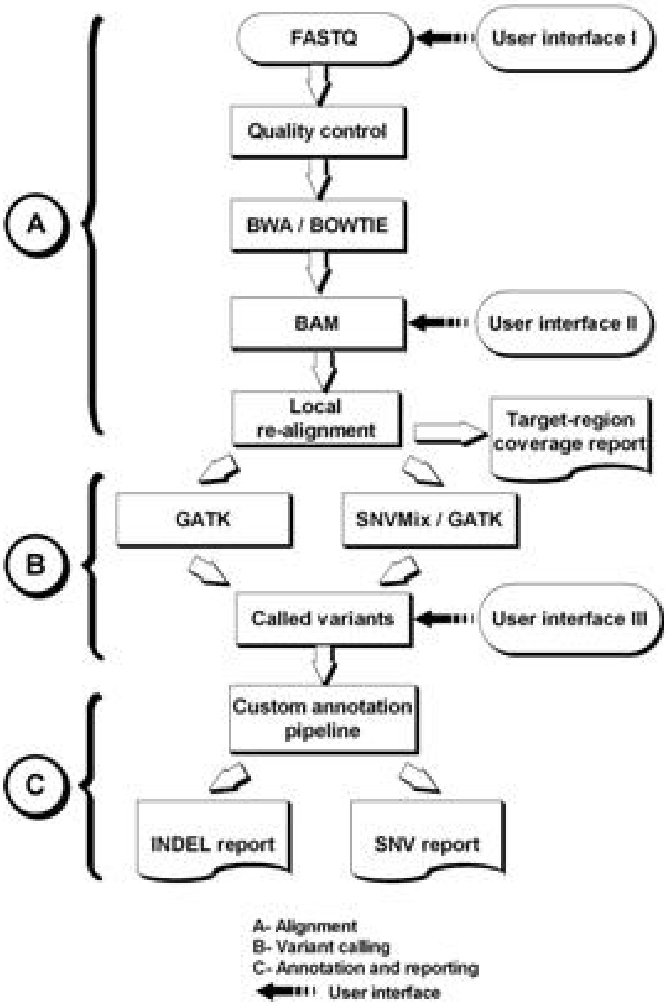
Files and Folder structure for TREAT

TREAT workflow comprises of 4 different modules

1. Alignment
2. Variant Calling
3. Annotation and Reports
4. Sample Statistics



Analytic TREAT workflow for Exome Capture analysis

NOTE:

Generated Intermediate files are there only for 60 days from the date of delivery. If user/PI wants to keep for longer term then they have to achieve these files or let the PI support group know about it. The Files we keep for longer term are merged variant reports, per sample variant reports and recalibrated and realigned BAM’s for visualization.

How to run the workflow

TREAT is just a command line tool. Run treat.sh to get the usage information.

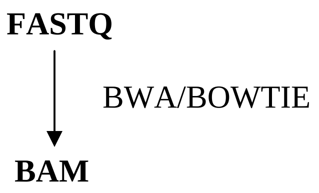
```
$ /path_to_workflow/treat.sh
Usage:
        1. <run_info file>
```

Before running the workflow user need to manually make three reference files

1. configuration file (path to various tools and references)
2. sample info file (names for the fastq, BAM, variant files for all the samples)
3. run information file (includes meta data information and information about the run)

Example for each reference file is available in an example folder with in the TREAT current version folder.

Alignment



Input:
Takes Fastq’s as input (all the Fastq’s should be in same folder)

Output:
Sorted and indexed BAM

Folders created:
Output_folder/alignment

Consist of a folder for each sample which contains the sorted and index bam file for that sample

```
<output_folder>
      | _ alignment
            | _ sample1
                  | _ sample-sorted.bam
                  | _ sample-sorted.bam.bai
            | _ sample2
```

Useful files:

```
      alignment/Sample/sample-sorted.bam
      alignment/Sample/sample-sorted.bam.bai
```

(Unfiltered BAM; contains all the reads (aligned and unaligned)

Potential Use of Intermediate data:

a) Extract Fastq

```
bam2fastq.sh
Usage: for PE: bam2fastq.sh <input bam>(provide full path) <name for read1> <name for read2>
       for SR: bam2fastq.sh <input bam>(provide full path) <name for read1>
```

b) To get the alignment numbers

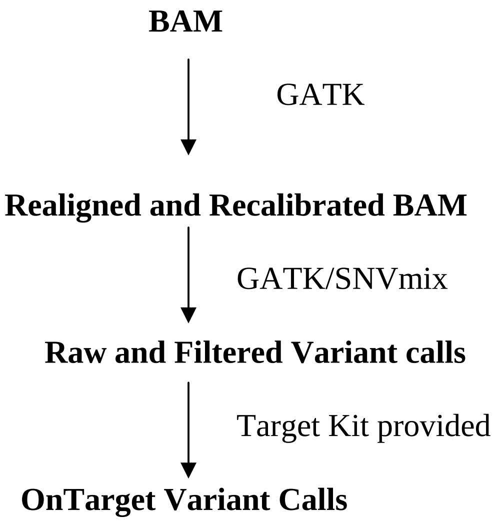
There is a file named `sample.flagstat` which gives out alignment numbers such as number of reads, mapped reads etc.

c) To get information about the BAM

```
samtools view -H output_folder/alignment/sample/sample-sorted.bam
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:247249719
@SQ     SN:chr2 LN:242951149
@SQ     SN:chr3 LN:199501827
```

```
@SQ     SN:chr4 LN:191273063
@PG     ID:bwa  PN:bwa  VN:0.5.9-r16
@PG     ID:BWA
@RG     ID:6-HBPN-Land-JL         SM:6-HBPN-Land-JL
LB:/data1/bsi/refdata/bictools/sequence/human/ncbi/36.49/indexed/BWA_Indexed_reference_includi
ng_ChrM/allChr.fa       PL:illumina     CN:MAYO
```

Variant Calling


BAM
GATK
Realigned and Recalibrated BAM
GATK/SNVmix
Raw and Filtered Variant calls
Target Kit provided
OnTarget Variant Calls

Input:
Sorted and indexed BAM from the alignment module or if user wants to use BAM as an input to the workflow then user inputted BAM. **To reduce the computational time we chop the BAM into per chromosome and execute whole module in parallel for each chromosome**.

Output:
OnTarget variant calls

Folders created:
Output_folder/realignment

This folder contains per sample folder, which includes per chromosome realigned and recalibrated and indexed BAM obtained using GATK.

```
<output_folder>
      | _ realignment
            | _ sample1
                  | _ chr*-sorted.bam
                  | _ chr*-sorted.bam.bai
                  | _ chr*.pileup
                  | _ chr*.flagstat
            | _ sample12
```

Useful files:

```
realignment/Sample/chr*-sorted.bam
realignment/Sample/chr*-sorted.bam.bai
realignment/Sample/chr*.pileup
```

Output_folder/variants

Folder contains all the per chromosome raw and filtered variant calls per sample.

```
<output_folder>
      | _ variants
         | _ SNV
            | _ sample.chr*.snvs.raw.snvmix / sample.chr*.snps.raw.gatk.vcf
            | _ sample.chr*.raw.snvs
         | _ INDEL
            | _ sample.chr*.indel.gatk.vcf
            | _ sample.chr*.raw.indels
         | _ logs
```

Useful files:

- output_folder/variants/SNV

  Folder contains raw and filtered per chromosome variant calls for each sample

  RAW

  `sample.chr*.snvs.raw.snvmix / sample.chr*.snps.raw.gatk.vcf`

  FILTERED

  `sample.chr*.raw.snvs`
- output_folder/variants/INDEL

  RAW

  `sample.chr*.indel.gatk.vcf`

  FILTERED

  `sample.chr*.raw.indels`

Output_folder/OnTarget

Folder contains OnTarget variant, pileup and BAM

```
<output_folder>
      | _ OnTarget
         | _ sample.chr*.raw.snvs.bed.i.ToMerge
         | _ sample.chr*.raw.indels.bed.i.ToMerge
         | _ sample.chr*.pileup.bed.i
         | _ sample.chr*.cleaned-sorted.bam.i
      | _ logs
```

Useful files:

- OnTarget SNVs

  `sample.chr*.raw.snvs.bed.i.ToMerge`
- OnTarget INDELs

  `sample.chr*.raw.indels.bed.i.ToMerge`
- OnTarget Pileup

  `sample.chr*.pileup.bed.i`
- OnTarget BAM

  `sample.chr*.cleaned-sorted.bam.i`

Output_folder/realigned_data

This folder contains per sample folder, which includes realigned and recalibrated and indexed BAM for IGV visualization obtained using GATK

```
<output_folder>
      | _ realigned_data
            | _ sample1
                  | _ sample.igv-sorted.bam
```

```
          | _ sample.igv-sorted.bam.bai
      | _ sample2
```

Useful files:

`realigned_data/sample/sample.igv-sorted.bam`

Filters Used:

1. Realigned and recalibrated BAM are filtered with quality (MAPQ >= 20)
2. filtered SNVs
   a. mapping and base quality >=20
   b. probability >= 0.8
   c. Reference Homozygous calls thrown out

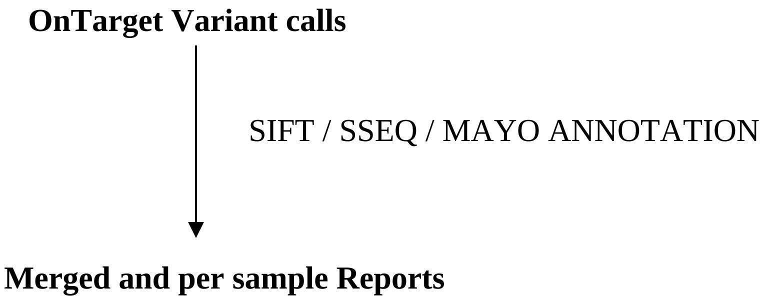
Potential Use of Intermediate data:

a) Intermediate raw variants calls to change the default filtering criterion purely study specific
Raw files can be obtained from `output_folder/variants/` and description of the file names is stated above.

b) To know how realignment and recalibration helps
In `output_folder/realignment` each sample contains `chr*.flagstat` files which has mapping numbers after realignment.

c) Interested in knowing the coverage on the Target regions
In `output_folder/OnTarget` each sample have OnTarget BAM's and Pileup and expected file names are described above.

Annotation and Reports



OnTarget Variant calls

SIFT / SSEQ / MAYO ANNOTATION

Merged and per sample Reports

Input:
OnTarget variant calls for each sample

Output:
Merged and per sample reports and variant distance for unique variants over all the samples

Folders created:
Output_folder/annotation

Folder contains annotation from SIFT and Seattle Seq

```
<output_folder>
      |_ annotation
         | _ SIFT
            |_ siftids
            |_ sift.out.allsamples.chr*.merge
            |_ sample.chr*.raw.snvs.bed.i.ToMerge.sift
            |_ sift folder (per chromosome per sample)
         | _ SSEQ
            |_ sample.chr*.snv.sseq
            |_ sample.chr*.indels.sseq
            |_ sseq.snvs.out.allsamples.chr*.merge
            |_ sseq.indels.out.allsamples.chr*.merge
         | _ logs
```

Useful files:

- output_folder/annotation/SSEQ
  This folder has per chromosome Seattle seq results for each sample
  `sample.chr*.snv.sseq` and `sample.chr*.indels.sseq`
  And this folder also contains merged Indels and SNVs annotation form Seattle seq for each chromosome which is used to create merged report
  `sseq.snvs.out.allsamples.chr*.merge` and `sseq.indels.out.allsamples.chr*.merge`

- output_folder/annotation/SIFT
  This folder has per chromosome SIFT annotation. When we run SIFT it gives out a number to a folder where it stores the annotation which we record in a file names `siftids` in the same folder so as to track which folder belongs to which set of variants. The merged sift annotation per chromosome are alos listed as `sift.out.allsamples.chr*.merge`

Output_folder/TempReports

This folder contains all the temporary files generated during adding columns to the OnTarget variant calls for various annotations. Used mainly to debug if something went wrong during the analysis

Output_folder/Reports

Folder contains Merged INDEL and SNV reports with variant distance fro the splice sites for INDELs and SNVs. For variants there are two reports we deliver i.e. Unfiltered and filtered

```
<output_folder>
      |_ Reports
            |_ *.cleaned_annot.xls
            |_ *.cleaned_annot_filtered.xls
            |_ *.report
            |_ *.cleaned.report
            |_ variantLocation_INDELs
            |_ variantLocation_SNVs
```

Useful files:

SNV `SNV.cleaned_annot.xls (file)` and `SNV.cleaned_annot_filtered.xls (file)`
INDEL `INDEL.cleaned_annot.xls (file)` and `INDEL.cleaned_annot_filtered.xls (file)`

Output_folder/Reports_per_Sample

Folder contains per sample INDEL and SNV reports (filtered and unfiltered)

```
<output_folder>
      |_ sample.*.report
      |_ sample.*.cleaned.report
      |_ sample.*.cleaned_annot.xls
      |_ sample.cleaned_annot_filtered.xls
```

Useful files:

SNV `sample.SNV.cleaned_annot.xls (file)` and `sample.SNV.cleaned_annot_filtered.xls (file)`
INDEL `sample.INDEL.cleaned_annot.xls (file)` and `sample.INDEL.cleaned_annot_filtered.xls (file)`

Filters Used:

1. Used to create the Filtered Reports (merged and per sample)
   a. dbSNP130 column does not have an rs ID (novel), OR
   b. functionGVS column having 'missense', 'nonsense', 'splice-3', 'splice-5', 'coding-notMod3', 'utr-3' or 'utr-5' (intron, intergenic and coding-synonymous removed using SeattleSeq annotation - (http://gvs.gs.washington.edu/GVS/HelpSNPSummary.jsp)), OR
   c. any variant reported within +/-2bp of an exon edge using 'distance' report for variants
2. Only OnTarget variant calls are annotated

Potential Use of Intermediate data:

a) Merged and Per sample reports for tertiary analysis
Results can be found in `Reports` and `Reports_per_Sample` folder

b) Above files can be used to filter the number of candidate variants by looking at the functional GVS from Seattle seq column in the reports

c) Detailed column description can be found in `output_folder/ColumnDescription_Reports.xls (file)` for SNVs and INDELs

Sample Statistics

Input:
All the modules completed

Output:
It gives out the quantitative analysis for all the samples.

Folders created:
Output_folder/numbers

Folder contains two files per sample to be used to generate HTML report

```
<output_folder>
      | _ numbers
            | _ sample.out
            | _ sample.coverage.out
      | _ logs
```

Useful files:

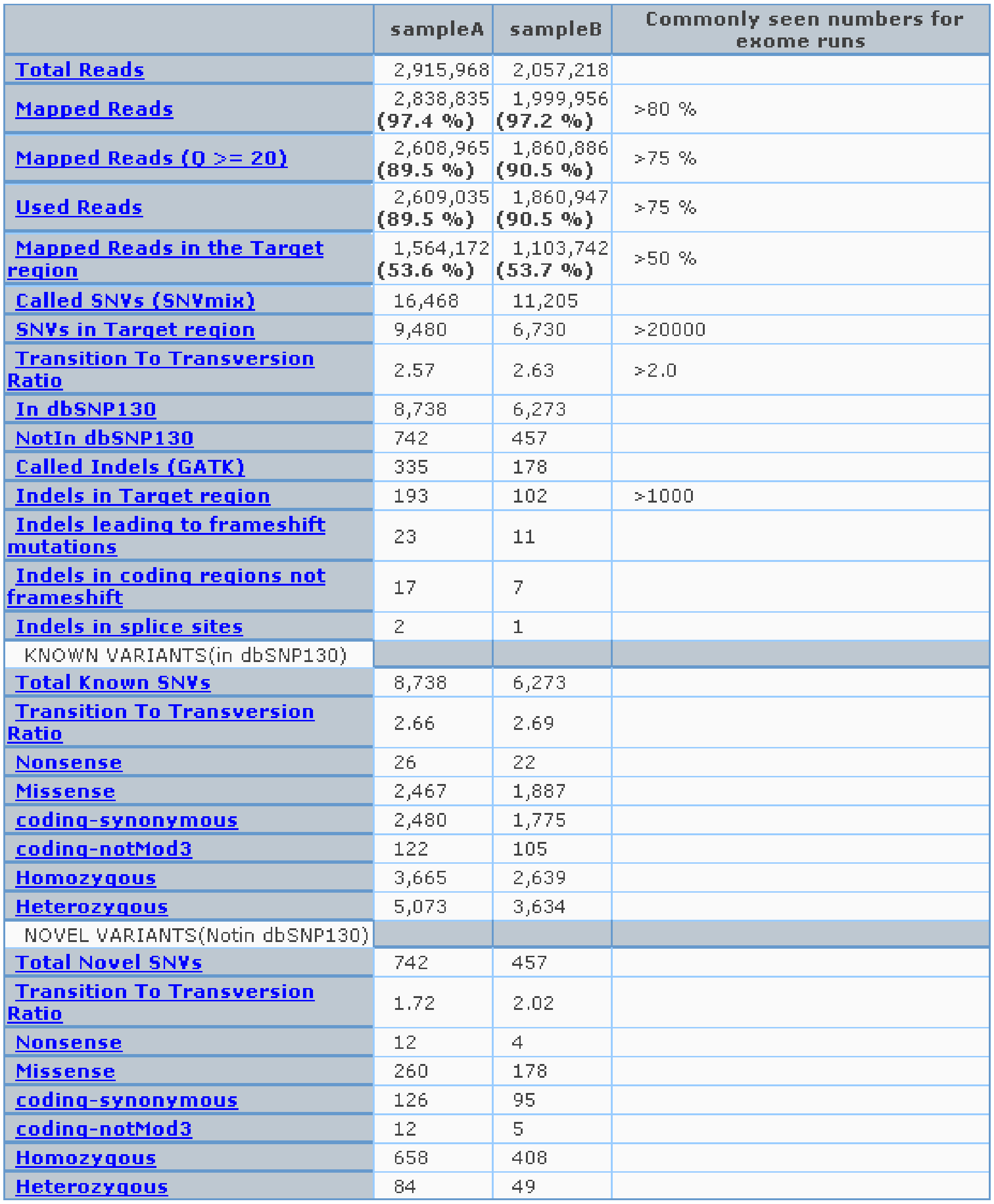
- Contains mapping numbers, annotation numbers, variant numbers, etc.
  `sample.out (file)`
- Contains numbers for coverage at 1 to 40X
  `sample.coverage.out (file)`

Potential Use of Intermediate data:

a) Used to create HTML report, named as Main_Document.html can be found in $output_folder. This report is an intuitive way to summarize all the findings during the analysis. (ClickMe)

b) To get the plot for coverage analysis. (ClickMe)

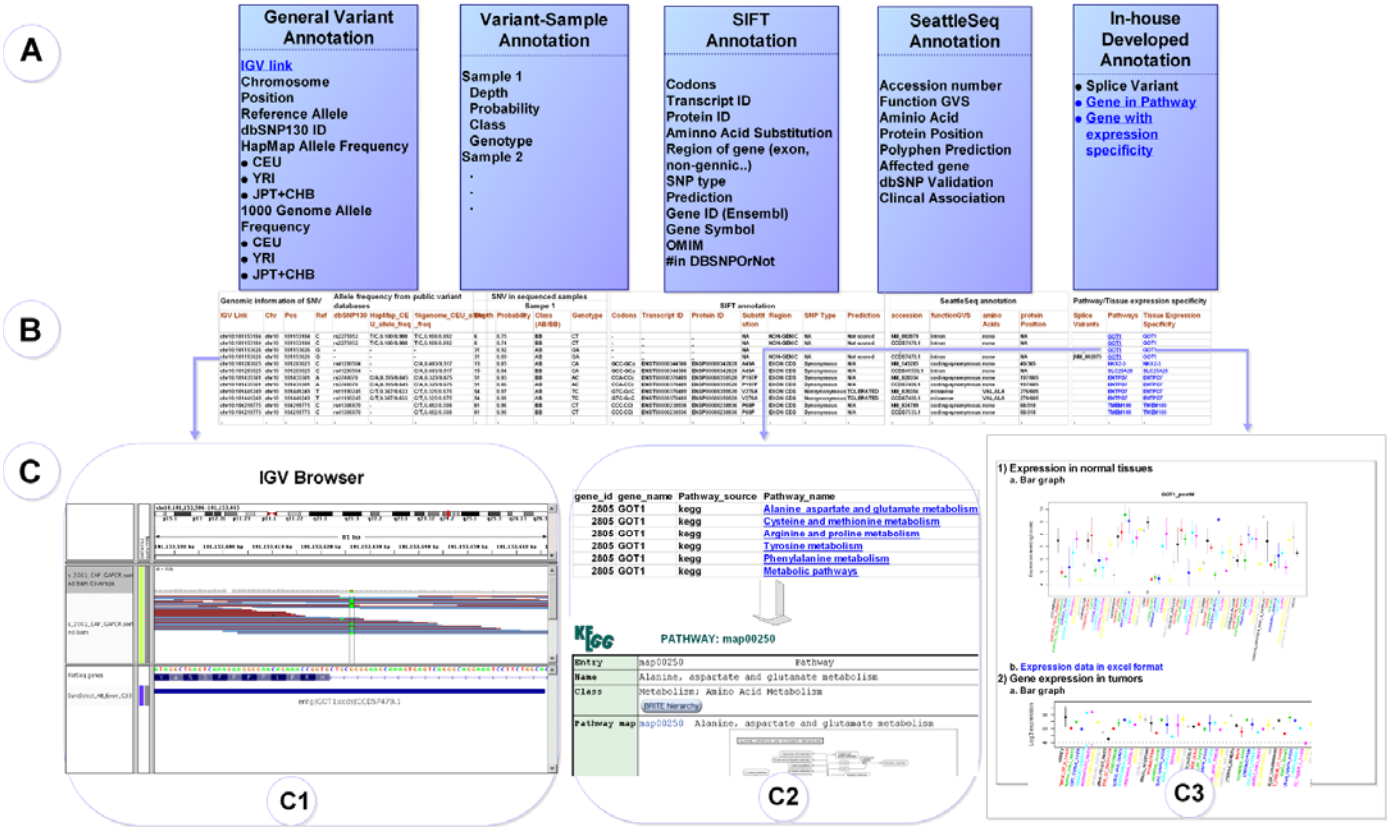
c) Statistical Summary Tables:



	sampleA	sampleB	Commonly seen numbers for exome runs
Total Reads	2,915,968	2,057,218	
Mapped Reads	2,838,835 **(97.4 %)**	1,999,956 **(97.2 %)**	>80 %
Mapped Reads (Q >= 20)	2,608,965 **(89.5 %)**	1,860,886 **(90.5 %)**	>75 %
Used Reads	2,609,035 **(89.5 %)**	1,860,947 **(90.5 %)**	>75 %
Mapped Reads in the Target region	1,564,172 **(53.6 %)**	1,103,742 **(53.7 %)**	>50 %
Called SNVs (SNVmix)	16,468	11,205	
SNVs in Target region	9,480	6,730	>20000
Transition To Transversion Ratio	2.57	2.63	>2.0
In dbSNP130	8,738	6,273	
NotIn dbSNP130	742	457	
Called Indels (GATK)	335	178	
Indels in Target region	193	102	>1000
Indels leading to frameshift mutations	23	11	
Indels in coding regions not frameshift	17	7	
Indels in splice sites	2	1	
KNOWN VARIANTS(in dbSNP130)			
Total Known SNVs	8,738	6,273	
Transition To Transversion Ratio	2.66	2.69	
Nonsense	26	22	
Missense	2,467	1,887	
coding-synonymous	2,480	1,775	
coding-notMod3	122	105	
Homozygous	3,665	2,639	
Heterozygous	5,073	3,634	
NOVEL VARIANTS(Notin dbSNP130)			
Total Novel SNVs	742	457	
Transition To Transversion Ratio	1.72	2.02	
Nonsense	12	4	
Missense	260	178	
coding-synonymous	126	95	
coding-notMod3	12	5	
Homozygous	658	408	
Heterozygous	84	49	

d) Expected Reports format with column descriptions (ClickMe)

TREAT Report Metadata and Feature Illustration



General Variant Annotation

IGV link
Chromosome
Position
Reference Allele
dbSNP130 ID
HapMap Allele Frequency
- CEU
- YRI
- JPT+CHB

1000 Genome Allele Frequency
- CEU
- YRI
- JPT+CHB

Variant-Sample Annotation

Sample 1
Depth
Probability
Class
Genotype
Sample 2
.
.
.

SIFT Annotation

Codons
Transcript ID
Protein ID
Aminno Acid Substitution
Region of gene (exon, non-gennic..)
SNP type
Prediction
Gene ID (Ensembl)
Gene Symbol
OMIM
#in DBSNPOrNot

SeattleSeq Annotation

Accession number
Function GVS
Aminio Acid
Protein Position
Polyphen Prediction
Affected gene
dbSNP Validation
Clincal Association

In-house Developed Annotation

- Splice Variant
- Gene in Pathway
- Gene with expression specificity

IGV Browser

gene_id	gene_name	Pathway_source	Pathway_name
2805	GOT1	kegg	Alanine, aspartate and glutamate metabolism
2805	GOT1	kegg	Cysteine and methionine metabolism
2805	GOT1	kegg	Arginine and proline metabolism
2805	GOT1	kegg	Tyrosine metabolism
2805	GOT1	kegg	Phenylalanine metabolism
2805	GOT1	kegg	Metabolic pathways

PATHWAY: map00250

Entry	map00250	Pathway
Name	Alanine, aspartate and glutamate metabolism	
Class	Metabolism; Amino Acid Metabolism BRITE hierarchy	
Pathway map	map00250 Alanine, aspartate and glutamate metabolism	

1) Expression in normal tissues
a. Bar graph

b. Expression data in excel format

2) Gene expression in tumors
a. Bar graph