

SNP detection and Genome Wide Association Study using Hadoop-BAM, CrossBow and Apache HIVE in Hadoop Cluster

Jyotsna Singh – PGDBD201901006

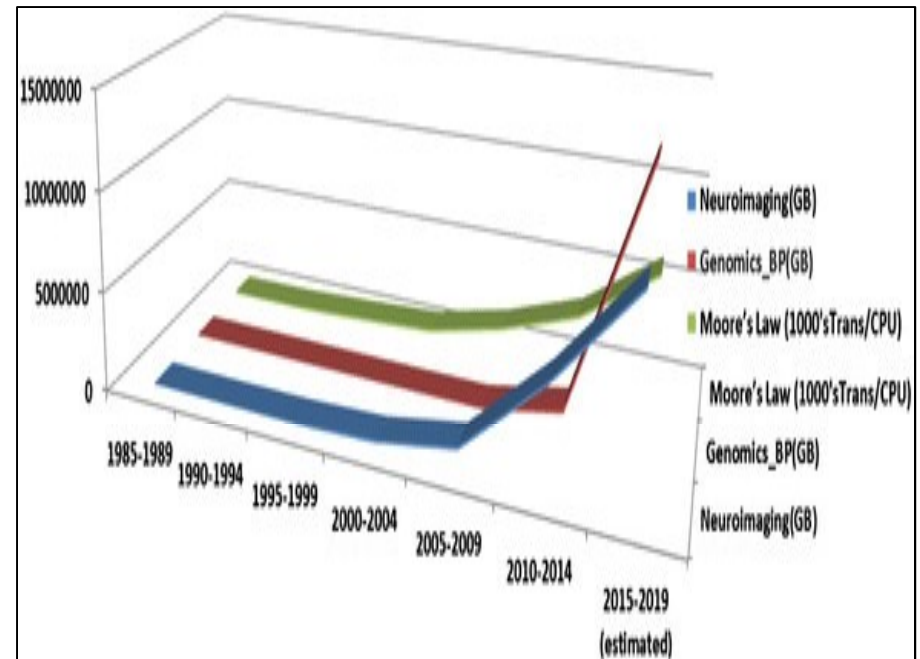
Paulami Das – PGDBD201901009

Poushali Gupta – PGDBD201901010

Institute of Bioinformatics & Applied Biotechnology,
Bangalore

Next Generation Sequencing & Big Data

- The amount of NGS Data worldwide is predicted to double every 5 months which is much faster than Moore's law
- 1000 Genomes project has Petabytes of human genome data sets
- In many GWAS and WGS studies multiple large files have to be processed sequentially



Kryder's law: Exponential growth of neuroimaging and genomics data, relative to increase of number of transistors per chip. By 2025 more than 100 PB of sequenced genome and 1 TB of neuroimaging data will be generated daily.

Different File Formats of Genomic Data

```
@HD VN:1.0 SO:coordinate
@SQ SN:chr20 LN:64444167
@PG ID:TopHat VN:2.0.14 CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-read-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714 16 chr20 190930 3 100M * 0 0
CCGTGTTTAAAGTGATGGCGTCACCTCCACGACTAGGCTTAGTGATTCTAGTTGGCCTAGGAATCCAGCTAGTCCTGTCTCTCAGTCCCCCTCT
C BBDDCCDDCCDDDDDDDDDDCCDCBCB?DDDDDDDDDDDDDDDDDDDDDDCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFDCC@@
AS:i:-15 XM:i:3 XO:i:0 XG:i:0 MD:Z:55C20C13A9 NM:i:3 NH:i:2 CC:Z:= CP:i:55352714 HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961 16 chr20 193953 50 100M * 0 0
TGCTGGATCATCTGTTAGTGCTTCTGACTCAGAGGACCTTCGTCCCTGGGGCAGTGAGCACTTCCAGTGATTCCTGCATAGAAGGGCATGGACGA
G DCDDDDEDDDDDDDDDDDDDDDDDDDDEEC>DFFFEJJJJJIGJJJIHGBHHGIJJJJJJJJJJJJJJJJJJJJHJJJJJJHHHHHHFFFFFCCC
AS:i:-16 XM:i:3 XO:i:0 XG:i:0 MD:Z:60G16T18T3 NM:i:3 NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030 16 chr20 270877 50 100M * 0 0
GGCTTTATTGGTAAAAAGGAATAGCAGATTTAATCAGAAATCCACCTGGGCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGCAGGAAAAAACCA
C DDDDDDDDDCCDDDDDDDDDEEEEEEEFFEFEFFEGHHHFGDJJIHJJIIJJJJIIIGGFJJJIHIIIJJJJJJIGHHFAHGFIJHFGGHFFDD@BB
AS:i:-11 XM:i:2 XO:i:0 XG:i:0 MD:Z:0A85G13 NM:i:2 NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699 0 chr20 271218 50 50M4700N50M * 0
0 GTGGCTCTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGTGCACTTGGGTCTCCGAAGCAGAACATCTCAAATATGACCTCTCG
accepted_hits.sam
```

SAM/BAM Files

ReferenceGenome	Homo_sapiens	(UCS)	(hg19)	Sequence	WholeGenome	FASTA
[Regions]						
Name	Chromosome	Start	End	Upstream Probe Length	Downstream Probe Len	
WASH5P-chr1-14363-14829	chr1	14363	14829	0		
WASH5P-chr1-14970-15038	chr1	14970	15038	0		
WASH5P-chr1-15796-15947	chr1	15796	15947	0		
WASH5P-chr1-16607-16765	chr1	16607	16765	0		
WASH5P-chr1-16858-17055	chr1	16858	17055	0		
WASH5P-chr1-17233-17368	chr1	17233	17368	0		
WASH5P-chr1-17606-17742	chr1	17606	17742	0		

BED Files

```
##fileformat=VCF4.2
##INFO=<ID=SVTYPE,Number=1,Type=String,
Description="Type of structure variant">
##INFO=<ID=END,Number=1,Type=Integer,
Description="End position of the variant described in this record">
#CHROM POS ID REF ALT QUAL FILTER INFO
```

```
1 160929435 rs7520618 G A . SVTYPE=SNP;END=160929436
1 160932043 rs113387749 A . SVTYPE=INS;END=160932043
1 160932206 rs5778188 C . SVTYPE=DEL;END=160932207
1 160932771 rs2256505 A G . SVTYPE=SNP;END=160932772
1 160934077 rs2481074 T A . SVTYPE=SNP;END=160934078
1 160934818 rs1023115 A G . SVTYPE=SNP;END=160934819
1 160935328 . AAA TGC . SVTYPE=SUB;END=160935331
1 160935334 rs75452934 AA TC . SVTYPE=SUB;END=160935336
```

VCF Files

```
>@HWI-ST216_0180:4:1101:1096:2196#GGCTAC/1
TTTTTCAGNGAATACTGCAAATCAATAAACTCTTTAG
>@HWI-ST216_0180:4:1101:1158:2236#GGCTAC/1
AAAAGCTCATTTCCTATAGTTAACAGGACATGCCTT
>@HWI-ST216_0180:4:1101:1448:2211#GGCTAC/1
ATTATATAAGATAGCGGCTTTTTCCGTTAGTTTCCT
>@HWI-ST216_0180:4:1101:1331:2227#GGCTAC/1
CACGTTCTCTGTCCCCAATGGTATTTGCATCCCTGT
>@HWI-ST216_0180:4:1101:1376:2237#GGCTAC/1
GCGTCCCTTAGCTGAACTACCCAAACGTACGAATGC
```

Fasta Files

```
@HWI-ST216_0180:4:1101:1096:2196#GGCTAC/1
TTTTCAGNGAATACTGCAAATCAATAAACTCTTTAG
+HWI-ST216_0180:4:1101:1096:2196#GGCTAC/1
ceedb]]B[[]]] [fffff\dddddedeedf_fbd
@HWI-ST216_0180:4:1101:1158:2236#GGCTAC/1
AAAAGCTCATTTCTATAGTTAACAGGACATGCCTT
+HWI-ST216_0180:4:1101:1158:2236#GGCTAC/1
ggggggggggggggggggggggggfffggggggggggfg
@HWI-ST216_0180:4:1101:1448:2211#GGCTAC/1
ATTATATAAGATAGCGGCTTTTCCGTTAGTTTCT
```

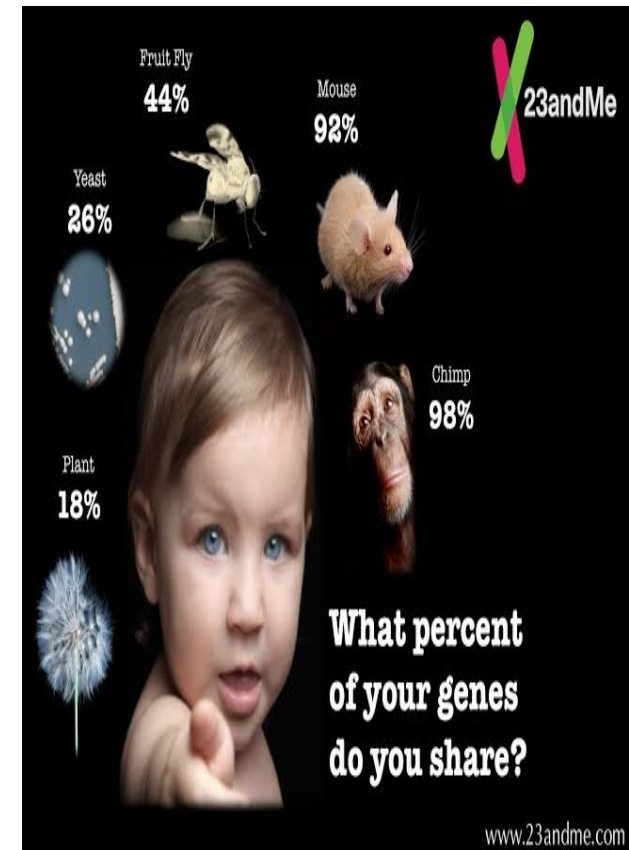
Fastq Files

Sequence	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr1	hg19_msk	exon	83886030	83886750	980	-	gene_id	"ERV1-E-nt";
chr1	hg19_msk	exon	192937729	192938064	491	+	gene_id	"MLT1";
chr1	hg19_msk	exon	242039324	242039681	7249	+	gene_id	"HLB1";
chr1	hg19_msk	exon	10485172	104866981	8848	-	gene_id	"LTR12C";
chr1	hg19_msk	exon	53477283	534773596	1005	+	gene_id	"MER21C";
chr1	hg19_msk	exon	74448847	74449973	9303	-	gene_id	"HLB1-nt";
chr1	hg19_msk	exon	82837212	82837259	461	-	gene_id	"ManRep152";
chr1	hg19_msk	exon	102760404	102760706	1754	+	gene_id	"THSD1";
chr1	hg19_msk	exon	145751968	145752401	1532	-	gene_id	"LTR47B";
chr1	hg19_msk	exon	153091848	153092533	4373	-	gene_id	"LTR78";
chr1	hg19_msk	exon	160423210	160425454	848	+	gene_id	"LTR85C";
chr1	hg19_msk	exon	210763724	210764168	1368	-	gene_id	"MLT28A";
chr1	hg19_msk	exon	211812124	211812354	1068	-	gene_id	"MER21A";
chr1	hg19_msk	exon	213809477	213809708	1378	-	gene_id	"MLT1F2";
chr1	hg19_msk	exon	238026629	238027031	816	+	gene_id	"MLT1M";
chr1	hg19_msk	exon	241172200	241172531	3210	-	gene_id	"LTR24C";
chr1	hg19_msk	exon	262105	262386	721	+	gene_id	"MER65D";
chr1	hg19_msk	exon	3538719	353968	2001	-	gene_id	"HLB1D";
chr1	hg19_msk	exon	4324984	4325379	1304	-	gene_id	"MLT1B";
chr1	hg19_msk	exon	5103832	5112709	28994	+	gene_id	"HERVH-nt";
chr1	hg19_msk	exon	12713599	12714399	808	-	gene_id	"MER52-nt";
chr1	hg19_msk	exon	12840257	12845900	27617	-	gene_id	"HERVK-nt";

GTF Files

Properties of Our Data Set

- Semi-Structured
- Reference Genome – Fasta Format
3.2 GB
62743362 bp
- Raw Data – Fastq Format
4.1 GB
27999799 bp
Sequence length - 36 bp
GC Content – 44%
- Output File – BAM Format, VCF Format
42 GB, 30 KB



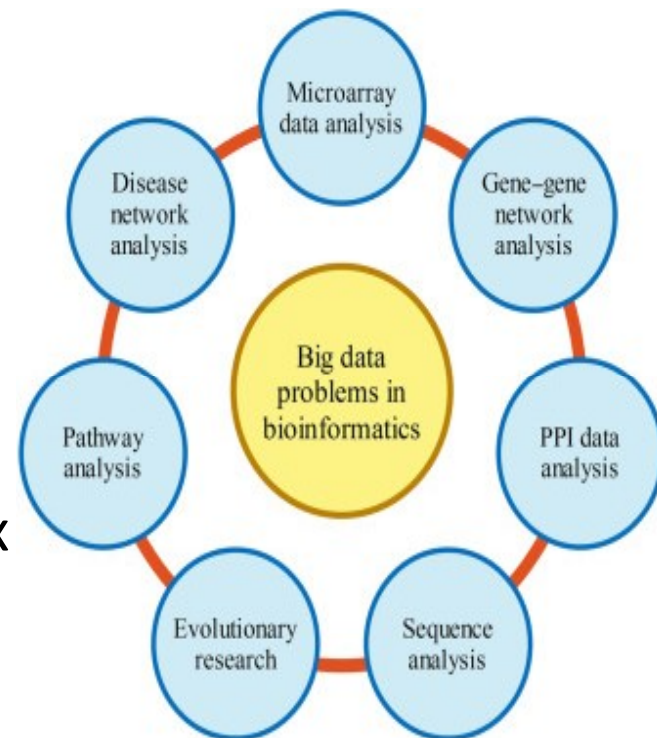
Relationship of Human
Genome with Other
Species

Advantages of Hadoop

- **Hadoop is Open Source distributed data processing system**
- Based on Google's **MapReduce** architecture design
- Cheap commodity hardware for storage
- Fault tolerant distributed filesystems: **HDFS**
- Batch processing systems: **Hadoop MapReduce, Apache Hive, Apache Pig (HDD), Apache Spark (RAM)**
- Parallel SQL implementations for analytics: **Apache Hive, Cloudera Impala, Apache Spark**
- Fault tolerant distributed database: **Hbase**
- **Distributed machine learning libraries, text indexing & search**

Hadoop in Different Biological Aspects

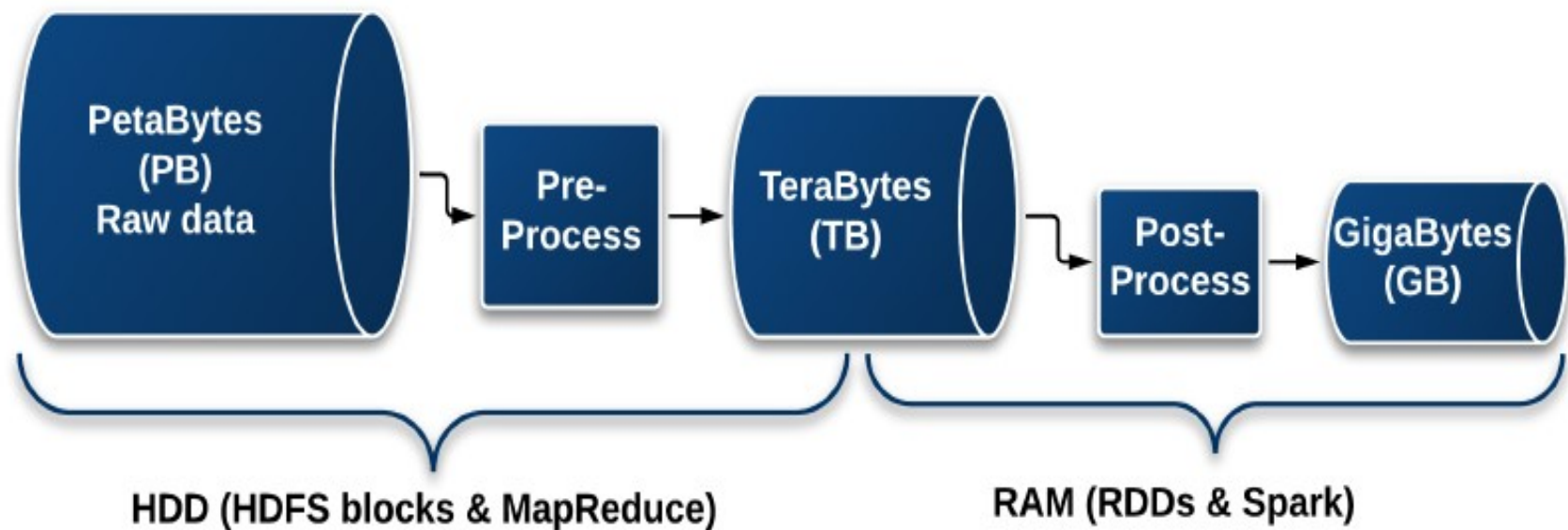
- In Cancer treatments
- In monitoring Patient Vitals
- In the Hospital Network
- In Healthcare Intelligence
- In Structural Bioinformatics –
 - 1) Molecular Docking
 - 2) Clustering of Protein-Ligand complex
 - 3) Structural Alignment
- In Genomic Data Analysis



Tools used in Hadoop For Biological Data Analysis

- **Cloud Burst** – Uses Hadoop as a platform for alignment of short reads.
- **Crossbow** – Uses Hadoop for SNP genotyping from short reads.
- **Contrail** – Uses Hadoop for denovo assembly from short sequencing reads
- **Myrna** – Uses Bowtie and R/Bioconductor for calculating differential gene expression from large RNASeq data sets
- **Cloud Blast** – Uses Gene Set Enrichment Analysis in Hadoop
- **BlueSNP** - Implements GWAS statistical tests in R & executes the calculations with Apache Hadoop using MapReduce formalism.
- **HadoopBAM** – A library for processing NGS data format in parallel with both Hadoop and Spark.
- **Amazon Elastic Compute Cloud & MapReduce.**

Typical Genomics Data Analysis Using HDFS



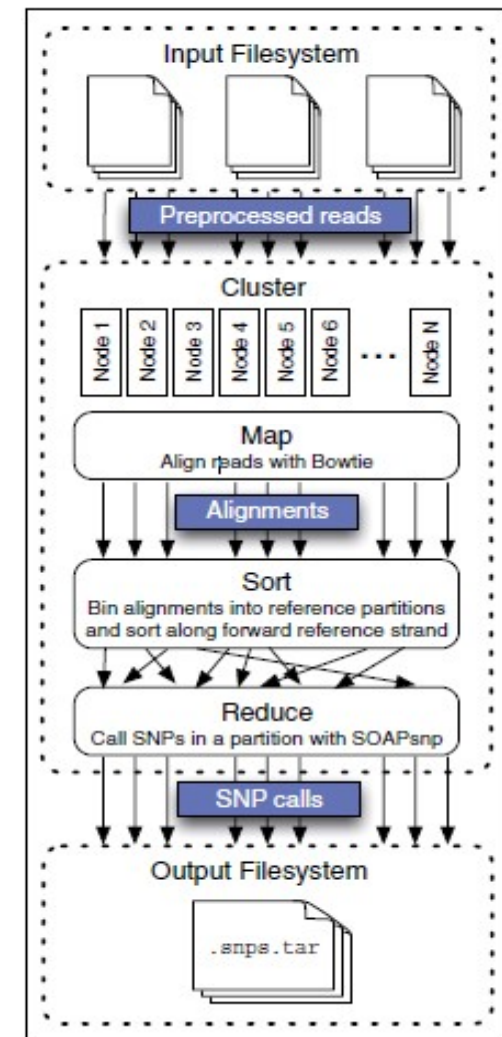
Processing Data in main memory instead of files in hard disks = minimal I/O operations. Map/Reduce data from Petabytes to Gigabytes (million times less in the end)

Project Proposal

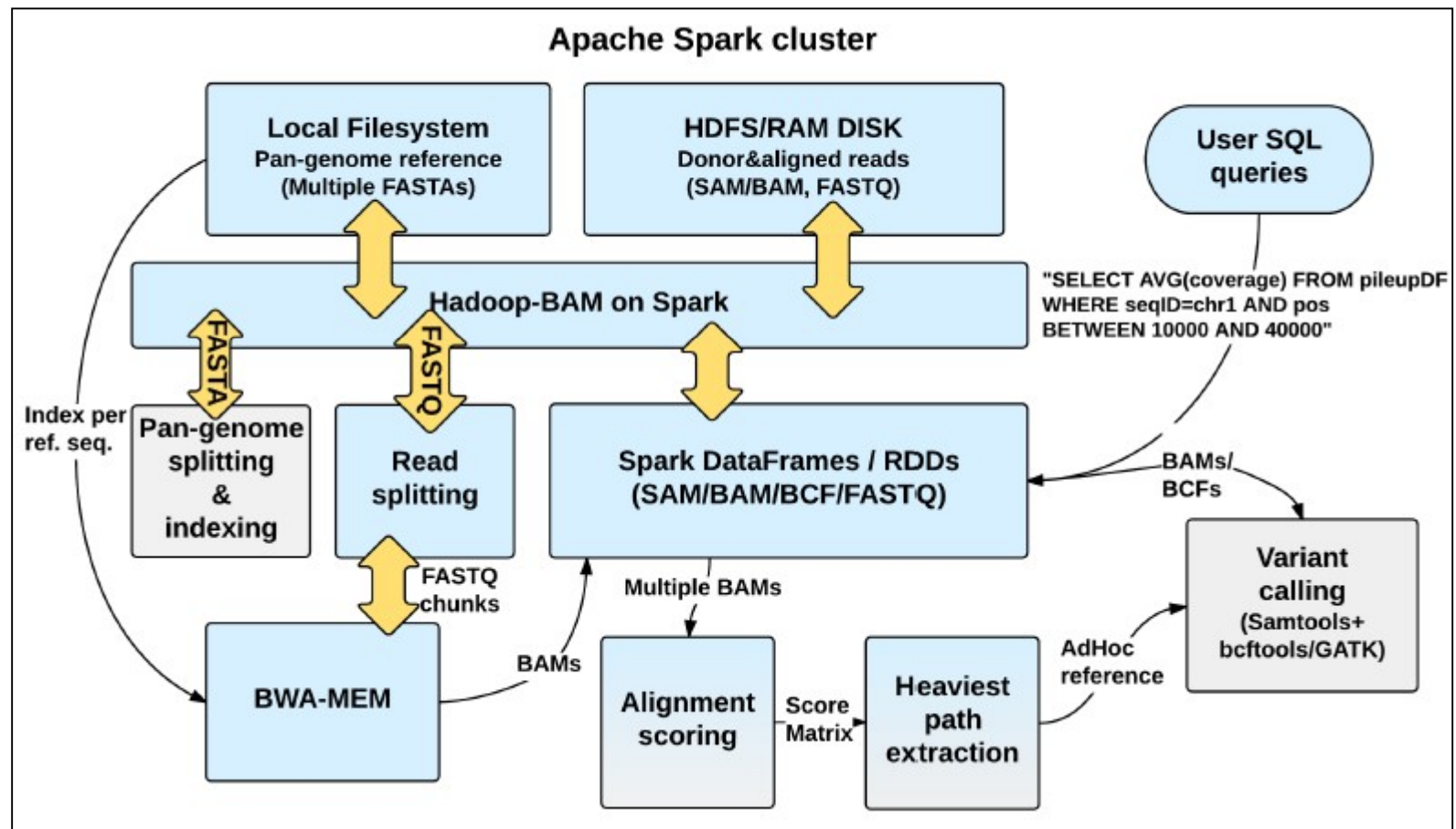
- Genome Alignment using HADOOP-BAM
- SNP Detection using Crossbow (Bowtie+SOAPsnp) and HadoopBAM and comparing both
- Genome Wide Association Study (GWAS) using Apache HIVE across Human Genome of Different Population

SNP Detection using MapReduce Algorithm in Crossbow

- Copying the Fastq raw data and Fasta reference genome from Local File System to HDFS
- Running the Crossbow pipeline in Hadoop Cluster
- Crossbow's Map phase align reads with Bowtie 2 which employs a compact index of reference sequence requiring about 3 GB of memory using HG19
- The index is distributed to all computers in cluster via hadoop file or by instructing each node to independently obtain the index from a shared file
- The reduce phase performs SOAPsnp
- The output of Reduce phase is SNP tuple which stored on the Clustered distributed File System which can be transferred to Local File System.

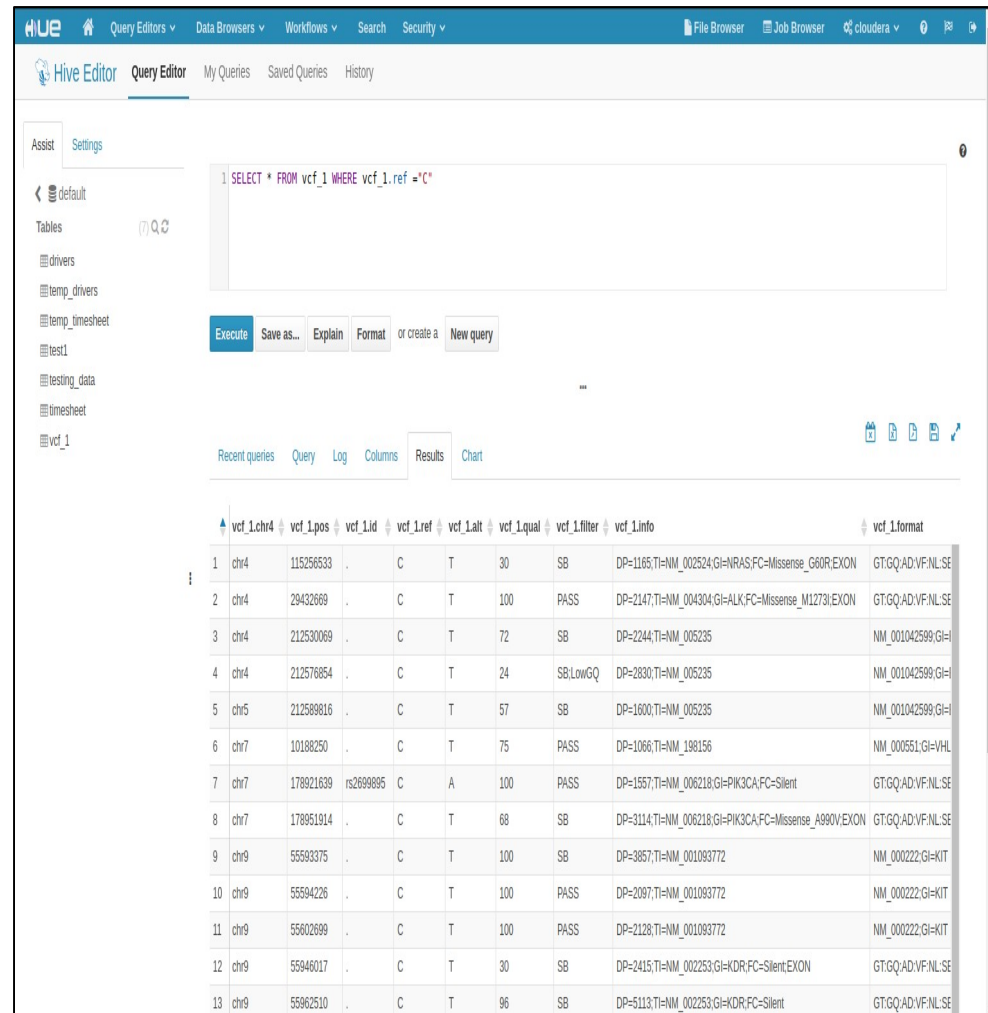


SNP Detection using HadoopBAM



Genome Wide Association Study using Apache HIVE

- Processing of VCF Files in Data Browser and query using Apache HIVE
- Counting the Allele Frequency
- Taking Input Data from different population and finding Genome wide association using Log odds ratio/Likelihood ratio/Chi-square test across different population

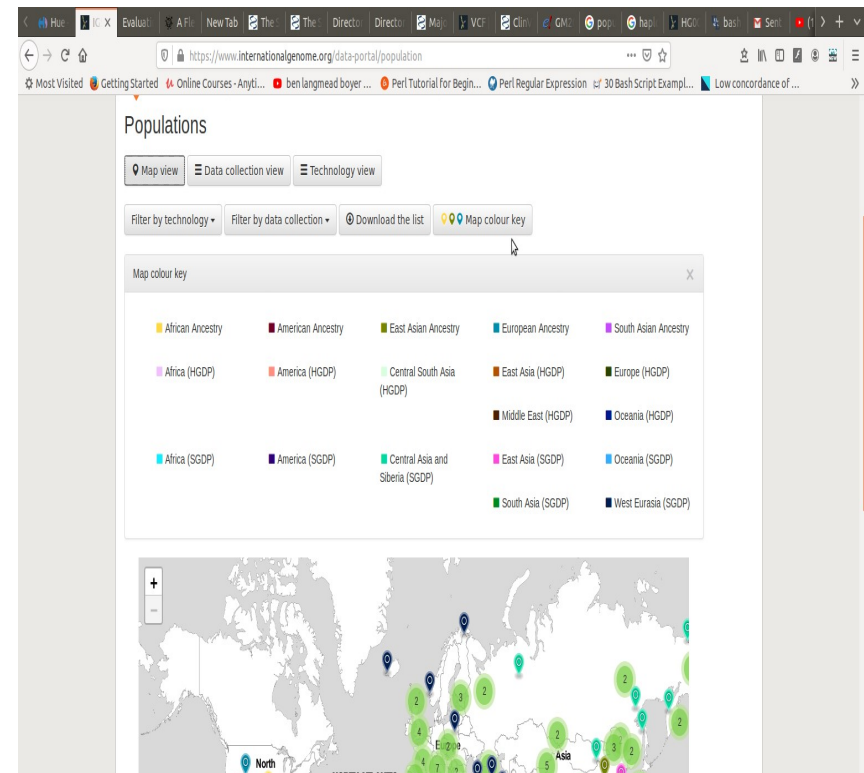
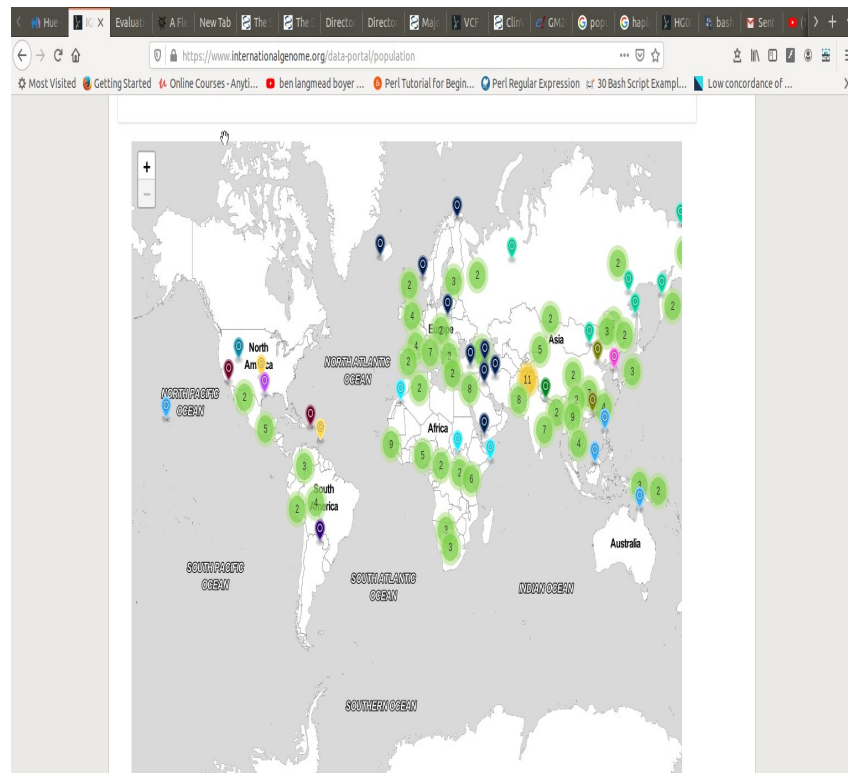


The screenshot shows the Apache HUE interface. At the top, there's a navigation bar with 'HUE' logo and various menu items like 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. Below this, the 'Hive Editor' tab is active, displaying a query editor with the text: `SELECT * FROM vcf_1 WHERE vcf_1.ref = 'C'`. Below the editor are buttons for 'Execute', 'Save as...', 'Explain', 'Format', and 'New query'. The 'Results' tab is selected, showing a table of query results. The table has columns: vcf_1.chr4, vcf_1.pos, vcf_1.id, vcf_1.ref, vcf_1.alt, vcf_1.qual, vcf_1.filter, vcf_1.info, and vcf_1.format. The results are listed in 13 rows, showing genomic data for chromosome 4 and 5.

	vcf_1.chr4	vcf_1.pos	vcf_1.id	vcf_1.ref	vcf_1.alt	vcf_1.qual	vcf_1.filter	vcf_1.info	vcf_1.format
1	chr4	115256533	.	C	T	30	SB	DP=1165;TI=NM_002524;GI=NRAS;FC=Missense_G60R;EXON	GT:GQ:AD:VF:NL:SE
2	chr4	29432669	.	C	T	100	PASS	DP=2147;TI=NM_004304;GI=ALK;FC=Missense_M1273I;EXON	GT:GQ:AD:VF:NL:SE
3	chr4	212530069	.	C	T	72	SB	DP=2244;TI=NM_005235	NM_001042599;GI=
4	chr4	212576854	.	C	T	24	SB,LowGQ	DP=2830;TI=NM_005235	NM_001042599;GI=
5	chr5	212589816	.	C	T	57	SB	DP=1600;TI=NM_005235	NM_001042599;GI=
6	chr7	10188250	.	C	T	75	PASS	DP=1066;TI=NM_198156	NM_000551;GI=VHL
7	chr7	178921639	rs2699895	C	A	100	PASS	DP=1557;TI=NM_006218;GI=PIK3CA;FC=Silent	GT:GQ:AD:VF:NL:SE
8	chr7	178951814	.	C	T	68	SB	DP=3114;TI=NM_006218;GI=PIK3CA;FC=Missense_A980V;EXON	GT:GQ:AD:VF:NL:SE
9	chr9	55593375	.	C	T	100	SB	DP=3857;TI=NM_001093772	NM_000222;GI=KIT
10	chr9	55594226	.	C	T	100	PASS	DP=2097;TI=NM_001093772	NM_000222;GI=KIT
11	chr9	55602699	.	C	T	100	PASS	DP=2128;TI=NM_001093772	NM_000222;GI=KIT
12	chr9	55946017	.	C	T	30	SB	DP=2415;TI=NM_002253;GI=KDR;FC=Silent;EXON	GT:GQ:AD:VF:NL:SE
13	chr9	55962510	.	C	T	96	SB	DP=5113;TI=NM_002253;GI=KDR;FC=Silent	GT:GQ:AD:VF:NL:SE

SNP Detection using Crossbow

- Downloaded fastq files from 1000 Genomes Project of different populations



Work Flow

Installed Crossbow, Bowtie, soapSNP, HadoopBAM



Indexed using Bowtie and Detected SNPs using soapSNP

```
-T <FILE> Only call consensus on regions specified in FILE. Format: ChrName\tStart\tEnd
ibab@IBAB-PGDBD-Comp07:~/Applications/AM_Assignment/GENOMICS/20feb/AlignmentFiles$ soapnp -B cancer_sorted.bam -o out
Authorized!
Illumina Fastq System Set
*****
*                               *
*      Software name :SOApsnp    *
*      Version       : v 1.05    *
*      Last Update   : 2011.9.25 *
*      Author:Core Development Group *
*      Fan Zhang & Bill Tang      *
*      E_mail : zhangfan@genomics.org.cn *
*      zhouguoyue@genomics.org.cn *
*      Copyright : BGI. All Rights Reserved. *
*****
Parameters:
-l <String> Input SORTED Soap Result or list
-s <String> Input SORTED SAM Result or list
-B <String> Input SORTED BAM Result or list
-l Specify this option will enable the alignments' filelist input
-P <Int> Set the thread number[1]
-d <String> Reference Sequence in fasta format
-o <String> (<DIR> if input is file list) Output consensus file Optional Parameters:(Default in [])
-z <char> ASCII character standing for quality=0 [0]
-g <Double> Global Error Dependency Coefficient, 0.0(complete dependent)-1.0(complete independent)[0.9]
-p <Double> PCR Error Dependency Coefficient, 0.0(complete dependent)-1.0(complete independent)[0.5]
-r <Double> novel althom prior probability [0.0005]
-e <Double> novel HET prior probability [0.0010]
-t set transition/transversion ratio to 2:1 in prior probability
-s <String> Pre-formated dbSNP information
-z2 specify this option will REFINE SNPs using dbSNPs information [Off]
-a <Double> Validated HET prior, if no allele frequency known [0.1]
-b <Double> Validated althom prior, if no allele frequency known [0.05]
-j <Double> Unvalidated HET prior, if no allele frequency known [0.02]
-k <Double> Unvalidated althom rate, if no allele frequency known [0.01]
-u Enable rank sum test to give HET further penalty for better accuracy [Off]
-m Enable monoploid calling mode, this will ensure all consensus as HOM and you probably should SPECIFY
  higher althom rate [Off]
-q Only output potential SNPs. Useful in Text output mode [Off]
-M <FILE>(<DIR> if set -l parameter) Output the quality calibration matrix;the matrix can be reused with -I
  if you rerun the program
-I <FILE>(<FILELIST> if set -l parameter) Input previous quality calibration matrix. It cannot be used
  simultaneously with -M
-L <Short> maximum length of read [45]
-Q <Short> maximum FASTQ quality score, using ASCII character standing for quality[h]
-F <Int> Output format. 0: Text; 1: GLEV2; 2: GPFV2.[0]
  0: Text; 1: GLEV2; 2: GPFV2; 3: GPFV2; 4: GPFV2; 5: GPFV2; 6: GPFV2; 7: GPFV2; 8: GPFV2; 9: GPFV2; 10: GPFV2; 11: GPFV2; 12: GPFV2; 13: GPFV2; 14: GPFV2; 15: GPFV2; 16: GPFV2; 17: GPFV2; 18: GPFV2; 19: GPFV2; 20: GPFV2; 21: GPFV2; 22: GPFV2; 23: GPFV2; 24: GPFV2; 25: GPFV2; 26: GPFV2; 27: GPFV2; 28: GPFV2; 29: GPFV2; 30: GPFV2; 31: GPFV2; 32: GPFV2; 33: GPFV2; 34: GPFV2; 35: GPFV2; 36: GPFV2; 37: GPFV2; 38: GPFV2; 39: GPFV2; 40: GPFV2; 41: GPFV2; 42: GPFV2; 43: GPFV2; 44: GPFV2; 45: GPFV2; 46: GPFV2; 47: GPFV2; 48: GPFV2; 49: GPFV2; 50: GPFV2; 51: GPFV2; 52: GPFV2; 53: GPFV2; 54: GPFV2; 55: GPFV2; 56: GPFV2; 57: GPFV2; 58: GPFV2; 59: GPFV2; 60: GPFV2; 61: GPFV2; 62: GPFV2; 63: GPFV2; 64: GPFV2; 65: GPFV2; 66: GPFV2; 67: GPFV2; 68: GPFV2; 69: GPFV2; 70: GPFV2; 71: GPFV2; 72: GPFV2; 73: GPFV2; 74: GPFV2; 75: GPFV2; 76: GPFV2; 77: GPFV2; 78: GPFV2; 79: GPFV2; 80: GPFV2; 81: GPFV2; 82: GPFV2; 83: GPFV2; 84: GPFV2; 85: GPFV2; 86: GPFV2; 87: GPFV2; 88: GPFV2; 89: GPFV2; 90: GPFV2; 91: GPFV2; 92: GPFV2; 93: GPFV2; 94: GPFV2; 95: GPFV2; 96: GPFV2; 97: GPFV2; 98: GPFV2; 99: GPFV2; 100: GPFV2; 101: GPFV2; 102: GPFV2; 103: GPFV2; 104: GPFV2; 105: GPFV2; 106: GPFV2; 107: GPFV2; 108: GPFV2; 109: GPFV2; 110: GPFV2; 111: GPFV2; 112: GPFV2; 113: GPFV2; 114: GPFV2; 115: GPFV2; 116: GPFV2; 117: GPFV2; 118: GPFV2; 119: GPFV2; 120: GPFV2; 121: GPFV2; 122: GPFV2; 123: GPFV2; 124: GPFV2; 125: GPFV2; 126: GPFV2; 127: GPFV2; 128: GPFV2; 129: GPFV2; 130: GPFV2; 131: GPFV2; 132: GPFV2; 133: GPFV2; 134: GPFV2; 135: GPFV2; 136: GPFV2; 137: GPFV2; 138: GPFV2; 139: GPFV2; 140: GPFV2; 141: GPFV2; 142: GPFV2; 143: GPFV2; 144: GPFV2; 145: GPFV2; 146: GPFV2; 147: GPFV2; 148: GPFV2; 149: GPFV2; 150: GPFV2; 151: GPFV2; 152: GPFV2; 153: GPFV2; 154: GPFV2; 155: GPFV2; 156: GPFV2; 157: GPFV2; 158: GPFV2; 159: GPFV2; 160: GPFV2; 161: GPFV2; 162: GPFV2; 163: GPFV2; 164: GPFV2; 165: GPFV2; 166: GPFV2; 167: GPFV2; 168: GPFV2; 169: GPFV2; 170: GPFV2; 171: GPFV2; 172: GPFV2; 173: GPFV2; 174: GPFV2; 175: GPFV2; 176: GPFV2; 177: GPFV2; 178: GPFV2; 179: GPFV2; 180: GPFV2; 181: GPFV2; 182: GPFV2; 183: GPFV2; 184: GPFV2; 185: GPFV2; 186: GPFV2; 187: GPFV2; 188: GPFV2; 189: GPFV2; 190: GPFV2; 191: GPFV2; 192: GPFV2; 193: GPFV2; 194: GPFV2; 195: GPFV2; 196: GPFV2; 197: GPFV2; 198: GPFV2; 199: GPFV2; 200: GPFV2; 201: GPFV2; 202: GPFV2; 203: GPFV2; 204: GPFV2; 205: GPFV2; 206: GPFV2; 207: GPFV2; 208: GPFV2; 209: GPFV2; 210: GPFV2; 211: GPFV2; 212: GPFV2; 213: GPFV2; 214: GPFV2; 215: GPFV2; 216: GPFV2; 217: GPFV2; 218: GPFV2; 219: GPFV2; 220: GPFV2; 221: GPFV2; 222: GPFV2; 223: GPFV2; 224: GPFV2; 225: GPFV2; 226: GPFV2; 227: GPFV2; 228: GPFV2; 229: GPFV2; 230: GPFV2; 231: GPFV2; 232: GPFV2; 233: GPFV2; 234: GPFV2; 235: GPFV2; 236: GPFV2; 237: GPFV2; 238: GPFV2; 239: GPFV2; 240: GPFV2; 241: GPFV2; 242: GPFV2; 243: GPFV2; 244: GPFV2; 245: GPFV2; 246: GPFV2; 247: GPFV2; 248: GPFV2; 249: GPFV2; 250: GPFV2; 251: GPFV2; 252: GPFV2; 253: GPFV2; 254: GPFV2; 255: GPFV2; 256: GPFV2; 257: GPFV2; 258: GPFV2; 259: GPFV2; 260: GPFV2; 261: GPFV2; 262: GPFV2; 263: GPFV2; 264: GPFV2; 265: GPFV2; 266: GPFV2; 267: GPFV2; 268: GPFV2; 269: GPFV2; 270: GPFV2; 271: GPFV2; 272: GPFV2; 273: GPFV2; 274: GPFV2; 275: GPFV2; 276: GPFV2; 277: GPFV2; 278: GPFV2; 279: GPFV2; 280: GPFV2; 281: GPFV2; 282: GPFV2; 283: GPFV2; 284: GPFV2; 285: GPFV2; 286: GPFV2; 287: GPFV2; 288: GPFV2; 289: GPFV2; 290: GPFV2; 291: GPFV2; 292: GPFV2; 293: GPFV2; 294: GPFV2; 295: GPFV2; 296: GPFV2; 297: GPFV2; 298: GPFV2; 299: GPFV2; 300: GPFV2; 301: GPFV2; 302: GPFV2; 303: GPFV2; 304: GPFV2; 305: GPFV2; 306: GPFV2; 307: GPFV2; 308: GPFV2; 309: GPFV2; 310: GPFV2; 311: GPFV2; 312: GPFV2; 313: GPFV2; 314: GPFV2; 315: GPFV2; 316: GPFV2; 317: GPFV2; 318: GPFV2; 319: GPFV2; 320: GPFV2; 321: GPFV2; 322: GPFV2; 323: GPFV2; 324: GPFV2; 325: GPFV2; 326: GPFV2; 327: GPFV2; 328: GPFV2; 329: GPFV2; 330: GPFV2; 331: GPFV2; 332: GPFV2; 333: GPFV2; 334: GPFV2; 335: GPFV2; 336: GPFV2; 337: GPFV2; 338: GPFV2; 339: GPFV2; 340: GPFV2; 341: GPFV2; 342: GPFV2; 343: GPFV2; 344: GPFV2; 345: GPFV2; 346: GPFV2; 347: GPFV2; 348: GPFV2; 349: GPFV2; 350: GPFV2; 351: GPFV2; 352: GPFV2; 353: GPFV2; 354: GPFV2; 355: GPFV2; 356: GPFV2; 357: GPFV2; 358: GPFV2; 359: GPFV2; 360: GPFV2; 361: GPFV2; 362: GPFV2; 363: GPFV2; 364: GPFV2; 365: GPFV2; 366: GPFV2; 367: GPFV2; 368: GPFV2; 369: GPFV2; 370: GPFV2; 371: GPFV2; 372: GPFV2; 373: GPFV2; 374: GPFV2; 375: GPFV2; 376: GPFV2; 377: GPFV2; 378: GPFV2; 379: GPFV2; 380: GPFV2; 381: GPFV2; 382: GPFV2; 383: GPFV2; 384: GPFV2; 385: GPFV2; 386: GPFV2; 387: GPFV2; 388: GPFV2; 389: GPFV2; 390: GPFV2; 391: GPFV2; 392: GPFV2; 393: GPFV2; 394: GPFV2; 395: GPFV2; 396: GPFV2; 397: GPFV2; 398: GPFV2; 399: GPFV2; 400: GPFV2; 401: GPFV2; 402: GPFV2; 403: GPFV2; 404: GPFV2; 405: GPFV2; 406: GPFV2; 407: GPFV2; 408: GPFV2; 409: GPFV2; 410: GPFV2; 411: GPFV2; 412: GPFV2; 413: GPFV2; 414: GPFV2; 415: GPFV2; 416: GPFV2; 417: GPFV2; 418: GPFV2; 419: GPFV2; 420: GPFV2; 421: GPFV2; 422: GPFV2; 423: GPFV2; 424: GPFV2; 425: GPFV2; 426: GPFV2; 427: GPFV2; 428: GPFV2; 429: GPFV2; 430: GPFV2; 431: GPFV2; 432: GPFV2; 433: GPFV2; 434: GPFV2; 435: GPFV2; 436: GPFV2; 437: GPFV2; 438: GPFV2; 439: GPFV2; 440: GPFV2; 441: GPFV2; 442: GPFV2; 443: GPFV2; 444: GPFV2; 445: GPFV2; 446: GPFV2; 447: GPFV2; 448: GPFV2; 449: GPFV2; 450: GPFV2; 451: GPFV2; 452: GPFV2; 453: GPFV2; 454: GPFV2; 455: GPFV2; 456: GPFV2; 457: GPFV2; 458: GPFV2; 459: GPFV2; 460: GPFV2; 461: GPFV2; 462: GPFV2; 463: GPFV2; 464: GPFV2; 465: GPFV2; 466: GPFV2; 467: GPFV2; 468: GPFV2; 469: GPFV2; 470: GPFV2; 471: GPFV2; 472: GPFV2; 473: GPFV2; 474: GPFV2; 475: GPFV2; 476: GPFV2; 477: GPFV2; 478: GPFV2; 479: GPFV2; 480: GPFV2; 481: GPFV2; 482: GPFV2; 483: GPFV2; 484: GPFV2; 485: GPFV2; 486: GPFV2; 487: GPFV2; 488: GPFV2; 489: GPFV2; 490: GPFV2; 491: GPFV2; 492: GPFV2; 493: GPFV2; 494: GPFV2; 495: GPFV2; 496: GPFV2; 497: GPFV2; 498: GPFV2; 499: GPFV2; 500: GPFV2; 501: GPFV2; 502: GPFV2; 503: GPFV2; 504: GPFV2; 505: GPFV2; 506: GPFV2; 507: GPFV2; 508: GPFV2; 509: GPFV2; 510: GPFV2; 511: GPFV2; 512: GPFV2; 513: GPFV2; 514: GPFV2; 515: GPFV2; 516: GPFV2; 517: GPFV2; 518: GPFV2; 519: GPFV2; 520: GPFV2; 521: GPFV2; 522: GPFV2; 523: GPFV2; 524: GPFV2; 525: GPFV2; 526: GPFV2; 527: GPFV2; 528: GPFV2; 529: GPFV2; 530: GPFV2; 531: GPFV2; 532: GPFV2; 533: GPFV2; 534: GPFV2; 535: GPFV2; 536: GPFV2; 537: GPFV2; 538: GPFV2; 539: GPFV2; 540: GPFV2; 541: GPFV2; 542: GPFV2; 543: GPFV2; 544: GPFV2; 545: GPFV2; 546: GPFV2; 547: GPFV2; 548: GPFV2; 549: GPFV2; 550: GPFV2; 551: GPFV2; 552: GPFV2; 553: GPFV2; 554: GPFV2; 555: GPFV2; 556: GPFV2; 557: GPFV2; 558: GPFV2; 559: GPFV2; 560: GPFV2; 561: GPFV2; 562: GPFV2; 563: GPFV2; 564: GPFV2; 565: GPFV2; 566: GPFV2; 567: GPFV2; 568: GPFV2; 569: GPFV2; 570: GPFV2; 571: GPFV2; 572: GPFV2; 573: GPFV2; 574: GPFV2; 575: GPFV2; 576: GPFV2; 577: GPFV2; 578: GPFV2; 579: GPFV2; 580: GPFV2; 581: GPFV2; 582: GPFV2; 583: GPFV2; 584: GPFV2; 585: GPFV2; 586: GPFV2; 587: GPFV2; 588: GPFV2; 589: GPFV2; 590: GPFV2; 591: GPFV2; 592: GPFV2; 593: GPFV2; 594: GPFV2; 595: GPFV2; 596: GPFV2; 597: GPFV2; 598: GPFV2; 599: GPFV2; 600: GPFV2; 601: GPFV2; 602: GPFV2; 603: GPFV2; 604: GPFV2; 605: GPFV2; 606: GPFV2; 607: GPFV2; 608: GPFV2; 609: GPFV2; 610: GPFV2; 611: GPFV2; 612: GPFV2; 613: GPFV2; 614: GPFV2; 615: GPFV2; 616: GPFV2; 617: GPFV2; 618: GPFV2; 619: GPFV2; 620: GPFV2; 621: GPFV2; 622: GPFV2; 623: GPFV2; 624: GPFV2; 625: GPFV2; 626: GPFV2; 627: GPFV2; 628: GPFV2; 629: GPFV2; 630: GPFV2; 631: GPFV2; 632: GPFV2; 633: GPFV2; 634: GPFV2; 635: GPFV2; 636: GPFV2; 637: GPFV2; 638: GPFV2; 639: GPFV2; 640: GPFV2; 641: GPFV2; 642: GPFV2; 643: GPFV2; 644: GPFV2; 645: GPFV2; 646: GPFV2; 647: GPFV2; 648: GPFV2; 649: GPFV2; 650: GPFV2; 651: GPFV2; 652: GPFV2; 653: GPFV2; 654: GPFV2; 655: GPFV2; 656: GPFV2; 657: GPFV2; 658: GPFV2; 659: GPFV2; 660: GPFV2; 661: GPFV2; 662: GPFV2; 663: GPFV2; 664: GPFV2; 665: GPFV2; 666: GPFV2; 667: GPFV2; 668: GPFV2; 669: GPFV2; 670: GPFV2; 671: GPFV2; 672: GPFV2; 673: GPFV2; 674: GPFV2; 675: GPFV2; 676: GPFV2; 677: GPFV2; 678: GPFV2; 679: GPFV2; 680: GPFV2; 681: GPFV2; 682: GPFV2; 683: GPFV2; 684: GPFV2; 685: GPFV2; 686: GPFV2; 687: GPFV2; 688: GPFV2; 689: GPFV2; 690: GPFV2; 691: GPFV2; 692: GPFV2; 693: GPFV2; 694: GPFV2; 695: GPFV2; 696: GPFV2; 697: GPFV2; 698: GPFV2; 699: GPFV2; 700: GPFV2; 701: GPFV2; 702: GPFV2; 703: GPFV2; 704: GPFV2; 705: GPFV2; 706: GPFV2; 707: GPFV2; 708: GPFV2; 709: GPFV2; 710: GPFV2; 711: GPFV2; 712: GPFV2; 713: GPFV2; 714: GPFV2; 715: GPFV2; 716: GPFV2; 717: GPFV2; 718: GPFV2; 719: GPFV2; 720: GPFV2; 721: GPFV2; 722: GPFV2; 723: GPFV2; 724: GPFV2; 725: GPFV2; 726: GPFV2; 727: GPFV2; 728: GPFV2; 729: GPFV2; 730: GPFV2; 731: GPFV2; 732: GPFV2; 733: GPFV2; 734: GPFV2; 735: GPFV2; 736: GPFV2; 737: GPFV2; 738: GPFV2; 739: GPFV2; 740: GPFV2; 741: GPFV2; 742: GPFV2; 743: GPFV2; 744: GPFV2; 745: GPFV2; 746: GPFV2; 747: GPFV2; 748: GPFV2; 749: GPFV2; 750: GPFV2; 751: GPFV2; 752: GPFV2; 753: GPFV2; 754: GPFV2; 755: GPFV2; 756: GPFV2; 757: GPFV2; 758: GPFV2; 759: GPFV2; 760: GPFV2; 761: GPFV2; 762: GPFV2; 763: GPFV2; 764: GPFV2; 765: GPFV2; 766: GPFV2; 767: GPFV2; 768: GPFV2; 769: GPFV2; 770: GPFV2; 771: GPFV2; 772: GPFV2; 773: GPFV2; 774: GPFV2; 775: GPFV2; 776: GPFV2; 777: GPFV2; 778: GPFV2; 779: GPFV2; 780: GPFV2; 781: GPFV2; 782: GPFV2; 783: GPFV2; 784: GPFV2; 785: GPFV2; 786: GPFV2; 787: GPFV2; 788: GPFV2; 789: GPFV2; 790: GPFV2; 791: GPFV2; 792: GPFV2; 793: GPFV2; 794: GPFV2; 795: GPFV2; 796: GPFV2; 797: GPFV2; 798: GPFV2; 799: GPFV2; 800: GPFV2; 801: GPFV2; 802: GPFV2; 803: GPFV2; 804: GPFV2; 805: GPFV2; 806: GPFV2; 807: GPFV2; 808: GPFV2; 809: GPFV2; 810: GPFV2; 811: GPFV2; 812: GPFV2; 813: GPFV2; 814: GPFV2; 815: GPFV2; 816: GPFV2; 817: GPFV2; 818: GPFV2; 819: GPFV2; 820: GPFV2; 821: GPFV2; 822: GPFV2; 823: GPFV2; 824: GPFV2; 825: GPFV2; 826: GPFV2; 827: GPFV2; 828: GPFV2; 829: GPFV2; 830: GPFV2; 831: GPFV2; 832: GPFV2; 833: GPFV2; 834: GPFV2; 835: GPFV2; 836: GPFV2; 837: GPFV2; 838: GPFV2; 839: GPFV2; 840: GPFV2; 841: GPFV2; 842: GPFV2; 843: GPFV2; 844: GPFV2; 845: GPFV2; 846: GPFV2; 847: GPFV2; 848: GPFV2; 849: GPFV2; 850: GPFV2; 851: GPFV2; 852: GPFV2; 853: GPFV2; 854: GPFV2; 855: GPFV2; 856: GPFV2; 857: GPFV2; 858: GPFV2; 859: GPFV2; 860: GPFV2; 861: GPFV2; 862: GPFV2; 863: GPFV2; 864: GPFV2; 865: GPFV2; 866: GPFV2; 867: GPFV2; 868: GPFV2; 869: GPFV2; 870: GPFV2; 871: GPFV2; 872: GPFV2; 873: GPFV2; 874: GPFV2; 875: GPFV2; 876: GPFV2; 877: GPFV2; 878: GPFV2; 879: GPFV2; 880: GPFV2; 881: GPFV2; 882: GPFV2; 883: GPFV2; 884: GPFV2; 885: GPFV2; 886: GPFV2; 887: GPFV2; 888: GPFV2; 889: GPFV2; 890: GPFV2; 891: GPFV2; 892: GPFV2; 893: GPFV2; 894: GPFV2; 895: GPFV2; 896: GPFV2; 897: GPFV2; 898: GPFV2; 899: GPFV2; 900: GPFV2; 901: GPFV2; 902: GPFV2; 903: GPFV2; 904: GPFV2; 905: GPFV2; 906: GPFV2; 907: GPFV2; 908: GPFV2; 909: GPFV2; 910: GPFV2; 911: GPFV2; 912: GPFV2; 913: GPFV2; 914: GPFV2; 915: GPFV2; 916: GPFV2; 917: GPFV2; 918: GPFV2; 919: GPFV2; 920: GPFV2; 921: GPFV2; 922: GPFV2; 923: GPFV2; 924: GPFV2; 925: GPFV2; 926: GPFV2; 927: GPFV2; 928: GPFV2; 929: GPFV2; 930: GPFV2; 931: GPFV2; 932: GPFV2; 933: GPFV2; 934: GPFV2; 935: GPFV2; 936: GPFV2; 937: GPFV2; 938: GPFV2; 939: GPFV2; 940: GPFV2; 941: GPFV2; 942: GPFV2; 943: GPFV2; 944: GPFV2; 945: GPFV2; 946: GPFV2; 947: GPFV2; 948: GPFV2; 949: GPFV2; 950: GPFV2; 951: GPFV2; 952: GPFV2; 953: GPFV2; 954: GPFV2; 955: GPFV2; 956: GPFV2; 957: GPFV2; 958: GPFV2; 959: GPFV2; 960: GPFV2; 961: GPFV2; 962: GPFV2; 963: GPFV2; 964: GPFV2; 965: GPFV2; 966: GPFV2; 967: GPFV2; 968: GPFV2; 969: GPFV2; 970: GPFV2; 971: GPFV2; 972: GPFV2; 973: GPFV2; 974: GPFV2; 975: GPFV2; 976: GPFV2; 977: GPFV2; 978: GPFV2; 979: GPFV2; 980: GPFV2; 981: GPFV2; 982: GPFV2; 983: GPFV2; 984: GPFV2; 985: GPFV2; 986: GPFV2; 987: GPFV2; 988: GPFV2; 989: GPFV2; 990: GPFV2; 991: GPFV2; 992: GPFV2; 993: GPFV2; 994: GPFV2; 995: GPFV2; 996: GPFV2; 997: GPFV2; 998: GPFV2; 999: GPFV2; 1000: GPFV2; 1001: GPFV2; 1002: GPFV2; 1003: GPFV2; 1004: GPFV2; 1005: GPFV2; 1006: GPFV2; 1007: GPFV2; 1008: GPFV2; 1009: GPFV2; 1010: GPFV2; 1011: GPFV2; 1012: GPFV2; 1013: GPFV2; 1014: GPFV2; 1015: GPFV2; 1016: GPFV2; 1017: GPFV2; 1018: GPFV2; 1019: GPFV2; 1020: GPFV2; 1021: GPFV2; 1022: GPFV2; 1023: GPFV2; 1024: GPFV2; 1025: GPFV2; 1026: GPFV2; 1027: GPFV2; 1028: GPFV2; 1029: GPFV2; 1030: GPFV2; 1031: GPFV2; 1032: GPFV2; 1033: GPFV2; 1034: GPFV2; 1035: GPFV2; 1036: GPFV2; 1037: GPFV2; 1038: GPFV2; 1039: GPFV2; 1040: GPFV2; 1041: GPFV2; 1042: GPFV2; 1043: GPFV2; 1044: GPFV2; 1045: GPFV2; 1046: GPFV2; 1047: GPFV2; 1048: GPFV2; 1049: GPFV2; 1050: GPFV2; 1051: GPFV2; 1052: GPFV2; 1053: GPFV2; 1054: GPFV2; 1055: GPFV2; 1056: GPFV2; 1057: GPFV2; 1058: GPFV2; 1059: GPFV2; 1060: GPFV2; 1061: GPFV2; 1062: GPFV2; 1063: GPFV2; 1064: GPFV2; 1065: GPFV2; 1066: GPFV2; 1067: GPFV2; 1068: GPFV2; 1069: GPFV2; 1070: GPFV2; 1071: GPFV2; 1072: GPFV2; 1073: GPFV2; 1074: GPFV2; 1075: GPFV2; 1076: GPFV2; 1077: GPFV2; 1078: GPFV2; 1079: GPFV2; 1080: GPFV2; 1081: GPFV2; 1082: GPFV2; 1083: GPFV2; 1084: GPFV2; 1085: GPFV2; 1086: GPFV2; 1087: GPFV2; 1088: GPFV2; 1089: GPFV2; 1090: GPFV2; 1091: GPFV2; 1092: GPFV2; 1093: GPFV2; 1094: GPFV2; 1095: GPFV2; 1096: GPFV2; 1097: GPFV2; 1098: GPFV2; 1099: GPFV2; 1100: GPFV2; 1101: GPFV2; 1102: GPFV2; 1103: GPFV2; 1104: GPFV2; 1105: GPFV2; 1106: GPFV2; 1107: GPFV2; 1108: GPFV2; 1109: GPFV2; 1110: GPFV2; 1111: GPFV2; 1112: GPFV2; 1113: GPFV2; 1114: GPFV2; 1115: GPFV2; 1116: GPFV2; 1117: GPFV2; 1118: GPFV2; 1119: GPFV2; 1120: GPFV2; 1121: GPFV2; 1122: GPFV2; 1123: GPFV2; 1124: GPFV2; 1125: GPFV2; 1126: GPFV2; 1127: GPFV2; 1128: GPFV2; 1129: GPFV2; 1130: GPFV2; 1131: GPFV2; 1132: GPFV2; 1133: G
```


GWAS using Apache Hive – Flow Chart

- From the generated VCF files we chose only two populations – 1) Bengalis in Bangladesh
2) British in Scotland
- Next we preprocessed the VCF files
- Randomly we chose chromosome 20 for further analysis

Preprocessing of VCF files

- Selection of lines which contain the information about SNP

```
ibab@IBAB-PGDBD-Comp07:~/SecondSemester/VCF$ cat ALL.chr20.phase3_shapeit2_mvnc  
all_integrated_v5a.20130502.genotypes.vcf | grep "^20" > british_in_scotland.csv
```

- Extracted relevant columns : chrom, pos, ref, alt, info, format

```
ibab@IBAB-PGDBD-Comp07:~/SecondSemester/VCF$ awk '{print $1,$2,$4,$5,$8,$9}' br  
itish_in_scotland.csv > british_in_scotland_1.csv
```

Continued...

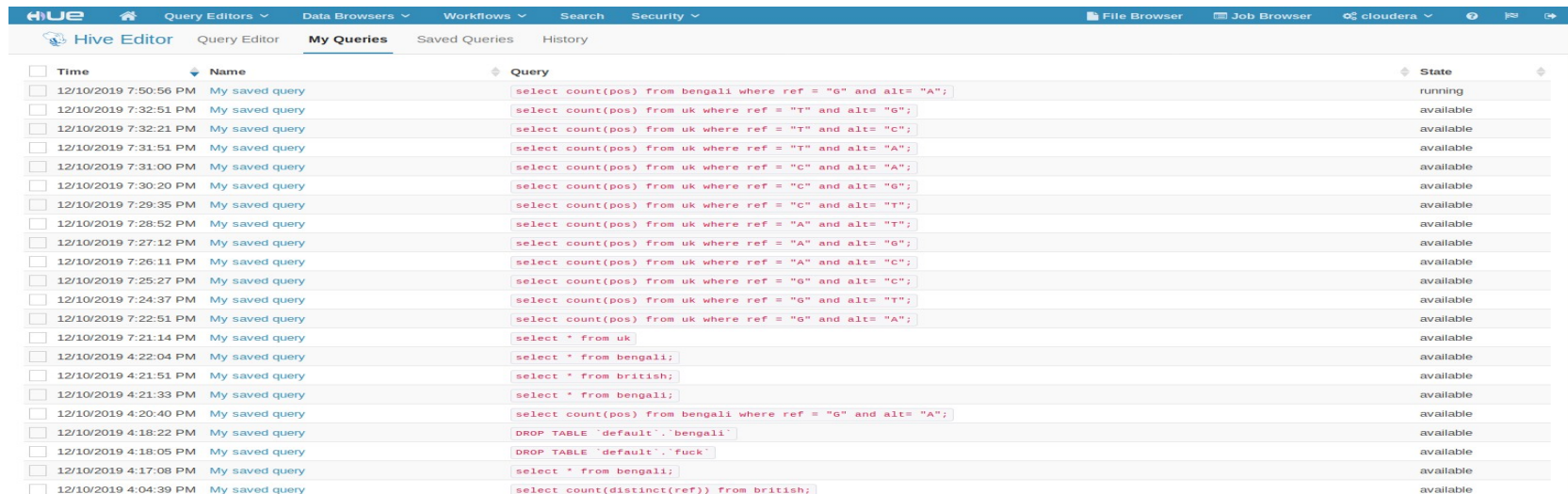
- Uploaded processed VCF files in Apache Hive

The screenshot displays the Hive Editor interface. The top navigation bar includes 'Hive Editor', 'Query Editor', 'My Queries', 'Saved Queries', and 'History'. The left sidebar shows a 'Tables' list with entries like 'bengali', 'british', 'drivers', 'newdrivers', 'newtimesheet', 'temp_drivers', 'temp_timesheet', 'testing', and 'uk'. The main query editor area contains the SQL query: `1 select * from uk`. Below the query editor are buttons for 'Execute', 'Save', 'Save as...', 'Explain', 'Format', and 'New query'. The 'Results' tab is active, showing a table with 12 rows of VCF data. The table has columns: 'uk.chrom', 'uk.pos', 'uk.ref', 'uk.alt', and 'uk.info'. The data represents genomic variants from the UK population.

	uk.chrom	uk.pos	uk.ref	uk.alt	uk.info
1	20	60343	G	A	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=20377;EAS_AF=0;AMR_AF=0.0014;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=.;VT=S
2	20	60419	A	G	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=19865;EAS_AF=0;AMR_AF=0;AFR_AF=0;EUR_AF=0;SAS_AF=0.001;AA=.;VT=SN
3	20	60479	C	T	AC=17;AF=0.00339457;AN=5008;NS=2504;DP=20218;EAS_AF=0;AMR_AF=0.0043;AFR_AF=0.0106;EUR_AF=0;SAS_AF=0;AA=.;VT=
4	20	60522	T	TC	AC=68;AF=0.0135783;AN=5008;NS=2504;DP=20754;EAS_AF=0;AMR_AF=0.0029;AFR_AF=0.0499;EUR_AF=0;SAS_AF=0;AA=.;VT=
5	20	60568	A	C	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=20728;EAS_AF=0;AMR_AF=0;AFR_AF=0.0008;EUR_AF=0;SAS_AF=0;AA=.;VT=S
6	20	60571	C	A	AC=10;AF=0.00199681;AN=5008;NS=2504;DP=20683;EAS_AF=0;AMR_AF=0.0014;AFR_AF=0.0068;EUR_AF=0;SAS_AF=0;AA=.;VT=
7	20	60579	G	A	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=20396;EAS_AF=0.001;AMR_AF=0;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=.;VT=SN
8	20	60649	A	G	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=20484;EAS_AF=0;AMR_AF=0.0014;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=.;VT=S
9	20	60778	A	G	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=21261;EAS_AF=0.001;AMR_AF=0;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=.;VT=SN
10	20	60795	G	C	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=21333;EAS_AF=0;AMR_AF=0;AFR_AF=0;EUR_AF=0.001;SAS_AF=0;AA=.;VT=SN
11	20	60808	G	A	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=21348;EAS_AF=0;AMR_AF=0;AFR_AF=0.0008;EUR_AF=0;SAS_AF=0;AA=.;VT=S
12	20	60810	G	GA	AC=4;AF=0.000798722;AN=5008;NS=2504;DP=21358;EAS_AF=0;AMR_AF=0.0058;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=.;VT=unknow

Queries in Hive

- Queried for all possible base changes : purine to purine (A -> G) , pyrimidine to pyrimidine (T -> C) , purine to pyrimidine(A ->T, A->C, G->T,G->C and vice versa)



The screenshot shows the Hive Editor interface with a table of queries. The table has four columns: Time, Name, Query, and State. The queries are listed in chronological order from top to bottom, showing various SQL commands used for data analysis and table management.

Time	Name	Query	State
12/10/2019 7:50:56 PM	My saved query	<code>select count(pos) from bengali where ref = "G" and alt= "A";</code>	running
12/10/2019 7:32:51 PM	My saved query	<code>select count(pos) from uk where ref = "T" and alt= "G";</code>	available
12/10/2019 7:32:21 PM	My saved query	<code>select count(pos) from uk where ref = "T" and alt= "C";</code>	available
12/10/2019 7:31:51 PM	My saved query	<code>select count(pos) from uk where ref = "T" and alt= "A";</code>	available
12/10/2019 7:31:00 PM	My saved query	<code>select count(pos) from uk where ref = "C" and alt= "A";</code>	available
12/10/2019 7:30:20 PM	My saved query	<code>select count(pos) from uk where ref = "C" and alt= "G";</code>	available
12/10/2019 7:29:35 PM	My saved query	<code>select count(pos) from uk where ref = "C" and alt= "T";</code>	available
12/10/2019 7:28:52 PM	My saved query	<code>select count(pos) from uk where ref = "A" and alt= "T";</code>	available
12/10/2019 7:27:12 PM	My saved query	<code>select count(pos) from uk where ref = "A" and alt= "G";</code>	available
12/10/2019 7:26:11 PM	My saved query	<code>select count(pos) from uk where ref = "A" and alt= "C";</code>	available
12/10/2019 7:25:27 PM	My saved query	<code>select count(pos) from uk where ref = "G" and alt= "C";</code>	available
12/10/2019 7:24:37 PM	My saved query	<code>select count(pos) from uk where ref = "G" and alt= "T";</code>	available
12/10/2019 7:22:51 PM	My saved query	<code>select count(pos) from uk where ref = "G" and alt= "A";</code>	available
12/10/2019 7:21:14 PM	My saved query	<code>select * from uk</code>	available
12/10/2019 4:22:04 PM	My saved query	<code>select * from bengali;</code>	available
12/10/2019 4:21:51 PM	My saved query	<code>select * from british;</code>	available
12/10/2019 4:21:33 PM	My saved query	<code>select * from bengali;</code>	available
12/10/2019 4:20:40 PM	My saved query	<code>select count(pos) from bengali where ref = "G" and alt= "A";</code>	available
12/10/2019 4:18:22 PM	My saved query	<code>DROP TABLE 'default'. 'bengali';</code>	available
12/10/2019 4:18:05 PM	My saved query	<code>DROP TABLE 'default'. 'fuck';</code>	available
12/10/2019 4:17:08 PM	My saved query	<code>select * from bengali;</code>	available
12/10/2019 4:04:39 PM	My saved query	<code>select count(distinct(ref)) from british;</code>	available

Frequency Table

Reference Allele	Alternative Allele	Count
T	A	53024
A	T	53692
C	G	72349
G	C	71959
C	T	373871
T	C	210950
T	G	56328
A	G	219805
A	C	56018
C	A	79395
G	T	81070
G	A	377981

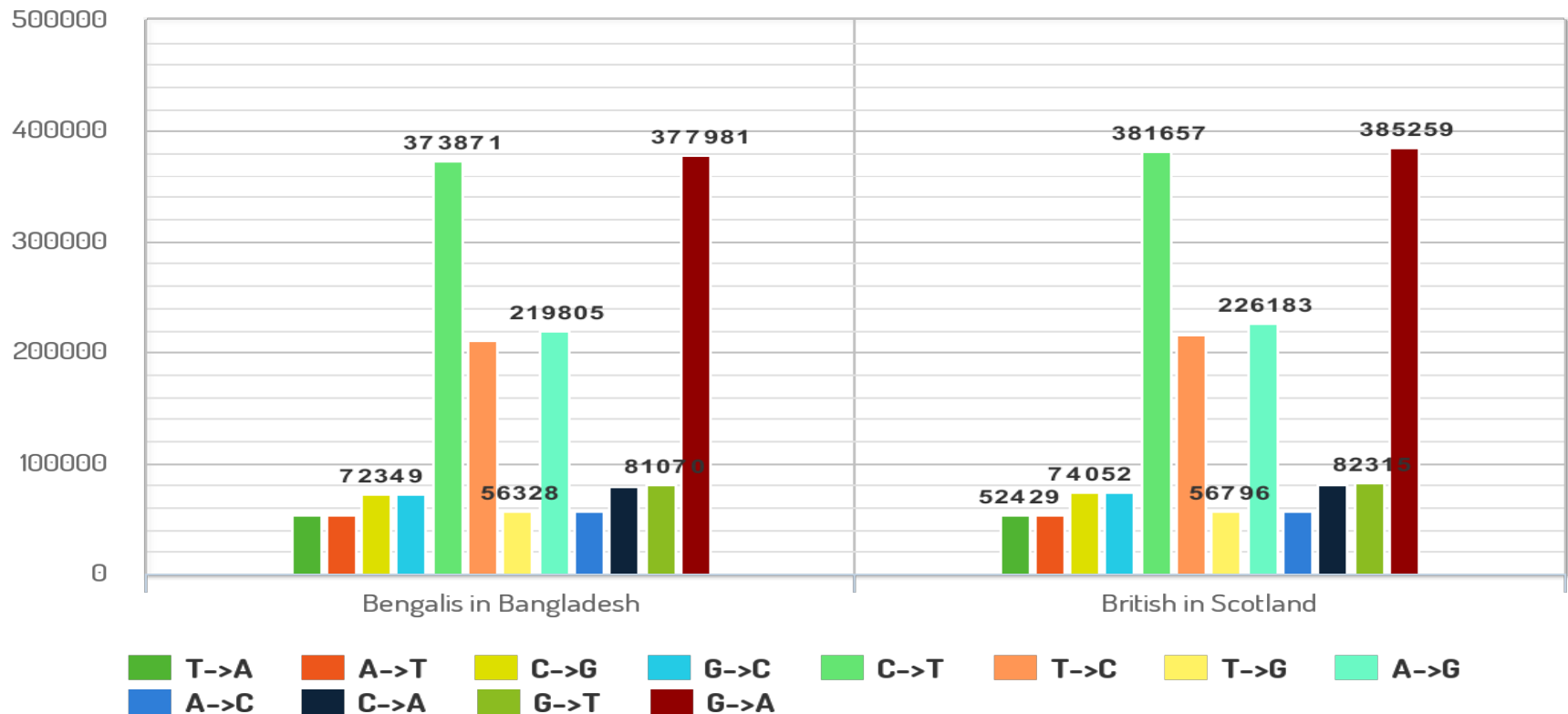
Benagali in Bangladesh

Reference Allele	Alternative Allele	Count
T	A	52429
A	T	53218
C	G	74052
G	C	73639
C	T	381657
T	C	216606
T	G	56796
A	G	226183
A	C	56717
C	A	80439
G	T	82315
G	A	385264

British in Scotland

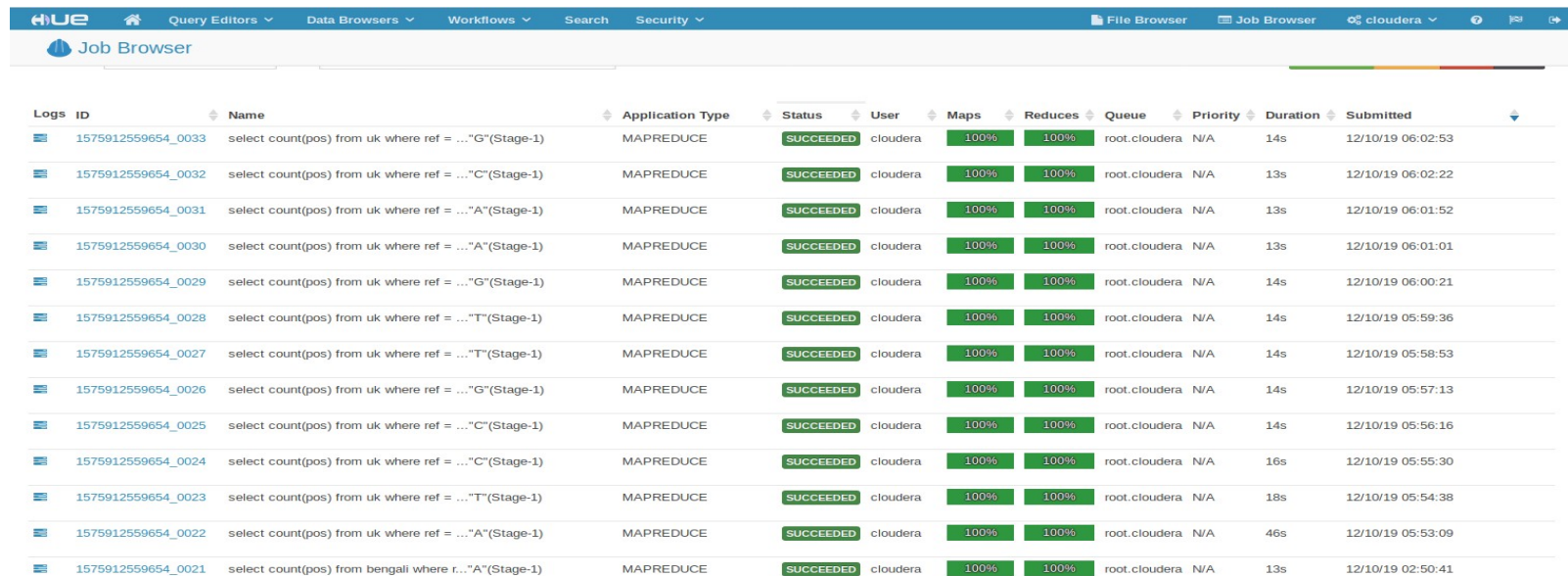
Visualization of the result

- Histogram of the VCF's of the two populations



Hive is Fast!

- Processing of files in hive was very fast as compared to running in local environment



The screenshot shows the Hue Job Browser interface. The top navigation bar includes links for Query Editors, Data Browsers, Workflows, Search, Security, File Browser, Job Browser, and a dropdown for cloudera. The main content area displays a table of jobs with columns for Logs, ID, Name, Application Type, Status, User, Maps, Reduces, Queue, Priority, Duration, and Submitted. All jobs listed are MAPREDUCE applications that have succeeded, with 100% completion for both maps and reduces. The jobs are sorted by duration, showing a range from 13s to 46s.

Logs	ID	Name	Application Type	Status	User	Maps	Reduces	Queue	Priority	Duration	Submitted
	1575912559654_0033	select count(pos) from uk where ref = ..."G"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	14s	12/10/19 06:02:53
	1575912559654_0032	select count(pos) from uk where ref = ..."C"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	13s	12/10/19 06:02:22
	1575912559654_0031	select count(pos) from uk where ref = ..."A"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	13s	12/10/19 06:01:52
	1575912559654_0030	select count(pos) from uk where ref = ..."A"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	13s	12/10/19 06:01:01
	1575912559654_0029	select count(pos) from uk where ref = ..."G"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	14s	12/10/19 06:00:21
	1575912559654_0028	select count(pos) from uk where ref = ..."T"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	14s	12/10/19 05:59:36
	1575912559654_0027	select count(pos) from uk where ref = ..."T"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	14s	12/10/19 05:58:53
	1575912559654_0026	select count(pos) from uk where ref = ..."G"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	14s	12/10/19 05:57:13
	1575912559654_0025	select count(pos) from uk where ref = ..."C"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	14s	12/10/19 05:56:16
	1575912559654_0024	select count(pos) from uk where ref = ..."C"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	16s	12/10/19 05:55:30
	1575912559654_0023	select count(pos) from uk where ref = ..."T"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	18s	12/10/19 05:54:38
	1575912559654_0022	select count(pos) from uk where ref = ..."A"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	46s	12/10/19 05:53:09
	1575912559654_0021	select count(pos) from bengali where r..."A"(Stage-1)	MAPREDUCE	SUCCEEDED	cloudera	100%	100%	root.cloudera	N/A	13s	12/10/19 02:50:41

References

- **Searching for SNPs with cloud computing**
Ben Langmead, Michael C Schatz, Jimmy Lin, Mihai Pop and Steven L Salzberg
- **The application of Hadoop in Structural Bioinformatics**
Jamie Alnasir, Hugh P. Shanahan
- **Big Data Processing for Genomics**
Altti Ilari Maarala, Keijo Heljanko, Andre Schumacher, Ridvan Dongelci, Luca Pireddu, Matti Niemenmaa, Aleksi Kallio, Eija Korpelainen and Gianluigi Zanetti

Thank You