Next Generation Mendelian Genetics

Sarah Ng

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:
Jay Shendure, Chair
Michael Bamshad
Deborah Nickerson

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Next Generation Mendelian Genetics

Sarah B Ng

Chair of the Supervisory Committee:
Associate Professor Jay Shendure
Department of Genome Sciences

The study of Mendelian disorders has been of immense utility in uncovering the genetic and molecular basis of numerous human traits, and has greatly furthered the identification of genes, the annotation of gene function and our understanding of biological pathways and cellular processes. Over the last 30 years, linkage analysis has been the most successful approach for finding the genes underlying Mendelian disorders, contributing to the identification of over 1,500 genes. However, thousands of disorders remain unsolved. Here I present a new paradigm to efficiently identify the genetic basis of Mendelian disorders, based on the direct observation of potentially causative mutations throughout the genome of affected individuals. This is enabled by massively parallel, or "next-generation", sequencing, which has made it increasingly feasible to generate large amounts of sequencing data at low cost. Although whole human genomes can now be sequenced, it is more cost effective to focus on specific regions of interest – for example, all the protein-coding regions (the "exome"), in which the majority of known Mendelian disease mutations are found. In this dissertation, I first describe a method for the efficient enrichment and sequencing of the human exome. I then validate this method by sequencing and describing the genetic variation in twelve human exomes – eight Hapmap samples and four samples with

Freeman-Sheldon syndrome (FSS) – and in a proof-of-concept experiment, show how the variation uncovered in exome data from the individuals affected with FSS can be filtered to identify the known causal gene. Next, I present the first successful applications of this approach to disorders of unknown genetic basis: a) Miller syndrome, a recessive disorder with only 40 described cases, and b) Kabuki syndrome, a dominant disorder where the majority of affected individuals are sporadic cases with no familial transmission. The development of exome sequencing and these filtering methods for exome data represents a new paradigm by which Mendelian disorders can be studied and new genes associated with disease can found, and as sequencing becomes ubiquitous, is likely to become a standard tool for the elucidation of the molecular basis of disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEDICATION

To Stan, my love.

CHAPTER 1 : INTRODUCTION

The study of Mendelian disorders has uncovered the function of thousands of human genes[1], and to date, nearly 3000 Mendelian phenotypes have been "solved" – that is, have had a genic cause associated with the phenotype[2,3]. Studies of such disorders have furthered our understanding of the genome on multiple fronts – from gene annotation and the identification of gene interactions and pathways, to elucidating the nature of mutations and mechanisms for pathogenesis.

1.1 What is a Mendelian disorder?

A Mendelian disorder is defined as one caused principally by one or more mutations in the genome of an affected individual, with little or no contribution by environmental effects. In most cases, Mendelian disorders are monogenic – i.e. mutations in a single gene are necessary and sufficient to cause disease; in very few cases, e.g. Bardet-Biedl syndrome[4], they are digenic – requiring two different genes to be inactivated before the phenotype is observed. When more and more genes contribute to a trait, such that mutations in them become neither necessary nor sufficient, but just predisposing, then a trait is multigenic and complex.

Although most Mendelian disorders are monogenic, the disorders are themselves each far from homogenous and the one-to-one gene-trait relationship does not always hold. First, as it is most often the case, there is allelic heterogeneity within a disease gene – that is to say, a "disease allele" could refer to any one of multiple inactivating mutations in the same gene. For example, there are currently 1,492 different mutations affecting F8 that are associated with haemophilia A[5], and while a single mutation in CFTR accounts for 66% of cystic fibrosis[6], over 1,900 other mutations in this gene have been implicated[7].

There is also genetic heterogeneity, where mutations in different genes can lead to the same phenotype (many genes – one trait). In an extreme example, mutations in over 45 different genes are known to cause retinitis pigmentosa[8], each gene with a specific inheritance pattern, be it autosomal or sex-linked, or recessive or dominant. It is worth noting that retinitis pigmentosa is still mainly a monogenic disorder, because any one of these mutations in any one of these genes is sufficient to cause the phenotype in a single individual.

On the other hand, there is also variable expressivity (that is, phenotypic heterogeneity), wherein a spectrum of phenotypes are observed, despite there being the same underlying mutated gene (one gene – many traits). In the gene *ABCA4*, different mutations can lead to a number of distinct ophthalmic disorders, ranging from retinitis pigmentosa to cone-rod dystrophy to age-related macular degeneration, which may be correlated with the severity of the mutation and gene activity (reviewed in ref. 9). Similarly, in cystic fibrosis, although in most cases *CFTR* is the mutated gene, and the mutation is ΔF508, there is a range of phenotypic effects including elevated salt in sweat, pancreatic insufficiency, lung disease, reduced fertility and meconium ileus, and the observation of each of these varies among affected individuals (reviewed in ref. 10). This is clearly due to *CFTR* having a cellular role in multiple organs, specifically in epithelial cells; at the same time, the range of phenotypic variability suggests that modifier genes, or even environmental factors, are present to modulate the effect of *CFTR* mutations in different organs[10].

Although Mendelian disorders are often presented on the opposite end of the spectrum from complex disease, it is clear they are far from simple. They are, however, by-and-large monogenic in nature, and based on this characterisation we expect causal mutations to have high penetrance and large effect.

1.2 Finding "disease genes"

The gene finding method that has found the broadest applicability in recent times is genetic mapping through linkage analysis, which is not dependent on any prior knowledge of biology or function, and is instead based purely on the inheritance of a trait in conjunction with the inheritance of chromosomal regions to pinpoint the location of disease-related genes.

The theory behind linkage analysis has been developed over the last 100 years, starting with the early observations by T. H. Morgan and others that traits that tend to be co-inherited are likely to be due to "factors" that are "linked" – i.e. genes that are in close proximity (e.g. ref. 11), because of a lower probability of recombination events happening between them. Based on this, Sturtevant (1913) created the first genetic map[12], using the results of experimental crosses of mutant flies to estimate the distance between trait-causing genes based on recombination fractions. Methods for genetic mapping were then further developed in experimental and agricultural biology systems, but because human genetic data are uncontrolled and observational by nature, it was only later that linkage methods were adapted to human genetics – first in nuclear families[13,14] and also in sibling pairs in the absence of parental data[15]. Likelihood ratios to include a comparison against the null hypothesis of no linkage were then introduced in multi-generational pedigrees[16], and eventually this was developed into the lod (log-odds or log-likelihood ratio) scores[17,18] that are currently used in linkage analysis. Smith (1953) also touched on identity-by-descent methods like sib-pair analyses and homozygosity mapping for recessive traits, as well as on the use of linkage disequilibrium (LD) to detect linkage in a population and in single-offspring pedigrees[17]. Arguably, these are all methods on the same theme – a likelihood approach to linkage detection[19], but are often presented now as separate approaches: linkage analysis, homozygosity mapping[20], the transmission-disequilibrium test[21] and association studies based on LD.

At the time the early statistical tools were developed, there was no way to do a genome-wide linkage study due to the paucity of human genetic markers. A typical linkage analysis was done with two loci that were known or assumed to be causal for discernible phenotypes – e.g. between loci underlying colour-blindness and haemophilia[16], or between the loci for haemophilia and G6PD electrophoretic variants[22]. This changed dramatically with the identification of classes of DNA markers that were polymorphic, found genome-wide, and easily assayed in individuals. The first such class was restriction fragment length polymorphisms, upon which a full human genetic map was proposed[23] and established[24]; later classes include the more highly polymorphic microsatellites[25] and the more abundant and easily assayable single nucleotide polymorphisms[26], which are currently highly utilised.

In the ideal case, the result of a linkage analysis would be the identification of a single gene within the linked locus, which could then be unambiguously shown to be causal for the phenotype. However, this hardly happens in practice. Rather, because there are limited meioses within human pedigrees, and hence a limited number of recombination events, linkage intervals that co-segregate with the trait tend to be relatively wide, spanning over 1 megabase (Mb) or more, and containing several genes. Before the completion of the human genome reference sequence[27], even determining what the genes in the region were was not a simple process, and to proceed from a mapped locus to a gene was very time-consuming and labour-intensive – for example, although Huntington's Disease was one of the earlier mapped diseases[28], it was not until 10 years later that the gene was cloned[29]. Today, with the human reference sequence, this process has been sped up, but is not necessarily simpler. In order to pin-point the causal mutations in the causal gene, investigators typically prioritise the genes in the region, if their function is known, and then Sanger sequence the genic regions in the locus in cases to look for potentially loss-of-function alleles – nonsense or frameshift insertions and deletions (indels), or those that alter

proteins – missense, splice affecting variants and in-frame indels, that could be responsible for the phenotype. Potential mutations are then screened in controls to identify those specific to cases and/or carriers, and ideally, each pedigree will have a mutation in a gene, common across pedigrees, that is functionally relevant and expressed in the relevant tissues.

It is worth noting that although linkage analysis to identify the causal genes for Mendelian disorders is popular, other approaches are also possible. For example, the gene responsible for haemophilia A was determined based on rescue of the clotting function of blood by a "globulin" isolated from normal blood (reviewed in ref. 30). Also, a candidate gene list can be determined based on function instead of location (as with linkage analysis), and cases and controls sequenced directly for potential mutations. In an example that will be referred to later in this dissertation, the mutations causal for Freeman-Sheldon syndrome (FSS), a distal arthrogryposis disorder, were identified in a screen of myosin heavy chain genes[31], these genes being chosen on the hypothesis that FSS was caused by mutations in contractile proteins, particularly those expressed during development. It is notable that this last candidate gene approach is made possible again only with the publication of the human genome reference sequence[27] and associated gene annotations, such that gene families or proteins with similar domains can be rapidly accessed based on sequence information.

While many rare disorders are highly amenable to linkage analysis, however, some disorders present a challenge for these methods. First, those which are extremely rare have only a few affected individuals and families per disorder, which result in underpowered analyses and/or large regions under the linkage peak(s). Second, these disorders are rare because the causal mutations are of large effect and under strong negative selection. As such, these mutations are not often transmitted through many generations and are, in fact, likely to be *de novo* events. Since linkage analysis is completely inheritance-dependent, these events are not ascertained at all by

such. To circumvent these challenges, it is necessary to identify these mutations directly through sequencing. Until recently, however, this has been highly resource-intensive and generally infeasible to do on a large scale or in a genome-wide manner.

1.3 "Next-generation" sequencing

Advances in sequencing technology have made it increasingly practical to generate large amounts of sequence data cost-effectively. Known as massively parallel or 'next-generation' sequencing, these technologies[32-35] enable investigators to obtain variant information down to single-base resolution in a rapid, high-throughput fashion on the scale of the whole human genome. In brief, a typical run consists of a complex DNA library being prepared, amplified and densely arrayed on a reaction plate or solid surface, in no particular order. This is followed by the incorporation of complementary bases by a polymerase or ligase, which is monitored by detection of fluorescent or chemical by-products. This enables the generation of a vast number of sequence reads (presently 10^5-10^8 per run) of length in the range of 100-500 nucleotides, again dependent on the technology, with more recent advances (e.g. ref. 36) showing the promise of read lengths upwards of a kilobase. These reads are typically aligned to a human genome reference sequence, and a consensus base called at positions which are covered by multiple reads with contributing bases of suitable quality.

Although the cost of sequencing has dropped dramatically, it is still not low enough to do whole-genome resequencing at a scale sufficient to power large scale genetic studies. Instead, it is more cost effective to focus on a relevant genomic subset, and sequence limited regions but for a larger sample size.

1.4 Target enrichment strategies

The challenge with only sequencing limited regions of the human genome is the need to first extract the regions of interest from a genomic DNA (gDNA) library, and there are a number of different methods by which this can be done.

The polymerase chain reaction (PCR) is the standard molecular technique to selectively amplify genomic regions of interest, with the largest scale projects amplifying all protein-coding exons[37-39]. PCR amplicons are a suitable input for massively parallel sequencers, and is feasible particularly if the number of regions amplified is modest (e.g. ref. 40). However, as the target size and number of amplicons increase, the efficiency – measured in terms of cost, feasibility and input requirements – of this method relative to that of the sequencing decreases rapidly. Multiplexing the PCR, both in terms of primers and samples, is one way to mitigate this disadvantage, and automated technology to do so has been developed[41].

Other approaches that are more matched in parallelism and cost have been described specifically to enrich massively parallel sequencing libraries for regions of interest. These can be broadly split into two categories – those based primarily on hybridization and those which include enzymatic activity to confer specificity.

In hybridization methods, oligonucleotide probes complementary to the regions of interest are designed and produced, incubated with a gDNA library, washed to remove non-binding DNA, and bound DNA eluted to result in a library enriched for target sequences. In the initial descriptions of this approach[42-44], DNA probes were made on programmable microarrays, capturing up to 10Mb of target nucleotide bases and requiring 20ug of input DNA. Later iterations of this method, including one described in this thesis (Chapter 2), allowed for capture of larger targets of up to 30Mb on a single, denser array (e.g., ref. 45), and more critically, moved the

capture reaction into solution by the use of biotinylated RNA probes[46] or DNA probes[47] which increased the scalability and specificity of the enrichment preparation.

There are two main enzymatic methods for capture – selective circularisation[48] and molecular inversion probes[49]. In the former, selector oligonucleotides consisting of a common internal double-stranded sequence with single-stranded flanking target-specific ends are incubated with a restriction digested gDNA library and hybridised target library molecules are ligated to the selectors by ligase, in some designs after endonucleaic cleavage of the 5' flap by *Taq* polymerase. The resultant circular molecules then undergo multi-template PCR with adaptor-primers to the common sequence to generate sequencer-appropriate libraries. The two enzymatic reactions required to make the circular molecules confer a high degree of specificity (~90%) to the capture reaction.

Molecular inversion probes (MIPs) follow a similar approach, starting with single-stranded oligonucleotides with a common internal sequence flanked by target-specific ends. These are incubated with the gDNA library, together with a ligase and a polymerase. The ends of the probes hybridise to the targets, forming a primer at the 5' end of the target for polymerisation of the target DNA, and a site for ligation at the 3' end of the target, resulting in circular molecules which are then PCR-ed with adaptors to produce sequencing libraries. Again, because of the two enzymatic steps, specificity is high and DNA input requirements are low, and with sufficient optimisation[50], the multiplex reaction can be tuned to be more uniform and efficient.

1.5 The "exome"

One genome-wide subset of interest is all the protein-coding regions (known as the "exome"), which comprises about 1-2% of the genome[51,52]. Particularly with respect to monogenic, Mendelian disease, the majority of known causal variants affect protein exons[53], which suggests

that Mendelian phenotypes are associated primarily with changes in protein sequence[54]. Hence, although focusing only on exome variants will give an incomplete picture of whole-genome variation, it will likely be sufficient to enrich for functional, disease-associated variants.

1.6 Interpretation of variant function

The major challenge remaining is the interpretation of these sequence data – how can background polymorphisms be distinguished from potentially disease-causing mutations? Typical mutation analyses on candidate genes prioritise variants based on segregation analyses, variant type, predicted effect on the protein or mRNA (for example, using Grantham scores[55], Polyphen [56,57] or SIFT[58]), and also by the prevalence of the mutation in the population[54,59]. The criteria of conservation and rare allele frequency has been used successfully to identify functional variants[60], while other studies have shown an excess of rare missense variants in candidate genes in individuals with extreme phenotypes[61,62]. Furthermore, various analyses suggest that alleles that are deleterious are more likely to be of lower frequency[63,64], as well as the converse, that rare missense variants are more likely to be deleterious[65]. Taken together, these suggest that searching for variants that are protein-altering, and rare or even unseen in the general population may be a useful starting point to identifying a disease gene.

1.7 Dissertation Aims

The general theme of this dissertation is to develop methods for the analysis of Mendelian disorders using massively parallel sequencing. Chapter 2 describes both the development of exome sequencing – specifically the improvements to capture the whole exome on a single microarray, as well as sequencing processing methods to more accurately call single-nucleotide variants as short indels – and also the approach to filter exome variant data from affected individuals to identify known causal mutations. The application of this approach to a rare

recessive disorder of unknown cause, Miller syndrome, is described in Chapter 3, and the application to a rare dominant disorder of unknown cause, Kabuki syndrome, is described in Chapter 4. The final chapter summarises this work and discusses current advances in the field.

CHAPTER 2 : TARGETED CAPTURE AND SEQUENCING OF 12 HUMAN EXOMES

This chapter was previously published as

Sarah B. Ng, Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E. Eichler, Michael Bamshad, Deborah A. Nickerson & Jay Shendure (2009) **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature.* 461(7261):272-6.

I performed the majority of the experimental work to generate exome libraries, with major help from Emily Turner in the development of the array capture protocol. Choli (Charlie) Lee and Michelle Wong performed the Illumina sequencing on the libraries. I also performed the majority of the algorithm development for probe design, short-read sequence processing and indel calling. Peggy Robertson and Tristan Shaffer designed a database for annotation of exome variants. Analysis of exome data was done by Jay Shendure and me, with contributions from Abigail Bigham and Michael Bamshad. The project was conceived by Emily Turner, Evan Eichler, Michael Bamshad, Deborah Nickerson, Jay Shendure and me.

2.1 Abstract

Genome-wide association studies suggest that common genetic variants explain only a modest fraction of heritable risk for common diseases, raising the question of whether rare variants account for a significant fraction of unexplained heritability[61,66]. Although DNA sequencing costs have fallen markedly[67], they remain far from what is necessary for rare and novel variants to be routinely identified at a genome-wide scale in large cohorts. We have therefore sought to develop second-generation methods for targeted sequencing of all protein-coding regions ('exomes'), to reduce costs while enriching for discovery of highly penetrant variants. Here we report on the targeted capture and massively parallel sequencing of the exomes of 12 humans. These include eight HapMap individuals representing three populations[68], and four unrelated individuals with a rare dominantly inherited disorder, Freeman–Sheldon syndrome (FSS)[31]. We demonstrate the sensitive and specific identification of rare and common variants in over 300 megabases of coding sequence. Using FSS as a proof-of-concept, we show that candidate genes for Mendelian disorders can be identified by exome sequencing of a small number of unrelated, affected individuals. This strategy may be extendable to diseases with more complex genetics through larger sample sizes and appropriate weighting of non-synonymous variants by predicted functional impact.

2.2 Exome sequencing of twelve individuals

Protein-coding regions constitute 1% of the human genome or 30 megabases (Mb), split across 180,000 exons. A brute-force approach to exome sequencing with conventional technology[38] is expensive relative to what may be possible with second-generation platforms[67]. However, the efficient isolation of this fragmentary genomic subset is technically challenging[69]. The enrichment of an exome by hybridization of shotgun libraries constructed from 140 mg of genomic DNA to seven microarrays was described previously[42]. To improve the practicality of

hybridization capture, we developed a protocol to enrich for coding sequences at a genome-wide scale starting with 10 mg of DNA and using two microarrays. Our initial target was 27.9Mb of coding sequence defined by CCDS (the NCBI Consensus Coding Sequence database)[51]. This curated set avoids the inclusion of spurious hypothetical genes that contaminate broader exome definitions[70]. The target is reduced to 26.6Mb on exclusion of regions that are poorly mapped with our anticipated read length owing to paralogous sequences elsewhere in the genome[a].

We captured and sequenced the exomes of eight individuals previously characterised by the HapMap[68] and Human Genome Structural Variation[71] projects. We also analysed four unrelated individuals affected with Freeman–Sheldon syndrome (FSS; Online Mendelian Inheritance in Man (OMIM) #193700), also called distal arthrogryposis type 2A, a rare autosomal dominant disorder caused by mutations in MYH3[31]. Unpaired, 76 base-pair (bp) reads[34] from post-enrichment shotgun libraries were aligned to the reference genome[72]. On average, 6.4 gigabases

Table 1 : Sequencing of twelve exome-enriched shotgun libraries

Individual	Mapped Bases	On Target Bases	Near Target Bases	Capture Specificity
NA18507 (YRI)	6,607,484,688	2,888,661,709	661,786,591	54%
NA18517 (YRI)	6,494,342,272	2,545,603,815	595,560,089	48%
NA19129 (YRI)	6,297,755,808	2,643,637,650	580,152,618	51%
NA19240 (YRI)	5,986,557,544	2,680,726,944	570,797,168	54%
NA18555 (CHB)	6,006,434,128	2,367,312,059	533,865,581	48%
NA18956 (JPT)	6,696,487,148	2,872,329,417	681,969,727	53%
NA12156 (CEU)	5,807,479,732	1,776,298,679	389,939,201	37%
NA12878 (CEU)	7,412,509,748	3,006,930,065	691,105,599	50%
FSS10066 (Eur)	6,213,695,628	2,724,817,939	522,381,165	52%
FSS10208 (Eur)	6,828,499,072	2,779,965,715	545,984,133	49%
FSS22194 (Eur)	6,710,279,780	2,139,816,034	523,364,262	40%
FSS24895 (Eur)	5,806,226,492	2,472,076,217	528,472,867	52%
Average	6,405,646,003	2,574,848,020	568,781,583	49%

Summary statistics on massively parallel sequencing are shown. All data was collected as unpaired 76 bp reads on the Illumina Genome Analyzer II platform (~10 lanes per individual). For each individual, we show the total number of mapped bases (Maq mapping score > 0), the number of these that align within ("On Target Bases") or near ("Near Target Bases") the 164,007 targeted intervals. Near target bases do not fall within a target, but are from reads that directly overlap a target. Capture specificity is calculated as the fraction of reads overlapping a target. CEU, CEPH HapMap; CHB, Chinese HapMap; Eur, European–American ancestry (non-HapMap); JPT, Japanese HapMap; YRI, Yoruba HapMap.

[a] Supplementary data 1 : A text file listing intervals excluded from consideration based on poor anticipated mappability with 76 bp single-end reads is found at http://www.nature.com/nature/journal/v461/n7261/extref/nature08250-s2.txt

(Gb) of mappable sequence was generated per individual (20-fold less than whole genome sequencing with the same platform[34]), and 49% of reads mapped to targets (Table 1). After removing duplicate reads that represent potential polymerase chain reaction artefacts[73], the average fold-coverage of each exome was 51× (Figure 1). On average per exome, 99.7% of targeted bases were covered at least once, and 96.3% (25.6 Mb) were covered sufficiently for variant calling (≥8× coverage and Phred-like[74] consensus quality ≥30). This corresponded to 78% of genes having >95% of their coding bases called[b] (Figure 2). The average pairwise correlation coefficient between individuals for gene-by gene coverage was 0.87, consistent with systematic bias in coverage between individual exomes.

False positives and false negatives are critical issues in genomic

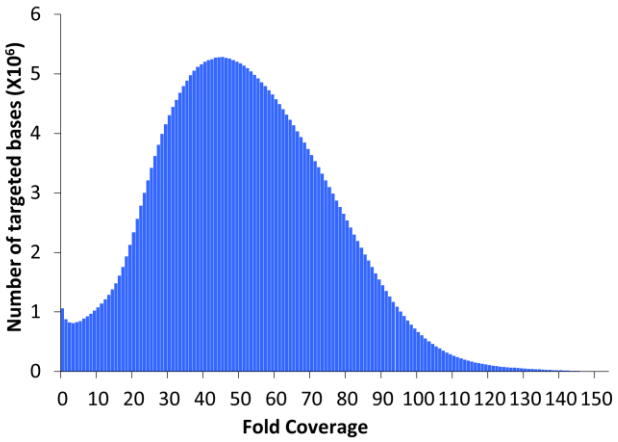

Figure 1 : Non-cumulative histogram of fold-coverage across twelve exomes.

The distribution of fold-coverage of mappable, targeted bases (26.6 Mb), summed across the twelve exome datasets (318 Mb aggregate target), is shown. As potential PCR duplicates (reads with the same start-point and orientation within the same genomic library) have been filtered out, the maximum possible coverage of any given position is 152X (i.e. reads from only 76 potential start-points X 2 orientations).

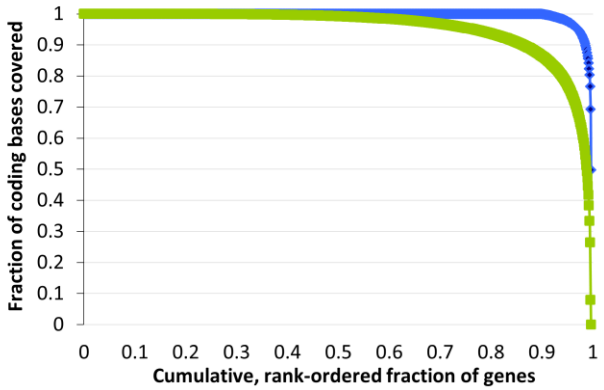

Figure 2 : Distribution of completeness on a gene-by-gene basis

We calculated the fraction of coding bases covered at least 1x (blue) or with sufficient coverage to variant call (green) on a gene-by-gene basis, for 197,952 genes (16,496 genes X 12 individuals). The cumulative, rank-ordered fraction of genes that meet minimum criteria are plotted above. For example, 98% of genes ≥1X coverage for at least 95% of their coding bases, while 78% of genes had sufficient coverage for variant calling for at least 95% of their coding bases.

[b] Supplementary data 2 : A text file listing the fraction of targeted coding bases in each gene that was covered in each of the twelve individuals (either with ≥ 1X coverage or with sufficient coverage to variant call) is found at http://www.nature.com/nature/journal/v461/n7261/extref/nature08250-s3.txt

Table 2 : Sequence coverage and array validation

Individual	≥1X Covered	Sequence called	Concordance with Illumina Human1M-Duo calls		
			Homozygous reference	Heterozygous	Homozygous non-reference
NA18507 (YRI)	26,477,161 (99.7%)	25,795,189 (97.1%)	23757/23762 (99.98%)	5553/5583 (99.46%)	3582/3592 (99.72%)
NA18517 (YRI)	26,476,761 (99.7%)	25,748,289 (97.0%)	23701/23705 (99.98%)	5575/5601 (99.54%)	3568/3579 (99.69%)
NA19129 (YRI)	26,491,035 (99.8%)	25,733,587 (96.9%)	23701/23708 (99.97%)	5482/5510 (99.49%)	3681/3690 (99.76%)
NA19240 (YRI)	26,486,481 (99.7%)	25,576,517 (96.3%)	23546/23551 (99.98%)	5600/5634 (99.40%)	3542/3549 (99.80%)
NA18555 (CHB)	26,475,665 (99.7%)	25,529,861 (96.1%)	23980/23984 (99.98%)	4877/4893 (99.67%)	3776/3786 (99.74%)
NA18956 (JPT)	26,454,942 (99.6%)	25,683,248 (96.7%)	24217/24221 (99.998%)	4890/4910 (99.59%)	3751/3760 (99.76%)
NA12156 (CEU)	26,476,155 (99.7%)	25,360,704 (95.5%)	23789/23794 (99.98%)	5493/5514 (99.62%)	3206/3213 (99.78%)
NA12878 (CEU)	26,439,953 (99.6%)	25,399,572 (95.6%)	23885/23891 (99.97%)	5413/5425 (99.78%)	3274/3292 (99.45%)
FSS10066 (Eur)	26,467,140 (99.7%)	25,546,738 (96.2%)	NA	NA	NA
FSS10208 (Eur)	26,461,768 (99.6%)	25,576,256 (96.3%)	NA	NA	NA
FSS22194 (Eur)	26,426,401 (99.5%)	25,454,551 (95.9%)	NA	NA	NA
FSS24895 (Eur)	26,478,775 (99.7%)	25,602,677 (96.4%)	NA	NA	NA

The number of coding bases covered at least 13 and with sufficient coverage to variant call (≥8X and consensus quality ≥30) are listed for each exome, with the fraction of the aggregate target (26.6 Mb) that this represents in parentheses. For the eight HapMap individuals, concordance with array genotyping (Illumina Human1M-Duo) is listed for positions that are homozygous for the reference allele, heterozygous or homozygous for the non-reference allele (according to the array genotype). NA, Not applicable.

resequencing. We assessed the quality of our exome data in four ways. First, comparing sequence-based calls for the eight HapMap exomes to array-based genotyping, we observed a high concordance with both homozygous (99.94%; n = 219,077) and heterozygous (99.57%; n = 43,070) genotypes (Table 2). Second, we compared our coding single-nucleotide polymorphism (cSNP) catalogue to ~1Mb of coding sequence determined in each of the eight HapMap individuals by molecular inversion probe (MIP) capture and direct resequencing[50]. At coordinates called in both data sets, 99.9% of all cSNPs (n = 4,620) and 100% of novel cSNPs (n = 334) identified here were concordant, consistent with a low false discovery rate. Third, we compared the NA18507 cSNPs identified here to those called by recent whole genome sequencing of this individual[34], and found substantial overlap (Figure 3). The relative numbers of cSNPs called by only one approach, and the proportions of these represented in dbSNP, indicate that exome sequencing has equivalent sensitivity for cSNP detection compared to whole genome sequencing. Fourth, we compared our data to cSNPs in high-quality Sanger sequence of single haplotype regions from fosmid clones of the same HapMap individuals[75]. Most fosmid defined cSNPs (38 of

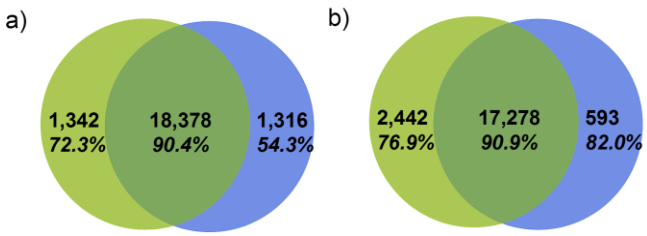

Figure 3 : Comparison of cSNPs from exome sequencing and whole genome sequencing of NA18507.

We identified 19,720 cSNPs by exome sequencing of this individual (green circles). In the above Venn diagrams, these are intersected with cSNPs from whole genome sequencing of this individual that overlapped with our exome target (Bentley et al., 2008) (blue circles). The percentage of each subset confirmed by dbSNP is given below the count. In (a), we used all cSNPs from Bentley et al. (2008) that were called by Maq (n = 19,694), whereas in (b), we used only the high-confidence cSNPs from Bentley et al. (2008), i.e. called by both Maq and Eland (n = 17,871). In (a), we observe similar numbers of variants that are only called in one data-set, suggesting both sets contain false negatives. The dbSNP-based confirmation rate is modestly higher for the 1,342 variants that are only called in our dataset as compared to the 1,316 variants only called from Bentley et al. (2008) (72.3% versus 54.3%). In (b), we call a much larger number of cSNPs than the high-confidence set from whole genome sequencing (19,720 vs. 17,871). The dbSNP-based confirmation rates for cSNPs called in only one dataset are similar (76.9% versus 82.0%). 78% of the 593 cSNPs called by Bentley et al. (2008) (Maq-Eland intersection) that were not identified here were at coordinates that had insufficient coverage to call in our data.

40) were at coordinates with sufficient coverage in our data for variant calling. Of these, 38 of 38 were correctly identified as variant.

A comparison of our data to past reports on exonic[44] or exomic[42] array-based capture revealed roughly equivalent capture specificity, but greater completeness in terms of coverage and variant calling (Table 3). These improvements probably arise from a combination of greater sequencing depth and differences in array designs and in experimental conditions for capture. Within the set of called positions, the high concordance with heterozygous array based genotypes (>99%) provides an estimate of our sensitivity for rare variant detection, as rare variants are overwhelmingly expected to be heterozygous. However, sensitivity was limited in that ~4% of known heterozygous genotypes were at coordinates where there was insufficient coverage to

Table 3 : Comparison to past reports on exonic[44] or exomic[42] array-based capture.

	Albert *et al.*2007	Hodges *et al.* 2007	This study
Reads mapping to exon target	36-76%	36-55%	37-54%
Target bases covered ≥ 1X	91.3-98.2%	25%	99.5-99.8%
Estimated sensitivity for variant calling	62.9-87.8%	60%*	955-97.1%

For Albert *et al.*, metrics for capture of 6,726 exonic regions are taken from Supplementary Tables 1 & 2, and we assume that at least 2 non-reference observations would be required to call a variant. For Hodges *et al.*, metrics for all-exome capture are taken from the text or Table 1. *reported for one of seven arrays only (EC5).

make a confident call.

There were 56,240 cSNPs called in one or more individuals, of which 13,347 were novel. On average, 17,272 cSNPs were called per individual, of which 92% were already annotated in a public database (dbSNP v129) (Table 4a). The proportion of previously annotated cSNPs was consistent by population, and higher for European (94%; n = 6) and Asian (93%; n = 2) than Yoruba (88%; n = 4) ancestry. These confirmation rates are ~10% higher than recent whole genome analyses[34,76-79]. The most likely explanation is that coding sequences have historically been more heavily

Table 4 : Coding variation across 12 human exomes

a) Summary statistics for observed cSNPs

Individual	cSNP calls	Number in dbSNP	Percentage in dbSNP	Number heterozygous	Number homozygous
NA18507 (YRI)	19,720	17,577	89.1%	12,896	6,824
NA18517 (YRI)	19,737	17,326	87.8%	13,039	6,698
NA19129 (YRI)	19,761	17,298	87.5%	12,845	6,916
NA19240 (YRI)	19,517	17,168	88.0%	12,866	6,651
NA18555 (CHB)	16,047	14,894	92.8%	9,181	6,866
NA18956 (JPT)	16,011	14,848	92.7%	9,132	6,879
NA12156 (CEU)	16,119	15,250	94.6%	10,179	5,940
NA12878 (CEU)	15,970	15,051	94.2%	9,928	6,042
FSS10066 (Eur)	16,229	15,144	93.3%	10,240	5,989
FSS10208 (Eur)	16,073	15,018	93.4%	9,966	6,107
FSS22194 (Eur)	16,094	15,128	94.0%	10,005	6,089
FSS24895 (Eur)	15,986	15,027	94.0%	9,920	6,066

b) Genome-wide estimates of cSNPs assuming a 30 Mb exome

Individual	Estimated total cSNPs	Estimated total heterozygous	Estimated total homozygous	Estimated total synonymous	Estimated total non-synonymous
NA18507 (YRI)	22,727	14,876	7,851	12,466	10,261
NA18517 (YRI)	22,841	15,135	7,706	12,550	10,291
NA19129 (YRI)	22,907	14,906	8,001	12,693	10,214
NA19240 (YRI)	22,814	15,063	7,751	12,565	10,249
NA18555 (CHB)	18,722	10,677	8,045	10,275	8,447
NA18956 (JPT)	18,523	10,585	7,938	10,072	8,451
NA12156 (CEU)	18,825	11,818	7,007	10,220	8,605
NA12878 (CEU)	18,544	11,455	7,089	10,110	8,434
FSS10066 (Eur)	18,836	11,795	7,041	10,240	8,596
FSS10208 (Eur)	18,591	11,444	7,147	10,075	8,516
FSS22194 (Eur)	18,667	11,539	7,128	10,144	8,523
FSS24895 (Eur)	18,508	11,466	7,042	10,169	8,339

For part a, cSNPs called in each individual, relative to the reference genome, are broken down by the fraction in dbSNP and by genotype. Part b shows extrapolation of observed numbers of cSNPs in each individual to an exactly 30Mb exome.

ascertained than noncoding sequences, although other factors such as dbSNP version, prior ascertainment of HapMap individuals and different false discovery rates may contribute as well. For the subset of cSNPs at coordinates with sufficient coverage for variant calling in all 12 individuals (n = 47,079), 32% of annotated variants and 86% of novel variants were singleton observations across 24 chromosomes (Figure 4a).

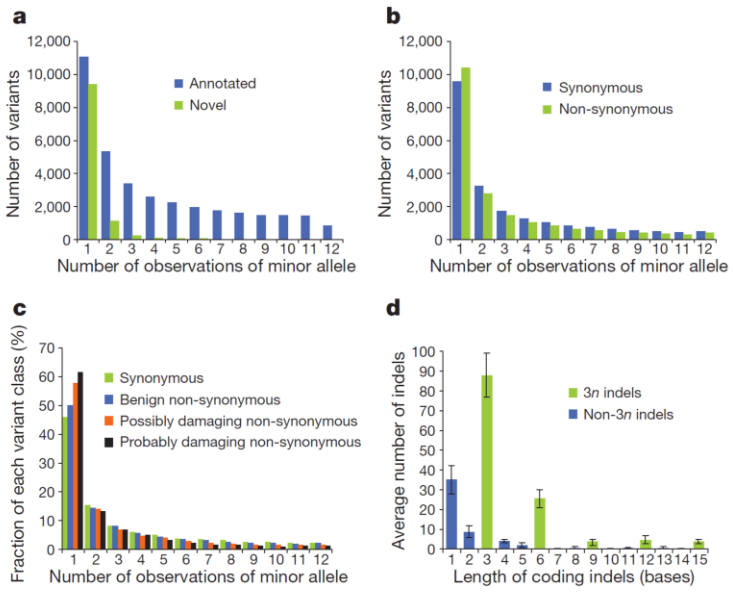

Figure 4 : Minor allele frequency and coding indel length distributions.

a, The distribution of minor allele frequencies is shown for previously annotated versus novel cSNPs. b, The distribution of minor allele frequencies is shown for synonymous versus non-synonymous cSNPs. c, The distribution of minor allele frequencies (by proportion, rather than count) is shown for synonymous cSNPs (n521,201) versus non-synonymous cSNPs predicted to be benign (n513,295), possibly damaging (n53,368), or probably damaging (n52,227) by PolyPhen[81]. d, The distribution of lengths of coding indel variants is shown (average numbers per exome). Error bars indicate s.d.

We also estimated the total number of cSNPs in each individual relative to the reference genome (Table 4b). As the precise and comprehensive definition of the human exome remains incomplete, we extrapolated our data to an estimated exome size of exactly 30 Mb. The results were remarkably consistent by population. As expected, a higher number of non-synonymous cSNPs were estimated for the Yoruba individuals (average 10,254; n = 4) than non-Africans (average 8,489; n = 8). More heterozygous cSNPs were estimated for the four Yoruba (average 14,995) than the six European Americans (average 11,586) and the two Asians (average 10,631). The ratio of synonymous to non-synonymous cSNPs was 1.2 within any single individual, and 1.1 when calculated for a non-redundant list of variants identified across all individuals. The difference results from the slightly shifted allele frequency distribution of non-synonymous variants (Figure

4b). Consistent with expectation[80], the trend is more pronounced for non-synonymous variants predicted to be damaging (by PolyPhen[81]) (Figure 4c).

Nonsense mutations and splice-site disruptions are often assumed to be deleterious, but have a broad range of potential fitness effects[82-84]. Our non-redundant cSNP catalogue included 225 nonsense mutations (112 novel) and 102 splice-site disruptions (49 novel). Excluding 86 nonsense alleles that are common in this data set (two or more observations) or in a recent study[82] (>5% allele frequency), our genome-wide estimate (projected to 30 Mb) for the average number of relatively rare mutations introducing premature nonsense codons in an individual genome was 10 for non-Africans (n = 8) and 20 for Yoruba (n = 4). However, these are probably overestimates, given that our catalogue of common nonsense mutations remains incomplete.

Short insertions and deletions (indels) in coding sequence are likely to be functionally important when they cause frameshifts, but are difficult to detect with short reads. We developed and applied an approach for identifying indels from our unpaired 76 bp reads. In total, 664 coding indels were called in one or more individuals. On average, 166 coding indels were called per

Table 5 : Coding indels across 12 human exomes

		Percentage			
Individual	**Number of coding indels**	**In dbSNP**	**Heterozygous**	**3n in length**	**Insertions**
NA18507 (YRI)	189	60%	61%	68%	40%
NA18517 (YRI)	204	58%	68%	70%	41%
NA19129 (YRI)	196	56%	65%	64%	42%
NA19240 (YRI)	183	57%	67%	70%	38%
NA18555 (CHB)	145	76%	53%	71%	47%
NA18956 (JPT)	139	71%	56%	71%	47%
NA12156 (CEU)	163	66%	61%	75%	48%
NA12878 (CEU)	146	64%	58%	66%	47%
FSS10066 (Eur)	170	59%	61%	70%	49%
FSS10208 (Eur)	165	59%	64%	70%	44%
FSS22194 (Eur)	152	66%	66%	68%	45%
FSS24895 (Eur)	142	65%	51%	65%	44%
Average	166	63%	61%	69%	44%

The number of called indels in each individual is listed, along with the proportion of these that are annotated in dbSNP (v129), the proportion that are heterozygous, the proportion that have a length that is a multiple of 3, and the proportion that are insertions as opposed to deletions, relative to the reference genome (UCSC hg18, NCBI Build 36.1).

individual, of which 63% were previously annotated in dbSNP (Table 5). To assess our sensitivity, we compared our data for NA18507 to data published previously[34]. The majority (73%) of their coding indels were also observed in our data (136 of 187). To assess specificity, we attempted PCR and Sanger sequencing of 28 novel coding indels chosen at random. Of 21 successful assays, 20 coding indels were verified and 1 was a false positive. We anticipate that future use of paired-end reads will improve detection of coding indels.

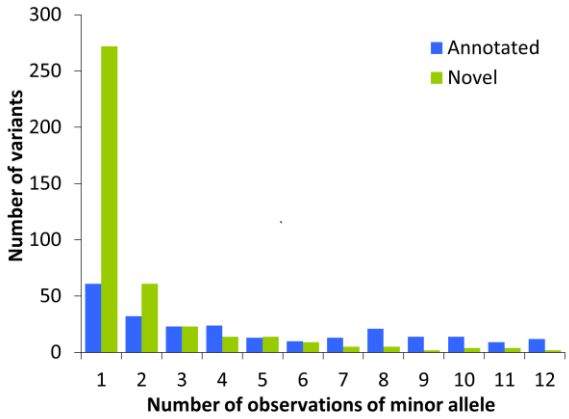

Figure 5: Minor allele frequencies for novel versus previously annotated coding indels.

A histogram of the number of observations of the minor allele for all observed coding indels is shown. Previously annotated (blue) and novel (green) coding indels are plotted separately.

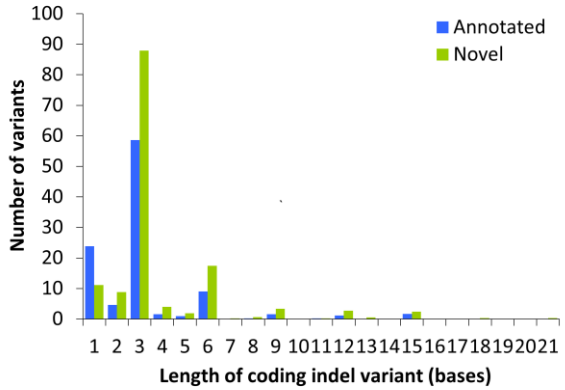

Figure 6 : Distribution of lengths of novel versus previously annotated coding indels.

The average number of coding indels called in each of the 12 exomes with various lengths is shown, plotted separately by annotation status.

The shape of the distribution of coding indel lengths was consistent with other studies[70,77] as well as across the 12 exomes (Figure 4d), demonstrating a preference for multiples of 3 ('3*n*'). Of the 664 coding indels observed here, 65% were 3*n* in length. The allele frequency distribution for novel indels relative to annotated indels was markedly shifted towards rarer variants (Figure 5). However, the length histograms for novel versus annotated coding indels were similar (Figure 6), reinforcing the notion that our set of novel coding indels is not excessively contaminated with false positives (as these would not be expected to have the observed 3n bias). Excluding indels that were common in this data set (two or more observations), the average

number of relatively rare frameshifting indels identified per individual was 8 for non-Africans (n = 8) and 17 for Yoruba (n = 4).

The number of synonymous, missense, nonsense, splice site, frameshifting indel and non-frameshifting indel variants observed in each individual (as well as the size of the subsets that are novel and singleton observations) is presented in Table 6. Also shown are the average numbers of variants of each class for non- Africans and Yoruba.

Table 6 : Number of variants observed in each individual

	cSNPs				Indels	
Individual	**Synonymous**	**Missense**	**Nonsense**	**Splice-site**	**In-frame**	**Frameshift**
NA18507 (YRI)	10817 (987,1620)	8862 (1143,1629)	41 (13,17)	18 (5,8)	128 (48,23)	61 (28,19)
NA18517 (YRI)	10845 (1111,1693)	8838 (1282,1644)	54 (18,18)	19 (6,9)	142 (60,24)	62 (26,16)
NA19129 (YRI)	10950 (1190,1699)	8761 (1258,1621)	50 (15,18)	18 (7,5)	125 (54,26)	71 (32,17)
NA19240 (YRI)	10749 (1109,1637)	8719 (1225,1582)	49 (15,16)	18 (8,7)	128 (56,22)	55 (23,15)
NA18555 (CHB)	8807 (489,653)	7198 (650,790)	42 (14,8)	21 (5,11)	103 (21,7)	42 (14,7)
NA18956 (JPT)	8706 (507,674)	7257 (643,785)	48 (13,11)	18 (1,3)	98 (30,13)	41 (11,8)
NA12156 (CEU)	8750 (355,540)	7322 (503,712)	47 (11,5)	15 (3,2)	123 (46,17)	40 (10,7)
NA12878 (CEU)	8706 (380,516)	7225 (529,632)	39 (10,10)	18 (8,10)	97 (33,13)	49 (20,5)
FSS10066 (Eur)	8822 (441,564)	7362 (629,703)	45 (15,12)	14 (2,2)	119 (49,18)	51 (21,12)
FSS10208 (Eur)	8709 (425,520)	7322 (617,668)	42 (13,8)	16 (2,6)	116 (48,18)	49 (20,12)
FSS22194 (Eur)	8745 (374,493)	7298 (580,652)	51 (12,13)	14 (2,3)	104 (34,10)	48 (18,6)
FSS24895 (Eur)	8783 (393,518)	7157 (553,579)	46 (13,8)	12 (6,5)	92 (31,7)	50 (19,9)
Average (YRI)	10840 (1099,1662)	8795 (1227,1619)	49 (15,17)	18 (7,7)	131 (55,24)	62 (27,17)
Average (Non-YRI)	8754 (421,560)	7268 (588,690)	45 (13,9)	16 (4,5)	107 (37,13)	46 (17,8)
Average	9449 (647,927)	7777 (801,1000)	46 (14,12)	17 (5,6)	115 (43,17)	52 (20,11)

A list of the number of observations of synonymous cSNPs, missense cSNPs, nonsense cSNPs, SNPs that disrupt canonical splice-site bases, in-frame coding indels, and frameshift indels. The first and second numbers in parentheses refer to novel variants (i.e. not in dbSNP) and singleton observations, respectively. Averages (Yoruba, non-African, and overall) are also shown.

2.3 Identification of causal mutations underlying a dominant Mendelian disorder

Phenotypes inherited in an apparently Mendelian pattern often lack sufficiently sized pedigrees to pinpoint the causal locus. We evaluated whether exome sequencing could be applied to identify directly the causative gene underlying a monogenic human disease (FSS), that is, with neither linkage data nor candidate gene analysis. Even in this simple scenario for 'whole exome/genome genetics', the key challenge that arises immediately is that the large number of apparently private mutations present by chance in any single human genome makes it difficult to

identify which variant is causal, even when only considering non-synonymous variants. This hurdle was overcome recently in the context of hereditary pancreatic cancer by restricting focus only to nonsense mutations and also resequencing tumour DNA from the same individual, but this approach greatly limits sensitivity and is only relevant to a subset of mechanisms within one disease class[39].

To quantify this background of non-causal variants in our exome data, we first investigated how many genes had one or more non-synonymous cSNPs, splice site disruptions or coding indels in one or several FSS exomes (Figure 7, row 1). Simply requiring that a gene contain variants in multiple affected individuals was clearly insufficient, as over 2,000 candidate genes remained even after intersecting four FSS exomes. We then applied filters to remove presumably common variants, as these are unlikely to be causative. Removing dbSNP-catalogued variants from consideration reduced the number of candidates considerably (Figure 7, row 2). Remarkably, the eight HapMap exomes provided a filter nearly equivalent to dbSNP (Figure 7, row 3). Combining the two catalogues had a synergistic effect (Figure 7, row 4), such that the candidate

Number of genes in which each affected has at least one...	FSS24895	FSS24895 FSS10208	FSS24895 FSS10208 FSS10066	FSS24895 FSS10208 FSS10066 FSS22194	ANY 3 OF 4 FSS24895 FSS10208 FSS10066 FSS22194
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
NS/SS/I not in dbSNP	513	128	71	53	119
NS/SS/I not in eight HapMap exomes	799	168	53	21	160
NS/SS/I neither in dbSNP nor eight HapMap exomes	360	38	8	1 (MYH3)	22
... and predicted to be damaging	160	10	2	1 (MYH3)	3

Figure 7 : Direct identification of the causal gene for a monogenic disorder by exome sequencing.
Boxes list the number of genes with one or more non-synonymous cSNP, splice-site SNP, or coding indel (NS/SS/I) meeting specified filters. Columns show the effect of requiring that one or more NS/SS/I variants be observed in each of one to four affected individuals. Rows show the effect of excluding from consideration variants found in dbSNP, the eight HapMap exomes, or both. Column five models limited genetic heterogeneity or data incompleteness by relaxing criteria such that variants need only be observed in any three of four exomes for a gene to qualify.

list could be narrowed to a single gene (MYH3, identified previously by a candidate gene approach as causative for FSS5). Specifically, MYH3 is the only gene where: (1) at least one (but not necessarily the same) non-synonymous cSNP, splice-site disruption or coding indel is observed in all four individuals with FSS; (2) the mutations are not in dbSNP, nor in the eight HapMap exomes. Taking the predicted deleteriousness of individual mutations into account served as an effective filter as well (Figure 7, row 5), but was not required to identify MYH3. Ranges of candidate list sizes when other permutations of individuals are used are shown in Figure 8. MYH3 was well covered in our data.

Number of genes in which each affected has at least one...	One Affected	Two Affecteds	Three Affecteds	Four Affecteds
Non-synonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510 – 4,617	3,284 - 3,373	2,765 – 2,808	2,479
NS/SS/I not in dbSNP	513 – 603	115 – 131	67 – 71	53
NS/SS/I not in eight HapMap exomes	799 – 912	150 – 191	49 – 54	21
NS/SS/I neither in dbSNP nor eight HapMap exomes	360 – 435	38	4 – 8	1 (MYH3)
...And predicted to be damaging	160 – 210	5 – 14	1 – 2	1 (MYH3)

Figure 8 : Direct identification of the causal gene - observation range.
Each box lists the number of genes with 1+ non-synonymous cSNP, splice-site variant, or coding indel ("NS/SS/I" variants) meeting specific criteria. In columns, we show the effect of requiring that 1+ NS/SS/I variants be observed in a given gene in all of 1, 2, 3 or 4 FSS-affected individuals. This figure is identical in format to Figure 7, except that we here provide ranges of observations that occur when all possible permutations of 1, 2, or 3 FSS-affected individuals are used.

To assess our sensitivity more globally, we calculated the probability that a mutation would have been identified in all four FSS-affected individuals for each gene, based on our overall coverage of that gene in each individual[c]. The average probability across all genes was 86%. This is probably still an overestimate of sensitivity, as functional noncoding or structural mutations would be missed. It also remains challenging to detect mutations in segmentally duplicated regions of the genome with short read sequencing. Nevertheless, our analysis suggests that direct sequencing of exomes of small numbers of unrelated individuals (but more than one) with a

[c] Supplementary data 2

shared monogenic disorder can serve as a genome-wide scan for the causative gene. The availability of the eight HapMap exomes was clearly helpful, suggesting that the power of this approach will improve as the 1000 Genomes Project[85] generates a catalogue of common variation that is more complete and evenly ascertained than dbSNP. Also, FSS is inherited in an autosomal dominant pattern so the presence of only one mutant allele is sufficient to cause disease. Applying this strategy to a recessive disease would probably be easier, because there are far fewer genes in each exome that are homozygous or compound heterozygous for rare non-synonymous variants. We also note that modelling of even a modest degree of genetic heterogeneity or data incompleteness is observed to have a significant impact on performance (Figure 7, column offset to the right). Moving along the spectrum from rare monogenic disorders to complex common diseases, it is likely that the increasing extent of genetic heterogeneity will need to be matched by increasingly large sample sizes30, and/or more sophisticated weighting of predicted mutational impact.

A clear limitation of exome sequencing is that it does not identify the structural and noncoding variants found by whole genome sequencing. At the same time, it allows a given amount of sequencing to be extended across at least 20 times as many samples compared to whole genome sequencing. In studies focused on identifying rare variants or somatic mutations with medical relevance, sample size and the interpretability of functional impact may be critical to achieving meaningful success. It is in the context of such studies that exome sequencing may be most valuable.

2.4 Conclusion

We demonstrate that targeted capture and massively parallel sequencing represents a cost-effective, reproducible and robust strategy for the sensitive and specific identification of variants causing protein-coding changes in individual human genomes. The 307Mb determined

here across 12 individuals is the largest data set reported so far of human coding sequence ascertained by second-generation sequencing methods. Finally, our successful demonstration that the causative gene for a Mendelian disorder can be identified directly by exome sequencing of several unrelated individuals provides increasing context to the possibility that exome or genome sequencing may represent a new approach for identifying gene–disease relationships.

2.5 Materials and methods

Genomic DNA samples. Targeted capture was performed on genomic DNA from eight HapMap individuals (four Yoruba (NA18507, NA18517, NA19129 and NA19240), two East Asians (NA18555 and NA18956) and two European- Americans (NA12156 and NA12878)) and four European-American individuals affected by Freeman–Sheldon syndrome (FSS10066, FSS10208, FSS22194 and FSS24895). Genomic DNA for HapMap individuals was obtained from Coriell Cell Repositories. Genomic DNA for FSS individuals was obtained by M.B.

Oligonucleotides and adaptors. All oligonucleotides were synthesised by Integrated DNA Technologies and resuspended in nuclease-free water to a stock concentration of 100 mM. Sequences are shown in Supplementary Table 5. Double-stranded library adaptors SLXA_1 and SLXA_2 were prepared to a final concentration of 50 mM by incubating equimolar amounts of

Table 7 : Sequences of oligonucleotides used for library construction or sequencing.

Oligonucleotide	Sequence	Function
SLXA_1_HI	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT	Library Adaptor
SLXA_1_LO	/5Phos/GAT CGG AAG AGC GTC GTG TAG GGA AAG AGT GT	Library Adaptor
SLXA_2_HI	CAA GCA GAA GAC GGC ATA CGA GCT CTT CCG ATC T	Library Adaptor
SLXA_2_LO	/5Phos/GAT CGG AAG AGC TCG TAT GCC GTC TTC TGC TTG	Library Adaptor
SLXA_FOR_AMP	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T	Library Amplification, Hybridization Blocker
SLXA_REV_AMP	CAA GCA GAA GAC GGC ATA CGA GCT CTT CCG ATC T	Library Amplification, Hybridization Blocker
SLXA_REV_AMP_rev	AGA TCG GAA GAG CTC GTA TGC CGT CTT CTG CTT G	Hybridization Blocker
SLXA_FOR_AMP_rev	AGA TCG GAA GAG CGT CGT GTA GGG AAA GAG TGT AGA TCT CGG TGG TCG CCG TAT CAT T	Hybridization Blocker

SLXA_1_HI and SLXA_1_LO together and SLXA_2_HI and SLXA_2_LO together at 95 °C for 3 min and then leaving the adaptors to cool to room temperature in the heat block.

Shotgun library construction. Shotgun libraries were generated from 10 µg of genomic DNA (gDNA) using protocols modified from the standard Illumina protocol[34]. Each library provided sufficient material for hybridization to two microarrays. For each sample, gDNA in 300 µl 1× Tris-EDTA was first sonicated for 30 min using a Bioruptor (Diagenode) set at high, then end-repaired for 45 min in a 100 µl reaction volume using 1× End-It Buffer, 10 µl dNTP mix and 10 µl ATP as supplied in the End-It DNA End-Repair Kit (Epicentre). The fragments were then A-tailed for 20 min at 70 °C in a 100 µl reaction volume with 1× PCR buffer (Applied Biosystems), 1.5 mM $MgCl_2$, 1mM dATP and 5U AmpliTaq DNA polymerase (Applied Biosystems). Next, library adaptors SLXA_1 and SLXA_2 were ligated to the A-tailed sample in a 90 µl reaction volume with 1× Quick Ligation Buffer (New England Biolabs) with 5 µl Quick T4 DNA Ligase (New England Biolabs) and each adaptor in 10× molar excess of sample. Samples were purified on QIAquick columns (Qiagen) after each of these four steps and DNA concentration determined on a Nanodrop- 1000 (Thermo Scientific) when necessary. Each sample was subsequently size-selected for fragments of size 150–250 bp using gel electrophoresis on a 6% TBE-polyacrylamide gel (Invitrogen). A gel slice containing the fragments of interest was then excised and transferred to a siliconized 0.5 ml microcentrifuge tube (Ambion) with a 20G needle-punched hole in the bottom. This tube was placed in a 1.5 ml siliconized microcentrifuge tube (Ambion), and centrifuged in a tabletop microcentrifuge at 16,110g for 5 min to create a gel slurry that was then resuspended in 200 µl 1× Tris-EDTA and incubated at 65 °C for 2 h, with periodic vortexing. This allowed for passive elution of DNA, and the aqueous phase was then separated from gel fragments by centrifugation through 0.2 µm NanoSep columns (Pall Life Sciences) and the DNA recovered using a standard ethanol precipitation. Recovered DNA was resuspended in elution buffer (EB; 10 mM Tris-Cl, pH

8.5, Qiagen) and the entire volume used in a 1 ml bulk PCR reaction volume with 1× iProof High-Fidelity Master Mix (Bio-Rad) and 0.5 μM each of primers SLXA_FOR_AMP and SLXA_REV_AMP with the conditions: 98 °C for 30 s, 20 cycles at 98 °C for 30 s, 65 °C for 10 s and 72 °C for 30 s, and finally 72 °C for 5 min. PCR products were purified across four QIAquick columns (Qiagen) and all the eluents pooled.

Design of exome capture arrays. We targeted all well-annotated protein-coding regions as defined by the CCDS (version 20080902). Coordinates were extracted from entries with 'public' status, and regions with overlapping coordinates were merged. This resulted in a target with 164,007 discontiguous regions summing to 27,931,548 bp. By comparison, coding sequence defined by all of RefSeq (NCBI 36.3) comprises 31.9Mb (14% larger). Hybridization probes against the target were designed primarily such that they were evenly spaced across each region. Probes were also constrained (1) to be relatively unique, such that the average occurrence of each 15-mer in the probe sequence is less than 100[42], (2) to be between 20 and 60 bases in length, with preference for longer probes, and (3) to have a calculated melting temperature (T_m)≤69 °C, with preference for higher T_m values. T_m was calculated by 64.9 + 41 × (number of G+Cs – 16.4)/length of probe. Two arrays (Agilent, 244K format) were designed and used per individual. The first array was common to all individuals, and contained 241,071 probes designed mainly against the subset of the target that was also found in a previous version of the CCDS (CCDS20070227). For most exomes, the second array was custom-designed specifically against target regions that had not been adequately represented after capture on the first array and subsequent sequencing. For two individuals (FSS10066, FSS10208), the matching was to a different individual's first-array data. However, this did not seem to have a significant effect on performance, probably because features capturing poorly on the first array largely did so consistently. Additionally, all of the second arrays

also targeted sequences found in CCDS20080902 that were not in CCDS20070227 and hence not targeted by the first array. A subset of arrays used lacked control grids.

Targeted capture by hybridization to DNA microarrays. Hybridizations to Agilent 244K arrays were performed following manufacturer's instructions with modifications. For each enrichment, a 520 µl hybridization solution containing 20 µg of the bulk-amplified genomic DNA library, 1× aCGH hybridization buffer (Agilent), 1× blocking agent (Agilent), 50 µg human CotI DNA (Invitrogen) and 0.92 nmol each of the blocking oligonucleotides SLXA_ FOR_AMP, SLXA_REV_AMP, SLXA_FOR_AMP_rev and SLXA_REV_ AMP_rev was incubated at 95 °C for 3 min and then at 37 °C for at least 30 min. The hybridization solution was then loaded and the hybridization chamber assembled following the manufacturer's instructions. Incubation was done at 65 °C for at least 66 h with rotation at 20 r.p.m. in a hybridization oven (Agilent). After hybridization, the slide-gasket sandwich was removed from the chamber and placed in a 50 ml conical tube filled with aCGH Wash Buffer 1 (Agilent). The slide was separated from the gasket while in the buffer and then washed, first with fresh aCGH Wash Buffer 1 at room temperature for 10 min on an orbital shaker (VWR) set on low speed, and then in pre-warmed aCGH Wash Buffer 2 (Agilent) at 37 °C for 5 min. Both washes were also done in 50 ml conical tubes. A Secure-Seal (SA2260, Grace Bio Labs) was then affixed firmly over the active area of the washed slide and heated briefly according to the manufacturer's instructions. One port was sealed with a seal tab and the seal chamber completely filled with approximately 1 ml of hot EB (95 °C). The other port was sealed and the slide incubated at 95 °C on a heat block. After 5 min, one port was unsealed and the solution recovered. DNA was purified from the solution using a standard ethanol precipitation. Precipitated DNA was resuspended in EB and the entire volume used in a 50 µl PCR volume comprising of 1× iTaq SYBR Green Supermix with ROX(Bio-Rad) and 0.2 µM each of primers SLXA_FOR_AMP and SLXA_REV_AMP. Thermal cycling was done in a MiniOpticon

Real-time PCR system (Bio-Rad) with the following programme: 95 °C for 5 min, then 30 cycles of 95 °C for 30 s, 55 °C for 2 min and 72 °C for 2 min. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. Samples were then purified on a QIAquick column (Qiagen) and sequenced.

Sequencing. All sequencing of post-enrichment shotgun libraries was carried out on an Illumina Genome Analyzer II as single-end 76 bp reads, following the manufacturer's protocols and using the standard sequencing primer. Image analysis and base calling was performed by the Genome Analyser Pipeline version 1.0 or 1.3 with default parameters, but with no pre-filtering of reads by quality. Quality values were recalibrated by alignment to the reference human genome with the Eland module.

Read mapping. The reference human genome used in these analyses was UCSC assembly hg18 (NCBI build 36.1), including unordered sequence (chrN_random.fa) but not including alternate haplotypes. For each lane, reads with calibrated qualities were extracted from the Eland export output. Base qualities were rescaled and reads mapped to the human reference genome using Maq (version 0.7.1)[72]. Unmapped reads were dumped using the –u option and subsequently used for indel mapping. Mapped reads that overlapped target regions ('target reads') were used for all other analyses.

Target masking. All possible 76-bp reads that overlapped the aggregate target were simulated, mapped using Maq and consensus called using Maq assemble with parameters *–q* 1 *–r* 0.2 *–t* 0.9. Target coordinates that had read depth < 76 (that is, half of the expected depth), reflecting a poor ability to have reads confidently mapped to them (Supplementary Data 1), were removed from consideration for downstream analyses, leaving a 26,553,795 bp target.

Variant calling. All reads with a map score of > 0 from each individual were merged and filtered for duplicates such that only the read with the highest aggregate base quality at any given start

position and orientation was retained. Sequence calls were obtained using Maq assemble with parameters *–r* 0.2 *–t* 0.9, and only coordinates with at least 8× coverage and an estimated Phred-like consensus quality value of at least 30 were used for downstream variant analyses.

Comparison of sequence calls to array genotypes, dbSNP and whole genome sequencing. For the eight HapMap individuals, sequence calls were compared to array-based genotyping data (Illumina Human1M-Duo) provided by Illumina. We excluded from consideration genotyping assays where all eight individuals were called by the arrays as homozygous non-reference as well as the MHC locus at chromosome 6:32500001–33300000, as both sets are likely to be error enriched in the genotyping data. We downloaded dbSNP(v129) from ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/chr_rpts on 13 May 2008. Approximately 14.2 million non-redundant coordinates were defined by this file set. For comparison of NA18507 cSNPs to whole genome data, variant lists were obtained from Illumina[34].

Identification of coding indels. Reads for which Maq was unsuccessful in identifying an ungapped alignment were converted to fasta format and mapped to the human reference genome with cross_match (v1.080812, http://www.phrap.org), using parameters *–gap_ext* 21 *–bandwidth* 10 *–minmatch* 20 *–maxmatch* 24. Output options *–tags –discrep_lists –alignments –score_hist* were also set. Alignments with an indel were then filtered for those that: (1) had a score at least 40 more than the next best alignment; (2) mapped at least 75 bases of the read; (3) had no substitutions in addition to the indel; and (4) overlapped a target region. Reads from filtered alignments that mapped to the negative strand were then reverse-complemented and, together with the rest of the filtered reads, re-mapped with cross_match using the same parameters. This was to reduce ambiguity in called indel positions due to different read orientations. After the second mapping, alignments were re-filtered using the same criteria (1) to (4). For each sample, a putative indel event was called if at least two filtered reads covered the same event. A fasta file

containing the sequences of all called events ±75 bp, as well as the reference sequence at the same positions, was then generated for each individual. All the reads from each individual were then mapped to its 'indel reference' with Maq using default parameters. Reads that mapped multiple times (map score 0) or had redundant start sites were removed, after which the number of reads mapping to either the reference or the non-reference allele was counted for each individual and indel. An indel was called if there were at least eight non-reference allele reads making up at least 30% of all reads at that genomic position. Indels were called as heterozygous if non-reference alleles were 30–70% of reads at that position, and homozygous non-reference if > 70%.

Variant annotation. For cSNP annotation, we constructed a local server that integrates data from NCBI (including dbSNP and Consensus CDS files) and from UCSC Genome Bioinformatics. We also generated PolyPhen predictions[81] for all cSNPs identified here, using the PolyPhen Grid Gateway and Perl scripts supplied by I. Adzhubey. The server reads files with SNP locations and alleles, and produces annotation files available for download. Annotation includes dbSNP rs IDs, overlapping-gene accession numbers, SNP function (for example, whether coding missense), conservation scores, HapMap minor-allele frequencies and various protein annotations (sequence, position, amino acid changes with physicochemical properties and PolyPhen classification). Indels were considered annotated by dbSNP if an entry was found with the same allele (or reverse complemented) within 1 bp of the variant position. This was to allow for ambiguities in calling the indel position.

Calculation of genome-wide estimates. Extrapolated estimates for the genome-wide number of cSNPs of various classes (Table 4b) were calculated based on the number of cSNP calls in that individual, the estimated sensitivity for making a variant call in that individual at any given position within the aggregate target (based on the fraction of array-based genotypes of that class that were successfully called; calculated separately for heterozygous and homozygous non-

reference variants), and extrapolation to an estimated exome size of exactly 30Mb (that is, multiplying by 30/26.6 = 1.13). A similar approach was taken to estimate the genome-wide number of uncommon cSNPs introducing nonsense codons, starting with the number observed in each individual and extrapolating based on estimated sensitivity for heterozygote detection and an estimated exome size of exactly 30 Mb.

Freeman–Sheldon syndrome mutations. For FSS10066, FSS22194 and FSS24895, the identified mutation was a CRT at chromosome 17:10485359, and the corresponding amino acid change was R672H. For FSS10208, the mutation was CRT at chromosome 17:10485360, and the corresponding amino acid change was R672C.

2.6 Acknowledgements

For discussions or assistance with genotyping data, we thank P. Green, J. Akey, R. Patwardhan, G. Cooper, J. Kidd, D. Gordon, J. Smith, I. Stanaway and M. Rieder. For assistance with project management, computation, data management and submission, we thank E. Torskey, S. Thompson, T. Amburg, B. McNally, S. Hearsey, M. Shumway and L. Hillier. For Human1M-Duo genotype data on HapMap samples, we thank Illumina. Our work was supported in part by grants from the National Institutes of Health/National Heart Lung and Blood Institute, the National Institutes of Health/National Human Genome Research Institute, National Institutes of Health/National Institute of Child Health and Human Development, and the Washington Research Foundation. S.B.N. is supported by the Agency for Science, Technology and Research, Singapore. E.H.T. and A.W.B. are supported by a training fellowship from the National Institutes of Health/National Human Genome Research Institute. E.E.E. is an investigator of the Howard Hughes Medical Institute.

CHAPTER 3 : EXOME SEQUENCING IDENTIFIES THE CAUSE OF A MENDELIAN DISORDER

This chapter previously published as

Sarah B Ng*, Kati J Buckingham*, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, Jay Shendure &Michael J Bamshad (2010) **Exome sequencing identifies the cause of a Mendelian disorder.** *Nature Genetics*. 42(1):30-5.

* contributed equally to the work

I performed the exome capture and generation of libraries, which were sequenced by Choli (Charlie) Lee. I also did the post-sequencing processing of short-reads, called variants and annotated them. Kati Buckingham performed the follow-up Sanger sequencing to verify mutations in the discovery and follow-up cohort. Clinical work and sample collection was done by Ethylin Jabs and Michael Bamshad. Data analysis of exome data was performed by Kati Buckingham, Abigail Bigham, Jay Shendure and me. This project was conceived by Michael Bamshad, Deborah Nickerson and Jay Shendure.

3.1 Abstract

We demonstrate the first successful application of exome sequencing to discover the gene for a rare Mendelian disorder of unknown cause, Miller syndrome (MIM%263750). For four affected individuals in three independent kindreds, we captured and sequenced coding regions to a mean coverage of 40× and sufficient depth to call variants at ~97% of each targeted exome. Filtering against public SNP databases and eight HapMap exomes for genes with two previously unknown variants in each of the four individuals identified a single candidate gene, *DHODH*, which encodes a key enzyme in the pyrimidine de novo biosynthesis pathway. Sanger sequencing confirmed the presence of *DHODH* mutations in three additional families with Miller syndrome. Exome sequencing of a small number of unrelated affected individuals is a powerful, efficient strategy for identifying the genes underlying rare Mendelian disorders and will likely transform the genetic analysis of monogenic traits.

3.2 Introduction

Rare monogenic diseases are of substantial interest because identification of their genetic bases provides important knowledge about disease mechanisms, biological pathways and potential therapeutic targets. However, to date, allelic variants underlying fewer than half of all monogenic disorders have been discovered. This is because the identification of allelic variants for many rare disorders is fundamentally limited by factors such as the availability of only a small number of affected individuals (cases) or families, locus heterogeneity, or substantially reduced reproductive fitness; each of these factors lessens the power of traditional positional cloning strategies and often restricts the analysis to a priori–identified candidate genes. In contrast, deep resequencing of all human genes for discovery of allelic variants could potentially identify the gene underlying any given rare monogenic disease. Massively parallel DNA sequencing technologies[67] have rendered the whole-genome resequencing of individual humans increasingly

practical, but cost remains a key consideration. An alternative approach involves the targeted resequencing of all protein-coding subsequences (that is, the exome)[42,86,87], which requires ~5% as much sequencing as a whole human genome[86].

Sequencing of the exome, rather than the entire human genome, is well justified as an efficient strategy to search for alleles underlying rare Mendelian disorders. First, positional cloning studies focused on protein- coding sequences have, when adequately powered, proven highly successful at identification of variants underlying monogenic diseases. Second, the clear majority of allelic variants known to underlie Mendelian disorders disrupt protein-coding sequences[88]. Splice acceptor and donor sites represent an additional class of sequences that are enriched for highly functional variation and are therefore targeted here as well. Third, a large fraction of rare non-synonymous variants in the human genome are predicted to be deleterious6. This contrasts with noncoding sequences, where variants are more likely to have neutral or weak effects on phenotypes, even in well-conserved noncoding sequences[65,89]. The exome therefore represents a highly enriched subset of the genome in which to search for variants with large effect sizes.

We recently showed how exome sequencing of a small number of affected, unrelated individuals could potentially be used to identify a causal gene underlying a monogenic disorder[86]. Specifically, we performed targeted enrichment of the exome by hybridization to programmable microarrays and then sequenced each enriched shotgun genomic library on an Illumina Genome Analyzer II. The exome was conservatively defined using the NCBI Consensus Coding Sequence (CCDS) database9 (version 20080902), which covers approximately 164,000 non-contiguous regions over 27.9 Mb, of which 26.6 Mb were 'mappable' using 76-bp single-end reads. Approximately 96% of targeted, mappable bases comprising the exomes of eight HapMap individuals and four individuals with Freeman-Sheldon syndrome (FSS; MIM#193700) were

successfully sequenced to high quality[86]. Using both dbSNP and HapMap exomes as filters to remove common variants, we showed that we could accurately identify the causal gene for FSS by exome sequencing alone. This effort demonstrated that low cost, high-throughput technologies for deep resequencing have the potential to rapidly accelerate the discovery of allelic variants for rare diseases. However, it provided only a proof of concept, as the causal gene for FSS had previously been identified[31]. A more recent report describes the application of exome sequencing to make an unanticipated genetic diagnosis of congenital chloride diarrhoea[87].

To evaluate the effectiveness of this strategy with a Mendelian condition of unknown cause, we sought to find the gene for a rare multiple malformation disorder named Miller syndrome[90], the cause of which has been intractable to standard approaches of discovery[91]. The clinical characteristics of Miller syndrome include severe micrognathia, cleft lip and/or palate, hypoplasia or aplasia of the posterior elements of the limbs, coloboma of the eyelids and supernumerary nipples (Figure 9a,b). Miller syndrome has been hypothesized to be an autosomal recessive disorder. However, only three multiplex families, each consisting of two affected siblings born to unaffected, nonconsanguineous parents, have been described among a total of ~30 reported cases of Miller syndrome for which substantial clinical information is available[90,92-96]. Accordingly, there has been speculation that Miller syndrome is an autosomal dominant disorder[97] and the rare occurrence of affected siblings is the result of

Figure 9 : Clinical characteristics of an individual with Miller syndrome and an individual with methotrexate embryopathy.
(**a**,**b**) A 9-year-old boy with Miller syndrome caused by mutations in *DHODH*. Facial anomalies (**a**) include cupped ears, coloboma of the lower eyelids, prominent nose, micrognathia and absence of the fifth digits of the feet (**b**). (**c**,**d**) A 26-year-old man with methotrexate embryopathy. Note the cupped ears, hypertelorism, sparse eyebrows and prominent nose (**c**) accompanied by absence of the fourth and fifth digits of the feet (**d**). **c** and **d** are reprinted with permission from ref. 107.

germline mosaicism. Although we thought it likely that Miller syndrome is recessive, we also considered a dominant model of inheritance.

3.3 Results

Exome sequencing identifies a candidate gene for Miller syndrome

We sequenced exomes in a total of two siblings with Miller syndrome (kindred 1 in Table 9) and two additional unrelated affected individuals (kindreds 2 and 3 in Table 9), totalling four affected individuals in three independent kindreds. An average of 5.1 Gb of sequence was generated per affected individual as single-end, 76-bp reads. After discarding reads that had duplicated start sites, we achieved ~40-fold coverage of the 26.6-Mb mappable, targeted exome defined by Ng *et al.*[86] (Table 8). About 97% of targeted bases were sufficiently covered to pass our

Table 8 : Summary statistics for exome sequencing of four individuals with Miller syndrome.

	Sequencing reads				Called coverage		
Kindred -sibling	Total	Uniquely mapping	Overlapping target	Non-duplicated	Mean coverage	Called bases	Percentage of CCDS
1-A	62,974,440	52,854,115	25,267,592	17,872,660	36.85	25,720,216	97%
1-B	72,539,306	61,940,123	40,335,280	21,971,509	44.24	25,825,104	97%
2	63,839,828	55,022,098	29,987,198	19,686,779	40.31	25,790,427	97%
3	68,690,600	57,970,901	36,180,596	19,649,281	39.81	25,617,361	96%

The total number of unpaired 76-bp sequencing reads per individual is reported (total), along with the number that map uniquely to the human genome (uniquely mapping, Maq map score > 0), the number that overlap at least one base of the target space (overlapping target) and the number left after removing reads with duplicate start sites (non-duplicated). Mean coverage over the whole of CCDS2008 is also given. Called bases refer to bases passing quality and coverage thresholds (Maq consensus quality ≥20 and read depth ≥8×). % of CCDS refers to the fraction of the mappable 26.6 Mb of CCDS2008 (that is, masked for poorly mappable coordinates, as described previously[86] that is called in each exome.

thresholds for variant calling. To distinguish potentially pathogenic mutations from other variants, we focused only on non-synonymous (NS) variants, splice acceptor and donor site mutations (SS), and short coding insertions or deletions (indels; I), anticipating that synonymous variants would be far less likely to be pathogenic. We also predicted that the variants responsible for Miller syndrome would be rare and therefore likely to be previously unidentified. A new variant was defined as one that did not exist in the datasets used for comparison, namely

dbSNP129, exome data from eight HapMap individuals sequenced in our previous study[86], and both groups combined (Table 9).

Each sibling (A and B) in kindred 1 was found to have at least a single NS/SS/I variant in ~4,600 genes and two or more NS/SS/I variants in ~2,800 genes. In our dominant model, each sibling was required to have at least one new NS/SS/I variant in the same gene, and filtering these variants against dbSNP129 and eight HapMap exomes reduced the candidate gene pool ~40-fold compared to the full CCDS gene set. In our recessive model, each sibling was required to have at least two new NS/SS/I variants in the same gene, and the candidate pool was reduced >500-fold compared to the full CCDS gene set. Both siblings were predicted to share the causal variant for Miller syndrome, so we next considered candidate genes shared between them. Under our dominant model, this reduced the pool of candidate genes to 228, and under our recessive model, the number of candidate genes was reduced to 9.

To further exclude candidate genes containing non-pathogenic variants, we next compared the candidate genes from both siblings in kindred 1 to those in two unrelated

Table 9 : Direct identification of the gene for a Mendelian disorder by exome resequencing.

	Kindred 1-A		Kindred 1-B		Kindred 1 (A+B)		Kindreds 1+2		Kindreds 1+2+3	
Filter	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive
NS/SS/I	4,670	2,863	4,687	2,859	3,940	2,362	3,099	1,810	2,654	1,525
Not in dbSNP129	641	102	647	114	369	53	105	25	63	21
Not in HapMap 8	898	123	923	128	506	46	117	7	38	4
Not in either	456	31	464	33	228	9	26	1*	8	1*
Predicted damaging	204	6	204	12	83	1	5	0	2	0

Each cell indicates the number of genes with non-synonymous (NS) variants, splice acceptor and donor site mutations (SS) and coding indels (I). Filtering either by requiring the presence of NS/SS/I variants in siblings (kindred 1 (A+B)) or of multiple unrelated individuals (columns) or by excluding annotated variants (rows) identifies 26 and 8 candidate genes under a dominant model and only a single candidate gene, *DHODH*, under a recessive model (light gray cells). Exclusion of mutations predicted to be benign using PolyPhen (row 5) increases sensitivity under a dominant model but excludes *DHODH* under a recessive model because a variant in kindred 1 is predicted to be benign. A single candidate gene is identified in kindred 1 under a recessive model and excluding benign mutations (dark gray cell), but this candidate is excluded in comparisons with unrelated cases of Miller syndrome. Mutations in this candidate, *DNAH5*, were found to cause a primary ciliary dyskinesia in kindred 1. The asterisk indicates that a second gene, *CDC27*, was also identified as a candidate gene, but this is due to the presence of multiple copies of a processed pseudogene that recurrently gave rise to a false positive signal in exome analyses.

individuals with Miller syndrome (kindreds 2 and 3). Using both dbSNP129 and the eight HapMap exomes as filters, comparison between the affected siblings in kindred 1 and the unrelated case in kindred 2 reduced the number of candidate genes to 26 under our dominant model. Under the autosomal recessive model, this comparison revealed that only a single gene, *DHODH*, which encodes the enzyme dihydroorotate dehydrogenase, was a shared candidate. Thus, comparison of exome data from two affected siblings and one unrelated affected individual was sufficient to identify *DHODH* as the sole candidate gene for Miller syndrome under our recessive model. Comparison between the siblings in kindred 1 and the unrelated cases in kindreds 2 and 3 reduced the number of candidate genes to eight under a dominant model, while retaining *DHODH* as the sole candidate under the recessive model.

We calculated a Bonferroni-corrected *P* value for the null hypothesis of seeing no deviation from the expected frequency of two variants in the same gene in three out of three unrelated, affected individuals over the ~17,000 genes in CCDS2008. Assuming all genes are of the same length and have the same mutation rate, the rate of new NS/SS/I variants per gene was 0.0309 (~526 new NS/SS/I variants per 17,000 genes). If the variants occur independently of one another, two variants occur in the same gene at a rate of $(0.0309)^2$, or 9.57×10^{-4}, so the *P* value is calculated as $((9.57 \times 10^{-4})^3 \times 17{,}000)$, or ~0.000015. Hence, even after correcting for searching across all genes, the result remains highly significant.

We also (i) examined the effect on the size of the candidate gene list when analysing the exomes of affected individuals in various pairwise or three-way combinations and (ii) examined the potential consequences of genetic heterogeneity by relaxing selection criteria such that only a subset of the exomes of affected individuals were required to contain new variants in a given gene for it to be considered as a candidate gene (Table 10). Heterogeneity clearly increases the number of candidate genes that must be considered under any fixed number of exomes analysed.

Table 10 : Number of candidate genes identified based on different filtering strategies.

	Number of affected exomes			Subsets of 3 exomes		Subsets of all 4 exomes		
	1	2	3	Any 1	Any 2	Any 1	Any 2	Any 3
Dominant model								
NS/SS/I	4,645–4,687	3,358–3,940	2,850–3,099	6,658	4,489	6,943	5,167	3,920
Not in dbSNP129	634–695	136–369	72–105	1617	274	1829	553	172
Not in HapMap 8	898–979	161–506	55–117	2336	409	2628	835	222
Not in either	453–528	40–228	10–26	1317	109	1516	333	44
Predicted damaging	204–284	10–83	3–6	682	37	787	126	11
Recessive model								
NS/SS/I	2,780–2,863	1,993–2,362	1,646–1,810	4,097	2,713	4,293	3,172	2,329
Not in dbSNP129	92–115	30–53	22–31	226	61	270	90	42
Not in HapMap 8	111–133	13–46	5–13	329	32	397	75	19
Not in either	31–45	2–9	2–3	100	6	121	14	4
Predicted damaging	6–16	0–2	0–1	35	2	44	4	1

Under the dominant model, at least one non-synonymous variant, splice acceptor or donor site variant or coding indel (NS/SS/I) in a gene was required in the gene. Under the recessive model, at least two novel variants were required, and these could be either at the same position (a homozygous variant) or at two different positions in the same gene (a potential compound heterozygote, though we are unable to ascertain phase at this stage). In each column are the range for the number of candidate genes for exomes considered individually (column 1) and all combinations of 2–4 exomes (columns 2–4). Note that the upper bound on the ranges may be inflated relative to what would be the case if four unrelated affected individuals had been used because the comparisons in which the two siblings were included provided reduced power compared to unrelated individuals. Columns 5–9 show the number of candidate genes when at least 1, 2 or 3 individuals is required to have one variant in a gene (dominant model) or two or more variants in a gene (recessive model). This is a simple model of genetic heterogeneity or incomplete data. For example, the total number of candidate genes common to any 3 of all 4 exomes is shown in column 9. For columns 5–6, one of the siblings (kindred 1-B) was not included in the analysis as siblings share 50% of variants.

However, this can likely be overcome by the inclusion of a greater number of cases with mutations in the same gene. Most variants underlying rare Mendelian diseases either affect highly conserved sequence and/or are predicted to be deleterious. Accordingly, we also sought to investigate to what extent the pool of candidate genes could be reduced by combining variant filtering with predictions of whether NS/SS/I variants were damaging. This strategy further reduced the pool of candidate genes for each of the comparisons made previously (Table 9). However, *DHODH* was not identified as a candidate under a recessive model in any of these comparisons. Review of predicted biophysical consequences of *DHODH* variants revealed that the effect of one variant, G605A, found in both siblings in kindred 1, was classified as benign. As a result, *DHODH* was eliminated from further consideration as a candidate under a recessive model

in kindred 1 and in all subsequent comparisons. However, because the other variant found in kindred 1, G454A, was predicted to be damaging, as was every other new *DHODH* variant identified in this study, *DHODH* was still one of only two candidate genes for Miller syndrome under a dominant model in a comparison of kindreds 1, 2 and 3 (Table 9). Nevertheless, the recessive model was favoured over the dominant model for Miller syndrome based on the observation that each case was a compound heterozygote for new *DHODH* mutations and five of six mutations were predicted to be damaging.

Combinatorial filtering supplemented by PolyPhen predictions initially identified a second candidate gene, *DNAH5*, in kindred 1 under a recessive model (Table 9). However, this candidate was excluded in subsequent comparisons to kindreds 2 and 3. *DNAH5* encodes a dynein heavy chain found in cilia, and mutations in *DNAH5* are a well-known cause of primary ciliary dyskinesia (PCD; MIM#608644), a disorder characterized by recurrent sinopulmonary infections, bronchiectasis and chronic lung disease. This was of particular interest because some of the clinical findings in the siblings in kindred 1 are unique among reported cases of Miller syndrome. Specifically, both siblings have recurrent lung infections, bronchiectasis and chronic obstructive pulmonary disease for which they have received medical management in a specialty clinic for

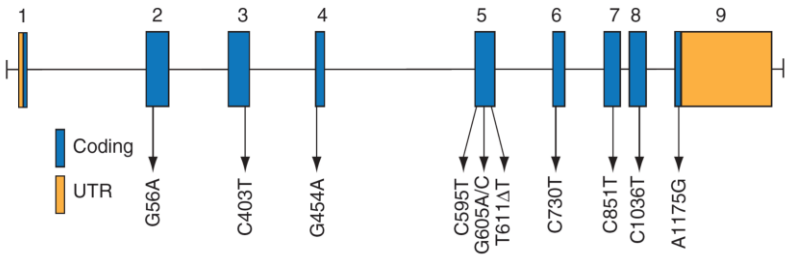

Figure 10 : Genomic structure of the exons encoding the open reading frame of *DHODH*.
DHODH is composed of nine exons that encode untranslated regions (UTR) (orange) and protein coding sequence (blue). Arrows indicate the locations of 11 different mutations found in 6 families with Miller syndrome.

individuals with cystic fibrosis. Accordingly, exome analysis revealed that both siblings in kindred 1 have, in addition to Miller syndrome, PCD due to mutations in *DNAH5*.

Sanger sequencing of implicated gene

To confirm that mutations in *DHODH* are responsible for Miller syndrome, we screened three additional unrelated kindreds (three simplex cases) and an affected sibling in kindred 2 by directed Sanger sequencing. All four individuals were found to be compound heterozygotes for missense mutations in *DHODH* that are predicted to be deleterious. Collectively, 11 different mutations in 6 kindreds with Miller syndrome were identified in *DHODH* by a combination of exome and targeted resequencing (Table 11 and Figure 10). Each parent of an affected individual who was tested was found to be a heterozygous carrier, and none of the mutations appeared to have arisen *de novo*. In the kindreds with affected siblings, none of the unaffected siblings were compound heterozygotes. None of these mutations were found in 200 control chromosomes from unaffected individuals of matched geographical ancestry that were genotyped. Ten of these mutations were missense mutations, two of which affected the same amino acid codon, and one was a 1-bp indel that is predicted to cause a frameshift resulting in a termination codon seven amino acids downstream. One mutation, C1036T, was shared between two unrelated individuals with Miller syndrome who are of different self-identified geographical ancestry. Each of the amino acid residues affected by a *DHODH* mutation is highly conserved among homologs studied to date (Figure 11). A single, validated non-

Table 11 : Summary of *DHODH* mutations in kindreds with Miller syndrome.

Kindred	Mutation	Exon	Amino acid change	Location
1	G454A	4	G152R	chr16: 70608443
	G605C	5	G202A	chr16: 70612611
2	C403T	3	R135C	chr16: 70606041
	C1036T	8	R346W	chr16: 70614936
3	C595T	5	R199C	chr16: 70612601
	611 delT	5	L204PfsX8	chr16: 70612617
4	G605A	5	G202D	chr16: 70612611
	C730T	6	R244W	chr16: 70613786
5	G56A	2	G19E	chr16: 70603484
	C1036T	8	R346W	chr16: 70614936
6	C851T	7	T284I	chr16: 70614596
	A1175G	9	D392G	chr16: 70615586

Mutations in kindreds 1-3 were originally identified by exome resequencing. Chromosomal position was determined using the March 2006 assembly from UCSC (hg18).

synonymous polymorphism in human *DHODH* has been studied previously[98]. This polymorphism causes a lysine-to-glutamine substitution in the relatively diverse N-terminal extension of dihydroorotate dehydrogenase that is responsible for the association of the enzyme with the inner mitochondrial membrane.

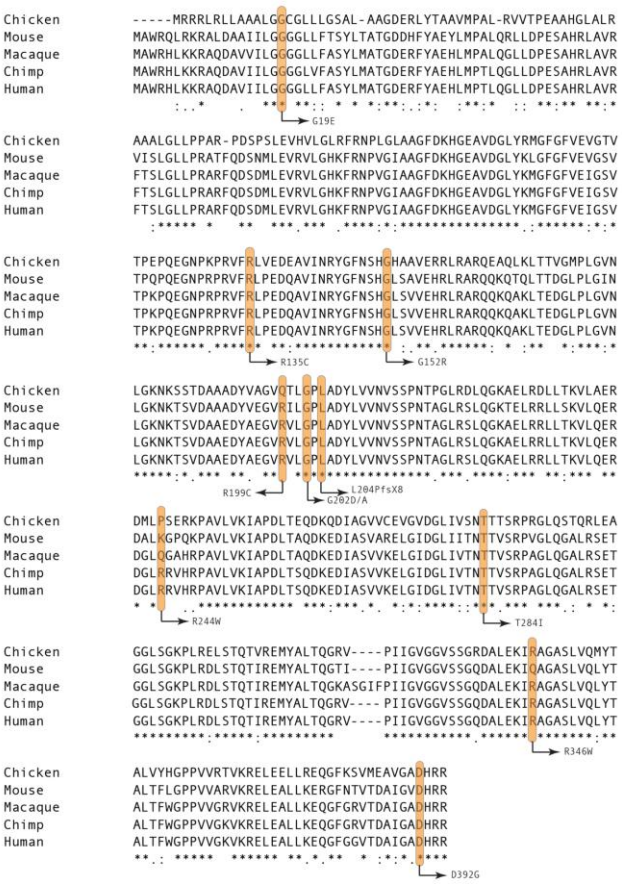

Figure 11 : Comparative protein alignment of dihydroorotate dehydrogenase in human, chimp, macaque, mouse and chicken.

The position of each variant identified in patients with Miller syndrome is boxed. Asterisks denote invariable sites and dots indicate conservative substitutions.

3.4 Discussion

We show that the sequencing of the exomes of affected individuals from a few unrelated kindreds, with appropriate filtering against public SNP databases and a small number of HapMap exomes, is sufficient to identify a single candidate gene for a monogenic disorder whose cause had previously been unknown, Miller syndrome. Several factors were important to the success of this study. First, Miller syndrome is a very rare disorder that is inherited in an autosomal recessive pattern. Therefore, the causal variants were unlikely to be found in public SNP databases or in control exomes. Second, genes for recessive diseases will, in general, be easier to find than genes for dominant disorders because fewer genes in any single individual have two or more new or rare non-synonymous variants. Third, we were fortunate that there was no genetic heterogeneity in our sample of individuals with Miller syndrome. In the presence of heterogeneity, it is possible to relax stringency by allowing genes common to subsets of all affected individuals to be considered

candidates, although this method will reduce power (Table 10). Fourth, all of the individuals with Miller syndrome for whom exomes were sequenced were of European ancestry. Sequencing exomes of affected individuals sampled from populations with a different geographical ancestry who have a higher number of novel and/or rare variants (for example, individuals with sub-Saharan African or East Asian ancestry) will make the identification of candidate genes more difficult. This will become less of an issue as databases of human polymorphisms become increasingly comprehensive.

Additional factors could facilitate the future application of this strategy. Mapping information, such as blocks of homozygosity, could focus the search to a smaller pool of candidates. The number of candidate variants can also be reduced further by comparison between variants in an affected individual to those found in each parent. For autosomal dominant disorders, this strategy can discover *de novo* coding variants, as neither parent is predicted to have a mutation that causes a fully penetrant dominant disorder; by contrast, in recessive disorders, parents are predicted to be carriers of the disease-causing variants.

There are at least three aspects of this approach where we see substantial scope for improvement. The first relates to missed variant calls, either due to low coverage or because some variants are not identified easily with current sequencing platforms—for example, those within repeat tracts in coding sequences. The second is that our filtering relied on a public SNP database (dbSNP) that has a highly uneven ascertainment of variation across the genome. It would be better to rely on catalogues of common variation that are ascertained in a single study either exome wide (as with the eight HapMap exomes[86]) or genome wide (for example, as with the 1,000 genomes project) and where estimates of allele frequency are available. Increasing the number of control exomes progressively reduces the relevance of dbSNP to this analysis (Figure 12). Furthermore, as increasingly deep catalogues of polymorphism become available, it may be

necessary to establish frequency-based thresholds for defining common variation that is unlikely to be causal for disease. A third concern is that the specificity of this approach is currently reduced by a subset of genes that recurrently appear to be enriched for new variants. These include long genes, but also genes that are subject to systematic technical artefacts (for example, mismapped reads due to duplicated or highly similar sequences in the genome). For sequences that are known to be duplicated or have paralogues (for example, genes from large gene families, or pseudogenes), these artefacts are mostly removed during read alignment (as reads with non-unique placements are removed from consideration). However, duplicated sequences not represented in the reference genome are not removed and spuriously appear as enriched for new variants (for example, *CDC27*).

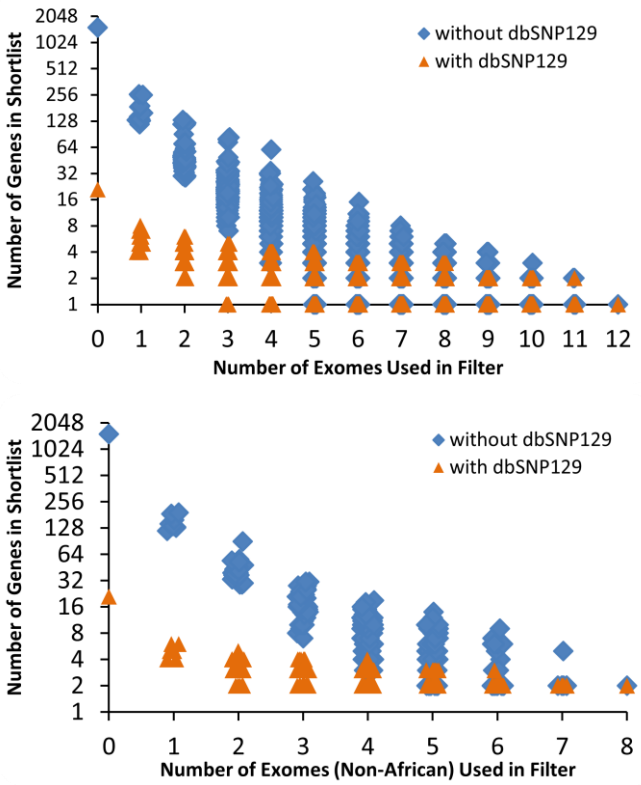

Figure 12 : Number of candidate genes vs. number of unaffected exomes in filter.

The results of adding more unaffected exomes as a filter for novel variants is shown for the recessive model, i.e. requiring at least two novel variants for a gene to qualify as a candidate. For this analysis, either all twelve previously sequenced exomes[86] were used (top), or only the eight non-Yoruba exomes from the same study (bottom), and results from all possible combinations of the specified number of exomes are represented. Results of the same filtering with (orange) and without dbSNP129 (blue) are also shown. The single gene identified when using all twelve exomes was *DHODH* (top); when only non-Yoruba exomes were used, the list also included *ESPNL*.

The mechanism by which mutations in *DHODH* cause Miller syndrome is unclear. The primary known function of dihydroorotate dehydrogenase is to catalyse the conversion of dihydroorotate to orotic acid, an intermediate in the pyrimidine *de novo* biosynthesis pathway

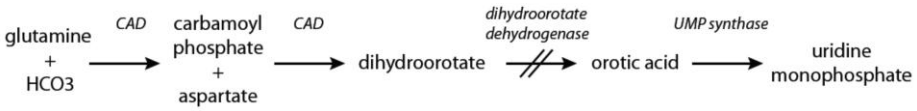

Figure 13 : Enzymatic steps controlling de novo pyrimidine synthesis.

Mutations in *DHODH* are predicted to diminish dihydroorotate dehydrogenase activity.

(Figure 13)[99]. Orotic acid is subsequently converted to uridine monophosphate (UMP) by UMP synthase. Pyrimidine biosynthesis might be particularly sensitive to the step mediated by dihydroorotate dehydrogenase[100], and the classical rudimentary phenotype in *Drosophila melanogaster*, reported by T.H. Morgan in 1910 and characterized by wing anomalies, defective oogenesis and malformed posterior legs, is caused by mutations affecting the same pathway[11,101,102]. However, the clinical characteristics of the other inborn errors of pyrimidine biosynthesis—such as orotic aciduria, caused by mutations in UMP synthase—do not include malformations. Indeed, inborn errors of metabolism are, in general, a rare cause of birth defects, so *DHODH* would be given little weight a priori as a candidate for a multiple malformation disorder. Thus, the discovery that mutations in *DHODH* cause Miller syndrome reveals both a new role for pyrimidine metabolism in craniofacial and limb development as well as a newly discovered function of dihydroorotate dehydrogenase that remains to be explored.

Selective inhibition of pyrimidine or purine biosynthesis has long been used as a therapeutic option to treat various cancers and autoimmune disorders. Leflunomide, a prodrug that is converted in the gastrointestinal tract to the active metabolite, A771726, reduces *de novo* pyrimidine biosynthesis by selectively inhibiting dihydroorotate dehydrogenase[100]. In mice, use of leflunomide during pregnancy causes a wide range of limb and craniofacial defects in the offspring, the most common of which are exencephaly, cleft palate and 'open eye' or failure of the eyelid to close[103]. These phenotypic characteristics recapitulate some of the malformations observed in individuals with Miller syndrome, providing further evidence that it is caused by mutations in *DHODH*.

The developmental pathways disrupted by leflunomide are unknown, but their elucidation could help us understand the mechanism by which *DHODH* mutations cause malformations. In the liver of mice treated with leflunomide, TNF-α production is repressed by the direct inhibition of NF-κB activity[104]. Interruption of NF-κB signalling during development can result in disrupted cell migration, diminished cellular proliferation and increased apoptosis[105]. Indeed, open-eye is a defect observed in mice with targeted disruption of *TNFA*28 . Furthermore, NF-κB has an important role in limb morphogenesis, specifically as a transducer of signals that regulate *Shh* (encoding the sonic hedgehog homolog) expression. *Shh* controls, in part, anterior-posterior patterning of the digits, and $Shh^{-/-}$ knockout mice fail to form digits 2–5 *(*ref. 106). These observations suggest that the malformations observed in individuals with Miller syndrome could be caused by perturbed NF-κB signalling due to loss of *DHODH* function.

The pattern of malformations observed in individuals with Miller syndrome is similar to those in individuals with foetal exposure to methotrexate (Figure 9c,d[107]). Methotrexate is a well-established inhibitor of *de novo* purine biosynthesis, and its anti-proliferative actions are thought to be due to its inhibition of dihydrofolate reductase and folate-dependent transmethylations. Accordingly, defects of both purine and pyrimidine biosynthesis appear to be capable of causing a similar pattern of birth defects. However, at low doses, methotrexate also decreases plasma levels of pyrimidines as well as purines. This observation raises the possibility that methotrexate embryopathy might indeed be caused by the drug's effects on pyrimidine rather than purine metabolism. Given that not all embryos exposed to methotrexate manifest birth defects, functional polymorphisms in *DHODH* or other genes encoding proteins in the *de novo* pyrimidine biosynthesis pathway could influence susceptibility to methotrexate embryopathy.

Individuals with Miller syndrome have similar phenotypic characteristics to those with Nager syndrome (MIM%154400), another rare monogenic disorder that primarily affects the

craniofacial skeleton. In contrast to Miller syndrome, the limb defects observed in individuals with Nager syndrome affect the anterior elements of the upper limb. Nevertheless, it has been hypothesized that Miller and Nager syndromes are caused by mutations in the same gene. We resequenced *DHODH* in 12 unrelated individuals diagnosed with Nager syndrome but found no pathogenic mutations (data not shown). Accordingly, either Nager syndrome and Miller syndrome are not allelic or Nager syndrome is caused exclusively by mutations in regulatory elements that alter the expression of *DHODH*.

Rare diseases are arbitrarily defined as those that affect fewer than 200,000 individuals in the United States. According to this definition, more than 7,000 rare diseases have been delineated, and in the aggregate, these affect more than 25 million people[108]. The majority of these diseases are considered genetic disorders, and many of them are thought to be monogenic. The bulk of genes underlying these rare monogenic diseases remain unknown. Lack of information about the genes and pathways that underlie rare monogenic diseases is a major gap in our scientific knowledge. Discovery of the genetic basis of a large collection of rare disorders that have, to date, been unyielding to analysis will substantially expand our understanding of the biology of rare diseases, facilitate accurate diagnosis and improved management, and provide initiative for further investigation of new therapeutics.

3.5 Conclusion

We have demonstrated that exome sequencing of a small number of affected family members or affected unrelated individuals is a powerful, efficient and cost-effective strategy for markedly reducing the pool of candidate genes for rare monogenic disorders and may even identify the responsible gene(s) specifically. This approach is likely to become a standard tool for the discovery of genes underlying rare monogenic diseases and to provide important guidance for developing an analytical framework for finding rare variants influencing risk of common disease.

3.6 Materials and methods

Patients and samples. For exome resequencing, we selected four individuals of self-reported European ancestry with Miller syndrome from three unrelated families. In two families, two siblings were affected (kindreds 1 and 2 in Table 9), and in one family a single individual had been diagnosed with Miller syndrome (kindreds 3 in Table 9). For validation, we selected samples from a sibling from kindred 2 and three simplex cases. All participants provided written consent, and the Institutional Review Boards of Seattle Children's Hospital and the University of Washington approved all studies. Separate informed consent was obtained from the individuals or their guardians to publish the photographs in Figure 9. A referral diagnosis of Miller syndrome made by a clinical geneticist was required for inclusion. The clinical characteristics of several of the individuals who had been diagnosed with Miller syndrome have been reported previously[90,92,93]. Phenotypic data were collected from review of medical records, phone interviews and photographs.

Targeted capture and massive parallel sequencing. Genomic DNA was extracted from peripheral blood lymphocytes, using Gentra Systems PUREGENE DNA purification kit and 10 µg of DNA from each of the four individuals with Miller syndrome in kindreds 1, 2 and 3 was used for construction of a shotgun sequencing library as described previously[86] using adaptors for single-end sequencing on an Illumina Genome Analyzer II (GAII). Each shotgun library was hybridized to two Agilent 244K microarrays for target enrichment, followed by washing, elution and additional amplification[86]. The first array targeted CCDS2007, whereas the second was designed against targets poorly captured by the first array plus updates to CCDS in 2008. Enriched libraries were then sequenced on a GAII.

Read mapping and variant analysis. Reads were mapped to the reference human genome (UCSC assembly hg18, NCBI Build 36.1), initially with efficient large-scale alignment of nucleotide

databases (ELAND) software (Illumina) for quality recalibration and then again with Maq[72]. Sequence calls were also performed by Maq and filtered to coordinates with ≥8× coverage and a Phred-like[72] consensus quality ≥20. Indels affecting coding sequence were identified as described previously[86]. Sequence calls were compared against eight HapMap individuals for whom we had previously reported exome data[86]. Annotations of variants were made using SeattleSeq Annotation[109] based on NCBI and UCSC databases, supplemented with PolyPhen Grid Gateway[110] predictions generated for nearly all non-synonymous SNPs. Any non-synonymous variant that was not assigned a 'benign' PolyPhen prediction was considered to be damaging, as were all splice acceptor and donor site mutations and all coding indels.

Mutation validation. Sanger sequencing of PCR amplicons from genomic DNA was used to confirm the presence and identity of variants in the candidate gene identified via exome sequencing and to screen the candidate gene in additional cases of Miller syndrome.

3.7 Acknowledgements

We thank the families for their participation and the Foundation of Nager and Miller Syndrome for their support. We thank M. McMillin for assistance with project coordination. We thank R. Scott, T. Cox, L. Cox, R. Jack, E. Eichler, M. Emond, G. Cooper, J. Kidd, R. Waterston and E. Wijsman for discussions. Our work was supported in part by grants from the National Heart, Lung, and Blood Institute, National Human Genome Research Institute and National Institute of Child Health and Human Development of the US National Institutes of Health, the Life Sciences Discovery Fund and the Washington Research Foundation. S.B.N. is supported by the Agency for Science Technology and Research, Singapore. A.W.B. is supported by a training fellowship from the National Human Genome Research Institute.

CHAPTER 4 : EXOME SEQUENCING IDENTIFIES *MLL2* MUTATIONS AS A CAUSE OF KABUKI SYNDROME

This chapter previously published as

Sarah B Ng*, Abigail W Bigham*, Kati J Buckingham, Mark C Hannibal, Margaret J McMillin, Heidi I Gildersleeve, Anita E Beck, Holly K Tabor, Gregory M Cooper, Heather C Mefford, Choli Lee, Emily H Turner, Joshua D Smith, Mark J Rieder, Koh-ichiro Yoshiura, Naomichi Matsumoto, Tohru Ohta, Norio Niikawa, Deborah A Nickerson, Michael J Bamshad & Jay Shendure (2011) **Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome.** *Nature Genetics* 42(9):790-3.

* contributed equally to the work

I performed the updated probe design and generation of exome libraries, which were sequenced by Choli (Charlie) Li. I also performed the post-sequencing processing of short reads, called variants and annotated them. Other experimental work was performed by Kati Buckingham, Heidi Gildersleeve, Anita Beck and Heather Mefford. Clinical work and sample collection was performed by Michael Bamshad, Mark Hannibal, Margaret McMillin, Koh-ichiro Yoshiura, Naomichi Matsumoto, Tohru Ohta, and Norio Niikawa. Analysis of exome data was performed by Abigail Bigham, Michael Bamshad, Greg Cooper, Jay Shendure and me. This project conceived by Michael Bamshad, Deborah Nickerson and Jay Shendure.

4.1 Abstract

We demonstrate the successful application of exome sequencing[86,87,111] to discover a gene for an autosomal dominant disorder, Kabuki syndrome (OMIM%147920). We subjected the exomes of ten unrelated probands to massively parallel sequencing. After filtering against existing SNP databases, there was no compelling candidate gene containing previously unknown variants in all affected individuals. Less stringent filtering criteria allowed for the presence of modest genetic heterogeneity or missing data but also identified multiple candidate genes. However, genotypic and phenotypic stratification highlighted *MLL2*, which encodes a Trithorax-group histone methyltransferase[112,113]: seven probands had newly identified nonsense or frameshift mutations in this gene. Follow-up Sanger sequencing detected *MLL2* mutations in two of the three remaining individuals with Kabuki syndrome (cases) and in 26 of 43 additional cases. In families where parental DNA was available, the mutation was confirmed to be *de novo* (n = 12) or transmitted (n = 2) in concordance with phenotype. Our results strongly suggest that mutations in *MLL2* are a major cause of Kabuki syndrome.

4.2 Introduction

Kabuki syndrome is a rare, multiple malformation disorder characterized by a distinctive facial appearance (Figure 14), cardiac anomalies, skeletal abnormalities, immunological defects and mild to moderate mental retardation. Originally described in 1981 (refs. 114,115), Kabuki syndrome has an estimated incidence of 1 in 32,000 (ref. 116), and approximately 400 cases have been reported worldwide. The vast majority of reported cases have been sporadic, but parent-to-child transmission in more than a half dozen instances[117] suggests that Kabuki syndrome is an autosomal dominant disorder. The relatively low number of cases, the lack of multiplex families and the phenotypic variability of Kabuki syndrome have made the identification of the gene(s)

Figure 14 : Photographs of the facial characteristics used to determine the subjective ranking of Kabuki phenotypes.

The phenotype ranking of the ten children with Kabuki syndrome is listed here from 1-10 based on similarity to the canonical phenotype of Kabuki syndrome. Asterisks indicate cases in which *MLL2* mutations were identified. Informed consent was obtained for publication of each of the facial photos shown.

underlying this disorder intractable to conventional approaches of gene discovery, despite aggressive efforts.

4.3 Results

We sequenced the exomes of ten unrelated individuals with Kabuki syndrome: seven of European ancestry, two of Hispanic ancestry and one of mixed European and Haitian ancestry

Table 12 : Clinical characteristics used to determine the subjective phenotype ranking of the 10 children with Kabuki syndrome.

	Cardiovascular	Spleen / Liver abnormality	Kidney abnormality / dysfunction	Hearing loss	Preauricular pits / tags	High arched / cleft	Hypotonia	Developmental delay
1	ASD/VSD, aortic coarctation, bicuspid valves, dysrhythmia	np	X	X	X	X	X	X
2	aortic coarctation, bicuspid valves	-	X	X	X	X	X	X
3	ASD/VSD, aortic coarctation, bicuspid valves	-	-	X	np	X	X	X
4	VSD, aortic coarctation	X	X	X	-	X	X	X
5	ASD, VSD	np	np	np	np	X	X	X
6	np	np	np	-	np	X	X	X
7	np	-	X	X	np	X	X	X
8	np	X	X	np	X	X	X	X
9	np	np	np	np	np	np	np	X
10	ASD, VSD, dysrhythmia	X	X	X	np	X	X	X

X denotes abnormality present; np denotes abnormality not present; - denotes data not found in medical record

(Figure 14 and Table 12). Enrichment was performed by hybridization of shotgun fragment libraries to custom microarrays followed by massively parallel sequencing[86,87,111]. On average, 6.3 gigabases of sequence were generated per sample to achieve 40× coverage of the mappable, targeted exome (31 Mb). As with our previous studies, we focused our analyses here primarily on non-synonymous variants, splice acceptor and donor site mutations and coding indels, anticipating that synonymous variants were far less likely to be pathogenic. We also predicted that variants underlying Kabuki syndrome are rare, and therefore likely to be previously unidentified. We defined variants as previously unidentified if they were absent from all datasets used for comparison, including dbSNP129, the 1000 Genomes Project, exome data from 16 individuals previously reported by us[86,111] and 10 exomes sequenced as part of the Environmental Genome Project (EGP).

Under a dominant model in which each case was required to have at least one previously unidentified non-synonymous variant, splice acceptor and donor site mutation or coding indel variant in the same gene, only a single candidate gene (*MUC16*) was shared across all ten exomes (Table 13). However, we considered *MUC16* as a likely false positive due to its extremely large size (14,507 amino acids). Potential explanations for our failure to find a compelling candidate gene in which newly identified variants were seen in all affected individuals included: (i) Kabuki syndrome is genetically heterogeneous and therefore not all affected individuals will have

Table 13 : Number of genes common to any subset of *x* affected individuals.

Subset analysis (any *x* of 10)	1	2	3	4	5	6	7	8	9	10
NS/SS/I	12,042	8,722	7,084	6,049	5,289	4,581	3,940	3,244	2,486	1,459
Not in dbSNP129 or 1000 genomes	7,419	2697	1057	488	288	192	128	88	60	34
Not in control exomes	7,827	2865	1025	399	184	90	50	22	7	2
Not in either	6,935	2227	701	242	104	44	16	6	3	1
Is loss-of-function (nonsense/frameshift indel)	753	49	7	3	2	2	1	0	0	0

The number of genes with at least one non-synonymous variant (NS), splice-site acceptor or donor variants (SS) or coding indel (I) are listed under various filters. Variants were filtered by presence in dbSNP or 1000 Genomes (not in dbSNP129 or 1000 genomes) and control exomes (not in control exomes) or both (not in either); control exomes refer to those from 8 Hapmap[86], 4 FSS[86], 4 Miller[111], and 10 EGP samples. The number of genes found using the union of the intersection of *x* individuals is given.

previously unidentified variants were shared in 3 genes, 6 genes and 16 genes, respectively (Table 13). However, there was no obvious way to rank these candidate genes.

We speculated that genotypic and/or phenotypic stratification would facilitate the prioritization of candidate genes identified by subset analysis. Specifically, we assigned a categorical rank to each individual with Kabuki syndrome based on a subjective assessment of the presence of, or similarity to, the canonical facial characteristics of Kabuki syndrome (Figure 14) and the presence of develo**p**mental delay and/or major birth defects (Table 12). The highest-ranked individual was one of a pair of monozygotic twins with Kabuki syndrome. We then categorized the functional impact (that is, nonsense versus non-synonymous substitution, splice-site disruption and frameshift compared to in-frame indel) of each newly identified variant in candidate genes shared by each subset of two or more ranked cases. Manual review of these data highlighted distinct, previously unidentified nonsense variants in *MLL2* in each of the four highest-ranked cases. After sequential analysis of phenotype-ranked cases with a loss-of-function filter, *MLL2* was the only candidate gene remaining after addition of the second individual (Table 14). We found no such variant in *MLL2* in the individual with Kabuki syndrome ranked fifth; hence, the number of candidate genes dropped to zero after the individual ranked fourth in the set (Table 14). However, we found a 4-bp deletion in the individual ranked sixth, and we found nonsense variants in the individuals ranked seventh and ninth. Thus, exome sequencing

Table 14 : Number of genes common in sequential analysis of phenotypically ranked individuals.

Sequential analysis	**1**	**+ 2**	**+ 3**	**+ 4**	**+ 5**	**+ 6**	**+ 7**	**+ 8**	**+ 9**	**+ 10**
NS/SS/I	5,282	3,850	3,250	2,354	2,028	1,899	1,772	1,686	1,600	1,459
Not in dbSNP129 or 1000 genomes	687	214	145	84	63	54	42	40	39	34
Not in control exomes	675	134	50	26	13	13	8	5	4	2
Not in either	467	89	34	18	9	8	4	4	3	1
Is loss-of-function (nonsense/frameshift indel)	25	1	1	1	0	0	0	0	0	0

Variants were filtered as in Table 13. Exomes were added sequentially to the analysis by ranked phenotype; for example, column "+3" shows the number of genes at the intersection of the three top ranked cases (Figure 14). The gene with at least one NS/SS/I in all individuals is *MUC16*, which is very likely to be a false positive due to its extreme length (14,507 amino acids).

identified a nonsense substitution or frameshift indel in *MLL2* in seven of the ten individuals with Kabuki syndrome analysed here. Retrospectively, we applied a loss-of-function filter to the subset analysis of exome data (Table 13), and at $x = 7$, found *MLL2* to be the only candidate gene.

In parallel with these analyses, we applied genomic evolutionary rate profiling (GERP)[118] to the exome data. GERP uses mammalian genome alignments to define a rejected substitution score for each variant regardless of functional class. We have previously shown that the quantitative ranking of candidate genes by the rejected substitution scores of their variants can facilitate the exome-based analysis of Mendelian disorders[119]. Following subset analysis with GERP-based ranking, *MLL2* remained on the candidate list up to $x = 8$, ranking third in a list of 11 candidate genes at this threshold (Table 15). Notably, the additional *MLL2* variant contributing to this analysis (such that *MLL2* was still considered at $x = 8$) was a synonymous substitution with a rejected substitution score of 0.368 in the individual ranked fifth.

We sought to confirm all newly identified variants in *MLL2*, particularly because loss-of-function variants identified through massively parallel sequencing have a high prior probability of being false positives. All seven loss-of-function variants in *MLL2* were validated by Sanger sequencing. We further analysed the three cases in which we did not initially find a loss-of-function variant in *MLL2*, first by array comparative genomic hybridization (aCGH) to determine any gross structural changes and then by Sanger sequencing of all exons of *MLL2* in case of false negatives by exome sequencing. Because an average of 96% of the coding bases in *MLL2* were called at sufficient quality and coverage for single nucleotide variant detection, we anticipated

Table 15 : Analysis of exome variants using genomic evolutionary rate profiling

GERP Score analysis (at least *x* of 10)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Variant RS score > 0	7,176	2,360	754	269	106	39	20	11	3	1
MLL2 Rank	3,732	1,232	399	136	47	14	6	3	NA	NA

The number of genes with at least a single previously unidentified variant with a rejected substitution score$_{10}$ > 0 in at least x individuals is given. A gene rank is assigned based on the average GERP score[118] over all newly identified variants with rejected substitution score > 0 in all affected individuals.

that any missed variants were more likely to be indels because of the higher coverage required for confident indel detection in short-read sequence data. Indeed, although aCGH did not find any structural variants in the region, Sanger sequencing did identify frameshift indels in two of these three cases (specifically, the cases ranked eighth and tenth).

Ultimately, loss-of-function mutations in *MLL2* were identified in nine out of ten cases in the discovery cohort (Figure 15), making this gene a compelling candidate for Kabuki syndrome. For validation, we screened all 54 exons of *MLL2* in 43 additional cases by Sanger sequencing. Previously unidentified non-synonymous, nonsense or frameshift mutations in *MLL2* were found in 26 of these 43 cases (Figure 15 and Table 16). In total, through either exome sequencing or targeted sequencing of *MLL2*, 33 distinct *MLL2* mutations were identified in 35 of 53 families (66%) with Kabuki syndrome (Figure 15 and Table 16). In each of 12 cases for which DNA from both parents was available, the *MLL2* variant was found to have occurred *de novo*. Three mutations were found in two individuals each. One of these three mutations was confirmed to have arisen *de novo* in one of the cases, indicating that some mutations in

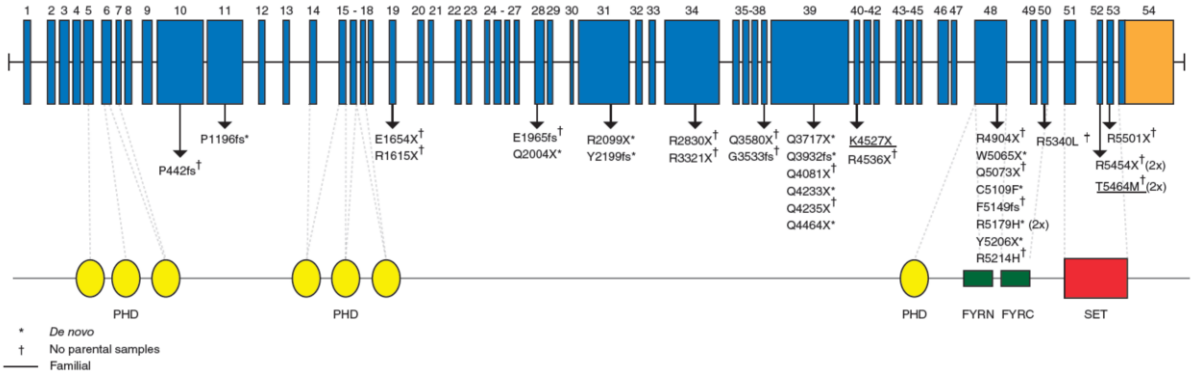

Figure 15 : Genomic structure and allelic spectrum of MLL2 mutations that cause Kabuki syndrome.

MLL2 is composed of 54 exons that encode untranslated regions (orange) and protein coding sequence (blue) including 7 PHD fingers (yellow), FYRN (green), FYRC (green) and a SET domain (red). Arrows indicate the locations of 32 different mutations found in 53 families with Kabuki syndrome including 20 nonsense mutations, 7 indels and 5 amino acid substitutions. Asterisks indicate mutations that were confirmed to be *de novo* and crosses indicate cases for which parental DNA was unavailable. The two underlined mutations were transmitted each within a family, from an affected parent to an affected child.

Table 16 : Annotation of all MLL2 mutations found in 53 Kabuki cases screened.

Kindred	Indiv	Exome Sequenced	Mutation	Exon	Predicted Amino Acid Change	Confirmed as *denovo*	Position
1		yes	c.G15195A	48	p.W5065X	+	chr12:47706821
2		yes	c.C6010T	28	p.Q2004X	+	chr12:47722238
3		yes	c.C12697T	39	p.Q4233X	+	chr12:47712058
4		yes	c.C8488T	34	p.R2830X	-	chr12:47718918
5		yes	--	--	--	--	--
6		yes	c.11794_11797delCAAC	39	p.Q3932SfsX46	+	chr12:47712958-61
7		yes	c.T15618G	48	p.Y5206X	+	chr12:47706398
8		yes	c.3585_3586insA	11	p.P1196TfsX11	+	chr12:47730053-54
9		yes	c.C6295T	31	p.R2099X	+	chr12:47721525
10		yes	c.6595delT	31	p.Y2199IfsX65	+	chr12:47721225
11		no	c.G15326T	48	p.C5109F	+	chr12:47706690
12		no	c.G15536A	48	p.R5179H	+	chr12:47706480
13		no	c.C11149T	39	p.Q3717X	+	chr12:47713606
14		no	c.15444_15445delTT	48	p.F5149CfsX9	-	chr12:47706571-72
15		no	c.C15217T	48	p.Q5073X	-	chr12:47706799
16		no	c.C9961T	34	p.R3321X	-	chr12:47717445
17		no	c.C14710T	48	p.R4904X	-	chr12:47707306
18		no	c.5875_5891dup17	28	p.E1965GfsX88	-	chr12:47722341-57
19		no	c.G15536A	48	p.R5179H	-	chr12:47706480
20		no	c.C12703T	39	p.Q4235X	-	chr12:47712152
21		no	c.C12241T	39	p.Q4081X	-	chr12:47712514
22		no	c.C13390T	39	p.Q4464X	+	chr12:47711365
23		no	c.G15641A	48	p.R5214H	-	chr12:47706375
24		no	--	--	--	--	--
25	1	no	c.A13580T	40	p.K4527X	*	chr12:47711035
	2	no	c.A13580T	40	p.K4527X	-	chr12:47711035
26		no	c.C16501T	53	p.R5501X	-	chr12:47702113
27	1	no	--	--	--	--	--
	2	no	--	--	--	--	--
28		no	--	--	--	--	--
29		no	c.C10738T	38	p.Q3580X	-	chr12:47714119
30		no	--	--	--	--	--
31		no	c.C16360T	52	p.R5454X	-	chr12:47702382
32		no	--	--	--	--	--
33		no	--	--	--	--	--
34		no	--	--	--	--	--
35		no	c.4956_4957insG	19	p.E1654X	-	chr12:47724800-01
36		no	c.10599_10630del32	38	p.V3534QfsX11	-	chr12:47714227-58
37		no	--	--	--	--	--
38		no	c.C13606T	40	p.R4536X	-	chr12:47711008
39		no	c.G16019T	50	p.R5340L	-	chr12:47704661
40		no	c.C4843T	19	p.R1615X	-	chr12:47724914
41		no	c.C16391T	52	p.T5464M	-	chr12:47702351
42		no	--	--	--	--	--
43		no	--	--	--	--	--
44		no	--	--	--	--	--
45		no	c.C16360T	52	p.R5454X	-	chr12:47702382
46		no	--	--	--	--	--
47		no	--	--	--	--	--
48		no	--	--	--	--	--
49		no	--	--	--	--	--
50		no	c.1324delC	10	p.P442HfsX487	-	chr12:47732409
51		no	--	--	--	--	--
52	1	no	c.C16391T	52	p.T5464M	*	chr12:47702351
	2	no	c.C16391T	52	p.T5464M	-	chr12:47702351
53		no	--	--	--	--	--

-- no mutation identified; + confirmed *de novo*; - no parental samples available; X stop codon; fs frameshift; * confirmed as inherited. Kindreds 25, 27 and 52 show dominant transmission of Kabuki syndrome from parent to child. Both affected individuals are listed here. Chromosomal position was determined using March 2006 assembly from UCSC (hg18).

individuals with Kabuki syndrome are recurrent. In addition, *MLL2* mutations (resulting in p.4527K>X and p.5464T>M) were also identified in each of two families in which Kabuki syndrome was transmitted from parent to child. None of the additional *MLL2* mutations was found in 190 control chromosomes from individuals of matched geographical ancestry.

4.4 Discussion

Our results strongly suggest that mutations in *MLL2* are a major cause of Kabuki syndrome. *MLL2* encodes a large 5,262-residue protein that is part of the SET family of proteins, of which Trithorax, the *Drosophila* homolog of MLL, is the best characterized[112]. The SET domain of MLL2 confers strong histone 3 lysine 4 methyltransferase activity and is important in the epigenetic control of active chromatin states[113]. In mice, loss of MLL2 results in embryonic lethality before embryonic day 10.5, whereas $MLL2^{+/-}$ mice are viable but are smaller than wild type (Kai Ge, personal communication).

Most of the *MLL2* variants identified in individuals with Kabuki syndrome are predicted to truncate the polypeptide chain before translation of the SET domain. Though it is not certain whether Kabuki syndrome results from haploinsufficiency or from a gain of function at *MLL2*, haploinsufficiency seems to be the more likely mechanism. Deletion of chromosome 12q12–q13.2, which encompasses *MLL2*, has been reported in a child with characteristics of Noonan syndrome[120]. However, we re-analysed this case using oligo aCGH (including 21 probes that cover *MLL2*) and found the distal breakpoint to be located ~700 kb proximal to *MLL2* (data not shown). Also, all of the pathogenic missense variants identified here are located in regions of *MLL2* that encode C-terminal domains. This suggests that missense variants elsewhere in *MLL2* may be better tolerated or, alternatively, may be embryonically lethal.

For the 18 of 53 cases for which no previously unidentified protein-altering variant was found, it is possible that noncoding or other missed mutations in *MLL2* are responsible for this disorder. Alternatively, Kabuki syndrome could be genetically heterogeneous, and further analysis of these cases by exome sequencing may elucidate additional genes for Kabuki syndrome and potentially explain some of the phenotypic heterogeneity seen in this disorder. Notably, 9 of 10 individuals in the discovery cohort (90%), but only 26 of 43 individuals in the replication cohort (60%), were ultimately found to have mutations in *MLL2*. It is therefore possible that the careful selection of canonical Kabuki cases for the discovery cohort enriched for a shared genetic basis. This underscores the importance of access to deeply phenotyped and well-characterized cases.

4.5 Conclusion

In summary, we applied exome sequencing of a small number of unrelated individuals with Kabuki syndrome to discover that mutations in *MLL2* underlie this disorder. As predicted in previous analyses[86,111], allowing for even a small degree of genetic heterogeneity or missing data substantially confounds exome analysis by increasing the number of candidate genes consistent with the model of inheritance. To facilitate the prioritization of genes under such criteria, we stratified data by ranked phenotypes and found that *MLL2* was prominent in the higher ranked cases. However, nine of the ten individuals with Kabuki syndrome in the discovery cohort were ultimately found to have *MLL2* mutations, such that stratification by phenotype was of less importance than originally appeared to have been the case. Nonetheless, the sequential analysis of ranked cases may have reduced the probability of confounding due to genetic heterogeneity. All of the *MLL2* mutations found in the discovery set via exome sequencing were loss-of-function variants. As a result, *MLL2* ranked highly among candidate genes assessed by predicted functional impact. Such a pattern will likely occur for some, but not all, Mendelian phenotypes subjected to this approach. We anticipate that the further development of strategies to stratify data at both the

genotypic and phenotypic level will be critical for exome and whole-genome sequencing to reach their full potential as tools for discovery of genes underlying Mendelian and complex diseases.

4.6 Materials and Methods

Cases and samples. For exome sequencing, we selected ten individuals of self-reported European, Hispanic or mixed European and Haitian ancestry with Kabuki syndrome from ten unrelated families. Phenotypic data were collected from review of medical records, phone interviews and photographs. All participants provided written consent, and the Institutional Review Boards of Seattle Children's Hospital and the University of Washington approved all studies. The clinical characteristics of the 43 individuals in the validation cohort who had been diagnosed with Kabuki syndrome have been reported previously[116]. Subjective assessment and ranking of the Kabuki phenotype was based on pictures of each subject (Figure 14) and clinical information (Table 12). Informed consent was obtained for publication of each of the facial photos shown.

Exome definition, array design and target masking. We targeted all protein-coding regions as defined by RefSeq 36.3. Entries were filtered for the following: (i) CDS as the feature type, (ii) transcript name starting with "NM_" or "-", (iii) reference as the group_label, (iv) not being on an unplaced contig (for example, 17|NT_113931.1). Overlapping coordinates were collapsed for a total of 31,922,798 bases over 186,040 discontiguous regions. A single custom array (Agilent, 1M features, aCGH format) was designed to have probes over these coordinates as previously described[86], except here, the maximum melting temperature (*T*m) was raised to 73 °C. The mappable exome was also determined as previously described[86] using this RefSeq exome definition instead. After masking for 'unmappable' regions, 30,923,460 bases were left as the mappable target.

Targeted capture and massive parallel sequencing. Genomic DNA was extracted from peripheral blood lymphocytes using standard protocols. Five micrograms of DNA from each of ten individuals with Kabuki syndrome was used for construction of a shotgun sequencing library as described previously3 using paired-end adaptors for sequencing on an Illumina Genome Analyzer II (GAII). Each shotgun library was hybridized to an array for target enrichment; this was then followed by washing, elution and additional amplification. Enriched libraries were then sequenced on a GAII to get either single-end or paired-end reads.

Read mapping and variant analysis. Reads were mapped and processed largely as previously described[86]. In brief, reads were quality recalibrated using Eland and then aligned to the reference human genome (UCSC assembly hg18, NCBI Build 36.1) using Maq. When reads with the same start site and orientation were filtered, paired-end reads were treated like separate single-end reads; this method is overly conservative and hence the actual coverage of the exomes is higher than reported here. Sequence calls were performed using Maq and these calls were filtered to coordinates with ≥8× coverage and consensus quality ≥ 20.

Indels affecting coding sequences were identified as previously described[86], but we used phaster[121] instead of cross_match and Maq. Specifically, unmapped reads from Maq were aligned to the reference sequence using phaster[121] (version 1.100122a) with the parameters *-max_ins*:21 *-max_del*:21 *-gapextend_ins*:-1 *-gapextend_del*:-1 *-match_report_type*:1. Reads were then filtered for those with at most two substitutions and one indel. Reads that mapped to the negative strand were reverse complemented and, together with the other filtered reads, were remapped using the same parameters to reduce ambiguity in the called indel positions. These reads were then filtered for (i) having a single indel more than 3 bp from the ends and (ii) having no other substitutions in the read. Putative indels were then called per individual if they were supported by at least two filtered reads that started from different positions. An 'indel reference' was generated as

previously described[86], and all the reads from each individual were mapped back to this reference using phaster with default settings and -match_report_type:1. Indel genotypes were called as previously described[86].

To determine the novelty of the variants, sequence calls were compared against 16 individuals for whom we had previously reported exome data[86,111] and 10 EGP exomes. Annotations of variants were based on NCBI and UCSC databases using an in-house server (SeattleSeqAnnotation[109]). Loss-of-function variants were defined as nonsense mutations (premature stop) or frameshifting indels. For each variant, we also generated constraint scores as implemented in GERP[119].

Mutation validation. Sanger sequencing of PCR amplicons from genomic DNA was used to confirm the presence and identity of variants in the candidate gene identified via exome sequencing and to screen the candidate gene in additional individuals with Kabuki syndrome.

Array comparative genomic hybridization (CGH). Samples were hybridized to commercially available whole-genome tiling arrays consisting of one million oligonucleotide probes with an average spacing of 2.6 kb throughout the genome (SurePrint G3 Human CGH Microarray 1x1M, Agilent Technologies). Twenty-one probes on this array covered *MLL2* specifically. Data were analysed using Genomics Workbench software according to the manufacturer's instructions.

dbGaP accession

http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000295.v1.p1

4.7 Acknowledgements

We thank the families for their participation and the Kabuki Syndrome Network for their support. We thank J. Allanson, J. Carey and M. Golabi for referral of cases and M. Emond for helpful discussion. We thank the 1000 Genomes Project for early data release that proved useful for filtering out common variants. Our work was supported in part by grants from the US National

Institutes of Health (NIH)–National Heart, Lung, and Blood Institute (5R01HL094976 to D.A.N.and J.S.), the NIH–National Human Genome Research Institute (5R21HG004749 to J.S., 1RC2HG005608 to M.J.B., D.A.N. and J.S.; and 5RO1HG004316 to H.K.T.), NIH–National Institute of Environmental Health Sciences (HHSN273200800010C to D.N. and M.J.R.), Ministry of Health, Labour and Welfare (K.Y., N.M., T.O. and N.N.), Japan Science and Technology Agency (N.M.), Society for the Promotion of Science (N.M.), the Life Sciences Discovery Fund (2065508 and 0905001), the Washington Research Foundation and the NIH–National Institute of Child Health and Human Development (1R01HD048895 to M.J.B.). S.B.N. is supported by the Agency for Science, Technology and Research, Singapore. A.W.B. is supported by a training fellowship from the NIH–National Human Genome Research Institute (T32HG00035).

CHAPTER 5 : MASSIVELY PARALLEL SEQUENCING AND RARE DISEASE

Parts of this chapter previously published as

Sarah B Ng, Deborah A Nickerson, Michael J Bamshad & Jay Shendure (2010) **Massively parallel sequencing and rare disease.** *Human Molecular Genetics*. 19(R2):R119-24.

Since the publication of the work in Chapter 2 in 2009[86], the use of massively parallel sequencing has gained incredible momentum, and has been applied successfully to >400 studies of Mendelian and other disease. Here, I summarize the previous chapters in the context of other sequencing studies, reviewing some of the strategies we and others have used to sift through variants to determine the causal mutations for several monogenic disorders[86,111,122-130 131-135], as well as to identify the molecular basis[136] or provide evidence for the clinical diagnosis[87] for other diseases.

5.1 Early Successes

5.1.1 Molecular diagnosis of mutations in known genes

The most straightforward method to identify causal mutations in an individual is by comparison with known mutations and disease-associated genes, like those curated in databases such as Online Mendelian Inheritance in Man (OMIM)[2] and the Human Gene Mutation Database (HGMD)[88]. This is of most utility in well-studied diseases and can be used to assist in molecular diagnoses, i.e. finding novel and known mutations in previously identified disease genes. For example, in whole-genome sequencing of an individual with Chacot–Marie–Tooth disease (CMT)[136], analysis of coding variants identified a nonsense mutation in SH3TC2 that was already

implicated in CMT, and consistent with the recessive mode of inheritance, a novel missense mutation in the same gene was also found.

In a similar vein, investigators have begun to consider the use of massively parallel sequencing as diagnostic assays for genetic disorders. Resequencing of candidate genes for neurofibromatosis type 1[137], ataxia[138], mitochondrial disorders[139] and ocular birth defects[140], as well as for HLA typing[141], has shown that targeted capture coupled with high-throughput sequencing is increasingly feasible for this task, although further improvements in accuracy and cost may be necessary before its widespread adoption in clinical laboratories.

5.1.2 Clinical diagnosis based on sequence data

Observing known mutations or novel mutations in disease genes may also assist in making a clinical diagnosis for patients. For example, annotation of homozygous variants identified in exome sequencing of a patient from a consanguineous union was used to make a genetic diagnosis of congenital chloride-losing diarrhoea (CLD)[87]. Initially, the patient was suspected of having Bartter syndrome, but no known disease loci were within homozygous segments. Instead, a homozygous single-base substitution was identified in the gene SLC26A3, in which mutations had previously been found to cause CLD. Clinical follow-up based on this molecular observation confirmed CLD as the primary diagnosis and led to a suitable treatment.

In another study, exome sequencing of four individuals affected with Miller syndrome[111] revealed that a subset of cases (siblings) alone had novel variants that were predicted to be damaging in a single gene, DNAH5, which had previously been implicated in primary ciliary dyskinesia. This led to the realisation that a cystic fibrosis-like phenotype unique to these siblings was, in fact, not related to Miller syndrome, but instead a superimposition of another disease phenotype caused by mutations in a separate gene.

These studies were the first diagnoses of monogenic disorders based on massively parallel sequencing and suggest its potential to assist in the evaluation and diagnosis of patients, particularly when the diagnosis is uncertain.

5.1.3 Novel disease gene discovery

Disorders without known mutations or disease-associated genes require different filtering strategies to isolate pathogenic mutations. In general, the following assumptions are made about causal mutations underlying 'simple', monogenic, Mendelian disease: (i) a single mutation is sufficient to cause the disease, which would (ii) be rare, and probably private to affected individuals, and (iii) because they are of large effect, they are most likely coding and (iv) highly penetrant. As such, investigators look first for variants that change protein sequence, i.e. missense and nonsense substitutions, coding indels as well as splice acceptor and donor site changes, and second for variants that are very rare or novel. Where necessary, a further assumption is often made that the disease is genetically homogenous, i.e. unrelated affected individuals have mutations in the same gene, at least for the individuals chosen for the study.

The first report of the potential of using massively parallel sequencing to identify mutations by this strategy was for Freeman–Sheldon syndrome (FSS)[86]. In this study, the exomes of four unrelated affected individuals were sequenced, and for each individual, genes that had at least one private protein-altering variant were shortlisted, consistent with a dominant disease model. After intersecting the genes from all four individuals, it was found that only one gene was common among all – *MYH3*, which had previously been shown to be causal for FSS. Although this was not a novel identification of the disease-associated gene, it was nonetheless a proof-of-concept experiment that showed how massively parallel sequencing could be applied on a genome-wide scale to find causal mutations for monogenic diseases even without linkage or pedigree information, nor any information related to disease mechanism.

The same strategy was employed successfully in studies for autosomal dominant disorders — for example for Schinzel–Giedion syndrome[124] and for Kabuki syndrome[126]. Notably, consistent with the sporadic nature of these syndromes, the majority of the causal mutations identified in these studies were found to be *de novo*, which highlights the advantage of using exome sequencing over linkage studies for these cases.

To extend the strategy to recessive disease, shortlisted genes were required to have at least two private protein-altering variants instead of just one. This accounts for two situations: under a simple recessive model, the disease mutation should be homozygous, and under a compound heterozygous model, two different mutations on different haplotypes are expected instead. This approach was applied to a presumed recessive disease, Miller syndrome[111]. Four affected individuals, of whom two were siblings, were exome sequenced, and a manual review of the intersecting two genes found compound heterozygous mutations in *DHODH* to be causal. A similar analysis in Fowler syndrome[131] also was successful in identifying compound heterozygous causal mutations in FLVCR2, using only two affected individuals.

It is not always necessary to do a genome-wide analysis, especially in the cases where linkage intervals have already been determined or other familial information is available. Sequencing only one or two individuals will often suffice, particularly because the potential causal variant lists are not long.

Using linkage information in whole-genome sequencing of a patient with metachondromatosis[125], exome sequencing of a proband and her mother in Joubert syndrome 2 (ref. 128) and exome sequencing of a proband with non-syndromic hearing loss DFNB82 (ref. 132) helped to narrow down and identify PTPN11, TMEM216 and GPSM2, respectively, as disease-associated genes.

Familial information was also used in whole-genome sequencing of Miller syndrome[130]. Two affected siblings and their parents were sequenced, which allowed the resolution of haplotype inheritance for each sibling. The investigators then focused their disease analysis on the 22% of the genome for which both maternal and paternal haplotypes were inherited identically in both siblings and could shortlist potential causal variants to within these intervals, even though no linkage analysis was available.

In other studies that utilised linkage information, only the linkage intervals were captured and sequenced, which is presently more cost-effective. In familial exudative vitreoretinopathy[123], clericuzio-type poikiloderma with neutropenia[127] and non-syndromic hearing loss DFNB79[129], a single proband was sequenced to find the causal mutation within linkage intervals; in sensory/motor neuropathy with ataxia[122], a proband with one or two parents were sequenced; and in X-linked TARP syndrome[133], only exons on the X chromosome were captured and sequenced from two carriers. As sequencing costs drop, however, it is possible that the cost of targeted capture becomes disproportionately high when compared with sequencing costs, which may favour more whole genome sequencing for single samples instead.

5.1.4 Filtering based on function

The primary filter most investigators use to identify potentially causal mutations is based on variant function—namely, if the variant affects coding regions (missense, nonsense, coding indels and splice acceptor and donor sites) or other non-coding RNA transcripts. The main rationale given for this is that these variants tend to be of larger effect than non-coding variants, and also because it is difficult to predict the effects of noncoding and synonymous variants with any certainty. As such, in order to reduce noise when analysing possible disease-causing variants, non-coding and synonymous variants are often ignored or greatly down-weighted.

For some disorders, it is possible to filter variants even further, by focusing only on those that are loss-of-function (i.e. nonsense and frameshift mutations). Since there are only a limited number of such mutations in any genome (< 50), the candidate list is shortened very quickly. This filter was particularly useful to identify mutations in RBM10 as causal for TARP syndrome[133], and also in *MLL2* for Kabuki syndrome[126].

For all the studies reviewed here, at least, restricting the analyses to coding variants has been justified—all have identified variants that affect protein function, mainly missense and nonsense mutants, and also frameshifting indels, and in one case a mutation at a splice acceptor site. This will most certainly not always be the case, as intronic, regulatory and synonymous variants are certainly known to affect disease (reviewed in ref. 142), and as more disorders are studied, there will be a growing need for functional annotation of non-coding regions and tools to analyse the same.

5.1.5 Ranking variants by effect and conservation

Variants can also be ranked by potential effect on protein structure and function, and also by conservation scores, as estimated by tools like SIFT[58], PolyPhen[56,57], CDPred[133], PhyloP[143] and GERP[118,119], with the rationale that mutations which are disruptive to proteins and/or at more conserved sites are more likely to be pathogenic. However, these tools have limited specificity and sensitivity[144], and mutations ultimately determined to be causal will rank highly, but potentially not first. As such, these rankings are normally used in conjunction with other strategies and not as a stand-alone filter.

5.1.6 Filtering for rare variants

Rare diseases, by definition, have an individual incidence of less than 1/1150–1/2000 in the population[108,145], and it is expected that mutations underlying these rare diseases will be at correspondingly rare frequencies, and most likely private to affected individuals. This is especially

so for mutations that are highly penetrant—these variants of major effect and distinctive phenotype are not expected to be found in the population at large, and hence will not be seen in genome-wide scans for variants [e.g. the 1000 Genomes Project[146]], nor in polymorphism repositories [e.g. dbSNP[147]]. Exclusion from these data sets is typically an important criterion in defining a rare, novel or private variant.

From empirical analysis of published exomes[86,111,126], we estimate that there are >20 000 single-nucleotide variants in a given exome (as defined by the Consensus Coding Sequence database, CCDS), with about half affecting protein sequence. Using dbSNP and 1000 genomes data together as a filter suggests the number of novel SNPs is at most 1/10th of that (Figure 16), which is a reflection of the breadth of ascertainment available in these data sets. However, a caveat to note is that phenotypic information is not always available for the samples used in these data sets, and it is possible that pathogenic mutations are present in them. In the case of recessive mutations, particularly, there is a chance that a normal carrier could have been genotyped and the recessive disease causing mutation deposited in the database.

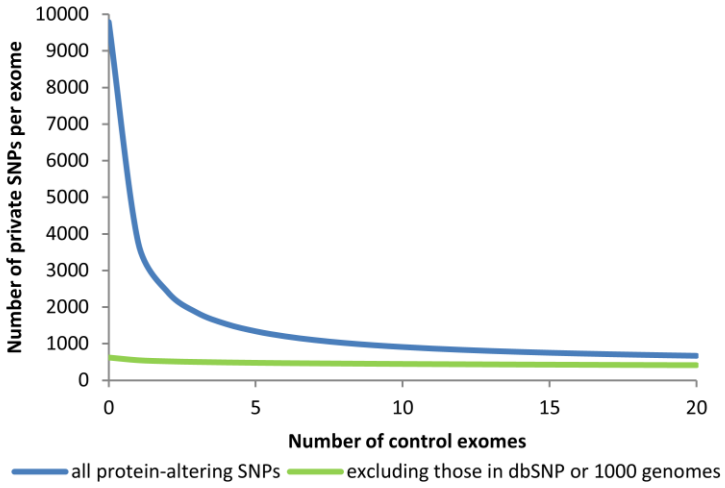

Figure 16 : Effect of increasing number of control exomes on private variants

Effect of increasing number of control exomes on private variants observed in a single exome, with and without the use of dbSNP and 1000 genomes data. The number of private mutations observed in an individual from sequential addition of control exomes was averaged from 10 000 permutations of 21 published exomes of non-African ancestry[86,109,124].

As a complementary approach then, control individuals with known phenotype and family history are often sequenced along with the affected cases, also to be used as a filter for common variants. This has the advantage of allowing for population-matched controls, especially for

populations that may be under-represented in the current databases, and also to control for technical artefacts that may arise during the sequencing or alignment of sequence reads.

The effect of adding more control exomes is illustrated in Figure 16, which shows that the average number of novel or private SNPs in a given individual drops exponentially as more control exomes are used and starts to plateau by the time 15–20 controls are added (Figure 16, blue line). This suggests that a limited number of control exomes is sufficient to filter for private variants by this method even without external datasets—adding beyond 20 controls to these data clearly has a diminishing rate of return.

5.1.7 Searching for de novo mutations

An emerging paradigm for finding disease-causing mutations is to take searching for rare variation to its logical extreme, and look for variants are present in probands but not in their unaffected parents, i.e. *de novo* mutations. Single-nucleotide *de novo* mutations are estimated to happen at a rate of 74 germline mutations per generation[148], of which about one on average is expected to be in a coding region[130]. At the same time, *de novo* mutations as a class have the potential of greatest deleterious effect, being subject to the least stringent selection[149]. In retrospect, the causal mutations in Schinzel–Giedion syndrome[124] and Kabuki syndrome[126] could have been predicted to be *de novo* based on the sporadic incidence of these syndromes. Prospectively, by focusing on these rare events, it is possible to quickly identify possible mutations of large effect in single families, which greatly improves power in disorders with genetic heterogeneity. For example, in a study of intellectual disability[135], ten patient-unaffected parent trios were exome sequenced, and nine Sanger-verified *de novo* mutations were found in nine different genes, of which two had been previously associated with the phenotype; in a separate study of autism spectrum disorder[134], 20 trios were sequenced, to identify *de novo* mutations in 11 genes, of which 4 had been previously associated with the phenotype. Under the

assumption of genetic homogeneity, none of these genes would have been identified with current sample sizes.

5.2 Current Limitations

5.2.1 Limited sequencing scope

Since most current analyses are restricted to protein-coding regions, a prior decision is made in some studies to restrict the sequencing scope to genic regions, whether within a linkage region or as a whole exome, primarily based on cost. It is notable that there are many potential definitions for the exome. Earlier studies used the genes from the CCDS[150], a set of well-annotated, highly conserved proteins, with the rationale being that although this is a relatively conservative set of genes, it is also less likely to contain pseudogenes and potentially unverified protein transcripts, which would result in spurious variant calls and add noise to the analysis of potential mutations.

However, the CCDS is by no means complete, and failure to find a disease gene could be the result of missing sequence or annotation. In one example, mutations in the gene associated with Kabuki syndrome[126] were identified only because the exome definition was expanded to the RefSeq database[151]. With the CCDS definition, this gene was not captured nor sequenced, and the mutations missed entirely[152]. Hence, to reduce the likelihood of this happening, more recent studies have started to use larger, more inclusive gene sets instead, like the aforementioned RefSeq database[151], the Ensembl database[153], genes from the UCSC browser[154] and the GENCODE set[155], particularly as capture technology limitations are overcome. It is worth noting that even when whole-genome sequencing becomes affordable enough such that targeted methods are not necessary, these definitions will still be important in the annotation of coding variants.

5.2.2 Spurious gene identifications

Using the aforementioned filters may not always be completely specific to the disease-associated gene. In a number of studies, spurious genes were also identified, but later dismissed upon manual review for various reasons.

In the Schinzel–Giedion syndrome study, all affected individuals had the exact same variant in 10 genes, suggesting that these could be polymorphisms that were missed in the controls[124], or possibly that there was a systematic artefact in variant calling. In another gene in the same study, CTBP2, the investigators note that it had a high variation rate, even in controls, suggesting that this could be due to the presence of other paralogous loci. A similar observation was made in the Miller syndrome study[111], where CDC27 was identified as a potential candidate, but noted to have an unannotated processed pseudogene that contributed to misalignments and an inflated variant call rate.

Another factor that could affect the number of variant calls is the length of the gene since many of these only have limited variant calls in the existing databases. For example, MUC16 was shortlisted in the Kabuki syndrome study[86] as the only gene to be common to all 10 cases, but was determined to be a false-positive due to its long length (>14 500 amino acids), which could have contributed to a high number of variant calls across all individuals.

5.2.3 Missing variant calls

It is also possible that not all affected individuals will have sequenced mutations in the same gene. This could be due to the disease model (e.g. genetic heterogeneity or non-coding variants) or technical issues (e.g. missing variant calls due to low coverage). As an example of the latter, of the 10 affected individuals in the Kabuki syndrome study[86], only 7 had causal mutations identified by exome sequencing. Misalignments and low coverage led to missed frameshift indel calls in two of the remaining three (later identified by Sanger sequencing), and the causal

mutation in the last is still unknown. To deal with these "missing" mutations, it is possible to sequence more affected individuals or to change the filtering approach to require that only a subset of individuals have a shared gene, rather than all, but the trade-off is that the candidate gene lists are greatly inflated 10-20 times[126].

5.3 Future directions

Exome sequencing has clearly been of great utility in identifying causal mutations for disease, but the actual success rate is difficult to measure, due to publication bias. One estimate is about 60%, based on studies at one centre[152]. The question remains: where or what are the mutations in the remaining disorders?

As previously mentioned, some of these mutations could be missed due to technical issues – examples include unannotated genes leading to missing capture products, systematic low capture or sequencing yields in high GC content regions, and insufficiently sensitive indel calling. Improvement in these areas are on-going, with exome capture reagents now covering >60 Mb of coding and non-coding transcripts[156,157], changes in library preparation and sequencing technology to reduce GC bias, and the use of local realignment and assembly programs to increase sensitivity in calling indels[158,159]. In general, as sequencing technology improves and costs decrease, it is anticipated that the generation of sequence data will no longer be a bottleneck but rather, the interpretation of this genome-scale data.

One major class of variants that will require more tools for functional analysis are the non-coding variants. Particularly because they are non-coding, they are, in general, of smaller effect size, and absolute phenotypic changes due to such mutations are not often observed, and as a result, these variants are often side-lined. It is noted that despite the numerous whole genomes that have been sequenced[76,78,136,160], the majority of studies that analysed potential associations

with disease focused primarily on coding variants, particularly with respect to Mendelian disorders. Useful resources that can give a clue to the function of such variants include genome-wide datasets of chromatin occupation and transcription factor binding (e.g. ENCODE[161]), but assays of functional effect are often too unwieldy to perform on a large scale. To this end, recent advances have suggested methods for high-throughput dissection of regulatory elements like promoters and enhancers[162,163], but such assays cannot currently be performed on a genome-wide scale.

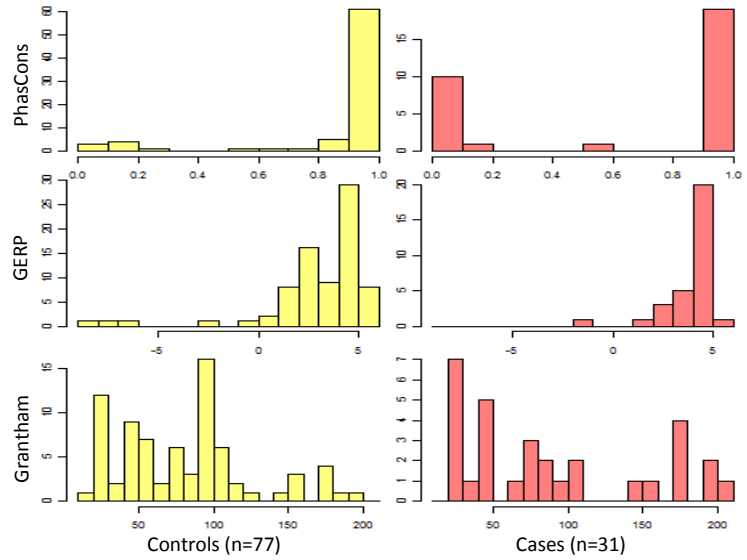

Figure 17 : Comparison of novel missense annotations between cases and controls.

Histograms of three different variant annotation scores based on novel missense variants found in individuals with (cases) and without (controls) Kabuki syndrome are compared and found not to be dissimilar. Missense mutations from cases were collated from resequencing studies of *MLL2*[162-165]. Missense variants in *MLL2* from unaffected individuals were obtained from the Exome Sequencing Project database[166], filtering for variants that were seen only in one out of > 5,000 chromosomes. All variants were annotated using SeattleSeq annotation, to determine PhastCons[167], GERP[116] and Grantham scores[55].

Even in the context of coding variants, it sometimes unclear when a given variant is pathogenic or benign. For example, in whole genome sequencing studies, missense variants that were annotated to be causal for serious Mendelian disorders were found in the asymptomatic patients[136,160], suggesting either that the variants are not completely penetrant, or more likely that the previous annotation was inaccurate. A further observation of novel missense mutations in *MLL2* between Kabuki patients[164-167] and unaffected controls[168] showed that bioinformatics classification methods[55,118,169] are unable to distinguish between the two (Figure 17), suggesting that more sensitive methods are necessary –

perhaps in the form of high-throughput functional protein assays – to determine actual functional significance of such variants.

5.4 Conclusions

Massively parallel – "next-generation" – sequencing has been applied successfully to find causal mutations for a number of Mendelian, monogenic disorders, including several that have been intractable to linkage analysis. In this thesis I described the use of exome sequencing to implicate mutations in *DHODH* for Miller syndrome, and in *MLL2* for Kabuki syndrome. In studies from other groups, massively parallel sequencing has also been used successfully to re-sequence genes within linkage intervals, and also to find variants on a whole-genome or exome scale.

Using simple filters based on variant function and frequency, as well as careful choices of cases and controls, has been highly useful in isolating pathogenic mutations from background polymorphisms. In particular, priority is given to variants that are private to affected individuals, under the assumption that the mutations underlying these monogenic disorders are highly penetrant and rare. A second assumption is that the underlying mutations have large effect and are likely to be coding, and third that the disorder is genetically homogeneous in the samples being studied.

As less simple syndromes and disorders are studied, these assumptions are less likely to hold. In cases where genetic heterogeneity is present, mutations are less penetrant and noncoding variation is causal, the current strategies outlined here will be too stringent. Allowing for these more complex models will inflate candidate lists, and more sophisticated approaches will need to be developed to conduct candidate prioritization. One approach that is successful even under

genetic heterogeneity is to search for mutations under a *de novo* paradigm, since only data from a trio – a single proband and unaffected parents – are necessary to make calls.

The last few years have clearly shown how massively parallel sequencing can accelerate the pace of disease gene discovery and revolutionize the study of the genetic bases on Mendelian disorders. Despite these numerous successes, on a whole, there is still much that is unknown about the variants found in an individual genome. As we move into an era where whole genome sequence data is ubiquitous, the interpretation of such data will still be the ultimate challenge, and advances in this area will be critical before we are able to comprehensively understand how individual genetic difference contribute to the incredible diversity of human phenotypes.

REFERENCES

1 Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. *Nat Rev Genet* **7**, 277-282, doi:10.1038/nrg1826 (2006).

2 McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80**, 588-604, doi:10.1086/514346 (2007).

3 *Online Mendelian Inheritance in Man*, <http://www.omim.org>.

4 Katsanis, N., Ansley, S. J., Badano, J. L., Eichers, E. R., Lewis, R. A., Hoskins, B. E., Scambler, P. J., Davidson, W. S., Beales, P. L. & Lupski, J. R. Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. *Science* **293**, 2256-2259, doi:10.1126/science.1063525 (2001).

5 *The Haemophilia A Mutation Database*, <http://hadb.org.uk>.

6 Zielenski, J. & Tsui, L. C. Cystic fibrosis: genotypic and phenotypic variations. *Annu Rev Genet* **29**, 777-807, doi:10.1146/annurev.ge.29.120195.004021 (1995).

7 *Cystic Fibrosis Mutation Database*, <http://www.genet.sickkids.on.ca>.

8 Hartong, D. T., Berson, E. L. & Dryja, T. P. Retinitis pigmentosa. *Lancet* **368**, 1795-1809, doi:10.1016/S0140-6736(06)69740-7 (2006).

9 Allikmets, R. Simple and complex ABCR: genetic predisposition to retinal disease. *Am J Hum Genet* **67**, 793-799, doi:10.1086/303100 (2000).

10 Rowntree, R. K. & Harris, A. The phenotypic consequences of CFTR mutations. *Ann Hum Genet* **67**, 471-485 (2003).

11 Morgan, T. H. Sex Limited Inheritance in Drosophila. *Science* **32**, 120-122, doi:10.1126/science.32.812.120 (1910).

12 Sturtevant, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43-59 (1913).

13 Fisher, R. A. The amount of information supplied by records of families as a function of the linkage in the population sampled. *Ann Eugen* **6**, 66-70 (1934).

14 Haldane, J. B. Methods for the detection of autosomal linkage in man. *Ann Eugen* **6**, 26-65 (1934).

15 Penrose, L. S. The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann Eugen* **6**, 133-138 (1935).

16 Haldane, J. B. & Smith, C. A. A new estimate of the linkage between the genes for colourblindness and haemophilia in man. *Ann Eugen* **14**, 10-31 (1947).

17 Smith, C. A. Detection of linkage in human genetics. *Journal of the Royal Statistical Society* **B 15**, 153-192 (1953).

18 Morton, N. E. Sequential tests for the detection of linkage. *Am J Hum Genet* **7**, 277-318 (1955).

19 Thompson, E. A. 1953: An unrecognized summit in human genetic linkage analysis. *Statistics Surveys* **1**, 1-15 (2007).

20 Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567-1570 (1987).

21 Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**, 506-516 (1993).

22 Boyer, S. H. & Graham, J. B. Linkage Between the X Chromosome Loci for Glucose-6-Phosphate Dehydrogenase Electrophoretic Variation and Hemophilia A. *Am J Hum Genet* **17**, 320-324 (1965).

23 Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**, 314-331 (1980).

24 Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T. P., Bowden, D. W., Smith, D. R., Lander, E. S. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319-337 (1987).

25 Weber, J. L. & May, P. E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* **44**, 388-396 (1989).

26 International SNP Map Working Group, Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933, doi:10.1038/35057149 (2001).

27 International Human Genome Sequencing Consortium, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

28 Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234-238 (1983).

29 Group, H. s. D. C. R. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971-983 (1993).

30 Ingram, G. I. The history of haemophilia. *J Clin Pathol* **29**, 469-479 (1976).

31 Toydemir, R. M., Rutherford, A., Whitby, F. G., Jorde, L. B., Carey, J. C. & Bamshad, M. J. Mutations in embryonic myosin heavy chain (MYH3) cause Freeman-Sheldon syndrome and Sheldon-Hall syndrome. *Nat Genet* **38**, 561-565, doi:10.1038/ng1775 (2006).

32 Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. & Church, G. M. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732, doi:10.1126/science.1117389 (2005).

33 Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380, doi:10.1038/nature03959 (2005).

34 Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:10.1038/nature07517 (2008).

35 Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K. *et al.* A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**, 1051-1063, doi:10.1101/gr.076463.108 (2008).

36 Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138, doi:10.1126/science.1162986 (2009).

37 Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-1157, doi:10.1038/nature04240 (2005).

38 Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274, doi:10.1126/science.1133427 (2006).

39 Jones, S., Hruban, R. H., Kamiyama, M., Borges, M., Zhang, X., Parsons, D. W., Lin, J. C., Palmisano, E., Brune, K., Jaffee, E. M. *et al.* Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. *Science* **324**, 217, doi:10.1126/science.1171202 (2009).

40 Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387-389, doi:10.1126/science.1167728 (2009).

41 Tewhey, R., Warner, J. B., Nakano, M., Libby, B., Medkova, M., David, P. H., Kotsopoulos, S. K., Samuels, M. L., Hutchison, J. B., Larson, J. W. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol* **27**, 1025-1031, doi:10.1038/nbt.1583 (2009).

42 Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**, 1522-1527, doi:10.1038/ng.2007.42 (2007).

43 Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J. & Zwick, M. E. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* **4**, 907-909, doi:10.1038/nmeth1109 (2007).

44 Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., Richmond, T. A., Middle, C. M., Rodesch, M. J., Packard, C. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4**, 903-905, doi:10.1038/nmeth1111 (2007).

45 Hodges, E., Rooks, M., Xuan, Z., Bhattacharjee, A., Benjamin Gordon, D., Brizuela, L., Richard McCombie, W. & Hannon, G. J. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**, 960-974, doi:10.1038/nprot.2009.68 (2009).

46 Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-189, doi:10.1038/nbt.1523 (2009).

47 Bainbridge, M. N., Wang, M., Burgess, D. L., Kovar, C., Rodesch, M. J., D'Ascenzo, M., Kitzman, J., Wu, Y. Q., Newsham, I., Richmond, T. A. *et al.* Whole exome capture in solution with 3 Gbp of data. *Genome Biol* **11**, R62, doi:10.1186/gb-2010-11-6-r62 (2010).

48 Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. & Nilsson, M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res* **33**, e71, doi:10.1093/nar/gni070 (2005).

49 Porreca, G. J., Zhang, K., Li, J. B., Xie, B., Austin, D., Vassallo, S. L., LeProust, E. M., Peck, B. J., Emig, C. J., Dahl, F. *et al.* Multiplex amplification of large sets of human exons. *Nat Methods* **4**, 931-936, doi:10.1038/nmeth1110 (2007).

50 Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* **6**, 315-316, doi:10.1038/nmeth.f.248 (2009).

51 National Center for Biotechnology Information. *Consensus CDS protein set*, <http://www.ncbi.nlm.nih.gov/projects/CCDS>.

52 National Center for Biotechnology Information. *The Reference Sequence collection (RefSeq)*, <http://www.ncbi.nlm.nih.gov/RefSeq/>.

53 *Human Gene Mutation Database*, <http://www.hgmd.cf.ac.uk>.

54 Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33 Suppl**, 228-237, doi:10.1038/ng1090 (2003).

55 Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-864 (1974).

56 Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**, 3894-3900 (2002).

57 Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

58 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).

59 Cotton, R. G. & Scriver, C. R. Proof of "disease causing" mutation. *Hum Mutat* **12**, 1-3, doi:10.1002/(SICI)1098-1004(1998)12:1<1::AID-HUMU1>3.0.CO;2-M (1998).

60 Ji, W., Foo, J. N., O'Roak, B. J., Zhao, H., Larson, M. G., Simon, D. B., Newton-Cheh, C., State, M. W., Levy, D. & Lifton, R. P. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* **40**, 592-599, doi:10.1038/ng.118 (2008).

61 Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R. & Hobbs, H. H. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869-872, doi:10.1126/science.1099870 (2004).

62 Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264-1272, doi:10.1056/NEJMoa054013 (2006).

63 Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-997, doi:10.1038/nature06611 (2008).

64 Zwick, M. E., Cutler, D. J. & Chakravarti, A. Patterns of genetic variation in Mendelian and complex traits. *Annu Rev Genomics Hum Genet* **1**, 387-407, doi:10.1146/annurev.genom.1.1.387 (2000).

65 Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* **80**, 727-739, doi:10.1086/513473 (2007).

66 Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**, 241-251, doi:10.1038/nrg2554 (2009).

67 Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145, doi:10.1038/nbt1486 (2008).

68 International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).

69 Olson, M. Enrichment of super-sized resequencing targets from the human genome. *Nat Methods* **4**, 891-892, doi:10.1038/nmeth1107-891 (2007).

70 Ng, P. C., Levy, S., Huang, J., Stockwell, T. B., Walenz, B. P., Li, K., Axelrod, N., Busam, D. A., Strausberg, R. L. & Venter, J. C. Genetic variation in an individual human exome. *PLoS Genet* **4**, e1000160, doi:10.1371/journal.pgen.1000160 (2008).

71 Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64, doi:10.1038/nature06862 (2008).

72 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858, doi:10.1101/gr.078212.108 (2008).

73 Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-729, doi:10.1038/ng.128 (2008).

74 Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194 (1998).

75 Kidd, J. M., Cheng, Z., Graves, T., Fulton, B., Wilson, R. K. & Eichler, E. E. Haplotype sorting using human fosmid clone end-sequence pairs. *Genome Res* **18**, 2016-2023, doi:10.1101/gr.081786.108 (2008).

76 Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876, doi:10.1038/nature06884 (2008).

77 Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60-65, doi:10.1038/nature07484 (2008).

78 Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254, doi:10.1371/journal.pbio.0050254 (2007).

79 Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72, doi:10.1038/nature07485 (2008).

80 Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R. *et al.* Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**, e1000083, doi:10.1371/journal.pgen.1000083 (2008).

81 Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., 3rd, Kondrashov, A. S. & Bork, P. Prediction of deleterious human alleles. *Hum Mol Genet* **10**, 591-597 (2001).

82 Yngvadottir, B., Xue, Y., Searle, S., Hunt, S., Delgado, M., Morrison, J., Whittaker, P., Deloukas, P. & Tyler-Smith, C. A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. *Am J Hum Genet* **84**, 224-234, doi:10.1016/j.ajhg.2009.01.008 (2009).

83 Olson, M. V. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* **64**, 18-23, doi:10.1086/302219 (1999).

84 Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K. & Hobbs, H. H. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* **37**, 161-165, doi:10.1038/ng1509 (2005).

85 Siva, N. 1000 Genomes project. *Nat Biotechnol* **26**, 256, doi:10.1038/nbt0308-256b (2008).

86 Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276, doi:10.1038/nature08250 (2009).

87 Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**, 19096-19101, doi:10.1073/pnas.0910672106 (2009).

88 Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S. & Cooper, D. N. The Human Gene Mutation Database: 2008 update. *Genome Med* **1**, 13, doi:10.1186/gm13 (2009).

89 Chen, C. T., Wang, J. C. & Cohen, B. A. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet* **80**, 692-704, doi:10.1086/513149 (2007).

90 Miller, M., Fineman, R. & Smith, D. W. Postaxial acrofacial dysostosis syndrome. *J Pediatr* **95**, 970-975 (1979).

91 Splendore, A., Passos-Bueno, M. R., Jabs, E. W., Van Maldergem, L. & Wulfsberg, E. A. TCOF1 mutations excluded from a role in other first and second branchial arch-related disorders. *Am J Med Genet* **111**, 324-327, doi:10.1002/ajmg.10567 (2002).

92 Fineman, R. M. Recurrence of the postaxial acrofacial dysostosis syndrome in a sibship: implications for genetic counseling. *J Pediatr* **98**, 87-88 (1981).

93 Ogilvy-Stuart, A. L. & Parsons, A. C. Miller syndrome (postaxial acrofacial dysostosis): further evidence for autosomal recessive inheritance and expansion of the phenotype. *J Med Genet* **28**, 695-700 (1991).

94 Donnai, D., Hughes, H. E. & Winter, R. M. Postaxial acrofacial dysostosis (Miller) syndrome. *J Med Genet* **24**, 422-425 (1987).

95 Genée, E. Une forme extensive de dysostose mandibulo-faciale. *J. Genet. Hum.* **17**, 45-52 (1969).

96 Pereira, S. C., Rocha, C. M., Guion-Almeida, M. L. & Richieri-Costa, A. Postaxial acrofacial dysostosis: report on two patients. *Am J Med Genet* **44**, 274-279, doi:10.1002/ajmg.1320440303 (1992).

97 Robinow, M., Johnson, G. F. & Apesos, J. Robin sequence and oligodactyly in mother and son. *Am J Med Genet* **25**, 293-297, doi:10.1002/ajmg.1320250214 (1986).

98 Grabar, P. B., Rozman, B., Logar, D., Praprotnik, S. & Dolzan, V. Dihydroorotate dehydrogenase polymorphism influences the toxicity of leflunomide treatment in patients with rheumatoid arthritis. *Ann Rheum Dis* **68**, 1367-1368, doi:10.1136/ard.2008.099093 (2009).

99 Brosnan, M. E. & Brosnan, J. T. Orotic acid excretion and arginine metabolism. *J Nutr* **137**, 1656S-1661S (2007).

100 Breedveld, F. C. & Dayer, J. M. Leflunomide: mode of action in the treatment of rheumatoid arthritis. *Ann Rheum Dis* **59**, 841-849 (2000).

101 Jarry, B. & Falk, D. Functional diversity within the rudimentary locus of Drosophila melanogaster. *Mol Gen Genet* **135**, 113-122 (1974).

102 Conner, T. W. & Rawls, J. M., Jr. Analysis of the phenotypes exhibited by rudimentary-like mutants of Drosophila melanogaster. *Biochem Genet* **20**, 607-619 (1982).

103 Fukushima, R., Kanamori, S., Hirashiba, M., Hishikawa, A., Muranaka, R. I., Kaneto, M., Nakamura, K. & Kato, I. Teratogenicity study of the dihydroorotate-dehydrogenase inhibitor and protein tyrosine kinase inhibitor Leflunomide in mice. *Reprod Toxicol* **24**, 310-316, doi:10.1016/j.reprotox.2007.05.006 (2007).

104 Imose, M., Nagaki, M., Kimura, K., Takai, S., Imao, M., Naiki, T., Osawa, Y., Asano, T., Hayashi, H. & Moriwaki, H. Leflunomide protects from T-cell-mediated liver injury in mice through inhibition of nuclear factor kappaB. *Hepatology* **40**, 1160-1169, doi:10.1002/hep.20438 (2004).

105 Bushdid, P. B., Brantley, D. M., Yull, F. E., Blaeuer, G. L., Hoffman, L. H., Niswander, L. & Kerr, L. D. Inhibition of NF-kappaB activity results in disruption of the apical ectodermal ridge and aberrant limb morphogenesis. *Nature* **392**, 615-618, doi:10.1038/33435 (1998).

106 Chiang, C., Litingtung, Y., Harris, M. P., Simandl, B. K., Li, Y., Beachy, P. A. & Fallon, J. F. Manifestation of the limb prepattern: limb development in the absence of sonic hedgehog function. *Dev Biol* **236**, 421-435, doi:10.1006/dbio.2001.0346 (2001).

107 Bawle, E. V., Conard, J. V. & Weiss, L. Adult and two children with fetal methotrexate syndrome. *Teratology* **57**, 51-55, doi:10.1002/(SICI)1096-9926(199802)57:2<51::AID-TERA2>3.0.CO;2-9 (1998).

108 *Rare Diseases Act of 2002*, <http://history.nih.gov/research/downloads/PL107-280.pdf>.

109 *SeattleSeq Annotation*, <http://snp.gs.washington.edu/SeattleSeqAnnotation/>.

110 *Polyphen-2*, <http://genetics.bwh.harvard.edu/pph2/>.

111 Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-35, doi:10.1038/ng.499 (2010).

112 Prasad, R., Zhadanov, A. B., Sedkov, Y., Bullrich, F., Druck, T., Rallapalli, R., Yano, T., Alder, H., Croce, C. M., Huebner, K. *et al.* Structure and expression pattern of human ALR, a novel gene with strong homology to ALL-1 involved in acute leukemia and to Drosophila trithorax. *Oncogene* **15**, 549-560, doi:10.1038/sj.onc.1201211 (1997).

113 Issaeva, I., Zonis, Y., Rozovskaia, T., Orlovsky, K., Croce, C. M., Nakamura, T., Mazo, A., Eisenbach, L. & Canaani, E. Knockdown of ALR (MLL2) reveals ALR target genes and leads to alterations in cell adhesion and growth. *Mol Cell Biol* **27**, 1889-1903, doi:10.1128/MCB.01506-06 (2007).

114 Niikawa, N., Matsuura, N., Fukushima, Y., Ohsawa, T. & Kajii, T. Kabuki make-up syndrome: a syndrome of mental retardation, unusual facies, large and protruding ears, and postnatal growth deficiency. *J Pediatr* **99**, 565-569 (1981).

115 Kuroki, Y., Suzuki, Y., Chyo, H., Hata, A. & Matsui, I. A new malformation syndrome of long palpebral fissures, large ears, depressed nasal tip, and skeletal anomalies associated with postnatal dwarfism and mental retardation. *J Pediatr* **99**, 570-573 (1981).

116 Niikawa, N., Kuroki, Y., Kajii, T., Matsuura, N., Ishikiriyama, S., Tonoki, H., Ishikawa, N., Yamada, Y., Fujita, M., Umemoto, H. *et al.* Kabuki make-up (Niikawa-Kuroki) syndrome: a study of 62 patients. *Am J Med Genet* **31**, 565-589, doi:10.1002/ajmg.1320310312 (1988).

117 Courtens, W., Rassart, A., Stene, J. J. & Vamos, E. Further evidence for autosomal dominant inheritance and ectodermal abnormalities in Kabuki syndrome. *Am J Med Genet* **93**, 244-249 (2000).

118 Cooper, G. M., Stone, E. A., Asimenos, G., Program, N. C. S., Green, E. D., Batzoglou, S. & Sidow, A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913, doi:10.1101/gr.3577405 (2005).

119 Cooper, G. M., Goode, D. L., Ng, S. B., Sidow, A., Bamshad, M. J., Shendure, J. & Nickerson, D. A. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods* **7**, 250-251, doi:10.1038/nmeth0410-250 (2010).

120 Tonoki, H., Saitoh, S. & Kobayashi, K. Patient with del(12)(q12q13.12) manifesting abnormalities compatible with Noonan syndrome. *Am J Med Genet* **75**, 416-418 (1998).

121 *Phaster*, <http://www.phrap.org>.

122 Brkanac, Z., Spencer, D., Shendure, J., Robertson, P. D., Matsushita, M., Vu, T., Bird, T. D., Olson, M. V. & Raskind, W. H. IFRD1 is a candidate gene for SMNA on chromosome 7q22-q23. *Am J Hum Genet* **84**, 692-697, doi:10.1016/j.ajhg.2009.04.008 (2009).

123 Nikopoulos, K., Gilissen, C., Hoischen, A., van Nouhuys, C. E., Boonstra, F. N., Blokland, E. A., Arts, P., Wieskamp, N., Strom, T. M., Ayuso, C. *et al.* Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. *Am J Hum Genet* **86**, 240-247, doi:10.1016/j.ajhg.2009.12.016 (2010).

124 Hoischen, A., van Bon, B. W., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., de Vries, P., de Reuver, R., Wieskamp, N., Mortier, G. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet* **42**, 483-485, doi:10.1038/ng.581 (2010).

125 Sobreira, N. L., Cirulli, E. T., Avramopoulos, D., Wohler, E., Oswald, G. L., Stevens, E. L., Ge, D., Shianna, K. V., Smith, J. P., Maia, J. M. *et al.* Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet* **6**, e1000991, doi:10.1371/journal.pgen.1000991 (2010).

126 Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., Beck, A. E., Tabor, H. K., Cooper, G. M., Mefford, H. C. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42**, 790-793, doi:10.1038/ng.646 (2010).

127 Volpi, L., Roversi, G., Colombo, E. A., Leijsten, N., Concolino, D., Calabria, A., Mencarelli, M. A., Fimiani, M., Macciardi, F., Pfundt, R. *et al.* Targeted next-generation sequencing appoints c16orf57 as clericuzio-type poikiloderma with neutropenia gene. *Am J Hum Genet* **86**, 72-76, doi:10.1016/j.ajhg.2009.11.014 (2010).

128 Edvardson, S., Shaag, A., Zenvirt, S., Erlich, Y., Hannon, G. J., Shanske, A. L., Gomori, J. M., Ekstein, J. & Elpeleg, O. Joubert syndrome 2 (JBTS2) in Ashkenazi Jews is associated with a TMEM216 mutation. *Am J Hum Genet* **86**, 93-97, doi:10.1016/j.ajhg.2009.12.007 (2010).

129 Rehman, A. U., Morell, R. J., Belyantseva, I. A., Khan, S. Y., Boger, E. T., Shahzad, M., Ahmed, Z. M., Riazuddin, S., Khan, S. N., Riazuddin, S. *et al.* Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. *Am J Hum Genet* **86**, 378-388, doi:10.1016/j.ajhg.2010.01.030 (2010).

130 Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-639, doi:10.1126/science.1186802 (2010).

131 Lalonde, E., Albrecht, S., Ha, K. C., Jacob, K., Bolduc, N., Polychronakos, C., Dechelotte, P., Majewski, J. & Jabado, N. Unexpected allelic heterogeneity and spectrum of mutations in

Fowler syndrome revealed by next-generation exome sequencing. *Hum Mutat* **31**, 918-923, doi:10.1002/humu.21293 (2010).

132 Walsh, T., Shahin, H., Elkan-Miller, T., Lee, M. K., Thornton, A. M., Roeb, W., Abu Rayyan, A., Loulus, S., Avraham, K. B., King, M. C. *et al.* Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *Am J Hum Genet* **87**, 90-94, doi:10.1016/j.ajhg.2010.05.010 (2010).

133 Johnston, J. J., Teer, J. K., Cherukuri, P. F., Hansen, N. F., Loftus, S. K., Center, N. I. H. I. S., Chong, K., Mullikin, J. C. & Biesecker, L. G. Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet* **86**, 743-748, doi:10.1016/j.ajhg.2010.04.007 (2010).

134 O'Roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S., Karakoc, E., Mackenzie, A. P., Ng, S. B., Baker, C. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* **43**, 585-589, doi:10.1038/ng.835 (2011).

135 Hoischen, A., van Bon, B. W., Rodriguez-Santiago, B., Gilissen, C., Vissers, L. E., de Vries, P., Janssen, I., van Lier, B., Hastings, R., Smithson, S. F. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat Genet* **43**, 729-731, doi:10.1038/ng.868 (2011).

136 Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D. A. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* **362**, 1181-1191, doi:10.1056/NEJMoa0908094 (2010).

137 Chou, L. S., Liu, C. S., Boese, B., Zhang, X. & Mao, R. DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model. *Clin Chem* **56**, 62-72, doi:10.1373/clinchem.2009.132639 (2010).

138 Hoischen, A., Gilissen, C., Arts, P., Wieskamp, N., van der Vliet, W., Vermeer, S., Steehouwer, M., de Vries, P., Meijer, R., Seiqueros, J. *et al.* Massively parallel sequencing of ataxia genes after array-based enrichment. *Hum Mutat* **31**, 494-499, doi:10.1002/humu.21221 (2010).

139 Vasta, V., Ng, S. B., Turner, E. H., Shendure, J. & Hahn, S. H. Next generation sequence analysis for mitochondrial disorders. *Genome Med* **1**, 100, doi:10.1186/gm100 (2009).

140 Raca, G., Jackson, C., Warman, B., Bair, T. & Schimmenti, L. A. Next generation sequencing in research and diagnostics of ocular birth defects. *Mol Genet Metab* **100**, 184-192, doi:10.1016/j.ymgme.2010.03.004 (2010).

141 Bentley, G., Higuchi, R., Hoglund, B., Goodridge, D., Sayer, D., Trachtenberg, E. A. & Erlich, H. A. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens* **74**, 393-403, doi:10.1111/j.1399-0039.2009.01345.x (2009).

142 Cooper, D. N., Chen, J. M., Ball, E. V., Howells, K., Mort, M., Phillips, A. D., Chuzhanova, N., Krawczak, M., Kehrer-Sawatzki, H. & Stenson, P. D. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum Mutat* **31**, 631-655, doi:10.1002/humu.21260 (2010).

143 Siepel, A., Pollard, K. & Haussler, D. New methods for detecting lineage-specific selection. *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)*, 190–205 (2006).

144 Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553-1561, doi:10.1101/gr.092619.109 (2009).

145 *EURORDIS: Rare Diseases Europe*, <http://www.eurordis.org/about-rare-diseases>.

146 *1000 Genomes Project*, <http://www.1000genomes.org>.

147 National Center for Biotechnology Information. *Database of Single Nucleotide Polymorphisms (dbSNP)*, <http://www.ncbi.nlm.nih.gov/SNP/>.

148 Conrad, D. F., Keebler, J. E., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**, 712-714, doi:10.1038/ng.862 (2011).

149 Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat Rev Genet* **13**, 565-575, doi:10.1038/nrg3241 (2012).

150 Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19**, 1316-1323, doi:10.1101/gr.080531.108 (2009).

151 Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**, D32-36, doi:10.1093/nar/gkn721 (2009).

152 Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Unlocking Mendelian disease using exome sequencing. *Genome Biol* **12**, 228, doi:10.1186/gb-2011-12-9-228 (2011).

153 Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L. *et al.* Ensembl 2009. *Nucleic Acids Res* **37**, D690-697, doi:10.1093/nar/gkn828 (2009).

154 Kuhn, R. M., Karolchik, D., Zweig, A. S., Wang, T., Smith, K. E., Rosenbloom, K. R., Rhead, B., Raney, B. J., Pohl, A., Pheasant, M. *et al.* The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* **37**, D755-761, doi:10.1093/nar/gkn875 (2009).

155 Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., Lagarde, J., Gilbert, J. G., Storey, R., Swarbreck, D. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**, S4 1-9, doi:10.1186/gb-2006-7-s1-s4 (2006).

156 Clark, M. J., Chen, R., Lam, H. Y., Karczewski, K. J., Chen, R., Euskirchen, G., Butte, A. J. & Snyder, M. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* **29**, 908-914, doi:10.1038/nbt.1975 (2011).

157 Coffey, A. J., Kokocinski, F., Calafato, M. S., Scott, C. E., Palta, P., Drury, E., Joyce, C. J., Leproust, E. M., Harrow, J., Hunt, S. *et al.* The GENCODE exome: sequencing the complete human exome. *Eur J Hum Genet* **19**, 827-831, doi:10.1038/ejhg.2011.28 (2011).

158 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

159 DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

160 Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., Dudley, J. T., Ormond, K. E., Pavlovic, A., Morgan, A. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525-1535, doi:10.1016/S0140-6736(10)60452-7 (2010).

161 Consortium, E. P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**, e1001046, doi:10.1371/journal.pbio.1001046 (2011).

162 Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D. & Shendure, J. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**, 1173-1175, doi:10.1038/nbt.1589 (2009).

163 Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S. I., Cooper, G. M. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265-270, doi:10.1038/nbt.2136 (2012).

164 Hannibal, M. C., Buckingham, K. J., Ng, S. B., Ming, J. E., Beck, A. E., McMillin, M. J., Gildersleeve, H. I., Bigham, A. W., Tabor, H. K., Mefford, H. C. *et al.* Spectrum of MLL2 (ALR) mutations in 110 cases of Kabuki syndrome. *Am J Med Genet A* **155A**, 1511-1516, doi:10.1002/ajmg.a.34074 (2011).

165 Micale, L., Augello, B., Fusco, C., Selicorni, A., Loviglio, M. N., Silengo, M. C., Reymond, A., Gumiero, B., Zucchetti, F., D'Addetta, E. V. *et al.* Mutation spectrum of MLL2 in a cohort of Kabuki syndrome patients. *Orphanet J Rare Dis* **6**, 38, doi:10.1186/1750-1172-6-38 (2011).

166 Li, Y., Bogershausen, N., Alanay, Y., Simsek Kiper, P. O., Plume, N., Keupp, K., Pohl, E., Pawlik, B., Rachwalski, M., Milz, E. *et al.* A mutation screen in patients with Kabuki syndrome. *Hum Genet* **130**, 715-724, doi:10.1007/s00439-011-1004-y (2011).

167 Paulussen, A. D., Stegmann, A. P., Blok, M. J., Tserpelis, D., Posma-Velter, C., Detisch, Y., Smeets, E. E., Wagemans, A., Schrander, J. J., van den Boogaard, M. J. *et al.* MLL2 mutation spectrum in 45 patients with Kabuki syndrome. *Hum Mutat* **32**, E2018-2025, doi:10.1002/humu.21416 (2011).

168 *NHLBI Exome Sequencing Project (ESP) Exome Variant Server*, <http://evs.gs.washington.edu/EVS>.

169 Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

VITA

Sarah B Ng was born in Singapore. She completed her secondary and pre-tertiary education there, after which she received a scholarship from the Agency for Science, Technology and Research, Singapore, to perform undergraduate and post-graduate studies overseas. She attended the University of Wisconsin-Madison where she earned a Bachelor of Science (Honors) in Molecular Biology. In 2012, she earned a Doctor of Philosophy at the University of Washington in Genome Sciences.