Organization and analysis of NGS variations.

Alireza Hadj Khodabakhshi

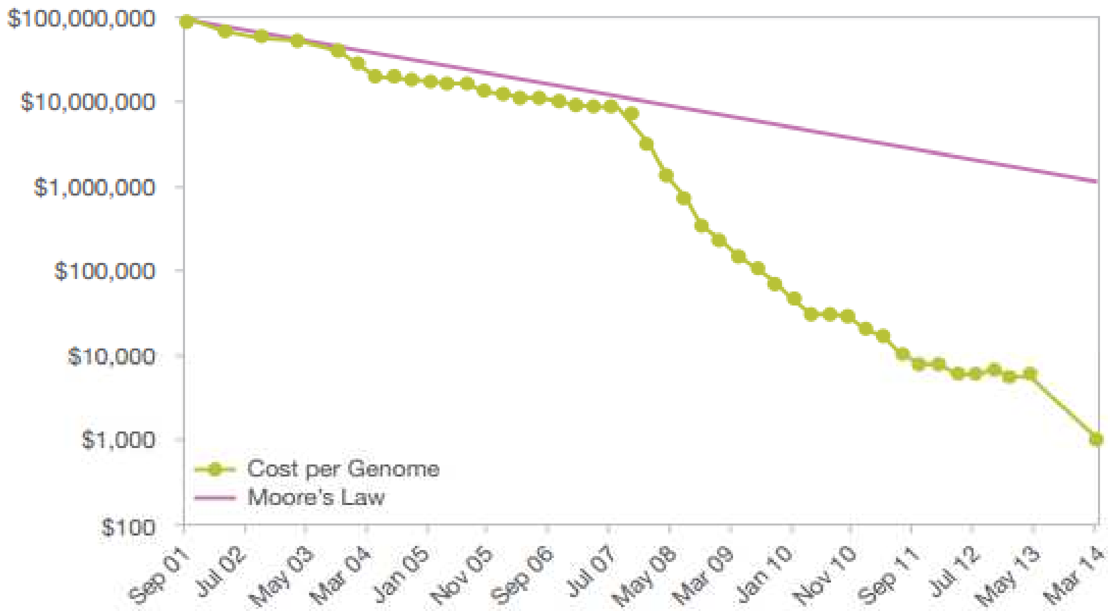
Research Investigator

Genomics Institute of the
Novartis Research
Foundation

Why is the NGS data processing a big challenge?

Computation cannot keep up with the Biology.

Figure 3: Illumina Sequencing Technology Outpaces Moore's Law for the Price of Whole Human Genome Sequencing



Illumina sequencing technology has been outpacing Moore's Law (purple line), which describes a long-term trend in the computer hardware industry where computing power doubles roughly every two years.[4] HiSeq X Ten dramatically continues this trend and is the first platform to break the $1000 barrier for a 30x human genome.

Source: illumin

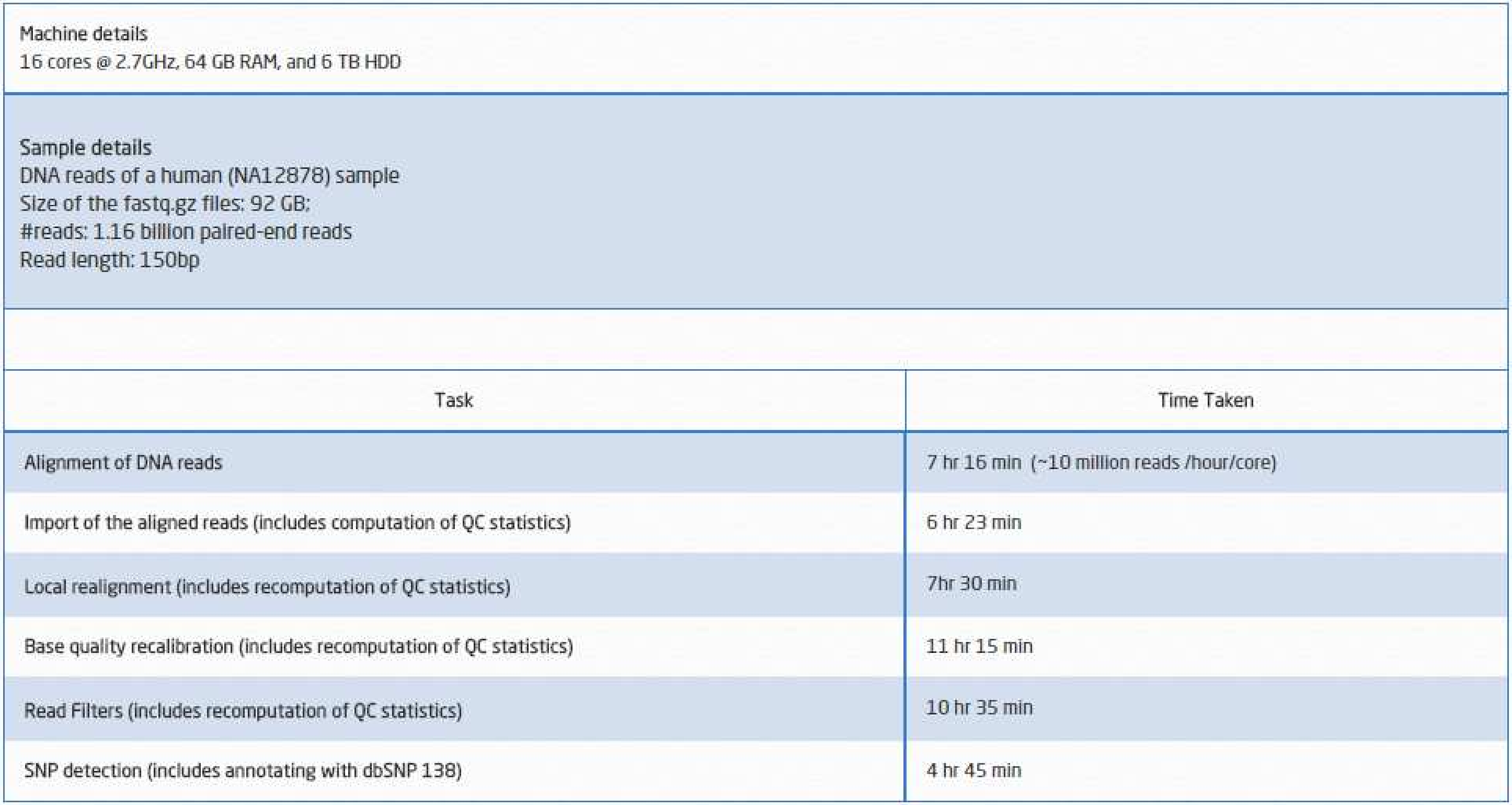
$1000 human gnome

- 50 whole genome per day
- 5 tera bytes (only mapped reads) per day

Figure 1: The HiSeq X Ten

The HiSeq X Ten, a set of 10 HiSeq X Sequencing Systems, is the only high-throughput sequencing system that can produce tens of thousands of human genomes a year for under $1000 per genome.

Source: illumin



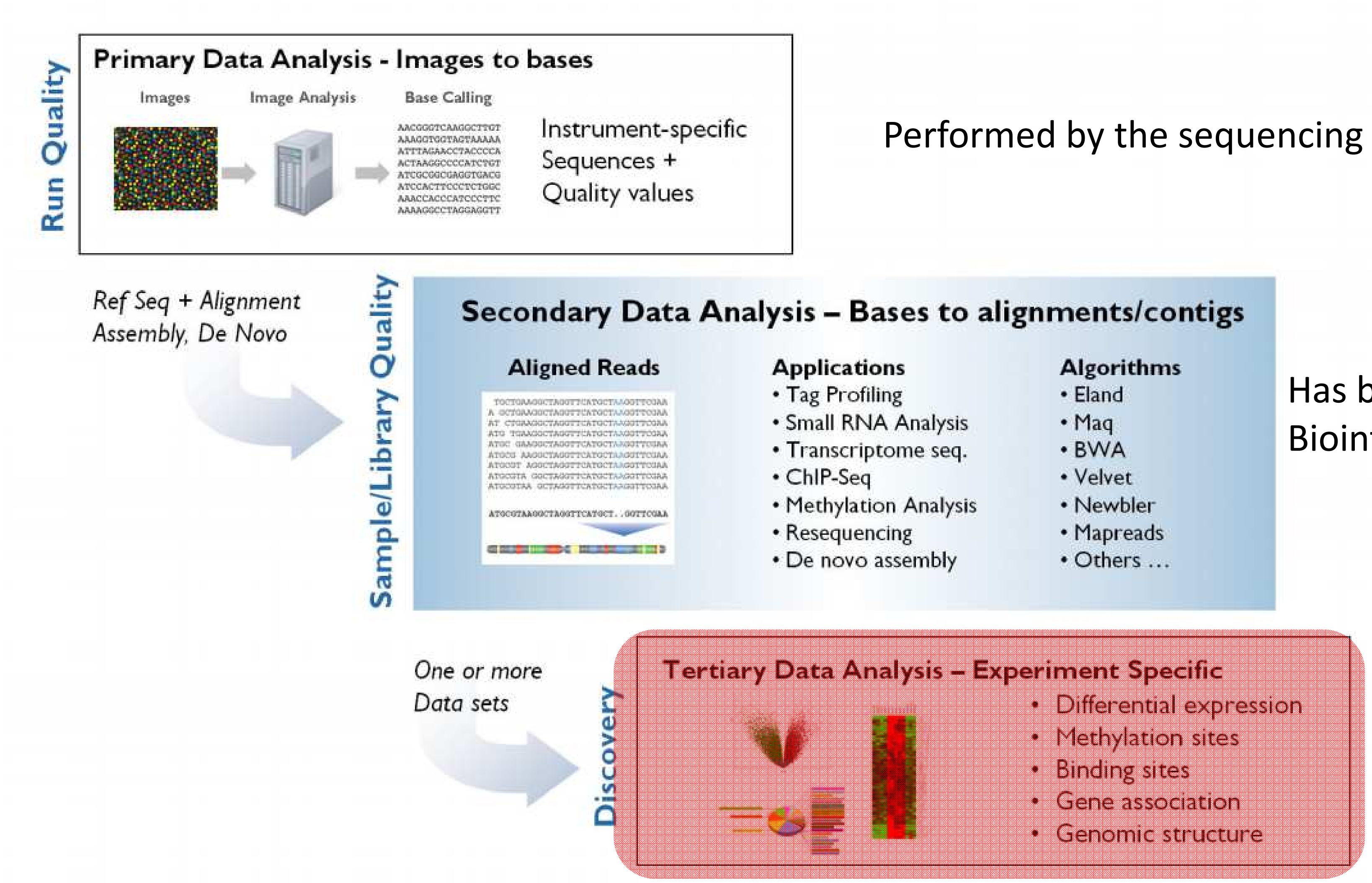
Machine details
16 cores @ 2.7GHz, 64 GB RAM, and 6 TB HDD

Sample details
DNA reads of a human (NA12878) sample
Size of the fastq.gz files: 92 GB;
#reads: 1.16 billion paired-end reads
Read length: 150bp

Task	Time Taken
Alignment of DNA reads	7 hr 16 min (~10 million reads /hour/core)
Import of the aligned reads (includes computation of QC statistics)	6 hr 23 min
Local realignment (includes recomputation of QC statistics)	7hr 30 min
Base quality recalibration (includes recomputation of QC statistics)	11 hr 15 min
Read Filters (includes recomputation of QC statistics)	10 hr 35 min
SNP detection (includes annotating with dbSNP 138)	4 hr 45 min

Source: strand-

Bioinformatics of NGS data



Run Quality

Primary Data Analysis - Images to bases

Instrument-specific Sequences + Quality values

Ref Seq + Alignment
Assembly, De Novo

Sample/Library Quality

Secondary Data Analysis – Bases to alignments/contigs

Applications
- Tag Profiling
- Small RNA Analysis
- Transcriptome seq.
- ChIP-Seq
- Methylation Analysis
- Resequencing
- De novo assembly

Algorithms
- Eland
- Maq
- BWA
- Velvet
- Newbler
- Mapreads
- Others …

One or more
Data sets

Discovery

Tertiary Data Analysis – Experiment Specific
- Differential expression
- Methylation sites
- Binding sites
- Gene association
- Genomic structure

Performed by the sequencing instrument.

Has been the main foc
Bioinformatics researc

Less tools are av

eneSifter: Next Generation Data Management and Analysis for Next Generation Sequencing.
tp://www.geospiza.com/

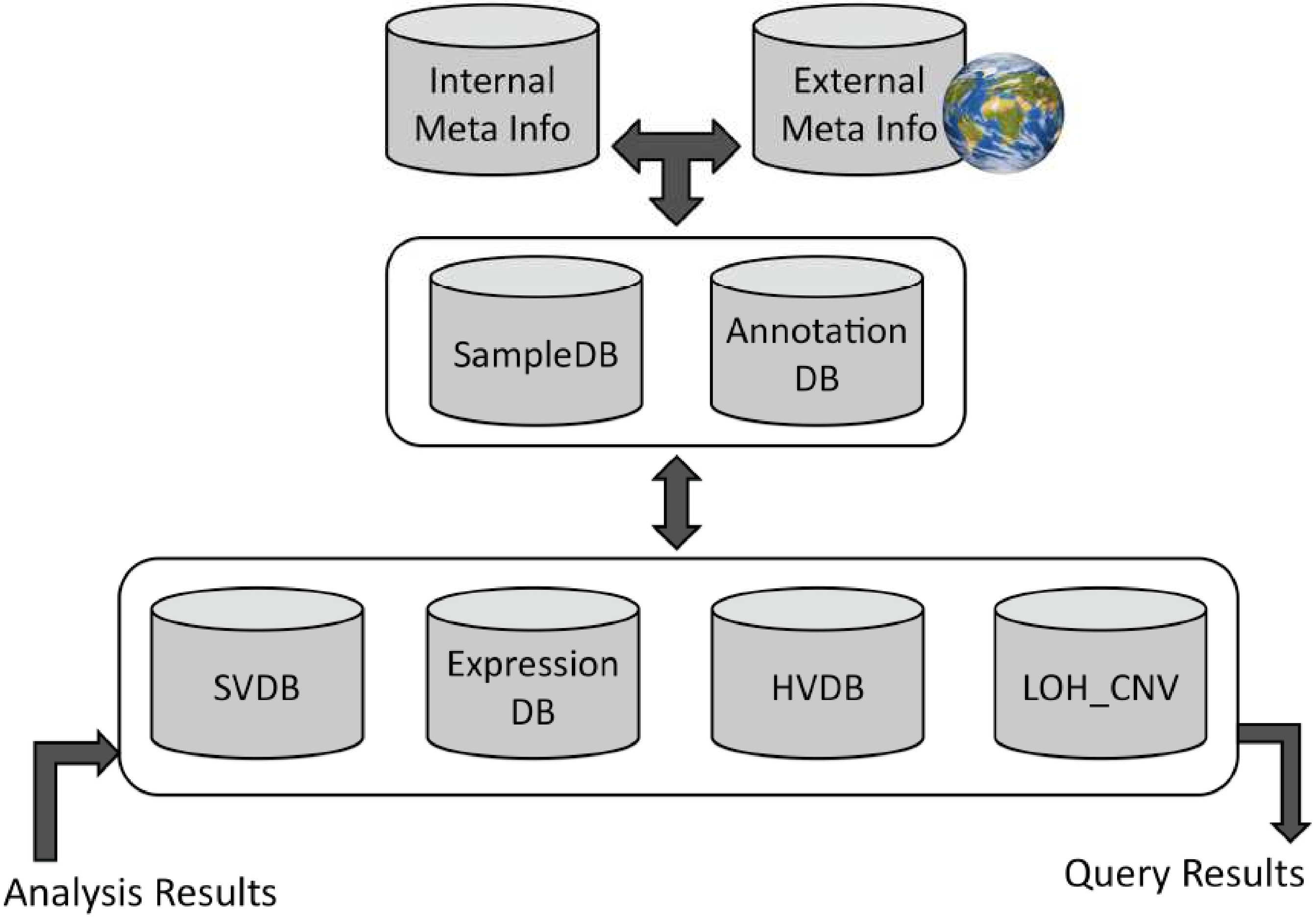
Organizing the variation data.

Scalable.

Enable insightful queries in a timely manner.

Support various NGS data (variations, expressions, annotations,…).

A consortium of databases for genomic discovery.



Sample Database(SampleDB):

clinical and experimental information of the samples (type of disease, pathology, age, sex,…).

Annotation Database(AnnotationDB):

annotations of genomic regions (sources: UCSC, Ensembl,….)

Structural Variation Database(SVDB):

genomic structural variations (translocations, inversions, large indels).

Expression Database (ExpressionDB):

expression levels of genomic regions (RPKM values).

Human Variation Database (HVDB):

small genomic variants (SNP, small indels)

Loss of Heterozygosity & Copy Number Variation (LOH_CNV)

Human Variation Database (HVDB)

Starting point of the consortium.

Stores SNPs and small indels.

Contains more than 4 billion variations across over 6000 samples.

Implemented with PostgreSQL and Java.

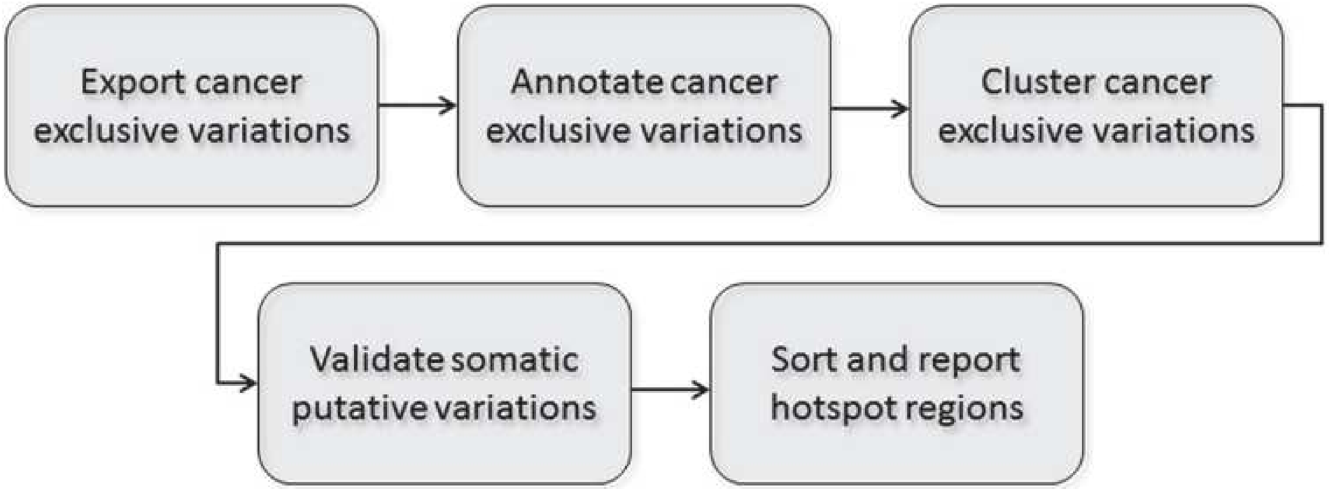
Its template and APIs are publically available.

Analyzing the data

- Mutated pathways in types of cancer.
- Variation hotspots.
- Correlation between various variation types (eg. correlation between SNVs and genomic translocations).
- Correlation between variations and expressions.

Mutation analysis pipeline:

- A high throughput pipeline on top of the genomic database consortium.
- Current version identifies statistically significant mutational hotspots.


Export cancer exclusive variations
Annotate cancer exclusive variations
Cluster cancer exclusive variations
Validate somatic putative variations
Sort and report hotspot regions

ure 1 Validation report. The major phases of the MuteProc mutation analysis pipeline.

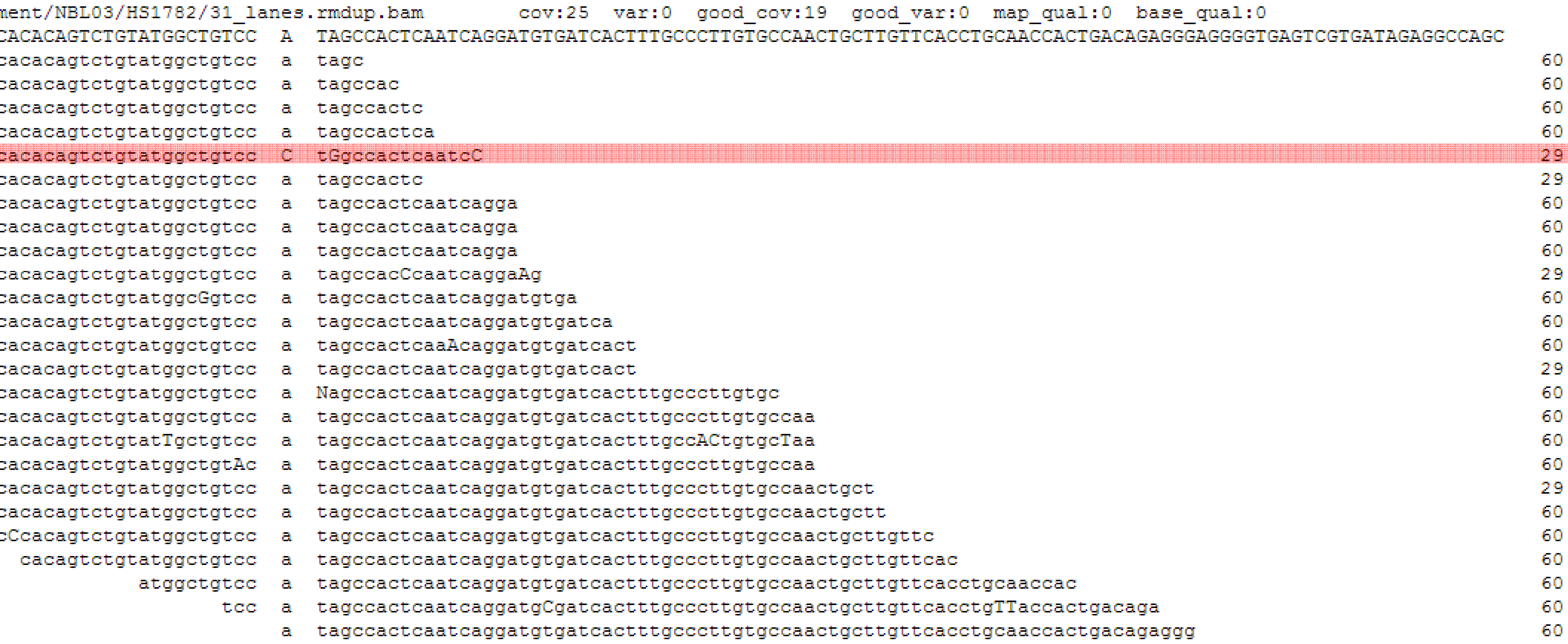
Validating the variations

Through the analysis of mapped read (raw data) at the variation site.

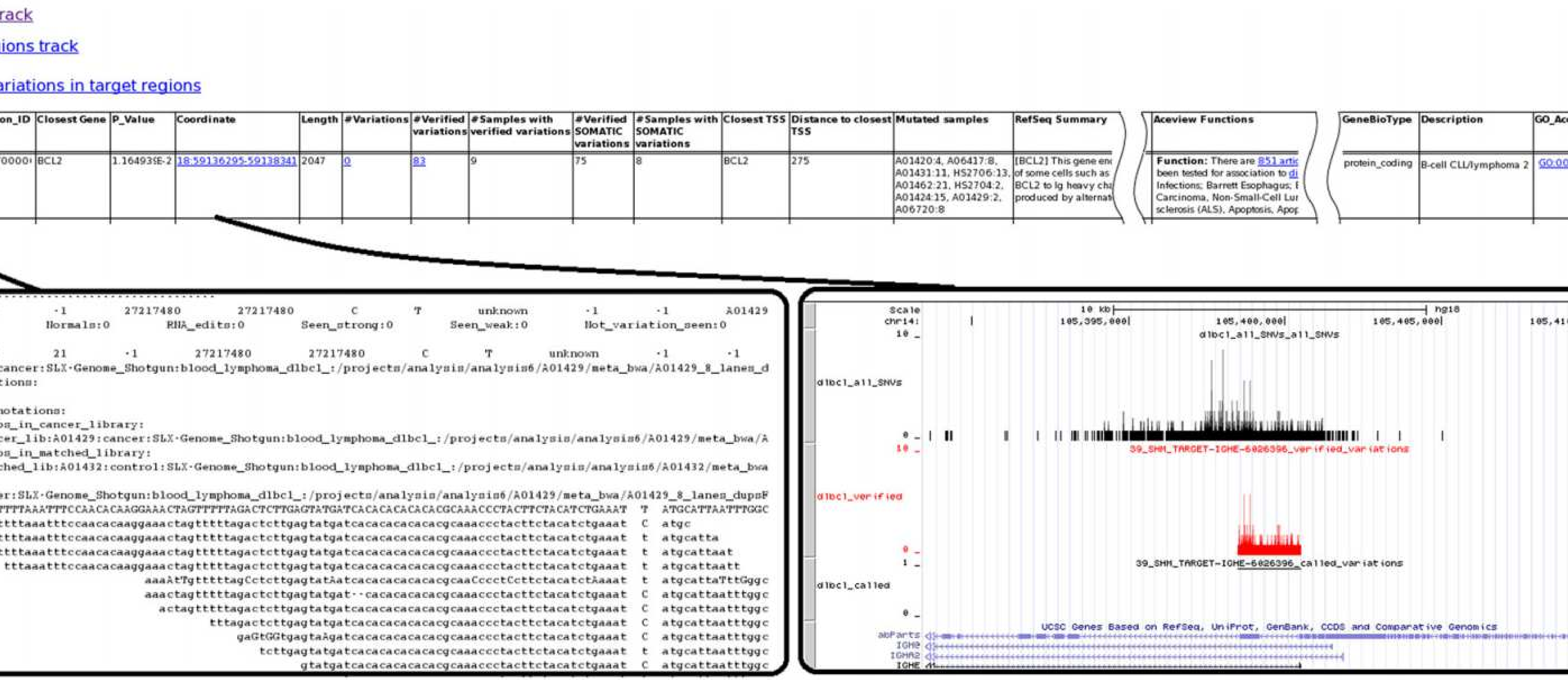
Calculates the confidence based the ratio of good reads that support the variation.

Uses the mapped read of the matched normal if available.

The process is performed on a computing cluster in a parallel way nment report.



```
ment/NBL03/HS1782/31_lanes.rmdup.bam        cov:25  var:0  good_cov:19  good_var:0  map_qual:0  base_qual:0
CACACAGTCTGTATGGCTGTCC  A  TAGCCACTCAATCAGGATGTGATCACTTTGCCCTTGTGCCAACTGCTTGTTCACCTGCAACCACTGACAGAGGGAGGGGTGAGTCGTGATAGAGGCCAGC
cacacagtctgtatggctgtcc  a  tagc                                                                                                   60
cacacagtctgtatggctgtcc  a  tagccac                                                                                                60
cacacagtctgtatggctgtcc  a  tagccactc                                                                                              60
cacacagtctgtatggctgtcc  a  tagccactca                                                                                             60
cacacagtctgtatggctgtcc  C  tGgccactcaatcC                                                                                         29
cacacagtctgtatggctgtcc  a  tagccactc                                                                                              29
cacacagtctgtatggctgtcc  a  tagccactcaatcagga                                                                                      60
cacacagtctgtatggctgtcc  a  tagccactcaatcagga                                                                                      60
cacacagtctgtatggctgtcc  a  tagccactcaatcagga                                                                                      60
cacacagtctgtatggctgtcc  a  tagccacCcaatcaggaAg                                                                                    29
cacacagtctgtatggcGgtcc  a  tagccactcaatcaggatgtga                                                                                 60
cacacagtctgtatggctgtcc  a  tagccactcaatcaggatgtgatca                                                                              60
cacacagtctgtatggctgtcc  a  tagccactcaaAcaggatgtgatcact                                                                            60
cacacagtctgtatggctgtcc  a  tagccactcaatcaggatgtgatcact                                                                            29
cacacagtctgtatggctgtcc  a  Nagccactcaatcaggatgtgatcactttgcccttgtgc                                                                60
cacacagtctgtatggctgtcc  a  tagccactcaatcaggatgtgatcactttgcccttgtgccaa                                                             60
cacacagtctgtatTgctgtcc  a  tagccactcaatcaggatgtgatcactttgccACtgtgcTaa                                                             60
cacacagtctgtatggctgtAc  a  tagccactcaatcaggatgtgatcactttgcccttgtgccaa                                                             60
cacacagtctgtatggctgtcc  a  tagccactcaatcaggatgtgatcactttgcccttgtgccaactgct                                                        29
cacacagtctgtatggctgtcc  a  tagccactcaatcaggatgtgatcactttgcccttgtgccaactgctt                                                       60
cCcacagtctgtatggctgtcc  a  tagccactcaatcaggatgtgatcactttgcccttgtgccaactgcttgttc                                                   60
  cacagtctgtatggctgtcc  a  tagccactcaatcaggatgtgatcactttgcccttgtgccaactgcttgttcac                                                 60
            atggctgtcc  a  tagccactcaatcaggatgtgatcactttgcccttgtgccaactgcttgttcacctgcaaccac                                       60
                   tcc  a  tagccactcaatcaggatgCgatcactttgcccttgtgccaactgcttgttcacctgTTaccactgacaga                                60
                        a  tagccactcaatcaggatgtgatcactttgcccttgtgccaactgcttgttcacctgcaaccactgacagaggg                             60
```



rack

ions track

ariations in target regions

on_ID	Closest Gene	P_Value	Coordinate	Length	#Variations	#Verified variations	#Samples with verified variations	#Verified SOMATIC variations	#Samples with SOMATIC variations	Closest TSS	Distance to closest TSS	Mutated samples	RefSeq Summary	Aceview Functions	GeneBioType	Description	GO_Ac
0000	BCL2	1.16493$E-2	18:59136295-59138341	2047	0	83	9	75	8	BCL2	275	A01420:4, A06417:8, A01431:11, HS2706:13, A01462:21, HS2704:2, A01424:15, A01429:2, A06720:8	[BCL2] This gene en of some cells such as BCL2 to Ig heavy cha produced by alternat	**Function:** There are 851 artic been tested for association to di Infections; Barrett Esophagus; I Carcinoma, Non-Small-Cell Lur sclerosis (ALS), Apoptosis, Apop	protein_coding	B-cell CLL/lymphoma 2	GO:00

```
--------------------------------------
  -1          27217480        27217480          C          T          unknown          -1           -1          A01429
       Normals:0          RNA_edits:0          Seen_strong:0          Seen_weak:0          Not_variation_seen:0

       21          -1          27217480          27217480          C          T          unknown          -1          -1
cancer:SLX-Genome_Shotgun:blood_lymphoma_dlbcl_:/projects/analysis/analysis6/A01429/meta_bwa/A01429_8_lanes_d
tions:

notations:
os_in_cancer_library:
cer_lib:A01429:cancer:SLX-Genome_Shotgun:blood_lymphoma_dlbcl_:/projects/analysis/analysis6/A01429/meta_bwa/A
os_in_matched_library:
ched_lib:A01432:control:SLX-Genome_Shotgun:blood_lymphoma_dlbcl_:/projects/analysis/analysis6/A01432/meta_bwa

er:SLX-Genome_Shotgun:blood_lymphoma_dlbcl_:/projects/analysis/analysis6/A01429/meta_bwa/A01429_8_lanes_dupsF
TTTTAAATTTCCAACACAAGGAAACTAGTTTTTAGACTCTTGAGTATGATCACACACACACACGCAAACCCTACTTCTACATCTGAAAT  T  ATGCATTAATTTGGC
ttttaaatttccaacacaaggaaactagtttttagactcttgagtatgatcacacacacacacgcaaaccctacttctacatctgaaat  C  atgc
ttttaaatttccaacacaaggaaactagtttttagactcttgagtatgatcacacacacacacgcaaaccctacttctacatctgaaat  t  atgcatta
ttttaaatttccaacacaaggaaactagtttttagactcttgagtatgatcacacacacacacgcaaaccctacttctacatctgaaat  t  atgcattaat
 tttaaatttccaacacaaggaaactagtttttagactcttgagtatgatcacacacacacacgcaaaccctacttctacatctgaaat  t  atgcattaatt
                     aaaAtTgtttttagCctcttgagtatAatcacacacacacacgcaaCccctCcttctacatctAaaat  t  atgcattaTttGggc
                     aaactagtttttagactcttgagtatgat--cacacacacacgcaaaccctacttctacatctgaaat  C  atgcattaatttggc
                       actagtttttagactcttgagtatgatcacacacacacacgcaaaccctacttctacatctgaaat  C  atgcattaatttggc
                              tttagactcttgagtatgatcacacacacacacgcaaaccctacttctacatctgaaat  C  atgcattaatttggc
                                  gaGtGGtgagtaAgatcacacacacacacgcaaaccctacttctacatctgaaat  C  atgcattaatttggc
                                     tcttgagtatgatcacacacacacacgcaaaccctacttctacatctgaaat  t  atgcattaatttggc
                                           gtatgatcacacacacacacgcaaaccctacttctacatctgaaat  C  atgcattaatttggc
```

Scale
chr14:
10 kb
hg18
105,395,000
105,400,000
105,405,000
105,41
dlbcl_all_SNVs_all_SNVs
dlbcl_all_SNVs
10
0
39_SHM_TARGET-IGHE-6026396_verified_variations
dlbcl_verified
10
0
1
39_SHM_TARGET-IGHE-6026396_called_variations
dlbcl_called
0
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
abParts
IGHG
IGHA2
IGHE

3 Final report. A snap shot of the final report generated by the pipeline. The links in the "rank" column point to the variation QC rep
esponding region. The three links on the top of the report, that is "All SNVs track", "Target regions track" and "Verified Variations in tar
", uploads the variation locations as custom tracks in the UCSC genome browser. Once these tracks are uploaded, clicking on the link
nate" column browse to the associated region in the UCSC genome browser where the variations are visible in the loaded variation t

Performance

ata size:

- 2.5 billion variations
- ~2000 cancer samples.
- ~2000 normal samples.

latform:

- Linux Centos
- PostgreSQL database.
- Java APIs.
- Database server: eight core Xeon ® 3.00 GHz, 64 GB memory
- Application machine: 4 core, 8 GB memory

he pipeline run completes in about 23 hours.

Analysis of 40 DLBCL genomes.

Goal: Identify mutational hotspots in DLBCL genome.

Cohort: 40 whole genome DLBCL samples and their matched normal amples.

Conclusion: Small regions in the promoter of certain genes harbor an extraordinary amount of somatic mutations.

These regions undergo somatic hypermutation.

Somatic HyperMutation (SHM)

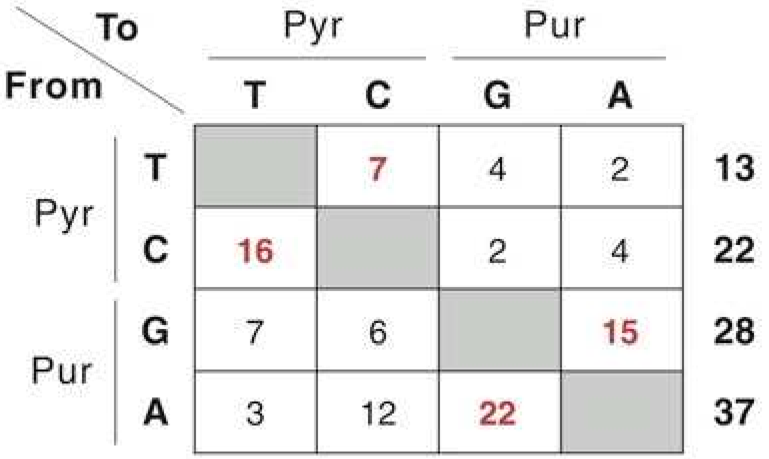
Naturally occurs in B-Cell development to generate diverse antibodies.

It occurs in variable region of immunoglobulin genes.

10^5-10^6 fold greater than the normal rate of mutation across the genome.

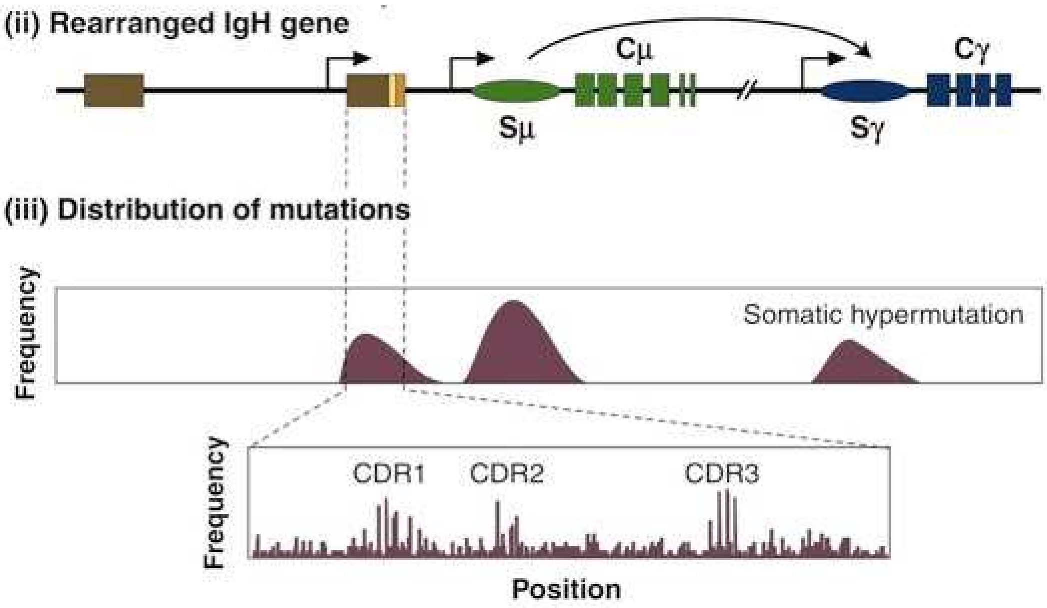
Mutations are mostly single base substitution (insertion and deletions are less common).

SHM Characteristics



	To	Pyr		Pur		
From		T	C	G	A	
Pyr	T		7	4	2	13
	C	16		2	4	22
Pur	G	7	6		15	28
	A	3	12	22		37

Di Noia JM and Neuberger MS.
Annu Rev Biochem. 76:1 (2007)



SHM has a tendency toward certain motifs in DNA sequence, most significantly **WRCY** (where W denotes A or T; R denotes A or G; and Y denotes C or T) or its reverse complement **RGYW**

SHM can aberrantly target proto-oncogenes (BCL6, PIM1, MYC, RHOH, PAX5) and tumor suppressors (CD95).

Such mistargeting of SHM (aSHM) contributes to the development of diffuse large B-ell lymphomas.

SHM also has a driving role in chromosomal translocations in B-cell lymphomas.

In the past decade twelve genes had been identified to have aSHM.

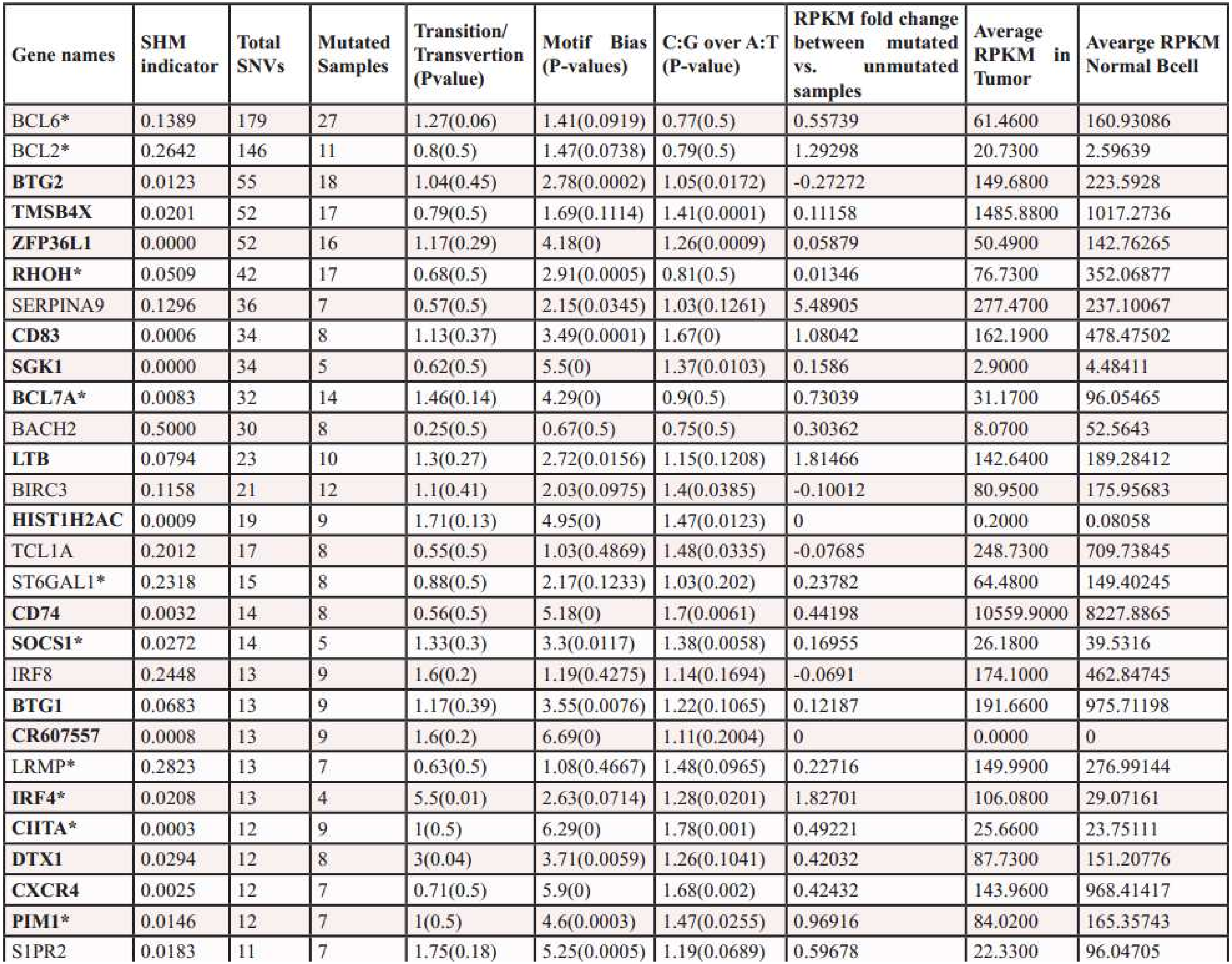
In addition to these genes our analysis identifies many more.

Are these novel genes really targeted by aSHM?

Do they show characteristics of SHM?

- More Transition than Transversion SNVs.
- Tendency toward **WRCY/RGYW** motif**.**
- More C:G mutations than A:T
- A bell shape mutation distribution around TSS.

Ve studied these characteristics for the genes that had similar or higher mutation ate than those known to be aSHM targets **(44 genes).**



Gene names	SHM indicator	Total SNVs	Mutated Samples	Transition/ Transvertion (Pvalue)	Motif Bias (P-values)	C:G over A:T (P-value)	RPKM fold change between mutated vs. unmutated samples	Average RPKM in Tumor	Avearge RPKM Normal Bcell
BCL6*	0.1389	179	27	1.27(0.06)	1.41(0.0919)	0.77(0.5)	0.55739	61.4600	160.93086
BCL2*	0.2642	146	11	0.8(0.5)	1.47(0.0738)	0.79(0.5)	1.29298	20.7300	2.59639
BTG2	0.0123	55	18	1.04(0.45)	2.78(0.0002)	1.05(0.0172)	-0.27272	149.6800	223.5928
TMSB4X	0.0201	52	17	0.79(0.5)	1.69(0.1114)	1.41(0.0001)	0.11158	1485.8800	1017.2736
ZFP36L1	0.0000	52	16	1.17(0.29)	4.18(0)	1.26(0.0009)	0.05879	50.4900	142.76265
RHOH*	0.0509	42	17	0.68(0.5)	2.91(0.0005)	0.81(0.5)	0.01346	76.7300	352.06877
SERPINA9	0.1296	36	7	0.57(0.5)	2.15(0.0345)	1.03(0.1261)	5.48905	277.4700	237.10067
CD83	0.0006	34	8	1.13(0.37)	3.49(0.0001)	1.67(0)	1.08042	162.1900	478.47502
SGK1	0.0000	34	5	0.62(0.5)	5.5(0)	1.37(0.0103)	0.1586	2.9000	4.48411
BCL7A*	0.0083	32	14	1.46(0.14)	4.29(0)	0.9(0.5)	0.73039	31.1700	96.05465
BACH2	0.5000	30	8	0.25(0.5)	0.67(0.5)	0.75(0.5)	0.30362	8.0700	52.5643
LTB	0.0794	23	10	1.3(0.27)	2.72(0.0156)	1.15(0.1208)	1.81466	142.6400	189.28412
BIRC3	0.1158	21	12	1.1(0.41)	2.03(0.0975)	1.4(0.0385)	-0.10012	80.9500	175.95683
HIST1H2AC	0.0009	19	9	1.71(0.13)	4.95(0)	1.47(0.0123)	0	0.2000	0.08058
TCL1A	0.2012	17	8	0.55(0.5)	1.03(0.4869)	1.48(0.0335)	-0.07685	248.7300	709.73845
ST6GAL1*	0.2318	15	8	0.88(0.5)	2.17(0.1233)	1.03(0.202)	0.23782	64.4800	149.40245
CD74	0.0032	14	8	0.56(0.5)	5.18(0)	1.7(0.0061)	0.44198	10559.9000	8227.8865
SOCS1*	0.0272	14	5	1.33(0.3)	3.3(0.0117)	1.38(0.0058)	0.16955	26.1800	39.5316
IRF8	0.2448	13	9	1.6(0.2)	1.19(0.4275)	1.14(0.1694)	-0.0691	174.1000	462.84745
BTG1	0.0683	13	9	1.17(0.39)	3.55(0.0076)	1.22(0.1065)	0.12187	191.6600	975.71198
CR607557	0.0008	13	9	1.6(0.2)	6.69(0)	1.11(0.2004)	0	0.0000	0
LRMP*	0.2823	13	7	0.63(0.5)	1.08(0.4667)	1.48(0.0965)	0.22716	149.9900	276.99144
IRF4*	0.0208	13	4	5.5(0.01)	2.63(0.0714)	1.28(0.0201)	1.82701	106.0800	29.07161
CIITA*	0.0003	12	9	1(0.5)	6.29(0)	1.78(0.001)	0.49221	25.6600	23.75111
DTX1	0.0294	12	8	3(0.04)	3.71(0.0059)	1.26(0.1041)	0.42032	87.7300	151.20776
CXCR4	0.0025	12	7	0.71(0.5)	5.9(0)	1.68(0.002)	0.42432	143.9600	968.41417
PIM1*	0.0146	12	7	1(0.5)	4.6(0.0003)	1.47(0.0255)	0.96916	84.0200	165.35743
S1PR2	0.0183	11	7	1.75(0.18)	5.25(0.0005)	1.19(0.0689)	0.59678	22.3300	96.04705

Oncotarge
3: 1308-13

All known targets of aSHM (12 genes) are in the list and 75% of them have a significant aSHM indicator (a good control for our analysis).

More than 81 and 90 percent of the SHM-targets showed a bias for SHM criteria "*Motif enriched*" and "*C:G vs A:T mutation bias*".

If these gene are enriched with aSHM mutations, a random mutated gene should have a significantly less aSHM indicator value.

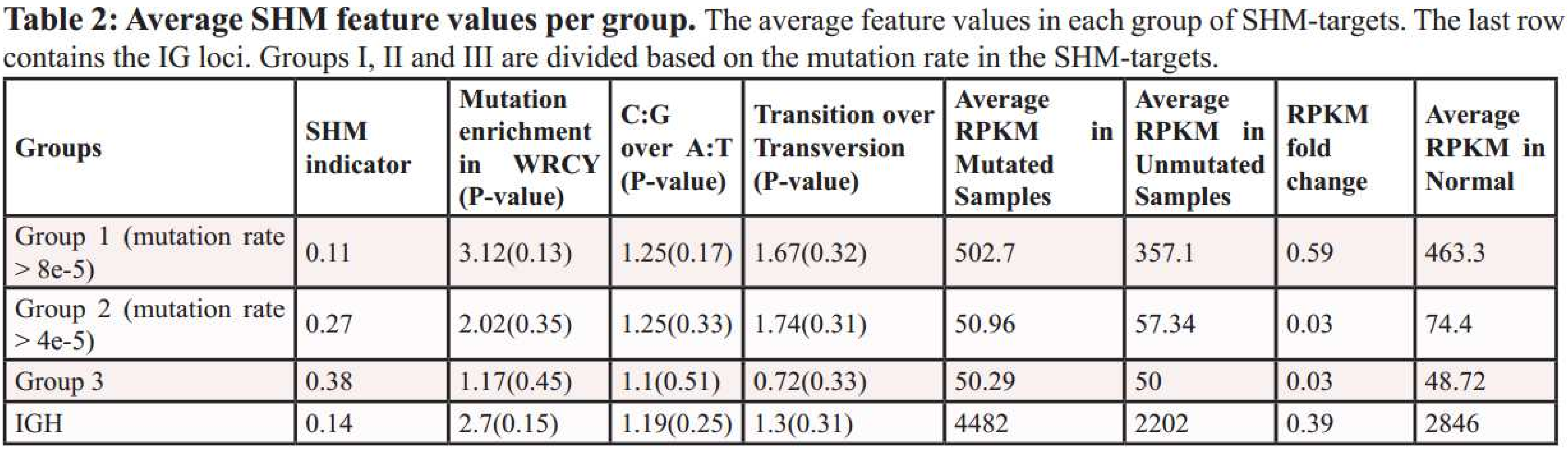
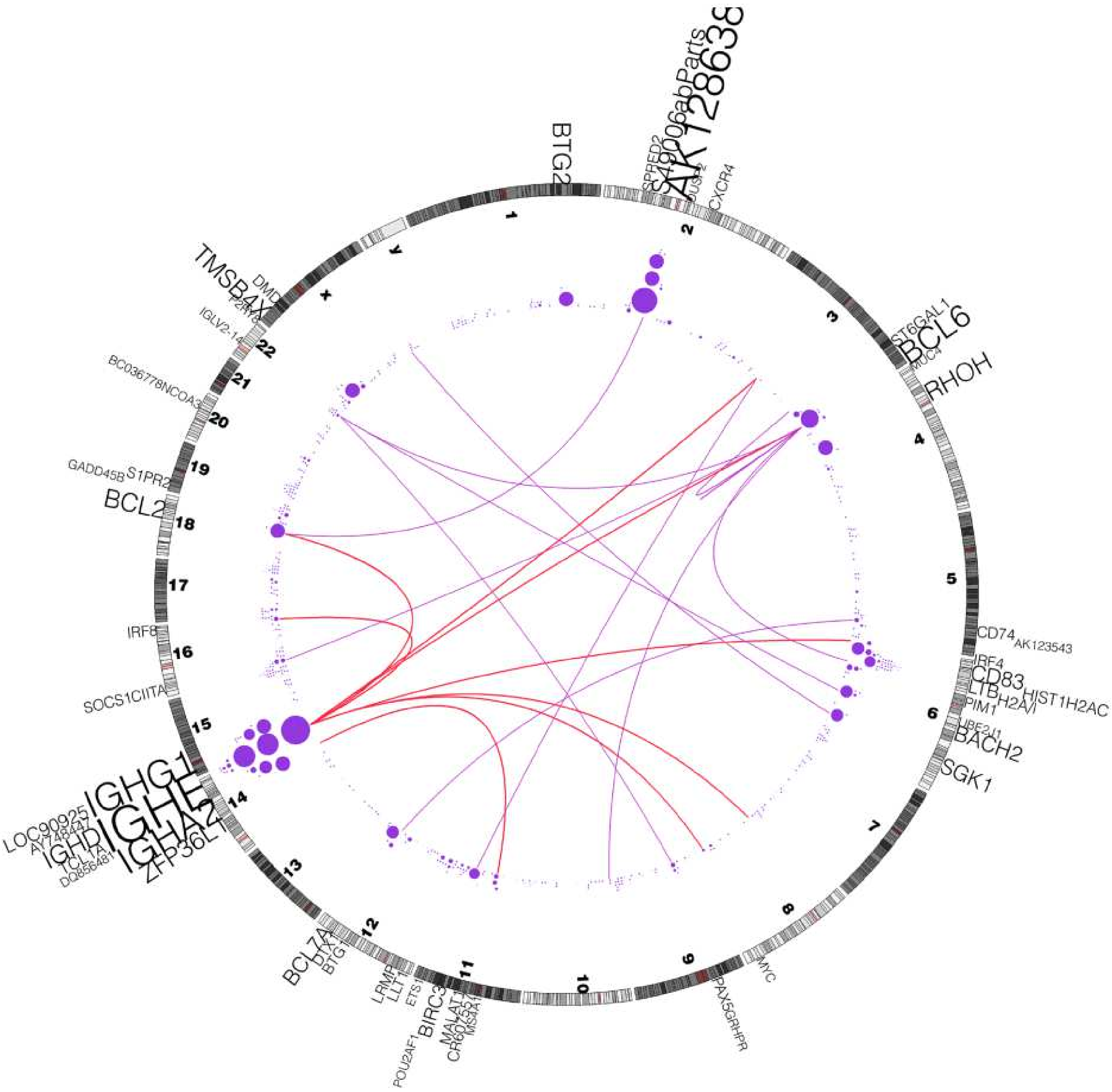


Table 2: Average SHM feature values per group. The average feature values in each group of SHM-targets. The last row contains the IG loci. Groups I, II and III are divided based on the mutation rate in the SHM-targets.

Groups	SHM indicator	Mutation enrichment in WRCY (P-value)	C:G over A:T (P-value)	Transition over Transversion (P-value)	Average RPKM in Mutated Samples	Average RPKM in Unmutated Samples	RPKM fold change	Average RPKM in Normal
Group 1 (mutation rate > 8e-5)	0.11	3.12(0.13)	1.25(0.17)	1.67(0.32)	502.7	357.1	0.59	463.3
Group 2 (mutation rate > 4e-5)	0.27	2.02(0.35)	1.25(0.33)	1.74(0.31)	50.96	57.34	0.03	74.4
Group 3	0.38	1.17(0.45)	1.1(0.51)	0.72(0.33)	50.29	50	0.03	48.72
IGH	0.14	2.7(0.15)	1.19(0.25)	1.3(0.31)	4482	2202	0.39	2846

The difference in RPKM values reflects a trend towards higher mRNA abundance of the mutated genes. This coincide with the observation that gene expression promotes SHM.

elation between tions and angements



Future works

- The processing pipeline is specially in early stages (include more analysis).
- Data visualization and GUI to browse results.
- Utilizing *Big Data* technologies to improve performance.
- Incorporate other data sources in a systematic way (pathways, PPI networks, …).
- Implement mechanisms to share data.

C Cancer Agency
ARE + RESEARCH
agency of the Provincial Health Services Authority

Research Centre
Canada's Michael Smith Genome Sciences Centre

Anthony Fejes

Steven Jones

Inanc Birol

Ryan Morin

Maria Mendez-Lago

Nina Thiessen

An He

Richard Varhol

Tina Wang

Richard Corbett

Misha Bilenky

Gordon Robertson

Andy Chu

Readman Chiu

Karen Mungall

Genomics Institute of t
Novartis Research
Foundation

Thanks