UCDAVIS Bioinformatics Core

So you want to do a Variant Analysis Project

Matt Settles

Director, Bioinformatics Core

Treating Bioinformatics as a Data Science

Seven stages to data science

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

Data science done well looks easy and that's a big problem for data scientists

simplystatistics.org
March 3, 2015 by Jeff Leek

What is Variant Analysis

The Focus on identification of variation in DNA from genomic, exome, targeted sequencing, or other reduced representation sequencing data.

Variants Include:

- Single Nucleotide Variants (SNV)
- Short insertion/deletions (INDEL)
- Copy number gain/loss (CNV)
- Large novel insertions (> 1kb)
- Inversions

http://varnomen.hgvs.org/recommendations/DNA/

Designing Experiments

Beginning with the question of interest (and working backwards)

- The final step of an analysis is the application of a model to each variant in your dataset.

  Traditional statistical considerations and basic principals of statistical design of experiments apply.

  - **Control** for effects of outside variables, avoid/consider possible biases, avoid confounding variables in sample preparation.
  - **Randomization** of samples, plots, etc.
  - **Replication** is essential (minimum here depends on MAF and how strongly the variant is associated with phenotype)
- You should know your final model and comparison of interest before beginning your experiment.
  - Here a variant predicts a specific phenotype effect which may need to account for background and other variables.

General rules for preparing samples

- Prepare more samples then you are going to need, i.e. expect some will be of poor quality, or fail
- Preparation stages should occur across all samples at the same time (or as close as possible) and by the same person
- Spend time practicing a new technique to produce the highest quality product you can, reliably
- Quality should be established using Fragment analysis traces (pseudo-gel images, for RNA RIN > 7.0)
- DNA/RNA should not be degraded
  - 260/280 ratios for RNA should be approximately 2.0 and 260/230 should be between 2.0 and 2.2. Values over 1.8 are acceptable
- Quantity should be determined with a Fluorometer, such as a Qubit.

Generating libraries

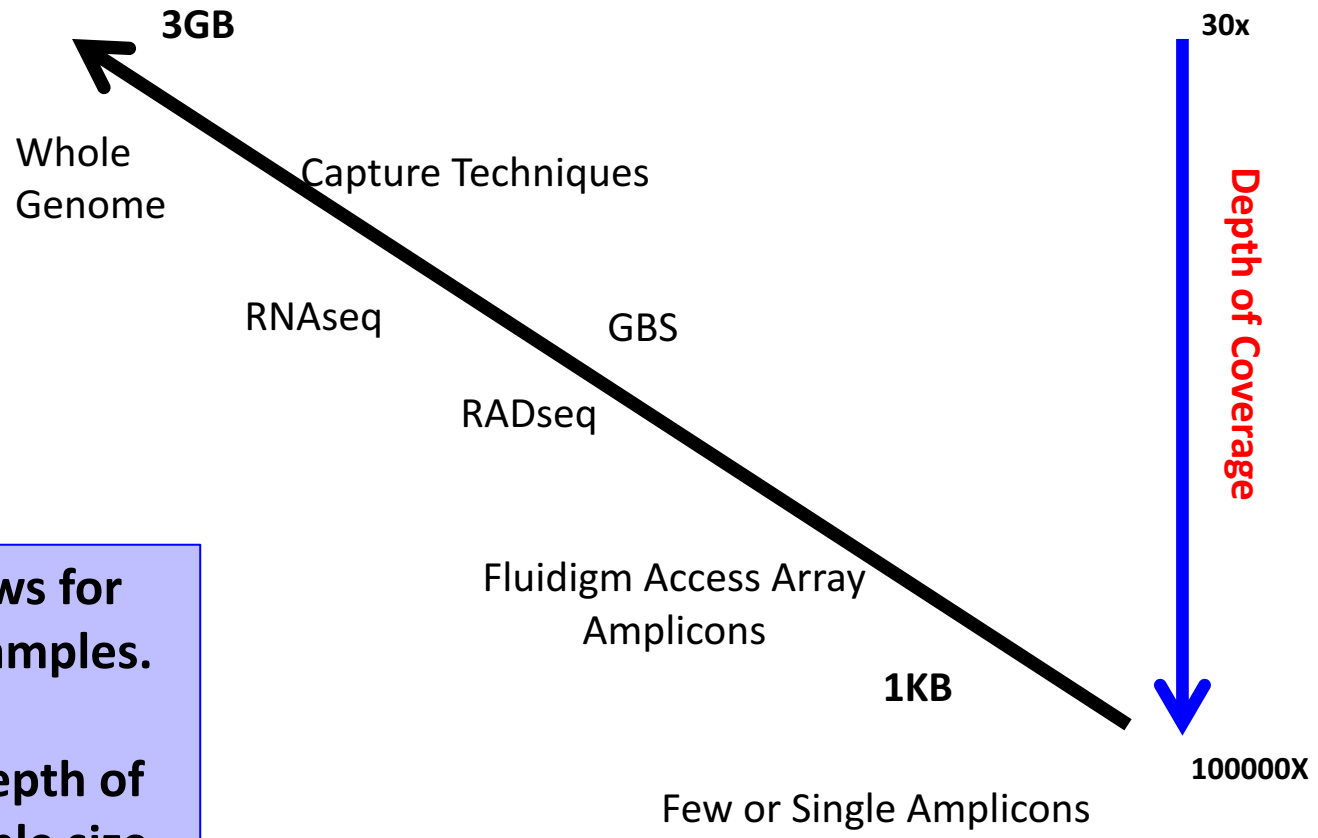
Types of Libraries

- Whole genome shotgun

Reduced Representation Libraries

- Exome capture
- RADseq
- GBS
- Many many others

Genomic Reduction



Genomic reduction allows for greater multiplexing of samples.

You can fine tune your depth of coverage needs and sample size with the reduction technique

Sequenced Basepairs per samples per lane

The first and most basic question is how many base pairs of sequence will I get

Factors to consider then are:

1. Number of reads being sequenced
2. Read length (if reads are paired, consider them as individuals for this calculations)
3. Number of samples being sequenced
4. Expected percentage of good bases/reads

$$\frac{bp}{sample} = \frac{readLength * (\#reads)}{\#samples} * 0.8$$

The number of reads and read length data are best obtained from the manufacturer's website (search for specifications) and always use the lower end of the estimate.

Genomic Coverage

Once you have the number of base pairs per sample you can then determine expected coverage

Factors to consider then are:

1. Length of the genome/capture/target
2. Any extra-genomic sequence (ie mitochondria, virus, plasmids, etc.). For bacteria in particular, these can become a significant percentage

$$\frac{ExpectedCoverage}{sample} = \frac{(readLength * numReads) * 0.8}{numSamples} * \text{num.lanes}}{TotalGenomicContent}$$

Genomic Coverage – Reduced Representation libraries

Once you have the number of base pairs per sample you can then determine expected coverage

Factors to consider then are:

1. Length of the genome/capture/target
2. Any extra-genomic sequence (ie mitochondria, virus, plasmids, etc.). For bacteria in particular, these can become a significant percentage
3. PCR duplication percentage [Use a higher fudge factor 0.5 to be conservative]

$$\frac{ExpectedCoverage}{sample} = \frac{\frac{(readLength * numReads) * 0.5}{numSamples} * \text{num.lanes}}{TotalGenomicContent}$$

Sequencing Depth – Counting based experiments

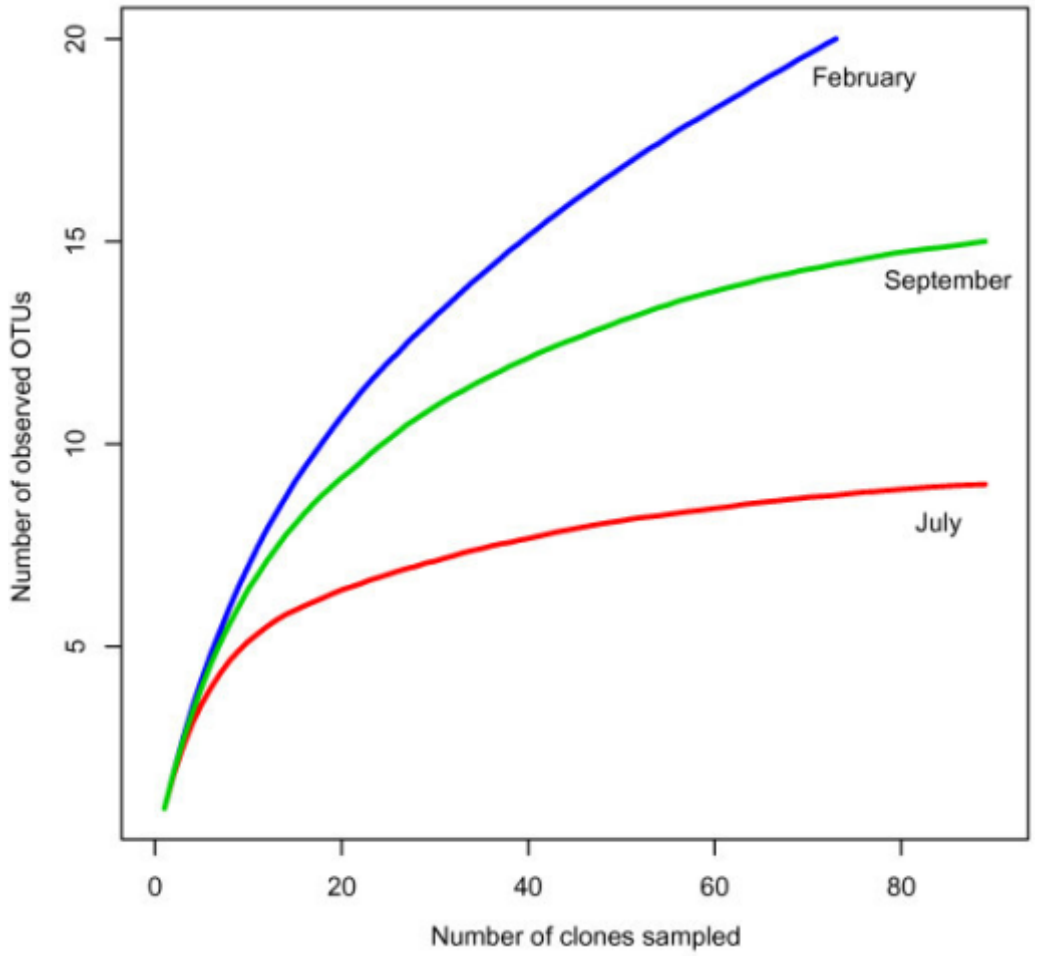
- Coverage is determined differently for ”Counting” based experiments (RAD, GBS, etc) where an expected number of reads per site per sample is typically more suitable.
- The first and most basic question is how many reads per sample will I get Factors to consider are (per lane):
  1. Number of reads being sequenced
  2. Number of samples being sequenced
  3. Expected percentage of usable data
  4. Number of lanes being sequenced
  5. Sites being targeted

$$\frac{reads/site}{sample} = \frac{\frac{reads.sequenced * 0.8}{num.sites}}{samples.pooled} * \text{num.lanes}$$

- Read length, or SE vs PE, does not factor into sequencing depth.

Sequencing Depth – Counting based experiments Did I sequence enough?

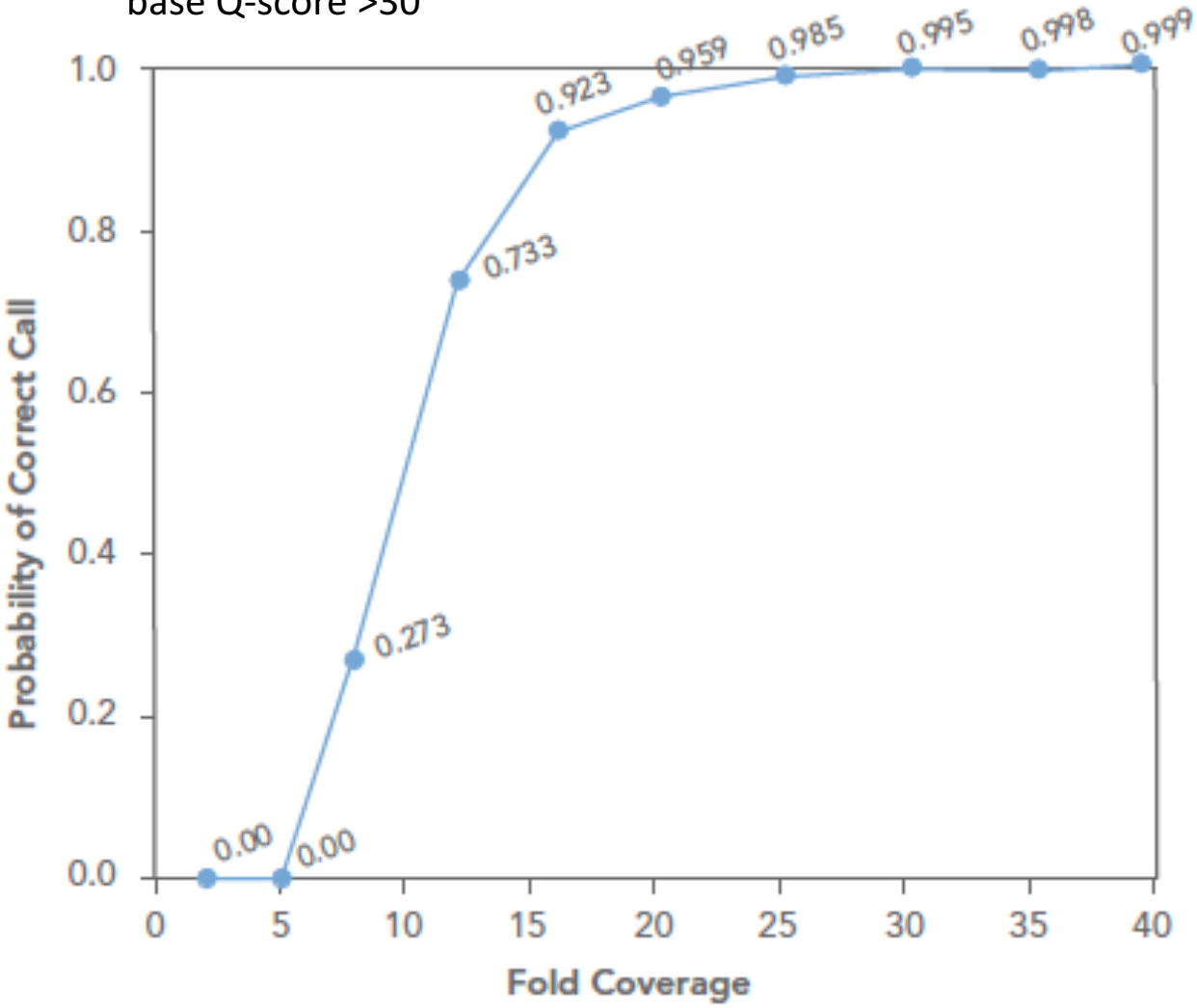
- ’Deep’ sequence a number of test samples
- Plot rarefactions curves of sites identifide, to determine if saturation is achieved



Variant Analysis

- Read length contributes to uniqueness of mapping
- Paired reads are required to identify structure changes
- For a single individual target > 30x coverage is desired.
- In population studies, the greater the number of samples less coverage per samples that is required. (ex. with 1000 samples 2x coverage per sample may be sufficient)

Probability of a correct SNP call, assuming base Q-score >30



Take Homes

- Experience and/or literature searches (other peoples experiences) will provide the best justification for estimates on needed depth.
- ‘Longer’ reads are better than short reads.
- Paired-end reads are more useful than single-end reads
- Libraries can be sequenced again, so do a pilot, perform a preliminary analysis, then sequence more accordingly.

Cost Estimation

- Extractions from tissue (DNA/RNA): cost per sample
- Sample quality assurance. Including quantification and sample degradation evaluation: cost per sample
- Library generation and quantification: cost per sample
- Pooling and quantification of libraries: cost per group
- Sequencing (type if sequencing PE/SE, length of reads, number of lanes / runs): cost per lane/run
- Bioinformatics, general rule is to estimate double your budget)

EX: http://dnatech.genomecenter.ucdavis.edu/prices/

UCDAVIS Bioinformatics Core

Bioinformatics Costs

Bioinformatics includes:

1. Storage of data
2. Access and use of computational resources and software
3. System Administration time
4. Bioinformatics Data Analysis time
5. Back and forth consultation/analysis to extract biological meaning

Rule of thumb:
Bioinformatics can and should cost as much (sometimes more) as the cost of data generation.

Barcodes and Pooling samples for sequencing

- Best to have as many barcodes as there are samples
  - Can purchase barcodes from vendor, generate them yourself and purchase from IDTdna (example), or consult with the DNA technologies core.
- Best to pool all samples into one large pool, then sequence multiple lanes
- IF you cannot generate enough barcodes, or pool into one large pool, RANDOMIZE samples into pools.
  - Bioinformatics core can produce a randomization scheme for you.
  - This must be considered/determined PRIOR to library preparation

Illumina Hiseq sequencing

- http://www.illumina.com/systems/hiseq-3000-4000/specifications.html

	HISEQ 3000 SYSTEM	HISEQ 4000 SYSTEM
No. of Flow Cells per Run	1	1 or 2
Data Yield: 2 × 150 bp 2 × 75 bp 1 × 50 bp	 650-750 Gb 325-375 Gb 105-125 Gb	 1300-1500 Gb 650-750 Gb 210-250 Gb
Clusters Passing Filter (Single Reads) (8 lanes per flow cell)	2.1-2.5 billion	4.3-5 billion
Quality Scores: 2 × 50 bp 2 × 75 bp 2 × 150 bp	 ≥ 85% bases above Q30 ≥ 80% bases above Q30 ≥ 75% bases above Q30	 ≥ 85% bases above Q30 ≥ 80% bases above Q30 ≥ 75% bases above Q30
Daily Throughput	> 200 Gb	> 400 Gb
Run Time	< 1-3.5 days	< 1-3.5 days
Human Genomes pe	up to 6	up to 12
Exomes per Run**	up to 48	up to 96
Transcriptomes per R	up to 50	up to 100

UCDAVIS Bioinformatics Core

Illumina Novaseq Sequencing

[Novaseq]

Sequencing Output per Flow Cell

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell Type	S1*	S2	S3*	S4*
2 × 50 bp	up to 167 Gb	280–333 Gb	NA**	NA**
2 × 100 bp	up to 333 Gb	560–667 Gb	NA**	NA**
2 × 150 bp	up to 500 Gb	850–1000 Gb	up to 2000 Gb	up to 3000 Gb

Specifications based on Illumina PhiX control library at supported cluster densities.
*The NovaSeq 5000 System, NovaSeq 5000 System Upgrade, and NovaSeq Reagent Kits with S1, S3, or S4 flow cells are not currently available for order.
** NA: not applicable

Reads Passing Filter

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell Type	S1*	S2	S3*	S4*
	up to 1.6 B	2.8–3.3 B	up to 6.6 B	up to 10 B

Cost Estimation (exercise 1)

- DNA/RNA extraction and QA/QC (Per sample)
- library preparation (Per sample)
  - Library QA/QC (Bioanalyzer and Qubit)
  - Any enrichment technique
- Sequencing (Number of lanes)
- Bioinformatics (General rule is to estimate the same amount as data generation, i.e. double your budget)

http://dnatech.genomecenter.ucdavis.edu/prices/

Example: DNA - 32 human samples, Agilent SureSelect V6 (use ~75Mb and $500/library), target ~ 100x coverage per exon, sequence on Hiseq 4000.

Cost Estimation (exercise 2)

Example: DNA- 18 human samples, whole genome library, target > 30x average coverage on the Novaseq – 2x150

Use $16,200 per Novaseq S2 2x150bp run

Cost Estimation (exercise 2)

Example: DNA- 18 human samples, whole genome library, target > 30x average coverage on the Novaseq – 2x150

Use $16,200 per Novaseq S2 2x150bp run

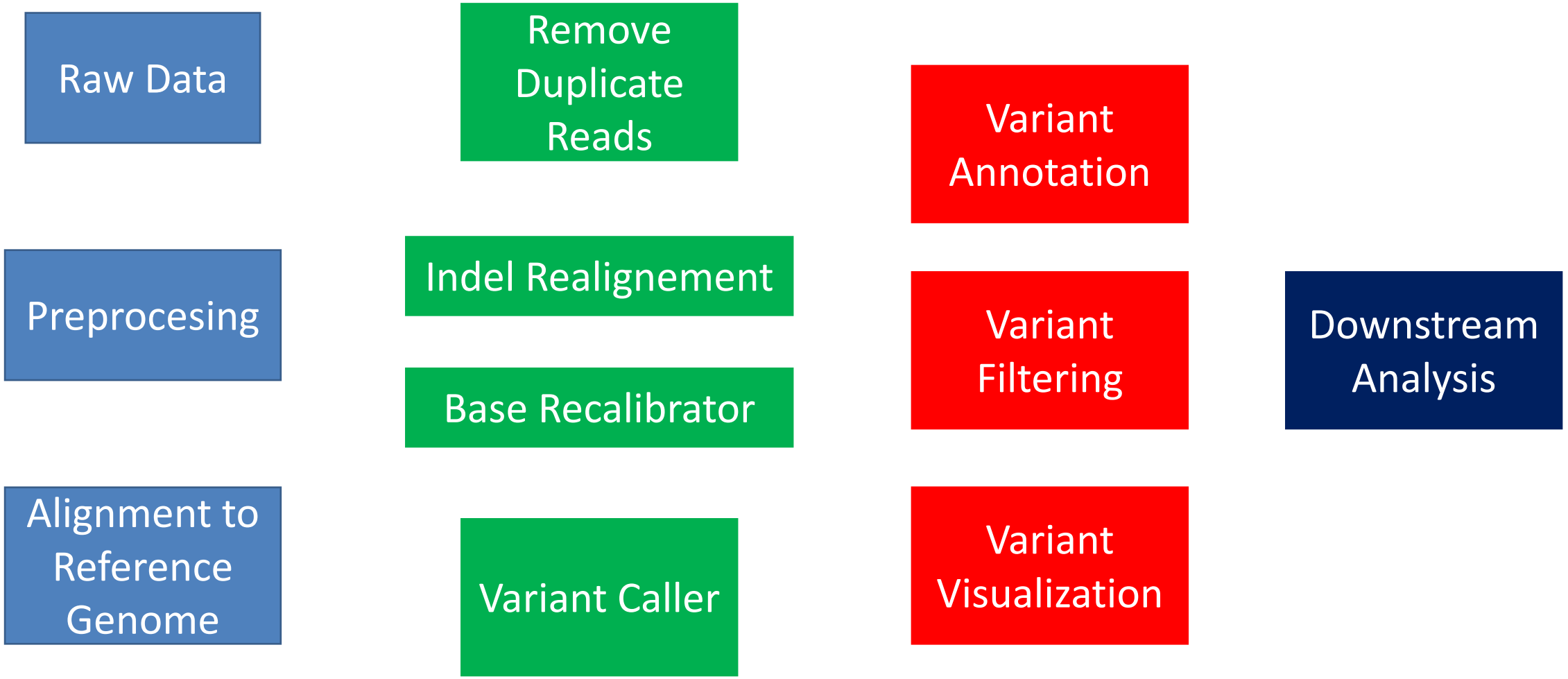
UCDAVIS Bioinformatics Core

Overview of Variant data analysis

Prerequisites

- Access to a multi-core (24 cpu or greater), ‘high’ memory 64Gb or greater Linux server.
- Familiarity with the ’command line’ and at least one programming language.
- Basic knowledge of how to install software
- Basic knowledge of R (or equivalent) and statistical programming
- Basic knowledge of Statistics and model building

Variant pipeline overview [GATK example]



Sequence Preprocessing: Why?

- We have found that aggressively "cleaning" and processing reads can make a large difference to the **speed** and **quality** of assembly and mapping results. Cleaning your reads means, removing reads/bases that are:
  - other unwanted sequence (polyA tails in RNA-seq data)
  - artificially added onto sequence of primary interest (vectors, adapters, primers)
  - join short overlapping paired-end reads
  - low quality bases
  - originate from PCR duplication
  - not of primary interest (contamination)
- Preprocessing also produces a number of statistics that are technical in nature that should be used to evaluate "experimental consistancy"

Sequence Preprocessing: QA/QC

- Beyond generating better data for downstream analysis, cleaning statistics also give you an idea as to the quality of the sample, library generation, and sequencing quality used to generate the data.
- This can help inform you of what you might do in the future.
- I've found it best to perform QA/QC on both the run as a whole (poor samples can affect other samples) and on the samples themselves as they compare to other samples (REMEMBER, BE CONSISTANT).
  - Reports such as Basespace for Illumina, are great ways to evaluate the runs as a whole.
  - PCA/MDS plots of the preprocessing summary are a great way to look for technical bias across your experiment

Mapping: Consideration

- Placing reads in regions that do not exist in the reference genome (reads extend off the end) [mitochondrial, plasmids, structural variants, etc.].
- Sequencing errors and variations: alignment between read and true source in genome may have more differences than alignment with some other copy of repeat.
- What if the closest fully sequenced genome is too divergent? (3% is a common assumed alignment capability)
- Placing reads in repetitive regions: Some algorithms only return 1 mapping; If multiple: map quality = 0
- Algorithms that use paired-end information => might prefer correct distance over correct alignment.

Mapping: Many Aligners to choose from

https://en.wikipedia.org/wiki/List_of_sequence_alignment_software

Generally BWA works well

- BWA aln for < 70bp reads
- BWA mem for >= 70bp reads

Mapping: QA/QC

- Mapper produce summary statistics, view the summary report (in a text editor) and compare across samples.
  - Other additional summary statistics can be produced with:

samtools flagstat

samtools idxstats

samtools stats

Indel Realignment, Base Recalibration

- Local realignment of insertions and deletions. Insertion-deletion (indel) mutations, with respect to the reference genome, can contain misalignments across BAM files. Misalignment of indel mutations, which can often be erroneously scored as substitutions, reduces the accuracy of downstream variant calling steps.
- A base quality score recalibration (BQSR). This step adjusts base quality scores based on detectable and systematic errors. This step also increases the accuracy of downstream variant calling algorithms.

Variant Caller – Haplotype callers

- Older techniques would try and call variants based directly on the alignment, one base at a time. Current ‘haplotype’ based callers use the reads directly and short sets of sequences to call variants.

- Identify variant bases, genotype likelihood and allele frequency while avoiding instrument noise.

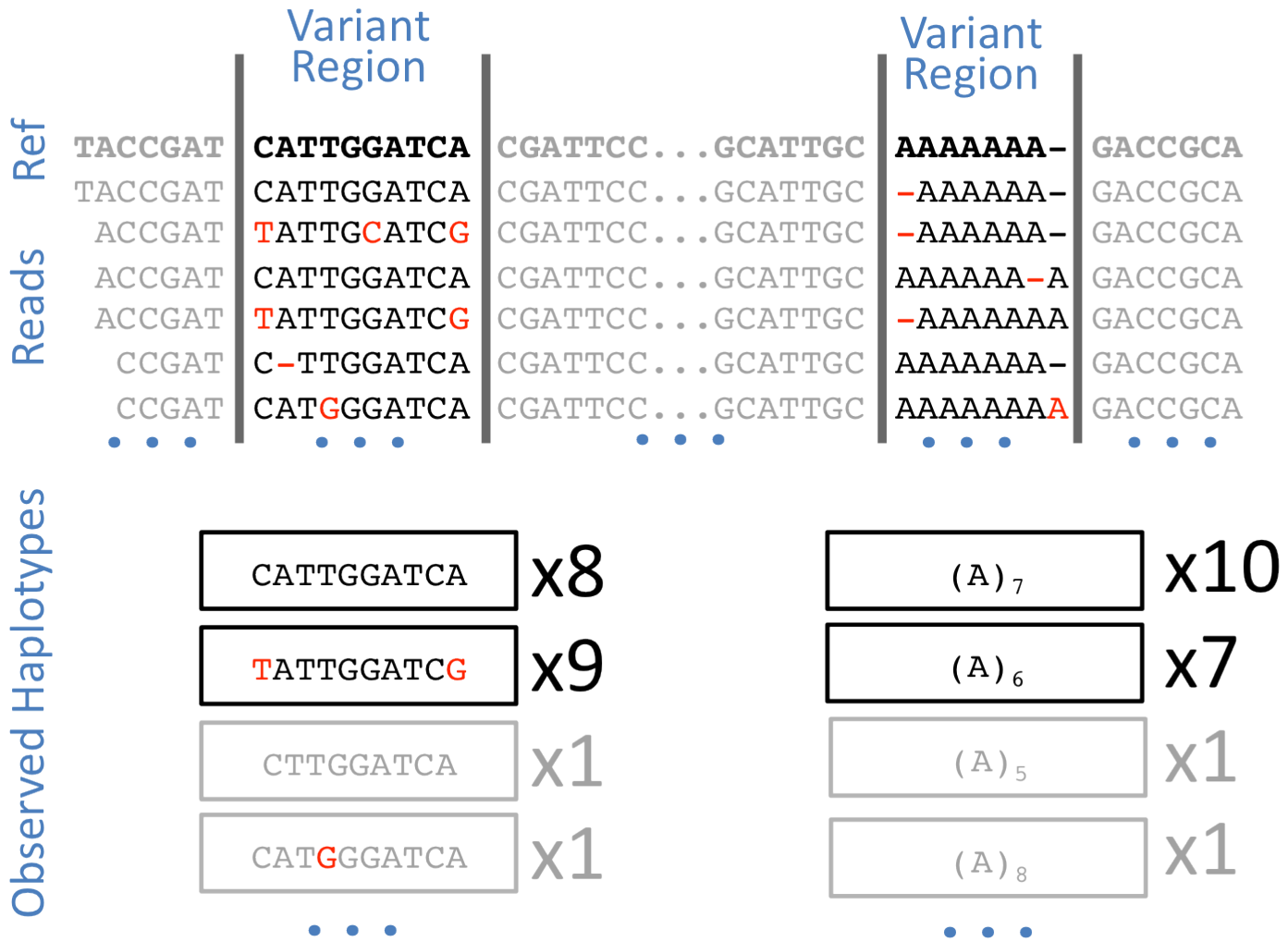
- Variant Analysis produces a VCF (Variant Call Format) file.

Variant Calling – Challenges

- Base Calling Errors:
  - Errors vary by technology/instrument, increase as the numbers of cycles increases and the sequence context (ex. GC)
- Low coverage sequence
  - Not enough sequence coverage to infer the alleles at a site
- Inaccurate mapping
  - Aligned reads should be reported with a mapping score

UCDAVIS Bioinformatics Core

Variant Calling: Alignments can be incorrect



VCF File format

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        N
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1
```

Variant Annotation – SNPeff

Variant annotation and effect prediction tool

- Using a genome fasta file (same used to map reads to) and a gtf genome annotation file, creates a database specific to an organism.
- Given a VCF file with predicted variants, SNPs, INDEL, etc.
- Produces an annotated VCF (Addition of the ANN field), calculates the effects they produce on known genes (e.g. amino acid changes).
- Variant annotation uses the coordinates and alleles in the VCF file to infer biological context for each variant including the location of each mutation, its biological consequence (frameshift/silent mutation), and any affected genes.

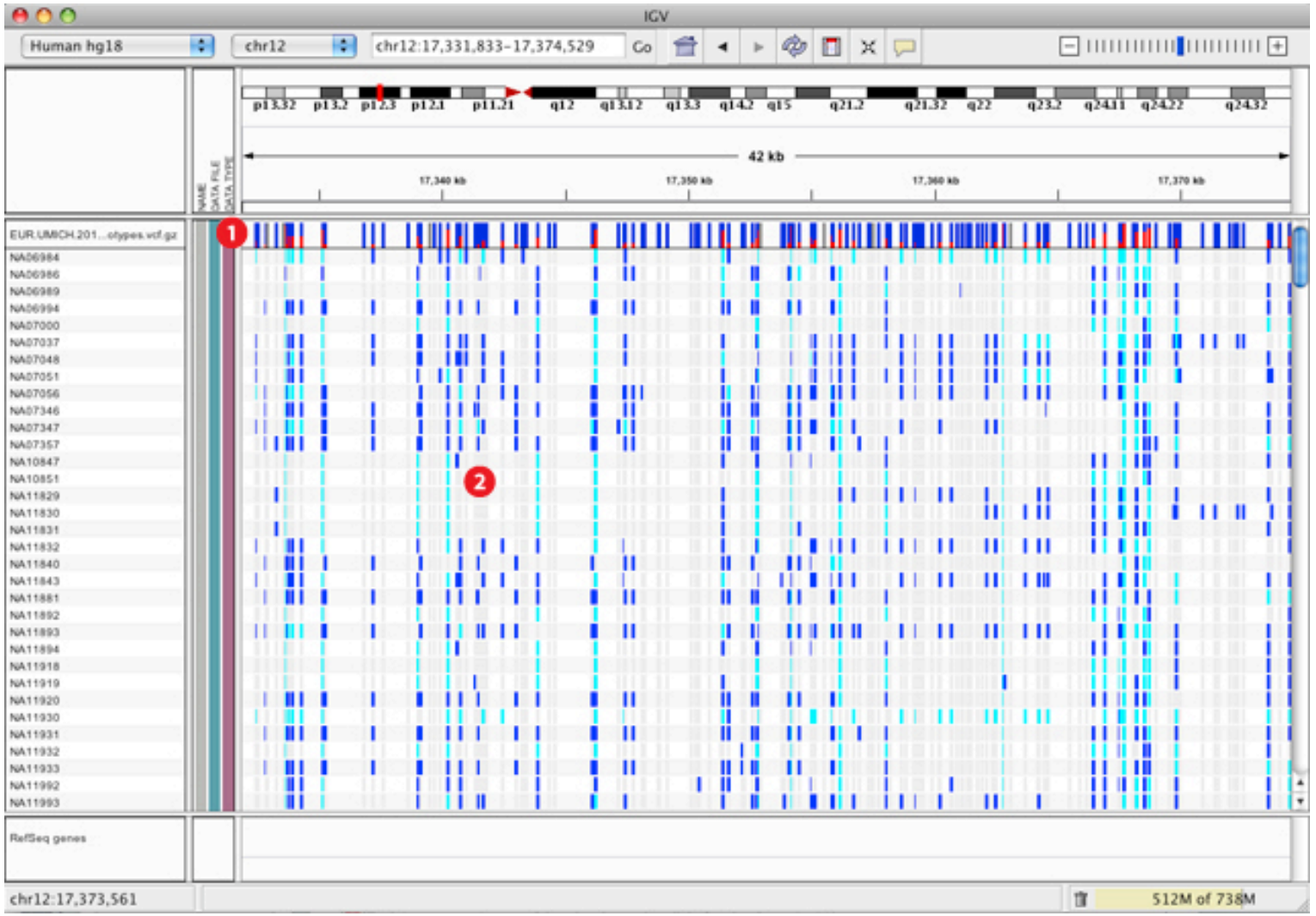
Variant Filtering

Variant analysis usually produces hundreds of thousands (euk) of candidate variants, not all are 'real' or 'interesting'. Filtering is usually experiment based, and takes into account expectations.

- Remove all low quality variants
- In gene/ not in gene
- Synonymous/non-Synonymous SNPs
- Consider only rare-mutations
- Comparison to public data

UCDAVIS Bioinformatics Core

Visualization - IGV



Downstream Analysis

- Downstream analysis is often to determine genotype-phenotype association from sequencing data (Based on GWAS)
- Application of statistical models to determine whether SNPs or Haplotypes are associated with a phenotype
- Gene Set Enrichment Analysis for variants in Pathways/GO

Software

Preprocessing:

- Python 2.7
  - Modules: argparse, optparse, distutils
- bowtie2 - contaminant screening
  - http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- Super-Deduper – Identify and remove PCR duplicates
  - https://github.com/dstreett/Super-Deduper
- Sickle – Trim low quality regions
  - https://github.com/dstreett/sickle
- Scythe – Identify and remove adapters in SE reads
  - https://github.com/ucdavis-bioinformatics/scythe
- FLASH2 – Join overlapping reads, identify and remove adapter in PE reads
  - https://github.com/dstreett/FLASH2

Software

Mapping:

- Bwa mem – map reads to a reference
  - http://sourceforge.net/projects/bio-bwa/files/
- samtools – processing of sam/bam file
  - http://www.htslib.org/

Mapping adjustments, Variant Analysis:

- GATK – Genome Analysis Tool Kit
  - https://software.broadinstitute.org/gatk/
- Freebayes
  - https://github.com/ekg/freebayes

Software

Variant Annotation and Filtering

- SNPeff
  - http://snpeff.sourceforge.net/
- SNPsift
  - http://snpeff.sourceforge.net/SnpSift.html
- Annovar
  - http://annovar.openbioinformatics.org/en/latest/

Visualization

- IGV - Integrative Genome Viewer
  - http://software.broadinstitute.org/software/igv/

Software

Analysis of Variants:

- Plink-seq
  - https://atgu.mgh.harvard.edu/plinkseq/
- EPACTS
  - https://genome.sph.umich.edu/wiki/EPACTS

- R http://www.r-project.org/
  - R Packages: TVTB, VariantTools
    - http://bioconductor.org/packages/release/bioc/html/TVTB.html
    - http://bioconductor.org/packages/release/bioc/html/VariantTools.html
- RStudio
  - https://www.rstudio.com/