Pacific Symposium on Biocomputing 14:368-379 (2009)

BIOFILTER: A KNOWLEDGE-INTEGRATION SYSTEM FOR THE MULTI-LOCUS ANALYSIS OF GENOME-WIDE ASSOCIATION STUDIES*

WILLIAM S. BUSH, SCOTT M. DUDEK, AND MARYLYN D. RITCHIE

Center for Human Genetics Research, Vanderbilt University
Nashville, TN 37232, USA

Genome-wide association studies provide an unprecedented opportunity to identify combinations of genetic variants that contribute to disease susceptibility. The combinatorial problem of jointly analyzing the millions of genetic variations accessible by high-throughput genotyping technologies is a difficult challenge. One approach to reducing the search space of this variable selection problem is to assess specific combinations of genetic variations based on prior statistical and biological knowledge. In this work, we provide a systematic approach to integrate multiple public databases of gene groupings and sets of disease-related genes to produce multi-SNP models that have an established biological foundation. This approach yields a collection of models which can be tested statistically in genome-wide data, along with an ordinal quantity describing the number of data sources that support any given model. Using this knowledge-driven approach reduces the computational and statistical burden of large-scale interaction analysis while simultaneously providing a biological foundation for the relevance of any significant statistical result that is found.

1. Introduction

1.1. *Genome-Wide Association Studies (GWAS)*

Over the last five years, genome-wide association studies (GWAS) have become a very popular study design for identifying genetic variants that incur disease risk in human populations. The overall strategy of the GWAS approach is inherently high-throughput, allowing investigators to blanket the genome with hundreds of thousands of single nucleotide polymorphisms (SNPs) in many individuals with the general goal of elucidating genetic causes of common human phenotypes – complex diseases in particular.

Traditional methods of genetic study design and analysis which excelled at identifying the rare mutations that cause Mendelian genetic disease have not performed as well for common complex disease, such as sporadic breast cancer or autism. Numerous candidate gene studies have been conducted for complex diseases, where particular genes of interest are investigated, but in many cases

* This work was supported by National Institutes of Health grants HL65962, and AG20135.

the results of these studies fail to replicate in other samples. While GWA studies are beginning to unravel the genetics of these complex diseases, one possible explanation for the lack of consistent findings from traditional studies is epistasis, or gene-gene interaction, and unless explicitly assessed, it may affect GWA studies also.

1.2. *Epistasis in GWA Studies*

Epistasis was first described by Bateson as the effect of one gene masking (or literally *standing upon*) the effect of another [1]. The Bateson view of epistasis has also been described as *biological epistasis* [2], where variation in the physical interaction of biomolecules affects a phenotype [3]. From a statistical perspective, epistasis was also observed as multi-allelic segregation patterns by Fisher who mathematically described the phenomenon as deviation from additivity in a linear model of genotypes [4]. Statistical epistasis and biological epistasis eventually converge as scientific understanding progresses. For example Bridges discovered statistical epistasis in *Drosophila* eye color, where collections of alleles Mendelize with various eye color phenotypes [5]. These alleles influence a common set of biochemical pathways controlling eye pigmentation that was elucidated many years later [6]. Epistasis can cause non-replication of single-SNP effects. If the effect of one allele is conditional on the presence of a second unknown allele, that second allele may not be present in a new population, and the effect of allele one will not replicate.

As epistasis is believed to play an important role in the genesis of complex disease, analysis strategies for detecting epistasis in large-scale data are increasingly important. A major hurdle in discovering epistasis, however, is the variable selection problem. Exhaustively evaluating all two-marker models in whole-genome data is a computational and statistical challenge, as processing the 5.00e11 possible two-marker models from a set of 1 million SNPs requires extensive computing resources and produces a plethora of statistically significant results with limited biological interpretability.

Two approaches are commonly suggested to address the variable selection problem. One approach is to select SNPs based on the strength of independent main effects, evaluating interactions only between SNPs that meet a certain effect size threshold. Another approach is to evaluate multi-marker combinations based on biological criteria [7]. Each of these strategies imposes a specific bias into the analysis, and neither strategy will be optimal in all cases. If we select or filter variables based on their main effects, we bias the analysis using statistical information, and assume that relevant interactions occur only between markers that independently have some effect on the phenotype alone.

Several studies have proposed complex theoretical penetrance models that influence the trait only through the interaction of two or more genetic variants [8-10], and filtering based on main effects would potentially miss these types of discoveries. If we filter variables using biological information – i.e. only examine interactions between SNPs in a common pathway or with a common structure or function – we bias the analysis in favor of models with an established biological foundation in the literature, and novel interactions between SNPs would be missed. Furthermore, the entire analysis is conditional upon the quality of the biological information used.

Several new tools have recently been developed to incorporate biological information with analytical approaches for GWAS data. Prioritizer is a Bayesian approach to incorporate multiple sources of gene interrelationships in a global "functional gene network". This network is used to prioritize significant single-SNP results by gene function [11]. Others methods use structured knowledge as a way to guide (but not restrict) variable selection for regression-based modeling. Province and Borecki propose a Bayesian re-sampling approach to select collections of SNPs that may have very small independent effects but function in aggregate explain a more substantial portion of trait variance [12]. Conti proposes a hierarchical modeling approach that uses an expert knowledge ontology to search for and test complex multi-SNP models. This Bayesian modeling process is flexible, allowing SNPs outside the knowledge-base to also be used in models (Pharmacogenomics Research Network Presentation 2008).

We propose a strategy that steps beyond the annotation and grouping of independent SNP effects, but does not attempt to jointly model large numbers of SNPs simultaneously. Also, we believe that ultimately data from multiple sources will better facilitate a comprehensive analysis, providing a biological foundation for testing specific multi-SNP association models in GWAS data. In this work, we present the Biofilter, a tool for knowledge-driven multi-SNP analysis of large scale SNP data. The Biofilter fundamentally differs from other methods in the way knowledge is incorporated into the analysis pipeline. The Biofilter uses biological information about gene-gene relationships and gene-disease relationships to construct multi-SNP models before conducting any statistical analysis. Rather than annotating the independent effect of each SNP in a GWAS dataset, the Biofilter allows the explicit detection and modeling of interactions between a set of SNPs. In this manner, the Biofilter process provides a tool to discover significant multi-SNP models with non-significant main effects that have established biological plausibility. This approach has the added benefit of reducing both the computational and statistical burden of exhaustively evaluating all possible multi-SNP models.

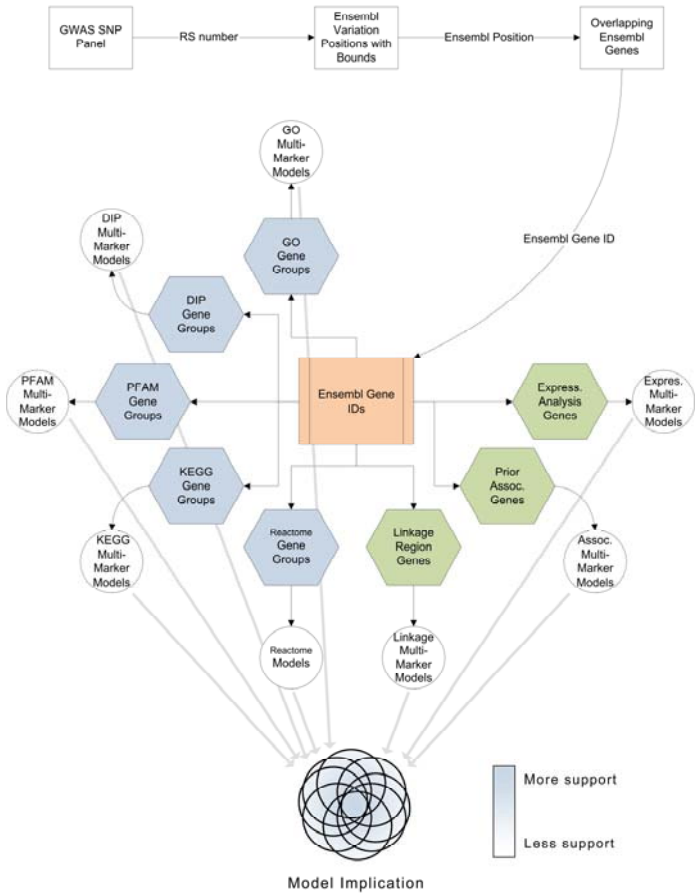

Figure 1. Overview of the Biofilter process. GWAS platform SNPs are mapped to Ensembl gene IDs and related to disease-independent sources (left) and to disease-dependent sources (right). Multi-marker models are generated from SNPs within knowledge-related genes. Derived models are overlaid to assess overall model implication.

2. Methods

2.1. *Overview*

An overview of the Biofilter is show in figure 1. The Biofilter model generation process is gene-centric, and as such, SNPs from GWAS genotyping platforms

must first be assigned to genes. Relationships between the genes represented by a platform can then be translated to multi-SNP models. Structured biological knowledge relevant to GWAS interaction analysis can come from various sources. We have partitioned relevant knowledge into two basic types: disease-dependent and disease-independent. *Disease-dependent* knowledge is information that relates a gene to the disease phenotype being studied, such as a previously associated SNP or a gene that is over-expressed in cases. *Disease-independent* knowledge is information that relates genes to one another, or defines collections of genes, such as a metabolic pathway or a common structural motif. These two types of information can be combined to form different classes of multi-SNP models and provide a measure of how strongly implicated a given model is based on the current available knowledge.

2.2. *Database Integration*

We chose Ensembl as our source of gene and SNP positional information due to ease of access and its clearly defined database schema. RS numbers for SNPs used in genotyping are matched to records within the Ensembl variation database (Release 49) to retrieve position information. The probe positions were joined to gene information tables to determine if the SNPs lie within one of the 32,000 known Ensembl genes that physically map to the autosomes or the X and Y chromosomes. This SNP-to-gene mapping is stored as a derived table in the Biofilter database.

Disease-dependent knowledge sources link individual genes to a disease phenotype. The goal of using disease-dependent knowledge is to identify genes that have some prior evidence of putative influence on the phenotype. One systematic source of disease-dependent knowledge is the Genetic Association Database (GAD). GAD is an archive of human genetic association studies of complex diseases and disorders established in 2004 [13]. GAD contains a list of prior associated genes for a variety of specific disease phenotypes and broader phenotype classes. Other types of disease-dependent knowledge may require manual selection from literature. Previous regions of genetic linkage, studies of differential gene expression, and hypothetical disease etiologies are possible sources of disease-gene relationships.

Disease-independent sources link two or more genes together. The goal of using disease-independent knowledge is to identify pairs of genes with some prior evidence of putative epistasis. The Gene Ontology project (GO, accessed on 3/16/08) is a collaborative effort to characterize and describe gene products in a collection of three hierarchical ontologies: cellular component, biological process, and molecular function. Because of its hierarchical structure, some broad ontology categories contain many hundreds of genes. For this analysis, smaller, more precisely defined gene categories (< 30 genes) were used as these

presumably contain stronger gene relationships. The Database of Interacting Proteins (DIP, 1/14/08 update) documents experimentally determined protein-protein interactions from more than 80 organisms [14]. We used the pair-wise human protein-protein interaction set contained in DIP to produce gene-gene pairs. The Protein Families Database (PFAM, Release 22) uses multiple sequence alignments and hidden Markov models to identify common protein domains and families based on structural and functional sequence patterns [15]. Generating pairs of genes using this data relies on the hypothesis that members of the same protein family are more likely to jointly influence disease risk. As such, we generated gene-gene pairs within proteins having the same domain, the same protein family, the same structural motif, and the same sequence repeat. The Kyoto Encyclopedia of Genes and Genomes (KEGG, 3/6/08 update) Pathway set is a collection of manually drawn pathway maps for a variety of metabolic and signaling pathways [16]. Reactome (Version 24) is a database of curated core pathways and reactions in human biology [17]. Netpath is a relatively new source of curated immune signaling and cancer pathways provided by the Pandey Lab at Johns Hopkins University and the Institute of Bioinformatics [18]. With these pathway collections, all possible gene-gene pairs were generated within each pathway-based gene group.

Relational data sources were downloaded and reconstituted in their original form within a MySQL database using Perl scripts. Using the schema for each data source, proteins and/or genes were translated to Ensembl gene IDs, and derivative tables containing gene groupings (such as protein families) were generated within the Biofilter database. Non-structured data sources, such as gene lists from publications, were manually imported into the Biofilter database. These gene lists were then translated to Ensembl gene IDs and used to establish gene groupings.

2.3. *Model Types and Generation*

Using both disease-dependent and disease-independent data sources, there are four types of two-SNP models possible: disease-independent, disease-dependent, hybrid with one disease-dependent gene, and hybrid with two disease-dependent genes. Figure 2 illustrates each of these model types. Disease-dependent information is based on a set of genes that are related to disease, visualized in the figure as a collection of dashed boxes, or unconnected nodes of a graph. Disease-dependent models are generated by exhaustively pairing all possible combinations of disease-related genes. Disease-independent information is based on relationships between sets of genes, visualized in the figure as a set of lines, or edges in a graph. To build disease-independent models, we generate pair-wise combinations of SNPs located in genes that are related, illustrated by as edges in figure 2. Hybrid models blend disease-

dependent and disease-independent information, and can contain either one or two disease-related genes. In the figure, one-gene hybrids must be connected by an edge (disease-independent connection) and contain at least one dashed square (disease-dependent gene). Two-gene hybrids must contain two dashed squares connected by an edge, meaning that there is evidence for biological interaction of two disease-related genes. For each pair-wise combination of genes, all possible two-SNP models across the two genes are built. For example, if there are two SNPs (A and B) in gene one and there is one SNP (C) in gene two, two models are generated (A,C and B,C).

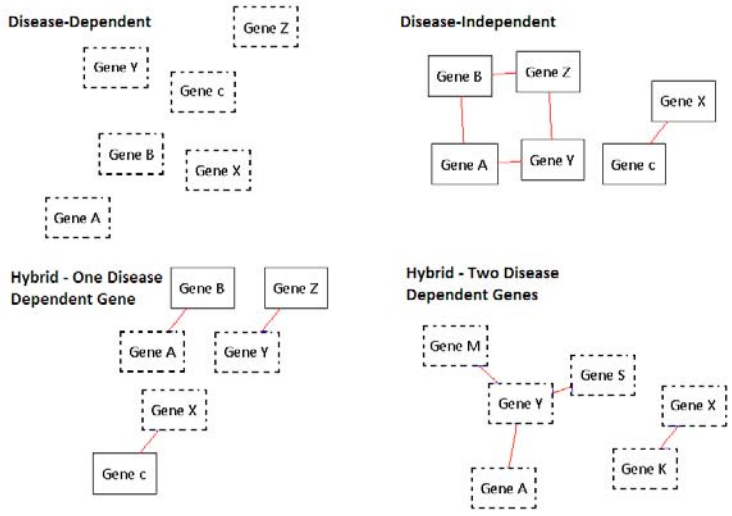

Figure 2. Two-gene model types. Each box represents a gene, and each line a connection between genes. Boxes that are dashed have been previously linked to disease by at least one data source.

2.4. *Model Implication*

Each model constructed has a set of Biofilter data sources that support it. If a combination of genes is supported by multiple data sources, it is likely more accepted by the scientific community and therefore may be more biologically plausible. We quantify the degree of knowledge-based support for a model with an *implication index*. The implication index is simply the number of data sources that provide evidence of gene-gene interaction or gene-disease relationship, and is calculated simply by summing the number of data sources supporting each of the two genes and the connection between them. For example, if one disease-related gene supported by two data sources is connected to another non-disease related gene, and that connection is supported by three

data sources, the implication index of the model is five. In this manner, the implication index provides an ordinal value representing the strength of the biological plausibility of a multi-SNP model.

Table 1. General GWAS platform statistics

	Illm1M	Affy60
Total SNPs (with RS IDs)	1,055,373	924,689
SNPs within genes	493,854	353,913
Genes represented	21,024	17,418
Common SNPs	267,900	
Common SNPs within genes	118,355	
Common genes represented	16,908	

3. Results

3.1. *GWAS Platform Representation*

Two large-scale genotyping platforms were assessed in this study: the Illumina Human1M-Duo BeadChip (Ilmn1M) and the Affymetrix Genome-Wide Human SNP Array 6.0 (Affy60). For the purposes of our assessment, only probes with vendor-specified Reference Sequence (RS) numbers were used in order to assure continuity of genomic position. General statistics for these two platforms are shown in Table 1.

As described in the methods, genes can be defined as disease-dependent and disease-independent. Because the disease-independent genes can be used in an analysis of any phenotype, we focused on how those genes are represented in the different data sources. Table 2 shows the number of gene pairs represented from each platform in each disease-independent data source. Table 3 shows the pair-wise overlap across the six different public databases.

3.2. *Generalized Disease Independent Models*

As disease-independent models are universally applicable to any phenotype, we provide files containing all derived gene-gene pairs, and platform specific two-SNP models for the Illm1M and Affy60 (available: http://chgr.mc.vanderbilt.edu/ritchielab/method.php?method=biofilter). Counts of these two-SNP models by implication index are shown in table 4. The Illm1M platform covers over a million more two-SNP models than the Affy60. This is most likely due to a lack of RS numbers for many Affymetrix probes, rather than a lack of gene coverage.

Table 2. Gene pairs represented by platform.

Data Source	Illm1M	Affy60
PFAM	14911	12837
DIP	747	638
GO	6129	5359
KEGG	4058	3543
Reactome	1799	1610
Netpath	3704	3246

Table 3. Pair-wise overlap of all known genes in disease-independent Biofilter data sources. The data source listed on each row contains genes that overlap the data source listed on each column. Cell values indicate the proportion of the genes in the column source that are represented in the row source.

	PFAM	DIP	GO	KEGG	Reactome	Netpath
PFAM	1	0.95	0.92	0.95	0.95	0.92
DIP	0.05	1	0.09	0.08	0.15	0.12
GO	0.37	0.73	1	0.63	0.64	0.56
KEGG	0.01	0.02	0.01	1	0.02	0.02
Reactome	0.12	0.36	0.19	0.25	1	0.21
Netpath	0.22	0.6	0.34	0.41	0.41	1

Performing an exhaustive analysis of all possible two-SNP models within genes represented by these two platforms would result in 1.22e11 models for the Illumina 1M and 6.26e10 models for the Affymetrix 1M. By reducing the interaction search space to only models with established biological plausibility via the disease-independent data sources, only 2.23e9 (Illm1M) and 1.2e9 (Affy60) model evaluations are required. Applying a Bonferroni correction to the exhaustive approach would require a model fit p-value of 4.10e-13 (Illm1M) and 7.98e-13 (Affy60) to be statistically significant. In contrast, using the knowledge-based approach, a Bonferroni correction of 2.25e-11 (Illm1M) or 4.16e-11 (Affy60) is required. In this manner, reducing the search space not only improves computation time, but also reduces the statistical burden of conducting biologically non-relevant statistical tests. Further model restriction (such as using models with an implication index > 1) would further reduce the Bonferroni adjusted significance threshold.

4. Discussion

When examining epistasis in genome-wide association studies, there are several variable selection strategies. Exhaustive evaluation of all multi-SNP models is generally computationally impractical. Exploring epistasis within a set of SNPs with detectable main effects may prevent the discovery of complex genetic

models where trait variance is explained largely by the interaction of SNPs. Using biological knowledge to perform SNP selection provides two key benefits simultaneously: it reduces the multi-SNP model search space, and it provides a biologically plausible foundation for the models to be evaluated. We developed the Biofilter to systematically reduce model search space based on multiple sources of structured biological knowledge.

We mapped the disease-independent models generated by the Biofilter to two GWAS genotyping platforms, the Affymetrix 1M and the Illumina 1M. The final evaluated model search space was 0.241% of the exhaustive model space for Affy60 and 0.40% of the exhaustive model space for Illm1M when requiring at least one source of structured biological knowledge connecting the two genes in a two-SNP model, with further reductions possible by adjusting the number of required knowledge sources implicating the model.

Table 4. Disease-independent gene pairs and model counts by implication index.

Implication Index	Illm1M Gene Pairs	Illm1M Two-SNP Models	Affy60 Gene Pairs	Affy60 Two-SNP Models
1	4,679,363	2,174,328,700	3,505,773	1,162,090,222
2	87,163	44,960,600	67,341	36,825,703
3	8,065	6,425,788	6,094	4,102,173
4	715	397,966	546	171,075
5	45	11,033	40	4,122
6	1	569	1	757
Total	4,775,352	2,226,124,656	3,579,795	1,203,194,052

The Biofilter method of variable selection can be implemented with a variety of analysis techniques, including logistic regression, classification and regression trees, and basic categorical statistics, among many others. To this end, the Biofilter is being developed as a knowledge-based filter in part of a larger analysis framework. The collection of multi-SNP models generated by the Biofilter can be passed seamlessly to several analytical methods. Statistical properties of each multi-SNP model are then stored, allowing retrieval of results with complete annotation of the SNPs, genes, gene grouping information, and in some cases, PubMed references to the original articles implicating the model. The end result of this analysis pipeline is a set of biologically plausible, statistically relevant multi-SNP genetic models.

Some approaches may be adapted to incorporate the implication index into the analysis plan. Prioritized subset analysis, for example, partitions statistical results based on prior biological knowledge. The false discovery rate (FDR) for

the "prioritized subset" is estimated separately, improving power when the prior knowledge is accurate [19]. Applying this strategy to subsets defined by the implication index could improve statistical power via p-value correction.

Ranking models based on the number of supporting data sources may introduce unknown literature-based biases. Some data sources may have inter-dependencies, where one source was referenced in the creation of another. The breakdown of gene overlap for the 6 disease-independent data sources shows the diversity of gene-pairs represented, though notably PFAM contains nearly all of the gene-pairs established by the other sources. This is likely because PFAM contains the largest number of genes. When using disease-dependent data sources, there are certainly many factors that influence the inclusion and promotion of specific genes in relation to a phenotype, such as reporting bias.

Overall, the Biofilter provides a systematic way to assess the level of knowledge-based support for a given genetic model, provide a ranked list of all possible knowledge-based models, and to statistically test each of these hypotheses in genome-wide association data.

Acknowledgments

This work was supported by National Institutes of Health grants HL65962, and AG20135.

References

1. Bateson W: *Mendel's Principles of Heredity.* Cambridge: Cambridge University Press; 1909.
2. Moore JH, Williams SM: **Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis.** *Bioessays* 2005, **27:**637-646.
3. Moore JH: **The ubiquitous nature of epistasis in determining susceptibility to common human diseases.** *Hum Hered* 2003, **56:**73-82.
4. Fisher RA: **The Correlation Between Relatives on the Supposition of Mendelian Inheritance.** *Transactions of the Royal Society of Edinburgh* 1918, **52:**399-433.
5. Bridges CB: **Specific modifiers of eosin eye color in Drosophila melanogaster.** *J Experimental Zoology* 1919, **28:**337-384.
6. Lloyd V, Ramaswami M, Kramer H: **Not just pretty eyes: Drosophila eye-colour mutations and lysosomal delivery.** *Trends Cell Biol* 1998, **8:**257-259.
7. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA: **Mapping complex disease loci in whole-genome association studies.** *Nature* 2004, **429:**446-452.

8. Culverhouse R, Suarez BK, Lin J, Reich T: **A perspective on epistasis: limits of models displaying no main effect.** *Am J Hum Genet* 2002, **70:**461-471.
9. Frankel WN, Schork NJ: **Who's afraid of epistasis?** *Nat Genet* 1996, **14:**371-373.
10. Moore J, Hahn L, Ritchie M, Thornton T, White B: **Routine discovery of complex genetic models using genetic algorithms.** *Applied Soft Computing* 2004, **4:**79-86.
11. Franke L, van BH, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.** *Am J Hum Genet* 2006, **78:**1011-1025.
12. Province MA, Borecki IB: **Gathering the gold dust: methods for assessing the aggregate impact of small effect genes in genomic scans.** *Pac Symp Biocomput* 2008,190-200.
13. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database.** *Nat Genet* 2004, **36:**431-432.
14. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.** *Nucleic Acids Res* 2002, **30:**303-305.
15. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL et al.: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36:**D281-D288.
16. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T et al.: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36:**D480-D484.
17. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de BB, Gillespie M, Jassal B, Lewis S et al.: **Reactome: a knowledge base of biologic pathways and processes.** *Genome Biol* 2007, **8:**R39.
18. **NetPath** [http://www.netpath.org]
19. Li C, Li M, Lange EM, Watanabe RM: **Prioritized subset analysis: improving power in genome-wide association studies.** *Hum Hered* 2008, **65:**129-141.