White Paper
Intel® Reference Architecture
Stanford Genomics

Accelerating Secondary Genome Analysis Using Intel® Reference Architecture

Weronika Sikora-Wohlfeld,
Division of Systems Medicine,
Department of Pediatrics,
Stanford University, Stanford, CA

Abhi K. Basu,
Intel Corporation,
Big Data Solutions/Data Center Group,
Portland, OR

Atul J. Butte,
Division of Systems Medicine,
Department of Pediatrics,
Stanford University, Stanford, CA

Monica Martinez-Canales,
Intel Corporation, Big Data Solutions/Data
Center Group, Santa Clara, CA

Executive Summary

The dramatic reduction in whole human genome sequencing costs, from USD 100 million per genome in 2001 to USD 4,500 per genome in 2014,[1] combined with the increasing performance gains in computing technology,[2] are revitalizing the healthcare and life sciences industries in ways only imagined a few years ago.

In fact, the healthcare and life sciences industries are reaching an exciting new inflection point, where they are shifting from population-based healthcare to personalized medicine,[3,4,5,6] and where diagnostics and treatments are prescribed based on each person's health history and genetic profile.

But many technical and policy challenges remain that must be addressed to enable ubiquitous genomics-based medicine and research. While recent U.S. and European laws have gone a long way in evolving healthcare and healthcare research policy, there is still much work to do on the technical infrastructure to enable ubiquitous genomics at scale.

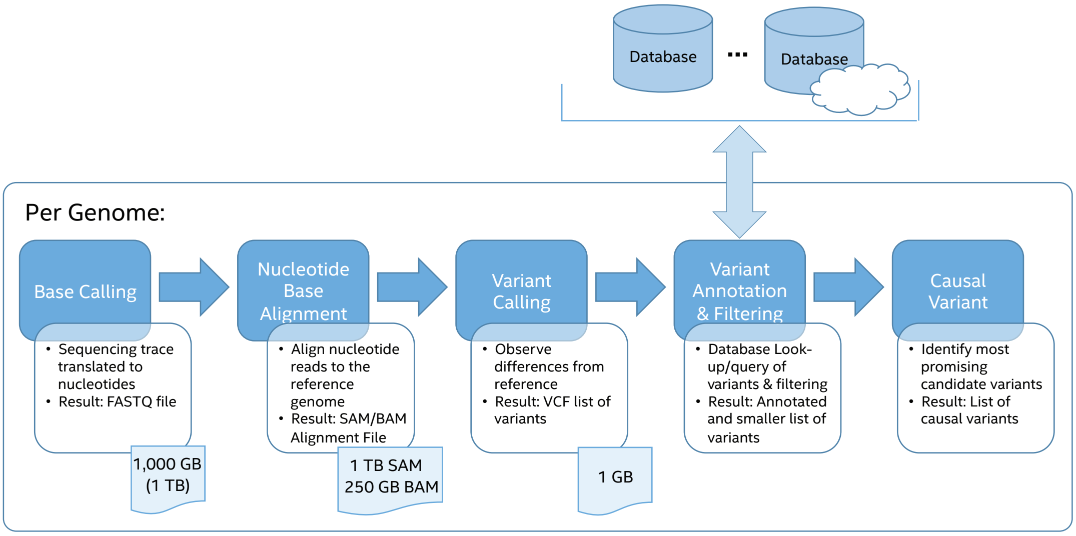
This paper begins to address one of those technical challenges that illustrates the need for data platform technology innovation.

Data Scale Challenges

Sequencing the genome of a person produces a large amount of data, about 1 TB of raw unfiltered data from a sequencing machine. For researchers to use that data, the sequence data must flow through a workflow process, generally described by Dolled-Filhart et al[7] (see Figure 1 based on Dolled-Filhart, but modified with additional detail). A single patient's genome data in raw text format (SAM) is approximately 1 TB in size. The most usable format, Variant Calling Format (VCF),[8] stores only the patient's gene sequence variations and is about 1 GB in size.

Medical or research interpretation requires the addition of many orthogonal biological data sets of equal or greater size. Analysis of genome-scale data is currently limited by the tools available to rapidly annotate, sort, and compare these large data sets. Additionally, the currently used legacy computing systems and data platforms are not capable of analyzing and manipulating these vast data sets. While many public data sets are readily available, archaic text processing tools add significant extract, transform, and load (ETL) overhead to research workflows. Such tools come with some query capabilities but do not provide SQL-like interactive query interfaces to interrogate the data.

Figure 1. Dolled-Filhart et al NGS workflow with data size and database detail.


Database
...
Database
Per Genome:
Base Calling
• Sequencing trace translated to nucleotides
• Result: FASTQ file
1,000 GB (1 TB)
Nucleotide Base Alignment
• Align nucleotide reads to the reference genome
• Result: SAM/BAM Alignment File
1 TB SAM 250 GB BAM
Variant Calling
• Observe differences from reference
• Result: VCF list of variants
1 GB
Variant Annotation & Filtering
• Database Look-up/query of variants & filtering
• Result: Annotated and smaller list of variants
Causal Variant
• Identify most promising candidate variants
• Result: List of causal variants

Citation: Marisa P. Dolled-Filhart, Michael Lee Jr., Chih-wen Ou-yang, Rajini Rani Haraksingh, and Jimmy Cheng-Ho Lin; *Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing*. The Scientific World Journal; Volume 2013, Article ID 730210, http://dx.doi.org/10.1155/2013/730210

A medical researcher often needs to compare the anonymized genome data of a local patient with genome data in other notable public data sets like 1,000 Genomes,[9] UK10K,[10] or NHLBI,[11] as well as with genome data in a private biobank.

To facilitate this common task, we investigated whether Intel's reference architecture met the following requirements:

1. Can it allow storing vast amounts of data in an accessible form and enable very fast queries of that data, thus constituting an ideal environment for performing this work?
2. Can it facilitate the rapid analysis of genomes of patients and research participants to better understand the genetics of human disease?
3. Can it provide the benefit of very efficient integration of multiple large data sets?

Intel Reference Architecture

Intel® Reference Architecture uses Intel's big data analytics platform as the backbone, supplementing it with various data analysis and statistical tools. The entire stack has been optimized to run best on Intel® architecture and provides many security and manageability capabilities.

Table 1 shows the Intel Reference Architecture components. Figure 2 illustrates the software components.

Table 2 shows the physical hardware specs for each node of the six-node Hadoop distribution.

TABLE 1. INTEL REFERENCE ARCHITECTURE COMPONENTS

COMPONENT TYPE	DETAILS
Cloudera* Distribution of Apache Hadoop* software (CDH5*)	• Data platform • Store all files once on Hadoop Distributed File System* (HDFS*) • Fault-tolerant and high throughput
Apache Hive Metastore*	• Flexible schema
SQL on Hadoop (Hive*—MapReduce*)	• ETL • Data management
SQL on Hadoop (Cloudera Impala*—in-memory)	• Interactive queries and near-real-time query performance • Supports ad hoc "what-if" investigations • Business intelligence tool integration
R*, R-Studio*, R-Impala* Connector	• Data analysis • Statistics • Preliminary data visualization
Python* (Pysam*, Impyla*, Biopython* libraries)	• Scripting • Data munging for Hive import

TABLE 2. PHYSICAL HARDWARE SPECS FOR EACH NODE

COMPONENT TYPE	PART NAME	QUANTITY
Knock-down kit	Intel® Server System R2312GZ4GC4 2U 12x3.5 SATA	1
CPU	Intel® Xeon® processor E5-2680 FC-LGA10 2.7 GHz 8.0 GT/s 20 MB 130W 8 cores CM8062107184424	2
Memory	8 GB 1,333 Reg ECC 1.5V DDR3 Romley	16
ATA hard drive	300 GB SSD 2.5-inch SATA 3 Gb/s 25nm Intel® Lyndonville SSDSA2BZ300G301 710 Series	1
ATA hard drive	2 TB HDD 3.5-inch SATA 6 Gb/s 7,200 RPM 64 MB Seagate Constellation* ES ST2000NM0011	12
Network adapter	NIC Niantic* X520-SR2 10 GBase-SR PCI-e dual-port E10G42BFSR or E10G42BFSRG1P5 Duplex Fiber Optic	1
Chassis component	Bezel—Intel A2UBEZEL Locking Bezel with two branding clip-on inserts	1
Power cord	6-foot power cord, 14AWG 15A with three conductors (C13/5-15P), black monoprice 5292	2
Add-in card	LSI* HBA LS100194 (9211-8i) 8-port 6 GB/s SATA +SAS PCIe 2.0 Raid LP	2

Comparing Database Queries: VCFtools Versus Intel Reference Architecture Performance

All data sets in use were in VCF text format. We used Python* for parsing/formatting the data for Hive* consumption. Hive was used to perform data munging and ETL processes before final tables were built. Impala* (in-memory SQL on Hadoop*) was used for interactive near-real-time queries and ad hoc investigation of data. This reference architecture provided the flexibility to pull in various data sources, store them once on Hadoop Distributed File System* (HDFS*), apply flexible schema in the Hive metastore, and then issue SQL queries over Hive (MapReduce*, disk-based) and Impala (in-memory).

VCFtools*[12] is one of the most popular tools for manipulating VCF files. Although the tool is very simple to use, computation time is the main limitation, since the tool parses text files on a single compute node. The 1000 Genome Project[13] Phase 1 data set consists of 1,092 individuals from 14 populations over four continents. Genetic background studies benefit from the 1000 Genome data set because of its population coverage and diversity. This data set provides variant calls, alignments, and raw sequence files. For our work, we used variant calls data, which measure 1.5 TB when uncompressed (VCF files).

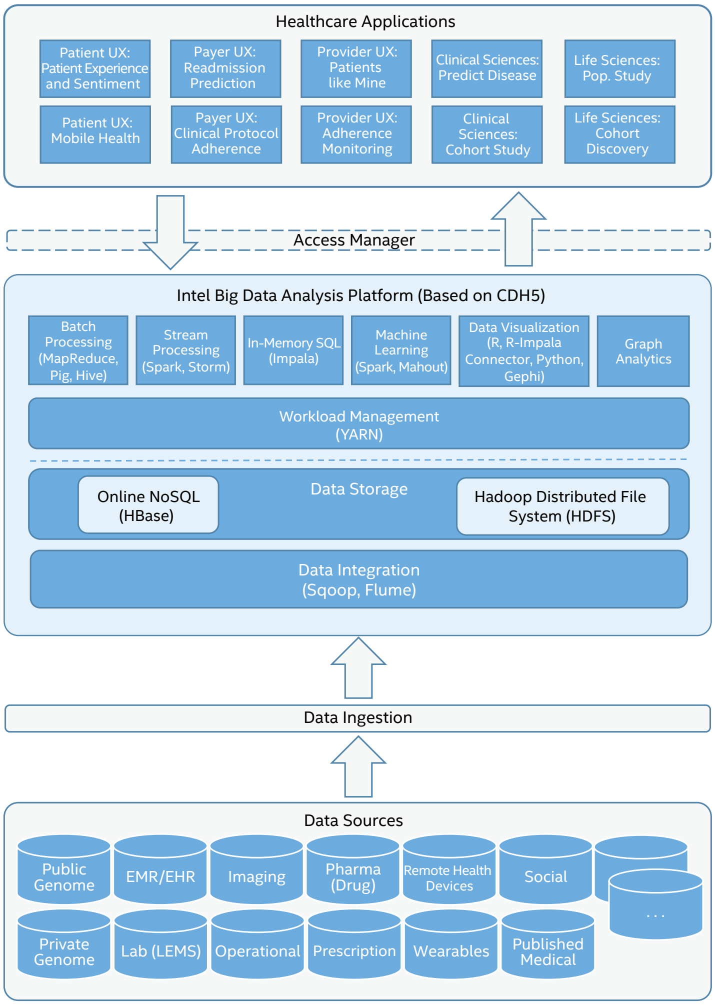

Healthcare Applications
Patient UX: Patient Experience and Sentiment
Payer UX: Readmission Prediction
Provider UX: Patients like Mine
Clinical Sciences: Predict Disease
Life Sciences: Pop. Study
Patient UX: Mobile Health
Payer UX: Clinical Protocol Adherence
Provider UX: Adherence Monitoring
Clinical Sciences: Cohort Study
Life Sciences: Cohort Discovery
Access Manager
Intel Big Data Analysis Platform (Based on CDH5)
Batch Processing (MapReduce, Pig, Hive)
Stream Processing (Spark, Storm)
In-Memory SQL (Impala)
Machine Learning (Spark, Mahout)
Data Visualization (R, R-Impala Connector, Python, Gephi)
Graph Analytics
Workload Management (YARN)
Online NoSQL (HBase)
Data Storage
Hadoop Distributed File System (HDFS)
Data Integration (Sqoop, Flume)
Data Ingestion
Data Sources
Public Genome
EMR/EHR
Imaging
Pharma (Drug)
Remote Health Devices
Social
. . .
Private Genome
Lab (LEMS)
Operational
Prescription
Wearables
Published Medical

Figure 2. Intel Reference Architecture Components.

Approach

Extract the following fields from the 1000 Genomes data, for all chromosomes (1-22 and X):

1. chrom (chromosome)
2. pos (position)
3. id (SNP id)
4. ref (reference allele)
5. alt (alternate allele)
6. info/VT (variant type)
7. info/EUR_AF (allele frequency in Europeans)

Implementation

Table 3 shows how we transformed a VCFtools query to the Impala SQL command for comparison. Each query was run three times for each chromosome and chosen search predicates.

Results

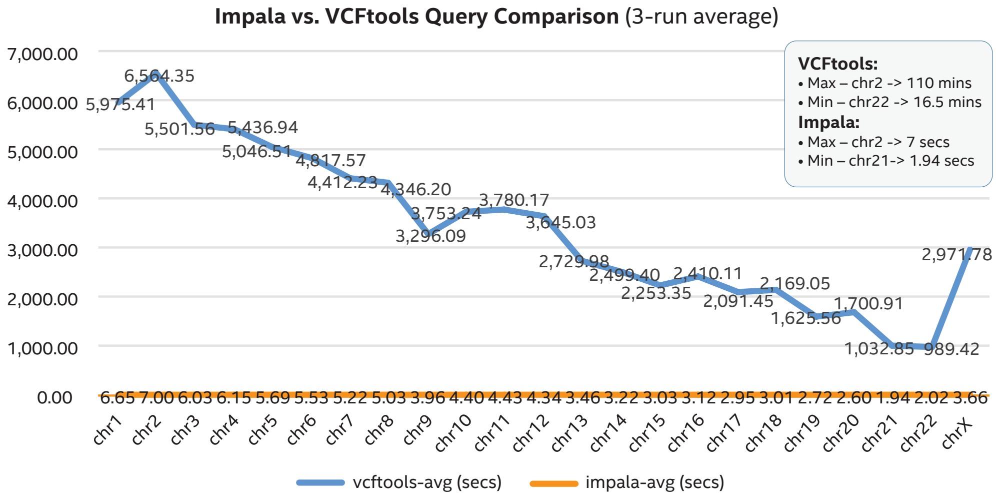
Figure 3 shows a query comparison of Impala and VCFtools. Table 4 shows the variant cardinality and the average query throughput.

Summary

1. VCFtools requires approximately 22 hours to run the queries on all chromosomes. Intel Reference Architecture completes all queries in approximately 1.6 minutes.
2. VCFtools queries range from 16.5 minutes (minimum) to 110 minutes (maximum) for one chromosome. Intel Reference Architecture queries range from approximately 2 seconds (minimum) to 7 seconds (maximum).
3. Intel architecture demonstrates predictable and consistent performance for all queries.

TABLE 3. TRANSFORMING VCFTOOLS QUERY TO IMPALA SQL COMMAND

VCFTOOLS QUERY	IMPALA QUERY
vcf-query input_file.vcf –f '%CHROM\t%POS\t%ID\t%REF\t%ALT\t%INFO/VT\t%INFO/EUR_AF\n' > output_file.txt	CREATE TABLE output_table AS SELECT chrom,pos,id,ref,alt, substr(strleft(substr(info, locate("VT", info)), if(locate(";", substr(info, locate("VT", info)))-1 > 0, locate(";", substr(info, locate("VT", info)))-1, length(substr(info, locate("VT", info))))),4) AS vt, substr(strleft(substr(info, locate("EUR_AF", info)), if(locate(";", substr(info, locate("EUR_AF", info)))-1 > 0, locate(";", substr(info, locate("EUR_AF", info)))-1, length(substr(info, locate("EUR_AF", info))))),8) AS eur_af FROM input_table;



VCFtools:
- Max – chr2 -> 110 mins
- Min – chr22 -> 16.5 mins

Impala:
- Max – chr2 -> 7 secs
- Min – chr21-> 1.94 secs

Figure 3. Impala versus VCFtools Query Comparison (Three-Run Average).

TABLE 4. PER CHROMOSOME VARIANT CARDINALITY AND AVERAGE QUERY THROUGHPUT

CHROMOSOME	NUMBER OF VARIANTS	AVERAGE TIME TO RUN VCFTOOLS (SECONDS)	AVERAGE TIME TO RUN IMPALA (SECONDS)
1	3,007,196	5,975.41	6.65
2	3,307,592	6,564.35	7.00
3	2,763,454	5,501.56	6.03
4	2,736,765	5,436.94	6.15
5	2,530,217	5,046.51	5.69
6	2,424,425	4,817.57	5.53
7	2,215,231	4,412.23	5.22
8	2,183,839	4,346.20	5.03
9	1,652,388	3,296.09	3.96
10	1,882,663	3,753.24	4.40
11	1,894,908	3,780.17	4.43
12	1,828,006	3,645.03	4.34
13	1,373,000	2,729.98	3.46
14	1,258,254	2,499.40	3.22
15	1,130,554	2,253.35	3.03
16	1,210,619	2,410.11	3.12
17	1,046,733	2,091.45	2.95
18	1,088,820	2,169.05	3.01
19	816,115	1,625.56	2.72
20	855,166	1,700.91	2.60
21	518,965	1,032.85	1.94
22	494,328	989.42	2.02
X	1,487,477	2,971.78	3.66
Total	**39,706,715**	**79,049.16**	**96.15**

Note: It is possible to simplify the Impala query using regular expressions; however, the overhead cost is the doubling of Impala's response time, even though the query performance is still an improvement over VCFtools.

The Benefits

Intel Reference Architecture enables usage of diverse data sets from multiple sources that can be stored on HDFS once and used for further analysis downstream. Hive/Impala allows very flexible schema application to data sets and the ability to view data in the form of tables. Impala (in-memory SQL on Hadoop) provides researchers with direct access to issue interactive near-real-time queries, saving results in the form of tables. Users have further flexibility to access the same data for statistics/analysis work using tools like R* and data visualization tools such as Gephi* and Tableau*.

The Future: Non-Volatile Memory

Future work will include investigating other genome data analysis tools and seeing how such workloads can be accelerated using big data technologies. With more reliance on in-memory applications (such as Impala and Spark*/Shark*), Intel is working on evaluating the effects of non-volatile memory and capability to allow much larger RAM on every node of a cluster on genomics and other workloads.

For more information on Intel® technology for Health and Life Sciences, visit **www.intel.com/content/www/us/en/healthcare-it/healthcare-overview**

[1] Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).* Available at: www.genome.gov/sequencingcosts. Accessed 31 July 2013.

[2] Shekhar Borkar and Andrew A. Chien. *The Future of Microprocessors*. Communications of the ACM, 2011. 54(5) [doi :10.1145/1941487.1941507]

[3] Stifel, Nicolaus & Company, *Initiating Coverage on Digital Healthcare; Building the Healthcare Cloud*. 18 July 2014.

[4] The President's Council of Advisors on Science and Technology, *Big Data and Privacy: A Technological Perspective.* Executive Office of the President, May 2014.

[5] Eric D. Green, Mark S. Guyer, and National Human Genome Research Institute, *Charting a course for genomic medicine from base pairs to bedside.* Nature, 10 Feb 2011, 470. p.204-213. [doi:10.1038/nature0976

[6] Marta Garcia Martinez de Lecea and Michael Rossbach, *Translational genomics in personalized medicine – scientific challenges en route to clinical practice.* The HUGO Journal 2012, 6(2). http://www.thehugojournal.com/content/6/1/2

[7] Marisa P. Dolled-Filhart, Michael Lee Jr., Chih-wen Ou-yang, Rajini Rani Haraksingh, and Jimmy Cheng-Ho Lin, *Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing.* The Scientific World Journal, 2013. Article ID 730210, 10 pages, 2013. http://dx.doi.org/10.1155/2013/730210

[8,12] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group, The Variant Call Format and VCFtools. Bioinformatics, 2011. 27(15): p. 2156-2158. [dx.doi.org/10.1093/bioinformatics/btr330]

[9,13] 1,000 Genomes Project Consortium, *An integrated map of genetic variation from 1,092 human genomes*. Nature 491, p56–65 (01 November 2012) [doi:10.1038/nature11632.]

[10] The UK10K Project. http://www.uk10k.org/

[11] NHLBI. Available at http://www.nhlbi.nih.gov/. Accessed on August 2, 2014.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web site at www.intel.com at www.intel.com.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Printed in USA 0914/MMC/TDA/XX/PDF ♻ Please Recycle 331199-001US