RNA-Seq data analysis based on Ion Torrent Personal Genome Machine data

Example analysis using Genomatix technologies to study two RNA-Seq data sets generated on an Ion Torrent™ Personal Genome Machine (PGM™).

Data source

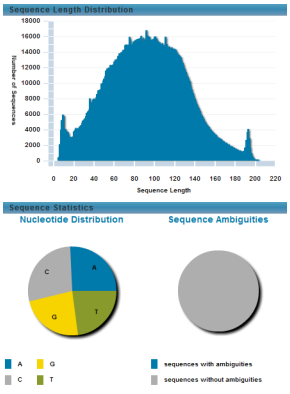
Data was obtained from the "Torrent Dev" section of life technologies' ion community website (http://lifetech-it.hosted.jivesoftware.com). We used the FASTQ files from two MAQC RNA data sets (Ambion® Human Brain Reference (HBR) and Stratagene® Universal Human Reference (UHR)). There were 5 files each for HBR (1.4 to 1.7 million reads, 5 to 203 bp) and UHR (1.6 to 2.3 million reads, 5 to 203 bp).

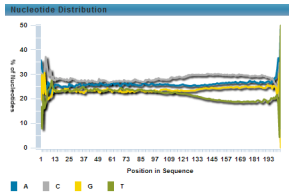
First level analysis

Mapping of reads was performed on the Genomatix Mining Station (GMS). The graphical user interface of the GMS allows biologist to run first level analyses of NGS data quickly and intuitively.

The Genomatix Mining Station interface showing the read classification results for a single human brain reference result.

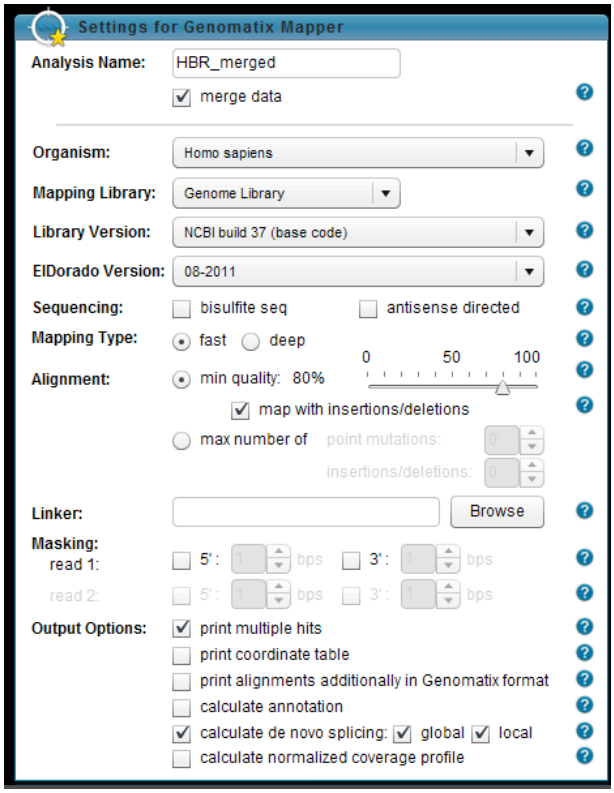
Below are some of the basic statistics that are automatically generated on the GMS after importing the data:





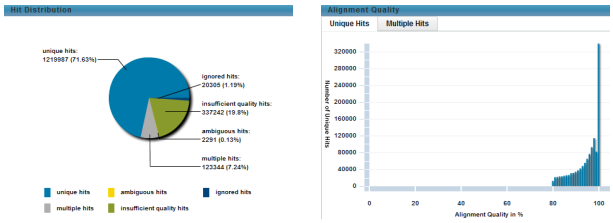
Sequence statistics on the GMS: histogram of read lengths (top left), plot of nucleotide distribution vs sequence position (top right), general nucleotide distribution (bottom left).

The parameters used for the mapping are shown in this screenshot:



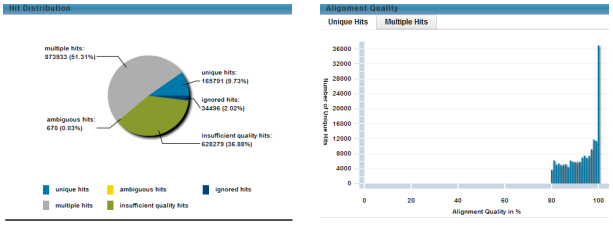
Parameters for the mapping: the 5 data files for each of the two experiments were mapped against the "genome" library; the long reads from the PGM allow fast mode; a minimum of 80% mapping quality and mapping with InDels were chosen to account for platform specific features; reporting of multiple matches and spliced alignment to detect reads spanning exon-exon junctions were also selected.

Mapping for the two full datasets took about 1 hour each with the majority of reads mapping uniquely to the human genome:



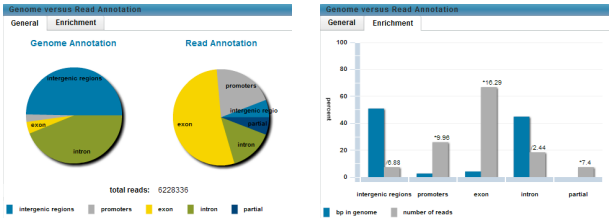
More than 70% of reads mapped uniquely, ~7% had multiple matches.

Mapping the datasets against the transcriptome library resulted in multiple matches for the majority of reads, reflecting the shared exon usage amongst alternative transcript variants:



Statistics from one of the transcriptome mapping results. All of the views on a GMS can be exported to PDF files.

Running the read classification task for the unique genomic hits showed a strong enrichment within exonic regions of the genome as expected for RNA-Seq experiments:

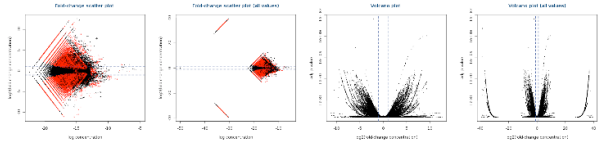


16-fold enrichment in exonic regions is shown in this read classification result. Users can view these as pie-chart comparison to genomic annotation (left) or as histograms (right). Reads are classified into promoters, exons, introns, intergenic regions and partial matches.

Downstream analysis

All results generated on a Genomatix Mining Station are immediately accessible on a connected Genomatix Genome Analyzer (GGA) for downstream analysis. The GGA offers a full set of tools and workflows for NGS data analysis including differential gene expression analysis, peak calling and classification, large scale SNP analysis, genome wide transcription factor (TF) analysis, tools for TF overrepresentation and TF association within peaks and regions. The visualization environment together with the comprehensive annotation, literature, transcription factor and pathway databases allow intuitive and coherent interpretation of experimental data in a biological context.

To compare the UHR and HBR datasets, a differential gene expression analysis using edgeR was performed for statistical evaluation (Audic/Claverie or DESeq are additional methods available on the GGA).



Fold change scatter plots and volcano plots generated on the GGA comparing the human brain RNA-Seq data against the universal human RNA-Seq data.

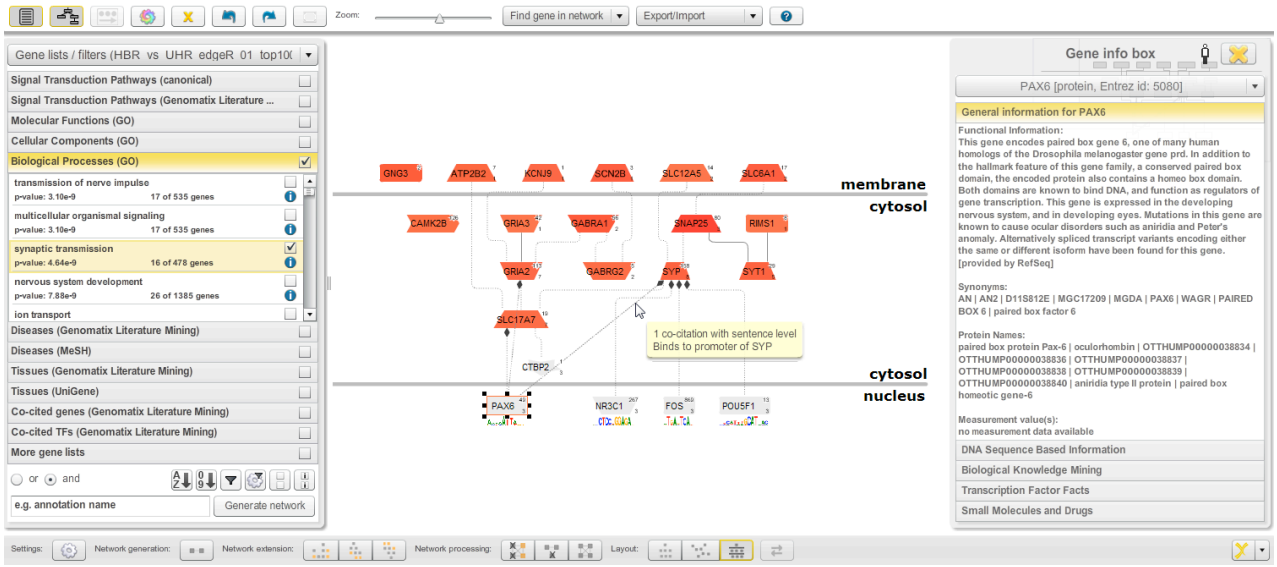
6,674 differentially expressed genes were identified by the analysis.

	Transcripts	Genes (known GeneId)
Total number analyzed	236890 download details (40Mb)	22413 download details (969Kb)
Differential expression	23607 download details (4.0Mb)	6674
Up-regulation	9767	2664 download details (204Kb)
Down-regulation	13840	4031 download details (316Kb)
Up- and down-regulated genes (with different transcripts)	-	21 download details (4.0Kb)

Overview table showing the number of analyzed genes and transcripts. Analysis is done on transcript level. "Up- and down-regulated genes" designates genes that have alternative transcripts with both, up- and down-regulation.

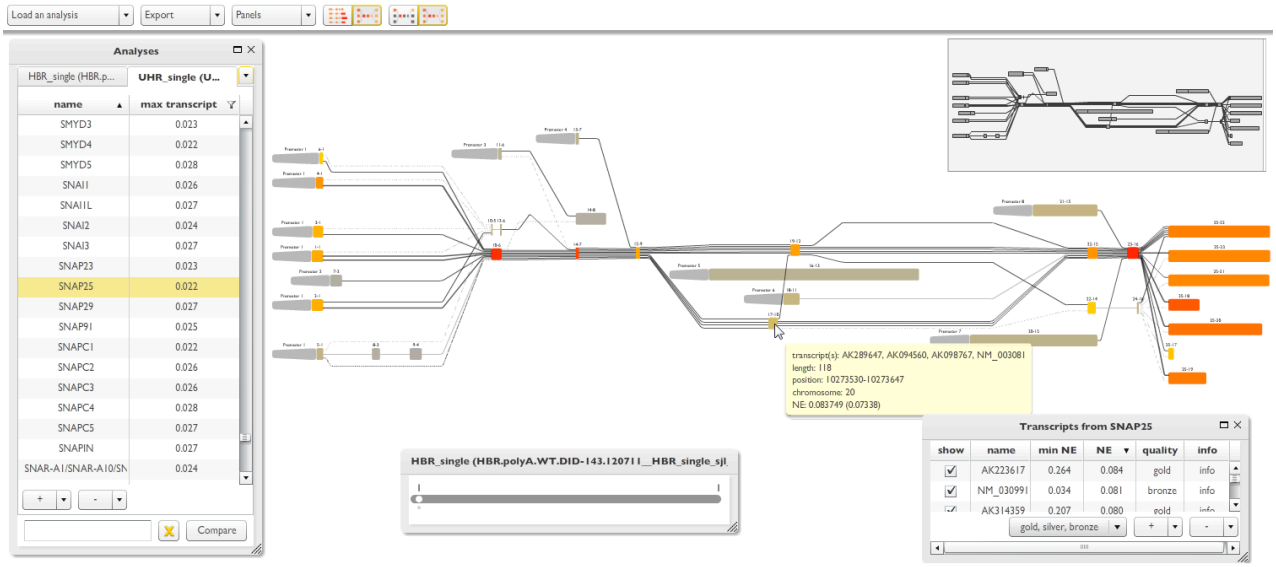
To interpret the data in a biological context, the differentially expressed genes can immediately be used as input for the Genomatix Pathway System (GePS). Gene sets can be classified based on canonical and literature based pathways. Associations with molecular functions, cellular components, biological processes, diseases and tissues based on GO- and MeSH-terms and extensive literature mining can be evaluated. Differential expression levels can be visualized. Networks can be complemented with meta-data like methylation or histone modifications. Extensive transcription factor data allows the exploration of gene regulatory connections within the networks.

Top scoring pathways and GO-terms for this analysis were brain related, as could be expected from the source of the experimental data. The top scoring disease and tissue were epilepsy and brain. The most co-cited gene is SNAP25, which is important for synaptic vesicle membrane docking and fusion. The most co-cited transcription factor REST encodes a transcriptional repressor that represses neuronal genes in non-neuronal tissues.

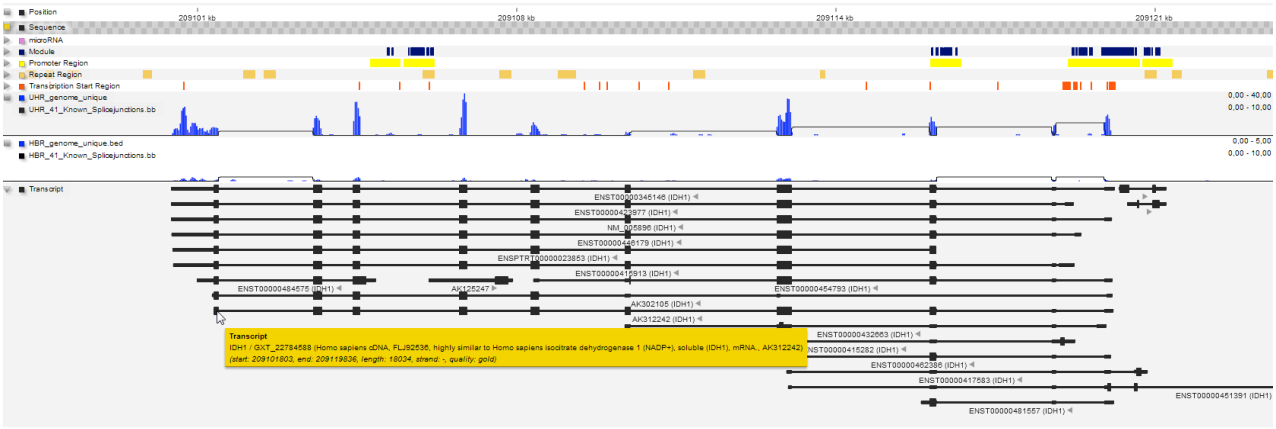


The Genomatix Pathway System (GePS). The main area in the center shows a network of those 16 genes from the input that are associated with synaptic transmission (the color saturation reflects the level of expression) plus 5 transcription (co-)factors that were added by automatic extension. Moving the pointer over an edge opens a popup containing the data known about the connection between the two nodes, moving it over a gene shows some information on the gene. Double-clicking an edge or node will show more detailed data in the information panel on the right (showing information on the selected PAX6 gene in this example). The panel on the left shows GO and MeSH associations and can be used to filter for pathways and associations and to combine filters for more restrictive analyses (e.g. "show overrepresented genes from the input set that have a tissue association with brain and a part of the nervous system development). Networks can be generated using different levels of co-citation or expert curation and can be extended automatically or manually. Several layouts are available.

In addition to differential expression between genes, RNA-Seq data can provide valuable insights into the differences in expression on transcript level. The Transcriptome Viewer (TViewer) on the GGA can be used to explore the alternative splicing variants of a gene overlayed with experimental data. Coverage of exons, known and de-novo splice junctions and promoter methylation can all be visualized for a detailed understanding of expression events. For data from paired end RNA-Seq experiments, distance profiles are available and gene fusions and read-throughs can be analyzed.



The Genomatix Transcriptome Viewer (TViewer). A merged view of all alternative transcripts of the SNAP25 gene is shown above. Exons shared between several transcripts are merged for a more condensed view. As in the Genomatix Pathway System the color saturation indicates the level of expression. The dark red exons are those with the highest normalized expression value. The thickness of the connection lines shows the level of splice junction coverage, with dotted lines indicating de-novo splice junctions. The panel on the left lists all genes and their maximum expression value. Different data sets can be accessed via the tabs on top. Comparisons between two datasets can be calculated on-the-fly to quickly identify differentially expressed genes. The central slider on the bottom allows users to switch between data from different conditions (cell lines, time points, patients, etc.). The panel on the bottom right shows all transcripts of the selected gene together with additional information including the expression value.



The Genomatix Genome Browser. The locus of the gene IDH1 is displayed together with annotation on known transcription factor modules (dark blue), promoters (yellow), repeat regions (orange), transcription start regions (red). The gene is on the reverse strand, therefore the promoters are at the right. There is one track each for the universal human reference (UHR) and the human brain reference (HBR) showing genomic (blue bars) and splice junction coverage (dark line). Underneath all transcripts for the locus are displayed. As can be clearly seen, expression of IDH1 in UHR is at a much higher level, both in terms of exon and splice junction coverage.

Interpreting data in their genomic context is possible via the Genomatix Genome Browser. Arbitrary user tracks for positional data can be added and visualized using different scaling, colors and display types. Overlaying of tracks is possible for an improved understanding of correlation between data (e.g. splice site coverage and expression level from RNA-Seq or DNA-Binding from ChIP-Seq and expression level from RNA-Seq, etc.). The zoom can be used to look into data from single nucleotide level to several million base pairs. In combination with Genomatix' extensive database of genomic annotations (containing transcripts, promoters, transcription start sites, transcription factor models, SNPs, microarray probe sets, repeats, microRNAs and SMARs) researchers can get all the information they need on their genes of interest.

For more information on Genomatix solutions and services, please visit:

http://www.genomatix.com

Visit

http://www.youtube.com/user/GenomatixWebcasts

for tutorials and demo videos.

Find us on facebook at:

http://www.facebook.com/genomatix

genomatix

http://www.genomatix.com

Contact Germany

Genomatix Software GmbH
Bayerstr. 85a
80335 Munich
Germany

phone +49 89 599766 0
email info@genomatix.de

Contact USA

Genomatix Software Inc.
3025 Boardwalk, Suite 160
Ann Arbor, MI 48108
USA

phone +1 877 436 6628
email sales-us@genomatix.com