. The pair with the largest shared segment had no known relationship and were identified in the foregoing analysis with identical deletions of the last exon of *BMPR2*. Family history records mentioned a possible case of PAH in a grandmother for one, and an aunt for the other subject. On request, the local research nurse recorded an extended family tree and confirmed the 5th degree relationship. In summary, the same deletion was observed in a pair of siblings, a 5th degree relationship to the one of the siblings and one additional subject with no known relatedness. A relationship could not be ruled out for the additional subject due to a lack of recorded family history and no calculated haplotype information available at the time. The length of the segments across the whole genome of the 5th degree relationship is shown in Fig. 3.20(c). The *BMPR2* loci had the longest shared segment followed by a large region on chromosome 8 overlapping with *SOX17*.

BMPR2 intronic deletion

The CNV analysis described above focussed on the deletion of protein coding regions of *BMPR2*. Next, I extended this analysis by focusing on CNV of the *BMPR2* intronic space and applied the same conservative filter strategy (deletions called by `Manta` and `Canvas`).



(https://commons.wikimedia.org/wiki/File:Table_of_Consanguinity_showing_degrees_of_relationship.png), „Table of Consanguinity showing degrees of relationship", modified to highlight 3rd degree of relatedness

Figure 3.20 Detection of distant relationship using shared segments going beyond standard methods. (a) degree of relationships with 3rd degree relationship detected by standard methods highlighted. (b) Length of shared segments between pairs of subjects overlapping the *BMPR2* loci labelled with the degree of relationship. (c) Largest shared segment between pair with confirmed 5th degree relationship overlapped with *BMPR2* followed by *SOX17* loci.

This analysis identified one PAH case compared to four unrelated controls with deletions in introns 1,3 and 10 of *BMPR2*. The longest intron (first intron) contains 60% of the deletions and the deletions were located closer to the start of the first intron. Fig. 3.21 shows regions with deletions in affected cases and control (including one related pair) and focused on the deletion in a PAH sample that overlapped with a regulatory region annotation. The deletion in controls was of a similar size in close proximity to the PAH subject deletion, but does not overlap the regulatory region. This regulatory region was annotated as open chromatin and was specific to human umbilical vein endothelial cells (HUVEC). The rare deletion in the *BMPR2* intronic space possibly explained one case of the PAH disease cohort, but requires functional validation.

BMPR2 mutation burden

The above described analysis identified different variant types in the protein coding and non-coding regions of *BMPR2* that were likely pathogenic. Focusing on rare variants of the *BMPR2* locus, I identified 163 subjects with putative disease causing variants of which 48.4% were PTVs (n=79), 34.3% missense (n=56) and 18.4% deletions of exons (n=30). One intronic deletion was identified but not taken into account due to the unknown impact. There were 2 subjects with two *BMPR2* variants of different consequence types (missense with PTV and missense with exonic deletion). The current analysis of the non-coding upstream region of *BMPR2* identified a further 5 subjects with variants passing the filter strategy, but these remain of uncertain functional significance.

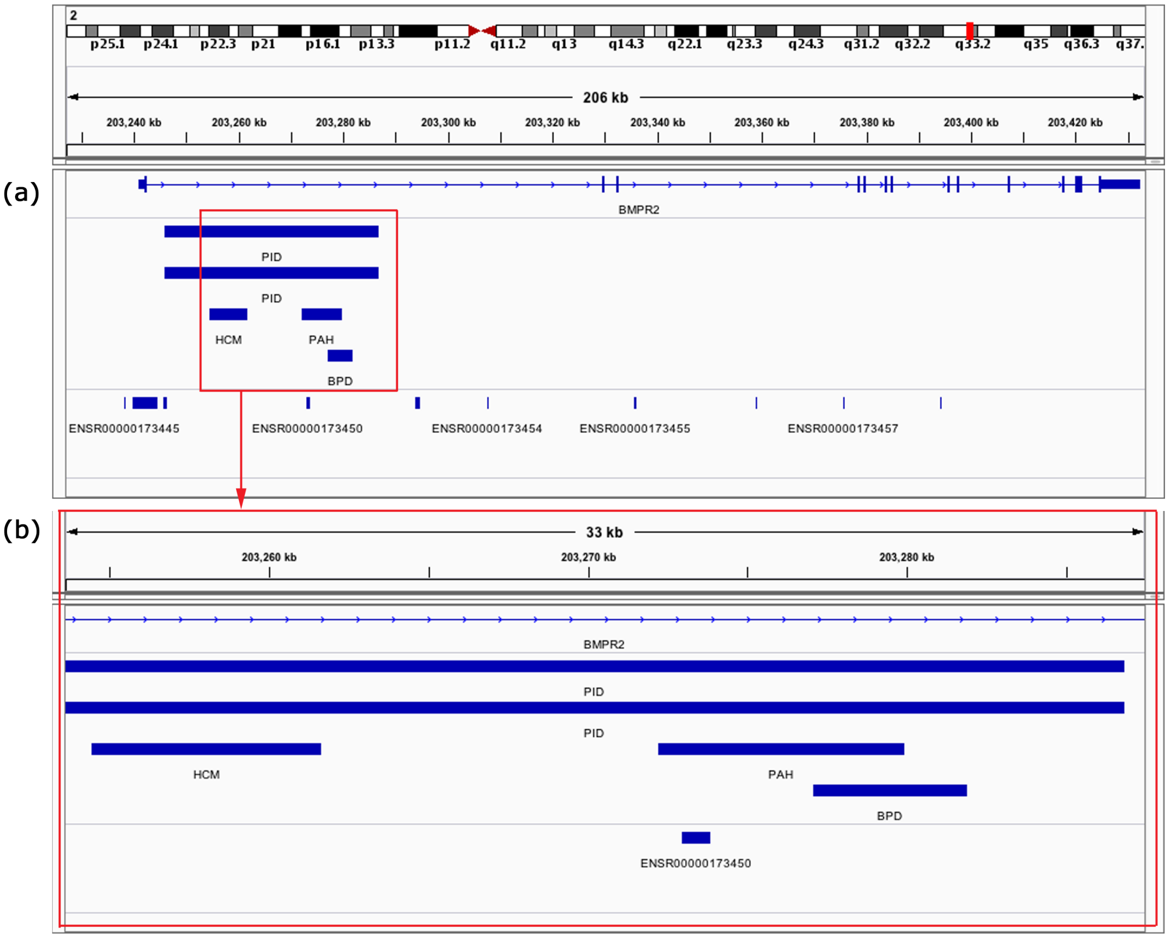


Figure 3.21 PAH specific deletion of open chromatin region. The deletions of (a) affected PAH and control subjects (incl. related controls) are shown for *BMPR2*. Regulatory regions are displayed below the deleted regions as a separate track and (b) focuses on the deletion of one regulatory region. The deleted regulatory region in (b) is specific to PAH except of two larger overlapping deletions from a related pair. The deleted regulatory region was annotated as open chromatin and specific to HUVEC.

Phenotype of *BMPR2* variant carriers

We collected phenotype information at the time of diagnosis for NIHR BR-RD subjects in OpenClinica and had information for 99.3% (n=1041) affected PAH adult subjects available in the OpenClinica release (see 2.3.9 on page 34). These information included 99.4% (n=162) of the subjects with likely disease causing *BMPR2* variants identified above. Due to missing information in the data set, I focused on measurements used for cohort characterisation, classification and diagnosis of PAH. The values included the recorded family history, gender ratio, age at diagnosis, cardiac output, mean pulmonary arterial pressure (mPAP), mean pulmonary arterial wedge pressure (mPAWP), left ventricular end-diastolic pressure (LVEDP), transfer coefficient for carbon monoxide and vasoreactivity. A summary of categorical data is listed in Tab. 3.8 while the distribution of numerical measurements is shown in Fig. 3.22 using standardised z-scores. *BMPR2* variant carriers compared to other affected subjects had a significantly younger age at diagnosis (p=1.5e-12), lower cardiac output (p=2.493e-12) and significantly increased mPAP (p=1.1e-08) and transfer coefficient (p=3.043e-14). The described PAH cohort matched previous reports with a female to male ratio of 2.2 to 1 and a family history of 70.4% for *BMPR2* carriers. The identified subjects with *BMPR2* variants were significantly younger, had more severe phenotypes (lower cardiac output, higher mean pulmonary arterial pressure) and had a significantly higher transfer coefficient (KCO) compared to other subjects diagnosed with PAH. The 5 subjects with non-coding upstream variants had a different phenotype pattern compared to *BMPR2* variant carriers, but the analysis was limited due to the low number of available data.

Measurement	*BMPR2*	other	% complete
Gender ratio (f/m)	2.1:1	2.2:1	99.3% (1041)
Family history	70.4%	29.5%	95.0% (995)
Functional class (I; II; III; IV)	1.3%; 21.7%; 59.8%; 17.1%	2.2%; 20.6%; 66.2%; 10.8%	87.4% (916)

Table 3.8 Classification of identified *BMPR2* variant carriers compared to other affected PAH subjects. Gender ratio, family history and functional classes are categorical data to classify PAH cases and are listed for identified *BMPR2* variant carriers and other affected PAH subjects. The higher female ratio is not specific to *BMPR2* variant carriers. A larger proportion of *BMPR2* cases are reported as functional class IV compared to other cases.

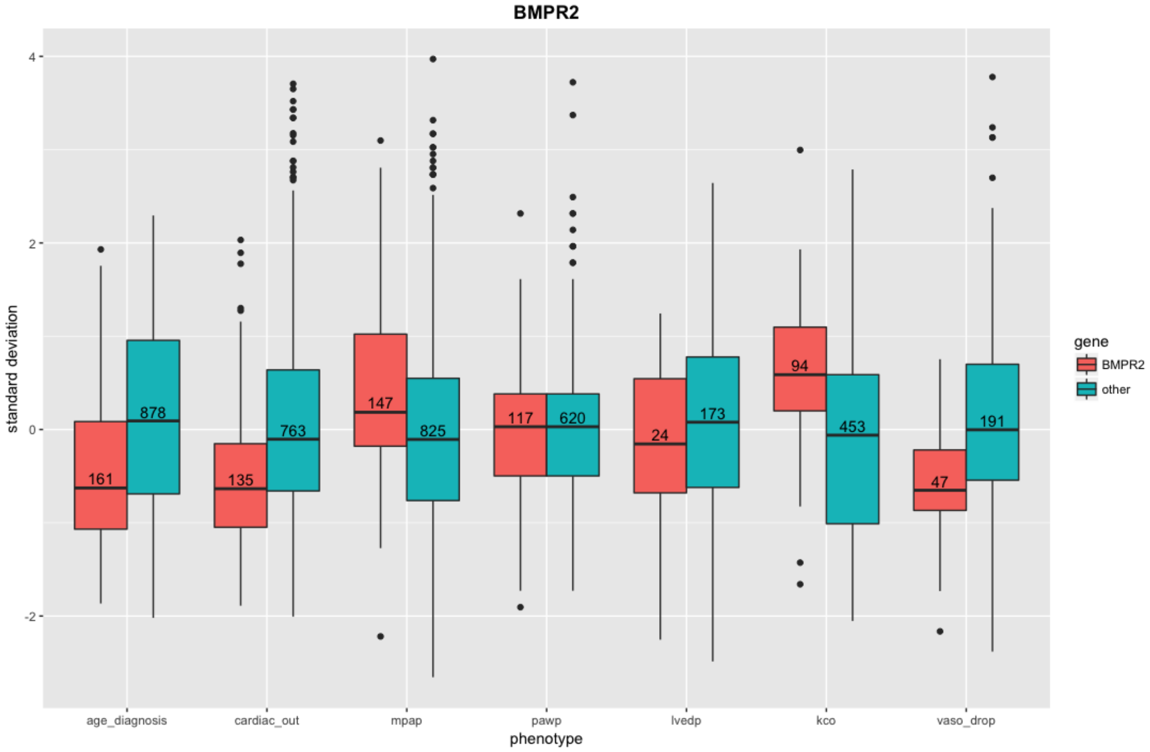


Figure 3.22 Normalised z-score distribution of selected measurements for identified *BMPR2* variant carriers and other affected PAH samples.

3.3.4 Variation in previously reported PAH genes

The chapter 3.3.3 focussed on protein truncating variants (PTV), missense variants and loss of copy number variation (CNV) in the autosomal dominantly inherited PAH disease gene *BMPR2*. My filtering strategy selected extremely rare (<1 in 10,000 or absent in population databases and control cohort) PTV and missense variants predicted to be deleterious (CADD $\geq$ 15 and without ambiguous SIFT and PolyPhen-2 predictions). Subjects with remaining variants in *BMPR2* were counted for each cohorts, affected PAH (1,048 subjects) and unrelated PAH control (6,385 subjects). The affected PAH cohort showed an enrichment of subjects (p-value 2.0E-94) containing the selected rare variants compared to the control cohort (see Tab. 3.6). Rare loss of CNVs (<1 in 1,000 in NIHR BR-RD and absent in population databases) were identified only in PAH cases (23 subjects). No subject in the unrelated PAH control cohort were found to have a loss of CNV overlapping a protein-coding region of *BMPR2*.

The filter strategy (see chapter 3.3.3) was extended beyond *BMPR2* to select likely deleterious variants in previously reported genes in PAH (*BMPR2*, *ACVRL1*, *ENG*, *CAV1*, *SMAD1*, *SMAD4*, *SMAD9*, *KCNK3*, *EIF2AK4*, *TBX4*) plus candidates (*TOPBP1*, *BMPR1B*, *KLF2*). Screened variants included PTV, missense variants and CNV. The number of subjects with one of these filtered variant types were counted per gene and cohort. In the detected variant set, I found PTVs, missense variants and CNV deletions in previously reported and candidate genes. Of the previously reported genes, *BMPR2* contained a significant overrepresentation of deleterious variants in PAH cases compared to unrelated PAH controls (see Fig. 3.23a). Significant higher subject frequency (p-value <0.05) was found in the six previously reported genes *BMPR2*, *EIF2AK4*, *TBX4*, *ACVRL1*, *ENG* and *SMAD9* (see Fig. 3.23b). No significant enrichment of affected PAH cases was found in *KCNK3* and *SMAD1*, with no pathogenic coding variation identified in the affected PAH cohort in *SMAD4* or *CAV1* (see Tab. 3.9). Eight previously reported genes contained likely disease causing variants for 229 subjects of which 6 subjects had variants in two previously reported genes. In four cases, *BMPR2* appeared once in combination with *ENG*, *SMAD1*, *SMAD9* or *EIF2AK4* and one case each of *EIF2AK4* with *SMAD9* and *ENG* with *TBX4*.

For 229 patients, rare and likely deleterious variation was found in at least one of the previously reported disease causing genes. Different levels of evidence were considered according to the guidelines of the American College of Medical Genetics and Genomics (ACMG) (Richards et al., 2015). Population, computation, functional and segregation data are considered as evidence to classify a variant as “pathogenic”, “likely pathogenic”, “uncertain significance”, “likely benign” or “benign”. Based on these categories, I assessed the reported PAH genes with the most and least p-value containing affected PAH samples,

	affected PAH cases					unrelated PAH controls					Fisher's	FDR
Gene	PTVs	Missense	Deletions	Total	frequency	PTVs	Missense	Deletions	Total	frequency	exact p-value	**P**$_{adj}$
BMPR2*	96	39	23	156	0.15	1	18	0	19	2.9E-3	5.3E-114	6.9E-113
EIF2AK4*	12	14	0	22	0.02	6	29	0	35	5.4E-3	3.9E-6	2.5E-5
TBX4*	8	7	0	15	0.01	0	17	0	17	2.6E-3	8.4E-6	3.6E-5
ACVRL1*	1	8	0	9	8.6E-3	2	9	0	11	1.7E-3	8.2E-4	0.003
ENG*	1	7	0	8	7.6E-3	0	18	0	18	2.8E-3	0.02	0.06
SMAD9*	2	5	0	7	6.7E-3	1	15	0	16	2.5E-3	0.03	0.07
TOPBP1#	0	6	0	6	5.7E-3	0	27	0	27	4.2E-3	0.3	0.4
BMPR1B#	0	5	0	5	4.8E-3	1	17	1	19	2.9E-3	0.2	0.3
KLF2#	0	5	0	5	4.8E-3	1	9	0	10	1.5E-3	0.05	0.09
KCNK3*	0	5	0	5	4.8E-3	0	13	0	13	2.0E-3	0.09	0.1
SMAD1*	1	1	0	2	1.9E-3	1	11	0	12	1.8E-3	0.6	0.7
SMAD4*	0	0	0	0	0	0	5	0	5	7.8E-4	1	1
CAV1*	0	0	0	0	0	0	5	0	5	7.8E-4	1	1

Table 3.9 Number of affected PAH subjects and unrelated controls with variants in previously reported (*) and putative (#) PAH genes grouped by variant type. Deletions called by Manta and Canvas and supported by at least one confident call were included.

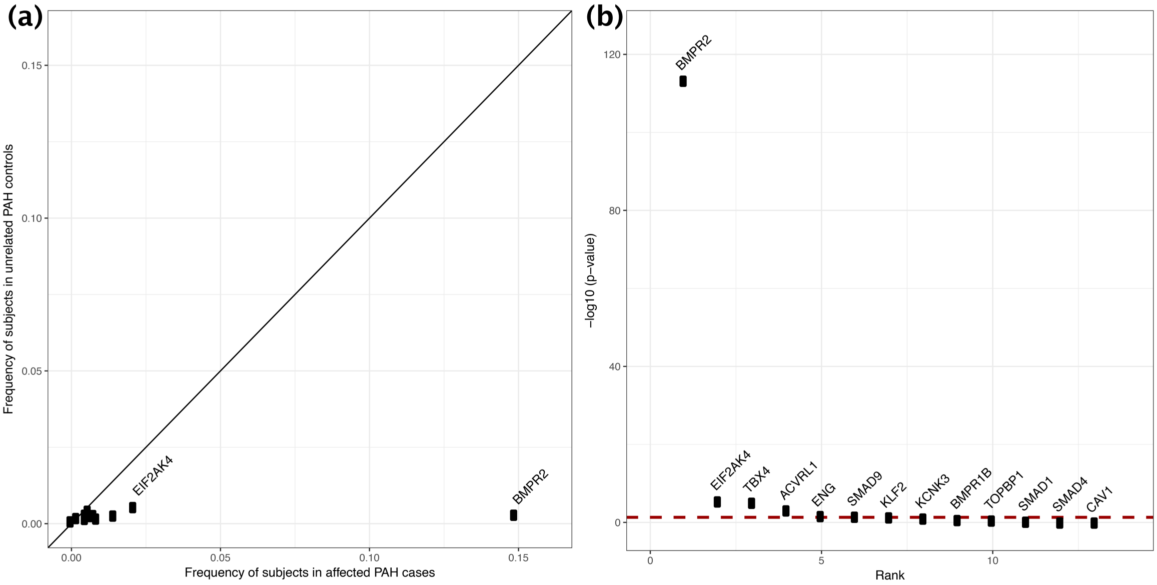


Figure 3.23 Enrichment of affected PAH cases with filtered variants in reported PAH genes. (a) The frequency of subjects for reported PAH genes are displayed for affected PAH cases and unrelated non-PAH control cohort. Each gene is represented by one point and labels are only shown for the two genes with the highest frequency in the affected PAH cohort due to space. (b) The number of subjects with filtered variants in each gene was tested for enrichment in affected PAH cases using Fisher's exact test. Ordered by significance, the negative decadic logarithm of unadjusted p-values is plotted for each gene. The p-value of 0.05 is indicated by the red line.

BMPR2 and *SMAD1* respectively. All PTVs in *BMPR2* were classified as “pathogenic” and supported by the variant type, the prevalence in affected individuals and absent of the variant in control cohorts. The LoF intolerance (pLI) and haploinsufficiency (%HI) are indicators that a single functional copy of a gene is likely insufficient to maintain normal function. The pLI ranges from 0-1 and a gene with a pLI >0.9 is extremely intolerant of loosing a copy, while %HI ranges from 0-100 and a gene with a %HI <10 is likely to exhibit haploinsufficiency. For *BMPR2*, the pLI (1.00) and %HI (1.47) strongly indicated the loss of normal function. Missense variants in *BMPR2* were also classified as “pathogenic” or “likely pathogenic” depending on whether the variant has been identified previously (case reports) in the same disease context, supported by the prevalence in affected individuals, well-established functional studies (Nishihara et al., 2002), absent of the variant in control cohorts and multiple lines of computational evidence (PVS1, PS3, PS4, PM2, PM4, PM6, PP2, PP3, PP4). In the gene *SMAD1* I identified a PTV and a missense variant. The subject carrying the PTV also had a PTV reported in *BMPR2* and was therefore not further considered. For *SMAD1*, the pLI (0.87) and %HI (13.31) were close to the suggested cut-offs and showed an indication for a loss of normal function. The missense variant was absent in control cohorts and supported by multiple lines of computational evidence. The patient’s diagnosis matches the initial case reports (Nasim et al., 2011; Shintani et al., 2009) of rare sequence variation in BMP signalling intermediaries, which provide additional evidence for a central role of dysregulated BMP signalling in PAH pathogenesis. In consideration of these evidence items, the resulting ACMG classification for the missense variant in *SMAD1* was “likely pathogenic” (PM2, PM6, PP2, PP3, PP4). The PTVs in *SMAD1* was classified as “uncertain significance” due to contradicting evidence, while the PTV in *BMPR2* in the same subject would still be regarded as “pathogenic”.

From the previously reported genes, PTVs and missense variants in the most significantly enriched gene were classified as “pathogenic” while the least significantly enriched gene classified missense variants as “likely pathogenic” and PTVs as “uncertain significance”. Future re-evaluation with new evidence could change the classification of *SMAD1* variants. For subjects with variants in previously reported genes, the strongest variant classification was either “pathogenic” or “likely pathogenic”, which provides a greater certainty that the disease in 229 patients is caused by rare deleterious variants in previously reported genes.

Classification of variants based on the ACMG standards and guidance provided greater certainty about the disease causing affect of the filtered variants in previously reported PAH disease genes. To describe the phenotypic differences, I extracted information for patients with variants in previously reported as well as putative genes, grouping patients by gene. The low number of subjects identified with variants per gene limited possible analyses in combination with the completeness of phenotype information (see chapter 2.3.9 on page 34). Despite these limitations, I focused on the phenotypes of age at diagnosis and transfer coefficient for carbon monoxide shown in Fig. 3.24. The youngest group had a median age of 26.4 and these subjects had variants in *KCNK3*, followed by *EIF2AK4* (median=37.6) and *BMPR2* (median=39.3). The oldest group was *SMAD1* (median=60.8). In contrast, the transfer coefficient was reduced in *EIF2AK4* (median=0.69) variant carriers compared to *BMPR2* (median=1.39). The single measurements for *KCNK3* and *SMAD1* also highlighted the issue of data completeness. The current method used the genotype information to identify distinct phenotypes, which either confirmed the diagnosis or could be used in future to distinguish between different subtypes of PAH or related diseases.

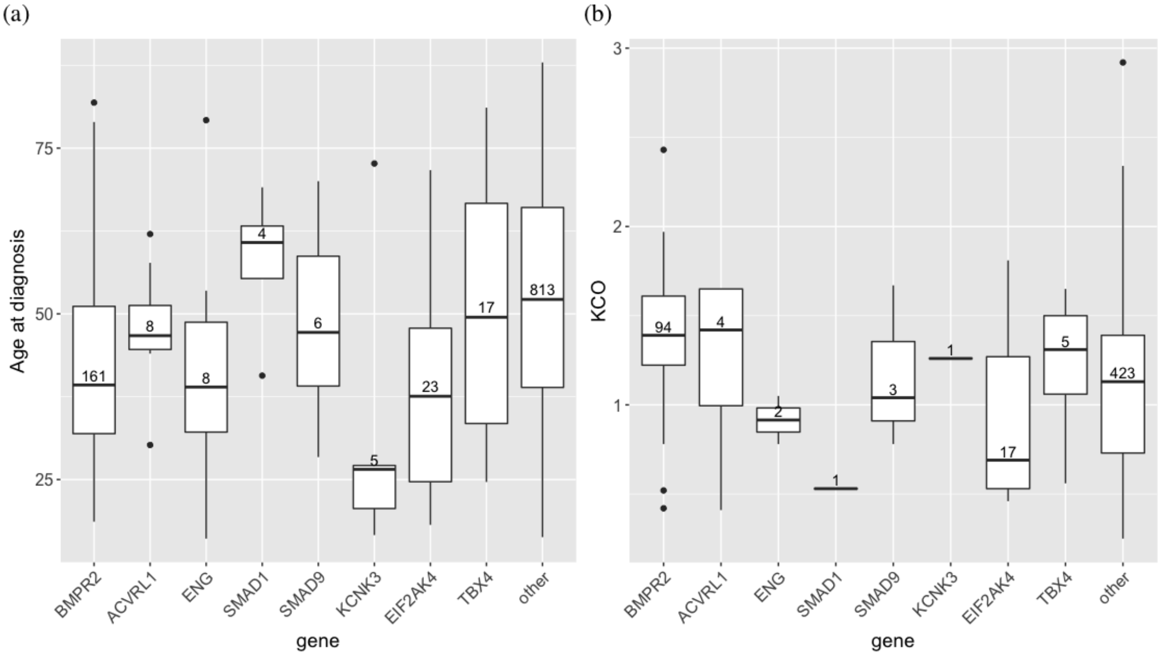


Figure 3.24 Distribution of age at diagnosis (a) and transfer coefficient for carbon monoxide (b).

3.3.5 Identification of novel disease-gene associations

The analysis selects rare protein truncating and missense variants from the aggregated VCF file. Consequence type-dependent filtering strategies group variants in a gene for cases and control subjects and test the gene for association with the disease. I performed the analysis on 1,038 PAHIDX cases and 6,385 UPAHC controls. In addition, 220 PAH cases were identified with variants in previously reported genes and removed from the additional analysis performed on 818 unrelated PAH index cases and 6,385 UPAHC controls. The same consequence type-dependent filtering strategy, developed and validated during the analysis of *BMPR2* variants, was applied to enrich for rare causal variants in protein coding genes. The present analysis selected subjects from the PAHIDX and UPAHC cohorts for a case-control comparison to identify novel PAH disease genes.

Protein truncating variation

In the first analysis I focused on likely high impact variation represented by PTVs, which, if present, are likely to be rare and deleterious events. A genome-wide analysis applied the defined MAF filter of 1 in 10,000 and selected PTVs in protein coding genes of canonical transcripts. The number of subjects with variants in PAHIDX and UPAHC were identified in each gene and tested for over-representation in PAH cases. This analysis identified an over-representation in *BMPR2*, *TBX4* and *EIF2AK4* with genome wide significance ($\mathbf{P}_{adj}$ <0.05) and an over-representation of PTVs in *ATP13A3*, *EVI5*, *SRM*, *KDR* and *PRR22*, compared to zero or one PTV identified in these genes in control subjects (see Fig. 3.25 and Tab. 3.10). The result confirmed the large genetic contribution of protein coding variants in *BMPR2* to PAH. The analysis also highlighted that beyond *BMPR2*, that additional genes with high impact rare variants are confined to small numbers of unrelated PAH cases.

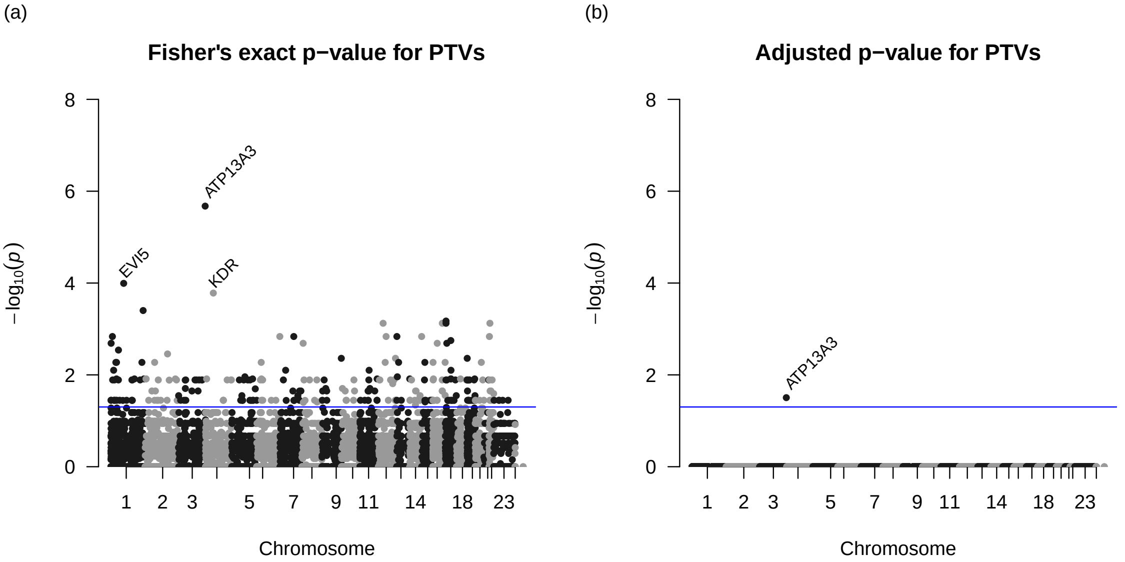


Figure 3.25 Manhattan plots of the PTV analysis, having excluded cases carrying rare variants in previously established PAH genes. Filtered variants were grouped per gene and tested for an excess of variants in PAH cases. The (a) Fisher's exact p-values and (b) adjusted p-values are plotted against the chromosomal location of each gene. The blue horizontal line indicates a p-value of 0.05. Chromosome X and Y are encoded as 23 and 24 respectively.

During this analysis we identified some individuals in the UPAHC group that carried rare PTVs in *BMPR2*, *EIF2AK4* and *EVI5*. One *BMPR2* frameshift variant was found in the control cohort, but this was located within the last exon of the gene. No PTVs were identified within the last exon for PAH cases and is thus likely to have low impact. Of the 12 cases with PTVs in *EIF2AK4*, 4 were identified with homozygous variants and 3 with biallelic variants, while UPAHC subjects only presented single heterozygous variants. The ethnicity of the PAH individuals carrying the 7 homozygous or two heterozygous variant cases were South-Asian (n=3), European (n=2) and other (n=2). The 6 control subjects were all Europeans. The control subject listed for *EVI5* included a stop gained variant at the carboxy terminus of the protein, while all PTVs reported in cases occurred before this position.

In an additional analysis, I excluded 220 (21.2%) cases identified earlier with likely causal variants in previously reported genes (*BMPR2*, *ACVRL1*, *ENG*, *SMAD1*, *SMAD9*, *KCNK3*, *EIF2AK4* and *TBX4*) and assessed the effect on the novel disease gene discovery (see Tab. 3.10). Following exclusion of these cases, *ATP13A3* reached genome wide significance ($\mathbf{P}_{adj}$ <0.05), whereas *SRM* and *PRR22* lost 1 and 2 cases respectively. The number of cases remained the same for *ATP13A3*, *EVI5* and *KDR*. Despite the low numbers of cases, for *SRM* and *PRR22* no control subjects were found with PTVs. We found individuals with combinations of PTVs in *SRM* with *BMPR2*, *PRR22* with *ENG* or *TBX4* each in one individual.

Missense variation

For the next analysis, I focused on moderate impact variation represented by missense variants, where the impact of variation on the encoded protein is less certain to be deleterious. The genome-wide analysis filtered out variants based on MAF and on ambiguous consequence predictions, as discussed before. A comparison between PAHIDX and UPAHC revealed a statistically significant ($\mathbf{P}_{adj}$ <0.05) higher frequency of cases with variants in *BMPR2*, *GDF2* and *TXNRD3*. Compared to the PTV analysis, the number of cases and controls carrying missense variants were both much higher and shows the challenge of developing an effective filtering strategy to select disease causing missense variants. Genes with a Fisher p-value <0.0001 are listed in Tab. 3.11 and also includes *FLNA*, which is located on the X chromosome (see Fig. 3.26).

Variants in *AQP1*, *C3orf20* and the X chromosome gene *FLNA* were overrepresented in PAH index cases (see Tab. 3.11). Exclusion of cases with mutations in previously reported genes reduced the number of variant carriers in all genes except *GDF2*. The adjusted p-value after the exclusion was not genome wide significant for *TXNRD3*. The coexistence

	All PAH index cases				Removed identified cases from PAH index			
HGNC	Cases	Controls	Fisher's exact p-value	$\mathbf{P}_{adj}$	Cases	Controls	Fisher's exact p-value	$\mathbf{P}_{adj}$
BMPR2	92 (8.86%)	1 (0.02%)	5.6E-79	8.4E-75				
TBX4	8 (0.77%)	0 (0.00%)	1.4E-07	2.1E-03				
EIF2AK4	12 (1.16%)	6 (0.09%)	4.3E-07	6.5E-03				
ATP13A3	6 (0.58%)	0 (0.00%)	7.4E-06	0.1	6 (0.73%)	0 (0.00%)	2.1E-06	0.03
EVI5	5 (0.48%)	1 (0.02%)	2.8E-04	1	5 (0.61%)	1 (0.02%)	1.0E-04	1
KDR	4 (0.39%)	0 (0.00%)	3.0E-04	1	4 (0.49%)	0 (0.00%)	2.6E-04	1
SRM	4 (0.39%)	0 (0.00%)	3.0E-04	1	3 (0.37%)	0 (0.00%)	1.4E-03	1
PRR22	4 (0.39%)	0 (0.00%)	3.0E-04	1	2 (0.24%)	0 (0.00%)	1.2E-02	1

Table 3.10 Results of PTV analysis listing PAH disease-gene associations. Only the top genes are listed for displaying purposes with a Fisher's exact p-value <0.0005 including all PAHIDX cases. The updated values are shown for PAHIDX cases without subjects in previously reported genes. Genes in bold reach a $\mathbf{P}_{adj}$ <0.05.

	All PAH index cases				Removed identified cases from PAH index			
HGNC	Cases	Controls	Fisher's exact p-value	$\mathbf{P}_{adj}$	Cases	Controls	Fisher's exact p-value	$\mathbf{P}_{adj}$
BMPR2	32 (3.08%)	18 (0.28%)	4.3E-16	8.1E-12				
GDF2	11 (1.06%)	5 (0.28%)	8.5E-07	0.02	11 (1.34%)	5 (1.34%)	9.7E-08	1.8E-03
TXNRD3	15 (1.06%)	13 (0.20%)	8.7E-07	0.02	7 (0.86%)	13 (0.20%)	4.9E-03	1
AQP1	8 (1.45%)	4 (0.06%)	4.2E-05	0.8	6 (0.73%)	4 (0.06%)	2.9E-04	1
FLNA	19 (0.77%)	35 (0.06%)	7.1E-05	1	14 (1.71%)	35 (0.55%)	7.9E-04	1
C3orf20	12 (1.83%)	14 (0.55%)	7.6E-05	1	9 (1.10%)	14 (0.22%)	5.6E-04	1

Table 3.11 Results of the missense variant analysis listing PAH disease-gene associations. Only for displaying purposes, genes with a Fisher's exact p-value <0.0001 for all PAH index cases are listed and the results after the exclusion of cases carrying variants in previously reported genes. Genes in bold reach a $\mathbf{P}_{adj}$ <0.05 and include *BMPR2*, *GDF2* and *TXNRD3*.

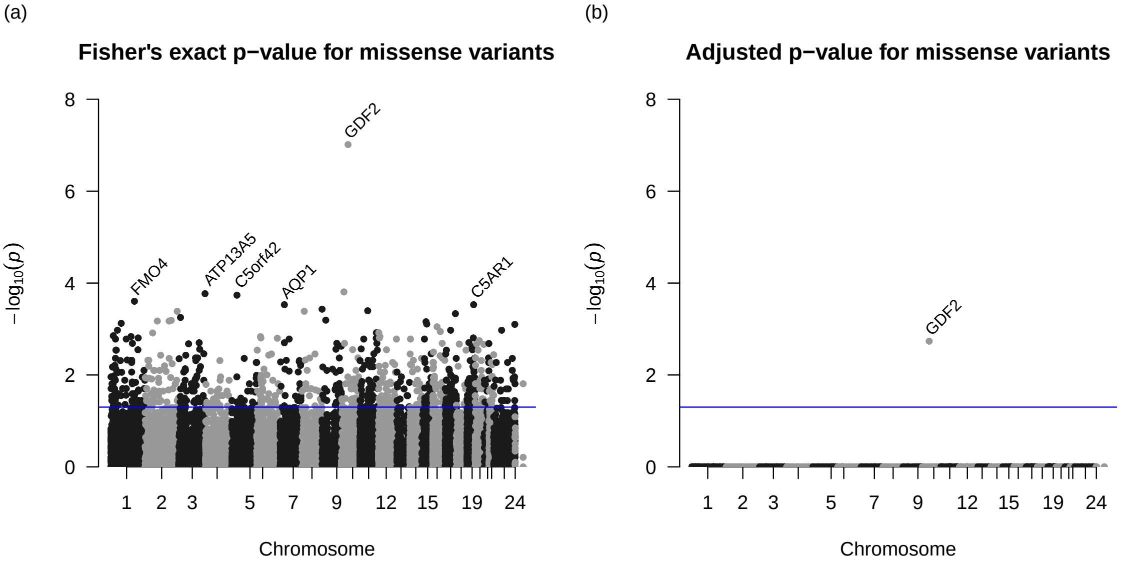


Figure 3.26 Manhattan plots of the missense variant analysis, having excluded cases carrying rare variants in previously established PAH genes. Filtered variants were grouped per gene and tested for an excess of variants in PAH cases. The (a) Fisher's exact p-values and (b) adjusted p-values are plotted against the chromosomal location of each gene. The blue horizontal line indicates a p-value of 0.05. Chromosome X and Y are encoded as 23 and 24 respectively.

of missense variant in putative genes with variants in previously reported genes was more prominent compared to PTVs.

Combined analysis

The combined analysis included all identified missense and PTV variants. A comparison between PAHIDX and UPAHC revealed a statistically significant ($\mathbf{P}_{adj}$ <0.05) higher frequency of cases with variants in *BMPR2*, *GDF2* and *TBX4* (see Fig. 3.27 and Tab. 3.12). The Fisher's exact p-value ranking suggests a separation between genes, which were mainly affected by PTVs (*KDR*, *PRR22*, *EVI5*), or missense (*GDF2*, *FLNA*, *TXNRD3*) or both variant types (*BMPR2*, *TBX4*, *EIF2AK4*). The gene symbol *AQP1* is listed twice caused by the assignment to two different Ensembl gene models in GRCh37. The Ensembl gene identifiers were ENSG00000240583 and ENSG00000250424 for the higher and lower ranked *AQP1* entry respectively. Exons shared between both models contained rare variants included in the analysis. The *AQP1* (ENSG00000240583) variant 7:30962212_C/T was present in 5 PAH subjects before and 3 PAH subjects after removing samples carrying variants in previously reported genes.

HGNC	Cases	Controls	Frequency cases	Frequency controls	Fisher’s exact p-value	$\mathbf{P}_{adj}$
BMPR2	124	19	0.1	3.0E-03	2.6E-87	4.95E-83
GDF2	12	6	0.01	9.4E-04	4.3E-07	8.4E-03
TBX4	16	17	0.02	2.7E-03	2.1E-06	0.04
EIF2AK4	22	35	0.02	5.5E-03	3.4E-06	0.07
TXNRD3	15	16	0.01	2.5E-03	4.6E-06	0.09
AQP1	9	5	8.7E-03	7.8E-04	2.0E-05	0.4
FLNA	20	36	0.02	5.6E-03	3.6E-05	0.7
ALPPL2	12	14	0.01	2.2E-03	7.6E-05	1
C3orf20	13	17	0.01	2.7E-03	8.6E-05	1
ATP13A3	11	14	0.01	2.2E-03	2.6E-04	1
IFT74	12	17	0.01	2.7E-03	2.7E-04	1
ARMCX2	5	1	4.8E-03	1.6E-04	2.8E-04	1
OR8U1	16	30	0.02	4.7E-03	3.0E-04	1
OTUB1	4	0	3.9E-03	0	3.8E-04	1
SPTBN5	40	128	0.04	2.0E-02	3.8E-04	1
AQP1	11	15	0.01	2.3E-03	3.9E-04	1
SPTBN1	18	38	0.02	6.0E-03	4.0E-04	1
SRM	6	3	5.8E-03	4.7E-04	4.2E-04	1
PIWIL1	9	10	8.7E-03	1.6E-03	4.8E-04	1
AC068533.7	9	10	8.7E-03	1.6E-03	4.8E-04	1
KDR	10	19	9.6E-03	3.0E-03	4.4E-03	1
PRR22	6	8	5.8E-03	1.3E-03	8.1E-03	1
EVI5	8	17	7.7E-03	2.7E-03	0.02	1

Table 3.12 Results of the combined analysis (PTV and missense) listing genes with a Fisher p-value <0.0005 or identified in the foregoing analyses. Rows highlighted in orange are genes identified in the separate PTV and missense analyses described above, and genes listed below the horizontal line are of no specific order. Genes in bold letters reach a $\mathbf{P}_{adj}$ <0.05 and include *BMPR2*, *GDF2* and *TBX4*.

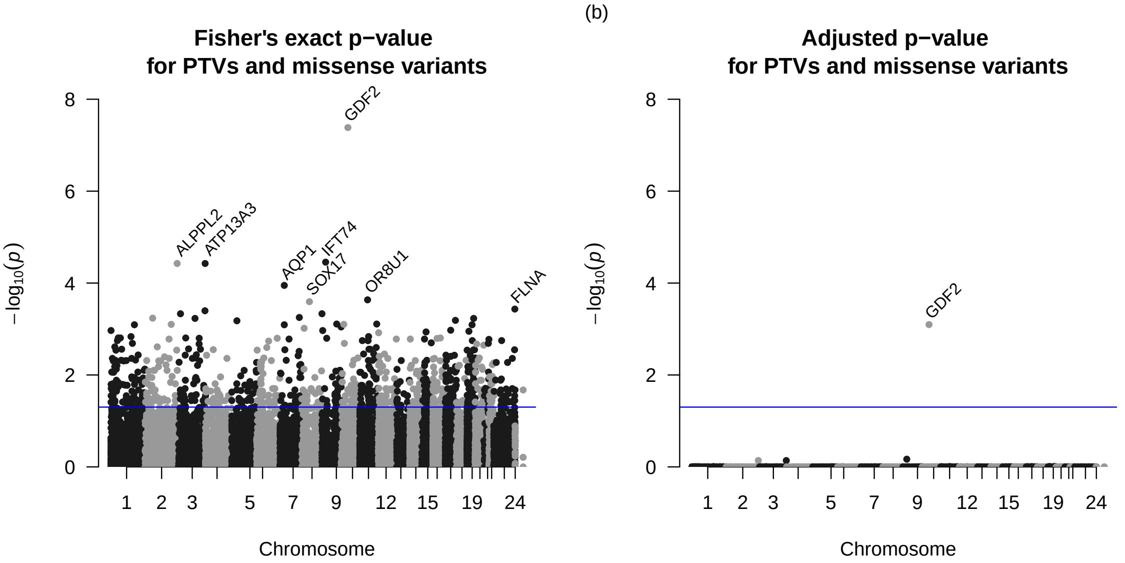


Figure 3.27 Manhattan plots of the combined variant analysis, having excluded cases carrying rare variants in previously established PAH genes. Filtered variants were grouped per gene and tested for an excess of variants in PAH cases. The (a) Fisher's exact p-values and (b) adjusted p-values are plotted against the chromosomal location of each gene. The blue horizontal line indicates a p-value of 0.05. Chromosome X and Y are encoded as 23 and 24 respectively.

After the exclusion of cases carrying deleterious variants in previous reported genes, the results are shown in Tab. 3.13. In addition, the analysis lists putative genes, which were not detected in the previous analyses (*IFT74*, *ALPPL2*, *OR8U1*, *SOX17*, *ATP13A5* and *DOCK8*).

Loss of copy number variation

After exploring the accumulation of small variation (SNV, INDELs), I focused on CNV events, and specifically on the deletion of large regions of the genome, ranging from 50 bp up to megabases (Mb). The aggregated deletion file from the data release contained deletions that were supported by `Manta` and `Canvas`, where one or both supported calls were labelled as ‘PASS’. The aggregated deletions were selected for a frequency of less than 1 in 1,000 in the NIHR BR-RD release (9,110 samples) and were required to demonstrate partial or full overlap with one or more protein coding exons. This genome wide analysis extracted the number of PAHIDX and UPAHC subjects with deletions to identify genes with an over-representation of deletions in cases. *BMPR2* was statistically significantly ($\mathbf{P}_{adj}$ <0.05) overrepresented and is listed in Tab. 3.14 with other putative genes. However, these other putative genes are located in close proximity to *BMPR2* and presented with exactly the

HGNC	Cases	Controls	Frequency cases	Frequency controls	Fisher's exact p-value	$\mathbf{P}_{adj}$
GDF2	12	6	0.01	9.3E-4	4.13E-08	8.0E-4
IFT74	12	17	0.01	3.0E-03	3.5E-05	0.68
ATP13A3	11	14	0.01	2.0E-03	3.75E-05	0.73
ALPPL2	11	14	0.01	2.0E-03	3.75E-05	0.73
AQP1	7	5	0.009	8.0E-04	1.0E-04	1
OR8U1	14	30	0.02	5.0E-03	2.0E-04	1
SOX17	8	9	0.01	1.0E-03	3.0E-04	1
FLNA	15	36	0.02	6.0E-03	4.0E-04	1
ATP13A5	13	28	0.02	4.0E-03	4.0E-04	1
DOCK8	20	60	0.02	9.0E-03	5.0E-04	1
C3orf20	10	17	0.01	3.0E-03	5.0E-04	1
KDR	9	19	0.01	3.0E-03	2.0E-03	1
EVI5	8	17	0.01	3.0E-03	5.0E-03	1
TXNRD3	7	16	0.009	3.0E-03	0.01	1
PRR22	4	8	0.005	1.0E-03	0.04	1
SRMS	3	29	0.004	5.0E-03	0.72	1

Table 3.13 Results of the combined analysis after the removal of cases with variants in the previously reported genes. Only for displaying purposes, genes identified in the PTV or nonsense analyses (orange) or with a fisher p-value <0.0005 are listed. Entries after the horizontal line are in no specific order. The gene *GDF2* reached a $\mathbf{P}_{adj}$ <0.05.

same number of cases. After removing the 220 cases with variants in previously reported genes, no statistically significant gene deletion was identified. Closer examination of the analysis revealed 5 deletions between 0.5 Mb and 3.7 Mb, which covered the wider *BMPR2* loci including surrounding genes.

3.3.6 Assessment of novel PAH gene associations

The disease-gene association tests identified *ATP13A3* and *GDF2* with genome wide significance as part of the PTV and missense analysis respectively. We considered filtered PTV and missense variants as possible contributing factors for both novel disease genes. The assessment of the functional impact was based on the protein location, functional domains annotation and likely changes in the protein structure. We performed the analyses on 1,048 affected PAH cases and 6,385 unrelated PAH controls and used the same PTV and missense filtering strategy as for the novel gene discovery.

ATP13A3

The PTV analysis identified the novel gene *ATP13A3*, which is part of the P5 subfamily of P-type transport ATPases and poorly characterised (Schultheis et al., 2004). Computationally inferred protein domain and sub-cellular location information were available and are visualised in Fig. 3.28. The filtered variant set contained 3 frameshift, 2 stop gained and 1 splice donor PTV (Fig. 3.28(a)), which were heterozygous and predicted to lead to a loss of the protein activity. In addition, 5 heterozygous missense variants (Fig. 3.28(b)) were identified in cases. One variant was shared with controls (Fig. 3.28(d)) at the start of the protein, while the remaining variants fall within or close to the same cytoplasmic region. Samples with identified variants in *ATP13A3* were sequenced with different pipelines and no pipeline bias was found (Fisher's exact test).

ATP13A3 copy number loss

The CNV analysis of deletions (see Loss of copy number variation in 3.3.5) revealed no deletion in *ATP13A3*. I analysed `Canvas` calls separately and identified 2 possible affected PAH subjects with heterozygous deletions of one or more exons in *ATP13A3*. The affected regions, coverage and read alignment are shown for deletions of length 1.6 Kb (Fig. 3.29(a)) and 9.2 Kb (Fig. 3.29(b)). Both deletions failed the Illumina quality metrics. The larger deleted region (Fig. 3.29(b)) was identified only once in NIHR BR-RD while the shorter region (Fig. 3.29(a)) was found deleted in 7 unrelated cases and discarded as technical artifact. In addition, visual inspection rejected the existence of the 1.6 Kb deletion but found sufficient

HGNC	Cases	Controls	Frequency cases	Frequency controls	Fisher’s exact p-value	$\mathbf{P}_{adj}$
BMPR2	22	0	0.02	0	1.32E-19	1.63E-16
NOP58	4	0	0.004	0	3.80E-4	0.47
SUMO1	4	0	0.004	0	3.80E-4	0.47
CARF	4	0	0.004	0	3.80E-4	0.47
FAM117B	4	0	0.004	0	3.80E-4	0.47
WDR12	4	0	0.004	0	3.80E-4	0.47

Table 3.14 Loss of copy number analysis identified an over-representation of deletions in *BMPR2* of genome wide significance. Other putative genes with a fisher p-value <0.0005 were located in close proximity to the *BMPR2* loci. These putative genes were covered by 5 deletions overlapped *BMPR2* and were not further considered.

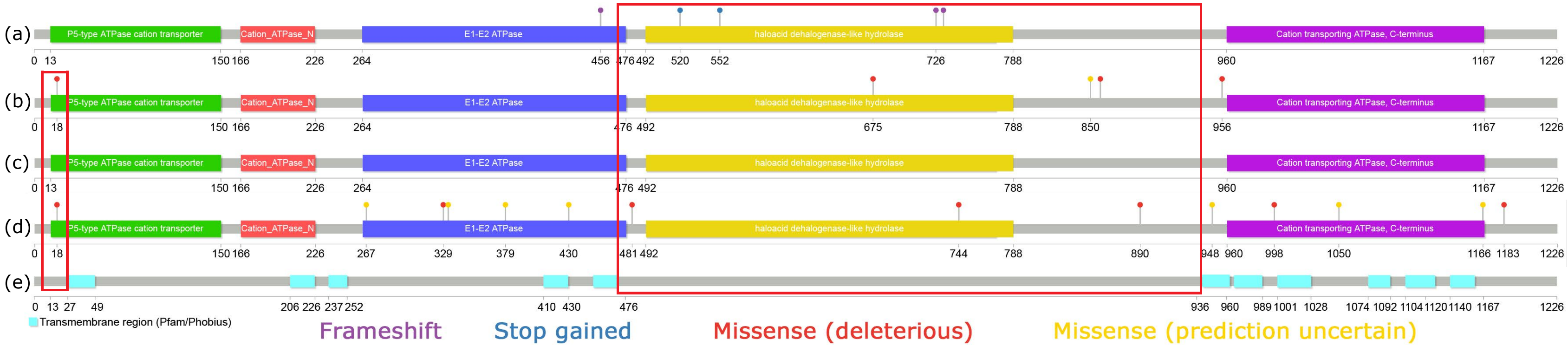


Figure 3.28 Protein truncating and missense variants highlighted on protein domains of *ATP13A3*. Lollipops indicate the variant position for (a,c) PTV and (b,d) missense variants for (a,b) PAH affected patients and (c,d) unrelated non-PAH controls. (e) Transmembrane regions are highlighted in blue. The colour representation of the lollipops are listed at the bottom. Missense (deleterious) are missense variants predicted to be deleterious and damaging in SIFT and PolyPhen-2 respectively with the remaining classified as uncertain prediction. The red box at the start of the protein shows one missense variant, which was observed in cases and controls. The red box in the center highlights a cytoplasmic region adjacent to transmembrane regions, which contain most of the observed PTVs and missense variants in cases. One splice donor variant is not displayed for an affected PAH patient, because the variant is located outside the exon.

support for the 9.2 Kb deletion. The additional analysis discovered 1 additional affected PAH subject with an exonic deletion in *ATP13A3*.

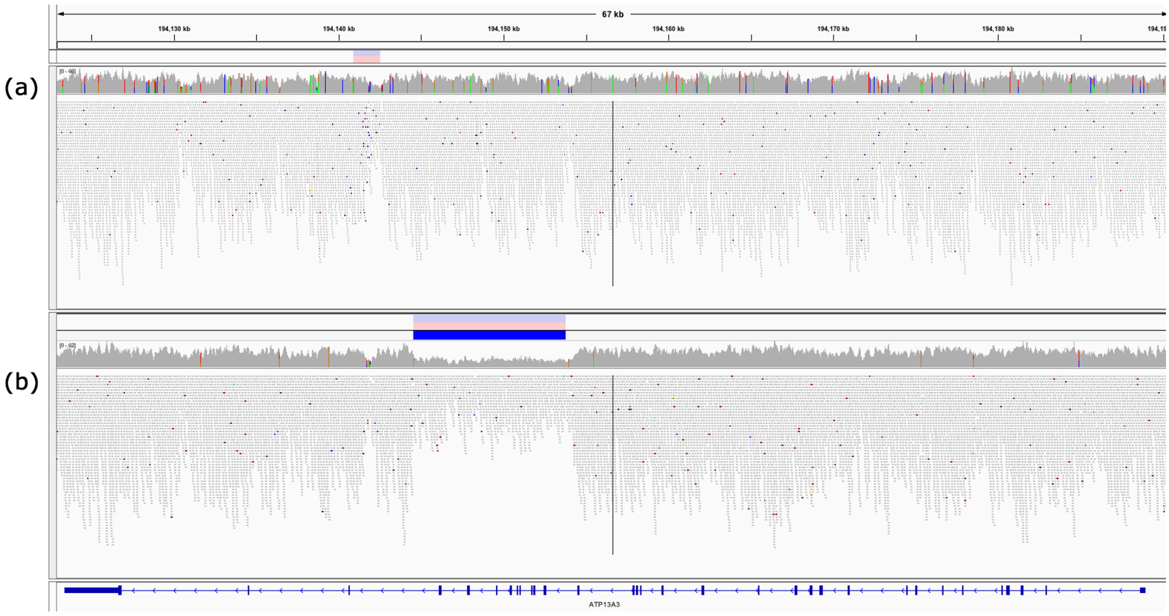


Figure 3.29 Loss of copy number in PAH cases of the *ATP13A3* loci. (a) The deleted region is highlighted as red and blue box followed by the read coverage and read alignment information. The visual inspection did not support the existence of a deletion. (b) The deleted region was supported by a drop of read coverage, but only some partially mapped reads surround the breakpoints.

ATP13A3 diagnostic descriptors

The previous analysis identified 11 subjects with PTV (n=6), missense (n=4) variants and deletions (n=1). I extracted the diagnostic descriptors for PAH from the OpenClinica phenotype release and compared *ATP13A3* with *BMPR2* variant carriers against the remaining PAH affected cohort. *ATP13A3* variant carriers were older compared to *BMPR2* variant carriers. Cardiac output, mPAP and KCO measurements were more extreme compared to *BMPR2*, but not significantly different (Fig. 3.30). The LVEDP distribution appeared elevated, but was based on two values only. No vasoreactivity information were available and no difference was found for PAWP in identified *ATP13A3* cases.

GDF2

The gene *GDF2* was identified as part of the missense analysis with significant genome wide over-representation. The gene is also known as bone morphogenetic protein 9 (*BMP9*),

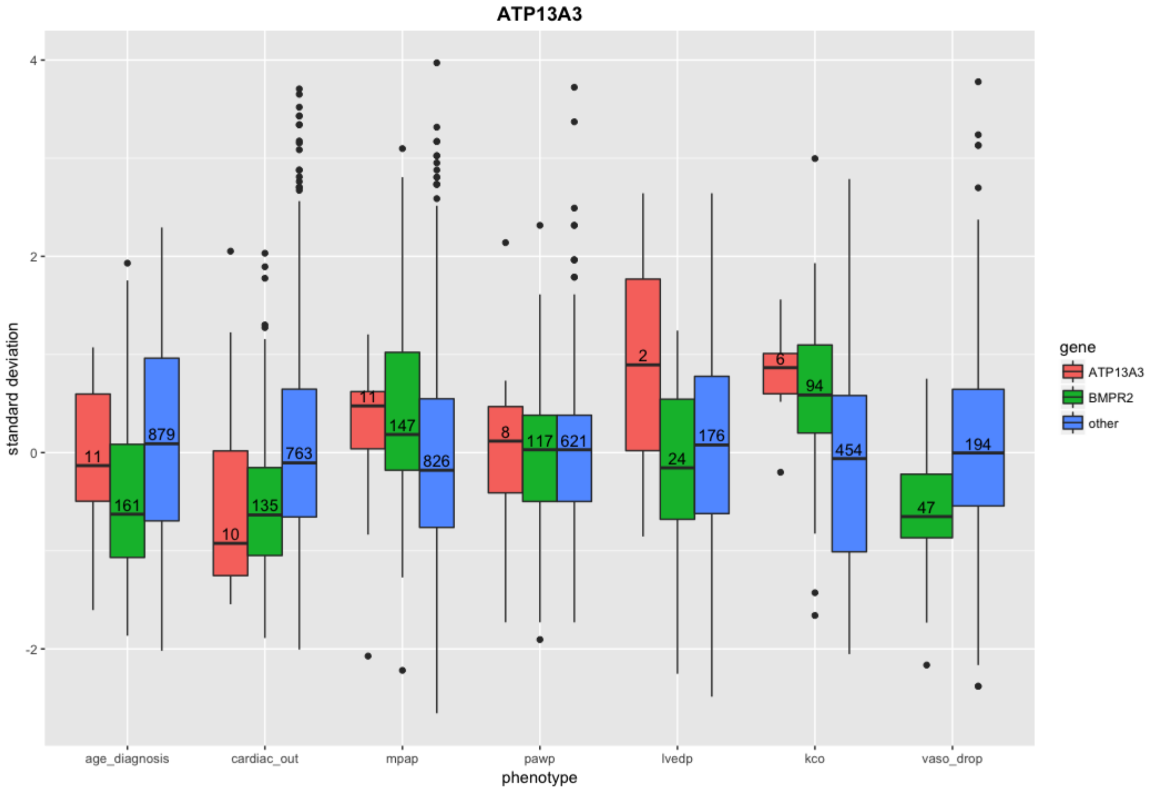


Figure 3.30 Diagnostic descriptors for subjects identified with variants in *ATP13A3* and *BMPR2*. For *ATP13A3* variant carriers, the age at diagnosis was in range to the remaining PAH cohort, but older than *BMPR2* variant carriers. Cardiac output, mPAP and KCO of *ATP13A3* deviated more than *BMPR2* from the remaining PAH cohort, but could be explained by fewer observations. No differences were found in PAWP, no vasoreactivity information were recorded and only two values were available for LVEDP in *ATP13A3* carriers. No significant difference was observed between *ATP13A3* and *BMPR2* variant carriers.

encodes the growth and differentiation factor 2 and was identified as a circulating ligand for the *BMPR2/ACVRL1* receptor complex (David et al., 2007). Heterozygous variants were present in 13 affected PAH subjects and included 1 frameshift, 1 splice site and 10 missense (11 subjects) variants. Two missense variants (3 subjects) found in PAH affected cases were shared with unrelated PAH controls. No deletion was identified in the CNV analysis (see Loss of copy number variation in 3.3.5). Identified variants are shown in Fig. 3.31 and variants shared with control subjects are highlighted. All missense variants in cases were predicted to be deleterious compared to controls and were located on or close to the transforming growth factor beta (TGF-β) propeptide and TGF-β domain. Samples with identified variants in *GDF2* were sequenced with different pipelines and no pipeline bias was found (Fisher's exact test).

GDF2 copy number loss

The CNV analysis (see Loss of copy number variation in 3.3.5) did not detect deletions for the *GDF2* loci. Additional analysis considered all deletions called by `Canvas` and identified 3 affected PAH subjects passing the quality metrics. Two larger deletions were 4.2 Mb and started and ended within 900 bp and 4 Kb of each other respectively. The shorter deletion was 1.1 Mb and started between the two larger deletions. In contrast, `Manta` called inversions for these 3 subjects covering the *GDF2* loci of 3.8 Mb size with start and end positions within 71 bp of each other. In addition, `Manta` also called at least 1 inversion for 91% (n=964) and 89% (n=5740) of cases and controls respectively. The difference was significant (p-value=0.01845), but did not take the read length into account. No relationship was known between these 3 individuals and were sequenced using different read lengths.

To confirm or reject possible deletions, I implemented alternative methods to validate deletions and investigated the possible cause for the disagreement between `Manta` and `Canvas`. The first alternative method focused on the read coverage of the protein coding exons of *GDF2* in comparison to *BMPR2* (see Fig. 3.32(a)). The analysis found a reduction of more than 50% for subjects with `Canvas` deletions overlapping *GDF2*. I found a significantly lower (Student's t-test p-value = 0.005) distribution of the coverage for samples with called *GDF2* deletions compared to other PAH samples. After the confirmation of a drop in coverage, the second alternative analysis focused on the number of heterozygous and homozygous variant calls for a 200 Kb region overlapping the *GDF2* loci. Fig. 3.32(b) shows the number of heterozygous and homozygous variant calls per sample in the PAH affected cohort. Subjects with `Canvas` deletions are highlighted in blue and found to have a significantly lower (Student's t-test p-value = 2.2E-16) distribution of ratio compared to other PAH

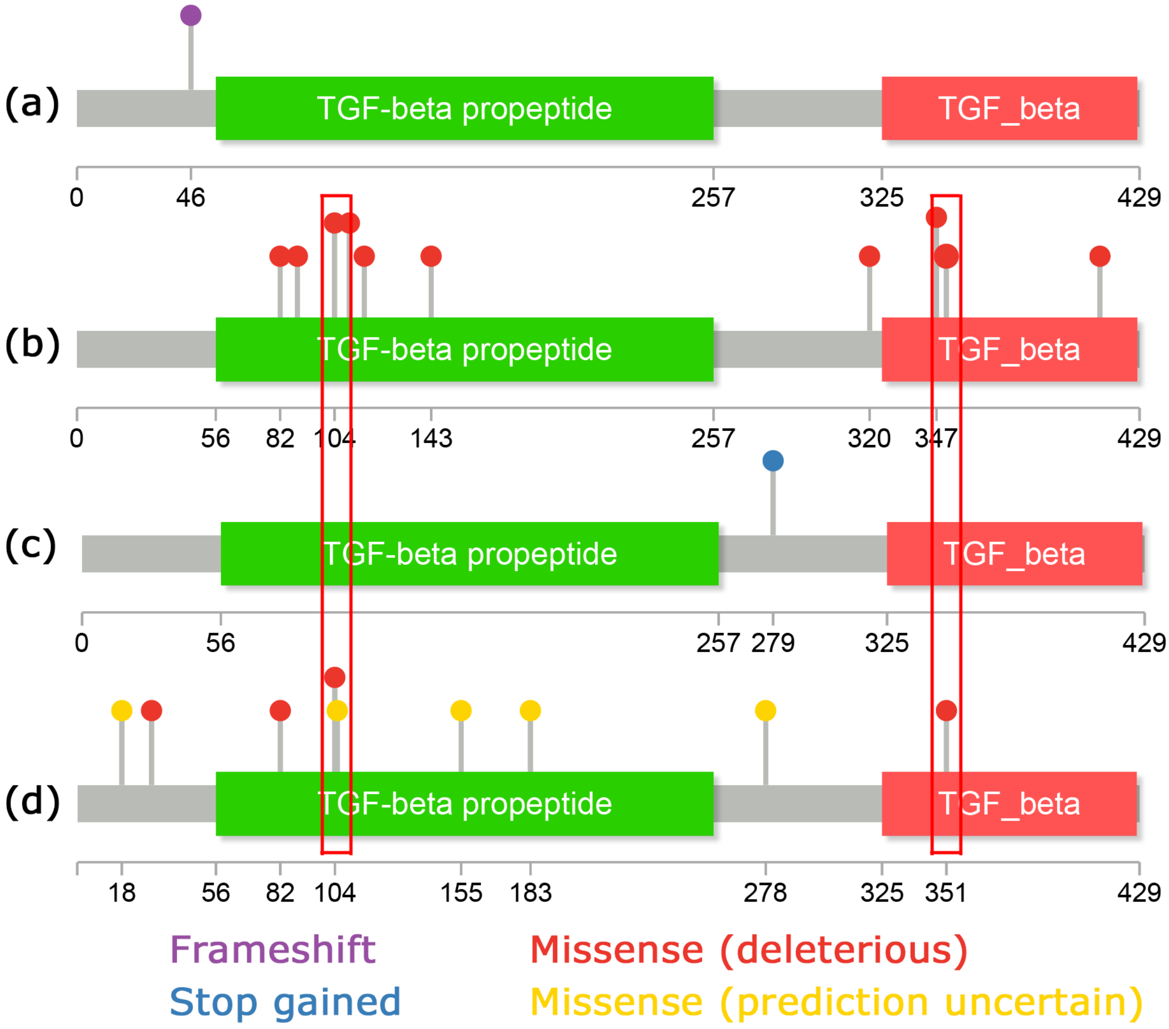


Figure 3.31 Protein position of identified variants in cases and controls. Lollipops indicate the position of (a,c) protein truncating and (b,d) missense variants in (a,b) affected PAH cases and (c,d) non-PAH controls. The colour representation of the lollipops are listed at the bottom. Missense (deleterious) are missense variants predicted to be deleterious and damaging in SIFT and PolyPhen-2 respectively with the remaining classified as uncertain prediction. Missense lollipops in (b) cases are all deleterious compared to (d) controls. Variants in cases cluster in the TGF-β propeptide and TGF-β domain. The red boxes surrounding missense variants at the protein position 104 and 351 were shared between cases and controls and were not regarded as deleterious. One splice acceptor variant is not displayed for PAH affected cases due to their genomic location.

samples. The lack of heterozygous variant calls for these 3 subjects with Canvas deletions in the *GDF2* region supported the absence of one allele.

Canvas deletions overlapped *GDF2* for 3 affected PAH patients and I confirmed the loss of one allele by analysing the relative coverage to *BMPR2* and the heterozygous/homozygous ratio in variant calls for an extended *GDF2* region. The analysis did not address the cause for the high number of Manta inversion calls, but found an increase in cases compared to controls.

Imperfect reference genome assembly affects *GDF2* deletion detection

Additional analysis identified and confirmed the existence of 3 PAH subjects with deletions of the *GDF2* loci, but also found large numbers of inversions called by Manta. The Manta calls are based on paired-end and partial read information compared to coverage information used by Canvas. Inversion calls would require the alignment of paired-end reads or partial alignment at both ends of the inverted region or both. Supporting read information were assessed in relation to the used reference genome representation. The call of an inversion in 90% (n=6704) of subjects indicated an alternative, incomplete or misrepresentation of the GRCh37 reference. To evaluate a possible misrepresentation, I extracted a 5 Mb region from the GRCh37 and GRCh38 reference genomes surrounding *GDF2* for analysis. A sequence similarity comparison (Fig. 3.33) identified an inversion of the *GDF2* region and a translocation of a larger region in GRCh38 compared to GRCh37. The large number of inversions were explained by an insufficient representation of GRCh37, which was corrected in GRCh38. The differences included rearrangement and inversion of large genomic regions.

GDF2 diagnostic descriptors

The previous analysis identified 12 subjects with PTV (n=2), missense variants (n=7) and deletions (n=3) in *GDF2*. The PAH diagnostic descriptors were extracted from the OpenClinica phenotype release and compared *GDF2* variant carriers, *BMPR2* variant carriers with the remaining PAH affected cohort. The average age (mean=44.71) was reduced to PAH affected (mean=51.18), but increased compared to *BMPR2* (mean=42.14). Cardiac output was higher (mean=4.4) and mPAP lower (mean=48.64) in *GDF2* carriers compared to remaining PAH affected cohort. Compared to *BMPR2* carriers, cardiac output and mPAP showed the opposite deviation and were significantly different (p=0.0099 and p=0.0032). PAWP (mean=8.22) and LVEDP (mean=8.66) were also reduced in *GDF2* carriers compared to the remaining affected PAH cohort.

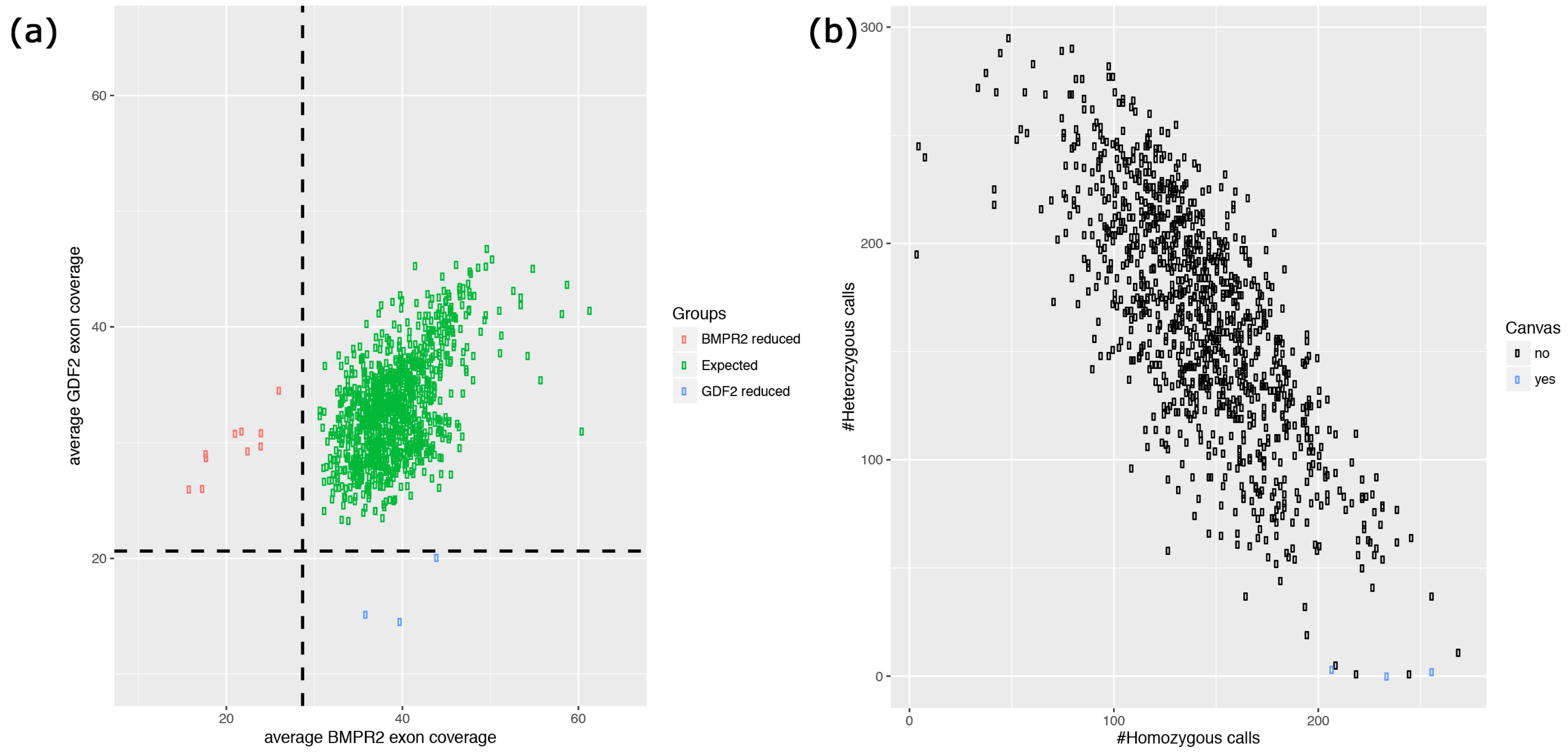


Figure 3.32 Coverage and heterozygous / homozygous ratio analysis confirms the existence of deletions. (a) The average exon read coverage for *BMPR2* was compared to *GDF2* for the PAH affected cohort. Subjects were highlighted in blue for reduced coverage in *GDF2*, red for reduced coverage in *BMPR2* and green for the remaining. The lines indicate the cut-off point (Q11.5xIQR). Subjects with a low *GDF2* coverage had deletions called by `Canvas`. (b) The number of homozygous was compared to heterozygous variant calls for a 200 Kb region covering *GDF2*. Subjects with `Canvas` deletions are highlighted in blue and contained close to 0 heterozygous variant calls compared to other affected PAH cases.

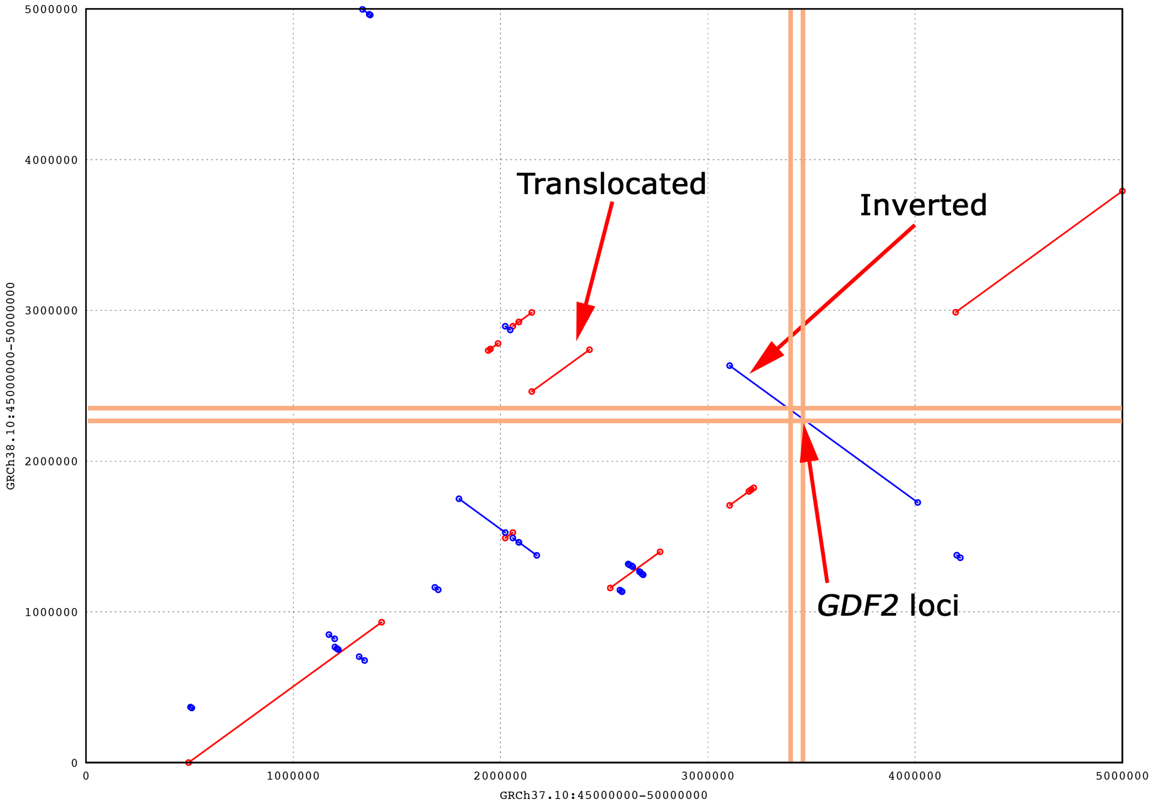


Figure 3.33 Comparison of GRCh37 and GRCh38 for a 5 Mb region surrounding the *GDF2* loci. The red and blue lines indicate matching regions on the same and opposite strand respectively. The start and end position of the *GDF2* gene is indicated for each release with an orange line and the *GDF2* region was found to be inverted in GRCh38 compared to GRCh37. In addition, a larger region moved from the start to the end of the inverted *GDF2* region.

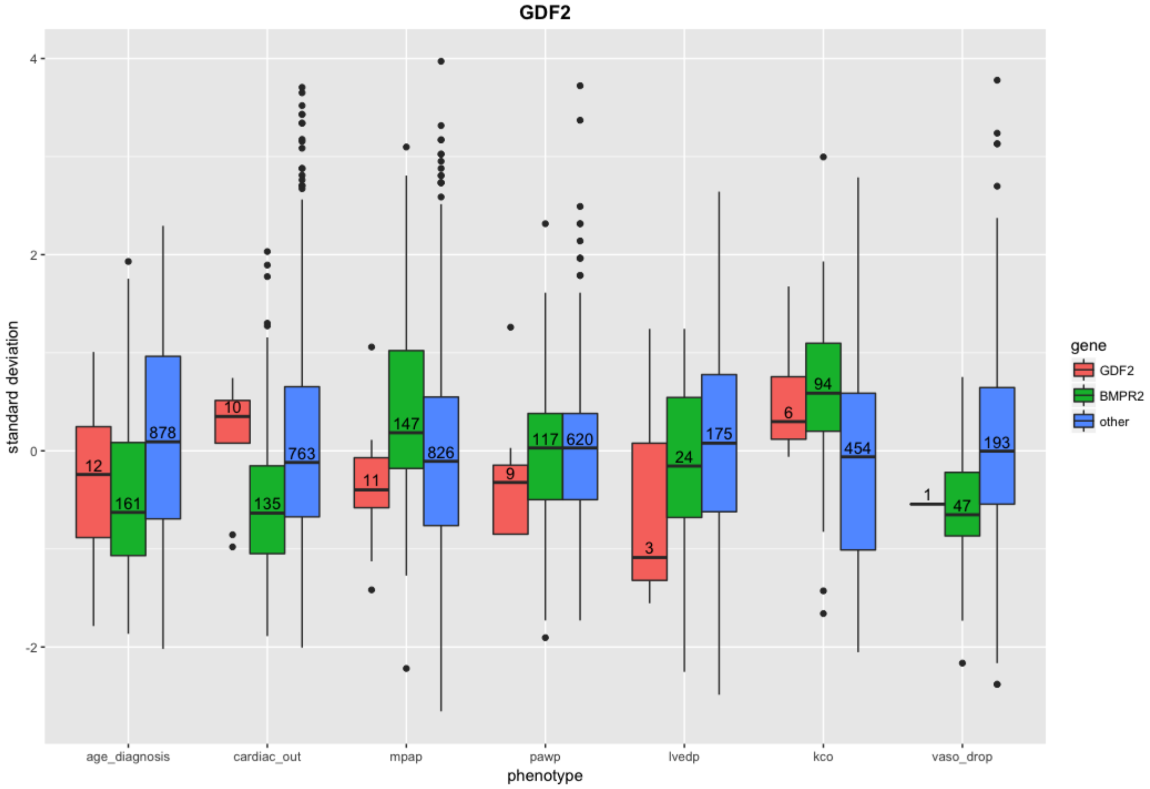


Figure 3.34 Diagnostic descriptors for identified *GDF2* and *BMPR2* variants carriers compared to the remaining PAH affected subjects. Cardiac output and mPAP deviate in opposite directions compared to *BMPR2* and, compared to *BMPR2*, were found to be statistically significant (p=0.0099 and p=0.0032). PAWP and LVEDP were reduced, KCO increased in *GDF2* carriers and vasoreactivity information was only recorded for one subject.

3.4 Discussion

We recruited the to date largest cohort of mainly unrelated cases with idiopathic and heritable forms of PAH to participate in the WGS NIHR BR-RD study. The sequenced samples were checked for quality, gender mismatch, relatedness and ethnicity before aggregation, annotation and filtering. These high quality variants were used to assess rare genetic variation in previously reported PAH disease genes. The assessment extended to the identification of distantly related subjects harbouring the same genetic variation. After characterisation of previously reported genes, the case / control study design aimed to discover novel disease causing genes in affected PAH index cases. The discovery effort included over-representation analyses of PTV, missense variants and large deletions in an extensively filtered variant set, enriched for disease causing variants. We confirmed *BMPR2* as the main disease causing gene in each analysis. The exclusion of cases with variants in previously reported genes revealed significant over-representation of PTVs in *ATP13A3* and missense variants in *GDF2*. Several other putative genes did not reach genome-wide significance, but were found to be highly enriched and in some cases specific to the PAH cohort. These genes were identified by a burden test counting individuals with heterozygous or homozygous variants. Further analysis of the novel disease genes revealed additional subjects with deletions. In particular the analysis of *GDF2* highlighted the need to upgrade to the latest human reference genome and indicated the limitations of the used software tools.

3.4.1 Whole genome sequencing

The NIHR BR-RD study used WGS instead of WES for variant detection in the cohort of rare disease. Previous studies have shown that WGS provides a better exome coverage than WES (Carss et al., 2017; Lelieveld et al., 2015; Turner et al., 2016). The choice of WGS enabled the use of PCR-free sequencing protocols and allowed the capture of genomic regions with extreme GC content without the loss of coverage. In addition, the detection of CNV events is superior in WGS compared to WES, in particular single exon deletions (Carss et al., 2017). WGS allows the precise characterisation of breakpoints falling within intronic and intergenic regions compared to WES.

The read alignment and variant calling of 10K WGS data sets is not a trivial and a computational intensive task using community accepted tools. For the NIHR BR-RD study, Illumina provided aligned read and variant calls based on their proprietary software products, which we used instead of a custom pipeline. Software versions of Illumina's tools changed during the course of the project, which was corrected by re-analysing the affected samples with the latest version. In addition, the underlying biochemistry changed during the project and transitioned from 100bp, 125bp to 150bp read length. Despite the software and biochemistry changes, the alignment and variant quality measurements were in the expected range and suggested that there would be no major gain in reprocessing samples on the same reference genome with a community accepted pipeline compared to the cost implications. Aggregated variant quality measurements indicated a large proportion of technical artefacts, which we addressed by the development of OPR based filtering using VCF files. The introduction of the Hadoop technology enabled further exploration of the variant quality within each read length cohort and the establishment of the minOPR. After detailed evaluation of different minOPR filtering values, we are confident that variants passing the minOPR 0.99 filter represent biological events. However, not all of the failed variants are technical noise, but the quantity of 54% rejected variants highlights the limitations with the current software and filtering strategy.

Population and family structure

The ethnic composition of recruited subjects is diverse and the aggregated variant set could contain variants specific to under-represented populations. However, we found a proportionate equal representation of all populations in affected PAH cases and non-PAH controls. The burden test does not stratify by population, but instead relies on the prior filtering strategy to remove population specific variants and enrich for rare disease causing variants. Novel genes were supported by population independent variants and confirms the sufficient representation of populations in the control cohort. In addition, inflated allele frequencies due to relatedness could lead to over-filtering of otherwise rare variants in non-PAH controls. Such over-filtering could skew the results in favour of PAH cases and increase the chance of false positive results. To improve the specificity, we established the family networks of related individuals up to 3$^{\text{rd}}$ degree. These networks allowed us to select the maximum number of unrelated subjects for an unbiased allele frequency in non-PAH controls.

3.4.2 Variant filter strategy

The variant filter strategy was developed to select rare, likely deleterious variants and used *BMPR2* as a training set. Allele frequency based filtering showed no effect on PTV and highlighted the evolutionary constraint *BMPR2* is under. In comparison, the number of subjects with missense variants changed significantly by introducing allele frequency based filtering, but some remained shared between cases and controls. Effect prediction based filtering was required to reject likely benign variants. The filter strategy was trained on an autosomal dominant gene, which might prioritise genes with a similar profile over recessive genes for the novel gene discovery. Variants affecting recessive genes require homozygous or bi-allelic heterozygous variants and might not be sufficiently enriched. However, the recessive gene *EIF2AK4* associated with pulmonary veno-occlusive disease (PVOD) was significantly overrepresented in the PTV analysis and suggests a balanced filtering approach.

We analysed copy number variation (CNV) after the removal of known benign and frequently observed deletions in external and internal database entries respectively. The genome-wide strategy is conservative and requires a deletion to be called by both variant callers including at least one 'PASS' call. For previously reported genes, the method was extended to include deletions called by the coverage based method `Canvas` only. Duplication, inversion and more complex SV were not considered due to the low specificity of the used tools. The analysis would benefit by the introduction of a community accepted method for a baseline measure, but would need to be run on all samples. The reanalysis would be more beneficial on GRCh38 because of the identified misrepresentation of the reference genome in the *GDF2* loci. This is a computational intensive tasks and will be performed at a later stage.

3.4.3 Variation in previously reported PAH genes

Genes previously associated with PAH were assessed based on the developed filtering strategy to identify possible biases. The genes *SMAD4* and *CAV1* did not bear any rare deleterious variants. For the recessive gene *EIF2AK4*, homozygous, bi-allelic as well as heterozygous variants were identified in PAH patients. These patients were excluded from the identification of novel PAH disease genes even though heterozygous variants were unlikely to cause the disease. The list of subjects with variants in previously reported genes should not be used to report clinical findings. A clinically assessment of these subjects would be required for a correct representation of patients. For the purpose of the burden test, these subjects are highly enriched for likely deleterious variants in previously reported genes and the removal of these subjects increases the chance to discover novel disease genes.

The american college of medical genetics and genomics (ACMG) standards and guidelines provided a framework to assess the pathogenicity of variants in known genes. The framework included the evaluation of multiple sources including computational, functional as well as population data. We found enrichment in 4 previously reported PAH genes (*BMPR2*, *EIF2AK4*, *TBX4*, *ACVRL1*) and no patient with variants in two genes(*CAV1*, *SMAD4*). The absence of variants in patients could be due to the rarity of PAH cases caused by the gene or the removal of causative variants by the filtering strategy. Identified associations for the PTV and missense variant analysis also highlighted differences in the gene function and how PTV, missense variants or the position of a variant affect the gene. However, some known PAH disease genes are not significant and have variants identified in patients as well as non-PAH control subjects, which question the causality of the gene. These genes would benefit from additional studies to evaluate the precise affect of identified variants in these patients to prove or reject the pathogenicity.

3.4.4 Identification of novel PAH disease genes

The burden test implemented for the novel gene discovery separately assesses PTV and missense variants before combining them into one dataset. Depending on the variant set, different genes showed a high over-representation in cases, except *BMPR2*, which was ranked highest for each variant set. The burden test of PTV identifies *TBX4* with genome-wide significance, was low ranked for missense variants and the results in the combined analysis suggests to be driven by PTVs. In contrast, the significance of *GDF2* was mainly supported by missense variants with only 1 rare PTV discovered in a PAH index case. These findings suggest that it is beneficial to test the variant burden on genes for PTVs, missense and combined variants separately in rare diseases. A gene only affected by one group of variants is informative in respect of their biological function (see *GDF2* in 3.3.6 on page 105) and interaction.

Novel disease genes

The novel PAH disease gene *ATP13A3* was identified with genome-wide significance in the PTV analysis. No variants were found to be shared with subjects for variants in previously reported genes. We compared the number of loss-of-function (LoF) variants reported in the exome aggregation consortium (ExAC) and found 8 LoF variants in 60,706 subjects. The listed LoF intolerance (pLI) probability score in ExAC was 1.00 for *ATP13A3* and genes with pLI >= 0.9 are considered to be extremely LoF intolerant genes. The gene encodes a poorly characterised P-type ATPase of the P5 subfamily and was found to be expressed in all mouse tissues, but highest expression in liver (Schultheis et al., 2004). The involvement in polyamine transport was demonstrated, but specifics are unknown (Madan et al., 2016). The gene expression is confirmed in pulmonary artery endothelial cell (PAEC) / pulmonary artery smooth muscle cells (PASMC), lung tissue and human umbilical vein endothelial cell (HUVEC) by the encyclopaedia of DNA elements (ENCODE), Genotype-Tissue Expression (GTEx) and BLUEPRINT project (Adams et al., 2012; Consortium, 2012; GTEx Consortium, 2013).

The missense analysis revealed the contribution of *GDF2* towards PAH in adult-onset cases and the exclusion of subjects with variants in previously reported genes did not reduce the number identified cases. The gene encodes the major circulating ligand for the endothelial BMPR2/ACVRL1 receptor complex and is expressed in the liver

Candidate disease genes

The burden test provided a selection of genes ranked by the unadjusted p-value with no genome-wide significance and included *AQP1*, *SOX17* and *FLNA*. Compared to *GDF2* and *ATP13A3*, the gene *AQP1* contains one heterozygous variant that is shared by 5 unrelated PAH index cases and not found in controls. These 5 subjects were recruited in participating centres across Europe (Amsterdam, Giessen, Papworth, Paris), unrelated (4th degree or more distant), sequenced at different times and are of european ancestry. We found co-occurrence of *AQP1* in two subjects, with a heterozytous variant in *EIF2AK4* for one subject and with a heterozygous *ENG* variant for the other subject. Assessment of these two variants categorised them as likely benign even though they passed the variant filtering. Aquaporin 1 is a membrane protein to facilitate water transport in response to osmotic gradients and is highly expressed in lung (GTEx) tissue, blood outgrowth endothelial cells (BOEC) and HUVEC (Adams et al., 2012). Nitric oxide (NO) was shown to be transported by *AQP1* and may play a role in controlling blood pressure (Herrera et al., 2006), while an other study implicates *AQP1* in the CO2 transport in blood (Hsu et al., 2017).

SOX17 harbours missense and nonsense variants and encodes SRY-box containing transcription factor 17. It has a key role in the vascular development and angiogenesis (Matsui et al., 2006), and conditional deletion leads to impaired formation of lung micro-vessels (Lange et al., 2014). Similar to *AQP1*, *SOX17* is highly expressed in relevant cells, which includes PAEC, HUVEC and BOEC (Adams et al., 2012). The implication of the vascular endothelium provides further evidence that this cell type has a major contribution in initiating the disease.

The X-linked gene *FLNA* was highly ranked after using the gender adjusted allele frequency cut-off (see Rare variant selection in 3.2.5 on page 56). Heterozygous and hemizygous variants were found for female and male PAH index cases respectively. *FLNA* encodes the widely expressed Filamin A protein and is central in providing a scaffold to anchor cytoplasmic signalling proteins (Stossel et al., 2001; van der Flier and Sonnenberg, 2001) and is implicated in different syndromic diseases (Gómez-Garre et al., 2006; Mariño-Enríquez et al., 2007). Compared to variants published for other diseases, variants of PAH patients cluster in a specific region, which was found to overlap with a SMAD binding domain. Heterozygous mutations in *FLNA* were also found in a case report and were suggested as the likely cause of the familial case of PAH in two females (Hirashiki et al., 2017). The identification of an X-linked gene could partially explain the over-representation of females in PAH.

Rare variant association tests

In addition to the burden test, we applied the established rare variant kernel-based SKAT-O method on the combined variant set. Excluding subjects with previously reported genes, the analysis identified *AQP1* ($\mathbf{P}_{adj}$ = 4.28x10-6), *MFRP* ($\mathbf{P}_{adj}$ = 1.30x10-5) and *SOX17* ($\mathbf{P}_{adj}$ = 6.69x10-5) as the top associated genes. In comparison, the unadjusted p-values from the burden test were 0.00011 (*AQP1*), 0.00252 (*MFRP*) and 0.00025 (*SOX17*). The identification of the same variant in 5 unrelated PAH index patients could explain the difference in ranking of *AQP1* between the burden test and SKAT-O. The gene *MFRP* contains one variant, again shared by 4 unrelated PAH index patients, but is questionable based on the association with retinal degeneration and a lack of gene expression in endothelial cells. In comparison, the genome-wide significance of *SOX17* using SKAT-O was based on singleton variants (not shared by unrelated individuals) and highlighted differences in the methodology of the burden test.

Rare disease gene discovery is a challenge with current tools and methods to reach genome-wide significance due to the small cohorts size. The discovery of *AQP1* and *SOX17* was based on an established statistical method and confirms the existence of novel disease genes, which do not reach genome-wide significance using a burden test. In contrast, the burden test identified *GDF2* and *ATP13A3* in different variant types and provides a granular ranking for potential candidate genes. The gender adjusted filter of the X chromosome was essential to identify possible X-linked genes. Further more, highly ranked candidate and novel PAH genes show co-occurrence with variants in subjects with previously reported genes. This observation could suggest di- or oligogenic inheritance and could explain incomplete penetrance in PAH.

Replication and validation

The discovery of novel disease-gene associations was based on the PAH patient cohort recruited as part of the NIHR BR-RD study. Recruiting centres included all specialised NHS PH centres in the UK and collaborating centres across Europe. The national audit of PH from 2017[2] reported 1,514 identified patient between 2009 and 2017 with idiopathic, heritable, or anorexigen-induced PAH. Considering that 50% of patients do not survive 4 years and excluding anorexigen-induced PAH, the NIHR BR-RD PAH project likely recruited all PAH

[2] National Audit of PH: https://files.digital.nhs.uk/pdf/h/8/national_audit_of_pulmonary_hypertension_8th_annual_report.pdf

patients that would like to participate in the UK. Replication of the disease-gene association test on a similar sized cohort is not possible in Europe in addition to financial implications. However, similar sized PH cohorts with basic phenotype information are available in the united states (US), which was partly exome sequenced. The sequence data were not available for analysis, but we are in discussion for a possible collaboration to replicate the analysis. The remaining samples could be screened for variants in selected genes to reduce costs and time, but the careful sample selection for a matching cohort will be essential.

Second, the novel disease-gene associations will be functional characterised using animal models. We applied to the genome editing mice for medicine (GEMM) program from the medical research council (MRC) for two genetically altered mice. The genes *ATP13A3* and *SOX17* should be altered in one mouse each. Variants from PAH patients were translated from the human genome coordinates to the mouse genome, computationally assessed for their impact and reviewed by experts for their suitability. Applications were submitted and approved for both genes and we are waiting for the edited mice to arrive.

Third, we study the function of cells using patient samples, samples from healthy donors and variants introduced into samples from healthy donors. The focus is on identified *GDF2*, *ATP13A3* or *SOX17* variants from patients in relevant cell types. Human pulmonary artery smooth muscle cells (PASMCs) were retrieved from explants dissected from lung resection specimens, human pulmonary artery endothelial cells (PAECs) were purchased from Lonza and blood outgrowth endothelial cells (BOECs) were derived from peripheral venous blood isolated from healthy subjects. The study was approved by the Cambridgeshire 3 Research Ethics Committee (Ref 11/EE/0297), and all subjects provided informed and written consent. Some results were included in a recent publication and further functional work is ongoing (Gräf et al., 2018).

3.4.5 Distant relatedness

We identified nearly identical deletions in 4 subjects of the last *BMPR2* exon. The relatedness of two subjects was established based on the family networks. Additional analysis identified a third subjects as a distantly related case, which was confirmed as a 5th degree relationship by a family tree. The status of the 4th sample requires re-analysis of the latest dataset, which was not performed at the time. Both distant related subjects reported a possible case of PAH in the family, which was identified as the same diseased person through an extended family tree. This possible person was 2nd and 3rd degree related to the different index cases. Based on the family structure and inheritance pattern, the parents must have been carriers of the *BMPR2* deletion and we are not aware of any PAH symptoms. A closer examination of the shared segments between the distantly related subjects revealed that the longest and second longest shared segment overlapped with *BMPR2* and *SOX17* respectively. No rare variants were found in *SOX17* for both subjects. Further work needs to be done to elucidate specific haplotypes of these shared regions and would require the phasing of the variants for all three related subjects for a co-segregation analysis. This would allow to analyse the wider locus of *SOX17* in more detail.

3.4.6 International collaboration

The NIHR BR-RD PAH project initially recruited PAH patients from NHS specialist centres in the United Kingdom. Prevalence and survival of IPAH patients limit the pool of IPAH patients available in the UK and was further reduced by patient participation and patients visiting satellite instead of recruiting centres. We established international collaborations to increase the cohort size to 1250 WGS samples. The growth in number of samples increases the power to discover novel disease causing genes based on rare variants, but these results can be biased by the ethnic origin of the patient. Stratified analyses will become more important to account for the ethnic diversity.

3.4.7 Beyond protein-coding regions

The novel disease gene discovery focused on protein coding gene regions, which covers only a small proportion of the human genome. The remaining 98% of the genome is ignored by the current strategy and would require cell type focused analyses to depict the hundred millions of variants. Comparing cell type specific functional data from patients would allow to detect differences in open chromatin regions or histone modification. Matching genomic and gene expression data would facilitate targeted analysis to identify variants effecting the transcription of a gene. Transcript specific differences would help elucidate the effect of suspected splice variants. For this purpose, novel machine learning approaches needs to be explored to further combine functional, expression and phenotype data with genomic information.

3.4.8 Phenotype integration

The rare variant analysis focused on the protein coding regions of the genome, which represents 2% of the human genome. *ATP13A3*, *GDF2*, *AQP1* and *SOX17* were discovered or confirmed as novel PAH genes by a burden or SKAT-O test respectively. Subjects with variants in these genes and previously reported genes account for 260 (25%) subjects of 1048 unrelated PAH index patients. The proportion of *de-novo* and inherited variants is not known and would require parental samples. The remaining 75% of patients are not explained yet and available phenotype information is too incomplete to support novel gene discovery. In contrast, the phenotype information is currently informed by the genotype to identify patterns for improved diagnosis. Going forward, the introduction of smart devices and automated sample processing would improve data completeness and allow continuous monitoring of patients. Patient specific trend lines would be more accurate to classify patients compared to a one-off measurement at the time of diagnosis and more beneficial to the genotype analysis.

The novel PAH genes were expressed or present in the relevant cell types. To elucidate the cause of PAH for the remaining cases, future analysis should be guided by cell type specific information. This would allow to identify active genomic regions in the non-coding space and changes in gene expression variants. The further integration of the omics and phenotype data with the genome is essential and ongoing.

Chapter 4

BigData infrastructure for human genome variation

4.1 Introduction

The general introduction (see chapter 1.1.3 on page 4) provided an overview of the challenges of applying whole genome sequencing (WGS) for large populations and discussed the computational burden of aggregating the variants for analysis. The rare disease pulmonary arterial hypertension (PAH) was described in chapter 2 on page 15 and known genetic causes with novel disease-gene associations were described in chapter 3.1 on page 43 using variants from WGS data. In this chapter, I discuss current data exchange standards, available technologies and methods for aggregating, annotating and analysing whole genome variation data for large populations.

4.1.1 Data sharing for global genomic research

Next generation sequencing revolutionised the field of genomics as well as clinical diagnostic and enabled the genome-wide analysis on scale (Koboldt et al., 2013). An increase in sequencing capacity and a drop in price allowed to produce more sequence data in shorter period of time (Pabinger et al., 2013). The cost of analysing the human genome with NGS technology decreased rapidly and made the application for clinical use feasible (Caulfield et al., 2013). At the same time, the availability of cloud based services provided the required secure storage and compute infrastructure to be able to handle and analyse genomic information on scale. Although secure environments provide several benefits such as lower costs and scalability, legal and ethical points needed to be considered (Dove et al., 2015). Besides the mentioned data security, these legal and ethical points also include data control and account-ability in geographically dispersed data centres. Other users sharing the same compute and storing infrastructure should not be able to access data, while sensitive clinical and genomic information should be shared with collaborating researchers. An international code of conduct was proposed to foster the secure and responsible sharing of clinical data (Knoppers et al., 2014, 2011; Mello et al., 2013). For this purpose, the global alliance for genomics and health (GA4GH) was established in 2013, bringing together hundreds of individuals and organizations. While data and compute are distributed around the world, application programming interfaces (APIs) can provide authorized access to virtually connected datasets for seamless analyses. The establishment of a federated model allowed globally collaborating researchers to apply a systematic approach to rare disease gene discovery (Philippakis et al., 2015). In addition, the NIH issued the genomic data sharing (GDS) policy in 2014 to share genomic data with the community in addition to other funding bodies (Health, 2014; Kaye et al., 2009).

The framework document developed by GA4GH regulatory and ethics working group (REWG) provides basic principles for responsible data sharing with the focus on the right of an individual to benefit from scientific and medical advances (Knoppers, 2014). Connecting healthcare providers and research centres allows to improve disease predictions as well as better informed medical decisions by clinicians. For this purpose, the data working group (DWG) developed APIs to globally communicate and exchange information. One developed technical specifications is the beacon project (https://beacon-project.io/) that defines checking the presence or absence of a specific allele. A simple boolean answer is provided protecting the privacy of single individuals. Aggregated access is available through the federated search engine (http://beacon-network.org/) and allows to interrogate a whole network of accessible beacons. A different specification focuses on the retrieval of a comprehensive set of information. The large scale genomics (LCG) work stream defined standardized methods to access distributed genomic data through an API and additional protocols. After successful authorization, encrypted data are shared and include read (BAM/CRAM/SAM), variation (VCF/BCF) and annotation information, in standard file formats where possible.

Matchmaker exchange (MME) is a collaborative effort supported by members from the rare diseases community (Philippakis et al., 2015). The project adopted a federated model to study patients with a clear missing etiology after the initial analysis. Connecting databases through the use of common APIs in a federated network allows the systematic discovery of rare disease genes. Matchmaker Services implementing the MME API store genome variation information and phenotypic abnormalities in standardised human phenotype ontology (HPO) terms (Robinson et al., 2008). Patients similarity are determined by matching identical or ontologically similar terms as well as genotypes. Notifications are automatically sent to clinicians or researchers about matching cases for further analysis.

The open-source for computational biology (OpenCB) initiative (https://github.com/opencb) is a collection of software for high-throughput genomic data and provides implementations of the developed GA4GH standards and data models (https://github.com/opencb/ga4gh). OpenCB comprises data models, APIs and platforms to provide biological data and analyse as well as retrieve genomic information. Biological information are collated by CellBase (https://github.com/opencb/cellbase), which is available through an API and includes a service to annotate variants (Bleda et al., 2012). Annotated variants are returned in standard data models in JavaScript object notation (JSON) format. The open computational genomics analysis (OpenCGA) project (https://github.com/opencb/opencga) provides an authenticated service for genomics data. Patients, samples, variation data and clinical information can be managed and analysed through the API. OpenCGA supplied the platform for the human genome variation archive (HGVA) and serves variant information from large scale projects with variant annotations from CellBase (Lopez et al., 2017). Client-side tools facilitate the API interaction and the interactive variant analysis (IVA) web-interface allows the visual exploration of the data (https://github.com/opencb/iva). HGVA is registered with the beacon network and provides access to aggregated variant information.

Cloud computing platforms have emerged for genomic researchers to store and process large scale data, which takes advantage of scalable services (Afgan et al., 2011; Heath et al., 2014; Reid et al., 2014). Access is provided using cloud specific APIs or the web browser. As a member of the GA4GH consortium, Google developed the genomics API (https://cloud.google.com/genomics) based on the GA4GH framework as an extension to the Google Cloud service. The genomics API supports authentication and allows the storing, processing, exploring, and sharing of genomic data. Client-side tools are available in different languages like Java, Python, Ruby to facilitate the interaction with the API. Imported files can be accessed by region and retrieved in JSON format.

4.1.2 Storage infrastructure

Traditional methods stored files on a central file system with hundreds of gigabytes in size ready for analysis and accessed these files by processing tools on different compute nodes. This centralised storage model required the transfer of terabytes (TB) of data across the network, which is a bottleneck. Distributed storage is the fundamental infrastructure to provide fast access to the data and was adopted by Google for a scalable system (Ghemawat et al., 2003). The MapReduce programming model was proposed to analyse large data sets (Dean and Ghemawat, 2004) and adapted by the open-source projects called Hadoop (White, 2009).

The decentralised storage followed the storage and fast access of structured data by BigTable and HBase (Chang et al., 2006; George, 2011). These developments provided the scale, speed and flexibility required by Google for a storage and analysis platform, while being resilient against hardware failure and data loss.

The advantages of a distributed analysis platform are provided as a service by multiple cloud providers, including Amazon Web Services (AWS), Google Cloud, Nimbus, Eucalyptus, Microsoft Azure, GoGrid and Rackspace (Barr, 2006; Google, 2008; Martin, 2014). The main advantage of these cloud services is the dynamic scaling and readjusting depending on the demand. Security and the physical location of the data were major concerns to store and process genomic information in the cloud (Dove et al., 2015). Focusing on the AWS and Google Cloud, these concerns were partly addressed by the introduction of dedicated storage locations and default server side encryption (Amazon, 2013; Google, 2015). Specialised cloud computing companies have emerged to store and process large scale genomic data based on the infrastructure provided by Amazon or other cloud provides (Reid et al., 2014).

4.1.3 Distributed variant analysis

The genome analysis toolkit (GATK) was designed based on the MapReduce programming model to handle and analyse high throughput sequence (HTS) related data and utilize the distributed architecture (McKenna et al., 2010). The computational approach was adapted to store HTS data in Hadoop and to create analysis workflows (Massie et al., 2013; Schumacher et al., 2014). Applications range from quality control of HTS reads (Robinson et al., 2011b), short read alignment (Pandey and Schlötterer, 2013; Pireddu et al., 2011) and variant calling (Langmead et al., 2009a) using the Hadoop framework. Other implementations focus entirely on the variant analysis and loading variant information from VCF files. O'Connor et al. (2010) demonstrated that query times using the distributed platform outperform traditional database systems with increasing number of variants. The variant analysis frameworks SEQSpark and Hail included rare variant association tests for cohort analyses, but requires multi-sample variant files to be loaded (Hail, 2018; Zhang et al., 2017). In contrast to other methods, the software package ADAM combines read alignment and variant calling (Massie et al., 2013). Specialised file formats were developed for efficient distributed processing of data with the availability of deep-learning libraries. Recent additions to ADAM were copy number as well as single sample and population variant callers (Avocado, 2018; DECA, 2018). Most of the analyses have moved to a distributed environment utilising the parallel computation, but an incremental approach does not exist to merge a single sample or a group of samples with a larger cohort and specifically annotate novel variants to reduce the computational burden.

4.1.4 Scalable variant annotation

The field of genomics applies next generation sequencing (NGS) technology to analyse the genetic diversity in large populations and highlights the importance of rare variation in disease (Lek et al., 2016; Song et al., 2016). In addition to the allele frequency, the identification of disease causing variants requires the consequence type as well as biological relevant information (see Prioritisation of genome variation on page 9). CellBase (Bleda et al., 2012) is a comprehensive repository of gene model, regulatory, functional, genomic and system biology information. The data are stored in a not only SQL (NoSQL) database, which can distribute data and computation across multiple servers for fast response times. A well defined RESTful web interface enables consistent access to required information and provides a variant annotation service that allows the submission of large number of requests. CellBase returns information in JavaScript object notation (JSON) format and includes the predicted consequence as well as a comprehensive set of biological relevant information. Annotation in VCF format is not currently supported by CellBase. The distributed analysis framework SEQSpark (Zhang et al., 2017) combines annotation and analysis that allows to mitigate scalability issues by sharing the computation across an entire compute cluster. In contrast to CellBase, the supported annotation is limited to the prediction of the consequence based on the gene model. The distributed analysis frameworks Hail and ADAM require the annotation of variants outside of their framework and load the prepared annotation files into their infrastructure to be included in the analysis (Hail, 2018; Massie et al., 2013).

4.2 Methods

The NIHR BR-RD recruited subjects and submitted samples for sequencing over a four year period. Single sample sequence and variant data were continuously delivered by Illumina in batches of various sizes. The same analyses were required to run after each batch to identify rare, disease causing variants as well as discover novel disease-gene associations. I assessed the characteristics of variant information and the suitability of distributed platforms to store and analyse variant. Based on these findings, the variant infrastructure for loading, merging, annotating and analysing human genomes (VILMAA) was developed to provide population wide analyses on scale. The results for disease-gene association for protein-truncating variants were compared with traditional methods (see Identification of novel disease-gene associations on page 93) using the latest available NIHR BR-RD samples (see Software and data release information on page 128).

4.2.1 Ethical approval

Samples included in the analysis were recruited for the national institute for health research (NIHR) BioResource - Rare Diseases(BR-RD) study and part of the same ethical approval as discussed in chapter 2.2.2 on page 18 and chapter 3.2.2 on page 48.

4.2.2 Software and data release information

The analyses were based on the reference data sets and software versions listed in Tab. 4.1 and 4.2 respectively unless stated otherwise.

Name	Version	Description
CADD	v1.3	CADD score whole genome SNV and INDELs
Ensembl 37way GERP	75	Conserved regions in humans based on eutherian mammals
ExAC	r0.3	Whole Exome frequencies
GERP	hg19	Downloaded BigWig file from UCSC FTP
Human GRCh37 reference	75	Human autosomes, X and Y downloaded from Ensembl FTP
NIHR BR-RD	20170104-A	Variant release of the NIHR BioResource – Rare Diseases
PhyloP	hg19	Downloaded 100way PhyloP BigWig file from UCSC FTP
PhastCons	hg19	Downloaded 100way PhastCons BigWig file from UCSC FTP
UK10K	20130411	Exome and whole genome frequencies

Table 4.1 Reference data release versions used throughout the project.

Name	Version	Info
Apache Hadoop	2.7.3	part of HORTONWORKS
Apache HBase	1.1.2	part of HORTONWORKS
Apache Phoenix	4.7.0	part of HORTONWORKS
BCFtools	1.3.1	including git commit bdb01d8
HORTONWORKS	HDP 2.5.0	Distributed data platform release
JAVA	8u66	Java Development Kit (JDK)
Python	3.4.1	
R	3.2.4	

Table 4.2 Software release versions used throughout the project

4.2.3 Data structure and data characteristic analysis

The evaluation was based on the merged VCF file from the NIHR BR-RD. The coordinates of exonic regions of protein coding genes were extracted from the Ensembl gene annotation file (see Tab. 4.1). The merged VCF file was filtered using these exonic regions only including ‘PASS’ variants. The defined exonic regions from protein coding regions were shuffled to get a random selection of genomic regions that is representative of the whole human genome. Variants from the exonic regions were extracted from the before prepared VCF file. The first 100K entries based on the unsorted regions were taken into account. The extracted fields were compressed using gzip and the resulting file size was measured. The fields included chromosome, position, reference and alternate for the coordinates, and the info field for the annotation. Genotypes were counted by recording each appearance for all samples in selected variants and heterozygous genotype counts for ‘0/1’ and ‘1/0’ were combined. The minor allele frequency (MAF) was calculated for the unrelated PAH control (UPAHC) and unrelated WGS10K (UWGS10K) cohort (see section 3.2.4 on page 55) using `BCFtools` and selected for the rare and common analysis respectively. Allele frequency based filtering was performed with `BCFtools` and annotation based filtering was performed with the `ensemblVEP R` package.

4.2.4 Hadoop cluster configuration

The Hadoop infrastructure was part of the high performance computing service (HPCS) of the University of Cambridge. Researchers with approved access to the NIHR BR-RD genomic data (see chapter 3.2.3 on page 3.2.3) were granted access to the dedicated Hadoop cluster. The Hadoop cluster consisted of 2 head and 32 worker nodes and the primary software install was ‘Scientific Linux’ version 7.2. Head nodes were equipped with 24 cores, 32 GB of memory and 2 disks of 2TB each per node. Worker nodes provided 48 cores, 64 GB of memory and 16 disks of 2TB each that amounts to 32 TB per node. The Hortonworks distributed data platform release was installed on the cluster and primary installed packages included the hadoop file system (HDFS), YARN, MapReduce2 and HBase. Total encrypted HDFS storage provided was 342 TB after a replication factor of 3. The HBase configuration was adjusted to optimise the JAVA garbage collection by adding `-XX:+UseG1GC -XX:MaxGCPauseMillis=100 -XX:+ParallelRefProcEnabled` to the `SERVER_GC_OPTS` environment. The YARN configuration was modified to allocate 32 GB of memory and 19 cores to be occupied by containers. Approved researchers were granted permission to log into the Hadoop infrastructure using the authentication service Kerberos and provided access to a dedicated storage directory on the Hadoop file system (HDFS).

4.2.5 Normalisation, transformation and loading of gVCFs

The variants were normalised using a two step process to normalise INDELs using `BCFtools` followed by the normalisation and transformation into the data model using `OpenCGA`. The `BCFtools norm` tool was applied to left align and normalise INDELs in genome VCF (gVCF) files using the `-cw` and `-cs` options to warn and set/fix incorrect/missing reference alleles respectively. The reference sequence was provided to allow the relocation of an variant to left most position of a repetitive region. Resulting variants were transformed from gVCF format into *proto* format using `OpenCGA`. During the transformation, `OpenCGA` performed the normalisation of INDELs by removing the anchoring reference bases. The generated *proto* files were loaded into a specified `HBase` table by performing 'PUT' operations using the `HBase` application programming interface (API). The Apache Phoenix library was used for the efficient encoding of the row key, which provided character and number conversion functionality.

4.2.6 Incremental variant merge in `HBase`

Single sample variant information were loaded into a `HBase` table grouped by regions of the genome. A region in `HBase` contained a group of variant information per sample stored in separate columns. The `MapReduce` job to merge variants processed the `HBase` table row per row and retrieved the columns containing the required sample data. The columns were selected using the 'Scan' object to only extract the required columns. Variants from each sample were checked for conflicts and resolved, if required. The resulting variants from the selected samples were decomposed and the information encoded as Apache Phoenix arrays. Row, column and array information were added to an 'Append' object that was submitted to `HBase` using an asynchronous thread.

4.2.7 Variant annotation

Prior to the analysis, variants in HBase required to be annotated with consequence predictions and biological relevant information, which reached hundreds of millions of variants. Due to the speed and RESTful web service, CellBase (Bleda et al., 2012) was used to annotate the variants hosted by the high-performance computing service (HPCS) and included allele frequencies, deleteriousness and conservation scores. The processes were executed as `MapReduce` jobs and configured to run 10 parallel executors with 2,560 MB memory, scanner timeout of 1,200,000, caching of 10,000 entries, provide `java` options to use a maximum memory of 2048 MB and to use the G1GC garbage collector Rows with variant annotations were ignored, unless re-annotation was forced. The row key was decoded as a variant object without genotype information. Batches of 200 variants were submitted for annotation to Cellbase using the REST API. The JavaScript object notation (JSON) formatted response is defined as part of the CellBase API and was stored as JSON text in the annotation column. The population allele frequencies were provided for the 1000G project (Sudmant et al., 2015), UK10K project (UK10K Consortium et al., 2015) and the exome aggregation consortium (ExAC) / genome aggregation database (gnomAD) (Song et al., 2016). Deleteriousness predictions were provided for the combined annotation dependent depletion (CADD) score (Kircher et al., 2014), SIFT (Ng and Henikoff, 2003) and PolyPhen-2 (Adzhubei et al., 2013). The conservation annotations included the genomic evolutionary rate profiling (GERP) (Cooper et al., 2005), PhastCons (Siepel et al., 2005) and PhyloP (Pollard et al., 2010). The variant type (SNV, MNV, insertion, deletion, complex) was inferred and stored in a separate column by the developed variant infrastructure for loading, merging, annotating and analysing human genomes (VILMAA).

4.2.8 Calculation of cohort summary statistic

Calculated summary statistics from different NIHR BR-RD cohorts were used for variant filtering, selection and quality control. Frequently used values were pre-calculated and stored for a list of NIHR BR-RD cohorts in `HBase` (see Tab. 4.3). The different cohorts were defined based on read length, gender, maximum unrelated set, ethnicity and project information provided by the NIHR BR-RD release. A file was created for each cohort with the selected sample identifiers and loaded into `OpenCGA` using the command line API. `VILMAA` replicated the updated `OpenCGA` configuration to the *Analysis* table in `HBase`. A `MapReduce` job was submitted to process each row in the *Analysis* table. The number of sample identifiers for each allele, no-call or failed call was used to calculate the reference allele count, alternate allele count, genotype count, MAF, minor genotype frequency (MGF), Hardy-Weinberg equilibrium (HWE), call rate (CR), pass rate (PR) and overall pass rate (OPR). The HWE p-values were calculated using the `HardyWeinbergCalculation` class part of the HTSJDK package.

The object model from OpenCB BioData was extended by `VILMAA` to include PR, CR and OPR values. The `VariantStats` class contained the variant statistics, transformed into the *proto* model and stored in `HBase`. The additional PR, CR and OPR values were stored as separate columns in HBase.

Cohort name	Samples	Description
W10K	9,110	NIHR BR-RD samples
UW10K	7,493	Unrelated NIHR BR-RD samples
UW10K_EUR	6,041	Unrelated European NIHR BR-RD samples
BRG	7,147	NIHR BR-RD without GEL samples
UBRG	6,214	Unrelated NIHR BR-RD samples without GEL
UBRG_EUR	4,922	Unrelated European NIHR BR-RD samples without GEL
UPAHC	6,385	Unrelated PAH controls
TEC_100	386	NIHR BR-RD samples with 100 bp read length
TEC_125	3,093	NIHR BR-RD samples with 125 bp read length
TEC_150	5,631	NIHR BR-RD samples with 150 bp read length
TEC_100_F	213	Female NIHR BR-RD samples with 100 bp read length
TEC_125_F	1,652	Female NIHR BR-RD samples with 125 bp read length
TEC_150_F	3,324	Female NIHR BR-RD samples with 150 bp read length
TEC_100_M	158	Male NIHR BR-RD samples with 100 bp read length
TEC_125_M	1,441	Male NIHR BR-RD samples with 125 bp read length
TEC_150_M	2,307	Male NIHR BR-RD samples with 150 bp read length

Table 4.3 List of NIHR BR-RD cohorts defined in `HBase`. Frequently used summary statistics were pre-calculated and stored separately for each cohort.

4.2.9 Novel disease causing gene discovery using `HBase`

The identification of novel disease-gene associations was described in chapter 3.3.5 on page **??** using VCF files and traditional command-line tools. A comparable analysis on protein-truncating variants was performed on the same NIHR BR-RD release with `VILMAA`. The same MAF filter of 1 in 10,000 was selected for unrelated non-PAH control, 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015), UK10K (UK10K Consortium et al., 2015) and ExAC (Lek et al., 2016) as described in chapter 3.3.3 on page 71. In addition, the gnomAD (Lek et al., 2016) MAF was also included as a filter of 1 in 10,000. The ‘Scan’ object was configured to select variants with an unrelated PAH control MAF >0.0001 and executed as a `MapReduce` job. The ‘Mapper’ step of the `MapReduce` job filtered variant objects by population frequency, biotype and consequence type. All populations were filtered by a MAF of 0.0001 in control data sets including ExAC (ALL), UK10K (ALL), 1000 Genomes (ALL), gnomAD (ALL). The biotype was restricted to protein coding and the protein-truncating consequence type included frameshift, start lost, stop gained / lost, splice donor / acceptor and transcript ablation / amplification variants. Sample ids with alternate alleles were extracted, grouped by cases and controls and submitted for each transcript. The ‘Reducer’ of the `MapReduce` job collated the cases and control samples per transcript and stored the result on the hadoop file system (HDFS). The number of samples were counted for each group, the canonical transcripts were selected and a one-tailed (greater) Fisher's exact test was performed in R.

4.2.10 Performance comparison

The speed and quality performance of VILMAA was compared to the VCF based method described in chapter 3.3 on page 61. Analysis in both methods were executed in a two step process and included the calculation of cohort summary statistics followed by a burden test of filtered variants. Cohort statistics, filtering, aggregation and burden test were consistent between both methods and further details can be found in the relevant chapters (see chapters 3.2.6, 4.2.8 and 4.2.9). The VCF based method stored variants in one file per chromosome, processed in parallel and the wall time reported for processing the chromosome. Total runtime was the sum of chromosome specific wall times. Chromosome 2 was the longest running process for each step. Time limiting factor was found to be the reading and writing of VCF file compared to the CPU performance. For these reasons, the BCFtool was limited to 2 or 3 cores for the cohort summary statistics and variant filtering respectively. Aggregation of samples per gene and the association test was performed in R using one core. For VILMAA, the wall time of the MapReduce job was extracted from the log file. The chapter 4.2.8 and 4.2.8 provide further details about the processing step.

Information to assess the variant quality were based on the NIHR BR-RD VCF and VILMAA release and comprised chromosome, variant type, reference allele, alternate allele, OPR/minOPR, WGS10K MAF, WGS10K unrelated EUR MAF, WGS10K unrelated EUR HWE For VCF files, BCFtools was used to extract the relevant measurements in tab delimited format for all variants and loaded into R for filtering, aggregation and visualisation. A MapReduce job was launched to extract information from VILMAA and stored as tab delimited format on the HDFS. Apache Spark loaded the tab delimited format, filtered and aggregated the information into minOPR bins before exporting the results in tab delimited format to the Linux file system. The results from VILMAA were visualised using R.

4.3 Results

The NIHR BR-RD release included the whole genomes from 9,110 individuals and contained 291M variants. The analyses are currently performed with file based methods, which are slow, time consuming and can not keep pace with the growth of the data. A fast and scalable solution does not exist to store data, merge samples, annotate variants and efficiently filter millions of variants. In order to overcome the file based bottleneck, I assessed the characteristics of variant information and the performance of not only SQL (NoSQL) technologies. Based on the assessment, I developed a variant infrastructure for loading, merging, annotating and analysing human genomes (VILMAA). The technology of VILMAA is based on the distributed processing framework Hadoop with the integrated HBase database as an analysis platform and variant store respectively. For responsible data sharing and the integration into the global genomic research infrastructure, I collaborated with the open source software for computational biology (OpenCB) initiative (see 4.1.1 on page 123) and extended the OpenCGA platform with VILMAA as a module (see Fig. 4.1). OpenCGA supports a standardised API, that allows the seamless authentication, loading, analysis and retrieval of genomic and clinical data. Users can interact with the API using OpenCGA command line tools or visually explore the data through the web-based integrative variant analysis (IVA) browser (https://github.com/opencb/iva). Variants stored in VILMAA were sent to CellBase for annotation and the API returned the results in data models compatible with OpenCGA as JSON. I analysed the annotated variants in VILMAA for novel disease-gene associations and compared the performance to the file based method (see **??** on page 93).

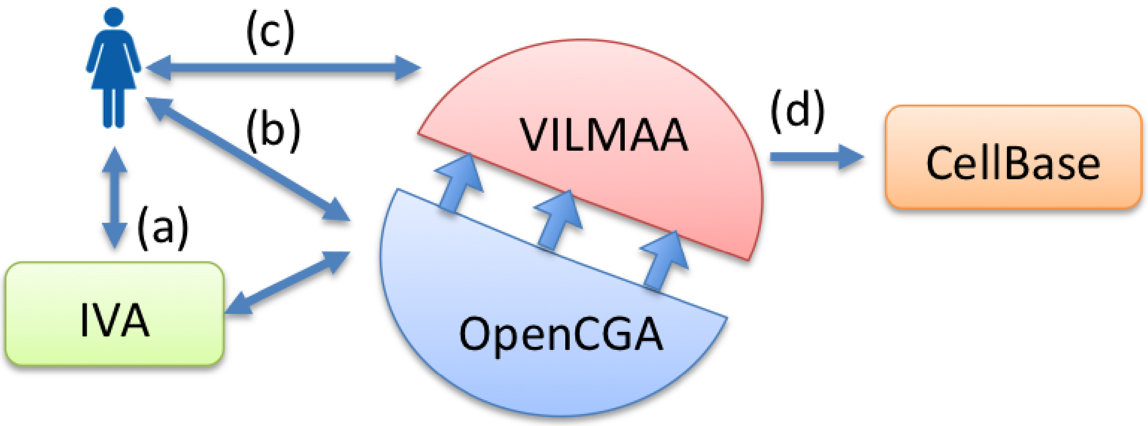
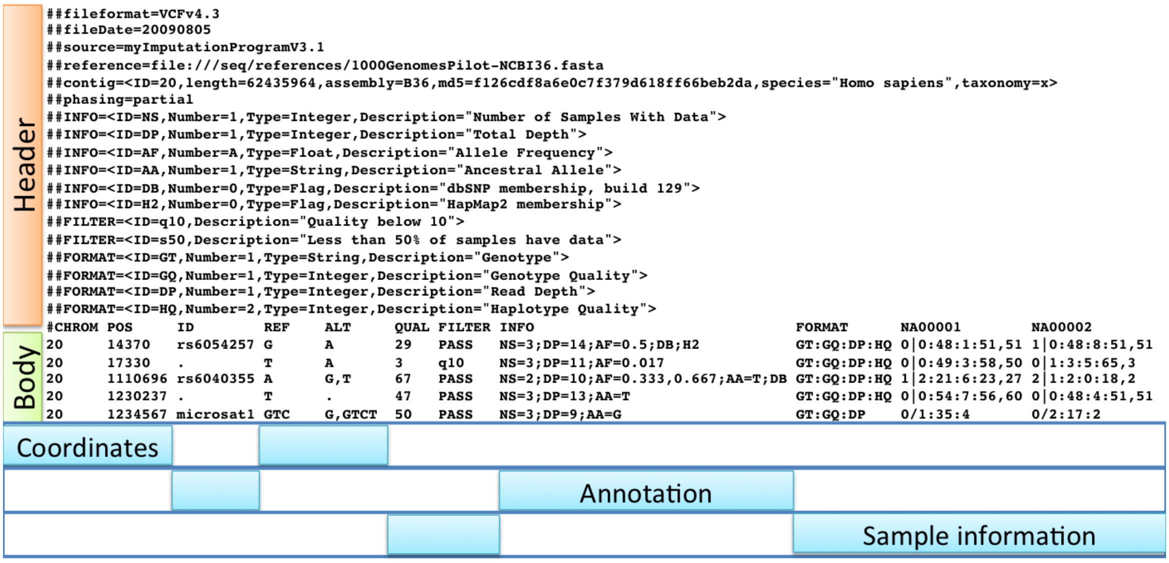


Figure 4.1 VILMAA integrates with various software modules. The user can interact with the system (a) through graphical user interfaces like the web-based integrative variant analysis (IVA) browser, (b) programmatically using the API or (c) by executing an analysis on the Hadoop framework. VILMAA sends requests to (d) CellBase for variant annotation purposes.

4.3.1 Data structure and characteristics

In order to design a scalable genome variant store, I analysed the data characteristics of variant entries in a VCF file and access patterns for genome wide analyses. The results were based on the January 2017 release of the NIHR BR-RD containing 9,110 samples. The VCF file structure is separated into a header and body section shown in Fig. 4.2. The header defines the available data and data types, while the values of the fields are stored in the body. For the analysis, the body was further separated into coordinates, sample information and annotation. Coordinates included positional information (chromosome, position), reference and alternate bases of the allele change and were relative to the selected reference genome. Sample information provided in the release file included genotypes, genotype quality, depth and allele depth per sample. I regarded the variant quality score (QUAL) and filter flags (FILTER) as sample information, since it was based on aggregated sample values. The coordinate and sample information were determined from WGS data while variant annotations relied on publicly available resources and additional software products. The INFO field stored these variant annotations and included diverse data types ranging from single numeric values to free text. The most extensive annotation was provided as text annotation by Ensembl VEP and represented a nested object structure with a mixture of numerical, categorical and textual information (see 3.2.4 on page 55).



```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2
```

Figure 4.2 Data type separation in VCF. The VCF file format contains coordinate, annotation and sample information, which are stored in separate columns. The columns for each data type are highlighted as blue blocks in separate tracks. Modified from Samtools organisation (2017).

Coordinates, annotation and sample data were stored together per row in the body of a VCF file. The annotated and compressed (gzip) VCF file for the NIHR BR-RD release was 3.9 TB, containing 291M entries and included 9,110 samples. On average, the size of each compressed variant entry was 13 KB. The removal of sample information reduced the file size to 0.97% (36.7 GB) and an average entry size of 0.126 KB. A detailed analysis of 100K random exonic variants without sample information found that 98.6% (26 MB) of compressed space was occupied by annotations compared to 0.35 MB (1.3%) for coordinates. The same 100K variants contained 911M genotype calls, which were homozygous reference (98.3%), heterozygous (1%), homozygous variant (0.6%) and no-call (0.02%).

Data access patterns of the VCF file were analysed for rare variants as part of a novel disease gene discovery, common variants as part of a genome wide association study and the merging of additional samples with existing variant data. First, the novel disease gene discovery (see Identification of novel disease-gene associations on page 93) extracted exonic regions of the genome from the VCF file based on coordinate ranges to reduce the number of 165M 'PASS' variants down to 2.6% (n=4.4M) exonic variants. The variant annotations were accessed to filter on sample summary statistics and functional predictions. These annotations further reduced the considered variants down to 0.2% (n=368K) rare, predicted deleterious variants. On average, the alternate allele was present in 1.6 samples (0.02%) per variant. Second, annotation based filtering was applied to obtain common variants (UWGS10K MAF >0.05) for the whole genome and retrieved 4.4% (n=7.2M) of the variants. On average, the alternate allele was present in 49% (n=4,479) of samples for a variant. Last, the access pattern for merging variants was estimated based on the required data types. These data types included the coordinates and sample information for the identification of the variant and the corresponding sample genotype data respectively. During the merge process, every data type of a variant entry was accessed and re-written as an entry in a new VCF file with additional sample information and updated sample summary annotation.

In summary, coordinate and annotation information occupied the smallest disk space of a VCF file and were found to be the most used filter option to reduce the number of variants for further analyses. Pre-calculated sample summary statistics were applied as a filter for the common and rare analysis as part of the annotation. Sample information were not directly used for filtering, but account for 99.3% of space with 98.3% occupied by reference calls. Merging of additional samples required reading and re-writing of all information.

4.3.2 Assessment of distributed NoSQL storage infrastructure

Different NoSQL storage implementations were evaluated due to their suitability for storing and processing variant information. The above identified VCF data type characteristics were used in the evaluation process. Traditional relational database management system (RDBMS) were designed to work on one single server and the data were stored in a specific schema designed for one purpose. Vertical scaling allowed to increase the capacity of RDBMS on one server by increasing the central processing unit (CPU) speed, number of cores, amount

Category	Implementation	Definition
Key-Value	Redis	Stores and retrieves documents based on a key.
Wide Column Stores	HBase, Cassandra	Similar to Key-Value store with the addition of separating information by dynamic columns.
Document storage	MongoDB	Stores structured (JSON) documents and indexes specific fields for fast access.
Graph database	Neo4J	Stores highly connected data

Table 4.4 High ranking NoSQL systems grouped by category. A NoSQL systems is selected for each category based on the ranked (solid IT gmbh, 2013).

of random-access memory (RAM) and improve the speed of the local disk space. In contrast, NoSQL systems were designed for horizontal scaling, which described the distribution of data across a group of servers. The requirement of data type or schema definition depended on the NoSQL category and specific implementation. The currently highest ranked NoSQL categories (solid IT gmbh, 2013) and their representatives are listed in Tab. 4.4. The `Neo4J` implementation is a graph database and optimised to store highly connected data. Variants are independently defined events and would not benefit using the graph structure. `Redis` stores key-value entries and was developed to cache specific web queries. The data storage and query facilities were optimised to retrieve binary data based on one specific key. This specific access pattern was not compatible with the access requirements of variant data. `MongoDB` is a structured document storage system which allows to create indices on specific fields to query the data. These indexes are similar to VCF position indices and can be applied on any predefined field. These fields would be searchable genome-wide and return the results in seconds, which would providing the required search functionality. `HBase` and `Cassandra` are wide column stores which provide a two dimensional key storage for binary data, where the row key and dynamic column keys are used to access information. `Cassandra` requires the exact row key to access data and does not provide a facility to access regions of rows, but allows to scan columns for regions. Rows in `HBase` are sorted and provide range queries for regional searches. The documentation[1] states that `HBase` scales to billions of rows and millions of columns and provides fast regional access as well as providing the computational back-end to scan the whole dataset for specific annotations. `HBase` is a module of the distributed processing Hadoop framework and provides the computational infrastructure to process the stored data locally.

[1] http://hbase.apache.org/book.html#arch.overview.when

MongoDB and HBase were further subjected to a more detailed evaluation because of their capabilities and likely choice as a variant store. The evaluation was based on the Yahoo! Cloud Serving Benchmark (YCSB) framework (Cooper et al., 2010). The YCSB is an open source implementation to evaluate data-serving and data-storing systems. In 2013, the company Altoros assessed the performance of HBase and MondoDB amongst others in different scenarios using the YCSB framework (Altoros Systems, 2013). The framework generated 100M entries of 1KB size entries and simulated different read / update / write workloads. I selected the simulated workloads A (50% read, 50% update) and B (5% read with 95% update) to estimate the usage during merging and annotating variants respectively shown in Fig. 4.3. The latency of the update operations in both workflows did not increase for HBase, while MongoDB showed rising latency by increasing the number of operations. The difference for read operations were not significant in workflow A, but showed a performance degrade of the MongoDB throughput in workflow B.

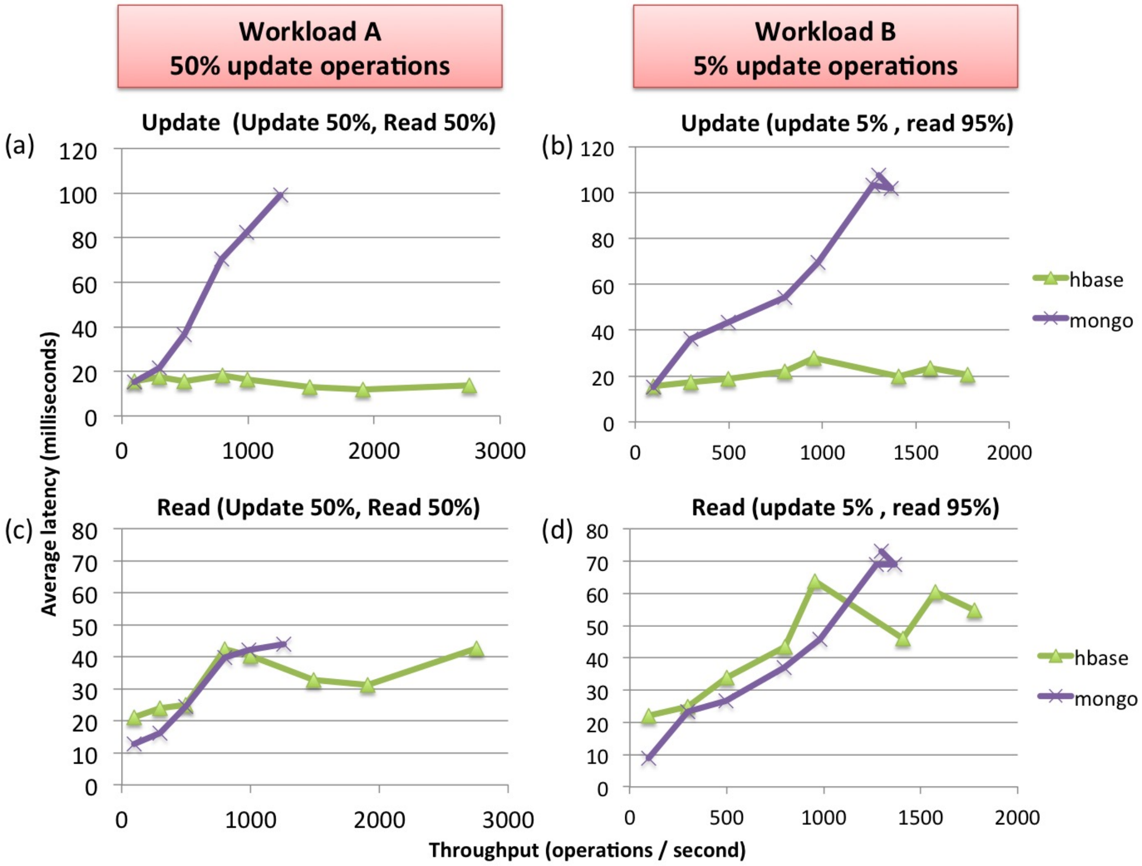


Figure 4.3 MongoDB and HBase performance comparison. The latency of update and read performance is displayed for two different workload models. Update operations are measured in (a) and (b), read operations are measured in (c) and (d) for workload models A and B respectively. Adapted from Altoros Systems (2013).

The evaluation of NoSQL storage implementations showed significant differences between NoSQL category representatives and even within a category. `MongoDB` and `HBase` were further tested due to their suitability to store and query variant information. `HBase` showed consistent lower latency during simulated update and read operations and was selected as the NoSQL variant storage backend.

4.3.3 Normalise, transform and load single sample gVCF into HBase

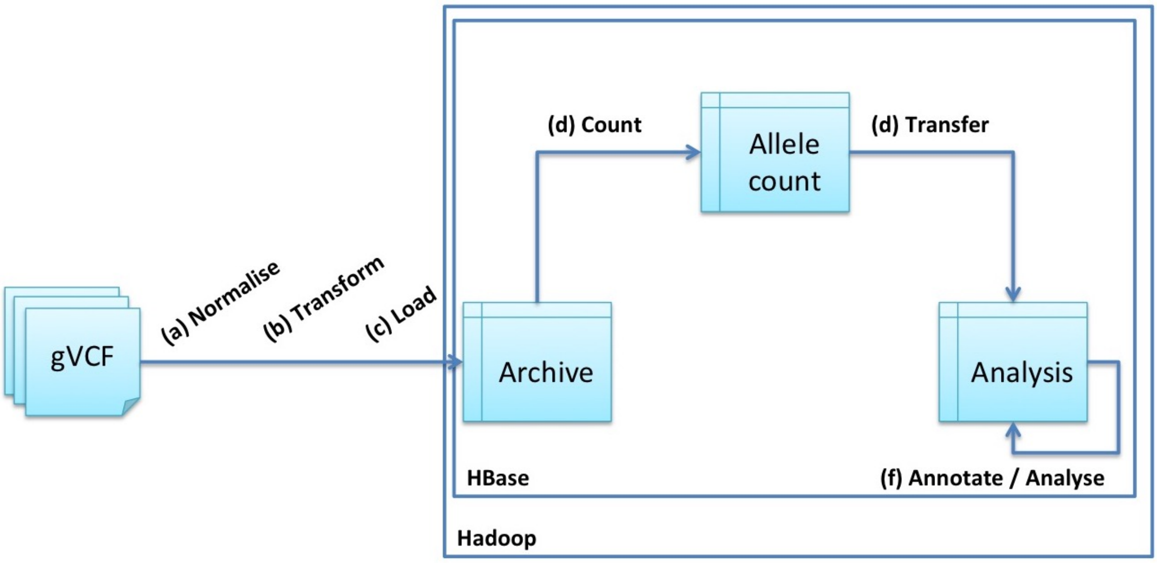


Figure 4.4 VILMAA variant workflow. Workflow shows required steps to move from a locally stored gVCF to an analysis ready `HBase` table. Analyses can be performed in minutes on the merged and annotated *Analysis* table.

The distributed database `HBase` was selected to store and provide fast access to genome variation information and their annotations. First, I developed an `HBase` schema to store a complete representation of 10K single sample genome variation data and found an efficient way to load the normalised gVCF files into the schema. The storage architecture provided by `HBase` can be described as a spreadsheet with rows, columns and values stored in cells. In contrast to spreadsheets, the names of columns and rows are user defined and contain information. Each cell can be accessed by the combination of row and column names and the access time to these cells are in the order of milliseconds. Based on this information, I designed the *Archive* table to provide positional and sample specific variant information described in Tab. 4.5. The column name identifies the gVCF file of the sample, the row stores the position of a 1 Kb region (slice) and the actual variants of the slices are stored as binary data in the cell. The VCF header definition is stored with other meta information

Description	Row	Column	Type
Variant information per file for region	<chr>:<slice position>	<study_id>_<file_id>	proto object
VCF header information per file	_METADATA	<study_id>_<file_id>	JSON
List of loaded file ids separated by ‘;’	_METADATA	_METADATA	string

Table 4.5 *Archive* table schema. Variant information of an individual file are grouped in slices. The row key stores the position of the slice region and the column name identifies the specific file in a study. The value of a cell contains the actual variant information of a slice and is stored as protocol buffer. VCF header and other meta information are stored in JSON format in the additional metadata row. A metadata row keeps track of all loaded files.

per sample in a separate row. I determined the 1 Kb slice size by comparing the number of variant entries bridging across slices while keeping sufficient number of rows for efficient parallel processing of slices (Tab. 4.6). Variants bridging across multiple slices were stored in all slices to avoid additional database requests.

The data structure of the slice information was designed using a data structure serialisation method called protocol buffers, also known as *proto* (https://developers.google.com/protocol-buffers/). The serialisation method allows to generate JAVA classes from the designed schema, which are filled with data. The filled information are then serialized to be stored in a file as binary format using the provided read and write functionality. *Proto* is optimised for speed and compact representation of data. The developed *VcfSlice* schema (see Listing A.1 on page 189) required the loss-less representation and fast access to the object model.

	100 bp	1 Kb	10 Kb
Number of slices	30 M	3 M	309 K
Average variants per slice	1	13	125
Batch size of 500 samples	625	6,250	62,500
Slices with bridging variants (1 sample)	30%	86%	91%

Table 4.6 Slice size assessment. The average number of variants are calculated for different slice sizes. Variants with a start and end position in different slices are regarded as bridging variants and are stored in all affected slices.

After designing the database schema, I focused on the normalisation, transformation and efficient loading of gVCF files into `HBase`. The normalisation (Fig. 4.4(a)) of variants was performed in two steps. First, the left alignment and removal of redundant bases of variants was performed for single sample gVCF files using `BCFtools`. The normalisation took on average 35 minutes, was performed in parallel and the results were stored as gVCF files.

Second, INDELs were normalised, the anchoring reference bases were removed, transformed (Fig. 4.4(b)) and stored as *proto* files using the OpenCGA package. The normalisation, transformation, grouping into slices and serialisation to *proto* took on average 15 minutes per file. *Proto* files were optimised for data loading and grouped into 1 Kb slices ready to be loaded into the *Archive* table in `HBase`. The loading process (Fig. 4.4(c)) took on average 106 seconds per sample and 4 files were loaded in parallel without performance reduction.

The normalisation, transformation and loading of one file took 52 minutes and occupied 1 CPU. Processing 16 files in parallel reduced the overall run-time significantly down to 10.2 minutes per file with further parallelisation possible. Such reduced run-times make it feasible to process 10K samples.

4.3.4 Variant conflict resolution

Single sample gVCF data were loaded into `HBase` with the assumption that every base of the genome is represented, but only represented once. I implemented a consistency check to highlight conflicting variant calls and missing regions. These conflicting calls involved overlapping or duplicated INDEL calls of lower quality. Fig. 4.5 shows two variant calls in the same individual, where an insertion is called twice, but differently. I implemented a rule engine to resolve these conflicts. The pseudocode for the rule engine is shown in Listing 4.1 that selects the most likely calls first and rejects following conflicting calls. Regions with missing calls were filled with no-calls.

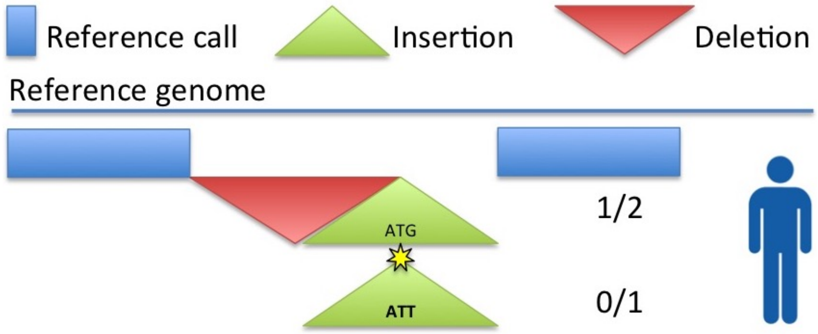


Figure 4.5 Example of an insertion conflict. Two insertions (1/2 and 0/1) are called at the same position. The conflict is highlighted by the star symbol. The variant calls provide different genotypes and alternates for the insertion (ATT versus ATG).

Listing 4.1 Conflict resolution pseudo code

```
function SortCalls:
Prefer variant calls over reference calls
Prefer PASS over other flags
Prefer other flags over SiteConfict
Prefer higher quality scores
Prefer lower start
Prefer lower end
if not decided
Random decision

FOR EACH individual IN individuals
FOR EACH conflictingCallSet IN individual
SortCalls: conflictingCallSet
WHILE call IN conflictingCallSet
IF noConflict: call , finalCallSet
ADD call TO finalCallSet
ELSE
## do nothing —> reject call
fillMissingRegionsWithNoCall: finalCallSet
```

4.3.5 Incremental variant merge using Hadoop infrastructure

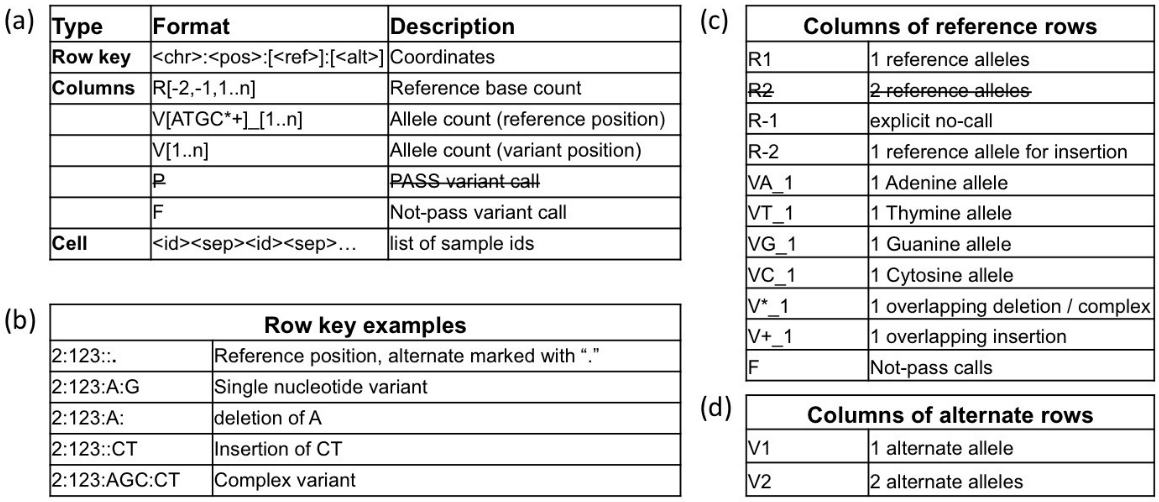
Single sample variant information were loaded into the *Archive* table, which was designed to provide fast regional access grouped into slices. One slice contains th information for a 1Kb region. The columns of each slice identified individual samples with their respective variant information. Variant information included no-call, reference and non-reference variant calls. Merging one specific position required the overlapping variants from all individuals to be available at the same time. Adding one additional sample with variants not observed before requires the reloading of data from all individuals for these positions. On average, one sample contains 13 calls (reference, no-call or variant call) for a 1Kb region. Genotype and associated data were converted into the *proto* format per slice, which required 681 bytes of disk space. For an average slice, the file size for 10K samples was 6 MB, contained 130K calls and was efficient to load and process. Regions of 1Kb with high number of variation increased the file size for one sample from 681 bytes to 25 KB containing 500 calls or more. These regions required 240 MB of compressed disk space for 10K samples and encoded 5M calls. Loading and reprocessing these information was not feasible for the growing amount of data.

An incremental merge approach was required to process variant information only once per sample. For this purpose, I designed an incremental two step process, that utilises the scalability and update functionality of HBase. Variants are decomposed into the separate alleles and the observed alleles for each individual are counted based on the recorded genotype. Tab. 4.7 illustrates the decomposition of variant calls at one position for different samples and the process of grouping samples with the same number of present alleles. The allele count representation allows the incremental adding of additional samples to each group and contains sufficient information to recreate the original genotype for each sample. Insertions and deletions are not fully represented in Tab. 4.7 and the differences for storing variants other than SNVs is addressed in the below table schema definition.

Sample	Variant	GT	A	T	C	DEL	INS
S1	1:123_A/T	0/1	1x	1x			
S2	1:123_A/C	1/1			2x		
S3	1:123_A/.	0/0	2x				
S4	1:123_A/-	0/1	1x			1x	
S5	1:123_A/AG	0/1	1x				1x
S6	1:123_A/.	0/0	2x				
Count	**A**:1x[S1,S4,S5]; **A**:2x[S3,S6]; **IND**:1x[S5] **T**:1x[S1]; **C**:2x[S2]; **DEL**:1x[S4];						

Table 4.7 Example variant decomposition and allele count for one position. Variants from 6 samples of the sample position are decomposed into their individual alleles. These decomposed information are aggregated and the final count listed in the last row.

Variants were decomposed to allow incremental adding of samples and this allele count information exists for every base of the genome. Fig. 4.6 describes the *Allele_count* schema, which I designed to store a sparse representation of the decomposed information. The reference row (Fig. 4.6(c)) contains the count information. Homozygous reference calls are inferred instead of stored, which were identified as the most common genotype (see Data structure and characteristics in 4.3.1). The reference row contains the count for overlapping deletions (including complex alleles) and anchoring position for insertions, but does not specify the precise allele. The overlap of the anchoring position in insertions requires to store the reference allele of INDELs separately to avoid over-counting the number of reference alleles. The precise alterations are stored in separate rows, where the columns (Fig. 4.6(d)) contain the allele count, only for the specific variants. The quality of calls is captured in the non-pass (any other flag than ‘PASS’) column in the reference row, while PASS calls are inferred. After counting, sample ids are appended to the required columns without prior access to the table. This incremental update is achieved through the ‘Append’ command in HBase, which creates required rows and columns, if needed.



(a)

Type	Format	Description
Row key	<chr>:<pos>:[<ref>]:[<alt>]	Coordinates
Columns	R[-2,-1,1..n]	Reference base count
	V[ATGC*+]_[1..n]	Allele count (reference position)
	V[1..n]	Allele count (variant position)
	~~P~~	~~PASS variant call~~
	F	Not-pass variant call
Cell	<id><sep><id><sep>…	list of sample ids

(b)

Row key examples	
2:123::.	Reference position, alternate marked with “.”
2:123:A:G	Single nucleotide variant
2:123:A:	deletion of A
2:123::CT	Insertion of CT
2:123:AGC:CT	Complex variant

(c)

Columns of reference rows	
R1	1 reference alleles
~~R2~~	~~2 reference alleles~~
R-1	explicit no-call
R-2	1 reference allele for insertion
VA_1	1 Adenine allele
VT_1	1 Thymine allele
VG_1	1 Guanine allele
VC_1	1 Cytosine allele
V*_1	1 overlapping deletion / complex
V+_1	1 overlapping insertion
F	Not-pass calls

(d)

Columns of alternate rows	
V1	1 alternate allele
V2	2 alternate alleles

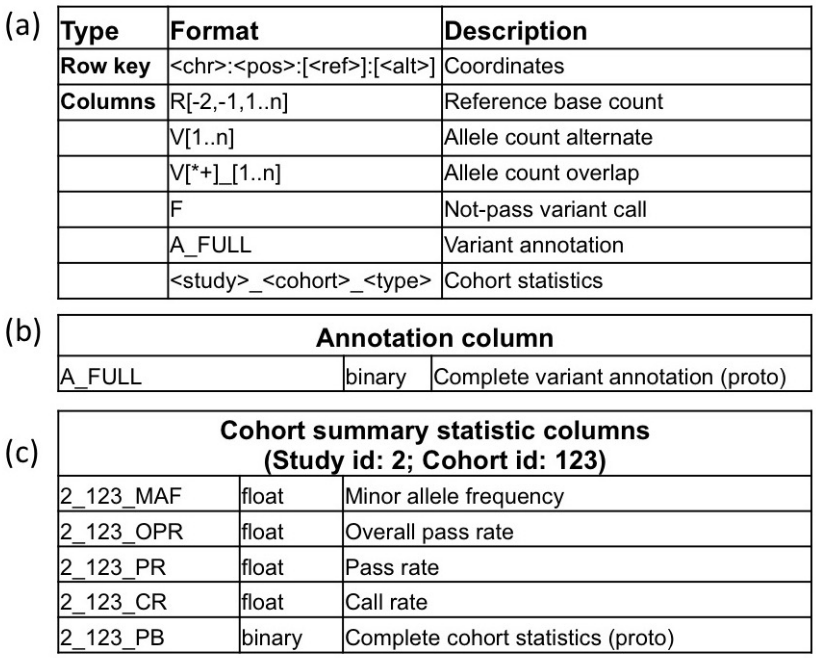
Figure 4.6 *Allele_count* table schema description. The (a) structure of the table shows the row and columns. All cells store a list of sample identifies using the array delimiter from Apache Phoenix. Genomic position, reference and alternate are encoded in the row key, while the sample genotypes are stored in the columns. Variant ‘PASS’ information are not saved, but is inferred from the not-pass variant calls. (b) Reference and alternate bases are encoded in the row key. Reference positions stores no reference bases and encode ‘.’ as the alternate. (c) Reference row entries contain allele counts of all observed single bases (reference and alternate), overlapping insertions and deletions encoded as ‘+’ and ‘*’ respectively. Samples with homozygous reference calls are not recorded and are inferred from the variant context. (d) Alternate row entries store samples with variant alleles only.

The complete process of allele counting of samples involved the conflict resolution, decomposition and the incremental adding of sample ids to HBase. This process was run for 12,551 samples and Tab. 4.8 shows that a speed of 42 seconds per sample was achieved. A table size of 3.6 billion rows is feasible in HBase without performance deterioration. Larger collections of samples perform better due to a reduced waiting time during the submission using the append command. The submission method was implemented as an asynchronous process to maximise the uninterrupted processing of the data.

The count table contained reference and alternate allele counts with 3.02 billion rows for incremental merging of samples. For analysis purposes, the alternate alleles were extracted into the *Analysis* table described in Fig. 4.7, which is of similar structure to the *Allele_count* table. The *Analysis* table enables the storage of (Fig. 4.7(b)) variant annotations and (Fig. 4.7(c)) pre-calculated cohort statistics. Specific cohort frequencies are stored in separate columns for improved query times. I implemented the transfer as a MapReduce job,

Samples new	Samples total	Time (total)	Time / sample	Rows (total)
5,000	5,000	59 hours	42 seconds	2.27 billion
3,729	8,729	46 hours	44 seconds	2.65 billion
928	9,657	19.5 hours	75 seconds	2.83 billion
2,894	12,551	33 hours	41 seconds	3.02 billion

Table 4.8 Incremental allele count processing times. The wall times required to incrementally add variants to the *Allele_count* table.



(a)

Type	Format	Description
Row key	<chr>:<pos>:[<ref>]:[<alt>]	Coordinates
Columns	R[-2,-1,1..n]	Reference base count
	V[1..n]	Allele count alternate
	V[*+]_[1..n]	Allele count overlap
	F	Not-pass variant call
	A_FULL	Variant annotation
	<study>_<cohort>_<type>	Cohort statistics

(b)

Annotation column		
A_FULL	binary	Complete variant annotation (proto)

(c)

Cohort summary statistic columns (Study id: 2; Cohort id: 123)		
2_123_MAF	float	Minor allele frequency
2_123_OPR	float	Overall pass rate
2_123_PR	float	Pass rate
2_123_CR	float	Call rate
2_123_PB	binary	Complete cohort statistics (proto)

Figure 4.7 *Analysis* table schema description. The (a) structure of the table shows the row key composition and the encoding of the column names. Allele count columns start with ‘R’ or ‘V’ for reference or variant respectively. Overlapping insertions and deletions are represented as ‘+’ and ‘*’ respectively. The number at the end of the column name represents the number of observed alleles while ‘R-1’ and ‘R-2’ represent a no-call and one reference allele present in an insertion respectively. Variant and reference allele cells contain a list of sample identifiers for the reference or alternate allele. (b) Variant annotation and population frequencies are stored in the column ‘A_FULL’ as proto. (c) Study and cohort identifiers are encoded with some statistical measurement into the column name. The complete allele summary statistics are stored in ‘2_123_PB’ as proto.

which processes each row from the *Allele_count* table separately. Rows are read in sorted order that ensures the reference row appears just before a variant row entry. Reference and variant information are extracted and the combined allele counts submitted to the *Analysis* table. Tab. 4.9 shows the transfer times with resulting variant counts for two sample sets and highlights that the HBase was capable of repopulating 347 million variants in less than 3 hours. Existing variant annotations were not overwritten during this process.

Samples	Count rows	Variant rows	Time
8,729	2.65 B	290 M	2h 28m
12,551	3.02 B	347 M	2h 58m

Table 4.9 Variant transfer times. The required wall time to detect and transfer variant calls from the *Allele_count* to the *Analysis* table.

4.3.6 Variant annotation and cohort statistics

Variants were merged across 12,551 samples and stored in the *Analysis* table. The resulting 347M variants required variant annotations and cohort summary statistic calculation for further analysis. For this purpose, the online CellBase annotation service was utilised to retrieve available population frequencies, gene model and variant predictions. I implemented a MapReduce job to annotate all variants in the *Analysis* table with CellBase, which took 2 days and 18 hours. The annotation process extracts variant (position, reference and alternate) information from the row key and submits variants in batches of 200 to CellBase for annotation. Annotation data returned from CellBase are stored in the *Analysis* table. Summary statistics were calculated based on samples included in the NIHR BR-RD release. The required cohorts (see Tab. 4.3 on page 132) were first defined in OpenCGA before the updated configuration was synchronised with HBase. I implemented a MapReduce job to calculate the reference allele count, alternate allele count, genotype count, minor allele frequency (MAF), minor genotype frequency (MGF), Hardy-Weinberg equilibrium (HWE), call rate (CR), pass rate (PR) and overall pass rate (OPR). An allele count focused design of the *Analysis* table allowed the efficient calculation of the allele frequencies. The HWE calculation was performed by the open source HTSJDK implementation. The calculation of the summary statistic took 2 hours and 46 minutes, and included 347 M variants, 16 cohorts and 9 values per cohort.

4.3.7 Variant noise reduction

The variants were normalised, merged and annotated in HBase as described above. To distinguish between biological and technical variants, I calculated the overall pass rate (OPR = CR x PR), which is composed of the call rate (CR) and pass rate (PR). The OPR calculated in HBase differs to the OPR calculated using the AGG tool (see Variant noise reduction in 3.3.1 on page 67), which uses the pass frequency (PF). The OPR was calculated for each technical cohort (100, 125 and 150 bp read length) and separately for each gender (male and female). Fig. 4.8(a) shows the sensitivity of the OPR to protocol changes. Due to technical differences, I selected the minimum OPR (minOPR) of all technical cohorts for filtering. Only female samples and only male samples were included for the minOPR calculation on the X and Y chromosome respectively. The analysis in Fig. 4.8(b) shows the distribution of common and rare variants for the minOPR. In general, rare variants accumulated closer to the extreme minOPR (0 and 1) and the separation was more distinct for SNV compared to INDELs (not shown). I observed 27.6% of SNVs with a minOPR <0.5 compared to 23% of insertions and 14% of deletions.

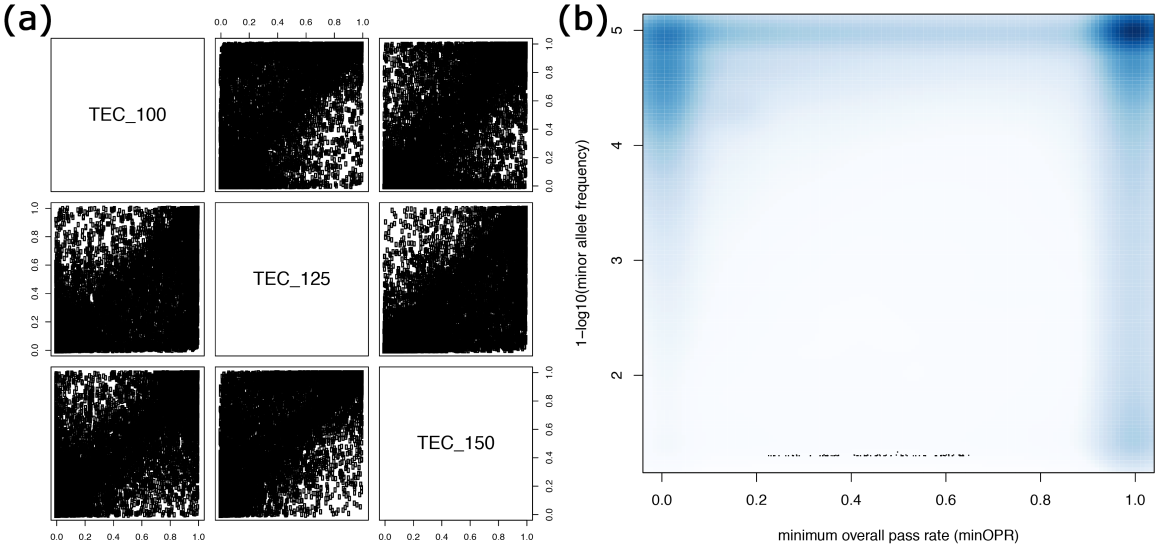


Figure 4.8 Technical OPR bias and minOPR MAF distribution. (A) The calculated OPR for the same variants are compared between different technical cohorts. The comparison shows variants that reach a high OPR in one cohort, but a low OPR in another cohort, which creates read length specific pattern. (B) The minOPR value is selected from the technical cohort and compared to the MAF of SNVs observed in the NIHR BR-RD cohort. The darker areas indicate an enrichment of variants close to 0 for rare and close to 1 of the minOPR for rare and common variants.

I determined the biological relevance of the observed separation of minOPR values by using the Ts/Tv ratio and HWE to assess biological and population specific properties for variants in different minOPR bins. The analysis in Fig. 4.9 confirms an enrichment of non-random base changes in minOPR bins closer to 1 for both Ts/Tv (a) and HWE (b) measurements. A focused analysis of the top three minOPR bins confirmed a conservative minOPR cutoff of ≥ 0.99 for the internal data release. We also found that insertions were in the expected HWE range until an accumulation of insertions. Deletions are excluded from the HWE analysis because overlapping deletions skew the HWE calculation. The composition of variant types are recorded in Tab. 4.10 and I found that 46% of variants were retained after filtering with a minOPR ≥ 0.99.

	SNV	MNV	INDELs	Out of total
All variants	84%	1%	15%	100%
≥ 0.99 OPR	90.5%	0.5%	9%	46%

Table 4.10 Breakdown of different variant types with and without filtering for high confident (≥ 0.99 minOPR) variants.

4.3.8 Novel disease-gene associations using HBase

The variants of 12,551 samples were loaded, merged, annotated and cohort specific statistics calculated in HBase. The purpose of the infrastructure development was to facilitate genome wide analysis and accelerate novel gene discovery in rare diseases. With the acceleration in mind, I reimplemented the novel gene discovery for protein-truncating variants in the PAH cohort as described in Identification of novel disease-gene associations on page 93 using the Hadoop infrastructure. The implementation included the selection of protein-truncating variants, filter variants based on population and control cohort frequencies and aggregate samples into unrelated PAH controls and unrelated PAH index cases. Samples were aggregated per transcript and tested for over-representation in unrelated PAH index cases. The analysis was executed as a MapReduce job on the Hadoop cluster, processed 347M variants in 12,551 samples and returned the aggregated results in 32 minutes. Tab. 4.11 shows the results for the HBase and VCF based analysis and highlights the differences in the number variants found comparing cases with controls. In the first 10 most significant results from HBase, there are only two genes with matching counts. I investigated these differences and found that the minOPR was below 0.99 for 10 variants (12 subjects), the VCF OPR was below 0.8 for 2 variants (2 subjects), UK10KWES annotations were missing for 5 variants (5 subjects), gnomAD annotation not part of VCF 1 variant (1 subject) and

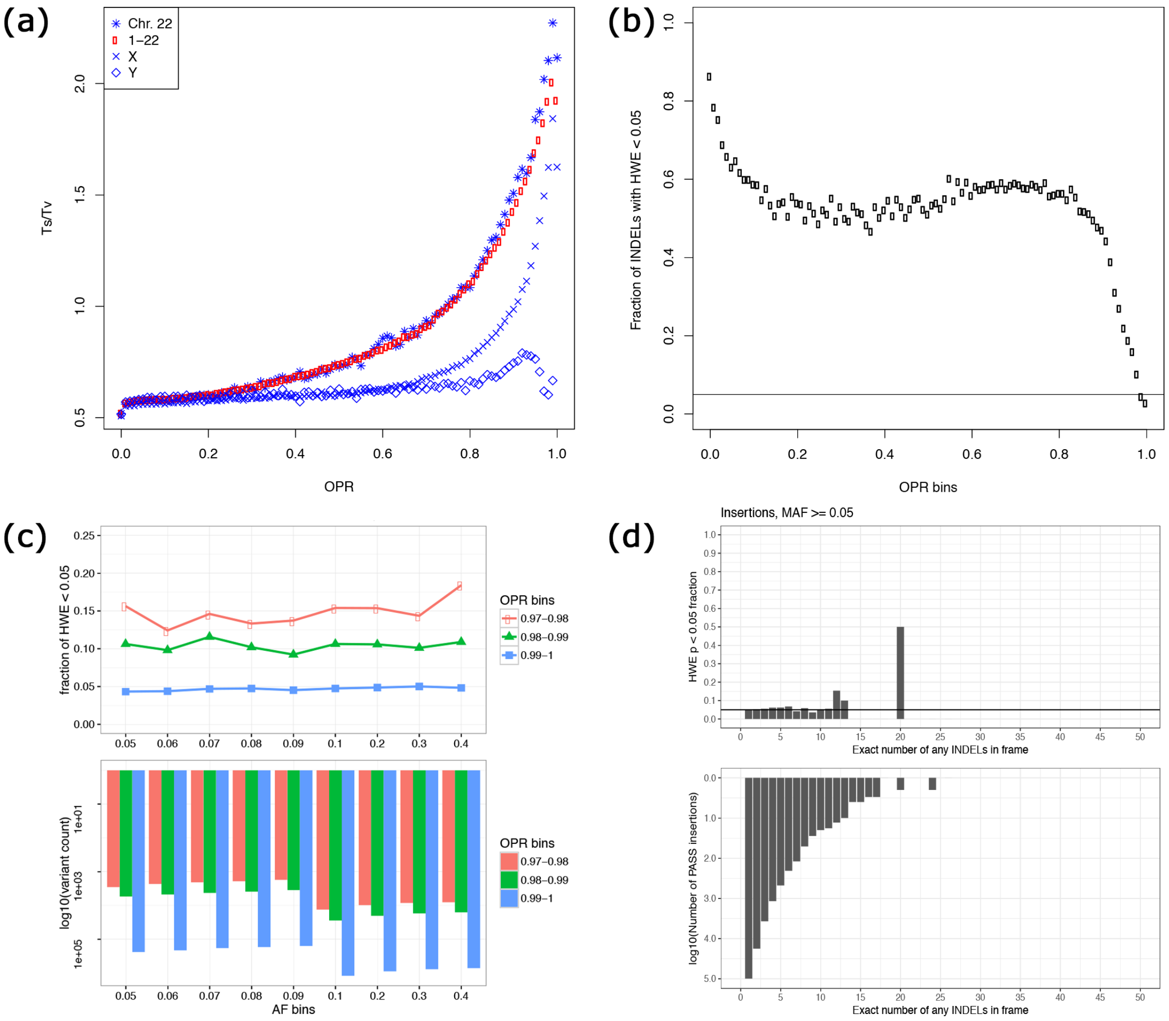


Figure 4.9 (a) The Ts/Tv ratio was calculated on SNVs per chromosome for different minOPR bins and shows a consistent pattern. (b) The fraction of Hardy-Weinberg equilibrium (HWE) less than 0.05 was calculated for INDELs in minOPR bins. The horizontal line is the expected value (0.05) and shows an improvement of the HWE fraction with increasing minOPR values. (c) The fraction of HWE less than 0.05 was calculated for SNVs and analysed for the top three minOPR bins across different MAF collections, with the number of available SNVs shown underneath. The expected values were reached with a minOPR cut-off greater or equal to 0.99. The analysis of insertions gives comparable results (data not shown). (d) HWE and frequency of insertions are consistent across different lengths after filtering. The insertions were collected in a frame size of 10bp.

protein-truncating consequence types in the VCF file did not have annotations in HBase for 2 variants (2 subjects). The majority of missing variants were due to different OPR cutoffs and OPR calculations. The UK10KWES annotation was shared by collaborators and not publicly available, which explained the missing frequencies in HBase. GnomAD frequencies were not part of the VCF data release and not used for the VCF file based analysis. Protein-truncating consequence types annotations should not be missing and was down to the variant representation in CellBase. CellBase does not provide consequence types for MNV but instead normalised the variants into SNVs and provides individual annotations for each SNV.

VIMLAA confirmed the scalability of HBase to store and process variant information for 10K samples. The incremental merging process utilised the full potential of HBase stable performance and diminishes several iterations of data processing.

	VILMAA protein truncating analysis			VCF based protein truncating analysis		
HGNC	Cases	Controls	Fisher's exact p-value	Cases	Controls	Fisher's exact p-value
BMPR2	89	2	1.18E-74	92	1	5.58E-79
EIF2AK4	13	7	2.22E-07	12	6	4.33E-07
ATP13A3	6	1	4.60E-05	6	0	7.38E-06
EVI5	5	1	0.00028	5	1	0.00028
SLC36A2	5	1	0.00028	3	2	0.02188
KDR	4	0	0.00038	4	0	0.00038
TBX4	4	0	0.00038	8	0	1.43E-07
KIF4B	4	0	0.00038	4	1	0.001690
SRM	3	0	0.00274	4	0	0.00038
PRR22	1	0	0.14011	4	0	0.00038

Table 4.11 Protein-truncating variant analysis comparison. Results with a p-value <0.0004 were included from the VILMAA and VCF based analysis.

4.3.9 Performance comparison

Variants from NIHR BR-RD subjects were analysed using the VCF file based (see chapter 3.3 on page 61) and the distributed compute infrastructure VILMAA (see chapter 4.3.8) method. Both methods selected the same subjects for the case and control cohort, and performed an association test in a two step process. Firstly the cohort frequencies were calculated and secondly subjects with rare deleterious variants were selected and tested for a disease-gene association. Using the available resources, I measured the performance of these two steps (see Tab. 4.12) for both methods. VILMAA compared to VCF was found to be 77 and 1000 times faster for calculating summary statistics and filtering variants including the burden test respectively. For VCF, chromosome 2 was found to have the longest runtime by processing chromosomes individually and indicative for a parallel computational approach with standard tools. The runtime for VCF on chromosome 2 was listed separately as a reference value and VILMAA performed 6 and 78 times faster for calculating summary statistics and filtering variants including the burden test respectively. VILMAA showed a speed improvement for each benchmark and was able to perform a burden test including variant filtering in minutes compared to days.

	VCF		VILMAA
	total	chromosome 2	total
Summary statistics	8d 23h 16m	18h 4m	2h 46m
Filter variants & burden test	22d 5h 50m	1d 17h 51m	32m

Table 4.12 Runtime comparison of rare variant burden test. The wall time was measured for calculating the summary statistics for defined cohorts for all variants. Variants were filtered for rare deleterious variants by considering calculated allele frequencies as well as variant annotations. Samples of the remaining variants were aggregated and included in the burden test. Separate measurements were provided for chromosome 2, which had the longest runtime and as a reference value for analysing chromosomes in parallel. The VCF method used 2 cores for calculating the summary statistics and 3 cores for variant filtering followed by the burden test. VILMAA showed a speed improvement for each listed benchmark.

After confirming the speed improvements, I compared the performance of the overall pass rate (OPR) and minimum OPR (minOPR) quality measurement, which were differently calculated (see chapter 3.3.1 and 4.3.6). The chapter 3.3.1 (Variant noise reduction) on page 67 describes the VCF based calculation of the OPR value including the call rate (CR) and pass rate (PR). VILMAA extended the OPR method by selecting the minimum OPR (minOPR) from OPRs calculated for different technical cohorts to reduce pipeline biases (see chapter 4.3.6 on page 146). The differences between the VCF based OPR and VILMAA based minOPR were explored in two different ways. Firstly I calculated the transition (Ts) and transversion(Tv) ratio from SNVs in different OPR or minOPR bins (see Fig. 4.10). For substitution events, there are two possible transitions and four possible transversion. If substitutions occur randomly, then the Ts/Tv ratio should be 0.5. Non-random events for the whole human genome occurs at a ratio of around 2.0. The OPR bin of 0 was found to have a Ts/Tv of 0.5 for both methods and an increase of the radio enriches OPR bins for biological events. The minOPR (VILMAA) shows a slow increase in Ts/Tv ratio with increasing OPR

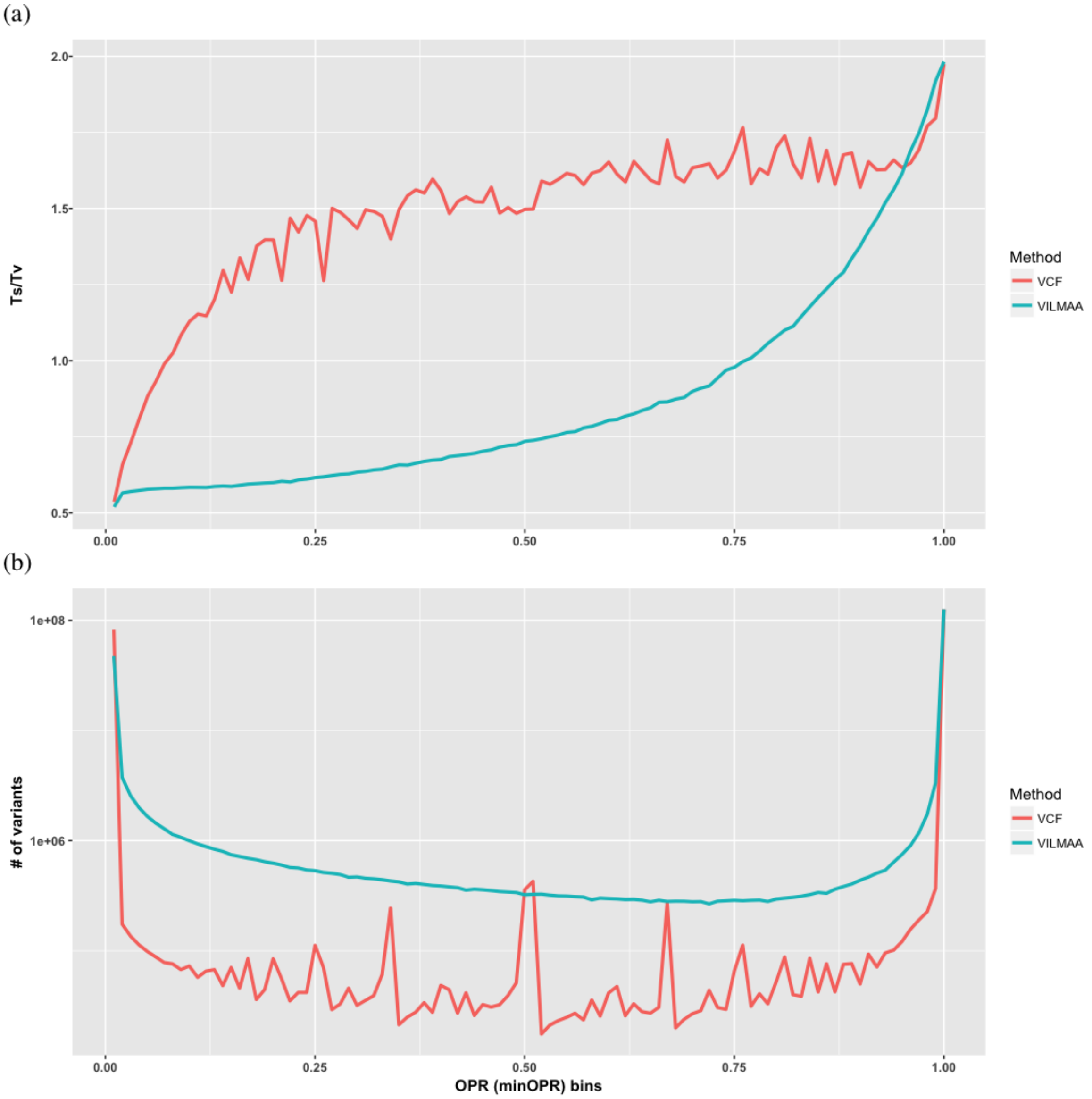


Figure 4.10 Transition/Transversion (Ts/Tv) ratio was calculated in different OPR bins for each method. OPR and minOPR values were used to create OPR bins for the VCF and VILMAA based method respectively. (a) The method specific Ts/Tv ratio is displayed by lines and an increase in OPR shows a method specific pattern. VILMAA shows a gradual rise in Ts/Tv for increasing OPR values compared to VCF. (b) Each lines shows the number of SNVs included in the Ts/Tv calculation for the OPR bin specific to the method.

value compared to a the OPR (VCF). The Ts/Tv of 1 was reached at the OPR bin of 0.76 and 0.07 for VILMAA and VCF respectively. The gradual Ts/Tv increase in minOPR suggests a better separation of technical and biological events compared to OPR. In both cases, the Ts/Tv comes close to the expected 2.0 ratio in the maximum OPR bin. Secondly I utilise the assumption of Hardy–Weinberg equilibrium (HWE), which states that the frequencies of heterozygous and homozygous genotypes in a population will remain constant across generations. The fraction of variants with a HWE of <0.05 in a large collection is expected to be 0.05 for common variants (MAF $\geq$0.05). Figure 4.11 shows that variants with an OPR between 0.99 and 1 reach the expected fraction of 0.05 in common variants. The minOPR (VILMAA) is more consistent compared to OPR (VCF) across different MAF bins. A reduction of the OPR shows less of an affect in minOPR compared to OPR. The quality measurements of Ts/Tv and HWE suggests a better performance in selecting variants based on the minOPR provided by VILMAA.

4.4 Discussion

The NIHR BR-RD received 13K variant data sets from Illumina over a 4 year period. Available single sample genome VCF (gVCF) files were aggregated into a multi-sample VCF file and the variants annotated with Ensembl variant effect predictor (VEP). The rising number of samples increased the file size and the preparation time. Longer analyses times using available software tools hindered the exploration of the genome and a new scalable approach was required to keep pace with the soaring amount of data. I evaluated variant information accessed by different analyses and characterised the stored data types in the merged VCF file. The evaluation highlighted the contrast between the high and low frequency use of the minority and majority of the data respectively.

4.4.1 NoSql technologies

My search for new technologies to store variant information was inspired by Google, Facebook and Twitter. All these companies applied different NoSql technologies to provide access to petabytes of information in seconds. I selected the most popular implementation for each NoSql category and evaluated the suitability to store, aggregate and analyse variant information for 13K individuals. The evaluation was based on the previous identified analysis and data characteristics of multi-sample VCF files, which found `HBase` as the best candidate. One major component was the computational scalability provided by the Hadoop framework, which `HBase` utilises. The large-scale data processing engine Apache Spark

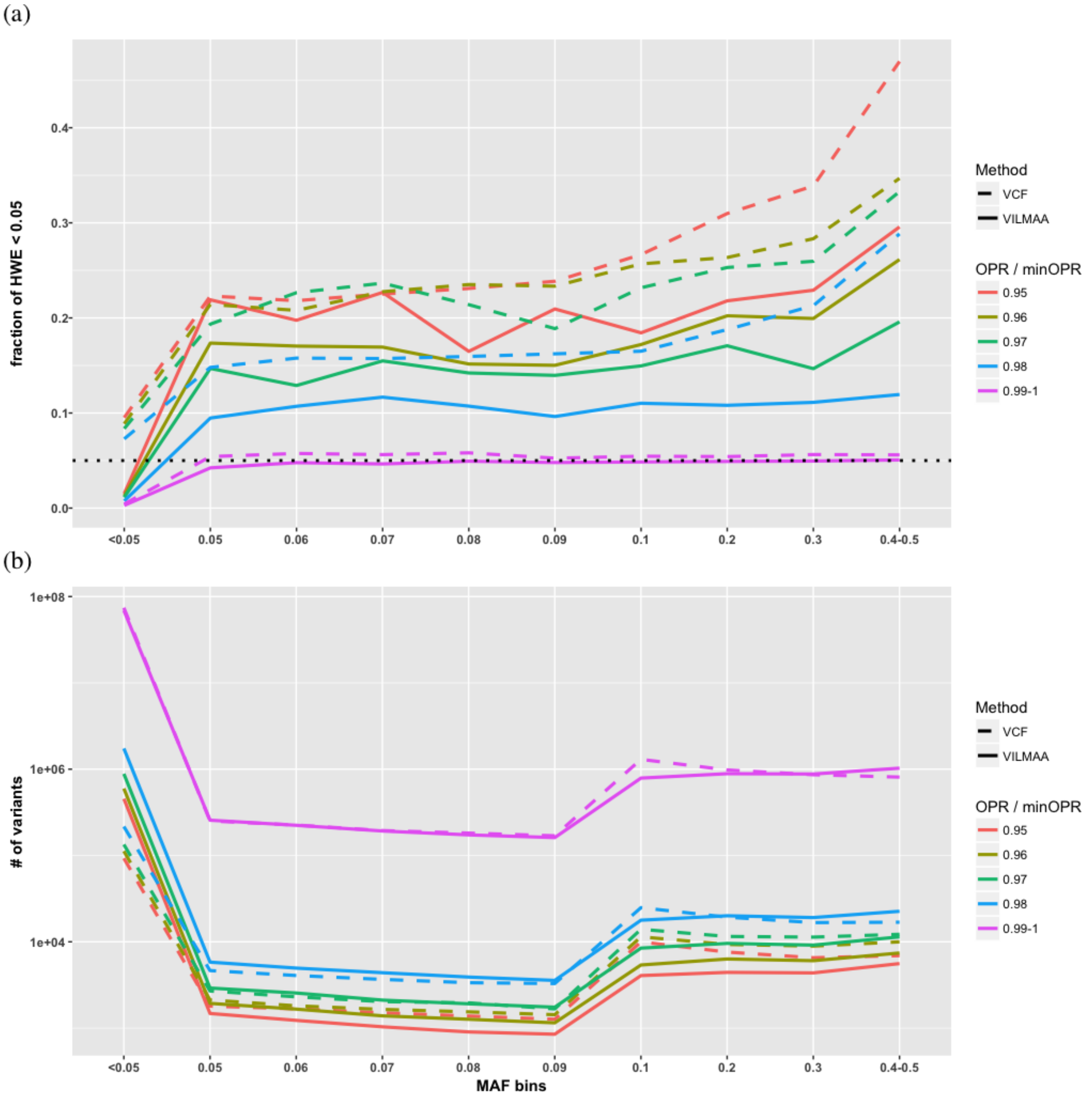


Figure 4.11 MAF affect on HWE in different OPR bins. OPR and minOPR value were used to create OPR bins for the VCF and VILMAA based method respectively. (a) Each line represents the fraction of HWE <0.05 for different OPR ranges and are calculated for MAF bins (WGS10K unrelated EUR). The black dotted line highlights the expected value of 0.05. VILMAA shows a more consistent pattern compared to the VCF based method and a lower OPR value displays less of an affect on the fraction of HWE <0.05 for VILMAA. (b) The number of SNVs used to calculate the fraction of HWE <0.05 is displayed for each OPR bin and method.

can also be deployed on the Hadoop framework and formed part of the technology used by SEQSpark (Zhang et al., 2017). Spark is a processing framework and stores information in structured files, which were not evaluated at the time.

4.4.2 Incremental variant aggregation

Variant information from samples arrived at different time intervals over the course of the project. During the course of the study, several months could pass between the arrival of sample data and the availability of these sample information as part of a release. The delay was a combination of waiting for sufficient samples to arrive to justify a new release and the release process itself. A new release re-merged samples and re-annotated variants, which were already processed before, for the previous release. Taken advantage of the scalability of `HBase`, I implemented an incremental aggregation and annotation approach that enabled a new release of 1K additional samples in a day. An incremental approach is different to current variant merging or aggregated variant call tools implemented by `agg` (Illumina) or `HaplotypeCaller` (GATK) respectively (Illumina, Inc., 2015; McKenna et al., 2010) and allows a fast release cycle. Integrated samples benefit from already annotated variants and the allele frequencies were recalculated as part of the release for different disease and ethnic cohorts. These information are essential to identify causal variants in a clinical setting and would move WGS a step closer from the bench to the bedside of patients.

Chapter 5

Concluding remarks

A disease is classified as rare, if it affects less than 5 in 10,000 people. An estimated 8,000 rare disorders exist in Europe[1], with some as rare as 1 in 100,000. Rare-disease diagnosis and research is being transformed by recent technical advances and availability of next generation sequencing (NGS). Identification, recruitment and analysis of sufficient cases with the same underlying cause remains a challenge.

5.0.1 Data sharing infrastructure

The NIHR BR-RD study sequenced the whole genome of 13,000 subjects harbouring billions of shared and rare variants. To analyse and identify rare disease causing variants, I developed a computational infrastructure to store, provide and analyse the billions of collated variants. The developed software was integrated with the OpenCGA framework to provides authenticated access through standardised application programming interfaces (APIs). These APIs enable the responsible and global sharing of genomic and phenotype information by implementing technical specifications developed by GA4GH working groups. The integration with GA4GH frameworks will allow to collate geographically distributed individual patients in one virtual place.

In addition to the genomic variants, we collected phenotype and clinical information from PAH patients in OpenClinica. Measurements in OpenClinica included numerical values in different units and are not easy to standardise and compare. The detailed collection would benefit from a translation to human phenotype ontology (HPO), which is a standardised ontology of phenotypic abnormalities. Such HPO terms are globally shared with genomic variants to describe patients and find similarities. Responsible sharing of a detailed description

[1] https://ec.europa.eu/health/non_communicable_diseases/rare_diseases_en

increases the chance in identifying possible other patients with the same disease and help diagnose undiagnosed cases.

5.0.2 NGS in clinical practice

The NIHR BR-RD PAH project aimed at the identification of novel disease-gene associations as well as the identification of patients with variants in known genes. My analysis found enrichment in 4 previously reported PAH genes, which questions the causality of the remaining known PAH gene. The american college of medical genetics and genomics (ACMG) published standards and guidelines for the clinical application of sequencing technologies. The assessment of variants following the ACMG guidelines provided strong evidence for causality in *BMPR2* but also concluded a “likely pathogenic” status for the known PAH gene *SMAD1* with no enrichment in the PAH cohort. This example highlights the need for cautious examination of evidence for clinical diagnosis. Clinical services and ultimately their patients will benefit from GA4GH frameworks to have access as well as contribute data globally. The framework ensures the confidentiality of patient information while joining and contributing to a rich pool of genotype and phenotype data. For time critical cases, the availability of the latest information to a clinician can be life saving. Specially critically ill children can benefit from rapid diagnosis and it is possible to go from a sample to a diagnosis within 4 days (Mestek-Boukhibar et al., 2018).

5.0.3 Personalised medicine

The wider integration of patients into the diagnosis process could be beneficial for patients and accelerate the understanding of rare diseases. Self assessment and the continuous monitoring of patients is possible by the use of smart and mobile devices. Remote monitoring of a patient could save resources by changing regular check-up visits to on-demand follow-ups. The continuous data flow would allow to establish personalised trends and could lead to sub-classification of diseases.

Medical records should be available as electronical health records and linked with genotype information. The electronic medical records and genomics (eMERGE) consortium focuses on the development of guidelines and methods to utilise electronic medical records (EMR) for large scale genomic research (Gottesman et al., 2013). At the same time, such medical records from individuals could provide genomically guided advice. Diagnoses or drugs could be suggested based on clinical data to assist point-of-care decisions and match drugs to a patient’s genome based on actionable pharmacogenomic variants (Collins, 2009). To support such decisions, the pharmacogenomic field requires a comprehensive knowledge basel, which is one of the goals of the clinical pharmacogenetics implementation consortium (CPIC) of the national institutes of health's (NIH) pharmacogenomics research network (PRN) (Relling and Klein, 2011). In my view, the challenge is translating the gain of information into the benefit of patients.

5.0.4 Validation of candidate genes

Responsibly sharing of patient information is possible through frameworks developed by the GA4GH consortium and enables to identify matching patients globally. However, the validation of candidate genes is a challenge for rare diseases due to the limited number of patients. Instead, functional confirmation of potentially disease-causing genes can be performed in cell lines or model organisms. Specific gene mutations could be studied for a molecular diagnosis of the disease. The rare diseases models and mechanisms network (RDMM) provides a platform to match novel rare-disease gene discoveries in patients with basic scientists for interrogation. Genes are matched to a list of registered model organism scientists and a financial incentive is provided to accelerate the assessment. I propose to open the network to the global research community to allow the suggestion of genes from specialised centres in other countries as well as for the distribution of the initial assessment. A coordinated approach would address the possible backlog of required gene assessments, collate negative results for genes and could encourage collaborations between research centres for the efficient use of resources.

References

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Abdalla, S. A., Gallione, C. J., Barst, R. J., Horn, E. M., Knowles, J. A., Marchuk, D. A., Letarte, M., and Morse, J. H. (2004). Primary pulmonary hypertension in families with hereditary haemorrhagic telangiectasia. *The European respiratory journal*, 23(3):373–377.

Abyzov, A. and Gerstein, M. (2011). AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics*, 27(5):595–603.

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., Dahl, F., Dermitzakis, E. T., Enver, T., Esteller, M., Estivill, X., Ferguson-Smith, A., Fitzgibbon, J., Flicek, P., Giehl, C., Graf, T., Grosveld, F., Guigó, R., Gut, I., Helin, K., Jarvius, J., Küppers, R., Lehrach, H., Lengauer, T., Lernmark, Å., Leslie, D., Loeffler, M., Macintyre, E., Mai, A., Martens, J. H. A., Minucci, S., Ouwehand, W. H., Pelicci, P. G., Pendeville, H., Porse, B., Rakyan, V., Reik, W., Schrappe, M., Schübeler, D., Seifert, M., Siebert, R., Simmons, D., Soranzo, N., Spicuglia, S., Stratton, M., Stunnenberg, H. G., Tanay, A., Torrents, D., Valencia, A., Vellenga, E., Vingron, M., Walter, J., and Willcocks, S. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology*, 30(3):224–226.

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*, Chapter 7:Unit7.20.

Afgan, E., Baker, D., Coraor, N., Goto, H., Paul, I. M., Makova, K. D., Nekrutenko, A., and Taylor, J. (2011). Harnessing cloud computing with Galaxy Cloud. *Nature biotechnology*, 29(11):972–974.

Aldred, M. A., Vijayakrishnan, J., James, V., Soubrier, F., Gomez-Sanchez, M. A., Martensson, G., Galie, N., Manes, A., Corris, P., Simonneau, G., Humbert, M., Morrell, N. W., and Trembath, R. C. (2006). BMPR2 gene rearrangements account for a significant proportion of mutations in familial and idiopathic pulmonary arterial hypertension. *Human Mutation*, 27(2):212–213.

Altoros Systems (2013). Evaluating NoSQL performance. https://www.slideshare.net/jaxLondonConference/evaluating-no-sql-performance-which-database-is-right-for-your-data-sergey-sverchkovaltoros. [Online; accessed 12-May-2014].

Amazon (2013). Aws storage gateway. [Online; accessed August 21, 2018].

Anthony, T. G., McDaniel, B. J., Byerley, R. L., McGrath, B. C., Cavener, D. R., McNurlan, M. A., and Wek, R. C. (2004). Preservation of liver protein synthesis during dietary leucine deprivation occurs at the expense of skeletal muscle mass in mice deleted for eIF2 kinase GCN2. *The Journal of biological chemistry*, 279(35):36553–36561.

Attisano, L., Cárcamo, J., Ventura, F., Weis, F. M., Massagué, J., and Wrana, J. L. (1993). Identification of human activin and TGF beta type I receptors that form heteromeric kinase complexes with type II receptors. *Cell*, 75(4):671–680.

Austin, E. D., Ma, L., LeDuc, C., Berman Rosenzweig, E., Borczuk, A., Phillips, J. A., Palomero, T., Sumazin, P., Kim, H. R., Talati, M. H., West, J., Loyd, J. E., and Chung, W. K. (2012). Whole exome sequencing to identify a novel gene (caveolin-1) associated with human pulmonary arterial hypertension. *Circulation. Cardiovascular genetics*, 5(3):336–343.

Avocado (2018). Avocado is a variant caller built on top of Apache Spark to allow rapid variant calling on cluster/cloud computing environments. http://bdg-avocado.readthedocs.io/en/latest/. [Online; accessed 18-Jul-2018].

Bahcall, O. G. (2016). Genetic variation: ExAC boosts clinical variant interpretation in rare diseases. *Nature reviews. Genetics*, 17(10):584.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome research*, 11(6):1005–1017.

Bamshad, M., Lin, R. C., Law, D. J., Watkins, W. C., Krakowiak, P. A., Moore, M. E., Franceschini, P., Lala, R., Holmes, L. B., Gebuhr, T. C., Bruneau, B. G., Schinzel, A., Seidman, J. G., Seidman, C. E., and Jorde, L. B. (1997). Mutations in human TBX3 alter limb, apocrine and genital development in ulnar-mammary syndrome. *Nature genetics*, 16(3):311–315.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics*, 12(11):745–755.

Barr, J. (2006). Amazon ec2 beta. [Online; accessed August 21, 2018].

Barst, R. J., McGoon, M., Torbicki, A., Sitbon, O., Krowka, M. J., Olschewski, H., and Gaine, S. (2004). Diagnosis and differential assessment of pulmonary arterial hypertension. *Journal of the American College of Cardiology*, 43(12 Suppl S):40S–47S.

Basson, C. T., Bachinsky, D. R., Lin, R. C., Levi, T., Elkins, J. A., Soults, J., Grayzel, D., Kroumpouzou, E., Traill, T. A., Leblanc-Straceski, J., Renault, B., Kucherlapati, R., Seidman, J. G., and Seidman, C. E. (1997). Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. *Nature genetics*, 15(1):30–35.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Benza, R. L., Miller, D. P., Barst, R. J., Badesch, D. B., Frost, A. E., and McGoon, M. D. (2012). An evaluation of long-term survival from time of diagnosis in pulmonary arterial hypertension from the REVEAL Registry. *Chest*, 142(2):448–456.

Best, D. H., Sumner, K. L., Austin, E. D., Chung, W. K., Brown, L. M., Borczuk, A. C., Rosenzweig, E. B., Bayrak-Toydemir, P., Mao, R., Cahill, B. C., Tazelaar, H. D., Leslie, K. O., Hemnes, A. R., Robbins, I. M., and Elliott, C. G. (2014). EIF2AK4 mutations in pulmonary capillary hemangiomatosis. *Chest*, 145(2):231–236.

Best, D. H., Vaughn, C., McDonald, J., Damjanovich, K., Runo, J. R., Chibuk, J. M., and Bayrak-Toydemir, P. (2011). Mosaic ACVRL1 and ENG mutations in hereditary haemorrhagic telangiectasia patients. *Journal of medical genetics*, 48(5):358–360.

Bleda, M., Tárraga, J., de Maria, A., Salavert, F., Garcia-Alonso, L., Celma, M., Martin, A., Dopazo, J., and Medina, I. (2012). CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. *Nucleic acids research*, 40(Web Server issue):W609–14.

Blekhman, R., Man, O., Herrmann, L., Boyko, A. R., Indap, A., Kosiol, C., Bustamante, C. D., Teshima, K. M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. *Current biology : CB*, 18(12):883–889.

Bongers, E. M. H. F., Duijf, P. H. G., van Beersum, S. E. M., Schoots, J., Van Kampen, A., Burckhardt, A., Hamel, B. C. J., Losan, F., Hoefsloot, L. H., Yntema, H. G., Knoers, N. V. A. M., and van Bokhoven, H. (2004). Mutations in the human TBX4 gene cause small patella syndrome. *American journal of human genetics*, 74(6):1239–1248.

Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics*, 84(2):210–223.

Budhiraja, R., Tuder, R. M., and Hassoun, P. M. (2004). Endothelial dysfunction in pulmonary hypertension. *Circulation*, 109(2):159–165.

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C., Teague, J. W., Menzies, A., Goodhead, I., Turner, D. J., Clee, C. M., Quail, M. A., Cox, A., Brown, C., Durbin, R., Hurles, M. E., Edwards, P. A. W., Bignell, G. R., Stratton, M. R., and Futreal, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*, 40(6):722–729.

Carss, K. J., Arno, G., Erwood, M., Stephens, J., Sanchis-Juan, A., Hull, S., Megy, K., Grozeva, D., Dewhurst, E., Malka, S., Plagnol, V., Penkett, C., Stirrups, K., Rizzo, R., Wright, G., Josifova, D., Bitner-Glindzicz, M., Scott, R. H., Clement, E., Allen, L., Armstrong, R., Brady, A. F., Carmichael, J., Chitre, M., Henderson, R. H. H., Hurst, J., MacLaren, R. E., Murphy, E., Paterson, J., Rosser, E., Thompson, D. A., Wakeling, E., Ouwehand, W. H., Michaelides, M., Moore, A. T., NIHR-BioResource Rare Diseases Consortium, Webster, A. R., and Raymond, F. L. (2017). Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. *American journal of human genetics*, 100(1):75–90.

Caulfield, T., Evans, J., McGuire, A., McCabe, C., Bubela, T., Cook-Deegan, R., Fishman, J., Hogarth, S., Miller, F. A., Ravitsky, V., Biesecker, B., Borry, P., Cho, M. K., Carroll, J. C., Etchegary, H., Joly, Y., Kato, K., Lee, S. S.-J., Rothenberg, K., Sankar, P., Szego, M. J., Ossorio, P., Pullman, D., Rousseau, F., Ungar, W. J., and Wilson, B. (2013). Reflections on the cost of "low-cost" whole genome sequencing: framing the health policy debate. *PLoS biology*, 11(11):e1001699.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., and Gruber, R. E. (2006). Bigtable: A distributed storage system for structured data. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation - Volume 7*, OSDI '06, pages 15–15, Berkeley, CA, USA. USENIX Association.

Chaouat, A., Coulet, F., Favre, C., Simonneau, G., Weitzenblum, E., Soubrier, F., and Humbert, M. (2004). Endoglin germline mutation in a patient with hereditary haemorrhagic telangiectasia and dexfenfluramine associated pulmonary arterial hypertension. *Thorax*, 59(5):446–448.

Charalampopoulos, A., Howard, L. S., Tzoulaki, I., Gin-Sing, W., Grapsa, J., Wilkins, M. R., Davies, R. J., Nihoyannopoulos, P., Connolly, S. B., and Gibbs, J. S. R. (2014). Response to pulmonary arterial hypertension drug therapies in patients with pulmonary arterial hypertension and cardiovascular risk factors. *Pulmonary Circulation*, 4(4):669–678.

Chaveroux, C., Lambert-Langlais, S., Parry, L., Carraro, V., Jousse, C., Maurin, A.-C., Bruhat, A., Marceau, G., Sapin, V., Averous, J., and Fafournoux, P. (2011). Identification of GCN2 as new redox regulator for oxidative stress prevention in vivo. *Biochemical and biophysical research communications*, 415(1):120–124.

Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., Zhou, H., Tian, L., Prakash, O., Lemire, M., Sleiman, P., Cheng, W.-Y., Chen, W., Shah, H., Shen, Y., Fromer, M., Omberg, L., Deardorff, M. A., Zackai, E., Bobe, J. R., Levin, E., Hudson, T. J., Groop, L., Wang, J., Hakonarson, H., Wojcicki, A., Diaz, G. A., Edelmann, L., Schadt, E. E., and Friend, S. H. (2016a). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature Biotechnology*, 34(5):531–538.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., and Saunders, C. T. (2016b). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222.

Chen, Y.-J., Yang, Q.-H., Liu, D., Liu, Q.-Q., Eyries, M., Wen, L., Wu, W.-H., Jiang, X., Yuan, P., Zhang, R., Soubrier, F., and Jing, Z.-C. (2013). Clinical and genetic characteristics of Chinese patients with hereditary haemorrhagic telangiectasia-associated pulmonary hypertension. *European journal of clinical investigation*, 43(10):1016–1024.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103.

Chida, A., Shintani, M., Yagi, H., Fujiwara, M., Kojima, Y., Sato, H., Imamura, S., Yokozawa, M., Onodera, N., Horigome, H., Kobayashi, T., Hatai, Y., Nakayama, T., Fukushima, H., Nishiyama, M., Doi, S., Ono, Y., Yasukouchi, S., Ichida, F., Fujimoto, K., Ohtsuki, S., Teshima, H., Kawano, T., Nomura, Y., Gu, H., Ishiwata, T., Furutani, Y., Inai, K., Saji, T., Matsuoka, R., Nonoyama, S., and Nakanishi, T. (2012). Outcomes of childhood pulmonary arterial hypertension in BMPR2 and ALK1 mutation carriers. *The American journal of cardiology*, 110(4):586–593.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E. E., Weinstock, G., Mardis, E. R., Wilson, R. K., Howe, K., Flicek, P., and Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS biology*, 9(7):e1001091.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews. Genetics*, 11(6):415–425.

Cogan, J. D., Pauciulo, M. W., Batchman, A. P., Prince, M. A., Robbins, I. M., Hedges, L. K., Stanton, K. C., Wheeler, L. A., Phillips, J. A., Loyd, J. E., and Nichols, W. C. (2006). High frequency of BMPR2 exonic deletions/duplications in familial pulmonary arterial hypertension. *American journal of respiratory and critical care medicine*, 174(5):590–598.

Cohen, A. W., Hnasko, R., Schubert, W., and Lisanti, M. P. (2004). Role of caveolae and caveolins in health and disease. *Physiological reviews*, 84(4):1341–1379.

Collins, F. (2009). Opportunities and challenges for the NIH–an interview with Francis Collins. Interview by Robert Steinbrook.

Conomos, M. P., Miller, M. B., and Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, 39(4):276–293.

Conomos, M. P., Reiner, A. P., Weir, B. S., and Thornton, T. A. (2016). Model-free Estimation of Recent Genetic Relatedness. *American journal of human genetics*, 98(1):127–148.

Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Cooper, B. F., Silberstein, A., Tam, E., Ramakrishnan, R., and Sears, R. (2010). Benchmarking cloud serving systems with ycsb. In *Proceedings of the 1st ACM Symposium on Cloud Computing*, SoCC '10, pages 143–154, New York, NY, USA. ACM.

Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, 15(7):901–913.

Cornish, A. and Guda, C. (2015). A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Research International*, 2015:456479.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M. J., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R., and Flicek, P. (2015). Ensembl 2015. *Nucleic acids research*, 43(Database issue):D662–9.

Czirják, G. and Enyedi, P. (2002). Formation of functional heterodimers between the TASK-1 and TASK-3 two-pore domain potassium channel subunits. *The Journal of biological chemistry*, 277(7):5426–5432.

D'Alonzo, G. E., Barst, R. J., Ayres, S. M., Bergofsky, E. H., Brundage, B. H., Detre, K. M., Fishman, A. P., Goldring, R. M., Groves, B. M., and Kernis, J. T. (1991). Survival in patients with primary pulmonary hypertension. Results from a national prospective registry. *Annals of internal medicine*, 115(5):343–349.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

David, L., Mallet, C., Mazerbourg, S., Feige, J.-J., and Bailly, S. (2007). Identification of BMP9 and BMP10 as functional activators of the orphan activin receptor-like kinase 1 (ALK1) in endothelial cells. *Blood*, 109(5):1953–1961.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, 6(12):e1001025.

Dean, J. and Ghemawat, S. (2004). Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, OSDI'04, pages 10–10, Berkeley, CA, USA. USENIX Association.

DECA (2018). DECA is a copy number variant caller built on top of Apache Spark to allow rapid variant calling on cluster/cloud computing environments. http://bdg-deca.readthedocs.io/en/latest/. [Online; accessed 18-Jul-2018].

Deng, Z., Morse, J. H., Slager, S. L., Cuervo, N., Moore, K. J., Venetos, G., Kalachikov, S., Cayanis, E., Fischer, S. G., Barst, R. J., Hodge, S. E., and Knowles, J. A. (2000). Familial Primary Pulmonary Hypertension (Gene PPH1) Is Caused by Mutations in the Bone Morphogenetic Protein Receptor–II Gene. *The American Journal of Human Genetics*, 67(3):737–744.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498.

Donnelly, N., Gorman, A. M., Gupta, S., and Samali, A. (2013). The eIF2α kinases: their structures and functions. *Cellular and molecular life sciences : CMLS*, 70(19):3493–3511.

Dove, E. S., Joly, Y., Tassé, A.-M., Public Population Project in Genomics and Society (P3G) International Steering Committee, International Cancer Genome Consortium (ICGC) Ethics and Policy Committee, and Knoppers, B. M. (2015). Genomic cloud computing: legal and ethical points to consider. *European journal of human genetics : EJHG*, 23(10):1271–1278.

Drake, K. M., Zygmunt, D., Mavrakis, L., Harbor, P., Wang, L., Comhair, S. A., Erzurum, S. C., and Aldred, M. A. (2011). Altered MicroRNA processing in heritable pulmonary arterial hypertension: an important role for Smad-8. *American journal of respiratory and critical care medicine*, 184(12):1400–1408.

Edwards, S. L., Beesley, J., French, J. D., and Dunning, A. M. (2013). Beyond GWASs: illuminating the dark road from association to function. *American journal of human genetics*, 93(5):779–797.

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K. E., Jonasdottir, A., Jonasdottir, A., Jonsdottir, I., Gudbjartsson, D. F., Melsted, P., Stefansson, K., and Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature genetics*, 49(11):1654–1660.

Elliott, C. G., Glissmeyer, E. W., Havlena, G. T., Carlquist, J., McKinney, J. T., Rich, S., McGoon, M. D., Scholand, M. B., Kim, M., Jensen, R. L., Schmidt, J. W., and Ward, K. (2006). Relationship of BMPR2 mutations to vasoreactivity in pulmonary arterial hypertension. *Circulation*, 113(21):2509–2515.

Engelman, J. A., Wykoff, C. C., Yasuhara, S., Song, K. S., Okamoto, T., and Lisanti, M. P. (1997). Recombinant expression of caveolin-1 in oncogenically transformed cells abrogates anchorage-independent growth. *The Journal of biological chemistry*, 272(26):16374–16381.

Engelman, J. A., Zhang, X. L., Galbiati, F., and Lisanti, M. P. (1998). Chromosomal localization, genomic organization, and developmental expression of the murine caveolin gene family (Cav-1, -2, and -3). Cav-1 and Cav-2 genes map to a known tumor suppressor locus (6-A2/7q31). *FEBS letters*, 429(3):330–336.

Ensembl (2015). Calculated variant consequences. [Online; accessed August 31, 2017].

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49.

Escribano-Subias, P., Blanco, I., López-Meseguer, M., Lopez-Guarch, C. J., Roman, A., Morales, P., Castillo-Palma, M. J., Segovia, J., Gómez-Sanchez, M. A., Barberà, J. A., and REHAP investigators (2012). Survival in pulmonary hypertension in Spain: insights from the Spanish registry. *The European respiratory journal*, 40(3):596–603.

Evans, J. D. W., Girerd, B., Montani, D., Wang, X.-J., Galie, N., Austin, E. D., Elliott, G., Asano, K., Grünig, E., Yan, Y., Jing, Z.-C., Manes, A., Palazzini, M., Wheeler, L. A., Nakayama, I., Satoh, T., Eichstaedt, C., Hinderhofer, K., Wolf, M., Rosenzweig, E. B., Chung, W. K., Soubrier, F., Simonneau, G., Sitbon, O., Gräf, S., Kaptoge, S., Di Angelantonio, E., Humbert, M., and Morrell, N. W. (2016). BMPR2 mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. *The Lancet. Respiratory medicine*, 4(2):129–137.

Eyries, M., Coulet, F., Girerd, B., Montani, D., Humbert, M., Lacombe, P., Chinet, T., Gouya, L., Roume, J., Axford, M. M., Pearson, C. E., and Soubrier, F. (2012). ACVRL1 germinal mosaic with two mutant alleles in hereditary hemorrhagic telangiectasia associated with pulmonary arterial hypertension. *Clinical genetics*, 82(2):173–179.

Eyries, M., Montani, D., Girerd, B., Perret, C., Leroy, A., Lonjou, C., Chelghoum, N., Coulet, F., Bonnet, D., Dorfmüller, P., Fadel, E., Sitbon, O., Simonneau, G., Trégouët, D.-A., Humbert, M., and Soubrier, F. (2014). EIF2AK4 mutations cause pulmonary veno-occlusive disease, a recessive form of pulmonary hypertension. *Nature Genetics*, 46(1):65–69.

Fessel, J. P., Flynn, C. R., Robinson, L. J., Penner, N. L., Gladson, S., Kang, C. J., Wasserman, D. H., Hemnes, A. R., and West, J. D. (2013). Hyperoxia synergizes with mutant bone morphogenic protein receptor 2 to cause metabolic stress, oxidant injury, and pulmonary hypertension. *American journal of respiratory cell and molecular biology*, 49(5):778–787.

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97.

Fletcher, J. (2015). Ethical approval for all studies involving human participants. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 187(2):91.

Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Koscielny, G., Kulesha, E., Lawson, D., Longden, I., Massingham, T., McLaren, W., Megy, K., Overduin, B., Pritchard, B., Rios, D., Ruffier, M., Schuster, M., Slater, G., Smedley, D., Spudich, G., Tang, Y. A., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S. P., Zadissa, A., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Smith, J., and Searle, S. M. J. (2010). Ensembl's 10th year. *Nucleic acids research*, 38(Database issue):D557–62.

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–251.

Frost, A. E., Badesch, D. B., Barst, R. J., Benza, R. L., Elliott, C. G., Farber, H. W., Krichman, A., Liou, T. G., Raskob, G. E., Wason, P., Feldkircher, K., Turner, M., and McGoon, M. D. (2011). The changing picture of patients with pulmonary arterial hypertension in the United States: how REVEAL differs from historic and non-US Contemporary Registries. *Chest*, 139(1):128–137.

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., NHLBI Exome Sequencing Project, and Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220.

Fujiwara, M., Yagi, H., Matsuoka, R., Akimoto, K., Furutani, M., Imamura, S.-i., Uehara, R., Nakayama, T., Takao, A., Nakazawa, M., and Saji, T. (2008). Implications of mutations of activin receptor-like kinase 1 gene (ALK1) in addition to bone morphogenetic protein receptor II gene (BMPR2) in children with pulmonary arterial hypertension. *Circulation journal : official journal of the Japanese Circulation Society*, 72(1):127–133.

Galie, N., Humbert, M., Vachiery, J.-L., Gibbs, S., Lang, I., Torbicki, A., Simonneau, G., Peacock, A., Vonk-Noordegraaf, A., Beghetti, M., Ghofrani, A., Gomez Sanchez, M. A., Hansmann, G., Klepetko, W., Lancellotti, P., Matucci, M., McDonagh, T., Pierard, L. A., Trindade, P. T., Zompatori, M., and Hoeper, M. (2015). 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT).

Galiè, N., Torbicki, A., Barst, R., Dartevelle, P., Haworth, S., Higenbottam, T., Olschewski, H., Peacock, A., Pietra, G., Rubin, L. J., Simonneau, G., Priori, S. G., Garcia, M. A. A., Blanc, J.-J., Budaj, A., Cowie, M., Dean, V., Deckers, J., Burgos, E. F., Lekakis, J., Lindahl, B., Mazzotta, G., McGregor, K., Morais, J., Oto, A., Smiseth, O. A., Barbera, J. A., Gibbs, S., Hoeper, M., Humbert, M., Naeije, R., Pepke-Zaba, J., and Task Force (2004). Guidelines on diagnosis and treatment of pulmonary arterial hypertension. The Task Force on Diagnosis and Treatment of Pulmonary Arterial Hypertension of the European Society of Cardiology.

Gambin, T., Jhangiani, S. N., Below, J. E., Campbell, I. M., Wiszniewski, W., Muzny, D. M., Staples, J., Morrison, A. C., Bainbridge, M. N., Penney, S., McGuire, A. L., Gibbs, R. A., Lupski, J. R., and Boerwinkle, E. (2015). Secondary findings and carrier test frequencies in a large multiethnic sample. *Genome medicine*, 7(1):54.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv e-prints*.

Genome Reference Consortium (2013). Announcing grch38. [Online; accessed Feb 15, 2018].

George, L. (2011). *HBase: The Definitive Guide*. O'Reilly Media, 1 edition.

Ghemawat, S., Gobioff, H., and Leung, S.-T. (2003). The google file system. *SIGOPS Oper. Syst. Rev.*, 37(5):29–43.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature reviews. Genetics*, 13(2):135–145.

Girerd, B., Montani, D., Coulet, F., Sztrymf, B., Yaici, A., Jaïs, X., Tregouet, D., Reis, A., Drouin-Garraud, V., Fraisse, A., Sitbon, O., O'Callaghan, D. S., Simonneau, G., Soubrier, F., and Humbert, M. (2010). Clinical outcomes of pulmonary arterial hypertension in patients carrying an ACVRL1 (ALK1) mutation. *American journal of respiratory and critical care medicine*, 181(8):851–861.

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., and Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC genomics*, 14 Suppl 3:S7.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1513–1518.

Gómez-Garre, P., Seijo, M., Gutiérrez-Delicado, E., Castro del Río, M., de la Torre, C., Gómez-Abad, C., Morales-Corraliza, J., Puig, M., and Serratosa, J. M. (2006). Ehlers-Danlos syndrome and periventricular nodular heterotopia in a Spanish family with a single FLNA mutation. *Journal of Medical Genetics*, 43(3):232–237.

Google (2008). Introducing google app engine. [Online; accessed August 21, 2018].

Google (2015). Data encryption options. [Online; accessed August 21, 2018].

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., Sanderson, S. C., Kannry, J., Zinberg, R., Basford, M. A., Brilliant, M., Carey, D. J., Chisholm, R. L., Chute, C. G., Connolly, J. J., Crosslin, D., Denny, J. C., Gallego, C. J., Haines, J. L., Hakonarson, H., Harley, J., Jarvik, G. P., Kohane, I., Kullo, I. J., Larson, E. B., McCarty, C., Ritchie, M. D., Roden, D. M., Smith, M. E., Böttinger, E. P., Williams, M. S., and eMERGE Network (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in medicine : official journal of the American College of Medical Genetics*, 15(10):761–771.

Gräf, S., Haimel, M., Bleda, M., Hadinnapola, C., Southgate, L., Li, W., Hodgson, J., Liu, B., Salmon, R. M., Southwood, M., Machado, R. D., Martin, J. M., Treacy, C. M., Yates, K., Daugherty, L. C., Shamardina, O., Whitehorn, D., Holden, S., Aldred, M., Bogaard, H. J., Church, C., Coghlan, G., Condliffe, R., Corris, P. A., Danesino, C., Eyries, M., Gall, H., Ghio, S., Ghofrani, H.-A., Gibbs, J. S. R., Girerd, B., Houweling, A. C., Howard, L., Humbert, M., Kiely, D. G., Kovacs, G., MacKenzie Ross, R. V., Moledina, S., Montani, D., Newnham, M., Olschewski, A., Olschewski, H., Peacock, A. J., Pepke-Zaba, J., Prokopenko, I., Rhodes, C. J., Scelsi, L., Seeger, W., Soubrier, F., Stein, D. F., Suntharalingam, J., Swietlik, E. M., Toshner, M. R., van Heel, D. A., Vonk Noordegraaf, A., Waisfisz, Q., Wharton, J., Wort, S. J., Ouwehand, W. H., Soranzo, N., Lawrie, A., Upton, P. D., Wilkins, M. R., Trembath, R. C., and Morrell, N. W. (2018). Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. *Nature communications*, 9(1):1416.

Grateau, G., Hentgen, V., Stojanovic, K. S., Jéru, I., Amselem, S., and Steichen, O. (2013). How should we approach classification of autoinflammatory diseases? *Nature reviews. Rheumatology*, 9(10):624–629.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585.

Gyles, C. (2008). The DNA revolution. *The Canadian veterinary journal = La revue veterinaire canadienne*, 49(8):745–746.

Hail (2018). An open-source, scalable framework for exploring and analyzing genomic data. https://github.com/hail-is/hail. [Online; accessed 18-Jul-2018].

Harrison, R. E., Flanagan, J. A., Sankelo, M., Abdalla, S. A., Rowell, J., Machado, R. D., Elliott, C. G., Robbins, I. M., Olschewski, H., McLaughlin, V., Gruenig, E., Kermeen, F., Halme, M., Räisänen-Sokolowski, A., Laitinen, T., Morrell, N. W., and Trembath, R. C. (2003). Molecular and functional analysis identifies ALK-1 as the predominant cause of pulmonary hypertension related to hereditary haemorrhagic telangiectasia. *Journal of Medical Genetics*, 40(12):865–871.

Hartness, M. E., Lewis, A., Searle, G. J., O'Kelly, I., Peers, C., and Kemp, P. J. (2001). Combined antisense and pharmacological approaches implicate hTASK as an airway O(2) sensing K(+) channel. *The Journal of biological chemistry*, 276(28):26499–26508.

Hayashi, K., Matsuda, S., Machida, K., Yamamoto, T., Fukuda, Y., Nimura, Y., Hayakawa, T., and Hamaguchi, M. (2001). Invasion activating caveolin-1 mutation in human scirrhous breast cancers. *Cancer research*, 61(6):2361–2364.

Health, N. (2014). Genomic data sharing. [Online; accessed August 21, 2016].

Heath, A. P., Greenway, M., Powell, R., Spring, J., Suarez, R., Hanley, D., Bandlamudi, C., McNerney, M. E., White, K. P., and Grossman, R. L. (2014). Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. *Journal of the American Medical Informatics Association : JAMIA*, 21(6):969–975.

Herrera, M., Hong, N. J., and Garvin, J. L. (2006). Aquaporin-1 transports NO across cell membranes. *Hypertension (Dallas, Tex. : 1979)*, 48(1):157–164.

Hicks, S., Wheeler, D. A., Plon, S. E., and Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Human mutation*, 32(6):661–668.

Hirashiki, A., Adachi, S., Nakano, Y., Kamimura, Y., Ogo, T., Nakanishi, N., Morisaki, T., Morisaki, H., Shimizu, A., Toba, K., Murohara, T., and Kondo, T. (2017). Left main coronary artery compression by a dilated main pulmonary artery and left coronary sinus of Valsalva aneurysm in a patient with heritable pulmonary arterial hypertension and FLNA mutation. *Pulmonary Circulation*, page 2045893217716107.

Hoeper, M. M., Bogaard, H. J., Condliffe, R., Frantz, R., Khanna, D., Kurzyna, M., Langleben, D., Manes, A., Satoh, T., Torres, F., Wilkins, M. R., and Badesch, D. B. (2013). Definitions and Diagnosis of Pulmonary Hypertension. *Journal of the American College of Cardiology*, 62(25_S) : $--$.

Holbrook, J. A., Neu-Yilik, G., Hentze, M. W., and Kulozik, A. E. (2004). Nonsense-mediated decay approaches the clinic. *Nature Genetics*, 36(8):801–808.

Hoogaars, W. M. H., Barnett, P., Moorman, A. F. M., and Christoffels, V. M. (2007). T-box factors determine cardiac design. *Cellular and molecular life sciences : CMLS*, 64(6):646–660.

Hsu, K., Lee, T.-Y., Periasamy, A., Kao, F.-J., Li, L.-T., Lin, C.-Y., Lin, H.-J., and Lin, M. (2017). Adaptable interaction between aquaporin-1 and band 3 reveals a potential role of water channel in blood CO2 transport. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. (2002). The Ensembl genome database project. *Nucleic acids research*, 30(1):38–41.

Humbert, M., Sitbon, O., Chaouat, A., Bertocchi, M., Habib, G., Gressin, V., Yaïci, A., Weitzenblum, E., Cordier, J.-F., Chabot, F., Dromer, C., Pison, C., Reynaud-Gaubert, M., Haloun, A., Laurent, M., Hachulla, E., Cottin, V., Degano, B., Jaïs, X., Montani, D., Souza, R., and Simonneau, G. (2010). Survival in patients with idiopathic, familial, and anorexigen-associated pulmonary arterial hypertension in the modern management era. *Circulation*, 122(2):156–163.

Humbert, M., Sitbon, O., Chaouat, A., Bertocchi, M., Habib, G., Gressin, V., Yaïci, A., Weitzenblum, E., Cordier, J.-F., Chabot, F., Dromer, C., Pison, C., Reynaud-Gaubert, M., Haloun, A., Laurent, M., Hachulla, E., and Simonneau, G. (2006). Pulmonary arterial hypertension in France: results from a national registry. *American journal of respiratory and critical care medicine*, 173(9):1023–1030.

Illumina, Inc. (2015). agg: a utility for aggregating Illumina-style GVCFs. https://github.com/Illumina/agg. [Online; accessed 15-Feb-2017].

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Bonnen, P. E., Gibbs, R. A., Gonzaga-Jauregui, C., Keinan, A., Price, A. L., Yu, F., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S. F., Zhang, Q., Ghori, M. J. R., McGinnis, R., McLaren, W., Pollack, S., Price, A. L., Schaffner, S. F., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., and McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58.

International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426(6968):789–796.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

International PPH Consortium, Lane, K. B., Machado, R. D., Pauciulo, M. W., Thomson, J. R., Phillips, J. A., Loyd, J. E., Nichols, W. C., and Trembath, R. C. (2000). Heterozygous germline mutations in BMPR2, encoding a TGF-beta receptor, cause familial primary pulmonary hypertension. *Nature Genetics*, 26(1):81–84.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232.

Ishiwata, T., Terada, J., Tanabe, N., Abe, M., Sugiura, T., Tsushima, K., Tada, Y., Sakao, S., Kasahara, Y., Nakanishi, N., Morisaki, H., and Tatsumi, K. (2014). Pulmonary arterial hypertension as the first manifestation in a patient with hereditary hemorrhagic telangiectasia. *Internal medicine (Tokyo, Japan)*, 53(20):2359–2363.

Johnson, D. W., Berg, J. N., Baldwin, M. A., Gallione, C. J., Marondel, I., Yoon, S. J., Stenzel, T. T., Speer, M., Pericak-Vance, M. A., Diamond, A., Guttmacher, A. E., Jackson, C. E., Attisano, L., Kucherlapati, R., Porteous, M. E., and Marchuk, D. A. (1996). Mutations in the activin receptor-like kinase 1 gene in hereditary haemorrhagic telangiectasia type 2. *Nature Genetics*, 13(2):189–195.

Jones, G., Robertson, L., Harrison, R., Ridout, C., and Vasudevan, P. (2014). Somatic mosaicism in ACVRL1 with transmission to several offspring affected with severe pulmonary arterial hypertension. *American journal of medical genetics. Part A*, 164A(8):2121–2123.

Kaye, J., Heeney, C., Hawkins, N., de Vries, J., and Boddington, P. (2009). Data sharing in genomics–re-shaping scientific practice. *Nature reviews. Genetics*, 10(5):331–335.

Kent, W. J. and Haussler, D. (2001). Assembly of the working draft of the human genome with GigAssembler. *Genome research*, 11(9):1541–1548.

Kerstjens-Frederikse, W. S., Bongers, E. M. H. F., Roofthooft, M. T. R., Leter, E. M., Douwes, J. M., Van Dijk, A., Vonk-Noordegraaf, A., Dijk-Bos, K. K., Hoefsloot, L. H., Hoendermis, E. S., Gille, J. J. P., Sikkema-Raddatz, B., Hofstra, R. M. W., and Berger, R. M. F. (2013). TBX4 mutations (small patella syndrome) are associated with childhood-onset pulmonary arterial hypertension. *Journal of Medical Genetics*, 50(8):500–506.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315.

Kirk, E. P., Sunde, M., Costa, M. W., Rankin, S. A., Wolstein, O., Castro, M. L., Butler, T. L., Hyun, C., Guo, G., Otway, R., Mackay, J. P., Waddell, L. B., Cole, A. D., Hayward, C., Keogh, A., Macdonald, P., Griffiths, L., Fatkin, D., Sholler, G. F., Zorn, A. M., Feneley, M. P., Winlaw, D. S., and Harvey, R. P. (2007). Mutations in cardiac T-box factor gene TBX20 are associated with diverse cardiac pathologies, including defects of septation and valvulogenesis and cardiomyopathy. *American journal of human genetics*, 81(2):280–291.

Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. *The HUGO journal*, 8(1):3.

Knoppers, B. M., Harris, J. R., Budin-Ljøsne, I., and Dove, E. S. (2014). A human rights approach to an international code of conduct for genomic and clinical data sharing. *Human genetics*, 133(7):895–903.

Knoppers, B. M., Harris, J. R., Tassé, A. M., Budin-Ljøsne, I., Kaye, J., Deschênes, M., and Zawati, M. H. (2011). Towards a data sharing Code of Conduct for international genomic research. *Genome medicine*, 3(7):46.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38.

Koehler, R., Grünig, E., Pauciulo, M. W., Hoeper, M. M., Olschewski, H., Wilkens, H., Halank, M., Winkler, J., Ewert, R., Bremer, H., Kreuscher, S., Janssen, B., and Nichols, W. C. (2004). Low frequency of BMPR2 mutations in a German cohort of patients with sporadic idiopathic pulmonary arterial hypertension. *Journal of Medical Genetics*, 41(12):e127.

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods*, 6(4):291–295.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., and Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–8.

Lange, A. W., Haitchi, H. M., LeCras, T. D., Sridharan, A., Xu, Y., Wert, S. E., James, J., Udell, N., Thurner, P. J., and Whitsett, J. A. (2014). Sox17 is required for normal pulmonary vascular morphogenesis. *Developmental biology*, 387(1):109–120.

Langmead, B., Schatz, M. C., Lin, J., Pop, M., and Salzberg, S. L. (2009a). Searching for SNPs with cloud computing. *Genome biology*, 10(11):R134.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009b). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.

Larkin, E. K., Newman, J. H., Austin, E. D., Hemnes, A. R., Wheeler, L., Robbins, I. M., West, J. D., Phillips, III, J. A., Hamid, R., and Loyd, J. E. (2012). Longitudinal Analysis Casts Doubt on the Presence of Genetic Anticipation in Heritable Pulmonary Arterial Hypertension. *American journal of respiratory and critical care medicine*, 186(9):892–896.

Ledergerber, C. and Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. *Briefings in bioinformatics*, 12(5):489–497.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics*, 95(1):5–23.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G., and Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291.

Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A., and Gilissen, C. (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human Mutation*, 36(8):815–822.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., Macdonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y.-H., Frazier, M. E., Scherer, S. W., Strausberg, R. L., and Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS biology*, 5(10):e254.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20(2):265–272.

Lopez, J., Coll, J., Haimel, M., Kandasamy, S., Tarraga, J., Furio-Tari, P., Bari, W., Bleda, M., Rueda, A., Gräf, S., Rendon, A., Dopazo, J., and Medina, I. (2017). HGVA: the Human Genome Variation Archive. *Nucleic acids research*, 45(W1):W189–W194.

Ma, L. and Chung, W. K. (2017). The role of genetics in pulmonary arterial hypertension. *The Journal of pathology*, 241(2):273–280.

Ma, L., Roman-Campos, D., Austin, E. D., Eyries, M., Sampson, K. S., Soubrier, F., Germain, M., Trégouët, D.-A., Borczuk, A., Rosenzweig, E. B., Girerd, B., Montani, D., Humbert, M., Loyd, J. E., Kass, R. S., and Chung, W. K. (2013). A novel channelopathy in pulmonary arterial hypertension. *The New England journal of medicine*, 369(4):351–361.

Machado, R. D., Eickelberg, O., Elliott, C. G., Geraci, M. W., Hanaoka, M., Loyd, J. E., Newman, J. H., Phillips, III, J. A., Soubrier, F., Trembath, R. C., and Chung, W. K. (2009). Genetics and Genomics of Pulmonary Arterial Hypertension. *Journal of the American College of Cardiology*, 54(1):S32–S42.

Machado, R. D., Southgate, L., Eichstaedt, C. A., Aldred, M. A., Austin, E. D., Best, D. H., Chung, W. K., Benjamin, N., Elliott, C. G., Eyries, M., Fischer, C., Gräf, S., Hinderhofer, K., Humbert, M., Keiles, S. B., Loyd, J. E., Morrell, N. W., Newman, J. H., Soubrier, F., Trembath, R. C., Viales, R. R., and Grünig, E. (2015). Pulmonary Arterial Hypertension: A Current Perspective on Established and Emerging Molecular Genetic Defects. *Human Mutation*, 36(12):1113–1127.

Mache, C. J., Gamillscheg, A., Popper, H. H., and Haworth, S. G. (2008). Early-life pulmonary arterial hypertension with subsequent development of diffuse pulmonary arteriovenous malformations in hereditary haemorrhagic telangiectasia type 1. *Thorax*, 63(1):85–86.

Madan, M., Patel, A., Skruber, K., Geerts, D., Altomare, D. A., and Iv, O. P. (2016). ATP13A3 and caveolin-1 as potential biomarkers for difluoromethylornithine-based therapies in pancreatic cancers. *American journal of cancer research*, 6(6):1231–1252.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Margulies, E. H., Blanchette, M., NISC Comparative Sequencing Program, Haussler, D., and Green, E. D. (2003). Identification and characterization of multi-species conserved sequences. *Genome research*, 13(12):2507–2518.

Mariño-Enríquez, A., Lapunzina, P., Robertson, S. P., and Rodríguez, J. I. (2007). Otopalatodigital syndrome type 2 in two siblings with a novel filamin A 629G>T mutation: clinical, pathological, and molecular findings. *American journal of medical genetics. Part A*, 143A(10):1120–1125.

Martin, S. (2014). Microsoft azure. [Online; accessed August 21, 2018].

Massagué, J. (1998). TGF-beta signal transduction. *Annual review of biochemistry*, 67:753–791.

Massie, M., Nothaft, F., Hartl, C., Kozanitis, C., Schumacher, A., Joseph, A. D., and Patterson, D. A. (2013). ADAM: Genomics formats and processing patterns for cloud scale computing. *... Berkeley*.

Mathew, R., Huang, J., Shah, M., Patel, K., Gewitz, M., and Sehgal, P. B. (2004). Disruption of endothelial-cell caveolin-1alpha/raft scaffolding during development of monocrotaline-induced pulmonary hypertension. *Circulation*, 110(11):1499–1506.

Matsui, T., Kanai-Azuma, M., Hara, K., Matoba, S., Hiramatsu, R., Kawakami, H., Kurohmaru, M., Koopman, P., and Kanai, Y. (2006). Redundant roles of Sox17 and Sox18 in postnatal angiogenesis in mice. *Journal of cell science*, 119(Pt 17):3513–3526.

McAllister, K. A., Grogg, K. M., Johnson, D. W., Gallione, C. J., Baldwin, M. A., Jackson, C. E., Helmbold, E. A., Markel, D. S., McKinnon, W. C., and Murrell, J. (1994). Endoglin, a TGF-beta binding protein of endothelial cells, is the gene for hereditary haemorrhagic telangiectasia type 1. *Nature genetics*, 8(4):345–351.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De La Vega, F. M., and Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research*, 19(9):1527–1541.

McKusick, V. A. (2007). Mendelian Inheritance in Man and its online version, OMIM. *American journal of human genetics*, 80(4):588–604.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology*, 17(1):122.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–2070.

Mefford, H. C. and Eichler, E. E. (2009). Duplication hotspots, rare genomic disorders, and common disease. *Current opinion in genetics & development*, 19(3):196–204.

Mello, M. M., Francer, J. K., Wilenzick, M., Teden, P., Bierer, B. E., and Barnes, M. (2013). Preparing for responsible sharing of clinical trial data. *The New England journal of medicine*, 369(17):1651–1658.

Mestek-Boukhibar, L., Clement, E., Jones, W. D., Drury, S., Ocaka, L., Gagunashvili, A., Le Quesne Stabej, P., Bacchelli, C., Jani, N., Rahman, S., Jenkins, L., Hurst, J. A., Bitner-Glindzicz, M., Peters, M., Beales, P. L., and Williams, H. J. (2018). Rapid Paediatric Sequencing (RaPS): comprehensive real-life workflow for rapid diagnosis of critically ill children. *Journal of medical genetics*.

Miosge, L. A., Field, M. A., Sontani, Y., Cho, V., Johnson, S., Palkova, A., Balakishnan, B., Liang, R., Zhang, Y., Lyon, S., Beutler, B., Whittle, B., Bertram, E. M., Enders, A., Goodnow, C. C., and Andrews, T. D. (2015). Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 112(37):E5189–98.

Montani, D., Achouh, L., Dorfmüller, P., Le Pavec, J., Sztrymf, B., Tchérakian, C., Rabiller, A., Haque, R., Sitbon, O., Jaïs, X., Dartevelle, P., Maître, S., Capron, F., Musset, D., Simonneau, G., and Humbert, M. (2008). Pulmonary veno-occlusive disease: clinical, functional, radiologic, and hemodynamic characteristics and outcome of 24 cases confirmed by histology. *Medicine*, 87(4):220–233.

Morse, J. H., Jones, A. C., Barst, R. J., Hodge, S. E., Wilhelmsen, K. C., and Nygaard, T. G. (1997). Mapping of familial primary pulmonary hypertension locus (PPH1) to chromosome 2q31-q32. *Circulation*, 95(12):2603–2606.

Naiche, L. A. and Papaioannou, V. E. (2003). Loss of Tbx4 blocks hindlimb development and affects vascularization and fusion of the allantois. *Development (Cambridge, England)*, 130(12):2681–2693.

Nasim, M. T., Ogo, T., Ahmed, M., Randall, R., Chowdhury, H. M., Snape, K. M., Bradshaw, T. Y., Southgate, L., Lee, G. J., Jackson, I., Lord, G. M., Gibbs, J. S. R., Wilkins, M. R., Ohta-Ogo, K., Nakamura, K., Girerd, B., Coulet, F., Soubrier, F., Humbert, M., Morrell, N. W., Trembath, R. C., and Machado, R. D. (2011). Molecular genetic characterization of SMAD signaling molecules in pulmonary arterial hypertension. *Human Mutation*, 32(12):1385–1389.

Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814.

Nichols, W. C., Koller, D. L., Slovis, B., Foroud, T., Terry, V. H., Arnold, N. D., Siemieniak, D. R., Wheeler, L., Phillips, J. A., Newman, J. H., Conneally, P. M., Ginsburg, D., and Loyd, J. E. (1997). Localization of the gene for familial primary pulmonary hypertension to chromosome 2q31-32. *Nature Genetics*, 15(3):277–280.

Nishihara, A., Watabe, T., Imamura, T., and Miyazono, K. (2002). Functional heterogeneity of bone morphogenetic protein receptor-II mutants found in patients with primary pulmonary hypertension. *Molecular biology of the cell*, 13(9):3055–3063.

O'Connor, B. D., Merriman, B., and Nelson, S. F. (2010). SeqWare Query Engine: storing and searching sequence data in the cloud. *BMC bioinformatics*, 11 Suppl 12:S2.

Olschewski, A., Li, Y., Tang, B., Hanze, J., Eul, B., Bohle, R. M., Wilhelm, J., Morty, R. E., Brau, M. E., Weir, E. K., Kwapiszewska, G., Klepetko, W., Seeger, W., and Olschewski, H. (2006). Impact of TASK-1 in human pulmonary artery smooth muscle cells. *Circulation research*, 98(8):1072–1080.

Pääbo, S. (2003). The mosaic that is our genome. *Nature*, 421(6921):409–412.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., and Trajanoski, Z. (2013). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*.

Packham, E. A. and Brook, J. D. (2003). T-box genes in human disorders. *Human molecular genetics*, 12 Spec No 1:R37–44.

Paila, U., Chapman, B. A., Kirchner, R., and Quinlan, A. R. (2013). GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS computational biology*, 9(7):e1003153.

Pandey, R. V. and Schlötterer, C. (2013). DistMap: a toolkit for distributed short read mapping on a Hadoop cluster. *PloS one*, 8(8):e72614.

Patel, H. H., Zhang, S., Murray, F., Suda, R. Y. S., Head, B. P., Yokoyama, U., Swaney, J. S., Niesman, I. R., Schermuly, R. T., Pullamsetti, S. S., Thistlethwaite, P. A., Miyanohara, A., Farquhar, M. G., Yuan, J. X.-J., and Insel, P. A. (2007). Increased smooth muscle cell expression of caveolin-1 and caveolae contribute to the pathophysiology of idiopathic pulmonary arterial hypertension. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 21(11):2970–2979.

Pathak, S., McDermott, M. F., and Savic, S. (2017). Autoinflammatory diseases: update on classification diagnosis and management. *Journal of clinical pathology*, 70(1):1–8.

Peacock, A. J., Murphy, N. F., McMurray, J. J. V., Caballero, L., and Stewart, S. (2007). An epidemiological study of pulmonary arterial hypertension. *The European respiratory journal*, 30(1):104–109.

Pfarr, N., Fischer, C., Ehlken, N., Becker-Grünig, T., López-González, V., Gorenflo, M., Hager, A., Hinderhofer, K., Miera, O., Nagel, C., Schranz, D., and Grünig, E. (2013). Hemodynamic and genetic analysis in children with idiopathic, heritable, and congenital heart disease associated pulmonary arterial hypertension. *Respiratory research*, 14:3.

Philippakis, A. A., Azzariti, D. R., Beltran, S., Brookes, A. J., Brownstein, C. A., Brudno, M., Brunner, H. G., Buske, O. J., Carey, K., Doll, C., Dumitriu, S., Dyke, S. O. M., den Dunnen, J. T., Firth, H. V., Gibbs, R. A., Girdea, M., Gonzalez, M., Haendel, M. A., Hamosh, A., Holm, I. A., Huang, L., Hurles, M. E., Hutton, B., Krier, J. B., Misyura, A., Mungall, C. J., Paschall, J., Paten, B., Robinson, P. N., Schiettecatte, F., Sobreira, N. L., Swaminathan, G. J., Taschner, P. E., Terry, S. F., Washington, N. L., Züchner, S., Boycott, K. M., and Rehm, H. L. (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. *Human mutation*, 36(10):915–921.

Picard (2017). Java command line tools for manipulating high-throughput sequencing (HTS) data and formats. http://broadinstitute.github.io/picard/. [Online; accessed 21-Feb-2017].

Pireddu, L., Leo, S., and Zanetti, G. (2011). SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics*, 27(15):2159–2160.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–575.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raczy, C., Petrovski, R., Saunders, C. T., Chorny, I., Kruglyak, S., Margulies, E. H., Chuang, H.-Y., Källberg, M., Kumar, S. A., Liao, A., Little, K. M., Strömberg, M. P., and Tanner, S. W. (2013). Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*, 29(16):2041–2043.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339.

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L., Plon, S. E., Ramos, E. M., Sherry, S. T., Watson, M. S., and ClinGen (2015). ClinGen–the Clinical Genome Resource. *The New England journal of medicine*, 372(23):2235–2242.

Reid, J. G., Carroll, A., Veeraraghavan, N., Dahdouli, M., Sundquist, A., English, A., Bainbridge, M., White, S., Salerno, W., Buhay, C., Yu, F., Muzny, D., Daly, R., Duyk, G., Gibbs, R. A., and Boerwinkle, E. (2014). Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC bioinformatics*, 15:30.

Relling, M. V. and Klein, T. E. (2011). CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clinical pharmacology and therapeutics*, 89(3):464–467.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., and ACMG Laboratory Quality Assurance Committee (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. In *Genetics in medicine : official journal of the American College of Medical Genetics*, pages 405–424. Department of Molecular and Medical Genetics, Knight Diagnostic Laboratories, Oregon Health & Science University, Portland, Oregon, USA.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., WGS500 Consortium, Wilkie, A. O. M., McVean, G., and Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8):912–918.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011a). Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, 83(5):610–615.

Robinson, T., Killcoyne, S., Bressler, R., and Boyle, J. (2011b). SAMQA: error classification and validation of high-throughput sequenced read data. *BMC Genomics*, 12:419.

Roller, E., Ivakhno, S., Lee, S., Royce, T., and Tanner, S. (2016). Canvas: versatile and scalable detection of copy number variants. *Bioinformatics*, 32(15):2375–2377.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome biology*, 14(5):R51.

Runo, J. R., Vnencak-Jones, C. L., Prince, M., Loyd, J. E., Wheeler, L., Robbins, I. M., Lane, K. B., Newman, J. H., Johnson, J., Nichols, W. C., and Phillips, J. A. (2003). Pulmonary veno-occlusive disease caused by an inherited mutation in bone morphogenetic protein receptor II. *American journal of respiratory and critical care medicine*, 167(6):889–894.

Sakiyama, J.-I., Yamagishi, A., and Kuroiwa, A. (2003). Tbx4-Fgf10 system controls lung bud formation during chicken embryonic development. *Development (Cambridge, England)*, 130(7):1225–1234.

Samtools organisation (2017). The Variant Call Format (VCF) Version 4.2 Specification. http://samtools.github.io/hts-specs/VCFv4.2.pdf. [Online; accessed 17-Aug-2017].

Scharpfenecker, M., van Dinther, M., Liu, Z., van Bezooijen, R. L., Zhao, Q., Pukac, L., Löwik, C. W. G. M., and ten Dijke, P. (2007). BMP-9 signals via ALK1 and inhibits bFGF-induced endothelial cell proliferation and VEGF-stimulated angiogenesis. *Journal of cell science*, 120(Pt 6):964–972.

Schultheis, P. J., Hagen, T. T., O'Toole, K. K., Tachibana, A., Burke, C. R., McGill, D. L., Okunade, G. W., and Shull, G. E. (2004). Characterization of the P5 subfamily of P-type transport ATPases in mice. *Biochemical and biophysical research communications*, 323(3):731–738.

Schumacher, A., Pireddu, L., Niemenmaa, M., Kallio, A., Korpelainen, E., Zanetti, G., and Heljanko, K. (2014). SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop. *Bioinformatics*, 30(1):119–120.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.)*, 305(5683):525–528.

Shintani, M., Yagi, H., Nakayama, T., Saji, T., and Matsuoka, R. (2009). A new nonsense mutation of SMAD8 associated with pulmonary arterial hypertension. *Journal of Medical Genetics*, 46(5):331–337.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050.

Simonneau, G., Gatzoulis, M. A., Adatia, I., Celermajer, D., Denton, C., Ghofrani, A., Gomez Sanchez, M. A., Krishna Kumar, R., Landzberg, M., Machado, R. F., Olschewski, H., Robbins, I. M., and Souza, R. (2013). Updated clinical classification of pulmonary hypertension. *Journal of the American College of Cardiology*, 62(25 Suppl):D34–41.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123.

Smoot, L. B., Obler, D., McElhinney, D. B., Boardman, K., Wu, B.-L., Lip, V., and Mullen, M. P. (2009). Clinical features of pulmonary arterial hypertension in young people with an ALK1 mutation and hereditary haemorrhagic telangiectasia. *Archives of disease in childhood*, 94(7):506–511.

solid IT gmbh (2013). DB-Engines Ranking. https://db-engines.com/en/ranking. [Online; accessed 12-May-2014].

Song, W., Gardner, S. A., Hovhannisyan, H., Natalizio, A., Weymouth, K. S., Chen, W., Thibodeau, I., Bogdanova, E., Letovsky, S., Willis, A., and Nagan, N. (2016). Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. *Genetics in medicine : official journal of the American College of Medical Genetics*, 18(8):850–854.

Staples, J., Qiao, D., Cho, M. H., Silverman, E. K., University of Washington Center for Mendelian Genomics, Nickerson, D. A., and Below, J. E. (2014). PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *American journal of human genetics*, 95(5):553–564.

Stennard, F. A. and Harvey, R. P. (2005). T-box transcription factors and their roles in regulatory hierarchies in the developing heart. *Development (Cambridge, England)*, 132(22):4897–4910.

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., Abeysinghe, S., Krawczak, M., and Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Human Mutation*, 21(6):577–581.

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., and Cooper, D. N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, 133(1):1–9.

Stossel, T. P., Condeelis, J., Cooley, L., Hartwig, J. H., Noegel, A., Schleicher, M., and Shapiro, S. S. (2001). Filamins as integrators of cell mechanics and signalling. *Nature reviews. Molecular cell biology*, 2(2):138–145.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Mu, X. J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., and Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.

Sztrymf, B., Coulet, F., Girerd, B., Yaici, A., Jais, X., Sitbon, O., Montani, D., Souza, R., Simonneau, G., Soubrier, F., and Humbert, M. (2008). Clinical outcomes of pulmonary arterial hypertension in carriers of BMPR2 mutation. *American journal of respiratory and critical care medicine*, 177(12):1377–1383.

Tan, A., Abecasis, G. R., and Kang, H. M. (2015). Unified representation of genetic variants. *Bioinformatics*, 31(13):2202–2204.

Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., and Biesecker, L. G. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in medicine : official journal of the American College of Medical Genetics*.

Thomson, J. R., Machado, R. D., Pauciulo, M. W., Morgan, N. V., Humbert, M., Elliott, G. C., Ward, K., Yacoub, M., Mikhail, G., Rogers, P., Newman, J., Wheeler, L., Higenbottam, T., Gibbs, J. S., Egan, J., Crozier, A., Peacock, A., Allcock, R., Corris, P., Loyd, J. E., Trembath, R. C., and Nichols, W. C. (2000). Sporadic primary pulmonary hypertension is associated with germline mutations of the gene encoding BMPR-II, a receptor member of the TGF-beta family. *Journal of Medical Genetics*, 37(10):741–745.

Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human mutation*, 32(4):358–368.

Trembath, R. C., Thomson, J. R., Machado, R. D., Morgan, N. V., Atkinson, C., Winship, I., Simonneau, G., Galie, N., Loyd, J. E., Humbert, M., Nichols, W. C., Berg, J., Manes, A., McGaughran, J., Pauciulo, M., Wheeler, L., and Morrell, N. W. (2001). Clinical and Molecular Genetic Features of Pulmonary Hypertension in Patients with Hereditary Hemorrhagic Telangiectasia. *The New England journal of medicine*, 345(5):325–334.

Turner, T. N., Hormozdiari, F., Duyzend, M. H., McClymont, S. A., Hook, P. W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H. A., Zody, M. C., Nelson, B. J., Huddleston, J., Sandstrom, R., Smith, J. D., Hanna, D., Swanson, J. M., Faustman, E. M., Bamshad, M. J., Stamatoyannopoulos, J., Nickerson, D. A., McCallion, A. S., Darnell, R., and Eichler, E. E. (2016). Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *American journal of human genetics*, 98(1):58–74.

UK10K Consortium, Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., Humphries, S. E., Hurles, M. E., Zeggini, E., Barrett, J. C., Plagnol, V., Richards, J. B., Greenwood, C. M. T., Timpson, N. J., Durbin, R., and Soranzo, N. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics*, 43:11.10.1–33.

van der Flier, A. and Sonnenberg, A. (2001). Structural and functional aspects of filamins. *Biochimica et biophysica acta*, 1538(2-3):99–117.

Walsh, R., Thomson, K. L., Ware, J. S., Funke, B. H., Woodley, J., McGuire, K. J., Mazzarotto, F., Blair, E., Seller, A., Taylor, J. C., Minikel, E. V., Exome Aggregation Consortium, MacArthur, D. G., Farrall, M., Cook, S. A., and Watkins, H. (2017). Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genetics in medicine : official journal of the American College of Medical Genetics*, 19(2):192–203.

Wang, G., Knight, L., Ji, R., Lawrence, P., Kanaan, U., Li, L., Das, A., Cui, B., Zou, W., Penny, D. J., and Fan, Y. (2014). Early onset severe pulmonary arterial hypertension with 'two-hit' digenic mutations in both BMPR2 and KCNA5 genes. *International journal of cardiology*, 177(3):e167–9.

Wang, J., Mullighan, C. G., Easton, J., Roberts, S., Heatley, S. L., Ma, J., Rusch, M. C., Chen, K., Harris, C. C., Ding, L., Holmfeldt, L., Payne-Turner, D., Fan, X., Wei, L., Zhao, D., Obenauer, J. C., Naeve, C., Mardis, E. R., Wilson, R. K., Downing, J. R., and Zhang, J. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature methods*, 8(8):652–654.

Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., and Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics*, 31(3):318–323.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G. K.-S., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H., and Wang, J. (2008). The diploid genome sequence of an Asian individual. *Nature*, 456(7218):60–65.

Wang, M. and Wei, L. (2016). iFish: predicting the pathogenicity of human nonsynonymous variants using gene-specific/family-specific attributes and classifiers. *Scientific reports*, 6:31321.

Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic acids research*, 39(19):e132.

Westbury, S. K., Turro, E., Greene, D., Lentaigne, C., Kelly, A. M., Bariana, T. K., Simeoni, I., Pillois, X., Attwood, A., Austin, S., Jansen, S. B., Bakchoul, T., Crisp-Hihn, A., Erber, W. N., Favier, R., Foad, N., Gattens, M., Jolley, J. D., Liesner, R., Meacham, S., Millar,

C. M., Nurden, A. T., Peerlinck, K., Perry, D. J., Poudel, P., Schulman, S., Schulze, H., Stephens, J. C., Furie, B., Robinson, P. N., van Geet, C., Rendon, A., Gomez, K., Laffan, M. A., Lambert, M. P., Nurden, P., Ouwehand, W. H., Richardson, S., Mumford, A. D., Freson, K., and BRIDGE-BPD Consortium (2015). Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome medicine*, 7(1):36.

Wetterstrand, KA (2015). Dna sequencing costs: Data from the nhgri genome sequencing program (gsp). [Online; accessed August 31, 2017].

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A., and Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876.

White, T. (2009). *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 1st edition.

Wierda, E. and Hoftijzer, H. (2016). Right heart catheterization. https://www.pcipedia.org/images/thumb/6/64/RightHeart_Waveforms_Fig1.svg/800px-RightHeart_Waveforms_Fig1.svg.png. [Online; accessed 16-July-2018].

Wikiversity (2014). Blausen gallery 2014 — wikiversity,. https://en.wikiversity.org/w/index.php?title=Blausen_gallery_2014&oldid=1277099#/media/File:Blausen_0055_ArteryWallStructure.png. [Online; accessed 5-January-2015].

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., Stenson, P. D., Cooper, D. N., Tyler-Smith, C., and 1000 Genomes Project Consortium (2012). Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *American journal of human genetics*, 91(6):1022–1032.

Yagi, H., Furutani, Y., Hamada, H., Sasaki, T., Asakawa, S., Minoshima, S., Ichida, F., Joo, K., Kimura, M., Imamura, S.-i., Kamatani, N., Momma, K., Takao, A., Nakazawa, M., Shimizu, N., and Matsuoka, R. (2003). Role of TBX1 in human del22q11.2 syndrome. *Lancet (London, England)*, 362(9393):1366–1373.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871.

Zarrei, M., Macdonald, J. R., Merico, D., and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3):172–183.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829.

Zhang, D., Zhao, L., Li, B., He, Z., Wang, G. T., Liu, D. J., and Leal, S. M. (2017). SEQSpark: A Complete Analysis Tool for Large-Scale Rare Variant Association Studies Using Whole-Genome and Exome Sequence Data. *American journal of human genetics*, 101(1):115–122.

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, 32(3):246–251.

Appendix A

Schema definitions

Variant schemas in HBase

Listing A.1 Slice proto schema definition

```
syntax = "proto3";
package protobuf.opencb;
option java_package = "org.opencb.biodata.models.variant.protobuf";
option java_outer_classname = "VcfSliceProtos";
option java_generate_equals_and_hash = true;
import "protobuf/opencb/variant.proto";

message VcfSample {
repeated string sample_values = 1;
// GT is mandatory.
// Saving it separately can create a map of genotypes in Fields
uint32 gt_index = 2;
}

message VcfRecord {
// 1 based
// May contain negative values but it's not likely
int32 relative_start = 1;
// May contain negative values but it's not likely
int32 relative_end = 2;
string reference = 3;
string alternate = 4;
float quality = 5;
VariantType type = 12;
string call = 13;
uint32 filter_index = 6;
repeated string id_non_default = 7;
repeated uint32 info_key_index = 8 [packed=true];
repeated string info_value = 9;
uint32 formatIndex = 10;
repeated VcfSample samples = 11;
repeated AlternateCoordinate secondaryAlternates = 14;
}

message Fields {
repeated string info_keys = 1;
repeated uint32 default_info_keys = 2;
// Possible filter compositions. Delimited by ';'.
// The first filter is the default one
repeated string filters = 3;
// Possible formats compositions. Delimited by ':'.
// The first format is the default one
repeated string formats = 5;
// Possible genotypes seen on the slice.
// The first GT is the default one
repeated string gts = 6;
}

message VcfSlice {
string chromosome = 1;
uint32 position = 2;
repeated VcfRecord records = 3; // List of records (lines)
Fields fields = 4;
}
```