2 | Developing, testing and applying analysis pipelines for family-based exome studies

2.1 Introduction

Although a rare genetic disorder, by definition (according to the European Commission), has a frequency of 1 in 2000, collectively rare diseases affect 6-10% of the population [248]. Rare genetic disorders are associated with high mortality rates, may account for 51% of deaths in children under 1 year [249], add a significant burden to the health care system in terms of cost (accounted for 184% more hospital charges than children who were hospitalized for other reasons [250]) and often under diagnosed [251].

Studying rare genetic disorders is essential to improve the quality of health care services and to obtain a precise and early diagnosis to these patients. Additionally, the insights from rare genetic disorders have helped to improve our understanding of many novel genes and molecular phenomena such as uniparental disomy, parental imprinting and epistatic interactions. These insights have also improved our understanding of the etiology of the risk and pathology of complex disease. For example, studying severe forms of familial insulin resistance has revealed important key genes when studying the common form of Diabetes Mellitus Type II [252].

In the last few decades, researchers have used different approaches to find the underlying genetic causes of rare disorders, such as positional cloning, linkage analysis and candidate gene resequencing among other methods. Despite these great efforts, the Online Mendelian Inheritance in Man (OMIM) [253] database lists 3,675 suspected Mendelian phenotypes without any known molecular basis , as of January 7th 2013. This large number of unidentified disorders shows the limitation of the traditional tools in identifying their genetic causes.

Next Generation Sequencing (NGS) platforms promise to accelerate this process. In 2005, the 454 Roche sequencer was introduced to the scientific community and soon other similar platforms followed, such as the Genome Analyzer from Illumina, SOLiD from Life Technologies and many others (discussed in chapter 1). These NGS platforms are able to generate unprecedented high-throughput DNA sequencing from whole genome or targeted sequences (e.g. exome or linkage regions) in a very short time and at an affordable cost. The first successful example of finding causal variants in a novel gene was published in 2010 when Sarah Ng *et al.* [174] used NGS to sequence the whole exome of four patients with Miller syndrome (OMIM #263750) and showed that mutations in the *DHODH* gene cause this recessive disorder. Soon afterwards, other groups around the world started using NGS to discover the causes of more than 100 novel genes in less than 3 years (Figure 2-1). This number is expected to grow as more researchers adopt NGS platforms for gene discovery in other monogenic disorders [202, 254] (discussed in monogenic disorder section in chapter 1).

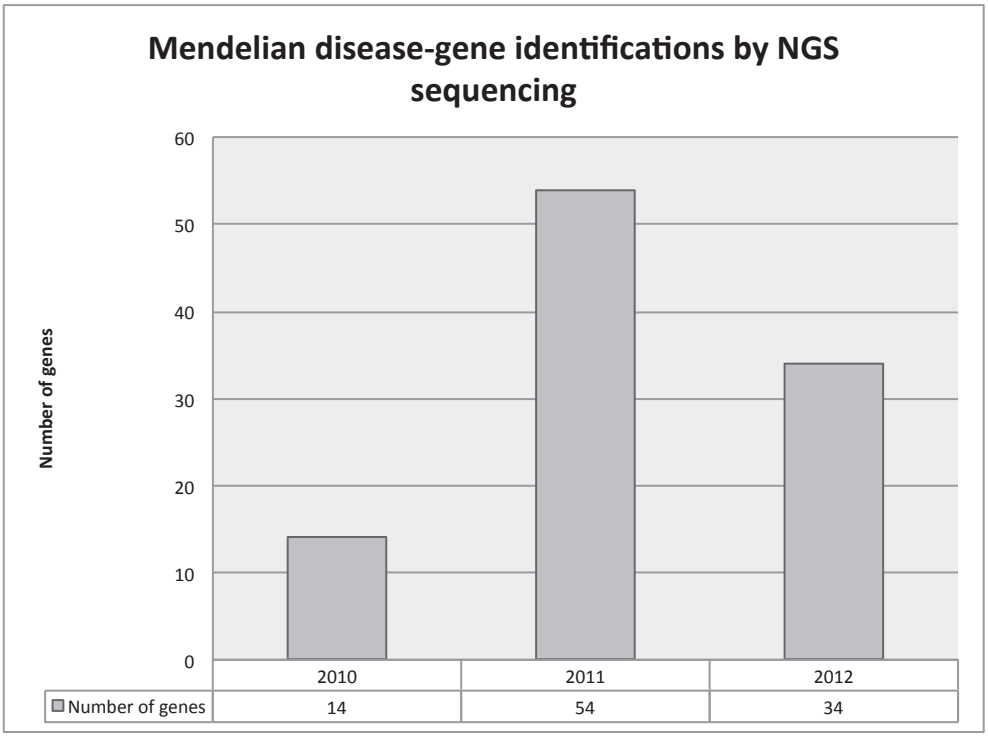


Figure 2-1 Number of Mendelian disease genes identified by NGS 2010 to mid of 2012 [254]

Congenital heart defects (CHD) are considered the most common birth defects worldwide when taken collectively [14]. However, they are considered rare disorders when considered separately (CHD prevalence is review in chapter 1). Inspired by the success of NGS in finding the genetic causes in other rare disorders, I approached CHD using family-based study designs combined with NGS.

However, since the genetic architecture of CHD is not currently clear, I have considered both Mendelian and non-Mendelian contributions to CHD. Not all pathogenic mechanisms can be evaluated using exome sequencing since it targets a small proportion of the genome (only coding DNA regions or < ~1-2% of the human genome size (Table 2-1). Cryptic splice sites, intragenic and long-range promoter variants that affect gene regulation cannot be studied using exome sequencing alone, and as such as they do not fall within the scope of this thesis. The existing examples of genetic causation of CHD are diverse, with respect to both their modes of inheritance and molecular mechanisms, and so investigation of CHD by exome sequencing requires a suite of tools capable of exploring different scenarios.

Table 2-1 lists the major inheritance patterns with syndromic or / and isolated CHD examples from literature, and whether they are amenable to analysis in whole exome sequence data (WES) or not, using tools I developed or implemented to scrutinize the candidate variants.

Table 2-1 Selected patterns of Mendelian and non-Mendelian inheritance and whether they are amenable to analysis using whole exome data. * Indicates mechanisms that have been evaluated in this thesis.

	Inheritance pattern	Example of syndromic and/or isolated CHD	Can be evaluated with WES?	Software	Explored in this thesis?
Mendelian	Autosomal recessive *	Adams-Oliver syndrome OMIM # 100300	Yes	FEVA	Chapter 2, 3 and 4
	Autosomal Recessive (compound heterozygous) *	five affected children with right atrial isomerism were compound heterozygotes for truncating mutations in *GDF1* gene [255]	Yes	FEVA	
	Autosomal dominant *	Alagille syndrome OMIM # 118450	Yes	FEVA	
	X-linked dominant *	Opitz GBBB syndrome OMIM # 300000	Yes	FEVA	
	X-linked recessive *	X-linked heterotaxy OMIM # 306955	Yes	FEVA	
	Y-linked	No reported CHD cases. Unlikely to harbor heart developmental genes	Yes	FEVA	Not explored
Non-Mendelian*	Recurrent *de novo* mutations *	*De novo* mutations in histone-modifying genes in isolated and syndromic CHD cases using exome data [256]	Yes, if in coding regions	DenovoGear	Chapter 3 and 4
	Digenic inheritance *	No reported CHD cases. But as an example: long QT syndrome	Yes	Digenic module	Chapter 3
	Polygenic inheritance	Tetralogy of Fallot [257]	Only with large sample size (in thousands), case/control analysis	Case/Control analysis	Not explored
	Imprinting	Prader-Willi syndrome OMIM # 176270 [258]	Yes, if large segment.	Uniparental Disomy (UPD) caller by Dan King,	Not explored
	Excess affected cases (segregation distortion) *	*MTHFR* C677T polymorphisms may contribute to the risk of CHDs [259]	Yes, in trio based studies	Rare collapsed TDT module	Chapter 3

2.1.1 Chapter overview

The main goal of this chapter is to describe the pipelines and analytical tools I developed and then applied to evaluate the utility of four family-based study designs (index cases with linkage analysis, affected sib-pairs, trios and affected parent-child). The lessons learnt from these analyses were subsequently applied to two CHD subtypes (Tetralogy of Fallot and Atrioventricular Septal Defects) in chapters 3 and 4, respectively. Figure 2-1 shows the main analytical components required for family-based exome studies.

In this chapter, first, I describe the **three pipelines used to call SNVs and indels** from all CHD samples included in this thesis. My colleagues at the Wellcome Trust Sanger Institute implemented two of the three pipelines (the Genome Analysis Production Informatics (GAPI) and the (UK10K) pipelines whilst I implemented the third one to call *de novo* variants, which was later adapted by Ray Miller for the Deciphering Developmental Disorders (DDD) project [260].

Each pipeline outputs a large number of variants including many false positive variants that would adversely affect any downstream analysis. At the beginning of my work on exome sequencing three years ago, it was not clear what best practices I should use to **improve the sensitivity and specificity** of variant calling. In the second part of the results, I describe how I chose **various filters** such as strand bias, phred-like quality scores among other filters to improve the sensitivity and specificity of the variant calls. Choosing the right filters is a dynamic research area and the best practices are expected to change to reflect new statistical models for variant calling. Many of the results I describe in this section do not reflect the current best practices but they represent examples of how to approach and set proper filter thresholds in exome-based studies. In addition to these filters, I discuss how I merged the variant calls from **multiple callers** to enhance sensitivity. I show that the precise manner in which the outputs from these callers are combined can have an unexpectedly large effect on the number of candidate variants

Once I have obtained a high quality set of variants for each sample, I describe in the third part of the results, how I used minor allele frequency and additional family data to minimize the search space for causal variants. These combined steps reduce the search space for causal variants to a few tens or hundreds instead of tens of thousands of variants.

Finally, I describe a suite of tools that I have designed to automate many steps discussed above. Although similar software, such as SVA, EVA and VarSift [261-263], have been published during my PhD, none of them were able to fulfill the

needs for my studies. One of the main drawbacks of these tools is that they are not suitable for high-throughput analysis. Additionally, most of them use hard coded filters, which is not practical to explore new filters. For these reasons, I developed a suite of tools called **Family-based Exome Variants Analysis (FEVA)** that reports candidate variants under different modes of inheritance (autosomal recessive, autosomal dominant and X-linked) for different study designs (index cases, affected sib-pairs, affected parent-child, and trios). In the last part of this chapter, I show how I used FEVA to identify pathogenic and candidate pathogenic genes under different study designs using real examples.

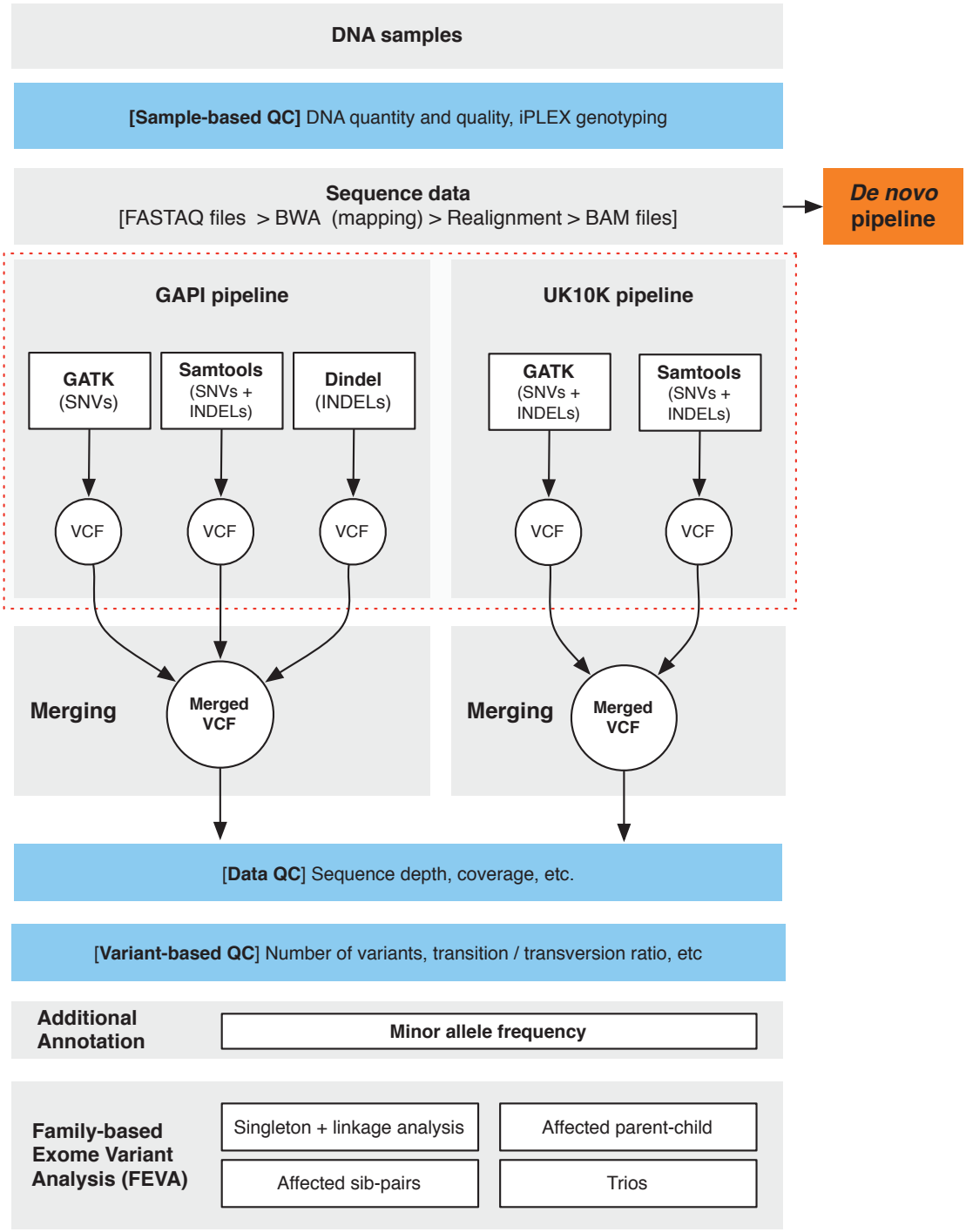


Figure 2-2 Overview of pipelines, tools and annotation discussed in this chapter.
Blue boxed are quality control tests that are performed at different stages of the workflow. The two main pipelines used to call variants from sequence data are GAPI and UK10K. A third one, the de novo pipeline (orange box), uses the sequence data (BAM files) and includes further steps described in Figure 2-9. Additional descriptions of these steps are available in Table 2-2.GAPI: the Genome Analysis Production Informatics pipeline, UK10K: UK10K variant calling, SNVs: single nucleotide variants, INDELs: insertion and deletion, QC: quality control.

Table 2-2 A list of main analytical tasks described in this chapter with a short description of each section.

Task	Section	Description
Variant calling pipelines	Genome Analysis Production Informatics (GAPI) pipeline	To call single nucleotide (SNVs) and insertion/deletion variants (INDELs) using three callers (Samtools, GATK and Dindel) in 381 CHD samples
	UK10K pipeline	Used to call SNVs and INDELs variants using two callers (Samtools and GATK) in 125 CHD samples.
	De novo variant calling pipeline	Used to call *de novo* SNVs and INDELs variants using one caller (DenovoGear) in 252 CHD trios
Improving sensitivity and specificity	Sample-based DNA quality test (DNA samples)	Various tests to detect the quantity and quality of the DNA samples and any possible sample contamination and swapping issues.
	Sample-based data quality test (Sequencing data)	Quality of NGS sequencing data in terms of depth, coverage and other parameters.
	Variant-based quality tests	Quality of variant calling based on the number of variants, genotypes, variants predicted effect on the protein and other quality ratios.
	Filtering low quality variants	Multiple filters based on thresholds of quality metrics used to exclude low quality variants
	Using multiple callers	Combining multiple variant callers (e.g. Samtools, GATK and Dindel) to overcome the deficiencies of individual callers
Minimizing the search space for causal variants	Minor allele frequency (MAF)	Using different population-based MAF resources to exclude common variants (>1%) and the effect of allele matching algorithm.
	Family-based designs	The effect of considering additional members of the family (either healthy or affected) on the final number of candidate variants and genes
Applications	FEVA suite	An easy to use suite of programs I developed to automate many of the steps discussed above (minimize the search space for causal variants and prioritization). These tools are available for small scale use with a graphical user interface and as common-line tools for high-throughput analysis.
	Simple monogenic diseases combined with linkage analysis	Use of FEVA to find pathogenic variants from four different index cases within linkage intervals for different neurodevelopmental monogenic disorders
	Affected sib-pairs	Use of FEVA to analyze CHD in affected sib-pairs from eight non-consanguineous and two consanguineous families.
	Affected parent-child	Using FEVA to analyze CHD in three affected parent-child pairs.
	Example of affected trios combined with candidate gene screening	Use of FEVA to analyze 1,080 trios from Deciphering Developmental Disorders (DDD) project trios and screen 1,142 candidate genes.

2.2 Methods

2.2.1 Samples and phenotypes

Table 2-3 summarises the different sample collections that I analyzed to evaluate the utility of different study designs. These sample collections were accessed through collaboration with various researchers and clinicians from the UK, Europe and Canada. All samples were collected from the families after obtaining informed consents and approved by the Ethical Review Boards of their respective organizations. Not all of the analyses of these sample sets are described in detail in this thesis.

Table 2-3 Samples and family-based study designs included in this thesis.
* Sample cohorts discussed in this chapter. GO-CHD: Genetic Origins of Congenital Heart Disease Study, DDD: Deciphering Developmental Disorders project, AVSD: atrioventricular septal defects. TOF: tetralogy of Fallot.

Design	Targeted Region	Cohort	Origin	Consanguineous	Phenotype	Number of families or samples
Index cases	Whole exome	GO-CHD	UK	No	Various CHD	110
		Toronto	Canada	No	AVSD	78
	Linkage region	Amish*	USA	No	Various Neurodevelopmental	4
Trios	Whole exome	GO-CHD	UK	No	Various CHD	2
		Newcastle	UK	No	TOF	30
		Toronto	Canada	No	AVSD	3
		Leuven	Belgium	No	AVSD	10
		DDD	UK	No	Developmental	1,080
	Candidate genes	Newcastle	UK	No	TOF	250
Affected sib-pairs	Whole exome	Toronto	Canada	No	AVSD	1
		Birmingham*	UK	Yes	Various CHD	2
		Birmingham*	UK	No	Various CHD	8
		GO-CHD*	UK	No	Various CHD	1
Affected parent-child	Whole exome	GO-CHD*	UK	No	Various CHD	3

2.2.2 DNA preparation and Quality Control

Our collaborators extracted the DNA from the patients' blood and / or saliva and sent the samples to the Sanger Institute for quality control before they were submitted for sequencing. The DNA sample quality control included three tests. The first was to determine the amount and concentration of DNA, which was analyzed by gel or picogram. The second test detected the sample's gender by genotyping SNPs on the sex chromosomes and compared it to the supplier sheet in order to detect any potential gender mismatches. The third test was to check for the possibility of sample contamination or swapping by genotyping another 30 SNPs. The genotyping was done using Sequenom platform and any sample, which failed one of these tests, was flagged for replacement or exclusion. The Sample Logistic Team at the Sanger Institute performed these quality control tests.

2.2.3 Target capturing and sequencing

DNA (1-3µg) was sheared to 100-400 bp using a Covaris E210 or LE220 (Covaris, Woburn, MA, USA). Sheared DNA was subjected to Illumina paired-end DNA library preparation and enriched for target sequences (Agilent Technologies; Human All Exon 50 Mb - ELID S02972011) according to manufacturer's recommendations (Agilent Technologies; SureSelectXT Automated Target Enrichment for Illumina Paired-End Multiplexed Sequencing). Enriched libraries were sequenced using the HiSeq platform (Illumina) as paired-end 75 base reads according to manufacturer's protocol.

2.3 Results

2.3.1 Assessing variant calling pipelines

2.3.1.1 Genome Analysis Production Informatics (GAPI) and UK10K pipelines

There are several pipelines deployed at the Wellcome Trust Sanger Institute (WTSI) to call variants from human whole genome and / or whole exome data. The majority of samples analyzed in this thesis were processed through the Genome Analysis Production Informatics (GAPI) pipeline (managed by Carol Scott *et al.*) except 125 samples that formed part of the UK10K RARE project, which were processed through the UK10K pipeline (managed by Shane McCarthy *et al.*) [264]. Both pipelines are used to call single nucleotide variants (SNVs) as well as insertion/deletion variants (INDELs). The GAPI pipeline provided single-sample calling only while UK10K pipeline provided both single and multi-sample calling. Although, the latter has some potential advantages, I decided to use single-sample calling only in order to be able to compare variants from both pipelines.

However, differences between these pipelines led to variability in the type and numbers of variants (Table 2-4, Table 2-5 and Figure 2-5). Data that were processed through the GAPI pipeline tend to have a larger number of SNVs and INDELs compared to UK10K pipeline. GAPI sequence data had 60% more SNVs compared with UK10K data although most of these differences can be attributed to non-coding variants which include intronic, intragenic, downstream, upstream and variants in untranslated regions UTRs).

To see if using different filters and thresholds in Table 2-5 caused the difference seen in SNVs counts between the two pipelines, I applied UK10K's filters on samples from the GAPI pipeline. First, I created a new set of samples called GAPI-II by merging variants from GATK and Samtools only and excluding Dindel calls since it is not part of the UK10K pipeline. This set of samples showed a similar number of coding and non-coding variants between both pipelines (Figure 2-4)

except for loss-of-function variants (LOF) where the UK10K pipeline has almost double the number of LOF variants compared with GAPI or GAPI-II (t test, *P* value < 2.2×10^{-16}). A difference in a caller version and its underlying statistical model is likely to cause this variation. This is more readily observed in LOF counts since they are fewer than missense variants and have a lower number of true variants and so are more sensitive to calling errors.

On the other hand, INDELs show larger differences between GAPI and UK10K pipelines (Figure 2-5). GAPI calls almost two to three times more INDELs than UK10K or GAPI-II (Figure 2-5-A). This is true regardless of the location of the indel with respect to coding sequences (Figure 2-5 B, C and D). One explanation for this observation would be the use of an additional caller specifically designed to call INDELs, called as Dindel, in the GAPI pipeline but not in the UK10K pipeline. Dindel is a dedicated caller for INDELs that uses a probabilistic realignment model to account for base-calling errors, mapping errors, and for increased sequencing error INDEL rates in long homopolymer runs [158]. Dindel's superior performance comes at a price of high computation demands, and the same underlying model has been incorporated into later versions of SAMtools, which is why the UK10K informatics team has refrained from using it on large numbers of samples.

Table 2-4 Similarities and differences between the components of Genome Analysis Production Informatics (GAPI) pipeline and the UK10K pipeline. Multiple factors are likely contribute to the differences in the number of variants generated by GAPI compared with UK10K pipeline such as the number of used callers, different software versions which usually reflect subtle changes in the underlying statistical models, filters and thresholds and how the output from different callers is merged (i.e. the order of callers from the most to least preferred, see section 2.3.2.2 for details)

Step	Goal / Description	GAPI	UK10K
Reference genome	Which version of the human reference genome used	GRCh37 (hs37d3) 1000 genome phase II reference	GRh37 (human_g1k_b37) 1000 Genomes Phase 1 reference
Align sequence reads to reference genome	Generate SAM/BAM files	BWA (v0.5.9-r16)	BWA (v0.5.9-r16)
Mark duplicates	To mitigate the effects of PCR amplification bias introduced during library construction.	Picard tools (v1.46)	Picard tools (v1.46)
Realignment around indels	Enhance variant calling	GATK (v1.4-15)	GATK (v1.1-5-g6f432841)
Base quality score recalibration	Recalibrate base quality scores of reads according to the base features (e.g., reported quality score, the position within the read)	GATK (v1.4-15)	GATK (v1.1-5-g6f43284)
Calling target region	Calling variants is limited to the coding regions plus variable flanking region	Exon bait regions plus or minus a 100bp window	Exon bait regions plus or minus a 100bp window
SNV calling	Single nucleotide variants calling programs	Samtools (v0.1.16) GATK (v1.0.15777)	Samtools (v0.1.17) GATK (v1.3-21)
INDEL calling	Insertion and deletion variants calling programs	Samtools (v0.1.16) Dindel (v1.01)	Samtools (v0.1.17) GATK (v1.3-21)
Variant predicted effect	The effect of variant on the protein is predicted by VEP	VEP 2.2 to 2.4	VEP 2.6 to 2.8
Caller merging	The order of which variants called by different callers are merged	Dindel > GATK > Samtools	GATK > Samtools
General filters	Filters applied during variant calling	See Table 2-5 for details	

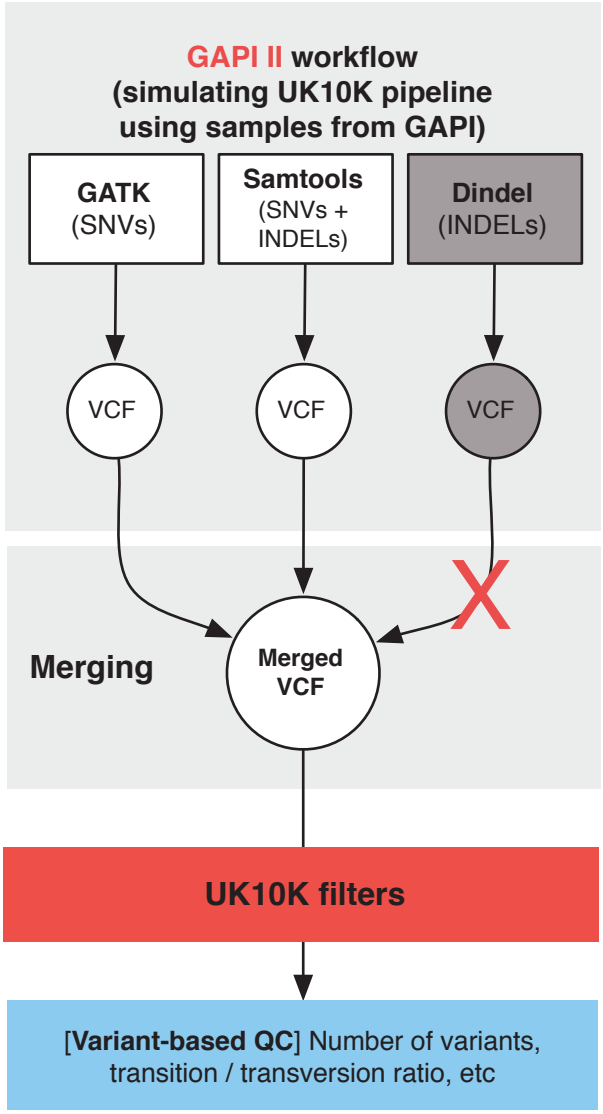


Figure 2-3 A workflow diagram to describe how I generated VCF files for GAPI-II set. The main goal is to use files from GAPI pipeline and apply similar workflow to UK10K and see if this would be enough to explain the differences between the pipelines.
Each sample from the original GAPI pipeline has three VCF files of variants called by GATK, Samtools and Dindel. I merged VCF files from GATK and Samtools but not from Dindel. Next, I applied the same filters used by UK10K to exclude low quality variants (filters were supplied by Shane McCarthy). A list of UK10K filters is available in Table 2-5.

Table 2-5 Filters and thresholds applied on variants from UK10K and GAPI pipelines.

Variant callers	Filters	Variant type	Pipelines	
			GAPI	UK10K
Samtools	Depth at locus (DP)	SNVs	4 < DP and DP > 1200	4 < DP and DP > 2000
		INDELs	4 < DP and DP > 1200	4 < DP and DP > 2000
	Mapping quality (MQ)	SNVs	MQ <=10	MQ <= 25
		INDELs	MQ <= 10	MQ <= 25
	Genotype quality (GQ)	SNVs	NA	GQ <= 25
		INDELs	NA	GQ <= 60
	Variant quality (QUAL)	SNVs	NA	QUAL <= 30
		INDELs	NA	QUAL <= 60
	StrandBiasPval	SNVs	StrandBiasPval < 0.0001	NA
		INDELs	StrandBiasPval < 0.0001	NA
	BaseqBiasPval	SNVs	BaseqBiasPval < 1e-100	NA
		INDELs	BaseqBiasPval < 1e-100	NA
	MapqBiasPval	SNVs	MapqBiasPval < 0	NA
		INDELs	MapqBiasPval < 0	NA
	EndDistBiasPval	SNVs	EndDistBiasPval < 0.0001	NA
		INDELs	EndDistBiasPval < 0.0001	NA
	MinbpfromGap	SNVs	MinbpfromGap < 10	NA
		INDELs	MinbpfromGap < 10	NA
GATK	Variant quality (QUAL)	SNVs	QUAL < 30	QUAL < 30
		INDELs	NA	NA
	Quality by Depth (QD)	SNVs	QD < 5.0	QD < 5
		INDELs	NA	QD < 2
	Homopolymer run length (Hrun)	SNVs	HRun > 5	Hrun > 5
		INDELs	NA	NA
	Strand bias (SB)	SNVs	SB > 10	SB > -0.1
		INDELs	NA	NA
	Fishers p-value (FS)	SNVs	NA	FS > 60
		INDELs	NA	FS > 200
	ReadPosRankSum	SNVs	NA	NA
		INDELs	NA	< -20
	InbreedingCoeff	SNVs	NA	NA
		INDELs	NA	< -0.8
	InDel	SNVs	Filtered if site covered by known indel mask file	Filtered if site covered by known indel mask file
		INDELs	NA	NA
	LowQual	SNVs	Repeat of QUAL < 30 (applied at calling)	NA
		INDELs	NA	NA
	SnpCluster	SNVs	Filtered if 3 SNPs within a 10bp window	NA
		INDELs	NA	NA
	Depth at locus (DP)	SNVs	4 < DP and DP > 1200	NA
		INDELs	4 < DP and DP > 1200	NA
	Hard to validate	SNVs	MQ0 >= 4 and (MQ0/(1.0*DP))	MQ0 >= 4 and (MQ0/(1.0*DP))
		INDELs	NA	NA
Dindel	Homopolymer run length (hp10)	INDELs	HRun > 10	NA
	Variant quality (q20)	INDELs	QUAL < 20	NA
	Non-reference allele (fr0)	INDELs	Not covered by at least one read on both strands	NA
	Multiple indels in the same window (wv)	INDELs	Other indel in window had higher likelihood	NA

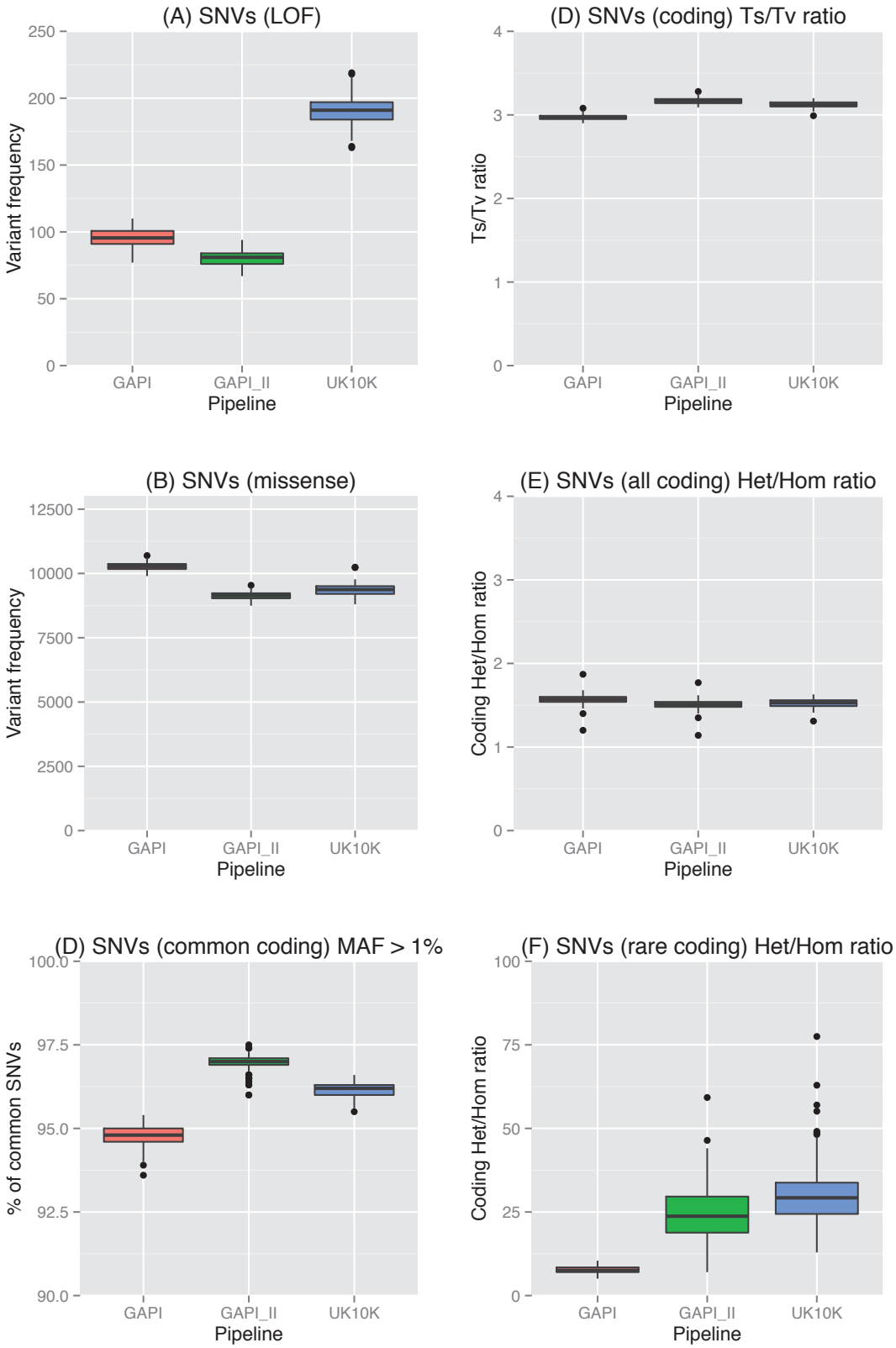


Figure 2-4 Differences in the counts of coding single nucleotide variant (SNVs) between GAPI and UK10K pipeline and GAPI_II, which include the same sample in GAPI but subjected to UK10K's filters (i.e. I applied the UK10K filter in Table 2-5 on GAPI samples).
LOF: loss-of-function variants include stop gain and variant disturbing donor or acceptor splice sites. Ts/Tv: Transition/Transversion ratio. Hom/Het: Homozygous/ Heterozygous ratio.

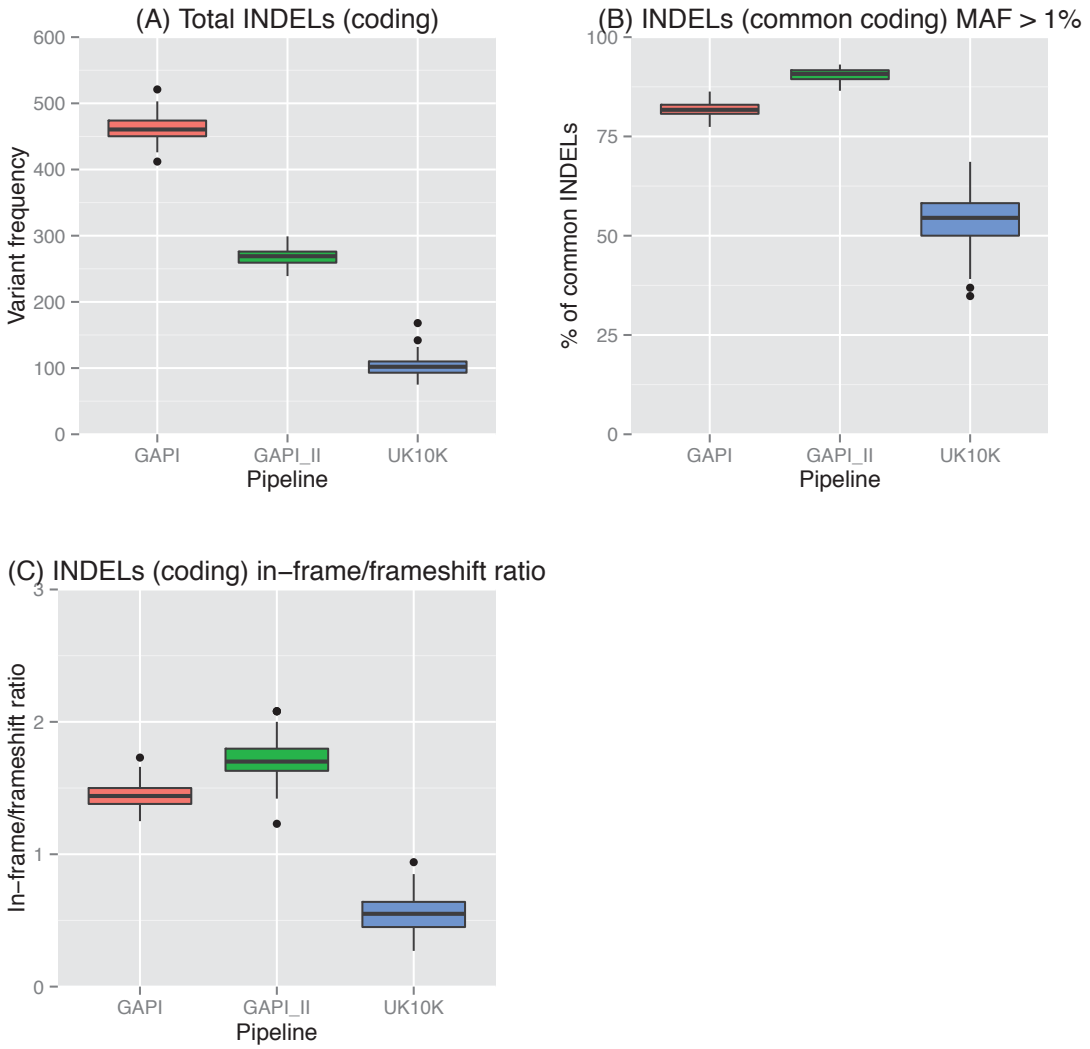


Figure 2-5 Differences of insertion-deletion variant (INDELs) counts between GAPI, UK10K pipeline and GAPI_II which are the same sample in GAPI but subjected to UK10K's filters).

2.3.1.2 Differences between GAPI releases

Since most of the samples analyzed in this thesis went through the GAPI pipeline at different points of my PhD, I sought to examine the effect of different releases of GAPI pipelines on the samples from three CHD cohorts (Figure 2-6 and Figure 2-7). The first cohort includes 94 samples of mostly atrioventricular septal defects (AVSD) children collected from SickKids hospital, Toronto, Canada (labeled as CHDT). The second cohort includes 90 samples of Tetralogy of Fallot (TOF) affected trios from the University of Newcastle while the third cohort includes 24 samples of affected sib-pairs of samples affected with various CHD

subtypes (about a quarter of these samples are from consanguineous families of a Pakistani origin). I found the variant counts were consistent between these cohorts even though they were generated at different times and with different versions of the GAPI pipeline. Small variations may occur as a result of systemic differences caused by the depth of the sequencing, or the population ancestry of the samples (e.g. samples with African ancestry are expected to have more variants than non-African samples).

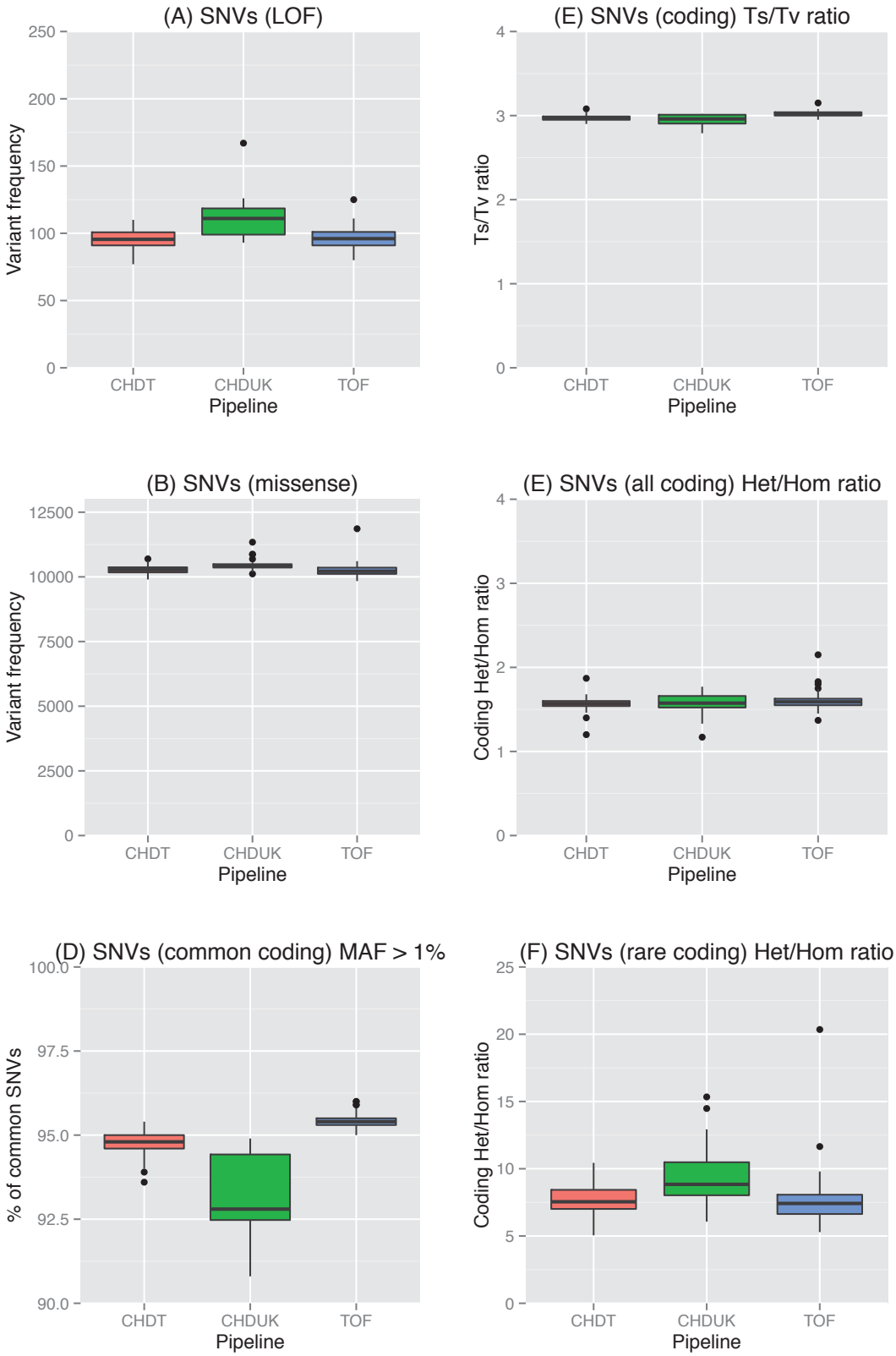


Figure 2-6 Differences of single nucleotide variant (SNVs) counts between GAPI studies. CHDT: Congenital heart defect samples from Toronto (discussed in chapter 4). CHDUK: Congenital heart defect samples from UK (discussed in application section in this chapter), TOF (Tetralogy of Fallot samples discussed in chapter 3). Ts/Tv: Transition/ Transversion ratio. Hom/Het: Homozygous/Heterozygous ratio.

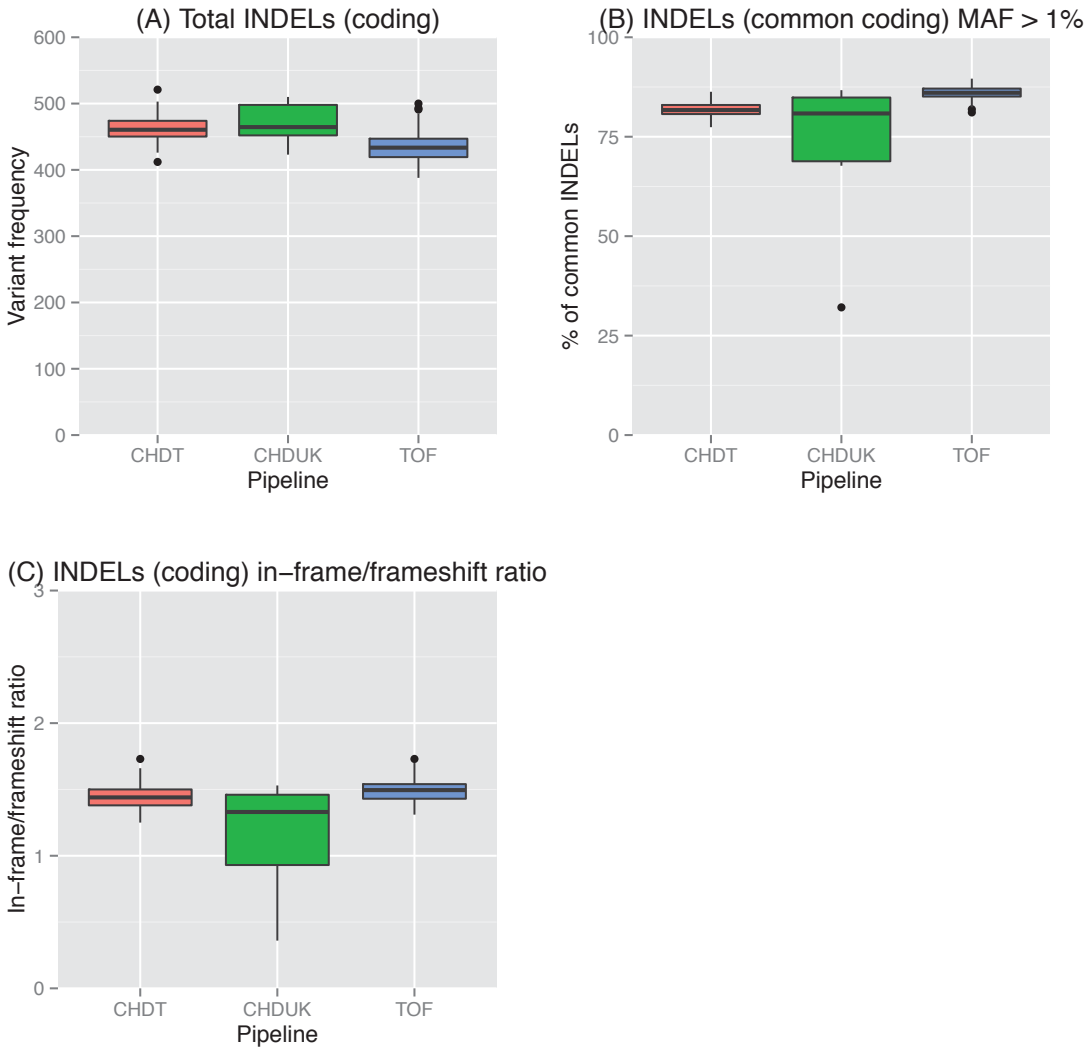


Figure 2-7 Differences of insertion-deletion variant (INDELs) counts between GAPI studies. CHDT: Congenital heart defect samples from Toronto (discussed in chapter 4). CHDUK: Congenital heart defect samples from UK (discussed in application section in this chapter), TOF (Tetralogy of Fallot samples discussed in chapter 3).

2.3.1.3 Implementing a de novo variant calling pipeline

Initially, I tried to identify potential *de novo* variants based on the variants called by either GAPI or UK10K pipelines in the child and not in parents. However, this approach yields a large number of candidate *de novo* variants per trio. A more efficient approach is to discover potential *de novo* variants from the child and his parents in a unified statistical framework. I designed and implemented a pipeline to call, filter, annotate and visualize *de novo* variants from trio-based studies based on DenovoGear program [265, 266]. This software was developed by Don Conrad and adopts a Bayesian approach to calculate the posterior probability of a *de novo* mutation at a single locus using the joint likelihood of the read-level data for all three trio members. DenovoGear outputs ~170 plausible *de novo* variants (with a posterior probability of greater than 0.001) per trio on average. However, most of these candidate variants are false positive since the expected number of *de novo* coding variants is ~1 according to published studies [190, 267-271].

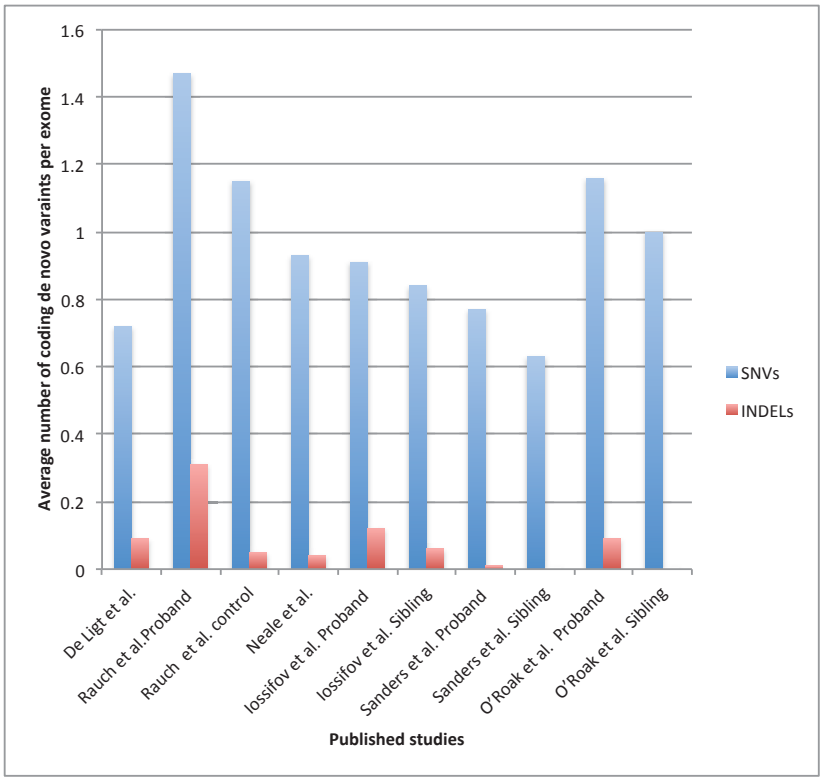


Figure 2-8 Average number of coding *de novo* variants per exome in different trio-based studies [190, 267-271]. (The literature survey and data are courtesy of Dr. Matthew Hurles)

In order to keep the number of false positive variants as small as possible, I applied five filters to exclude: (i) variants in tandem repeat or segmental duplication regions, (ii) common variants with minor allele frequency > 1% in the 1000 genomes [155], NHLBI-ESP exome project [199] and the UK10K Twins cohort [264], (iii) when > 10% of the reads in either parent support the alternate allele (i.e. the variant is more likely to be inherited from a parent), (iv) variants not called by an independent caller such as SamTools, Dindel or GATK, and (v) variants predicted to be non-coding by the VEP tool [170]. Collectively, these filters effectively remove ~98.8% of the original candidate *de novo* variants (leaving ~1.8 coding plausible *de novo* candidate per exome).

This pipeline was used to automate several tasks designed to obtain high quality sets of candidate *de novo* variants from trios. This first step is calling candidate *de novo* variants from whole genome or whole exome data from human or mouse trio samples, followed by applying various filters to improve the specificity of the calls. The pipeline was designed in a modular fashion where each step generates intermediate files that are used as input for subsequent steps (steps are listed in Figure 2-9). This design allows the end user to change the pipeline by modifying steps and files or add new steps in order to customize the pipeline to suit the need of different studies.

One of the challenges faced by this pipeline is the run time per trio (~12 hours for whole exome data and up to 36 hours for whole genome data). To make the pipeline run faster, especially for large-scale project, I modified the code (which I wrote in Python programming language) to split sequence data in each sample into 24 segments (by the chromosome) and run them in parallel. This has shortened the run time to 2-3 hours for whole exome data and 10-12 hours for whole genome data. Moreover, another layer of parallelism is achievable by running multiple trios at the same time, which is suitable for large-scale projects such as the Deciphering Developmental Disorders (DDD) project with thousands of trios.

I used this pipeline to call *de novo* variants in 238 trios affected with Tetralogy of Fallot in the third chapter and in 13 trios with atrioventricular septal defect in the fourth chapter. Moreover, this pipeline has been used successfully in several whole genome sequencing projects in human and mouse pedigrees that are investigating the factors influencing rates of germline mutation.

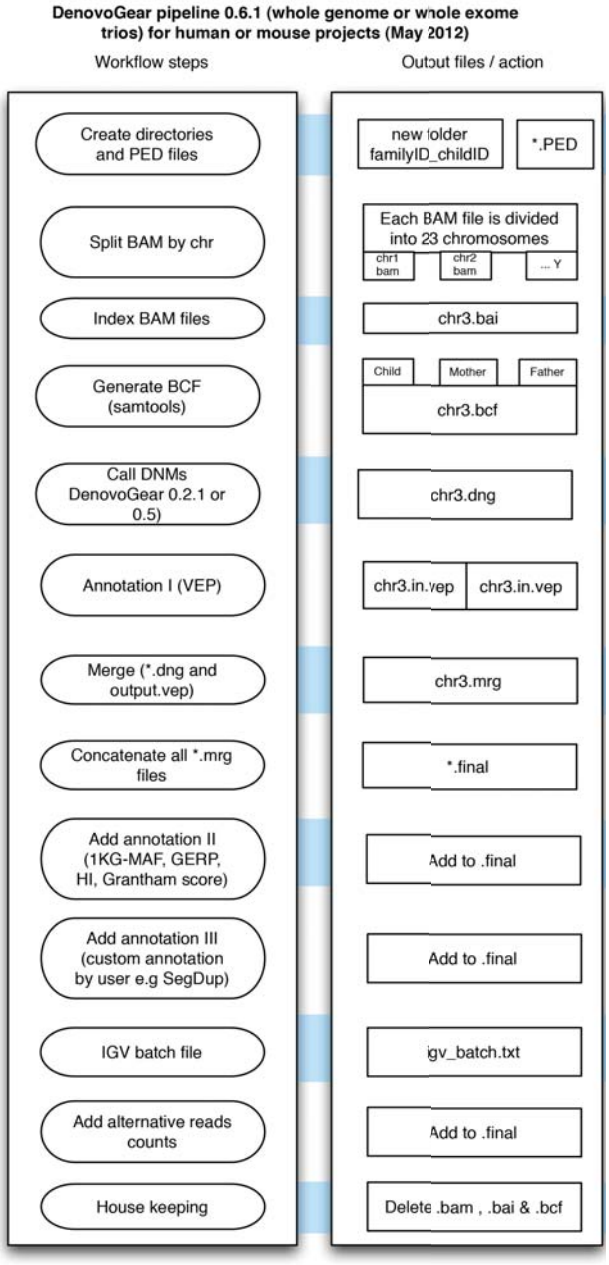


Figure 2-9 **The workflow of the DenovoGear pipeline.**
PED: pedigree files. BCF are binary files of VCF (variant call format) that are generated by Samtools mpileup with genotype likelihoods required by DenovoGear [272]. DNMs: *de novo* mutations. VEP: variant effect predictor [170]. 1KG-MAF: 1000 genomes minor allele frequency. GERP: Genomic Evolutionary Rate Profiling scores [164]. HI: haploinsufficiency scores [273].

2.3.2 Minimizing the rate of false positive variants

2.3.2.1 Variant-based filters

At the beginning of my PhD studies, it was not clear what were the best practices I should use to improve the sensitivity and specificity of variant calling from exome data. To investigate this aspect of data analysis, I tested different filters in order to determine the best callset possible from CHD samples called by the UK10K pipeline. These callsets include raw unfiltered variants called by GATK (G), Samtools (S), or both callers (GS). In this analysis, I focused mainly on SNVs since they are the most abundant variants and represent a large proportion of the known pathogenic variants [274]. More importantly, there are many high quality training SNVs data sets available to improve variant quality (e.g. HapMap). On the other hand, indels were, and still are, more difficult to call and tend to have a higher false positive rate [155].

SNVs are thought to be among the easiest variant classes to call from NGS data but nonetheless sequencing errors can generate false positive calls. Sequencing error rates depend on factors such as the context of the DNA sequence, depth of sequencing, and the type of substituted bases among other factors [143]. To control for these biases in the exome NGS data, I examined the relationship between strand bias (SB), quality by depth (QD), genotype quality (GQ) and variant quality (QUAL) with transition/transversion ratio (Ts/Tv). This ratio has been used by different groups in the 1000 genomes consortium as a quality control test and typically ranged between 2.9-3.3 in coding regions based on sequence data from different NGS platforms. I used the Ts/Tv ratio as the truth measurement to determine the proper thresholds values for each one of the four filters.

Variant quality (QUAL)

The QUAL parameter is the phred-scaled quality score probability of the alternative allele at a given site in sequencing data being wrong. This scale is calculated as:

$$QUAL = -10 * \log (1-p)$$

where p is a base-calling error probability. A value of 10 indicates one in 10 chance of error, while a value of 100 indicates one in 100 chance. Higher QUAL values indicate higher confident in the variant calls. I plotted the QUAL scores for eight different callsets based on filtered and unfiltered variants from Samtools, GATK or both against the Ts/Tv ratio (Figure 2-10). The Ts/Tv ratio was at its highest when variants are called by both GATK and Samtools and pass the callers internal filters (Figure 2-10, dashed red line) and dropped slightly below 3 when the QUAL was < 30, which I used as the minimum accepted threshold.

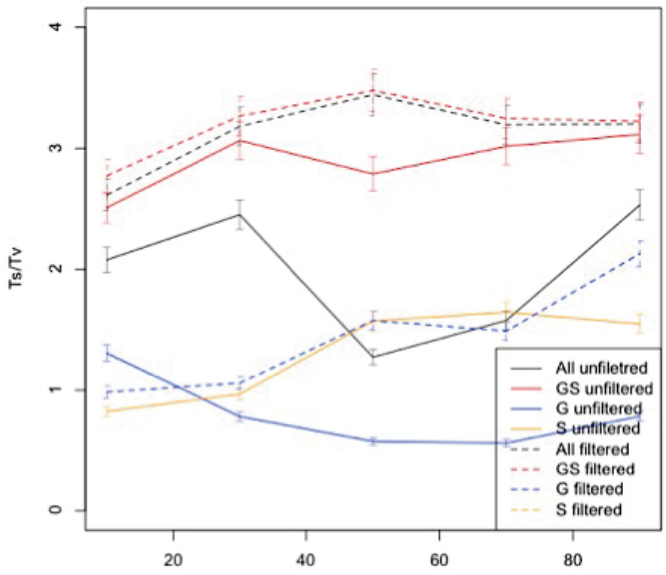


Figure 2-10 The relationship between variant calling quality (QUAL) and the transition/ transversion ration (Ts/Tv) of coding SNVs. The plot shows eight different callsets based on variants called by a single caller or two callers and whether the internal filters of a caller were applied (filtered) or not (unfiltered). These internal filters are usually part of the pipeline itself. (S) is a variant callset called by Samtools alone, (G) variants called by GATK alone, (GS) variants called by both Samtools and GATK, and (All) is a callset composed of variants from the previous three callsets. The GS filtered callset (dashed red line) is the only callset that shows a Ts/Tv ratio close to the expected range (2.9-3.3). However, since the Ts/Tv ratio of this callset drops below QUAL of 30, I used this value as the minimum threshold of high quality variants. Any variants with QUAL < 30 were excluded from the downstream analyses.

Quality by depth (QD)

The QD is a simple statistic to quantify the variant confidence given as 'variant confidence' (from the QUAL field) divided by 'unfiltered depth of non-reference samples' where low QD scores are indicative of false positive calls [275]. QD is only available for variants called by GATK only and thus I was not able to test variants called by Samtools (Figure 2-11). Similar to the QUAL metric above, the variant callset closest to the expected Ts/Tv ratio is the one called by both GATK and Samtools and has passed their internal filters (dashed red line). Unfiltered variants with QD < 5 has significantly lower Ts/Tv ratio below 2.0, which is the minimum accepted threshold I chose for QD (Figure 2-11).

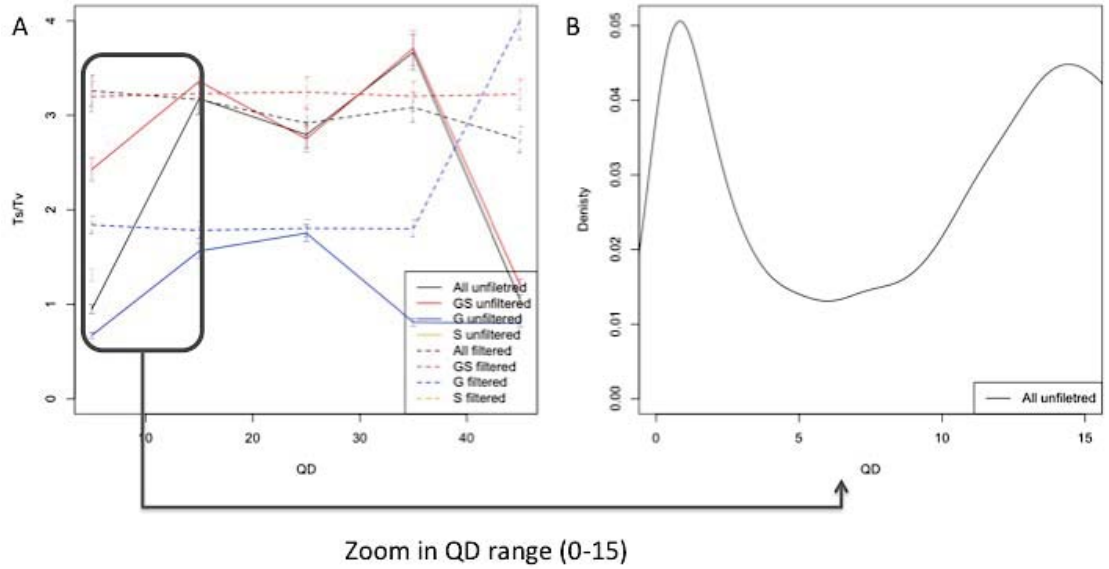


Figure 2-11 The relationship between quality by depth (QD) and the transition/ transversion ration (Ts/Tv) of coding SNVs. (A) I plotted QD values from eight different callsets as described in the previous figure (Figure 2-10). QD values are available for GATK variants, thus variants called by Samtools alone are not shown. The GS filtered callset (dashed red line) the closest Ts/Tv ratio to the expected range (2.9-3.1) is and was consentient along QD values on the X axis. (B) To choose the appropriate minimum QD threshold, I plotted the QD values of all variants, regardless of the caller, from unfiltered callset (All unfiltered, black dashed line in plot A) and restricted the QD to values between 0-15. This shows variants with QD < 5 are enriched for low quality variants (i.e. did not pass the internal filters).

Strand bias (SB)

The third filter I assessed was the strand bias (SB) metric, which quantifies the evidence of a variant being seen on only the forward or only the reverse strand in the sequencing reads. Higher SB values > 0 denote significant strand bias and

are associated with lower values of Ts/Tv ratio, therefore they are more likely to indicate false positive calls (Figure 2-12).

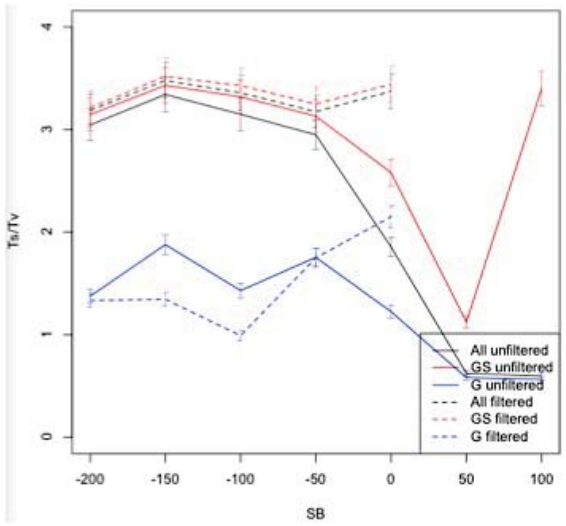


Figure 2-12 The relationship between strand bias (SB) and the transition/ transversion ration (Ts/Tv) of coding SNVs. I plotted SB values from eight different callsets as described in the previous figure (Figure 2-10). At the time, SB values were available for GATK variants only and thus variants called by Samtools are not shown. The callset with closet Ts/Tv ratio to the expected range (2.9-3.1) is the GS filtered callset (dashed red line) and was consentient along SB values (-0.01 to -200). The Ts/Tv ratio values drop dramatically when SB > 0 (solid lines).

Genotype quality (GQ)

Finally, the GQ is another phred-scaled score that represents the confidence of the true genotype at a certain locus. In a diploid genome, the homozygous reference, heterozygous , and homozygous non-reference genotypes are denoted ('0/0' , '0/1' and '1/1') respectively in the variant call format files (VCF files). For a heterozygous genotype (0/1), the genotype quality (GQ) is calculated as :

$$\frac{L(0/1)/L(0/0)}{L(0/1)/L(1/1)}$$

where L is the likelihood of a genotype given the NGS sequence data at that locus. Variants with a GQ of < 30 tend to have lower Ts/Tv ration (~2.7) and hence I used this as the minimum cutoff (Figure 2-13)

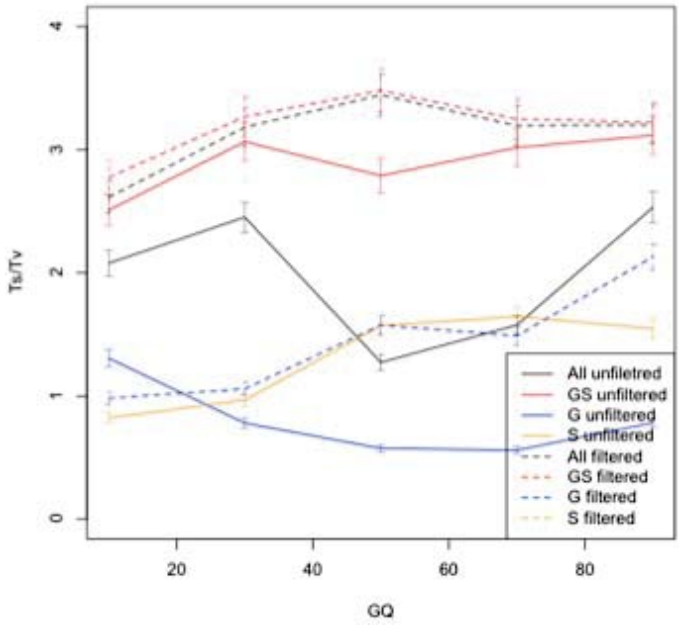


Figure 2-13 The relationship between genotype quality (GQ) and the transition/ transversion ration (Ts/Tv) of coding SNVs. I plotted GQ values from eight different callsets as described in the previous figure (Figure 2-10). The callset with closet Ts/Tv ratio to the expected range (2.9-3.1) is the GS filtered callset (dashed red line) when GQ values > 30.

These four filters where used at the early stages of my analyses of UK10K data to improve the sensitivity and specificity of SNVs calling. It is important, however, to notice that choosing the best filters with highest sensitivity and specificity remains an active area of research. As the developers keep tuning the underlying statistical models in their variant calling programs, these filters need to be adjusted accordingly to reflect the current best practices. More importantly, reviewing the results of validation experiments using capillary sequencing periodically is essential to gain insights about the performance of each filter.

2.3.2.2 Merging caller sets and caller priority

In order to increase the confidence of variant calls, the GAPI pipeline used two independent callers with different underlying probabilistic statistical models to detect SNVs and two callers for INDELs [152-154, 276]. GATK and Samtools were used to call SNVs and while Samtools and Dindel are used to call INDELs. Since Samtools are used to call both SNVs and indels, the GAPI pipeline generates three

files, one from each caller in a variant call format (known as VCF files) [161], per sample.

Using three files separately would complicate downstream analyses since two callers do not agree on the total number of variants, genotypes, and alternative alleles. For example, two SNV callers may detect different alternative alleles at a given locus or report different genotypes (e.g heterozygous by one and homozygous non-reference by the other). To overcome this issue, I decided to merge the three VCF files into a single file per sample. This would have been an easy task if the two callers agreed on all variants, but since this is not the case, I needed to decide on which caller of the two, generated a more reliable set of variants and thus should be used in the conflict cases.

To answer this question, I generated seven different callsets, (Table 2-6 first column) where each callset is composed of at least one group of variants from five scenarios (from 1 to 5). These five scenarios are based on the variant's status according to the two callers (A and B). A variant status can have one of three possible values: (PASS) when a variant is called and passes the caller's filters, (Non-PASS) when a variant is called but does not pass the caller's filters (e.g. when a variant has a low genotype quality), and third status (Not called) is when a variant is missed completely by the caller. Based on the variant status in the two callers, there are five scenarios and each callset is composed of variants from one or more scenarios.

One benefit of organizing variants in these callsets is to test various levels of stringency. For example, the callset named 'Any PASS' includes variants from all five scenarios regardless of the variant status. On the other hand, the callset named "both PASS" includes only variants that pass the called and pass the filters of both callers. These different levels of stringency allowed some callsets to have more variants than other and thus reflected different levels sensitivity and specificity. Moreover, I generated these callsets for both SNVs and INDELs separately (Table 2-7) since SNVs are called by GATK (G) and Samtools (S) while INDELs are called by GATK (G) and Dindel (D).

To decide which callset has the most desirable properties, I measured three different ratios. First, I used the Ts/Tv ratio for the SNVs the expected values ranges between (2.9-3.3) based on different sequencing projects at the Wellcome Trust Sanger Institute and 1000 genomes consortium. For INDELs, I used the coding in-frame/frameshift (n3/nn3) ratio, which was expected to be above 1 where the premise is coding frameshift variants are under much stronger negative selection. The third ratio I used was the rare/common ratio for both SNVs and INDELs (rare variants are defined as MAF < 1%).

Table 2-6 The criteria of choosing different variant callsets in order to determine the closest set to the truth measurements (Ts/Tv, n3/nn3 and rare/common ratios).

Scenarios	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**
Callset name	Caller A: PASS	Caller A: PASS	Caller A: Non PASS	Caller A: PASS	Caller A: Not Called
	Caller B: PASS	Caller B: Non PASS	Caller B: PASS	Caller B: Not Called	Caller B: PASS
Both PASS	Yes	-	-	-	-
Any PASS	Yes	Yes	Yes	Yes	Yes
Priority PASS (single Caller)	Yes	Yes	-	Yes	-
Any PASS (stringent)	Yes	Yes	Yes	-	-
Priority PASS (stringent)	Yes	Yes	-	-	-
Priority PASS (plus)	Yes	Yes	-	Yes	Yes
No Conflicts	Yes	-	-	Yes	Yes

The total number of SNVs varies between the callsets (Figure 2-14-A). The variation in coding SNVs was observed in the Ts/Tv ratio as well as rare/common ratio (Figure 2-14-B and C). As expected, the most stringent callset (bothPASS), that includes a variant only if it is called by both callers (GATK and SamTools) and passes both of their filters (i.e. PASS), has the highest Ts/Tv ratio (~3.18) while (anyPass) callset has the lowest Ts/Tv ratio (~3.01).

Table 2-7 A list of callsets in each call set based on the caller and if the pass the caller's internal filters (i.e. PASS).

SNVs		INDELs	
Callset Name	**Callset included**	**Callset Name**	**Callset included**
Both PASS	GS	Both PASS	DS
Any PASS	GS, Gs, gS, G., .S	Any PASS	DS, Ds, dS, D., .S
G Priority PASS	GS, Gs, G.	D Priority PASS	DS, Ds, D.
S Priority PASS	GS, gS, .S	S Priority PASS	DS, dS, .S
Any PASS (stringent)	GS, Gs, 'gS'	Any PASS stringent	DS, Ds, 'dS'
G Priority PASS (stringent)	GS, Gs	D Priority PASS (stringent)	DS, Ds
S Priority PASS (stringent)	GS, gS	S Priority PASS (stringent)	DS, dS
G Priority PASS (plus)	GS, Gs, G., .S	D Priority PASS (plus)	DS, Ds, D., .S
S Priority PASS (plus)	GS, gS, G., .S	S Priority PASS (plus)	DS, dS, D., .S
No Conflicts	GS, G., .S	No Conflicts	DS, D., .S

Keys: A single letter denotes each caller. For example "G" denotes GATK, "S" for Samtools and "D" for Dindel. Capital letter means the variant is a PASS (i.e. passed the caller internal filters) and a small letter if does not pass. The "." means the variant was not called by the caller. As an example, the callset named "G Priority PASS (stringent)" under SNVs includes two types of variants (GS) and (Gs). The (GS) is all variants that are called as PASS in both GATK and Samtools while (Gs) includes all variants that are called by GATK as PASS but called as non-PASS by Samtools.

On the other hand, the rare/common ratio of loss-of-function (or functional variant) shows the opposite trend; "bothPass" callset has the lowest rare/common ratio (~0.09) and "anyPass" showed the highest (~0.15). The benefit of using rare/common ratio is that it can tell us if a certain callset is enriched for rare variant more than expected. Since single-sample variant callers are not aware of the variant frequencies (i.e. whether it is common or rare) one would not expect the callers to be biased towards either rare or common variants. However, the variants called by Samtools seem to be enriched for rare variants mainly in three callsets that use Samtools as the dominant caller (S_Priority, S_ PriorityPASSplus and S_PriorityPASSstringent). What is even more interesting is that the Ts/Tv and rare/common ratios are inversely correlated (Figure 2-14-D). The higher Ts/Tv ratio gets, the lower the rare/common ratio becomes. Additionally, this correlation is also seen in other classes of variants such as functional (missense), silent (synonymous) and intronic variant (data not shown).

Similarly for INDELs, I examined different callsets derived from two callers, Dindel and Samtools (Table 2-6 and Table 2-7). The truth measurement I used for INDELs includes coding in-frame/frameshift (n3/nn3) and the rare/common ratios. Not surprisingly, the most stringent callset is "bothPASS" which includes INDELs that are called both callers and pass their internal filters. This callset performs well on both matrices (the n3/3nn ratio is ~1.66 and the rare/common ratio is ~0.10, see Figure 2-15 A-C). Here again, we see inverse correlation between these two ratios as we saw between the Ts/Tv and rare/common in the SNVs (Figure 2-15-D).

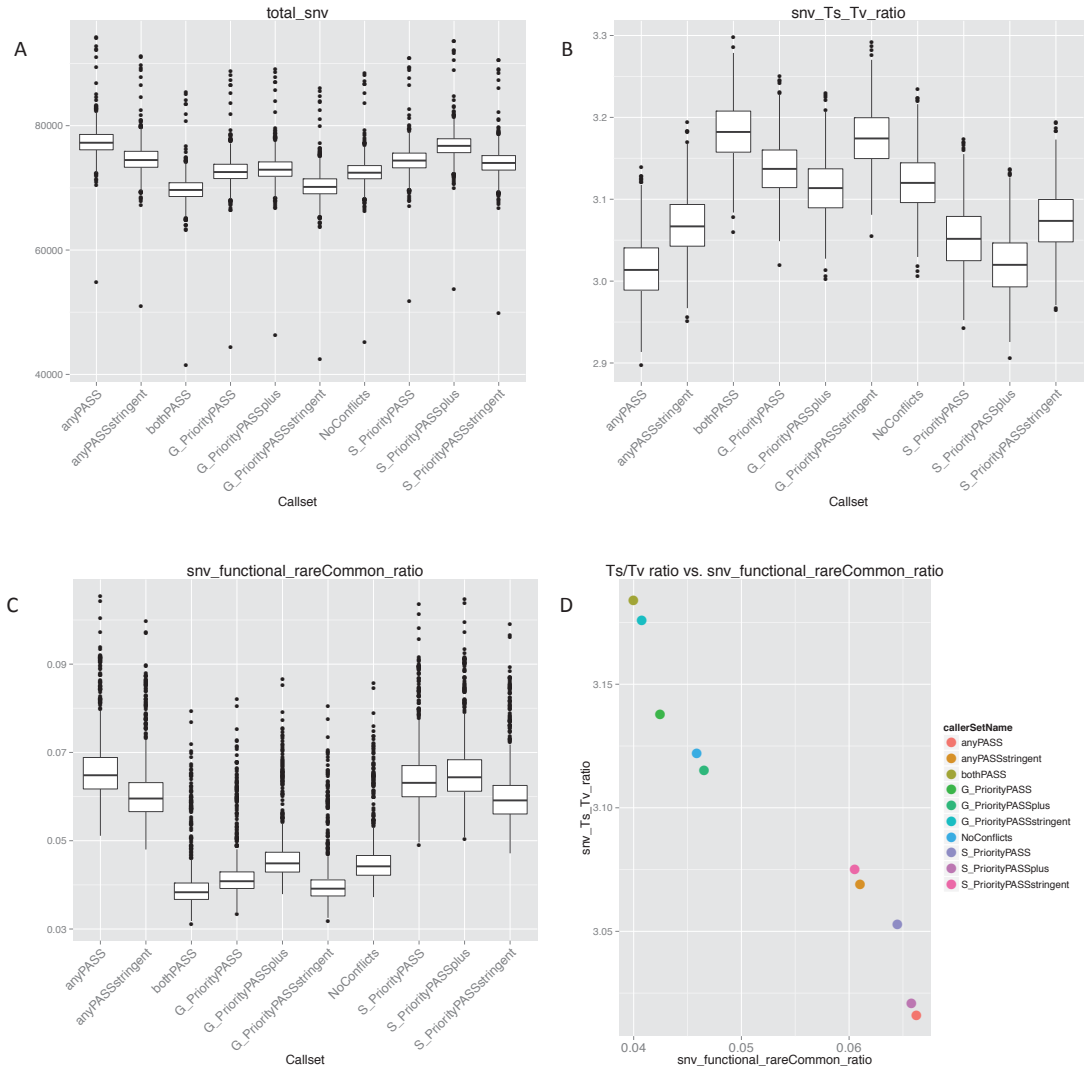


Figure 2-14 **Comparison of SNV callsets from GATK and Samtools.**
(A) Shows the total number of variants in each call set (n=960 samples) and most are comparable. (B) Ts/Tv ratios of functional variants (missense) SNVs per callset. (C) Rare/common ratios of functional variants (missense) SNVs per callset. (D) The relationship between Ts/Tv and rare/common ratios per callset.

Although these analyses were very informative, they were not enough to determine which caller contributed the most to the false positive rate (in terms of low Ts/Tv, n3/nn3 and / or rare/common ratios). The final piece of information was obtained by dissecting each callset to its basic five scenarios as defined in (Table 2-6). For example, SNVs variants can be grouped into five groups (GS, Gs, gS, G. and S.). Similarly, for INDELs, there are five classes (DS, Ds, dS, D. and .S) (see Figure 2-16). This analysis shows that Samtools tends to call more rare variants (in both SNVs and INDELs) and generally performed worse than other callers.

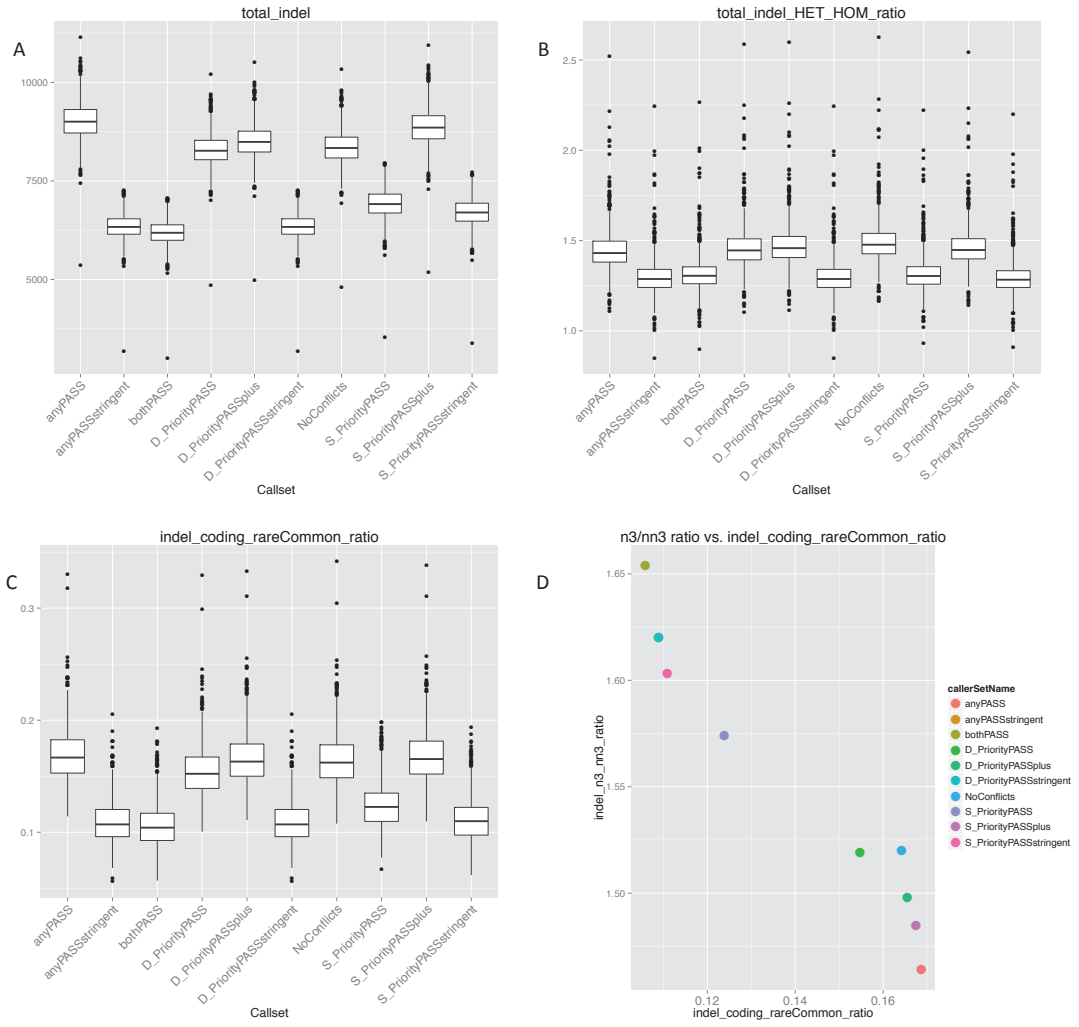


Figure 2-15 Comparison of INDEL callsets from Dindel and Samtools callers.
(A) Shows the total number of variants in each call set (n=960 samples) and most are comparable. (B) In-frame/frameshift (n3/nn3) ratios of coding INDEL variants per callset. (C) Rare/common ratios of coding INDEL variants per callset. (D) The relationship between n3/nn3 and rare/common ratios per callset.

This has a very important consequence on the downstream analysis since, on average, Samtools contributes 2.5 rare loss-of-function SNVs, four rare missense and two rare coding INDELs per sample. These might seem small for the number of candidates in one sample, but in a project with 100 or 1000 samples, this has a tremendous effect on the number of candidate variants needed to be validated or sent for functional studies.

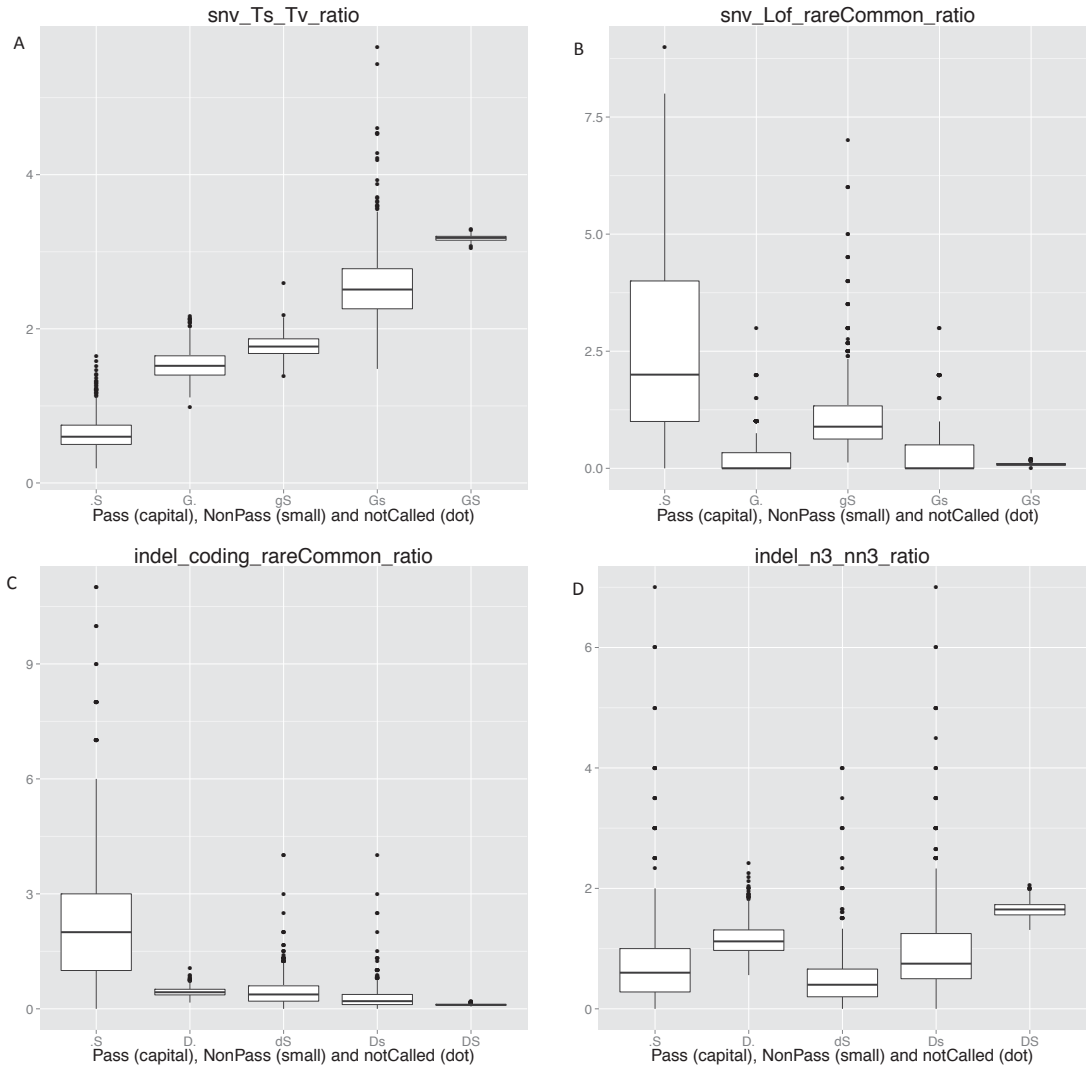


Figure 2-16 **Comparing callsets by callers.**
(A) Ts/Tv ratio of functional (missense) SNVs. (B) Rare/common ratio of loss-of-function SNVs (includes stop gain and variants that disturb the acceptor or donor splice sites). (C) Rare/common ratio of coding INDELs. (D) In-frame/frameshift (n3/nn3) ratio for coding indels. A single letter denotes each caller: "G" denotes GATK, "S" for Samtools and "D" for Dindel. Capital letter means the variant is a PASS (i.e. passed the caller internal filters) and a small letter if does not pass. The "." means the variant was not called by the caller. As an example, the callset named "G Priority PASS (stringent)" under SNVs includes two types of variants (GS) and (Gs). The (GS) is all variants that are called as PASS in both GATK and Samtools while (Gs) includes all variants that are called by GATK as PASS but called as non-PASS by Samtools.

Collectively, these results suggested the importance of discarding or flagging the rare coding variants called by Samtools alone (both SNVs and INDELs) in order to decrease the false positive rare candidate variants. It is important to notice that these observations are true for the specific older version of Samtools and for the filters used in the pipeline and may change accordingly.

2.3.2.3 Sample and data quality control tests

Before obtaining a set of high quality DNA variants for any downstream analysis, several tests are required to detect any quality issues such as contamination, sample swapping or failed sequencing experiments at the level of DNA samples, sequence data (BAM files) and called variants (VCF files).

DNA sample quality tests

The sample logistic team at the Wellcome Trust Sanger Institute tested the DNA quality of each sample using an electrophoretic gel to exclude samples with degraded DNA. The team also tested DNA volume and concentration using PicoGreen assay [277] to make sure every sample met the minimum requirements of exome sequencing. Additionally, 26 autosomal and four sex chromosomes SNPs were genotyped as part of the iPLEX assay from Sequenom (USA). This test helps to determine the gender discrepancies or possible contamination issues. Occasionally, the relatedness between sample and the family membership may need to be tested using the genotype of SNPs in iPLEX assay from the sample sequence data. An example of relatedness test from sequence data is discussed in chapter 3 (part of a replication study of 250 trios with tetralogy of Fallot).

Sequence data quality tests

The second group of quality tests was performed on the sequence reads generated by the next-generation sequencing platform. Carol Scott from the Genome Analysis Production Informatics (GAPI) team performed these tests to detect samples with low sequence coverage.

Variant quality tests

The third group of quality control tests targets the called variants that are stored in the Variant Call Format (VCF) files [161]. The aim of these tests is to detect the outlier samples based on the counts of single nucleotide variants (SNV) or insertion/ deletion variants (INDEL) in comparison to other published and / or internal projects (Figure 2-17 for SNV and for Figure 2-18 for INDEL variants). These plots are based on 94 CHD samples generated by GAPI pipeline and these plots are generated for each CHD project in chapter 3 and 4. These serve to monitor the consistency of variant calling between samples from the same project and also between different projects. Samples that show extreme low or high values above 2-3 standard deviations of the mean values are flagged for further investigations to determine the possible causes (e.g. contamination issues, poor sequence data, etc.)

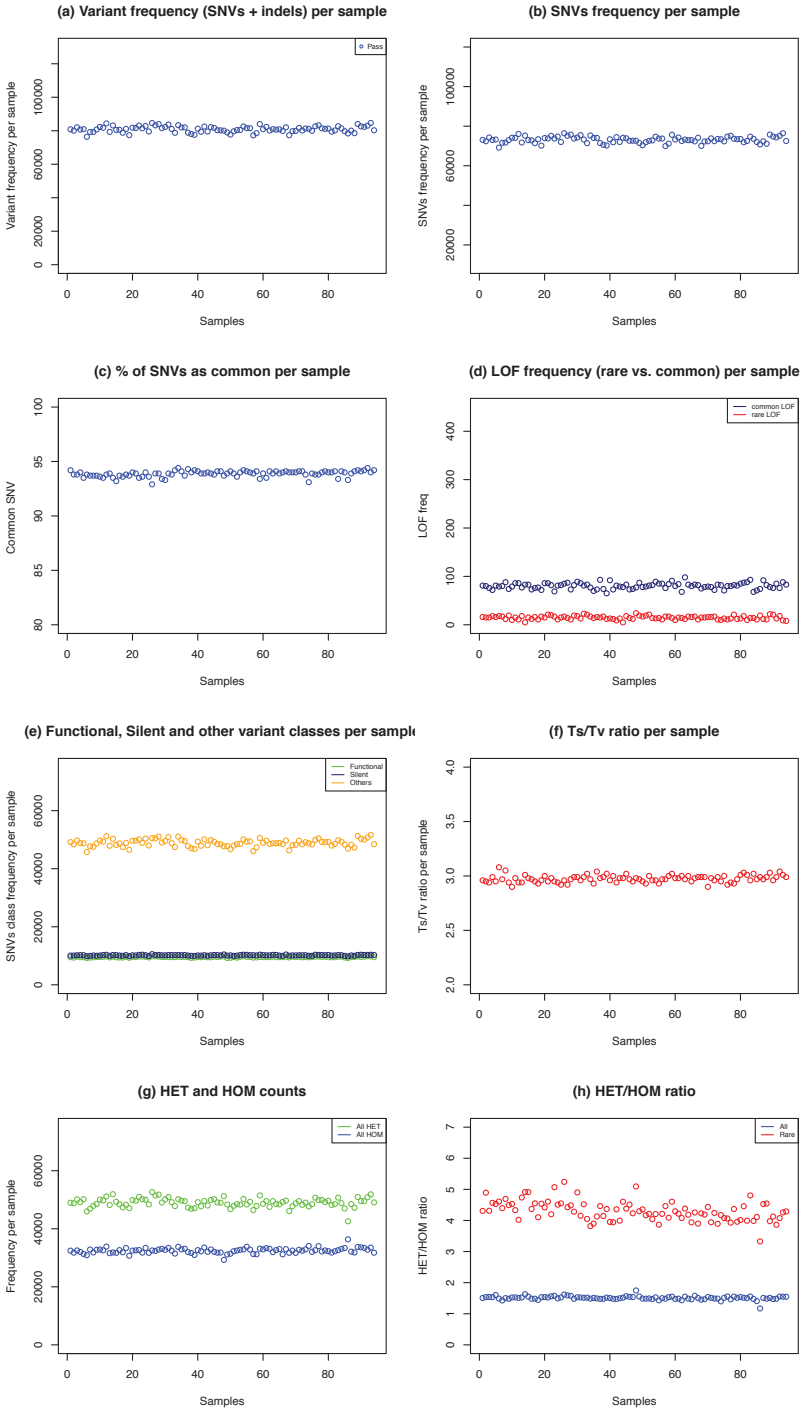


Figure 2-17 An example of QC plots I routinely generate for all samples in each study.
Variant counts per sample (n=94 selected CHD samples). (a) Total number of variants, both SNVs and INDELs, that pass caller internal filters (i.e. PASS). (b) Total number of single nucleotide variants only. (c) Percentage of common variants (MAF ≥ 1% in 1000 genomes project). (d) Number of rare and common loss-of-function (includes stop gain and variants that disturb the acceptor or donor splice sites). (e) Number of functional (missense), silent (synonymous) or others (include non-coding variants such as intronic and variants in untranslated regions, UTR). (f) Transition/transversion ratio of coding SNVs. (g) Count of heterozygous and homozygous variants. (h) Homozygous/heterozygous ratio of all or rare variants.

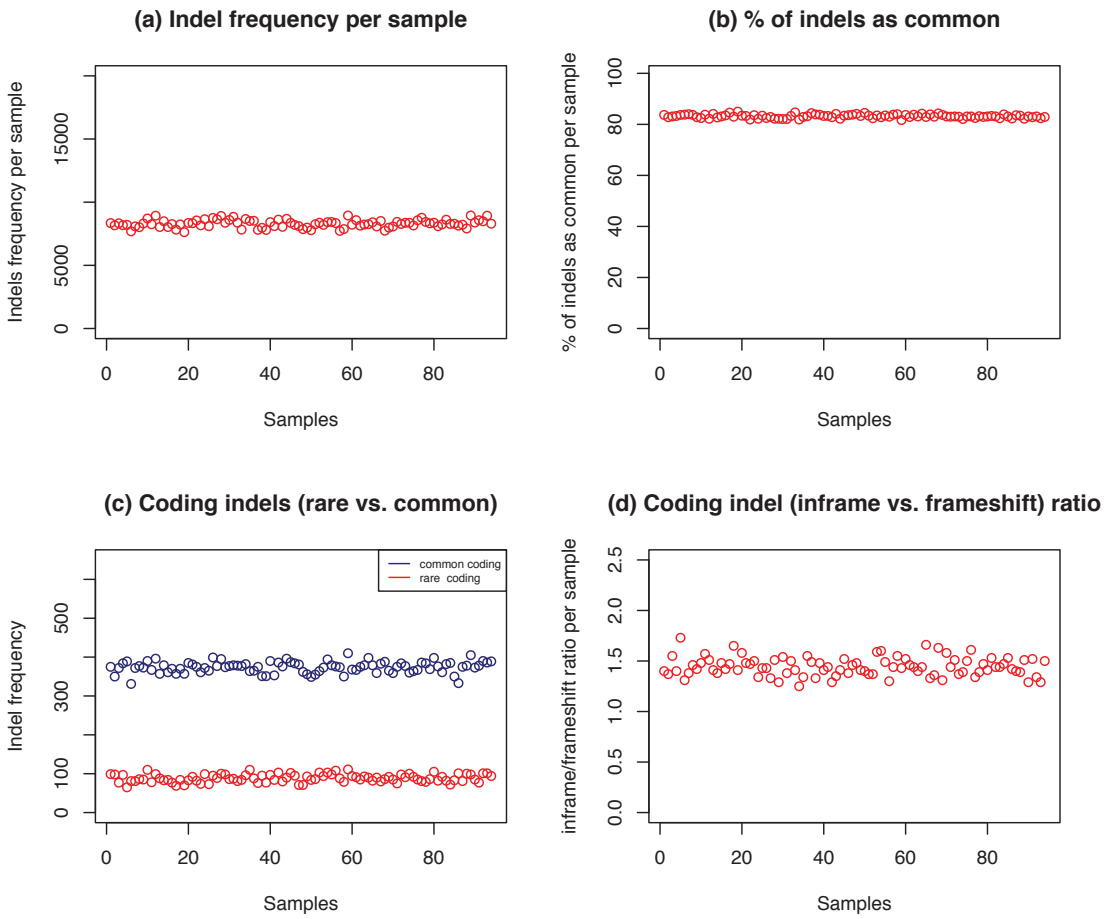


Figure 2-18 Count of INDEL variants per sample (n=94 selected CHD samples). (a) Total number of INDELs that pass caller internal filters (i.e. PASS). (b) Percentage of common variants (MAF ≥ 1% in 1000 genomes project). (c) Number of rare and common INDELs. (d) Coding In-frame/frameshift ratio (n3/nn3).

2.3.3 Minimizing the search space for causal variants

2.3.3.1 Minor allele frequency

In this thesis I have assumed that highly penetrant genetic causes of CHD are rare in the population given the fact that CHD affects usually less than 1% of the population and highly penetrant alleles should be strongly selected against. This makes annotating variants in CHD samples with allele frequency in matching population highly important for downstream analyses such as the family-based co-segregation, case/control and many other analyses. In this section, I describe the different resources of population allele frequencies that I used and their effect on the final number of rare candidate variants.

It is generally accepted that rare variants are defined as the variants with a minor allele frequency of 1% or less [278]. Currently, there are three major projects from which the allele frequency is available in a large number of samples. The first is the 1000 genomes that include 1,092 samples from different populations and used low-depth whole genome sequencing and high-depth whole exome sequencing [155]. The second is the NHLBI Exome Sequencing Project and includes 6,015 individuals of European American and African American ancestry and uses high-depth whole exome sequencing [199]. The third MAF resource is the UK10K cohort of low-depth whole genome sequencing from ~4,000 individuals of European ancestry [264]. While the individuals from the 1000 genomes and UK10K Cohort are presumably healthy, the NHLBI Exome Sequencing Project includes affected patients with various different phenotypes. This led me to disregard the MAF from NHLBI-ESP samples since I cannot rule out the possibility that some samples may have congenital heart defects. Additionally, the captured exome data in NHLBI-ESP project is based on a smaller set of genes (~17,000 genes compared with ~20,000 genes captured in the exome data in my samples), which can adversely affect many downstream analyses such as the case/control analysis by generating spurious false positive signals.

In addition to publicly available MAF resources, I generated an internal MAF based on 576 healthy parents from the Deciphering Developmental Disorders (DDD) project. The main goal of using the internal MAF is to exclude variants that appear as rare according to population MAF recourse but appear in > 1% of the samples. These are expected to be novel 'common' variants or, possibly more likely, sequencing / pipeline errors.

At the time of writing this thesis, there was no general consensus on the best strategy to match the exome sequence variants with variants in population frequency resources, especially the indels, in our internal pipelines (GAPI and UK10K) nor in other external sequencing centers like the Broad institute in the USA (Shane McCarthy, personal communication). Some groups match variants in

their projects with MAF from public resources if both have the same chromosome and position only while others expand this matching strategy by matching variants in a window of 10-30bp to the closest variant.

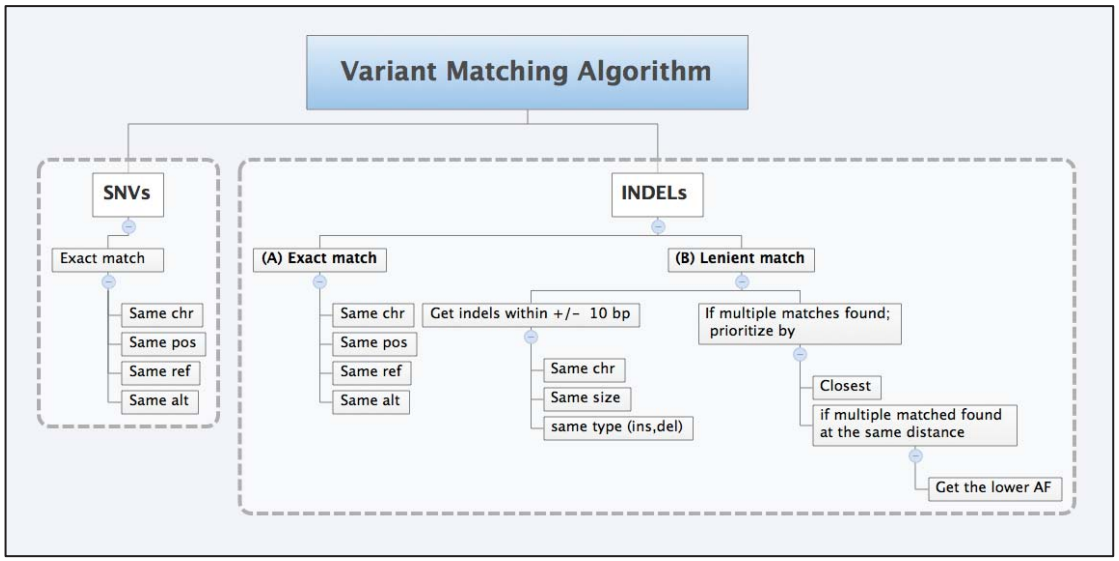


Figure 2-19 The variant matching algorithm between alleles in exome data and alleles from MAF resources.

I designed a hierarchical algorithm that matches between the source files (UK10K, 1KG and ESP) and the target files (CHD samples or other samples like DDD) (see Figure 2-19). The goal of this algorithm is to make sure I match the right allele in my CHD samples with the corresponding alleles in the MAF resources. This algorithm generates two keys; one from the source file (e.g. CHD sample) and the second key is generated from the target file (1000 genomes MAF file) and then tests if both keys match each other (see Figure 2-20 for examples).

In the case of SNVs, I constructed the key using four values (chromosome, position, reference allele and alternative allele) and called this an "exact I" matching. On other hand, INDELs are harder to annotate because callers might call the INDEL alleles differently especially in repeat regions. To accommodate these different scenarios, I tested three different matching definitions. The first is "exact I" which is similar to the SNVs and is considered the most stringent approach. The second strategy is called "exact II" where I construct a key, also using four values (chromosome, position, slice and direction). This key requires

both INDELs in the target and source files to be at the same locus (chromosome and position) while 'slice' is computed based on the DNA sequence difference between the reference and alternative alleles and 'direction' is either deletion or insertion. Although "exact II" matching may look different to "exact I", it is also a stringent matching that tries to accommodate the differences imposed by different callers when they call the same INDEL.

When a matching algorithm fails to find any results using "exact I or II" strategies, it switches to a lenient matching mode where it expands the search for similar INDELs within 10-30bp flanking window. If the algorithm finds more than one INDEL that meet its criteria, it chooses the nearest matching INDEL to the target locus and if it finds multiple INDELs at the same distance, it picks the one with lower MAF value, to be conservative.

(A)

Chr	Pos	Ref	Alt	Direction	Slice
chr1	866511	CCCCT	CCCCTCCCT	ins	CCCT

(B)

Retrieve all variants in 1KG at region chr1: 866511 +/- 10bp flanking region

Chr	Pos	Ref	Alt	Direction	Slice	AF
chr1	866505	G	A	-	-	0.00023
chr1	866511	CCCCT	C	del	CCCT	0.00176
chr1	866511	CCCCT	CCCCTCCCT	ins	CCCT	0.58993
chr1	866517	C	G	-	-	0.00431

Figure 2-20 Example of how MAF matching algorithm works. (A) The chromosome (Chr), position (Pos), reference (Ref) and alternative (Alt) alleles from a source file (e.g. VCF file of a CHD sample). (B) Possible matching alleles within + 10bp flanking region extracted from the MAF resource file from 1000 genomes project. The direction of the allele can be either insertion or deletion in case of INDELs and '-' for SNVs (i.e. point mutation). 'Slice' (red) is the DNA sequence difference between reference and alternative alleles and computed for INDELs only. In this example, since the VCF file contain an INDEL, the matching algorithm will try to look for "exact I" matching key (same chromosome, position, reference and alternative alleles). If this failed, it will start matching using "exact II" strategy (i.e. same chromosome, position, direction and slice), which corresponds to the third record in the (B) where the allele frequency is (0.58993) in the 1000 genomes.

To test the algorithm performance under each mode (exact I, II and lenient), I tested the correlation between three MAF resources (1KG, UK10K and ESP) with DDD internal MAF described above (

Table 2-8). My assumption is that the vast majority of variants should have similar allele frequency in the DDD samples as in the three MAF recourses (except for private or extremely rare variants and sequence errors). A proper matching algorithm should be able to match same alleles and thus the MAF values should show a strong correlation between the DDD samples and the other MAF resources. Both exact I and exact II strategies show a strong correlation between the allele frequencies in 1KG, UK10K or ESP with DDD internal allele frequencies (correlation coefficient > 0.8) but not the lenient strategy for declaring a match (correlation coefficient -0.03 to 0.008).

After I showed that both 'exact I and II' algorithms are well suited for matching alleles in samples sequenced locally with alleles available in public resources, I decided to test the effect of using MAF from different resources on the number of rare coding variants per sample. To evaluate the effect of these MAF recourses, I selected 288 samples from DDD project and annotated them with allele frequency from four MAF recourses (1KG, UK10K, ESP and DDD's internal MAF) (Figure 2-21) in order to eliminate common variants (MAF > 1%). The number of variants left after excluding common variants based on MAF from the 1000 genomes project or the UK10K project was comparable (616 and 631 respectively). The MAF from ESP on the other hand do not appear to be very effective for filtering. This is not unexpected since the ESP sequence data are based on a smaller version of the exome compared with the whole genome data in the 1000 genomes and UK10K projects. However, using all three MAF resources together was more effective than using each separately (~428 rare variants per sample).

Table 2-8 Correlation values between "allele frequencies" of ~9,000 INDELs on chromosome 1 from DDD (n=576 samples) and the corresponding allele frequencies from three population-based projects: 1000 genomes, UK10K twins cohort (n=~4000), and ESP projects (n=~6500) using three matching strategies (exact I, II and lenient). 1KG: 1000 genomes, COHROT: UK10K twins cohort, ESP: NHLBI Exome Sequencing Project, cor=correlation coefficient.

Population-based Projects	Matching strategy		
	Exact type I	Exact type II	Lenient
1KG	0.80	0.83	-0.03
COHROT	0.92	0.73	0.01
ESP	0.89	0.88	0.01

Surprisingly, using the internal MAF from healthy parents in DDD project was even more effective than using all three public MAF together (~419 rare variants per sample when used alone and 327 when used in addition to the other three MAF resources). A possible explanation is that alleles with MAF > 1% and specific to a given project are likely to be sequence or pipeline errors, otherwise they would have been identified in large-scale projects such as the 1000 genomes, which aims to discover alleles with low allele frequency of at least 1% in the populations studied [155].

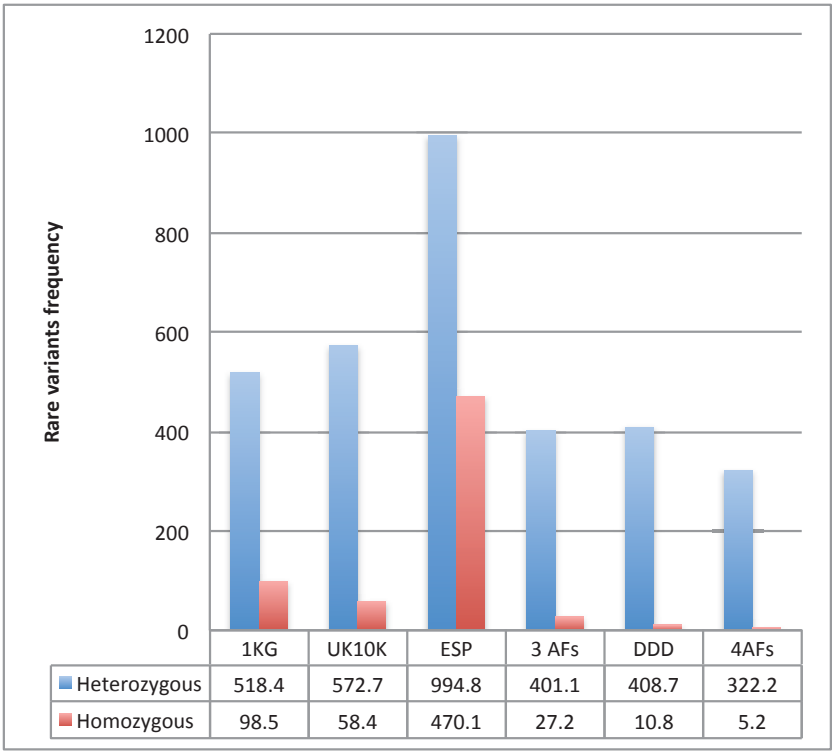


Figure 2-21 average number of autosomal rare variant when filtering based on < 1% minor allele frequencies from different resources. The data are based on 288 samples from the Deciphering Developmental Disorders (DDD) project. 1KG: 1000 genomes, UK10K: 4,000 healthy twins from UK10K cohort, ESP: 6,015 samples from NHLBI Exome Sequencing Project. 3 AFs includes rare

variants in (1KG, UK10K and ESP). DDD is an internal allele frequencies based on 576 healthy parents from DDD project. 4 AFs includes rare variants in 1KG, UK10K, ESP and DDD.

2.3.4 Family-based study designs in CHD

There are many family-based designs one can consider when studying CHD, such as singletons, affected sib-pairs, parent-offspring trios, affected parent-child and multiplex families. However, since the mode of inheritance in CHD is poorly understood in general, there is no obviously optimal study design.

Each design has advantages and disadvantages, for example, in terms of the feasibility of the sample collection and the availability of suitable analytical approaches (Table 2-9). Singletons (or index cases) are the easiest to collect but each sample has several hundreds of rare coding variants if analyzed separately, which makes the task of finding likely pathogenic variants difficult. On the other hand, trio family designs are usually more difficult to collect but they offer a chance to detect *de novo* and definitive compound heterozygous variants in the affected child, which are not feasible in singleton or affected-sib pair designs.

To see how different study designs may affect the final number of candidate genes, I selected one family of healthy parents and three affected children (two females and one male, Figure 2-22) to estimate the number of rare, functional coding variants under different designs and inheritance scenarios. Variants were defined as rare if they have a minor allele frequency < 1% in the 1000 genomes [155] and in 2,172 parents from the Deciphering Developmental Disorders (DDD) project [260] (this analysis was performed more recently with a newer version of the DDD project which include a larger number of parents compared with analysis described in previous sections where I included 576 parents only). Functional coding variants are defined as variants predicted by VEP tool [170] to be either loss of function (stop gain, frameshift or variants affected donor or acceptor splice sites) or functional (missense or stop lost). I excluded silent (synonymous) variants from the analysis.

Table 2-9 Overview of study designs and analytical approaches

Study Design	Advantages	Disadvantages	Analytical approaches
Index cases	-Easy to collect	-Lack of family genotype information means larger search space for causal variant(s).	-Case/control (collapsed, weighted, etc.)
Extended families	- Co-segregated variants that are absent from control provide strong evidence for causality.	- Rare to find and collect samples.	- Linkage analysis and then targeted sequencing.
Trios	- Utilize parental genotype to detected *de novo* variants - Compound heterozygous mutations can be detected - Avoid population stratification bias (e.g. TDT tests)	- More difficult to collect	- *De novo* - Co-segregation - Transmission disequilibrium test (TDT)
Affected-sib pairs	- Suggestive of autosomal recessive disorders. - Small search space due to few autosomal recessive candidates and siblings share only half of the variants.	- The lack of parental genotype information inflates the number of homozygous variant candidates.	- Runs of homozygosity - Co-segregation - Identical By Decent (IBD) analysis (Autozygosity) - Identical by State (IBS) analysis (Allozygous)
Multiplex families **(parents plus > 1 affected child)**	-Combine the power both trios and affected sib-pairs - Smaller search space for variant with more affected children.	- Difficult to analyze when affected members have heterogeneous phenotypes - Less common families than the trios.	Same as trios in addition to the affected sib-pairs
Affected parent-child	- Suggestive of autosomal dominant disorders.	- The variant search space is larger than in trios.	- Co-segregation of heterozygous variants

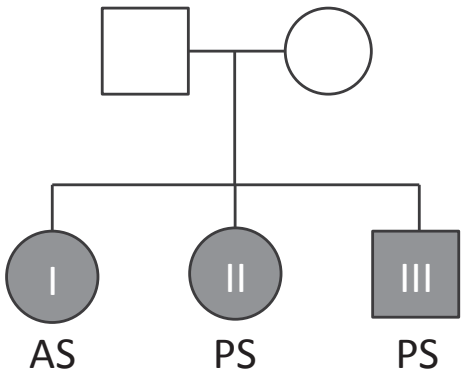


Figure 2-22 Pedigree chart of a multiplex family (three affected children and their healthy parents) used to count the number of candidate genes with rare coding under different inheritance scenarios.
AS: aortic stenosis, PS: pulmonary stenosis.

Initially, I analyzed each affected child separately to test the singleton design and found two rare coding homozygous, 10 compound heterozygous and 381 heterozygous variants on average (Table 2-10). If I consider two children as an affected sib-pair and look for shared rare coding variants, the number of rare coding heterozygous variants drops to less than half and less than a quarter of recessive variants (both homozygous and compound heterozygous) compared with the singleton design. Combining all three affected sibs at the same time shows only 75 rare coding heterozygous variants shared among them, which represents 80% less than singleton and 50% less than two affected sib-pairs but no recessive variants are shared between all three sibs.

On the other hand, the number of candidate genes with rare coding variants drops dramatically to just a handful of genes in the trio design when I consider the parents and assume complete penetrance. This is mainly because the parents' exome data provides additional genotype information to exclude most heterozygous variants (see Table 2-11 for details).

These empirical numbers of rare coding variants shared between different family members are in general agreement with what I would predict from Mendelian inheritance. For example, since the number of rare coding heterozygous variants observed in each child is ~381 on average, two affected sibs should share 50% (IBD=1) or 190 variants which is not far from what I observed in the three affected sib-pairs in this family (~153). Similarly for the rare coding homozygous variants, the observed average in each child is ~10 and each sib-pair is expected to share 25% (IBD=2) or 2.5 homozygous variants, which is very close to the observed value (~2.67).

The variation between the observed and the expected numbers of shared variants under Mendelian inheritance laws is likely caused by under-calling the same variant in one more member. I found the same broad agreement between the average numbers of variants in the affected parent-child pairs (~157) compared with the expected numbers under Mendielian inheritance laws (~190).

Table 2-10 Number of rare coding variants in affected children under different study designs (see family pedigree Figure 2-22).
Singleton: each affected case is analyzed independently. Affected sib-pairs: shared variants between two or more affected sibs without parental information. Trios: each child is analyzed with his/her healthy parents and assuming complete penetrance (see Table 2-11 for the full list of allowed genotypes). Multiple: analysis of two or more children with their healthy parents and assuming complete penetrance.
* Indicates the average number of one affected parent (father or mother) and any child of the three. NA: not applicable (e.g. no autosomal recessive variants are allowed in affected parent-child design).

Family study design	**Samples**	**Number of candidate genes with rare coding variants**		
		Recessive (homozygous)	**Recessive (compound)**	**Dominant**
Singleton	Child I	1	11	373
	Child II	1	12	413
	Child III	4	8	357
Affected sibs	Shared between sibs (I and II)	0	5	162
	Shared between sibs (I and III)	1	1	171
	Shared between sibs (II and III)	0	2	126
	Shared between sibs (I, II and III)	0	0	75
Affected parent-child	One affected parent and one affected child	NA	NA	157*
	One affected parent and two affected children	NA	NA	74*
	One affected parent and three affected children	NA	NA	37*
Trios	Trio (child I)	0	3	1
	Trio (child II)	0	5	0
	Trio (child III)	0	5	0
Multiplex	Shared between trios (I and II)	0	4	0
	Shared between trios (I and III)	0	0	0
	Shared between trios (II and III)	0	4	0
	Shared between trios (I, II, III)	0	0	0

Finally, I consider the shared rare coding variants between two or more trios (i.e. multiplex family design). This study design has identified four genes only with compound heterozygous that are shared between child-I and child-II and another four genes between child-II and child-III. No rare coding variants were detected when all three sibs and their parents were analysed at the same time. This may suggest either a possible under-calling of a monogenic variant (i.e. missed by the callers) or an oligogenic nature of the disease (i.e. multiple genes with different rare causal variants). Nonetheless, the trio design is clearly superior to the affected-sib pairs or singleton designs since it identifies very small number of candidate genes.

Table 2-11 The accepted genotype combinations in a complete trio are the genotypes that are compatible with Mendialin inherence laws and also in agreement with the assumption of complete penetrance. Each trio includes an affected child (male or female) and two healthy parents. Each cell in the first column "genotype combinations" represents three genotypes in child, mother and father. "0" indicates a homozygous reference genotype, "1" is a heterozygous genotype, and "2" is a homozygous genotype in diploid chromosome (autosomal) or hemizygous in a haploid chromosome (e.g. X-chromosome in a male child). Y-chromosome and mitochondrial DNA are omitted from the table. Empty cells indicate that a given genotype combination is incompatible with Mendelian laws (e.g. 1,0,0 is *de novo*) or not expected under complete penetrance assumption (e.g. 1,1,1 is heterozygous in both the affected child and his parents). Only three genotype combinations were considered when I performed trios or multiplex analysis.

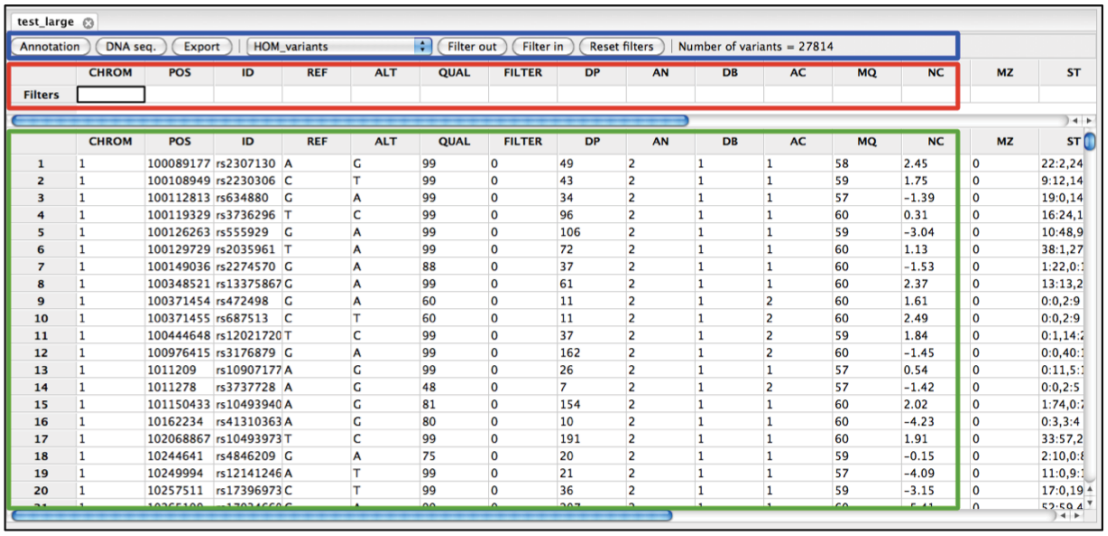
Genotype combinations	**Autosomal**	**X- chromosome in an affected male child**	**X- chromosome in an affected female child**
(1, 0, 0)			
(1, 0, 1)			
(1, 0, 2)			
(1, 1, 0)			
(1, 1, 1)			
(1, 1, 2)			
(1, 2, 0)			
(1, 2, 1)			
(1, 2, 2)			
(2, 0, 0)			
(2, 0, 1)			
(2, 0, 2)			
(2, 1, 0)		Hemizygous inherited from a carrier mother	
(2, 1, 1)	Homozygous in child and inherited from carrier parents		
(2, 1, 2)			
(2, 2, 0)			
(2, 2, 1)			
(2, 2, 2)			
(1,0,1) and (1,1,0)	Compound heterozygous in the child in a given gene		

2.3.5 Family-based Exome Variant Analysis (FEVA) suite

To generate a list of candidate genes from exome data of a given rare, putatively monogenic, disorder, one needs to go through multiple steps that include excluding low quality variants based on various filters, excluding incompatible genotype combinations with either the study design or the plausible inheritance models (see Table 2-11 for an example of incompatible genotypes with a trio design) and filtering common variants (MAF > 1%) as well as non-coding variants since rare coding variants (except silent) are more likely to have a measurable effect on the phenotype. Performing these steps manually in non-specialized software, such as Microsoft Excel, is time consuming and error prone due to the large number of variants. This is clearly not suitable for large-scale projects of hundreds of samples with different family structures.

To automate the analysis and variant reporting under different Mendelian inheritance models I designed a 'Family-based Exome Variant Analysis' tool. FEVA is a suite of tools that enable users to generate a list of candidate genes under various study designs. FEVA offers two interfaces for the end user. The first interface is a Command Line Interface (CLI) suitable for high-throughput analysis, which can be incorporated into automated data analysis pipelines. The second interface is a graphical user interface (GUI) aimed for low-throughput analysis that is easy to use with minimal training (Figure 2-23). I designed the GUI version of FEVA three years ago when many sequencing projects, such as the UK10K RARE project, was just starting at the Wellcome Trust Sanger Institute. At that time, there was no GUI available for our collaborators to explore variants files (VCF files) with ease. I coded most FEVA components in the Python programing language, which I chose for its readability and agility for prototyping. Since Python is a high-level programming language; it can be slow when performing computer intensive tasks (such as parsing large files which are commonly used in the next-generation sequencing era). However, Python is easily extendable by other low-level statically typed, and thus quite fast, programming languages to overcome this limitation. For example, I have used many C and C++ libraries to parse large exome/genome files. Moreover, I used

graphical user interface components, which are written in C++ (QT library) for fast viewing.



	CHROM	POS	ID	REF	ALT	QUAL	FILTER	DP	AN	DB	AC	MQ	NC	MZ	ST
1	1	100089177	rs2307130	A	G	99	0	49	2	1	1	58	2.45	0	22:2,24
2	1	100108949	rs2230306	C	T	99	0	43	2	1	1	59	1.75	0	9:12,14
3	1	100112813	rs634880	G	A	99	0	34	2	1	1	57	-1.39	0	19:0,14
4	1	100119329	rs3736296	T	C	99	0	96	2	1	1	60	0.31	0	16:24,1
5	1	100126263	rs555929	G	A	99	0	106	2	1	1	59	-3.04	0	10:48,9
6	1	100129729	rs2035961	T	A	99	0	72	2	1	1	60	1.13	0	38:1,27
7	1	100149036	rs2274570	G	A	88	0	37	2	1	1	60	-1.53	0	1:22,0:
8	1	100348521	rs13375867	G	A	99	0	61	2	1	1	60	2.37	0	13:13,2
9	1	100371454	rs472498	G	A	60	0	11	2	1	2	60	1.61	0	0:0,2:9
10	1	100371455	rs687513	C	T	60	0	11	2	1	2	60	2.49	0	0:0,2:9
11	1	100444648	rs12021720	T	C	99	0	37	2	1	2	59	1.84	0	0:1,14:2
12	1	100976415	rs3176879	G	A	99	0	162	2	1	2	60	-1.45	0	0:0,40:1
13	1	1011209	rs10907177	A	G	99	0	26	2	1	1	57	0.54	0	0:11,5:1
14	1	1011278	rs3737728	A	G	48	0	7	2	1	2	57	-1.42	0	0:0,2:5
15	1	101150433	rs10493940	A	G	81	0	154	2	1	1	60	2.02	0	1:74,0:2
16	1	10162234	rs41310363	A	G	80	0	10	2	1	1	60	-4.23	0	0:3,3:4
17	1	102068867	rs10493973	T	C	99	0	191	2	1	1	60	1.91	0	33:57,2
18	1	10244641	rs4846209	G	A	75	0	20	2	1	1	59	-0.15	0	2:10,0:8
19	1	10249994	rs12141246	A	T	99	0	21	2	1	1	57	-4.09	0	11:0,9:1
20	1	10257511	rs17396973	C	T	99	0	36	2	1	1	59	-3.15	0	17:0,19

Figure 2-23 **Screen print of FEVA graphical user interface (GUI).**
This simple interface shows three parts. The green rectangle shows a list of variants and their annotations. Each row represents one variant along with its quality scores and biological information such as gene, variant type, effect on protein, etc. The red rectangle is where the user can enter filter conditions to exclude or include rows. The blue rectangle includes additional functions such as applying a set of pre-defined filters or to export a list of candidate variants to other programs.

Although other tools have been published during my work with similar functionality, such as SVA, EVA and VarSift [261-263], none of them were able to fulfill the needs for my projects. One limitation common to these tools it that they are not suitable for both interactive and high-throughput analysis. Additionally, many of them have hard coded filters, and so lack flexibility, or require a certain formatting that is not necessarily compatible with the VCF files generated by the GAPI or UK10K pipelines (see Table 2-12 for comparisons with FEVA).

The family-based analyses in FEVA go through three steps (Figure 2-24): (1) reduce the search space by applying quality and MAF filters (e.g. exclude common variants, low quality, etc.), (2) identify co-segregating variants in family members (e.g. exclude variants in healthy sib or shared variants between affected parent-child), (3) Group the possibly pathogenic variants by the inheritance model (e.g. recessive or dominant).

Table 2-12 Comparison of four freely available graphical user interface applications for genome or exome analysis. N/A: not available.

Features	FEVA	EVA [262]	SVA [261]	VarSifter [263]
Desktop application	Yes	No	Yes	Yes
User custom annotation	Yes	No	Yes	No
Visualization	No	Basic	Advanced	No
Custom filters	Yes	Hard-coded	Hard-coded	Hard-coded
Whole genome	Yes	No	Yes	No
Accepts compressed files	Yes	No	N/A	No
Family Based analysis	Yes	Yes	No	No (Var-MD)
Memory usage (RAM)	Minimal	N/A	Large	N/A
QC statistics	External module	Yes	Yes	No
Has command-line tools	Yes	No	No	No
Input files	VCF	VCF	VCF & bco	VCF
Cross-platform	Yes	N/A	Yes	Yes

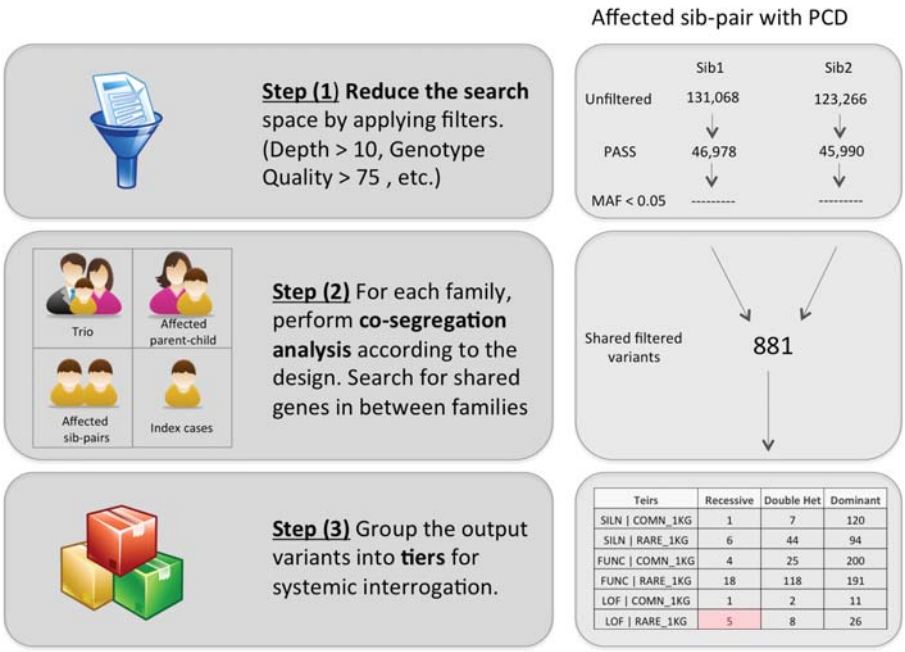


Figure 2-24 **FEVA workflow.**
An example of one sib-pair affected with Primary Ciliary Dyskinesia (PCD), which has been sequenced as part of the ciliopathies study in the UK10K RARE project. The user supplies the variants files and chooses which family design and FEVA performs three tasks automatically. First, FEVA excludes low quality variants and common variants using a MAF threshold supplied by the user. In the next step, FEVA applies the rules of co-segregation designed for affected sib pairs (i.e. shared variants in both sibs). Finally, FEVA groups shared variants under recessive (homozygous or compound heterozygous) and dominant models. Furthermore, FEVA can divide the candidate variants into loss-of-function and functional classes according to the user settings. Almost all steps described here are adjustable by the end user, which enable FEVA to accommodate different needs and scenarios.

The rules of co-segregation vary according to the family design (e.g. singleton, trio of healthy parent or trio of affected father-child, etc.) and can be made more or less stringent. These models are configurable by the user to suit a unique study design (only in the command-line version of FEVA). In the next section, I will describe how I used FEVA with different study designs to identify pathogenic and candidate pathogenic genes for different disorders.

2.3.6 Application of FEVA in rare disease studies

Application 1: Targeted sequencing of linkage regions (monogenic disease)

Dr. Andrew Crosby and his team at St. George's University of London have previously detailed the clinical features of members of a large UK family affected by dominantly transmitted distal hereditary motor neuropathy type VII (OMIM 158580). The team had previously mapped the gene responsible to chromosomal region 2q14 in a family of 14 affected and 12 unaffected members and I collaborated with them to analyze the exome sequence data of one affected family member.

Coding regions were captured with SureSelect All Exons (50 Mb) and sequenced by Illumina HiSeq at the Wellcome Trust Sanger Institute, yielding 9.8 Gb data (~130 million reads) corresponding to 91% target coverage with a mean depth of 1,073 and identifying 52,806 variants. Based on previous linkage analysis [279, 280], I used the FEVA software to report rare coding variants in two regions (~13.5 Mb) with high LOD scores (Table 2-13).

Table 2-13 **Genome coordinates of microsatellite marker**

Regions	Size	Marker ID	Locus in human genome
Region (1)	9.2Mb	AC084377	Chr2:99560750
		D2S160	Chr:2:112998734
Region (2)	4.3Mb	D2S2970	Chr2:118948333
		D2S2969	Ch2:123237183

After filtering common and non-coding variants (Table 2-14), I identified only one loss of function variant within the critical region; this was a single base deletion (c.1497delG) in *SLC5A7* gene encoding the Na^+/Cl^- dependent, high-affinity choline transporter. This novel variant was found to co-segregate in all affected members using capillary sequencing and this work was published in the American Journal of Human Genetics [281].

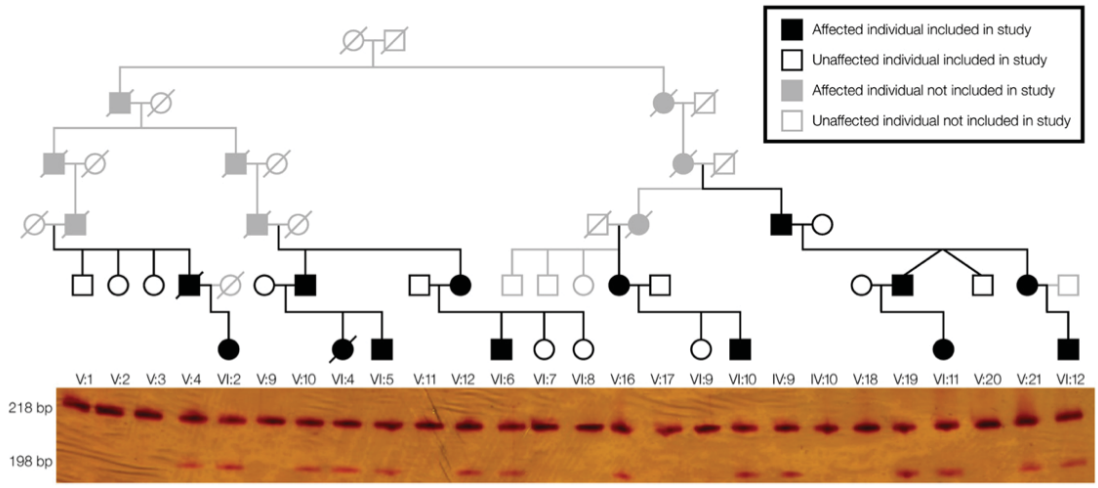


Figure 2-25 Family pedigree and c.1497delG cosegregation in *SLC5A7* gene [281]. The c.1497delG variant results in the creation of a novel *Ssp*I restriction site that facilitates cosegregation analysis by restriction digestion of exon 9 PCR products resolved by polyacrylamide gel electrophoresis. (Image and caption are adapted from [281])

Table 2-14 Number of variants in two linkage regions (~total size of 13.5 Mb). The variants are classified based on genotype (heterozygous or homozygous), by the predicted effect on protein to functional (missense) or loss-of-function (LOF class includes stop gain, frameshift and variants that disturb acceptor or donor splice sites). Only one rare LOF variant, a coding frameshift, found in *SLC5A7* gene that encodes for choline transporter protein.

Genotype	All variants	Common		Rare	
		Functional	LOF	Functional	LOF
Heterozygous	134	32	2	23	1
Homozygous	77	24	1	0	0

Similar to the analytical strategy I used to discover causal mutations in *SLC5A7* gene, I utilized FEVA to analyze data from other monogenic diseases under an autosomal recessive model in collaboration with Dr. Crosby and his team (Table 2-15). In all of these cases, I used the linkage analysis information to guide FEVA while filtering for rare coding homozygous or compound heterozygous variants.

These analyses were usually straightforward since FEVA reported only one or two candidate variants per sample because of the small linkage intervals.

Table 2-15 Results from other monogenic phenotypes where linkage analysis was used to guide the variant filtering of variants from whole exome or custom designed data (using FEVA).

Phenotype	Hereditary spastic paraplegia	Developmental delay with macrocephaly	Microlissencephaly
Mendelian model	Autosomal recessive	Autosomal recessive	Autosomal recessive
Linkage analysis	14.3Mb (chr12)	19q.13.32	2.36Mb (chr19)
Sequencing region	Custom design	Whole exome	Custom design
Number of samples	1	1	1
Candidate gene	*B4GALNT1*	*KPTN*	*WDR62*
Casual variant	c.1458insA	c.776C>T	c.1562T>A and c.4038-4039delAA
Project status	Published in [282]	Published in [283]	Manuscript is being prepared

Application 2: Affected trio families combined with candidate gene screening

The aim of the Deciphering Developmental Disorders (DDD) project is to collect DNA and clinical information from undiagnosed children in the UK with developmental disorders and their parents [260]. I used FEVA to test its performance in high-throughput on 1,080 trios of affected children with various developmental disorders and also to estimate the number of candidate genes, assuming healthy parents and complete penetrance of rare coding variants (Table 2-16).

FEVA was able to report rare coding variants according to the genotype rules in (Table 2-11) under autosomal recessive (homozygous or compound heterozygous) and X-linked models (separately for male and female children). The rare variants are defined as variants with MAF < 1% in the 1000 genomes project and in parental MAF from DDD (n=2,172). Regardless of gender, each child has, on average, four candidate genes with autosomal recessive rare coding

variants (excluding silent) and another three candidate genes on the X chromosome.

I also tested FEVA's ability to screen candidate genes for the presence of rare or novel coding variants (Table 2-16, DDG2P genes). DDG2P is a list of 1,148 manually curated genes with strong evidence supporting involvement in development disorders (the DDG2P gene list was developed by the DDD team). The screening analysis revealed, on average, only one autosomal rare coding variant, one X-linked in females and 0.18 X-linked in males. However, the DDD team implements additional filtering steps for their clinical reporting pipeline. These steps involve matching the phenotype and family history to the genotype (i.e. compatibility with the Mendelian rules), which lowers the number of candidate genes per child still further.

Table 2-16 Number of candidate variants in 1,080 affected DDD trios assuming healthy parents and complete penetrance (558 males and 522 females).
LOF: loss-of-function (include strop gain, variants disturbing acceptor or donor splice sites and frameshift), functional (includes missense). DDG2P: a list of 1,148 manually curated genes with strong evidence supporting involvement in development disorders (the DDG2P gene list is a courtesy of DDD team).

Variant	Chromosome	Genotype	All genes (n=~20,000)		DDG2P genes (n=1,148)	
			LOF	Functional	LOF	Functional
SNVs	Autosomal	Homozygous	0.02	1.01		0.08
		Compound heterozygous	0.13	2.99	0.01	0.42
	X-chromosome (male child)	Homozygous	0.1	3.28	0.02	0.63
INDELs	Autosomal	Homozygous	0.03	0.03		0.08
		Compound heterozygous	0.11	0.07	0.01	0.43
	X-chromosome (male child)	Homozygous	0.07	0.12	0.03	0.66
Total candidate genes in a female child			0.29	4.1	0.02	1.01
Total candidate genes in a male child			0.46	7.5	0.07	2.3

FEVA requires 1-3 minutes to generate a report of candidate genes for one trio. When run in parallel, FEVA can generate reports of candidate genes for thousands of exomes in a few hours with minimum memory usage (< 50 Mb per

trio). This feature makes FEVA suitable for large-scale projects such as the DDD, which aims to analyze the exome data from 12,000 trios in the next couple of years.

Application 3: Affected sib-pairs in UK CHD families

In collaboration with Prof. Eamonn Maher at the University of Birmingham, I analyzed the exome data of 10 families with at least two CHD affected sibs. Two of these families are consanguineous (from Birmingham Pakistani population). All families have two affected sibs except family CHD1 and CHD16 where each has three affected sibs of various CHD phenotypes.

I used FEVA software to generate reports of rare coding variants that are shared between at least two sibs (Table 2-17). The rare variants are defined as variants with MAF < 1% in 1000 genomes and the internal MAF of 2,172 parents from DDD project. As expected, affected sib-pairs from consanguineous families (CHD1 and CHD4) have more candidate genes with autosomal recessive rare coding variants than non-consanguineous families. On average, each family's FEVA output lists 3.5 gene candidate genes with homozygous rare coding variants and 25 candidate genes with compound heterozygous rare coding variants.

Initially, I focused my search for candidate genes with rare loss of function (stop gained, frameshift or variants disturbing acceptor or donor splice sites) (Table 2-18). The top recurrent five genes that appear in most of the families (*ANKRD36C, LINC00955, CDC27, OR4C5,* and *MUC3A*) are unlikely to be linked to the CHD phenotypes since they have compound heterozygous LOF in almost all families. Most of the remaining genes do not have knockout mouse models except three genes (*TTN, PLA2G1B* and *RBMX*) and *TTN* is the only gene that shows structural cardiac defects in the mouse models. Since it not expected to identify recurrent pathogenic genes in such a small study with variable CHD phenotypes, I only considered genes that appear in one affected sib-pair only. I also excluded genes with frameshift variants (INDELs) since they tend to have a

higher false positive rate. Only two genes, *GMFG* and *TAS2R43,* met all filters. TAS2R43 gene encodes a taste receptor and it is unlikely to have a rule in CHD. On the other hand, *GMFG* harbors a rare homozygous stop gain variant (p.Arg24X) in two sibs diagnosed with tetralogy of Fallot in family CHD1 (Figure 2-26). Upon validation with capillary sequencing (carried out by my colleague Chirag Patel), the same homozygous variant co-segregate in the third affected child with TOF (IV:4) but heterozygous in both parents not seen in the fourth child with ventricular septal defect (IV:3) . This variant was absent from ~200 ethnically matched control chromosomes.

Table 2-17 Number of candidate genes with shared coding rare variants, in at least two sibs, under autosomal recessive model.
* Numbers in parenthesis are number of gene candidates with rare coding variants shared between all three sibs.

Family ID	**Consanguineous family**	**Child / Phenotypes**	**Number of sibs**	**Number of candidate genes**	
				Homozygous	Compound heterozygous
CHD1	Yes	Child 1:TOF Child 2:VSD Child 3: VSD, PA (TOF spectrum)	3	23 (1)*	36 (29)*
CHD4		Child 1: VSD, PA (TOF spectrum) Child 2: AS	2	18	24
CHD5	No	Child 1: VSD, RV hypoplasia Child 2: ASD, RV hypoplasia	2	3	21
CHD6		Child 1: TOF Child 2: TOF	2	6	25
CHD11		Child 1: VSD Child 2:AS, BAV	2	1	29
CHD13		Child 1: TGA, VSD, PS Child 2: TGA	2	0	25
CHD16		Child 1: TOF Child 2: VSD, CoA, BAV Child 3: ASD	3	39 (1)*	36 (28)*
CHD20		Child 1: Tricuspid Atresisa Child 2: TGA, RV hypoplasia	2	1	29
CHD22		Child 1: HLHS Child 2: VSD	2	4	19
CHD23		Child 1: AS, subaortic stenosis Child 2: AS, subaortic stenosis	2	0	23

ASD: Atrial Septal Defects, AS: Aortic stenosis, BAV: Bicuspid Aortic Valve, CoA: Coarctation of Aorta, HLHS: Hypoplastic left heart syndrome, PA: Pulmonary Atresia, RV: Right Ventricle, TGA: Transposition of the Great Arteries, TOF: Tetralogy of Fallot, VSD: Ventricular Septal Defects.

GMFG was initially identified as a growth and differentiation factor acting on neurons and glia in vertebrate brain [284]. *GMFG* encodes a small protein of 142 amino acids an actin-binding protein predominantly expressed in microvascular endothelial cells and inflammatory cells [285, 286]. The expression of *GMFG* was found to be unregulated at the site injury during the heart regeneration in

zebrafish models[287]. However, its role in the heart development in mammals has not been studied yet. A knockout mouse of *GMGF* is being modelled at the Wellcome Trust Sanger Institute to investigate further its role during the development of the heart.

Table 2-18 List of candidate genes with rare loss-of-function variants shared in between at least two affected sibs. Genes in red harbor stop gained (SNVs) variants while the rest have frameshift. The phenotypes in knockout mouse models from the Mouse Genome Database [288].

Gene	**Number of families with candidate genes carrying**		**Phenotypes in mouse knockout mouse models Mouse Genome Database**
	Homozygous	Compound Heterozygous	
ANKRD36C		10	NA
LINC00955		10	NA
CDC27		10	NA
OR4C5		9	NA
MUC3A		9	NA
RBMX		5	Decreased lean body mass
CCDC144NL		4	NA
FAM182A		1	NA
TTN		1	First branchial arch and somites, vascular, cardiac and skeletal muscle defects.
MUC4		1	NA
PLA2G1B		1	Abnormalities in lipid absorption and increased insulin sensitivity.
KLHL24		1	NA
ROPN1		1	NA
PITPNC1		1	NA
GMFG	1		NA
TAS2R43	1		NA
ZNF717	1		NA

Next, I performed the same analysis but for shared rare missense variants and identified 119 genes with homozygous and / or compound heterozygous variants in these families (Table 2-19). The majority of the genes appear only in one affected sib-pair while a few appear in all of them (mainly genes from the Olfactory or Mucin gene families which are unlikely to be causal in CHD).

Two of these genes are well known CHD genes such *NOTCH2* and *TBX20* although as dominant genes. Other genes knockout mouse models exhibit structural heart defects (*UTY*, *HSPG2*, *CTBP2*, and *ADAM12*).

Table 2-19 List of candidate genes with rare missense variants shared between at least two sibs. Genes in red have a knockout mouse models that exhibit structural heart defects [288].

Number of Affected sib-pairs	Homozygous	Compound heterozygous
1	*ZC3H13, PGLYRP2, FAM182A, PLCH2, KIAA1683, ZFX, NPIPB1P, PSG6, HR, SHROOM4, PSG11, GMIP, GUCY2F, IKBKG, LPAR4, OR11H6, SPTBN4, UTY, FCGBP, TRGC2, GPKOW, TAS2R43, SLITRK2, MUC16, CXorf61, CXorf64, GPR112, LYNX1, ZNF431, MEGF6, IL12RB1, LRBA, NADK, ZNF30, NKX2-1, ASXL3, OR11H7, MCOLN1, VCX2, OR4L1, TUBGCP5, NDUFA13, HSPG2, TRIT1, OR4K13, PKN2, AQP12A, HNRNPA1L2*	*CTBP2, MYEOV, FILIP1L, FAM182A, TMC2, LRSAM1, CMYA5, KANK1, FAT1, TYRO3, IGHV5-51, MYOCD, TBX20, STIL, SPTBN5, NRCAM, GPR108, MYO15A, PITPNM1, ADAM12, MYO7B, GCOM1, FRAS1, PLA2G1B, LAMB2, RANBP2, IQGAP1, AHRR, PRRC2B, PTGFRN, ODZ4, TRIOBP, HNRNPCL1, KIAA2022, IGHV3-38, NOTCH2, FRG2B, PDHX, AHNAK2*
2	*MUC4*	*FRG1, SRRM2, FAM27E1, USP6, DNAH14*
3	*SLC9B1P1*	*ATM, IGHV7-81, MUC16, ARSD*
4		*PRSS1, CCDC144NL*
5		*TTN, TRGC2, LINC00273*
6		*IGHV2-70, IGLV5-45*
7		*CEP89, NCOR1, RBMX*
8		*TAS2R31*
9		*MUC4*
10		*MUC6, TRBV6-5, ANKRD36C, MUC3A, BCLAF1, OR9G1, CDC27, AQP7, LINC00955, KCNJ12, MUC3A, OR4C5, OR4C3*

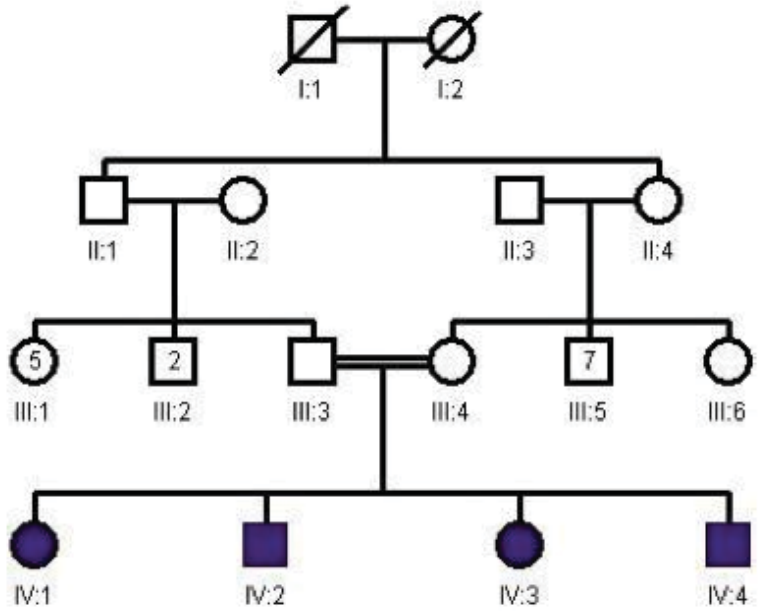


Figure 2-26 **Pedigree chart of family CHD1.**
Four affected sibs from a consanguineous family of a Pakistani origin. Only three sibs had their exome sequenced in this study (IV:1, IV:2 and IV:3). All sibs are diagnosed with tetralogy of Fallot except IV:3 who is diagnosed with ventricle septal defect (VSD). The homozygous stop gain variant was detected in two sibs with TOF (IV:1, IV:2) and capillary sequencing confirmed the presence of the same homozygous stop gain variant in the third sib with TOF (IV:4). Both parents are heterozygous for this variant and in 200 ethnically matched control chromosomes but not see in the child with VSD (IV:3). (Dr. Chirag Patel at the University of Birmingham performed the validation work).

Application 4: Affected parent-child pairs in UK10K CHD families

Most of the samples in UK10K (RARE CHD) are index cases (110 out of 124 samples) except for a few related samples (three affected parent-child pairs, one affected sib-pair and two parent-offspring trios). In this analysis, I focused on the affected parent-child only as this family structure is not covered in the analyses described above. In such a family design, I only looked for inherited rare coding and heterozygous variants shared between the parent and the child.

I used FEVA software to report rare coding heterozygous variants shared between the parent and the child. I defined rare as variants with MAF < 1% in 1000 genomes and the internal MAF of 2,172 parents from DDD project. On average, each affected parent-child pair shared 230 candidate genes (Table 2-20), which is much higher than the number of candidate genes in affected sib-pairs or complete trios (28 and 7 candidate genes, respectively). It is important to note that the number of candidate genes in these families is even larger (47% more) than the number of candidate genes from the simulated parent-child family (see Table 2-10 for details), which has 157 candidate genes on average. This is likely to be as a result in the differences in the calling pipelines (UK10K vs. GAPI). The internal MAF from the 2,172 is based on GAPI pipeline and it is likely to be less effective on samples that went through the UK10K pipeline and thus have more candidate genes per family.

Table 2-20 Number of candidate genes with rare coding heterozygous variants shared between affected parent and child in three CHD families form UK10K RARE CHD project. Loss of function class includes (stop gain, frameshift, variants that disturb acceptor or donor splice sites), functional class includes (missense, in-frame deletion or insertion and stop lost).

Family Id	CHD phenotype		Number of candidate genes	
	Child	Parent	Loss of function	Functional
UK10K_CHD_0015	Atrial septal defect	Atrial septal defect	23	219
UK10K_CHD_0060	Atrioventricular septal defects	Ebstein's anomaly	24	208
UK10K_CHD_0067	Pulmonary stenosis and Atrial septal defect	Pulmonary stenosis	15	201

Since the number of genes with rare functional variants is large in each affected parent-child pair (~200), I focused my search for genes with rare heterozygous loss of function variants (this class includes stop gain, frame-shift, variants that disturb acceptor or donor splice sites) and are shared between the affected parent and the child (Table 2-21). The heart phenotypes observed in these families are varied from family to family and thus I did not expect to see the same gene appear more than once. There are 29 genes where each one has a single loss of function in a single family (first row in Table 2-21). Only one gene, *CCDC39*, shows heart phenotypes in knockout mouse models. This gene harbors a rare frame-shift (c.610_614delTTAGAinsA) in a parent with Ebstein's anomaly and a child with atrioventricular septal defect (family id: UK10K_CHD_0060).

CCDC39 gene encodes a protein that localizes to ciliary axonemes and is essential for the assembly of inner dynein arms and the dynein regulatory complex [289]. Recessive loss of function variants have been found to cause a large proportion of primary ciliary dyskinesia in human. However, the knockdown of Ccdc39 in zebrafish embryos at the 2-cell stage caused a dose-dependent increase in heart looping defects and other laterality defects may suggest a possible *CCDC39* haploinsufficiency [289]. Moreover, a knockout mouse model submitted to the Mouse Genome Database (MGI:5445973) [288] shows double outlet right ventricle, atrial septal defect and dextrocardia but it has not been published. These findings suggest the involvement in *CCDC39* in the development of the heart but further work is required to confirm the role of this heterozygous frame-shift variant in causing the heart phenotypes observed in this family.

Table 2-21 List of genes with rare loss of function (stop gain, frameshift, variants that disturb acceptor or donor splice sites) variants shared between affected parent and child.

Number of affected parent-child pairs	**Genes**
1	*ATXN3L, AXDND1, CCDC39, CCDC7, CCL8, CD5L, COL6A5, CYP2C8, AC061992.1, ERAP1, F5, FAM49A, FHAD1, FLG2, GPLD1, MUC19, NDUFA10, NLRP5, OR51E1, OR51T1, OR5AN1, POLR1A, SERGEF, SMYD4, TAS1R3, TAS2R43, VNN2, VPS8, ZNF211*
2	*PRSS3, RBMX*
3	*CDC27, LINC00955, FRG1B, MUC3A, OR4C5*

2.4 Discussion

NGS has accelerated gene discovery in rare monogenic disorders in the last few years. More than 180 novel genes have been identified using whole genome or whole exome sequence data generated by NGS platforms so far. Based on the current rate of novel gene discovery, it has been estimated recently that most of the disease-causing genes of rare monogenic diseases will be identified by the year 2020 [202].

The success of NGS with rare monogenic disorders inspired me to apply the exome sequencing strategy for studying congenital heart defects (CHD). However, applying NGS to CHD cases is not straightforward since the inheritance model for CHD is not well defined. Evidence from genetic epidemiology and genome-wide association studies has supported the polygenic model [112, 115] and at the same time several monogenic examples of isolated and familial forms of CHD have been reported in the literature [14]. There is no general consensus on what is the most plausible inheritance model that can explain CHD. For this reason, I explored four different family-based study designs in order to evaluate the power of each design to identify rare coding variants that might explain the monogenic CHD cases.

This chapter describes the tools and pipelines used to call single nucleotide (SNVs) and insertion/deletion (INDELs) variants from exome data. One major challenge I addressed is how to improve the sensitivity and specificity of variant calling from exome data. The issue of sensitivity and specificity stem from the underlying probabilistic statistical models implemented by different variant callers. These models are being actively developed and thus it is expected that the best practices for filtering and cleaning up exome data will keep changing for the foreseeable future, especially for indels.

In this thesis, two pipelines have been used to call variants from exome data: GAPI and UK10K pipelines. Both of these pipelines use different callers and

filters to generate the variants. Although they have been able to detect a relatively comparable number of coding SNVs, the number and type of INDELs varied substantially in both pipelines. This is most likely caused by the use of an additional caller, Dindel, to detect INDELs in the GAPI pipeline. On the other hand, the intra-pipeline comparisons between GAPI sample releases at different time points show minimal differences. These findings highlight the need to use only one pipeline for consistency and to avoid unnecessary complications for the downstream analysis (such as case/control analysis using the samples from different pipelines as discussed in chapter 4).

To improve the sensitivity and specificity of SNV calls generated by UK10K pipeline as an example, I tested the relationship between strand bias (SB), quality by depth (QD), genotype quality (GQ) and variant quality (QUAL) with transition/transversion ratio (Ts/Tv) to chose the proper filtering thresholds. Applying these filters has helped me to eliminate low quality variant calls in a systematic fashion. However, this method of variant filtering using hard cut-offs is no longer considered the best practice and newer filters based on sophisticated statistical models that integrate several quality metrics simultaneously have now been used. One example is the Variant Quality Score Recalibration (VQSR) scores recently implemented in GATK, which seems to be superior to other filtering methods. However, VQSR is not so successful for filtering indel callsets since it is sutable for SNV callsets only.

It is not uncommon to use more than one variant caller to detected SNVs and / or INDELs to improve the sensitivity and specificity of variant calling. Theoretically, callers that utilize different probabilistic models to call variants independently, are most appropriate. However, it was not clear how to resolve conflicts that arise when a variant passes the filters of one caller but not the other, or when a variant is missed by one of them. My analysis of 14 different datasets (seven INDELs and seven SNVs) based on different scenarios shows that INDELs called by Dindel were superior to Samtools calls, as they show in-frame/frameshift (n3/nn3) ratio closer to the exacted ~1.5 ratio. Similarly, GATK SNVs calls were superior to Samtools calls in terms of transition /transversion (Ts/Tv) and

rare/common ratios. These results have led me to change the order of caller when I merge calls in the final variant call format files (i.e. I used Dindel as the default caller for INDELs and GATK as the default caller for SNVs). Such a small decision has a large effect on the final number of rare coding variants. For example, Samtools calls more rare loss of function variants than GATK or Dindel. Such that, in large-scale projects, this could mean hundreds of false positive candidate variants that would slow down any downstream analysis or functional studies.

Once an optimal callset of variants is obtained, it is important to exclude common variants based on **minor allele frequencies** (MAF) to minimize the number of candidate variants. There are many population-based MAF recourses available to facilitate this step such as 1000 genomes (1KG), UK10K Twins cohort (UK10K) and the NHLBI Exome Sequencing Project (ESP). Additionally, I generated a fourth MAF resources (called internal DDD MAF) based on 2,172 parental samples generated by GAPI pipeline to target variants that appear as rare variants in the public MAF recourse but are common in the internal samples which likely indicate that they are sequence or calling errors.

Matching alleles between sequenced samples (e.g. DDD or CHD samples) and the population variation resources (e.g. 1000 genomes project) in order to obtain the correct minor allele frequency is straightforward for SNVs but more difficult for INDELs since they can be called differently due to the genomic context such as homopolymer runs for example. To assign the correct MAF, I tested three allele-matching strategies (two exact matching algorithms and one lenient algorithm based on 10-30bp matching window) and I used the correlation between the observed minor allele frequency in DDD samples and the population allele frequency from all three MAF population resources as a metric to compare different matching strategies. I showed that the exact strategies have a stronger correlation between the observed minor allele frequency from DDD samples and population allele frequency from all three MAF population resources.

Using the **exact matching algorithm**, I evaluated the consequence of applying each MAF resource independently and combined on the final number of rare candidate variants in 288 affected samples from the DDD project. This analysis showed that the internal frequency from the DDD project alone was able to eliminate most common variants compared with other combined public MAF recourses. Combining two or more MAF is more effective than using each individually. However, using allele frequencies from ESP and UK10K has some drawbacks. First, ESP includes many affected samples with unpublished phenotype, which may include CHDs and thus cannot be used as controls. Moreover, the targeted exome in ESP is smaller than the exome design used to sequence CHD samples in my thesis, (~16,000 genes and ~20,000 genes, respectively. Similarly, the MAF from the UK10K Twin cohort does not include variants on X-chromosome. For these reasons, I decided on a MAF filtering strategy using the 1000 genomes project data combined with the internal allele frequencies from healthy parents in DDD project to exclude common variants and pipeline errors.

Another factor that affects the final number of candidate variants/genes is the **family design**. I performed a simulation analysis using one multiplex family of three affected sibs and two parents and showed how the number of candidate variants varied between singletons, sib-pairs, parent-child, and complete trios study designs within the same family.

The **Singleton** study design generates the largest number of candidate variants per sample compared with other family-based study designs, unless it is combined with linkage analysis to limit the search in a smaller region. The example of 'distal hereditary motor neuropathies type VII' with two small linkage regions (9.2 and 4.3 Mb) has identified only one candidate gene, *SLC5A7*. This example, in addition to another three genes identified using the same strategy (*B4GALNT1*, *KPTN* and *WDR62*), indicates that finding causal genes by combining NGS and linkage analysis can be powerful and relatively straightforward. Without linkage analysis, the number of candidate genes per sample is usually large especially for dominant disorders. In the absence of

linkage analysis information, sequencing multiple unrelated cases may help to identify the causal gene in monogenic disorder, but can be challenging for extremely genetically heterogeneous disorders such as intellectual disabilities and CHD. In such disorders, a case/control analysis might be more suitable but requires a large number of samples.

The affected sib-pairs design is helpful when looking for shared homozygous or compound heterozygous candidate genes in non-consanguineous families or homozygous candidates in consanguineous families. This analysis has highlighted variants in a few known CHD genes such as *NOTCH2* and *TBX20,* but these genes are mostly known to cause CHD under a dominant model while they have been reported here to harbor rare and presumably recessive variants. It remains to be seen if these variants are pathogenic. Additionally,_I identified novel genes such as *GMFG* with a homozygous stop gain shared between three affected sibs in the same consanguineous family of a Pakistani origin. These candidate genes were found in a single sib-pair only and thus require additional families sharing the same candidate genes to be identified and / or to be confirmed by functional studies. Nonetheless, the number of recessive candidate genes in this design is manageable and provides a chance to investigate the recessive model in different CHD subtypes.

The **trio and multiplex designs** identify far fewer candidate genes than the other designs because of the additional information from the parents. Assuming healthy parents and complete penetrance, each trio has, on average, seven rare inherited coding variants and a smaller number in multiplex families. The small number of candidate genes per trio makes most downstream analyses amenable to further investigations either *in silico* or by functional experiments (e.g. modeling in zebrafish). The design is also suitable for *de novo* analysis, as I will discuss in the next two chapters.

Many of the steps described above are time consuming and error prone when performed manually in non-specialized software such as Microsoft Excel. I designed the **"Family-based Exome Variant Analysis" (FEVA)** tools to

automate applying various quality filters and to report candidate genes from different study designs. FEVA reports candidate variants under different models of inheritance and can be customized by the end users to accommodate new family designs not covered by the program default settings. I used FEVA successfully to find causal genes in monogenic disorders from single cases such as the *SLC5A7* gene in distal hereditary motor neuropathy (type VII) [281] and another three genes (*B4GALNT1, KPTN* and *WDR62*) in various neurodevelopmental disorders (manuscripts were submitted or are being prepared). Other groups at the Wellcome Trust Sanger Institute as well as external groups from Cambridge University, University College London and other institutes, working with different rare disorders such as ciliopathies [290-292], neuromuscular, thyroid disorders and familial hyperlipidemia, have used FEVA to identify mutations in novel or known genes. Moreover, FEVA is also being used in large-scale projects with hundreds of families, such as in the Deciphering Developmental Disorders (DDD) project [260].

The results from this chapter show that at every step of the analysis pipeline small, seemingly insignificant, changes can have a big impact on the numbers of candidate variants being explored. Planning an upgrade of a pipeline, implementing a new version of a caller, modifying a filter threshold are some of the decisions that should not be taken lightly without careful consideration of how such a decision would affect the output. This is especially true in clinical settings where maximum levels of sensitivity and specificity are required for a definitive diagnosis.