

A REVOLUTIONARY APPROACH TO GWAS ANALYSIS

Wendell Jones, Steve McPhail, and Joel Parker
Expression Analysis, Inc, Durham, NC USA

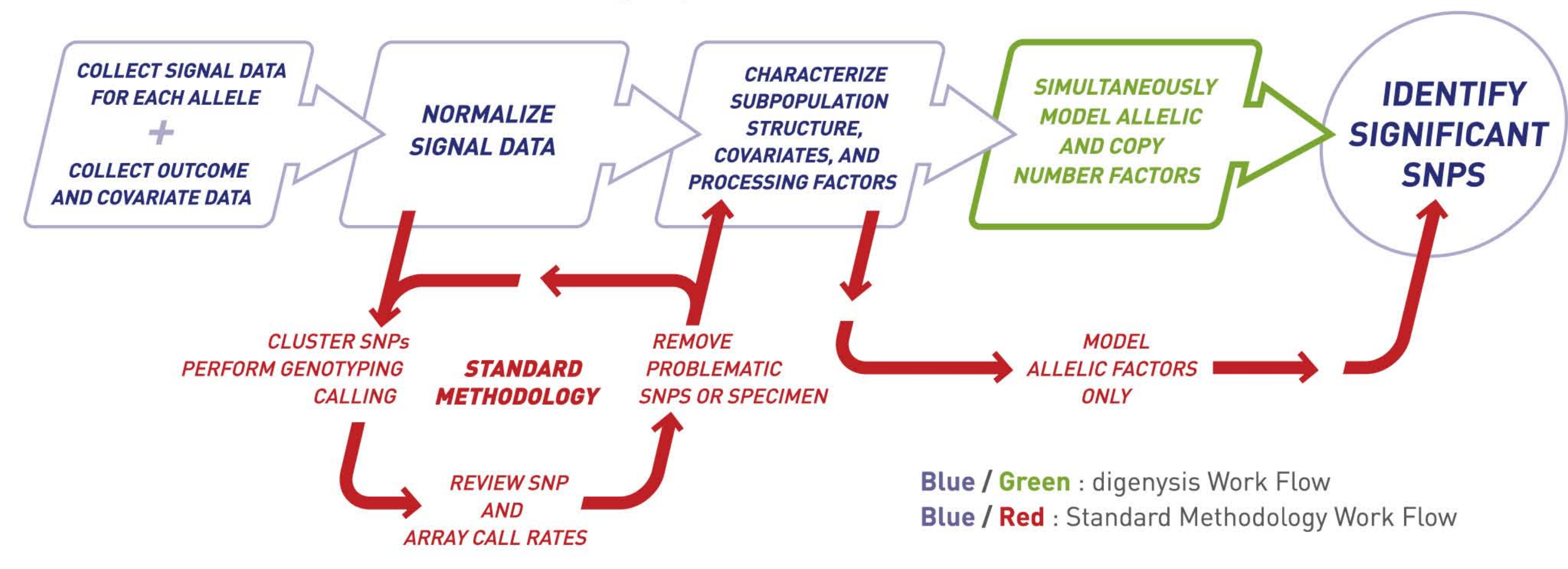


INTRODUCTION

GWAS studies are a useful method for examining and identifying genes and genomic regions that are significantly associated with disease susceptibility or other phenotypic traits. They frequently point to areas that are at the root cause of disease or other phenotypic aspects such as height or hair color. Typically, GWAS involve measuring hundreds to thousands of individual biological samples in a case-control experiment of unrelated individuals. The measurements are typically made by whole-genome SNP/CNV micro-arrays, which can target hundreds of thousands to millions of known variants throughout the genome. Computational and statistical methods are then applied to the data to yield associations between outcomes (cases) and genomic characteristics (genotypes) in the presence of other potentially influential factors, such as population traits, age, environmental exposures, etc. GWAS, as opposed to candidate gene studies, permit a high-coverage scan of the genome in an unbiased fashion and thus have the potential to identify previously unknown susceptibility factors.



Efficient GWAS Work Flow with digenysis

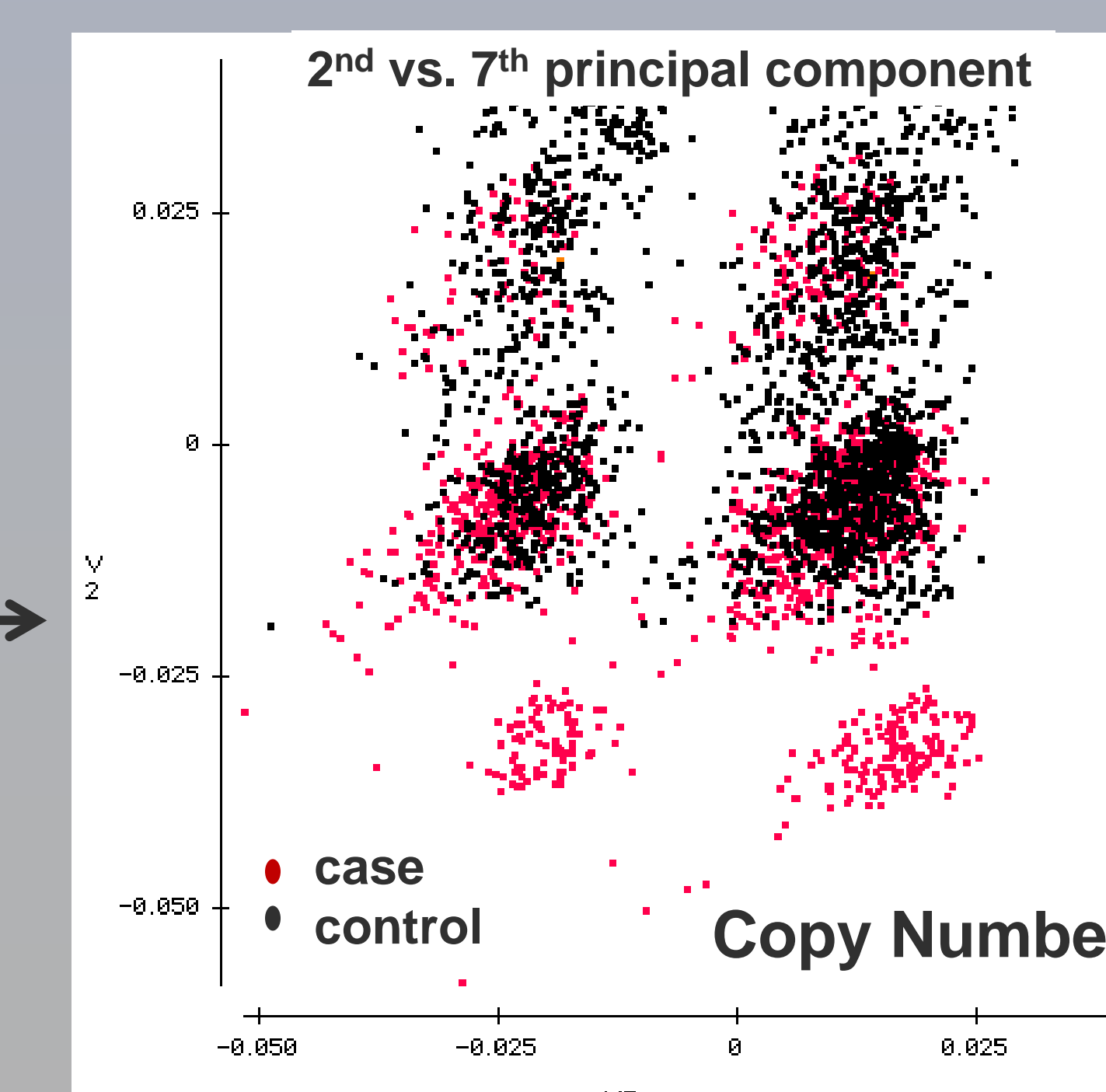
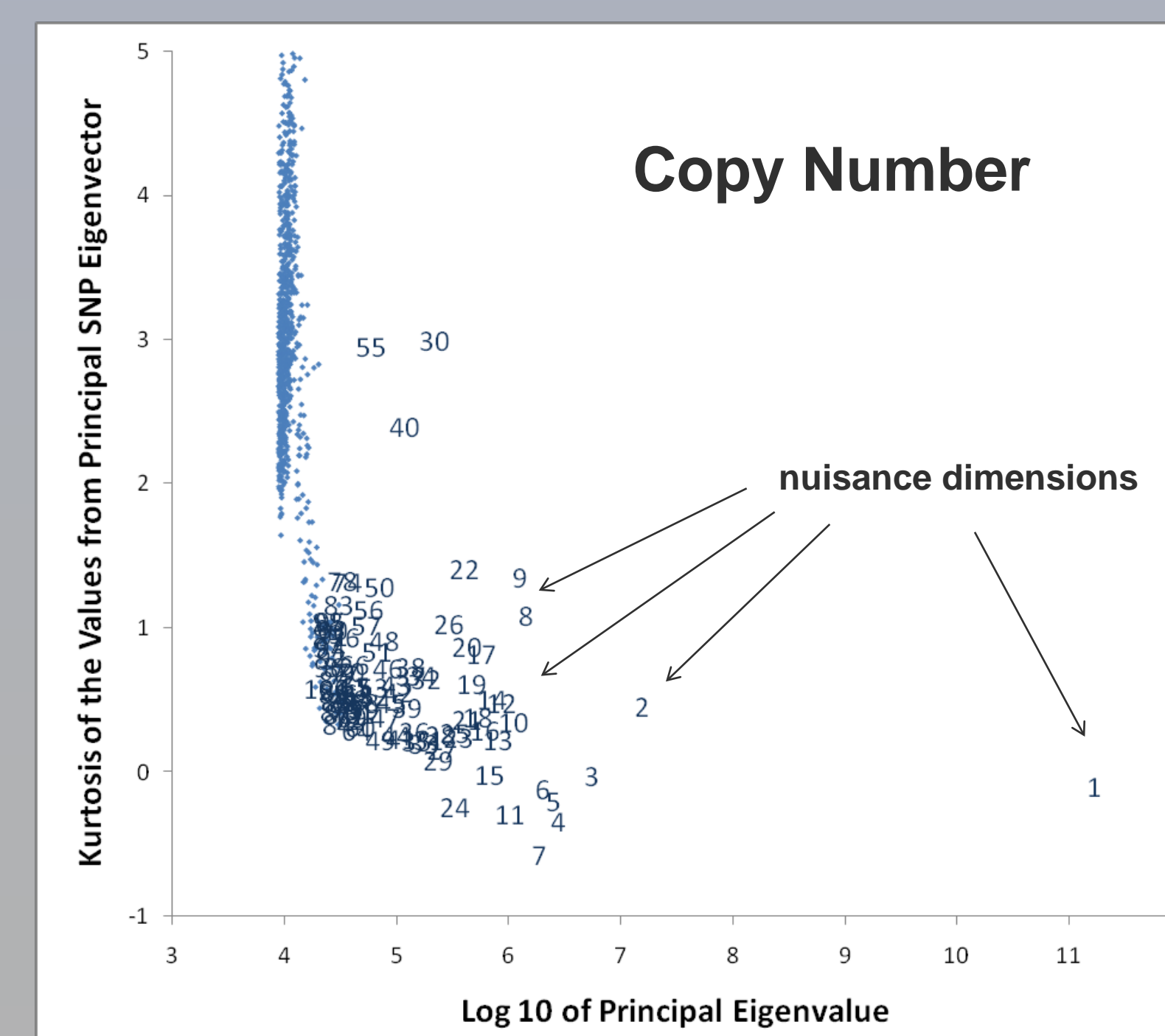
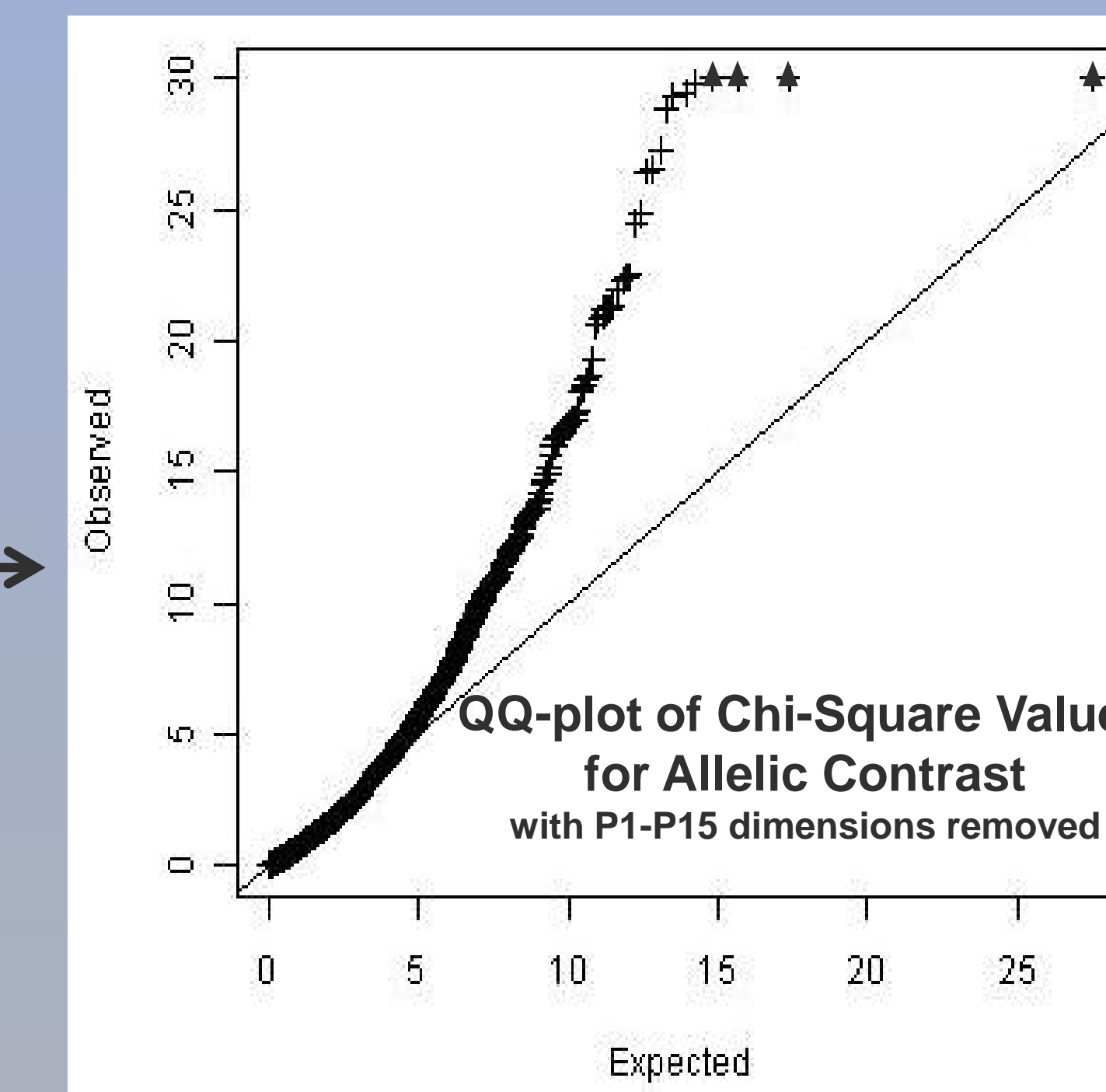
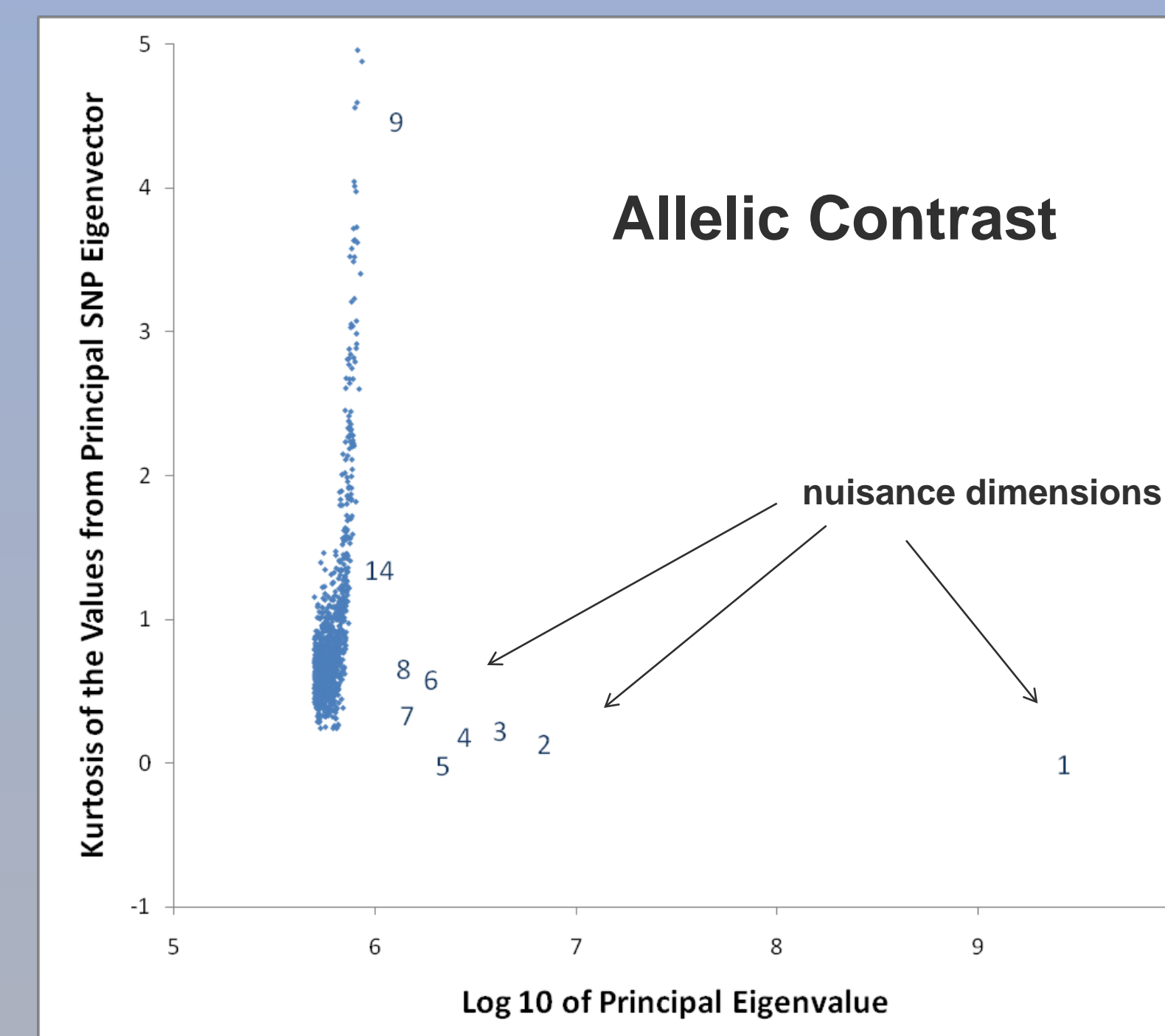


digenysis FACILITATES DETECTION AND REMOVAL OF TECHNICAL ARTIFACTS AND UNWANTED POPULATION EFFECTS

Technical artifacts and population effects can have dramatic impacts on GWAS and the WTCC studies are a good example. We present here examples of technically-driven or population-driven structure in the allele intensity values. We know this because of the association of the size of the principal dimensions of the intensity data with the eigenstructure (the distribution of values from the corresponding eigenvectors associated with SNPs) of the intensity data. Pictured below are values of kurtosis of the distribution of values within the SNP-oriented eigenvectors. Kurtosis is a useful statistic as smaller or negative values of kurtosis imply that the distribution of the values are more uniformly distributed. Our graphs indicate that the top eigenvalues are strongly associated with SNP-based eigenvectors which have a more uniform distribution of values across the genome. This implies that a series of major effects in the SNP data are influencing allele intensities genome-wide. We will assume that the source of these genome-wide effects are either real subpopulation effects or technical artifacts due to batch processing of samples that should be removed/accounted for to reduce unwanted bias in the study, leading to false positives.

EXAMPLE – WTCC GWAS – Rheumatoid Arthritis

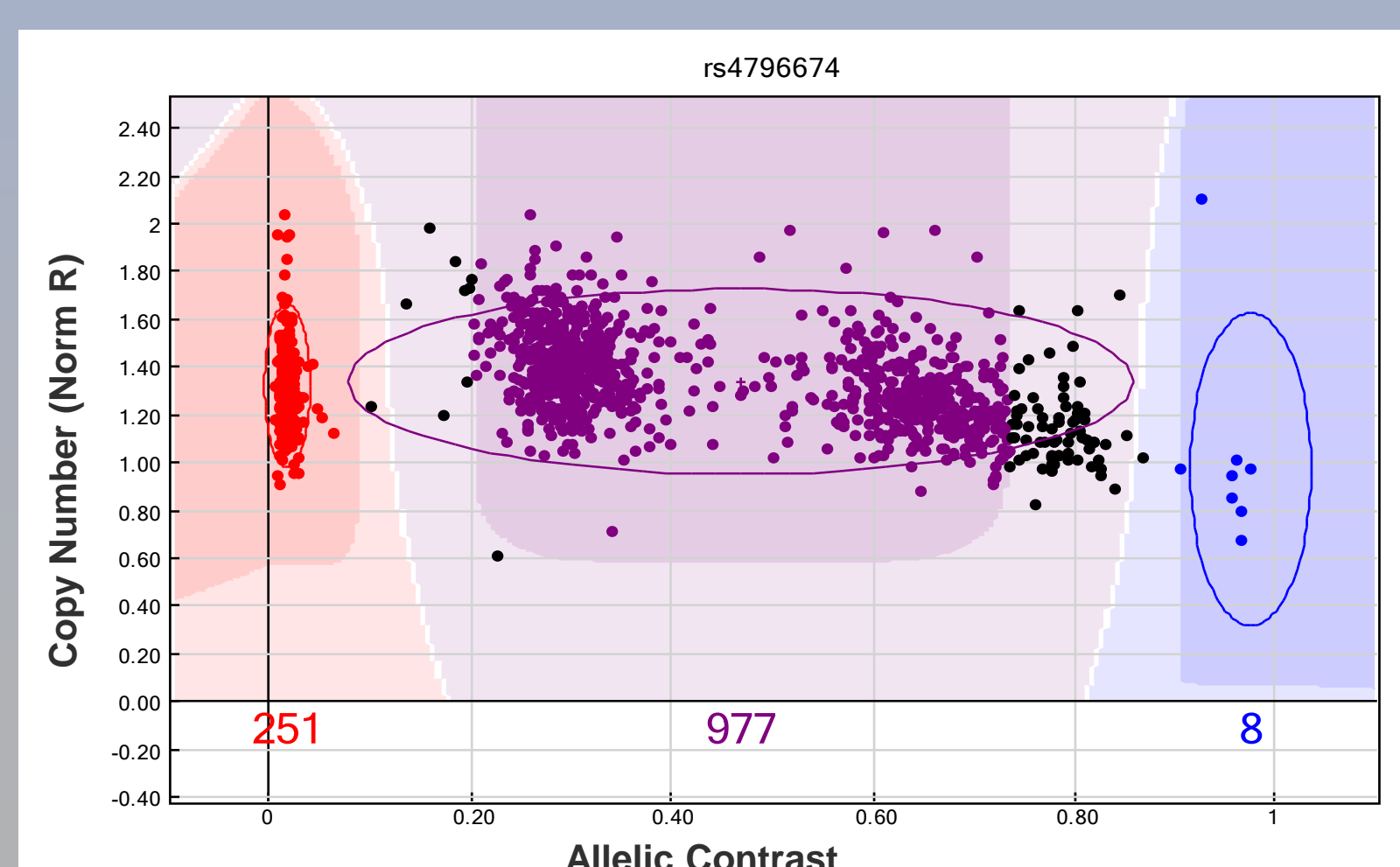
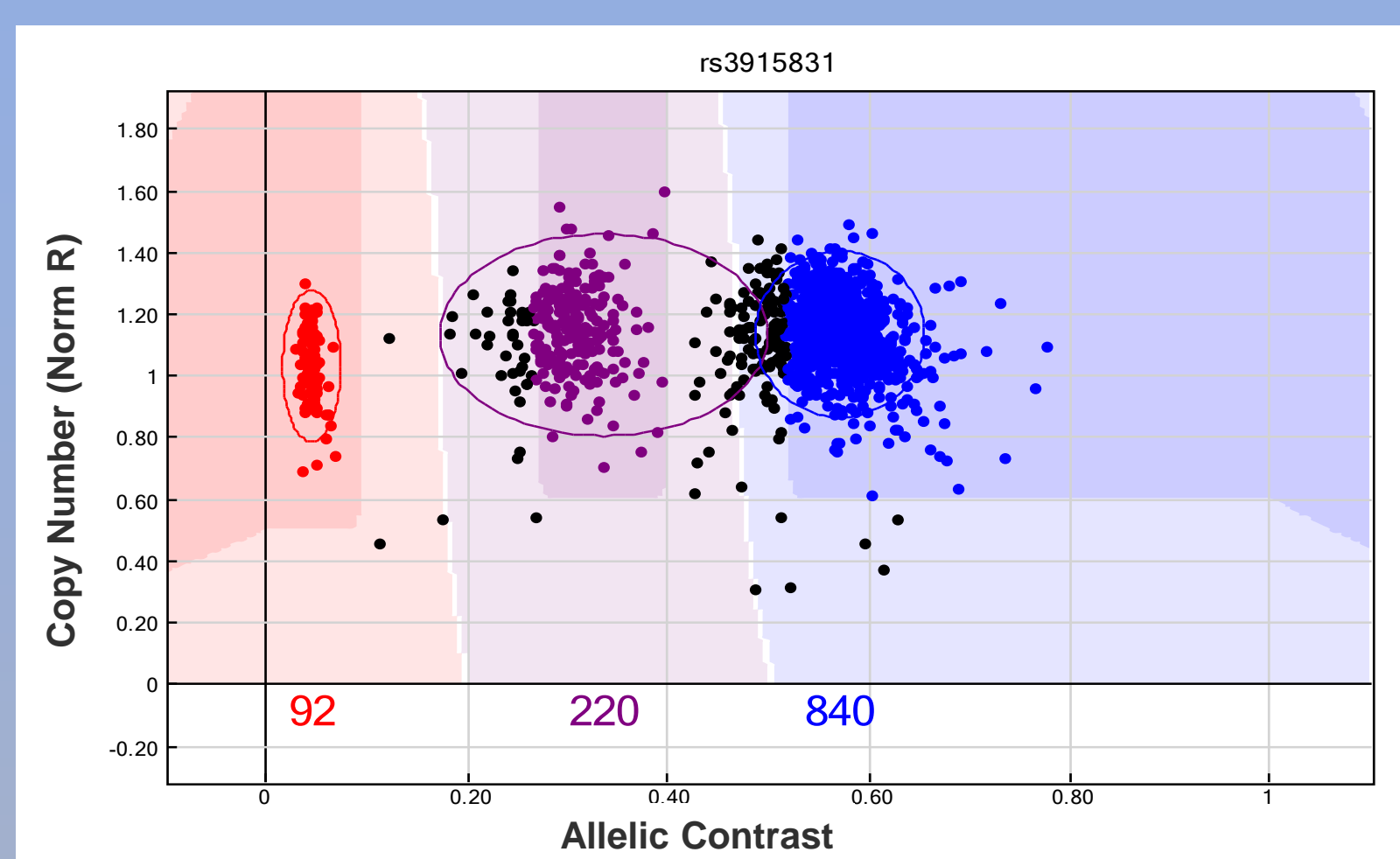
Kurtosis of SNP eigenvectors from matrix of genetic measurements (allelic contrast or copy number) versus the principal eigenvalues of the matrix



BENEFITS OF OUR APPROACH

- Simultaneous analysis of allelic and copy number association with outcome
- Identification of nuisance dimensions in the data
- Fast turnaround time – days instead of months
- Fewer assumptions are made implying your study can
 - utilize more SNPs (SNPs are not routinely removed except for monomorphisms)
 - utilize more samples (marginal quality samples are usually retained)
 - no constraints regarding AA, AB, BB canonical assumptions of genotype
- Directly address subpopulation effects
- Easily integrate covariates, such as age, weight, env. exposure
- Implementations of method are flexible and includes testing that is equivalent to trend and/or dominant-recessive tests
- Platform independent
 - Analysis can be done on Affy 6.0, 5.0, 500k and any Illumina platform (1M, 1MDuo, 660, 610, 550, etc.) as well as other genotyping platforms

ISSUES WITH ASSIGNING GENOTYPES TO EVERY SNP



If a) there are unusual genomic phenomena (such as chromosome loss or gain),
or b) MAF is smaller (e.g., < 15%)
or c) the study is of insufficient size for the minor allele to appear to define the het or MA homo cluster
or d) there is poorer quality or biases in the sample processing,

then there is a much greater likelihood that the SNP will be genotyped incorrectly. When properly assessed, these SNPs may be some of the most interesting to examine in a follow-on to your genome-wide study.

Since our approach does not force a genotypic call, we are not constrained by the implications of incorrect calls, yet our method has the same or better power as when the calls are made correctly (from simulations and empirical evidence).

REFERENCES

- Korn, JJ, Kuruvilla, FG, McCarroll, SA, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs, *Nat Genet*, advance on-line Sept.
- WTCC Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145):661-78.