# Enrichment Analysis with BaseSpace™ Correlation Engine

An overview of bioset ranking algorithms for genomic analysis.

## Introduction

BaseSpace Correlation Engine (formerly known as NextBio™ Research) provides life science researchers insightful scientific tools (eg, Body Atlas, Disease Atlas, Pharmaco Atlas, Knockdown Atlas, Genetic Markers, Meta-analysis) and unprecedented access to vast numbers of high quality, whole-genomic studies curated from public sources.

Designed to take advantage of continuously expanding content, a simple intuitive graphical interface (Figure 1) enables researchers to identify novel correlations with ease and efficiency. This powerful software as a service (SaaS) application allows users to share data seamlessly across multiple organizations and geographic locations, enabling impactful collaborations and partnerships.

This technical note describes how studies qualify for curation into BaseSpace Correlation Engine, how correlation scores are achieved, and how the Running Fisher algorithm provides flexibility in obtaining correlations from different sets of data.

## How BaseSpace Correlation Engine works

Public data are evaluated both programmatically and by stringent scientific review before they are uploaded into BaseSpace Engine. To date, over 80,000 studies have been evaluated to arrive at more than 22,000 included studies. A study must be of a supported species, have sufficient replicates, use a supported whole genome technology (eg, not PCR), have recognizable features (if array-based), have a sensible study design (ie, treated vs. untreated, disease vs. normal, etc.), use unique samples (not reanalysis), and pass data quality assurance.

For next-generation sequencing (NGS) studies, raw data is required, for which minimal read depth and RNA quality standards are imposed. For arrays, raw data is preferred, and processed data must be untransformed and normalized in a way comparable to a per-chip median normalization process (eg, RMA, MASS5) with no evident batch effects. Authors are contacted whenever annotations are unclear or discrepancies are found. New studies are added on a continuous basis.[1]

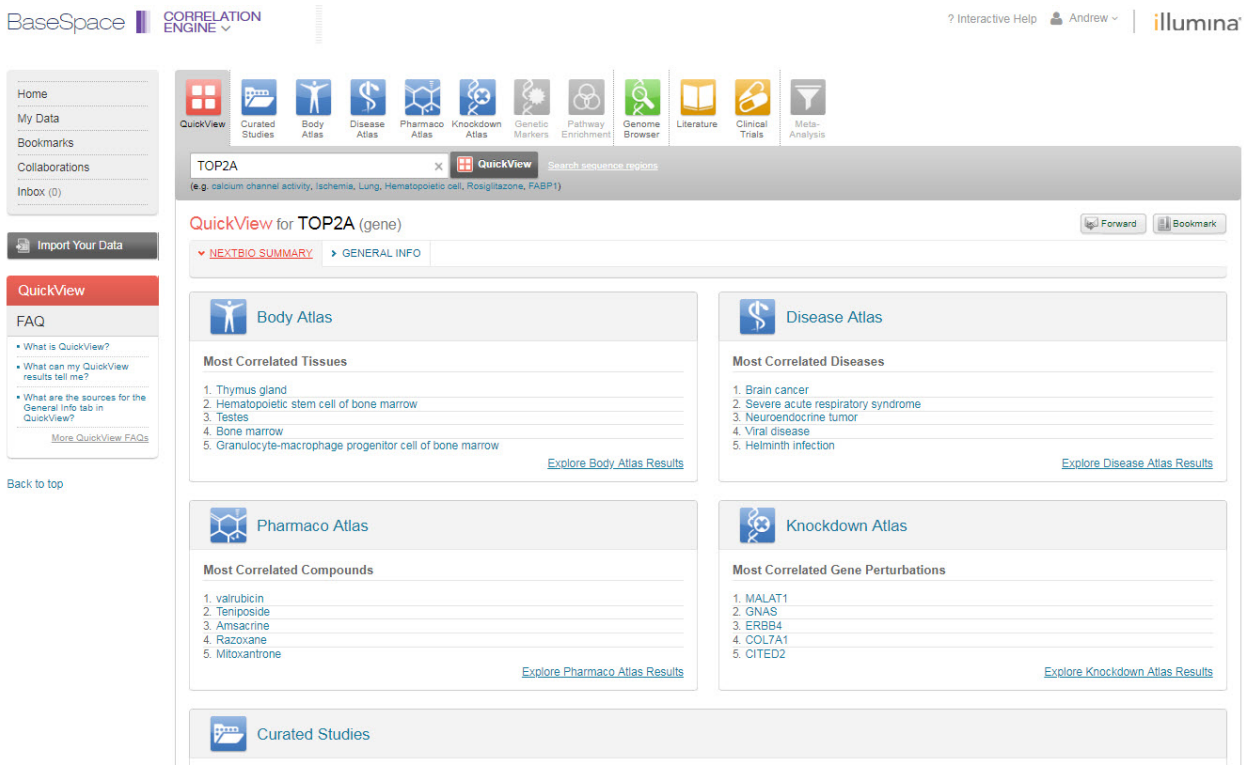


**Figure 1: BaseSpace Correlation Engine user interface enables queries for numerous association types —** Novel correlations and associations are quickly identified for a given query, revealing data driven connections between genes, diseases, compounds, tissues, pathways and literature.

## Biosets

The primary analysis result is referred to as a bioset, a list of elements such as genes, probes, proteins, compounds, small nucleotide variants (SNVs), sequence regions, etc. A bioset may consist of ranked or unranked elements corresponding to a given treatment, condition versus a baseline or reference condition, in an experiment.

For a gene expression experiment, biosets are derived for all relevant experimental factors within a study. Resulting gene signature lists have associated fold-change values and statistical information, such as p-values and q-values. Only statistically significant results will be reported for a given experiment. For example, an RNA sequencing (RNA-Seq) dataset is subject to statistical thresholds set at a minimum of 1.2 absolute fold change and a maximum adjusted p-value (or q-value) of 0.05. Then each bioset is associated with an array of ontology-based biomedical concepts that determine how query results are organized in the system.

The statistics associated with gene elements in a bioset is used to determine their ranks and directionality.[2] In brief, statistics selected by the user (eg, fold change) are converted to non-parametric ranks, while fold change direction gives those ranks a positive or negative sign (Figure 2). A mapping logic is defined to derive corresponding genes when studies measure chromosomal regions or SNPs.[2] Subsequent enrichment analysis is performed on the derived set of mapped genes and associated ranks (Figure 2). In cases where a user imports a custom gene list without ranks or direction, a simplified enrichment analysis is applied.[2]

## The Running Fisher algorithm

Illumina has developed a specialized gene set enrichment algorithm that forms the basis for a variety of analyses. The general design of the algorithm (Running Fisher) aims to compare a query signature with a target signature, which is analogous to the Gene Set Enrichment Analysis (GSEA) method.[3,4] Our Running Fisher algorithm dynamically detects the most significant enrichment signal in a ranked signature set by systematically scoring subsets of that list down to a preselected statistical cutoff. A comprehensive collection of gene lists is generated to report the most significant findings. 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 (Figure 3). Also, unlike GSEA, the Running Fisher can evaluate gene sets with both up and down regulated genes.

Overall, the advantage of this approach is the flexibility of being able to compute correlation scores for data of different types, sizes, and filter thresholds. The algorithm involves identifying subset pairs defined by available directionality in the data. The Running Fisher algorithm is applied to each subset pair (up vs. up, down vs. down, up vs. down, down vs. up) (Figure 3). The genes in the query subset are scanned from top to bottom in the rank order to identify each rank with a gene matching a member in the target signature. If the subset is unranked, all the genes in the subset are retrieved at the first scan. At the end of the scan, the best p-value is retained, and the negative log of the p-value is a score for the subset pair. Next, the Running Fisher algorithm is performed with the previous query signature as the target and the previous target signature as the query signature. The average of the two best scores is taken to measure the magnitude of similarity between the two subsets.

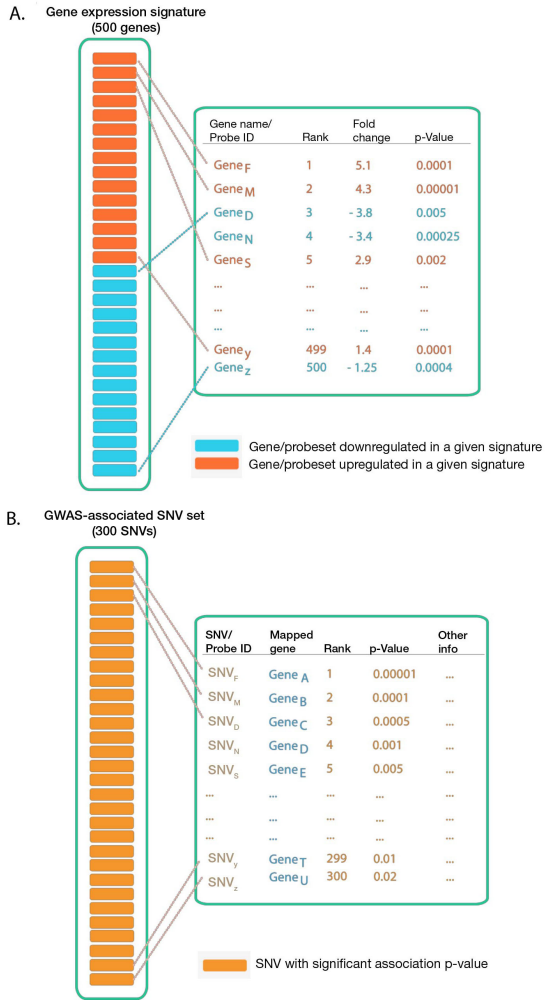


**Figure 2: Bioset ranking and directionality—** Examples of ranking in different bioset types. (A) Ranks based on fold change and directionality (fold change direction). (B) GWAS association bioset with a set of mapped genes, used in the enrichment analysis. Ranks for SNPs and mapped genes are derived from the corresponding p-value.

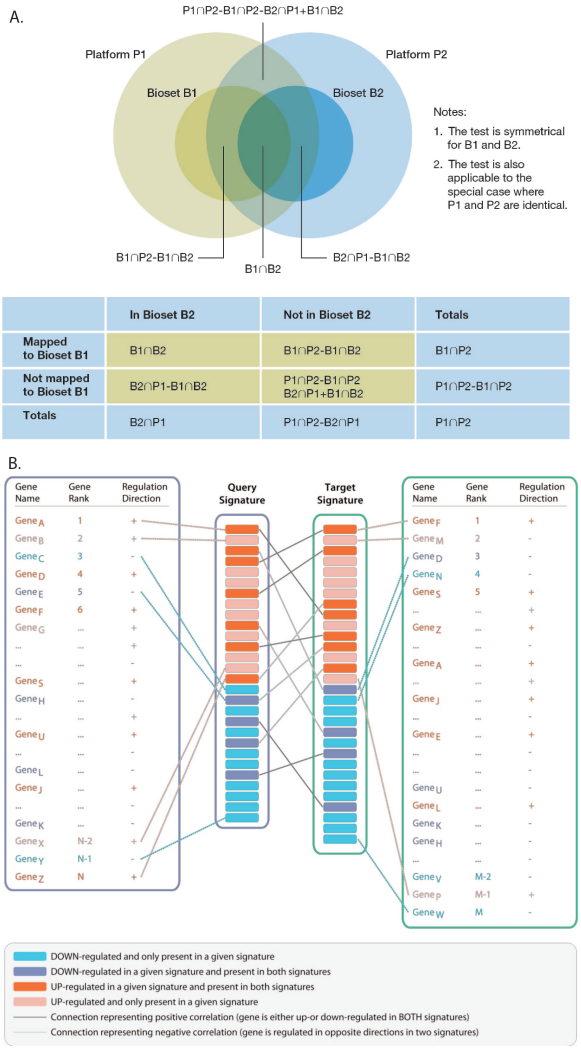


**Figure 3: Pairwise correlation between biosets —** Ebiosets—(A) Fisher's exact test parameters for bioset vs. bioset: assessment of statistical significance of enrichment of one bioset within another bioset. When the bioset is directional, the direction-specific subset (b1+ or b1- in place of B1; b2+ 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, where K is a dynamic number from 1 to N in the query bioset or 1 to M in the target. (B) An outline of pairwise directional enrichment analysis between two signatures.

The overall correlation score is the sum of directional subset scores, and the sign of the sum determines whether the two signatures are positively or negatively correlated (Figure 4). If one gene set is ranked and directional while the other is not, enrichment is computed for each directional subset (b+, b-) in the ranked, directional bioset. The overall enrichment score is the sum of the subset scores (Figure 4). In the application, scores are presented in top-level results normalized to a linear 0–100 scale with 100 representing the highest score. The magnitude of this overall enrichment score is also used to rank the query to every other bioset in the system (including user data in their private domain) and rendered from top to bottom in the Curated Studies app.[3,5]

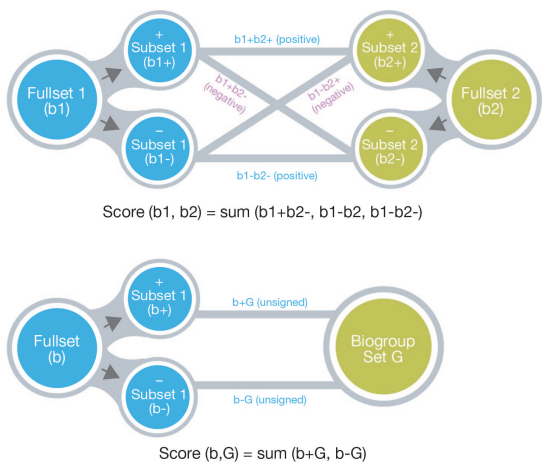


**Figure 4: Rank-based directional enrichment —** Enrichment scores are computed for the directional subsets, followed by a summation over all the subset scores. (A) Bioset vs. Bioset: the subsets are formed for both b1 and b2, and subset-subset enrichment scores are computed. (B) Bioset vs. unranked gene set: the subsets are formed for the bioset, and subset-biogroup enrichment scores are computed.

## Ontology-driven meta-analysis

Every bioset entered into the system is specifically 'tagged' with ontology-based biomedical terms for associating and ranking top data-driven diseases, compound treatments, and genetic perturbations. The bioset-bioset scores are used to compute scores against these ontology terms, which in turn are used to rank results in the Disease Atlas, Pharmaco Atlas, and Knockdown Atlas. For a gene or region query, the inverse of the normalized rank of the gene or gene region in the target biosets is used. When a gene or gene signature is queried, the first result is the rank ordered list of biosets scored from best down to a $10^{-6}$ cutoff. These biosets will then be categorized based on their tags and used to compute scores for each of the associated ontology terms (Figure 5). In the Disease Atlas, for example, this will produce a rank ordered list of disease categories based on a score computed from the tagged biosets.

A number of factors feed into the concept score (Figure 6). The Normalized bioset count is the sum of the ratio of scores to the best score (or in the case of a gene query, the sum of the ratio of inverse ranks to the best rank). This is divided by the Background count, which is the total number of biosets tagged with a given concept, both scored and below the scoring cutoff. This important factor balances against undue significance being given to over-represented concepts in the system. The concept score is further reduced by the Average weighted rank, which is a measure of how well the query signature scores against the concept data compared to all signatures in the system. Queries that perform better against the concept versus all signatures will be have an Average weighted rank closer to one and a better overall score.

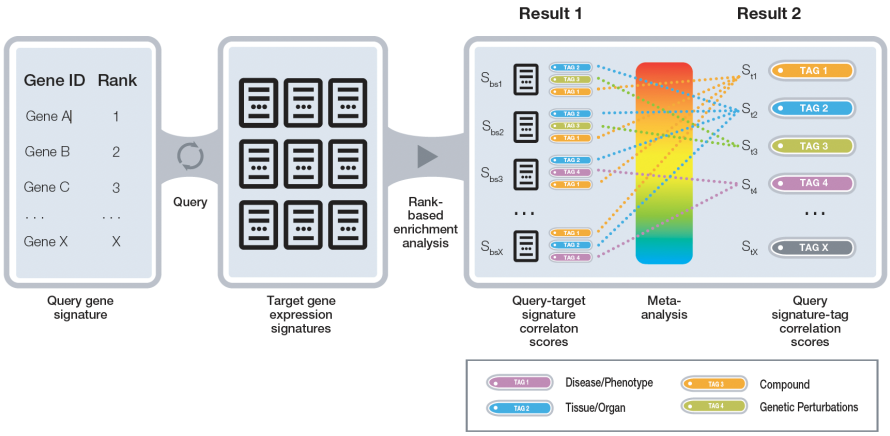


**Figure 5: Data correlation and meta-analysis —** Pairwise gene bioset correlation scores (computed using rank-based enrichment statistics) are computed, follow by meta-analysis of individual score-tag pairs to compute overall tag scores.

$$Score_{concept\ or\ gene} = \frac{Normalized\ bioset\ count}{Background\ count\ x\ Average\ weighted\ rank}$$

**Figure 6: Calculation for Concept score**

## Start today

The public data available in BaseSpace Correlation Engine is just the starting point for discovery. Users can upload their own data and query it against itself and the public data. Enterprise account holders can share their results within their private domain and add their results to the meta-analysis applications to generate unique correlations. Private data is inaccessible across domains and results are kept safe and private in an ISO27001, SOC1, SOC2, SOC3, PCI DSS certified environment.

## Learn more

To purchase, start a free trial, and learn more go to www.illumina.com/basespacecorrelationengine

For additional contact information, please see www.illumina.com/company/contact-us.html

Special academic pricing is available

## References

1. BaseSpace Correlation Engine Support page. support.illumina.com/sequencing/sequencing_software/basespace-correlation-engine.html. Accessed September 27, 2018.
2. Illumina (2014) Ranking of Genes, SNVs, and Sequence Regions. (support.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote-ranking-snvs.pdf).
3. Kupershmidt I, Su QJ, Grewal A, et al. Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS One*. 2010;5(9). pii: e13066. doi: 10.1371/journal.pone.0013066.
4. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.
5. Illumina (2014) Data Correlation Details: Enrichment Analysis. (support.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote-data-correlation-enrichment.pdf).

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



illumina®