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

Jyotsna Singh – PGDBD201901006

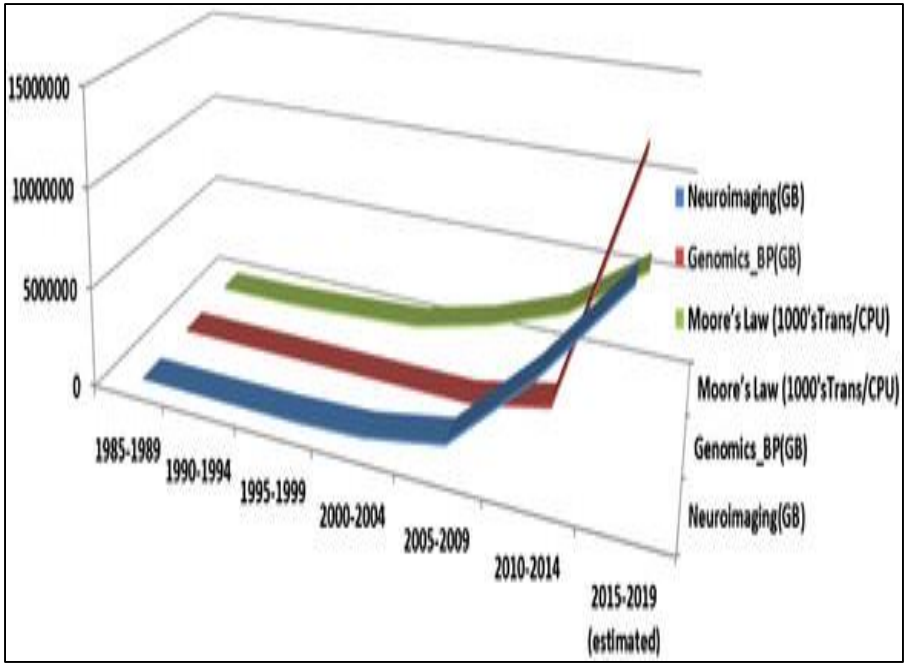
Paulami Das – PGDBD201901009

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 must faster than Moore's law
- 1000 Genomes projects has Petabytes of human genome data sets
- In many GWAS and WGS studies multiple large files has 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-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /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
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    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
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G     DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFCCC
    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
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C     DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@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       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

SAM/BAM Files

```
>@HWI-ST216_0180:4:1101:1096:2196#GGCTAC/1
TTTTCAGNGAATACTGCAAATCAATAAACTCTTTAG
>@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\ddddedeeedf_fbd
@HWI-ST216_0180:4:1101:1158:2236#GGCTAC/1
AAAAGCTCATTTCCTATAGTTAACAGGACATGCCTT
+HWI-ST216_0180:4:1101:1158:2236#GGCTAC/1
gggggggggggggggggeggfffggggggggfg
@HWI-ST216_0180:4:1101:1448:2211#GGCTAC/1
ATTATATAAGATAGCGGCTTTTTCCGTTAGTTTCCT
```

Fastq Files

ReferenceGenome	Homo_sapiens\UCSC\hg19\Sequence\WholeGenomeFASTA				
[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
```

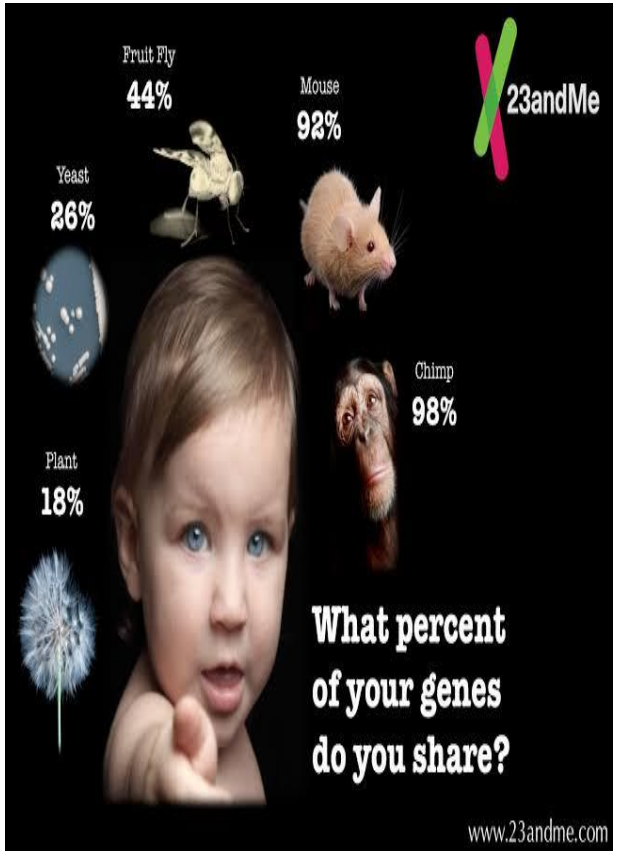
VCF Files

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr1	hg19_rmsk	exon	83886030	83886750	980	-	.	gene_id "ERVL-E-int";
chr1	hg19_rmsk	exon	192937729	192938046	491	+	.	gene_id "MLT1J";
chr1	hg19_rmsk	exon	243269324	243269678	2249	+	.	gene_id "THE1B";
chr1	hg19_rmsk	exon	10485712	10486981	8848	-	.	gene_id "LTR12C";
chr1	hg19_rmsk	exon	53477263	53477536	1005	+	.	gene_id "MER21C";
chr1	hg19_rmsk	exon	74448477	74449973	9303	-	.	gene_id "THE1B-int";
chr1	hg19_rmsk	exon	82837212	82837529	461	-	.	gene_id "MamRep1527";
chr1	hg19_rmsk	exon	102760404	102760736	1754	+	.	gene_id "THE1D";
chr1	hg19_rmsk	exon	145751968	145752401	1532	-	.	gene_id "LTR47B";
chr1	hg19_rmsk	exon	153091848	153092533	4373	-	.	gene_id "LTR76";
chr1	hg19_rmsk	exon	160432100	160432454	848	+	.	gene_id "LTR85c";
chr1	hg19_rmsk	exon	210763724	210764168	1368	-	.	gene_id "MLT2B4";
chr1	hg19_rmsk	exon	211812124	211812354	1068	-	.	gene_id "MER21A";
chr1	hg19_rmsk	exon	213909477	213910008	1378	-	.	gene_id "MLT1F2";
chr1	hg19_rmsk	exon	238026629	238027031	816	+	.	gene_id "MLT1M";
chr1	hg19_rmsk	exon	241172240	241172753	3210	+	.	gene_id "LTR24C";
chr1	hg19_rmsk	exon	262105	262386	721	+	.	gene_id "MER65D";
chr1	hg19_rmsk	exon	3538719	3539068	2001	+	.	gene_id "THE1D";
chr1	hg19_rmsk	exon	4324984	4325379	1304	-	.	gene_id "MLT1B";
chr1	hg19_rmsk	exon	5108932	5112709	28994	+	.	gene_id "HERVH-int";
chr1	hg19_rmsk	exon	12713599	12714399	808	-	.	gene_id "MER52-int";
chr1	hg19_rmsk	exon	12840257	12845090	27617	-	.	gene_id "HERVK-int";

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
  Read length - 150 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**

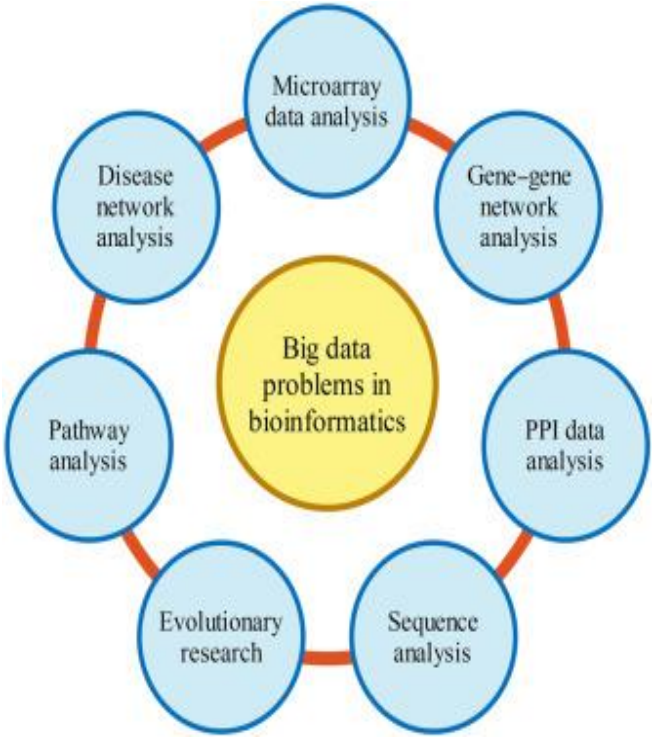
Hadoop in Different Biological Aspects

- 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

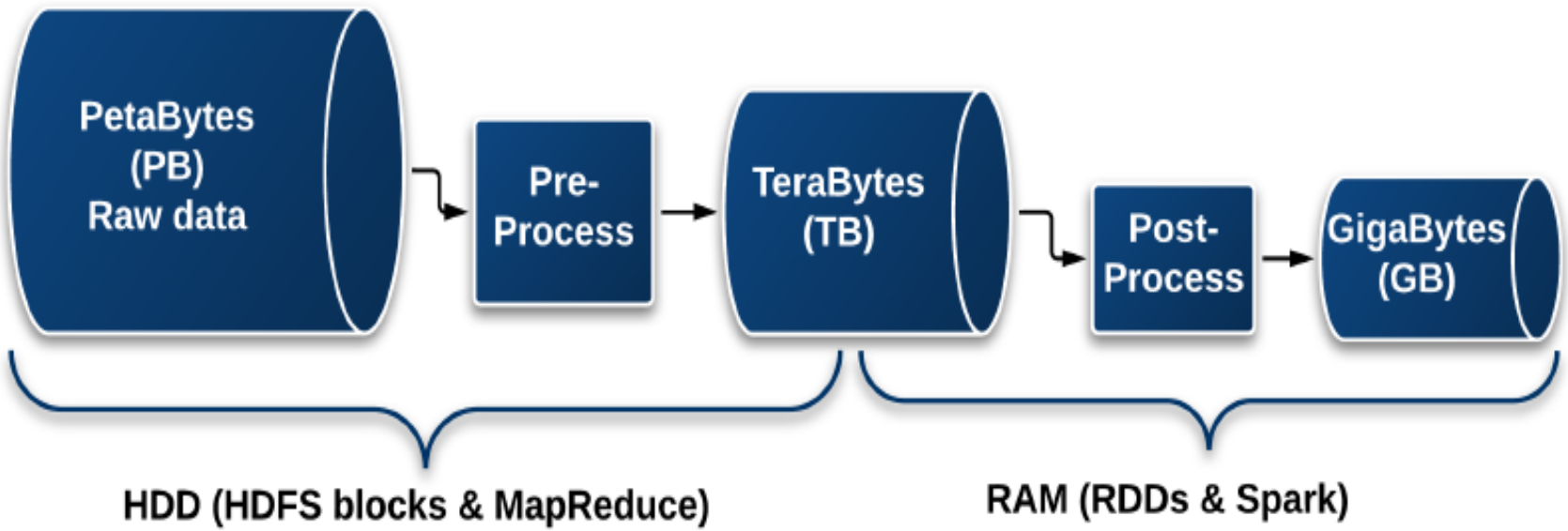
  3) Structural Alignment
- In Genomic Data Analysis



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

↓

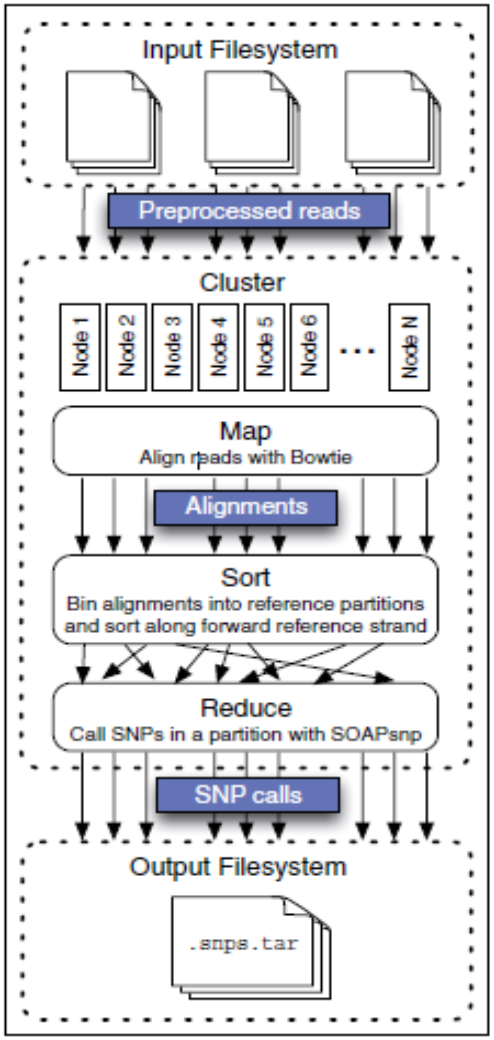
- The index is distributed to all computers in cluster via hadoop file or by instructing each node to independently obtain the index from a shared file

↓

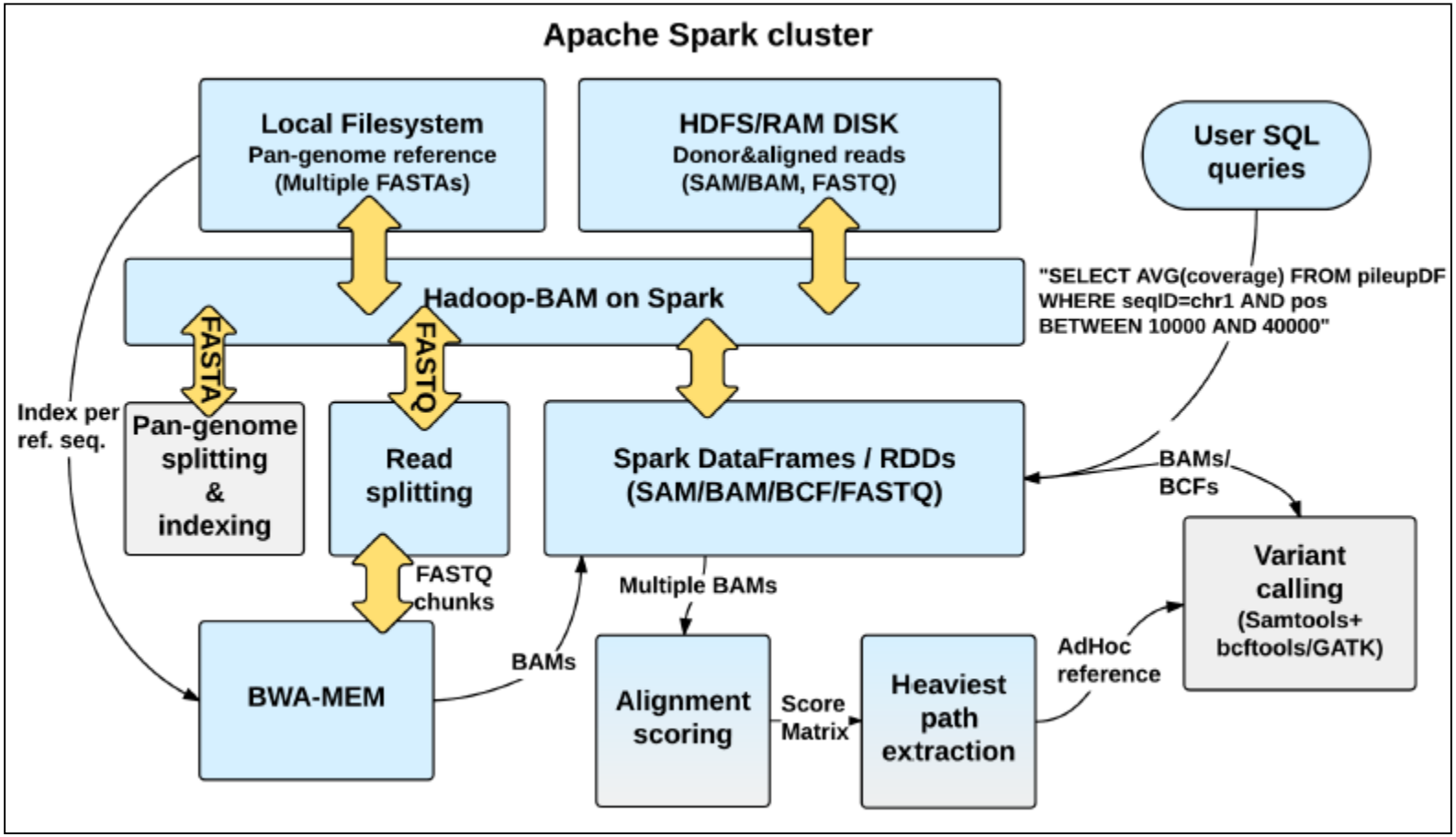
- The reduce phase performs SOAPsnp

↓

- The output of Reduce phase is SNP tuple which stored on the Clustered distributed File System which can be transferred to Local File System.

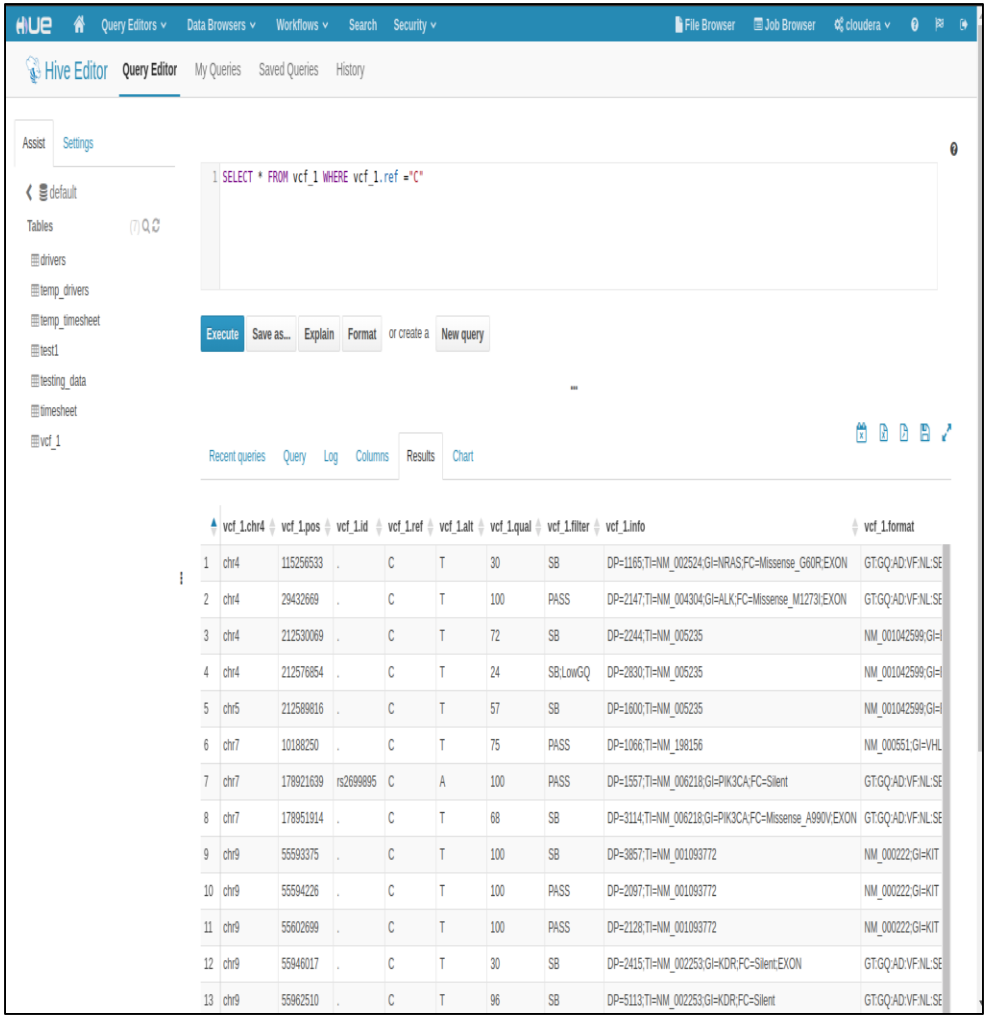


SNP Detection using HadoopBAM



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



	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=I
4	chr4	212576854	.	C	T	24	SB;LowGQ	DP=2830;TI=NM_005235	NM_001042599;GI=I
5	chr5	212589816	.	C	T	57	SB	DP=1600;TI=NM_005235	NM_001042599;GI=I
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	178951914	.	C	T	68	SB	DP=3114;TI=NM_006218;GI=PIK3CA;FC=Missense_A990V;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

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