SNP Detection Analysis

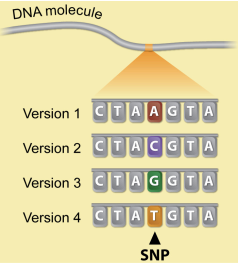
Manpreet S. Katari

PTC Tasting

- Phenylthiocarbamide is a chemical that not everyone can taste.
- Generally about 70% of individuals can taste the bitterness.
- Can we use genetics to explain this phenomenon ?

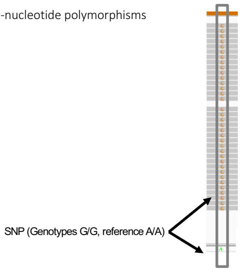
- In order to keep the dataset small we are going to focus on one gene that is known to important in this.
- The sequence for this gene was obtained from four individuals who also performed the taste test.

Single nucleotide polymorphisms (SNPs)



Types of variants

- Single-nucleotide polymorphisms

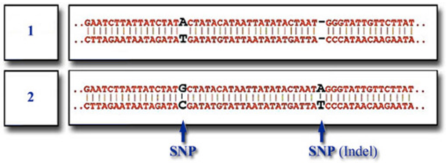


Types of base substitutions

- Transition (Ti):
- purine -> purine mutation
- Or
- pyrimidine -> pyrimidine mutation

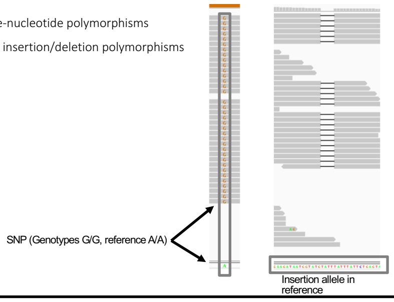
- Transversion (Tv):
- Purine -> pyrimidine mutation
- Or
- Pyrimidine -> purine mutation

Small insertion deletions (INDELS)


1
..GAATCTTATTATCTATACTATACATAATTATATACTAAT-GGGTATTGTTCTTAT..
..CTTAGAATAATAGATATGATATGTATTAATATATGATTA-CCCATAACAAGAATA..
2
..GAATCTTATTATCTATGCTATACATAATTATATACTAATAGGGTATTGTTCTTAT..
..CTTAGAATAATAGATACGATATGTATTAATATATGATTATCCCATAACAAGAATA..
SNP
SNP (Indel)

Types of variants

- Single-nucleotide polymorphisms
- Small insertion/deletion polymorphisms


SNP (Genotypes G/G, reference A/A)

Insertion allele in reference

Mutation detection pipeline

- Alignment: SAM format (BAM = binary)

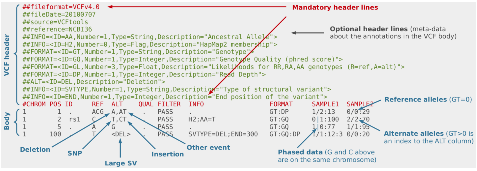
↓

- SNP detection: VCF format (BCF = binary)

↓

- Functional assessment

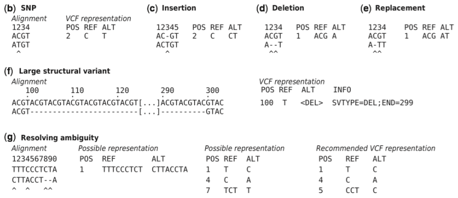
Variant Call Format (VCFs): summary of variants in a genome(s)



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1   SAMPLE2
1        1 .      ACG  A,AT      .  PASS    .                    GT:DP      1/2:13    0/0:29
1        2 rs1    C    T,CT      .  PASS    H2;AA=T              GT:GQ      0|1:100   2/2:70
1        5 .      A    G         .  PASS    .                    GT:GQ      1|0:77    1/1:95
1      100 .      T    <DEL>     .  PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3  0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Phased data (G and C above are on the same chromosome)
Deletion
SNP
Large SV
Insertion
Other event

Variant Call Format (VCFs): representation of variants



```
(b) SNP                                      (c) Insertion        (d) Deletion         (e) Replacement
 Alignment      VCF representation
 1234           POS REF ALT                   12345   POS REF ALT   1234   POS REF ALT   1234   POS REF ALT
 ACGT           2   C   T                     AC-GT   2   C   CT    ACGT   1   ACG A     ACGT   1   ACG AT
 ATGT                                         ACTGT                 A--T                 A-TT
  ^                                             ^                    ^^                   ^^

(f) Large structural variant
 Alignment                                                  VCF representation
    100       110       120       290       300             POS REF  ALT    INFO
 ACGTACGTACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC            100  T   <DEL>  SVTYPE=DEL;END=299
 ACGT-----------------------[...]----------GTAC

(g) Resolving ambiguity
 Alignment     Possible representation        Possible representation   Recommended VCF representation
 1234567890    POS  REF        ALT            POS REF  ALT              POS   REF   ALT
 TTTCCCTCTA    1    TTTCCCTCT  CTTACCTA       1   T    C                1     T     C
 CTTACCT--A                                   4   C    A                4     C     A
 ^  ^   ^^                                    7   TCT  T                5     CCT   C
```

VCF line example

- CHROM: PTC_Human
- POS: 245
- ID: .
- REF: T
- ALT: C
- QUAL: 999
- FILTER: .
- INFO: DP=7609;VDB=0.0040;AF1=0.625;AC1=5;DP4=1985,855,2346,2276;MQ=59;FQ=999;PV4=2e-60,1,4.6e-05,1
- FORMAT: GT:PL:GQ GT=genotype, PL=genotype likelihood,GQ=genotype quality
- 0/1:244,0,248:99
- 0/0:0,255,255:99
- 1/1:255,255,0:99
- 1/1:255,255,0:99

Whole genome-resequencing: snp-calling

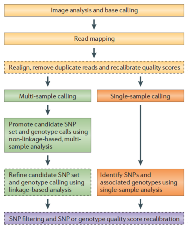


Image analysis and base calling
Read mapping
Realign, remove duplicate reads and recalibrate quality scores
Multi-sample calling
Single-sample calling
Promote candidate SNP set and genotype calls using non-linkage-based, multi-sample analysis
Refine candidate SNP set and genotype calling using linkage-based analysis
Identify SNPs and associated genotypes using single-sample analysis
SNP filtering and SNP or genotype quality score recalibration

Requires assembled genome

Steps to reduce the "false discovery rate" (FDR)

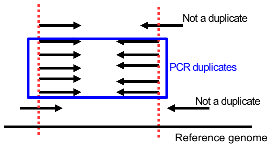
Multi-sample algorithms improve SNP-calling in population studies

Hard-filtering of suspect SNPs is required in most snp-calling and genotyping applications

Nielsen, Nat. Rev. Genetics 2011

Duplicate handling: PCR Duplicates

- What is a PCR duplicate?



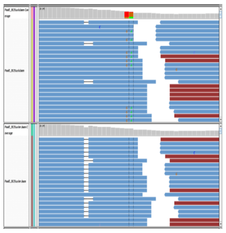
Indel Re-alignment

- Why is it necessary to re-align reads?
- Outcome is refinement of insertion-deletion positioning and reduction in false positive SNPs
- Example

DINDEL

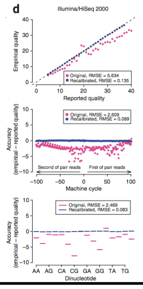
GATK
IndelRealignerTargetCreator /
IndelRealigner

Before:



After:

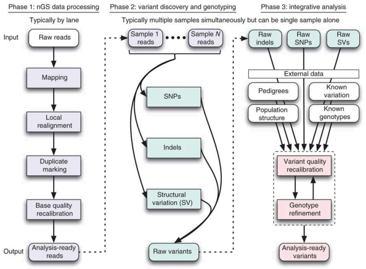
Base Quality Score Recalibration



Resequencing workflow

- QC of sequencing data
- Align reads - generates SAM/BAM alignment
- Coordinate sort reads
- Mark duplicate reads
- Re-alignment around insertions/deletions
- SNP-calling
- Filtering / Quality control
- Assess predicted effects of SNPs

GATK framework



Phase 1: nGS data processing

Typically by lane

Input: Raw reads → Mapping → Local realignment → Duplicate marking → Base quality recalibration → Output: Analysis-ready reads

Phase 2: variant discovery and genotyping

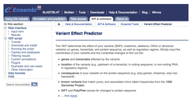
Typically multiple samples simultaneously but can be single sample alone

Sample 1 reads •••• Sample *N* reads → SNPs, Indels, Structural variation (SV) → Raw variants

Phase 3: integrative analysis

Raw indels, Raw SNPs, Raw SVs; External data: Pedigrees, Known variation, Population structure, Known genotypes → Variant quality recalibration ⇄ Genotype refinement → Analysis-ready variants

Annotating Variants



Variant Effect Predictor

The VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. Simply input the coordinates of your variants and the nucleotide changes to find out the:

- **genes** and **transcripts** affected by the variants
- **location** of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- **consequence** of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- **known variants** that match yours, and associated minor allele frequencies from the **1000 Genomes Project**
- **SIFT** and **PolyPhen** scores for changes to protein sequence
- ... And more!

▶ Launch VE!P

Fly (Austin). 2012 Apr 1; 6(2): 80–92.
Published online 2012 Apr 1. doi: 10.4161/fly.19695

PMCID: PMC3679285

A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff

SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3

Pablo Cingolani,[1, 2, 3] Adrian Platts,[4] Le Lily Wang,[1] Melissa Coon,[2] Tung Nguyen,[5] Luan Wang,[1, 2] Susan J. Land,[2] Xiangyi Lu,[1] and Douglas M. Ruden[1, 2, *]

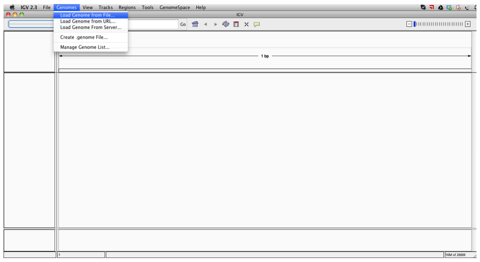
Author information ▶ Article notes ▶ Copyright and License information ▶

Do the analysis

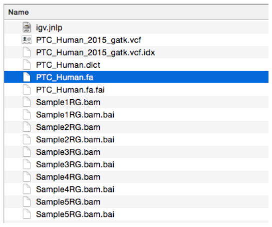
IGV: Integrative Genomics Viewer

- http://www.broadinstitute.org/igv/
- Standalone java program
  - Does not require a mysql database server or an apache web server
  - Limited to the resources of the machine that it is running on.
  - More interactive compared to Gbrowse.
  - Both IGV and Gbrowse can use GFF file format.

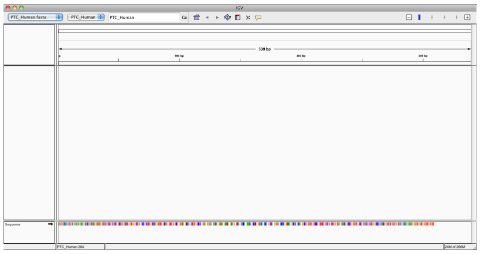
Load the fasta file as the genome (reference sequence)



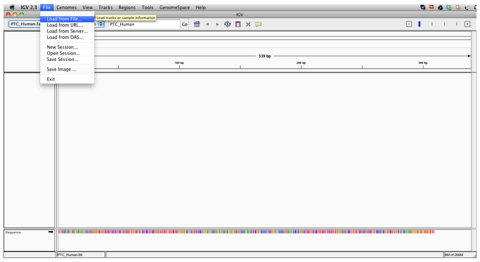
Select the PTC_Human.fa file


Name
igv.jnlp
PTC_Human_2015_gatk.vcf
PTC_Human_2015_gatk.vcf.idx
PTC_Human.dict
PTC_Human.fa
PTC_Human.fa.fai
Sample1RG.bam
Sample1RG.bam.bai
Sample2RG.bam
Sample2RG.bam.bai
Sample3RG.bam
Sample3RG.bam.bai
Sample4RG.bam
Sample4RG.bam.bai
Sample5RG.bam
Sample5RG.bam.bai

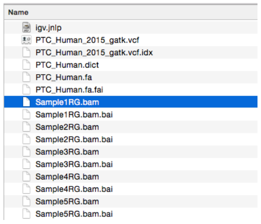
Notice the sequence at the bottom



Now load the Alignment files (.bam)

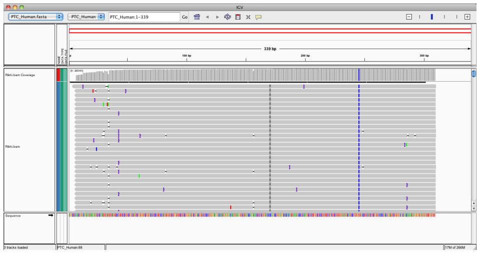


Now load the Alignment files (.bam)



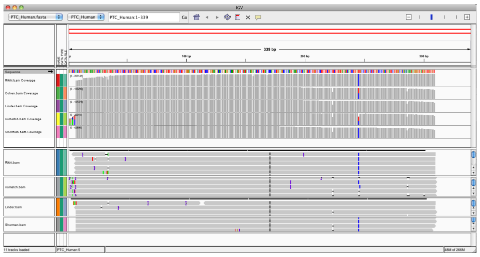
Name
igv.jnlp
PTC_Human_2015_gatk.vcf
PTC_Human_2015_gatk.vcf.idx
PTC_Human.dict
PTC_Human.fa
PTC_Human.fa.fai
Sample1RG.bam
Sample1RG.bam.bai
Sample2RG.bam
Sample2RG.bam.bai
Sample3RG.bam
Sample3RG.bam.bai
Sample4RG.bam
Sample4RG.bam.bai
Sample5RG.bam
Sample5RG.bam.bai

Coverage is the consensus of all sequences together



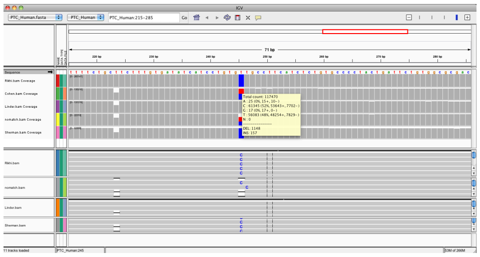
Gray means sequence is the same as the genome, color shows a change from reference.

Load all the bam files



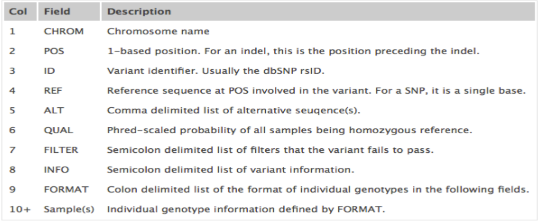
You can move the tracks by clicking on the name and dragging them.

Zoom into region of interest



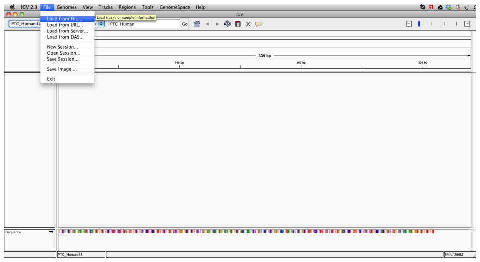
Put your mouse over each base to get more statistics about each base.

VCF Format

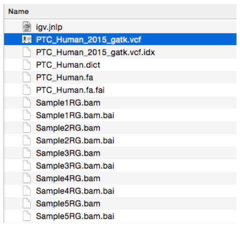


Col	Field	Description
1	CHROM	Chromosome name
2	POS	1-based position. For an indel, this is the position preceding the indel.
3	ID	Variant identifier. Usually the dbSNP rsID.
4	REF	Reference sequence at POS involved in the variant. For a SNP, it is a single base.
5	ALT	Comma delimited list of alternative seuqence(s).
6	QUAL	Phred-scaled probability of all samples being homozygous reference.
7	FILTER	Semicolon delimited list of filters that the variant fails to pass.
8	INFO	Semicolon delimited list of variant information.
9	FORMAT	Colon delimited list of the format of individual genotypes in the following fields.
10+	Sample(s)	Individual genotype information defined by FORMAT.

Load the VCF file



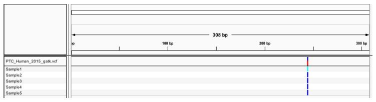
Load the VCF file

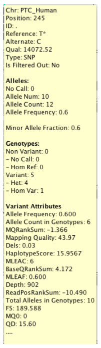


Name

igv.jnlp
PTC_Human_2015_gatk.vcf
PTC_Human_2015_gatk.vcf.idx
PTC_Human.dict
PTC_Human.fa
PTC_Human.fa.fai
Sample1RG.bam
Sample1RG.bam.bai
Sample2RG.bam
Sample2RG.bam.bai
Sample3RG.bam
Sample3RG.bam.bai
Sample4RG.bam
Sample4RG.bam.bai
Sample5RG.bam
Sample5RG.bam.bai

Details of SNP





Chr: PTC_Human
Position: 245
ID: .
Reference: T*
Alternate: C
Qual: 14072.52
Type: SNP
Is Filtered Out: No

Alleles:
No Call: 0
Allele Num: 10
Allele Count: 12
Allele Frequency: 0.6

Minor Allele Fraction: 0.6

Genotypes:
Non Variant: 0
– No Call: 0
– Hom Ref: 0
Variant: 5
– Het: 4
– Hom Var: 1

Variant Attributes
Allele Frequency: 0.600
Allele Count in Genotypes: 6
MQRankSum: -1.366
Mapping Quality: 43.97
Dels: 0.03
HaplotypeScore: 15.9567
MLEAC: 6
BaseQRankSum: 4.172
MLEAF: 0.600
Depth: 902
ReadPosRankSum: -10.490
Total Alleles in Genotypes: 10
FS: 189.588
MQ0: 0
QD: 15.60
....

Place your mouse over the SNP to get the details.

Are the conclusions the same ?
What additional filtering should we apply?