frontiers in Genetics

BRIEF RESEARCH REPORT
published: 01 December 2020
doi: 10.3389/fgene.2020.612515

A Distributed Whole Genome Sequencing Benchmark Study

Richard D. Corbett[1], Robert Eveleigh[2], Joe Whitney[3], Namrata Barai[3], Mathieu Bourgey[2], Eric Chuah[1], Joanne Johnson[1], Richard A. Moore[1], Neda Moradin[3], Karen L. Mungall[1], Sergio Pereira[3], Miriam S. Reuter[4], Bhooma Thiruvahindrapuram[3], Richard F. Wintle[3], Jiannis Ragoussis[2], Lisa J. Strug[3], Jo-Anne Herbrick[3], Naveed Aziz[4], Steven J. M. Jones[1], Mark Lathrop[2], Stephen W. Scherer[3], Alfredo Staffa[2] and Andrew J. Mungall[1]**

[1] Canada's Michael Smith Genome Sciences Centre, BC Cancer Research Institute, Provincial Health Services Authority, Vancouver, BC, Canada, [2] McGill Genome Centre, McGill University, Montreal, QC, Canada, [3] The Centre for Applied Genomics, The Hospital for Sick Children and University of Toronto, Toronto, ON, Canada, [4] Canada's Genomics Enterprise (CGEn), The Hospital for Sick Children, Toronto, ON, Canada

Population sequencing often requires collaboration across a distributed network of sequencing centers for the timely processing of thousands of samples. In such massive efforts, it is important that participating scientists can be confident that the accuracy of the sequence data produced is not affected by which center generates the data. A study was conducted across three established sequencing centers, located in Montreal, Toronto, and Vancouver, constituting Canada's Genomics Enterprise (www.cgen.ca). Whole genome sequencing was performed at each center, on three genomic DNA replicates from three well-characterized cell lines. Secondary analysis pipelines employed by each site were applied to sequence data from each of the sites, resulting in three datasets for each of four variables (cell line, replicate, sequencing center, and analysis pipeline), for a total of 81 datasets. These datasets were each assessed according to multiple quality metrics including concordance with benchmark variant truth sets to assess consistent quality across all three conditions for each variable. Three-way concordance analysis of variants across conditions for each variable was performed. Our results showed that the variant concordance between datasets differing only by sequencing center was similar to the concordance for datasets differing only by replicate, using the same analysis pipeline. We also showed that the statistically significant differences between datasets result from the analysis pipeline used, which can be unified and updated as new approaches become available. We conclude that genome sequencing projects can rely on the quality and reproducibility of aggregate data generated across a network of distributed sites.

Keywords: whole genome sequencing, genome, benchmark, informatics, comparison, variant

OPEN ACCESS

Edited by:
Youri I. Pavlov, University of Nebraska Medical Center, United States

Reviewed by:
Igor B. Rogozin, National Institutes of Health (NIH), United States
Elena Stepchenkova, Vavilov Institute of General Genetics, Russian Academy of Sciences, Russia

***Correspondence:**
Stephen W. Scherer
Stephen.Scherer@sickkids.ca
Andrew J. Mungall
amungall@bcgsc.ca

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: *30 September 2020*
Accepted: *10 November 2020*
Published: *01 December 2020*

Citation:
Corbett RD, Eveleigh R, Whitney J, Barai N, Bourgey M, Chuah E, Johnson J, Moore RA, Moradin N, Mungall KL, Pereira S, Reuter MS, Thiruvahindrapuram B, Wintle RF, Ragoussis J, Strug LJ, Herbrick J-A, Aziz N, Jones SJM, Lathrop M, Scherer SW, Staffa A and Mungall AJ (2020) A Distributed Whole Genome Sequencing Benchmark Study. Front. Genet. 11:612515. doi: 10.3389/fgene.2020.612515

INTRODUCTION

The global sequencing market is valued at approximately $10 billion[1]. To date, more than 500,000 human genomes have been sequenced[2] and deposited in public databases as part of previous large-scale genome projects (Auton et al., 2015; Turro et al., 2020), personal genome projects (Beck et al., 2018; Reuter et al., 2018; Jeon et al., 2020) or sizeable aggregation projects across larger populations (Karczewski et al., 2019). The genomes of another two million individuals are expected to be sequenced under current projects[3,4]. To date, such data have been used to increase understanding of the underlying genetic architecture in disease (Yuen et al., 2017; Bailey et al., 2018; Priestley et al., 2019; Pleasance et al., 2020; Trost et al., 2020) and are increasingly being used in clinical genetics settings (Stavropoulos et al., 2016; Lionel et al., 2018).

For large-scale projects where expansive data are to be collected across populations, the resources of many institutions may be pooled to meet sequencing capacity demands, as well as to satisfy possible jurisdictional requirements, ethno-cultural and anthropological considerations (Knoppers et al., 2014), as well as ethical or legal restrictions on sample transfer (Mascalzoni et al., 2015), or requirements for grant funds to be spent locally. As genome sequences become increasingly used as the foundational biological reference point for national precision medicine initiatives, multi-site participation will only increase (Stark et al., 2019). In such projects, it is important to identify and quantify any differences in results that may arise due to different methodological and analytical procedures used across sites. While there are methods to evaluate and correct for batch effects once data have been generated (Tom et al., 2017; Baskurt et al., 2020) for whole genome sequencing projects, genetic variants for example cannot be reproducibly called if the appropriate reads are not sampled on a given sequencing instrument. Therefore, generation of consistently comparable data is preferred.

To facilitate the evaluation of whole genome assays, the genome in a bottle (GIAB) consortium combines sequence data from multiple centers along with results from several variant calling algorithms to provide consensus variant calls and importantly, regions of confident genotyping for each of the model samples (Zook et al., 2016). The consortium enables sequencing centers to routinely assess the precision and sensitivity of single nucleotide variants (SNVs) and insertion and deletions (Indels) detected in their analyses by sequencing GIAB reference samples (Cleary et al., 2015).

In order to prepare to support national genome sequencing initiatives of the highest quality for sharing in open-science databases (Rahimzadeh et al., 2016) three GIAB reference cell lines were sequenced in triplicate at each of Canada's Michael Smith Genome Sciences Centre at BC Cancer in Vancouver, The Centre for Applied Genomics at The Hospital for Sick Children in Toronto, and the McGill Genome Centre in Montreal. Importantly, all processes were performed using current best practice approaches as determined at each center, to allow us to accurately assess differences observed under production conditions. Assessing the results for the 81 openly accessible whole genome data sets generated from combinations of samples, replicates, sequencing center, and analysis center allowed us to rank the variables in order of the associated variability in results. Our results inform our own, and any other multi-site projects, on how to collectively yield the most accurate genome sequence data and genetic variant calls.

[1]https://www.grandviewresearch.com/industry-analysis/next-generation-sequencing-market

[2]https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet/#f5df45c386d2

[3]https://ec.europa.eu/digital-single-market/en/news/germany-joins-1million-genomes-initiative

[4]https://www.nih.gov/news-events/news-releases/nih-funds-new-all-us-research-program-genome-center-test-advanced-sequencing-tools

METHODS

Each of the three sequencing centers used the Illumina HiSeq X technology to generate short-read genome sequence data of at least 30X coverage, using DNA from three GIAB reference cell lines (see below). These resulting 27 datasets were then processed through the bioinformatic pipelines in use at each center to create 81 datasets defined by four variables: unique cell line, replicate, sequencing center, and analysis pipeline. **Figure 1** provides an overview of the combinatorial study design, which leveraged the benchmark data provided by the GIAB consortium (Zook et al., 2016). The genome sequence data are submitted to the NCBI SRA database under accession SRP278908.

Samples

The samples used were from the National Institute of Standards and Technology (NIST) reference material 8392. These are further described as "Human DNA for Whole-Genome Variant Assessment (Family Trio of Eastern Europe Ashkenazi Jewish Ancestry) (HG002, HG003, HG004)" (Zook et al., 2016). DNAs were obtained from large homogenized growths of B lymphoblastoid cell lines from the Human Genetic Cell Repository at Coriell Institute for Medical Research. To eliminate any potential variability from differences in DNA preparation between sites, the samples sequenced at each site were aliquots from the same primary preparation.

PCR-Free Whole Genome Sequencing

Each center performed DNA quality control (QC), library construction, and sequencing steps following their own standard procedures (summarized in **Table 1**), some of which are the same, with other components being different. Of note, two different PCR-free library preparation kits and DNA input amounts were employed, and target insert sizes and input starting amounts of DNA also differed across centers. As indicated, all sequencing was performed on Illumina HiSeq X instruments. While 1% PhiX spike-in was used in both Montreal and Toronto, Vancouver used its standard method of including a plasmid-based sample tracking spike-in.

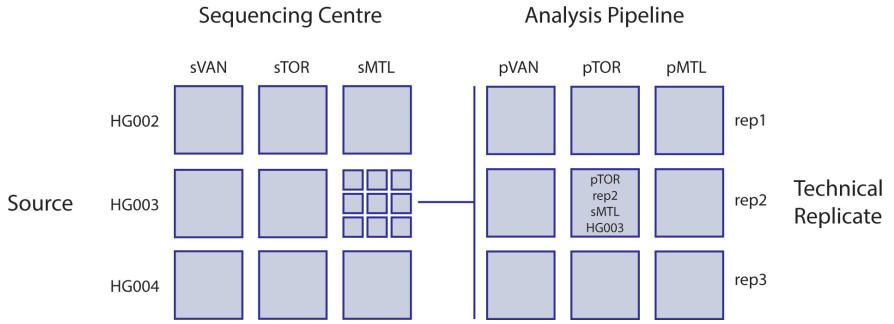

FIGURE 1 | Study design. Sequence data for three different reference DNA sample sources (the Ashkenazim trio; son HG002, father HG003, and mother HG004) were generated in triplicate at each center then analyzed by each center for a total of 81 total combinations of sequencing center, DNA Source, analysis pipeline, and technical replicate. Sequencing centers (s) and analysis pipelines (p): VAN, Vancouver; MTL, Montreal; TOR, Toronto.

TABLE 1 | Laboratory methods as performed at each of the three centers.

		Sequencing center (s)		
		sVAN	**sMTL**	**sTOR**
gDNA QC assays	Integrity	Agarose gel	Agarose gel or Tapestation	Agarose gel or Tapestation
	Quantification	Qubit or Quant-it DNA HS assays	Qubit DNA HS assay	Qubit DNA HS assay
	Purity	Not applicable	A260/280 between 1.8 and 2.0	A260/280 between 1.8 and 2.0
WGS library construction	PCR-free library prep kit	NEB paired-end sample prep Premix kit	Illumina TruSeq PCR-free library prep kit	Illumina TruSeq PCR-free library prep kit
	Input DNA amount (ng)	500	500	700
	DNA fragmentation	Covaris LE220	Covaris LE220	Covaris LE220
	Target size range (bp)	300–400	300	400
	Size selection	PCRClean DX (Aline Biosciences)	Ampure beads (Beckman Coulter)	Ampure beads (Beckman Coulter)
Library QC	Library validation (sizing)	Agilent Bioanalyzer DNA high sensitivity assay	Agilent Bioanalyzer DNA high sensitivity assay	Agilent Bioanalyzer DNA high sensitivity assay
	Library validation (quantification)	KAPA qPCR library quant kit	KAPA qPCR library quant kit	KAPA qPCR library quant kit
Sequencing	Sequencer (reads)	HiSeq X (2 × 150)	HiSeq X (2 × 150)	HiSeq X (2 × 150)
	Genomes (library) per lane	1	1	1
	Spike-in controls	Tracking plasmid (Moore et al., 2020)	1% PhiX	1% PhiX

sVAN, Vancouver; sMTL, Montreal; sTOR, Toronto.

Genetic Variant Calling and Informatics

Analysis methods for germline, PCR-free genomes that were performed at each center are reported in **Table 2**. All analyses were performed against each center's chosen human genome reference assembly based on NCBI's Genome Reference Consortium human build 37 (GRCh37), each performing alignments using BWA mem (Li, 2013). Of note, two centers (Montreal and Toronto) employed GATK 3.7+ and associated best-practice workflows (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013) while Vancouver used Strelka 2 (Kim et al., 2018) without any explicit steps for base recalibration or Indel realignment.

To assess differences observed in the data in advance of variant calling, aligned reads from each pipeline were processed with Picard[5], Qualimap (García-Alcalde et al., 2012), and SAMtools (Li et al., 2009) to identify quality differences. Variant calling results were assessed using version 3.6.2 of RTGTools vcfeval (Cleary et al., 2015) using release 3.3.2 of the GIAB references for HG002, HG003, and HG004.

[5]http://broadinstitute.github.io/picard

RESULTS

Assessing each of the 81 BAM files for quality, we detected some notable differences in the data yielded by each center. The average read coverage across the 81 datasets was 36.5X, with both the lowest (30.9X) and highest (42.1X) coverage for a single lane of data coming from Montreal's sequencing pipeline. Mean insert sizes were consistently lower in the data from Montreal, whose data had both more AT dropout and less GC dropout than data from the other

TABLE 2 | Informatics tools and settings employed at each center.

	Sequencing center analysis pipeline (p)		
	pVAN	**pMTL**	**pTOR**
Reference	Hg19a[6]	hs37d5[7]	hs37d5[7]
Read trimming	Custom trimmed to 150 bp	Skewer 0.2.2	Not applicable
Alignment	BWA mem 0.7.6a-M	BWA mem 0.7.12	BWA mem 0.7.12
BAM sorting	Sambamba 0.5.5	Sambamba 0.6.6	Picard 2.5.0 SortSam
BAM duplicate marking	Sambamba 0.5.5	Sambamba 0.6.6	Picard 2.5.0 MarkDuplicates
BAM calibration	Not applicable	GATK 3.8 BQSR and IR	GATK 3.7.0 BQSR and IR
Variant calling	Strelka 2.9.2	GATK 3.8 HaplotypeCaller	GATK 3.7.0 HaplotypeCaller
Variant filtering	Not applicable	Not applicable	GATK 3.7.0 VQSR

pVAN, Vancouver; pMTL, Montreal; pTOR, Toronto. [6]*https://www.bcgsc.ca/downloads/genomes/9606/hg19/1000genomes/bwa_ind/genome/README.GRCh37-lite*
[7]*http://www.imsbio.co.jp/RGM/R_rdfile?f=BSgenome.Hsapiens.1000genomes.hs37d5/man/package.Rd&d=R_BC*

centers (**Supplementary Figure 1**). While non-uniformity of read depth likely has little impact on the identification of SNVs, it can have notable effects on the sensitivity and specificity of CNV detection from whole genome sequence data, particularly when using read-depth based methods (Trost et al., 2019).

Concordance Against Benchmark Data

The corresponding variant calls for each of the 81 BAM sequence files were compared to available benchmark data, and across data sets. Although there were significant differences in the raw data metrics, the primary focus for this project was the final concordance of resulting variant calls. When comparing results to the available benchmark data, final VCF files from all combinations of unique cell line, replicate, sequencing center, and analysis pipeline yielded sensitivity measures above 98.9% and precision values above 99.5%. **Supplementary Figure 2** shows the full set of 81 sensitivity, precision, and F1 (model accuracy) values. Overall, the analysis pipelines from Montreal and Toronto, both of which employ the GATK based pipeline, had consistently higher sensitivity and lower precision than Vancouver, which employs the Strelka2 based pipeline. There was also a higher variance in F1 scores at Montreal where the results for the second replicate of HG003 and HG004 yielded reduced sensitivity in comparison to the other sets while the precision remained high. The sequencing results that generated consistently lower sensitivity had mean genome coverage numbers of 31.6X and 30.9X, while all others from the same center had mean coverage of 37.3X or higher. The raw data for the low coverage samples also had the highest estimated base error rates of the samples from the same lab (**Supplementary Table 1**).

Intersection of Genetic Variant Calls

In addition to the need for production of high-quality genetic variant calls across a network of centers, equally important is that the variants called within each center must also be as consistent and reproducible as possible. SNVs and Indels generated for each of the three samples were intersected to assess the level of difference between pipeline configurations (**Supplementary Table 2**). To achieve this, we treated each of the samples independently, and for each we held two of the three remaining variables (sequencing center, analysis center, and replicate) constant to evaluate the amount of change observed when considering just one variable. For example, for sample HG002 sequenced in Montreal and analyzed at Toronto, three sets of results were produced (one for each of the three replicates). Those three sets of results were intersected to evaluate the level of discordance across sequencing replicates. This type of analysis was repeated 27 times to cover all the combinations of sample, sequencing center, and analysis center to generate a distribution of expected differences due exclusively to the replicates.

A summary of the intersection analysis is presented in **Figure 2**, where the fraction of variants from each three-way comparison that are common to all three sets was collected. The fraction of variant calls that were common across sequencing replicates (median = 93.6) was slightly higher than that for the sequencing center (median = 93.2), and significantly greater (Mann–Whitney-Wilcoxon test, p-value = 1.027e-15) than that for the analysis pipeline (median = 89.5). While investigating each variable, there were multiple datasets with higher discordance than the common distribution for the 27 data points contributing to each curve. In each case, the more variable results occurred when comparing variant calls that included the second technical replicates from Montreal, which had an average coverage near 30X while other datasets from the same center had closer to 39X coverage. Historically genome sequencing studies have used a threshold of 30X coverage although this has, in part, been driven by sequencing costs and target density when loading flowcells. As expected, our results confirm the benefit of deeper sequence coverage for accurate variant calling. As sequencing costs continue to decrease, in principle, we expect to generate higher average genome coverage and therefore, higher variant calling accuracy.

DISCUSSION

Our results indicate that performing whole genome sequencing, using the technology platform tested across multiple sites

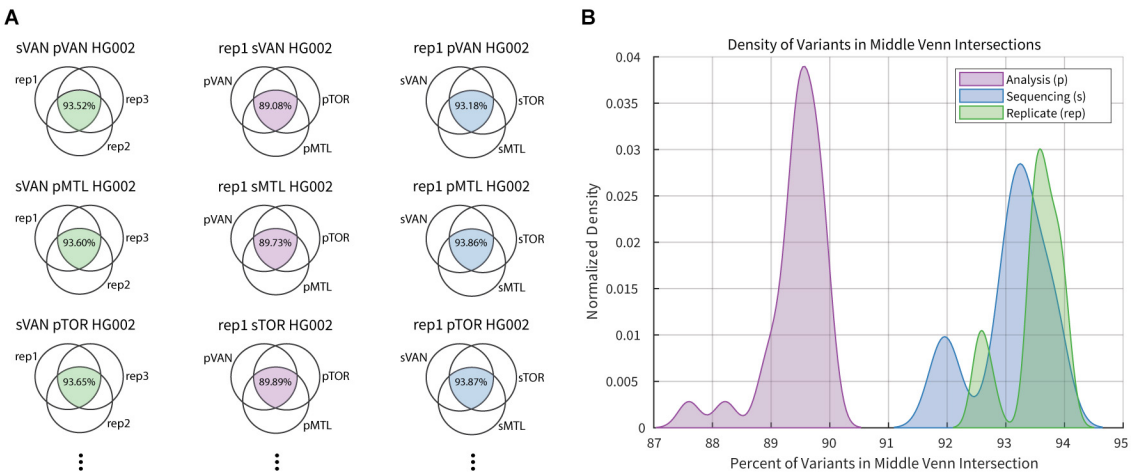

FIGURE 2 | Example of intersection analysis for select HG002 datasets. **(A)** Column 1 (green) Venn diagrams compare three replicate sequence datasets generated in Vancouver (sVAN) and analyzed by all three center's pipelines. Column 2 (purple) Venn diagrams compare the results from the three analysis pipelines for replicate 1 (rep1) of HG002 sequenced in Vancouver (sVAN), Montreal (sMTL) or Toronto (sTOR). Column 3 (blue) Venn diagrams compare results from the three sequencing centers for replicate 1 of HG002 analyzed by the three pipelines in Vancouver (pVAN), Montreal (pMTL) and Toronto (pTOR). In total there were 81 3-way intersections completed. **(B)** Percentage of variants common across sequencing replicates (green), analysis pipelines (purple) and sequencing centers (blue) are summarized in the density plots.

is an acceptable approach when trying to maximize sample size, for example, for large-scale population or disease studies. We presented a framework for testing multiple variables controlled by the sequencing centers, and found the most significant differences when different analysis pipelines were implemented. This underlines the robustness of the library preparation protocols, sequencing and imaging applied in this study, which minimizes experimental errors identified in short read sequencing (Robasky et al., 2014). We have made our data publicly available for additional testing.

Much of the evaluation was completed by comparing variant calls to benchmark data provided by the GIAB Consortium, where millions of true positive SNVs and Indels are known for each sample and, critically, large regions of confident non-variant positions allowing for the assessment of precision. However, it should be noted that these regions do not cover all classes of genetic variants, nor the entire genome, and studies such as this one cannot assess the precision or quality of variant calls within the missing regions. For example, version 3.3.2 of the available benchmark data for HG002 lists confident genotype information for 2.358 Gb of the genome but does not contain any information for variants on the X or Y chromosomes. Moreover, copy number and structural genetic variation datasets were not yet examined (Scherer et al., 2007). It is also important to consider the source of the DNA sample (Trost et al., 2019), which can influence the quality and amount of input DNA used for sequencing. Each of these factors may in turn impact all aspects of data generated, in particular when long-read technologies are used rather than the short-read sequencing presented here (Wang et al., 2019; Thibodeau et al., 2020).

There were two samples in the second replicate run, originating from one site, that yielded lower average coverage (31.6X and 30.9X) than its other samples, all of which had an average coverage greater than 37X. As expected, these particular samples had the lowest variant calling sensitivity suggesting that an average genome coverage nearing 30X may compromise sensitivity in germline studies, and higher coverage could be recommended. The average coverage numbers, however, do not explain all of the differences that are observed. In **Figure 2**, the low coverage samples cause the peaks at the lower end of each distribution while the larger distributions show that the choice of analysis pipeline can have a large impact on the consistency of variant results, as has been described by us and others (Craig et al., 2016; Chen et al., 2019; Kumaran et al., 2019). A consistent analysis pipeline is expected to improve across-center consistency by up to 5%, assuming that the variance among replicates represents the maximal reproducibility across datasets.

Since each of the three participating centers developed their pipelines largely independently (although there were some ongoing, cross-site projects sharing concepts), it was encouraging to find that overall genome variant calling results were both of high quality and consistent between sites. Our study did reveal minor differences in approaches, such as the selection of the version of the reference sequence used; two centers used an identical reference (hs37d5), but the third typically used hg19a (see section "Methods").

In summary, the employment of different standard analysis pipelines was thus determined as the main source of variation between datasets generated by the three centers. Fortunately, this aspect of the sequencing process can be easily controlled, either prospectively, or retrospectively. Major technology developments and operational guidelines for this purpose have been put forth in recent years, motivated precisely by reproducibility challenges in genomic data generation. Here, we add to this growing body of literature, arriving at recommendations for our own path forward, in which we suggest the three

centers implement containerization using Singularity (Kurtzer et al., 2017) and portable workflows using workflow definition language (WDL) (Voss et al., 2017) in each local high-performance computing facility. With these capabilities in place, our distributed sequencing network is poised to generate consistent, high-quality, whole-genome datasets for national, as well as international-scale, projects.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, SRP278908.

AUTHOR CONTRIBUTIONS

AM, RC, SS, RW, and JW wrote the manuscript. AS and AM supervised the study. RC, RE, JW, NB, MB, EC, JJ, RM, KM, NM, SP, MR, BT, RW, JR, LS, JA-H, NA, AS, and AM are members of CGEn's technical experts committee that conceived of and executed the study. SJ, ML, NA, and SS are the scientific directors of CGEn and oversaw all aspects of this project. All authors reviewed and approved the final manuscript.

FUNDING

CGEn is a national sequencing facility supported by the Canada Foundation for Innovation's Major Science Initiatives. Leveraged co-funding supporting CGEn and the experiments was provided by Genome Canada through Genome BC, Ontario Genomics, and Génome Quebec, and by the University of Toronto McLaughlin Centre.

ACKNOWLEDGMENTS

We are grateful to technical personnel within the library preparation, sequencing, and bioinformatics groups of CGEn for their expertise, and Kirstin Brown for editorial assistance. SS holds the Canadian Institutes of Health Research (CIHR) GlaxoSmithKline Endowed Chair for Genome Sciences at The Hospital for Sick Children and University of Toronto.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.612515/full#supplementary-material

Supplementary Figure 1 | Quality assessment of the 81 BAM files. For all four plots Ashkenazim trio DNA samples are listed on x-axis. Mean X coverage achieved for the genomes **(A)**, adenine and thymine percentage, or AT, dropout rate **(B)**, average insert size in base pairs **(C)**, and guanine and cytosine, or GC, percentage dropout rate **(D)**.

Supplementary Figure 2 | Sensitivity, precision, and F1 values for the 81 datasets. Fractional values are provided for each comparison; F1 measure (gray bars), sensitivity (pink squares), and specificity (blue squares). Sequencing centers (s) and analysis pipelines (p): VAN, Vancouver; MTL, Montreal; TOR, Toronto. Reference DNA sample sources (the Ashkenazim trio; son HG002, father HG003, and mother HG004).

REFERENCES

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371.e18–385.e18. doi: 10.1016/j.cell.2018.02.060

Baskurt, Z., Mastromatteo, S., Gong, J., Wintle, R. F., Scherer, S. W., and Strug, L. J. (2020). VikNGS: a C++ variant integration kit for next generation sequencing association analysis. *Bioinforma. Oxf. Engl.* 36, 1283–1285. doi: 10.1093/bioinformatics/btz716

Beck, S., Berner, A. M., Bignell, G., Bond, M., Callanan, M. J., Chervova, O., et al. (2018). Personal Genome Project UK (PGP-UK): a research and citizen science hybrid project in support of personalized medicine. *BMC Med. Genomics* 11:108. doi: 10.1186/s12920-018-0423-1

Chen, J., Li, X., Zhong, H., Meng, Y., and Du, H. (2019). Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci. Rep.* 9:9345. doi: 10.1038/s41598-019-45835-3

Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., et al. (2015). Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *bioRxiv* [Preprint]. doi: 10.1101/023754

Craig, D. W., Nasser, S., Corbett, R., Chan, S. K., Murray, L., Legendre, C., et al. (2016). A somatic reference standard for cancer genome sequencing. *Sci. Rep.* 6:24607. doi: 10.1038/srep24607

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., et al. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 28, 2678–2679. doi: 10.1093/bioinformatics/bts503

Jeon, S., Bhak, Y., Choi, Y., Jeon, Y., Kim, S., Jang, J., et al. (2020). Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci. Adv.* 6:eaaz7835. doi: 10.1126/sciadv.aaz7835

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* [Preprint]. doi: 10.1101/531210

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* 15, 591–594. doi: 10.1038/s41592-018-0051-x

Knoppers, B. M., Harris, J. R., Budin-Ljøsne, I., and Dove, E. S. (2014). A human rights approach to an international code of conduct for genomic and clinical data sharing. *Hum. Genet.* 133, 895–903. doi: 10.1007/s00439-014-1432-6

Kumaran, M., Subramanian, U., and Devarajan, B. (2019). Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. *BMC Bioinformatics* 20:342. doi: 10.1186/s12859-019-2928-9

Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: scientific containers for mobility of compute. *PLoS One* 12:e0177459. doi: 10.1371/journal.pone.0177459

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* [Preprint]. Available online at: http://arxiv.org/abs/1303.3997 (accessed March 9, 2020).

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., et al. (2018). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 20, 435–443. doi: 10.1038/gim.2017.119

Mascalzoni, D., Dove, E. S., Rubinstein, Y., Dawkins, H. J. S., Kole, A., McCormack, P., et al. (2015). International Charter of principles for sharing bio-specimens and data. *Eur. J. Hum. Genet.* 23, 721–728. doi: 10.1038/ejhg.2014.197

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Moore, R. A., Zeng, T., Docking, T. R., Bosdet, I., Butterfield, Y. S., Munro, S., et al. (2020). Sample tracking using unique sequence controls. *J. Mol. Diagn. JMD* 22, 141–146. doi: 10.1016/j.jmoldx.2019.10.011

Pleasance, E., Titmuss, E., Williamson, L., Kwan, H., Culibrk, L., Zhao, E. Y., et al. (2020). Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* 1, 452–468. doi: 10.1038/s43018-020-0050-6

Priestley, P., Baber, J., Lolkema, M. P., Steeghs, N., de Bruijn, E., Shale, C., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* 575, 210–216. doi: 10.1038/s41586-019-1689-y

Rahimzadeh, V., Dyke, S. O. M., and Knoppers, B. M. (2016). An international framework for data sharing: moving forward with the global alliance for genomics and health. *Biopreservation Biobanking* 14, 256–259. doi: 10.1089/bio.2016.0005

Reuter, M. S., Walker, S., Thiruvahindrapuram, B., Whitney, J., Cohn, I., Sondheimer, N., et al. (2018). The personal genome project Canada: findings from whole genome sequences of the inaugural 56 participants. *CMAJ Can. Med. Assoc. J.* 190, E126–E136. doi: 10.1503/cmaj.171151

Robasky, K., Lewis, N. E., and Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15, 56–62. doi: 10.1038/nrg3655

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., et al. (2007). Challenges and standards in integrating surveys of structural variation. *Nat. Genet.* 39, S7–S15. doi: 10.1038/ng2093

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656. doi: 10.1038/s41576-019-0150-2

Stavropoulos, D. J., Merico, D., Jobling, R., Bowdin, S., Monfared, N., Thiruvahindrapuram, B., et al. (2016). Whole genome sequencing expands diagnostic utility and improves clinical management in pediatric medicine. *NPJ Genomic Med.* 1:15012. doi: 10.1038/npjgenmed.2015.12

Thibodeau, M. L., O'Neill, K., Dixon, K., Reisle, C., Mungall, K. L., Krzywinski, M., et al. (2020). Improved structural variant interpretation for hereditary cancer susceptibility using long-read sequencing. *Genet. Med.* 22, 1892–1897. doi: 10.1038/s41436-020-0880-8

Tom, J. A., Reeder, J., Forrest, W. F., Graham, R. R., Hunkapiller, J., Behrens, T. W., et al. (2017). Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics* 18:351. doi: 10.1186/s12859-017-1756-z

Trost, B., Engchuan, W., Nguyen, C. M., Thiruvahindrapuram, B., Dolzhenko, E., Backstrom, I., et al. (2020). Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* 586, 80–86. doi: 10.1038/s41586-020-2579-z

Trost, B., Walker, S., Haider, S. A., Sung, W. W. L., Pereira, S., Phillips, C. L., et al. (2019). Impact of DNA source on genetic variant detection from human whole-genome sequencing data. *J. Med. Genet.* 56, 809–817. doi: 10.1136/jmedgenet-2019-106281

Turro, E., Astle, W. J., Megy, K., Gräf, S., Greene, D., Shamardina, O., et al. (2020). Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* 583, 96–102. doi: 10.1038/s41586-020-2434-2

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* 43, 11.10.1–11.10.33. doi: 10.1002/0471250953.bi1110s43

Voss, K., Gentry, J., and Auwera, G. V. D. (2017). Full-stack genomics pipelining with GATK4 + WDL + cromwell. *F1000Research* 6:e126144. doi: 10.7490/f1000research.1114631.1

Wang, Y.-C., Olson, N. D., Deikus, G., Shah, H., Wenger, A. M., Trow, J., et al. (2019). High-coverage, long-read sequencing of Han Chinese trio reference samples. *Sci. Data* 6:91. doi: 10.1038/s41597-019-0098-2

Yuen, R. K. C., Merico, D., Bookman, M., Howe, J. L., Thiruvahindrapuram, B., Patel, R. V., et al. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* 20, 602–611. doi: 10.1038/nn.4524

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* 3:160025. doi: 10.1038/sdata.2016.25

Conflict of Interest: SS is on the Scientific Advisory Boards of Deep Genomics and Population Bio and intellectual property arising from his research held at The Hospital for Sick Children is licensed by Athena Diagnostics and Lineagen. The Center for Applied Genomics, directed by SS and The Hospital for Sick Children has benefited from joint relationships with Illumina and other suppliers, but none of these arrangements have impacted the studies described in this paper. SS holds the Canadian Institutes of Health Research (CIHR) GlaxoSmithKline Endowed Chair for Genome Sciences at The Hospital for Sick Children and University of Toronto.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.