Realizing the Potential of Genomics across the Continuum of Precision Health Care

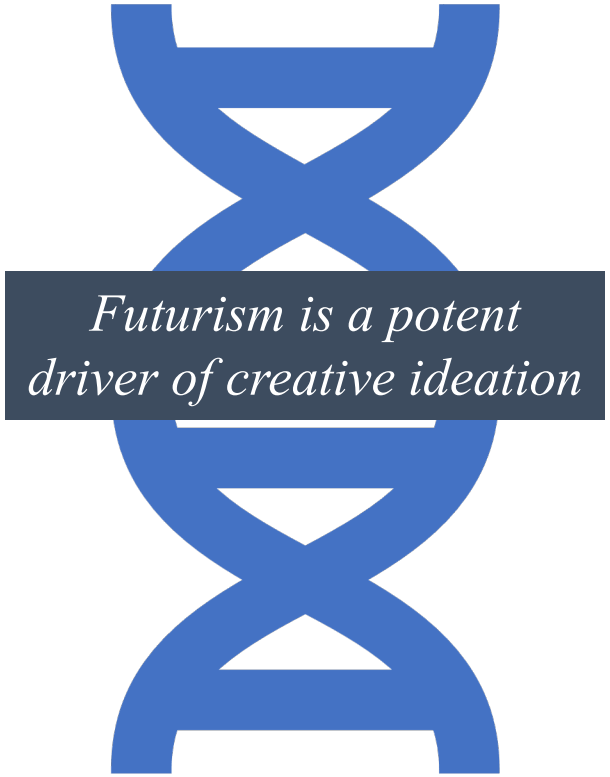
How far have we come — and how far do we have to go?

Euan Ashley BSc, MB ChB, FRCP, DPhil
Roger and Joelle Burnell Professor of Genomics and Precision Health
Associate Dean
Stanford University

The dream…

A genome

- for *anyone* who wants it
- *when* they want it
- integrated into their medical record
- ready to warn of risk for thousands of diseases
- ready to inform prescribing
- seamlessly and securely connected to the genomes and medical information of millions around the world
- linked to a constantly updated anthology of medical evidence connecting genetic variation and risk of disease



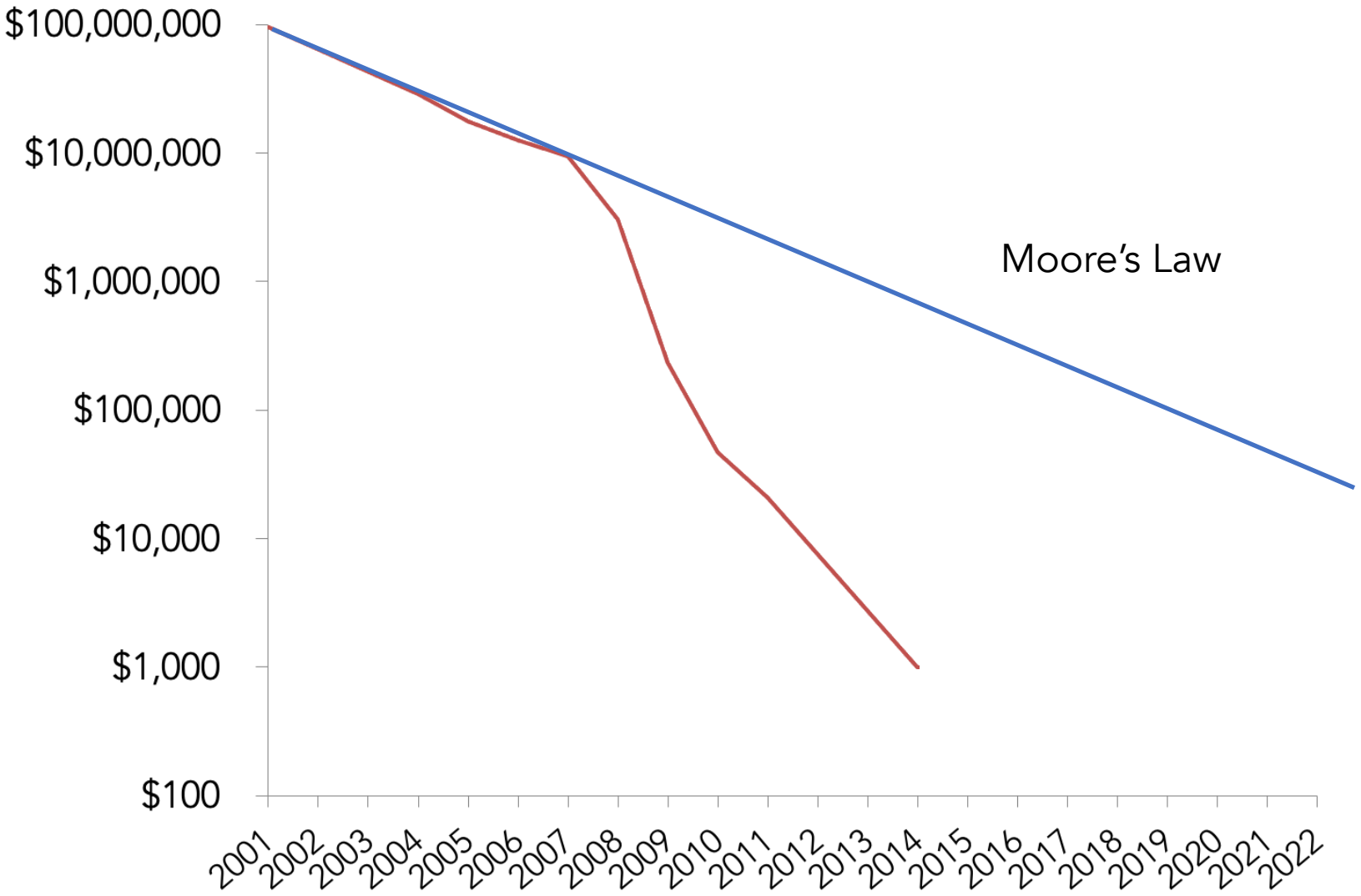
COST

ACCURACY

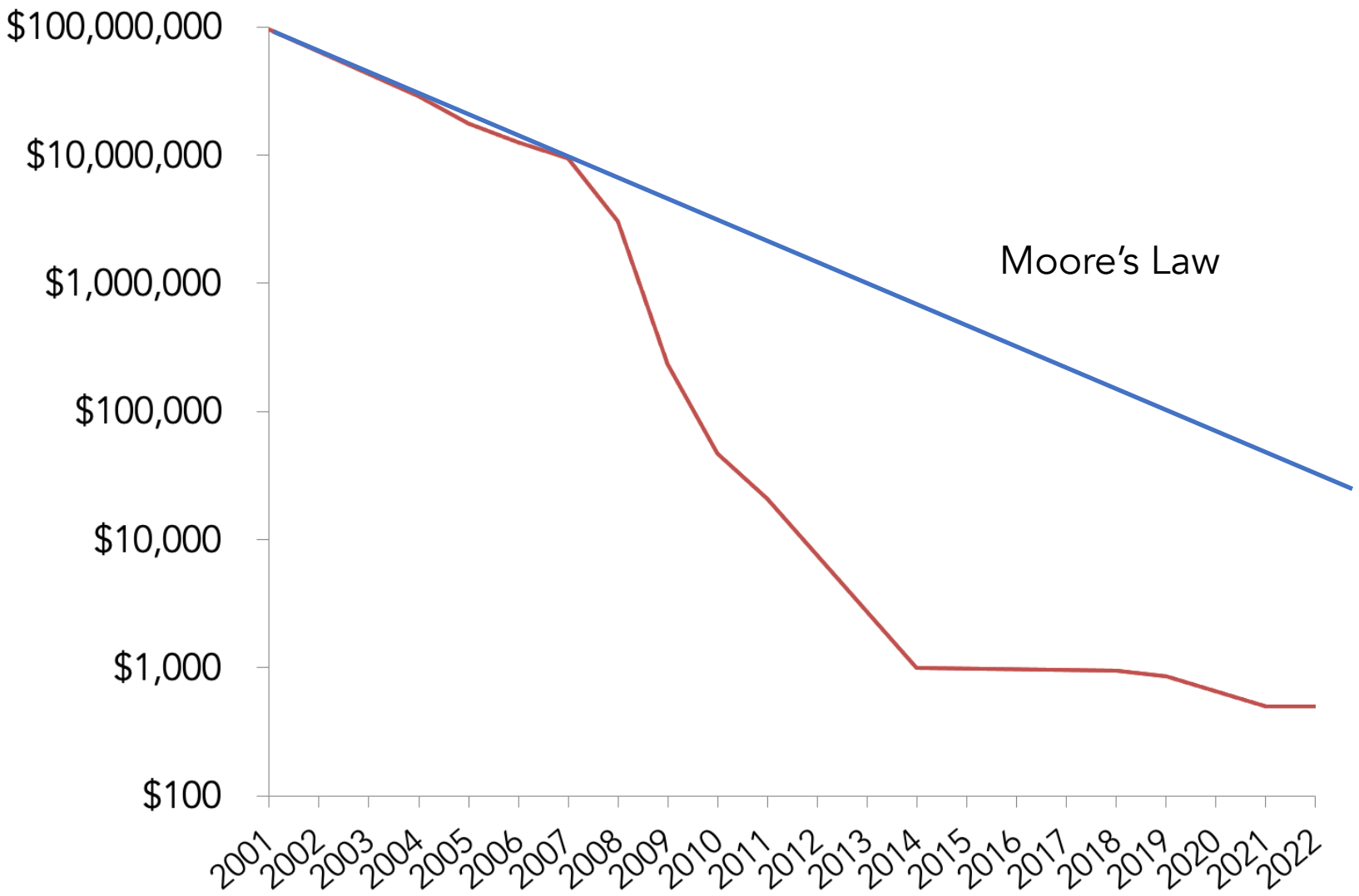
SPEED

IMPLEMENTATION &
INTEGRATION

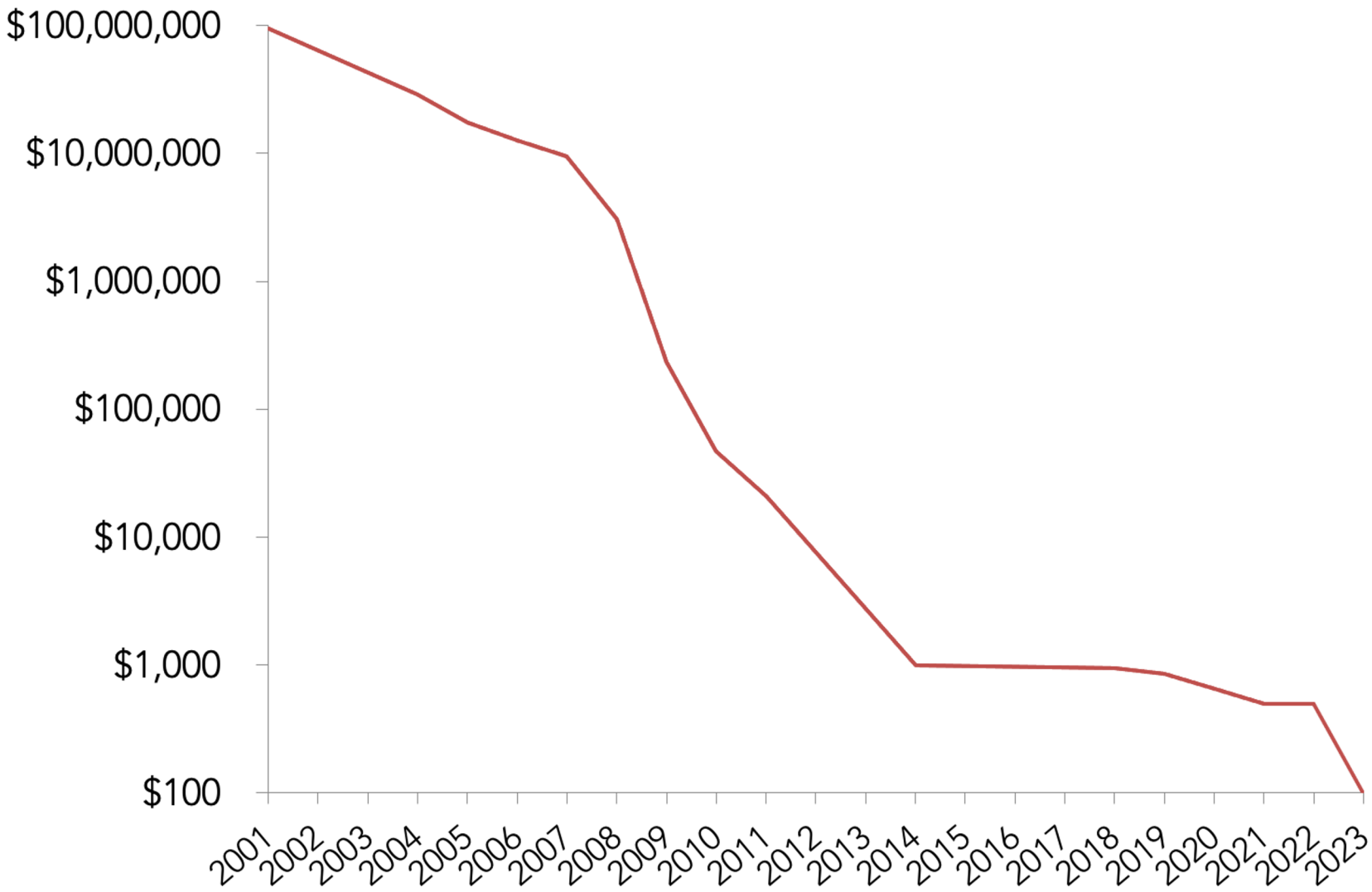
Cost per high depth genome



Cost per high depth genome



Cost per high depth genome



Search Inventory Search

MASERATI

Ferrari Silicon Valley

2750 El Camino Real Redwood City, CA 94061

Sales: (888) 378-7586
Service: (888) 377-1063
Parts & Accessories: (888) 430-4670
Body Shop: (866) 981-0953

Inventory | Service | Parts & Accessories | Body Shop | News & Events | Racing Team | About Us

2013 Ferrari 458 Spider

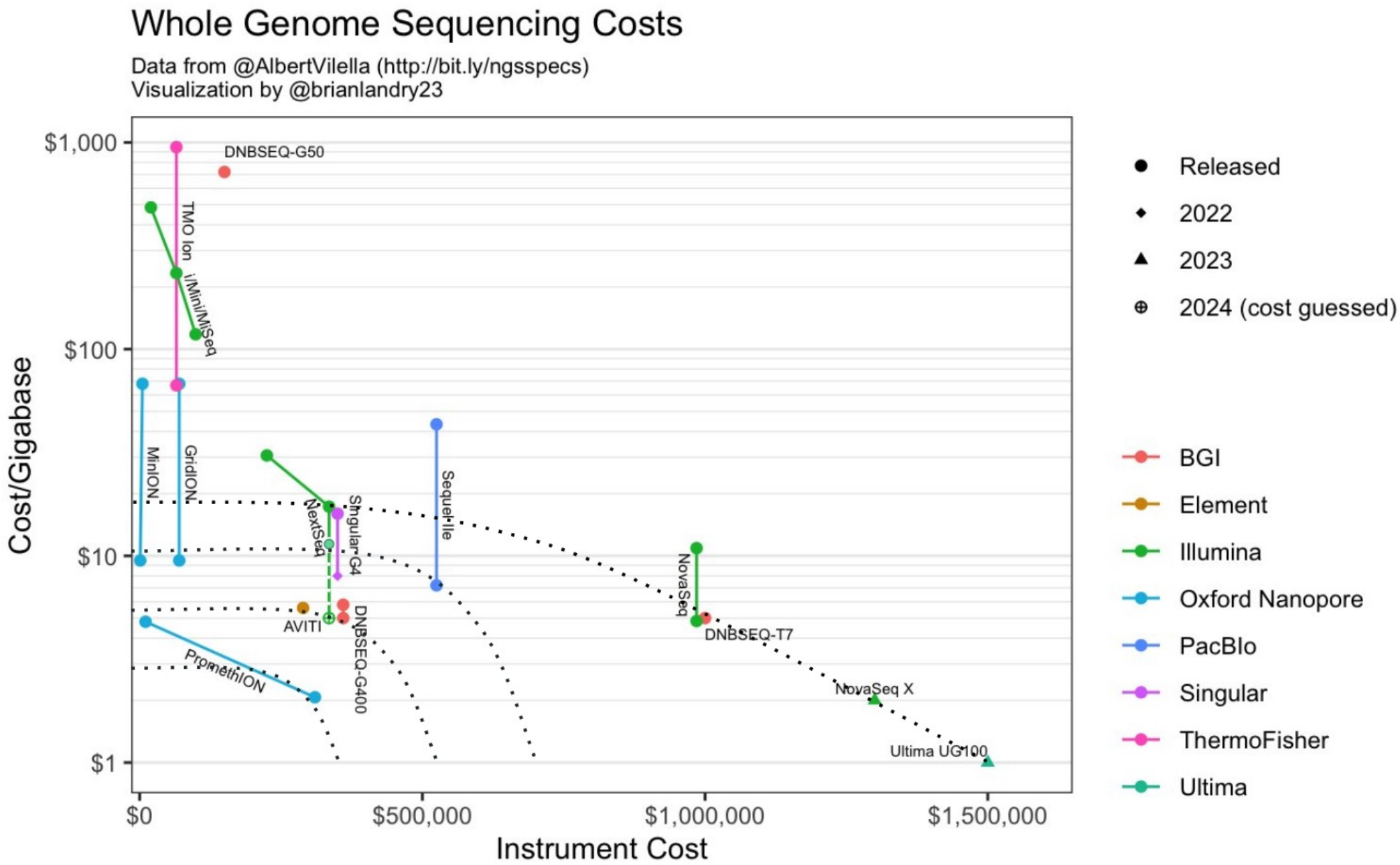
Ferrari Videos

Share

$398,000

↓

$0.01


Whole Genome Sequencing Costs
Data from @AlbertVilella (http://bit.ly/ngsspecs)
Visualization by @brianlandry23
Cost/Gigabase
$1,000
$100
$10
$1
$0
$500,000
$1,000,000
$1,500,000
Instrument Cost
DNBSEQ-G50
TMO Ion
i/Mini/MiSeq
MinION
GridION
NextSeq
Singular G4
Sequel IIe
AVITI
DNBSEQ-G400
PromethION
NovaSeq
DNBSEQ-T7
NovaSeq X
Ultima UG100
Released
2022
2023
2024 (cost guessed)
BGI
Element
Illumina
Oxford Nanopore
PacBio
Singular
ThermoFisher
Ultima

Cost – discussion points

Sequencing includes long and short read technology

Recent (2022) patent expiry of SBS led to entry of several new players to the market. Cost is predicted but accuracy of these technologies is not yet established.

Applications include sequencing high depth, low depth, and counting applications.

Cell free DNA/RNA is short, so applications are more about cost/Gb which favors short read technology

Germline and somatic sequencing requires accuracy in low complexity, highly polymorphic, or structurally variant areas of the genome which favors long read technology

Although sequencing costs have reduced, ancillary costs have come down less:

- analysis costs
- cloud/compute costs
- human curation costs

Most clinical genomes are still billed at thousands of dollars (~3-12k)

Major sequencing providers	Sequencing approach
Illumina (/CLR)	short (/long “synthetic”)
Pacific Biosciences (/Omniome)	long HiFi (short SBB)
Oxford Nanopore	long nanopore
Element Biosciences	Short (SBB)
Ultima Genomics	Short (large, circular wafer)
Singular	short
MGI	short
Thermo Fisher / Iontorrent	short
Genapsys	short

COST

ACCURACY

SPEED

IMPLEMENTATION &
INTEGRATION

Precision medicine needs to be accurate medicine

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Towards precision medicine

Euan A. Ashley

Abstract | There is great potential for genome sequencing to enhance patient care through improved diagnostic sensitivity and more precise therapeutic targeting. To maximize this potential, genomics strategies that have been developed for genetic discovery — including DNA-sequencing technologies and analysis algorithms — need to be adapted to fit clinical needs. This will require the optimization of alignment algorithms, attention to quality-coverage metrics, tailored solutions for paralogous or low-complexity areas of the genome, and the adoption of consensus standards for variant calling and interpretation. Global sharing of this more accurate genotypic and phenotypic data will accelerate the determination of causality for novel genes or variants. Thus, a deeper understanding of disease will be realized that will allow its targeting with much greater therapeutic precision.

The sequencing of the human genome led many to speculate on the near-term potential for clinical medicine[1]. Understanding the genetic basis of disease was naturally expected to lead to better targeted therapies. Indeed, the steep decline in the cost of sequencing, pursuant to the invention of 'next-generation' technologies, facilitated the discovery of many more causative genes[2,3] and, more recently, application to individual patients, including several widely reported examples of genome-driven medical decision making[4–6]. Pilot studies explored the use of genomic information more broadly in patient care[7–9] and the US National Human Genome Research Institute (NHGRI) laid out a 20-year plan for translating insights from genomics to medicine[10,11]. Additionally, direct-to-consumer companies put genotypes in the hands of interested participants[12]. However, the brightest spotlight was provided in 2015 by President Obama in his State of the Union address where he laid out a vision for a national Precision Medicine Initiative in the United States[13,14].

The term 'precision medicine' (BOX 1) was first given prominence by a publication from the US National Research Council that sought to inspire a new taxonomy for disease classification via a knowledge network[15]. In the appendix of that publication, the authors clarify that its coining, as opposed to the more commonly used term 'personalized medicine', was intended to convey the principle that although therapeutics were rarely developed for single individuals, increasingly, subgroups of patients could be defined, often by genomics, and targeted in more specific ways. Worldwide internet searches for the term increased dramatically after the State of the Union address and have remained at similar levels to that of 'personalized medicine' ever since (FIG. 1a).

The timing does seem right for a new approach: genomic data are more readily available, we have a greater understanding of population-scale genetic variation[16,17], and approaches to data integration with electronic medical records will lead to much improved characterization of phenotypes[18]. However, for precision medicine to succeed it also needs to be more accurate. The current algorithms for genome analysis were developed for population or cohort variant discovery where the consequences of reduced accuracy are a lost opportunity for discovery. By contrast, an inaccurate clinical genetic test could lead to very serious consequences for individuals and families with genetic disease. In this Review, I describe promising applications of precision medicine as it currently exists then move on to discuss the challenges our community needs to face, in the areas of sequencing technology, algorithm development and data sharing, to bring genomics up to clinical grade.

Promising applications of precision medicine

Cystic fibrosis. In the State of the Union address, President Obama specifically gave as an example the drug ivacaftor, which was developed for patients with cystic fibrosis. Cystic fibrosis is an autosomal recessive disease that affects approximately 70,000 people worldwide and that is caused by variants in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. The protein product of this gene is an epithelial ion channel located on the cell surface where it regulates cellular chloride transit. Mutations of *CFTR* cause abnormal regulation of salt and water, which particularly affects the function of the lungs, pancreas and sweat glands. Recurrent pulmonary disease and resistant infection represent the major therapeutic challenges of cystic fibrosis, and traditional therapies have focused entirely

Center for Inherited Cardiovascular Disease, Falk Cardiovascular Research Building, Stanford Medicine, 870 Quarry Road, Stanford, California 94305, USA.
euan@stanford.edu

doi:10.1038/nrg.2016.86
Published online 16 Aug 2016

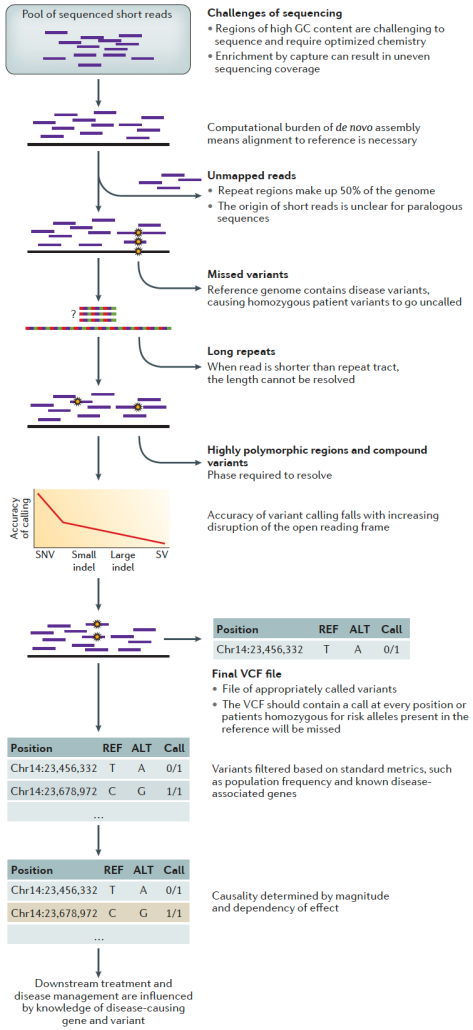


Figure 2 | **Origins of reduced accuracy in clinical genomics from short sequencing reads.** Accuracy can be optimized at multiple steps in the route from DNA to variant calling and reporting. Regions of high GC content require tailored approaches both for capture and for sequencing. Enrichment by capture leads to uneven coverage. Alignment to the haploid reference sequence is required for short reads because of the computational burden of *de novo* assembly. Paralogous sequence is common throughout the genome, and the origin of a short read cannot be determined in 5% of cases. For diseases such as Huntington disease that are caused by repeat tracts where the most severe disease is associated with tracts longer than the short reads, length cannot be resolved. Similarly, highly polymorphic regions such as the major histocompatibility complex (MHC), which is used for HLA typing in transplantation and for risk quantification for multiple immune diseases, cannot be resolved. Along with compound variants such as multiple nucleotide variants, these cannot be adequately resolved without phasing. The accuracy of calling decreases with increasing disruption of the open reading frame. However, variants that are more disruptive of the open reading frame, such as structural variants (SVs), are generally more likely to cause disease. As the human reference sequence is made up of DNA from multiple individuals — and contains risk variants that reduce the accuracy of alignment and result in missed calls for homozygous risk alleles such as factor V Leiden — a call at every position should be included in the variant call file. Finally, causality is a complex construct with final effects determined by magnitude and dependency of the variant effect. Causal variants can often lead to changes in clinical management: in some cases, precision therapy for the patient and in other cases changes in screening are recommended for the family. ALT, alternative allele; indel, small insertion or deletion; REF, reference allele; SNV, single nucleotide variant; VCF, variant call format.

Nature Reviews Genetics 17 (507-522)

Original Investigation

Clinical Interpretation and Implications of Whole-Genome Sequencing

Frederick E. Dewey, MD; Megan E. Grove, MS; Cuiping Pan, PhD; Benjamin A. Goldstein, PhD; Jonathan A. Bernstein, MD, PhD; Hassan Chaib, PhD; Jason D. Merker, MD, PhD; Rachel L. Goldfeder, BS; Gregory M. Enns, MB, ChB; Sean P. David, MD, DPhil; Neda Pakdaman, MD; Kelly E. Ormond, MS; Colleen Caleshu, MS; Kerry Kingham, MS; Teri E. Klein, PhD; Michelle Whirl-Carrillo, PhD; Kenneth Sakamoto, MD; Matthew T. Wheeler, MD, PhD; Atul J. Butte, MD, PhD; James M. Ford, MD, PhD; Linda Boxer, MD; John P. A. Ioannidis, MD, PhD; Alan C. Yeung, MD; Russ B. Altman, MD, PhD; Themistocles L. Assimes, MD, PhD; Michael Snyder, PhD; Euan A. Ashley, MRCP, DPhil; Thomas Quertermous, MD

IMPORTANCE Whole-genome sequencing (WGS) is increasingly applied in clinical medicine and is expected to uncover clinically significant findings regardless of sequencing indication.

OBJECTIVES To examine coverage and concordance of clinically relevant genetic variation provided by WGS technologies; to quantitate inherited disease risk and pharmacogenomic findings in WGS data and resources required for their discovery and interpretation; and to evaluate clinical action prompted by WGS findings.

DESIGN, SETTING, AND PARTICIPANTS An exploratory study of 12 adult participants recruited at Stanford University Medical Center who underwent WGS between November 2011 and March 2012. A multidisciplinary team reviewed all potentially reportable genetic findings. Five physicians proposed initial clinical follow-up based on the genetic findings.

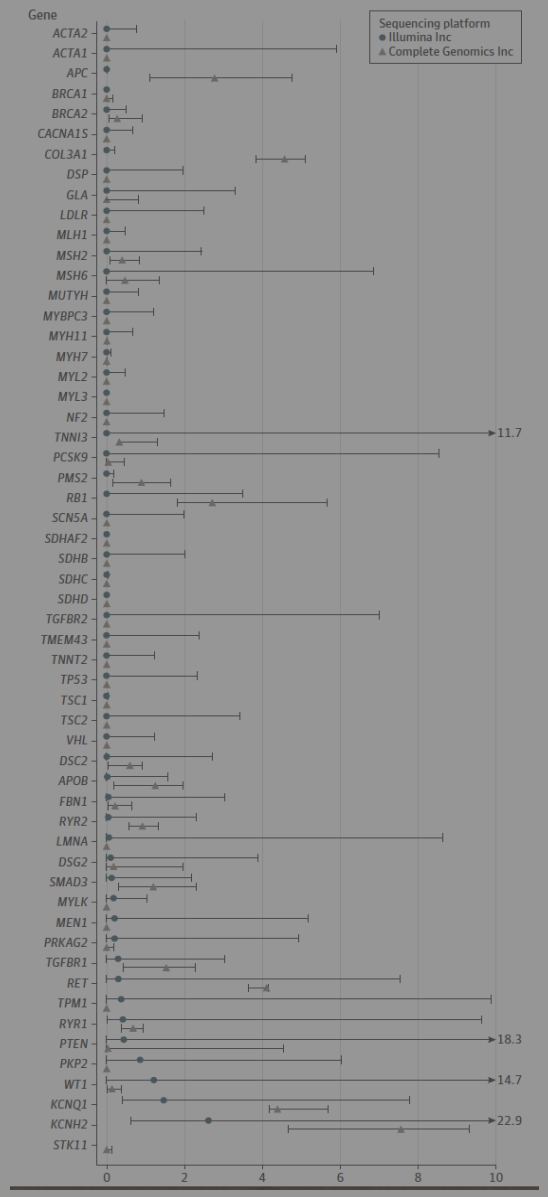
MAIN OUTCOMES AND MEASURES Genome coverage and sequencing platform concordance in different categories of genetic disease risk, person-hours spent curating candidate disease-risk variants, interpretation agreement between trained curators and disease genetics databases, burden of inherited disease risk and pharmacogenomic findings, and burden and interrater agreement of proposed clinical follow-up.

RESULTS Depending on sequencing platform, 10% to 19% of inherited disease genes were not covered to accepted standards for single nucleotide variant discovery. Genotype concordance was high for previously described single nucleotide genetic variants (99%-100%) but low for small insertion/deletion variants (53%-59%). Curation of 90 to 127 genetic variants in each participant required a median of 54 minutes (range, 5-223 minutes) per genetic variant, resulted in moderate classification agreement between professionals (Gross κ, 0.52; 95% CI, 0.40-0.64), and reclassified 69% of genetic variants cataloged as disease causing in mutation databases to variants of uncertain or lesser significance. Two to 6 personal disease-risk findings were discovered in each participant, including 1 frameshift deletion in the *BRCA1* gene implicated in hereditary breast and ovarian cancer. Physician review of sequencing findings prompted consideration of a median of 1 to 3 initial diagnostic tests and referrals per participant, with fair interrater agreement about the suitability of WGS findings for clinical follow-up (Fleiss κ, 0.24; $P < 001$).

CONCLUSIONS AND RELEVANCE In this exploratory study of 12 volunteer adults, the use of WGS was associated with incomplete coverage of inherited disease genes, low reproducibility of detection of genetic variation with the highest potential clinical effects, and uncertainty about clinically reportable findings. In certain cases, WGS will identify clinically actionable genetic variants warranting early medical intervention. These issues should be considered when determining the role of WGS in clinical medicine.

JAMA 2014 311:1035-44 2014

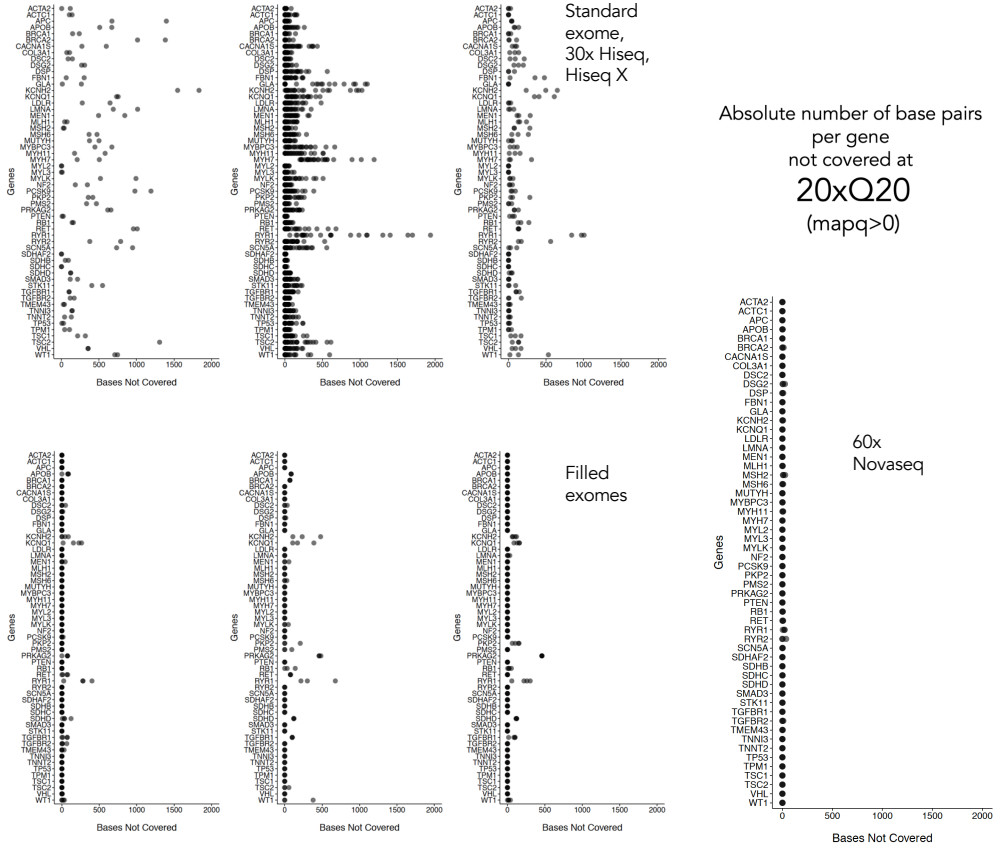
Figure 2. Missing Coverage of 56 Genes the ACMG Recommends for Pathogenic Variant Discovery, Review, and Reporting in WGS



A precision metric for clinical genome sequencing

Rachel L. Goldfeder, Euan A. Ashley

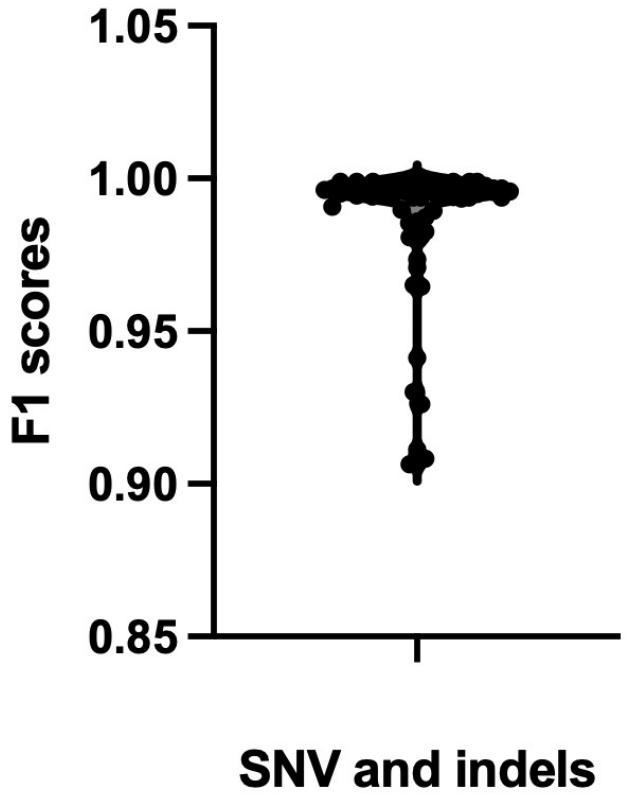
Stanford University, Department of Medicine

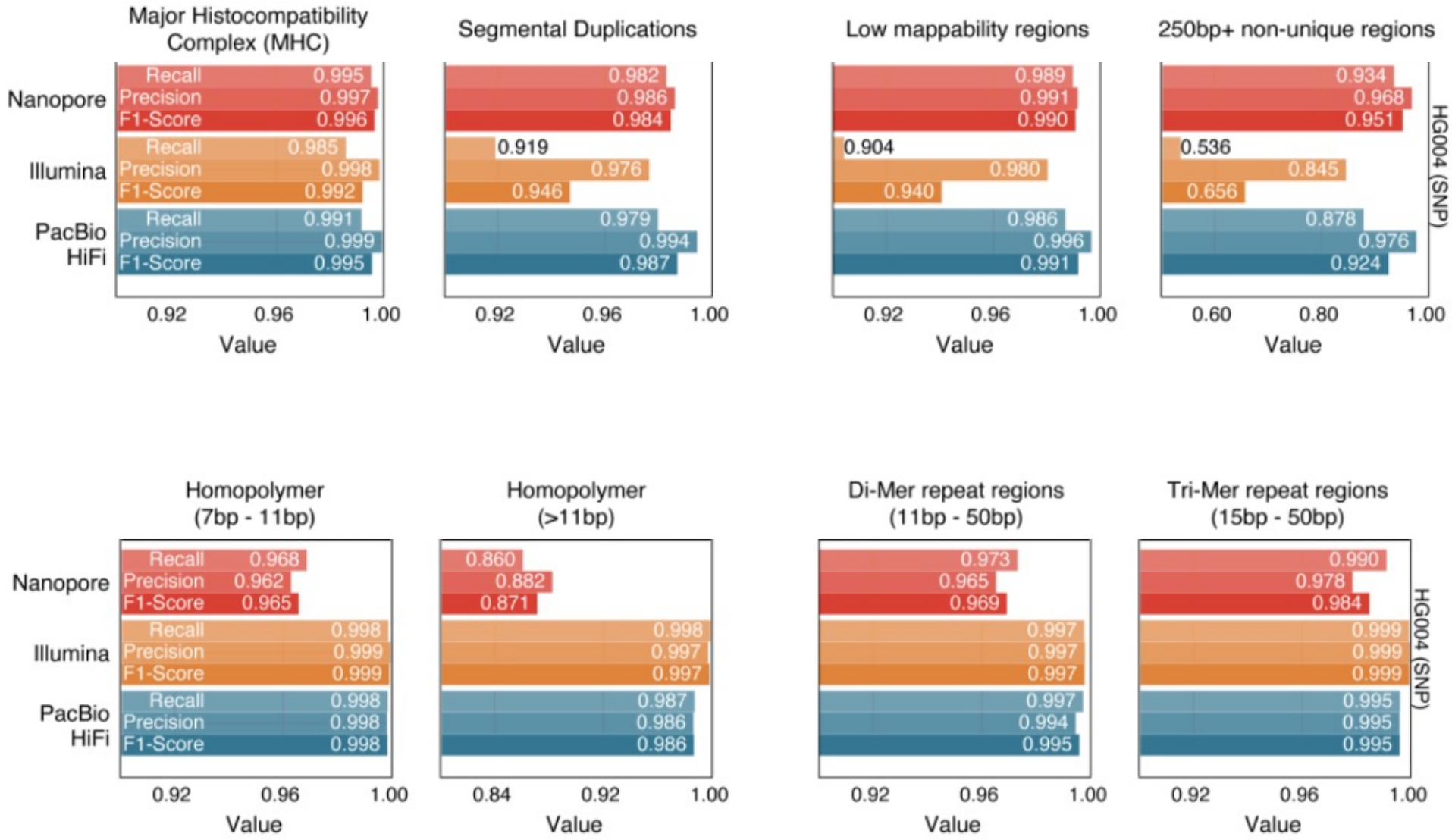


Major improvements in accuracy

GIAB: Truth competition v2, June 2020

- Data (fastqs) from Illumina NovaSeq, PacBio HiFi (Sequel 2), Nanopore PromethION
- Trio data
- Benchmark dataset greatly expanded including with indels
- Dataset included difficult to sequence areas including MHC
- 64 submissions, 20 teams
- Winners: Sentieon, DRAGEN, Google (Genomics/Health), Seven Bridges, UCSC, Wang Lab, Roche Sequencing Solutions




Major Histocompatibility Complex (MHC)
Segmental Duplications
Low mappability regions
250bp+ non-unique regions
Homopolymer (7bp - 11bp)
Homopolymer (>11bp)
Di-Mer repeat regions (11bp - 50bp)
Tri-Mer repeat regions (15bp - 50bp)
Nanopore
Illumina
PacBio HiFi
Recall
Precision
F1-Score
0.995
0.997
0.996
0.985
0.998
0.992
0.991
0.999
0.995
0.982
0.986
0.984
0.919
0.976
0.946
0.979
0.994
0.987
0.989
0.991
0.990
0.904
0.980
0.940
0.986
0.996
0.991
0.934
0.968
0.951
0.536
0.845
0.656
0.878
0.976
0.924
0.968
0.962
0.965
0.998
0.999
0.999
0.998
0.998
0.998
0.860
0.882
0.871
0.998
0.997
0.997
0.987
0.986
0.986
0.973
0.965
0.969
0.997
0.997
0.997
0.997
0.994
0.995
0.990
0.978
0.984
0.999
0.999
0.999
0.995
0.995
0.995
HG004 (SNP)
HG004 (SNP)
0.92
0.96
1.00
0.60
0.80
0.84
Value

SPECIAL SECTION COMPLETING THE HUMAN GENOME

RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

Sergey Nurk[1†], Sergey Koren[1†], Arang Rhie[1†], Mikko Rautiainen[1†], Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9‡], Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[20], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[25], Erik Garrison[11,25], Patrick G. S. Grady[10], Tina A. Graves-Lindsay[26], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[29], Michael W. Hunkapiller[29], Chirag Jain[1,30], Miten Jain[11], Erich D. Jarvis[22,23], Peter Kerpedjiev[31], Melanie Kirsche[9], Mikhail Kolmogorov[32], Jonas Korlach[33], Milinn Kremitzki[26], Heng Li[16,17], Valerie V. Maduro[13], Tobias Marschall[34], Ann M. McCartney[1], Jennifer McDaniel[35], Danny E. Miller[4,36], James C. Mullikin[14,37], Eugene W. Myers[31], Nathan D. Olson[35], Benedict Paten[11], Paul Peluso[33], Pavel A. Pevzner[32], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,38,39], Jeffrey A. Rosenfeld[40], Steven L. Salzberg[8,41], Valerie A. Schneider[42], Fritz J. Sedlazeck[43], Kishwar Shafin[11], Colin J. Shew[44], Alaina Shumate[45], Ying Sims[19], Arian F. A. Smit[46], Daniela C. Soto[44], Ivan Sović[29,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[42], James Torrance[19], Justin Wagner[35], Brian P. Walenz[1], Aaron Wenger[28], Jonathan M. D. Wood[19], Chunlin Xiao[42], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[45], Megan Y. Dennis[44], Ivan A. Alexandrov[6,7,51], Jennifer L. Gerton[13,52], Rachel J. O'Neill[10], Winston Timp[8,41], Justin M. Zook[35], Michael C. Schatz[9,49], Evan E. Eichler[4,53*], Karen H. Miga[11,54*], Adam M. Phillippy[1*]

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

The current human reference genome was released by the Genome Reference Consortium (GRC) in 2013 and most recently patched in 2019 (GRCh38.p13) (*1*). This reference traces its origin to the publicly funded Human Genome Project (*2*) and has been continually improved over the past two decades. Unlike the competing Celera effort (*3*) and most modern sequencing projects based on "shotgun" sequence assembly (*4*), the GRC assembly was constructed from sequenced bacterial artificial chromosomes (BACs) that were ordered and oriented along the human genome by means of radiation hybrid, genetic linkage, and fingerprint maps. However, limitations of BAC cloning led to an underrepresentation of repetitive sequences, and the opportunistic assembly of BACs derived from multiple individuals resulted in a mosaic of haplotypes. As a result, several GRC assembly gaps are unsolvable because of incompatible structural polymorphisms on their flanks, and many other repetitive and polymorphic regions were left unfinished or incorrectly assembled (*5*).

The GRCh38 reference assembly contains 151 mega–base pairs (Mbp) of unknown sequence distributed throughout the genome, including pericentromeric and subtelomeric regions, recent segmental duplications, ampliconic gene arrays, and ribosomal DNA (rDNA) arrays, all of which are necessary for fundamental cellular processes (Fig. 1A). Some of the largest reference gaps include human satellite (HSat) repeat arrays and the short arms of all five acrocentric chromosomes, which are represented in GRCh38 as multimegabase stretches of unknown bases (Fig. 1, B and C). In addition to these apparent gaps, other regions of GRCh38 are artificial or are otherwise incorrect. For example, the pericentromeric alpha satellite arrays are represented as computationally generated models of alpha satellite monomers to serve as decoys for resequencing analyses (*6*), and sequence assigned to the short arm of chromosome 21 appears falsely duplicated and poorly assembled (*7*). When compared with other human genomes, GRCh38 also shows a genome-wide deletion bias that is indicative of incomplete assembly (*8*). Despite finishing efforts from both the Human Genome Project (*9*) and GRC (*1*) that improved the quality of the reference, there was limited

[1]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [2]Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, CA, USA. [3]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University, Saint Petersburg, Russia. [4]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [5]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA. [6]Sirius University of Science and Technology, Sochi, Russia. [7]Vavilov Institute of General Genetics, Moscow, Russia. [8]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD, USA. [9]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [10]Institute for Systems Genomics and Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA. [11]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [12]University of Geneva Medical School, Geneva, Switzerland. [13]Stowers Institute for Medical Research, Kansas City, MO, USA. [14]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [15]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [16]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA. [17]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [18]DNAnexus, Mountain View, CA, USA. [19]Wellcome Sanger Institute, Cambridge, UK. [20]Department of Genetics, University of Cambridge, Cambridge, UK. [21]Inscripta, Boulder, CO, USA. [22]Laboratory of Neurogenetics of Language and The Vertebrate Genome Lab, The Rockefeller University, New York, NY, USA. [23]Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. [24]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. [25]University of Tennessee Health Science Center, Memphis, TN, USA. [26]McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO, USA. [27]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA. [28]Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [29]Pacific Biosciences, Menlo Park, CA, USA. [30]Department of Computational and Data Sciences, Indian Institute of Science, Bangalore KA, India. [31]Reservoir Genomics LLC, Oakland, CA, USA. [32]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [33]Undiagnosed Diseases Program, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [34]Heinrich Heine University Düsseldorf, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Düsseldorf, Germany. [35]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, USA. [36]Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA, USA. [37]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [38]Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA, USA. [39]Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia. [40]Cancer Institute of New Jersey, New Brunswick, NJ, USA. [41]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [42]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [43]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [44]Genome Center, MIND Institute, Department of Biochemistry and Molecular Medicine, University of California, Davis, CA, USA. [45]Institute for Systems Biology, Seattle, WA, USA. [46]Digital BioLogic d.o.o., Ivanić-Grad, Croatia. [47]Chan Zuckerberg Biohub, San Francisco, CA, USA. [48]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC, USA. [49]Department of Biology, Johns Hopkins University, Baltimore, MD, USA. [50]Department of Pathology, University of Pittsburgh, Pittsburgh, PA, USA. [51]Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia. [52]Department of Biochemistry and Molecular Biology, University of Kansas Medical School, Kansas City, MO, USA. [53]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [54]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA.
*Corresponding author. Email: eee@gs.washington.edu (E.E.E.); khmiga@ucsc.edu (K.H.M.); adam.phillippy@nih.gov (A.M.P.)
†These authors contributed equally to this work. ‡Present address: Oxford Nanopore Technologies Inc., Lexington, MA, USA.

ARTICLES
https://doi.org/10.1038/s41587-021-01158-1
nature biotechnology
Check for updates

Curated variation benchmarks for challenging medically relevant autosomal genes

Justin Wagner[1], Nathan D. Olson[1], Lindsay Harris[1], Jennifer McDaniel[1], Haoyu Cheng[2], Arkarachai Fungtammasan[3], Yih-Chii Hwang[3], Richa Gupta[3], Aaron M. Wenger[4], William J. Rowell[4], Ziad M. Khan[5], Jesse Farek[5], Yiming Zhu[5], Aishwarya Pisupati[5], Medhat Mahmoud[5], Chunlin Xiao[6], Byunggil Yoo[7], Sayed Mohammad Ebrahim Sahraeian[8], Danny E. Miller[9,10], David Jáspez[11], José M. Lorenzo-Salazar[11], Adrián Muñoz-Barrera[11], Luis A. Rubio-Rodríguez[11], Carlos Flores[11,12,13], Giuseppe Narzisi[14], Uday Shanker Evani[14], Wayne E. Clarke[14], Joyce Lee[15], Christopher E. Mason[16], Stephen E. Lincoln[17], Karen H. Miga[18], Mark T. W. Ebbert[19,20,21], Alaina Shumate[22,23], Heng Li[2], Chen-Shan Chin[3,24] ✉, Justin M. Zook[1,24] ✉ and Fritz J. Sedlazeck[5,24] ✉

The repetitive nature and complexity of some medically relevant genes poses a challenge for their accurate analysis in a clinical setting. The Genome in a Bottle Consortium has provided variant benchmark sets, but these exclude nearly 400 medically relevant genes due to their repetitiveness or polymorphic complexity. Here, we characterize 273 of these 395 challenging autosomal genes using a haplotype-resolved whole-genome assembly. This curated benchmark reports over 17,000 single-nucleotide variations, 3,600 insertions and deletions and 200 structural variations each for human genome reference GRCh37 and GRCh38 across HG002. We show that false duplications in either GRCh37 or GRCh38 result in reference-specific, missed variants for short- and long-read technologies in medically relevant genes, including *CBS*, *CRYAA* and *KCNE1*. When masking these false duplications, variant recall can improve from 8% to 100%. Forming benchmarks from a haplotype-resolved whole-genome assembly may become a prototype for future benchmarks covering the whole genome.

Authoritative benchmark samples are driving the development of technologies and the discovery of new variants, enabling highly accurate clinical genome sequencing and advancing our detection and understanding of the impact of many genomic variations on human disease at scale. With recent improvements in sequencing technologies[1], assembly algorithms[2–4] and variant-calling methods[5], genomics offers more insights into challenging genes associated with human diseases across a higher number of patients[6]. Still, challenges remain for medically relevant genes that are often repetitive or highly polymorphic[7,8]. In fact, a recent study found that 13.8% (17,561) of pathogenic variants identified by a high-throughput clinical laboratory were challenging to detect with short-read sequencing[9]. These included challenging variants such as variants 15–49 bp in size, small copy-number variations (CNVs), complex variants and variants in low-complexity or segmentally duplicated regions.

The Genome in a Bottle (GIAB) consortium develops benchmarks to advance accurate human genomic research and clinical applications of sequencing. GIAB provides highly curated benchmark sets for single-nucleotide variant (SNV)[10], small insertion and deletion (INDEL)[11] and structural variant (SV) calling[12]. Here, we define SNVs as single base substitutions, while INDELs are defined as insertions and deletions smaller than 50 bp, in contrast to insertions and deletions larger than 50 bp, which we refer to as SVs. Furthermore, GIAB and the Food and Drug Administration (FDA) host periodic precisionFDA challenges providing a snapshot and recommendations for small variant calling enabling the high precision and sensitivity required for clinical research, with

[1]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA. [2]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [3]DNAnexus, Inc., Mountain View, CA, USA. [4]Pacific Biosciences, Menlo Park, CA, USA. [5]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [7]Genomic Medicine Center, Children's Mercy Kansas City, Kansas City, MO, USA. [8]Roche Sequencing Solutions, Santa Clara, CA, USA. [9]Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA, USA. [10]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [11]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain. [12]CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain. [13]Research Unit, Hospital Universitario N.S. de Candelaria, Santa Cruz de Tenerife, Spain. [14]New York Genome Center, New York, NY, USA. [15]Bionano Genomics, San Diego, CA, USA. [16]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [17]Invitae, San Francisco, CA, USA. [18]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [19]Sanders-Brown Center on Aging, University of Kentucky, Lexington, KY, USA. [20]Department of Internal Medicine, Division of Biomedical Informatics, University of Kentucky, Lexington, KY, USA. [21]Department of Neuroscience, University of Kentucky, Lexington, KY, USA. [22]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [23]Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA. [24]These authors contributed equally: Chen-Shan Chin, Justin M. Zook, Fritz J. Sedlazeck. ✉e-mail: jchin@dnanexus.com; jzook@nist.gov; Fritz.Sedlazeck@bcm.edu

Clinical Chemistry 00:0
1–3 (2022)

Perspective

Mind the Gap: The Complete Human Genome Unlocks Benefits for Clinical Genomics

Daniel Seung Kim,[a,†] Laurens Wiel,[a,†] and Euan A. Ashley[a,b,c,*]

The initial draft of the human reference genome was assembled in 2001 and, since then, nearly all human genomes sequenced have been mapped to this monoploid reference. With genome or exome sequencing becoming more routine for clinical diagnostics in the last 2 decades, the impact of the initial reference genome has grown. However, there have remained gaps in the reference, particularly around centromeres and telomeres, resulting predominantly from limitations in sequencing technology. Over the course of several releases, many of these gaps were addressed by the Genome Reference Consortium, most recently human build 38 (GRCh38). However, it is remarkable to consider that, until this year, over 200 megabases (mb), or approximately 8% of the human genome, remained missing from reference GRCh38 (1).

The GRCh38 human genome draft was the first to characterize sequence information of the centromeres, albeit with limited coverage and accuracy. However, the telomeres remained mostly elusive. To address this, the telomere-to-telomere (T2T) consortium used DNA from a complete hydatidiform mole, taking advantage of its diploid androgenetic nature. The structure of the hydatidiform mole genome allowed the T2T investigators to better resolve genetic regions that vary considerably between parental chromosomes in each individual, creating a high-fidelity reference map. The T2T consortium also leveraged recent advances in long-read and ultra-long-read sequencing, allowing for sequencing of highly repetitive sequences that are prevalent in both centromeres and telomeres, to accurately assemble the most complete human genome draft to date (1).

The T2T consortium found 3604 new genes, of which 1956 mapped to regions within the nearly 200 mb of sequence missing from the previous genome drafts (1). Notably, the complete sequence of a human Y chromosome (not included in the hydatidiform mole cell line) was subsequently released by the T2T consortium after publication of their initial landmark manuscripts. Additionally, they reported enrichment of segmental duplications within the short arms of acrocentric chromosomes (13, 14, 15, 21, and 22). Interestingly, they also revealed these were enriched with euchromatic segmental duplications [a possible cause of chromosomal translocations (2)] and approximately 400 ribosomal DNAs (1).

The implications of the new T2T reference human genome (T2T-CHM13) for clinical genomics are many. Perhaps most important is the significant improvement in sequence alignment and variant calling. Using data from the 1000 Genomes Project, investigators identified an average of 1 million additional high-quality variants per sample using the T2T-CHM13 reference sequence as compared to GRCh38, many of which were found in regions corresponding to the newly sequenced telomeric, centromeric, and short arms of the acrocentric chromosomes (3). Importantly, they were also able to reduce the number of spurious, false-positive, genetic variants identified by approximately 10 000 per sample (3).

T2T investigators were able to fully characterize the gene structure of several genes that had previously been unresolved in GRCh38 due to repetitive genetic elements. An important example is *LPA*, which encodes lipoprotein(a), a molecule causally implicated in coronary artery disease, valvular disease, and atrial fibrillation. In addition to fully characterizing the number of kringle IV repeat domains in *LPA* (which determines in an inverse manner the plasma levels of lipoprotein(a) and thereby cardiovascular disease risk), investigators also found rare coding variation within *LPA* and other genes that had previously been unresolved [e.g., *TBC1D3*, which is implicated in neurodevelopment and speciation between great apes and humans (2)].

In addition to improved genome variant calling and resolution of gene structure of repetitive genes, T2T investigators were also able to characterize the total number of duplicated genes, some of which are critically important for human growth and development. As an example, T2T-CHM13 contained 23 copies of *FRG1* (an increase from 9 copies in a prior reference). *FRG1* is a centromeric gene on chromosome 4q35, and is

[a]Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA; [b]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA; [c]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA.
*Address correspondence to this author at: Falk CVRC, 300 Pasteur Dr., Stanford, CA 94305. E-mail euan@stanford.edu.
†Denotes equal contribution
Received June 17, 2022; accepted June 23, 2022.
https://doi.org/10.1093/clinchem/hvac133

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease

K. Splinter, D.R. Adams, C.A. Bacino, H.J. Bellen, J.A. Bernstein, A.M. Cheatle-Jarvela, C.M. Eng, C. Esteves, W.A. Gahl, R. Hamid, H.J. Jacob, B. Kikani, D.M. Koeller, I.S. Kohane, B.H. Lee, J. Loscalzo, X. Luo, A.T. McCray, T.O. Metz, J.J. Mulvihill, S.F. Nelson, C.G.S. Palmer, J.A. Phillips III, L. Pick, J.H. Postlethwait, C. Reuter, V. Shashi, D.A. Sweetser, C.J. Tifft, N.M. Walley, M.F. Wangler, M. Westerfield, M.T. Wheeler, A.L. Wise, E.A. Worthey, S. Yamamoto, and E.A. Ashley, for the Undiagnosed Diseases Network*

ABSTRACT

BACKGROUND

Many patients remain without a diagnosis despite extensive medical evaluation. The Undiagnosed Diseases Network (UDN) was established to apply a multidisciplinary model in the evaluation of the most challenging cases and to identify the biologic characteristics of newly discovered diseases. The UDN, which is funded by the National Institutes of Health, was formed in 2014 as a network of seven clinical sites, two sequencing cores, and a coordinating center. Later, a central biorepository, a metabolomics core, and a model organisms screening center were added.

METHODS

We evaluated patients who were referred to the UDN over a period of 20 months. The patients were required to have an undiagnosed condition despite thorough evaluation by a health care provider. We determined the rate of diagnosis among patients who subsequently had a complete evaluation, and we observed the effect of diagnosis on medical care.

RESULTS

A total of 1519 patients (53% female) were referred to the UDN, of whom 601 (40%) were accepted for evaluation. Of the accepted patients, 192 (32%) had previously undergone exome sequencing. Symptoms were neurologic in 40% of the applicants, musculoskeletal in 10%, immunologic in 7%, gastrointestinal in 7%, and rheumatologic in 6%. Of the 382 patients who had a complete evaluation, 132 received a diagnosis, yielding a rate of diagnosis of 35%. A total of 15 diagnoses (11%) were made by clinical review alone, and 98 (74%) were made by exome or genome sequencing. Of the diagnoses, 21% led to recommendations regarding changes in therapy, 37% led to changes in diagnostic testing, and 36% led to variant-specific genetic counseling. We defined 31 new syndromes.

CONCLUSIONS

The UDN established a diagnosis in 132 of the 382 patients who had a complete evaluation, yielding a rate of diagnosis of 35%. (Funded by the National Institutes of Health Common Fund.)

The New York Times

The Week in Good News

Dee Reynolds has Niemann-Pick Type C, a genetic disease that went undetected until adulthood.
Julia Rendleman for The New York Times

When an illness is a mystery, patients turn to these detectives.

The Undiagnosed Diseases Network is the last resort of people with diseases or ailments that have confounded doctors and experts. Researchers at this federally funded project pursue every possible clue to discover what's wrong.

"Patients find it really valuable even just to give a name to the enemy," said Dr. Euan Ashley, a geneticist and co-director of the network.

Those who come away without a diagnosis or treatment are told that if the science improves and an answer emerges, the network will contact them.

"We never give up," Dr. Ashley said. *Read more »*

Most rare disease genome studies:

- Show a diagnosis rate for ES or GS of 30-50%
- Show that most findings are actionable (79% in UDN study)
- Show medical costs are reduced (by 94% in UDN study)

COST

ACCURACY

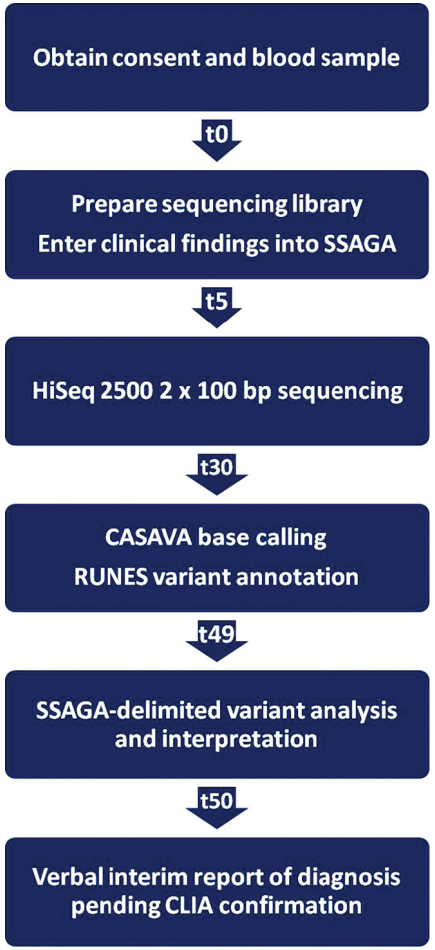
SPEED

IMPLEMENTATION &
INTEGRATION

Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units

Carol Jean Saunders,[1,2,3,4,5*] Neil Andrew Miller,[1,2,4*] Sarah Elizabeth Soden,[1,2,4*] Darrell Lee Dinwiddie,[1,2,3,4,5*] Aaron Noll,[1] Noor Abu Alnadi,[4] Nevene Andraws,[3] Melanie LeAnn Patterson,[1,3] Lisa Ann Krivohlavek,[1,3] Joel Fellis,[6] Sean Humphray,[6] Peter Saffrey,[6] Zoya Kingsbury,[6] Jacqueline Claire Weir,[6] Jason Betley,[6] Russell James Grocock,[6] Elliott Harrison Margulies,[6] Emily Gwendolyn Farrow,[1] Michael Artman,[2,4] Nicole Pauline Safina,[1,4] Joshua Erin Petrikin,[2,3] Kevin Peter Hall,[6] Stephen Francis Kingsmore[1,2,3,4,5†]

2012: 50 hours
2014: 42 hours
2015: 26 hours
2018: 19.5 hours



New GUINNESS WORLD RECORDS™ Title Set for Fastest Genetic Diagnosis

Innovations in whole genome sequencing speed answers and hope for newborns and children with rare, genetic diseases

Dr. Kingsmore receives the GUINNESS WORLD RECORDS™ certificate for the fastest genetic diagnosis.

San Diego—Feb. 12, 2018–Scientists at the Rady Children's Institute for Genomic Medicine (RCIGM) have compressed the time needed to decode rare genetic disorders in newborns through DNA sequencing to less than a day.

Through close collaboration with leading technology and data-science developers — Illumina, Alexion, Clinithink, Edico Genome, Fabric Genomics and Diploid—the RCIGM team has engineered a seamless process—enhanced by in-house expertise—to deliver analysis and interpretation of life-threatening genetic variations in just 19.5 hours.

The GUINNESS WORLD RECORDS™ title was achieved on Feb. 3, 2018. The previous speed record of 26 hours was set in 2015 by Stephen Kingsmore, M.D., D.Sc., president and CEO of Rady Children's Institute for Genomic Medicine. Dr. Kingsmore has pioneered the rapid turnaround and delivery of genetic test results to neonatal and pediatric intensive care (NICU/PICU) physicians.

Sci Tr Med 2012: 154: 1-13
Heart Rhythm 2014 11:1707-13
PNAS 113 (41) 11555-11560

CORRESPONDENCE

Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting

TO THE EDITOR: Rapid genetic diagnosis can guide clinical management, improve prognosis, and reduce costs in critically ill patients.[1,2] Although most critical care decisions must be made in hours, traditional testing requires weeks and rapid testing requires days. We have found that nanopore genome sequencing can accurately and rapidly provide genetic diagnoses. Our workflow combines streamlined preparation of commercial nanopore sequencing, distributed Cloud-based bioinformatics, and a custom variant-prioritization approach (Fig. 1).[3]

Between December 2020 and May 2021, at two hospitals in Stanford, California, we enrolled 12 patients who were generally representative of persons living in the United States with respect to race, ethnic group, and sex (Tables S1 and S2 in the Supplementary Appendix, available with the full text of this letter at NEJM.org). We obtained an initial genetic diagnosis in 5 of the patients (Table S3). The shortest time from arrival of the blood sample in the laboratory to the initial diagnosis was 7 hours 18 minutes.

After establishing a diagnosis in Patient 1, we updated our bioinformatics framework to permit the transfer of terabytes of raw signal data to Cloud storage in real time and distributed the data across multiple Cloud computing machines to achieve near real-time base calling and alignment, a step that reduced the postsequencing run time (base calling through alignment) by 93%, from 7 hours 21 minutes to 34 minutes (the average of postsequencing run times for Patients 2 to 12) (Table S5).

Flow cells were washed and reused until exhaustion to reduce the sequencing cost per sample. Libraries were bar-coded in Patients 1 through 7 to prevent carryover from one sample to the next. After processing the sample obtained from Patient 7, we benchmarked and adopted a bar-code–free method to rapidly generate genome sequences.[3] Removing the bar-coding process accelerated sample preparation by 37 minutes, to an average of 2.5 hours, and enabled us to load a greater amount of patients' DNA into each flow cell (333 ng vs. 155 ng) and increase pore occupancy (to 82% from 64%) (Figs. S1 and S2 and Table S4). Our sequencing workflow generated 173 to 236 Gb of data per genome using 48 flow cells, with an alignment identity of 94% (Fig. S3) and 46 to 64× autosomal coverage (i.e., each base of each autosome was represented in 46 to 64 sequence reads) (Fig. S4). Half the sequencing throughput was in reads that were 25 kb or longer (Table S6).

Small variants and structural variants were called after the reads were aligned to the GRCh37 human reference genome, which generated a median of 4,490,490 single-nucleotide variants and small insertions and deletions (indels).[4,5] Custom filtration and prioritization of variants with an ultrarapid scoring system (Fig. S5) substantially decreased the number of candidate variants for manual review to a median of 29 (range, 16 to 53) for small variants and 22 (range, 11 to 37) for structural variants (Table S2).

Each initial diagnosis was immediately reviewed by study and bedside physicians, and a consensus was reached as to whether the proposed variant represented the primary cause of the patient's presentation. Diagnostic variants were identified in 5 of the 12 patients, who ranged in age from 3 months to 57 years. The findings were immediately confirmed by a laboratory certified by the Clinical Laboratory Improvement Amendments (CLIA) process and informed clinical management (including sympathectomy, heart transplantation, screening, and changes in medication) for each of the 5 patients or their family members.

In one patient, a 3-month-old full-term infant who presented in status epilepticus, seizure semiology included right gaze deviation with bilateral upper-extremity clonic jerking and perioral myoclonic twitching. Interictal electroencephalography revealed abundant predominantly posterior

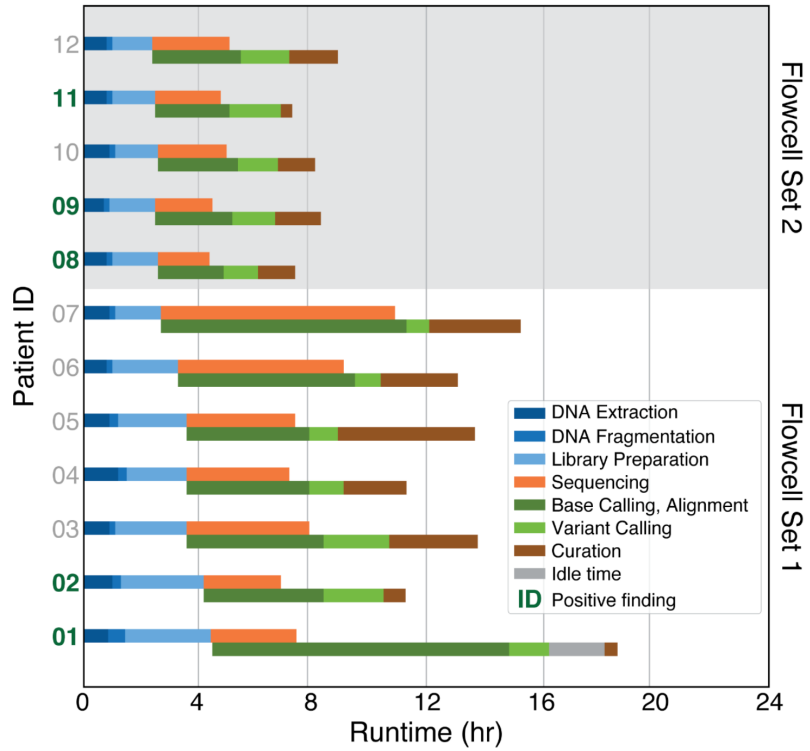
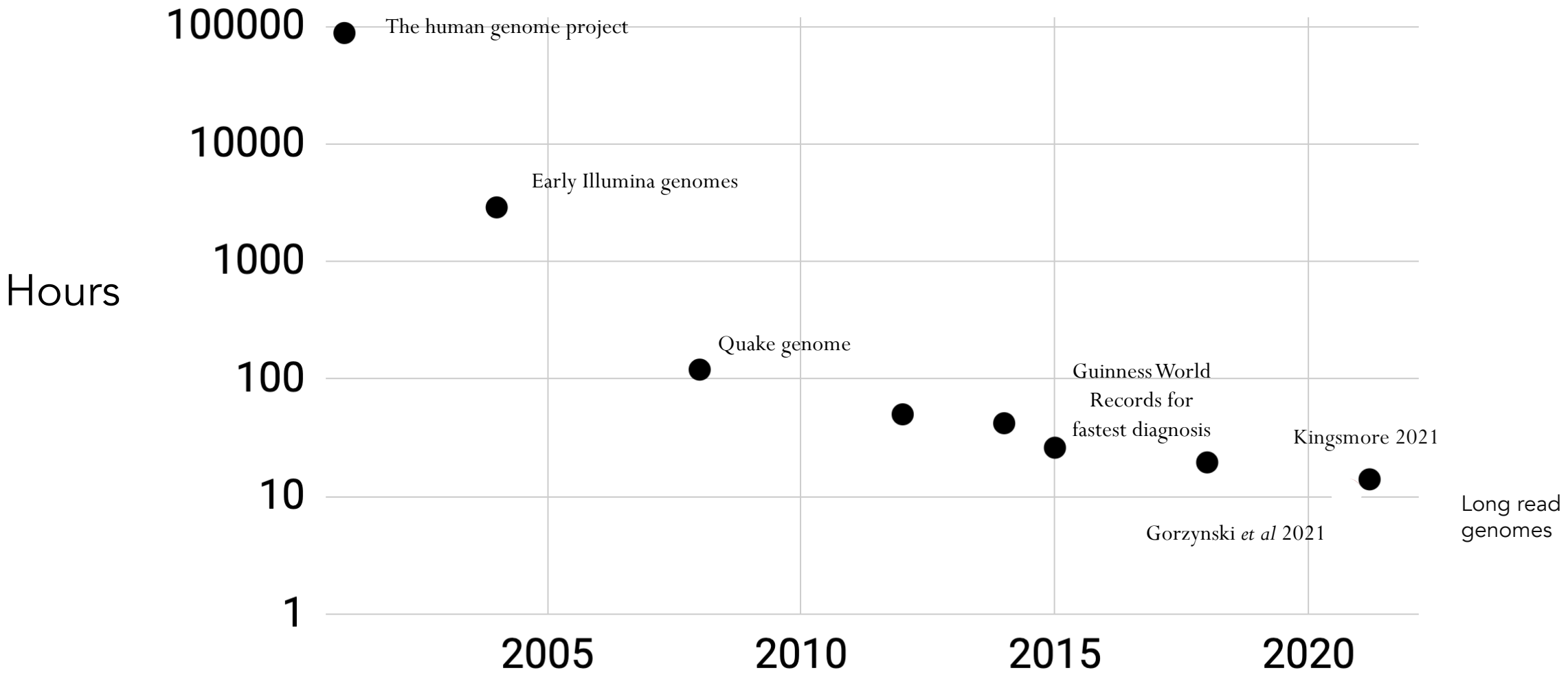


Figure 1. (a) The ultra-rapid whole genome sequencing pipeline. The schema depicts all processes from sample collection to a diagnosis. Vertically stacked processes are run in parallel. (b) The performance of the pipeline on twelve patients in two phases. Run-time of individual components are shown by corresponding color from panel (a). The fastest runtime was 7:18 hours (Patient 11) with a positive diagnosis.


Hours
100000
10000
1000
100
10
1
2005
2010
2015
2020
The human genome project
Early Illumina genomes
Quake genome
Guinness World Records for fastest diagnosis
Gorzynski et al 2021
Kingsmore 2021
Long read genomes

Search Inventory Search

MASERATI

Ferrari Silicon Valley

2750 El Camino Real Redwood City, CA 94061

Sales: (888) 378-7586
Service: (888) 377-1063
Parts & Accessories: (888) 430-4670
Body Shop: (866) 981-0953

Inventory | Service | Parts & Accessories | Body Shop | News & Events | Racing Team | About Us

2013 Ferrari 458 Spider

Ferrari Videos

Share

~~Top speed: 199mph~~

New top speed: 32,000,000 mph

New comparator vehicle required

COST

ACCURACY

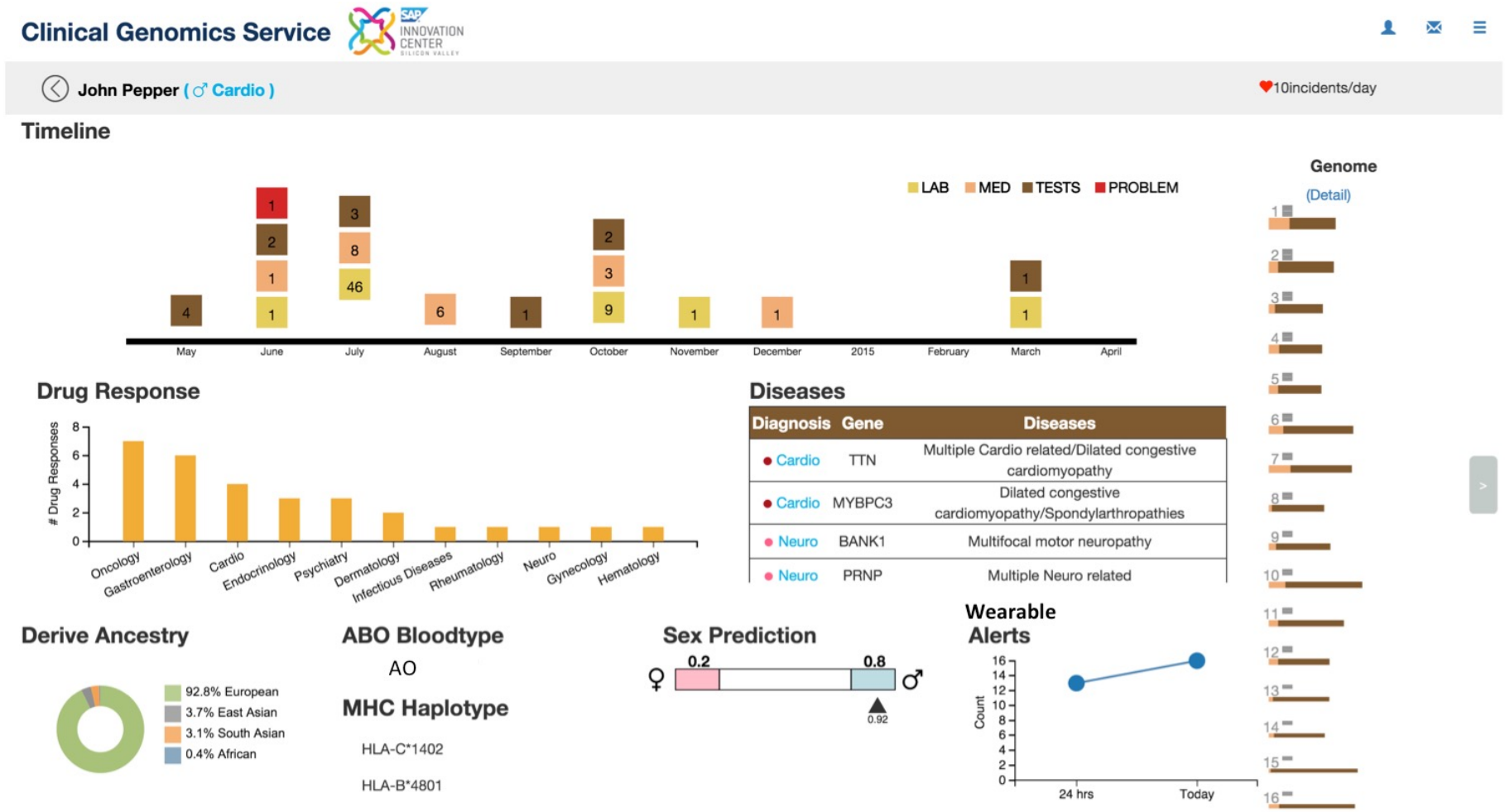
SPEED

IMPLEMENTATION &
INTEGRATION

The Next Generation Precision Medical Record - A Framework for Integrating Genomes and Wearable Sensors with Medical Records

https://www.biorxiv.org/content/early/2016/02/29/039651

The Next Generation Precision Medical Record - A Framework for Integrating Genomes and Wearable Sensors with Medical Records



https://www.biorxiv.org/content/early/2016/02/29/039651

genomeweb

Epic EMR Users Begin Integrating Genomics Data for Clinical Decision Support

Feb 25, 2020 | Neil Versel

"We're trying to help liberate genomics from the PDF report"

Alan Hutchison, Epic's VP of population health, noted that Epic has a series of Health Level 7 International (HL7) and Fast Healthcare Interoperability Resources (FHIR) application programming interfaces, as well as a set of specifications and guidelines for exchanging genetic testing orders and results. The firm has also created what it has dubbed the Epic Specialty Diagnostics Suite of applications specifically for labs.

Major genomics labs that already use or are in the process of installing Epic's Specialty Diagnostics Suite include Exact Science, Foundation Medicine, Natera, Tempus, Guardant Health, and Caris Life Sciences, according to Hutchison. All will have access to the new functionality Epic is building with Myriad by next year.

Guardant is also currently working on an integration of its cancer tests into the Epic platform, which the company said should be complete this quarter.

Stanford Medicine launches in-house service for whole genome sequencing

Stanford Medicine now provides a service that harnesses the power of genome sequencing to identify the source of diseases and help target treatments.

Amy Hower with her husband and children. Hower turned to Stanford Health Care to find out the genetic cause of her heart condition. “The results I got were life-changing for me,” she said. *Courtesy of Amy Hower*

February 2021

Whole genome backbone for all genetic tests

- 40-45x short read coverage
- *in silico* exome
- *in silico* cardiovascular/renal panels
- higher accuracy
- improved SV detection

+ Polygenic risk scores to panel testing

- Cardiovascular (CAD -> DM, LVH, EDV, AF)
- Germline cancer panels (prostate, breast/ovarian)

+ Pharmacogenomics

- Pain (PPI, anti-emetics)
- Depression
- cardiovascular (statin, P2Y12)

Headwinds and Challenges

Diagnostics remains grossly under-valued compared to therapeutics.

It remains unclear who should pay for diagnostics

- health care systems
- payors incl CMS
- pharma
- self

There remains very little integration of genomic data with health care data in delivery

Rx	Market cap (billions $)		Dx	Market cap (billions $)
Johnson & Johnson	422		Labcorp	19
Eli Lilly	309		Quest	14
Pfizer	235		Exact	6
Abbvie	243		Guardant	5
Merck	229		Natera	4
Abbott	177		Invitae	0.5
BMS	147		Sema4/GeneDx	0.4

Data as of 10/10/2022

COST

ACCURACY

SPEED

IMPLEMENTATION &
INTEGRATION

A call for an integrated approach to improve efficiency, equity and sustainability in rare disease research in the United States

To build a more efficient, equitable and sustainable approach to rare disease research in the United States, we must prioritize integrated research infrastructure and approaches that focus on understanding connections across rare diseases.

Meghan C. Halley, Hadley Stevens Smith, Euan A. Ashley, Aaron J. Goldenberg and Holly K. Tabor

The cumulative burden of rare diseases is immense, with the over 7,000 identified rare diseases together affecting an estimated 25–30 million Americans[1]. Most have serious impacts on individuals' physical and/or cognitive functioning, and many are life-threatening or fatal. Healthcare spending in rare diseases reached nearly US$966 billion in 2019, well exceeding the amount spent for some of the most common chronic diseases[1]. Despite this tremendous physical, psychological and economic burden, over 90% of rare diseases lack an approved therapy[1].

At the same time, recent advances in genomic sequencing, molecular biology and machine learning suggest that notable progress for the rare disease community could be on the horizon. As an estimated 80% of rare diseases are genetic in etiology[1], advances in genomic sequencing are helping to determine the genetic bases and specific biological mechanisms underlying many of these conditions. Advances in molecular biology have further opened up the possibilities for new types of therapy (such as antisense oligonucleotides, gene and cell therapies), even for ultra-rare diseases[2]. When these developments are paired with new machine learning approaches for identifying patterns in large volumes of data, the possibilities for diagnostic and therapeutic advances increase dramatically[3]. For the first time, we can imagine a world in which many more rare diseases have effective treatments, or even cures.

However, to ensure that rare disease patients benefit from these advances, we need to critically examine our current approach to rare disease research in the United States. Although many of the challenges faced in the United States apply, at least in part, internationally, the diversity of healthcare systems in other countries contributes specificities that warrant their own investigation beyond the scope of this paper. Here we examine the challenges of efficiency, equity and sustainability in our current approach to rare disease research in the United States, which together limit our ability to ensure that scientific innovations benefit all patients with rare diseases. To address these challenges, we suggest that decision-makers prioritize integrated research infrastructure and approaches that focus on understanding connections within and across rare diseases as a basis for a more equitable, evidence-based allocation of research resources.

The challenge of efficiency in rare disease research

Although the excitement surrounding the potential for advances in rare disease research is certainly warranted, in the United States this research largely remains siloed around individual rare diseases. Although efforts such as the Rare Disease Clinical Research Network (RDCRN) do provide some coordination across diseases, the RDCRN still includes only 2% of the estimated 7,000 rare diseases[4], and data silos built around single or small groups of rare diseases limit our understanding of connections across rare diseases[5]. Research on a single rare disease will always have an important place in science and translational medicine. However, overreliance on this approach limits our ability to identify more efficient strategies to advance rare disease diagnostics, therapeutic agents, outcomes measurement, epidemiology, public health and health services research that may benefit more than one rare disease, or even all rare diseases.

In a single-disease approach, we cannot easily assess the relative value of new diagnostic technologies for subsets of the large and diverse rare disease community. For example, we cannot easily assess the costs and benefits of increased clinical availability of genomic sequencing on health outcomes or quality of life for rare disease patients with diverse phenotypic characteristics[6]. Although some such analyses have been possible in large single-payer healthcare systems in other countries (for example, ref. [7]), in the United States they have been limited to circumscribed contexts, such as the neonatal intensive care unit (for example, ref. [8]), specific subgroups of rare diseases (for example, ref. [9]) and/or evaluation based on intermediate outcomes such as diagnostic yield (for example, ref. [10]).

In therapeutics, several rare diseases can involve the same biological pathways, and either a downstream or an upstream modulator could potentially treat multiple different conditions. For example, a recent umbrella trial tested a monoclonal antibody in three different rare diseases all characterized by overproduction of the same cytokine protein[11]. Although the frequency of such shared pathways remains unknown, a single-disease approach does not facilitate the data sharing necessary to easily identify and leverage such connections. With a single-disease approach, it may take years — or even decades — for drugs deemed safe and effective in one rare disease to be repurposed for another.

Our current approach also misses the opportunity to increase efficiency through coordination of outcomes development and testing across rare diseases with similar phenotypes, etiologies or trajectories, which could dramatically accelerate the pace of drug approvals. As a result, researchers utilize extensive resources developing outcomes that are only applicable in a single rare (or even ultra-rare) disease. For example, Luxturna, the first gene therapy for a monogenic disease approved by the US Food and Drug Administration (FDA), was designed to treat an ultra-rare subtype of inherited retinal disease. As part of the drug development process, scientists developed a

Supporting undiagnosed participants when clinical genomics studies end

Many large research initiatives have cumulatively enrolled thousands of patients with a range of complex medical issues but no clear genetic etiology. However, it is unclear how researchers, institutions and funders should manage the data and relationships with those participants who remain undiagnosed when these studies end. In this Comment, we outline the current literature relevant to post-study obligations in clinical genomics research and discuss the application of current guidelines to research with undiagnosed participants.

Meghan C. Halley, Euan A. Ashley and Holly K. Tabor

The past 15 years have witnessed the rapid expansion of clinical genomics research focused on applying advances in genomic sequencing to the diagnosis of patients afflicted by rare genetic diseases. These 'genomic diagnosis' studies include many large research efforts, such as the Centers for Mendelian Genomics, the Undiagnosed Diseases Network, and multiple sites of the Clinical Sequencing Evidence-Generating Research consortium in the United States[1–3]. International efforts include components of Solve-RD in the European Union, the 100,000 Genomes Pilot on Rare-Disease Diagnosis in the United Kingdom, and sites affiliated with Undiagnosed Diseases Network International, among others worldwide[4–6]. These studies exist at the boundary between research and clinical care, pursuing the scientific aim of understanding the contribution of genetics to human disease through the provision of genomic diagnoses for patients.

Measured against the twin goals of advancing research and providing individual diagnoses, these studies have been very successful, identifying diagnoses for approximately 30% of enrolled participants and culminating in the publication of hundreds of papers that advance knowledge of the genomic underpinnings of rare diseases[5,7]. However, as a number of these studies near their end, many questions remain unanswered about how researchers, institutions and funders should manage the data and relationships with the thousands of participants who remain undiagnosed. For example, if a variant of unknown significance is recognized as pathogenic next year or 5 years from now, will participants or their families be notified? If new initiatives are launched to evaluate new diagnostic technologies (for example, the GREGoR Consortium; https://gregorconsortium.org/), will participants from other genomic diagnosis studies be eligible to enroll? Will prior participation in genomic diagnosis research limit access to subsequent diagnostic evaluation in the clinical setting, due to either lack of payer coverage or a provider's perception that their patient's diagnostic needs had been met through research? Without answers to these and other questions, participants enrolled in genomic diagnosis research are at risk of being marooned by the biomedical research establishment when these studies end, left without a diagnosis and, potentially, with no way to pursue one.

Why would post-study obligations apply?

The ethical guidelines for research in the United States are based on the core principles laid out in the Belmont Report, including respect for persons, beneficence and justice[8]. Although the Belmont Report does not explicitly address the question of post-study obligations, its broad principles remain relevant. For example, in adhering to the principle of respect for persons, researchers have the obligation to treat their research participants as whole persons, not simply as a means for achieving research gains. Thus, if researchers are working with participants with unmet or ongoing medical needs — particularly needs that fall within the scope of the original study — the researchers, their institutions and funders together have an obligation to facilitate a pathway to address these needs even when studies end[9]. Researchers also are obligated to maximize benefits and minimize harm to participants, including ensuring that participation in research — or the ending of participation in this research — does not inadvertently harm participants in the post-study period[10]. Furthermore, the principle of justice asserts that the benefits and burdens of research must be equitably distributed. Thus, post-study obligations are particularly critical for protecting those participants who already face challenges to healthcare access, as they may be additionally dependent on research as a source of healthcare, and may be additionally vulnerable to harms due to study closure[11].

The Declaration of Helsinki, which provides the most widely accepted guidance on ethical research internationally, directly addresses the question of post-study obligations, stating that "sponsors, researchers and host country governments should make provisions for post-trial access for all participants who still need an intervention identified as beneficial in the trial."[12] Although concern for post-study obligations originally emerged in the context of therapeutic clinical trials, these obligations are now recognized as extending beyond post-trial access to experimental treatments[10]. However, the Declaration of Helsinki provides only this limited directive, and more-detailed guidance developed by groups such as the Multi-Regional Clinical Trials Center (https://mrctcenter.org/) remains firmly anchored in the language of therapeutic clinical trials and thus requires adaptation to the context of genomic diagnosis research. For example, in genomic diagnosis research, the 'intervention' typically includes a range of advanced testing that is tailored to individual participants' diagnostic needs, rather than standardized treatment protocols. Furthermore, the 'benefits' of a diagnosis are likely to vary widely and unpredictably depending on individual participant characteristics, including a range of personal benefits to both the participant and their family (for example, through reproductive decision-making) and may accrue many years down the road[13]. The question of heterogeneity of potential benefits of diagnosis within the undiagnosed community is one that deserves further examination and is beyond the scope of this Comment. However, given that at

Rare disease

Molecular classification of known disease

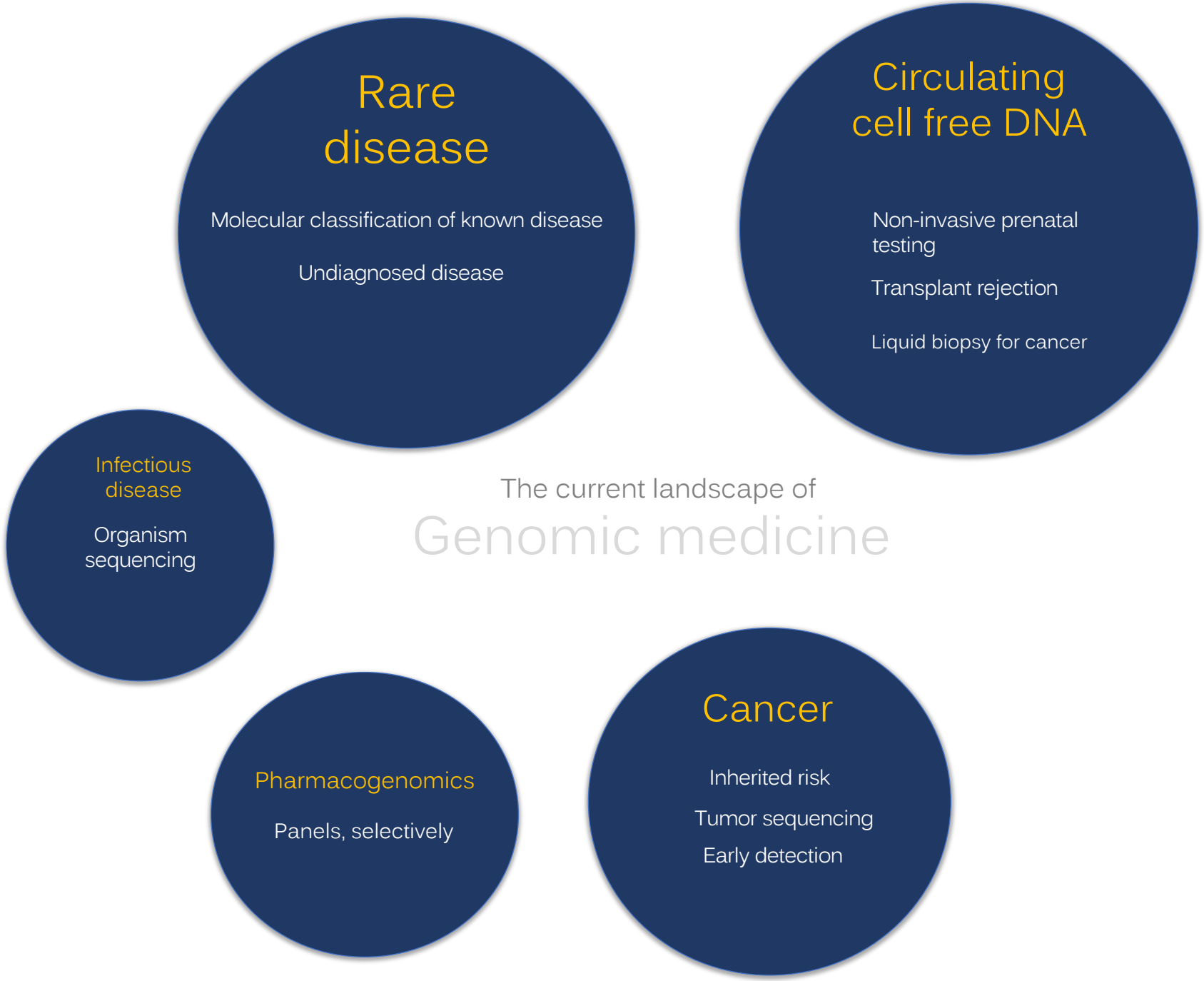
Undiagnosed disease

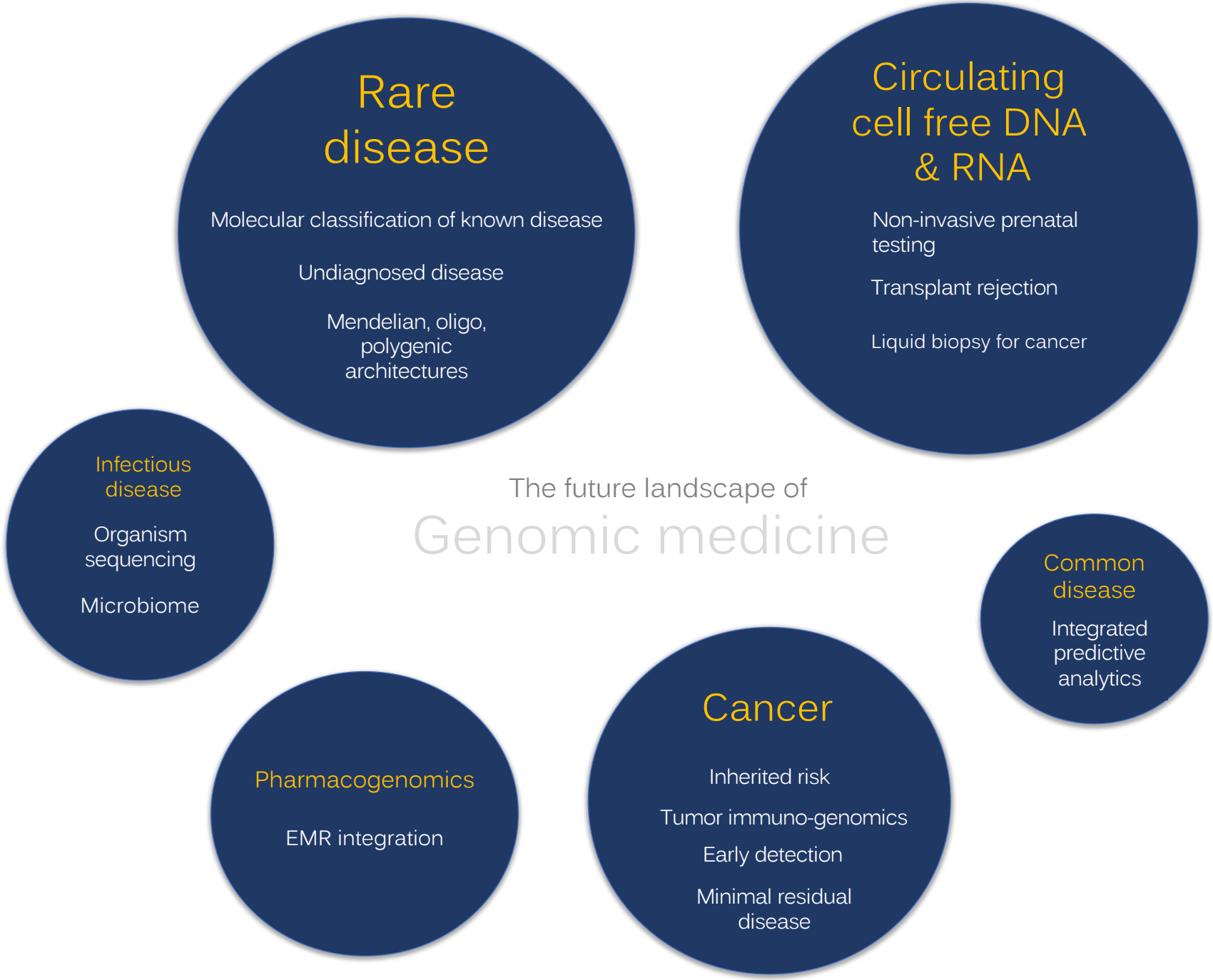
The historical landscape of

Genomic medicine

Cancer

Inherited risk

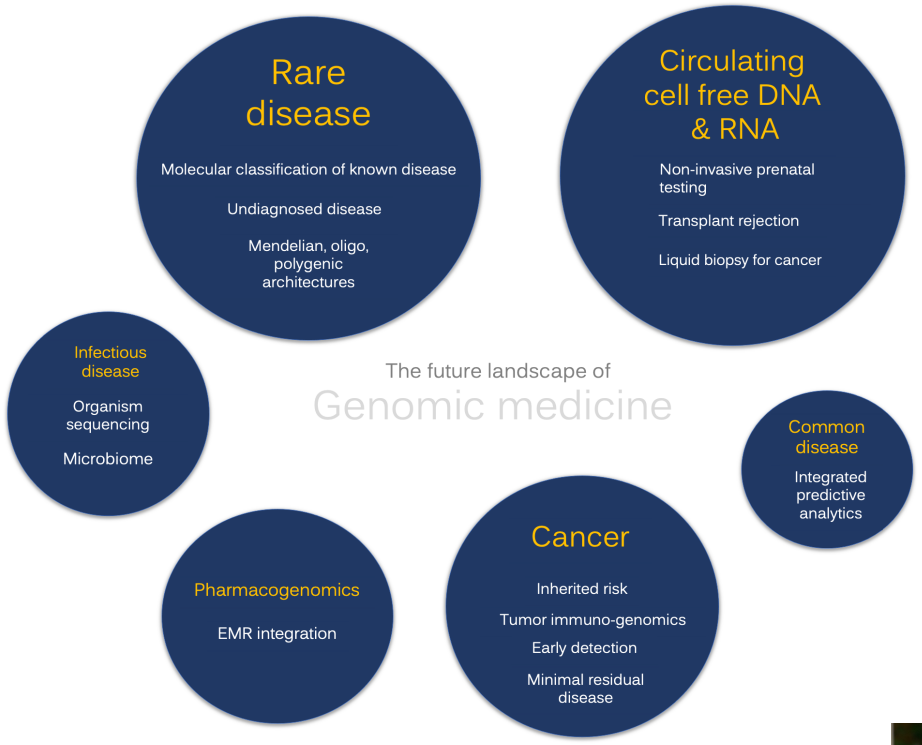

Rare disease
Molecular classification of known disease
Undiagnosed disease
Circulating cell free DNA
Non-invasive prenatal testing
Transplant rejection
Liquid biopsy for cancer
Infectious disease
Organism sequencing
The current landscape of
Genomic medicine
Pharmacogenomics
Panels, selectively
Cancer
Inherited risk
Tumor sequencing
Early detection

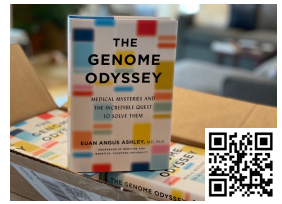

Rare disease
Molecular classification of known disease
Undiagnosed disease
Mendelian, oligo, polygenic architectures
Circulating cell free DNA & RNA
Non-invasive prenatal testing
Transplant rejection
Liquid biopsy for cancer
Infectious disease
Organism sequencing
Microbiome
The future landscape of
Genomic medicine
Common disease
Integrated predictive analytics
Pharmacogenomics
EMR integration
Cancer
Inherited risk
Tumor immuno-genomics
Early detection
Minimal residual disease

Conclusions

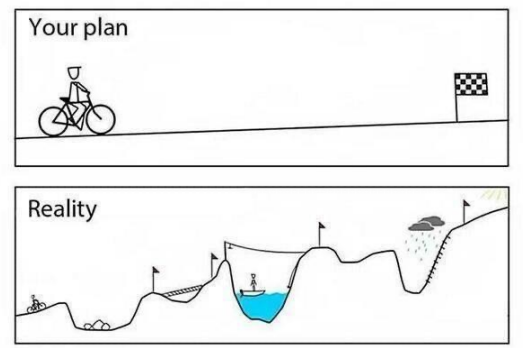
- We have come a long way in a short time
- The genome: a $3bn multinational, 10-year effort can now be completed for single individuals for $100 in a few hours
- Accuracy has improved dramatically
- Long read sequencing provides advantages in clinical genomics
- While today's major application is in rare and undiagnosed disease, near-term applications beyond this are expansive

- There is still work to do
- *Equity*: a genome for everyone who wants one
- *Diversity*: population cohorts must represent genetic diversity
- *Inclusion*: patients and families should be at the center of all our planning and decision-making
- Efforts should be focused on accelerating integration with health care processes beyond lab billing but actually at the point of care
- It is long past time that payors realize the patient benefits and cost savings of implementing genomics





genomebook.info


Your plan
Reality