NAIST-IS-DD1361017

Doctoral Dissertation

Protein sequence modeling and transcription regulation network analysis towards integrative big data biology

Kibinge Nelson Kipchirchir

March 14, 2016

Department of Applied Informatics
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation

submitted to the Graduate School of Information Science,

Nara Institute of Science and Technology

in partial fulfillment of the requirements for the degree of

Doctor of Science.

Thesis Committee:

Professor Shigehiko KANAYA	(Supervisor)
Professor Keiichi YASUMOTO	(Co-supervisor)
Associate Professor MD. ALTAF-UL-AMIN	(Co-supervisor)
Associate Professor Tadao SUGIURA	(Co-supervisor)
Assistant Professor Naoaki ONO	(Co-supervisor)

Protein sequence modeling and transcription regulation network analysis towards integrative big data biology*

Kibinge Nelson Kipchirchir

Abstract

Informatics has played key roles across the so called 'omics' studies including genomics, transcriptomics and proteomics. Research has recently focused on platforms that combine multiple sources of high-throughput biological data in the course of analysis. In this work, present two frameworks, one of which focuses on protein sequence representation and the other focusing on transcription regulation networks. Both applications; as we will describe in the course of this dissertation, will bolster the idea that integrative frameworks for molecular data are likely to improve interpretation and hypothesis generation. Such strategies increase 'biological objectivity' of informatics oriented tools such as sequence-based machine learning and molecular network mining.

In the Chapter 3, we present a residue representation procedure for protein sequence analysis applications. We utilized the amino acid index to determine a set of physical properties that quantitatively model protein sequence contents. This method was tested through application of these selected biochemical and physical attributes in examination of peptide sequence diversity of terpenoid synthases accumulated in the KNApSAcK motorcycle database. In the Chapter 4, we describe a transcription regulation network analysis framework based on pathway-stratification and multi-layered network mining modules. This workflow was implemented as a web application with functions including pathway visualization and comparison of transcription factor activity between sample conditions defined in the experimental design. The network mining components comprised of differential expression, network construction, pathway-based abstraction, clustering and visualization applications. This framework was tested through application in analysis of expression datasets related to cancer.

Keywords:

protein sequence representation, transcription regulation and network analaysis

*Doctoral Dissertation, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1361017

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Big data biology

Big data biology can be loosely be defined as the branch of biology dealing with collection, curation and value extraction in high-throughput molecular datasets of various forms with the leverage of computational and bioinformatics resources. As a subject of intense research in recent times, it has played a pivotal role in various research projects ranging from database development, indexing, ontology creation and data-mining. Owing to the range of possibilities in this rather nascent field of study, its scope and relevance has not been sufficiently emphasized. As prices of instruments drop, even small laboratories become hubs for generation of huge datasets. Databases and software tools have become accessible and computational infrastructure too is becoming more readily available and affordable. Cloud computing technologies have enabled easy sharing of data thus giving researchers quick access to information. One such platform is the European Bioinformatics Institute (EBI) mirror sites on the Amazon Web Services Elastic Compute Cloud (EC2).

1.1.1 Role and relevance of big data and computational biology

The enormous complexity of biological systems is a platform upon which the flexibility of computational resources can be applied to increase capacity of experimental data collection, storage and processing. Computational biology is generally split into two main branches: data mining/knowledge discovery which aims to uncover hidden patterns in large datasets and simulation computational biology which applies principles of biology to build and test in hypotheses *in silico*. Data mining is used extensively for purposes such as gene prediction,

structure and function prediction, transcription regulation network inference and other objectives. This type of computational biology relies on sophisticated algorithms primarily from information sciences e.g Hidden Markov Models (HMMs). Simulation computational biology similarly applies statistical, mathematical and information science based concepts to construct hypotheses and test computational models and map their meaningfulness in the context of biology principles. Instrumentation has increased the capacity to churn out data in different forms and at varying levels.

Sequence and functional data

Approaches excluding the need for culturing in the sequencing flow have enabled scientists and clinicians to efficiently generate data at a rate that could not have been achieved before. This has rapidly improved our understanding of biology. Molecular data collection through technologies such as next generation sequencing (NGS) has been applied in deciphering genomes, transcriptomes, proteomes and antibiotic sequence information. This often involves discrete processes whose output is almost unachievable without automation and precise algorithmic functions. Sequence assembly for instance is akin to piecing together jig-saw puzzles from large numbers of sequence fragments. Another area of significance to bioinformatics is the sequence preprocessing tools including shell scripts and other file processing applications. Computational biology certainly plays an indespensable role especially as the cost of sequencing lowers and the throughput of sequencing increases. Sequence databases such as genebank (Benson *et al.*, 2000), Universal protein resource (UniProt) (Consortium *et al.*, 2008), Saccharomyces Genome Database (SGD) (Cherry *et al.*, 1998) and similar DBs are computational resources that curate, charaterize, systematize and disseminate sequence data from research all over the world. Other areas of application of computational biology at the sequence level include, sequence alignment, gene prediction, domain prediction and mapping transcription start sites. Functional differences at the sequence level is often examined by technologies such as gene expression profiling and genomewide association studies (GWAS). These rely on bioinformatics for data analysis. As the sequencing technology improves, instruments will produce even longer reads at significantly lower costs. An immediate future objective will be the implementation of rapid and precise sequencing tools usable at the clinical setting to help in improving health strategies in curing and preventing diseases.

Metabolome and pathway data

Another area of massive progress in throughput is metabolome quantification. Tools such as mass spectrometry, nuclear magnetic reasonance and other compound quantification methods have been applied in metabolite metrification, structure elucidation, volatile organic compound detection and novel compound identification inspite of the numerous challenges mostly associated with high signal to noise ratio in such metabolome technologies. The resulting data is accumulated in databases such as KNApSAcK (Afendi *et al.*, 2012), KEGG (Kanehisa and Goto, 2000), Metacyc (Caspi *et al.*, 2008) and other metabolome databases. At our laboratory, we have built the KNApSAcK family database to profile, species-species interactions, volatile organic compounds (VOC) as well as food and herbal medicines from around the world. The data in these repositories can be mined for value to further understand the underlying features and to solve metabolome level challenges. Bionformatics comes into play at this level in areas such as database development and pathway enrichment analysis. In addition to this, data cataloguing, processing and conversion, feature detection, normalization, compound identification and quality control employ computational biology. Vendor-specific formats require bioinformatics to facilitate data sharing and standardization. Metabolome informatics also plays a role in feature detection through algorithms such as Gaussian model fitting and tranformation. This has been applied in separation of noisy peaks from signal peaks. A number of other algorithms have been developed for quality checking of spectral data. Determining novel metabolites also depends on filtration algorithms.

Data mining

There is a need for innovative, novel and sophisticated techniques for extracting value from the datasets. Computational biology is a tool in itself that can be applied to shape and refine hypotheses. More importantly it is applicable in advancing the role of science in accelerating discovery of new drugs, vaccines and disease management technologies. This has a direct bearing towards personalized medicine, understanding evolutionary footprints in genomes, quantitative ecology and in solving global issues such as development of alternative energy sources such bio-fuels.

Computational biology is a key component in the drug design process. Computation assisted drug design hase been extensively applied in pharmaceautical research and development. Together with this, identification of biomarkers for diseases is largely dependent on data-mining based

on tools such as machine learning algorithms and annotation platforms such as Gene Ontology (GO) (Ashburner *et al.*, 2000). In this context, an amalgamation of statistical, mathematical and informatics subjects interact with biology in a broader quantitative sense.

1.1.2 Achievements of computational biology

Increased computational power has enabled applications that would have been impossible a few years ago. Among the earlier mainstays of bionformatics, basic local alignment search tool (BLAST) (Altschul *et al.*, 1990) stands out the most. As databases and data technologies broadened towards the end of the twentieth century, such computational tools have been extensively utilized for sequence comparison. BLAST has been particularly vital for function elucidation, prediction, annotation and mapping evolutionary footprints in sequence data of newly sequenced genomes. This has been made possible, fast and efficient by BLAST's inherent heuristic criteria of matching sequences.

Another notable milestone achievement is the ENCODE (ENCyclopedia of DNA Elements) project (Consortium *et al.*, 2004). The main objective of ENCODE is to map functional elements of the human genome. The project has contributed towards changing the previous notion that majority of the DNA was junk. Recent publication showed that about 62 percent of the genome shows active RNA transcription (Consortium *et al.*, 2012). Computational biology has played a role in this project by disseminating the data through the projects data portal www.encodeproject.org.

1.1.3 Objectives

On a broader scale, bionformatics aims to develop tools and systems that facilitate storage, analysis and inference deduction from biological data. Our main subject in this dissertation is to report usable bionformatics applications and pipelines developed in the course of this work. We will in subsequent chapters introduce two tools; one of which centered on analyzing protein sequences and the other on data-mining of transcription regulation networks from expression datasets through an integrated framework.

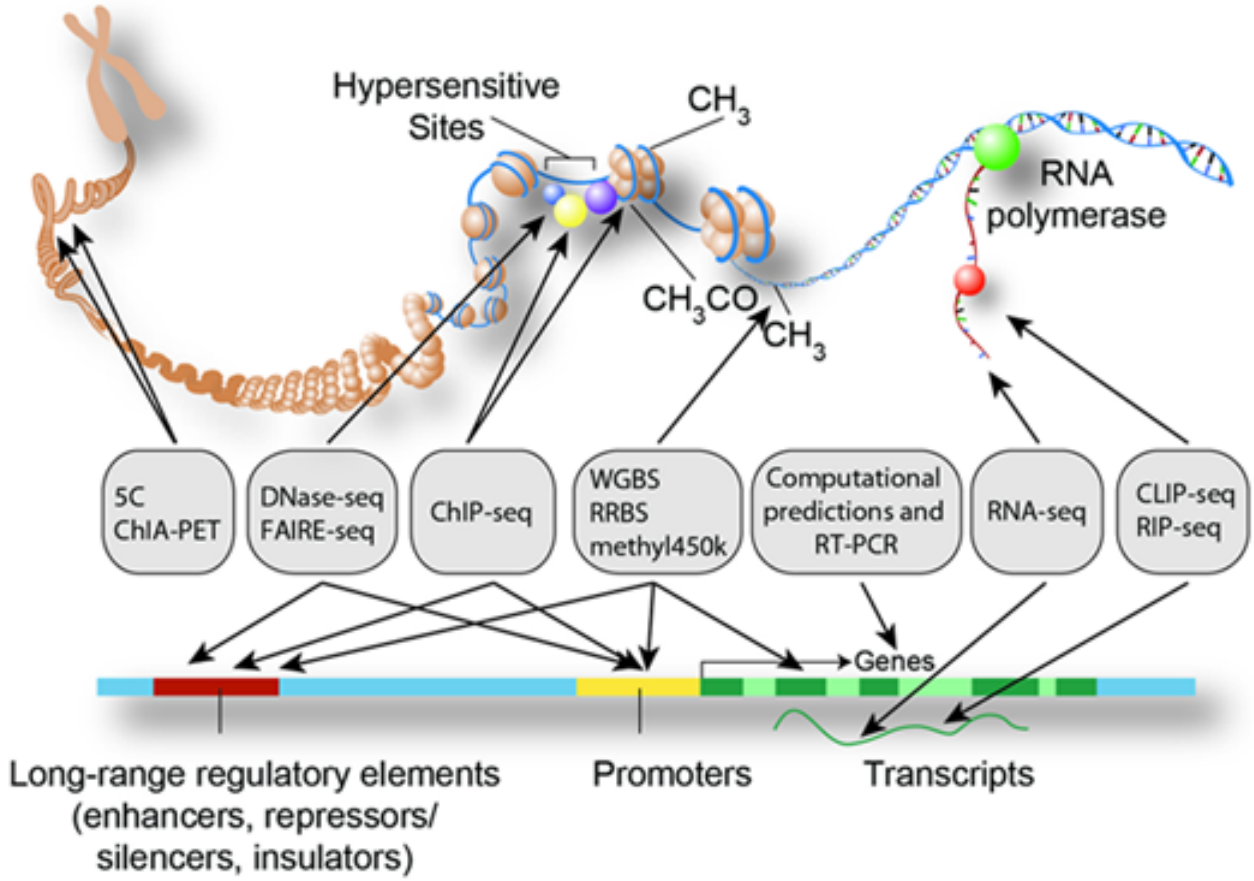

Figure 1.1: *The ENCODE projects aims to identify all functional elements in a human genome. (Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI). Downloaded from www.encodeproject.org on 5th September 2015)*

Table 1.1: A search for the term 'big data biology' at the NCBI website

Category	Counts	Description
Literature	33074	Books, reports, journals, abstracts, PubMed indexing and other forms of literature
Health	51	Human variations of clinical significance, genotypes, clinical effectiveness, disease and drug reports
Genomes	2467	Bioprojects, Genome assembly, DNA and RNA sequences, high throughput read archives and taxonomy and databases
Genes	192	Gene loci information, functional genome studies e.g gene expression and similar
Proteins	643	Protein sequences, domains and clusters
Chemicals	11	Metabolites, pathways, bioactivity, and chemical structures

1.2 Motivation

The use of machines, advanced technologies and knowledge has enabled efficient data collection, improved precision and resolution of research targets and hypotheses in modern science and more so in biology. Table 1.1 summarizes the outcome of a search of the term 'big data biology' on the National Center for Biotechnology Information (NCBI) website. This reveals the extent of data deluge characterizing modern biology and is true for other facets of research. A relatively recent review of biology in this context of large data details the rapid incline in cloud computing infrastructural progress and the arising challenges and opportunities that come along with such biological data explosion (Marx, 2013).

1.2.1 Shaping biology through informatics tools

The present work has been influenced by the gap between molecular biology and information science. Previous experience has shown that wet-bench biologists often have little interest in technical content of statistics, mathematics and computer sciences. Similarly mathematicians, computer scientists and statisticians have for a long time had little interest and knowledge of biology including basics notions such as the central dogma. The realization of the indispensable role played by computers in sciences including biology requires a bridge of gap by biologists adopting a flexible attitude towards information sciences. As a biologist, I have been driven by the endeavour to learn and create a niche for information science in biology. I have tried to contribute towards development of integrated platforms facilitating this gap-bridging process. The 'wet-bench' will only make leaps, if computational biology is fully adopted as a tool to shape the next generation of biological research.

1.2.2 Application of data-mining tools in biology

Deriving value from biological data requires algorithmic tools, pipelines, and software frameworks, whose development is often out of the biology scope although it inevitably requires detailed biological knowledge. Examples of recent themes in need of data mining are the so called 'personalized medicine' and the trans-omics integration.

Personalized medicine is the quest towards creating health care solutions of individuals based on genomic fingerprint. This has recently targeted diseases such as cancer and other genome-specific diseases and disorders, some of which scientists believe can be cured by precise

targeting techniques.

Trans-omics integration describes the recent merging of the omics levels by holistic examination of multi-level omics data for example mapping the transcriptome to the metabolome (Li *et al.*, 2015) or mapping the proteome to the metabolome. Bioinformatics glues these levels together and there is room for research involving development of algorithms and software tools that facilitate this process. In this dissertation, we will discuss tools within the trans-omics realm.

1.3 Dissertation outline

This chapter introduced big data biology, major achievements of bioinformatics and the role and relevance of informatics tools in understanding molecular data, objectives and motivation behind this work. Chapter 2 lays the essential foundation for detailed understanding of contribution of this work. Therein, I introduce protein sequence analysis and transcription regulation network mining. Chapter 3 delves into incorporation of residue attributes of protein sequences for analytical application. We introduce numeric scale quantified metrics for protein sequence representation and illustrate how it can be applied for sequence based quantitative analyses. Chapter 4 introduces transcription regulation networks based on pathway modularization of gene expression datasets. We describe an application that merges distint analytical tools such as differential gene expression, network construction and transcription regulation network analysis into a singular framework. Chapter 5 draws the main conclusions and observations of this work. Overall, this manuscript describes the role of integrative computational biology in improving biological objectivity of informatics based analyses of molecular datasets.

Chapter 2

Essential Foundation

Chapter Summary

Computational biology in general has a wide range of applications. This chapter builds the theoretical concepts needed to understand the contribution of this dissertation by discussing the following:

- Molecular Sequence analysis
- Applications of Statistics and data-mining in sequence analysis
- Challenges and opportunities for research in Sequence data-mining
- Transcriptomics technologies
- Applications of data-mining to transcriptome data
- Challenges and opportunites in computational transcriptomics
- Problem and significance

Besides database development, another purpose of computational biology is application of statistical and computational algorithms to understand molecular data. Two such subfields where this is prominent are molecular sequence analysis and transcriptome data analysis.

2.1 Sequence analysis

Research in sequence analysis evidently plays a crucial role in understanding genome function. Molecular sequence analysis refers to the process of digging information from nucleotide or amino acid sequences by using computational (*in silico*) tools. Recognizing novel genes and proteins remains one of the most pressing challenges in genome analysis. Techniques for gene and protein prediction continue to be developed and several tools are available for these tasks.

Some strategies apply *ab initio* statistical prediction whereas others apply homology based methods which depend on inference of similarity in function. Statistical theory for sequence comparison continues to develop for application to both DNA and protein sequences. There however remains many challenges such as low complexity patterns in some eukaryotic genomes. Databases of sequences are publicly and commercially available. A wide range of software tools are also available for sequence analysis but opportunities remain for development of new tools.

2.1.1 Molecular sequence databases

The three most popular public sequence databases are:

- GenBank: hosted at National Center for Biotechnology Information(NCBI) USA
- DNA Data Bank of Japan (DDBJ): hosted at National Institute of Genetics (NIG) Japan
- EMBL: hosted at the European Molecular Biology Laboratory(EMBL), UK

These 3 databases collaborate and form part of the so called 'International Nucleotide Sequence Database Collaboration'. Each of them accumulates new data and annotation information from scientists all over the world. They are frequently updated and are very comprehensive resources. A comprehensive list of molecular databases storing DNA and protein sequences is shown in Table 2.1

Table 2.1: A list of DNA and Protein sequence databases.
(Descriptions adopted from http://sbkb.org/page/show/dna-and-protein-sequence-databases)

Database	Description
AGD	genome/transcriptome database containing gene annotation and high-density oligonucleotide microarray expression data for protein-coding genes from *Ashbya gossypii* and *Saccharomyces cerevisiae*.
BioCyc	each database in the BioCyc collection describes the genome and metabolic pathways of a single organism.
CleanEx	portal which provides access to multiple curated public gene expression data resources
CYGD	comprehensive yeast genome database
Dictybase	full genomics, material, and networking resource for the Dictyostelid community
EchoBase	integrated post-genomic database for Escherichia coli K-12 strain MG1655
EcoGene	database of Escherichia coli Sequence and Function

continued from previous page

Database	Description
euHCVdb	euHCVdb is oriented towards protein sequence, structure and function analyses and structural biology of HCV
EvoTrace	creates an integrated report about the evolutionary propensity of individual residues
FlyBase	genome annotation and phenotype image database for Drosophila melanogastor (fruit fly)
GeneCards	searchable, integrated database of human genes that provides concise genomic related information on all known and predicted human genes
GeneDB	Provides access access to the latest sequence data and annotation/curation of over 40 pathogenic organisms
GeneFarm	genome annotation database for Arabidopsis thaliana (watercress
GenoList	integrated environment for comparative exploration of over 700 microbial genomes
Gramene	comparative genome annotation database for several Grass species
HAMAP	semi-automatic annotation of proteins that are part of well-conserved families or subfamilies
HGNC	unique gene symbols and names to over 33,000 human loci
HInv-DB	curated annotations of human genes
HOGENOM	database of complete genome homologous genes families
KEGG	database resource for understanding high level functions and utilities of the biological system
KNApSAcK Family	A group of repositories for understanding species-species, species-metabolite interactions and other health and medically vital associations
MaizeGDB	community-oriented informatics service to researchers focused on the crop plant and model organism Zea mays
MEROPS	information resource for peptidases (also termed proteases, proteinases and proteolytic enzymes) and the proteins that inhibit them
MGD	integrated, community-driven data resource on mouse genes, genome features, and phenotypes
NMPDR	curated annotations of food-based and sexually-transmitted pathogens
NCBI Nucleotide db	collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.
NCBI RefSeq	comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins
PANTHER	Protein ANalysis THrough Evolutionary Relationships, library of protein families and subfamilies indexed by function
PCCDB	data repository and searchable archive for Circular Dichroism spectra from proteins
PeptideAtlas	multi-organism proteomics data
PeroxiBase	database of manual annotation of peroxidase superfamilies encoding sequences

continued from previous page

Database	Description
Pfam	database that curates protein sequence families, each represented by multiple sequence alignments and hidden Markov models (HMMs).
PhosphoSitePlus	systems biology resource providing comprehensive information and tools for the study of protein post-translational modifications
PlasmoDB	genomic and proteomic data for different species of the parasitic eukaryote Plasmodium, the cause of Malaria.
PptaseDB	prokaryotic protein phosphatase database
PRINTS	compendium of conserved sequence motifs used to characterise a protein family
ProDom	comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases
ProMEX	tryptic peptide fragmentation mass spectra derived from plants
ProSite	protein domains, families and functional sites as well as associated patterns and profiles to identify them
PseudoCAP	comparative analysis of Pseudomonas aeruginosa with other species
RGD	integrated, community-driven data resource on rat genes,genome features, and phenotypes
SGD	comprehensive integrated biological information for budding yeast along with search and analysis tools to explore these data
TAIR	database of genetic and molecular biology data for the model higher plant Arabidopsis thaliana
NCBI Taxonomy	curated classification and nomenclature for all of the organisms in the public sequence databases
TIGR/JCVI	Genome annotation projects from the J. Craig Venter Institute
UniGene	computationally identifies transcripts from the same locus; analyzes expression by tissue, age, and health status; and reports related proteins (protEST) and clone resources
UniProt	UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins.
VectorBase	NIAID Bioinformatics Resource Center for Invertebrate Vectors of Human Pathogens
World-2DPage	known 2-D PAGE database servers, as well as to 2-D PAGE related servers and services.
WormBase	community-driven integrated resource of biology and genome of C. elegans
ZFIN	community-driven integrated resource of biology and genome of D. rerio

2.1.2 Protein sequence bioinformatics

The repositories listed in the Table 2.1 provide access to massive amounts of sequence data. Computational analysis of these datasets is done for purposes including functional annotation, understanding gene regulation components such as conserved transcription factor binding sites and mapping evolutionary patterns of sequence data by phylogenetics etc. DNA and protein sequences are often examined with different objectives. Proteomics has from inception to date changed significantly. The concept in biology encompasses assays of protein samples quantitatively or methodologicaly. In bioinformatics however, proteomics contains a wide array of tools for cataloguing, quantitative assessments and data mining studies. Prediction of secondary and tertiary structures of protein sequences is an extensively studied area of bioinformatics. In this dissertation we are particulary interested in protein sequence analysis. We will introduce a proposed amino acid residue property annotation model for application in quantitative-oriented computational algorithms.

There are many reasons why protein characterization requires more emphasis than DNA sequences. These include larger number of building blocks (20 amino acids rather than 4 nucleotides), lower signal to noise ratio when searching proteins from databases, functional context of protein sequences and the availability of curated data. In computational biology, proteins can be characterized in different contexts for example, sequence representation, function prediction, structure prediction and functionl level interactions amongst other aspects. In machine learning, protein sequence representation is particularly crucial for purposes such as prediction of motifs, domains, structures and interactions.

The first step in analysis of a newly sequenced protein is the search for similar or closely related sequences in databases. UniProt (Universal Protein Resource) is the most common reference protein database. Similarity searches determine statistical significance of database sequences to the query sequences. Information on function importance can be inferred by the extent of similarity of functional domains in the query and subject sequences. Proteins are universally represented as strings of Alphabetical characters. The basic information of a protein therefore comes from its sequence and thus it is difficult to make inferences about a protein from just a single sequence. Related proteins are often aligned together to discover useful domains. At a machine level, this involves use of features like regular expressions to identify conserved patterns. Other methods like hidden Markov models (HMMs) apply probabilistic scoring systems to elucidate patterns in sequences.

2.1.3 Computational applications to protein sequences

Due the functional nature in cells, proteins are studied in many computational perspectives particularly through application of machine learning (ML) algorithms. ML will increasingly play a crucial role in understanding the role of proteins. Supervised forms of ML involve exploration of patterns and functional roles of primary amino acid sequences. This approach has been attractive especially in protein structure prediction, residue-residue interaction mapping and protein classification. ML infers patterns from data and gives an output that is not intrinsically programmed into the algorithm. This makes it attractive for application in sequence data analysis. More clearly, given model parameters an algorithm ML takes a sequence s and gives an output prediction y. A simplified illustration of an ML system for protein sequences is shown in the figure below.

Predicting protein structure

Experimental techniques such as X-ray crystallography and nuclear magnetic resonance attempt to elucidate protein structure. The earliest recorded structure prediction was in the works of (Sanger and Thompson, 1953; Perutz *et al.*, 1960; Kendrew *et al.*, 1960; Anfinsen, 1972). Despite being considered as gold standard measures of structure determination, empirical methods are limited by cost, and throughput levels. Computational predictions thus offer viable and complementary alternatives. Many studies attempt to predict protein structure (1D-4D) from sequence data using ML techniques. These methods cover secondary structure prediction, solvent property prediction, binding and functional site prediction and transmembrane helix prediction (Rost and Sander, 1993a,b; Bryson *et al.*, 2007; Tress *et al.*, 2007; Cheng, 2007). A detailed review of computational structure prediction from protein sequences has been documented by Cheng *et al.* (2008).

Predicting protein function

An essential goal of proteomics is to predict protein function based on sequences. ML approaches are more sophisticated and are independent of sequence alignment workflows like ortholog prediction from alignments. Examples of ML tools that have been applied in function prediction include support vector machines (SVM)(Cai *et al.*, 2003), naive Bayes's(NB) classifiers (Borro *et al.*, 2006), neural networks (NN) Jensen *et al.* (2002) and ensemble classifiers (Chen and Liu, 2005; Guan *et al.*, 2008) have been used for function prediction. Cai *et al.* (2003), applied SVM

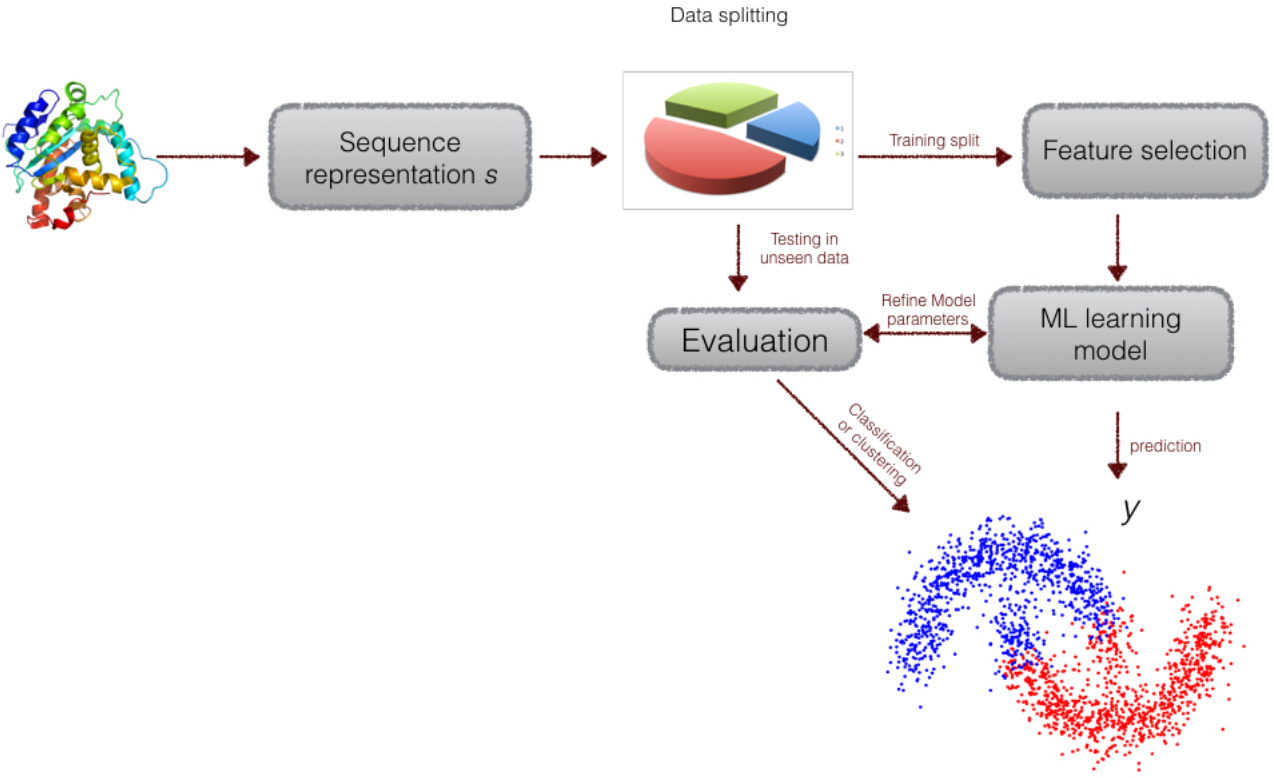

Figure 2.1: *Typical flow of protein sequence ML where input sequence representations are defined and data split into training, testing and evaluation sets. ML models are refined through feature selection*

to predict protein function with an accuracy of 84-96% and concluded that biochemical and physical properties particularly secondary structure, solvent accessibility, secondary structure, polarity and van der Waals interactions amongst other features played critical roles in function prediction.

Predicting protein interactions

Accurate elucidation of protein-protein interactions is a key step in understanding function. ML also finds applications in this context. (Zubek *et al.*, 2015) used protein structures from the PDB database by systematic characterization using random forests (RF) (Breiman, 2001) and support vector machines (Vapnik and Vapnik, 1998). A study has attempted to summarize some of the evolutionary metrics such as homology conservation proportion and residue property distance which can be applied in prediction of interactions Aumentado-Armstrong *et al.* (2015). A recent study (Hamp and Rost, 2015) evaluated some of the problems associated with ML techniques using sequences for interaction prediction. These challenges include low redundancy reduction, non-standard cross validation and the influence of negative interactions. A separate study (Walsh *et al.*, 2015) cites a few other problems such as low data quality and represantativeness, less sequence diversity and overfitting and underfitting challenges.

2.1.4 Proteomics challenge addressed in this dissertation

An underlying mechanism at a lower level of ML and other computational biology applications, is the nature of protein sequence representation. The link between statistical analysis and data mining of protein sequences face a critical gap in the nature of protein representation i.e., the use of alphabetic letters to define sequence elements. Letters lack an underlying natural quantitative metric for ease of objective comparison and amenability to computational procedures. This can be exemplified by the, amino acid Histidine (H) which is known to be similar in properties to Lysine (K) and Arginine (R) in terms of polarity. Alphabetical coding cannot explicitly capture this information. Application of letters codes in sequence analyses omit critical information about biochemical and physical properties of amino acids and the proteins indirectly. There is need for alternative forms of coding.

Some very early studies in the 60s and 70s applied objective metric representation in varying forms e.g by contextual annotation of amino acid variation as *ad hoc* indices (Sneath, 1966). This is however prone to underestimation of variation and thus may not explain the source of

residue diversity. In addition, it may not represent associations among amino acids.

In recent studies, information theory indices such as entropy and mutual information have been used as proxy indices of variation among amino acids (Herzel and Große, 1995; Roman-Roldan *et al.*, 1996). Building on this, other studies have performed mutual information based analysis of variance in statistical modeling of multiple sequence alignments as multivariate datasets whose variation can be examined by dimension reduction algorithms such as principal component analysis (PCA). This is a significant improvement but it still does not consider presence of negative associations such as those driven by mutational changes in certain sites. It also does not accurately uncover whether a non-monotonic association is causal or correlational.

In this dissertation we build on work of others along the lines of multivariate modeling of sequence alignments to describe a novel solution to the protein sequence representation problem. We take advantage of data accumulated over the years in the amino acid index database (Kawashima and Kanehisa, 2000) to build a subset of biochemical and physical properties (BPP) that can be directly encoded to sequence data in ML applications.

Classifying terpenoids enzymes in KNApSAcK database

Closely tied to function, metabolic pathways in plants are driven by proteins (enzymes). Among these metabolite categories are terpenoids whose synthesis is driven by terpenoid synthases. At our laboratory, we have developed the KNApSAcK family of databases; one of which is the Motorcycle DB (Ikeda *et al.*, 2013) which stores enzyme reactions.

Terpenes are the largest group of plant natural products, with a variety of core chemical structures comprising at least 30,000 compounds (Connolly and Hill, 1991). Terpene diversity is caused by the large number of different terpene synthases used in the first step of terpene synthesis. Some terpene synthases produce multiple products (Degenhardt *et al.*, 2009). Terpene synthases are generally classified according to the number of carbons in their substrates.

As an example application of our proposed representation model, we explored terpenoid synthase sequences through a multivariate statistical procedure. We address this issue comprehensively in chapter 3.

2.2 Transcriptome technologies

Detailed history and scope of transcriptomics has been presented by Morozova *et al.* (2009). One of the most fascinating facts of molecular biology is that all cell types are made up of the same genetic content but their phenotypes and function are different despite similarity in genetic content make up. Such diversity is due to the different sets of active genes in different cells leading to different roles and functions. The process of genetic information transcription from DNA to RNA is responsible for phenotypic differences. Studies continue to fully comprehend the mechanics behind this phenomena.

2.2.1 Evolution of transcriptome quantification technologies

Technology that has drastically transformed precision and scope of transcriptome research.

Nothern blotting

The earliest technology for transcriptome studies was based on candidate gene assays. Nothern blot technology (Alwine *et al.*, 1977) is a low throughput technique that implemented radioactive tracing and use of large amounts of RNA. This requirement therefore limited the application of Nothern blots to detection of few known transcripts at a time.

Reverse Transcription Quantitative PCR (RT-qPCR)

The advent of RT-qPCR (Becker-Andre and Hahlbrock, 1989) increased detection of transcripts and throughput at the same time reducing the quantity of RNA required. Old as this technology may be, It remain small scale considering its restriction of magnitude to assays of hundreds of genes at a time.

Microarray technology

Candidate gene assays were soon supplanted by microarray technology (Schena *et al.*, 1995) which allowed simultenous characterization of thousands of transcripts using probes attached to glass surfaces. This technology enabled systematic study of expression markers and the scope of use expanded to include features such as single nucleotide polymorphisms (SNPs). This cheap technology is the most commonly used tool for transcriptome quantification. Despite the significant gain from previous methods, in terms of throughput, efficiency and cost, microarrays

still do not answer fundamental questions of transcriptomics for example detection of novel transcripts. In addition to this, it is remains an indirect method of quantification based on hybridization intensity and does reflect the actual counts or frequency of the target probe. This is besides technical limitations such as reduced hybridization.

Sequencing-based transcriptome quantification

Driven by ability to detect novel transcripts, DNA sequencing based approaches were invented as alternatives. Besides novelty of transcripts, these methods were able to detect direct abundance unlike indirect inference characteristic of Microarrays. Earlier among the sequence based technologies was the cDNA sequencing which proved to be expensive. Expression sequence tag (EST) sequencing was later developed as the cheaper version but remained overly expensive nonetheless. Subsequent technologies included Serial Analysis of Gene Expression (SAGE) (Velculescu *et al.*, 1995) which enabled comparison of multiple samples. The procedure however remained intensive of labour due to the necessity to clone. Gradual progress gave way to the so called tag-based 'next generation sequencing' platforms. These tag-based sequencing technologies reduced the cost drastically.

Next generation sequencing (NGS) is popularly known as an increasingly powerful platform for application in a broad range of studies. Significant challenges still remain but, translation of these technologies into useful tools of examining transcriptomes cannot be understated. Currently available NGS machines include illumina, 454/Roche, Applied Biosystems and Helicos BioSciences. Dideoxynucleotide chain termination technique popularly known as the Sanger method dominated sequencing until recently. Several variants of NGS for transcriptome sequencing exist. These include DeepSuperSAGE (Matsumura *et al.*, 2012), cap analysis of gene expression (CAGE) (Shiraki *et al.*, 2003), deepCAGE (de Hoon and Hayashizaki, 2008) RNA-seq (Mortazavi *et al.*, 2008) and Tag-seq (Morrissy *et al.*, 2009).

Table 2.2: The expected output of modern devices such as Roche 454, Illumina, ABi SOLiD and Pacific Biosciences

	FLX	Titanium	GS.Junior	I	II	IIx	X.2000.	X1	X2	X3	X4
Reads(Millions)	0.50	1.25	0.10	28	100	250	2000	40	115	320	1400
Read length(fragment)	200	400	400	35	50	100	125	25	35	50	75
Read length(paired end)	200	400	400	2x35	2x50	2x100	2x125	2x25	2x35	2x50	2x75

Cap Analysis of Gene Expression (CAGE) is a notable example of NGS based sequencing for transcriptome data. Originally developed at RIKEN Japan by Shiraki *et al.* (2003), it has

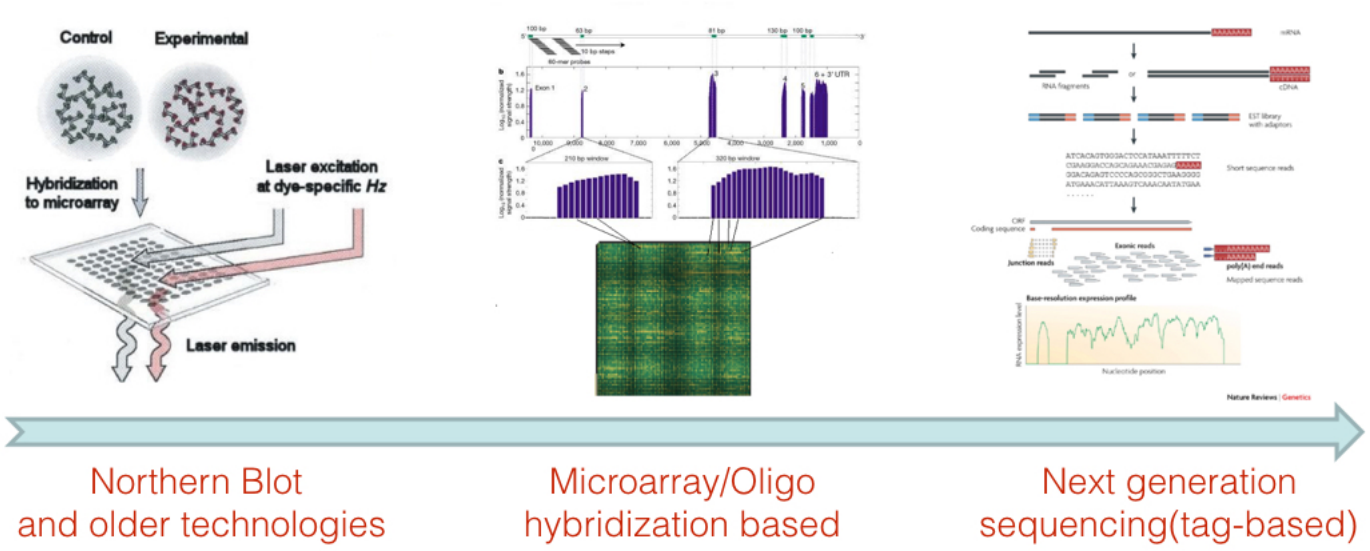

Figure 2.2: Evolution of transcriptome quantification

been primarily used for mapping transcription start sites (TSS). CAGE is especially promising in the study of isomorphs of mRNA resulting from differential use of TSS. CAGE has been the main assay method in so called Functional Annotation of Mammalian cDNA (FANTOM) (Carninci *et al.*, 2005) project which has been through 5 phases so far. FANTOM 6 phase is due to begin in 2016. Both CAGE and SAGE produce fragments of RNA as tags but the concepts are different. CAGE describes the genome positions where RNA transcription begins.

2.2.2 Computational requirements for transcriptome datasets

Computational infrastructure required for data analysis is extensive. With the data deluge generated by NGS technologies, it is imperative to apply informatics and computational facilities to meet arising challenges. Technologies evolve over time and throughput seems to increase exponentially. Instrument manufacturers such Roche, Illumina and PacBiosciences provide elementary software. Data from the instruments is often captured from either a small server or a workstation attached to the machine. Some other manufacturers provide additional tools for analysis. As data accumulate however, the need for structured cataloguing, quality control and comprehensive mining becomes necessary. Bionformatics scales up the vendor provided tools to enhance value extraction. Sequencing itself is the entry point of computational biology. Primary raw sequence reads from the machine require processing to separate tags from linkers and adapter sequences. Tag extraction is the computational process of retreiving tags from the 'dirty' tags used during library preparation. Some sequencing technologies eg the Roche 454 have sequencing errors in filtering read quality and therefore primary raw reads may often be contaminated at a rate of 0.5-1.0 % (Arner, 2010). Tag extraction from contaminated sequences is a string matching computational problem that requires the use of regular expressions. Annotation of tags to known sequence transcripts is the subsequent computational step. This step in known as tag mapping and is the foundation for all downstream data analyses. Several criteria are applied depending on the sequence of interest. Once mapping is complete, reads are mapped to a genomic location. This can be visualized computationally using genome browsers. Tertiary analyses such as differential expression, gene enrichment and biomarker extraction using various computational workflows enable value extraction from gene expression data. Gene expression molecular fingerprints determined through the use of ensemble computational tools enable comparative evaluation between conditions, cell types, time-series or other researcher dependent variables.

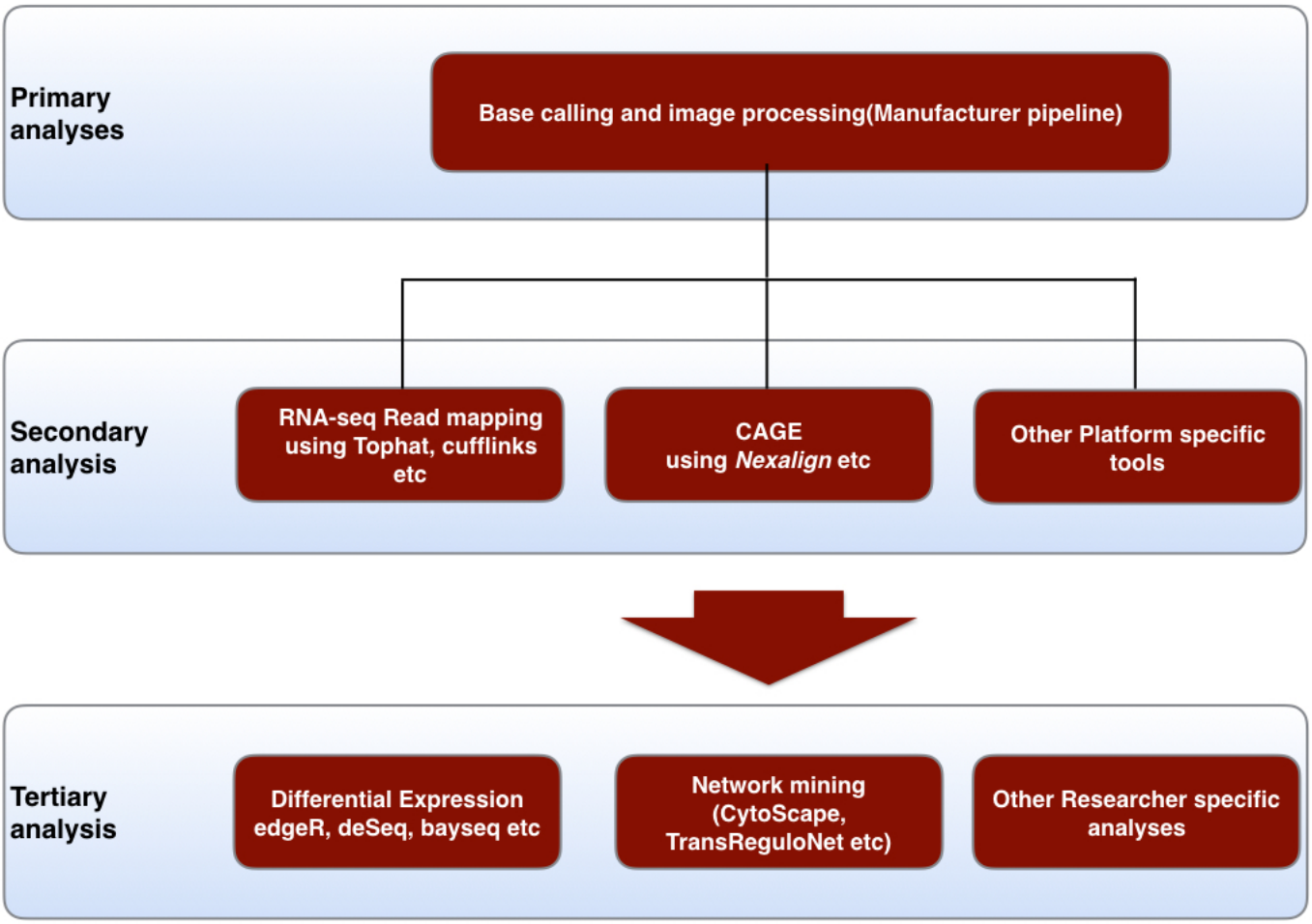

Figure 2.3: Stages of computation in transcriptomics

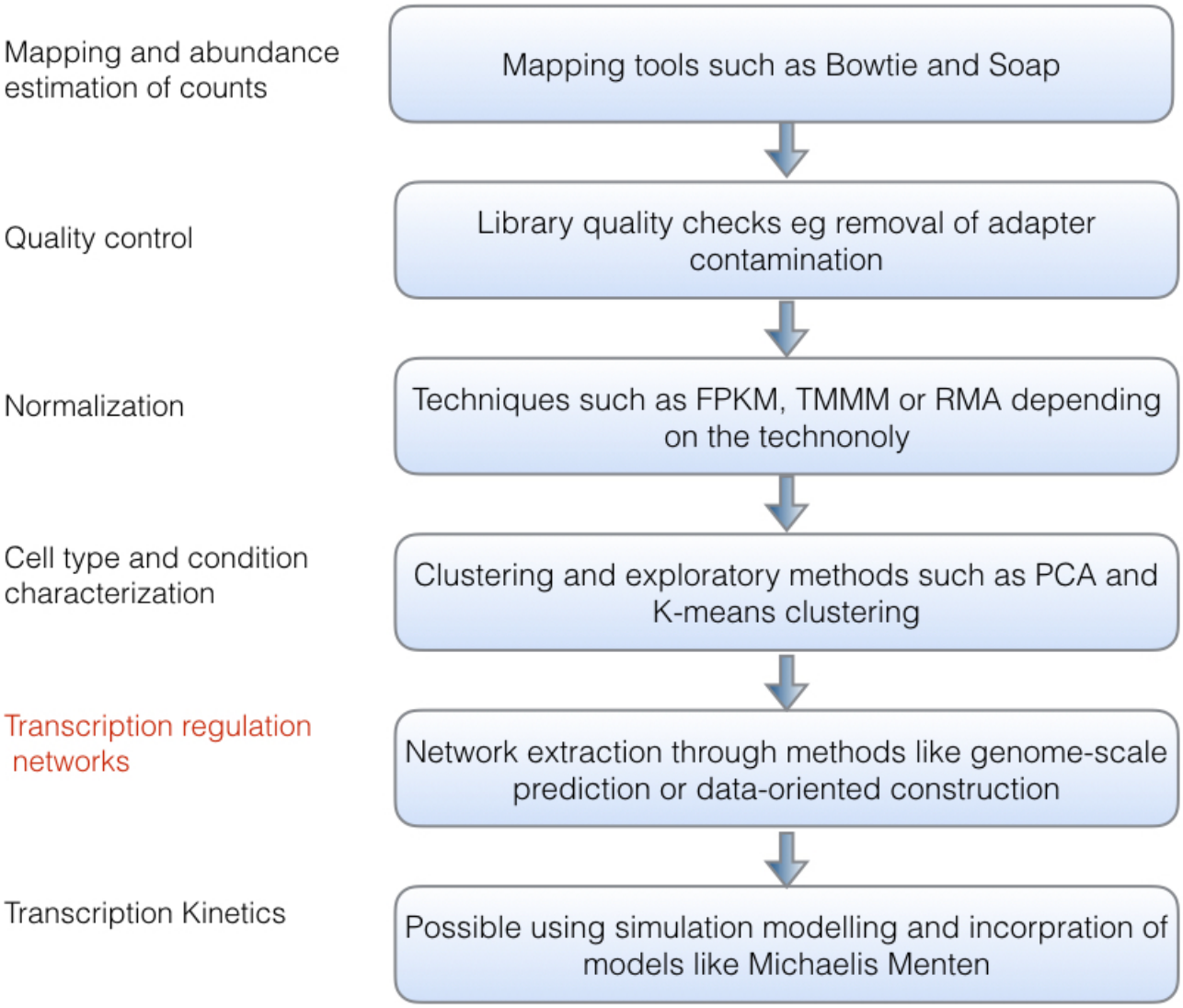

Figure 2.4: Computational processes required for transcriptome analysis

Figure 2.5: *Schematic view of the gene regulation region (promoter). The black arrows represent the transcription start sites. Genes may have isoforms often attributed to differences in positions of TSS. The binding sites near the TSS are known as proximal promoters and those further to the left are know as distal promoters*

2.2.3 Transcriptome informatics methods for gene regulation analysis

Transcription regulation processes play crucial roles in determining specificity of the underlying drivers behind the phenotypic variation.

In this work, we shall introduce our proposed method for data mining of gene expression data by integrated analysis. Appropriate packages and toolkits that unify heterogeneity of expression profiles exist in platforms such as Galaxy (Giardine *et al.*, 2005), R-Bioconductor (Gentleman *et al.*, 2004) and BioGPS (Wu *et al.*, 2009). In order to understand the need for computational tools in elucidating gene regulation, it is necessary to first understand the architecture of typical gene promoters.

Molecular events preceeding transcription fundamentally start with a transcription factor (a protein), binding a specific recognition site along the upstream (5') region of a gene. This happens in conjunction with cofactors that link protein-DNA interactions. microRNA is the other element of transcription regulation. It modulates binding events resulting in a functional or negative effect (Filipowicz *et al.*, 2008). Under the common linear model, distal and proximal regulation elements constitute the *cis*-regulation, a feature that is exploited in computational biology. Genomescale *in silico* prediction of the *cis*-regulation elements is increasing in resolution and precision. Motif matching by use of position weight matrices is by far the most common approach applied in detection of binding sites. This is efficient for only small sets of genes but with gradual increase, computational complexity increases substantially. A slightly faster alternative of TFBS prediction is the use of information theory principles (Hertz and Stormo, 1999). Another fast algorithm based on probability framework was invented by van Nimwegen (2007) and applied in studies on chromatin boundaries. Sequence based motif detection leverage on statistical tests such as overrepresentation analysis (Bajic *et al.*, 2003) for identification of functional motifs.

Toolkits for Gene Regulation analysis

Many computational tools are available but the underlying assumption for each of them is different, making it the researcher's prerogative to determine appropriateness depending on research goals in consideration of limitations of specific tools. Both Microarray and NGS data can be analyzed using open source software packages available in R-Bioconductor (Gentleman *et al.*, 2004). Model based methods like the one presented by Xie *et al.* (2005) provide network

estimates using coregulated genes. A matrix-based toolkit like Pscan (Zambelli *et al.*, 2009) employs a statistically robust strategy to infer proximal regulatory elements. Other tools for *de novo* motif discovery include MEME (Bailey *et al.*, 1994) and WEEDER (Pavesi *et al.*, 2004) which uses ChIP-seq data to extract peaks. Other applications for trancriptome informatics focus on annotation and enrichment analysis. Web platforms eg DAVID (Huang *et al.*, 2008), GSEA (Subramanian *et al.*, 2005) and GAGE (Luo *et al.*, 2009) are example tools that are applied in gene enrichment studies.

2.2.4 Regulation network analysis in context of this work

Despite the presence of toolkits and other frameworks applied in the general area of analyzing gene regulation networks, alot remains to be done to improve the overall output and intepretation. Example fundamental challenges include the lack of a *de facto* standard approach for construction of transcription networks regulation (TRNs). Static genome scale TRNs can be constructed on the basis of motif mapping within promoter regions of all known genes but this larger network is not directly useful for inference because transcription is condition-dependent and cell-type specific. Static networks do not explicitly capture conditional variation in transcription regulation dynamics. On the other hand this limitation has been countered by most of the popular TRN inference toolkits such as Conservative Causal Core Networks (C3NET) (Altay and Emmert-Streib, 2010), Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (Margolin *et al.*, 2006), relevance networks (RN) (Butte *et al.*, 2000) and Maximum relevance/minimum redundancy Network (MRNET) (Meyer *et al.*, 2007) which apply data centric modeling of TRNs reliant on evaluation of causal or correlational tendencies of genes within expression profiles. These tools can be applied to crossplatform data and can handle Microarray as well as NGS transcriptome profiles. According to our evaluation challenges in TRN analysis arise at three levels:

1. Data handling
2. Method/Software tools integration
3. Biological intepretability.

Public repositories including Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002), Array Express (Brazma *et al.*, 2003) and FANTOM database (Bono *et al.*, 2002) store data of diverse nature. GEO and ArrayExpress for instance, accumulate cross-platform data including Microarray,

RNA-seq and other NGS variants of transcriptome quantification. FANTOM database is biased towards CAGE-derived expression profiles. The nature of diversity in these databases require expert bioinformatics preprocessing skills which often lack in wet-bench laboratories. This leads to accumulation of data with little reuse. Besides this, there is also a significant lack of standard structure for the datasets stored in these repositories thereby limiting the range of combined analysis on datasets of similar condition or cell types.

There is also a need for method and software integration to improve inference deduction from transcription regulation networks. Currently, most software focus on separate aspects such as differential expression (DE), gene set enrichment (GSE), data-oriented network construction and visualization. Tools handle these aspects separately for example edgeR (Robinson *et al.*, 2010), which is one of the open source Bioconductor package, specializes in extraction of significantly expressed genes from an expression dataset. Geneset Enrichment Analysis tool (GSEA) (Subramanian *et al.*, 2005) and Generally Applicable Gene Enrichment (Luo *et al.*, 2009) software focus on gene set analysis whereas Cytoscape and most of some of its plugins (Shannon *et al.*, 2003) focus on transcription regulation network visualization. This abstraction can be complimented by an integrative combination of tools into a singular platform that can be easily used by biologists to draw inferences from TRNs.

A more compound problem with TRN bioinformatics is the biological intepretability of analyses. Most informatics tools are 'black box' in nature, requiring a user to plug in data and obtain results without necessarily knowing the mechanics of the application. The statistical philosophy 'garbage in garbage out', makes it is imperative for users to ensure that data meet standards of quality for meaningful output to be obtained. There is a gap between experimental design and downstream analyses due lack of pre-experiment consideration of both the wet-bench objectives and the requirement of downstream software tools. To this extent, making inferences from analyses often requires background skills in both areas. A particular focus of this work is the inference of key regulating biomarkers particularly transcription factors. During collection of gene expression profiles, wet-bench labs may not focus on separating transcription factors from target genes during assays. At the same time, not much is known about many transcription factors in terms of binding site preferences and other activity related properties. Out of the over 1500 transcription factors known, less than 300 of them have published information about their preferred motifs. Inferring importance of these markers is therefore limited to the well studied markers. Similar limitations exist for other forms of biological interpretation from transcription

regulation networks.

TransReguloNet: An integrated platform for transcription regulation network studies

In Chapter 4 of this dissertation, we shall present an application which touches on the above limitations namely data handling, software/tool integration and biological intepretability. We shall describe data dependent filtration of static transcription regulation networks as an alternative to *de novo* data independent network construction strategies. Data currently loaded in the TransReguloNet server is retrieved from publicly available databases and expert-curated thus cutting the need for user preprocessing. Downstream analyses combines gene differential analysis and network construction within the pipeline. Biological intepretability is implemented through functional pathway abstraction of gene expression profiles. The application targets transcription factor driven regulation.

Chapter 3

Protein Sequence Representation: Integrating residue attributes

Chapter Summary

The sequence representation problem introduced in the last chapter is especially significant for protein sequences. This chapter proposes a metric representation approach. In summary, the following is covered:

- Nature and progress of sequence representation in bioinformatics
- Amino acid index and its role
- Selecting important indices by random forest
- Example application in data from KNApSAcK motorcycle DB

Chapter Abstract

Progress in the 'omics' fields including genomics, transcriptomics, proteomics and metabolomics has engendered a need for innovative analytical techniques to derive meaningful information from the increasing molecular data in public repositories. KNApSAcK motorcycle DB is a popular database for enzymes related to secondary metabolic pathways in plants. One of the challenges in analyses of protein sequence data in such repositories is the standard notation of sequences as strings of alphabetical characters. This has created lack of an intepretable and comprehensive metric of representation that eases amenability to computation while conserving biological relevance. In view of this requirement, we applied novel integration of selected biochemical and physical attributes (quantified in numerical scale) of amino acids derived from the amino acid index, to examine diversity of peptide sequences of terpenoid synthases accumulated in KNApSAcK motorcyle DB. We initially generated a reduced amino acid index table. This is a set of biochemical and physical properties obtained by random forest feature selection of important indices from the amino acid index. Principal component analysis was then applied for characterization of enzymes involved in synthesis of terpenoids. The variance explained was increased by incorporation of residue attributes for analyses.

3.1 Background

Biology and other modern sciences have become data intensive and infact data-driven biology is now a fully-fledged domain of specialization among the life sciences. Among these accruing data are the protein sequences whose usefulness in the eukaryotic cells vary from structural importance at a lower scale to clinical relevance at a phenotypic level. One of the main drivers of protein sequence research is the Munich Information center for Protein Sequences (MIPS), which curates and mantains various databases of protein sequences allowing for proteomics oriented bioinformatics. Two interesting databases MIPS curate and mantain are the Comprehensive Resource for Mammalian Protein complexes (CORUM) (Ruepp *et al.*, 2009) and the Similarity Matrix of Proteins (SIMAP) (Rattei *et al.*, 2010). CORUM utilizes sequence information to infer protein complexes and allow users to trace cellular localization and function of the complexes. SIMAP on the other hand uses protein sequence data to draw statistical similarities between protein sequences using amino acid residue contents. Such resources enable data-driven analyses of sequence data and are expected to significantly refine our knowledge of phenomena related

to protein sequence composition.

Despite the availability of a wide range of database resources for proteomics, data heterogeneity still pose significant challenges due to the amorphous nature of sequence data and the minimal trans-level interaction. Various databases have continually been developed to allow systematization and integration (Stein, 2003). Alot however still needs to be done to characterize these compilations into meaningful information especially by linking proteomics with other levels of 'omics' sciences such as metabolomics. KNApSAcK database describes species-metabolite relationships, and within the KNApSAcK family, we have developed an enzyme-reaction database called KNApSAcK motorcycle DB containing curated experimental evidence of secondary plant metabolite reactions and the corresponding enzyme peptide sequences. This database describes the relationships between species and their metabolites and as linked by the proteome layer and is therefore essential for both proteomics and metabolomics research due to its systematic curation of enormous numbers of protein sequences involved in various metabolic pathways. This is especially so because of the often uncharacterized structural information of both the proteins and metabolites.

Increasingly, the need to analyse data in such repositories has made advanced mathematical and statistical tools a mainstay of bioinformatics in recent years moreso in sequence-based analyses (He and Petoukhov, 2011). The use of mathematical principles in sequence analysis is not new. Infact in the early 90s, Steipe *et al.* (1994) applied mathematical definitions of evolutionary models based on amino acid residue frequency distribution in immunoglobins to develop a concept they defined as 'statistical free energy' to estimate mutational pressures on functional domains. As discussed in Chapter 2, statistics, mathematics and computational algorithms play a role in three aspects of protein bionformatics. These are; predicting interactions (complexes), prediction of structure and prediction of function. The work of Steipe *et al.* (1994) is an early example of function prediction using mathematical models. Given the indispensable role of computationally oriented tools in proteomics, low level challenges still exist and sequence representation is particularly noteworthy.

Computational models of representation are increasingly applied for the storage of sequences of amino acid residues and for their evaluation. For this purpose, a one-letter code has been adopted as the "gold standard" representation for labeling of individual amino-acid residues. In view of the increasing number of different notations and the representation problems for protein sequences the international union of pure and applied chemistry and international union of biochemistry (IUPAC-IUB) commission on Biochemical Nomeclature undertook the task

Table 3.1: IUPAC-IUB amino acid letter code

	Amino acid	Three Letter code	One Letter code
1	Alanine	Ala	A
2	Arginine	Arg	R
3	Asparagine	Asn	N
4	Aspartic acid	Asp	D
5	Asparagine or Aspartic acid	Asx	B
6	Cysteine	Cys	C
7	Glutamine	Gln	Q
8	Glutamic acid	Glu	E
9	Glutamine or Glutamic acid	Glx	Z
10	Glycine	Gly	G
11	Histidine	His	H
12	Isoleucine	Ile	I
13	Leucine	Leu	L
14	Lysine	Lys	K
15	Methionine	Met	M
16	Phenylalanine	Phe	F
17	proline	Pro	P
18	Serine	Ser	S
19	Threonine	Thr	T
20	Tryptophan	Trp	W
21	Tyrosine	Tyr	Y
22	Valine	Val	V

of drafting a single notation for one-letter symbols. In 1968, the IUPAC IUB Commission on Biochemical Nomenclature (CBN) set definitive rules for 'letter notation' of amino acid sequences (IUPAC-IUB, 1971). They defined the letter code annotation as in the Table 3.1

For molecular sequence data especially proteins, this standard alphabetical notation of sequence information may not explicitly capture aspects such as their biochemical and physico-chemical properties (BPP) such as those in Table 3.2 and may to some extent, limit tractability to mathematical analyses. Ideally, computational analyses of the often heterogenous datasets require theoretical representations in forms suitable for various data processing models. This formal representation has been defined as sequence feature-coding (Kong *et al.*, 2007). There has been no standard method for direct encoding of quantifiable protein sequences BPP hitherto. A key research question has thus been how to quantitatively characterize such data for computation whilst considering these aspects and other sequence metadata (Vendruscolo and Tartaglia, 2008). In the present study, we introduce a BPP subset for encoding amino acid residue properties into protein sequences during analyses. We found that this increases the flexibility of computational analyses focusing on facets of biochemical, physical and evolutionary attributes of sequence data. Here, integration of BPP information is employed in examination of diversity in enzymes related to secondary metabolite pathways, specifically those involved in terpenoid synthesis.

Researchers have proposed schemes to ensure amenability of sequences to computation, but it remains difficult to achieve computational objectivity while mantaining biological intepretability. Bit representation such as 8-bit, 5-bit and 3-bit binary feature coding of amino acids in peptide sequences acids have been applied in studies such as in the work of Coghlan *et al.* (2001). White *et al.* also used 20-bit transformation for neural network application for translation of proteins (White and Seffens, 1998). A limitation of binary feature-coding is the minimal biological information with respect to amino acid diversity. This is because bit-coding does not account for relative similarities or differences between amino acids and neither is it flexible to integration of BPP information. In addition, binary notation of highly conserved protein sets may also pose numerical difficulties to probability-based models.

Information theory has also been exploited as an alternative, where mutual information and entropy are estimated by the shannon-weiner index (Henikoff and Henikoff, 1992; Weiss *et al.*, 2000). This way, distance and variation between amino acid units is estimable and therefore is an improvement over the binary coding method. It however does not directly represent characteristic attributes such as polarity, molecular size and residue features.

Table 3.2: Classification of amino acids based on physical properties

Hydrophobic amino acids	Uncharged Polar amino acids	Acidic amino acids	Basic amino acids
Ala	Asn	Asp	Arg
Ile	Cys	Glu	His
Leu	Gly		Lys
Met	Gln		
Phe	Ser		
Pro	Thr		
Trp			
Val			

More recently, the amino acid index (AAindex) database has accumulated published data of amino acid properties (Kawashima and Kanehisa, 2000). Each index has a set of 20 numerical values of a BPP quantified and published in research literature. Currently, AAindex contains 544 indices describing quantifiable amino acids residue properties. This provides a foundation for feature coding proteins by assigning 'scale-measured' attributes of amino acids.

AAindex in its entirety (raw form) is not merited for feature coding since it is highly redundant and has some missing values. Atchley and colleagues proposed index reduction using multivariate factor analysis reducing it to five compressed factors (Atchley *et al.*, 2005). This methodology is a useful solution for AAindex-based metrification of amino acid residues but factor analysis (FA) reduction complicates biological intepretability in downstream analyses. This is because FA just like principal component analysis (PCA) assume an underlying linear independence of variables whose coefficients, also called 'factors', as a proxy intepretation of AAindex variables (Krzanowski, 2000). This means that 'factors' derived are pseudo-variables of actual original properties and in a way adds to complexity of biological intepretation in subsequent downstream steps of sequence analyses.

In light of these challenges in systematization of protein sequences, the present work applies a slightly different variable selection criteria from the AAindex for the purpose of encoding BPP information into sequence data. This was achieved by use of random forest (RF) algorithm (Breiman, 2001) to reduce redundance and to maximize amino acid metadata captured in the AAindex. Eight BPP indices describing variability of amino acids were selected based on our experimental results. The derived reduced AAindex (rAAindex) is a subset of the original AAindex after elimination of redundancies. We further integrated the rAAindex in characterization of protein sequence diversity in the KNApSAcK motorcycle DB. The enzymes characterised are involved in secondary metabolic pathways of terpenoids and include, monoterpenoid

synthases, diterpenoid synthases, triterpenoid synthases and sesquiterpenoid synthases.

3.2 Materials and methods

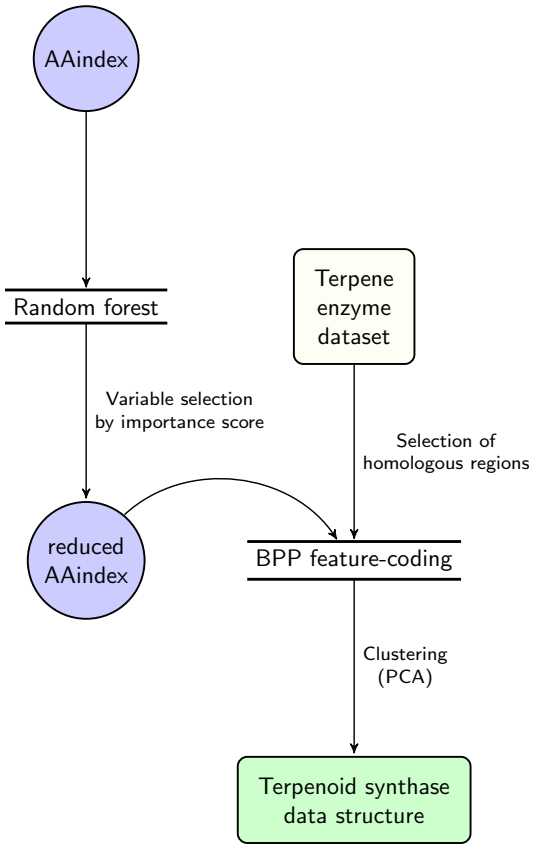

Figure 3.1: *Depicts a generalized procedure of the analysis in this chapter. The starting point is the amino acid index database which is reduced to a small subset based on variable importance scores derived from random forest algorithm. The reduced set (rAAindex) is used to encode biochemical and physical properties into protein sequences for examination of the data structure of the sub-families terpene synthases.*

3.2.1 Amino acid index and random forest selection of biochemical and physical properties

The nature of protein structure and functional specificity can be attributed to the combination of the 20 different amino acids residues coded for in the genetic code. These building blocks of proteins have unique features including shape, size, and chemical properties among other characteristics. Experiments in computational biology research is continually being performed to understand biochemical and physical properties of each residue. For the 20 amino acids, each of these attributes can be quantified in a scale thus yielding a measurable index that describes

the features of protein building blocks in numerically intepretable form. The amino acid index (AAindex) database is a well known curated set of numerical indices describing biochemical and physical attributes of the 20 amino acids (Kawashima and Kanehisa, 2000). The AAindex is publicly available and can be used to generate a more robust sequence representation approach. It provides a plausible starting point for intepreting peptide BPPs numerically through its 'building blocks'. We selected a set of important indices that broadly characterize amino acid BPP variation. The RF algorithm (Breiman, 2001) was used for index selection.

We denote by X the set of amino acid indices as the explanatory variables. N denotes the set of amino acids (AA). We determined that the categorical predictor variable that best defines the AA population is its qualitative attribute in aqueous solutions. This implies that every amino acid is described as either hydrophobic or hydrophilic. We therefore denote Y as describing hydrophobic or hydrophilic properties of an amino acid. The i^{th} amino acid; n_i, is thus described by a vector $(x_1, ..., x_m, y_i)$.

Raw AAindex is highly redundant and multi-collinear. We initially processed it by removal of indices that had missing values for any amino acid. Redundant indices were eliminated by backward elimination of variables whose correlation co-efficient was above a threshold of 0.85. 283 indices were retained for RF variable selection.

Random forest (RF) (Breiman, 2001) is a popular algorithm for feature selection in statistics and bioinformatics for two reasons:

- It is a powerful classification and regression tree (CART) tool that generates ensembles of decision trees. RF and other decision tree-based classifiers are non-parametric. They do not assume underlying structure in datasets and are therefore useful for classification and regression modeling of complex biological data.

- RF implements a mechanism of calculating variable importance scores (VIM) by permutation testing. These measures are useful in feature selection and provides an advantage which we explore in this work.

Besides these two advantages, further application to biological research has been documented by Boulesteix *et al.* (2012). Detailed mechanism of the RF algorithm is described in the work of Breiman (2001) and Svetnik *et al.* (2003) although a generalized outlook of its concepts is illustrated in **figure 3.2**. RF was implemented for selection of a reduced AAindex (rAAindex) consisting of indices describing BPPs explaining the sufficient variation of amino acids.

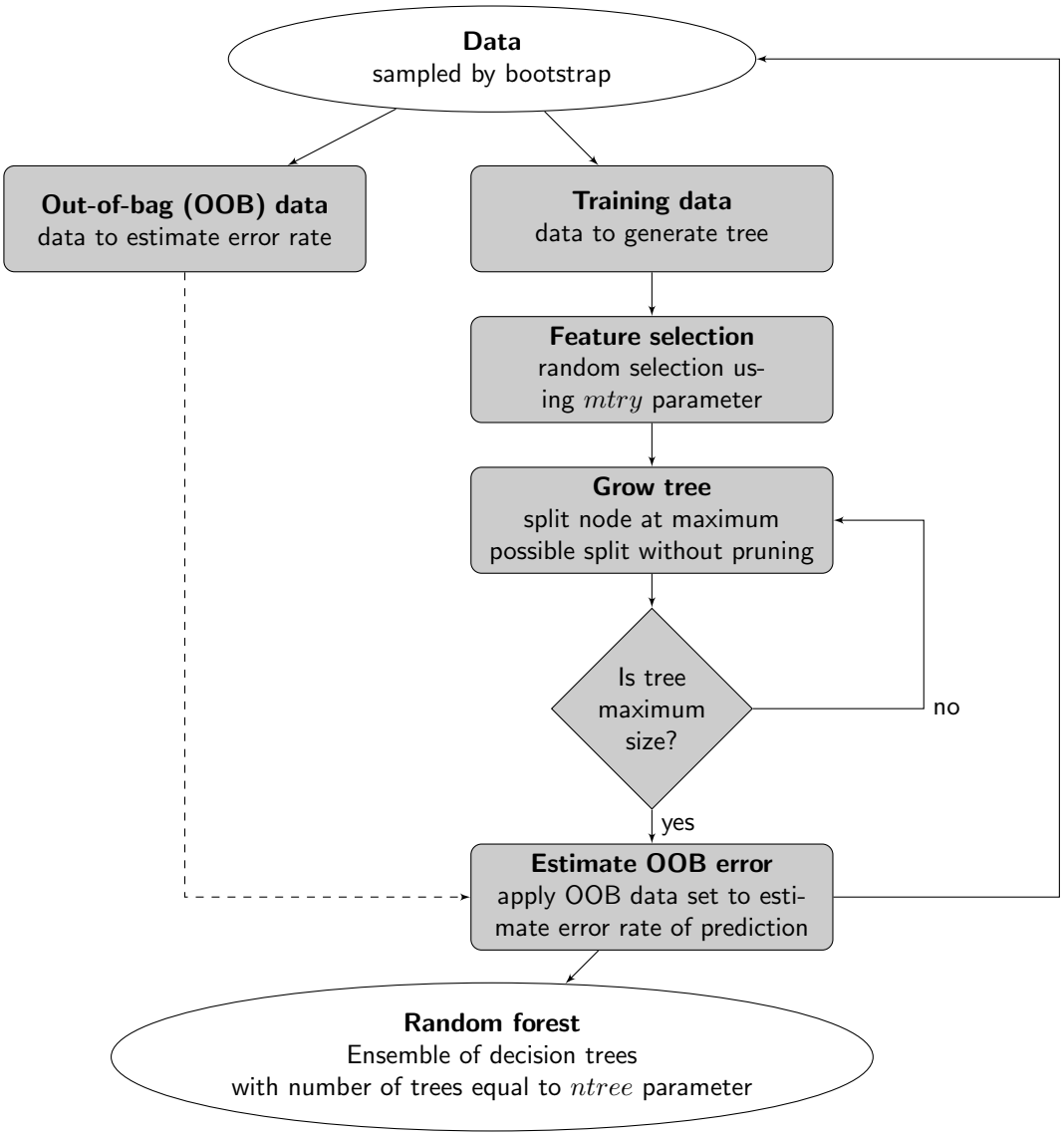

Figure 3.2: *Mechanism of the random forest (RF) algorithm starting from the data selection by bootstrapping upto variable importance calculation. The amino acid index data containing physico-chemical metrics of amino acids was subjected to RF for index selection*

In RF, sampling by boostrap creates an 'out of the bag' (OOB) sample which is an important feature due to its usefulness in estimation of VIM. These scores are derived by permutation testing of the OOB data in the error estimation step. VIM scores of index x_i is described as the mean error-rate over all trees in the RF ensemble. Detailed information on VIM calculation is described elsewhere (Breiman, 2001), but for descriptive purposes, we simplify the formal representation of this measure as:

$$VIM(x_i) = \frac{1}{ntree} \sum_{1}^{ntree} (\widetilde{err.OOB_t} - err.OOB_t)$$

Where, $VIM(x_i)$ is a function estimating the VIM score for variable x_i and $ntree$ is the number of trees in the RF ensemble whereas $\widetilde{err.OOB_t}$ is the number of misclassifications tested on a tree t where the input was permuted values of variable x_i. Conversely, $err.OOB_t$ is the number of misclassifications tested on a tree t whose input was the non-permuted values of variable x_i.

Validity of permutation test derivation of VIM in the RF algorithm operates on the premise that if a variable is important, then permuting its values (realistically) leads to reduced accuracy of class prediction. Variables were selected using the method described by Genuer *et al.* (2010) although variable selection procedures based on variable importance ranking in RF tree models have been discussed in other works. Selection can be perfomed by score thresholding of variable importance or reduction of features using prediction performance. For the case of classification, RF variable reduction has been applied for feature reduction for example in the work of Aguas *et al.* (2013) who determined discriminant features of influenza proteins to determine functional relevance. The most implemented strategy is the recursive feature selection addressed in the work of Díaz-Uriarte and De Andres (2006). Genuer *et al.* (2010) proposed a simple and intuitive RF-based feature selection for datasets whose structure is $n < p$. In this procedure, variable importance is computed conditionally based on classifications of observations (in our case amino acid properties in Table 3.2). The overall objective is to attain variables that are highly related to the response variable and more importantly find a small number of variables sufficient for a parsimonious prediction of the response variable. Two steps in this procedure are therefore:

- Preliminary elimination and ranking of variables:

  Sort variables in decreasing RF scores and then remove variables of small importance

- Variable Selection:

  Construct nested collection of random forest models involving the first k features for $k = 1$

to m. Predict starting from the ordered features and construct a sequence of RF models andtesting variables in a step-wise procedure.

In the present work, we initially ran 1000 RF models of classification trials of AAindex, at each run recording the mean decrease in accuracy (VIM score). Indices were then ranked on a decreasing score order. The variation of these VIM scores was obtained and the point of minimum variance was initialized as a threshold, from which 93 amino acid indices were retained for further index reduction by nested RF feature selection approach described in detail by Genuer *et al.* (2010). The threshold of significant deviation in the increasing error rates from the nested RF modeling was set to the number of variables above which the error rates significantly increase above the threshold of 0.02 percent meaning that at most, only a single amino acid misclassification could be accepted in the nested RF model. A reduced amino acid index (rAAindex), was thus derived and its usefulness as a representation of amino acid information tested on data from our KNApSAcK motorcycle database.

3.2.2 KNApSAcK Motorcycle DB: Peptide sequence and metabolic reaction relationship DB

It is necessary to extend the species–metabolite relationship DB by incorporating a secondary metabolite pathway DB that include pathways with detected enzymatic reactions and other actual or predicted peptide sequences that may be involved in these pathways. We surveyed reactions of secondary metabolites in scientific literature, and amino acid sequences involved in secondary metabolism were obtained from public databases in PubMed. All the data comprising 2,881 secondary metabolic reactions were accumulated in the KNApSAcK Motorcycle DB (http://kanaya.naist.jp/motorcycle/top2.html) as shown in the main window of the KNApSAcK Motorcycle DB (**figure 3.3**), enzyme reactions can be retrieved using keywords of enzymes, species, genes, metabolites and peptide sequences obtained from a BLASTP search. For metabolite search using its keywords, we obtain information on enzyme name, reaction involved, compound class and subclass of metabolic reactions and reaction mechanisms. Using BLASTP search on this DB, we can predict reaction equations for a targeted peptide sequence using information on the class and subclass of metabolic pathways. Thus, the Motorcycle DB makes it possible to predict enzyme reactions based on the class and subclass of metabolic reactions evidenced by experiments mentioned in scientific literature. This differentiates it from KEGG (Kanehisa and Goto, 2000) and BioCyc (Caspi *et al.*, 2008). We have thus far obtained 596,974

protein sequences of 59,165 plant species and 124,292 protein sequences of 66 bacterial species from the Non-redundant protein sequences of PlantGDB.

For analytical purposes of the presently developed method, we narrowed our test dataset to terpenoid synthase peptide sequences with > 200 amino acid residues. Terpenoids are organic metabolites of plants that have been shown to have insect-pesticide properties among other roles. Terpenoid synthases sequence from the KNApSAcK database (Shinbo *et al.*, 2006) was examined for patterns in diversity. Enzymes annotated to four families namely monoterpenoid synthases, diterpenoid synthases, triterpenoid synthases and sesquiterpenoid synthases, were examined by PCA. Understanding the data structure of these terpenoid enzyme sub-families is important for annotation of similar organic compounds (Bohlmann *et al.*, 1998). Multiple sequence alignment and gap removal was carried out to extract homologous regions of sequences from the four sub-families of terpenoid synthases. These domains had a length of 28 residues for the 283 sequences. Binary and rAAindex (BPP) feature-coding of amino acid residues in these sequences were compared.

3.2.3 Sequence diversity characterization based on principal component analyses (PCA)

The present work attempts to examine diversity of secondary metabolic enzyme groups using datasets from the KNApSAcK Motorcycle DB by integrating amino acid attributes represented in the rAAindex where PCA was used to analyse variation. PCA is a technique that enables efficient interpretation of variation and relationship between variables in a huge dataset represented by higher dimensional vectors (Jolliffe, 2005). It is widely applied in bioinformatics as exemplified by Tatusov *et al.* (2001) who phylogenetically classified genomes by protein function. For comparative purposes, a BPP integrated dataset was analysed in comparison to 8-bit binary encoding of the same sequence set. We initially generated lattices representing individual sequences encoded by both rAAindex and 8-bit binary feature coding.

3.3 Results and discussion

3.3.1 Reduced amino acid index

Physico-chemical properties of amino acids quantitatively describe the overall biochemical behaviour of peptide and protein sequences (Grantham, 1974). The amino acid index database (Kawashima

Figure 3.3: *Enzyme-reaction database. (A) The main window of Motorcycle. (B) An example of a keyword search. (C) An example of a BLASTP search.*

and Kanehisa, 2000) has collected properties of amino acids measured by various researchers since the 1970s using scientific instruments and quantifiable metrics. It is essential to consider these properties in objective analyses of sequence data. Numeric quantification is also pivotal because it gives a flexible way of integrating this information in a mathematically and statistically amenable form different from the alphabetical string representation.

The AAindex database is highly redundant and has some missing values for certain properties. In its raw form, the AAindex is not suitable for direct BPP encoding. Ideally, a reduced set would work for most sequence analyses. Researchers have utilized various compression techniques to reduce the AAindex. Atchley and colleagues used a multivariate factor analysis approach to propose a compressed variable set of five vectors describing amino acids in a multi-dimensional space (Atchley *et al.*, 2005). More recently, fuzzy c-means algorithm has been applied in clustering the AAindex indices and the resultant clusters incorporated in a support vector machine modeling experiment to predict DNA-binding domains(Huang *et al.*, 2011).

Factor analysis (FA) is a useful approach for purpose of defining a minimal set of 'factors' that simplify intepretion of protein sequence characteristics. Random forest (RF) variable reduction however differs from FA by selection of important variables without compressing the whole variable set into fewer descriptive factors. RF has actually been proven to be a useful tool for biological data as described by Díaz-Uriarte and De Andres (2006). Here, BPP selection entails minimizing the original variables in the AAindex by elimination of redundancy, high collinearity and less informative variables whilst mantaining a sufficiently parsimonious set of the original BPP properties represented in the AAindex. Compression (as in FA) on the other hand is re-definition of original AAindex variables into new components by multivariate techniques such as PCA (Jolliffe, 2005) and factor analysis (Kline, 1994). We argue against compression in the context of AAindex. While a minimized descriptive set is achieved, there results a new challenge with regards to the complexity of biological intepretation when the redefined variables are applied in subsequent downstream mathematical, statistical or computational analyses.

From the initial 544 properties contained in the amino acid index database, 13 of which had missing values were dropped. The redundancy was further reduced by dropping variables with a correlation coefficient greater than 0.85, further trimming the set to 283 amino acid indices. The retained indices were then subjected to the RF algorithm (1000 trials) and variable reduction performed using the technique proposed by Genuer *et al.* (2010). Variable importance scores

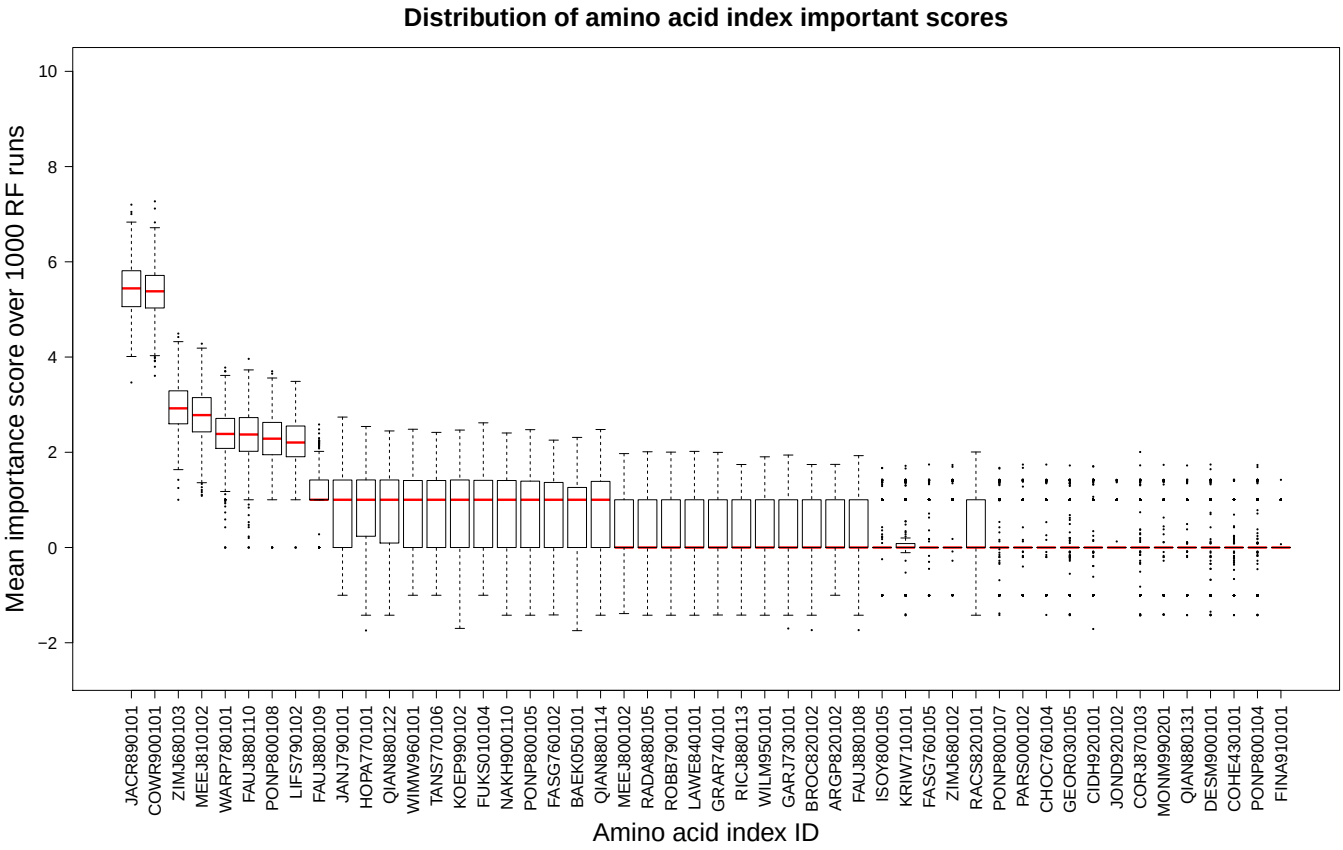

Figure 3.4: *Variable importance score distribution for 1000 runs of random forest classification on the amino acid index. Each boxplot represents distribution of each property (also called variable) represented on the horizontal axis. The properties have been ordered in decreasing order of the median score (red line in boxplot). For easier visualization, the set has been truncated to show the top 50 properties. The corresponding properties are shown in table 3.4*

(VIM) were ranked in decreasing order(**figure 3.4**).

The "importance" score of a BPP illustrates its significance with regard to amino acid classification. Figure 3.4 illustrates the VIM score distribution for 1000 runs of random forest classification of the AAindex. Each boxplot represents the distribution of each BPP (also called variable) represented on the horizontal axis. These properties have been ordered in decreasing order of the median score (red line in boxplot). For easier visualization, the set has been truncated to show the top 50 properties by mean VIM ranking. The corresponding properties are shown in table 3.4.

Variation of the ranked scores was observed as shown in Fig. 3.5. Standard deviation of the importance scores of the properties (y axis), models contribution of each property towards performance of the RF algorithm. Those variables with a close to zero variation are less 'important'. At the tail end variance is higher than zero but is largely due to chance (p-value > 0.05). Variables with a mean VIM score of 0 or less were dropped, lowering the retained variables to 93 indices.

Nested RF modeling was performed using the ranked indices starting with the highest ranked variable followed by subsequent stepwise addition of the remaining at each step. Error rates were estimated at each level in the nested model RF reduction. Details of nested RF are explained in the work of Genuer *et al.* (2010). The threshold of acceptable error rate was set to 2 percent. Fig. 3.6 shows that when the first 8 indices are used for classification, the error rate remained under the 2 percent error rate (horizontal red line in figure) threshold whereas it significantly rose with subsequent addition of indices.

An RF-reduced subset of the amino acid index; rAAindex (**table 3.3**), with these 8 most important BPPs is therefore proposed for use in BPP-encoding especially for statistical learning and other mathematical tasks involving protein sequences. The properties retained are shown in table 3.5.

Significance of Random Forest based redundancy reduction

Besides its user-friendly implementation and intepretability of RF feature selection with respect to selecting important BPP, the non-parametric data structure assumption in the RF classification is a vital point especially with regard to variable selection. Variable importance quantification by GINI index (impurity reduction) has been found to exhibit some biases (Strobl *et al.*, 2007) and can also randomly select a variable amongst co-linear features without regards to its performance

Table 3.3: Selected properties based on nested random forest variable selection

Amino acid	JACR890101	COWR900101	ZIMJ680103	MEEJ810102	FAUJ880110	WARP780101	PONP800108	LIFS790102
Ala	0.18	0.42	0.00	1.00	0.00	10.04	6.05	1.00
Arg	-5.40	-1.56	52.00	-2.00	3.00	6.18	5.70	0.68
Asn	-1.30	-1.03	3.38	-3.00	3.00	5.63	5.04	0.54
Asp	-2.36	-0.51	49.70	-0.50	4.00	5.76	4.95	0.50
Cys	0.27	0.84	1.48	4.60	0.00	8.89	7.86	0.91
Gln	-1.22	-0.96	3.53	-2.00	3.00	5.41	5.45	0.28
Glu	-2.10	-0.37	49.90	1.10	4.00	5.37	5.10	0.59
Gly	0.09	0.00	0.00	0.20	0.00	7.99	6.16	0.79
His	-1.48	-2.28	51.60	-2.20	1.00	7.49	5.80	0.38
Ile	0.37	1.81	0.13	7.00	0.00	8.72	7.51	2.60
Leu	0.41	1.80	0.13	9.60	0.00	8.79	7.37	1.42
Lys	-2.53	-2.03	49.50	-3.00	1.00	4.40	4.88	0.59
Met	0.44	1.18	1.43	4.00	0.00	9.15	6.39	1.49
Phe	0.50	1.74	0.35	12.60	0.00	7.98	6.62	1.30
Pro	-0.20	0.86	1.58	3.10	0.00	7.79	5.65	0.35
Ser	-0.40	-0.64	1.67	-2.90	2.00	7.08	5.53	0.70
Thr	-0.34	-0.26	1.66	-0.60	2.00	7.00	5.81	0.59
Trp	-0.01	1.46	2.10	15.10	0.00	8.07	6.98	0.89
Tyr	-0.08	0.51	1.61	6.70	2.00	6.90	6.73	1.08
Val	0.32	1.34	0.13	4.60	0.00	8.88	7.62	2.63

Table 3.4: Top 50 indices ranked based on random forest feature selection

	ID	Property
1	RADA880101	Information value for accessibility; average fraction 35% (Biou et al., 1988)
2	ROSM880101	Information value for accessibility; average fraction 23% (Biou et al., 1988)
3	KIDA850101	Retention coefficient in TFA (Browne et al., 1982)
4	EISD840101	Normalized hydrophobicity scales for alpha+beta-proteins (Cid et al., 1992)
5	JACR890101	Normalized hydrophobicity scales for alpha/beta-proteins (Cid et al., 1992)
6	COWR900101	Consensus normalized hydrophobicity scale (Eisenberg, 1984)
7	BLAS910101	Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)
8	MEEJ810101	Hydrophobic parameter pi (Fauchere-Pliska, 1983)
9	CIDH920104	Number of hydrogen bond donors (Fauchere et al., 1988)
10	GRAR740102	Number of full nonbonding orbitals (Fauchere et al., 1988)
11	ZIMJ680103	Polarity (Grantham, 1974)
12	MEEJ810102	Hydration number (Hopfinger, 1971), Cited by Charton-Charton (1982)
13	RADA880104	Hydropathy index (Kyte-Doolittle, 1982)
14	KUHL950101	Hydrophobic parameter (Levitt, 1976)
15	FAUJ880110	Conformational preference for parallel beta-strands (Lifson-Sander, 1979)
16	RADA880107	Average surrounding hydrophobicity (Manavalan-Ponnuswamy, 1978)
17	WARP780101	Retention coefficient in NaClO4 (Meek-Rossetti, 1981)
18	BIOV880102	Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981)
19	BIOV880101	8 A contact number (Nishikawa-Ooi, 1980)
20	FASG890101	Partition coefficient (Pliska et al., 1981)
21	PONP800108	Average number of surrounding residues (Ponnuswamy et al., 1980)
22	FAUJ830101	Hydrophobicity (Prabhakaran, 1990)
23	CORJ870101	Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)
24	MANP780101	Transfer free energy from chx to wat (Radzicka-Wolfenden, 1988)
25	LIFS790102	Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)
26	GUOD860101	Energy transfer from out to in(95%buried) (Radzicka-Wolfenden, 1988)
27	WOLS870101	Mean polarity (Radzicka-Wolfenden, 1988)
28	WOLR790101	Side chain hydropathy, uncorrected for solvation (Roseman, 1988)
29	LEVM760101	Side chain hydropathy, corrected for solvation (Roseman, 1988)
30	WOLR810101	Normalized frequency of chain reversal D (Tanaka-Scheraga, 1977)
31	ROSM880102	Average interactions per side chain atom (Warme-Morgan, 1978)
32	WOEC730101	Polar requirement (Woese, 1973)
33	PLIV810101	Hydration potential (Wolfenden et al., 1981)
34	RADA880108	Principal property value z1 (Wold et al., 1987)
35	FAUJ880109	Polarity (Zimmerman et al., 1968)
36	HOPA770101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White, 1996)
37	CIDH920103	Hydropathy scale based on self-information values in the two-state model (9% accessibility) (Naderi-Manesh et al., 2001)
38	TANS770106	Hydropathy scale based on self-information values in the two-state model (16% accessibility) (Naderi-Manesh et al., 2001)
39	PRAM900101	Hydrophilicity scale (Kuhn et al., 1995)
40	ENGD860101	Retention coefficient at pH 2 (Guo et al., 1986)
41	BROC820101	Modified Kyte-Doolittle hydrophobicity scale (Juretic et al., 1998)
42	NADH010103	Knowledge-based membrane-propensity scale from 1D_Helix in MPtopo databases (Punta-Maritan, 2003)
43	NADH010102	Hydrophobicity index (Wolfenden et al., 1979)
44	KYTJ820101	Hydrophobicity-related index (Kidera et al., 1985)
45	EISD860103	Weights from the IFH scale (Jacobs-White, 1989)
46	NISK800101	Hydrophobicity index, 3.0 pH (Cowan-Whittaker, 1990)
47	JURD980101	Scaled side chain hydrophobicity values (Black-Mould, 1991)
48	WIMW960101	NNEIG index (Cornette et al., 1987)
49	QIAN880122	Hydrophobicity index (Engelman et al., 1986)
50	PUNT030101	Hydrophobicity index (Fasman, 1989)

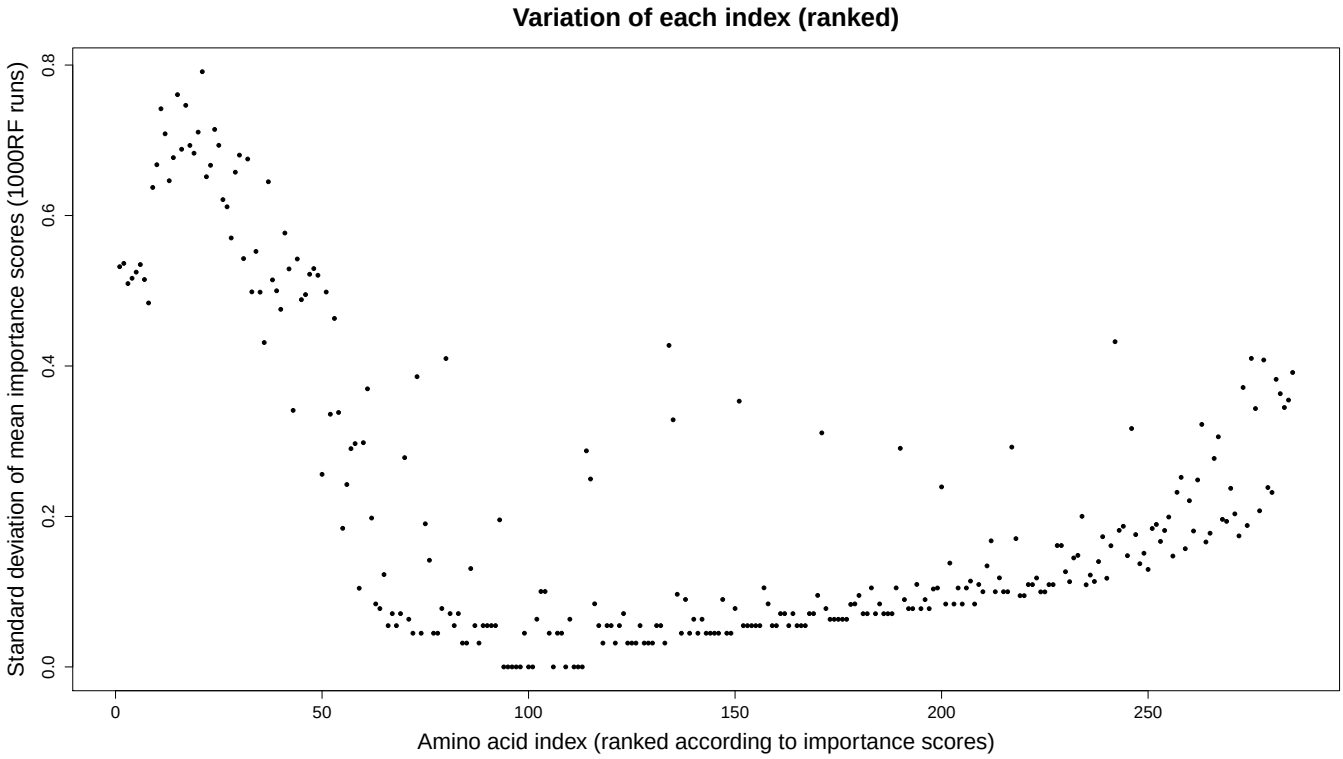

Figure 3.5: *Standard deviation of the importance scores of the properties (y axis), models contribution of each property towards performance of the RF algorithm. Those variables with a close to zero variation are less 'important'. At the tail end variance is higher zero but is largely due to chance.*

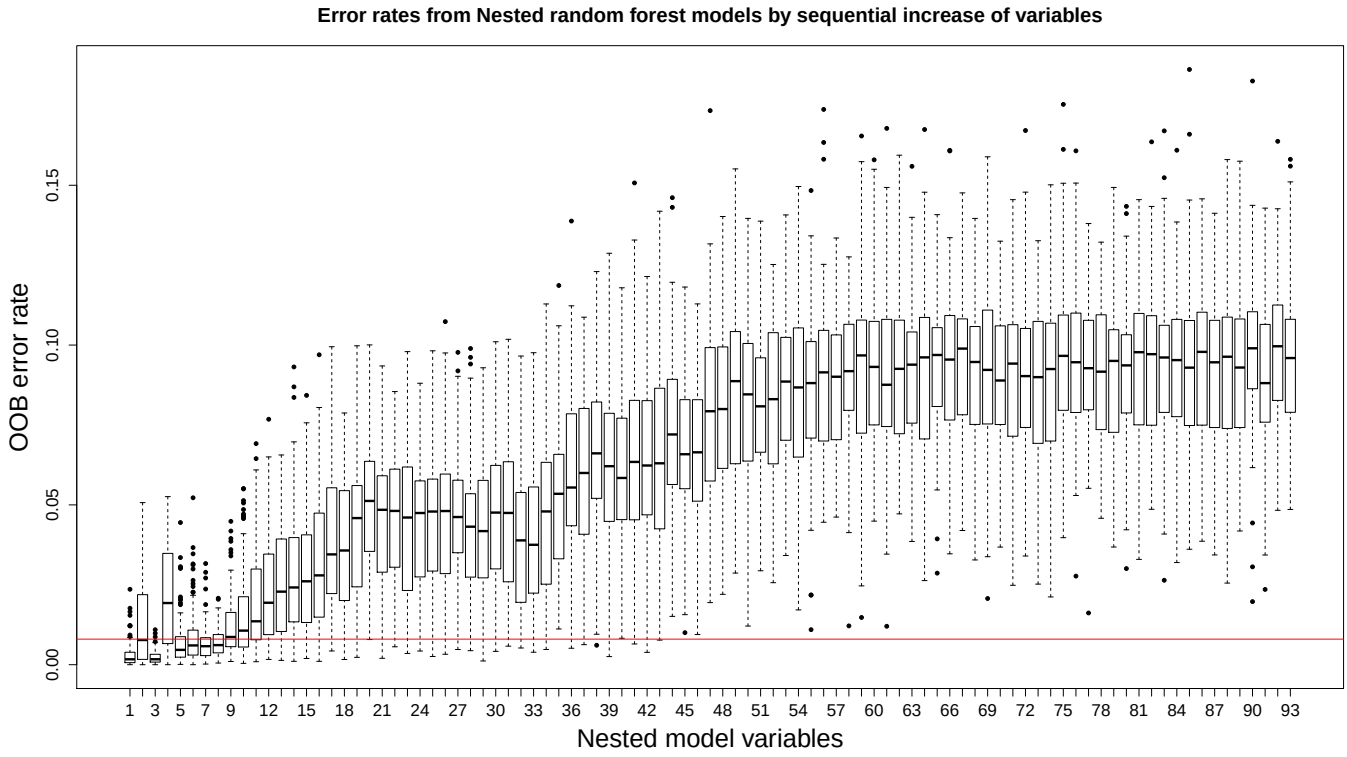

Figure 3.6: *Nested random forest variable selection: Variables have been ordered by their importance scores, new RF models are built by single variable addition in the nested RF set up and RF error rates are measured. In this experiment, each box and whisker plot represents the distribution of error rates from 100 trials at each nested RF step. In total, there were 93 steps corresponding to the 93 top indices from previous step. The y axis shows the error rates as a percentage. The threshold of acceptable mean error rate was set at 2 percent shown by the red horizontal line*

as predictor. This can lead to mis-intepretation of data. This can however be reduced by the random selection of features at nodes during tree building.

In our feature selection model, we applied variable importance estimation by mean decrease in accuracy. This feature of RF implements permutation of values for each feature with the simple logic that a predictor model (RF), should perform poorly when values of an important feature are permuted whereas this process of permutation should have no effect on the performance of the model when values of a 'non-important' feature are permuted.

Simply stated, For the rAAindex with the selected features in listed table 3.5, the nested RF model determined these variables from the amino acid index as the minimum number of variables that can classify the 20 amino acids in the BPP feature space close enough to the grouping illustrated in figure 3.7.

Table 3.5: Selected indices and their correspoding biochemical and physical attributes

	ID	Property
1	JACR890101	Number of full nonbonding orbitals (Fauchere et al., 1988)
2	COWR900101	Conformational preference for parallel beta-strands (Lifson-Sander, 1979)
3	ZIMJ680103	Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981)
4	MEEJ810102	Average number of surrounding residues (Ponnuswamy et al., 1980)
5	WARP780101	Average interactions per side chain atom (Warme-Morgan, 1978)
6	FAUJ880110	Polarity (Zimmerman et al., 1968)
7	PONP800108	Weights from the IFH scale (Jacobs-White, 1989)
8	LIFS790102	Hydrophobicity index, 3.0 pH (Cowan-Whittaker, 1990)

The selected features are further described below:

JACR890101: Number of full non-bonding orbitals

This BPP represents amino acid residues with carbonyl groups carrying two non bonding pairs of oxygen atoms. This influences hybridization or binding of residues to other atoms therefore influencing the structural properties of a peptide sequence especially in solvents. This property exhibited by residue side-chains therefore not only determines if an amino acid is hydrophobic or hydrophilic but also its overall structures.

COWR900101: Conformational preference for parallel beta-strands

A β-strand secondary conformation, requires non-local interactions between regions of the protein chain that are not necessarily consecutive in the amino acid sequence (Steward and Thornton, 2002). It has been shown that secondary structures with parallel β conformation require a mechanism to minimize interaction with solvent on the inner parts while fostering solvent interaction through hydrogen bonds on the exterior. This creates a strict requirement

for molecular shape of peptide sequences, thereby having an implication on protein folding patterns.

ZIMJ680103: Retention coefficient in NaH2PO4

Protein sequences have a varied retention time in high-pressure liquid chromatography (HPLC). This is dependent on the amino acid side-chains. Again, this property determines the behaviour of a residue in solvents (eg. NaHPO4). The cummulative influence of each residue contributes directly to the retention co-efficient values of the entire protein/peptide sequence (Meek, 1980). This capability is exploited in HPLC systems designed for protein purification. (Peptide separation).

MEEJ810102: Average number of surrounding residues

Residues surrounding a functional site influence conformation and function of a sequence. Gong *et al.* (2000) examined the impact of residues surrounding the *haem* group of cytochrome f. They concluded that the surrounding residues may change hydrophobicity properties of peptides by influencing the rates of electron transfer to plastocyanin through change of binding constants and altering of redox potential balance.

WARP780101: Average interactions per side chain atom

Side chains of the 20 amino acids interact with each other resulting in a possible 400 types of interactions. The atoms involved in the interactions are however alot more than this. This property influences the packing potential of a protein sequence. The higher the number of interacting atoms, the higher the packing capacity. This therefore determines structural properties, function and stability of sequences.

FAUJ880110: Polarity

Polar residues are those whose side chains reside in an aqueous environment. These amino acids are mostly found on the surface of a protein. Amino acids can be classified as either polar or non-polar. There are three types of polar amino acids: those with no charge for example serine, threonine and cysteine, those with negative charge for example aspartic acid and glutamic acid and those with positive charge eg histidine, arginine and lysine. Polarity influences hydrophobic/hydrophilic properties.

PONP800108: Weights from the IFH scale Interfacial Hydrophybicity (IFH) scale (Jacobs and White, 1989), quantifies the behaviour of intermembrane proteins that span between lipid layers of the membrane. Unlike polarity, IFH estimates the hydrogen bonding behaviour in lipid bilayer proteins with α-helix secondary structure.

Hydrophobicity index, 3.0 pH

This index estimates retention time of amino acid chains in high performance liquid chromatography at pH 3.0 whose results proved to be significantly different from at pH 7.5 but with high corellation for constants determined for water/octanol partitioning of N terminal protected amino acids (Cowan and Whittaker, 1989).

Using these properties, a number of general principles of protein sequences can be analysed to understand characteristics such as mutation in context of the biochemical and physical pressures.

Cellular environment related physical properties of proteins

Different parts of the cell have distinct chemical enviroments thus resulting in varied protein behaviours. The biggest cellular location influence is perhaps on those properties related to solubility, especially in membrane proteins. Soluble proteins are hydrophilic in nature and are surrounded by water molecules whereas membrane proteins are hydrophobic and are surrounded by lipid molecules. Soluble proteins still vary in kind. Extracellular soluble proteins exist outside of the cell whereas cytosolic soluble proteins exist within the cell. Cytosol environment is more aqueous than extracellular environments. These two types differ especially in protein density requirements a feature that can be attributed directly to amino acid contents particularly Cysteine. Extra-cellularly, Cysteine molecules in proximal distance form di-sulphide bonds which is important for folding. Formation of di-sulphide bonds in cytosol is difficult and therefore it can be argued that cysteine does not influence protein structure intra-cellularly as it does extra-cellularly and may therefore be subject to mutational pressure. It would be great for a study to compare cysteine contents of intracellular proteins to membrane proteins to determine the impact of selection against its role in the cytosol. The gene ontology consortium give a detailed classification of cellular location aspects of protein sequences, thus enabling further research into the role of cellular factors driving protein properties.

Influence of properties on protein structure

Proteins contain micro-environments resulting from its structural features which are actually determined by the amino acid content. Soluble proteins, have a surface that interacts with water and therefore is consists mostly of polar or charged amino acids than the inner side of the protein micro-environment which is more likely to contain hydrophobic amino acids.

Role of Biochemical and physical properties on sequence Evolution

Functional proteins in different organisms often belong to homologous groups. Understanding features of these families give insights on possible functions. Sequences in these families undergo evolutionary processes to achieve homology. Speciation and duplication are the two main processes involved in these mechanisms. Orthologs are proteins related together by speciation whereas duplication results in paralogous proteins. Over time, paralogs are likely to evolve to perform different functions. Functional similarity is often correlated with sequence identity. For human beings, greater sequence identity ($< 80\%$) likely mean orthology. In distantly related organisms, such high similarity is unlikely to be achieved and thus *ad hoc* rules of thumb are often applied (commonly $< 40\%$ identity). When considering evolutionary pressures resulting in mutation, certain amino acids are likely fo play key structural roles only in orthologs conferring specificity thus meaning they vary in all orthologs. One way of classifying amino acids is by mutational matrices. These matrices contain sets of numbers that describe propensities of exchanging one amino acid for another. Mutational matrices are derived from alignment of large numbers of sequences and counting frequencies for a particular substitution and the values often estimated using a model of evolutionary time. The best known among the matrices are the Point Accepted Mutation (PAM) and BLOSUM matrices.

These matrices are derived from alignments based on sequence identity. (Based on alphabetical sequence representation). They do not therefore account for likelihood and effects of certain substitutions at particular sites of the sequences, especially in absence of 3D structural information. A more robust classification of sequences requires the application of physical, chemical and structural properties of amino acids

Some research work has described the main physical factors of amino acid residues as the side chain properties and hydrophobicity. Effects of different amino acids on protein structure can account for mutation data when these physico-chemical properties do not. Hydrophobicity and size differ widely between glycine, proline, aspartic acid and glutamic acid but the distance between them according to mutational matrices is small because these residues prefer sharply turning regions on the surface of the protein; the phi and psi bonds of glycine are unconstrained by any side chain, proline forces a sharp turn because its side chain is bonded to the backbone nitrogen as well as to carbon, and aspartate and glutamate prefer to expose their charged side chains to solvent. The Taylor classification of residues based on properties is illustrated in the figure

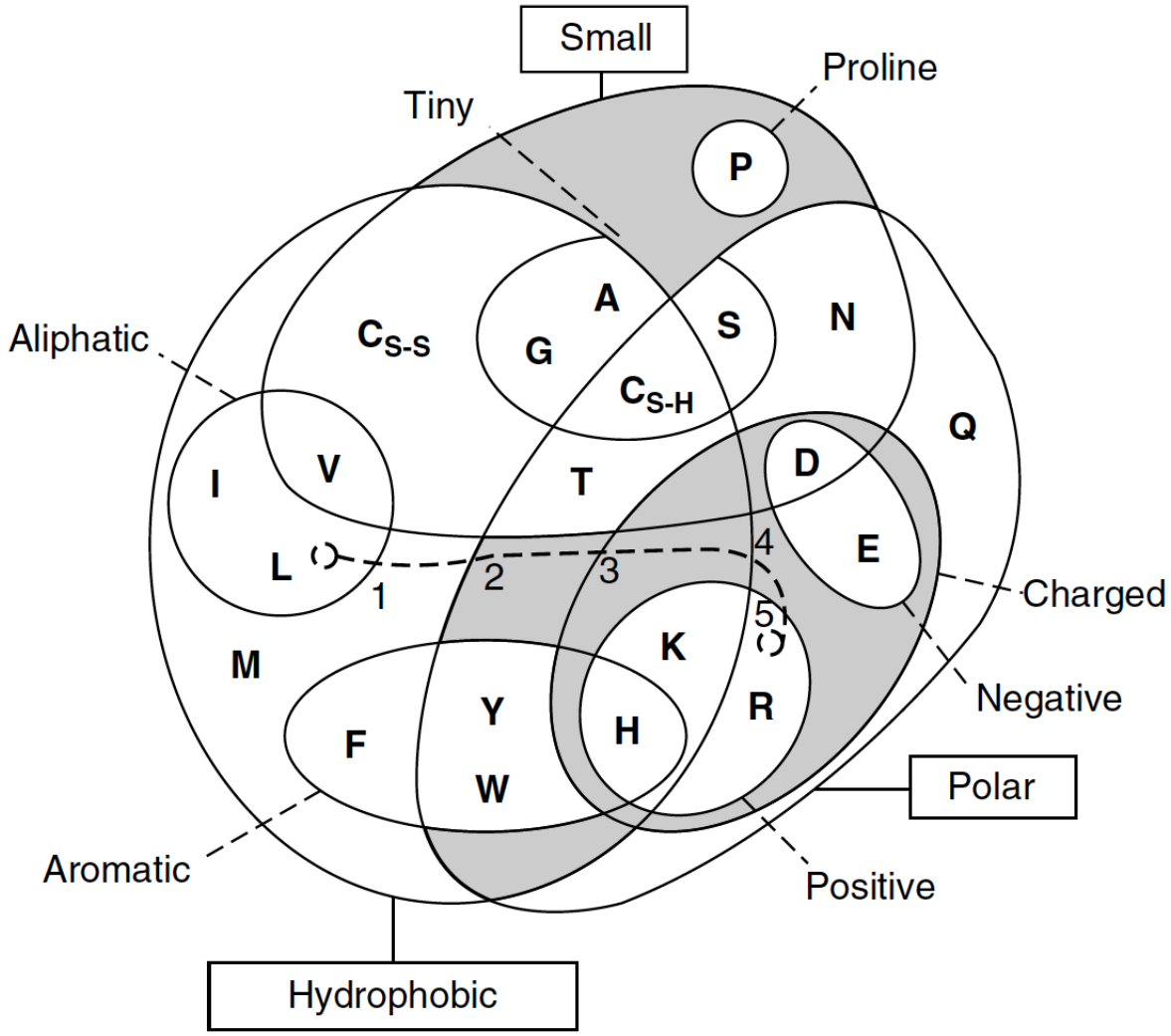

Figure 3.7: *The amino acids are positioned on this by multidimensional scaling of Dayhoffs mutation matrix, and then grouped by common physico-chemical properties. Size is subcategorized into small and tiny (with large included by implication). Affinity for water is described by several sets: polar and hydrophobic, which overlap, and charged, which is divided into positive and negative. Sets of aromatic and aliphatic amino acids are also marked. These properties were enough to distinguish between most amino acids. However, properties such as hydrogen-bonding ability and the previously mentioned propensity for sharply turning regions are not described well. Although these factors are less important on average, and would confuse the effects of more important properties if included on the diagram, the dangers of relying on simple classifications are apparent. This can be overcome somewhat by listing all amino acids which belong to each subset (defined as an intersection or union of the sets) in the diagram, for example small and non-polar, and including extra subsets to describe important additional properties. These subsets can be used to give qualitative descriptions of each position in a multiple alignment, by associating the positions with the smallest subset that includes all the amino acids found at that position. (Description and Illustration adopted from Betts and Russell (2003))*

3.3.2 Characterization of terpenene synthase sequence diversity

Terpenes are the largest group of plant natural products with a variety of core chemical structures comprising at least 30,000 compounds and synthesized majorly by terpenoid synthases (Connolly and Hill, 1991). Terpene diversity is caused by the large number of different terpene synthases used in the first step of terpene synthesis, and some terpene synthases produce multiple products (Degenhardt *et al.*, 2009). Terpene synthases are generally classified according to the number of carbons in their substrates, that is, geranyl diphosphate (C10, GPP) for monoterpene synthases, farnesyl diphosphate (C15, FPP) for sesquiterpene synthases, geranylgeranyl diphosphate (C20, GGPP) for diterpene synthases and squalene for triterpene synthases (C30). The rather limited similarity of plant terpenoids (Bohlmann *et al.*, 1998) complicates annotation of their enzymes. Clustering algorithms improve the resolution to some extent. PCA was used to analyse variation among 4 terpene synthase sub families. We examined the perfomance of PCA classification when rAAindex BPP encoding is implemented,in comparison to the commonly applied 8-bit binary string representation. According to the overall result, the combined variance explained by the first two principal components is 30.02 (**figure 3.8 left**) percent whereas the variance explained by the first two components in the binary encoded set is 14.84 percent(**figure 3.8 right**). Triterpenoid synthases can be clearly distinguished from the other categories, which was also consistent with our previous findings (Ikeda *et al.*, 2013).

We compared the performances of rAAindex to 8-bit binary encoding of amino acids(Staden, 1977) in an actual dataset (described in the data and methods section). Binary coding is the most popular representation scheme for machine learning tasks of protein data (Coghlan *et al.*, 2001; White and Seffens, 1998) and was utilized as a benchmark of comparison to rAAindex encoding. Figure 3.8 (left) shows the fragment distribution of terpene synthases coded by 8-bit binary notation mainly clustered into five regions. Similarly, figure 3.8 (right) shows the distribution of terpene synthases coded by rAAindex into the same 5 groups. Principal component analysis of the terpenoid synthase sub-families where amino acid residues are encoded in 8-bit binary method. PC1 and PC2 show a combined variance of 14.84 percent explained variance. The second part of the figure illustrates PCA of the same data set encoding the biochemical and physical properties of amino acid residues described above. PC1 and PC2 in this case explain 30.02 percent variance showing that more sequence information is described by the BPP subset. The four sub-families clustered are monoterpenoid, diterpenoid, triterpenoid and sesquiterpenoid synthase. Triterpenoid synthases and subgroups of terpenoid synthases are

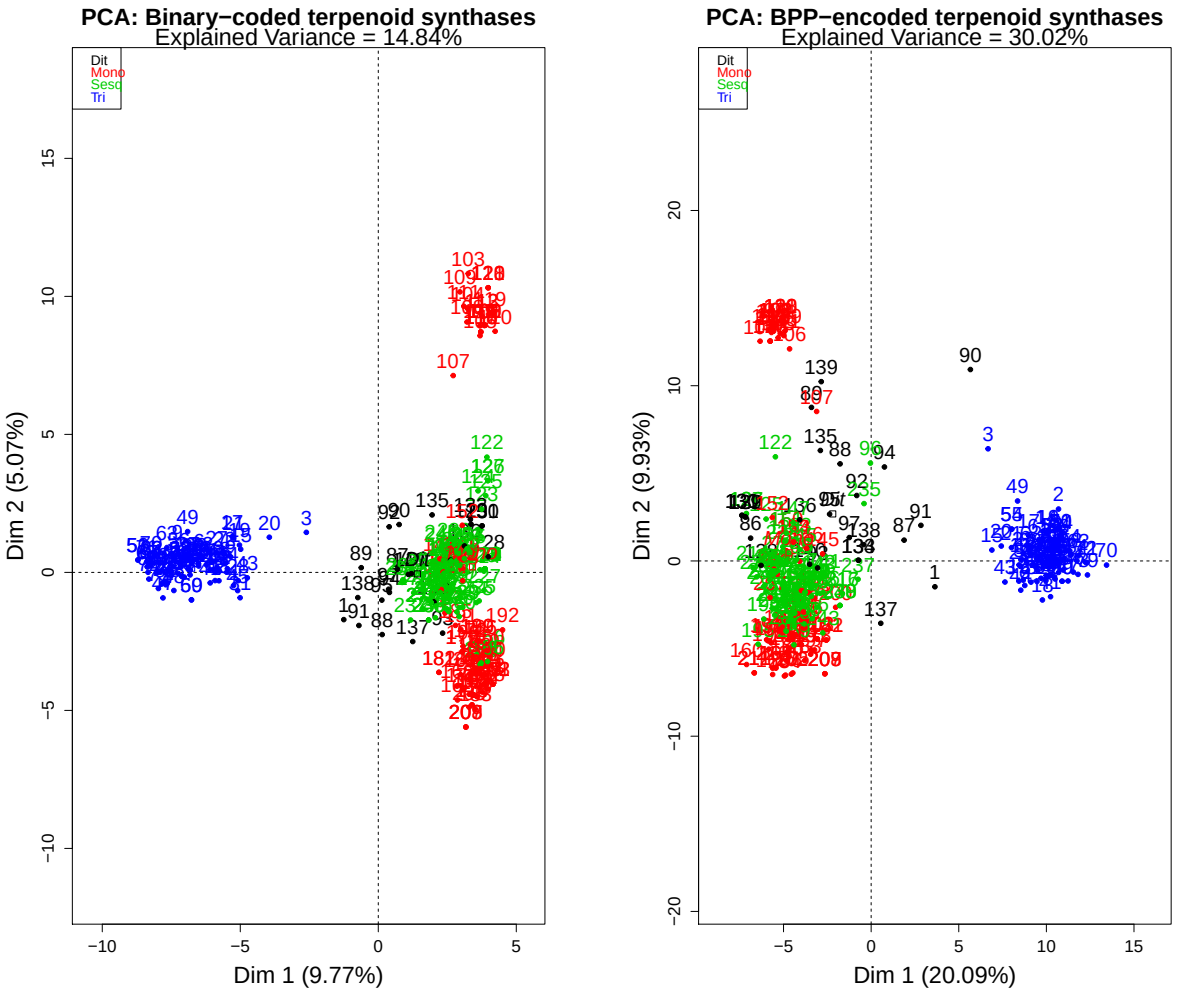

Figure 3.8: ***Left****: Principal component analysis of the terpenoid synthase sub-families where amino acid residues are encoded in 8-bit binary representation. PC1 and PC2 show a combined variance of 14.84 percent explained variance.* ***Right:*** *Principal component analysis of the same data set encoded using the biochemical and physical properties of amino acid residues. PC1 and PC2 in this case explain 30.02 percent variance. The four sub-families clustered are monoterpenoid, diterpenoid, triterpenoid and sesquiterpenoid synthase. Triterpenoid synthases and subgroups of terpenoid synthases are distinctly different in structure compared the other synthases.*

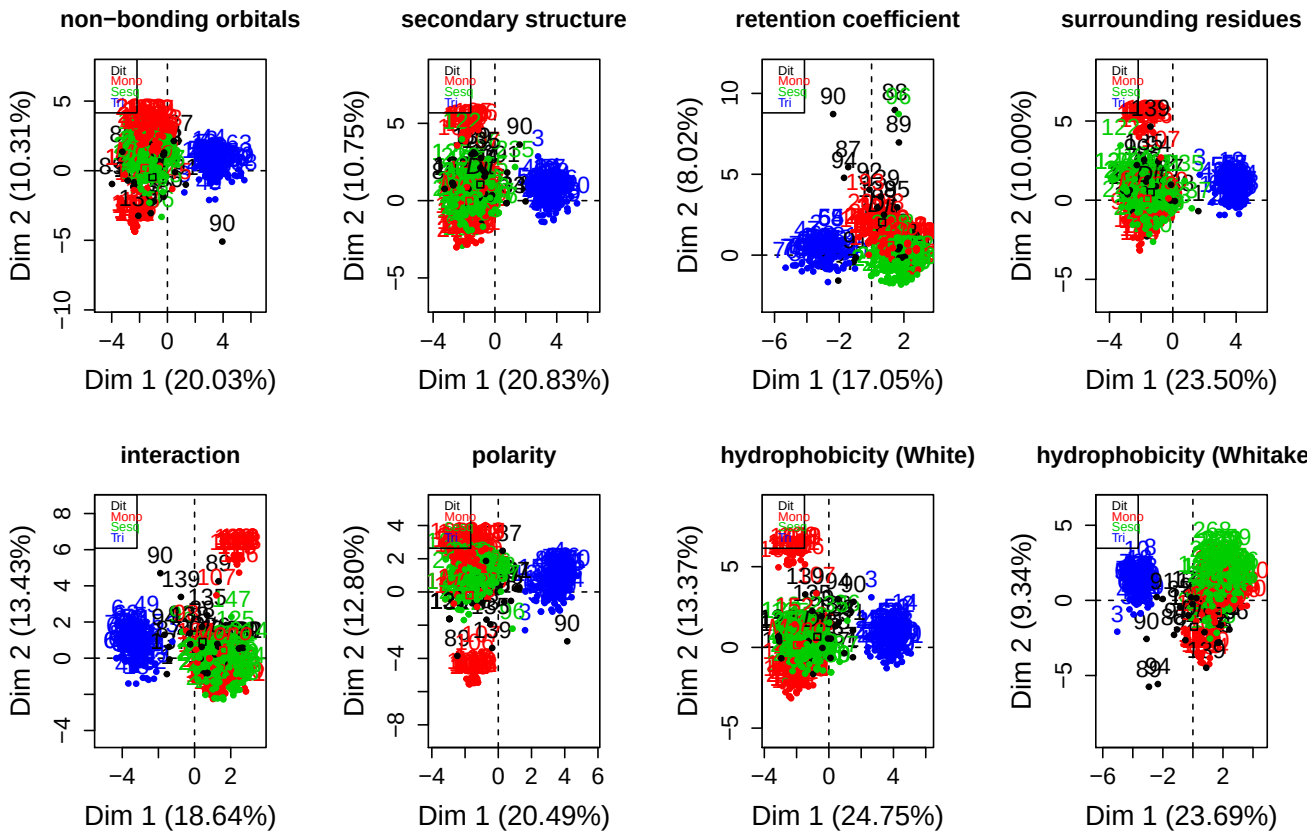

Figure 3.9: *The selected 8 properties were each evaluated for their contribution in the Terpene synthases datasets*

distinctly different in structure compared the other synthases. It is noted that three types of terpene synthases except for the diterpene synthases are less divergent at the peptide sequence level, that is, small changes in peptide sequences of the terpene synthase make it possible to synthesize many different terpenoid compounds. The orders of fragments from the N- to the C-terminus in enzymes are arranged in two clusters for monoterpene synthases and are a single cluster for the other categories. Thus, monoterpene and sesquiterpene synthases are very similar in arrangement of peptide fragments, which is consistent with the similarity of the 3D structures in monoterpene and sesquiterpene synthases (Hyatt *et al.*, 2007; Nagegowda *et al.*, 2008), and with the fact that several bifunctional enzymes possess both sesquiterpene and monoterpene synthase activities (Nieuwenhuizen *et al.*, 2009; Nagegowda *et al.*, 2008). Diterpene and triterpene synthases have inherent structures specified by a single cluster for diterpene synthases, and two clusters from the N- to C-terminus for triterpene synthases.

Hypothetical explanation of why property integration is useful

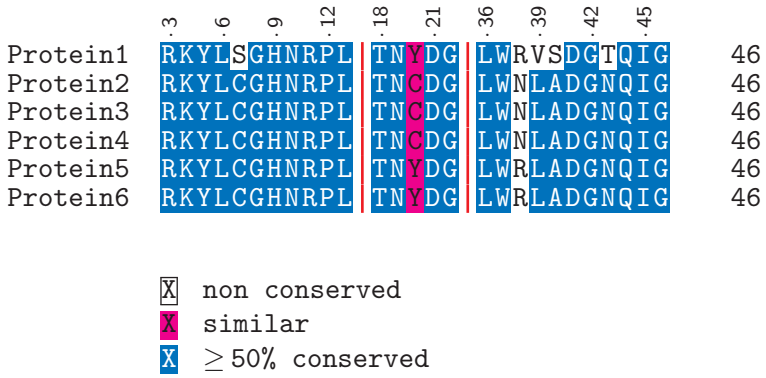

Figure 3.10: *An example of a typical protein multiple sequence alignment showing an example of orthologs. Blue colored regions show identical residues whereas red colored columns represent similar residues. Un-colored columns shows un-conserved regions*

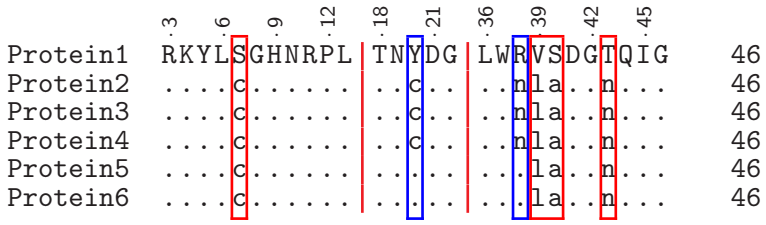

Figure 3.11: variable sites of figure 3.10. Site identities are shown at top

Site 7, 39, 40 and **43** differ by a single amino acid residue while **site 20** and **38** differ by three amino acids. Alphabetical representation in this sense cannot capture variation in these columns in the same way as numeric representation (Table 3.3). Our approach also enables mathematical determination of variance which explains measurable site-specific differences. It is possible to make biological inferences from this information. For example, **site 20** and **site 38** both differ by 3 amino acid residues but the variance in **site 38** is lower than that of **site 20** meaning that, **site 20** is less susceptible to mutation compared to **site 38**.

3.4 Conclusion

This chapter has introduced a subset of eight biochemical and physical attributes of amino acids, that can be encoded in protein sequences when performing sequence based analysis. These features quantify attributes of individual residues in numerical metrics that improve amenability to mathematical and statistical tasks and also enhances biological intepretability of such tasks. Terpenoid synthases protein set was used as an example to evaluate the encoding of these attributes when performing PCA. The terpenoid synthase sub-families classification established that more variance is explained when BPP are encoded compared to the commonly used binary encoding which does not integrate physico-chemical aspects of protein sequences. Both eukaryotic and prokaryotic protein sequences are difficult to computationally characterize. Due to the diverse attributes of amino acid residues, protein similarity often transcends beyond residue identity as identified by alphabetical notation. Even within a protein family the diversity of sequences is so high such that the similarity between sequence pairs is so low that simple comparison using tools such as BLAST cannot give meaningful results. Our proposed approach has shown that to characterize these sequences which often display such low similarity requires a novel approach for their representation in order to cover higher percentage of variance resulting from the residue differences. A similar method had been discussed by Atchley *et al.* (2005) but with methodological difference in choice of feature selection. In protein sequence computational biology such as those mapping biological function and phylogeny are further improved by additional incorportation of quantifiable features of residues. For example the exploratory classification of terpenoid synthase families as presented in this work, the clear separation of sub-family structures and the increase in percentage of variance explained suggests more

usability of feature encoding in sequence analysis. This can also be extended to phylogenetic studies.

Chapter 4

Integrated platforms for transcription regulation network analysis

Chapter Summary

Transcription regulation is the all important process that distinguishes phenotypes. It is the key element that defines uniqueness of characteristics at the cellular level. Simply stated, transcription regulation is the main reason why cell types differ from each other. In this chapter we focus on gene regulation networks from a bioinformatics perspective. This chapter will describe the following:

- The nature of transcription regulation networks
- Integrated methods for gene regulation analysis
- Proposed integrated method for mining transcription regulation networks
- Application on a few example gene expression datasets

Abstract

Systems for analysis of transcription regulation networks (TRNs) vary considerably in scope, underlying models and implementation. It has recently become necessary to merge biological objectives into software and method development for data-oriented inference of transcription regulation networks. Conventionally, workflows examining TRNs from gene expression data often involve distinct analytical steps. There is need for pipelines that unify data mining and inference deduction into a singular framework to enhance intepretation and hypotheses generation. We propose a workflow that merges network construction with gene expression data mining focusing on regulation processes in the context of transcription factor driven gene regulation. The pipeline implements pathway-based modularization of expression profiles into functional units to improve biological intepretation. This knowledge-based utilization of well-defined biologically functional units is essential for both experimental biology and computational biology. The integrated workflow was implemented as a web application software (TransReguloNet) with functions that enable pathway visualization and comparison of transcription factor activity between sample conditions defined in the experimental design. The pipeline merges differential expression, network construction, pathway-based abstraction, clustering and visualization. The framework was applied in analysis of expression profiles related to cancer. Additionally we also combined a few datasets with pathway enrichment analysis, where we identified 23 enriched pathways that can be further studied for biomarker extraction and identification of key regulation of gene expression processes in the lung cancer context.

4.1 A common framework for handling gene expression profiles

A typical gene expression analysis spans both the experimental and computational procedures as illustrated in the pipeline in Fig. 4.1. Quantitative estimation is the entry point for bioinformatics but is preceeded by the pivotal 'wet-bench' experimental steps which vary depending on objectives. Sequencing assays for instance, involve sample preparation (extracting nucleic acids) followed by library preparation which may include procedures such as adapter ligation and so on. The sequencing process in the next step involve instrument driven generation of raw sequence reads as base-calls and quality scores. The computational steps then handle the data analysis with the starting point being the quantitative estimation of experimental assay output. This stage involves matching the machine extracted information (eg array intensity of

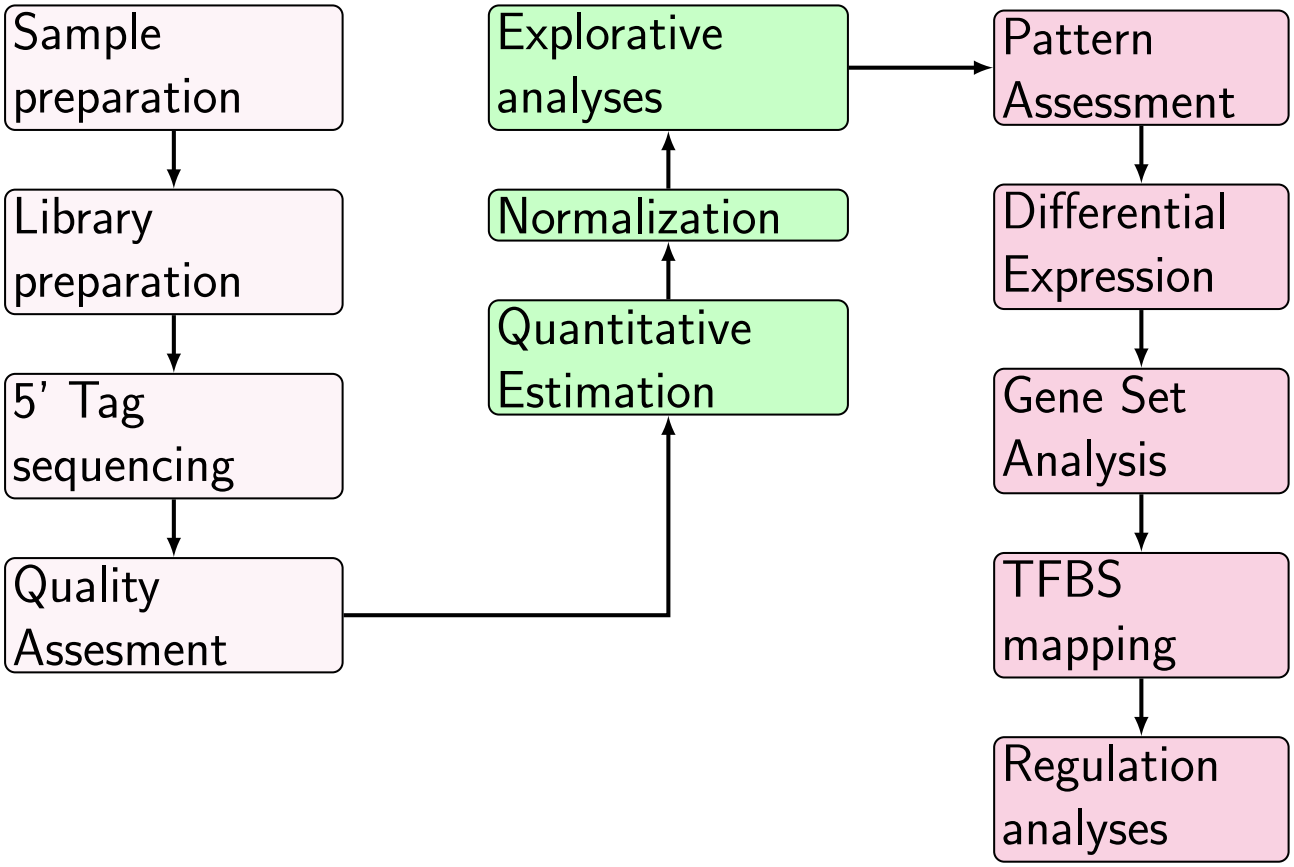

Figure 4.1: *Expression data-oriented framework from biological preparation to computational mining to examine transcriptional regulation*

microarray or raw reads for sequencing assays) to the genes or probes to determine the level of expression. For microarray assays, quantitative estimation determines relative expression whereas the more recent sequencing technologies quantify counts of expressed transcripts. In order to compare libraries, the raw quantified expression levels require normalization. We describe details of these function related analysis of gene expression profiles in the next few subsections.

Exploration and pattern assessment

Expression profile exploration is applied to examine underlying structures in the data based on a researcher's expertise, pre-conception and current knowledge. The most obvious goal is to identify individual genes with similar expression patterns across samples and also to find unknown groups within samples. The initial step is to identify outliers which seem to exhibit uniquely different patterns. Researchers have devised methods to explore profiles by application of distribution examination using tools such as box plots or similar graphics, clustering, principal component analysis and other tools. Such techniques can be used to uncover novel uncharacterized genes for different sample conditions as covered in the experimental design. Many software tools can be used for exploration. Clustering is an example of a commonly used exploratory tool to elucidate resemblances among groups of genes or groups of sample conditions. It is however limited because it cannot be used to determine differentially regulated genes. Clustering is modeled on the assumption that differences in gene expression profiles can be measured as distances eg Manhattan, Euclidian or similar metrics. A researcher must therefore knowledgeably decide how to compute these distances. Different approaches require varyings procedures to extract forms of desired similarities and this often gives different clustering results. Scaling or transformation is also used for this purpose. Appropriate methods for transformation based on multivariate statistics have been devised. Linear scaling is frequently used but is influenced by groups of co-expressed gene sets which is characteristic of common 'house keeping' genes. Log-scaling is an alternative to linear scaling. This can however amplify the noise for genes with low expression levels. Another objective of exploration by clustering is gene filtering. Transcription is a context dependent process and therefore it is required that genes with low signal levels can be selected and removed to remove the uncertainty or noise levels of such genes in the dataset of interest. For most purposes, hierarchical clustering is the most intuitive and computationally inexpensive for small datasets. Other algorithms that

can be applied to gene expression data include k-means and self organizing maps (SOM). The problem with these two is that they require a user to pre-determine the number of clusters to be constructed. Principal components analysis is also an exploratory tool for transcription datasets. PCA can find coherent patterns of genes and also distinguishes groups based on variation.

4.2 An introduction to transcription regulation network analysis

The human genome encodes over 20000 genes and these occupy only about 1.5 percent of the genome, thus leaving the function of a large percentage (about 98.5 percent) of the genome largely unknown. An interesting feature of human genomes is the complexity of instructions that direct gene expression (turning on or off). Generally, there exists similarities of gene expression between prokaryotes and eukaryotes for example in the feature that transcription is carried out by transcription factors which bind the 5' region of genes to initiate transcription with the help of RNA polymerase. This process nonetheless depends on other spatio-temporal factors.

Transcription factors (TFs) are proteins that activate or repress gene expression by binding to specific sequences known as recognition sites that are often close upstream (proximal) or further upstream (distal) to the transcription start site. The specific sequences bound by TFs are defined as motifs (DNA-binding domains) with high affinity for binding to the TF compared to other parts of the genome. These motifs are conserved across species and can be used to stratify known TFs into families, for example SOX proteins. These TFs can also be categorized by structure into groups like zinc finger proteins. TFs are very specific to their motifs. Transcription-factor induction of expression is considered the most important mechanism of transcription regulation in both prokaryotes and eukaryotes. There are cases where genes are regulated at every step but for the majority of genes, regulation normally occur at the point of initation because this is energetically the most energy efficient point to regulate.

In Eukaryotes, transcription regulation is combinatorial machinery involving many complex proteins and genes interactions forming a system commonly called 'transcription regulation networks'. Cellular processes such as differentiation or other phenotypes associated with cells are characterized by unique topologies of transcription regulation networks. To understand the principles behind such processes, it is necessary to engineer ways to determine the identity and

expression levels of transcription factors and target genes and its spatio-temporal interactions throughout the regulation process. Combinatorial complexity of regulation makes elucidation of networks difficult both *in vitro* and *in silico*. Mechanisms of specific transcriptional response for particular transcription factors and target specificity are therefore hard to understand. These key questions are currently hard to directly answer from gene expression quantification technologies such as RNA-seq, CAGE and Microarrays. Recently, there has been progress in development of complimentary methods to address these challenges. Experiments in genomics have contributed to and have been applied for genome-wide expression profiling in combination with chromatin immunoprecipitation (ChIP) to identify protein-gene interactions obviously depending on the experimental design. This has enhanced the use of expression profiling for application in constructing TRNs.

Briefly stated, transcription regulation is a complex process controlled by a combination of regulators such as transcription factors and miRNA amongst other driving factors (Jacob and Monod, 1961). The regulation process is heterogeneous and can vary significantly between samples (Kumar *et al.*, 2014). In cancer for example, some phenotypes such as metastasis issue from regulatory differences relative to non-cancer cells (Ablett *et al.*, 2014).

Many projects including The Cancer Genome Atlas (TCGA) (Weinstein *et al.*, 2013) and Functional Annotation of The Mammalian Genome (FANTOM) (Consortium *et al.*, 2014) aim to profile gene expression of thousands of tumors at various levels thus contributing to further understanding of pivotal regulation elements such as transcription factors. These projects generate vast amounts of data therefore create a need for informatics tools and pipelines for purposes such as expression analyses, transcription regulation network (TRN) mining and visualization of data to leverage interpretation and provide insight into genome-wide transcription regulation processes.

Due to the significance and specificity of motif-transcription factor interaction in controlling gene expression, it is has become increasingly useful to apply computational means to study binding site motifs and to apply them for prediction of generalized TRNs. These motif-based predictions rely on knowlegde about TF-TG preference or could alternatively apply *de novo* elucidation of novel motifs by application of computational methods. In knowledge based prediction, position weight matices (PWMs) are used to scan for unique sequences in the promoter regions. Being a purely *in silico* approach, sequence-based prediction does not consider factors like multi-factor binding or other sequence features like conformation and opposing

or synergistic nature of TFs. In the case of *de-novo* prediction, computational techniques identify short stretches of sequences whose existence cannot be attributed to chance alone (statistical significance). One such technique is MEME (Bailey *et al.*, 2006). When used in combination with expression profiling, the capability of computational prediction of networks can be enhanced by providing more useful filtration of gene networks to suit expression profiles of a given condition or phenotype especially in situations where the promoter of a gene has not been already defined.

The application of expression data and computational determination of interacting transcription factors and binding sites enable context specific elucidation of networks which can be mined in downstream analyses. It has therefore become possible to examine expression profiles to infer hidden patterns of interactions between transcription factors and binding sites. The basic attributes of a TRN can be studied to understand two aspects: First, the network architecture and secondly the key network biomarkers. Network mining involves extracting patterns of interactions, hubs and other crucial nodes that uniquely characterize a network. Computational tools and pipelines for studying gene regulation networks ought to provide clinically important information that can inform gene expression research designs and their application to healthcare needs.

In gene expression data mining, one common study approach is to infer global-scale networks of genes that characterize a phenotype of interest, for example cancer (Qin *et al.*, 2015; Turner *et al.*, 2007; Cordero *et al.*, 2014; Ergün *et al.*, 2007). Recently, omics approaches for instance those linking gene-expression to metabolic pathways are becoming popular for systematizating expression data to identify key regulators, patterns and biologically meaningful network biomarkers. This is especially because most experimental designs in gene expression research are often comparative in nature, matching sample conditions. (e.g variant to normal or mutant to wildtype) (Yao *et al.*, 2015). Most workflows in this regard however treat gene expression data mining and complementary analyses such as exploration and differential expression analysis as separate facets from genome-scale transcription regulation network construction and mining (Allison *et al.*, 2006; Saeed *et al.*, 2003). One such example is the differential expression analysis and regulation network construction which are often considered as two sets of analysis. While differential expression is aimed at extracting significantly upregulated or downregulated genes and probes, expression network construction is aimed at inferring regulatory interactions between genes and transcription factors. This abstraction is largely due to perceived independence

of these network-mining steps. Several studies take an integrative approach but often limited to unsupervised learning algorithms such as hierarchical clustering which may lack sufficient prior biological foundation therefore making their results less interpretable (Chipman *et al.*, 2003). In addition to these challenges, there is still a need for methods that merge network construction, mining, gene expression data integration, pathway-specific visualization and simplified inference deduction.

Data-oriented extraction of TRNs has been shown by methods such as those developed by (Margolin *et al.*, 2006). A particularly interesting question is whether heterogenous steps involved in TRN inference can be integrated analytically in order to give a coherent view of gene regulation across datasets. Systems for TRN data mining for example cytoscape often require a user to prepare their network in advance or can use plugin components to extract pre-composed networks. To the best of our knowledge, only a few methods combine the data-based network inference with other network annotation and mining techniques in their work flows. A few studies have attempted to combine analyses of gene expression data. Isik *et al.* (2012) for example combined the use of microarray data, ChIP seq data and pathway enrichment methods to develop a simulation algorithm for cellular signalling based on ranking of gene-scores. While this method is integrative in nature, it does not explicitly identify gene regulation mechanisms on the presumption of transcription factor driven regulation. A separate study by Kutmon *et al.* (2013) presented a system that allowed explicit extension of TRNs by addition of regulatory interactions which can be automatically retreived from third party databases. This is certainly an improvement on the previous approach but still lacked a more targeted knowledge-based network mining strategy.

To further enhance value derivation from data, knowledge-based leverage of TRN analysis is required. Earlier attempts as shown by 'plugin' applications such as CytoKEGG were the best attempts at integration of biologically knowledge for transcriptome data. In that framework, a user is able to define pathway data from expression profile and also trace pathways that a gene is involved in. It does not apply to gene regulation in itself but it captures the whole picture of interacting genes (including regulatory interactions).

In this work, we introduce TransReguloNet; a computational pipeline that integrates gene expression profiles with knowledge-based and predicted transcription factor-target gene (TF-TG) interactions to elucidate expression data-specific TRN architecture. We incorporated stratification of TRNs based on functional units (pathways) to enhance biological interpretation. The rationale

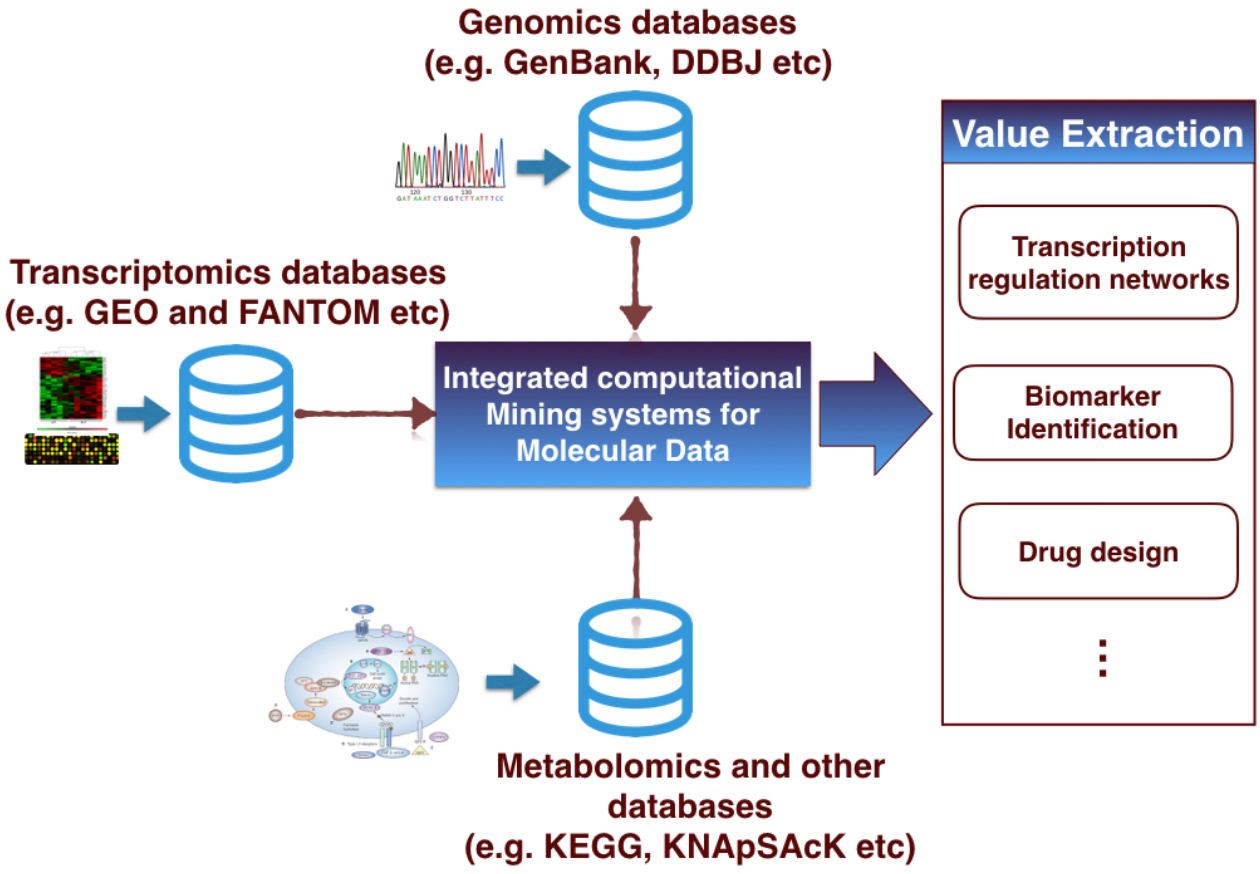

Figure 4.2: *data-oriented integrated computational mining pipelines to enhance value extraction from transcriptomics datasets*

behind the study design is incorporation of TRN stratification based on functional units (pathways) to enhance and simplify biological interpretation. This enables objective examination of gene sets in expression data for unique patterns and influence of regulators on genes within known functional roles. For this purpose we utilized published human KEGG pathways for modularization of trancription regulation networks. The proposed flow is however flexible for use with other pathway definitions such as the Gene Ontology pathways or other user-defined pathways. In addition to the workflow, we developed a user-friendly tool (TransReguloNet) for pathway-specific visualization, pathway-based mapping of transcription factors influence on their target genes and cluster visualization of pathway-specific expression profiles. This was implemented as a web application.

In the present study, we also addressed the systematic understanding of gene expression datasets of various types of cancer including lung, breast and prostrate cancer to examine the usability of the TransReguloNet application. This pipeline is however flexible and scalable for application to other transcriptome profiles. Lung cancer is known for salient phenotypes that can be linked to unique gene regulation mechanisms (Fu *et al.*, 2015; Saito *et al.*, 2013). These tumors are hypothetically regulated uniquely by certain transcription factors, therefore functional molecular pathways are likely to be influenced differently by transcription factors. Cancer cell pathways may not be entirely unique but may be defined by abberant characteristics in critical roles such as energy derivation which differentiate them from functional pathways in normal cells. Using the proposed framework, we sought to find out peculiarities of gene regulation in cancer relative to normal cells in some instances and in drug related comparison in some instances, from the viewpoint of TF-TG interactions at a pathway level. The datasets used were from different expression quantification technologies, in particular microarray and cap analysis of gene expression (CAGE) (Kodzius *et al.*, 2006). Studies have shown that pathway-level modularization of networks improves ease of deriving value from expression data (Steinfeld *et al.*, 2015). In the present work, we target the problem of insufficient integration of gene expression data in network mining. We propose a workflow that includes a user-friendly visualization platform.

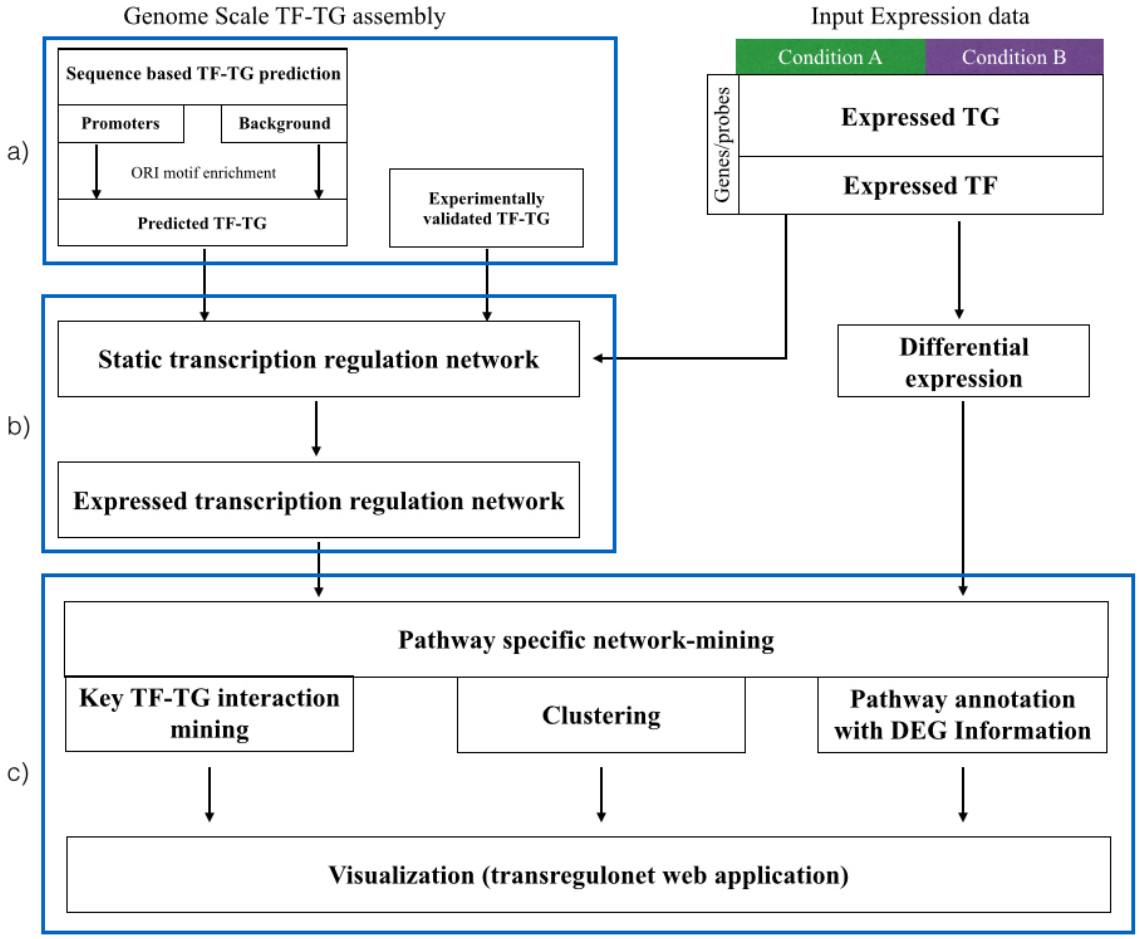

Figure 4.3: *TransReguloNet Workflow: (a) An initial static transcription regulation network is constructed and (b) filtered to represent the nodes present in the expression data. The current implementation of the TransReguloNet pipeline assumes comparative expression profiling (such as tumor vs normal or condition A vs condition B). Differential expression is initially computed and input into the network mining flow. (c) KEGG pathway abstraction of the expressed TRN is applied to enhance functional intepretation. Subsequent steps include network annotation with differentially expressed genes, clustering and quantification of interaction strength between TFs and TGs. A web application (TransReguloNet) encompassing these steps was developed to enable visualization*

4.3 Materials and methods

4.3.1 Summary

A general overview of the TransReguloNet pipeline is illustrated in Fig.4.3. This system was tested on gene expression datasets related to cancer (Table 4.1). Descriptions of the data include the number of samples, the experimental design and the corresponding expression profiling technology.

Six datasets namely, GSE43346 (Sato *et al.*, 2013). GSE10245 (Kuner *et al.*, 2009), GSE19804 (Lu *et al.*, 2010), FANTOM5 (Consortium *et al.*, 2014), GSE38376 (Komurov *et al.*, 2012) and GSE55945 (Arredouani *et al.*, 2009) were examined. The datasets evaluated encompassed different variants of cancer including small cell lung cancer (SCLC), non-small cell lung cancer (NSCLC) and squamous cell lung cancer (SCC), breast cancer and prostrate cancer. For instance GSE43346 compares expression between normal and small cell lung cancer whereas the FANTOM5 data consisted of 19 lung cancer lines and 264 non-lung cancer cell lines. Data was obtained from the Gene Expression Omnibus (GEO) database (Edgar *et al.*, 2002) and FANTOM project database (Kawaji *et al.*, 2009), both of which accumulate data derived from various transcription quantification technologies. FANTOM 5 project accumulates mostly CAGE technology derived mammalian gene expression data while GEO stores expression profiles from diverse transcriptome technologies including micro-array and sequence-based expression quantification methods. In the pre-processing steps, micro-array based datasets were quantile normalized whereas trimmed mean of M-values (TMM) normalization procedure was applied to the CAGE-based data. Differential expression analysis was performed under the assumption of negative binomial (NB) distribution (Robinson *et al.*, 2010) for count data whereas quantile-adjusted conditional maximum likely-hood (qCML) exact test was performed for the microarray based datasets.

Table 4.1: Cancer related gene expression datasets analysed

DATA ID	SOURCE	DESCRIPTION	CONDITIONS	SAMPLES	TECHNOLOGY
GSE43346	GEO	Small cell lung cancer (SCLC)	42 Normal 26 clinical SCLC	68	Microarray (Affymetrix)
GSE10245	GEO	Non-small cell lung cancer (NSCLC)	2 Disease States (Adenocarcinoma (AC) and Squamous cell carcinoma (SCC))	58	Microarray (Affymetrix)
GSE19804	GEO	Non Small cell Lung cancer in female smokers	Paired Tumor- Normal	120	Microarray (Affymetrix)
FANTOM5	FANTOM5	Human cultured cell lines	Lung cancer v. Non Cancer cells	282	Tag based sequencing (CAGE)
GSE38376	GEO	Breast Cancer Lapatinib drug response	Lapatinib - Control	18	Microarray (Illumina HumanHT)
GSE55945	GEO	Prostrate cancer dataset	Prostate cancer- Normal	21	Microarray (Affymetrix)

4.3.2 Whole genome static transcription regulation network assembly

A static regulatory network of the human genome was derived as an amalgamation of TF-TG interactions from two sources. Firstly, knowledge-based interactions from experimentally validated and published sources were obtained from human transcription regulation interaction database (HTRIdb) Bovolenta *et al.* (2012). This dataset consisted of 284 unique transcription factors and 52467 interactions. The experimentally verified interactions in the HTRIdb repository are detected by small and mid-scale techniques such as chromatin immunoprecipitation, concatenate chromatin immunoprecipitation, CpG chromatin immunoprecipitation, DNA affinity chromatography, DNA affinity precipitation assay, DNase I footprinting, electrophoretic mobility shift assay, southwestern blotting, streptavidin chromatin immunoprecipitation, surface plasmon resonance and yeast one-hybrid assay and chromatin immunoprecipitation coupled with microarray (ChIP-chip) or chromatin immunoprecipitation coupled with deep sequencing (ChIP-seq). The second category of TF-TG interactions consisted of predictions from sequence-based string matching of transcription factor binding sites (TFBS) considering 1000 bases upstream of over 18000 human genes. For known human transcription factors, position frequency matrices corresponding to each binding site motifs were obtained from MotifDB (Shannon *et al.*, 2013). Coordinates of transcripts of all genes in the human genome were obtained from the *TxDb.Hsapiens.UCSC.hg19.knownGene* annotation package. Predicted TFBS in promoter sequences were evaluated against a background sequence set by motif enrichment analysis using a modified over-representation index (ORI) (Bajic *et al.*, 2003) measure illustrated in the equation:

$$ORI(TF)_{ig} = \frac{Counts(TF)_{ig}}{Counts(TF)_{ib}} \times \frac{Total(TF)_{iG}}{Total(TF)_{iB}}$$

Where $ORI(TF)_{ig}$ represents the ORI score of the binding site of transcription factor i upstream of gene g. $Counts(TF)_{ig}$ is the number of predicted binding sites of the i^{th} transcription factor upstream of gene g while $Counts(TF)_{ib}$ is the number of binding sites predicted in the random background sequence downstream of a gene. $Total(TF)_{iG}$ is the overall count of all possible binding sites of the i^{th} TF in the promoter sequence set G whereas $Total(TF)_{iB}$ represents the overall count of binding sites in random background sequence set B. Significant ORI scores were used to select predicted TF-TG interactions. The idea of ORI score enrichment in the sets compared (promoters and background sets) is to identify transcription factor binding sites that are most distinct in promoters relative to the background set. A simpler intuitive explanation of this metric is that if the number of sequences in a population

of sequences is N, out of which M are linked to a transcription factor binding site $TFBS$, it is possible to select K sequences randomly from the whole sequence set without replacement and the probability of getting x motifs (sequences) associated with a $TFBS$ in K can be modeled as an hypergeometric distribution (Johnson et al, 2005). ORI provides a numerical index that associates presence of a TFBS in a promoter region as an indicator of enrichment. The genome scale static TRN was constituted by combining the experimentally validated and computationally predicted interactions.

4.3.3 Network mining

Pathway-based network stratification In the workflow implemented here, the static network which consists of both predicted and validated TF-TG interactions was first reduced to the set of nodes (genes and transcription factors) expressed in the specific expression dataset being analyzed. The subsequent steps of network mining therefore focus on only the nodes expressed in the expression profile. This network is then modularized into subnetworks based on genes corresponding to KEGG pathways (Kanehisa and Goto, 2000), thus representing functional units that can be independently examined. Other modules such as GO gene sets or user defined gene sets can be used but the current implementation only defines subnetworks of KEGG pathways.

Quantifying TF-TG interaction level The level of interaction between TFs and TGs within the pathway-based sub-networks was evaluated by use of mutual information(MI). Mutual information rates the statistical dependence of TF expression on that of its target genes. More precisely, given a joint probability distribution between a TF and TG; $P(TF, TG)$, the mutual information between TF and TG; $MI(TF, TG)$, is defined as:

$$\mathrm{MI(TF, TG)} = \sum_{tf \in TF} \sum_{tg \in TG} \mathrm{P(tf, tg)} * \log \frac{\mathrm{P(tf,tg)}}{\mathrm{P(tf)P(tg)}}$$

Intuitively MI can be defined as the distance between the joint density of a transcription factor tf and a target gene tg and the product of their individual densities. MI can quantify non-monotonic relationships (including causal association). MI score quantifies the strength of interaction between the TF and TG. This score is non-negative and yields 0 if the TF and TG are statistically independent. Compared to other associative metrics like correlation, MI tends to be more general in scope and measures reduction of uncertainty in dependence of expression

of a gene tg on a transcription factor tf. Normalized Mutual Information (NMI) is an intuitively intepretable standardized MI score ranging from 0 to 1 (Romano *et al.*, 2014).

$$\mathrm{NMI_{ij}} = \frac{MI_{ij} - \min(MI)}{\max(MI) - \min(MI)}$$

Where NMI_{ij} is the standardized MI score of i^{th} transcription factor and its j target gene. NMI was applied in annotation of the TF-TG edges of the TRN in the TransReguloNet application. To further enhance the meaningfulness of this score, we comparatively evaluated strengths of Transcription factors to its target genes between the sample conditions. The concept of mutual information developed by Shannon (1948) for application in communication theory. MI can be applied to measure coherence in two datasets. Consider for example a gene TG and a transcription factor TF, both of whose expression quantities have been measured by an expression profiling assay. They may not necessarily ne independent of each other (ie, their expression may be related in the sense that TF may be function in regulating TG). MI values are non-negative but without an upper limit in the amount of information. For intuitive application in the TransReguloNet system we normalized MI values using a well known statistical strategy known as range normalization. This technique is for bringing all the data points into proportion with one another. MI and NMI were considered for application in TransReguloNet for two reasons:

- Non-negative score of information content
- Unexpected expression values (for instance outliers) have high information content due to their low probability and conversely frequent expression values have low information content due to their high probability

Pathway-based clustering Clustering was performed by standard hierarchical clustering analysis (HCA) for each subset of the expression data corresponding to the specified KEGG pathway. Application of HCA to expression data has been discussed in detail by Sturn *et al.* (2002). Briefly, we applied it to examine between sample condition expression patterns.

4.3.4 Visualization

Some datasets (see Table 4.1) are loaded in the current web implementation of our application (http://transregulonet.naist.jp). Briefly described, these datasets represent expression profiles related to various lung cancer forms. We developed a user-friendly web application tool to

visualize networks of functional pathways of these profiles. Inputs into the visualization application include a static regulatory network, actual expression data and the differential gene expression output (based on the experimental design of each dataset). The static network is initially reduced to represent the nodes expressed in the dataset. Subsequently, human functional pathways are retreived automatically from KEGG database. Network nodes and edges are color-coded based on differential expression analysis. Node shapes also distunguish transcription factors from genes. In addition, edges are color-coded to differentiate computationally predicted interactions from validated interactions. Networks of a chosen transcription factor and its level of interaction with target gene association can also be visualized comparatively between the sample conditions. In addition to this, hierarchical clustering heatmaps of pathway-specific subsets of the expression data can be visualized.

4.3.5 Pathway enrichment analysis

One of the key objectives of pathway analysis is to determine to enriched gene sets that works as a unit. We utilized GAGE (generally applicable gene enrichment) algorithm (Luo *et al.*, 2009). In the datasets, the gene sets are defined as the KEGG gene categories corresponding the pathways under investigation (Supplementary table A.1). These pathways have been curated independently. One of the advantages of GAGE algorithm is the separation of each gene set as either cannonical pathways or experimentally derived differential expression pathway sets. Using the experimental set, the test statistic (t-test) is computed based on the average per-gene tests. For the canonical pathway set, GAGE calculated the average absolute values of the per-gene statistics to measure both up regulation and down regulation. In order to test for significance (ie whether a gene set is significantly correlated to a certain sample condition defined in the experimental design), fold changes of gene expression levels are examined for all the sample conditions represented in the datasets (eg. cancer v. normal). The corresponding hypothesis is whether the mean FC of a gene set is significantly different between the sample conditions. The model t test function used is:

$$t = \frac{(m-M)}{\sqrt{\frac{s^2}{n}+\frac{S^2}{n}}}$$

Where m, s and n are the mean fold change (log ratio of expression levels), standard deviation, and number of genes in a particular gene set, and M and S are the mean fold change and standard deviation for all of the genes in the dataset. This is a classic two sample t-test

between sample conditions

For the case of multiple libraries as represented in our dataset, GAGE derives multiple t-statistics and p-values from inter-library comparisons. Subsequently, a global p-value is derived by combining individual p-values. The test assumption is that these Individual p-value follow a uniform distribution. The algorithm then estimates a meta p-value by the following function:

$$x = -\sum_{k} \log p_k$$

GAGE analysis assume that p-values come from independent comparisons. However, this may not be true for gene expression data or unpaired experiments with k number of samples with one condition (eg cancer) and l number of samples with the other condition (eg Normal). A global p-value is derived as:

$$x = -\frac{1}{L}\sum_{k} \log p_k$$

The resulting p-value is corrected for false discovery rate (FDR).

4.4 Results and discussion

4.4.1 Overview

As illustrated in Fig.1, expression profiles are initially subjected to differential gene expression analysis. This step defines genes as upregulated, down-regulated or unchanged. This is then applied in downstream annotation of transcription factors and target genes during the network visualization step. The workflow then filters the genome-scale static TRN to obtain a subset representing only the nodes 'expressed' in the profiles before subsequent further reduction of the network systematically into subnetworks of genes corresponding to the defined human KEGG pathways (supplementary Table A1). The flow also includes a step of quantification of TF-TG interaction levels in each profile by MI (see methods). Mutual information(MI) is applied to examine the influence of each TF on its respective target gene where the resulting MI score is an indirect measure of interaction strength based on the expression density of a TF relative to its target gene. An integrated platform for visualizing these steps was implemented as a web usable application (available online.

4.4.2 Transcription regulation network construction

TRANSFAC database (Matys *et al.*, 2003) accumulates information about transcription factors and their target genes. Genome-scale mapping of transcription factor binding sites (TFBS) was performed by sequence-based motif matching using known position weight matrices (PWM) of 128 TFs from TRANSFAC. 1000 base pairs upstream was defined as the promoter region. Instances of predicted motif presence were assessed by evaluation in comparison to occurrence in background sequences (non-promoter regions) and an enrichment score (Bajic *et al.*, 2003) computed to determine significance of predicted interactions. In total 522184 interactions were determined as significant. In addition, experimentally validated interactions were obtained from the HTRIdb (Bovolenta *et al.*, 2012). 284 unique transcription factors regulating over 18200 genes with 52467 interactions were merged with predicted interactions to create a whole genome static TRN.

4.4.3 Visualization and intepretation

To the best of our knowledge, few software applications offer the functionality of visualizing data specific TRNs by pathway abstraction directly from gene expression data. Cytoscape (Shannon *et al.*, 2003) for example requires a user to input pre-constituted networks and is therefore not versatile to automatically learn data-specific or pathway-specific transcription regulation networks. To generalize the comparison between TransReguloNet and the commonly used network mining tools such as Cytoscape and BioLayoutExpress3D, we have included the summarized comparison in Fig.4.5. The proposed TransReguloNet application takes advantage of the available resources such as KEGG pathways and publicly available expression datasets to understand regulation patterns in different conditions. Starting from a TF-TG based TRN, the initial static network is filtered to represent only the nodes present in the expression data set. Innovative features of the pipeline include the integrated approach and profile modularization based on known functional units. For visualization, the nodes in the constructed TRN are annotated using differential gene expression analysis.

Fig.4.4 illustrates the visual interface for a user chosen dataset and pathway (in this case GSE43346 and Glycolysis/gluconeogenesis respectively) visualized using TransReguloNet. Differential expression information for each node is annotated by use of different colors thus a user can visually extract expression patterns in a pathway of interest. Red nodes represent upregulated genes or transcription factors, whereas blue nodes represent down-regulated genes or transcription

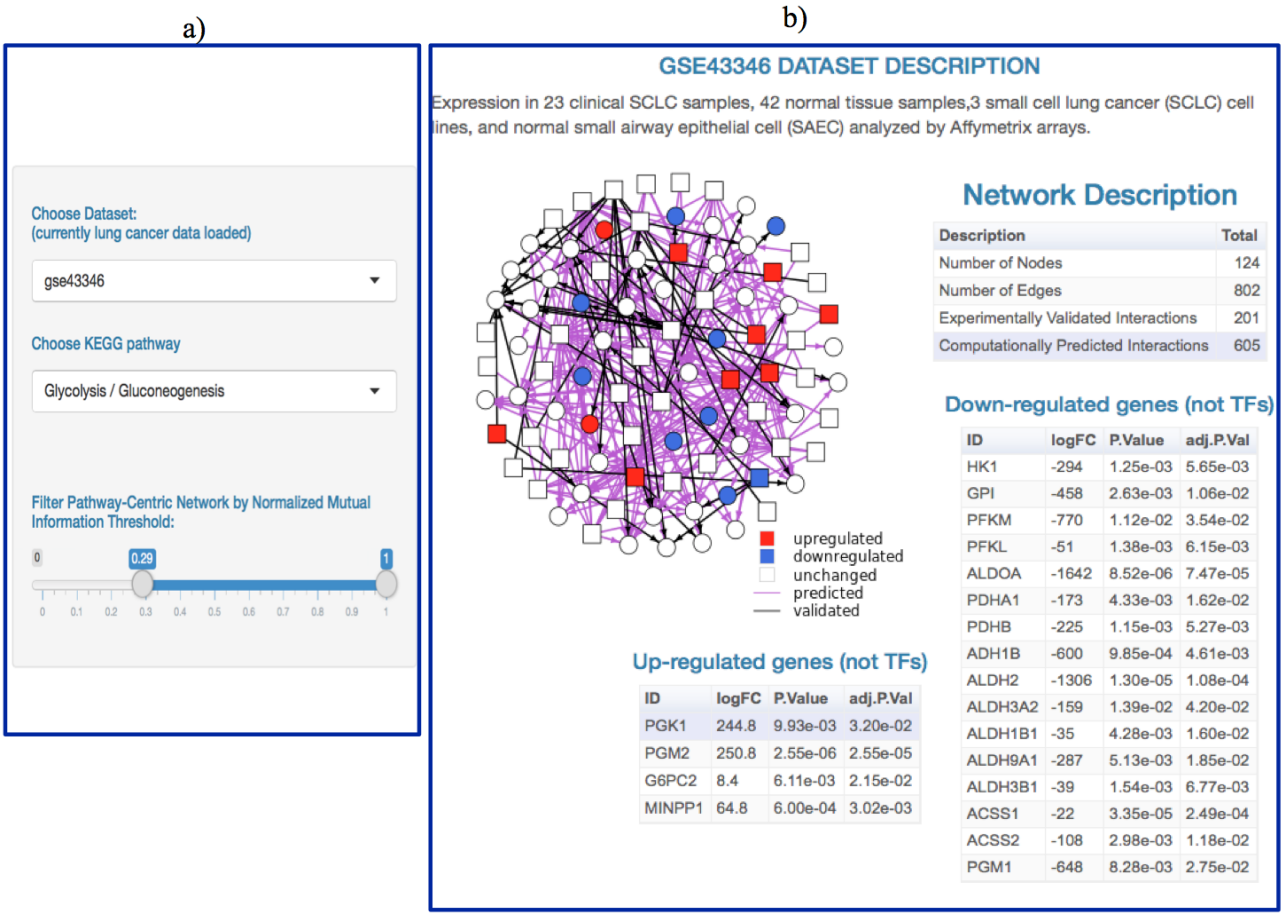

Figure 4.4: *Pathway modularization: a) The 'Pathway Network' page in the TransReguloNet interface, consists of a side panel where a user select a dataset and KEGG pathway of interest. It also consists of sliding bar ranging from 0 to 1, to filter edges by the interaction strength as determined by the normalized mutual information score enabling a user to reduce network size nodes based on desired threshold of interaction strength. b) The main panel consists of a description of the chosen dataset, a visual network of the selected pathway annotated with differential expression information and network metadata such as number of nodes and interactions within the pathway specific network. Tables listing upto a maximum of top 20 downregulated and top 20 upregulated genes within the pathway network are also output for easier user intepretation.*

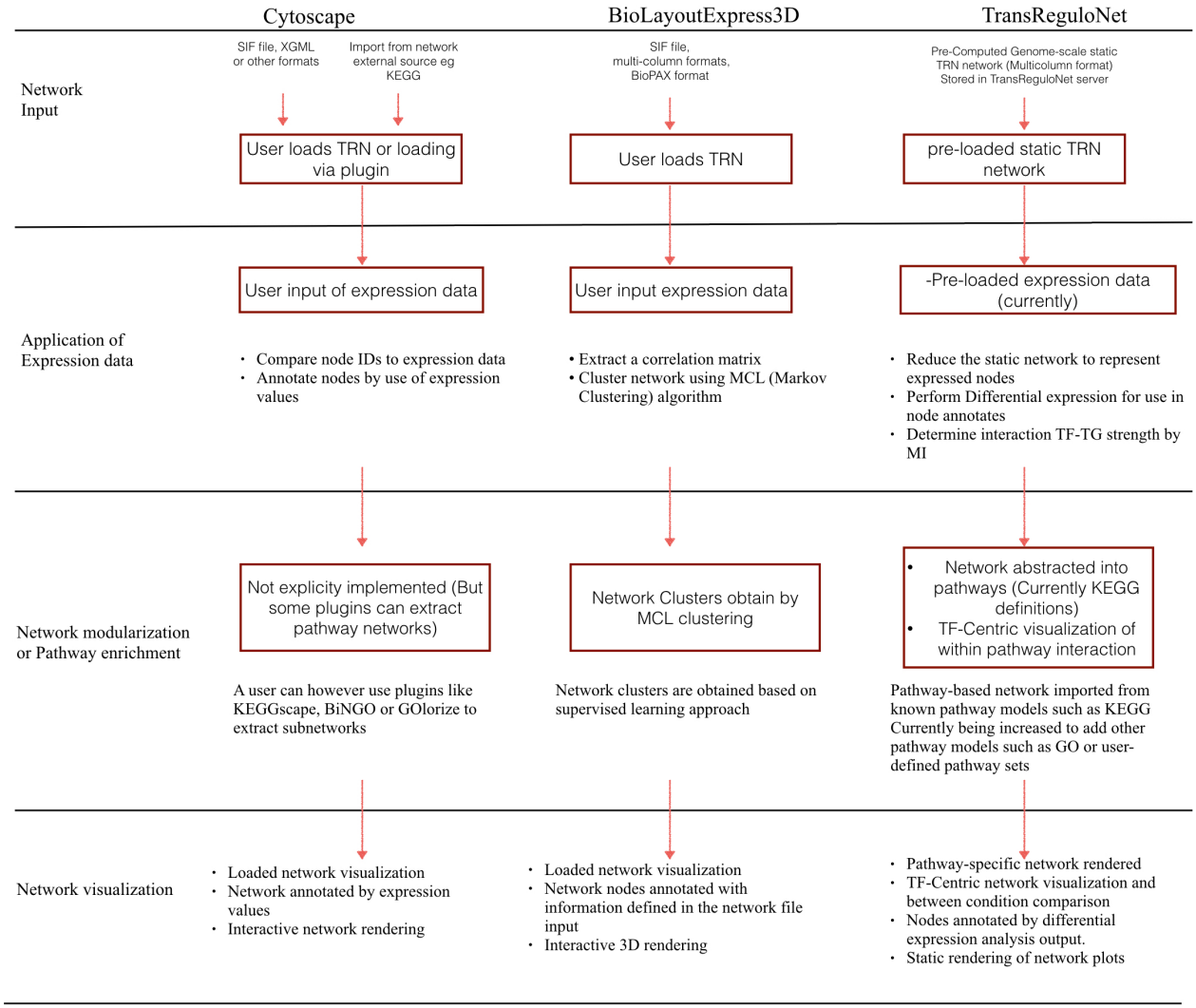

Figure 4.5: Comparing TransReguloNet to cytoscape and BioLayoutExpress3D

factors. Furthermore, the shape of nodes distinguish TFs from TGs. Rectangular nodes represent transcription factors and circular nodes represent target genes. In addition to this, the edges denoting TF-TG interactions are also color coded such that black arrows denote experimentally validated interactions as obtained from HTRIdb (Bovolenta *et al.*, 2012) whereas the magenta colored arrows denote interactions predicted by sequence-based motif matching for transcription factors whose position weight matrices have been determined and whose ORI score met the threshold of high significance. Fig.4.6 shows comparative networks for a chosen transcription factor and its target genes in the different conditions (eg normal vs lung cancer in this case as determined by the expression profile design).

To illustrate one of the ways in which the pipeline can be used for studies, we compared the effects of certain transcription factors in particular pathways for the different sample conditions covered in the expression profiling assay. Fig.4.6 shows comparative networks for a chosen transcription factor and its target genes in the different conditions (eg normal vs cancer in this case as determined by the expression profile design). For the case of GSE43346 dataset in the illustrated example, interactions of the PRDM14 transcription factor in the glycolysis/gluconeogenesis pathway is compared in lung cancer and normal samples. The PRDM14 node is at the center and is radially bound to target 9 genes, which have been experimentally validated as shown by the black edge color. Two target genes ACSS2 and PGM1 are up-regulated as shown by the red node colors. In addition to this, the binding strength of the TFs to its target is quantified by the normalized mutual information score annotating the edges. In each of the two cases of the up-regulated genes ACSS2 and PGM1 in the example, there seems to be a marginal increase in the association strength in the lung cancer samples relative to the normal samples from 0.036 to 0.0512 and 0.0095 to 0.0171 respectively, this may indicate a peripheral role in the differences between glycolysis pathway in tumors v. normal cells but this is insufficient evidence to draw such a conclusion. This however provides a basis for new hypothesis creation for further experiments and provides avenues for new research design for experiments focusing on differences in glycolysis or other energy metabolism pathways and other hallmarks of cancer Hanahan and Weinberg (2000).

Another example utility of TransReguloNet illustrated in Fig.4.7 is the hierachical clustering exploration of pathway-specific subset of the expression data.The columns in the output heatmap represent the samples and the color bar differentiates the sample conditions represented in the data. In the specific example shown in Fig.4.7, the blue portions on the color bar correspond

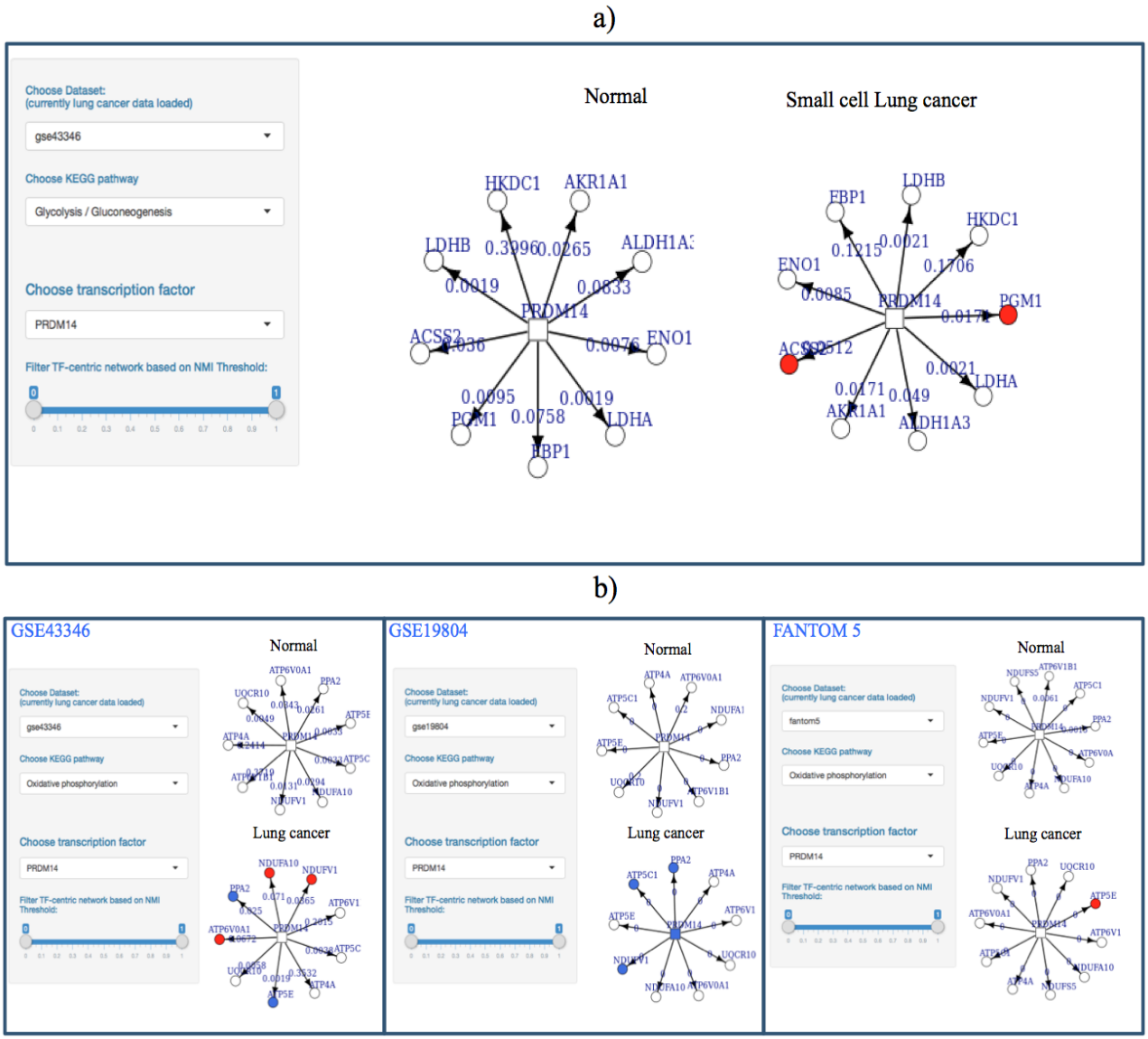

Figure 4.6: *Comparing TF centric subnetworks a) The figure illustrates the TF-centric visualization of interactions and the strength of the association to its target genes quantified by normalized mutual information (NMI) shown as the edge weight. The sidebar showing the chosen dataset and the pathway in focus and an extra drop-down menu of the TFs linked with the genes in the selected pathway. The main panel contains plots corresponding to sample conditions represented in the expression profile. The NMI is calculated independently for the conditions represented in the dataset. The example presented for instance shows a plot of PRDM14 transcription factor between matched lung cancer v. normal condition. b) Oxidative phosphorylation pathway is one of the aberrant mechanisms in cancer. Using TransReguloNet, we compare the PRDM14 transcription factor in three datasets, GSE43346, GSE19804 and FANTOM5*

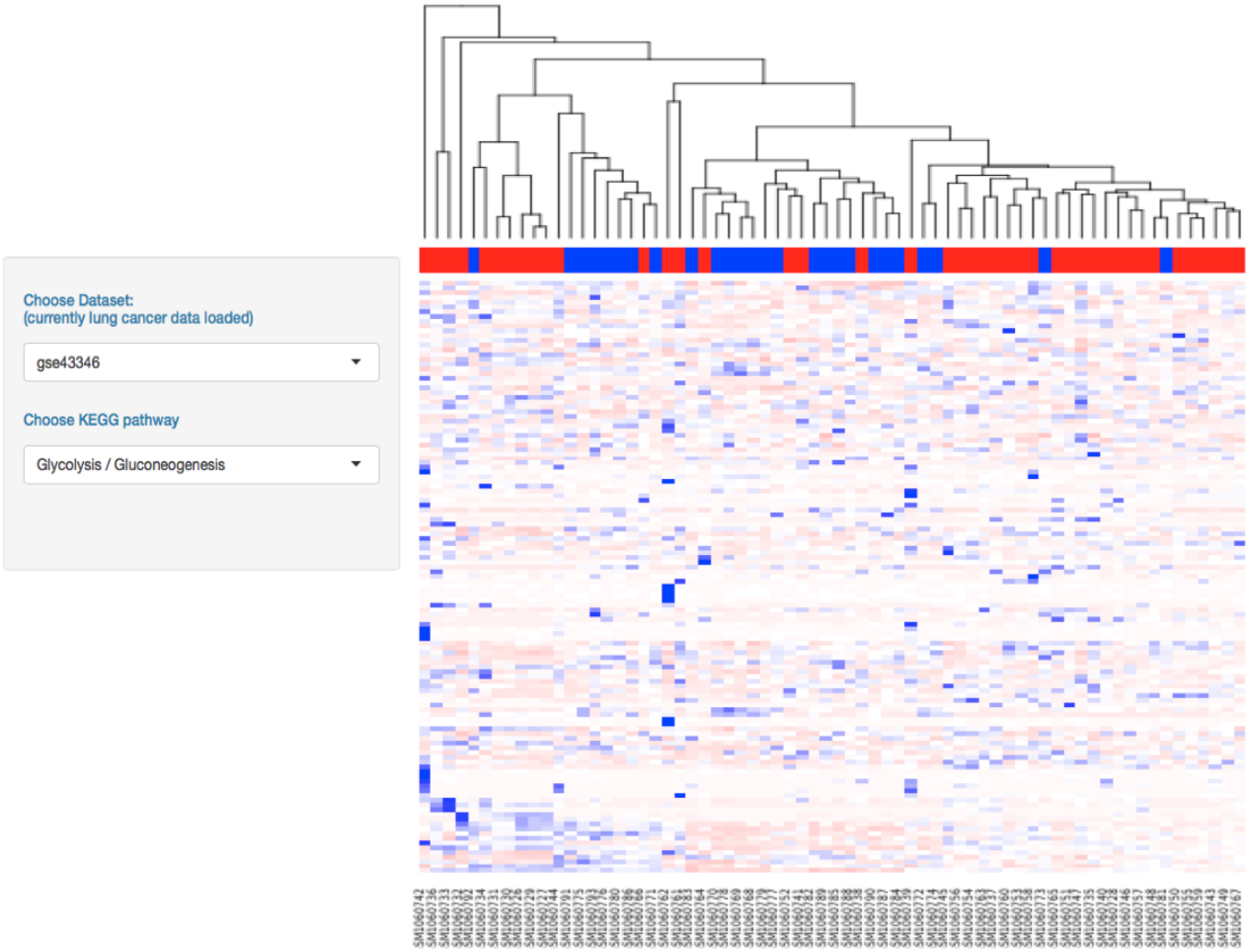

Figure 4.7: *Pathway specific expression data clustering: Heatmap plot derived from hierachical clustering of the expression data subset corresponding to the genes within the pathway being analysed. The blue and red colored frame on top shows sample conditions. The blue bars correspond to SCLC samples on the gse43346 dataset and the red columns correspond to the normal cell samples in the same dataset*

to SCLC samples in dataset GSE43346 (Table 4.1) while red column portions correspond to the normal cell samples in the same dataset. One of the hallmark features of cancer is reprogramming of energy metabolism into the inefficient aerobic glycolysis (Warburg *et al.*, 1956). Due to this so called 'Warburg effect', it is expected that expression of gene sets related to the glycolysis pathway can clearly distinguish cancer cells from normal cells. This generalized heatmap illustration as can be visualized in the TransReguloNet application does not clearly show the expected pattern although it provides an exploratory visualization to understand the general features of the genesets associated with the pathway as captured by the GSE43346 assay.

Such example application of TransReguloNet in studying expression profiles of various cancer forms demonstrate that obtaining biological knowledge from regulation networks is often complicated by large network sizes unlike the objectivity achieved by modularization implemented in this workflow. Studies have attempted to infer network structure and function often based on de-novo strategies such as un-supervised learning for example clustering (Monti *et al.*, 2003; Medvedovic *et al.*, 2004; Ouyang *et al.*, 2004).

A more empirical approach however is to apply well-defined pathways from sources such as KEGG, GO or other pathway models such as BioCyc as means of TRN abstraction into manageable and interpretable functional modules. In the current implementation, we have only implemented KEGG pathway gene sets. Overall, 295 pathways from KEGG can be examined using our application. To augment the usability of TransReguloNet with other pathway models, we further performed separate pathway enrichment tests to examine 3 datasets namely FANTOM5, GSE43346 and GSE19804 (Table 4.1) whose assay design compared normal to lung cancer samples. Here, we examined pathways to determine significantly altered sets of genes in 3 of the expression profiles.

A pivotal goal of health bionformatics research is to define cellular behavior and disease. Gene expression profiling provide a fundamental starting point to this cause. An Ideal framework to contextualize gene expression patterns with functional pathways is to examine systems level coordination of gene signalling processes. As such a systems level approach gives an informative view (compared to gene-based analysis) of the behaviours. This therefore allows a researcher to understand the bigger picture of molecular function thus improving the precision of function targeting.

Pathway enrichment is popular strategy for gene expression data analysis based on pathway

network mining.This method focuses on sets of related genes and has some crucial advantages: Pathway enrichment methods are better at filtering relevant molecular information and a solid intepretation regardless of expression data types. It also uses the entire gene expression data, rather than pre-filtering only differentially expressed genes. This therefore extracts, small coordinated gene expression changes which may have causal effects in determining cellular behaviour. Pathway enrichment also uses knowledge-based definition of biological pathways such as KEGG, GO or other expertly curated pathway gene models.

Interestingly, for our study, the pathway enrichment tests on 3 of the datasets matching tumor and normal phenotypes, i.e FANTOM5, GSE43346 and GSE19804 yielded 23 pathways identified as significantly upregulated or down regulated as summarized in Table 4.2.

The 23 pathways in this table describe pathways that were found to be commonly enriched in the three datasets namely FANTOM5, GSE43346 and GSE19804 whose experimental design constituted matched tumor-normal comparison. The statistical significance from the pathway enrichment tests is shown by the p-value and q-values where the threshold of significance was set at $q-\text{value} < 0.05$. The set size is the number of genes directly or indirectly linked to the named pathways. The results suggest an inherently similar trend of upregulation and down-regulation patterns despite the distinctness of the datasets.

Summarized contribution of the TransReguloNet System

1. Data-Oriented inference and biological objectivity

TransReguloNet implements a filtration mechanism for selecting a genome-scale network based on only the nodes represented in a gene expression dataset. Unlike total reliance on a static whole network, this approach enables examination of relevant interactions within a specified dataset. Biological relevance is also refined by focusing on modular subsets such as the functional molecular pathways as exemplified by the use of KEGG pathway gene sets in the current implementation.

2. Integrated methodology in a single platform

When studying a specific gene expression data set, the application of such a unified pipeline consisting of differential expression, normalization, pathway specific visualization and clustering in a single platform is convenient for a biological researcher with limited interest in technical aspects of such analytical methods. TransReguloNet facilitates objective and qualitative refining of biological hypotheses centered on TF-oriented gene regulation mechanisms.

Table 4.2: Commonly enriched pathways in three lung cancer expression profiles

	GSE43346GSE43346			GSE19804GSE19804			FANTOM5FANTOM5		
Pathway	set size	p-value	q-value	set size	p-value	q-value	set size	p-value	q-value
Ribosome	29	1.53E-14	1.90E-13	44	1.22E-04	1.52E-03	21	8.53E-03	1.06E-01
Spliceosome	23	7.47E-13	7.88E-12	123	1.22E-03	1.29E-02	11	2.34E-02	2.47E-01
Huntington's disease	22	1.17E-20	2.92E-19	73	1.84E-11	4.47E-10	145	3.22E-08	8.01E-07
Proteasome	123	5.07E-17	7.72E-16	112	9.93E-06	1.49E-04	72	2.66E-03	4.04E-02
RNA transport	128	6.47E-14	7.38E-13	14	3.80E-04	4.34E-03	82	1.58E-02	1.80E-01
Alcoholism	36	9.35E-47	1.14E-44	126	7.00E-39	9.59E-37	125	2.23E-13	3.06E-11
Ribosome biogenesis in eukaryotes	48	5.48E-14	6.53E-13	162	2.25E-04	2.67E-03	42	1.47E-02	1.75E-01
Aminoacyl-tRNA biosynthesis	123	2.46E-42	1.69E-40	43	1.15E-26	7.85E-25	181	5.41E-10	2.75E-08
DNA replication	150	1.87E-23	5.69E-22	99	4.60E-13	1.40E-11	145	5.79E-09	1.76E-07
Protein export	68	4.43E-15	6.07E-14	68	1.71E-05	2.35E-04	166	4.95E-03	6.78E-02
Focal adhesion	110	2.38E-52	6.53E-51	91	4.11E-14	1.13E-12	238	6.37E-04	1.74E-02
Amoebiasis	129	6.69E-58	3.67E-56	252	1.03E-15	4.84E-14	85	1.85E-04	1.02E-02
Chemokine signaling pathway	116	3.40E-56	1.33E-54	171	1.90E-15	6.65E-14	128	4.40E-04	1.54E-02
Protein digestion and absorption	152	1.44E-39	2.08E-38	73	2.84E-11	4.10E-10	109	2.36E-03	3.39E-02
cAMP signaling pathway	253	8.68E-58	3.96E-56	196	1.06E-15	4.84E-14	194	2.73E-04	1.25E-02
Platelet activation	80	5.88E-44	9.47E-43	200	1.99E-12	3.21E-11	156	2.15E-03	3.39E-02
Staphylococcus aureus infection	134	1.06E-33	1.08E-32	116	1.49E-10	1.44E-09	41	3.49E-03	3.55E-02
Salivary secretion	36	5.94E-35	6.51E-34	80	1.31E-10	1.43E-09	86	3.29E-03	3.55E-02
Neuroactive ligand-receptor interaction	57	1.56E-51	3.56E-50	55	1.48E-13	3.38E-12	112	8.68E-04	1.77E-02
Rheumatoid arthritis	46	1.91E-38	2.49E-37	149	4.03E-11	5.26E-10	47	2.64E-03	3.39E-02
Vascular smooth muscle contraction	57	7.55E-30	7.14E-29	109	1.53E-10	1.44E-09	100	3.82E-03	3.61E-02
Oxytocin signaling pathway	71	3.03E-46	5.54E-45	101	7.52E-13	1.37E-11	112	1.56E-03	2.86E-02
Adrenergic signaling in cardiomyocytes	124	3.22E-76	8.81E-74	155	3.38E-18	9.26E-16	205	2.92E-05	8.01E-03

3. Flexibility and Scalability

Application of TransReguloNet platform beyond cancer related expression datasets is a key advantage. As demonstrated in this study, an array of cancer datasets of different forms was examined. TransReguloNet also facilitates flexible use of data from diverse quantitative assay technologies such as Micro-array and sequencing based expression profiling. The span of network mining consisted of visualization, clustering and differential expression but the modular nature of the workflow enables scalability to include other quantitative methods that a researcher may need for instance inclusion of explorative statistical tools such as data distribution assessment, outlier checking and pattern extraction.

Limitations

TransReguloNet system is however not without current limitations. The obvious future intention of our pipeline is to develop a 'smart' system that leverages on automatic extraction expression sets published in specialized expression data repositories such as GEO, FANTOM and ArrayExpress databases, such that a researcher focusing on transcription regulation networks can compare their data to those in published works. This will require a standardized cross-platform representation of expression profiles in these databases or alternatively an extra filtration system to standardize these sets to facilitate automation. In the current implementation, we largely rely on in-house scripts for retreival of data prior to pre-processing before loading in the TransReguloNet server. A closely related limitation is the lower diversity of data and the bias towards human genomes as currently available in the system. As described however, this application is intended to illustrate the potential benefits of integrated pathway-based modularization and flexibility of visualization.

4.5 Conclusion

With the proposed TransReguloNet pipeline and visualization application, we provide users a new platform for studying interaction networks by a combination of mining steps and diverse data types. An innovative feature of this pipeline is the combination of network mining procedures including network construction, pathway-based abstraction and interaction strength analysis. The integration of gene expression data is implemented in an efficient approach enabling a user to observe interaction and regulation patterns of a pathway of interest. Another

additional feature implemented is the mutual information inference of interaction between transcription factors and their target genes. This is estimated from expression levels of the interacting TF and TG. The framework enhances biological intepretation from a pathway perspective. Using the visualization application we performed tests to examine usefulness of the workflow using actual gene expression data related to cancer. The application is however scalable and applicable to other expression datasets. We leave it to the user to determine the best way to infer meaningful pathway-specific patterns from the perspective of TF-centric expression regulation networks and to set their hypotheses. The relevance of integrative analysis in the context of TRN mining cannot be quantified but the network modularization presented herein refines the focus within which a user can set and test hypotheses centered on TF importance from an expression profile based on inferences deducted from using such an integrated pipeline. In the future versions improvements will be made in terms of expanding the pathway models to encompass Gene Ontology definitions, BioCyc or user-defined gene sets. Programmatic access to the TransReguloNet server is also under consideration in subsequent developments.

Chapter 5

Conclusion, Perspective and Future Work

This work has presented multilevel analytical approaches for proteomics and transcription regulation network analysis. These two fields are characterized by highly diverse data types and necessity for integrative computational methods.

Proteins are macromolecular structures made up of unique building blocks (amino acids). These molecules exist in various spatial, temporal and conditional scales. To better understand the contribution these components, robust analytical strategies in 'omics' studies need to be adopted for a wide range of applications. This will facilitate derivation of informative and biologically driven inferences. In the first part, we presented a novel method with a potential to improve analytical strategies in computational sequence representation of proteins. The more widely used method is alphabetical notation. This however does not fully describe sequence data and thus the need for computational applications that adopt detailed representation strategies such as the one presented in Chapter 3 of this dissertation.

The other theme we considered in this work was transcription regulation network analysis. Many studies have been and continue to be carried out to elucidate interaction of regulation elements at systems scale. In the computational biology context, there is a relatively recent trend relying on integrative analyses of regulation networks through a combination of several analytical tools and experimental ('wet-bench') data. This differs from earlier approaches that solely relied on computational work. There is a niche for development of multi-faceted platforms that enable the use of high-throughput datasets with a combination of computational tools for mining and visualization while taking into consideration 'wet-bench' objectives, particularly those related to elucidation of molecular and network biomarkers of diseases. This form of integrative research has been driven by platforms such as the Galaxy project (Hillman-Jackson

et al., 2012) which offer functionality for analyzing transcription regulation networks through workflows such as Cistrome (Liu *et al.*, 2011).

Since the publication of the amino acid index in 2006 (Kawashima and Kanehisa, 2000), some studies have made an attempt to apply this useful proteomics resource but there haven't been many reports on explicit integration of amino acid properties in protein sequence representation despite its obvious benefits and potential. This is relevant and ideal for machine-learning oriented protein sequence research and similar applications. Similarly, in transcription regulation network research, there are a few integrative applications centered on data-oriented TRN evaluation.

In Chapter 1 and 2, we introduced the conceptual framework of our research. These building blocks provide the fundamental information necessary for understanding our contribution. The overall key advantage of our work is the flexibility for application on diverse data although we illustrated a single example for each of the two methods in Chapter 3 and 4.

In Chapter 3, we established a comprehensive approach for sequence representation of protein sequences. Our idea was to introduce residue property integration in a procedure exemplified in Fig 3.1. Random forest algorithm was utilized for property selection. We tested our method by application in terpenoid synthase family characterization where we examined classification patterns of terpenoid synthase subgroups based on protein sequence data represented using our model. In the flow, we initially examined the amino acid index database for properties that characterize individual amino acids. Currently, 544 properties can be obtained but some of the indices are either redundant or incomplete. In their raw form, these indices cannot be used for sequence representation. Proteins can be classified as illustrated in figure 3.7. Individual residue attributes are insufficient for whole sequence property determination, however, it is possible to predict the general propensity of a sequence based on cummulative grouping of elemental properties. In computationally based sequence analyses, integration of properties such as, number of full nonbonding orbitals (Fauchere et al., 1988), conformational preference for parallel beta-strands (Lifson-Sander, 1979), retention coefficient in $NaH2PO4$ (Meek-Rossetti, 1981), average number of surrounding residues (Ponnuswamy et al., 1980), average interactions per side chain atom (Warme-Morgan, 1978), polarity (Zimmerman et al., 1968), weights from the IFH scale (Jacobs-White, 1989) and hydrophobicity index, 3.0 pH (Cowan-Whittaker, 1990) selected in our approach, enable holistic evaluation of sequence data in addition to residue identity. Random forest algorithm revealed indices that had greater influence in determining

residue diversity. We initially constructed a classification task of amino acids. This procedure selected indices whose error rate in prediction amino acid class was less than 2 percent (i.e 98 percent accuracy). As a case example, we applied this in terpenoid synthase enzyme classification. We explored using principal component analysis (PCA), the patterns exhibited by different subcategories of terpenoid synthases. We found that integrating residue attributes contributed to an increased amount of variance explained compared to bit encoding (alphabetical notation). We think that such consideration of extra attributes can be used to examine mutational pressures at the genome level by providing a traceback mechanism for evaluation of residue mutational propensities due to biochemical property closeness to function. I suppose that the proposed subset of amino acid indices can be an important tool for functional proteomics research.

In chapter 4, we interlinked TRNs to data specific gene expression profiles. We merged various indendent steps such as sequence based TF-TG prediction, condition-dependent differential gene expression analysis and knowledge-based network mining. In addition a web visualization tool was developed. We applied this framework to a few cancer related gene expression profiles.

At the initial network construction step, we created a 'whole-genome' TRN by merging experimentally validated TF-TG interactions with computationally predicted interactions. An overrepresentation metric was applied in weighting significance of predicted TF binding sites by comparison of motifs in promoter sequences to background sequence sets. At the subsequent step, we create data-specific contexts by utilization of expression datasets to extract 'expressed' subset of the static network. Three important features implemented at the network mining level included KEGG pathway specific modularization, transcription factor centric network visualization and explorative evaluation of pathway-related subsets. Another innovative feature of this proposed method is the interaction strength assessment through the use of mutual information. We performed KEGG pathway modularization in order to understand functional regulations instead of the often studied global evaluation. In combination with pathway enrichment tests, we observed that certain pathways (Table 4.2), were highly enriched in 3 lung cancer datasets suggesting their likely pivotal role in lung cancer mechanisms. Our approach is an effective way to characterize transcription networks using pathways from a functional perspective unlike extraction of patterns at the global cellular scale. TF-TG interaction strength ascertains the trend of regulation within the specific dataset of interest. This relies on the assumption that mutual information unlike other association metrics like correlation, take into account

non-monotonic associations including causal interactions. For the TransReguloNet pipeline visualization platform depicted in figure 4.6 we were able to compare conditional influence of transcription factors on target genes. This work presents the first report of a framework demonstrating data specificity in evaluation of TF-centered modules of TRNs. This integrated approach is an effective means to unravel key elements of transcription factor driven regulation networks.

With regards to the TransReguloNet application presented in chapter 4, the system is not without current limitations. The obvious future intention of the pipeline is to achieve a 'smart' system status leveraging on automatic retrieval of expression sets published in specialized expression data repositories such as GEO, FANTOM and ArrayExpress databases. This will enable a researcher using the application to compare data to those in published works. This however requires a standardized cross-platform representation of expression profiles in these databases or alternatively extra filtration systems to 'standardize' these sets to enable automation. In the current implementation, we largely rely on in-house scripts for retrieval of data prior to manual preprocessing and loading into the TransReguloNet server. A closely related limitation is the lower diversity of data and the bias towards human genomes. As described however, the application is intended to illustrate the potential benefits of integrated pathway-based modularization and flexibility of visualization. At this moment, we have clearly not achieved this yet but evidently, TransReguloNet lays the groundwork for further improvements towards this course.

"Trans-omics" is an emerging area in bioinformatics. It must continue to full scale to facilitate discovery of clinically useful software tools. It will also be applicable in complementary research activities in order to extend our understanding of multilevel systems biology. We have demonstrated the wide scope of application of integrated frameworks in the proteomics and transcriptomics. Further improvements of technologies and workflows deciphering transcription regulation networks will likely develop simultenously with the increase in expression profiling capacities. Although we reported multi-faceted bioinformatics techniques for handling protein sequences and gene regulation data, it is inherent that the scope we covered is very small considering the many possibilities in integrative 'omics'. Some of the arising questions include whether such systems can be scaled and automated for data retrieval and analysis to enable users to not only have the access to datasets of interest (or within their knowledge), but also get access to datasets submitted by other researchers. Obviously, this will require alot of computational

and financial resources for large scale system development. In the context of future work, we predict a emergence of such 'smart' systems for mining proteomics and transcriptomics datasets. It is therefore crucial to keep datasets publicly available for the development of systems biology as a field.

Bibliography

Ablett, M. P., O'Brien, C. S., Sims, A. H., Farnie, G., and Clarke, R. B. (2014). A differential role for cxcr4 in the regulation of normal versus malignant breast stem cell activity. *Oncotarget*, 5(3):599.

Afendi, F. M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M., Darusman, L. K., *et al.* (2012). Knapsack family databases: integrated metabolite–plant species databases for multifaceted plant research. *Plant and Cell Physiology*, 53(2):e1–e1.

Aguas, R., Ferguson, N. M., and Pond, S. L. K. (2013). Feature selection methods for identifying genetic determinants of host species in rna viruses. *PLoS computational biology*, 9(10):e1003254.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews genetics*, 7(1):55–65.

Altay, G. and Emmert-Streib, F. (2010). Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology*, 4(1):132.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. *Proceedings of the National Academy of Sciences*, 74(12):5350–5354.

Anfinsen, C. B. (1972). Studies on the principles that govern the folding of protein chains.

Arner, E. amd Lassman, T. (2010). chapter 7: Extraction and quality control of CAGE tags (Cap analysis of Gene Expression). Pan Stanford Publishing, Singapore.

Arredouani, M. S., Lu, B., Bhasin, M., Eljanne, M., Yue, W., Mosquera, J.-M., Bubley, G. J., Li, V., Rubin, M. A., Libermann, T. A., *et al.* (2009). Identification of the transcription factor single-minded homologue 2 as a potential biomarker and immunotherapy target in prostate cancer. *Clinical Cancer Research*, 15(18):5794–5802.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Atchley, W. R., Zhao, J., Fernandes, A. D., and Drüke, T. (2005). Solving the protein sequence metric problem. *Proc Natl Acad Sci USA*, 102(18):6395–6400.

Aumentado-Armstrong, T. T., Istrate, B., and Murgita, R. A. (2015). Algorithmic approaches to protein-protein interaction site prediction. *Algorithms for Molecular Biology*, 10(1):7.

Bailey, T. L., Elkan, C., *et al.* (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic acids research*, 34(suppl 2):W369–W373.

Bajic, V. B., Choudhary, V., and Hock, C. K. (2003). Content analysis of the core promoter region of human genes. *In silico biology*, 4(2):109–125.

Becker-Andre, M. and Hahlbrock, K. (1989). Absolute mrna quantification using the polymerase chain reaction (pcr). a novel approach by a pcr aided transcipt titration assay (patty). *Nucleic acids research*, 17(22):9437–9446.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. (2000). Genbank. *Nucleic acids research*, 28(1):15–18.

Betts, M. J. and Russell, R. B. (2003). Amino acid properties and consequences of substitutions. *Bioinformatics for geneticists*, 317:289.

Bohlmann, J., Meyer-Gauen, G., and Croteau, R. (1998). Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proc Natl Acad Sci USA*, 95(8):4126–4133.

Bono, H., Kasukawa, T., Furuno, M., Hayashizaki, Y., and Okazaki, Y. (2002). Fantom db: Database of functional annotation of riken mouse cdna clones. *Nucleic acids research*, 30(1):116–118.

Borro, L. C., Oliveira, S. R., Yamagishi, M. E., Mancini, A. L., Jardine, J. G., Mazoni, I., dos Santos, E. H., Higa, R. H., Kuser, P. R., and Neshich, G. (2006). Predicting enzyme class from protein structure using bayesian classification. *Genetics and molecular research: GMR*, 5(1):193–202.

Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov*, 2(6):493–507.

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). Htridb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC genomics*, 13(1):405.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., *et al.* (2003). Arrayexpressa public repository for microarray gene expression data at the ebi. *Nucleic acids research*, 31(1):68–71.

Breiman, L. (2001). Random forests. *Mach Learn*, 45(1):5–32.

Bryson, K., Cozzetto, D., and Jones, D. T. (2007). Computer-assisted protein domain boundary prediction using the dom-pred server. *Current Protein and Peptide Science*, 8(2):181–188.

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186.

Cai, C., Wang, W., Sun, L., and Chen, Y. (2003). Protein function classification via support vector machine approach. *Mathematical biosciences*, 185(2):111–122.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., *et al.* (2005). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., *et al.* (2008). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 36(suppl 1):D623–D631.

Chen, X.-W. and Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–4400.

Cheng, J. (2007). Domac: an accurate, hybrid protein domain prediction server. *Nucleic acids research*, 35(suppl 2):W354–W356.

Cheng, J., Tegge, A. N., and Baldi, P. (2008). Machine learning methods for protein structure prediction. *Biomedical Engineering, IEEE Reviews in*, 1:41–49.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., *et al.* (1998). Sgd: Saccharomyces genome database. *Nucleic acids research*, 26(1):73–79.

Chipman, H., Hastie, T. J., and Tibshirani, R. (2003). Clustering microarray data. *Statistical analysis of gene expression microarray data*, 1:159–200.

Coghlan, A., Mac Donaill, D. A., and Buttimore, N. H. (2001). Representation of amino acids as five-bit or three-bit patterns for filtering protein databases. *Bioinformatics*, 17(8):676–685.

Connolly, J. D. and Hill, R. A. (1991). *Dictionary of terpenoids. 1. Mono-and sesquiterpenoids*, volume 1. CRC Press.

Consortium, E. P. *et al.* (2004). The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640.

Consortium, E. P. *et al.* (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Consortium, T. F. *et al.* (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.

Consortium, U. *et al.* (2008). The universal protein resource (uniprot). *Nucleic acids research*, 36(suppl 1):D190–D195.

Cordero, D., Solé, X., Crous-Bou, M., Sanz-Pamplona, R., Paré, L., Guinó, E., Olivares, D., Berenguer, A., Santos, C., Salazar, R., *et al.* (2014). Large differences in global transcriptional regulatory programs of normal and tumor colon cells. *BMC cancer*, 14(1):708.

Cowan, R. and Whittaker, R. G. (1989). Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography. *Peptide research*, 3(2):75–80.

de Hoon, M. and Hayashizaki, Y. (2008). Deep cap analysis gene expression (cage): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques*, 44(5):627.

Degenhardt, J., Köllner, T. G., and Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry*, 70(15):1621–1637.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210.

Ergün, A., Lawrence, C. A., Kohanski, M. A., Brennan, T. A., and Collins, J. J. (2007). A network biology approach to prostate cancer. *Molecular systems biology*, 3(1):82.

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? *Nature Reviews Genetics*, 9(2):102–114.

Fu, J., Khaybullin, R., Liang, X., Morin, M., Xia, A., Yeh, A., and Qi, X. (2015). Discovery of gene regulation pattern in lung cancer by gene expression profiling using human tissues. *Genomics data*, 3:112–115.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognit Lett*, 31(14):2225–2236.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., *et al.* (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–1455.

Gong, X.-S., Wen, J. Q., and Gray, J. C. (2000). The role of amino-acid residues in the hydrophobic patch surrounding the haem group of cytochrome f in the interaction with plastocyanin. *European Journal of Biochemistry*, 267(6):1732–1742.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864.

Guan, Y., Myers, C. L., Hess, D. C., Barutcuoglu, Z., Caudy, A., and Troyanskaya, O. (2008). Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome biology*, 9(Suppl 1):S3.

Hamp, T. and Rost, B. (2015). More challenges for machine learning protein interactions. *Bioinformatics*, page btu857.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *cell*, 100(1):57–70.

He, M. and Petoukhov, S. (2011). *Mathematics of Bioinformatics: Theory, Methods and Applications*, volume 19. Wiley.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*, 89(22):10915–10919.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7):563–577.

Herzel, H. and Große, I. (1995). Measuring correlations in symbol sequences. *Physica A: Statistical Mechanics and its Applications*, 216(4):518–542.

Hillman-Jackson, J., Clements, D., Blankenberg, D., Taylor, J., Nekrutenko, A., and Team, G. (2012). Using galaxy to perform large-scale interactive data analyses. *Current protocols in bioinformatics*, pages 10–5.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57.

Huang, H.-L., Lin, I.-C., Liou, Y.-F., Tsai, C.-T., Hsu, K.-T., Huang, W.-L., Ho, S.-J., and Ho, S.-Y. (2011). Predicting and analyzing dna-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. *BMC bioinformatics*, 12(Suppl 1):S47.

Hyatt, D. C., Youn, B., Zhao, Y., Santhamma, B., Coates, R. M., Croteau, R. B., and Kang, C. (2007). Structure of limonene synthase, a simple model for terpenoid cyclase catalysis. *Proceedings of the National Academy of Sciences*, 104(13):5360–5365.

Ikeda, S., Abe, T., Nakamura, Y., Kibinge, N., Morita, A. H., Nakatani, A., Ono, N., Ikemura, T., Nakamura, K., Altaf-Ul-Amin, M., *et al.* (2013). Systematization of the protein sequence diversity in enzymes related to secondary metabolic pathways in plants, in the context of big data biology inspired by the knapsack motorcycle database. *Plant Cell Physiol*, 54(5):711–727.

Isik, Z., Ersahin, T., Atalay, V., Aykanat, C., and Cetin-Atalay, R. (2012). A signal transduction score flow algorithm for cyclic cellular pathway analysis, which combines transcriptome and chip-seq data. *Molecular bioSystems*, 8(12):3224–3231.

IUPAC-IUB (1971). A one letter notation for amino acid sequences(definitive rules). *IUPAC*.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356.

Jacobs, R. E. and White, S. H. (1989). The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices. *Biochemistry*, 28(8):3421–3437.

Jensen, L. J., Skovgaard, M., and Brunak, S. (2002). Prediction of novel archaeal enzymes from sequence-derived features. *Protein Science*, 11(12):2894–2898.

Jolliffe, I. (2005). *Principal component analysis*. Wiley Online Library.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kawaji, H., Severin, J., Lizio, M., Waterhouse, A., Katayama, S., Irvine, K. M., Hume, D. A., Forrest, A., Suzuki, H., Carninci, P., *et al.* (2009). The fantom web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol*, 10(4):R40.

Kawashima, S. and Kanehisa, M. (2000). Aaindex: amino acid index database. *Nucleic Acids Res*, 28(1):374–374.

Kendrew, J., Dickerson, R., Strandberg, B., Hart, R., Davies, D., Phillips, D., and Shore, V. (1960). Structure of myoglobin: A three-dimensional fourier synthesis at 2 a. resolution. *Nature*, 185(4711):422–427.

Kline, P. (1994). *An easy guide to factor analysis*. Routledge.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., *et al.* (2006). Cage: cap analysis of gene expression. *Nature methods*, 3(3):211–222.

Komurov, K., Tseng, J.-T., Muller, M., Seviour, E. G., Moss, T. J., Yang, L., Nagrath, D., and Ram, P. T. (2012). The glucose-deprivation network counteracts lapatinib-induced toxicity in resistant erbb2-positive breast cancer cells. *Molecular systems biology*, 8(1):596.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G. (2007). Cpc: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*, 35(suppl 2):W345–W349.

Krzanowski, W. J. (2000). *Principles of multivariate analysis*. Oxford University Press.

Kumar, R. M., Cahan, P., Shalek, A. K., Satija, R., DaleyKeyser, A. J., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J. J., *et al.* (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, 516(7529):56–61.

Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C., Schnabel, P., Warth, A., Poustka, A., Sültmann, H., *et al.* (2009). Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung cancer*, 63(1):32–38.

Kutmon, M., Kelder, T., Mandaviya, P., Evelo, C. T., and Coort, S. L. (2013). Cytargetlinker: a cytoscape app to integrate regulatory interactions in network analysis. *PloS one*, 8(12):e82160.

Li, D., Ono, N., Sato, T., Sugiura, T., Altaf-Ul-Amin, M., Ohta, D., Suzuki, H., Arita, M., Tanaka, K., Ma, Z., *et al.* (2015). Targeted integration of rna-seq and metabolite data to elucidate curcuminoid biosynthesis in four curcuma species. *Plant and Cell Physiology*, page pcv008.

Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., Shin, H., Wong, S. S., Ma, J., Lei, Y., *et al.* (2011). Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol*, 12(8):R83.

Lu, T.-P., Tsai, M.-H., Lee, J.-M., Hsu, C.-P., Chen, P.-C., Lin, C.-W., Shih, J.-Y., Yang, P.-C., Hsiao, C. K., Lai, L.-C., *et al.* (2010). Identification of a novel biomarker, sema5a, for non–small cell lung carcinoma in nonsmoking women. *Cancer Epidemiology Biomarkers & Prevention*, 19(10):2590–2597.

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). Gage: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, 10(1):161.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7.

Marx, V. (2013). Biology: The big challenges of big data. *Nature*, 498(7453):255–260.

Matsumura, H., Molina, C., Krüger, D. H., Terauchi, R., and Kahl, G. (2012). Deepsupersage: High-throughput transcriptome sequencing with now-and next-generation sequencing technologies. *Tag-Based Next Generation Sequencing*, pages 1–21.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., *et al.* (2003). Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):374–378.

Medvedovic, M., Yeung, K. Y., and Bumgarner, R. E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232.

Meek, J. L. (1980). Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proceedings of the National Academy of Sciences*, 77(3):1632–1636.

Meyer, P. E., Kontos, K., Lafitte, F., and Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, 2007:8–8.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118.

Morozova, O., Hirst, M., and Marra, M. A. (2009). Applications of new sequencing technologies for transcriptome analysis. *Annual review of genomics and human genetics*, 10:135–151.

Morrissy, A. S., Morin, R. D., Delaney, A., Zeng, T., McDonald, H., Jones, S., Zhao, Y., Hirst, M., and Marra, M. A. (2009). Next-generation tag sequencing for cancer gene expression profiling. *Genome research*, 19(10):1825–1835.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628.

Nagegowda, D. A., Gutensohn, M., Wilkerson, C. G., and Dudareva, N. (2008). Two nearly identical terpene synthases catalyze the formation of nerolidol and linalool in snapdragon flowers. *The Plant Journal*, 55(2):224–239.

Nieuwenhuizen, N. J., Wang, M. Y., Matich, A. J., Green, S. A., Chen, X., Yauk, Y.-K., Beuning, L. L., Nagegowda, D. A., Dudareva, N., and Atkinson, R. G. (2009). Two terpene synthases are responsible for the major sesquiterpenes emitted from the flowers of kiwifruit (actinidia deliciosa). *Journal of experimental botany*, 60(11):3203–3219.

Ouyang, M., Welsh, W. J., and Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, 20(6):917–923.

Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. (2004). Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic acids research*, 32(suppl 2):W199–W203.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., and Will, G. (1960). Structure of hæmoglobin: a three-dimensional fourier synthesis at 5.5-å. resolution, obtained by x-ray analysis. *Nature*, 185:416–422.

Qin, S., Ma, F., and Chen, L. (2015). Gene regulatory networks by transcription factors and micrornas in breast cancer. *Bioinformatics*, 31(1):76–83.

Rattei, T., Tischler, P., Götz, S., Jehl, M.-A., Hoser, J., Arnold, R., Conesa, A., and Mewes, H.-W. (2010). Simapa comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic acids research*, 38(suppl 1):D223–D226.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Roman-Roldan, R., Bernaola-Galvan, P., and Oliver, J. (1996). Application of information theory to dna sequence analysis: a review. *Pattern recognition*, 29(7):1187–1194.

Romano, S., Bailey, J., Nguyen, V., and Verspoor, K. (2014). Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1143–1151.

Rost, B. and Sander, C. (1993a). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences*, 90(16):7558–7562.

Rost, B. and Sander, C. (1993b). Prediction of protein secondary structure at better than 70% accuracy. *Journal of molecular biology*, 232(2):584–599.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2009). Corum: the comprehensive resource of mammalian protein complexes2009. *Nucleic acids research*, page gkp914.

Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., *et al.* (2003). Tm4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2):374.

Saito, A., Suzuki, H. I., Horie, M., Ohshima, M., Morishita, Y., Abiko, Y., and Nagase, T. (2013). An integrated expression profiling reveals target genes of tgf-beta and tnf-alpha possibly mediated by microrna s in lung cancer cells. *PloS one*, 8(2):e56587.

Sanger, F. and Thompson, E. (1953). The amino-acid sequence in the glycyl chain of insulin. 1. the identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 53(3):353.

Sato, T., Kaneda, A., Tsuji, S., Isagawa, T., Yamamoto, S., Fujita, T., Yamanaka, R., Tanaka, Y., Nukiwa, T., Marquez, V. E., *et al.* (2013). Prc2 overexpression and prc2-target gene repression relating to poorer prognosis in small cell lung cancer. *Scientific reports*, 3.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

Shannon, P., Shannon, M. P., RUnit, S., and biocViews GenomicSequence, M. (2013). Package motifdb.

Shinbo, Y., Nakamura, Y., Altaf-Ul-Amin, M., Asahi, H., Kurokawa, K., Arita, M., Saito, K., Ohta, D., Shibata, D., and Kanaya, S. (2006). Knapsack: a comprehensive species-metabolite relationship database. In *Plant Metabolomics*, pages 165–181. Springer.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., *et al.* (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781.

Sneath, P. (1966). Relations between chemical structure and biological activity in peptides. *Journal of theoretical biology*, 12(2):157–195.

Staden, R. (1977). Sequence data handling by computer. *Nucleic Acids Res*, 4(11):4037–4052.

Stein, L. D. (2003). Integrating biological databases. *Nature Reviews Genetics*, 4(5):337–345.

Steinfeld, I., Navon, R., Creech, M. L., Yakhini, Z., and Tsalenko, A. (2015). Enviz: a cytoscape app for integrated statistical analysis and visualization of sample-matched data with multiple data types. *Bioinformatics*, page btu853.

Steipe, B., Schiller, B., Plückthun, A., and Steinbacher, S. (1994). Sequence statistics reliably predict stabilizing mutations in a protein domain. *Journal of molecular biology*, 240(3):188–192.

Steward, R. E. and Thornton, J. M. (2002). Prediction of strand pairing in antiparallel and parallel β-sheets using information theory. *Proteins: Structure, Function, and Bioinformatics*, 48(2):178–191.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25.

Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics*, 18(1):207–208.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *J Chem Inf Comput Sci*, 43(6):1947–1958.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research*, 29(1):22–28.

Tress, M., Cheng, J., Baldi, P., Joo, K., Lee, J., Seo, J.-H., Lee, J., Baker, D., Chivian, D., Kim, D., *et al.* (2007). Assessment of predictions submitted for the casp7 domain prediction category. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):137–151.

Turner, D. P., Findlay, V. J., Moussa, O., and Watson, D. K. (2007). Defining ets transcription regulatory networks and their contribution to breast cancer progression. *Journal of cellular biochemistry*, 102(3):549–559.

van Nimwegen, E. (2007). Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC bioinformatics*, 8(Suppl 6):S4.

Vapnik, V. N. and Vapnik, V. (1998). *Statistical learning theory*, volume 1. Wiley New York.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235):484–487.

Vendruscolo, M. and Tartaglia, G. G. (2008). Towards quantitative predictions in cell biology using chemical properties of proteins. *Mol Biosyst*, 4(12):1170–1175.

Walsh, I., Pollastri, G., and Tosatto, S. C. (2015). Correct machine learning on protein sequences: a peer-reviewing perspective. *Briefings in bioinformatics*, page bbv082.

Warburg, O. *et al.* (1956). On the origin of cancer cells. *Science*, 123(3191):309–314.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., *et al.* (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

Weiss, O., Jimenez-Montano, M. A., and Herzel, H. (2000). Information content of protein sequences. *J Theor Biol*, 206(3):379–386.

White, G. and Seffens, W. (1998). Using a neural network to backtranslate amino acid sequences. *Electron J Biotechnol*, 1(3):17–18.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C. L., Haase, J., Janes, J., Huss, J. W., *et al.* (2009). Biogps: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol*, 10(11):R130.

Xie, X., Lu, J., Kulbokas, E., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3 utrs by comparison of several mammals. *Nature*, 434(7031):338–345.

Yao, L., Shen, H., Laird, P., Farnham, P., and Berman, B. (2015). Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome biology*, 16(1):105–105.

Zambelli, F., Pesole, G., and Pavesi, G. (2009). Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic acids research*, 37(suppl 2):W247–W252.

Zubek, J., Tatjewski, M., Boniecki, A., Mnich, M., Basu, S., and Plewczynski, D. (2015). Multi-level machine learning prediction of protein–protein interactions in saccharomyces cerevisiae. *PeerJ*, 3:e1041.

Achievements

Journal Publications

Related to Thesis

1. **Nelson Kibinge**, Shun Ikeda, Naoaki Ono, Md Altaf-Ul-Amin, and Shigehiko Kanaya. Integration of residue attributes for sequence diversity characterization of terpenoid enzymes (May 2014). *BioMed research international* May 2014 doi:10.1155/2014/753428 (in Chapter 3).

2. **Nelson Kibinge**, Naoaki Ono, Masafumi Horie, Tetsuo Sato, Tadao Sugiura, Md. Altaf-Ul-Amin, Akira Saito and Shigehiko Kanaya (2015). Integrated-pathway based transcription regulation network mining and visualization based on gene expression profiles. *Journal of Biomedical Informatics*, *Under Revision* (in Chapter 4).

3. Shun Ikeda, Takashi Abe, Yukiko Nakamura, **Nelson Kibinge**, Aki Hirai Morita, Atsushi Nakatani, Naoaki Ono, Tomishi Ikemura, Kensuke Nakamura, Md.Altaf-Ul-Amin and Shigehiko Kanaya (March 2013). Systematization of the protein sequence diversity in enzymes related to secondary metabolic pathways in plants, in the context of big data biology inspired by the KNApSAcK Motorcycle database. *Plant and cell physiology*: pct041 (in Chapter 3)

Other Publications

1. Yuki Ohtana, Azian Azamimi Abdullah, Md. AltafUlAmin, Ming Huang, Naoaki Ono, Tetsuo Sato, Tadao Sugiura, Hisayuki Horai, Yukiko Nakamura, Aki Morita, Klaus W. Lange, **Nelson K. Kibinge** and Shigehiko Kanaya (Jan 2014). Clustering of 3DStructure

Similarity Based Network of Secondary Metabolites Reveals Their Relationships with Biological Activities. *Molecular Informatics*, 33(1112), 790-801.

2. Noriko Tanaka, Nor Aini B. Zakaria , **Nelson K. Kibinge**, Kanaya Shigehiko, Toshiyo Tamura and Masaki Yoshida (2014). Fall-risk classification of the timed up-and-go test with principal component analysis. *International journal of neurorehabilitation*, 1, e106-e106.

Conference Publications

1. **Nelson Kibinge**, Naoaki Ono, Masafumi Horie, Tetsuo Sato, Tadao Sugiura, MD Altaf-Ul-Amin, Akira Saito and Shigehiko Kanaya (2015). Integrated pathway-based transcription regulation network mining and visualization tool for gene expression data *Functional Genomics and Systems Biology: From Model Organisms to Human Health 2015*. Hinxton Cambridge United Kingdom, October 28-30 2015 (in Chapter 4).

2. **Nelson Kibinge**, Naoaki Ono, Masafumi Horie, Ming Huang, Tetsuo Sato, Tadao Sugiura, Md. Altaf-Ul-Amin, Saito Akira and Shigehiko Kanaya (2014), A systems mapping of transcription regulation in genes and modules of genes in lung cancer pathways. *GIW-ISCB 2014*, Odaiba Tokyo Japan, December 14-17, 2014 (in Chapter 4)

3. **Nelson Kibinge**, Masafumi Horie, Md.Altaf-Ul-Amin, Shigehiko Kanaya, Saito Akira and Naoaki Ono. Dynamics of transcription initiation and promoter usage during development *3rd International Conference and Exhibition on Metabolomics and Systems Biology*. San Antonio Texas USA, March 24-26 2014 (in Chapter 4).

4. **Nelson Kibinge**, Shun Ikeda, Naoaki Ono, Nakamura Kensuke, MD Altaf-Ul-Amin and Shigehiko Kanaya (2013). Utilizing metric properties of amino acids in distance matrix methods for protein sequence analysis *Brain and Health Informatics BHI2013*. Maebashi Gunma Japan, October 29-31 2013 (in Chapter 3).

Acknowledgment

Profound gratitude my supervisor Prof. Shigehiko Kanaya for giving me the chance to do research in his. Without his guidance this would not have been possible. I extend my gratitude to assistant Prof. Naoaki Ono who has directly supervised this research and made significant contributions towards shaping the general direction we eventually took. Thanks to all the professors in the CSB lab including associate Prof. Md. Altaf-Ul-Amin, associate Prof. Tadao Sugiura and assistant Prof. Tetsuo Sato for their valuable discussions and questions. I appreciate the thesis committee particularly Prof. Keiichi Yasumoto for taking his time to review my thesis and for their insightful recommendations. I am thankful to all the colleagues who in one way or another shared in the challenge of graduate research. MEXT made it all happen. I cannot thank the ministry enough for shaping my future so profoundly. I reserve the heartiest of gratitudes for my family. For the unfailing love and support, thanks dearest Loly for the patience and for ensuring we had something to eat especially on those late nights despite the busy schedule. My little boy MKC, has been a major motivation. Welcome to the world son, its not as bad as they say it is out there. I hope you too will experience the journey. And to my parents who have made immense sacrifices, brothers and sisters who have looked upto me, I hope I haven't let you down. Sincere appreciation to everyone else who directly or indirectly made a contribution towards my 5 years of grad school.

March 2016

Nelson Kipchirchir Kibinge

Appendices

Appendix A

Supplementary tables

Table A.1: KEGG pathway geneset definitions

Pathway ID	Description	Gene Sets
path:hsa00010	Glycolysis / Gluconeogenesis	HK3 HK1 HK2 HKDC1 GCK GPI PFKM PFKP PFKL FBP1 FBP2 ALDOC ALDOA ALDOB TPI1 GAPDH GAPDHS PGK2 PGK1 PGAM1 PGAM2 PGAM4 ENO3 ENO2 ENO1 PKM PKLR PDHA2 PDHA1 PDHB DLAT DLD LDHAL6A LDHAL6B LDHA LDHB LDHC ADH1A ADH1B ADH1C ADH7 ADH4 ADH5 ADH6 AKR1A1 ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1 ALDH3B1 ALDH3B2 ALDH1A3 ALDH3A1 ACSS1 ACSS2 GALM PGM1 PGM2 G6PC G6PC2 G6PC3 ADPGK BPGM MINPP1 PCK1 PCK2
path:hsa00020	Citrate cycle (TCA cycle)	CS ACLY ACO2 ACO1 IDH1 IDH2 IDH3B IDH3G IDH3A OGDHL OGDH DLST DLD SUCLG1 SUCLG2 SUCLA2 SDHA SDHB SDHC SDHD FH MDH1 MDH2 PC PCK1 PCK2 PDHA2 PDHA1 PDHB DLAT
path:hsa00030	Pentose phosphate pathway	GPI G6PD PGLS H6PD PGD RPE RPEL1 TKT TKTL2 TKTL1 TALDO1 RPIA DERA RBKS PGM1 PGM2 PRPS1L1 PRPS2 PRPS1 RGN IDNK ALDOC ALDOA ALDOB FBP1 FBP2 PFKM PFKP PFKL
path:hsa00040	Pentose and glucuronate interconversions	GUSB KL UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 UGDH UGP2 AKR1A1 CRYL1 RPE RPEL1 XYLB AKR1B1 AKR1B10 DCXR SORD DHDH ALDH2 ALDH3A2 ALDH1B1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00051	Fructose and mannose metabolism	MPI PMM2 PMM1 GMPPB GMPPA GMDS TSTA3 FPGT FUK ENOSF1 HK3 HK1 HK2 HKDC1 PFKM PFKP PFKL FBP1 FBP2 PFKFB1 PFKFB2 PFKFB3 PFKFB4 C12orf5 KHK SORD AKR1B1 AKR1B10 ALDOC ALDOA ALDOB TPI1
path:hsa00052	Galactose metabolism	GALM GALK1 GALT GALE UGP2 PGM1 PGM2 HK3 HK1 HK2 HKDC1 GCK G6PC G6PC2 G6PC3 GLB1 LCT LALBA B4GALT1 B4GALT2 GLA AKR1B1 AKR1B10 PFKM PFKP PFKL MGAM GAA GANC SI
path:hsa00053	Ascorbate and aldarate metabolism	UGDH UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 MIOX RGN ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1
path:hsa00061	Fatty acid biosynthesis	ACACA ACACB MCAT FASN OXSM OLAH ACSL6 ACSL4 ACSL1 ACSL5 ACSL3 ACSBG1 ACSBG2
path:hsa00062	Fatty acid elongation	ACAA2 HADHB HADH HADHA ECHS1 MECR PPT1 PPT2 ELOVL1 ELOVL2 ELOVL3 ELOVL4 ELOVL5 ELOVL6 ELOVL7 HSD17B12 HACD2 HACD1 HACD4 HACD3 TECR ACOT4 ACOT2 ACOT1 ACOT7
path:hsa00071	Fatty acid degradation	ACAT2 ACAT1 ACAA1 ACAA2 HADHB HADH HADHA EHHADH ECHS1 ACOX3 ACOX1 ACADS ACADM ACADL ACADSB ACADVL GCDH ACSL6 ACSL4 ACSL1 ACSL5 ACSL3 ACSBG1 ACSBG2 CPT1A CPT1B CPT1C CPT2 ECI1 ECI2 CYP4A11 CYP4A22 ADH1A ADH1B ADH1C ADH7 ADH4 ADH5 ADH6 ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1
path:hsa00072	Synthesis and degradation of ketone bodies	HMGCS1 HMGCS2 HMGCL HMGCLL1 OXCT1 OXCT2 ACAT2 ACAT1 BDH1 BDH2
path:hsa00100	Steroid biosynthesis	FDFT1 SQLE LSS CYP51A1 TM7SF2 MSMO1 FAXDC2 NSDHL HSD17B7 EBP DHCR24 SC5D DHCR7 LIPA CEL SOAT2 SOAT1 CYP2R1 CYP27B1 CYP24A1
path:hsa00120	Primary bile acid biosynthesis	CYP46A1 CYP39A1 HSD3B7 CH25H CYP7B1 CYP7A1 CYP27A1 CYP8B1 AKR1D1 AKR1C4 SLC27A5 AMACR ACOX2 HSD17B4 SCP2 ACOT8 BAAT
path:hsa00130	Ubiquinone and other terpenoid-quinone biosynthesis	TAT COQ2 COQ3 COQ6 COQ5 COQ7 NQO1 GGCX VKORC1 VKORC1L1 HPD

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00140	Steroid hormone biosynthesis	CYP11A1 CYP17A1 STS SULT2B1 CYP21A2 HSD3B1 HSD3B2 SRD5A1 SRD5A2 SRD5A3 AKR1C2 AKR1C4 AKR1C1 AKR1C3 CYP11B1 CYP11B2 AKR1D1 HSD11B1 HSD11B2 CYP7B1 SULT1E1 HSD17B1 HSD17B2 HSD17B6 HSD17B7 HSD17B8 HSD17B12 CYP1A1 CYP1A2 CYP3A5 CYP3A7 CYP3A7-CYP3A51P CYP2E1 CYP3A4 CYP1B1 CYP19A1 CYP7A1 UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 COMT HSD17B3
path:hsa00190	Oxidative phosphorylation	ND1 ND2 ND3 ND4 ND4L ND5 ND6 NDUFS1 NDUFS2 NDUFS3 NDUFS4 NDUFS5 NDUFS6 NDUFS7 NDUFS8 NDUFV1 NDUFV2 NDUFV3 NDUFA1 NDUFA2 NDUFA3 NDUFA4 NDUFA4L2 NDUFA5 NDUFA6 NDUFA7 NDUFA8 NDUFA9 NDUFA10 NDUFAB1 NDUFA11 NDUFA12 NDUFA13 NDUFB1 NDUFB2 NDUFB3 NDUFB4 NDUFB5 NDUFB6 NDUFB7 NDUFB8 NDUFB9 NDUFB10 NDUFB11 NDUFC1 NDUFC2 NDUFC2-KCTD14 SDHA SDHB SDHC SDHD UQCRFS1 CYTB CYC1 UQCRC1 UQCRC2 UQCRH UQCRHL UQCRB UQCRQ UQCR10 UQCR11 COX10 COX3 COX1 COX2 COX4I2 COX4I1 COX5A COX5B COX6A1 COX6A2 COX6B1 COX6B2 COX6C COX7A1 COX7A2 COX7A2L COX7B COX7B2 COX7C COX8C COX8A COX11 COX15 COX17 ATP5A1 ATP5B ATP5C1 ATP5D ATP5E ATP5O ATP6 ATP5F1 ATP5G1 ATP5G2 ATP5G3 ATP5H ATP5I ATP5J2 ATP5L ATP5J ATP8 ATP6V1A ATP6V1B1 ATP6V1B2 ATP6V1C2 ATP6V1C1 ATP6V1D ATP6V1E2 ATP6V1E1 ATP6V1F ATP6V1G1 ATP6V1G3 ATP6V1G2 ATP6V1H TCIRG1 ATP6V0A2 ATP6V0A4 ATP6V0A1 ATP6V0C ATP6V0B ATP6V0D1 ATP6V0D2 ATP6V0E1 ATP6V0E2 ATP6AP1 ATP4A ATP4B ATP12A PPA2 PPA1 LHPP
path:hsa00220	Arginine biosynthesis	OTC ASS1 ASL ARG2 ARG1 NOS1 NOS2 NOS3 GLS2 GLS GLUL GLUD2 GLUD1 CPS1 GOT1 GOT2 GPT2 GPT NAGS ACY1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00230	Purine metabolism	NUDT9 ADPRM NUDT5 PGM1 PGM2 PRPS1L1 PRPS2 PRPS1 PPAT GART PFAS PAICS ADSL ATIC APRT NT5C2 NT5C1A NT5E NT5C1B NT5C NT5M NT5C3A NT5C3B NT5C1B-RDH14 PNP HPRT1 IMPDH1 IMPDH2 NME6 NME7 NME2 NME4 NME1 NME5 NME3 NME1-NME2 AK9 ENTPD3 ENTPD8 ENTPD1 CANT1 ENTPD4 ENTPD5 ENTPD6 NUDT16 ITPA XDH NUDT2 GMPS GMPR GMPR2 GDA GUK1 PKM PKLR RRM1 RRM2B RRM2 DGUOK POLR1A POLR1B ZNRD1 TWISTNB POLR1E POLR2A POLR2B POLR2C POLR2D POLR2E POLR2F POLR2G POLR2H POLR2I POLR2L POLR2J POLR2J3 POLR2J2 POLR2K POLR3A POLR3B POLR3C POLR3D POLR3E LOC101060521 POLR1C POLR3K POLR1D POLR3H POLR3GL POLR3G POLR3F POLA1 POLA2 PRIM1 PRIM2 POLD1 POLD2 POLD3 POLD4 POLE POLE2 POLE3 POLE4 HDDC3 PRUNE ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 ADCY10 GUCY1A2 GUCY1A3 GUCY1B3 GUCY2C GUCY2D GUCY2F NPR1 NPR2 PDE1A PDE1B PDE1C PDE2A PDE3A PDE3B PDE5A PDE6A PDE6B PDE6C PDE6D PDE6G PDE6H PDE9A PDE10A PDE11A ADSSL1 ADSS AMPD2 AMPD3 AMPD1 ADK DCK ADA CECR1 AK7 AK4 AK5 AK2 AK1 LOC390877 AK8 AK6 AK3 ENTPD2 NTPCR PNPT1 PDE4A PDE4B PDE4C PDE4D PDE7A PDE7B PDE8B PDE8A FHIT ENPP4 PAPSS2 PAPSS1 ENPP1 ENPP3 URAD ALLC
path:hsa00232	Caffeine metabolism	CYP1A2 NAT2 NAT1 CYP2A6 XDH

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00240	Pyrimidine metabolism	CAD DHODH UMPS CMPK1 CMPK2 NME6 NME7 NME2 NME4 NME1 NME5 NME3 NME1-NME2 AK9 PNPT1 ENTPD3 ENTPD8 ENTPD1 CANT1 ENTPD4 ENTPD5 ENTPD6 ITPA CTPS1 CTPS2 POLR1A POLR1B ZNRD1 TWISTNB POLR1E POLR2A POLR2B POLR2C POLR2D POLR2E POLR2F POLR2G POLR2H POLR2I POLR2L POLR2J POLR2J3 POLR2J2 POLR2K POLR3A POLR3B POLR3C POLR3D POLR3E LOC101060521 POLR1C POLR3K POLR1D POLR3H POLR3GL POLR3G POLR3F POLA1 POLA2 PRIM1 PRIM2 POLD1 POLD2 POLD3 POLD4 POLE POLE2 POLE3 POLE4 UCK1 UCK2 UCKL1 NT5C2 NT5C1A NT5E NT5C1B NT5C NT5M NT5C3A NT5C3B NT5C1B-RDH14 UPP2 UPP1 UPRT DPYD DPYS UPB1 TXNRD1 TXNRD2 TXNRD3 RRM1 RRM2B RRM2 DCTPP1 DUT TYMS CDA TYMP PNP DCK DCTD TK2 TK1 DTYMK NUDT2
path:hsa00250	Alanine, aspartate and glutamate metabolism	GOT1 GOT2 IL4I1 DDO ASNS NIT2 GPT2 GPT AGXT AGXT2 ASS1 ASL ADSSL1 ADSS ADSL NAT8L RIMKLB RIMKLA FOLH1 ASPA GAD1 GAD2 ABAT ALDH5A1 GLUD2 GLUD1 ALDH4A1 GLUL CAD GLS2 GLS CPS1 GFPT2 GFPT1 PPAT
path:hsa00260	Glycine, serine and threonine metabolism	SHMT2 SHMT1 PSPH PSAT1 PHGDH GLYCTK PGAM1 PGAM2 PGAM4 BPGM GRHPR GCAT ALAS1 ALAS2 MAOB MAOA AOC3 AOC2 GLDC AMT DLD GCSH DAO AGXT AGXT2 GATM GAMT CHDH ALDH7A1 BHMT DMGDH PIPOX SARDH GNMT CBS LOC102724560 CTH SDS SDSL SRR
path:hsa00270	Cysteine and methionine metabolism	CTH CBS LOC102724560 BHMT MTR MAT2B MAT1A MAT2A AMD1 SRM SMS MTAP MRI1 APIP ENOPH1 ADI1 TAT IL4I1 DNMT1 DNMT3A DNMT3B AHCYL2 AHCYL1 AHCY BCAT2 BCAT1 AGXT2 GCLC GCLM GSS CDO1 GOT1 GOT2 MPST TST LDHAL6A LDHAL6B LDHA LDHB LDHC MDH1 MDH2 SDS SDSL
path:hsa00280	Valine, leucine and isoleucine degradation	BCAT2 BCAT1 IL4I1 BCKDHA BCKDHB DBT DLD ACADS ACADM IVD ACADSB ACAD8 HADHA EHHADH ECHS1 HADH HSD17B10 ACAA1 ACAA2 HADHB PCCA PCCB MCEE MUT HIBCH HIBADH ALDH6A1 ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1 AOX1 ACSF3 ABAT MCCC1 MCCC2 AUH HMGCL HMGCLL1 OXCT1 OXCT2 AACS ACAT2 ACAT1 HMGCS1 HMGCS2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00290	Valine, leucine and isoleucine biosynthesis	SDS SDSL BCAT2 BCAT1
path:hsa00300	Lysine biosynthesis	AADAT ALDH7A1
path:hsa00310	Lysine degradation	AASS ALDH7A1 AADAT OGDHL OGDH DLST GCDH HADHA EHHADH ECHS1 HADH ACAT2 ACAT1 HYKK PHYKPL PIPOX ASH1L DOT1L EHMT2 EHMT1 NSD1 WHSC1 WHSC1L1 SETD1B SETD1A SETD2 SETD7 SETD8 SETDB1 SETDB2 SETMAR SUV39H1 SUV39H2 SUV420H2 SUV420H1 KMT2A KMT2D KMT2C KMT2B KMT2E CAMKMT TMLHE ALDH2 ALDH3A2 ALDH1B1 ALDH9A1 BBOX1 PLOD1 PLOD2 PLOD3 COLGALT1 COLGALT2
path:hsa00330	Arginine and proline metabolism	GATM GAMT CKM CKMT1A CKMT2 CKB CKMT1B AZIN2 AGMAT ODC1 SRM SMS AMD1 AOC1 SMOX ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1 CNDP1 CNDP2 CARNS1 SAT2 SAT1 MAOB MAOA NOS1 NOS2 NOS3 ARG2 ARG1 OAT PYCRL PYCR2 PYCR1 PRODH LOC102724788 ALDH4A1 ALDH18A1 LAP3 P4HA2 P4HA3 P4HA1 PRODH2 GOT1 GOT2 HOGA1 DAO L3HYPDH
path:hsa00340	Histidine metabolism	HAL UROC1 AMDHD1 FTCD HDC DDC AOC1 ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1 ASPA HNMT MAOB MAOA ALDH3B1 ALDH3B2 ALDH1A3 ALDH3A1 CARNS1 CNDP2 CNDP1
path:hsa00350	Tyrosine metabolism	GOT1 GOT2 TAT IL4I1 HPD HGD GSTZ1 FAH TYR TH DCT TYRP1 DDC DBH PNMT COMT MAOB MAOA AOC3 AOC2 ALDH3B1 ALDH3B2 ALDH1A3 ALDH3A1 ADH1A ADH1B ADH1C ADH7 ADH4 ADH5 ADH6 TPO AOX1 FAHD1 MIF
path:hsa00360	Phenylalanine metabolism	PAH DDC AOC3 AOC2 MAOB MAOA ALDH3B1 ALDH3B2 ALDH1A3 ALDH3A1 GLYAT GOT1 GOT2 TAT IL4I1 HPD MIF
path:hsa00380	Tryptophan metabolism	TDO2 IDO1 IDO2 AFMID KMO KYNU HAAO ACMSD OGDHL OGDH GCDH HADHA EHHADH ECHS1 HADH ACAT2 ACAT1 CCBL2 CCBL1 AADAT TPH2 TPH1 DDC MAOB MAOA ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1 AOX1 ASMT AANAT CYP1A1 CYP1A2 CYP1B1 INMT IL4I1 AOC1 CAT
path:hsa00400	Phenylalanine, tyrosine and tryptophan biosynthesis	GOT1 GOT2 TAT IL4I1 PAH

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00410	beta-Alanine metabolism	GADL1 GAD1 GAD2 CNDP1 CARNS1 CNDP2 ABAT SRM SMS SMOX AOC3 AOC2 ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1 ALDH3B1 ALDH3B2 ALDH1A3 ALDH3A1 DPYD DPYS UPB1 HIBCH HADHA EHHADH ECHS1 ACADM MLYCD ALDH6A1
path:hsa00430	Taurine and hypotaurine metabolism	CDO1 GAD1 GAD2 CSAD GADL1 ADO GGT7 GGT6 GGT1 GGT5 BAAT
path:hsa00450	Selenocompound metabolism	MTR CTH SCLY CCBL2 CCBL1 TXNRD1 TXNRD2 TXNRD3 INMT PAPSS2 PAPSS1 SEPHS2 SEPHS1 PSTK SEPSECS MARS MARS2
path:hsa00460	Cyanoamino acid metabolism	GBA3 GGT7 GGT6 GGT1 GGT5 SHMT2 SHMT1
path:hsa00471	D-Glutamine and D-glutamate metabolism	GLS2 GLS GLUD2 GLUD1
path:hsa00472	D-Arginine and D-ornithine metabolism	DAO
path:hsa00480	Glutathione metabolism	GGT7 GGT6 GGT1 GGT5 GGCT OPLAH GCLC GCLM GSS LAP3 ANPEP GSTA5 GSTA2 GSTA4 GSTO2 GSTM4 GSTT2 GSTT1 GSTM3 MGST1 MGST3 GSTP1 GSTM1 GSTM5 MGST2 GSTA1 GSTM2 GSTA3 GSTO1 GSTT2B GSTK1 GSR IDH1 IDH2 PGD G6PD TXNDC12 GPX6 GPX7 GPX2 GPX3 GPX1 GPX5 GPX8 GPX4 ODC1 SRM SMS RRM1 RRM2B RRM2
path:hsa00500	Starch and sucrose metabolism	MGAM GAA GANC SI GBA3 TREH UGDH UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 GUSB KL UXS1 UGP2 PGM1 PGM2 HK3 HK1 HK2 HKDC1 GCK G6PC G6PC2 G6PC3 PGM2L1 GPI GBE1 PYGL PYGM PYGB GYS2 GYS1 AMY1C AMY2A AMY1A AMY2B AMY1B AGL ENPP1 ENPP3
path:hsa00510	N-Glycan biosynthesis	DOLK DPAGT1 ALG5 ALG13 ALG14 DPM1 DPM2 DPM3 ALG1 ALG2 ALG11 ALG3 ALG9 ALG12 ALG6 ALG8 ALG10 ALG10B STT3A STT3B RPN1 RPN2 DAD1 TUSC3 DDOST DOLPP1 MOGS GANAB MAN1B1 MAN1A2 MAN1C1 MAN1A1 MGAT1 MAN2A1 MAN2A2 MGAT2 FUT8 B4GALT1 B4GALT2 B4GALT3 ST6GAL1 ST6GAL2 MGAT3 MGAT4A MGAT4B MGAT4D MGAT5 MGAT5B MGAT4C
path:hsa00511	Other glycan degradation	NEU1 NEU3 NEU4 NEU2 GLB1 HEXA HEXB HEXDC MAN2C1 MAN2B1 MAN2B2 MANBA ENGASE FUCA1 FUCA2 AGA GBA GBA2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00512	Mucin type O-Glycan biosynthesis	GALNTL6 GALNT5 WBSCR17 GALNT11 GALNT12 GALNT13 GALNT14 GALNT16 GALNT15 GALNT18 GALNTL5 GALNT10 GALNT2 GALNT3 GALNT1 GALNT6 GALNT4 GALNT9 GALNT7 GALNT8 POC1B-GALNT4 C1GALT1 C1GALT1C1 GCNT1 GCNT3 GCNT4 ST3GAL1 ST3GAL2 ST6GALNAC1 B3GNT6 B4GALT5
path:hsa00514	Other types of O-glycan biosynthesis	POMT1 POMT2 POMGNT1 B4GALT1 B4GALT2 B4GALT3 ST3GAL3 FUT9 FUT4 FUT7 MGAT5B B3GALT4 B3GAT1 B3GAT2 CHST10 OGT EOGT POFUT1 POFUT2 MFNG LFNG RFNG ST6GAL1 ST6GAL2 B3GALTL POGLUT1 GXYLT1 GXYLT2 COLGALT1 COLGALT2 PLOD3
path:hsa00520	Amino sugar and nucleotide sugar metabolism	CHIA CHIT1 HEXA HEXB NAGK PGM3 UAP1 UAP1L1 GNE RENBP NANS NANP NPL CMAS CYB5R1 CYB5R3 CYB5R2 CYB5RL CYB5R4 HK3 HK1 HK2 HKDC1 AMDHD2 GNPNAT1 GNPDA1 GNPDA2 GFPT2 GFPT1 UXS1 GCK GPI PGM1 PGM2 UGP2 UGDH GALK1 GALT GALE PMM2 PMM1 GMPPB GMPPA GMDS MPI FUK FPGT TSTA3
path:hsa00524	Butirosin and neomycin biosynthesis	HK3 HK1 HK2 HKDC1 GCK
path:hsa00531	Glycosaminoglycan degradation	HYAL2 HYAL1 SPAM1 HYAL4 HYAL3 GUSB IDS IDUA ARSB HPSE HPSE2 SGSH HGSNAT NAGLU GALNS GLB1 GNS HEXA HEXB
path:hsa00532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	XYLT1 XYLT2 B4GALT7 B3GALT6 B3GAT3 CSGALNACT1 CSGALNACT2 CHSY3 CHSY1 CHPF CHPF2 DSE CHST11 CHST12 CHST13 CHST3 CHST7 CHST15 UST CHST14
path:hsa00533	Glycosaminoglycan biosynthesis - keratan sulfate	FUT8 B4GALT1 B4GALT2 B4GALT3 B3GNT2 B4GAT1 CHST6 B4GALT4 B3GNT7 CHST1 ST3GAL3 CHST2 CHST4 ST3GAL1 ST3GAL2
path:hsa00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	XYLT1 XYLT2 B4GALT7 B3GALT6 B3GAT3 EXTL2 EXTL3 EXTL1 EXT1 EXT2 NDST1 NDST2 NDST3 NDST4 GLCE HS2ST1 HS6ST1 HS6ST2 HS6ST3 HS3ST1 HS3ST2 HS3ST3B1 HS3ST3A1 HS3ST5

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00561	Glycerolipid metabolism	GLYCTK ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1 AKR1B1 AKR1B10 AKR1A1 DAK GK2 GK GPAM GPAT2 AGPAT6 AGPAT9 AGPAT1 AGPAT2 AGPAT5 LCLAT1 MBOAT1 MBOAT2 AGPAT3 AGPAT4 PPAP2A PPAP2B PPAP2C LPIN1 LPIN3 LPIN2 DGKZ DGKD DGKI DGKA DGKE DGKB DGKH DGKG DGKQ DGKK DGAT1 DGAT2 MOGAT3 LIPG LIPC CEL PNPLA2 PNPLA3 PNLIP PNLIPRP1 PNLIPRP2 PNLIPRP3 LIPF LPL AGK MGLL MOGAT1 MOGAT2 GLA
path:hsa00562	Inositol phosphate metabolism	PIK3C3 MTM1 MTMR1 MTMR2 MTMR3 MTMR4 MTMR8 MTMR6 MTMR7 MTMR14 PI4KA PI4KB PI4K2A PI4K2B PIP5K1C PIP5K1A PIP5K1B PIP5KL1 SYNJ1 SYNJ2 OCRL INPP5E INPP5B PIK3CA PIK3CD PIK3CB PIK3CG PTEN INPP5D INPPL1 PIK3C2G PIK3C2A PIK3C2B INPP4A INPP4B PIKFYVE PIP4K2C PIP4K2A PIP4K2B PLCB1 PLCB2 PLCB3 PLCB4 PLCD1 PLCD3 PLCD4 PLCE1 PLCG1 PLCG2 PLCZ1 PLCH1 PLCH2 IMPA2 IMPA1 IMPAD1 CDIPT ISYNA1 INPP1 INPP5K INPP5J INPP5A MINPP1 ITPKB ITPKA ITPKC ITPK1 IPMK IPPK MIOX ALDH6A1 TPI1
path:hsa00563	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	PIGA PIGH PIGC PIGQ PIGP PIGY DPM2 PIGL PIGW GPLD1 PIGM PIGX PIGV PIGN PIGB PIGF PIGO GPAA1 PIGK PIGS PIGT PIGU PGAP1 PIGZ PIGG
path:hsa00564	Glycerophospholipid metabolism	GPD1L GPD1 GPD2 GPAM GPAT2 AGPAT6 AGPAT9 AGPAT1 AGPAT2 AGPAT5 LCLAT1 MBOAT1 MBOAT2 AGPAT3 AGPAT4 GNPAT ADPRM PPAP2A PPAP2B PPAP2C LPIN1 LPIN3 LPIN2 DGKZ DGKD DGKI DGKA DGKE DGKB DGKH DGKG DGKQ DGKK CHPT1 CEPT1 PLD1 PLD2 PLD3 PLD4 LCAT PLA2G10 PLA2G2D PLA2G2E PLA2G3 PLA2G2F PLA2G12A PLA2G12B PLA2G1B PLA2G5 PLA2G2A PLA2G2C PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLA2G6 PLB1 PLA2G16 LPCAT2 LPCAT1 LPCAT4 LPCAT3 LYPLA1 PLA2G15 LYPLA2 PNPLA6 PNPLA7 GPCPD1 CHAT ACHE CHKA CHKB PHOSPHO1 PCYT1B PCYT1A EPT1 ETNK1 ETNK2 PCYT2 ETNPPL PEMT CDS1 CDS2 PTDSS1 PTDSS2 PISD PGS1 CRLS1 TAZ LPGAT1 CDIPT MBOAT7

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00565	Ether lipid metabolism	AGPS PPAP2A PPAP2B PPAP2C EPT1 CHPT1 CEPT1 PLA2G10 PLA2G2D PLA2G2E PLA2G3 PLA2G2F PLA2G12A PLA2G12B PLA2G1B PLA2G5 PLA2G2A PLA2G2C PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLA2G6 PLB1 PLA2G16 LPCAT4 ENPP6 ENPP2 TMEM86B PLD1 PLD2 PLD3 PLD4 UGT8 GAL3ST1 LPCAT2 LPCAT1 PAFAH1B1 PAFAH1B2 PAFAH1B3 PLA2G7 PAFAH2
path:hsa00590	Arachidonic acid metabolism	PLA2G10 PLA2G2D PLA2G2E PLA2G3 PLA2G2F PLA2G12A PLA2G12B PLA2G1B PLA2G5 PLA2G2A PLA2G2C PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLA2G6 PLB1 PLA2G16 PTGS1 PTGS2 PTGES PTGES2 PTGES3 CBR1 CBR3 FAM213B TBXAS1 HPGDS PTGDS AKR1C3 PTGIS ALOX5 LTA4H CYP4F2 CYP4F3 LTC4S GGT1 GGT5 GPX6 GPX7 GPX2 GPX3 GPX1 GPX5 GPX8 CYP2E1 CYP2J2 CYP2U1 CYP4A11 CYP2C19 CYP4F8 ALOX12 ALOX12B ALOX15B CYP2B6 CYP2C8 CYP2C9 EPHX2 ALOX15
path:hsa00591	Linoleic acid metabolism	PLA2G10 PLA2G2D PLA2G2E PLA2G3 PLA2G2F PLA2G12A PLA2G12B PLA2G1B PLA2G5 PLA2G2A PLA2G2C PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLA2G6 PLB1 PLA2G16 ALOX15 CYP1A2 CYP2C8 CYP2C9 CYP2C19 CYP2J2 CYP2E1 CYP3A4
path:hsa00592	alpha-Linolenic acid metabolism	PLA2G10 PLA2G2D PLA2G2E PLA2G3 PLA2G2F PLA2G12A PLA2G12B PLA2G1B PLA2G5 PLA2G2A PLA2G2C PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLA2G6 PLB1 PLA2G16 FADS2 ACOX3 ACOX1 ACAA1
path:hsa00600	Sphingolipid metabolism	SPTLC1 SPTLC2 SPTLC3 KDSR CERS2 CERS3 CERS6 CERS1 CERS4 CERS5 ASAH1 ASAH2 ACER2 ACER1 ACER3 DEGS1 DEGS2 SGMS1 SGMS2 SMPD1 SMPD2 SMPD3 SMPD4 ENPP7 CERK PPAP2A PPAP2B PPAP2C SGPP1 SGPP2 SPHK1 SPHK2 SGPL1 UGCG GBA GBA2 B4GALT6 GLB1 UGT8 GALC GAL3ST1 ARSA NEU1 NEU3 NEU4 NEU2 GLA
path:hsa00601	Glycosphingolipid biosynthesis - lacto and neolacto series	B3GNT5 B3GALT1 B3GALT2 B3GALT5 FUT1 FUT2 FUT3 ST3GAL3 ST3GAL4 ABO B4GALT1 B4GALT2 B4GALT3 B4GALT4 FUT9 FUT4 FUT5 FUT6 FUT7 ST3GAL6 ST8SIA1 B3GNT2 B4GAT1 B3GNT3 B3GNT4 GCNT2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00603	Glycosphingolipid biosynthesis - globo series	A4GALT GLA B3GALNT1 HEXA HEXB B3GALT5 GBGT1 NAGA FUT1 FUT2 ST3GAL1 ST3GAL2 ST8SIA1 FUT9
path:hsa00604	Glycosphingolipid biosynthesis - ganglio series	B4GALNT1 B3GALT4 ST3GAL2 ST3GAL1 ST8SIA5 ST3GAL5 ST8SIA1 SLC33A1 ST6GALNAC3 ST6GALNAC4 ST6GALNAC5 ST6GALNAC6 GLB1 HEXA HEXB
path:hsa00620	Pyruvate metabolism	ACSS1 ACSS2 PDHA2 PDHA1 PDHB DLAT DLD PKM PKLR ACACA ACACB ACYP2 ACYP1 ALDH2 ALDH3A2 ALDH1B1 ALDH7A1 ALDH9A1 ACOT12 LDHAL6A LDHAL6B LDHA LDHB LDHC LDHD GLO1 HAGHL HAGH GRHPR ME2 ME3 ME1 PC MDH1 MDH2 FH PCK1 PCK2 ACAT2 ACAT1
path:hsa00630	Glyoxylate and dicarboxylate metabolism	MDH1 MDH2 CS ACO2 ACO1 ACAT2 ACAT1 MCEE PCCA PCCB MUT HAO2 HAO1 CAT GRHPR PGP AGXT GLUL SHMT2 SHMT1 GLDC AMT DLD GCSH HYI GLYCTK HOGA1 AFMID
path:hsa00640	Propanoate metabolism	ACSS1 ACSS2 ACSS3 BCKDHA BCKDHB DBT DLD ACADM HADHA EHHADH ECHS1 HIBCH ACACA ACACB MLYCD ABAT PCCA PCCB ECHDC1 MCEE MUT SUCLG1 SUCLG2 SUCLA2 ALDH6A1 LDHAL6A LDHAL6B LDHA LDHB LDHC ACAT2 ACAT1
path:hsa00650	Butanoate metabolism	ACAT2 ACAT1 HADH HADHA EHHADH ECHS1 ACADS ACSM1 ACSM2A ACSM4 ACSM5 ACSM3 ACSM2B L2HGDH GAD1 GAD2 ABAT ALDH5A1 HMGCS1 HMGCS2 HMGCL HMGCLL1 OXCT1 OXCT2 AACS BDH1 BDH2
path:hsa00670	One carbon pool by folate	DHFR DHFRL1 MTHFD1L MTHFD1 MTHFD2 MTHFD2L SHMT2 SHMT1 GART ATIC FTCD MTFMT AMT MTR TYMS ALDH1L1 ALDH1L2 MTHFR MTHFS ST20-MTHFS
path:hsa00730	Thiamine metabolism	NFS1 NTPCR TPK1 THTPA
path:hsa00740	Riboflavin metabolism	RFK FLAD1 BLVRB TYR
path:hsa00750	Vitamin B6 metabolism	PNPO PDXK PDXP PHOSPHO2 AOX1 PSAT1
path:hsa00760	Nicotinate and nicotinamide metabolism	QPRT NAPRT PNP NMRK1 NMRK2 NT5C2 NT5C1A NT5E NT5C1B NT5C NT5M NT5C3A NT5C3B NT5C1B-RDH14 NMNAT3 NMNAT1 NMNAT2 ENPP1 ENPP3 NUDT12 NADSYN1 NAMPT CD38 BST1 NADK NADK2 NNT NNMT AOX1
path:hsa00770	Pantothenate and CoA biosynthesis	PANK1 PANK4 PANK3 PANK2 PPCS PPCDC ENPP1 ENPP3 COASY AASDHPPT VNN1 VNN2 BCAT2 BCAT1 DPYD DPYS UPB1 GADL1
path:hsa00780	Biotin metabolism	OXSM HLCS BTD
path:hsa00785	Lipoic acid metabolism	LIAS LIPT2 LIPT1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00790	Folate biosynthesis	GCH1 ALPL ALPI ALPP ALPPL2 DHFR DHFRL1 FPGS GGH PTS SPR QDPR MOCS1 MOCS2
path:hsa00830	Retinol metabolism	BCO1 ADH1A ADH1B ADH1C ADH7 ADH4 ADH5 ADH6 DHRS3 DHRS4 DHRS4L1 DHRS4L2 DHRS9 RDH8 RDH10 RDH11 RDH12 RDH16 SDR16C5 HSD17B6 LRAT DGAT1 AWAT2 PNPLA4 RPE65 RDH5 AOX1 ALDH1A2 ALDH1A1 CYP26A1 CYP26B1 CYP26C1 CYP1A1 CYP1A2 CYP2A6 CYP2B6 CYP2C8 CYP2C9 CYP2C18 CYP2S1 CYP3A4 CYP3A5 CYP3A7 CYP3A7-CYP3A51P CYP4A11 UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 RETSAT
path:hsa00860	Porphyrin and chlorophyll metabolism	ALAS1 ALAS2 EARS2 EPRS ALAD HMBS UROS UROD CPOX PPOX FECH COX10 COX15 MMAB HMOX1 HMOX2 BLVRA BLVRB UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 GUSB HCCS CP HEPH FXN
path:hsa00900	Terpenoid backbone biosynthesis	ACAT2 ACAT1 HMGCS1 HMGCS2 HMGCR MVK PMVK MVD IDI1 IDI2 FDPS GGPS1 PDSS1 PDSS2 DHDDS NUS1 FNTA FNTB RCE1 ZMPSTE24 ICMT PCYOX1
path:hsa00910	Nitrogen metabolism	GLUD2 GLUD1 GLUL CPS1 CA13 CA1 CA6 CA7 CA12 CA5B CA14 CA9 CA3 CA5A CA8 CA2 CA4
path:hsa00920	Sulfur metabolism	PAPSS2 PAPSS1 BPNT1 IMPAD1 SUOX CYCS ETHE1 MPST TST SQRDL
path:hsa00970	Aminoacyl-tRNA biosynthesis	TRNA TRNR TRNN TRND TRNC TRNQ TRNE TRNG TRNH TRNI TRNL2 TRNL1 TRNK TRNM TRNF TRNP TRNS2 TRNS1 TRNT TRNW TRNY TRNV EARS2 EPRS QRSL1 GATB GATC QARS AARS2 AARS DARS2 DARS NARS2 NARS GARS TARS2 TARS TARSL2 SARS2 SARS PSTK SEPSECS CARS CARS2 MARS MARS2 MTFMT VARS2 VARS LARS LARS2 IARS IARS2 KARS RARS2 RARS PARS2 HARS HARS2 FARSA FARS2 FARSB YARS YARS2 WARS2 WARS

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa00980	Metabolism of xenobiotics by cytochrome P450	CYP1A1 CYP2C9 CYP3A4 CYP1B1 GSTA5 GSTA2 GSTA4 GSTO2 GSTM4 GSTT2 GSTT1 GSTM3 MGST1 MGST3 GSTP1 GSTM1 GSTM5 MGST2 GSTA1 GSTM2 GSTA3 GSTO1 GSTT2B GSTK1 EPHX1 CYP2B6 SULT2A1 CYP1A2 CYP2A6 CYP2E1 CYP2F1 CYP2S1 AKR1C2 AKR1C4 AKR1C1 DHDH CYP2A13 CYP2D6 HSD11B1 CBR1 CBR3 UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 CYP3A5 AKR7A2 AKR7A3 ALDH3B1 ALDH3B2 ALDH1A3 ALDH3A1 ADH1A ADH1B ADH1C ADH7 ADH4 ADH5 ADH6
path:hsa00982	Drug metabolism - cytochrome P450	CYP2D6 CYP2C9 CYP3A4 FMO1 FMO2 FMO5 FMO3 FMO4 CYP2C19 CYP2B6 CYP3A5 GSTA5 GSTA2 GSTA4 GSTO2 GSTM4 GSTT2 GSTT1 GSTM3 MGST1 MGST3 GSTP1 GSTM1 GSTM5 MGST2 GSTA1 GSTM2 GSTA3 GSTO1 GSTT2B GSTK1 ADH1A ADH1B ADH1C ADH7 ADH4 ADH5 ADH6 ALDH3B1 ALDH3B2 ALDH1A3 ALDH3A1 MAOB MAOA AOX1 UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 CYP1A2 CYP2E1 CYP2C8 CYP2A6
path:hsa00983	Drug metabolism - other enzymes	HPRT1 IMPDH1 IMPDH2 GMPS TPMT XDH ITPA CES1 CES2 UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 GUSB CYP3A4 CDA TYMP DPYD DPYS UPB1 CYP2A6 UPP2 UPP1 UCK1 UCK2 UCKL1 TK2 TK1 UMPS NAT2 NAT1
path:hsa01040	Biosynthesis of unsaturated fatty acids	ELOVL6 HSD17B12 HACD2 HACD1 HACD4 HACD3 TECR PECR SCD SCD5 FADS2 ELOVL5 FADS1 ELOVL2 ACOX3 ACOX1 HADHA ACAA1 ACOT4 ACOT2 ACOT1 ACOT7 BAAT
path:hsa01100	Metabolic pathways	NA
path:hsa01200	Carbon metabolism	NA
path:hsa01210	2-Oxocarboxylic acid metabolism	NA
path:hsa01212	Fatty acid metabolism	NA
path:hsa01220	Degradation of aromatic compounds	NA

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa01230	Biosynthesis of amino acids	NA
path:hsa02010	ABC transporters	ABCA1 ABCA2 ABCA3 ABCA4 ABCA7 ABCA12 ABCA13 ABCA5 ABCA6 ABCA8 ABCA9 ABCA10 TAP1 TAP2 ABCB8 ABCB9 ABCB10 ABCB1 ABCB4 ABCB5 ABCB6 ABCB7 ABCB11 ABCC1 ABCC2 ABCC3 ABCC5 ABCC6 ABCC8 ABCC9 ABCC11 ABCC12 ABCC4 CFTR ABCC10 ABCD1 ABCD2 ABCD3 ABCD4 ABCG1 ABCG4 ABCG2 ABCG5 ABCG8
path:hsa03008	Ribosome biogenesis in eukaryotes	CSNK2B CSNK2A1 CSNK2A2 NOL6 RRP7A WDR43 CIRH1A UTP15 HEATR1 WDR75 UTP18 WDR36 TBL3 WDR3 UTP6 PWP2 MPHOSPH10 LOC643802 IMP3 IMP4 TCOF1 FBL NOP56 NOP58 NHP2L1 DKC1 NHP2 GAR1 NOP10 FCF1 UTP14C UTP14A DROSHA EMG1 BMS1 NAT10 RCL1 POP1 RPP38 POP4 POP5 RPP25L RPP25 POP7 RPP30 RPP40 LOC100288562 REXO1 LOC101929601 LOC101930111 LOC81691 LOC101929627 REXO2 XRN1 XRN2 GTPBP4 GNL2 GNL3 GNL3L NVL MDN1 RBM28 NOB1 RAN XPO1 NMD3 NXF1 NXF2 NXF2B NXF5 NXF3 NXT1 NXT2 EIF6 SBDS EFTUD1 LSG1 SPATA5 AK6 RIOK1 RIOK2 RNR1 RNR2 SNORD3A SNORD3C SNORD3B-2 SNORD3B-1 RMRP

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa03010	Ribosome	MRPS2 MRPS5 MRPS6 MRPS7 MRPS9 MRPS10 MRPS11 MRPS12 MRPS14 MRPS15 MRPS16 MRPS17 MRPS18C MRPS18A MRPS21 RPS2 RPS3 RPS3A RPS4Y1 RPS4X RPS4Y2 RPS5 RPS6 RPS7 RPS8 RPS9 RPS10 RPS10-NUDT3 RPS11 RPS12 RPS13 RPS14 RPS15 RPS15A RPS16 RPS17 RPS18 RPS19 RPS20 RPS21 RPS23 RPS24 RPS25 RPS26 LOC101929876 RPS27 RPS27L RPS27A RPS28 RPS29 FAU RPSA MRPL1 MRPL2 MRPL3 MRPL4 MRPL12 MRPL9 MRPL10 MRPL11 MRPL13 MRPL14 MRPL15 MRPL16 MRPL17 MRPL18 MRPL19 MRPL20 MRPL21 MRPL22 MRPL23 LOC102724828 MRPL24 MRPL27 MRPL28 MRPL30 MRPL32 MRPL33 MRPL34 MRPL35 MRPL36 RPL3L RPL3 RPL4 RPL5 RPL6 RPL7 RPL7A RPL8 RPL9 RPL10L RPL10 RPL10A RPL11 RPL12 RPL13 RPL13A RPL14 RPL15 RPL17 RPL18 RPL18A RPL19 RPL21 RPL22L1 RPL22 RPL23 RPL23A RSL24D1 RPL24 RPL26 RPL26L1 RPL27 RPL27A RPL28 RPL29 RPL30 RPL31 RPL32 RPL34 RPL35 RPL35A RPL36 RPL37 RPL37A RPL38 RPL39 UBA52 RPL41 RPL36AL RPL36A RPL36A-HNRNPH2 RPLP0 RPLP1 RPLP2 RNR1 RNR2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa03013	RNA transport	POP1 RPP38 POP4 POP5 RPP25L RPP25 POP7 RPP14 RPP30 RPP21 RPP40 ELAC2 ELAC1 TRNT1 RAN XPOT XPO5 EEF1A1 EEF1A2 XPO1 PHAX NCBP1 NCBP2 CLNS1A PRMT5 TGS1 KPNB1 SNUPN NMD3 WIBG TPR NUP50 NUP153 SENP2 NUP98 RAE1 SEC13 SEH1L NUP133 NUP107 NUP37 NUP160 NUP85 NUP43 NUP62 NUPL1 NUP54 NUP93 NUP205 NUP188 NUP155 NUP35 NUP210 NUP210L NDC1 POM121C POM121 POM121L2 NUP214 NUP88 RANBP2 RGPD4 RGPD1 RGPD3 RGPD8 RGPD2 RGPD5 RANGAP1 UBE2I SUMO3 SUMO2 SUMO1 SUMO4 LOC101929087 AAAS NUPL2 SMN1 SMN2 GEMIN2 DDX20 GEMIN4 GEMIN5 GEMIN6 GEMIN7 GEMIN8 STRAP EIF3J EIF3I EIF3H EIF3G EIF3F EIF3E EIF3D EIF3C EIF3CL EIF3B EIF3A EIF1AY EIF1AX EIF1B EIF1 EIF5 EIF2S1 EIF2S2 EIF2S3 EIF2B1 EIF2B2 EIF2B4 EIF2B3 EIF2B5 EIF5B EIF4G3 EIF4G1 EIF4G2 EIF4E EIF4E2 EIF4E1B EIF4A1 EIF4A2 EIF4B PABPC1 PABPC5 PABPC3 PABPC1L2B PABPC1L PABPC1L2A PABPC4 PABPC4L PAIP1 EIF4EBP1 EIF4EBP2 EIF4EBP3 TACC3 CYFIP1 CYFIP2 FMR1 FXR1 FXR2 RBM8A MAGOH MAGOHB CASC3 EIF4A3 SAP18 PNN ACIN1 RNPS1 ALYREF DDX39B THOC1 THOC2 THOC5 THOC6 THOC7 THOC3 SRRM1 NXF1 NXF2 NXF2B NXF5 NXF3 NXT1 NXT2 UPF1 UPF2 UPF3B UPF3A RNU6-1 SNORD3A SNORD3C SNORD3B-2 SNORD3B-1
path:hsa03015	mRNA surveillance pathway	NCBP1 NCBP2 UPF3B UPF3A RBM8A MAGOH MAGOHB CASC3 EIF4A3 SAP18 PNN ACIN1 RNPS1 ALYREF NXF1 NXF2 NXF2B NXF5 NXF3 NXT1 NXT2 DDX39B WIBG SRRM1 RNGTT RNMT PABPN1 PABPN1L BCL2L2-PABPN1 NUDT21 CPSF6 CPSF7 PAPOLG PAPOLB PAPOLA CLP1 PCF11 CPSF1 CPSF2 CPSF3 CPSF4 FIP1L1 WDR33 WDR82 PPP1CA PPP1CB PPP1CC SSU72 CSTF1 CSTF2 CSTF2T CSTF3 SYMPK DAZAP1 MSI1 MSI2 PABPC1 PABPC5 PABPC3 PABPC1L2B PABPC1L PABPC1L2A PABPC4 PABPC4L ETF1 GSPT2 GSPT1 UPF1 UPF2 SMG1 SMG7 SMG5 SMG6 PPP2CA PPP2CB PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D PPP3R3B PPP2R3C PPP2R3A PPP2R5B PPP2R5C PPP2R5D PPP2R5E PPP2R5A HBS1L PELO

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa03018	RNA degradation	DCPS EXOSC1 EXOSC2 EXOSC3 EXOSC8 EXOSC6 EXOSC7 EXOSC4 EXOSC5 EXOSC9 DIS3 EXOSC10 C1D MPHOSPH6 DIS3L PAPD7 PAPD5 ZCCHC7 SKIV2L2 SKIV2L TTC37 WDR61 CNOT6 CNOT6L CNOT1 CNOT2 CNOT3 CNOT4 CNOT7 CNOT8 RQCD1 CNOT10 DHX36 PARN TOB1 TOB2 BTG3 BTG4 BTG1 BTG2 PABPC1 PABPC5 PABPC3 PABPC1L2B PABPC1L PABPC1L2A PABPC4 PABPC4L PAN2 PAN3 DCP1A DCP1B DCP2 DDX6 EDC3 EDC4 PATL1 XRN1 XRN2 NUDT16 LSM1 LSM2 LSM3 LSM4 LSM5 LSM6 LSM7 LSM8 ENO3 ENO2 ENO1 PNPT1 PFKM PFKP PFKL HSPA9 HSPD1
path:hsa03020	RNA polymerase	POLR2B POLR2A POLR2C POLR2J POLR2J3 POLR2J2 POLR2D POLR2G POLR2I POLR2E POLR2F POLR2H POLR2K POLR2L POLR3B POLR3A POLR1D POLR1C POLR3C POLR3D POLR3E LOC101060521 POLR3K POLR3H POLR3GL POLR3G POLR3F POLR1B POLR1A ZNRD1 POLR1E TWISTNB
path:hsa03022	Basal transcription factors	GTF2A1L GTF2A1 GTF2A2 GTF2B TBPL2 TBPL1 TBP TAF1 TAF1L TAF2 TAF7 TAF7L TAF8 TAF3 TAF10 TAF5L TAF5 TAF4B TAF4 TAF12 TAF6 TAF6L TAF9 TAF9B TAF11 TAF13 TAF15 GTF2E1 GTF2E2 GTF2F1 GTF2F2 GTF2H1 GTF2H2 GTF2H2C_2 GTF2H2C GTF2H3 GTF2H4 ERCC3 ERCC2 GTF2H5 CDK7 MNAT1 CCNH GTF2IRD1 GTF2I
path:hsa03030	DNA replication	SSBP1 RNASEH1 RPA1 PCNA DNA2 FEN1 LIG1 POLA1 POLA2 PRIM1 PRIM2 POLD1 POLD2 POLD3 POLD4 POLE POLE2 POLE3 POLE4 MCM2 MCM3 MCM4 MCM5 MCM6 MCM7 RPA2 RPA4 RPA3 RFC1 RFC4 RFC2 RFC5 RFC3 RNASEH2A RNASEH2B RNASEH2C

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa03040	Spliceosome	RNU6-1 DDX46 DDX39B DHX16 DHX38 CDC40 PRPF18 DHX8 SLU7 DHX15 SNRPB SNRPD1 SNRPD2 SNRPD3 SNRPE SNRPF SNRPG SNRNP70 SNRPA SNRPC PRPF40B PRPF40A RBM25 DDX5 TCERG1 SNRPA1 SNRPB2 SF3A1 SF3A2 SF3A3 SF3B1 SF3B2 SF3B3 SF3B4 SF3B5 SF3B6 PHF5A DDX42 U2AF1 LOC102724594 U2AF1L4 U2AF2 PUF60 SMNDC1 RBM17 CHERP U2SURP RP9 LSM2 LSM3 LSM4 LSM5 LSM6 LSM7 LSM8 PRPF3 PRPF4 PPIH PRPF31 NHP2L1 SNRNP27 USP39 SART1 ZMAT2 PRPF38A PRPF38B EFTUD2 SNRNP200 PRPF6 PRPF8 SNRNP40 DDX23 TXNL4A PRPF19 CDC5L BCAS2 PLRG1 CWC15 CTNNBL1 HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 PQBP1 WBP11 SNW1 XAB2 SYF2 CRNKL1 ISY1 ISY1-RAB43 PPIL1 PPIE CCDC12 RBM22 BUD31 AQR ACIN1 EIF4A3 RBM8A MAGOH MAGOHB THOC1 THOC2 THOC3 ALYREF NCBP1 NCBP2 HNRNPA3 HNRNPA1 HNRNPA1L2 HNRNPC RBMX RBMXL1 RBMXL2 RBMXL3 HNRNPK HNRNPM HNRNPU PCBP1 SRSF1 SRSF9 SRSF2 SRSF8 SRSF3 SRSF4 SRSF5 SRSF6 SRSF7 TRA2A TRA2B SRSF10
path:hsa03050	Proteasome	PSMD3 PSMD12 PSMD11 PSMD6 PSMD7 PSMD13 PSMD14 PSMD8 SHFM1 PSMD4 PSMD2 PSMD1 PSMC2 PSMC1 PSMC5 PSMC6 PSMC3 PSMC4 PSME1 PSME2 PSME3 PSME4 PSMA6 PSMA2 PSMA4 PSMA7 PSMA8 PSMA5 PSMA1 PSMA3 PSMB6 PSMB7 PSMB3 PSMB2 PSMB5 PSMB1 PSMB4 PSMB9 PSMB10 PSMB8 PSMB11 IFNG PSMF1 POMP
path:hsa03060	Protein export	OXA1L SRP54 SEC61A1 SEC61A2 SEC61B SEC61G SEC62 SEC63 HSPA5 SRP9 SRP14 SRP72 SRP68 SRP19 SRPR SRPRB SPCS1 SPCS2 SPCS3 SEC11C SEC11A IMMP1L IMMP2L
path:hsa03320	PPAR signaling pathway	CD36 SLC27A1 SLC27A4 SLC27A2 SLC27A5 SLC27A6 FABP1 FABP2 FABP3 FABP4 FABP5 FABP6 FABP7 PPARA RXRA RXRB RXRG PPARD PPARG APOA1 APOA2 APOC3 APOA5 PLTP FADS2 SCD SCD5 CYP7A1 CYP8B1 NR1H3 CYP27A1 DBI LPL ACSL6 ACSL4 ACSL1 ACSL5 ACSL3 ACSBG1 ACSBG2 OLR1 EHHADH ACAA1 SCP2 ACOX3 ACOX1 ACOX2 CPT1A CPT1B CPT1C CPT2 ACADL ACADM ANGPTL4 SORBS1 PLIN1 ADIPOQ MMP1 UCP1 ILK PDPK1 UBC PCK1 PCK2 GK2 GK AQP7 HMGCS2 ME1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa03410	Base excision repair	OGG1 NTHL1 NEIL1 NEIL2 NEIL3 UNG TDG SMUG1 MUTYH MPG MBD4 APEX1 APEX2 POLB POLL HMGB1 XRCC1 PCNA POLD1 POLD2 POLD3 POLD4 POLE POLE2 POLE3 POLE4 LIG1 LIG3 PARP2 PARP1 PARP3 PARP4 FEN1
path:hsa03420	Nucleotide excision repair	RBX1 CUL4B CUL4A DDB1 DDB2 XPC RAD23B RAD23A CETN2 ERCC8 ERCC6 CDK7 MNAT1 CCNH ERCC3 ERCC2 GTF2H5 GTF2H1 GTF2H2 GTF2H2C_2 GTF2H2C GTF2H3 GTF2H4 ERCC5 BIVM-ERCC5 XPA RPA1 RPA2 RPA3 RPA4 ERCC4 ERCC1 POLD1 POLD2 POLD3 POLD4 POLE POLE2 POLE3 POLE4 PCNA RFC1 RFC4 RFC2 RFC5 RFC3 LIG1
path:hsa03430	Mismatch repair	SSBP1 PMS2 MLH1 MSH6 MSH2 MSH3 MLH3 RFC1 RFC4 RFC2 RFC5 RFC3 PCNA EXO1 RPA1 RPA2 RPA3 RPA4 POLD1 POLD2 POLD3 POLD4 LIG1
path:hsa03440	Homologous recombination	SSBP1 RAD50 MRE11A NBN RPA1 RPA2 RPA3 RPA4 RAD51 RAD52 BRCA2 SHFM1 SYCP3 RAD51B RAD51C RAD51D XRCC2 XRCC3 RAD54L RAD54B POLD1 POLD2 POLD3 POLD4 BLM TOP3A TOP3B MUS81 EME1
path:hsa03450	Non-homologous end-joining	XRCC6 XRCC5 DCLRE1C PRKDC POLL POLM DNTT LIG4 XRCC4 NHEJ1 RAD50 MRE11A FEN1
path:hsa03460	Fanconi anemia pathway	ATRIP ATR FANCM C19orf40 APITD1 STRA13 TELO2 HES1 C17orf70 FANCA FANCB FANCC FANCE FANCF FANCG FANCL WDR48 USP1 UBE2T FANCI FANCD2 BRCA2 PALB2 RAD51C RAD51 BRCA1 BRIP1 FAN1 MLH1 PMS2 REV1 REV3L POLH POLI POLK POLN RMI1 RMI2 TOP3A TOP3B BLM RPA1 RPA2 RPA3 RPA4 MUS81 EME1 EME2 ERCC4 ERCC1 SLX1A SLX1B SLX4

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04010	MAPK signaling pathway	CACNA1A CACNA1B CACNA1C CACNA1D CACNA1E CACNA1F CACNA1G CACNA1H CACNA1I CACNA1S CACNA2D1 CACNA2D2 CACNA2D3 CACNA2D4 CACNB1 CACNB2 CACNB3 CACNB4 CACNG1 CACNG2 CACNG3 CACNG4 CACNG5 CACNG6 CACNG7 CACNG8 PRKACA PRKACB PRKACG PRKCA PRKCB PRKCG GNA12 GNG12 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 RASGRF1 RASGRF2 RASGRP1 RASGRP2 RASGRP3 RASGRP4 RAPGEF2 NF1 RASA1 RASA2 RAP1A RAP1B NGF BDNF NTF3 NTF4 EGF FGF1 FGF2 FGF3 FGF4 FGF17 FGF6 FGF7 FGF8 FGF9 FGF10 FGF11 FGF12 FGF13 FGF14 FGF16 FGF5 FGF18 FGF19 FGF20 FGF21 FGF22 FGF23 PDGFA PDGFB NTRK1 NTRK2 EGFR FGFR1 FGFR2 FGFR3 FGFR4 PDGFRA PDGFRB GRB2 SOS1 SOS2 HRAS KRAS NRAS RRAS RRAS2 MRAS BRAF RAF1 MOS MAP2K1 MAP2K2 LAMTOR3 MAPK1 MAPK3 MKNK1 MKNK2 RPS6KA3 RPS6KA1 RPS6KA2 RPS6KA6 ATF4 ELK1 ELK4 MYC SRF FOS MAPT STMN1 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F TNF IL1A IL1B TGFB1 TGFB2 TGFB3 TNFRSF1A IL1R1 IL1R2 TGFBR1 TGFBR2 FASLG FAS CD14 RAC1 RAC2 RAC3 CDC42 CASP3 TRAF2 DAXX TRAF6 GADD45A GADD45B GADD45G TAB1 TAB2 ECSIT MAP4K3 MAP4K4 MAP4K1 PAK1 PAK2 STK4 STK3 MAP4K2 MAP3K8 MAP3K1 MAP3K11 MAP3K2 MAP3K3 MAP3K13 MAP3K12 ZAK MAP3K6 MAP3K5 MAP3K7 MAP3K4 TAOK2 TAOK3 TAOK1 MAP2K4 MAP2K7 MAP2K3 MAP2K6 MAPK8IP1 MAPK8IP2 MAPK8IP3 FLNA FLNC FLNB CRK CRKL ARRB1 ARRB2 MAPK8 MAPK10 MAPK9 MAPK11 MAPK12 MAPK13 MAPK14 MAPKAPK5 MAPKAPK2 MAPKAPK3 RPS6KA5 RPS6KA4 CDC25B NFATC1 NFATC3 JUN JUND ATF2 TP53 DDIT3 MAX MEF2C HSPB1 AKT1 AKT2 AKT3 PPM1A PTPRR PTPN5 PTPN7 DUSP1 DUSP4 DUSP2 DUSP7 DUSP8 DUSP5 DUSP6 DUSP10 DUSP16 DUSP9 DUSP3 PPP5C PPP5D1 PPM1B HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 MECOM MAP2K5 MAPK7 NR4A1 MAP3K14 CHUK IKBKB IKBKG NLK NFKB1 NFKB2 RELA RELB

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04012	ErbB signaling pathway	EGF TGFA AREG EGFR ERBB2 PLCG1 PLCG2 CAMK2A CAMK2D CAMK2B CAMK2G PRKCA PRKCB PRKCG CBLC CBL CBLB STAT5A STAT5B SRC PTK2 CRK CRKL ABL1 ABL2 NCK1 NCK2 PAK1 PAK2 PAK3 PAK4 PAK6 PAK7 MAP2K4 MAP2K7 MAPK8 MAPK10 MAPK9 JUN ELK1 BTC HBEGF EREG ERBB3 NRG1 NRG2 ERBB4 SHC1 SHC2 SHC3 SHC4 GRB2 SOS1 SOS2 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 MYC GAB1 NRG3 NRG4 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 MTOR RPS6KB1 RPS6KB2 EIF4EBP1 BAD GSK3B CDKN1B CDKN1A

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04014	Ras signaling pathway	EGF FGF1 FGF2 FGF3 FGF4 FGF17 FGF6 FGF7 FGF8 FGF9 FGF10 FGF11 FGF12 FGF13 FGF14 FGF16 FGF5 FGF18 FGF19 FGF20 FGF21 FGF22 FGF23 NGF INS IGF1 PDGFA PDGFB PDGFC PDGFD CSF1 KITLG VEGFA VEGFB PGF VEGFC FIGF HGF ANGPT1 ANGPT2 ANGPT4 EFNA1 EFNA2 EFNA3 EFNA4 EFNA5 EGFR FGFR1 FGFR2 FGFR3 FGFR4 NGFR INSR IGF1R PDGFRA PDGFRB CSF1R KIT FLT1 FLT4 KDR MET TEK EPHA2 GRB2 GAB1 GAB2 SHC1 SHC2 SHC3 SHC4 PTPN11 SOS1 SOS2 PLCG1 PLCG2 RASGRP1 RASGRP2 RASGRP3 RASGRP4 ZAP70 LAT HTR7 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 PRKACA PRKACB PRKACG RASGRF1 RASGRF2 GRIN1 GRIN2A GRIN2B CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 HRAS KRAS NRAS MRAS RRAS RRAS2 NF1 RASA1 RASA2 RASA3 RASA4 RASA4B SYNGAP1 RASAL1 RASAL2 RASAL3 RASSF1 RASSF5 STK4 TIAM1 RAC1 RAC2 RAC3 PAK1 PAK2 PAK3 PAK4 PAK6 PAK7 RHOA PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 IKBKG CHUK IKBKB NFKB1 RELA BAD BCL2L1 FOXO4 FASLG MLLT4 SHOC2 RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 PLA1A PLA2G10 PLA2G2D PLA2G2E PLA2G3 PLA2G2F PLA2G12A PLA2G12B PLA2G1B PLA2G5 PLA2G2A PLA2G2C PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLA2G6 PLA2G16 ELK1 ETS1 ETS2 BRAP KSR1 KSR2 RAPGEF5 RAP1A RAP1B RALGDS RGL1 RGL2 RALA RALB MAPK8 MAPK10 MAPK9 EXOC2 TBK1 REL PLD1 PLD2 RALBP1 CDC42 PLCE1 PRKCA PRKCB PRKCG RIN1 ABL1 ABL2 RAB5A RAB5B RAB5C ARF6

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04015	Rap1 signaling pathway	RAP1A RAP1B MRAS RAPGEF5 DOCK4 GRIN1 GRIN2A GRIN2B CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 F2R F2RL3 P2RY1 FPR1 LPAR1 LPAR2 LPAR3 LPAR4 LPAR5 ADORA2A ADORA2B GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 RAPGEF3 RAPGEF4 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 RASGRP3 RASGRP2 EGF FGF1 FGF2 FGF3 FGF4 FGF17 FGF6 FGF7 FGF8 FGF9 FGF10 FGF11 FGF12 FGF13 FGF14 FGF16 FGF5 FGF18 FGF19 FGF20 FGF21 FGF22 FGF23 NGF INS IGF1 PDGFA PDGFB PDGFC PDGFD CSF1 KITLG VEGFA VEGFB PGF VEGFC FIGF HGF ANGPT1 ANGPT2 ANGPT4 EFNA1 EFNA2 EFNA3 EFNA4 EFNA5 EGFR FGFR1 FGFR2 FGFR3 FGFR4 NGFR INSR IGF1R PDGFRA PDGFRB CSF1R KIT FLT1 FLT4 KDR MET TEK EPHA2 CRK CRKL RAPGEF1 BCAR1 CDH1 CTNNB1 MAGI1 MAGI2 MAGI3 RAPGEF2 RAPGEF6 LAT FYB LCP2 SKAP1 PLCG1 PRKCA PRKCB PRKCG PRKD1 PRKD3 PRKD2 DRD2 CNR1 GNAI1 GNAI3 GNAI2 GNAO1 RAP1GAP SIPA1L1 SIPA1 SIPA1L2 SIPA1L3 ID1 THBS1 RALGDS RALA RALB RAC1 RAC2 RAC3 APBB1IP TLN1 TLN2 ITGA2B ITGB3 PFN3 PFN1 PFN2 PFN4 VASP ACTB ACTG1 ARAP3 RHOA SRC FARP2 CDC42 TIAM1 PARD3 PARD6A PARD6G PARD6B PRKCZ PRKCI MLLT4 CTNND1 KRIT1 RGS14 RASSF5 ITGAL ITGAM ITGB2 ITGB1 BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 MAP2K3 MAP2K6 MAPK11 MAPK12 MAPK13 MAPK14 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 PLCE1 HRAS KRAS NRAS RRAS VAV2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04020	Calcium signaling pathway	SLC8A1 SLC8A2 SLC8A3 ATP2B1 ATP2B3 ATP2B4 ATP2B2 CHRM1 CHRM3 CHRM5 ADORA2A ADORA2B ADRB1 ADRB2 ADRB3 DRD1 DRD5 HRH2 HTR4 HTR5A HTR6 HTR7 GNAS GNAL ADCY1 ADCY2 ADCY3 ADCY4 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG PLN ATP2A1 ATP2A3 ATP2A2 STIM1 STIM2 ORAI1 ORAI2 ORAI3 CACNA1C CACNA1D CACNA1F CACNA1S CACNA1A CACNA1B CACNA1E CACNA1G CACNA1H CACNA1I CHRNA7 P2RX1 P2RX2 P2RX3 P2RX4 P2RX5 P2RX7 P2RX6 GRIN1 GRIN2A GRIN2C GRIN2D RYR1 RYR2 RYR3 CYSLTR1 CYSLTR2 CHRM2 ADRA1A ADRA1B ADRA1D AGTR1 EDNRA EDNRB F2R GRM1 GRM5 GRPR HRH1 HTR2A HTR2B HTR2C LHCGR NTSR1 OXTR AVPR1A AVPR1B LTB4R2 PTAFR PTGER1 PTGER3 PTGFR BDKRB1 BDKRB2 TACR1 TACR2 TACR3 TBXA2R TRHR CCKAR CCKBR EGFR ERBB2 ERBB3 ERBB4 PDGFRA PDGFRB GNAQ GNA11 GNA14 GNA15 PLCD1 PLCD3 PLCD4 PLCB1 PLCB2 PLCB3 PLCB4 PLCG1 PLCG2 PLCE1 PLCZ1 ITPR1 ITPR2 ITPR3 CD38 SPHK1 SPHK2 VDAC1 VDAC2 VDAC3 SLC25A4 SLC25A5 SLC25A6 SLC25A31 PPIF TNNC1 TNNC2 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 PHKG1 PHKG2 PHKB PHKA2 PHKA1 MYLK MYLK2 MYLK3 MYLK4 CAMK2A CAMK2D CAMK2B CAMK2G CAMK4 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 NOS1 NOS2 NOS3 PDE1A PDE1B PDE1C PTK2B ITPKB ITPKA ITPKC PRKCA PRKCB PRKCG

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04022	cGMP-PKG signaling pathway	AGTR1 EDNRA EDNRB ADRA1A ADRA1B ADRA1D ADRA2A ADRA2B ADRA2C GNAQ GNA11 GNA12 GNA13 TRPC6 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 MEF2A MEF2BNB-MEF2B MEF2B MEF2C MEF2D NFATC1 NFATC2 NFATC3 NFATC4 SRF GATA4 NPPB NPR1 NPR2 PRKG1 PRKG2 CACNA1C CACNA1D CACNA1F CACNA1S KCNMA1 KCNU1 KCNMB1 KCNMB2 KCNMB3 KCNMB4 SLC8A1 SLC8A2 SLC8A3 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 ATP2B1 ATP2B3 ATP2B4 ATP2B2 RGS2 GTF2IRD1 GTF2I PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 MRVI1 PLN ATP2A1 ATP2A3 ATP2A2 RHOA ROCK1 ROCK2 PPP1R12A PPP1CA PPP1CB PPP1CC CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 MYLK MYLK2 MYLK3 MYLK4 MYL9 CNGA1 CNGB1 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PDE2A PDE3A PDE3B PDE5A GUCY1A2 GUCY1A3 GUCY1B3 BDKRB2 OPRD1 ADORA1 ADORA3 GNAI1 GNAI3 GNAI2 INS INSR IRS1 IRS2 IRS4 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 NOS3 ADRB1 ADRB2 ADRB3 PRKCE KCNJ8 VDAC1 VDAC2 VDAC3 SLC25A4 SLC25A5 SLC25A6 SLC25A31 PPIF CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B BAD VASP RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 MYH7

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04024	cAMP signaling pathway	FSHB ADRB1 DRD1 DRD5 ADORA2A HTR4 HTR6 PTGER2 ADCYAP1R1 VIPR2 TSHR MC2R GLP1R GIPR GPR119 FSHR NPR1 GNAS NPY GHRL ADRB2 HTR1A HTR1B HTR1D HTR1E HTR1F CHRM1 CHRM2 DRD2 GABBR1 GABBR2 ADORA1 EDNRA NPY1R SSTR1 SSTR2 SSTR5 HCAR1 HCAR2 HCAR3 FFAR2 SUCNR1 PTGER3 OXTR GHSR GNAI1 GNAI3 GNAI2 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 ADCY10 HCN2 HCN4 CNGA1 CNGA2 CNGA3 CNGA4 CNGB1 CNGB3 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G CAMK4 ABCC4 RAPGEF3 RAPGEF4 RRAS RRAS2 PLD1 PLD2 PLCE1 MAPK8 MAPK10 MAPK9 RAP1A RAP1B TIAM1 VAV3 VAV1 VAV2 RAC1 RAC2 RAC3 PAK1 ARAP3 RHOA MLLT4 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 PRKACA PRKACB PRKACG PPP1R1B PPP1CA PPP1CB PPP1CC CREB1 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 CREBBP EP300 BDNF FOS JUN GLI3 GLI1 PTCH1 HHIP NFKBIA NFKB1 RELA SOX9 AMH PPARA ACOX3 ACOX1 NFATC1 F2R BAD LIPE ROCK1 ROCK2 PPP1R12A MYL9 TNNI3 PLN RYR2 GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D GRIN3A GRIN3B GRIA1 GRIA2 GRIA3 GRIA4 CFTR ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 FXYD1 SLC9A1 ORAI1 ATP2B1 ATP2B3 ATP2B4 ATP2B2 CACNA1C CACNA1D CACNA1F CACNA1S PDE3A PDE3B PDE4A PDE4B PDE4C PDE4D ATP2A2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04060	Cytokine-cytokine receptor interaction	CXCL1 CXCL2 CXCL3 CXCL5 CXCL6 PPBP CXCL8 CXCL9 CXCL10 CXCL11 CXCL12 CXCL13 CXCL16 PF4 PF4V1 CXCL14 XCL1 XCL2 CX3CL1 CCL1 CCL14 CCL16 CCL17 CCL18 CCL22 CCL24 CCL26 CCL20 CCL25 CCL19 CCL21 CCL2 CCL4 CCL4L2 CCL4L1 CCL3 CCL3L1 CCL3L3 CCL13 CCL7 CCL5 CCL15 CCL23 CCL8 CCL11 CCL28 CCL27 IL6 IL11 OSM LIF CNTF CLCF1 CTF1 CSF3 LEP IL4 IL13 IL12A IL12B IL23A CSF2 IL3 IL5 IL2 IL7 IL9 IL15 IL21 TSLP EPO GH1 GH2 PRL THPO PDGFC PDGFA PDGFB VEGFA VEGFB VEGFC FIGF HGF EGF CSF1 KITLG FLT3LG IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 IFNW1 IFNK IFNE IFNG IL10 IL19 IL20 IL24 IL22 IFNL2 IFNL3 IFNL1 IL26 TNFSF15 TNFSF10 TNFSF11 TNFSF12 TNF LTA LTB TNFSF14 FASLG CD40LG CD70 TNFSF8 TNFSF9 TNFSF4 TNFSF18 TNFSF13 TNFSF13B EDA TGFB1 TGFB2 TGFB3 INHBA INHBB INHBC INHBE AMH BMP2 BMP7 GDF5 IL17A IL17B IL25 IL1A IL1B IL18 CXCR2 CXCR1 CXCR3 CXCR4 CXCR5 CXCR6 ACKR3 XCR1 CX3CR1 CCR8 CCR6 CCR9 CCR4 CCR7 CCR2 CCR5 CCR1 CCR3 CCR10 IL6R IL6ST IL11RA LIFR OSMR CNTFR CSF3R LEPR IL4R IL13RA1 IL12RB1 IL12RB2 IL23R CSF2RA CSF2RB IL3RA IL5RA IL2RA IL2RB IL2RG IL7R IL9R IL15RA IL21R CRLF2 EPOR GHR PRLR MPL PDGFRA PDGFRB FLT1 KDR FLT4 MET EGFR CSF1R KIT FLT3 IFNAR1 IFNAR2 IFNGR1 IFNGR2 IL10RA IL10RB IL20RA IL20RB IL22RA1 IL22RA2 IFNLR1 TNFRSF10A TNFRSF10B TNFRSF10C TNFRSF10D TNFRSF11B TNFRSF11A TNFRSF25 TNFRSF12A TNFRSF21 NGFR TNFRSF1B TNFRSF1A LTBR TNFRSF14 TNFRSF6B FAS CD40 CD27 TNFRSF8 TNFRSF9 TNFRSF4 TNFRSF18 TNFRSF17 TNFRSF13B TNFRSF13C EDAR EDA2R TNFRSF19 RELT TGFBR2 TGFBR1 ACVR2A ACVR2B ACVR1 ACVR1B AMHR2 BMPR2 BMPR1A BMPR1B IL17RA IL17RB IL1R1 IL1RAP IL1R2 IL18R1 IL18RAP PLEKHO2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04062	Chemokine signaling pathway	CXCL1 CXCL2 CXCL3 CXCL5 CXCL6 PPBP CXCL8 CXCL9 CXCL10 CXCL11 CXCL12 CXCL13 CXCL16 PF4 PF4V1 CXCL14 XCL1 XCL2 CX3CL1 CCL2 CCL3 CCL3L1 CCL3L3 CCL4 CCL4L2 CCL4L1 CCL5 CCL7 CCL8 CCL11 CCL13 CCL15 CCL23 CCL19 CCL20 CCL21 CCL25 CCL27 CCL28 CCL1 CCL14 CCL16 CCL17 CCL18 CCL22 CCL24 CCL26 CXCR2 CXCR1 CXCR3 CXCR4 CXCR5 CXCR6 XCR1 CX3CR1 CCR8 CCR6 CCR9 CCR4 CCR7 CCR2 CCR5 CCR1 CCR3 CCR10 JAK2 JAK3 STAT1 STAT2 STAT3 STAT5B GNAI1 GNAI3 GNAI2 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG LYN HCK FGR SRC SHC1 SHC2 SHC3 SHC4 GRB2 SOS1 SOS2 HRAS KRAS NRAS RAF1 BRAF MAP2K1 MAPK1 MAPK3 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG PRKCZ AKT1 AKT2 AKT3 FOXO3 CHUK IKBKB IKBKG NFKBIA NFKBIB NFKB1 RELA GSK3A GSK3B ITK VAV3 VAV1 VAV2 RAC1 RAC2 PAK1 CDC42 WAS WASL RHOA ROCK1 ROCK2 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 PREX1 ELMO1 DOCK2 PTK2 PXN BCAR1 CRK CRKL PTK2B PLCB1 PLCB2 PLCB3 PLCB4 RASGRP2 RAP1A RAP1B PARD3 TIAM1 PRKCB PRKCD NCF1 GRK7 GRK1 ADRBK1 ADRBK2 GRK4 GRK5 GRK6 ARRB1 ARRB2
path:hsa04064	NF-kappa B signaling pathway	LCK ZAP70 LAT PLCG1 PRKCQ SYK LYN BLNK BTK PLCG2 PRKCB CARD11 BCL10 MALT1 IL1B IL1R1 MYD88 IRAK1 IRAK4 TRAF6 TNF TNFRSF1A RIPK1 TRADD TRAF2 TRAF5 BIRC2 BIRC3 DDX58 TRIM25 LBP CD14 TLR4 LY96 TIRAP TICAM2 TICAM1 CD40LG CD40 TRAF3 TNFSF11 TNFRSF11A LTA LTB TNFSF14 LTBR MAP3K14 MAP3K7 TAB1 TAB2 TAB3 TNFSF13B TNFRSF13C IKBKG CHUK IKBKB PIAS4 UBE2I ATM PIDD1 ERC1 NFKBIA NFKB1 RELA CFLAR XIAP BCL2L1 BCL2 TRAF1 BCL2A1 NFKB2 CXCL8 TNFAIP3 PTGS2 CCL4 CCL4L2 CCL4L1 VCAM1 PLAU CSNK2A1 CSNK2A2 CSNK2B RELB CCL13 CCL19 CCL21 CXCL12 ICAM1 PARP1 CXCL2 GADD45B

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04066	HIF-1 signaling pathway	IL6 IL6R STAT3 TLR4 IFNG IFNGR1 IFNGR2 RELA NFKB1 INS EGF IGF1 INSR EGFR IGF1R ERBB2 MAP2K1 MAP2K2 MAPK1 MAPK3 MKNK1 MKNK2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 MTOR EIF4EBP1 EIF4E EIF4E2 EIF4E1B RPS6KB1 RPS6KB2 RPS6 HIF1A VHL RBX1 TCEB1 TCEB2 CUL2 EGLN1 EGLN3 EGLN2 ARNT CREBBP EP300 NOX3 NOX1 CYBB PLCG1 PLCG2 PRKCA PRKCB PRKCG CAMK2A CAMK2D CAMK2B CAMK2G TIMP1 LTBR EPO TF TFRC VEGFA FLT1 SERPINE1 ANGPT1 ANGPT2 ANGPT4 TEK EDN1 NOS2 NOS3 NPPA SLC2A1 PDK1 HK3 HK1 HK2 HKDC1 GAPDH ENO3 ENO2 ENO1 PFKFB3 BCL2 CDKN1A CDKN1B PDHA2 PDHA1 PDHB ALDOA HMOX1 LDHA PFKL PGK1
path:hsa04068	FoxO signaling pathway	TGFB1 TGFB2 TGFB3 TGFBR1 TGFBR2 SMAD4 SMAD2 SMAD3 STK11 PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 NLK SKP2 MDM2 SETD7 CREBBP EP300 SIRT1 USP7 SLC2A4 MAPK8 MAPK10 MAPK9 STK4 IGF1 IGF1R INS INSR IRS1 IRS2 IRS4 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PDPK1 CHUK IKBKB PTEN AKT1 AKT2 AKT3 SGK1 SGK2 SGK3 C8orf44-SGK3 FOXO1 FOXO3 FOXO4 FOXO6 GRM1 HOMER1 HOMER2 HOMER3 AGAP2 CSNK1E LOC400927-CSNK1E IL6 IL10 STAT3 EGF EGFR GRB2 SOS1 SOS2 HRAS KRAS NRAS BRAF RAF1 ARAF MAP2K1 MAP2K2 MAPK1 MAPK3 MAPK11 MAPK12 MAPK13 MAPK14 CDK2 PRMT1 FOXG1 CCNB1 CCNB2 CCNB3 CCND1 CCND2 CCNG2 CDKN2B CDKN2D CDKN1A CDKN1B RBL2 PLK1 PLK2 PLK3 PLK4 GADD45A GADD45B GADD45G FASLG BCL2L11 TNFSF10 BCL6 BNIP3 ATG12 GABARAP GABARAPL1 GABARAPL2 CAT SOD2 ATM PCK1 PCK2 G6PC G6PC2 G6PC3 IL7R KLF2 S1PR1 S1PR4 RAG1 RAG2 FBXO25 FBXO32

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04070	Phosphatidylinositol signaling system	PI4KA PI4KB PI4K2A PI4K2B PIP5K1C PIP5K1A PIP5K1B SYNJ1 SYNJ2 OCRL INPP5E INPP5B PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PTEN INPP5D INPPL1 INPP4A INPP4B PIKFYVE MTM1 MTMR1 MTMR2 MTMR3 MTMR4 MTMR8 MTMR6 MTMR7 MTMR14 PIP4K2C PIP4K2A PIP4K2B TMEM55B TMEM55A PIK3C3 PIK3C2G PIK3C2A PIK3C2B PLCB1 PLCB2 PLCB3 PLCB4 PLCD1 PLCD3 PLCD4 PLCE1 PLCG1 PLCG2 PLCZ1 IPMK ITPKB ITPKA ITPKC INPP5K INPP5J INPP5A ITPK1 IPPK IP6K1 IP6K2 IP6K3 PPIP5K1 PPIP5K2 INPP1 IMPA2 IMPA1 IMPAD1 DGKZ DGKD DGKI DGKA DGKE DGKB DGKH DGKG DGKQ DGKK CDS1 CDS2 CDIPT ITPR1 ITPR2 ITPR3 PRKCA PRKCB PRKCG CALML3 CALM2 CALM3 CALM1 CALML6 CALML5
path:hsa04071	Sphingolipid signaling pathway	SPTLC1 SPTLC2 SPTLC3 CERS2 CERS3 CERS6 CERS1 CERS4 CERS5 DEGS1 DEGS2 SMPD1 SGMS1 SGMS2 TNF TNFRSF1A NSMAF SMPD2 ASAH1 ASAH2 ACER2 ACER1 TRADD TRAF2 CTSD BID BAX MAP3K5 MAPK8 MAPK10 MAPK9 MAPK11 MAPK12 MAPK13 MAPK14 PRKCZ AKT1 AKT2 AKT3 BCL2 PPP2CA PPP2CB PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D PPP2R3B PPP2R3C PPP2R3A PPP2R5B PPP2R5C PPP2R5D PPP2R5E PPP2R5A TP53 FCER1A MS4A2 FCER1G FYN GAB2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PLD1 PLD2 SPHK1 SPHK2 ADORA1 ADORA3 BDKRB2 OPRD1 GNAI1 GNAI3 GNAI2 PLCB1 PLCB2 PLCB3 PLCB4 PRKCE MAPK1 MAPK3 SGPP1 SGPP2 ABCC1 SGPL1 S1PR1 S1PR2 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 PDPK1 NOS3 RAC1 RAC2 RAC3 S1PR3 GNAQ PRKCA PRKCB PRKCG S1PR4 S1PR5 GNA12 GNA13 RHOA ROCK1 ROCK2 PTEN NFKB1 RELA

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04072	Phospholipase D signaling pathway	EGF PDGFA PDGFB PDGFC PDGFD KITLG INS EGFR PDGFRA PDGFRB KIT INSR GRB2 GAB1 GAB2 SHC1 SHC2 SHC3 SHC4 PTPN11 SOS1 SOS2 HRAS KRAS NRAS MRAS RRAS RRAS2 RALGDS RALA RALB PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 TSC1 TSC2 RHEB MTOR RHOA ARF1 ARF6 PLCG1 PLCG2 FCER1A MS4A2 FCER1G FYN SYK PRKCA PLD1 PLD2 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F AGPAT1 AGPAT2 AGPAT3 AGPAT4 AGPAT5 RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 PIP5K1C PIP5K1A PIP5K1B DGKZ DGKD DGKI DGKA DGKE DGKB DGKH DGKG DGKQ DGKK PPAP2A PPAP2B PPAP2C AVP CXCL8 AGTR1 F2R GRM1 GRM2 GRM3 GRM4 GRM5 GRM6 GRM7 GRM8 AVPR1A AVPR1B AVPR2 LPAR1 LPAR2 LPAR3 LPAR4 LPAR5 LPAR6 PTGFR CXCR1 CXCR2 PLCB1 PLCB2 PLCB3 PLCB4 GNA12 PTK2B GNA13 CYTH3 CYTH4 CYTH2 CYTH1 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 RAPGEF3 RAPGEF4 DNM1 DNM2 DNM3 SPHK1

Continued on next page

continued from previous page

Pathway ID	Description		Gene Sets
path:hsa04080	Neuroactive interaction	ligand-receptor	CHRM1 CHRM2 CHRM3 CHRM4 CHRM5 ADRA1A ADRA1B ADRA1D ADRA2A ADRA2B ADRA2C ADRB1 ADRB2 ADRB3 DRD1 DRD2 DRD3 DRD4 DRD5 HRH1 HRH2 HRH3 HRH4 HTR1A HTR1B HTR1D HTR1E HTR1F HTR2A HTR2B HTR2C HTR4 HTR5A HTR6 HTR7 TAAR9 TAAR6 TAAR1 TAAR8 TAAR5 TAAR2 AGTR1 AGTR2 NMBR GRPR BRS3 BDKRB1 BDKRB2 C5AR1 C3AR1 FPR1 FPR3 FPR2 APLNR CCKAR CCKBR EDNRA EDNRB MC1R MC2R MC3R MC4R MC5R NPY1R NPY2R NPY4R LOC100996758 NPY5R GPR83 NTSR1 NTSR2 OPRD1 OPRK1 OPRM1 OPRL1 SSTR1 SSTR2 SSTR3 SSTR4 SSTR5 TACR1 TACR2 TACR3 AVPR1A AVPR1B AVPR2 OXTR GALR1 GALR2 GALR3 F2 CTSG GZMA PLG PRSS3 PRSS2 PRSS1 F2R F2RL1 F2RL2 F2RL3 PARD3 HCRTR1 HCRTR2 NPFFR1 NPFFR2 UTS2R NMUR1 NMUR2 MCHR1 MCHR2 FSHB FSHR TSHB LHCGR TSHR PTGDR PTGER1 PTGER2 PTGER3 PTGER4 PTGFR PTGIR TBXA2R ADORA1 ADORA2A ADORA2B ADORA3 P2RY2 P2RY1 P2RY4 P2RY6 LPAR6 P2RY10 P2RY14 P2RY8 P2RY11 P2RY13 LPAR4 GPR35 CNR1 CNR2 PTAFR GNRHR TRHR GHSR KISS1R MTNR1A MTNR1B GPR50 S1PR1 LPAR1 S1PR3 LPAR2 S1PR2 S1PR4 LPAR3 S1PR5 LTB4R LTB4R2 MAS1 RXFP1 RXFP2 RXFP3 RXFP4 CYSLTR1 CYSLTR2 CALCR CALCRL CRHR1 CRHR2 GIPR GLP1R GLP2R GCGR GHRHR PTH1R PTH2R ADCYAP1R1 SCTR VIPR1 VIPR2 GRM1 GRM2 GRM3 GRM4 GRM5 GRM6 GRM7 GRM8 GABBR1 GABBR2 GPR156 MLNR NPBWR1 NPBWR2 PRLHR GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D GRIN3A GRIN3B GABRA1 GABRA2 GABRA3 GABRA4 GABRA5 GABRA6 GABRB1 GABRB3 GABRB2 GABRD GABRE GABRG1 GABRG2 GABRG3 GABRP GABRR1 GABRR2 GABRR3 GABRQ CHRNA1 CHRNA2 CHRNA3 CHRNA4 CHRNA5 CHRNA6 CHRNA7 CHRNA9 CHRNA10 CHRNB1 CHRNB2 CHRNB3 CHRNB4 CHRND CHRNE CHRNG P2RX1 P2RX2 P2RX3 P2RX4 P2RX5 P2RX7 P2RX6 CGA LHB GRIA1 GRIA2 GRIA3 GRIA4 GRIK1 GRIK2 GRIK3 GRIK4 GRIK5 GRID1 GRID2 GLRA1 GLRA2 GLRA3 GLRB TRPV1 TSPO NR3C1 GH1 GH2 CSH1 CSH2 GHR THRA THRB LEP LEPR PRL PRLR

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04110	Cell cycle	CCND1 CCND2 CCND3 CDK4 CDK6 RB1 RBL1 RBL2 ABL1 HDAC1 HDAC2 E2F1 E2F2 E2F3 E2F4 E2F5 TFDP1 TFDP2 GSK3B TGFB1 TGFB2 TGFB3 SMAD2 SMAD3 SMAD4 MYC ZBTB17 CDKN2A CDKN2B CDKN2C CDKN2D CDKN1B CDKN1C CDKN1A CCNE1 CCNE2 CDK2 SKP1 CUL1 RBX1 SKP2 CCNA2 CCNA1 CDC6 CDC45 CDC7 DBF4 CDK1 CCNB1 CCNB2 CCNB3 CDC25B CDC25C YWHAZ YWHAB YWHAQ YWHAE YWHAH YWHAG PLK1 WEE1 WEE2 PKMYT1 CCNH CDK7 ANAPC1 ANAPC2 CDC27 ANAPC4 ANAPC5 CDC16 ANAPC7 CDC23 ANAPC10 ANAPC11 CDC26 ANAPC13 CDC20 PTTG1 PTTG2 ESPL1 SMC1A SMC1B SMC3 STAG2 STAG1 RAD21 TTK BUB1 BUB3 BUB1B MAD1L1 MAD2L1 MAD2L2 FZR1 CDC14B CDC14A ATR ATM TP53 CHEK1 CHEK2 CREBBP EP300 PRKDC MDM2 GADD45A GADD45B GADD45G PCNA SFN CDC25A ORC1 ORC2 ORC3 ORC4 ORC5 ORC6 MCM2 MCM3 MCM4 MCM5 MCM6 MCM7
path:hsa04114	Oocyte meiosis	INS IGF1 IGF1R PGR AR ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG AURKA CPEB1 CPEB2 CPEB3 CPEB4 MOS MAP2K1 MAPK1 MAPK3 RPS6KA3 RPS6KA1 RPS6KA2 RPS6KA6 PKMYT1 YWHAZ YWHAB YWHAQ YWHAE YWHAH YWHAG SPDYA SPDYC CDK1 CDK2 PPP1CA PPP1CB PPP1CC SLK PLK1 CDC25C RBX1 CUL1 SKP1 BTRC FBXW11 FBXO5 ANAPC1 ANAPC2 CDC27 ANAPC4 ANAPC5 CDC16 ANAPC7 CDC23 ANAPC10 ANAPC11 CDC26 ANAPC13 CDC20 FBXO43 PTTG1 PTTG2 ESPL1 REC8 STAG3 SMC1A SMC1B SMC3 BUB1 MAD2L1 MAD2L2 CCNE1 CCNE2 SGOL1 PPP2R5B PPP2R5C PPP2R5D PPP2R5E PPP2R5A PPP2R1B PPP2R1A PPP2CA PPP2CB PLCZ1 ITPR1 ITPR2 ITPR3 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 CAMK2A CAMK2D CAMK2B CAMK2G MAPK12 CCNB1 CCNB2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04115	p53 signaling pathway	ATM CHEK2 ATR CHEK1 CDKN2A MDM2 MDM4 TP53 CDKN1A CCND1 CCND2 CCND3 CDK4 CDK6 CCNE1 CCNE2 CDK2 SFN RPRM CCNB1 CCNB2 CCNB3 CDK1 GADD45A GADD45B GADD45G GTSE1 FAS PIDD1 CASP8 BID BAX PMAIP1 BBC3 TP53AIP1 TP53I3 EI24 SHISA5 PERP ZMAT3 SIAH1 CYCS APAF1 CASP9 CASP3 IGFBP3 IGF1 SERPINE1 ADGRB1 CD82 THBS1 SERPINB5 DDB2 RRM2B RRM2 SESN1 SESN2 SESN3 PTEN TSC2 STEAP3 RFWD2 RCHY1 CCNG1 CCNG2 PPM1D TP73 TNFRSF10B
path:hsa04120	Ubiquitin mediated proteolysis	UBA1 SAE1 UBA2 UBA3 UBA7 UBA6 UBE2A UBE2B UBE2C UBE2E3 UBE2D4 UBE2D1 UBE2D2 UBE2D3 UBE2E1 UBE2E2 UBE2F UBE2G1 UBE2G2 UBE2H UBE2I UBE2J1 UBE2J2 UBE2L3 UBE2L6 UBE2M UBE2NL UBE2N UBE2O UBE2Q1 UBE2Q2 UBE2QL1 UBE2R2 CDC34 UBE2S UBE2U UBE2W UBE2Z UBE2K BIRC6 UBE3A UBE3B UBE3C SMURF1 SMURF2 ITCH WWP1 WWP2 TRIP12 NEDD4 NEDD4L HUWE1 UBR5 HERC1 HERC2 HERC3 HERC4 UBE4A UBE4B STUB1 PPIL2 PRPF19 UBOX5 MDM2 CBLC CBL CBLB PARK2 SIAH1 PML TRAF6 MAP3K1 RFWD2 RCHY1 BIRC2 BIRC3 XIAP BIRC7 BIRC8 PIAS1 PIAS2 PIAS3 PIAS4 SYVN1 NHLRC1 AIRE MGRN1 BRCA1 FANCL MID1 TRIM32 TRIM37 RBX1 CUL1 SKP1 BTRC FBXW11 SKP2 FBXW7 FBXO2 FBXO4 CUL2 TCEB1 TCEB2 VHL CUL3 KEAP1 KLHL9 KLHL13 RHOBTB2 RHOBTB1 CUL4B CUL4A DDB1 DDB2 ERCC8 DET1 RNF7 CUL5 SOCS1 SOCS3 CUL7 FBXW8 ANAPC11 ANAPC2 CDC20 FZR1 ANAPC1 CDC27 ANAPC4 ANAPC5 CDC16 ANAPC7 CDC23 ANAPC10 CDC26 ANAPC13
path:hsa04122	Sulfur relay system	NFS1 MPST TST URM1 MOCS3 CTU1 CTU2 TRMU MOCS2 MOCS1
path:hsa04130	SNARE interactions in vesicular transport	STX1A STX2 STX3 STX1B STX4 STX19 STX11 STX7 STX16 STX5 STX17 STX18 VTI1B VTI1A GOSR1 GOSR2 BNIP1 STX6 STX10 STX8 BET1L BET1 USE1 SNAP23 SNAP29 VAMP1 VAMP2 VAMP3 VAMP8 VAMP4 VAMP5 VAMP7 YKT6 SEC22B

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04140	Regulation of autophagy	INS PRKAA1 PRKAA2 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNG ULK3 ULK1 ULK2 ATG13 PIK3R4 BECN1 ATG14 PIK3C3 ATG12 ATG7 ATG10 ATG5 ATG16L1 ATG16L2 GABARAP GABARAPL1 GABARAPL2 ATG4A ATG4B ATG4C ATG4D ATG3
path:hsa04141	Protein processing in endoplasmic reticulum	SEC61A1 SEC61A2 SEC61B SEC61G SEC62 SEC63 RPN1 RPN2 DAD1 TUSC3 DDOST STT3A STT3B MOGS CKAP4 RRBP1 SIL1 HYOU1 HSPA5 DNAJB11 DNAJC1 DNAJC3 DNAJC10 HSP90B1 GANAB PRKCSH CANX PDIA3 CALR MAN1B1 MAN1A2 MAN1C1 MAN1A1 LMAN2 LMAN1 LMAN1L PREB SAR1A SAR1B SEC13 SEC31A SEC31B SEC23B SEC23A SEC24B SEC24A SEC24C SEC24D UGGT2 UGGT1 EDEM1 EDEM2 EDEM3 P4HB PDIA4 PDIA6 ERP29 TXNDC5 ERO1L ERO1LB OS9 ERLEC1 SSR1 SSR2 SSR3 SSR4 BCAP31 TRAM1 DERL1 DERL2 DERL3 UBXN6 NSFL1C SVIP VCP NPLOC4 UFD1L HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 DNAJA1 DNAJA2 DNAJB1 DNAJB2 DNAJB12 DNAJC5 DNAJC5B DNAJC5G HSP90AA1 HSP90AB1 HSPH1 HSPA4L BAG1 BAG2 HSPBP1 CRYAA LOC102724652 CRYAB YOD1 PLAA RAD23B RAD23A UBQLN1 UBQLN2 UBQLN3 UBQLN4 UBQLNL NGLY1 ATXN3 ATXN3L UBE4B EIF2AK1 EIF2AK2 EIF2AK3 EIF2AK4 EIF2S1 NFE2L2 ATF4 PPP1R15A DDIT3 BCL2 ATF6 ATF6B WFS1 MBTPS1 MBTPS2 XBP1 ERN1 TRAF2 MAP3K5 MAP2K7 MAPK8 MAPK10 MAPK9 BAX BAK1 CAPN1 CAPN2 CASP12 MARCH6 UBE2J1 UBE2J2 UBE2G1 UBE2G2 SYVN1 RNF5 RNF185 VIMP SEL1L SEL1L2 HERPUD1 AMFR STUB1 UBE2E3 UBE2D4 UBE2D1 UBE2D2 UBE2D3 UBE2E1 UBE2E2 PARK2 RBX1 CUL1 SKP1 FBXO2 FBXO6

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04142	Lysosome	TCIRG1 ATP6V0A2 ATP6V0A4 ATP6V0A1 ATP6V0D1 ATP6V0D2 ATP6V1H ATP6AP1 ATP6V0C ATP6V0B CTSA CTSB CTSC CTSD CTSE CTSF CTSG CTSH CTSK CTSL CTSO CTSS CTSV CTSW CTSZ NAPSA LGMN TPP1 GLA GLB1 GAA GBA IDUA NAGA NAGLU GALC GUSB HEXA HEXB MANBA MAN2B1 NEU1 ARSA ARSB ARSG GALNS GNS IDS SGSH LIPA PLA2G15 DNASE2 DNASE2B ACP2 ACP5 SMPD1 ASAH1 AGA PSAP PSAPL1 GM2A PPT1 PPT2 LAMP1 LAMP2 LAMP3 CD68 CD63 SCARB2 NPC1 NPC2 CTNS SLC17A5 SLC11A1 SLC11A2 LAPTM4B LAPTM5 LAPTM4A ABCA2 ABCB9 CD164 ENTPD4 SORT1 CLN3 CLN5 MFSD8 HGSNAT SUMF1 GNPTAB GNPTG NAGPA IGF2R M6PR CLTA CLTB CLTC CLTCL1 AP1G1 AP1G2 AP1B1 AP1M1 AP1M2 AP1S1 AP1S2 AP1S3 AP3D1 AP3B2 AP3B1 AP3M1 AP3M2 AP3S2 AP3S1 AP4E1 AP4B1 AP4M1 AP4S1 GGA2 GGA3 GGA1 MCOLN1 LITAF FUCA1 HYAL1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04144	Endocytosis	TGFB1 TGFB2 TGFB3 TGFBR1 TGFBR2 ZFYVE9 ZFYVE16 PML SMAD2 SMAD3 TFRC EGFR FLT1 KDR PDGFRA FGFR2 FGFR3 FGFR4 IGF1R MET NTRK1 ERBB3 ERBB4 KIT RET CSF1R EGF DNM1 DNM2 DNM3 ARF6 PIP5K1C PIP5K1A PIP5K1B PIP5KL1 PLD1 PLD2 CLTA CLTB CLTC CLTCL1 AP2A2 AP2A1 AP2B1 AP2M1 AP2S1 EHD2 CBLC CBL CBLB NEDD4 NEDD4L MDM2 TRAF6 RNF41 ITCH SMURF1 SMURF2 WWP1 SH3KBP1 SH3GL3 SH3GL2 SH3GL1 SH3GLB1 SH3GLB2 AMPH BIN1 EPN1 EPN3 EPN2 EPS15 EPS15L1 SPG20 LDLR LDLRAP1 DAB2 CCR5 CXCR1 CXCR2 CXCR4 F2R ADRB1 ADRB2 ADRB3 GRK7 GRK1 ADRBK1 ADRBK2 GRK4 GRK5 GRK6 ARRB1 ARRB2 WAS WASL ARPC1B ARPC1A ARPC2 ARPC3 ARPC4 ARPC5 ARPC5L WIPF2 WIPF1 HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 DNAJC6 KIAA1033 KIAA0196 FAM21C FAM21A WASH1 CCDC53 CAPZA1 CAPZA2 CAPZA3 CAPZB VPS29 VPS26B VPS26A VPS35 SNX12 SNX3 USP8 STAMBP RAB7A HGS STAM2 STAM TSG101 MVB12A MVB12B VPS28 VPS37D VPS37A VPS37B VPS37C SNF8 VPS36 VPS25 CHMP6 CHMP4C CHMP4B CHMP4A CHMP3 RNF103-CHMP3 CHMP2A CHMP2B CHMP7 VPS4B VPS4A VTA1 CHMP1B CHMP5 IST1 PDCD6IP SPG21 IGF2R SNX1 SNX2 SNX32 SNX5 SNX6 SRC HRAS IL2RA IL2RB IL2RG RHOA HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E FOLR1 FOLR2 FOLR3 IZUMO1R CAV1 CAV2 CAV3 EHD3 EHD4 EEA1 RAB5A RAB5B RAB5C RABEP1 RBSN RAB4A RUFY1 EHD1 VPS45 SNX4 RAB22A RAB31 RAB10 RAB8A RAB35 RAB11A RAB11B RAB11FIP2 RAB11FIP1 RAB11FIP5 ZFYVE27 KIF5A KIF5B KIF5C RAB11FIP4 RAB11FIP3 PARD3 PARD6A PARD6G PARD6B PRKCZ PRKCI CDC42 SMAP2 SMAP1 GIT1 GIT2 ASAP1 ASAP3 ASAP2 ACAP3 ACAP2 ACAP1 ARAP1 ARAP2 ARAP3 AGAP1 AGAP3 AGAP2 ARFGAP1 ARFGAP3 ARFGAP2 PSD3 PSD4 PSD PSD2 IQSEC2 IQSEC1 IQSEC3 CYTH3 CYTH4 CYTH2 CYTH1 ARFGEF2 ARFGEF1 GBF1 ARF1 ARF3 ARF5

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04145	Phagosome	VAMP3 STX12 STX7 ACTB ACTG1 CORO1A STX18 SEC22B HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 RAB5A RAB5B RAB5C EEA1 PIK3C3 TFRC HGS ATP6V1A ATP6V1B1 ATP6V1B2 ATP6V1C2 ATP6V1C1 ATP6V1D ATP6V1E2 ATP6V1E1 ATP6V1F ATP6V1G1 ATP6V1G3 ATP6V1G2 ATP6V0E1 ATP6V0E2 TCIRG1 ATP6V0A2 ATP6V0A4 ATP6V0A1 ATP6V0D1 ATP6V0D2 ATP6V1H ATP6V0C ATP6V0B ATP6AP1 RAB7A RAB7B RILP DYNC1H1 DYNC2H1 DYNC1I1 DYNC1I2 DYNC1LI2 DYNC1LI1 TUBA1B TUBA4A TUBA3C TUBA1A TUBA1C TUBA8 TUBA3E TUBA3D TUBAL3 TUBB6 TUBB TUBB1 TUBB2A TUBB3 TUBB4A TUBB8 TUBB2B TUBB4B LAMP1 LAMP2 PIKFYVE M6PR NOS1 MPO CTSL CTSS SEC61A1 SEC61A2 SEC61B SEC61G TAP1 TAP2 CALR CANX FCAR FCGR1A FCGR2A FCGR2B FCGR2C FCGR3A FCGR3B C1R ITGAM ITGB2 C3 COLEC11 COLEC12 MBL2 SFTPA1 SFTPA2 SFTPD ITGAV ITGA2 ITGA5 ITGB1 ITGB3 ITGB5 THBS1 COMP THBS2 THBS3 THBS4 TLR2 TLR6 TLR4 CD14 PLA2R1 MRC1 MRC2 CLEC4M CD209 CLEC7A MSR1 MARCO OLR1 SCARB1 CD36 CYBA NOX3 NOX1 CYBB RAC1 NCF1 NCF2 NCF4
path:hsa04146	Peroxisome	PEX16 PEX3 PEX19 ABCD3 PEX1 PEX6 PEX26 PEX7 PEX5 PEX5L PEX14 PEX13 PEX12 PEX10 PEX2 PXMP2 MPV17 MPV17L2 MPV17L PXMP4 PEX11A PEX11B PEX11G SLC25A17 HACL1 AMACR PHYH ACOX3 ACOX1 ACOX2 HSD17B4 SCP2 BAAT EHHADH ACAA1 DECR2 ECH1 ABCD1 ABCD2 ABCD4 SLC27A2 ACSL6 ACSL4 ACSL1 ACSL5 ACSL3 PECR ECI2 NUDT7 NUDT12 NUDT19 ACOT8 CRAT CROT MLYCD GNPAT AGPS FAR2 FAR1 MVK PMVK AGXT DAO DDO IDH1 IDH2 PAOX PIPOX HMGCL HMGCLL1 HAO2 HAO1 CAT PRDX5 SOD1 SOD2 NOS2 PRDX1 EPHX2 GSTK1 XDH DHRS4 DHRS4L1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04150	mTOR signaling pathway	INS IGF1 IRS1 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG PTEN PDPK1 AKT1 AKT2 AKT3 AKT1S1 TSC1 TSC2 RHEB MLST8 RRAGA RRAGB RRAGC RRAGD MTOR RPTOR RICTOR HIF1A VEGFA RPS6KB1 RPS6KB2 EIF4B RPS6 EIF4EBP1 EIF4E EIF4E2 EIF4E1B ULK3 ULK1 ULK2 MAPK1 MAPK3 RPS6KA3 RPS6KA1 RPS6KA2 RPS6KA6 DDIT4 TNF IKBKB STK11 STRADA CAB39 CAB39L PRKAA1 PRKAA2 BRAF PRKCA PRKCB PRKCG

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04151	PI3K-Akt signaling pathway	EGF FGF1 FGF2 FGF3 FGF4 FGF17 FGF6 FGF7 FGF8 FGF9 FGF10 FGF11 FGF12 FGF13 FGF14 FGF16 FGF5 FGF18 FGF19 FGF20 FGF21 FGF22 FGF23 NGF INS IGF1 PDGFA PDGFB PDGFC PDGFD CSF1 KITLG VEGFA VEGFB PGF VEGFC FIGF HGF ANGPT1 ANGPT2 ANGPT4 EFNA1 EFNA2 EFNA3 EFNA4 EFNA5 EGFR FGFR1 FGFR2 FGFR3 FGFR4 NGFR INSR IGF1R PDGFRA PDGFRB CSF1R KIT FLT1 FLT4 KDR MET TEK EPHA2 GRB2 SOS1 SOS2 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 IRS1 TLR2 TLR4 RAC1 SYK CD19 PIK3AP1 GH1 GH2 CSH1 CSH2 PRL OSM IL2 IL3 IL6 IL4 IL7 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 EPO CSF3 GHR PRLR OSMR IL2RA IL2RB IL2RG IL3RA IL6R IL4R IL7R IFNAR1 IFNAR2 EPOR CSF3R JAK1 JAK2 JAK3 COL1A1 COL1A2 COL2A1 COL4A2 COL4A4 COL4A6 COL4A1 COL4A5 COL4A3 COL6A1 COL6A2 COL6A3 COL6A6 COL6A5 COL9A1 COL9A2 COL9A3 LAMA1 LAMA2 LAMA3 LAMA5 LAMA4 LAMB1 LAMB2 LAMB3 LAMB4 LAMC1 LAMC2 LAMC3 CHAD RELN THBS1 COMP THBS2 THBS3 THBS4 FN1 SPP1 VTN TNC TNN TNR TNXB VWF IBSP ITGA1 ITGA2 ITGA2B ITGA3 ITGA4 ITGA5 ITGA6 ITGA7 ITGA8 ITGA9 ITGA10 ITGA11 ITGAV ITGB1 ITGB3 ITGB4 ITGB5 ITGB6 ITGB7 ITGB8 PTK2 F2R CHRM1 CHRM2 LPAR1 LPAR2 LPAR3 LPAR4 LPAR5 LPAR6 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 PDPK1 STK11 PRKAA1 PRKAA2 DDIT4 TSC1 TSC2 RHEB MLST8 MTOR RPTOR EIF4EBP1 EIF4E EIF4E2 EIF4E1B RPS6KB1 RPS6KB2 EIF4B RPS6 PRKCA PKN1 PKN3 PKN2 SGK1 SGK2 SGK3 C8orf44-SGK3 AKT1 AKT2 AKT3 PTEN THEM4 PPP2CA PPP2CB PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D PPP2R3B PPP2R3C PPP2R3A PPP2R5B PPP2R5C PPP2R5D PPP2R5E PPP2R5A HSP90AA1 HSP90AB1 HSP90B1 CDC37 CRTC2 PHLPP1 PHLPP2 TCL1A TCL1B MTCP1 NOS3 BRCA1 GSK3B GYS2 GYS1 PCK1 PCK2 G6PC G6PC2 G6PC3 MYC CCND1 CDKN1A CDKN1B CDK2 CDK4 CDK6 CCND2 CCND3 CCNE1 CCNE2 FOXO3 RBL2 FASLG BCL2L11 YWHAZ YWHAB YWHAQ YWHAE YWHAH YWHAG BAD BCL2L1 BCL2 CASP9 CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B MCL1 RXRA NR4A1 IKBKG CHUK IKBKB RELA NFKB1 MYB MDM2 TP53 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04152	AMPK signaling pathway	MAP3K7 STK11 CAB39 CAB39L STRADA STRADB LEP LEPR CAMKK1 CAMKK2 ADRA1A ADIPOQ ADIPOR1 ADIPOR2 PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 PFKFB1 PFKFB2 PFKFB3 PFKFB4 FBP1 FBP2 PFKM PFKP PFKL HNF4A G6PC G6PC2 G6PC3 PCK1 PCK2 CRTC2 CREB1 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 PPARGC1A ELAVL1 CCND1 CCNA2 CCNA1 EEF2K LOC101930123 EEF2 FOXO1 FOXO3 SIRT1 SLC2A4 GYS2 GYS1 LIPE HMGCR SREBF1 FASN ACACA SCD SCD5 MLYCD CPT1A CPT1B CPT1C PPP2CA PPP2CB PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D PPP2R3B PPP2R3C PPP2R3A PPP2R5B PPP2R5C PPP2R5D PPP2R5E PPP2R5A TBC1D1 RAB2A RAB8A RAB10 RAB11B RAB14 CFTR CD36 IGF1 IGF1R INS INSR IRS1 IRS2 IRS4 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PDPK1 AKT1 AKT2 AKT3 TSC2 TSC1 RHEB MTOR RPTOR AKT1S1 RPS6KB1 RPS6KB2 EIF4EBP1 PPARG ACACB ULK1
path:hsa04210	Apoptosis	FASLG FAS TNFSF10 TNFRSF10A TNFRSF10B TNFRSF10C TNFRSF10D TNF TNFRSF1A IL1A IL1B IL1R1 IL1RAP FADD TRADD CFLAR CASP10 CASP8 CASP3 CASP7 CASP6 BIRC2 BIRC3 XIAP BIRC7 BIRC8 BID BCL2 BCL2L1 CYCS APAF1 CASP9 DFFA DFFB AIFM1 ENDOG ATM TP53 RIPK1 TRAF2 MYD88 IRAK1 IRAK2 IRAK3 IRAK4 MAP3K14 CHUK IKBKB IKBKG NFKBIA NFKB1 RELA NGF NTRK1 IL3 IL3RA CSF2RB PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 PRKAR1A PRKAR2A PRKAR2B PRKAR1B PRKACA PRKACB PRKACG BAD BAX PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 CAPN1 CAPN2 CASP12

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04260	Cardiac muscle contraction	CACNA1C CACNA1D CACNA1F CACNA1S CACNB1 CACNB2 CACNB3 CACNB4 CACNA2D1 CACNA2D2 CACNA2D3 CACNA2D4 CACNG1 CACNG2 CACNG3 CACNG4 CACNG5 CACNG6 CACNG7 CACNG8 RYR2 TNNC1 TNNI3 TNNT2 TPM1 TPM2 TPM3 TPM4 ACTC1 MYH7 MYH6 MYL2 MYL3 MYL4 UQCRFS1 CYTB CYC1 UQCRC1 UQCRC2 UQCRH UQCRHL UQCRB UQCRQ UQCR10 UQCR11 COX3 COX1 COX2 COX4I2 COX4I1 COX5A COX5B COX6A1 COX6A2 COX6B1 COX6B2 COX6C COX7A1 COX7A2 COX7A2L COX7B COX7B2 COX7C COX8C COX8A ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 SLC9A1 SLC9A6 ATP2A2 SLC8A1
path:hsa04261	Adrenergic signaling in cardiomyocytes	ADRB1 ADRB2 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 SLC9A1 KCNQ1 KCNE1 ATP2B1 ATP2B3 ATP2B4 ATP2B2 SCN5A SCN7A SCN1B SCN4B CACNA1C CACNA1D CACNA1F CACNA1S CACNB1 CACNB2 CACNB3 CACNB4 CACNA2D1 CACNA2D2 CACNA2D3 CACNA2D4 CACNG1 CACNG2 CACNG3 CACNG4 CACNG5 CACNG6 CACNG7 CACNG8 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G RYR2 PPP2CA PPP2CB PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D PPP2R3B PPP2R3C PPP2R3A PPP2R5B PPP2R5C PPP2R5D PPP2R5E PPP2R5A PPP1CA PPP1CB PPP1CC TNNC1 TNNI3 TNNT2 TPM1 TPM2 TPM3 TPM4 ACTC1 MYH7 MYH6 MYL2 MYL3 MYL4 PLN AGTR1 AGTR2 ADRA1A ADRA1B ADRA1D GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PPP1R1A MAPK1 MAPK3 RPS6KA5 CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B CREM BCL2 RAPGEF3 RAPGEF4 MAPK11 MAPK12 MAPK13 MAPK14 GNAI1 GNAI3 GNAI2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 ATP2A2 SLC8A1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04270	Vascular smooth muscle contraction	ADRA1A ADRA1B ADRA1D AGTR1 EDNRA AVPR1A AVPR1B GNAQ GNA11 PLA2G10 PLA2G2D PLA2G2E PLA2G3 PLA2G2F PLA2G12A PLA2G12B PLA2G1B PLA2G5 PLA2G2A PLA2G2C PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLA2G6 CYP4A11 CYP4A22 KCNMA1 KCNU1 KCNMB1 KCNMB2 KCNMB3 KCNMB4 CACNA1C CACNA1D CACNA1F CACNA1S CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 MYLK MYLK2 MYLK3 MYLK4 MYL6B MYL6 MYL9 PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 PRKCA PRKCB PRKCG PRKCD PRKCE PRKCH PRKCQ ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 CALD1 ACTA2 ACTG2 PPP1R14A PPP1CA PPP1CB PPP1CC PPP1R12A PPP1R12B PPP1R12C GNA12 GNA13 ARHGEF12 ARHGEF1 ARHGEF11 RHOA ROCK1 ROCK2 ADORA2A ADORA2B PTGIR CALCRL RAMP1 RAMP2 RAMP3 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG NPR1 NPR2 PRKG1 MRVI1 GUCY1A2 GUCY1A3 GUCY1B3 MYH11

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04310	Wnt signaling pathway	PORCN WNT1 WNT2 WNT2B WNT3 WNT3A WNT4 WNT5A WNT5B WNT6 WNT7A WNT7B WNT8A WNT8B WNT9A WNT9B WNT10B WNT10A WNT11 WNT16 CER1 WIF1 SERPINF1 SOST DKK1 DKK2 DKK4 SFRP1 SFRP2 SFRP4 SFRP5 FZD1 FZD7 FZD2 FZD3 FZD4 FZD5 FZD8 FZD6 FZD10 FZD9 LRP5 LRP6 BAMBI CSNK1E LOC400927-CSNK1E DVL3 DVL2 DVL1 FRAT1 FRAT2 CSNK2A1 CSNK2A2 CSNK2B NKD1 NKD2 CXXC4 SENP2 GSK3B CTNNB1 AXIN1 AXIN2 APC APC2 CSNK1A1L CSNK1A1 TCF7 TCF7L1 TCF7L2 LEF1 CTNNBIP1 CHD8 SOX17 CTBP1 CTBP2 CREBBP EP300 RUVBL1 SMAD4 MAP3K7 NLK MYC JUN FOSL1 CCND1 CCND2 CCND3 PPARD MMP7 PSEN1 PRKACA PRKACB PRKACG TP53 SIAH1 CACYBP SKP1 TBL1X TBL1Y TBL1XR1 BTRC FBXW11 CUL1 RBX1 GPC4 VANGL2 VANGL1 PRICKLE1 PRICKLE2 INVS DAAM1 DAAM2 RHOA ROCK2 RAC1 RAC2 RAC3 MAPK8 MAPK10 MAPK9 PLCB1 PLCB2 PLCB3 PLCB4 CAMK2A CAMK2D CAMK2B CAMK2G PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 PRKCA PRKCB PRKCG NFATC1 NFATC2 NFATC3 NFATC4 SMAD3
path:hsa04320	Dorso-ventral axis formation	EGFR GRB2 SOS1 SOS2 KRAS MAP2K1 MAPK1 MAPK3 ETV6 ETV7 ETS1 ETS2 NOTCH1 NOTCH2 NOTCH3 NOTCH4 CPEB1 CPEB2 CPEB3 CPEB4 PIWIL4 PIWIL3 PIWIL2 PIWIL1 FMN2 SPIRE1 SPIRE2
path:hsa04330	Notch signaling pathway	DLL3 DLL1 DLL4 JAG1 JAG2 MFNG LFNG RFNG NOTCH1 NOTCH2 NOTCH3 NOTCH4 RBPJL RBPJ HES1 HES5 PTCRA DVL3 DVL2 DVL1 NUMB NUMBL DTX2 DTX3L DTX1 DTX3 DTX4 ADAM17 PSEN1 PSEN2 PSENEN NCSTN APH1A APH1B MAML3 MAML2 MAML1 CREBBP EP300 KAT2B KAT2A SNW1 CTBP1 CTBP2 NCOR2 CIR1 HDAC1 HDAC2
path:hsa04340	Hedgehog signaling pathway	SHH IHH DHH PTCH1 PTCH2 SMO STK36 SUFU GLI1 GLI2 GLI3 WNT1 WNT2 WNT2B WNT3 WNT3A WNT4 WNT5A WNT5B WNT6 WNT7A WNT7B WNT8A WNT8B WNT9A WNT9B WNT10B WNT10A WNT11 WNT16 BMP2 BMP4 HHIP GAS1 LRP2 RAB23 PRKACA PRKACB PRKACG GSK3B CSNK1A1L CSNK1A1 CSNK1G2 CSNK1G3 CSNK1G1 CSNK1D CSNK1E LOC400927-CSNK1E BTRC FBXW11 ZIC2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04350	TGF-beta signaling pathway	CHRD NOG NBL1 MINOS1-NBL1 THBS1 DCN LEFTY1 LEFTY2 FST BMP2 BMP4 BMP5 BMP6 BMP7 BMP8B BMP8A GDF7 GDF5 GDF6 AMH LTBP1 TGFB1 TGFB2 TGFB3 INHBA INHBB INHBC INHBE NODAL BMPR2 AMHR2 TGFBR2 ACVR2A ACVR2B BMPR1A BMPR1B ACVR1 TGFBR1 ACVR1B ACVR1C BAMBI SMAD1 SMAD5 SMAD9 SMAD2 SMAD3 SMAD4 SMAD6 SMAD7 SMURF1 SMURF2 ZFYVE9 ZFYVE16 ID1 ID2 ID3 ID4 RBL1 E2F4 E2F5 TFDP1 CREBBP EP300 SP1 TGIF1 TGIF2 MYC CDKN2B PITX2 RBX1 CUL1 SKP1 MAPK1 MAPK3 IFNG TNF RHOA ROCK1 PPP2R1B PPP2R1A PPP2CA PPP2CB RPS6KB1 RPS6KB2
path:hsa04360	Axon guidance	NTN1 DCC RAC1 RAC2 RAC3 ABLIM1 ABLIM3 ABLIM2 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 NFATC2 NFATC3 NFATC4 RHOA ROCK1 ROCK2 UNC5A UNC5B UNC5C UNC5D NTN3 NTN4 NTNG1 NTNG2 LRRC4C LRRC4 EFNA1 EFNA2 EFNA3 EFNA4 EFNA5 FYN EFNB1 EFNB2 EFNB3 PTK2 RGS3 EPHA1 EPHA2 EPHA3 EPHA4 EPHA5 EPHA6 EPHA7 EPHA8 EPHB1 EPHB2 EPHB3 EPHB4 EPHB6 ABL1 NGEF CDC42 PAK1 PAK2 PAK3 PAK4 PAK6 PAK7 RASA1 HRAS KRAS NRAS MAPK1 MAPK3 CXCR4 GNAI1 GNAI3 GNAI2 SLIT1 SLIT2 SLIT3 ROBO1 ROBO2 ROBO3 SRGAP2 SRGAP1 SRGAP3 SEMA3A SEMA3B SEMA3C SEMA3D SEMA3E SEMA3F SEMA3G PLXNA2 PLXNA3 PLXNA1 NRP1 L1CAM GSK3B CDK5 FES DPYSL2 DPYSL5 LIMK1 LIMK2 CFL1 CFL2 RHOD RND1 SEMA4F SEMA4D SEMA4B SEMA4C SEMA4G SEMA4A SEMA5A SEMA5B SEMA6A SEMA6B SEMA6C SEMA6D PLXNB1 PLXNB2 PLXNB3 MET ARHGEF12 SEMA7A ITGB1 PLXNC1 NCK1 CXCL12 NCK2
path:hsa04370	VEGF signaling pathway	VEGFA KDR SH2D2A PLCG1 PLCG2 PRKCA PRKCB PRKCG SPHK1 SPHK2 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 NFATC2 PTGS2 PTK2 SHC2 PXN CDC42 MAPK11 MAPK12 MAPK13 MAPK14 MAPKAPK2 MAPKAPK3 HSPB1 SRC PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 RAC1 RAC2 RAC3 AKT1 AKT2 AKT3 NOS3 CASP9 BAD

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04380	Osteoclast differentiation	CSF1 CSF1R GRB2 MAPK1 MAPK3 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 IFNG IFNGR1 IFNGR2 STAT1 IL1A IL1B IL1R1 TNF TNFRSF1A TGFB1 TGFB2 TGFBR1 TGFBR2 TNFSF11 TNFRSF11A TNFRSF11B TRAF2 TRAF6 LCK FYN MAP3K14 CHUK RELB NFKB2 MAP3K7 TAB1 TAB2 IKBKG IKBKB NFKBIA RELA NFKB1 NFATC2 NFATC1 IFNB1 MAP2K1 MAP2K6 MAPK11 MAPK12 MAPK13 MAPK14 MAP2K7 MAPK8 MAPK10 MAPK9 FOS FOSB FOSL2 FOSL1 JUN JUND JUNB RAC1 NOX3 NOX1 CYBB CYBA NCF2 NCF1 NCF4 BTK TEC OSCAR LILRB2 LILRB1 LILRB5 LILRB4 LILRA1 LILRB3 LILRA3 LILRA2 LILRA4 LILRA6 LILRA5 FCGR1A FCGR2A FCGR2B FCGR2C FCGR3A FCGR3B TREM2 SIRPA SIRPG SIRPB1 TYROBP SYK BLNK LCP2 PLCG2 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 CAMK4 CREB1 SPI1 MITF CTSK ACP5 CALCR ITGB3 PPARG IFNAR1 IFNAR2 JAK1 TYK2 STAT2 IRF9 SOCS1 SOCS3 GAB2 FHL2 CYLD SQSTM1
path:hsa04390	Hippo signaling pathway	CRB2 CRB1 PARD3 PARD6A PARD6G PARD6B PRKCZ PRKCI INADL MPP5 AMOT YAP1 WWTR1 CDH1 LIMD1 AJUBA WTIP NF2 WWC1 FRMD1 FRMD6 SAV1 STK3 RASSF6 RASSF1 PPP2CA PPP2CB PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D LATS2 LATS1 MOB1A MOB1B PPP1CA PPP1CB PPP1CC TP53BP2 LLGL2 LLGL1 SCRIB DLG1 DLG4 DLG2 DLG3 CSNK1D CSNK1E LOC400927-CSNK1E BTRC FBXW11 TP73 BBC3 TEAD1 TEAD4 TEAD3 TEAD2 CTGF GLI2 AREG BIRC5 AFP ITGB2 FGF1 TGFB1 TGFB2 TGFB3 TGFBR1 TGFBR2 SMAD7 SMAD2 SMAD3 SMAD4 SERPINE1 BMP2 BMP4 BMP5 BMP6 BMP7 BMP8B BMP8A GDF7 GDF5 GDF6 AMH BMPR1A BMPR1B BMPR2 SMAD1 ID1 ID2 WNT1 WNT2 WNT2B WNT3 WNT3A WNT4 WNT5A WNT5B WNT6 WNT7A WNT7B WNT8A WNT8B WNT9A WNT9B WNT10B WNT10A WNT11 WNT16 FZD1 FZD7 FZD2 FZD3 FZD4 FZD5 FZD8 FZD6 FZD10 FZD9 DVL3 DVL2 DVL1 YWHAZ YWHAB YWHAQ YWHAE YWHAH YWHAG GSK3B CTNNB1 APC APC2 AXIN1 AXIN2 TCF7 TCF7L1 TCF7L2 LEF1 MYC CCND1 CCND2 CCND3 SOX2 SNAI2 ACTB ACTG1 CTNNA3 CTNNA1 CTNNA2 BIRC2 NKD1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04510	Focal adhesion	COL1A1 COL1A2 COL2A1 COL4A2 COL4A4 COL4A6 COL4A1 COL4A5 COL4A3 COL6A1 COL6A2 COL6A3 COL6A6 COL6A5 COL9A1 COL9A2 COL9A3 LAMA1 LAMA2 LAMA3 LAMA5 LAMA4 LAMB1 LAMB2 LAMB3 LAMB4 LAMC1 LAMC2 LAMC3 CHAD RELN THBS1 COMP THBS2 THBS3 THBS4 FN1 SPP1 VTN TNC TNN TNR TNXB VWF IBSP ITGA1 ITGA2 ITGA2B ITGA3 ITGA4 ITGA5 ITGA6 ITGA7 ITGA8 ITGA9 ITGA10 ITGA11 ITGAV ITGB1 ITGB3 ITGB4 ITGB5 ITGB6 ITGB7 ITGB8 PDGFA PDGFB PDGFC PDGFD EGF IGF1 VEGFA VEGFB PGF VEGFC FIGF HGF PDGFRA PDGFRB IGF1R KDR EGFR FLT1 FLT4 MET ERBB2 SRC ARHGAP35 ARHGAP5 RHOA DIAPH1 ROCK1 ROCK2 MYL2 MYL5 MYL7 MYL9 MYL10 MYL12B MYL12A MYLPF PPP1CA PPP1CB PPP1CC PPP1R12A PPP1R12B PPP1R12C MYLK MYLK2 MYLK3 MYLK4 ACTB ACTG1 RASGRF1 CAPN2 ACTN1 ACTN2 ACTN3 ACTN4 TLN1 TLN2 FLNA FLNC FLNB PXN ILK ZYX VASP VCL PARVB PARVA PARVG PDPK1 AKT1 AKT2 AKT3 GSK3B CTNNB1 PRKCA PRKCB PRKCG PTK2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PTEN VAV3 VAV1 VAV2 RAC1 RAC2 RAC3 PAK1 PAK2 PAK3 PAK4 PAK6 PAK7 CDC42 BCAR1 CRK CRKL DOCK1 RAPGEF1 RAP1A RAP1B MAPK8 MAPK10 MAPK9 JUN BRAF CAV1 CAV2 CAV3 FYN SHC1 SHC2 SHC3 SHC4 GRB2 SOS1 SOS2 HRAS RAF1 MAP2K1 MAPK1 MAPK3 ELK1 CCND1 CCND2 CCND3 BIRC2 BIRC3 XIAP BAD BCL2 PIP5K1C
path:hsa04512	ECM-receptor interaction	COL1A1 COL1A2 COL2A1 COL4A2 COL4A4 COL4A6 COL4A1 COL4A5 COL4A3 COL6A1 COL6A2 COL6A3 COL6A6 COL6A5 COL9A1 COL9A2 COL9A3 LAMA1 LAMA2 LAMA3 LAMA5 LAMA4 LAMB1 LAMB2 LAMB3 LAMB4 LAMC1 LAMC2 LAMC3 CHAD RELN THBS1 COMP THBS2 THBS3 THBS4 FN1 SPP1 VTN TNC TNN TNR TNXB VWF IBSP AGRN HSPG2 ITGA1 ITGA2 ITGA2B ITGA3 ITGA4 ITGA5 ITGA6 ITGA7 ITGA8 ITGA9 ITGA10 ITGA11 ITGAV ITGB1 ITGB3 ITGB4 ITGB5 ITGB6 ITGB7 ITGB8 CD44 SDC1 SDC4 SV2C SV2B SV2A CD36 GP5 GP1BA GP1BB GP9 GP6 DAG1 CD47 HMMR

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04514	Cell adhesion molecules (CAMs)	CD58 CD2 CD80 CD274 CD28 CD86 CTLA4 ICOSLG ICOS HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD4 HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E CD8A CD8B PDCD1LG2 CD276 VTCN1 PDCD1 CD40 CD40LG ALCAM CD6 PVR CD226 PVRL2 TIGIT ITGAL ITGB2 ICAM1 ICAM2 ICAM3 CD22 PTPRC SIGLEC1 SPN PVRL3 CLDN4 CLDN3 CLDN7 CLDN19 CLDN16 CLDN14 CLDN15 CLDN17 CLDN20 CLDN11 CLDN18 CLDN22 CLDN5 CLDN10 CLDN8 CLDN6 CLDN2 CLDN1 CLDN9 CLDN23 CLDN25 CLDN24 OCLN F11R JAM2 JAM3 ESAM CDH5 PECAM1 CD99 ITGAM SELPLG SELP ITGA4 ITGB1 ITGA9 VCAM1 ITGB7 MADCAM1 SELL CD34 GLG1 SELE PVRL1 CDH2 NCAM1 NCAM2 L1CAM CADM1 NEGR1 NTNG1 LRRC4C NTNG2 LRRC4 PTPRF LRRC4B SDC1 SDC2 SDC3 SDC4 ITGAV ITGB8 ITGA8 NRXN1 NRXN2 NRXN3 NLGN1 NLGN2 NLGN3 NLGN4X CADM3 NRCAM CNTN1 PTPRM CNTN2 NFASC CNTNAP1 CNTNAP2 MPZ MPZL1 MAG CDH1 VCAN ITGA6 CDH3 CDH4 CDH15 NEO1
path:hsa04520	Adherens junction	PVRL1 PVRL2 PVRL3 PVRL4 PARD3 SRC FARP2 CDC42 RAC1 RAC2 RAC3 WAS WASL IQGAP1 BAIAP2 WASF1 WASF2 WASF3 MLLT4 LMO7 SSX2IP SORBS1 ACTN1 ACTN2 ACTN3 ACTN4 VCL TJP1 CDH1 CTNND1 CTNNB1 CTNNA3 CTNNA1 CTNNA2 ACTB ACTG1 RHOA PTPRM PTPRB PTPRF PTPN1 PTPN6 PTPRJ CSNK2A1 CSNK2A2 CSNK2B TCF7 TCF7L1 TCF7L2 LEF1 IGF1R INSR MET EGFR ERBB2 FGFR1 FYN YES1 MAPK1 MAPK3 SNAI2 SNAI1 TGFBR1 TGFBR2 SMAD2 SMAD3 SMAD4 CREBBP EP300 MAP3K7 NLK FER ACP1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04530	Tight junction	CLDN4 CLDN3 CLDN7 CLDN19 CLDN16 CLDN14 CLDN15 CLDN17 CLDN20 CLDN11 CLDN18 CLDN22 CLDN5 CLDN10 CLDN8 CLDN6 CLDN2 CLDN1 CLDN9 CLDN23 CLDN25 CLDN24 OCLN F11R JAM2 JAM3 IGSF5 CRB3 MPP5 INADL PARD6A PARD6G PARD6B PRKCZ PRKCI PARD3 LLGL2 LLGL1 CDC42 PPP2CA PPP2CB PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D PRKCA PRKCB PRKCG PRKCD PRKCE PRKCH PRKCQ CSNK2A1 CSNK2A2 CSNK2B GNAI1 GNAI3 GNAI2 MPDZ VAPA TJP3 TJP1 TJP2 ZAK CTNNB1 YBX3 SYMPK ASH1L CDK4 CGN ACTB ACTG1 MYL2 MYL5 MYL7 MYL9 MYL10 MYL12B MYL12A MYLPF MYH15 MYH1 MYH2 MYH3 MYH4 MYH8 MYH9 MYH10 MYH11 MYH7B MYH14 MYH13 MYH7 MYH6 RHOA MLLT4 HRAS KRAS NRAS RRAS RRAS2 MRAS CASK AMOTL1 TJAP1 CTTN HCLS1 CTNNA3 CTNNA1 CTNNA2 EPB41 EPB41L1 EPB41L2 EPB41L3 ACTN1 ACTN2 ACTN3 ACTN4 SHROOM1 SHROOM2 SHROOM3 SHROOM4 SRC YES1 RAB3B RAB13 EXOC3 EXOC4 MAGI1 MAGI2 MAGI3 PTEN AKT1 AKT2 AKT3 SPTAN1
path:hsa04540	Gap junction	GJA1 GJD2 LPAR1 GNAI1 GNAI3 GNAI2 PDGFA PDGFB PDGFC PDGFD EGF PDGFRA PDGFRB EGFR GRB2 SOS1 SOS2 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 SRC MAP3K2 MAP2K5 MAPK7 TUBA1B TUBA4A TUBA3C TUBA1A TUBA1C TUBA8 TUBA3E TUBA3D TUBAL3 TUBB6 TUBB TUBB1 TUBB2A TUBB3 TUBB4A TUBB8 TUBB2B TUBB4B CSNK1D CDK1 TJP1 ADRB1 DRD1 GNAS DRD2 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG HTR2A HTR2B HTR2C GRM1 GRM5 GNA11 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 PRKCA PRKCB PRKCG GUCY1A2 GUCY1A3 GUCY1B3 PRKG1 PRKG2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04550	Signaling pathways regulating pluripotency of stem cells	LIF LIFR IL6ST JAK1 JAK2 JAK3 STAT3 KLF4 SOX2 MYC GRB2 MAP2K1 MAP2K2 MAPK1 MAPK3 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 TBX3 NANOG INHBA INHBB INHBC INHBE NODAL ACVR1B ACVR1C ACVR2A ACVR2B SMAD2 SMAD3 BMP2 BMP4 BMPR1A BMPR1B ACVR1 BMPR2 SMAD1 SMAD5 SMAD9 SMAD4 ID1 ID2 ID3 ID4 DUSP9 MAPK11 MAPK12 MAPK13 MAPK14 WNT1 WNT2 WNT2B WNT3 WNT3A WNT4 WNT5A WNT5B WNT6 WNT7A WNT7B WNT8A WNT8B WNT9A WNT9B WNT10B WNT10A WNT11 WNT16 FZD1 FZD7 FZD2 FZD3 FZD4 FZD5 FZD8 FZD6 FZD10 FZD9 DVL3 DVL2 DVL1 GSK3B AXIN1 AXIN2 APC APC2 CTNNB1 TCF3 ESRRB HNF1A POU5F1 POU5F1B FGF2 FGFR1 FGFR2 FGFR3 FGFR4 HRAS KRAS NRAS RAF1 IGF1 IGF1R HESX1 ZIC3 SKIL SMARCAD1 KAT6A SETDB1 JARID2 REST RIF1 PCGF1 PCGF2 PCGF3 BMI1 COMMD3-BMI1 PCGF5 PCGF6 PAX6 MEIS1 LHX5 OTX1 NEUROG1 HAND1 DLX5 MYF5 ONECUT1 ISL1 ZFHX3 ESX1 HOXB1 LEFTY2
path:hsa04610	Complement and coagulation cascades	F12 KLKB1 KNG1 BDKRB1 BDKRB2 F11 F9 F8 VWF F3 F7 F10 F5 F2 FGA FGB FGG F13A1 F13B THBD PROC PROS1 PLAT PLAU PLG TFPI SERPIND1 CPB2 SERPINE1 SERPINF2 SERPINC1 SERPINA5 A2M SERPINA1 F2R PLAUR C1QA C1QB C1QC C1R C1S C2 C4A C4B C3 CFB CFD MBL2 MASP1 MASP2 C5 C6 C7 C8A C8B C8G C9 SERPING1 C4BPA C4BPB CFH CFI CD55 CD46 CD59 C3AR1 C5AR1 CR1 CR2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04611	Platelet activation	TBXA2R F2R F2RL3 GNA13 ARHGEF1 ARHGEF12 RHOA ROCK1 ROCK2 PPP1CA PPP1CB PPP1CC PPP1R12A MYL12B MYL12A P2RX1 ORAI1 ITPR1 ITPR2 ITPR3 STIM1 MYLK MYLK2 MYLK3 MYLK4 P2RY1 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCZ PRKCI RASGRP1 RASGRP2 RAP1A RAP1B APBB1IP TLN1 TLN2 ITGA2B ITGB3 FERMT3 FGA FGB FGG P2RY12 GNAI1 GNAI3 GNAI2 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PTGIR GNAS PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PRKACA PRKACB PRKACG VASP ACTB ACTG1 SRC ARHGAP35 FCGR2A SYK COL1A1 COL1A2 COL3A1 GP6 FCER1G LYN FYN ITGA2 ITGB1 LCP2 PLCG2 BTK VWF GP5 GP1BA GP1BB GP9 AKT1 AKT2 AKT3 NOS3 GUCY1A2 GUCY1A3 GUCY1B3 PRKG1 PRKG2 MAPK11 MAPK12 MAPK13 MAPK14 MAPK1 MAPK3 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PTGS1 TBXAS1 SNAP23 VAMP8
path:hsa04612	Antigen processing and presentation	IFNG TNF PSME1 PSME2 PSME3 HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 HSPA4 HSP90AA1 HSP90AB1 HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E CANX B2M PDIA3 CALR TAPBP TAP1 TAP2 CD8A CD8B KIR3DL2 KIR3DL1 KIR3DL3 KIR2DL2 KIR2DL1 KIR2DL3 KIR2DL4 KIR2DL5A KLRC1 KLRC2 KLRC3 KLRC4 KLRD1 KIR2DS1 KIR2DS3 KIR2DS4 KIR2DS5 KIR2DS2 IFI30 LGMN CTSB HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD74 CTSL CTSS CD4 CIITA RFX5 RFXANK RFXAP CREB1 NFYA NFYB NFYC HSPA5
path:hsa04614	Renin-angiotensin system	AGT REN ACE CMA1 CTSG KLK2 KLK1 ENPEP ANPEP PREP ACE2 CTSA CPA3 MME THOP1 NLN PRCP MAS1 MRGPRD AGTR1 AGTR2 LNPEP ATP6AP2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04620	Toll-like receptor signaling pathway	TLR1 TLR2 TLR6 LBP CD14 LY96 TLR3 TLR4 TLR5 TLR7 TLR8 CTSK TLR9 RAC1 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 TOLLIP MYD88 TIRAP FADD CASP8 IRAK4 IRAK1 TRAF6 TAB1 TAB2 MAP3K7 IKBKG CHUK IKBKB NFKBIA NFKB1 RELA MAP3K8 MAP2K1 MAP2K2 MAPK1 MAPK3 MAP2K3 MAP2K6 MAP2K4 MAP2K7 MAPK11 MAPK12 MAPK13 MAPK14 MAPK8 MAPK10 MAPK9 JUN FOS TNF IL1B IL6 IL12A IL12B CXCL8 CCL5 CCL3 CCL3L1 CCL3L3 CCL4 CCL4L2 CCL4L1 TICAM2 TICAM1 RIPK1 IRF5 IRF7 SPP1 IKBKE TBK1 TRAF3 IRF3 CD40 CD80 CD86 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 IFNAR1 IFNAR2 STAT1 CXCL10 CXCL9 CXCL11
path:hsa04621	NOD-like receptor signaling pathway	NOD1 NOD2 TRIP6 RIPK2 TRAF6 IKBKG CHUK IKBKB NFKBIB NFKBIA NFKB1 RELA CXCL8 CCL2 CCL5 TNF IL6 MAP3K7 TAB1 TAB2 TAB3 MAPK1 MAPK3 MAPK8 MAPK10 MAPK9 MAPK11 MAPK12 MAPK13 MAPK14 CARD9 CASP8 SUGT1 ERBB2IP CARD6 BIRC2 BIRC3 TNFAIP3 NLRP1 PYCARD CASP1 CASP5 NLRP3 CARD8 PYDC1 MEFV PSTPIP1 HSP90AA1 HSP90AB1 HSP90B1 NLRC4 CARD18 NAIP IL1B IL18 CXCL1 CXCL2
path:hsa04622	RIG-I-like receptor signaling pathway	DDX58 IFIH1 MAVS DHX58 TRAF3 TANK AZI2 TBKBP1 IKBKG TBK1 IKBKE IRF3 IRF7 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 IFNW1 IFNE IFNK TRADD FADD RIPK1 CASP8 CASP10 CHUK IKBKB NFKBIB NFKBIA NFKB1 RELA TRAF2 MAP3K7 TRAF6 MAP3K1 MAPK8 MAPK10 MAPK9 MAPK11 MAPK12 MAPK13 MAPK14 CXCL8 TNF IL12A IL12B CXCL10 TRIM25 CYLD RNF125 ISG15 ATG5 ATG12 NLRX1 TMEM173 OTUD5 SIKE1 DDX3X PIN1 DAK

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04623	Cytosolic DNA-sensing pathway	POLR3A POLR3B POLR3C POLR3D POLR3E LOC101060521 POLR1C POLR3K POLR1D POLR3H POLR3GL POLR3G POLR3F POLR2E POLR2F POLR2H POLR2K POLR2L DDX58 MAVS NFKB1 RELA IL6 MB21D1 TMEM173 TBK1 IKBKE IRF3 IRF7 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 ZBP1 RIPK1 RIPK3 IKBKG CHUK IKBKB NFKBIB NFKBIA CCL4 CCL4L2 CCL4L1 CCL5 CXCL10 AIM2 PYCARD CASP1 IL1B IL18 IL33 TREX1 ADAR
path:hsa04630	Jak-STAT signaling pathway	IL2 IL3 IL4 IL5 IL6 IL7 IL9 IL10 IL11 IL12A IL12B IL13 IL15 IL19 IL20 IL24 IL21 IL22 IL23A IL17D IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 IFNG IFNE IFNK IFNL2 IFNL3 IFNL1 IFNW1 OSM LIF TSLP CTF1 CSF2 CNTF CSF3 EPO GH1 GH2 CSH1 CSH2 LEP THPO PRL IL2RA IL2RB IL2RG IL3RA IL4R IL5RA IL6R IL7R IL9R IL10RA IL10RB IL11RA IL12RB1 IL12RB2 IL13RA1 IL13RA2 IL15RA IL20RA IL20RB IL21R IL22RA1 IL22RA2 IL23R IL27RA IL6ST IFNAR1 IFNAR2 IFNGR1 IFNGR2 IFNLR1 OSMR LIFR CRLF2 CNTFR CSF2RA CSF2RB CSF3R EPOR GHR LEPR MPL PRLR JAK1 JAK2 JAK3 TYK2 STAT1 STAT2 STAT3 STAT4 STAT5A STAT5B STAT6 CISH SOCS1 SOCS2 SOCS3 SOCS4 SOCS5 SOCS7 SOCS6 BCL2 MCL1 BCL2L1 PIM1 MYC CCND1 CCND2 CCND3 CDKN1A AOX1 GFAP STAM2 STAM PTPN2 PTPN6 IRF9 CREBBP EP300 PIAS1 PIAS2 PIAS3 PIAS4 FHL1 PTPN11 GRB2 SOS1 SOS2 HRAS RAF1 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 MTOR
path:hsa04640	Hematopoietic cell lineage	KITLG IL7 IL4 CSF2 FLT3LG IL5 CSF3 IL3 IL6 IL11 IL1A IL1B TNF CSF1 EPO THPO CD34 FLT3 DNTT CD44 KIT IL2RA IL7R TFRC CD7 CD2 CD5 CD1A CD1B CD1C CD1D CD1E CD4 CD8A CD8B CD3D CD3E CD3G MME CD9 CD19 CD22 CD24 MS4A1 CR2 CD37 FCER2 CR1 CSF2RA IL3RA CD33 IL4R IL6R FCGR1A CSF1R ANPEP ITGAM CD14 IL9R IL1R1 IL1R2 CSF3R IL5RA EPOR CD36 CD55 CD59 IL11RA ITGB3 ITGA2B GP9 GP1BA GP1BB GP5 ITGA1 ITGA2 ITGA3 ITGA4 ITGA5 ITGA6 GYPA HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD38

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04650	Natural killer cell mediated cytotoxicity	HLA-A HLA-B HLA-C HLA-G HLA-E KIR3DL2 KIR3DL1 KIR2DL2 KIR2DL1 KIR2DL3 KIR2DL4 KIR2DL5A KLRC1 KLRC2 KLRC3 KLRD1 PTPN6 PTPN11 ICAM1 ICAM2 ITGAL ITGB2 PTK2B VAV3 VAV1 VAV2 RAC1 RAC2 RAC3 PAK1 MAP2K1 MAP2K2 MAPK1 MAPK3 TNF CSF2 IFNG KIR2DS1 KIR2DS3 KIR2DS4 KIR2DS5 KIR2DS2 NCR2 TYROBP LCK FCGR3A FCGR3B NCR1 NCR3 FCER1G CD247 ZAP70 SYK LCP2 LAT PLCG1 PLCG2 SH3BP2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 FYN SHC1 SHC2 SHC3 SHC4 GRB2 SOS1 SOS2 HRAS KRAS NRAS ARAF BRAF RAF1 MICB MICA ULBP1 ULBP2 ULBP3 RAET1G RAET1L RAET1E KLRK1 KLRC4-KLRK1 HCST CD48 CD244 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 NFATC1 NFATC2 PRKCA PRKCB PRKCG SH2D1B SH2D1A IFNGR1 IFNGR2 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 IFNAR1 IFNAR2 TNFSF10 TNFRSF10A TNFRSF10B TNFRSF10C TNFRSF10D FASLG FAS GZMB PRF1 CASP3 BID
path:hsa04660	T cell receptor signaling pathway	CD3D CD3E CD3G CD247 CD4 CD8A CD8B PTPRC LCK FYN ZAP70 LCP2 LAT ITK TEC NCK1 NCK2 VAV3 VAV1 VAV2 GRAP2 GRB2 PAK1 PAK2 PAK3 PAK4 PAK6 PAK7 RHOA CDC42 DLG1 MAPK11 MAPK12 MAPK13 MAPK14 PLCG1 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 NFATC1 NFATC2 NFATC3 SOS1 SOS2 RASGRP1 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 FOS JUN PRKCQ CARD11 BCL10 MALT1 MAP3K7 MAP2K7 CHUK IKBKB IKBKG NFKB1 RELA NFKBIA NFKBIB NFKBIE CD28 ICOS CD40LG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG PDPK1 AKT1 AKT2 AKT3 MAP3K8 MAP3K14 GSK3B PDCD1 CTLA4 PTPN6 CBLC CBL CBLB IL2 IL4 IL5 IL10 IFNG CSF2 TNF CDK4 MAPK9

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04662	B cell receptor signaling pathway	CD79A CD79B LYN SYK BTK DAPP1 BLNK VAV3 VAV1 VAV2 RAC1 RAC2 RAC3 PLCG2 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 NFATC1 NFATC2 NFATC3 GRB2 SOS1 SOS2 RASGRP3 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 FOS JUN PRKCB CARD11 BCL10 MALT1 CHUK IKBKB IKBKG NFKB1 RELA NFKBIA NFKBIB NFKBIE CD81 CD19 CR2 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG AKT1 AKT2 AKT3 GSK3B INPP5D INPPL1 CD22 CD72 PTPN6 PIK3AP1 LILRB3 FCGR2B IFITM1
path:hsa04664	Fc epsilon RI signaling pathway	FCER1A MS4A2 FCER1G SYK LYN BTK INPP5D PLCG1 PLCG2 PRKCA PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PDPK1 AKT1 AKT2 AKT3 RAC1 RAC2 RAC3 MAP2K4 MAP2K7 MAP2K3 MAP2K6 MAPK8 MAPK10 MAPK9 MAPK11 MAPK12 MAPK13 MAPK14 IL4 IL13 IL3 IL5 CSF2 TNF LCP2 VAV3 VAV1 VAV2 FYN GAB2 LAT GRB2 SOS1 SOS2 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F
path:hsa04666	Fc gamma R-mediated phagocytosis	FCGR1A FCGR2A PTPRC HCK LYN SYK PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 RPS6KB1 RPS6KB2 PLCG1 PLCG2 PRKCD PRKCE RAF1 MAP2K1 MAPK1 MAPK3 MARCKS MARCKSL1 PLD1 PLD2 PPAP2A PPAP2B PPAP2C SPHK1 SPHK2 PRKCA PRKCB PRKCG NCF1 GSN SCIN VAV3 VAV1 VAV2 CDC42 WAS WASL VASP ARPC5 ARPC5L ARPC4 ARPC3 ARPC1B ARPC1A ARPC2 RAC1 RAC2 WASF1 WASF2 WASF3 PAK1 LIMK1 LIMK2 CFL1 CFL2 PIP5K1C PIP5K1A PIP5K1B ARF6 CRK CRKL DOCK2 ASAP1 ASAP3 ASAP2 FCGR2B INPP5D INPPL1 GAB2 LAT AMPH BIN1 MYO10 PLA2G4B PLA2G4E DNM2 FCGR3A PLA2G4F PLA2G4D PLA2G4A PLA2G6

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04668	TNF signaling pathway	TNF TNFRSF1A BAG4 TRADD TRAF2 TRAF5 RIPK1 BIRC2 BIRC3 MAP3K7 TAB1 TAB2 TAB3 MAP2K4 MAP2K7 MAPK8 MAPK10 MAPK9 JUN ITCH CFLAR MAP2K3 MAP2K6 MAPK11 MAPK12 MAPK13 MAPK14 CEBPB MAP3K5 MAP3K14 IKBKG IKBKB CHUK NFKBIA RELA NFKB1 MAP3K8 MAP2K1 MAPK1 MAPK3 RPS6KA5 RPS6KA4 CREB1 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 ATF2 ATF4 CREB5 ATF6B RIPK3 MLKL PGAM5 DNM1L FADD CASP8 CASP10 CASP7 CASP3 CCL2 CCL5 CCL20 CXCL1 CXCL2 CXCL3 CXCL10 CX3CL1 CSF1 CSF2 FAS IL18R1 IL1B IL6 IL15 LIF LTA BCL3 SOCS3 TNFAIP3 TRAF1 FOS JUNB MMP3 MMP9 MMP14 EDN1 NOD2 ICAM1 SELE VCAM1 PTGS2 TNFRSF1B TRAF3 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 MAGI2 JAG1 CXCL5 VEGFC
path:hsa04670	Leukocyte transendothelial migration	JAM3 ITGAM ITGB2 JAM2 ITGA4 ITGB1 PECAM1 CD99 ITGAL F11R CDH5 CLDN4 CLDN3 CLDN7 CLDN19 CLDN16 CLDN14 CLDN15 CLDN17 CLDN20 CLDN11 CLDN18 CLDN22 CLDN5 CLDN10 CLDN8 CLDN6 CLDN2 CLDN1 CLDN9 CLDN23 CLDN25 CLDN24 OCLN ESAM VCAM1 EZR MSN ACTB ACTG1 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 RAC1 NOX3 NOX1 CYBB CYBA NCF2 NCF1 NCF4 CTNNB1 CTNND1 CTNNA3 CTNNA1 CTNNA2 PTPN11 MMP2 MMP9 MAPK11 MAPK12 MAPK13 MAPK14 ICAM1 PLCG1 PLCG2 PRKCA PRKCB PRKCG PTK2 PXN BCAR1 THY1 ARHGAP35 ARHGAP5 RHOA ROCK1 ROCK2 MYL2 MYL5 MYL7 MYL9 MYL10 MYL12B MYL12A MYLPF MLLT4 RAP1A RAP1B SIPA1 VASP ACTN1 ACTN2 ACTN3 ACTN4 VCL CXCL12 CXCR4 GNAI1 GNAI3 GNAI2 RAPGEF3 RAPGEF4 RASSF5 PTK2B ITK TXK VAV3 VAV1 VAV2 RAC2 CDC42 RHOH
path:hsa04672	Intestinal immune network for IgA production	CD80 CD86 HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD28 IL2 IL4 IL5 IL6 IL10 TGFB1 TNFSF13 TNFSF13B TNFRSF13B TNFRSF17 TNFRSF13C AICDA CD40 CD40LG ICOS ICOSLG CCR9 ITGA4 ITGB7 CXCL12 CXCR4 CCL28 CCR10 CCL25 MADCAM1 LTBR MAP3K14 IL15 IL15RA PIGR

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04710	Circadian rhythm	CSNK1D CSNK1E LOC400927-CSNK1E PER1 PER3 PER2 CRY1 CRY2 ARNTL CLOCK NPAS2 NR1D1 RORA RORB RORC BHLHE40 BHLHE41 RBX1 CUL1 SKP1 BTRC FBXW11 FBXL3 PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 CREB1
path:hsa04713	Circadian entrainment	GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G NOS1 MAPK1 MAPK3 RPS6KA5 CREB1 ITPR1 ITPR3 RYR1 RYR2 RYR3 NOS1AP RASD1 GNAI1 GNAI3 GNAI2 GNAO1 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 GRIA1 GRIA2 GRIA3 GRIA4 CACNA1G CACNA1H CACNA1I ADCYAP1R1 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 ADCY10 PRKACA PRKACB PRKACG GUCY1A2 GUCY1A3 GUCY1B3 PRKG1 PRKG2 CACNA1C CACNA1D MTNR1B MTNR1A GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG KCNJ3 KCNJ5 KCNJ6 KCNJ9 PER1 PER3 PER2 FOS
path:hsa04720	Long-term potentiation	GRIA1 GRIA2 ADCY1 ADCY8 PRKACA PRKACB PRKACG PPP1R1A PPP1CA PPP1CB PPP1CC CAMK2A CAMK2D CAMK2B CAMK2G RAPGEF3 RAP1A RAP1B PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D CACNA1C CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CREBBP EP300 ATF4 CAMK4 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 RPS6KA3 RPS6KA1 RPS6KA2 RPS6KA6 ITPR1 ITPR2 ITPR3 PRKCA PRKCB PRKCG GRM1 GRM5 GNAQ PLCB1 PLCB2 PLCB3 PLCB4

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04721	Synaptic vesicle cycle	SLC32A1 SLC18A1 SLC18A2 SLC18A3 SLC17A6 SLC17A8 SLC17A7 SYT1 VAMP2 RAB3A RIMS1 STX1A STX2 STX3 STX1B STXBP1 UNC13A UNC13B UNC13C SNAP25 CPLX1 CPLX2 CPLX3 CPLX4 CACNA1A CACNA1B NSF NAPA DNM1 DNM2 DNM3 CLTA CLTB CLTC CLTCL1 AP2A2 AP2A1 AP2B1 AP2M1 AP2S1 ATP6V1A ATP6V1B1 ATP6V1B2 ATP6V1C2 ATP6V1C1 ATP6V1D ATP6V1E2 ATP6V1E1 ATP6V1F ATP6V1G1 ATP6V1G3 ATP6V1G2 ATP6V0E1 ATP6V0E2 TCIRG1 ATP6V0A2 ATP6V0A4 ATP6V0A1 ATP6V0D1 ATP6V0D2 ATP6V1H ATP6V0C ATP6V0B
path:hsa04722	Neurotrophin signaling pathway	NGF BDNF NTF4 NTF3 NTRK1 NTRK2 NTRK3 SH2B2 SH2B1 SH2B3 GRB2 SOS1 SOS2 HRAS KRAS NRAS RAF1 BRAF MAP2K1 MAP2K2 MAPK1 MAPK3 RPS6KA3 RPS6KA1 RPS6KA2 RPS6KA6 RPS6KA5 ATF4 BCL2 KIDINS220 FRS2 CRK CRKL RAPGEF1 RAP1A RAP1B MAP3K3 MAP2K5 MAPK7 MAPK11 MAPK12 MAPK13 MAPK14 MAPKAPK2 SHC1 SHC2 SHC3 SHC4 GAB1 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG AKT1 AKT2 AKT3 NFKBIB NFKBIA NFKBIE NFKB1 RELA FOXO3 FASLG BAD GSK3B PDPK1 IRS1 PLCG1 PLCG2 PRKCD CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G CAMK4 ABL1 PTPN11 MATK NGFR ARHGDIB ARHGDIA ARHGDIG RHOA CDC42 RAC1 MAP3K1 MAP3K5 MAP2K7 MAPK8 MAPK10 MAPK9 JUN TP53 BAX TP73 TRAF6 ZNF274 PRDM4 MAGED1 NGFRAP1 YWHAE RIPK2 IRAK1 IRAK2 IRAK3 IRAK4 IKBKB SORT1 PSEN1 PSEN2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04723	Retrograde endocannabinoid signaling	SLC17A6 SLC17A8 SLC17A7 GRIA1 GRIA2 GRIA3 GRIA4 CACNA1C CACNA1D CACNA1F CACNA1S GRM1 GRM5 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 NAPEPLD PRKCA PRKCB PRKCG DAGLA DAGLB PTGS2 ABHD6 FAAH CNR1 GNAI1 GNAI3 GNAI2 GNAO1 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 KCNJ3 KCNJ6 KCNJ9 KCNJ5 CACNA1A CACNA1B MAPK1 MAPK3 MAPK8 MAPK10 MAPK9 MAPK11 MAPK12 MAPK13 MAPK14 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG MGLL SLC32A1 RIMS1 GABRA1 GABRA2 GABRA3 GABRA4 GABRA5 GABRA6 GABRB1 GABRB3 GABRB2 GABRG1 GABRG2 GABRG3 GABRD GABRE GABRQ GABRP GABRR1 GABRR2 GABRR3
path:hsa04724	Glutamatergic synapse	SLC38A1 SLC38A2 GLS2 GLS SLC17A6 SLC17A8 SLC17A7 GRIK1 GRIK2 GRIK3 GRIK4 GRIK5 GRIA1 GRIA2 GRIA3 GRIA4 GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D GRIN3A GRIN3B PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 DLG4 DLGAP1 SHANK3 SHANK1 SHANK2 TRPC1 GRM1 GRM5 HOMER1 HOMER2 HOMER3 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG ITPR1 ITPR2 ITPR3 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLD1 PLD2 MAPK1 MAPK3 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG GRM2 GRM3 GRM4 GRM6 GRM7 GRM8 GNAI1 GNAI3 GNAI2 GNAO1 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 ADRBK1 ADRBK2 KCNJ3 CACNA1A SLC1A2 SLC1A7 SLC1A1 SLC1A6 CACNA1C CACNA1D SLC1A3 GLUL SLC38A3

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04725	Cholinergic synapse	CHAT ACHE SLC18A3 CHRM1 CHRM3 CHRM5 GNAQ GNA11 PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 PRKCA PRKCB PRKCG KCNQ1 KCNQ2 KCNQ3 KCNQ4 KCNQ5 KCNJ2 KCNJ12 KCNJ4 KCNJ14 CHRM2 CHRM4 GNAI1 GNAI3 GNAI2 GNAO1 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 KCNJ3 KCNJ6 HRAS KRAS NRAS MAP2K1 MAPK1 MAPK3 FOS CHRNA7 CHRNA4 CHRNB2 CHRNA3 CHRNB4 CHRNA6 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG CREB1 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 CAMK2A CAMK2D CAMK2B CAMK2G CAMK4 JAK2 FYN PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 BCL2 CACNA1A CACNA1B CACNA1C CACNA1D CACNA1F CACNA1S SLC5A7
path:hsa04726	Serotonergic synapse	TPH2 TPH1 DDC SLC18A1 SLC18A2 CACNA1C CACNA1D CACNA1F CACNA1S HTR2A HTR2B HTR2C GNAQ PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 PRKCA PRKCB PRKCG MAPK1 MAPK3 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F CYP2C8 CYP2C9 CYP2C18 CYP2C19 CYP2D6 CYP2J2 CYP4X1 ALOX5 ALOX12 ALOX12B ALOX15 ALOX15B PTGS1 PTGS2 HTR3A HTR3E HTR3D HTR3C HTR3B HTR4 HTR6 HTR7 GNAS ADCY5 PRKACA PRKACB PRKACG KCNN2 KCND2 GABRB1 GABRB3 GABRB2 RAPGEF3 APP HTR1A HTR1B HTR1D HTR1E HTR1F HTR5A GNAI1 GNAI3 GNAI2 GNAO1 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 CASP3 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 CACNA1A CACNA1B KCNJ3 KCNJ6 KCNJ9 KCNJ5 TRPC1 SLC6A4 MAOB MAOA DUSP1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04727	GABAergic synapse	SLC38A1 SLC38A2 GLS2 GLS GAD1 GAD2 SLC32A1 ABAT GABRA1 GABRA2 GABRA3 GABRA4 GABRA5 GABRA6 GABRB1 GABRB3 GABRB2 GABRG1 GABRG2 GABRG3 GABRD GABRE GABRQ GABRP PRKACA PRKACB PRKACG SRC PRKCA PRKCB PRKCG HAP1 GABARAP GABARAPL1 GABARAPL2 NSF TRAK2 PLCL1 GPHN GABRR1 GABRR2 GABRR3 CACNA1A CACNA1B CACNA1C CACNA1D CACNA1F CACNA1S GABBR1 GABBR2 GNAI1 GNAI3 GNAI2 GNAO1 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 KCNJ6 SLC6A1 GLUL SLC38A3 SLC38A5 SLC12A5 SLC6A11 SLC6A13
path:hsa04728	Dopaminergic synapse	TH DDC SLC18A1 SLC18A2 DRD1 DRD5 CALY GNAQ PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G PPP3CA PPP3CB PPP3CC PRKCA PRKCB PRKCG FOS GNAS GNAL ADCY5 PRKACA PRKACB PRKACG CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B MAPK11 MAPK12 MAPK13 MAPK14 MAPK8 MAPK10 MAPK9 PPP1R1B PPP1CA PPP1CB PPP1CC SCN1A CACNA1C CACNA1D CACNA1A CACNA1B KCNJ3 KCNJ6 KCNJ9 KCNJ5 DRD3 DRD4 GNAI1 GNAI3 GNAI2 GNAO1 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 DRD2 PPP2CA PPP2CB PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D PPP2R3B PPP2R3C PPP2R3A PPP2R5B PPP2R5C PPP2R5D PPP2R5E PPP2R5A AKT1 AKT2 AKT3 GSK3A GSK3B GRIN2A GRIN2B GRIA1 GRIA2 GRIA3 GRIA4 KIF5A KIF5B KIF5C CLOCK ARNTL SLC6A3 MAOB MAOA COMT ARRB2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04730	Long-term depression	NOS1 GUCY1A2 GUCY1A3 GUCY1B3 PRKG1 PRKG2 PPP1R17 PPP2R1B PPP2R1A PPP2CA PPP2CB HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 GRID2 GRM1 GNAI1 GNAI3 GNAI2 GNAO1 GNAZ GNAS GNA12 GNA13 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PRKCA PRKCB PRKCG GNAQ GNA11 PLCB1 PLCB2 PLCB3 PLCB4 GRIA1 GRIA2 GRIA3 LYN CACNA1A ITPR1 ITPR2 ITPR3 RYR1 CRH CRHR1 IGF1 IGF1R

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04740	Olfactory transduction	OR2J3 OR14J1 OR10C1 OR2A2 OR52K2 OR5P2 OR5P3 OR8I2 OR2D3 OR52E2 OR52J3 OR51L1 OR51A7 OR51S1 OR51F2 OR52R1 OR4C46 OR4X2 OR52M1 OR1E1 OR1E2 OR1D2 OR1D5 OR1G1 OR1A1 OR1A2 OR2J2 OR2A4 OR2C1 OR2F1 OR2W1 OR3A1 OR3A2 OR3A3 OR5I1 OR5V1 OR6A2 OR7A17 OR7A5 OR10H1 OR10H2 OR10H3 OR10J1 OR11A1 OR12D2 OR12D3 OR51E2 OR52A1 OR51B4 OR51B2 OR4N4 OR5B3 OR9K2 OR4Q3 OR4M1 OR13G1 OR2L13 OR52E6 OR52E8 OR52E4 OR56A3 OR56A5 OR10A6 OR4X1 OR5D13 OR5D16 OR8H2 OR8H3 OR5T3 OR5T1 OR8K1 OR5M9 OR5M10 OR5M1 OR9G1 OR2AG1 OR6B3 OR1Q1 OR7D2 OR56B4 OR8U1 OR4C16 OR4C11 OR4S2 OR4C6 OR5D14 OR5L1 OR5D18 OR5AS1 OR8K5 OR5T2 OR8H1 OR8K3 OR8J1 OR5R1 OR5M3 OR5M8 OR5AR1 OR8B12 OR8G5 OR10G8 OR10G9 OR10S1 OR6T1 OR4D5 OR6Q1 OR9I1 OR9Q1 OR9Q2 OR1S2 OR1S1 OR10Q1 OR5B17 OR5B21 OR5A2 OR5A1 OR4D6 OR4D11 OR6C74 OR6C3 OR1L4 OR52B2 OR4C3 OR4S1 OR51F1 OR1C1 OR2B6 OR1J4 OR2M4 OR2L2 OR2K2 OR5L2 OR5K1 OR8G2 OR8B8 OR10A3 OR7C2 OR7C1 OR4D1 OR2T1 OR2H1 OR4C13 OR4C12 OR51V1 OR8D1 OR8D2 OR8B4 OR9G4 OR10A4 OR6C6 OR2Z1 OR10H5 OR14A16 OR2V2 OR13C9 OR13D1 OR8D4 OR5F1 OR5AP2 OR52L1 OR2AG2 OR52B6 OR2AT4 OR10A2 OR6C2 OR6C4 OR8S1 OR6S1 OR6F1 OR2T3 OR10R2 OR2T29 OR6V1 OR2A12 OR2A1 OR1J1 OR1B1 OR13H1 OR56B1 OR52K1 OR52I1 OR51D1 OR52A5 OR51B6 OR2D2 OR52W1 OR56A4 OR56A1 OR10P1 OR10AD1 OR10A7 OR4K14 OR4L1 OR11H6 OR4D2 OR7D4 OR7G1 OR1M1 OR1I1 OR10H4 OR2M5 OR2M3 OR2T12 OR14C36 OR2T34 OR2T10 OR2T4 OR2T11 OR10J5 OR2B11 OR10T2 OR10X1 OR10Z1 OR6K6 OR6N1 OR9A4 OR2Y1 OR9A2 OR2A14 OR6B1 OR2F2 OR13C5 OR13C8 OR13C3 OR13C4 OR13F1 OR1L8 OR1N2 OR1N1 OR52B4 OR52I2 OR10A5 OR51M1 OR51Q1 OR51I1 OR51I2 OR52D1 OR52H1 OR52N4 OR52N5 OR52N2 OR5AK2 OR5B12 OR5AN1 OR4D10 OR4D9 OR10V1 OR6X1 OR6M1 OR10G4 OR10G7 OR8A1 OR6C1 OR6C75 OR6C76 OR6C70 OR4N2 OR4K2 OR4K13 OR4K17 OR4N5 OR11G2 OR11H4 OR5AU1 OR4M2 OR4F6 OR4F15 OR7G2 OR7G3 OR7A10 OR10K2 OR10K1 OR6Y1 OR6K3 OR11L1 OR2L8 OR2AK2 OR2L3 OR2M2 OR2T33 OR2M7 OR2G6 OR2A25 OR13J1 OR13C2 OR1L6 OR5C1 OR1K1 OR2A5 OR2A7 OR51T1 OR51A4 OR51A2 OR2T2 OR2T5 OR14I1 OR5K2 OR2A42 OR1F1 OR2S2 OR13A1 OR2H2 OR2C3 OR2B2 OR4B1 OR5M11 OR2T6 OR51E1 OR8G1 OR10G3 OR10G2 OR4F4 OR4F3 OR4E2 OR1L3 OR1L1 OR1J2

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04742	Taste transduction	SCNN1A SCNN1B SCNN1G ASIC2 CACNA1A CACNA1B GRM4 TAS1R1 TAS1R3 GNB1 GNB3 GNG3 GNG13 TAS2R39 TAS2R40 TAS2R41 TAS2R43 TAS2R31 TAS2R45 TAS2R46 TAS2R19 TAS2R20 TAS2R50 TAS2R60 TAS2R42 TAS2R3 TAS2R4 TAS2R16 TAS2R1 TAS2R9 TAS2R8 TAS2R7 TAS2R13 TAS2R10 TAS2R14 TAS2R5 TAS2R38 TAS2R30 ITPR3 TRPM5 TAS1R2 GNAS ADCY4 ADCY6 ADCY8 PRKACA PRKACB PRKACG KCNB1 GNAT3 PDE1A PLCB2
path:hsa04744	Phototransduction	RHO GRK7 GRK1 RCVRN SAG GNAT2 GNAT3 GNAT1 GNB1 GNGT1 RGS9 PDE6A PDE6B PDE6G GUCY2D GUCY2F GUCA1A GUCA1B GUCA1C SLC24A1 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CNGA1 CNGB1
path:hsa04750	Inflammatory mediator regulation of TRP channels	BDKRB1 BDKRB2 HTR2A HTR2B HTR2C HRH1 P2RY2 GNAQ PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLA2G6 ALOX12 PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 PRKCE PPP1CA PPP1CB PPP1CC TRPA1 TRPV1 ASIC1 ASIC2 ASIC3 ASIC4 ASIC5 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G IL1B IL1R1 IL1RAP MAP2K3 MAP2K6 MAPK11 MAPK12 MAPK13 MAPK14 MAPK8 MAPK10 MAPK9 NGF NTRK1 PLCG1 PLCG2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PRKCD SRC TRPM8 PTGER2 PTGER4 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG TRPV4 F2RL1 CYP2J2 PRKCA PRKCB PRKCG PRKCH PRKCQ IGF1 TRPV2 TRPV3

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04810	Regulation of actin cytoskeleton	F2 F2R CD14 INS EGF FGF1 FGF2 FGF3 FGF4 FGF17 FGF6 FGF7 FGF8 FGF9 FGF10 FGF11 FGF12 FGF13 FGF14 FGF16 FGF5 FGF18 FGF19 FGF20 FGF21 FGF22 FGF23 PDGFA PDGFB PDGFC PDGFD INSRR EGFR FGFR1 FGFR2 FGFR3 FGFR4 PDGFRA PDGFRB FN1 ITGA1 ITGA2 ITGA2B ITGA3 ITGA4 ITGA5 ITGA6 ITGA7 ITGA8 ITGA9 ITGA10 ITGA11 ITGAV ITGAL ITGAM ITGAX ITGAD ITGAE ITGB1 ITGB2 ITGB3 ITGB4 ITGB5 ITGB6 ITGB7 ITGB8 BDKRB1 BDKRB2 CHRM1 CHRM2 CHRM3 CHRM4 CHRM5 GNA12 GNA13 GNG12 FGD1 FGD3 PTK2 BCAR1 CRK CRKL DOCK1 SRC SOS1 SOS2 HRAS KRAS NRAS RRAS RRAS2 MRAS ARHGEF6 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 VAV3 VAV1 VAV2 TIAM1 ARAF BRAF RAF1 MOS MAP2K1 MAP2K2 MAPK1 MAPK3 ARHGEF1 ARHGEF12 ARHGAP35 RHOA RAC1 RAC2 RAC3 CDC42 PAK1 PAK2 PAK3 PAK4 PAK6 PAK7 ARHGEF7 GIT1 ROCK1 ROCK2 MYLK MYLK2 MYLK3 MYLK4 PPP1CA PPP1CB PPP1CC PPP1R12A PPP1R12B PPP1R12C MYL2 MYL5 MYL7 MYL9 MYL10 MYL12B MYL12A MYLPF DIAPH1 DIAPH2 SLC9A1 PIP5K1C PIP5K1A PIP5K1B PIP4K2C PIP4K2A PIP4K2B PIKFYVE LIMK1 LIMK2 DIAPH3 BAIAP2 ENAH WAS WASL WASF2 CYFIP1 CYFIP2 NCKAP1 NCKAP1L ABI2 BRK1 WASF1 ARPC5 ARPC5L ARPC4 ARPC3 ARPC1B ARPC1A ARPC2 ACTB ACTG1 PFN3 PFN1 PFN2 PFN4 PXN EZR RDX MSN TMSB4X TMSB4Y CFL1 CFL2 SSH1 SSH3 SSH2 VCL IQGAP1 IQGAP2 IQGAP3 GSN SCIN ACTN1 ACTN2 ACTN3 ACTN4 APC APC2 ARHGEF4 MYH9 MYH10 MYH14

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04910	Insulin signaling pathway	INS INSR IRS1 IRS2 IRS4 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG PDPK1 AKT1 AKT2 AKT3 GSK3B GYS2 GYS1 PPP1CA PPP1CB PPP1CC PPP1R3A PPP1R3C PPP1R3D PPP1R3B PPP1R3E PPP1R3F PHKG1 PHKG2 PHKB PHKA2 PHKA1 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 PYGL PYGM PYGB PDE3B PRKACA PRKACB PRKACG PRKAR1A PRKAR2A PRKAR2B PRKAR1B LIPE PRKCZ PRKCI SLC2A4 FLOT2 FLOT1 SH2B2 SORBS1 CBLC CBL CBLB CRK CRKL RAPGEF1 RHOQ EXOC7 TRIP10 SREBF1 ACACA ACACB FASN PKLR HK3 HK1 HK2 HKDC1 GCK PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 FOXO1 PPARGC1A G6PC G6PC2 G6PC3 FBP1 FBP2 PCK1 PCK2 MTOR RPTOR RPS6KB1 RPS6KB2 RPS6 EIF4EBP1 EIF4E EIF4E2 EIF4E1B TSC1 TSC2 RHEB BAD SHC1 SHC2 SHC3 SHC4 GRB2 SOS1 SOS2 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 MKNK1 MKNK2 ELK1 SOCS1 SOCS2 SOCS3 SOCS4 PTPN1 PTPRF MAPK8 MAPK10 MAPK9 IKBKB INPPL1 INPP5K
path:hsa04911	Insulin secretion	SLC2A1 SLC2A2 GCK TRPM4 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 ABCC8 KCNJ11 CACNA1C CACNA1D CACNA1F CACNA1S CAMK2A CAMK2D CAMK2B CAMK2G KCNMA1 KCNU1 KCNMB1 KCNMB2 KCNMB3 KCNMB4 KCNN1 KCNN2 KCNN3 KCNN4 GCG GLP1R GIP GPR119 ADCYAP1 ADCYAP1R1 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B PDX1 INS RAPGEF4 RIMS2 PCLO RAB3A CCKAR CHRM3 FFAR1 GNAQ GNA11 PLCB1 PLCB2 PLCB3 PLCB4 ITPR3 RYR2 VAMP2 STX1A SNAP25 PRKCA PRKCB PRKCG

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04912	GnRH signaling pathway	GNRH1 GNRH2 GNRHR GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG ATF4 GNAQ GNA11 PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G PRKCA PRKCB PRKCD CACNA1C CACNA1D CACNA1F CACNA1S MAP3K1 MAP3K2 MAP3K3 MAP3K4 MAP2K3 MAP2K6 MAPK11 MAPK12 MAPK13 MAPK14 MAPK7 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PLD1 PLD2 SRC CDC42 MAP2K4 MAP2K7 MAPK8 MAPK10 MAPK9 JUN LHB CGA FSHB PTK2B MMP2 MMP14 HBEGF EGFR GRB2 SOS1 SOS2 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 ELK1
path:hsa04913	Ovarian steroidogenesis	CGA LHB LHCGR GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F ALOX5 CYP2J2 PTGS2 INS INSR IGF1 IGF1R LDLR SCARB1 STAR CYP11A1 CYP17A1 HSD17B1 HSD17B2 AKR1C3 HSD17B7 HSD3B1 HSD3B2 FSHB FSHR CYP19A1 CYP1B1 CYP1A1 GDF9 BMP15 BMP6 ACOT2
path:hsa04914	Progesterone-mediated oocyte maturation	PGR PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG MAPK8 MAPK10 MAPK9 MAPK11 MAPK12 MAPK13 MAPK14 GNAI1 GNAI3 GNAI2 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG CPEB1 CPEB2 CPEB3 CPEB4 SPDYA SPDYC CDK2 CDK1 CCNA2 CCNA1 MOS HSP90AA1 HSP90AB1 MAP2K1 MAPK1 MAPK3 RPS6KA3 RPS6KA1 RPS6KA2 RPS6KA6 PKMYT1 CCNB1 CCNB2 CCNB3 PLK1 CDC25A CDC25B CDC25C INS IGF1 IGF1R AKT1 AKT2 AKT3 PDE3B KRAS ARAF BRAF RAF1 BUB1 MAD1L1 MAD2L1 MAD2L2 FZR1 ANAPC1 ANAPC2 CDC27 ANAPC4 ANAPC5 CDC16 ANAPC7 CDC23 ANAPC10 ANAPC11 CDC26 ANAPC13

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04915	Estrogen signaling pathway	ESR1 ESR2 HSP90AA1 HSP90AB1 HSP90B1 FKBP4 FKBP5 HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 FOS JUN SP1 GPER1 GNAS SRC MMP2 MMP9 HBEGF ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B EGFR SHC1 SHC2 SHC3 SHC4 GRB2 SOS1 SOS2 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 GABBR1 GABBR2 GNAI1 GNAI3 GNAI2 GNAO1 KCNJ3 KCNJ6 KCNJ9 KCNJ5 OPRM1 PRKCD GNAQ PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 NOS3 GRM1
path:hsa04916	Melanogenesis	POMC ASIP MC1R GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG CREB1 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREBBP EP300 MITF WNT1 WNT2 WNT2B WNT3 WNT3A WNT4 WNT5A WNT5B WNT6 WNT7A WNT7B WNT8A WNT8B WNT9A WNT9B WNT10B WNT10A WNT11 WNT16 FZD1 FZD7 FZD2 FZD3 FZD4 FZD5 FZD8 FZD6 FZD10 FZD9 GNAO1 GNAQ DVL3 DVL2 DVL1 GSK3B CTNNB1 TCF7 TCF7L1 TCF7L2 LEF1 KITLG KIT HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 TYR TYRP1 DCT EDN1 EDNRB GNAI1 GNAI3 GNAI2 PLCB1 PLCB2 PLCB3 PLCB4 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G PRKCA PRKCB PRKCG
path:hsa04917	Prolactin signaling pathway	PRL PRLR JAK2 SHC1 SHC2 SHC3 SHC4 SRC PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 FOXO3 GALT GRB2 SOS1 SOS2 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 FOS MAPK8 MAPK10 MAPK9 MAPK11 MAPK12 MAPK13 MAPK14 CCND1 GSK3B CISH SOCS1 SOCS2 SOCS3 SOCS4 SOCS5 SOCS7 SOCS6 STAT1 STAT3 STAT5A STAT5B TNFSF11 TNFRSF11A NFKB1 RELA ELF5 CSN2 CGA LHB LHCGR CYP17A1 ESR1 ESR2 IRF1 TH SLC2A2 GCK CCND2 INS

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04918	Thyroid hormone synthesis	TSHB CGA TSHR GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG TTF1 TTF2 PAX8 CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B TG HSPA5 HSP90B1 CANX PDIA4 ASGR1 ASGR2 ITPR1 ITPR2 ITPR3 SLC5A5 SLC26A4 TPO ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 IYD DUOXA2 GPX6 GPX7 GPX2 GPX3 GPX1 GPX5 GPX8 GSR LRP2 DUOX2
path:hsa04919	Thyroid hormone signaling pathway	PRKACA PRKACB PRKACG ITGAV ITGB3 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 STAT1 TP53 ESR1 THRB NCOR1 SIN3A HDAC1 HDAC2 HDAC3 THRA RXRA RXRB RXRG KAT2B KAT2A NCOA1 NCOA2 NCOA3 CREBBP EP300 MED4 MED12L MED12 MED13L MED13 MED14 MED16 MED17 MED24 MED27 MED30 MED1 CCND1 GATA4 RCAN1 HIF1A MYC PLN WNT4 CTNNB1 NOTCH1 NOTCH2 NOTCH3 NOTCH4 PLCB1 PLCB2 PLCB3 PLCB4 PLCD1 PLCD3 PLCD4 PLCE1 PLCG1 PLCG2 PLCZ1 PRKCA PRKCB PRKCG SLC9A1 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 SLC16A2 SLC16A10 SRC PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PDPK1 AKT1 AKT2 AKT3 GSK3B MDM2 TSC2 RHEB MTOR TBC1D4 PFKFB2 BAD FOXO1 CASP9 RCAN2 SLC2A1 SLCO1C1 DIO1 DIO2 DIO3 ACTB ACTG1 MYH6 ATP2A2 PFKP BMP4
path:hsa04920	Adipocytokine signaling pathway	TNF TNFRSF1A TRADD TNFRSF1B TRAF2 MTOR MAPK8 MAPK10 MAPK9 CHUK IKBKB IKBKG NFKBIA NFKBIB NFKBIE NFKB1 RELA SOCS3 IRS1 IRS2 IRS4 AKT1 AKT2 AKT3 CD36 ACSL6 ACSL4 ACSL1 ACSL5 ACSL3 ACSBG1 ACSBG2 PRKCQ LEP LEPR JAK2 STAT3 POMC PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 AGRP NPY PPARGC1A PCK1 PCK2 G6PC G6PC2 G6PC3 PTPN11 PPARA RXRA RXRB RXRG ADIPOQ ADIPOR1 ADIPOR2 STK11 CAMKK1 CAMKK2 ACACB CPT1A CPT1B CPT1C SLC2A1 SLC2A4

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04921	Oxytocin signaling pathway	OXT OXTR GNAQ HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PTGS2 MAP2K5 MAPK7 JUN FOS MEF2C CCND1 ELK1 RYR1 RYR2 RYR3 CD38 TRPM2 KCNJ2 KCNJ12 KCNJ4 KCNJ14 PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG EEF2K LOC101930123 EEF2 CACNA1C CACNA1D CACNA1F CACNA1S CACNB1 CACNB2 CACNB3 CACNB4 CACNA2D1 CACNA2D2 CACNA2D3 CACNA2D4 CACNG1 CACNG2 CACNG3 CACNG4 CACNG5 CACNG6 CACNG7 CACNG8 ITPR1 ITPR2 ITPR3 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 NFATC1 NFATC2 NFATC3 NFATC4 RGS2 RCAN1 CAMKK1 CAMKK2 PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 CAMK1D CAMK1G CAMK1 CAMK2A CAMK2D CAMK2B CAMK2G CAMK4 NOS3 GUCY1A2 GUCY1A3 GUCY1B3 NPR1 NPR2 MYLK MYLK2 MYLK3 MYLK4 MYL6B MYL6 MYL9 ACTB ACTG1 RHOA ROCK1 ROCK2 PPP1CA PPP1CB PPP1CC PPP1R12A PPP1R12B PPP1R12C GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG GNAI1 GNAI3 GNAI2 GNAO1 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 SRC KCNJ3 KCNJ6 KCNJ9 KCNJ5 EGFR CDKN1A
path:hsa04922	Glucagon signaling pathway	AKT1 AKT2 AKT3 PDE3B GCG GCGR GNAS ADCY2 PRKACA PRKACB PRKACG CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 SMEK1 SMEK2 PPP4C CRTC2 SIK2 CREBBP EP300 PPARGC1A SIK1 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 CAMK2A CAMK2D CAMK2B CAMK2G FOXO1 SIRT1 PPARA PRMT1 G6PC G6PC2 G6PC3 PCK1 PCK2 CPT1A CPT1B CPT1C GYS2 GYS1 PHKB PHKA2 PHKA1 PHKG1 PHKG2 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 PYGL PYGM PYGB PFKFB1 PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 ACACA ACACB SLC2A1 SLC2A2 GCK PGAM1 PGAM2 PGAM4 PKM PDHA2 PDHA1 PDHB LDHAL6A LDHAL6B LDHA LDHB LDHC FBP1 PFKL

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04923	Regulation of lipolysis in adipocytes	TSHB CGA TSHR ADRB1 ADRB2 ADRB3 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG NPR1 PRKG1 PRKG2 PLIN1 LIPE PNPLA2 ABHD5 MGLL FABP4 AQP7 INS INSR IRS1 IRS2 IRS4 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 PDE3B PLA2G16 PTGS1 PTGS2 ADORA1 PTGER3 NPY NPY1R GNAI1 GNAI3 GNAI2
path:hsa04924	Renin secretion	ADRB1 ADRB2 ADRB3 ADCYAP1 ADCYAP1R1 AQP1 PTGER2 PTGER4 GNAS ADCY5 ADCY6 PRKACA PRKACB PRKACG CREB1 KCNMA1 PDE1A PDE1B PDE1C PDE3A PDE3B REN ADORA1 GNAI1 GNAI3 GNAI2 ORAI1 CLCA1 CLCA2 CLCA4 CACNA1C CACNA1D CACNA1F CACNA1S KCNJ2 EDNRA AGTR1 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 GUCY1A2 GUCY1A3 GUCY1B3 NPR1 PRKG2 CTSB AGT ACE
path:hsa04925	Aldosterone synthesis and secretion	AGTR1 GNAQ GNA11 PLCB1 PLCB2 PLCB3 PLCB4 ITPR1 ITPR2 ITPR3 PRKCA PRKCB PRKCG PRKCE PRKD1 PRKD3 PRKD2 CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B ATF1 DAGLA DAGLB KCNK3 KCNK9 CACNA1G CACNA1H CACNA1I CACNA1C CACNA1D CACNA1F CACNA1S ORAI1 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK1D CAMK1G CAMK1 CAMK2A CAMK2D CAMK2B CAMK2G CAMK4 LIPE LDLR SCARB1 MC2R GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG CYP11B2 STAR NR4A2 NR4A1 CYP11A1 HSD3B1 HSD3B2 CYP21A2 NPR1 PDE2A
path:hsa04930	Type II diabetes mellitus	INS INSR IRS1 IRS2 IRS4 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG SLC2A4 ADIPOQ MAPK1 MAPK3 MTOR PRKCZ SOCS1 SOCS2 SOCS3 SOCS4 IKBKB MAPK8 MAPK10 MAPK9 TNF PRKCD PRKCE PDX1 MAFA SLC2A2 HK3 HK1 HK2 HKDC1 GCK PKM PKLR KCNJ11 ABCC8 CACNA1C CACNA1D CACNA1A CACNA1B CACNA1E CACNA1G

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04931	Insulin resistance	INS RPS6KA3 RPS6KA1 RPS6KA2 RPS6KA6 PPP1CA PPP1CB PPP1CC PPP1R3A PPP1R3C PPP1R3D PPP1R3B PPP1R3E INSR IRS1 PTPN1 PTPN11 PTPRF PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 MTOR RPS6KB1 RPS6KB2 PRKCB IL6 STAT3 SOCS3 GSK3B GYS2 GYS1 TNF TNFRSF1A MAPK8 MAPK10 MAPK9 IKBKB NFKBIA NFKB1 RELA NOS3 PRKCZ TBC1D4 SLC2A4 PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 ACACB CPT1B SLC27A1 SLC27A4 SLC27A2 SLC27A3 SLC27A5 SLC27A6 CD36 PRKCQ PRKCD PPP2R4 SREBF1 IRS2 TRIB3 PYGL PYGM PYGB FOXO1 PCK1 PCK2 G6PC G6PC2 G6PC3 SLC2A2 NR1H3 NR1H2 PPARGC1B MLX MLXIP MLXIPL CPT1A PRKCE PPARGC1A PPARA PDPK1 AKT1 AKT2 AKT3 PTEN MGEA5 OGT SLC2A1 GFPT2 GFPT1 CRTC2 CREB1 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	IL6 IL6R SOCS3 TNF TNFRSF1A NFKB1 RELA INS INSR IRS1 IRS2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 GSK3A GSK3B NR1H3 RXRA SREBF1 MLX MLXIP MLXIPL PKLR LEP LEPR ADIPOQ ADIPOR1 ADIPOR2 PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 PPARA CDC42 RAC1 MAP3K11 MAPK8 MAPK10 MAPK9 ITCH ERN1 TRAF2 MAP3K5 JUN IL1A IL1B IKBKB XBP1 CEBPA CYP2E1 FASLG CXCL8 TGFB1 EIF2AK3 EIF2S1 ATF4 DDIT3 BCL2L11 BAX FAS CASP8 BID CYCS CASP3 CASP7 NDUFV1 NDUFV2 NDUFV3 NDUFA1 NDUFA2 NDUFA3 NDUFA4 NDUFA4L2 NDUFA5 NDUFA6 NDUFA7 NDUFA8 NDUFA9 NDUFA10 NDUFAB1 NDUFA11 NDUFA12 NDUFA13 NDUFB1 NDUFB2 NDUFB3 NDUFB4 NDUFB5 NDUFB6 NDUFB7 NDUFB8 NDUFB9 NDUFB10 NDUFB11 NDUFS1 NDUFS2 NDUFS3 NDUFS4 NDUFS5 NDUFS6 NDUFS7 NDUFS8 NDUFC1 NDUFC2 NDUFC2-KCTD14 SDHA SDHB SDHC SDHD UQCRFS1 CYTB CYC1 UQCRC1 UQCRC2 UQCRH UQCRHL UQCRB UQCRQ UQCR10 UQCR11 COX3 COX1 COX2 COX4I2 COX4I1 COX5A COX5B COX6A1 COX6A2 COX6B1 COX6B2 COX6C COX7A1 COX7A2 COX7A2L COX7B COX7B2 COX7C COX8C COX8A

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04933	AGE-RAGE signaling pathway in diabetic complications	TGFB2 TGFB3 TGFB1 TGFBR1 TGFBR2 SMAD2 SMAD3 SMAD4 CDKN1B FN1 COL1A1 COL1A2 COL3A1 COL4A2 COL4A4 COL4A6 COL4A1 COL4A5 COL4A3 AGTR1 AGER NOX3 NOX1 CYBB PLCD1 PLCD3 PLCD4 PLCB1 PLCB2 PLCB3 PLCB4 PLCG1 PLCG2 PLCE1 PRKCA PRKCB PRKCD PRKCE PRKCZ MAPK1 MAPK3 JUN VEGFA VEGFB VEGFC FIGF CCL2 SERPINE1 SELE VCAM1 ICAM1 MMP2 IL1A IL1B IL6 CXCL8 TNF F3 EDN1 THBD MAPK11 MAPK12 MAPK13 MAPK14 RELA NFKB1 MAPK8 MAPK10 MAPK9 DIAPH1 RAC1 HRAS KRAS NRAS PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 NOS3 FOXO1 BCL2 BAX CASP3 JAK2 STAT3 PIM1 NFATC1 STAT1 STAT5A STAT5B CCND1 CDK4 EGR1 CDC42
path:hsa04940	Type I diabetes mellitus	INS GAD1 GAD2 PTPRN PTPRN2 CPE HSPD1 HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD80 CD86 CD28 IL12A IL12B IL2 IFNG HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E FASLG FAS PRF1 GZMB LTA TNF IL1A IL1B ICA1
path:hsa04950	Maturity onset diabetes of the young	HHEX MNX1 ONECUT1 PDX1 NR5A2 NEUROG3 NKX2-2 NKX6-1 PAX6 PAX4 NEUROD1 RFX6 HES1 HNF1B FOXA2 MAFA HNF4A HNF1A HNF4G FOXA3 PKLR SLC2A2 INS IAPP GCK BHLHA15
path:hsa04960	Aldosterone-regulated sodium reabsorption	HSD11B2 NR3C2 SCNN1A SCNN1B SCNN1G SGK1 KCNJ1 SLC9A3R2 NEDD4L SFN KRAS ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 FXYD4 INS IGF1 INSR IRS1 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG PDPK1 PRKCA PRKCB PRKCG MAPK1 MAPK3
path:hsa04961	Endocrine and other factor-regulated calcium reabsorption	PTH1R GNAS ADCY6 ADCY9 PRKACA PRKACB PRKACG VDR ESR1 TRPV5 KL DNM1 DNM2 DNM3 AP2A2 AP2A1 AP2B1 AP2M1 AP2S1 CLTA CLTB CLTC CLTCL1 RAB11A CALB1 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 KLK2 KLK1 BDKRB2 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG ATP2B1 SLC8A1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04962	Vasopressin-regulated water reabsorption	AVP AVPR2 GNAS ADCY6 ADCY3 ADCY9 PRKACA PRKACB PRKACG ARHGDIB ARHGDIA ARHGDIG CREB1 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 AQP2 RAB11A RAB11B DYNC1H1 DYNC2H1 DYNC1I1 DYNC1I2 DYNC1LI2 DYNC1LI1 DYNC2LI1 DYNLL1 DYNLL2 DCTN1 DCTN2 DCTN4 DCTN5 DCTN6 VAMP2 NSF STX4 RAB5A RAB5B RAB5C AQP4 AQP3
path:hsa04964	Proximal tubule bicarbonate reclamation	SLC9A3 CA4 AQP1 CA2 SLC4A4 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 SLC38A3 GLS2 GLS GLUD2 GLUD1 SLC25A10 MDH1 PCK1 PCK2
path:hsa04966	Collecting duct acid secretion	CA2 ATP6V1A ATP6V1B1 ATP6V1B2 ATP6V1G1 ATP6V1G3 ATP6V1G2 ATP6V1C2 ATP6V1C1 ATP6V1D ATP6V1E2 ATP6V1E1 ATP6V1F ATP6V0E1 ATP6V0E2 TCIRG1 ATP6V0A2 ATP6V0A4 ATP6V0A1 ATP6V0C ATP6V0D1 ATP6V0D2 ATP4A ATP4B SLC4A1 SLC12A7 CLCNKB
path:hsa04970	Salivary secretion	ADRB1 ADRB2 ADRB3 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG VAMP2 ADRA1A ADRA1B ADRA1D CHRM3 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG ITPR1 ITPR2 ITPR3 RYR3 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 NOS1 GUCY1A2 GUCY1A3 GUCY1B3 PRKG1 PRKG2 CD38 BST1 BEST2 AQP5 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 ATP2B1 ATP2B3 ATP2B4 ATP2B2 SLC12A2 KCNN4 KCNMA1 SLC4A2 SLC9A1 TRPV6 MUC5B MUC7 PRH2 PRH1 PRB1 PRB2 DMBT1 CST1 CST2 CST3 CST4 CST5 HTN1 HTN3 STATH LYZ LPO CAMP AMY1A AMY1B AMY1C
path:hsa04971	Gastric acid secretion	CHRM3 GAST CCKBR GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG EZR ITPR1 ITPR2 ITPR3 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G MYLK MYLK2 MYLK3 MYLK4 HRH2 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 SST SSTR2 GNAI1 GNAI3 GNAI2 PRKACA PRKACB PRKACG ATP4A ATP4B KCNE2 KCNQ1 KCNJ1 KCNJ2 KCNJ10 KCNJ15 KCNJ16 CFTR KCNK2 KCNK10 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 SLC9A1 SLC9A4 CA2 SLC26A7 SLC4A2 ACTB

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04972	Pancreatic secretion	CHRM3 CCKAR GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG RAB3D RAB8A RAB11A RAB27B RAP1A RAP1B RHOA RAC1 PRSS3 PRSS2 PRSS1 CTRL CELA2A CELA2B CELA3A CELA3B CPA1 CPA2 CPA3 CPB1 CPB2 PNLIP PNLIPRP1 PNLIPRP2 CEL PLA2G10 PLA2G2D PLA2G2E PLA2G3 PLA2G2F PLA2G12A PLA2G12B PLA2G1B PLA2G5 PLA2G2A PLA2G2C ITPR1 ITPR2 ITPR3 RYR2 CD38 BST1 TPCN2 ATP2A1 ATP2A3 ATP2A2 CLCA1 CLCA2 CLCA4 TRPC1 SLC12A2 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 KCNQ1 KCNMA1 ATP2B1 ATP2B3 ATP2B4 ATP2B2 SLC9A1 SLC4A2 SCTR GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 CFTR SLC26A3 SLC4A4 CA2 CTRB1 AMY2A AMY2B
path:hsa04973	Carbohydrate digestion and absorption	AMY1C AMY2A AMY1A AMY2B AMY1B LCT MGAM SI SLC5A1 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 SLC2A2 HK3 HK1 HK2 HKDC1 SLC37A4 G6PC G6PC2 G6PC3 SLC2A5 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG AKT1 AKT2 AKT3 CACNA1D PRKCB TAS1R2 TAS1R3 GNAT3 PLCB2
path:hsa04974	Protein digestion and absorption	PGA5 PGA4 PGA3 PRSS3 PRSS2 PRSS1 CTRL CELA2A CELA2B CELA3A CELA3B CPA1 CPA2 CPA3 CPB1 CPB2 SLC9A3 SLC15A1 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 KCNN4 KCNK5 KCNE3 KCNQ1 KCNJ13 SLC38A2 MME MEP1A MEP1B ACE2 PRCP DPP4 XPNPEP2 SLC1A1 SLC1A5 SLC8A1 SLC8A2 SLC8A3 SLC6A19 SLC3A2 SLC7A8 SLC16A10 SLC3A1 SLC7A9 SLC7A7 ELN COL1A1 COL1A2 COL2A1 COL3A1 COL4A2 COL4A4 COL4A6 COL4A1 COL4A5 COL4A3 COL5A1 COL5A2 COL5A3 COL11A1 COL24A1 COL27A1 COL11A2 COL6A1 COL6A2 COL6A3 COL6A6 COL6A5 COL7A1 COL9A1 COL9A2 COL9A3 COL10A1 COL12A1 COL13A1 COL14A1 COL15A1 COL17A1 COL18A1 COL21A1 COL22A1 CTRB1 SLC36A1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa04975	Fat digestion and absorption	LIPF PNLIP PNLIPRP1 PNLIPRP2 CEL PLA2G10 PLA2G2D PLA2G2E PLA2G3 PLA2G2F PLA2G12A PLA2G12B PLA2G1B PLA2G5 PLA2G2A PLA2G2C CLPS CD36 GOT2 FABP2 FABP1 AGPAT1 AGPAT2 PPAP2A PPAP2B PPAP2C MOGAT2 MOGAT3 DGAT1 DGAT2 SCARB1 NPC1L1 ABCG5 ABCG8 APOA4 APOB MTTP APOA1 ABCA1 SLC27A4 ACAT2
path:hsa04976	Bile secretion	EPHX1 SLCO1A2 SLCO1B3 SLCO1B1 SLCO1B7 SLC10A1 SLC22A7 SLC22A8 SLC22A1 HMGCR SCARB1 LDLR NCEH1 CYP7A1 NR1H4 RXRA NR0B2 SULT2A1 SLC27A5 BAAT AQP9 AQP8 ABCG5 ABCG8 ABCB11 ABCC2 ABCG2 CA2 SLC4A2 ABCB1 ABCB4 SLC9A1 SLC4A5 ABCC3 ABCC4 SLC51A SLC51B KCNN2 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 SCTR GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG CFTR SLC4A4 AQP4 AQP1 SLC10A2 SLC5A1 SLC9A3 SLC2A1 CYP3A4 UGT2B4
path:hsa04977	Vitamin digestion and absorption	FOLH1 BTD GIF CUBN LMBRD1 MMACHC ABCC1 TCN2 SLC19A3 SLC19A2 SLC52A3 SLC5A6 SLC19A1 SCARB1 PNLIP PLB1 RBP2 AWAT2 LRAT APOA4 APOB APOA1 SLC46A1 SLC23A1
path:hsa04978	Mineral absorption	VDR TRPM6 TRPM7 TRPV6 S100G SLC26A3 SLC26A6 SLC26A9 CLCN2 SLC9A3 SLC5A1 SLC6A19 ATP1A1 ATP1A2 ATP1A3 ATP1A4 ATP1B4 ATP1B1 ATP1B2 ATP1B3 FXYD2 HMOX1 HMOX2 CYBRD1 SLC11A1 SLC11A2 FTH1 FTL SLC40A1 HEPH TF SLC39A4 SLC30A1 SLC31A1 STEAP1 STEAP2 MT1B MT1E MT1F MT1G MT1H MT1M MT1X MT2A MT1A ATOX1 SLC34A2 SLC46A1 ATP2B1 ATP7A SLC8A1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05010	Alzheimer's disease	ADAM10 ADAM17 APP NAE1 APBB1 GAPDH BACE1 BACE2 PSENEN PSEN1 PSEN2 NCSTN APH1A APH1B IDE MME NDUFV1 NDUFV2 NDUFV3 NDUFA1 NDUFA2 NDUFA3 NDUFA4 NDUFA4L2 NDUFA5 NDUFA6 NDUFA7 NDUFA8 NDUFA9 NDUFA10 NDUFAB1 NDUFA11 NDUFA12 NDUFA13 NDUFB1 NDUFB2 NDUFB3 NDUFB4 NDUFB5 NDUFB6 NDUFB7 NDUFB8 NDUFB9 NDUFB10 NDUFB11 NDUFS1 NDUFS2 NDUFS3 NDUFS4 NDUFS5 NDUFS6 NDUFS7 NDUFS8 NDUFC1 NDUFC2 NDUFC2-KCTD14 SDHA SDHB SDHC SDHD UQCRFS1 CYTB CYC1 UQCRC1 UQCRC2 UQCRH UQCRHL UQCRB UQCRQ UQCR10 UQCR11 COX3 COX1 COX2 COX4I2 COX4I1 COX5A COX5B COX6A1 COX6A2 COX6B1 COX6B2 COX6C COX7A1 COX7A2 COX7A2L COX7B COX7B2 COX7C COX8C COX8A ATP5A1 ATP5B ATP5C1 ATP5D ATP5E ATP6 ATP5F1 ATP5G1 ATP5G2 ATP5G3 ATP5H ATP5O ATP5J ATP8 HSD17B10 LPL APOE LRP1 FAS TNFRSF1A FADD CASP8 BID CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 BAD CYCS APAF1 CASP9 CASP3 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D CACNA1C CACNA1D CACNA1F CACNA1S MAPK1 MAPK3 RYR3 ITPR1 ITPR2 ITPR3 ATP2A1 ATP2A3 ATP2A2 ATF6 ERN1 EIF2AK3 CASP12 NOS1 CAPN1 CAPN2 CDK5R1 CDK5 MAPT GSK3B CASP7 SNCA TNF IL1B

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05012	Parkinson's disease	ADORA2A GNAL DRD2 GNAI1 GNAI3 GNAI2 ADCY5 PRKACA PRKACB PRKACG DRD1 UBA1 UBA7 UBB UBE2L3 UBE2L6 UBE2J2 UBE2J1 UBE2G2 UBE2G1 PARK2 SNCA GPR37 SEPT5 SNCAIP UCHL1 TH SLC6A3 SLC18A1 SLC18A2 ND1 ND2 ND3 ND4 ND4L ND5 ND6 NDUFV1 NDUFV2 NDUFV3 NDUFA1 NDUFA2 NDUFA3 NDUFA4 NDUFA4L2 NDUFA5 NDUFA6 NDUFA7 NDUFA8 NDUFA9 NDUFA10 NDUFAB1 NDUFA11 NDUFA12 NDUFA13 NDUFB1 NDUFB2 NDUFB3 NDUFB4 NDUFB5 NDUFB6 NDUFB7 NDUFB8 NDUFB9 NDUFB10 NDUFB11 NDUFS1 NDUFS2 NDUFS3 NDUFS4 NDUFS5 NDUFS6 NDUFS7 NDUFS8 NDUFC1 NDUFC2 NDUFC2-KCTD14 SDHA SDHB SDHC SDHD UQCRFS1 CYTB CYC1 UQCRC1 UQCRC2 UQCRH UQCRHL UQCRB UQCRQ UQCR10 UQCR11 COX3 COX1 COX2 COX4I2 COX4I1 COX5A COX5B COX6A1 COX6A2 COX6B1 COX6B2 COX6C COX7A1 COX7A2 COX7A2L COX7B COX7B2 COX7C COX8C COX8A ATP5A1 ATP5B ATP5C1 ATP5D ATP5E ATP6 ATP5F1 ATP5G1 ATP5G2 ATP5G3 ATP5H ATP5O ATP5J ATP8 LRRK2 PINK1 PARK7 HTRA2 VDAC1 VDAC2 VDAC3 SLC25A4 SLC25A5 SLC25A6 SLC25A31 PPIF CYCS APAF1 CASP9 CASP3
path:hsa05014	Amyotrophic lateral sclerosis (ALS)	TP53 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 CASP1 BID BCL2 BCL2L1 BAX BAD APAF1 CYCS CASP9 CASP3 SOD1 TOMM40 TOMM40L DERL1 CASP12 MAP3K5 TNF TNFRSF1A TNFRSF1B DAXX MAP2K3 MAP2K6 MAPK11 MAPK12 MAPK13 MAPK14 NOS1 CCS CAT PRPH NEFL NEFM NEFH ALS2 RAB5A RAC1 SLC1A2 GRIA1 GRIA2 GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D GPX1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05016	Huntington's disease	CLTA CLTB CLTC CLTCL1 HIP1 AP2A2 AP2A1 AP2B1 AP2M1 AP2S1 HTT IFT57 CASP8 CASP3 GRM5 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 GRIN1 GRIN2B ITPR1 DLG4 TGM2 REST SIN3A RCOR1 HDAC1 HDAC2 POLR2A POLR2B POLR2C POLR2D POLR2E POLR2F POLR2G POLR2H POLR2I POLR2L POLR2J POLR2J3 POLR2J2 POLR2K BDNF DCTN1 DCTN2 DCTN4 HAP1 DNAH1 DNAH3 DNAH2 DNAH7 DNAH5 DNAH6 DNAH9 DNAH11 DNAH12 DNAH14 DNAH8 DNAH10 DNAH17 DNALI1 DNAI1 DNAI2 DNAL1 DNAL4 CREBBP EP300 TAF4B TAF4 TBPL2 TBPL1 TBP SP1 CREB1 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 PPARGC1A PPARG TFAM NRF1 SOD1 SOD2 UCP1 TP53 BAX BBC3 NDUFV1 NDUFV2 NDUFV3 NDUFA1 NDUFA2 NDUFA3 NDUFA4 NDUFA4L2 NDUFA5 NDUFA6 NDUFA7 NDUFA8 NDUFA9 NDUFA10 NDUFAB1 NDUFA11 NDUFA12 NDUFA13 NDUFB1 NDUFB2 NDUFB3 NDUFB4 NDUFB5 NDUFB6 NDUFB7 NDUFB8 NDUFB9 NDUFB10 NDUFB11 NDUFS1 NDUFS2 NDUFS3 NDUFS4 NDUFS5 NDUFS6 NDUFS7 NDUFS8 NDUFC1 NDUFC2 NDUFC2-KCTD14 SDHA SDHB SDHC SDHD UQCRFS1 CYTB CYC1 UQCRC1 UQCRC2 UQCRH UQCRHL UQCRB UQCRQ UQCR10 UQCR11 COX3 COX1 COX2 COX4I2 COX4I1 COX5A COX5B COX6A1 COX6A2 COX6B1 COX6B2 COX6C COX7A1 COX7A2 COX7A2L COX7B COX7B2 COX7C COX8C COX8A ATP5A1 ATP5B ATP5C1 ATP5D ATP5E ATP6 ATP5F1 ATP5G1 ATP5G2 ATP5G3 ATP5H ATP5O ATP5J ATP8 VDAC1 VDAC2 VDAC3 SLC25A4 SLC25A5 SLC25A6 SLC25A31 PPIF CYCS APAF1 CASP9 GPX1
path:hsa05020	Prion diseases	PRNP HSPA5 NCAM1 NCAM2 LAMC1 SOD1 STIP1 PRKACA PRKACB PRKACG BAX FYN CASP12 C5 C6 C7 C8A C8B C8G C9 C1QA C1QB C1QC MAP2K1 MAP2K2 MAPK1 MAPK3 ELK1 EGR1 CCL5 IL1A IL1B IL6 HSPA1A NOTCH1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05030	Cocaine addiction	TH DDC SLC18A1 SLC18A2 MAOB MAOA SLC6A3 DRD1 GNAS GRM2 GRM3 GNAI1 GNAI3 GNAI2 DRD2 ADCY5 PRKACA PRKACB PRKACG GPSM1 RGS9 CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B PDYN BDNF JUN FOSB NFKB1 RELA CDK5 CDK5R1 GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D GRIN3A GRIN3B DLG4 PPP1R1B GRIA2
path:hsa05031	Amphetamine addiction	TH DDC SLC18A1 SLC18A2 MAOB MAOA SLC6A3 CACNA1C CACNA1D GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D GRIN3A GRIN3B GRIA1 GRIA2 GRIA3 GRIA4 DRD1 PRKCA PRKCB PRKCG STX1A CAMK2A CAMK2D CAMK2B CAMK2G CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK4 GNAS ADCY5 PRKACA PRKACB PRKACG PPP1R1B PPP1CA PPP1CB PPP1CC CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B PDYN ARC FOS PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 JUN FOSB SIRT1 HDAC1
path:hsa05032	Morphine addiction	OPRM1 DRD1 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 SLC32A1 KCNJ3 KCNJ6 KCNJ9 KCNJ5 CACNA1A CACNA1B GABRA1 GABRA2 GABRA3 GABRA4 GABRA5 GABRA6 GABRB1 GABRB3 GABRB2 GABRG1 GABRG2 GABRG3 GABRD GABRE GABRQ GABRP GABRR1 GABRR2 GABRR3 GABBR1 GABBR2 GNAI1 GNAI3 GNAI2 GNAO1 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 GNAS ADRBK1 ADRBK2 GRK4 GRK5 GRK6 ARRB1 ARRB2 PRKCA PRKCB PRKCG PRKACA PRKACB PRKACG PDE1A PDE1B PDE1C PDE2A PDE3A PDE3B PDE4A PDE4B PDE4C PDE4D PDE7A PDE7B PDE8B PDE8A PDE10A PDE11A ADORA1
path:hsa05033	Nicotine addiction	SLC32A1 CHRNA6 CACNA1A CACNA1B CHRNA4 CHRNB2 GABRA1 GABRA2 GABRA3 GABRA4 GABRA5 GABRA6 GABRB1 GABRB3 GABRB2 GABRG1 GABRG2 GABRG3 GABRD GABRE GABRQ GABRP GABRR1 GABRR2 GABRR3 SLC17A6 SLC17A8 SLC17A7 CHRNA7 GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D GRIN3A GRIN3B GRIA1 GRIA2 GRIA3 GRIA4

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05034	Alcoholism	TH DDC SLC18A1 SLC18A2 MAOB MAOA SLC6A3 GRIN1 GRIN2A GRIN2B GRIN2C GRIN2D GRIN3A GRIN3B HDAC1 HDAC2 HDAC3 HDAC4 HDAC5 HDAC6 HDAC7 HDAC8 HDAC9 HDAC10 HDAC11 H2AFX HIST2H2AC HIST1H2AH HIST1H2AA HIST3H2A H2AFB3 HIST1H2AE HIST1H2AB H2AFY2 H2AFY HIST2H2AA4 H2AFJ H2AFB1 HIST1H2AM HIST2H2AA3 HIST1H2AG HIST2H2AB H2AFV HIST1H2AD H2AFZ HIST1H2AK HIST1H2AC HIST1H2AI HIST1H2AJ HIST1H2AL H2AFB2 HIST1H2BN HIST1H2BH HIST1H2BM H2BFWT HIST1H2BA HIST2H2BF HIST1H2BO HIST1H2BB HIST1H2BG HIST1H2BL HIST1H2BC HIST1H2BD HIST1H2BE HIST1H2BI HIST2H2BE HIST1H2BJ HIST1H2BK HIST3H2BB HIST1H2BF H2BFM H3F3C H3F3B HIST1H3D HIST1H3C HIST1H3A H3F3A HIST3H3 HIST2H3C HIST2H3A HIST2H3D HIST1H3E HIST1H3I HIST1H3G HIST1H3J HIST1H3H HIST1H3B HIST1H3F HIST4H4 HIST2H4B HIST1H4I HIST1H4D HIST1H4F HIST1H4K HIST1H4J HIST1H4C HIST1H4H HIST1H4B HIST1H4E HIST1H4L HIST2H4A HIST1H4A HIST1H4G HAT1 DRD1 GNAS DRD2 GNAI1 GNAI3 GNAI2 GNAO1 GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 ADORA2A ADORA2B ADCY5 CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B CRH NPY BDNF NTRK2 SHC1 SHC2 SHC3 SHC4 GRB2 SOS1 SOS2 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAPK1 MAPK3 PDYN CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMKK1 CAMKK2 CAMK4 FOSB PPP1R1B PPP1CA PPP1CB PPP1CC PKIA SLC29A1 PRKACA

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05100	Bacterial invasion of epithelial cells	CDH1 CTNNB1 CTNNA3 CTNNA1 CTNNA2 ARHGAP10 MET GAB1 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG CRK CRKL DOCK1 CDC42 RAC1 WAS WASL WASF1 WASF2 ARPC1B ARPC1A ARPC2 ARPC3 ARPC4 ARPC5 ARPC5L ACTB ACTG1 SEPT1 SEPT2 SEPT9 SEPT12 SEPT3 SEPT6 SEPT11 SEPT8 CBLC CBL CBLB SHC1 SHC2 SHC3 SHC4 CD2AP DNM1 DNM2 DNM3 CLTA CLTB CLTC CLTCL1 CAV1 CAV2 CAV3 CTTN HCLS1 SRC PTK2 BCAR1 PXN FN1 ITGA5 ITGB1 ILK MAD2L2 ARHGEF26 RHOG ELMO1 ELMO2 ELMO3 RHOA VCL
path:hsa05110	Vibrio cholerae infection	ATP6V1A ATP6V1B1 ATP6V1B2 ATP6V1C2 ATP6V1C1 ATP6V1D ATP6V1E2 ATP6V1E1 ATP6V1F ATP6V1G1 ATP6V1G3 ATP6V1G2 ATP6V0E1 ATP6V0E2 TCIRG1 ATP6V0A2 ATP6V0A4 ATP6V0A1 ATP6V0D1 ATP6V0D2 ATP6V1H ATP6AP1 ATP6V0C ATP6V0B KDELR1 KDELR2 KDELR3 PDIA4 ERO1L SEC61A1 SEC61A2 SEC61B SEC61G ARF1 GNAS ADCY3 ADCY9 PRKACA PRKACB PRKACG CFTR KCNQ1 SLC12A2 ACTB ACTG1 PLCG1 PLCG2 PRKCA TJP1 TJP2 MUC2
path:hsa05120	Epithelial cell signaling in Helicobacter pylori infection	PTPN11 MET PLCG1 PLCG2 TJP1 F11R JAM2 JAM3 IGSF5 ADAM17 HBEGF EGFR CXCL8 CXCR1 CXCR2 ADAM10 SRC LYN CSK RAC1 CDC42 MAPK11 MAPK12 MAPK13 MAPK14 PAK1 MAP2K4 MAPK8 MAPK10 MAPK9 JUN MAP3K14 CHUK IKBKB IKBKG NFKBIA NFKB1 RELA NOD1 CCL5 PTPRZ1 GIT1 CASP3 ATP6V1A ATP6V1B1 ATP6V1B2 ATP6V1C2 ATP6V1C1 ATP6V1D ATP6V1E2 ATP6V1E1 ATP6V1F ATP6V1G1 ATP6V1G3 ATP6V1G2 ATP6V0E1 ATP6V0E2 TCIRG1 ATP6V0A2 ATP6V0A4 ATP6V0A1 ATP6V0D1 ATP6V0D2 ATP6V1H ATP6AP1 ATP6V0C ATP6V0B CXCL1
path:hsa05130	Pathogenic Escherichia coli infection	TLR5 CD14 TLR4 LY96 TUBA1B TUBA4A TUBA3C TUBA1A TUBA1C TUBA8 TUBA3E TUBA3D TUBAL3 TUBB6 TUBB TUBB1 TUBB2A TUBB3 TUBB4A TUBB8 TUBB2B TUBB4B ARHGEF2 RHOA ROCK1 ROCK2 NCK1 NCK2 WAS WASL ARPC1B ARPC1A ARPC2 ARPC3 ARPC4 ARPC5 ARPC5L ACTB ACTG1 FYN ABL1 CTTN HCLS1 NCL ITGB1 CDC42 OCLN EZR PRKCA CDH1 CTNNB1 YWHAQ KRT18 YWHAZ CLDN1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05131	Shigellosis	ITGA5 ITGB1 CD44 RHOG ELMO1 ELMO2 ELMO3 DOCK1 RAC1 WASF1 WASF2 ARPC5 ARPC5L ARPC4 ARPC3 ARPC1B ARPC1A ARPC2 SRC CTTN HCLS1 ABL1 CRK CRKL CDC42 VCL DIAPH1 ROCK1 ROCK2 WAS WASL ACTB ACTG1 PFN3 PFN1 PFN2 PFN4 ATG5 NOD1 NOD2 RIPK2 MAPK8 MAPK10 MAPK9 MAPK1 MAPK3 MAPK11 MAPK12 MAPK13 MAPK14 IKBKG CHUK IKBKB NFKBIB NFKBIA NFKB1 RELA CXCL8 U2AF1 LOC102724594 U2AF1L4 BTRC FBXW11 MAD2L2 UBE2D2
path:hsa05132	Salmonella infection	CD14 TLR4 LBP MYD88 TLR5 NLRC4 PYCARD CASP1 RAC1 CDC42 RHOG MAPK11 MAPK12 MAPK13 MAPK14 MAPK1 MAPK3 MAPK8 MAPK10 MAPK9 FOS JUN NFKB1 RELA IL18 IL1B IL1A IL6 CXCL8 CCL3 CCL3L1 CCL3L3 CCL4 CCL4L2 CCL4L1 CXCL1 CXCL2 CXCL3 CSF2 WASF1 WASF2 WAS WASL ARPC5 ARPC5L ARPC4 ARPC3 ARPC1B ARPC1A ARPC2 TJP1 ACTB ACTG1 PKN1 PKN3 PKN2 ROCK1 ROCK2 PLEKHM2 KLC3 KLC1 KLC2 KLC4 RAB7A RAB7B RILP DYNC1H1 DYNC2H1 DYNC1I1 DYNC1I2 DYNC1LI2 DYNC1LI1 PFN3 PFN1 PFN2 PFN4 FLNA FLNC FLNB IFNG IFNGR1 IFNGR2 NOS2 MYH9 MYH10 MYH14
path:hsa05133	Pertussis	ITGA5 ITGB1 ITGAM ITGB2 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 RHOA CFL1 CFL2 CASP3 CASP7 CASP1 PYCARD NLRP3 IL1B C1QA C1QB C1QC C1R C1S C2 C4A C4B C3 C5 SERPING1 C4BPA C4BPB LY96 TLR4 CD14 TIRAP MYD88 IRAK4 IRAK1 TRAF6 NFKB1 RELA TICAM2 TICAM1 IRF3 MAPK11 MAPK12 MAPK13 MAPK14 MAPK8 MAPK10 MAPK9 MAPK1 MAPK3 FOS JUN IL6 IL12A IL12B IL23A TNF IL10 IRF1 IRF8 SFTPA1 SFTPA2 GNAI1 GNAI3 GNAI2 CXCL8 CXCL5 CXCL6 NOD1 IL1A NOS2
path:hsa05134	Legionellosis	HSPD1 C3 CR1 ITGAM ITGB2 ARF1 RAB1A RAB1B SEC22B SAR1A SAR1B VCP APAF1 CYCS CASP9 CASP3 CASP8 BNIP3 BCL2L13 NFKB2 NFKBIA NFKB1 RELA HBS1L EEF1A1 EEF1A2 HSF1 HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 EEF1G NAIP CASP7 NLRC4 PYCARD CASP1 IL18 IL1B TLR5 TLR2 CD14 TLR4 MYD88 TNF IL6 IL12A IL12B CXCL8 CXCL1 CXCL2 CXCL3 CLK1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05140	Leishmaniasis	TLR2 TLR4 MYD88 IRAK1 IRAK4 TRAF6 MAP3K7 TAB1 TAB2 NFKBIB NFKBIA NFKB1 RELA IL1A IL1B IL12A IL12B TNF IL4 NOS2 IL10 TGFB1 TGFB2 TGFB3 C3 CR1 ITGAM ITGB2 FCGR1A FCGR2A FCGR2C FCGR3A FCGR3B ITGA4 ITGB1 PTGS2 PRKCB NCF1 NCF2 NCF4 CYBA MAPK1 MAPK3 ELK1 FOS JUN MAPK11 MAPK12 MAPK13 MAPK14 IFNG IFNGR1 IFNGR2 JAK1 JAK2 STAT1 HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 MARCKSL1 PTPN6
path:hsa05142	Chagas disease (American trypanosomiasis)	TLR4 TICAM1 TLR2 TLR6 TLR9 MYD88 IRAK1 IRAK4 TRAF6 MAP2K4 MAPK8 MAPK10 MAPK9 MAPK1 MAPK3 MAPK11 MAPK12 MAPK13 MAPK14 FOS JUN IKBKG CHUK IKBKB NFKBIA NFKB1 RELA IFNB1 CCL5 CCL2 TNF IL12A IL12B IL6 CXCL8 CCL3 CCL3L1 CCL3L3 IL1B PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D PPP2CA PPP2CB IFNG IFNGR1 IFNGR2 TNFRSF1A NOS2 C3 CALR C1QA C1QB C1QC IL10 TGFB1 TGFB2 TGFB3 ACE BDKRB2 GNAQ GNA11 GNA14 GNA15 PLCB1 PLCB2 PLCB3 PLCB4 GNAI1 GNAI3 GNAI2 GNAO1 GNAS GNAL ADCY1 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG AKT1 AKT2 AKT3 TGFBR2 TGFBR1 SMAD2 SMAD3 SERPINE1 FASLG FAS FADD CASP8 CFLAR CD3D CD3E CD3G CD247 IL2
path:hsa05143	African trypanosomiasis	TLR9 MYD88 THOP1 APOA1 HPR HBA1 HBA2 HBB IL10 IL12A IL12B IL18 IFNG TNF IL1B IL6 FASLG FAS VCAM1 SELE ICAM1 LAMA4 IDO1 IDO2 F2RL1 GNAQ PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG APOL1
path:hsa05144	Malaria	SDC1 SDC2 LRP1 CD81 HGF MET ACKR1 GYPA GYPB GYPC HBA1 HBA2 HBB TLR9 MYD88 TLR2 TLR4 IL10 TGFB1 TGFB2 TGFB3 CXCL8 IL6 CCL2 CSF3 IL1B TNF IL12A IL18 KLRB1 KLRK1 KLRC4-KLRK1 IFNG CD36 CR1 PECAM1 VCAM1 THBS1 COMP THBS2 THBS3 THBS4 SELE ITGAL ITGB2 ICAM1 CD40LG CD40 SELP

Continued on next page

continued from previous page

Pathway ID	Description		Gene Sets
path:hsa05145	Toxoplasmosis		IFNG IFNGR1 IFNGR2 JAK1 JAK2 STAT1 CIITA HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 IRGM NOS2 SOCS1 MYD88 IRAK1 IRAK4 TRAF6 MAP3K7 TAB1 TAB2 MAPK1 MAPK3 MAPK8 MAPK10 MAPK9 MAP2K3 MAP2K6 MAPK11 MAPK12 MAPK13 MAPK14 CHUK IKBKB IKBKG NFKBIA NFKBIB NFKB1 RELA IL12A IL12B TNF BCL2 BCL2L1 BIRC2 BIRC3 XIAP BIRC7 BIRC8 LY96 TLR4 HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 TLR2 LAMA1 LAMA2 LAMA3 LAMA5 LAMA4 LAMB1 LAMB2 LAMB3 LAMB4 LAMC1 LAMC2 LAMC3 ITGA6 ITGB1 LDLR TNFRSF1A CASP8 CASP3 CYCS CASP9 GNAI1 GNAI3 GNAI2 GNAO1 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PDPK1 AKT1 AKT2 AKT3 BAD CD40LG CD40 PPIF CCR5 ALOX5 IL10 IL10RA IL10RB TYK2 STAT3 TGFB1 TGFB2 TGFB3
path:hsa05146	Amoebiasis		IL1B IL1R1 IL1R2 NFKB1 RELA HSPB1 MUC2 COL1A1 COL1A2 COL4A2 COL4A4 COL4A6 COL4A1 COL4A5 COL4A3 FN1 LAMA1 LAMA2 LAMA3 LAMA5 LAMA4 LAMB1 LAMB2 LAMB3 LAMB4 LAMC1 LAMC2 LAMC3 CASP3 GNAQ GNA11 GNA14 GNA15 PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG GNAS GNAL ADCY1 PRKACA PRKACB PRKACG RAB5A RAB5B RAB5C RAB7A RAB7B TLR2 TLR4 CD14 IL6 CSF2 CXCL8 TNF IL12A IL12B IFNG COL3A1 PTK2 VCL ACTN1 ACTN2 ACTN3 ACTN4 ARG2 ARG1 NOS2 ITGB2 ITGAM PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG SERPINB1 SERPINB2 SERPINB3 SERPINB4 SERPINB6 SERPINB9 SERPINB10 SERPINB13 CTSG IL10 TGFB1 TGFB2 TGFB3 C8A C8B C8G C9 CXCL1 CD1D
path:hsa05150	Staphylococcus infection	aureus	FGG C3 CFB CFD CFH MBL2 MASP1 MASP2 C1QA C1QB C1QC C1R C1S C2 C4A C4B C5 C3AR1 C5AR1 FCGR1A FCGR2A FCGR2B FCGR2C FCGR3A FCGR3B FCAR FPR3 FPR2 FPR1 PLG CFI SELPLG SELP ICAM1 ITGAL ITGAM ITGB2 DSG1 HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 PTAFR IL10 KRT10

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05152	Tuberculosis	TNF TNFRSF1A TRADD FADD CASP8 CASP10 CASP3 BID BAX CYCS CASP9 APAF1 AKT1 AKT2 AKT3 BAD BCL2 CAMK2A CAMK2D CAMK2B CAMK2G IFNG IFNGR1 IFNGR2 JAK1 JAK2 STAT1 CIITA RFX5 RFXANK RFXAP NFYA NFYB NFYC CREB1 HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD74 CREBBP EP300 IL10 IL10RA IL10RB CTSS CLEC4E FCER1G CLEC7A SRC SYK CARD9 MALT1 BCL10 NOD2 RIPK2 HSPA9 HSPD1 LBP TLR2 TLR1 TLR6 TLR4 CD14 TIRAP MYD88 IRAK4 IRAK1 IRAK2 TRAF6 NFKB1 RELA MAPK11 MAPK12 MAPK13 MAPK14 MAPK1 MAPK3 MAPK8 MAPK10 MAPK9 NOS2 IL6 IL12A IL12B IL18 IL23A IL1A IL1B CEBPB CEBPG TLR9 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 CLEC4M CD209 ARHGEF12 RHOA LSP1 PLK3 KSR1 RAF1 TGFB1 TGFB2 TGFB3 CYP27B1 VDR CAMP C3 CR1 ITGAX ITGB2 ITGAM PLA2R1 MRC1 MRC2 SPHK1 SPHK2 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 PIK3C3 RAB5A RAB5B RAB5C EEA1 RAB7A CTSD TCIRG1 ATP6V0A2 ATP6V0A4 ATP6V0A1 ATP6V0D1 ATP6V0D2 ATP6V1H ATP6AP1 ATP6V0C ATP6V0B LAMP1 LAMP2 PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 CORO1A FCGR1A FCGR2A FCGR2B FCGR2C FCGR3A FCGR3B

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05160	Hepatitis C	LDLR SCARB1 CD81 CLDN4 CLDN3 CLDN7 CLDN19 CLDN16 CLDN14 CLDN15 CLDN17 CLDN20 CLDN11 CLDN18 CLDN22 CLDN5 CLDN10 CLDN8 CLDN6 CLDN2 CLDN1 CLDN9 CLDN23 CLDN25 CLDN24 OCLN OAS1 OAS2 OAS3 RNASEL DDX58 MAVS TRAF3 TBK1 IKBKE IRF3 IRF7 IFNB1 TLR3 TICAM1 TRAF6 RIPK1 CHUK IKBKB IKBKG NFKBIA NFKB1 RELA EIF2AK1 EIF2AK2 EIF2AK3 EIF2AK4 IRF1 EIF2S1 CXCL8 STAT3 IFIT1 IFIT1B EIF3E IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNAR1 IFNAR2 JAK1 TYK2 STAT1 STAT2 IRF9 SOCS3 PPP2R1B PPP2R1A PPP2R2A PPP2R2B PPP2R2C PPP2R2D PPP2CA PPP2CB PIAS1 EGF EGFR GRB2 SOS1 SOS2 HRAS KRAS NRAS BRAF RAF1 ARAF MAPK11 MAPK12 MAPK13 MAPK14 MAPK8 MAPK10 MAPK9 MAPK1 MAPK3 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG PDPK1 AKT1 AKT2 AKT3 GSK3B BAD TP53 CDKN1A TNF TNFRSF1A TRAF2 TRADD PSME3 RXRA PPARA NR1H3
path:hsa05161	Hepatitis B	HSPG2 MAPK8 MAPK10 MAPK9 TGFB1 TGFB2 TGFB3 TGFBR1 CDKN1A MYC EGR2 EGR3 CREBBP EP300 FASLG FAS FADD CASP8 CASP10 DDB1 DDB2 MAP3K1 YWHAZ YWHAB YWHAQ MAP2K4 NFKBIA NFKB1 RELA MMP9 BCL2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 CHUK IKBKB IKBKG BAD TP53 PTEN CASP12 APAF1 LAMTOR5 BIRC5 BAX CYCS CASP9 CASP3 VDAC3 NFATC1 NFATC2 NFATC3 NFATC4 TNF PTK2B SRC GRB2 HRAS KRAS NRAS JUN SMAD4 ATF2 CXCL8 PCNA JAK1 STAT1 STAT2 STAT3 STAT4 STAT5A STAT5B STAT6 ATP6AP1 PRKCA PRKCB PRKCG RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 CREB1 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 ATF4 CREB5 ATF6B FOS ELK1 DDX3X IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNAR1 TLR3 TICAM1 DDX58 IFIH1 MAVS IKBKE TBK1 IRF3 IRF7 TLR4 TICAM2 IFNB1 TLR2 TIRAP MYD88 IL6 CCND1 CDK4 CDK6 RB1 E2F1 E2F2 E2F3 CDKN1B CCNE1 CCNE2 CDK2 CCNA2 CCNA1 SMAD3

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05162	Measles	MSN CLEC4M CD209 TACR1 CD46 SLAMF1 SH2D1A FYN DOK1 PRKCQ IL13 IL4 CD3D CD3E CD3G CD28 IL2 IL2RA IL2RB IL2RG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG AKT1 AKT2 AKT3 GSK3B JAK1 JAK3 STAT3 STAT5A STAT5B DDX58 IFIH1 MAVS IKBKE TBK1 IRF3 TLR7 TLR9 MYD88 IRAK1 IRAK4 IRF7 CHUK TLR2 TLR4 TRAF6 MAP3K7 TAB2 NFKBIB NFKBIA NFKB1 RELA CSNK2A1 CSNK2A2 CSNK2B TNFAIP3 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 IL1A IL1B IL6 IL12A IL12B FCGR2B IFNAR1 IFNAR2 GNB2L1 TYK2 STAT1 STAT2 IRF9 IFNG IFNGR1 IFNGR2 JAK2 OAS1 OAS2 OAS3 ADAR MX1 CCND1 CCND2 CCND3 CCNE1 CCNE2 CDK4 CDK6 CDK2 CDKN1B RAB9A RAB9B HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 EIF3H EIF2AK1 EIF2AK2 EIF2AK3 EIF2AK4 EIF2S1 RCHY1 TP53 TP73 BBC3 FASLG FAS TNFSF10 TNFRSF10A TNFRSF10B TNFRSF10C TNFRSF10D CBLB

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05164	Influenza A	PRSS3 PRSS2 PRSS1 TMPRSS2 TMPRSS4 OAS1 OAS2 OAS3 RNASEL DNAJB1 DNAJC3 HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 EIF2AK1 EIF2AK2 EIF2AK3 EIF2AK4 EIF2S1 MAP2K3 MAP2K6 MAPK11 MAPK12 MAPK13 MAPK14 MAP2K4 MAP2K7 MAPK8 MAPK10 MAPK9 JUN ATF2 DDX58 IFIH1 MAVS NLRX1 IKBKB NFKBIB NFKBIA NFKB1 RELA TRIM25 TLR3 TICAM1 TBK1 IKBKE IRF3 CREBBP EP300 PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG AKT1 AKT2 AKT3 TLR4 TLR7 MYD88 IRAK4 IRF7 IL1A IL1B IL6 IL12A IL12B TNF CXCL8 CCL2 CCL5 CXCL10 ICAM1 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFNB1 IFNAR1 IFNAR2 JAK1 TYK2 STAT1 STAT2 IRF9 SOCS3 MX1 ADAR PML IFNG IFNGR1 IFNGR2 JAK2 CIITA HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 NLRP3 PYCARD CASP1 IL18 IL33 TNFSF10 TNFRSF10A TNFRSF10B TNFRSF10C TNFRSF10D FASLG FAS TNFRSF1A VDAC1 CYCS CASP9 GSK3B KPNA1 KPNA2 DDX39B XPO1 AGFG1 PRKCA RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 RSAD2 FDPS ACTB ACTG1 IVNS1ABP CPSF4 PABPN1 PABPN1L BCL2L2-PABPN1 NXF1 NXF2 NXF2B NXF5 NXF3 NXT1 NXT2 RAE1 HNRNPUL1 NUP98 FURIN TMPRSS13 PLG SLC25A6 PRKCB

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05166	HTLV-I infection	TGFB1 TGFB2 TGFB3 SLC2A1 NRP1 VCAM1 CD3D CD3E CD3G ICAM1 ITGAL ITGB2 TLN1 TLN2 MYC TRRAP KAT5 CCND2 RANBP3 RAN XPO1 HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E CALR PPP3CA PPP3CB PPP3CC PPP3R1 PPP3R2 NFATC1 NFATC2 NFATC3 NFATC4 CANX IL2RB IL2RG JAK1 JAK3 STAT5A STAT5B FDPS HRAS KRAS NRAS RRAS RRAS2 MRAS VDAC1 SLC25A4 SLC25A5 SLC25A6 SLC25A31 VDAC2 VDAC3 TSPO VAC14 RANBP1 MAD2L1 BUB1B BUB3 CDC20 ANAPC1 ANAPC2 CDC27 ANAPC4 ANAPC5 CDC16 ANAPC7 CDC23 ANAPC10 ANAPC11 CDC26 PTTG1 PTTG2 CDKN1A PCNA POLD1 POLD2 POLD3 POLD4 POLE POLE2 POLE3 POLE4 CCND1 CCND3 CDK4 RB1 E2F1 E2F2 E2F3 CDKN2A CDKN2B WNT1 WNT2 WNT2B WNT3 WNT3A WNT4 WNT5A WNT5B WNT6 WNT7A WNT7B WNT8A WNT8B WNT9A WNT9B WNT10B WNT10A WNT11 WNT16 FZD1 FZD7 FZD2 FZD3 FZD4 FZD5 FZD8 FZD6 FZD10 FZD9 DVL3 DVL2 DVL1 GSK3B CTNNB1 APC APC2 DLG1 PDGFA PDGFB PDGFRA PDGFRB PIK3R1 PIK3R5 PIK3R2 PIK3R3 PIK3CA PIK3CD PIK3CB PIK3CG AKT1 AKT2 AKT3 TNF TNFRSF1A MAP3K1 MAP2K4 JUN GPS2 IL1R1 IL1R2 MAP3K3 CHUK IKBKB IKBKG NFKBIA NFKB1 RELA LTBR CD40 TNFRSF13C MAP3K14 RELB NFKB2 IL2 IL2RA IL15 IL15RA IL6 CSF2 LTA BCL2L1 XIAP ZFP36 NFYB HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 SRF ELK4 ELK1 SPI1 ETS1 ETS2 TBPL2 TBPL1 TBP FOS EGR1 EGR2 FOSL1 ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG CREB1 CREM ATF1 ATF2 ATF3 ATF4 XBP1 CRTC1 CRTC2 CRTC3 CREBBP EP300 KAT2B KAT2A TCF3 POLB CDKN2C LCK TERT MYB MYBL1 MYBL2 TGFBR1 TGFBR2 SMAD2 SMAD3 SMAD4 ATM ATR CHEK1 CHEK2 TP53 BAX TP53INP1 MSX1 MSX2 MAPK8 CCNB2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05168	Herpes simplex infection	LTA TNFSF14 TNFRSF14 TRAF1 TRAF2 TRAF3 TRAF5 MAPK8 MAPK10 MAPK9 JUN FOS IKBKB CHUK IKBKG NFKBIA NFKBIB NFKB1 RELA PVRL1 PVRL2 PILRA PTPN11 TLR2 MYD88 TRAF6 TAB1 TAB2 MAP3K7 TLR9 TLR3 TICAM1 IRF3 IRF7 DDX58 IFIH1 MAVS IKBKE TBK1 CCL2 CCL5 TNF IL1B IL6 IL12A IL12B IL15 IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 IFIT1 IFIT1B USP7 EIF2AK1 EIF2AK2 EIF2AK3 EIF2AK4 EIF2S1 PPP1CA PPP1CB PPP1CC OAS1 OAS2 OAS3 RNASEL C3 C5 CFP SKP1 UBE2R2 CDC34 CUL1 SKP2 EEF1D GLTSCR2 MCRS1 POLR2A SRPK1 SRSF1 SRSF9 SRSF2 SRSF8 SRSF3 SRSF4 SRSF5 SRSF6 SRSF7 PML SP100 CREBBP EP300 TP53 DAXX CDK1 CDK2 CSNK2A1 CSNK2A2 CSNK2B HCFC2 HCFC1 TBPL2 TBPL1 TBP TAF9B TAF6 TAF6L TAF5L TAF5 TAF4B TAF4 TAF3 TAF13 TAF10 MED8 GTF2IRD1 GTF2I HMGN1 ALYREF NXF1 NXF2 NXF2B NXF5 NXF3 C1QBP HNRNPK HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD74 TAP1 TAP2 HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E IFNB1 IFNAR1 IFNAR2 JAK1 TYK2 STAT1 STAT2 IRF9 SOCS3 IFNG IFNGR1 IFNGR2 JAK2 FASLG FAS TNFRSF1A FADD CASP8 CASP3 CYCS ARNTL CLOCK PER1 PER3 PER2 POU2F3

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05169	Epstein-Barr virus infection	CR2 POLR2A POLR2B POLR2C POLR2J POLR2J3 POLR2J2 POLR2D POLR2G POLR2I POLR2E POLR2F POLR2H POLR2K POLR2L POLR3A POLR3B POLR1C POLR1D POLR3C POLR3D POLR3E LOC101060521 POLR3F POLR3GL POLR3G POLR3H POLR3K XPO1 RAN NUP214 HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E CD38 SPN CDK1 AKAP8L HSPA8 HSPA1A HSPA2 HSPA1L HSPA1B HSPA6 HSPB1 HSPB2 YWHAZ YWHAB YWHAQ YWHAE YWHAH YWHAG GTF2B SND1 GTF2E1 GTF2E2 HDAC4 HDAC5 SNW1 NCOR2 RBPJL RBPJ SPI1 HDAC1 HDAC2 CREBBP EP300 PTMA FCER2 FGR MYC TP53 TBPL2 TBPL1 TBP CDKN1A MDM2 PSMC2 PSMC1 PSMC4 PSMC6 PSMC3 PSMC5 PSMD2 PSMD1 PSMD3 PSMD12 PSMD11 PSMD6 PSMD7 PSMD13 PSMD4 PSMD14 PSMD8 SHFM1 USP7 PRKACA PRKACB PRKACG CSNK2A1 CSNK2A2 CSNK2B CDKN1B CCNA2 CCNA1 CDK2 RB1 SKP2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 GSK3B TRAF1 TRAF2 TRAF3 TRAF5 MAP3K14 CHUK RELB NFKB2 TRADD RIPK1 IKBKG IKBKB NFKBIB NFKBIA NFKBIE NFKB1 RELA IRAK1 TRAF6 TAB1 TAB2 MAP3K7 MAP2K3 MAP2K6 MAPK11 MAPK12 MAPK13 MAPK14 MAP2K4 MAP2K7 MAPK8 MAPK10 MAPK9 JUN ATF2 JAK3 ENTPD3 ENTPD8 ENTPD1 CD40 CD44 ITGAL CD58 ICAM1 VIM TNFAIP3 BCL2 LYN SYK PLCG1 PLCG2 CD19 NEDD4 DDX58 IFNG TBK1 IRF3 IL10 EIF2AK1 EIF2AK2 EIF2AK3 EIF2AK4 IL10RA IL10RB JAK1 TYK2 STAT3 HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05200	Pathways in cancer	DCC CASP3 CASP9 APPL1 CDH1 CTNNB1 CTNNA3 CTNNA1 CTNNA2 AXIN1 AXIN2 APC APC2 GSK3B TCF7 TCF7L1 TCF7L2 LEF1 BIRC5 MYC CCND1 WNT1 WNT2 WNT2B WNT3 WNT3A WNT4 WNT5A WNT5B WNT6 WNT7A WNT7B WNT8A WNT8B WNT9A WNT9B WNT10B WNT10A WNT11 WNT16 FZD1 FZD7 FZD2 FZD3 FZD4 FZD5 FZD8 FZD6 FZD10 FZD9 DVL3 DVL2 DVL1 F2R F2RL3 LPAR1 LPAR2 LPAR3 LPAR4 LPAR5 LPAR6 AGTR1 GNA12 GNA13 ARHGEF12 ARHGEF11 ARHGEF1 PLEKHG5 RHOA ROCK1 ROCK2 CXCL12 CXCR4 GNAI1 GNAI3 GNAI2 PTGER1 PTGER2 PTGER3 PTGER4 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG GNB1 GNB2 GNB3 GNB4 GNB5 GNG2 GNG3 GNG4 GNG5 GNG7 GNG8 GNG10 GNG11 GNG12 GNG13 GNGT1 GNGT2 COL4A2 COL4A4 COL4A6 COL4A1 COL4A5 COL4A3 LAMA1 LAMA2 LAMA3 LAMA5 LAMA4 LAMB1 LAMB2 LAMB3 LAMB4 LAMC1 LAMC2 LAMC3 FN1 ITGA2 ITGA2B ITGA3 ITGA6 ITGAV ITGB1 PTK2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PTEN NKX3-1 AKT1 AKT2 AKT3 CHUK IKBKB IKBKG NFKBIA NFKB1 NFKB2 RELA PTGS2 NOS2 BCL2 BIRC2 BIRC3 XIAP BIRC7 BIRC8 BCL2L1 TRAF1 TRAF2 TRAF3 TRAF4 TRAF5 TRAF6 MTOR BAD FOXO1 MDM2 TP53 CDKN1B CDKN1A BCR ABL1 CRK CRKL CBLC CBL CBLB STAT5A STAT5B BDKRB1 BDKRB2 EDNRA EDNRB GNAQ GNA11 PLCB1 PLCB2 PLCB3 PLCB4 PRKCA PRKCB PRKCG RASGRP1 RASGRP2 RASGRP3 RASGRP4 JAK1 STAT3 STAT1 VEGFA VEGFB PGF VEGFC FIGF TGFA EGF EGFR ERBB2 PDGFA PDGFB PDGFRA PDGFRB IGF1 IGF1R KITLG KIT FLT3LG FLT3 HGF MET FGF1 FGF2 FGF3 FGF4 FGF17 FGF6 FGF7 FGF8 FGF9 FGF10 FGF11 FGF12 FGF13 FGF14 FGF16 FGF5 FGF18 FGF19 FGF20 FGF21 FGF22 FGF23 FGFR1 FGFR2 FGFR3 GRB2 SOS1 SOS2 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 JUN FOS MMP1 MMP2 MMP9 CXCL8 CDK4 RET CCDC6 NCOA4 NTRK1 TPM3 TPR TFG RASSF1 RASSF5 STK4 DAPK1 DAPK3 DAPK2 PLCG1 PLCG2 RALGDS RALA RALB RALBP1 CDC42 RAC1 RAC2 RAC3 MAPK8 MAPK10 MAPK9 PAX8 PPARG RXRA RXRB RXRG RARB PPARD JUP ZBTB16 PML RARA RUNX1 RUNX1T1 SPI1 CEBPA CSF2RA CSF3R CSF1R IL6 CDKN2A E2F1 E2F2 E2F3 MAX PIAS2 CDKN2B CDK6 CKS1B CKS2 SKP2 CDK2 CCNE1 CCNE2 RB1 MITF TGFB1 TGFB2 TGFB3 TGFBR1 TGFBR2 SMAD2 SMAD3 SMAD4 MECOM CTBP1 CTBP2 HDAC1

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05202	Transcriptional misregulation in cancer	RUNX1 CSF1R MPO CSF2 IL3 RUNX1T1 HDAC1 HDAC2 SIN3A NCOR1 CEBPA SPI1 CD14 ITGAM FCGR1A JUP PML RARA CEBPE BCL2A1 ZBTB16 MYC TCF3 PBX1 WNT16 ETV6 ETV7 ELANE GZMB KMT2A AFF1 CDK9 CCNT1 CCNT2 MLLT1 MLLT3 DOT1L LMO2 PBX3 RUNX2 SMAD1 KLF3 MEF2C JMJD1C HMGA2 KDM6A UTY SUPT3H PROM1 FLT3 BMP2K IGF1R CDKN1B CDK14 MEIS1 SIX1 SIX4 EYA1 CDKN2C HPGD GRIA3 FUT8 TLX3 TLX1 CCR7 LDB1 LYL1 HHEX PTCRA REL TRAF1 BCL2L1 CD86 CD40 BCL6 MAF CCND2 ITGB7 WHSC1 H3F3C H3F3B HIST1H3D HIST1H3C HIST1H3A H3F3A HIST3H3 HIST2H3C HIST2H3A HIST2H3D HIST1H3E HIST1H3I HIST1H3G HIST1H3J HIST1H3H HIST1H3B HIST1H3F PAX5 PAX8 PPARG RXRA RXRB RXRG PRCC TFE3 CDKN1A TMPRSS2 ERG PLAU PLAT MMP3 MMP9 SPINT1 IL1R2 ETV1 ETV4 ETV5 SLC45A3 ELK4 DDX5 MYCN MAX MDM2 PTK2 TP53 BMI1 COMMD3-BMI1 SP1 ZBTB17 NTRK1 NGFR MEN1 EWSR1 FLI1 IGF1 ID2 TGFBR2 IGFBP3 FEV ATF1 ARNT2 ATM GOLPH3 GOLPH3L WT1 PDGFA IL2RB BAIAP3 TSPAN7 MLF1 NR4A3 TAF15 FUS DDIT3 CEBPB IL6 NFKBIZ NFKB1 RELA CXCL8 FOXO1 FLT1 SS18 SSX1 SSX2 SSX2B NUPR1 ASPSCR1 MET DEFA3 DUSP6 HOXA9 HOXA10 HOXA11 BIRC3 PAX3 PAX7 ZEB1 PER2 CCNA1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05203	Viral carcinogenesis	SRC HRAS KRAS NRAS MAPK1 MAPK3 CREB1 ATF2 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 ATF6B DDB1 YWHAZ YWHAB YWHAQ YWHAE YWHAH YWHAG JAK1 STAT3 STAT5A STAT5B CASP3 HPN PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 RELA REL NFKB2 NFKB1 TP53 CCNE1 CCNE2 CDK2 EGR2 EGR3 VDAC3 LTBR SP100 HNRNPK CDKN1A DDX3X EIF2AK2 PRKACA PRKACB PRKACG HIST1H2BN HIST1H2BH HIST1H2BM H2BFWT HIST1H2BA HIST2H2BF HIST1H2BO HIST1H2BB HIST1H2BG HIST1H2BL HIST1H2BC HIST1H2BD HIST1H2BE HIST1H2BI HIST2H2BE HIST1H2BJ HIST1H2BK HIST3H2BB HIST1H2BF H2BFM HIST4H4 HIST2H4B HIST1H4I HIST1H4D HIST1H4F HIST1H4K HIST1H4J HIST1H4C HIST1H4H HIST1H4B HIST1H4E HIST1H4L HIST2H4A HIST1H4A HIST1H4G BAX GRB2 TRADD CDK1 TBPL2 TBPL1 TBP USP7 MDM2 RBPJL RBPJ SNW1 GTF2H1 GTF2H2 GTF2H2C_2 GTF2H2C GTF2H3 GTF2H4 GTF2E1 GTF2E2 SND1 CREBBP EP300 GTF2B CCND1 CCND2 CCND3 CDK6 CCNA2 CCNA1 SKP2 MRPS18B TRAF1 TRAF2 TRAF3 TRAF5 JAK3 LYN SYK ATP6V0D1 ATP6V0D2 UBE3A DLG1 SCRIB PXN BAK1 IRF3 PSMC1 RB1 RBL1 RBL2 CDKN1B UBR4 JUN CHD4 HDAC1 HDAC2 HDAC3 HDAC4 HDAC5 HDAC6 HDAC7 HDAC8 HDAC9 HDAC10 HDAC11 PKM IRF9 DNAJA3 KAT2B KAT2A GTF2A1L GTF2A1 GTF2A2 SRF IKBKG NFKBIA RASA2 CDC42 RAC1 RHOA GSN SCIN ACTN1 ACTN2 ACTN3 ACTN4 CDK4 CDKN2A CDKN2B CHEK1 CDC20 MAD1L1 VAC14 RANBP1 POLB IRF7 IL6ST CCR5 CCR8 CCR3 CCR4 MAPKAPK2 CASP8 BAD PMAIP1 C3 HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05204	Chemical carcinogenesis	CYP1A1 CYP3A4 CYP3A5 CYP3A7 CYP3A7-CYP3A51P CYP3A43 PTGS2 NAT2 NAT1 CYP1A2 SULT1A2 SULT1A1 SULT1A3 SULT1A4 CYP1B1 CYP2C8 CYP2C9 CYP2C18 CYP2C19 EPHX1 ARNT GSTA5 GSTA2 GSTA4 GSTO2 GSTM4 GSTT2 GSTT1 GSTM3 MGST1 MGST3 GSTP1 GSTM1 GSTM5 MGST2 GSTA1 GSTM2 GSTA3 GSTO1 GSTT2B GSTK1 SULT2A1 CHRNA7 CBR1 HSD11B1 CYP2A6 CYP2A13 UGT2A1 UGT2A3 UGT2B17 UGT2B11 UGT2B28 UGT1A6 UGT1A4 UGT1A1 UGT1A3 UGT2B10 UGT1A9 UGT2B7 UGT1A10 UGT1A8 UGT1A5 UGT2B15 UGT1A7 UGT2B4 UGT2A2 CYP2E1 ALDH3B1 ALDH3B2 ALDH1A3 ALDH3A1 ADH1A ADH1B ADH1C ADH7 ADH4 ADH6 ADH5 CCBL2 CCBL1 AKR1C2
path:hsa05205	Proteoglycans in cancer	CD44 SRC CTTN HCLS1 ERBB2 GRB2 HRAS KRAS NRAS RRAS RRAS2 MRAS BRAF RAF1 ARAF MAP2K1 MAP2K2 MAPK1 MAPK3 RAC1 IQGAP1 CDC42 ELK1 ESR1 CCND1 ACTB ACTG1 FLNA FLNC FLNB PAK1 TIAM1 ARHGEF1 RHOA ROCK1 ROCK2 ANK1 ANK2 ANK3 GAB1 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 SLC9A1 PPP1CA PPP1CB PPP1CC PPP1R12A PPP1R12B PPP1R12C ARHGEF12 PLCE1 ITPR1 ITPR2 ITPR3 CAMK2A CAMK2D CAMK2B CAMK2G NANOG DDX5 DROSHA STAT3 MIR21 TWIST1 TWIST2 MIR10A MIR10B HOXD10 DCN IGF1 IGF1R MTOR PDPK1 RPS6KB1 RPS6KB2 EIF4B RPS6 EGFR CAV1 CAV2 CAV3 CD63 CDKN1A CASP3 TGFB1 TLR2 TLR4 PDCD4 TNF IL12B ERBB3 ERBB4 MYC CTNNB1 HIF1A TFAP4 VEGFA KDR MET CBLC CBL CBLB TIMP3 THBS1 MMP2 MMP9 LUM FASLG FAS MDM2 TP53 TGFB2 HPSE HPSE2 SDC1 PLAU PLAUR ITGA2 ITGB1 VTN ITGAV ITGB3 FN1 ITGB5 HGF WNT1 FZD1 FZD7 FZD2 FZD3 FZD4 FZD5 FZD8 FZD6 FZD10 FZD9 ITGA5 SDC2 EZR RDX MSN SDC4 FGF2 FGFR1 FRS2 PTPN11 SOS1 SOS2 PLCG1 PLCG2 PRKCA PRKCB PRKCG NUDT16L1 PXN PTK2 HBEGF GPC1 IGF2 GPC3 WNT2 WNT2B WNT3 WNT3A WNT4 WNT5A WNT5B WNT6 WNT7A WNT7B WNT8A WNT8B WNT9A WNT9B WNT10B WNT10A WNT11 WNT16 PTCH1 SMO HSPG2 CTSL PRKACA PRKACB PRKACG MAPK11 MAPK12 MAPK13 MAPK14 HSPB2 PTPN6 SMAD2 VAV2 COL21A1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05206	MicroRNAs in cancer	MIR17 MIR18A MIR19B2 MIR19A MIR19B1 MIR20A MIR92A1 MIR92A2 PTEN MIR221 MIR222 CDKN1B MIR21 TIMP3 MIRLET7A1 MIRLET7A2 MIRLET7A3 MIRLET7B MIRLET7D MIRLET7E MIRLET7F1 MIRLET7F2 MIRLET7G MIRLET7I MIRLET7C KRAS CDK6 CDC25A CDC25B CDC25C HMGA2 MIR107 MIR15A MIR15B MIR16-1 MIR16-2 CCND1 CCND2 CCNE1 CCNE2 MIR126 CRK CRKL MIR1-1 MIR1-2 PIM1 MET MIR29A MIR29B1 MIR29B2 MIR29C DNMT3A DNMT3B MIR133A1 MIR133A2 MIR133B MCL1 BCL2L2 MIR145 EGFR VEGFA MIR200C MIR141 TGFB2 MIR183 EZR MIR125A MIR125B1 MIR125B2 PDCD4 MIR143 MIR192 MIR194-1 MIR194-2 CREBBP EP300 MIR215 MIR28 MIR30E MIR30D MIR30A MIR30B MIR30C1 MIR30C2 TPM1 SPRY2 BCL2 SERPINB5 MIR155 MIR203A MIR203B MIR205 ZEB1 ZEB2 MIR27A MIR27B ST14 CYP1B1 MIR31 MIR99A MIR100 MTOR RPTOR CYP24A1 ERBB2 ERBB3 ABL1 TP63 MIR375 MIR602 RASSF1 MIR10B HOXD10 DDIT4 BMF MIR224 PAK4 MMP9 MIR363 MIR494 MIR615 MIR625 MIR25 MIR34A MIR96 MIR10A MIR23A MIR23B MIR23C MIR122 SLC7A1 CCNG1 MIR150 MIR223 STMN1 MIR342 MIR423 STAT3 MYC CASP3 HRAS NRAS MIR199B MIR324 MIR483 MIR26A1 MIR26A2 MIR26B MIR152 DNMT1 MIR135B MIR135A2 MIR135A1 APC APC2 IRS1 MAPK7 MIR200A MIR200B HNRNPK RECK MIR124-1 MIR124-3 MIR124-2 MIR137 MIR128-1 MIR128-2 E2F3 MIR7-1 MIR7-2 MIR7-3 IRS2 MIR326 NOTCH1 NOTCH2 NOTCH3 NOTCH4 BMI1 COMMD3-BMI1 MIR146A MMP16 MIR181A2 MIR181A1 MIR181B1 MIR181B2 MIR181C MIR181D ATM SOCS1 MIR373 MIR520A MIR520G MIR520H CD44 MIR103A1 MIR103B2 MIR103A2 MIR103B1 DICER1 FZD3 ITGA5 RDX RHOA MIR193B PLAU BAK1 ZFPM2 ITGB3 UBE2I MIR335 TNC TNN TNR TNXB MIR206 MIR451A ABCB1 MIR345 ABCC1 MIR214 BRCA1 MIR106B CDKN1A MIR210 THBS1 IGF2BP1 TRIM71 MIR9-1 MIR9-2 MIR9-3 NFKB1 MIR199A1 MIR199A2 IKBKB PTGS2 PRKCA PRKCB PRKCG GRB2 FGFR3 RPS6KA5 TP53 MDM2 MDM4 PDGFRA PLCG1 PLCG2 MIR101-1 MIR101-2 RAF1 MIR195 MAP2K1 MAP2K2 SOS1 SOS2 SHC1 PDGFA MIR129-1 MIR129-2 SHC4 PDGFB PDGFRB CDKN2A E2F1 E2F2 MIR32 BCL2L11 MARCKS BMPR2 MIR449A EZH2 GLS2 GLS SIRT1 WNT3 WNT3A MIR330 MIR331 CDCA5 KIF23 MIR520C ROCK1 PRKCE SLC45A3 VIM EFNA3 FOXP1 HDAC1 HMOX1 MIR34B MIR34C PIK3CA PIK3R2 MAPK1 FSCN1 SOX4 HDAC4

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05210	Colorectal cancer	GSK3B AXIN1 AXIN2 CTNNB1 APC APC2 TCF7 TCF7L1 TCF7L2 LEF1 BIRC5 MYC CCND1 KRAS PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 BAD CASP9 ARAF BRAF RAF1 MAP2K1 MAPK1 MAPK3 JUN FOS RALGDS RAC1 RAC2 RAC3 RHOA MAPK8 MAPK10 MAPK9 DCC CASP3 APPL1 TGFB1 TGFB2 TGFB3 TGFBR1 TGFBR2 SMAD2 SMAD3 SMAD4 MLH1 MSH2 MSH3 MSH6 BAX BCL2 CYCS TP53
path:hsa05211	Renal cell carcinoma	HIF1A EPAS1 EGLN1 EGLN3 EGLN2 VHL TCEB1 TCEB2 RBX1 CUL2 ARNT ARNT2 CREBBP EP300 SLC2A1 VEGFA TGFB1 TGFB2 TGFB3 PDGFB TGFA HGF MET GAB1 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 CRK CRKL RAPGEF1 RAP1A RAP1B PTPN11 GRB2 SOS1 SOS2 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 JUN RAC1 CDC42 PAK1 PAK2 PAK3 PAK4 PAK6 PAK7 FH FLCN ETS1
path:hsa05212	Pancreatic cancer	KRAS PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 ARHGEF6 RAC1 RAC2 RAC3 NFKB1 RELA AKT1 AKT2 AKT3 CHUK IKBKB IKBKG BAD BCL2L1 CASP9 ARAF BRAF RAF1 MAP2K1 MAPK1 MAPK3 MAPK8 MAPK10 MAPK9 RALGDS RALA RALB RALBP1 CDC42 TGFA EGF EGFR ERBB2 JAK1 STAT3 STAT1 VEGFA CDKN2A CDK4 CDK6 CCND1 RB1 E2F1 E2F2 E2F3 TP53 TGFB1 TGFB2 TGFB3 TGFBR1 TGFBR2 SMAD2 SMAD3 SMAD4 BRCA2 RAD51 PLD1
path:hsa05213	Endometrial cancer	EGF EGFR PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PTEN PDPK1 ILK AKT1 AKT2 AKT3 CASP9 BAD FOXO3 GRB2 SOS1 SOS2 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 ELK1 MLH1 CDH1 CTNNB1 CTNNA3 CTNNA1 CTNNA2 AXIN1 AXIN2 APC APC2 GSK3B TCF7 TCF7L1 TCF7L2 LEF1 MYC CCND1 TP53 ERBB2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05214	Glioma	EGF TGFA EGFR PDGFA PDGFB PDGFRA PDGFRB IGF1 IGF1R PLCG1 PLCG2 CALML3 CALM2 CALM3 CALM1 CALML6 CALML5 CAMK2A CAMK2D CAMK2B CAMK2G PRKCA PRKCB PRKCG SHC1 SHC2 SHC3 SHC4 GRB2 SOS1 SOS2 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 MTOR PTEN CDKN2A MDM2 TP53 CDKN1A CCND1 CDK4 CDK6 RB1 E2F1 E2F2 E2F3
path:hsa05215	Prostate cancer	CDKN1B CDK2 CCNE1 CCNE2 RB1 E2F1 E2F2 E2F3 INS PDGFA PDGFB PDGFC PDGFD EGF TGFA IGF1 INSRR PDGFRA PDGFRB FGFR1 FGFR2 EGFR ERBB2 IGF1R PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PTEN PDPK1 NKX3-1 AKT1 AKT2 AKT3 CASP9 BAD FOXO1 CDKN1A MDM2 TP53 GSK3B CREB1 ATF4 CREB3 CREB3L1 CREB3L2 CREB3L3 CREB3L4 CREB5 CTNNB1 CREBBP EP300 TCF7 TCF7L1 TCF7L2 LEF1 CCND1 CHUK IKBKB IKBKG NFKBIA NFKB1 RELA BCL2 MTOR GRB2 SOS1 SOS2 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 SRD5A2 AR HSP90AA1 HSP90AB1 HSP90B1 KLK3 GSTP1
path:hsa05216	Thyroid cancer	RET CCDC6 NCOA4 NTRK1 TPM3 TPR TFG HRAS NRAS KRAS BRAF MAP2K1 MAP2K2 MAPK1 MAPK3 PAX8 PPARG RXRA RXRB RXRG TP53 CDH1 CTNNB1 TCF7 TCF7L1 TCF7L2 LEF1 MYC CCND1
path:hsa05217	Basal cell carcinoma	TP53 SHH PTCH1 SMO STK36 SUFU GLI1 GLI2 GLI3 BMP2 BMP4 HHIP PTCH2 WNT1 WNT2 WNT2B WNT3 WNT3A WNT4 WNT5A WNT5B WNT6 WNT7A WNT7B WNT8A WNT8B WNT9A WNT9B WNT10B WNT10A WNT11 WNT16 FZD1 FZD7 FZD2 FZD3 FZD4 FZD5 FZD8 FZD6 FZD10 FZD9 DVL3 DVL2 DVL1 GSK3B AXIN1 AXIN2 APC APC2 CTNNB1 TCF7 TCF7L1 TCF7L2 LEF1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05218	Melanoma	FGF1 FGF2 FGF3 FGF4 FGF17 FGF6 FGF7 FGF8 FGF9 FGF10 FGF11 FGF12 FGF13 FGF14 FGF16 FGF5 FGF18 FGF19 FGF20 FGF21 FGF22 FGF23 HGF IGF1 PDGFA PDGFB PDGFC PDGFD EGF FGFR1 MET IGF1R PDGFRA PDGFRB EGFR HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 CDK4 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 BAD PTEN CDKN2A MDM2 TP53 CDKN1A CCND1 CDK6 RB1 E2F1 E2F2 E2F3 CDH1 MITF
path:hsa05219	Bladder cancer	FGFR3 HRAS KRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 RPS6KA5 MYC RASSF1 DAPK1 DAPK3 DAPK2 CDKN2A MDM2 TP53 CDKN1A CCND1 CDK4 RB1 E2F1 E2F2 E2F3 THBS1 HBEGF MMP2 MMP9 UPK3A SRC EGF ERBB2 EGFR TYMP VEGFA MMP1 CXCL8 CDH1
path:hsa05220	Chronic myeloid leukemia	BCR ABL1 CRK CRKL CBLC CBL CBLB PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 BAD BCL2L1 CHUK IKBKB IKBKG NFKBIA NFKB1 RELA MDM2 TP53 CDKN1B GRB2 GAB2 PTPN11 SOS1 SOS2 HRAS KRAS NRAS RAF1 ARAF BRAF MAP2K1 MAP2K2 MAPK1 MAPK3 SHC1 SHC2 SHC3 SHC4 MYC STAT5A STAT5B CDKN2A CDKN1A CCND1 CDK4 CDK6 RB1 E2F1 E2F2 E2F3 TGFB1 TGFB2 TGFB3 TGFBR1 TGFBR2 SMAD4 MECOM RUNX1 CTBP1 CTBP2 HDAC1 HDAC2 SMAD3
path:hsa05221	Acute myeloid leukemia	KIT FLT3 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 CHUK IKBKB IKBKG NFKB1 RELA BAD MTOR EIF4EBP1 RPS6KB1 RPS6KB2 GRB2 SOS1 SOS2 HRAS NRAS KRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 STAT3 STAT5A STAT5B PIM1 PIM2 RUNX1 RUNX1T1 PML RARA ZBTB16 CEBPA SPI1 JUP TCF7 TCF7L1 TCF7L2 LEF1 CCND1 MYC PPARD CCNA1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05222	Small cell lung cancer	FHIT RARB RXRA RXRB RXRG TP53 BCL2 CYCS APAF1 CASP9 MYC MAX PIAS2 CDKN2B CDK4 CDK6 CCND1 CKS1B CKS2 SKP2 CDK2 CCNE1 CCNE2 RB1 E2F1 E2F2 E2F3 COL4A2 COL4A4 COL4A6 COL4A1 COL4A5 COL4A3 LAMA1 LAMA2 LAMA3 LAMA5 LAMA4 LAMB1 LAMB2 LAMB3 LAMB4 LAMC1 LAMC2 LAMC3 FN1 ITGA2 ITGA2B ITGA3 ITGA6 ITGAV ITGB1 PTK2 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PTEN AKT1 AKT2 AKT3 CHUK IKBKB IKBKG NFKBIA NFKB1 RELA BIRC2 BIRC3 XIAP BIRC7 BIRC8 BCL2L1 TRAF1 TRAF2 TRAF3 TRAF4 TRAF5 TRAF6 PTGS2 NOS2 CDKN1B
path:hsa05223	Non-small cell lung cancer	FHIT RARB RXRA RXRB RXRG CDKN2A CDK4 CDK6 CCND1 RB1 E2F1 E2F2 E2F3 KRAS RASSF1 RASSF5 STK4 PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PDPK1 AKT1 AKT2 AKT3 BAD CASP9 FOXO3 EGF TGFA EGFR ERBB2 GRB2 SOS1 SOS2 HRAS NRAS ARAF BRAF RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 PLCG1 PLCG2 PRKCA PRKCB PRKCG TP53 EML4 ALK
path:hsa05230	Central carbon metabolism in cancer	SLC2A1 SLC2A2 GCK PKM PDHA2 PDHA1 PDHB PDK1 TP53 SLC1A5 SIRT3 SLC16A3 SIRT6 MYC HIF1A KIT MET RET EGFR ERBB2 NTRK1 NTRK3 PDGFRA PDGFRB FGFR1 FGFR2 FGFR3 FLT3 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 GLS2 GLS G6PD HK3 HK1 HK2 HKDC1 PFKM PFKP PFKL PGAM1 PGAM2 PGAM4 C12orf5 PTEN PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 AKT1 AKT2 AKT3 MTOR SLC7A5 IDH1 LDHA SCO2

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05231	Choline metabolism in cancer	EGF PDGFA PDGFB PDGFC PDGFD EGFR PDGFRA PDGFRB GRB2 SOS1 SOS2 HRAS KRAS NRAS RAF1 MAP2K1 MAP2K2 MAPK1 MAPK3 RALGDS MAPK8 MAPK10 MAPK9 PLA2G4E PLA2G4A JMJD7-PLA2G4B PLA2G4B PLA2G4C PLA2G4D PLA2G4F PIK3CA PIK3CD PIK3CB PIK3CG PIK3R1 PIK3R5 PIK3R2 PIK3R3 PDPK1 AKT1 AKT2 AKT3 TSC1 TSC2 RHEB MTOR RPS6KB1 RPS6KB2 EIF4EBP1 PIP5K1C PIP5K1A PIP5K1B WAS WASL RAC1 RAC2 RAC3 WASF1 WASF2 WASF3 SP1 PLD1 PLD2 SLC5A7 SLC44A1 SLC44A4 SLC44A5 SLC44A2 SLC44A3 SLC22A1 SLC22A2 SLC22A3 SLC22A5 SLC22A4 CHKA CHKB HIF1A JUN FOS PCYT1B PCYT1A CHPT1 PLCG1 PPAP2A PPAP2B PPAP2C DGKZ DGKD DGKI DGKA DGKE DGKB DGKH DGKG DGKQ DGKK PRKCA PRKCB PRKCG LYPLA1 GPCPD1
path:hsa05310	Asthma	HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 IL4 CD40LG CD40 FCER1A MS4A2 FCER1G IL9 IL10 IL13 IL5 CCL11 TNF IL3 PRG2 RNASE3 EPX
path:hsa05320	Autoimmune thyroid disease	IFNA1 IFNA2 IFNA4 IFNA5 IFNA6 IFNA7 IFNA8 IFNA10 IFNA13 IFNA14 IFNA16 IFNA17 IFNA21 CTLA4 TPO TG HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD80 CD86 CD28 IL2 HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E FASLG FAS PRF1 GZMB CD40LG CD40 IL4 IL5 IL10 TSHR CGA TSHB
path:hsa05321	Inflammatory bowel disease (IBD)	TLR2 TLR4 TLR5 NFKB1 RELA NOD2 HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 IFNG IFNGR1 IFNGR2 STAT1 TBX21 IL4 IL5 IL12A IL12B IL12RB2 IL12RB1 STAT4 IL2 IL18 IL18R1 IL18RAP JUN TNF IL6 IL1A IL1B TGFB1 TGFB2 TGFB3 SMAD2 SMAD3 STAT3 IL21 IL21R IL23A IL23R RORC RORA FOXP3 IL17A IL17F IL22 IL4R IL2RG STAT6 IL10 IL13 MAF NFATC1 GATA3

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05322	Systemic lupus erythematosus	C1QA C1QB C1QC C2 C4A C4B H2AFX HIST2H2AC HIST1H2AH HIST1H2AA HIST3H2A H2AFB3 HIST1H2AE HIST1H2AB H2AFY2 H2AFY HIST2H2AA4 H2AFJ H2AFB1 HIST1H2AM HIST2H2AA3 HIST1H2AG HIST2H2AB H2AFV HIST1H2AD H2AFZ HIST1H2AK HIST1H2AC HIST1H2AI HIST1H2AJ HIST1H2AL H2AFB2 HIST1H2BN HIST1H2BH HIST1H2BM H2BFWT HIST1H2BA HIST2H2BF HIST1H2BO HIST1H2BB HIST1H2BG HIST1H2BL HIST1H2BC HIST1H2BD HIST1H2BE HIST1H2BI HIST2H2BE HIST1H2BJ HIST1H2BK HIST3H2BB HIST1H2BF H2BFM H3F3C H3F3B HIST1H3D HIST1H3C HIST1H3A H3F3A HIST3H3 HIST2H3C HIST2H3A HIST2H3D HIST1H3E HIST1H3I HIST1H3G HIST1H3J HIST1H3H HIST1H3B HIST1H3F HIST4H4 HIST2H4B HIST1H4I HIST1H4D HIST1H4F HIST1H4K HIST1H4J HIST1H4C HIST1H4H HIST1H4B HIST1H4E HIST1H4L HIST2H4A HIST1H4A HIST1H4G SNRPB SNRPD1 SNRPD3 GRIN2A GRIN2B TRIM21 TROVE2 SSB ACTN1 ACTN2 ACTN3 ACTN4 HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD80 CD86 CD28 CD40LG CD40 TNF IFNG IL10 C1R C1S C3 C5 C6 C7 C8A C8B C8G C9 CTSG ELANE FCGR1A FCGR2A FCGR3A FCGR3B
path:hsa05323	Rheumatoid arthritis	CD80 CD86 CD28 CTLA4 HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 ITGAL ITGB2 ICAM1 IL15 TNFSF13 TNFSF13B LTB TNF IL1A IL1B IL6 IL11 IL18 TLR2 TLR4 JUN FOS TGFB1 TGFB2 TGFB3 IL23A IL17A CSF1 TNFSF11 TNFRSF11A ATP6V1A ATP6V1B1 ATP6V1B2 ATP6V1C2 ATP6V1C1 ATP6V1D ATP6V1E2 ATP6V1E1 ATP6V1F ATP6V1G1 ATP6V1G3 ATP6V1G2 ATP6V0E1 ATP6V0E2 TCIRG1 ATP6V0A2 ATP6V0A4 ATP6V0A1 ATP6V0D1 ATP6V0D2 ATP6V1H ATP6AP1 ATP6V0C ATP6V0B CTSK ACP5 MMP1 MMP3 CTSL CSF2 CCL5 CCL2 CCL3 CCL3L1 CCL3L3 CCL20 CXCL5 CXCL6 CXCL8 CXCL12 VEGFA FLT1 ANGPT1 TEK IFNG CXCL1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05330	Allograft rejection	HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E CD80 CD86 CD28 IL12A IL12B FASLG FAS PRF1 GZMB HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 IL2 CD40LG CD40 IFNG TNF IL4 IL5 IL10
path:hsa05332	Graft-versus-host disease	IL6 IL1A IL1B TNF HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 CD80 CD86 CD28 IL2 HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E FASLG FAS PRF1 GZMB IFNG KLRD1 KIR2DL1 KIR2DL2 KIR2DL3 KIR3DL1 KIR3DL2 KLRC1 KIR2DL5A
path:hsa05340	Primary immunodeficiency	ADA IL7R IL2RG DCLRE1C RAG1 RAG2 CD3D CD3E PTPRC CD4 CD8A CD8B AIRE TAP1 TAP2 LCK ZAP70 RFX5 RFXAP RFXANK CIITA ORAI1 CD79A BLNK BTK IKBKG CD40 CD40LG UNG AICDA ICOS TNFRSF13C CD19 TNFRSF13B IGLL1 JAK3
path:hsa05410	Hypertrophic cardiomyopathy (HCM)	ITGA1 ITGA2 ITGA2B ITGA3 ITGA4 ITGA5 ITGA6 ITGA7 ITGA8 ITGA9 ITGA10 ITGA11 ITGAV ITGB1 ITGB3 ITGB4 ITGB5 ITGB6 ITGB7 ITGB8 SGCD SGCG SGCA SGCB DAG1 DES DMD ACTB ACTG1 TTN TNNT2 TNNC1 TNNI3 ACTC1 TPM1 TPM2 TPM3 TPM4 MYBPC3 MYL3 MYL2 EMD LMNA CACNA1C CACNA1D CACNA1F CACNA1S CACNB1 CACNB2 CACNB3 CACNB4 CACNA2D1 CACNA2D2 CACNA2D3 CACNA2D4 CACNG1 CACNG2 CACNG3 CACNG4 CACNG5 CACNG6 CACNG7 CACNG8 RYR2 PRKAA1 PRKAA2 PRKAB1 PRKAB2 PRKAG1 PRKAG3 PRKAG2 ACE IGF1 TGFB1 TGFB2 TGFB3 TNF IL6 LAMA2 MYH6 MYH7 ATP2A2 SLC8A1

Continued on next page

continued from previous page

Pathway ID	Description	Gene Sets
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	ITGA1 ITGA2 ITGA2B ITGA3 ITGA4 ITGA5 ITGA6 ITGA7 ITGA8 ITGA9 ITGA10 ITGA11 ITGAV ITGB1 ITGB3 ITGB4 ITGB5 ITGB6 ITGB7 ITGB8 SGCD SGCG SGCA SGCB DAG1 DMD DES ACTB ACTG1 EMD LMNA CTNNB1 TCF7 TCF7L1 TCF7L2 LEF1 CACNA1C CACNA1D CACNA1F CACNA1S CACNB1 CACNB2 CACNB3 CACNB4 CACNA2D1 CACNA2D2 CACNA2D3 CACNA2D4 CACNG1 CACNG2 CACNG3 CACNG4 CACNG5 CACNG6 CACNG7 CACNG8 RYR2 CDH2 JUP CTNNA3 CTNNA1 CTNNA2 ACTN1 ACTN2 ACTN3 ACTN4 DSC2 PKP2 DSP DSG2 GJA1 LAMA2 ATP2A2 SLC8A1
path:hsa05414	Dilated cardiomyopathy	ITGA1 ITGA2 ITGA2B ITGA3 ITGA4 ITGA5 ITGA6 ITGA7 ITGA8 ITGA9 ITGA10 ITGA11 ITGAV ITGB1 ITGB3 ITGB4 ITGB5 ITGB6 ITGB7 ITGB8 SGCD SGCG SGCA SGCB DAG1 DES DMD ACTB ACTG1 TTN TNNT2 TNNC1 TNNI3 ACTC1 TPM1 TPM2 TPM3 TPM4 MYBPC3 MYL3 MYL2 EMD LMNA ADRB1 GNAS ADCY1 ADCY2 ADCY3 ADCY4 ADCY5 ADCY6 ADCY7 ADCY8 ADCY9 PRKACA PRKACB PRKACG CACNA1C CACNA1D CACNA1F CACNA1S CACNB1 CACNB2 CACNB3 CACNB4 CACNA2D1 CACNA2D2 CACNA2D3 CACNA2D4 CACNG1 CACNG2 CACNG3 CACNG4 CACNG5 CACNG6 CACNG7 CACNG8 RYR2 PLN IGF1 TGFB1 TGFB2 TGFB3 TNF LAMA2 MYH6 MYH7 ATP2A2 SLC8A1
path:hsa05416	Viral myocarditis	CXADR CD55 FYN CAV1 ABL1 ABL2 RAC1 RAC2 RAC3 SGCD SGCG SGCA SGCB DAG1 DMD ACTB ACTG1 EIF4G3 EIF4G1 EIF4G2 CCND1 CASP8 BID CYCS CASP9 CASP3 MYH7 MYH6 CD40LG CD40 HLA-DMA HLA-DMB HLA-DOA HLA-DOB HLA-DPA1 HLA-DPB1 HLA-DQA1 HLA-DQA2 HLA-DQB1 HLA-DRA HLA-DRB1 HLA-DRB3 HLA-DRB4 HLA-DRB5 HLA-A HLA-B HLA-C HLA-F HLA-G HLA-E CD80 CD86 CD28 PRF1 ITGAL ITGB2 ICAM1 LAMA2