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.
1. First, semiautomated crawlers collect public data from diverse sources, such as Gene Expression Omnibus (GEO), Array Express, 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.
3. 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 biost 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...
gene identifiers from different sources. They are mapped to a standard National Center for Biotechnology Information (NCBI) Gene reference. 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), Broad Cancer Genomics, National Cancer Informatics Program (NCIP), database of Genotypes and Phenotypes (dbGaP), 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—e.g., 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 Gene, UniGene, Ensembl, RefSeq, or GenBank 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 Labs, HomoloGene at NCBI, and Ensembl. 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.
3. 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
Technical Note: Informatics

Geometric Mean p values

Conceptually, the problem is that of defining a score for each concept based on an aggregated p-value score that is computed using the geometric mean of the p-values retrieved, taking into account the number of studies tagged with that concept.

The normalized bioset count for a concept is the sum of the number of biosets in the BaseSpace Engine 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.

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.

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 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.

\[
\text{AggregatedP}_{\text{Concept}} = \frac{\text{GeometricMean(p-values)}}{10^{\text{ScoreNorm}}} = \frac{\prod_{i=1}^{n} p_i^{1/n}}{10^{\sum_{i=1}^{n} \text{ScoreNorm}_i}}
\]

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

\[
\text{Score}_{\text{Gene}} = \frac{\text{Score}_{\text{Bioset}} \times \text{NormalizedBiosetCount}^* \times \text{BackgroundCount}^*}{\text{AverageWeightedRank}^*}
\]
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.

\[ \text{Score}_{\text{Gene or Biogroup}} = \frac{\text{AverageRawScore} \times \text{BiosetCount} \times \text{StudyCount}}{\text{BackgroundCount}} \]  

(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.

\[ \text{AggregatedP}_{\text{SNP}} = \text{GeometricMean}(p \text{-values}) \times \frac{\text{BackgroundCount}}{10^{\text{StudyCount}-1}} \]  

(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 (e.g., 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 negative score is obtained by aggregating the scores for the positively 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, respectively, to the total number of contributing 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 \( I_g \) (see cross-species comparison).
2. 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.
   a. 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. Based on tags or tag combinations, biosets retrieved satisfy some basic criteria to qualify for contributing to a given disease, tissue, or treatment concept.
2. Retrieve the set of mapped ortholog clusters for the query region.
3. Retrieve the set of mapped regions for the query region.
4. 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.
5. Compute the score for each concept with which the list of biosets is tagged, using Equation 1.

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:

1. 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 (e.g., pathway) that comprises a set of genes is queried, the processing workflow to identify concept correlations is as follows:

1. 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 gene-centric form. Correlated SNVs are ranked based on the aggregated p-value score computed using Equation 4.

**References**

2. www.ebi.ac.uk/arrayexpress/
4. smd.princeton.edu/
6. cbit.nci.nih.gov/ncip
10. www.ensembl.org
12. www.informatics.jax.org