Genet. Res., Camb. (2015), vol. 97, e15. © Cambridge University Press 2015
doi:10.1017/S0016672315000166

The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders

TONY SHEN[1,2], ARIEL LEE[1,3], CAROL SHEN[1,2] AND C.JIMMY LIN[1]*
[1]*Rare Genomics Institute, 5225 Pooks Hills Road, Suite 1701N, Bethesda, MD 20814, USA*
[2]*Washington University School of Medicine, 660 South Euclid Avenue, Saint Louis, MO 63110, USA*
[3]*Nova Southeastern University, College of Osteopathic Medicine, 3301 College Avenue, Ft. Lauderdale, FL 333314-796, USA*

(*Received 23 February 2014; revised 25 June 2015; accepted 29 June 2015*)

Summary

There are an estimated 6000–8000 rare Mendelian diseases that collectively affect 30 million individuals in the United States. The low incidence and prevalence of these diseases present significant challenges to improving diagnostics and treatments. Next-generation sequencing (NGS) technologies have revolutionized research of rare diseases. This article will first comment on the effectiveness of NGS through the lens of long-tailed economics. We then provide an overview of recent developments and challenges of NGS-based research on rare diseases. As the quality of NGS studies improve and the cost of sequencing decreases, NGS will continue to make a significant impact on the study of rare diseases moving forward.

1. Introduction and background

Rare diseases, or orphan diseases, collectively affect millions of individuals worldwide. There currently exists an estimated 6000–8000 rare diseases, 75% of which affect children. An estimated 30 million people in the United States and 30 million in the European Union are diagnosed with a rare disease. A total of 30% of affected individuals die before 5 years of age. In the United States, rare disease is defined as a condition that affects fewer than 200,000 people. Historically, the low incidence and prevalence of these diseases have presented major challenges to the development of diagnostics and treatments (http://rarediseases.info.nih.gov/).

The increasingly widespread use of NGS technologies has revolutionized the study of rare diseases, of which 80% have genetic etiologies (Yaneva–Deliverska, 2011). For Mendelian disorders, sequencing enables researchers to understand specific diseases in great detail and informs the development of new treatments. Between 2007 and 2014, the number of disease phenotypes with characterized genetic causes has more than doubled (Koboldt *et al.*, 2013). Whole-exome sequencing (WES) and whole-genome sequencing (WGS) strategies allow researchers to study a wide range of diseases through a common work flow.

With NGS in place as an effective tool for the study of rare diseases, coordinated research efforts play a significant role in advancing research (Griggs *et al.*, 2009). The first systematic effort to address rare disease in the United States began with the Orphan Drug Act of 1983 (Orphan Drug Act, 2049 vols, United States of America, 1983). This legislation, administered by the FDA Office of Orphan Products Development (OOPD), created incentives for the development of drugs that specifically targeted rare diseases. Additionally, the Orphan Drug Act allowed for the repurposing of available drugs originally indicated for other conditions. Since this legislation was enacted, more than 300 drugs have been developed to treat rare diseases (Griggs *et al.*, 2009). Many more organizations now exist for the purpose of advancing rare disease research, including the Undiagnosed Disease Program (NIH) and the International Rare Disease Research Consortium (Danielsson *et al.*, 2014). Moving forward, the study of rare diseases will require the coordination of technological advances, institutional collaboration and financial resources.

The development of rare disease research can be described as a "long-tailed" problem. Originally developed to understand the rise of internet retailers,

*Corresponding author: E-mail: jimmy.lin@raregenomics.org

long-tail concepts broadly describe the challenges of organizing and accessing disparate collections of items. We believe these economic concepts are helpful for understanding the recent advances and challenges of rare disease research. In this article, we will first provide a discussion of the long tail and its applicability to this field. We will then review the impact of the latest advances in NGS technology and research infrastructure.

2. Rare disease as a long-tailed problem

To understand the relevance of long-tail economics in the context of rare disease research, we must first examine its emergence. The long-tail concept has recently gained attention for its usefulness in understanding the recent success of internet-based retailers. In business, the Pareto model (or 80-20 rule) describes how roughly 20% of products typically account for 80% of income, serving as a reasonable estimate of the relationship between products and revenue (Pareto & Businо, 1964; Arnold, 2004; Anderson, 2006) (Fig. 1). For example, out of all the films produced by the entertainment industry, only a handful of blockbusters account for the majority of revenue generated. The lesser known films, though greatly outnumbering the blockbusters, generate far less revenue per film. This principle is generalizable to numerous other industries as well.

The success of these companies may be understood within the framework of supply and demand. On the supply side, centralized warehouses or servers significantly decrease the cost of maintaining a large library of niche products. While brick-and-mortar stores need to carefully stock items based on their popularity, centralized supply chain management allows for economically justifiable stocking of niche products. On the demand side, improved search and prediction algorithms not only allow businesses to tap into existing demand for niche products, but also generates new demand based on the individual consumer's tastes. In addition, increased efficiency from electronic payment systems decrease the transactional costs of selling niche items to economically feasible levels (Bakos, 1998). Taken together, these forces "open up" the long tail for business, connecting consumers to a previously inaccessible body of products.

For the purposes of this article, we derive two themes from the technologies outlined above: increased access and reduced cost. Increased access refers to the ability of search algorithms to guide the consumer through a sea of products in the long tail and find what they are looking for. Reduced cost refers to the facilitation of online transactions or reduction in overhead costs by centralizing resources.

In the context of rare Mendelian disease research, access and cost summarize two of the greatest challenges facing the field today – challenges that are being

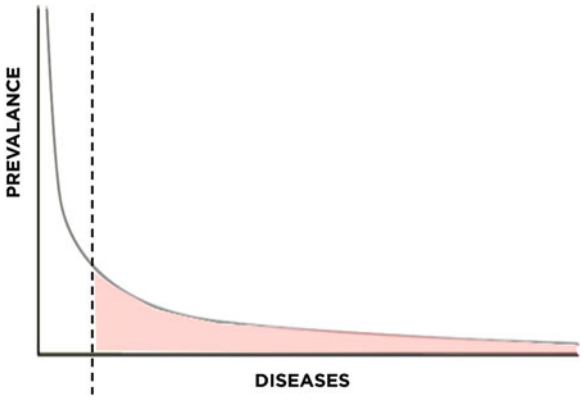

Fig. 1. Pareto distribution with long tail highlighted.

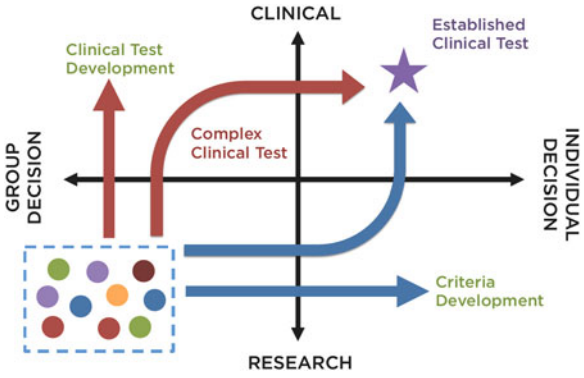

Fig. 2. Participant selection for genome sequencing. The x-axis describes decision-making for participant selection. The y-axis describes the purpose of sequencing the participant in question.

overcome by NGS technologies. The principles derived from long-tail economics shape our understanding of the recent development of this field and offer insight towards needed improvements. Traditional research has focused on the 20% of genetic conditions that account for 80% of disease, neglecting the remaining 80% of genetic conditions that account for a nontrivial 20% of disease. Long-tail concepts illustrate that the demand for this 20% of disease is significant; opening up this long tail will require integrated technological, organizational and financial solutions.

3. The impact of next-generation sequencing

(i) *Participant selection*

NGS for rare diseases begins with selecting the participants who will be sequenced. From a macroscopic perspective, changes in participant selection occur on two axes: group/individual and research/clinical (Fig. 2). Here, we overview the shifts in participant selection and propose that the field needs to move towards an

individually-initiated and clinically-oriented mode of determining candidates for sequencing.

Before NGS, researchers relied on methods such as chromosomal linkage association within families to identify Mendelian diseases (Ku *et al.*, 2011). This type of study would be placed in the lower left quadrant of Fig. 2 as a research focused endeavor with data restricted to the investigators. Early genomic studies such as micro-array-based genome-wide association studies (GWAS) also fall into the same quadrant, with notable examples such as the Wellcome Trust Consortium study that examined 14 000 cases of common diseases (Wellcome Trust Case Control Consortium, 2007). These studies are often large-scale projects involving one or more research centers. Participant selection depended upon the research aims of the project, allowing only suitable candidates to undergo genome sequencing. For rare diseases, the prevalence may be too low for large studies. *De novo* mutations may also occur in unrelated individuals. Thus, the decision of who can be sequenced needs to shift from research groups to individual patients in order to better focus research efforts on a broader number of rare diseases (Fig. 2). This movement requires new research criteria that can accommodate low sample sizes, unrelated patients and the opportunity to provide more patients with genomic sequencing. NGS has enabled investigators to identify a large number of disease-causing genes. Table 1 shows the number of entries in the Online Mendelian Inheritance in Man (OMIM) database for which the molecular basis of a particular phenotype is known. Between 2007 and 2014, the number of entries more than doubled, with 428 new entries added between 2013 and 2014. The investigators sequenced the exomes of four patients to identify DHODH as a candidate gene and later confirmed this finding in three other families by Sanger sequencing. Because NGS enables researchers to discover disease-causing genes from such small sample sizes, the threshold for offering sequencing for affected patients has decreased dramatically. However, these studies still depend on research initiatives and are limited by the logistical challenge of connecting patients to studies.

In addition to the shift from group to individual decisions, participant selection for rare disease research must also shift from being research focused to clinically focused. This movement will eventually lead to readily available established clinical tests for rare diseases based on NGS (Boyd, 2013). Currently, NGS is not regularly used as a primary diagnostic tool. However, as the number of discovered gene–phenotype associations increase, clinicians will be more likely to diagnose patients based on their sequencing data, blurring the line between research and clinical genetic testing (Boycott *et al.*, 2013; Delanty & Goldstein, 2013). Already, clinicians have used NGS as a supplementary diagnostic tool in limited contexts. Exome sequencing has been used to diagnose congenital chloride diarrhea in a cohort of patients suspected to have Bartter syndrome (Choi *et al.*, 2009). In this particular case, all six patients were found to have a deletion in SLC26A3, a chloride anion exchanger, leading to the first ever diagnosis based on exome sequencing (Rizzo & Buck, 2012). Exome sequencing has also been used to diagnose a child presenting with inflammatory bowel disease with a mutation in X-linked inhibitor of apoptosis protein (XIAP) (Worthey *et al.*, 2011). In this particular case, finding this mutation led clinicians to perform a haematopoetic progenitor cell transplant because mutations in XIAP increase risk of death due to haemophagocytic lymphohistiocytosis. The decision to perform this invasive yet ultimately effective procedure would not have been made without NGS.

We have focused our attention thus far on rare Mendelian diseases. However, we must remember that there remain many other diseases for which NGS does not drive a similar degree of change. Figure 3 highlights the segment of diseases particularly suited for NGS-based research. For the patients in the starred area, NGS is driving a trend towards clinically-oriented testing initiated by the patient.

Table 1. *The number of OMIM phenotypes for which the molecular basis is known since 2007 (Koboldt et al., 2013; Online Mendelian Inheritance in Man).*

Inheritance pattern	January 2007	July 2013	July 2014
Autosomal	1851	3525	3852
X linked	169	277	287
Y linked	2	4	4
Mitochondrial	26	28	28
Total	2048	3843	4171

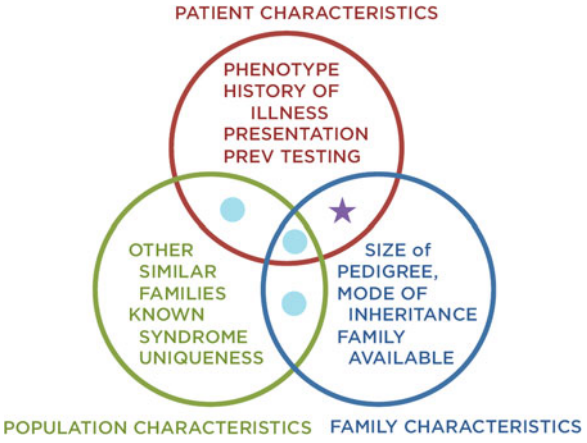

Fig. 3. Overview of disease qualities. The starred region represents diseases most likely to benefit from genomic sequencing.

Genetic testing has a long history prior to the development of NGS. First generation Sanger sequencing, still considered the gold standard for accuracy, has long been used for single-gene studies. Examples include BRCA1 and BRCA2 testing for women with family histories of breast cancer (Wooster *et al.*, 1995; Nelson *et al.*, 2005). Gene panels build upon the single-gene study by testing for multiple candidate genes at once. The ability to efficiently sequence for hundreds of candidate genes produced a large number of GWAS studies. Clinically, more and more institutions and companies offer gene panels for cancer patients in order to generate a more precise genetic profile. For rare diseases, the GWAS study approach is severely limited because there must be a known candidate gene to target. Additionally, most GWAS studies followed a case-control study design in which conclusions were based on genetic differences between the case and control groups. An association between a mutation and a disease could only be made with large enough sample sizes. This type of classical epidemiological study design is unfeasible with rare disease research due to an inherent lack of available research subjects. NGS-based research instead relies on a number of bioinformatic strategies to correctly identify rare variants within a small sample size (Boyd, 2013). Examples of these statistical methods include the burden test and variance-component test, which are reviewed in greater depth by Lee *et al.* (2014).

WES has dominated rare disease research in recent years. Compared to WGS, exome sequencing covers only the 1% (~30Mb) of the genome that is translated into protein (Bamshad *et al.*, 2011). Compared to WGS, WES offers a significantly more cost-effective and time-effective method of collecting and analyzing genomic data. Ng *et al.* first sequenced the exomes of 12 individuals with Freeman-Sheldon syndrome (OMIM 193700) in order to demonstrate the feasibility of exome sequencing as a method, identifying rare and common variants in both related and unrelated individuals (Ng *et al.*, 2009). The same investigators later used this method to identify DHODH as the causative gene for Miller syndrome (OMIM 263750) and MLL2 for Kabuki syndrome (OMIM 147920) (Ng *et al.*, 2010 *a*; Ng *et al.*, 2010 *b*). Interestingly, the initial analysis for the Kabuki syndrome study did not reveal any candidate genes. Due to the phenotypic heterogeneity of Kabuki syndrome, the investigators accounted for phenotypic severity by assigning a qualitative score to each patient based on physical features of the disease. After factoring this score into the analysis, MLL2 emerged as the sole candidate gene (Ng *et al.*, 2010 *a*). These classic examples look for shared mutations between unrelated individuals. Other examples of this strategy include the identification of SETB1 for Schinzel-Giedion syndrome (OMIM 269150) and ASXL1 for Bohring-Opitz syndrome (OMIM 605039) (Hoischen *et al.*, 2010; Hoischen *et al.*, 2011). Investigators have also successfully used other strategies of interpreting exome sequencing data to identify disease-causing genes. A summary of these other strategies are reviewed by Boyd (2013) and Koboldt *et al.* (2013).

Though exome sequencing has proven to be a productive study method, these studies do not cover the remaining 99% of the genome which is non-coding. The NIH-curated catalog of GWAS studies shows that the majority of GWAS loci lie in non-coding regions (Hindorff *et al.*, 2009; Lee *et al.*, 2014). In addition, initiatives such as the ENCODE Project are beginning to elucidate the functions of introns (ENCODE Project Consortium, 2012). Thus, WGS holds great potential for discovering disease-causing mutants in regions outside of the exon. The primary barrier to this study method has been its prohibitive cost, though the continuous decreasing cost of NGS has enabled investigators to use WGS to identify disease-causing genes. In 2013, Wang *et al.* used a combination of WGS and WES to identify mutations in RBCK1 as the cause for a novel Mendelian disease with cardiac and neuromuscular involvement (Wang *et al.*, 2013). While exome sequencing may be particularly suited to solving known Mendelian diseases for which the genetic aetiology is unknown, the authors claim that the data generated from WGS is more suited to the task of discovering the genetic basis of yet unknown diseases. Another group also used a combination of WGS and WES to identify a frameshift mutation in HMGB3 as the cause for X-linked colobomatous microphthalmia (OMIM 309800) (Scott *et al.*, 2014). The authors write that the increased coverage generated from multiple orthogonal sequencing methods improved their ability to identify variants over a single-technique approach (Scott *et al.*, 2014). Enns *et al.* also used a combination of WGS and WES to identify NGLY-1 deficiency as the cause of a glycosylation disorder found in eight patients in 2014 (Enns *et al.*, 2014). As a side note, this particular study gained significant attention in mainstream media, bringing NGS into the spotlight (Might & Wilsey, 2014; Mnookin, 2014). While these studies demonstrate the promise of WGS as a study method, there remain some challenges to routine implementation. An exploratory study to assess the clinical significance of WGS findings revealed that coverage for up to 19% of inherited disease genes were not up to accepted standards (Dewey *et al.*, 2014). An assessment of exome sequencing performance revealed that WES could actually capture small variants missed by WGS (Clark *et al.*, 2011).

As NGS technologies develop, there will be greater diversity in research methodology. Table 2 provides an overview of various strategies currently in use.

Table 2. *Summary of advantages and disadvantages of NGS study methods, adapted from Lee et al. (2014).*

	Advantages	Disadvantages
GWAS Chip	Inexpensive	Low accuracy for rare variants, relies on large sample size
Exome Chip	Less costly than WES	Limited to target regions; cannot identify extremely rare variants
WES	Ability to identify all variants in the exome, less costly than WGS, strong record of gene discovery	Does not sequence non-coding regionsUnable to identify structure variation
Low-depth WGS	Cost-effective and useful for association mapping	Limited accuracy for rare-variant identification
High-depth WGS	Ability to identify nearly all variants in the entire genome with high confidence	Cost-prohibitive for large-scale studiesDifficulty detecting copy number repeats

There is a general tradeoff between breadth and depth of coverage. The ideal method, high-depth WGS, would cover the entire genome with sufficient depth, but this method is currently cost-prohibitive. The field of rare disease research has generally equilibrated around WES as the most practical balance of breadth, depth and cost. Table 3 summarizes selected studies from 2013–2014 that have identified disease-causing genes for Mendelian disorders using NGS.

(ii) *Sample preparation and enrichment*

Until WGS becomes a routine genetic test in the laboratory and clinic, there will be a need to selectively enrich target areas of the genome (Mamanova *et al.*, 2010). Enrichment techniques generally fall into two categories: amplification and hybridization-capture. Amplification techniques rely on PCR. Because PCR-based enrichment requires primers, this technique is able to enrich targeted sequences with high specificity. However, this technique does not scale up efficiently as the number of target regions increases. The commercially available RDT1000 (RainDance Technologies) addresses this limitation with a multiplex droplet system. Each microdroplet houses distinct PCR reactions, facilitating parallel enrichment of thousands of target sequences (Tewhey *et al.*, 2009). Amplification techniques are useful for studies in which there are fewer sequencing targets and has been used by clinical laboratories for diagnosis (Valencia *et al.*, 2013).

Hybrid-capture methods are the preferred method today for the efficient enrichment of the exome. Genomic DNA is first sheared and the library prepared with appropriate adaptors. Specialized probes then hybridize with target regions. The DNA-probe hybrids may be purified using a solid-phase (microarray) or solution-based method. Today, solution-based systems are the preferred method of exome capture because the procedure can be accomplished using common laboratory equipment. There are three main solution-based systems commercially available today: SeqCap EZ® (Roche NimbleGen), SureSelect® (Agilent Technologies) and TruSeq® (Illumina). Several investigators have analysed the technical performances of each of these systems (Asan *et al.*, 2011; Clark *et al.*, 2011; Parla *et al.*, 2011). These kits generally use the same workflow, differing mostly on probe design. The NimbleGen system covers fewer genomic regions, but requires the least amount of coverage to sensitively detect SNPs and small indels. Thus, the NimbleGen system is well suited to research within defined genomic regions. The Agilent and Illumina systems cover more variants than NimbleGen with additional sequencing. Notably, only the Illumina platform is able to enrich untranlated regions (Clark *et al.*, 2011). The platforms available today represent an equilibrium between cost and coverage. As the cost of sequencing continues to decrease, future enrichment systems are likely to focus on wide and high-quality capture of target DNA.

(iii) *Sequencing technology*

First generation DNA sequencing platforms relied on the Sanger dideoxy method. In the age of NGS, this method retains significant purpose in sequencing predetermined genes with high accuracy. Most NGS studies use Sanger sequencing to confirm the validity of the newly identified candidate gene. The main shortcoming of automated Sanger sequencing is the limited number of fragments that may be sequenced simultaneously.

A number of technologies overcame this challenge to achieve massively parallel sequencing. In general, NGS platforms begin with the preparation of a library of DNA fragments, which are then clonally amplified. Different strategies are then used to determine the sequence of each fragment, which are performed in parallel. The details of NGS platforms are reviewed extensively elsewhere (Shendure & Ji, 2008; Metzker, 2010; Liu *et al.*, 2012; Mardis, 2013). The following

Table 3. *Summary of disease-causing genes identified using NGS, 2013–2014.*

Disorder	OMIM	Gene(s)	Method	Purification	Sequencer	Citation
Mitochondrial infantile cardiomyopathy		*MLRP44*	WES	NimbleGen 2·1M Human Exome V2·0	Illumina Genome Analyzer	Carroll *et al.* (2013)
Spinocerebellar Ataxia 38		*ELOVL5*	Exome/linkage	Agilent SureSelect	ABI SOLiD	Di Gregorio *et al.* (2014)
Lenz microphthalmia	309800	*NAA10*	WES	Illumina Truseq	Illumina HiSeq 2000	Esmailpour *et al.* (2014)
Frontotemporal dementia		*TREM2*	WES	NimbleGen SeqCap EZ Exome Library	Illumina HiSeq 2000	Guerreiro *et al.* (2013)
Caroli disease		*PKHD1*	WES	Agilent SureSelectXT	Illumina HiSeq 2000	Hao *et al.* (2014)
Metacarpal 4–5 fusion	309630	*FGF16*	WES	Agilent SureSelect	Illumina HiSeq 2000	Jamsheer *et al.* (2013)
Macrophage activation syndrome; juvenile ideopathic arthritis		*LYST, MUNC13-4, STXBP2*	WES	Agilent SureSelectXT	Illuma HiSeq 2000	Kaufman *et al.* (2014)
Malignant hyperthermia	145600	*RYR1, CACNA1S*	WES	NimbleGen EZ Human Exome Library v2·0	Illuma HiSeq 2000	Kim *et al.* (2013)
Spinocerebellar ataxia		*TGM6*	WES	NimbleGen 2·1M Human Exome V2·0	Illumina Genome Analyzer II	Li *et al.* (2013)
Mandibulofacial dysotosis with microcephaly	610536	*EFTUD2*	WES	NimbleGen SeqCap EZ Exome Library	Illumina HiSeq 2000	Luquetti *et al.* (2013)
Infantile myofibromatosis	228550	*PDGFRB, NOTCH3*	WES	Agilent SureSelect All Exon V4 + UTR	Illumina HiSeq 2000	Martignetti *et al.* (2013)
Neonatal cholestasis		*AKR1D1, SKIV2L*	WES/ genotyping	In supplement	In supplement	Morgan *et al.* (2013)
Amnestic mild cognitive impairment		*CARD10, PARP1*	WES/imaging	Agilent SureSelect All Exon	Illumina HiSeq 2000	Nho *et al.* (2013)
Congenital hyperinsulinism	256450	*ABCC8, GLUD1, HNF1A, KGNH6, GNAS, ACABC, NOTCH2, RYR3, TRPV3, TRPC5, CAMK2D, PIK3R3, CDKAL1, SGN8A, KCNJ10, PDE4C, NOS2, SLC24A6, CAGNA1A, PC*	WES/SNP genotyping	Agilent SureSelect All Exon	Illumina Genome Analyzer Iix	Proverbio *et al.* (2013)
X-linked colobomatous microphthalmia syndrome	309800	*HMGB3*	WGS/WES	Agilent SureSelect X-exome	Illumina Genome Analyzer IIx, Complete Genomics	Scott *et al.* (2014)
Neu-Laxova Syndrome	256520	*PHGDH*	WES	In supplement	In supplement	Shaheen *et al.* (2014)
Saethre-Chotzen syndrome		*TCF12*	WES	In supplement	In supplement	Sharma *et al.* (2013)
Autosomal dominant autoinflammatory disorder		*MEFV*	WES	Agilent SureSelect All Exon V2	Illumina HiSeq 2000	Stoffels *et al.* (2014)

Familial cortical myoclonic tremor with epilepsy	*CNTN2*		WES/linkage	Agilent SureSelect	Illumina Genome Analyzer IIx	Stogmann *et al.* (2013)
Limb-girdle muscular dystrophy 1 F	*TNPO3*		WES	Agilent SureSelect Human All Exon	ABI SOLiD	Torella *et al.* (2013)
Novel disease with neuromuscular and cardiac involvement	*RBCK1*		WGS/WES/ SNP genotyping	Agilent SureSelect	Illumina Genome Analyzer II, Complete Genomics	Wang *et al.* (2013)
Ossification of the posterior longitudinal ligament	602475	*PTCH1, COL17A1*	WES	Agilent SureSelect	Illumina HiSeq 2000	Wei *et al.* (2014)
Congenital disorder of glycosylation (NGLY-1 deficiency)		*NGLY1*	WGS/WES	Agilent SureSelect	Illumina HiSeq 2000, Complete Genomics	Enns *et al.* (2014)

section will provide a basic overview of several systems in use today, as well as "third-generation" systems in development.

The Illumina HiSeq and Genome Analyzer platforms use a sequencing-by-synthesis (SBS) strategy. Library fragments are first clonally amplified in oil droplets. The amplified fragments then undergo stepwise elongation using modified fluorescent dNTPs. The dNTPs function as "reversible terminators," which allow elongation to pause and continue following the addition of a single nucleotide (Bentley *et al.*, 2008). As Each dNTP is cycled through, a high-resolution image sensor records fluorescent signals from millions of amplicons simultaneously. This cycle is repeated to generate a sequence for every amplified DNA fragment. The Illumina HiSeq platform has dominated NGS research in recent years.

The Roche/454 platform relies on the detection of pyrophosphate released during nucleotide incorporation. Using a SBS strategy, library fragments are clonally amplified and then elongated one nucleotide at a time. A system of luciferase, luciferin, ATP sulfurylase and adenosine-5-phosphosulfate respond to pyrophosphate release by emitting photons, which are detected by a camera (Liu *et al.*, 2012; Valencia *et al.*, 2013). Similar to the Illumina platform, each dNTP is cycled through with imaging following each dNTP addition. The pattern of photon emission can then be used to produce sequences for all the amplicons in parallel. The use of this platform has been declining, as Roche announced in 2013 that the 454 sequencing division would be discontinued.

Ion Torrent (Life Technologies) also uses a SBS strategy. Instead of fluorescence or pyrophosphate, the platform uses a pH-sensitive semiconductor to detect proton release following nucleotide incorporation. Library fragments are first amplified on beads and deposited onto a pH sensitive chip. As each dNTP cycles through, the pH sensor detects which amplicons underwent nucleotide incorporation. This technology notably avoids the imaging step used in Illumina HiSeq or Roche/454, enabling significantly shorter run times (Valencia *et al.*, 2013).

The ABI/SOLiD platform relies on ligation between DNA library fragments and specially-designed DNA probes. Amplified samples are incubated with single-stranded target DNA which are ligated to fluorescent probes. The system determines sequence based on changes in fluorescence, which are dependent on the ligation pattern of the sample to target DNA (Shendure & Ji, 2008).

New third-generation sequencing platforms improve upon current NGS technologies in several ways. First, third-generation technologies avoid amplification of library fragments by sequencing single molecules. This reduces biased reading of regions of the genome that were preferentially amplified and

also allows for the detection of DNA modifications. Second, third-generation technologies produce significantly longer read lengths (Chin *et al.*, 2013; Mardis, 2013). This decreases our reliance on alignment to a reference genome and enables sequencing of highly repetitive intronic regions. Finally, third-generation technologies have decreased run times, increasing the efficiency of genomic research.

Single-molecule real-time sequencing (SMRT, Pacific Biosciences) uses a system of fluorescent probes to detect nucleotide incorporation by DNA polymerase. The platform produces read lengths in the range of ~5000–6000 bp (English *et al.*, 2012; Chin *et al.*, 2013). Nanopore sequencing (Oxford Nanopore) exhibits similar features, though the technology is currently not widely available. The nanopore platform detects voltage changes across a lipid bilayer as a DNA strand is elongated through an α-haemolysin nanopore (Eid *et al.*, 2009). Read lengths fall around 4500 bp (Branton *et al.*, 2008; Laszlo *et al.*, 2014).

(iv) *Bioinformatics*

Bioinformatics refers to the computational processing and analysis of raw sequencing data. Detailed reviews of the NGS bioinformatics pipeline may be found elsewhere (Dolled-Filhart *et al.*, 2013; Hong *et al.*, 2013). In this section, we will briefly outline three general steps of bioinformatics analysis: alignment, variant calling and filtering/annotation. We will also discuss the development of cloud-based computational architectures as a strategy to increase efficiency and reduce cost.

The sequencing reads produced by NGS must first be mapped to a reference genome. To accomplish this, algorithms are designed to match fragment sequences with a reference while accounting for variations and errors (Rizzo & Buck, 2012; Shang *et al.*, 2014). This process is computationally intensive. There are two general types of alignment algorithms: hash-table and Burrows-Wheeler Transform (BWT). Examples of hash-table aligners include SeqMap, PASS, MAQ, GASSST, RMAP, PErM, GenomeMapper, BOAT and mrsFAST (Shang *et al.*, 2014).

Variant calling refers to the process of detecting differences, or variants, between the sample and reference sequences. Variant calling programs must distinguish between sequencing errors and true variants. Single-nucleotide polymorphisms, insertions and deletions are types of variation that may be detected in the sample DNA, each with different computational approaches. Popular programs used to detect SNPs include the Genome Analysis Toolkit (GATK), SOAPsnp and VarScan (Dolled-Filhart *et al.*, 2013; Hong *et al.*, 2013). Pindel, dindel and GATK are programs used to detect insertions and deletions (Dolled-Filhart *et al.*, 2013).

Following the generation of a list of variants, investigators need to identify the variants with a higher likelihood of contributing to disease. Filtering refers to the process of eliminating variants that may be explained by a specific genetic model. This can be accomplished by evaluating the subject's pedigree or comparing the sample sequencing to a normal control. Annotation refers to the process of identifying variants for which the biological function is known. Effective annotation requires the curation of a database of known variants. A list of programs for filtering and annotation may be found in a review by Dolled-Filhart *et al.* (2013).

The rate at which NGS throughput increases far outpaces the increase in computational performance (Schatz *et al.*, 2010). In order to handle the increasing volume of data generated by NGS, many investigators have turned to cloud computing architectures. Cloud computing enables efficient distribution of computational resources and allows for parallel work flows. To demonstrate the effectiveness of parallelized programming and cloud computing, Maji *et al.* modified an aligner to use parallel computations in place of serial ones to decrease execution time by 41% (Maji *et al.*, 2014). Investigators at Baylor University and the University of Minnesota have implemented cloud-based bioinformatics workflows, demonstrating increased efficiency and scalability (Onsongo *et al.*, 2014; Reid *et al.*, 2014).

(v) *Reporting results*

The rise in genomic data has produced a need for efficient platforms for data curation and sharing (No authors listed, 2014). While individual journals may have systems for managing published data, there is no central organization managing data between peer-reviewed journals (Tenopir *et al.*, 2011). There are over 600 subject-specific databases available, indicating probable redundancy between different databases (No authors listed, 2014). As computational methods continue to mature, a consistent system for genomic data will greatly facilitate data sharing. For the study of rare disease, OMIM is one of the most important databases keeping track of discovered genetic causes of Mendelian disorders . The 1000 Genomes Project organizes whole-genome data gathered internationally to serve as a reference for future genomic research (Siva, 2008). The NIH manages Gene Expression Omnibus (GEO) and RefSeq as repositories for sequencing data.

The increase in genomic data introduces ethical and legal questions regarding the "ownership" of data. Current regulations prevent investigators from sharing results with research subjects. Gholson Lyon reported being unable to share the results from an NGS study of Ogden syndrome with a participating family, a situation

that is detrimental to both the patient and investigator (Lyon, 2012). At least three challenges need to be addressed to improve communication between scientists and patients: logistical feasibility, data quality standards and availability of interventions (Lee & Lin, 2013).

First, reporting results to all participants may be logistically unfeasible for large studies. Researchers and patients need to develop reliable, private and secure means of communication. One way of addressing these challenges would be to develop sophisticated databases that can be accessed only by researchers. While database development may facilitate collaboration between investigators, there remains the issue of patient access to their own research data. Another strategy, patient-centric initiatives, addresses these challenges by allowing patients to determine access to their own research data. This approach depends on the development of specialized information technologies and has shown potential. While this approach shows promise, there needs to be a significant cultural shift in research and clinical practice in order for this concept to become widespread (Kaye *et al.*, 2012).

Second, the results of genomic research need to be held to a standard before they can be communicated to the patient. In the United States, the Clinical Laboratory Improvement Amendments (CLIA) regulates the quality and reliability of laboratory results. Given the rapidly changing nature of sequencing technologies and volume of new discoveries, a new standard is needed to certify actionable data in the context of genomics. In July 2014, the United States Food and Drug Administration (United States Food and Drug Administration, 2014) announced that they would increase their involvement in regulating diagnostic tests. Though some reacted to the announcement with trepidation, it remains to be seen what the impact of increased FDA regulation will be on ensuring the reliability of diagnostic sequencing data (Pollack, 2014).

Because many rare Mendelian disorders do not have interventions available, the availability of treatments must be considered when reporting results to a patient. A survey of patients indicated that 90% would prefer to know all individual results, including those that are unactionable (Kohane & Taylor, 2010). A total of 75% reported reduced willingness to participate in studies that do not report all results (Kohane & Taylor, 2010). Though no formal guidelines exist for the communication of unactionable test results, informed consent to all tests is crucial (Hunter *et al.*, 2012; Lee & Lin, 2013).

Genome sequencing also carries potential for incidental or secondary findings. Unlike unactionable findings, secondary findings may have implications for patient care. In 2013, the American College of Medical Genetics and Genomics (ACMG) published a group of genetic findings that should be reported if found secondarily (Green *et al.*, 2013). However, the ACMG acknowledged that this working list is imperfect and likely to change with new data (Green *et al.*, 2013). Prior to genetic testing, the patient should be counselled regarding likelihood of secondary findings and the type of results that will be disclosed (ACMG Board of Directors, 2012). In a survey of 200 patients who underwent diagnostic exome sequencing, 93·5% chose to receive secondary results (Shahmirzadi *et al.*, 2014). Moving forward, patient preference and better clinical data will continue to guidelines for disclosure of secondary findings.

(vi) *Reimbursement*

The dramatic decrease in the cost of sequencing has led to changes in the economics of rare disease research. Due to the small number of individuals affected by any particular rare disease, the cost of research has traditionally been a challenge in the field. For-profit organizations in particular face the difficulty of justifying the cost of research given the small return on investment. Government-funded research, though not motivated by profit, must also perform a cost-benefit analysis when awarding grants. Here, we frame the effects of NGS from two perspectives: top-down costs and bottom-up funding. We define top-down costs as the overall cost of research, while bottom-up funding refers to new initiatives to fund the increasing number of potential research projects made possible by NGS.

While the effect of decreasing sequencing costs on the amount of rare disease research is undeniable, there remains some debate on the true cost of sequencing. In 2009, the estimated cost for sequencing a human genome was $100 000. By 2014, Illumina has claimed to reach the $1000 genome threshold (Sboner *et al.*, 2011). During this time, thousands of genes underlying Mendelian diseases were discovered, marking an accelerated period of discovery (Table 1). In addition to decreased cost, we see the development of a common workflow for many exome- or genome-sequencing studies. Table 3 shows the recent dominance of the Illumina HiSeq system with Agilent and Roche Nimblegen as popular enrichment platforms. Aided by the availability of commercial kits at each step of the research process, rare disease research can benefit from economies of scale. While the expansion of rare disease research in the nascent years of genomic research is clear, researchers debate how to measure the true cost of sequencing. As sequencing becomes more widespread, new costs such as data management or computationally-intensive analysis will continue to emerge (Sboner *et al.*, 2011). Two literature reviews reveal a lack of high-quality economic data and thus the inability to form any conclusions with regard to the cost of expanding genomic research (Frank *et al.*, 2013; Gordon *et al.*, 2014).

As a long-tailed problem, continued discovery of rare diseases requires a funding infrastructure that can sustainably support the work needed to identify the great number of rare diseases. Government funding agencies face certain limitations due to its centralized bureaucratic organization. As such, the traditional funding model may not be ideal for rare disease research. The research requirements for a particular patient with a rare disease may not fit well into an existing grant. In these cases, crowdfunding, a model that leverages contributions from interested individuals, offers key advantages. Especially as research funding in the United States decreases, there is an increasing need for alternative funding sources. Just as internet retailers "opened up" the long tail by making niche products available to interested consumers, crowdfunding platforms connect highly-invested individuals to a particular research project. Dragojlovic and Lynd followed five crowdfunding campaigns in 2013, reporting that five out of six met or exceeded their goal (Dragojlovic & Lynd, 2014). The Rare Genomics Institute also uses crowdfunding as a key source of financial support for research projects.

4. Concluding remarks

In this article, we reviewed the impact of NGS on the study of rare Mendelian disorders. We identified two trends from long-tail concepts that are useful for describing NGS and rare disease research: increased access and reduced cost. Genome sequencing provides a relatively uniform workflow capable of studying a wide range of genetic diseases. This has led to an expansion of sequencing centers world-wide, providing more patients with access to genome sequencing. Data from these studies are stored and curated on publically accessible databases. The expanding body of identified genes for Mendelian disorders will lead to better diagnostics and will form the basis of new therapies. The cost of sequencing has reduced dramatically since the introduction of NGS. Additionally, the value of a sequenced genome continues to increase as more disease-causing genetic variants are identified. Taken together, it is clear that NGS has revolutionized the study of rare diseases and will continue to do so moving forward.

Declaration of Interest

None.

References

ACMG Board of Directors (2012). Points to consider in the clinical application of genomic sequencing. *Genetics in Medicine* **14**, 759–761.

Anderson, C. (2006). *The Long Tail: Why the Future of Business is Selling Less of More*. New York, USA: Hyperion.

Arnold, B. C. (2004). Pareto Distribution. In *Encyclopedia of Statistical Sciences*. New York, USA: John Wiley & Sons, Inc.

Asan, Xu, Y., Jiang, H., Tyler-Smith, C., Xue, Y., Jiang, T., Wang, J., Wu, M., Liu, X., Tian, G., Wang, J., Wang, J., Yang, H. & Zhang, X. (2011). Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biology* **12**, R95.

Bakos, Y. (1998). The emerging role of electronic marketplaces on the Internet. *Commununications of the ACM* **41**, 35–42.

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

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

Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. & Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59.

Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews. Genetics* **14**, 681–691.

Boyd, S. D. (2013). Diagnostic applications of high-throughput DNA sequencing. *Annual Review of Pathology: Mechanisms of Disease* **8**, 381–410.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V., Ramsey, J. M., Riehn, R., Soni, G. V., Tabard-Cossa, V., Wanunu, M., Wiggin, M. & Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nature Biotechnology* **26**, 1146–1153.

Carroll, C. J., Isohanni, P., Pöyhönen, R., Euro, L., Richter, U., Brilhante, V., Götz, A., Lahtinen, T., Paetau, A., Pihko, H., Battersby, B. J., Tyynismaa, H. & Suomalainen, A. (2013). Whole-exome sequencing identifies a mutation in the mitochondrial ribosome protein MRPL44 to underlie mitochondrial infantile cardiomyopathy. *Journal of Medical Genetics* **50**, 151–159.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W. & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**, 563–569.

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Ozen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S. & Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 19096–19101.

Clark, M. J., Chen, R., Lam, H. Y., Karczewski, K. J., Chen, R., Euskirchen, G., Butte, A. J., Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology* **29**, 908–914.

Danielsson, K., Mun, L. J., Lordemann, A., Mao, J. & Lin, C.-H. J. (2014). Next-generation sequencing applied to rare diseases genomics. *Expert Review of Molecular Diagnostics* **14**, 469–487.

Delanty, N. & Goldstein, D. B. (2013). Diagnostic exome sequencing: a new paradigm in neurology. *Neuron* **80**, 841–843.

Dewey, F. E., Grove, M. E., Pan, C., Goldstein, B. A., Bernstein, J. A., Chaib, H., Merker, J. D., Goldfeder, R. L., Enns, G. M., David, S. P., Pakdaman, N., Ormond, K. E., Caleshu, C., Kingham, K., Klein, T. E., Whirl-Carrillo, M., Sakamoto, K., Wheeler, M. T., Butte, A. J., Ford, J. M., Boxer, L., Ioannidis, J. P., Yeung, A. C., Altman, R. B., Assimes, T. L., Snyder, M., Ashley, E. A. & Quertermous, T. (2014). Clinical interpretation and implications of whole-genome sequencing. *JAMA* **311**, 1035–1045.

Di Gregorio, E., Borroni, B., Giorgio, E., Lacerenza, D., Ferrero, M., Lo Buono, N., Ragusa, N., Mancini, C., Gaussen, M., Calcia, A., Mitro, N., Hoxha, E., Mura, I., Coviello, D. A., Moon, Y. A., Tesson, C., Vaula, G., Couarch, P., Orsi, L., Duregon, E., Papotti, M. G., Deleuze, J. F., Imbert, J., Costanzi, C., Padovani, A., Giunti, P., Maillet-Vioud, M., Durr, A., Brice, A., Tempia, F., Funaro, A., Boccone, L., Caruso, D., Stevanin, G. & Brusco, A. (2014). ELOVL5 mutations cause spinocerebellar ataxia 38. *The American Journal of Human Genetics*. **95**, 209–217.

Dolled-Filhart, M. P., Lee, M., Ou-Yang, C.-W., Haraksingh, R. R. & Lin, J. C.-H. (2013). Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. *The Scientific World Journal* **2013**, 10.

Dragojlovic, N. & Lynd, L. D. (2014). Crowdfunding drug development: the state of play in oncology and rare diseases. *Drug Discovery Today* **19**, 1775–1780.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. & Turner, S. (2009). Real-Time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C. & Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS ONE* **7**, e47768.

Enns, G. M., Shashi, V., Bainbridge, M., Gambello, M. J., Zahir, F. R., Bast, T., Crimian, R., Schoch, K., Platt, J., Cox, R., Bernstein, J. A., Scavina, M., Walter, R. S., Bibb, A., Jones, M., Hegde, M., Graham, B. H., Need, A. C., Oviedo, A., Schaaf, C. P., Boyle, S., Butte, A. J., Chen, R., Clark, M. J., Haraksingh, R.; FORGE Canada Consortium, Cowan, T. M., He, P., Langlois, S., Zoghbi, H. Y., Snyder, M., Gibbs, R. A., Freeze, H. H. & Goldstein, D. B. (2014). Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genetics in Medicine* **16**, 751–758.

Esmailpour, T., Riazifar, H., Liu, L., Donkervoort, S., Huang, V. H., Madaan, S., Shoucri, B. M., Busch, A., Wu, J., Towbin, A., Chadwick, R. B., Sequeira, A., Vawter, M. P., Sun, G., Johnston, J. J., Biesecker, L. G., Kawaguchi, R., Sun, H., Kimonis, V. & Huang, T. (2014). A splice donor mutation in NAA10 results in the dysregulation of the retinoic acid signalling pathway and causes Lenz microphthalmia syndrome. *Journal of Medical Genetics* **51**, 185–196.

Frank, M., Prenzler, A., Eils, R. & Graf von der Schulenburg, J.-M. (2013). Genome sequencing: a systematic review of health economic evidence. *Health Economics Review* **3**, 29.

Gordon, J. E., Leiman, J. M., Deland, E. L. & Pardes, H. (2014). Delivering value: provider efforts to improve the quality and reduce the cost of health care. *Annual Review of Medicine* **65**, 447–458.

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., McGuire, A. L., Nussbaum, R. L., O'Daniel, J. M., Ormond, K. E., Rehm, H. L., Watson, M. S., Williams, M. S., Biesecker, L. G & American

College of Medical Genetics and Genomics (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine* **15**, 565–574.

Griggs, R. C., Batshaw, M., Dunkle, M., Gopal-Srivastava, R., Kaye, E., Krischer, J., Nguyen, T., Paulus, K., Merkel, P. A. & Rare Diseases Clinical Research Network (2009). Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular Genetics and Metabolism* **96**, 20–26.

Guerreiro, R. J., Lohmann, E., Brás, J. M., Gibbs, J. R., Rohrer, J. D., Gurunlian, N., Dursun, B., Bilgic, B., Hanagasi, H., Gurvit, H., Emre, M., Singleton, A. & Hardy, J. (2013). Using exome sequencing to reveal mutations in TREM2 presenting as a frontotemporal dementia-like syndrome without bone involvement. *JAMA Neurology* **70**, 78–84.

Hao, X., Liu, S., Dong, Q., Zhang, H., Zhao, J. & Su, L. (2014). Whole exome sequencing identifies recessive PKHD1 mutations in a Chinese twin family with Caroli disease. *PLoS ONE* **9**, e92661.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362–9367.

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., Devriendt, K., Amorim, M. Z., Revencu, N., Kidd, A., Barbosa, M., Turner, A., Smith, J., Oley, C., Henderson, A., Hayes, I. M., Thompson, E. M., Brunner, H. G., de Vries, B. B. & Veltman, J. A. (2010). *De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature Genetics* **42**, 483–485.

Hoischen, A., van Bon, B. W., Rodríguez-Santiago, B., Gilissen, C., Vissers, L. E., de Vries, P., Janssen, I., van Lier, B., Hastings, R., Smithson, S. F., Newbury-Ecob, R., Kjaergaard, S., Goodship, J., McGowan, R., Bartholdi, D., Rauch, A., Peippo, M., Cobben, J. M., Wieczorek, D., Gillessen-Kaesbach, G., Veltman, J. A., Brunner, H. G. & de Vries, B. B. (2011). *De novo* nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nature Genetics* **43**, 729–731.

Hong, H., Zhang, W., Shen, J., Su, Z., Ning, B., Han, T., Perkins, R., Shi, L. & Tong, W. (2013). Critical role of bioinformatics in translating huge amounts of next-generation sequencing data into personalized medicine. *Science China Life Sciences* **56**, 110–118.

Hunter, L. E., Hopfer, C., Terry, S. F. & Coors, M. E. (2012). Reporting actionable research results: shared secrets can save lives. *Science Translational Medicine* **4**, 143cm8.

Jamsheer, A., Zemojtel, T., Kolanczyk, M., Stricker, S., Hecht, J., Krawitz, P., Doelken, S. C., Glazar, R., Socha, M. & Mundlos, S. (2013). Whole exome sequencing identifies FGF16 nonsense mutations as the cause of X-linked recessive metacarpal 4/5 fusion. *Journal of Medical Genetics* **50**, 579–584.

Kaufman, K. M., Linghu, B., Szustakowski, J. D., Husami, A., Yang, F., Zhang, K., Filipovich, A. H., Fall, N., Harley, J. B., Nirmala, N. R. & Grom, A. A. (2014). Whole exome sequencing reveals overlap between macrophage activation syndrome in systemic juvenile idiopathic arthritis and familial hemophagocytic lymphohistiocytosis. *Arthritis & Rheumatology* **66**, 3486–3495

Kaye, J., Curren, L., Anderson, N., Edwards, K., Fullerton, S. M., Kanellopoulou, N., Lund, D., MacArthur, D. G., Mascalzoni, D., Shepherd, J., Taylor, P. L., Terry, S. F. & Winter, S. F. (2012). From patients to partners: participant-centric initiatives in biomedical research. *Nature Reviews. Genetics* **13**, 371–376.

Kim, J. H., Jarvik, G. P., Browning, B. L., Rajagopalan, R., Gordon, A. S., Rieder, M. J., Robertson, P. D., Nickerson, D. A., Fisher, N. A. & Hopkins, P. M. (2013). Exome sequencing reveals novel rare variants in the ryanodine receptor and calcium channel genes in malignant hyperthermia families. *Anesthesiology* **119**, 1054–1065.

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

Kohane, I. S. & Taylor, P. L. (2010). Multidimensional results reporting to participants in genomic studies: getting it right. *Science Translational Medicine* **2**, 37cm19.

Ku, C.-S., Naidoo, N. & Pawitan, Y. (2011). Revisiting Mendelian disorders through exome sequencing. *Human Genetics* **129**, 351–370.

Laszlo, A. H., Derrington, I. M., Ross, B. C., Brinkerhoff, H., Adey, A., Nova, I. C., Craig, J. M., Langford, K. W., Samson, J. M., Daza, R., Doering, K., Shendure, J. & Gundlach, J. H. (2014). Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnology* **32**, 829–833.

Lee, M. Jr & Lin, J. C.-H. (2013). Overcoming the obstacles to returning genomic research results. *Genetics Research* **95**, 45–50.

Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics* **95**, 5–23.

Li, M., Pang, S., Song, Y., Kung, M., Ho, S. L. & Sham, P. C. (2013). Whole exome sequencing identifies a novel mutation in the transglutaminase 6 gene for spinocerebellar ataxia in a Chinese family. *Clinical Genetics* **83**, 269–273.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* **2012**, 251364.

Luquetti, D. V., Hing, A. V., Rieder, M. J., Nickerson, D. A., Turner, E. H., Smith, J., Park, S., Cunningham, M. L. (2013). "Mandibulofacial dysostosis with microcephaly" caused by EFTUD2 mutations: expanding the phenotype. *American Journal of Medical Genetics Part A* **161**, 108–113.

Lyon, G. J. (2012). Personalized medicine: bring clinical standards to human-genetics research. *Nature* **482**, 300–301.

Maji, R. K., Sarkar, A., Khatua, S., Dasgupta, S. & Ghosh, Z. (2014). PVT: an efficient computational procedure to speed up next-generation sequence analysis. *BMC Bioinformatics* **15**, 167.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J. & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**, 111–118.

Mardis, E. R. (2013). Next-generation sequencing platforms. Annual Review of Analytical Chemistry *(*Palo Alto Calif.*)* **6**, 287–303.

Martignetti, J. A., Tian, L., Li, D., Ramirez, M. C., Camacho-Vanegas, O., Camacho, S. C., Guo, Y., Zand, D. J., Bernstein, A. M., Masur, S. K., Kim, C. E., Otieno, F. G., Hou, C., Abdel-Magid, N., Tweddale, B., Metry, D., Fournet, J. C., Papp, E., McPherson, E. W., Zabel, C., Vaksmann, G., Morisot, C., Keating, B., Sleiman, P. M., Cleveland, J. A., Everman, D. B., Zackai, E. & Hakonarson, H. (2013). Mutations in

PDGFRB cause autosomal-dominant infantile myofibromatosis. *The American Journal of Human Genetics* **92**, 1001–1007.

Metzker, M. L. (2010). Sequencing technologies – the next generation. *Nature Reviews. Genetics* **11**, 31–46.

Might, M. & Wilsey, M. (2014). The shifting model in clinical diagnostics: how next-generation sequencing and families are altering the way rare diseases are discovered, studied, and treated. *Genetics in Medicine*. **16**, 736–737.

Mnookin, S. (2014). One of a Kind. In *The New Yorker*. 21 July 2014.

Morgan, N. V., Hartley, J. L., Setchell, K. D., Simpson, M. A., Brown, R., Tee, L., Kirkham, S., Pasha, S., Trembath, R. C., Maher, E. R., Gissen, P. & Kelly, D. A. (2013). A combination of mutations in AKR1D1 and SKIV2L in a family with severe infantile liver. *Orphanet Journal of Rare Diseases* **8**, 74.

Nelson, H. D., Huffman, L. H., Fu, R. & Harris, E. L. (2005). Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: systematic evidence review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine* **143**, 362–379.

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., Lee, C., Turner, E. H., Smith, J. D., Rieder, M. J., Yoshiura, K., Matsumoto, N., Ohta, T., Niikawa, N., Nickerson, D. A., Bamshad, M. J. & Shendure, J. (2010 *a*). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics* **42**, 790–793.

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., Shendure, J. & Bamshad, M. J. (2010 *b*). Exome sequencing identifies the cause of a Mendelian disorder. *Nature Genetics* **42**, 30–35.

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., Bamshad, M., Nickerson, D. A. & Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276.

Nho, K., Corneveaux, J., Kim, S., Lin, H., Risacher, S., Shen, L., Swaminathan, S., Ramanan, V. K., Liu, Y., Foroud, T., Inlow, M. H., Siniard, A. L., Reiman, R. A., Aisen, P. S., Petersen, R. C., Green, R. C., Jack, C. R., Weiner, M. W., Baldwin, C. T., Lunetta, K., Farrer, L. A.; Multi-Institutional Research on Alzheimer Genetic Epidemiology (MIRAGE) Study, Furney, S. J., Lovestone, S., Simmons, A., Mecocci, P., Vellas, B., Tsolaki, M., Kloszewska, I., Soininen, H.; AddNeuroMed Consortium, McDonald, B. C., Farlow, M. R., Ghetti, B.; Indiana Memory and Aging Study, Huentelman, M. J., Saykin, A. J. & Alzheimer's Disease Neuroimaging Initiative (ADNI) (2013). Whole-exome sequencing and imaging genetics identify functional variants for rate of change in hippocampal volume in mild cognitive impairment. *Molecular Psychiatry* **18**, 781–787.

No authors listed (2014). Share alike. *Nature* **507**, 140.

Online Mendelian Inheritance in Man, OMIM. Available at www.omim.org (Accessed July 2015).

Onsongo, G., Erdmann, J., Spears, M. D., Chilton, J., Beckman, K. B., Hauge, A., Yohe, S., Schomaker, M., Bower, M., Silverstein, K. A. & Thyagarajan, B. (2014). Implementation of Cloud based next generation sequencing data analysis in a clinical laboratory. *BMC Research Notes* **7**, 314.

Pareto, V. & Busino, G. (1964). *Œuvres complètes*. Droz, Genève.

Parla, J., Iossifov, I., Grabill, I., Spector, M., Kramer, M. & McCombie, W. R. (2011). A comparative analysis of exome capture. *Genome Biology* **12**, R97.

Pollack, A. (2014). F.D.A. Acts on Lab Tests Developed In-House. In *The New York Times*. 31 July 2014.

Proverbio, M. C., Mangano, E., Gessi, A., Bordoni, R., Spinelli, R., Asselta, R., Valin, P. S., Di Candia, S., Zamproni, I., Diceglie, C., Mora, S., Caruso-Nicoletti, M., Salvatoni, A., De Bellis, G. & Battaglia, C. (2013). Whole genome SNP genotyping and exome sequencing reveal novel genetic variants and putative causative genes in congenital hyperinsulinism. *PLoS ONE* **8**, e68740.

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

Rizzo, J. M. & Buck, M. J. (2012). Key principles and clinical applications of "next-generation" DNA sequencing. Cancer Prevention Research *(*Philadelphia, Pa.*)* **5**, 887–900.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome Biology* **12**, 125.

Schatz, M. C., Langmead, B. & Salzberg, S. L. (2010). Cloud computing and the DNA data race. *Nature Biotechnology* **28**, 691–693.

Scott, A. F., Mohr, D. W., Kasch, L. M., Barton, J. A., Pittiglio, R., Ingersoll, R., Craig, B., Marosy, B. A., Doheny, K. F., Bromley, W. C., Roderick, T. H., Chassaing, N., Calvas, P., Prabhu, S. S. & Jabs, E. W. (2014). Identification of an HMGB3 frameshift mutation in a family with an X-linked colobomatous microphthalmia syndrome using whole-genome and X-exome sequencing. *JAMA Ophthalmology* **132**, 1215–1220.

Shaheen, R., Rahbeeni, Z., Alhashem, A., Faqeih, E., Zhao, Q., Xiong, Y., Almoisheer, A., Al-Qattan, S. M., Almadani, H. A., Al-Onazi, N., Al-Baqawi, B. S., Saleh, M. A., Alkuraya, F. S. (2014). Neu-Laxova syndrome, an inborn error of serine metabolism, is caused by mutations in PHGDH. *The American Journal of Human Genetics* **94**, 898–904.

Shahmirzadi, L., Chao, E. C., Palmaer, E., Parra, M. C., Tang, S. & Gonzalez, K. D. F. (2014). Patient decisions for disclosure of secondary findings among the first 200 individuals undergoing clinical diagnostic exome sequencing. *Genetics in Medicine* **16**, 395–399.

Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W. & Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *BioMed Research International* **2014**, 16.

Sharma, V. P., Fenwick, A. L., Brockop, M. S., McGowan, S. J., Goos, J. A., Hoogeboom, A. J., Brady, A. F., Jeelani, N. O., Lynch, S. A., Mulliken, J. B., Murray, D. J., Phipps, J. M., Sweeney, E., Tomkins, S. E., Wilson, L. C., Bennett, S., Cornall, R. J., Broxholme, J., Kanapin, A.; 500 Whole-Genome Sequences (WGS500) Consortium, Johnson, D., Wall, S. A., van der Spek, P. J., Mathijssen, I. M., Maxson, R. E., Twigg, S. R. & Wilkie, A. O. (2013). Mutations in TCF12, encoding a basic helix-loop-helix partner of TWIST1, are a frequent cause of coronal craniosynostosis. *Nature Genetics* **45**, 304–307.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135–1145.

Siva, N. (2008). 1000 Genomes Project. *Nature Biotechnology* **26**, 256–256.

Stoffels, M., Szperl, A., Simon, A., Netea, M. G., Plantinga, T. S., van Deuren, M., Kamphuis, S., Lachmann, H. J., Cuppen, E., Kloosterman, W. P., Frenkel, J., van Diemen, C. C., Wijmenga, C., van Gijn, M. & van der Meer, J. W. (2014). MEFV mutations affecting pyrin amino acid 577 cause autosomal dominant autoinflammatory disease. *Annals of the Rheumatic Diseases* **73**, 455–461.

Stogmann, E., Reinthaler, E., El Tawil, S., El Etribi, M. A., Hemeda, M., El Nahhas, N., Gaber, A. M., Fouad, A., Edris, S., Benet-Pages, A., Eck, S. H., Pataraia, E., Mei, D., Brice, A., Lesage, S., Guerrini, R., Zimprich, F., Strom, T. M. & Zimprich, A. (2013). Autosomal recessive cortical myoclonic tremor and epilepsy: association with a mutation in the potassium channel associated gene CNTN2. *Brain*, **136**, 1155–1160.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M. & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLoS ONE* **6**, e21101.

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., Topol, E. J., Weiner, M. P., Harismendy, O., Olson, J., Link, D. R. & Frazer, K. A. (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology* **27**, 1025–1031.

Torella, A., Fanin, M., Mutarelli, M., Peterle, E., Del Vecchio Blanco, F., Rispoli, R., Savarese, M., Garofalo, A., Piluso, G., Morandi, L., Ricci, G., Siciliano, G., Angelini, C. & Nigro, V. (2013). Next-generation sequencing identifies transportin 3 as the causative gene for LGMD1F. *PLoS ONE* **8**, e63536.

United States Food and Drug Administration (2014). FDA takes steps to help ensure the reliability of certain diagnostic tests. In *Reinforces Agency's Commitment to Fostering Personalized Medicine*. MD, USA: United States Food and Drug Administration.

Valencia, C. A., Pervaiz, M. A., Husami, A., Qian, Y. & Zhang, K. (2013). A Review of DNA Enrichment Technologies. In *Next Generation Sequencing Technologies in Medical Genetics*, p. 25–32. New York, USA: Springer.

Wang, K., Kim, C., Bradfield, J., Guo, Y., Toskala, E., Otieno, F. G., Hou, C., Thomas, K., Cardinale, C., Lyon, G. J., Golhar, R. & Hakonarson, H. (2013). Whole-genome DNA/RNA sequencing identifies truncating mutations in RBCK1 in a novel Mendelian disease with neuromuscular and cardiac involvement. *Genome Medicine* **5**, 67.

Wei, W., He, H., Chen, C., Zhao, Y., Jiang, H., Liu, W., Du, Z. F., Chen, X. L., Shi, S. Y. & Zhang, X. N. (2014). Whole exome sequencing implicates PTCH1 and COL17A1 genes in ossification of the posterior longitudinal ligament of the cervical spine in Chinese patients. *Genetics and Molecular Research* **13**, 1794–1804.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C. & Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792.

Worthey, E. A., Mayer, A. N., Syverson, G. D., Helbling, D., Bonacci, B. B., Decker, B., Serpe, J. M., Dasu, T., Tschannen, M. R., Veith, R. L., Basehore, M. J., Broeckel, U., Tomita-Mitchell, A., Arca, M. J., Casper, J. T., Margolis, D. A., Bick, D. P., Hessner, M. J., Routes, J. M., Verbsky, J. W., Jacob, H. J. & Dimmock, D. P. (2011). Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in Medicine* **13**, 255–262.

Yaneva–Deliverska, M. (2011). Rare diseases and genetic discrimination. *Journal of IMAB – Annual Proceeding Scientific Papers* **17**, 116–119.