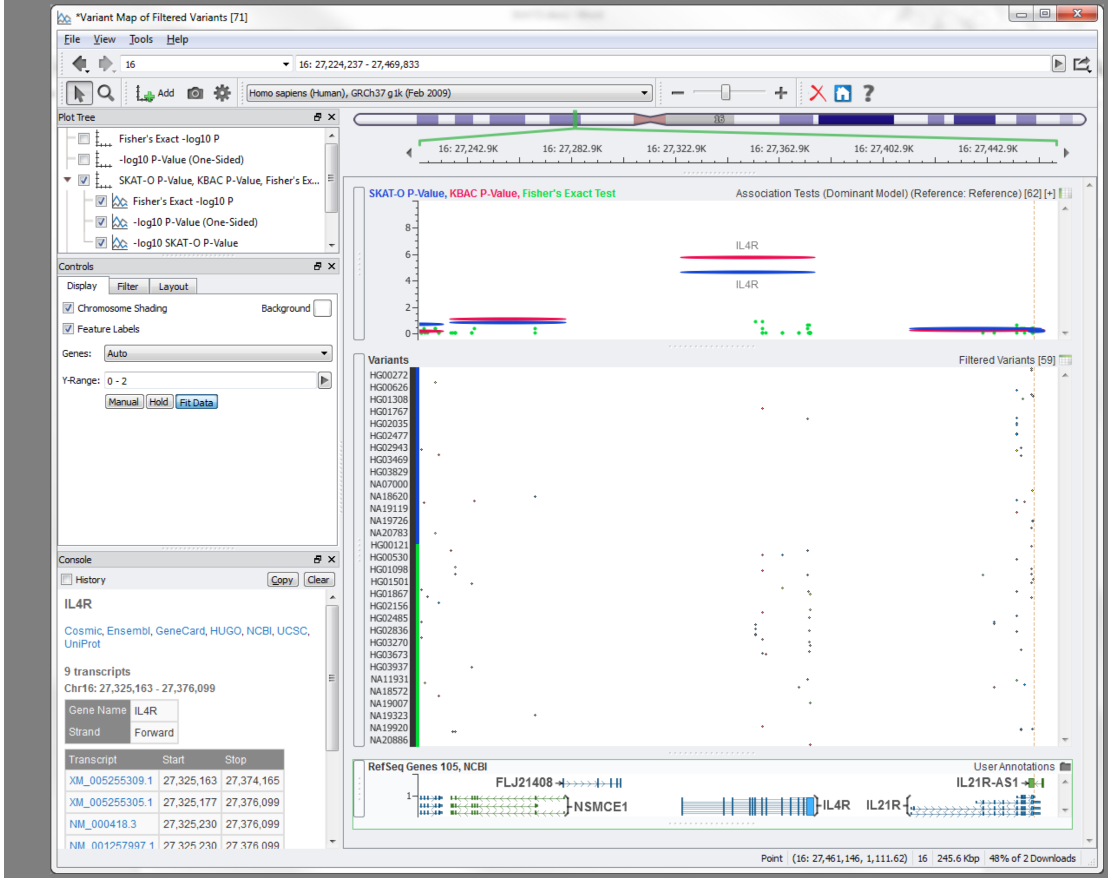


Chromosome: 16

Location: 16: 27,224,237 - 27,469,833

Genome: Homo sapiens (Human), GRCh37 g1k (Feb 2009)

Plot Tree:
- ☐ Fisher's Exact -log10 P
- ☐ -log10 P-Value (One-Sided)
- ☑ SKAT-O P-Value, KBAC P-Value, Fisher's Ex...
  - ☑ Fisher's Exact -log10 P
  - ☑ -log10 P-Value (One-Sided)
  - ☑ -log10 SKAT-O P-Value

☑ Chromosome Shading

Background □

☑ Feature Labels

Genes: Auto

Y-Range: 0 - 2

☐ History

IL4R

Cosmic, Ensembl, GeneCard, HUGO, NCBI, UCSC, UniProt

9 transcripts

Chr16: 27,325,163 - 27,376,099

Gene Name	IL4R
Strand	Forward

Transcript	Start	Stop
XM_005255309.1	27,325,163	27,374,165
XM_005255305.1	27,325,177	27,376,099
NM_000418.3	27,325,230	27,376,099
NM_001257997.1	27,325,230	27,376,099

Point (16: 27,461,146, 1,111.62) 16 245.6 Kbp 48% of 2 Downloads

Population-Based DNA Variant Analysis with Golden Helix SVS

Jan 21, 2015

Bryce Christensen, PhD

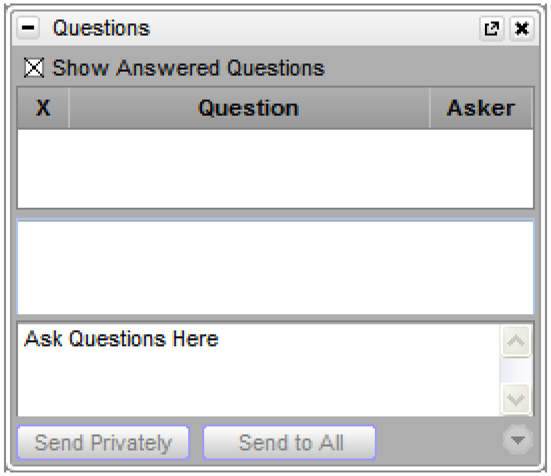
Statistical Geneticist / Director of Services

GOLDEN HELIX

Accelerating the Quest for Significance™

Questions during the presentation

Use the Questions pane in your GoToWebinar window



GOLDEN HELIX
Accelerating the Quest for Significance™

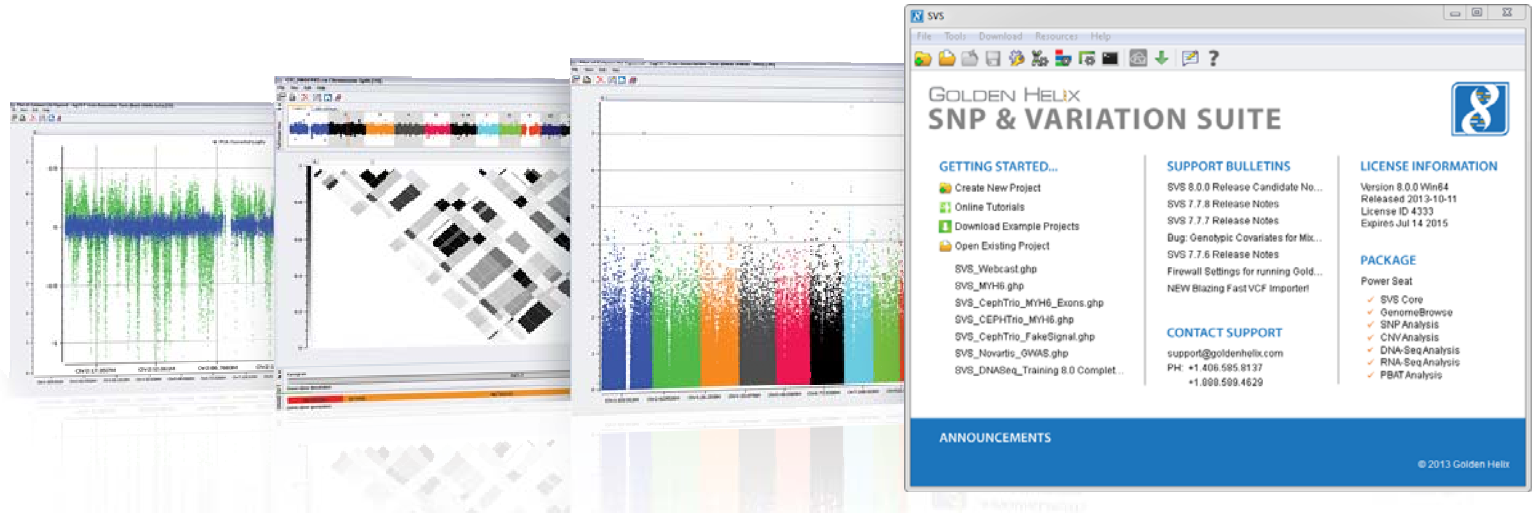
About Golden Helix

Leaders in Genetic Analytics

- Founded in 1998
- Multi-disciplinary: computer science, bioinformatics, statistics, genetics
- Software and analytic services

GOLDEN HELIX
Accelerating the Quest for Significance™

SNP & Variation Suite (SVS)



GOLDEN HELIX
SNP & VARIATION SUITE

GETTING STARTED...

- Create New Project
- Online Tutorials
- Download Example Projects
- Open Existing Project

SVS_Webcast.ghp
SVS_MYH6.ghp
SVS_CephTrio_MYH6_Exons.ghp
SVS_CEPHTrio_MYH6.ghp
SVS_CephTrio_FakeSignal.ghp
SVS_Novartis_GWAS.ghp
SVS_DNASeq_Training 8.0 Complet...

SUPPORT BULLETINS

SVS 8.0.0 Release Candidate No...
SVS 7.7.8 Release Notes
SVS 7.7.7 Release Notes
Bug: Genotypic Covariates for Mix...
SVS 7.7.6 Release Notes
Firewall Settings for running Gold...
NEW Blazing Fast VCF Importer!

CONTACT SUPPORT

support@goldenhelix.com
PH: +1.406.585.8137
+1.888.589.4629

LICENSE INFORMATION

Version 8.0.0 Win64
Released 2013-10-11
License ID 4333
Expires Jul 14 2015

PACKAGE

Power Seat

- ✓ SVS Core
- ✓ GenomeBrowse
- ✓ SNP Analysis
- ✓ CNV Analysis
- ✓ DNA-Seq Analysis
- ✓ RNA-Seq Analysis
- ✓ PBAT Analysis

ANNOUNCEMENTS

© 2013 Golden Helix

Core Features

- Powerful Data Management
- Rich Visualizations
- Robust Statistics
- Flexible
- Easy-to-use

Applications

- Genotype Analysis
- DNA sequence analysis
- CNV Analysis
- RNA-seq differential expression
- Family Based Association

GOLDEN HELIX
Accelerating the Quest for Significance™

Agenda

1 Define the problem: What is rare variant (RV) analysis?

2 Overview of RV analysis methods

3 NGS workflow design in SVS

4 Method Comparisons

5 Upstream analysis and QC considerations

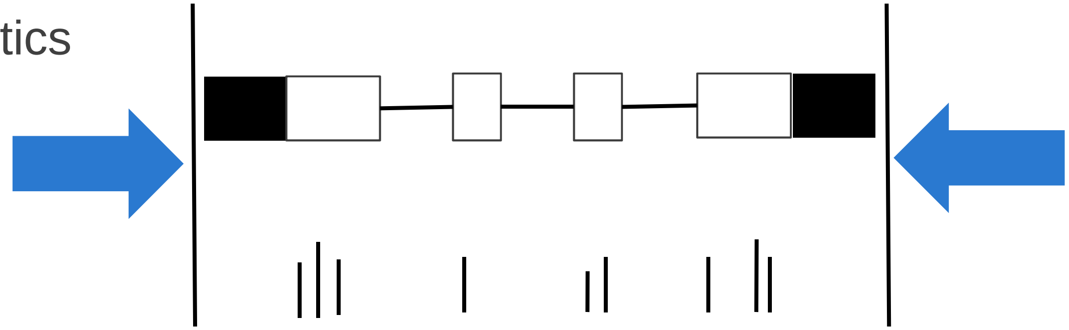
6 Q&A

The Problem

- Array-based GWAS has been the primary technology for gene-finding research for the past decade
- NGS technology, particularly whole-exome sequencing, makes it possible to include rare variants (RVs) in the analysis
- Individual RVs lack statistical power for standard GWAS approaches
  - How do we utilize that information?
- Proposed solution: combine RVs into logical groups and analyze them as a single unit
  - AKA “Collapsing” or “Burden” tests.

Two Primary Approaches

- **Direct search for susceptibility variants**
  - Assume highly penetrant variant and/or Mendelian disease
  - Extensive reliance on bioinformatics for variant annotation and filtering
  - Sample sizes usually small
- **Rare Variant (RV) "collapsing" methods**
  - Increasingly common in complex disease research
    - May require very large sample sizes!
  - Assume that any of several LOF variants in a susceptibility gene may lead to same disease or trait
  - Many statistical tests available
  - Also relies heavily on bioinformatics



Families of Collapsing Tests

- **Burden Tests**
  - Combine minor alleles across multiple variant sites…
    - Without weighting (**CMC**, CAST, CMAT)
    - With fixed weights based on allele frequency (WSS, RWAS)
    - With data-adaptive weights (Lin/Tang, **KBAC**)
    - With data-adaptive thresholds (Step-Up, VT)
    - With extensions to allow for <u>effects in either direction</u> (Ionita-Laza/Lange, C-alpha)
- **Kernel Tests**
  - Allow for individual variant effects in either direction and permit covariate adjustment based on kernel regression
    - Kwee et al., *AJHG,* 2008
    - SKAT
    - **SKAT-O**

Credit: Schaid et al., *Genet Epi*, 2013

Burden Test Methods in SVS

- CMC: Combined Multivariate and Collapsing test
  - Multivariate test: simultaneous test for association of common and rare variants in gene
  - Testing methods include Hotelling T^2 and Regression
  - Li and Leal, *AJHG*, 2008
- KBAC: Kernel-Based Adaptive Clustering
  - Test models the risk associated with multi-locus genotypes in gene regions
  - Adaptive weighting procedure that gives higher weights to genotypes with higher sample risks
  - Permutation testing, regression or mixed-model significance testing options
  - Liu and Leal, *PLoS Genetics*, 2010

KBAC: Kernel Based Adaptive Clustering

- Test models the risk associated with multi-locus genotypes at a per-gene level
- Adaptive weighting procedure that gives higher weights to genotypes with higher sample risks
- SVS implementation includes option for 1- or 2-tailed test
  - But most powerful when all variants in gene have unidirectional effect
- Permutation testing, regression or mixed-model significance testing options
- Liu and Leal, *PLoS Genetics*, 2010

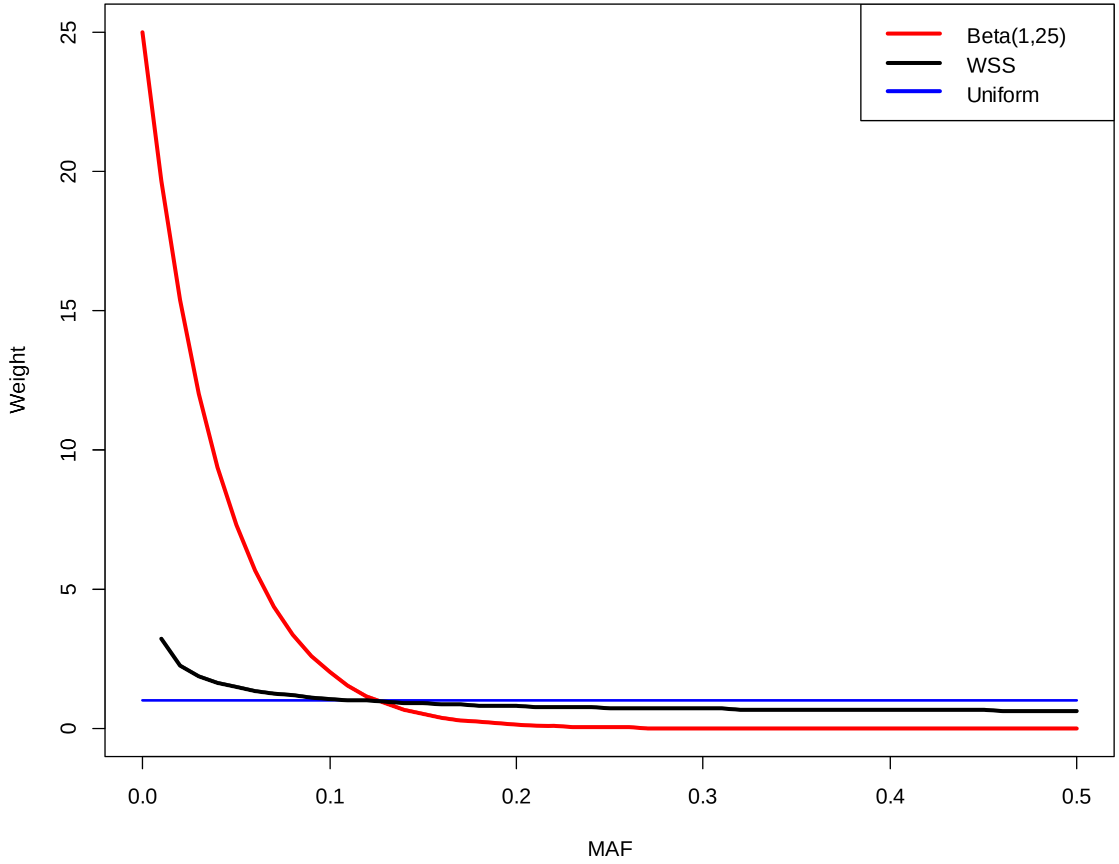
SKAT: Sequence Kernel Association Test

- Utilizes kernel machine methods
- Aggregates test statistics of variants over gene region to compute region level p-values
- Many extensions of the method
- “This method can be more powerful when causal variants have bidirectional effects and/or a large proportion of the variants within gene region are non-causal.”
- “SKAT is less powerful than burden tests when causal variant effects are unidirectional.”
  - Liu and Leal, *PLoS Genetics,* 2012

SKAT-O: Optimized SKAT approach

- Combines a burden test with SKAT test in unified approach
- Burden tests are more powerful when most variants in a region are causal
- SKAT test is more powerful when variants have effects in different directions or some variants have no effect
- SKAT-O optimizes power by adaptively weighting the SKAT and burden results in a combined test

Variant Weighting in SKAT Tests



- **Default weighting scheme is based on Beta(1,25) distribution**
- **Gives much greater weight to the rarest variants**

Bioinformatic Filtering and Annotation

- The genomics community has spent years producing vast resources of data about DNA sequence variants
  - Some data is observational, like variant frequencies from the 1000 genomes project or the NHLBI Exome Sequencing Project
  - Other data is based on predictive algorithms, like PolyPhen or SIFT.
  - Even “simple” annotations, like mapping data for genes, segmental duplications and other sequence features are extremely valuable for analytic workflows.
- These data sources can be used to annotate variants identified in an NGS experiment
  - Annotations may be used for both QC and analysis purposes.
- Once annotated, variants may be filtered, sorted, and prioritized to help us identify disease-causing mutations

Interactive Demonstration

- **SVS 8.3**
- **Exploratory analysis workflow**
  - Simulate the development of a burden test
- **Formal RV association test workflows**
  - SKAT and SKAT-O
  - KBAC and MM-KBAC
  - CMC

NGS Analysis Workflow Development in SVS

- SVS is very flexible in workflow design.
- SVS includes a broad range of tools for data manipulation and variant annotation and visualization that can be used together to guide us on an interactive exploration of the data.
- We begin by defining the final goal and the steps needed to help us reach that goal:
  - Are we looking for a very rare, non-synonymous variant that causes a dominant Mendelian trait?
  - Are we looking for a gene with excess rare variation in cases vs controls?
- Once we know what we are looking for, we can identify the available annotation sources that will help us answer the question.

Data Simulation Process

- **Begin with 2504 samples from 1000 Genomes Phase 3 data.**
- **Define LOF variants as:**
  - MAF<0.01 in NHLBI/ESP 6500 Exomes data
  - MAF<0.01 according to dbSNP
  - Predicted damaging by at least one of 5 prediction methods in dbNSFP
    - (Automatically excludes synonymous, non-coding, and InDel variants)
  - Results in 395k variant sites
- **Create random binary phenotype**
- **Adjust phenotypes based on carrying LOF variants in any of eight genes previously implicated in asthma**
- **Final: 1338 cases, 1166 controls**
- **About 18k genes in tests**

Data Simulation

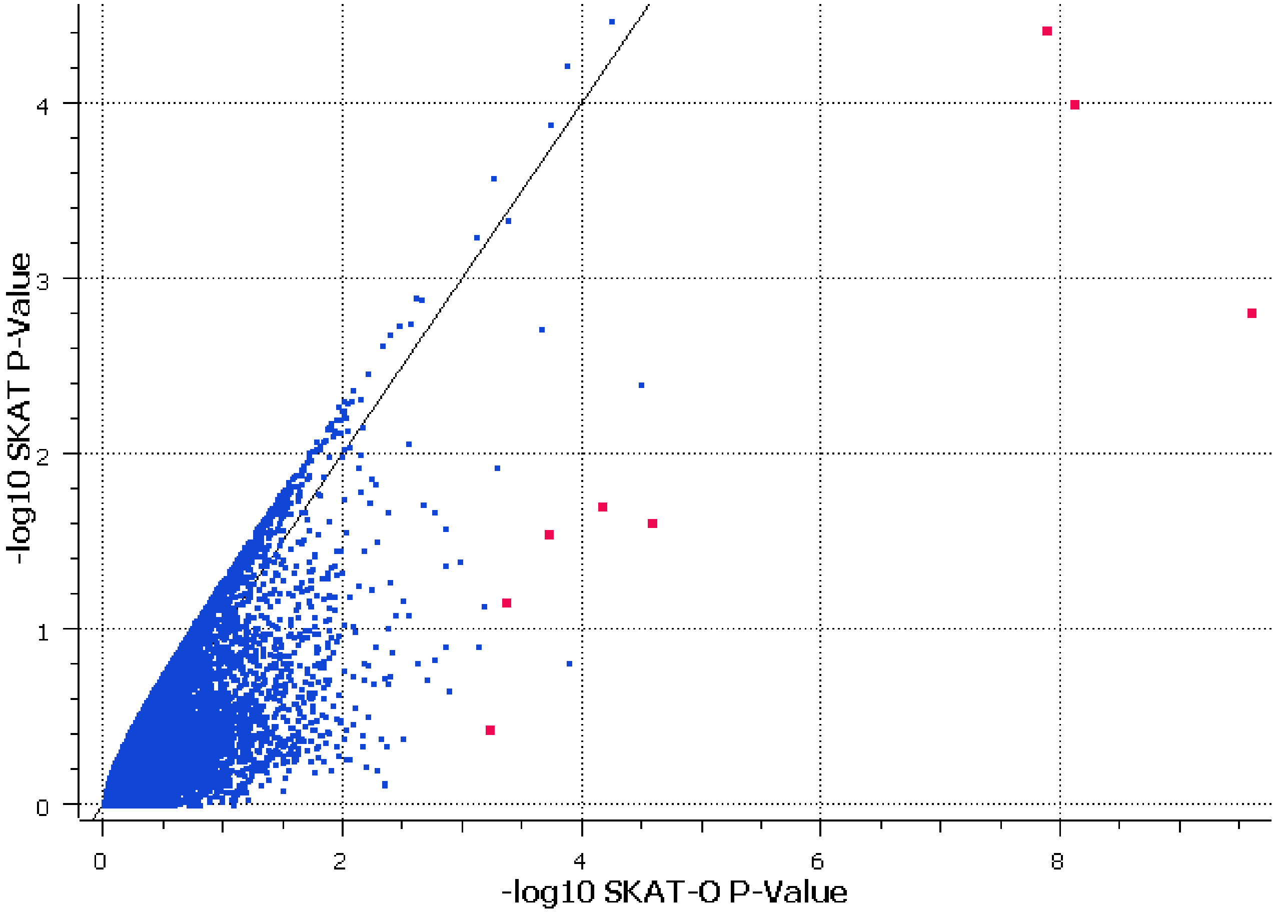
Gene	Chr	Rare NS Variants	LOF Variants	Samples w/ LOF variant	Cases w/ LOF variant	% of Total Cases
IL12B	5	18	11	19	19	1.42
TNF	6	11	9	17	17	1.27
COL26A1	7	23	23	58	53	3.96
TPSG1	16	48	48	80	72	5.38
TPSAB1	16	22	19	80	68	5.08
TPSD1	16	12	11	23	21	1.57
IL4R	16	39	20	30	28	2.09
DHX8	17	20	18	26	23	1.72

Demonstration

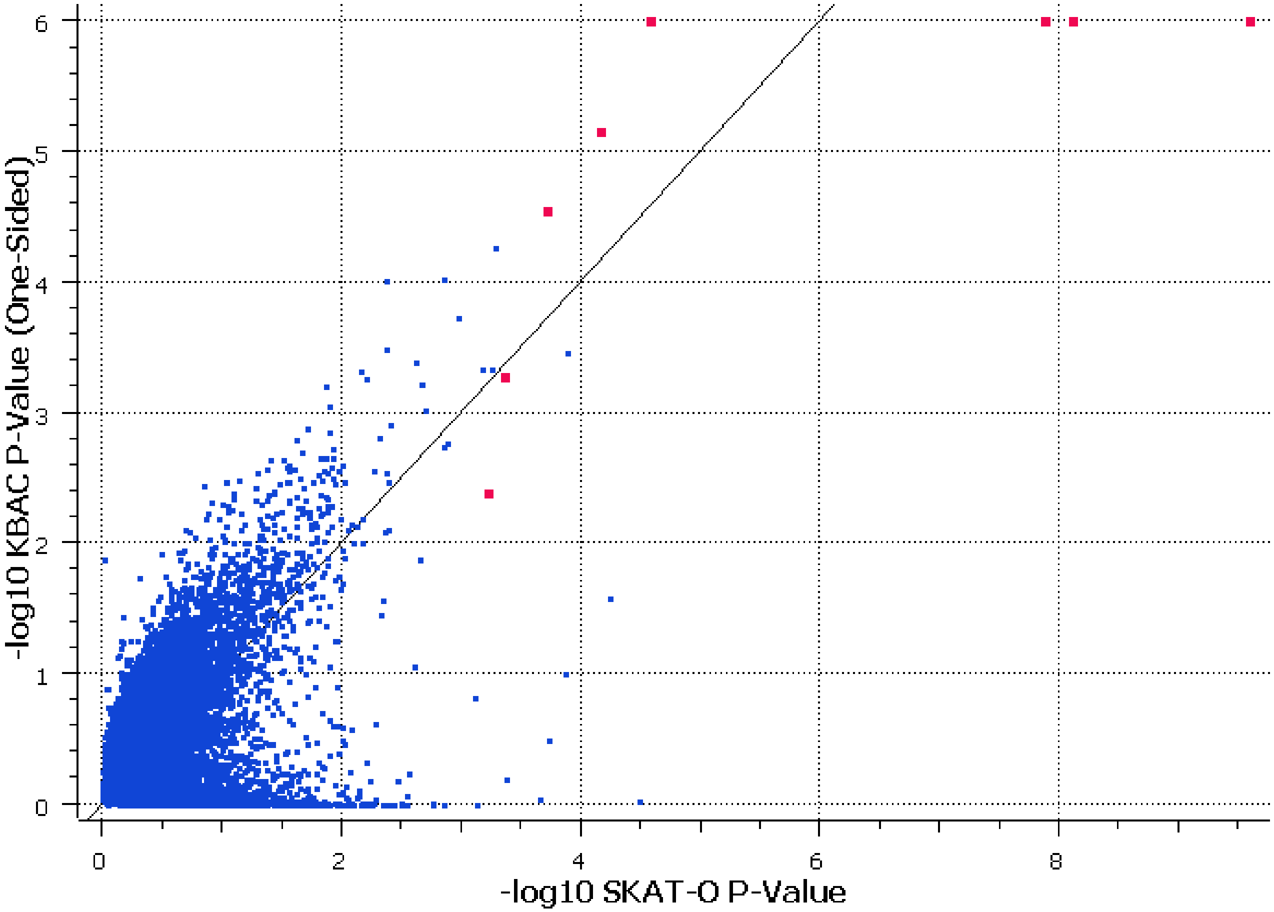
Results in Ideal Conditions (only rare LOF variants)

Gene	Samples w/ variant	Cases w/ variant	KBAC p (1M sims)	SKAT p	SKAT-O p
IL12B	19	19	7e-6	0.02	6.8e-5
TNF	17	17	2.8e-5	0.03	2.0e-4
COL26A1	58	53	1e-6	9.9e-5	7.8e-9
TPSG1	80	72	1e-6	0.001	2.6e-10
TPSAB1	80	68	1e-6	3.7e-5	1.3e-8
TPSD1	23	21	5.3e-4	0.07	4.3e-4
IL4R	30	28	1e-6	0.25	2.6e-5
DHX8	26	23	0.004	0.37	6.1e-4

SKAT vs SKAT-O



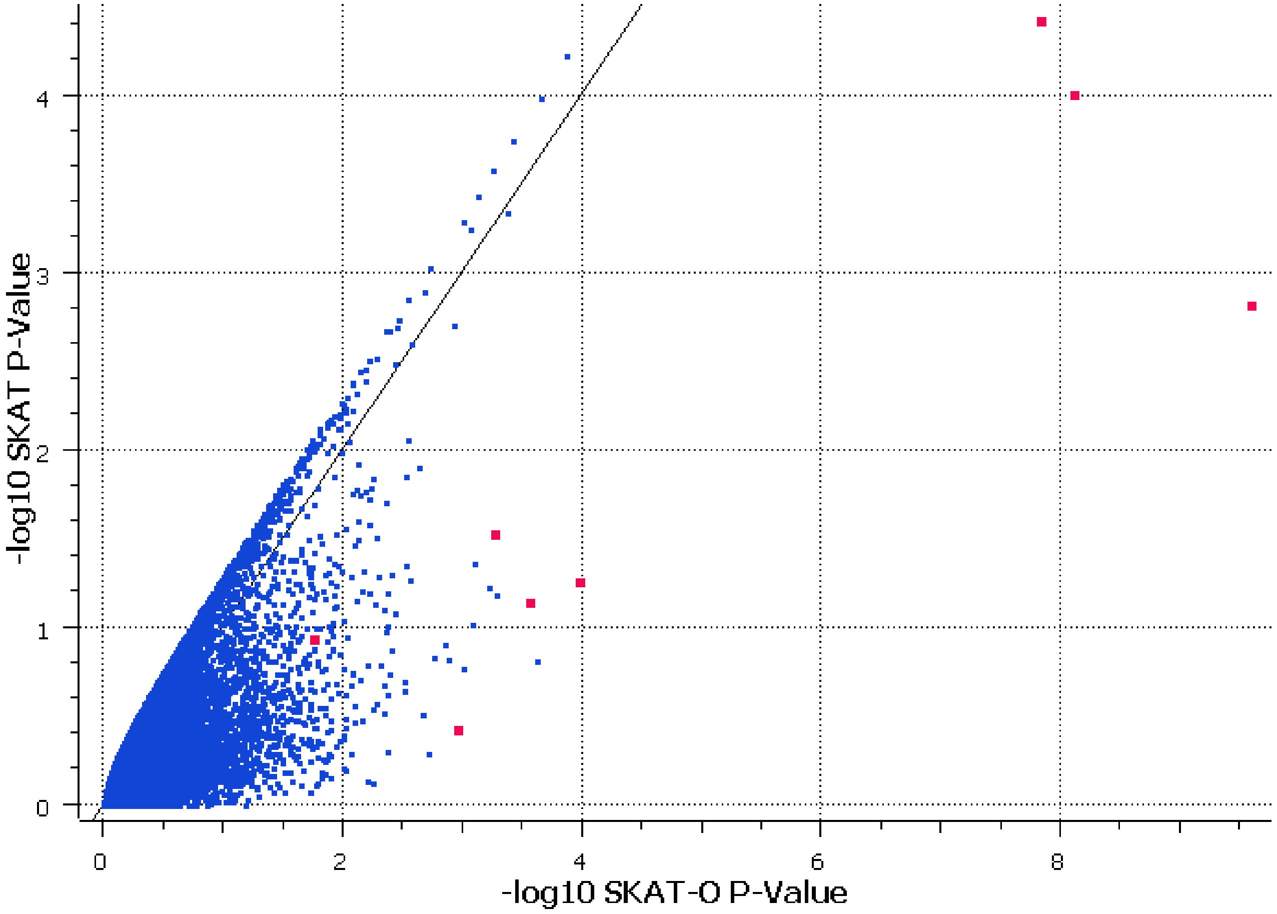
KBAC vs SKAT-O



Results in Noisy Conditions (All rare NS variants)

Gene	Rare NS Variants	LOF Variants	KBAC p (1M sims)	SKAT p	SKAT-O p
IL12B	18	11	2.6e-4	0.06	1.0e-4
TNF	11	9	4.5e-5	0.03	5.4e-4
COL26A1	23	23	1e-6	9.9e-5	7.8e-9
TPSG1	48	48	1e-6	0.002	2.6e-10
TPSAB1	22	19	1e-6	3.8e-5	1.5e-8
TPSD1	12	11	4.1e-4	0.07	2.8e-5
IL4R	39	20	0.0037	0.12	0.017
DHX8	20	18	0.0038	0.43	0.001

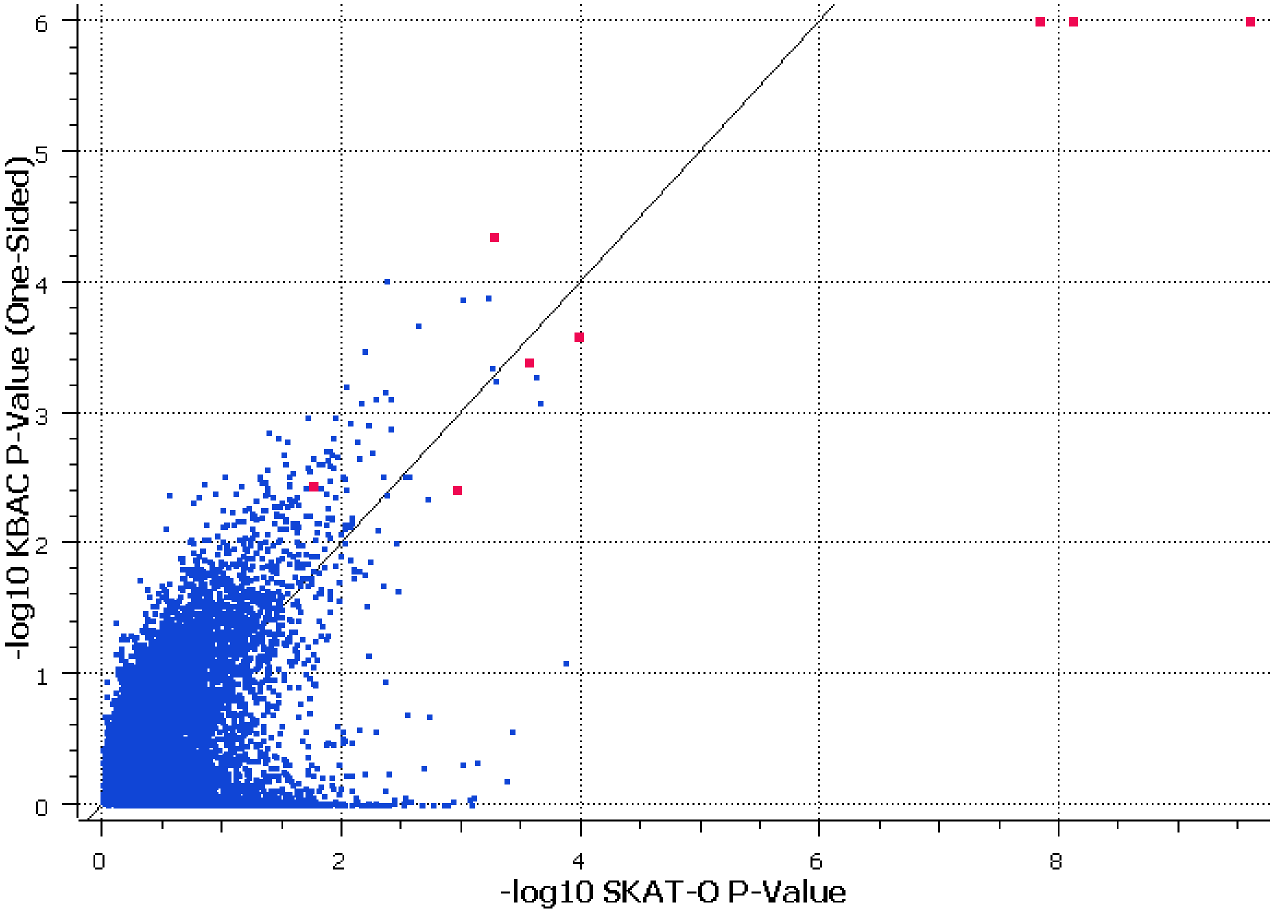
SKAT vs SKAT-O



Golden Helix

Accelerating the Quest for Significance™

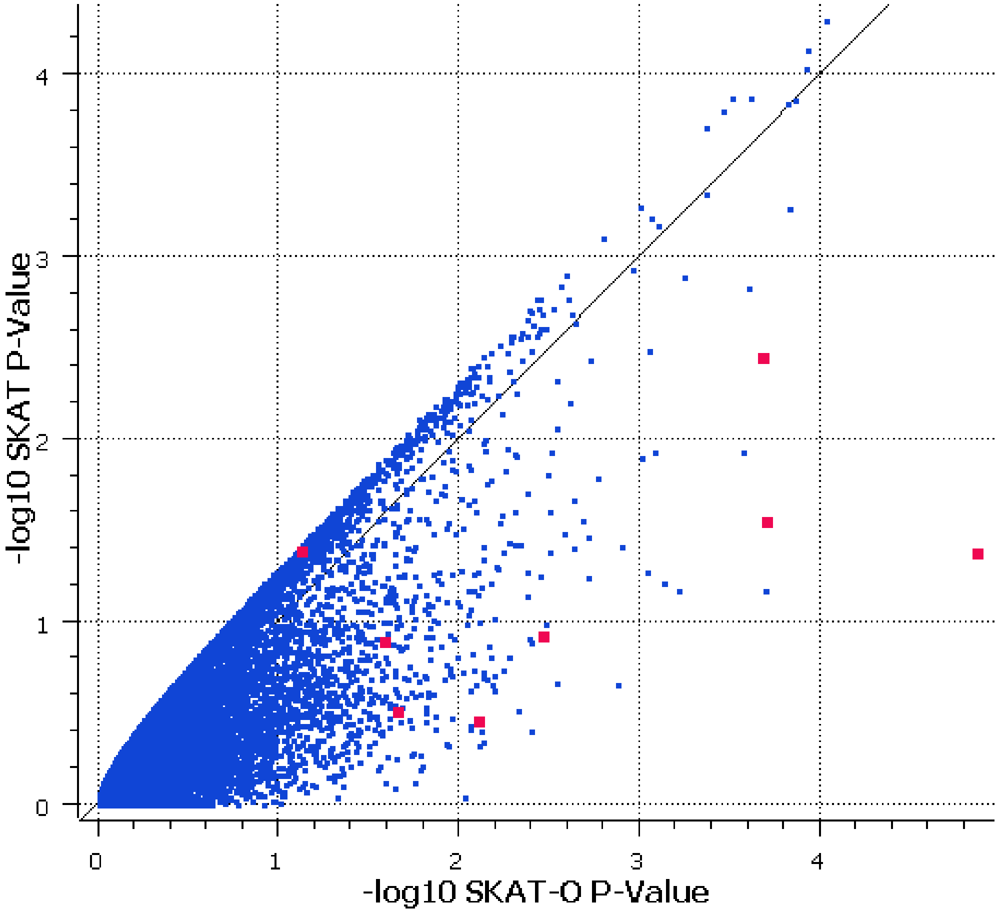
KBAC vs SKAT-O



Results in Noisy Conditions (All Functional Variants)

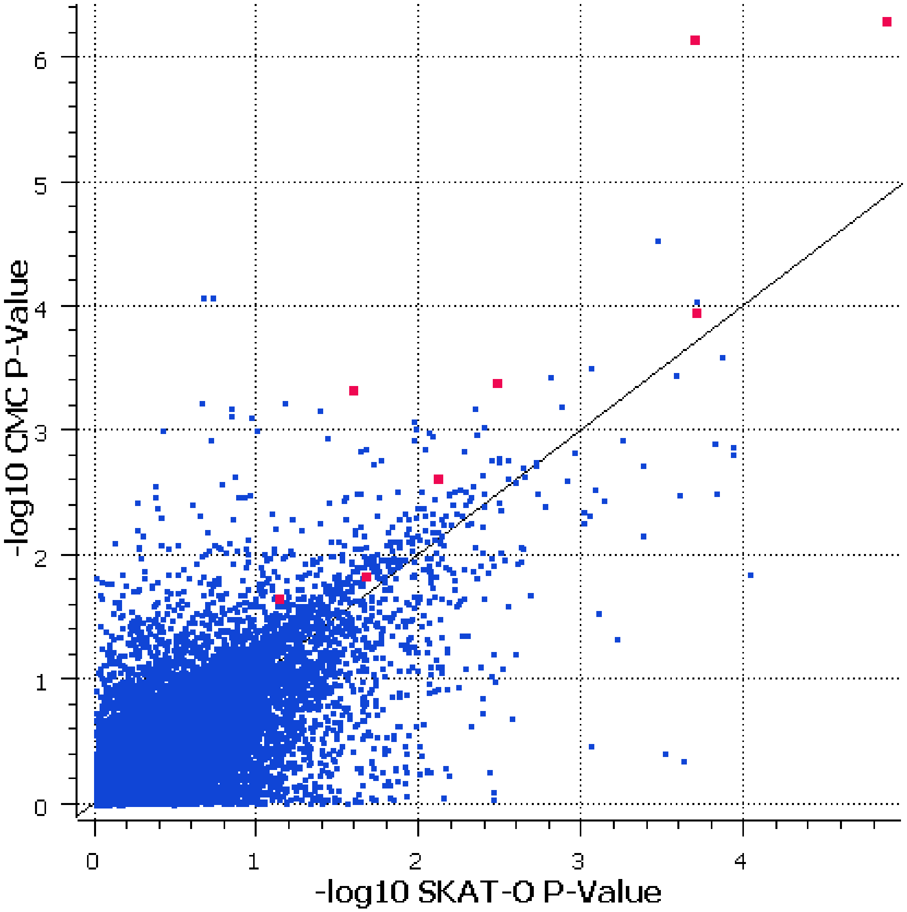
Gene	Functional Variants	LOF Variants	CMC p	SKAT p	SKAT-O p
IL12B	14	11	4.6e-4	0.13	0.026
TNF	9	9	1.1e-4	0.029	2.0e-4
COL26A1	28	23	5.1e-7	0.043	1.3e-5
TPSG1	67	48	0.002	0.35	0.0078
TPSAB1	22	19	7.2e-7	0.004	2.0e-4
TPSD1	21	11	4.1e-4	0.12	0.0033
IL4R	28	20	0.022	0.04	0.075
DHX8	20	18	0.014	0.31	0.021

SKAT vs SKAT-O

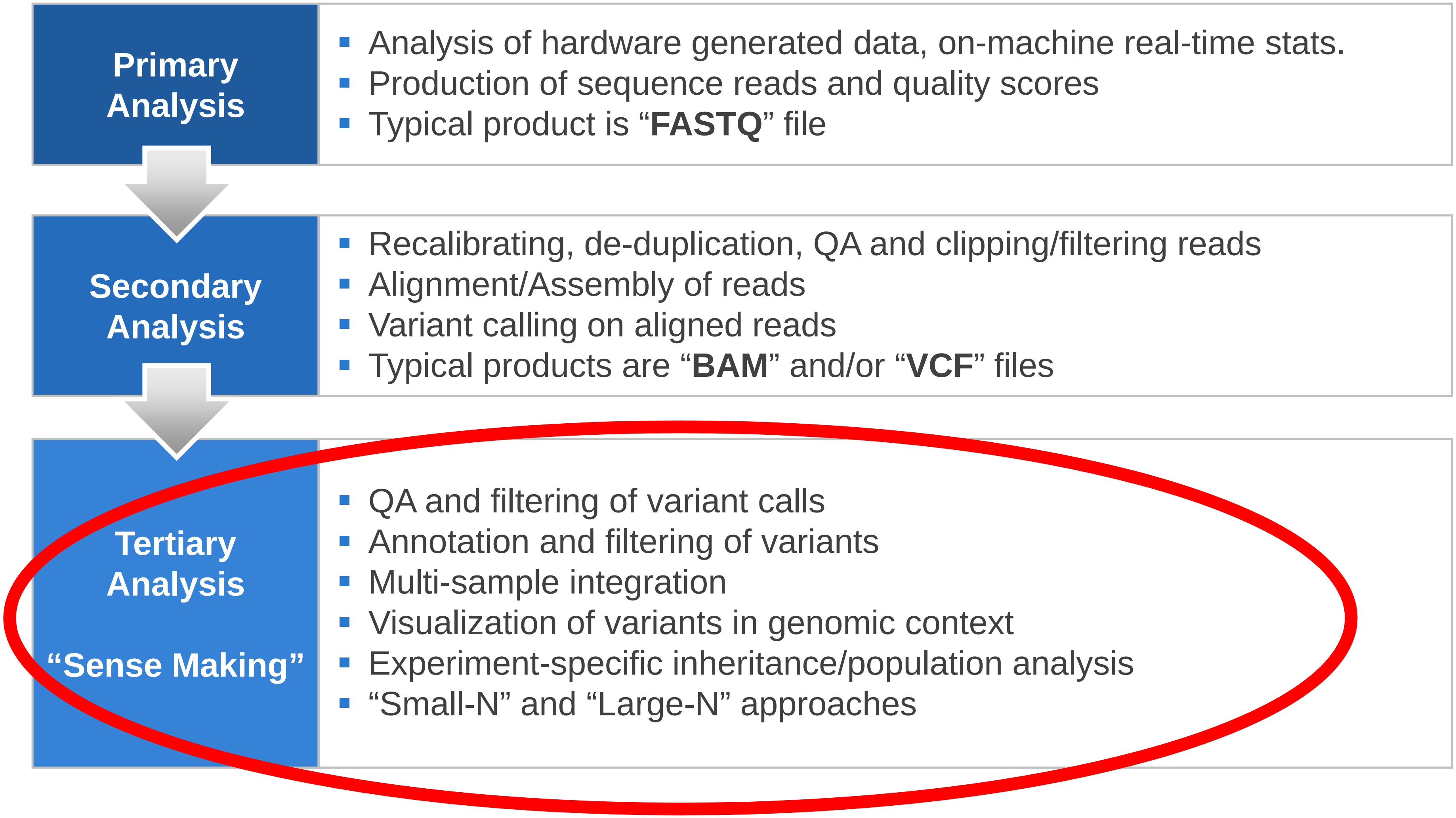


Golden Helix
Accelerating the Quest for Significance™

CMC vs SKAT-O



NGS Analysis



Primary Analysis

- Analysis of hardware generated data, on-machine real-time stats.
- Production of sequence reads and quality scores
- Typical product is “**FASTQ**” file

Secondary Analysis

- Recalibrating, de-duplication, QA and clipping/filtering reads
- Alignment/Assembly of reads
- Variant calling on aligned reads
- Typical products are “**BAM**” and/or “**VCF**” files

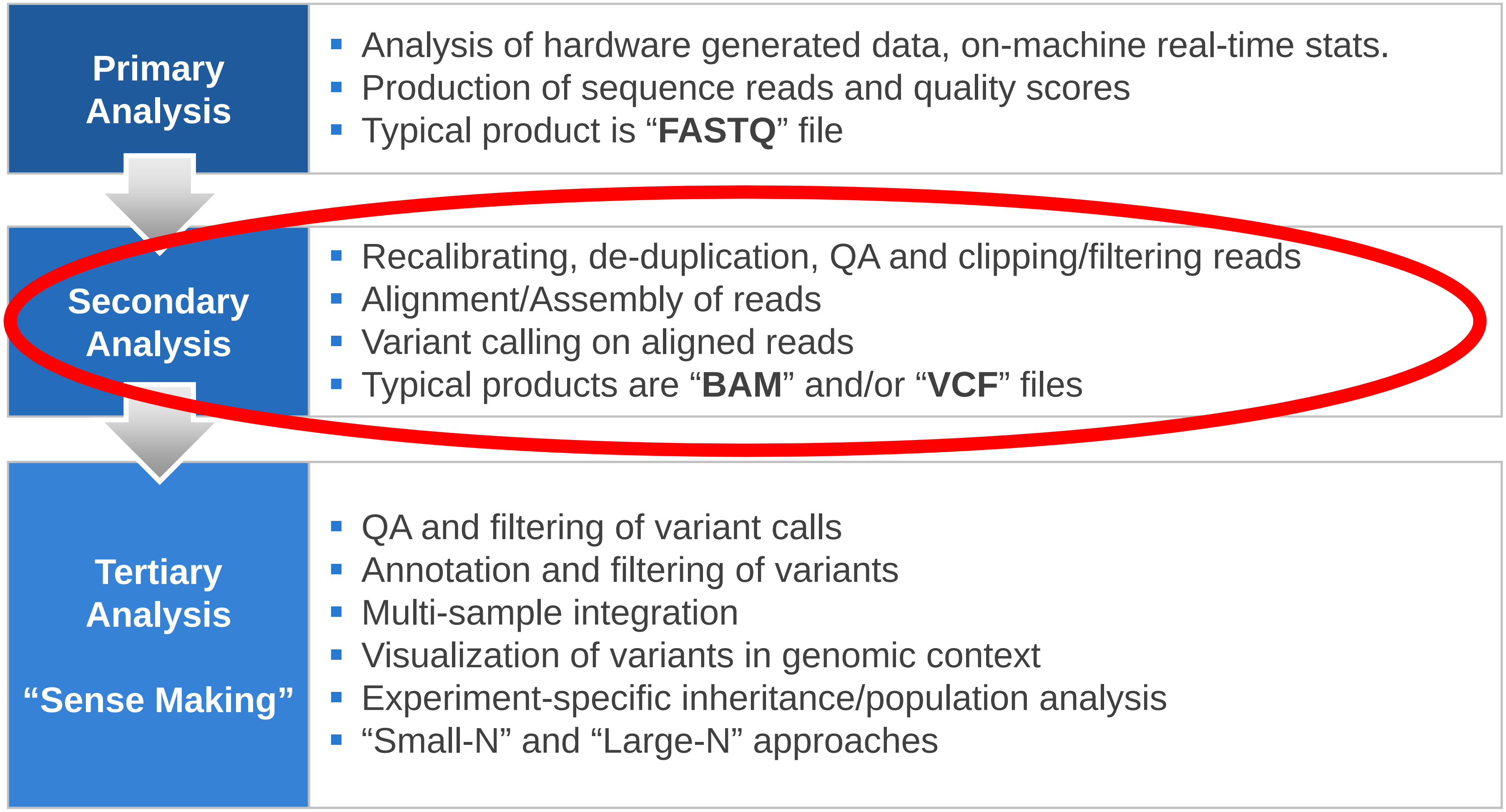
Tertiary Analysis

“Sense Making”

- QA and filtering of variant calls
- Annotation and filtering of variants
- Multi-sample integration
- Visualization of variants in genomic context
- Experiment-specific inheritance/population analysis
- “Small-N” and “Large-N” approaches

GOLDEN HELIX
Accelerating the Quest for Significance™

NGS Analysis



Primary Analysis

- Analysis of hardware generated data, on-machine real-time stats.
- Production of sequence reads and quality scores
- Typical product is “**FASTQ**” file

Secondary Analysis

- Recalibrating, de-duplication, QA and clipping/filtering reads
- Alignment/Assembly of reads
- Variant calling on aligned reads
- Typical products are “**BAM**” and/or “**VCF**” files

Tertiary Analysis

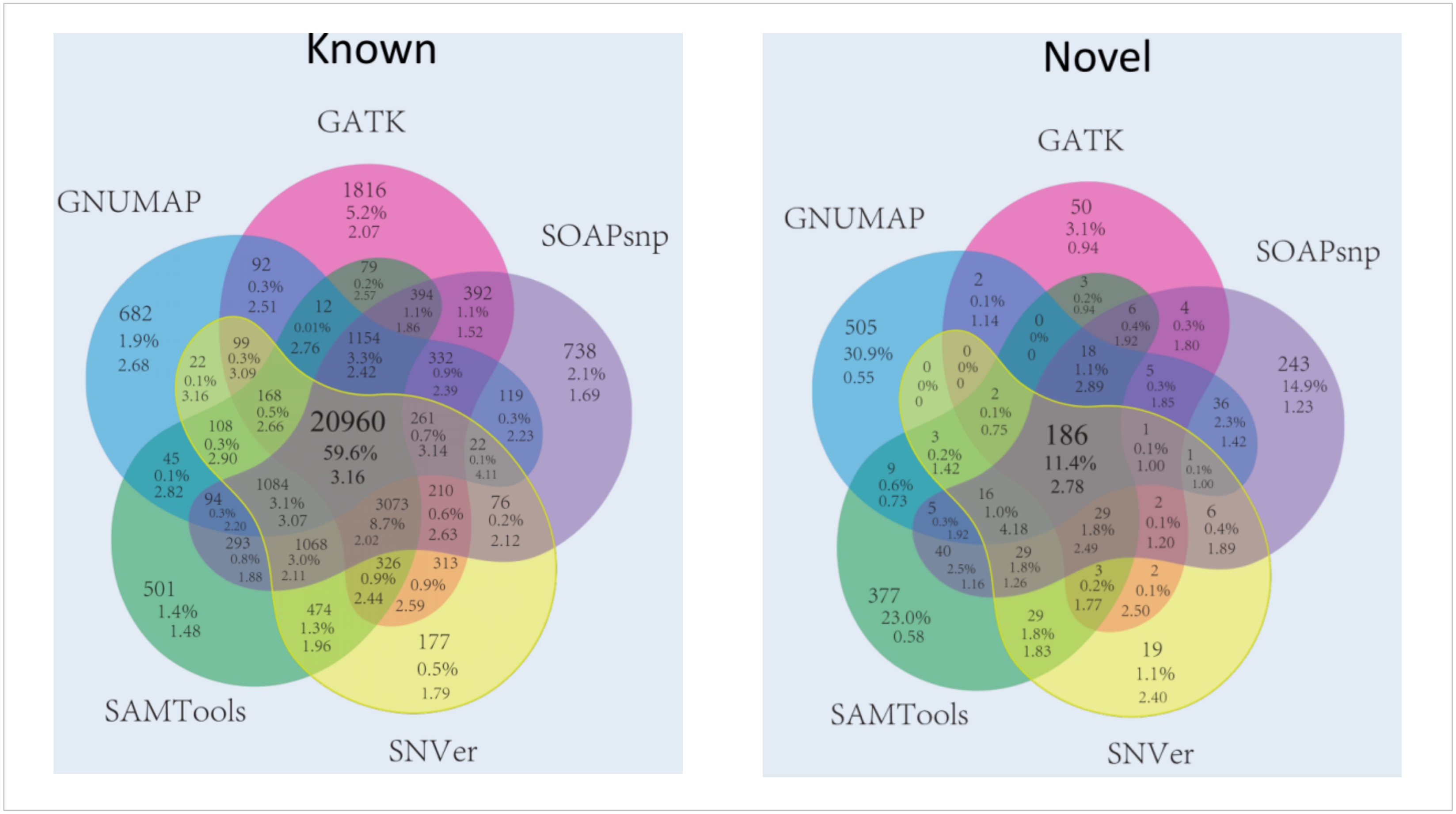
“Sense Making”

- QA and filtering of variant calls
- Annotation and filtering of variants
- Multi-sample integration
- Visualization of variants in genomic context
- Experiment-specific inheritance/population analysis
- “Small-N” and “Large-N” approaches

Secondary Analysis and QC Considerations

- **What did we do in GWAS?**
  - Call rate
  - HWE
  - MAF
  - But those aren't really applicable for RV analysis…
- **What do we use for NGS?**
  - Coverage depth
  - Quality scores per variant and per genotype call
  - Singleton counts
  - Ts/Tv ratios
  - Mappability of the region

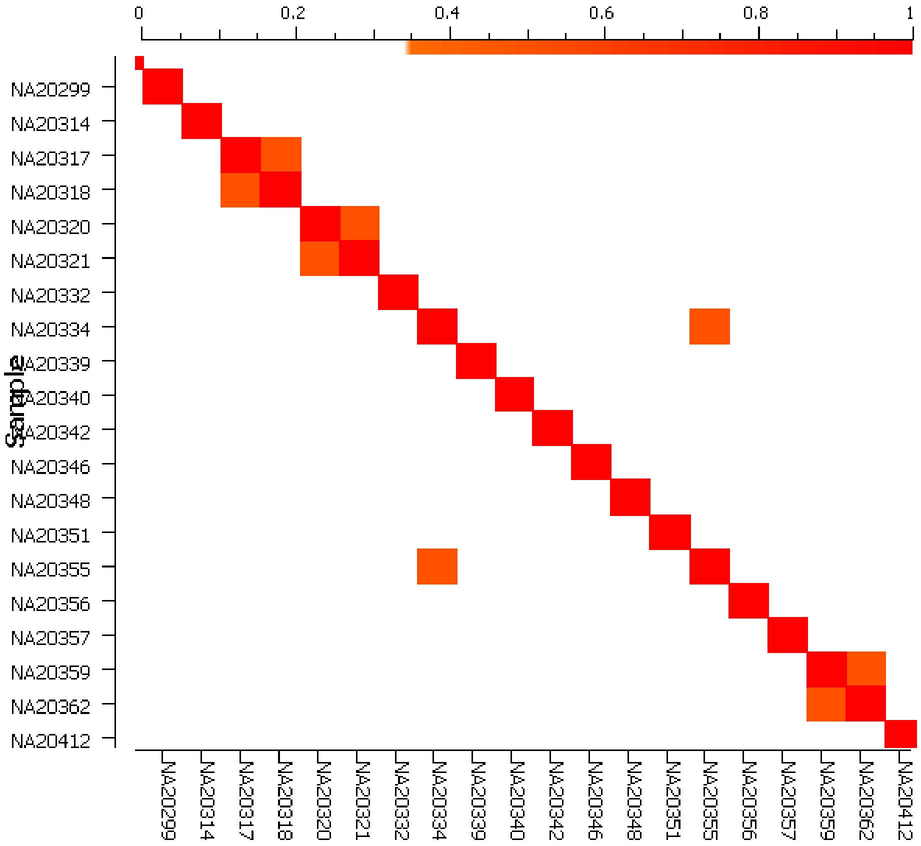
Most Importantly: Be Consistent!



Gholson Lyon, 2012

What about common variants?

- **Yes, you can run GWAS-style common variant analysis procedures with sequence data**
- **Helpful to have homozygous reference genotype calls in the data**
- **Certain procedures may require careful consideration in terms of variant selection**
  - PCA
  - IBD analysis
  - Mixed model regression



Marker Selection Process

Affymetrix SNP6

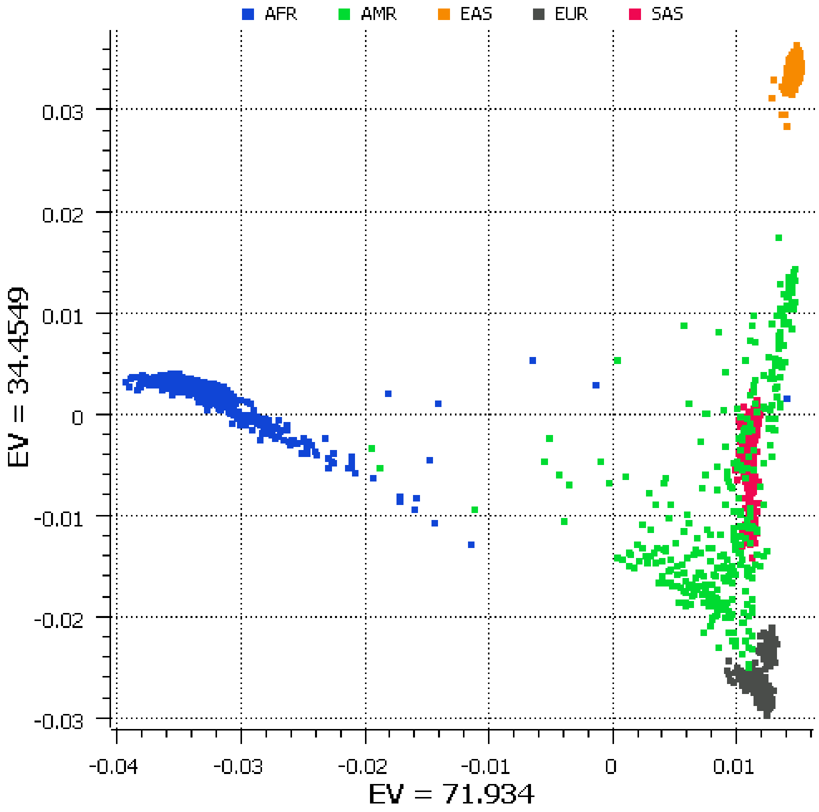
Full content:	906k
Autosomes:	867k
MAF>0.01, CR>0.99:	806k
LD Pruned:	74k

1kG Phase 3

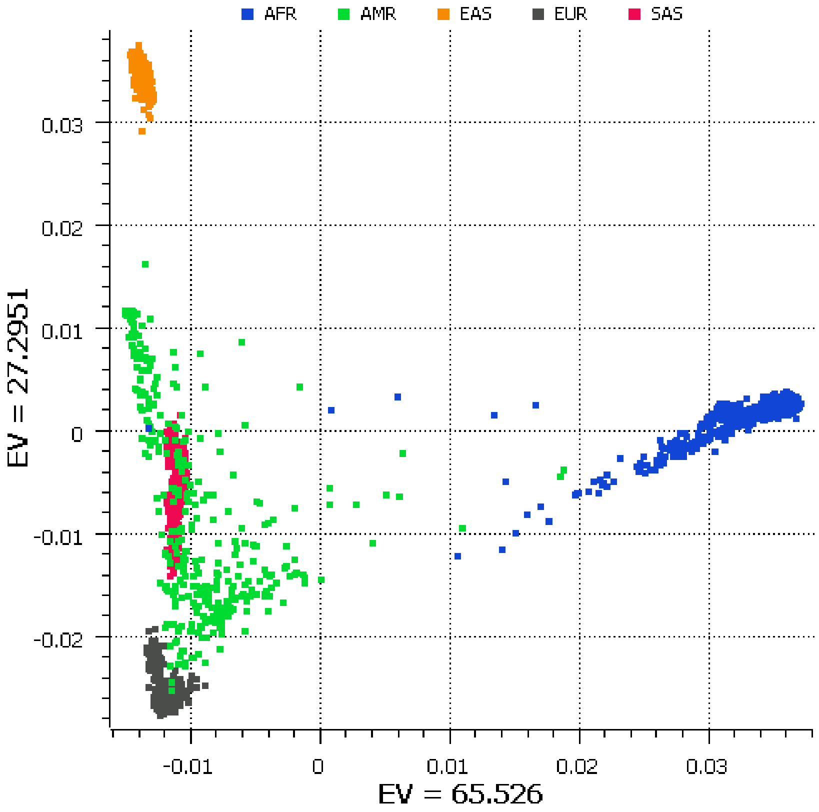
Whole genome:	81M
Exons +/- 5bp:	2.2M
MAF>0.01:	263k
LD-Pruned:	70k

PCA Comparison

Affymetrix SNP6

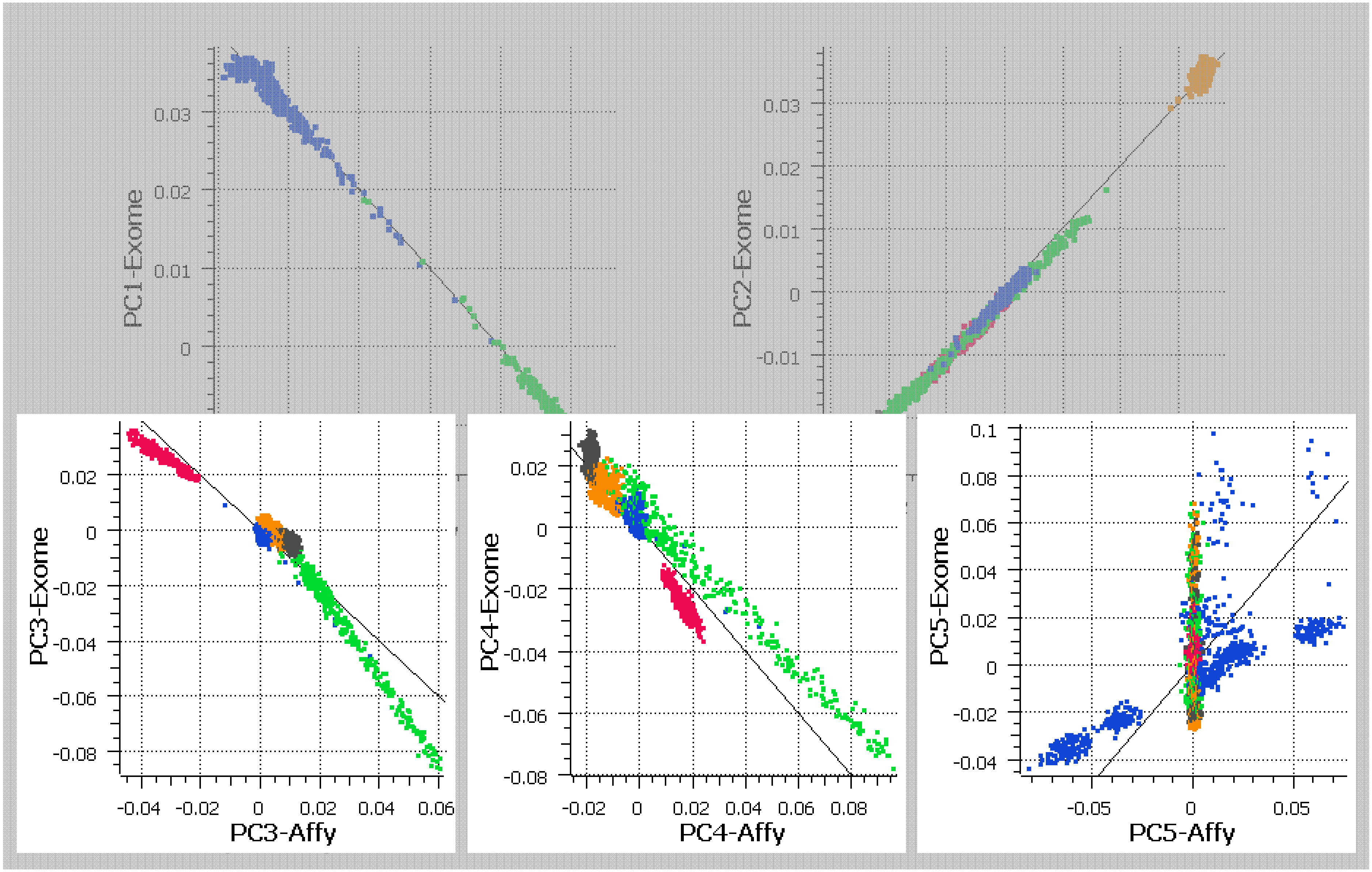


1kG Phase 3



GOLDEN HELIX
Accelerating the Quest for Significance™

Measuring the same thing?



Conclusions

- **SVS is a comprehensive platform for analysis of common and rare sequence variants**
- **SKAT-O is an optimized combination of SKAT and burden test approaches**
- **The power of collapsing tests is affected by many factors:**
  - Variant weighting schemes
  - Variant filtering/selection process
  - Causal vs. “passenger” variants
  - Sample size
  - The true underlying biology of the disease

Questions or more info:

- info@goldenhelix.com
- Request a copy of SVS at www.goldenhelix.com
- Download GenomeBrowse for free at www.GenomeBrowse.com

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

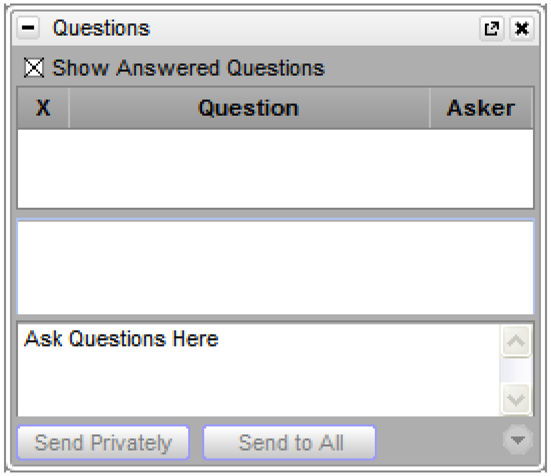
Annotation, filtering, and interpretation of variants with easy to use workflows.

Explore your NGS alignment and variant data. For FREE.

GOLDEN HELIX
Accelerating the Quest for Significance™

Any Questions?

Use the Questions pane in your GoToWebinar window



GOLDEN HELIX
Accelerating the Quest for Significance™