CluGene: A Bioinformatics Framework for the Identification of Co-Localized, Co-Expressed and Co-Regulated Genes Aimed at the Investigation of Transcriptional Regulatory Networks from High-Throughput Expression Data

Tania Dottorini[1], Pietro Palladino[2], Nicola Senin[3], Tania Persampieri[2], Roberta Spaccapelo[2], Andrea Crisanti[1,2]*

[1] Department of Biological Sciences, Imperial College London, London, United Kingdom, [2] Department of Experimental Medicine, University of Perugia, Perugia, Italy, [3] Department of Industrial Engineering, University of Perugia, Perugia, Italy

Abstract

The full understanding of the mechanisms underlying transcriptional regulatory networks requires unravelling of complex causal relationships. Genome high-throughput technologies produce a huge amount of information pertaining gene expression and regulation; however, the complexity of the available data is often overwhelming and tools are needed to extract and organize the relevant information. This work starts from the assumption that the observation of *co-occurrent* events (in particular co-localization, co-expression and co-regulation) may provide a powerful starting point to begin unravelling transcriptional regulatory networks. Co-expressed genes often imply shared functional pathways; co-expressed and functionally related genes are often co-localized, too; moreover, co-expressed and co-localized genes are also potential targets for co-regulation; finally, co-regulation seems more frequent for genes mapped to proximal chromosome regions. Despite the recognized importance of analysing co-occurrent events, no bioinformatics solution allowing the simultaneous analysis of co-expression, co-localization and co-regulation is currently available. Our work resulted in developing and valuating CluGene, a software providing tools to analyze multiple types of co-occurrences within a single interactive environment allowing the interactive investigation of combined co-expression, co-localization and co-regulation of genes. The use of CluGene will enhance the power of testing hypothesis and experimental approaches aimed at unravelling transcriptional regulatory networks. The software is freely available at http://bioinfolab.unipg.it/.

Citation: Dottorini T, Palladino P, Senin N, Persampieri T, Spaccapelo R, et al. (2013) CluGene: A Bioinformatics Framework for the Identification of Co-Localized, Co-Expressed and Co-Regulated Genes Aimed at the Investigation of Transcriptional Regulatory Networks from High-Throughput Expression Data. PLoS ONE 8(6): e66196. doi:10.1371/journal.pone.0066196

Editor: Christian Schönbach, Kyushu Institute of Technology, Japan

Received November 12, 2012; **Accepted** May 5, 2013; **Published** June 18, 2013

Funding: This work was supported by the European Union, [PIEF-GA-2010-273268], acronym Vectrap and [218164], acronym FIGHTMAL. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: acrs@imperial.ac.uk

Introduction

The maintenance and modulation of cellular activities is a complex process controlled by the co-ordinated expression and action of different gene networks. This composite process is regulated by a higher-order genome structure arranged into transcriptional regulatory modules. Understanding the organization of these networks is crucial for elucidating how different gene combinations work together in accomplishing distinct cellular functions. Although genome-wide transcriptional analysis and in general genome high-throughput experimental technologies are producing a vast amount of information, the relationships existing amongst genome organization, gene regulation and the diverse facets of cell activity still remain largely unknown.

Experimental observations indicate that genes sharing similar expression patterns are correlated either from a functional or evolutionary standpoint [1,2,3]. A key question regarding co-expressed genes is whether they are under common transcription regulatory controls (i.e. gene co-regulation), that is, genes that by sharing common *cis*-regulatory elements in their promoters are highly likely controlled by the same transcription factors.

Recent experimental evidence revealed that transcription regulation may reflect also an underlying chromosomal gene positional order [4,5,6,7,8]. Interestingly, in several eukaryotic organisms, functionally linked and co-transcribed genes have been reported to cluster together in physical proximity unravelling a gene organization with operon-like features, i.e. co-regulation of neighbouring genes [9]. This is supported by the observation that proximity of co-expressed genes does not occur randomly, but rather underlies a selection process involving genes working together in signalling and metabolic pathways [9,10,11,12,13,14]. Consistently, experimental investigations revealed that functionally related, co-localized genes show higher probability of being co-expressed than random pairs of genes [15,16,17]. The existence of a non-random organization of the genome has been confirmed by the observation that co-localized genes targeted by the same transcription factor are subjected to stronger co-regulation when compared to randomly distributed targets [18]. It has been proposed that the complex eukaryotic transcriptional regulatory mechanism has imposed a significant constraint on gene positional organization along and across eukaryotic chromosomes [19]. Clusters of co-expressed and co-localized genes have been recognized as playing key roles in a number of functional pathways and adaptive responses including organism development, differentiation, disease states and aging.

Although the importance of identifying genes localized in close proximity to each other on the chromosome (gene positional clusters) is widely recognized as a means for unravelling unsuspected functional and transcriptional clues, only few bioinformatics tools and methods are available to support the analysis of large-scale gene datasets [5,20,21,22,23,24]. However, they do not allow the investigation of relationships between positional organization of genes within the genome and transcription control. Current methods for the analysis of genome-wide expression data allow to determine which genes show similar increased or decreased expression profiles (i.e. co-expression) in a specific experimental condition, over a number of conditions or throughout a time course. To identify transcriptional networks, co-expressed genes are searched at a later stage for locating common transcription regulatory controls (i.e. co-regulation). Recently, many transcription factor predictor tools have become available. They mostly focus on searching for transcription factor binding sites (TFBS) by using position specific scoring matrices (PSSMs) [25,26] and by using a comparative genomic approach to identify conserved sequences among homologous promoters, namely phylogenetic footprinting [27,28,29,30,31]. At present, these approaches still suffer from a high rate of false positive and false negative results. It has been estimated that only 10% of the overall predictions matches to biologically functional binding sites [32]. Integrated approaches using both TFBSs search methods (PSSMs and Phylogenetic footprinting) and gene expression data were recently implemented, leading to a dramatic improvement in specificity while identifying over-represented TFBSs in sets of co-expressed genes [33,34]. Interestingly, integration of gene expression data has been demonstrated to increase TFBSs prediction specificity, namely enrichment of functional binding sites versus false positives [33]. Another interesting approach [35] attempts at identifying regulatory modules and their cognate regulators from gene expression data using the module network algorithm; although of remarkable importance, the method is characterized by potential limitations, such as: transcription factor post-transcriptional modifications, low regulator expression variation, or the identification of only one regulator out of potentially many.

Here we report on the development of the first standalone software application named CluGene for the identification of co-regulated, co-localized and co-expressed genes.

CluGene integrates gene expression analysis profiling with automated search and identification of co-expressed and co-localized genes while searching for transcriptional regulatory modules. The software combines a number of pre and post processing functionalities together with statistical tools that significantly facilitate and expedite the analysis of gene network co-regulation on a global scale. In addition, CluGene provides a plug-in mechanism so that user-defined functionality can be seamlessly integrated within the software framework, allowing the replication of a multitude of literature algorithms and approaches.

Materials and Methods

In this section CluGene main functionalities will be described. Further details and more technical details are reported in Text S1. CluGene is available for download from the website of the Bioinformatics Laboratory (BioInfoLab) at the University of Perugia (http://bioinfolab.unipg.it/) and can be freely used for research purposes. CluGene is implemented with the Java Web Start technology and can run on any operating system (Windows, Mac or Linux) where the Java Runtime Environment (JRE) is installed (version 1.6 or higher is required).

2.1. Architecture and Basic Principles of Operation of CluGene

A typical CluGene work session revolves around a *project* where gene-related data is managed, analyses are run, and results are generated and visualized. Projects can be saved to files so that any work session can be restored at a later time.

Within a project, gene-related information is collected into one or multiple *gene datasets*. Essentially, a gene dataset is a list of *ENSEMBL geneID* names; however, the dataset is structured so that additional information can be associated to each gene, depending on current analysis needs, such as: the *chromosome* the gene maps to; the genomic localization, its DNA strand orientation, its promoter, one or multiple *transcription factors* matching the promoter and one or multiple *expression levels* as obtained by multiple sources including microarray, EST, qRT-PCR, SAGE, RNA-seq, or other high-throughput sequencing techniques.

A typical work session begins by creating a new project and by importing one or more user-defined lists of geneIDs from CSV or Excel files. The following organisms are currently supported by CluGene: *Anopheles gambiae*, *Culex quinquefasciatus, Aedes aegypti, Drosophila Melanogaster, Homo sapiens, Mus musculus, Caenorhabditis elegans, Saccharomices cerevisiae, Plasmodium falciparum and Plasmodium berghei;* multiple organisms can be managed in a single CluGene project. Additional organisms retrieved from Biomart can be included in CluGene, the process is rather easy, and can be done upon user request to the developers. The possibility of letting the user update the set of organisms within CluGene is being considered for future releases of the software.

In CluGene chromosome names, positional information, promoters and transcription factors are automatically retrieved from online web-services, while expression levels must be loaded into the project from user-defined CSV files. Genomic expression data coming from any online source can be imported into CluGene, as long as it is first formatted into properly structured input files. It is the user's responsibility to make sure that expression data coming from heterogeneous sources is comparable. This can be achieved by applying several known techniques for normalizing data, e.g. turning expression raw data into fold-changes with respect to given controls.

2.2. Functionality Provided by CluGene

CluGene provides multiple options for operating on gene datasets. These can be conveniently organized into the following macro-categories:

2.2.1. Name-based gene dataset processing. This category contains functions aimed at processing lists of genes with set operations (merge, intersect, subtract).

2.2.2. Expression-based gene dataset processing. This set of functions allows the analysis of gene datasets by taking into account expression-related information. The following main types of operation are possible: 1) adding to the current dataset gene expression information loaded from different sources; 2) searching, sorting and visualizing genes based on expression levels; 3) combining multiple expression levels into single scalar values; 3) filtering datasets based on expression thresholds; and 4) comparing datasets based on expression levels. In Figure 1a an illustrative visualization of genes with multiple expressions is shown, where colour is used to indicate expression level. In Figure 1b an illustrative interactive dialog is shown for gene filtering based on expression level thresholding and Boolean combination of expressions.

2.2.3. TF-based gene dataset processing. This category of functions operates with transcription factors (TF) and with TF related information including promoter sequences and transcription factor binding sites retrieved from different online web-services. See Text S1 for detailed information on used web services. For each gene it is possible to identify the position, the number and the identity of the TFBS, and consequently deduce which TFs can interact with them (Figure S1a). The software also allows the reverse search, e.g. by selecting a TF and searching for all the putative promoter sequences containing the corresponding TFBS, which are returned by the software together with their associated recognition frequencies (Figure S1b). Finally, distinct datasets can be compared in terms of predicted TF binding sides (Figure S1c).

2.2.4. Position-based gene dataset processing. Functions are provided for retrieving gene position in the chromosome and gene length; for retrieving the local density of genes along each chromosome, for computing what percentage of a chromosome is represented by a set of genes; for computing the min, max and average lengths of the genes within a chromosome; the min, max and average distance between consecutive genes in a chromosome, etc. In Figure 1c an interactive window is shown where a chromosome is visualized along with the local gene density plot (see later for details) together with statistical information comparing the selected chromosome with the rest of the genome. Datasets can be processed by positional clustering analysis (i.e., identification of groups of genes -clusters- that are in close proximity with each other in the chromosome). This is an example of *co-occurrence analysis* (in this case specifically addressing co-localization) and represents a typical application for which CluGene has been designed.

2.2.5. Multi-attribute gene dataset processing for co-occurrence analysis. This group of functions embodies the main innovative features of CluGene, i.e. the capability of analysing *co-occurrences* (co-localization, co-expression and co-regulation) of genes-related observations for studying transcriptional regulatory networks. These functions revolve around *gene clustering*, i.e. a set of statistical methods for collecting genes into groups (clusters) based on similarity of attributes. The approach followed by CluGene is designed around the following routines:

- *Pre-processing of the gene dataset.* Genes are filtered out using thresholds of expression levels, or by rejecting those genes that either do not contain specific TFBS, or whose corresponding sequences are too short or too far apart, etc. Filters can be customized by taking into account expression, regulation and localization information to better focus on more interesting co-occurrences;
- *Application of positional clustering.* Genes are collected into clusters based on mutual proximity in the chromosome (i.e. co-localization). Clustering can be also driven by additional measures of similarity that take into account expression level and/or TF-related information, in order to obtain groups of genes that are co-localized, co-expressed and co-regulated.
- *Post-processing and analysis of gene clusters. The* clustering results are filtered based on different thresholding criteria that include expression, regulation or position; statistical analyses are performed to determine the significance of the identified clusters in the specific application context.

In the following sections details are provided on positional clustering and analysis of clustering results.

2.3. Gene Positional Clustering

Positional clustering plays a central role in the analysis of gene co-occurrences. It operates in different ways, depending on the following customised configurations:

2.3.1. Gene distance metric. Genes are grouped according to similarity assessed by a *distance metric*, which is a function that defines how gene attributes should be accounted for and combined in order to obtain a quantitative measure of similarity.

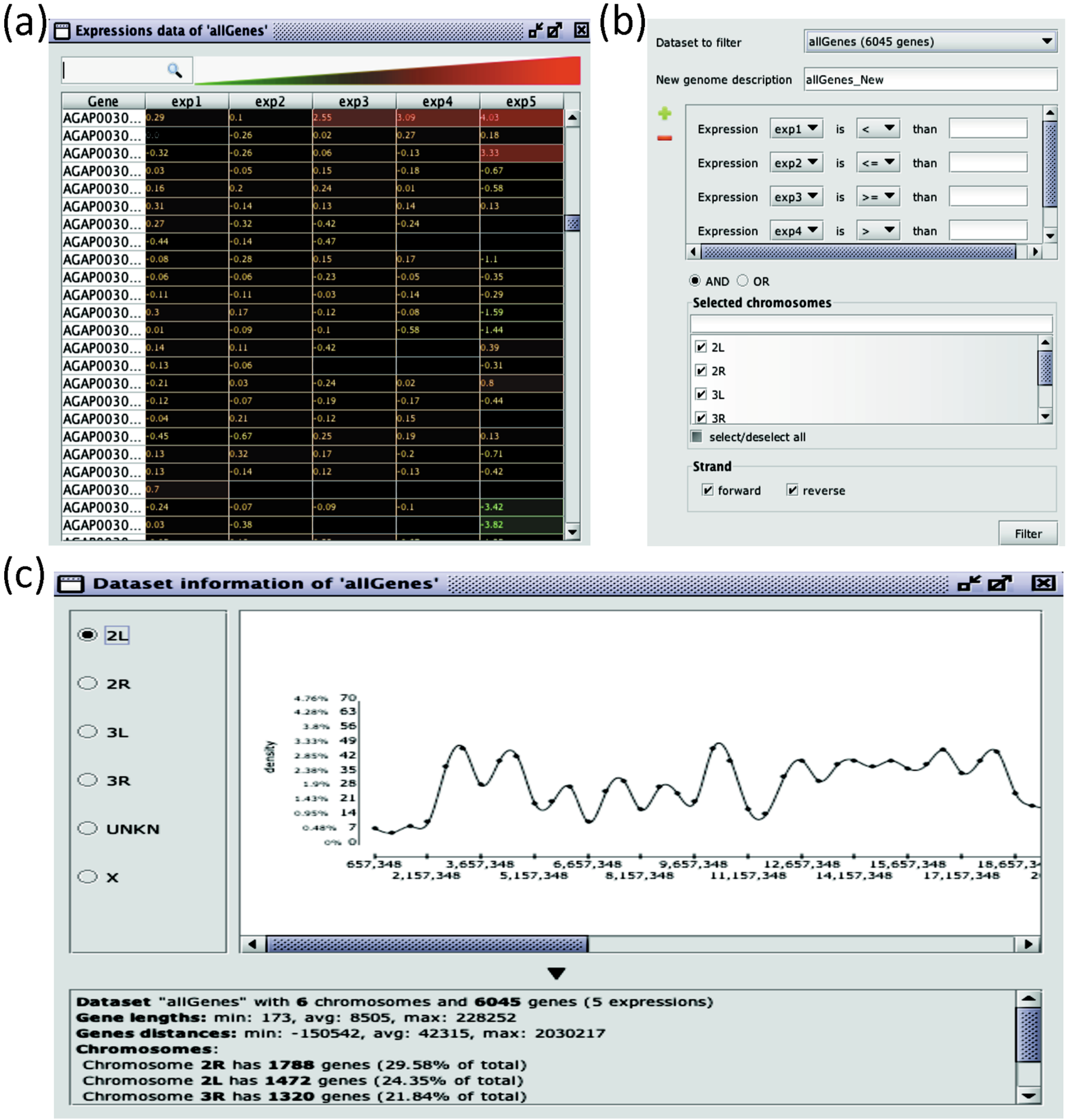
For all the distance metrics implemented by CluGene, the *starting point* of a gene is the starting point of the first exon of any of the transcripts for that gene; the *end point* is the position of the last exon of all the transcripts that belong to that gene. In Biomart, the coordinates do not change for the -1 strand.

The distance metrics have all constant algorithmic complexity O(1), and are the following:

- *positional (default)*: the distance of two genes is the distance between their starting points in the chromosome; where the starting point coordinates are directly taken from the default Biomart output (which does not make any difference between forward and reverse strands) [5];
- *functional*: the distance of two genes is the distance between the positions of their respective promoters; the promoter is physically located at the starting point of the gene; however, to take into account the behaviour of Biomart which reverses the genes on the -1 strand, the coordinates of the end point of those genes are used as the promoter positions;
- *centre of genes*: the distance of two genes is the distance of their two centres;
- *intragenes*: the distance of two genes is the minimum distance between them (i.e. the distance between the starting point of the second and the ending point of the first)

2.3.2. Gene distance weighting strategy. Weighting strategies can be used to modify the distance metric so that it takes into account also other gene attributes. The weighting strategies implemented by CluGene are the following:

- *uniform weighting:* weights are not used.
- *density based:* this technique tries to capture the concept that a simple gene distance measured in bp (base pairs) may have different functional relevance depending on local gene density. See again Fig. 1c for an example plot of a density function for a chromosome.
- *expression based:* the overall similarity of two genes depends on both their actual proximity on the chromosome, and on the similarity of their expression levels. Two options are available for considering expression similarity. When multiple expressions are available for each gene, and these can be assumed to have a normal distribution within each set, the ANOVA can be used to test the equality of the means among the different groups. The p-value resulting from the test is complemented and used as a weight. Alternatively, a custom weighting function has been designed that works both with multiple and single expression values.



Expressions data of 'allGenes'

Gene	exp1	exp2	exp3	exp4	exp5
AGAP0030...	0.29	0.1	2.55	3.09	4.03
AGAP0030...	0.0	-0.26	0.02	0.27	0.18
AGAP0030...	-0.32	-0.26	0.06	-0.13	3.33
AGAP0030...	0.03	-0.05	0.15	-0.18	-0.67
AGAP0030...	0.16	0.2	0.24	0.01	-0.58
AGAP0030...	0.31	-0.14	0.13	0.14	0.13
AGAP0030...	0.27	-0.32	-0.42	-0.24	
AGAP0030...	-0.44	-0.14	-0.47		
AGAP0030...	-0.08	-0.28	0.15	0.17	-1.1
AGAP0030...	-0.06	-0.06	-0.23	-0.05	-0.35
AGAP0030...	-0.11	-0.11	-0.03	-0.14	-0.29
AGAP0030...	0.3	0.17	-0.12	-0.08	-1.59
AGAP0030...	0.01	-0.09	-0.1	-0.58	-1.44
AGAP0030...	0.14	0.11	-0.42		0.39
AGAP0030...	-0.13	-0.06			-0.31
AGAP0030...	-0.21	0.03	-0.24	0.02	0.8
AGAP0030...	-0.12	-0.07	-0.19	-0.17	-0.44
AGAP0030...	-0.04	0.21	-0.12	0.15	
AGAP0030...	-0.45	-0.67	0.25	0.19	0.13
AGAP0030...	0.13	0.32	0.17	-0.2	-0.71
AGAP0030...	0.13	-0.14	0.12	-0.13	-0.42
AGAP0030...	0.7				
AGAP0030...	-0.24	-0.07	-0.09	-0.1	-3.42
AGAP0030...	0.03	-0.38			-3.82

Dataset to filter: allGenes (6045 genes)

New genome description: allGenes_New

Expression exp1 is < than

Expression exp2 is <= than

Expression exp3 is >= than

Expression exp4 is > than

● AND ○ OR

Selected chromosomes

☑ 2L
☑ 2R
☑ 3L
☑ 3R
□ select/deselect all

Strand

☑ forward ☑ reverse

Filter

Dataset information of 'allGenes'

● 2L
○ 2R
○ 3L
○ 3R
○ UNKN
○ X

Dataset "allGenes" with **6** chromosomes and **6045** genes (5 expressions)
Gene lengths: min: 173, avg: 8505, max: 228252
Genes distances: min: -150542, avg: 42315, max: 2030217
Chromosomes:
Chromosome **2R** has **1788** genes (29.58% of total)
Chromosome **2L** has **1472** genes (24.35% of total)
Chromosome **3R** has **1320** genes (21.84% of total)

Figure 1. Expression-based gene dataset processing and Position-based gene dataset processing. (**a**). An example visualization of genes with multiple expressions is shown. Expression levels are depicted by the means of different colours (black = 0, gray = not available, from green to red = increased expression values). Data are referring to transcription regulation of sex-biased genes during ontogeny in the malaria vector *Anopheles gambiae* [74]. The 5 different expression datasets are referring to (1st, 2nd, 3rd larvae stages, pupae and adults). (**b**). An example interactive dialog is shown for gene filtering based on expression level thresholding and Boolean combination of expressions. (**c**). An example interactive window is shown where a chromosome is rendered along with the local gene density plot. Different functions are available for retrieving gene position in the chromosome and their lengths; for retrieving the density of genes along each chromosome, for computing what percentage of a chromosome is represented by a gene dataset; for computing the min, max and average lengths of the genes populating a chromosome; the min, max and average distance between consecutive genes in a chromosome, etc. The dataset considered in this example is the *Anopheles Gambiae* genome.
doi:10.1371/journal.pone.0066196.g001

- *transcription factor (TF) based*: promoters recognized by the same TFs significantly affect the overall similarity of two genes. To implement this behaviour, the list of TFs for each gene is retrieved, then the number of shared TFs between the two genes is computed (by intersecting the two lists) and transformed into a weighting factor.
- *user-defined:* a plug-in mechanism implemented within CluGene allows users to define their own weighting functions, which can

make use of any available gene information (expression, TFBS similarity, chromosome local density, etc.).

2.3.3. Clustering algorithm. The last parameter that can be selected for gene positional clustering is the type of clustering algorithm. Three options are provided by CluGene:

- Clustering by sequential gene processing
- K-means clustering
- Max-gap Clusters by Multiple Sequence Comparison

In clustering by sequential gene processing, genes are scanned sequentially according to the order they are located in the chromosome. For each gene, its (weighted) distance to the previous is computed. If such distance is below a predefined threshold, then the gene is added to the same cluster of the previous one, otherwise it forms a new cluster. It is important to point out that the threshold is applied to the weighted distance, therefore it must be chosen with care to keep into account also the effects of the weights. This clustering algorithm is the evolution of a previous implementation by the authors [5]. Since genes are ordered by position in the chromosome it is sufficient for the algorithm to check only one pair of adjacent genes at a time, and so the complexity of this algorithm is O(n) where n is the number of genes, the complexities of the distance metric evaluation and distance weighting function should be added to that in order to obtain a final figure of the overall complexity of the procedure. As a final consideration, it should be noted that one of the combinations available within CluGene, namely positional clustering with density-based weighting, shares many similarities with general-purpose density-based clustering algorithms such as DBSCAN [36] and in particular OPTICS [37]. However, the positional clustering algorithms implemented within CluGene can take full advantage of the fact that distances (and density) are computed over a one-dimensional space (i.e. along the chromosome), which ultimately leads to an algorithmic complexity of O(n), a significant improvement over the typical O(n^2) of DBSCAN and OPTICS.

The second clustering algorithm available in CluGene is K-means clustering. The k-means algorithm is a popular solution for partitioning the items of a dataset into k clusters based on similarity of item attributes. For gene clustering, once a proper distance metric (and weighting function, if required) is defined, the k-means algorithm can be used straightforwardly.

Differently from clustering by sequential gene processing, in this case individual chromosomes are not scanned, and no distance threshold must be manually defined. Typical of k-means algorithms, the operator is asked to input the number of desired clusters k that should be located in the dataset. In addition, CluGene provides also an implementation of the method by [38], which is used to predict the optimal number of clusters, which k-means should search for. After having set the number k of clusters to be found, the k-means algorithm consists of an iterative process where gene groups are initially formed and increasingly refined. The process starts from an initial assessment regarding the positions of the cluster k centroids: at each iteration all the genes are assigned to a cluster, then the centroids are updated according to the updated cluster contents, and the process is repeated until either the centroids have converged to stable positions, or until the maximum number of allowed iterations is reached. The most critical aspect of k-means is in the initial guess of cluster centroids; for this reason, multiple solutions are implemented in CluGene, as listed in the following:

- *random* [39]: k random points are chosen among all valid genes positions;
- *uniform*: k points uniformly distributed on the chromosome are chosen, even if not assigned to valid genes;
- *plusplus* [40]: to spread the k initial centroids away from each other, the first is chosen from a uniform random distribution on gene positions, and each other one is chosen from the remaining positions with probability proportional to its squared distance to the closest existing centroid.

If the plusplus initialization is chosen, the k-means algorithm is guaranteed to find a solution that is O(log k) competitive to the optimal k-means solution.

The third clustering algorithm available in CluGene implements the method known as: Max-gap Clusters by Multiple Sequence Comparison. This algorithm implements the proposal originally from Xu Ling, Xin He and Dong Xin [21]. Our version of algorithm works on pairs of genomes which have orthologous genes in common. Information about orthologous genes is retrieved from Biomart. Given a dataset of one of the supported genomes the user can specify three parameters:

- *Maxgap*: maximum number of non-orthologous genes between two genes in a cluster;
- *Minsize*: minimum size (number of genes) in a cluster;
- *Comparison organism*: the other organism type on which orthologous should be computed.

The algorithm is based on the following steps:

- all orthologous genes, ordered by chromosome position, are computed and genes without an orthologous are removed;
- for each pair of genes in the dataset, they belong to the same cluster if and only if the distance between the two genes computed as the number of non-orthologous genes between them is lower or equal to *maxgap;*
- all clusters with less than *minsize* genes are discarded.

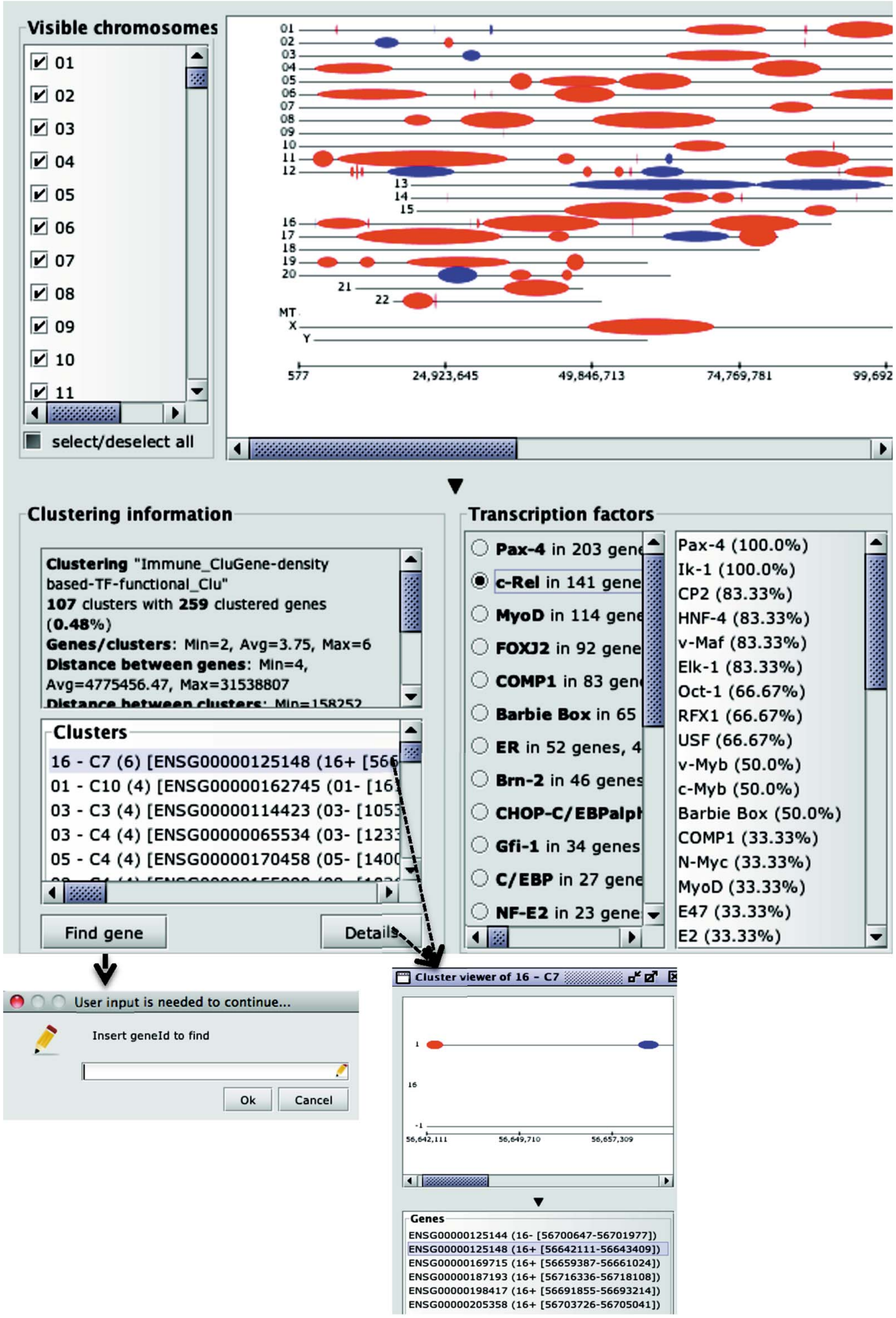
Currently the Maxgap method within CluGene only supports the genomes of *Anopheles Gambiae* and *Drosophila Melanogaster*.

An exemplifying cluster analysis session within CluGene is illustrated in Figure 2, based on sequential gene processing, functional distance metric and a weighting function combining local density and TF similarity.

2.4. Analysis of Statistical Significance of Gene Clusters

CluGene offers functions for post-processing the clustering results in order to highlight significant aspects. Two main alternatives are provided: *comparison to a random set*, and *neighbourhood model and clustering statistics*.

2.4.1. Comparison to a random set. The clustering results obtained from the available dataset are compared with the results of the same clustering process applied to a random set of genes. The goal is to see whether the clustering of the available genes produces a number of clusters which is significantly different than what could be obtained by aggregating random genes. To perform this analysis, CluGene needs to produce an alternative set containing a random set of genes, as large as the original set (size: n). The random set is created by randomly extracting n genes from a user-selected set (size>n). No constraints are imposed on the choice of such set, which may or may not even contain some or all the genes of the dataset being scrutinized. Once the random set is ready, CluGene runs the same clustering analysis that was performed onto the original dataset and the entire process



Clustering "Immune_CluGene-density based-TF-functional_Clu"
107 clusters with **259** clustered genes (**0.48%**)
Genes/clusters: Min=2, Avg=3.75, Max=6
Distance between genes: Min=4, Avg=4775456.47, Max=31538807
Distance between clusters: Min=158252

Clusters
16 - C7 (6) [ENSG00000125148 (16+ [566
01 - C10 (4) [ENSG00000162745 (01- [16
03 - C3 (4) [ENSG00000114423 (03- [105
03 - C4 (4) [ENSG00000065534 (03- [123
05 - C4 (4) [ENSG00000170458 (05- [140

○ **Pax-4** in 203 gene
● **c-Rel** in 141 gene
○ **MyoD** in 114 gene
○ **FOXJ2** in 92 gene
○ **COMP1** in 83 gen
○ **Barbie Box** in 65
○ **ER** in 52 genes, 4
○ **Brn-2** in 46 genes
○ **CHOP-C/EBPalph
○ **Gfi-1** in 34 genes
○ **C/EBP** in 27 gene
○ **NF-E2** in 23 gene

Pax-4 (100.0%)
Ik-1 (100.0%)
CP2 (83.33%)
HNF-4 (83.33%)
v-Maf (83.33%)
Elk-1 (83.33%)
Oct-1 (66.67%)
RFX1 (66.67%)
USF (66.67%)
v-Myb (50.0%)
c-Myb (50.0%)
Barbie Box (50.0%)
COMP1 (33.33%)
N-Myc (33.33%)
MyoD (33.33%)
E47 (33.33%)
E2 (33.33%)

User input is needed to continue...
Insert geneId to find

Genes
ENSG00000125144 (16- [56700647-56701977])
ENSG00000125148 (16+ [56642111-56643409])
ENSG00000169715 (16+ [56659387-56661024])
ENSG00000187193 (16+ [56716336-56718108])
ENSG00000198417 (16+ [56691855-56693214])
ENSG00000205358 (16+ [56703726-56705041])

Figure 2. An example of cluster analysis session within CluGene, based on sequential gene processing, functional distance metric and a weighting function combining local density and TF similarity. Clusters are visualized as oval-shapes calculated on the different chromosomes; information concerning the number of clusters and the number of clustered genes is also provided, along with the minimum, average and maximum number of genes present in the clusters, and the minimum, average and maximum distance (bp) between clustered genes and clusters. Detailed information of the distribution of cluster sizes is provided as well. For each cluster, it is possible to visualize the number of contained genes, together with their location and the list of TFs with their associated frequencies. Besides, it is possible to visualize, for entire clustered dataset, the calculated TFs together with associated frequencies. For each TF it is possible to identify the number of recognized target genes and the relative number of clusters. Once a TF is selected, the clusters positive for the specified TF are interactively highlighted in red whereas negative clusters remain blue (default colour).

doi:10.1371/journal.pone.0066196.g002

(selection of random genes and clustering) can be repeated multiple times. Once the process is finished, the results from clustering of random genes are aggregated and compared with the "real" clustering results. The comparison is done through hypothesis testing: CluGene provides the one-sample t-Student test to compare number of clusters. Ideally, for the "real" clustering to be significant, a different number of clusters should be generated in the "real" case with respect to the random case, and such difference would be statistically significant (low p-value). In its current implementation, CluGene can only assess if the number of clusters in the test case is significantly larger than the random case, and if such difference is statistically significant (one-tailed, one-sample t-test). This approach may be limited, and may fail to flag as interesting those cases where the opposite happens (e.g. gene sets belonging to the same functional pathway may be more likely be localized within the same cluster). Future CluGene implementations will consider the adoption of a two-tailed test to assess significant differences, regardless of the sign. Also, a more complete offering is planned for future implementations of CluGene, which includes non-parametric tests, as equally viable and sometimes preferable alternatives to parametric tests such as the one-sample t-test. In the current implementation, any other test which is not the one-sample t-test can still be performed, but this requires manually exporting the CluGene results into separate software applications.

2.4.2. Neighbourhood model and clustering statistics. CluGene implements an additional algorithm to test statistical significance of the computed clusters, as originally defined by Li et al [23]. In a sense, this is similar to the previous comparison to a random set. For each chromosome, the mean number of adjacent genes in a random arrangement and the standard deviation is computed.

Results

In order to study the capabilities of CluGene we performed two different analyses (case studies) testing the different clustering algorithms and analysis tools of CluGene. Organisms of the genuses *Homo* and *Plasmodium* were used in the validation process.

3.1. Case Study 1: Co-localization of the Binding Sites of Oct-1, c-REL, NF-KappaB, IK-1, BSAP and CP2

Specific transcription factors contribute to immunoglobulin and other relevant genes expression crucial for B cell development and function. They include octamer-binding transcription factor 2 (Oct-2), B cell Oct-binding protein 1 (BOB.1), B cell–specific activator protein (BSAP), BCL-6, MUM1/IRF4, PU.1 (also known as Spi-1), Oct-1, c-REL, NF-KAPPAB, IK1 and CP2 [41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57]. Unbalance in the synthesis of these transcription factors or the presence of mutations can lead to suboptimal activation of immunoglobulin or of other relevant gene promoters in different B cell tumors [43,58,59,60] [61,62,63]. B-cell transcription factors show diverse expression patterns in different lymphomas (i.e. Hodgkin lymphoma (CHL) from diffuse large B-cell lymphoma (DLBCL) [41] or in non–germinal center small B-cell lymphomas [57] (a heterogeneous group of non-Hodgkin lymphomas).

Being able to carry out a detailed stratification analysis based in the identification of individual combinations of transcription factors would be of great clinical value in terms of diagnosis, prognosis and response to therapy.

To test the capability of CluGene in disclosing networks of co-regulated genes, the human genome data and whole genome arrays from diffuse large B-cell lymphoma (DBLCL) specimens were considered.

3.1.1. Datasets for the diffuse large B-cell lymphoma. In previous work [64], a combined approach of transcriptional profiling and gene set enrichment analysis had highlighted the existence of three distinct DLBCL sets containing a significant number of functionally related genes, referred to as: *oxidative phosphorylation (*OXPHOS), *B-cell receptor/proliferation* (BCR/proliferation) and *host response* (HR) (downloadable from http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi). The human gene IDs in the Ensembl format were obtained using Clone/Gene ID converter [65].

3.1.2. Identification of the main TFBS motifs. We first investigated which of the annotated B cell TFBS motifs were represented in the promoter sequences of the three gene subsets. The putative promoter regions needed for the functional distance metric were retrieved directly from Biomart and correspond to the 1000 bp flanking the 5′ end of each gene. According to the Match^{TM} prediction functionality provided by CluGene the binding sites for transcription factors Oct-1, NF-KAPPA-B, IK-1, c-REL, BSAP and CP2 occurred in the promoters of the three subsets with different frequencies (Figure S2a). In particular, in the OXPHOS subset, the Oct-1 TF showed the highest number of binding sites. For assessing whether the presence of such a high number of binding sites had a functional significance, we compared the result with the number of Oct-1 TFBS found in a dataset of genes randomly extracted from the human genome. The random gene set (same size of the original dataset) was generated through the functions available within CluGene, and analyzed with the same Match^{TM} prediction function. All the results were exported from CluGene in order to apply the one-sample Student's t-test provided by a commercial statistical software application (SPSS) to test for significant differences between the study and random datasets. The test showed that the higher number of Oct-1 TFBS found for the real dataset was significantly different from the amount of binding sites found for the random set (Figure S2b).

3.1.3. Dataset clustering behavior under the application of different distance metrics. To begin analyzing the relationships between co-regulation and co-localization, a first investigation was aimed at identifying interesting clustering behaviors of the datasets subjected to sequential clustering, under the application of three main distance metrics implemented within CluGene:

- first metric: positional with no weights;
- second metric: functional (promoter distance), weighted by local density;
- third metric: functional (promoter distance), weighted by local density and TF similarity.

A threshold value of 83,000 bp [66] on the distance metric was adopted as the rule for cluster formation for all strategies. The number of identified clusters was tested for statistical significance with one-sample t-test (through a function available within CluGene) against a random set of genes. Since the population may be nonnormal the power of the one-sample t-test may be slightly reduced, albeit not that much, given the sample size. The third metric generated a significantly larger number of clusters compared to the random set (Figure 3a), this applied consistently for all three datasets. Also, CluGene post-process analysis of the clusters produced with the third metric showed that a high percentage of genes targeted by the selected list of TFs are co-localized within the genome (Figure 3b).

(a)

Clustering procedure	Number of clusters in the OXPHOS dataset	Number of clusters compute d from the random dataset (avg.) [a]	One sample T test- t value [b]	One sample T test Sig. (one-tailed) [c]	Number of clustered genes in the OXPHOS dataset	Number of clustered genes in the random dataset (avg.) [a]	One sample T test- t value [b]	One sample T test Sig. (one-tailed) [c]	Conclusion [d]
Positional clustering	22	12.1	-13.07	<0.02	45	24.8	-13.49	<0.02	obs>rnd
Functional clustering LD	101	93.05	-5.18	<0.02	237	230.05	-1.90	>0.02	rnd>obs
Functional clustering LD-TF	72	15.65	-100.04	<0.02	179	34.55	-111.74	<0.02	obs>rnd

Clustering procedure	Number of clusters in the BRC dataset	Number of clusters computed from the random dataset (avg.) [a]	One sample T test- t value [b]	One sample T test Sig. (one-tailed) [c]	Number of clustered genes in the BRC dataset	Number of clustered genes in the random dataset (avg.) [a]	One sample T test- t value [b]	One sample T test Sig. (one-tailed) [c]	Conclusion [d]
Positional clustering	15	13.6	-1.69	>0.02	34	28	-3.71	<0.02	obs=rnd
Functional clustering LD	84	104.39	13.24	<0.02	200	253.6	18.18	<0.02	rnd>obs
Functional clustering LD-TF	67	19.9	-60.04	<0.02	153	43.7	-69.48	<0.02	obs>rnd

Clustering procedure	Number of clusters in the HR dataset	Number of clusters compute d from the random dataset (avg.) [a]	One sample T test- t value [b]	One sample T test Sig. (one-tailed) [c]	Number of clustered genes in the HR dataset	Number of clustered genes in the random dataset (avg.) [a]	One sample T test- t value [b]	One sample T test Sig. (one-tailed) [c]	Conclusion [d]
Positional clustering	41	21.95	-21.62	<0.02	92	45.95	-23.93	<0.02	obs>rnd
Functional clustering LD	121	125.9	3.33	<0.02	283	298.3	3.97	<0.02	rnd>obs
Functional clustering LD-TF	107	25.6	-103.11	<0.02	259	56.3	-122.39	<0.02	obs>rnd

[a] At each run, CluGene randomly extracted genes to from a random dataset the same size of the OXPHOS dataset, and did the transcription factors analysis on such set. The results were averaged over 20 runs. The random gene selection was done on the entire human genome as retrieved by Biomart.

[b,c] One-sample Student's t-test statistics obtained with CluGene.

[d] obs>rnd: the n. of clusters computed from the real dataset (observed) is larger than the average number of clusters computed from the random (rnd) datasets; rnd>obs: the opposite is true.

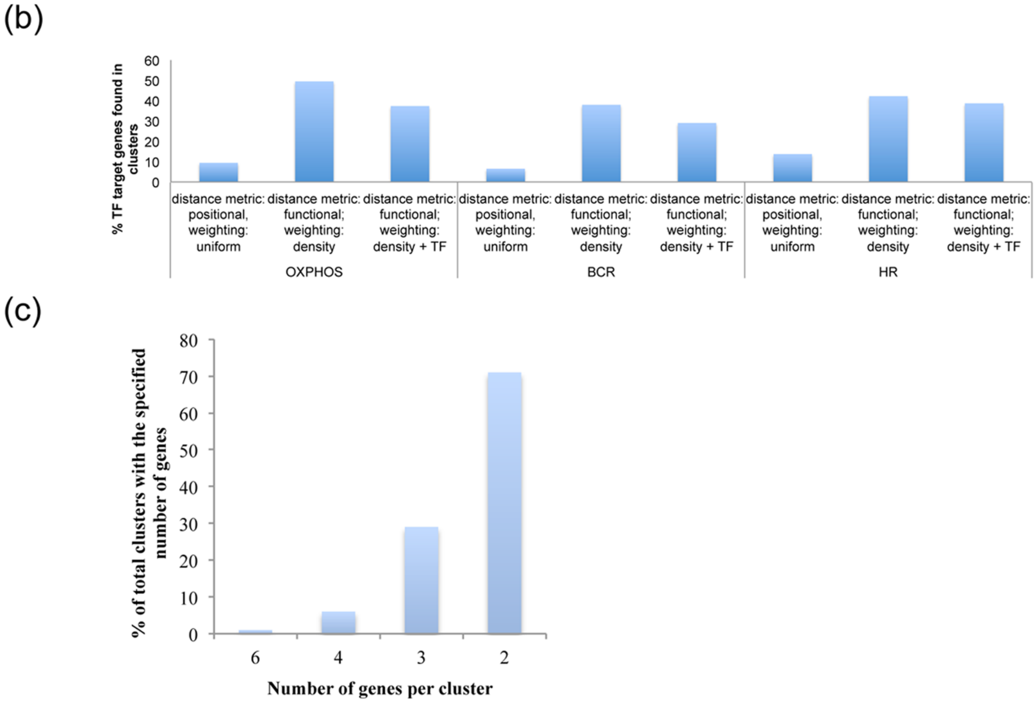


Figure 3. Different clustering behaviours of the DLBCL subsets subjected to sequential clustering and different distance metrics. (**a**). The one sample t-test was used to validate the statistical significance of the number of clusters found in the test datasets, vs. number of clusters found in random datasets. (**b**). % target genes found in clusters with: simple positional distance metric, functional with local density and functional with local density and TF-similarity. (**c**) A histogram showing the frequency distribution of clusters is shown.
doi:10.1371/journal.pone.0066196.g003

The application of the third metric highlighted also that in the HR subset there is the highest percentage of genes (45.17%), which can be found in large clusters (defined as clusters containing three or more genes). A histogram showing the frequency distribution of clusters is shown in figure 3c. This preliminary result may indicate that the HR subset is under transcriptional co-regulation by the same set of TFs.

To better investigate the benefits of utilising the third metric (promoter distance, weighted by local density and TF similarity), we considered the identified B cell transcription factors (Oct-1, NF-KAPPA B, IK-1, c-REL, BSAP and CP2) and the co-localization of their cognate target genes along the genome. Notably, a higher amount of co-localized positive annotated B-cell transcription factors target genes was identified with the third metric (Figure 4) compared to the application of simple positional clustering (first metric). In particular, with slight variations depending on the subset, the percentage of genes, originally predicted as targets for the TFs and found in clusters obtained with the third metric reached 70%, while the same percentage obtained through simple positional clustering was not higher than 18.8%. This result suggests that DLBCL co-regulated genes are physically associated in clusters. Such clusters are more subtly defined than what can be detected by means of proximity alone. The percentages of TF target genes found in the clusters are shown in Figure S3.

3.1.4. Investigations on the co-localization of the binding sites of the selected TFs. We also investigated whether Oct-1, c-REL, NF-KappaB, IK-1, BSAP and CP2 co-localize themselves within the same genomic sites, which may imply a possible co-operative regulatory effect. Again by using the functionality provided within CluGene, we investigated if specific pairs of TFs would co-occur more frequently than others within clusters. As a starting point, the clustering results obtained with the more promising third distance metric were considered. The analysis specifically highlighted a statistically more frequent co-occurrence of Oct-1 with c-REL and CP2 as judged by the contingency tables created to compare any two pairs in terms of number of clusters featuring them (Figure 5); contingency tables were created to compare any two pairs of genes in terms of number of clusters featuring them, to see whether specific pairs of TFs appeared more often within the same clusters. The contingency tables were created using the GraphPad QuickCalcs Web site (http://www.graphpad.com/quickcalcs/Contingency1.cfm). However the test for statistical significance of this result, which was based on running within Clugene the same type of analysis on clusters obtained from random genes, did not pass. While it is yet unknown if the observed associations might ultimately provide a contribution to the understanding of B-cell lymphomas pathogenesis, it is interesting to note that the identified co-occurrences include genes which have been already recognized as being part of the same functional pathways in large B-cell lymphoma, and this finding applies to all three different DLBCL subsets.

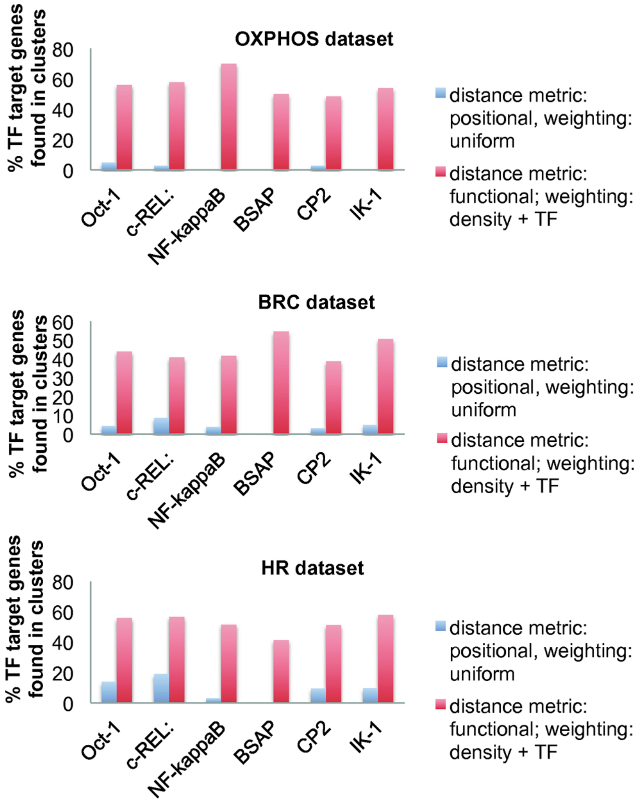
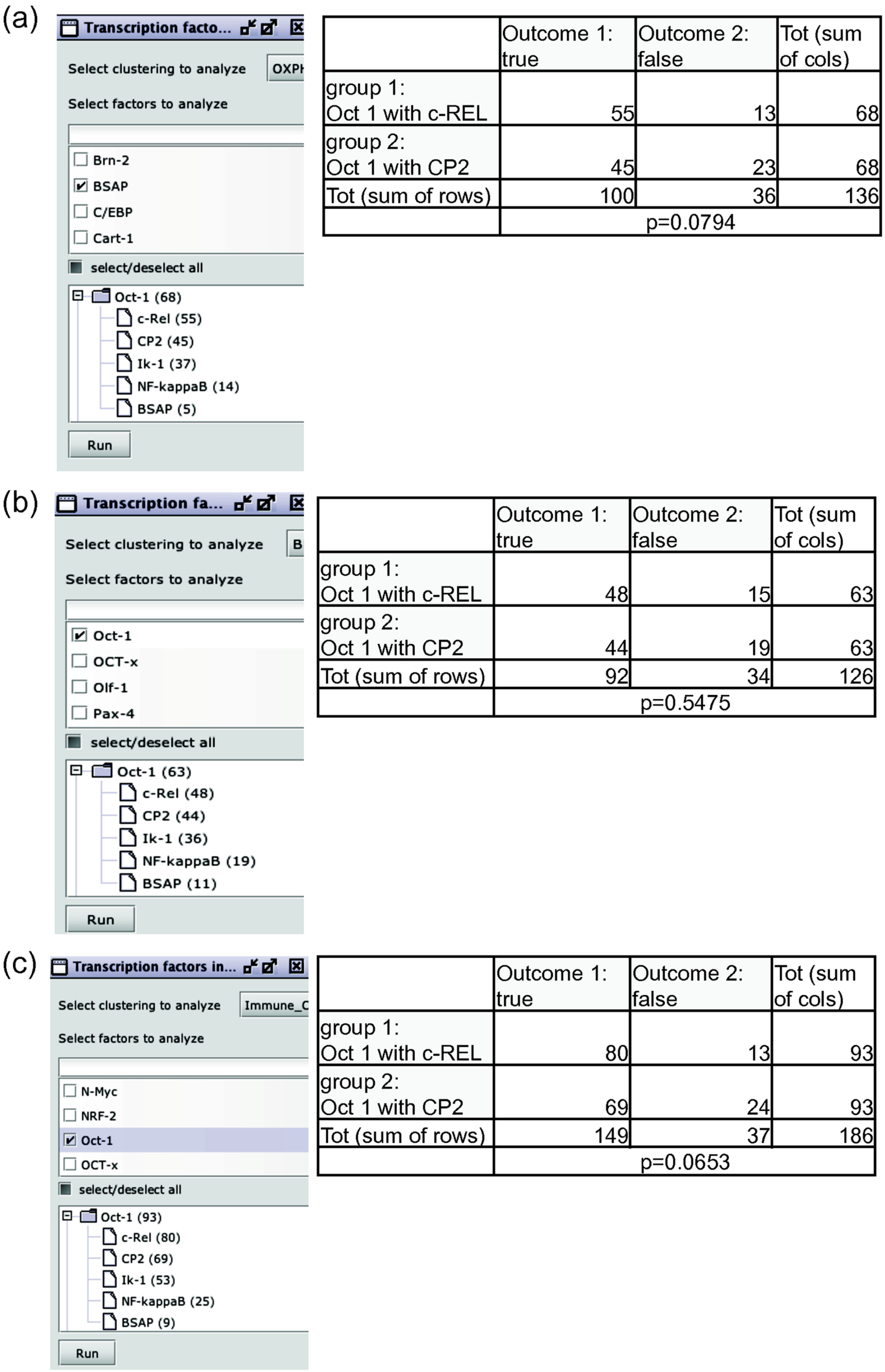


Figure 4. Percentage of of co-localized positive annotated B-cell transcription factors. Percentage of Oct-1, NF-KAPPA B, IK-1, c-REL, BSAP and CP2 target genes found in clusters obtained with the functional distance metric weighted by local gene density and TF similarity and with the simple positional distance metric, out of the total amount of recognized TF target genes.
doi:10.1371/journal.pone.0066196.g004

3.1.5. Identification of specific diagnostic markers of DLBCL. Additional investigations were made on the datasets with the ultimate aim of providing support to the studies addressing the identification of specific diagnostic markers of DLBCL. As pointed out by Monti and co-authors [64], the HR subset represents an interesting source of potential treatment targets as a consequence of the role of the immune response featuring this DBLCL subset. To this aim we have applied the same distance metric illustrated above (promoter-based distance metric with weights based on local density and TF similarity) to the genes of the HR subset, and then we used CluGene functions to compute the average number of TFBSs located within the promoters of the clustered genes. Interestingly, this analysis indicated that both Oct-1 and NF-Kappa B had -on average- more than two regulatory cognate site regions per promoter (Figure 6a). Then, by investigating the percentage of clustered HR genes, which had also the signature of immune/inflammatory response, we were able to estimate the number of immunity-annotated genes that mapped to proximal chromosomal locations (figure 6b). Interestingly almost 40% of the immune markers appeared to be co-localized.

Finally, we examined whether co-regulated and co-localized genes in the HR subset would also show similar expression levels, which may indicate similar transcriptional activity.

Previous literature work had outlined that co-localized, TF target genes are subjected to tighter co-regulation compared to



(a)

Transcription facto...

Select clustering to analyze: OXPI

Select factors to analyze

- ☐ Brn-2
- ☑ BSAP
- ☐ C/EBP
- ☐ Cart-1

■ select/deselect all

- Oct-1 (68)
  - c-Rel (55)
  - CP2 (45)
  - Ik-1 (37)
  - NF-kappaB (14)
  - BSAP (5)

Run

	Outcome 1: true	Outcome 2: false	Tot (sum of cols)
group 1: Oct 1 with c-REL	55	13	68
group 2: Oct 1 with CP2	45	23	68
Tot (sum of rows)	100	36	136
	p=0.0794		

(b)

Transcription fa...

Select clustering to analyze: B

Select factors to analyze

- ☑ Oct-1
- ☐ OCT-x
- ☐ Olf-1
- ☐ Pax-4

■ select/deselect all

- Oct-1 (63)
  - c-Rel (48)
  - CP2 (44)
  - Ik-1 (36)
  - NF-kappaB (19)
  - BSAP (11)

Run

	Outcome 1: true	Outcome 2: false	Tot (sum of cols)
group 1: Oct 1 with c-REL	48	15	63
group 2: Oct 1 with CP2	44	19	63
Tot (sum of rows)	92	34	126
	p=0.5475		

(c)

Transcription factors in...

Select clustering to analyze: Immune_C

Select factors to analyze

- ☐ N-Myc
- ☐ NRF-2
- ☑ Oct-1
- ☐ OCT-x

■ select/deselect all

- Oct-1 (93)
  - c-Rel (80)
  - CP2 (69)
  - Ik-1 (53)
  - NF-kappaB (25)
  - BSAP (9)

Run

	Outcome 1: true	Outcome 2: false	Tot (sum of cols)
group 1: Oct 1 with c-REL	80	13	93
group 2: Oct 1 with CP2	69	24	93
Tot (sum of rows)	149	37	186
	p=0.0653		

Figure 5. Analysis of co-operative regulatory effect of Oct-1, c-REL, NF-KappaB, IK-1, BSAP and CP2 in shared DLBCL genomic sites. Co-operative regulatory regions in target genes were identified using the functional distance metric weighted by local density and TF similarity. The panels refer to an investigation aimed at identifying TFs appearing more often together in clusters. The values in the contingency tables (right panels) are the frequencies of clusters featuring either both TFs of a given pair (true), or only one of them (false). The tables contain also the p-values of the Fisher's exact test, indicating whether there is some statistically significant difference between the co-occurrences of TF pairs within clusters. Panels refer specifically to the (a). Oxphos; (b). BRC and (c). HR subsets. Only statistically significant results are shown. These results emphasize the role of c-REL, CP2 and Oct-1 association in shared diffuse large B-cell lymphoma regulatory regions.
doi:10.1371/journal.pone.0066196.g005

isolated genes, and genes being co-regulated by common TFs tend to have a similar expression profile [18,35,67].

For each DLBCL transcription factor identified as illustrated above, we first separated co-localized genes being recognized by that TF (test group) from those that although being co-localized were not recognized by it (control group). Since some of the expression profiles for the two groups were nonnormal, for each TF we compared the gene expression profiles of the two groups using the Mann-Whitney U test at a threshold of $p<0.05$ (SPSS Statistics v.17.0).

Notably, co-localized genes containing both Oct-1 and IK-1 binding sites showed statistically significant difference in the expression patterns (higher expressions) when compared to the control group (Mann-Whitney U test $p_{Oct-1} = 0.03$; Mann-Whitney U test $p_{IK-1}<0.01$) (Figure 7a). This behaviour was not shared by the other TFs (NF-KAPPAB, c-REL, BSAP and CP2), which did not show expression differences between the two groups.

To investigate whether gene co-expression was a consequence of either sharing specific transcription factor binding sites; or being in chromosomal proximity, we compared the expression profiles of Oct-1 target genes that showed co-localization patterns with those that did not show such pattern by using the Mann-Whitney U test at a threshold of $p<0.05$. Notably, Oct-1 co-localized target genes differed in their expression –higher expression- from those that were not co-localized (Mann-Whitney U test $p = 0.017$) (Figure 7b).

Taken all together, these results suggest that co-localized genes are more likely to be co-regulated than isolated genes. As the observed correspondence between co-regulation and co-localization was not shared by all examined TFs, we expect that not all annotated TFs have a stringent or relevant role in DLBCL or either that the changes in expression patterns are not sufficient to be detected.

As a side note, it should be pointed out that another valid non-parametric test that could have been adopted for the investigations illustrated above is the Kolmogorov-Smirnov (KS) test. While the KS test is capable of detecting a wider array of differences in the shape and localization of two distributions, in this specific case, we were only interested in simple comparisons of the means of expression values (i.e. figuring out whether one group would have significantly higher expression patterns than the other), which is why the Mann-Whitney U test was adopted.

3.2. Case Study 2: Specificity of Genes Containing the Host Targeting (HT) Motif in *P. Falciparum*

To investigate the capabilities of CluGene clustering algorithms, we first analysed the entire genome of *Plasmodium Falciparum* by considering genes having different cellular localization attributes. Genes exhibiting different localization attributes were downloaded from PlasmoDB website (http://plasmodb.org/plasmo/), using the following queries: (i) Genes with signal peptide [68]: genes that are predicted with the of signal peptide cleavage sites in amino acid sequences, predictions are made with the SignalP program; (ii) Genes with transmembrane domains [69]: genes with at least a transmembrane helix in encoded proteins, prediction are made with the TMHMM2 program; (iii) Apicoplast Genes [70]: genes targeted to the apicoplast membrane, a derived non-photosynthetic plastid found in most Apicomplexa; (iv) Genes with Host-targeting (HT) motif [71]: genes with sequence motifs identified by Hiller and co-workers [71] as being associated with proteins secreted to the human erythrocyte. Proteins secreted to the host erythrocyte represent important modulators of the antigenic and adhesive modifications in the infected erythrocytes [71].

To compare results from different research analyses and from different Plasmodial species, we also considered an additional dataset [72] containing genes of *P. Falciparum* and *Plasmodium Berghei* exhibiting the HT motif.

First, all downloaded dataset were imported into CluGene system and then CluGene's visualization tools were used to compute the gene densities present across all chromosomes in the different datasets. All datasets were then clustered using the default options: positional clustering algorithm with constant threshold, using a positional gene distance strategy (gene distance defined as the distance between genes' starts) and subsequently the obtained clustered datasets were compared to each other. From the output of CluGene analysis for PlasmoDB datasets (genes containing signal peptide, genes containing transmembrane domains, genes containing the Apicoplast internalization signal and *P. Falciparum* genes containing the HT motif) we found that gene density is almost linear for all chromosomes and that genes are distributed almost uniformly on all chromosomes with the exception of genes containing the HT motif (see Figure S4 (a) for density and distribution examples). The gene density of genes containing the HT motif is always higher at chromosome ends and almost zero in the centre of chromosomes on all chromosomes. HT genes, in fact, are mostly distributed at chromosome ends (see Figure S4 (b)), when computing a positional clustering with CluGene we observed a co-localization of 192 clustered genes in 36 clusters (90% of the total number of genes, see Figure S4 (c)).

The significance of this clustering was assessed against generated random clusters that resulted in having at most 50 genes in 30 clusters (just 23% of the total number of genes).

When we analysed the *P. Falciparum* and *P. Berghei* HT datasets described by Sargeant [72], for comparison, we found a similar clustering behaviour for both *P. Falciparum* and *P. Berghei* gene datasets (data not shown). CluGene clustering features and visualization tools helped in identifying this inter-species clustering behaviour in a very simple way.

Discussion

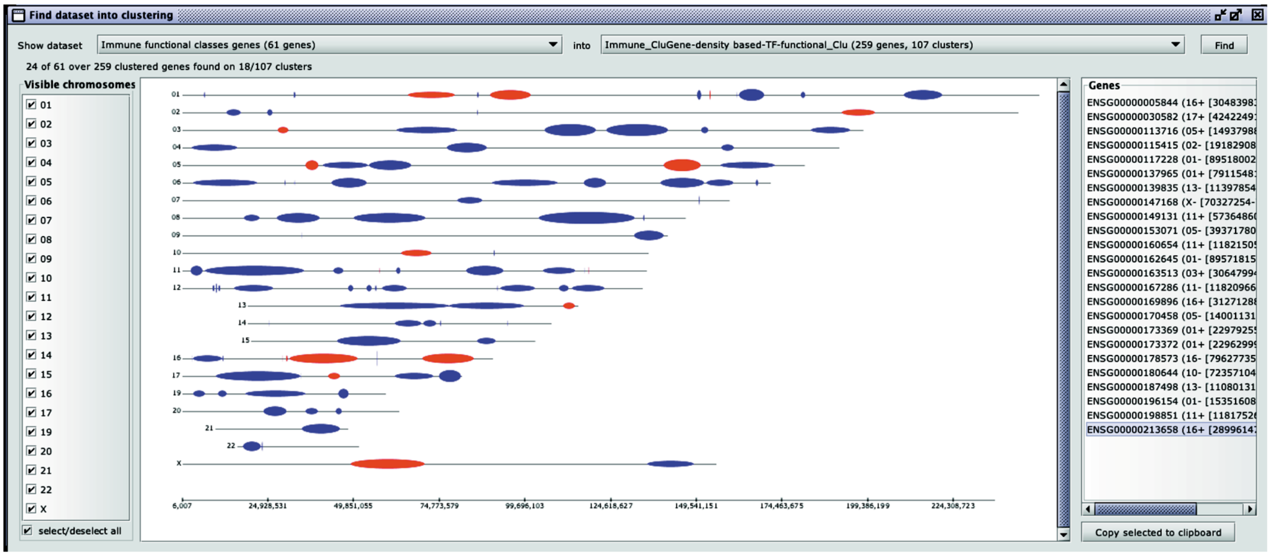
The importance of adding positional information to the data generated by high-throughput techniques, such as histone marks, chip-seq, and DNAseI hypersensitivity, is already recognized. Adding gene localization information could also provide benefit to projects such as the Encyclopedia of DNA Elements (ENCODE) project [73].

Despite the benefits of positional information, obtaining a comprehensive picture of transcriptional regulatory networks is still beyond our reach. Additional insight can be gained by studying co-occurrences, i.e. events involving activation/deactivation of clusters of genes that not only are localized within the same

(a)

Oct-1		c-REL		IK-1		NF-KAPPA B		BSAP		CP2	
Total number of TF	*Average TF per site*	*Total number of TF*	*Average TF per site*	*Total number of TF*	*Average TF per site*	*Total number of TF*	*Average TF per site*	*Total number of TF*	*Average TF per site*	*Total number of TF*	*Average TF per site*
183	**2.91**	141	1.56	85	1.22	34	**2.73**	12	1.16	114	1.29

(b)



Find dataset into clustering

Show dataset: Immune functional classes genes (61 genes) into Immune_CluGene-density based-TF-functional_Clu (259 genes, 107 clusters) [Find]

24 of 61 over 259 clustered genes found on 18/107 clusters

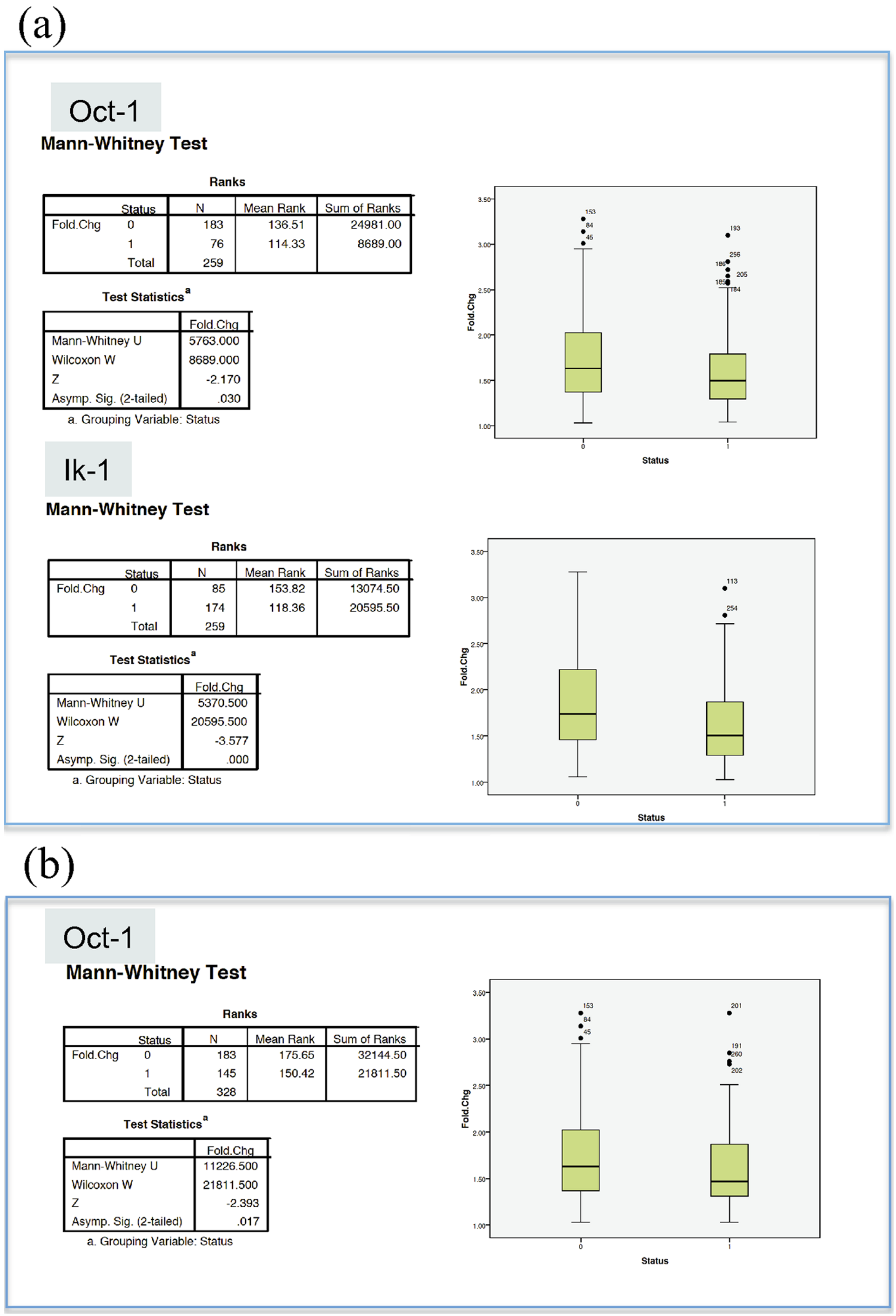
Genes

ENSG00000005844 (16+ [3048398
ENSG00000030582 (17+ [4242249
ENSG00000113716 (05+ [1493798
ENSG00000115415 (02- [19182908
ENSG00000117228 (01- [89518002
ENSG00000137965 (01+ [7911548
ENSG00000139835 (13- [11397854
ENSG00000147168 (X- [70327254-
ENSG00000149131 (11+ [5736486
ENSG00000153071 (05- [39371780
ENSG00000160654 (11+ [1182150
ENSG00000162645 (01- [89571815
ENSG00000163513 (03+ [3064799
ENSG00000167286 (11- [11820966
ENSG00000169896 (16+ [3127128
ENSG00000170458 (05- [14001131
ENSG00000173369 (01+ [2297925
ENSG00000173372 (01+ [2296299
ENSG00000178573 (16- [79627735
ENSG00000180644 (10- [72357104
ENSG00000187498 (13- [11080131
ENSG00000196154 (01- [15351608
ENSG00000198851 (11+ [1181752
ENSG00000213658 (16+ [2899614

ENSG00000173372	Complement component 1 qA
ENSG00000173369	Complement component 1 qB
ENSG00000162433	Adenylate kinase 4
ENSG00000137965	Interferon-induced protein 44
ENSG00000117228	Guanylate binding protein 1, interferon-inducible
ENSG00000162645	Guanylate binding protein 2, interferon-inducible (GBP2
ENSG00000162692	Vascular cell adhesion molecule 1
ENSG00000163220	S100 calcium binding protein A9
ENSG00000143546	S100 calcium binding protein A8
ENSG00000196154	S100 calcium binding protein A4
ENSG00000115415	Signal transducer and activator of transcription (STAT1)
ENSG00000196141	Spermatogenesis associated, serine-rich 2-like
ENSG00000163508	Eomesodermin
ENSG00000163513	Transforming growth factor, beta receptor II
ENSG00000168685	Interleukin 7 receptor
ENSG00000082074	FYN binding protein [Source:HGNC Symbol;Acc:4036]
ENSG00000153071	Disabled homolog 2, mitogen-responsive phosphoprotein
ENSG00000170458	CD14 molecule
ENSG00000113716	Colony stimulating factor 1 recepto (CSF-1)
ENSG00000197043	Annexin A6
ENSG00000196743	Annexin A6
ENSG00000150347	AT rich interactive domain 5B (MRF1-like)
ENSG00000180644	Perforin 1
ENSG00000156587	Ubiquitin-conjugating enzyme E2L 6
ENSG00000149131	Serine (or cysteine) proteinase inhibitor, clade G (C1 inhibitor)

ENSG00000198851	CD3e molecule, epsilon (CD3-TCR complex)
ENSG00000167286	CD3D antigen, delta polypeptide
ENSG00000160654	CD3g molecule, gamma
ENSG00000187498	Collagen, type IV, alpha 1
ENSG00000139835	Lysosomal-associated membrane protein 1 (LAMP1)
ENSG00000176953	Nuclear factor of activated T-cells
ENSG00000213658	Linker for activation of T cells (LAT)
ENSG00000132207	SLX1 structure-specific endonuclease subunit homolog A
ENSG00000005844	Integrin, alpha L
ENSG00000103490	PYD and CARD domain containing
ENSG00000169896	Integrin alpha M
ENSG00000083799	Cylindromatosis (turban tumor syndrome)
ENSG00000198373	WW domain containing E3 ubiquitin protein ligase 2
ENSG00000178573	c-MAF
ENSG00000103187	Coactosin-like 1
ENSG00000030582	Granulin
ENSG00000108424	Karyopherin (importin) beta 1
ENSG00000012211	PRICKLE3
ENSG00000155659	V-set and immunoglobulin domain containing 4
ENSG00000147168	Interleukin 2 receptor, gamma

Figure 6. Characterization of the immune/inflammatory response genes of the HR subset. (**a**). Total and average number of TFBSs present in the promoters of genes expressed in the HR subset. Frequencies were calculated on the genes clustered using the functional distance metric weighted by local density and TF similarity. In bold are shown TFs (Oct-1 and NF-KAPPA B) with an average number value exceeding 2 sites/promoter. (**b**). Analysis of localization within clusters of HR genes having the signature of immune/inflammatory response. The *Show dataset into clusters* utility of CluGene (top panel), allows to identify which clusters (red) out of the total (blue) contain a specified list of genes (immune annotated genes). The list of genes contained in clusters and their location is given on the right part of the panel. The tables (bottom) include the identified clustered immunity genes (bold), their function and the size and what other genes are grouped within the same cluster.
doi:10.1371/journal.pone.0066196.g006



Oct-1

Mann-Whitney Test

Ranks

	Status	N	Mean Rank	Sum of Ranks
Fold.Chg	0	183	136.51	24981.00
	1	76	114.33	8689.00
	Total	259		

Test Statistics[a]

	Fold.Chg
Mann-Whitney U	5763.000
Wilcoxon W	8689.000
Z	-2.170
Asymp. Sig. (2-tailed)	.030

a. Grouping Variable: Status

Ik-1

Mann-Whitney Test

Ranks

	Status	N	Mean Rank	Sum of Ranks
Fold.Chg	0	85	153.82	13074.50
	1	174	118.36	20595.50
	Total	259		

Test Statistics[a]

	Fold.Chg
Mann-Whitney U	5370.500
Wilcoxon W	20595.500
Z	-3.577
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: Status

Oct-1

Mann-Whitney Test

Ranks

	Status	N	Mean Rank	Sum of Ranks
Fold.Chg	0	183	175.65	32144.50
	1	145	150.42	21811.50
	Total	328		

Test Statistics[a]

	Fold.Chg
Mann-Whitney U	11226.500
Wilcoxon W	21811.500
Z	-2.393
Asymp. Sig. (2-tailed)	.017

a. Grouping Variable: Status

Figure 7. Co-localization of TF target genes positively affects co-regulation. (**a**). Comparison of expression profiles of genes showing both co-localization patterns and recognition patterns by a specific DLBCL annotated transcription factor (Status = 0) from those that although being co-localized were not recognized by the specified transcription factor (Status = 1). The comparison of the gene expression profiles between the two groups was done using the Mann-Whitney U test at a threshold of $p<0.05$ and results were illustrated as boxplot. Only TF co-localized target genes differing in their expression patterns with respect to those not recognized by the specific TFs have been detailed (Oct-1 and IK-1). (**b**). Both co-regulation by a common TF and co-localization patterns determine co-expression of TF target genes. The expression profiles of Oct-1 co-localized target genes (Status = 0) were compared to those that although recognized by Oct-1did not cluster together (Status = 1); notably Oct-1 co-localized target genes differed in their expression from those that did not cluster as shown by the Mann-Whitney U test at a threshold of $p<0.05$.
doi:10.1371/journal.pone.0066196.g007

confined regions on the chromosomes, but also respond to the same transcription factor families, and are similarly expressed.

No currently available bioinformatics tool allows for multi-attribute gene searches where co-localization, co-expression and co-regulation can be considered at the same time.

CluGene offers such possibility for the first time, and provides flexible integration mechanisms to assign varying relative importance to co-localization, co-expression and co-regulation to accommodate the needs of many types of analysis processes.

It is particularly important to stress out the *concurrency* of the search criteria (localization, expression and regulation) allowed by CluGene. A possible alternative would consist in running multi-step searches that proceed by further refinements: e.g. a first positional clustering to identify groups of genes that are located close to each other, followed by filtering on expression levels, and again followed by searching for common TFBSs in order to identify those subsets of co-localized, co-expressed and co-regulated genes. It goes without saying that such a sequential approach would greatly limit the amount of promising genes that in the end survive the bioinformatics selection process, each step potentially eliminating promising candidates only because of the order the selection operations were performed. On the contrary, with CluGene, such steps can be integrated so that all the criteria can be applied at the same time, and with weights, so that the relative importance of each can be controlled.

The case studies have shown that there may not necessarily be a unique analysis protocol that works equally well for any type of problem; which means that in general, for a new problem, many pathways may be explored involving numerous combinations of multi-attribute clustering and weighting strategies before important novel insight can be gained, and emerging behaviour can be detected. This brings to the table the second main distinguishing feature of CluGene, i.e., the software being an interactive framework where different solutions can be rapidly explored and compared. Furthermore, through the plug-in mechanism, CluGene allows for user-defined algorithms and procedures to be integrated within the framework to expand its functionalities, reaching an optimal compromise between the accessibility of a user-friendly solution and the flexibility of a programming environment.

The types of investigations that may be run with CluGene are numerous, ranging from simpler searches trying to determine if a specific TF regulates nearby genes, or analysing the tendency of a TF to preferentially regulate genes located on one or a few specific chromosomes or chromosome regions, to more complex analysis tasks, such as determining whether TF regulatory mechanisms have lead to evolutionary constraints favouring the formation of specific positional clusters of target genes, and whether clusters of co-regulated genes show some tendency of producing similar expression levels as well.

A limitation of CluGene concerning TF analysis is that it currently supports only the analysis of binding sites in promoters; in future releases binding information will be searchable also in other locations across the genome, such as enhancers, introns, etc. (e.g. analysis of ChIP-Seq data).

Moreover, despite the capacity of CluGene to provide novel viewpoints for understanding transcription regulation, it still fails to provide insight on higher-order phenomena. Specifically, the chromatin remodelling events leading to inactive and active states and the three-dimensional, spatial organization of the chromosomes within the nucleus (i.e. chromosome territories) are currently beyond the analysis capabilities of CluGene. Our method typically identifies only the first level of transcription regulation: further applications will certainly need to take into consideration higher-order transcription regulation in order to provide more tools for understanding transcription regulatory mechanisms.

Nevertheless, it is our belief that as genome-wide transcriptional analysis and in general genome high-throughput experimental technologies will keep on providing a growing amount of information on different organisms, the application of software tools such as CluGene will become increasingly more important.

The availability of techniques that allow for the identification of relationships between gene properties and regulation phenomena will reveal itself as a fundamental prerequisite for the development of original and performing search tools to shed some more light on transcriptional regulatory networks.

Supporting Information

Figure S1 TF-based gene dataset processing. (**a**). For each gene present within a dataset (left panel) it is possible to identify number and types of TFBS, and consequently, what TFs can recognize them. (**b**). It is also possible by selecting a specific TF to search for all genes being recognized by it together with associated frequencies, a list of all TFs recognizing each gene in dataset is provided as well. (**c**). Distinct datasets can be compared in terms of predicted TFBs. The comparison can be ordered as a function of difference or relevance. TFs are coloured differently in the compared datasets depending on the outcome: present at an higher extent (green), present at a lower extent (red) and absent (black).
(TIF)

Figure S2 Transcription factor binding site motifs over-represented in sequences from the three DLBCL subsets (BRC, HR and OXPHOS). (**a**). List of TFs with associated frequencies for each DLBCL subset according to the MatchTM predictor embedded in CluGene. Annotated B cell transcription factors are boxed. TFs are coloured differently in the compared datasets depending on the outcome: Higher transcription factors binding site motifs percentages are coloured green whereas lower percentage are marked in red, absent TFBSs are coloured balck. The last column shows the maximum difference values amongst TF frequencies within the three subsets. (**b**). Statistical significance assessment for the results obtained with selected TFs in the OXPHOS dataset with respect to a random set of genes. Transcription factors results on the OXPHOS dataset were compared with average transcription results on 10 datasets of randomly selected genes. Ten random sets of genes were extracted form the human genome, with a number of genes same as OXPHOS. The One sample student's t-test (SPSS) was used for the analysis.
(TIF)

Figure S3 Percentages of genes that are TF targets. Number of TF target genes out of the total number of genes present within the three DLBCL subsets. A comparison between the number of TF target genes in the three DLBCL subsets and different clustering strategies output is presented. Gene clustering was performed using both the *un-weighted positional clustering* (Positional clustering) followed by TF prediction and the *functional clustering weighted by local density and TF similarity*. For the positional and functional clustering cases, the percentage refers to genes found in clusters.
(TIF)

Figure S4 Localization of genes containing the HT motif in *P. falciparum*. The gene density and the cluster of genes containing specific localization attributes were predicted through-

out the *Plasmodium Falciparum* genome using different source datasets. (a) Calculation of the gene density of genes containing the signal peptide (PlasmoDB) as localization attribute, a total of 1216 genes were considered for the analysis. The left-hand panel shows the distribution of the gene density on chromosome 5 as an example, while the right-hand panel shows the localization of genes containing the signal peptide on all chromosomes. The relative position of the individual genes (blue bars) are shown along the chromosomes (horizontal lines). The positions of the genes along the chromosomes are also indicated with a label that represents the position (in kbp). (b) Calculation of the gene density of *P. falciparum* genes containing the HT localization motif (PlasmoDB), a total of 214 genes were considered for the analysis, the left-hand panel shows the calculated gene density on chromosome 5 as an example, while the right-hand panel shows the localization of the genes containing the HT motif along the different chromosomes: genes tend to locate mostly at the end of the chromosomes. (c) Clustering analysis of the dataset containing the genes with the HT motif (same dataset as in (b)). The relative position and the size of individual genes clusters (blue-oval shape) are shown along the chromosomes (horizontal lines). The size of each cluster is proportional to its population size. Each cluster is also flagged with a label that indicates the position (in kbp) of the first gene. A total of 192 genes in 36 clusters were found out of the total 214. Gene clustering was performed using default positional clustering options.
(TIF)

Text S1 Supplementary information about the CluGene software. Download details; system requirements; third-party libraries used by CluGene; genome and transcription factor online databases accessible from within the software; technical details on the clustering algorithms available in CluGene.
(DOCX)

Acknowledgments

The authors do not have any conflicts of interest. The authors thank the anonymous reviewers for the constructive comments.

Author Contributions

Conceived and designed the experiments: TD NS AC. Performed the experiments: TD PP. Analyzed the data: TD PP NS. Contributed reagents/materials/analysis tools: TP RS. Wrote the paper: TD NS AC. Conceived and designed the software: TD NS PP. Implemented the software: PP. Wrote the software documentation: PP TP.

References

1. Heyer LJ, Kruglyak S, Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Res 9: 1106–1115.
2. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9: 3273–3297.
3. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.
4. Dottorini T, Persampieri T, Palladino P, Baker DA, Spaccapelo R, et al. (2013) Regulation of Anopheles gambiae male accessory gland genes influences postmating response in female. FASEB J 27: 86–97.
5. Dottorini T, Senin N, Mazzoleni G, Magnusson K, Crisanti A (2011) Gepoclu: a software tool for identifying and analyzing gene positional clusters in large-scale gene expression analysis. BMC Bioinformatics 12: 34.
6. Dottorini T, Nicolaides L, Ranson H, Rogers DW, Crisanti A, et al. (2007) A genome-wide analysis in Anopheles gambiae mosquitoes reveals 46 male accessory gland genes, possible modulators of female behavior. Proc Natl Acad Sci U S A 104: 16215–16220.
7. Hershberg R, Yeger-Lotem E, Margalit H (2005) Chromosomal organization is shaped by the transcription regulatory network. Trends Genet 21: 138–142.
8. Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI (2002) Large clusters of co-expressed genes in the Drosophila genome. Nature 420: 666–669.
9. Osbourn AE, Field B (2009) Operons. Cell Mol Life Sci 66: 3755–3775.
10. Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5: 299–310.
11. Koonin EV (2009) Evolution of genome architecture. Int J Biochem Cell Biol 41: 298–306.
12. Lawrence JG (2002) Shared strategies in gene organization among prokaryotes and eukaryotes. Cell 110: 407–413.
13. Michalak P (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics 91: 243–248.
14. Sproul D, Gilbert N, Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet 6: 775–781.
15. Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet 26: 183–186.
16. Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol 1: 5.
17. Williams EJ, Bowles DJ (2004) Coexpression of neighboring genes in the genome of Arabidopsis thaliana. Genome Res 14: 1060–1067.
18. Dai Z, Dai X (2011) Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. Nucleic Acids Res.
19. Janga SC, Collado-Vides J, Babu MM (2008) Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. Proc Natl Acad Sci U S A 105: 15761–15766.
20. He X, Goldwasser MH (2005) Identifying conserved gene clusters in the presence of homology families. JComputBiol 12: 638–656.
21. Ling X, He X, Xin D (2009) Detecting gene clusters under evolutionary constraint in a large number of genomes. Bioinformatics 25: 571–577.
22. Luc N, Risler JL, Bergeron A, Raffinot M (2003) Gene teams: a new formalization of gene clusters for comparative genomics. Comput Biol Chem 27: 59–67.
23. Li Q, Lee BT, Zhang L (2005) Genome-scale analysis of positional clustering of mouse testis-specific genes. BMC Genomics 6: 7.
24. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
25. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16–23.
26. Stormo GD (1990) Consensus patterns in DNA. Methods Enzymol 183: 211–221.
27. Wasserman WW, Krivan W (2003) In silico identification of metazoan transcriptional regulatory regions. Naturwissenschaften 90: 156–166.
28. Koop BF (1995) Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. Trends Genet 11: 367–371.
29. Hardison RC, Oeltjen J, Miller W (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res 7: 959–966.
30. Duret L, Bucher P (1997) Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol 7: 399–406.
31. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE (2000) Human-mouse genome comparisons to locate regulatory sites. Nat Genet 26: 225–228.
32. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137–144.
33. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, et al. (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res 33: 3154–3164.
34. Hestand MS, van Galen M, Villerius MP, van Ommen GJ, den Dunnen JT, et al. (2008) CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes. BMC Bioinformatics 9: 495.
35. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166–176.
36. Ester M, Kriegel H, Jörg Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise; 1996.
37. Ankerst M, Breunig MM, Kriegel H, Sander J (1999) OPTICS: Ordering Points To Identify the Clustering Structure. ACM SIGMOD international conference on Management of data: ACM press. 49–60.
38. Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis: Academic Press.
39. Hamerly G, Elkna C. Alternatives to the k-means algorithm that find better clusterings; 2002.
40. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding; 2007. 1027–1035.
41. Browne P, Petrosyan K, Hernandez A, Chan JA (2003) The B-cell transcription factors BSAP, Oct-2, and BOB.1 and the pan-B-cell markers CD20, CD22, and

CD79a are useful in the differential diagnosis of classic Hodgkin lymphoma. Am J Clin Pathol 120: 767–777.

42. Flenghi L, Bigerna B, Fizzotti M, Venturi S, Pasqualucci L, et al. (1996) Monoclonal antibodies PG-B6a and PG-B6p recognize, respectively, a highly conserved and a formol-resistant epitope on the human BCL-6 protein amino-terminal region. Am J Pathol 148: 1543–1555.
43. Stein H, Marafioti T, Foss HD, Laumen H, Hummel M, et al. (2001) Down-regulation of BOB.1/OBF.1 and Oct2 in classical Hodgkin disease but not in lymphocyte predominant Hodgkin disease correlates with immunoglobulin transcription. Blood 97: 496–501.
44. Falini B, Fizzotti M, Pucciarini A, Bigerna B, Marafioti T, et al. (2000) A monoclonal antibody (MUM1p) detects expression of the MUM1/IRF4 protein in a subset of germinal center B cells, plasma cells, and activated T cells. Blood 95: 2084–2092.
45. Foss HD, Reusch R, Demel G, Lenz G, Anagnostopoulos I, et al. (1999) Frequent expression of the B-cell-specific activator protein in Reed-Sternberg cells of classical Hodgkin's disease provides further evidence for its B-cell origin. Blood 94: 3108–3113.
46. Hamada T, Yonetani N, Ueda C, Maesako Y, Akasaka H, et al. (1998) Expression of the PAX5/BSAP transcription factor in haematological tumour cells and further molecular characterization of the t(9;14)(p13;q32) translocation in B-cell non-Hodgkin's lymphoma. Br J Haematol 102: 691–700.
47. Jundt F, Kley K, Anagnostopoulos I, Schulze Probsting K, Greiner A, et al. (2002) Loss of PU.1 expression is associated with defective immunoglobulin transcription in Hodgkin and Reed-Sternberg cells of classical Hodgkin disease. Blood 99: 3060–3062.
48. Torlakovic E, Tierens A, Dang HD, Delabie J (2001) The transcription factor PU.1, necessary for B-cell development is expressed in lymphocyte predominance, but not classical Hodgkin's disease. Am J Pathol 159: 1807–1814.
49. Saez AI, Artiga MJ, Sanchez-Beato M, Sanchez-Verde L, Garcia JF, et al. (2002) Analysis of octamer-binding transcription factors Oct2 and Oct1 and their coactivator BOB.1/OBF.1 in lymphomas. Mod Pathol 15: 211–220.
50. McCune RC, Syrbu SI, Vasef MA (2006) Expression profiling of transcription factors Pax-5, Oct-1, Oct-2, BOB.1, and PU.1 in Hodgkin's and non-Hodgkin's lymphomas: a comparative study using high throughput tissue microarrays. Mod Pathol 19: 1010–1018.
51. Garbati MR, Alco G, Gilmore TD (2010) Histone acetyltransferase p300 is a coactivator for transcription factor REL and is C-terminally truncated in the human diffuse large B-cell lymphoma cell line RC-K8. Cancer Lett 291: 237–245.
52. Gilmore TD, Kalaitzidis D, Liang MC, Starczynowski DT (2004) The c-Rel transcription factor and B-cell proliferation: a deal with the devil. Oncogene 23: 2275–2286.
53. Kontgen F, Grumont RJ, Strasser A, Metcalf D, Li R, et al. (1995) Mice lacking the c-rel proto-oncogene exhibit defects in lymphocyte proliferation, humoral immunity, and interleukin-2 expression. Genes Dev 9: 1965–1977.
54. Starczynowski DT, Trautmann H, Pott C, Harder L, Arnold N, et al. (2007) Mutation of an IKK phosphorylation site within the transactivation domain of REL in two patients with B-cell lymphoma enhances REL's in vitro transforming activity. Oncogene 26: 2685–2694.
55. Bavi P, Uddin S, Bu R, Ahmed M, Abubaker J, et al. (2011) The biological and clinical impact of inhibition of NF-kappaB-initiated apoptosis in diffuse large B cell lymphoma (DLBCL). J Pathol 224: 355–366.
56. Georgopoulos K, Bigby M, Wang JH, Molnar A, Wu P, et al. (1994) The Ikaros gene is required for the development of all lymphoid lineages. Cell 79: 143–156.
57. Thieblemont C, Nasser V, Felman P, Leroy K, Gazzo S, et al. (2004) Small lymphocytic lymphoma, marginal zone B-cell lymphoma, and mantle cell lymphoma exhibit distinct gene-expression profiles allowing molecular diagnosis. Blood 103: 2727–2737.
58. Marafioti T, Hummel M, Foss HD, Laumen H, Korbjuhn P, et al. (2000) Hodgkin and reed-sternberg cells represent an expansion of a single clone originating from a germinal center B-cell with functional immunoglobulin gene rearrangements but defective immunoglobulin transcription. Blood 95: 1443–1450.
59. Theil J, Laumen H, Marafioti T, Hummel M, Lenz G, et al. (2001) Defective octamer-dependent transcription is responsible for silenced immunoglobulin transcription in Reed-Sternberg cells. Blood 97: 3191–3196.
60. Re D, Muschen M, Ahmadi T, Wickenhauser C, Staratschek-Jox A, et al. (2001) Oct-2 and Bob-1 deficiency in Hodgkin and Reed Sternberg cells. Cancer Res 61: 2080–2084.
61. Urbanek P, Wang ZQ, Fetka I, Wagner EF, Busslinger M (1994) Complete block of early B cell differentiation and altered patterning of the posterior midbrain in mice lacking Pax5/BSAP. Cell 79: 901–912.
62. Kozmik Z, Wang S, Dorfler P, Adams B, Busslinger M (1992) The promoter of the CD19 gene is a target for the B-cell-specific transcription factor BSAP. Mol Cell Biol 12: 2662–2672.
63. Pham LV, Tamayo AT, Yoshimura LC, Lin-Lee YC, Ford RJ (2005) Constitutive NF-kappaB and NFAT activation in aggressive B-cell lymphomas synergistically activates the CD154 gene and maintains lymphoma cell survival. Blood 106: 3940–3947.
64. Monti S, Savage KJ, Kutok JL, Feuerhake F, Kurtin P, et al. (2005) Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood 105: 1851–1861.
65. Alibes A, Yankilevich P, Canada A, Diaz-Uriarte R (2007) IDconverter and IDClight: conversion and annotation of gene and protein IDs. BMC Bioinformatics 8: 9.
66. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
67. Pilpel Y, Sudarsanam P, Church GM (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 29: 153–159.
68. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8: 785–786.
69. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567–580.
70. Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, et al. (2003) Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum. Science 299: 705–708.
71. Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T, et al. (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. Science 306: 1934–1937.
72. Sargeant TJ, Marti M, Caler E, Carlton JM, Simpson K, et al. (2006) Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. Genome Biol 7: R12.
73. Maher B (2012) ENCODE: The human encyclopaedia. Nature 489: 46–48.
74. Magnusson K, Mendes AM, Windbichler N, Papathanos PA, Nolan T, et al. (2011) Transcription regulation of sex-biased genes during ontogeny in the malaria vector Anopheles gambiae. PLoS One 6: e21572.