Biclustering analysis of co-regulation patterns in nuclear encoded mitochondrial genes and metabolic pathways

Running title: Biclustering analysis of mitochondrial genes

Robert B. Bentham[1]*, Kevin Bryson[2] and Gyorgy Szabadkai[1,3,4]

[1]Department of Cell and Developmental Biology, Consortium for Mitochondrial Research, University College London, WC1E 6BT London, UK

[2]Department of Computer Sciences, University College London, WC1E 6BT London, UK

[3]Department of Biomedical Sciences, University of Padua, 35131 Padua, Italy

[4]The Francis Crick Institute, 1 Midland Road, London NW1 1AT, United Kingdom

* To whom correspondence should be addressed. Tel:+44(0)2086798362 ; Email: robert.bentham.11@ucl.ac.uk

Abstract

Transcription of a large set of nuclear encoded genes underlies biogenesis of mitochondria, regulated by a complex network of transcription factors and co-regulators. A remarkable heterogeneity can be detected in the expression of these genes in different cell types and tissues, and the recent availability of large gene expression compendiums allows the quantification of specific mitochondrial biogenesis patterns. We have developed a method to effectively perform this task. Massively Correlated biclustering (MCbiclust) is a novel bioinformatics method that has been successfully applied to identify co-regulation patterns in large genesets, underlying essential cellular functions and determining cell types. The method has been recently evaluated and made available as a package in Bioconductor for R. One of the potential applications of the method is to compare expression of nuclear encoded mitochondrial genes, or larger sets of metabolism related genes between different cell types or cellular metabolic states. Here we describe the essential steps to use MCbiclust as a tool to investigate co-regulation of mitochondrial genes and metabolic pathways.

Keywords: biclustering, MCbiclust, mitochondria, metabolism, gene expression.

1. Introduction

Mammalian mitochondria are estimated to be composed of as many as 1500 genes (1) encoded in the nucleus along with the 13 protein-coding genes of the mitochondrial genome (mtDNA). To maintain proper mitochondrial function, the expression of the two genomes must be both co-ordinated and able to adapt to highly variable energetic demands. This results in a remarkable heterogeneity of mitochondrial composition, as detailed in numerous recent studies exploring the startling variety of mitochondrial function, physiology and proteome make-up across different tissues and cell types (2–4). Accordingly, the transcriptional regulation of mitochondrial biogenesis has been shown to be a highly complex process (see e.g. (5, 6)), involving numerous transcription factors and co-regulators, forming a complex interaction network, which is also highly adaptable via post-transcriptional modifications. While physiological regulation of mitochondrial biogenesis and composition vary enormously across healthy tissues, it is also known to contribute to major disease states. Mitochondrial dysfunction due to defects in the mitochondrial biogenesis pathway is known to be an important factor in cancer, neuromuscular degenerative disease and cardiomyopathies (7, 8). Whether these changes are the primary cause of the disease or the result of adaptation or maladaptation is an important open question in many cases. For these reasons bioinformatics tools to investigate the co-regulation of nuclear encoded mitochondrial genes not only have the potential to examine how physiological regulation works but also to reveal underlying factors that contribute to disease.

While the direct examination of the total mitochondrial proteome affected by the transcription factor network is often technically unfeasible, the availability of good quality, high coverage gene expression (microarray or RNAseq) data make it realistic to study the output of this network at the mRNA level. However, the success of this analysis relies on the ability of the applied methods to identify gene-sample ‘biclusters’ of similar mitochondrial co-regulation, since a single dataset often contain multiple modes of control in diverse mitochondrial gene groups. Here we discuss how a recently developed novel method MCbiclust (9) can be used for this task.

2. Materials

In the following sections we will refer to these software/manuals/datasets:

1. MCbiclust (doi:10.18129/B9.bioc.MCbiclust, current version 1.2.1), an R package available in Bioconductor (10), an open source platform for software in bioinformatics.
2. The MCbiclust package introductory manual (IM) accessed on the Bioconductor website (https://bioconductor.org/packages/release/bioc/vignettes/MCbiclust/inst/doc/MCbiclust_vignette.html).
3. The MCbiclust reference manual (RM) providing documentation to the R functions involved, accessed on the Bioconductor website (https://bioconductor.org/packages/release/bioc/manuals/MCbiclust/man/MCbiclust.pdf).
4. The MitoCarta 1.0 (2) mitochondrial geneset used in the IM.
5. The microarray dataset from the Cancer Cell Line Encyclopedia (11) also used in the IM.

3. Methods

In this section we will discuss in detail only the implications for applying the methodology to the analysis of mitochondrial biogenesis patterns. For complete understanding of the method, the theoretical considerations and benchmarking against other algorithms, please refer to Bentham et al. (9).

3.1 Choosing a geneset

Once a dataset has been chosen (for details on choosing your dataset and judging whether it is suitable see Note 1), the first step of using MCbiclust is to select a suitable geneset representing mitochondrial function with the scope of discovering co-regulation patterns in nuclear encoded mitochondrial genes. This is not a trivial problem as there are genes with different confidence levels of evidence relating them to mitochondria, as well as genes that while not being mitochondrial are highly co-regulated with mitochondrial processes. We consider two alternative methods for geneset selection.

3.1.1. Established databases with mitochondrial genesets

i) MitoCarta (12) in its latest version (2.0) contains 1158 human and mouse genes with strong support of mitochondrial localization.
ii) MitoMiner 4.0 (13) is an integrated web resource of mitochondrial localisation evidence and phenotype data for mammals, zebrafish and yeast. The team behind MitoMiner developed the Integrated Mitochondrial Protein Index (IMPI), which in its current version (Q3 2017) includes 1550 genes.
iii) Genes associated with the Gene Ontology (14) term "mitochondrion", which contains 1647 genes; genes in the dataset, however, have varying evidence with many inferred from in silico analysis.

The user can decide whether to use one of these data sets in order to select the mitochondrial genes to be analysed. Alternatively, the intersection (985 genes) or union of all three datasets (1997 genes) could be used. The size of the geneset is an important factor for determining the speed at which MCbiclust completes the analysis. However, an increased geneset size does not necessarily bring any benefits (see Note 2).

3.1.2 Interaction networks of mitochondrial genes

An alternative strategy to using public lists of known or predicted mitochondrial genes is to compose a list by using a single well established mitochondrial gene and determine its interactions from the existing correlation structure in the dataset of interest. By taking a single, well established mitochondrial gene, e.g. a component of the electron transport chain or the mitochondrial ribosome, the remaining genes can be ordered by the strength of the Pearson's correlation coefficient to the expression of this gene across all of the samples (see Note 3 for details). The geneset, used by MCbiclust to initiate the analysis, can then be selected as the top genes correlated with the mitochondrial gene of interest. The advantages of this method are that (i) it is more likely to include genes that are strongly co-regulated with mitochondrial processes, thus representing a specific function; (ii) it is more likely to identify biclusters that are associated with a single mitochondrial gene of interest; and (iii) the geneset can be specifically tailored for each dataset. The disadvantage of using this strategy is that the geneset will differ in each user case, thus comparison of results will become more complex or even unfeasible.
Overall, there is no 'correct' way to choose a geneset, and the appropriate way should be decided on a case by case basis, according to the precise biclusters that are being sought.

Nor should an investigator be limited to running a single geneset as the results of MCbiclust using multiple genesets can be compared (see Note 4).

3.2 Running MCbiclust to identify co-regulation of nuclear encoded mitochondrial genes

Following the selection and loading the sample set and initial geneset(s) (IM 3.1), `FindSeed` is used to identify a 'seed' of samples with high Pearson correlations between the genes in the geneset (IM 3.2; 3.3). Importantly, this method is stochastic and identifies the samples by a greedy search. Thus, in order to find an exhaustive and representative sample set, it is required to run `FindSeed` multiple times. The different strategies to perform this task are discussed in Note 5.

Multiple runs of `FindSeed` result in a number of sample seeds. Once a suitable number of sample seeds have been found, the next step is to identify how many distinct modes of regulation of the geneset have been found, i.e. which samples are included, and how genes are correlated in these sample seeds. Clearly, if the samples are identical in different seeds, they represent the same pattern, but it is not clear if different samples between seeds represent fundamental differences in regulation or the seed has selected different samples that are representative of the same pattern. For this reason, the different outputs of MCbiclust must be compared at the gene-level using a parameter that is called the correlation vector (CV, see IM 3.4). The CV is a vector that quantifies the correlation of each gene measured in the dataset to the average expression of a group of genes in the chosen geneset that are selected as 'highly representative' of the bicluster. The CV for each run can then be compared to one another, after which the runs are clustered and then the Silhouette method (15) is used to identify the number of distinct biclusters found in the analysis (IM 4. and RM: SilhouetteClustGroups). The CVs can be averaged across each distinct bicluster and consequently the samples can be ranked by how well they preserve the correlation within the geneset. The final output of MCbiclust for each bicluster found is a correlation vector describing the strength of the correlation of each gene to the bicluster and an ordered list of samples (IM 4). Accordingly, the biclusters can be visualised with a distinctive 'Fork plot' with the ranked samples on the x-axis plotted against the PC1 value from a PCA analysis of the samples within the seed, with the PC1 value being fitted to the remaining samples (IM 3.10). At the beginning of the ranking the samples separate into an

upper and lower fork. By convention, the sign of the PC1 value being chosen is such that the upper fork samples will have genes with a positive correlation vector that are up-regulated and genes with a negative correlation vector value that are down-regulated. The lower fork samples have the opposite phenotype.

3.3 Analysing the resulting biclusters

The analysis of the resulting biclusters involves the separate analysis of genes and samples. Sample analysis is dataset-specific and involves associating samples in the distribution plot with the different properties (metadata) of sample groups made available for the dataset (for previous examples, see Figs. 5, 6, 7 and 8 from Bentham et al. (9)). In patient derived gene expression samples, this typically includes clinical outcome, genetic and histological subtypes of the disease. Thus, biclusters are the basis of stratification, that is, classification of disease states according to mitochondrial gene expression patterns.

On the other hand, the methods for the analysis of genes can be generalised for different biological applications and are listed below.

3.3.1 Geneset enrichment analysis

The simplest analysis is a geneset enrichment analysis on the values of the correlation vector (IM 3.5). The correlation vector can be viewed as a ranked list of genes with values between +1 and -1, and thus geneset enrichment analysis can be run on the entire ranked list, or on selected genesets, e.g. the top positive or negative correlation vector values. At this point, any geneset enrichment method can be used (e.g. DAVID (16), GSEA(17), gProfiler(18)). The MCbiclust package comes with a specifically designed method that uses the entire correlation vector and applies the Mann-Whitney test to identify gene ontology terms that have significantly different distributions (either more positive or negative) as compared to the entire distribution of values. The output gives the average CV value for each significant term, thus terms that are positive in average (i.e. up-regulated in the upper fork, down-regulated in the lower fork) can be distinguished from those that are negative in average (i.e. down-regulated in the upper fork and up-regulated in the lower fork). Interpretation of the significant terms can be challenging, since standard terms often give no other detail than the list of genes that are generally related to ‘mitochondria’. For fine grain understanding of the differences in pathways, the individual genes involved must be examined. Different mitochondrial pathways of interest, such as the metabolic enzymes,

can each be examined individually. For these metabolic pathways, it is also possible to build diagrams of the pathways to show which parts have been regulated in different ways, e.g. with the pathview R package (19). On the other hand, geneset enrichment analysis can be useful for identifying non-mitochondrial pathways that are also being simultaneously co-regulated with mitochondria, providing further insight into the biology behind the underlying process.

3.3.2 Comparison of genesets across biclusters

In cases where two or more biclusters are found, it is appropriate to compare the differences in the co-regulation of the genes in the biclusters. In order to identify a module of genes that are regulated in the same way across different biclusters, different visualisation techniques can be applied. First, co-regulation of genes in different biclusters can be compared using the CVplot function in MCbiclust (IM 4, RM: CVplot). This function plots the values of the correlation vectors against each other for all the genes, as well as genes in any chosen geneset (e.g. mitochondrial genes). In this way, modules of co-regulated genes across different biclusters can be identified. Alternatively, these groups can be identified through examining the intersection of genesets (e.g. up-regulated in bicluster 1, up-regulated in bicluster 2, etc.), using Venn diagrams for a small number of groups. If the number of different biclusters is large, a different technique such as UpSet plot (18) can also be used. Examples of these visualisation techniques are shown in **Figure 1**.

3.4 Identification of samples in other datasets matching the bicluster.

Once a bicluster has been identified and associated with a particular type of mitochondrial function, a further aim is to determine whether this type of gene expression pattern can be identified in additional data sets. Theoretically this could be achieved by running the entire MCbiclust pipeline on this new dataset and comparing the resulting correlation vectors to understand whether a similar bicluster is present. However, this approach might be time consuming and often datasets are not large enough for MCbiclust to reliably identify biclusters (see discussion on the required dataset size in Bentham et al. (9)). Thus, ideally a method is required that can take a small dataset or single sample and determine whether these samples fall into a particular bicluster and whether they belong to a particular branch in the fork distribution.

3.4.1 Point score algorithm

A method of choice included in the MCbiclust package to achieve the classification of single samples is the PointScore algorithm (RM: PointScore). This method uses the two genesets (A and B) determining the distribution of samples in the fork pattern (see Note 6 for how these genesets are chosen). Geneset A includes genes up-regulated in the upper fork and down-regulated in the lower fork, and geneset B contains genes down-regulated in the upper fork and up-regulated in the lower fork. 'PointScore' scores samples based on how well they match this regulation by comparing the genes in the genesets to the median value across the entire dataset. Importantly, this method requires that the dataset contains samples that are representative of all types of regulation seen in the original dataset (where the bicluster was identified), so that the median of the genes can be used as a dividing line for resolving up- or down-regulation in samples. For this reason, the PointScore algorithm cannot be used for single or too few samples.

For single samples or datasets with very few samples there are two further solutions detailed in sub-sections 3.4.2 and 3.4.3.

3.4.2. Single sample GSEA (ssGSEA)

Single sample GSEA (20), from the Bioconductor package GSVA (21), can be applied by taking the same genesets as used in the PointScore method and calculating the ssGSEA score, based on how the genes in each geneset are up or down regulated, compared to other genes in the samples. Therefore, for an upper fork sample, the ssGSEA score for geneset A will be positive and the score for geneset B will be negative.

3.4.3. First principal component values

It is possible to calculate the PC1 value of the sample (using the R function lsfit from the known PC1 loadings), and compare it directly to the initial bicluster. This technique requires that this sample (or small dataset) is normalised to the original dataset. This is only reliable when the datasets are all measured on the same platform, quantile-normalisation is performed and any possible batch effects are removed between experiments (for example by using ComBat (22)).

4. Notes

1. As a method MCbiclust is agnostic towards the data platform and can be run on both microarray and RNASeq data. However, for a successful run, the data must meet one

important requirement, that the dataset contains enough samples. As a rule of thumb, at an absolute minimum there should be at least 100 samples in the sample set. In general, the more samples are in a data set, the more likely MCbiclust is able to find significant biclusters. If the dataset contains few samples, it can be analysed by comparing to previously analysed larger sets, as described in 3.4.

2. MCbiclust calculates the correlation matrices of the chosen geneset repeatedly. Thus, the larger the geneset chosen, the more computation time is needed to perform MCbiclust. In general, a geneset containing more than 1000 genes is sub-optimal and significantly slows down the computation. There is also little advantage to augment the size of the geneset past a certain point, since the biclusters we seek to find are large; as long as a significant number of genes in the geneset are contained in them, they will be found. Additionally, genes outside the geneset can easily be found to be associated with the bicluster in the correlation vector stage of the method (see section 3.2). Thus in general, there is no need for genesets significantly larger than 1000.
3. This can be achieved simply using base functions such as the apply and cor function in R e.g. vec1 <- apply(data, MARGIN = 1, FUN = function(x) cor(x, as.numeric(data[gene.loc,])) and then selecting the genes that have the highest correlation, e.g. hicor.loc <- order(abs(vec1), decreasing = TRUE)[seq_len(1000)].
4. Since the choice of the initial geneset is an important factor in determining the results of MCbiclust, running MCbiclust on different initial genesets, e.g. a general mitochondrial one from MitoCarta, as well as various different genesets made up of genes that are strongly correlated with mitochondrial genes of interest is a good and recommended strategy.
5. FindSeed should be run enough times to identify all significant biclusters present in the dataset. Typically, this number should be at least 100. However, some biclusters are only identified rarely by random search, and to find these, it is necessary to run FindSeed a very large number of times. In these cases, it is of help to use high performance computing to run the FindSeed algorithm. An alternative way to find these rare biclusters is to run FindSeed on different initial genesets or run FindSeed on the dataset after removing the most commonly selected samples in the final seed. This way the final seed is forced to include samples not yet chosen.

6. Genesets that represent the upper and lower fork can be created directly from the correlation vector selecting genes with a value greater than a certain threshold e.g. > 0.9 for upper fork and < -0.9 for lower fork.

Acknowledgements

Funding was provided by University College London COMPLeX/British Heart Foundation Fund (SP/08/004), the Biochemical and Biophysical Research Council (BB/L020874/1, BB/P018726/1), the Wellcome Trust (097815/Z/11/Z) in the UK, and the Association for Cancer Research (AIRC, IG13447) in Italy.

References

1. Lopez,M.F., Kristal,B.S., Chernokalskaya,E., Lazarev,A., Shestopalov,A.I., Bogdanova,A. and Robinson,M. (2000) High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis*, **21**, 3427–3440.
2. Pagliarini,D.J., Calvo,S.E., Chang,B., Sheth,S.A., Vafai,S.B., Ong,S.-E., Walford,G.A., Sugiana,C., Boneh,A., Chen,W.K., *et al.* (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, **134**, 112–23.
3. Thor Johnson,D., Harris,R.A., French,S., Blair,P. V, You,J., Bemis,K.G., Wang,M. and Balaban,R.S. (2007) Tissue heterogeneity of the mammalian mitochondrial proteome. *Am J Physiol Cell Physiol*, **292**, 689–697.
4. Kuznetsov,A. V., Hermann,M., Saks,V., Hengster,P. and Margreiter,R. (2009) The cell-type specificity of mitochondrial dynamics. *Int. J. Biochem. Cell Biol.*, **41**, 1928–1939.
5. Scarpulla,R.C. (2008) Transcriptional paradigms in mammalian mitochondrial biogenesis and function. *Physiol Rev*, **88**, 611–638.
6. Hock,M.B. and Kralli,A. (2009) Transcriptional Control of Mitochondrial Biogenesis and Function. *Annu. Rev. Physiol.*, **71**, 177–203.
7. Duchen,M.R. and Szabadkai,G. (2010) Roles of mitochondria in human disease: Figure 1. *Essays Biochem.*, **47**, 115–137.
8. Jones,A.W.E., Yao,Z., Vicencio,J.M., Karkucinska-Wieckowska,A. and Szabadkai,G. (2012) PGC-1 family coactivators and cell fate: roles in cancer, neurodegeneration,

cardiovascular disease and retrograde mitochondria-nucleus signalling. *Mitochondrion*, **12**, 86–99.

9. Bentham,R.B., Bryson,K. and Szabadkai,G. (2017) MCbiclust: a novel algorithm to discover large-scale functionally related genesets from massive transcriptomics data collections. *Nucleic Acids Res.*, **45**, 8712–8730.
10. Gentleman,R., Carey,V., Bates,D., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J., *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
11. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G. V, Sonkin,D., *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
12. Calvo,S.E., Clauser,K.R. and Mootha,V.K. (2015) MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.*, 10.1093/nar/gkv1003.
13. Smith,A.C., Blackshaw,J. a and Robinson,A.J. (2012) MitoMiner: a data warehouse for mitochondrial proteomics data. *Nucleic Acids Res.*, **40**, D1160-7.
14. Consortium,T.G.O. (2000) Gene ontology: Tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
15. Rousseeuw,P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
16. Huang,D.W., Lempicki,R. a and Sherman,B.T. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
17. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Geneset enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.
18. Reimand,J., Arak,T., Adler,P., Kolberg,L., Reisberg,S., Peterson,H. and Vilo,J. (2016) g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.*, **44**, W83–W89.
19. Luo,W. and Brouwer,C. (2013) Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, **29**, 1830–1831.
20. Barbie,D.A., Tamayo,P., Boehm,J.S., Kim,S.Y., Susan,E., Dunn,I.F., Schinzel,A.C., Sandy,P., Meylan,E., Fröhling,S., *et al.* (2010) Systematic RNA interference reveals that oncogenic

KRAS- driven cancers require TBK1. *Nature*, **462**, 108–112.

21. Hanzelmann,S., Castelo,R. and Guinney,J. (2013) GSVA: geneset variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, **14**, 7.
22. Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Legend to Figures

Figure 1. Comparison of genesets across biclusters using CVplot and Upset plots.
Plots produced from a RNA-Seq dataset from CoMMpass (Relating Clinical Outcomes in MM to Personal Assessment of Genetic Profile) IA9 study (NCT01454297) produced by the Multiple Myeloma Research Foundation (MMRF) containing transcriptomics from 734 patient samples. **A.** shows an output of CVplot comparing the correlation vectors from three different runs of MCbiclust with mitochondrial genes from Mitocarta (Mito), a gene set based on the most correlated genes to mitochondrial gene MRPL58 (ICT1) and random (Rand) gene sets. The lower diagonal plots (cyan) represent values of the non-mitochondrial genes in the correlation vector while the upper diagonal plots (red) represent the mitochondrial genes in the correlation vector. In this case a very similar bicluster (in terms of the genes which are most strongly correlated to it) is found from all three initial gene sets used. Plots in the diagonal axis show the frequency distribution of mitochondrial (red traces) and non-mitochondrial (cyan traces) genes across the correlation values in the three biclusters. **B.** shows the output from the UpSet R package to determine the intersections of the significant genes identified in each of these correlation vectors from MCbiclust's custom gene set enrichment method (see 3.3.1). The significant gene sets found in each bicluster have been split into two groups (pos and neg) depending on whether they are associated with genes with positive or negative correlation vector values. The majority of significant terms are shared between all three biclusters, again indicating that these the three biclusters are close to identical.

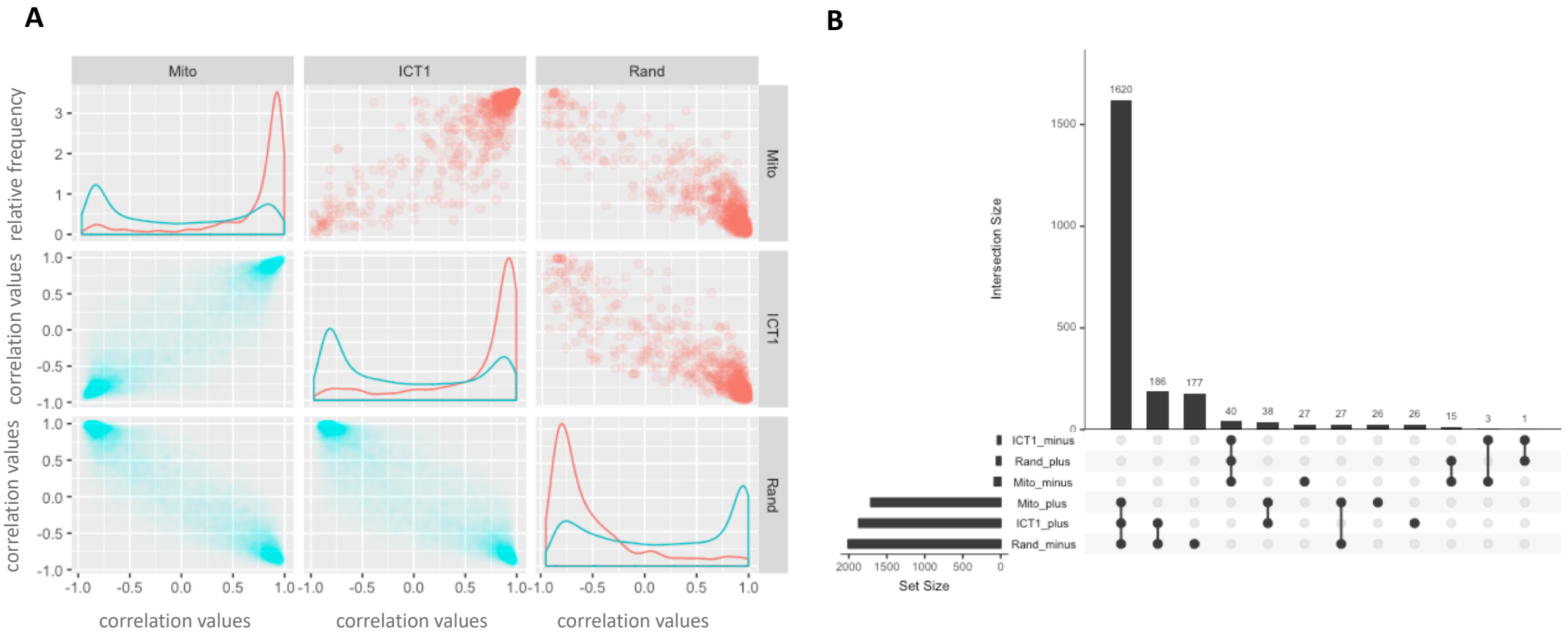
A
Mito
ICT1
Rand
relative frequency
correlation values
correlation values
correlation values
correlation values
correlation values
correlation values
B
Intersection Size
1620
186
177
40
38
27
27
26
26
15
3
1
ICT1_minus
Rand_plus
Mito_minus
Mito_plus
ICT1_plus
Rand_minus
Set Size

Figure 1.