National Institute of Allergy and Infectious Diseases

Variants calling and Exome-seq

Mariam Quiñones, PhD

Computational Molecular Biology Specialist

Bioinformatics and Computational Biosciences Branch
OCICB/OSMO/OD/NIAID/NIH

February 4th, 2015

NIAID

BCBB: A Branch Devoted to Bioinformatics and Computational Biosciences

- Researchers' time is increasingly important
- BCBB saves our collaborators time and effort
- Researchers speed projects to completion using BCBB consultation and development services
- No need to hire extra post docs or use external consultants or developers

What makes us different?

Computational Biologists

Software Developers

PMs and Analysts

NIH National Institute of Allergy and Infectious Diseases

NIAID

BCBB

- "NIH Users: Access a menu of BCBB services on the NIAID Intranet:
  - http://bioinformatics.niaid.nih.gov/
- Outside of NIH –
  - search "BCBB" on the NIAID Public Internet Page: www.niaid.nih.gov
- Email us at:
  - ScienceApps@niaid.nih.gov

National Institute of Allergy and Infectious Diseases

Remember Sanger?

- Sanger introduced the “dideoxy method” (also known as Sanger sequencing) in December 1977

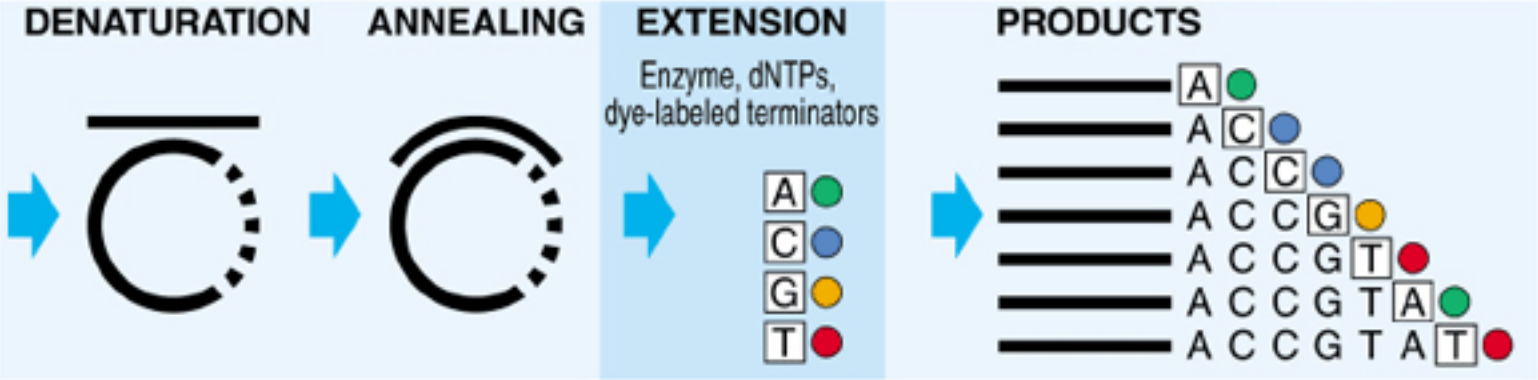
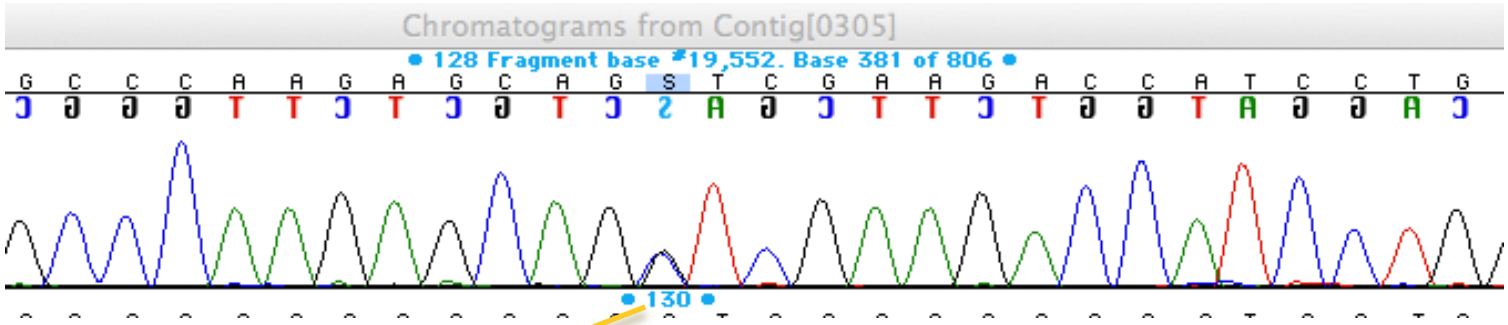


IMAGE: http://www.lifetechnologies.com

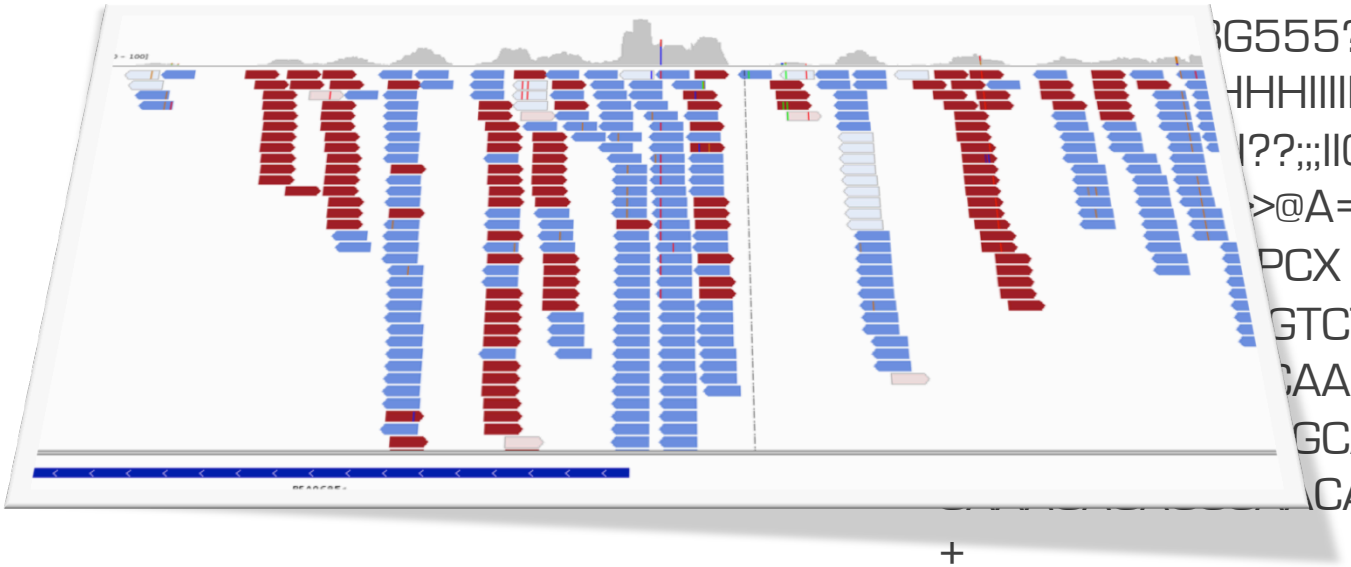


National Institute of Allergy and Infectious Diseases

Alignment of reads using tools such as ‘Sequencher’

NIAID

Understanding file formats



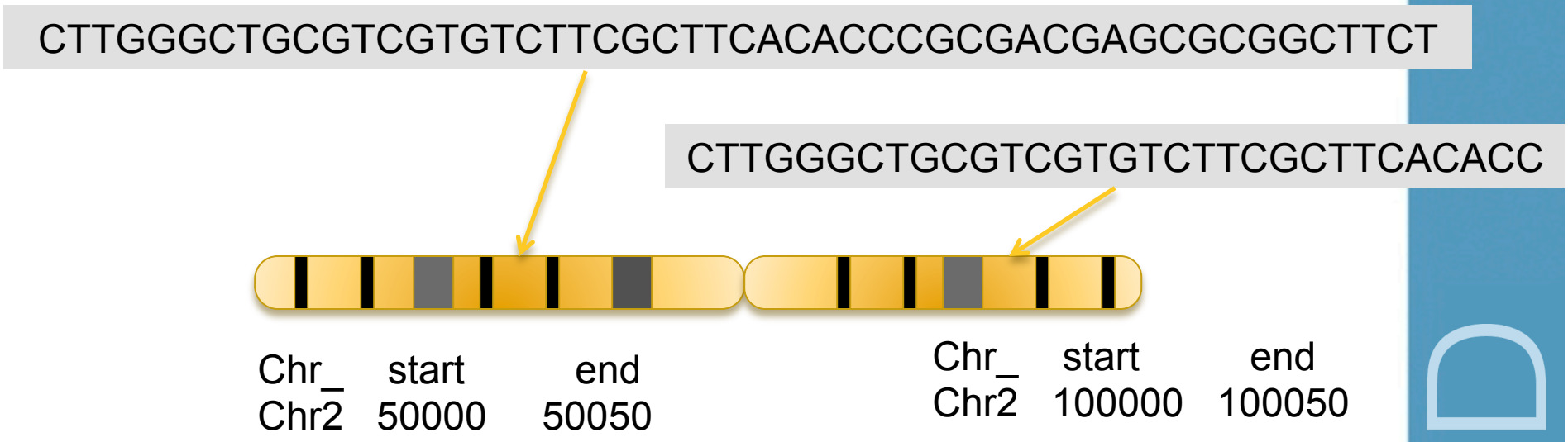
@F29EPBU01CZU4O
GCTCCGTCGTAAAAGGGG
+
24469:666811//..,,
@F29EPBU01D6OZF
CTCGTTCTTGATTAATGAAACATTCTTGGCAAA
TGCTTTCGCTCTGGTCCGTCTTGCGCCGGTCCA
AGAATTTCACCTCTAGCGGCGCAATACGAATG
CCCAAACACACCCAACACACCA
+
3G555?
HHHIIIIIIHHHIIHHHIIIIIIIIH99;;CB
I??;;;IIGDBCEA?
>@A=BEIEEE
PCX
GTCTTGGAAGCTTGACTACCC
CAAATGGACCTTGAGAGCTTG
GCAGGGGAGCGCATCTCCC
CACACCA
+
IIIIIIIIIIIIIIIIIIIIHHHHIIIIHHHIIIIIIIIIIIIHHHIIIIIIIIIIIIIIIH
HHIIIIIIIIEIIB94422=4GEEEEEIBBBBHHHFIH??
?CII=?AEEEE
@F29EPBU01DER7Q
TGACGTGCAAATCGGTCGTCCGACCTCGGTAT

NIH National Institute of Allergy and Infectious Diseases

NIAID

File formats for aligned reads

- SAM (sequence alignment map)



National Institute of Allergy and Infectious Diseases

NIAID

Most commonly used alignment file formats

- SAM (sequence alignment map)

Unified format for storing alignments to a reference genome

- BAM (binary version of SAM) – used commonly to deliver data

Compressed SAM file, is normally indexed

- BED

*Commonly used to report features described by **chrom**, **start**, **end, name, score,** and **strand**.*

For example:

chr1 11873 14409 uc001aaa.3 0 +

Variant Analysis – Class Topics

National Institute of Allergy and Infectious Diseases

NIAID

Efforts at creating databases of variants:

- Project that started on 2002 with the goal of describing patterns of human genetic variation and create a haplotype map using SNPs present in at least 1% of the population, which were deposited in dbSNPs.

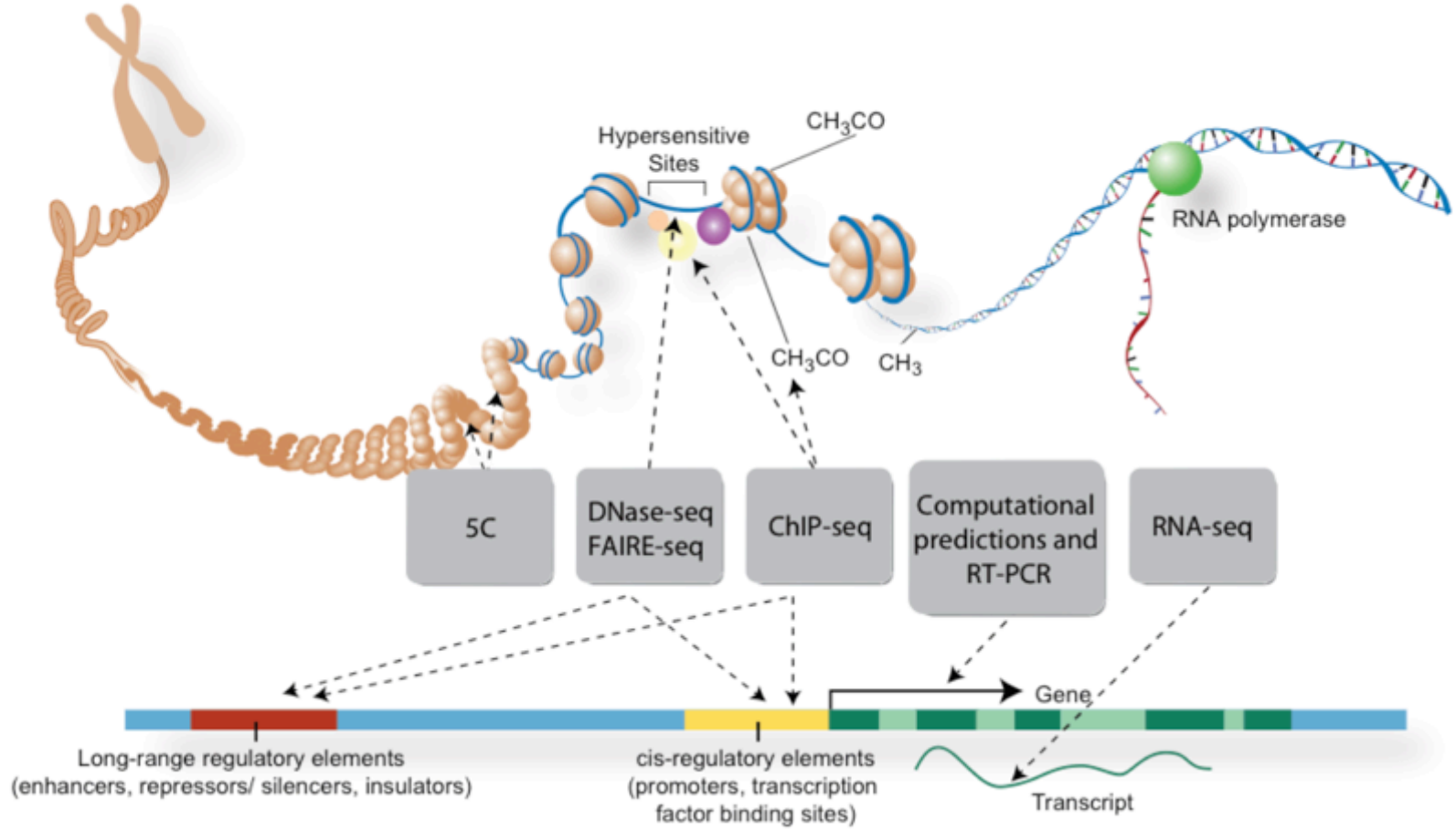
- Started in 2008 with a goal of using at least 1000 individuals (about 2,500 samples at 4X coverage), interrogate 1000 gene regions in 900 samples (exome analysis), find most genetic variants with allele frequencies above 1% and to a 0.1% if in coding regions as well as Indels and structural variants
- Make data available to the public ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/
  or via Amazon Cloud http://s3.amazonaws.com/1000genomes

NIAID

Variants are not only found in coding sequences

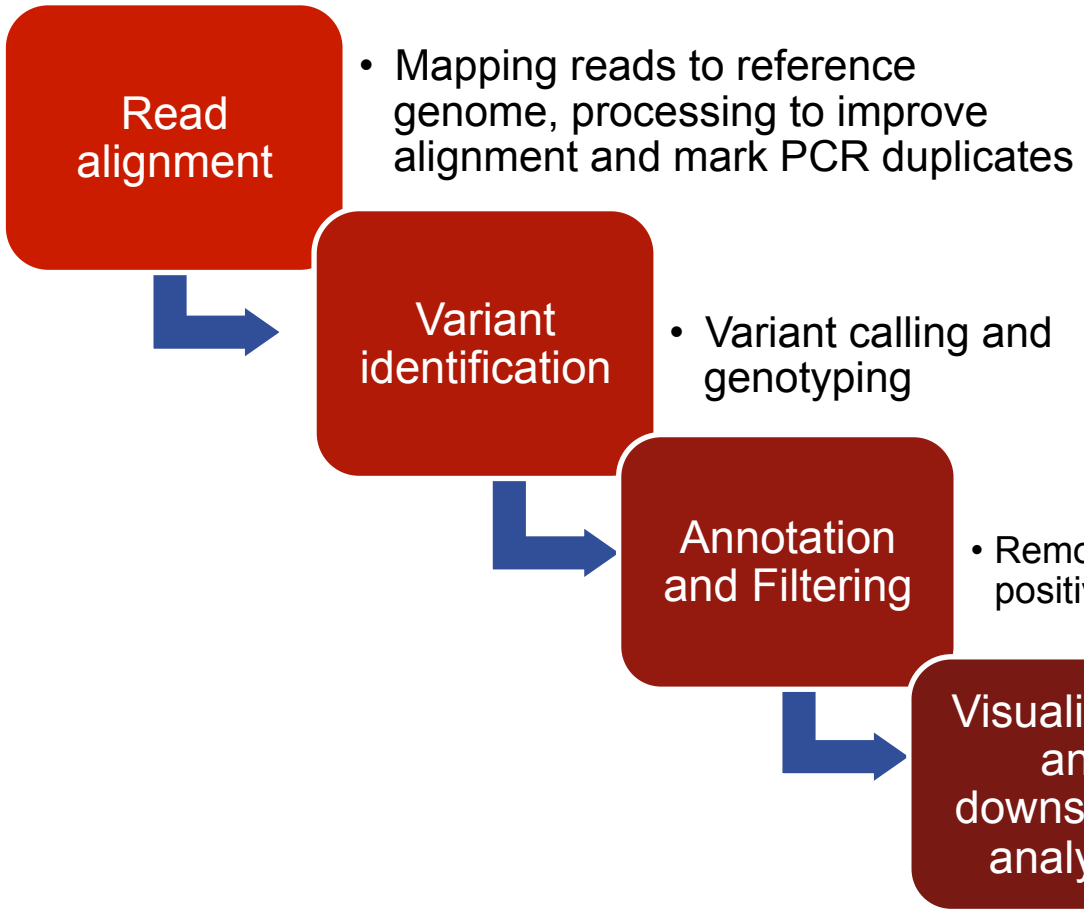
Main Goal:

- Find all functional elements in the genome



NIAID

A typical variant calling pipeline



National Institute of Allergy and Infectious Diseases

NIAID

Read alignment

Mapping reads to reference

Whole Genome

- Useful for identifying SNPs but also structural variants in any genomic region.

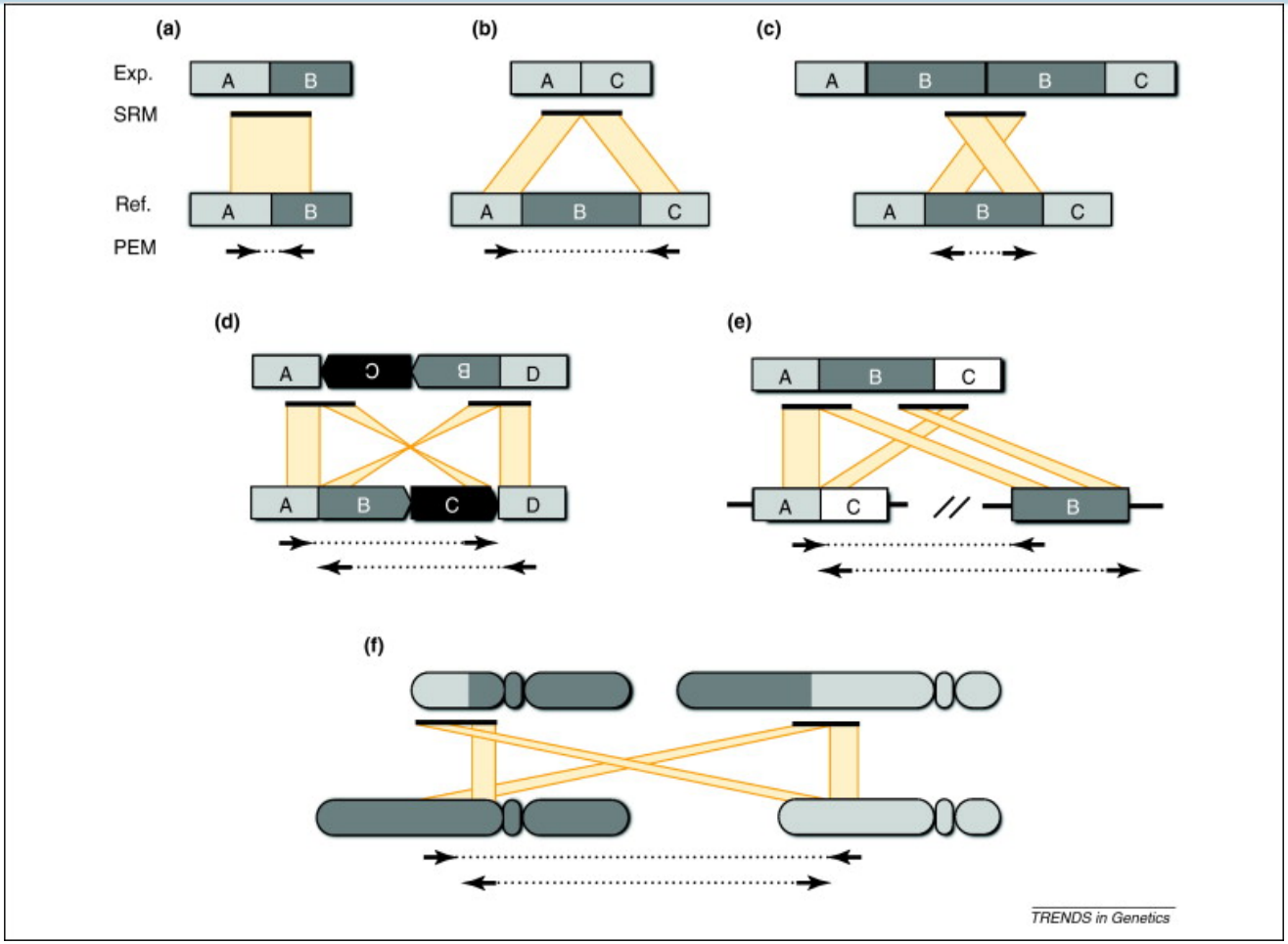
Exome-Seq

- Interrogates variants in specific genomic regions, usually coding sequences.
- Cheaper and provides high coverage

Recommended alignment tools: BWA-MEM, Bowtie2, ssaha2, novoalign.
See more tools:

- Seq Answers Forum: http://seqanswers.com/wiki/Software/list

How to identify structural variants?



http://www.sciencedirect.com/science/article/pii/S0168952511001685

National Institute of Allergy and Infectious Diseases

Methods and SV tools

1- Read Pair – use pair ends. GASVPro (Sindi 2012), BreakDancer (Chen 2009)

2- Read depth – CNVnator (Abyzov, 2010)

3- Split read – PinDEL (Ye 2009)

4- Denovo assembly

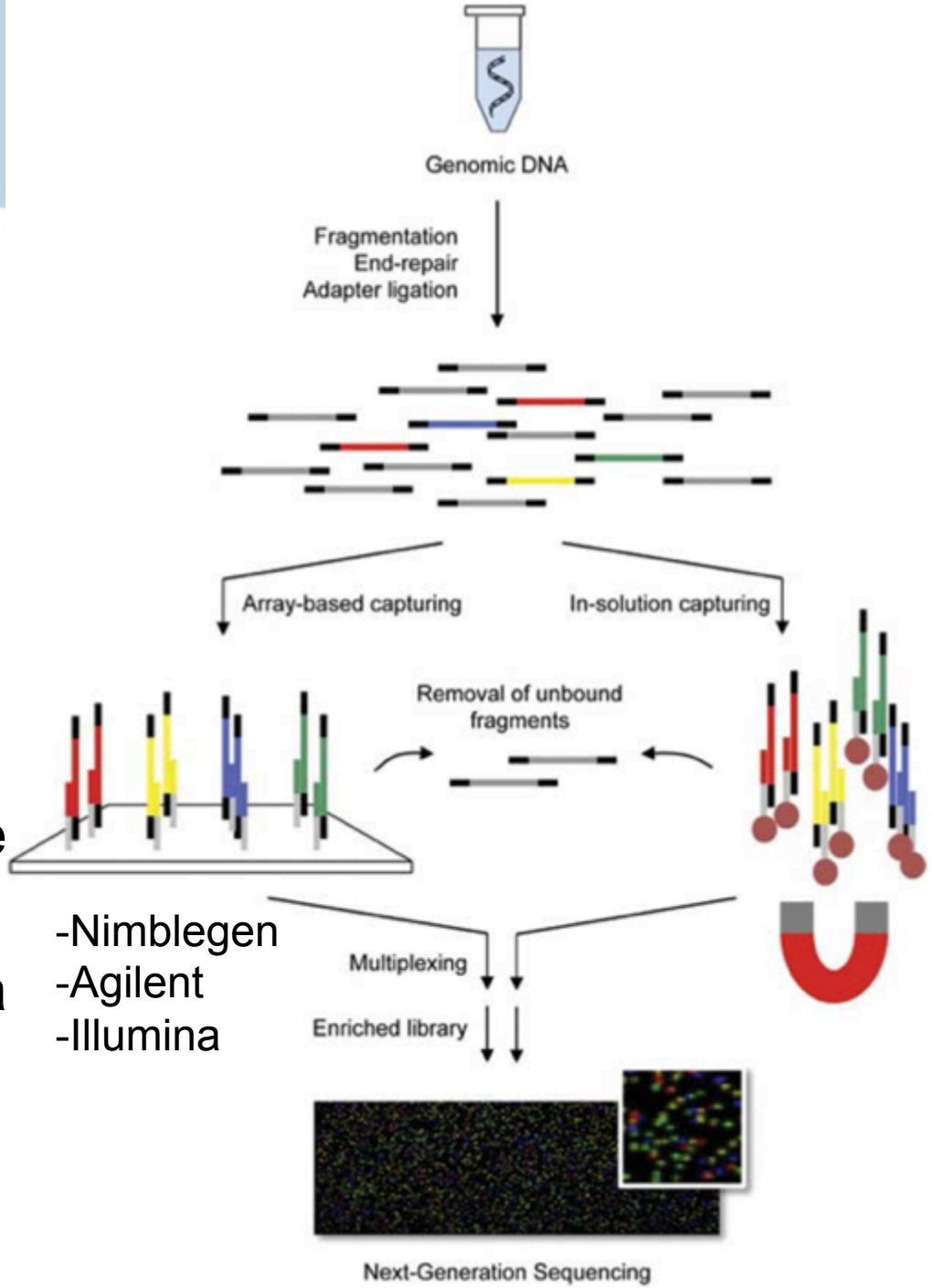
Learn more:
http://bit.ly/16A8FKR

Also see references:
http://bit.ly/1z7xnol - SV
http://bit.ly/qin294 - CNV
http://bit.ly/rLB7jc
http://bit.ly/17fdZTc

Exome-Seq

Targeted exome capture

- targets ~20,000 variants near coding sequences and a few rare missense or loss of function variants
- Provides high depth of coverage for more accurate variant calling
- It is starting to be used as a diagnostic tool



Ann Neurol. 2012 Jan;71(1):5-14.

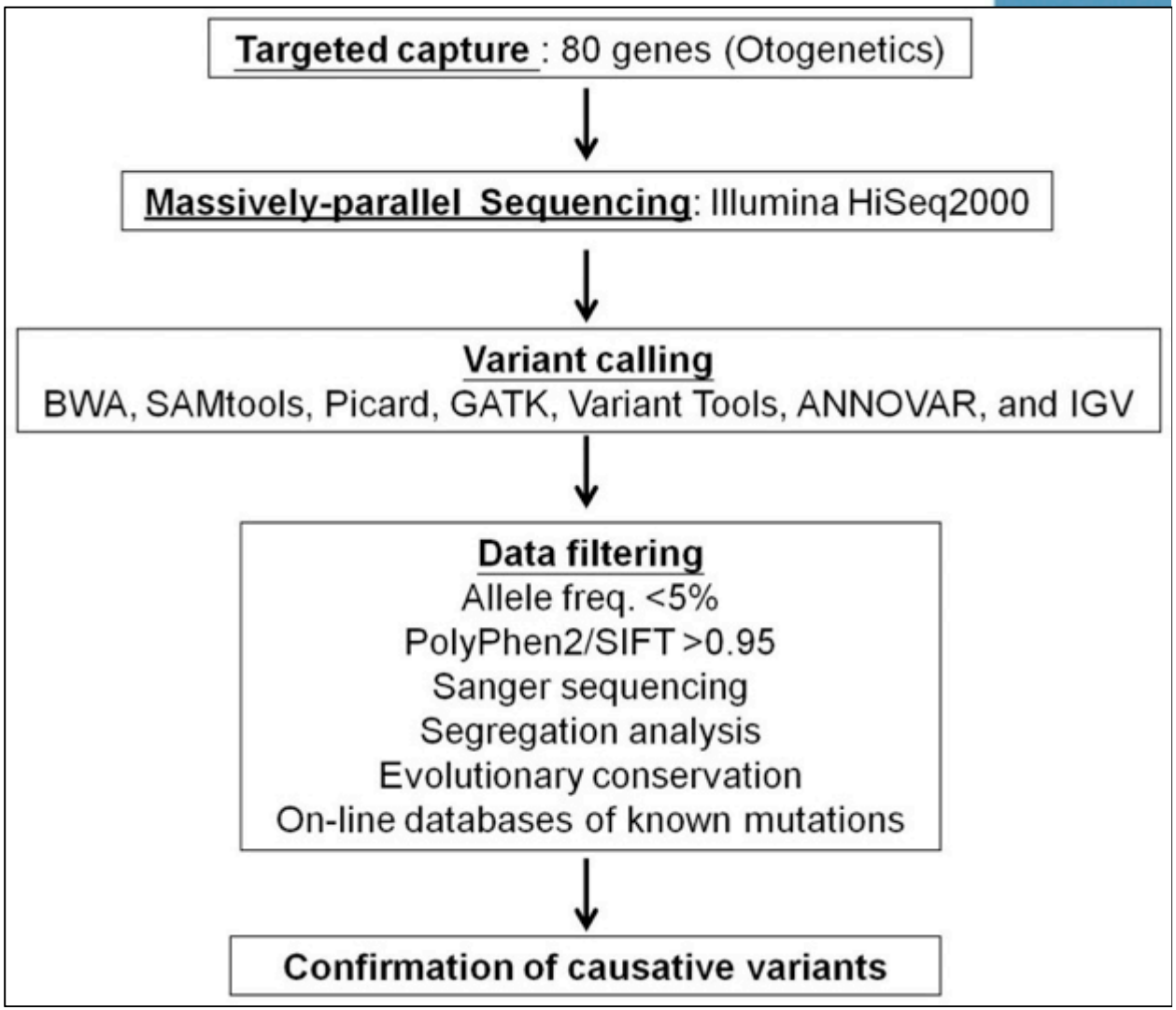
PLoS One. 2013;8(2):e57369. doi: 10.1371/journal.pone.0057369. Epub 2013 Feb 22.

Application of massively parallel sequencing to genetic diagnosis in multiplex families with idiopathic sensorineural hearing impairment.

Wu CC, Lin YH, Lu YC, Chen PJ, Yang WS, Hsu CJ, Chen PL.

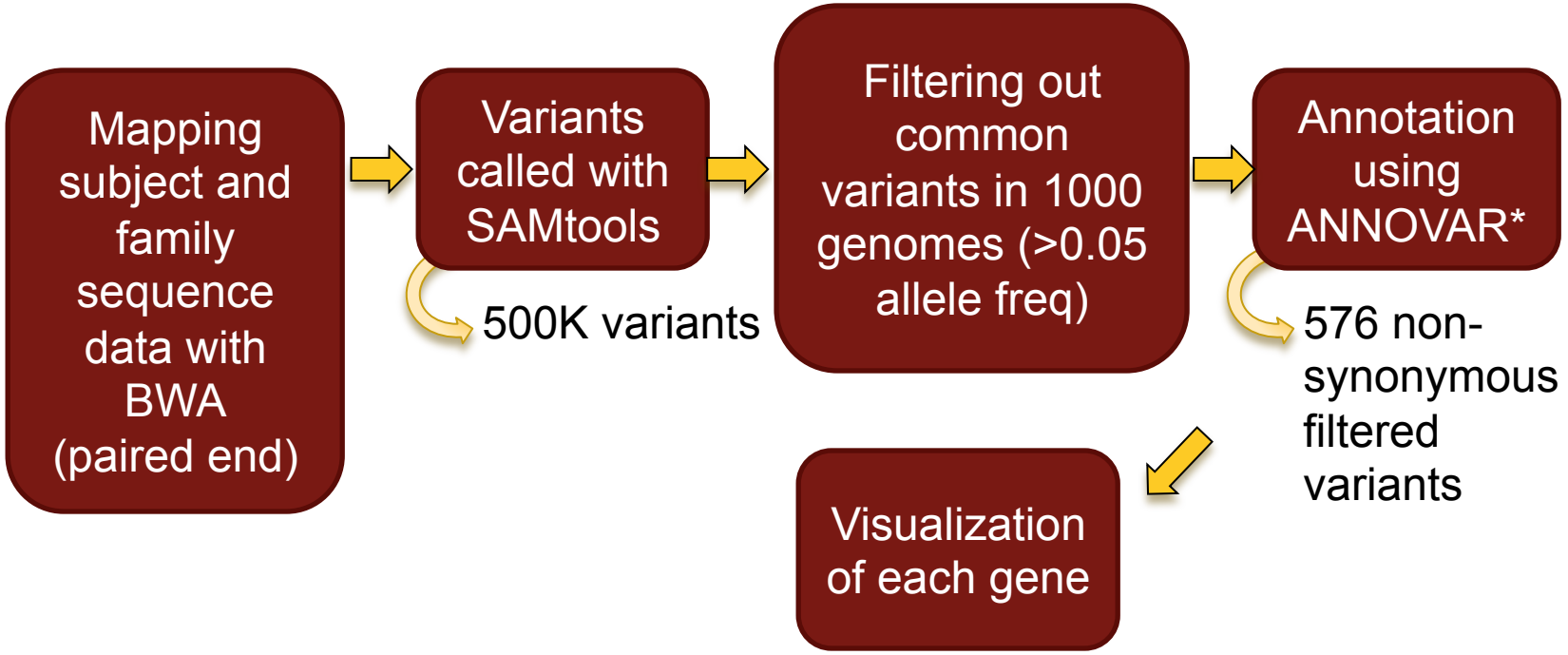
Department of Otolaryngology, National Taiwan University Hospital, Taipei, Taiwan ; Department of Medical Genetics, National Taiwan University Hospital, Taipei, Taiwan.

Analysis Workflow →



NIH National Institute of Allergy and Infectious Diseases

Whole-exome sequencing reveals a heterozygous LRP5 mutation in a 6-year-old boy with vertebral compression fractures and low trabecular bone density Fahiminiya S, Majewski J, Roughley P, Roschger P, Klaushofer K, Rauch F. Bone: July 23, 2013



*Annovar annotated: the type of mutations, presence in dbSNP132, minor allele frequency in the 1000 Genomes project, SIFT, Polyphen-2 and PHASTCONS scores

NIH National Institute of Allergy and Infectious Diseases

NIAID

Exome-seq

- Also used for finding somatic mutations

Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin- remodeling and ubiquitin ligase complex genes

VOLUME 44 | NUMBER 12 | DECEMBER 2012 Nature Genetics

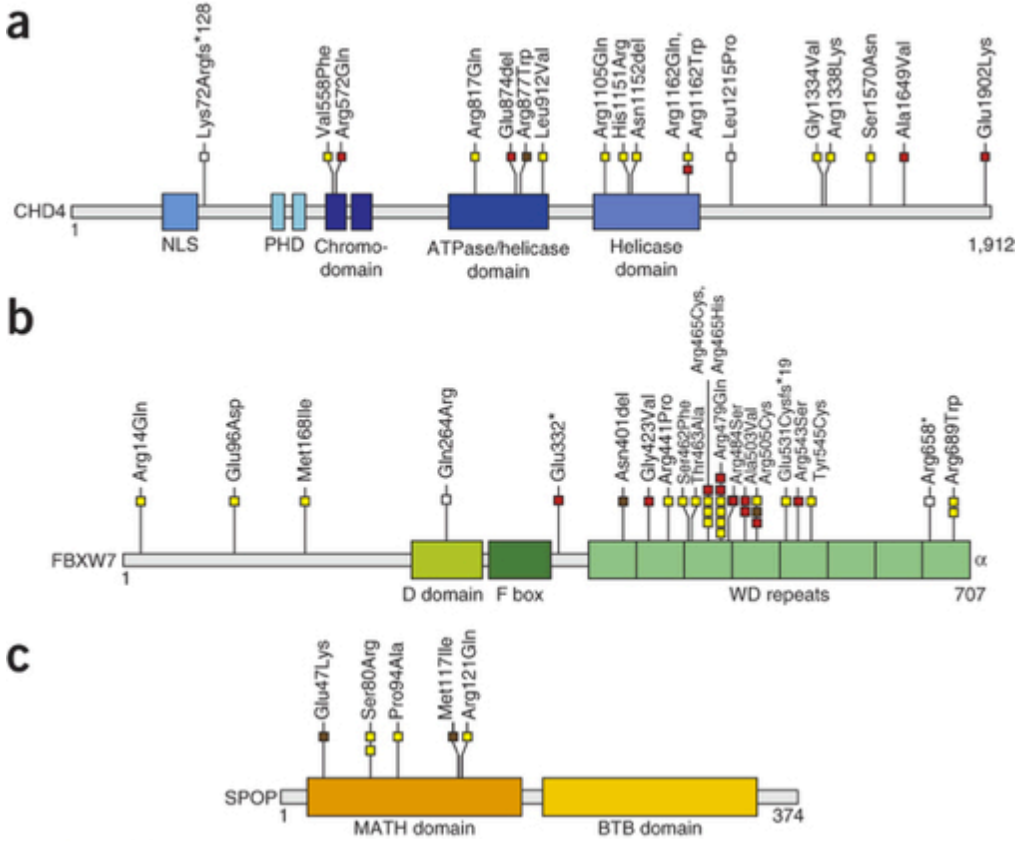


Figure 1: Somatic mutations in *CHD4*, *FBXW7* and *SPOP* cluster within important functional domains of the encoded proteins.

NIAID

Variant Analysis – Class Topics

National Institute of Allergy and Infectious Diseases

NIAID

Variant identification

Goals:

- Identify variant bases, genotype likelihood and allele frequency while avoiding instrument noise.
- Generate an output file in vcf format (variant call format) with genotypes assigned to each sample

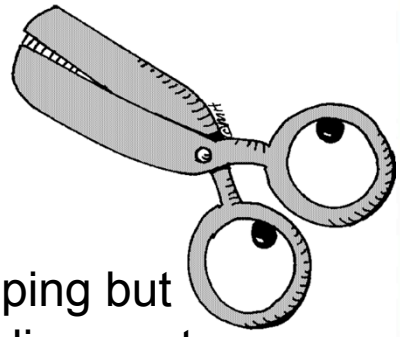
Challenges of finding variants using NGS data

- Base calling errors
  - Different types of errors that vary by technology, sequence cycle and sequence context
- Low coverage sequencing
  - Lack of sequence from two chromosomes of a diploid individual at a site
- Inaccurate mapping
  - Aligned reads should be reported with mapping quality score

Pre-Processing…

Processing reads for Variant Analysis

- Trimming:
  - Unless quality of read data is poor, low quality end trimming of short reads (e.g. Illumina) prior to mapping is usually not required.
  - If adapters are present in a high number of reads, these could be trimmed to improve mapping.
    - *Recommended tools: Prinseq, Btrim64, FastQC*



- Marking duplicates:
  - It is not required to remove duplicate reads prior to mapping but instead it is recommended to mark duplicates after the alignment.
    - *Recommended tool: Picard.*
  - A library that is composed mainly of PCR duplicates could produce inaccurate variant calling.

National Institute of Allergy and Infectious Diseases

http://www.biomedcentral.com/1471-2164/13/S8/S8

Clip: http://bestclipartblog.com

NIAID

Picard Tools (integrates with GATK)

AddOrReplaceReadGroups.jar
BamIndexStats.jar
BamToBfq.jar
BuildBamIndex.jar
CalculateHsMetrics.jar
CleanSam.jar
CollectAlignmentSummaryMetrics.jar
CollectCDnaMetrics.jar
CollectGcBiasMetrics.jar
CollectInsertSizeMetrics.jar
CollectMultipleMetrics.jar
CompareSAMs.jar
CreateSequenceDictionary.jar
EstimateLibraryComplexity.jar
ExtractIlluminaBarcodes.jar
ExtractSequences.jar
FastqToSam.jar
FixMateInformation.jar
IlluminaBasecallsToSam.jar
MarkDuplicates.jar
MeanQualityByCycle.jar
MergeBamAlignment.jar
MergeSamFiles.jar
NormalizeFasta.jar
picard-1.45.jar
QualityScoreDistribution.jar
ReorderSam.jar
ReplaceSamHeader.jar
RevertSam.jar
sam-1.45.jar
SamFormatConverter.jar
SamToFastq.jar
SortSam.jar
ValidateSamFile.jar
ViewSam.jar

```
java -jar QualityScoreDistribution.jar I=file.bam CHART=file.pdf
```

National Institute of Allergy and Infectious Diseases

http://picard.sourceforge.net/command-line-overview.shtml

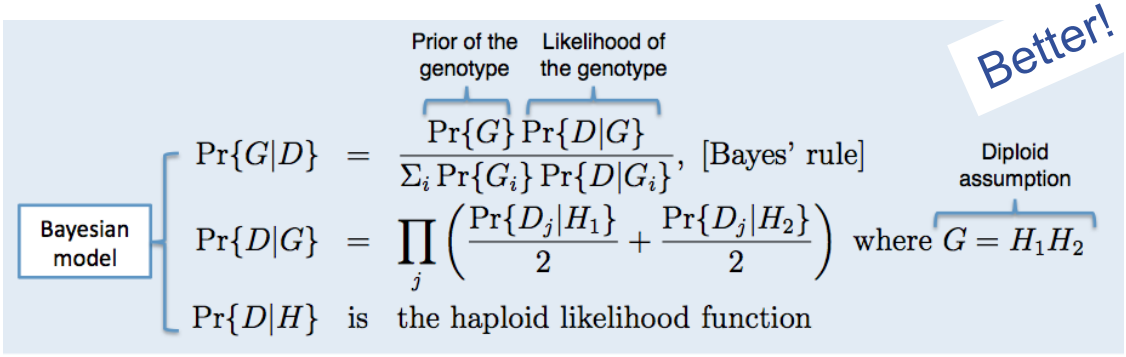
NIAID

SNP Calling Methods

Early SNP callers and some commercial packages use a simple method of counting reads for each allele that have passed a mapping quality threshold. **** This is not good enough in particular when coverage is low.**

It is best to use advanced SNP callers which add more statistics for more accurate variant calling of low coverage datasets and indels.

-e.g. GATK



National Institute of Allergy and Infectious Diseases

NIAID

SNP and Genotype calling

Germline Variant Callers

Table 1
Variant identification

Name	OS	BAM/SAM input	Other inputs	Output	Identifies
Germline callers					
CRISP	Lin	Yes	–	VCF	SNP, INDEL
GATK (UnifiedGenotyper)	Lin	Yes	–	VCF	SNP, INDEL
SAMtools	Lin	Yes	FASTA	VCF	SNP, INDEL
SNVer	Lin, Mac, Win	Yes	–	VCF	SNP, INDEL
VarScan 2	Lin, Mac, Win	No	pileup/mpileup	VCF, VarScan CSV	SNP, INDEL

Somatic Variant Callers

- Varscan, MuTEC

National Institute of Allergy and Infectious Diseases

http://bib.oxfordjournals.org/content/early/2013/01/21/bib.bbs086/T1.expansion.html

Samtools: https://github.com/samtools/bcftools/wiki/HOWTOs

NIAID

VCF format (version 4.0)

- Format used to report information about a position in the genome
- Use by 1000 genomes to report all variants

VCF FORMAT

The Variant Call Format (VCF) is a TAB-delimited format with each data line consists of the following fields:

Col	Field	Description
1	CHROM	CHROMosome name
2	POS	the left-most POSition of the variant
3	ID	unique variant IDentifier
4	REF	the REFerence allele
5	ALT	the ALTernate allele(s), separated by comma
6	QUAL	variant/reference QUALity
7	FILTER	FILTers applied
8	INFO	INFOrmation related to the variant, separated by semi-colon
9	FORMAT	FORMAT of the genotype fields, separated by colon (optional)
10+	SAMPLE	SAMPLE genotypes and per-sample information (optional)

VCF format

http://www.broadinstitute.org/gsa/wiki/index.php/Understanding_the_Unified_Genotyper's_VCF_files

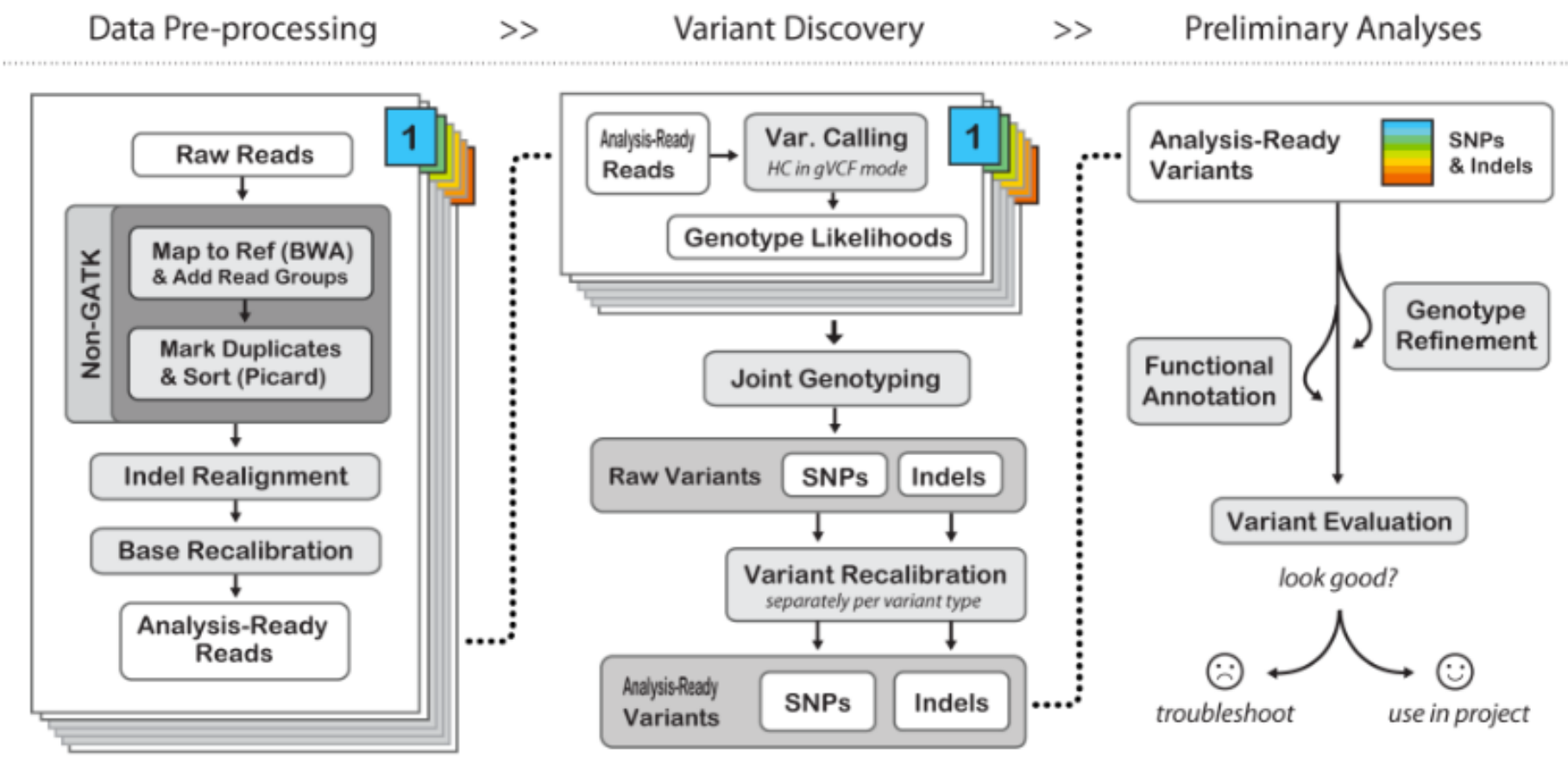
```
[HEADER LINES]
#CHROM  POS     ID              REF     ALT     QUAL    FILTER  INFO            FORMAT          NA12878
chr1    873762  .               T       G       5231.78 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
chr1    877664  rs3828047       A       G       3931.66 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
chr1    899282  rs28548431      C       T       71.77   PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:1,3:4:25.92:103,0,26
chr1    974165  rs9442391       T       C       29.84   LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:14,4:14:60.91:61,0,255
```

How variation is represented in a VCF

Each line represents one variant (here everything is a SNP, but some could be indels or CNVs) as well as the genotype of our sample, NA12878, at that variant. I've chosen these four variants because they each represent an important aspect in interpreting a VCF file:

- chr1:873762 is a novel T/G polymorphism, found with very high confidence (QUAL = 5231.78).
- chr1:877664 is a known A/G SNP (rs3828047), found with very high confidence (QUAL = 3931.66)
- chr1:899282 is a known C/T SNP (rs28548431), but has a relative low confidence (QUAL = 71.77)
- chr1:974165 is a known T/C SNP but we have so little evidence for this variant in our data that although we write out a record for it (book keeping, really) our statistical evidence is so low that we filter the record out as a bad site "LowQual".

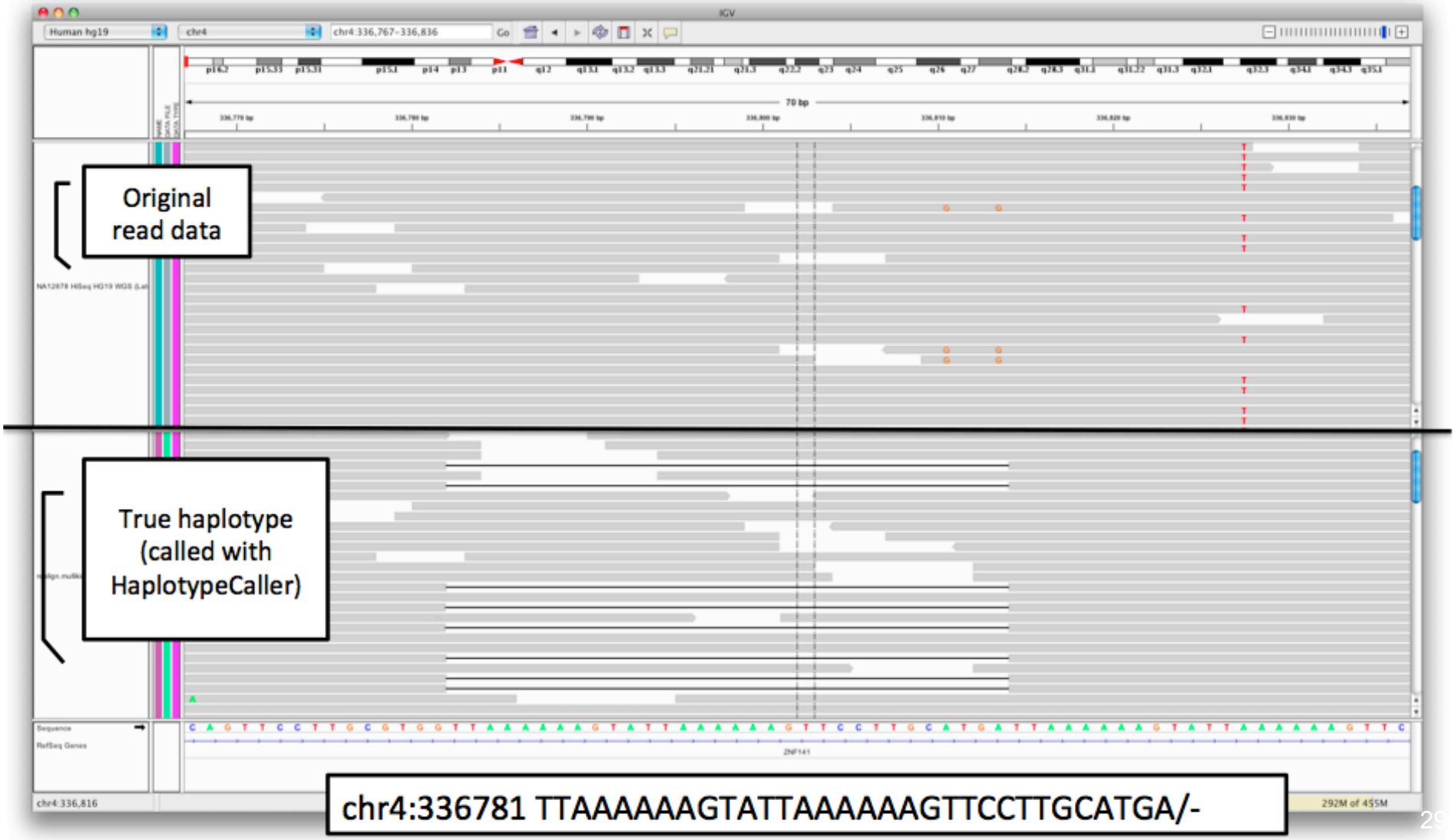
GATK pipeline (BROAD) for variant calling



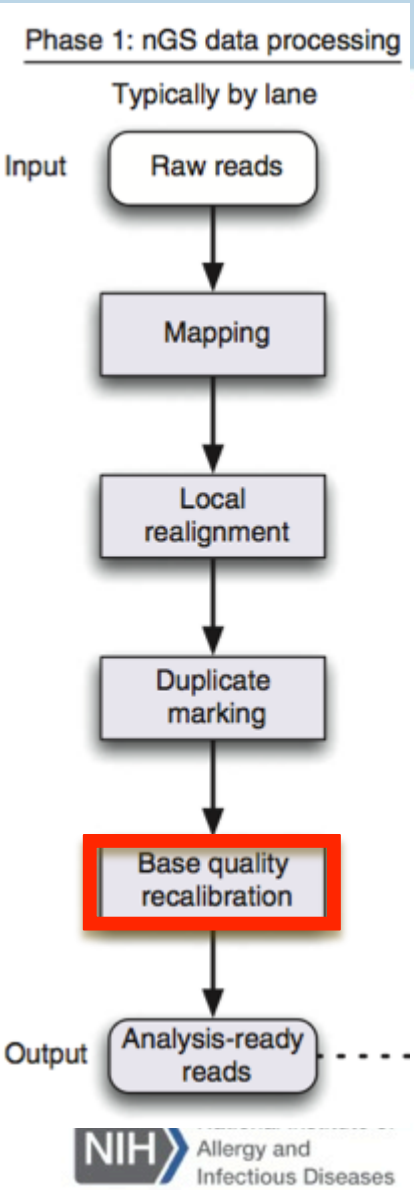
National Institute of Allergy and Infectious Diseases

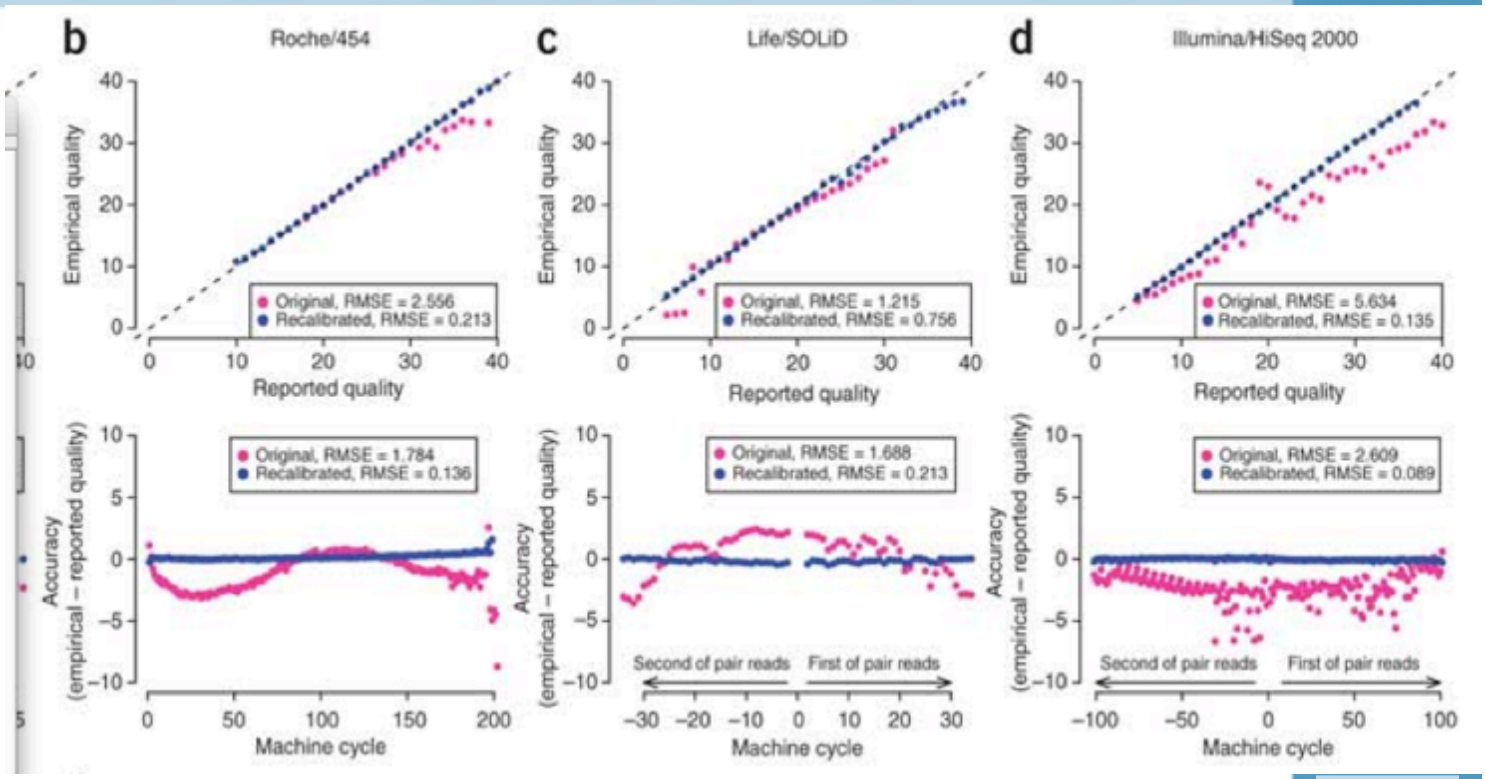
http://www.broadinstitute.org/gatk/guide/topic?name=methods-and-workflows

GATK Haplotype Caller shows improvements on detecting misalignments around gaps and reducing false positives



GATK uses a Base Quality Recalibration method





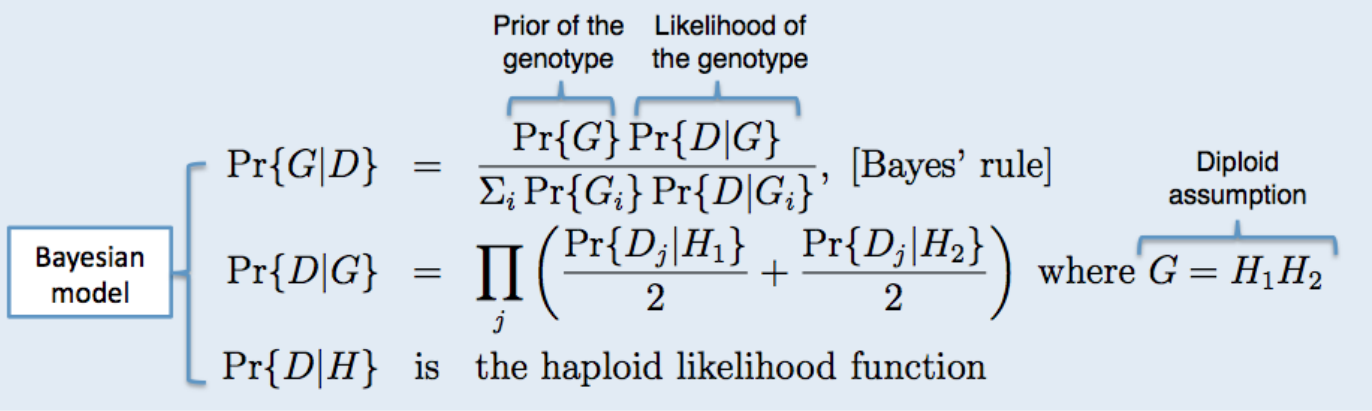
Covariates

- the raw quality score
- the position of the base in the read
- the dinucleotide context and the read group

http://www.nature.com/ng/journal/v43/n5/full/ng.806.html

NIH Allergy and Infectious Diseases

GATK uses a bayesian model



- It determines possible SNP and indel alleles
- Computes, for each sample, for each genotype, likelihoods of data given genotypes
- Computes the allele frequency distribution to determine most likely allele count, and emit a variant call if determined

– When it reports a variant, it assign a genotype to each sample

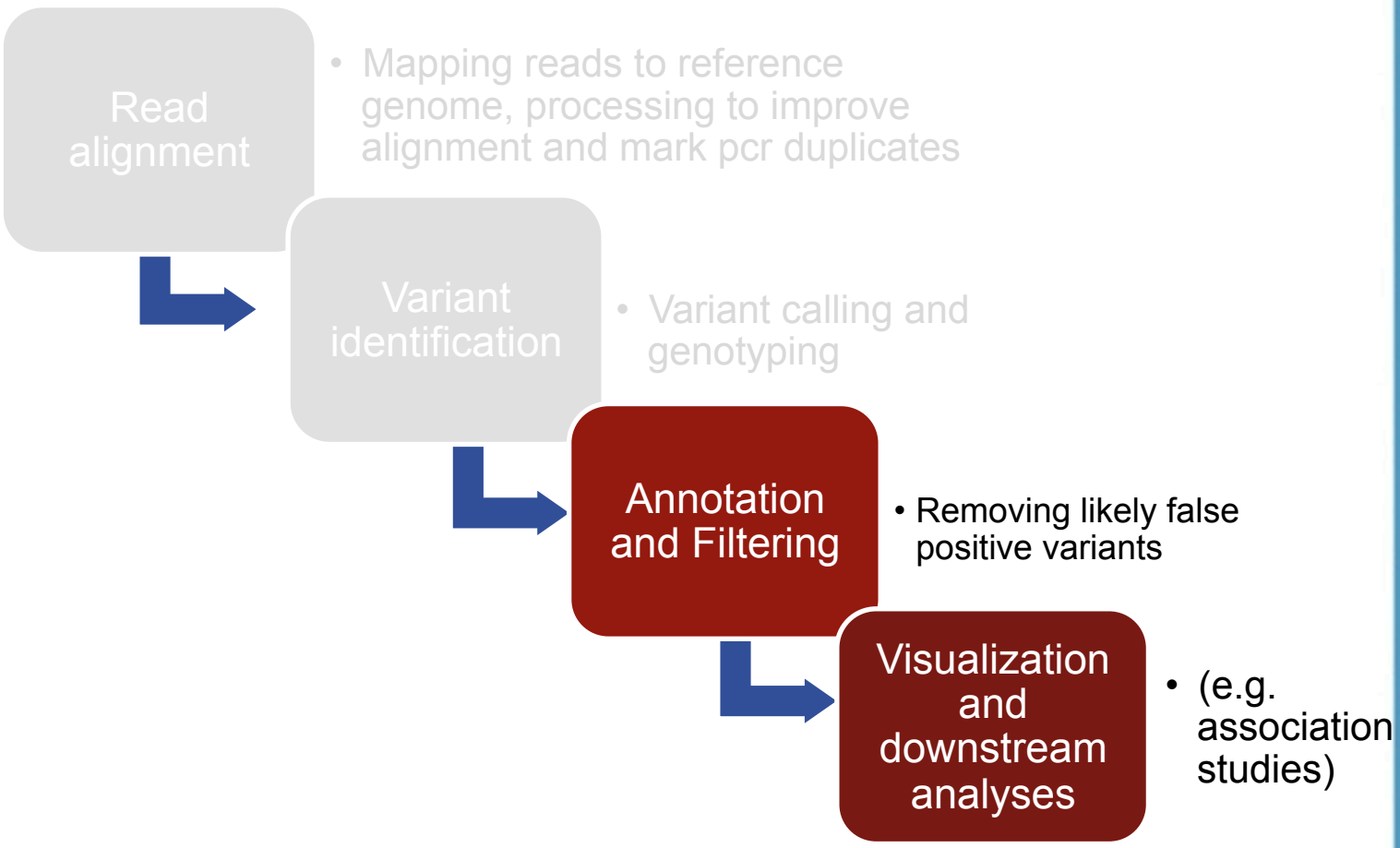
http://www.broadinstitute.org/gatk//events/2038/GATKwh0-BP-5-Variant_calling.pdf

NIAID

From the output of thousands of variants, which ones should you consider?

- Various important considerations
  - Is the variant call of good quality?
  - If you expect a rare mutation, is the variant commonly found in the general population?
  - What is the predicted effect of the variant?
    - Non-synonymous
    - Detrimental for function

A typical variant calling pipeline

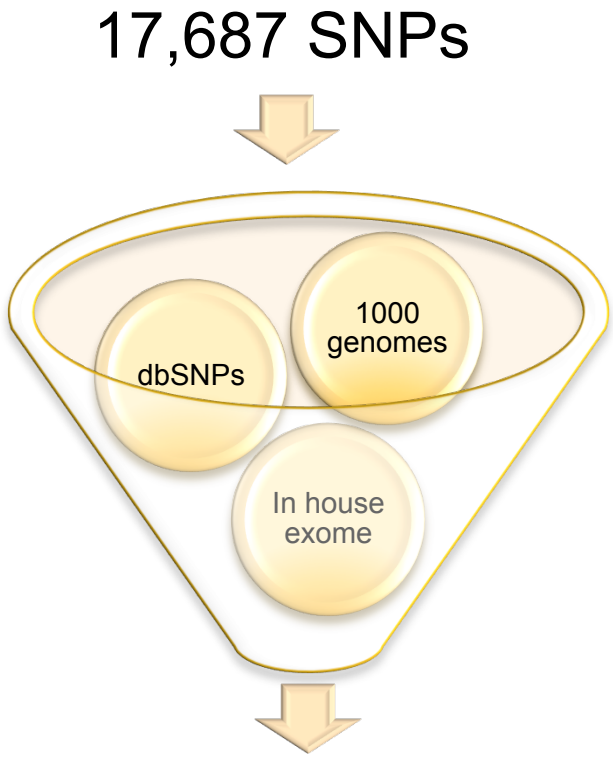


National Institute of Allergy and Infectious Diseases

NIAID

Filtering variants

- The initial set of variants is usually filtered extensively in hopes of removing false positives.
- More filtering can include public data or custom filters based on in-house data

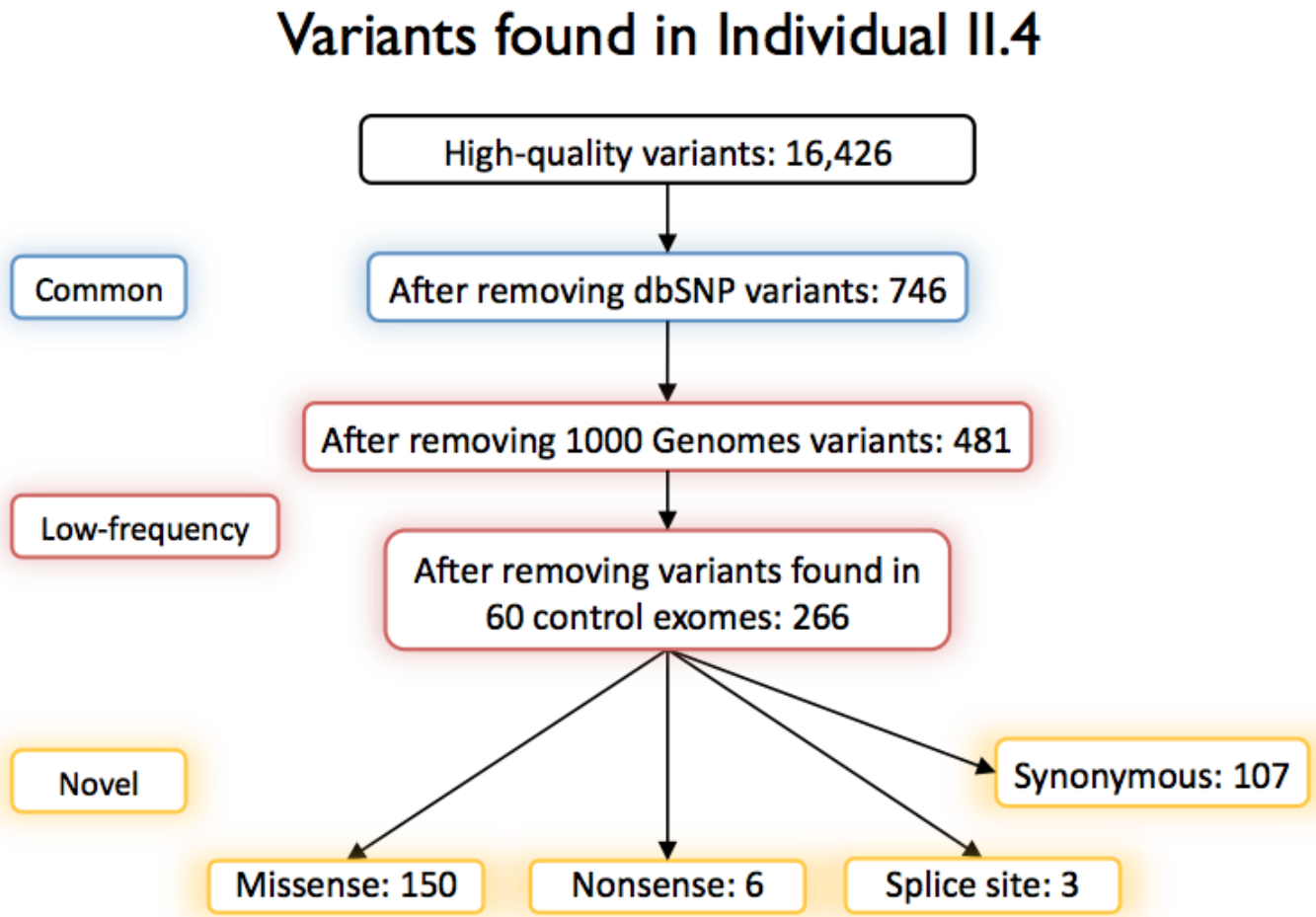


National Institute of Allergy and Infectious Diseases

PLoS One. 2012;7(1):e29708

NIAID

Removing the common and low frequency to get the novel SNPs



http://www.cardiovasculairegeneeskunde.nl/bestanden/Kathiresan.pdf

Musunuru*, Pirruccello*, Do*, *N Engl J Med* 2010

NIAID

Filtering variants

- The initial set of variants is usually filtered extensively in hopes of removing false positives.
- More filtering can include public data or custom filters based on in-house data

- Most exome studies will then filter common variants (>1%) such as those in dbSNP database
  - Good tools for filtering and annotation: Annovar, SNPeff
  - Reports genes at or near variants as well as the type (non-synonymous coding, UTR, splicing, etc.)

Filtering variants

- The initial set of variants is usually filtered extensively in hopes of removing false positives.
- More filtering can include public data or custom filters based on in-house data
- Most exome studies will then filter common variants (>1%) such as those in dbSNP database
  - Good tools for filtering and annotation: Annovar, SNPeff
  - Reports genes at or near variants as well as the type (non-synonymous coding, UTR, splicing, etc.)
- To filter by variant consequence, use SIFT and Polyphen2. These are available via Annovar or ENSEMBL VEP http://www.ensembl.org/Homo_sapiens/Tools/VEP
- For a more flexible annotation workflow, explore GEMINI

  http://gemini.readthedocs.org/en/latest/

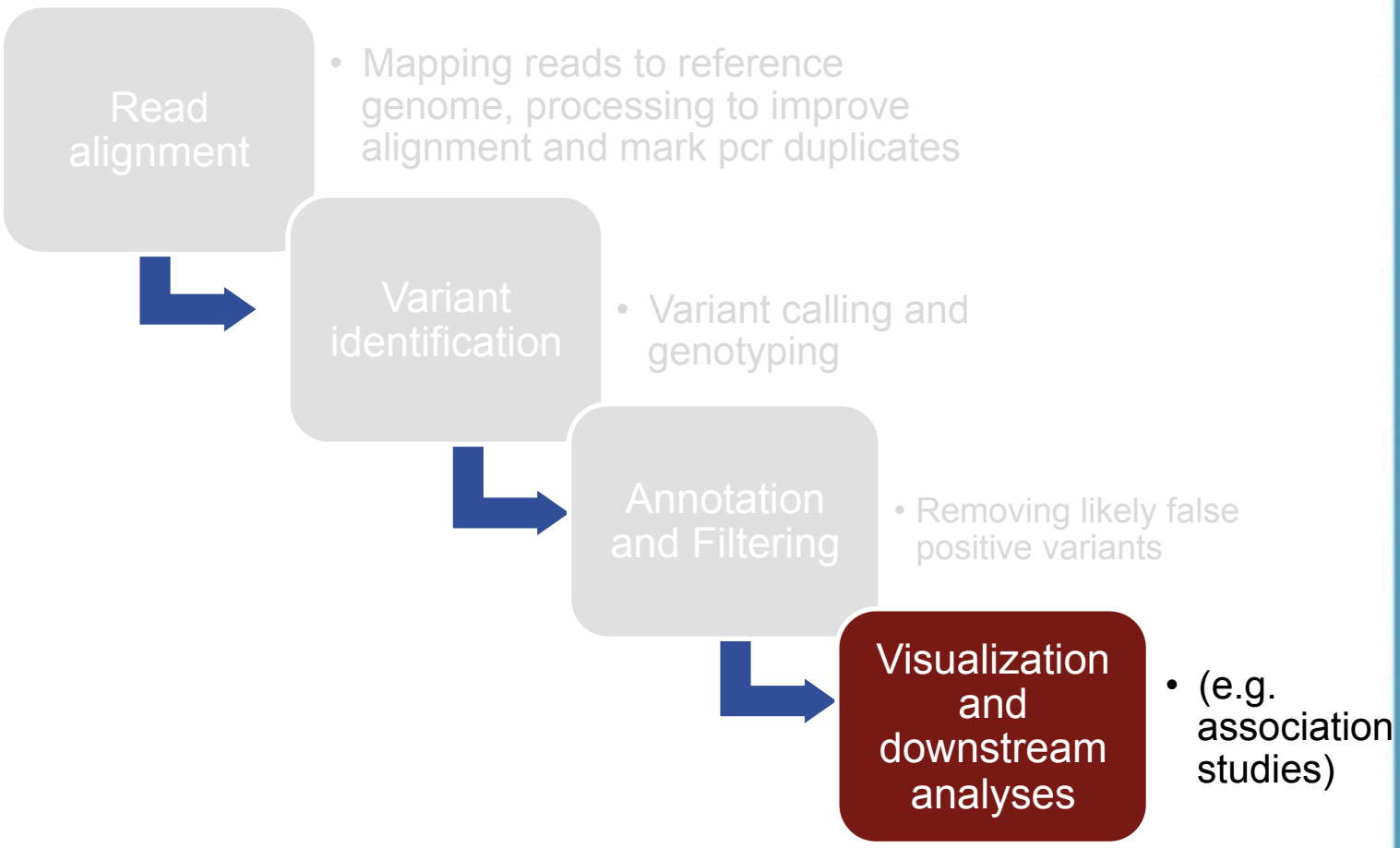
Evaluating Variant Consequence

SIFT predictions - SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids.

PolyPhen predictions - PolyPhen is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.

Condel consensus predictions - Condel computes a weighed average of the scores (WAS) of several computational tools aimed at classifying missense mutations as likely deleterious or likely neutral. The VEP currently presents a Condel WAS from SIFT and PolyPhen.

A typical variant calling pipeline



Viewing reads in browser

- If your genome is available via the UCSC genome browser http://genome.ucsc.edu/, import bam format file to the UCSC genome browser by hosting the file on a server and providing the link.

- If your genome is not in UCSC, use another browser such as IGV http://www.broadinstitute.org/igv/ , or IGB http://bioviz.org/igb/
  - Import genome (fasta)
  - Import annotations (gff3 or bed format)
  - Import data (bam)

Variant Analysis – Class Topics

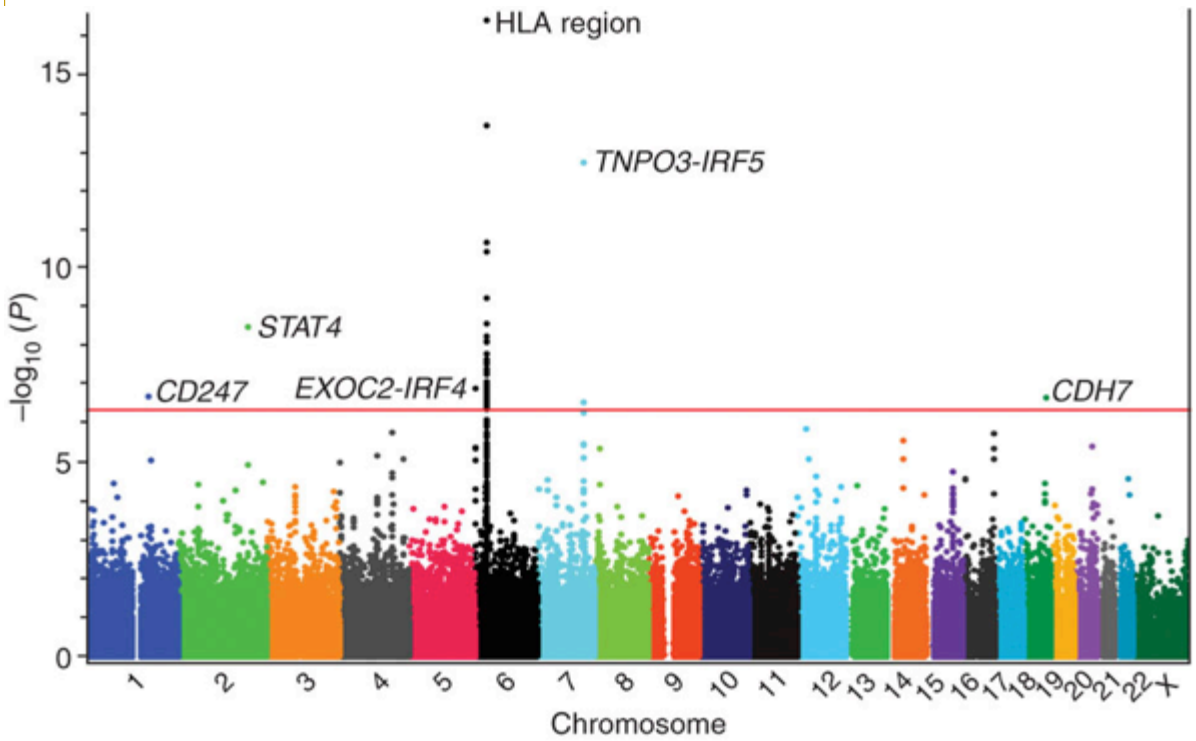
National Institute of Allergy and Infectious Diseases

NIAID

Before exome-seq there was GWAS

GWAS – Genome wide association study

- Examination of common genetic variants in individuals and their association with a trait
- Analysis is usually association to disease (Case/control)

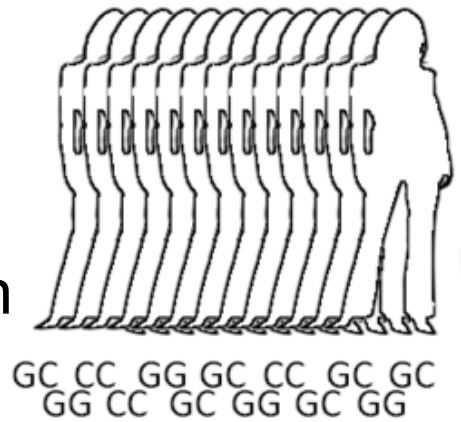


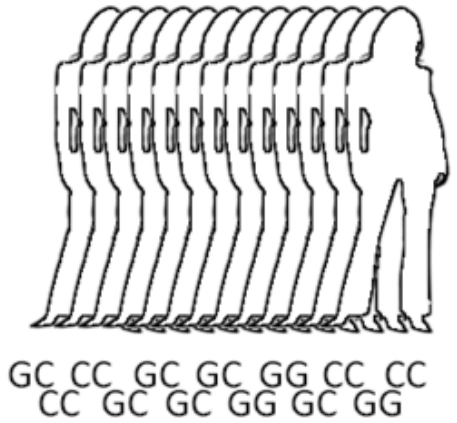
Manhatan Plot – each dot is a SNP, Y-axis shows association level

Genome-wide association study of systemic sclerosis identifies *CD247* as a new susceptibility locus
Nature Genetics **42**, 426–429 (2010)

GWAS Association studies

- A typical analysis:
  - Identify SNPs where one allele is significantly more common in cases than controls
  - Hardy-Weinberg Chi Square





SNP1	*SNP2*	*SNP …*
Cases Count of G: 2104 of 4000	**Cases** Count of G: 1648 of 4000	*Repeat for all SNPs*
Frequency of G: 52.6%	Frequency of G: 41.2%	
Controls Count of G: 2676 of 6000	**Controls** Count of G: 2532 of 6000	
Frequency of G: 44.6%	Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

National Institute of Allergy and Infectious Diseases

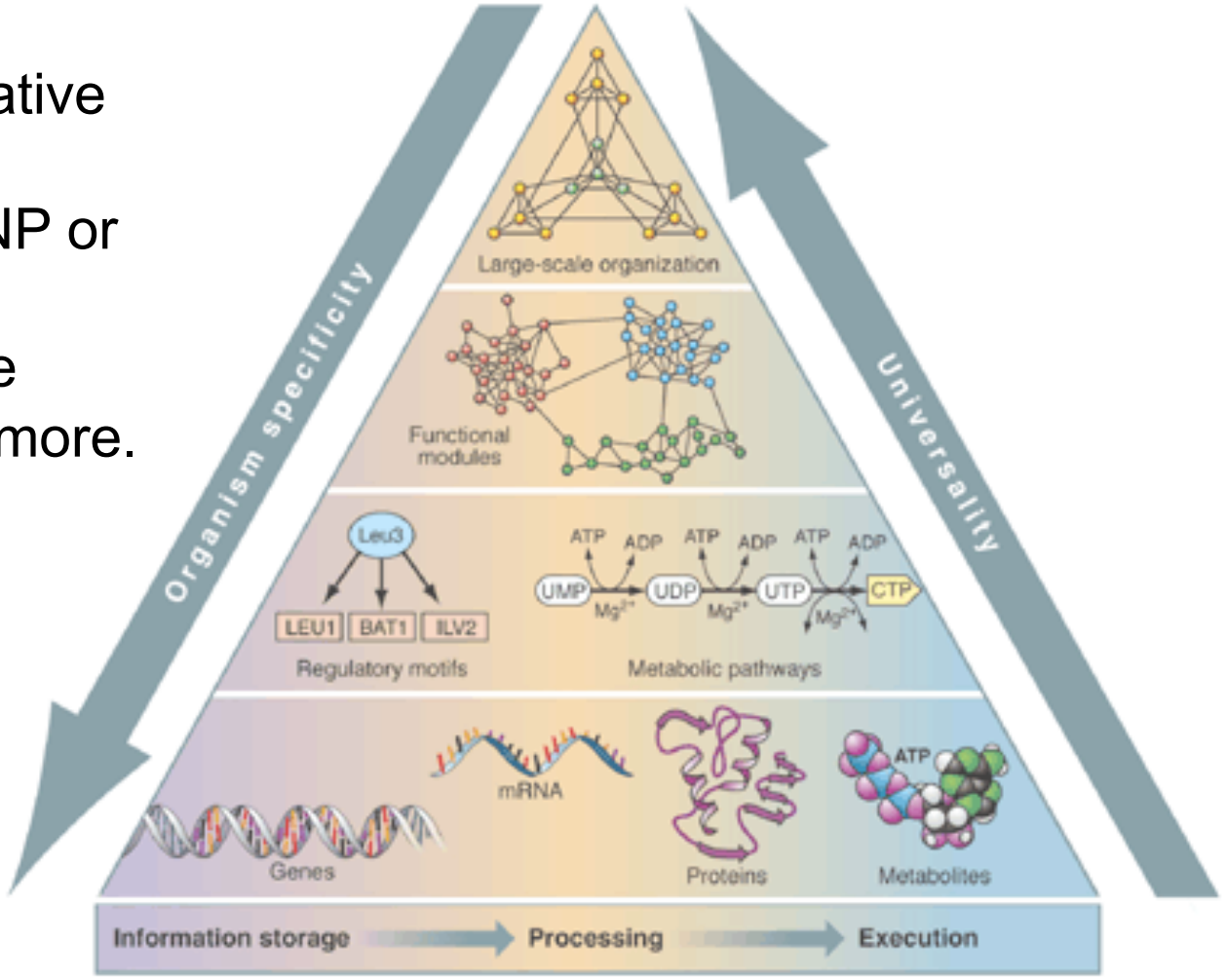
http://en.wikipedia.org/wiki/Genome-wide_association_study

Analysis Tools for genotype-phenotype association from sequencing data

- Plink-seq - http://atgu.mgh.harvard.edu/plinkseq/
- EPACTS - http://genome.sph.umich.edu/wiki/EPACTS
- SNPTestv2 / GRANVIL - http://www.well.ox.ac.uk/GRANVIL/

How will I know if the candidate variant is the one I am looking for?

Consider a more integrative approach, which could include adding more SNP or exome data, biological function, pedigree, gene exprssion network and more.



Science 2002 Oltvai et al., pp. 763 - 764

National Institute of Allergy and Infectious Diseases

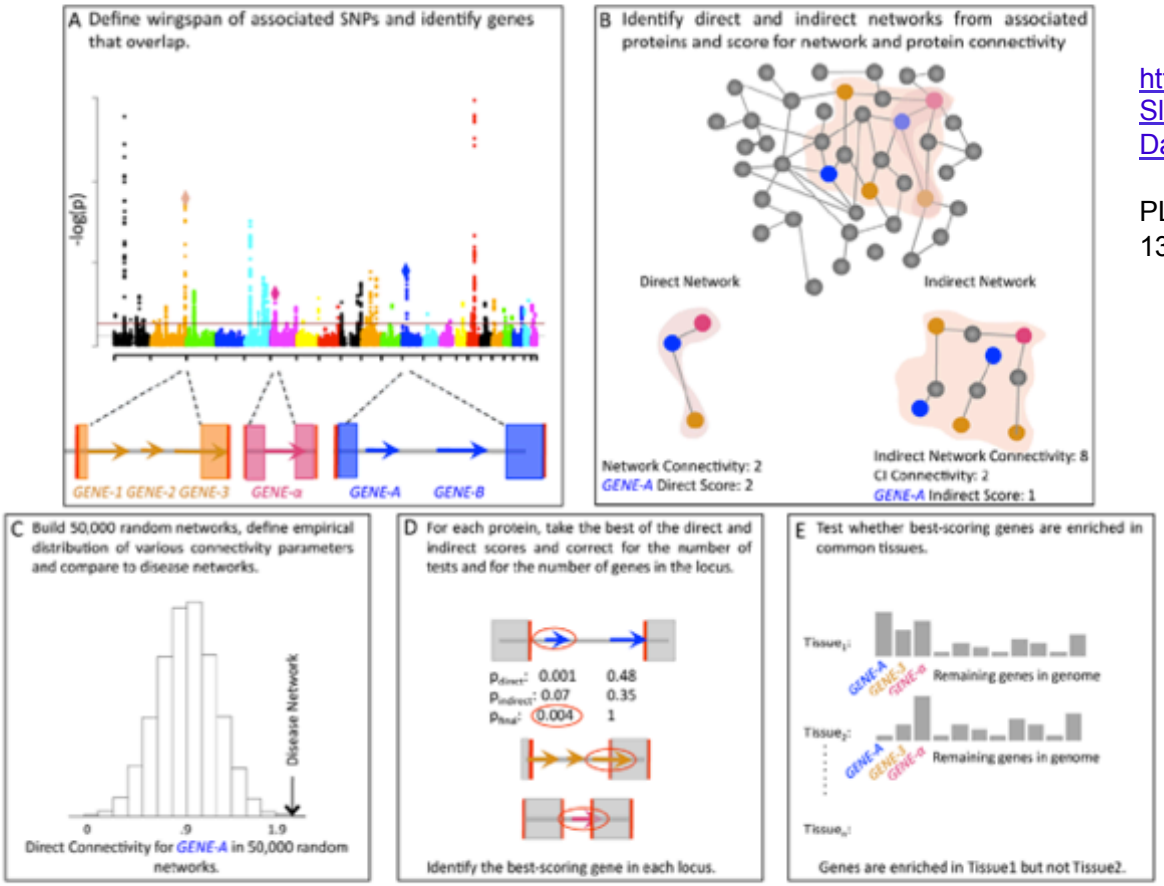
A more integrative approach

OPEN ACCESS Freely available online

PLoS GENETICS

Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology

Elizabeth J. Rossin[1,2,3,4,5], Kasper Lage[2,3,6,7], Soumya Raychaudhuri[1,2,8], Ramnik J. Xavier[1,2,3], Diana Tatar[6], Yair Benita[1], International Inflammatory Bowel Disease Genetics Constortium[¶], Chris Cotsapas[1,2*], Mark J. Daly[1,2,3,4,5*]



http://www.genome.gov/Multimedia/Slides/SequenceVariants2012/Day2_IntegratedApproachWG.pdf

PLoS Genet. 2011 Jan 13;7(1):e1001273

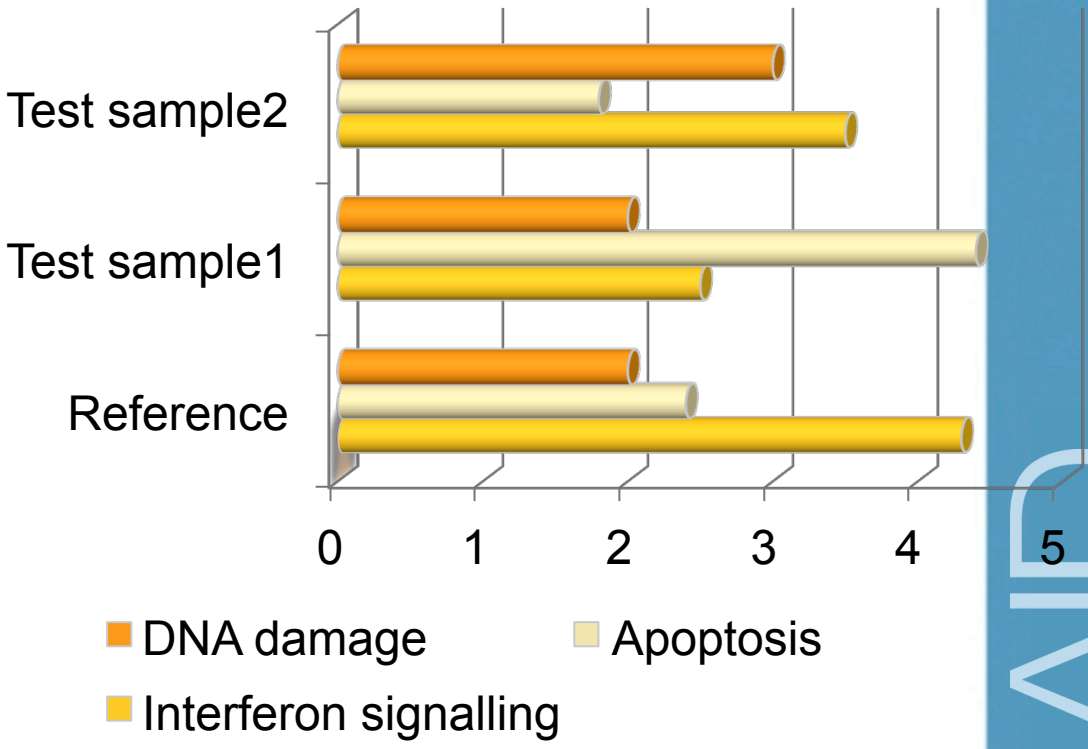
NIAID

Pathways Analysis

Enrichment and Network

With pathways analysis we can study:

- How genes and molecules communicate
- The effect of a disruption to a pathway in relation to the associated phenotype or disease
- Which pathways are likely to be affected an affected individual



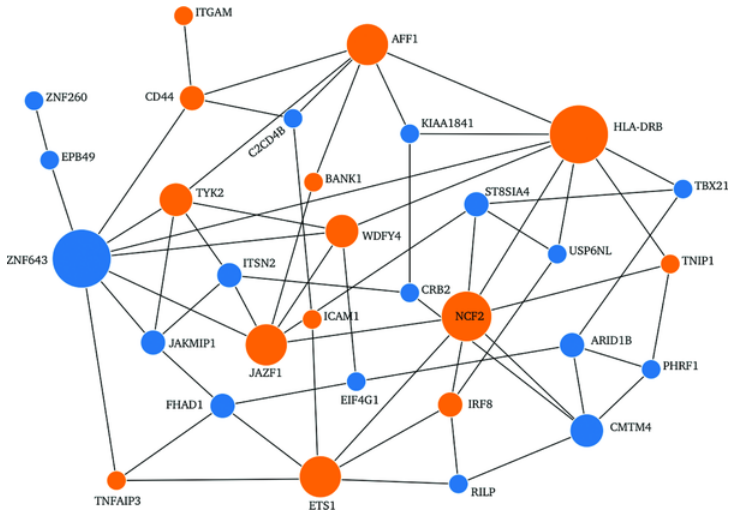
Networks Analysis

- **Encore: Genetic Association Interaction Network Centrality Pipeline and Application to SLE Exome Data**

Nicholas A. Davis[1,†,‡], Caleb A. Lareau[2,3,‡], Bill C. White[1], Ahwan Pandey[1], Graham Wiley[3], Courtney G. Montgomery[3], Patrick M. Gaffney[3], B. A. McKinney[1,2,*]

Article first published online: 5 JUN 2013

  - open source network analysis pipeline for genome-wide association studies and rare variant data

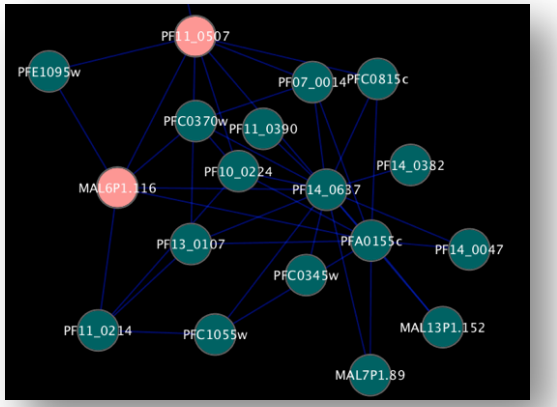


National Institute of Allergy and Infectious Diseases

NIAID

Visualization

- Browsers
  - IGV - http://www.broadinstitute.org/igv/ - see slide 51
  - UCSC Genome browser - http://genome.ucsc.edu/
- Other
  - Cytoscape – can be used for example to visualize variants in the context of affected genes

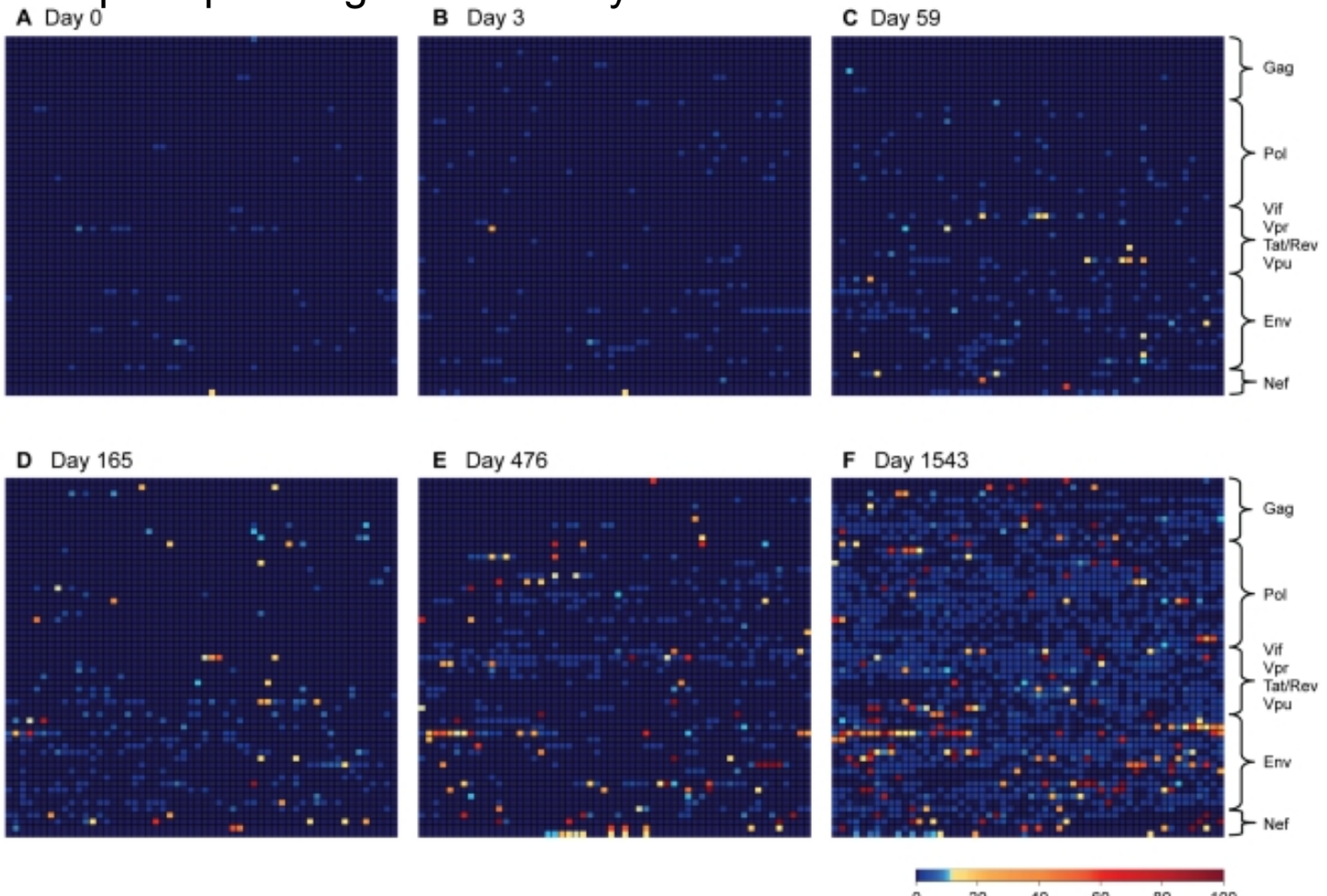


National Institute of Allergy and Infectious Diseases

NIAID

Other applications of variant calling in Deep Sequence: Virus evolution

Deep Sequencing detects early variants in HIV evolution



http://www.broadinstitute.org/scientific-community/science/projects/viral-genomics/v-phaser

Infectious Diseases

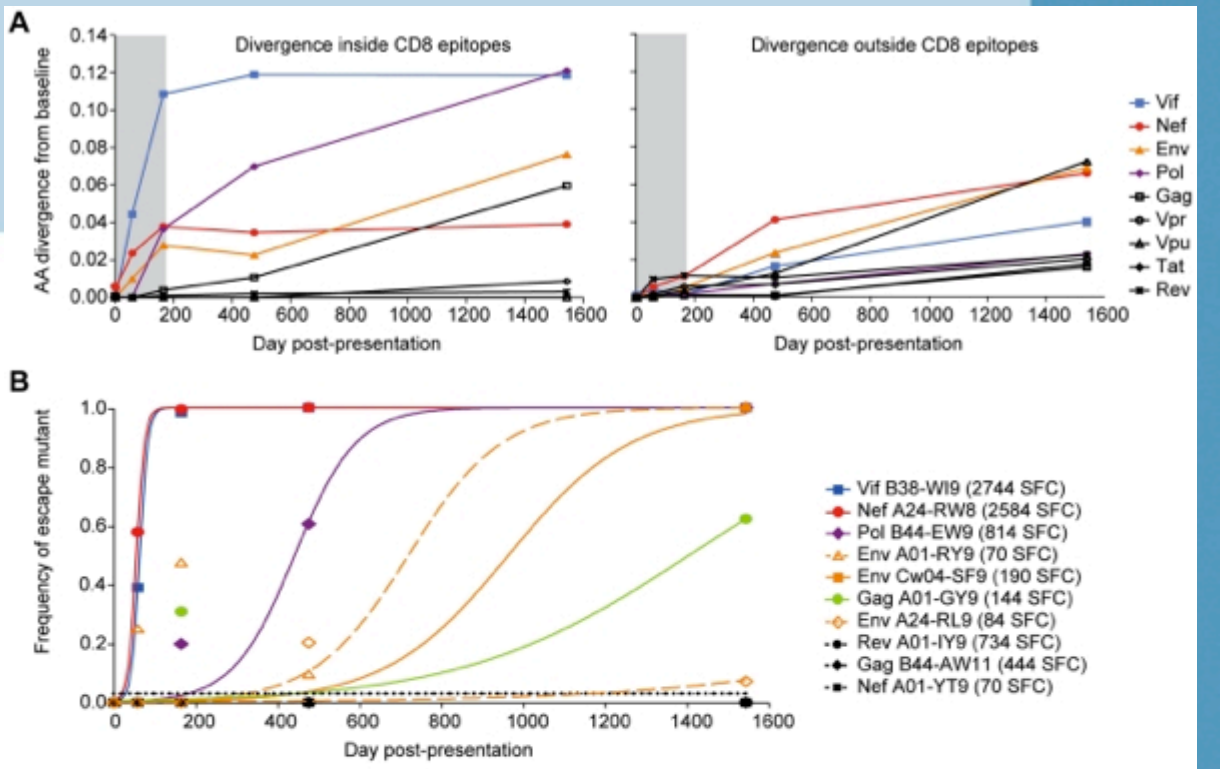
http://www.plospathogens.org/article/info%3Adoi%2F10.1371%2Fjournal.ppat.1002529

NIAID

Quasispecies evolution

Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection

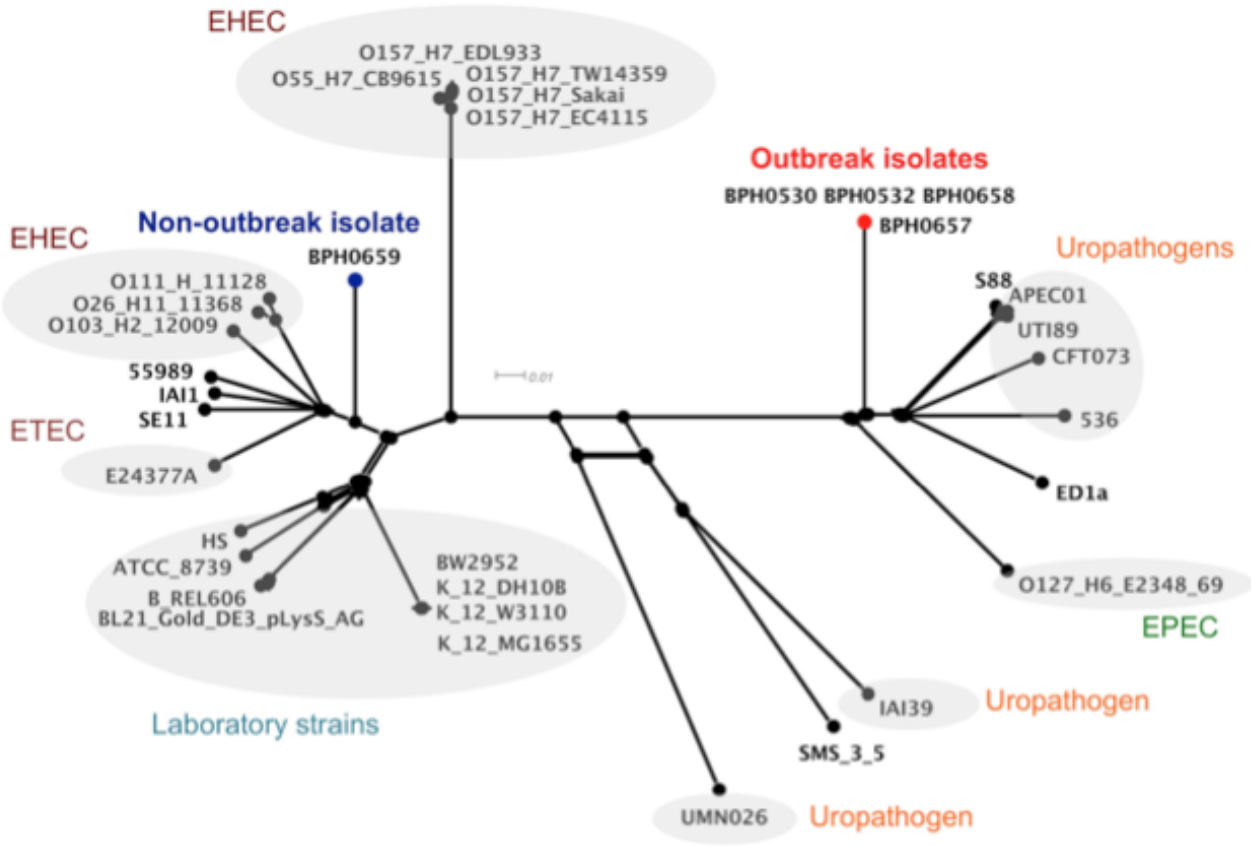
- They concluded that the early control of HIV-1 replication by immunodominant CD8+ T cell responses may be substantially influenced by rapid, low frequency viral adaptations not detected by conventional sequencing approaches, which warrants further investigation.



http://www.plospathogens.org/article/info%3Adoi%2F10.1371%2Fjournal.ppat.1002529

Other applications of variant calling in Deep Sequence: Fast genotyping

Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory. J Clin Microbiol. 2013 Feb 13



Ion Torrent was used!

NIH Infectious Diseases

NIAID

Links for variant calling and filtering tools

- **Variant Annotation**
- Annovar http://www.openbioinformatics.org/annovar/
- snpEff http://snpeff.sourceforge.net
- SeattleSeq Annotation http://snp.gs.washington.edu/SeattleSeqAnnotation137/

- **Variant Consequence**
- SIFT: http://sift.jcvi.org/
- Polyphen-2: http://genetics.bwh.harvard.edu/pph2/

- Multiple tools for VCF file manipulation, variant calling, phenotype association
- Galaxy - https://main.g2.bx.psu.edu/root

Thank You

Question or Comments please contact:

quinonesm@niaid.nih.gov

ScienceApps@niaid.nih.gov

National Institute of Allergy and Infectious Diseases

NIAID