Chapter 6

Gene Set Enrichment Analysis

Charles A. Tilford and Nathan O. Siemers

Abstract

Set enrichment analytical methods have become commonplace tools applied to the analysis and interpretation of biological data. The statistical techniques are used to identify categorical biases within lists of genes, proteins, or metabolites. The goal is to discover the shared functions or properties of the biological items represented within the lists. Application of these methods can provide great biological insight, including the discovery of participation in the same biological activity or pathway, shared interacting genes or regulators, common cellular compartmentalization, or association with disease. The methods require ordered or unordered lists of biological items as input, understanding of the reference set from which the lists were selected, categorical classifiers describing the items, and a statistical algorithm to assess bias of each classifier. Due to the complexity of most algorithms and the number of calculations performed, computer software is almost always used for execution of the algorithm, as well as for presentation of the results.

This chapter will provide an overview of the statistical methods used to perform an enrichment analysis. Guidelines for assembly of the requisite information will be presented, with a focus on careful definition of the sets used by the statistical algorithms. The need for multiple test correction when working with large libraries of classifiers is emphasized, and we outline several options for performing the corrections. Finally, interpreting the results of such analysis will be discussed along with examples of recent research utilizing the techniques.

Key words: Gene set enrichment analysis, statistics, software.

1. Introduction

Enrichment analysis is a statistical approach used to discover unusual representation of a categorical class within a selection of items from a heterogeneous population. It can be applied in any situation where bias is suspected in the choice of a subset of members from a larger discrete list. In the biological sciences, it has gained

Yuri Nikolsky, Julie Bryant (eds.), *Protein Networks and Pathway Analysis, vol. 563*

DOI 10.1007/978-1-60761-175-2_6 Springerprotocols.com

popularity in the analysis of experimentally defined gene lists, generally for the purposes of describing the list or discovering common characteristics that generally describe the members in aggregate. As an example, analysis of a gel band might reveal eight distinct proteins, four of which are known to be localized within the mitochondrial lumen. Is such an observation likely to occur by chance, or does the band contain an unusually high (or low) number of lumen genes?

Two major challenges face the researcher who wishes to perform an enrichment analysis. While a single enrichment calculation is fairly trivial, a typical analysis will require calculations on thousands of sets against thousands of candidate classifiers, and generate a very large result set. Fortunately there are many software packages available to researchers to manage both the calculations and presentation of results. Some of these packages are commercial, some are available for use online, and others are fully open sourced and can be compiled locally on the researcher's own computer.

The second challenge is more daunting; however, like any statistical approach, enrichment analysis can suffer if the starting assumptions, application of analysis, or interpretation of results are not carefully considered. This primary value of this chapter may be in raising awareness of these issues, and providing options to avoid or mitigate them. The major concerns can be summarized as follows:

1. Careful definition of the reference set. It is important to rigorously define the "world" from which the selection list was chosen. This can have a dramatic impact when calculating classifiers. If it is impossible for a gene to ever be selected (for example, because the detection technology is focused only on specific genes), then the gene should not be present in the reference set. Similarly, the initial analytic method might report multiple gene identifiers (splice variants, redundant probes) that should be considered the same gene, particularly if the categories being tested were created at the gene level. The reference set and selected list should be normalized such that each gene is present only once.
2. Accommodating non-independent association of set members. The null hypothesis for most of the statistical algorithms is that each gene was added to the list randomly and independently of all other genes. However, it is well established that genes, proteins, and other biological components often display dependent and correlated behavior. In addition, our existing classification schemes are based on and biased by biology that is well understood, which introduces further challenges to the independence assumption. Failure to account for dependencies between genes can result in the significance of a classifier being over or under estimated.

3. Correcting for multiple testing. Enrichment analysis is commonly used as a data mining tool, meaning that tens of thousands of potential classifiers can be surveyed across large numbers of lists. The need to adjust expectation thresholds in such circumstances is generally accepted, but can be difficult to implement. In particular, lack of independence among the classifiers (for example, in the hierarchical organization of many ontologies) complicates estimation of the correction factor to use.

1.1. Definition of Concepts Used in This Chapter

Set or **List** – a set in the strict mathematical sense; a discrete, non-redundant collection of objects.

Member – an individual object within a set.

Ranked List or **Ranked Set** – an ordered set; the order in which members have been added to the set is informative.

Selected Set or **Selection** – a set of objects that is of interest to the researcher, presumably identified through one or more experimental procedures. The selection could be ranked or unranked.

Reference Set – the "world" from which the selection was chosen. The reference set should contain all objects that have a non-zero chance of being selected, and exclude any object with a zero chance of selection (so the selection will be a subset of the reference set). In this chapter the reference set will always be unranked.

Accession – an identifier representing an entity within a database. Accessions can generally be recognized as identifying a single, specific entity even in the absence of any other context, for example NP_001225 or GO:0051925. Gene symbols, such as CAV3, are generally not considered accessions, since they are usually ambiguous as to the species they describe and poorly standardized. While it is certainly possible to work with sets composed of objects other than accessions, they are strongly recommended for any analysis you perform.

Namespace – a set of identifiers, usually accessions, from a single source. For example, RefSeq, Swiss-Prot, and LocusLink Gene are distinct namespaces, each of which has one or more accessions representing CAV3 (NP_001225/NP_203123, P56539/Q3T1A4, LOC859).

Gene – for convenience, "gene" in the context of a "gene set" will be used here to refer to any biological entity that a researcher may conceivably investigate; this includes RNAs, proteins, probes as well as classical gene loci. Any of these identifiers are equally mathematically suited to enrichment analysis, so long as consistent namespaces are used. However, we will see that

choosing more finely grained identifiers such as proteins or probes can result in problems of artificial over-representation within a gene set.

Classifier or **Class** – a categorization or description that could be applied to the genes in the reference set. Each classifier is itself an unranked set, and when utilized should be a subset of the reference set (that is, any members that are not part of the reference set should be discarded). We will consider only binary classifiers; a gene is either a full member of a classifier or not part of the classifier at all. Any conceivable subset of the reference set could be used as a classifier.

Query Class – a single specific classifier being considered at the moment.

Ontology – Ontologies are structured collections of classifiers. They are generally curated and frequently organized as a directed hierarchy, with parent classifiers being more generic than their child classifiers (which are always subsets of their parents). The most commonly used ontology is the Gene Ontology (GO), consisting of three distinct descriptive hierarchies.

For the analyses described here, the concepts above fully capture the descriptive information we will use for our calculations. The primary experimental data is simply the gene membership in the selected set and possibly rank order within the selection. Supporting data (presumably from an external source unrelated to the experiment) will be gene membership in the reference and gene membership in one or more classifiers. The analyses will ignore any other quantitative or qualitative data captured during the experiment. The reference and class sets are solely represented by their membership, any higher-order relationships or qualifiers between or within the sets are discarded (although such information may be used to generate additional classifier sets).

1.2. Standard Mathematical Variables

N – the size (total count of set members) of your selection

i – the number of selection members that belong to the query class

n – the number of reference set members that belong to the query class

m – the number of reference set members that do *not* belong to the class. Since the classes are binary descriptors and every set member is either a member of the class or not, $n + m$ is the total size of the reference set.

T – the total number of statistical tests performed, relevant when correcting for multiple testing.

2. Preparation of Analysis

2.1. Definition of the Selected Set

Any mechanism that produces a list of genes is suitable for generating the selected gene set. The only mathematical restriction on the selected set is that it be a subset of the reference set. Generally the set will be derived by an experiment measuring the absolute or relative quantity of a gene-associated product (typically RNA or protein). If absolute gene levels are being measured, subsets can be derived by application of a threshold level, or all or part of the reference set can be selected as a ranked list. Relative gene levels are captured when one or more experimental results are compared to one or more controls. Lists can again be derived by thresholding or as ranked lists.

As with any experiment, the utility and reliability of the gene sets in subsequent analyses will be dependent on the care taken in experimental design, technical quality of the experiment, the mechanisms used to determine the quantity of gene present in any particular experiment, and the procedures used to convert raw gene levels to the selected or ranked lists. The greater the experimental or technical error in any of these components, the more likely the interpretation of the enrichment results will be confounded, or the capacity to identify truly differentially enriched populations reduced.

Expression studies are one of the more common sources for defining selection sets, where the selection is generally derived from a comparison of experimental and control samples, choice being made via the presence of a gene in one or the other, or by a measure of difference between the two. In many cases the selected set can be ranked based on the magnitude of a measured difference.

An increasing number of gene sets are becoming available through publication by researchers throughout the world. While these sets are mathematically suitable for direct enrichment analysis, they are often provided without even basic experimental descriptors, let alone the raw data used to generate the gene set. Lacking an understanding of the experimental protocol will generally make meaningful interpretation impossible. In addition, a gene set without the raw experimental data used to derive it denies the use of label-based permutation, which will be discussed later in this chapter.

2.2. Ranked Versus Unranked Selections

The selected set can be unranked, in which case a single enrichment calculation will be performed for each classifier. If an unranked set is being used, selecting the entirety of the reference set ($N = m + n$) is always insignificant (if you buy all available lottery tickets, it should not come as a surprise that you win all the prizes).

Analyzing ranked sets can provide useful analysis even if the entire reference set is used, however. Ranked set analysis will effectively consider all possible sublists starting with the set of just the top ranked member, and then expanded by iterative addition of each subsequent member from the ranked list. This can be accomplished by generating a running score that is modified with addition of each new set member, or by treating each subset as a discrete unranked set. The results generated at each point of this process can be reported (frequently as a plot reflecting the running sum or a statistical measure at each point), or the single "best" result encountered can be reported for the entire ranked list.

2.3. Definition of the Reference Set

Properly defining the reference set, which represents the world from which the selected set is drawn, is critical. In very simple terms, the reference set is the "denominator" of the analysis – failing to define it properly can dramatically alter the presumptive odds of observing your selection by chance. Consider the example mentioned earlier – analysis of a gel spot reveals eight proteins, half of which are annotated as being from the mitochondrial lumen. If the eight proteins could have come from any of 36,000 loci in the human genome (100 of which are localized to the mitochondrial lumen; $N = 8$, $i = 4$, $m + n = 36{,}000$, $n = 100$), using Fisher's exact test would predict that our observation has a p-value of roughly 3.9×10^{-9} given a null hypothesis of totally random selection from the genome. However, if experimental constraints were such that the gel spot could at most have had only 30 genes, 6 of which were localized to the lumen ($N = 8$, $i = 4$, $m + n = 30$, $n = 6$), our p-value degrades to a mere 0.03.

In general, if there is no chance that a gene could be drawn into your selection, it should be removed from the reference set. A common scenario would be an experimental protocol that involves a sensor designed to detect specific, discrete genes (e.g., an RNA hybridization array). The reference set should only contain those genes that are represented on the sensor.

Additionally, you may wish to remove set members that have known technical biases. For example, perhaps you have purified a whole cell protein preparation using a column chemistry that you know will greatly deplete an entire class of proteins, let us say those embedded in a plasma membrane. You may find it useful to remove those entries from both the reference and selected sets, even if they still have the potential to be detected. Assuming the column worked properly, your selected gene lists should be depleted for membrane proteins, and may then also be enriched for at least some classifiers that are *not* associated with membrane proteins. If the number of genes removed by the column is large enough, analytic methods that assume random selection as the null hypothesis will be able to detect the perturbation and report it. While this is not a "wrong" answer (your sample is in fact perturbed) it is presumably not useful

either, and may cause unrelated classes to be more significantly enriched. By excluding membrane proteins from your gene lists altogether you can focus analysis on discovery of novel enriched classifiers.

Any time a group of genes has almost no chance of being selected (or is always at the very bottom of a ranked list) you should consider if it should be removed from your gene lists (both reference and selections). In addition to technical aspects of your experiment, there may be recognized biological constraints, such as the tissue or developmental stage of the sample being used. If permutation techniques are available to provide a control on significance calculations, then consistently unselected genes will not be a problem, but as we will see permutation is not always available.

2.4. Classifier Sets

The only mathematical restriction on classifier sets is that they be subsets of the reference set. Otherwise, any set of genes can be used as a classifier. Classifiers can be constructed computationally (for example through sequence similarity searches) but often have some level of expert human curation applied to them.

- **Functional ontologies** – Describe the behavior of the gene in broad terms. Gene Ontology (1) is probably the best known and most popular, with a massive curated hierarchy of classifiers. OMIM (2) classifies genes by the diseases they are suspected of causing.
- **Domain or motif classifiers** – Group genes by shared sequence features. InterPro (3) is an excellent example organized as a hierarchy, and integrates roughly a dozen motif and structural databases.
- **Networks** – Allow genes to be grouped by their connection within a known pathway. Distilling a pathway into discrete sets of genes generally requires relatively sophisticated rules; How far away can two genes be in order to be in the same set? How strong or well-documented do we require connections to be? Do we care about the direction of the connections, or if they are effectors or suppressors? Careful mining of networks can produce very useful gene sets, however. For example, building sets of genes that are directly or indirectly downstream of regulatory genes can allow intuition of a small set of effector genes from their downstream effects.
- **Experimentally derived gene lists** – Gene sets from other experiments can themselves serve as classifier sets. In such a scenario using a previously generated selection as a classifier against the current selection is effectively asking "does my selection show an unusual enrichment or depletion of genes specified by the prior selection." Such tests can be used as a measure of similarity between two selections, and could highlight similar biology behind the generation of both.

- **Generic co-occurrence** – Guilt by association, assigning genes to a common classifier if they are encountered in the same context. For example, **PubMed** (4) will note genes that are explicitly mentioned in a particular publication. A journal article can then serve as a gene set, and may be able to describe a commonality not captured by a single domain or ontology term.

If a hierarchy of classifiers is being used, it is possible to test either all classes within the hierarchy, or only those that are leaf nodes (the most specific classifiers, with children of their own). While testing leaves alone provides some benefit in reducing the multiple testing correction needed, it prevents detection of enrichment within more general categories of the hierarchy. If exploratory analysis is being performed, it is generally advised for all classes of a hierarchy to be tested.

The quality of the classifier will affect its utility in an analysis. If a classifier is improperly assigned to many irrelevant genes, then the true signal defined by the classifier will be diluted, and may fail to be observed. Some classifier sets include quality assessments that can be used to filter out unreliable assignments. For example, Gene Ontology assignments include an evidence code that indicates how the assignment was made (5); TAS (Traceable Author Statement) entries are supported by published literature, whereas IEA (Inferred from Electronic Annotation) are machine-assigned, generally by sequence similarity. One could use only TAS assignments or discard all those coded with IEA.

However, if a classifier has been assigned to only a fraction of the genes it should represent, there may be too few assignments to generate an enrichment of adequate significance to pass your threshold. In most cases, only a small fraction of true assignments are supported in the literature, so relying on TAS generally results in a very small number of ontology assignments. Similarly, while annotation through sequence similarity can introduce a sizable number of erroneous assignments, it usually has a good success rate and can greatly expand the number of annotations available.

2.5. Specific Enrichment Tests

All enrichment tests will attempt to identify classifiers that appear to be unusually enriched or depleted in the selected set compared to the reference set, and attempt to quantify or qualify the magnitude of the effect.

2.5.1. Sample Enrichment

The simplest quantity to calculate is the relative enrichment of your sample:

$$\text{Enrichment} = \frac{i/N}{n/(n+m)} = \frac{i(m+n)}{Nn}$$

An enrichment of exactly 1 indicates that the selection has the same proportion of class members as the reference set. Enrichment values greater than 1 indicate that the selection is enriched for the

classifier, while values less than 1 indicate depletion. While helpful, the enrichment alone does not provide any indication of how unusual such an enrichment may be. Very large enrichment values are generally not significant when dealing with very small values of N or n, and similarly a modest enrichment can be very significant when N or n is large.

2.5.2. Odds of Random, Independent Selection

If we assume a null hypothesis that every gene was selected at random and independently of all others we can generate a statistical measure of our observation for each class member using a 2×2 contingency table:

Class member	Selected set *i*	Not selected *n – i*	Class totals *n*
Not a class member	$N - i$	$m + i - N$	m
Set totals	N	$m + n - N$	$m + n$

The probability of such an observation is reported by Fischer's exact test (6), which in this case is the hypergeometric distribution (7):

$$\mathrm{p} = \frac{\binom{n}{i}\binom{m}{N-i}}{\binom{m+n}{N}} = \frac{m!n!N!(m+n-N)}{i!(n-i)!(m+i-N)!(N-i)!(m+n)!}$$

We can calculate a one-tailed *p*-value by adding the odds of all events more extreme than this to our first calculation. If the enrichment is greater than 1, this will be all events with a larger i, while enrichment values less than 1 should consider all events of smaller i. When enrichment equals 1, a consistent choice of one of these options should be made; most systems will likely take all larger i values. A two-tailed *p*-value can also be calculated; this would add all events that are less likely than our observation, including those with an alternative enrichment.

Most modern computers can perform even a fairly large number of such calculations without problems. When using large ontology sets, however, or when presented with many experiments, "fairly large" can quickly become "extremely large". In such cases it may be advantageous to use a simpler distribution to estimate the *p*-value. By ignoring the total size of the reference set $(m + n)$ and considering only the average occurrence of a classifier $(n/(m + n))$, we can calculate approximate *p*-values using either the binomial discrete distribution (8) or chi-squared continuous distribution (9). These tests are appealing as they are easy to implement, trivially accommodate changes in the size of the

selection or reference sets, and require no other information other than the basic composition of the sets being used (no scaling factors, for example).

The serious shortcoming with all the above approaches is the assumption that set members are independent; this is clearly incorrect for almost all biological gene sets that a researcher might consider. It is very important to bear this in mind when interpreting any results returned from such an analysis; the reported likelihood is fundamentally a measure of how independent the members of your selection are. Dependency can be generated by mechanisms that would generally not be considered "interesting" or informative. Some common examples are as follows:

- "Multiple voting" in the experimental protocol. Many gene identification experiments contain redundancies for some or all the genes being tested. If your set namespace is based on the probes, you would expect multiple probes from the same gene to be tightly dependent on each other. It is always desirable to collapse redundant identifiers down to discrete loci. This is true even if the redundancy is identical for all probes. For example, consider a reference set of 1,000 probes, 100 of which are assigned to a classifier ($m + n = 1{,}000$, $n = 100$). If we select 100 probes, 20 of which match the classifier ($N = 100$, $i = 20$), the hypergeometric distribution reports a p-value of 0.001. If we then redesign our experiment to precisely duplicate every single probe ($m + n = 2{,}000$, $n = 200$) and find the comparable "scaled" result in a repeated experiment ($N = 200$, $i = 40$), we now get a p-value of 4.8×10^{-6}. By allowing redundancy, we have greatly inflated the significance of our observation.
- "Multiple voting" in a biological context. For example, there are 23 human tubulin genes, many of which are tightly co-expressed. All 23 are associated with the classifier GO:0007018 (microtubule-based movement), which describes a total of 124 loci in the human genome. If we select 100 loci from a reference of 36,000, and find that 8 are assigned to GO:0007018, the hypergeometric distribution would assign a p-value of 2.3×10^{-9} to this observation ($i = 8$, $N = 100$, $n = 124$, $n + m = 36{,}000$). If all eight loci were tubulins, however, we may be concerned that this observation is merely highlighting the importance of tublins in "microtuble-based movement", a revelation that is hardly earth-shattering. If we collapse all 23 tubulins into a single "placeholder", and repeat the calculation ($i = 1$, $N = 93$, $n = 102$, $n + m = 35{,}978$), we now find that our p-value has degraded to 0.23, suggesting that while the set is tubulin-heavy, it is not significantly associated with microtubule-based movement specifically.

The assumption of random selection is obviously wrong, too. In any given experiment, we would not presume that all genes have an equal probability of being selected. In particular, there are some genes that will always have a very low, or even zero, probability of being selected in an experiment. As for dependency, we can find uninformative enrichment arising from both experimental and biological sources:

- Experimental bias. For whole-genome assays there are generally genes that cannot be detected, perhaps because they are depleted or removed during sample preparation. Even for discrete detection methods such as microarrays, design errors can result in probes that fail to detect the gene they were based on. As mentioned in the prior discussion on defining the reference set, such genes should be removed entirely from the analysis.
- Biological bias. Experimental samples are generally from a fairly well-defined biological source, such as "adult male liver" or a specific cell line, and often under controlled conditions. It is possible that entire classes of genes would be absent from your analysis; for example, genes active only during embryonic development would presumably not be found in adult somatic tissue.

The affect of such biases depends on the number of genes in question, as well as the distribution of the classifier within the affected genes. If the classifier occurs in the biased genes with the same frequency as the whole of the reference set, the bias is unlikely to have a notable impact on the calculations. If the biased genes also contain biases in classifier annotation, however, they can skew the enrichment calculations for selected genes. As an example, many microarray designs contain probes for "putative" genes which are determined to be spurious after fabrication of the microarrays, or probe. In the case of DNA microarrays, it is also possible for genes with only minimal EST support to have probes accidentally produced against the wrong strand, resulting in a "dead" probe even when the gene is ultimately validated. Putative genes are often very poorly annotated and as a group are under-enriched in most classifiers. The presence of a large number of "dead" probes (which are either not selected into unranked lists or at the very bottom of ranked lists) effectively leads to minor (or occasionally major) enrichment of all classes in the selected genes.

For a concrete example, assume a microarray that has probes for 13,000 genes, 4,500 of which are classed as cytoplasmic; 1,100 classed as membrane integral; and only 200 classed as both. We perform an experiment that results in the loss of any membrane-associated protein during preparation of all samples, and then generate a selection set of 5,000 genes, 1,800 of which are classed as cytoplasmic. Using the entire reference set, the hypergeometric

distribution reports a *p*-value of 0.005 ($i = 1{,}800$, $N = 5{,}000$, $n = 4{,}500$, $n + m = 13{,}000$), suggesting that cytoplasmic genes are unlikely to have been randomly and independently added to our selection. If we instead first exclude all membrane integral genes from both sets ($i = 1{,}800$, $N = 5{,}000$, $n = 4{,}300$, $m + n = 11{,}900$), the *p*-value now degrades to 0.4, indicating that the enrichment in cytoplasmic genes was almost certainly due to the depletion of membrane proteins, which are depleted for the cytoplasm classifier.

It is important to emphasize that these effects do not produce "wrong" statistics, but rather could result in confounding the analysis or misinterpreting the results. As for any statistical technique, the interpretation of the result must be made in the context of the null hypothesis – when these tests are performed, significant results indicate classes of genes that are likely to be dependent on one another and/or non-randomly selected. In the prior example, analysis with an uncorrected reference set tested not only for differences between treatment and control, but also identified genes that were affected by the sample preparation. Cytoplasmic genes were not chosen at random since they were preferentially passed by the column, a bias that was appropriately detected by the analysis as failing the null hypothesis.

Being too zealous in the removal of "unobtainable" genes can cause truly interesting biology to be overlooked, however; the point of an enrichment analysis is generally to identify unexpected features of your selection set. For example, purging all early development genes from the reference set used to analyze adult tumors would limit the opportunity to detect pathological activation of an embryonic pathway. Technically impossible genes (such as those referenced by improperly designed probes) should always be removed. In almost all cases it will be prudent to retain biologically improbable genes within the reference set, however.

Aside from the limitations of the null hypothesis, the appropriateness of the application in general has itself been questioned. Goeman and Bühlmann point out that the technique above treats the *genes* as samples, while classically the samples have been individual organisms, cell lines, experimental treatments, etc. (10). New gene sets do not really represent new samples, but rather a re-ordering of the same fixed reference set.

2.5.3. Odds Via Permutation

An alternative approach to the methods described above is permutation (11). Permutation relies on randomization of your experimental data to provide an empirical distribution of a scoring metric in your real data. In general, the process is as follows:

- A scoring metric is chosen for testing a classifier against a gene set.
- The selected set is generated from the experimental data.

- The experimental data is randomized (permuted) and another selected set is generated with the same methodology used to make the "true" selected set. This process is repeated as many times as possible to generate a series of gene sets derived from the permuted data.
- For each classifier of interest, the scoring metric is applied to both the true selected set and every permuted set. The number of permuted sets with a score equal to or better than that of the true set are tallied and reported as a fraction of the total number of permutations. This fraction represents an empirical odds of observing such a score within the context of your experiment.

The most powerful form of permutation shuffles the "treatment" and "control" labels (the "class labels", not to be confused with the classes in the classifier sets) assigned to raw experimental samples. That is, if your experiment is composed of 5 treatment and 5 control samples, the permuted sets are generated by randomizing the assignment of "treatment" and "control" to the samples; if we maintain 5 samples in each group, this allows for 126 discrete permutations (half of "10 choose 5", one of which is the true experimental set). This level of randomization preserves dependencies between genes, as well as general biases in the likelihood of selecting any given gene in the first place. As such, class label permutation corrects for the limitations of the "random, independent" null hypothesis described above. The null hypothesis for class label permutation would be "gene sets generated by comparing the treatment and control data are no different than those generated by arbitrary comparisons between data sets, based on the scoring metric." Note that this hypothesis is fundamentally addressing experimental samples, not the gene sets themselves.

Permutation of the class labels is limited by availability of the raw data, however. In some cases, a researcher will not have access to the original data sets, making permutation impossible. In other cases, the number of control and treatment will be so small as to provide only a limited number of permutations. If treatment and control experiments were each performed in triplicate, this allows for only 10 permutations (31 if we allow the number of samples in each class to vary). Ten permutations allows at best for a likelihood of 0.1 (1 out of 10), which fails to meet most confidence thresholds.

Permutation can still be employed even if the raw data is unavailable or too sparse, however. For example, if you have access to raw data for unrelated experiments, random samples chosen from the unrelated experiments can be used to generate permuted gene sets. Such sets will obviously fail to capture dependencies or biases that are specific to each experiment, but they can still account for commonalities between all experiments (such as common experimental procedures or fundamental biological relationships). The

strength of this approach is that it can leverage a very large pool of experimental data, allowing a massive number of permutations to be generated. A lab that has performed 100 experiments would be able to generate 75 million distinct, arbitrary groups of 5 experiments, resulting in roughly 10^{15} available faux "treatment/control" pairs.

The scoring metric can be as simple as counting the number of set members within a gene set. It is important that the metric be universally comparable between the true set and all permutations, however. If a simple count was being used, it would require the size of all selected gene sets to be constant. If the process for generating the gene set results in lists of varying sizes (e.g., a ranked list cutoff based on fold change), metrics resilient to set size should be used; *p*-values from the chi-squared and hypergeometric distributions would be a reasonable choice.

2.5.4. Odds Via Simulation

If no raw data are available, attempts to capture dependencies and biases within an experiment can be made through simulation. In a fashion similar to permutation, simulated gene sets are produced and used to generate an explicit distribution, against which the experimental results are compared. Note that the simulation consisting of the random and independent ordering of genes is effectively a Monte Carlo approach to re-generating the hypergeometric distribution (which should be used directly instead). Simulations can attempt to capture biases in selection likelihood, as well as dependencies between pairs or small groups of genes. Generally these relationships would be codified in a series of rules or algorithms, and then parameters for each gene or group of genes would be determined by analyzing a collection of previously generated experimental gene sets.

Simulations have the same limitations as permutations over pooled, unrelated experiments. In addition, while they can generally do a good job at defining general selection biases for single genes, they will be incapable of defining dependencies between gene groups larger than two or three, and will likely do a mediocre job for even pair-wise dependencies. Finally, attempting to define the parameters defining the simulation can result in errors being built into the model itself. For these reasons, results derived from a simulation should be compared to those from the hypergeometric distribution; if the simulation appears flawed, it is likely preferable to just use the hypergeometric distribution.

2.6. Multiple Testing

It is uncommon for an enrichment analysis to be performed against a single classifier. More commonly, a large collection of classifiers is considered, the goal being to identify previously unrecognized enrichment within the selection. If multiple analyses are being performed, our expectation of significance will need to be adjusted, a process known as multiple testing correction.

As a simple analogy, consider the discovery of an unusual cancer in three unrelated individuals living near a cell phone tower. If the incidence of the cancer in the general population is 1 in 1,000, and we tested all 250 people near the tower, the probability of finding 3 or more cases of the cancer is roughly 0.002. If we have chosen a significance threshold (α) of 0.05, we would deem the observation significant.

However, significance thresholds are usually chosen in the context of a "per comparison" error rate, the error we are willing to accept for any single test. If the researchers had gone through a medical textbook and considered 80 different rare tumors, the discovery was not found by a single comparison. Instead, we should apply a more stringent threshold to account for the fact that multiple tests were considered. The simplest form of multiple testing correction, the Bonferroni correction (12), scales the original threshold by the number of tests performed, so our corrected threshold (α_c) would be $0.05/80 = 0.000625$, and we find that we can no longer reject the null hypothesis. Such scaling based on the overall number of tests performed controls for the family-wise (within a set of tests) error rate.

Testing of multiple classifiers in an enrichment analysis presents the same problem. A Bonferroni correction is trivial to implement (simply correct by the number of classifiers tested), but is often overly conservative. Bonferroni assumes that the tests were entirely independent; in the case of hierarchical ontologies like GO, this assumption is obviously false, where child terms are completely contained within their parents. Alternative approaches include

- Sidak correction (13): Sometimes called Dunn-Sidak. The corrected significance threshold for T independent tests is $\alpha_c = 1 - (1 - \alpha)^{1/T}$. The Bonferroni correction represents an approximation of the Sidak correction. Sidak is marginally less strict than Bonferroni, but similar enough that Bonferroni should suffice as a reasonable substitute in effectively all cases.
- Holm–Bonferroni correction (14): An iterative approach; the *p*-values for all tests are rank ordered and checked from smallest to largest. Each test is compared to a corrected significance threshold as per Bonferroni, but the magnitude of the correction will decrease for each test. So for T tests, the best *p*-value will be corrected by $1/T$, the second best by $1/(T-1)$, and so forth. The process is continued so long as the corrected *p*-value is as good or better than the original significance threshold; as soon as a corrected *p*-value exceeds the original threshold, the process is halted, and only the prior *p*-values (if any) are accepted as meeting the corrected threshold.
- False discovery rate (FDR) (15): As defined by Benjamini and Hochberg, implementing FDR is very similar to Holm–Bonferroni, as it iteratively tests ordered *p*-values while

altering the correction factor for each test. For T tests, the best p-value is corrected by $1/T$, the second best by $2/T$, and so forth. Again, the process is halted when a p-value fails to meet the corrected threshold (the correction is somewhat different for negatively correlated dependent tests).

- Positive false discovery rate (pFDR) (16): The above approaches make very simple assumptions about the distribution of results for the performed tests. The pFDR, as defined by Storey, converts p-values into "q-values" by a process which considers the distribution of all the p-values actually observed during testing. Choosing a q-value threshold allows a researcher to then control the number of false positives actually selected. That is, a q-value cutoff of 0.05 indicates that the researcher wishes to limit the number of false positives in the *positives reported as significant* to 1 in 20; by contrast, a p-value cutoff of 0.05 indicates that *any* given test (not just those being reported) has a 1 in 20 chance of being a false positive.
- Permutation or simulation (17). If gene lists are available through permutation or another form of randomized simulation, the empirical observation of the likelihood of finding a classifier significant can serve as the corrected likelihood.

pFDR represents an excellent mechanism for controlling the number of false positives reported in an enrichment analysis. Implementation is non-trivial; fortunately, the Storey lab has provided an R module for calculating q-values. FDR and Bonferroni will both control for multiple testing, but will be excessively strict in most circumstances.

3. Software Options

A large number of software packages are available for performing enrichment analysis. Khatri and Draghici (18) provide an excellent overview of over a dozen programs, including critical analysis of the limitations and merits of each. Their own platform (OntoExpress) is freely available after registration and it provides a fairly sophisticated interface for performing enrichment analysis on Gene Ontology terms and chromosome location.

The following programs are examples which provide a reasonably rigorous approach to enrichment analysis. Given that many well-designed tools are available, only those without access restrictions are shown; academic users may find additional resources available at no cost. Hyperlinks are provided, but these may change over time; it is suggested that a web search be used to verify that the most recent site is being used. Further, since many tools are

under active development, it is advisable to occasionally run a generic web search to look for newly created or updated tools ("set enrichment Gene Ontology" works well as a query, since nearly all tools utilize Gene Ontology).

- **GSEA-P** (11, 19) – http://www.broad.mit.edu/gsea/. Rigorous permutation based enrichment analysis, provided as a desktop application, jar file or R implementation. Source code available from the Broad Institute. Uses raw expression results as input.
- **Onto-Express** (20) – http://vortex.cs.wayne.edu. Nice Java interface provides multiple options for namespaces, enrichment algorithms, and multiple testing correction.
- **ermineJ** (21) – http://www.bioinformatics.ubc.ca/ermineJ/. Standalone Java tool using multiple analysis methods with FDR multiple testing correction. Focused on microarray analysis, but allows set customization.
- **GeneTrail** (22) – http://genetrail.bioinf.uni-sb.de/. Accommodates multiple namespaces, custom reference sets, custom ontologies. Includes a large selection of common ontologies, plus some uncommon ones. Multiple testing correction by Bonferroni or FDR. Results are presented in summary text form, as well as graphically.
- **FUNC** (23) – http://func.eva.mpg.de/. Web portal or GPL-released command line tool, provides four algorithms to choose from. Estimates *p*-values from permutation of gene labels, also provides FDR corrections. Very flexible input allows arbitrary sets to be tested, but utilizes a specific format for each algorithm (additional user effort needed to initiate an analysis).
- **FuncAssociate** (17) – http://llama.med.harvard.edu/cgi/func1/funcassociate. Easy-to-use Gene Ontology analysis, with automated namespace resolution. Ranked or unranked lists are tested using Fisher's Exact Test. Multiple testing correction is achieved through simulation.
- **GOstat** (24) – http://gostat.wehi.edu.au/. Recognizes multiple namespaces, tests against Gene Ontology using chi-squared or Fisher's Exact Test. FDR and Holm–Bonferroni corrections are available. Results presentation is somewhat sparse. *Caution* – the program will use all annotated genes as the reference set by default; this will almost always be inappropriate.

Again, the above tools represent a sample of the available programs. Your choice on which tool(s) to use should be primarily determined by the quality of the analysis (algorithms used, appropriate choice of reference set and multiple testing correction, in particular). Ease of use, flexibility of configuration, speed of calculation, format of output, and availability for local use vary widely between tools, and may also help decide

which one is the most important. It is recommended that a testing data set with known enrichment be tried in multiple tools in order to compare results.

4. Interpretation

While the final interpretation of an enrichment analysis will always depend on the specific context of the original experiment, we can offer a few guidelines for focusing the process. Most important of these is to recognize the null hypothesis that you are testing. For label permutation, this will be "the distribution of class members is not different between experimental and control samples". For essentially all other analyses, it will be "the distribution of class members is independent and random". As mentioned in several sections above, there are many "uninteresting" reasons why gene classes may distribute non-randomly. You should consider any technical artifacts or biases that might lead to a class being identified as significantly enriched.

While some enrichment tools will simply report classifiers that are significantly different than the null hypothesis, most will distinguish between over-represented (enriched) and under-represented (depleted) classifiers. Interpreting the meaning of a depleted class can be more challenging than one that is enriched. Sometimes the observation can be explained by the co-occurrence of a mutual exclusive enriched class – depletion of "cytoplasmic" genes should not be surprising if you already have an enrichment of "extracellular".

When considering specific results, the *p*-value will help guide your confidence that the null hypothesis has been rejected for a class. If you are using *q*-values, you will also have a good estimate of the fraction of results that are likely to be false positives. Unless you are using a very conservative threshold, *p*-values near the threshold should be treated with skepticism. Marginal *p*-values will generally occur due to one or more small values in a contingency table (for example, a small selection or a class with only a few members, either overall or in your selection).

Failure to find any enriched classifiers (or expected ones) can be due to several reasons. The possibility that enrichment is not present (or too small to be considered significant) should always be considered. However, it is also possible that configuration of the analysis has led to a truly significant enrichment being missed:

- Significance threshold is too conservative. It is generally trivial to relax the threshold and re-run the analysis. Of course, you will incur increasingly large numbers of false positives in your reported results.

- Multiple testing correction is too aggressive. Bonferroni will over-correct for the vast majority of analyses being run. Choosing an alternative correction algorithm may allow less obvious enrichments to pass your threshold.
- Enrichment algorithm is not appropriate. The algorithm may be assuming a distribution that is inappropriate for your data or is reporting inaccurate results due to estimation. Trying another algorithm may provide more sensitivity.
- The classifier sets were not capable of finding the signal. Are you certain that the classifiers you have chosen adequately capture the kinds of classes you are interested in? Even if the appropriate class exists, it may be too poorly annotated, or be diluted with too many inappropriate annotations (*see* Vivanco et al. in the Applications section further for an example of this).
- Too many classifiers are being tested. Being able to test tens of thousands of classifiers is an extremely helpful tool in exploratory research, but it comes with the cost of a greater magnitude multiple testing correction. If you are not interested in some classes, or have an interest in a specific subset of classes, it is generally a good idea to test with a smaller set (an example of utilizing a focused classifier set is mentioned in Li et al. *(27)*).

Any modifications to the classifier sets should be made *before* running an analysis in order for the multiple testing correction to represent an true attempt at controlling false positives. If a test of the entire Gene Ontology tree yields no significant corrected results, re-running the analysis with just the terms that barely failed your threshold is not appropriate; you cannot pick a lottery number after the draw has been performed. This is one reason why it is advisable to familiarize yourself with a tool or ontology prior to analyzing your own data sets. Published gene sets can be found through web searches or can be obtained from colleagues. Determining optimal search parameters and classifier set membership with other researcher's data sets can help avoid tainting the multiple testing correction on your own data. Analyzing gene sets with previously reported enrichment can also help train your intuition for interpreting reported *p*-values in results.

5. Applications of Gene Set Enrichment Methods

The power of gene set analysis tools, both in terms of increased sensitivity in detecting effects in analysis of genomic data sets as well as conceptualization of the results, has led to their wide use in genomic research. Google Scholar estimates that there are roughly 71,000 articles indexed with the phrase "gene set enrichment",

although it is likely this measure is redundant. We present here a small selection of applications of gene set enrichment analysis. These include seminal studies, more recent applications, and examples where enrichment tools led to the generation of new hypotheses that were prospectively tested and, in some cases, confirmed. While enrichment has primarily been used to interpret RNA profiling data, examples outside this realm are also presented.

The seminal article by Mootha et al. uses permutation (via the Broad Institute's GSEA-P software) against RNA profiling samples to dissect changes associated with diabetic state in human muscle in 119 compiled gene sets (19). The most significant enrichment score was obtained from a gene set representing oxidative phosphorylation, including multiple subunits of mitochondrial ATP synthase, cytochrome C oxidase, and NADH dehydrogenase. These data were used to generate and test the hypothesis that these perturbations would be correlated with maximal oxygen uptake of individuals, ultimately finding a modest but significant correlation.

Houstis et al. apply GSEA-P in a study of the mechanisms of insulin resistance in 3T3-L1 adipocytes treated with Tumor Necrosis Factor or dexamethasone (25). Both agents are known to cause insulin resistance though their mechanisms of action are believed to be different; TNF having potent pro-inflammatory properties while dexamethasone is anti-inflammatory via glucocorticoid receptor agonism. Gene sets based on profile changes by each of the two treatments identify a single dominant category shared by both, reactive oxygen species biology (ROS). The hypothetical role of ROS in insulin resistance was then supported through measurement of ROS levels and insulin resistance in cells perturbed with known modulators of ROS. One of these agents was demonstrated to affect glucose homeostasis and insulin sensitivity in a murine in vivo experiment.

Vivanco et al. turned to enrichment analysis after mouse genetic models indicated that ATK activation alone was not sufficient to account for all tumorigenic effects observed following PTEN loss (26). GSEA-P analysis of PTEN-deficient cell lines identified enrichment in the JNK pathway activation. The paper is also a good example of how careful management of classifier sets can result in better enrichment signal and improved biological interpretation. The researchers noted that their human-curated JNK pathway class included both kinases and phosphatases. Given that these proteins generally operate in opposition to one another, they extracted the phosphatases into their own class. The resulting JNK phosphatase class showed significant enrichment and the residual JNK class demonstrated an improved significance as well.

Li et al. used a focused GSEA analysis to explore hepatic free fatty acid (FFA) toxicity (27). Gene sets were derived from human hepatoblastoma cells treated with various FFAs and tested against

37 classifiers previously associated with challenge by FFA or TNF-α. Of these, 14 showed significant enrichment in their data, including oxidative stress and energy generation. These 14 classes were then used in partial least squares regression to identify the genes most associated with cytotoxicity. NADH dehydrogenases were implicated and later experimentally validated in vitro. Also experimentally validated was the failure of ceramide metabolism to show significant enrichment in the GSEA analysis.

Grasser et al. applied enrichment analysis to elucidate the mechanism of BMP-initiated bone formation (28). Gene sets from oophrectomized mice treated with BMP-6 were analyzed with GSEA-P against 522 classifier sets. The analysis found significant enrichment in the IGF-I, EGF, and IL1R1 pathways. Upregulated expression of EGF and IGF-I ligands was experimentally confirmed using human osteoblast cells treated with BMP-6.

Radich et al. apply the hypergeometric distribution to gene sets derived from transcriptional changes associated with the progression of chronic myeloid leukemia (29). Testing both Gene Ontology and KEGG, they observed enrichment in Wnt signaling, myeloid differentiation, and apoptosis, among others.

A number of modern studies exploit combinations of genetic/epigenetic studies with analysis of transcription and gene set analysis methods to provide an integrated analysis of disease. Dehan et al. performed a comprehensive analysis and integration of genome-wide comparative genomic hybridization and RNA profiling across 23 nonsmall-cell lung cancer cell lines (30). Their approach identified dominant genomic amplifications and deletions associated with transformation of these lines as well as transcripts that were coordinatively regulated in relation to their physical genomic location. Upon analysis of their data with gene set methods and the Gene Ontology, a number of biological process terms were enriched. Notably, there were striking chromosomal regions that were amplified and enriched in genes responsible for degradation of extracellular matrix, while other regions were deleted that were associated with cell adhesion and cell junctions.

Finally, Dixon et al. have undertaken extensive genome-wide association studies in humans, comparing and correlating genotype with RNA transcription in derived lymphoblastoid cell lines derived from 400 subjects recruited from a study of childhood asthma (31). The resulting transcripts whose mRNA levels were associated with genotype (28% of transcripts tested) were also subjected to gene set analysis to the Gene Ontology, to test whether the identified heritability of transcription was associated with specific biological functions. The top Gene Ontology hits for transcripts that displayed heritability included response to unfolded protein, regulation of progression through cell cycle, and RNA processing. These results raise interesting and unanswered questions about the evolutionary advantage of polymorphisms in these biological processes.

6. Summary

Gene set enrichment is proving to be an exceptionally useful tool for molecular biologists in the post-genomic era. Enough interest exists in the field that multiple software solutions are available free of charge, and computational technology is such that even complex analyses can be performed in a reasonable length of time on inexpensive equipment. While the technique has become easier than ever for a bench scientist to perform, it remains important that the researcher be aware of the limitations of the technique and possesses a basic understanding of the assumptions made by the tools being used. In particular, the null hypothesis defined by the analysis should always be understood, care should be taken when defining the gene sets used, and statistical results should be properly controlled for multiple testing.

Acknowledgments

We would like to thank Roumyana Yordanova for her thoughts and advice on statistical matters.

References

1. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. and Sherlock, G. (2000) Gene Ontology: Tool for the unification of biology. Nature. 25, 25–29.
2. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) http://www.ncbi.nlm.nih.gov/Omim/
3. Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A., Orchard, S., Orengo, C., Petryszak, R., Selengut, J., Sigrist, C., Thomas, P., Valentin, F., Wilson, D., Wu, C. and Yeats, C. (2007) New developments in the InterPro database. Nucleic Acids Research. 35, D224.
4. PubMed. http://www.pubmed.gov/
5. Guide to GO Evidence Codes. Gene Ontology http://www.geneontology.org/GO.evidence.shtml
6. Fisher's exact test. Wikipedia http://en.wikipedia.org/wiki/Fisher's_exact_test
7. Hypergeometric distribution. Wikipedia http://en.wikipedia.org/wiki/Hypergeometric_distribution
8. Binomial distribution. Wikipedia http://en.wikipedia.org/wiki/Binomial_distribution
9. Chi-square distribution. Wikipedia http://en.wikipedia.org/wiki/Chi-square_distribution
10. Goeman, J. and Buhlmann, P. (2007) Analyzing gene expression data in terms of gene sets: Methodological issues. Bioinformatics. 23, 980–987.

11. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E. and Mesirov, P. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 102, 15545–15550.
12. Bonferroni correction. Wikipedia http://en.wikipedia.org/wiki/Bonferroni_correction
13. Ury, H. (1976) A comparison of four procedures for multiple comparisons among means (pairwise contrasts) for arbitrary sample sizes. Technometrics. 18, 89–97.
14. Holm, S. (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 6, 65–70.
15. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 57, 289–300.
16. Storey, J. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. Annals of Statistics. 31, 2013–2035.
17. Berriz, G., King, O., Bryant, B., Sander, C. and Roth, P. (2003) Characterizing gene sets with FuncAssociate. Bioinformatics. 19, 2502–2504.
18. Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: Current tools, limitations, and open problems. Bioinformatics. 21, 3587–3595.
19. Mootha, V., Lindgren, C., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., RidderstrÃ¥le, M., Laurila, E., Houstis, N., Daly, M., Patterson, N., Mesirov, J., Golub, T., Tamayo, P., Spiegelman, B., Lander, E., Hirschhorn, J., Altshuler, D. and Groop, L. (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics. 34, 267–273.
20. Khatri, P., Draghici, S., Ostermeier, G. and Krawetz, S. (2002) Profiling gene expression using onto-express. Genomics. 79, 266–270.
21. Lee, H., Braynen, W., Keshav, K. and Pavlidis, P. (2005) ErmineJ: Tool for functional analysis of gene expression data sets. BMC Bioinformatics. 6, 269.
22. Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y., Mueller, R., Meese, E. and Lenhof, H.-P. (2007) GeneTrail – Advanced gene set enrichment analysis. Nucleic Acids Research. 35, 186.
23. Prufer, K., Muetzel, B., Do, H.-H., Weiss, G., Khaitovich, P., Rahm, E., Paabo, S., Lachmann, M. and Enard, W. (2007) FUNC: a package for detecting significant associations between gene sets and ontological annotations. BMC Bioinformatics. 8, 41.
24. Beißbarth, T. and Speed, T. (2004) GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics. 20, 1464–1465.
25. Houstis, N., Rosen, E. and Lander, E. (2006) Reactive oxygen species have a causal role in multiple forms of insulin resistance. Nature. 440, 944–948.
26. Vivanco, I., Palaskas, N., Tran, C., Finn, S., Getz, G., Kennedy, N., Jiao, J., Rose, J., Xie, W., Loda, M., Golub, T., Mellinghoff, I., Davis, R. and Sawyers, C. (2007) Identification of the JNK signaling pathway as a functional target of the tumor suppressor PTEN. Cancer Cell. 11, 555–569.
27. Li, Z., Srivastava, S., Yang, X., Mittal, S., Norton, P., Resau, J., Haab, B. and Chan, C. (2007) A hierarchical approach employing metabolic and gene expression profiles to identify the pathways that confer cytotoxicity in HepG2 cells. BMC Systems Biology. 1, 21.
28. Grasser, W., Orlic, I., Borovecki, F., Riccardi, K., Simic, P., Vukicevic, S. and Paralkar, V. (2007) BMP-6 exerts its osteoinductive effect through activation of IGF-I and EGF pathways. International Orthopaedics. 31, 759–765.
29. Radich, J., Dai, H., Mao, M., Oehler, V., Schelter, J., Druker, B., Sawyers, C., Shah, N., Stock, W., Willman, C., Friend, S. and Linsley, P. (2006) Gene expression changes associated with progression and response in chronic myeloid leukemia. Proceedings of the National Academy of Sciences. 103, 2794.
30. Dehan, E., Ben-Dor, A., Liao, W., Lipson, D., Frimer, H., Rienstein, S., Simansky, D., Krupsky, M., Yaron, P., Friedman, E., Rechavi, G., Perlman, M., Aviram-Goldring, A., Izraeli, S., Bittner, M., Yakhini, Z. and Kaminski, N. (2007) Chromosomal aberrations and gene expression profiles in non-small cell lung cancer. Lung Cancer. 56, 175–184.
31. Dixon, A., Liang, L., Moffatt, M., Chen, W., Heath, S., Wong, K., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. and Cookson, W. (2007) A genome-wide association study of global gene expression. Nature Genetics. Advanced online publication. 39, 1202–1207.