

DNA Microarray Data Analysis

IIRIS HOVATTA, KATJA KIMPPA, ANTTI LEHMUSSOLA, TOMI PASANEN,
JANNA SAARELA, ILANA SAARIKKO, JUHA SAHARINEN, PEKKA TIIKKAINEN
TEEMU TOIVANEN, MARTTI TOLVANEN, MAUNO VIHINEN AND GARRY WONG
EDITORS JARNO TUIMALA AND M. MINNA LAINE

CSC

CSC – Scientific Computing Ltd. is a non-profit organization for high-performance computing and networking in Finland. CSC is owned by the Ministry of Education. CSC runs a national large-scale facility for computational science and engineering and supports the university and research community. CSC is also responsible for the operations of the Finnish University and Research Network (FUNET).

All rights reserved. The PDF version of this book or parts of it can be used in Finnish universities as course material, provided that this copyright notice is included. However, this publication may not be sold or included as part of other publications without permission of the publisher.

© The authors and
CSC – Scientific Computing Ltd.
2005

Second edition

ISBN 952-5520-11-0 (print)

ISBN 952-5520-12-9 (PDF)

<http://www.csc.fi/oppaat/siru/>

<http://www.csc.fi/molbio/arraybook/>

Printed at
Picaset Oy
Helsinki 2005

List of Contributors

Iiris Hovatta
National Public Health Institute
Haartmaninkatu 8
FI-00290 Helsinki
Finland

Katja Kimppa
PerkinElmer Life Sciences and Analytical Sciences
- Wallac Oy
P.O.Box 10
FI-20101 Turku
Finland

M. Minna Laine
CSC, the Finnish IT center for science
Keilaranta 14
FI-02101 Espoo
Finland

Antti Lehmussola
Tampere University of Technology
P.O.Box 553
FI-33101 Tampere
Finland

Tomi Pasanen
University of Helsinki
P.O.Box 68
FI-00014 University of Helsinki
Finland

Janna Saarela
Biomedicum Biochip Center
Haartmaninkatu 8
FI-00290 Helsinki
Finland

Ilana Saarikko
University of Helsinki
P.O.Box 68
FI-00014 University of Helsinki
Finland

Juha Saharinen
National Public Health Institute
Haartmaninkatu 8
FI-00290 Helsinki
Finland

Pekka Tiikkainen
VTT
P.O.Box 106
FI-20521 Turku
Finland

Teemu Toivanen
Centre for Biotechnology
Tykistökatu 6
FI-20521 Turku
Finland

Martti Tolvanen
Institute of Medical Technology
Biokatu 8
FI-33520 Tampere
Finland

Jarno Tuimala
CSC, the Finnish IT center for science
Keilaranta 14
FI-02101 Espoo
Finland

Mauno Vihinen
Institute of Medical Technology
Biokatu 8
FI-33520 Tampere
Finland

Garry Wong
A. I. Virtanen -institute
University of Kuopio
FI-70211 Kuopio
Finland

1 Web extra: Data mining for promoter sequences

Martti Tolvanen, Jarno Tuimala and Mauno Vihinen

1.1 Using BioMart to retrieve promoter regions

The preferred option for retrieving upstream sequences is found at the BioMart service (<http://www.ensembl.org/Multi/martview>) of the Ensembl project. This is limited only by the number of Ensembl genes that are annotated and by the accuracy of the annotation. Many genes still have many multiple entries in Ensembl, and some entries are for pseudogenes. However, the service is very easy to use. Note that you will want to use primarily Ensembl data, not Vega (which is manually annotated but still very incomplete data).

For some microarrays (especially Affymetrix chips), BioMart provides direct mappings. Such mappings with microarray contents have become more common in the genome sites, both at Ensembl and at UCSC, and the annotations provided by the manufacturers have improved, too. Therefore, retrieving the upstream sequences is less of a technical problem now, but the problem of correct transcription start sites remains a serious one, especially in the case of alternative promoters..

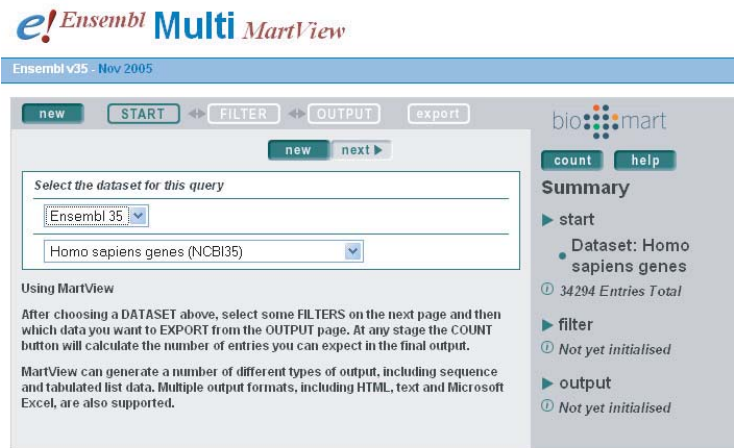


Figure 1.1: BioMart start screen.

Next, enter your list of gene identifiers in the Filter step:

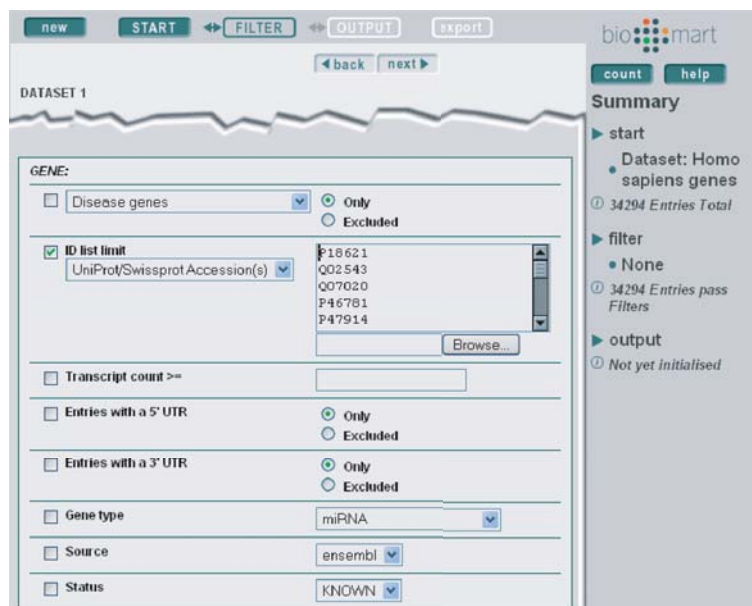


Figure 1.2: BioMart filtering. There are many options, and you should be aware that some mappings are more complete than others. Next to internal Ensembl references, RefSeq is your best choice. This example comes from data in which reliable SwissProt codes were provided. A long list of other filtering options is omitted from the figure.

Then, in the following output phase, you can optionally first choose your output as Features, to check the completeness and consistency of your results, and finally as Sequences to retrieve the data. The Features export gives a tabular out-

put which can be imported easily to other programs, e.g. by a direct copy/paste to MS-Excel..



Figure 1.3: Changing between Features and Sequences.

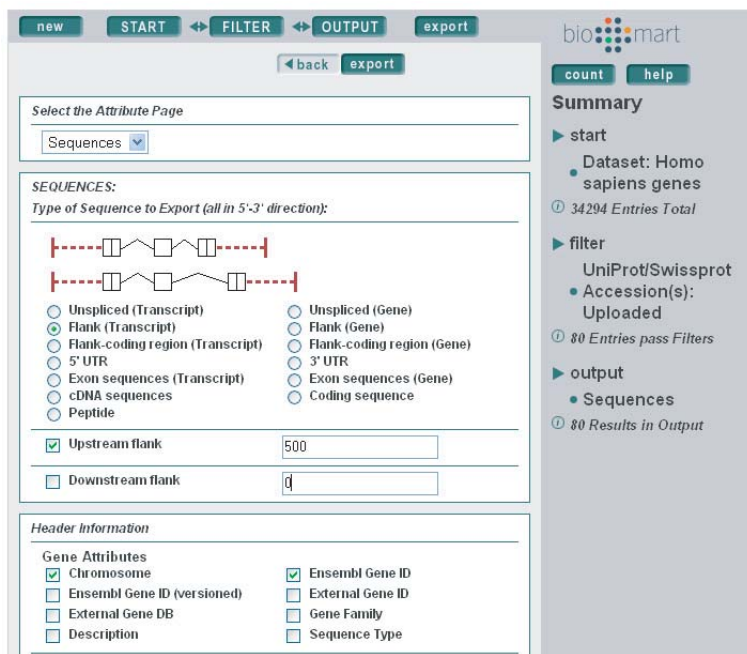


Figure 1.4: Sequence output options for obtaining only the 5'-upstream flank of your found genes in BioMart. You may want to add more options in the Header information and/or select a longer sequence region. Additional options for data compression, saving locally etc. are not shown.

script variants than the current RefSeqs. UCSC genome browser relies directly to the RefSeq start sites in their upstream data sets.

- Ensembl tries to extend the 5'-end as far as possible, but in a few cases this leads erratically long sequences, due to seemingly mis-spliced mRNA versions

1.2 GeneSpring and promoter analysis

GeneSpring includes a promoter analysis tool, which can be used for finding novel common regulatory sequences in a gene list, or to search for a known sequence. The tool can be invoked from *Tools->Find potential regulatory sequences*. In order to search for potential regulatory sequences, you need to have a whole genomic sequence of the organism under study. In principle, if only a partial genome of the organism is known, it is not possible to search for regulatory elements (GeneSpring forbids the use of the tool), because the statistical support and frequency values of the elements would be erroneous. However, there is a trick, which allows the analysis of partial genomes. For more information, see the tech note at http://www.silicongenetics.com/cgi/TNgen.cgi/GeneSpring/GSnotes/Notes/how_contig.

The tool opens a new window (Figure 1.7). First, you need to select a genelist you want to study, but do not use the “all genes” or “all genomic elements” list, because then you would compare the whole genome against itself, which is not a viable analysis. From the pull-down menu, select whether you want to search for new sequences or for a specific sequence. You can also select the length of the sequence to be considered a promoter region, how long a regulatory element is being searched, and how many unknown bases are allowed. The longer the sequence, and the larger the number of unknown bases, the longer the analysis time. You have control over the probability statistics: The p -value cut-off for a significant pattern can be modified. Whether the sequence is relative to the sequence upstream of other genes or relative to the whole genomic sequence can also be modified. The first option is far more common.

After the analysis have completed, or you stop the search, the results are reported. They appear on right side of the toolbox. Potential regulatory sequences, the number of genes they were detected in, and the detection p -value are reported. The best findings are reported first.

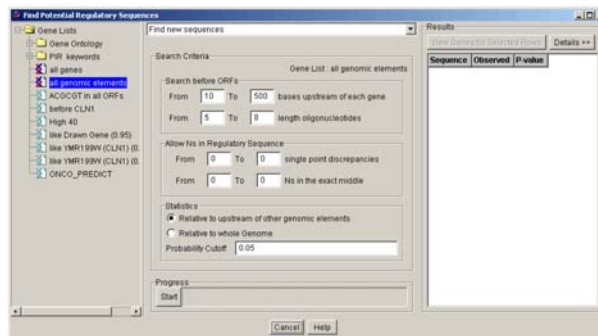


Figure 1.7: Find potential regulatory elements -tool in GeneSpring.