Published in final edited form as:

Curr Opin Pediatr. 2017 October ; 29(5): 513–519. doi:10.1097/MOP.0000000000000532.

A Primer to Clinical Genome Sequencing

James R Priest, MD

Abstract

Purpose of review—Genome sequencing is now available as a clinical diagnostic test. There is a significant knowledge and translation gap for non-genetic specialists of the processes necessary to generate and interpret clinical genome sequencing. The purpose of this review is to provide a primer on contemporary clinical genome sequencing for non-genetic specialists describing the human genome project, current techniques and applications in genome sequencing, limitations of current technology, and techniques on the horizon.

Recent Findings—As currently implemented, genome sequencing compares short pieces of an individual's genome to a reference sequence developed by the human genome project. Genome sequencing may be used for obtaining timely diagnostic information, cancer pharmacogenomics, or in clinical cases when previous genetic testing has not revealed a clear diagnosis. At present, the implementation of clinical genome sequencing is limited by the availability of clinicians qualified for interpretation, and current techniques in used clinical testing do not detect all types of genetic variation present in a single genome.

Summary—Clinicians considering a genetic diagnosis have wide array of testing choices which now includes genome sequencing. Though not a comprehensive test in its current form, genome sequencing offers more information than gene-panel or exome sequencing and has the potential to replace targeted single-gene or gene-panel testing in many clinical scenarios.

Introduction

With decreases in cost and increasing availability, clinical genome sequencing is no longer a technique of the future but a test available to any informed clinician. While there are many working definitions of precision medicine, they share the fundamental idea that the health and medical services for each individual could be personalized using unique information about that individual such as the genetic information encoded by their genome sequence (1). The aims of this review are to A. Summarize the history of genome sequencing with attention to creation of a reference sequence, B. Outline current methods in genome sequencing in a clinical setting, and C. Introduce future methods for obtaining more accurate and personalized genetic information.

Corresponding Author: James R. Priest MD, Assistant Professor, Pediatric Cardiology, Stanford University School of Medicine, Lucile Packard Children's Hospital, 750 Welch Road Suite 300, Stanford, CA 94305, Phone: (650) 795-9812, jpriest@stanford.edu, Twitter: jpriest@JamesRPriestMD.

Conflicts of interest

I have no conflicts of interest to disclose.

THE HUMAN GENOME PROJECT AND REFERENCE SEQUENCE

To use a metaphor, the human genome is like a book with 3 billion characters represented as A C T G with 23 chapters. Each chapter corresponds to a chromosome, and every person has two copies of each chapter with one copy received from each parent. Among the 3 billion letters only about 1.5% actually codes for the 20,000 genes in the human genome, and variation in some genes can cause genetic diseases, predispositions to disease, or change how a person responds to treatment. How the “reference” book of the human genome was created is the first step understanding clinical genome sequencing.

The 4-letter code of DNA was famously described in the 1960’s and the first chemical process of reading DNA sequences, (which retained the eponym Sanger-sequencing) became automated in the 1980’s (2–4). Concurrently during the 1960’s and 70’s, the field of cytogenetics made the first physical maps of human chromosomes (5), which made possible the placement of single genes to specific cytogenetically defined locations upon individual human chromosomes (6,7). As detailed physical maps of human genetic information converged with the automation of DNA sequencing, the research community conceptualized a collaborative, government-funded effort to obtain the entire sequence the human genome, termed the human genome project, (8).

With primary support from the United States government, the human genome project used a specific and orderly process 1. Each chromosome was divided into fragments (9–11) 2. Fragments were mapped to a cytogenetic location upon each chromosome (12), 3. Each individual fragment was sequenced and assembled (13) 4. Fragments from adjacent sections were linked into long linear stretches of letters (14). Ultimately the sequencing of each piece was performed with first generation automated Sanger-sequencing in lengths of approximately 1000 base pairs. These long linear stretches of DNA stitched together by chromosomes formed the original “reference sequence”.

The human genome contains a large proportion of repetitive sequence and difficult to map regions not present in the first release of the human reference sequence; these problematic regions have required an additional decade of iterative fine mapping and finishing work to obtain more accurate and complete genome sequence.(15) The final product is the “human reference sequence” which underlies the most widely used contemporary framework for clinical and research genome sequencing. A key point is that the reference sequence comes from a group of individuals from different ethnic backgrounds, which introduces complications in the current methodologies for genome sequencing as noted below (16).

CURRENT METHODS IN GENOME SEQUENCING

When looking for a “misspelled” word within genes that can cause disease, newer and cheaper techniques to read the genome are now available. The technical costs of sequencing a human genome are now less than \$1,500 per person. A single chromosome can be as long as 250 million letters, the first step is to cut up the DNA into much small fragments of approximately 1000 to 10000 letters, then to “read” 50 to 250 letters from each end of the fragment. Once all of the “reads” are available, a software tool places these “reads” in order

relative to the human reference sequence. Next, a separate software tool is used to look for places where the reads do not match the reference, and a statistical hypothesis is computed to determine whether the mismatch represents a true genetic difference, or some type of error. Finally we use knowledge about how a gene works to see if the genetic differences observed changes the function of the protein. Genetic counselors play an indispensable role before testing by preparing patients and families to about the process and outcomes of testing, and afterwards in interpreting genetic differences in the appropriate clinical context.

Sequencing and Mapping

Commercial efforts to reduce the cost of DNA sequencing resulted in a handful of related technologies significantly cheaper than the first-generation Sanger-sequencing chemistry employed in the initial drafts of the human genome. These technologies are alternately referred to as next-generation sequencing (NGS), short-read technology, or sequencing-by-synthesis chemistry (17).

From the competing technologies encompassing NGS, the chemistries and equipment marketed by Illumina Corporation (San Diego, California) are the most widely used for clinical and research purposes (18). However a recent report of systematic errors from one widely-used Illumina product illustrate the problems associated with relying on a single commercial provider of sequencing technology (19). The basic principle of the NGS involves optical detection of the base incorporated in the synthesis of a strand complementary to the sample DNA (20). Relative to Sanger-sequencing, NGS has shorter read-lengths (50–250 base pairs compared to 1000 base pairs in Sanger sequencing) and higher error rates as a trade-off for the reduction in cost, a compromise which further complicates the steps of detecting genetic variation (Fig. 1A). A DNA sample obtained from a patient is prepared by cutting the DNA into pieces of small fragments of around 1000 to 10000 base pairs and the NGS technology then reads 50 to 250 base pairs (the limit of this type of technology) from either end of the fragment; thus each read is “paired” with the read from the opposite end of the fragment- and termed “paired-end” reads. To roughly illustrate the scale of a typical genome sequencing run for one person; a sequencing test or “run” would be optimized to cover each nucleotide in the human genome 30 times, with 3 billion base pairs in the human genome such a run would produce approximately 90 billion letters requiring subsequent processing.

Read Mapping

Alignment algorithms to detect the best match between two or more fragments of genetic sequence were developed and applied initially for determining the evolutionary relationship of nucleotide sequences from different species (21). Subsequent adaptations for speed, higher-throughput, and paired-end reads were necessary to handle the large amount of data arising from NGS (22,23). The results from sequence alignment can be visualized as the original sequencing reads ordered by their best match to a position on the reference sequence (Fig. 1B). Given that there are many regions of the human genome that share identical or highly similar sequences errors in alignment are inherent from the process of short-read mapping. Additional alignment errors may also arise from the chemistry errors

within the NGS data, as well as imperfections or differences in the human reference sequence relative to the sequence data that is undergoing alignment.

Variant Calling

Bioinformatic algorithms for detecting genetic differences, a process termed “variant-calling”, rely upon detecting mismatches between a reference sequence and the mapped reads (Fig. 1B). The most widely used software for variant calling is the Genome Analysis Tool Kit (GATK), a comprehensive set of tools authored and maintained by the Broad Institute (Cambridge, Massachusetts) (24). Single nucleotide variants (SNVs) are relatively simple to detect. However with current approaches the likelihood of detecting insertions and deletions of DNA (indels) is inversely proportional to the size of the indel (25). A number of integrated open source software tools exist as alternatives to GATK including FreeBayes and the RTG suite, each which may offer advantages over GATK in areas such as detection of indels, calling variants within familial pedigrees, and processing time (26,27).

Mismatches between the reference sequence and mapped reads can arise from the presence of a true variant, but mismatches frequently come from errors in sequencing chemistry, biases in the reference sequence related to differences in the genetic background between the mapped reads and individuals originally contributing to the human reference sequence, or errors in the alignment process (28). Therefore, filtering of variants is a necessary step after variant calling. Variant calling algorithms employ statistical models of potential errors which incorporate information from the sequencing chemistry, read alignments, and reference sequence in order to assign a likelihood that a detected mismatch represents a true genotype, and these statistical models are used for filtering variants. No matter which analytical framework is applied, all variant filtering balances sensitivity and specificity, removing false variants at the cost of excluding real genetic differences (29).

Annotation and Interpretation

To understand the role a genetic variant might play in human health or disease, the addition of functional information is necessary in a process called “variant annotation”. At a minimum annotation incorporates information about the presence of a variant within a gene and the predicted function of the variant upon the protein product of that gene (30). Many commonly used annotation tools offer a variety of other types of information about genetic variants such as describing the presence of the variant within human population databases, evolutionary conservation of the variant among different species, and aspects of local genomic structure where the variant is located (31). These data may then variously be incorporated into a clinical interpretation (Fig. 1C).

In a clinical setting, interpretation of genome sequencing is dependent on the expertise of the clinicians involved and the medical context under which a test was ordered. Genomic testing for specific diseases falls under a widely accepted medical paradigm for obtaining timely diagnostic information (32,33), cancer genetics and pharmacogenomics (34), and is increasingly being used in clinical cases where previous targeted genetic testing has failed to yield a clear diagnosis (35). In all use-cases, variants are most appropriately interpreted by

clinicians with domain-specific expertise in the genetic basis of the specific disease under consideration.

Clinical genomic testing of asymptomatic healthy individuals for purposes of disease prediction or risk stratification is not yet supported by evidence (36), though a handful of ongoing studies seek to address the cost and efficacy of predictive genomic testing in both children and adults (37,38). In addition to regulatory, ethical, and legal considerations, frameworks from professional bodies have been constructed to guide experienced clinicians in their interpretation and reporting of genomic information (39,40). Data-sharing efforts by researchers and providers of clinical genetic testing are increasing the body of knowledge for specific genes and variants, information that is fundamental to the process of interpreting genetic variation in an informed clinical context (41,42).

Genetic counseling is standard of care before and after testing to ensure that patient and parental consent is truly informed and to provide appropriate context and follow up for the results. Additionally genetic counselors fill an integral role in performing variant interpretation with comprehensive expertise for specific genes and conditions, determining which variants merit reporting, and conducting cascade genetic screening of family members when appropriate. The term “genomic counseling” has been coined to describe these essential roles, which typically fall outside the scope of practice for most physicians (43,44).

CURRENT LIMITATIONS AND FUTURE METHODS

Extending the metaphor of a book to illustrate different kinds of genetic variation, in addition to misspelled words, the genome of each person has large differences with unique sentences and paragraphs placed in different positions throughout any of the chapters. The short reads of current genome sequencing technology can detect the misspelled words (SNPs- (single nucleotide polymorphism) and small indels) reasonably well, but does not currently detect missing or extra sentences and paragraphs (microdeletions or duplications). Current genome sequencing technologies compares one book to another to provide a list of locations where an individual genome has small differences from the reference genome. Future applications have the potential to provide not just a list of differences, but rather a complete and entire book for each person.

Limitations of Current Methods

Compared to gene-panel or exome sequencing, genome sequencing provides a larger quantity of information. However in its current form genome sequencing is not a comprehensive genomic test. Genome sequencing does not allow direct determination of genetic inheritance, referred to as “phasing”; follow up parental testing is often necessary to ascertain diagnoses of conditions caused by recessive inheritance. Clinical genome sequencing does not yet reliably detect copy number variations (CNVs), for example micro deletions or duplications such as 22q11 (DiGeorge) deletion syndrome, which are best assessed in a clinical setting by array comparative genomic hybridization (arrayCGH) or targeted FISH testing (45). Additionally there is a large class of genetic variation, indels between 100 and 100,000 base pairs, which is poorly detected by current genetic and genomic technologies and the impact of such variation on human health and disease is

simply not known at this time. Conversely, genome sequencing may offer a window into classes of variation such as somatic mosaicism which have been relatively difficult to detect (46), and the role of genetic variation in sequence which does not code for genes conferring risk for disease (47).

A key limitation in the application of genome sequencing is the small number of qualified clinical personnel with training and expertise. Clinicians who are ordering genome sequencing may alternately have training in medical genetics, or may be physicians with strong background knowledge in genomics, or training in genetic counseling. However the increased availability of genetic testing in general has outpaced the availability of qualified clinicians of all types (48). In order to meet the projected demands for genomic testing in general, new and integrated educational models in the use and interpretation of genomic testing are needed for clinicians in training and practice (49).

Future Methods

Due to the limitations of reference based short-read genome sequencing, an active area of research seeks to perform a “human genome project” each individual. The newest sequencing technologies are based upon electrochemical reading of bases of DNA as they pass through microscopic pores, and have extended the read length to 2,000 or more base pairs (50). Though long-read sequencing is currently highly prone to errors, the combination of short and long read sequencing allows the assembly of an individual genome at a relatively rapid time and low cost (51). In theory, a personal genome assembly should be better able to detect larger indels and CNVs in addition to the small mutations which are detected by the current paradigm (52).

Conclusions

Genome sequencing has rapidly evolved over the last 10 years from a research tool to a clinically available diagnostic test. Though this review has focused on techniques in clinical genome sequencing, the basic concepts presented here apply to gene-panel tests and to exome-sequencing which utilize the same basic strategies to interrogate individual or multiple genes. Currently clinicians considering a genetic diagnosis have wide array of testing choices which now includes genome sequencing. The rapid changes in our understanding of genetic disease frustrates the use of gene-panel tests; when a new gene or mechanism for disease is reported in the literature, a patient with a negative genetic test requires repeat testing with a second genetic test to include the new information. It is not difficult to imagine a future where a single comprehensive genetic test is ordered, such as genome sequencing, and the patient’s report is updated annually with new findings as our understanding of genetic diseases evolves (53,54).

Acknowledgments

I would like to thank Dr. Dan Bernstein, Kelsey C. Priest MPH, and Benjamin R. Priest MS for useful commentary on this manuscript and Nancy C. Helmsworth MFA for providing original artwork.

Financial support and sponsorship

This work was supported by the Department of Pediatrics and Betty Irene Moore Children's Heart Center at Stanford University School of Medicine, and by NIH Grants K99-HL130523.

Glossary

Indel	An insertion or deletion of DNA. Example: Reference ACTGCGT, Insertion ACTGTCTGCGT, Deletion ACGT
SNP	single nucleotide polymorphism. Example: Reference ACTGCGT, Alternative allele ACTACGT
Reference sequence	A version of the human genome produced by the human genome project from multiple individuals used as a baseline for comparison to sequencing information from single individuals.
CNV	copy number variation, large insertions or deletions of DNA often defined as greater than 100,000 base pairs
Base Pairs	Adenine, Cytosine, Guanine, and Thymine which represent the four chemical letters or base pairs making up human genomic information
Gene Panel Test	The DNA from one or more genes related to a pathogenic condition are selected by a process of chemical hybridization or targeted amplification which is then followed by next-generation sequencing.
Exome Sequencing	All 20,000 genes (approximately 1.5% of the human genome) are selected by a process of chemical hybridization, followed by next-generation sequencing. Regions of the genome which do not encode for genes are not assayed.

References

- 1*. Ashley EA. Towards precision medicine. *Nat Rev Genet.* 2016 Aug 16; 17(9):507–22. A comprehensive review of fundamental precepts of precision medicine and challenges to achieving these goals. [PubMed: 27528417]
2. Watson JD, Crick FH. The structure of DNA. *Cold Spring Harb Symp Quant Biol.* 1953; 18:123–31. [PubMed: 13168976]
3. Sanger F. The early days of DNA sequences. *Nat Med.* 2001 Mar; 7(3):267–8. [PubMed: 11231611]
4. Hood LE, Hunkapiller MW, Smith LM. Automated DNA sequencing and analysis of the human genome. *Genomics.* 1987 Nov; 1(3):201–12. [PubMed: 3328736]
5. Paris Conference (1971): Standardization in human cytogenetics. *Cytogenetics.* 1972; 11(5):317–62. [PubMed: 4647417]
6. Sparkes RS, Carrel RE, Paglia DE. Probable localization of a triosephosphate isomerase gene to the short arm of the number 5 human chromosome. *Nature.* 1969 Oct 25; 224(5217):367–8. [PubMed: 5343881]
7. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science.* 1989 Sep 8; 245(4922):1059–65. [PubMed: 2772657]

8. Mapping and Sequencing the Human Genome. Washington (DC): National Academies Press (US); 1988. National Research Council (US) Committee on Mapping and Sequencing the Human Genome.
9. Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, et al. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci USA*. 1992 Sep 15; 89(18):8794–7. [PubMed: 1528894]
10. Anand R, Villasante A, Tyler-Smith C. Construction of yeast artificial chromosome libraries with large inserts using fractionation by pulsed-field gel electrophoresis. *Nucleic Acids Res*. 1989 May 11; 17(9):3425–33. [PubMed: 2542900]
11. Cantor CR. Orchestrating the Human Genome Project. *Science*. 1990 Apr 6; 248(4951):49–51. [PubMed: 2181666]
12. Burke DT. The role of yeast artificial chromosome clones in generating genome maps. *Current Opinion in Genetics & Development*. 1991 Jun; 1(1):69–74. [PubMed: 1840880]
13. Nickerson DA, Tobe VO, Taylor SL. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*. 1997 Jul 15; 25(14):2745–51. [PubMed: 9207020]
14. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. *Genome Res*. 1998 Mar; 8(3):195–202. [PubMed: 9521923]
15. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. 2004 Oct 21; 431(7011):931–45. [PubMed: 15496913]
- 16*. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*. 2017 May; 27(5):849–64. A technical but clear illustration of how the reference assembly influences the current paradigm of genome sequencing using short-read technology. [PubMed: 28396521]
17. Niedergringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. *Anal Chem*. 2011 Jun 15; 83(12):4327–41. [PubMed: 21612267]
18. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016 May 17; 17(6):333–51. [PubMed: 27184599]
- 19*. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, et al. Index Switching Causes “Spreading-Of-Signal” Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. 2017 An example of how chemical/technical errors in sequencing technology can be propagated and become a source of measurement bias in projects involving short-read sequencing technologies.
20. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008 Nov 6; 456(7218):53–9. [PubMed: 18987734]
21. Vingron M, Waterman MS. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J Mol Biol*. 1994 Jan 7; 235(1):1–12. [PubMed: 8289235]
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009 Jul 15; 25(14):1754–60. [PubMed: 19451168]
23. Reinert K, Langmead B, Weese D, Evers DJ. Alignment of Next-Generation Sequencing Reads. *Annu Rev Genomics Hum Genet*. 2015; 16(1):133–51. [PubMed: 25939052]
24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010 Sep; 20(9):1297–303. [PubMed: 20644199]
25. Narzisi G, Schatz MC. The challenge of small-scale repeats for indel discovery. *Front Bioeng Biotechnol*. 2015; 3(3):8. [PubMed: 25674564]
26. Garrison, E., Marth, GT. Haplotype-based variant detection from short-read sequencing [Internet]. Available from: <https://arxiv.org/abs/1207.3907>
27. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, et al. Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *J Comput Biol*. 2014 Jun; 21(6):405–19. [PubMed: 24874280]

- 28**. Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, et al. Medical implications of technical accuracy in genome sequencing. *Genome Medicine*. 2016 Mar 2.8(1): 24. A useful inquiry into the testing characteristics of exome and genome sequencing and their implications for establishment of diagnoses in the clinical setting. [PubMed: 26932475]
29. Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen J-B, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics*. 2014 May 2.15(1):125. [PubMed: 24884706]
30. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010 Sep.38(16):e164. [PubMed: 20601685]
31. Dewey, FE., Grove, ME., Priest, JR., Waggott, D., Batra, P., Miller, CL., et al. Sequence to Medical Phenotypes: A Framework for Interpretation of Human Whole Genome DNA Sequence Data. In: Funke, B., editor. *PLoS Genet*. Vol. 11. Public Library of Science; 2015 Oct. p. e1005496
32. Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Alnadi NA, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med*. 2012 Oct 3.4(154):154ra135.
33. Priest JR, Ceresnak SR, Dewey FE, Malloy-Walton LE, Dunn K, Grove ME, et al. Molecular diagnosis of long QT syndrome at 10 days of life by rapid whole genome sequencing. *Heart Rhythm*. 2014 Oct; 11(10):1707–13. [PubMed: 24973560]
34. Parsons DW, Roy A, Yang Y, Wang T, Scollon S, Bergstrom K, et al. Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors. *JAMA Oncol*. 2016 Jan 28.2(5):616.
35. Ramoni RB, Mulvihill JJ, Adams DR, Allard P, Ashley EA, Bernstein JA, et al. The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. *Am J Hum Genet*. 2017 Feb 2; 100(2):185–92. [PubMed: 28157539]
36. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, et al. Clinical interpretation and implications of whole-genome sequencing. *JAMA*. 2014 Mar 12; 311(10):1035–45. [PubMed: 24618965]
37. Ceyhan-Birsoy, O., Kalia, SS., Yu, TW., Agrawal, PB., Holm, IA., McGuire, A., et al. The BabySeq Project. San Diego: Establishing distinctive criteria for reporting genomic sequencing results in healthy versus ill newborns.
38. Vassy, JL., Lautenbach, DM., McLaughlin, HM., Kong, SW., Christensen, KD., Krier, J., et al. *Trials*. Vol. 15. BioMed Central Ltd; 2014. The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine; p. 85
39. Directors ABO. *Genetics in Medicine*. Vol. 17. Nature Publishing Group; 2014 Nov 13. ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing; p. 68-9.
- 40**. Frankel, LA., Pereira, S., McGuire, AL. *Pediatrics*. Vol. 137. American Academy of Pediatrics; 2016 Jan 1. Potential Psychosocial Risks of Sequencing Newborns; p. S24-9. A useful discussion of genetic testing in children prioritizing flexibility and family preference over mandatory reporting of deleterious mutations as proposed by the ACMG.
- 41*. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res*. 2017 Jan 4; 45(D1):D840–5. A large collection of exome studies describing protein coding variation across all human genes. Useful as both a clinical and research reference. [PubMed: 27899611]
42. Kirkpatrick, BE., Riggs, ER., Azzariti, DR., Miller, VR., Ledbetter, DH., Miller, DT., et al. GenomeConnect: matchmaking between patients, clinical laboratories, and researchers to improve genomic knowledge. In: Boycott, K.Hamosh, A., Rehm, H., editors. *Hum Mutat*. Vol. 36. 2015 Oct. p. 974-8.
43. Grove ME, Wolpert MN, Cho MK, Lee SS-J, Ormond KE. Views of genetics health professionals on the return of genomic results. *J Genet Couns*. 2014 Aug; 23(4):531–8. [PubMed: 23728783]
44. Caleshu C, Kasparian NA, Edwards KS, Yeates L, Semsarian C, Perez M, et al. Interdisciplinary psychosocial care for families with inherited cardiovascular diseases. *Trends Cardiovasc Med*. 2016 Oct; 26(7):647–53. [PubMed: 27256036]

45. Wou K, Levy B, Wapner RJ. Chromosomal Microarrays for the Prenatal Detection of Microdeletions and Microduplications. *Clin Lab Med*. 2016 Jun; 36(2):261–76. [PubMed: 27235911]
- 46*. Priest JR, Gawad C, Kahlig KM, Yu JK, O'Hara T, Boyle PM, et al. Early somatic mosaicism is a rare cause of long-QT syndrome. *Proc Natl Acad Sci USA*. 2016 Oct 11; 113(41):11555–60. An example of how careful analysis of comprehensive genome sequence data from a single patient can change our fundamental understanding of the pathogenesis of disease. [PubMed: 27681629]
47. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. De novo mutations in regulatory elements cause neurodevelopmental disorders. 2017
48. Smith AJ, Oswald D, Bodurtha J. Trends in Unmet Need for Genetic Counseling Among Children With Special Health Care Needs, 2001–2010. *Acad Pediatr*. 2015 Sep; 15(5):544–50. [PubMed: 26162247]
49. Bowdin S, Gilbert A, Bedoukian E, Carew C, Adam MP, Belmont J, et al. Recommendations for the integration of genomics into clinical practice. *Genet Med*. 2016 Nov; 18(11):1075–84. [PubMed: 27171546]
50. McGinn S, Bauer D, Brefort T, Dong L, El-Sagheer A, Elsharawy A, et al. New technologies for DNA analysis--a review of the READNA Project. *N Biotechnol*. 2016 May 25; 33(3):311–30. [PubMed: 26514324]
51. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017 May; 27(5):722–36. [PubMed: 28298431]
52. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res*. 2017 May; 27(5):677–85. [PubMed: 27895111]
53. Sweet K, Sturm AC, Schmidlen T, McElroy J, Scheinfeldt L, Manickam K, et al. Outcomes of a Randomized Controlled Trial of Genomic Counseling for Patients Receiving Personalized and Actionable Complex Disease Reports. *J Genet Couns*. 2017 Mar 27; 340(12):c2697.
54. Hunter JE, Irving SA, Biesecker LG, Buchanan A, Jensen B, Lee K, et al. A standardized, evidence-based protocol to assess clinical actionability of genetic disorders associated with genomic variation. *Genet Med*. 2016 Dec; 18(12):1258–68. [PubMed: 27124788]

Key Points

Genome sequencing is now available as a clinical genetic test, with key limitations on the type of genetic information that can be detected.

The current implementation of genome sequencing is based on looking for mismatches between short pieces of a patient's genome and the reference sequence generated by the human genome project.

Genetic counselors and other qualified clinicians are essential for the appropriate and ethical use of genome sequencing in the clinical setting.

Figure 1. An Illustration of the Steps in Clinical Genome Sequencing

A. Generating Raw Sequencing Data: After obtaining informed consent from the patient and family which includes discussion of family preferences regarding incidental findings, a sample of DNA is obtained from the patient and subject to fragmentation and preparation into a “library” for sequencing. The library is subject to sequencing which yields raw data in the form of unordered “reads” of sequence between 50–250 base pairs in length. **B.**

Mapping and Variant Calling: The raw data is mapped to the reference sequence from the human genome project resulting in an ordered alignment of the reads tiled across the entirety of the known reference sequence. In this representation of an alignment, matches to the reference sequence are denoted by dots while mismatches are represented by a capital letter. Some reads may fail to be aligned or alternately may be aligned incorrectly in the wrong location. When a mismatch between the reads and the reference genome is detected in enough reads to provide a threshold of statistical confidence, a variant is reported. In this example a Cytosine is altered to a Thymine at chromosome 5, position 172,660,039 of the hg19 human reference sequence. Note that other mismatches appearing only once or twice

among the mapped reads are not reported and could represent errors in sequencing, errors in alignment, or true variants not meeting the statistical threshold for variant calling. **C.**

Annotation and Interpretation: Functional information is added to describe the predicted role this variant might play in changing the function of a protein product of a gene, in this example the gene is *NKX2-5* a key transcription factor in heart development which when altered can cause congenital heart disease. The functional annotation utilizes a model of gene based on a variety of information which are organized and maintained by organizations such as the National Center for Biotechnology Information (NCBI) at the National Institutes of Health. In this example the annotation includes the presence of this genetic difference is two important databases of human variation. The ExAC database is maintained by the Broad Institute, and this variant is not present among the 63500 adults who have undergone exome sequencing and are ostensibly free of major malformations and early onset disease. The ClinVar database is maintained by the NCBI and curates genetic variation in genes associated with human disease with information contributed by providers and expert users of genetic testing services; from this resource we see that the variant is present has been reported to be Pathogenic on two separate occasions in two different patients with congenital heart disease. Finally, the qualified clinician involved integrates all available annotations of the genetic variant along to interpret the variant in the context the clinical scenario and determine key next steps; reporting, testing of family members, and the known disease or treatment correlates with the genetic variant.