illumina®

The Body Atlas Application

Data processing methodology and case studies for analyzing gene expression.

Introduction

The Body Atlas application within the BaseSpace Correlation Engine presents an aggregated analysis of gene expression across various tissues, cell types, and cell lines. It enables users to investigate expression of individual genes and gene sets.

The BaseSpace Engine library contains over 113,725 biosets obtained by mining the vast amounts of publicly available genomic data from sources such as the Gene Expression Omnibus, ArrayExpress, and other repositories. These data go through a systematic screening, curation, and analysis process (Figure 1).

A subset of these studies containing samples from normal tissues, cell types, and cancer cell lines are selected for inclusion into the aggregated Body Atlas analysis application. This technical note presents an overview of the data processing methodology for the Body Atlas application.

Body Atlas Data Collection

Table 1 shows the subset of studies collated under the Body Atlas application.

Studies are selected for inclusion based on the following criteria:

- Sample measurements generated from Affymetrix HG-U133A or HG-U133 Plus 2 platforms for human data, and Affymetrix GeneChip Mouse Genome 430A 2.0 Arrays for mouse studies.
- RNA samples obtained from normal tissues, stem cells, primary or cultured cell types, or cancer cell lines.

Data Processing

Expression Summary

dChip[1], MAS5[2], or RMA[3] typically preprocess raw data from various studies. A systematic curation process annotates each sample as subtypes of normal tissues, cell types, stem cells, or cancer cell lines.

Normalization

Expression intensity signals for all probesets in the 430A 2.0 platform (for mouse studies), and U133A and U133 Plus 2 platforms (for human studies) are normalized to a median value of 500. The intensities of probesets common to all platforms are used as a reference, and the data are compared across different studies using a **reference distribution transformation**. This approach has been shown to improve cross-platform analysis of microarray data sets.[4,5] Assuming that the distribution of gene expression is similar across all samples, a **universal reference** signal distribution is computed using all studies in the **tissue** category. It is then applied to 4 data sets (one each for tissues, cell types, cell lines, and stem cells). This results in 4 data sets whose expression intensities are normalized to the same universal reference.

Investigating Batch Effects

To assess study or batch effects after signal transformation, a combination of and ComBat[6] hierarchical clustering was applied to measures from common subtypes across studies. Samples in the same group or subtype tended to cluster closer than samples from the same study.

Average and Relative Gene Expression

Replicate samples for the same subtype (eg, brain cerebellum tissue, mononuclear cell of bone marrow) are grouped and the mean expression and standard deviation are computed for each probeset. Expression levels for a specific gene are comparable across all entries within the same panel: tissues in the **Tissues** tab, cell types in the **Cell Types** tab, cell lines in the **Cell Lines** tab, and stem cells in the **Stem Cells** tab. However, the overall expression range for each gene query (**Expression Level** column) is different. In addition, when switching between different Body Atlas categories, the gene expression scale is adjusted from minimum to maximum expression of a gene across all items in that specific category.

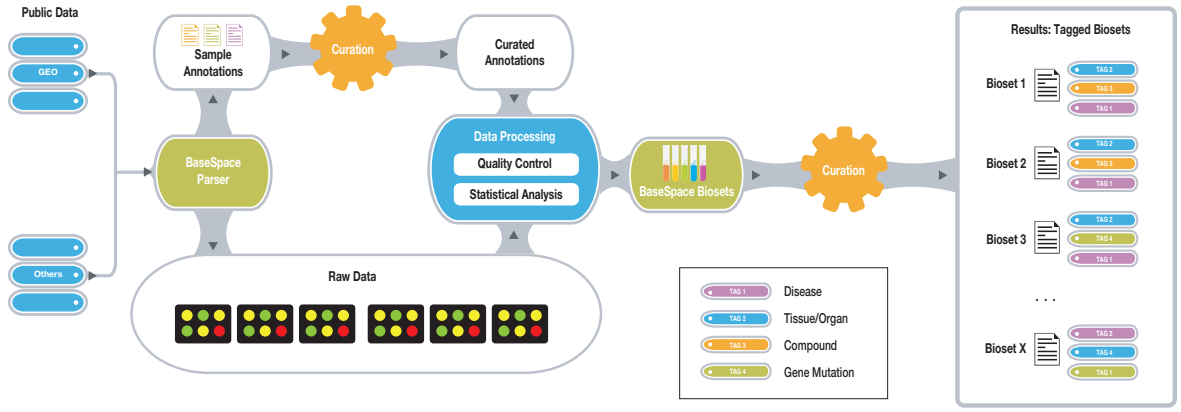

Figure 1: BaseSpace Correlation Engine Data Processing Workflow.

Table 1: Summary of Body Atlas Data[a]

Organism	Category	No. of Strains	No. of BaseSpace Engine Subtypes	No. of Studies	No. of Arrays
Mouse	Tissues	91	151	310	2730
	Cell Types	63	409	227	1585
Human	Tissues	NA	128	12	1068
	Cell Types	NA	170	98	1125
	Cell Lines	NA	748	47	881
	Stem Cells	NA	52	19	141

a. Current as of October 2014

Tissue- or cell-type–specific expression for genes can be identified in the Body Atlas application. Such genes form a gene expression signature for that tissue, cell type, stem cell, or cell line and can be compared with biogroups, internal biosets, or biosets uploaded by users.

Expression signatures for genes in a particular category are computed with reference to a median signal for each tissue group, or cell line and cell type. To minimize any bias resulting from an overpresented body subsystem, subtypes within a category are grouped to obtain a mean expression intensity signal for each subgroup. The median across each category is computed based on this mean signal.

Rank within a category denotes the relative importance of the query term within that bioset. Probesets are ranked by a score function that accounts for differential expression relative to the control and the corresponding standard deviation.

Tissue-Specific Gene Ranking

Tissue-specific genes are ranked based on 3 parameters: rank, percentage of maximum expression, and tissue-specificity index. For each tissue, specific genes of equal rank can be compared by their percentage of maximum expression. If 2 genes are equally significant based on the first 2 parameters, tissue-specificity index can be used as a measure to compare the relative specificity of expression.

Rank

Rank denotes the rank of the queried tissue for that particular gene.

Percentage of Maximum Expression

For each gene, tissues are ranked based on expression intensity and the maximum expression level identified. The percentage of maximum expression for all subsequent tissues is calculated as a fraction relative to this maximum expression of the gene.

The expression intensity of a gene listed under **Tissue-Specific Genes** in the queried tissue is at least 20% of the maximum expression for that gene.

Tissue-Specificity Index

In a list of tissues ranked by expression intensity for a particular gene, the decrease of expression going from one tissue (T1) to the one ranked immediately below (T2) is recorded as the specificity index.

$$Specificity\ Index = \frac{(Expression\ in\ \mathrm{T1} - Expression\ in\ \mathrm{T2})}{Expression\ in\ \mathrm{T2}}$$

For a given gene, if the specificity index for a particular tissue T1 is higher than the threshold (currently set to 0.8) and T1 is ranked within the top 10 tissues, the gene is considered to be specifically expressed in T1 and all tissues ranked above T1. In addition, a requirement of at least 20% of the maximum expression is imposed for any gene listed under **Tissue-Specific Gene**.

Body Atlas Use Cases

I. Querying Body Atlas with a Gene

Querying Body Atlas with the name of a gene, such as *ESR1*, results in the mean expression and standard deviation of all probesets mapped to the *ESR1* gene across all 4 panels: tissues, cell lines, cell types, and stem cells. The median expression for all subtypes within a panel is also indicated on each expression level bar.

Tissue subtypes in the category **Tissues** are organized into body systems. The **View By** menu organizes results either by body system (Figure 2) or rank. Selecting a particular body system within the ranks view will highlight tissues from that body system in the ranked list of tissues below (Figure 3). Clicking the arrow next to a tissue name opens the next level of results, providing probeset-specific information.

II: Querying Body Atlas with a Bioset

Body Atlas can also be queried using biosets curated by the BaseSpace Engine or created by users. The BaseSpace Engine computes enrichment p-values and an overall correlation direction. A positive value implies gene expression changes in the same direction in the query biosets and the tissue. A negative value implies that the expression in the query bioset and the tissue are in opposite directions.

In the example shown in Figure 4, Body Atlas is queried using the bioset *Primary hepatocyte _vs_ adult hepatic progenitor cell_ GPL10333*, which compares RNA expression differences in mouse primary hepatocytes and adult liver progenitor cells.

A strong positive correlation with liver tissue and hepatocyte cell types in both human and mouse validates the experiment that this particular bioset represents. This approach can be used to explore normal tissue similarities using cell types, or identify cell lines with similar patterns of expression to genes represented in a particular bioset.

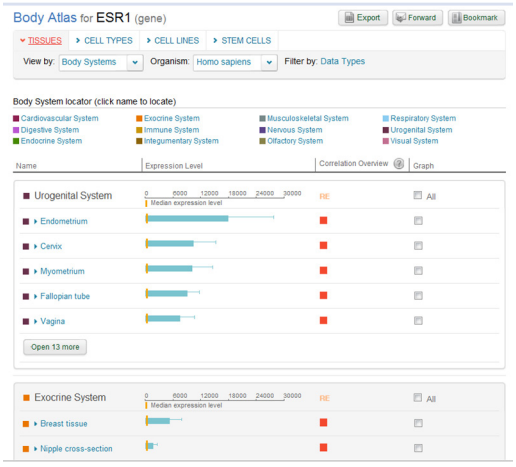

Figure 2: Querying with a Gene—Expression of probesets of the ESR1 gene across all tissues, organized by body system. The thick bar represents the overall expression level, while the thin bar represents standard deviation. Clicking the arrow next to the tissue name opens the next level of results that provide probeset-specific information.

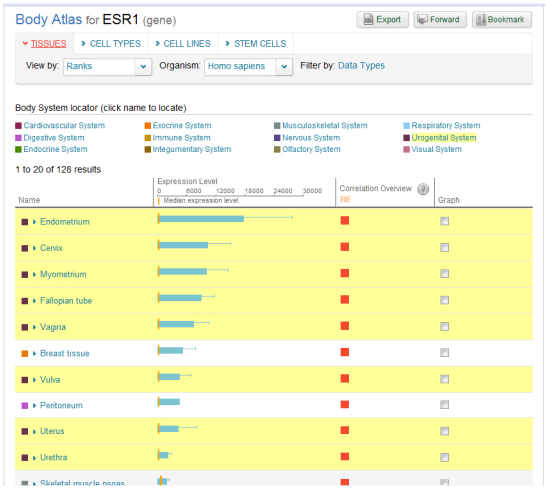

Figure 3: Querying with a Gene: Ranked Order—Expression profile of the ESR1 gene across a panel of normal tissues in their ranked order. The current Body System of interest (urogenital system) is selected, which automatically highlights the urogenital system tissues in the ranked list.

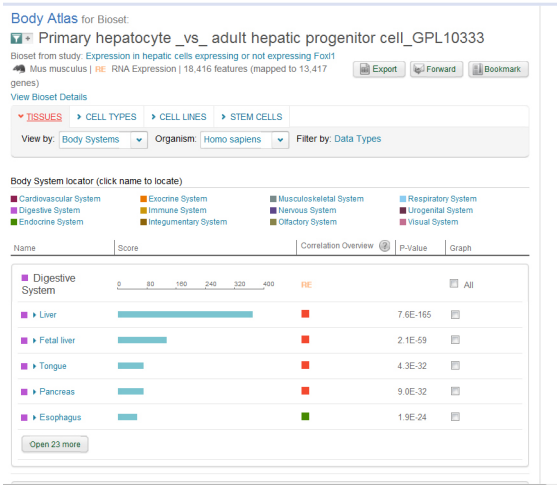

Figure 4: Querying with a Bioset—Query of a bioset across Body Atlas identifies strong positive correlation with digestive system–related cell types.

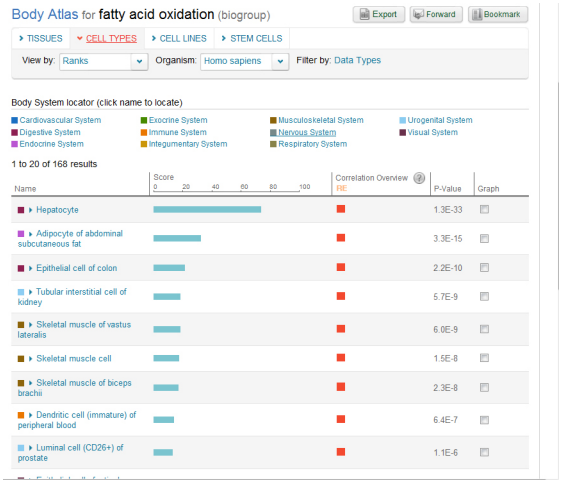

Figure 5: Querying with a Biogroup—Analysis of the fatty acid oxidation gene set (biogroup) across all tissues indicates hepatocytes as the top tissue in which fatty oxidation genes are upregulated.

Cell line and stem cell models identified could be used to design future experiments.

III. Querying Body Atlas with an Ontology Group (Biogroup)

Body Atlas can be used to identify tissues, cell types, cell lines, and stem cells where a particular set of genes from a given pathway or ontology group (a **biogroup**) is highly up- or downregulated relative to other subtypes within a panel.

Querying Body Atlas with a biogroup like *Fatty acid oxidation* uses the relative, ranked expression of all genes in the biogroup to compute enrichment p-values and overall direction of relative expression in tissues, cell types, cell lines, and stem cells (Figure 5).

Thus, when querying the biogroup **Fatty acid oxidation** across Body Atlas, the top-ranked results in cell types are hepatocytes and adipocytes. When filtering results by the direction of regulation, T lymphocytes have some of the lowest relative expression levels of this particular biogroup (Figure 6).

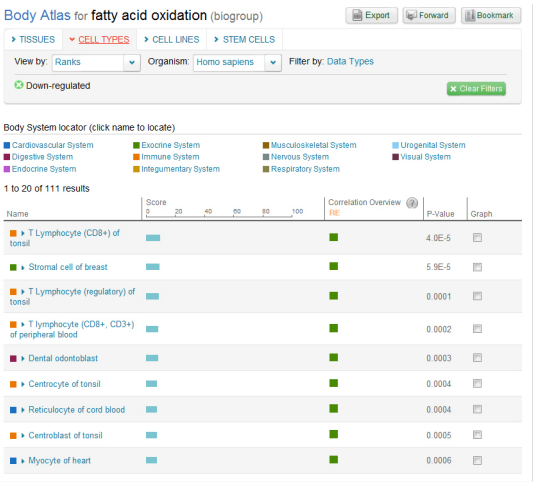

Figure 6: Querying with a Bioset: Filtered by Direction of Regulation— T-lymphocytes are the top tissue in which fatty acid oxidation genes are downregulated.

IV. Querying with a Tissue, Cell Type, or Cell Line

Querying Body Atlas with a tissue, cell type, or cell line will result in a ranked list of genes organized in 2 tabs: one for all genes, and one for tissue-specific genes. Results for cell line queries are displayed under 3 tabs that describe gene signatures for the cell line, namely **Gene Expression**, **Amplifications/Deletions**, and **Mutations**.

In the example shown in Figure 7, the query term **hypothalamus** was used to identify gene expression levels in this tissue. Results can be viewed for expression intensity of all genes in the hypothalamus, or tissue- specific genes only. The **Tissue-specific** genes tab includes genes that are specifically expressed in this tissue, ordered first by rank, followed by percentage of maximum expression, and lastly by specificity index. Within query results, 2 genes of equal rank can be compared for relative enrichment in a tissue by the percentage of maximum expression. If these 2 terms are identical, as for the *PMCH* and *OXT* genes in this example (Figure 7), the 2 genes can be compared by their specificity index. This index provides a measure of how much more this tissue (hypothalamus) expresses *PMCH* or *OXT* relative to the second-ranking tissue (superior cervical ganglion for *PMCH*, globus pallidum external for *OXT*).

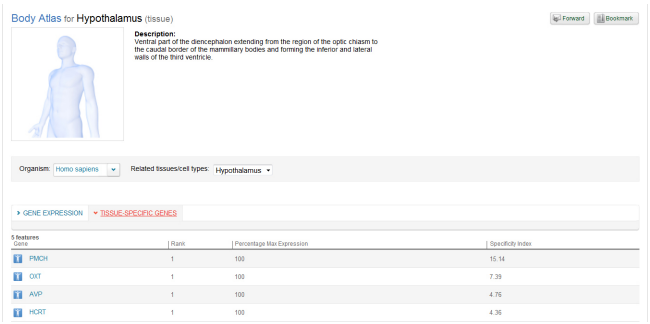

Figure 7: Querying with a Tissue— Tissue-specific genes, such as PMCH and OXT, can be compared by specificity index when both are expressed at the maximum level in a particular tissue.

Although hypothalamus is the top expressing tissue for both PMCH and OXT, the specificity of PMCH in hypothalamus is higher than that of OXT. The specificity indexes indicate that PMCH is expressed ~16-fold over the second- ranking tissue (superior cervical ganglion) while OXT is expressed ~8-fold over the second-ranking tissue (globus pallidum).

References

1. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA*. 2001; 98: 31-36.
2. Support by Product. Affymetrix web site. www.affymetrix.com/support/technical/byproduct.affx?product=mas. Accessed October 30, 2014.
3. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003; 4: 249-264.
4. Jiang H, Deng Y, Chen HS, et al. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics.* 2004; 5: 81.
5. Warnat P, Eils R, Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*. 2005; 6: 265.
6. Johnson WE, Rabinovic A, Li C. Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics*. 2005; 8(1): 118–127.

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

illumina®