UCDAVIS Bioinformatics Core

GWAS/Cancer Studies

Matt Settles

UC Davis Bioinformatics Core

What is GWAS

- Most commonly performed using genotype arrays (SNP arrays), but methods are now translated to sequence data.
- Goal is to test alleles, genotypes, or haplotypes of these SNPs directly for association with a phenotype known to be heritable (hiehgt, diabetes, etc.).
  - Identify statistical connections between regions in the genome and the phenotype, in order to drive hypothesis for biological studies of genes/regions.
  - Generate insides in to the architecture of the phenotype. Ie. many small associated genes, or few large and concentrated.
  - Build models to predict phenotype from genotype.

Approach

- Collect subject, number needed is dependent on the effect size of the phenotype, heredity, and the expected complexity.
- Measure genotype across many location genome-wide. Arrays measure 10^5-10^6, moving to whole genome sequencing.
- Produces a matrix of samples by genotype, typically expressed as subjects as rows and SNPs as columns
- Association testing
  - Find SNPs that are typically associated with subject phenotype
  - Can be thought of as x separate statistical tests (where x is the number of SNPs) run on the matrix

UCDAVIS Bioinformatics Core

Associated SNPs

Cases:

AGAGCAGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCGCTAGAGCCGTGAGATCGACATGATAGCC
AGAGCCGTCGACATGTATAGTCTACATGAGATCGACATGAGATCGCTAGAGCAGTGAGATCGACATGATAGTC
AGAGCAGTCGACAGGTATAGTCTACATGAGATCGACATGAGATCGCTAGAGCCGTGAGATCGACATGATAGCC
AGAGCAGTCGACAGGTATAGCCTACATGAGATCAACATGAGATCGCTAGAGCAGTGAGATCGACATGATAGCC
AGAGCCGTCGACATGTATAGCCTACATGAGATCGACATGAGATCGCTAGAGCCGTGAGATCAACATGATAGCC
AGAGCCGTCGACATGTATAGCCTACATGAGATCGACATGAGATCGCTAGAGCAGTGAGATCAACATGATAGCC
AGAGCCGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCGCTAGAGCAGTGAGATCAACATGATAGTC
AGAGCAGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCTCTAGAGCCGTGAGATCGACATGATAGCC

Controls:

AGAGCAGTCGACATGTATAGTCTACATGAGATCGACATGAGATCGCTAGAGCAGTGAGATCAACATGATAGCC
AGAGCAGTCGACATGTATAGTCTACATGAGATCAACATGAGATCTCTAGAGCCGTGAGATCGACATGATAGCC
AGAGCAGTCGACATGTATAGCCTACATGAGATCGACATGAGATCTCTAGAGCCGTGAGATCAACATGATAGCC
AGAGCCGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCTCTAGAGCCGTGAGATCGACATGATAGTC
AGAGCCGTCGACAGGTATAGTCTACATGAGATCGACATGAGATCTCTAGAGCCGTGAGATCAACATGATAGCC
AGAGCAGTCGACAGGTATAGTCTACATGAGATCGACATGAGATCTCTAGAGCAGTGAGATCGACATGATAGCC
AGAGCCGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCTCTAGAGCCGTGAGATCGACATGATAGCC
AGAGCCGTCGACAGGTATAGTCTACATGAGATCAACATGAGATCTCTAGAGCAGTGAGATCGACATGATAGTC

Associated SNP

Associated SNPs

2x2 contingency table

	G	**T**
Case	7	1
Control	1	7

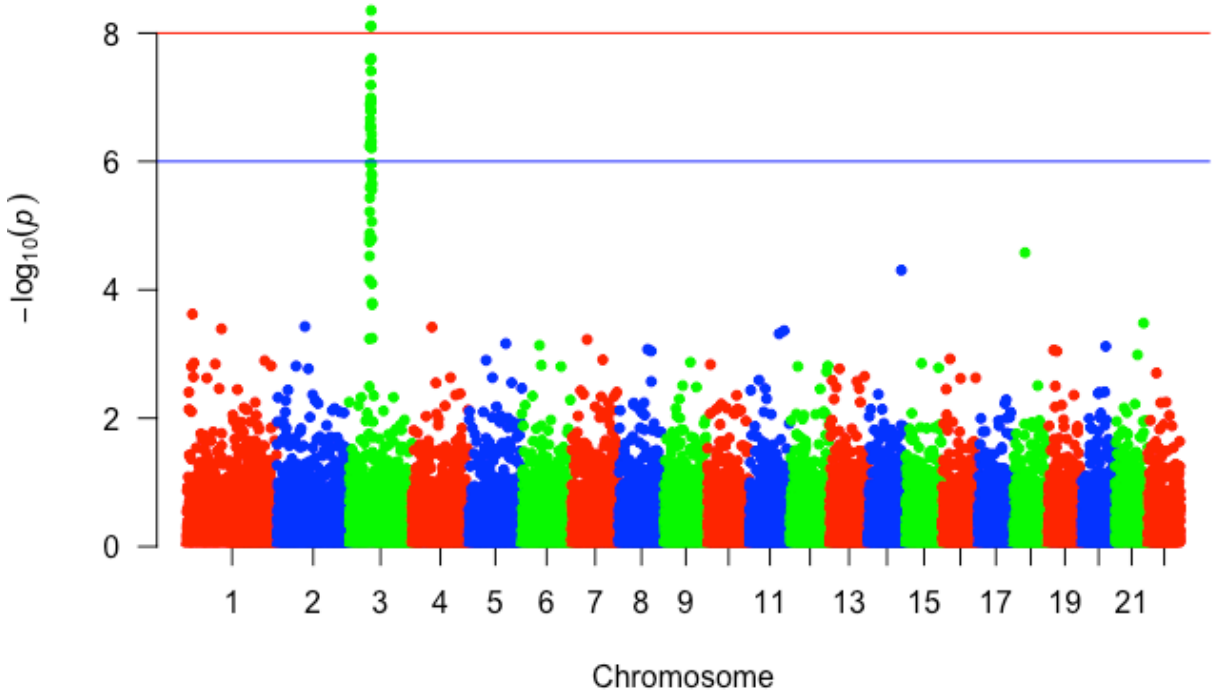
Perform Case-Control Association test (Pearson Chi Square

”Manhattan Plots” for GWAS

If we use a p-value threshold of 0.05 We would get many tens of thousands of ‘significant’ results

So people tend to be very selective in what results we declare as significant. Common thresholds are 10^{-6} and/or 10^{-8}

Declaring only one associated region in Chr3



Association testing with sequence data

- Plink/SEQ
  - https://atgu.mgh.harvard.edu/plinkseq/index.shtml
  - C code, fast, very flexible
- GenABEL
  - http://www.genabel.org/
  - Suite of R packages

Cancer

Cancer targets

- Sequence: mutations, polymorphisms
- Structure: copy number variation, translocation, loss-of-heterozygosity
- Function: expression, exon-usage
- External: Viruses, bacteria

Cancer DNA Sequencing Approaches

- Whole genome sequencing (WGS)
- Whole exome sequencing (WES)
- Targeted gene sequencing
- Targeted variant genotyping
- Epigenome modification (bisulphite)

DNA

- Transcriptome sequencing (RNAseq)
- miRNA sequencing

RNA

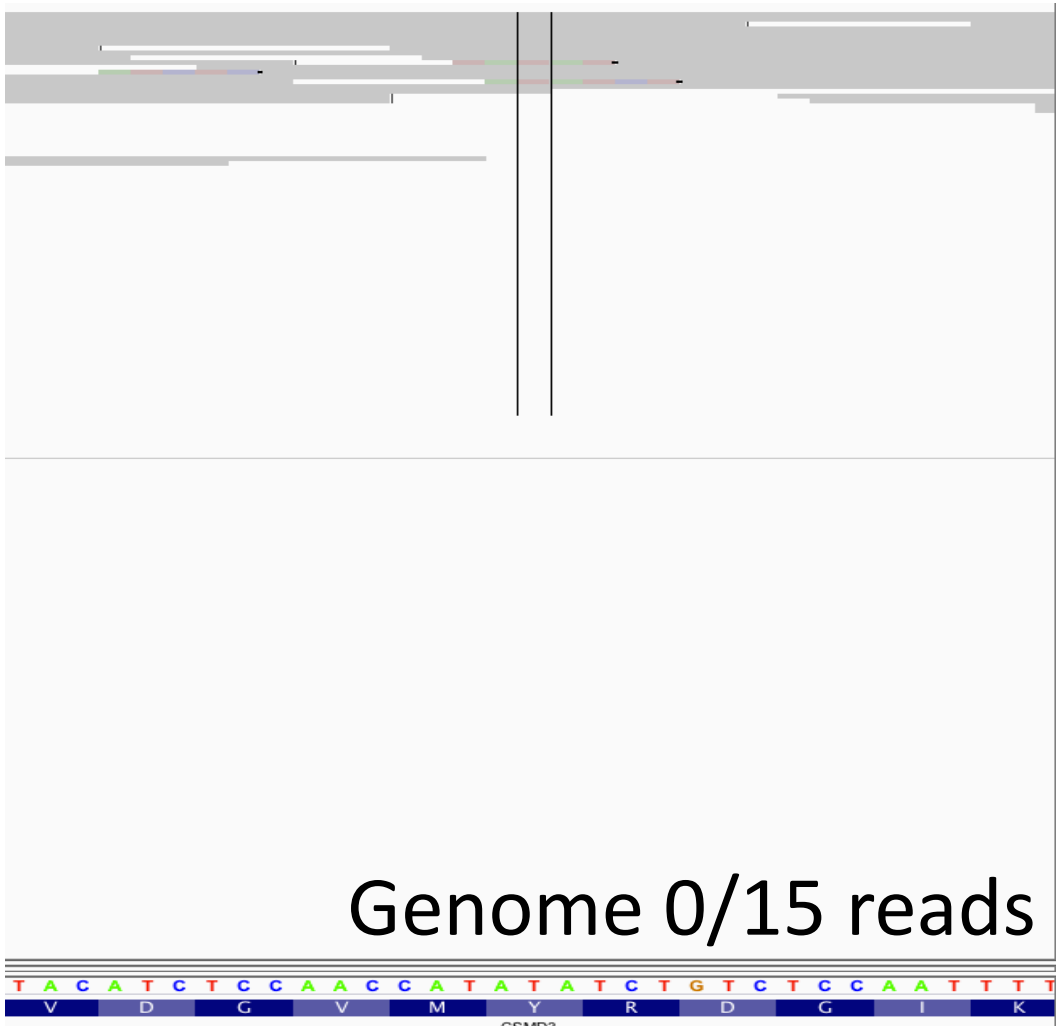
Somatic Mutations

- Tumors are a mix of cancer and “normal” cells.

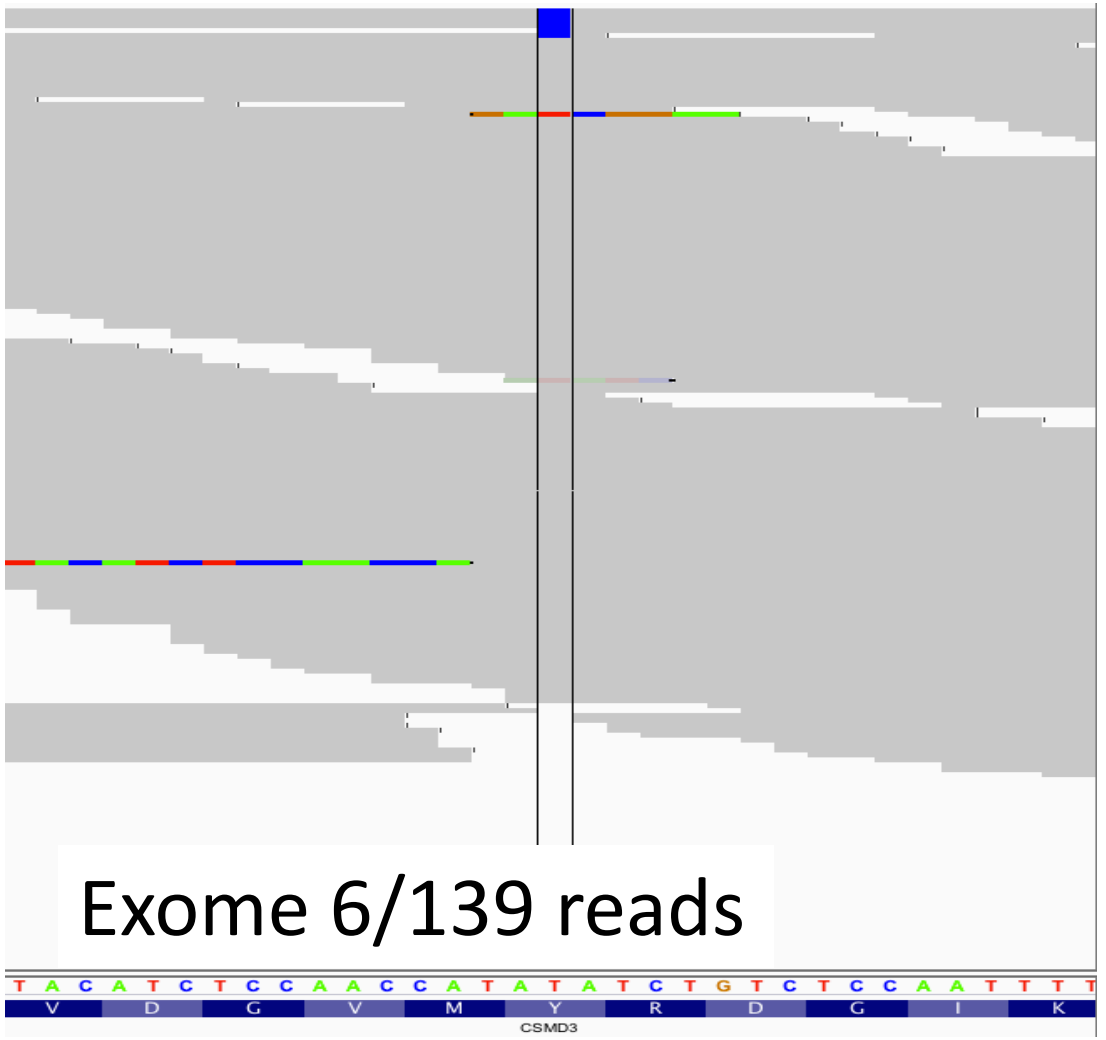
- Identifying somatic variants within a **tumor** sample can be accomplished by performing whole-genome **sequencing** (WGS) of DNA extracted from a **cancer** sample and DNA from a matching **normal** sample. **Cancer**-specific variants are those observed in the **tumor** sample but absent from the **normal** sample.

UCDAVIS Bioinformatics Core

Deep coverage is necessary to detect mutations in low purity



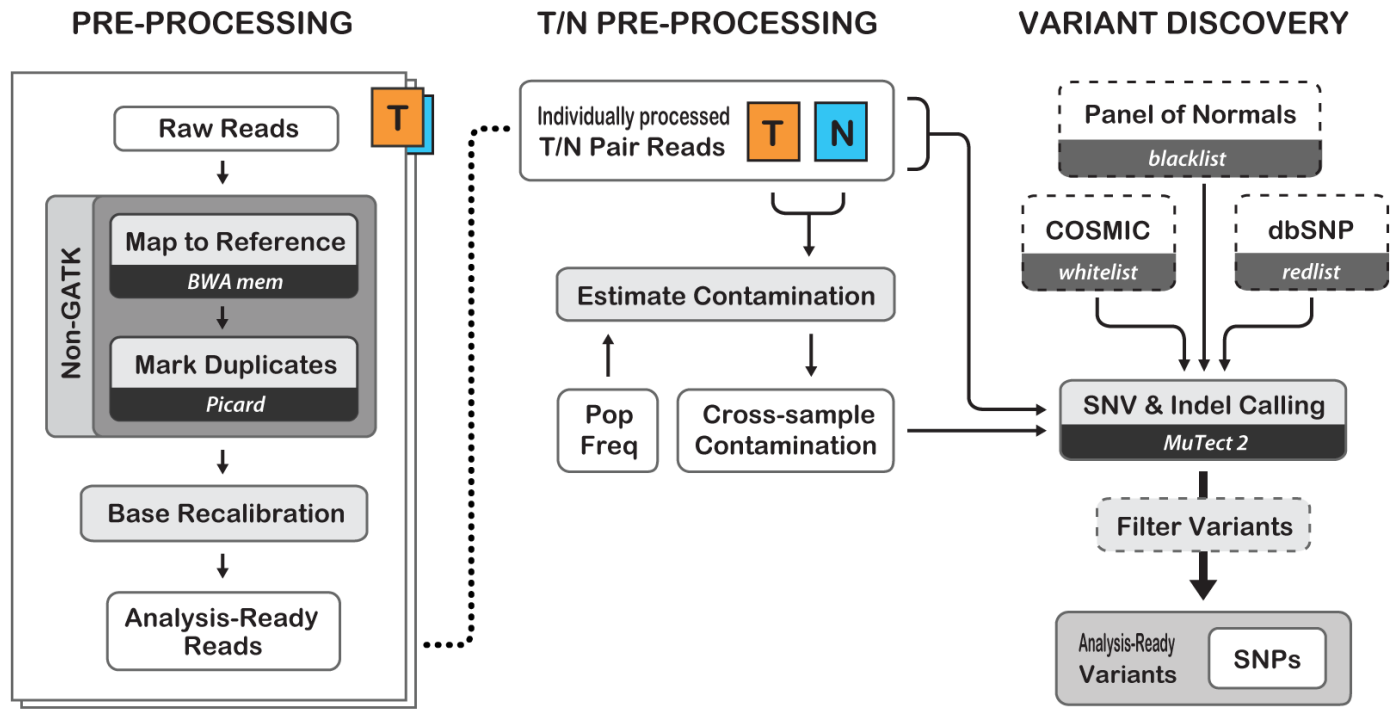
Genome 0/15 reads



Exome 6/139 reads

Cancer Variant Analysis

- Broad has MUTECT2
- https://software.broadinstitute.org/gatk/best-practices/



Best Practices for Somatic SNVs and Indels in Whole Genomes and Exomes - BETA

Cancer Variant Analysis

- Freebayes

freebayes -f reference.fasta --pooled-continuous --pooled-discrete -F 0.03 -C 2 tumor.bam normal.bam >out.vcf

--pooled-continuous outputs all alleles which pass filters regardless of genotype outcome

--pooled-discrete is a recommended flags for somatic variant calling

-F is the lowest fraction of observations supporting an alternative allele within an individual to evaluate the position

-C is the lowest number of reads supporting an alternative allele in an individual in order to evaluate the position.