Exome sequencing: a transformative technology

Andrew B Singleton

Lancet Neurol 2011; 10: 942–46

Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
(A B Singleton PhD)

Correspondence to:
Dr Andrew B Singleton, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20837, USA
singleta@mail.nih.gov

Summary

Background **Much basic research into disease mechanisms has made use of genetic findings to model and understand aetiology. Broad success has been achieved in finding disease-linked mutations with traditional positional cloning approaches; however, because of the requirements of this method, these successes have been limited by the availability of large, well characterised families. Because of these and other restrictions the genetic basis of many diseases, and diseases in many families, remains unknown.**

Recent developments **Exome sequencing uses DNA-enrichment methods and massively parallel nucleotide sequencing to comprehensively identify and type protein-coding variants throughout the genome. Coupled with growing databases that contain known variants, exome sequencing makes identification of genetic mutations and risk factors possible in families and samples that were deemed insufficiently informative for previous genetic studies. Not only does exome sequencing enable identification of mutations in families that were undetectable with linkage and positional cloning methods, but compared with these methods, it is also much quicker and cheaper. Use of exome sequencing has so far been successful in many rare diseases.**

Where next? **Exome sequencing is being adopted widely and we can expect an abundance of mutation discovery, similar to the deluge of genome-wide-association findings reported over the past 5 years; it is expected to enable the discovery of not only rare causal variants, but also protein-coding risk variants. This method will have application in both the research and clinical arenas and sets the scene for the use of whole-genome sequencing.**

Introduction

Progress towards a full resolution of the genetic basis of disease is being substantially aided by a fast-moving technological development, exome sequencing. This method promises to speed up discovery of the genetic causes of disease in both the research and the clinical setting.

The method of exome sequencing has been covered elsewhere.[1] Although several methods exist, they all use a similar principle: reducing a genomic DNA sample to one that is enriched for the protein-coding regions of the genome (exons), followed by very high-throughput sequencing of the exon-enriched sample (figure). In short, this is a method for rapidly identifying protein-coding mutations, including missense, non-sense, splice site, and small deletion or insertion mutations.

Exome sequencing uses second-generation sequencing, which generates sequence data from hundreds of millions of short DNA fragments in parallel. The sequencing of input libraries is, to all intents and purposes, random; each of the fragments that happens to be in the DNA library and is applied to the sequencer has about an equal chance of being sequenced. Thus, directive sequencing of specific DNA fragments is determined by creating a DNA library solely consisting of, or enriched for, the DNA regions of interest. In the context of exome sequencing, this target selection is done with one of several enrichment products, each of which is intended to produce a DNA sample in which the content consists of the protein-coding and regulatory regions of the genome. This method has some limitations: first, coverage of regions of interest is not complete. In early experiments about 8% of the regions of interest were not captured by the enrichment strategy,[2] and although the coverage has improved, it will probably never reach 100%; second, at present, this method is not useful for identifying repeat mutations (such as triplet repeats in spinocerebellar ataxia); and third, copy number variants are difficult to detect with exome sequencing. However, in view of the distribution of variants, exome-sequencing remains an efficient way to identify most mutations altering protein sequence in any single DNA sample. Although exome sequencing is quite new to the market, it has been rapidly adopted by the research community.

In this Rapid Review, I aim to review the likely use and success of exome sequencing in research and clinical settings.

Mendelian disorders

The primary successes for exome sequencing have been in finding mutations that cause rare, familial forms of disease. The strength of this approach lies in the comprehensive discovery of protein-coding variants throughout the genome. This means that DNA samples collected from small families and isolated affected individuals, which could not be used for mutation identification through traditional linkage and positional cloning, can now be used to discover mutations that cause disease.

Exome sequencing of a DNA sample from a single individual will typically reveal about 25 000 variants; the challenge then lies not in finding variants, but in identifying the particular mutation responsible for disease. A common subsequent step, when looking for extremely rare causal mutations, involves filtering against variants

Tecan Ex. 2020, Page 1 of 5
IDT, Inc. v. Tecan Genomics - IPR2025-00015

known to exist in the general population, because these variants are unlikely to cause disease. Genome-wide association studies (GWAS) have shown that generation of large, accessible, reference datasets is an economical way to produce control data,[3] which is a model that is being emulated in exome sequencing. The most broadly used reference panel for exome sequencing is the data derived from the 1000 Genomes Project, an initiative that will ultimately involve the sequencing the genomes of about 2500 individuals.[4] The aim of this project is to provide a comprehensive resource on human genetic variation across a large series of individuals. Filtering the 25 000 or so variants identified in a single sample against the 1000 Genomes data and the Single Nucleotide Polymorphism database (dbSNP) variants typically reduces the list of candidate mutations down to about 700. Further filtering steps often include removing variants that do not appreciably alter aminoacid sequence, do not fit with the mode of disease inheritance, do not segregate with disease (in families), and that are predicted computationally to have minimal consequences in terms of protein structure. This formulaic approach to reducing the size of a candidate list of variants to define the causal variants has proven successful in many diseases,[2,5–22] as in the identification of mutations that cause Parkinson's disease,[23,24] for example.

As more sequence data accumulate in the public domain, the list of candidate variants after filtering will decrease, but with this the likelihood of erroneously filtering out a disease-causing mutation increases. Such an error is perhaps less likely in very young-onset severe diseases, but probably quite likely in later-onset diseases, diseases with benign phenotypes, or those with mutations that have a low penetrance. Mutation carriers who are too young to express disease, who do not show symptoms because of decreased penetrance, or who have a phenotype that might not be noticed could be sequenced and have their data deposited in reference panels. Furthermore, some (perhaps all) reference data are associated with very limited phenotypic data, so patients with, or who will get, common diseases are likely to be included but not identified.

Beyond Mendelian disorders

There is increasing interest in taking advantage of the power of exome sequencing in diseases that do not exhibit a simple Mendelian mode of transmission. Exome sequencing has particular potential in diseases caused by non-inherited or de-novo mutations. Typically, this potential can be achieved with complementary approaches: the first, and most simple, is to sequence a group of cases suspected to have disease-causing de-novo or non-inherited mutations and look for a gene that is commonly mutated; the second, and perhaps more compelling approach, is to sequence parent-affected child trios to establish the variant(s) that are present only in the affected children. De-novo and mosaic mutations were particularly resistant to previous efforts aimed at gene localisation and identification. In this context, exome sequencing will not only help identify novel genetic causes of disease, but could also show that de-novo mutations are a more common cause of disease than previously expected. Some success in identifying these mutations has been seen with exome sequencing for varied diseases, including several malformation and mental retardation disorders, and most recently Proteus syndrome.[11,17,25–27]

In diseases with a complex genetic component, exome sequencing is used to identify variability in protein-coding regions that alters risk for, rather than causes, disease. Many examples of this type of coding risk variant exist, although generally those that have been discovered so far are common variants and thus detectable with GWAS, or were identified through positional and functional candidate-based screening strategies. Classic examples include *APOE* in Alzheimer's disease, *CFH* in age-related macular degeneration, and *GBA*/*LRRK2* in Parkinson's disease.[28–32] The power of exome sequencing is that it will allow the identification of common protein-coding risk alleles and, with suitable statistical power and study design, the identification of rare risk alleles. This latter approach centres on a complicated analytical design that involves assessment of the collective burden of multiple risk alleles at a locus. Although this approach remains largely unproven, initial work in epilepsy suggests that this method can reveal complex relations

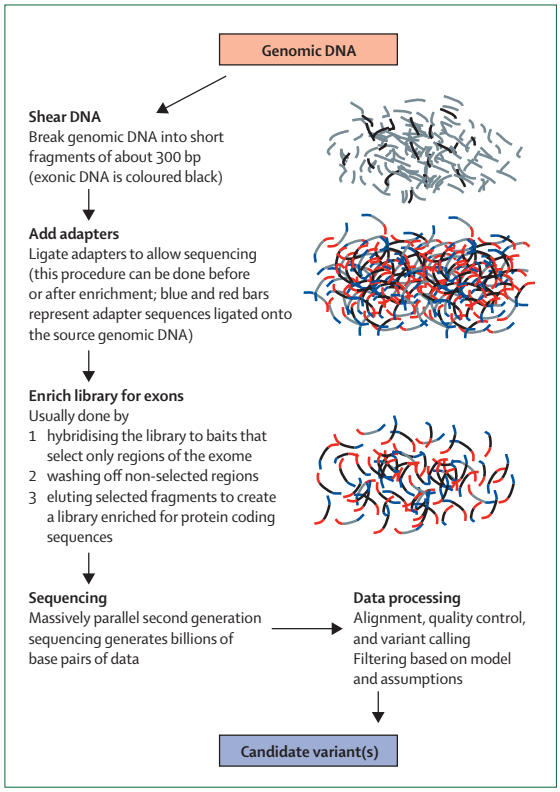

Figure: **Current method of exome sequencing**

For more on the **1000 Genomes Project** see http://www.1000genomes.org/

For the **Single Nucleotide Polymorphism database** see http://www.ncbi.nlm.nih.gov/SNP/

Tecan Ex. 2020, Page 2 of 5
IDT, Inc. v. Tecan Genomics - IPR2025-00015

between established disease-associated variants, and enable risk-prediction modelling.[33]

Challenges and opportunities

Exome sequencing, and also genome sequencing, will probably have a substantial effect in the clinical setting, beyond the identification of genes that were previously not known to contain disease-causing mutations. In the context of genetic testing, instead of screening an inherently limited panel of genes for a specific set of diseases, why not just sequence the whole exome? This method opens the door to many possibilities—eg, rapid genetic diagnosis and screening, when known disease-causing mutations are detected, and an inevitable expansion of the phenotypic range of diseases associated with specific mutations or mutations in specific genes.

Exome sequencing has already proven its worth in rapid genetic diagnosis and screening; for many neurological diseases, which have a long list of candidate genes and loci, identification of mutations by exome sequencing is often cheaper and quicker than it is by traditional sequencing approaches. For many diseases with a high degree of genetic heterogeneity, screenings are often designed to catch only the most common mutations or alterations in the most frequently mutated genes. Montenegro and colleagues[34] showed the power of exome sequencing in exactly this situation, with the analysis of a family with Charcot-Marie-Tooth disease. Rather than screening the 35 genes known to contain mutations causing this disease, the investigators used exome sequencing in two affected family members; with these data they were able to identify *GJB1* mutations as a cause of Charcot-Marie-Tooth disease in this family. This case is also illustrative: a *GJB1* mutation would have been ruled out for screening because of a reported male-to-male inheritance, incompatible with a mutation of this gene, which lies on chromosome X. The use of exome sequencing probably saved time and money in reaching a genetic diagnosis. The increasing adoption of exome sequencing in the research setting means that these data are becoming easier to process and analyse, in addition to becoming much cheaper to generate; many laboratories are now able to generate an exome for less than US$1500 a sample. When compared with the costs and time required for conventional screening, such a comprehensive approach represents value for money. Despite still being a new technology, exome sequencing in genetically heterogeneous neurological diseases has already been used to identify *TECR* mutations in non-syndromic mental retardation,[8] *WDR62* mutations in severe brain development malformations,[35] *TGM6* mutations in ataxia,[21] and a *WRN* mutation in atypical Werner's syndrome.[36]

As mentioned above, exome sequencing is likely to help with identification of mutations in genes previously linked to disease but associated with a phenotype distinct from the one being tested, as shown with the identification of *VCP* mutations as a cause of amyotrophic lateral sclerosis (ALS).[13] Johnson and colleagues[13] had used exome sequencing to identify the genetic lesion responsible for an autosomal dominant form of ALS in a large pedigree from Italy. Surprisingly, *VCP* mutations, including the same aminoacid change identified in this ALS kindred, had previously been linked to Paget's disease, inclusion body myopathy, and frontotemporal dementia. Therefore, this finding broadened the clinical and pathological phenotype of *VCP* mutations to include ALS and TDP-43 inclusions. Notably, further work by this group showed that *VCP* mutations substantially contribute to the cause of familial ALS, being responsible for about 2% of cases in this group.[13] As this work shows, broadening the phenotype associated with mutations has the potential to provide information on the aetiological basis of disorders by uniting what is known about the biological underpinnings of apparently unrelated disorders into a single model.

One might also predict that as exome data accumulate, our understanding of the role of mutations in disease will improve. Exome data will not only help to identify pathogenic variants, including those of previously unknown or equivocal pathogenicity, but it will also help to determine penetrance, expressivity, and prevalence of mutations in specific populations. Related to the notion of penetrance, in the sequencing of normal individuals we might expect to find variants that were previously thought to cause fully penetrant disease, and these data will call into question the pathogenicity of some published variants. Recent research describing the high phenocopy rate in families with apparently monogenic Parkinson's disease has shown that even in families with an apparently simple monogenic cause, confounding genetic, environmental, or stochastic factors probably affect presentation and penetrance.[37] Introduction of exome sequencing will raise many questions and often reveal an apparently confusing relation between genetic variants and disease, but, with time, accumulating data will help to bring resolution to many unanswered questions about clinical-genetic relations.

One clinical challenge that is specific to comprehensive genetic approaches comes with the inevitable discovery of mutations unrelated to the disorder in question. These secondary findings will be common in exome sequencing, and indeed one of the first publications on exome sequencing, which sought to identify mutations for Miller syndrome, also described the identification of mutations causing ciliary dyskinesia within the same family.[18,38] The questions then arise as to what a clinician should do when they identify a mutation of proven or even potential clinical relevance and how the disclosure of these mutations should be handled. These are not necessarily simple issues, especially when considering the confounding factors and problems such as disclosing carrier status, non-paternity, reduced penetrance, absence of viable treatment options, and interpretation

Tecan Ex. 2020, Page 3 of 5
IDT, Inc. v. Tecan Genomics - IPR2025-00015

Search strategy and selection criteria

I searched PubMed for articles published between Jan 1, 2000, and June 15, 2011, using the search term "exome". Results were manually assessed to identify articles that were relevant to the topic of this Rapid Review. Only articles published in English were selected.

of risk factors. As has been argued previously,[38] these issues will require thoughtfully constructed research and continuing education of both health-care providers and the general public.

Where next?

Some hold the view that there is little purpose in finding new genetic causes of disease because we have made little progress in understanding the consequences of the mutations that we already know about. However, examples of gains in knowledge from the identification of genes suggest that this argument is ill-founded; for instance, the role of amyloid processing in Alzheimer's disease was indicated by the discovery of *APP*, *PS1*, and *PS2* mutations.[39] This view also does not take into account non-research aspects of mutation identification, especially our responsibility to find a cause of disease for patients, most of whom desire a diagnosis even if it does not come with a treatment. Moreover, this logic centres on the idea that aetiology-based therapies are not likely to be developed or are not important, or both, because such therapies would need to be grounded in a complete understanding of the disease process. Such an understanding requires the foundation of knowledge provided by defined disease-initiating events. This highlights a philosophical difference between those who believe in something of a systems-based approach, whereby the more information (genes) we have to put in a model, the more reliable and complete is our understanding of the network implicated, and those who believe in a more reductionist approach to science, whereby the consequences of a single perturbation to a system should be studied one by one. However, even a truly reductionist approach would allow the separate study of perturbations in several genes and a subsequent search for overlap in the network of downstream effects. In either case, the more genes that we can link to a disease entity, the more complete our understanding will become of the affected systems and the probable consequences of therapeutic interventions directed at those systems.

A second argument against applying exome sequencing seems to revolve around the notion that a better genetic technique will be developed soon. Many genomic technologies are indeed transient in nature, and exome sequencing is probably no exception. The next logical step, taking us to whole-genome sequencing, offers the advantages of comprehensive genomic coverage, easier identification of structural genomic mutations than with conventional sequencing, and no intervening library enrichment steps. Factors inhibiting this transition are the high price and large analytical burden (about 100 times the data of an exome). Critics have suggested that assessment of non-coding sequences is associated with too much uncertainty and at too high a price to ever make whole-genome sequencing viable; however, per base-pair costs for genomic technologies continue to drop precipitously and scientists have never stayed away from large datasets for much time. Clearly, as both the cost and analytical limitations decrease, whole-genome sequencing will become the dominant technology. Indeed, this method has already been successful in the identification of genetic perturbations underlying disease, including the identification of a novel cause of Charcot-Marie-Tooth disease.[40–43] In view of this success, it might be reasonable to ask why we should not skip exome sequencing and wait for whole-genome sequencing to be more widely available. My view on this is simple: we cannot afford to wait. We are in a race to cure diseases. We know exome sequencing works, and it is affordable. The earlier we make genetic findings, the earlier we can begin to integrate these discoveries into an understanding of the disease process. This surely places us on the road towards aetiology-based therapies sooner and with a more complete set of analytical methods.

Conflicts of interest

I declare that I have no conflicts of interest.

Acknowledgments

This work was supported by the Intramural Research Program of the National Institute on Aging, US National Institutes of Health, Department of Health and Human Services; project Z01 AG000958-08.

References

1 Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet* 2010; **19:** R145–51.
2 Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009; **461:** 272–76.
3 Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447:** 661–78.
4 Via M, Gignoux C, Burchard EG. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med* 2010; **2:** 3.
5 Sanna-Cherchi S, Burgess KE, Nees SN, et al. Exome sequencing identified MYO1E and NEIL1 as candidate genes for human autosomal recessive steroid-resistant nephrotic syndrome. *Kidney Int* 2011; **80:** 389–96.
6 Hanson D, Murray PG, O'Sullivan J, et al. Exome sequencing identifies CCDC8 mutations in 3-M syndrome, suggesting that CCDC8 contributes in a pathway with CUL7 and OBSL1 to control human growth. *Am J Hum Genet* 2011; **15:** 148–53.
7 Bolze A, Byun M, McDonald D, et al. Whole-exome-sequencing-based discovery of human FADD deficiency. *Am J Hum Genet* 2010; **87:** 873–81.
8 Caliskan M, Chong JX, Uricchio L, et al. Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13. *Hum Mol Genet* 2011; **20:** 1285–89.
9 Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 2009; **106:** 19096–101.

Tecan Ex. 2020, Page 4 of 5
IDT, Inc. v. Tecan Genomics - IPR2025-00015

10 Gilissen C, Arts HH, Hoischen A, et al. Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *Am J Hum Genet* 2010; **87:** 418–23.

11 Hoischen A, van Bon BW, Gilissen C, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet* 2010; **42:** 483–85.

12 Johnson JO, Gibbs JR, Van Maldergem L, Houlden H, Singleton AB. Exome sequencing in Brown-Vialetto-van Laere syndrome. *Am J Hum Genet* 2010; **87**: 567–69.

13 Johnson JO, Mandrioli J, Benatar M, et al. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* 2010; **68:** 857–64.

14 Krawitz PM, Schweiger MR, Rodelsperger C, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* 2010; **42:** 827–29.

15 Lalonde E, Albrecht S, Ha KC, et al. Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. *Hum Mutat* 2010; **31:** 918–23.

16 Musunuru K, Pirruccello JP, Do R, et al. Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N Engl J Med* 2010; **363:** 2220–27.

17 Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010; **42:** 790–93.

18 Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010; **42:** 30–35.

19 Pierce SB, Walsh T, Chisholm KM, et al. Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault syndrome. *Am J Hum Genet* 2010; **87:** 282–88.

20 Walsh T, Shahin H, Elkan-Miller T, et al. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *Am J Hum Genet* 2010; **87:** 90–94.

21 Wang JL, Yang X, Xia K, et al. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain* 2010; **133:** 3510–18.

22 Shi Y, Li Y, Zhang D, et al. Exome sequencing identifies ZNF644 mutations in high myopia. *PLoS Genet* 2011; **7:** e1002084.

23 Vilarino-Guell C, Wider C, Ross OA, et al. VPS35 mutations in Parkinson disease. *Am J Hum Genet* 2011; **89:** 162–67.

24 Zimprich A, Benet-Pages A, Struhal W, et al. A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. *Am J Hum Genet* 2011; **89:** 168–75.

25 O'Roak BJ, Deriziotis P, Lee C, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* 2011; **43:** 585–89.

26 Vissers LE, de Ligt J, Gilissen C, et al. A de novo paradigm for mental retardation. *Nat Genet* 2010; **42:** 1109–12.

27 Lindhurst MJ, Sapp JC, Teer JK, et al. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med* 2011; **365:** 611–19.

28 Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 1993; **261:** 921–23.

29 Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308:** 385–89.

30 Tan EK. Identification of a common genetic risk variant (LRRK2 Gly2385Arg) in Parkinson's disease. *Ann Acad Med Singapore* 2006; **35:** 840–42.

31 Hernandez DG, Paisan-Ruiz C, McInerney-Leo A, et al. Clinical and positron emission tomography of Parkinson's disease caused by LRRK2. *Ann Neurol* 2005; **57:** 453–56.

32 Sidransky E, Nalls MA, Aasly JO, et al. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N Engl J Med* 2009; **361:** 1651–61.

33 Klassen T, Davis C, Goldman A, et al. Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. *Cell* 2011; **145:** 1036–48.

34 Montenegro G, Powell E, Huang J, et al. Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family. *Ann Neurol* 2011; **69:** 464–70.

35 Bilguvar K, Ozturk AK, Louvi A, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* 2010; **467:** 207–10.

36 Raffan E, Hurst LA, Turki SA, et al. Early diagnosis of Werner's syndrome using exome-wide sequencing in a single, atypical patient. *Front Endocrinol* 2011; **2:** 1–8.

37 Klein C, Chuang R, Marras C, Lang AE. The curious case of phenocopies in families with genetic Parkinson's disease. *Mov Disord* 2011. Published online July 6. DOI:10.1002/mds.23853.

38 Biesecker LG. Exome sequencing makes medical genomics a reality. *Nat Genet* 2010; **42:** 13–14.

39 Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* 2002; **297:** 353–56.

40 Pelak K, Shianna KV, Ge D, et al. The characterization of twenty sequenced human genomes. *PLoS Genet* 2010; **6:** e1001111.

41 Sobreira NL, Cirulli ET, Avramopoulos D, et al. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet* 2010; **6:** e1000991.

42 Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 2010; **362:** 1181–91.

43 Rios J, Stein E, Shendure J, Hobbs HH, Cohen JC. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum Mol Genet* 2010; **19:** 4313–18.

Tecan Ex. 2020, Page 5 of 5
IDT, Inc. v. Tecan Genomics - IPR2025-00015