Bioinformatics, 33(23), 2017, 3709–3715
doi: 10.1093/bioinformatics/btx468
Advance Access Publication Date: 26 July 2017
Original Paper

OXFORD

Genome analysis

Cloud-based interactive analytics for terabytes of genomic variants data

Cuiping Pan[1,2,*,†], Gregory McInnes[1,3,†], Nicole Deflaux[4,5,†], Michael Snyder[2,3], Jonathan Bingham[4,5], Somalee Datta[1,3] and Philip S. Tsao[1,6,*]

[1]VA Palo Alto Health Care System, Palo Alto Epidemiology Research and Information Center for Genomics, CA 94304, USA, [2]Department of Genetics, [3]Stanford Center for Genomics and Personalized Medicine, Stanford University, CA 94305, USA, [4]Google, Mountain View, CA 94043, USA, [5]Verily Life Sciences, South San Francisco, CA 94080, USA and [6]Division of Cardiovascular Medicine, Stanford University, Stanford, CA 94305, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.
Associate Editor: Inanc Birol

Received on March 2, 2017; revised on June 30, 2017; editorial decision on July 17, 2017; accepted on July 25, 2017

Abstract

Motivation: Large scale genomic sequencing is now widely used to decipher questions in diverse realms such as biological function, human diseases, evolution, ecosystems, and agriculture. With the quantity and diversity these data harbor, a robust and scalable data handling and analysis solution is desired.

Results: We present interactive analytics using a cloud-based columnar database built on Dremel to perform information compression, comprehensive quality controls, and biological information retrieval in large volumes of genomic data. We demonstrate such Big Data computing paradigms can provide orders of magnitude faster turnaround for common genomic analyses, transforming long-running batch jobs submitted via a Linux shell into questions that can be asked from a web browser in seconds. Using this method, we assessed a study population of 475 deeply sequenced human genomes for genomic call rate, genotype and allele frequency distribution, variant density across the genome, and pharmacogenomic information.

Availability and implementation: Our analysis framework is implemented in Google Cloud Platform and BigQuery. Codes are available at https://github.com/StanfordBioinformatics/mvp_aaa_codelabs.

Contact: cuiping@stanford.edu or ptsao@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic sequencing projects have grown from sequencing a few genomes to thousands or even tens of thousands of genomes in the past decade (1000 Genomes Project Consortium *et al.*, 2015; Ball *et al.*, 2014; Telenti *et al.*, 2016), generating massive amount of data that present challenges ranging from affordable long-term storage, controlled data sharing, flexible data retrieval, fast and scalable data processing, and interactive mining of biological information. Recently, jointly analyzing genomic data from multiple studies (Fortney *et al.*, 2015; Fuchsberger *et al.*, 2016; van Rheenen *et al.*, 2016) as well as with other types of data (Abul-Husn *et al.*, 2016; Akbani *et al.*, 2014; GTEx Consortium, 2015) has proven to be invaluable in improving study power and thus yielded important new discoveries. These data integration efforts have highlighted the need for a scalable analysis platform that can combine various information sources.

Published by Oxford University Press 2017. This work is written by US Government employees and is in the public domain in the United States.

Downloaded from https://academic.oup.com/bioinformatics/article/33/23/3709/4036385 by guest on 20 August 2022

Standard public tools, run on local files on fixed-size computer clusters, do not readily scale for large studies. Public cloud platforms, with sufficient computing capacity for large batch analysis jobs, can provide part of the solution and have demonstrated early potential to grow into mature solutions for processing large-scale genomic data (Afgan *et al.*, 2015; Calabrese and Cannataro, 2016; Huang *et al.*, 2013; Karczewski *et al.*, 2014; Reid *et al.*, 2014; Wilkinson and Almeida, 2014; Shringarpure *et al.*, 2015; Souilmi *et al.*, 2015). Another part of the solution may be new frameworks for interactive analytics based on distributed computing approaches, such as Dremel, a SQL query engine based on a columnar database (Melnik et al., 2011).

Here we present a new paradigm for cloud-based genomic computation using a Dremel database to effectively structure dense genomic information and perform complex analytics for large volumes of genomic data (Fig. 1). Our implementation was primarily SQL queries, but also used other distributed computing approaches. We applied this framework to analyzing 475 deeply sequenced human genomes, assessed its performance with larger simulated datasets, and achieved interactive queries in a web browser in seconds for terabytes of variants data.

2 Materials and methods

2.1 Study sample, DNA sequencing and variant calling

The study protocols were approved by the IRB committee at Stanford University. A total of 475 unrelated study subjects were recruited and consented through three local hospitals (VAPAHCS, Stanford Hospitals and Clinics, and Kaiser Permanente). Study IDs were given to each subject for de-identification purpose, which were used throughout this research project.

Blood collection and DNA preparation followed standard protocols. DNA were sequenced using the 101 base-pair pair-end reversible terminator massively parallel sequencing on the HiSeq 2000 instrument to an average genome coverage of 50×. Alignment and variant calling from sequence reads were performed by Genome Analysis Toolkit (GATK) Best Practices (DePristo *et al.*, 2011), via the Google Pipelines Application Program Interface (API), a cloud-based task runner similar to Grid Engine. Briefly, reads were aligned with BWA-MEM v0.7.10 to decoy human reference genome hs37d5, and variants were called for each genome by HaplotypeCaller and recalibrated by Variant Quality Score Recalibration in GATK v3.3. The turnaround for individual sample alignment and variant calling was ~2 days and the total wall-time for all samples, based on our parallelization strategy, was ~5 days.

2.2 Representing genomic calls in Dremel

To collectively analyze a cohort of genomes, we attempted to represent genotypes along with their variant calling metrics in a single table. The current VCF file format provides a good template in which the first five columns denote positional information, followed by three columns denoting variant calling metrics, one column specifying the genotypic format, and the last columns presenting sample-specific genotypes. An apparent limitation of such file format is, should there be more than one genome documented, the variant calling metrics will only represent the first genome and therefore losing the granularity that one wishes to have to examine each individual call across all the genomes. To overcome this problem, we made use of the repeated and nested features of the Dremel database, with which the key information for every called position across all genomes could be retained. Dremel was originally built by Google to analyze petabyte scale log files (Melnik et al., 2011). Its available implementations include Apache Drill, Cloudera Impala, Amazon Athena, and Google BigQuery. In our test case, we used BigQuery in our implementation.

The schema we designed used nested and repeated fields to organize variant calling results in a tree structure (Supplementary Fig. S1). Generally, chromosomal positions for a genomic event, such as reference block, single nucleotide variant (SNV) or short insertion and deletion (INDEL), were recorded as one "record" in the Dremel database. Within a record, definitive parameters such as chromosome name, positions and reference bases were listed as single variables and parameters that could vary across genomes, such as alternate bases, were repeated to accommodate all existing scenarios that appeared in the cohort. For example, in a chromosomal position, if an A to C variation was detected for some genomes, an A to G variation was detected for other genomes, and the rest genomes had a base pair matching to the reference genome, then the alternate bases would be reported in three repeated lines, denoting as C, G

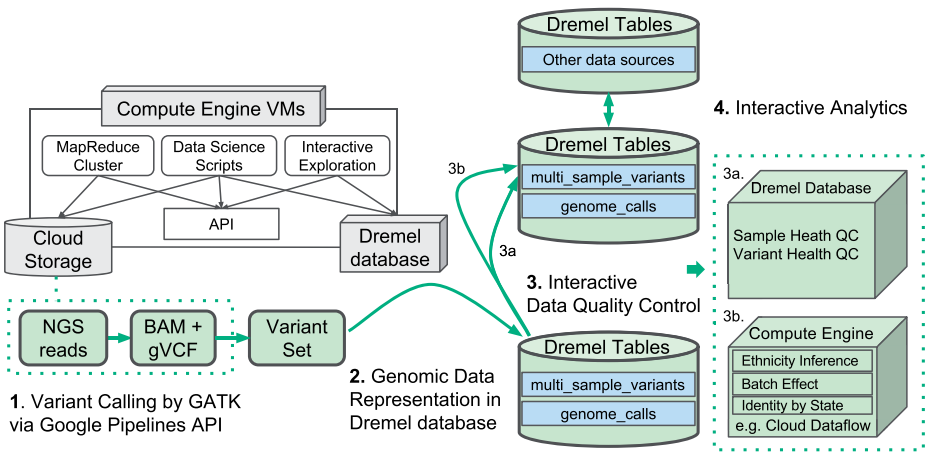

Fig. 1. The computational paradigm of cloud platform-based data processing and Dremel-based interactive analytics for large-scale genomic data. Shown here is the Google Cloud Platform-enabled solution. Variant calling from raw reads to genotypes is performed by GATK via Google Genomics API in Compute Engine Virtual Machines (VMs). Genomic data is represented in a Dremel database to enable interactive data QC and analytics

Downloaded from https://academic.oup.com/bioinformatics/article/33/23/3709/4036385 by guest on 20 August 2022

and null, and the corresponding genome IDs were listed for each case.

We also hierarchically organized records into flat and nested fields. For example, call was nested to reflect multiple parameters from read alignment and variant calling. This hierarchical structure facilitated the retrieval of cohort information by traversing less data and therefore reducing both computation time and cost. The nested fields could be unfolded to access detailed information. Such feature of repeated and nested fields in Dremel databases enabled preservation of read alignment and variant calling details for each genome in a single table.

Finally, to achieve intuitive queries and interactive performance, we designed two types of Dremel tables to represent genomic data. The *genome_calls* table denoted all detected positions, capturing both reference calls and variant calls. In this table, consecutive reference calls were presented as blocks of chromosomal positions ('reference blocking'), a feature resembling the GATK HaplotypeCaller results. On the other hand, the *multi_sample_variants* table centered on only the variable positions, i.e. where at least one genome had a DNA variant, whereas positions with only reference calls were omitted. This variant-centric table thus overcomes the issue brought by reference blocking, i.e. a position under examination might be in the middle of a reference block. By explicitly extracting all calls for the positions of interest, one can conveniently examine its values across all genomes. For our study dataset of 475 genomes, the genome_call table was 1.2 terabytes and the multi_sample_variants table was 1.4 terabytes.

2.3 Simulation of larger genomic datasets

In order to test the scalability of the BigQuery implementation, we simulated large genomic datasets containing 1000, 2500 and 5000 genomes using an Apache Beam pipeline (aka Dataflow). For genomic positions where minor allele frequencies >0.5% were observed in the 1000 Genomes, we randomly generated genotypes for each genome but maintained the same allele frequency rates as in the 1000 Genomes. Additionally, to simulate rarer and unique variation in each individual in concordance with other cohorts of this size (1000 Genomes Project Consortium *et al.*, 2015), we added 20 000 singleton SNVs to each genome according to a uniform distribution across all genomic coordinates. The resulting multi_sample_variants tables in BigQuery were 3.6 terabytes for 1000 genomes, 14.6 terabytes for 2500 genomes, and 48.4 terabytes for 5000 genomes.

2.4 Other distributed computing approaches

We used the GA4GH Genomics APIs, e.g. the variant API, for preprocessing and developed other analytical methods using distributed computing approaches such as Apache Beam and Apache Spark, where SQL query was found to be less optimal for the analyses. Primarily, these non-SQL approaches were used in, but not limited to, data quality control (QC) steps (Supplementary Material).

3 Results

Our study started with sequencing 475 whole human genomes to an average genome coverage of 50× and resulted in 48 terabytes of aligned reads in BAM format and 1.1 terabytes of genotypic information in the compressed gVCF format. We structured these genotypes along with a few preselected variant calling metrics, which we regarded as important for downstream interrogation, in two forms of Dremel tables: the genome_calls table and the multi_sample_variants table. The former captured all reference and variant calls and therefore presented the most comprehensive genotypic information of the cohort; the latter recorded positions where at least one variant had to be present across all genomes, hence variant-centric. These two tables were 1.2 and 1.4 terabytes, respectively. We then used these tables to assess our analytical approaches by its capability in enabling biological discoveries and by systems performance such as runtime, scalability and cost.

3.1 Interactive genomic analytics enabled by Dremel database for biological discoveries

We implemented SQL-based, comprehensive QC steps on the genomic tables following recommendations for genome-wide association studies (GWAS) (Turner *et al.*, 2011) and derived a good quality dataset containing 461 genomes with 26 948 498 SNVs and 8 656 234 INDELs (Supplementary Table S1, Supplementary Fig. S2 and S3, method validation see Supplementary Material). Fourteen genomes were excluded from downstream analysis due to various reasons (Supplementary Table S2). This sample dropout rate is consistent with other large-scale genomic sequencing studies adopting similar QC methods (Fuchsberger *et al.*, 2016; Guo *et al.*, 2017; Kenna *et al.*, 2016). Notably, most of our QC computations, which were carried out on database tables of one terabytes large, were completed in tens of seconds. Next, we sought to gain deep understanding of these 461 genomes for deriving statistical, biological, and medical information.

3.1.1 Near-complete call rate

First, we assessed the call rate in each genome with regard to all positions in the human reference genome hs37d5 (Fig. 2). Briefly, genomic positions on the reference genome were classified into uncalled and called positions. Uncalled positions were those not reported by variant calling. The called positions were further divided into low quality and high quality sub-groups, depending on if they failed or passed QC steps (Fig. 2A). In our deep sequencing genomic dataset, 99.98% of the base pairs in the reference genome were detected, demonstrating the capability of deep sequencing to access almost all positions in the human reference genome (Fig. 2B). The 0.02% uncalled positions were enriched in 1–9 bp, with the longest uncalled region (UR) of 323 bp, suggesting our sequencing left mostly very short gaps in the genome (Fig. 2C left). These URs did not display significant difference among chromosomes (Fig. 2C middle), most of them occurred only once across the study population, and the longer the uncalled gaps, the rarer they became (Fig. 2C right). When overlaying these uncalled gaps with the ENCODE blacklisted regions, we found little overlap (Supplementary Fig. S4). Our analysis suggested little systematic bias of variant detection. Among the called genomic positions, we examined variant counts at different QC levels (Fig. 2D) and in different categories (Fig. 2E), and observed that after complete QC, 99.78% of the genomic positions were reference calls, whereas 0.12% base pairs had an SNV event and 0.08% base pairs had INDEL events (Fig. 2E).

3.1.2 Genomic statistics for individual genomes

We carried on to compute genomic statistics for each genome. Overall, about 3.6 million SNVs and 800 000 INDELs were detected in each genome, of which more than 97% SNVs and 85% INDELs had been previously reported in the dbSNP135 database (Supplementary Fig. 5A and B). In each genome, about 22 000 SNV and 8 500 INDEL events were private, i.e. only occurred to a single genome in this study population. Interestingly many private INDEL calls were almost exclusively heterozygous with enrichment in 1/2

Downloaded from https://academic.oup.com/bioinformatics/article/33/23/3709/4036385 by guest on 20 August 2022

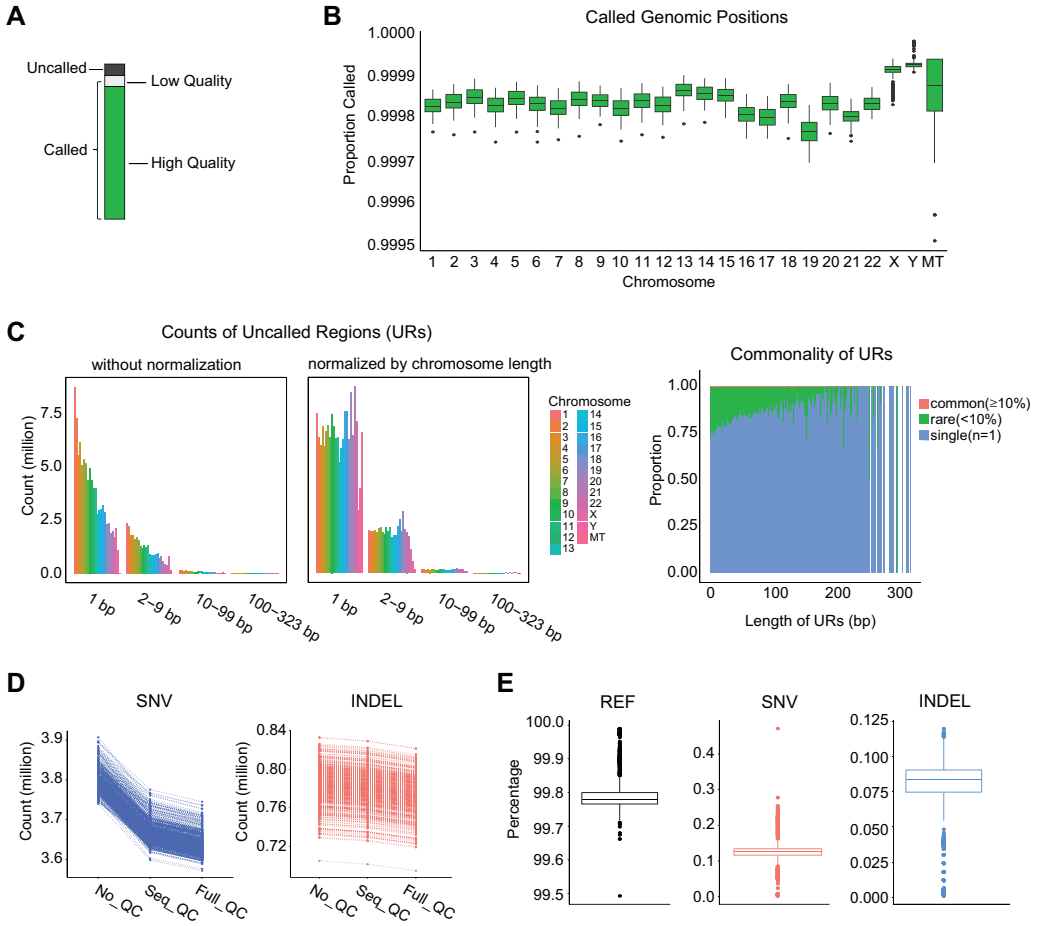

Fig. 2. Callability assessment. **(A)** Diagram shows the overall categories of genomic positions by callability and quality. **(B)** Percentage of detected genomic positions in each chromosome for each genome in this WGS study, based on all calls reported by GATK. **(C)** Uncalled regions (URs) and the length distribution, (left) number of URs per chromosome in each genome, categorized by different length groups; (middle) number of URs per chromosome in each genome, normalized by chromosome lengths; (right): commonality of URs across all genomes. **(D)** Numbers of SNVs and INDELs passing different QC levels. No_QC: all calls by GATK without any filtering. Seq_QC: calls passing the VQSR filtering in GATK. Full_QC: calls passing all levels of QC. **(E)** Percentage of genomic positions called as reference bases, SNVs and INDELs

genotypes, indicating these variations could be rather heterogeneous (Supplementary Fig. S5C and D). For all the variants detected in each genome, the ratio of transition to transversion was about 2.05, and the ratio of heterozygous variants to homozygous variants was around 1.46 (Supplementary Fig. S5E and F). These parameters matched what have been reported in literature (Lam *et al.*, 2011).

3.1.3 Distribution of variants and allele frequencies

Our study subjects, except a few, had genetic ancestry of European (Supplementary Fig. S3E). We computed variant allele frequencies for all genomes in this study, and observed that 25% SNVs and 22% INDELs were common (minor allele frequency [MAF] > 5%), 17% SNVs and 28% INDELs had low frequency (MAF between 0.5 and 5%), and 58% SNVs and 50% INDELs were rare (MAF < 0.5%). There have been multiple deep sequencing efforts in various populations, e.g. the Southeast Asian Malays (Wong *et al.*, 2013), Dutch population (Genome of the Netherlands Consortium, 2014), Icelandic population (Gudbjartsson *et al.*, 2015), and Japanese population (Nagasaki *et al.*, 2015), and all of them reported over half of the variants being rare. Our study, together with these population-based deep sequencing studies, demonstrated that low frequency to rare variants are the main components of an individual genome and should therefore be considered in a comprehensive genomic study, whether in population-based diseases mapping or in clinical interpretation of personal genomes.

Previously, the 1000 Genomes project reported sequencing 503 genomes of European ancestry with average genome coverage of 7.4× (1000 Genomes Project Consortium *et al.*, 2015). When comparing it with our dataset, we found that our dataset had similar number of SNVs but significantly larger number of INDELs (Supplementary Fig. S6). Likely this difference was caused by sequencing depths. While SNVs were easier to detect, detection of INDELs was much more challenging and often required higher sequencing coverages. Despite the difference in absolute counts, the variant distribution pattern across chromosomes was similar. When normalizing by chromosome lengths, in both datasets, chromosomes 9 and 15 consistently displayed proportionally fewer variants than their neighbouring chromosomes. This suggested that our genomic dataset was of high quality and suitable for further discovery.

Next, we assess the saturation call rate in our dataset by the accumulative variant counts with increasing number of genomes. The most rapid increase of SNVs and INDELs occurred to the initial dozens of genomes, and slowed down when genomes further

Downloaded from https://academic.oup.com/bioinformatics/article/33/23/3709/4036385 by guest on 20 August 2022

accumulated, although not reaching saturation at the maximal number of genomes in this study (Fig. 3A). Different allele frequency groups displayed distinct distribution pattern, with rare variants and singletons (i.e. allele count of one) rapidly accumulating. We expect further sequencing beyond this study will reveal many more variants of low to rare frequency. Interestingly, SNV displayed a stronger increase momentum than INDEL, suggesting variation at single nucleotide level occurred more often than those involving longer nucleotides. We observed no obvious difference between the allele frequency distributions for SNV and INDEL (Fig. 3B and C).

3.1.4 Most variable regions of the genome

We surveyed the density of variants by one million bp window across the genome and observed that the most variable regions located on chromosome 6 and chromosome 8 (Supplementary Fig. S7A). The chromosome 6 region exclusively corresponded to the human leukocyte antigen (HLA) genes, which encoded major histocompatibility complex proteins in humans that were responsible for regulating the immune system (Supplementary Fig. S7B). The excessive polymorphism on HLA genes has been known for producing highly variable peptides in the antigen docking regions of MHC that are responsible for docking diverse antigens to the cell surface, thus forming a strategy for our immune system to cope with a broad spectrum of pathogens. On the highly variable region on chromosome 8 variants were more evenly distributed, although an enrichment in the CSMD1 gene was observed (Supplementary Fig. S7C). CSMD1 encodes the CUB and sushi domain-containing protein 1, whose function have been indicated in various human diseases including cancer (Escudero-Esparza *et al.*, 2016; Sun *et al.*, 2001), inflammation (Chandran, 2013), and neurological diseases (Athanasiu *et al.*, 2017). It is unknown whether our study population was enriched in any of the CSMD1-related diseases. We suspect the high polymorphism in this gene was related to its versatile cellular functions.

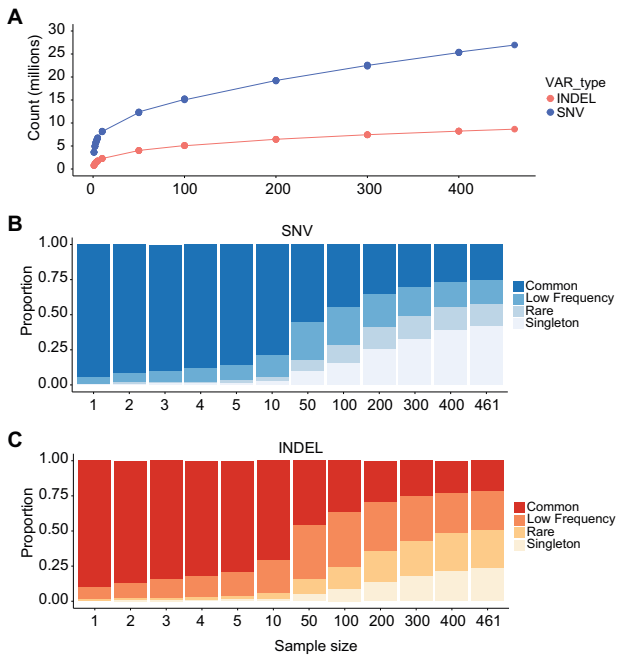

Fig. 3. Saturation call rate for SNVs and INDELs. **(A)** Number of unique SNVs and INDELs by increasing number of genomes. **(B)** Distribution of allele frequencies for SNVs. **(C)** Distribution of allele frequencies for INDELs

3.1.5 Pharmacogenomics suggested for reduced Warfarin dosage

Last, in lieu of the increasing efforts to understand medical implications from genomic sequencing data (Caudle *et al.*, 2017; Dewey *et al.*, 2014; Kalia *et al.*, 2017; Thompson *et al.*, 2014), we examined the genomes in our dataset for pharmacogenomic information. Warfarin is an anticoagulant whose excessive dosing could lead to lethal bleeding, and according to the PharmGKB knowledge base (Caudle *et al.*, 2016), variation patterns of more than 50 genomic positions on CYP2C9 gene and VKORC1 gene can affect sensitivity to Warfarin to different degrees. Particularly, rs1799853 and rs1057910 on CYP2C9 gene, and rs9923231 on VKORC1 gene play important roles in decreasing Warfarin metabolism, which in turn leads to extended accumulation of Warfarin in blood and therefore excessive dosing. We queried these genomic positions in our dataset and concluded that 30% of our study subjects harbored the variation patterns that would conferred them higher sensitivity to Warfarin (Supplementary Fig. S8).

3.2 Scalability and cost analysis

3.2.1 Scalability assessment

In our own experience, batch mode public tools such as VCFtools and bcftools scaled roughly linearly with the number of genomes. For example, when sample sizes increased from 5 to 461 genomes, runtime of the public tools increased from interactive mode (i.e. 1–2 min) to batch mode (i.e. 1–2 h) (Supplementary Table S3). Cloud computation, given the parallel computational paradigm it enables, is known to better overcome the scalability problem. The limit of scalability is thus likely to reside in our data schema design.

We addressed the scalability of our method by comparing performance on 5, 100, 200, 300, 400 and 461 genomes, sampled from our study subjects and represented in multi_sample_variants tables in Dremel with sizes of 19.8 GB, 320 GB, 636 GB, 952 GB, 1.24 TB and 1.43 TB, respectively. The non-linear increase in table size was due to the inclusion of rare variants, which caused increasing sparsity in the tables. We chose to evaluate four queries: two queries spanned across all genomes (total variant counts and missingness per site) and two queries ran on individual genomes (Ti/Tv summary and heterozygosity). For most of the queries examined, we observed the run-time of seconds to tens of seconds (Fig. 4A).

In a further testing, we simulated datasets of 1000, 2500 and 5000 genomes resulting in multi_sample_variants tables in Dremel of 3.6, 14.6 and 48.4 TB, respectively ('Method' section). The queries remained interactive for 1000 genomes, but slowed down when the dataset increased to 5000 genomes (Fig. 4B). For example, computing the ratio of heterozygous calls to homozygous calls for each of the 5000 genomes simultaneously took 30 min in total. This indicated that our schema design, with the intent to retain many key variant calling metrics in a single table, was no longer achieving interactive speeds when unfolding the nested and repeated records for 5000 genomes. Future improvement could consider alternate schemas. For example, another schema has since been developed to store genotypes only when they differ from the reference allele and noting the identity of samples matching the reference in a separate column. This schema drastically reduces the sparsity introduced by large numbers of low frequency or rare variants that are uncovered by whole genome sequencing, and was proven to be able to handle 5205 genomes with interactive speed (Yuan *et al.*, 2017).

3.2.2 Cost assessment

The cost of running queries on the BigQuery database is determined by the amount of data that the SQL computation traverses for each query.

Downloaded from https://academic.oup.com/bioinformatics/article/33/23/3709/4036385 by guest on 20 August 2022

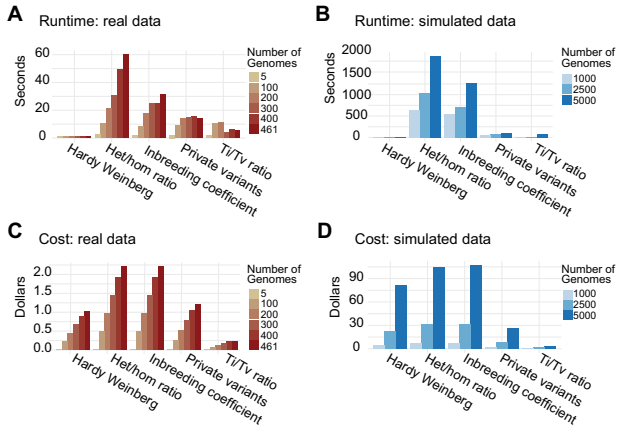

Fig. 4. Dremel query performance assessment on real and simulated genomic data. Shown here are the query run times and cost for five representative queries on tables containing real genomic data, ranging from 5 to 461 genomes (A and C), and simulated genomic data, ranging from 1000 to 5000 genomes (B and D)

Based on the unit cost at the time of writing i.e. $5/TB, we computed the cost of running the four queries in differently sized datasets in the BigQuery and presented them along the runtime (Fig. 4C and D).

Costs of platforms were more difficult to compare because of the integrative computational environment. We singled out the cost based on runtime with the following considerations. For batch mode tools VCFtools and bcftools, (i) we estimated cost based on core-hours and chose the minimum time taken between the tools, (ii) we present results from running the tools on local high-performance computing clusters at near full capacity (on-premise), as well as on Google Compute Engine which resembled the local computing clusters environment to a large extent (cloud) and (iii) we assumed $0.05/core-h for both on-premise and cloud environments. We ignored any fixed or temporary storage costs associated with storing 1.1 terabytes data in the compressed gVCF format. For the Dremel database, we estimated cost based on the amount of data that the queries traversed and ignored the cost for storing the Dremel tables. Analyzing four different queries, we found a tradeoff between cost, scalability, and wall time (Supplementary Table S4). Though the cost for BigQuery was higher, the difference was modest given the pronounced performance gains at orders of magnitude.

Notably, benchmarking of cost is complex and requires some simplifications and caveats. First, the units across which costs were measured differ. Server costs are typically provided in units of core-hours, which is the number of hours the CPU core is engaged during computation taking data transfer into account, whereas on BigQuery, setting aside the cost of data storage itself, the cost of a query is defined by data traversed by the query. These were fundamentally different measurements. Second, cost of an on-premise server is difficult to generalize across all academic centers as different settings and financial models, such as subsidies and charge back, often exist. Third, a fair comparison between on-premise and cloud computation requires taking the entire infrastructure into account such as server cost, storage cost, network cost, and system administration costs for the duration of the project. A complete analysis of system level cost comparison is beyond the scope of this publication.

4 Discussion

In this study, we presented an accessible and novel Big Data oriented computational paradigm combining cloud-based distributed database and computation to address the scalability and interactivity in large genomic data analytics. With the actual data of 475 deeply sequenced human genomes and simulated datasets expanding to 5000 human genomes, we demonstrated that such solutions can greatly shorten the cycle of data analysis and hypothesis testing, transforming long-running batch jobs into questions that can be asked from a web browser interactively. Further, we developed a wide variety of SQL methods to extensively interrogate the 475 genomes from various aspects, leading to insight on the data and novel biological discoveries.

Notably, our analyses were implemented in short, standard SQL code in a browser window without requiring further software development, and most queries completed running in tens of seconds. This fast runtime was partially achieved by the hierarchical and nested schema that we designed to structure genomic data in the database, and partially by parallel computation implemented on the levels of both distributed storage and an elastic computational cluster. A significant advantage of cloud-based computing is that the map-reduce implementation of parallelization was available as a standard infrastructure feature, leaving us to focus on data analysis itself without the need to worry about performance optimization of the computing cluster. Furthermore, various tools are natively available in public clouds to facilitate visualization, documentation and more complex computation, such as hosted Apache Spark and Apache Beam.

Our solution used GA4GH APIs extensively to support interoperability between datasets and systems. We also provided tools to connect our approaches with existing tools, e.g. converting existing VCF data to our proposed schema, and transforming the Dremel table multi_sample_variants to the standard multi-sample VCF file. Our experience showed that the use of interoperable standards simplified data exploration by bringing code simplification, standardization, and analytical transparency. We believe that the API centric approach will allow for development of 'behind-the-scene' data compression schemes, perhaps even from third party providers, that will further reduce cost and enhance performance.

In this study, though we tested a cloud-based service implementing Dremel, researchers restricted from using public clouds can choose to install and operate one of the multiple implementations on-premises, and benefit from much of the performance, subject to the constraints of cluster size, and utilization. The performance and scalability of these Big Data oriented distributed databases make them specifically applicable to large sequencing data.

Acknowledgements

We acknowledge the Genetics Bioinformatics Service Center (GBSC), a Stanford School of Medicine Service Center that provides computational infrastructure for genomics research. GBSC provided the dbGaP compliant on-premise cluster and cloud gateway for this research. We thank Alicia Deng and Hassan Chaib from Stanford University for preparing the DNA for sequencing, Denis Salins, Isaac Liao, and Paul Douglas Billing-Ross from Stanford University for bioinformatics support, Elmer Garduno, Danyao Wang, Asha Rostamianfar and Samuel Gross from Google for bioinformatics support on Google Cloud, Shannon Rego from Stanford University for discussing results, and David Glazer from Google for insightful advice along the project.

Funding

This work was supported by research grants from the National Institutes of Health (1P50HL083800 and 1R01HL101388) and from the Veterans Affairs Office of Research and Development.

Conflict of Interest: N.D. and J.B. are employees of Google and Verily Inc. Google provided the computational platform in which our analytical methods were built. Other authors declare no conflict of interest.

Downloaded from https://academic.oup.com/bioinformatics/article/33/23/3709/4036385 by guest on 20 August 2022

References

1000 Genomes Project Consortium. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Abul-Husn,N.S. *et al.* (2016) Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science (New York, N.Y.)*, **354**.

Afgan,E. *et al.* (2015) Genomics virtual laboratory: a practical bioinformatics workbench for the cloud. *PloS One*, **10**, e0140829.

Akbani,R. *et al.* (2014) A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat. Commun.*, **5**, 3887.

Athanasiu,L. *et al.* (2017) A genetic association study of CSMD1 and CSMD2 with cognitive function. *Brain Behav. Immun.*, **61**, 209–216.

Ball,M.P. *et al.* (2014) Harvard Personal Genome Project: lessons from participatory public research. *Genome Med.*, **6**, 10.

Calabrese,B. and Cannataro,M. (2016) Bioinformatics and microarray data analysis on the cloud. *Methods Mol. Biol. (Clifton, N.J.)*, **1375**, 25–39.

Caudle,K.E. *et al.* (2017) Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). *Genet. Med.*, **19**, 215–223.

Caudle,K.E. *et al.* (2016) Evidence and resources to implement pharmacogenetic knowledge for precision medicine. *Am. J. Health Syst. Pharm.*, **73**, 1977–1985.

Chandran,V. (2013) The genetics of psoriasis and psoriatic arthritis. *Clin. Rev. Allergy Immunol.*, **44**, 149–156.

Datta,S. *et al.* (2016) Secure cloud computing for genomic data. *Nat. Biotechnol.*, **34**, 588–591.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Dewey,F.E. *et al.* (2014) Clinical interpretation and implications of whole-genome sequencing. *JAMA*, **311**, 1035–1045.

Escudero-Esparza,A. *et al.* (2016) Complement inhibitor CSMD1 acts as tumor suppressor in human breast cancer. *Oncotarget*, **7**, 76920–76933.

Fortney,K. *et al.* (2015) Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS Genet.*, **11**, e1005728.

Fuchsberger,C. *et al.* (2016) The genetic architecture of type 2 diabetes. *Nature*, **536**, 41–47.

Genome of the Netherlands Consortium. (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.*, **46**, 818–825.

GTEx Consortium. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, **348**, 648–660.

Gudbjartsson,D.F. *et al.* (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.*, **47**, 435–444.

Guo,M.H. *et al.* (2017) Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E327–E336.

Huang,Z. *et al.* (2013) Cloud processing of 1000 genomes sequencing data using Amazon Web Service. In *2013 IEEE Global Conference on Signal and Information Processing, Austin, TX, USA* (pp. 49–52).

Kalia,S.S. *et al.* (2017) Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.*, **19**, 249–255.

Karczewski,K.J. *et al.* (2014) STORMSeq: an open-source, user-friendly pipeline for processing personal genomics data in the cloud. *PloS One*, **9**, e84860.

Kenna,K.P. *et al.* (2016) NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.*, **48**, 1037–1042.

Lam,H.Y.K. *et al.* (2011) Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.*, **30**, 78–82.

Melnik,S. *et al.* (2011) Dremel: Interactive Analysis of Web-Scale Datasets. *Commun. ACM*, **54**, 114–123.

Nagasaki,M. *et al.* (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.*, **6**, 8018.

Reid,J.G. *et al.* (2014) Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics*, **15**, 30.

Shringarpure,S.S. *et al.* (2015) Inexpensive and highly reproducible cloud-based variant calling of 2,535 human genomes. *PloS One*, **10**, e0129277.

Souilmi,Y. *et al.* (2015) Scalable and cost-effective NGS genotyping in the cloud. *BMC Med. Genomics*, **8**, 64.

Sun,P.C. *et al.* (2001) Transcript map of the 8p23 putative tumor suppressor region. *Genomics*, **75**, 17–25.

Telenti,A. *et al.* (2016) Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 11901–11906.

Thompson,B.A. *et al.* (2014) Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat. Genet.*, **46**, 107–115.

Turner,S. *et al.* (2011) Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics*, *Chapter 1*, Unit1.19. doi: 10.1002/0471142905.hg0119s68.

van Rheenen,W. *et al.* (2016) Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.*, **48**, 1043–1048.

Wilkinson,S.R. and Almeida,J.S. (2014) QMachine: commodity supercomputing in web browsers. *BMC Bioinformatics*, **15**, 176.

Wong,L.-P. *et al.* (2013) Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.*, **92**, 52–66.

Yuan,R.K.C. *et al.* (2017) Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*, **20**, 602–611.

Downloaded from https://academic.oup.com/bioinformatics/article/33/23/3709/4036385 by guest on 20 August 2022