VSClinical®

Using the GRCh38 reference assembly for clinical interpretation in VSClinical

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!

gsk GlaxoSmithKline

VarSEQ®

Variant Calling
Filtering and Annotation
Variant Interpretation
Clinical Reports
CNV Analysis
Pipeline: Run Workflows

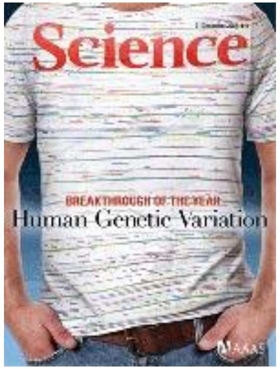
WAREHOUSE

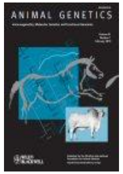
Variant Warehouse
Centralized Annotations
Hosted Reports
Sharing and Integration

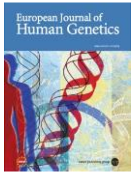
SNP & VARIATION SUITE

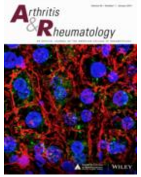
CNV Analysis
GWAS
Genomic Prediction
Large-N-Population Studies
RNA-Seq
Large-N CNV-Analysis

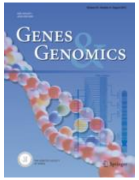
Cited in over 1,300 peer-reviewed publications











Over 350 customers globally

SickKids
THE HOSPITAL FOR
SICK CHILDREN

Cincinnati Children's
Hospital Medical Center

CH
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

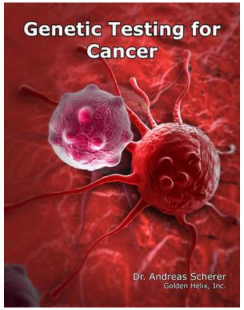
USDA
USDA Agricultural Research Service

FDA

Golden Helix – Who We Are

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

- REPUTATION
- TRUST
- EXPERIENCE



Genetic Testing for Cancer

Dr. Andreas Scherer
Golden Helix, Inc.

- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

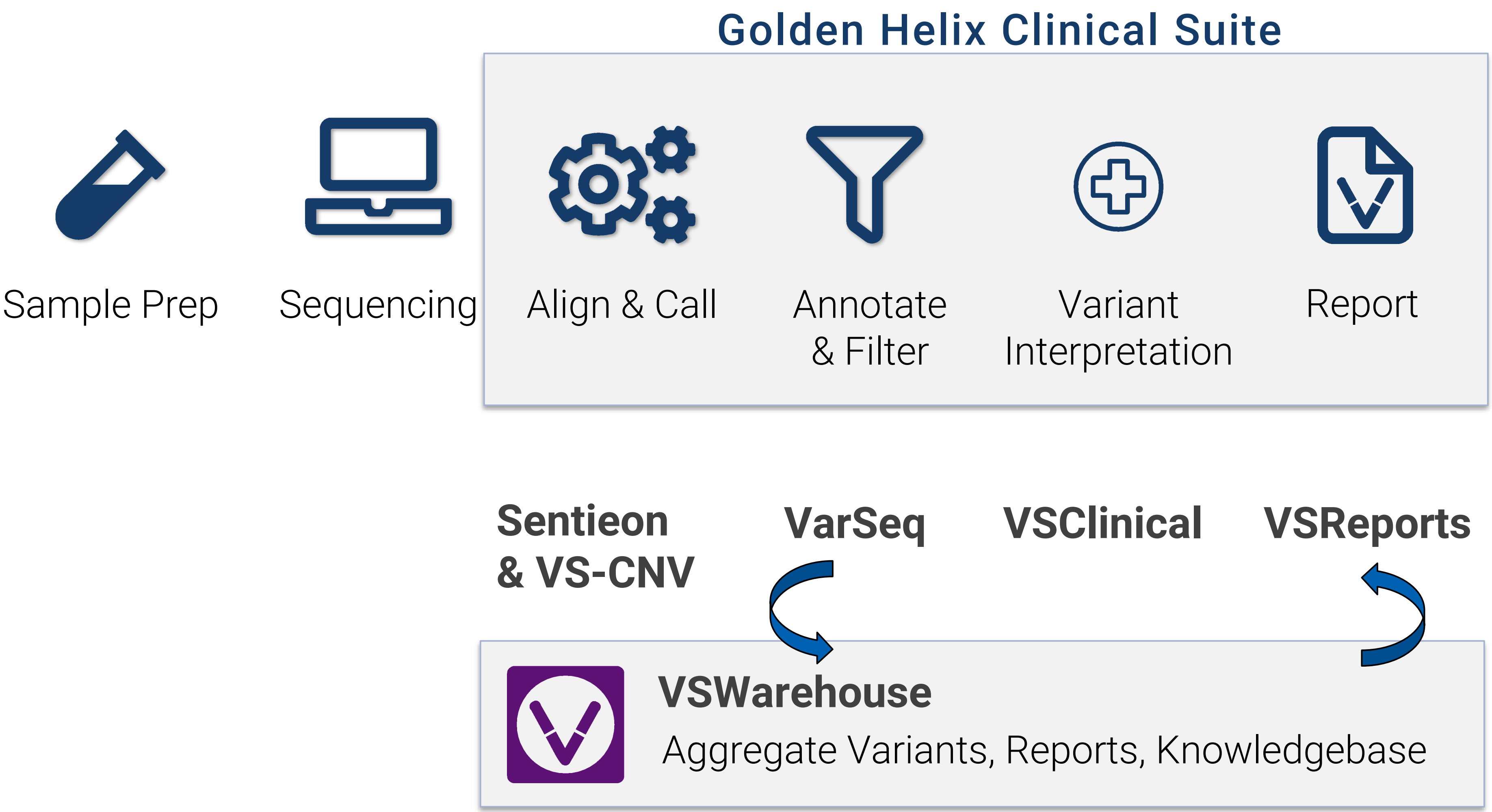
- TRAINING
- SUPPORT
- RESPONSIVENES S

- INNOVATION and SPEED
- CUSTOMIZATIONS

Agenda

- Genetic Testing with NGS
- Variant Representation
- Human Reference Genomes
- Implications for Variant Interpretation
- Demo using VarSeq + VSClinical
- Motivation for Using GRCh38
- Other Lab Considerations
- Thanks! / Q&A

NGS Genetics Testing Process



Golden Helix Clinical Suite

Sample Prep

Sequencing

Align & Call

Annotate & Filter

Variant Interpretation

Report

Sentieon & VS-CNV

VarSeq

VSClinical

VSReports

VSWarehouse

Aggregate Variants, Reports, Knowledgebase

Representing a Variant

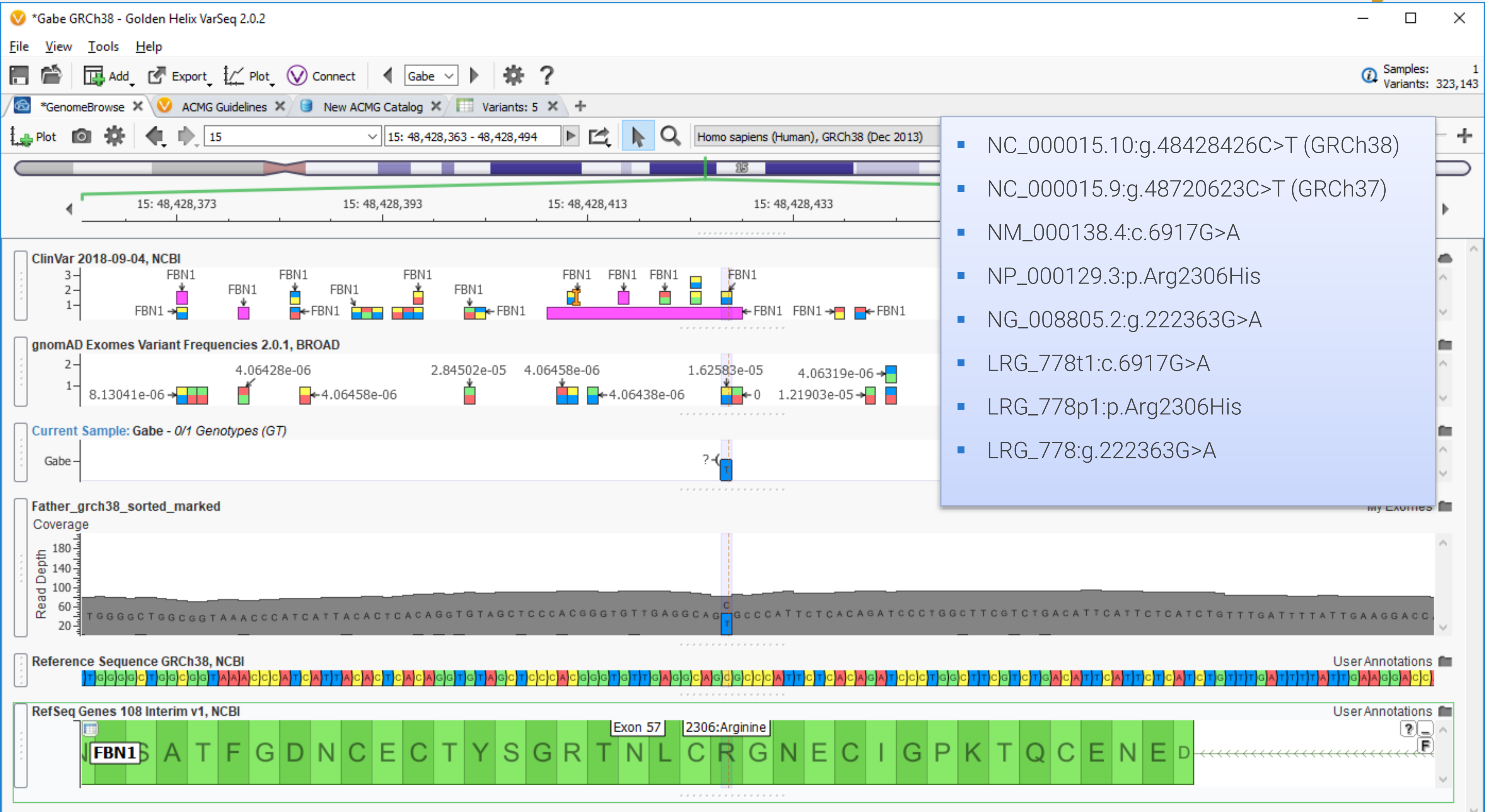
- **Genomic:**
  - chr2: 47,641,560 A/T
  - NC_000002.11:g.47641560A>T
  - chr14: 51,378,590 TT/T
  - NC_000014.8 :g.51378593delT
- **Gene Coding Sequence:**
  - BRAF c.1799T>A
  - NM_058197.4:c.105dupG
  - LRG_218t1(*MSH2*):c.942+3A>T
- **Gene Protein Sequence**
  - DYX1C1 p.E417*
  - NP_000483.3:p.Phe508del

- **Genomic Representation Enables**
  - Precise lookup of annotations
  - Overlap / relation to genomic features
  - Representation of non-genic variants
- **Coding Representation Enables**
  - Genomic reference independent
  - UTR and Intronic variants
  - Informative representation of coding change
- **Protein Representation Enables**
  - Grouping of variants that result in same protein
  - Descriptive of effect on protein
  - Coordinates match domains and protein DBs

Genomes Are Just a Means to an End (Genes)

- **RefSeqGenes – mRNA sequence archive, with mappings to genomes**
  - Provided mappings to Locus Reference Gene (LRG) database
  - Use genome mappings by NCBI (through genome annotation builds). NOT UCSC
  - “Clinically Relevant” transcript in VarSeq:
    - Most commonly submitted to ClinVar
    - LRG if available, longest if tied
- **Ensembl – defined directly against the human genome**
  - More inclusive of genes discovered with high-throughput methods
  - Gencode subset – similar to RefSeqGenes in size / definition
- **Each have unique Accessions and Version Numbers**
  - Newer releases are provided only on GRCh38
  - GRCh37 mappings not being updated (“105 Interim” by special request)

Variant Representation and the Reference Genome



*Gabe GRCh38 - Golden Helix VarSeq 2.0.2

Samples: 1
Variants: 323,143

Chromosome: 15

15: 48,428,363 - 48,428,494

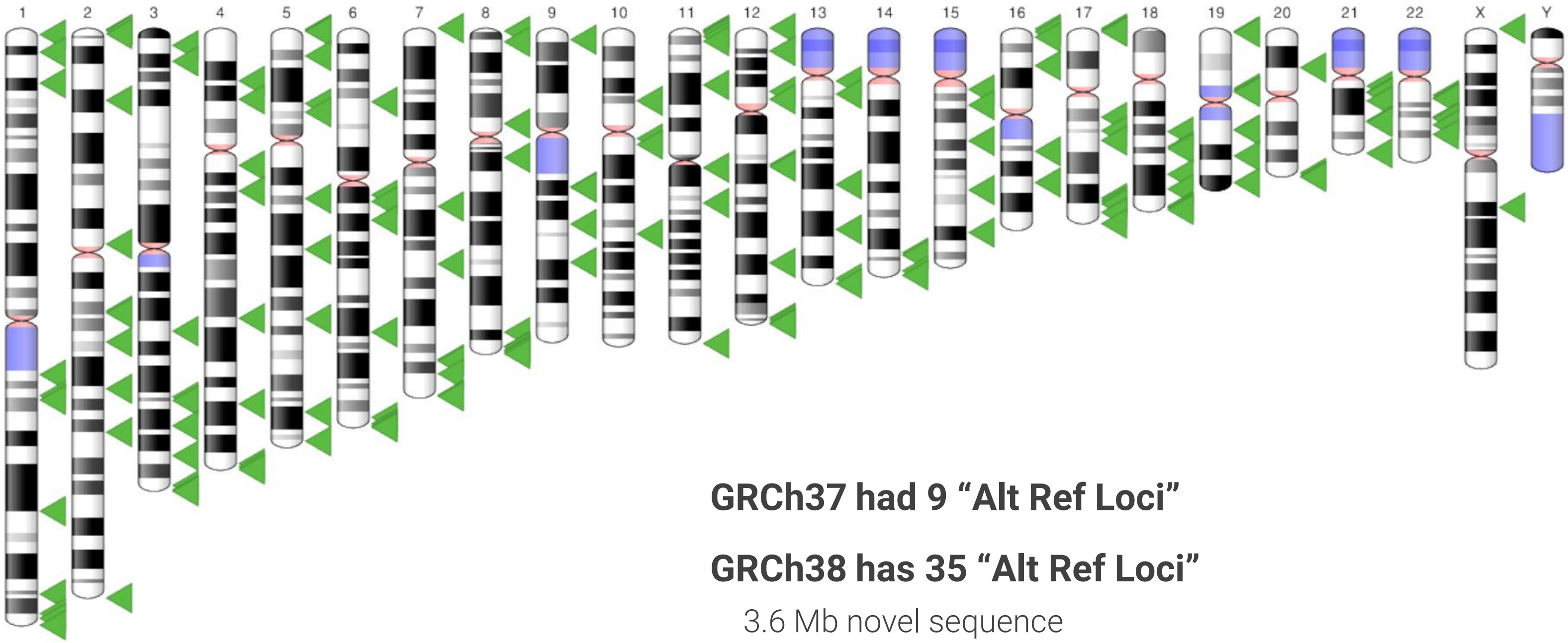
Homo sapiens (Human), GRCh38 (Dec 2013)

- NC_000015.10:g.48428426C>T (GRCh38)
- NC_000015.9:g.48720623C>T (GRCh37)
- NM_000138.4:c.6917G>A
- NP_000129.3:p.Arg2306His
- NG_008805.2:g.222363G>A
- LRG_778t1:c.6917G>A
- LRG_778p1:p.Arg2306His
- LRG_778:g.222363G>A

History of the Human Reference Genome

- **2003: Human Genome Project Declared Done**
- **2006: NCBI36 (hg18)**
  - Produced by the International Human Genome Sequencing Consortium
  - Used by first high throughput sequencers (Illumina GAII), pilot project of 1000 Genomes
  - UCSC uses its own sequential versioning, calling this hg18
- **2009: GRCh37 (hg19)**
  - Handed over to the Genome Reference Consortium (GRCh)
  - Used by the 1000 genome project (Phase I/II/III) in the era of the HiSeq 2000
- **2013: GRCh38 (hg38)**
  - ~100 assembly gaps updated, ~2000 erroneous alleles fixed
  - Included centromere models, mitochondrial reference, **alternate sequences**

Alternative Loci / “Haplotypes”



GRCh37 had 9 “Alt Ref Loci”

GRCh38 has 35 “Alt Ref Loci”

3.6 Mb novel sequence

153 genes

Up to 25% of these genes hare medically interpretable

Alignment support

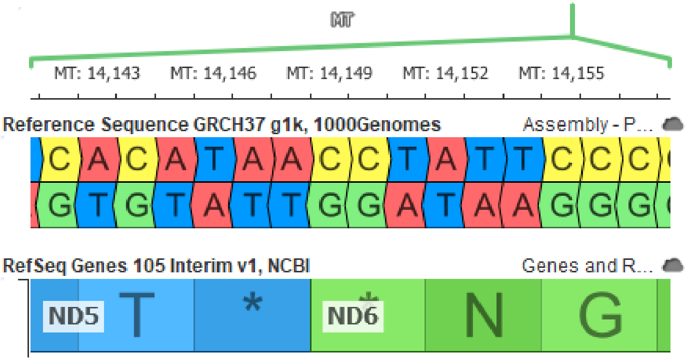
Before using, ensure aligner can support alt loci without flagging “multi-alignment” codes that cause reads to be filtered out / lost. BWA-MEM supports alt loci.

More than Chromosomes in your FASTA

- **Other bits of the reference:**
  - Un-localized scaffolds assigned to chromosomes
  - Unplaced scaffolds (not assigned to chromosomes
  - Patches Releases (i.e. GRCh37.p13, GRCh38.p12)
    - Types of “alt”, “fix” or “novel”
    - Not applied, and do not change the primary sequence
    - You can think of them as “known issues, with proposed fixes for next major release”
- **Other useful things to add for alignment purposes:**
  - A “decoy” reference genome segment as primary reference
    - DNA virus: human herpesvirus 4, type 1, aka Epstein-Barr virus (EBV)
    - Unique sequence found in HuRef (Craig Venter’s genome) or de novo assemblies
    - Other novel unaccounted for (or “novel” patch) sequence
  - Full set of HLA “haplotype” sequences, marked as “alternates”
- **Mitochondrial!**

The Human Mitochondrial

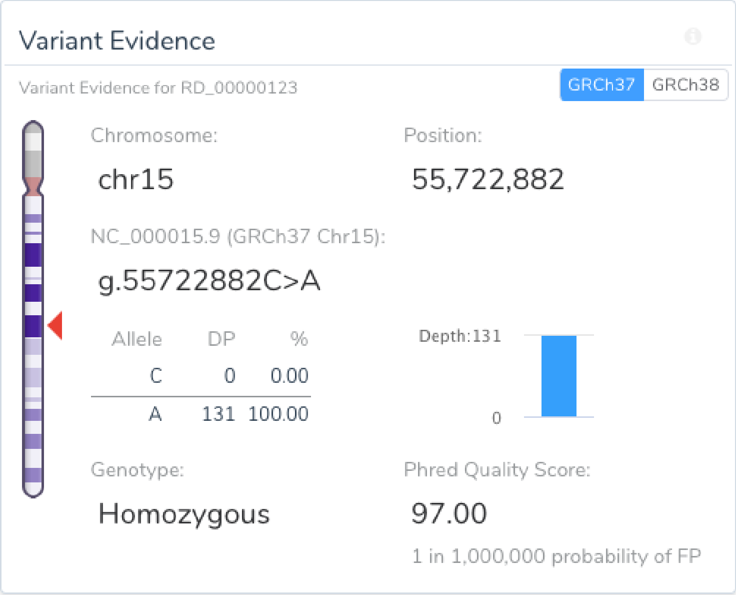
- **Our second genome:**
  - Only 16Kb long
  - Encodes 37 genes (product of energy and its storage in ATP)
  - Slightly different genetic code than nuclear genes:
    - UGA = tryptophan, AUA = methionine, and AGA and AGG = stop


MT
MT: 14,143
MT: 14,146
MT: 14,149
MT: 14,152
MT: 14,155
Reference Sequence GRCH37 g1k, 1000Genomes
Assembly - P...
C A C A T A A C C T A T T C C C
G T G T A T T G G A T A A G G G
RefSeq Genes 105 Interim v1, NCBI
Genes and R...
ND5
T
*
ND6
N
G

- **Sequence in 1981 as the "Cambridge Reference Sequence" (before HGP)**
  - 2014: "revised Cambridge Reference Sequence" or rCRS
    - 16,569bp long
    - 1000 genome project used with GRCh37 +decoy to create the "g1k" reference
    - This is the default for Golden Helix Sentieon pipeline and VarSeq interpretation
- **NCBI36 (hg18) Included a MT reference NC_001807 in 2006:**
  - Derived from a African (Yoruba) Individual
  - 16,571bp long, differing from the rCRS by 40 variants
  - Removed from GenBank, don't publish with this M!
  - UCSC hg19 includes NC_001807 as "M" and still uses it today!
  - Next VarSeq version drops support for this "hg19" genome

Variant Interpretation in VSClinical

- **Evaluate and Classify Variants using ACMG Guidelines:**
  - Focused workflow to evaluate criteria relevant to each variant, resulting in final classification
  - Aggregates annotations from population and clinical resources
  - Customized visualizations and annotation presentations
  - Allows easy look-up and cross reference
- **Save Interpretations into Assessment Catalogs:**
  - New samples have previous classifications brought in
  - See previous interpretations, review and update
  - Can be potted for regional context
- **Use VarSeq's Filter, GenomeBrowse, VSReports:**
  - Customize to lab specific QC, annotation and filtering
  - Genomic context of variant vital to assess
  - VSReports allows custom presentation of VSClinical output



Variant Evidence

Variant Evidence for RD_00000123

Chromosome: chr15

Position: 55,722,882

NC_000015.9 (GRCh37 Chr15): g.55722882C>A

Allele	DP	%
C	0	0.00
A	131	100.00

Genotype: Homozygous

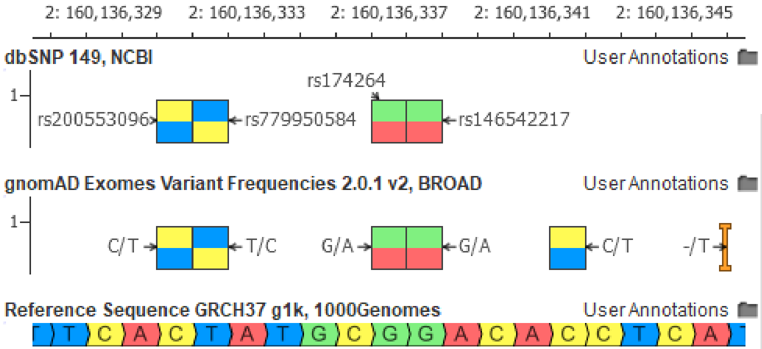
Phred Quality Score: 97.00

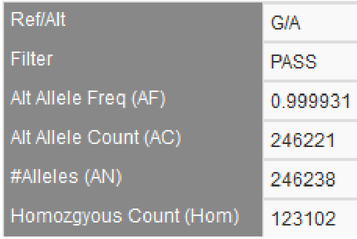
1 in 1,000,000 probability of FP

GRCh38: Implications for Variant Interpretation

- **Assembly Regions:**
  - Multiple Species Alignment
  - Repeat Regions / Low Complexity Regions
  - Genomic “Super Dups”
- **Genes (and Annotations)**
  - Functional Domains
  - Transcript Counts of Gene Constraint
- **Population Catalogs on GRCh37**
  - dbSNP
  - 1000 Genomes
  - ExAC / gnomAD Exomes / Genomes
- **Clinical Annotations**
  - ClinVar
  - CIVIC
  - OMIM (variants, genes, phenotypes)
- **Functional Annotations / Conservation**
  - CADD
  - SIFT/Polyphen/Missense Badness
  - Conservation scores

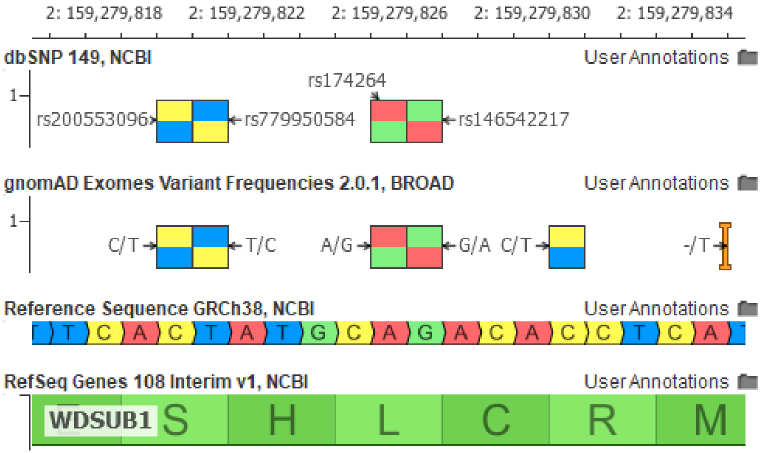
GRCh37: rs174264

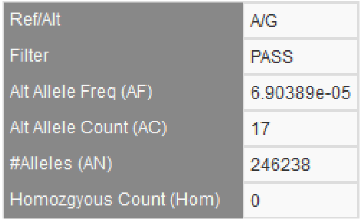




Ref/Alt	G/A
Filter	PASS
Alt Allele Freq (AF)	0.999931
Alt Allele Count (AC)	246221
#Alleles (AN)	246238
Homozgyous Count (Hom)	123102

GRCh37: rs174264



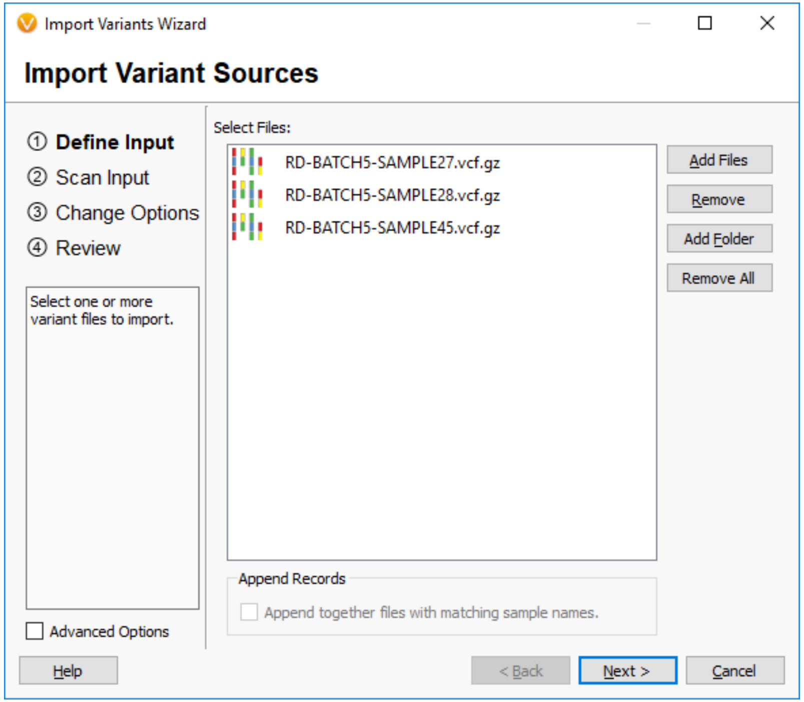


Ref/Alt	A/G
Filter	PASS
Alt Allele Freq (AF)	6.90389e-05
Alt Allele Count (AC)	17
#Alleles (AN)	246238
Homozgyous Count (Hom)	0

Substitution Leu (leucine) → Pro (proline) at 173
Leucine conserved in all vertebrates!

VarSeq Import LiftOver

Start with GRCh37 VCFs:



Import Variants Wizard

Import Variant Sources

① **Define Input**
② Scan Input
③ Change Options
④ Review

Select one or more variant files to import.

□ Advanced Options

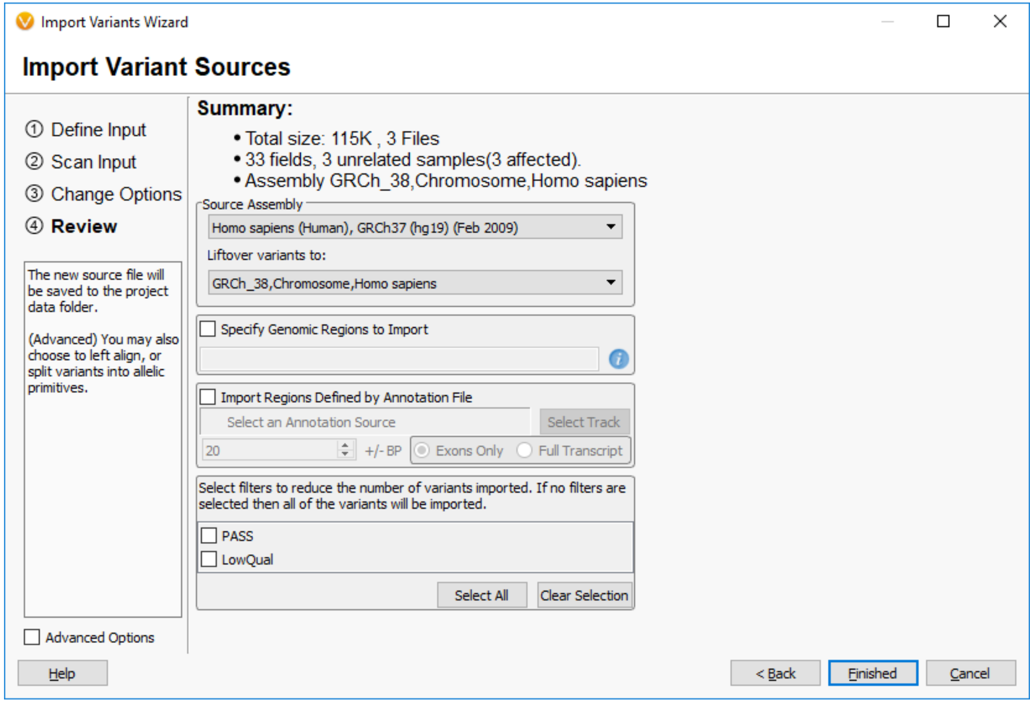
Select Files:

- RD-BATCH5-SAMPLE27.vcf.gz
- RD-BATCH5-SAMPLE28.vcf.gz
- RD-BATCH5-SAMPLE45.vcf.gz

Append Records

□ Append together files with matching sample names.

LiftOver to GRCh38:



Import Variants Wizard

Import Variant Sources

① Define Input
② Scan Input
③ Change Options
④ **Review**

The new source file will be saved to the project data folder.

(Advanced) You may also choose to left align, or split variants into allelic primitives.

□ Advanced Options

Summary:

- Total size: 115K , 3 Files
- 33 fields, 3 unrelated samples(3 affected).
- Assembly GRCh_38,Chromosome,Homo sapiens

Source Assembly: Homo sapiens (Human), GRCh37 (hg19) (Feb 2009)

Liftover variants to: GRCh_38,Chromosome,Homo sapiens

□ Specify Genomic Regions to Import

□ Import Regions Defined by Annotation File

Select an Annotation Source | Select Track

20 +/- BP ○ Exons Only ○ Full Transcript

Select filters to reduce the number of variants imported. If no filters are selected then all of the variants will be imported.

□ PASS
□ LowQual

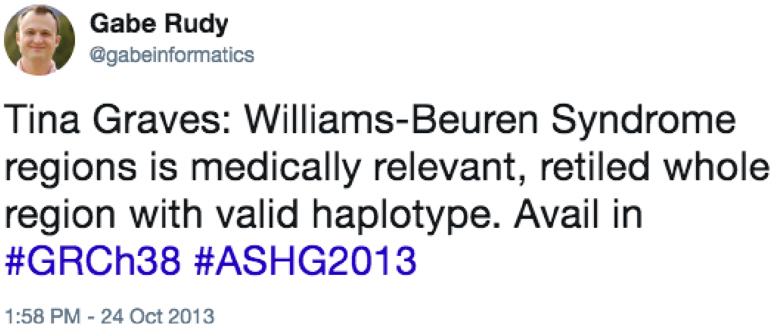
Or the Other Way Around! GRCh38 => GRCh37

VSClinical®

[Demo in VarSeq]

Reasons to Switch to GRCh38

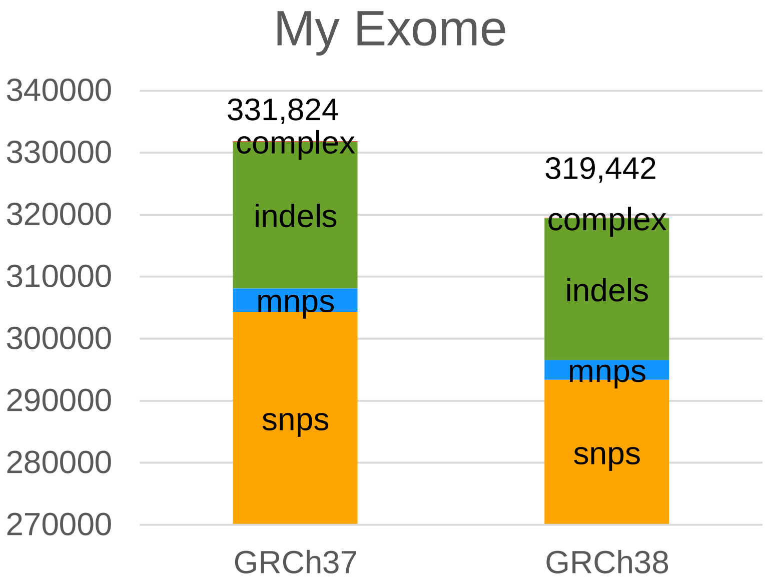
- **Better for alignment**
  - More reads mapped
  - Fewer variants called
- **Better gene representations**
  - Fewer “frame-fixing” introns
  - Some genes fixed/improved
- **Newer annotations are GRCh38**
  - Large consortiums are switching to GRCh38 first:
    - Cancer: ICGC, COSMIC
    - TopMed (65K WGS)



Gabe Rudy
@gabeinformatics

Tina Graves: Williams-Beuren Syndrome regions is medically relevant, retiled whole region with valid haplotype. Avail in #GRCh38 #ASHG2013

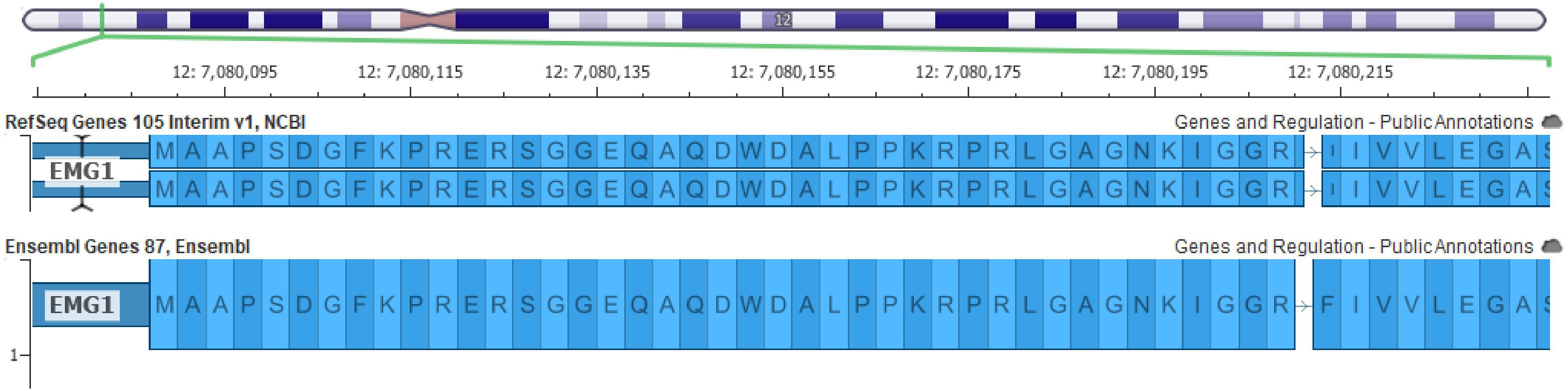
1:58 PM - 24 Oct 2013



Better Gene Representation

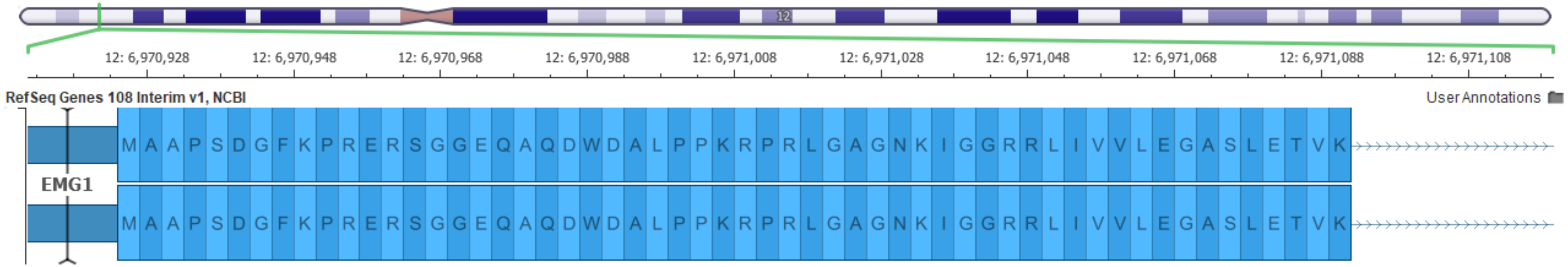
- The human genome does not necessarily contain the mRNA sequence in RefSeq
- “Frame-fixing” intron introduced in alignment of mRNA coding sequence to human reference:

EMG1 on GRCh37:



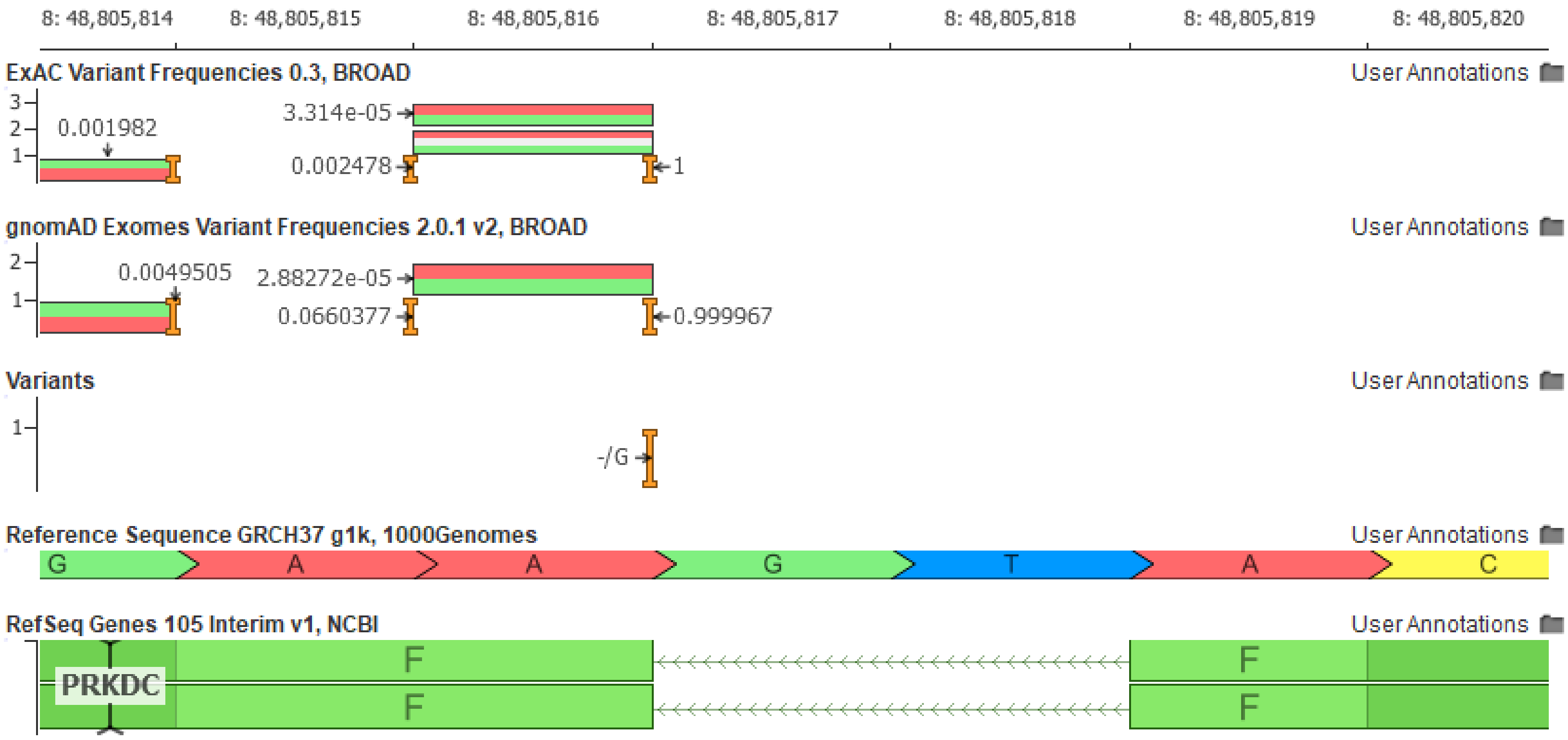
12
12: 7,080,095
12: 7,080,115
12: 7,080,135
12: 7,080,155
12: 7,080,175
12: 7,080,195
12: 7,080,215
RefSeq Genes 105 Interim v1, NCBI
Genes and Regulation - Public Annotations
EMG1
M A A P S D G F K P R E R S G G E Q A Q D W D A L P P K R P R L G A G N K I G G R I I V V L E G A S
M A A P S D G F K P R E R S G G E Q A Q D W D A L P P K R P R L G A G N K I G G R I I V V L E G A S
Ensembl Genes 87, Ensembl
Genes and Regulation - Public Annotations
EMG1
M A A P S D G F K P R E R S G G E Q A Q D W D A L P P K R P R L G A G N K I G G R F I V V L E G A S
1

EMG1 on GRCh38:

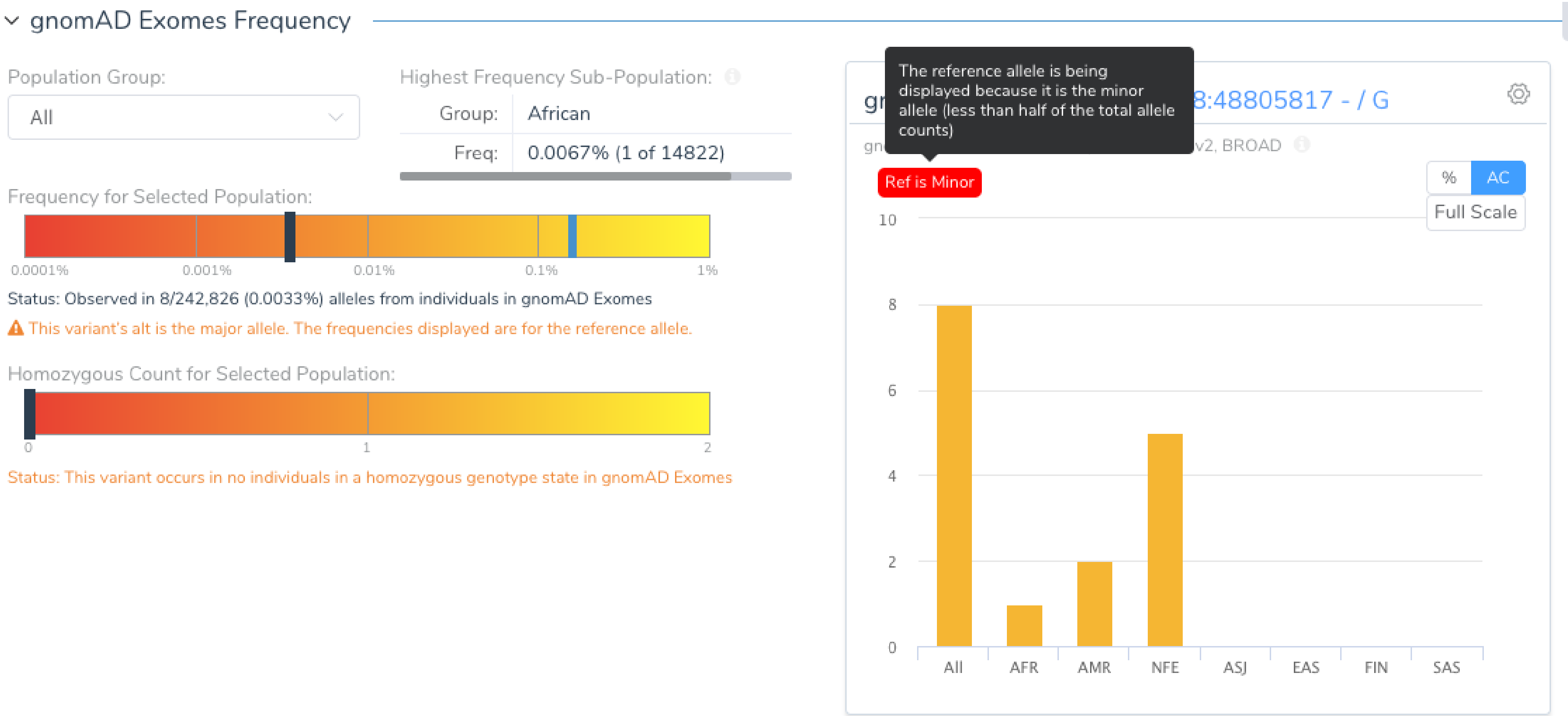


12
12: 6,970,928
12: 6,970,948
12: 6,970,968
12: 6,970,988
12: 6,971,008
12: 6,971,028
12: 6,971,048
12: 6,971,068
12: 6,971,088
12: 6,971,108
RefSeq Genes 108 Interim v1, NCBI
User Annotations
EMG1
M A A P S D G F K P R E R S G G E Q A Q D W D A L P P K R P R L G A G N K I G G R R L I V V L E G A S L E T V K
M A A P S D G F K P R E R S G G E Q A Q D W D A L P P K R P R L G A G N K I G G R R L I V V L E G A S L E T V K

Some Variants are Pure “Reference Artifacts”



Some Variants are Pure “Reference Artifacts”



gnomAD Exomes Frequency

Population Group: All

Highest Frequency Sub-Population:

Group:	African
Freq:	0.0067% (1 of 14822)

Frequency for Selected Population:

Status: Observed in 8/242,826 (0.0033%) alleles from individuals in gnomAD Exomes

⚠ This variant's alt is the major allele. The frequencies displayed are for the reference allele.

Homozygous Count for Selected Population:

Status: This variant occurs in no individuals in a homozygous genotype state in gnomAD Exomes

gr 8:48805817 - / G

gn v2, BROAD

The reference allele is being displayed because it is the minor allele (less than half of the total allele counts)

Ref is Minor

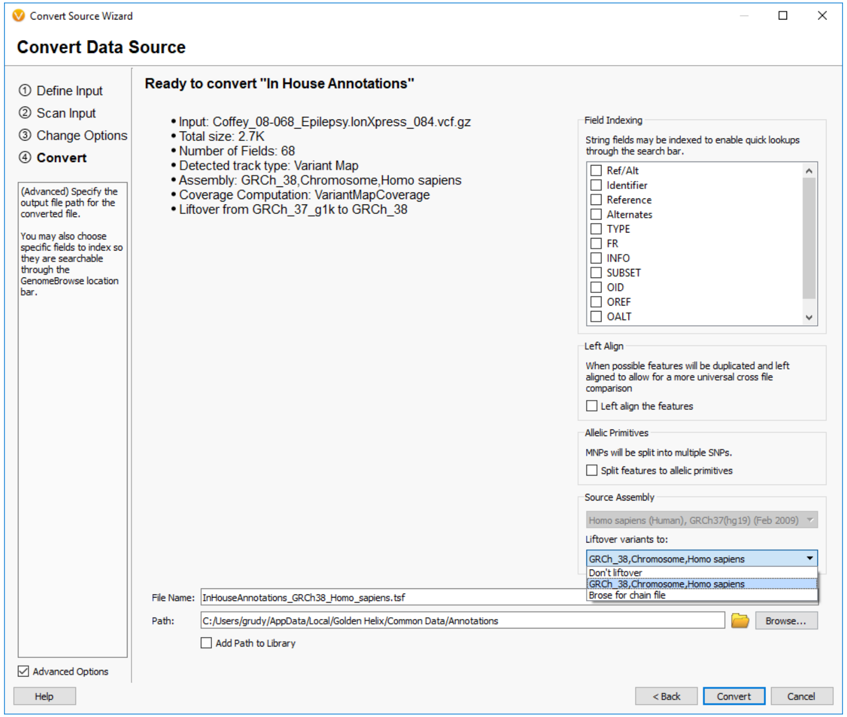
% AC

Full Scale

Considerations for Transitioning your Lab

- **Switching your Secondary Pipeline**
- **Your Genomic Variants Being Saved:**
  - VSClinical Catalog / Assessment Catalogs
  - Catalog of Observed CNVs
  - VSWarehouse Projects (all variants from samples)
  - Target capture annotations
  - Custom in-house annotations
- **Converting Existing Data:**
  - Re-import variants using import Liftover
  - Export/import catalogs using Liftover
  - Convert custom annotations using Liftover

Liftover Using Our Convert Wizard:



Convert Source Wizard

Convert Data Source

① Define Input
② Scan Input
③ Change Options
④ **Convert**

(Advanced) Specify the output file path for the converted file.

You may also choose specific fields to index so they are searchable through the GenomeBrowse location bar.

Ready to convert "In House Annotations"

- Input: Coffey_08-068_Epilepsy.IonXpress_084.vcf.gz
- Total size: 2.7K
- Number of Fields: 68
- Detected track type: Variant Map
- Assembly: GRCh_38,Chromosome,Homo sapiens
- Coverage Computation: VariantMapCoverage
- Liftover from GRCh_37_g1k to GRCh_38

Field Indexing

String fields may be indexed to enable quick lookups through the search bar.

□ Ref/Alt
□ Identifier
□ Reference
□ Alternates
□ TYPE
□ FR
□ INFO
□ SUBSET
□ OID
□ OREF
□ OALT

Left Align

When possible features will be duplicated and left aligned to allow for a more universal cross file comparison

□ Left align the features

Allelic Primitives

MNPs will be split into multiple SNPs.

□ Split features to allelic primitives

Source Assembly

Homo sapiens (Human), GRCh37(hg19) (Feb 2009)

Liftover variants to:

GRCh_38,Chromosome,Homo sapiens

Don't liftover
GRCh_38,Chromosome,Homo sapiens
Brose for chain file

File Name: InHouseAnnotations_GRCh38_Homo_sapiens.tsf

Path: C:/Users/grudy/AppData/Local/Golden Helix/Common Data/Annotations

□ Add Path to Library

☑ Advanced Options

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.

ASHG 2018

- Booth 408
- Live demos and CoLab Sessions
- Unveiling our new t-shirt designs
- Chance to win some iPads

ASHG2018

SAN DIEGO • OCTOBER 16-20, 2018

SHARING DISCOVERIES. SHAPING OUR FUTURE.

See us at ASHG 2018 Booth #408

Exhibit Dates: October 17-19