Opportunities and challenges in NGS for diagnosis of rare paediatric diseases

Chiara Bacchelli[1] and Hywel Williams[2]

1 – Chiara Bacchelli, PhD

Senior Lecurer in Personalised Medicine, NIHR Investigator, Head of Experimental & Personalised Medicine Section, Genetics and Genomic Medicine Programme, UCL Institute of Child Health, 30 Guilford Street, London WC1N 1EH. T: +44 (0)207 905 2108, F: +44 (0)207 404 6191, E: c.bacchelli@ucl.ac.uk

2 – Dr Hywel Williams, GOSgene, Genetics and Genomic Medicine Programme, UCL Institute of Child Health, 30 Guilford Street, London WC1N 1EH. T: +44 (0)207 905 2625, Fax: +44 (0)207 404 6191, E: hywel.williams@ucl.ac.uk. Corresponding author.

Abstract:

Introduction: Rare paediatric diseases are clinically severe with high rates of mortality and morbidity. This review outlines how Next Generation Sequencing (NGS)can be used to greatly advance the identification of the underlying genetic causes.

Areas Covered: This review is a blend of evidence obtained from literature searches from PubMed and rare disease related websites, laboratory experience and the author's opinions. The review covers the current state of the field and identifies where the challenges lie and how they are being overcome, using up-to-date references.

Expert Opinion: The field of NGS is still relatively new but it has already transformed the field of rare disease research. Technological advances in instrumentation and computational hardware and software have resulted in the identification of many causative genes but as sequencing moves into population-scale initiatives standardisation and data sharing is going to be paramount to ensure we derive the maximum benefit for patients.

Keywords:

Rare disease

Next Generation Sequencing

Whole genome sequencing

Whole exome sequencing

Phenotype

Paediatric

Mutation

Bioinformatics

Personalised medicine

Network analysis

1. Introduction:

1.1. *The challenges of diagnosing rare paediatric diseases:*

Rare diseases are typically severe, genetic in origin and in the vast majority of cases affect children. In Europe a rare disease (RD), as defined by the European Commission, is one that occurs in less than 1 in 2,000 individuals. The total number of RDs that fall within this definition is difficult to determine exactly but best estimates put number at over 7000 [1]. One reason for this uncertainty is that current nosology uses umbrella terms such as intellectual disability to aggregate samples that probably have multiple and distinct aetiological routes to causation and thus leads to an underestimate of the true diversity of diseases. Another reason is that due to their individual rarity not all rare diseases have yet been documented and those that have are typically derived from developed regions with good healthcare provision suggesting that many more diseases are currently unreported due to a lack of healthcare in undeveloped regions of the world. What's more, given that the majority of human genes show high conservation across the mammalian lineage, it is probable that specific mutations in a large proportion of all human genes will result in a phenotypic outcome [2]. This will probably also be the case for a subset of the less-conserved genes that mark inter-species differences. Current observations, with mutations in over 2,900 human genes already characterised [3-5] and an increasing number of novel disease genes being published weekly, lends credence to this view.

Collectively RDs pose a significant psychological burden [6] to patients and their families and a significant financial burden to the healthcare system. The psychological burden comes from the

obvious emotions associated with having or caring for someone with a RD but is amplified by the observation that as many as 75% of RDs have their onset at birth or in childhood and that 15% of affected children will not live beyond 5 years of age. The financial burden to the healthcare system comes from the fact that in aggregate RDs affect around 1 in 17 of the population, which in the UK which equates to 3.5 million people (www.raredisease.org.uk). Moreover, the typically early age of onset for RDs results in a substantial lifelong burden of disability which has been estimated to cost in the region of €2,000,000 per patient [7]. This high cost results, in part, from our lack of understanding of the pathobiological aetiology of most RDs, leading to a dearth of specific therapies. This leads to a situation whereby clinical treatments are focused on ameliorating the key symptoms rather than targeting the underlying pathobiology.

To help reduce the psychological and financial burden associated with RDs we need to be able to make a rapid and accurate diagnoses and, when a diagnosis is made, we need to have targeted therapies available to treat or alleviate the symptoms. Currently however, the causative gene has been identified in less than half of all known RDs, meaning there are still over 3,500 diseases where no gene has yet been discovered. Furthermore, for many RDs where known genes have been identified, only a subset of patients will have mutations within these genes meaning that, even for the 3,500 RDs where genes have been identified, there are still a substantial number of patients diagnosed with a RD but without a genetic diagnosis.

There is a specific group of RD patients for whom obtaining a diagnosis is even more difficult. These are the uncharacterised disease patients, who have an array of phenotypic features that make them unique and unlike any known documented RD. For these patients getting a diagnosis usually only comes through the identification of other phenotypically identical patients and the identification of a novel causative gene [8]. As these patients tend to be the ‘rarest of the rare’ finding matching patients is not always possible and so many of them have to live without a definitive diagnosis and with an unknown prognosis. For RD patients the term ‘diagnostic odyssey’ is well known and

describes the time taken to go from first clinical evaluation to a diagnosis. In a UK based study it has been shown that the average rare disease patient will consult with 5 doctors, receive 3 misdiagnoses and wait 4 years until they get the correct diagnosis (www.raredisease.org.uk). For families, receiving the correct diagnosis is important because without it patients find it difficult to access the correct specialists and receive appropriate therapies, which can result in un-necessary patient suffering. Furthermore, delayed treatment can result in continued pathology that will increase their long-term morbidity and suffering. The psychological burden can also be increased in patients without a diagnosis as there is the constant uncertainty as to the future prognosis. To help RD patients and their families there are a large number of patient groups and charities (e.g. Rare Disease UK) that give patients and their families access to help, advice and support which many find invaluable. These networks are able to lobby on behalf of RD patients and their families and have led to many changes in government and in health authorities to raise the awareness of RDs and to improve patient treatment. An example of this comes from the work of the charity Rare Diseases UK that has resulted in a unified Implementation Strategy for Rare Diseases in the UK (www.raredisease.org.uk), this is a document listing specific targets the devolved health authorities across the UK should adhere to so as to ensure continuity of treatment. For patients without a diagnosis it can be difficult to obtain similar support as many of these charities are dedicated towards a single or group of related RDs, however there are some notable exceptions (UNIQUE:www.rarechromo.co.uk and SWAN: undiagnosed.org.uk) that bring together patients without a diagnosis so that they can share experiences and help to support each other.

One of the key challenges of diagnosing patients with RDs comes from the way in which a diagnosis is typically made. To make a diagnosis the reviewing clinician will assess the patient's phenotype which includes the patients' physical characteristics plus any diagnostic test results, then using their wealth of clinical knowledge make a subjective judgement. In many instances this is sufficient to provide a definitive diagnosis, especially where the phenotypic or diagnostic test data is striking and defines an obvious RD. However, the sheer number of RDs combined with the fact that many are so

rare that a clinician will be unlikely to have ever personally seen another patient means it is not always so simple.

It is also worth remembering that each individual is unique and their phenotype results from the complex interplay between their genetic background and environmental influences. This can give rise to the phenomenon referred to as phenotypic heterogeneity, that is, where a group of patients with mutations in the same RD gene display a spectrum of phenotypes some of which do not overlap and others that may overlap with other RDs [9] or locus heterogeneity where patients with the same RD phenotype have mutations in different genes [10] and incomplete penetrance, where individuals with known disease causing alleles do not display a disease phenotype [11]. A further complication in diagnosis comes with the diagnoses of very young children and babies as their phenotype can commonly look different to that of older children and adults, resulting in their RD as not being recognised.

It has been estimated that in a standard clinical setting the success in diagnosing RD patients can range from approximately 50% [1,12] to as low as 34% (11% in children) [13]. Indeed, within our own group GOSgene, that focuses on the diagnosis of children with rare undiagnosed diseases we too have achieved a diagnostic rate of ~45% (unpublished findings). Therefore, more than half of all patients with a RD fail to obtain a diagnosis which represents a huge unmet need and is a major challenge for clinical genetics to overcome.

1.2. *How to overcome these challenges:*

Firstly, it should be recognised that the achievements to date in the field of RD research have been outstanding. That we are able to diagnose almost half of all patients with a RD is the result of many decades worth of endeavour by scientists and clinicians, many of whom have focused their careers towards elucidating the cause and treatments for specific RDs. Through the identification of specific mutations that link genetic variation to altered protein function/dynamics and observable phenotypes we have been able to progress our understanding of a human development and

physiology in both health and disease. Moreover, in many ways the identification of mutations in RD research has been the driving force in understanding the link between gene function and phenotype diversity in humans. By finding out what has gone wrong we have been able to elucidate biological systems previously unknown.

1.3. *Technological innervation:*

A major turning point in RD research came in 2009 with the publication of the first manuscript that described the use of NGS for the identification of RD genes [14]. The NGS technique employed in this study was that of Whole Exome Sequencing (WES). This technique involves the use of a hybridisation step that uses baits that are targeted specifically to the coding 1-2% of the human genome (the exons). By targeting just the exonic region of the genome researchers are able to sequence only these regions at a higher coverage and/or more samples to be combined and sequenced together. This is a far more cost effective method compared to that performing Whole Genome Sequencing (WGS) which was still prohibitively expensive at that time. Also, there was a sound scientific rationale for focusing on coding regions as studies to date, including those following up on linkage signlas, have shown that ~80% of all RD mutations identified have resided within this coding sequence [15]. A technical description of the different NGS techniques and their evolution is beyond the scope of this manuscript but readers are encouraged to read the recent review by Goodwin and colleagues that presents a comprehensive overview [16]. The study by Ng and colleagues demonstrated the proof-of-concept of using WES as a cost effective technique for gene identification through the elucidation of candidate genes for Freeman-Sheldon syndrome (FSS) in 4 unrelated affected individuals and in the process also demonstrated the utility of having data from unaffected control individuals to help filter and prioritise candidate variants [14].

1.4. *Sharing is good:*

Since the publication of the first WES study to identify a RD mutation the number of studies and RD genes identified has increased to the point where every week a number of novel RD gene

publications are added to PubMed. Driving this surge in RD gene identification has been the falling price of NGS and the concomitant realisation of groups around the world that NGS is becoming a cost effective diagnostic tool for RD gene identification [1,2,12,13,17-30]. The widespread use of NGS has led to the formation of large programs, such as Finding of Rare Disease Genes(FORGE) [1] or Deciphering Developmental Disorders (DDD) [31], that aim to collaborate together and share information on RD patients, thus overcoming one of the major problems associated with analysing NGS data for RD patients, that of having a list of potential causative variants but requiring a second, phenotypically similar family for confirmation. To aid in the identification of phenotypically similar patients there is the need to adopt standard phenotypic nomenclature, for example: Human Phenotype Ontology (HPO) [32] that can be used throughout the RD community to ensure that all poignant phenotypic features are recorded in an electronic format. Such data can then be used to search for patients with matching phenotypic overlaps.

Undiagnosed patients still form a substantial fraction of those that have undergone NGS and strategies need to be put in place to allow for better collaborations to be made. Indeed, there has been a shift by many groups away from keeping their data locked in silos as they realise that by sharing their data in a collaborative manner everyone can benefit, especially the patients who may finally get to receive a diagnosis. This is not trivial task however as the ethics required to share such information is commonly not available, especially for historical samples, and it also raises pertinent questions regarding the protection of patient data and that of their families.

The largest such collaborative endeavour, known as the Matchmaker Exchange (MME) [33], acts as a meta-repository linking various other RD gene identification groups together through an application programming interface (API). It offers a glimpse into the future of RD research where the overarching aim is to identify genes for patient benefit through collaboration not isolation. Such endeavours do however raise pertinent questions that revolve around patient confidentiality and ethics. To this end MME is working closely with groups such as the Global Alliance for Genomics and

Health (GA4GH) and International Rare Diseases Research Consortium (IRDiRC) to ensure consensus standards and best practices are maintained.

To help users adhere to local consent agreements for data sharing MME has set up two levels of matchmaking; Level 1: Requires no additional consent and includes the use of broad phenotype and HPO terms with candidate gene names (+/- variant type) whereas, Level 2: Requires consent and included the use of unique or sensitive identifying phenotypic descriptions and sequence level variant information [34]. By formulating this structure now MME aims to set in place the foundations of a harmonised system that can be used on a global basis to aid in the sharing of RD data going into the future.

1.5. *Setting the standards now will benefit future studies:*

1.5.1. *Sequencing hardware:*

To share RD variant data globally requires the use of standardised systems to measure, analyse and annotate the huge amounts of genomic data being produced. None of this work would be possible if it were not for the Human Genome Project (HGP) [35,36]. Firstly, the HGP gives us a standard genomic reference that can be referred to and secondly, the technological advances made during its completion are what led to the development of NGS techniques. It is worth noting that since the concept of NGS was realised there have been a number of commercial companies employing vastly different chemistries and technologies to provide high quality, cost effective sequencers. Although this innovation is still proceeding, albeit in more niche applications, it is obvious that the field of NGS has become dominated by the use of sequencing machines built by one company, that of Illumina (www.illumina.com).

1.5.2. *Sequencing software:*

As the use and diversity of NGS hardware increased so did the number of computational programs designed to process the data. Unlike the commercial dominance of the NGS hardware market, the

software market was a free enterprise with the majority of programs being written by research teams from the academic sector and, as such the majority were open source and free to use. A downside of open source software is that it can be cumbersome to use, inefficient and highly prone to crashing due to compatibility bugs. Nonetheless, some of the larger sequencing centres from around the world have worked to build file formats and analysis tools that are robust and free to use for academic users and in turn these have been adopted by a significant percentage of NGS users. As such, the analysis of NGS data is typically performed using the program BWA-MEM [37] to align the sequence reads to the human reference genome and the software package Genome Analysis Tool Kit (GATK) [38,39] or SAMtools [40] for variant calling. A typical NGS experiment, utilising the above hardware and software, will result in a large amount of data for each individual, stored in three standard file formats; a .fastq file containing the raw sequence data from the sequencing machine, a .bam file that represents the .fastq data aligned to the human reference genome and a .vcf file that contains details of just those nucleotides that differ between the individual sequenced and the reference genome. There are a number of alternative software programs available to map and call variants that could be included here and the interested reader is advised to refer to relevant literature such as the comparison by Hwang, S and colleagues to gain more detailed insights [41].

A positive outlook on the consequence arising from the dominance in NGS hardware and software is that it has given laboratories from across the globe the ability to generate and analyse their data using a common set of tools that are on the whole compatible. This has allowed large amounts of data from tens and thousands of samples to be combined while keeping technical artefacts to a minimum(for example ExAC [42]).

1.5.3. *Filtering sequence data:*

Once the sequence data for an individual has been generated and processed the next step is to interpret the findings in an attempt to identify the causative gene. Although there are a myriad of ways to annotate and analyse the variant data a general workflow is highlighted in figure 1. Briefly,

the data is first filtered to include only high quality variants (high read depth and call/mapping quality), that are rare in the general population (<0.5% in ExAC [42] for example) and which are likely to have a functional impact on a candidate gene (missense, nonsense, frameshift and canonical splice-site). Familial data can then be used to determine if the candidate variants fit with the suspected inheritance pattern for that patient (figure 2) and for this step is hugely beneficial to have sequence data available from both parents (trio), especially for the detection of *de novo* variants. It is at this stage at which phenotypic and clinical test data can be incorporated to try to distinguish between those variants that look potentially damaging but which are in fact benign and those variants that are causative. As discussed above, this type of gene identification workflow can, at best, identify causative variants in around half of patients but as the NGS technology improves and the analysis and interpretation of genome data improves this percentage is set to increase.

1.5.4. *The inexorable rise in the use of Whole Genome Sequencing (WGS):*

The overwhelming majority of NGS studies in the field of RD gene identification to date have utilised WES but as the cost of NGS continues to fall it will reach a point where the price differential between performing WES and WGS becomes negligible. This time is fast approaching and we are soon to reach the so-called '$1000 Genome', which is predicted to be the time when WGS will be more cost effective than WES [43]. It is worth at this point considering the advantages and disadvantages of WGS versus WES, the main points are highlighted in table 1. This demonstrates that although for most metrics WGS is the better option, the main metric that restricts researchers is the cost, and as WES is currently cheaper it is, therefore, still the default option for most.

1.6. *Factors affecting RD gene identification:*

It is worthwhile considering where the potential bottlenecks are likely to be in the gene identification pathway for RDs. In the first instance, in the pre-NGS era, the technology was the major barrier, as witnessed by the time and cost of using Sanger sequencing to complete the HGP. Next, NGS technologies were developed that overcame the limitations of Sanger based sequencing

but they were initially very expensive, restricting sample numbers. As a consequence targeted sequencing approaches such as WES were developed as a way to minimise costs and increase usage. The popularity and success of WES continued to drive innovation in the field of NGS technology, allowing more data to be produced in shorter time-frames and at lower costs which has led to the situation now where the price of NGS has reached the point where WGS is sure to become the more favourable technique soon.

1.6.1. *Data storage:*

Although NGS has many advantages it can potentially result in a data analysis and storage bottlenecks due to the increased number of samples that can now be analysed. This is no minor problem, a typical WES or WGS run will generate 25GB or 75GBof data respectively for a single sample and, when this is scaled up to include multiple samples it soon becomes too large to be stored on a standard desktop computer or hard drive. This then requires the use of large computational servers for the long-term, safe storage of the data which is not trivial and has to be factored in to the overall cost of an NGS project. More problematic than the storage issue however, is the processing of the raw (.fastq) data to produce the .bam and .vcf files as the software for this analysis requires UNIX based operating systems and the use of high performance computing servers or clusters. For most large academic institutions the high performance computing infrastructures are in place to allow the simultaneous analysis of a reasonably large set of samples (~30) to be performed in a timely manner but again the costs associated with the running and maintenance of such infrastructure has to be factored in to the overall cost of the experiment. Some centres on the other hand operate on a scale far greater, for example the sequencers at the Broad Institute of MIT and Harvard generate 20TB of data a day 365 days a year. In an attempt to efficiently handle such huge amounts of data they are trying innovative ideas and have teamed up with Google Genomics to develop a cloud based computing solution. The increasing use of high powered computing hardware and the personnel to run it is minimising the bottleneck surrounding NGS data processing and

storage but again it is pushing that bottleneck further downstream so that the bottleneck now becomes the interpretation of the data.

1.6.2. *Data Interpretation:*

An example of a typical interpretation pipeline for NGS variant data is shown in figure 1.This can be used for both WES and WGS data and many of the steps can be automated to streamline the process. The difficulty typically arises when a list of candidate genes remains that fulfil the selection criteria as these have to be manually assessed on a gene-by-gene basis to determine their role in the RD being studied. The time taken to complete this process can take anywhere from an hour to days or weeks even, depending on the amount of follow-up work required. The cost associated with this level of interpretation can far out way the cost of generating the data, and for diagnostic laboratories can result in a substantial proportion of their budget, especially when the time taken to produce a diagnostic report is factored in.

As shown above, the success rate for this type of analysis ranges but at best, in at least half of all samples no genetic variant will be located. For WES based studies there is little more that can be done with such data. In comparison, if WGS was performed and no coding variant could be found then WGS has the major advantage of capturing all the noncoding variation as well as giving far more robust Copy Number Variant (CNV) data and information relating to Structural Variation (SV) such as inversions and translocations. This is a vast amount of additional data, for example in our laboratory, just in terms of variants; WES generates ~50,000 variants whereas WGS generates ~3,500,000 variants and it is the large number of noncoding variants which poses huge problems in terms of annotation and understanding.

We know from the large number of causative RD variants found to date that they are mostly coding, rare and loss of function (LOF) and we understand in great detail how changes to the coding sequence can change protein structure or alter binding characteristics and lead to the phenotypes of the affected patients. However, when the causative variant is not coding we are suddenly faced with

whole new set of problems. We can filter our variants by quality (for example, mapping quality) as the same statistics are relevant but when it comes to filtering by frequency we have a far smaller dataset that can be used, for example the 1000 Genomes Project [44] data. How we then annotate the >3,000,000 variants that are noncoding is something that projects such as ENCODE(Encyclopaedia of DNA Elements) [45] have been striving to answer. The goal of the ENCODE project, as stated on their website (https://www.encodeproject.org/) is "to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active." By using this data we can start to annotate our noncoding variants to see if they reside within a region of the genome predicted to alter the function of a gene and as we learn more about how to perform these prioritisations we will become better at identifying candidate genes that we can then apply our phenotypic knowledge to in order to identify noncoding causative variants for RDs. In practical terms, the ability to use the ENCODE data in a meaningful way in studies aimed at RD gene identification is still very much in its infancy and still restricted to research based projects instead of diagnostic laboratories but it does hold the potential to identify causative regulatory variants we know exist [46] and it offers much hope that in the future we will be able to identify the cause for the >50% of RD patients who are refractory to current gene identification studies.

1.7. *Future directions:*

1.7.1. *Population-scale projects:*

Although the utilisation of WGS in a systematic way for population-scale RD research projects has not yet been performed, there are examples of projects that are currently underway that will change this and which will revolutionise the use of genomics in the healthcare environment. Presently, the most advanced such project is taking place in the UK, this is the 100,000 Genomes Project (100KGP) which will perform WGS on 100,000 samples. Of relevance here, half of the genomes sequenced will be related to RD gene identification and will be achieved through sequencing ~17,000 RD patients

and their parents in the form of trios, the remaining half of the project will be based on cancer and infection related studies. The project is funded by the UK government and is being run through the National Health Service (NHS), thus allowing the electronic Health Records (eHRs) of patients and their whole-genome data to be combined on a large scale for the first time. One of the aims of the project is to build the NHS infrastructure that is able to utilise the advances in NGS technology for patient benefit and to stimulate innovation in the genomic healthcare system by working with commercial partners and academic groups. A high security computational datacentre has been built to house all the eHR and genomic data and a number of disease specific clinical/academic partnerships have been developed to analyse the genomic data of patients that do not receive a definitive clinical diagnosis following analysis by clinical scientist. By combining the resources of the NHS with that of academics and commercial partners the 100KGP is likely to have a profound positive impact on the study of RDs and will no doubt lead to unexpected advances in our understanding of human genomics.

The 100KGP project has spearheaded a major global shift in the use of WGS for clinical RD research and already President Barack Obama has set up the personalised medicine initiative in the USA that plans to perform WGS on 1 million people while Qatar have also launched a population-scale genomic sequencing project. Sequencing projects on such an unprecedented scale will allow us to better understand variation across populations, drive innovation and technological advances and; allow the development of standard methods to analyse, interpret and share data. This latter point is essential in RD research as there are currently wide discrepancies across diagnostic laboratories in the interpretation and classification of variants [47] and it will help overcome false positive results being published on novel variants that are based on allele frequencies derived from inappropriate populations or insufficient sample sizes [48].

1.7.2. *Advanced analytical techniques:*

With access to large homogeneously phenotyped, sequenced and analysed datasets of RD patients and their families, we will be able to apply higher analytical methods such as those employing bio-statistical algorithms or machine learning techniques to identify novel variants and better elucidate the underlying pathobiology. An example of just such an application was demonstrated by Akawi, N and colleagues who utilised the power of the DDD (Deciphering Developmental Disorders) dataset by developing a set of statistical tests to identify four novel recessive phenotypes [49]. Other techniques such as pathway and network analyses hold great promise for the identification of novel biologically relevant interactomes that can highlight novel biological pathways which could in turn be targeted by therapeutics [50,51]. In this context, WGS data will allow us to build networks combining data not only from coding LOF variants but allow us incorporate data from noncoding regulatory variants that influence both known and novel genes. A consequence of such networks is that when applied to large well phenotyped datasets they have the potential to delineate phenotypic subgroups of patients that may be more responsive to a particular drug or therapeutic intervention.

1.8. *The elephant in the room:*

The one thing that cannot be ignored when discussing the challenges of RD research is the fact that usually the singly most difficult part of proving a novel candidate variant is causative is demonstrating the functionality of the identified variant. There are a suite of prediction programs and databases that can be interrogated to provide evidence of candidacy but the final test requires the ability to show in a model system that the identified variant causes a functional effect that mimics the phenotype seen in the patient. Since the completion of the HGP, the cost and time taken to perform NGS has fallen to the point now where a whole human genome can be sequenced and analysed within 26 hours [52] at a cost of a few thousand dollars, whereas the concomitant time and cost of functionally characterising a candidate variant is still measured in terms of months or even

years and the cost can easily be in excess of tens of thousands of dollars. To overcome this bottleneck is going to take huge investment and coordination by laboratories on national and international scales because as projects such as 100KGP come online there are likely to be hundreds of novel variants that will need functional validation every month. This is the biggest bottleneck in translating the results of an NGS experiment into a definitive answer for many patients and although for many, just receiving a genetic diagnosis is something which is a huge relief, it does not lead to an improvement in how they are clinically managed or how their symptoms are treated. There have recently been some major advances in the field of functional biology such as CRISPR/cas9 gene editing technology and the development of induced pluripotent stem cells (iPSc) reprogramming techniques that will undoubtedly help expedite the functional characterisation of genetic variants but the time scales involved are still large.

The reason why it is so important to overcome the bottleneck related to the functional characterisation of genes is that it provides categorical evidence for the causality of the identified variant and gives an insight into the underlying pathobiology. Armed with this knowledge it is then possible to look at therapies that can be used to treat the patients, this may involve relieving the symptoms or ideally, preventing them from occurring in the first place. This can involve strategies that use for example; drugs, dietary supplements or gene therapy techniques. The effectiveness of these interventions however varies widely and therefore many groups are now working on ways to utilise the genomic knowledge to get a better understanding of the optimal therapeutic strategy on an individual basis, what is termed personalised medicine.

It should nevertheless be remembered that even with advances in therapeutics, some patients are going to be refractory to treatment especially if the critical period occurs during development.

1.9. *Personalised Medicine:*

Personalised medicine has been defined by the US Food and Drug Administration (FDA) as the tailoring of medical treatment to the individual characteristics, needs and preferences of a patient

during all stages of care, including prevention, diagnosis, treatment and follow-up (http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/PersonalisedMedicine/UCM372421.pdf). The aim is to tailor healthcare to each individual patient or group of patients. The way this is achieved is by stratifying patients according to their genetic makeup into subgroups with for example different responses to drugs, different disease risks or other clinically relevant factors. The stratification results in more cost-effective treatments based on the underlying cause of disease allowing early intervention in the disease process or preventing the disease from occurring at all. Patients' management, surgical interventions and treatments will be informed by variants identified in their genomes. An example of this is represented by some form of immunodeficiency or very-early-onset inflammatory bowel disease (VEO-IBD) [53]. By knowing which type of gene or mutation is responsible for the disease one can inform patient management towards hematopoietic stem cell transplantation, immunomodulation or gene therapy.

Patients affected by a RD often have few or no drugs available to treat their conditions. In recent years there has been a shift in the attention of pharmaceutical companies towards RDs drug development research. This is partly due to the success of the FDA Orphan Drug Designation which creates incentives and quicker approval process for pharmaceutical companies working on development of RD drugs. Equally, the advances in genomic research by NGS technologies have increased the capacity and understanding of diseases. Once the exact causative mutation responsible for a RD has been identified (i.e. resulting for example in a deficiency or an absence of an enzyme), the development of a targeted treatment can be easily developed.

Current data from the FDA shows that about 47% of the novel drugs approved in 2015 (21 of 45) were approved to treat rare or "orphan" diseases. One example is Kanuma (sebelipase alfa) produced by Alexion Pharmaceuticals Inc., which received Orphan Drug Designation by the FDA for the treatment of a rare condition known as lysosomal acid lipase (LAL) deficiency [54]. Patients with LAL deficiency (also known as Wolman disease and cholesteryl ester storage disease [CESD]) have

little or no LAL enzyme activity resulting in a build-up of fats within the cells of various tissues. This can lead to liver and cardiovascular disease and other complications. Wolman disease is a progressive severe disorder that often presents during infancy (around 2 to 4 months of age) and patients rarely survive beyond the first year of life. Kanuma is a recombinant rhLAL protein that functions in place of the missing, partially active or inactive LAL protein in the patients and is administrated intravenously once a week in LAL patients. Other examples are from companies like Novartis and Genzyme and products for RDs like Gleevec for Chronic Myeloid Leukemia and Cerezyme for type 1 Gaucher disease.

Great hopes for RDs treatment has also been put into personalised gene therapy. While genomics studies provide the identification of the genetic defect, gene therapy provides the correction of the defect. A number of successful trials in the last two decades have employed ex-vivo haematopoietic stem cell gene therapy as a therapeutic option to treat specific rare inherited immune deficiencies, including adenosine deaminase deficiency, X-linked severe combined immunodeficiency, chronic granulomatous disease and Wiskott Aldrich syndrome [55]. In more recent years, other RD patients have been involved in gene therapy trials for retinitis pigmentosa, X-linked Chronic Granulomatous Disease, Severe Haemophilia A amongst others. For a complete list of gene therapy trials see http://www.genetherapynet.com/clinicaltrialsgov.html.

2. Conclusions:

In the field of paediatric Rare Disease (RD) gene discovery there is an optimistic feeling that many of the obstacles that prevent a diagnosis being made or therapy being found are beginning to be overcome. It is over a decade since the completion of the Human Genome Project (HGP) [35,36] which had, as one of its primary aims, the goal of improving of our understanding of human disease. This objective has been greatly advanced as a direct result of the utilisation of Next Generation Sequencing (NGS) technologies. NGS has transformed our ability to perform unbiased, genome-wide

sequencing on large numbers of RD patients and their families and has led to the identification of the genetic cause in approximately half of the 7000 RDs currently classified. As this technology advances and the price continues to fall, we will soon be in a position to perform Whole Genome Sequencing (WGS) as a standard research tool which will help realise the goals of groups such as the International Rare Diseases Research Consortium (IRDiRC) of identifying the genetic cause of all RDs by 2020.

Projects such as the 100,000 Genomes Project (100KGP) will provide a huge boost to the goals of IRDiRC and lay the foundations of introducing genomics into the routine healthcare environment as well as producing an unprecedented data source that will drive the implementation of higher order analyses, such as network analyses to mine for novel disease genes and phenotypic modifiers. The goal of such analyses is to lead to a better understanding of the underlying pathobiology which can be modelled in cell or animal systems with the aim that armed with this knowledge we can develop specific drugs or therapies that alleviate disease symptoms or even prevent them occurring at all. By understanding how our genetic background can modify a disease phenotype we can begin to treat patients on an individualised nature which is the goal of personalised medicine.

3. Expert commentary:

The field of Rare Disease research and diagnosis is being transformed through the use of Next Generation Sequencing (NGS). In the research realm the falling cost of NGS, especially Whole Genome Sequencing (WGS), is leading to a position where it will be the method of choice for most groups within the next few years. This is has resulted in an ever increasing number of RD genes being identified which is also driving the use of molecular techniques to functionally characterise these genes. Concomitantly, in the diagnostic realm a point will be reached soon where WGS becomes the preferred option as gene panels are growing to a point where it will be more efficient to perform WGS and apply virtual gene panels. To this end, the 100,000 Genomes Project in the UK is set to

revolutionise the implementation of genomics into healthcare and will set the standard practises that will be the blueprint for countries around the world. For RD patients this will lead to improved rates of diagnosis and with it the hope that it will drive innovation in the fields of drug discovery and therapeutic interventions to allow the realisation of personalised medicines.

4. Five-year view:

The field of Next Generation Sequencing (NGS) is still relatively new, it has only been with us for around 10 years and although its use is increasing it is still mainly restricted to specialised research groups. In the field of Rare Disease (RD) research the use of NGS has led to the identification of a huge number of novel genes and that number is increasing all the time. The increased use of NGS is being driven by technological advances which result in ever falling costs and have reached the point where we are almost at the '$1000 Genome' threshold. These reduced costs have allowed projects such as the 100,000 Genomes Project (100KGP) in the UK to be financially viable and through these large scale endeavours the rate of novel gene discovery is set to increase drastically over the next 5 years. Importantly, the 100KGP will set a benchmark that can be followed by other groups because if we want this data to be accessible to the whole community then standards have to be put in place that make the data generated from different groups compatible. One of the main outputs of the 100KGP is likely to be the integration of Whole Genome Sequencing (WGS) data into a standard healthcare system so that WGS moves away from being the reserve of research groups and becomes a standard diagnostic clinical tool. The accumulation of the huge amounts of genome wide data on hundreds of thousands humans from a range of populations will drive the innovation of higher analytical techniques to mine the data and lead to a far greater understanding of human genetic diversity.

5. Key issues:

- There are at least 7000 rare diseases with the causative gene identified in about half.
- Rarity of many diseases makes finding families to replicate candidate mutations difficult.

- Many patients have unique phenotypes that make them intractable to standard diagnoses which is a huge barrier to them receiving appropriate treatment.
- Next Generation Sequencing technologies have had a profound influence in the rare disease field and the rate of novel genes being discovered is increasing massively.
- Being able to accurately phenotype patients in a standardised manner, using electronic records is a necessity for the future.
- The decreasing cost of NGS, especially whole genome sequencing, is allowing projects such as the 100,000 Genomes Project in the UK to be conducted which will revolutionise healthcare in the future.
- The huge amounts of data being produced result in an unprecedented computational burden and require solutions to allow the safe storage, analysis and interpretation of NGS data.
- Higher order analytical techniques such as those employing bio-statistical algorithms, machine learning techniques and network analyses will be required to mine the vast amounts of data to find causative and modifying genes.

Acknowledgements:

The authors are supported by the National Institute for Health Research Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London.

References:

1. Beaulieu CL, Majewski J, Schwartzentruber J et al. FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. Am J Hum Genet, 94(6), 809-817 (2014). *Provides details of large national project to identify rare disease genes on a large scale and provides details of diagnosis rates.
2. Chong JX, Buckingham KJ, Jhangiani SN et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet, 97(2), 199-215 (2015).

3. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic acids research, 43(Database issue), D789-798 (2015).

4. Kaiser J. Human genetics. Affordable 'exomes' fill gaps in a catalog of rare diseases. Science, 330(6006), 903 (2010).

5. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annual review of medicine, 61, 437-455 (2010).

6. Anderson M, Elliott EJ, Zurynski YA. Australian families living with rare disease: experiences of diagnosis, health services use and needs for psychosocial support. Orphanet journal of rare diseases, 8, 22 (2013).

7. Angelis A, Tordrup D, Kanavos P. Socio-economic burden of rare diseases: A systematic review of cost of illness evidence. Health policy (Amsterdam, Netherlands), 119(7), 964-979 (2015).

8. Thomas AC, Williams H, Seto-Salvia N et al. Mutations in SNX14 cause a distinctive autosomal-recessive cerebellar ataxia and intellectual disability syndrome. Am J Hum Genet, 95(5), 611-621 (2014).

9. Seri M, Cusano R, Gangarossa S et al. Mutations in MYH9 result in the May-Hegglin anomaly, and Fechtner and Sebastian syndromes. The May-Hegglin/Fechtner Syndrome Consortium. Nat Genet, 26(1), 103-105 (2000).

10. Hartong DT, Berson EL, Dryja TP. Retinitis pigmentosa. Lancet, 368(9549), 1795-1809 (2006).

11. Chen R, Shi L, Hakenberg J et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nature biotechnology, 34(5), 531-538 (2016).

12. Shashi V, McConkie-Rosell A, Rosell B et al. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. Genetics in medicine : official journal of the American College of Medical Genetics, 16(2), 176-182 (2014).

13. Gahl WA, Mulvihill JJ, Toro C et al. The NIH Undiagnosed Diseases Program and Network: Applications to modern medicine. Molecular genetics and metabolism, (2016).

*Provides a thorough overview of large national study and provides detailed metrics of rates of diagnosis.

14. Ng SB, Turner EH, Robertson PD et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 461(7261), 272-276 (2009).

15. Cooper DN, Chen JM, Ball EV et al. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Human mutation, 31(6), 631-655 (2010).

16. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet, 17(6), 333-351 (2016).

17. Bloss CS, Zeeland AA, Topol SE et al. A genome sequencing program for novel undiagnosed diseases. Genetics in medicine : official journal of the American College of Medical Genetics, 17(12), 995-1001 (2015).

18. Boycott KM, Dyment DA, Sawyer SL, Vanstone MR, Beaulieu CL. Identification of genes for childhood heritable diseases. Annual review of medicine, 65, 19-31 (2014).

19. Boycott KM, Vanstone MR, Bulman DE, Mackenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet, 14(10), 681-691 (2013).

20. Delaney SK, Hultner ML, Jacob HJ et al. Toward clinical genomics in everyday medicine: perspectives and recommendations. Expert review of molecular diagnostics, 1-12 (2016).

21. Gahl WA, Wise AL, Ashley EA. The Undiagnosed Diseases Network of the National Institutes of Health: A National Extension. Jama, 314(17), 1797-1798 (2015).

22. Grozeva D, Carss K, Spasic-Boskovic O et al. Targeted Next-Generation Sequencing Analysis of 1,000 Individuals with Intellectual Disability. Human mutation, 36(12), 1197-1204 (2015).

23. Prada CE, Gonzaga-Jauregui C, Tannenbaum R et al. Clinical utility of whole-exome sequencing in rare diseases: Galactosialidosis. European journal of medical genetics, 57(7), 339-344 (2014).

24. Taylor JC, Martin HC, Lise S et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet, 47(7), 717-726 (2015).

*Study describing the use of whole genome sequencing for the detection of rare disease genes and its utility for diagnosis.

25. Tennessen JA, Bigham AW, O'Connor TD et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science, 337(6090), 64-69 (2012).

26. Volk A, Conboy E, Wical B, Patterson M, Kirmani S. Whole-Exome Sequencing in the Clinic: Lessons from Six Consecutive Cases from the Clinician's Perspective. Molecular syndromology, 6(1), 23-31 (2015).

27. Walter K, Min JL, Huang J et al. The UK10K project identifies rare variants in health and disease. Nature, 526(7571), 82-90 (2015).

28. Williams HJ, Hurst JR, Ocaka L et al. The use of whole-exome sequencing to disentangle complex phenotypes. Eur J Hum Genet, (2015).

29. Wright CF, Fitzgerald TW, Jones WD et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet, 385(9975), 1305-1314 (2015).

30. Yang Y, Muzny DM, Reid JG et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med, 369(16), 1502-1511 (2013).

31. Study DDD. Large-scale discovery of novel genetic causes of developmental disorders. Nature, 519(7542), 223-228 (2015).

32. Kohler S, Doelken SC, Mungall CJ et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research, 42(Database issue), D966-974 (2014).

33. Buske OJ, Schiettecatte F, Hutton B et al. The Matchmaker Exchange API: Automating Patient Matching Through the Exchange of Structured Phenotypic and Genotypic Profiles. Human mutation, (2015).

34. Philippakis AA, Azzariti DR, Beltran S et al. The Matchmaker Exchange: a platform for rare disease gene discovery. Human mutation, 36(10), 915-921 (2015).

35. Lander ES, Linton LM, Birren B et al. Initial sequencing and analysis of the human genome. Nature, 409(6822), 860-921 (2001).

36. Venter JC, Adams MD, Myers EW et al. The sequence of the human genome. Science, 291(5507), 1304-1351 (2001).

37. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26(5), 589-595 (2010).

38. DePristo MA, Banks E, Poplin R et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet, 43(5), 491-498 (2011).

39. McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20(9), 1297-1303 (2010).

40. Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079 (2009).

41. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific reports, 5, 17875 (2015).

42. (ExAc) EAC. Analysis of protein-coding genetic variation in 60,706 humans. BioRxiv, (2016).

43. Hayden EC. Technology: The $1,000 genome. Nature, 507(7492), 294-295 (2014).

44. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature, 467(7319), 1061-1073 (2010).

45. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57-74 (2012).

**There is a huge amount of associated literature describing the characterisation of the functional elements contained across the genome. This is a must for the analysis and interpretation of Whole Genome datasets

46. Makrythanasis P, Antonarakis SE. Pathogenic variants in non-protein-coding sequences. Clinical genetics, 84(5), 422-428 (2013).

47. Amendola LM, Jarvik GP, Leo MC et al. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet, (2016).

48. Andreasen C, Refsgaard L, Nielsen JB et al. Mutations in genes encoding cardiac ion channels previously associated with sudden infant death syndrome (SIDS) are present with high frequency in new exome data. The Canadian journal of cardiology, 29(9), 1104-1109 (2013).

49. Akawi N, McRae J, Ansari M et al. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nature Genetics, 47(1546-1718 (Electronic)) (2015).

*Study highlighting the use of bio-statistical models to identify novel rare disease genes from large cohorts of patients.

50. Boldt K, van Reeuwijk J, Lu Q et al. An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. Nature communications, 7, 11491 (2016).

51. Menche J, Sharma A, Kitsak M et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science, 347(6224), 1257601 (2015).

*Use of network analysis to delineate discreet disease related interactomes that have the potential to identify novel pathobiological mechanisms in patients.

52. Miller NA, Farrow EG, Gibson M et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. Genome medicine, 7(1), 100 (2015).

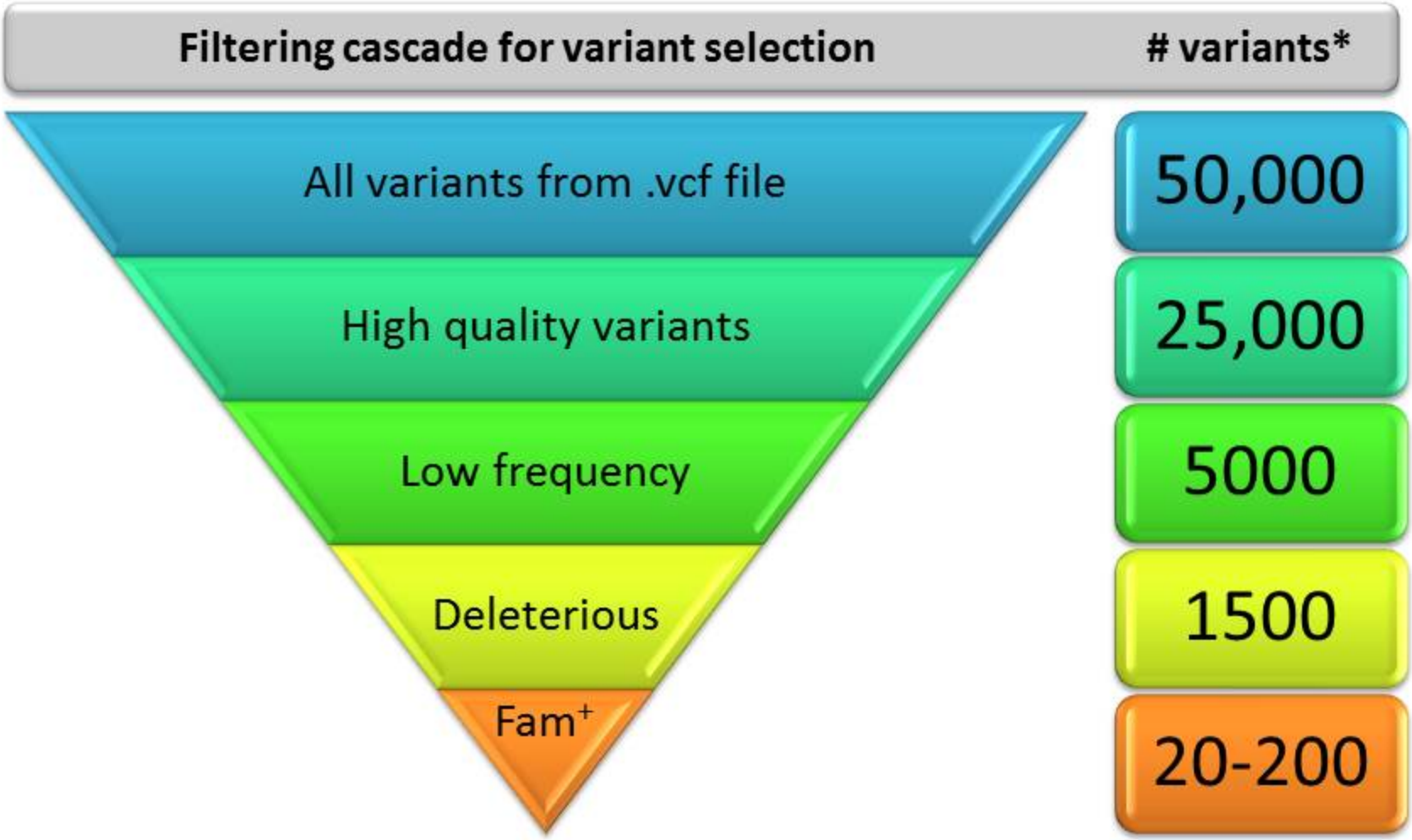
**Study demonstrating the application of Whole Genome Sequencing to an acute diagnostic setting for improved clinical management.

53. Kammermeier J, Drury S, James CT et al. Targeted gene panel sequencing in children with very early onset inflammatory bowel disease--evaluation and prospective analysis. J Med Genet, 51(11), 748-755 (2014).

54. Shirley M. Sebelipase alfa: first global approval. Drugs, 75(16), 1935-1940 (2015).

55. Booth C, Gaspar HB, Thrasher AJ. Treating Immunodeficiency through HSC Gene Therapy. Trends in molecular medicine, 22(4), 317-327 (2016).

Figure 1: Overview of the filters used to select relevant pathogenic variants from a Whole Exome Sequencing experiment.



* - numbers correspond to whole exome sequencing study

+ - Filter based on suspected familial mode of inheritance (see Figure 2)

Figure 2: Suspected familial inheritance patterns can be used to filter variants

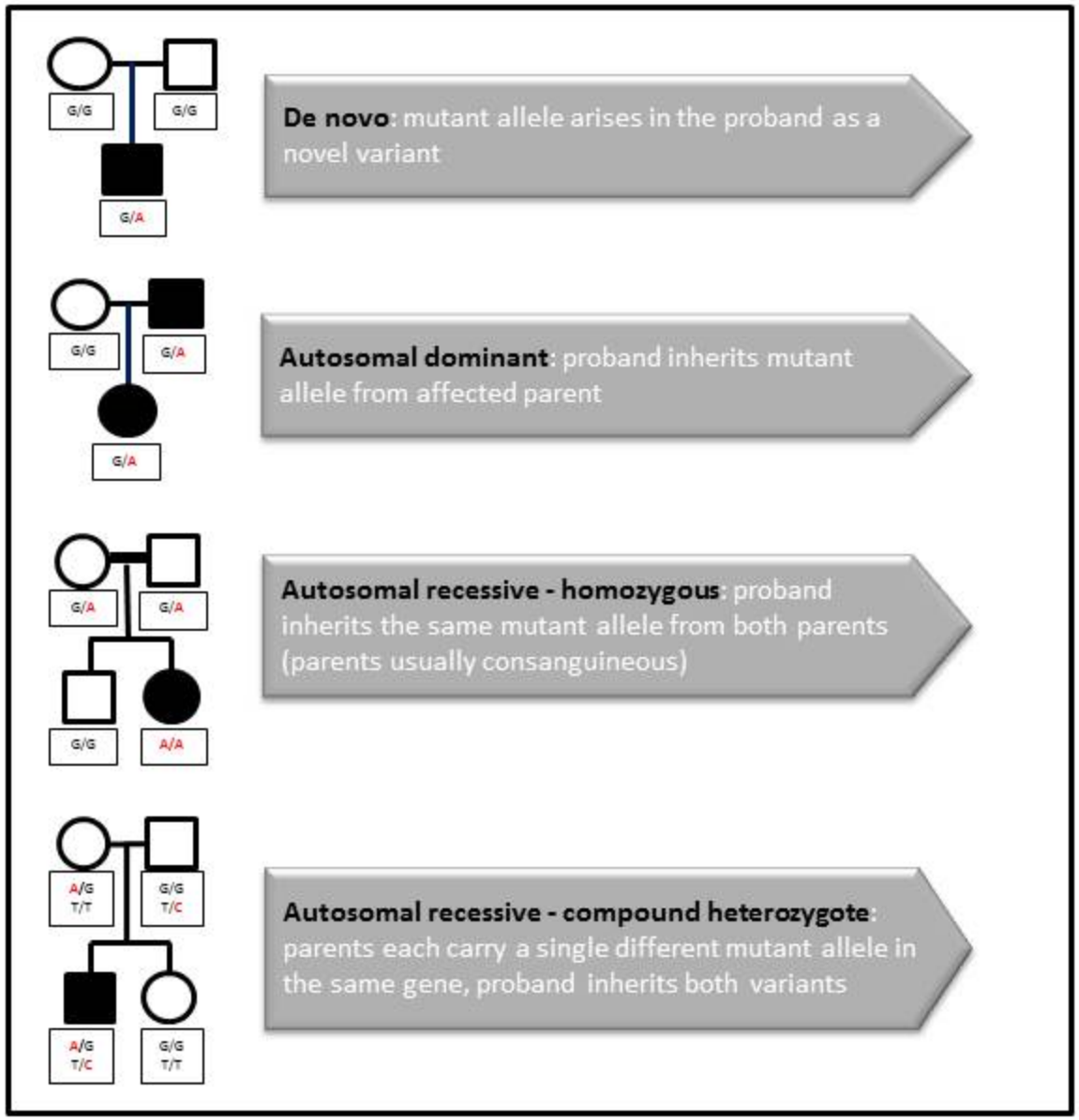


Table 1: Comparison of the advantages/disadvantages between whole exome sequencing and whole genome sequencing.

	Whole Exome Sequencing (WES)	Whole Genome Sequencing (WGS)	Optimal method
Sample preparation	Requires PCR and hybridisation	PCR free and no hybridisation	WGS
Cost(£)[1]	300-500	1,200-1,500	WES
Coverage	Uneven and poor for high GC regions	Uniform	WGS
Variants called	Coding and flanking intronic	Whole genome	WGS
Copy number variation	Poor quality	High quality	WGS
Structural variation	Not possible	High quality	WGS
Data storage[2]	25GB	75GB	WES

[1] equivalent to 50X for WES and 30X for WGS; [2] typical data size for .fastq, .bam and .vcf files from a single sample