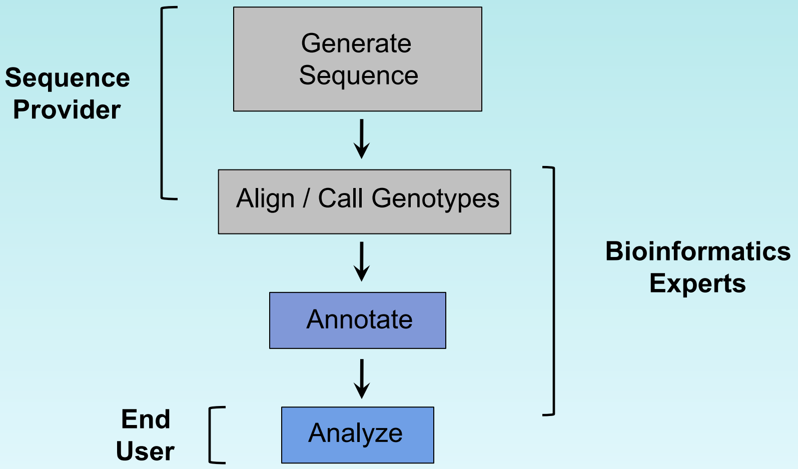
Variant Annotation
and
Viewing Exome Sequencing Data

Jamie K. Teer
Exomes 101
9/28/2011

Data Workflow



Sequence Provider

Generate Sequence

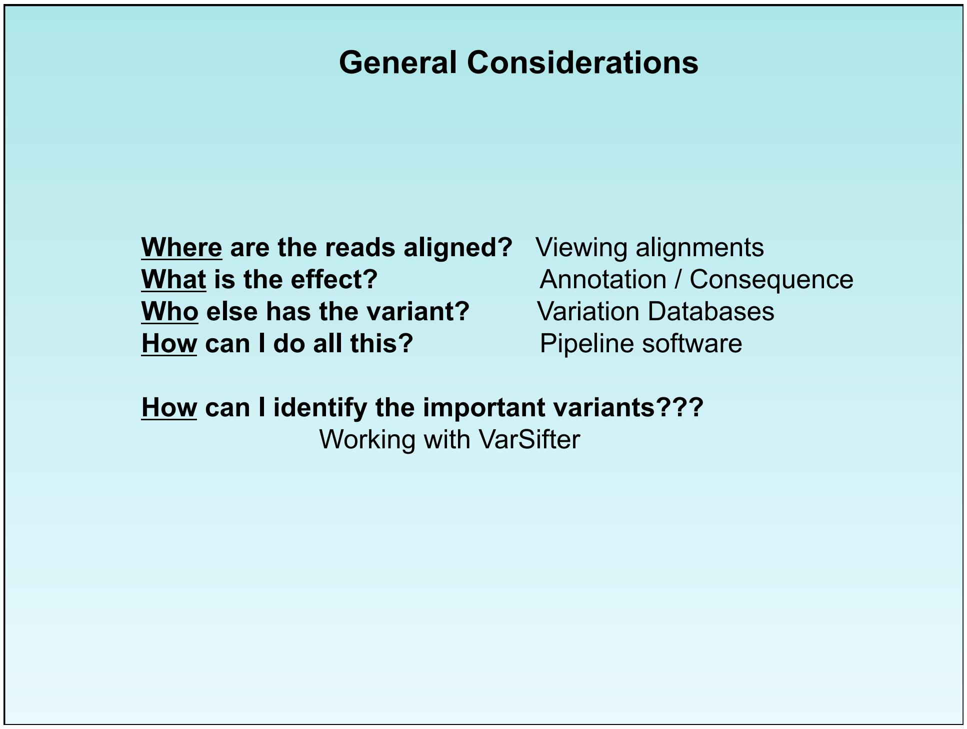
Align / Call Genotypes

Annotate

Analyze

Bioinformatics Experts

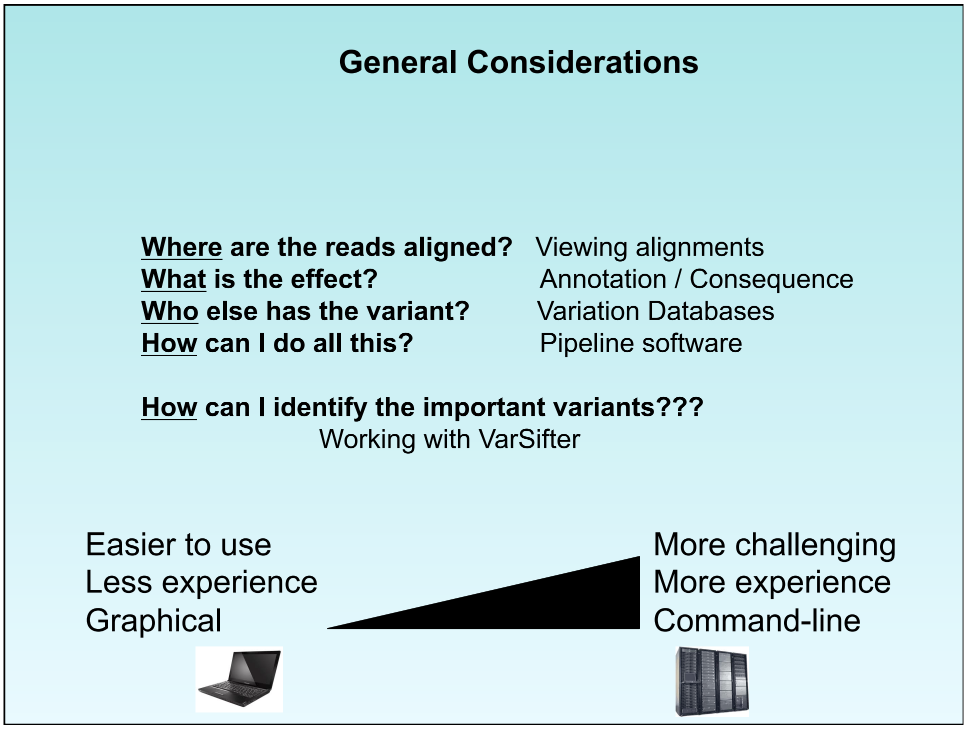
End User



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	More challenging
Less experience	More experience
Graphical	Command-line

WHERE

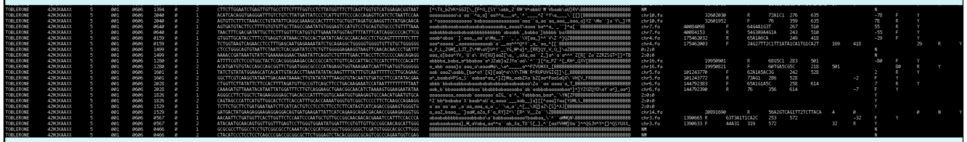
are the reads?

Are they aligned correctly?

What do the alignments look like?

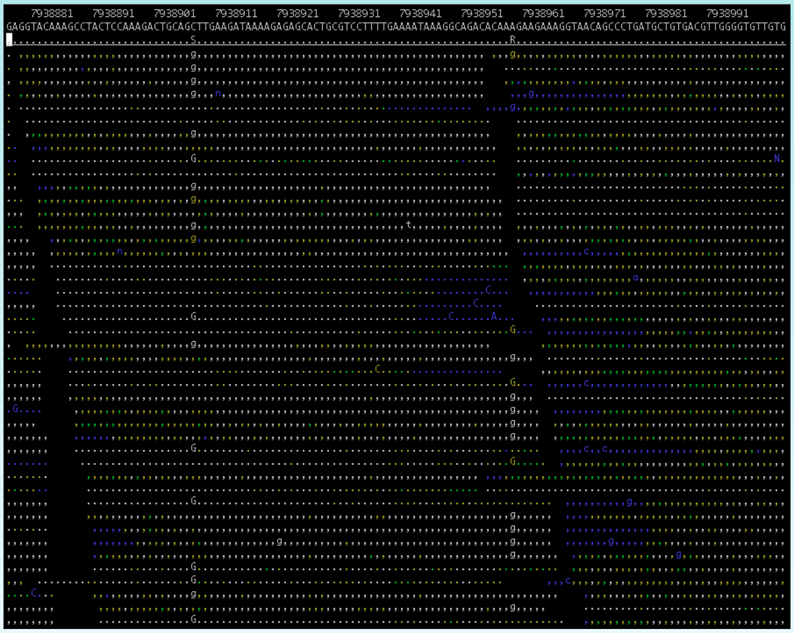
File Formats

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



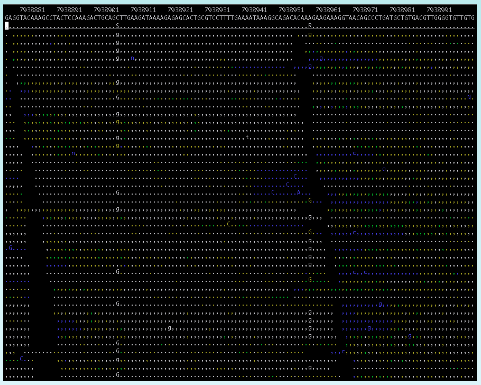
```
TOBLERONE  42WJKAAXX  5  001  0606  1394  0  2  CTTCTTGGAATCTGAGTTGTTGCCTTTCTTTTGGTCCTCTTATGGTTTCTTCAGTTGGTGTCATGGAGACGGTAAT  [^\TX_bZVK^GGI[\_[F^O_[SY`\abb_Z`RW`V^abbU`M`Vbaab\aU]RV\BBBBBBBBBBBBBBBBBBBB  NM                                                        N
TOBLERONE  42WJKAAXX  5  001  0606  2040  0  1  ACATCACAGGTGAGGGATTTGTCTGTCTTATGATTATCCCTCATTGTTTCCCACCAGAGTTCATCTCTAATTCCAA  aaaaaaaaaa\a`aa``^a_a]`aaY^a___a^_`^`____^Z`^Y\a____\`[[][BBBBBBBBBBBBBBBBBBB  chr10.fa     32602030    R   72A1C1  276  635               -78  F  Y
TOBLERONE  42WJKAAXX  5  001  0606  2040  0  2  AGTGTTCTTTCTAAACCCTGTATATTCAGGCGAAAGCCACTTTTCTGCTGGTTAGATGCAAGGTTCTATGACAACA  a`^aaaaaaaaaaaaa`babaaaaaaaaaaaaaa`aaa``a_aa`aa_aaa__aaa_a]YZ`\Wa``]a`Y\_]YR  chr10.fa     32601952    F   76  359  635                  78   R  Y
TOBLERONE  42WJKAAXX  5  001  0606  0640  0  1  AGTGATGTACTTCCTTTTGTGTGTTGCTTAGCCGAATATGTGGGAGTCCATGTCCTGCAGTGTGCCCTGTTTTAAA  aaababbaaabaaabaaababaaaaaba_aaaaX\aaaa`ba^^a_`_``]^a[``BBBBBBBBBBBBBBBBBBBB  chr7.fa   40084098   F   64G4A1G3T  267  510              55   R  Y
TOBLERONE  42WJKAAXX  5  001  0606  0640  0  2  TAACTTTCGACGATATTGCTTCTTTGGTTTCATGGTGTTGAAATATGGTAGTTTTATTTCATCAGGCCCCACTTCG  aababbbabaababaabbbbbbbbbbb`abaabb`_babba`]bbbbb`baXBBBBBBBBBBBBBBBBBBBBBBBB  chr7.fa   40084153   R   54G3A9A4G1A  243  510            -55  F  Y
TOBLERONE  42WJKAAXX  5  001  0606  0159  0  1  GTGGTTGCATACCTTTCCTGAGGTCATAAACCTGCCACTGATATCAACGCCAACAGCCTCTGCAGTTTTTTTTCTTT  aaab^abaa``]`aaa__aa`a\Wa__T```_`_`\V[aa_]^^`V\Z`^]QZBBBBBBBBBBBBBBBBBBBBBBB  chr4.fa   175462032  R   65A1A6CA  249  418              -29  F  Y
TOBLERONE  42WJKAAXX  5  001  0606  0159  0  2  TCTGGTAAGTCAGAACCTCCTTTGAGCAATGAGAAAATATCTGCAGAGGCTGGGGGTGGGGTGTTGTGCTGGGGGG  aaa^aaaaa`_aaaaaaaaaaaab`a`__aa^^^Q^Y`_a_`aa_^[[`BBBBBBBBBBBBBBBBBBBBBBBBBBB  chr4.fa   175462003  F   24G27TT2C1TT1ATA1CA1TG1CA2T  169  418     29  R  Y
TOBLERONE  42WJKAAXX  5  001  0606  1499  0  1  CTCCTGGGCAGTGTAATTCTAATCTCACGGATATCCTCTGTTGGGGGAGAAGGTAAGTTCAACACAACCCTGATTT  a_P_L_ZGW[_LJT_Z\^W\aV]U^Y`__YG_W^a]^_[RR]QY_V_O_]\aZBBBBBBBBBBBBBBBBBBBBBBB  0:2:0                                                     N
TOBLERONE  42WJKAAXX  5  001  0606  1499  0  2  AAAACCGATAAATGTATCTGAAAAAGGAACTAAATATTCAGGTCTGTTTTGAACTTACCTTCTCCCCCAACAGAGG  aaa_a]baa^YV_`U`a\`XV[XU]aaZ[\a_`_aXa_aa``Z_]a^`a`a^^^`Z[R[`Za`ZZRZSGT^]]^TB  2:0:0                                                     N
TOBLERONE  42WJKAAXX  5  001  0606  1280  0  1  ATTTTCGTCTCCGTGGCTACTCCACGGGGAAGACCACCGCCATCTTGTTCACCATTACCTCCATCTTTCCCACATT  abbbba_baba_a^bbabaa`a^JZab]aZJYa`aa\`^``][^a_PZ`^[_RH^_QXV[BBBBBBBBBBBBBBB  chr16.fa     19950901    R   68G5C1  283  501               -80  F  Y
TOBLERONE  42WJKAAXX  5  001  0606  1280  0  2  ACATGATGTGTACCAGCCAGCGGTTCTGGATGGGCGCCCCATAGAGGTGGTAAAGAATCAATTTGATGGTGGGGGG  a_abb`aaaa]a`aaa_a\aaaaWa\_\a^_____a^^PZVUHXX_[BBBBBBBBBBBBBBBBBBBBBBBBBBBB  chr16.fa     19950821    F   60TGA5CG5C  218  501            80   R  Y
TOBLERONE  42WJKAAXX  5  001  0606  1601  0  1  TATCTGTATATGGAAGCATCACATTGTACACCTTAAATATATACAAGTTTTATTTGTCAATTTTTCCTTGCAGAAC  aab`aaaZ\aabb_[ba^aY`[G][aaQ^a\\Y\THN`R^RUPUVGIG]]\[_BBBBBBBBBBBBBBBBBBBBBB  chr5.fa   101243770  F   62A1A5AC3G  242  528             2    R  Y
TOBLERONE  42WJKAAXX  5  001  0606  1601  0  2  GGCTTCGTCAAGGTATAATTGACAAATAAAACTTGTATATATTTAAGGTGTACAATGTGATGCTTCCATATACGAA  a^_baabaVPSa_S``aabaa^aa_^]Z]Wa_aabZXa`bZ[aa^PaaSaQX\`VH[V_FRBBBBBBBBBBBBBBB  chr5.fa   101243772  R   73AG1  286  528            -2   F  Y
TOBLERONE  42WJKAAXX  5  001  0606  2028  0  1  CTGGTTCTTATATTCTTCCACTTTTAGATGTTGCCGTTCAGCTTCCTGACAAGAAATCCATAATTTTTTTTTTAAT  abbb`bbabaabbaabaaababababbaabaabaabaa\Za``aa`aaaa`a_T_BBBBBBBBBBBBBBBBBBBB  chr6.fa   144792383  F   65A1G1A5C  258  614             7    R  Y
TOBLERONE  42WJKAAXX  5  001  0606  2028  0  2  CAAAGATGTTAAATACATAATTATGGATTTCTTGTCAGGAAGCTGAACGGCAACATCTAAAAGTGGAAGAATATAA  aab_b`bbaaaabbabbaa`bbabbbabaaaaba`ab`aabbabaaaabaa^]^]YZOZQYD\aY`a^]_aa^]  chr6.fa   144792390  R   76  356  614                 -7   F  Y
TOBLERONE  42WJKAAXX  5  001  0606  1826  0  1  AGGGCCTTCTGGCTCTAGAAGGGGAGCTGACACCCATTTTGGTGCAAATGGTGAAGAGTGCCAACATGAATGTGCA  aaaaaaaaa_aaaaab`aaaaaaa`aZG_`a`^_`Yabbbaa_baa^_`\YN[ZPBBBBBBBBBBBBBBBBBBBBB  2:0:0                                                     Y
TOBLERONE  42WJKAAXX  5  001  0606  1826  0  2  CAGTAGCCCATTCATGTTGGCACTCTTCACCATTTGCACCAAAATGGGTGTCGGCTCCCCTTCTCAAGCCAGAAGG  ^Z`bb^bababa`X`baab^aU`a_aaaa`___aab__]a][[^aaa]Yaa[YUWLS_BBBBBBBBBBBBBBBBB  2:0:0                                                     Y
TOBLERONE  42WJKAAXX  5  001  0606  1529  0  1  TCTTCTGCTTCTGATGAATAATCTTCATCACTGTCCTCCTCTTCCTCTTCATAGTCATCAGACCGGAAGTGGGGTG  a`aa`aa`aa`_a`aa_aaa_a_a`_``\a_a`_^[__\XQ]aZ\[S]^XJ[BBBBBBBBBBBBBBBBBBBBBBBB  2:0:0                                                     Y
TOBLERONE  42WJKAAXX  5  001  0606  1529  0  2  GATGACTATGAAGAGGAAGAGGAGGACAGTGATGAAGATTATTCATCAGAAGCAGACGACAACCGGAAGAGGGTGG  a_aaZ_baaa__`]aaW_aZa_F`a_RY]Z^\`[R^`V__Za``\ZBBBBBBBBBBBBBBBBBBBBBBBBBBBBB  chr14.fa     20891690    R   56A2GTCAG1TT2TCTTACA  4  0         0   N  Y
TOBLERONE  42WJKAAXX  5  001  0606  0567  0  1  AACAATTCTGATGGTTCACTTGGTTCTCCAATCCCAATGCTGTTGCCGGCAACAACACGAAATCCATTTCCACCCA  abaababbbbbaaaaabbaba\a`babbaaabaaaaYbaabaa_\`^``aWWQN\BBBBBBBBBBBBBBBBBBBB  chr3.fa   1390665 R   63T3A1T1CA2C  253  572              -32  F  Y
TOBLERONE  42WJKAAXX  5  001  0606  0567  0  2  ATACAATGCAACAGTGGTTGGTTTGAGTCCTTGGGTGGAATATGGATTTCGTGTTGTTGCCGGCAACAGCATTGGG  aaabaaabaaaa]_W_aVaba_aa^^a``ab_Xa_TU`S[_]_^`[aaYVHW]Ua`]^^QGJV^Y^[]^QSYUXX_  chr3.fa   1390633 F   44A31  319  572                32   R    Y
TOBLERONE  42WJKAAXX  5  001  0606  0466  0  1  GCGCGCCTTGGCCTCCTGTCGGCGCCTCAAATCACCGCATGGCGGCTGGGCGGGCTCGATGTGGGCACGCCTTGGG  BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB  NM                                                        N
TOBLERONE  42WJKAAXX  5  001  0606  0466  0  2  CTACATCCCTCCTCCTCAGCCCGACCGCGGCGCTTCTGGGAGTCTACACGGGGCGCAGTCGCCCAGAATGGTCGAG  BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB  NM                                                        N
```

Analysis and Visualization – Samtools tview



```
    7938881   7938891   7938901   7938911   7938921   7938931   7938941   7938951   7938961   7938971   7938981   7938991
GAGGTACAAAGCCTACTCCAAAGACTGCAGCTTGAAGATAAAAGAGAGCACTGCGTCCTTTTGAAAATAAAGGCAGACACAAAGAAGAAAGGTAACAGCCCTGATGCTGTGACGTTGGGGTGTTGTG
```

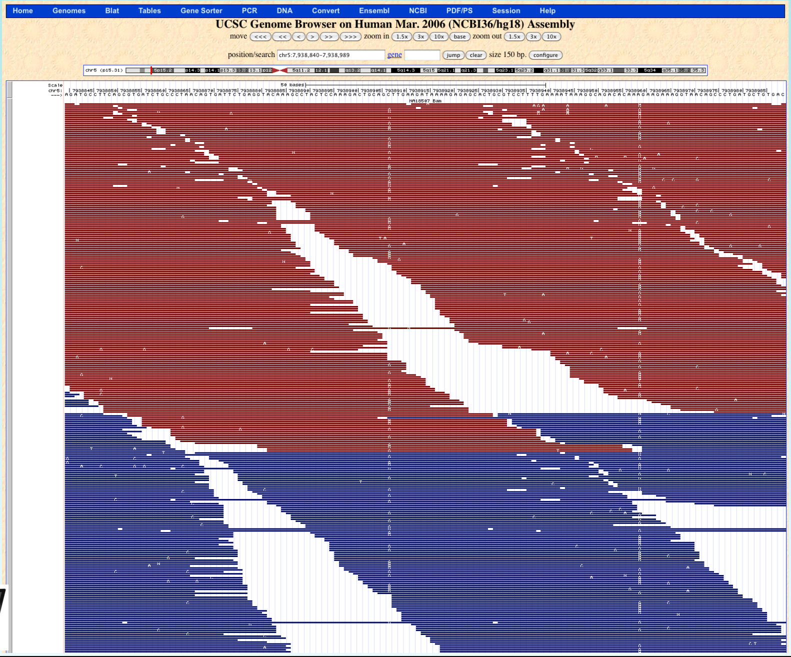
Analysis and Visualization – Samtools tview



- FAST!
- Text-based
- Basic functionality

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

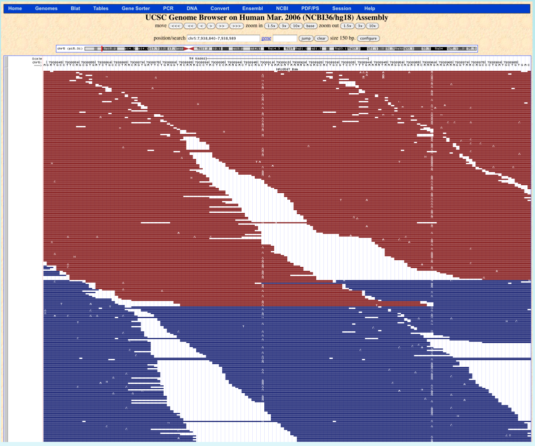
Analysis and Visualization – UCSC Genome Browser



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

position/search chr5:7,938,840-7,938,989 gene size 150 bp.

Analysis and Visualization – UCSC Genome Browser

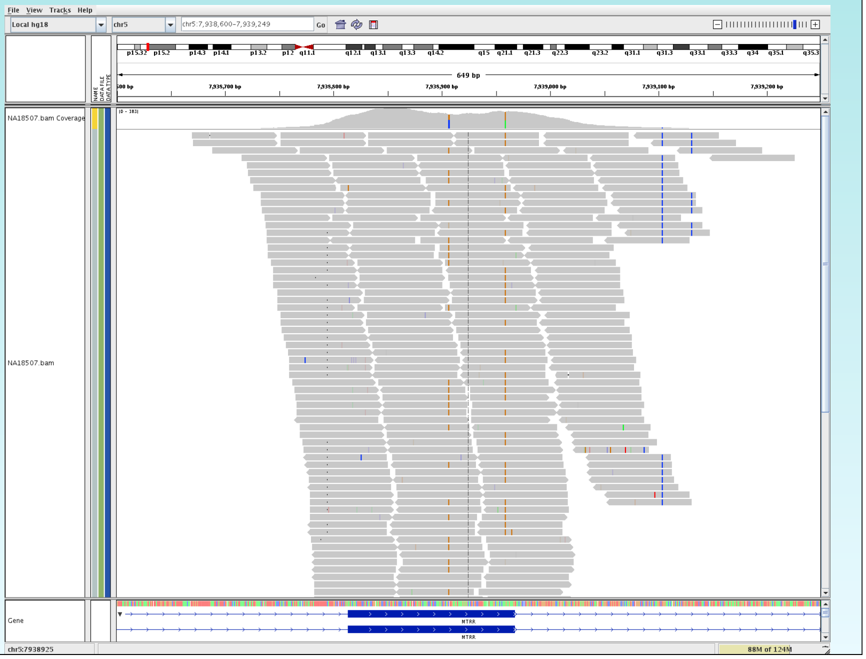


UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

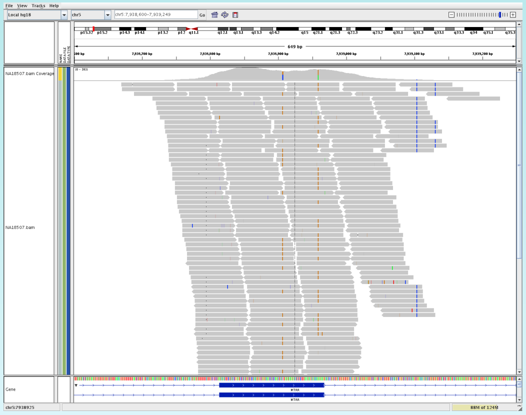
position/search chr5:7,938,840-7,938,989 gene size 150 bp.

- 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?

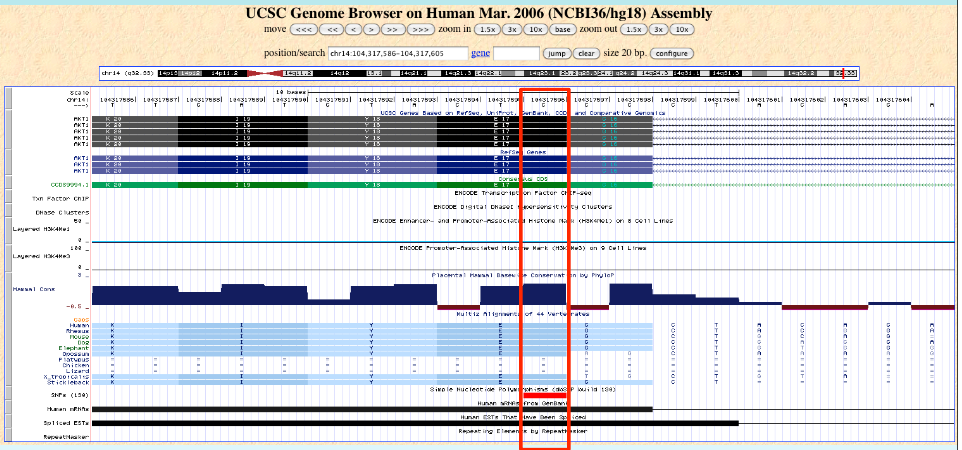
In a gene?

Amino-acid change?

Coding?

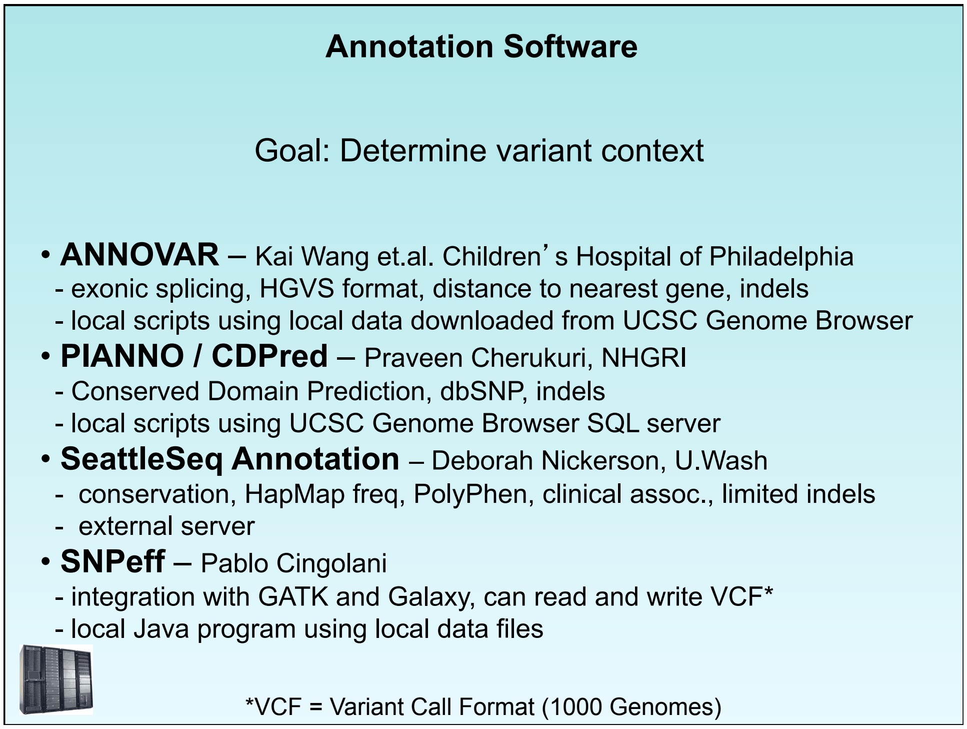
Detrimental for function?

Annotation



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

position/search chr14:104,317,586-104,317,605 size 20 bp.

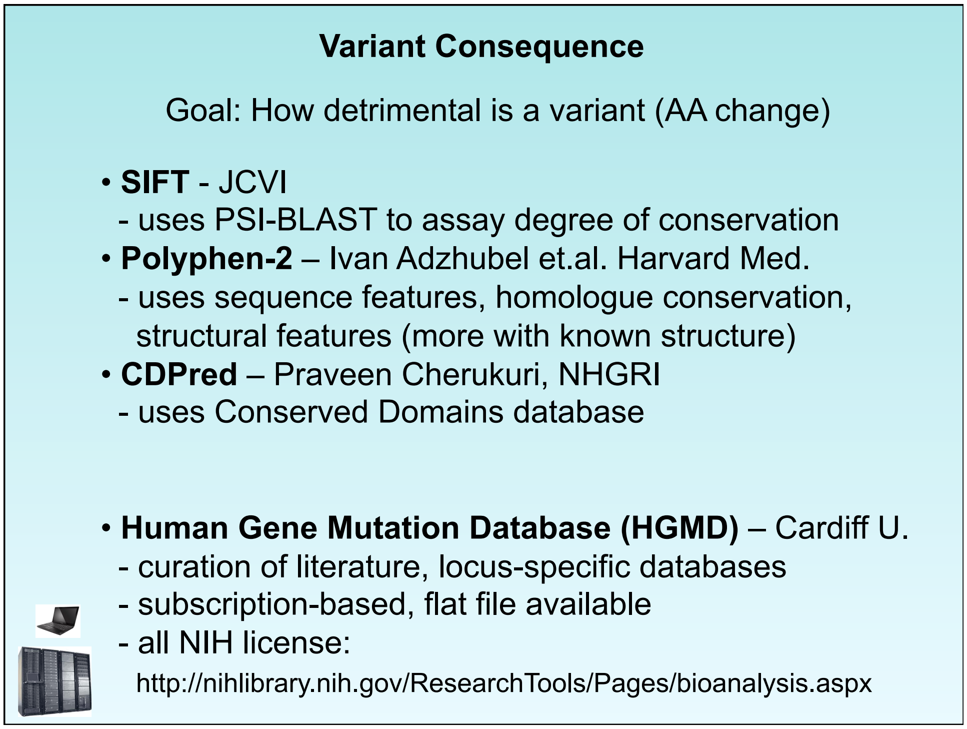


Annotation Software

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

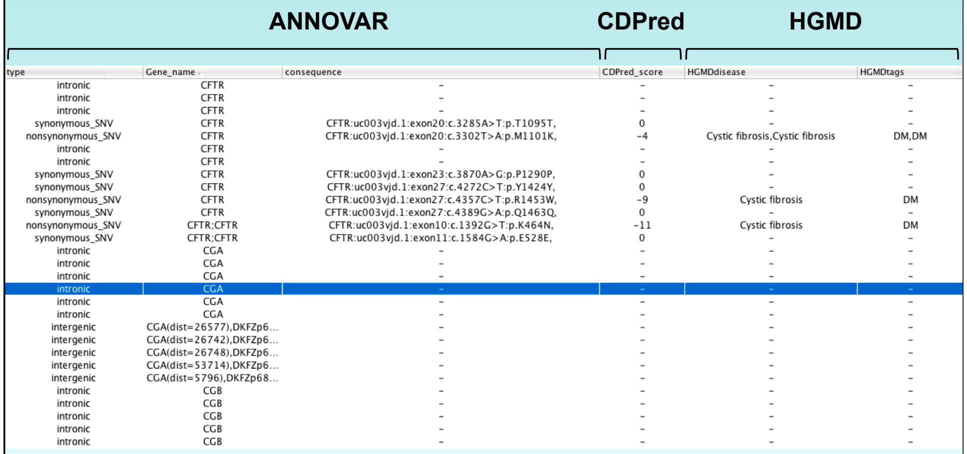
Goal: How detrimental is a variant (AA change)

- **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

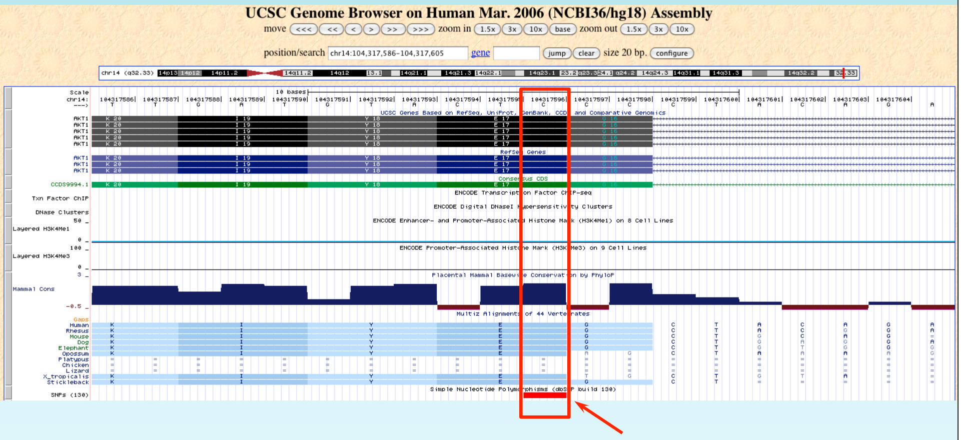
dbSNP
- Includes everything!!
- SNPs have information about origin
- VCF available

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!

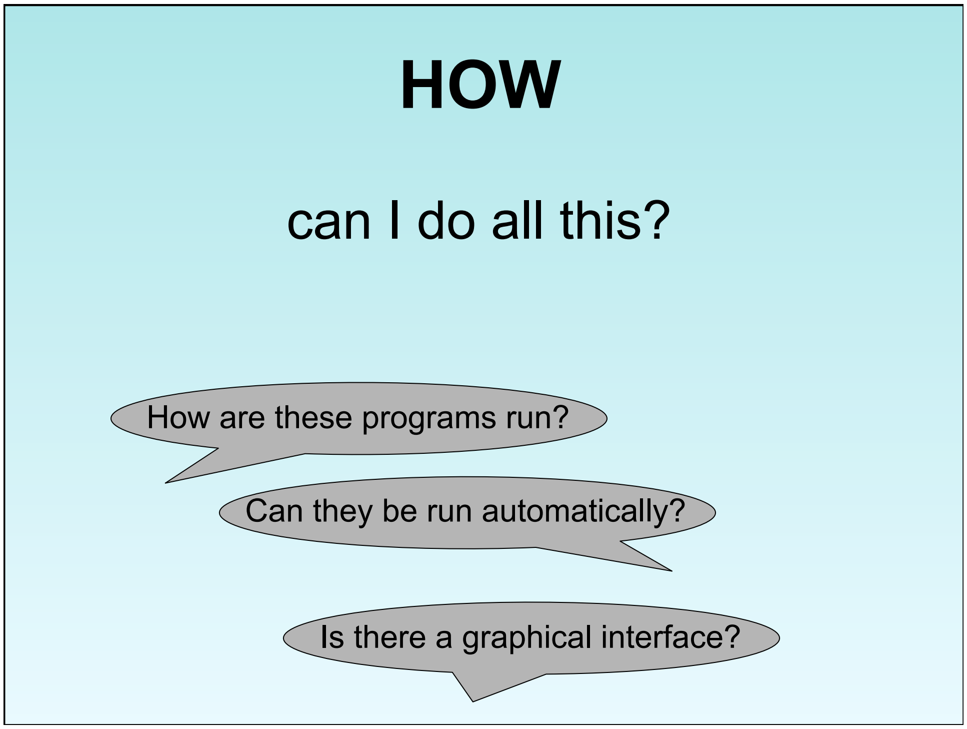


UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

position/search chr14:104,317,586-104,317,605 size 20 bp.

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
gene
jump
clear
configure
chr14 (q32.33)
Scale
10 bases
chr14:
104317586 104317587 104317588 104317589 104317590 104317591 104317592 104317593 104317594 104317595 104317596 104317597 104317598 104317599 104317600 104317601 104317602 104317603 104317604
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
AKT1
K 20
I 19
Y 18
E 17
G 16
RefSeq Genes
Consensus CDS
CCDS9994.1
ENCODE Transcription Factor ChIP-seq
Txn Factor ChIP
ENCODE Digital DNaseI Hypersensitivity Clusters
DNase Clusters
ENCODE Enhancer- and Promoter-Associated Histone Mark (H3K4Me1) on 8 Cell Lines
Layered H3K4Me1
ENCODE Promoter-Associated Histone Mark (H3K4Me3) on 9 Cell Lines
Layered H3K4Me3
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 44 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Opossum
Platypus
Chicken
Lizard
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 130)
SNPs (130)

dbSNP is not exclusively common polymorphisms!!!



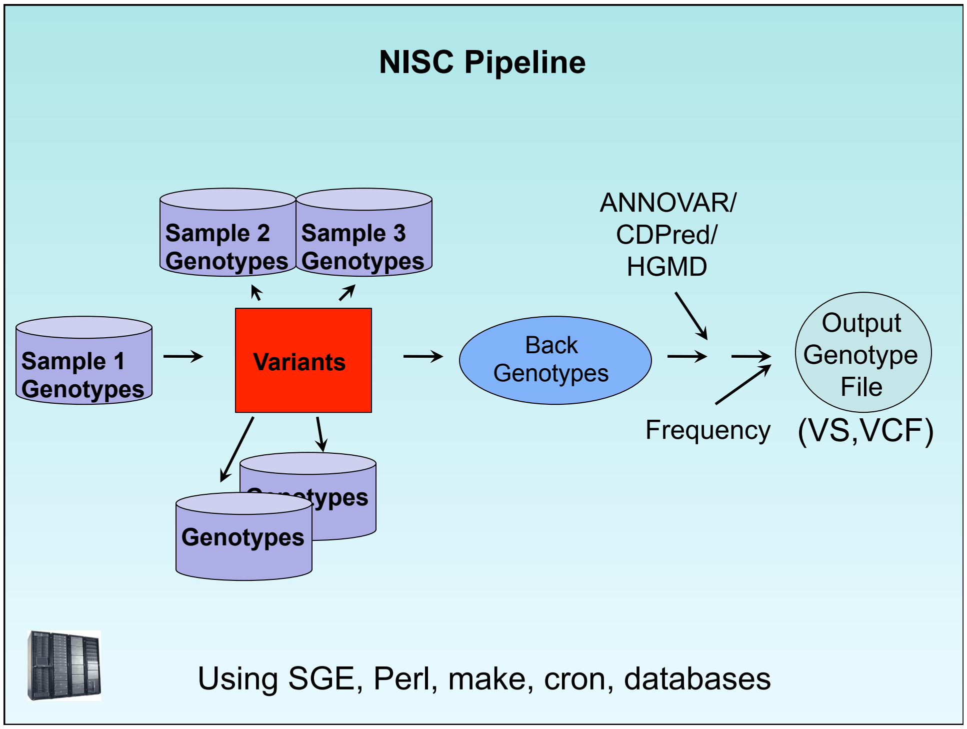
HOW

can I do all this?

How are these programs run?

Can they be run automatically?

Is there a graphical interface?



NISC Pipeline

Sample 2 Genotypes

Sample 3 Genotypes

Sample 1 Genotypes

Variants

Genotypes

Genotypes

Back Genotypes

ANNOVAR/
CDPred/
HGMD

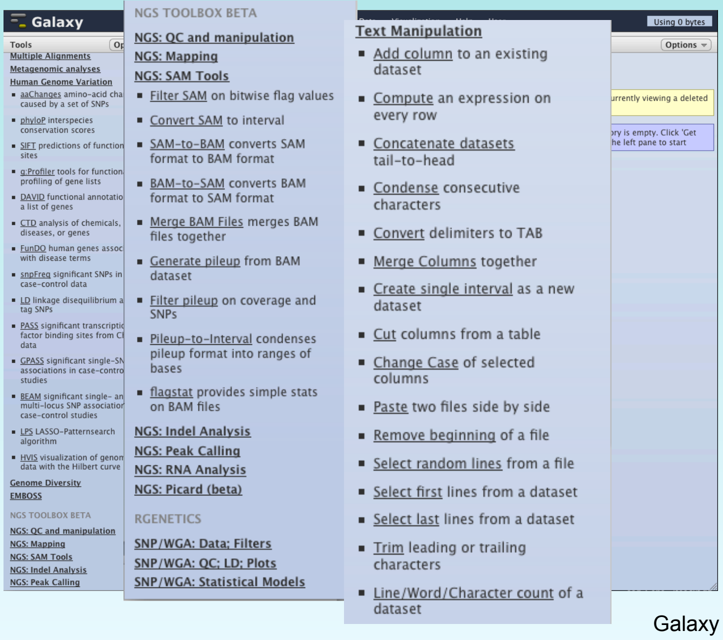
Frequency

Output Genotype File

(VS,VCF)

Using SGE, Perl, make, cron, databases

Pipelines - Galaxy



Multiple Alignments
Metagenomic analyses
Human Genome Variation

- aaChanges amino-acid cha caused by a set of SNPs
- phyloP interspecies conservation scores
- SIFT predictions of function sites
- g:Profiler tools for function profiling of gene lists
- DAVID functional annotatio a list of genes
- CTD analysis of chemicals, diseases, or genes
- FunDO human genes assoc with disease terms
- snpFreq significant SNPs in case-control data
- LD linkage disequilibrium a tag SNPs
- PASS significant transcriptic factor binding sites from Cl data
- GPASS significant single-SN associations in case-contro studies
- BEAM significant single- an multi-locus SNP associatior case-control studies
- LPS LASSO-Patternsearch algorithm
- HVIS visualization of genon data with the Hilbert curve

Genome Diversity
EMBOSS

NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files

NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
NGS: Picard (beta)

RGENETICS

SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models

Text Manipulation

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Condense consecutive characters
- Convert delimiters to TAB
- Merge Columns together
- Create single interval as a new dataset
- Cut columns from a table
- Change Case of selected columns
- Paste two files side by side
- Remove beginning of a file
- Select random lines from a file
- Select first lines from a dataset
- Select last lines from a dataset
- Trim leading or trailing characters
- Line/Word/Character count of a dataset

urrently viewing a deleted

ory is empty. Click 'Get he left pane to start

Galaxy Team

HOW

can I identify
the important variants?

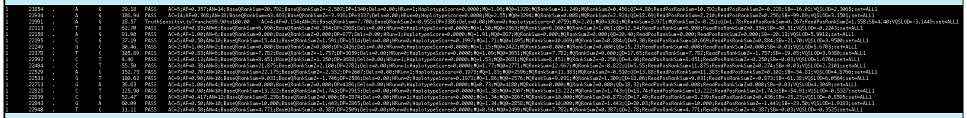
Ab initio?

User-guided?

Easy to use?

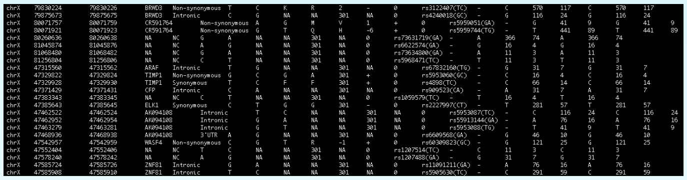
Variant File Formats

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



```
1	21854	.	A	G	29.18	PASS	AC=5;AF=0.357;AN=14;BaseQRankSum=20.792;BaseQRankSumZ=-2.507;DP=1340;Dels=0.00;HRun=1;HaplotypeScore=0.0000;MQ=1.96;MQ0=1329;MQRankSum=11.249;MQRankSumZ=0.456;QD=4.80;ReadPosRankSum=10.792;ReadPosRankSumZ=-0.228;SB=-16.02;VQSLOD=2.3065;set=ALL1
1	21934	.	A	G	186.94	PASS	AC=14;AF=0.368;AN=38;BaseQRankSum=43.463;BaseQRankSumZ=-3.916;DP=3337;Dels=0.00;HRun=0;HaplotypeScore=0.0000;MQ=2.55;MQ0=3294;MQRankSum=0.008;MQRankSumZ=2.924;QD=18.69;ReadPosRankSum=2.210;ReadPosRankSumZ=0.256;SB=-99.89;VQSLOD=3.1501;set=ALL1
1	21991	.	A	G	18.57	TruthSensitivityTranche99.90to100.00	AC=4;AF=0.154;AN=26;BaseQRankSum=7.700;BaseQRankSumZ=-0.955;DP=3386;Dels=0.00;HRun=0;HaplotypeScore=0.0759;MQ=2.01;MQ0=3361;MQRankSum=3.971;MQRankSumZ=-0.251;QD=1.78;ReadPosRankSum=0.267;ReadPosRankSumZ=1.558;SB=4.40;VQSLOD=-3.1449;set=ALL1
1	22113	.	T	G	62.70	PASS	AC=5;AF=0.83;AN=6;BaseQRankSum=6.990;BaseQRankSumZ=0.354;DP=2368;Dels=0.00;HRun=0;HaplotypeScore=0.0000;MQ=1.50;MQ0=2363;MQRankSum=6.990;MQRankSumZ=1.061;QD=12.54;ReadPosRankSum=6.990;ReadPosRankSumZ=0.354;SB=-30.51;VQSLOD=-0.2243;set=ALL1
1	22158	.	A	G	91.90	PASS	AC=6;AF=1.00;AN=6;BaseQRankSum=0.000;BaseQRankSumZ=0.000;DP=877;Dels=0.00;HRun=1;HaplotypeScore=0.0000;MQ=1.91;MQ0=867;MQRankSum=0.000;MQRankSumZ=0.000;QD=20.40;ReadPosRankSum=0.000;ReadPosRankSumZ=0.000;SB=-20.13;VQSLOD=5.9912;set=ALL1
1	22195	.	C	G	37.19	PASS	AC=5;AF=0.50;AN=10;BaseQRankSum=15.441;BaseQRankSumZ=1.591;DP=1514;Dels=0.00;HRun=0;HaplotypeScore=0.1997;MQ=1.71;MQ0=1495;MQRankSum=10.669;MQRankSumZ=0.884;QD=9.30;ReadPosRankSum=10.669;ReadPosRankSumZ=0.884;SB=-21.70;VQSLOD=3.9500;set=ALL1
1	22249	.	G	C	30.46	PASS	AC=2;AF=1.00;AN=2;BaseQRankSum=0.000;BaseQRankSumZ=0.000;DP=2426;Dels=0.00;HRun=0;HaplotypeScore=0.0000;MQ=1.15;MQ0=2422;MQRankSum=0.000;MQRankSumZ=0.000;QD=15.23;ReadPosRankSum=0.000;ReadPosRankSumZ=0.000;SB=-0.03;VQSLOD=5.6703;set=ALL1
1	22378	.	T	C	105.89	PASS	AC=5;AF=0.83;AN=6;BaseQRankSum=7.782;BaseQRankSumZ=-1.757;DP=3659;Dels=0.00;HRun=0;HaplotypeScore=0.0000;MQ=1.09;MQ0=3651;MQRankSum=7.782;MQRankSumZ=0.000;QD=17.65;ReadPosRankSum=7.782;ReadPosRankSumZ=-1.757;SB=-19.05;VQSLOD=3.9380;set=ALL1
1	22392	.	C	T	4.46	PASS	AC=1;AF=0.13;AN=8;BaseQRankSum=8.451;BaseQRankSumZ=1.250;DP=3688;Dels=0.00;HRun=1;HaplotypeScore=0.0000;MQ=1.53;MQ0=3681;MQRankSum=8.451;MQRankSumZ=-0.250;QD=4.46;ReadPosRankSum=8.451;ReadPosRankSumZ=-0.250;SB=-0.03;VQSLOD=1.6764;set=ALL1
1	22494	.	C	T	55.50	PASS	AC=3;AF=0.38;AN=8;BaseQRankSum=21.875;BaseQRankSumZ=2.100;DP=2782;Dels=0.00;HRun=1;HaplotypeScore=0.0000;MQ=1.77;MQ0=2771;MQRankSum=12.667;MQRankSumZ=-0.822;QD=5.55;ReadPosRankSum=11.875;ReadPosRankSumZ=0.274;SB=-0.03;VQSLOD=2.2101;set=ALL1
1	22529	.	A	C	152.73	PASS	AC=7;AF=0.70;AN=10;BaseQRankSum=22.175;BaseQRankSumZ=-2.552;DP=2607;Dels=0.00;HRun=1;HaplotypeScore=0.1973;MQ=1.83;MQ0=2596;MQRankSum=11.383;MQRankSumZ=-0.510;QD=13.88;ReadPosRankSum=11.383;ReadPosRankSumZ=0.102;SB=-54.81;VQSLOD=4.8766;set=ALL1
1	22533	.	G	T	180.62	PASS	AC=9;AF=0.90;AN=10;BaseQRankSum=9.031;BaseQRankSumZ=-1.746;DP=2586;Dels=0.00;HRun=0;HaplotypeScore=0.1973;MQ=1.86;MQ0=2576;MQRankSum=9.031;MQRankSumZ=1.309;QD=18.06;ReadPosRankSum=9.031;ReadPosRankSumZ=-0.873;SB=-61.38;VQSLOD=6.0564;set=ALL1
1	22713	.	G	A	36.63	PASS	AC=4;AF=1.00;AN=4;BaseQRankSum=0.000;BaseQRankSumZ=0.000;DP=4186;Dels=0.00;HRun=0;HaplotypeScore=0.0000;MQ=0.73;MQ0=4180;MQRankSum=0.000;MQRankSumZ=0.000;QD=18.32;ReadPosRankSum=0.000;ReadPosRankSumZ=0.000;SB=-0.03;VQSLOD=4.3049;set=ALL1
1	22825	.	C	T	125.90	PASS	AC=9;AF=0.90;AN=10;BaseQRankSum=13.222;BaseQRankSumZ=1.743;DP=2915;Dels=0.00;HRun=0;HaplotypeScore=0.0000;MQ=1.38;MQ0=2907;MQRankSum=13.222;MQRankSumZ=1.743;QD=15.74;ReadPosRankSum=13.222;ReadPosRankSumZ=1.743;SB=-54.91;VQSLOD=-0.5327;set=ALL1
1	22839	.	C	G	52.47	PASS	AC=5;AF=0.417;AN=12;BaseQRankSum=8.239;BaseQRankSumZ=0.000;DP=2874;Dels=0.00;HRun=1;HaplotypeScore=0.0000;MQ=1.34;MQ0=2867;MQRankSum=10.000;MQRankSumZ=0.873;QD=17.49;ReadPosRankSum=8.239;ReadPosRankSumZ=0.436;SB=-25.23;VQSLOD=-0.8595;set=ALL1
1	22843	.	G	A	60.09	PASS	AC=5;AF=0.50;AN=10;BaseQRankSum=10.000;BaseQRankSumZ=1.443;DP=2865;Dels=0.00;HRun=0;HaplotypeScore=0.0000;MQ=1.34;MQ0=2858;MQRankSum=10.000;MQRankSumZ=1.443;QD=20.03;ReadPosRankSum=10.000;ReadPosRankSumZ=-1.443;SB=-23.50;VQSLOD=1.9183;set=ALL1
1	22948	.	T	C	11.11	PASS	AC=2;AF=0.50;AN=4;BaseQRankSum=4.771;BaseQRankSumZ=-0.387;DP=2509;Dels=0.00;HRun=0;HaplotypeScore=0.0000;MQ=0.94;MQ0=2499;MQRankSum=7.782;MQRankSumZ=0.387;QD=2.78;ReadPosRankSum=4.771;ReadPosRankSumZ=-0.387;SB=-0.03;VQSLOD=-0.3525;set=ALL1
```

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



```
chrX	79830224	79830226	BRWD3	Non-synonymous	T	C	K	R	2	-	0	rs3122407(TC)	-	C	570	117	C	570	117
chrX	79875673	79875675	BRWD3	Intronic	C	G	NA	NA	301	NA	0	rs4240018(GC)	-	G	116	24	G	116	24
chrX	80071757	80071759	CR591764	Non-synonymous	A	G	M	V	1	+	0	rs5959051(GA)	-	G	41	9	G	41	9
chrX	80071921	80071923	CR591764	Non-synonymous	G	T	Q	H	-6	+	0	rs5959744(TG)	-	T	441	89	T	441	89
chrX	80260636	80260638	NA	NC	G	A	NA	NA	301	NA	0	rs73631719(GA)	-	A	366	74	A	366	74
chrX	81045874	81045876	NA	NC	A	G	NA	NA	301	NA	0	rs6622574(GA)	-	G	16	4	G	16	4
chrX	81068480	81068482	NA	NC	G	A	NA	NA	301	NA	0	rs73634800(GA)	-	A	11	3	A	11	3
chrX	81256804	81256806	NA	NC	C	T	NA	NA	301	NA	0	rs5968471(TC)	-	T	11	3	T	11	3
chrX	47315560	47315562	ARAF	Intronic	T	G	NA	NA	301	NA	0	rs67832160(TG)	-	G	31	7	G	31	7
chrX	47329822	47329824	TIMP1	Non-synonymous	G	C	G	A	301	+	0	rs5953060(GC)	-	C	16	4	C	16	4
chrX	47329928	47329930	TIMP1	Synonymous	T	C	F	F	301	+	0	rs4898(TC)	-	C	66	14	C	66	14
chrX	47371429	47371431	CFP	Intronic	C	A	NA	NA	301	NA	0	rs909523(CA)	-	A	31	7	A	31	7
chrX	47383343	47383345	NA	NC	C	T	NA	NA	301	NA	0	rs1059579(TC)	-	T	16	4	T	16	4
chrX	47385643	47385645	ELK1	Synonymous	C	T	G	G	301	-	0	rs2227997(CT)	-	T	281	57	T	281	57
chrX	47462522	47462524	AK094108	Intronic	T	C	NA	NA	301	NA	0	rs5953087(TC)	-	C	116	24	C	116	24
chrX	47462952	47462954	AK094108	Intronic	G	A	NA	NA	301	NA	0	rs55913144(GA)	-	A	76	16	A	76	16
chrX	47463279	47463281	AK094108	Intronic	G	T	NA	NA	301	NA	0	rs5953088(TG)	-	T	41	9	T	41	9
chrX	47468936	47468938	AK094108	3'UTR	A	G	NA	NA	301	NA	0	rs6609568(GA)	-	G	46	10	G	46	10
chrX	47542957	47542959	WASF4	Non-synonymous	C	G	T	R	-1	+	0	rs60309823(GC)	-	G	121	25	G	121	25
chrX	47552404	47552406	NA	NC	T	C	NA	NA	301	NA	0	rs1207514(TC)	-	C	11	3	C	11	3
chrX	47578240	47578242	NA	NC	A	G	NA	NA	301	NA	0	rs1207488(GA)	-	G	31	7	G	31	7
chrX	47585724	47585726	ZNF81	Intronic	G	A	NA	NA	301	NA	0	rs11091211(GA)	-	A	76	16	A	76	16
chrX	47585908	47585910	ZNF81	Intronic	T	C	NA	NA	301	NA	0	rs5905630(TC)	-	C	291	59	C	291	59
```

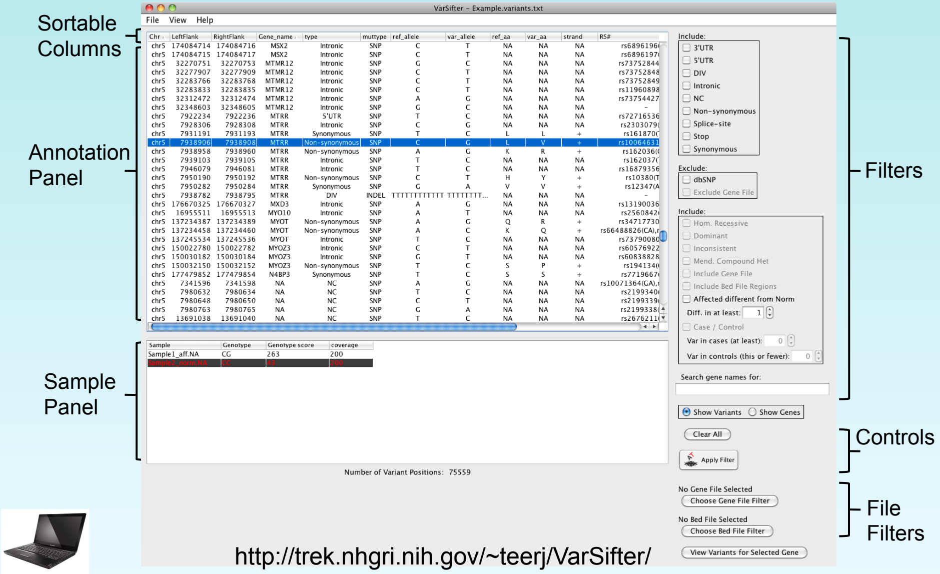
Variant Prioritization

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



Sortable Columns

Annotation Panel

Sample Panel

Filters

Controls

File Filters

VarSifter – Example.variants.txt

Chr	LeftFlank	RightFlank	Gene_name	type	muttype	ref_allele	var_allele	ref_aa	var_aa	strand	RS#
chr5	174084714	174084716	MSX2	Intronic	SNP	C	T	NA	NA	NA	rs6896196(
chr5	174084715	174084717	MSX2	Intronic	SNP	C	T	NA	NA	NA	rs6896197(
chr5	32270751	32270753	MTMR12	Intronic	SNP	G	C	NA	NA	NA	rs73752844
chr5	32277907	32277909	MTMR12	Intronic	SNP	C	T	NA	NA	NA	rs73752848
chr5	32283766	32283768	MTMR12	Intronic	SNP	C	T	NA	NA	NA	rs73752849
chr5	32283833	32283835	MTMR12	Intronic	SNP	C	T	NA	NA	NA	rs11960898
chr5	32312472	32312474	MTMR12	Intronic	SNP	A	G	NA	NA	NA	rs73754427
chr5	32348603	32348605	MTMR12	Intronic	SNP	G	C	NA	NA	NA	–
chr5	7922234	7922236	MTRR	5'UTR	SNP	T	C	NA	NA	NA	rs72716536
chr5	7928306	7928308	MTRR	Intronic	SNP	C	G	NA	NA	NA	rs2303079(
chr5	7931191	7931193	MTRR	Synonymous	SNP	T	C	L	L	+	rs161870(
chr5	7938906	7938908	MTRR	Non-synonymous	SNP	C	G	L	V	+	rs10064631
chr5	7938958	7938960	MTRR	Non-synonymous	SNP	A	G	K	R	+	rs162036(
chr5	7939103	7939105	MTRR	Intronic	SNP	T	C	NA	NA	NA	rs162037(
chr5	7946079	7946081	MTRR	Intronic	SNP	T	C	NA	NA	NA	rs16879356
chr5	7950190	7950192	MTRR	Non-synonymous	SNP	C	T	H	Y	+	rs10380(T
chr5	7950282	7950284	MTRR	Synonymous	SNP	G	A	V	V	+	rs12347(A
chr5	7938782	7938795	MTRR	DIV	INDEL	TTTTTTTTTTTT	TTTTTTTT...	NA	NA	NA	–
chr5	176670325	176670327	MXD3	Intronic	SNP	A	G	NA	NA	NA	rs13190036
chr5	16955511	16955513	MYO10	Intronic	SNP	A	T	NA	NA	NA	rs2560842(
chr5	137234387	137234389	MYOT	Non-synonymous	SNP	A	G	Q	R	+	rs34717730
chr5	137234458	137234460	MYOT	Non-synonymous	SNP	A	C	K	Q	+	rs66488826(CA),r
chr5	137245534	137245536	MYOT	Intronic	SNP	T	C	NA	NA	NA	rs73790080
chr5	150022780	150022782	MYOZ3	Intronic	SNP	C	T	NA	NA	NA	rs60576922
chr5	150030182	150030184	MYOZ3	Intronic	SNP	G	T	NA	NA	NA	rs60838828
chr5	150032150	150032152	MYOZ3	Non-synonymous	SNP	T	C	S	P	+	rs194134(
chr5	177479852	177479854	N4BP3	Synonymous	SNP	T	C	S	S	+	rs7719667(
chr5	7341596	7341598	NA	NC	SNP	A	G	NA	NA	NA	rs10071364(GA),r
chr5	7980632	7980634	NA	NC	SNP	T	C	NA	NA	NA	rs2199340(
chr5	7980648	7980650	NA	NC	SNP	C	T	NA	NA	NA	rs2199339(
chr5	7980763	7980765	NA	NC	SNP	G	A	NA	NA	NA	rs2199338(
chr5	13691038	13691040	NA	NC	SNP	T	C	NA	NA	NA	rs2676211(

Sample	Genotype	Genotype score	coverage
Sample1_aff.NA	CG	263	200
Sample2_norm.NA	CC	43	200

Number of Variant Positions: 75559

Include:
- □ 3'UTR
- □ 5'UTR
- □ DIV
- □ Intronic
- □ NC
- □ Non-synonymous
- □ Splice-site
- □ Stop
- □ Synonymous

Exclude:
- □ dbSNP
- □ Exclude Gene File

Include:
- □ Hom. Recessive
- □ Dominant
- □ Inconsistent
- □ Mend. Compound Het
- □ Include Gene File
- □ Include Bed File Regions
- □ Affected different from Norm

Diff. in at least: 1

- □ Case / Control

Var in cases (at least): 0

Var in controls (this or fewer): 0

Search gene names for:

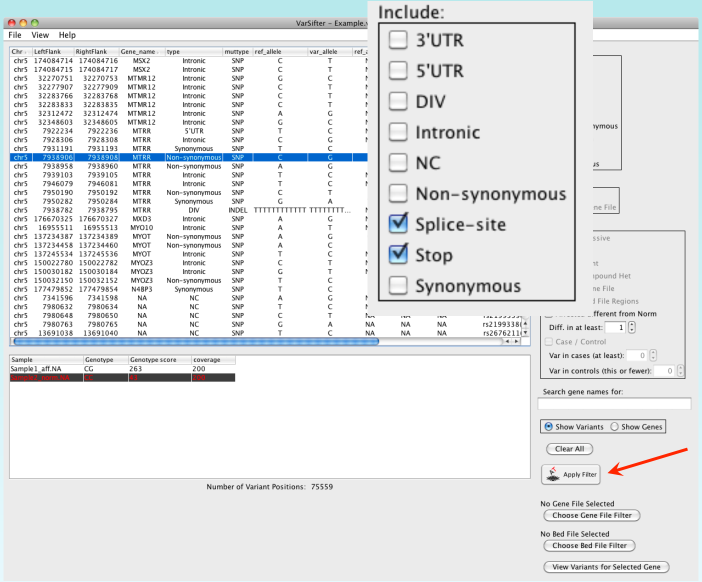
● Show Variants ○ Show Genes

No Gene File Selected

No Bed File Selected

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

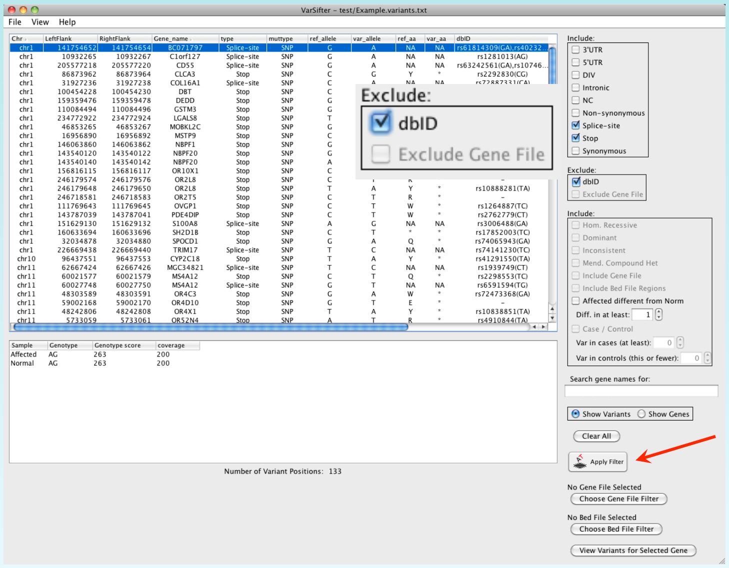
Filtering – Mutation Type



Include:
- □ 3'UTR
- □ 5'UTR
- □ DIV
- □ Intronic
- □ NC
- □ Non-synonymous
- ☑ Splice-site
- ☑ Stop
- □ Synonymous

Number of Variant Positions: 75559

Filtering – Mutation Type



VarSifter – test/Example.variants.txt

Chr	LeftFlank	RightFlank	Gene_name	type	muttype	ref_allele	var_allele	ref_aa	var_aa	dbID
chr1	141754652	141754654	BC071797	Splice-site	SNP	G	A	NA	NA	rs61814309(GA),rs40232...
chr1	10932265	10932267	C1orf127	Splice-site	SNP	G	A	NA	NA	rs1281013(AG)
chr1	205577218	205577220	CD55	Splice-site	SNP	G	A	NA	NA	rs63242561(GA),rs10746...
chr1	86873962	86873964	CLCA3	Stop	SNP	C	G	Y	*	rs2292830(CG)
chr1	31927236	31927238	COL16A1	Splice-site	SNP	C	A	NA	NA	rs72887331(CA)
chr1	100454228	100454230	DBT	Stop	SNP	C				
chr1	159359476	159359478	DEDD	Stop	SNP	G				
chr1	110084494	110084496	GSTM3	Stop	SNP	G				
chr1	234772922	234772924	LGALS8	Stop	SNP	T				
chr1	46853265	46853267	MOBKL2C	Stop	SNP	G				
chr1	16956890	16956892	MSTP9	Stop	SNP	C				
chr1	146063860	146063862	NBPF1	Stop	SNP	G				
chr1	143540120	143540122	NBPF20	Stop	SNP	G				
chr1	143540140	143540142	NBPF20	Stop	SNP	A				
chr1	156816115	156816117	OR10X1	Stop	SNP	C				
chr1	246179574	246179576	OR2L8	Stop	SNP	C	T	R		–
chr1	246179648	246179650	OR2L8	Stop	SNP	T	A	Y	*	rs10888281(TA)
chr1	246718581	246718583	OR2T5	Stop	SNP	C	T	R	*	–
chr1	111769643	111769645	OVGP1	Stop	SNP	C	T	W	*	rs1264887(TC)
chr1	143787039	143787041	PDE4DIP	Stop	SNP	C	T	W	*	rs2762779(CT)
chr1	151629130	151629132	S100A8	Splice-site	SNP	A	G	NA	NA	rs3006488(GA)
chr1	160633694	160633696	SH2D1B	Stop	SNP	C	T	*	*	rs17852003(TC)
chr1	32034878	32034880	SPOCD1	Stop	SNP	G	A	Q	*	rs74065943(GA)
chr1	226669438	226669440	TRIM17	Splice-site	SNP	T	C	NA	NA	rs74141230(TC)
chr10	96437551	96437553	CYP2C18	Stop	SNP	T	A	Y	*	rs41291550(TA)
chr11	62667424	62667426	MGC34821	Splice-site	SNP	T	C	NA	NA	rs1939749(CT)
chr11	60021577	60021579	MS4A12	Stop	SNP	C	T	Q	*	rs2298553(TC)
chr11	60027748	60027750	MS4A12	Splice-site	SNP	G	T	NA	NA	rs6591594(TG)
chr11	48303589	48303591	OR4C3	Stop	SNP	G	A	W	*	rs72473368(GA)
chr11	59002168	59002170	OR4D10	Stop	SNP	G	T	E	*	–
chr11	48242806	48242808	OR4X1	Stop	SNP	T	A	Y	*	rs10838851(TA)
chr11	5733059	5733061	OR52N4	Stop	SNP	A	T	R	*	rs4910844(TA)

Sample	Genotype	Genotype score	coverage
Affected	AG	263	200
Normal	AG	263	200

Number of Variant Positions: 133

Exclude:
☑ dbID
☐ Exclude Gene File

Include:
☐ 3'UTR
☐ 5'UTR
☐ DIV
☐ Intronic
☐ NC
☐ Non-synonymous
☑ Splice-site
☑ Stop
☐ Synonymous

Exclude:
☑ dbID
☐ Exclude Gene File

Include:
☐ Hom. Recessive
☐ Dominant
☐ Inconsistent
☐ Mend. Compound Het
☐ Include Gene File
☐ Include Bed File Regions
☐ Affected different from Norm
Diff. in at least: 1
☐ Case / Control
Var in cases (at least): 0
Var in controls (this or fewer): 0

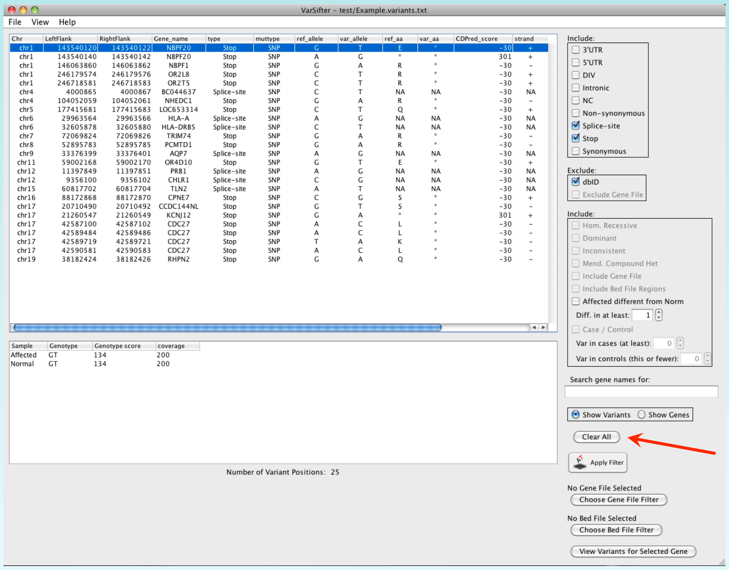
Search gene names for:

● Show Variants ○ Show Genes

No Gene File Selected

No Bed File Selected

Filtering – database



VarSifter – test/Example.variants.txt

Chr	LeftFlank	RightFlank	Gene_name	type	muttype	ref_allele	var_allele	ref_aa	var_aa	CDPred_score	strand
chr1	143540120	143540122	NBPF20	Stop	SNP	G	T	E	*	-30	+
chr1	143540140	143540142	NBPF20	Stop	SNP	A	G	*	*	301	+
chr1	146063860	146063862	NBPF1	Stop	SNP	G	A	R	*	-30	-
chr1	246179574	246179576	OR2L8	Stop	SNP	C	T	R	*	-30	+
chr1	246718581	246718583	OR2T5	Stop	SNP	C	T	R	*	-30	+
chr4	4000865	4000867	BC044637	Splice-site	SNP	C	T	NA	NA	-30	NA
chr4	104052059	104052061	NHEDC1	Stop	SNP	G	A	R	*	-30	-
chr5	177415681	177415683	LOC653314	Stop	SNP	C	T	Q	*	-30	+
chr6	29963564	29963566	HLA-A	Splice-site	SNP	A	G	NA	NA	-30	NA
chr6	32605878	32605880	HLA-DRB5	Splice-site	SNP	C	T	NA	NA	-30	NA
chr7	72069824	72069826	TRIM74	Stop	SNP	G	A	R	*	-30	-
chr8	52895783	52895785	PCMTD1	Stop	SNP	G	A	R	*	-30	-
chr9	33376399	33376401	AQP7	Splice-site	SNP	A	G	NA	NA	-30	NA
chr11	59002168	59002170	OR4D10	Stop	SNP	G	T	E	*	-30	+
chr12	11397849	11397851	PRB1	Splice-site	SNP	A	G	NA	NA	-30	NA
chr12	9356100	9356102	CHLR1	Splice-site	SNP	C	G	NA	NA	-30	NA
chr15	60817702	60817704	TLN2	Splice-site	SNP	A	T	NA	NA	-30	NA
chr16	88172868	88172870	CPNE7	Stop	SNP	C	G	S	*	-30	+
chr17	20710490	20710492	CCDC144NL	Stop	SNP	G	T	S	*	-30	-
chr17	21260547	21260549	KCNJ12	Stop	SNP	G	A	*	*	301	+
chr17	42587100	42587102	CDC27	Stop	SNP	A	C	L	*	-30	-
chr17	42589484	42589486	CDC27	Stop	SNP	A	C	L	*	-30	-
chr17	42589719	42589721	CDC27	Stop	SNP	T	A	K	*	-30	-
chr17	42590581	42590583	CDC27	Stop	SNP	A	C	L	*	-30	-
chr19	38182424	38182426	RHPN2	Stop	SNP	G	A	Q	*	-30	-

Sample	Genotype	Genotype score	coverage
Affected	GT	134	200
Normal	GT	134	200

Number of Variant Positions: 25

Include:
☐ 3'UTR
☐ 5'UTR
☐ DIV
☐ Intronic
☐ NC
☐ Non-synonymous
☑ Splice-site
☑ Stop
☐ Synonymous

Exclude:
☑ dbID
☐ Exclude Gene File

Include:
☐ Hom. Recessive
☐ Dominant
☐ Inconsistent
☐ Mend. Compound Het
☐ Include Gene File
☐ Include Bed File Regions
☐ Affected different from Norm
Diff. in at least: 1
☐ Case / Control
Var in cases (at least): 0
Var in controls (this or fewer): 0

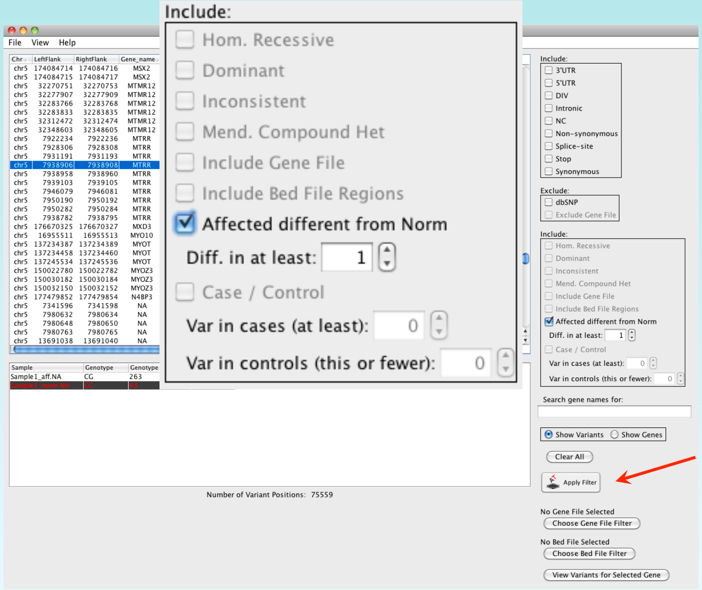
Search gene names for:

● Show Variants ○ Show Genes

No Gene File Selected

No Bed File Selected

Filtering – Affected/Normal



Chr	LeftFlank	RightFlank	Gene_name
chr5	174084714	174084716	MSX2
chr5	174084715	174084717	MSX2
chr5	32270751	32270753	MTMR12
chr5	32277907	32277909	MTMR12
chr5	32283766	32283768	MTMR12
chr5	32283833	32283835	MTMR12
chr5	32312472	32312474	MTMR12
chr5	32348603	32348605	MTMR12
chr5	7922234	7922236	MTRR
chr5	7928306	7928308	MTRR
chr5	7931191	7931193	MTRR
chr5	7938906	7938908	MTRR
chr5	7938958	7938960	MTRR
chr5	7939103	7939105	MTRR
chr5	7946079	7946081	MTRR
chr5	7950190	7950192	MTRR
chr5	7950282	7950284	MTRR
chr5	7938782	7938795	MTRR
chr5	176670325	176670327	MXD3
chr5	16955511	16955513	MYO10
chr5	137234387	137234389	MYOT
chr5	137234458	137234460	MYOT
chr5	137245534	137245536	MYOT
chr5	150022780	150022782	MYOZ3
chr5	150030182	150030184	MYOZ3
chr5	150032150	150032152	MYOZ3
chr5	177479852	177479854	N4BP3
chr5	7341596	7341598	NA
chr5	7980632	7980634	NA
chr5	7980648	7980650	NA
chr5	7980763	7980765	NA
chr5	13691038	13691040	NA

Sample	Genotype	Genotype
Sample1_aff.NA	CG	263
Sample2_norm.NA	CC	43

Include:
- □ Hom. Recessive
- □ Dominant
- □ Inconsistent
- □ Mend. Compound Het
- □ Include Gene File
- □ Include Bed File Regions
- ☑ Affected different from Norm
- Diff. in at least: 1
- □ Case / Control
- Var in cases (at least): 0
- Var in controls (this or fewer): 0

Include:
- □ 3'UTR
- □ 5'UTR
- □ DIV
- □ Intronic
- □ NC
- □ Non-synonymous
- □ Splice-site
- □ Stop
- □ Synonymous

Exclude:
- □ dbSNP
- □ Exclude Gene File

Include:
- □ Hom. Recessive
- □ Dominant
- □ Inconsistent
- □ Mend. Compound Het
- □ Include Gene File
- □ Include Bed File Regions
- ☑ Affected different from Norm
- Diff. in at least: 1
- □ Case / Control
- Var in cases (at least): 0
- Var in controls (this or fewer): 0

Search gene names for:

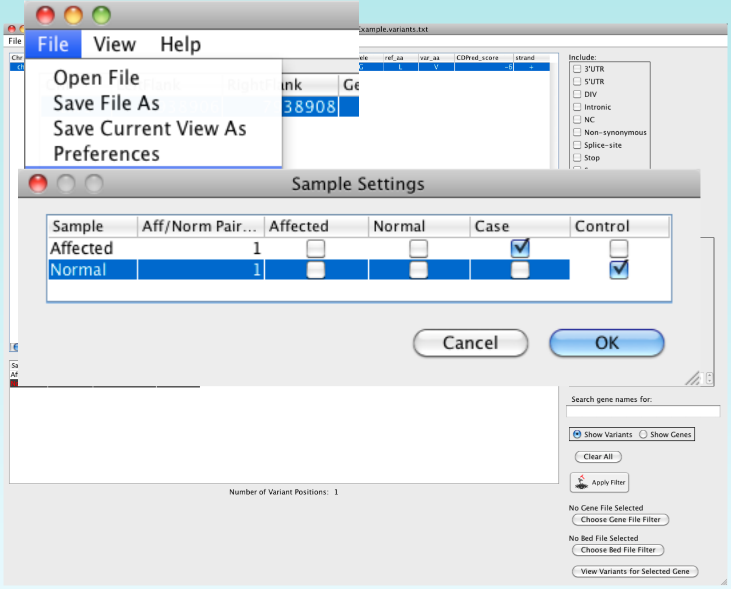
● Show Variants ○ Show Genes

No Gene File Selected

No Bed File Selected

Number of Variant Positions: 75559

Filtering – Affected/Normal



Example.variants.txt

ele	ref_aa	var_aa	CDPred_score	strand
G	L	V	-6	+

Sample Settings

Sample	Aff/Norm Pair…	Affected	Normal	Case	Control
Affected	1	□	□	☑	□
Normal	1	□	□	□	☑

Include:
- □ 3'UTR
- □ 5'UTR
- □ DIV
- □ Intronic
- □ NC
- □ Non-synonymous
- □ Splice-site
- □ Stop

Search gene names for:

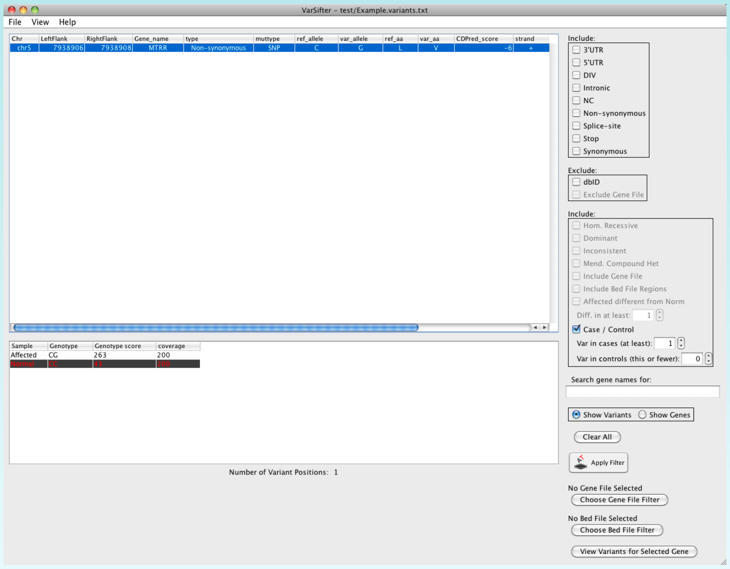
● Show Variants ○ Show Genes

No Gene File Selected

No Bed File Selected

Number of Variant Positions: 1

Filtering – Affected/Normal



VarSifter – test/Example.variants.txt

Chr	LeftFlank	RightFlank	Gene_name	type	muttype	ref_allele	var_allele	ref_aa	var_aa	CDPred_score	strand
chr5	7938906	7938908	MTRR	Non-synonymous	SNP	C	G	L	V	-6	+

Sample	Genotype	Genotype score	coverage
Affected	CG	263	200
Normal	CC	43	200

Number of Variant Positions: 1

Include:
- ☐ 3'UTR
- ☐ 5'UTR
- ☐ DIV
- ☐ Intronic
- ☐ NC
- ☐ Non-synonymous
- ☐ Splice-site
- ☐ Stop
- ☐ Synonymous

Exclude:
- ☐ dbID
- ☐ Exclude Gene File

Include:
- ☐ Hom. Recessive
- ☐ Dominant
- ☐ Inconsistent
- ☐ Mend. Compound Het
- ☐ Include Gene File
- ☐ Include Bed File Regions
- ☐ Affected different from Norm
- Diff. in at least: 1
- ☑ Case / Control
- Var in cases (at least): 1
- Var in controls (this or fewer): 0

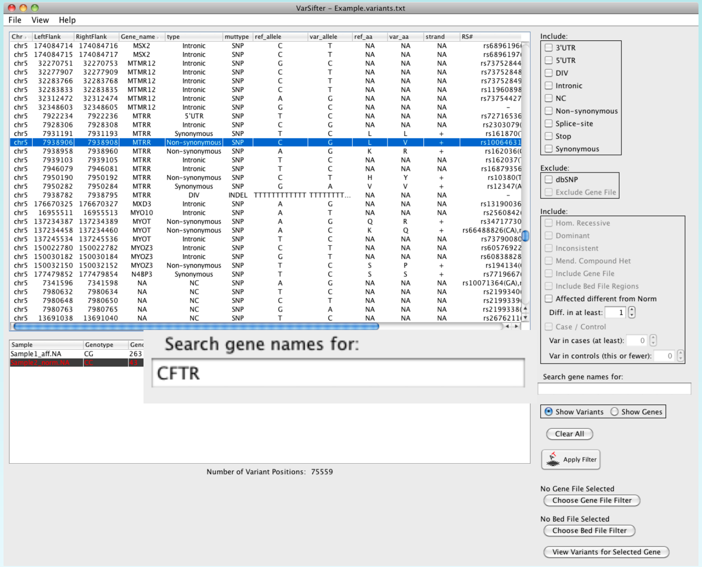
Search gene names for:

● Show Variants ○ Show Genes

No Gene File Selected

No Bed File Selected

Filtering – Gene Name



VarSifter – Example.variants.txt

Chr	LeftFlank	RightFlank	Gene_name	type	muttype	ref_allele	var_allele	ref_aa	var_aa	strand	RS#
chr5	174084714	174084716	MSX2	Intronic	SNP	C	T	NA	NA	NA	rs6896196(
chr5	174084715	174084717	MSX2	Intronic	SNP	C	T	NA	NA	NA	rs6896197(
chr5	32270751	32270753	MTMR12	Intronic	SNP	G	C	NA	NA	NA	rs73752844
chr5	32277907	32277909	MTMR12	Intronic	SNP	C	T	NA	NA	NA	rs73752848
chr5	32283766	32283768	MTMR12	Intronic	SNP	C	T	NA	NA	NA	rs73752849
chr5	32283833	32283835	MTMR12	Intronic	SNP	C	T	NA	NA	NA	rs11960898
chr5	32312472	32312474	MTMR12	Intronic	SNP	A	G	NA	NA	NA	rs73754427
chr5	32348603	32348605	MTMR12	Intronic	SNP	G	C	NA	NA	NA	-
chr5	7922234	7922236	MTRR	5'UTR	SNP	T	C	NA	NA	NA	rs72716536
chr5	7928306	7928308	MTRR	Intronic	SNP	C	G	NA	NA	NA	rs2303079(
chr5	7931191	7931193	MTRR	Synonymous	SNP	T	C	L	L	+	rs161870(
chr5	7938906	7938908	MTRR	Non-synonymous	SNP	C	G	L	V	+	rs10064631
chr5	7938958	7938960	MTRR	Non-synonymous	SNP	A	G	K	R	+	rs162036(
chr5	7939103	7939105	MTRR	Intronic	SNP	T	C	NA	NA	NA	rs162037(
chr5	7946079	7946081	MTRR	Intronic	SNP	T	C	NA	NA	NA	rs16879356
chr5	7950190	7950192	MTRR	Non-synonymous	SNP	C	T	H	Y	+	rs10380(T
chr5	7950282	7950284	MTRR	Synonymous	SNP	G	A	V	V	+	rs12347(A
chr5	7938782	7938795	MTRR	DIV	INDEL	TTTTTTTTTTTT	TTTTTTTT...	NA	NA	NA	-
chr5	176670325	176670327	MXD3	Intronic	SNP	A	G	NA	NA	NA	rs13190036
chr5	16955511	16955513	MYO10	Intronic	SNP	A	T	NA	NA	NA	rs2560842(
chr5	137234387	137234389	MYOT	Non-synonymous	SNP	A	G	Q	R	+	rs34717730
chr5	137234458	137234460	MYOT	Non-synonymous	SNP	A	C	K	Q	+	rs66488826(CA),r
chr5	137245534	137245536	MYOT	Intronic	SNP	T	C	NA	NA	NA	rs73790080
chr5	150022780	150022782	MYOZ3	Intronic	SNP	C	T	NA	NA	NA	rs60576922
chr5	150030182	150030184	MYOZ3	Intronic	SNP	G	T	NA	NA	NA	rs60838828
chr5	150032150	150032152	MYOZ3	Non-synonymous	SNP	T	C	S	P	+	rs194134(
chr5	177479852	177479854	N4BP3	Synonymous	SNP	T	C	S	S	+	rs7719667(
chr5	7341596	7341598	NA	NC	SNP	A	G	NA	NA	NA	rs10071364(GA),r
chr5	7980632	7980634	NA	NC	SNP	T	C	NA	NA	NA	rs2199340(
chr5	7980648	7980650	NA	NC	SNP	C	T	NA	NA	NA	rs2199339(
chr5	7980763	7980765	NA	NC	SNP	G	A	NA	NA	NA	rs2199338(
chr5	13691038	13691040	NA	NC	SNP	T	C	NA	NA	NA	rs2676211(

Sample	Genotype	Genc
Sample1_aff.NA	CG	263
Sample2_norm.NA	CC	43

Search gene names for:

CFTR

Number of Variant Positions: 75559

Include:
- ☐ 3'UTR
- ☐ 5'UTR
- ☐ DIV
- ☐ Intronic
- ☐ NC
- ☐ Non-synonymous
- ☐ Splice-site
- ☐ Stop
- ☐ Synonymous

Exclude:
- ☐ dbSNP
- ☐ Exclude Gene File

Include:
- ☐ Hom. Recessive
- ☐ Dominant
- ☐ Inconsistent
- ☐ Mend. Compound Het
- ☐ Include Gene File
- ☐ Include Bed File Regions
- ☐ Affected different from Norm
- Diff. in at least: 1
- ☐ Case / Control
- Var in cases (at least): 0
- Var in controls (this or fewer): 0

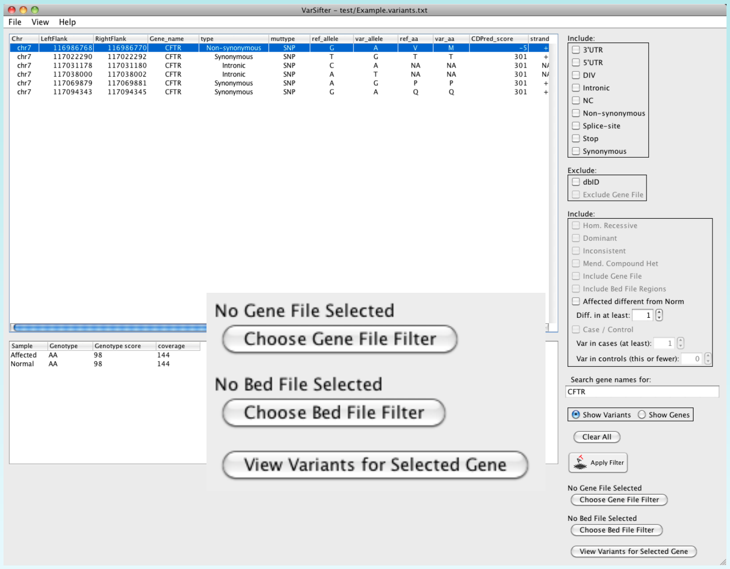
Search gene names for:

● Show Variants ○ Show Genes

No Gene File Selected

No Bed File Selected

Filtering – Gene Name



VarSifter – test/Example.variants.txt

Chr	LeftFlank	RightFlank	Gene_name	type	muttype	ref_allele	var_allele	ref_aa	var_aa	CDPred_score	strand
chr7	116986768	116986770	CFTR	Non-synonymous	SNP	G	A	V	M	-5	+
chr7	117022290	117022292	CFTR	Synonymous	SNP	T	G	T	T	301	+
chr7	117031178	117031180	CFTR	Intronic	SNP	C	A	NA	NA	301	NA
chr7	117038000	117038002	CFTR	Intronic	SNP	A	T	NA	NA	301	NA
chr7	117069879	117069881	CFTR	Synonymous	SNP	A	G	P	P	301	+
chr7	117094343	117094345	CFTR	Synonymous	SNP	G	A	Q	Q	301	+

Sample	Genotype	Genotype score	coverage
Affected	AA	98	144
Normal	AA	98	144

Include:
- □ 3'UTR
- □ 5'UTR
- □ DIV
- □ Intronic
- □ NC
- □ Non-synonymous
- □ Splice-site
- □ Stop
- □ Synonymous

Exclude:
- □ dbID
- □ Exclude Gene File

Include:
- □ Hom. Recessive
- □ Dominant
- □ Inconsistent
- □ Mend. Compound Het
- □ Include Gene File
- □ Include Bed File Regions
- □ Affected different from Norm
- Diff. in at least: 1
- □ Case / Control
- Var in cases (at least): 1
- Var in controls (this or fewer): 0

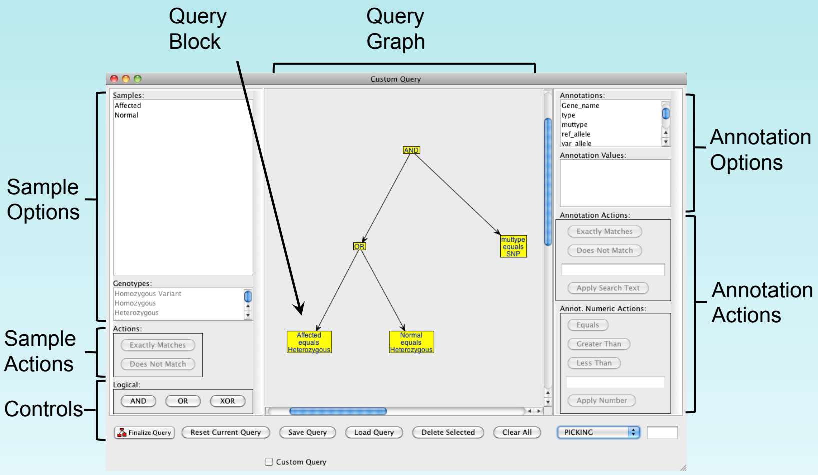
Search gene names for: CFTR

● Show Variants ○ Show Genes

No Gene File Selected

No Bed File Selected

Custom Query Filters



Query Block

Query Graph

Sample Options

Sample Actions

Controls

Annotation Options

Annotation Actions

Custom Query

Samples: Affected, Normal

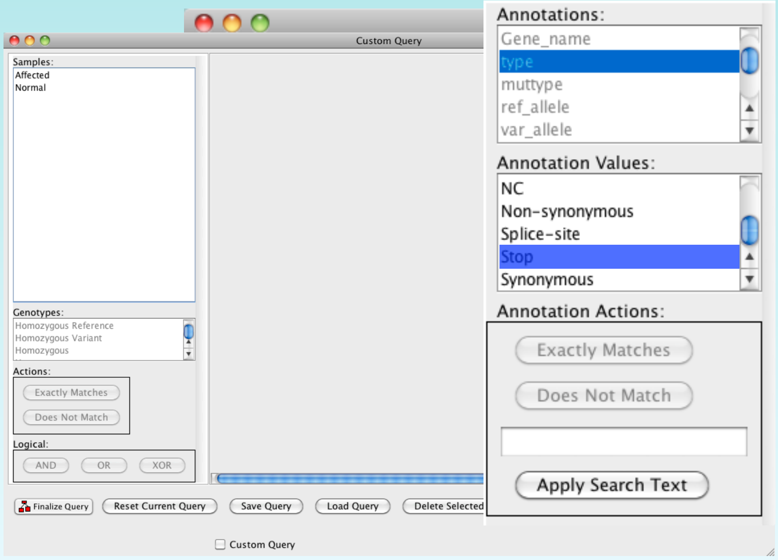
Genotypes: Homozygous Variant, Homozygous, Heterozygous

Annotations: Gene_name, type, muttype, ref_allele, var_allele

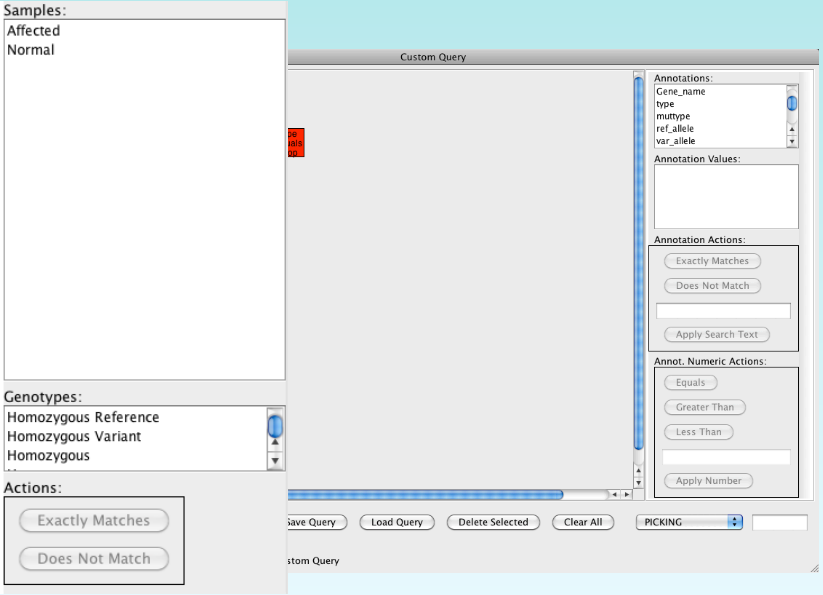
Annotation Values:

□ Custom Query

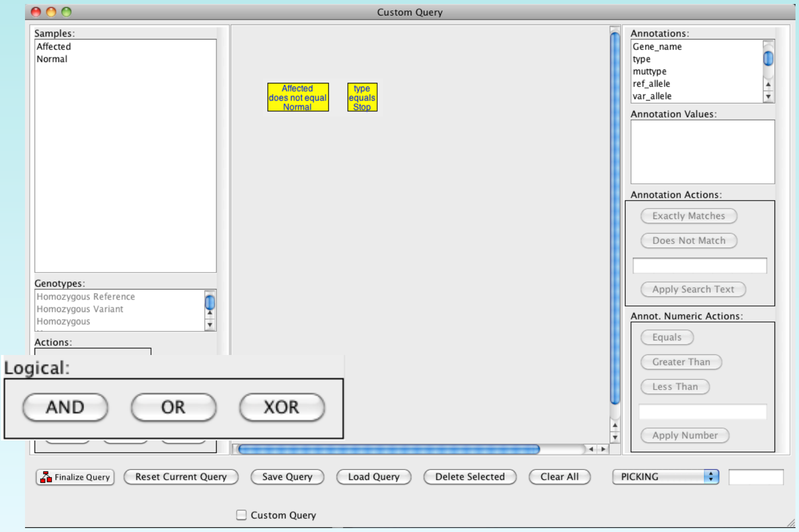
Custom Query Filters - Annotation



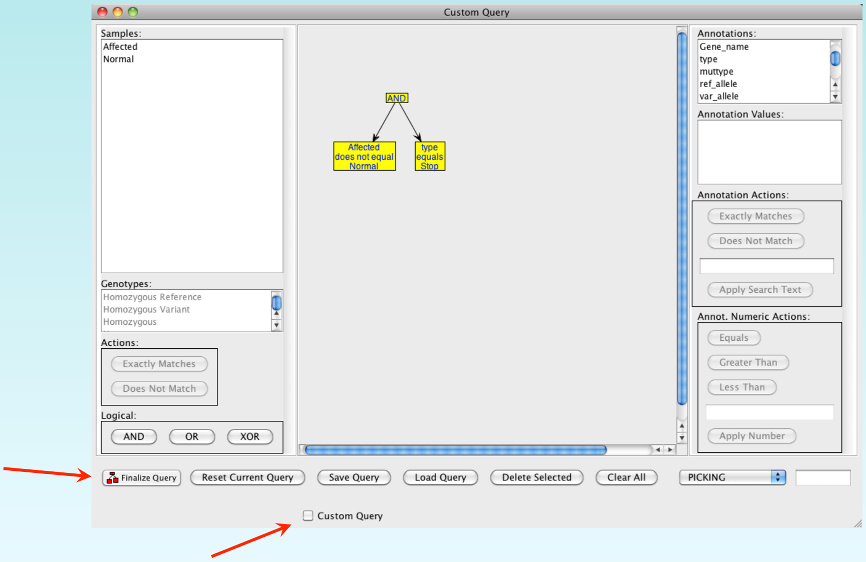
Custom Query Filters - Sample



Custom Query Filters - Linking



Custom Query Filters - Linking



Summary

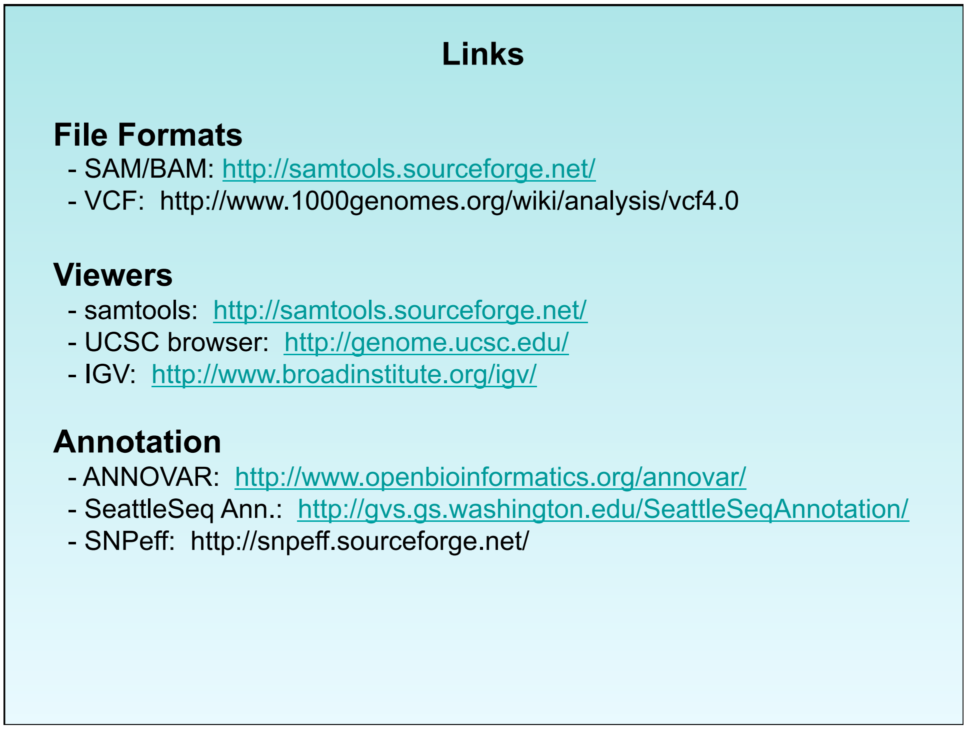
- Annotation gives context
- Consequence prediction can guide analysis
- Varying experience required
- Prioritization tools return “black box” answer
- Visualization can allow guided, informed analysis
- VarSifter is a powerful tool for “hands-on” analysis

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

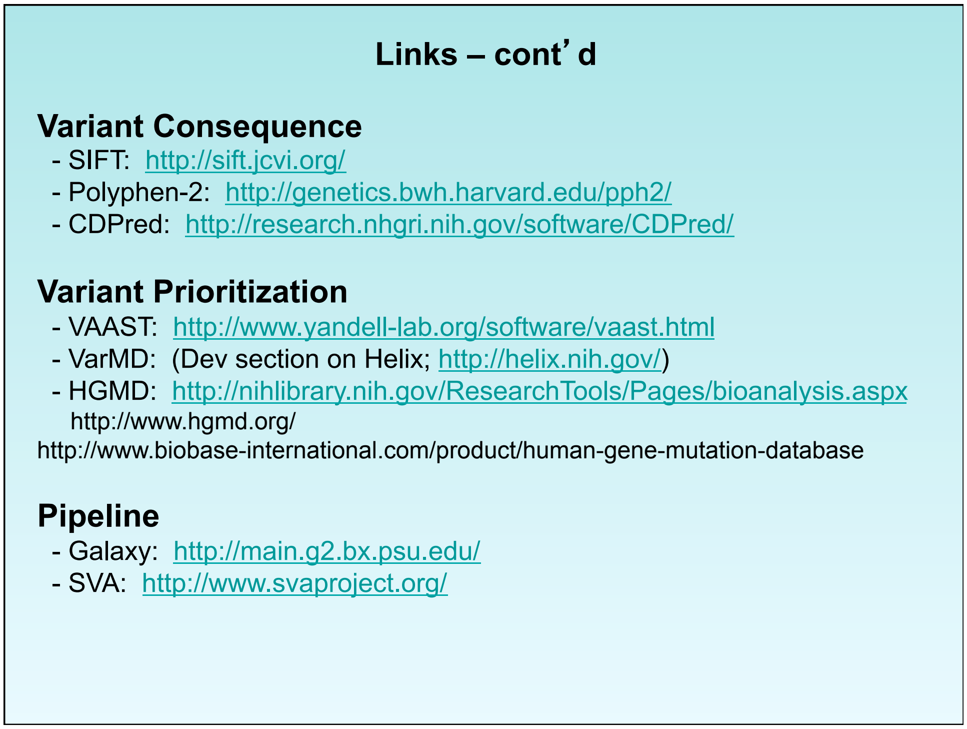
- SAM/BAM: http://samtools.sourceforge.net/
- VCF: http://www.1000genomes.org/wiki/analysis/vcf4.0

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

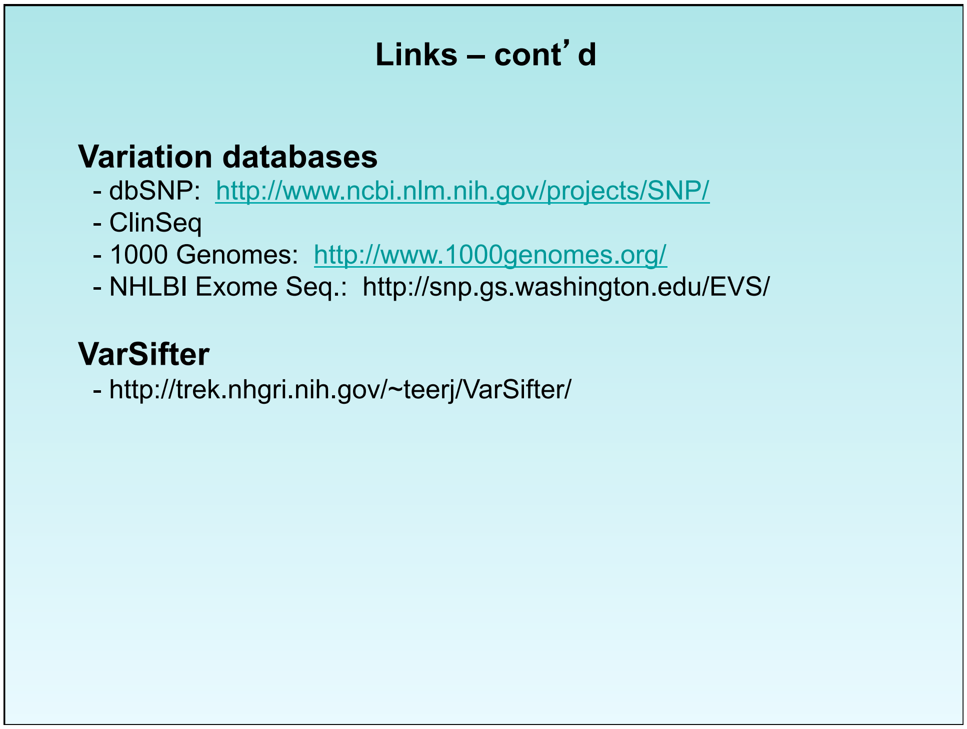
- SIFT: http://sift.jcvi.org/
- Polyphen-2: http://genetics.bwh.harvard.edu/pph2/
- CDPred: http://research.nhgri.nih.gov/software/CDPred/

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/