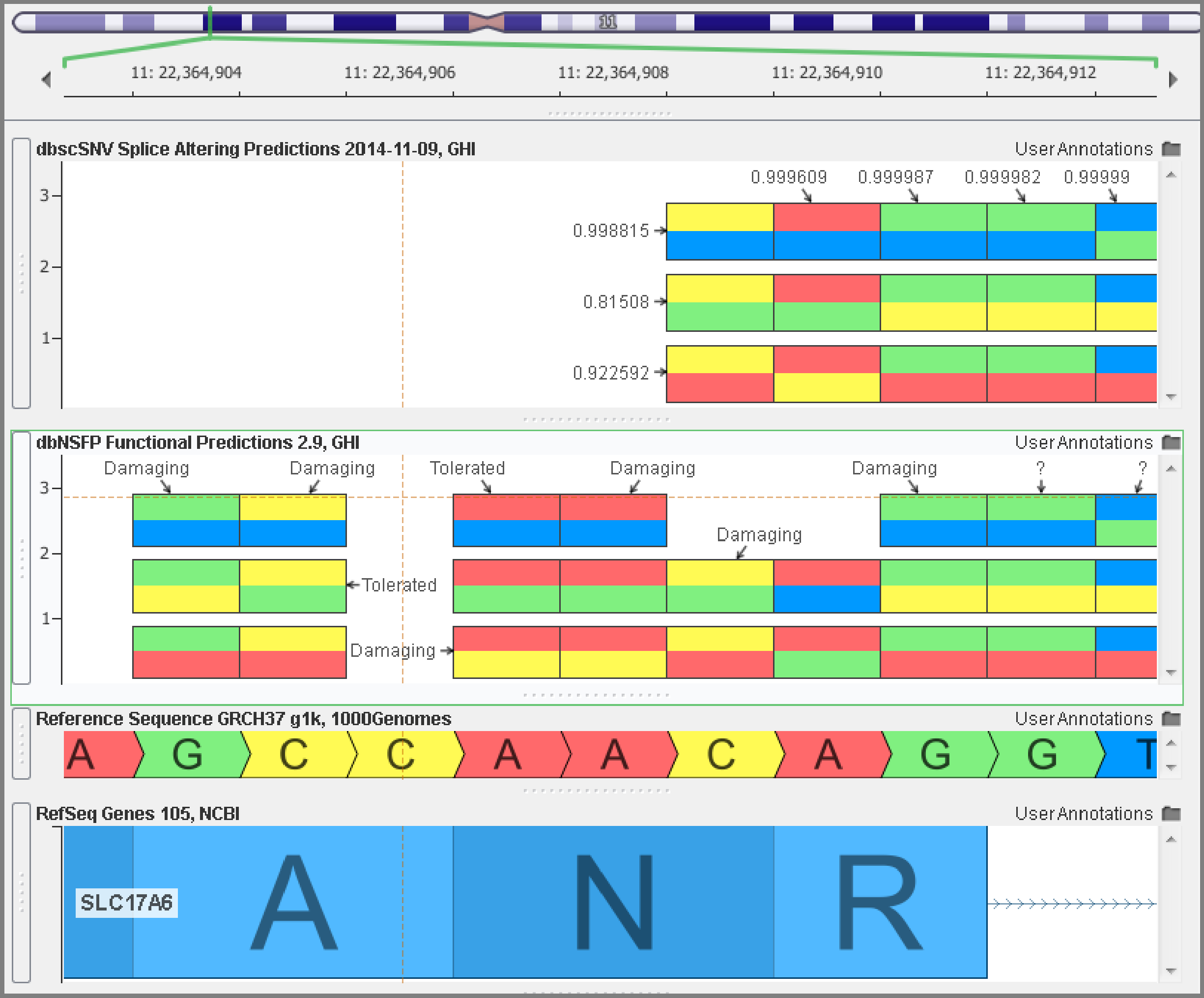


Using VarSeq to Improve Variant Analysis Research

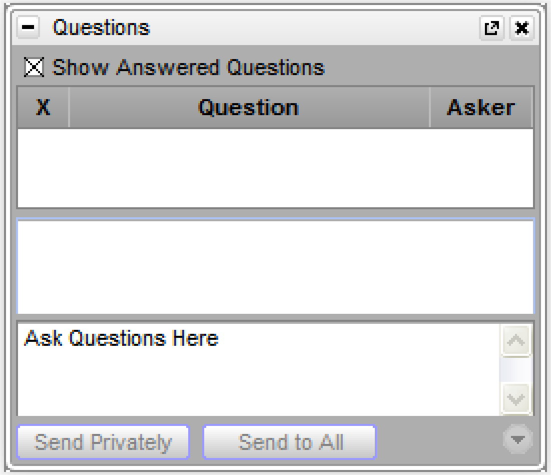
June 10, 2015

G Bryce Christensen
Director of Services

Accelerating the Quest for Significance™

Questions during the presentation

Use the Questions pane in your GoToWebinar window



Questions

☒ Show Answered Questions

X	Question	Asker

Ask Questions Here

Agenda

1 Variant analysis workflows

2 What makes a damaging variant?

3 QC Considerations

4 VarSeq Interactive Demonstration

What is VarSeq?

Flexible

Simple

Scalable

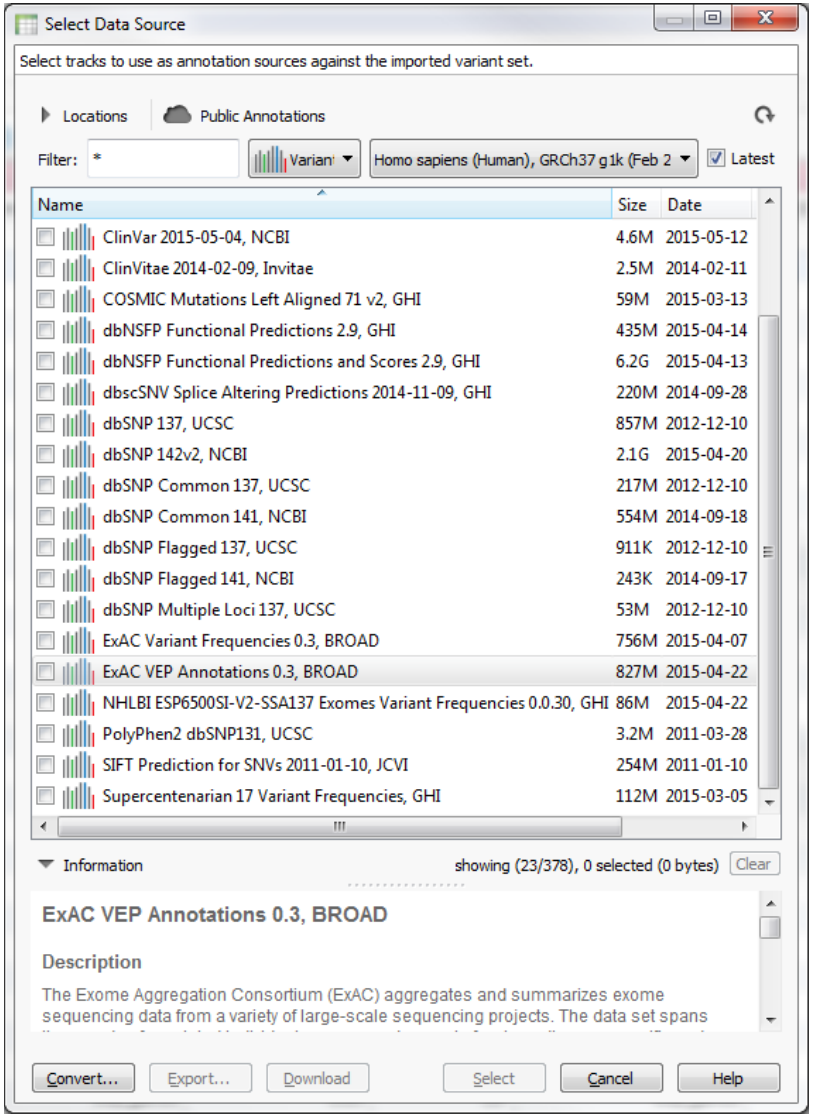
- Variant annotation, filtering and ranking
- Repeatable workflows
- Rich visualizations with GenomeBrowse integration
- Powerful GUI and command-line interfaces

Workflow Development Process in VarSeq

1. Begin from one or many VCF files
2. Annotate variants using public data sources curated by Golden Helix and/or annotate with custom data sources.
3. Run additional computation algorithms
   - Allele counts, genotype zygosity, gene list matching, etc
4. Construct filter chain to identify candidate variants
   - May use combinations of logical operators in filters
   - May have multiple independent filter chains and/or endpoints
5. Process results
   - Gene Ranking with PhoRank
   - Review variant QC
   - Vizualization with GenomeBrowse
   - Commit variants to local database
   - Etc.

Annotations are the key

- Good variant analysis begins with accurate annotations.
- Golden Helix invests extensive time and effort in validating and maintaining data sources.
- Annotation data sources may be used for either quality control or analytic purposes.



Select Data Source

Select tracks to use as annotation sources against the imported variant set.

Filter: *   Variant   Homo sapiens (Human), GRCh37 g1k (Feb 2   ☑ Latest

	Name	Size	Date
□	ClinVar 2015-05-04, NCBI	4.6M	2015-05-12
□	ClinVitae 2014-02-09, Invitae	2.5M	2014-02-11
□	COSMIC Mutations Left Aligned 71 v2, GHI	59M	2015-03-13
□	dbNSFP Functional Predictions 2.9, GHI	435M	2015-04-14
□	dbNSFP Functional Predictions and Scores 2.9, GHI	6.2G	2015-04-13
□	dbscSNV Splice Altering Predictions 2014-11-09, GHI	220M	2014-09-28
□	dbSNP 137, UCSC	857M	2012-12-10
□	dbSNP 142v2, NCBI	2.1G	2015-04-20
□	dbSNP Common 137, UCSC	217M	2012-12-10
□	dbSNP Common 141, NCBI	554M	2014-09-18
□	dbSNP Flagged 137, UCSC	911K	2012-12-10
□	dbSNP Flagged 141, NCBI	243K	2014-09-17
□	dbSNP Multiple Loci 137, UCSC	53M	2012-12-10
□	ExAC Variant Frequencies 0.3, BROAD	756M	2015-04-07
□	ExAC VEP Annotations 0.3, BROAD	827M	2015-04-22
□	NHLBI ESP6500SI-V2-SSA137 Exomes Variant Frequencies 0.0.30, GHI	86M	2015-04-22
□	PolyPhen2 dbSNP131, UCSC	3.2M	2011-03-28
□	SIFT Prediction for SNVs 2011-01-10, JCVI	254M	2011-01-10
□	Supercentenarian 17 Variant Frequencies, GHI	112M	2015-03-05

Information   showing (23/378), 0 selected (0 bytes)

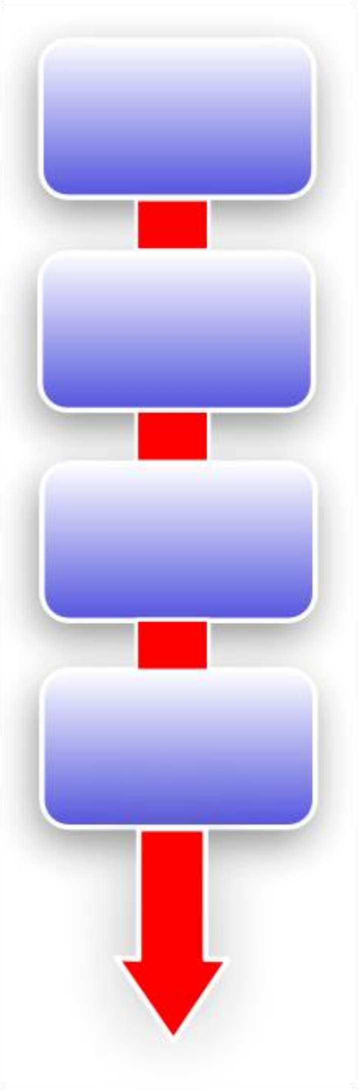
ExAC VEP Annotations 0.3, BROAD

Description

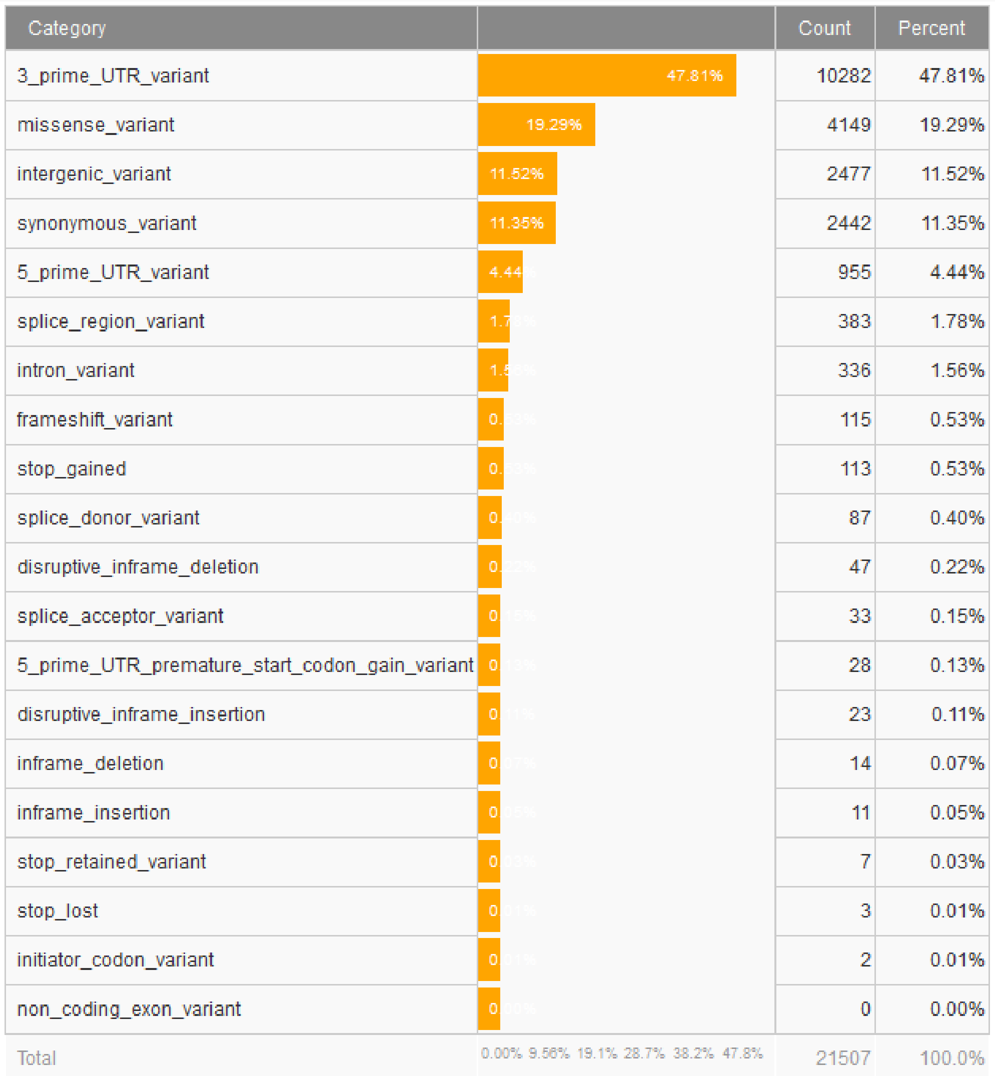
The Exome Aggregation Consortium (ExAC) aggregates and summarizes exome sequencing data from a variety of large-scale sequencing projects. The data set spans

Defining Deleteriousness

- **What makes a variant potentially damaging?**
- **Start by defining the search space:**
  - Rare, non-synonymous, homozygous variants?
  - DeNovo mutations in highly conserved genes?
  - Splice-site mutations?
  - Etc.
- **Review annotations for remaining variants to identify causal candidates**
- **Which annotations to use?**



Variant Classification

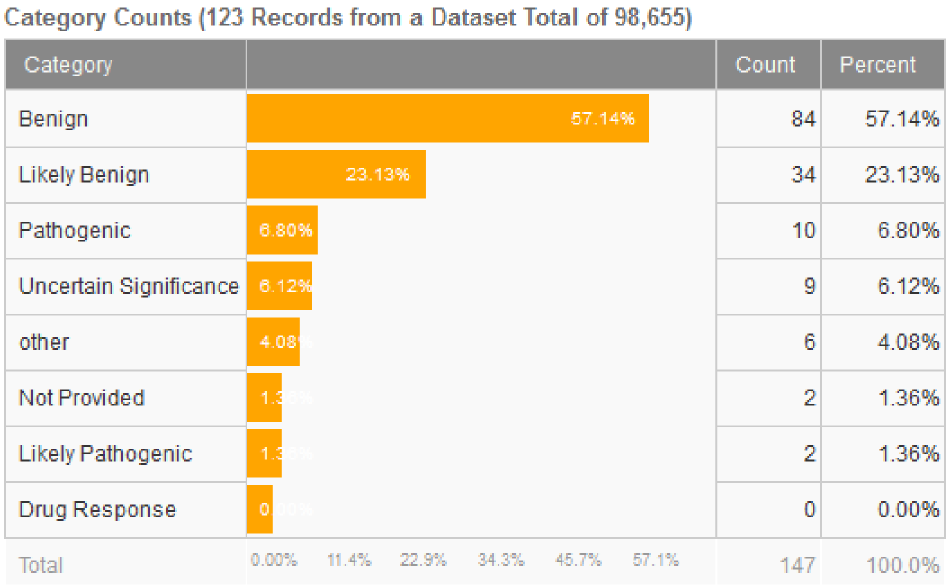


Category		Count	Percent
3_prime_UTR_variant	47.81%	10282	47.81%
missense_variant	19.29%	4149	19.29%
intergenic_variant	11.52%	2477	11.52%
synonymous_variant	11.35%	2442	11.35%
5_prime_UTR_variant		955	4.44%
splice_region_variant		383	1.78%
intron_variant		336	1.56%
frameshift_variant		115	0.53%
stop_gained		113	0.53%
splice_donor_variant		87	0.40%
disruptive_inframe_deletion		47	0.22%
splice_acceptor_variant		33	0.15%
5_prime_UTR_premature_start_codon_gain_variant		28	0.13%
disruptive_inframe_insertion		23	0.11%
inframe_deletion		14	0.07%
inframe_insertion		11	0.05%
stop_retained_variant		7	0.03%
stop_lost		3	0.01%
initiator_codon_variant		2	0.01%
non_coding_exon_variant		0	0.00%
Total		21507	100.0%

- **VarSeq classifies variants into 20+ different categories**
- **The categories are further grouped as:**
  - Loss of Function
  - Missense
  - Other
- **Choice of gene transcript reference**
  - RefSeq
  - Ensembl
  - Others

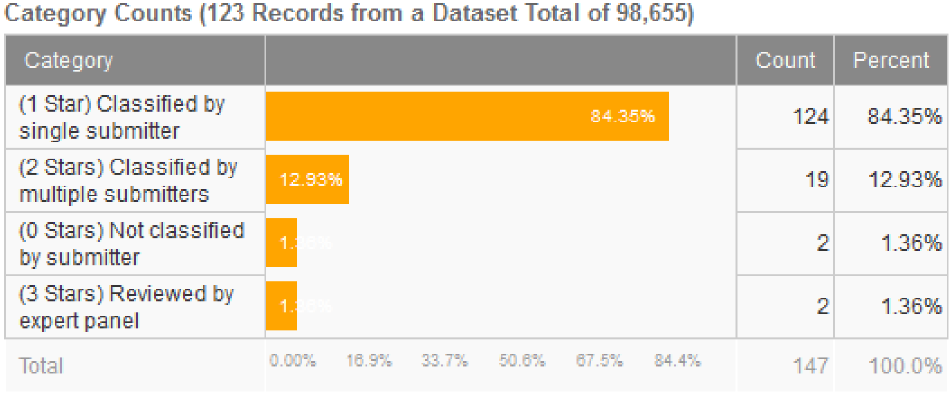
ClinVar

- ClinVar is a public archive of variants evaluated for potential causal relationships to diseases
- Submissions from many sources, including major clinical laboratories
- Over 100k records
- Updated monthly



Category Counts (123 Records from a Dataset Total of 98,655)

Category		Count	Percent
Benign	57.14%	84	57.14%
Likely Benign	23.13%	34	23.13%
Pathogenic	6.80%	10	6.80%
Uncertain Significance	6.12%	9	6.12%
other	4.08%	6	4.08%
Not Provided	1.36%	2	1.36%
Likely Pathogenic	1.36%	2	1.36%
Drug Response	0.00%	0	0.00%
Total	0.00% 11.4% 22.9% 34.3% 45.7% 57.1%	147	100.0%



Category Counts (123 Records from a Dataset Total of 98,655)

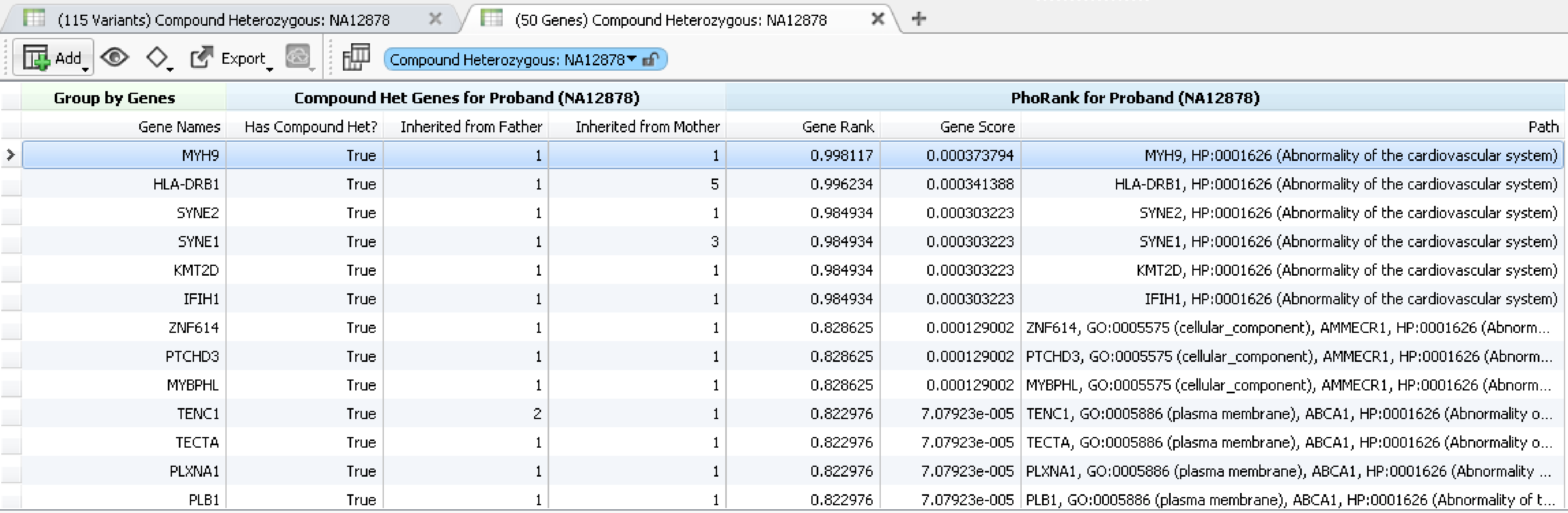
Category		Count	Percent
(1 Star) Classified by single submitter	84.35%	124	84.35%
(2 Stars) Classified by multiple submitters	12.93%	19	12.93%
(0 Stars) Not classified by submitter	1.36%	2	1.36%
(3 Stars) Reviewed by expert panel	1.36%	2	1.36%
Total	0.00% 16.9% 33.7% 50.6% 67.5% 84.4%	147	100.0%

Functional Predictions

- **Functional predictions use algorithms to determine the expected consequence of variants (or the resulting amino acid substitutions).**
- **dbNSFP**
  - The Database for NonSynonymous Functional Predictions (dbNSFP) is a free tool developed by Dr. Xiaoming Liu.
  - Catalogs pre-computed conservation and functional prediction scores for all possible missense SNVs in the genome
  - Methods include SIFT, PolyPhen-2, MutationTaster, MutationAssessor, FATHMM, more
- **dbscSNV**
  - Companion to dbNSFP that scores variants in splice consensus regions
  - Variants in these regions may disrupt normal gene expression and/or function
- **dbNSFP and dbscSNV are both accessible in VarSeq**

Variant/Gene Ranking

- PhoRank algorithm in VarSeq uses HPO and GO terminology to score relationships between genes and phenotypes
- Very useful to prioritize a long list of variants for individual review
- Based on PHEVOR method.


(115 Variants) Compound Heterozygous: NA12878
(50 Genes) Compound Heterozygous: NA12878
Add
Export
Compound Heterozygous: NA12878

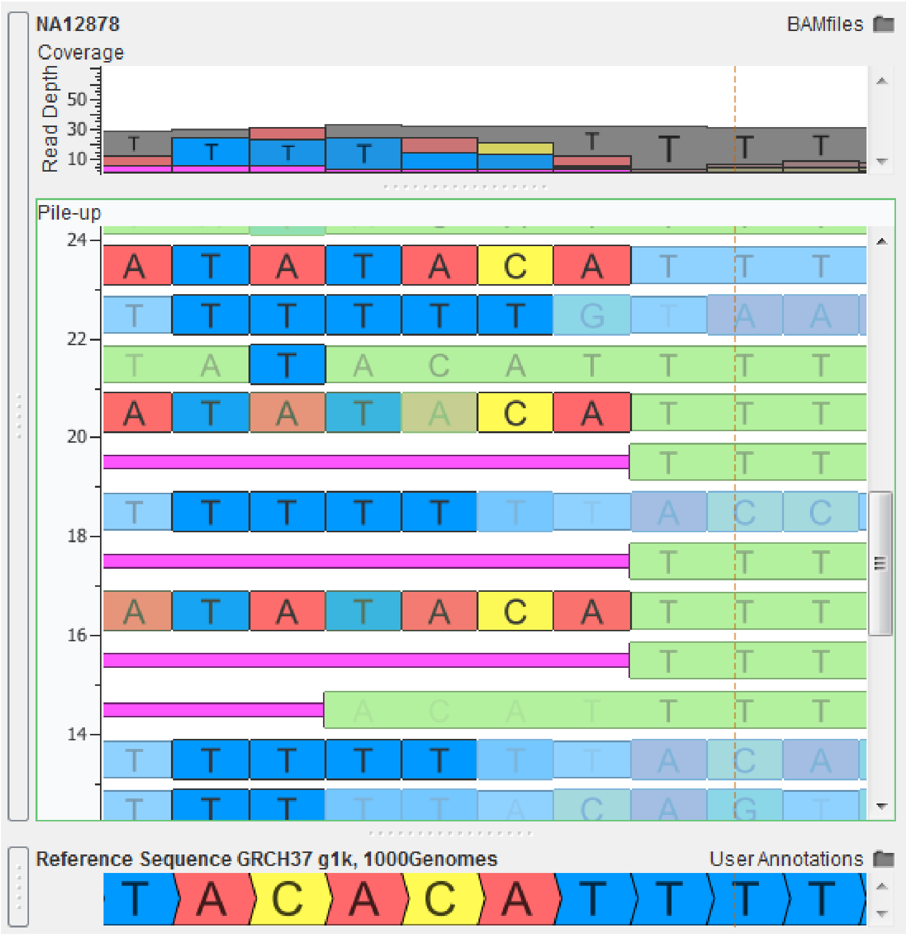
Group by Genes	Compound Het Genes for Proband (NA12878)			PhoRank for Proband (NA12878)		
Gene Names	Has Compound Het?	Inherited from Father	Inherited from Mother	Gene Rank	Gene Score	Path
MYH9	True	1	1	0.998117	0.000373794	MYH9, HP:0001626 (Abnormality of the cardiovascular system)
HLA-DRB1	True	1	5	0.996234	0.000341388	HLA-DRB1, HP:0001626 (Abnormality of the cardiovascular system)
SYNE2	True	1	1	0.984934	0.000303223	SYNE2, HP:0001626 (Abnormality of the cardiovascular system)
SYNE1	True	1	3	0.984934	0.000303223	SYNE1, HP:0001626 (Abnormality of the cardiovascular system)
KMT2D	True	1	1	0.984934	0.000303223	KMT2D, HP:0001626 (Abnormality of the cardiovascular system)
IFIH1	True	1	1	0.984934	0.000303223	IFIH1, HP:0001626 (Abnormality of the cardiovascular system)
ZNF614	True	1	1	0.828625	0.000129002	ZNF614, GO:0005575 (cellular_component), AMMECR1, HP:0001626 (Abnorm...
PTCHD3	True	1	1	0.828625	0.000129002	PTCHD3, GO:0005575 (cellular_component), AMMECR1, HP:0001626 (Abnorm...
MYBPHL	True	1	1	0.828625	0.000129002	MYBPHL, GO:0005575 (cellular_component), AMMECR1, HP:0001626 (Abnorm...
TENC1	True	2	1	0.822976	7.07923e-005	TENC1, GO:0005886 (plasma membrane), ABCA1, HP:0001626 (Abnormality o...
TECTA	True	1	1	0.822976	7.07923e-005	TECTA, GO:0005886 (plasma membrane), ABCA1, HP:0001626 (Abnormality o...
PLXNA1	True	1	1	0.822976	7.07923e-005	PLXNA1, GO:0005886 (plasma membrane), ABCA1, HP:0001626 (Abnormality ...
PLB1	True	1	1	0.822976	7.07923e-005	PLB1, GO:0005886 (plasma membrane), ABCA1, HP:0001626 (Abnormality of t...

QC Considerations

- **Variant QC**
- **Rare variants deserve special attention**
- **VCF/BAM Data:**
  - Depth - DP
  - Quality - GQ
  - Strand bias
  - Etc.
- **Public Annotations:**
  - “Mappability”

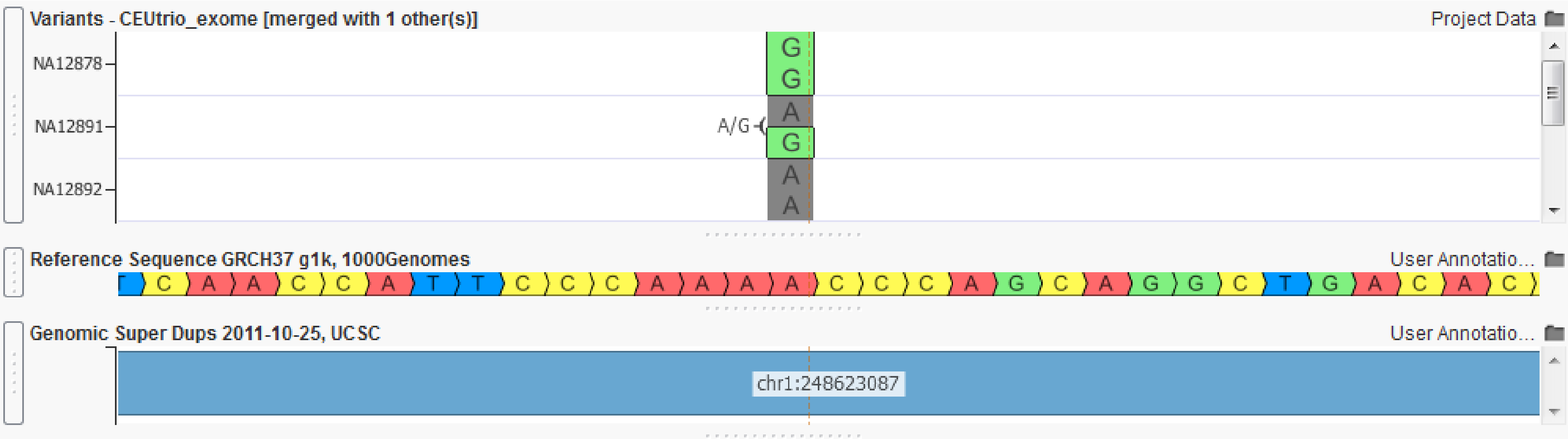
Mappability Annotations

- **The human reference genome has assembly gaps and other “difficult” regions**
- **NGS technology sequences short DNA fragments which are the aligned to the reference genome**
  - Most sequences are aligned correctly
  - Some sequences can’t be aligned uniquely
  - Some sequences may be incorrectly aligned
- **Luckily, we can predict many of the trouble spots**



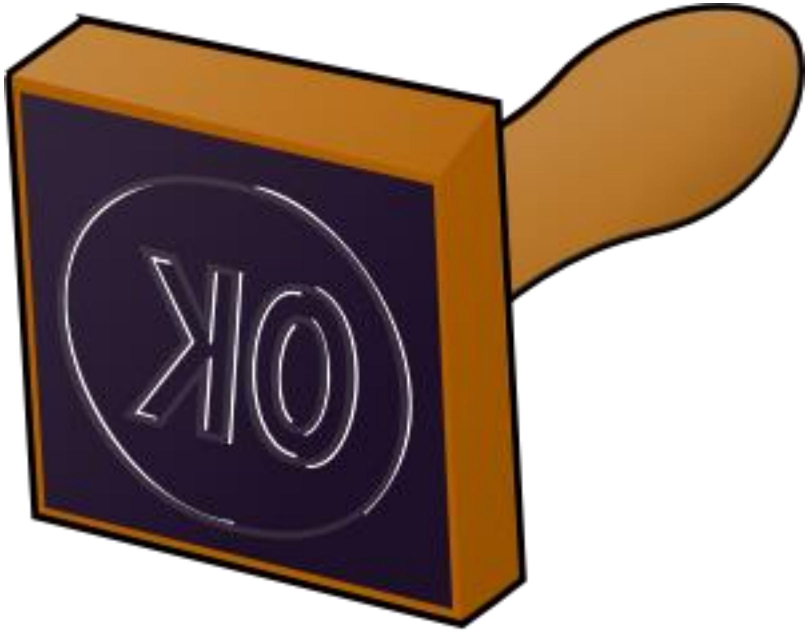
Segmental Duplications

- **Segmental duplications are a common confounder**
- **UCSC “Genomic Super Dups” annotation available through VarSeq**
- **Recent Example (below):**
  - Apparent UPD feature in family trio was determined to be an artifact of seg. duplication
  - Large chromosome segment duplicated elsewhere with >98% similarity

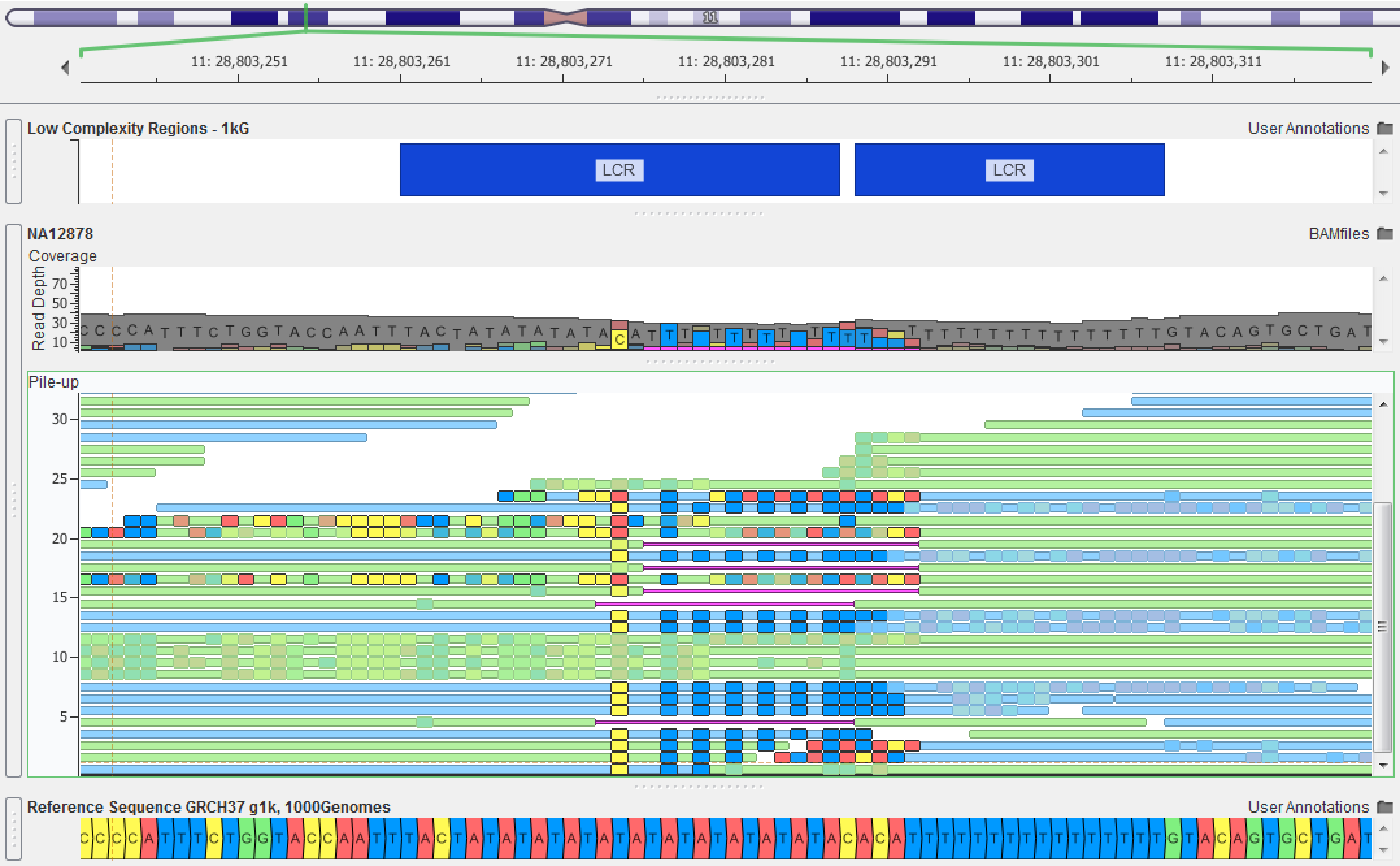

Variants - CEUtrio_exome [merged with 1 other(s)]
Project Data
NA12878
NA12891
NA12892
G
G
A/G
A
G
A
A
Reference Sequence GRCH37 g1k, 1000Genomes
User Annotatio...
T C A A C C A T T C C C A A A A C C C A G C A G G C T G A C A C
Genomic Super Dups 2011-10-25, UCSC
User Annotatio...
chr1:248623087

Emerging Standards

- **Several organizations working on best practices guidelines for genome mappability**
  - 1000 Genomes Project
  - Genome in a Bottle Consortium
  - Global Alliance for Genomics and Health (GA4GH)
  - National Institute of Standards and Technology
- **Downloadable annotations available for many types of features:**
  - Mappability by read length
  - High G-C content regions
  - Low complexity
  - Segmental duplications
  - Etc.

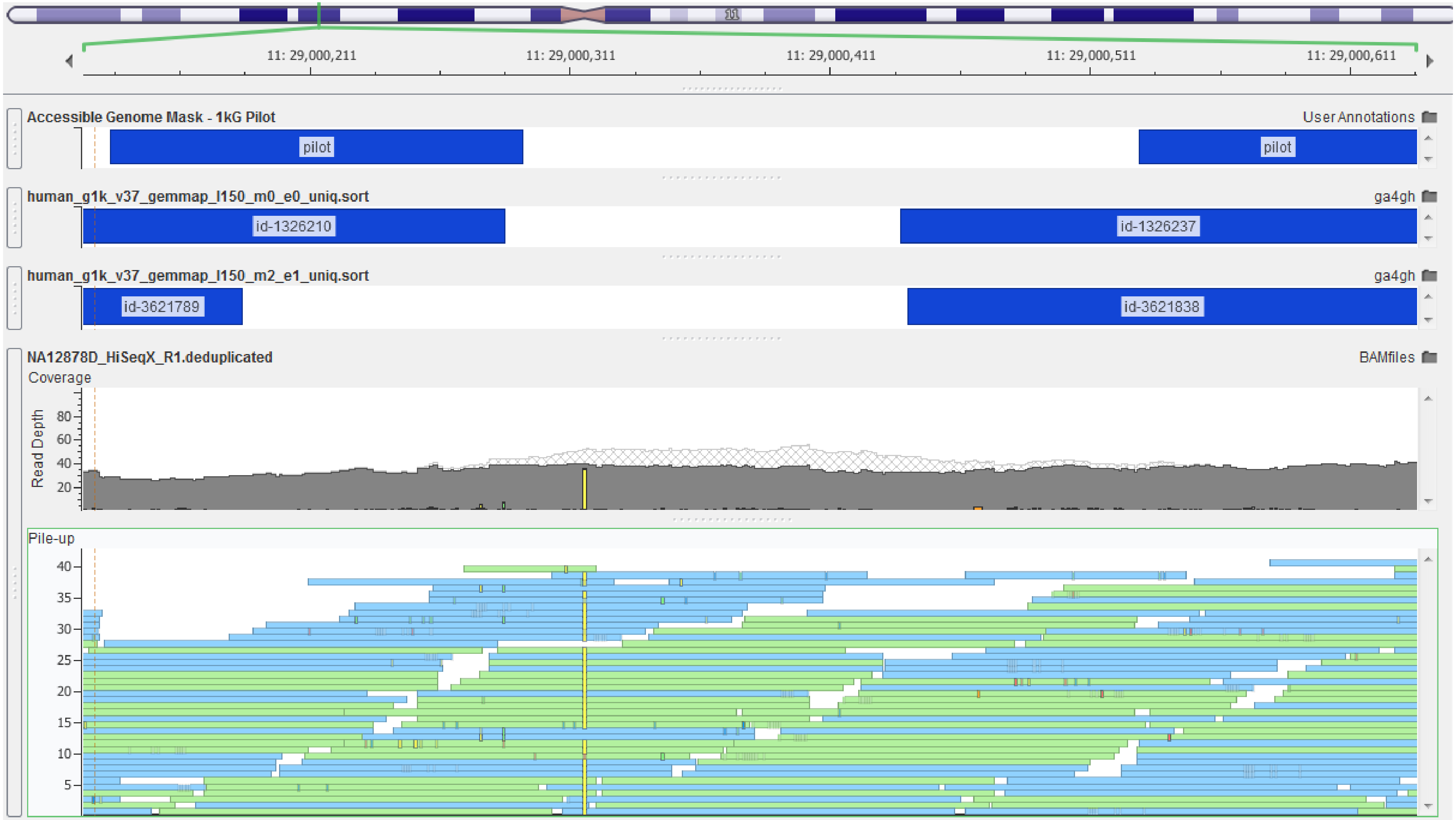


Example: 1kG Low Complexity Regions



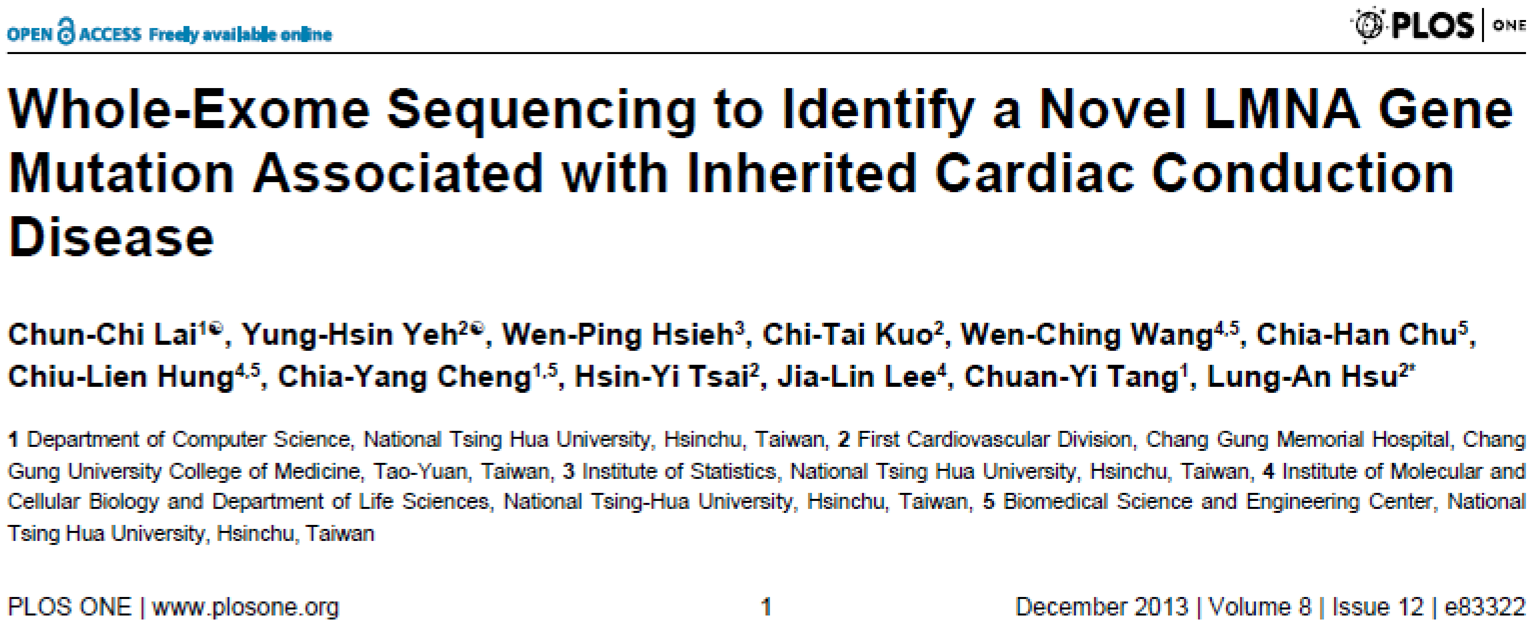
11
11: 28,803,251
11: 28,803,261
11: 28,803,271
11: 28,803,281
11: 28,803,291
11: 28,803,301
11: 28,803,311
Low Complexity Regions - 1kG
User Annotations
LCR
LCR
NA12878
BAMfiles
Coverage
Read Depth
70
50
30
10
Pile-up
30
25
20
15
10
5
Reference Sequence GRCH37 g1k, 1000Genomes
User Annotations
C C C C A T T T C T G G T A C C A A T T T A C T A T A T A T A T A T A T A T A T A T A T A C A C A T T T T T T T T T T T T T T T T G T A C A G T G C T G A T

Example: GA4GH 150-bp Mappability



VarSeq Demonstration Data

- Exome sequencing of five individuals from family with familial cardiac conduction disease (CCD)
- Raw sequence data obtained from SRA



OPEN ACCESS Freely available online

PLOS | ONE

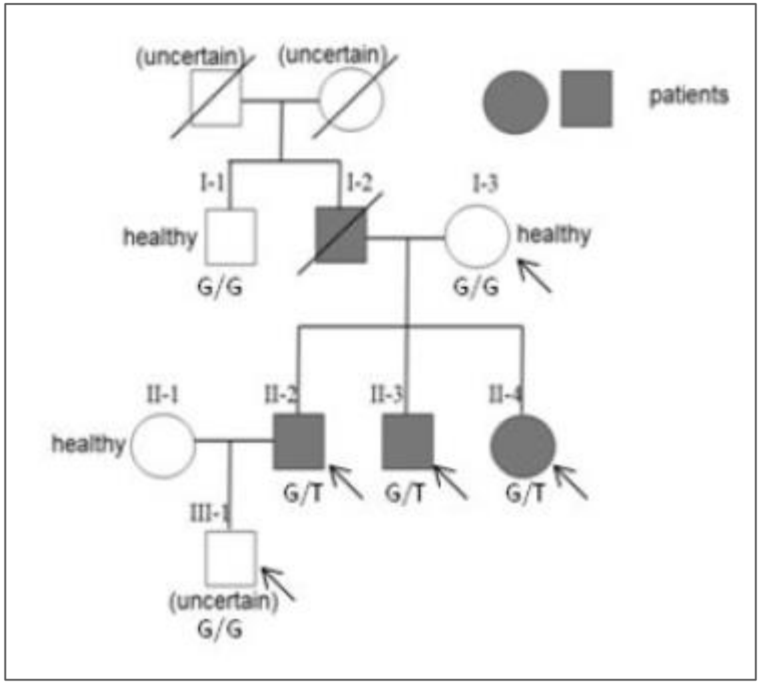
Whole-Exome Sequencing to Identify a Novel LMNA Gene Mutation Associated with Inherited Cardiac Conduction Disease

Chun-Chi Lai[1¶], Yung-Hsin Yeh[2¶], Wen-Ping Hsieh[3], Chi-Tai Kuo[2], Wen-Ching Wang[4,5], Chia-Han Chu[5], Chiu-Lien Hung[4,5], Chia-Yang Cheng[1,5], Hsin-Yi Tsai[2], Jia-Lin Lee[4], Chuan-Yi Tang[1], Lung-An Hsu[2*]

1 Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan, **2** First Cardiovascular Division, Chang Gung Memorial Hospital, Chang Gung University College of Medicine, Tao-Yuan, Taiwan, **3** Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan, **4** Institute of Molecular and Cellular Biology and Department of Life Sciences, National Tsing-Hua University, Hsinchu, Taiwan, **5** Biomedical Science and Engineering Center, National Tsing Hua University, Hsinchu, Taiwan

PLOS ONE | www.plosone.org 1 December 2013 | Volume 8 | Issue 12 | e83322

Workflow Discussion Points



- Male-to-male transmission makes X-linked model unlikely
- May follow dominant or recessive transmission
- Inherited forms of CCD are rare
- Family has East Asian ancestry

varSEQ™

[Demonstration]

Why VarSeq?

Flexible

Simple

Scalable

- Variant annotation, filtering and ranking
- Exploratory analysis
- Powerful GUI with immediate feedback
- Rich visualizations with GenomeBrowse integration

Questions or more info:

- Email info@goldenhelix.com
- Request an evaluation of the software at www.goldenhelix.com

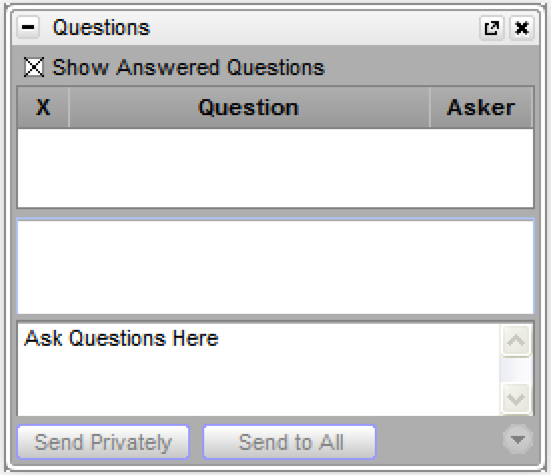
Complete toolset for SNP, CNV, and DNA & RNA Sequencing analysis

Annotation, filtering, and interpretation of variants with easy to use workflows.

Explore your NGS alignment and variant data. For FREE.

Questions?

Use the Questions pane in your GoToWebinar window



Questions
☒ Show Answered Questions
X
Question
Asker
Ask Questions Here
Send Privately
Send to All