Finding and Calling Genome Variants

Whitehead Institute

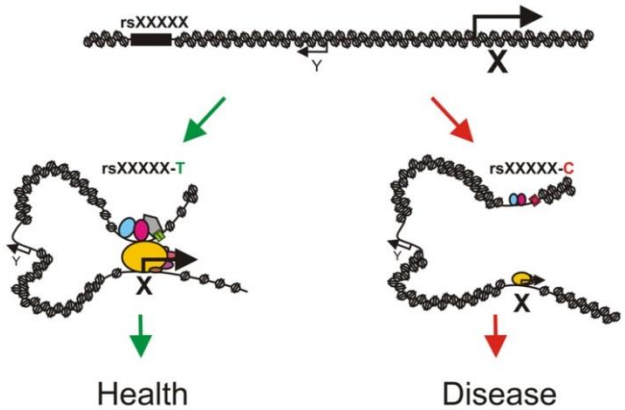
BaRC Hot Topics

Outline

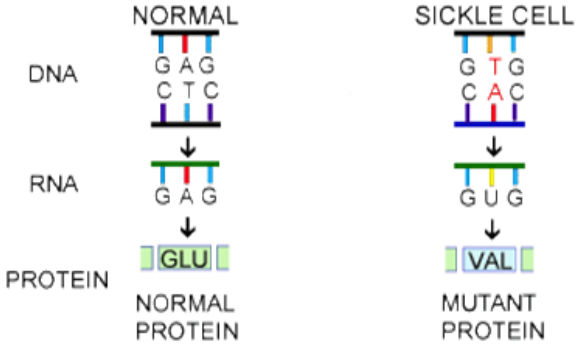
- Genome variants overview
- Mining variants from databases
  - ❖dbSNP
  - ❖HapMap
  - ❖1000 Genomes
  - ❖Disease/Clinical variants databases
- Calling variants using your own data
  - ❖Samtools (mpileup/bcftools)

Genomic Variation

- Population genetics
  - Measure/explain diversity
- Disease susceptibility
  - GWAS
  - Biomarkers
- Variants may cause a particular trait
  - Regulatory element (eg. promoter, enhancer, etc.)
  - Protein coding sequence (eg. silent, missense, or nonsense mutation)



Palstra, RJ. et al (2012)



http://evolution.berkeley.edu/evolibrary/article/mutations_06

Genomic Variation: Sequence vs Structural Variation

- Sequence Variants

Type	Description	Example (Reference / Alternative)	
SNP	Single Nucleotide Polymorphism	Ref: ...TTG**A**CGTA...	Alt: ...TTG**G**CGTA...
Insertion	Insertion of one or several nucleotides	Ref: ...TTGACGTA...	Alt: ...TTGA**TG**CGTA...
Deletion	Deletion of one or several nucleotides	Ref: ...TTG**AC**GTA...	Alt: ...TTGGTA...
Substitution	A sequence alteration where the length of the change in the variant is the same as that of the reference.	Ref: ...TTG**AC**GTA...	Alt: ...TTG**TA**GTA...

- Structural Variants (>50 bases or more)

Type	Description	Example (Reference / Alternative)	
CNV	Copy Number Variation: increases or decreases the copy number of a given region		"Gain" of one copy:  "Loss" of one copy: 
Inversion	A continuous nucleotide sequence is inverted in the *same* position		
Translocation	A region of nucleotide sequence that has translocated to a *new* position (eg. BCR-ABL fusion gene)		

http://www.ensembl.org/info/genome/variation

Genome Variation: Individual and Population

- SNP vs SNV (Single Nucleotide Variant)
- SNP: present in more than 1% of the population
  - MAF* > 1% common SNP
  - MAF* < 1% rare SNP
  - Some definitions use 5% as threshold
- On average one variant every 1200 bases (based on HapMap)

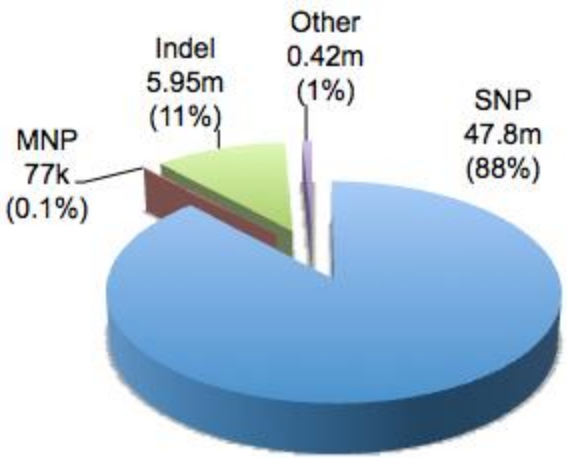
*Minor Allele Frequency

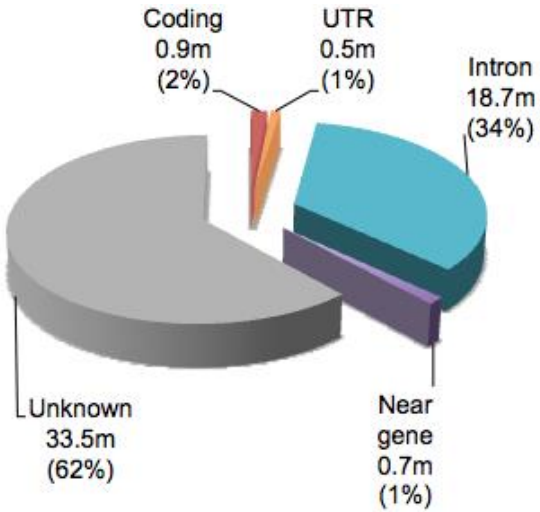
Genome Variation: Reference

Organism	Description/Strain	Assembly*
Human	DNA isolated from WBC of 4 anonymous individuals (2 males and 2 females). However, the majority of the sequence came from one of the male donors	GRCh37/GRCh38
Mouse	C57BL/6J	GRCm37/GRCm38
C.*elegans*	N2	WormBase v WS220
Fruit fly	ISO1	BDGP Release 5
Yeast	S288C	SGD Feb 2011
A.*thaliana*	Col ecotype	TAIR10

*Available in /nfs/genomes

Genome Variation: Distribution of Variants*

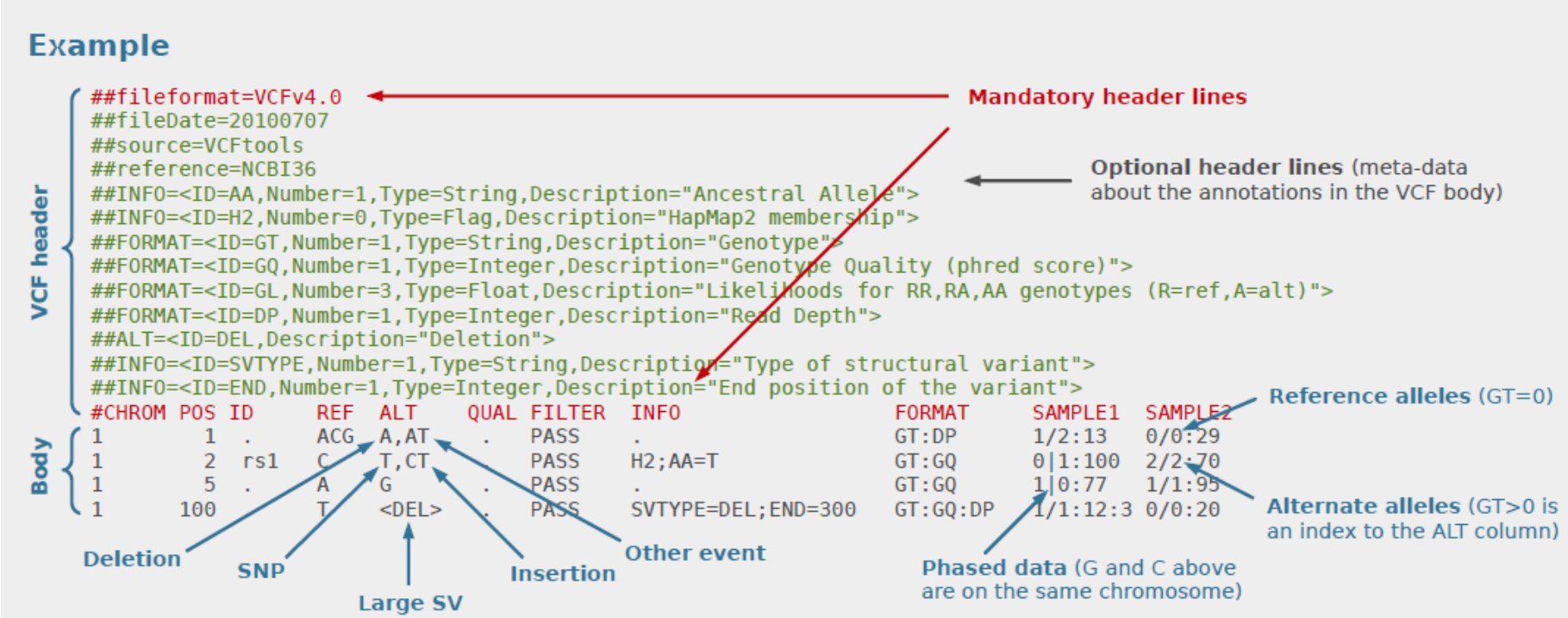




*based on dbSNP 135 (http://massgenomics.org)
MNP: Multiple Nucleotide Polymorphism

VCF Format

- Variant Call Format (VCF); BCF ➔ binary version of VCF
- Text file format with meta-information and header lines, followed by data lines containing information about a position in the genome.



www.1000genomes.org

Whitehead Institute

BaRC

Describing/Annotating Variants

- General guidelines*
  - no position 0
  - range indicated by "_" (eg. 586_591)
- DNA
  - g.957A>T (to include chromosome use chr9:g.957A>T)
  - g.413delG
  - g.451_452insT
  - In CDS,
    - c.23G>C
    - +1 is A of ATG (start codon); -1 is the previous/upstream nucleotide
    - "*" is the stop codon (eg. *1 is the first nucleotide of the stop codon)
- RNA
  - r.957a>u
- Protein (three/one letter aa)
  - p.His78Gln

Chr	Position	Ref	Alt	Source	g.change : rsID : Depth=AvgSampleReadDepth : FunctionGVS : hgvsProteinVariant
16	89824989	G	T	EVS	g.89824989G>T : rs140823801 : Depth=141 : missense : p.Q993K

*For complete list/guidelines see hgvs.org

Genome Variation Databases: dbSNP

- Repository for SNPs and short sequence variation (<50 bases)
- Current build: dbSNP 143 (Mar 2015)
  - Approx. 88M validated rs#'s for human
    - Mostly germline mutations (smaller subset of somatic)
    - Contains rare variants as well
  - Various organisms
- Each SNP, or record, is identified by an rs# that includes
  - Summary attributes
  - NCBI resources (linked to ClinVar, GenBank, etc.)
  - External resources (linked to OMIM and NHGRI GWAS)
- Submissions are made from public laboratories and private organizations (ss#'s), and identical records are clustered into a single record (rs#'s).
- rs id is same for different assemblies (eg. GRCh37/38), but chromosomal coordinates may differ!

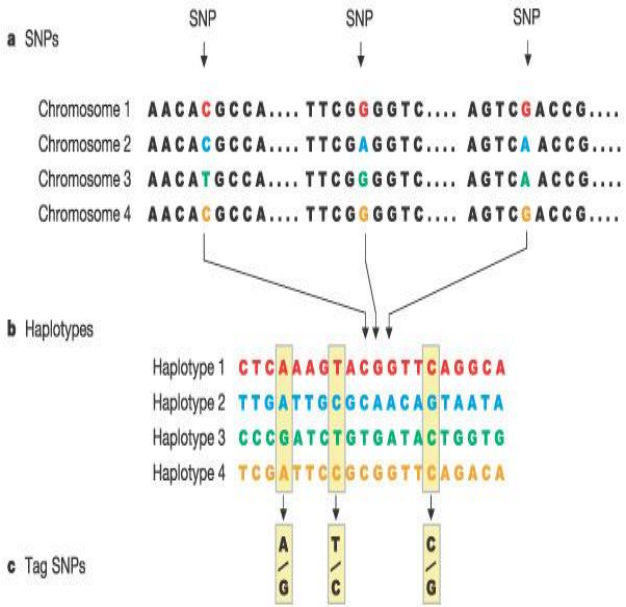
Hands-on: dbSNP

- Finding SNPs for your favorite gene in dbSNP and UCSC Genome Browser

Genome Variation Databases: HapMap Project

- Created to find distribution of SNPs and other genetic variants (GRCh37 coordinates)
- Phase I: 269 samples (2005)
- Phase II: 270 samples (2007)
- Phase III: 1115 samples (2008)
- Samples were from various populations
  - CHB (Han Chinese)
  - MEX (Mexican ancestry in Los Angeles, CA)
  - JPT (Japanese)

  etc.



Genome Variation Databases: 1000 Genomes Project

- Extension of the HapMap in 2008 to catalogue genetic variation by sequencing at least 1000 participants
- Discover population level human genetic variations
- Initially consisted of whole genome low coverage (4X) and high coverage exome (20X) sequencing
- VCF file was developed, and initially maintained, for the project

Mining Disease/Clinical Variants

Database	Link
Catalog of Published GWAS (NHGRI)	genome.gov/26525384
GWAS Central	gwascentral.org
ClinVar (NCBI)	ncbi.nlm.nih.gov/clinvar
PheGenI (NCBI)	ncbi.nlm.nih.gov/gap/phegeni
SNPedia	snpedia.com

Mining Disease/Clinical Variants in Cancer: cBioPortal

- Visualization and analysis of cancer genomics data sets
- Currently 89 data sets (including TCGA) consists of various studies
- Tools/Features:
  - OncoPrint: graphical summary of genomic alterations
  - Mutual Exclusivity
  - Correlation Plots
  - MutationMapper ("lollipop plots")

Hands-on: cBioPortal

- Mining cancer variants in cBioPortal website

Mining Disease/Clinical Variants in Cancer: COSMIC

- Catalog of Somatic Mutations in Cancer (COSMIC) created in 2005
- v70 (Aug 2014) had ~2M coding point mutations
- Datasets are curated from published literature and other databases (eg. TCGA, ICGC)
- Available in both GRCh37/38 coordinates
- Tools/Features
  - Cancer Gene Census (currently 572 genes)
  - Browser: Cancer/Cell Line
  - COSMIC Mart (similar to BioMart)

Calling Variants

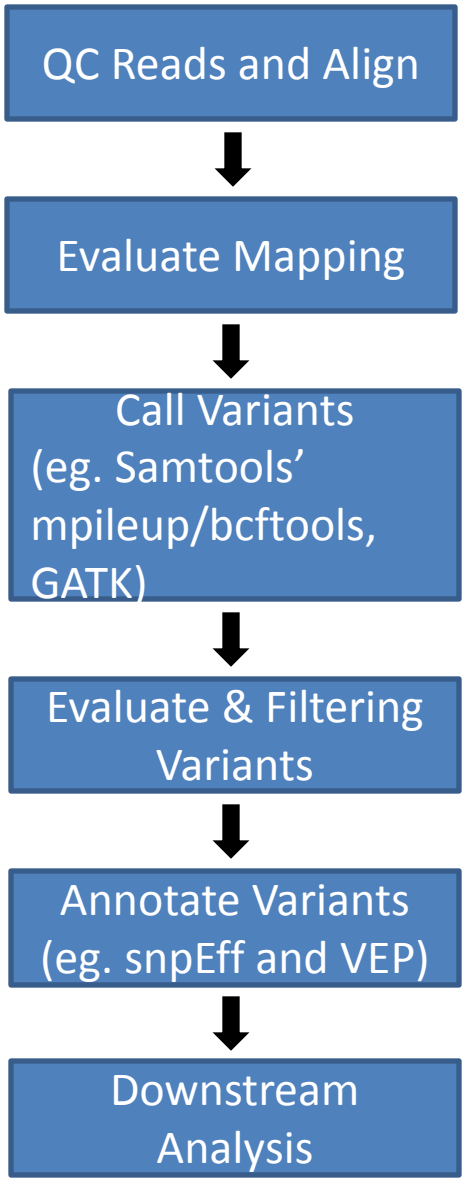
- Align reads to reference* and call variants
- Popular tools include Samtools and GATK (from Broad)
- Germline vs Somatic mutations
- Samtools:
  - Samtools's mpileup (formerly pileup) computes genotype likelihoods supported by the aligned reads (BAM file) and stores in binary call format (BCF) file.
  - Bcftools applies the priors (from above) and calls variants (SNPs and indels). Bcftools can be used to filter VCF files.

*reference-free variant calling software are available (eg. CORTEX)

Calling Variants: Samtools (cont.)

- Removes duplicate reads (eg. from PCR)
- Both unique and multi-mapped reads are used for calling variants
- Recalibrates quality scores to take into account sequencing errors

Calling Variants: Workflow



- Use a sensitive (gapped) aligner eg. BWA

Calling Variants

Calling Variants: Questionable Calls

Calling Variants: Evaluating

- Percent, or number of, reads containing variant vs reference
  - ❖View in a browser (eg. IGV)
- Base quality (eg. at least Q30)
- Mapping quality (depends on aligner)
- Coverage across region(s) of interest
  - ➢ depth of coverage (eg. at least 5X)
  - ➢ 20X considered high-coverage
- Strand bias

Calling Variants: Filtering/Querying

- vcf-annotate (from VCFtools*)
  - `vcf-annotate -f + myFile.vcf > myFile_annot.vcf`
  - "+" applies several filters with default values, eg.
    - Strand bias: test if variant bases tend to come from one strand (Fisher's Test)
    - End distance bias: test if variant bases tend to occur at a fixed distance from the end of reads (t-test)
- Remove common SNPs (eg. use dbSNP)
  - bedtools
- bcftools
  - `bcftools view -i 'DP>100' myFile.vcf`
  - `bcftools view -i 'GT[0]=="1/1"' myFile.vcf`

*vcftools.sourceforge.net

Hands-on: Samtools' mpileup/bcftools

- Calling variants from an alignment (bam) file
  - ❖ `samtools mpileup [options] [file1.bam file2.bam …]`

```
samtools mpileup -d100000 -Buf
/nfs/genomes/human_gp_feb_09_no_random/fasta/chr1.fa -o
chr1_Subset_1000Genomes_CEU.bcf chr1_Subset_1000Genomes_CEU.bam
```

  - ❖ `bcftools call [options] [file | STDIN]`

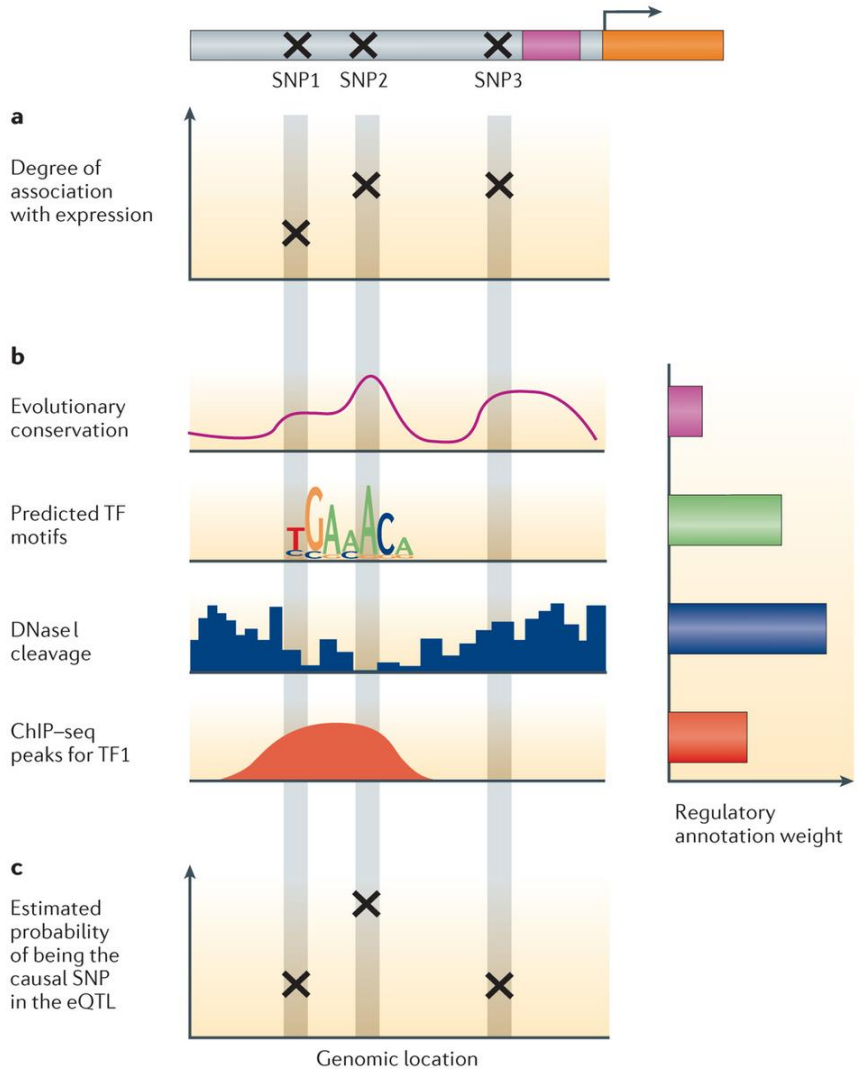
```
bcftools call -vmO v -o chr1_Subset_1000Genomes_CEU.vcf
chr1_Subset_1000Genomes_CEU.bcf
```

For list of options see: htslib.org/doc/samtools-1.1.html and http://samtools.github.io/bcftools/bcftools.html

Calling Variants: Annotation

- Annotate variants with (functional) consequence
  eg. chr12:g25232372A>G is a missense variant
- Popular tools include snpEff, and Variant Effect Predictor (VEP) from Ensembl
- Choice of annotation may affect variant annotation
  - RefSeq
  - Ensembl
  - GENCODE

Further/Downstream Analysis



Nature Reviews | Genetics
Levo, M. and Segal, E. (2014)

BaRC SOP

- Variant calling using Samtools and GATK. Manipulating/interpreting VCF files

http://barcwiki/wiki/SOPs under

Variant calling and analysis

Resources For Mining Variants

Database	Link
dbSNP	www.ncbi.nlm.nih.gov/SNP
HapMap	hapmap.ncbi.nlm.nih.gov
1000 Genomes	1000genomes.org
UK10K	uk10k.org
Exome Variant Server (EVS)	evs.gs.washington.edu/EVS
Personal Genome Project (Harvard)	personalgenomes.org
ExAC Browser (Broad)	exac.broadinstitute.org

Resources For Mining Variants: Cancer

Database	Link
International Cancer Genome Consortium (ICGC)	icgc.org
Catalogue of Somatic Mutation in Cancer (COSMIC)	cancer.sanger.ac.uk
cBioPortal for Cancer Genomics	cbioportal.org
Cancer Cell Line Encyclopedia (CCLE)	broadinstitute.org/ccle

Resources For Mining Variants: Plants

- 1001 Genomes (A.*thaliana* 1001 strains)
  - 1001genomes.org
- 1000 Genomes (large-scale gene sequencing of at least 1000 plant species)
  - www.onekp.com