# Supplementary Material for The Genotypic Imperative: Unraveling Disease-Permittivity in Functional Modules of Complex Diseases

Abdoul K. Kaba, Kelly L. Vomo-Donfack, Ian Morilla

December 10, 2023

## 1 Supplementary Results

In this section, we present additional results that complement and enhance the understanding of the issues addressed in the main manuscript. We will follow the same order as in the results section of the primary document.

### 1.1 Removing Unwanted Variation

Before conducting any calculations, we performed data denoising analyses at various levels of statistical and genomic information, as illustrated in Figures S1 to S6. To carry out this analysis, we followed the guidelines provided in the *RUVcorr* vignette, in conjunction with its corresponding R-package available at http://bioconductor.org/packages/release/bioc/vignettes/RUVcorr/inst/doc/Vignette.html [1].

Assuming the common scenario of dependence between biological datasets (denoted as $X$) and a consistent source of noise (referred to as $N$) [2, 3], we determined the effectiveness of the data cleaning process based on the ridge parameter $\nu$ and the dimensionality of $N$, denoted as $k$. As introduced earlier, $N$ is the matrix of unobserved covariates that accounts for the noise, following a known gene correlation model:

$$Y = X\beta + N\alpha + \epsilon \tag{1}$$

Then $Y$ is the matrix of gene expression in the MRCA dataset and $X$ is the matrix of the unobserved variables.

On initial observation, Figures S2 and S3 suggest $k = 4$ corrects the wide range of the distribution of the correlations between random MRCA genes, with some little non-zero variance in the global MRCA dataset. To figure out what is going on those non-zero correlations, relative log expression (RLE) plots (see Figure S4) are useful approaches.

The variation fixed to $\nu = 5000$ seems to overtake any eventual performance of the model using the other values. However, the selected $\nu$ should do remove the variation derived from known source. Thus, the next figure may help with this task. In effect, Figure S5 confirms that after empirical batch-correction the remaining noise set is a null set. Now, we are able to leverage this cleaned dataset and making robust inference on a large set of analyses such as expression profile correlations, mechanistic network models of disease, etc.

## 2 e-QTL Analyses

From Figures S7-S10, we can observe how the gene expression vs genotype trend observed in all the MRCA samples resembles that of their SNPS co-expression introduced in the main manuscript.

## 3 Model and Bell Inequalities Count

This section shows Figure S11, which summarises various model behaviours depending on the se-lected $TAD$ and DNA-binding domains of transcription factors. Yet Figure S12 describes the cor-rections in counts of the $TP53$ and $MDM2$ model after considering the geometric mean as a good

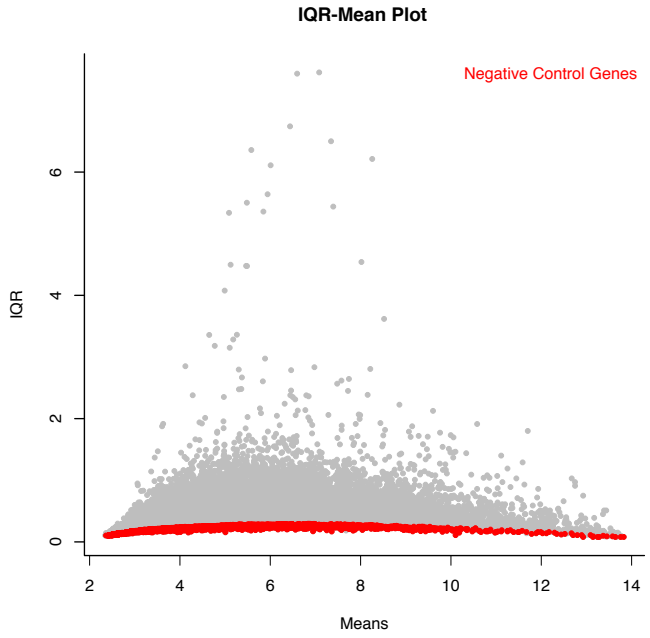


Figure S1: Interquartile range (IQR) vs Means plot of MRCA dataset. The IQR estimates quanti-tatively the spanning between gene sample extreme expressions respect their means. Highlighted in red the empirically chosen negative control or housekeeping genes.

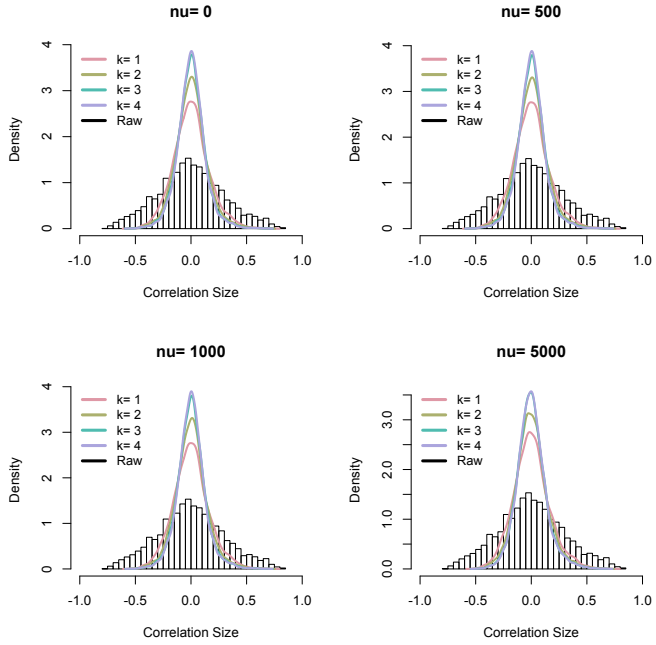


Figure S2: Impact on the MRCA dataset cleaning of different levels of background correction. Clockwise from top-left: correlation densities for $\nu = \{0, 500, 1000, 5000\}$ and $k = 1, 2, 3, 4$. Highlighted in black, the density of the randomly selected genes based on the raw data is displayed as a histogram.

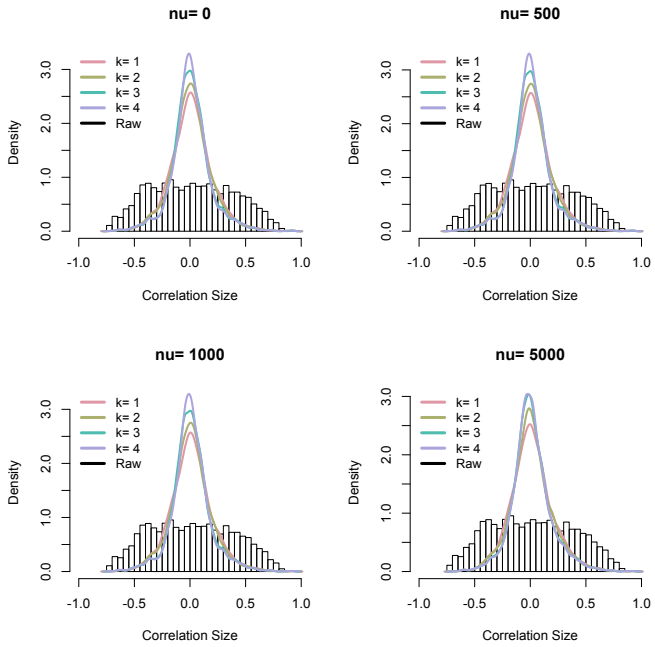


Figure S3: Correlation densities of the above parameter choices, but for a subset of randomly selected genes in the MRCA samples. Histograms of each panel estimate correlations density of a randomly selected subset of genes in the raw data.

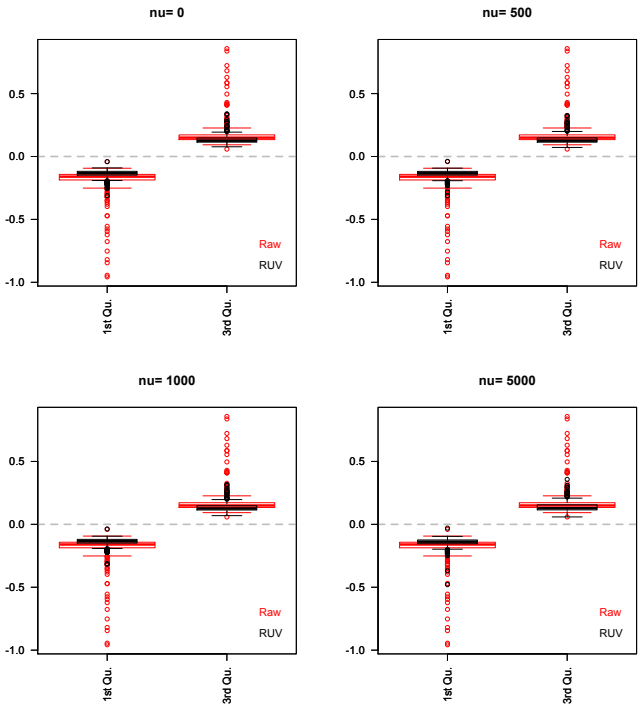


Figure S4: Relative log expression (RLE) plots comparing different values of $\nu$ for a given $k = 3$. Boxplots describe the 1st and 3rd quantile of the quantified gene expression vs study median spanning distance in the whole samples. Highlighted in red and black the raw data and RUV with $k = 4$ respectively.

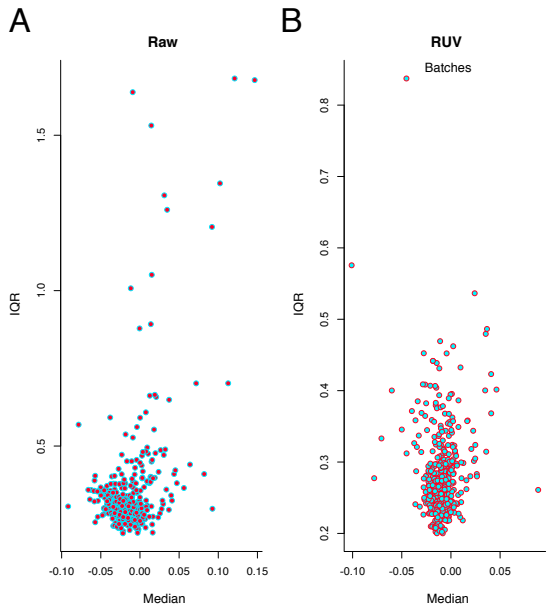


Figure S5: RLE plots to remove the variation of known source in MRCA gene expression profiles. RUV correction based on above statistics with parameters $\nu = 5000$ and $k = 4$. (A) RLE plot of denoised gene expressions before removing an empirical batch effect. (B) RLE plot normalised and batch-corrected of the MRCA panel. Median vs IQR based on inter-difference between MRCA gene expressions and their study mean.

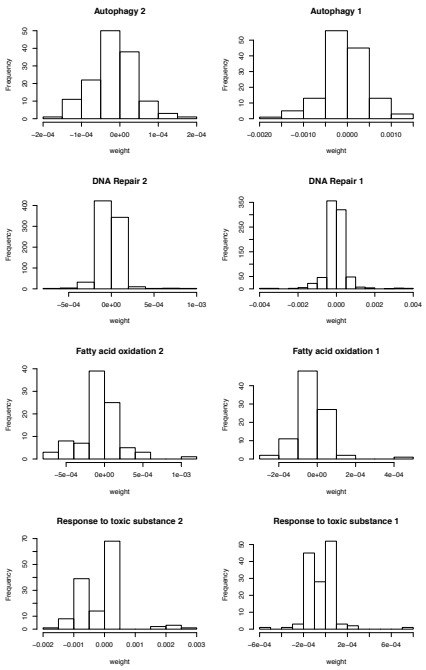


Figure S6: Genotype distribution per sample and GO functional module after cleaning process. Clockwise from top-left: samples type I, type II, Autophagy, DNA repair, Fatty acid oxidation, and Response to toxic substance.

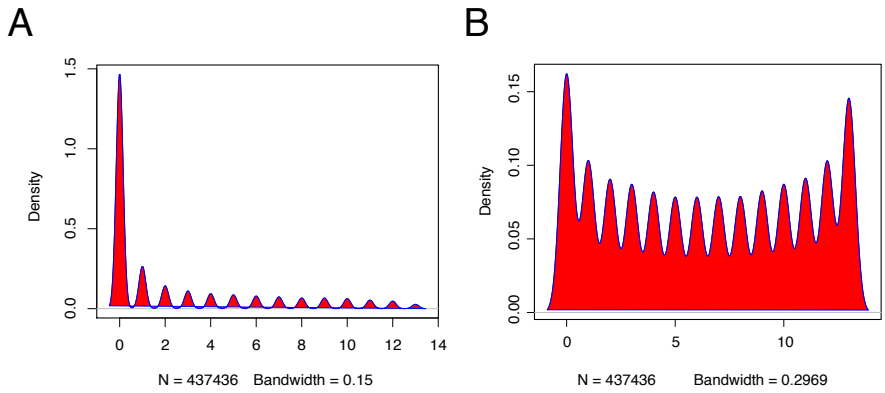


Figure S7: Genotypic variables abundance after counting samples in MRCA dataset in the e-QTL analyses. (A) Kernel density of frequency samples vs SNPs in presence of $A$. (B) Kernel density of frequency samples vs SNPs in presence of $a$.

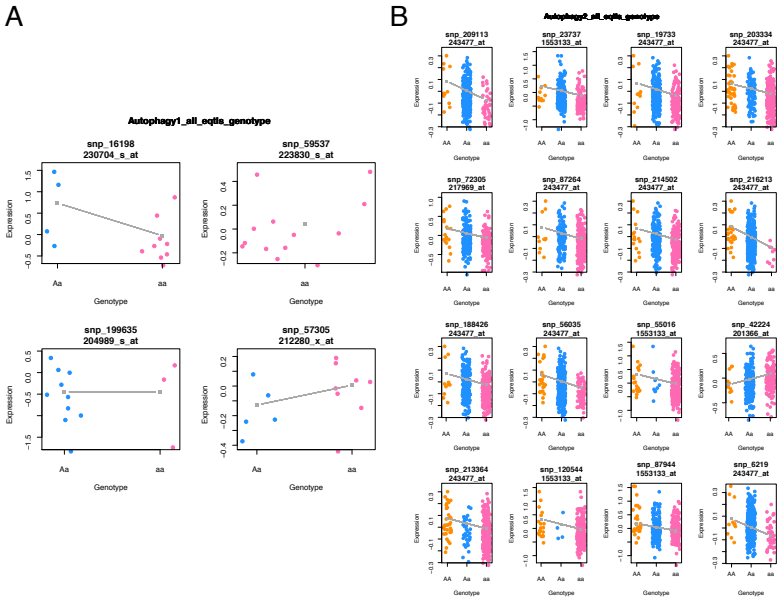


Figure S8: Stratification of individuals at low levels of expression. (A) e-QTL analysis of the class I community respect the autophagy functional module (B) Autophagy expression analysis of the samples class II. Colours grid: orange: $AA$ genotype; blue: $Aa$ genotype; pink: $aa$ genotype.

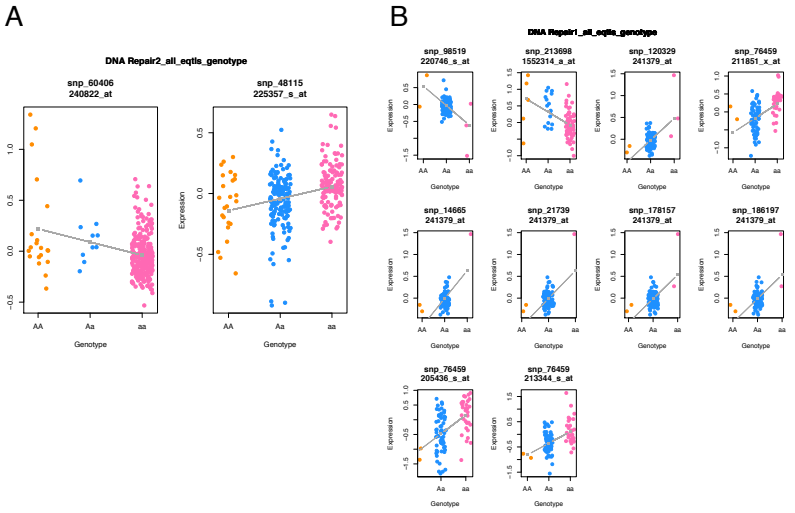


Figure S9: Stratification of individuals at low levels of expression. (A) e-QTL analysis of the class I community respect the DNA repair functional module (B) DNA repair expression analysis of the samples class II. Colours grid: orange: *AA* genotype; blue: *Aa* genotype; pink: *aa* genotype.

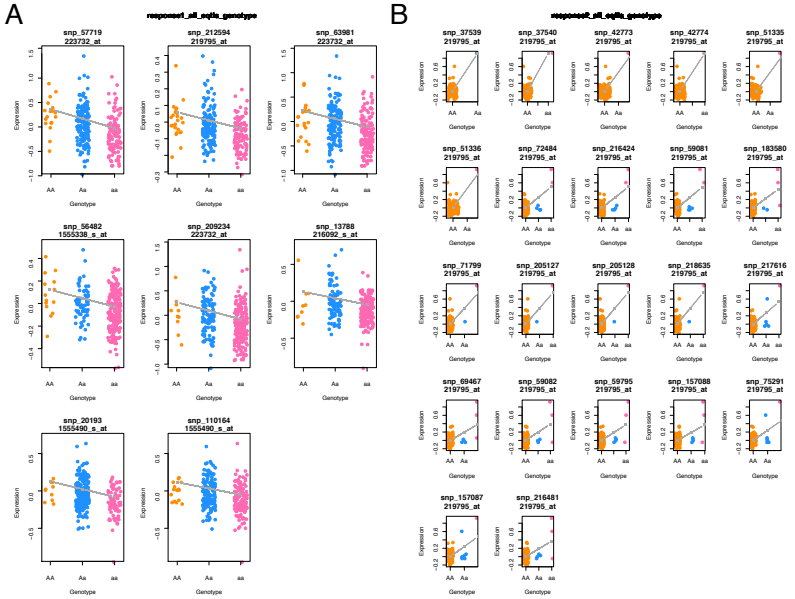


Figure S10: Stratification of individuals at low levels of expression. (A) e-QTL analysis of the class I community respect the response to toxic substance functional module (B) Response to toxic substance expression analysis of the samples class II. Colours grid: orange: *AA* genotype; blue: *Aa* genotype; pink: *aa* genotype.

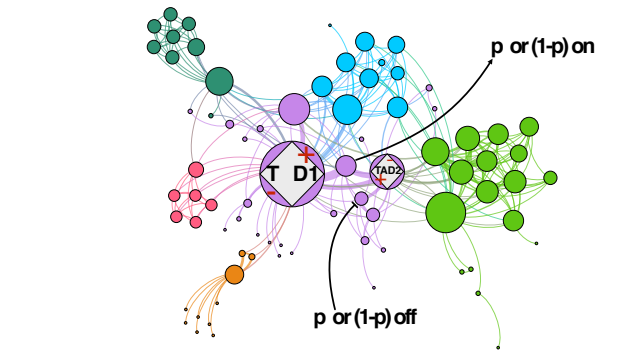


Figure S11: Graphical representation of our model. In violet an example of a community where $TAD1$ (i.e. TP53) and $TAD2$ (i.e. MDM2) may be correlated or anti-correlated with uncertainty $p$ or $1-p$ depending on the sense of the transcription factor. Such (anti)correlation must be imposed by other force than genotype and would make the module permissive or not to a disease start.

enough statistic to prevent count overestimations in presence of imbalanced gene distributions of the samples.

| Setups | TP53 | MDM2 | p21-up | p21-down | Bax-up | Bax-down |
|---|---|---|---|---|---|---|
| 1 | GC.CC | GT.TT | TRUE | TRUE | TRUE | TRUE |
| 2 | GC.CC | GT.TT | FALSE | TRUE | FALSE | FALSE |
| 3 | 0 0 | 0 0 | TRUE | TRUE | TRUE | TRUE |
| 4 | GC.CC | GG | TRUE | TRUE | FALSE | FALSE |
| 5 | GC.CC | GT.TT | FALSE | FALSE | FALSE | FALSE |
| 6 | GC.CC | GT.TT | FALSE | TRUE | TRUE | TRUE |

Figure S12: Samples counting after geometric mean correction. $(TP53, MDM2)$ model applied to the couple of gene variants $(Bax, p21)$. Hint: True amounts to presence and False to absence in the genotype.

# References


[1] S. Freytag (2020). *RUVcorr: Removal of unwanted variation for gene-gene correlations and related analysis*. Bioconductor, R package version 1.20.0.

[2] S. Freytag (2015). *Simulating and cleaning gene expression data using RUVcorr in the context of inferring gene co-expression*. Bioconductor, R package, version 1.0.1.

[3] S. Freytag, J. Gagnon-Bartsch, T.P. Speed & M. Bahlo (2015). *Systematic noise degrades gene co-expression signals but can be corrected. BMC Bioinformatics* 16, 309.