CNV, GWAS and Clinical Analysis Advancements in SVS

Dr. Eli Sward | FAS

Gabe Rudy | VP of Product & Engineering

NIH Grant Funding Acknowledgments

- **Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under:**
  - Award Number R43GM128485
  - Award Number 2R44 GM125432-01
  - Award Number 2R44 GM125432-02

- **PI is Dr. Andreas Scherer, CEO Golden Helix.**
- **The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.**

Q & A

Please enter your questions into your GoToWebinar Panel

Golden Helix – Who We Are

Golden Helix is a global bioinformatics company founded in 1998, celebrating our 20th year!

GlaxoSmithKline

varSEQ®

- Variant Calling
- Filtering and Annotation
- ACMG Guidelines
- Clinical Reports
- CNV Analysis
- Pipeline: Run Workflows

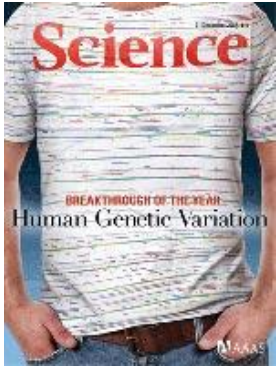
VARIANT WAREHOUSE

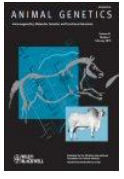
- Variant Warehouse
- Centralized Annotations
- Hosted Reports
- Sharing and Integration

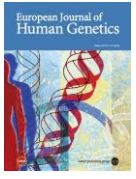
SNP & VARIATION SUITE

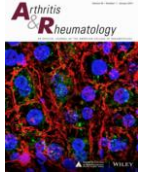
- GWAS
- Genomic Prediction
- Large-N-Population Studies
- RNA-Seq
- Large-N CNV-Analysis

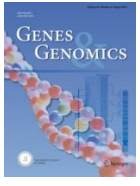
Cited in over 1,300 peer-reviewed publications

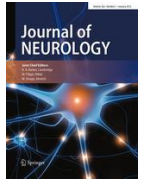












Over 400 customers globally

SickKids
THE HOSPITAL FOR SICK CHILDREN

Cincinnati Children's
Hospital Medical Center

The Children's Hospital of Philadelphia

Stanford University

UBC

Ucla

Yale

The Feinstein Institute for Medical Research
North Shore-Long Island Jewish Health System
North Shore LIJ

PREVENTION GENETICS
DISEASE PREVENTION THROUGH GENETIC TESTING

MAYO CLINIC

LINEAGEN

MOFFITT CANCER CENTER

Dartmouth-Hitchcock
NORRIS COTTON CANCER CENTER

BAYER

Lilly

abbvie

TEVA

NATIONAL CANCER INSTITUTE

USDA
USDA Agricultural Research Service

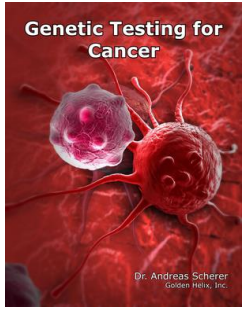
FDA

GOLDEN HELIX
Enabling Precision Medicine

Golden Helix – Who We Are

When you choose a Golden Helix solution, you get more than just software

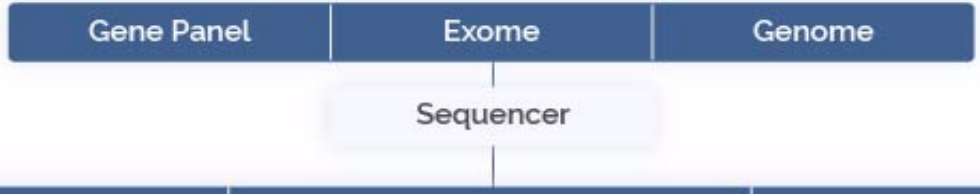
- REPUTATION
- TRUST
- EXPERIENCE



- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

- TRAINING
- SUPPORT
- RESPONSIVENESS

- INNOVATION
- SPEED



Products	Bioinformatics Pipeline	Function
DNAseq (Sentieon) TNseq (Sentieon) VS-CNV	FASTQ BAM VCF	▸ Single nucleotide variation ▸ Copy number variation & loss of heterozygosity ▸ Chromosomal aberration
Annotations	Annotated VCF	▸ Public & commercial annotations to enrich genomic data sets
VarSeq VSReports VSPipeline	Clinical Report	▸ Annotate & filter ▸ Visually inspect alignments ▸ Variant prioritization ▸ Clinical assessment
VSClinical	Automated ACMG Guidelines	▸ Clinical variant interpretation in concordance with ACMG Guidelines
VSWarehouse	Data Warehousing Web-Enabled Interface + Powerful API: JSON, XML, TSV, CSV, SQL, FHIR	▸ Clinical assessment catalog ▸ Advanced data querying ▸ Versioning ▸ Interoperability ▸ Compliance with HIPPA, CLIA & CAP data discovery

Agenda

Dr. Eli Sward (FAS)

- Background on detecting CNVs
- Targeted vs. Binned Region approach
- Software demonstration of VarSeq
- Software demonstration in SVS
- Downstream applications of CNV analysis in SVS

Gabe Rudy (VP of Product)

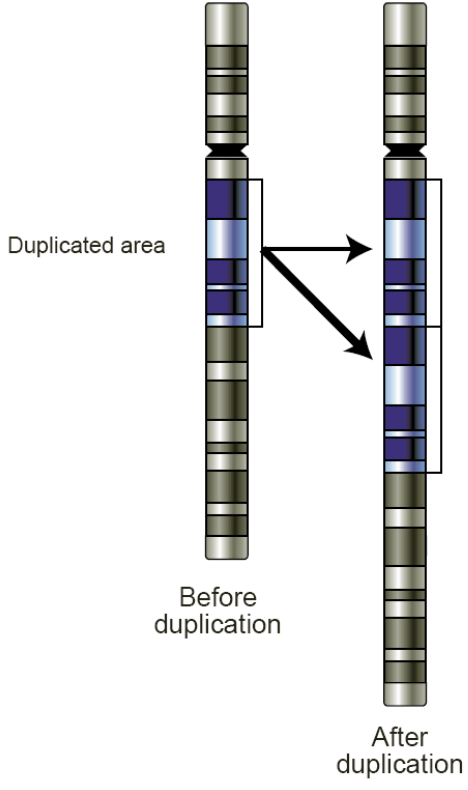
- GWAS improvement and explanations
- SVS-Clinical Variant Scoring
- Splice Site Predictions
- Functional Predictions & Conservation
- Other updates to SVS

CNVs in Clinical and Research settings

- **Clinical – Diagnostic testing**
  - Common drivers in specific cancers and causal agents in hereditary variation
    - EGFR Exon 19 deletion (J Clin Oncol. 2011 May 20; 29(15): 2066–2070.)
    - PIK3CA Amplification breast cancer (Mol Cytogenet. 2018; 11: 5.)
  - Large Events commonly seen in disorders
    - Autism Spectrum Disorder
    - Developmental Delay
- **Research – Large population based discovery**
  - Large microdeletions associated with schizophrenia (Nature. 2008 Sep 11; 455(7210): 232–236.)
  - Genome wide CNV analysis for growth rates *Bos indicus* (BMC Genomics. 2016 Jun 1;17:419.)
  - Discovery leads to testing for association

Detecting CNVs

- **Chromosomal microarray**
  - Current best practice
  - Slow
  - Additional expense
  - Only detects large events
- **CNV calling from NGS data**
  - Calls CNVs from existing coverage data
  - Detects large and small single-exon events
  - Can be applied to both **Clinical** and **Research** settings
    - Implement a targeted approach (exomes and gene panels)
    - Or binned region approach (whole genome data)



Targeted vs. Binned Region Approach

Targeted (BED file)

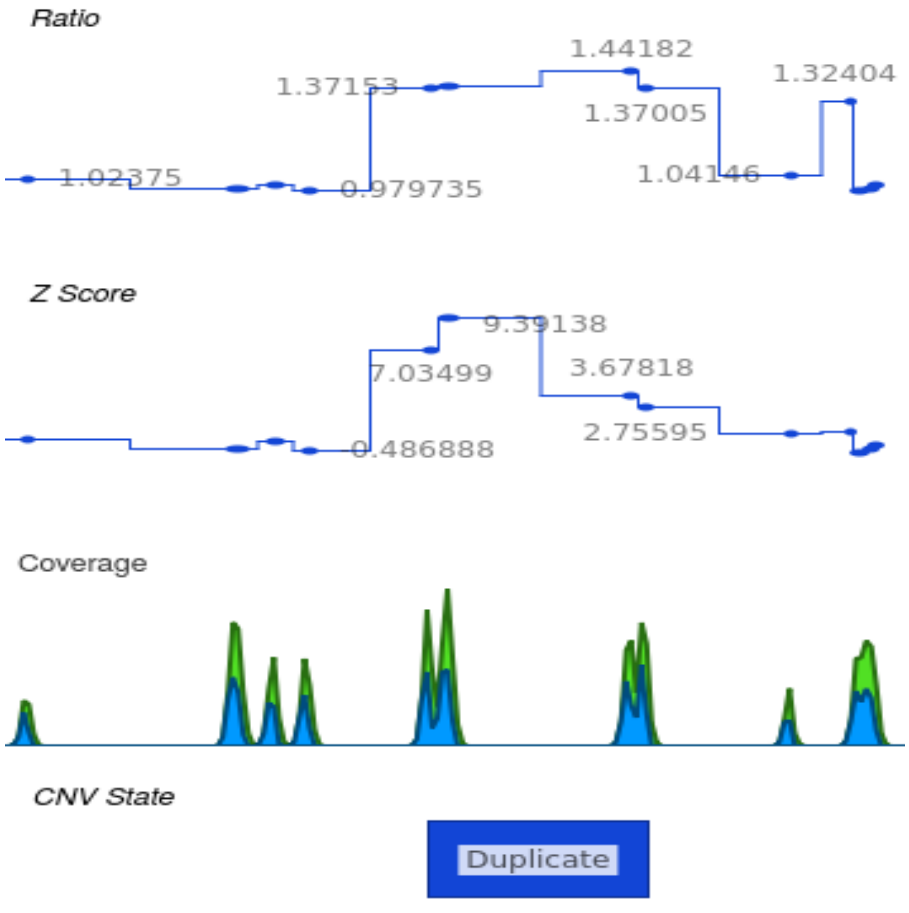
- Gene Panels, Whole Exome Sequencing
- 30+ Samples (same library and preparation)
- Normalization
- ≥ 100X coverage
- ≥ Single exon level
- Metrics (Z-score, Ratio, p-value)

Binned (Binned widths)

- Shallow Whole Genomes
- 30+ Samples (same library and preparation)
- Normalization
- ≥ 0.5X Coverage
- ≥ 100Kbp CNV events
- Metrics (Z-score, Ratio, p-value)

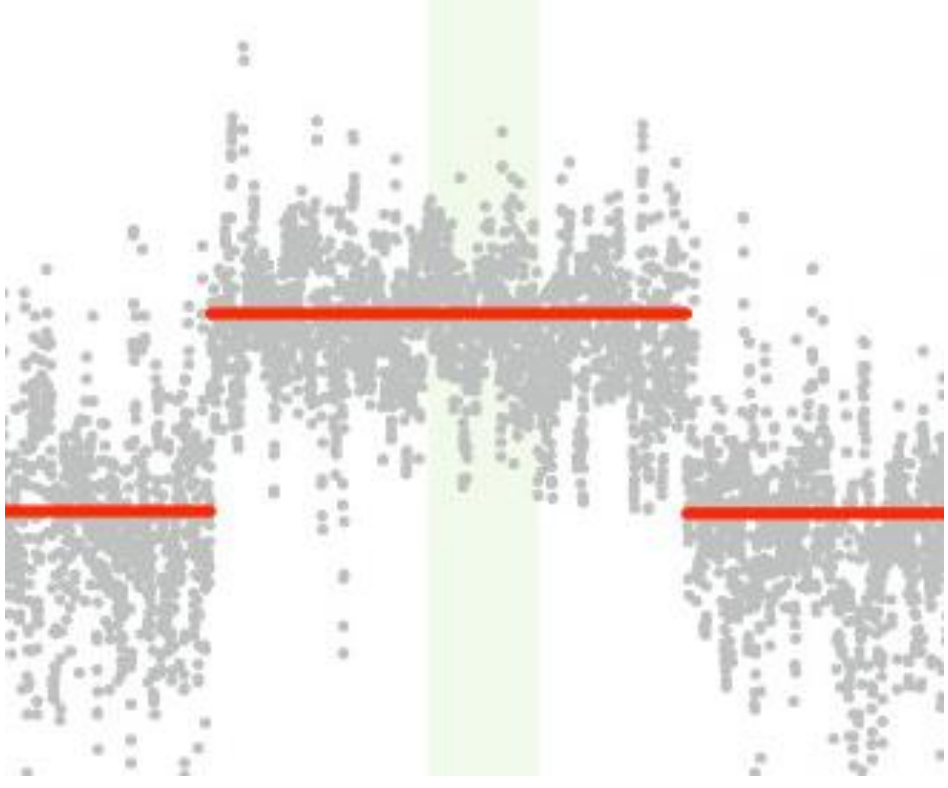
CNV Calling in VarSeq and SVS

- **Metrics**
  - Z-score number of s.d from reference sample mean
  - Ratio sample coverage/reference sample mean
  - Variant allele frequency (supporting metric)
- **Flags**
  - CNV events
    - Low coverage, high variation, noise
  - Samples
    - Mismatch to reference, low coverage, high variance across regions
- **Improves precision**



Calling CNVs from WGS implements segmentation

- **Metrics are noisy over large regions**
- **Outliers cause large events to be called as many small events**
- **Address this using CNAM Optimal Segmentation**
  - Regions containing many events are segmented
  - Small events sharing a segmented region are merged



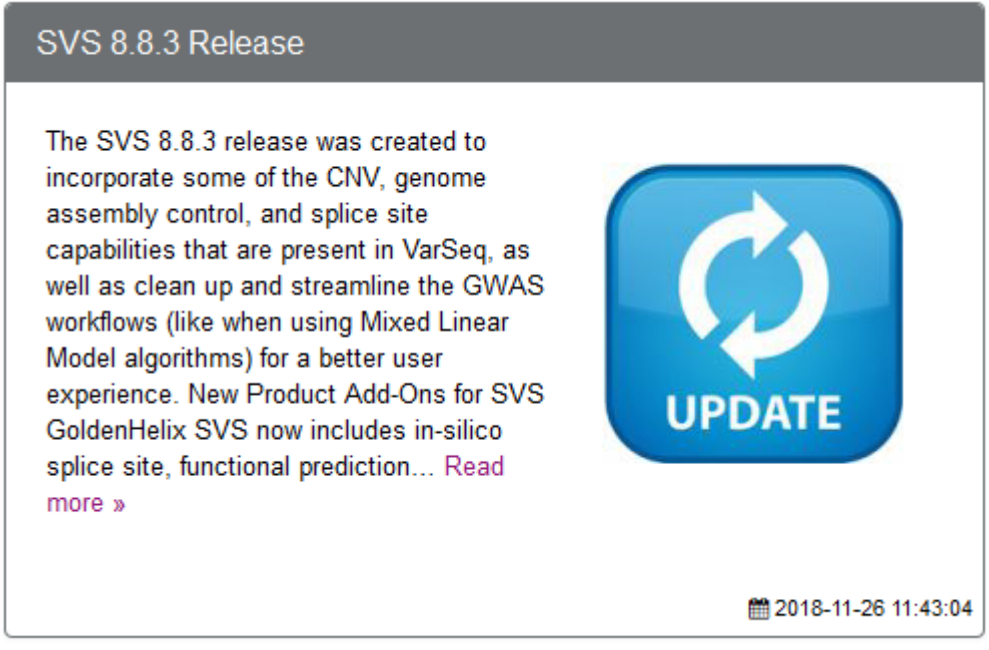
Workflow and Demonstration

- Filter CNVs from WGS
- Annotate
- Visualize > Sample specific
- Catalog

- Filter CNVs from WGS
- Annotate
- Visualize > Cohort
- Discuss downstream analyses

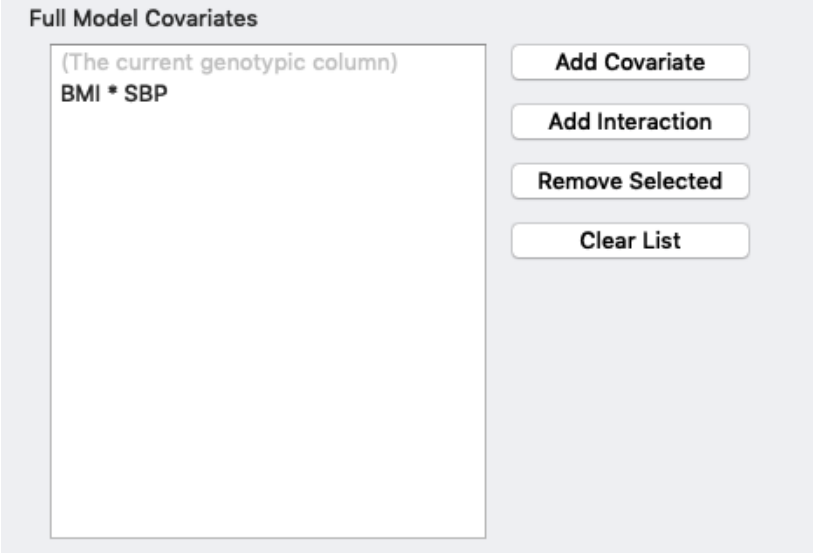
Highlights of Latest SVS Release

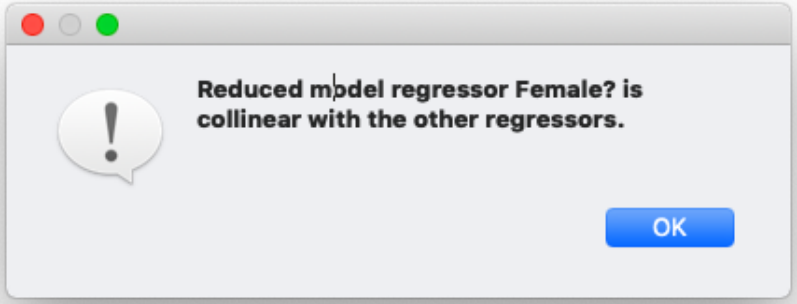
- **Whole Genome CNV Calling**
- **GWAS and Mixed Model Improvements**
- **Clinical Variant Scoring Module**



SVS: GWAS Improvements

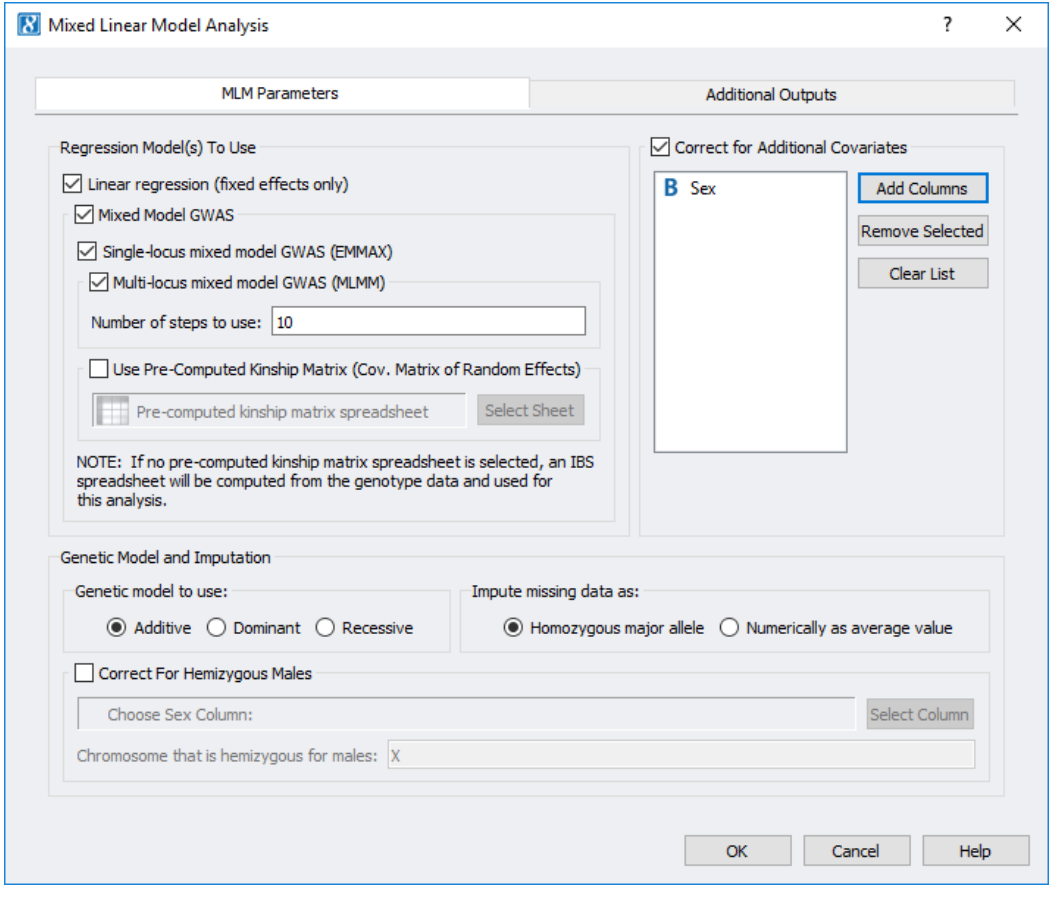
- **Genotype Regression**
  - Does not require recoding genotypes to numeric
  - Includes marker statistics
- **Improved Regression:**
  - Display of “implicit” full & reduced model terms
  - Collinear detection and warnings
  - Don’t allow double-adding terms
- **Output Improvements:**
  - Covariate term names in column headers for Beta / Beta-Standard-Error columns





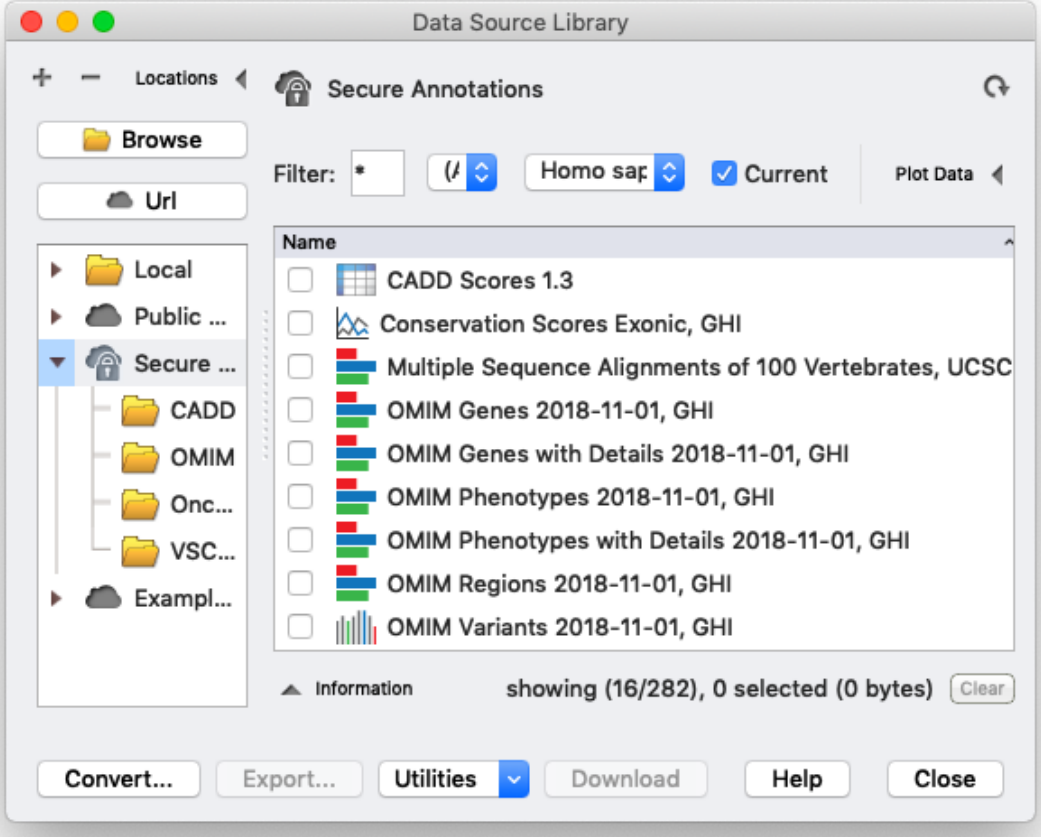
SVS: GWAS Mixed Model Improvements

- **GBLUP Improvements**
  - General speedup (2-5X)
  - Precision tightened from .0001 to .000001
- **Mixed Model Analysis:**
  - Collinear detection with informative diagnosis
  - All Beta values and their Standard Error output
  - Multi-allelic genotypes supported
  - Various numerical ”edge cases” now supported disorders



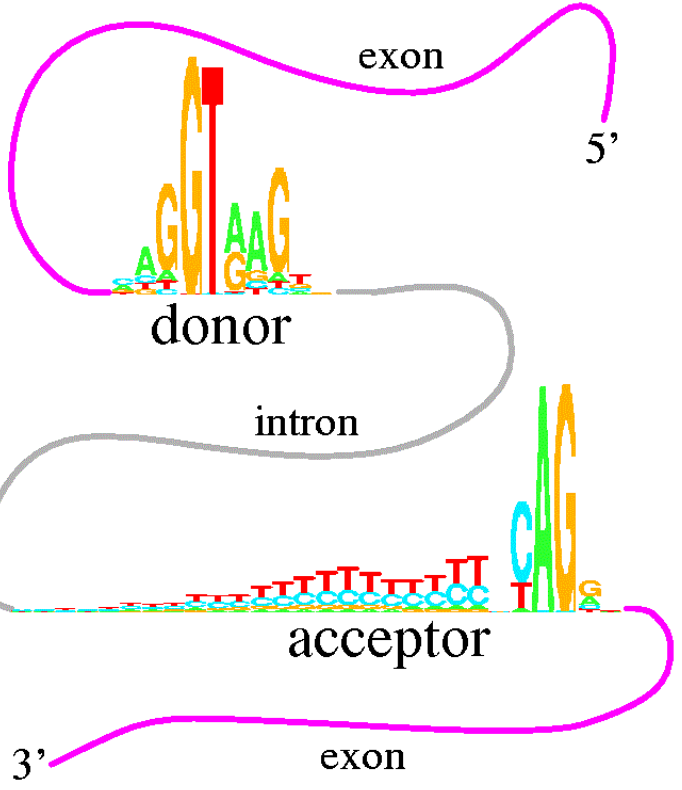
SVS: Clinical Variant Scoring Module

- **Premium Annotations for Clinical Workflows**
  - Splice Site Predictions
  - Multiple Sequence Alignment SIFT/Polyphen2
  - Conservation Scores: GERP++/PhyloP
- **Other Premium SVS Annotations:**
  - **CADD**: Best option for annotating whole genomes and indels. Updated 1.4 coming soon!
  - **OMIM**: Variants and gene annotations for genomic disorders



Splice Sites

- **Introns have distinct nucleotide pairs at each end**
  - GT at the 5’ end (Donor Site)
  - AG at the 3’ end (Acceptor Site)
- **Sequences around splice sites are highly variable**
- **Machine learning and probabilistic methods are used to identify sites**

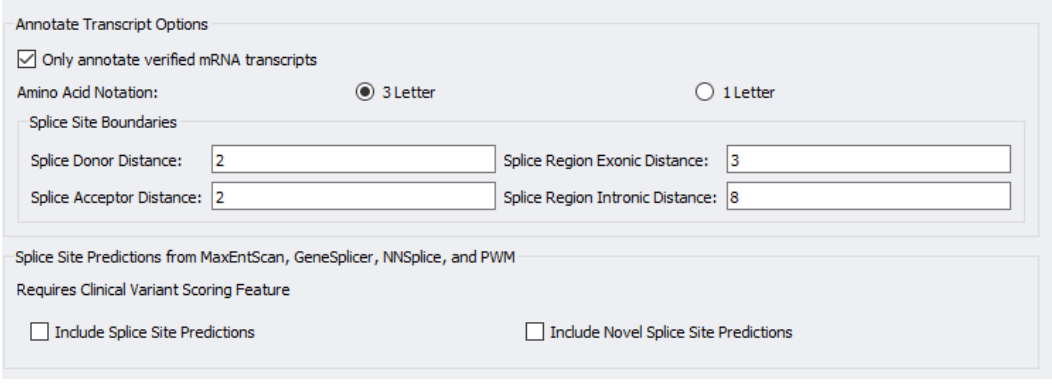


Splice Site Algorithms

- **VSClinical supports four splice site prediction algorithms**
  - PWM: Uses position weight matrix similar to SpliceSiteFinder and Human Splice Finder
  - MaxEntScan: Approximates sequence motifs using Maximum Entropy Distribution
  - NNSplice: Identifies splice sites using neural networks
  - GeneSplicer: Uses Markov models combined with maximal dependence decomposition

Splice Site Algorithms in SVS

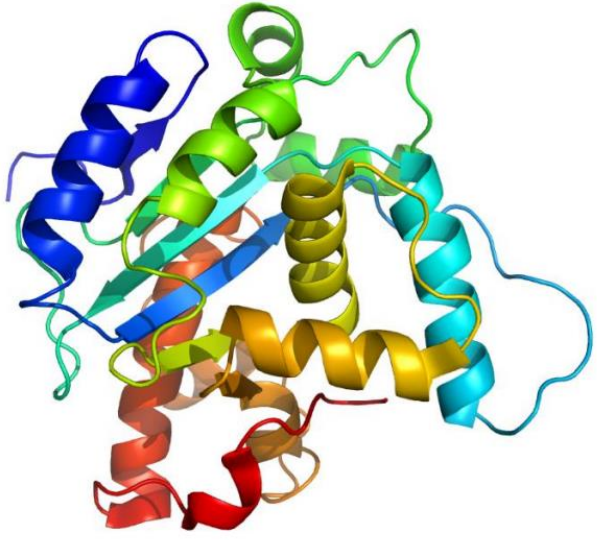
- **Runs Transcript Annotation**
- **Two Modes:**
  - Predicted Splicing Disrupted
  - Novel Splice Site Detection



C 11	C 12	C 13	I 14
Novel Splice Site	Novel Splice Type	N of 4 Predicted Novel Splice Site	Distance to Novel Splice Site
1:17301791	Donor	1 of 4 Predicted Splicing Disrupted	-11
1:40555086	Donor	4 of 4 Predicted Splicing Disrupted	-1
1:41284287	Donor	4 of 4 Predicted Splicing Disrupted	-5
1:116283389	Donor	4 of 4 Predicted Splicing Disrupted	-1
1:155207936	Donor	3 of 4 Predicted Splicing Disrupted	0
1:158645957	Donor	4 of 4 Predicted Splicing Disrupted	0
1:161276536	Donor	4 of 4 Predicted Splicing Disrupted	-1
1:209791295	Donor	4 of 4 Predicted Splicing Disrupted	-1
1:216498839	Donor	3 of 4 Predicted Splicing Disrupted	2
2:166895936	Donor	1 of 4 Predicted Splicing Disrupted	0
2:169791747	Donor	4 of 4 Predicted Splicing Disrupted	0
2:219674478	Donor	2 of 4 Predicted Splicing Disrupted	-1
2:219677824	Donor	1 of 4 Predicted Splicing Disrupted	5
3:15495354	Donor	2 of 4 Predicted Splicing Disrupted	-1
3:48618331	Donor	4 of 4 Predicted Splicing Disrupted	-1

Functional Prediction Algorithms

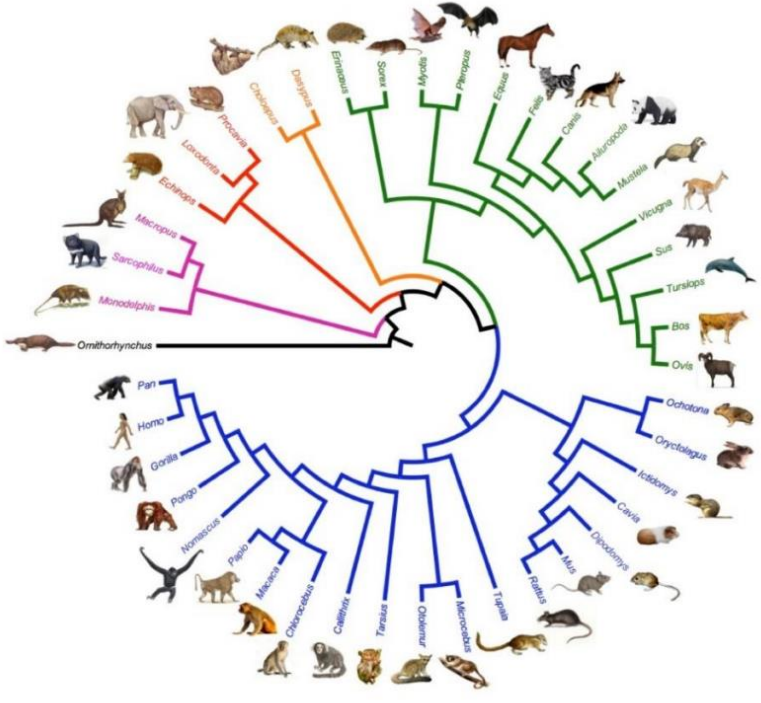
- **SIFT**
  - Uses matrix to encode the probability of each amino acid at each position of the protein
  - Probabilities computed from protein sequence alignment
  - Other Premium SVS Annotations:
- **PolyPhen2**
  - Naïve Bayes multi-evidence approach
  - Uses similar protein alignment-based probability score called PSIC
  - Incorporates nine other metrics



Functional Prediction Algorithms

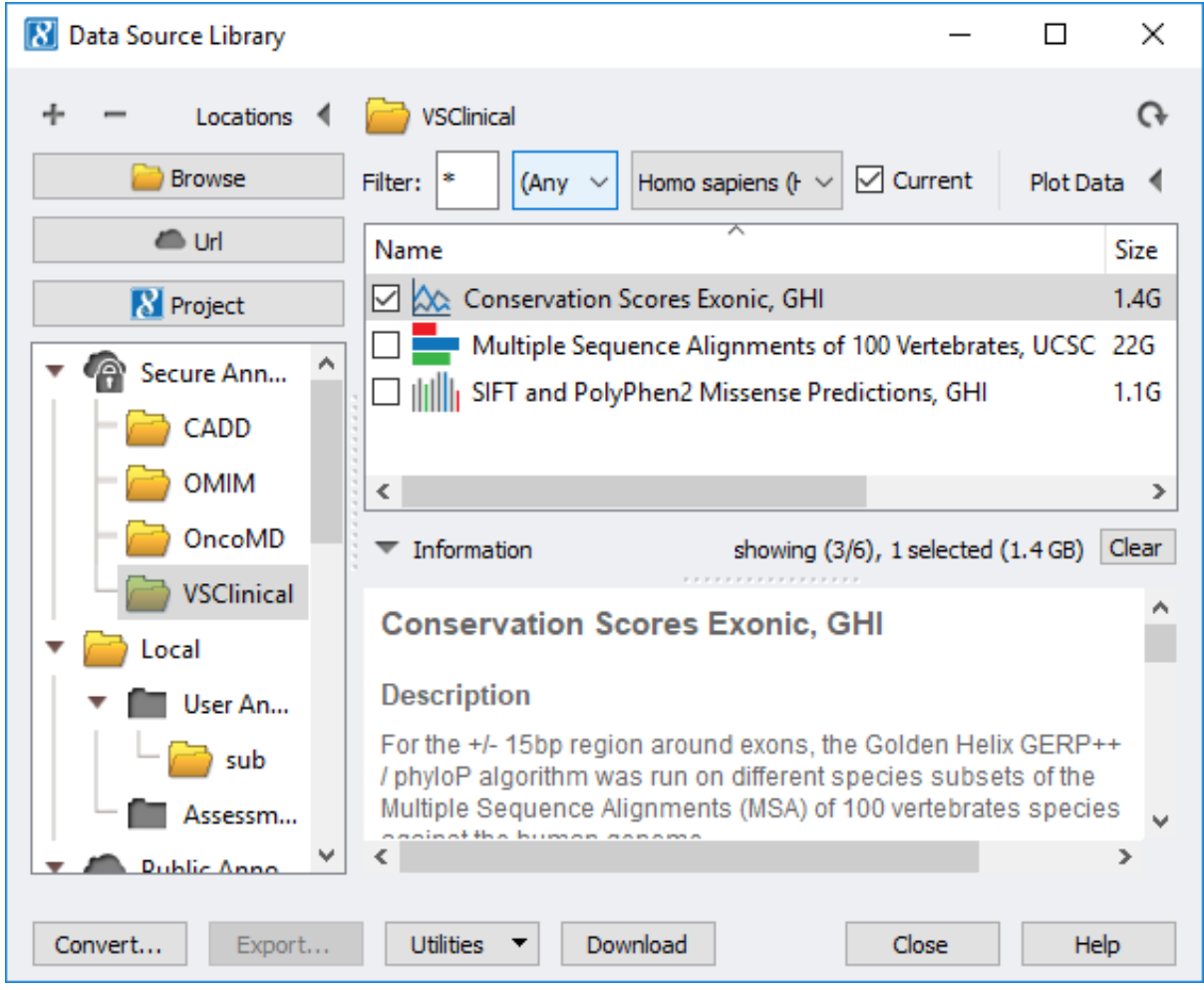
- **Conservation Scores**
  - Use phylogenetic models
  - Find maximum likelihood scaling factor for model given an alignment
- **GERP++**
  - Uses rejected substitutions (RS) as test statistic
  - RS value is computed from the neutral rate **n** and the maximum likelihood scaling factor **θ**

    RS = n(1- θ)
- **PhyloP LR**
  - Two times the difference in log likelihood between
    - null hypothesis (no scaling factor)
    - alternative hypothesis (maximum likelihood scaling factor)



Conservation Scores in SVS

- **Precomputed Scores for Exon Regions**
  - RefSeq exons +/- 15bp
  - Conservation Scores Exonic track
- **Annotate Any Variant (Non-Exonic)**
  - Uses Multiple Sequence Alignment to compute conservation
  - Download the (large) MSA for performant annotation



Demo

SNP & VARIATION SUITE

[DEMONSTRATION]

Q & A

Please enter your questions into your GoToWebinar Panel

Thank you!

- **Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under:**
  - Award Number R43GM128485
  - Award Number 2R44 GM125432-01
  - Award Number 2R44 GM125432-02

- **PI is Dr. Andreas Scherer, CEO Golden Helix.**
- **The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.**

End of Year Bundles - 2018

- 3 remain ~~(5)~~ SNP & Variation Suite (SVS) w/ CADD & OMIM | 3-Seat License - $5,995
- 4 remain ~~(5)~~ SNP & Variation Suite (SVS) w/ CNV & VarSeq w/ CNV | 1-Seat License - $9,995
- 2 remain ~~(3)~~ SNP & Variation Suite (SVS) Imputation Module | 2-Seat License - $7,995
- 2 remain ~~(3)~~ VarSeq w/ CADD & OMIM | 3-Seat License - $8,995
- 3 remain ~~(5)~~ VarSeq CNV PowerPack & Sentieon Tier One | 2-Seat License - $17,495
- 4 remain ~~(5)~~ VSClinical, CNV & Sentieon Tier One | 2-Seat License - $24k
- 2 remain ~~(2)~~ Small Lab Data Warehouse Package w/ CNV, VSClinical & Sentieon Tier One | 2-Seat License - $42k

Bundles End December 21st

bit.ly/EOYbundles

GOLDEN HELIX
Enabling Precision Medicine

Q & A

Please enter your questions into your GoToWebinar Panel