SNP calling

Jose Blanca
COMAV institute
bioinf.comav.upv.es

SNP calling


IGV
Human hg18
chr1
chr1:19,332,457-19,334,716
Go
chr 1
chr 7
chr1:19,332,457-19,334,716
chr7:66,172,833-66,175,092
NAME
DATA FILE
DATA TYPE
NA12878 SLX (CEU daughter) Coverage
NA12878 SLX (CEU daughter)
NA12878 SOLID (CEU daughter) Coverage
NA12878 SOLID (CEU daughter)
RefSeq genes
UBR4
TYW1
6 tracks loaded
346M of 734M

Genotype matrix

		Marker Set Dimension						
		rs10001	rs10002	rs10003	rs10004	rs10005	rs10006	...
Sample Set Dimension	Sample1	AA	GG	CC	AC	GG	CC	
	Sample2	AA	GG	CT	AC	CG	CC	
	Sample3	AT	GG	TT	AA	CG	TT	
	Sample4	TT	GG	TT	AA	GG	CC	
	Sample5	AA	GG	TT	AA	GG	TT	
	Sample6	AA	GA	CC	AC	GG	CC	
	Sample7	AT	GG	CC	CC	GG	CC	
	Sample8	AA	GG	CT	AA	GG	CT	
	...							

Genotype Dimension

Genotype matrix: Samples x SNPs

SNPs and errors

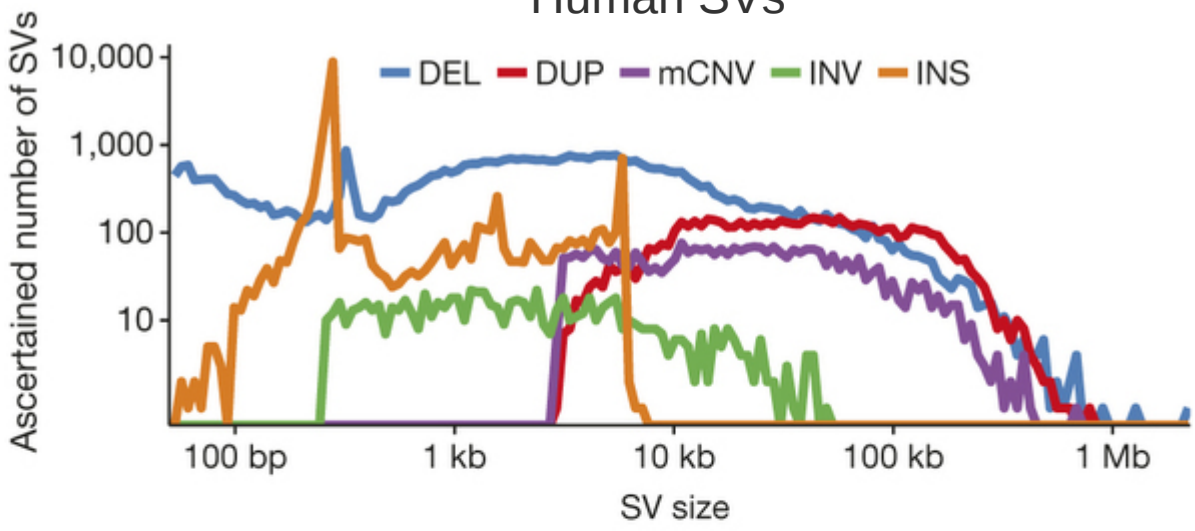
A change in a read may due to:

- Sample contamination
- Cloning or PCR artifacts
  - Homopolymer stuttering
  - Hexamers in RNASeq libraries
- Sequencing errors
- Mapping errors
  - Problems related to structural variants
  - Alignment errors
- Reference genome errors
- Real SNVs

Two kinds of errors:

- False positives
- False negatives

Structural variants are frequent!!!

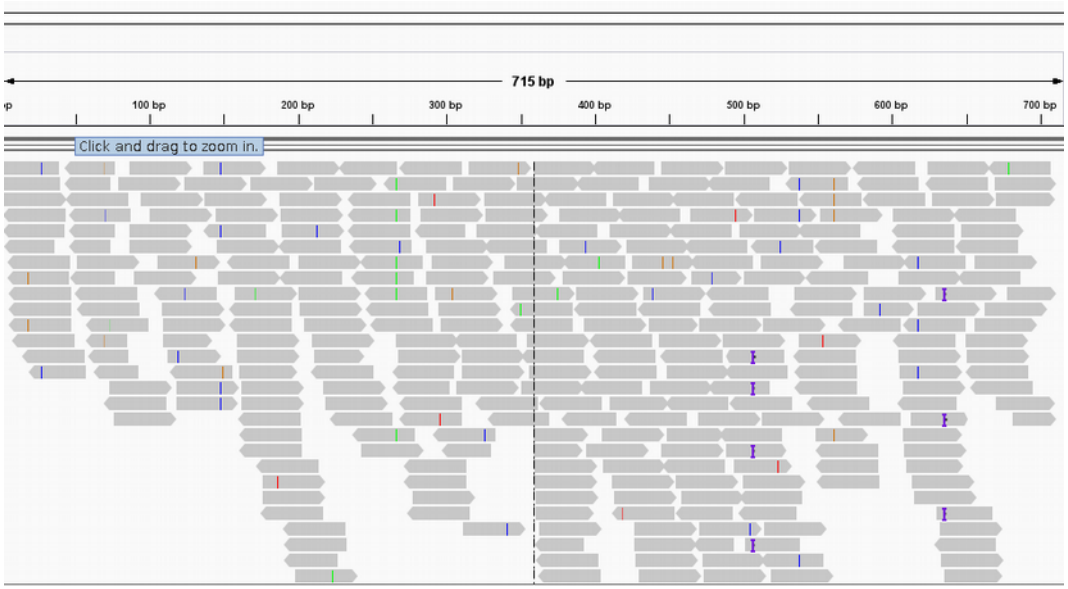


SNP calling errors

In a Hiseq 2500 line there are

if Q= 30:

- 150 millions of sequencing errors



SNP calling alignment errors

```
Ref       ...aggttttataaaac----aattaagtctacagagcaacta...
Sample    ...aggttttataaaacAAATaattaagtctacagagcaacta...
Read1     ...aggttttataaaac****aaAtaa
Read2     ....ggttttataaaac****aaAtaaTt
Read3     ........ttataaaacAAATaattaagtctaca............
read4                 CaaaT****aattaagtctacagagcaac......
read5                   aaT****aattaagtctacagagcaact.....
read6                     T****aattaagtctacagagcaacta....
```

Strategies to mitigate this problem:

- Fix the problem.
  - GATK, GLIA realignment.
  - It realigns the problematic regions (lots of SNPs or some indels).
  - Computationally slow.
  - It does not fix all problems.

- Avoid using the misaligned positions.
  - Samtools BAQ (calmd).
    - For each position It calculates the probability of being misaligned.
  - Avoid read ends

SNP callers

Simple algorithms based on:

- Allele counts in reads
- Allele rate
- Total coverage

Complex bayesian algorithms add

- Hardy-Weimberg equilibrium
- Linkage disequilibrium

More common SNP callers:

- GATK
- Freebayes

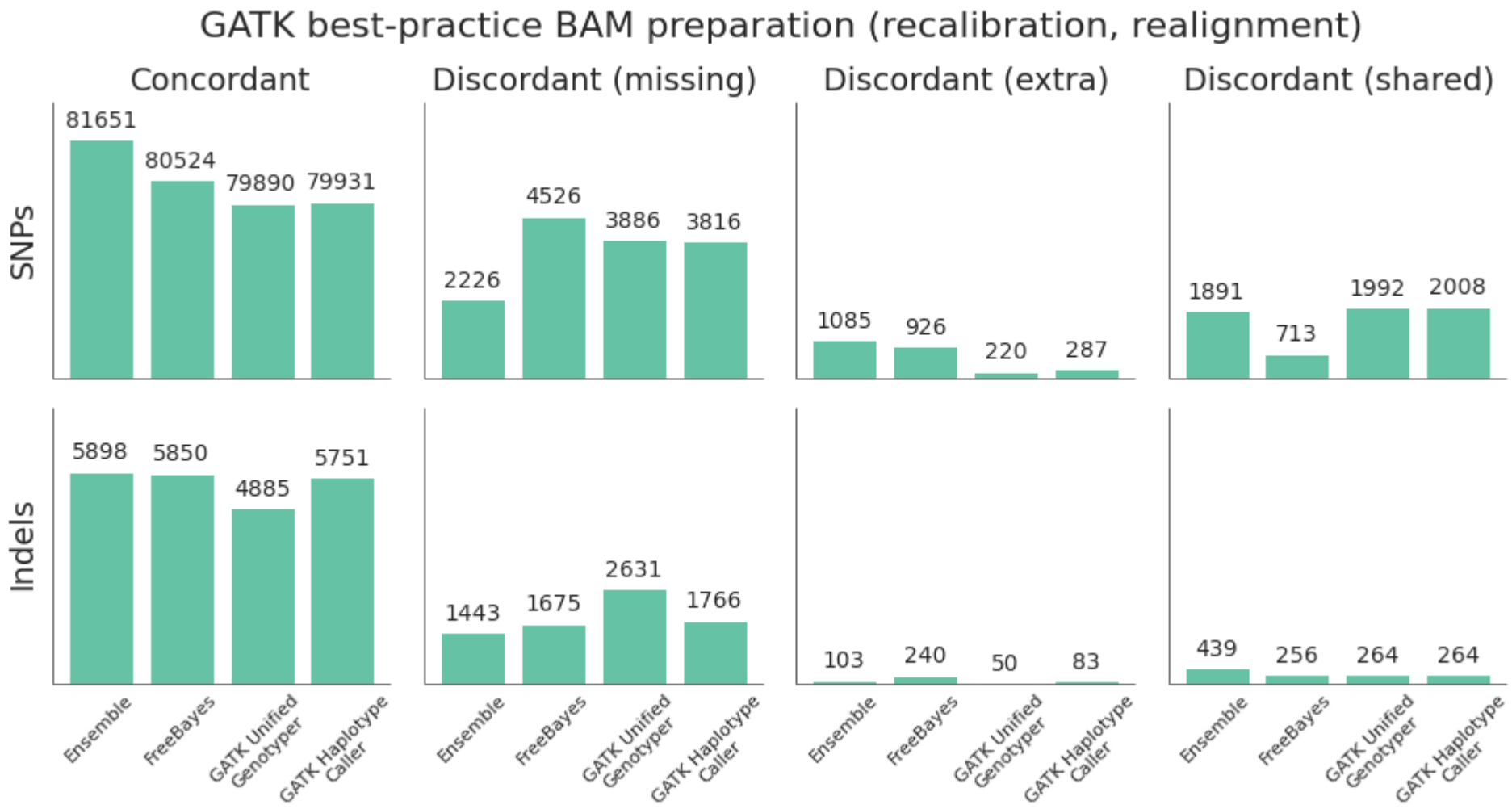
SNP calling

Brad Chapman has a very interesting piece comparing the use of aligners and SNP callers.

- de-duplication with Picard MarkDuplicates, GATK base quality score recalibration and GATK realignment
- Minimal post-processing, with de-duplication using samtools rmdup and no realignment or recalibration.

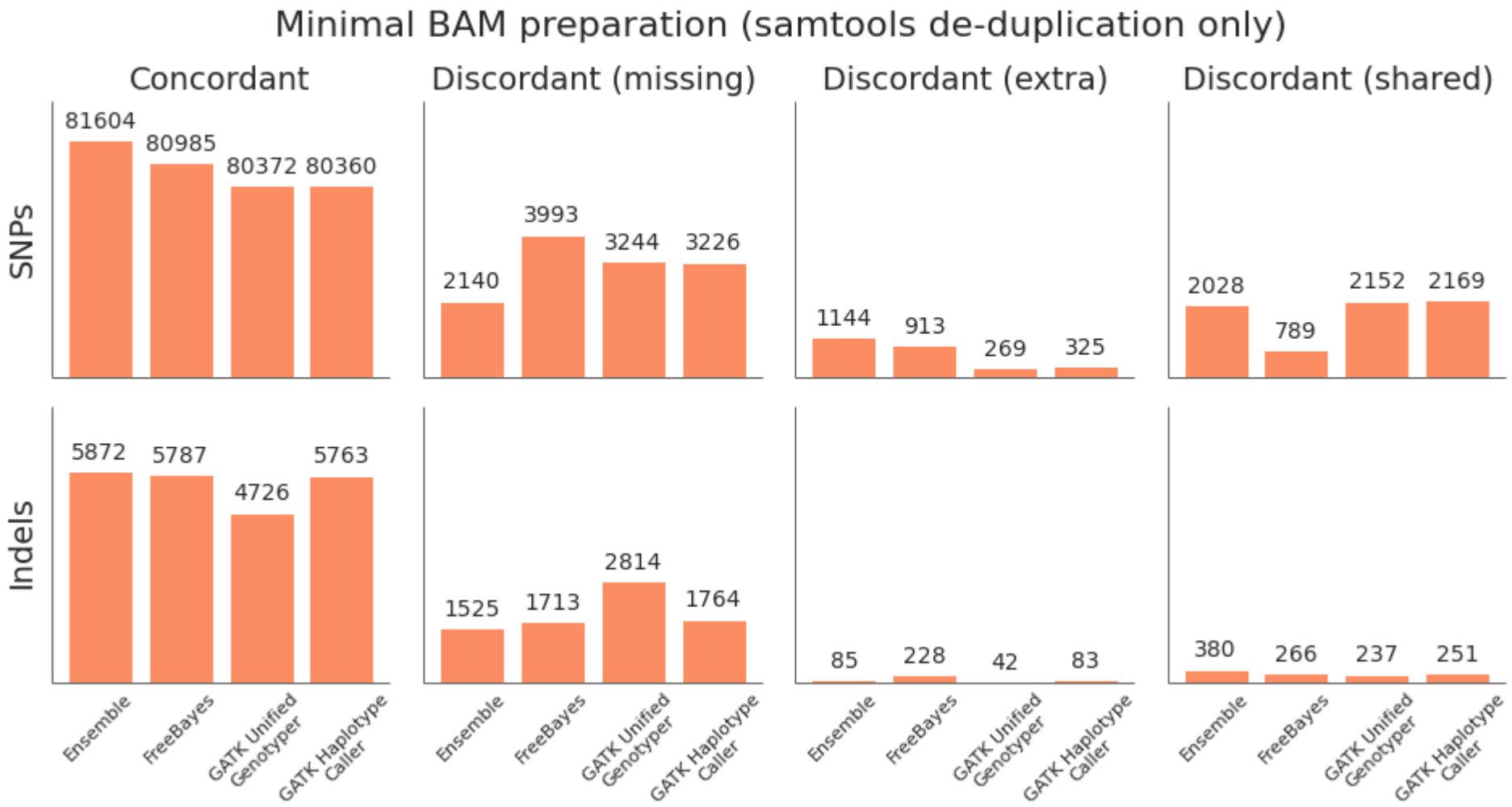
- FreeBayes (v0.9.9.2-18): A haplotype-based Bayesian caller from the Marth Lab.
- GATK UnifiedGenotyper (2.7-2): GATK's widely used Bayesian caller.
- GATK HaplotypeCaller (2.7-2): GATK's more recently developed haplotype caller which provides local assembly around variant regions

SNP calling



http://bcbio.wordpress.com/

SNP calling



http://bcbio.wordpress.com/

Required coverage

Depends:

- Confidence required
  - SNP vs Genotypes
  - Clinical diagnosis vs marker search
- Sample complexity
  - Diploid
  - Polyploid
  - Pooled samples
  - Heterogeneous cancer
- Ploidy
- Population structure
  - e.g. F2 population vs random mating population

SNP quality assesment

% of missing calls per SNP

SNP depth or Genotype depth

SNP observed heterozygosity

SNP quality

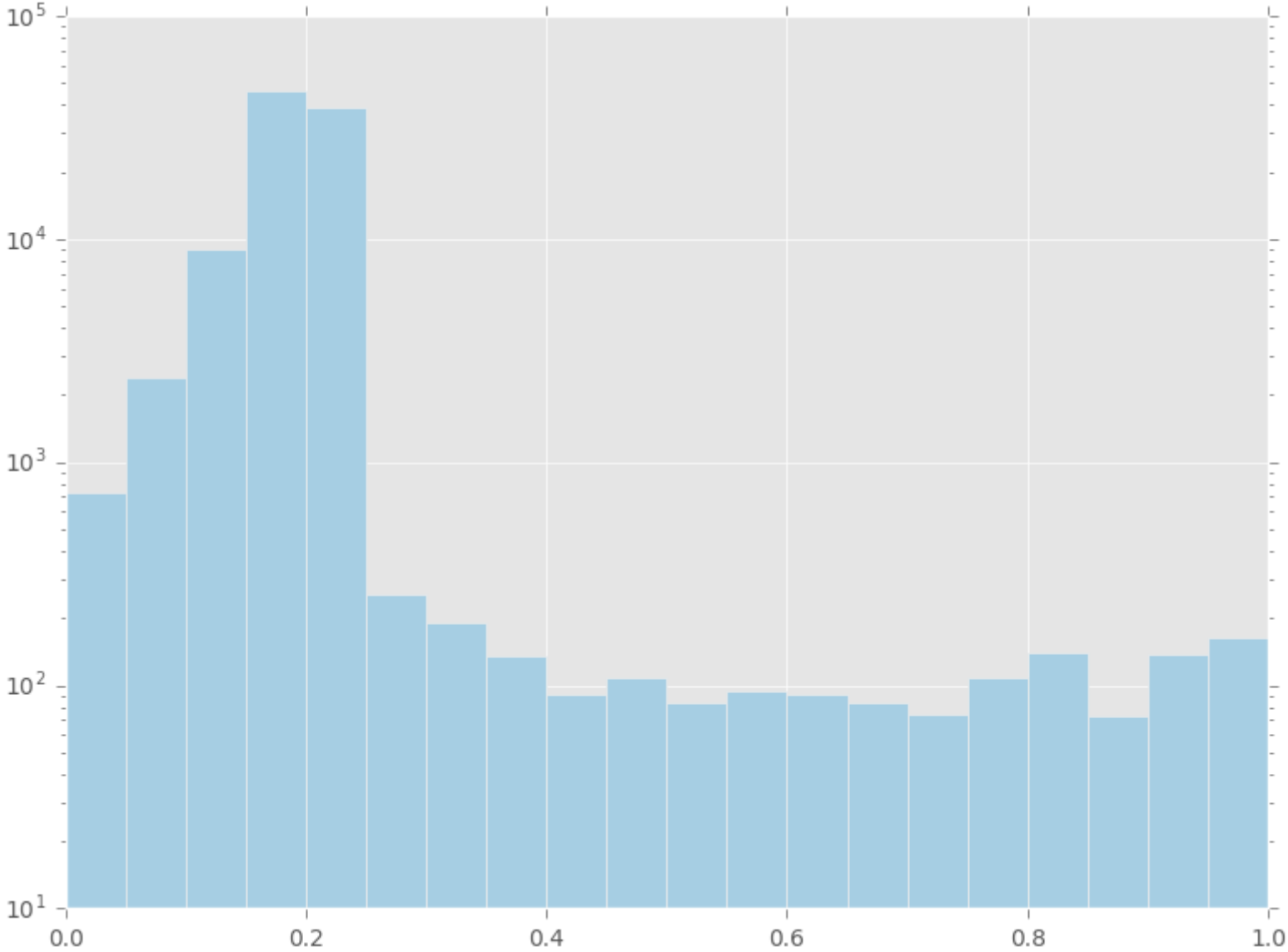
SNP density

Sample depth

% of missing calls per sample

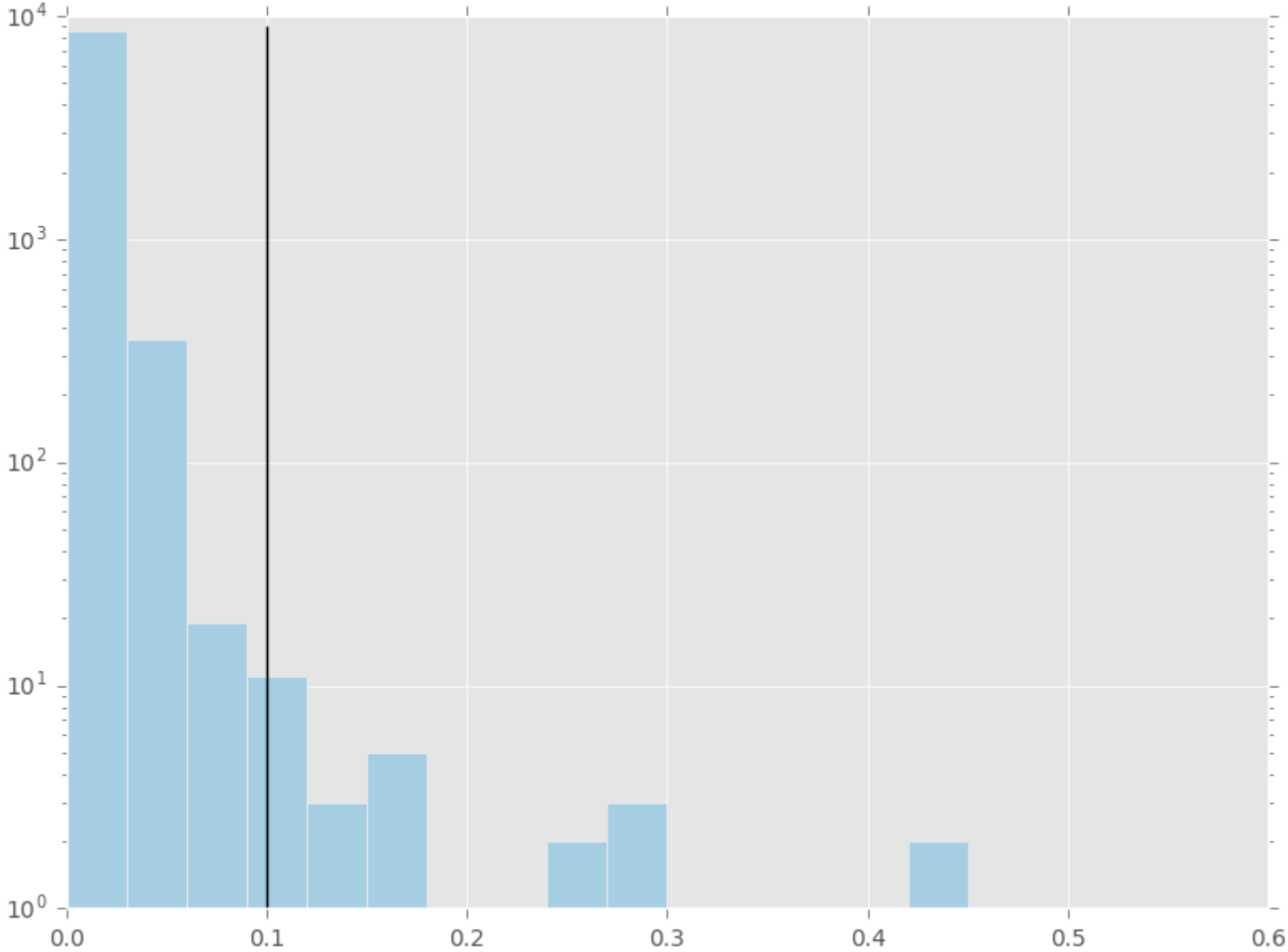
Sample observed heterozygosity

SNP quality assesment



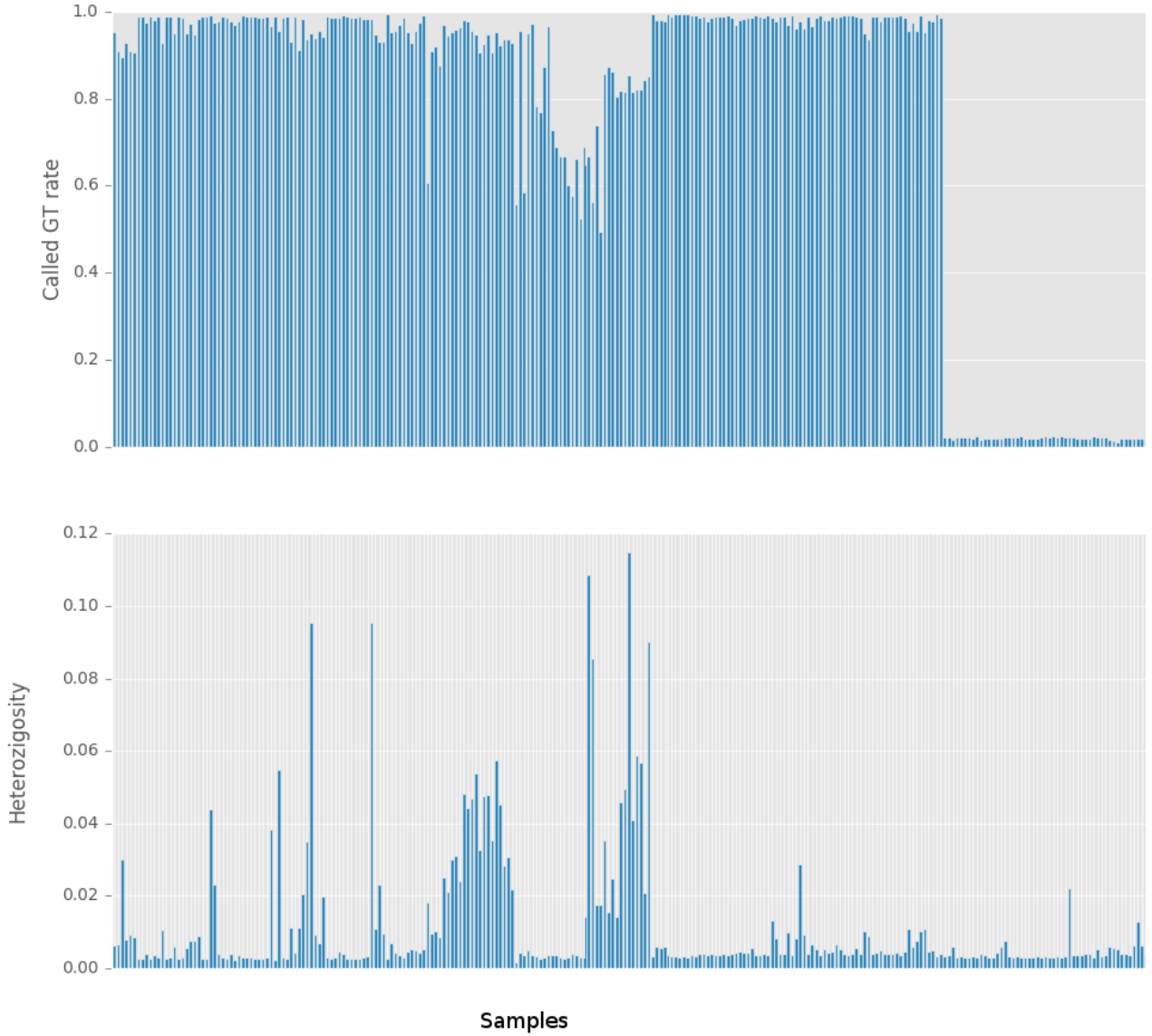
Number of SNPs called at different missing rates

SNP quality assesment



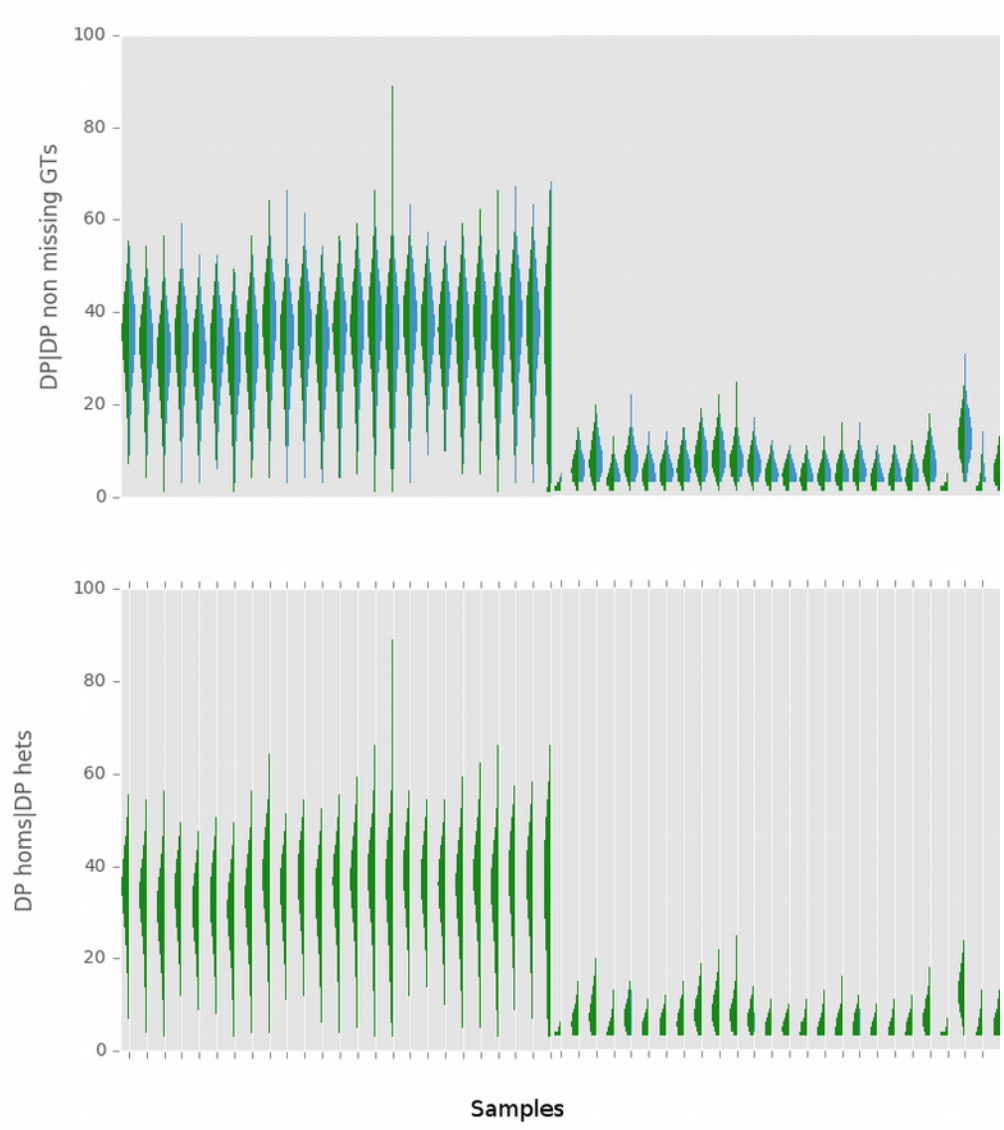
Number of SNPs with different observed heterozygosities

SNP quality assesment



Observed heterozygosity and called GT rate per sample

SNP quality assesment



Depth distributions per sample for all genotype calls and for the non-missing genotype calls

SNP filtering

Filtering the SNPs after the SNP calling is a critical task.

The objetive is to remove false positives, but keeping the correct SNPs.

The filters and parameteres depend of the data and the application.

Vcftools, vcffilter, Annovar…

Home-made filters

SNP filtering

SNPs Low quality:

SNP callers usually assign a quality (probability) to the SNPs. We can filter out the SNPs with lower qualities.

Genotype Low quality:

It is also possible to filter out not SNPs, but genotypes. In this case the genotype is usually set to not determined

Missing data:

We could filter the SNPs with large amount of missing genotypes.

SNP filtering

Number of alleles:

It is possible to remove the monomorphic SNPs or to filter out the SNPs that are not biallelic.

Minor (or Major) Allele Frequency (MAF):

Observed Heterozygosity:

High heterozygosities in populations are suspicious

SNP filtering

High Coverage:

SNPs with an excessive coverage can be false positives due to duplicated regions in the sequenced sample not found in the reference genome.

Highly Variable Region:

Having regions with too many SNPs it is also a sign that we are piling up reads from repeated regions.

Low Complexity Region:

It has been shown that due to problems with the PCR and the alignment the low complexity regions are particularly prone to false positive SNPs.

SNP filtering

Linkage Disequilibrium:

If we have genotype a segregant population it could be useful to filter out the SNPs that are not in linkage disequilibrium with their closest SNPs.

Kind

We can filter the SNVs according to its type: SNV, indel, complex or structural variation.

SNP filtering

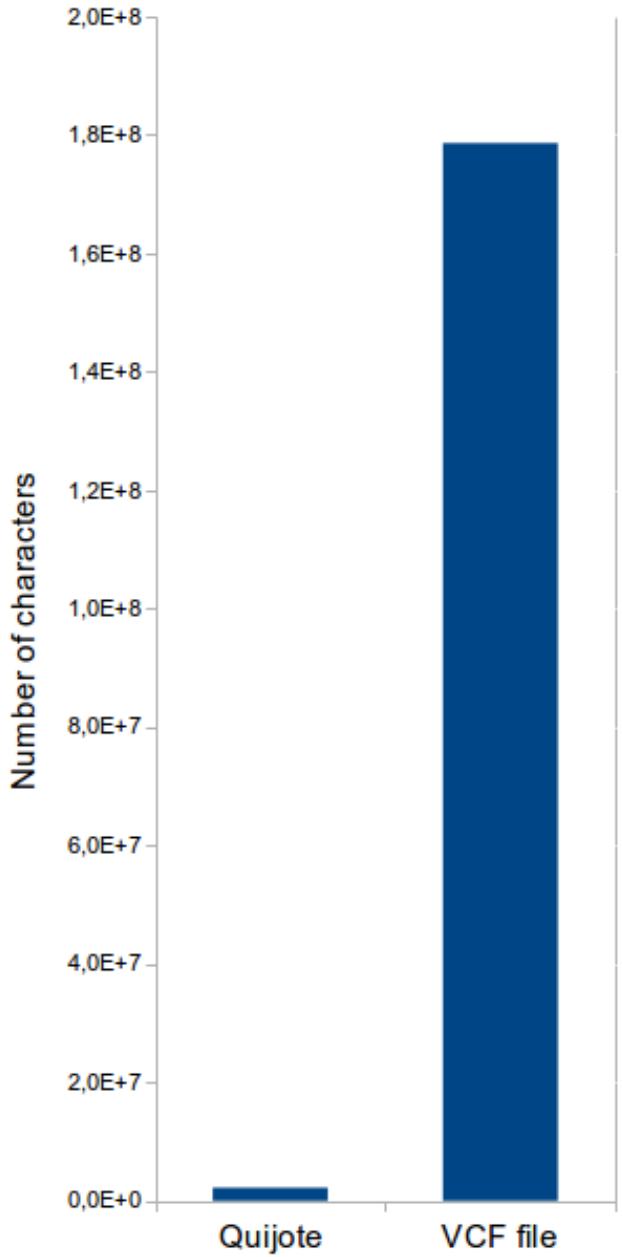
Aminoacid change:

We can select the SNPs with large impacts in the coded proteins.

By genome localization:

We can choose SNPs in a determinate region or kind of features (exon, UTR….)

SNP annotation



filter

NHLBI Exome Sequencing Project (ESP)

Exome Variant Server

OMIM

1000 Genomes

A Deep Catalog of Human Genetic Variation

LOVD

Leiden Open Variation Database

dbSNP

Short Genetic Variations

SNP annotation

RESEARCH ARTICLE

Genetic Variation in an Individual Human Exome

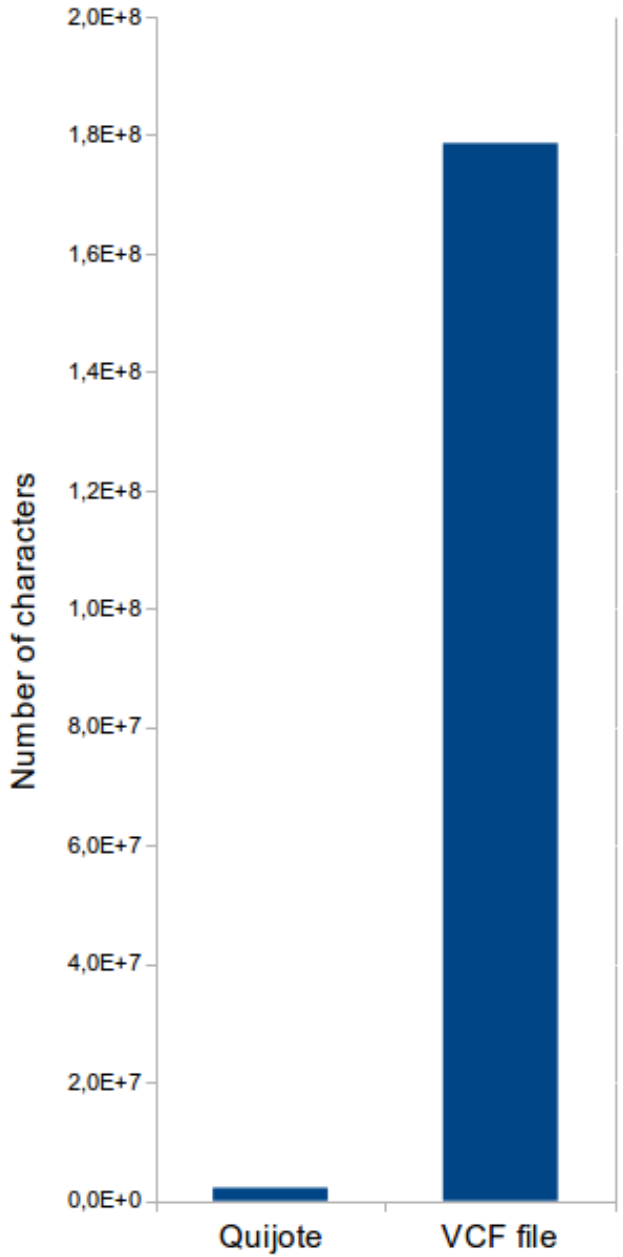
Pauline C. Ng, Samuel Levy, Jiaqi Huang, Timothy B. Stockwell, Brian P. Walenz, K...

Synonymous		10,413
Heterozygous	Novel	551
	dbSNP	5,183
Homozygous	Novel	98
	dbSNP	4,581
Nonsynonymous		10,389
Heterozygous*	Novel	557
	dbSNP	5,047
Homozygous	Novel	215
	dbSNP	4,570

*All heterozygous novel nonsynonymous SNPs were manually inspected (see Methods).

doi:10.1371/journal.pgen.1000160.t001

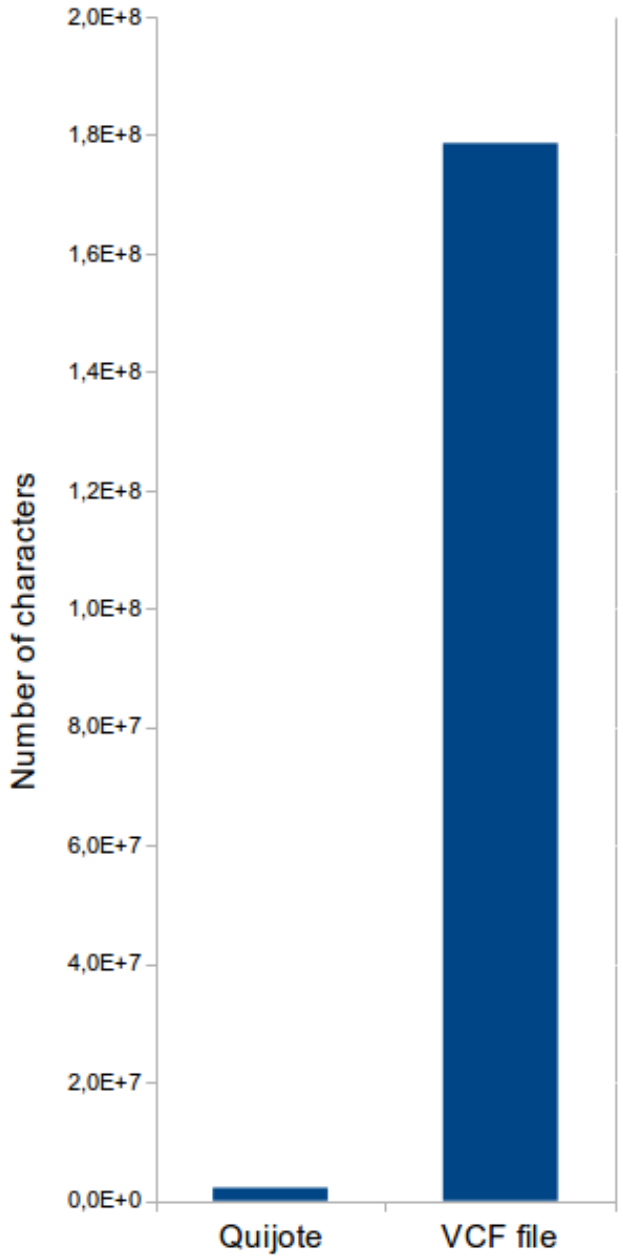
Mutation filtering



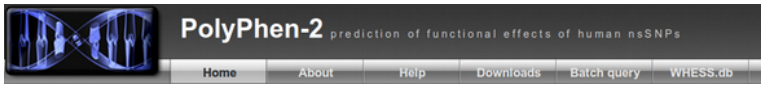
filter

NEW

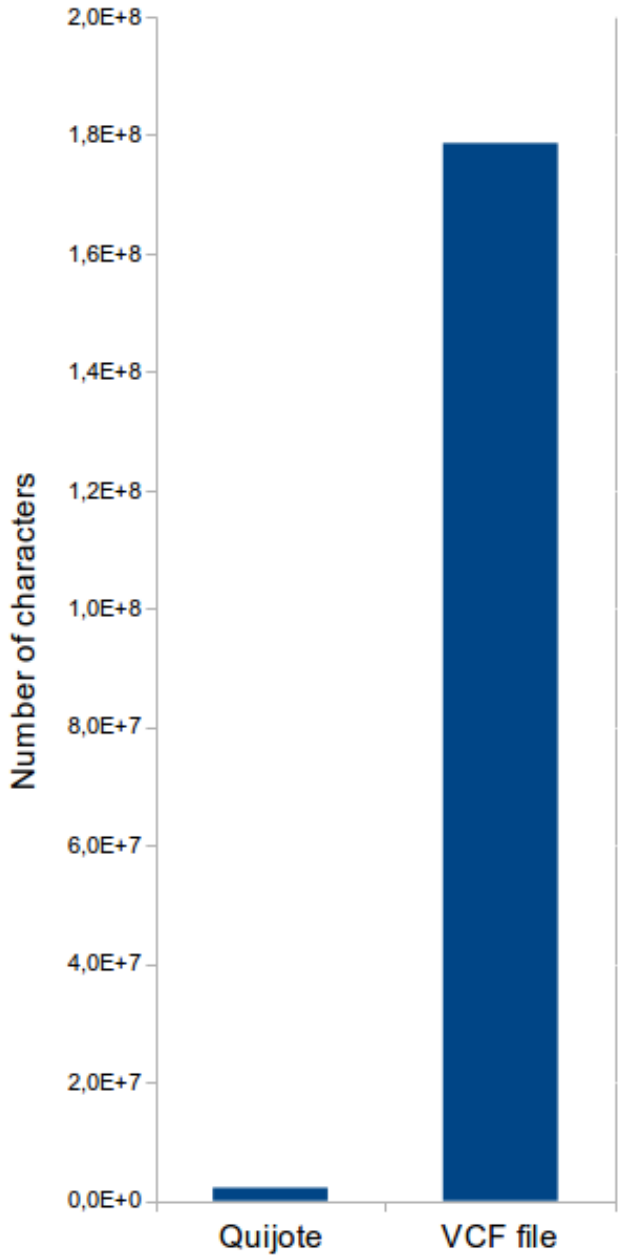
Mutation filtering



New



Mutation filtering



analysis

ANNOVAR™

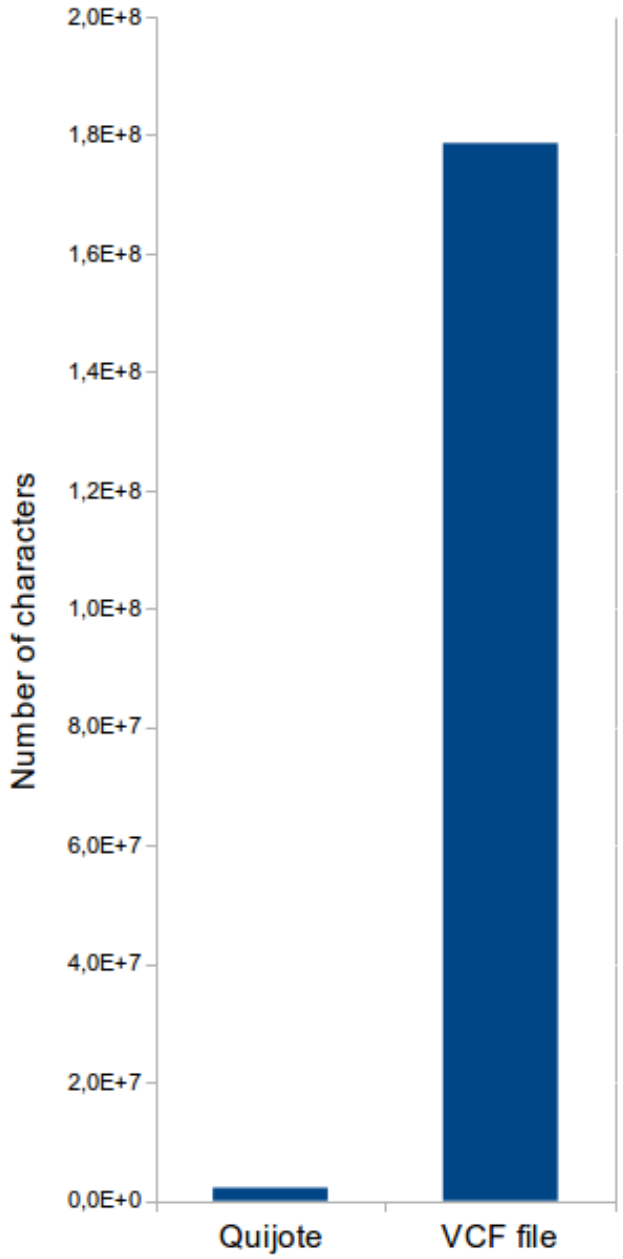
SeattleSeq Annotation 138

AnnTools

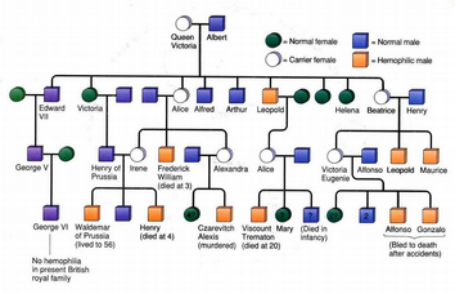
SnpEff

Genetic variant annotation and effect prediction toolbox.

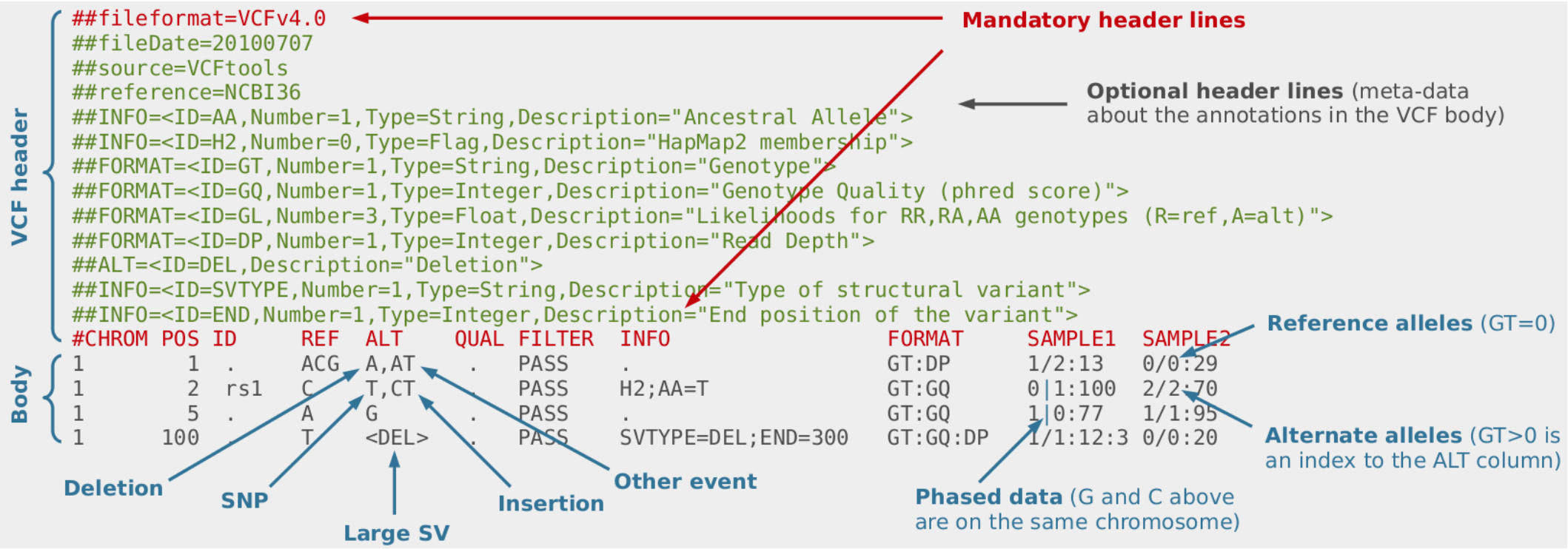
Mutation filtering



Candidate



File format: VCF



File format: VCF

```
SNPs
Alignment  VCF representation
ACGT       POS REF ALT
AtGT       2   C   T

Insertions
Alignment  VCF representation
AC-GT       POS REF ALT
ACtGT       2   C   CT

Deletions
Alignment  VCF representation
ACGT       POS REF ALT
A--T       1   ACG A

Complex events
Alignment  VCF representation
ACGT       POS REF ALT
A-tT       1   ACG AT

Large structural variants
VCF representation
POS REF ALT   INFO
100 T   <DEL> SVTYPE=DEL;END=300
```

File format: GFF

Based on a sequence ontology

Annotation of regions in sequences

```
##gff-version 3
##sequence-region ctg123 1 1497228
ctg123 . gene 1000 9000 . + . ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000 1012 . + . ID=tfbs00001;Parent=gene00001
ctg123 . mRNA 1050 9000 . + . ID=mRNA00001;Parent=gene00001;Name=EDEN.1
```

File format: BED

Annotation of regions in sequences

```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000
```

The BED format uses 0-based coordinates for the starts and 1-based for the ends. So the 1st base on chromosome 1 would be:

```
chr1    0    1    first_base
```