illumina®

Data Correlation Details: Enrichment Analysis

An overview of bioset ranking algorithms for genomic analysis.

Introduction

One of the primary applications of the BaseSpace® Correlation Engine is performing metaanalyses that harness large amounts of genomic, proteomic, and assay data. Metaanalysis enables researchers to look for potentially novel and interesting results that cannot necessarily be seen by looking at a single, existing experiment. These results may be interesting in themselves (eg, associations between different treatment factors, or between a treatment and an existing known pathway or protein family), or may be used to guide further research and experimentation.[1,2]

The primary entity in these analyses is the **bioset**, a ranked list of elements (genes, probes, proteins, compounds, single-nucleotide variants [SNVs], sequence regions, etc.) that corresponds to a given treatment or condition in an experiment or an assay. For a gene expression experiment, the biosets will consist of gene lists with associated change values and statistical information for each relevant experimental factor. For example, a bioset could consist of a list of Affymetrix probesets, fold change difference values between treated and control groups, and p-values produced by a t test between treated and control sample probesets' intensity values. A sequence-centric bioset from a copy-number analysis experiment will contain ranked sequence regions with associated gain and loss statistics. An SNV bioset from a genome-wide association study (GWAS) will contain ranked SNVs with associated p-values.

In summary (Figure 1), biosets may contain one of many of the following columns:

- Identifier of an entity—gene, SNV, or sequence region (required)
- Other entity identifiers—chromosome, position (start/stop)
- Summary statistics—p-value, fold-change, score, rank, odds ratio

Bioset Ranking and Directionality

There are 2 major types of biosets:

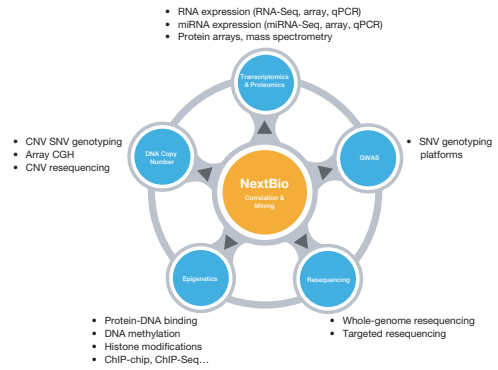
- Gene-centric (genes, proteins)
- Sequence-centric (SNVs, sequence regions)

The BaseSpace Engine enables enrichment analysis across both gene- and sequence-centric data.

Gene-Centric Biosets

The statistics associated with gene elements in a bioset are used to determine their ranks and directionality (see *Technical Note: Ranking of Genes, SNVs, and Sequence Regions*). In essence, statistics selected by the user (eg, fold change) are converted to nonparametric ranks, while fold-change direction gives those ranks a positive or negative sign (Figure 2A). In cases where a user imports a custom gene list without ranks or direction, a simplified enrichment analysis is applied. One of the important parameters associated with the bioset is the platform used to generate corresponding biosets (eg, Affymetrix U133 Plus 2.0).

Figure 1: Overview of Biosets



Biosets within the BaseSpace Engine comprise gene-centric or sequence-centric signatures from orthogonal types of biological experiments.

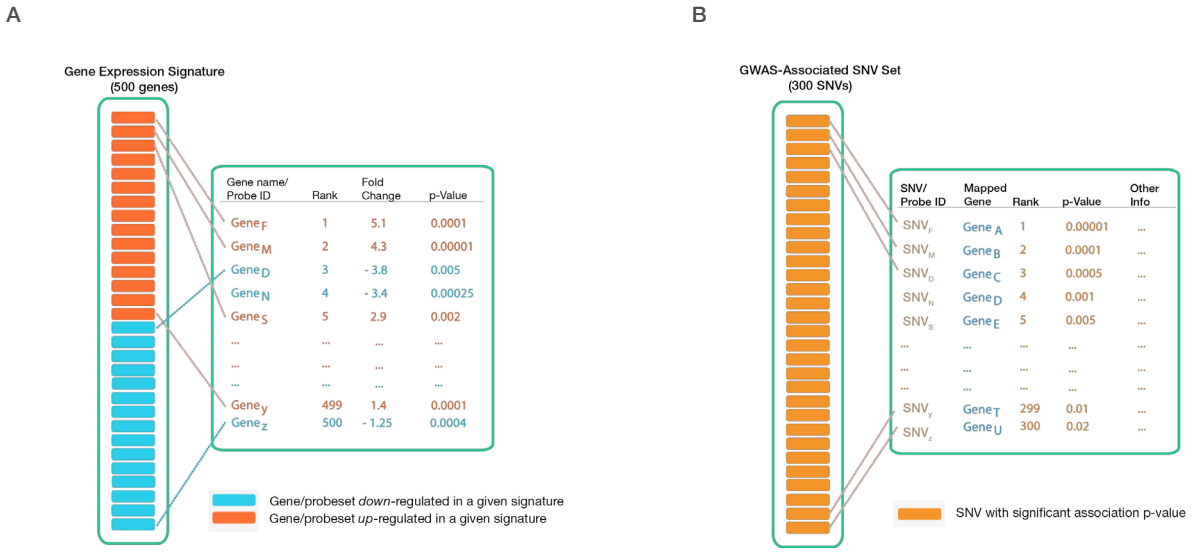
SNV/Sequence-Centric Biosets

For sequence-centric biosets that contain nongene elements, a mapping logic is defined to derive corresponding gene mappings (see *Technical Note: Ranking of Genes, SNVs, and Sequence Regions*). Subsequent enrichment analysis is performed on the derived set of mapped genes and associated ranks (Figure 2B). Future iterations of the algorithm will also enable users to apply enrichment analysis directly on SNVs/sequence regions within a bioset, independent of gene associations.

Rank-Based Directional Enrichment Algorithm

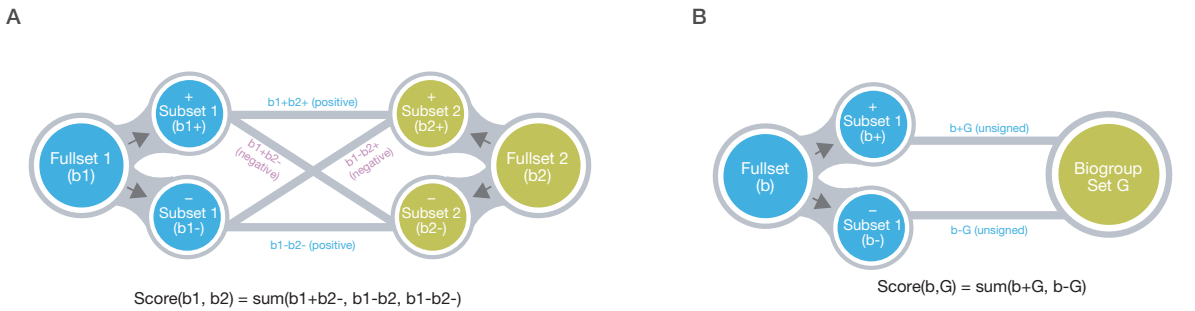
Illumina has developed a specialized geneset enrichment algorithm that is used to perform various analyses. The general design of the algorithm (Running Fisher) is analogous to the Gene Set Enrichment Analysis (GSEA) method.[3] Similar to GSEA, Running Fisher dynamically detects the most significant enrichment signal in a ranked signature set, allowing the signature set to contain relatively more comprehensive collections of genes at a preselected statistical cutoff. This dynamic detection approach overcomes the limitations of selection approaches where a cutoff that is too stringent might lead to potential loss of significant information. Alternatively, cutoffs that are too relaxed might include insignificant data into the evaluation.[4] The Running Fisher algorithm differs from GSEA in the assessment of the statistical significance, where p-values are computed by a Fisher's exact test rather than by permutations. Overall, the advantage of this approach is the flexibility of being able to compute correlation scores for data of different sizes and filter thresholds.

Figure 2: Bioset Ranking and Directionality



A) Results of different gene expression analysis biosets, showing ranks (based on fold change) and directionality (fold-change direction). B) GWAS association results bioset with a set of mapped genes, used in the enrichment analysis. Ranks for SNPs and mapped genes are derived from the corresponding p-value.

Figure 3: Rank-Based Directional Enrichment



Enrichment scores are computed for the directional subsets, followed by a summation over all the subset scores. A) Bioset vs. Bioset: the subsets are formed for both *b1* and *b2*, and subset-subset enrichment scores are computed. B) Bioset vs. Biogroup: the subsets are formed for the bioset, and subset-biogroup enrichment scores are computed.

The directional relationship between the 2 signatures is captured by the sign of the correlation score. Upregulated genes (*b+*) and downregulated genes (*b-*) are separated into directional subsets, and correlation scores are computed for each directional subset from one signature (*b1+*, *b1-*) against each subset from the other signature (*b2+*, *b2-*). A positive sign is given to a subset pair of the same direction (*b1+b2+*, *b1-b2-*), and a negative sign is given to a subset pair of opposite directions (*b1+b2-*, *b1-b2+*). The overall correlation score is the sum of directional subset scores, and the sign of the sum determines whether the 2 signatures are positively or negatively correlated (Figure 3A).

BaseSpace Engine biogroups are genesets that are not directional and not ranked. The enrichment of a biogroup in a bioset is computed for each directional subset (*b+*, *b-*) comprising the bioset. Because the biogroups are not directional, the subset scores are unsigned and the overall enrichment score is the sum of the subset scores (Figure 3B).

Computing Pairwise Correlation Scores

The correlation scores between gene signature (bioset) pairs are computed with the Running Fisher algorithm, a nonparametric rank-based statistical approach. Detailed steps for 2 signature sets (*b1, b2*) are as follows:

First, each gene signature set is rank-ordered according to fold change, p-value, or a particular score. If appropriate metrics are not provided, then the gene signature set is unranked. The upregulated genes and downregulated genes are noted with positive and negative signs to imply directionality, respectively. A directional subset is generated for each direction, such as *b1+*, *b1-*, *b2+*, and *b2-* (Figure 3A). If no directional data are provided, then the gene signature set is not directional and only one subset is formed with the whole signature set, such as *b1o*, or *b2o*.

Second, all the subset pairs are identified: *b1Di*, *b2Dj*, where *Di* and *Dj* are the available directions (+, -, or o) in *b1 and b2*, respectively. The Running Fisher algorithm is applied to each subset pair. The top 1% of genes in the platform in the first subset *b1Di* are collected as a group *G*. The second subset, *b2Dj*, is scanned top to bottom in the rank order to identify each rank with a gene matching a member in the group *G*. If the subset is unranked, all the genes in the subset are retrieved at the first scan.

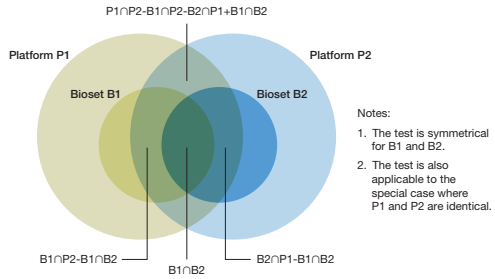
At each matching rank *K*, the scanned portion of the second subset *b2Dj* consists of *N* genes, and the overlap between group *G* and *N* genes is *M*. A Fisher's exact test is performed at rank *K*, to evaluate the statistical significance of observing *M* overlaps between a set of size *G* and a set of size *N*. In this analysis, the set of size *G* comes from platform *P1* and the set of size *N* comes from platform *P2*, given the sizes of *P1* and *P2* as well as the overlap between *P1* and *P2*. The parameters for Fisher's exact test are determined as shown in Figure 4.

At the end of the scan, the best p-value is retained, and the negative log of the best p-value is a score for the subset pair.

Next, the Running Fisher algorithm is performed in the reverse direction: the top-ranking genes in the second subset *b2Dj* are collected as a group *G*, and the first subset *b1Di* is scanned in the rank order. The same procedure in this reverse direction produces another score for the same subset pair. The 2 scores are averaged to represent the magnitude of the similarity between the 2 subsets. A positive sign is given to the final subset pair score if *Di and Dj* are the same. A negative sign is given if *Di and Dj* are opposite. The score is unsigned if any of *Di and Dj* are not directional.

Finally, the overall score is computed by summing up all directional subset pair scores (Figure 3A). The sign of the sum determines whether the 2 signatures are positively or negatively correlated. If both signatures are not directional, the overall score is the same as the only subset pair score that is unsigned. If one of the 2 signature sets is directional and the other is not directional, the overall score is represented by the larger of the 2 subset pair scores, annotated with the contributing direction from the directional signature. The matching genes between 2 typical gene signatures are depicted in Figure 5. The directionality and rank for each gene is shown.

Figure 4: Pairwise Correlation between Biosets



	In Bioset B2	Not in Bioset B2	Totals
Mapped to Bioset B1	B1∩B2	B1∩P2-B1∩B2	B1∩P2
Not mapped to Bioset B1	B2∩P1-B1∩B2	P1∩P2-B1∩P2 B2∩P1+B1∩B2	P1∩P2-B1∩P2
Totals	B2∩P1	P1∩P2-B2∩P1	P1∩P2

Fisher's exact test parameters for Bioset vs. Bioset—assessment of statistical significance on a bioset's enrichment in another bioset. When the bioset is directional, the direction-specific subset (*b1+* or *b1-* in place of B1; *b2+* or *b2-* in place of B2) is used. The scan of the bioset dynamically cuts off the top K rows of the gene set for evaluation at rank K.

Extending the Bioset-Bioset Scoring Algorithm

The bioset-bioset scoring algorithm can be extended to compute the correlation score between 2 biosets of other data types. For example, the algorithm can be used to compute correlation scores between 2 exon data sets, after the following requirements are satisfied:

1. Identification of exons from 2 different platforms can be mapped to a common standard set of exon IDs.
2. All exons in each platform are stored so that the overlap between one exon set with the platform for the other exon set can be computed, as well as the overlap between the 2 platforms.
3. Rank and directionality for each exon are defined and retrievable.

Similarly, other data types—such as isoform-specific transcript data sets or microRNA (miRNA) data sets—can be compared against any other data set within the space of same data types. Also, the mapping from these nongene data types to genes generates a gene-level representation of the data set, so that all the data sets can be compared at the gene level.

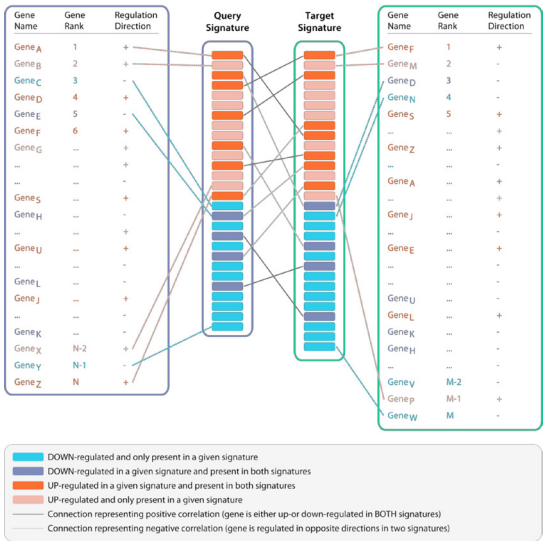
Computing the Enrichment Score for a Biogroup in a Bioset

The enrichment score for a biogroup in a bioset is computed with the Running Fisher algorithm. The detailed steps given a signature set (*b*) and a geneset in a biogroup (*C*) are as follows:

First, the gene signature set is rank-ordered and assigned directionality as previously described for biosets. Directional subsets are also generated as described (Figure 3B).

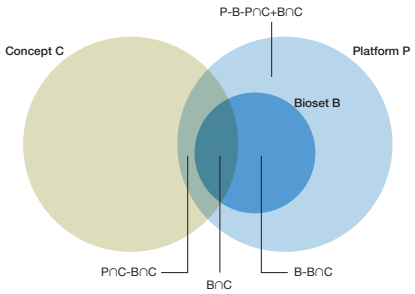
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AATCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACC
AACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTA
TTAACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGAT
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AAGATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGA
AACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAG

Figure 5: Pairwise Signature Correlation Scores



An outline of pairwise directional enrichment analysis between 2 signatures.

Figure 6: Bioset vs. Biogroup



	In Bioset B	Not in Bioset B	Totals
Mapped to C	B∩C	P∩C-B∩C	P∩C
Not mapped to C	B-B∩C	P-B-P∩C+B∩C	P-P∩C
Totals	B	P-B	P

Fisher's Exact Test parameters for Bioset vs. Biogroup—assessment of statistical significance on a biogroup's enrichment in a bioset. When the bioset is directional, the direction-specific subset (*b+* or *b-* in place of B) is used. The scan of the bioset dynamically cuts off the top K rows of the gene set for evaluation at rank K.

Second, the Running Fisher algorithm is applied to compute the enrichment of the biogroup genes in each subset. The subset *b+ or b-* is scanned top to bottom in the rank order to identify each rank with a gene matching a member in the biogroup *C*. If the subset is unranked, all the genes in the subset are retrieved at the first scan.

At each matching rank *K*, the scanned portion of the subset *b+ or b-* consists of *N* genes, and the overlap between group *C* and *N* genes is *M*. A Fisher's exact test is performed at rank *K*, to evaluate the statistical significance of observing *M* overlaps between a set of size *C* and a set of size *N*, where the set of size *N* comes from platform *P*, given the size of *P* and the overlap between *P* and *C*. The parameters for Fisher's exact test are determined as in Figure 6.

At the end of the scan, the best p-value is retained. The negative natural log of the best p-value is the enrichment score for the biogroup in the subset.

Finally, the overall score is computed by summing up the subset pair scores (Figure 3B). The sum score is unsigned, as are the subset scores.

References

1. Kupershmidt I, Su QJ, Grewal A, Sundaresh S, Halperin I, et al. (2010). Ontology-based metaanalysis of global collections of high-throughput public data. PLoS ONE 5(9): e13066.
2. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313(5795): 1929–1935.
3. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge- based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43): 15545–15550.
4. Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat 1(1): 85–106.

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

For Research Use Only. Not for use in diagnostic procedures.

© 2016 Illumina, Inc. All rights reserved. Illumina, BaseSpace, and the pumpkin orange color are trademarks of Illumina Inc., and its affiliate(s) in the U.S. and/or in other countries. Pub. No. 970-2014-007 Current as of 28 March 2016.

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA