Rapid multivariate analysis of 269 Hapmap subjects and 1 million SNPs using ‘taxonomy3’

Rémi Lebret [1], Serge Iovleff [1], Christophe Biernacki [1], Julien Jacques [1], Cristian Preda [1], Alun McCarthy [2], Olivier Delrieu [2*]
1: Université Lille 1, UFR de Mathématiques, UMR CNRS 8524, 59655 Villeneuve d'Ascq Cedex, France.
2: Pharmacogenomic Innovative Solutions Ltd, Aston Court, Kingsmead Business Park, High Wycombe, Bucks, HP11 1LA, United Kingdom.
***Correspondence to: olivier.delrieu@pgxis.com**

PGXiS PharmacoGenomic Innovative Solutions
Part of CytoPathfinder Inc

ABSTRACT

‘Taxonomy3’ is a novel mathematical method for the **multivariate analysis of complex datasets**. It is based on correlations of individualized divergences named Log Bayes Factors (LBFs), and their Eigen decomposition. We applied this method to **269 subjects** of the Hapmap project (Africans , Caucasians, Chinese and Japanese) genotyped for **more than 1 million SNPs** (Illumina 1Mduo chip). We used **newly developed software** able to efficiently analyse such large datasets. Results show significant distinctions between ethnic groups, and sets of markers of importance for these distinctions. Multivariate models based on all available SNPs, accurately predict subjects’ ethnicity. This confirms the benefits of this new method: **powerful signal detection** with small number of subjects, **sub-group identification** facilitating personalized medicine and ability to build **multivariate predictive models using the whole genome**. We intend to apply this method and software to other large and complex disease and pharmacogenetic datasets.

INTRODUCTION

Taxonomy3 ('Tax3') is a novel mathematical method for the analysis of complex datasets [1,2].

Tax3 is multivariate, and does not require correction for multiple testing, as only one single statistical test is performed. Hence, it can detect significant differences between small cohorts, even when large number of variables are investigated.

Tax3 probes heterogeneity: it investigates correlation patterns in cohorts, and not only delivers significant variables for distinguishing cohorts as a whole, but also reveals meaningful subsets of subjects and significant variables distinguishing them.

OBJECTIVES

Our objective was to develop software able to perform rapid and cost effective tax3 analysis of large datasets generated by the latest genotyping technology.

SUBJECTS AND MATERIAL

We used a publicly available (Illumina FTP site) dataset composed of 269 Hapmap subjects: 90 sub-Saharan Africans used as reference group (controls) and 89 Caucasians, 45 Chinese, 45 Japanese used as 'cases'. Subjects were genotyped with the Illumina 1Mduov3 chip. A total of 1,152,058 SNPs with a genotyping success rate of >90% were analysed.

SNP LEVEL ANALYSIS

Figure 1 - Biplot

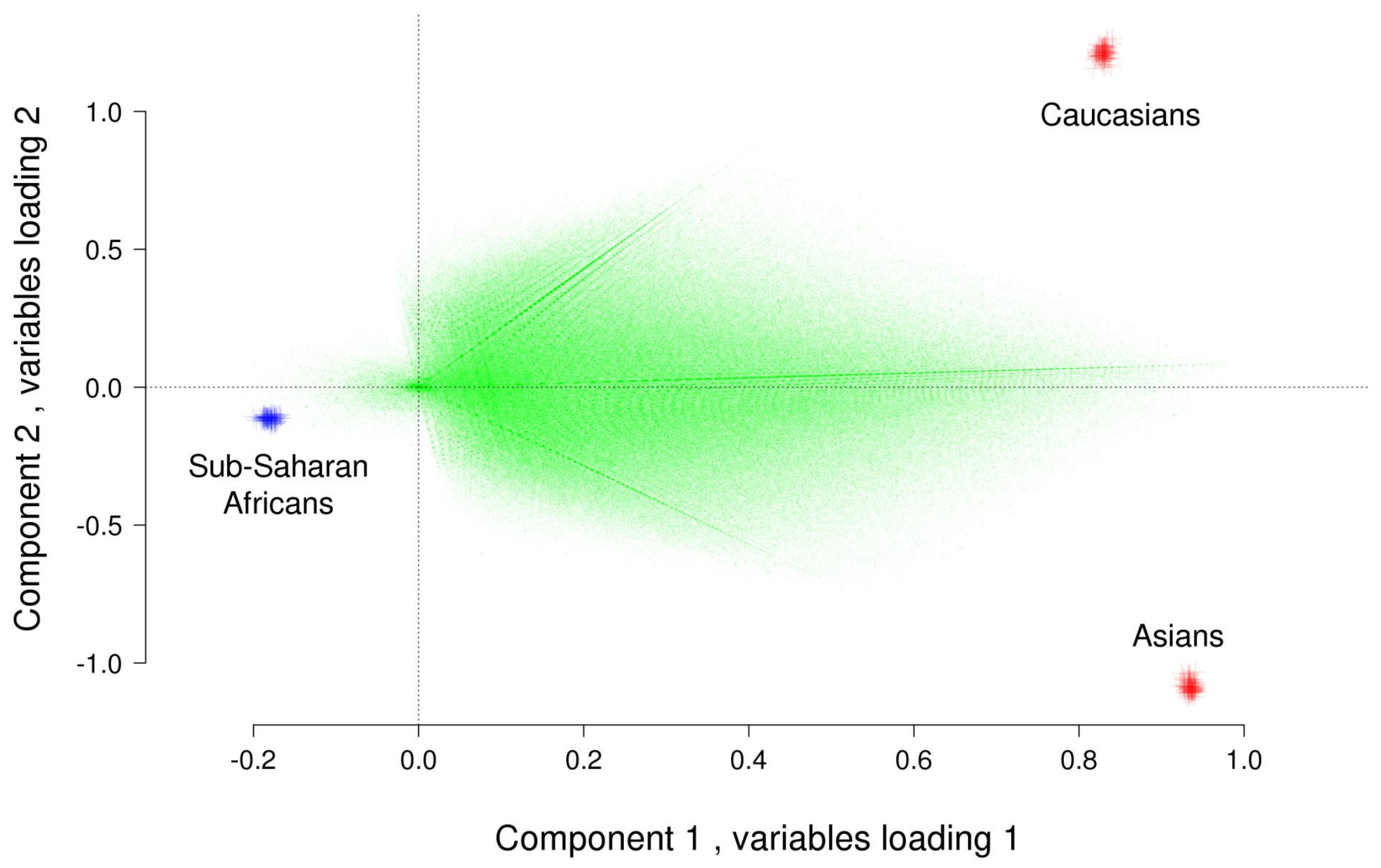


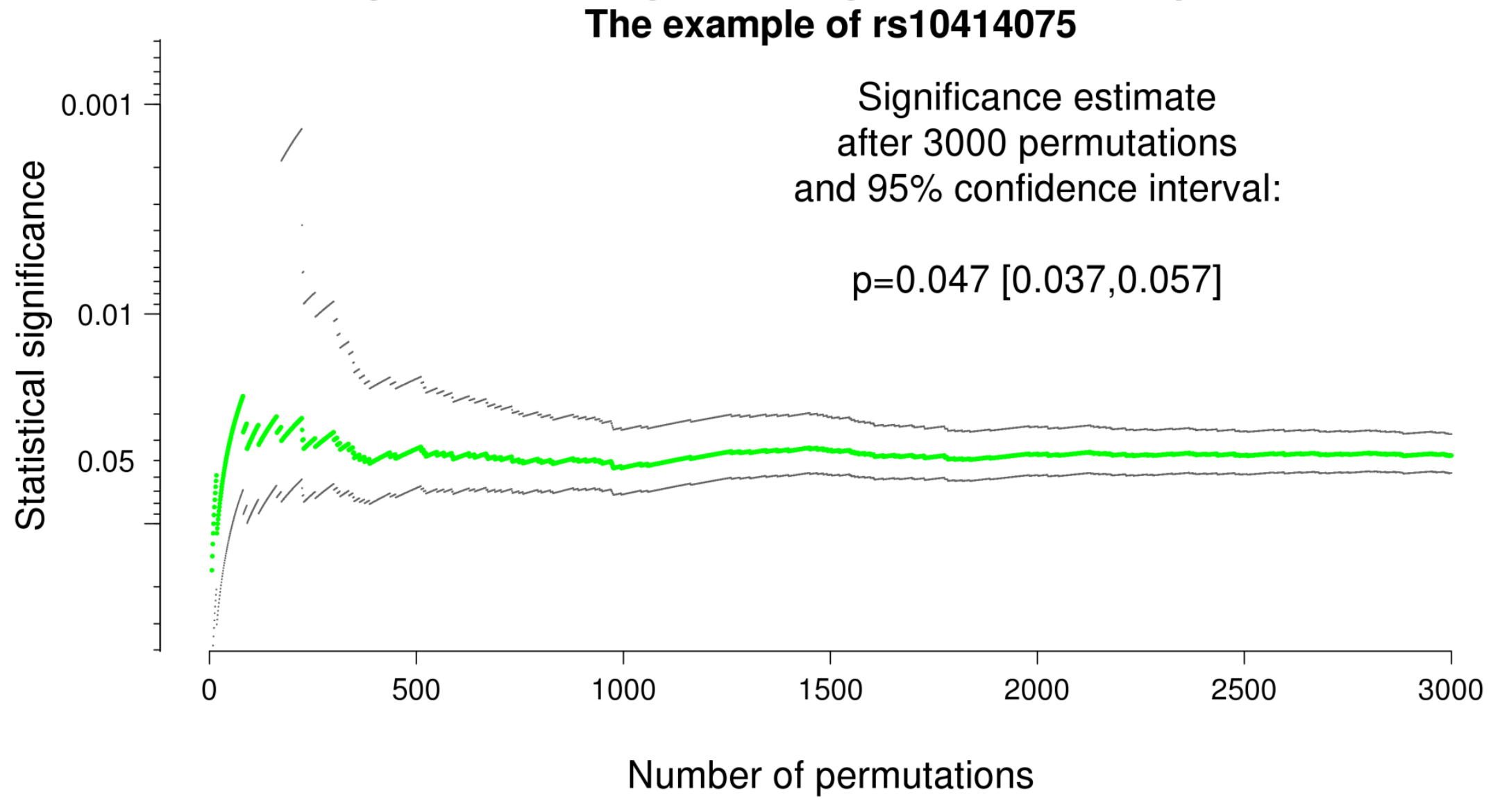
Figure 1 shows the two components explaining most of the dataset variability. The non-spherical shape of the variables cloud (green) indicates a strong signal differentiating controls (blue) from cases (red).

Variables having a large loading have a large effect on between-cohorts distinction, and are of potential interest.

The method (blinded to subjects’ ethnic status) reveals strong heterogeneity within cases, tight clusters of Caucasians and Asians, as seen on the 2nd component.

SIGNIFICANCE ASSESSMENT

Figure 2 - Convergence of significance on component 1
The example of rs10414075



The effect size (loadings) represented in Figure 1 should be translated into statistical significance (“p-values”), to confirm that there is a significant signal differentiating cases vs. controls and to identify which variables are playing a significant role.

This assessment is carried out by randomly permuting subjects' case/control status. This gives, for each variable, the probability of observing by chance a higher effect. About 3000 permutations are needed to obtain good estimates when p~0.05, as seen in Figure 2.

MULTIVARIATE SIGNIFICANCE

Figure 3 - Statistical significance at 0.05 level

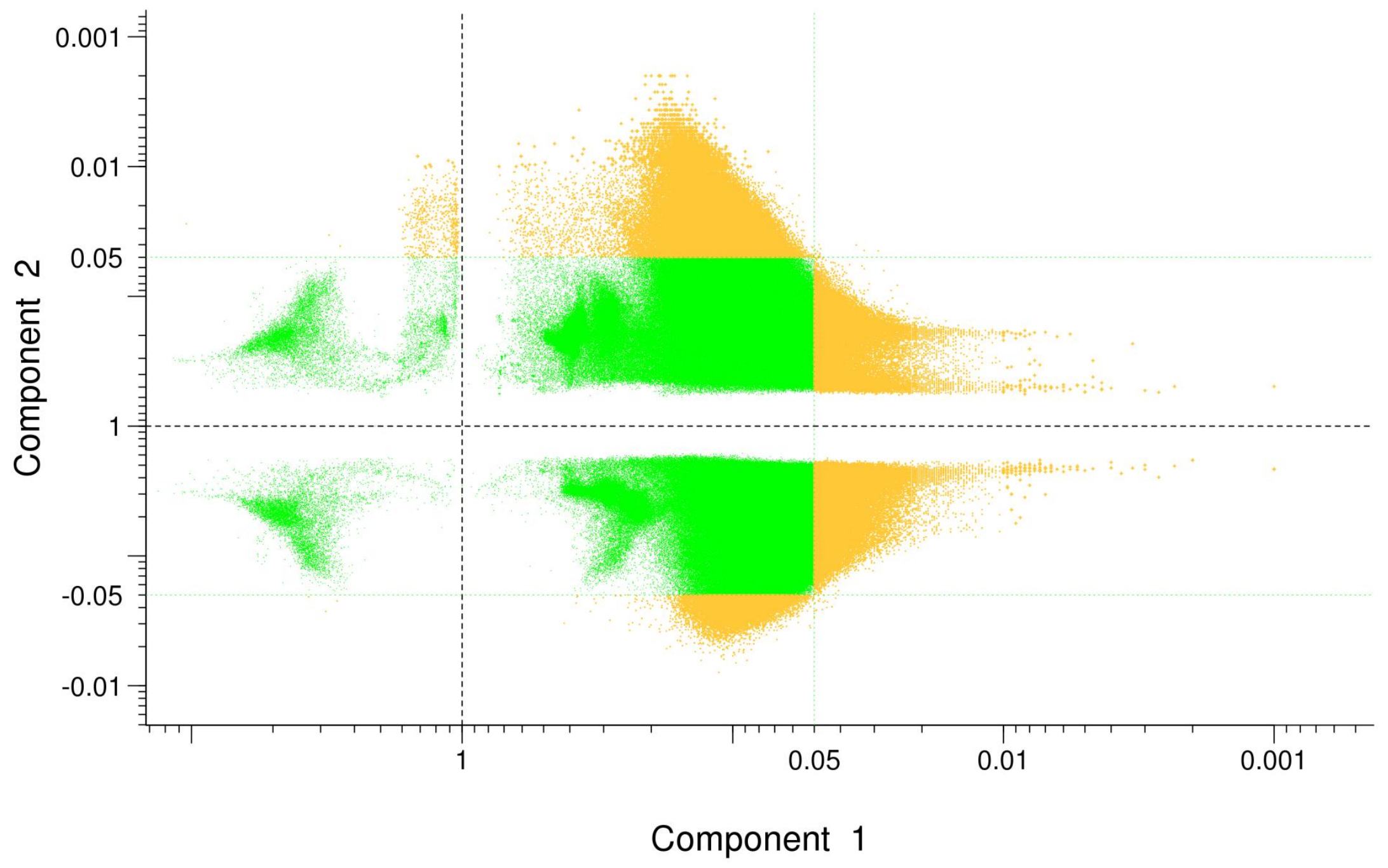
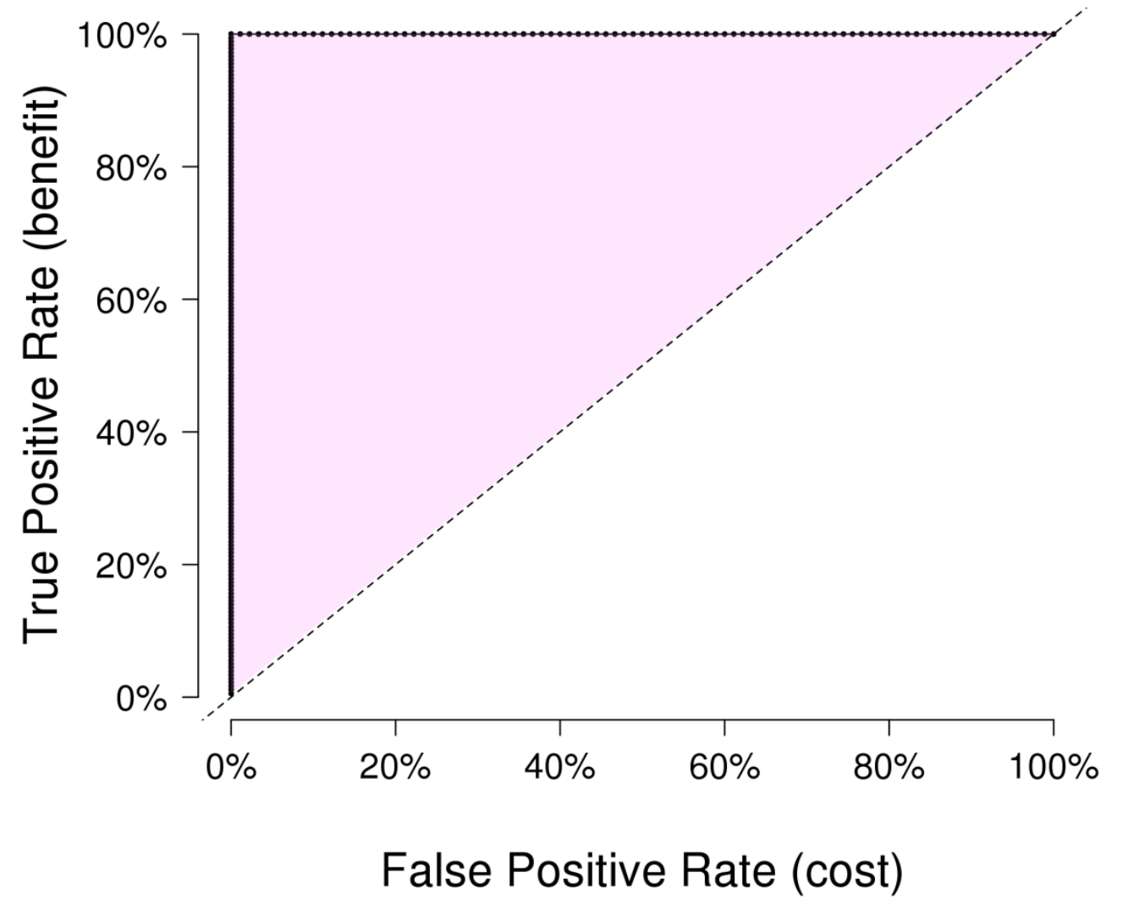


Figure 3 shows significance of each variable on the first two components. Variables below the usual 0.05 level are displayed in orange (a negative significance simply means a negative loading)

This confirms the strong signal differentiating cohorts, and the large number of SNPs which significantly diverged between the analyzed populations: a total of 218,895 SNPs are significant: 109,147 on component 1 (controls vs all cases) and 109,781 on component 2 (specific to controls vs Caucasians or controls vs Asians).

STATUS INFERENCE

Figure 4
Receiver Operating Characteristics curve



Subject case/control status can be inferred using all available data from other subjects. If all subjects are sequentially given an ‘unknown’ status, the predictive characteristics of the dataset can be represented as an ROC curve, as seen in figure 4.

This effectively builds a whole-genome-scan predictive model, based on variables loadings.

Figure 4 shows that subjects’ case/control status (African vs non-African) can be perfectly predicted, with 100% True Positives and 0% False Positives.

METHODS

Taxonomy3 is based on correlations of individualized divergences named Log Bayes Factors (LBFs), and their Eigen decomposition.

The initial data matrix is transformed into a LBF matrix of the same dimension, representing the *information gain* provided by each subject and SNP pertaining to the overall case/control distinction:

$$LBF(i,j) = \log \frac{L_{cases}(j,k)}{L_{controls}(j,k)}$$

where *k* is the genotype of subject *i* for SNP *j*, and *L* the Bayesian likelihood estimate of *k*.

Eigen decomposition of correlations of LBFs is the multivariate analysis of choice as it produces independent sets of variables, leading to phenomenological insight and understanding (Figure 1).

Assessment of significance is carried out by case/control status resampling (Figures 2 and 3). For each variable, relationship between significance *p*, its precision ϵ for a type I error α and *n* permutations is given by:

$$n = p(1-p)\left[\frac{Z_{1-\alpha/2}}{\varepsilon}\right]^2 .$$

Subjects of unknown status are attributed LBFs given their genotypes. In the Eigen decomposition, their position relative to cases and controls allows to infer their status. Leave-One-Out Cross-Validation gives dataset predictive characteristics (Figure 4).

SOFTWARE

Our proprietary software is written in C, multi-threaded and encodes SNPs at the binary level.

The single tax3 analysis displayed in Figure 1 was performed in a few minutes with 8 CPUs and about 500MB RAM.

Our virtual Linux cluster, with 160 CPUs, performed 3000 permutations (Figures 2 and 3) in a few hours.

CONCLUSION

The taxonomy3 method is extremely promising for signal detection, inference of disease status and probing heterogeneity. Our software and IT framework allow rapid and efficient analyses. We intend to apply this method and our know-how of other large and complex disease and pharmacogenomic datasets.

REFERENCES

1. On using the correlations of divergences. Delrieu O and Bowman C. In: Barber, S, Baxter, P D and Mardia, K V (eds) (2007). Systems Biology and Statistical Bioinformatics. University of Leeds pp 27-35

2. Visualizing gene determinants of disease in drug discovery. Delrieu O and Bowman C. Pharmacogenomics. 2006 Apr;7(3):311-29.