

Article

Entropic Ranks: A Methodology for Enhanced, Threshold-Free, Information-Rich Data Partition and Interpretation

Hector - Xavier de Lastic ^{1,2}, Irene Liampa ^{2,3}, Alexandros G. Georgakilas ¹, Michalis Zervakis ³, Aristotelis Chatziioannou ^{4,5,*}

¹ DNA Damage Laboratory, Physics Department, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Athens, 15780, Greece

² Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens 11635, Greece

³ Digital Image and Signal Processing Laboratory (DISPLAY), School of Electrical and Computer Engineering, Technical University of Crete, Chania 73100, Greece

⁴ e-NIOS PC, Kallithea-Athens, 17671, Greece

⁵ Center of Systems Biology, Biomedical Research Foundation of the Academy of Athens (BRFAA), 11527, Athens, Greece

* Correspondence: achatzi@bioacademy.gr

Featured Application: The generic applicability of the entropy-empowered RP calculation score supports the utilization of this non-parametric, threshold-free methodology in different kinds of data. This is not restricted only in meta-analysis of different data sets [1], but could serve as a key methodology for data integration of different sources of information, in the quest for highly automated, systemic Big Data biological interpretation.

Abstract: Background: Traditional omic analysis relies on p-value and fold change as selection criteria. There is an ongoing debate on their effectiveness in delivering systemic and robust interpretation, due to their dependence on assumptions of conformity with various parametric distributions. Here, we propose a threshold-free selection method based on robust, non-parametric statistics, ensuring independence from the statistical distribution properties and broad applicability. Such methods could adapt to different initial data distributions, contrary to statistical techniques based on fixed thresholds.

Methods: Our work extends the Rank Products methodology with a neutral selection method of high information-extraction capacity. We introduce the calculation of the RP distribution's entropy to isolate the features of interest by their contribution to the distribution's information content. The aim is a methodology performing threshold-free identification of the differentially expressed features, which are highly informative about the phenomenon under scrutiny.

Conclusions: Applying the proposed method on microarray (transcriptomic and DNA methylation) and RNAseq count data of varying sizes and noise presence, we observe robust convergence for the different parameterisations to stable cutoff points. Functional analysis through BioInfoMiner and EnrichR was used to evaluate the information potency of the resulting feature lists. Overall, the derived functional terms provide a systemic description highly compatible with the results of traditional statistical hypothesis testing techniques. The methodology behaves consistently across different data types. The feature lists are compact and information-rich, indicating phenotypic aspects specific to the tissue and biological phenomenon investigated. Selection by information content measures efficiently addresses problems, emerging from arbitrary thresholding, thus facilitating the full automation of the analysis.

Keywords: data analysis; threshold-free; differential analysis

1. Introduction

Data analysis of high-throughput technologies (microarrays, next generation sequencing) commonly predicates on the adoption of arbitrary p-value and fold change thresholds to define the reliability and relevance of a set of features, in order to partition the initial distribution into two sets. The first set (S) is further investigated for its phenotypic relevance, whereas the other is exempted from further analysis, considered to be the baseline distribution with noise, either biological (coexisting, causally unrelated processes) or technical [2]. This approach as a selection philosophy is currently being debated. Specifically, regarding p-value thresholding, critics raise the issues of incomplete information, misrepresentation, misinterpretation and bias [3,4], whereas for fold change thresholding, the issues cited include the adoption of arbitrary thresholds [1], consequent the potential of strong bias [5], with no theoretical underpinning for the threshold values. Ideally, selection thresholds should take into account the form of data distributions, the presence of confounders, and the complexity of phenomenon under investigation.

1.1. From Rank Products to Entropic Ranks

The rank product method partially addresses these issues through a frequentist approach, by measuring the consistency of behaviour across the sample groups, using the fold change (FC) criterion. When testing for upregulation, the Rank Product (RP_g^{up}) of a gene (g) is calculated as

$$RP_g^{up} = \prod_{i=1}^k (r_{i,g}^{up} N_i) \quad (1)$$

where N_i is the total number of features and $r_{i,g}^{up}$ (for single-colour transcriptomics) is the rank of gene g in the decreasing-FC-ordered list of genes in the i th pair of control VS case samples (i.e. $r^{up} = 1$ for a gene consistently more overexpressed than any other in all cases VS all control samples). Similarly calculated is the RP_g^{down} , over the increasing-FC-ordered lists. The percentage of false positive (pfp) value for each Rank Product (RP) score is estimated through a permutation-based procedure outlined in the initial publication of the method [1].

The initial Rank Product implementation in Bioinformatics, whilst addressing a number of the aforementioned issues, retains arbitrariness by trimming the final list either by a calculated pfp/p-value thresholding or by directly choosing the number of genes. Moreover, it tends to behave over-optimistically with increasing data dimensionality, as shown in recent works [6] (see also "Supplementary Material 1 - Method").

Our work aims to measure the information content of the RP distribution, implement a data-driven, non-parametric partitioning, provide a workflow for high-throughput data analysis and unbiased information extraction (Figure 1), and generalize the approach to various data types. We introduce the calculation of entropy over a transformation of the RP distribution, followed by a clustering procedure in order to identify the most consistent cutoff point successfully separating information-rich features from noise-dominated ones, without human intervention. The methodology operates upon the non-parametric RP distribution, relegating pfp and p-value (which may be calculated empirically or parametrically) to quality indicators of the analysis instead of decision criteria. Consequently, it allows for improvement of the pfp calculation methods and adoption of new approaches [6,7], ensuring result reproducibility and comparability as long as the RP calculation process itself remains unchanged.

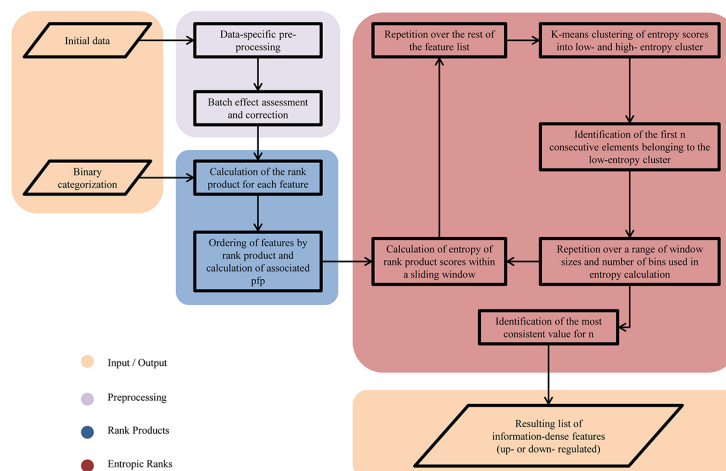


Figure 1. Entropic Ranks workflow. Entropic Ranks builds upon the Rank Products methodology and assesses its results in terms of the progression of Shannon entropy over a sliding window.

1.2. Entropy in Biostatistics

The utilization of Entropy in bioinformatics has been tentative, for the most part, leading to specific implementations instead of producing consistent classes of methodologies, governed by standard practices. Nevertheless, we identify three main categories of such implementations:

Firstly, the usage of entropy as a measure for the optimization of classification techniques. In most implementations, this is achieved prior to the classification, through dimensionality reduction approaches driven by the evaluation of entropy in order to reduce feature redundancy [8–10]. However, there is a recently introduced approach [11] in which entropy is used to directly weight and rank FRFT coefficients, upon which further clustering approach is based.

Secondly, the evaluation of entropy has been used as a main component of the decision process. Various approaches include patient stratification [12], inferring regulatory networks using transfer entropy [13] and identification of periodical biological processes in time series data [14]. However, there exists an approach more closely based on Shannon entropy and conceptually closer to our work than others, which attempts to identify differential expression in RNA count data [15].

Lastly, a unique implementation utilizes entropy evaluation upon the variability of genome regions [16] in order to evaluate their information content in contrast to uniformity, a contrast we also use during the partitioning of the Rank Product distribution in the proposed methodology.

2. Methods

2.1. Rank Product Requirements

Rank product methodology is applied upon four basic premises [1] considered valid in high-throughput signal distributions:

- that $S \ll N$ (N : the full set of features)
- independence of measurements between replicate arrays
- the intensity of each feature over the range of samples is largely homoscedastic
- the majority of non-zero fold changes between the sample groups are independent of each other

2.2. The segmentation problem

Our working hypothesis is a direct corollary of the ordered RP distribution being an ordered set; namely, that the first n elements of the RP distribution correspond to features of high information content in respect to the phenomenon under scrutiny, and subsequent ones are to be excluded from

further analysis. Our partitioning process is functionally identical to a threshold-driven usage of Rank Products, using optimal pfp thresholds, chosen individually for each experiment (see Table 3 and "Supplementary material 1 - Method"). The generalized approach we aim to create should automatically adapt to each experiment, consistently separating signal from noise while eschewing the bias introduced by thresholding approaches.

2.3. Partitioning the RP distribution

The RP distribution (i.e for upregulated features) starts with a steep incline, converging to a linear distribution for features following the null hypothesis and finally diverging again for the last few features (i.e downregulated) (see "Supplementary material 1 - Method"). The set of features most significant in describing the phenomenon investigated consists of the first n elements of that distribution, encompassing at least part of the steep incline. Conversely, the linear part of the distribution corresponds to stochastically behaving features, which should be excluded. Moreover, the RP score distribution carries information regarding each feature's consistency of behaviour across replicates and is rigidly determined by the structure and form of the original data, deflecting external computational bias (in contrast to pfp calculation). Consequently, n will be robustly determined by partitioning the RP distribution.

2.4. Entropy of differences

Given that the RP distribution is ordered, we can calculate the distribution of the differences between consecutive terms, shown in the lower part of Figure 2. Differences stochastically oscillate around a descending trend, converging near zero. We observe that the area corresponding to the first elements highlighted by rank product calculation is dominated by the descending distribution. Conversely, the area of features to be excluded from analysis corresponds to an area dominated by stochastic oscillations. Given that RP scores reward systematic behaviour of features across replicates (and thus convey information on it), these two different patterns can be considered to represent the difference between systematic behaviour, attributable to differentiated processes, and noise-dominated behaviour following the null Rank Product hypothesis. Furthermore, the calculation of entropy can partition this distribution into an initial, low-entropy and information-rich area and a high-entropy, information-poor subsequent area. The former will be considered to be the set S , containing n features. The latter will be considered the set of features with a low signal-to-noise ratio, excluded from further analysis.

2.5. Partitioning the entropy distribution

Using a sliding window over the distribution of differences, we calculate the entropy of values within in nats, using the Shannon entropy:

$$H = - \sum_{k=1}^p \theta_k \log(\theta_k) \quad (2)$$

as modified by a Dirichlet-multinomial regularization resulting in

$$\hat{H}^{Bayes} = - \sum_{k=1}^p \hat{\theta}_k^{Bayes} \log(\hat{\theta}_k^{Bayes}) \quad (3)$$

where $\hat{\theta}_k^{Bayes} = y_k + a_k n + A$, $A = \sum_{k=1}^p a_k$ and $a_k = 1$,

as per the Bayes-Laplace uniform prior, reflecting our prior knowledge for the bin counts used during discretization [17]. The upper part of Figure 2 shows the resulting entropy score distribution for the GSE60767 gene expression data set. The initial, low-entropy area represents S . Subsequently, entropy rises sharply and oscillates around higher values, due to (small in absolute value) stochastic oscillations of the RP differences. This pattern holds across different parameters of the sliding window

and bin count, as well as across different data types, with consistency analogous to that of the RP distribution's overall shape.

We note that there may be numerous low-entropy areas over the entropy distribution. The RP distribution is an ordered set, with ordering corresponding to decreasingly statistically significant and consistent differential behaviour under the fold change criterion. Thus, our hypothesis is that the very first low-entropy area corresponds to features with consistently differentiated behaviour between the two populations, hence more strongly tied to the phenomenon under study compared to subsequent areas. Contrarily, the high-entropy areas (including local minima which could result from random oscillations in low-signal areas as seen in the right part of the distribution) correspond to features associated with the Rank Products null hypothesis (achieving ranks with a uniform probability distribution [18]), thus not informative on the phenomenon. Consequently, we now need to consistently identify the first low-entropy area in an unsupervised manner. This segmentation should be achieved while eschewing the need for user-defined arbitrary thresholds, thus avoiding the re-introduction of bias inherent in thresholding approaches. The finely detailed structure of the entropy distribution prohibits stable convergence of change-point algorithms, which would be the first approach to partitioning ordered distributions. They exhibit high variation of performance, wholly reliant on the specifics of the sliding window and bin count. This class of algorithms also fails to converge reliably regarding the RP distribution. See also "Supplementary Material 1 - Method".

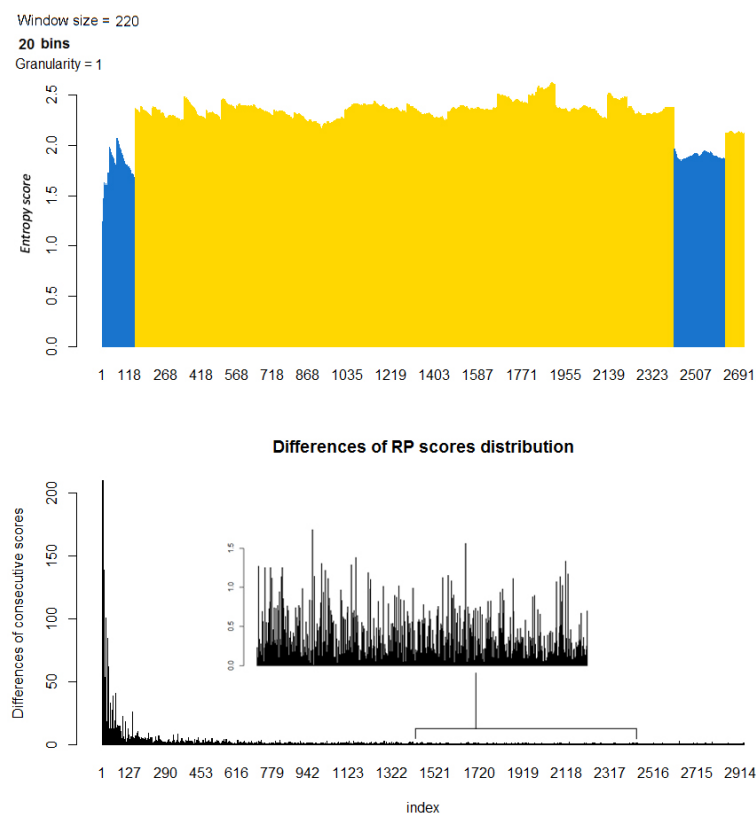


Figure 2. Upper graph: A characteristic entropy distribution for GSE60767 data, Winter 2009, upregulated features. Parametres used were: window size 220, 20 bins for discretization and advancement of the sliding window gene-by-gene. The initial, low-entropy set is clearly defined (initial blue-colored area). The low-entropy area near the end is attributed to a single difference value oscillation, lowering the entropy of the 220 windows containing it.

Lower graph: The corresponding distribution of consecutive RP score differences. The initial, signal-dominated area always overlaps with genes selected by the original Rank Products. The latter, noise-dominated area represents features to be excluded.

2.6. K-means clustering to calculate n

To reliably overcome the partitioning problem of the RP distribution, we evaluated the performance of a selection of clustering methodologies, aiming to categorize the entropy values into a high-entropy and a low-entropy cluster. Maximum consistency against outliers and incidental complexity was exhibited by the K-means algorithm. Direct corollary of our hypotheses so far is that if the low-entropy cluster starts at the beginning of the RP score list (which is an ordered set), then there is at least a single, highly informative, differentially behaving feature. If so, n will be defined as the number of features preceding the first element of the high-entropy cluster (see the upper part of Figure 2, leftmost blue area). The rationale is that the first high-entropy value will correspond to the first sliding window containing no information-rich (in terms of behavioural pattern relevant to the comparison performed) features. Due to the ordering of the RP distribution, all features following a rejection are to be rejected as well. This procedure proves highly resistant to perturbations concerning the specifics of the sliding window and the number of bins used in entropy calculation. Reiterating this procedure over a range of window sizes and bin counts yields a small set of suggested values for n , one of which exhibits prominent consistency. This will mark the cutoff defining the set of highly informative features, forgoing the need for external thresholds (to pfp or even entropy values), thereby decoupling the selection process from calculated statistical scores, relegating their use to post-selection result assessment.

2.7. Implementation

The development, actualization, testing, and verification through analyses were all performed in R v3.4.1, using open source R packages from Bioconductor in RStudio and usegalaxy.eu, in order to ensure transparency and reproducibility. Full citation of the packages is offered in "Supplementary material 1 - Method". In the interests of reproducibility and platform independence afforded by a dockerized implementation, the dockerfile for the tool is hosted on https://github.com/Hector-Xavier/Entropic_Ranks_docker

3. Results

3.1. Evaluation criteria

A threshold-free, adaptive, generalized selection process, like the one proposed, should be evaluated according to the following criteria: (a) specificity of the selected features in terms of biological relevance, (b) sensitivity to weak biological signals, (c) performance on data sets of varying noise content, and (d) its generality in terms of reliable performance across different types of experiments. To test against these criteria, we selected and analyzed a range of published, publicly accessible data sets, each tied to one or more of the aforementioned criteria. The chosen data sets (Table 1) and workflows used during analysis are presented more fully in "Supplementary material 2 - Data sets". Biological relevance of the results was assessed with Gene Set Enrichment Analysis [19] tools EnrichR [20,21] and BioInfoMiner [22].

As an additional measure of verification, we compared its performance with the Rank Products and Rank Sums, both in their original form [1] and their recent implementation [18], as well as with the implementation of our entropic analysis upon the Rank Sum statistic (referred to as "Entropic Sums"). Overall, our approach exhibited consistent behaviour on real as well as simulated data. Functional analysis of the derived feature lists showed that Entropic Ranks provides increased specificity at a concise list size, supporting the argument of efficient rejection of noise-dominated features. Discussion of the implementations and summaries of results are in "Supplementary Material" 1 & 2. All files, tables and plots created are contained in "Supplementary material 3 - Output".

This approach was adopted due to the fact that standard methodologies of comparison, such as list overlap, are inappropriate for two specific reasons. Firstly, they allow the assessment of

interchangeability of two methods given similar thresholding choices, whereas Entropic Ranks was created to be a threshold-independent feature selection process. Secondly, statistical testing methodologies evaluate the value distributions in each population, whereas our approach is driven by patterned behaviour, thus preferring genes with higher fold changes as a byproduct of its function, instead of its main focus.

Table 1. GEO data sets used for verification.

GEO Accession	Organism	Type of Samples	Array	Samples	Cases	Controls
GSE12288	H. sapiens	Total RNA from leukocytes in peripheral blood	Affymetrix Human Genome U133A Array	222	110	112
GSE69486	H. sapiens	Total RNA from fibroblasts from skin biopsies	Illumina HumanHT-12 V4.0 expression beadchip	12	10	2
GSE60767	H. sapiens	Total RNA was extracted from leukocytes	Illumina HumanHT-12 V3.0 expression beadchip	466	312	154
GSE42861	H. sapiens	DNA from Blood leukocytes	Illumina HumanMethylation450 BeadChip	44	20	24
SRP127667	H. sapiens	DNA from sorted cardiac myocytenuclei	Illumina HiSeq 2500	13	10	3

3.2. Simulated data [23]

In order to assess performance on simulated data with known truth values, as is standard practice, we elected an RNAseq count table simulation data set with spiked values. It has been created using a random number generator for the express purpose of benchmarking differential expression methodologies and consists of 10 samples, 5 "cases" and 5 "controls". Out of the 12500 features, 1250 known features have spiked values.

Our method highlights features in a manner different than usual statistical testing: instead of relying on statistical value thresholding, it trims the resulting lists according to the identification of pockets of organized and consistent behavior among the features investigated. Consequently, direct comparison to approaches such as t-testing can be difficult, especially on simulated data tailored to fit hypothesis testing. Simulated data sets created with random number generators exhibit none of the underlying biological constraints present in real data. Moreover, the differentially expressed features in biological systems tend to be organized in consistent networks.

These differences lead us to expect that our method, which detects patterns of expression instead of statistical distributions, will underperform compared to hypothesis testing approaches in a simulated, spiked value data set. Even more importantly, our method aims to assess the information content of features' behavior across populations. Simulated data created using random number generators by definition exhibit highly stochastic behaviour in the absence of biological constraints which are difficult to model. Assessment of such a data set in terms of information content should be expected to return few findings, if any, as there are no consistent patterns to be detected. Moreover, removal of the spiked values should reduce the findings even further, possibly eliminating them altogether.

Indeed, Entropic Ranks underperformed in the identification of the spiked features when compared to limma on the simulated data, as can be seen in "Supplementary data 3 - output". Testing the null hypothesis by removing the spiked values reduced performance even further, leading to the identification of 23 differentially expressed features by Entropic Ranks compared to a single differentially expressed feature returned by limma. However, investigation of these 23 differentially "false positives" showed that they represented rows for which the random number engine had failed to create properly uniform distributions across samples. Instead, these features were highly differentially expressed (15-fold or more) between populations, usually with a single outlying value in one of the two populations. Both under the null hypothesis and when using the full data, Entropic Ranks was more robust than Rank Products and "Entropic Sums" against false positive results. Its robustness was comparable with Rank Sums, which has convergence issues with high data dimensionality (see "Supplementary Material 1 - Method", Part 5).

This level of sensitivity to patterned behaviour and robustness against false positive discoveries should be considered features of the method. Moreover, the plots created by Entropic Ranks during entropy calculation show very high entropy near the beginning of the distribution, and tend to oscillate around lower values further on. This pattern shows an initial low-entropy cluster behaving similar to

the noise-dominated areas than we see in real data sets. Such behaviour could help the experimenter identify poorly structured data sets.

3.3. Series GSE12288 [24]

The set provides microarray gene expression data of leukocytes from 110 patients with Duke Coronary Artery Disease (CAD index >23) and 112 control subjects (CADi=0). It is provided as a studied data set of a known pathology upon which the baseline specificity of the method can be assessed. Moreover, we can compare our method to a mainstream analysis workflow by comparing our results to the list of 160 genes identified in the original publication as significantly ($\rho>0.2$, $P<0.0027$) correlated with the CAD index.

A comparison of cardiovascular diseases associated with the gene lists identified in the original publication and through our methodology using CTD [25] set analyzer is presented in Table 2. The gene lists do not overlap, with the exception of a single gene, CDC42, which has been shown to function as an anti-hypertrophic molecular switch in the heart [26,27]. BioInfoMiner was used to map the list onto the Human Phenotype [28] and MGI Mammalian [29,30] ontologies, highlighting a number of inflammatory response terms and T-cell activation processes, associated with abnormalities of the hematopoietic system. Mapping our list onto the Reactome ontology [31,32] through BioInfoMiner highlights a Selenocysteine synthesis process, which has been shown to have antioxidant effects [33]. EnrichR mapped the resulting gene list onto dbGaP [34], ranking “Hypertension” as the top term by combined score. At the same time, the sampled tissue was clearly identified through Jensen TISSUES [35], ARCHS4 Tissues [36] and Human Gene Atlas [37] as “Blood”, “Peripheral Blood” and “Whole Blood”, respectively. When we mapped the 160-gene list from the original publication using BioInfoMiner, it returned relevant results, especially in the Human Phenotype and MGI Mammalian ontologies, but it failed to achieve both the breadth and specificity of the results returned by the list generated using our approach (see “Supplementary data 3 - Output”, under GSE12288 for the full results).

Table 2. GSE12288: Results of the comparison of differentially expressed genes identified by the initial study and Entropic Ranks with cardiovascular diseases through CTD.

Disease Name	Disease ID	Disease Categories	Original Analysis Corrected P-value	Entropic Ranks Corrected P-value
Stroke	MESH:D020521	Cardiovascular disease ! Nervous system disease	0,00408	0,00504
Infarction, Middle Cerebral Artery	MESH:D020244	Cardiovascular disease ! Nervous system disease	5,88e-6	0,00712
Cardiovascular Diseases	MESH:D002318	Cardiovascular disease	-	0,01264
Cerebral Arterial Diseases	MESH:D002539	Cardiovascular disease ! Nervous system disease	-	0,01484
Cerebral Infarction	MESH:D002544	Cardiovascular disease ! Nervous system disease	-	0,01484
Intracranial Arterial Diseases	MESH:D020765	Cardiovascular disease ! Nervous system disease	-	0,01829
Vascular Diseases	MESH:D014652	Cardiovascular disease	1,23e-4	0,01950
Brain Infarction	MESH:D020520	Cardiovascular disease ! Nervous system disease	-	0,02086
Hypertension	MESH:D006973	Cardiovascular disease	0,01236	-

3.4. Series GSE69486

In order to assess the specificity even when the input signals are weak due to technical or biological (phenotype associated) reasons, we used a gene expression data set containing microarray gene expression data of fibroblast cells from 10 samples of patients with bipolar disease and two control samples. A way to assess specificity is the capability of the resulting differentially expressed list to identify the the cell population. Sensitivity is evaluated if functional analysis of the gene list identified ontological terms associated with the neurological pathologies underlying bipolar condition (as was the hypothesis of the original study).

EnrichR successfully identified the cell population as “Fibroblast” through ARCHS4. Achilles Fitness Decrease [38] highlighted “GB1-central nervous system” as the most significant term by rank-based. More phenotype-specific results were achieved through BioInfoMiner-mediated mapping of the list onto MGI Mammalian, which is densely described. This mapping includes a distinct branch of terms related to nervous system abnormalities. Of note is the presence of MMP3 in the highly connected gene list produced by BioInfoMiner, as it has been shown to be tied to bipolar disease [39].

Moreover, Reactome highlighted the term for “Synthesis of Prostaglandins (PG) and Thromboxanes (TX)”, which have been tied to bipolar disease [40] and are used as markers in relevant pharmacological research [41].

3.5. Series GSE60767 [42]

Chosen to assess performance in data of high noise content and to assess the sensitivity of the proposed method, this data set contains gene expression data from 312 leukocyte samples of healthy adult males from the highly polluted industrial region of Ostrava and 154 healthy male control samples from Prague. The study aimed to investigate differential gene expression induced by chronic exposure to elevated pollution levels. Due to the weakness of the biological signal and the need to address a significant batch effect induced by beadchip performance, standardized t-testing identified no statistically significant (p-value <0.05) differentially expressed genes within each of the three sampling seasons, even with very low long fold change (lfc) thresholds (<0.1).

Our methodology was able to identify genes with differential behaviour in regards to the city of origin in all three sampling seasons, as seen in Table 3. Mapping the resulting gene lists through EnrichR consistently identified the “diabetes melitus, type 2” through OMIM Disease and OMIM Expanded [43], a connection supported by past research [44]. Bioinforminer mapping of the gene lists onto Gene Ontology [45,46] and MGI Mammalian ontologies highlights terms characteristic of a response to increased levels of particulate matter and associated pollutants. There are generalized inflammation indicators which have been tied to cardiovascular syndromes and lung cancer [47], as well as terms relating to the development of the nervous system, which can be attributed to the toxic metal load of particulate matter particles [48].

Table 3. Number of differentially expressed features identified in the initial study compared with Entropic Rank results. The last column notes the maximum p-value calculated through Rank Products for these features. The number of features shows no correlation with sample size and the number identified in the initial research.

GEO accession	Type	Original publication S	Entropic ranks S	RP maximum p-value
GSE12288	Differentially expressed	160	18*	< 0.0001
			86**	< 0.0001
GSE69486	Upregulated	127	135	0.0027
	Downregulated	330	89	0.0083
GSE60767 [winter 2009]	Upregulated	0	29	< 0.0001
	Downregulated	0	53	< 0.0001
GSE60767 [summer 2009]	Upregulated	0	33	< 0.0001
	Downregulated	0	26	< 0.0001
GSE60767 [winter 2010]	Upregulated	0	34	< 0.0001
	Downregulated	0	41	< 0.0001
GSE42861	Hypermethylated	NA	24	< 0.0001
	Hypomethylated	NA	98	< 0.0001
SRP127667	Upregulated	NA	32	< 0.0001
	Downregulated	NA	37	< 0.0001

*: Upregulated, **: Downregulated

3.6. Series GSE42861 [49,50]

This data set was selected in order to test the performance and specificity of the proposed method on DNA methylation data profiles of a known pathology. This will also allow evaluation of the generality of the proposed method, given that DNA methylation platforms contain many more probes, have different distribution of values (M-values) and are also greatly influenced by blood cell population perturbations between samples. The study explores the methylation profiles of peripheral blood leukocytes from patients with rheumatoid arthritis compared to healthy controls. We opted to apply our methodology onto the subpopulation of the data set consisting of samples taken from 50 to 60 years old men and women who had never been smokers, in order to reduce potential confounders.

Out of the 44 samples thus selected, 20 were of patients with rheumatoid arthritis and 24 were control samples.

Using EnrichR, Jensen DISEASES identified rheumatoid arthritis as the most significant term by rank based ranking. Of particular note is the presence of allograft rejection terms at the top of the lists of both KEGG 2016 [51] and WikiPathways 2016 [52], pointing to the triggering of the same basic mechanisms in the course of the disease. BioInfoMiner mapping of the list onto Gene Ontology, Human Phenotype, MGI Mammalian and Reactome provides highly overlapping results. There is an overarching inflammatory response with terms specific to T-cell activation. Gene Ontology highlights the “telomere maintenance” (Figure 3) term, which has been an area of active study as to its implication in autoimmune syndromes [53–55]. Furthermore, in the Human Phenotype ontology, highlighted terms include autoimmunity and rheumatoid arthritis. Lastly, the highly connected genes identified through BioInfoMiner for these four ontologies have a strong presence of the major histocompatibility complex family (HLA-C, HLA-DRB1, HLA-DQB1, HLA-DQA1, HLA-DRB1), a finding in agreement with one of the studies citing the data set [50]. Even with a reduced sample number, the proposed method extracted a biological signal highly relevant to the phenotype and in line with the findings of the original study for the full data set.

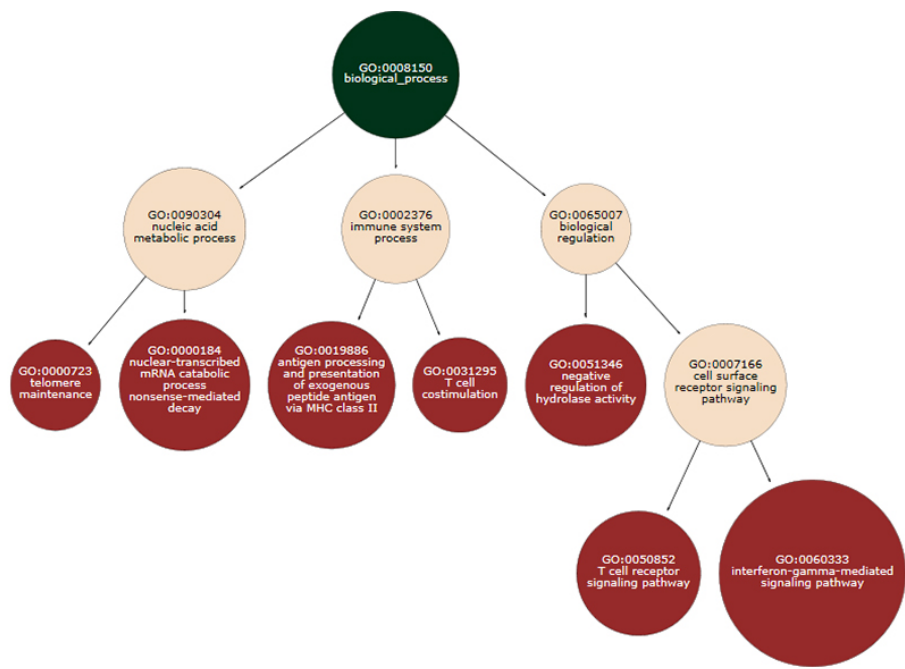


Figure 3. GSE42861: differentially methylated genes mapped onto Gene Ontology using BioInfoMiner

3.7. Series SRP127667

In order to further test the generality of the proposed method, we applied it on RNAseq count table data, which follow different distributions than gene expression and methylation data. The application of rank statistics on RNAseq count data is novel within the Bioinformatics field. Nevertheless, in the field of Astrophysics RPs have been successfully used in pipelines performing occultation [56] and gravitational wave [57] event verification. These "discovery enumeration" phenomena are Poisson point processes, similar to the discovery-based formation of RNAseq count tables from RNA reads. Given that the requirements of RP hold for the count tables except for the independence of variance (which has been shown to affect statistical threshold selection [58], which we do not perform), we extend the verification testing of our non-parametric methodology to RNAseq data, aiming to assess the specificity of the method, despite the biological and computational problems generated by RNAseq data as opposed to transcriptomics. We compared the gene counts obtained from RNA sequencing of

cardiac myocytes from 10 adult patients with terminal heart failure to three control samples from the BioProject study SRP127667.

EnrichR was used to map the differentially expressed genes to Panther 2016 [59], Jensen DISEASES and Reactome. Panther 2016 highlighted as the first term by combined ranking the Wnt signaling pathway, which has been implicated in cardiovascular syndromes [60]. Jensen DISEASES terms ranked first by combined score were “Hypertension”, “Coronary artery disease” and “Cerebrovascular disease”. Reactome highlighted as the second term by combined ranking the Ca^{2+} pathway. Using BioInfoMiner to map the list onto Gene Ontology and Human Phenotype Ontology showed terms related to thrombosis abnormalities, tyrosine phosphorylation of Stat3 protein, and the regulation of body fluid levels through the urinary system - the last of which which is a known regulator mechanism of blood pressure (Figure 4). Lastly, the highly connected genes identified through BioInfoMiner are specifically associated with pharmaceuticals prescribed for cardiovascular conditions.

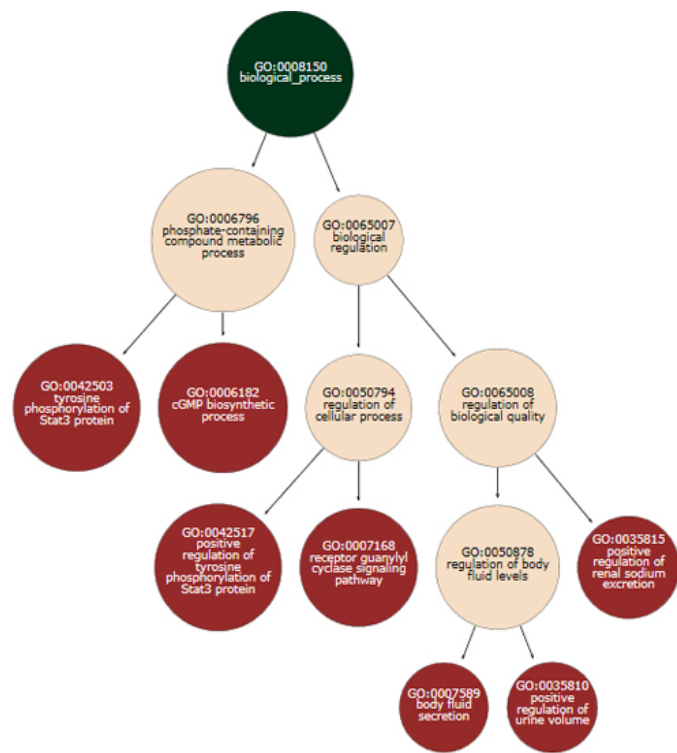


Figure 4. SRP127667: differentially expressed genes mapped onto Gene Ontology using BioInfoMiner

4. Conclusion

We present and evaluate a methodology which builds upon the Rank Products foundations to create a generalized method for threshold-independent selection of differentially expressed features, according to the information content of their behaviour.

The biological interpretation of the functional analysis performed on each data set supports the capability of our method to separate information-rich data from noise, eschewing the limitations and plights of fold change and p-value thresholding approaches which are inherent in statistical testing approaches such as t-testing. Fold change and pfp are computed, but are relegated to quality indicators for the evaluation of the experiments and subsequent analysis instead of being used as decision criteria.

The analytic workflow we compose exploits solely the elementary preprocessing, normalization and signal correction bioinformatic techniques, to ensure reproducibility of results and transparency of the comparative evaluation. No further processing steps, aiming to force values to conform to a specific kind of distribution were used, alluding to the generalized character of the proposed method, as well as the case of its introduction for broader data analytic scenarios.

4.1. Features of Entropic Ranks

In summary, the proposed method extends the rank product methodology by incorporating the measurement of information content as an integral part of the analysis and interpretation. Firstly, the selection of significant genes is based on the distribution of all genes over the entire populations, rather than evaluating each gene independently. Information-poor data, such as simulated data sets [23] with a very low signal-to-noise ratio exhibit a starkly different entropy distribution, without a defined, initial, low-entropy area followed by stably high entropy area with minimal oscillations (see "Supplementary material 2 - Data sets" and "Supplementary material 3 - Output"). Secondly, the methodology is applicable to a broad array of selection problems and data types, as long as they conform to basic assumptions made by the rank products methodology. Thirdly, it departs from the adoption of arbitrary or empirical statistical thresholds, exploring the information density of the distribution and cherry picking clusters of high information content, through rigorous entropic analysis. Fourthly, the automation of the partition process is possible, allowing for unsupervised and unbiased analytical processes to be applied. The fifth feature is the ability to freely adjust the analytic granularity (by changing the sliding window step) to more refined or coarser inspection, enabling solutions of varying computational cost and level of convergence. Lastly, another advantage of this method is the potential for integration of data from different sources or dissection levels into the same analysis, as long as they can be transformed to similar, ranked value distributions.

Funding: Irene Liampa's PhD thesis is supported by a scholarship from the State Scholarship Foundation in Greece (IKY) (Operational Program "Human Resources Development - Education and Lifelong Learning" Partnership Agreement (PA) 2014-2020).

We acknowledge support of this work by the project "ELIXIR-GR: Hellenic Research Infrastructure for the Management and Analysis of Data from the Biological Sciences" (MIS 5002780) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Acknowledgments: Aristotelis Chatziioannou is co-founder of e-NIOS PC.

Author Contributions: Conceptualization, A.C.; methodology, H-X. de L. and E.L.; software, H-X.; validation, E.L. and H-X. de L.; writing—original draft preparation, H-X. de L. and E.L.; writing—review and editing, A.G.G., M.Z. and A.C.; visualization, X.X.; supervision, A.C., A.G.G. and M.Z.; All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

RP Rank Product

References

1. Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P. Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters* **2004**, doi:10.1016/j.febslet.2004.07.055
2. Tsimring, L., S. (2014). Noise in Biology. *Rep Prog Phys.* **2014**, 77(2): 026601, doi:10.1088/0034-4885/77/2/026601
3. Leek, J., McShane, B.B., Gelman, A., Colquhoun, D., Nuijten, M.B., Goodman, S. Five ways to fix statistics. *Nature* **2017**, doi:10.1038/d41586-017-07522-z
4. Chawla, D.S. 'One-size-fits-all' threshold for P values under fire. *Nature* **2017**, doi:10.1038/nature.2017.22625
5. Dalman, M.R., Deeter, A., Nimishakavi, G., and Duan, Z. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics* **2012**, doi:10.1186/1471-2105-13-s2-s11

6. Del Carratore, F., Jankevics, A., Eisinga, R., Heskes, T., Hong, F., Breitling, R. RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. *Bioinformatics* **2017**, doi:10.1093/bioinformatics/btx292
7. Yang, T., Y. A Simple Rank Product Approach for Analyzing Two Classes. *Bioinformatics and Biology Insights* **2015**, doi:10.4137/BBI.S26414
8. Liu, X., Krishnan, A., Mondry, A. An Entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics* **2005**, 6:76, doi: <https://doi.org/10.1186/1471-2105-6-76>
9. Wang, Y., Yan, H. Entropy based sub-dimensional evaluation and selection method for DNA microarray data classification. *Bioinformation* **2008**, 3(3):124-9
10. Furlanello, C., Serafini, M., Merler, S., Jurman, G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics* **2003**, 4: 54, <https://doi.org/10.1186/1471-2105-4-54>
11. Guo, Z., Xin, Y., Zhao, Y. Cancer classification using entropy analysis in fractional Fourier domain of gene expression profile. *Biotechnology & Biotechnological Equipment* **2018**, 32:4, 1042-1046, doi: 10.1080/13102818.2017.1413596
12. Liu, H., Zhao, R., Fang, H., Cheng, F., Fu, Y., Liu, Y. Entropy-based consensus clustering for patient stratification. *Bioinformatics* **2000**, <https://doi.org/10.1093/bioinformatics/btx167>
13. Tung, T. Q., Ryu, T., Lee, K. H., Lee, D. Inferring Gene Regulatory Networks from Microarray Time Series Data Using Transfer Entropy. *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS07)* **2007**, doi:10.1109/cbms.2007.60
14. Langmead, C., Mcclung, C., Donald, B. A maximum entropy algorithm for rhythmic analysis of genome-wide expression patterns. *Proceedings. IEEE Computer Society Bioinformatics Conference* **2002** doi:10.1109/csb.2002.1039346
15. Zambelli, F., Mastropasqua, F., Picardi, E., D'Erchia, A. M., Pesole, G., Pavesi, G. RNentropy: an entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. *Nucleic Acids Research* **2018**, <https://doi.org/10.1093/nar/gky055>
16. Batista, M.V., Ferreira, T.A., Freitas, A.C., Balbino, V.Q. An entropy-based approach for the identification of phylogenetically informative genomic regions of Papillomavirus. *Infection, Genetics and Evolution* **2011**, 11(8), 2026-2033. doi:10.1016/j.meegid.2011.09.013
17. Hausser, J., Strimmer, K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn.* **2009**, doi:10.1.1.165.2233
18. Eisinga, R., Breitling, R., Heskes, T. The exact probability distribution of the rank product statistics for replicated experiments. *FEBS Lett.* **2013**, 18;587(6):677-82, doi: 10.1016/j.febslet.2013.01.037
19. 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., Mesirov, J.P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **2005**, 102 (43) 15545-15550, doi:10.1073/pnas.0506580102
20. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., Koplev, S., Jenkins, S.L., Jagodnik, K. M., Lachmann A., McDermott, M.G., Monteiro, C.D., Gundersen G.M., Ma'Ayan, A. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **2013**, doi:10.1186/1471-2105-14-128
21. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W., Maayan, A. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **2016**, doi:10.1093/nar/gkw377
22. Koutsandreas T, Binenbaum I, Pilalis E, Valavanis I, Papadodima O, Chatziioannou A. Analyzing and visualizing genomic complexity for the derivation of the emergent molecular networks. *Int J Monit Surveill Technol* **2016**, doi:10.4018/IJMSTR.2016040103
23. UZH, Statistical Bioinformatics Group, data hosted at http://imlspenticton.uzh.ch/robinson_lab/benchmark_collection/
24. Sinnaeve, P.R., Donahue, M.P., Grass, P., Seo, D., Vonderscher, J., Chibout, S.-D., Kraus, W.E., Sketch M. Jr., Nelson, C., Ginsburg, G.S., Goldschmidt-Clermont, P.J., Granger, C.B. Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS One* **2009**, doi:10.1371/journal.pone.0007037

25. Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., King, B.L., McMorran, R., Wiegers, J., Wiegers, T.C., Mattingly, C.J. The Comparative Toxicogenomics Database: Update 2017. *Nucleic Acids Research* **2016**, doi:10.1093/nar/
26. Mailliet, M., Lynch, J.M., Sanna, B., York, A.J., Zheng, Y., Molkentin, J.D. Cdc42 is an antihypertrophic molecular switch in the mouse heart. *Journal of Clinical Investigation* **2009**, doi:10.1172/jci37694
27. Gu, R., Zheng, D., Bai, J., Xie, J., Dai, Q., and Xu, B. Altered melusin pathways involved in cardiac remodeling following acute myocardial infarction. *Cardiovascular Pathology* **2012**, doi:10.1016/j.carpath.2011.03.002
28. Köhler, S., Vasilevsky, N., Engelstad, M., Foster, E., et al. The Human Phenotype Ontology. *Nucl. Acids Res.* **2017**, doi:10.1093/nar/gkw1039
29. Blake, J.A., Eppig, J.T., Kadin, J.A., Richardson, J.E., Smith, C.L., Bult, C.J. Mouse Genome Database (MGD)-2017: Community knowledge resource for the laboratory mouse. *Nucleic Acids Research* **2016**, doi:10.1093/nar/gkw1040
30. Ringwald, M. The Mouse Gene Expression Database (GXD). *Nucleic Acids Research* **2001**, doi:10.1093/nar/29.1.98
31. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., Deustachio, P. The Reactome pathway knowledgebase. *Nucleic Acids Research* **2013**, doi:10.1093/nar/gkt1102
32. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Robin, H., Bijay, J., Florian, K., Bruce, M., Marija, M., Corina, D.R., Karen, R., Cristoffer, S., Veronica, S., Solomon, S., Thawfeek, V., Guilherme, V., Joel, W., Guanming, W., Lincoln, S., Henning, H., D'Eustachio, P. The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **2017**, doi:10.1093/nar/gkx1132
33. Suh, N., Lee, E. Antioxidant effects of selenocysteine on replicative senescence in human adipose-derived mesenchymal stem cells. *BMB Reports*, 50(11) **2017**, doi:10.5483/bmbrep.2017.50.11.174
34. dbGaP/database of Genotypes and Phenotypes National Center for Biotechnology Information, National Library of Medicine (NCBI/NLM). <https://www.ncbi.nlm.nih.gov/gap>
35. Santos, A., Tsafou, K., Stolte, C., Pletscher-Frankild, S., O'Donoghue, S.I., Jensen, L.J. Comprehensive comparison of large-scale tissue expression datasets **2015**, doi:10.7287/peerj.preprints.1072v1
36. Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Silverstein, M.C., Wang, L., Maayan, A. Massive Mining of Publicly Available RNA-seq Data from Human and Mouse. **2017**, doi:10.1101/189092
37. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M.P. Walker, J.R., Hogenesch, J.B. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, doi:10.1073/pnas.0400782101
38. Cowley, G.S., Weir, B.A., Vazquez, F., Tamayo, P., Scott, J.A., Rusin, S., East-Seletsky, A., Ali, L.D., Gerath, W.F., Pantel, S.E., Lizotte, P.H., Jiang, G., Hsiao, J., Tsherniak, A., Dwinell, E., Aoyama, S., Okamoto, M., Harrington, W., Gelfand, E., Green, T.M., Tomko, M.J., Gopal, S., Wong, T.C., Li, H., Howell, S., Stransky, N., Liefeld, T., Jang, D., Bistline, J., Hill Meyers, B., Armstrong, S.A., Anderson, K.C., Stegmaier, K., Reich, M., Pellman, D., Boehm, J.S., Mesirov, J.P., Golub, T.R., Root, D.E., Hahn, W.C. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific Data* **2014**, doi:10.1038/sdata.2014.35
39. Kucukali, C.I., Aydin, M., Ozkok, E., Bilge, E., Orhan, N., Zengin, A., and Kara, I. Do schizophrenia and bipolar disorders share a common disease susceptibility variant at the MMP3 gene? *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **2009**, doi:10.1016/j.pnpbp.2009.02.012
40. Gurvich, A., Begemann, M., Dahm, L., Sargin, D., Miskowiak, K., Ehrenreich, H. A role for prostaglandins in rapid cycling suggested by episode-specific gene expression shifts in peripheral blood mononuclear cells: a preliminary report. *Bipolar Disord* **2014**, doi: 10.1111/bdi.12223
41. Savitz, J. B., Teague, T. K., Misaki, M., Macaluso, M., Wurfel, B. E., Meyer, M., Drevets, D., Yates, W., Gleason, O., Drevets, W.C., Preskorn S.H. Treatment of bipolar depression with minocycline and/or aspirin: An adaptive, 2x2 double-blind, randomized, placebo-controlled, phase IIA clinical trial. *Translational Psychiatry* **2018**, doi:10.1038/s41398-017-0073-7
42. Rossner, P., Tulupova, E., Rossnerova, A., Libalova, H., Honkova, K., Gmuender, H., Pastorkova, A., Svecova, V., Topinka, J., Sram, R.J. Reduced gene expression levels after chronic exposure to high

- concentrations of air pollutants. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **2015**, doi:10.1016/j.mrfmmm.2015.08.001
43. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). Retrieved from <http://www.ncbi.nlm.nih.gov/omim/>
 44. Rajagopalan S., Brook R., D. Air pollution and type 2 diabetes: mechanistic insights. *Diabetes* **2012**, doi:10.2337/db12-0190
 45. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. Gene Ontology: Tool for the unification of biology. *Nature Genetics* **2000**, doi:10.1038/75556
 46. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **2016**, doi:10.1093/nar/gkw1108
 47. Iii, C.A. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *Jama* **2002**, doi:10.1001/jama.287.9.1132
 48. Huiming, L., Xin, Q., and Qin'geng, W. Heavy Metals in Atmospheric Particulate Matter: A Comprehensive Understanding Is Needed for Monitoring and Risk Mitigation. *American Chemical Society* **2013**, doi:10.1021/es404751a
 49. Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekström, T.J., Feinberg, A. P. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* **2013**, doi:10.1038/nbt.2487
 50. Kular, L., Liu, Y., Ruhrmann, S., et al. DNA methylation as a mediator of HLA-DRB1*15:01 and a protective variant in multiple sclerosis. *Nat Commun.* **2018**, doi: 10.1038/s41467-018-04732-5.
 51. Kanehisa, M., Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000**
 52. Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., Ehrhart, F., Giesbertz, P., Kalafati, M., Martens, M., Miller, R., Nishida, K., Rieswijk, L., Waagmeester, A., Eijssen, L.M.T., Evelo, C.T., Pico, A.R., Willighagen, E.L. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* **2017**, doi:10.1093/nar/gkx1064
 53. Hohensinner, P., J., Goronzy, J., J., Weyand, C., M. Telomere dysfunction, autoimmunity and aging. *Aging Dis* **2011**
 54. Hohensinner, P., J., Goronzy, J., J., Weyand, C., M. Targets of immune regeneration in rheumatoid arthritis. *Mayo Clin Proc* **2014**, doi:10.1016/j.mayocp.2014.01.020
 55. Georgin-Lavialle, S., Aouba, A., Mouthon, L., Londono-Vallejo, J., A., Lepelletier, Y., Gabet, A., S., Hermine, O. The telomere/telomerase system in autoimmune and systemic immune-mediated diseases. *Autoimmun Rev.* **2010**, doi:10.1016/j.autrev.2010.04.004
 56. Lehner, M. J., Coehlo, N. K., Zhang, Z., Bianco, F. B., Wang, J., Rice, J. A., Protopapas, P., Alcock, C., Axelrod, T., Byun, Y.-I., Chen, W. P., Cook, K. H., de Pater, I., Kim, D.-W., King, S.-K., Lee, T., Marshall, S. L., Schwamb, M. E., Wang, S.-Y., Wen, C. The TAOS Project: Statistical Analysis of Multi-Telescope Time Series Data. *Publications of the Astronomical Society of the Pacific* **2010**, 122(894), 959–975, doi: 10.1086/655443
 57. Aasi, J. et al. (LIGO Scientific Collaboration and Virgo Collaboration). First low frequency all-sky search for continuous gravitational wave signals *Phys. Rev. D* **2016**, 93, 042007, doi:10.1103/PhysRevD.93.042007
 58. Breitling, R., and Herzyk, P. Rank-Based Methods As A Non-Parametric Alternative Of The T-Statistic For The Analysis Of Biological Microarray Data. *Journal of Bioinformatics and Computational Biology* **2005**, 03(05), 1171-1189, doi:10.1142/s0219720005001442
 59. Thomas, P., D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., Narechania, A. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **2003**, doi:10.1101/gr.772403
 60. Hermans, K., C., Blankesteijn, W., M. Wnt Signaling in Cardiac Disease. *Comprehensive Physiology* **2015**, doi:10.1002/cphy.c140060

