illumına[®]

Metaanalysis Framework

An overview of techniques for analysis of global genomic data sets.

Introduction

This technical note describes the ontology-based metaanalysis framework developed by Illumina, which enables analysis of global high-throughput gene- and sequence-centric data sets. Queries enabled within this metaanalysis framework include the user data sets, as well as genes, single-nucleotide variants (SNVs), sequence regions, or **biogroups** (ie, pathways) of interest. This metaanalysis is enabled across data from different species, technologies, and platforms.

The framework uses a combination of rank-based enrichment algorithms, ontologies, and metaanalysis techniques to compute correlation scores between a gene, SNV, sequence region, biogroup, or a bioset with thousands of public studies and user data. An output produces a ranked set of studies and biomedical ontology concepts or **metaconcepts**, representing diseases, normal tissues, compound treatments, and genetic perturbations (gene mutations, knockouts, siRNA knockdowns) with strong association to a query of interest.

Data Pre-Processing Overview

The data-mining strategy can be divided into 3 parts.

 First, semiautomated crawlers collect public data from diverse sources, such as Gene Expression Omnibus (GEO)¹, Array Express², and other repositories (Figure 1). The data analysis step produces sets of differentially expressed genes associated with biological conditions investigated in each experiment, such as disease versus normal. Signatures (called **biosets**) are tagged with relevant ontology terms reflecting associated tissues, disease/ phenotype, compound treatment, or genetic perturbation (eg, gene mutation, knockout, siRNA knockdown).

- 2. Second, within each bioset, genomic entities (eg, SNVs, sequence regions, genes, and proteins) are ranked to indicate their relative significance to the tagged ontology term, compared to the other entities within the bioset. For example, in gene signatures, the ranking criterion used is the fold change. The technical note *Ranking of Genes, SNVs, and Sequence Regions* provides details of the ranking criteria for each type of bioset, as well as the "normalized" ranking function that enables comparison of ranks across various data types. For entity-based queries (eg, gene, SNV, sequence region, and protein), the"average" normalized rank of a genomic entity in biosets tagged with a given ontology term is an important factor in determining its correlation with that ontology term.
- Third, rank-based enrichment statistics are applied to compute pairwise correlation scores between all biosets and between biogroups and biosets. These scores are used by the metaanalysis engine to compute bioset and biogroup correlation scores with ontology terms or concepts (diseases, tissues, and treatments) (Figure 2).

To make sure that data across different platforms can be compared, a universal gene translation dictionary is applied to cross-reference







Pairwise gene bioset correlation scores (computed using rank-based enrichment statistics) are computed, followed by metaanalysis of individual score-tag pairs to compute overall tag scores.

gene identifiers from different sources. They are mapped to a standard National Center for Biotechnology Information (NCBI) Gene reference³. A cross-organism translation dictionary, based on precomputed ortholog clusters, is applied before computation to translate gene identifiers for biosets from different species.

Data Sources and Types

The BaseSpace[®] Correlation Engine contains data that were collected and preprocessed from several major public sources, including NCBI GEO, Array Express, Stanford Microarray Database (SMD)⁴, Broad Cancer Genomics⁵, National Cancer Informatics Program (NCIP)⁶, database of Genotypes and Phenotypes (dbGaP)⁷, and others.

Raw Data Pre-Processing

Figure 1 shows a typical data process pipeline for BaseSpace Engine bioset generation. All the raw data go through rounds of quality control, normalization, and statistical analysis that depend on the data type, and on the platform and experimental design used to generate these data. Several quality assessment methods are employed to review sample-level and data set-level integrity, and basic cutoffs are applied to filter out irrelevant features.

Different studies from different platforms exhibit variable expression profiles and statistics with different dynamic ranges—eg, distributions of fold changes and p-values—that reflect the technologies used. To allow interstudy comparability, a nonparametric approach has been established so that, for each bioset, ranks are assigned to each gene, SNV, or region based on selected metrics. Ranks are then further normalized to eliminate any bias due to various platform sizes.

Correlation Infrastructure

Cross-Platform Comparisons

To compare microarray data from different platforms, an index of microarray platforms is compiled. The index enables recognition of commonly used public gene identifiers and specific vendor identifiers. It provides a standardized mapping from individual identifiers to reference identifiers such as NCBI Entrez Gene⁸, UniGene⁹, Ensembl¹⁰, RefSeq³, or GenBank¹¹ accession numbers. The conversion of platform-specific identifiers into reference identifiers allows gene comparisons across different platforms.

Cross-Species Comparisons

To enable seamless comparison across different species, orthologs are identified for each pair of organisms and are grouped into ortholog clusters. Ortholog information is derived from Mouse Genome Informatics (MGI) at Jackson Labs¹², HomoloGene at NCBI¹³, and Ensembl¹⁰. Ortholog clusters are generated as follows:

- 1. The manually curated pairwise ortholog data among human, mouse, and rat from MGI are retrieved and clustered to form initial ortholog clusters.
- 2. The homology group data among human, mouse, rat, fly, and worm are analyzed to remove those in conflict with MGI data. The filtered homology group data are then entered into the ortholog clusters.
- The whole-genome pairwise sequence similarity data from Ensembl are processed to identify reciprocal best hits as candidate orthologs for all pairwise organisms among human, mouse, rat, fly, worm, and yeast. The candidate orthologs are

prioritized based on the percentage sequence identity and examined against the existing ortholog cluster. Qualified ortholog candidates are then entered into the ortholog cluster.

Ontology

The BaseSpace Knowledge Network contains an extensive vocabulary of tissues, diseases, compounds, and genomic entities derived from well-known ontology sources (Table 1).

SNV to Gene Mapping

The BaseSpace Knowledge Network contains precomputed associations between SNVs and genes. Associations are annotated based on the location, proximity, and function of the SNV relative to the genes.

Sequence-Centric Bioset Mapping

Sequence-centric biosets go through a mapping phase where regions are mapped to corresponding genes, resulting in a "gene-centric" version of the sequence-centric bioset. These biosets thus reside in the BaseSpace Knowledge Network in a sequence-centric as well as gene-centric form. The gene-centric form of a sequence-centric bioset comprises genes that are located within the region and extending up to 5,000 bp on either end. The genes in the gene-centric bioset inherit the ranks of the corresponding sequence region to which they map. Metaanalysis uses both these forms in different contexts.

Computing Pairwise Bioset Correlation Scores

Before metaanalysis, the BaseSpace Engine computes pair-wise correlation scores between all biosets by applying rank-based enrichment statistics. This algorithm is described in the Technical Note *Data Correlation Details: Enrichment Analysis.*

Metaanalysis: Computing Bioset-Ontology Correlations

Correlation of Genes, Sequence Regions, and Biosets with Ontology Terms

Given a single gene, a bioset representing the set of genes of interest, or a sequence region from a given experiment, a metaanalysis of the tens of thousands of tagged gene biosets in the BaseSpace Engine can be used to determine all strongly associated tissues, diseases, and compounds¹⁴. Conceptually, the problem is that of ranking ontology terms (concepts) based on how strongly enriched biosets tagged with those concepts are with a set of genes of interest. For a set of genes, ranked or unranked, the gene set enrichment analysis described here is used to identify other strongly associated biosets. Based on the strength of association, as determined using the pairwise bioset correlation score, the aggregated scores for their associated semantic concepts are computed.

Metaanalysis scores are computed separately for tissue, disease, and compound categories. Given a query bioset, a list of contributing biosets is obtained for each category based on 2 criteria:

- 1. They have enrichment scores with the query bioset above a predetermined threshold.
- 2. They pass an initial screening logic that makes sure that they are tagged with the appropriate combination of concepts to allow them to contribute to that category.

Based on the list of contributing biosets, along with the tagged concepts and enrichment scores, Equation 1 describes the various factors considered in determining a score for a concept.

$$Score_{Concept} = \frac{NormalizedBiosetCount^{a}}{BackgroundCount^{b} \times AverageWeightedRank^{c}}$$
(1)

The normalized bioset count for a concept is the sum of the associated score of each target bioset tagged with that concept divided by the overall best association score. For a gene or region query, the score represents the inverse of the normalized rank of the gene or region in the target bioset. For a bioset query, the score represents the pairwise bioset correlation score of the bioset with other target biosets. The background count of a concept is the number of biosets in the BaseSpace Engine tagged with that concept. It is an important factor that reduces the bias toward popular concepts that have more associated biosets than others. Finally, the average weighted rank represents the average rank of a tagged bioset relative to all other correlated biosets weighted by the associated normalized score.

This metaanalysis scoring process results in a ranked list of ontological terms for each tissue, disease, and compound category. Note that concepts may be part of a hierarchical ontological framework. When that is the case, enrichment scores for biosets tagged with specialized concepts are accordingly propagated to more general parent concepts in the hierarchy. The advantage of this strategy is that related but different ontology tags can still be associated with each other semantically. For example, **heart** and **left heart ventricle** can both contribute to the **heart** concept metaanalysis score, because **heart** is the parent concepts of **heart ventricle**. After scores for all concepts are computed, children concepts with lower scores than parent concepts are clustered and presented.

Correlation of SNVs with Ontology Terms

For each SNV, a correlation score is computed with a target concept (eg, disease). This score addresses the question: Given an SNV, what are the closely associated diseases, tissues, and compounds? First, all SNV biosets containing the SNV are retrieved, along with the associated p-values. An aggregation function ranks target concepts based on an aggregated p-value score that is computed using the p-values retrieved, taking into account the number of studies tagged with the concept (Equation 2). Concepts are ranked in **ascending** order of aggregated p-value.

$$AggregatedP_{concept} = \frac{GeometricMean(p-values)}{10^{StudyCount-1}}$$
(2)

Correlation of Ontology Terms with Genes and Biogroups

For each concept (eg, disease), a correlation score is computed with associated genes and biogroups. This score addresses the question: Given a disease, tissue or treatment, what are the highly relevant genes and biogroups?

First, all biosets tagged with the concept query are retrieved. For each gene in the biosets retrieved, an aggregated score is computed (Equation 3). This score takes into account the normalized rank of the gene in the bioset, the number of biosets containing the gene, the number of studies, and a background count. The average raw score is proportional to the sum of the inverse normalized score divided by the number of biosets. The background count is the number of concepts that are tagged to the set of biosets containing this gene. Only biosets that contain the gene with a normalized rank less than a predefined cutoff are allowed to contribute to the aggregating function. Concepts are finally ranked in **descending** order of score.

 $Score_{Gene_or_Biogroup} \quad \frac{AverageRawScore \times BiosetCount^a \times StudyCount^b}{BackgroundCount^c} (3)$

Biogroup scores are computed similarly. Only biosets that correlate with a biogroup having a correlation score greater than a predefined cutoff are allowed to contribute to the aggregating function.

Correlation of Ontology Terms with SNVs

This score addresses the question: Given a disease, tissue, or treatment, what are the closely associated SNVs? First, all SNV biosets tagged with the concept are retrieved, along with the associated p-values. An aggregation function ranks SNVs based on an aggregated p-value score that is computed using the p-values retrieved, and taking into account the number of studies containing the SNV (Equation 4). SNVs are ranked in **ascending** order of aggregated p-value.

$$AggregatedP_{SNP} = \frac{GeometricMean(p-values)}{10^{StudyCount-1}}$$
(4)

Dynamic Metaanalysis

The BaseSpace Engine is dynamic: the distribution of data from various species, platforms, data types, and semantic categories changes on a continuous basis. Thus, it is not obvious at the outset what the relative contributions of each of these factors should be in determining an optimal overall scoring function for top ranked concepts. These scores are determined empirically by optimizing the model described previously, using gold standard use cases and tuning parameters (eg, *a*, *b*, *c* in Equations 1 and 3) for each of the factors.

With every new bioset added, metaanalysis computations for existing biosets in the system are also updated. This ongoing process ensures that, at a given time, the most up-to-date results of the metaanalysis are produced, based on the current state of the BaseSpace Knowledge Network. When a query is performed, a total collection of metacategories, as well as individual biosets, are scanned to identify top-ranking normal tissues, diseases, and compounds.

Direction of Correlation

The overall direction of association (positive or negative) between a query bioset and a concept refers to the *net* correlation of the query bioset with a set of contributing biosets tagged with that concept. These contributing biosets may have positive or negative pairwise correlation scores with the query bioset. A *cumulative* positive and negatively associated biosets separately. To minimize the effect of any spuriously strong bioset association, the cumulative scores are each *weighted* by the ratio of the number of positively and negatively associated biosets. Finally, the overall direction is called depending on which

weighted cumulative score is greater, as long as a minimum difference threshold is met.

BaseSpace Correlation Engine Queries

Gene Query

The process for finding concepts that are highly correlated with a gene query is as follows. Sequence-centric biosets in the BaseSpace Knowledge Network also contribute to the correlation score in their gene-centric form.

- 1. Translate the gene query into its ortholog cluster identifier $I_{\rm g}$ (see cross-species comparison).
- Retrieve all biosets that contain I_g, along with the corresponding normalized rank. If I_g appears multiple times in a bioset, retrieve the smallest normalized rank.
 - Based on tags or tag combinations, biosets retrieved satisfy some basic criteria to qualify for contributing to a given disease, tissue, or treatment concept.
 - b. For gene-centric biosets originating from a sequence-centric platform, normalized ranks may be weighted, in the future, by the specific gene-region mapping.
- 3. Sort the biosets by normalized rank.
- 4. Compute the score for each concept with which the list of biosets is tagged, using Equation 1.

SNV Query

The process for finding correlated concepts is described in the section *Correlation of SNVs with Ontology Terms*, using Equation 2.

Region Query

The processing workflow for the region-concept correlation is similar to the one used for gene-concept correlation, with the main difference being query translation. Instead of obtaining correlation for only one ortholog cluster identifier, sets of: a) mapped ortholog clusters/genes; b) mapped regions; and c) mapped SNVs are obtained first as follows:

- 1. Translate the region into a set of mapped genes:
 - a. Retrieve the set of mapped ortholog clusters for the query region.
 - b. Retrieve the set of mapped regions for the query region.
 - c. Retrieve the set of mapped SNVs within a fixed window size for the query region. Retrieve the smallest normalized rank for each bioset containing *any* of the mapped ortholog cluster identifiers.
- 2. Retrieve the smallest normalized rank for each sequencecentric bioset (with qualifying basic category) containing regions *overlapping* with the query region. These sequence-centric biosets belong to the same species as the region queried.
- Retrieve the smallest normalized rank for each SNV data set (with qualifying basic category) containing *any* of mapped SNVs.
 Note: Biosets retrieved satisfy some basic criteria to qualify for contributing to a given disease, tissue, or treatment concept.
- 4. Sort all biosets retrieved in steps 2–4 by normalized rank.

5. Compute the score for each concept for which the list of biosets is tagged, using Equation 1.

Bioset Query (Gene- and Sequence-Centric)

Biosets comprise gene sets or sequence-region sets, which are internally mapped into gene sets. When a bioset is queried to obtain correlations with other concepts, the processing workflow to identify concept correlations is as follows:

- Retrieve the summary scores for each bioset having a pairwise bioset correlation score ≥ minimum threshold against the query bioset.
- 2. Sort the biosets by absolute value of summary score.
- 3. Compute the score for each concept for which the list of biosets is tagged, using Equation 1.

Biogroup Query

When a biogroup (eg, pathway) that comprises a set of genes is queried, the processing workflow to identify concept correlations is as follows:

- Retrieve the summary scores for each bioset having a pairwise bioset-biogroup correlation score ≥ minimum threshold against the query bioset.
- 2. Sort the biosets by summary score.
- 3. Compute the score for each concept for which the list of biosets is tagged, using Equation 1.

Concept (Disease/Tissue/Treatment) Query

When users query a disease, tissue, or treatment, the associated genes and biogroups are ranked and presented. The basic form of the aggregation function that assigns an association score to each gene or biogroup is presented in Equation 3. The target bioset count includes both gene-centric biosets and sequence-centric biosets in their genecentric form. Correlated SNVs are ranked based on the aggregated p-value score computed using Equation 4.

References

- 1. www.ncbi.nlm.nih.gov/geo/
- 2. www.ebi.ac.uk/arrayexpress/
- 3. www.ncbi.nlm.nih.gov/refseq/
- 4. smd.princeton.edu/
- Park YR, Lee HW, Kim JH (2005) Integrating microarray gene expression object model and clinical document architecture for cancer genomics research. AMIA Annu Symp Proc: 1073.
- 6. cbiit.nci.nih.gov/ncip
- 7. www.ncbi.nlm.nih.gov/gap
- 8. www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Entrez/
- 9. www.ncbi.nlm.nih.gov/unigene/
- 10. www.ensembl.org
- 11. www.ncbi.nlm.nih.gov/genbank/
- 12. www.informatics.jax.org
- 13. www.ncbi.nlm.nih.gov
- 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.

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/or its affiliate(s) in the U.S. and/or other countries. Pub. No. 970-2014-008 Current as of 28 March 2016

