

HELIXTREE

HelixTree is the core module of the SNP & Variation Suite. Its unique set of conventional tools empower you to quickly and easily perform a broad array of workflows for genetic association studies.

Data Management

The core architecture of HelixTree has been completely reinvented to efficiently handle datasets of virtually any size and type on a desktop computer. Supported dependent variables include case-control and quantitative traits. Predictors can be binary, continuous, ordinal, categorical, nominal, and genetic (bi- and multi-allelic SNPs, log ratios, CNVs, microsatellites, etc.). Further, direct support for most common file formats streamlines data import, ensuring you spend most of your time on the more important aspects of analysis. Real-time spreadsheet manipulation, data editing, and enrichment help eliminate the hassles of working with large-scale, complex data.

Genetic Association Testing

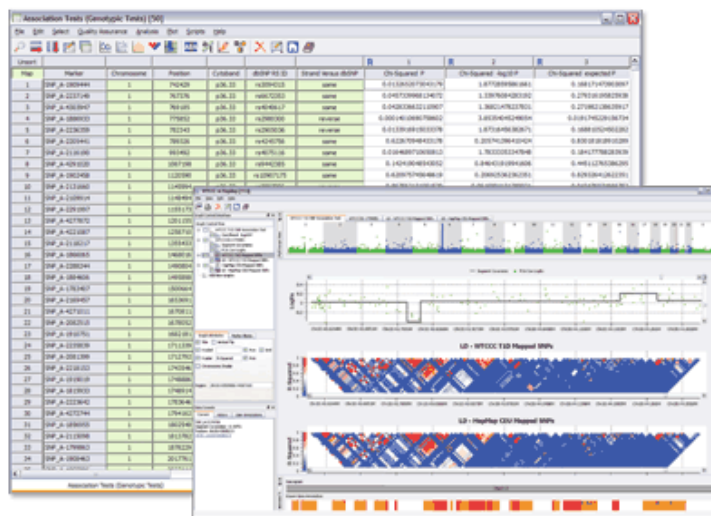
Find more associations with the most extensive collection of genetic association tests including allele, genotype, haplotype, CNV, ROH, LD, and advanced regression-based testing. Within a streamlined interface you can test against either cases vs. controls or quantitative traits using a variety of statistical measures under any one of several genetic model assumptions. Tests can be run individually or simultaneously while also correcting for stratification and applying a number of multiple testing corrections, including permutation testing.

LD and Haplotype Analysis

Interactively explore LD and haplotype analysis in an innovative and powerful new interface. You can view LD plots from one or more populations and explore them side-by-side with association results. For haplotype analysis it is easy to define and modify haplotype blocks from an LD plot or spreadsheet, compute haplotype and diplotype frequency tables, and perform a number of haplotype association tests, including per-block and per-haplotype methods.

Quality Assurance

High quality data is key to quality results. Considerable effort has been made to enhance quality assurance at every step. You can now easily generate a number of genotype statistics, view cluster plots of allele intensities, check gender and marker concordance, perform variance analysis on log ratios, filter poor



quality markers and samples, and more. HelixTree also offers a powerful principal component analysis (PCA) approach for both SNP and CNV data to both detect and correct for batch effects and population stratification.

Richly Interactive Data Visualization

View results from a whole new perspective with richly interactive visualization and unprecedented whole-genome navigation. Apply data transformations in real-time, or customize plots for more enhanced presentation. You can combine multiple data series, such as SNP, CNV, and haplotype results or compare graphs side-by-side as with two or more LD plots from differing populations. After you attain the desired view, you can easily save it to a number of publication quality formats, including scalable vector graphics.

Fully Programmatic Scripting and Automation

Automate workflows, incorporate custom methods, or interoperate with other programs. These are just a few examples of how you can enhance the utility of HelixTree and other SVS 7 modules with a fully programmatic Python scripting interface. New in SVS 7 is an integrated Python editor that makes it easy to read and write scripts helping even novice users realize the power of scripting.

COPY NUMBER ANALYSIS MODULE (CNAM)

CNAM, in conjunction with other SVS 7 modules, offers a complete set of tools for processing raw intensity data, identifying regions of copy number variation, visualizing copy number data, and performing association analyses on a variety of CNV covariates.

CNV Data Processing

CNAM offers direct import of log ratio data from a number of providers, including Affymetrix, Agilent, and Illumina. For Affymetrix CEL files (500K, 5.0, and 6.0) a powerful processing tool enables you to run quantile normalization on the A and B probe intensities, including virtual array generation to merge CN and SNP probes or multiple arrays (e.g. NSP and STY). This process scales to thousands of samples, and can use any sample set as a reference.

CNV Quality Assurance

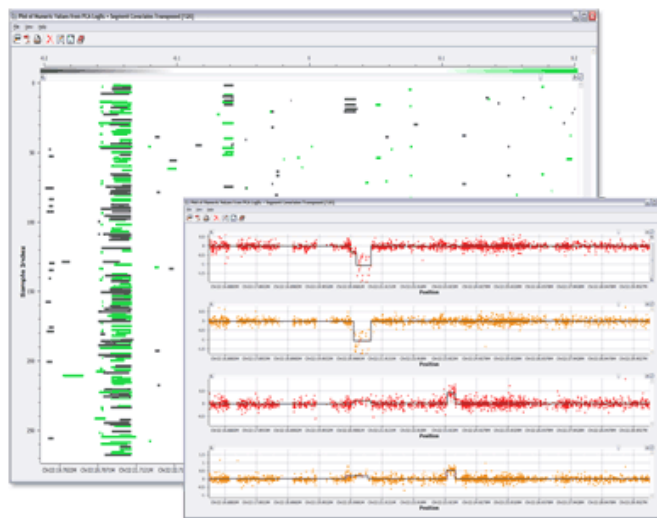
For both microarray and aCGH data, significant bias can be introduced by batch effects (plate, machine, and site variation), genomics waves, and population stratification. Other sources of variation include sample extraction and preparation procedures, cell types, temperature fluctuation, and even ambient ozone levels in a lab. These can lead to complications ranging from poorly defined segments to false and non-replicable findings. Utilizing a powerful PCA approach enables you to simultaneously correct for all experimental artifacts, while significantly improving signal-to-noise ratios.

CNV Association Testing

A number of covariate generation procedures enable you to perform association testing on raw or PCA-corrected log ratios, CNV segment means, and discretized values based on three- and two-state models representing loss, neutral, and gain. Perform numeric association tests or advanced linear and logistic regression with CNV covariates alone or in combination with other genetic markers and phenotypic variables.

CNV Data Visualization

"Seeing is believing" with richly interactive data visualization that provides unprecedented whole genome views and easy navigation of your data. Visually detect CNVs across many samples or confirm optimal segmenting results. Generate cluster plots of allele intensities to filter poor quality markers. Visualize CNV association p-values alongside SNP p-values.



And when you finalize the views you want, you can save them to a number of publication quality formats, including scalable vector graphics.

CNAM Optimal Segmenting

CNAM employs a powerful optimal segmenting algorithm using dynamic programming to detect inherited and de novo CNVs on a per-sample (univariate) and multi-sample (multivariate) basis. Unlike Hidden Markov Models, which assume the means of different copy number states are consistent, optimal segmenting properly delineates CNV boundaries in the presence of mosaicism, even at a single probe level, and with controllable sensitivity and false discovery rate.

CNAM Optimal Segmenting now incorporates a new parallelized, unbiased randomization permutation procedure that uses all available cores on your computer. The new permutation procedure replaces a naïve, potentially biased randomization procedure with the unbiased Fisher and Yates method (also known as the Knuth shuffle). An added option allows you to further refine your segments by efficiently removing univariate outliers during the segmentation process.

PBAT

Developed in collaboration with Dr. Christophe Lange of Harvard's School of Public Health, Golden Helix PBAT delivers an exclusive and extensive array of advanced statistical routines for the design and analysis of family-based SNP and CNV studies.

Pre-Study Power Calculations

The PBAT capabilities for power calculations are a software implementation of the approaches to analytical power calculations for FBATs by [Lange 2002a, Lange 2002b]. They enable you to assess the power of both family and non family-based association tests for many different study designs

Family-Based Quality Control

The latest version of PBAT incorporates a novel test that assesses the genotyping quality of individual probands in family-based association studies. Published in PLoS Genetics [Fardo, 2009] these tests are "ideally suited as the final layer of quality control filters in the cleaning process of genome-wide association studies." You can also assess Mendelian errors, Hardy-Weinberg Equilibrium and Call Rates per Marker.

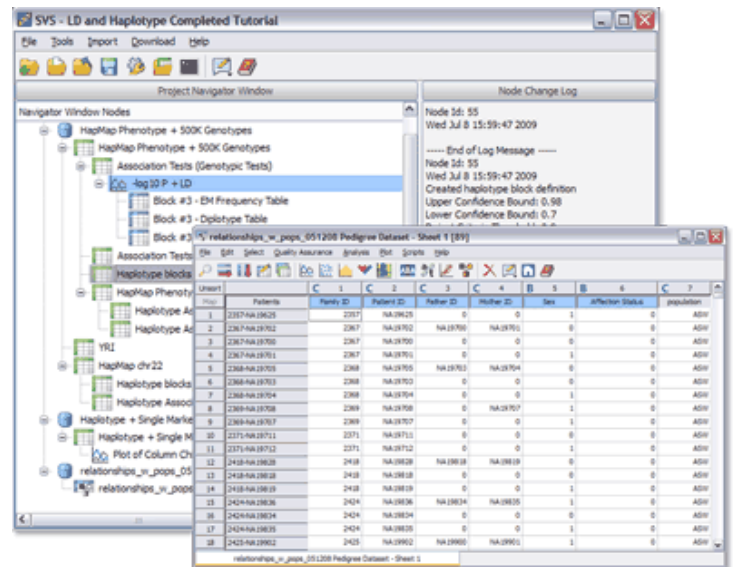
Family-Based SNP and CNV Analysis

Golden Helix PBAT offers a unified approach to the FBAT statistic, a generalization of the transmission disequilibrium test (TDT), to cover different genetic models, tests of different sampling designs, tests involving different disease phenotypes, tests with missing parents and tests of different null hypotheses, all in the same framework. PBAT also supports the testing for copy number variation (CNV) in a family-based setting. All robustness properties of the FBAT approach are maintained as in PBAT for SNP analysis. In addition, all previously-developed FBAT extensions, including FBATs for time-to-onset, multivariate FBATs, and FBAT-testing strategies, can be directly applied to the analysis of CNVs.

Screening Based on Conditional Mean Model

The key concept of PBAT's screening technique is the conditional mean model approach [Lange 2002a, Lange 2002b], for which the data space is considered to be partitioned into two independent testing sets. This approach may be described as follows:

-First, find which combination of phenotypes as a group and markers have the highest power when tested against, not actual genotypes, but those predicted from the parents' genotypes.



-Second, perform the appropriate FBAT test for the selected combinations of phenotypes and markers on the actual genotypes of the patients, both as a group and individually. This allows one to control the type I error rates and to overcome one of the most important statistical hurdles when analyzing genome-wide association studies - the multiple comparison problem. PBAT's screening methods are only minimally affected by the non-causal SNPs. In addition, they are robust against effects of population stratification and admixture.

Rapid Extended Pedigree Algorithm

Golden Helix PBAT includes a new option when doing family-based analysis to use an alternative rapid extended pedigree algorithm that can speed up analysis significantly. It can be applied to SNP, haplotype and copy number analyses.

^{1a} Lange, C. et al. *Power and Design Considerations for a General Class of Family-Based Association Tests: Quantitative Traits*. American Journal of Human Genetics. December 2002.

^{1b} Lange, C. et al. *On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations*. Genetic Epidemiology. September 2002.

WHOLE GENOME ANALYSIS MODULE

The Whole Genome Analysis Module incorporates several technologies and methods designed to overcome the statistical and computational challenges of large-scale whole genome analysis.

Sparse Data Storage Technology

The WGA Module internally compresses SNP data into proprietary sparse storage formats that use a fraction of both system memory and disk space compared to standard genotype file formats. These formats also improve the speed and efficiency of both data import and analysis.

Genome Browser

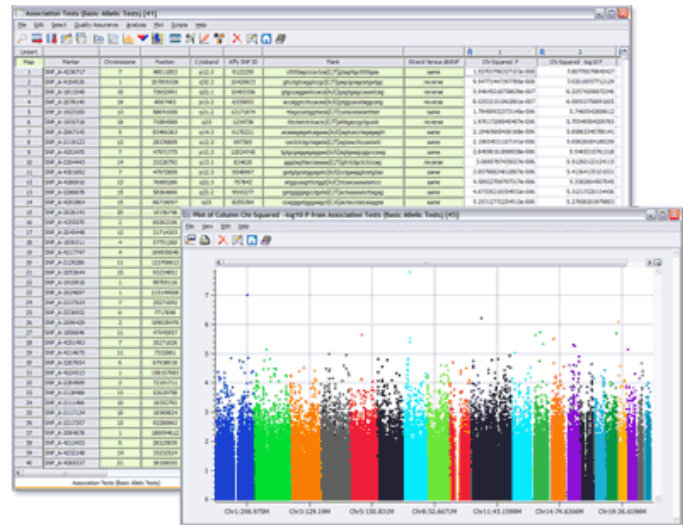
If a genetic marker map is applied to your spreadsheet, you can view results in chromosome and position order with integrated karyogram view and gene annotation tracks. The genome browser immediately puts your data in genomic context and enables you to link to online databases for further investigation of a region, gene, or marker.

Multi-Processor Support

Certain analyses in genome-wide association studies are computationally intensive and can take many hours or, in some cases, days to run (such as CNAM optimal segmenting). Adding the WGA module enables you to utilize all available processors and cores on your machine, which can significantly speed up supported analyses.

Runs of Homozygosity (ROH) Analysis

The recent development of microarray platforms, capable of genotyping hundreds of thousands of single nucleotide polymorphisms (SNPs), has provided an opportunity to rapidly identify novel susceptibility genes for complex phenotypes. Studies employing genotyping microarrays have typically utilized a whole genome association approach, in which each SNP is examined individually for association with disease. While this approach has resulted in several important breakthroughs in the past few years, it is biased towards detecting common alleles with additive effects. At the same time, structural properties of WGA datasets, including patterns of linkage disequilibrium (LD), have not yet been exploited in



these analyses.

Consequently, Dr. Todd Lencz, Associate Director of Research at The Zucker Hillside Hospital, working in collaboration with Dr. Christophe Lambert of Golden Helix, has developed a novel analytic approach that first identifies patterned clusters of SNPs demonstrating extended homozygosity (runs of homozygosity or “ROHs”) and then employs both genome-wide and regionally-specific statistical tests for association to disease. This approach can identify chromosomal segments that may harbor rare, penetrant recessive loci.

In a recent study published in the Proceedings of the National Academy of Sciences¹, Dr. Lencz outlines the runs of homozygosity association methodology and how it was used to identify nine genetic risk loci for schizophrenia.

SVS 7 includes all the functionality in the paper in an easy to use interface. You can plot ROH results, generate different ROH covariates, and run ROH association tests.

¹ Lencz, T. et al. *Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia*. Proceedings of the National Academy of Sciences. October 22, 2007.

