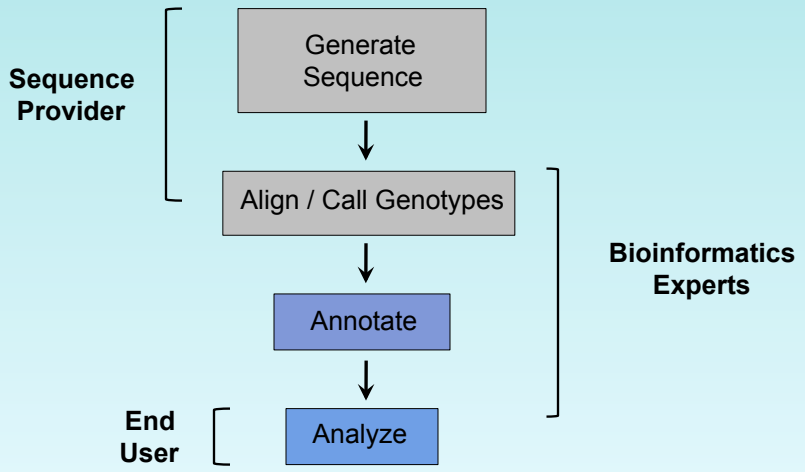
# Variant Annotation and Viewing Exome Sequencing Data

Jamie K. Teer
Exomes 101
9/28/2011

## Data Workflow



## General Considerations

**Where are the reads aligned?** Viewing alignments
**What is the effect?** Annotation / Consequence
**Who else has the variant?** Variation Databases
**How can I do all this?** Pipeline software

**How can I identify the important variants???**
Working with VarSifter

## General Considerations

**Where are the reads aligned?** Viewing alignments
**What is the effect?** Annotation / Consequence
**Who else has the variant?** Variation Databases
**How can I do all this?** Pipeline software

**How can I identify the important variants???**
Working with VarSifter

Easier to use
Less experience
Graphical

More challenging
More experience
Command-line

# WHERE

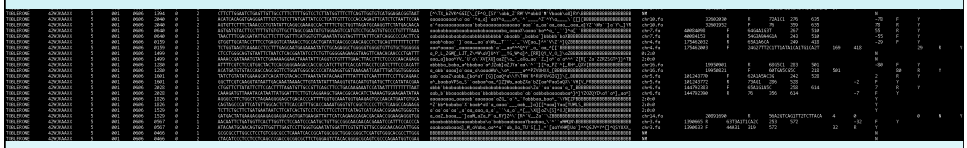
## are the reads?

Are they aligned correctly?

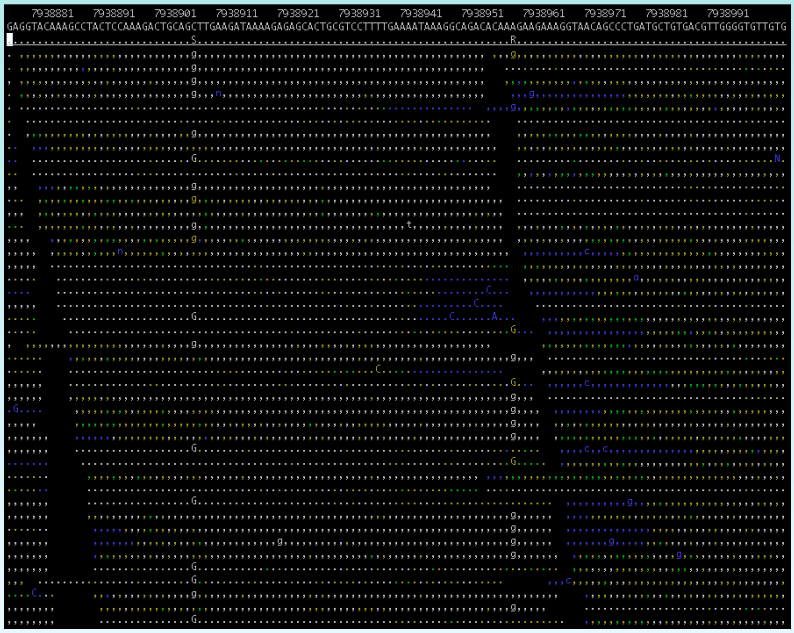
What do the alignments look like?

### File Formats

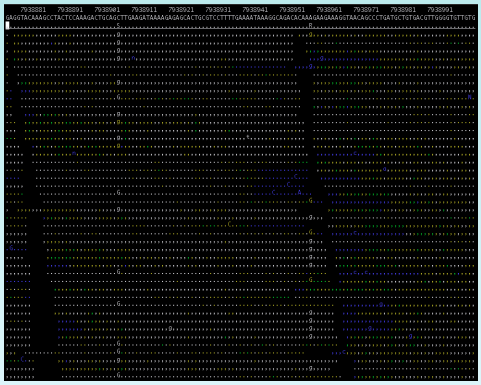
- **SAM/BAM** – Alignments
  - 20-100+ million lines per sample
  - viewing and manipulation programs:
    samtools (C), Picard (Java), Bio-SamTools (Perl), Pysam (Python)



## Analysis and Visualization – Samtools tview



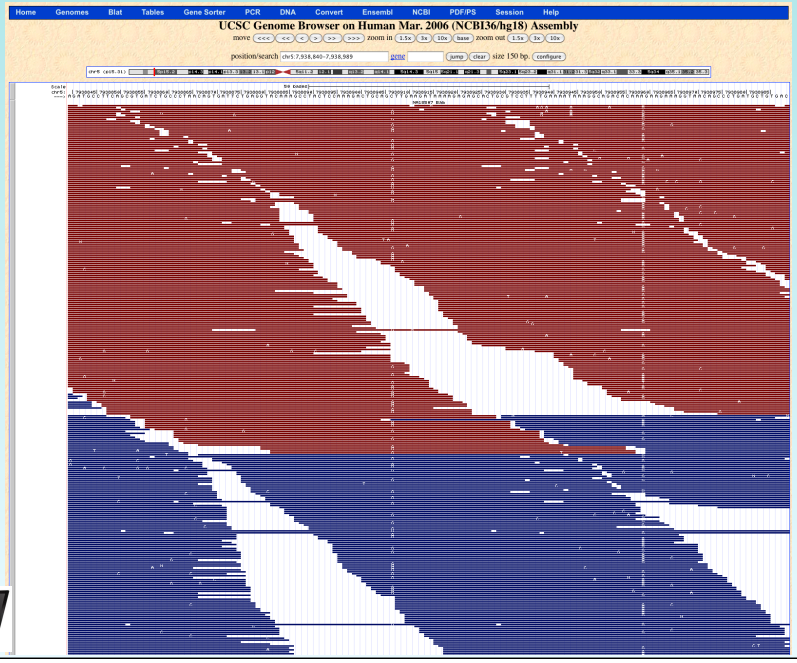

## Analysis and Visualization – Samtools tview



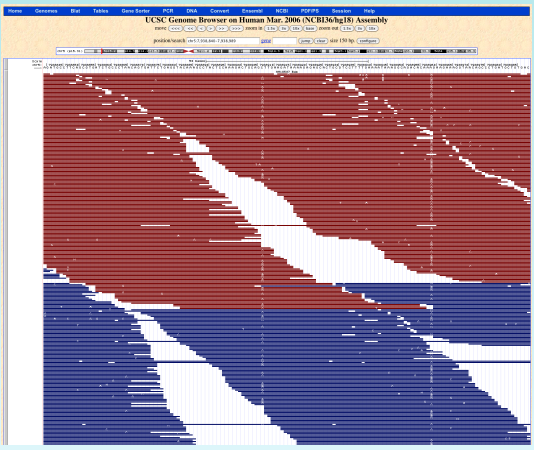
- FAST!
- Text-based
- Basic functionality

Li, H *et.al.* The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 2009

## Analysis and Visualization – UCSC Genome Browser

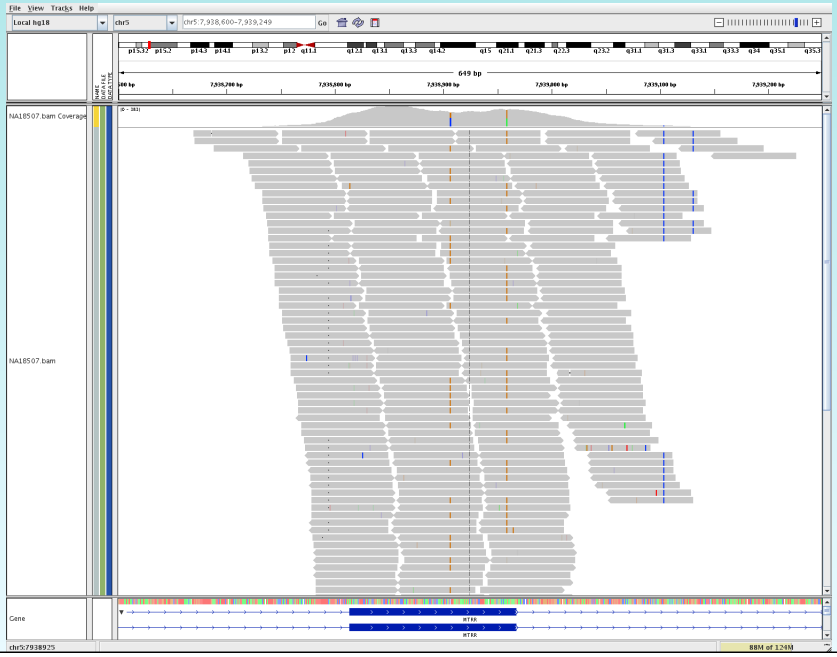



## Analysis and Visualization – UCSC Genome Browser

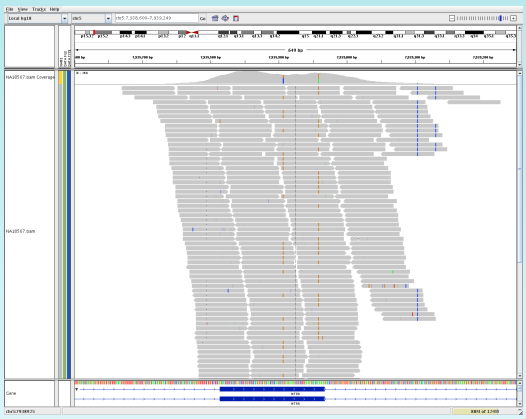



- View with UCSC tracks
- Need public facing server to hold data
- Limited viewing options

## Analysis and Visualization – IGV




## Analysis and Visualization – IGV



- Zooming
- Highlight reads to get more info
- Many features
- Web launcher

Robinson, JT *et.al*. Integrative Genomics Viewer. Nature Biotechnology, 2011

# WHAT

## is the effect?

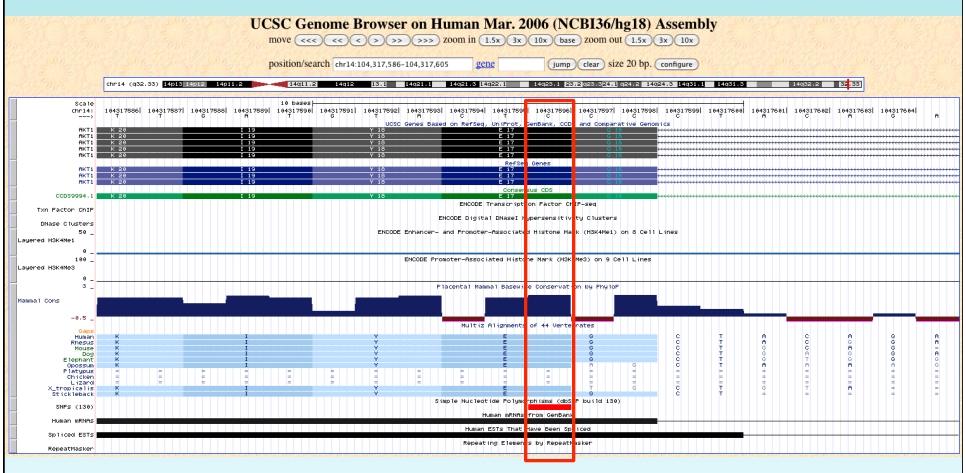
In a gene?

Coding?

Amino-acid change?

Detrimental for function?

## Annotation



## Annotation Software

Goal: Determine variant context

- **ANNOVAR** – Kai Wang et.al. Children's Hospital of Philadelphia
  - exonic splicing, HGVS format, distance to nearest gene, indels
  - local scripts using local data downloaded from UCSC Genome Browser
- **PIANNO / CDPred** – Praveen Cherukuri, NHGRI
  - Conserved Domain Prediction, dbSNP, indels
  - local scripts using UCSC Genome Browser SQL server
- **SeattleSeq Annotation** – Deborah Nickerson, U.Wash
  - conservation, HapMap freq, PolyPhen, clinical assoc., limited indels
  - external server
- **SNPeff** – Pablo Cingolani
  - integration with GATK and Galaxy, can read and write VCF*
  - local Java program using local data files

*VCF = Variant Call Format (1000 Genomes)

## Variant Consequence

Goal: How detrimental is a variant (AA change)

- **SIFT** - JCVI
  - uses PSI-BLAST to assay degree of conservation
- **Polyphen-2** – Ivan Adzhubel et.al. Harvard Med.
  - uses sequence features, homologue conservation, structural features (more with known structure)
- **CDPred** – Praveen Cherukuri, NHGRI
  - uses Conserved Domains database

- **Human Gene Mutation Database (HGMD)** – Cardiff U.
  - curation of literature, locus-specific databases
  - subscription-based, flat file available
  - all NIH license:

http://nihlibrary.nih.gov/ResearchTools/Pages/bioanalysis.aspx

## Annotation / Consequence at NISC

| ANNOVAR | | | CDPred | HGMD | |
|---|---|---|---|---|---|
| type | Gene_name | consequence | CDPred_score | HGMDdisease | HGMDtags |
| intronic | CFTR | - | - | - | - |
| intronic | CFTR | - | - | - | - |
| intronic | CFTR | - | - | - | - |
| synonymous_SNV | CFTR | CFTR:uc003vjd.1:exon20:c.3285A>T:p.T1095T, | 0 | - | - |
| nonsynonymous_SNV | CFTR | CFTR:uc003vjd.1:exon20:c.3302T>A:p.M1101K, | -4 | Cystic fibrosis,Cystic fibrosis | DM,DM |
| intronic | CFTR | - | - | - | - |
| intronic | CFTR | - | - | - | - |
| synonymous_SNV | CFTR | CFTR:uc003vjd.1:exon23:c.3870A>G:p.P1290P, | 0 | - | - |
| synonymous_SNV | CFTR | CFTR:uc003vjd.1:exon27:c.4272C>T:p.Y1424Y, | 0 | - | - |
| nonsynonymous_SNV | CFTR | CFTR:uc003vjd.1:exon27:c.4357C>T:p.R1453W, | -9 | Cystic fibrosis | DM |
| synonymous_SNV | CFTR | CFTR:uc003vjd.1:exon27:c.4389G>A:p.Q1463Q, | 0 | - | - |
| nonsynonymous_SNV | CFTR;CFTR | CFTR:uc003vjd.1:exon10:c.1392G>T:p.K464N, | -11 | Cystic fibrosis | DM |
| synonymous_SNV | CFTR;CFTR | CFTR:uc003vjd.1:exon11:c.1584G>A:p.E528E, | 0 | - | - |
| intronic | CGA | - | - | - | - |
| intronic | CGA | - | - | - | - |
| intronic | CGA | - | - | - | - |
| intronic | CGA | - | - | - | - |
| intronic | CGA | - | - | - | - |
| intronic | CGA | - | - | - | - |
| intergenic | CGA(dist=26577),DKFZp6... | - | - | - | - |
| intergenic | CGA(dist=26742),DKFZp6... | - | - | - | - |
| intergenic | CGA(dist=26748),DKFZp6... | - | - | - | - |
| intergenic | CGA(dist=53714),DKFZp6... | - | - | - | - |
| intergenic | CGA(dist=5796),DKFZp68... | - | - | - | - |
| intronic | CGB | - | - | - | - |
| intronic | CGB | - | - | - | - |
| intronic | CGB | - | - | - | - |
| intronic | CGB | - | - | - | - |
| intronic | CGB | - | - | - | - |

# WHO

else has the variant?

Is it a common variant?

Is it seen in certain populations?

Has it been observed in a disease cohort?

## Human Variation Databases

Goal: Determine presence of variant in others

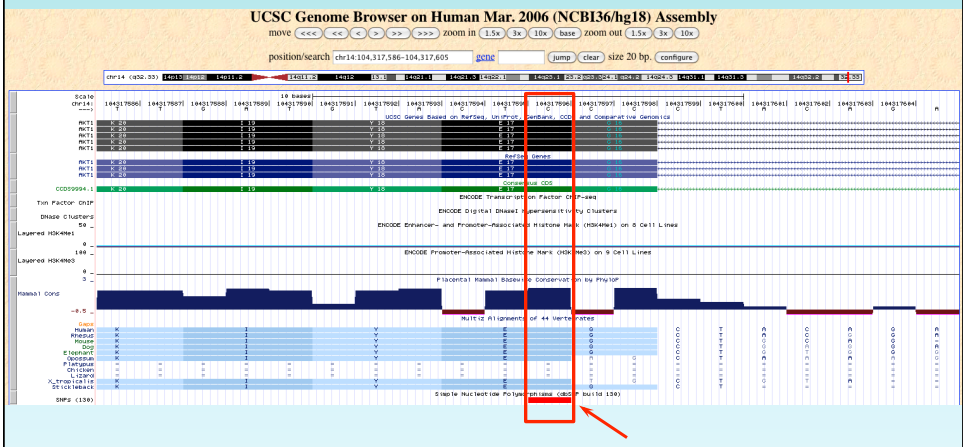
**dbSNP**
- Includes everything!!
- SNPs have information about origin
- VCF available

**1000 Genomes**
- 1,094 low coverage genomes (~4x)
- >=1% sensitivity
- In dbSNP, VCF available
- 822 Exomes – CCDS (coming soon)

**ClinSeq**
- 650 Exomes with phen. (dbGaP)
- CCDS and knownGenes
- soon in dbSNP, VCF to be available

**NHLBI Exome Sequencing Project**
- 2,500 Exomes with phen. (dbGaP)
- in dbSNP, VCF available

## dbSNP – Caution!




**dbSNP is not exclusively common polymorphisms!!!**

# HOW

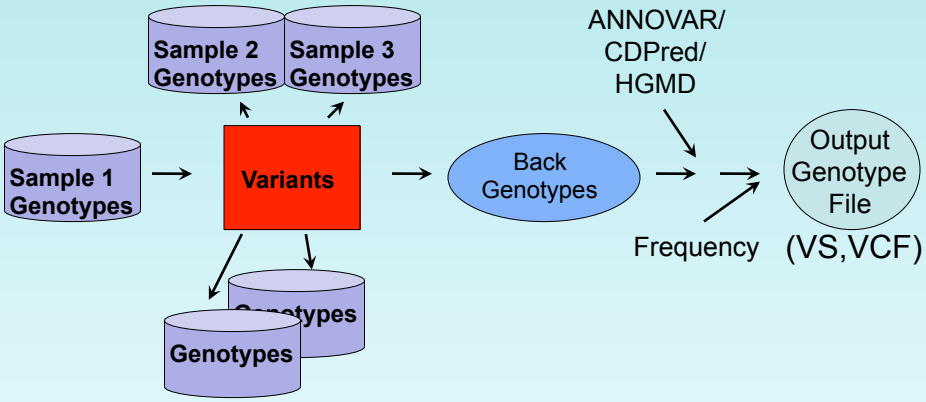
can I do all this?

How are these programs run?

Can they be run automatically?

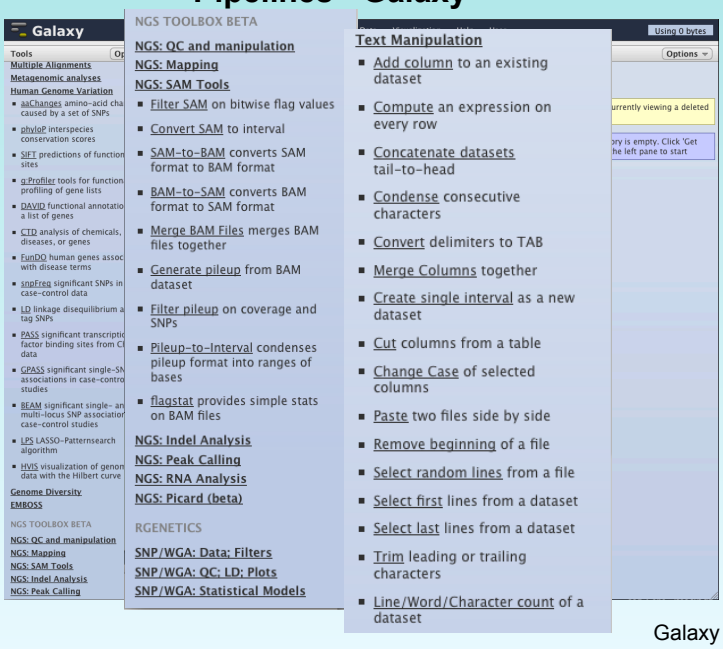
Is there a graphical interface?

## NISC Pipeline



Using SGE, Perl, make, cron, databases

## Pipelines - Galaxy



Galaxy Team

# HOW

## can I identify the important variants?

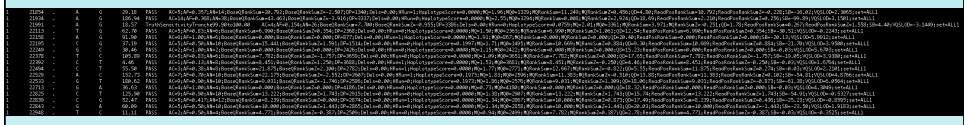
*Ab initio*?

User-guided?

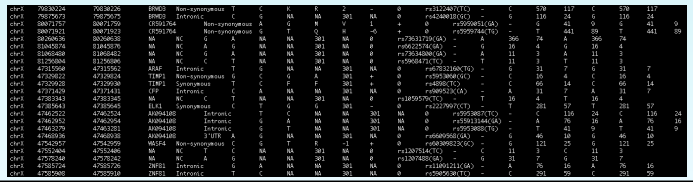
Easy to use?

## Variant File Formats

- **VCF** – genotypes (100,000+)
  - BGZIP indexing using Tabix (samtools)
  - viewing and manipulation with VCFtools



- **Structured Text** – genotypes (100,000+)
  - Header line
    - Annotation, sample names
  - Certain annotations handled specially



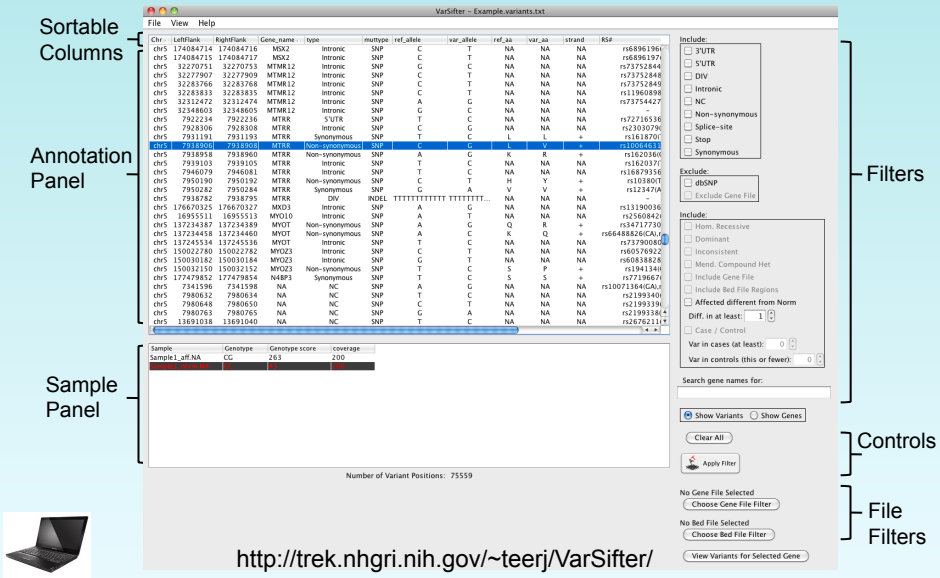
## Variant Prioritization

Goal: Identify most interesting variants *ab initio*

- **VAAST** – Mark Yandell et.al., Univ. Utah, Omicia
  - prioritize variants using a probabilistic approach
  - uses AA substitution, aggregation, inheritance
  - free for academic research use
- **VarMD** – Murat Sincan et.al., NHGRI
  - prioritize variants using inheritance
  - available on Helix / Galaxy Dev

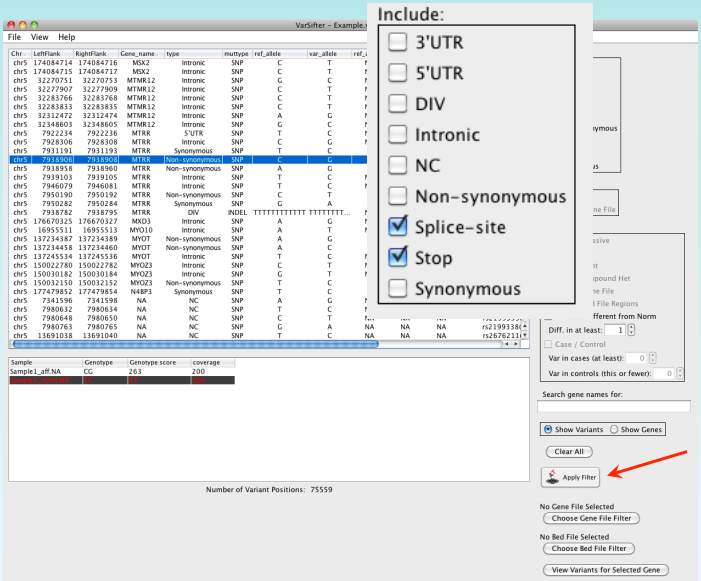
## VarSifter – viewing, sorting, and filtering variants

Goal: Hands-on analysis

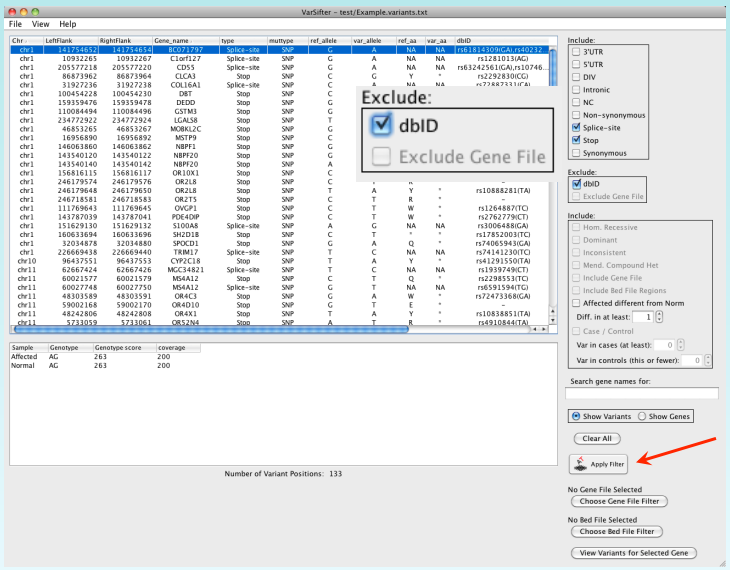



http://trek.nhgri.nih.gov/~teerj/VarSifter/

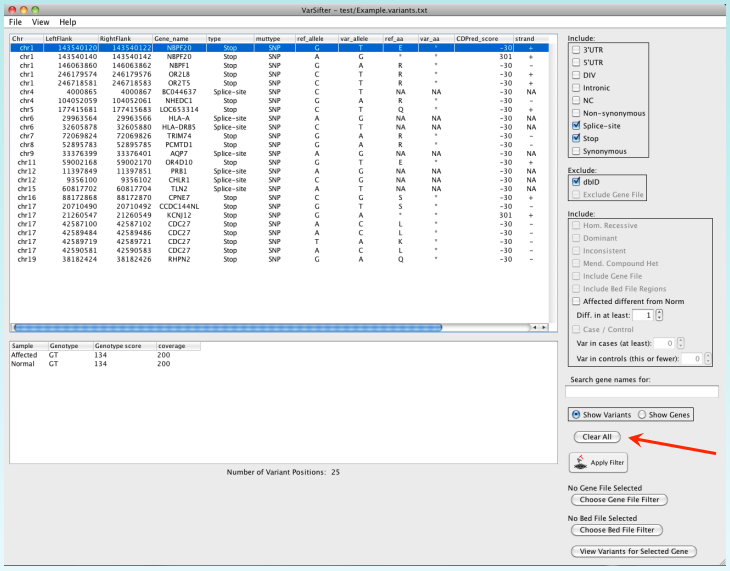
## Filtering – Mutation Type



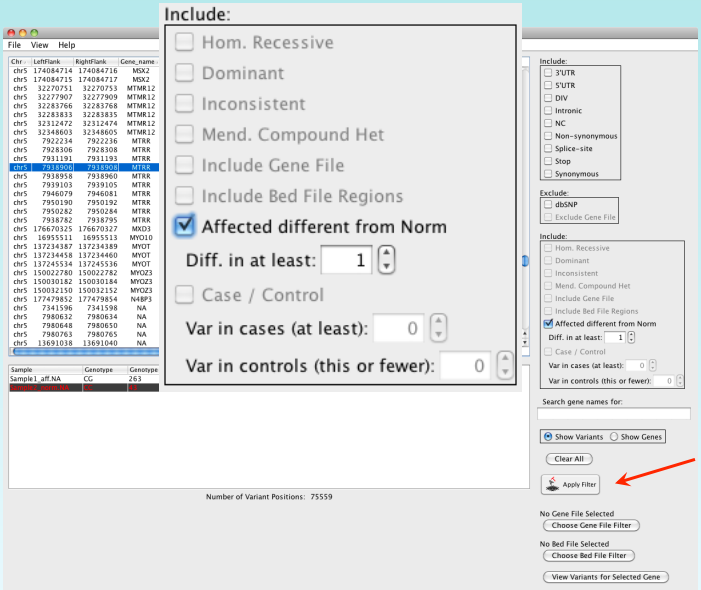
## Filtering – Mutation Type



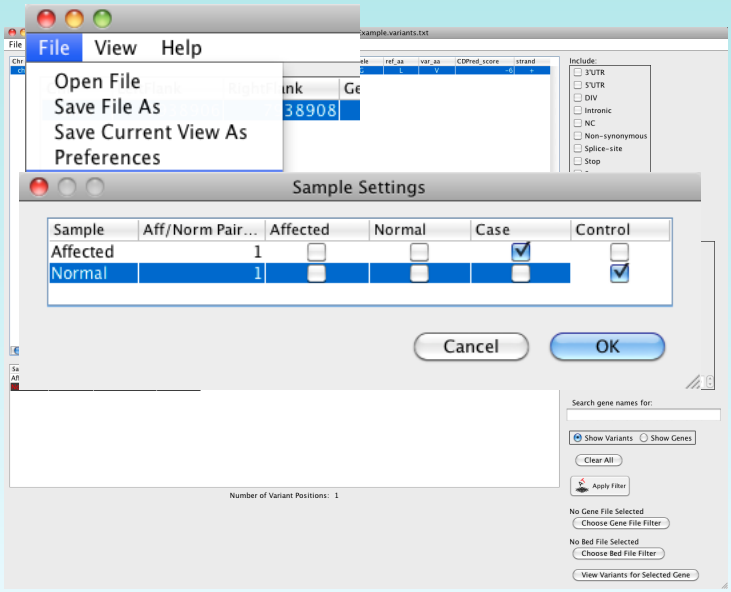

## Filtering – database



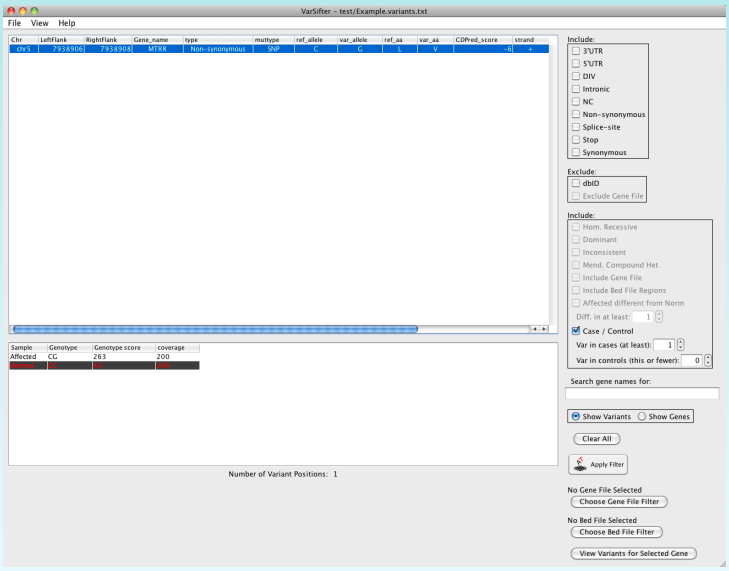
## Filtering – Affected/Normal



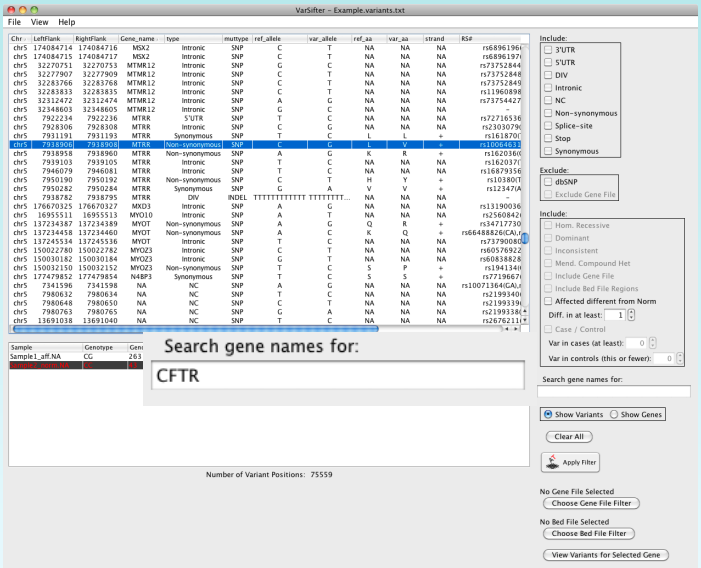

## Filtering – Affected/Normal



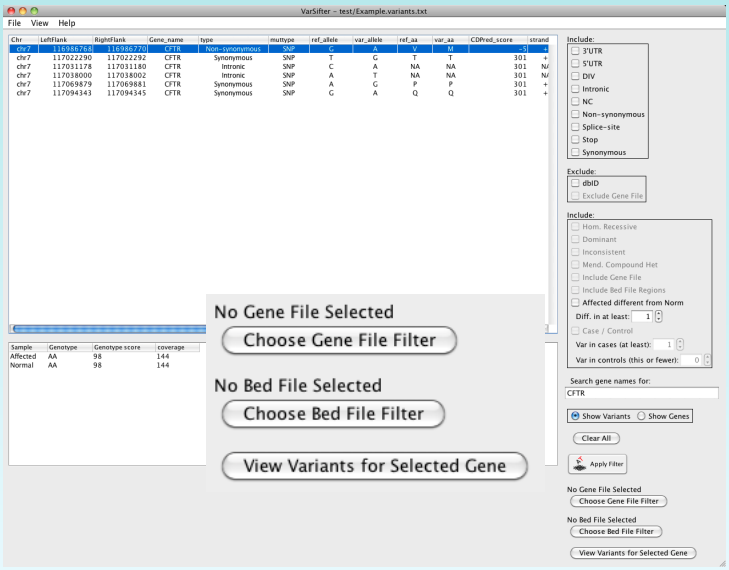
## Filtering – Affected/Normal



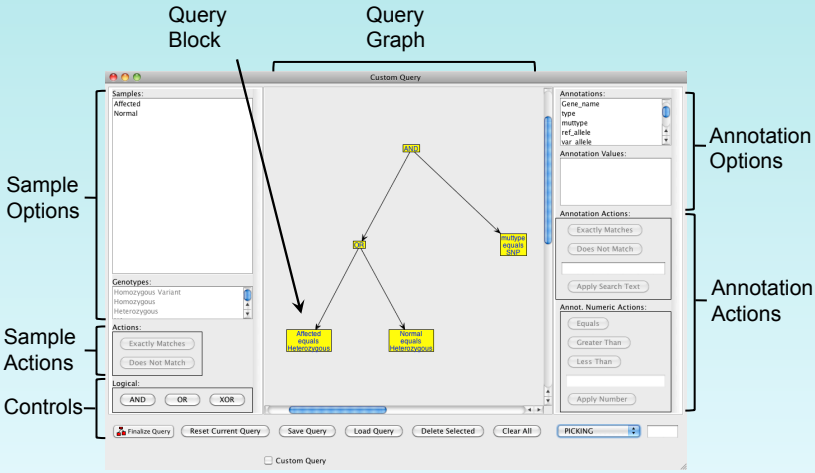

## Filtering – Gene Name



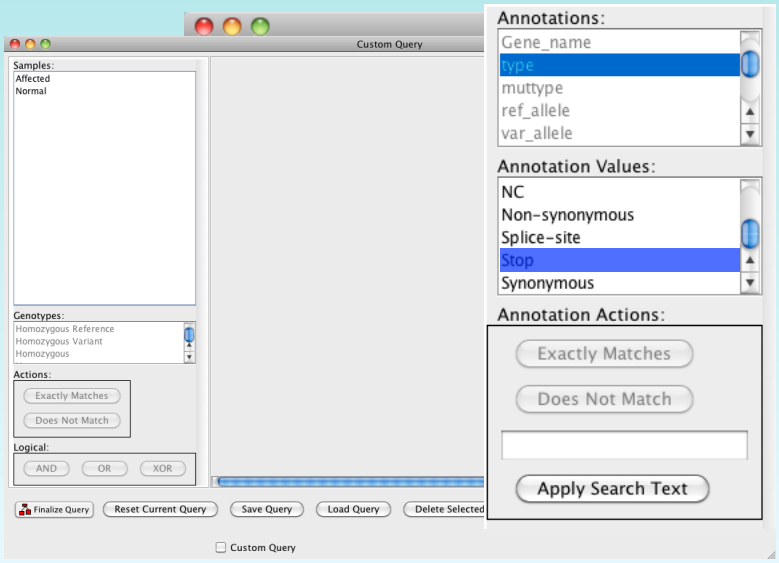
## Filtering – Gene Name



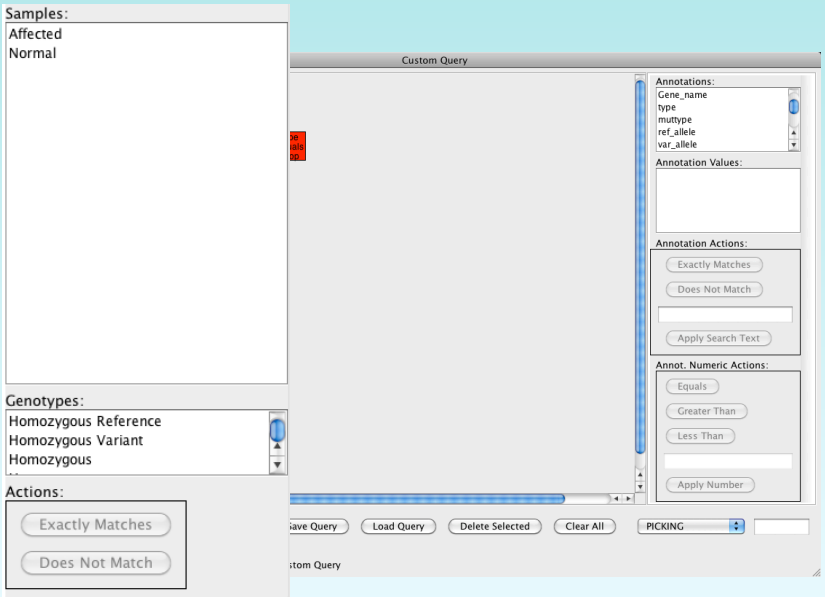

## Custom Query Filters



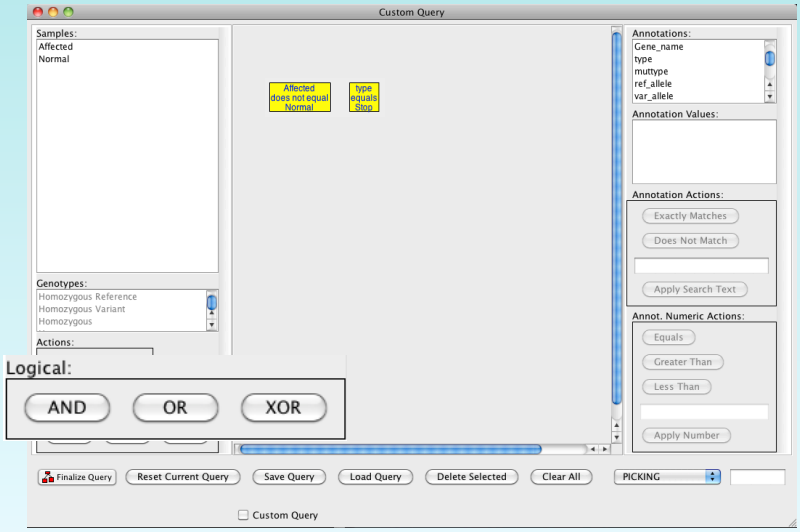
## Custom Query Filters - Annotation



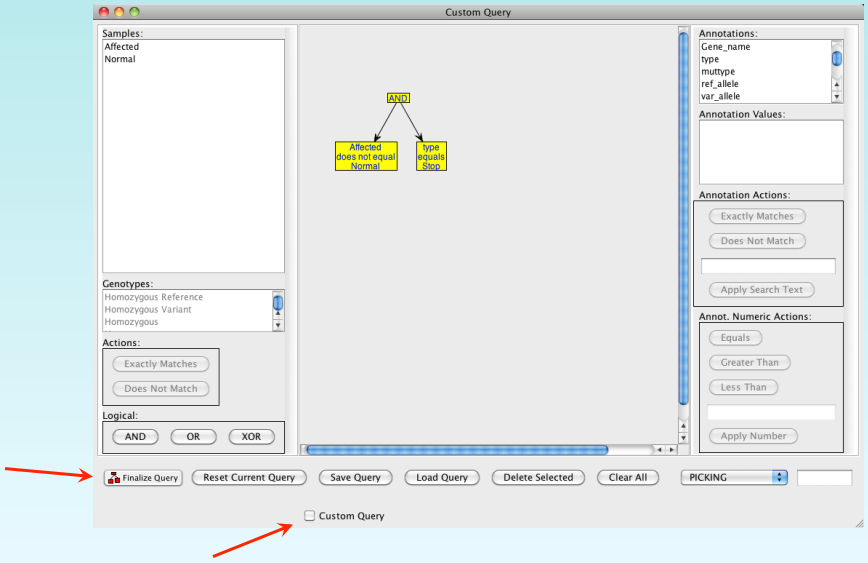

## Custom Query Filters - Sample



## Custom Query Filters - Linking




## Custom Query Filters - Linking



## Summary

- Annotation gives context
- Consequence prediction can guide analysis
- Varying experience required
- Prioritization tools return “black box” answer
- Visualization can allow guided, informed analysis
- VarSifter is a powerful tool for “hands-on” analysis


## Acknowledgements

**NIH Intramural Sequencing Center**

- **Sequencing Operations**
  - Bob Blakesley
  - Alice Young
  - Lab Staff
- **Bioinformatics**
  - Meg Vemulapalli
- **IT Linux Support**
  - Jesse Becker
  - Matt Lesko
- **Mullikin Lab**
  - Nancy Hansen
  - Pedro Cruz
  - Praveen Cherukuri
- **Biesecker Lab**
  - Jennifer Johnston
  - Marjorie Lindhurst
  - David Ng
  - Steve Gonsalves
- **UDP**
  - David Adams
  - Thomas Markello
  - Karin Feuntes Fajardo
  - Murat Sincan
- Yardena Samuels
- Peter Chines

## Links

### File Formats
- SAM/BAM: http://samtools.sourceforge.net/
- VCF: http://www.1000genomes.org/wiki/analysis/vcf4.0

### Viewers
- samtools: http://samtools.sourceforge.net/
- UCSC browser: http://genome.ucsc.edu/
- IGV: http://www.broadinstitute.org/igv/

### Annotation
- ANNOVAR: http://www.openbioinformatics.org/annovar/
- SeattleSeq Ann.: http://gvs.gs.washington.edu/SeattleSeqAnnotation/
- SNPeff: http://snpeff.sourceforge.net/

## Links – cont'd

### Variant Consequence
- SIFT: http://sift.jcvi.org/
- Polyphen-2: http://genetics.bwh.harvard.edu/pph2/
- CDPred: http://research.nhgri.nih.gov/software/CDPred/

### Variant Prioritization
- VAAST: http://www.yandell-lab.org/software/vaast.html
- VarMD: (Dev section on Helix; http://helix.nih.gov/)
- HGMD: http://nihlibrary.nih.gov/ResearchTools/Pages/bioanalysis.aspx
  http://www.hgmd.org/

http://www.biobase-international.com/product/human-gene-mutation-database

### Pipeline
- Galaxy: http://main.g2.bx.psu.edu/
- SVA: http://www.svaproject.org/

## Links – cont'd

**Variation databases**

- dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/
- ClinSeq
- 1000 Genomes: http://www.1000genomes.org/
- NHLBI Exome Seq.: http://snp.gs.washington.edu/EVS/

**VarSifter**

- http://trek.nhgri.nih.gov/~teerj/VarSifter/