1 Introduction

Rare genetic disorders often have a classical Mendelian pattern of inheritance, and they are often caused by a single high-penetrance variant. There are at least 6000-7000 rare genetic disorders, meaning that collectively they are in fact common, and the causes of around half have been identified thus far (1). While numerous different phenotypes are associated with rare genetic disorders, they often affect development, and first manifest *in utero*, in infancy, or in childhood.

There are two reasons why the study of rare developmental disorders is of great importance. First, it directly improves the lives of patients and their families. Occasionally, identification of the genetic cause of a disorder will lead to improved treatment or a new therapy for a patient (2). It also often allows patients and their families to access additional social and educational services, and it can allow estimation of recurrence risk for future pregnancies. Families affected by a rare developmental disorder often go through a 'diagnostic odyssey' that can last a decade or more, during which many different individual medical and genetic tests are performed in an attempt to identify the cause of the disorder (3). Therefore, finally receiving a genetic diagnosis can bring great peace of mind, even if it would not influence treatment.

The second reason to study rare developmental disorders is that they often give insights into relevant biological processes, and into the aetiology of more common forms of disease. This has been recognised for centuries. In 1657 Dr William Harvey observed that "there is no better way to advance the proper practice of medicine than to give our minds to the discovery of the usual law of nature by the careful investigation of cases of rarer forms of disease." For example, pathogenic variants in *PFN1* can cause familial amyotrophic lateral sclerosis (ALS), and they have also been implicated in the sporadic form of the disorder (4). Furthermore, this finding suggested that dysregulation of cytoskeletal machinery has role in the aetiology of ALS. In another example, pathogenic variants in several member of the SWI/SNF complex, which is involved in chromatin remodelling, can cause Coffin-Siris syndrome, highlighting the importance of appropriate chromatin remodelling (5).

Historically, identification of genes associated with rare developmental disorders relied on linkage mapping followed by positional cloning or painstaking Sanger sequencing of candidate genes. Many genes were identified in this way, including *CFTR* in cystic fibrosis, to name but one example (6). However, this method requires large families with multiple affected individuals, a relatively homogeneous and high-penetrance phenotype, and often knowledge of the function of candidate genes, which severely limits the utility of this approach. However, in recent years, the development of next generation sequencing (NGS) has enabled the entire genome (or selected portions of it such as the exome) to be sequenced in a rapid, systematic, high-throughput, and relatively cheap manner. This has led to nothing less than a revolution in the field of rare developmental disorder diagnostics and gene discovery.

The first example of NGS to identify a novel rare disorder-associated gene came in 2010, when pathogenic variants in *DHODH* were found to cause Miller syndrome (7). Since then, at least one hundred other rare disorder-associated genes have been identified through the application of NGS, bringing many advantages both directly to the lives of those patients, and indirectly to the wider understanding of the pathogenesis of developmental disorders (8). Several consortia around the world have been established to sequence the exomes or genomes of cohorts of patients with rare genetic disorders on a large scale, including the Deciphering Developmental Disorders (DDD) project, the UK10K project, the Finding of Rare Disease Genes (FORGE) Canada Consortium and others (3, 9-11).

Recently, there has been much discussion surrounding the exact extent and nature of the evidence required in order to state that a given gene is indeed associated with a rare genetic disorder. While there are still contentions in this area, the importance of a consistent and stringent approach is increasingly being recognised, and a preliminary set of guidelines for this purpose was recently published (12). Identification of a loss of function variant that segregates with a rare disorder in a single family is not on its own sufficient evidence that the variant causes that disorder, particularly because loss of function variants in many genes are not uncommon in healthy individuals (12, 13). Therefore, statistical or functional follow-up experiments are also required.

One very important and commonly used statistical follow-up approach is to identify potentially pathogenic variants in the same gene in multiple unrelated affected individuals (3). There is no one rule as to the number of unrelated individuals required to statistically demonstrate that the occurrence of a particular number of variants in a

particular gene is highly unlikely to occur by chance. Instead, the number required depends on various factors including the size of the gene, and its mutation rate. Another relevant statistical follow-up approach that can be used is identification of a significant burden of variants in cases compared to controls (14).

Functional follow-up approaches can be an alternative or complementary method to statistical follow-up approaches. Examples of such approaches include *in silico* experiments such as computational modelling of the effect of a variant on the structure of a protein (15), *in vitro* experiments such as investigation of the affect of a variant in human cells (16), and *in vivo* experiments such as recapitulation of aspects of patients' phenotypes using an appropriate animal model (17). Selection of appropriate statistical or functional follow-up experiments for the study of putative rare disorder-associated genes is of great importance, and depends on many factors including the availability of additional patients with overlapping phenotypes, the mode of inheritance of the phenotype, the predicted mechanism of action of the variant, and current knowledge of gene function.

In this dissertation I describe three distinct projects in which NGS was used to identify variants that cause rare developmental disorders, followed by statistical or functional follow-up approaches to validate or further explore the results. Because the projects are distinct, the following three chapters are self-contained, and the majority of the introductory and discursive material is located within each chapter.

The aim of the project described in chapter 2 was to explore how well exome sequencing performs as a method for identifying variants that cause abnormal fetal development, by performing exome sequencing on 30 parent-fetus trios where the fetuses had a diverse range of structural abnormalities. In chapter 2 I will describe the analysis of these data, different methods of interpreting variants, and the identification of causal and possibly causal variants. This project demonstrates that exome sequencing is a promising method for prenatal genetic diagnostics.

In chapter 3 I will describe a targeted resequencing study that was performed on a cohort of patients with intellectual disability (ID) as part of the UK10K project. The aims of this project were to identify causal variants in known ID-associated genes in the cohort, to identify novel ID-associated genes, and to ascertain whether there is a burden of variants in ID-associated genes in ID patients compared to controls. Statistical follow-up approaches such as the case-control enrichment analyses that I

will describe in chapter 3 can be a valuable method to give insights into the genetic aetiology of developmental disorders such as ID.

In chapter 4 I will describe a project in which two candidate dystroglycanopathy-associated genes, *B3GALNT2* and *GMPPB*, were identified using exome sequencing as part of the UK10K project. The aim of my work was to make zebrafish models of dystroglycanopathy by inhibiting the expression of each of these genes, and then to determine the extent to which the phenotype of these models recapitulated the phenotypes of the patients. I will demonstrate that zebrafish are an appropriate model for this purpose, and I will show that modelling candidate genes in zebrafish embryos is a functional follow-up approach that can help to determine whether a candidate gene is truly associated with a developmental disorder, and to give further insights into the pathology of that disorder.

The zebrafish project described in chapter 4 was carried out first (May 2011- February 2013), closely followed by the abnormal fetal development project described in chapter 2 (September 2011- November 2013) and then the project on the ID group of the UK10K project, described in chapter 3 (June 2013-August 2014). In this dissertation, I have described these projects in a non-chronological order, because the parts I played in each project flow more logically in the order in which I present them here. That is, for the abnormal fetal development project I was responsible for the majority of the analysis and interpretation of the exome sequencing data itself, for the ID project I was responsible for data analysis and also further statistical follow-up investigations, and for the zebrafish project I was responsible for functional follow-up of exome sequencing results using an animal model. All three projects serve to emphasise the importance of NGS for the diagnosis of rare developmental disorders, and for the identification of causal variants in novel genes.