

# ADVANCED GENOTYPE CALLING AND IMPUTATION

## OVERVIEW

With our steadfast goal of helping genetic researchers continuously enhance the quality of their association findings, Golden Helix recently initiated the first state-of-the-art genotype calling and imputation solution employing the renowned BEAGLE and BEAGLECALL software programs within a comprehensive service offering.

Designed for maximum accuracy and quick turnaround for genetic researchers who have limited budgets and time, our advanced genotyping and imputation services will increase the power of a study, improve association results, and reduce the scope of follow-up fine-mapping efforts. This is done by increasing sample size and marker density by an order of magnitude, while significantly reducing the number of SNPs discarded due to poor quality.

With our service, optimal results can be achieved within a matter of days and at a fraction of the cost of acquiring additional samples, regenotyping on higher density platforms, or building the high-performance computing environment necessary to support advanced genotype calling and imputation.

## ADVANCED GENOTYPE CALLING

### Benefits:

1. Improves genotype accuracy
2. Improves genotype call rates
3. Reduces false-positives

Over the past two years, Golden Helix has analyzed over 30 whole-genome data sets. Of these, over 90% suffered from some form of experimental design error. This is usually caused by cases and controls not being balanced across plates, trios and families divided across multiple plates, multi-site genotyping where phenotypes are not randomized, and the merging of two or more datasets from different projects (one of the main benefits of imputation as discussed later). The impact is detrimental, as poor experimental design can cause endless struggles with genotyping artifacts, high type I error rates, the discarding of as much as 10% – 30% of genotypes, and more.

To illustrate a typical example where genotypes were miscalled due to batch effects, observe the cluster plots of A and B allele

intensities in Figure 1. The top plot is colored by genotype call. If no genotyping errors were present, each cluster would be a solid color. However, the middle cluster, representing genotypes that should be called heterozygous A\_B (green), actually contains homozygous A\_A calls (blue). The bottom plot is the same cluster plot colored by batch. In this particular study there were two control batches and two case batches. As evidenced by the identical pattern of blue points in the middle cluster, the miscalled genotypes were due to Case Batch 1.

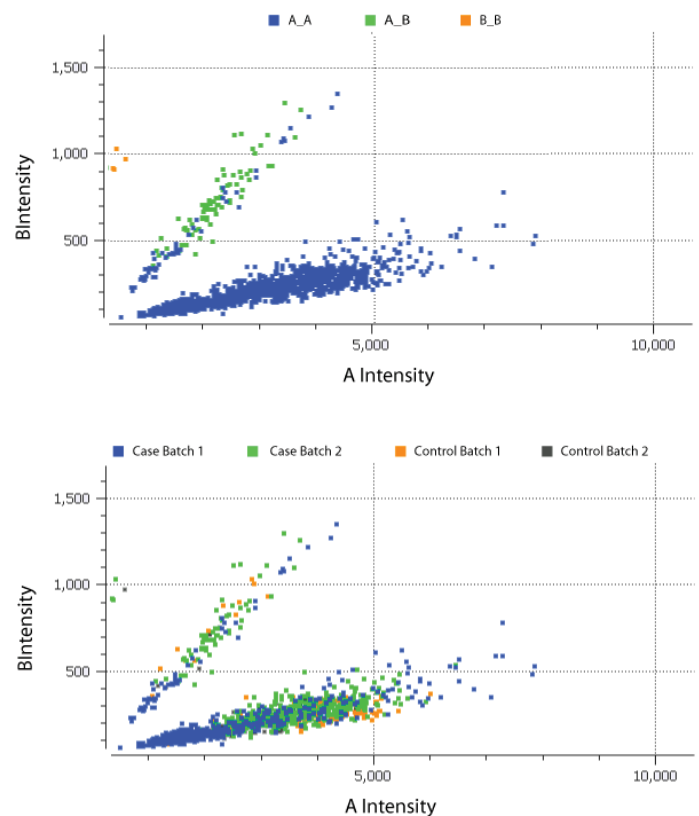


Figure 1: Cluster plots of A and B intensities for an unidentified SNP.

Aside from batch differences, genotyping errors can also result from noisy intensity data with poor cluster resolution. Such SNPs often have poor call rates, and the calls that are made tend to be questionable.

If data has already been acquired and the experiments cannot be rerun, how are the impacts of genotyping artifacts mitigated? In practice, most researchers adhere to very strict quality control measures to filter samples and genotypes that display poor

quality. Simply eliminating this data is not very encouraging, however, given the time and resources it takes to acquire. Other measures include principal component analysis (PCA). Though effective in some cases, PCA correction is as much art as science, and it is difficult to determine if gains in quality come at the expense of lost signals. Our goal, therefore, is to improve the accuracy of the genotype calls themselves using advanced calling algorithms that remove genotyping artifacts and simultaneously control for plate or batch effects.

BEAGLECALL<sup>2</sup>, developed by Dr. Brian Browning of the University of Auckland, is a novel analytic program designed to greatly

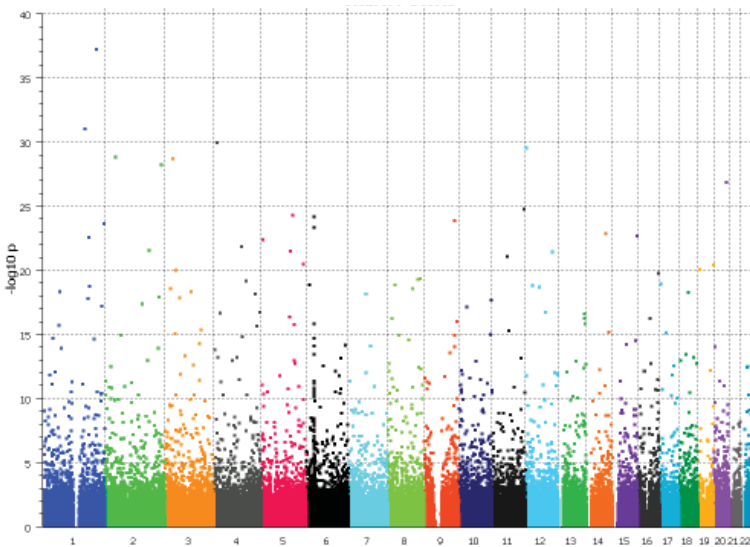
improve the accuracy of genotype calls and reduce false-positives. Whereas most genotype calling algorithms work exclusively with the A and B allele intensities, BEAGLECALL also uses haplotype frequency models – determined using pre-existing genotype calls – in conjunction with the raw intensities, to calculate optimal genotype calls for each sample. It also provides a solution for “rehabilitating” SNPs with poor call rates.

Figure 2, below, illustrates how BEAGLECALL increases the quality of genotype calls and removes batch effects. The two pairs of plots show genome-wide association results from two recent studies conducted by Golden Helix. The plots on the

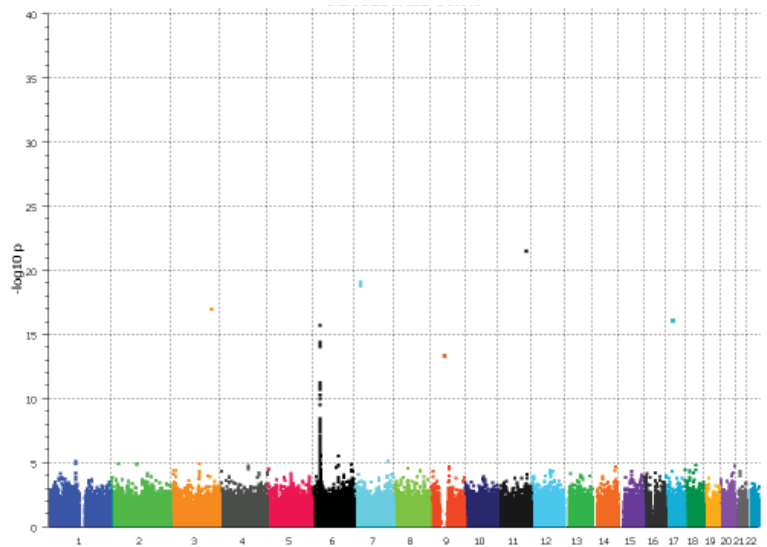
### Study 1

(Significant non-HLA regions have been moved to preserve identity of novel findings.)

CRLMM



BEAGLECALL



### Study 2

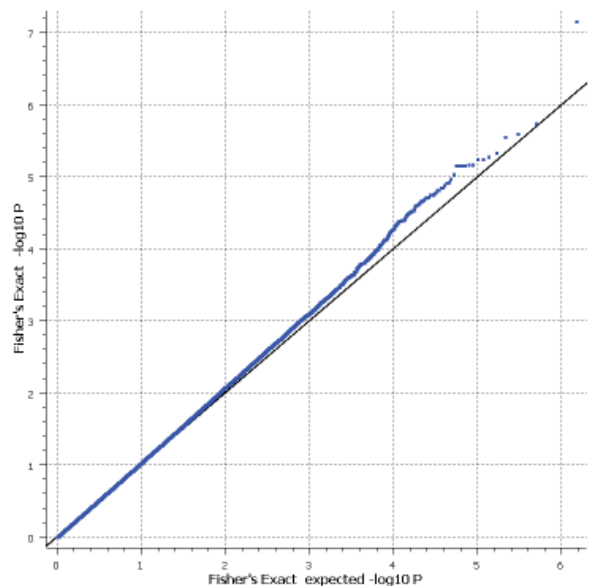
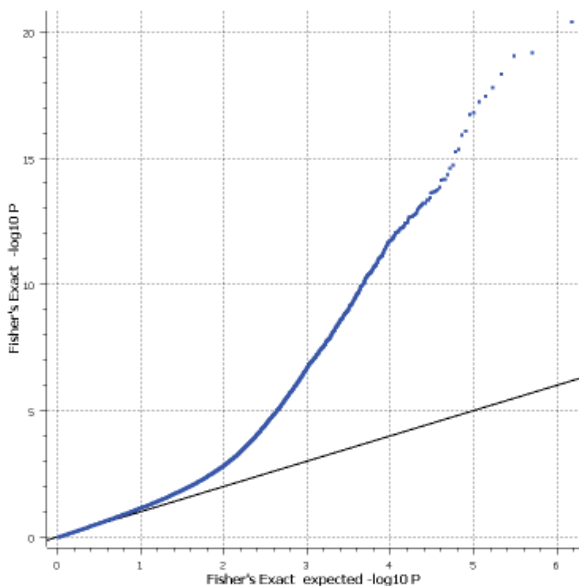


Figure 2: Manhattan and Q-Q plots for two different studies comparing p-values from CRLMM (left) genotypes and BEAGLECALL genotypes (right).

right show results from using BEAGLECALL for genotypes versus those on the left generated by the popular CRLMM<sup>6</sup> genotype calling algorithm. The Manhattan plots for Study 1 clearly show how BEAGLECALL increases the quality of signals and eliminate spurious associations. Likewise, the Q-Q plot from BEAGLECALL in Study 2 is much more encouraging than that from CRLMM.

To see how BEAGLECALL can rehabilitate SNPs with poor call rates, observe the two plots in Figure 3, discussed in the supplemental data section of Browning and Yu's 2009 American Journal of Human Genetics article<sup>2</sup>. These plots compare CHIAMO<sup>8</sup> genotype calls with BEAGLECALL calls for a marker (rs5015480) that was associated with type 2 diabetes in the WTCCC study<sup>9</sup>. Genotyped using the Affymetrix 500K array, CHIAMO had 123 uncalled genotypes versus BEAGLECALL's 33.

## IMPUTATION

### Benefits:

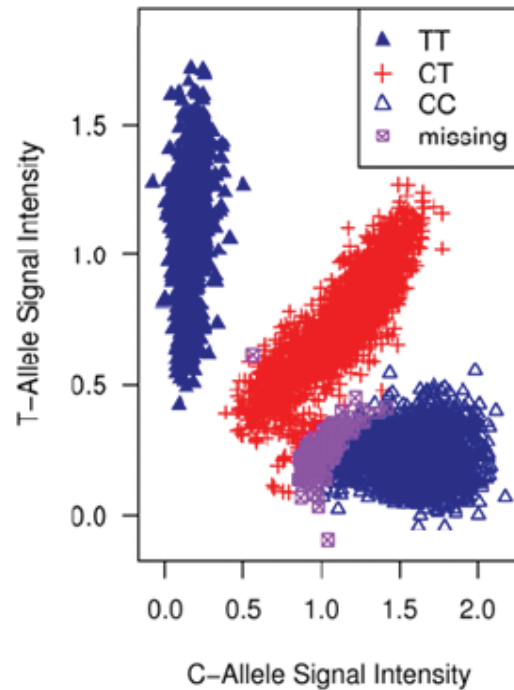
1. Enables the combining of two or more datasets, regardless of platforms or arrays
2. Increases the density of existing data with high-density reference data
3. Improves existing association results and enables the identification of novel association signals
4. Reduces post-analysis fine-mapping efforts

If the first generation of genome-wide association studies (GWAS) taught us anything, it was that large sample size and high density of markers are critical to finding significant associations that replicate. But again, even with the cost of whole genome microarrays and next generation sequencing technology in steady decline, it is still cost-prohibitive for many groups to conduct GWAS on a scale necessary to achieve publishable results. Therefore, to increase the power of a study within a limited budget, researchers are turning to external public and private datasets to supplement their own, and are using imputation methods to help combine them.

Given the abundance of publically available large-scale datasets (e.g. dbGaP, GAIN, and iControl), combining external data sources with internal data is a very cost effective approach to increasing power. In fact, researchers are now empowered to effectively perform meaningful GWAS analysis *without spending anything on data acquisition!* Furthermore, imputing data using HapMap and the 1000 Genomes Project as reference data, one can increase the density of a smaller GWAS study by an order of magnitude.

However, combining two or more datasets poses its own set of challenges. One such mega-analysis project being conducted by the Psychiatric GWAS Consortium is attempting to analyze a pooled set of approximately 12,000 cases and 9,000 controls.

### CHIAMO Calls



### BEAGLE Calls

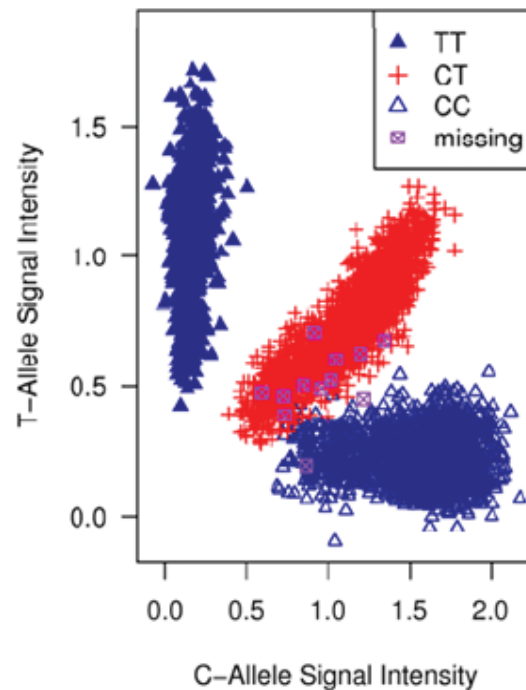


Figure 3: CHIAMO vs. BEAGLECALL genotype calls on SNP rs4015480 from the WTCCC study<sup>7</sup>.

Should they decide to use the 1000 Genomes as a reference, the resulting dataset could be over 21,000 total samples and up to 8 million SNPs for each. Beyond the data management and computational challenges of this extremely large dataset,

combining datasets that were never intended to be combined introduces the worst form of batch effects. In mega-analysis, one has to consider that every added data source was generated at different times, on different arrays, with different quality control procedures, using different strands to make calls, etc.

Additionally, researchers are tasked with sequencing or fine-mapping regions around significant variants in order to validate GWAS findings. The greater the number of regions and genomic distance around each significant region, the more costly fine-mapping becomes. Availability of higher density probes enables researchers to more accurately determine which regions to investigate further and actually narrow each region on which they should perform fine-mapping. The challenge, again, is obtaining higher density data within a limited budget.

BEAGLE<sup>3</sup>, also developed by Dr. Brian Browning of the University of Auckland, is designed to impute both non-overlapping genotypes when combining two datasets (where one or both is used as a reference) and sporadic missing data within a single dataset (no reference).

BEAGLE is well suited for large sample sizes, ideal considering most publically available reference panels consist of over a thousand samples. Most imputation algorithms are based on complex haplotype frequency models, where the larger the sample size, the more complex the models. Though larger reference panels, and therefore higher complexity, help with accuracy, significant computational resources are needed. In order to achieve results in an efficient manner (which could still be weeks or months on a whole-genome scale), competing methods must limit haplotype model complexity for larger samples sizes. BEAGLE does not, and therefore delivers superior quality for larger sample sizes. Further, despite not having to limit complexity, BEAGLE is typically one to two orders of magnitude faster than competing methods!

## COMBINING ADVANCED GENOTYPE CALLING WITH IMPUTATION

Incorporating both advanced genotype calling and imputation in an integrated approach is extremely important because the quality of imputation results is significantly dependent on the accuracy of the original genotype calls. Though BEAGLE (nor any other imputation algorithm, for that matter) does not directly address the problem of batch effects when combining two datasets, the option to run BEAGLECALL on each prior to imputation mitigates the added complexity. Internal and external comparisons have found that BEAGLECALL, which can work on any array platform, is the most accurate genotype calling algorithm as compared to CRLMM<sup>6</sup>, BRLMM<sup>1</sup>, Birdseed<sup>5</sup>, CHIAMO<sup>7</sup>, ILLUMINUS<sup>7</sup> and others<sup>2</sup>. It provides the highest call rates while nearly eliminating genotyping artifacts. Further, BEAGLE provides best-in-class accuracy and unrivaled speed for imputation as compared to Mach, Impute, PLINK, and others<sup>3,4</sup>.

## GOLDEN HELIX'S SOLUTION

The primary drawback with BEAGLECALL and BEAGLE is that they require significant resources, expertise, and high-performance computing capabilities to achieve optimal results within a reasonable timeframe. Having worked closely with Dr. Browning on the implementation of our solution, we have a detailed understanding of all the various BEAGLECALL and BEAGLE parameters and have optimized the programs to run in a cloud-enabled grid environment. Our collaborative approach ensures that the most appropriate reference data are used, the optimal parameters are employed (given each researcher's unique study design and phenotypes), strand issues are resolved, and results are delivered quickly in a ready-to-analyze format.

Using Golden Helix services, researchers can focus on their research while leveraging our high-performance computing infrastructure and the experience gained from dozens of whole genome studies and imputation projects. Within a matter of days of our receiving data, researchers can obtain optimal genotype calling and imputation results without having to learn complex command-line programs and with absolutely no investment in expensive equipment.

## REFERENCES

1. BRLMM White Paper, [http://www.affymetrix.com/support/technical/whitepapers/brlmm\\_whitepaper.pdf](http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf)
2. Browning and Yu. Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association..., *The American Journal of Human Genetics* (2009), doi:10.1016/j.ajhg.2009.11.004.
3. Browning and Browning. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals, *The American Journal of Human Genetics* (2009), doi:10.1016/j.ajhg.2009.01.005.
4. Chang, Lin, Tang, and Hsieh. Comparison of Genotype Imputation Methods for SNP Array Data, Poster #481: 2009 American Society of Human Genetics Conference.
5. Korn, Kuruvilla, McCarroll, Wysoker, Nemesh, Cawley, Hubbell, Veitch, Collins, Darvishi, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* (2008). 40, 1253–1260.
6. Lin, Carvalho, Cutler, Arking, Chakravarti, and Irizarry. Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays. *Genome Biology*. (2008) Apr 3;9(4):R63.
7. Teo, Inouye, Small, Gwilliam, Deloukas, Kwiatkowski, and Clark. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* (2007) 23, 2741–2746.
8. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* (2007) 447, 661–678.