Post analysis of SNV calls: Annotating, filtering and quality assessment

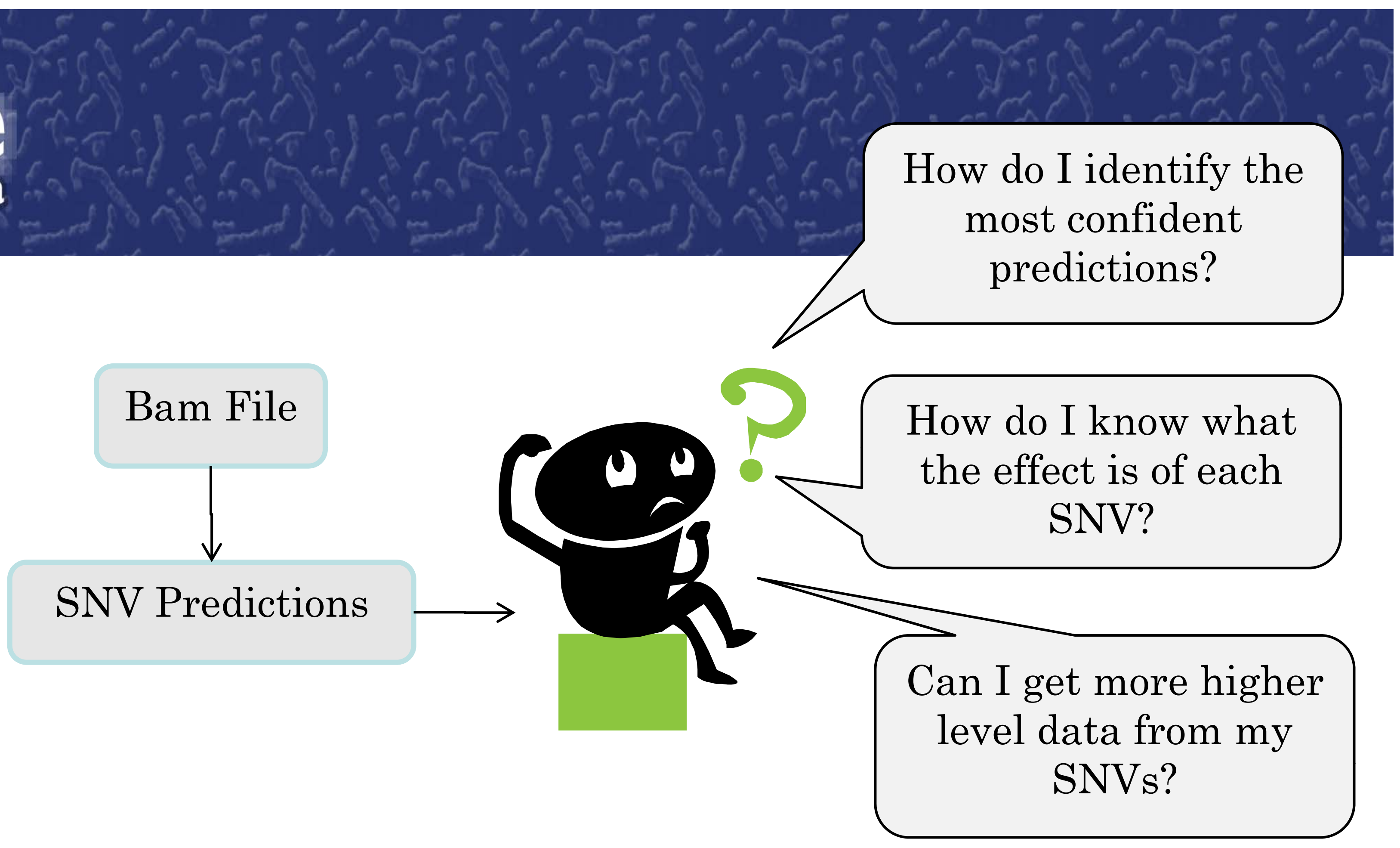
Yaron S. Butterfield*, Richard Corbett*, Steve JM Jones, İnanç Birol

Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada
* These authors contributed equally.

Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca

Abstract

With single nucleotide variations (SNVs) being identified in ever increasing numbers with next generation sequencing data, we are often interested in a breakdown of the effect of the mutations, and if a particular change is somatic or germline. We are also interested in high quality true positive mutations, which are increasingly important considering their use in clinical decisions. Here we present a suite of tools we developed to help us address these needs in our high-throughput sequencing and analysis environment at the British Columbia Genome Sciences Centre.


Bam File

SNV Predictions

How do I identify the most confident predictions?

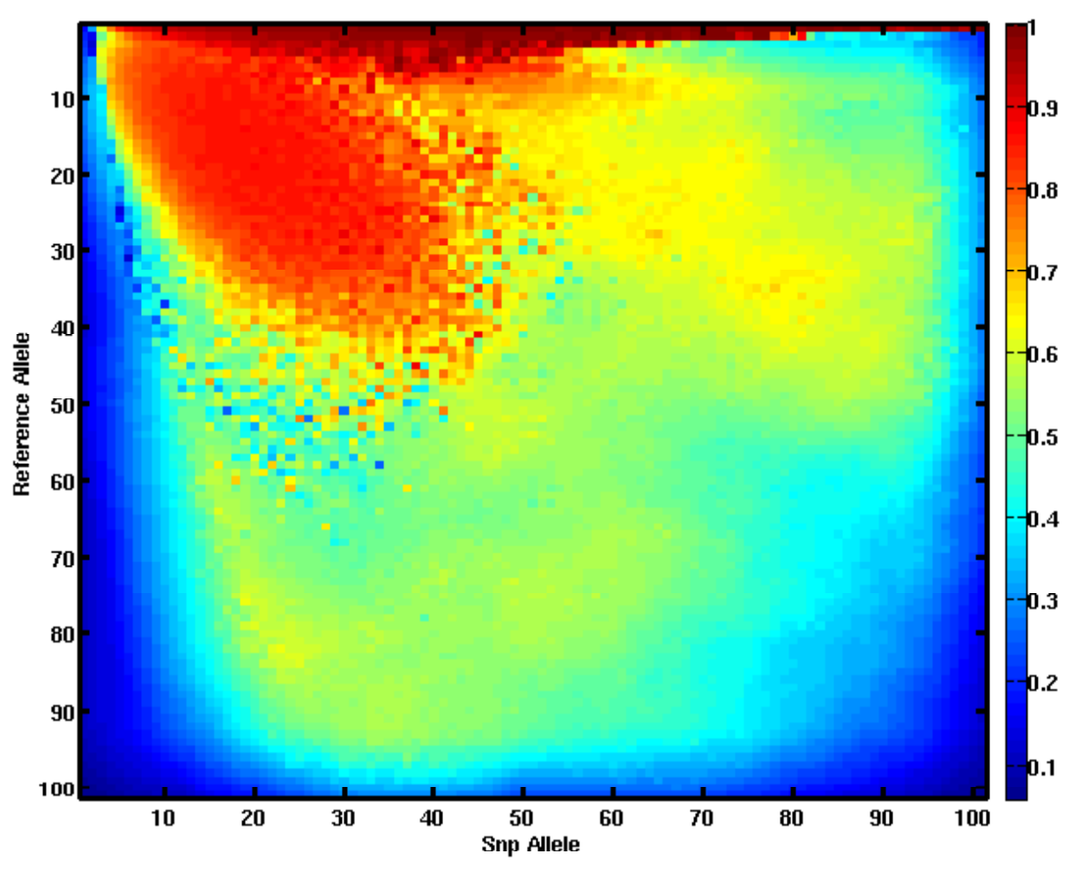
How do I know what the effect is of each SNV?

Can I get more higher level data from my SNVs?

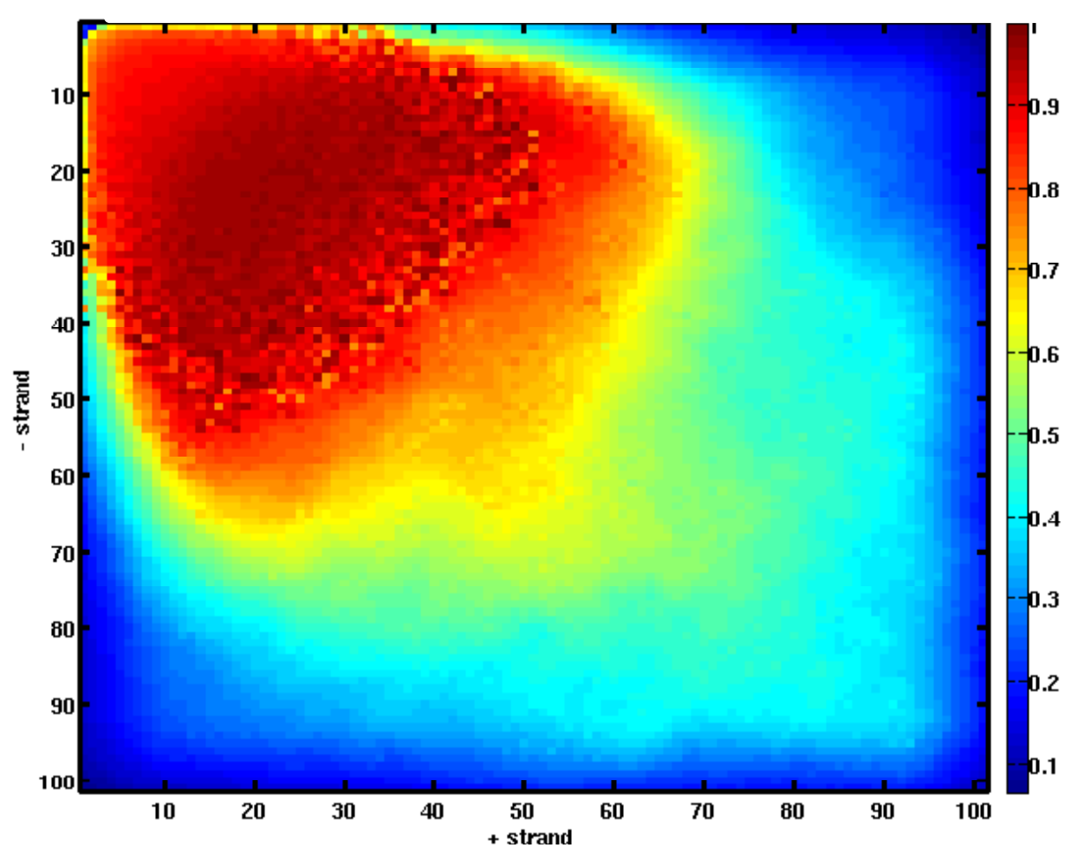
SNV filtering

Before SNVs are annotated, a level of filtering can be applied to improve confidence and decrease false positives. SNVs are scored by a combination of allelic frequencies, strandedness, read positions, and SNV caller-reported quality. For each mutation, we count how many times the SNV or the reference allele is present, as well as which strand and read position the mutant base comes from. We then partition the read pileup file into groups of unique value for each of the metrics. By calculating the dbSNP concordance along the full scale of each of our metrics, we are able build a lookup table that to estimate this concordance for any SNV identified in a similar library. This allows us to take a pileup file and rank the identified SNVs by order of confidence. Below we show how each of these metrics relates to dbSNP concordance.

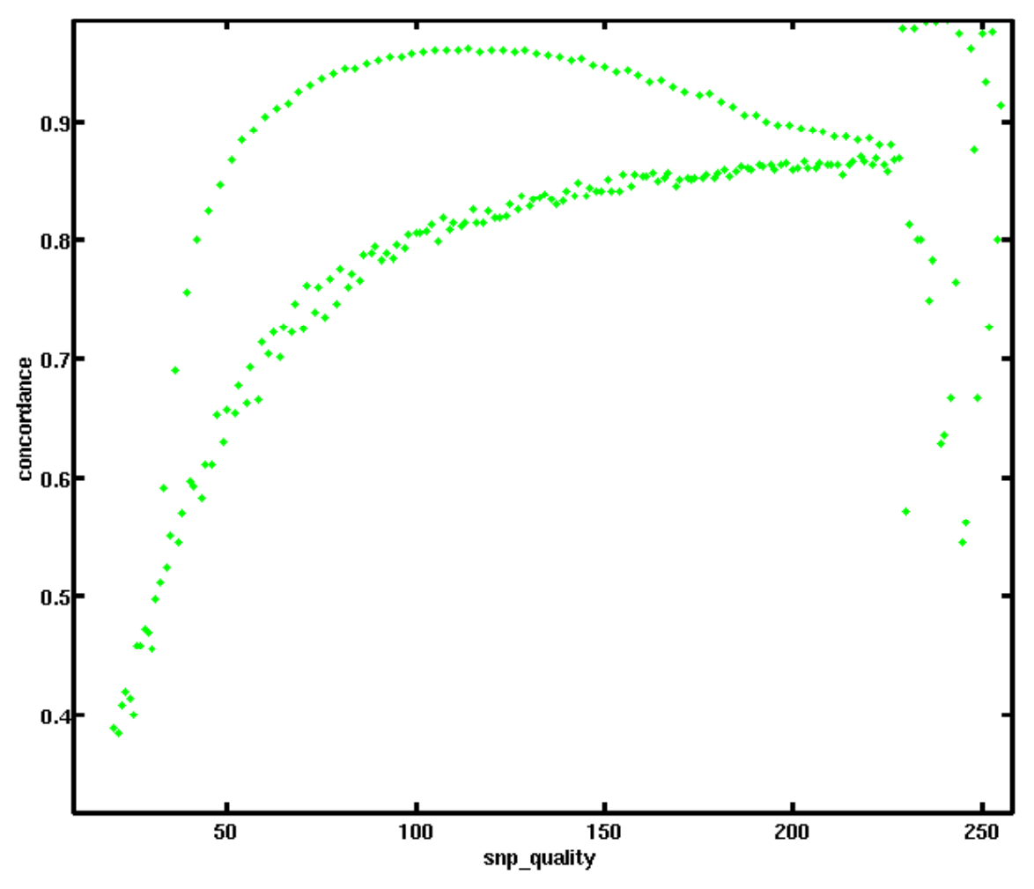
SNP Concordance at Allele Ratios



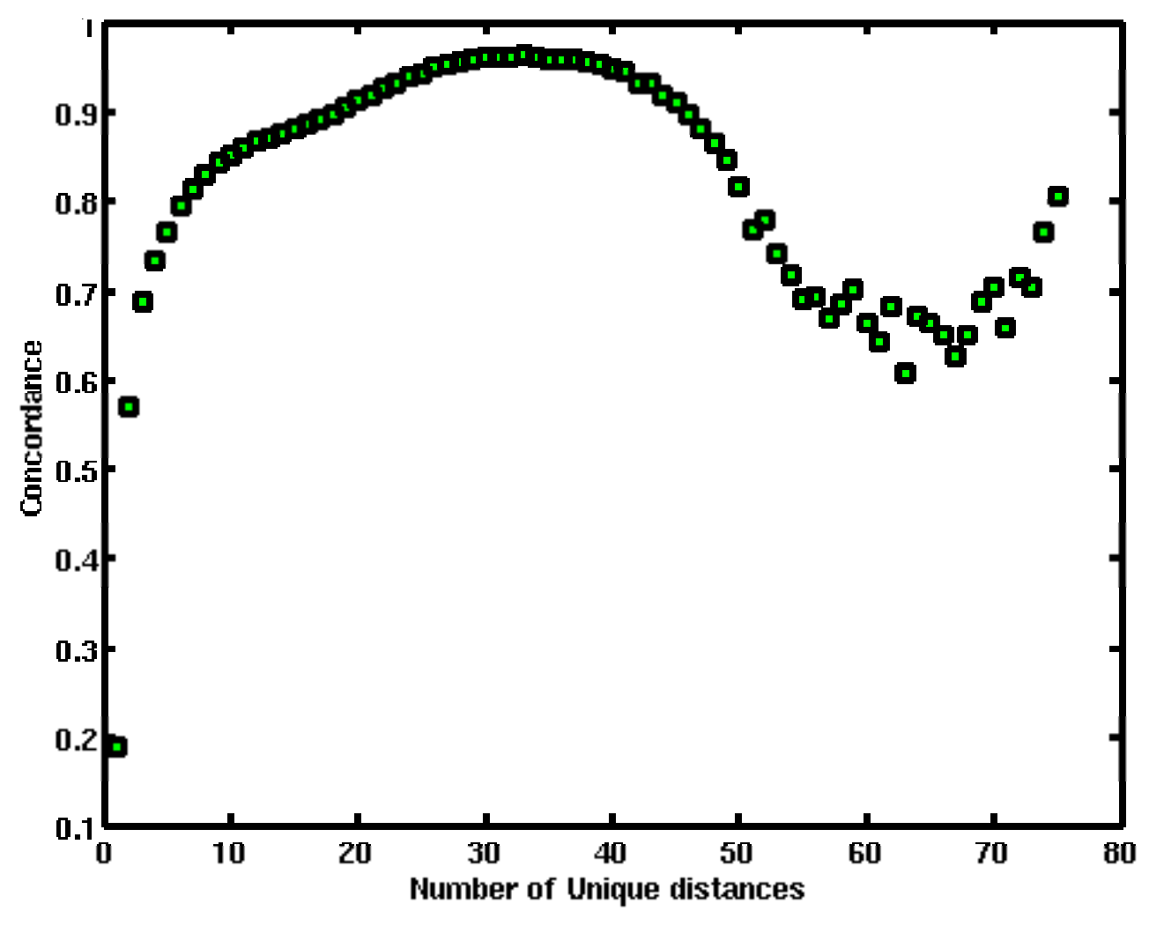
SNP Concordance at Strand Ratios



SNP Concordance vs SAMtools SNP Quality

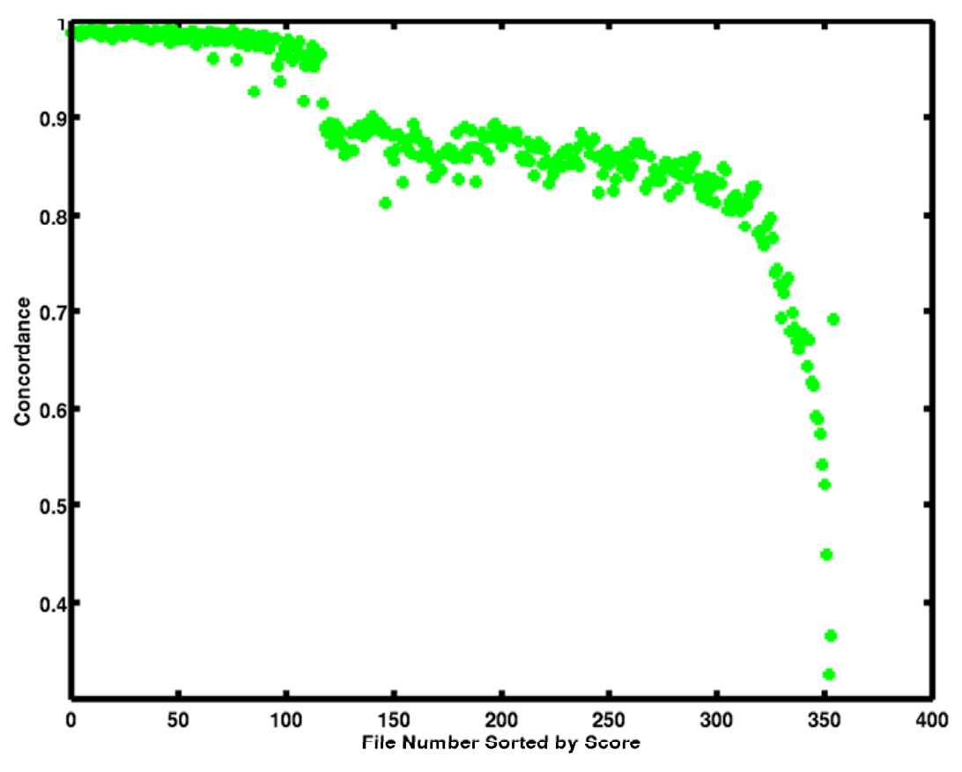


SNP Concordance vs # Unique Read Positions



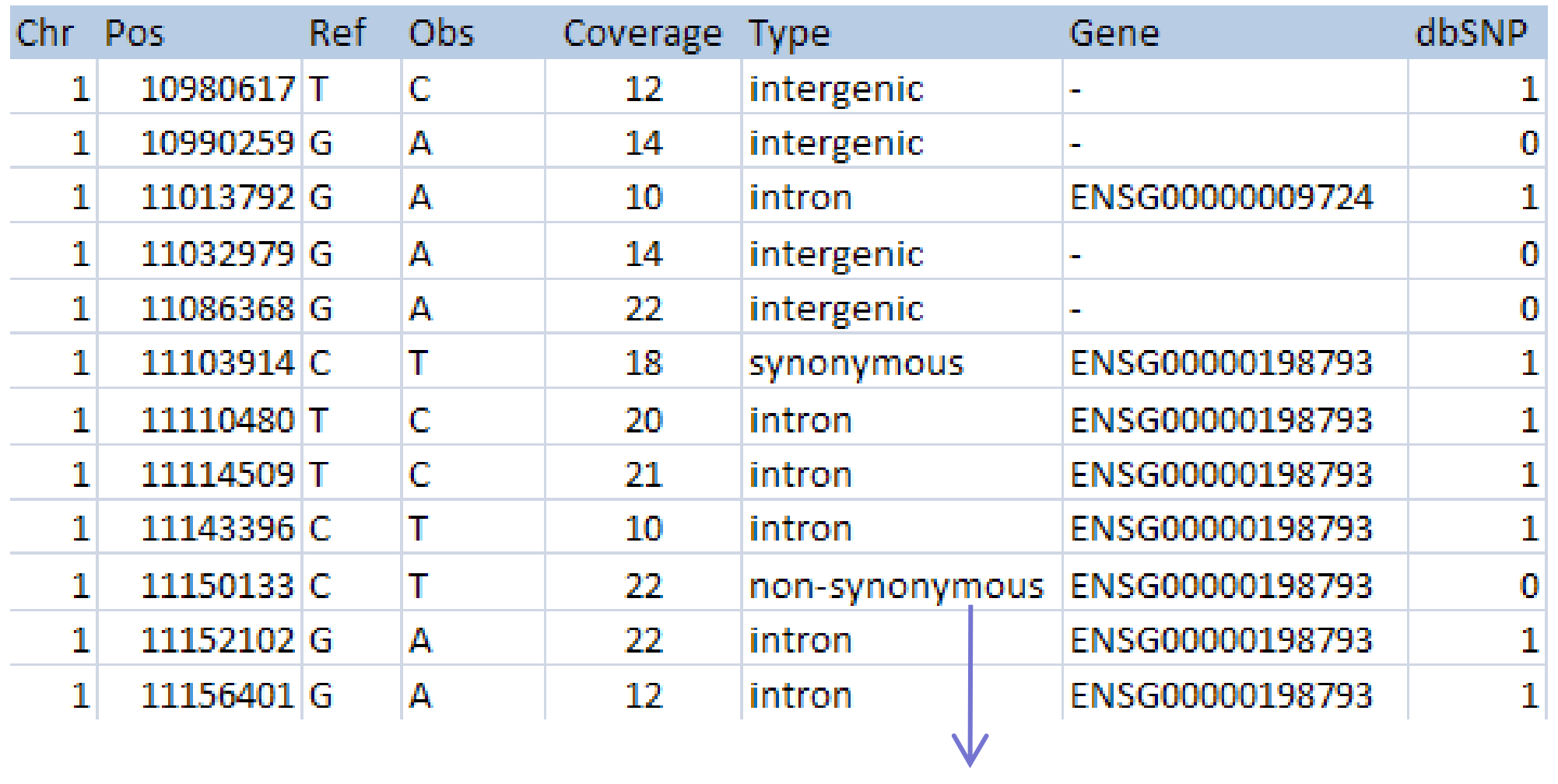
Finally we show how our meta score stratifies the SNPs in a library not used for training.

SNP Concordance vs Min(SNP Scores)



SNV annotator

SNVannotator is a python script part of the PASsiT package that takes a list of mutations from various sources such as Samtools' Pileup/varFilter, GFF, VarScan, VCF, or SNVMix, and classifies the effect of each SNV. Using reference data from Ensembl, SNVannotator outputs an annotated list of SNVs identifying if a mutation is intergenic or intragenic; and if the latter, what gene it is in, and if it's in the 5' or 3' UTR, intron or exon. If in an exon, the mutation is marked as synonymous or non-synonymous according to its effect. SNVs are identified as known or novel with respect to the list of polymorphisms recorded in the dbSNP repository. In addition, a list of novel non-synonymous SNVs is generated with the associated protein and its amino acid change resulting from the mutation. If a matched tumor/normal pair is given, SNVannotator identifies somatic non-synonymous mutations by comparing the SNV calls on both. See example output below.



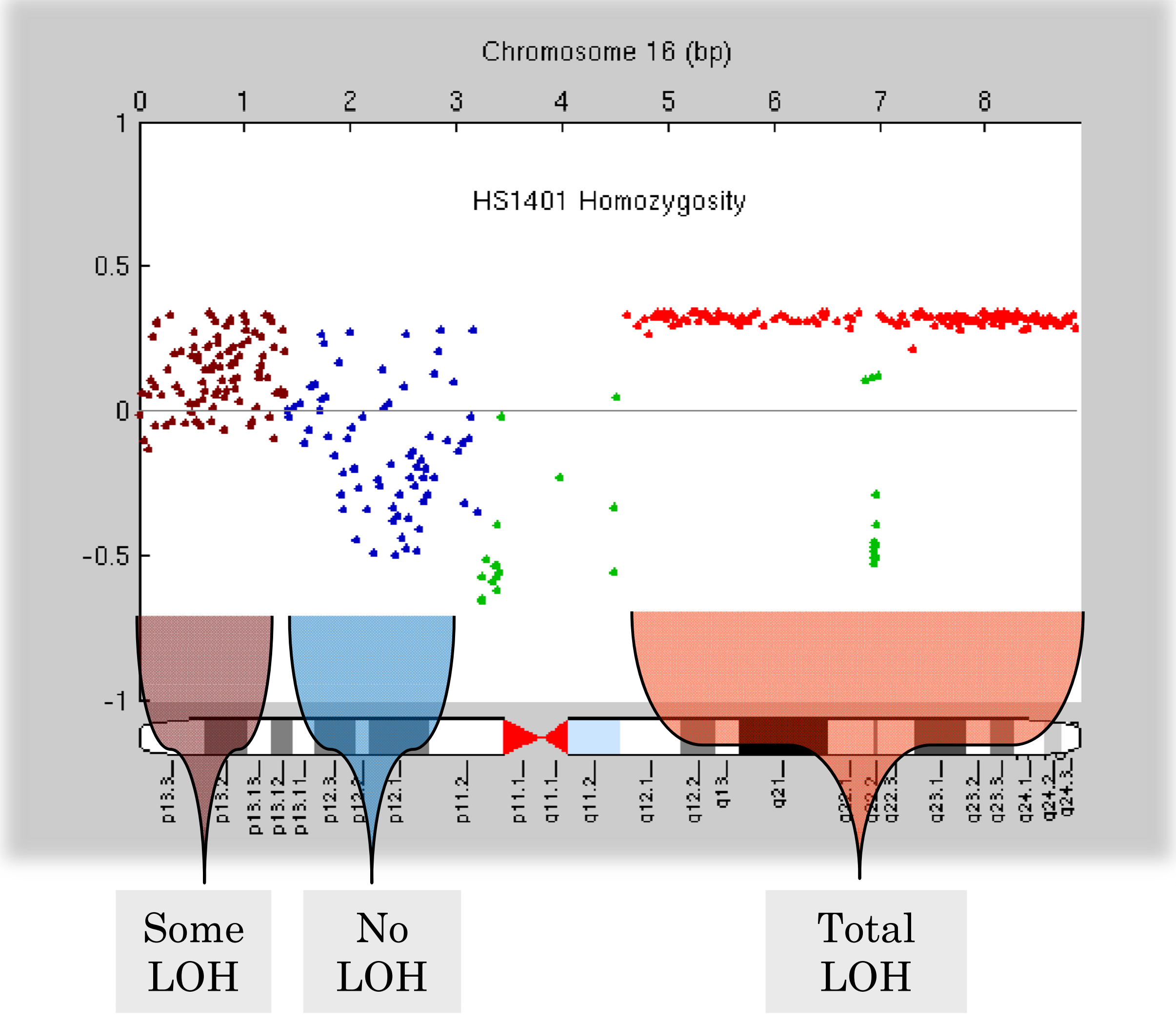
Chr	Pos	Ref	Obs	Coverage	Type	Gene	dbSNP
1	10980617	T	C	12	intergenic	-	1
1	10990259	G	A	14	intergenic	-	0
1	11013792	G	A	10	intron	ENSG00000009724	1
1	11032979	G	A	14	intergenic	-	0
1	11086368	G	A	22	intergenic	-	0
1	11103914	C	T	18	synonymous	ENSG00000198793	1
1	11110480	T	C	20	intron	ENSG00000198793	1
1	11114509	T	C	21	intron	ENSG00000198793	1
1	11143396	C	T	10	intron	ENSG00000198793	1
1	11150133	C	T	22	non-synonymous	ENSG00000198793	0
1	11152102	G	A	22	intron	ENSG00000198793	1
1	11156401	G	A	12	intron	ENSG00000198793	1

Total number of SNPs: 2165750
dbSNP 129 concordance: 0.9044
Intergenic: 1353654
Intragenic: 812096
  not coding: 8503
  5-UTR: 1977
  3-UTR: 20610
  intron: 764327
  synonymous: 8521
  non-synonymous: 8158
novel non-synonymous: 792
somatic novel non-synonymous: 192

1 11150133 C T - 22 FRAP1 ENSG00000198793 ENST00000361445 4282 1428 A T EFQKGPTPAILESLISINNKLQQPE=A=AAGVLEYAMKHFGELEIQATWYEKL FKBP12-rapamycin complex-associated protein (FK506-binding protein 12-rapamycin complex-associated protein 1)(Rapamycin target protein)(RAPT1)(Mammalian target of rapamycin)(mTOR) [Source:UniProtKB/Swiss-Prot;Acc:P42345] FRAP1

LOH detection

Using the called SNVs and their estimated zygosity states, we also identify regions of loss of heterozygosity (LOH). For each sample, genomic bins of consistent SNV zygosity states are used by a hidden Markov model (HMM) to identify genomic regions of consistent rates of heterozygosity. The HMM partitions each tumor genome into three states: normal heterozygosity, increased homozygosity (low), and total homozygosity (high), where the intermediate state of low homozygosity represents a genomic region where only a portion of the cellular population sampled had lost one of the alleles.



Availability

PASsiT
Post Alignment SNV Tools

http://www.bcgsc.ca/platform/bioinfo/software/passit

For more information on the application of this tool, please see:
Butterfield Y. Integrative Genomic and Transcriptome Analysis of Oligodendroglioma Using Next Generation Sequencing Technology, RECOMB 2011.

Annotation of SNVs were done by an in-house Python package