THE UNIVERSITY OF UTAH

UNIVERSITY OF UTAH SCHOOL OF MEDICINE

VAAST

Variant Annotation Analysis and Search Tool

CSHL Programming for Biology Oct 2014

Barry Moore
Director, Science & Research
USTAR Center for Genetic Discovery
Department of Human Genetics
University of Utah

Outline

- Historical Perspective
- Variant Calling (Follow the Probabilities)
- VAAST
  - VAT
  - VST
  - VAAST 2.0
- Rare Disease Applications
- Common Disease Applications
- Future Directions

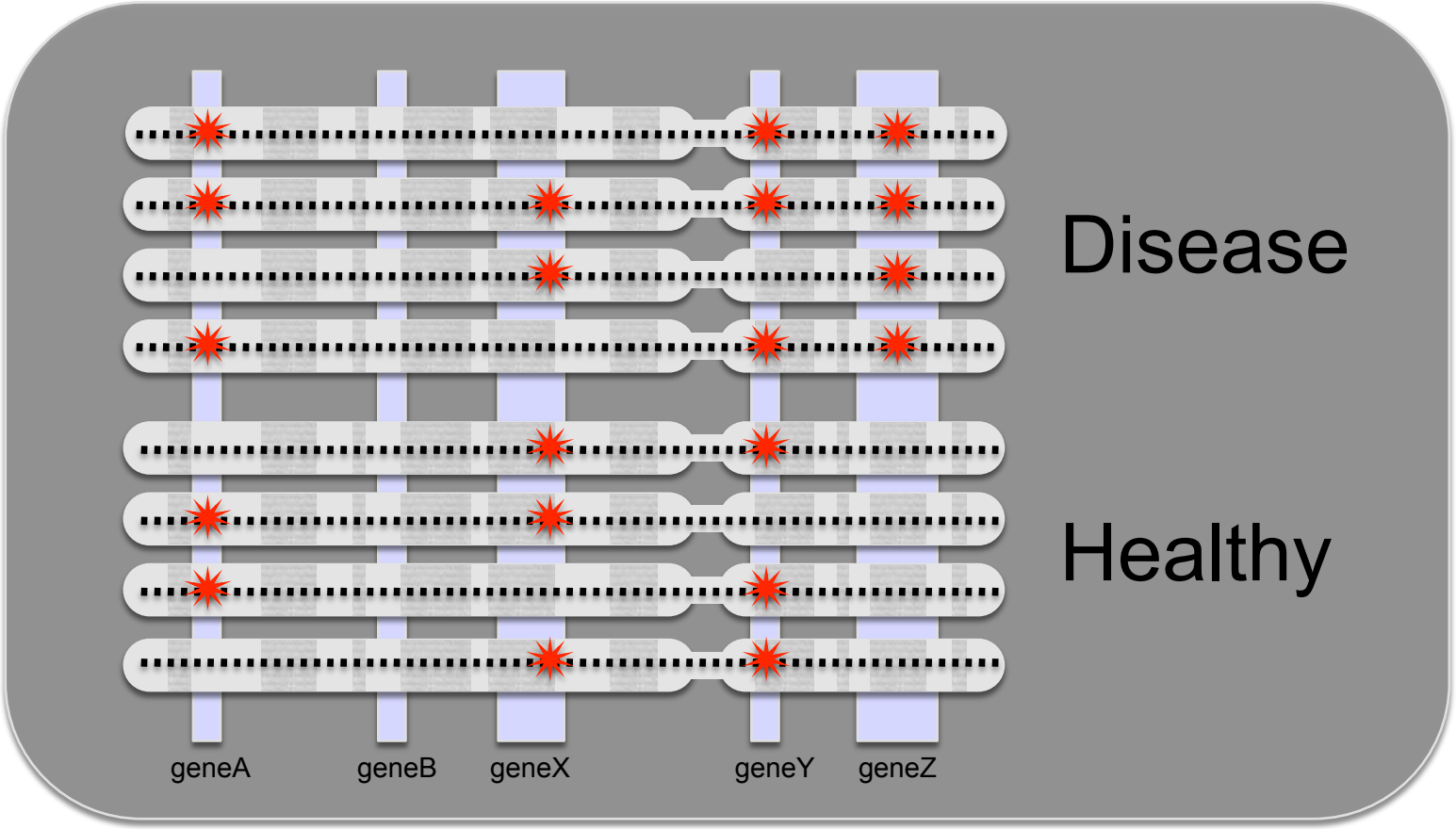
Motivation

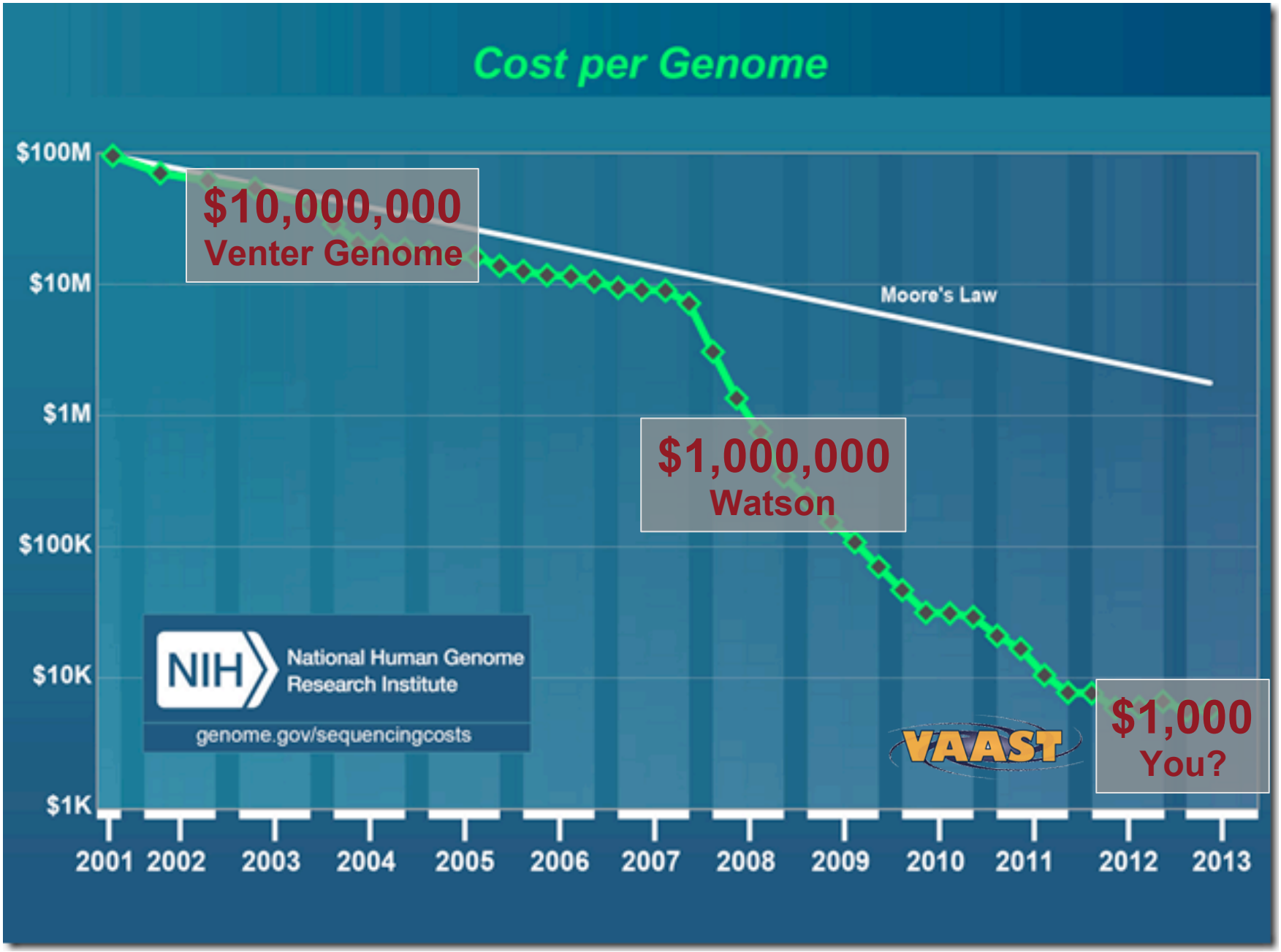
- Billions of years of evolution have fine tuned our DNA sequence.
- Genetic alterations to that sequence can cause disease.
- Knowing which mutations provides clues to understanding the disease.
- What are the mutations – developing technology.
- Which mutation – developing analysis methodolgies.

A Very Breif History of Medical Genetics

- Cytogenetics
  - Downs Syndrome - 1959
- Linkage Mapping
  - HTT (Huntington's Disease) gene mapped – 1983
- Positional Cloning
  - CFTR (Cystic Fibrosis) gene discovered - 1989
- Sequencing, microarrays and GWAS
  - ARMD, CD, MI, IBD – 2005-2006
- Next generation sequencing – personalized genomics
  - Charcot-Marie-Tooth, Miller Syndrome, Ogden Syndrome 2010-2011

Genome Wide Association




Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
$10,000,000
Venter Genome
Moore's Law
$1,000,000
Watson
$1,000
You?
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
VAAST
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

VAAST Overview

- Probabilistic tool for disease gene discovery
- Aggragative variant analysis – feature based
- Both allele and AAAS frequencies
- Conservation-controlled AAS
- Impliments numerous filters
- Standardized ontology based formats
- Modular and flexible in design

Outline

- Historical Perspective
- Variant Calling (Follow the Probabilities)
- VAAST
  - VAT
  - VST
  - VAAST 2.0
- Rare Disease Applications
- Common Disease Applications
- Future Directions

NGS Data Analysis

- Trillions of glowing DNA fragments produce base calls (reads).
- 100s of millions of sequence reads produce alignments.
- 10s of millions of variant sites produce variant calls.
- 10s of thousands of variants analyzed for association with disease.
- 1 gene causes disease.

Follow the Probabilities

- Base calling – Bayesian inference
- Base quality score recalibration – Covariant analysis
- Mapping quality – Pseudo-probabilistic
- Variant calling – Bayesian inference
- Variant quality score recalibration – LOD ratio based on a trained Gaussian mixture model
- VAAST - CLRT

Variant calling - Individual

- What is the probability that this site is reference vs. variant given the reads aligned at this site.

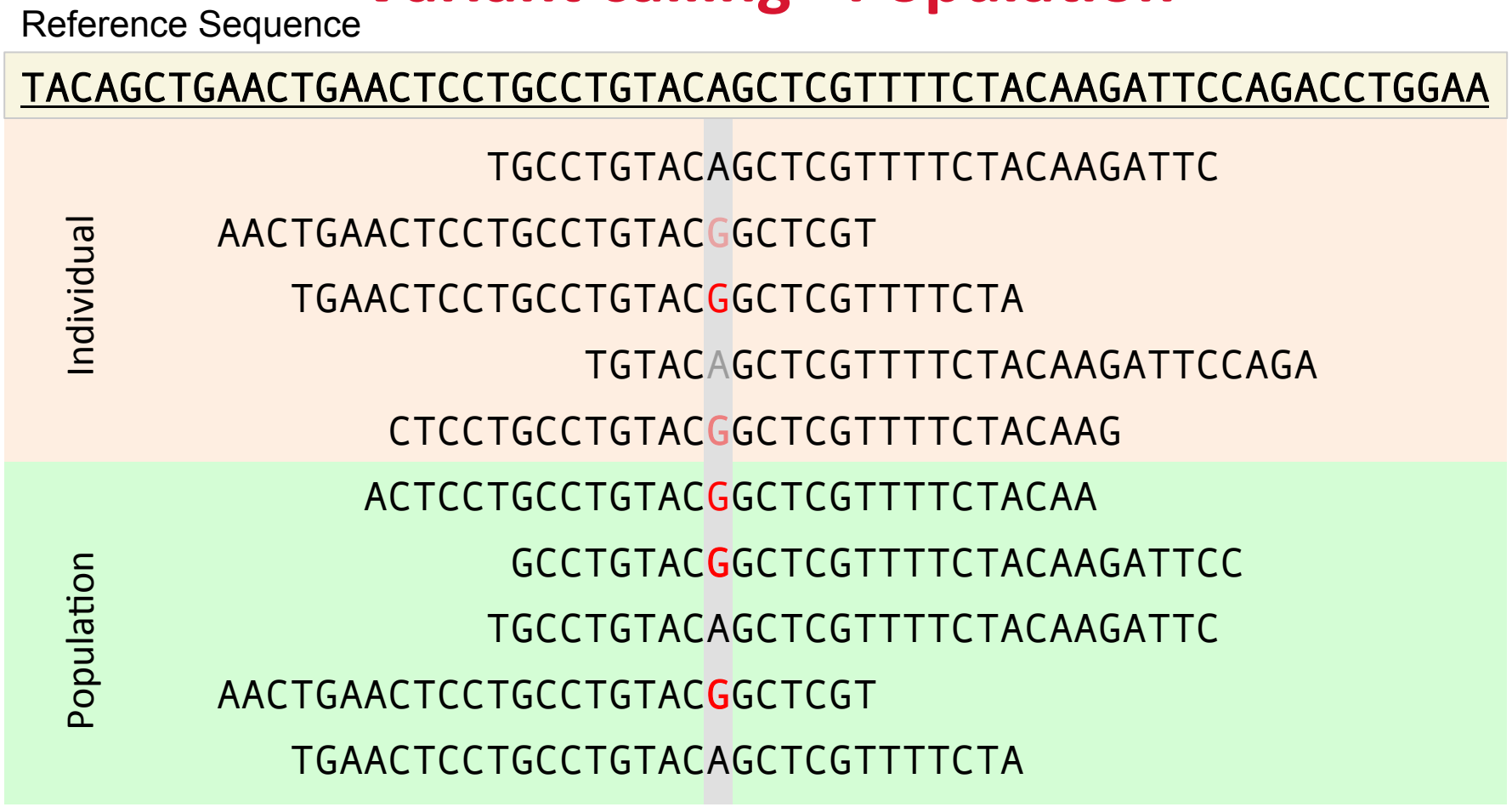
- What is the probability that this site has homozygous reference genotype vs. heterozygous genotype given the reads aligned at this site.

Variant calling - Population

- What is the probability that this site is reference (**FOR THE POPULATION**) given:
  - The reads aligned at this site **FOR THE POPULATION**

- What is the probability that the genotype is homoygous reference (**FOR THE INDIVIDUAL**) given:
  - The reads aligned at this site **FOR THE INDIVIDUAL**
  - The probability that this site is variant **FOR THE POPULATION**

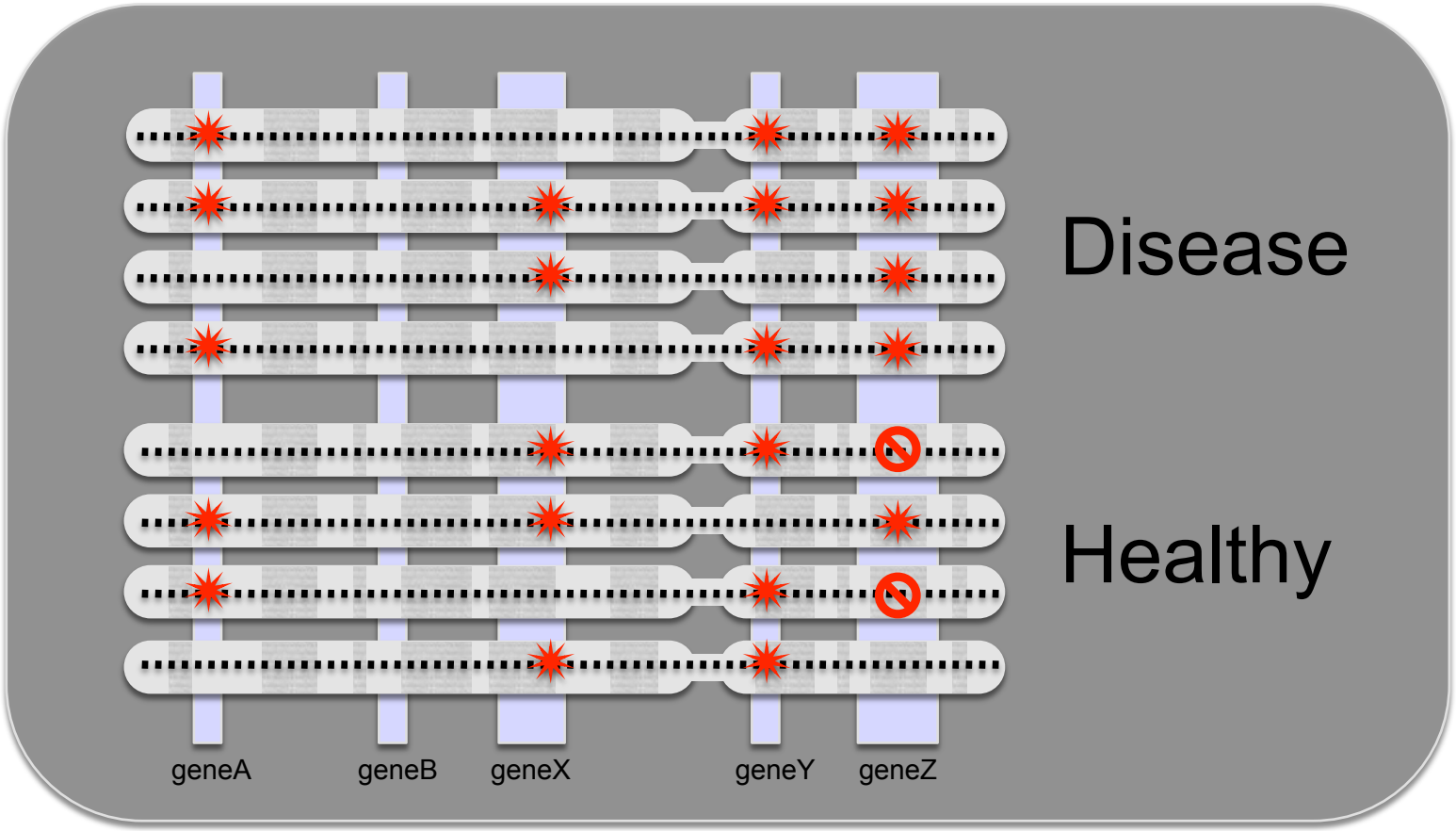
Variant calling - Population



Missing data

- Low/no sequence coverage
- Low base qualities
- Variant callers typically emit only variant sites
- What happens when we don't distinguish between missing data and reference sites
- Annotating no-calls can help solve this problem
- Population based variant calling provides this

Missing data



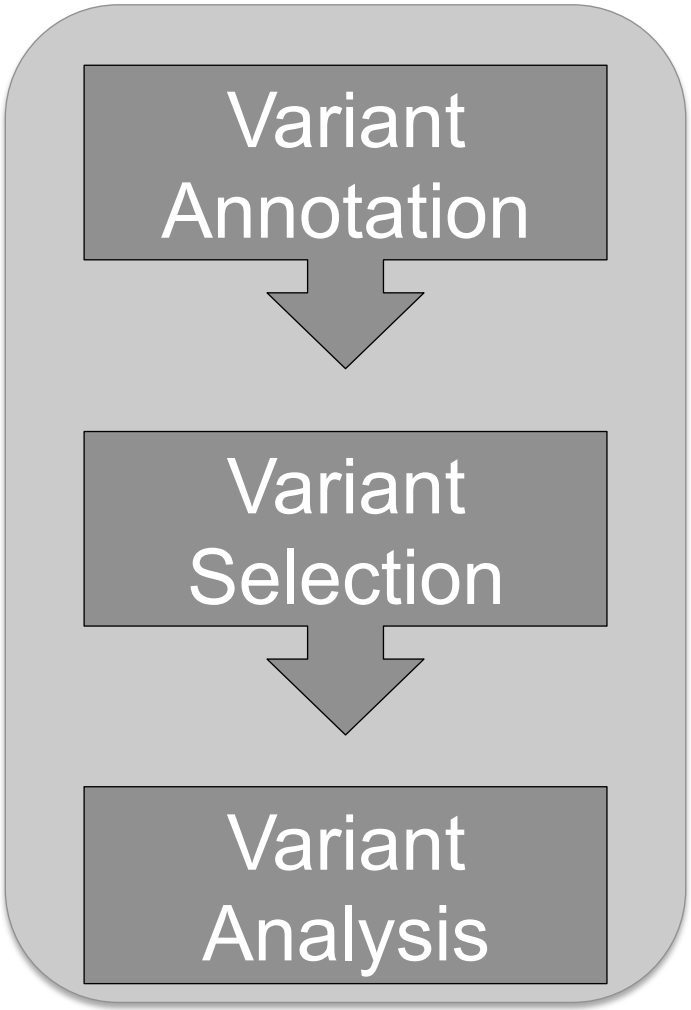
Variant Calling Format (VCF)

```
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth">
##contig=<ID=1,length=249250621>
```

CHROM	POS	ID	REF	ALT	QAUL	FILTER	INFO	FORMAT	INDIVIDUAL	INDIVIDUAL
1	745370	.	TA	T	1310.90	PASS	DP=210;	GT:GQ	0/1:99	0/1:99
1	749592	.	G	A	20	LowQual	DP=7;	GT:GQ	./.	1/1:6
1	749683	.	C	T	602.40	PASS	DP=69;	GT:GQ	./.	1/1:13
1	749856	.	C	T	261.37	PASS	DP=79;	GT:GQ	1/1:9	0/1:99
1	749899	.	G	A	302.28	PASS	DP=53;	GT:GQ	0/0:9	0/1:99
1	752566	.	G	A	1047.91	PASS	DP=47;	GT:GQ	1/1:30	0/1:29
1	752721	.	A	G	7625.90	PASS	DP=360;	GT:GQ	1/1:99	0/1:99

VAAST

- A tool for identifying disease genes and variants
- Collaboratatively developed
  - Mark Yandell (University of Utah)
  - Chad Huff (MD Anderson)
  - Martin Reese (Omicia Inc.)
- Inputs
  - Target genome variant files
  - Background genome variant files
  - Genomic features (gene models)
  - Genomic sequence
- Outputs a prioritized list of features (genes/ transcripts) associated with the disease genomes.
  - VAAST Score
  - P-value
  - Confidence interval

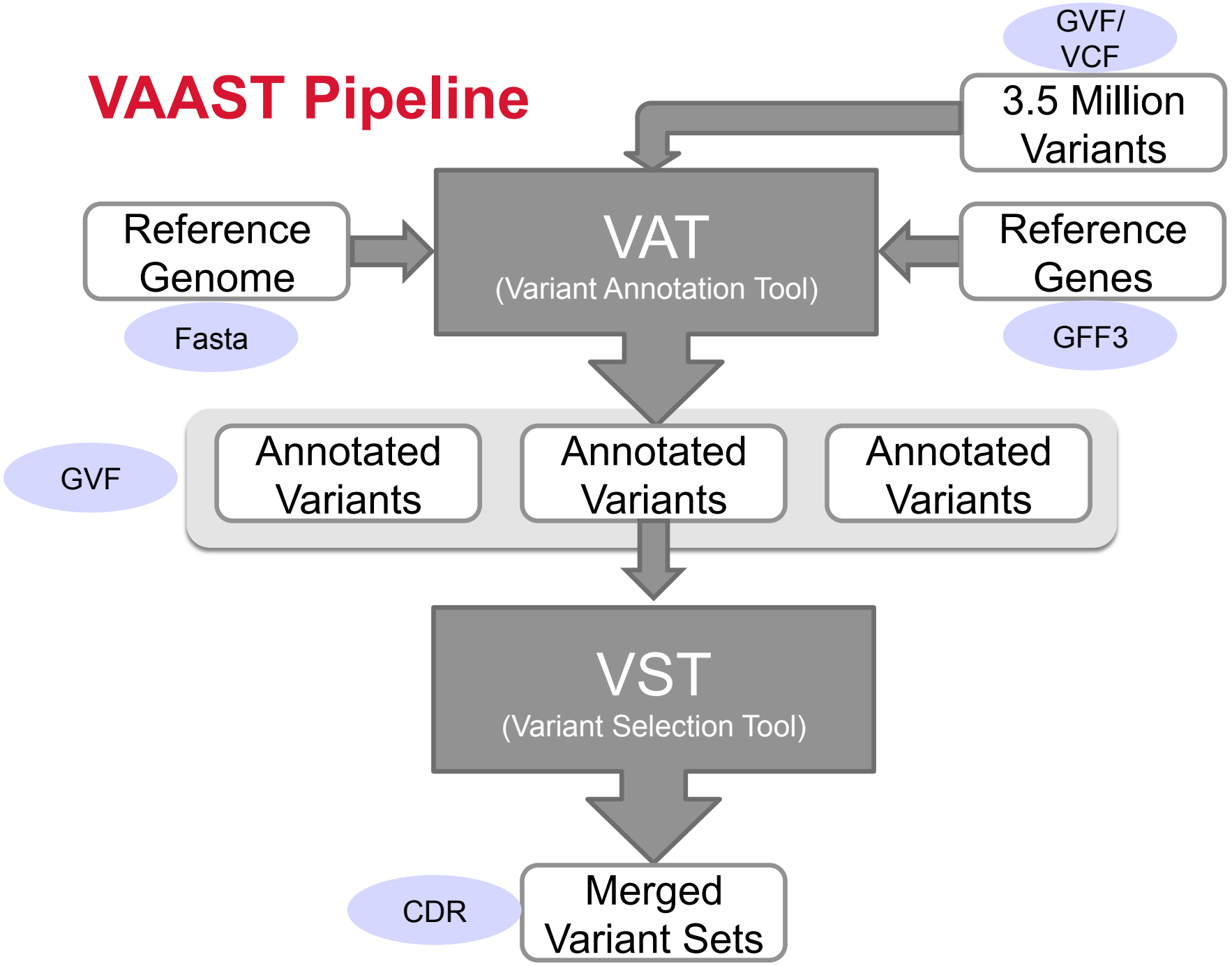


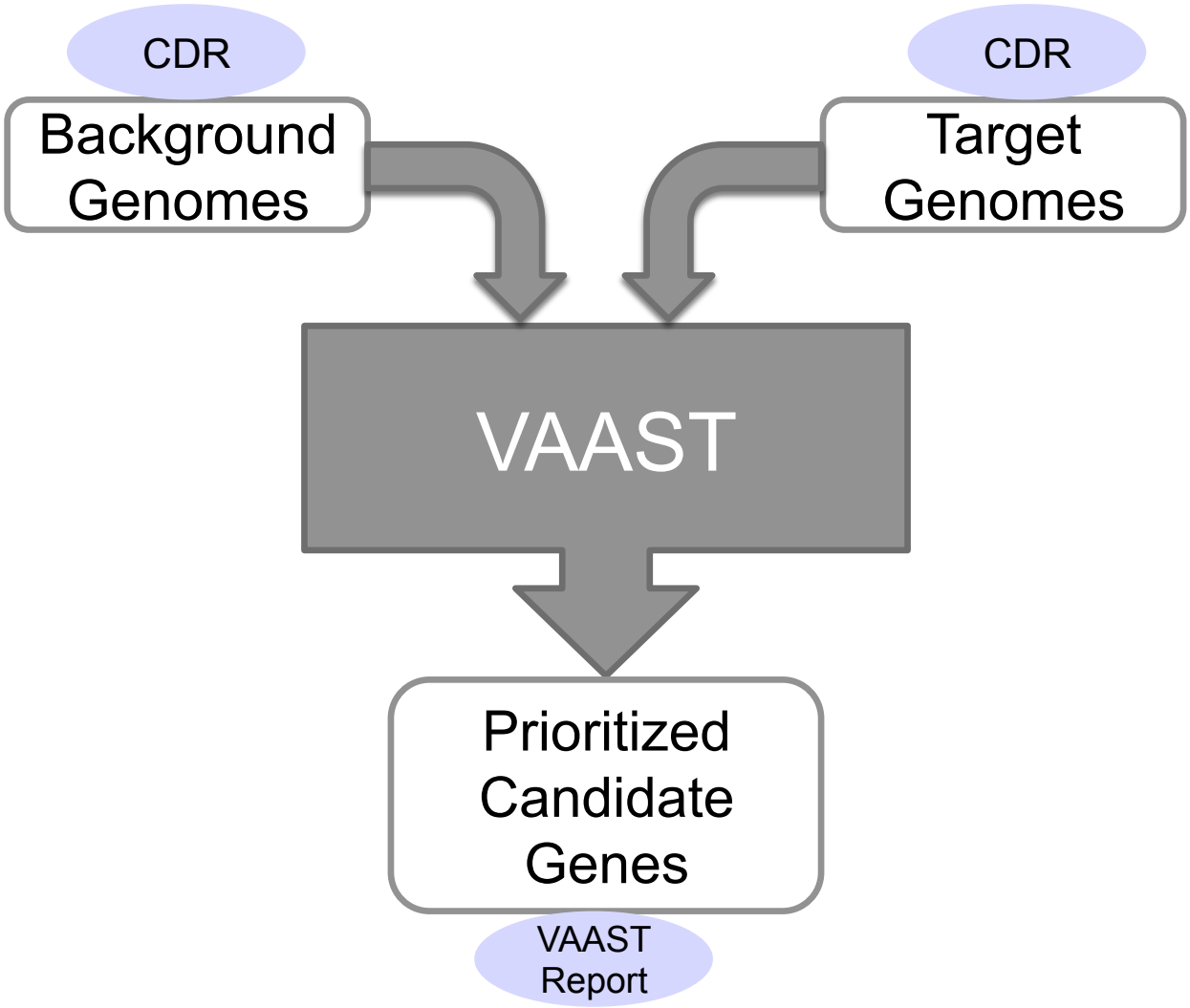
V*ariant*
A*nnotation*
T*ool*

V*ariant*
S*election*
T*ool*

V*ariant*
A*nnotation*
A*nalysis*
S*earch*
T*ool*

VAAST Pipeline




CDR
Background Genomes
CDR
Target Genomes
VAAST
Prioritized Candidate Genes
VAAST Report

VAAST File Types

- Input
  - Fasta - Genome
  - GFF3 - Genes
  - GVF - Variants
- Output
  - GVF – Annotated Variants
  - CDR – Population Variants
  - VAAST – Prioritized gene list

Fasta

```
>chr1
TAACAAAATAAGATCCAGAAACTTTCCATTAGCGTGGGGGTGACCATGAA
ATGCCTGGTCAAAAACCCGGGCACTGATTGTCATAACCATTATGCAACTG
GTGTTGCGTCCATCAGAATCTAGTTTAAGAATACTCTTCTCTCTATAGGA
GTCTTCGCGGCAGACCTAGCCTGCTCTGTGTCCTCCTGAAATGAAGGAAT
GTTCTCTCCCATTATTTCTTCTAACAGCTTGGTTAGCAAGCTCCGCCCTC
TTCTTTATCTGACCTTCTAACGACCTCACCAGATGTGTGAAGCAGCCCGG
CTCCATGTGTATCAGgcacgcacgcacacacgcacgcacacCAACCTGCA
AAGGAAATAACGGGGCAGCCCTGCAGTGTGAAAAGCAATGGGATTTTGTG
GGTTCCACCTCCTCACCTAAGCATCCCTGGTCTACGCTATGTCACGACCC
TCTGCTGAACCACGTCAGGGTGAACCCCNNNNNNNNNNNNN
```


www.sequenceontology.org
Google
Barry's Wiki
Yandell Lab Home
Yandbeck Wiki
The Sequence O...
Kronos
Campus Informa...
The Joss Bioinfo...
National Weathe...
The Sequence Ontology Project
Home
Browser
Wiki
GFF3
GVF
Resources
Software
About
Request A Term
Site Map
Welcome to the Sequence Ontology
This is the home page of the Sequence Ontology (SO). SO is a collaborative ontology project for the definition of sequence features used in biological sequence annotation. SO was initially developed by the Gene Ontology Consortium. Contributors to SO include the GMOD community, model organism database groups such as WormBase, FlyBase, Mouse Genome Informatics group, and institutes such as the Sanger Institute and the EBI. Input to SO is welcomed from the sequence annotation community. SO is also part of the Open Biomedical Ontologies library. Our aim is to develop an ontology suitable for describing the features of biological sequences. For questions, please send mail to the SO developers mailing list. For new term suggestions, please use the Term Tracker.
Introduction
The Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence. SO includes different kinds of features which can be located on the sequence. Biological features are those which are defined by their disposition to be involved in a biological process. Examples are binding_site and exon. Biomaterial features are those which are intended for use in an experiment such as aptamer and PCR_product. There are also experimental features which are the result of an experiment. SO also provides a rich set of attributes to describe these features such as "polycistronic" and "maternally imprinted".
News
October 2013 GVF was used in the clinical annotation of a whole genome, for precision medicine. Integrating precision medicine in the study and clinical treatment of a severely mentally ill person
September 2013 The SO development team and the Monarch Initiative held a collaborative workshop in Portland to align the SO and the GENO ontologies for better annotation of phenotypes. This meeting was partially funded by the Phenotype RCN .

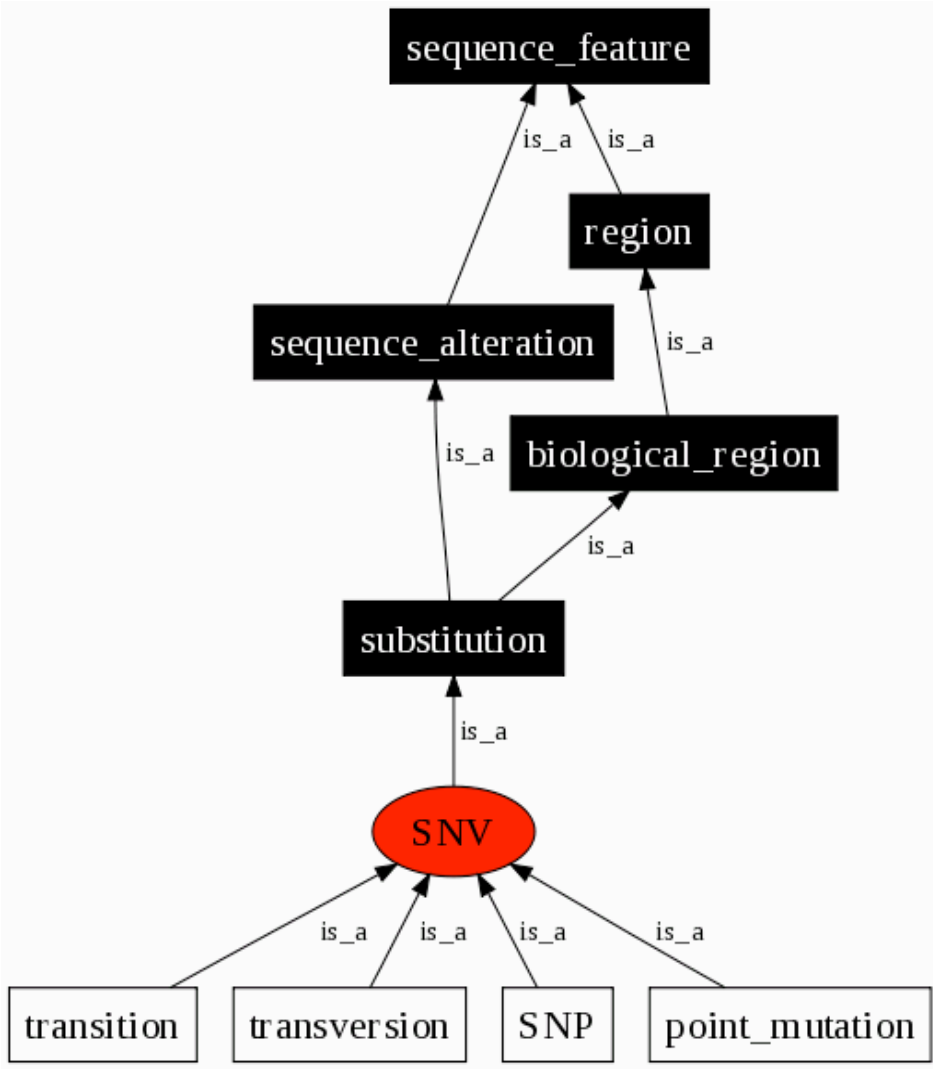
Variants, Features and Effects

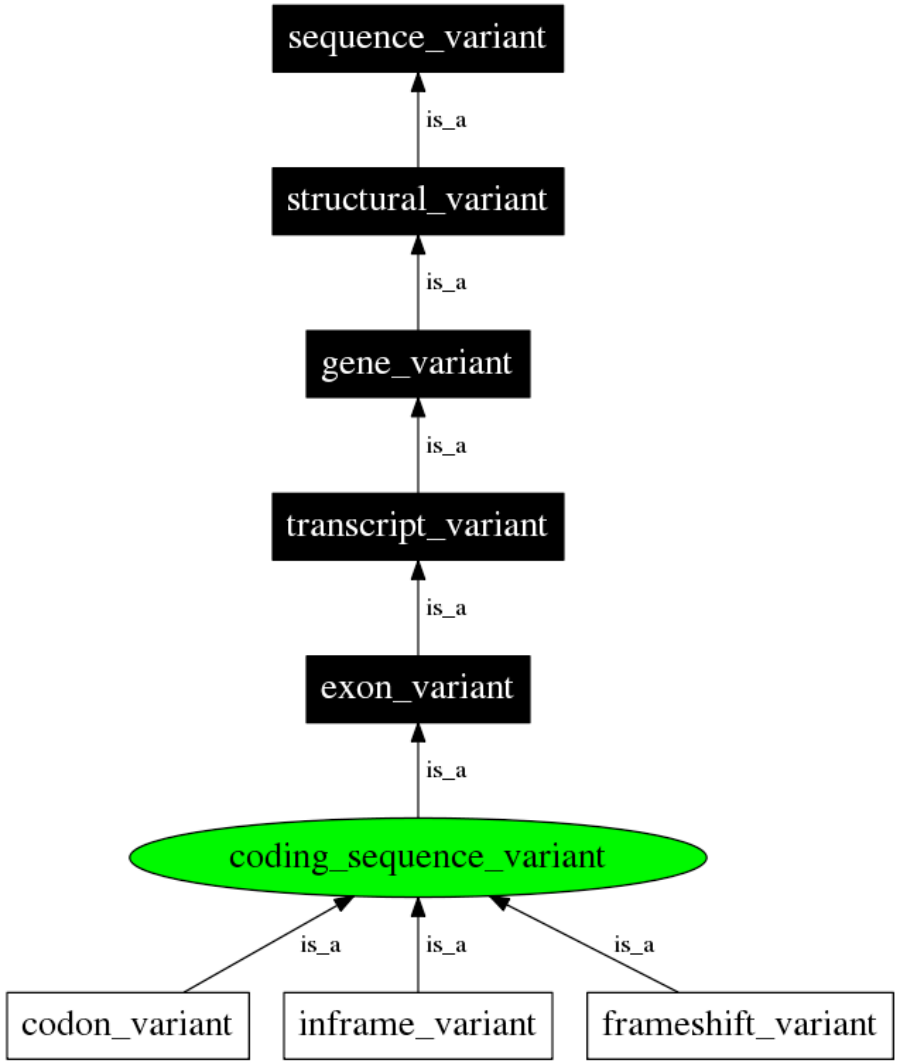
Variant Type
- sequence_alteration
- deletion
- insertion
- duplication
- inversion
- substitution
- **SNV**
- MNP
- complex substitution
- translocation

Feature Type
- sequence_feature
- gene
- mRNA
- exon
- **CDS**
- splice site
- ncRNA

Variant Effect
- sequence_variant
- gene_variant
- five_prime_UTR_variant
- three_prime_UTR_variant
- exon_variant
- splice_region_variant
- splice_donor_variant
- splice_acceptor_variant
- intron_variant
- coding_sequence_variant
- stop_retained
- stop_lost
- stop_gained
- synonymous_variant
- **missense_variant**
- amino_acid_substitution
- frameshift_variant
- inframe_variant

Sequence Ontology





Generic Feature Format (GFF3)

```
##gff-version 3
##sequence-region chr1 1 1497228
chr1	.	gene	1000	9000	.	+	.	ID=gene00001;Name=EDEN 3
chr1	.	mRNA	1050	9000	.	+	.	ID=mRNA00001;Parent=gene00001;Name=EDEN.1
chr1	.	mRNA	1050	9000	.	+	.	ID=mRNA00002;Parent=gene00001;Name=EDEN.2
chr1	.	mRNA	1300	9000	.	+	.	ID=mRNA00003;Parent=gene00001;Name=EDEN.3
chr1	.	exon	1300	1500	.	+	.	ID=exon00001;Parent=mRNA00003
chr1	.	exon	1050	1500	.	+	.	ID=exon00002;Parent=mRNA00001,mRNA00002
chr1	.	exon	3000	3902	.	+	.	ID=exon00003;Parent=mRNA00001,mRNA00003
chr1	.	exon	5000	5500	.	+	.	ID=exon00004;Parent=mRNA00001,mRNAchr1  mRNA00003
chr1	.	exon	7000	9000	.	+	.	ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
```

Seqid Source Type Start End Score Strand Phase Attributes

Genome Variation Format

```
##gvf-version 1.06
##genome-build GRCh37.1
##sequence-region chr16 1 88827254
chr16 UG SNV 291141 291141 33 + . ID=ID_1;Variant_seq=A,G;Reference_seq=G;
chr16 UG SNV 291360 291360 17 + . ID=ID_2;Variant_seq=G;Reference_seq=C;
chr16 UG SNV 302125 302125 67 + . ID=ID_3;Variant_seq=T,C;Reference_seq=C;
chr16 UG SNV 302365 302365 43 + . ID=ID_4;Variant_seq=G,C;Reference_seq=C;
chr16 UG SNV 302700 302700 75 + . ID=ID_5;Variant_seq=T;Reference_seq=C;
chr16 UG SNV 303084 303084 16 + . ID=ID_6;Variant_seq=G,T;Reference_seq=T;
chr16 UG SNV 303156 303156 90 + . ID=ID_7;Variant_seq=T,C;Reference_seq=C;
chr16 UG SNV 303427 303427 52 + . ID=ID_8;Variant_seq=T,C;Reference_seq=C;
chr16 UG SNV 303596 303596 66 + . ID=ID_9;Variant_seq=T,C;Reference_seq=C;
```

Seqid Source Type Start End Score Strand Phase Attributes

Variant Annotation Tool (VAT)

- Adds functional annotation of the effect of sequence alterations (SNV, insertion,deletion) on sequence features (genes, mRNA)
- Takes as input a reference sequence (Fasta file), and set of gene models (GFF3 file) and a set of variants (GVF)
- Produces an annotated GVF file as output

```
VAT -a genome.fasta -f gene.gff3 variants.gvf > variants.vat.gvf
```

Variant_effect attribute

- Describes the effect of a sequence alteration on a sequence feature
  - The sequence_variant (the effect)
  - The Variant_seq allele index (which allele causes this effect)
  - The sequence_feature (what type of feature is affected)
  - The feature IDs (which features are affected

```
Variant_seq=A,T;
Variant_effect=missense_variant 0 mRNA NM_001160184 NM_032129;
```

Variant Selection Tool (VST)

- Applies complex set operations (intersection, union etc) to GVF files and produces a condensed representation of the genotypes.
- Takes as input a description of the set operation and a group of GVF files.
- Outputs population genotypes in CDR format.

```
VST -o 'I(0,1)' exome1.gvf exome2.gvf > affected.cdr

VST -o 'C(0,U(1,2))' kid.gvf mom.gvf dad.gvf > denovo.cdr
```

VST set operations

- **(U)nion**: All variants in all files.
- **(I)ntersection**: Variants shared by all files.
- **(C)omplement**: The left relative complement or variants unique to the first file (set).
- **(D)ifference**: The symetric difference or variants unique to any one file (set).
- (**S)hared**: Variants shared by n files. S(n,0..2);

  'S(">2",0..2)').
  - = Exactly n files share the variant.
  - > Greater than n files share the variant.
  - < Less than n files share the variant.

VAAST Condenser File (CDR)

Seqid	Start	End	Type	Effect	Reference	Genotypes	
chr1	877831	877831	SNV	missense_variant	T|W	0-3|C:C|R:R	
chr1	881627	881627	SNV	synonymous_variant	G|L	0,2-3|A:A|L:L	1|A:G|L:L
chr1	881918	881918	SNV	missense_variant	G|S	2|A:G|L:S	
chr1	887801	887801	SNV	synonymous_variant	A|T	1|A:G|T:T	0,2-3|G:G|T:T
chr1	888639	888639	SNV	synonymous_variant	T|E	0|C:C|E:E	1|C:T|E:E
chr1	888659	888659	SNV	missense_variant	T|I	0,1-2|C:C|V:V	3|C:T|V:I
chr1	889238	889238	SNV	missense_variant	G|A	1|A:G|V:A	
chr1	897325	897325	SNV	synonymous_variant	G|A	0,2|C:C|A:A	1|C:G|A:A
chr1	897738	897738	SNV	synonymous_variant	C|L	1|C:T|L:L	
chr1	900505	900505	SNV	synonymous_variant	G|V	3|C:C|V:V	2|C:G|V:V
chr1	900972	900972	SNV	3_prime_UTR_variant	T	0,2|G:G	1|G:T
chr1	901023	901023	SNV	3_prime_UTR_variant	T	0,2-3|C:C	1|C:T

```
##      GENOME-LENGTH   914121104
##      GENOME-COUNT    7
##      GENDER  F:0-1,3       M:2
##      FILE-INDEX      0       A12.vat.gvf
##      FILE-INDEX      1       B34.vat.gvf
##      FILE-INDEX      2       C56.vat.gvf
##      FILE-INDEX      2       D78.vat.gvf
```

VAAST

- Scores and prioritizes features in a probabalistic fashion for their likelihood of being associated with a disease phenotype.
- Takes as input a set of gene models (GFF3), a set of variants for background/healthy genomes (CDR) and a set of variants for target/disease genomes (CDR).

```
VAAST -m lrt -o Output_name genes.gff3 background.cdr target.cdr
```

VAAST Uses Variant Frequencies in a Probabilistic Fashion

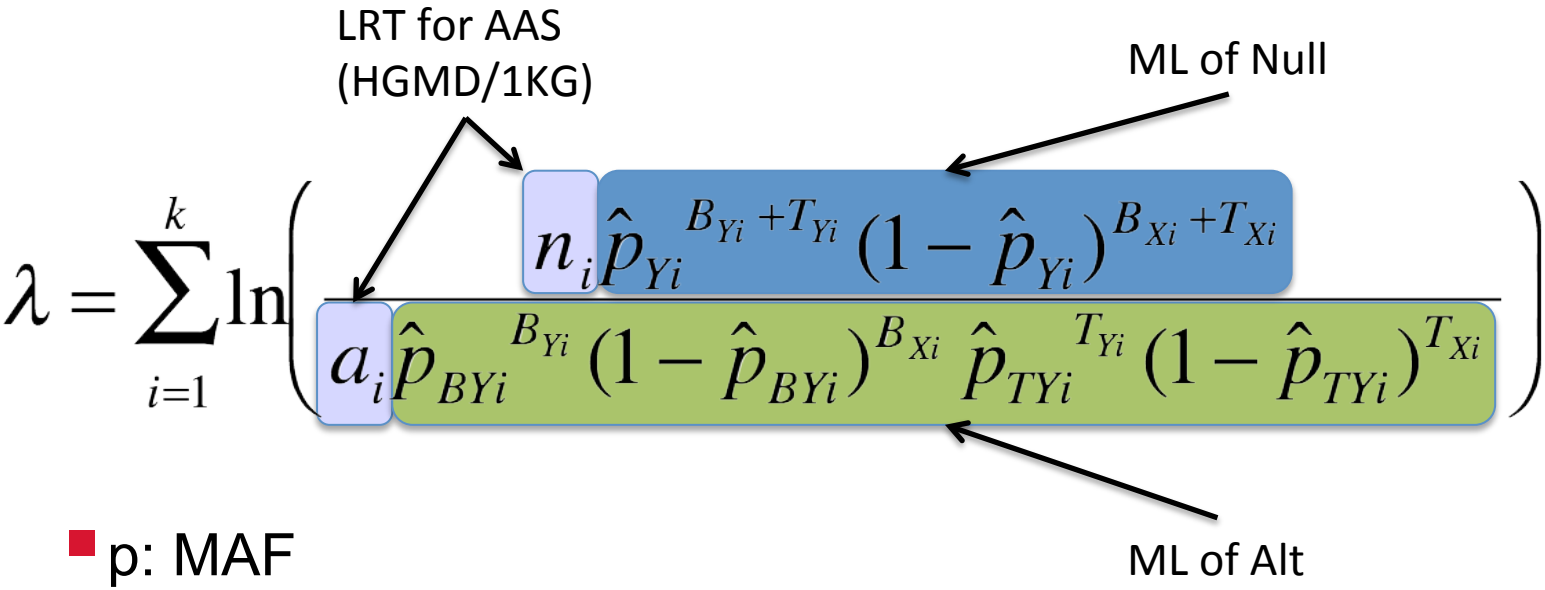
Composite Likelihood Ratio Test

$$\lambda = \ln\left(\frac{L_{Null}}{L_{Alt}}\right)$$

Maximum Likelihood of the Null Model (*No Difference*)

Maximum Likelihood of the Alternate Model (*There is Difference*)

VAAST Uses Variant Frequencies in a Probabilistic Fashion



- p: MAF
- B/T: Backgroud/Target Allele Counts
- X/Y: Minor/Major Allele

Conservation-controlled amino acid scoring matrix – CASM

- The amino acid severity parameter is adjusted to account for the degree of phylogenetic conservation at a site.
- Conservation data comes from PhastCons scores which estimate a probability of a site being under negative selection (conserved).
- Each AAS type is scored at all sites with a PhastCons score of 0 and 1 (the two extremes).
- The AAS severity parameter at any PhastCons score is then lineraly interpolated between those extremes.
- The effect is that the AAS severity paremeter dimishes with diminishing conservation.

VAAST Uses Variant Frequencies in a Probabilistic Fashion

- VAAST gives us the likelihood of the composite genotype of a given gene in the target given the background.
- Do allele and AAS frequencies differ between background and target genomes within a given gene or feature?
- Composite likelihood calculation assumes independence across sites. To control for LD, statistical significance is estimated by permutation test.
- Multiple test correction for number of features (~20,000) is two orders of magnitude better than for the number of variants (~3,500,000).

VAAST: highly accurate variant prioritization

	Percent Judged Deleterious					
	Dream tool	SIFT[2]	ANNOVAR[3]	PolyPhen2	Mutation Taster	VAAST
Disease alleles[A]	100%	58%	71%	84%	84%	99%
Healthy alleles[B]	0%	12%	1%	16%	16%	10%
Accuracy (Sn + Sp)/2	100%	80%	88%	86%	86%	95%

[A]1454 high quality, published disease-causing/predisposing OMIM alleles[1]
[B]1454 Variants randomly selected from 5 different healthy CEU individuals' genomes

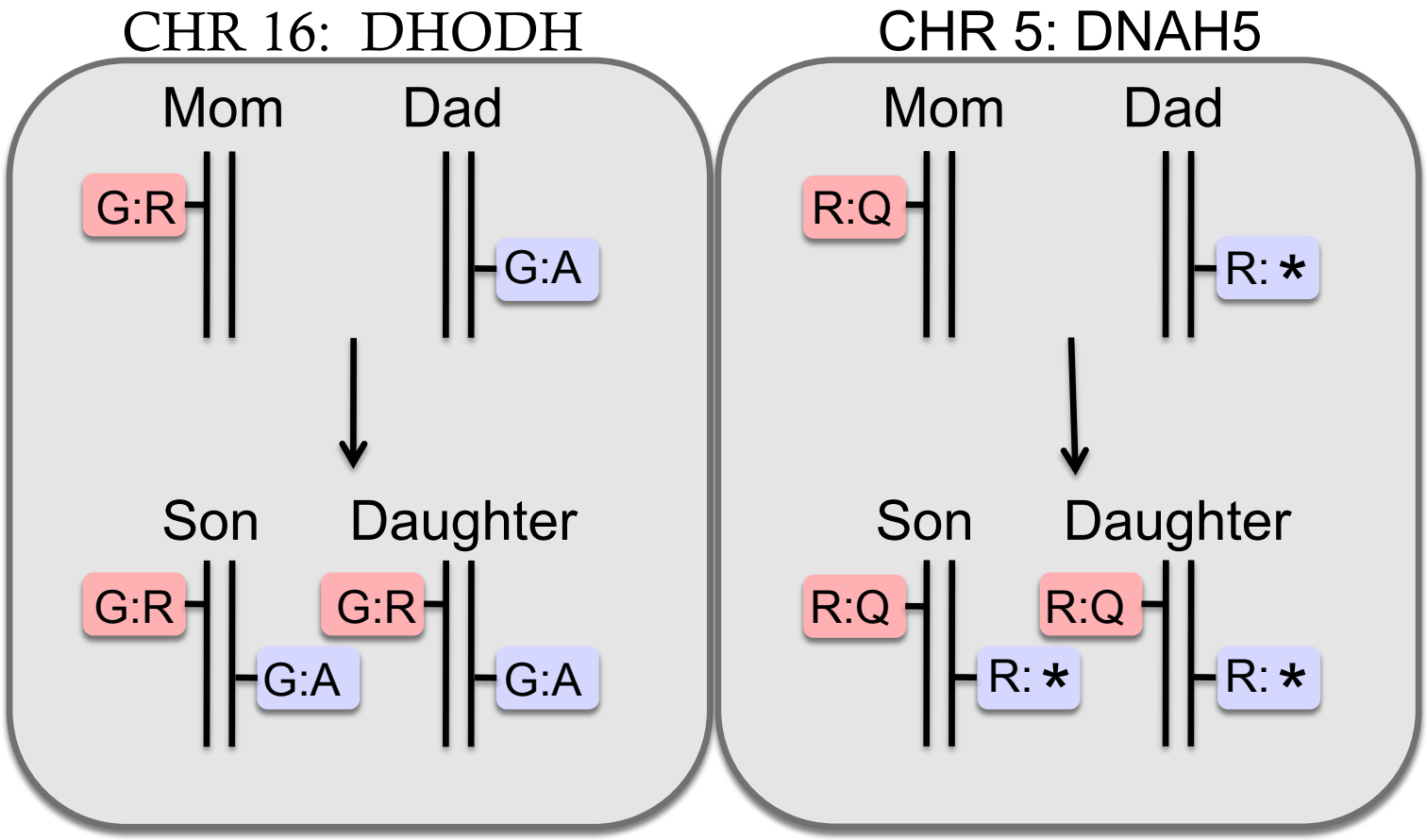
VAAST Filters

- Inheritance model
  - Dominant: Only score one allele per feature
  - Recessive: Only score two alleles per feature
- Locus heterogeneity
  - Require all affected individuals to have a scoring allele
- Complete penetrance
  - Don't score a feature if anyone in the background shares it's scoring alleles.
- Rate/PAR
  - Hold the MAF in the background/target below the given value.

Dangers of filtering

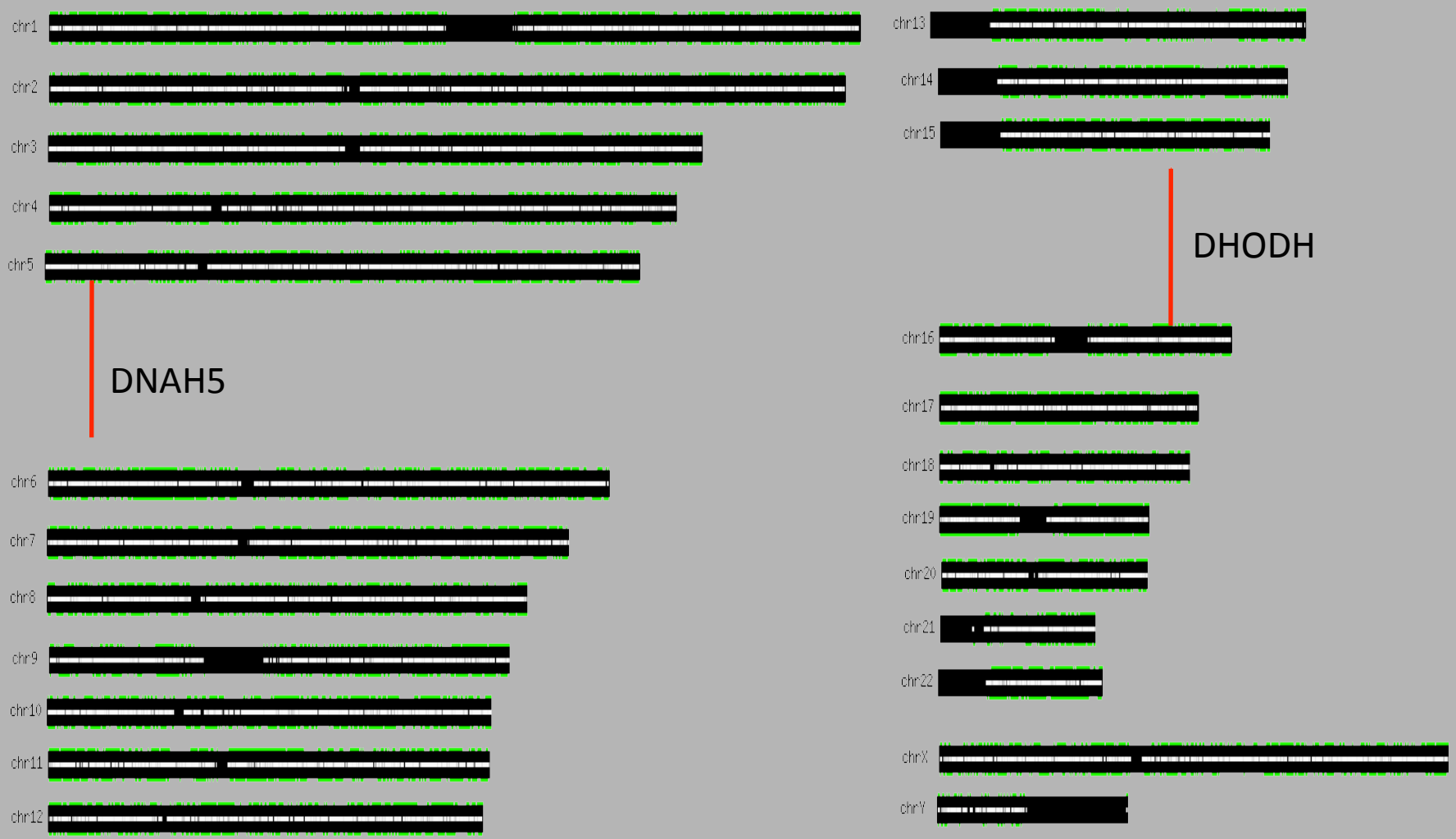
- Filters are binary – you either pass or fail.
- Filters (usually) don't consider missing data.
- Filters aren't able to incoporate multiple factors into the equation.

Alleles Responsible for Miller Syndrome in Utah Kindred



•Ng *et al,* Nature Genetics 42, 30–35, 2010
•Roach, *et al*, Science 328 636, 2010

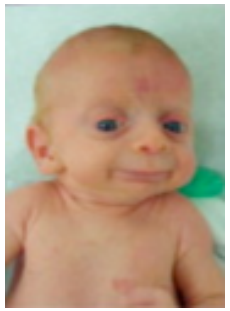
Schematic of VAAST Analysis of Utah Miller Kindred Using a Single Quartet



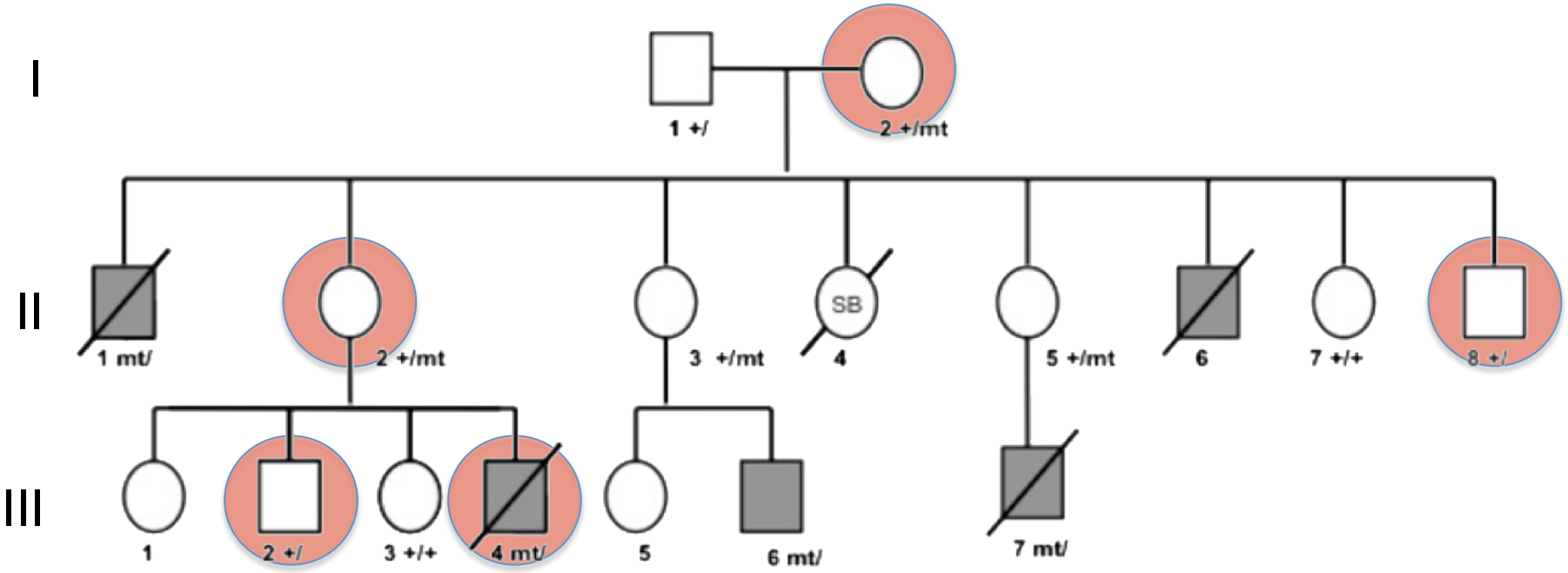
A rare X-linked mendelian disorder

- A Utah family coming to the University Hospital for 20+ years
- About half of the male offspring die around 1 year of age
- Aged appearance
- Craniofacial anomalies
- Hypotonia
- Global developmental delays
- Cardiac arrhythmias



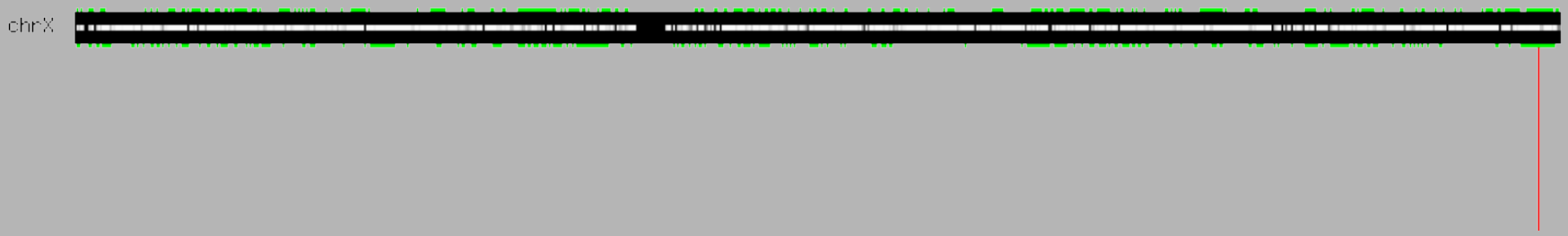



Four Affected Boys over Two Generations



Identifying Candidate Genes

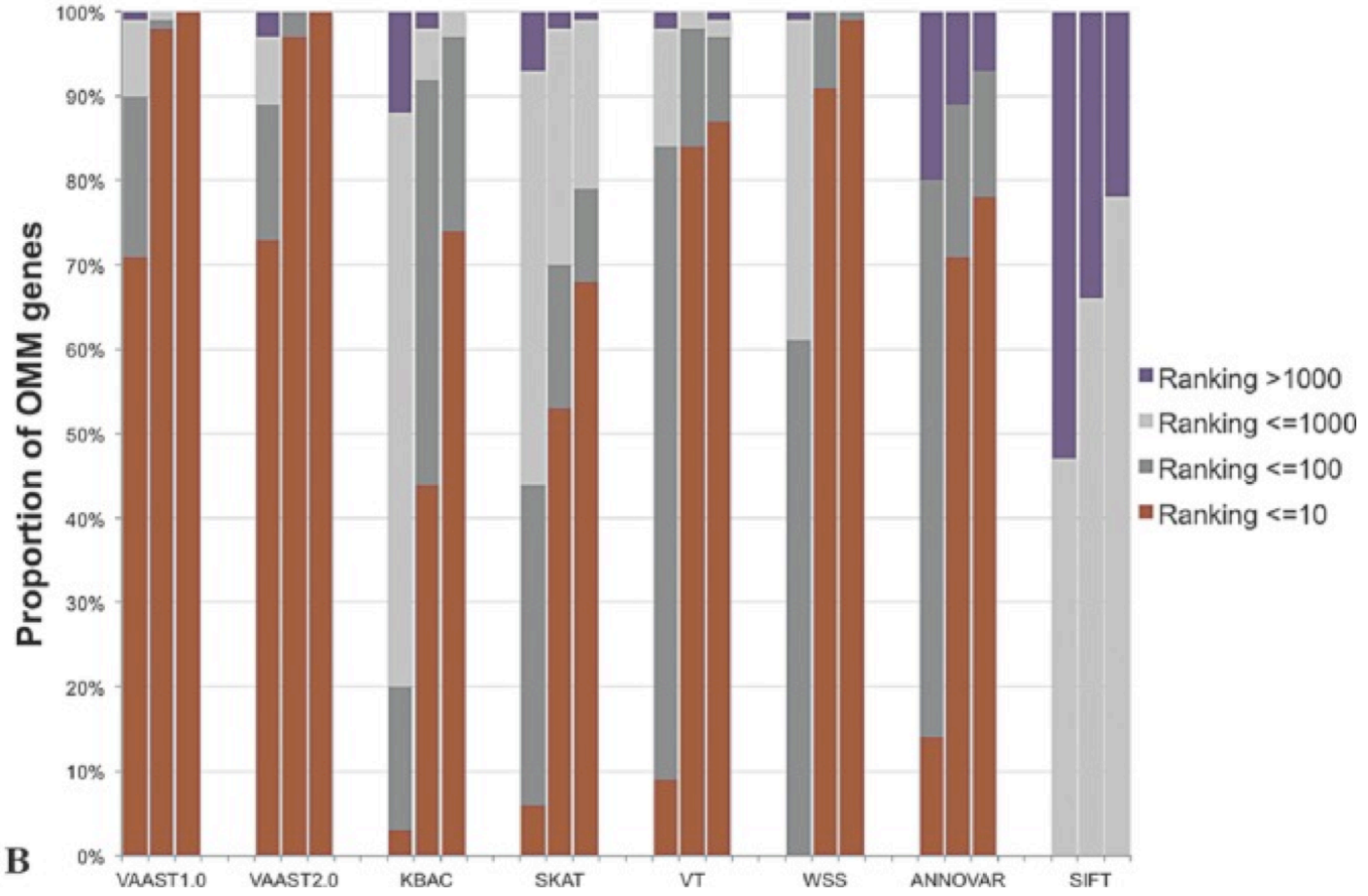
- VAAST identified NAA10 as candidate gene
  - Run entire pipeline in an afternoon
  - 3 candidate genes (NAA10 ranked 2) proband only
  - 1 candidate gene (NAA10) with pedigree



VAAST benchmark – 100 OMIM diseases

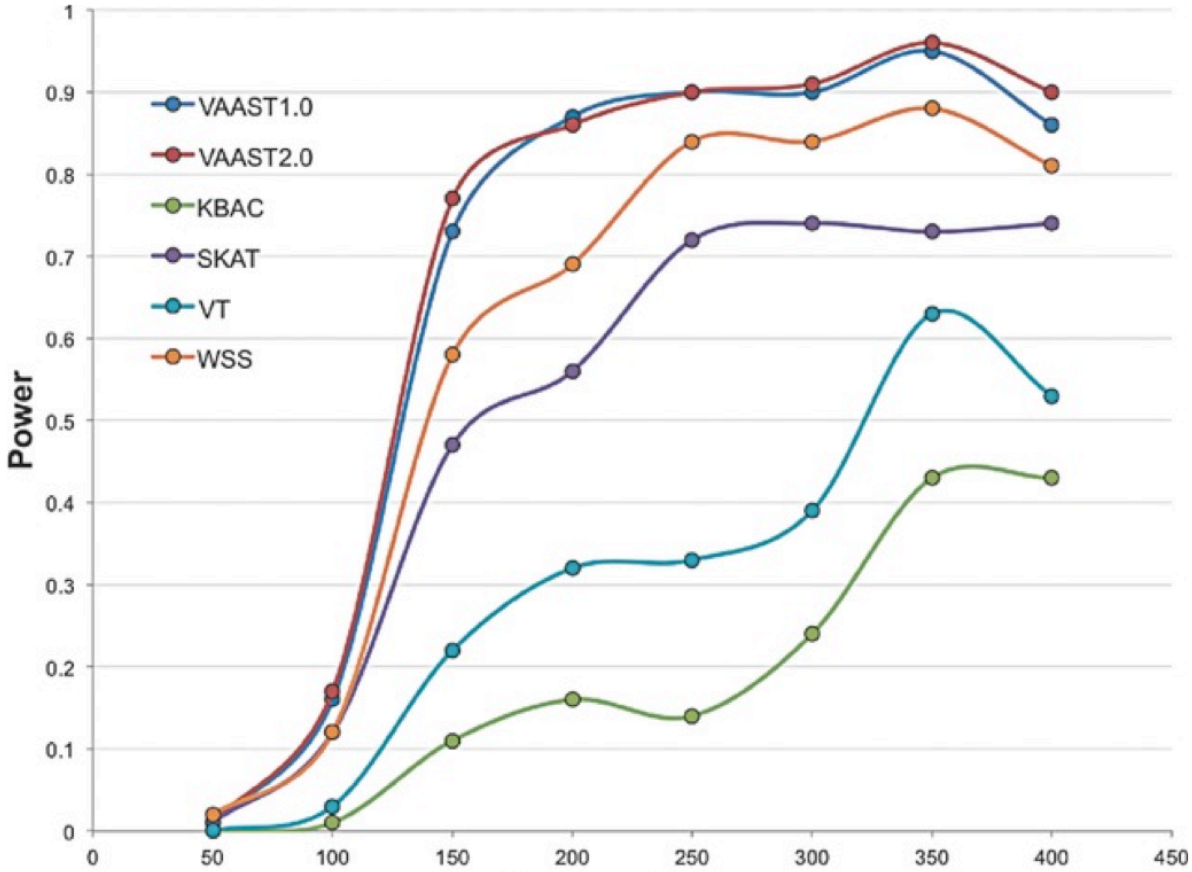
- Randomly choose known disease genes from OMIM
- Randomly insert one or more published disease causing variants for that gene into a personal exome
- Assay the ability of VAAST, SIFT1 and ANNOVAR2 to identify the disease gene in a genome-wide screen
- Repeat for 100 different genes under a variety of different scenarios, e.g. dominant, recessive, various case cohort sizes etc.

VAAST benchmark – 100 OMIM Diseases



Hu et al. Genet Epidemiol. 2013

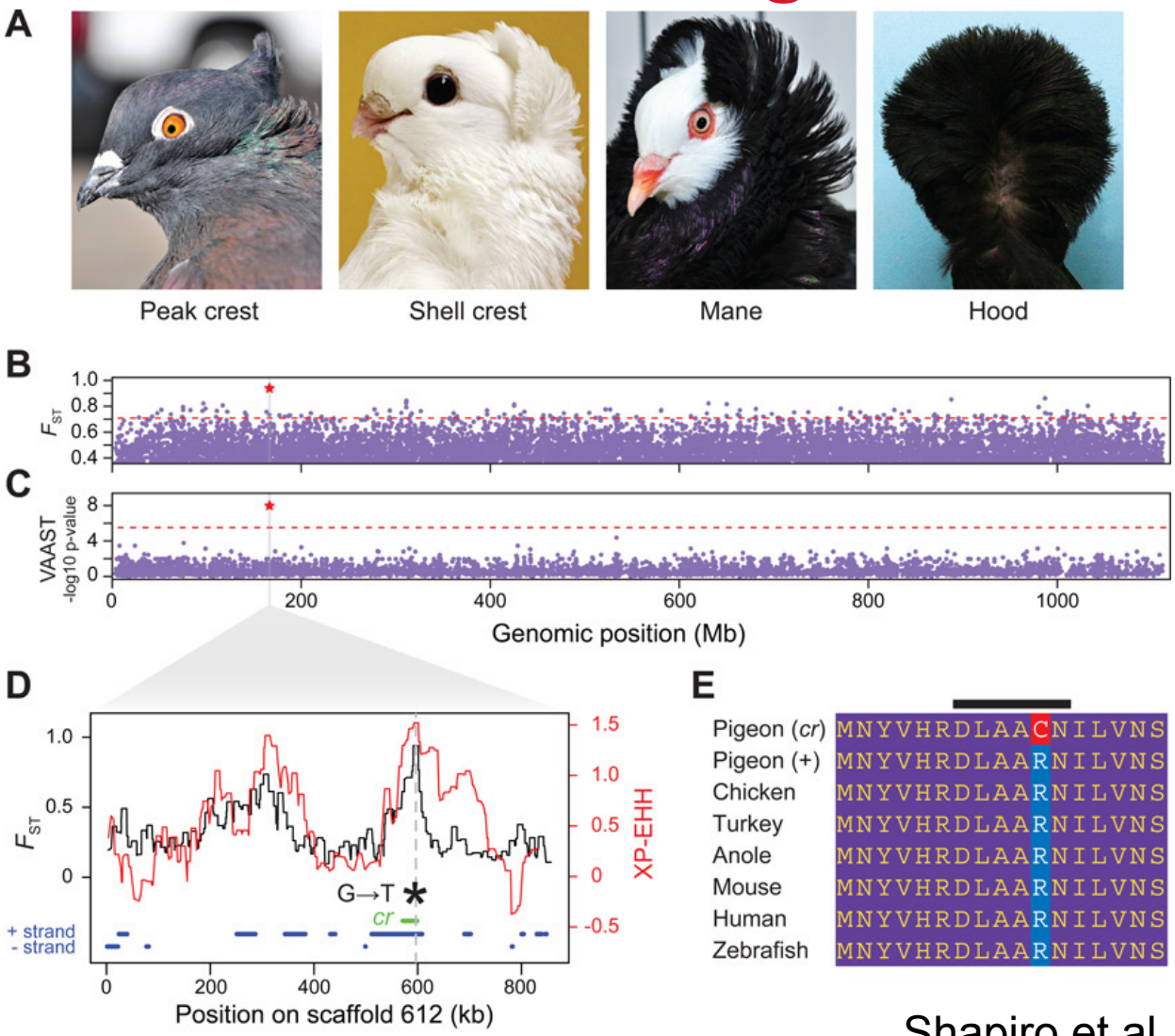
VAAST benchmark – common disease



Power comparisons over published LPL
(Power is calculated based on 100 bootstraps)

Hu et al. Genet Epidemiol. 2013

VAAST in non-human organisms

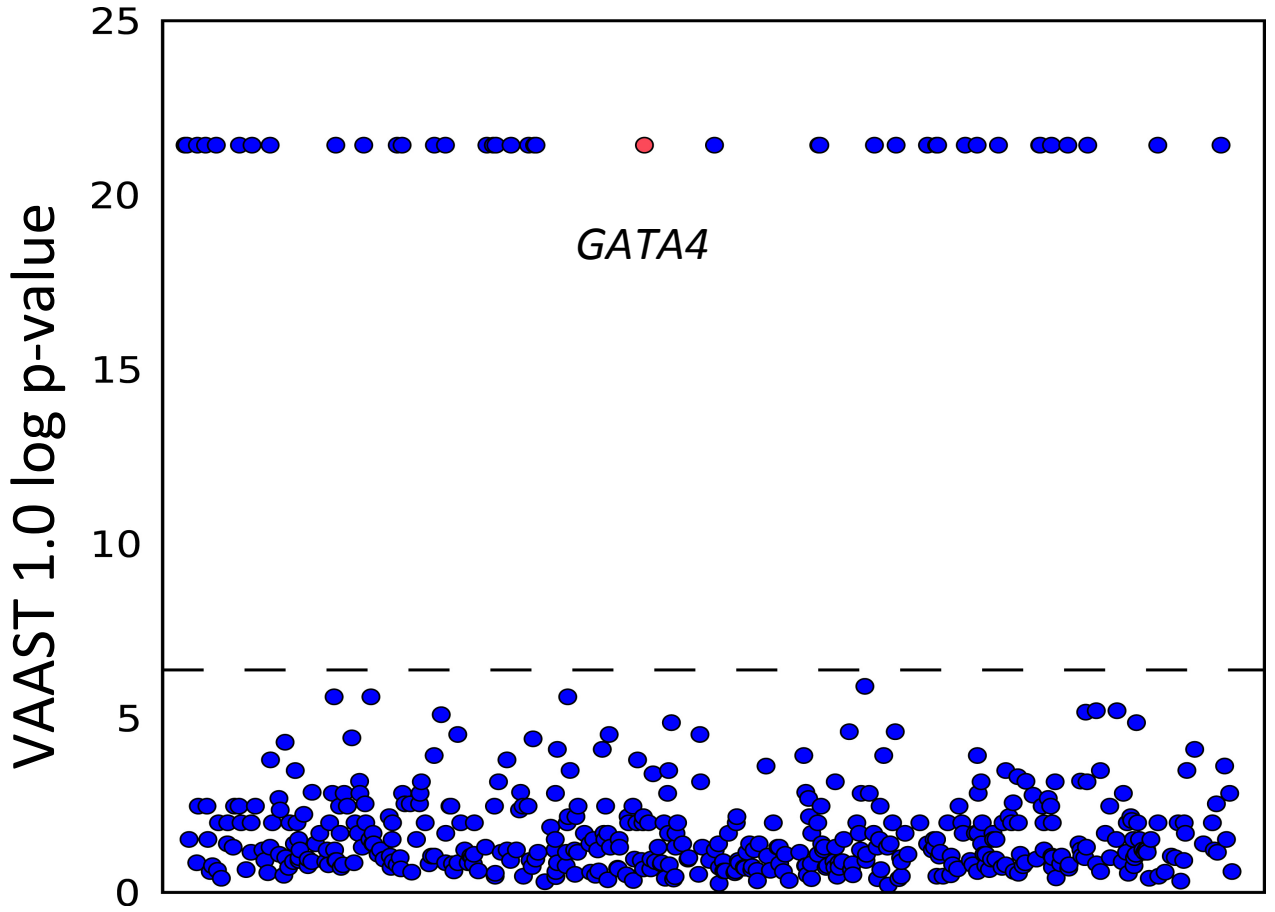


Shapiro et al. Science. 2013

pVAAST for pedigree analyses

- Extends VAAST to incorporate family data (pedigrees)
- pVAAST performs linkage analysis by calculating a gene-based LOD designed for NGS
- The LOD score at each locus is incorporated directly into the CLRT.
- In large-scale simulation studies and re-analysis of known disease pedigress, pVAAST had significantly higher statistical power compared other tools – including VAAST.

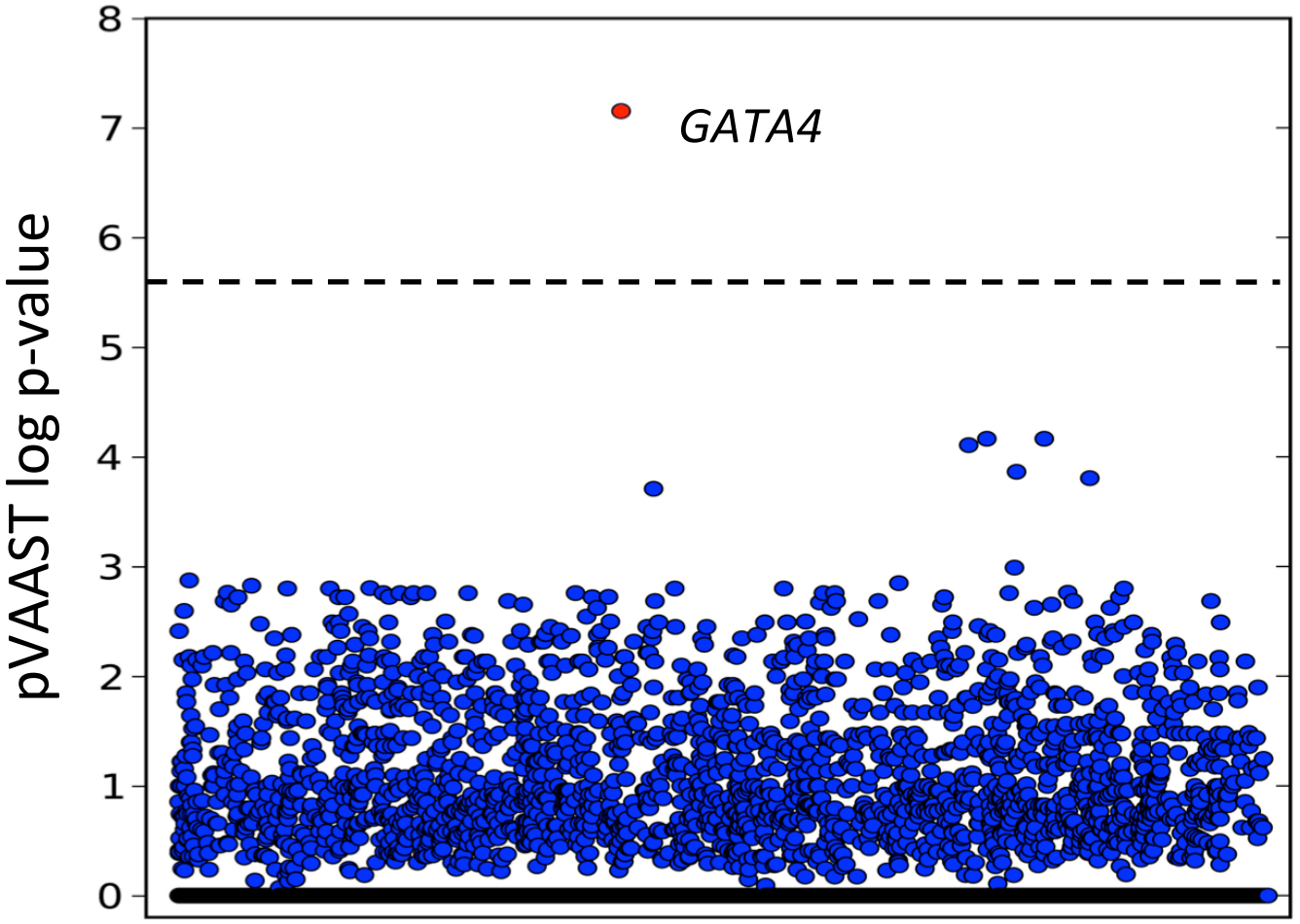
VAAST 1.0 cardiac septal defect



Run time: 24 hours on 42 CPUs

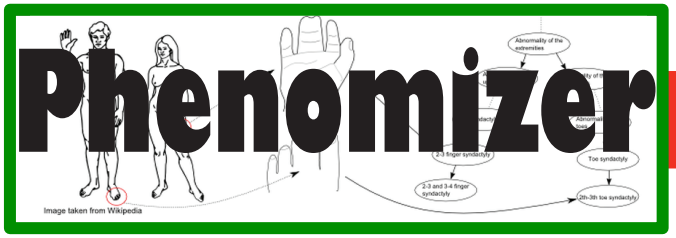
Hu et al submitted, Original study: Garg et al. Nature. 2003

pVAAST - cardiac septal defect



Run time: 4 hours on 42 CPUs

Phevor



Gene Properties

- Gene Ontology
  - biological process
  - molecular function
  - cellular component

Disease Phenotype

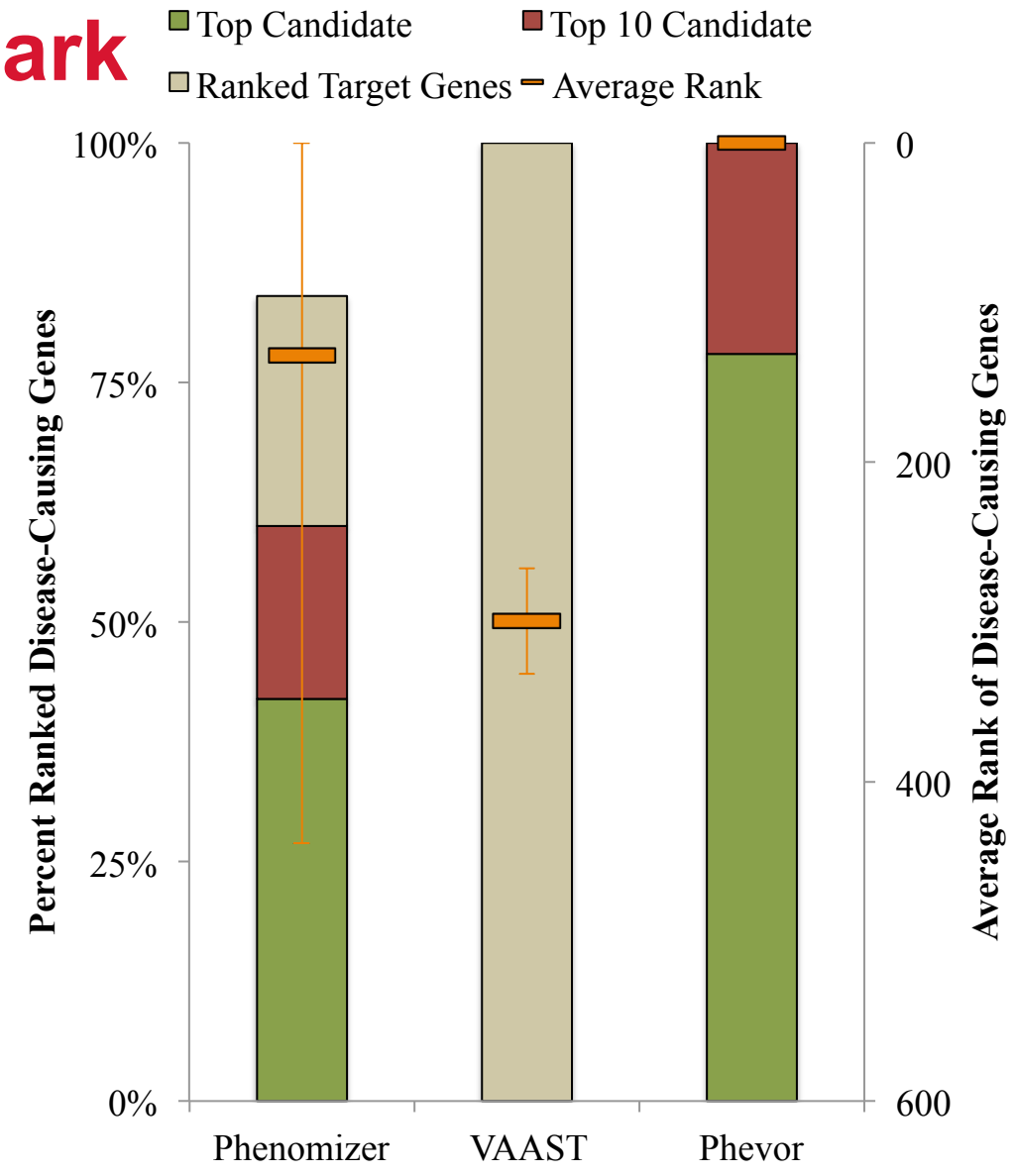
- Human Phenotype Ontology
- Mammalian Phenotype Ontology
- Disease Ontology

Candidate Gene Ranking

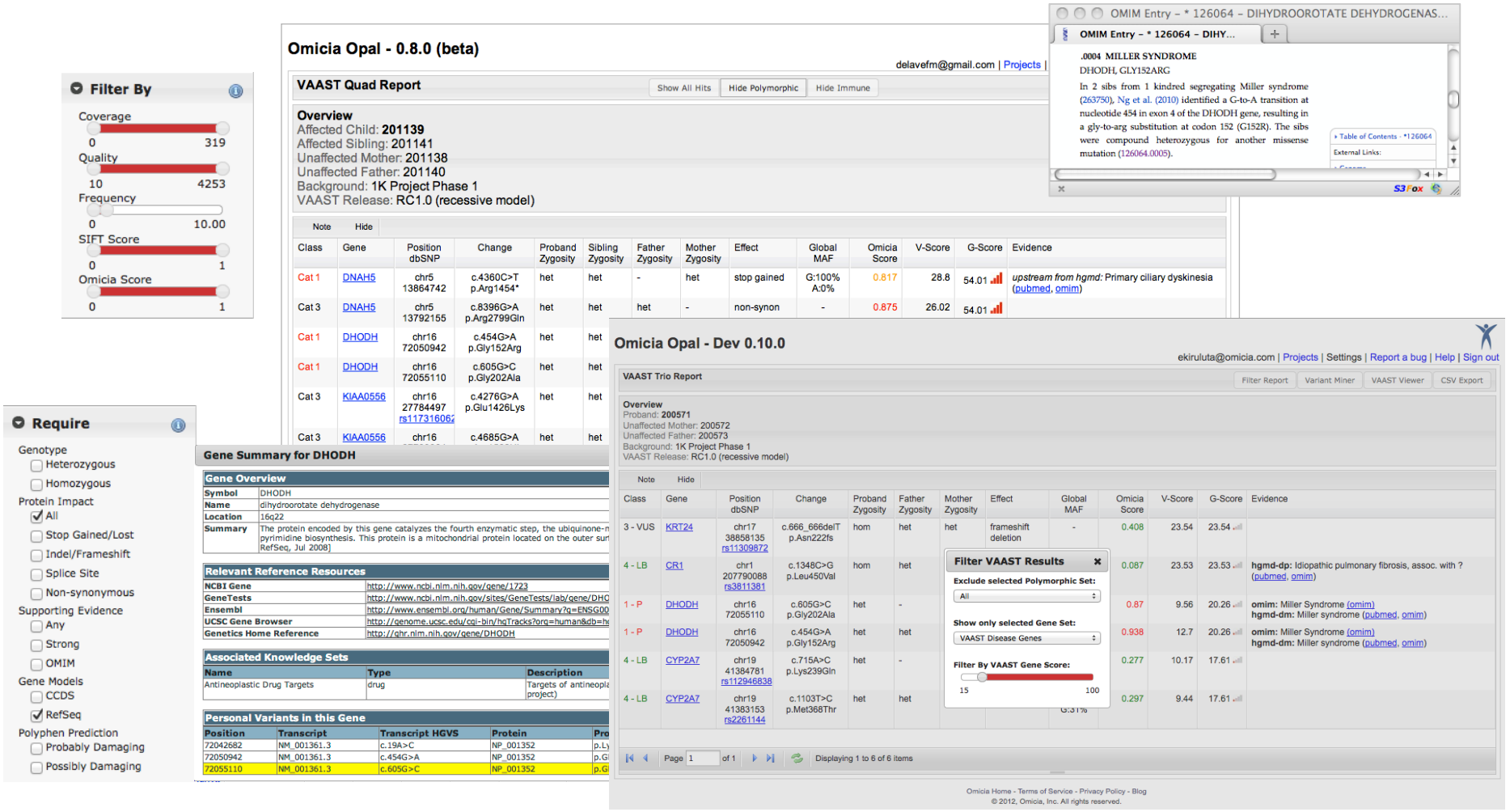
Phevor benchmark

- 50 Dominant Disease-Causing Variants
- 50 Corresponding Phenomizer Reports
- Spiked into a single healthy exome

- Average Ranks
  - Phenomizer – 134*
  - VAAST – 300
  - Phevor – 1.3

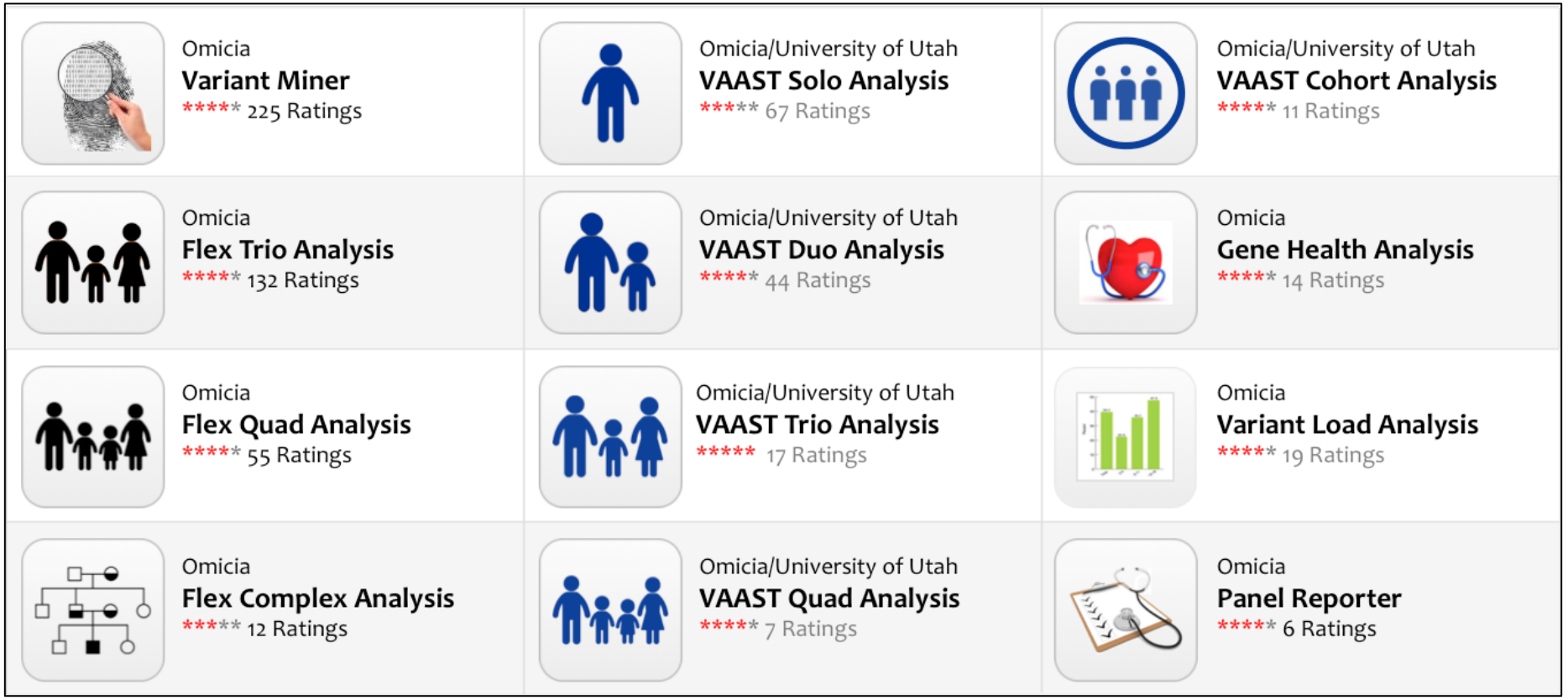


Omicia - Opal



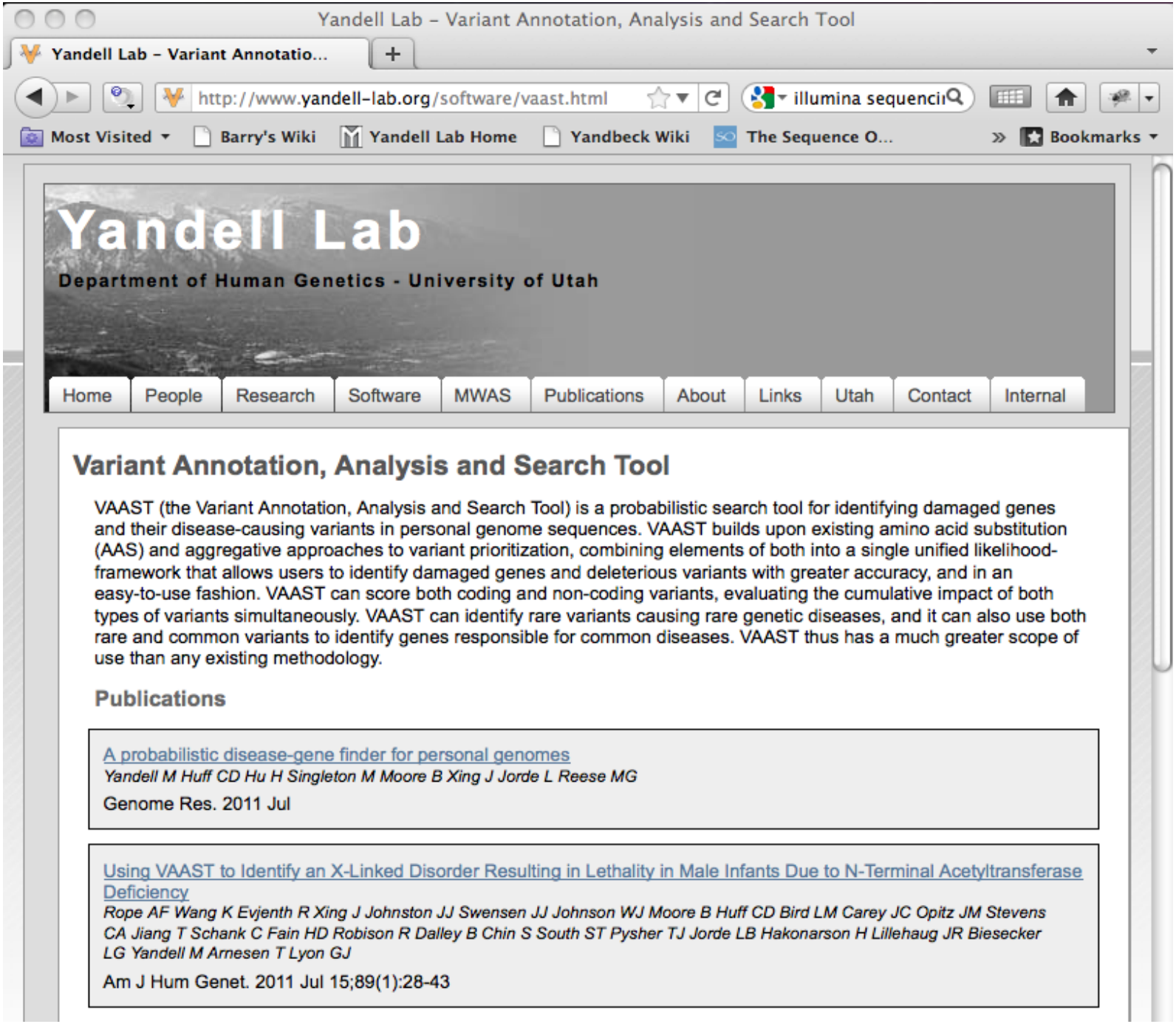
Development supported by NIH SBIR grants 1R4HG003667 to Omicia/Yandell, SBIR 1R44HG002991 to Omicia

VAAST in Opal



VAAST in Summary

- Probabilistic Disease Gene Finder
- Feature Based
- Both Allele and AAS Frequencies
- Considers the Inheritance Model
- As few as 1-2 target genomes can be sufficient to identify causative gene.
- Complete analysis pipeline
- Many parameters allow fine-grained control of analysis
- pVAAST and Phevor on the way for otherwise under-powered analyses.


Yandell Lab – Variant Annotation, Analysis and Search Tool
Yandell Lab – Variant Annotatio...
http://www.yandell-lab.org/software/vaast.html
illumina sequenci
Most Visited
Barry's Wiki
Yandell Lab Home
Yandbeck Wiki
The Sequence O...
Bookmarks
Yandell Lab
Department of Human Genetics - University of Utah
Home
People
Research
Software
MWAS
Publications
About
Links
Utah
Contact
Internal
Variant Annotation, Analysis and Search Tool
VAAST (the Variant Annotation, Analysis and Search Tool) is a probabilistic search tool for identifying damaged genes and their disease-causing variants in personal genome sequences. VAAST builds upon existing amino acid substitution (AAS) and aggregative approaches to variant prioritization, combining elements of both into a single unified likelihood-framework that allows users to identify damaged genes and deleterious variants with greater accuracy, and in an easy-to-use fashion. VAAST can score both coding and non-coding variants, evaluating the cumulative impact of both types of variants simultaneously. VAAST can identify rare variants causing rare genetic diseases, and it can also use both rare and common variants to identify genes responsible for common diseases. VAAST thus has a much greater scope of use than any existing methodology.
Publications
A probabilistic disease-gene finder for personal genomes
Yandell M Huff CD Hu H Singleton M Moore B Xing J Jorde L Reese MG
Genome Res. 2011 Jul
Using VAAST to Identify an X-Linked Disorder Resulting in Lethality in Male Infants Due to N-Terminal Acetyltransferase Deficiency
Rope AF Wang K Evjenth R Xing J Johnston JJ Swensen JJ Johnson WJ Moore B Huff CD Bird LM Carey JC Opitz JM Stevens CA Jiang T Schank C Fain HD Robison R Dalley B Chin S South ST Pysher TJ Jorde LB Hakonarson H Lillehaug JR Biesecker LG Yandell M Arnesen T Lyon GJ
Am J Hum Genet. 2011 Jul 15;89(1):28-43

Acknowledgements

VAAST Development

- Chad Huff
- *Hao Hu*
- Lynn Jorde
- *Edward Kirulata*
- *Marco Falcioni*
- *Barry Moore*
- Martin Reese
- Jinchuan Xing
- Mark Yandell

Yandell Lab

- Michael Campbell
- Daniel Ence
- Steven Flygare
- Carson Holt
- Brett Kennedy
- Zev Kronenberg
- Qing Li
- Gordon Lemmon
- Barry Moore
- EJ Osbourne
- Scott Watkins
- Mark Yandell

Omicia

- Marco Falcioni
- Edward Kirulata
- Martin Reese
- Jeff Rule
- Charlene Rigby
- Andrew Gao

Sequence Ontology

- Mike Bada
- Colin Batchelor
- Karen Eilbeck
- Barry Moore
- Shawn Reynearson

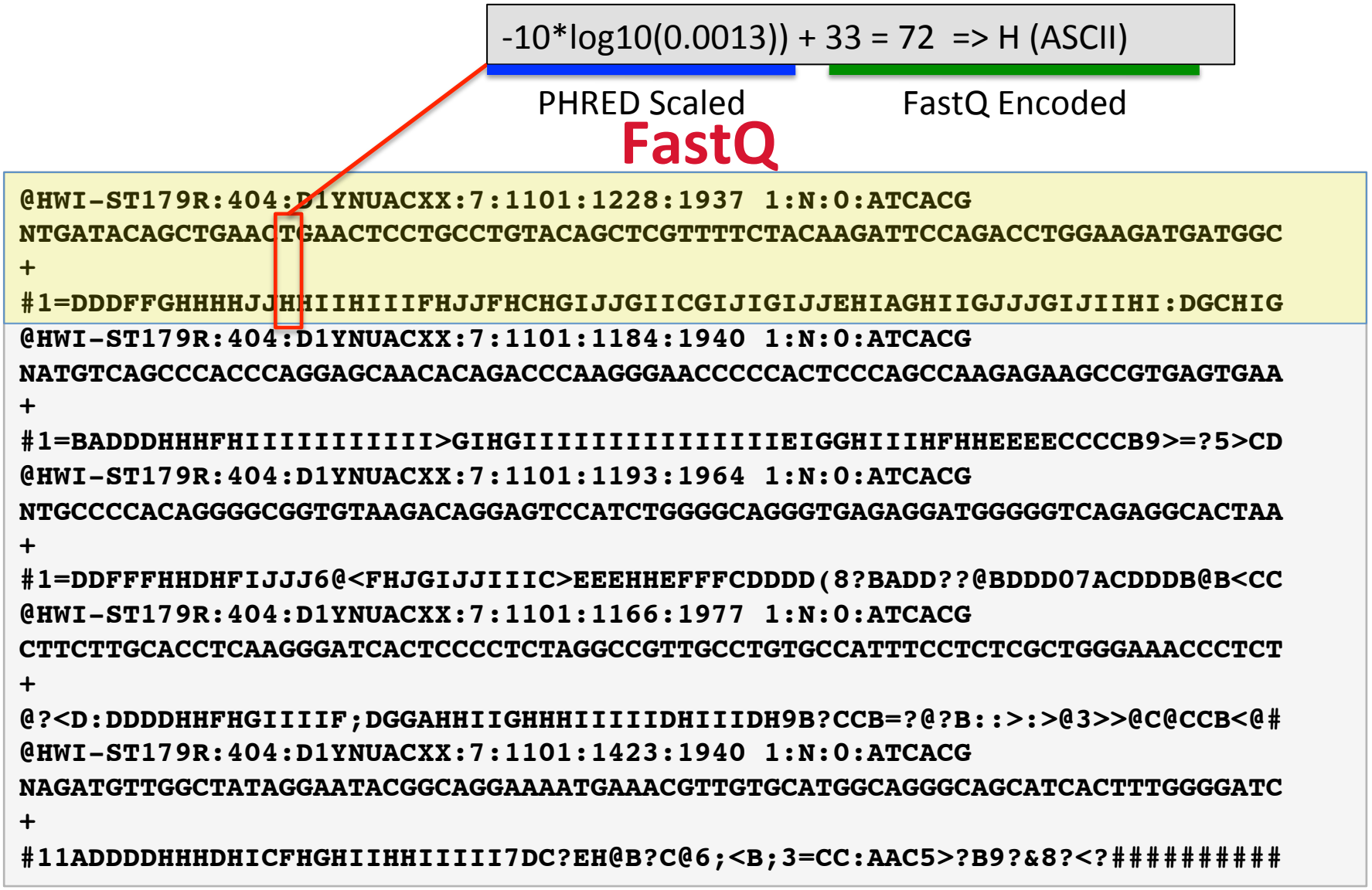
Funding

- NHGRI

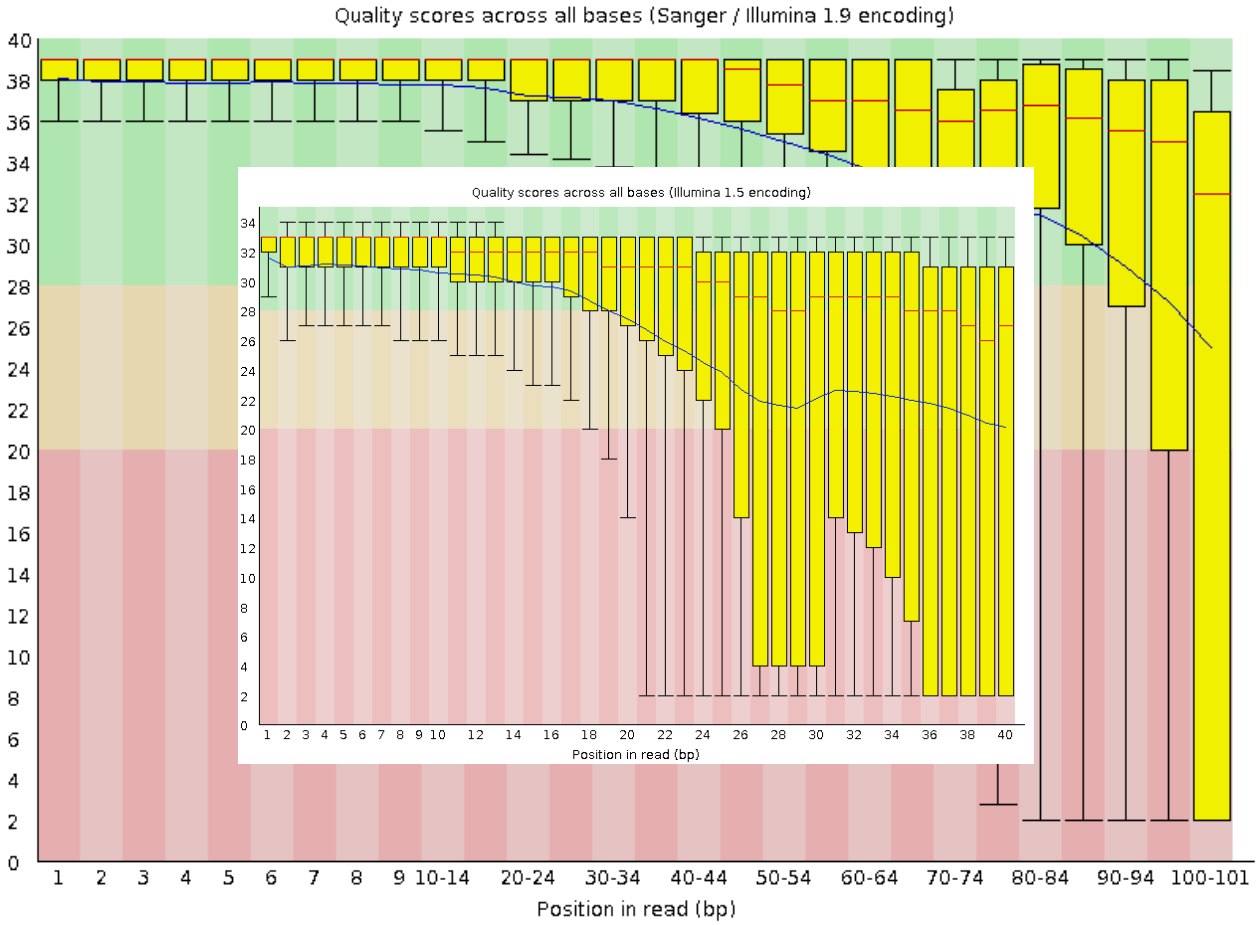
Acknowledgements

Base Quality Score

- Base-calling considers many possible sources of error in the sequencing process:
  - Mixed clusters
  - Out of phase clusters
  - Overlapping emission spectra
- The base quality score (BQS) is based on the probability that the base was called wrong.
- The quality score is "Phred scaled"
  - $-10 * \log_{10}(\text{probability of miscall})$


-10*log10(0.0013)) + 33 = 72 => H (ASCII)
PHRED Scaled
FastQ Encoded
FastQ
@HWI-ST179R:404:D1YNUACXX:7:1101:1228:1937 1:N:0:ATCACG
NTGATACAGCTGAACTGAACTCCTGCCTGTACAGCTCGTTTTCTACAAGATTCCAGACCTGGAAGATGATGGC
+
#1=DDDFFGHHHHJJHHIIHIIIFHJJFHCHGIJJGIICGIJIGIJJEHIAGHIIGJJJGIJIIHI:DGCHIG
@HWI-ST179R:404:D1YNUACXX:7:1101:1184:1940 1:N:0:ATCACG
NATGTCAGCCCACCCAGGAGCAACACAGACCCAAGGGAACCCCCACTCCCAGCCAAGAGAAGCCGTGAGTGAA
+
#1=BADDDHHHFHIIIIIIIIIII>GIHGIIIIIIIIIIIIIIIEIGGHIIIHFHHEEEECCCCB9>=?5>CD
@HWI-ST179R:404:D1YNUACXX:7:1101:1193:1964 1:N:0:ATCACG
NTGCCCCACAGGGGCGGTGTAAGACAGGAGTCCATCTGGGGCAGGGTGAGAGGATGGGGGTCAGAGGCACTAA
+
#1=DDFFFHHDHFIJJJ6@<FHJGIJJIIIC>EEEHHEFFFCDDDD(8?BADD??@BDDD07ACDDDB@B<CC
@HWI-ST179R:404:D1YNUACXX:7:1101:1166:1977 1:N:0:ATCACG
CTTCTTGCACCTCAAGGGATCACTCCCCTCTAGGCCGTTGCCTGTGCCATTTCCTCTCGCTGGGAAACCCTCT
+
@?<D:DDDDHHFHGIIIIF;DGGAHHIIGHHHIIIIIDHIIIDH9B?CCB=?@?B::>:>@3>>@C@CCB<@#
@HWI-ST179R:404:D1YNUACXX:7:1101:1423:1940 1:N:0:ATCACG
NAGATGTTGGCTATAGGAATACGGCAGGAAAATGAAACGTTGTGCATGGCAGGGCAGCATCACTTTGGGGATC
+
#11ADDDDHHHDHICFHGHIIHHIIIII7DC?EH@B?C@6;<B;3=CC:AAC5>?B9?&8?<?##########

FastQC Report

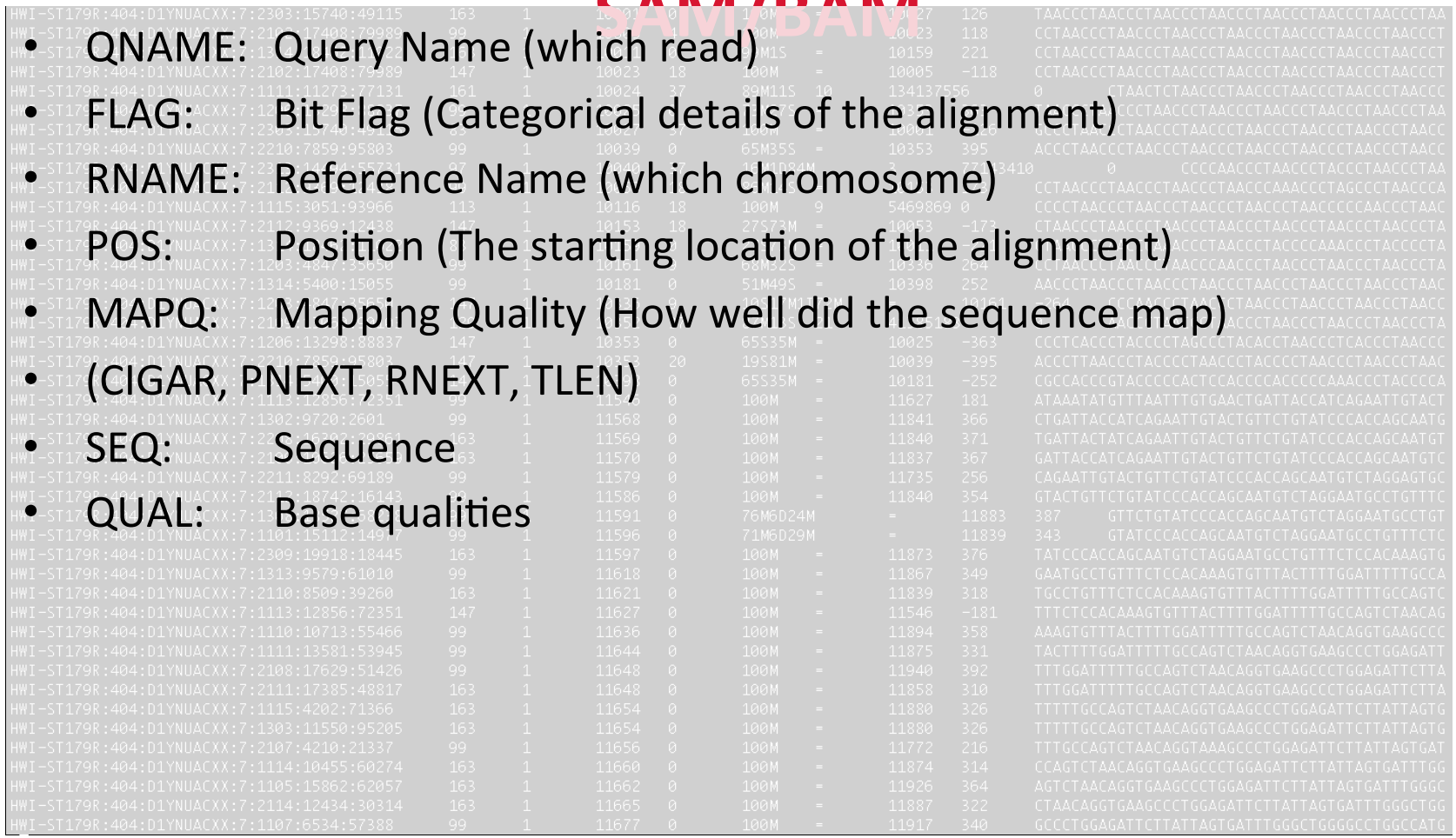


Sequence Alignment

- BLAST
  - Fast (at least we used to think so)
  - Accurate – Smith-Waterman algorithm guaruntees optiomal alignment.
  - Seed and extend (SW dynamic)
- Short-read aligners
  - Fast
  - Sort of accurate – Many hueristics
  - Seeding and extension (hueristics terminate unlikely extenstions).

SAM/BAM

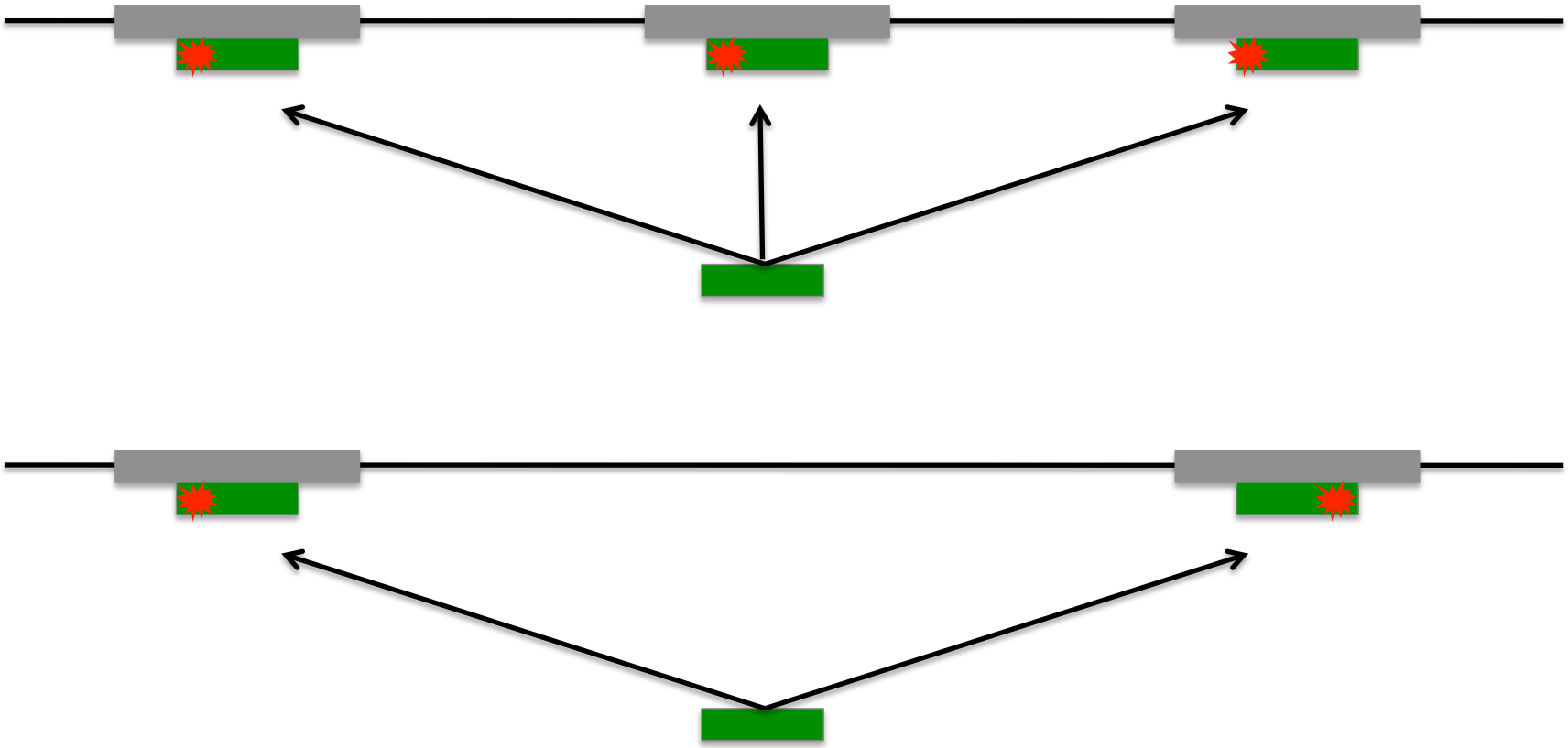
- QNAME: Query Name (which read)
- FLAG: Bit Flag (Categorical details of the alignment)
- RNAME: Reference Name (which chromosome)
- POS: Position (The starting location of the alignment)
- MAPQ: Mapping Quality (How well did the sequence map)
- (CIGAR, PNEXT, RNEXT, TLEN)
- SEQ: Sequence
- QUAL: Base qualities



Challenges for short-read aligners

- Repetative regions
  - Multiple algnments with the same score
  - How many of multiple alignments to score
  - Which SNV should I choose
- Insertions and deletions
  - Aligners hate to open gaps
  - Working one read doesn't allow context
  - Indel edges are inconsistent
  - A halo of bogus SNVs can be induced

Multi-mapping reads



Challenges for short-read aligners

- Repetative regions
  - Multiple algnments with the same score
  - How many of multiple alignments to score
  - Which SNV should I choose
- Insertions and deletions
  - Aligners hate to open gaps
  - Working one read at a time doesn't allow context
  - Indel placements are inconsistent
  - A halo of bogus SNVs can be induced

Indel alignment

Reference Sequence

TACAGCTGAACTGAACTCCTGCCTGTACAGCTCGTTTTCTACAAGATTCCAGACCTGGAA

```
                  TGCCTGTACAGCTCGTT-TCTACAAGATTC
AACTGAACTCCTGCCTGTACAGCTCGT
   TGAACTCCTGCCTGTACAGCTCG-TTTCTA
                     TGTACAGCTCGTTT-CTACAAGATTCCAGA
      CTCCTGCCTGTACAGCTCGTTTTCTACAAG
     ACTCCTGCCTGTACAGCTCGTTTC-TACAA
           GCCTGTACAGCTCGT-TTCTACAAGATTCC
```

Polishing Alignments

- Base quality recalibration – covariant analysis
  - Readgroup
  - Cycle
  - Sequence context
- Local re-alignment around indels (Local assembly)
  - Improves indel calls
  - Reduces SNV fales positive

Covariant based BQS recalibration

Reference Sequence

TACAGCTGAACTGAACTCCTGCCTGTACAGCTCGTTTTCTACAAGATTCCAGACCTGGAA

TGCCTGTACAGCTCGTTTTCTACGAGATTC

AACTGAACTCCTGCCTGTACAGCTCGT

TGAACTCCTGCCTGTACAGCTCGTTTTCTA

TGTACAGCTCGTTTTCTACGAGATTCCAGA

CTCCTGCCTGTACAGCTCGTTTTCTACGAG

ACTCCTGCCTGTACAGCTCGTTTTCTACG

GCCTGTACAGCTCGTTTTCTACGAGATTCC

Polishing Alignments

- Base quality recalibration – Covariant analysis
  - Readgroup
  - Cycle
  - Sequence context
- Local re-alignment around indels (Local assembly)
  - Improves indel calls
  - Reduces SNV fales positive

Indel alignment

Reference Sequence

TACAGCTGAACTGAACTCCTGCCTGTACAGCTCGTTTTCTACAAGATTCCAGACCTGGAA

```
                  TGCCTGTACAGCTCGTT-TCTACAAGATTC
AACTGAACTCCTGCCTGTACAGCTCGT
   TGAACTCCTGCCTGTACAGCTCG-TTTCTA
                     TGTACAGCTCGTTT-CTACAAGATTCCAGA
      CTCCTGCCTGTACAGCTCGTTTTCTACAAG
     ACTCCTGCCTGTACAGCTCGTTTC-TACAA
           GCCTGTACAGCTCGT-TTCTACAAGATTCC
```

Indel alignment

Reference Sequence

TACAGCTGAACTGAACTCCTGCCTGTACAGCTCGTTTTCTACAAGATTCCAGACCTGGAA

```
                    TGCCTGTACAGCTCGTT-TCTACAAGATTC
AACTGAACTCCTGCCTGTACAGCTCGT
   TGAACTCCTGCCTGTACAGCTCGTT-TCTA
                       TGTACAGCTCGTT-TCTACAAGATTCCAGA
                CTCCTGCCTGTACAGCTCGTT-TCTACAAG
               ACTCCTGCCTGTACAGCTCGTT-TCTACAA
                     GCCTGTACAGCTCGTT-TCTACAAGATTCC
```

Variant Calling

Reference Sequence

TACAGCTGAACTGAACTCCTGCCTGTACAGCTCGTTTTCTACAAGATTCCAGACCTGGAA

TGCCTGTACAGCTCGTTTTCTACAAGATTC

AACTGAACTCCTGCCTGTACGGCTCGT

TGAACTCCTGCCTGTACAGCTCGTTTTCTA

TGTACAGCTCGTTTTCTACAAGATTCCAGA

CTCCTGCCTGTACGGCTCGTTTTCTACAAG

ACTCCTGCCTGTACAGCTCGTTTTCTACAA

GCCTGTACGGCTCGTTTTCTACAAGATTCC

Bayes Theorem

$$P(Ref|Data) = \frac{P(Data|Ref) * P(Ref)}{P(Data)}$$

- Incorporates the probability of our data given the hypotyhesis
- Incorporates prior information – the probability this site is reference.
- Incorporates the probability of the data under all hypotheses
- Provides a probability of belief in the hypothesis

Variant quality score

- The probability that the site was incorrectly called
- Phred scaled so its more intuitive
- Can be used to filter low quality sites

Genotype quality score

- The probability that the site was incorrectly genotyped.
- Is similar to VQS when calling single individual's variants – but not when calling population variants.
- Can be used to filter low quality sites on an individual basis when calling variants on a population.
- This is the value that you want to use to evaluate the quality of an variant.

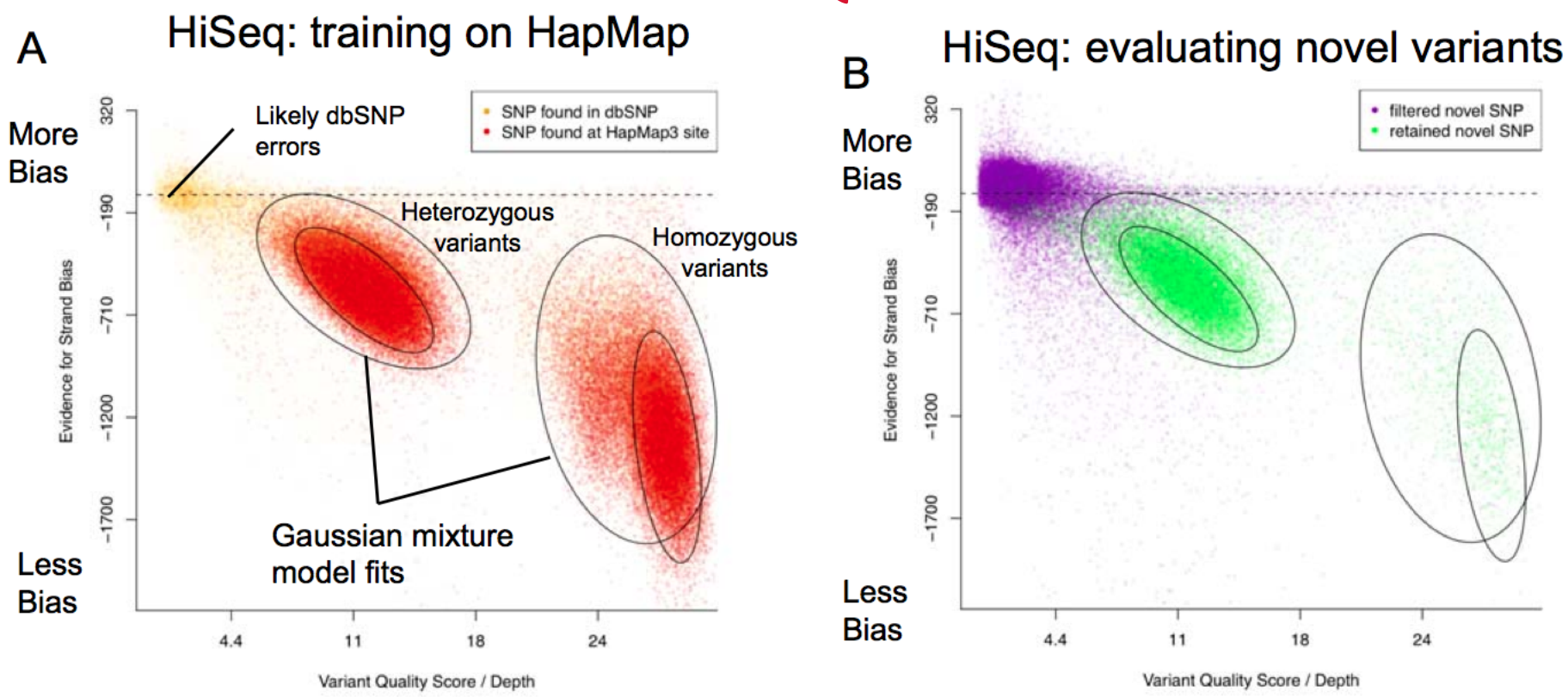
Variant filtering and recalibration

- Filtering on VQS or GQS is so yesterday!
- Population based variant calling allows population based covariant analysis of VQS
- Variant calling programs are implimenting these recalibrations
  - GATK – VQSR
  - Real Time Genomics – AVR Score
  - FreeBayes

VQSR

- Considering only high-quality sites (HapMap3 concordant) build a mixture model considering various co-varying parameters
  - Allele Balance (ref/alt in hets)
  - Homopolymer run
  - MAPQ score
  - Strand Bias
  - Depth of coverage
- Apply that model to all variants until a given number of known sites have been recovered.

VQSR



DePreisto et al, Nature Genetics 2011

Evaluating variant call sets

- Variant counts (~3.5-4M WGS, ~50K capture, ~20K coding, ~10K missense)
- Ti/Tv Ratio (~2.1-2.8)
- Concordance with known sites (HapMap3)

- NIST/GCAT – Tool for comparing pipelines
  - http://www.bioplanet.com/gcat