# Chapter 1
# Functional Genomics, Proteomics, Metabolomics and Bioinformatics for Systems Biology

**Stéphane Ballereau, Enrico Glaab, Alexei Kolodkin, Amphun Chaiboonchoe, Maria Biryukov, Nikos Vlassis, Hassan Ahmed, Johann Pellet, Nitin Baliga, Leroy Hood, Reinhard Schneider, Rudi Balling and Charles Auffray**

**Abstract** This chapter introduces Systems Biology, its context, aims, concepts and strategies, then describes approaches used in genomics, epigenomics, transcriptomics, proteomics, metabolomics and lipidomics, and how recent technological advances in these fields have moved the bottleneck from data production to data analysis. Methods for clustering, feature selection, prediction analysis, text mining and pathway analysis used to analyse and integrate the data produced are then presented.

**Keywords** Emergence · Holistic · Bottom-up · Top-down · Middle-out · Interactions · Data integration · Mathematical model · Functional genomics · High-throughput · Epigenomics · Transcriptomics · Proteomics · Metabolomics · Next generation sequencing · Mass spectrometry · Bioinformatics · Knowledge management · Ontology · Pathway · Network · High-dimensionality · Curse of dimensionality · Clustering · Feature selection · Prediction analysis · Text-mining

S. Ballereau (✉) · A. Chaiboonchoe · H. Ahmed · J. Pellet · C. Auffray
European Institute for Systems Biology & Medicine, CNRS-UCBL—Université de Lyon,
50 Avenue Tony Garnier, 69007 Lyon, France
e-mail: sballereau@eisbm.org, achaiboonchoe@eisbm.org, hahmed@eisbm.org,
jpellet@eisbm.org, cauffray@eisbm.org

E. Glaab · A. Kolodkin · M. Biryukov · N. Vlassis · R. Schneider · R. Balling
Luxembourg Centre for Systems Biomedicine, University of Luxembourg,
7 Avenue des Hauts-Fourneaux, 4362 Esch-sur-Alzette, Luxembourg
e-mail: enrico.glaab@uni.lu, alexey.kolodkin@uni.lu, maria.biryukov@uni.lu,
nikos.vlassis@uni.lu, reinhard.schneider@uni.lu, rudi.balling@uni.lu

N. Baliga · L. Hood
Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109–5234, USA
e-mail: nbaliga@systemsbiology.org, lhood@systemsbiology.org

A. Prokop and B. Csukás (eds.), *Systems Biology*,
DOI: 10.1007/978-94-007-6803-1_1,

## Abbreviations

| | |
|---|---|
| BASE | BioArray Software Environment |
| BS | BiSulphite |
| CATCH-IT | Covalent Attachment of Tags to Capture Histones and Identify Turnover |
| CFS | Correlation-based Feature Selection |
| CHARM | Comprehensive High-throughput Array for Relative Methylation |
| ChIA-PET | Chromatin Interaction Analysis by Paired-End Tag |
| ChIP | Chromatin ImmunoPrecipitation |
| CLIP | Crosslinking immunoprecipitation |
| DHS | DNAse I hypersensitivity |
| DNA | DeoxyriboNucleic Acid |
| EFS | Ensemble Feature Selection |
| ELISA | Enzyme-Linked ImmunoSorbent Assays |
| ENCODE | ENCyclopedia Of DNA Elements |
| ESI | ElectroSpray Ionisation |
| EWAS | Epigenome-Wide Association Studies |
| FAB | Fast Atom Bombardment |
| FAIRE | Formaldehyde-assisted isolation of regulatory elements |
| FDR | False Discovery Rate |
| FT-ICR | Fourier Transform Ion Cyclotron Resonance |
| FUGE | Functional Genomics Experiment data model |
| GAGE | Generally Applicable Gene-set Enrichment |
| GC | Gas Chromatography |
| GEO | Gene Expression Omnibus |
| GO | Gene Ontology |
| GSEA | Gene Set Enrichment Analysis |
| GWAS | Genome-Wide Association Studies |
| HITS-CLIP | HIgh-Throughput Sequencing of RNAs isolated by CrossLinking ImmunoPrecipitation |
| HMM | Hidden Markov Models |
| HPLC | High Performance Liquid Chromatography |
| IMS | Imaging Mass Spectrometry |
| IP | ImmunoPrecipitation |
| iTRAQ | Isobaric Tags for Relative and Absolute Quantitation |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| kNN | k-Nearest Neighbor |
| LC | Liquid Chromatography |
| MALDI | Matrix Assisted Laser Desorption Ionisation |
| MBD | Methyl-CpG Binding Domain |
| MCAM | Multiple Clustering Analysis Methodology |
| MeDIP | Methylated DNA Immunoprecipitation |
| MGDE | Microarray Gene Expression Data |
| MIAME | Minimum Information About a Microarray Experiment |

| | |
|---|---|
| MIAPE | Minimum Information About a Proteomics Experiment |
| MINSEQE | Minimum INformation about a high-throughput SeQuencing Experiment |
| MMASS | Microarray-based Methylation Assessment of Single Samples |
| MN | Microccocal Nuclease |
| MRM | Multiple Reaction Monitoring |
| mRNA | Messenger RiboNucleic Acid |
| MS | Mass Spectrometry |
| NCBI | National Center for Biotechnology Information |
| NER | Named-Entity Recognition |
| NGS | Next Generation Sequencing |
| NIH | National Institutes of Health |
| NMR | Nuclear Magnetic Resonance |
| PaGE | Patterns from Gene Expression |
| PCR | Polymerase Chain Reaction |
| PRIDE | PRoteomics IDEntifications |
| PSM | Peptide-Spectrum Match |
| QMS | Quadrupole Mass Analyser |
| RNA | RiboNucleic Acid |
| RRBS | Reduced Representation Bisulphite Sequencing |
| RT-qPCR | Reverse Transcription quantitative PCR |
| SAGE | Serial Analysis of Gene Expression |
| SELDI | Surface Enhanced Laser Desorption Ionization |
| SILAC | Stable Isotope Labeling by Amino acids in Cell culture |
| SNP | Single Nucleotide Polymorphism |
| SRM | Selected Reaction Monitoring |
| SUMCOV | SUM of COVariances |
| SVM | Support Vector Machine |
| ToF | Time-of-Flight |
| UCSC | University of California, Santa Cruz |
| VOCs | Volatile Organic Compounds |

## 1.1 Background

### *1.1.1 Context*

Life in a broad scientific context can be defined as the phenomenon that emerges from particles of inorganic matter organised in molecules which interact with each other within a cell [1]. This property is systemic because it only appears in the system and not in its parts [2]. Living systems are complex, modular and hierarchical structures. Indeed, a multicellular organism consists of molecules, such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA), proteins, lipids and

metabolites involved in chemical reactions and structures of cells. Cells are organised in tissues forming organs with specific functions that are required for the health of the organism. Systemic properties appear at each level, for example homeostasis and response to stimuli in a single intracellular network, metabolism, growth, adaptation, reproduction in a single cell.

Information that defines an organism and its ability to react to its environment is encoded in its DNA and is expressed differentially in space and time throughout life. Typical studies in biology have until recently used the reductionist approach and addressed specific issues employing one or a few types of molecules at a small scale, each shedding light on only a small fraction of vastly complex phenomena. Some findings were remarkable, such as the discovery of the structure of DNA, and later of the way genetic information stored in DNA is transcribed in messenger RNA (mRNA) then translated in proteins, essential components of the cell machinery and the engines of life. The accumulation of such knowledge on molecules and mechanisms led to the *'bottom-up'* approach to modeling biological systems, using genes as core elements to simulate cells, organs and the whole organism. This was complementary to the *'top-down'* view of an organism as a physiological system integrating information from its various constituents and their interaction with the environment.

Major technological advances have in the last 15 years enabled biologists to eventually gather information on a larger scale in various tissues, including samples obtained with non-invasive methods, such as the collection of blood and urine. The massive increase in throughput has had several consequences. First, biologists can now study the vast majority of constituents, i.e. 'ome', of a given element, e.g. genes, of a system be it an organism, organ or cell, e.g. all genes in its genome. Second, the sheer size of data sets implies that their analysis relies increasingly on computational tools and power available to analysts. Third, because characterisation of several 'omes', e.g. genome, transcriptome, proteome and metabolome, progresses rapidly along with other disciplines such as imaging and in particular pharmaceutical research with cheminformatics, compound libraries, high throughput screening, safety and clinical data [3–5], one can now attempt to disentangle interactions between the different elements of a biological system, or 'interactome', to understand its behavior across several scales in a holistic manner, in health and disease.

### *1.1.2 Aims and Concepts*

Systems Biology is the integrative study of complex systems in life with a *holistic* approach now based on large-scale data sets analyzed iteratively with mathematical models and simulation tools [6, 7]. Understanding each component of a complex system in isolation is not sufficient to characterise the system. Indeed, properties of the system are not only defined by the simple addition of elementary functions but also emerge from the *interactions* between the elements [7–9]. These emergent properties are studied by inferring networks of interactions between

these constituents, e.g. genes, proteins and ligands, and by unraveling their regulatory mechanisms. Because of the very large number of elements in these networks, such an endeavor relies on concepts defined in the framework of the theory of complex systems [10]. Systems Biology not only aims at understanding the relationships between different levels of the expression of genetic information, via *data integration*, but also at defining the system as a whole and producing a convincing *mathematical model* of it, linking the highly complex interactions between its components to its *emergent properties* [11–14]. In this context, disease can be viewed as a shift of homeostasis from the normal range due to a large set of perturbations in the network of interacting biomolecules in the whole organism. Distinct perturbations may therefore result in a single disease phenotype, in agreement with our understanding of complex diseases. Conversely, shifting the system back to healthy homeostasis may be achieved in multiple ways and by targeting several points in the network [15, 16].

Systems Biology follows an integrative and iterative approach that relies on experimental and mathematical methods (Fig. 1.1). First, existing data relating to different hierarchical levels of the system are integrated into mathematical or graphical models to generate hypotheses towards understanding mechanisms at play and build predictions on the functions of that system. Some components of the system are then perturbed experimentally, such as in *in vitro* or *in vivo* models of a disease. The outcome is assessed in the context of the model and the initial hypotheses are revised accordingly. These revised hypotheses finally inform new perturbation experiments. The approach is repeated until the system's behaviour is faithfully simulated by the model [7]. Further complexity is added when one considers the environmental factors of the model.

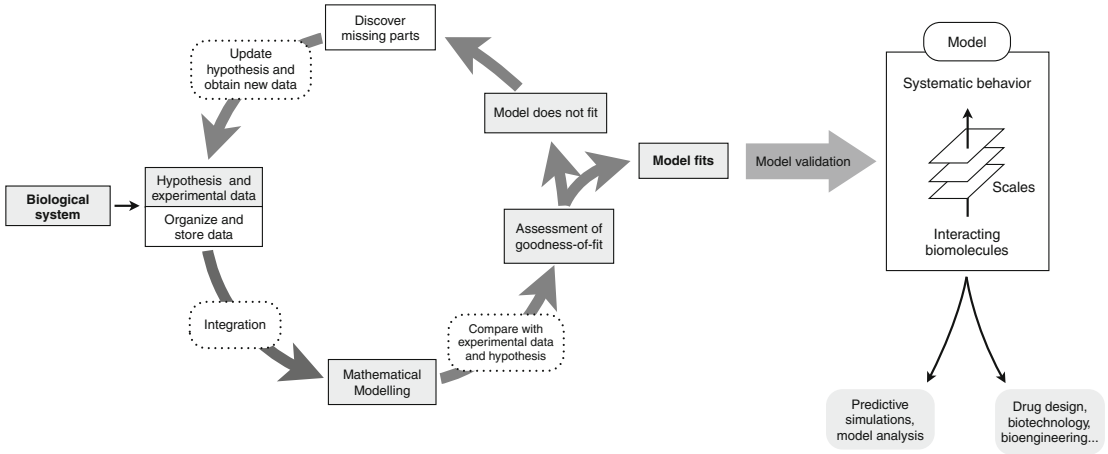


**Fig. 1.1** Modeling in Systems Biology. Modeling starts with the integration of different experimental data into a single knowledge base to organize and store data. Mathematical descriptions of the interaction between model elements allow (1) simulation of the emergent behavior of the system, (2) comparison of this simulated behavior with experimental data, (3) adjustment of the model and (4) design of further experiments. When the model fits experimental data, studying the role of particular design features may help identify mechanisms at play and design principles. The model may also be used in drug design, biotechnology or bioengineering for example

### 1.1.3 Strategies

Three main strategies aim to build the link between the system's components and its emerging properties: 'bottom-up', 'top-down' and 'middle-out' (Fig. 1.2). The main steps of the 'bottom-up' approach are to graphically or mathematically model relationships between the components of the system, starting with those at the lowest level of the multiscale structure, hence 'bottom', e.g. genes and proteins, set model parameters using experimental values and verify the model by comparing its systemic behavior with the behavior of a real system. The term bottom-up also refers to the direction chosen: from known or assumed properties of the components one deduces system functions [17]. This molecular biology strategy has been successful in modeling biological systems with relatively low number of components, e.g. a single intracellular network or a single prokaryotic cell. It may however not be suited to reconstruction of the emergence of larger systems, e.g. the whole body physiological behavior in Mammals. In contrast, the 'top-down' or physiology approach relies on the systemic behavior. It first involves defining ways the complicated systemic function of interest varies with conditions and/or time, and then inferring hypothetical structures responsible for this function. The system behavior is perturbed and the effects studied at the level of the system components, i.e. genome, transcriptome, proteome and metabolome. This strategy is limited to an extent by the challenge of inferring DNA sequences

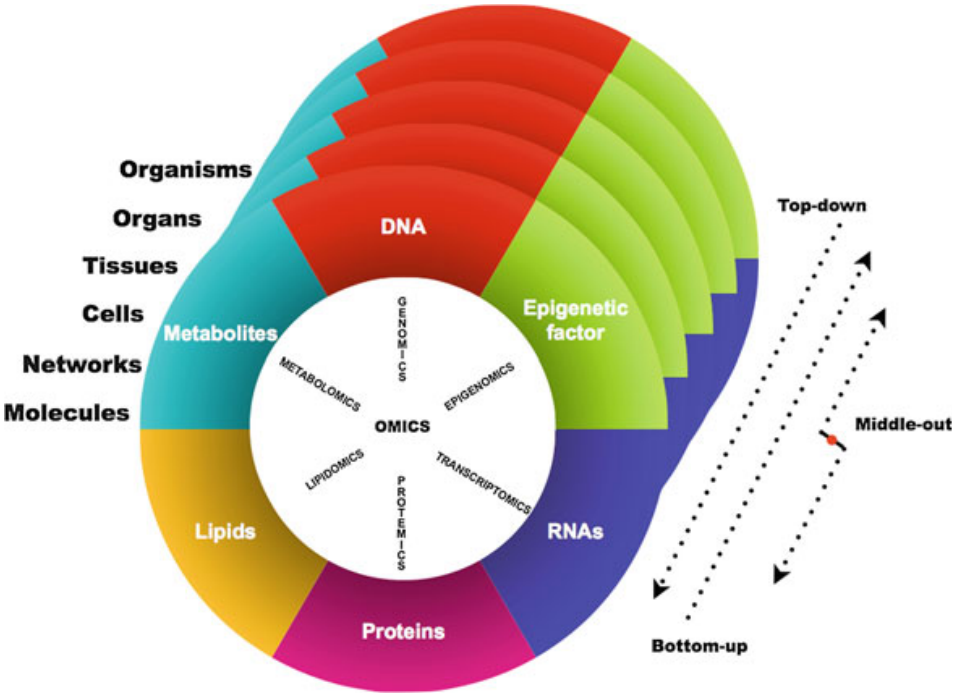


**Fig. 1.2** Multiple scale strategies in Systems Biology. Starting at the molecular level, interactions between DNA, epigenetic factors, RNA, proteins, lipids and metabolites define the core biological processes required for higher order functions. These processes are defined by molecular interaction networks, which communicate with each other within a given cell, between cells in the same tissue or distinct tissues, or between organs of a complex organism

from phenotypes. Also, models built with top-down approaches must be updated with every new experiment using all existing experiments, making the analytical and computational challenges increasingly difficult. In contrast, models built with the bottom-up approach such as an *in silico* cell model comprise modules which are updated independently of each other [18]. The '*middle-out*' strategy intends to overcome the intrinsic limitations of the above approaches, taking into account that chains of causality can operate in biological systems in both directions, starting at any levels of biological organization. The behavior of a single functional system is thus modeled in terms of interactions between entities at a level sufficiently well described by experimental data ('middle'), typically of the lower levels of organization but not necessarily down to molecules. The model is then extended to higher and lower levels ('out') iteratively by combining 'bottom-up' and 'top-down' approaches. It was successfully implemented in the Physiome project [19, 20].

Systems Biology will play a crucial role in the development of personalized medicine as it will enable integration of different types of data to profile patients, identify unbiased biomarkers and produce precise disease phenotypes. It will hence help prevention, diagnosis and treatment, or Systems Medicine [21, 22].

## 1.2 Introduction to Functional Genomics, Proteomics, Metabolomics and Bioinformatics

Genomics is the study of the sequence, structure and content of the genome, in particular the genes and their number, structure, function and organisation along the genome. *Functional genomics* is the study of the function of genes and the regulation of their expression at the level of the cell, organ or organism, spatially and at different time points and/or health status, by deciphering the dynamics of gene transcription, translation and protein–protein interactions on a genome-wide scale using *high-throughput* technologies. The main large-scale experimental tools used to study epigenetics (*epigenomics*) and gene expression (*transcriptomics*) have so far involved microarrays and more recently next-generation sequencing. Mass spectrometry is widely used to study proteins (*proteomics*), metabolites (*metabolomics*), and more recently volatile organic compounds (VOCs) in exhaled breath condensate (breathomics). Technical advances also led to the development of computational tools to handle and analyse their output.

### *1.2.1 Sequencing Technologies*

Whole genome sequencing started with the sequencing of a bacteriophage in 1977 using the Sanger sequencing technique. The development and maturation of

4-color automated Sanger sequencing produced the instruments that sequenced the human genome (Smith et al. 1986). Several high-throughput sequencing techniques, or *Next Generation Sequencing* (NGS), arose subsequently which were each inferior to the more established automated Sanger technique, being slower per run, less accurate, with shorter read length and more expensive, but far superior by virtue of the vastly larger number of nucleotides read [23–25]. Now 3rd generation sequencing strategies employ nanopores and single molecule reads, and promise to increase the throughput and decrease the cost of sequencing strikingly. Computational tools are being developed to process the very large amount of NGS short, low quality reads and assemble them into a genome sequence [26]. Genome sequences of over sixty pro- and eu-karyotes are annotated in online public genome browsers [27, 28]. Knowledge of whole genomes also enabled the large-scale study of gene expression and the development of functional genomics. NGS can indeed be used for DNA or RNA sequence analyses and has several advantages over microarrays: it does not require array design, enables wider scale, whole-genome studies, improved resolution, more flexibility, allele-specificity, lower cost and amount of input material. NGS now also enables routine discovery of variants in entire exomes and even large genomes [29, 30] as in Human with the 1000 Genomes Project [31], in cancer research [32, 33] and studies of allele specificity in gene expression [34]. NGS also catalyzed the massive development of metagenomics [35] and will thus help decipher host-gene-microbial interactions [36]. NGS is however not mature enough for routine use in clinical field [37]. The ever increasing speed, quality and range of applications of sequencing methods have created a huge flow of data and related challenging requirements not only for computing power, memory and storage [38–40] but also data sharing [41]. Reads mapped onto a reference genome can be displayed with other sources of annotation such as NCBI [42] with Ensembl [28] and UCSC browsers [43].

### *1.2.2 Mass Spectrometry*

*Mass spectrometry* (MS) relies on deflection of charged atoms by magnetic fields in a vacuum to measure their mass/charge (m/z) ratio. A typical experiment follows five steps: (1) introduction of the sample, (2) ionisation of its particles, (3) acceleration, (4) deflection proportional to the mass and charge of the ion, and (5) detection, recorded as a spectrum showing peaks on a plot of relative quantity as a function of the m/z ratio.

Several methods for introduction, ionisation and types of spectrometers enable a wide range of analyses. Introduction methods are Gas chromatography (CG) for thermally stable mixtures, liquid chromatography (LC) for thermally labile mixtures, and solid probes. Some compounds such as large proteins and polymers must be ionized directly. Ionisation methods can be hard or soft. Hard ionisation introduces high amount of energy in the molecules that results in fragmentation and thus helps identify the compound but resulting spectra rarely contain the

molecular ion. ElectroSpray Ionisation (ESI) uses high voltage to disperse and ionise macromolecules through a spray nozzle. It is soft, limits fragmentation and produces multiply charged ions, allowing detection of large compounds at lower mass/charge value, and hence increases the analyser's mass range. ESI is often coupled with LC/MS. Mixtures containing non-volatile molecules can also be analysed with Fast Atom Bombardment (FAB) and Matrix Assisted Laser Desorption Ionisation (MALDI). MALDI is used to analyse extremely large molecules, up to 200,000 Da, often coupled with time-of-flight (ToF) MS. Surface Enhanced Laser Desorption Ionization Mass Spectrometry (SELDI-MS) separates protein subsets fixed onto a surface according to specific biophysical properties, e.g. hydrophobicity. Thus, analysis of proteins, peptides and nucleotides can be performed with ESI, SELDI, MALDI, and FAB [44].

Several types of analysers exist. In a quadrupole mass analyser (QMS) ions are deflected by oscillating positive and negative electric fields. A triple-QMS contains three QMS one after the other where the first QMS enables the identification of known compounds, the second its fragmentation, and the third the identification of the fragments, thereby elucidating the compound structure. Other types of analysers include ion trap, ToF, Orbitrap, and Fourier Transform Ion Cyclotron Resonance (FT-ICR) with increasing mass resolution and accuracy. Orbitraps are cheaper, more robust and have a higher-throughput than FT-ICRs. Tandem-MS involves several steps of selection of compound using MS. MS methods mentioned above vary in throughput, robustness, sensitivity, selectivity and ease of use [44].

### *1.2.3 Bioinformatics*

*Bioinformatics* comprises mathematical approaches and algorithms applied to biology and medicine using Information Technology tools, e.g. databases and mining software [45, 46]. Analysis of omics data typically follows four steps: (1) data processing and identification of molecules, (2) statistical data analysis, (3) *pathway* and *network* analysis, and (4) system modelling. Examples include *de novo* genome assembly, genome annotation, identification of co- or differentially expressed genes at the level of transcripts or proteins and the inference of protein–protein interaction networks. Bioinformatics also enables integration of heterogeneous high-throughput data sets produced by a given study and existing data sets using knowledge management, annotation and text mining tools such as the two structured vocabularies Gene Ontology (GO) for genes and associated biological processes, cellular components and molecular functions [47, 48] and Microarray Gene Expression Data (MGED) ontology [49], the PRoteomics IDEntifications (PRIDE) database [50], Functional Genomics Experiment data model (FuGE) [51], the Systems Biology Markup Language [52], the Systems Biology Graphical Notation [53], BioMART [54, 55], tranSMART [56], bioXM [57], GARUDA [58], Nexbio [59], and includes Systems Biology [23]. Identification of pathways, and

network inference and analysis is covered in chapter 'Network analysis for systems biology'.

These efforts collectively aim at unraveling the molecular pathways underpinning physiology and at identifying biomarkers to describe a system with a combination of environmental, clinical, physiological measures to improve detection and monitoring of a phenomenon, such as diseases in medical research to facilitate diagnosis and therapy. Biomarker discovery relies on two types of studies: unbiased, which only depend on the technique used, and targeted, which focus on pre-defined biomarkers measured by specific methods. Experimental and bioinformatics methods and tools mentioned in the following text are listed in Tables 1.1 and 1.2.

## 1.3 Functional Genomics, Proteomics and Metabolomics

### *1.3.1 Epigenomics*

Epigenomics is the genome-wide study of modifications of chromatin, i.e. DNA and associated proteins, which play an important role in gene regulation, gene-

**Table 1.1** Examples of methods and tools for functional genomics, proteomics and metabolomics. This list is non exhaustive and only includes items mentioned in the text

| | |
|---|---|
| Epigenomics methods | DNA methylation [61]: Endonucleases (MMASS, CHARM, Methyl-seq), bisulphite (BS) conversion (RRBS, MethylC-seq), and affinity (MeDIP-chip, MeDIP-seq, MDB-seq). Methylation levels can then be measured with microarrays and sequencing techniques; |
| | Chromatin accessibility (DNAseI-seq, FAIRE–seq, Sono-seq, 3C, 4C, 5C, ChIA-PET); |
| | Nucleosome positioning (CATCH-IT, MNase-se, haploChIP) |
| Epigenomics tools | Encyclopedia Of DNA elements (ENCODE) project [63], the NIH Roadmap Epigenomics effort [64], the Human Epigenome Project [65] and recently BLUEPRINT [67] |
| Transcriptomics methods | DNA microarray, SAGE, RNA-seq, ChIP-seq, CLIP-seq [108, 113, 114, 117] |
| Transcriptomics tools | ArrayExpress [104], GEO [106], MIAME [107], MINSEQE [119]. See [26, 120] for reviews on downstream analysis. |
| Proteomics methods | ELISA, 2D gel electrophoresis, NMR, MS, iTRAQ, SILAC, SRM, SELDI-ToF [126–131] |
| Proteomics tools | MIAPE [134], TransProteomic pipeline, protein atlas, neXProt [139–141] |
| Metabolomics methods | NMR [143], MS [44], IMS [144, 147] |
| Metabolomics tools | MetabolomeExpress [150], metaP [151], KEGG [145], human metabolome project [142] |
| Lipidomics methods | MS [44, 161], orbitraps [160], IMS [144, 147] |
| Lipidomics tools | LIPID MAPS [165], XCMS [162], MZmine2 [163] |

**Table 1.2** Examples of methods and tools for bioinformatics. This list is non exhaustive and only includes items mentioned in the text

| | |
|---|---|
| Bioinformatics | Microarray gene expression data (MGED) ontology [49], the proteomics identifications (PRIDE) database [50], functional genomics experiment data model (FuGE) [51], the systems biology markup language [52], the systems biology graphical notation [53], BioMART [54, 55], tranSMART [56], bioXM [57], GARUDA [58], nexbio [59] |
| Clustering | Babelomics [176], BASE [177], MCAM [178] |
| Feature selection | Unsupervised [187], supervised [186]; filters (student's *t* test, Wilcoxon rank sum test, CFS, EFS, Markov blanket filtering) [188], wrappers (kNN [203], Naive Bayes [204], sequential forward search [205]), hybrid methods [202], mathematical programming [209], signal processing approaches [210] |
| Prediction analysis | Unsupervised (clustering, feature selection, dimension reduction, density estimation, and model structure learning, nonlinear dimension reduction methods) [211–213]; supervised (SVM [215], random forest [216]); semi-supervised [217]; time series (HMM [218]) |
| Networks from literature | NER [225], iHOP [232], FActa + [221], AliBaba [233], IntAct [234], CoPub [235] |
| Pathway analysis | Differential expression filtering, overrepresentation statistics [236], GSEA [240], PAGE [241], GAGE [242], ontologizer [243], GeneCodis [244], elementary flux analysis [245], extreme pathways [246] |

environment interactions, development and in diseases such as inflammation and cancer [60, 61]. Such modifications involve the DNA itself but not its sequence, i.e. a methylated cytosine (mC) adjacent to a guanine (CpG dinucleotides in mammals), and of chromatin proteins, i.e. methylation, acetylation and phosphorylation of histones. Epigenomics also covers chromatin accessibility, nucleosome remodelling, long-range chromatin interactions and allele-specific chromatin signatures. Technological advances are now enabling Epigenome-Wide Association Studies or EWAS, akin to Genome-Wide Association Studies or GWAS [62], and large scale studies in different cell types and tissues, as in the human ENCyclopedia Of DNA Elements (ENCODE) project [63], the NIH Roadmap Epigenomics effort [64], the Human Epigenome Project [65], [66] and recently BLUEPRINT that aims to determine the epigenome of 100 different blood cell types [67].

DNA methylation at CpG is widely studied as it mediates gene repression in a cell-specific manner by preventing the transcriptional machinery from accessing DNA. Methylated DNA can be detected with three types of DNA treatments, i.e. endonucleases, bisulphite (BS) conversion, and affinity. Methylation levels can then be measured with microarrays and sequencing techniques.

Endonucleases cleave DNA at specific sites, are sensitive to methylation and enable several DNA analyses techniques. Recent methods enable analysis of a single sample, e.g. microarray-based methylation assessment of single samples (MMASS), better statistical analyses and methods for array design, e.g. comprehensive high-throughput array for relative methylation (CHARM) [68] and the

widely used NGS sequencing of DNA enriched for CpG containing regions (Methyl-seq) [61].

BS conversion modifies unmethylated cytosine in CpGs into a uracil and thus transforms an epigenetic difference into a genetic one detectable by methylation specific DNA microarrays with single-nucleotide resolution [69, 70]. Except for mC, BS treated DNA comprises only three base types and hence has reduced sequence complexity and hybridization specificity. This is overcome by enriching for CpG-containing segments as in Reduced Representation Bisulphite Sequencing (RRBS) with BS treatment and NGS. Alternatives include whole-genome BS sequencing, although that is expensive, and the widely used MethylC-seq, i.e. NGS of BS treated DNA. Throughput and coverage may increase with nanopore sequencing which can sequence mC directly, without BS treatment [71].

Genome-wide identification of DNA binding-sites and corresponding binding proteins is mainly achieved with the affinity-based approach chromatin immunoprecipitation (ChIP) whereby DNA-binding proteins, e.g. histones and transcription factors, are cross-linked *in vivo* in cells that are then lysed. DNA is fragmented by sonification, recovered by heating DNA–protein complexes and detected with microarray (ChIP-chip) or NGS (ChIP-seq) [72, 73]. Methylated DNA Immunoprecipitation (MeDIP-chip and MeDIP-seq) uses monoclonal antibody against methylated cytosine to enrich single-strand methylated DNA. Some alternatives rely instead on high affinity binding of a Methyl-CpG Binding Domain (MBD) protein complex for double-strand methylated DNA (e.g. MDB-seq) [60, 74]. Transcription factor binding sites are then predicted in the sequences identified [75]. ChIP is also widely used to study patterns of histone modifications and chromatin modifiers [63, 76]. It can be integrated to other data sets, as with Segway [77], helping development of chromatin model [78]. ChIP coupled with quantitative real-time PCR allows the study of the dynamics of DNA and proteins interactions in living cells for up to several minutes, and has now been adapted to microfluidics technology reducing the number of cells and time required [79].

Across the three types of treatment, at least 13 array- and 10 seq-based analytical methods exist, the choice of which depends on their features, the required coverage and resolution, types of bias, accuracy and reproducibility, and also on the number of samples, available DNA quality (high for affinity techniques) and quantity (high for nuclease techniques), and in particular for array-based methods: the organism. The most widely used NGS-based methods rely on BS (RRBS and MethylC-seq) or affinity (MeDIP-seq and MBD-seq) approaches [61, 80, 81].

Microarray data processing addresses imaging and scanning artefacts, background correction, batch and array normalization, and correction for GC content and CpG density. The ratio of methylated to unmethylated molecules for a given locus is a widely used metric. It is analysed with tools developed for gene expression data, potentially wrongly since they rely on assumptions violated by DNA-methylation data, e.g. independence of the number of methylated and unmethylated sites, and similarity of signal strength across samples [61, 82–84]. Processing sequencing reads involves mapping of reads to the reference genome, counting and/or analysis of bisulphite data [85, 86].

Genomic regions of chromatin accessibility, i.e. low nucleosomal content and open chromatin structure, potentially harbour regulatory sequences and can be identified with high-throughput DNAse I hypersensitivity assay (DNAseI-seq aka DHS-seq) [87], formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE–seq) [88] and Sono-seq [89]. And long range chromosomal interaction are identified with chromosomal conformation capture (3C) [90, 91], 3C on chip (4C) [92], 3C carbon copy (5C) [93] and coupled with NGS as in using Hi-C [94] and ChIA-PET [95]. Nucleosome positioning and remodelling is studied with CATCH-IT [96] and MNase-seq [97] while haploChIP identifies allele-specific chromatin profiles [98, 99], including SNPS that affect gene expression [100].

Methods to integrate epigenomics data are recent and currently being developed. Examples include integration with gene expression data, using an empirical Bayes model [101] and clustering of DNA methylation data followed with non-linear regression analyses [102]. Visualisation tools can display raw data genome-wide as with Circos [103] or analysis output in a similar manner to that used for GWAS, using log10 *p*-value, but on two axes: test of difference in methylation status and test of difference in gene expression [83].

### *1.3.2 Transcriptomics*

Transcriptomics is the genome-wide identification and quantification of RNA species such as mRNAs, non-coding RNAs and small RNAs, in health and disease, and in response to external stimuli. With DNA microarrays, gene expression levels are measured as the amount of RNA in the sample that matches the set of probes fixed on the array; RNA molecules are fluorescently labelled and hybridised onto the array where the intensity of the signal measured for a given probe is assumed to be proportional to the quantity of RNA bound to it. Changes in expression levels between experimental conditions or samples with or without disease on one hand and similarity of expression pattern with a gene with known function on the other hand indicate the most likely functions of the genes. Two main public repositories for gene expression data sets exist: ArrayExpress [104, 105] and Gene Expression Omnibus (GEO) [106], both compliant with the 'Minimum information about a microarray experiment' (MIAME) guidelines [107]. Although microarrays are an established and very widely used technology [108], data processing and analysis methods are still being developed. For example, recent studies claim that models for background noise based on Gaussian distribution for computational efficiency may not be appropriate and non-parametric methods may harbour a lower false positive rate [109], while weighted average difference seems to be the best method to identify differentially expressed genes [110]. Two main sequencing-based alternatives exist which, unlike microarrays, do not rely on a set of pre-defined probes and are therefore considered unbiased: Serial Analysis of Gene Expression (SAGE) and genome-wide transcriptome NGS (RNA-seq).

SAGE entails sequencing tags that are unique to each gene and not defined *a priori*. SAGE was for example used to build expression profiles of long non-coding RNAs for 26 normal tissues and 19 cancers in human [111], shedding light on their poorly understood function [112]. The more recent RNA-seq provides whole transcript sequences, has very low background noise, offers a very large dynamic range, is highly accurate and reproducible, enables the discovery of novel exons, isoforms and transcripts. RNA-seq has already proved very promising but is not as mature as microarrays yet [113–115]. Rare and transient transcripts so far undetected by current methods were recently identified with targeted transcriptomics by capture on tiling array followed by NGS [116]. Currently, some experimental protocols may introduce bias due to amplification, fragmentation and ligation processes [117, 118]. Development of robust quality control standards and guidelines for microarrays occurred over a decade but should be faster for RNA-seq. Methods are being developed to describe experiments using MIAME-like 'Minimum Information about a high-throughput SeQuencing Experiment' (MINSEQE) guidelines [119], map the vast amount of short read sequences [26], assess expression levels and detect differentially expressed transcripts [120].

Estimates of expression levels of transcripts of interest must be validated by RT-qPCR and emerging techniques such as direct visualization and counting of RNA molecules [121]. These must however be standardised and applied across platforms [21]. Microarrays are still relatively cheaper than RNA-seq, their biases are known and analysis workflows are mature. They are therefore still preferred in drug discovery, though RNA-seq methods will probably replace them over the next years. Because gene expression profiles obtained with both methods correlate well, the vast amount of data acquired with microarrays is complementary to new data produced by RNA-seq [108].

Other techniques such as ChIP are also used to identify proteins binding DNA (ChIP-seq) [73] and RNA (CLIP-seq aka HITS-CLIP) [122]. These fast evolving high throughput methods are greatly improving our understanding of gene expression regulation [123, 124], at the transcriptional and post-transcriptional levels [125].

### *1.3.3 Proteomics*

Correlation between levels of transcripts and proteins is incomplete due to variation in speed and efficiency of translation and of mRNA degradation. Many proteins undergo posttranslational modifications, e.g. phosphorylation and ubiquitination, which modulate their activity and mediate signal transduction. Proteins also play their role as part of complexes with other proteins or nucleic acids. A recent study of a human cell line identified over 10,000 proteins, with concentrations ranging over seven orders of magnitude. The human proteome has been estimated to comprise several millions distinct species which cannot currently be amplified and reflect concentrations with a very wide dynamic range [126].

Proteins can be identified using low-throughput antibody methods, Enzyme-Linked ImmunoSorbent Assays (ELISAs) and 2D gel electrophoresis. Proteomics aims at defining all of the proteins present in a cell, a tissue, or an organism (or any other biological compartment) and employs large-scale, high-throughput studies of protein content, modifications, function, structure, localisation, and interactions using high-throughput techniques. Protein microarrays capture proteins using agents fixed on their surface, which can be antibodies but also peptides, receptors, antigens, nucleic acids. Detection and quantification are often fluorescence-based and identify interactions between proteins, kinase substrates, activators of transcription factors [127]. Nanoproteomics has the potential to provide fast, high-throughput and sensitive methods using only minute amount of samples [128]. However, MS is currently the main technique for large-scale whole-proteome study with precise measurements [129, 130].

Shotgun proteomics, i.e. shotgun LC coupled with tandem MS (LC–MS/MS) is the most widely used approach. The sample of peptides resulting from the trypsin (or other enzyme) digestion of proteins is separated by High Performance Liquid Chromatography (HPLC) and peptides are identified using tandem MS: peptides are ionised and separated, producing mass spectra with peaks corresponding to peptides (first MS), which are then identified using further fragmentation and separation of resulting peptide fragments (second MS). Inclusion of labelled synthetic peptides as spike-in or labelling samples chemically (iTRAQ) or metabolically (SILAC) improves quantification [131]. Mixture complexity is addressed by fractioning the mixture. Targeted proteomics allows one to identify 100-200 proteins in a complex mixture by previously identifying the "transition peptide fragments" through the use of a triple quadrupole mass spectrometer which separates the trypsin peptide fragments, then fragments these further into "transitions" that can be quantified in the third quadrupole. One attempts to choose transitions that are unique to individual proteins and spiking in isotopically labelled transition peptides greatly improves quantification. Targeted mass spectrometry is termed Selected Reaction Monitoring (SRM) or Multiple Reaction Monitoring (MRM). SRM assays for the entire human proteome (more than 20,000 proteins) have recently been developed (R. Mortiz, personal communication).

HPLC–MS is highly sensitive, specific and fast, and thus used for bioanalysis, in particular pharmacokinetics to measure speed of drug clearance by the body, and in urine sample analysis. Drawbacks however include a bias towards identification of most abundant peptides. SELDI-ToF is more accurate than shotgun approach and is thus better suited to biomarker quantification, but may not be accurate enough for clinical diagnostics [132].

Recent techniques produce data sets of approximately one million spectra, up to 100 Gb in size, where up to 8,000 proteins can be identified [133]. Pre-processing of raw spectra entails noise filtering, baseline subtraction, peak detection, and calibration and alignment of LC/MS maps. Analysis follows four steps: (1) identification of amino-acid sequences, peptides and proteins in Peptide-Spectrum Match (PSM), and detection, quantification, annotation and alignment of features, (2) peptide and protein significance analysis, (3) class discovery and prediction,

and (4) data integration and pathway analysis. Identification of amino-acid sequences mainly involves searching databases of spectra obtained experimentally or of spectra predicted from genomic sequences using *in silico* digestion, and reporting PSMs with the best scores. Statistical strength of predictions is indicated using the False Discovery Rate (FDR) computed using decoy databases, or models including the proportions of true and false identifications. Because many spectra map to many peptides and many peptides map to many proteins, identification of peptides and proteins is cumbersome and not completely solved. The issue is further complicated by post translational modifications and single amino-acid polymorphisms. Current methods identify approximately two thirds of tandem MS spectra. Proteins are reported on the basis of single-peptide match, or more stringently of match to protease specific peptides [133, 134]. Experiments are described using MIAME-like Minimum Information About a Proteomics Experiment (MIAPE) guidelines [135].

Difference in protein abundance is assessed with protein quantification (concentration estimate) and class comparison (change in abundance between conditions). The principle is to summarise all quantitative data relating to the protein by (1) spectral counting, where the number of spectra is assumed to reflect abundance with LC MS–MS, and is limited to large change for abundant proteins in low-complexity mixtures, or (2) probabilistic models incorporating all features of a protein and their variation. These models aim to address important issues, such as representation of the experimental design, treatment of missing data and control of FDR [134, 136]. Recent studies have shown convincing examples of quantitative proteomics efforts ran across different laboratories and using several experimental platforms. Currently, about two-third of human proteins predicted to exist have been detected with MS, hence the need to improve sensitivity, reproducibility of identification, and sensitivity and accuracy of quantification [133, 134, 136]. Protein–protein interactions and cell signalling cascades are mainly studied with the following approaches: yeast two-hybrid complementation, protein microarray, immunoaffinity chromatography and MS [137], and with a lower throughput by immunoprecipitation and mass spectrometry in Mammals [138, 139]. Attempts to integrate proteomics with other omics data are hindered by current drawbacks of proteomics analysis: proteome not completely sampled, uncertain identification of protein, difficulties in mapping identifiers across the different omics sources, hence the need for protein-centric knowledge bases such as TransProteomic Pipeline [140], Protein Atlas [141] and neXProt [142].

## *1.3.4 Metabolomics and Lipidomics*

### 1.3.4.1 Metabolomics

Metabolomics is the high-throughput characterisation of the mixture of all metabolites in a biological system, i.e. endogenous and exogenous small

molecules [143]. Metabolites are lipids, peptides, and amino, nucleic and organic acids. Metabolomics is now widely used in microbiology, nutrition, agriculture and environmental sciences, and clinical and pharmaceutical fields. Metabolites are the product of enzymatic reactions mediating complex biological processes and may therefore help understand phenotypes. They can be analysed using NMR spectroscopy although it lacks sensitivity [144] and MS (GC and LC) is usually preferred and used in targeted and untargeted approaches. Targeted strategies are specific and sensitive, allow absolute quantification and thus widely used in clinical diagnostics and drug development. Targeted approaches based on stable isotopes and models of metabolic networks allow estimation of the flux through biochemical pathways [145]. In contrast, untargeted approaches harbour a high coverage, though any metabolite identification is less specific and sensitive, and requires more intensive computational analysis. Features to use for identification are detected using univariate and multivariate analyses and then used to search databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [146, 147]. Further experiments to distinguish isomers and characterise unidentified metabolites using tandem MS or NMR are often required. Metabolomics also include identification of substrate in *in vitro* assays of three types: (1) the protein is fixed onto a surface and ligands screened, (2) the metabolite is fixed and serve as bait for interacting proteins, or (3) activity-based protein profiling using chemical probes and beads. Last but not least, location of metabolites within cells, tissues or bodies can be studied by coupling MALDI or matrix-free MS and imaging techniques (imaging mass spectrometry, IMS) to obtain spectra by scanning the biological sample with the laser and then compiling a map of metabolite content across that sample [145, 148].

Standards for experiment description and tools for processing and analysis of metabolomics data are actively being developed [149, 150]. For example, MetabolomeExpress [151] and metaP [152] both combine tools from raw data processing, i.e. MS peak detection, to multivariate analysis.

Development of biomarkers with metabolomics and comparison between data sets depend on: (1) the characterisation of technical MS artefacts and differences in compounds discriminating samples between analysers and (2) sample type and biological variability [153]. The Human Metabolome Project quantified over 4,000 metabolites in up to 70 samples [143] out of 6,826 identified by Wishart and colleagues [154]. Another recent large-scale targeted metabolomics study quantified 122 metabolites in 377 individuals, including type 2 diabetes patients and controls, and identified 25 metabolites in plasma and 15 more in serum with different concentrations in the two groups [155].

### 1.3.4.2 Lipidomics

Lipids play important roles in the signalling involved in metabolism, energy storage, and cell proliferation, migration and apoptosis [156]. They are also the main components of cellular membranes, together with membrane proteins.

They thereby maintain cellular architecture and mediate membrane trafficking by enabling protein machinery assembly, as for example in dynamic clusters gathering specific proteins in lipids rafts [157]. Lipids are very diverse in their structure, physical properties and quantity. For example, signalling and structural lipids are respectively found in low and high abundance. Lipidomes, the lipids present in biological structures, are currently poorly understood [158]. The Human lipidome may contain thousands of species [159] while only 20 % of all lipids may have been detectable with existing technologies, as in 2009 [154]. Lipidomics studies aim to characterise lipids content, localisation and activity in cells and tissues [160]. The vast majority of lipids are extracted from lysed cells and tissues, and analysed with MS either directly in the shotgun method, i.e. 'top-down' lipidomics with high resolution analysers such as Orbitraps, or with LC–MS/MS 'bottom-up' lipidomics to distinguish lipids with identical charge to mass ratio [161]. Lipids have also been analysed with MALDI IMS [162]. Lipidomics MS raw data can be analysed with tools used for metabolomics, such as XCMS [163] and MZmine 2 [164].

Lipids are identified and quantified using raw data processing and statistical analysis, followed by pathway analysis and modelling [165]. Major lipidomics intiatives include the 'Lipid Metabolites And Pathways Strategy' (LIPID MAPS) which has established standards and enabled absolute rather than relative quantification [166], and the Mouse Macrophage Lipidome [167]. Absolute quantities for proteomics and lipidomics will help characterise complexes comprising both proteins and lipids [145].

Future technical advances should aim for higher accuracy better consistency, and harmonisation of protocols. Analytical developments should include: (1) automated data processing and lipid identification and mining, (2) statistical data analysis to address high-dimensionality and platform-independent computation of lipid identification false discovery rate, (3) pathway analysis to identify biochemical, signalling and regulatory processes that involve the lipids of interest characterised in a sample set, and (4) modelling in time and space within the context of physiology and systems [168].

## 1.4 Methods and Tools

Current high-throughput technologies produce very large data sets and have shifted the bottleneck from data production to data analysis. *Knowledge management* tools are thus very valuable to organise, store and analyse data either directly with embedded software or indirectly by exporting the data in the required format. Recent data sets also harbour very high dimensionality. Data integration aims at combining such *high-dimensionality*, large data sets differing in the type of data collected. Unsupervised integration aims to reduce the dimensionality of large data sets, without introducing a bias inherent to prior knowledge and hypotheses. It helps detect patterns within and amongst data sets and complements standard

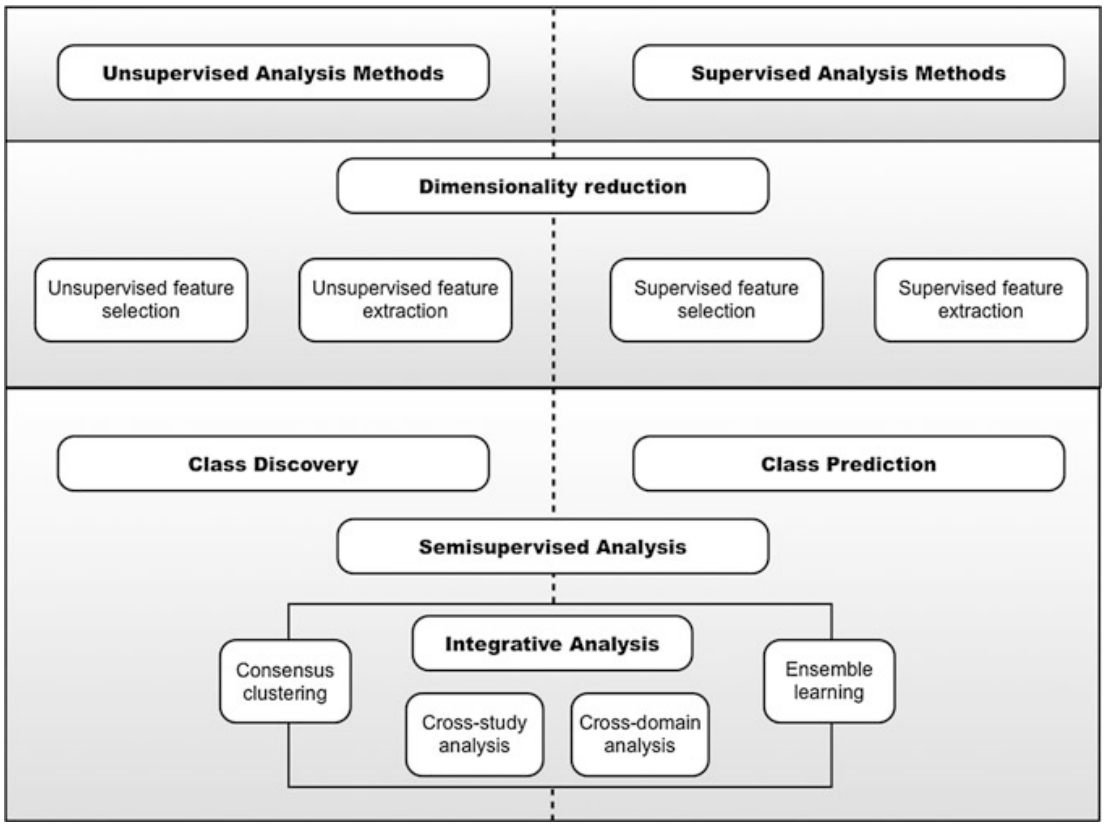


**Fig. 1.3** Overview of machine learning methods. Supervised and unsupervised methods range from lower level dimensionality reduction approaches to higher-level analytical techniques and their extensions for integrative data analysis [171]

observations in building hypotheses. These are then tested analytically with supervised methods, usually only using a fraction of the available dimensions, and experimentally [58, 169, 170]. Despite its power and promises data integration is only a means to an end, not an automatic engine to generate valuable findings. Indeed, answers to the questions asked in a scientific study directly depend on the experimental design, e.g. the types of data, controls, processing and analyses, and the size of samples, within financial and time constraints. The following section describes methods for clustering, feature selection, prediction analysis, text mining and pathway analysis (Fig. 1.3).

### *1.4.1 Clustering*

**Motivation:** Clustering is a data-exploration technique for multivariate analysis which divides data based on intrinsic groups without predefined labels. Clustering methods have been applied to various aspects of biomedical research, e.g. gene expression in cancer, to distinguish patients or genes subgroups based on expression levels of a set of differentially expressed genes. Clustered genes may have similar functions, be involved in the same cellular process or in similar pathways.

Such knowledge would improve our understanding of gene function and biological processes. Clustering methods can be used for visualization, hypothesis generation and selection of genes for further analysis.

**Pre-processing:** Clustering requires standard normalization methods for omics data [172–174]. Clustering specifically requires a prior dimensionality reduction and data standardization, e.g. filtering out genes or proteins with low variance across the samples, methods based on the maximization of a function of covariances as in the 'sum of covariances' (SUMCOV) method [175], and standardization of the data, e.g. mean absolute deviation standardization.

**State-of-the-art:** Numerous clustering tools have been developed. Several well-known clustering algorithms are: hierarchical clustering, partition and density-based clustering and fuzzy clustering. More recently developed clustering algorithms include: subspace or bi-clustering methods that cluster both genes and samples [176]. Automatic acquisition, pre-processing and clustering analysis via web-based tools is possible for several high-throughput technologies, e.g. Babelomics [177], BioArray Software Environment (BASE) [178] and Multiple Clustering Analysis Methodology (MCAM) [179]. Efficient cluster validation procedures are crucial for decision making with large number of genes in the absence of large amount of samples and will therefore be extremely useful to understand genetic interactions and design drug targets.

**Use cases**: Clustering is widely used in microarray data analysis and a wide choice of tools exists. Clustering of genes may identify a group of genes with similar functions while clustering of samples can suggest patient subgroups for stratification, response to treatments and disease subtypes or grade, e.g. childhood leukemia [180], breast cancer [181] and asthma [182, 183]. Clusters can also be integrated with pathway analysis [184].

### *1.4.2 Feature Selection*

**Motivation**: Feature or attribute selection methods have a wide range of applications in Systems Biology. They enable an experimenter to identify which genes or proteins are significantly differentially expressed across different biological conditions in a cell type of interest, and which subsets of genes or proteins provide the most promising combined set of biomarkers for discriminating between these conditions (see also the section on prediction analysis). Moreover, feature selection approaches are often used to reduce the dimension of the input data before applying other higher-level statistical analysis methods. This alleviates a variety of statistical problems referred to as the *curse of dimensionality* in the literature [185]. However, in contrast to feature transformation based dimension reduction methods [186], the original features of the data are preserved, which facilitates data interpretation in subsequent analyses.

Feature selection algorithms can be grouped into *supervised* [187] and *unsupervised* approaches [188], depending on whether they incorporate information

from class labels for the biological conditions. Moreover, feature selection algorithms employing prediction methods to score the informativeness of a feature subset are known as *wrappers*, whereas other univariate and combinatorial approaches to filter attributes are called *filters* [189].

**Pre-processing**: For most experimental platforms used in Systems Biology, several low-level pre-processing steps are required before applying feature selection methods. These include image processing [190, 191], normalisation [192] and summarisation approaches [193, 194], for microarray gene expression data [195], and raw data filtering [196], peak detection [197], peak alignment [198] and retention time normalisation methods for proteomics and metabolomics mass spectrometry data [199]. Moreover, some feature selection methods require a prior discretization of the data, e.g. if special association measures are used, such as mutual information [200].

**State-of-the-art**: The choice of the feature selection method depends both on the analysis goal (e.g. identifying individual biomarkers, or building a combinatorial predictive model for sample classification) and on the desired trade-off between efficiency (the run-time complexity of the algorithm) and accuracy (the predictive power of the selected features).

Among the filter approaches, simple univariate statistics like the parametric *Student's t test* and the non-parametric *Wilcoxon rank sum test* are still widely used, due to their advantages in terms of speed and the difficulty of estimating feature dependencies from noisy, high-dimensional data. More complex combinatorial methods such as *CFS* [201], *EFS* [202] and *Markov blanket filtering* [203] have recently gained influence.

Wrapper methods are becoming increasingly popular. They score feature subsets using prediction methods in combination with a search space exploration approach and their selections reach state-of-the art predictive performance in biological classification problems. Examples include combinations of fast and simple prediction methods, e.g. *kNN* [204] and Naïve Bayes [205], and search space exploration methods, e.g. sequential forward search [206]. These approaches are gradually being replaced by more complex algorithm combinations, including evolutionary algorithms [207] and kernel-based machine learning methods [208].

Finally, several recent techniques have improved the trade-off between speed and accuracy: (1) combination of filters [209], (2) combination of filters and wrappers into hybrid methods [203], (3) mathematical programming [210] and (4) signal processing approaches [211].

**Use cases**: Identification and prioritisation of gene, protein or metabolite biomarkers via feature selection techniques have three main aims: (1) distinguish biological conditions, e.g. presence of cancer, of viral infection, or tumor grades, (2) mediate early diagnostic, patient-tailored therapy, disease progression monitoring, and (3) help study treatment in a cell culture or animal model. However, feature selection methods are also used to filter datasets prior to the application of other higher-level data analysis methods, e.g. other machine learning methods, pathway overrepresentation analysis and network analysis. Finally, feature

selection is often integrated with classification and regression techniques to decrease the complexity of machine learning models and maximize their predictive accuracy.

### *1.4.3 Prediction Analysis*

**Motivation**: Prediction analysis refers to a family of methods that attempt to capture statistical dependencies and extract patterns from a set of measured data, to make predictions about future data. Such methods hold great promise in functional genomics, proteomics, metabolomics and bioinformatics, where the recent technologies provide a wealth of data such as gene and protein expression measurements, DNA and RNA sequence reads. The rate at which such data are produced makes automatic prediction analysis an indispensable tool for the biologist. Methods for prediction analysis can be unsupervised, semi-supervised, or supervised.

**State-of-the-art**: Unsupervised methods find regularities and hidden structure in the data. Typical approaches include clustering, feature selection, dimension reduction, density estimation, and model structure learning [212]. Classical linear dimension reduction methods are principal component analysis and independent component analysis, but recently some very powerful nonlinear dimension reduction methods have appeared [213, 214].

Supervised methods use data in the form of pairs (x, y) and estimate a function that predicts the value of y from a given input x. When y is a discrete quantity (for example a label of a number of distinct biological conditions) the method is called classification and when y is continuous the method is called regression. The key challenge is to ensure that the estimated function can generalize well to unseen situations [215]. Two methods are popular: (1) support vector machine (SVM) that estimates a discriminative function by maximizing class separation margin [216] and (2) random forest, based on tree ensembles and voting [217].

Semi-supervised methods combine ideas from supervised and unsupervised methods, to capture unsupervised structure in the data in order to boost classification performance [218].

Time series methods use data measured at different times to model and predict future values of the data, by capturing its structure and regularities and accounting for stochastic effects, e.g. with hidden Markov models (HMM) [219].

**Use cases**: A typical example is the classification of biological data such as gene expression data into different biological classes, e.g. disease and healthy, mostly using SVM and random forests. Prediction methods are also applied to pathway analysis, network decomposition and sequence annotation. They are often combined with a feature selection to extract the most relevant dimensions in the input data space [220].

### 1.4.4 Building Networks and Pathways from Literature

**Motivation**: Text mining joints efforts with the experimental sciences to help multifaceted disease-related research. Networks and connectivity maps are derived from text in an attempt to find connections and causal relations between components of complex biomedical systems, in order to elucidate disease mechanisms and detect co-morbidities [221, 222].

**Pre-processing**: Preparation of textual data consists of tokenization, removal of punctuation marks, part-of-speech tagging and sometimes syntactic parsing. Next, names of proteins, genes, chemicals, phenotypes and diseases are identified in the text. Management of biomedical terminology addresses several issues, such as appearance of new terms [221], heavy use of acronyms, abbreviations and general-purpose words that designate genes [223]. Synonymy and homonymy impose special challenges on the recognition process and complicate linking of a gene name to its unique identifier in the database [224, 225]. State-of-the-art named-entity recognition (NER) systems achieve F-measure of about 86 % [226] on biomedical corpus as opposed to 93 % on general purpose English texts [227].

**State-of-the-art**: Reconstruction of biological pathways from literature has evolved from undirected pairwise protein–protein co-occurrences [228] to complex biomedical events of typed and therefore directed interactions spanning multiple proteins [229–232]. The latter rely to a large extent on the richly annotated corpora, deep syntactic parsing and supervised machine learning techniques. Due to complexity of the natural language, accurate extraction of biomedical events remains a challenge. F-measure achieved by state-of-the-art systems varies from roughly 70–48 % depending largely on the event type being recognized.

**Use cases**: Many biomedical text-mining tools assist users at different stages of text processing, in particular for networks and pathways construction. Co-occurrence model has been successfully implemented in iHop, a hyperlinked network of genes and proteins mentioned in PubMed abstracts [233]. Facta + extends the pairwise co-occurrence model with event extraction and discovery of indirect associations between the biomedical concepts [222]. Based on PubMed abstracts, AliBaba builds networks of interacting proteins, genes—disease associations and subcellular location of proteins [234]. Networks extracted from text can be complemented with experimental data using IntAct [235] and CoPub [236].

### 1.4.5 Pathway Analysis

**Motivation**: Pathway analysis aims at identifying pathway deregulations to improve the understanding of complex phenotypes by leveraging information on known biomolecular interactions in pathways to guide the search through the space of possible functional associations. A wide range of methods exists, including enrichment analysis statistics, pathway-based disease gene prioritization methods,

convex metabolic pathway analysis and *in silico* pathway prediction/reconstruction methods [237].

**Pre-processing**: Because experimental measurement platforms and pathway databases tend to use different identifier formats, pathway analysis usually starts with the conversion of gene/protein names into a standard format [238–240], followed by normalisation and pre-processing of the experimental data.

**State-of-the-art**: Several novel approaches have recently been developed to infer changes in pathway activity from high-throughput data more accurately than by the classical combination of differential expression filtering with overrepresentation statistics like the Fisher exact test (for unordered datasets) or the Kolmogorov–Smirnov test (for ranked datasets). These include parametric and non-parametric approaches that take into account unfiltered gene expression level measurements, e.g. GSEA [241], PaGE [242], GAGE [243] or exploit information from ontology graphs, e.g. Ontologizer [244] and GeneCodis [245]. For the study of metabolic pathways, two related approaches using convex analysis have become increasingly important: Elementary flux modes [246] and extreme pathways [247]. Finally, as opposed to the classical human expert-based definition of pathways, various methods for pathway prediction/reconstruction using experimental data have been proposed recently [248, 249].

**Use cases**: Genome-wide pathway analyses have provided new insights on the aetiology of complex diseases that cannot be obtained from classical single-locus analyses [250]. Such analyses have indeed shown that different disruptions in a pathway can cause the same disease, as in colorectal cancer [251]. Metabolic pathway analysis is used in biomedical and biotechnological applications, e.g. to increase the production yield of microorganisms by metabolic engineering, i.e. the modification of selected pathways via recombinant DNA technologies [252]. Pathway analysis can also be integrated with network analysis to identify deregulated network modules in complex diseases [253].

## 1.5 Conclusions

Study of individual genes and their products in model systems has shifted to high-throughput studies in laboratories and often generated by large consortia. Each type of omic data is proving very valuable and their integration promises even greater rewards. Current techniques are very diverse and can analyse complex biological samples. They harbour high sensitivity and specificity, albeit not always sufficient, as in proteomics. Ongoing developments will increase accuracy, robustness, and flexibility while reducing cost. Current technical innovations continue shifting the bottleneck from data production to data analysis. Our understanding of biology will indeed increasingly rely on data and knowledge management, and informatics infrastructure to complement advances in mathematical and computational modelling for temporal and spatial analytical techniques, which are crucial to Systems Biology.

**Acknowledgments** This work was supported by the CNRS, the University of Luxembourg and the ISB, and in part by the EU grants to CA in the context of the MeDALL consortium (*Mechanisms of the Development of Allergy,* Grant Agreement FP7 N°264357) and the U-BIOPRED consortium (*Unbiased Biomarkers for the PREDiction of respiratory disease outcomes*, Grant Agreement IMI 115010).

# References

1. Gayon J, Malaterre C, Morange M, Raulin-Cerceau F, Tirard S (2010) Defining life: conference proceedings. Orig Life Evol Biosph 40(2):119–120
2. Westerhoff H, Hofmeyr J-H (2005) What is systems biology? From genes to function and back. In: Alberghina L, Westerhoff HV (eds) systems biology, vol 13. Springer, Berlin, pp 163–185
3. Kell DB (2006) Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discov Today 11(23–24):1085–1092
4. Lowe JA, Jones P, Wilson DM (2010) Network biology as a new approach to drug discovery. Curr Opin Drug Discov Devel 13(5):524–526
5. Pujol A, Mosca R, Farrés J, Aloy P (2010) Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci 31(3):115–123
6. Kitano H (2002) Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. Curr Genet 41(1):1–10
7. Auffray C, Imbeaud S, Roux-Rouquié M, Hood L (2003) From functional genomics to systems biology: concepts and practices. C R Biol 326(10–11):879–892
8. Van Regenmortel MHV (2004) Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism. EMBO Rep 5(11):1016–1020
9. Boogerd FC (2007) Systems biology: philosophical foundations. Elsevier, London, p 342
10. Wolkenhauer O (2001) Systems biology: the reincarnation of systems theory applied in biology? Brief Bioinf 2(3):258–270
11. Auffray C, Nottale L (2008) Scale relativity theory and integrative systems biology: 1. Founding principles and scale laws. Prog Biophys Mol Biol 97(1):79–114
12. Auffray C, Noble D (2009) Origins of systems biology in William Harvey's masterpiece on the movement of the heart and the blood in animals. Int J Mol Sci 10(4):1658–1669
13. Kohl P, Noble D (2009) Systems biology and the virtual physiological human. Mol Syst Biol 5:292
14. Westerhoff HV, Kolodkin A, Conradie R, Wilkinson SJ, Bruggeman FJ, Krab K, van Schuppen JH, Hardin H, Bakker BM, Moné MJ, Rybakova KN, Eijken M, van Leeuwen HJP, Snoep JL (2009) Systems biology towards life in silico: mathematics of the control of living cells. J Math Biol 58(1–2):7–34
15. Hornberg JJ, Bruggeman FJ, Westerhoff HV, Lankelma J (2006) Cancer: a systems biology disease. BioSystems 83(2–3):81–90
16. del Sol A, Balling R, Hood L, Galas D (2010) Diseases as network perturbations. Curr Opin Biotechnol 21(4):566–571
17. Westerhoff HV (2001) The silicon cell, not dead but live! Metab Eng 3(3):207–210
18. Kohl P, Crampin EJ, Quinn TA, Noble D (2010) Systems biology: an approach. Clin Pharmacol Ther 88(1):25–33
19. Hunter PJ, Borg TK (2003) Integration from proteins to organs: the physiome project. Nat Rev Mol Cell Biol 4(3):237–243
20. Hunter PJ, Crampin EJ, Nielsen PMF (2008) Bioinformatics, multiscale modeling and the IUPS physiome project. Brief Bioinf 9(4):333–343

21. Auffray C, Chen Z, Hood L (2009) Systems medicine: the future of medical genomics and healthcare. Genome Med 1(1):2
22. Hood L, Flores M (2012) A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. New Biotechnol 29:613
23. Horner DS, Pavesi G, Castrignanò T, De Meo PD, Liuni S, Sammeth M, Picardi E, Pesole G (2010) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinf 11(2):181–197
24. Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46
25. Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. J Genet Genomics 38(3):95–109
26. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12(10):671–682
27. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, James Kent W (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* 40(Database issue):D918–D923
28. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham L, Durbin R, Fernández-Suarez XM, Harrow J, Herrero J, Hubbard TJP, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SMJ (2012) Ensembl 2012. Nucleic Acids Res 40(Database issue):D84–D90
29. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6):415–425
30. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498
31. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073
32. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MMF, Knoppers BM, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SOM, Joly Y, Kato K, Kennedy KL, Nicolás P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, ···Zeps N, Lichter P, Biankin AV, Chabannon C, Chin L, Clément B, de Alava E, Degos F, Ferguson ML, Geary P, Hayes DN, Hudson TJ, Johns AL, Kasprzyk A, Nakagawa H, Penny R, Piris MA, Sarin R, Scarpa A, Shibata T, van de Vijver M, Futreal PA, Aburatani H, Bayés M, Botwell DDL, Campbell PJ, Estivill X, Gerhard DS, Grimmond SM, Gut I, Hirst M, López-Otín C, Majumder P, Marra M, McPherson JD, Nakagawa H, Ning Z, Puente XS, Ruan Y, Shibata T, Stratton MR, Stunnenberg HG, Swerdlow H, Velculescu VE, Wilson RK, Xue HH, Yang L, Spellman PT, Bader GD, Boutros PC, Campbell PJ, Flicek P, Getz G, Guigó R, Guo G, Haussler D, Heath S, Hubbard TJ, Jiang T, Jones SM, Li Q, López-Bigas N, Luo R, Muthuswamy L, Ouellette BFF, Pearson JV, Puente XS, Quesada V, Raphael BJ, Sander C, Shibata T, Speed TP, Stein LD, Stuart JM, Teague TW, Totoki Y, Tsunoda T, Valencia A, Wheeler DA, Wu H, Zhao S,

Zhou G, Stein LD, Guigó R, Hubbard TJ, Joly Y, Jones SM, Kasprzyk A, Lathrop M, López-Bigas N, Ouellette BFF, Spellman PT, Teague JW, Thomas G, Valencia A, Yoshida T, Kennedy KL, Axton M, Dyke SOM, Futreal PA, Gerhard DS, Gunter C, Guyer M, Hudson TJ, McPherson JD, Miller LJ, Ozenberger B, Shaw KM, Kasprzyk A, Stein LD, Zhang J, Haider SA, Wang J, Yung CK, Cros A, Cross A, Liang Y, Gnaneshan S, Guberman J, Hsu J, Bobrow M, Chalmers DRC, Hasel KW, Joly Y, Kaan TSH, Kennedy KL, Knoppers BM, Lowrance WW, Masui T, Nicolás P, Rial-Sebbag E, Rodriguez LL, Vergely C, Yoshida T, Grimmond SM, Biankin AV, Bowtell DDL, Cloonan N, deFazio A, Eshleman JR, Etemadmoghadam D, Gardiner BB, Gardiner BA, Kench JG, Scarpa A, Sutherland RL, Tempero MA, Waddell NJ, Wilson PJ, McPherson JD, Gallinger S, Tsao MS, Shaw PA, Petersen GM, Mukhopadhyay D, Chin L, DePinho RA, Thayer S, Muthuswamy L, Shazand K, Beck T, Sam M, Timms L, Ballin V, Lu Y, Ji J, Zhang X, Chen F, Hu X, Zhou G, Yang Q, Tian G, Zhang L, Xing X, Li X, Zhu Z, Yu Y, Yu J, Yang H, Lathrop M, Tost J, Brennan P, Holcatova I, Zaridze D, Brazma A, Egevard L, Prokhortchouk E, Banks RE, Uhlén M, Cambon-Thomsen A, Viksna J, Ponten F, Skryabin K, Stratton MR, Futreal PA, Birney E, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Martin S, Reis-Filho JS, Richardson AL, Sotiriou C, Stunnenberg HG, Thoms G, van de Vijver M, van't Veer L, Calvo F, Birnbaum D, Blanche H, Boucher P, Boyault S, Chabannon C, Gut I, Masson-Jacquemier JD, Lathrop M, Pauporté L, Pivot X, Vincent-Salomon A, Tabone E, Theillet C, Thomas G, Tost J, Treilleux I, Calvo F, Bioulac-Sage P, Clément B, Decaens T, Degos F, Franco D, Gut I, Gut M, Heath S, Lathrop M, Samuel D, Thomas G, Zucman-Rossi J, Lichter P, Eils R, Brors B, Korbel JO, Korshunov A, Landgraf P, Lehrach H, Pfister S, Radlwimmer B, Reifenberger G, Taylor MD, von Kalle C, Majumder PP, Sarin R, Rao TS, Bhan MK, Scarpa A, Pederzoli P, Lawlor RA, Delledonne M, Bardelli A, Biankin AV, Grimmond SM, Gress T, Klimstra D, Zamboni G, Shibata T, Nakamura Y, Nakagawa H, Kusada J, Tsunoda T, Miyano S, Aburatani H, Kato K, Fujimoto A, Yoshida T, Campo E, López-Otín C, Estivill X, Guigó R, de Sanjosé S, Piris MA, Montserrat E, González-Díaz M, Puente XS, Jares P, Valencia A, Himmelbauer H, Himmelbaue H, Quesada V, Bea S, Stratton MR, Futreal PA, Campbell PJ, Vincent-Salomon A, Richardson AL, Reis-Filho JS, van de Vijver M, Thomas G, Masson-Jacquemier JD, Aparicio S, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Stunnenberg HG, van't Veer L, Easton DF, Spellman PT, Martin S, Barker AD, Chin L, Collins FS, Compton CC, Ferguson ML, Gerhard DS, Getz G, Gunter C, Guttmacher A, Guyer M, Hayes DN, Lander ES, Ozenberger B, Penny R, Peterson J, Sander C, Shaw KM, Speed TP, Spellman PT, Vockley JG, Wheeler DA, Wilson RK, Hudson TJ, Chin L, Knoppers BM, Lander ES, Lichter P, Stein LD, Stratton MR, Anderson W, Barker AD, Bell C, Bobrow M, Burke W, Collins FS, Compton CC, DePinho RA, Easton DF, Futreal PA, Gerhard DS, Green AR, Guyer M, Hamilton SR, Hubbard TJ, Kallioniemi OP, Kennedy KL, Ley TJ, Liu ET, Lu Y, Majumder P, Marra M, Ozenberger B, Peterson J, Schafer AJ, Spellman PT, Stunnenberg HG, Wainwright BJ, Wilson RK, Yang H (2010) International network of cancer genome projects. Nature 464(7291):993–998

33. Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11(10):685–696
34. Pastinen T (2010) Genome-wide allele-specific analysis: insights into regulatory variation Nat Rev Genet 11(8):533–538
35. Thomas T, Gilbert J, Meyer F (2012) Metagenomics—a guide from sampling to data analysis. Microb Inf Exp 2(1):1–12
36. Virgin HW, Todd JA (2011) Metagenomics and personalized medicine. Cell 147(1):44–56
37. Desai N, Antonopoulos D, Gilbert JA, Glass EM, Meyer F (2012) From genomics to metagenomics. Curr Opin Biotechnol 23(1):72–76
38. Langmead B, Hansen KD, Leek JT (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol 11(8):R83
39. Stein LD (2010) The case for cloud computing in genome informatics. Genome Biol 11(5):207

40. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, Nelson K (2012) Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinf 13(1):42
41. Field D, Sansone S-A, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, Millard S, Mugabushaka A-M, Perrin N, Remacle JE, Remington K, Rocca-Serra P, Taylor CF, Thorley M, Tiwari B, Wilbanks J (2009) Megascience omics data sharing. Science 326(5950):234–236
42. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman LM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J (2011) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39(Database issue):D38–D51
43. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12(6):996–1006
44. Dunn WB (2011) Mass spectrometry in systems biology an introduction. Meth Enzymol 500:15–35
45. Hagen JB (2000) The origins of bioinformatics. Nat Rev Genet 1(3):231–236
46. Mount DR (2004) Bioinformatics: sequence and genome analysis. Cold Spring Harbor Laboratory Press, Second
47. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29
48. Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Res 11(8):425–1433
49. Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, Fragoso G, Game L, Heiskanen M, Morrison N, Rocca-Serra P, Sansone S-A, Taylor C, White J, Stoeckert CJ Jr (2006) The MGED ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 22(7):866–873
50. Csordas A, Ovelleiro D, Wang R, Foster JM, Ríos D, Vizcaíno JA, Hermjakob H (2012) PRIDE: quality control in a proteomics data repository. Database (Oxford) 2012:bas004
51. Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, Degreef J, Hardy N, Hermjakob H, Hubbard SJ, Hussey P, Igra M, Jenkins H, Julian RK Jr, Laursen K, Oliver SG, Paton NW, Sansone S-A, Sarkans U, Stoeckert CJ Jr, Taylor CF, Whetzel PL, White JA, Spellman P, Pizarro A (2007) The functional genomics experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol 25(10):1127–1133
52. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr J-H, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novère N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4):524–531
53. Novere NL, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villeger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H (2009) The systems biology graphical notation. Nat Biotech 27(8):735–741

54. Guberman JM, Ai J, Arnaiz O, Baran J, Blake A, Baldock R, Chelala C, Croft D, Cros A, Cutts RJ, Di Genova A, Forbes S, Fujisawa T, Gadaleta E, Goodstein DM, Gundem G, Haggarty B, Haider S, Hall M, Harris T, Haw R, Hu S, Hubbard S, Hsu J, Iyer V, Jones P, Katayama T, Kinsella R, Kong L, Lawson D, Liang Y, Lopez-Bigas N, Luo J, Lush M, Mason J, Moreews F, Ndegwa N, Oakley D, Perez-Llamas C, Primig M, Rivkin E, Rosanoff S, Shepherd R, Simon R, Skarnes B, Smedley D, Sperling L, Spooner W, Stevenson P, Stone K, Teague J, Wang J, Wang J, Whitty B, Wong DT, Wong-Erasmus M, Yao L, Youens-Clark K, Yung C, Zhang J, Kasprzyk A (2011) BioMart central portal: an open database network for the biological community. Database 2011:bar041
55. Zhang J, Haider S, Baran J, Cros A, Guberman JM, Hsu J, Liang Y, Yao L, Kasprzyk A (2011) BioMart: a data federation framework for large collaborative projects. Database (Oxford) 2011:bar038
56. Perakslis ED, Van Dam J, Szalma S (2010) How informatics can potentiate precompetitive open-source collaboration to jump-start drug discovery and development. Clin Pharmacol Ther 87(5):614–616
57. Maier D, Kalus W, Wolff M, Kalko SG, Roca J, Marin de Mas I, Turan N, Cascante M, Falciani F, Hernandez M, Villà-Freixa J, Losko S (2011) Knowledge management for systems biology a general and visually driven framework applied to translational medicine. BMC Syst Biol 5:38
58. Ghosh S, Matsuoka Y, Asai Y, Hsin K-Y, Kitano H (2011) Software for systems biology: from tools to integrated platforms. Nat Rev Genet 12(12):821–832
59. Kupershmidt I, Su QJ, Grewal A, Sundaresh S, Halperin I, Flynn J, Shekar M, Wang H, Park J, Cui W, Wall GD, Wisotzkey R, Alag S, Akhtari S, Ronaghi M (2010) Ontology-based meta-analysis of global collections of high-throughput public data. PLoS ONE 5(9):e13066
60. Huang Y-W, Huang TH-M, Wang L-S (2010) Profiling DNA methylomes from microarray to genome-scale sequencing. Technol Cancer Res Treat 9(2):139–147
61. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet 11(3):191–203
62. Rakyan VK, Down TA, Balding DJ, Beck S (2011) Epigenome-wide association studies for common human diseases. Nat Rev Genet 12(8):529–541
63. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SCJ, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung W-K, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei C-L, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D,

Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CWH, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JNS, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PIW, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrímsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VVB, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146):799–816

64. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA (2010) The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 28(10):1045–1048
65. Bradbury J (2003) Human epigenome project–up and running. PLoS Biol 1(3):E82
66. Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox AV, Andrews TD, Howe KL, Otto T, Olek A, Fischer J, Gut IG, Berlin K, Beck S (2004) DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. PLoS Biol 2(12):e405
67. Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, Bock C, Boehm B, Campo E, Caricasole A, Dahl F, Dermitzakis ET, Enver T, Esteller M, Estivill X, Ferguson-Smith A, Fitzgibbon J, Flicek P, Giehl C, Graf T, Grosveld F, Guigo R, Gut I, Helin K, Jarvius J, Kuppers R, Lehrach H, Lengauer T, Lernmark A, Leslie D, Loeffler M, Macintyre E, Mai A, Martens JH, Minucci S, Ouwehand WH, Pelicci PG, Pendeville H, Porse B, Rakyan V, Reik W, Schrappe M, Schubeler D, Seifert M, Siebert R, Simmons D, Soranzo N, Spicuglia S, Stratton M, Stunnenberg HG, Tanay A, Torrents D, Valencia A, Vellenga E, Vingron M, Walter J, Willcocks S (2012) BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotech 30(3):224–226
68. Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Jeddeloh JA, Wen B, Feinberg AP (2008) Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Res 18(5):780–790
69. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, Yamamoto H, Hidalgo M, Tan A-C, Galm O, Ferrer I, Sanchez-Cespedes M, Villanueva A, Carmona J, Sanchez-Mut JV, Berdasco M, Moreno V, Capella G, Monk D, Ballestar E, Ropero S, Martinez R, Sanchez-Carbayo M, Prosper F, Agirre X, Fraga MF, Graña O, Perez-Jurado L, Mora J, Puig S, Prat J, Badimon L, Puca AA, Meltzer SJ, Lengauer T, Bridgewater J, Bock C, Esteller M (2011) A DNA methylation fingerprint of 1628 human samples. Genome Res 22:407
70. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics 6(6):692–702
71. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M,

Wiggin M, Schloss JA (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26(10):1146–1153
72. Farnham PJ (2009) Insights from genomic profiling of transcription factors. Nat Rev Genet 10(9):605–616
73. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10(10):669–680
74. Serre D, Lee BH, Ting AH (2010) MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res 38(2):391–399
75. Macisaac KD, Fraenkel E (2010) Sequence analysis of chromatin immunoprecipitation data for transcription factors. Methods Mol Biol 674:179–193
76. Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12(1):7–18
77. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods 9:473
78. Korolev N, Fan Y, Lyubartsev AP, Nordenskiöld L (2012) Modelling chromatin structure and dynamics: status and prospects. Curr Opin Struct Biol 22(2):151–159
79. Geng T, Bao N, Litt MD, Glaros TG, Li L, Lu C (2011) Histone modification analysis by chromatin immunoprecipitation from a low number of cells on a microfluidic platform. Lab Chip 11(17):2842–2848
80. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, Jäger N, Gnirke A, Stunnenberg HG, Meissner A (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol 28(10):1106–1114
81. Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147(6):1408–1419
82. Muro EM, McCann JA, Rudnicki MA, Andrade-Navarro MA (2009) Use of SNP-arrays for ChIP assays: computational aspects. Methods Mol Biol 567:145–154
83. Siegmund KD (2011) Statistical approaches for the analysis of DNA methylation microarray data. Hum Genet 129(6):585–595
84. Sun S, Huang Y-W, Yan PS, Huang TH, Lin S (2011) Preprocessing differential methylation hybridization microarray data. BioData Min 4:13
85. Huss M (2010) Introduction into the analysis of high-throughput-sequencing based epigenome data. Brief Bioinf 11(5):512–523
86. Massie CE, Mills IG (2012) Mapping protein-DNA interactions using ChIP-sequencing. Methods Mol Biol 809:157–173
87. Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res 21(3):456–464
88. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. Genome Res 17(6):877–885
89. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrançois P, Struhl K, Gerstein M, Snyder M (2009) Mapping accessible chromatin regions using sono-seq. Proc Natl Acad Sci USA 106(35):14926–14931
90. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. Science 295(5558):1306–1311
91. Dean A (2011) In the loop: long range chromatin interactions and gene regulation. Brief Funct Genomics 10(1):3–10
92. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38(11):1348–1354
93. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J (2006) Chromosome conformation capture

carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16(10):1299–1309
94. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950):289–293
95. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EGY, Huang PYH, Welboren W-J, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KDSA, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RKM, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung W-K, Liu ET, Wei C-L, Cheung E, Ruan Y (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462(7269):58–64
96. Deal RB, Henikoff JG, Henikoff S (2010) Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. Science 328(5982):1161–1164
97. Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132(5):887–898
98. Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP (2003) In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. Nat Genet 33(4):469–475
99. McDaniell R, Lee B-K, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, Keefe D, Collins FS, Willard HF, Lieb JD, Furey TS, Crawford GE, Iyer VR, Birney E (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. Science 328(5975):235–239
100. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong M-Y, Karczewski KJ, Huber W, Weissman SM, Gerstein MB, Korbel JO, Snyder M (2010) Variation in transcription factor binding among humans. Science 328(5975):232–235
101. Jeong J, Li L, Liu Y, Nephew KP, Huang TH-M, Shen C (2010) An empirical Bayes model for gene expression and methylation profiles in antiestrogen resistant breast cancer. BMC Med Genomics 3:55
102. Loss LA, Sadanandam A, Durinck S, Nautiyal S, Flaucher D, Carlton VEH, Moorhead M, Lu Y, Gray JW, Faham M, Spellman P, Parvin B (2010) Prediction of epigenetically regulated genes in breast cancer cell lines. BMC Bioinf 11:305
103. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639–1645
104. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone S-A (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31(1):68–71
105. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A (2010) Gene expression atlas at the European Bioinformatics Institute. Nucleic Acids Res 38(Database issue):D690–D698
106. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim LF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A (2011) NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Res 39(Database issue):D1005–D1010
107. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29(4):365–371

108. Malone JH, Oliver B (2011) Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 9:34
109. Posekany A, Felsenstein K, Sykacek P (2011) Biological assessment of robust noise models in microarray data analysis. Bioinformatics 27(6):807–814
110. Kadota K, Shimizu K (2011) Evaluating methods for ranking differentially expressed genes applied to microarray quality control data. BMC Bioinf 12:227
111. Gibb EA, Vucic EA, Enfield KSS, Stewart GL, Lonergan KM, Kennett JY, Becker-Santos DD, MacAulay CE, Lam S, Brown CJ, Lam WL (2011) Human cancer long non-coding RNA transcriptomes. PLoS ONE 6(10):e25915
112. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10(3):155–159
113. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63
114. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464(7289):768–772
115. Hansen KD, Wu Z, Irizarry RA, Leek JT (2011) Sequencing technology does not eliminate biological variability. Nat Biotechnol 29(7):572–573
116. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol 30(1):99–104
117. Raz T, Kapranov P, Lipson D, Letovsky S, Milos PM, Thompson JF (2011) Protocol dependence of sequencing-based gene expression measurements. PLoS ONE 6(5):e19287
118. Schwartz S, Oren R, Ast G (2011) Detection and removal of biases in the analysis of next-generation sequencing reads. PLoS ONE 6(1):e16685
119. Brazma A (2009) Minimum information about a microarray experiment (MIAME)–successes, failures, challenges. Sci World J 9:420–423
120. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinf 11:94
121. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T, Ratcliffe AL, Webster PJ, Davidson EH, Hood L, Dimitrov K (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotech 26(3):317–325
122. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 460(7254):479–486
123. Jacquier A (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet 10(12):833–844
124. Licatalosi DD, Darnell RB (2010) RNA processing and its regulation: global insights into biological networks. Nat Rev Genet 11(1):75–87
125. Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 13(4):227–232
126. Beck M, Schmidt A, Malmstroem A, Claassen M, Ori A, Szymborska A, Herzog F, Rinner O, Ellenberg J, Aebersold R (2011) The quantitative proteome of a human cell line. Mol Syst Biol 7:549, ISSN:1744-4292. doi:10.1038/msb.2011.82, URL:http://www.ncbi.nlm.nih.gov/pubmed/22068332. Accessed 15 Apr 2012
127. DeLuca DS, Marina O, Ray S, Zhang GL, Wu CJ, Brusic V (2011) Data processing and analysis for protein microarrays. Methods Mol Biol 723:337–347
128. Ray S, Reddy PJ, Choudhary S, Raghu D, Srivastava S (2011) Emerging nanoproteomics approaches for disease biomarker detection: a current perspective. J Proteomics 74(12):2660–2681
129. Domon B, Aebersold R (2010) Options and considerations when selecting a quantitative proteomics strategy. Nat Biotech 28(7):710–721
130. Mallick P, Kuster B (2010) Proteomics: a pragmatic perspective. Nat Biotechnol 28(7):695–709

131. Kito K, Ito T (2008) Mass spectrometry-based approaches toward absolute quantitative proteomics. Curr Genomics 9(4):263–274
132. Diao L, Clarke CH, Coombes KR, Hamilton SR, Roth J, Mao L, Czerniak B, Baggerly KA, Morris JS, Fung ET, Bast RC Jr (2011) Reproducibility of SELDI spectra across time and laboratories. Cancer Inform 10:45–64
133. Matthiesen R, Azevedo L, Amorim A, Carvalho AS (2011) Discussion on common data analysis strategies used in MS-based proteomics. Proteomics 11(4):604–619
134. Käll L, Vitek O (2011) Computational mass spectrometry-based proteomics. PLoS Comput Biol 7(12):e1002277
135. Taylor CF (2006) Minimum reporting requirements for proteomics: a MIAPE primer. Proteomics 6(Suppl 2):39–44
136. Jacob RJ (2010) Bioinformatics for LC-MS/MS-based proteomics. Methods Mol Biol 658:61–91
137. Walhout AJ, Vidal M (2001) Protein interaction maps for model organisms. Nat Rev Mol Cell Biol 2(1):55–62
138. Rinner O, Mueller LN, Hubalek M, Muller M, Gstaiger M, Aebersold R (2007) An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. Nat Biotech 25(3):345–352
139. Hutchins JRA, Toyoda Y, Hegemann B, Poser I, Hériché J-K, Sykora MM, Augsburg M, Hudecz O, Buschhorn BA, Bulkescher J, Conrad C, Comartin D, Schleiffer A, Sarov M, Pozniakovsky A, Slabicki MM, Schloissnig S, Steinmacher I, Leuschner M, Ssykor A, Lawo S, Pelletier L, Stark H, Nasmyth K, Ellenberg J, Durbin R, Buchholz F, Mechtler K, Hyman AA, Peters J-M (2010) Systematic analysis of human protein complexes identifies chromosome segregation proteins. Science 328(5978):593–599
140. Deutsch EW, Shteynberg D, Lam H, Sun Z, Eng JK, Carapito C, von Haller PD, Tasman N, Mendoza L, Farrah T, Aebersold R (2010) Trans-Proteomic Pipeline supports and improves analysis of electron transfer dissociation data sets. Proteomics 10(6):1190–1195
141. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Bjorling L, Ponten F (2010) Towards a knowledge-based human protein atlas. Nat Biotech 28(12):1248–1250
142. Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, Gateau A, Gaudet P, Gleizes A, Masselot A, Zwahlen C, Bairoch A (2012) neXtProt: a knowledge platform for human proteins. Nucleic Acids Res 40(Database issue):D76–D83
143. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, Sinelnikov I, Krishnamurthy R, Eisner R, Gautam B, Young N, Xia J, Knox C, Dong E, Huang P, Hollander Z, Pedersen TL, Smith SR, Bamforth F, Greiner R, McManus B, Newman JW, Goodfriend T, Wishart DS (2011) The human serum metabolome. PLoS ONE 6(2):e16957
144. Beckonert O, Keun HC, Ebbels TMD, Bundy J, Holmes E, Lindon JC, Nicholson JK (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nat Protoc 2(11):2692–2703
145. Lee DY, Bowen BP, Northen TR (2010) Mass spectrometry-based metabolomics, analysis of metabolite-protein interactions, and imaging. Biotechniques 49(2):557–565
146. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30
147. Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. Methods Mol Biol 802:19–39
148. Sinha TK, Khatib-Shahidi S, Yankeelov TE, Mapara K, Ehtesham M, Cornett DS, Dawant BM, Caprioli RM, Gore JC (2008) Integrating spatially resolved three-dimensional MALDI IMS with in vivo magnetic resonance imaging. Nat Meth 5(1):57–59
149. Griffin JL, Steinbeck C (2010) 'So what have data standards ever done for us? The view from metabolomics. Genome Med 2(6):38

150. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M (2012) Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. Curr Bioinf 7(1):96–108
151. Carroll AJ, Badger MR, Harvey Millar A (2010) The metabolomeexpress project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. BMC Bioinf 11:376
152. Kastenmüller G, Römisch-Margl W, Wägele B, Altmaier E, Suhre K (2011) metaP-server: a web-based metabolomics data analysis tool. J Biomed Biotechnol 2011
153. Gika HG, Theodoridis GA, Earll M, Snyder RW, Sumner SJ, Wilson ID (2010) Does the mass spectrometer define the marker? A comparison of global metabolite profiling data generated simultaneously via UPLC-MS on two different mass spectrometers. Anal Chem 82(19):8226–8234
154. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia J, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P, Fang L, Peng J, Fradette R, Cheng D, Tzur D, Clements M, Lewis A, De Souza A, Zuniga A, Dawe M, Xiong Y, Clive D, Greiner R, Nazyrova A, Shaykhutdinov R, Li L, Vogel HJ, Forsythe I (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37(Database issue):D603–D610
155. Yu Z, Kastenmüller G, He Y, Belcredi P, Möller G, Prehn C, Mendes J, Wahl S, Roemisch-Margl W, Ceglarek U, Polonikov A, Dahmen N, Prokisch H, Xie L, Li Y, Wichmann H-E, Peters A, Kronenberg F, Suhre K, Adamski J, Illig T, Wang-Sattler R (2011) Differences between human plasma and serum metabolite profiles. PLoS ONE 6(7):e21230
156. Wymann MP, Schneiter R (2008) Lipid signalling in disease. Nat Rev Mol Cell Biol 9(2):162–176
157. van Meer G, Voelker DR, Feigenson GW (2008) Membrane lipids: where they are and how they behave. Nat Rev Mol Cell Biol 9(2):112–124
158. Shevchenko A. Simons K (2010) Lipidomics: coming to grips with lipid diversity. Nat Rev Mol Cell Biol 11(8):593–598
159. Quehenberger O, Armando AM, Brown AH, Milne SB, Myers DS, Merrill AH, Bandyopadhyay S, Jones KN, Kelly S, Shaner RL, Sullards CM, Wang E, Murphy RC, Barkley RM, Leiker TJ, Raetz CRH, Guan Z, Laird GM, Six DA, Russell DW, McDonald JG, Subramaniam S, Fahy E, Dennis EA (2010) Lipidomics reveals a remarkable diversity of lipids in human plasma. J Lipid Res 51(11):3299–3305
160. Han X, Gross RW (2003) Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. J Lipid Res 44(6):1071–1079
161. Jung HR, Sylvänne T, Koistinen KM, Tarasov K, Kauhanen D, Ekroos K (2011) High throughput quantitative molecular lipidomics. Biochim Biophys Acta 1811(11):925–934
162. Chaurand P, Cornett DS, Angel PM, Caprioli RM (2011) From whole-body sections down to cellular level, multiscale imaging of phospholipids by MALDI mass spectrometry. Mol Cell Proteomics 10(2):O110.004259
163. Nordström A, O'Maille G, Qin C, Siuzdak G (2006) Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. Anal Chem 78(10):3289–3295
164. Pluskal T, Castillo S, Villar-Briones A, Oresic M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf 11:395
165. Orešič M (2011) Informatics and computational strategies for the study of lipids. Biochim Biophys Acta 1811(11):991–999
166. Schmelzer K, Fahy E, Subramaniam S, Dennis EA (2007) The lipid maps initiative in lipidomics. Meth Enzymol 432:171–183
167. Dennis EA, Deems RA, Harkewicz R, Quehenberger O, Brown HA, Milne SB, Myers DS, Glass CK, Hardiman G, Reichart D, Merrill AH Jr, Sullards MC, Wang E, Murphy RC, Raetz CRH, Garrett TA, Guan Z, Ryan AC, Russell DW, McDonald JG, Thompson BM,

Shaw WA, Sud M, Zhao Y, Gupta S, Maurya MR, Fahy E, Subramaniam S (2010) A mouse macrophage lipidome. J Biol Chem 285(51):39976–39985
168. Niemelä PS, Castillo S, Sysi-Aho M, Oresic M (2009) Bioinformatics and computational methods for lipidomics. J Chromatogr B Analyt Technol Biomed Life Sci 877(26):2855–2862
169. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin A-C (2010) Visualization of omics data for systems biology. Nat Methods 7(3 Suppl):S56–S68
170. Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11(7):476–486
171. Glaab E (2011) Analysing functional genomics data using novel ensemble,consensus and data fusion techniques. University of Nottingham, Nottingham
172. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2):249–264
173. Mueller LN, Brusniak M-Y, Mani DR, Aebersold R (2008) An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res 7(1):51–61
174. Castillo S, Gopalacharyulu P, Yetukuri L, Orešič M (2011) Algorithms and tools for the preprocessing of LC–MS metabolomics data. Chemometrics Intell Lab Syst 108(1):23–32
175. Tritchler D, Parkhomenko E, Beyene J (2009) Filtering genes for cluster and network analysis. BMC Bioinf 10:193
176. Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinf 1(1):24–45
177. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tarraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, Garcia F, Marba M, Montaner D, Dopazo J (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res 38(Web Server):W210–W213
178. Vallon-Christersson J, Nordborg N, Svensson M, Häkkinen J (2009) BASE—2nd generation software for microarray data management and analysis. BMC Bioinf 10:330
179. Naegle KM, Welsch RE, Yaffe MB, White FM, Lauffenburger DA (2011) MCAM: multiple clustering analysis methodology for deriving hypotheses and insights from high-throughput proteomic datasets. PLoS Comput Biol 7(7):e1002119
180. Chaiboonchoe A, Samarasinghe S, Kulasiri D (2009) Machine learning for childhood acute lymphoblastic leukaemia gene expression data analysis: a review. Curr Bioinform 5(2):118–133
181. Schummer M, Green A, Beatty JD, Karlan BY, Karlan S, Gross J, Thornton S, McIntosh M, Urban N (2010) Comparison of breast cancer to healthy control tissue discovers novel markers with potential for prognosis and early detection. PLoS ONE 5(2):e9122
182. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, D'Agostino R, Castro M, Curran-Everett D, Fitzpatrick AM, Gaston B, Jarjour NN, Sorkness R, Calhoun WJ, Chung KF, Comhair SAA, Dweik RA, Israel E, Peters SP, Busse WW, Erzurum SC, Bleecker ER (2010) Identification of asthma phenotypes using cluster analysis in the severe asthma research program. Am J Respir Crit Care Med 181(4):315–323
183. Just J, Gouvis-Echraghi R, Rouve S, Wanin S, Moreau D, Annesi Maesano I (2012) Two novel severe asthma phenotypes identified during childhood using a clustering approach. Official Journal of the European Society for Clinical Respiratory Physiology, The European Respiratory Journal
184. Bjornsdottir US, Holgate ST, Reddy PS, Hill AA, McKee CM, Csimma CI, Weaver AA, Legault HM, Small CG, Ramsey RC, Ellis DK, Burke CM, Thompson PJ, Howarth PH, Wardlaw AJ, Bardin PG, Bernstein DI, Irving LB, Chupp GL, Bensch GW, Bensch GW, Stahlman JE, Karetzky M, Baker JW, Miller RL, Goodman BH, Raible DG, Goldman SJ, Miller DK, Ryan JL, Dorner AJ, Immermann FW, O'Toole M (2011) Pathways activated

during human asthma exacerbation as revealed by gene expression patterns in blood. PLoS ONE 6(7):e21902
185. Bellman R (1961) Adaptive control processes. Princeston University Press, Princeston
186. Kusiak A (2001) Feature transformation methods in data mining. IEEE Trans Electron Packaging Manuf 24(3):214–221
187. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. JMachine Learn Res 3:1157–1182
188. Dy JG (2008) Unsupervised feature selection. Comput Methods Feature Sel 2008:19–39
189. Tsamardinos I, Aliferis CF (2003) Towards principled feature selection: relevancy, filters and wrappers. In: Proceedings of the 9th international workshop on artificial intelligence and statistics, 2003
190. Bozinov D, Rahnenführer J (2002) Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering. Bioinformatics 18(5):747
191. Katzer M, Kummert F, Sagerer G (2003) Methods for automatic microarray image segmentation. IEEE Transactions on NanoBiosci 2(4):202–214
192. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2):185–193
193. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci 98(1):31
194. Lazaridis EN, Sinibaldi D, Bloom G, Mane S, Jove R (2002) A simple method to improve probe set estimates from oligonucleotide arrays. Math Biosci 176(1):53–58
195. Shakya K, Ruskin H, Kerr G, Crane M, Becker J (2010) Comparison of microarray preprocessing methods. In: Arabina HR (ed) Advances in computational biology. Springer, New York, pp 139–147
196. Katajamaa M, Oresic M (2007) Data processing for mass spectrometry-based metabolomics. J Chromatogr A 1158(1–2):318–328
197. Hastings CA, Norton SM, Roy S (2002) New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. Rapid Commun Mass Spectrom 16(5):462–467
198. Smith CA, Elizabeth J, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78(3):779–787
199. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 389(4):1017–1031
200. Hu Q, Pan W, An S, Ma P, Wei J (2010) An efficient gene selection technique for cancer recognition based on neighborhood mutual information. International Journal of Machine Learning and, Cybernetics, pp 1–12
201. Hall MA (1999) Correlation-based feature selection for machine learning. The University of Waikato, Waikato
202. Wu Y, Zhang A (2004) Feature selection for classifying high-dimensional numerical data. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition (CVPR 2004), vol. 2, p 251
203. Xing EP, Jordan MI, Karp RM, et al. (2001) Feature selection for high-dimensional genomic microarray data. In: Machine learning-international workshop then conference, 2001, pp 601–608
204. Li L, Pedersen LG, Darden TA, Weinberg CR (2002) Computational analysis of leukemia microarray expression data using the GA/KNN method. In: Methods of microarray data analysis: papers from CAMDA'00, pp 81–95
205. Blanco R, Larrañaga P, Inza I, Sierra B (2001) Selection of highly accurate genes for cancer classification by estimation of distribution algorithms. In: Workshop of Bayesian models in medicine. AIME 2001, pp 29–34

206. Inza I, Sierra B, Blanco R, Larrañaga P (2002) Gene selection by sequential search wrapper approaches in microarray cancer class prediction. J Intell Fuzzy Syst 12(1):25–33
207. Liu J, Iba H, Ishizuka M (2001) Selecting informative genes with parallel genetic algorithms in tissue classification. Genome Inform 12:14–23
208. Chen YW, Lin CJ (2006) Combining SVMs with various feature selection strategies. In: Isabella G, Andre E (eds) Feature extraction, Springer, Berlin, pp 315–324
209. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2011) An ensemble of filters and classifiers for microarray data classification. Pattern Recogn 45:531
210. Sun M, Xiong M (2003) A mathematical programming approach for gene selection and tissue classification. Bioinformatics 19(10):1243
211. Subramani P, Sahu R, Verma S (2006) Feature selection using haar wavelet power spectrum. BMC Bioinf 7(1):432
212. Koller D, Friedman N (2009) Probabilistic graphical models principles and techniques. MIT press, Cambridge
213. Lee JA, Verleysen M (2007) Nonlinear dimensionality reduction, 1st edn. Springer, Berlin
214. Maaten LVD, Hinton G (2008) Visualizing data using t-SNE. J Machine Learn Res 9(2579–2605):2579–2605
215. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Data mining, inference, and prediction. Springer, New York
216. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G (2008) Support vector machines and kernels for computational biology. PLoS Comput Biol 4(10):e1000173
217. Breiman L, Schapire E (2001) Random forests. Machine Learn 45:5–32
218. Chapelle O, Schölkopf B, Zien A (2010) Semi-supervised learning. MIT Press, Cambridge
219. Koski T (2002) Hidden Markov models of bioinformatics, 1st ed. Springer, Berlin
220. Bonneau R, Facciotti MT, Reiss DJ, Schmid AK, Pan M, Kaur A, Thorsson V, Shannon P, Johnson MH, Bare JC, Longabaugh W, Vuthoori M, Whitehead K, Madar A, Suzuki L, Mori T, Chang D-E, DiRuggiero J, Johnson CH, Hood L, Baliga NS (2007) A predictive model for transcriptional control of physiology in a free living cell. Cell 131(7):1354–1365
221. Krallinger M, Valencia A, Hirschman L (2008) Linking genes to literature: text mining information extraction, and retrieval applications for biology. Genome Biol 9(Suppl 2):S8
222. Tsuruoka Y, Miwa M, Hamamoto K, Tsujii J, Ananiadou S (2011) Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics 27(13):i111
223. Roche M, Prince V (2010) A web-mining approach to disambiguate biomedical acronym expansions. Informatica (Slovenia) 34(2):243–253
224. Hakenberg J, Plake C, Royer L, Strobelt H, Leser U, Schroeder M (2008) Gene normalization and interaction with context and sentence motifs. Genome Biol 9(Suppl 2):S14
225. Seringhaus MB, Cayting PD, Gerstein MB (2008) Uncovering trends in gene naming. Genome Biol 9(1):401
226. Leaman R, Gonzalez G (2008) BANNER: an executable survey of advances in biomedical named entity recognition. In: Pacific symposium on biocomputing, pp 652–663
227. Marsh E, Perzanowski D (1998) MUC-7 evaluation of IE technology: overview of results. In: Proceedings of the 7th message understanding conference (MUC-7)
228. Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A (2005) Text mining for metabolic pathways, signaling cascades, and protein networks. Sci STKE 2005(283):e21
229. Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii J (2011) Extracting bio-molecular events from literature—the BioNLP'09 shared task. Comput Intell 27(4):513–540
230. Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T (2010) Complex event extraction at PubMed scale. Bioinformatics [ISMB] 26(12):382–390
231. McClosky D, Surdeanu M, Manning CD (2011) Event extraction as dependency parsing. In: ACL 2011, pp 1626–1635
232. Riedel S, McCallum A (2011) Fast and robust joint models for biomedical event extraction. In: EMNLP 2011, pp 1–12

233. Hoffmann R, Valencia A (2005) Implementing the iHOP concept for navigation of biomedical literature. In: ECCB/JBI 2005, p 258
234. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U (2006) ALIBABA: PubMed as a graph. Bioinformatics 22(19):2444–2445
235. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A et al (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res 32(suppl 1):D452–D455
236. Fleuren WWM, Verhoeven S, Frijters R, Heupers B, Polman J, , van Schaik R, de Vlieg J, Alkema W (2011) CoPub update: CoPub 5.0 a text mining system to answer biological questions. Nucleic Acids Res 39(Web Server):W450–W454
237. Ramanan VK, Shen L, Moore JH, Saykin AJ (2012) Pathway analysis of genomic data: concepts, methods and prospects for future development. Trends Genet 28(7): 323–332, ISSN:0168-9525. doi:10.1016/j.tig.2012.03.004
238. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: database for annotation. Visualization, and integrated discovery. Genome Biol 4(5):P3
239. Alibés A, Yankilevich P et al (2007) IDconverter and IDClight: conversion and annotation of gene and protein IDs'. BMC Bioinf 8(1):9
240. Durinck S, Spellman PT, Birney E, Huber W (2009) Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomart. Nat Protoc 4(8):1184–1191
241. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545
242. Kim SY, Volsky D (2005) PAGE: parametric analysis of gene set enrichment. BMC Bioinf 6(1):144
243. Luo W, Friedman M, Shedden K, Hankenson K, Woolf P (2009) GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinf 10(1):161
244. Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24(14):1650
245. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol 8(1):R3
246. Schuster S, Hlgetag C (1994) On elementary flux modes in biochemical reaction systems at steady state. J Biol Syst 2(2):165–182
247. Schilling CH, Letscher D, Palsson BO (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J Theoretical Biol 203(3):229–248
248. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol 23(5):561–566
249. Ma X, Tarone AM, Li W (2008) Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. PLoS ONE 3(4):e1922
250. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol 33(5):419–431
251. Markowitz SD, Bertagnolli MM (2009) Molecular basis of colorectal cancer. N Engl J Med 361(25):2449–2460
252. Xu X, Cao L, Chen X (2008) Elementary flux mode analysis for optimized ethanol yield in anaerobic fermentation of glucose with Saccharomyces cerevisiae. Chin J Chem Eng 16(1):135–142
253. Cerami E, Demir E, Schultz N, Taylor BS, Sander C (2010) Automated network analysis identifies core pathways in glioblastoma. PLoS ONE 5(2):e8918