Journal of Proteomics & Bioinformatics - Open Access

Research Article JPB/Vol.2/January 2009

GExMap: An Intuitive Visual Tool to Detect and Analyze Genomic Distribution in Microarray-generated Lists of Differentially Expressed Genes

Nicolas Cagnard[1*], **Carlo Lucchesi**[2], **Gilles Chiocchia**[3]

[1]Plateforme Bioinformatique Paris Descartes, Université Paris Descartes,IRNEM IFR-94, Paris 75730
[2]Institut Curie, INSERM U830, Paris 75248
[3]Institut Cochin, Université Paris,Descartes, CNRS (UMR 8104), Paris, F-75014 France,
Inserm, U567, Département d'Immunologie, Paris, F-75014 France, AP-HP, Hôpital Ambroise Paré,
Service de Rhumatologie, Boulogne-Billancourt, F 92000, France

*Corresponding author: Nicolas Cagnard, Plateforme Bioinformatique Paris
Descartes, Université Paris Descartes,IRNEM IFR-94, Paris 75730

Received December 01, 2008; Accepted January 13, 2009; Published January 15, 2009

Citation: Nicolas C (2009) GExMap: An Intuitive Visual Tool to Detect and Analyze Genomic Distribution in Microarray-generated Lists of Differentially Expressed Genes. J Proteomics Bioinform 2: 051-059. doi:10.4172/jpb.1000060

Abstract

Background

High- throughput technologies, as DNA microarrays, generate a huge amount of data, which are difficult to interpret. Biologists require easy-to-use tools to analyze them and explore new scientific hypotheses. We propose a visual data mining tool to extract a new type of useful genomic information buried in gene lists generated by differential expression studies. We compare the genomic distribution that is observed within the gene list with the expected distribution, which is estimated from public genomic databases. An algorithm of research and a statistical test give reliable and optimal results. GExMap helps identify genomic regions that are enriched in genes whose differential expression is of potential interest for target diseases. This software is freely available[20] and easily customized. Since sources are frequently updated, it offers tools for updates at any moment. GExMap is usable by any commercially and publicly available microarray platform. Furthermore, GExMap helps in interpretation by showing an ordered list for each gene ontology. Presently, GExMap can also be used to analyse gene lists not only from the Homo sapiens genome but also the Mus musculus and the Rattus norvegicus genomes.

Results

GExMap is designed to be an easy-to-use software with visual and intuitive interpretation of results. We have chosen to develop the software with R[1] language to allow total compatibility with every operating system. ENSEMBL[2] was chosen as a reference identifier for genes, meaning that GExMap requires a pre-processing step to match and replace most of the common user's identifier types (Unigene, Affymetrix, Agilent, etc) by ENSEMBL identifiers. The choice of a reference genome for this pre-processing step allows compatibility with several databases and microarray identifiers. In the same step, GExMap creates a new table with all ENSEMBL available information. GExMap is an open source project. Furthermore, its source structure and simple annotation facilitate customization.

Conclusions

Biological interpretation of bioinformatics data is an ongoing and challenging task. Scientists need to access to types of information buried in biological databases. From microarray data, GExMap indicates the observed genomic distribution of the regulated genes and statistically compares it to the expected genomic distribution. Statistical analysis of relative genomic distribution is a potentially powerful and informative approach to microarray data interpretation. GExMap provides biologists with easy access to this new type of information.

Background

Microarray technology allows parallel and simultaneous monitoring of the whole genome. It is used to detect differentially expressed genes under different conditions. General studies of microarray gene expression generate lists of differentially expressed genes and molecular signatures of diseases. In some cases, the list of genes might be used as a powerful diagnostic and/or prognostic tool for the disease in question[-7].

Since its beginnings with cDNA microarrays[8-9], targeting only parts of the genome, microarray technology has been improved in accuracy[10-12, 21] and now targets the whole genome.

Microarray data analysis produces large lists of genes that are difficult to interpret. Although still imperfect, several programs are able to extract gene functions and bibliographic associations from ontology databases[18], while others are able to document[13-14] or propose biological pathways[15-16]. Few programs, however, can analyze the genomic distributions buried in the data. GExMap has been designed to help fill this gap, by providing solutions to specific requests, and mapping and analyzing several types of microarray data[17].

Implementation

GExMap is an R software package implementing visual and statistical tools to study and map genomic distribution of differentially expressed genes, in order to facilitate gene list interpretation. With visual representation, the genomic distribution observed in a gene list can be compared with the expected genomic distribution. The results obtained are also statistically tested to unravel genomic regions of potential interest. Moreover, GExMap produces a user-friendly documentation, widely annotated source code and a “visual technical manual” available online[20]. GExMap can also produce scored lists and graphical illustrations of Gene Ontology[18] identifiers to facilitate gene list interpretation and highlight specific molecular function, biological processes or cellular component.

GExMap produces one individual genomic map per chromosome, complete with statistical information. By default, chromosome maps are not created for chromosomes that contain fewer than five of the listed genes. In those cases, the genomic density comparisons will not be relevant. On the other hand, this parameter could be easily modified by the user, taking into account the increasing false-positive rate of potentially interesting regions.

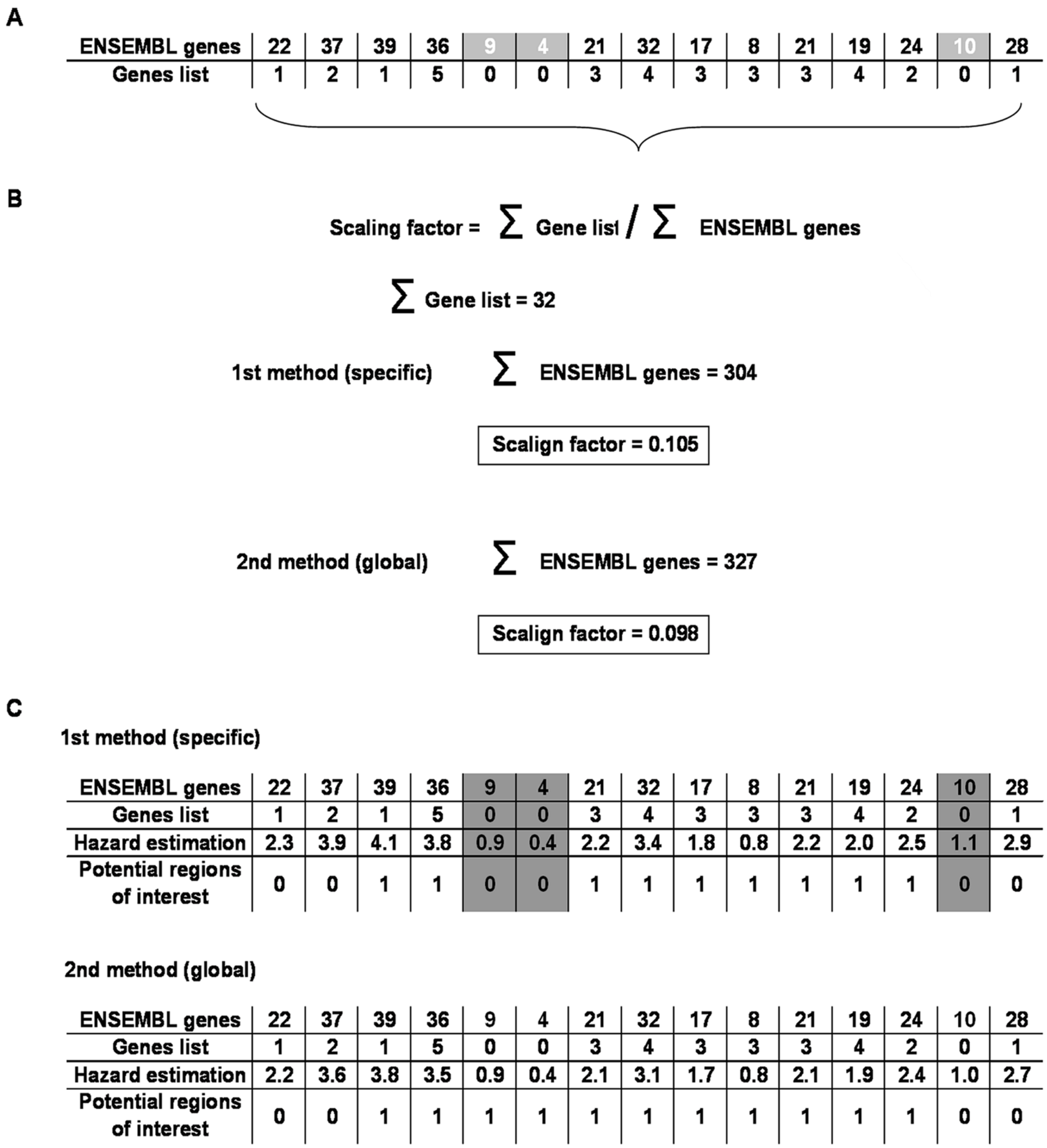
To facilitate correspondence between identifiers of several types of microarray and genome databases, we have chosen the ENSEMBL genome as the reference. Once the identifier type is recognized, the appropriate correspondence file is automatically loaded. Then, all identifiers are replaced by ENSEMBL annotation and appropriate genomic information. The ENSEMBL genome is presently composed of around 33 000 identifiers corresponding to all known genes and pseudogenes. The microarray probes do not always correspond to one known gene or to one ENSEMBL identifier. Meaning that probes with no correspondences cannot be allocated to genomic locations. This problem will be solved when the genome is fully and scrupulously annotated. Furthermore the unprocessed identifiers are listed and reported in the final report. The main GExMap data file will be frequently updated and freely available to limit this problem. We have chosen not to take into account identifiers without ENSEMBL correspondence for the subsequent steps.

There is virtually no limit to the size of the tested gene list. In terms of contents, we would recommend the gene list be generated by any rigorous mathematical, statistical or experimental approach designed to identify of differentially expressed genes. If any arbitrary selection has been performed, the gene list may import a significant bias for the results of further tests.

We define a genomic map associated with a gene list as the information about the genomic location of each gene of the list, in the corresponding genome draft.

GExMap produces the genomic map of a gene list, aggregates genomic locations into genomic units, calculates the observed gene frequency by unit, and then performs the statistical comparison of the observed genomic distribution to the one expected in the full genome. The graphical and computational unit can be chosen by the user. Default scaling has been set to 1 million bp. The hazard curve represents the expected genomic distribution which could be found when the genome-wide number of genes is scaled to the size of the tested gene list. Two hazard computations are available. The first one only takes into account ENSEMBL genes located in units where at least one gene of the tested list is located. The second type of hazard computation takes into account ENSEMBL genes of all genome (Figure 1). The hazard frequency is used to test the hypothesis that the expected and observed genomic distributions are not homogeneous, and the latter must vary in relation to the disease studied. The total number of genes is broken into two regulation curves showing the number of up- and downregulated genes per unit.

When the tested gene list curve is higher than the hazard curve, then one can expect that the cluster of genes encompassed in the genomic unit could be of interest. GExMap has been set to detect these clusters automatically. On the other hand, in some particular cases, it could be interesting



A

ENSEMBL genes	22	37	39	36	9	4	21	32	17	8	21	19	24	10	28
Genes list	1	2	1	5	0	0	3	4	3	3	3	4	2	0	1

B

Scaling factor = $\sum$ Gene list / $\sum$ ENSEMBL genes

$\sum$ Gene list = 32

1st method (specific) $\sum$ ENSEMBL genes = 304

Scalign factor = 0.105

2nd method (global) $\sum$ ENSEMBL genes = 327

Scalign factor = 0.098

C

1st method (specific)

ENSEMBL genes	22	37	39	36	9	4	21	32	17	8	21	19	24	10	28
Genes list	1	2	1	5	0	0	3	4	3	3	3	4	2	0	1
Hazard estimation	2.3	3.9	4.1	3.8	0.9	0.4	2.2	3.4	1.8	0.8	2.2	2.0	2.5	1.1	2.9
Potential regions of interest	0	0	1	1	0	0	1	1	1	1	1	1	1	0	0

2nd method (global)

ENSEMBL genes	22	37	39	36	9	4	21	32	17	8	21	19	24	10	28
Genes list	1	2	1	5	0	0	3	4	3	3	3	4	2	0	1
Hazard estimation	2.2	3.6	3.8	3.5	0.9	0.4	2.1	3.1	1.7	0.8	2.1	1.9	2.4	1.0	2.7
Potential regions of interest	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0

Figure 1: Hazard estimations. A: For the specific hazard estimation method, only units encompassing at least one gene from the gene list are taken into account for the hazard estimation. Units are not filtered for the global hazard estimation method. B: Scaling factor computation, by the ratio of the sums of the ENSEMBL genes and the tested list genes. C: Hazard estimation is computed by scaling the ENSEMBL genome density to the scale of the tested gene list.

to detect the regions of absence of genes. For this reason, this parameter can be customized thought the use of a variable. A statistical test is performed to validate each detected cluster, and the result is depicted graphically. Significant regions are highlighted on the map by a specific green line. The complete analysis process can be summarized into 6 simple steps (Figure 2).

Data Sources

The program loads the preformatted source file "Data.Rdata" containing all ENSEMBL identifiers and their genomic locations. The algorithm is designed to implement several processing modes. The user can define the scale of the resulting graphics, the path of the source file, and the

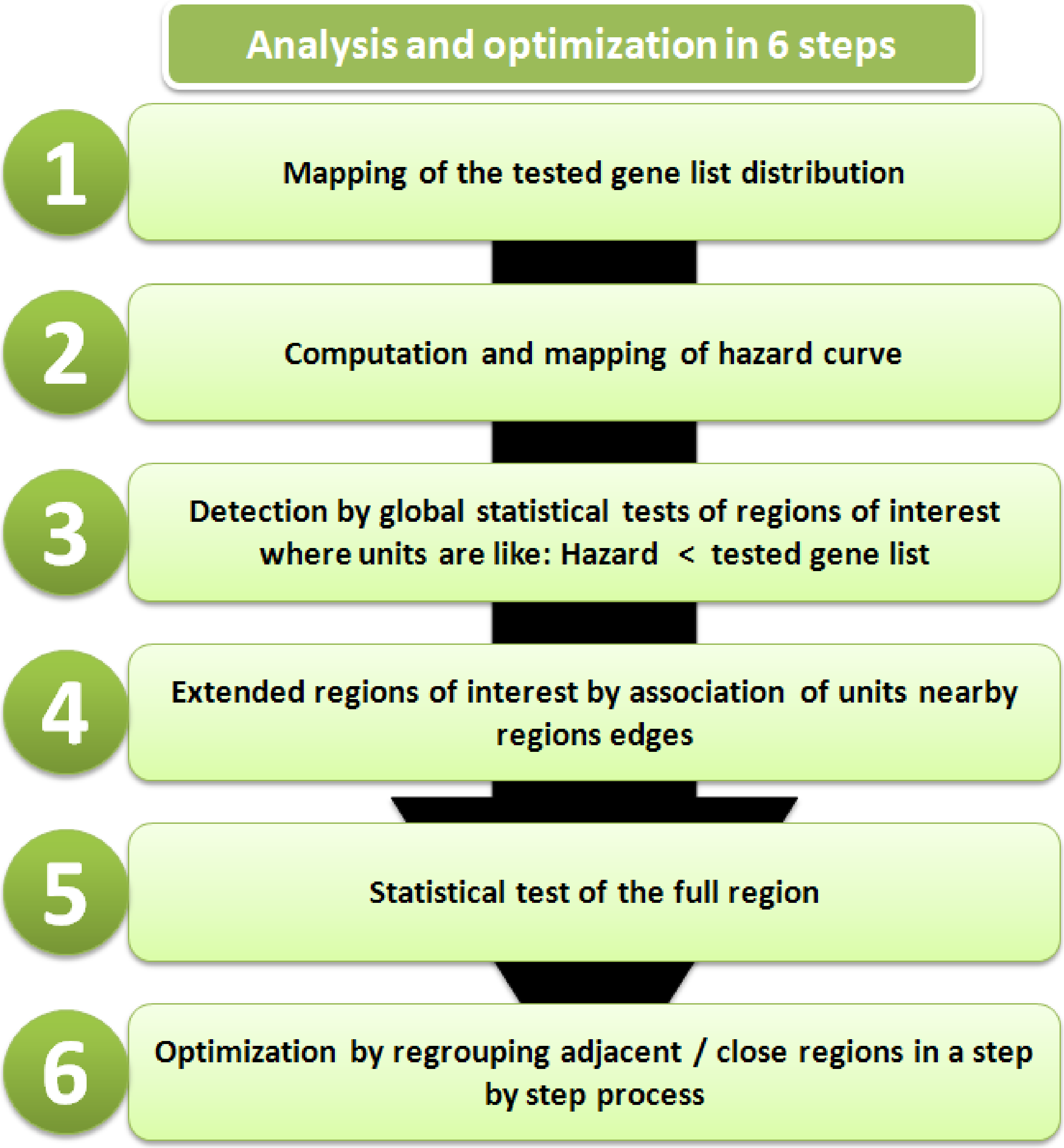

Analysis and optimization in 6 steps
1
Mapping of the tested gene list distribution
2
Computation and mapping of hazard curve
3
Detection by global statistical tests of regions of interest where units are like: Hazard < tested gene list
4
Extended regions of interest by association of units nearby regions edges
5
Statistical test of the full region
6
Optimization by regrouping adjacent / close regions in a step by step process

Figure 2: Local Analysis and optimization in six steps. 1: The first step is to construct a matrix of genomic localisation for the tested gene list. 2: Then the corresponding hazard distribution is computed. 3: A global test is performed to detect potential regions of interest at a large scale. Then each base unit are tested if they are located in large regions of interest and if there are more genes than expected (real number > hazard estimation). 4: A step of optimization test if extension of statistically significant regions of interest brings better results. 5: The final test is done after optimization. 6: Closer significant regions of interest can be merged if the resulting larger region improves statistical significance.

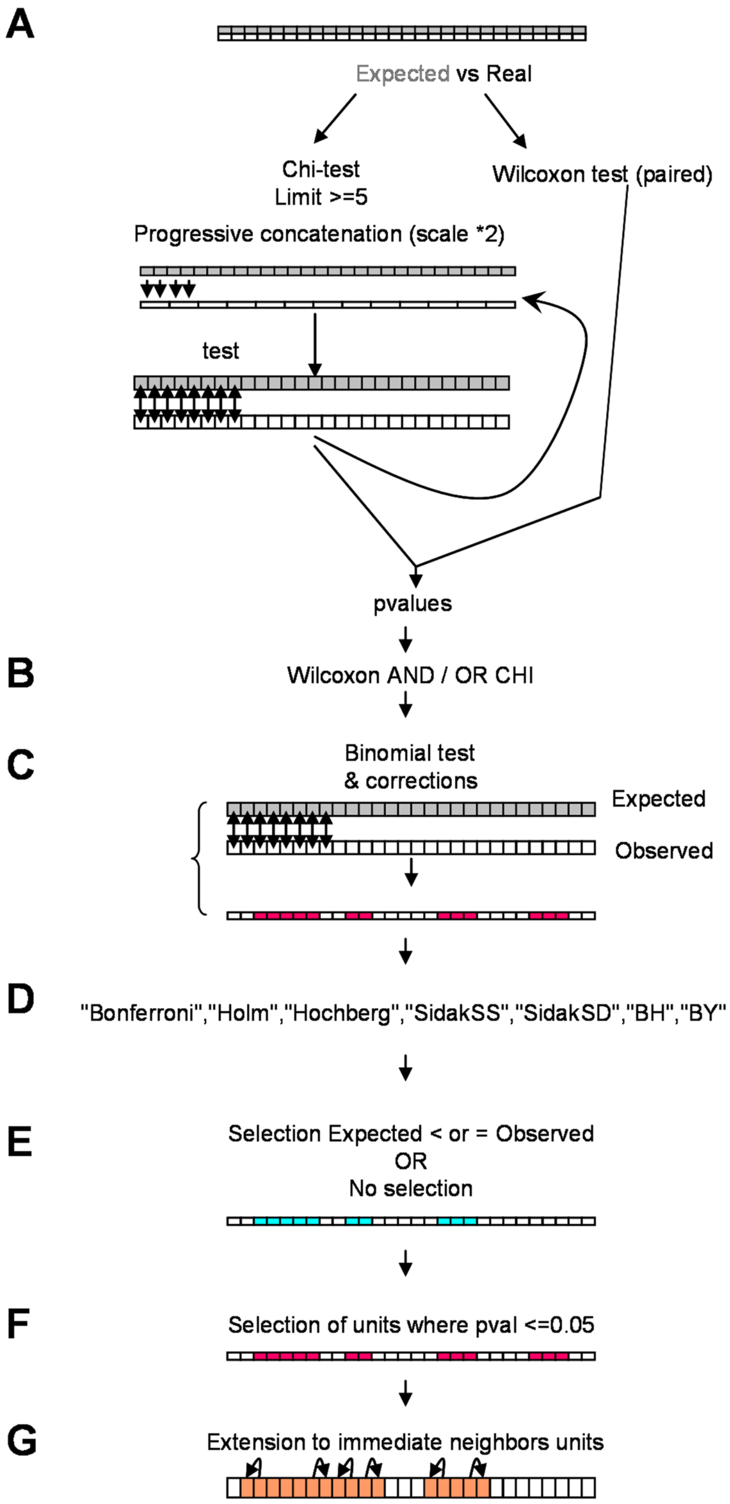

A
Expected vs Real
Chi-test
Limit >=5
Wilcoxon test (paired)
Progressive concatenation (scale *2)
test
pvalues
B
Wilcoxon AND / OR CHI
C
Binomial test
& corrections
Expected
Observed
D
"Bonferroni","Holm","Hochberg","SidakSS","SidakSD","BH","BY"
E
Selection Expected < or = Observed
OR
No selection
F
Selection of units where pval <=0.05
G
Extension to immediate neighbors units

Figure 3: Selection of regions to test. First, regions of interest are detected by comparison of observed and expected distributions (red squares). Second, the regions are 5' and 3' extended by one unit to allow a sufficient sample size to be statistically tested. A: Comparison of observed and expected distribution at the chromosome scale by two global tests: CHI-goodness of fit, Wilcoxon. B: Selection of the chromosomes according to the statistical results (global tests) or the user choice. C: Comparison of observed and expected distribution unit by unit throughout each selected chromosome with binomial local test. D: Local pval correction according to the user choice. E: Selection of the units showing observed effectives greater than expected according to the user choice. F: Selection of units of interest. G: Extension of regions of interest.

results folder. At the moment, there is only one default mode mapping differentially expressed gene lists. Other modes, as CGH and SNP mapping, are under development. The mode is automatically implemented by the program through the analysis of the source file's column names.

Data Identification

The appropriate file with correspondences between identifiers and ENSEMBL genome is loaded. The replicates are filtered and their numbers reported in a report text file. A table is then created with unique ENSEMBL identifiers for each selected gene, with their genomic locations. Once no existing correspondence can be found, the identifier will increment a "none mapped" list in the results folder. After filtering of the gene list from replicates and unassigned identifiers, a data matrix is created. Genomic locations are calculated by taking the center of each gene (start + end / 2). Genes are then classified by chromosomes and by chromosomal location.

Hazard Estimation

The matrix used to calculate the genomic distribution is computed chromosome by chromosome. The observed

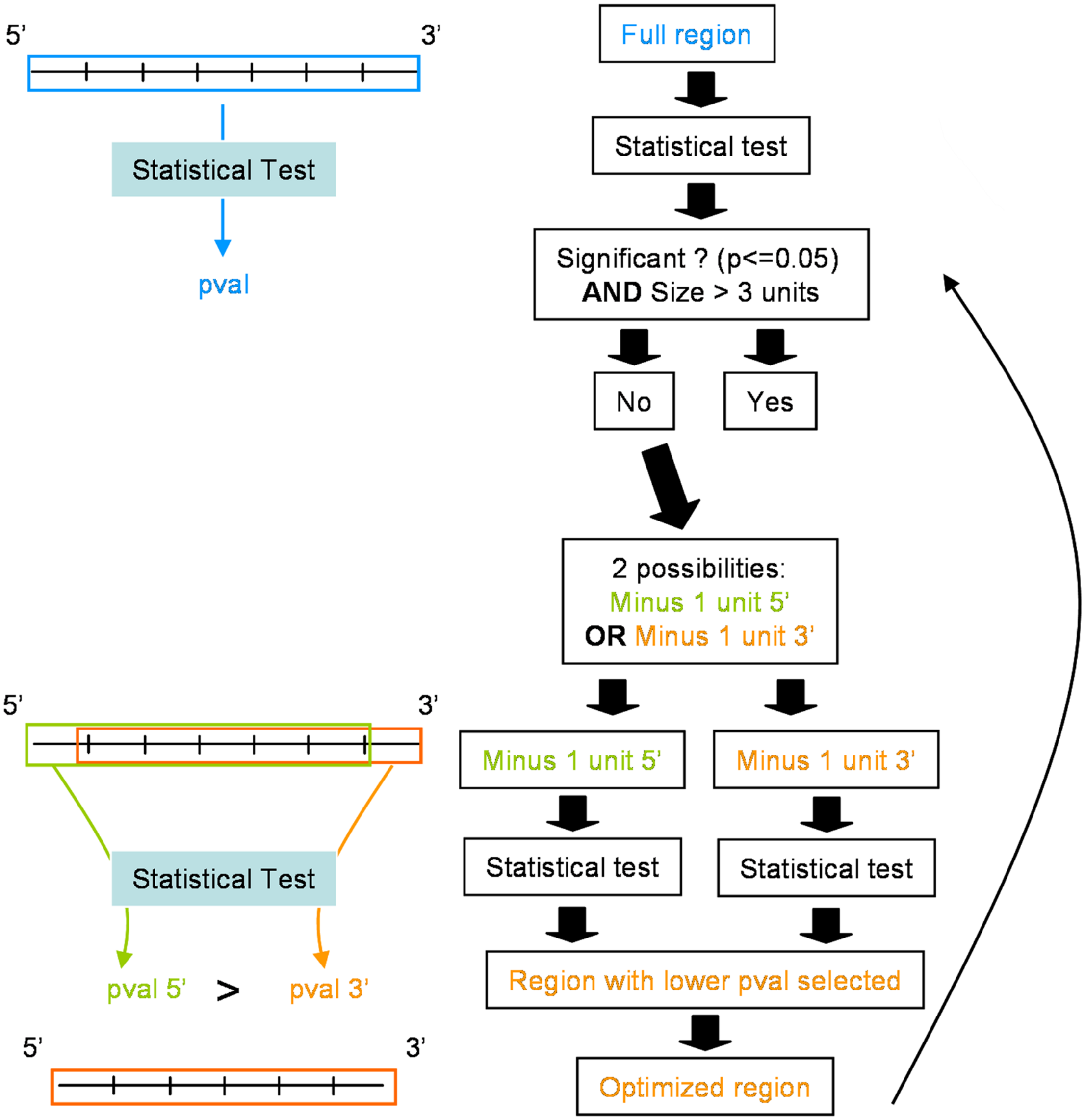


Figure 4: Local statistical test and progressive optimization. First of all, the full region of interest is statistically tested (blue). When the test invalidates the region a step of progressive optimization takes place. Progressive optimization consists in statistically testing the two potential reduced regions (green and orange). The reduced region allowing the best statistical result is used in the next steps. The optimization steps and test are performed as long as the result of the test is not significant (pval > 0.05) and the size of the tested region is longer than three units.

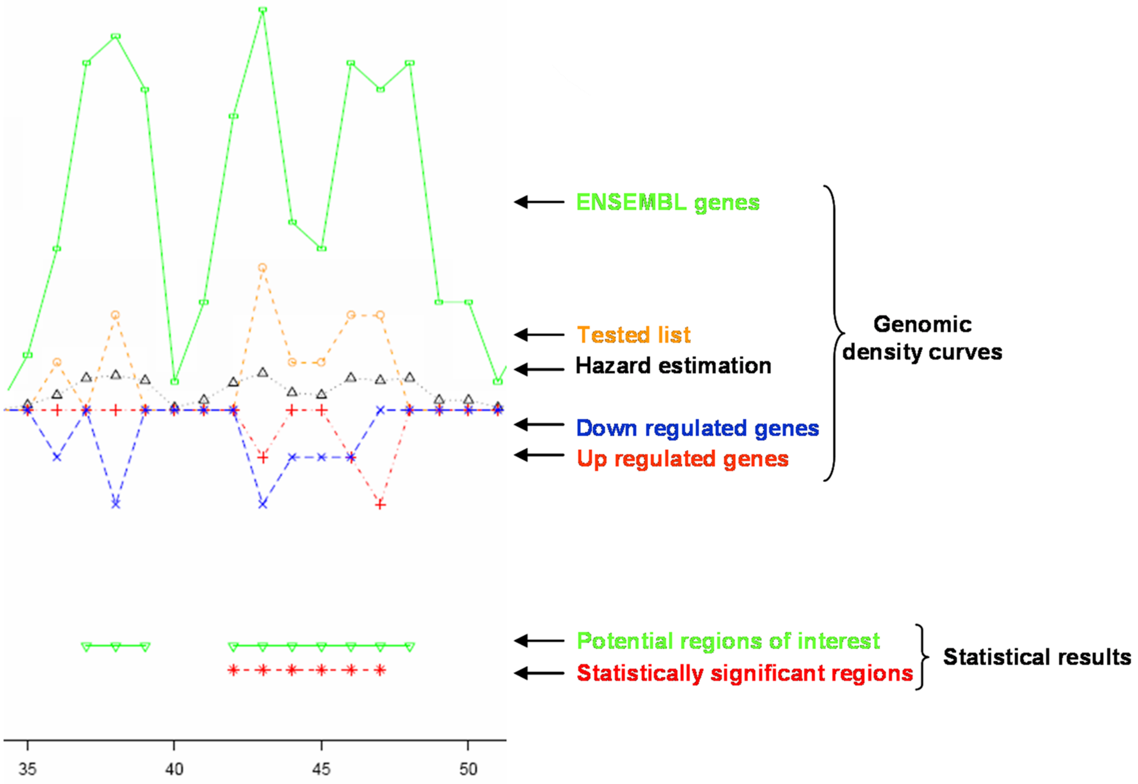
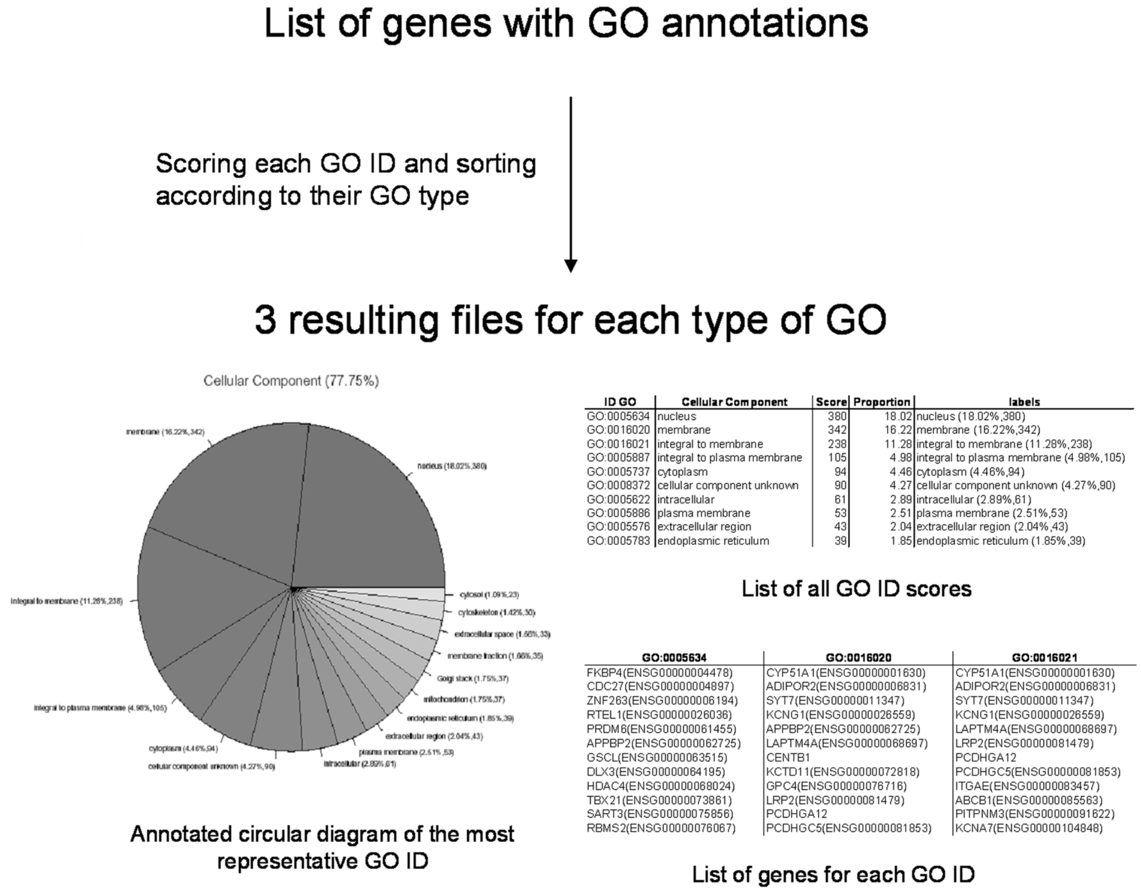


Figure 5: Genomic map and statistical results. One genetic map is generated for each chromosome. The map shows the density curves of the ENSEMBL genome (green), the tested gene list (orange) and the estimated hazard (black). The tested gene list curve is divided into up- (red) and down regulated (blue) curves drawn in mirror.


List of genes with GO annotations
Scoring each GO ID and sorting according to their GO type
3 resulting files for each type of GO
Cellular Component (77.75%)
membrane (16.22%,342)
nucleus (18.02%,380)
integral to membrane (11.28%,238)
cytosol (1.09%,23)
cytoskeleton (1.42%,30)
extracellular space (1.56%,33)
membrane fraction (1.66%,35)
Golgi stack (1.75%,37)
mitochondrion (1.75%,37)
endoplasmic reticulum (1.85%,39)
extracellular region (2.04%,43)
plasma membrane (2.51%,53)
intracellular (2.89%,61)
integral to plasma membrane (4.98%,105)
cytoplasm (4.46%,94)
cellular component unknown (4.27%,90)
Annotated circular diagram of the most representative GO ID
List of all GO ID scores
List of genes for each GO ID

ID GO	Cellular Component	Score	Proportion	labels
GO:0005634	nucleus	380	18.02	nucleus (18.02%,380)
GO:0016020	membrane	342	16.22	membrane (16.22%,342)
GO:0016021	integral to membrane	238	11.28	integral to membrane (11.28%,238)
GO:0005887	integral to plasma membrane	105	4.98	integral to plasma membrane (4.98%,105)
GO:0005737	cytoplasm	94	4.46	cytoplasm (4.46%,94)
GO:0008372	cellular component unknown	90	4.27	cellular component unknown (4.27%,90)
GO:0005622	intracellular	61	2.89	intracellular (2.89%,61)
GO:0005886	plasma membrane	53	2.51	plasma membrane (2.51%,53)
GO:0005576	extracellular region	43	2.04	extracellular region (2.04%,43)
GO:0005783	endoplasmic reticulum	39	1.85	endoplasmic reticulum (1.85%,39)

GO:0005634	GO:0016020	GO:0016021
FKBP4(ENSG00000004478)	CYP51A1(ENSG00000001630)	CYP51A1(ENSG00000001630)
CDC27(ENSG00000004897)	ADIPOR2(ENSG00000006831)	ADIPOR2(ENSG00000006831)
ZNF263(ENSG00000006194)	SYT7(ENSG00000011347)	SYT7(ENSG00000011347)
RTEL1(ENSG00000026036)	KCNG1(ENSG00000026559)	KCNG1(ENSG00000026559)
PRDM6(ENSG00000061455)	APPBP2(ENSG00000062725)	LAPTM4A(ENSG00000068697)
APPBP2(ENSG00000062725)	LAPTM4A(ENSG00000068697)	LRP2(ENSG00000081479)
GSCL(ENSG00000063515)	CENTB1	PCDHGA12
DLX3(ENSG00000064195)	KCTD11(ENSG00000072818)	PCDHGC5(ENSG00000081853)
HDAC4(ENSG00000068024)	GPC4(ENSG00000076716)	ITGAE(ENSG00000083457)
TBX21(ENSG00000073861)	LRP2(ENSG00000081479)	ABCB1(ENSG00000085563)
SART3(ENSG00000075856)	PCDHGA12	PITPNM3(ENSG00000091622)
RBMS2(ENSG00000076067)	PCDHGC5(ENSG00000081853)	KCNA7(ENSG00000104848)

Figure 6: Gene Ontology analysis. GO identifiers of each genes from the user gene list are processed to produce a scoredlist of all GO IDs, a graphical illustration of the most representative GO identifiers and the lists of genes for each GO ID.

number of genes from the tested list and the expected number of genes from the ENSEMBL genome are summarized per scale unit and stored in a matrix. At this stage the hazard computation can start. The hazard computation scales the ENSEMBL genomic distribution to the size of gene list generated by microarray analysis. For each unit, hazard is computed as follows:

Specific hazard of a unit = Number of ENSEMBL genes of the current unit * (Total number of tested genes / N ENSEMBL genes).

N ENSEMBL genes is computed by summarizing all genes of ENSEMBL genome for the first step and, in the second step, taking into account only ENSEMBL genes which are localized in units that contains at least one gene from the tested list (Figure 3).

Statistical tests and progressive optimization

First of all, two global statistical tests are performed to select chromosomes showing potential interest (Figure 3A). The CHI2-Goodness of fit test and the Wilcoxon paired sample test are the two global statistical tests used to compare expected with observed gene distributions, chromosome by chromosome.

Then a local test is performed for chromosomes validated by at least one global test (this step is customizable by the user) (Figure 3B). The binomial test is used to detect units in which observed gene population is statistically different from expected (Figure 3C to D). These units showing potential interest are then filtered, according to their significance and the difference between expected and observed distributions (this second selection is customizable by the user) before being extend to regions of potential interest (Figure 3E to G). A last step of optimization is performed to adjust statistically selected regions to seek the largest significant regions of interest (Figure 4).

To preserve clarity, only significant results are graphically reported. All statistical results, even those that are nonsignificant, are reported in the final result file.

To facilitate the interpretation of the results, the total effective curve is separated into curves of up- (red) and a curve of down regulated (blue) genes (Figure 5). In an optional part the annotated genes are sorted and analyzed by Gene Ontology (Figure 6).

All the data produced and formatted by GExMap are saved in a text file. These data allow the user to identify genes located in a region of interest, or to verify data used by the statistical tests.

Conclusions

GExMap provides helpful new types of information to interpret data generated by microarray studies. The software combines easy use and easy interpretation of results. Source code is freely available, optimized and annotated to facilitate customization, as implementation of different types of statistical tests or customized detection of interesting regions. All results are separately generated to preserve clarity of graphic representations. We have tested the software with a list of 3855 genes generated from microarray analysis[19]. GExMap allowed us to focus on a limited number of interesting regions which are now under biological investigation. Further optimizations are planned to complete the generated information. We work on integration of new statistical tests, as a t test adapted to small effective samples with a resampling pre-processing step. To improve compatibility with most of database identifiers and microarray types, a tool dedicated to the generation of correspondence files from text data files will be soon available.

Additional functions for comparisons between microarray gene lists or genomic association studies and to map and analyze several types of genetic markers, genetic association studies, and genetic maps of all kinds are also under development.

Availability and Requirements

GExMap is publicly accessible from the URL http://gexmap.site.voila.fr/. Questions and comments are welcomed through the site.

References

1. The R Project for Statistical Computing. » CrossRef » Pubmed » Google Scholar
2. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2008) Ensembl 2008. Nucleic Acids Res 36:D707–D714. » CrossRef » Pubmed » Google Scholar
3. Ensembl. » Google Scholar
4. Schena M, Shalon D, Davis RW, Brown P (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470. » CrossRef » Pubmed » Google Scholar
5. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression

monitoring. Science 286:531-7. » CrossRef » Pubmed » Google Scholar

6. Klein U, Tu Y, Stolovitzky GA, Mattioli M, Cattoretti G, et al. (2001) Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. J Exp Med 194: 1625-38. » CrossRef » Pubmed » Google Scholar

7. Hans CP, Weisenburger DD, Greiner TC, Gascoyne RD, Delabie J, et al. (2004) Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. Blood 103: 275-82. » CrossRef » Pubmed » Google Scholar

8. Devauchelle V, Marion S, Cagnard N, Mistou S, Falgarone G, et al. (2004) DNA microarray allows molecular profiling of rheumatoid arthritis and identification of pathophysiological targets. Genes Immun 5: 597-608. » CrossRef » Pubmed » Google Scholar

9. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14: 1675-80. » CrossRef » Pubmed » Google Scholar

10. Harrington CA, Rosenow C, Retief J (2000) Monitoring gene expression using DNA microarrays. Curr Opin Microbiol 3: 285-91. » CrossRef » Pubmed » Google Scholar

11. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J (2005) Independence and reproducibility across microarray platforms. Nat Methods 2: 329-30. » CrossRef » Pubmed » Google Scholar

12. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, et al. (2005) Multiple-laboratory comparison of microarray platforms. Nat Methods 2: 329-30. » CrossRef » Pubmed » Google Scholar

13. Petersen D, Chandramouli GV, Geoghegan J, Hilburn J, Paarlberg J, et al. (2005) Three microarray platforms: an analysis of their concordance in profiling gene expression. BMC Genomics 6: 63. » CrossRef » Pubmed » Google Scholar

14. Genespring 7.2. » CrossRef » Google Scholar

15. Ingenuity. » CrossRef » Google Scholar

16. Pathway Assist. » CrossRef » Google Scholar

17. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, et al. (2005) A network-based analysis of systemic inflammation in humans. Nature 437: 1032-7. » CrossRef » Pubmed » Google Scholar

18. Hoheisel JD (2006) Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet 7: 200-10. » CrossRef » Pubmed » Google Scholar

19. The Gene Ontology. » CrossRef » Google Scholar

20. Cagnard N, Letourneur F, Essabbani A, Devauchelle V, Mistou S, et al. (2005) Interleukin-32, CCL2, PF4F1 and GFD10 are the only cytokine / chemokine genes differentially expressed by in vitro cultured rheumatoid and osteoarthritis fibroblast-like synoviocytes. Eur Cytokine Netw 16: 289-92. » CrossRef » Pubmed » Google Scholar

21. 20- GExMap web site. » CrossRef » Google Scholar

22. MAQC consortium (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24: 1151-1161. » CrossRef » Pubmed » Google Scholar