



Case-Control GWAS analysis using digenysis

A faster, more robust, and more straightforward method for finding associations in case-control GWAS

The scientific aim of Genome-Wide Association Studies (GWAS) is to discover genetic factors that contribute to the development, progression, or treatment options for a particular disease or trait (such as high blood pressure or obesity). GWAS have proven to be particularly powerful for the study of common complex diseases such as asthma, cancer, diabetes, heart disease and mental illnesses where the individual genetic contributions to the disease are expected to be relatively weak.

The digenysis approach

- the genetic analysis of both allele variation and copy number
- diagenesis (root word) - a recombination of the constituents resulting in a new product

Standard methodologies are often statistically valid but are resource-intensive and sensitive to heterogeneous sources of variation such as measurement bias and the presence of copy number variants. Analysts work

diligently to "qualify" the SNP data but often throw away allelic measurements that contain useful, if partial, information. During the quality phase of the analysis, there can be give-and-take related to retention of subjects versus markers resulting in extra project and computational time spent performing reclustering whenever subjects are discarded.

The basic methodology of digenysis is a rigorous statistical approach that begins with the primary constituents of the study: the array measurements of the two alleles of each SNP. However, with digenysis, these measurements are not used to call genotypes. Instead, digenysis recombines and transforms these allele measurements into values that contrast as well as summarize the allelic intensities. It then performs statistical tests for allelic and copy number association directly while simultaneously accounting for latent factors such as subpopulation and processing biases. Figure 1 provides an overview of the standard analysis process flow and the alternative flow of digenysis.

Efficient GWAS Work Flow with digenysis

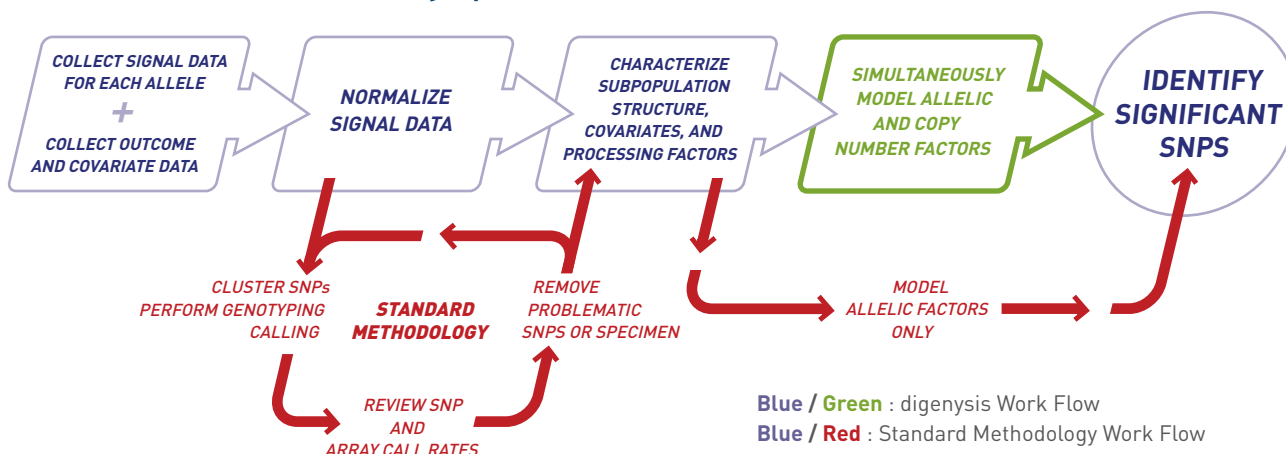


Figure 1 - digenysis analysis flow vs. the standard methodology work flow. With digenysis, the genotype-calling/SNP review process is not required, thereby greatly saving analysis resources and time. In addition, digenysis performs analysis of both allelic and copy number association.

A faster, more robust, and more straightforward method of finding associations in case-control GWAS

Case-control association studies are, in essence, population studies. Even though the current standard methodology requires the estimation of each individual's alleles/genotypes or CNVs, the final inference is not about the individual but rather about the population that was sampled. In addition, the genotype-calling process itself is expensive and frequently inaccurate due to the inconsistencies of the clusters (see Figure 2). This motivates our alternative and robust analysis to circumvent these issues.

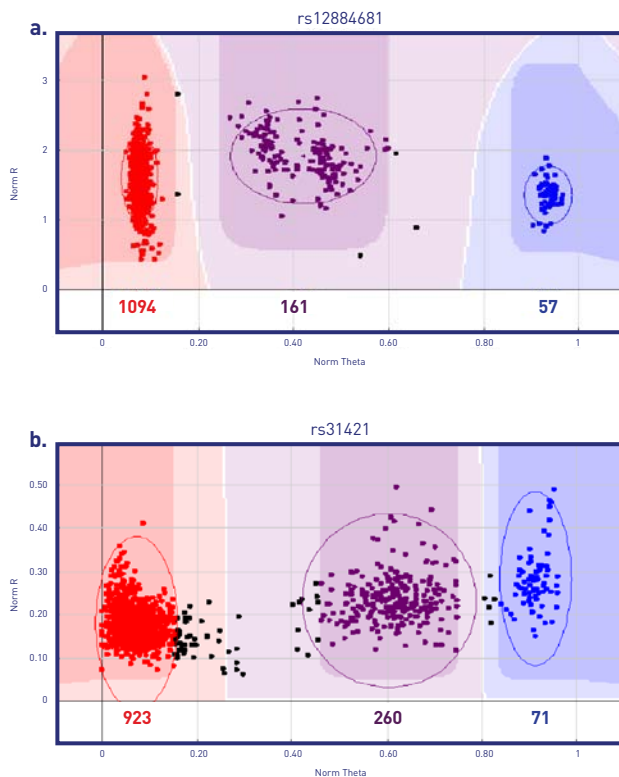


Figure 2 Examples of SNP clusters and calling. Homozygous clusters are at the extremes of "Norm Theta" with heterozygotes in-between. a) Typical SNP which has few missing calls and the clusters are well-defined. Nevertheless, it significantly deviates from HWE assumptions ($p < 10^{-15}$) and is typically removed from analysis. The other cluster graph (b) has a much higher no-call rate (12.2%) as indicated by the black points, yet it is clear that the missing genotypes convey useful information that would be lost, thereby reducing power, using the standard calling and analysis methodology. Overall, 2%-15% of SNPs are removed for these reasons, often unnecessarily.

Specific advantages include

- we do not remove SNPs due strictly to lower call rates as digenysis does not require genotype calls and the allelic intensities may be providing useful information, especially when CNV are present
- there are no samples that we discard due to low call rates as digenysis does not require genotype calls (samples may be removed for other reasons, such as reducing population heterogeneity)
- we do not remove SNPs simply because they violate HWE assumptions as the genotype calls necessary for HWE analysis may be in error—instead we focus on the association of the allelic intensities with outcome
- there is no need to recluster SNPs when samples are removed as digenysis does not require SNP clustering

Perhaps not surprisingly, using the original intensity data for association analysis is very similar to using the genotype calls. In fact, logistic regression using the allelic intensity contrast is statistically equivalent to the Cochran-Armitage trend test when there is no measurement or classification error. Bifurcation of the allelic contrast intensity data for dominant/recessive testing is also equivalent.

Summary of digenysis™ benefits:

- Power of approach typically matches and can better traditional approaches, especially as variation in signal measurements increase or as CNV are present
- Faster turnaround time (days as opposed to months)
- Simultaneous analysis of allelic and copy number association
- Covariates such as population, processing effects, and exposure are easily and seamlessly integrated (or removed)
- Retains more specimens and SNPs than the typical or standard statistical approaches
 - "Problematic" SNPs may sometimes be conferring important information
- Platform independent - the method requires only individual allele intensity measurements
- Ploidy independent

