White Paper
Intel Health & Life Sciences

Statistical Analysis of Genome Sequencing Data with Intel Reference Architecture

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

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

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

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

Abstract

Next generation sequencing (NGS) technologies generate vast amounts of variant data, the analysis of which poses a big computational challenge. Numerous currently undertaken research efforts, such as population genetics studies or association studies, require computing various statistics and performing statistical tests on the genome sequencing data. With the aim of facilitating such analyses, Intel has developed a specialized analytics platform, referred to as the Intel Reference Architecture. This platform provides a comprehensive set of solutions, which enable convenient storing, manipulating and analyzing the genome sequencing data. The intuitive representation of variant data in a table format and the SQL-like interactive query interface make the Intel Reference Architecture a very attractive alternative to the existing NGS analytics tools.

In this study, we present a set of exemplary queries, which allow executing commonly used operations, such as calculating allele and genotype frequencies, testing for Hardy-Weinberg equilibrium and for association between SNPs and a given condition. To illustrate these queries, we used the 1000 Genomes data and we applied the operations to a set of 12 SNPs, known to be associated with type 2 diabetes.

Table of Contents

In this study we developed a pipeline for statistical analysis of variant data reported in diverse human populations (Figure 1). To illustrate the functionality of this pipeline we applied it to investigate the differences in the allele and genotype distributions of the type 2 diabetes (T2D) associated single nucleotide polymorphisms (SNPs) in different human populations from the 1000 Genomes Project[1]. It has been previously reported that the allele and genotype distributions of T2D associated SNPs show very strong geographical differentiation. The largest differences were observed between East Asian and African populations[2]. In this work we aimed at testing the hypothesis that the genotype distributions of T2D associated SNPs are indeed statistically significantly different between East Asian and African populations.

Schematic representation of the workflow we followed is shown in Figure 1. Particular components are described in details in the following sections. Briefly, we first built a repository of T2D associated single nucleotide polymorphisms (SNPs) reported in literature (Figure 1A) and we identified 12 SNPs strongly associated with T2D in multiple human populations[2]. Then, we used the Intel Reference Architecture querying engine (Figure 1B) to extract the genotype data for these 12 SNPs in 911 individuals from the 1000 Genomes Project. The selected individuals belonged to one of the 11 subpopulations, which could be grouped into three main populations (continental groups): East Asian, European and African (Figure 1C). We calculated the T2D risk allele frequencies in all 11 subpopulations and analyzed the geographical pattern comparing it with previously reported findings. Next, we pooled the individual subpopulations into the three main populations and in each group we counted the genotypes for T2D associated SNPs. We further used the genotype counts as follows: 1) to calculate genotype frequencies and compare them with allele frequencies (Figure 1D), 2) to perform the Hardy-Weinberg equilibrium test (Figure 1F) and 3) to perform the association test (Figure 1E). The association test was performed to assess whether any of the 12 T2D associated SNPs show significant differences in genotype distributions between East Asian and Africans, i.e., whether any of these SNPs are 'associated' with ethnicity. The significant result of this test (defined as p-value below a certain threshold) would indicate that the investigated SNPs indeed show different genotype distributions between East Asian and African populations, which would confirm our initial hypothesis.

Variation in T2D risk allele frequencies across populations

Type 2 diabetes (T2D) is a complex metabolic disease with multifactorial etiology. In addition to environmental factors, which play a considerable role in the development of T2D, genetic susceptibility is a well-established risk factor[3]. A large number of T2D genome-wide association studies (GWAS) performed thus far have resulted in the identification of numerous T2D associated variants. We previously used the known associations to predict the individual's genetic risk of T2D[4, 5]. Furthermore, we evaluated the population-specific genetic risk of T2D using samples from the HapMap project

Methodology

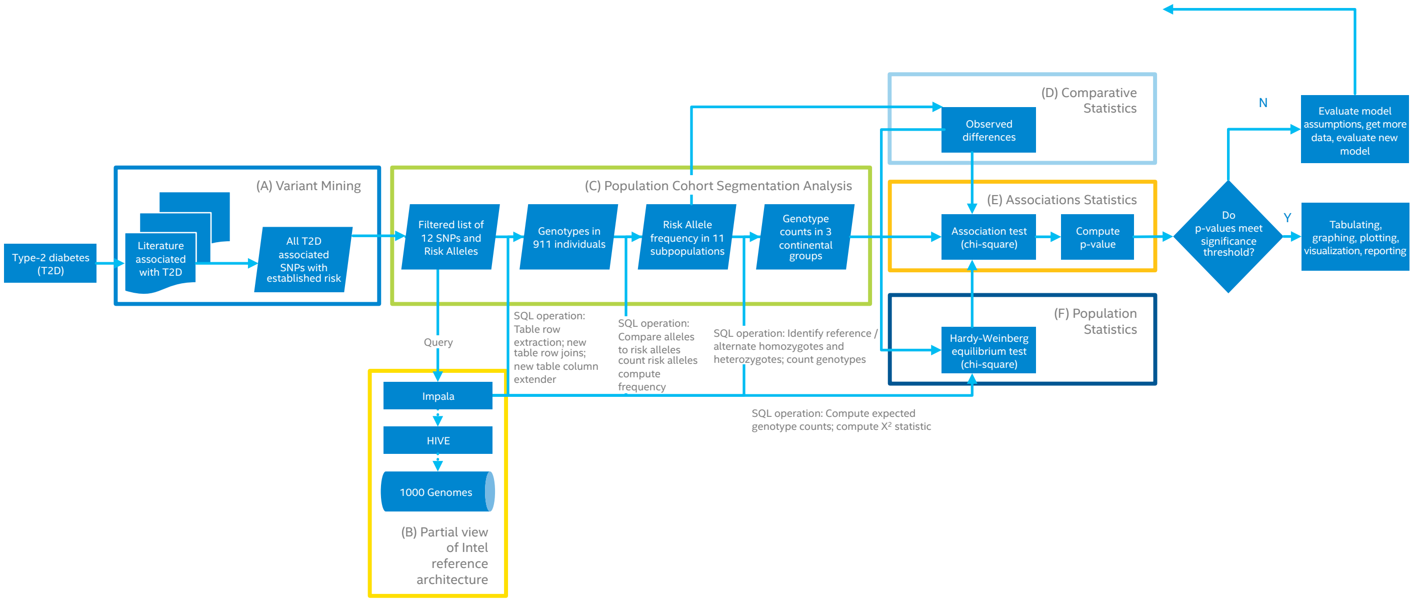
Type-2 diabetes (T2D)
(A) Variant Mining
Literature associated with T2D
All T2D associated SNPs with established risk
(C) Population Cohort Segmentation Analysis
Filtered list of 12 SNPs and Risk Alleles
Genotypes in 911 individuals
Risk Allele frequency in 11 subpopulations
Genotype counts in 3 continental groups
Query
Impala
HIVE
1000 Genomes
(B) Partial view of Intel reference architecture
SQL operation: Table row extraction; new table row joins; new table column extender
SQL operation: Compare alleles to risk alleles count risk alleles compute frequency
SQL operation: Identify reference / alternate homozygotes and heterozygotes; count genotypes
SQL operation: Compute expected genotype counts; compute X² statistic
(D) Comparative Statistics
Observed differences
(E) Associations Statistics
Association test (chi-square)
Compute p-value
(F) Population Statistics
Hardy-Weinberg equilibrium test (chi-square)
Do p-values meet significance threshold?
N
Y
Evaluate model assumptions, get more data, evaluate new model
Tabulating, graphing, plotting, visualization, reporting

Figure 1. Pipeline for statistical analysis of variant data across human populations.

and Human Genome Diversity Panel (HGDP)[2]. We found that the predicted genetic risk of T2D significantly differed between populations, being the highest in the African populations, the lowest in the Asian populations and intermediate in the European populations[2]. In particular, we identified 12 single nucleotide polymorphisms (SNPs) significantly associated with T2D genetic risk in multiple populations[2] (Table 1). We observed that a number of T2D associated SNPs consistently showed a characteristic geographic pattern of risk allele frequencies, decreasing from Sub-Saharan Africa through Europe to East Asia[2].

TABLE 1. Selected T2D associated SNPs.

SNP	RISK ALLELE
rs7903146	T
rs10811661	T
rs13266634	C
rs4402960	T
rs7754840	C
rs5219	T
rs1111875	C
rs11196205	C
rs8050136	A
rs2237892	C
rs7756992	G
rs2074196	G

In the current study, we analyzed the risk allele frequencies for the 12 T2D associated SNPs in human populations using the 1000 Genomes phase 1 data[1]. The 1000 Genomes data have been imported into the Intel Reference Architecture as previously described[6]. Briefly, the 1000 Genomes VCF files were downloaded from the project FTP site (**ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/integrated_call_sets/**) and copied into Hadoop Distributed File System (HDFS). Next, the data were loaded into Hive tables, which were directly available to Impala in-memory SQL queries. The 1000 Genomes data were split chromosome-wise and thus stored as 23 distinct tables. Each table contains hundreds of thousands of rows corresponding to the individual SNPs and the following nine columns:

1. **CHROM** (chromosome),
2. **POS** (SNP position),
3. **ID** (SNP id number),
4. **REF** (reference allele),
5. **ALT** (alternate allele),
6. **QUAL** (quality score),
7. **FILTER** (filtering information),
8. **INFO** (additional SNP annotation),
9. **FORMAT** (description of the genotype data format) followed by 1092 columns storing the genotype data of the individual samples.

In order to analyze the 12 T2D associated SNPs, we first extracted specific rows from the appropriate tables and joined them into a new table. We then extended this table by adding an additional column containing the risk allele information for each SNP as shown in Table 1.

First, we calculated risk allele frequencies for all T2D associated SNPs in each of the 11 subpopulations of the three main continental groups: East Asian, European and African. In the 1000 Genomes tables based on the original VCF files, the genotypes are encoded with the following symbols: 1) “0|0” referring to a reference homozygote, 2) “1|1” referring to an alternate homozygote and 3) “0|1” or “1|0” referring to a heterozygote. The number of risk alleles in a given population equals double the number of the risk homozygotes (individuals carrying two risk alleles) plus the number of the heterozygotes (individuals carrying one risk and one reference allele). We built a query to compare the reference and alternate alleles with the risk allele, count the number of risk alleles and calculate the risk allele frequency in each of the selected subpopulations (see Appendix). The results are shown in Figure 2. The risk allele frequency patterns observed in the 1000 Genomes populations coincide well with the findings reported before[2]. Majority of SNPs show the highest risk allele frequency in the African populations. Except for one SNP, rs5219, risk allele frequency in the East Asian populations is always lower than in the African populations and mostly also lower than in the European populations. This observation confirms that the T2D associated SNPs tend to show geographically diverse risk allele frequency distribution and despite some exceptions (e.g. rs5219), the cumulative genetic risk of T2D appears to be the lowest in the East Asian and the highest in the African populations.

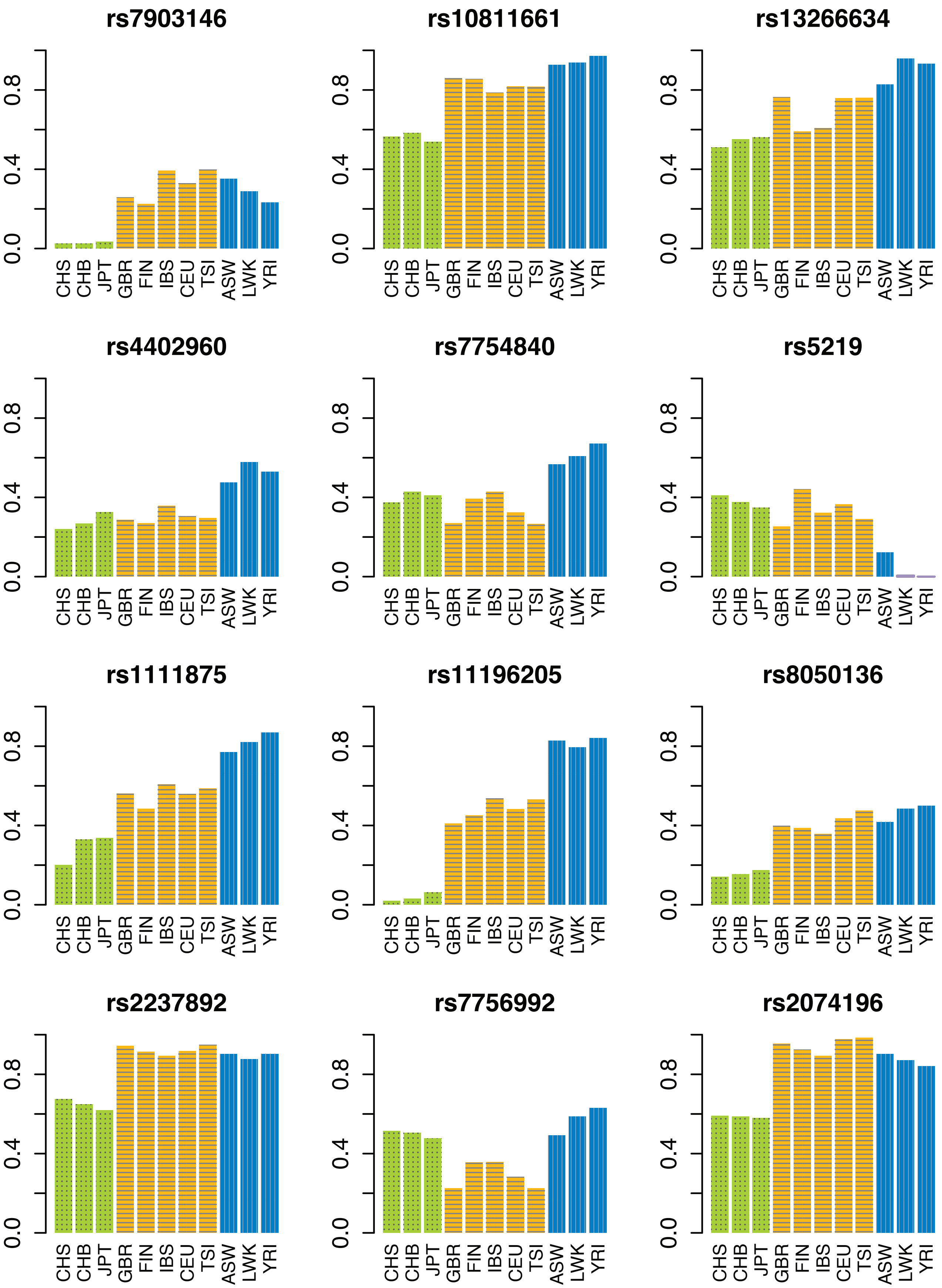

Figure 2. T2D risk allele frequencies in the 1000 Genomes populations segmented by continental groups. Individual subpopulations are grouped into three continental groups: East Asian (CHS, CHB, JPT), European (GBR, FIN, IBS, CEU, TSI) and African (ASW, LWK, YRI).

Genotype counts for T2D associated SNPs across populations

In the next step of the analysis, we calculated genotype counts. We pooled samples from individual populations into three categories corresponding to the main population groups: East Asian, European and African. For each T2D associated SNP, we computed counts of all three possible genotypes: 1) reference homozygote, 2) heterozygote and 3) alternate homozygote. We built a query to return the genotype counts in the three main population groups for all T2D associated SNPs (see Appendix). The results are listed in Table 2.

TABLE 2. Genotype counts for 12 T2D associated SNPs across the East Asian, European and African populations ('ref/ref,' 'ref/alt' and 'alt/alt' denote reference homozygote, heterozygote and alternate homozygote, respectively).

SNP	EAST ASIAN			EUROPEAN			AFRICAN		
	ref/ref	ref/alt	alt/alt	ref/ref	ref/alt	alt/alt	ref/ref	ref/alt	alt/alt
rs7903146	271	14	1	187	151	41	122	108	16
rs10811661	92	138	56	268	97	14	221	24	1
rs13266634	84	141	61	197	147	35	207	37	2
rs4402960	151	112	23	192	153	34	55	119	72
rs7754840	99	143	44	189	140	50	37	113	96
rs5219	40	137	109	37	181	161	2	12	232
rs1111875	26	112	148	119	179	81	166	74	6
rs11196205	266	19	1	103	194	82	11	67	168
rs8050136	207	69	10	136	166	77	69	121	56
rs2237892	116	139	31	326	53	0	194	51	1
rs7756992	69	148	69	210	130	39	47	113	86
rs2074196	89	157	40	347	32	0	186	55	5

Next, we exported the genotype counts table into R for further processing. We calculated the genotype frequencies for each SNP in each population group and we visualized the results as a barplot highlighting the risk homozygote frequency (Figure 3). As expected, the genotype frequency patterns follow the risk allele frequency patterns, showing lower risk homozygote frequency in the East Asian than in the African populations.

While the genotype counts data can be analyzed as such, they can also serve as intermediate results, used in more advanced statistical tests. In the following sections, we utilize the genotype counts data to perform the chi-square tests.

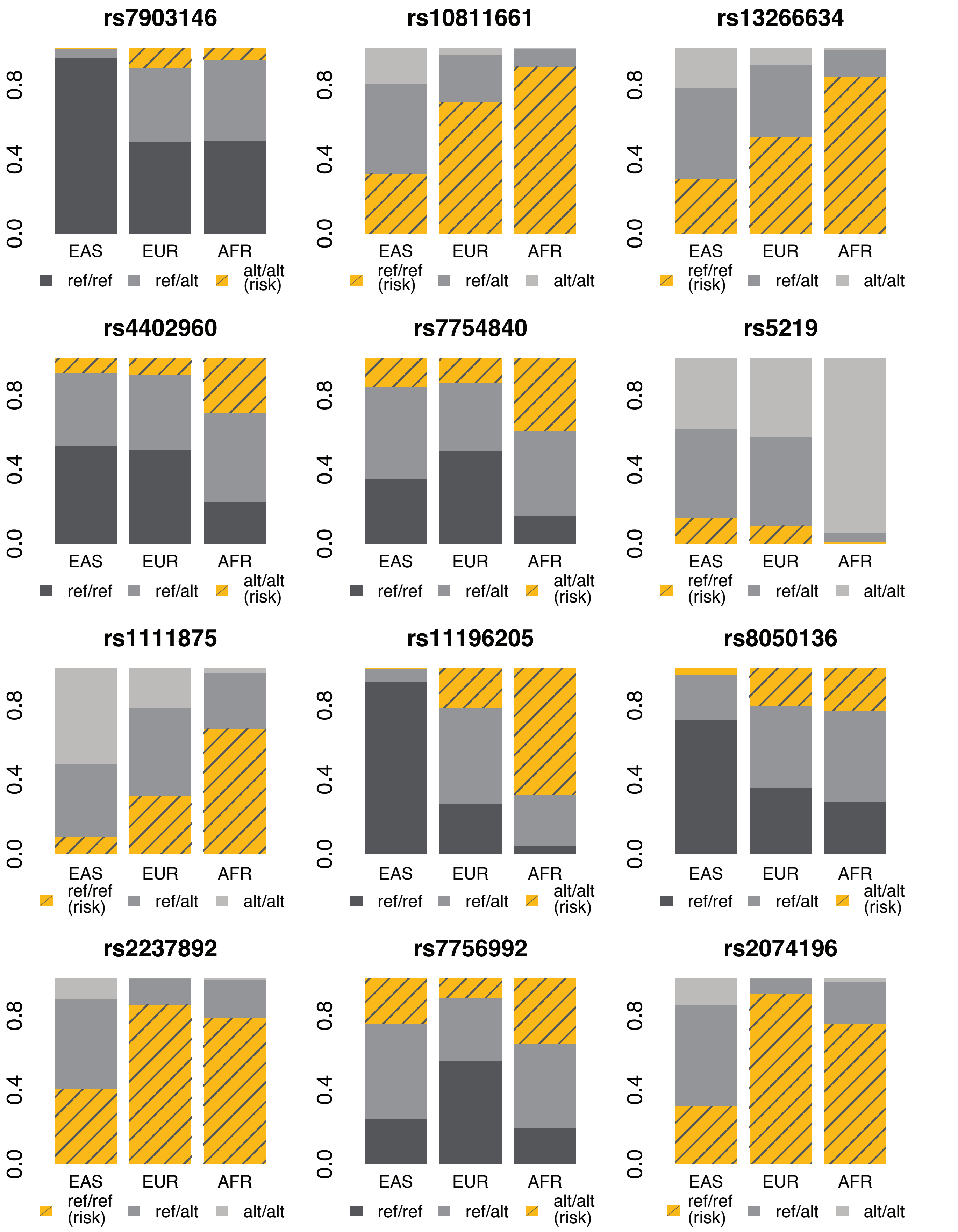

Figure 3. Genotype frequencies for T2D associated SNPs in the 1000 Genomes populations. 'EAS', 'EUR' and 'AFR' denote East Asian, European and African populations, respectively. 'ref/ref ', 'ref/alt' and 'alt/alt' denote reference homozygote, heterozygote and alternate homozygote, respectively. The risk homozygote is highlighted in goldenrod.

Hardy-Weinberg equilibrium test for T2D associated SNPs

The Hardy-Weinberg principle, also known as Hardy-Weinberg equilibrium (HWE), refers to the distribution of allele and genotype frequencies in a population. According to HWE, these frequencies remain constant in the absence of any evolutionary influences. The Hardy-Weinberg principle allows calculating the expected counts of all possible genotypes, given the counts of the individual alleles. Assuming a single locus with two alleles, *A* and *a* and their corresponding frequencies $f(A)=p$ and $f(a)=q$, the expected genotype counts are as follows: $f(AA)=p^2$ for AA homozygotes, $f(Aa)=2pq$ for heterozygotes and $f(aa)=q^2$ for aa homozygotes. If the observed genotype counts coincide with the expected genotype counts, such population is considered at equilibrium. Perfect equilibrium, however, would only occur in a theoretical, indefinitely large population, with no influence of evolutionary events. The real-world populations typically show some deviation from HWE, however the reasonably large populations, which follow the assumptions of HWE, should still show a congenial distribution of genotype counts. A significant deviation from HWE might indicate that some of the HWE assumptions are violated (e.g. random mating) or that the sample is not large enough to adequately represent all possible genotypes. In order to test whether the observed genotype counts significantly differ from the expected genotype counts, statistical tests, such as chi-square test, can be performed. HWE test often performed preceding other analyses (e.g. association test discussed in the next section) in order to prevent the findings being influenced by the violation of HWE.

In this study we demonstrate the SQL implementation of the chi-square test for HWE. We used the observed genotype counts computed in the previous section and then we built a set of queries to calculate the intermediate parameters, such as expected genotype counts, and eventually the chi-square statistic (see Appendix). Finally, we exported the chi-square statistic values into R to calculate the corresponding p-values using the `pchisq` function.

Table 3 shows the chi-square statistic values and the corresponding p-values for 12 T2D associated SNPs in three population groups: East Asian, European and African. Assuming the significance level as p-value=0.001, all SNPs indicate no significant deviation from HWE in any of the populations, except for rs5219 SNPs in the African populations. Given the extreme distribution of rs5219 alleles in the African populations, with very rare occurrence of one allele, the observed deviation from HWE might suggest that the analyzed sample is not big enough to adequately represent all genotypes.

TABLE 3. Chi-square test results for Hardy-Weinberg equilibrium in the East Asian, European and African populations. 'X^2' denotes the chi-square statistic calculated with Impala query and 'p-value' denotes the p-value calculated in R.

SNP	EAST ASIAN		EUROPEAN		AFRICAN	
	X^2	P-VALUE	X^2	P-VALUE	X^2	P-VALUE
rs7903146	2.850	0.091	1.568	0.211	1.506	0.220
rs10811661	0.108	0.742	1.898	0.168	0.159	0.690
rs13266634	0.016	0.898	0.980	0.322	0.059	0.808
rs4402960	0.121	0.727	0.197	0.657	0.191	0.662
rs7754840	0.422	0.516	8.123	0.004	0.157	0.692
rs5219	0.085	0.770	1.838	0.175	12.431	0.0004
rs1111875	0.518	0.472	0.796	0.372	0.449	0.503
rs11196205	1.056	0.304	0.274	0.600	1.612	0.204
rs8050136	1.917	0.166	3.963	0.047	0.045	0.832
rs2237892	1.254	0.263	2.142	0.143	1.513	0.219
rs7756992	0.350	0.554	7.284	0.007	0.817	0.366
rs2074196	4.916	0.027	0.736	0.391	0.154	0.695

Genotype frequency differences between East Asian and African populations for T2D associated SNPs

An objective of the association studies is to find SNPs, which differentiate two (or more) studied cohorts. The cohorts differ with respect to one particular trait, typically a disease state. First, the two groups of individuals, the healthy ones and the ones diagnosed with the disease of interest, are genotyped or sequenced. Then, for each SNP, the allelic or genotypic frequencies are compared between the two groups. Such comparison allows identifying SNPs, which show significant differences between the groups or, in other words, are 'associated' with disease status, and can thus serve as disease markers. Chi-square test is one of the statistical tests, which can be used to assess the significance of the observed differences.

In this study we apply the SQL implementation of the chi-square test to test the significance of the differences in genotypic frequencies between the two extreme population groups, East Asian and African (we perform an association test using ethnicity as a trait). We used the previously calculated genotype counts observed in the East Asian and African populations. We then built a set of queries to calculate the intermediate parameters, such as the expected genotype counts, and finally the chi-square statistic (see Appendix). We used the `pchisq` function of R to compute the corresponding p-values. The results are shown in Table 4. Except for one SNP, rs7756992, all SNPs yield very low p-values, confirming that genotype counts for these SNPs are significantly different between the East Asian and African populations.

TABLE 4. Chi-square test results for T2D associated SNPs. Chi-square test for difference in genotype frequencies between the East Asian and African populations was performed for each T2D associated SNP. 'X^2' denotes the chi-square statistic calculated with Impala query and 'p-value' denotes the p-value calculated in R.

SNP	X^2	P-VALUE
rs7903146	139.94	4.10e-31
rs10811661	184.49	8.66e-41
rs13266634	165.94	9.27e-37
rs4402960	67.60	2.10e-15
rs7754840	48.36	3.15e-11
rs5219	181.63	3.62e-40
rs1111875	239.13	1.19e-52
rs11196205	425.96	3.19e-93
rs8050136	112.92	3.01e-25
rs2237892	85.99	2.13e-19
rs7756992	7.77	2.05e-2
rs2074196	108.12	3.33e-24

Intel Reference Architecture

The Intel Reference Architecture for Genomics Data Analysis (Figure 4) is built from a hardware and software stack that has been optimized to run best on Intel Architecture and can cater to a wide variety of healthcare applications/use cases. The main software components in use were as follows: Cloudera Hadoop Distribution v5.1*[7], Hive* (for schema and ETL), Impala* (In-memory SQL-over-Hadoop), HDFS* (Hadoop Distributed File System), Python*, Python Libraries and R*. All datasets were stored on HDFS and flexible schema built using Hive. The queries were written in Python code and distributed over multiple nodes of the Hadoop cluster using Impyla* library[10] for Python, which allows executing SQL through Impala. The specification of each of the nodes of the 6-node Hadoop cluster is shown in Table 5. Figure 4 shows some of the future healthcare applications that can be developed (gray boxes).

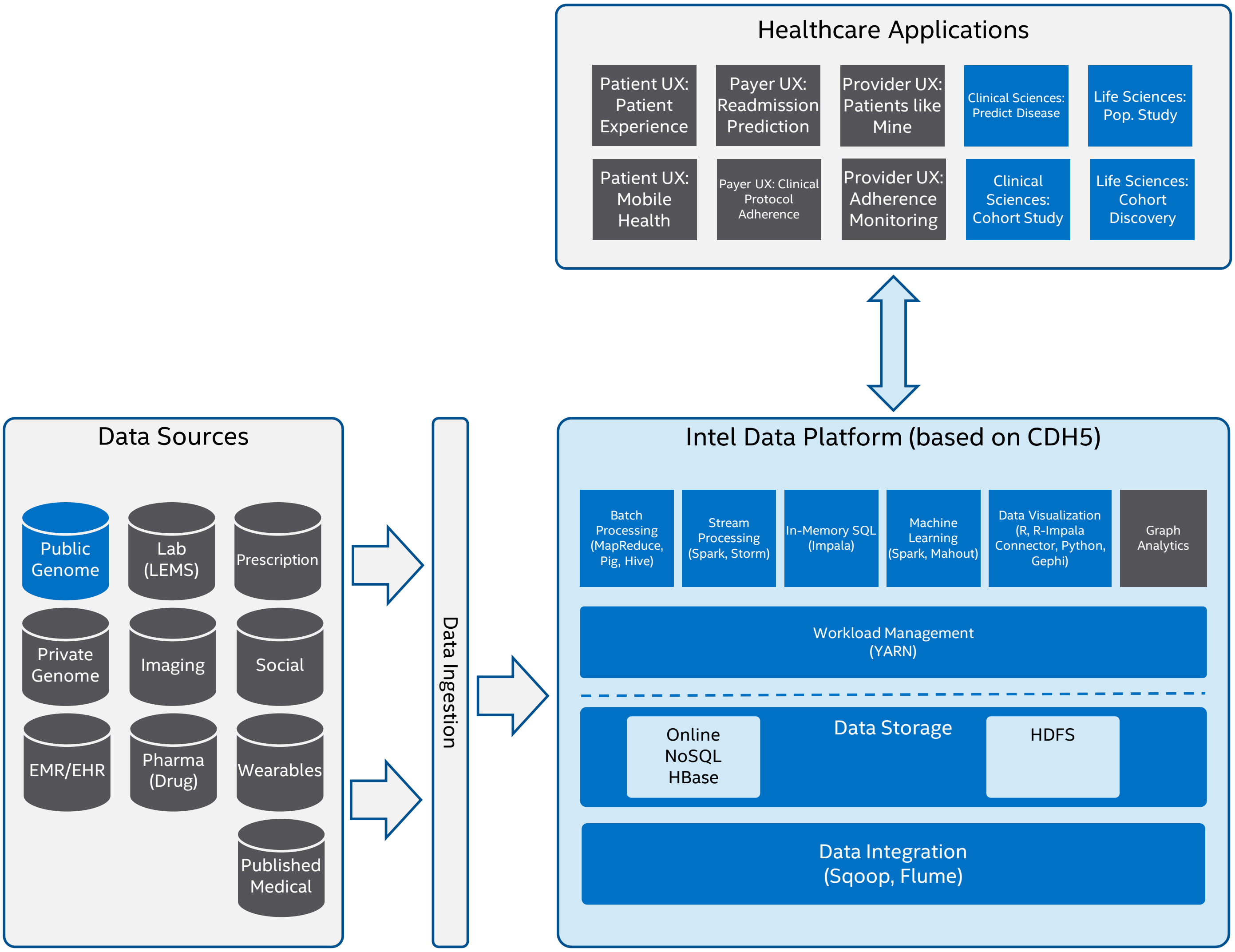
Healthcare Applications
Patient UX: Patient Experience
Payer UX: Readmission Prediction
Provider UX: Patients like Mine
Clinical Sciences: Predict Disease
Life Sciences: Pop. Study
Patient UX: Mobile Health
Payer UX: Clinical Protocol Adherence
Provider UX: Adherence Monitoring
Clinical Sciences: Cohort Study
Life Sciences: Cohort Discovery
Data Sources
Public Genome
Lab (LEMS)
Prescription
Private Genome
Imaging
Social
EMR/EHR
Pharma (Drug)
Wearables
Published Medical
Data Ingestion
Intel Data 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
HDFS
Data Integration (Sqoop, Flume)

Figure 4. Intel Reference Architecture for Genomics Data Analysis. Gray boxes denote future healthcare applications.

TABLE 5. Specification of a single node of the 6-node Hadoop Cluster.

COMPONENT TYPE	PART NAME	QTY
Server Platform	Intel® Server System R2312GZ4GC4 2U 12x3.5 SATA	1
CPU	Intel® Xeon® [8] Processor E5-2680 FC-LGA10 2.7GHz 8.0GT/s 20MB 130W 8 cores CM8062107184424	2
Memory	8GB 1333 Reg ECC 1.5V DDR3 Romley	16
ATA Hard Drive	300GB SSD 2.5in SATA 3Gb/s 25nm Intel Lyndonville SSDSA2BZ300G301 710 Series	1
ATA Hard Drive	2TB HDD 3.5in SATA 6Gb/s 7200 RPM 64MB Seagate Constellation ES ST2000NM0011	12
Network Adapter	NIC – Niantic X520-SR2 10GBase-SR PCI-e Dual Port E10G42BFSR or E10G42BFSRG1P5 Duplex Fiber Optic	1
Chassis Component	Bezel - Intel A2UBEZEL Locking Bezel w/ 2 Branding Clip-on Inserts	1
Power Cord	Pwr Cord - 6ft 14AWG 15A w/ 3 Conductors (C13/5-15P) Black Monoprice 5292	2
Add-in Card	LSI HBA LS100194 (9211-8i) 8 port 6GB/s SATA +SAS PCIe 2.0 Raid LP	2

Dataset and Processing

We used the 1000 Genome Project phase 1[1] dataset consisting of 1,092 individuals, which span four continents and 14 populations, because of the data diversity and population coverage. We used variant calls data, which measures about 1.5 TB when uncompressed (VCF files). The VCF files were copied to HDFS* (Hadoop Distributed File System) and converted to Hive* tables using flexible schema, and queried using Impala* through Python* libraries as described in the Intel Reference Architecture section.

Benefits of Intel Reference Architecture

Whole genome sequencing generates vast amounts of data per person. Depending on how much accuracy is desired (described as coverage), a single human's computer genome representation can start at around 1 TB (fastq format[9] generated from a Sequencer Machine like Illumina*[10]) and go through a software pipeline (e.g. BWA/GATK*[11]) where it is converted to a semi-structured VCF[12] file (about 100 GB) through a process of alignment and variant calling. Even though the VCF file is commonly the start point for downstream genome data analysis, we have to keep in mind that the original versions of genome representations still need to be retained, in case we needed to validate or annotate the data in the future. We can understand now the scale required for such a data management platform which allows us the flexibility to store genome data from millions of people (e.g. Bio Bank[13] infrastructure) and the ability to not have to move data between systems for analysis.

The Intel Reference Architecture provides Hadoop* (HDFS) as the data platform for storage and analysis of genome data. Commodity-level server nodes are cost effective, yet allow large connected storage and processing power on each node (24 TB storage, 128 GB RAM and 2-CPU Intel® Xeon®). When a Hadoop* cluster is built from such nodes, data is distributed over the various nodes (total storage = storage on 1 node x number of data nodes / 3 default replication factor) and processing occurs closest to where the data is located. Future need for more storage and processing capacity can be easily accommodated by 'horizontal scaling', which involves adding more nodes to the cluster and including them into the Hadoop framework. This architecture allows fast interactive SQL queries against vast datasets using a very flexible schema application in the Hive* metastore on any format of data stored on HDFS*. Impala* SQL queries are distributed over multiple nodes and run in Random Access Memory (RAM) and take advantage of RAM on all the Impala nodes as specified in the cluster. Easy integration with data analysis, statistics and visualization tools like R*, Python* and Tableau* allow researchers to integrate diverse sources of data and conduct rapid population and personal genomics research to identify and analyze various disease and patterns. Our intention was to showcase this architecture to an alternative to current High Performance Computing architectures in use today.

Conclusions

In this study, we demonstrate how the Intel Reference Architecture can be applied to quantitative analysis of the genome sequencing data. In particular, we present four different operations frequently performed in the course of such data analysis: allele frequency calculation, genotype counting, test for Hardy-Weinberg equilibrium and chi-square test for differences in genotype counts between populations (see Appendix). We show that these operations can be successfully performed using the Impala in-memory queries. We prove that the Intel Reference Architecture not only provides an integrated platform to store and manipulate the genome sequencing data, but also enables advanced statistical analysis on that data.

References

1. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. "An integrated map of genetic variation from 1,092 human genomes." *Nature*. 2012;491(7422):56-65. doi: 10.1038/nature11632.
2. Chen R, Corona E, Sikora M, Dudley JT, Morgan AA, Moreno-Estrada A, et al. "Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases." *PLoS Genet*. 2012;8(4):e1002621. doi: 10.1371/journal.pgen.1002621.
3. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, et al. "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes." *Nat Genet*. 2012;44(9):981-90. doi: 10.1038/ng.2383.
4. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. "Clinical assessment incorporating a personal genome." *Lancet*. 2010;375(9725):1525-35. doi: 10.1016/S0140-6736(10)60452-7.
5. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, et al. "Personal omics profiling reveals dynamic molecular and medical phenotypes." *Cell*. 2012;148(6):1293-307. doi: 10.1016/j.cell.2012.02.009.
6. Sikora-Wohlfeld W, Basu AK, Butte AJ, Martinez-Canales M. *Accelerating Secondary Genome Analysis Using Intel Reference Architecture*. **http://www.intel.com/content/www/us/en/high-performance-computing/high-performance-xeon-e5-genome-analysis-study.html.2014**.
7. Cloudera Impala. Available at **http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/impala.html**. Accessed on March 16, 2015.
8. Intel Xeon. Available at **http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-e7-family.html**. Accessed on March 16, 2015.
9. Fastq format. Available at **http://en.wikipedia.org/wiki/FASTQ_format**. Accessed on March 23, 2015.
10. Illumina - Whole Genome Sequencing. Available at **http://www.illumina.com/applications/sequencing/dna_sequencing/whole_genome_sequencing.html**. Accessed on March 23, 2015.
11. Genome Analysis Toolkit. Available at **https://www.broadinstitute.org/gatk/index.php**. Accessed on March 23, 2015.
12. Variant Call Format. Available at **http://www.1000genomes.org/node/101**. Accessed on March 23, 2015.
13. Biobank. Available at **http://en.wikipedia.org/wiki/Biobank**. Accessed on March 23, 2015.

Appendix: Python Code

Here we present the Python code written to conduct the calculations described in this study. The Impyla library was used to make connection to the Hadoop Cluster (Impala nodes) and the queries were executed over all the nodes.

```
# Pre-defined objects needed to execute the code

cursor (database cursor)
pop2ind (dictionary with 11 elements, each of them being a
list of column names corresponding to the individuals
belonging to one of the 11 populations)
superpop2ind (dictionary with three elements, each of them
being a list of column names corresponding to the
individuals belonging to a one of the three population
groups: East Asian, European and African)
t2d (table containing genotype data and risk allele
information in the ‘riskal’ column for 12 T2D associated
SNPs)

# Risk allele frequency in 11 populations

query = “CREATE TABLE t2d_raf AS SELECT id, ref, alt,
riskal”

for pop in pop2ind.keys():
     tmp00 = []
     for item in pop2ind[pop]:
         tmp00.append(‘if ‘ + ‘(‘ + item + ‘=\”0|0\”, 2,
0)’)

     tmp11 = []
     for item in pop2ind[pop]:
         tmp11.append(‘if ‘ + ‘(‘ + item + ‘=\”1|1\”, 2,
0)’)

     tmp01 = []
     for item in pop2ind[pop]:
         tmp01.append(‘if ‘ + ‘(‘ + item + ‘=\”0|1\” OR ‘
+ item + ‘=\”1|0\”, 1, 0)’)
     query = query + “, (CASE WHEN ref=riskal THEN (“ + ‘ +
‘.join(tmp00) + “) ELSE (“ + ‘ + ‘.join(tmp11) + “) END + “
+ “(“ +  ‘ + ‘.join(tmp01) + “)) / “ +
str(2*len(pop2ind[pop])) + “ AS “ + pop

query = query + “ FROM t2d”

cursor.execute(query)

# Genotype counts in East Asian, European and African
populations
```

```
for pop in superpop2ind.keys():
     query = “CREATE TABLE t2d _ gen _ ” + pop + “ AS SELECT
     id, ref, alt, riskal”
     tmp00 = []
     for item in superpop2ind[pop]:
         tmp00.append(‘if ‘ + ‘(‘ + item + ‘=\”0|0\”, 1,
0)’)
     tmp11 = []
     for item in superpop2ind[pop]:
         tmp11.append(‘if ‘ + ‘(‘ + item + ‘=\”1|1\”, 1,
0)’)

     tmp01 = []
     for item in superpop2ind[pop]:
         tmp01.append(‘if ‘ + ‘(‘ + item + ‘=\”0|1\” OR ‘
+ item + ‘=\”1|0\”, 1, 0)’)
     query = query + “, “ + ‘ + ‘.join(tmp00) + “ AS
ref _ ref, “ + ‘ + ‘.join(tmp01) + “ AS ref _ alt, “ + ‘ +
‘.join(tmp11) + “ AS alt _ alt FROM t2d”
     cursor.execute(query)

# Chi-square test for Hardy-Weinberg equilibrium in East
Asian, European and African populations

for pop in superpop2ind.keys():
     cursor.execute(“DROP TABLE IF EXISTS t2d _ hwe1 _ ” + pop)
     cursor.execute(“CREATE TABLE t2d _ hwe1 _ ” + pop + “ AS
SELECT id, ref _ ref AS x1, ref _ alt AS x2, alt _ alt AS x3,
ref _ ref+ref _ alt+alt _ alt AS n, 2*ref _ ref+ref _ alt AS a,
2*alt _ alt+ref _ alt AS b FROM t2d _ gen _ ” + pop)
     cursor.execute(“DROP TABLE IF EXISTS t2d _ hwe2 _ ” + pop)
     cursor.execute(“CREATE TABLE t2d _ hwe2 _ ” + pop + “ AS
SELECT *, a/(a+b) AS p, b/(a+b) AS q FROM t2d _ hwe1 _ ” + pop)
     cursor.execute(“DROP TABLE IF EXISTS t2d _ hwe3 _ ” + pop)
     cursor.execute(“CREATE TABLE t2d _ hwe3 _ ” + pop + “ AS
SELECT *, pow(p,2)*n AS ex1, 2*p*q*n AS ex2, pow(q,2)*n AS
ex3 FROM t2d _ hwe2 _ ” + pop)
     cursor.execute(“DROP TABLE IF EXISTS t2d _ hwe _ ” + pop)
     cursor.execute(“CREATE TABLE t2d _ hwe _ ” + pop + “ AS
SELECT *, pow((x1-ex1),2)/ex1 + pow((x2-ex2),2)/ex2 +
pow((x3-ex3),2)/ex3 AS chisq FROM t2d _ hwe3 _ ” + pop)
     cursor.execute(“DROP TABLE t2d _ hwe1 _ ” + pop)
     cursor.execute(“DROP TABLE t2d _ hwe2 _ ” + pop)
     cursor.execute(“DROP TABLE t2d _ hwe3 _ ” + pop)
```

```
# Chi-square test for differences in genotype frequencies
between East Asian and African populations

cursor.execute(“DROP TABLE IF EXISTS t2d _ chisq _ eas”)
cursor.execute(“CREATE TABLE t2d _ chisq _ eas AS SELECT id AS
id _ eas, x1, x2, x3 FROM t2d _ hwe _ eas”)
cursor.execute(“DROP TABLE IF EXISTS t2d _ chisq _ afr”)
cursor.execute(“CREATE TABLE t2d _ chisq _ afr AS SELECT id AS
id _ afr, x1 AS x4, x2 AS x5, x3 AS x6 FROM t2d _ hwe _ afr”)
cursor.execute(“DROP TABLE IF EXISTS t2d _ chisq1”)
cursor.execute(“CREATE TABLE t2d _ chisq1 AS SELECT id _ eas AS
id, x1, x2, x3, x4, x5, x6 FROM t2d _ chisq _ eas e JOIN
t2d _ chisq _ afr a ON (e.id _ eas=a.id _ afr)”)
cursor.execute(“DROP TABLE IF EXISTS t2d _ chisq2”)
cursor.execute(“CREATE TABLE t2d _ chisq2 AS SELECT *,
(x1+x2+x3) * (x1+x4) / (x1+x2+x3+x4+x5+x6) AS ex1,
(x1+x2+x3) * (x2+x5) / (x1+x2+x3+x4+x5+x6) AS ex2,
(x1+x2+x3) * (x3+x6) / (x1+x2+x3+x4+x5+x6) AS ex3,
(x4+x5+x6) * (x1+x4) / (x1+x2+x3+x4+x5+x6) AS ex4,
(x4+x5+x6) * (x2+x5) / (x1+x2+x3+x4+x5+x6) AS ex5,
(x4+x5+x6) * (x3+x6) / (x1+x2+x3+x4+x5+x6) AS ex6 FROM
t2d _ chisq1”)
cursor.execute(“DROP TABLE IF EXISTS t2d _ chisq”)
cursor.execute(“CREATE TABLE t2d _ chisq AS SELECT *,
pow((x1-ex1),2)/ex1 + pow((x2-ex2),2)/ex2 + pow((x3-
ex3),2)/ex3 + pow((x4-ex4),2)/ex4 + pow((x5-ex5),2)/ex5 +
pow((x6-ex6),2)/ex6 AS chi FROM t2d _ chisq2”)
cursor.execute(“DROP TABLE t2d _ chisq1”)
cursor.execute(“DROP TABLE t2d _ chisq2”)
```

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.