5 | Discussion

In this thesis I explored different subtypes of congenital heart defect (CHD) using next-generation sequencing (NGS) data with a focus on family-based study designs such as parent-offspring trios. Even with the relatively small sample sizes of the cohorts studied in this thesis, I was able to detect three clearly pathogenic genes: *NOTCH1* and *JAG1* in isolated tetralogy of Fallot and *NR2F2* in isolated atrioventricular septal defects.

What did I learn about exome analysis pipelines?

At the beginning of my PhD studies, variant calling from whole genome or whole exome sequencing data was still in its infancy. It was not clear what were the best practices, pipelines, tools or filtering strategies required to achieve high levels of sensitivity and specificity for variant identification. This led me to investigate different aspects of the variant calling workflow to determine appropriate callers and filters to achieve high specificity and sensitivity.

Initially, I assessed **sequence and variant calling parameters** such as phred-like quality (QUAL), strand bias (SB), quality-by-depth (QD) and genotype quality (GQ) in order to set thresholds to eliminate low quality variants. These filters and thresholds worked well for the early sample releases, but as the underlying probabilistic models for calling and filtering variants improved, these filters changed accordingly and they will probably continue to change in the foreseeable future. Newer parameters of sequence data and variant calling have emerged and they are replacing many previous filtering strategies (for example the Variant Quality Score Recalibration (VQSR) filter from GATK caller has been suggested as a superior quality filter for single nucleotide variants from exome sequencing, but not indels). Currently, choosing the right set of filters and thresholds is an area that needs to be revisited on a regular basis in order to adhere to the best practices available.

Another important part of variant calling workflows, which is usually overlooked, is **how to merge variants identified by two or more callers** (e.g. Samtools and GATK). If the two callers disagree on an alternative allele or a genotype, which caller should be used as the default? When I started my projects I decided, naively, to use GATK as the default caller over both Samtools and Dindel, for samples called by the Genome Analysis Production Informatics (GAPI) pipeline. However, when I investigated this issue in more detail later on, I discovered a complex relationship between the type of the caller used as a default caller, and the number and type of rare coding variants identified and reported for downstream analysis. For example, Samtools tends to call more rare loss of function variants (~8 per sample on average) that are either missed by GATK or have been flagged by GATK as low quality variants. I was able to show that these variants exhibit a low transition/transversion ratio, which is indeed a sign of being low quality variants. In studies with a small number of samples this might not be a major issue, but for large-scale projects with hundreds or thousands of samples such as the Deciphering Developmental Disorders project with 12,000 affected children, this can have a huge effect on the amount of downstream work required for validation and / or functional experiments. These findings hold true for the version of the callers used to call variants in my samples, but it is expected to change when using a different version of the same caller, and thus it is important to perform this detailed analysis whenever a newer version of a variant caller is implemented.

Small decisions such as what threshold of a filter should be used, or which is the default variant caller, can lead to big differences in the type, number and quality of the variants identified in whole exome data, especially the rare coding variants of greatest interest in rare disease studies. This was clearly manifested by the variant differences I identified between **two analytical pipelines** that were used to call variants from the CHD samples described in this thesis: Genome Analysis Production Informatics (GAPI) and UK10K. Both pipelines used different numbers and versions of the variant callers and they also adopted variable filters and thresholds. Each difference might have a small effect on its own, but their cumulative effects are appreciable. The most obvious differences I observed were in the number of rare coding variants in the GAPI pipeline which called (~42%) rare missense variants and almost 4.4-fold more coding insertion/deletion (indels) than the UK10K pipeline. When samples are used from both pipelines, as they were in the burden analysis of rare missense variants in **chapter 4**, I noticed an inflation of quantile-quantile (Q-Q) plots. An obvious explanation was that the inflation was caused by the high number of rare missense variants in the GAPI pipeline compared with the controls from the UK10K pipeline. However, it is likely that the explanation is probably more complex, and is caused by multiple factors. More work is required to investigate the origin of these differences.

What did I learn about tetralogy of Fallot?

The **two-stage study design** I used to investigate the genetic architecture of isolated **tetralogy of Fallot** enabled me to detect two clearly pathogenic genes: *NOTCH1* and its ligand *JAG1* in a cohort of 238 parent-offspring trios. Although both genes have been associated with congenital heart defects in the past, their involvement in the isolated tetralogy of Fallot is less well appreciated. Rare coding variants in *NOTCH1* have been linked to familial forms of left ventricular outflow tract malformations more often than with the malformations of the right side of the heart. Similarly, mutations in *JAG1* are usually associated with Alagille syndrome where CHD occurs in ~90% of the patients (6-17% are ToF) more often than with non-syndromic tetralogy of Fallot. I was able to detect *de novo* coding variants (except silent variants) in these genes in 2.5% of patients in this cohort. These variants included four *de novo* coding variants in the *NOTCH1* gene and two *de novo* coding variants in the *JAG1* gene. Interestingly, two-thirds of these *de novo* variants are loss-of-function, which showed up as a highly statistically significant burden of *de novo* loss-of-function in the *NOTCH1* gene (P=3.8 ×10^{-9}).

More interestingly, a **theme has emerged when I combined** *de novo* variant analysis with other analyses that target rare coding variants with presumably intermediate effect size (i.e. incomplete penetrance). I identified two genes, *NOTCH1* and *ARHGAP35*, both with *de novo* functional or loss-of-function variants, and both were also enriched for rare inherited missense variants. The case/control analysis identified *NOTCH1* as being enriched for rare missense variants (P=8.8 × 10^{-05}). On the other hand, the modified transmission disequilibrium test (TDT) identified an over-transmission of rare missense variants in the *ARHGAP35* (P=0.02).

Collectively, these genes have five *de novo* variants where all but one, are loss-of-function variants. This observation suggests that two classes of variants contribute to the isolated tetralogy of Fallot. The first group is rare coding variants with large effect size, mainly loss-of-function, that are able to cause the phenotype when they occur *de novo*. The second group is rare, typically missense, variants that increase the risk of isolated tetralogy of Fallot but are not sufficient to cause the phenotype by themselves. This group might require additional in *cis*- or *trans*- variants in order to cause the phenotype. One way to investigate this possibility is the digenic inheritance model that I described in **chapter 3**. Although the digenic inheritance analysis has identified a few interesting gene pairs such as *ZFPM2-CTBP2* that are enriched for rare missense variants in cases compared with 1,080 controls, the sample size is clearly underpowered, so I was not able to obtain signals that are statistically significant at the genome-wide level.

What did I learn about isolated atrioventricular septal defects?

Similarly, combining *de novo* analysis with case/control analysis enabled me to identify *NR2F2* as a novel candidate gene **for isolated atrioventricular septal defects** (AVSD) in human (**chapter 4**). Although the case/control analysis of a burden of rare missense variants burden did not, on its own, identify *NR2F2* as the most significant gene, it was the subsequent *de novo* analysis that identified this gene as the most intriguing candidate gene in this cohort.

NR2F2 is one of the most conserved genes across the genome and exhibits very little variation in populations, which supports its fundamental roles in the development of many organs, including the heart. Additionally, the published conditional knockout mouse model recapitulated many of the atrioventricular septal defects observed in human. These findings have been shown by others to be driven by defective endothelial-mesenchymal transformation (EMT) and the hypocellularity of the atrioventricular canal, accompanied by down regulation of the *Snai1* gene. Moreover, the results from luciferase assays (appendix B) performed by my colleague, Sebastian Gerety, indicate that all Nr2f2 coding sequence variants identified from the AVSD cohort had a measurable impact on transcriptional activation in at least one target gene. Further modelling work will be required to clarify whether these differences between target genes translate into distinct biological mechanisms of disease, affecting single or multiple molecular interactions required for heart morphogenesis.

Expanding the search for *NR2F2'*s mutations in other CHD subtypes revealed its involvement in tetralogy of Fallot, hypoplastic left heart syndrome and coarctation of the aorta. This analysis increased the total number of CHD families with *NR2F2* to eight (I have identified six CHD families while the other two CHD families were identified by my collaborators: David Wilson, David FitzPatrick and Catherine Mercer who identified a *de novo* balanced translocation in a child with coarctation of the aorta and Marc-Phillip Hitz who identified a *de novo* splice site in a child with hypoplastic left heart syndrome). These findings suggest *NR2F2* as a **novel dosage-sensitive regulator gene** involved in the CHD in human similar to other well-known CHD genes such as *TBX5*, *TBX1*, *NKX2-5* and *GATA4*. I hypothesise that these master regulators potentially play a key role in integrating genetic and environmental risk factors for abnormal cardiac development, although testing this hypothesis will require substantial downstream work.

What did I learn about study designs?

The two most **informative study designs** I evaluated in my thesis are the trio-based and the case/control designs. The trio family-based design is a versatile design since it is amenable to different analyses aimed to investigate rare coding variants with large size effect as well as variants with intermediate effect sizes. *De novo* analysis is the main test used to investigate variants with large effect size. Less commonly used, the modified transmission disequilibrium test (TDT) tries to identify over-transmission of rare variants from healthy parents to their affected children, as well as the digenic inheritance analysis which targets rare variants in affected children inherited from two different parents. The case/control analysis worked surprisingly well given the small size of the cohorts in this thesis. Its success is most likely attributed to being used in combination with the results from the *de novo* analysis. Nonetheless, performing case/control analysis in larger sample size of homogenous CHD cohorts is expected to identify additional genes involved in congenital heart defects. Other study designs I used such as affected parent-child and affected sib-pairs were not as successful, but this is likely to be due to the small sample size of these studies, and the difficulty in identifying additional families with similar mutations.

What were the limitations of my work?

Next-generation sequence (NGS) platforms have revolutionized the way we identify causal genes in monogenic disorders. This technology has helped me to identify different causal genes in two non-syndromic CHD subtypes. Nonetheless, NGS platforms impose some major **analytical challenges**. The most important one is the fact that my analysis, in common with all such analyses, has identified too many variants of unknown significance (VUS). This reflects our current state of very limited understanding of the function of most genes and the consequences of most variants. One way to overcome this problem in gene discovery analysis, will be to increase the sample size in order to increase the power of genetic analyses. International collaborations and data sharing will be important for increasing sample sizes. For VUS in known CHD genes, functional assays *in vivo* or *in vitro* may help to confirm their pathogenicity, although even these assays will have their associated false positives and false negatives.

How do my findings relate to other peoples work?

Recently, Zaidi *et al.* used NGS to sequence the whole exome in a trio cohort of 362 severe cases of syndromic and non-syndromic CHD and predicted that *de novo* point mutations in several hundreds of genes may contribute to ~10% of severe CHD cases [256]. This estimation is difficult to ascertain using the samples described in my thesis, since I have a much smaller sample size of trios (n=43 complete trios with whole exome sequence data). Nonetheless, I was able to identify likely pathogenic *de novo* variants in *NOTCH1* and *JAG1* in 2.5% of isolated tetralogy of Fallot (six out of 238 trios) and about ~12% in atrioventricular septal defects trios (two out of 16 complete trios that were available with either exome data or capillary sequencing) but given the other candidate genes that I identified with *de novo* variants (e.g. *ZMYM2, ARHGAP35, HDAC3*). My results are broadly consistent with the conclusion by Zaidi et al.

Future directions

Selecting an optimal variant calling pipeline is not an easy task and once one is implemented, any potential upgrade or new pipeline needs to be assessed in considerable detail to ensure that data quality is improved. Equally importantly, using a single, consistent, pipeline is essential in order to obtain consistent datasets, which helps to avoid complicating any downstream analyses.

Future CHD studies will require **larger sample sizes,** possibly of the order of a few thousand samples, in order to achieve enough power to identify a substantial fraction of recurrently mutated causal genes. Given the rarity of many CHD subtypes, a **national and international network of collaborators** is necessary to collect enough samples for parent-offspring complete trios and/ or case-control designs, both of which have been shown to be suitable study designs for isolated CHD.

Beside the genetic components required to support newly identified CHD genes in trios and case/control study designs, **functional experiments** are essential to confirm the pathogenic effect of genes in animal models using knockout or knockdown experiments in mouse and zebrafish models. Where appropriate, the pathogenic effect of specific variants can also be investigated using cell-based assays such as luciferase activity experiments. Moreover, integrating exome and genome sequence data with gene expression data using RNA-Seq from fetal heart tissues at different developmental stages are likely to be a helpful tool to prioritize candidate genes. Integrating high-throughput genetics, functional genomics and cellular and animal modeling will require concerted effort and collaboration.