Principles and methods of integrative genomic analyses in cancer

Vessela N. Kristensen[1–3], Ole Christian Lingjærde[2,4], Hege G. Russnes[1,2,5], Hans Kristian M. Vollan[1,2,6], Arnoldo Frigessi[7,8] and Anne-Lise Børresen-Dale[1,2]

Abstract | Combined analyses of molecular data, such as DNA copy-number alteration, mRNA and protein expression, point to biological functions and molecular pathways being deregulated in multiple cancers. Genomic, metabolomic and clinical data from various solid cancers and model systems are emerging and can be used to identify novel patient subgroups for tailored therapy and monitoring. The integrative genomics methodologies that are used to interpret these data require expertise in different disciplines, such as biology, medicine, mathematics, statistics and bioinformatics, and they can seem daunting. The objectives, methods and computational tools of integrative genomics that are available to date are reviewed here, as is their implementation in cancer research.

[1]*Department of Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway.*
[2]*K.G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, 0313 Oslo, Norway.*
[3]*Department of Clinical Molecular Oncology, Division of Medicine, Akershus University Hospital, 1478 Ahus, Norway.*
[4]*Division for Biomedical Informatics, Department of Computer Science, University of Oslo, 0316 Oslo, Norway.*
[5]*Department of Pathology, Oslo University Hospital, 0450 Oslo, Norway.*
[6]*Department of Oncology, Division of Cancer, Surgery and Transplantation, Oslo University Hospital, 0450 Oslo, Norway.*
[7]*Statistics for Innovation, Norwegian Computing Center, 0314 Oslo, Norway.*
[8]*Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, PO Box 1122 Blindern, 0317 Oslo, Norway.*
Correspondence to A.-L.B.-D.
e-mail: a.l.borresen-dale@medisin.uio.no
doi:10.1038/nrc3721

Leonardo da Vinci's collection of drawings entitled *Studies of the Human Body and Principles of Anatomy* started a renaissance in studying human anatomy and pathology that led to a better understanding of the mechanics, proportions and functions of the human body. Today, we live in an era when biological sciences are marked by the same exploratory drive, but this time it is at an invisible, molecular level. The accumulation of enormous quantities of molecular data has led to the emergence of 'systems biology' — a branch of science that discovers the principles that underlie the basic functional properties of living organisms, starting from interactions between macromolecules[1–4]. Integrative genomics is based on the fundamental principle that any biological mechanism builds upon multiple molecular phenomena, and only through the understanding of the interplay within and between different layers of genomic structures can one attempt to fully understand phenotypic traits. Therefore, principles of integrative genomics are based on the study of molecular events at different levels and on the attempt to integrate their effects in a functional or causal framework. Although perhaps less aesthetically pleasing than the drawings of Leonardo da Vinci, the new visualization tools based on mathematical models can present the 'digital universe of information' in a form that is of use for the treatment of a cancer patient[5] and for revealing the existence and principles of molecular interactions that govern fundamental biological mechanisms[6].

Cancer is currently one of the most well-characterized pathological systems at the molecular level. Most (if not all) cancers involve genetic aberrations in the germ line and/or at the somatic level. By producing a complete catalogue of inherited and acquired mutations, with functional consequences of each mutation with respect to tumour type, it is hoped that one can, for example, assess the metastatic potential of a tumour and suggest the most promising treatment[7,8]. Although data are rapidly accumulating from various cancer-profiling projects, interpreting these data is not easy. The development and progression of a tumour is a dynamic biological and evolutionary process. It involves composite organ systems, with genomes shaped by gene aberrations, epigenetic changes, the cellular biological context, characteristics that are specific to the individual patient, and environmental influences[9,10]. Sophisticated statistical and mathematical techniques have been developed for the analysis, interpretation and validation of biological data, and novel computational techniques and tools are continuously emerging. In principle, mathematical modelling of pattern formation — using methods from interacting particle systems, system dynamics and hierarchical models — can be used to study tumour formation and growth. In practice, statistics and information theory constitute essential methodologies in the analysis of biological data sets. These methodologies are the subject of this Review. We discuss the different computational models

Key points

- Genomic, metabolomic and clinical data on a range of solid cancers and model systems are emerging and can be used to identify novel patient subgroups for tailored therapy and monitoring.
- Molecular markers identified at the DNA, mRNA, microRNA and protein levels have been used to develop profiles associated with taxonomy, tumour aggressiveness, response to therapy and patient outcome.
- The information content is higher in integrated analysis than in any of the molecular levels studied separately, and a large number of statistical methods for the integration of 'omics' data have emerged.
- The access to large data sets that have been made available by the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) has made it possible to compare the performance of some of the statistical methods of omic data integration on the same data set.
- These recent developments will fundamentally alter the way that we statistically model and evaluate treatment strategies, from identifying patient groups that respond to treatment above random, to identifying pathways and biological entities that are druggable and altered above random.
- A shift from large randomized clinical trials towards treatment modalities that are tailored for stratified patient groups, down to N-of-1 trials, in which a single patient constitutes the entire trial, will require new statistical methods.
- Outsourcing data and searching for solutions in open competition will allow new ideas to instantly emerge to 'embrace the complexity' that is associated with the exponentially increasing amounts of data and find new ways of shared analysis.

that are being used to assess these data and how these models have been applied to better understand cancer development, progression and treatment response.

Multi-dimensional molecular data sources

Continuous improvements in the rate, accuracy and resolution of 'omics' data and biochemical features that can be observed in a tumour or a patient have set the stage for the integration of many sources of information, including data from epidemiological studies, clinical studies and genomic and metabolomic profiling (see FIG. 1, which uses breast cancer as an example). Much of these data are being housed in different databases. For example, more than 1.5 million individual mutations in 25,606 genes in almost 950,000 samples have been described in the Catalogue of Somatic Mutations in Cancer (COSMIC) database[11]. Compiling these types of databases is a primary goal of several consortia, such as the Cancer Genome Project[12], the International Cancer Genome Consortium (ICGC)[13] and The Cancer Genome Atlas (TCGA)[14]. Project Achilles aims to identify genetic vulnerabilities across large numbers of cancer cell lines by systematic loss-of-function studies[9], and the ENCyclopedia Of DNA Elements (ENCODE)[15] investigates structural and regulatory units in the human genome. Genome-wide association studies (GWAS) have identified numerous loci that are linked to cancer susceptibility, but the mechanism by which variations at these loci influence susceptibility remains unknown. Understanding how and why these variants influence subtype-specific cancer risk contributes to our understanding of cancer aetiology. For example, many recent studies emphasize that the genetic architecture of breast cancer is context specific, and integrated analysis of gene expression and chromatin remodelling in normal and tumour tissues will be required to explain the mechanisms of risk alleles[16]. In a network-based strategy, linking GWAS hits with transcription factors that are known to function as master regulators, Fletcher *et al.*[17] found that the risk associated with altered fibroblast growth factor receptor 2 (FGFR2) signalling is due to altered activity of the oestrogen receptor-α (ERα)-associated transcriptional network.

Various molecular markers, which have been identified at DNA, mRNA, microRNA (miRNA) and protein levels, have been used to develop profiles that are associated with taxonomy, tumour aggressiveness, response to therapy and patient outcome[18,19]. In addition, complex biological features at the cellular level, such as histopathological and radiological images, which were traditionally evaluated and scored visually by a trained expert, are now subjected to computational quantification[20,21]. However, small or no overlaps between predictive profiles from different sources persist because of the low statistical power of these studies and the different clinical strata used in each study, among other differences. Pooling data sets, combining profiles at various levels and analysing the data in a compendium — such as the GeneSapiens database[22], the Integrative Multi-Species Prediction (IMP) server[23], Search-Based Exploration of Expression Compendium (SEEK), ProfileChazer[24,25] Oncomine[26], Rembrandt[27] and similar tools — can lead to more reliable molecular signatures and thereby more specific diagnosis and treatment of cancer patients. The joint analysis of multiple data domains, each of which reflect various dimensions of a biological function, has the potential to generate explanatory power that cannot be obtained with one data type alone.

In order to access these data and to carry out some of the integrative analyses detailed below, storage and computing platforms such as the Bionimbus, Bioconductor[28], CytoScape[29,30], IntOGen[31], OncoDrive[32] and Synapse (see Further information) have been designed to enable scientists to exchange data sets, algorithms and mathematical models of cancer. Recently, the idea of sharing and interactive collaboration towards solving a certain biological problem was taken forward by Sage Bionetworks in the competition-based crowdsourcing Dialogue for Reverse Engineering Assessments and Methods (DREAM) breast cancer prognosis challenge (BCC)[33], indicating the need for databases that enable joint analysis and data exchange between researchers.

Bioinformatics tools for integrative analyses

In the context of this Review, integrative statistical analysis refers to the analysis of at least two different types of omics data[34]. The analysis can be restricted to molecular data (such as in expression quantitative trait loci (eQTL) studies, in which the relation between germ line variation and gene expression is investigated[35,36]) or it can involve clinical outcomes (for example, survival, stage and treatment response) or intermediate phenotypes and biomarkers. It is useful to distinguish three broad objectives of integrative analysis, which can be addressed by different statistical tools. The first objective is to understand molecular behaviours, mechanisms and relationships between and within the different types of molecular structures, including associations between these and various phenotypes,

Information theory
A branch of applied mathematics that quantifies the value of information in data.

Bioconductor
A free, open-source and open-development software project for the analysis of high-throughput genomic data. Based on the statistical programming language R, the project was started in 2001 and now contains more than 750 packages to carry out data handling, visualization and analysis.

Expression quantitative trait loci
(eQTL). Genomic loci that regulate expression levels of mRNAs or proteins.

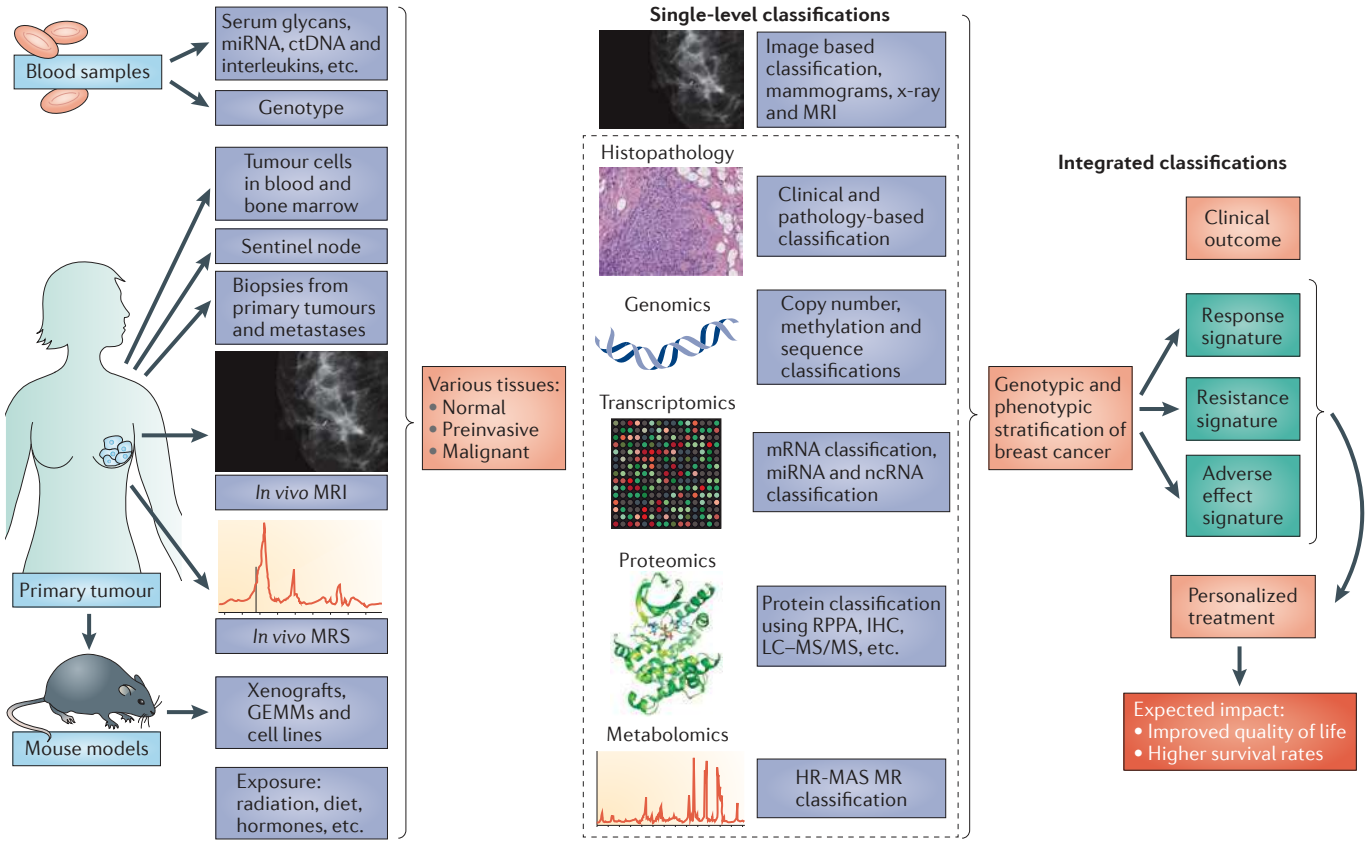

Figure 1 | **The systems biology of breast cancer.** Exploring the systems biology of breast cancer and strategies to investigate multi-dimensional interactions by integration of data from various sources at the indicated levels. ctDNA, circulating tumour DNA; GEMMs, genetically engineered mouse models; HR-MAS MR, high-resolution magic angle spinning magnetic resonance; IHC, immunohistochemisty; LC–MS, liquid chromatography–mass spectrometry; miRNA, microRNA; MRI, magnetic resonance imaging; MRS, magnetic resonance spectroscopy; ncRNA, non-coding RNA; RPPA, reverse phase protein array. Mammography image courtesy of M. M. Holmen of Oslo University Hospital, Oslo, Norway.

such as clinical outcomes, pathways, interactions, 'hot-spot' DNA mutations and mutations in genes that drive cancer development. The second objective is to understand the taxonomy of diseases, thereby classifying individuals (or samples) into latent classes of disease subtype; and the third objective is to predict an outcome or phenotype (such as survival or efficacy of therapy) for prospective patients. Some statistical methods are specialized to one type of question, and others can be used for several. These statistical methods are classified into broad groups (summarized in TABLE 1). Some of the tools, such as enrichment analysis, were originally designed to reveal features of genes and pathways, whereas others, such as integrative clustering, were designed to reveal features of patient subgroups; however, most of the tools discussed below can be applied to both, including integrative graphical models, which can be used to identify aberrant pathways and patient subgroups. The statistical methods discussed in this Review can be classified as unsupervised or supervised (for example, according to whether one proceeds in an exploratory manner or applies clinical labels to individual cases). Some methods use cross-validation or other model selection approaches to estimate the over-fitting in the training set.

Over-fitting
In statistics, over-fitting occurs when a statistical model describes random noise instead of the underlying relationship.

Sequential analysis: combining several distinct omics levels of evidence. This approach allows the confirmation or refinement of findings based on one data type, with additional analyses of further omics data obtained from the same set of samples. In this case, at least two types of omics data are analysed — for example, copy-number alterations (CNAs) and gene expression level data. To integrate two different levels of omics data from the same set of breast cancer samples, Chin *et al.*[37] identified genes whose expression levels were significantly deregulated by CNAs, as well as genes that are associated with metastasis and reduced survival. Lando *et al.*[38] used CNAs integrated with gene expression and gene ontology to identify genes representing five biological processes associated with poor outcome in cervical cancer after chemotherapy and radiotherapy. Moreover, Beroukhim *et al.*[39] combined data from 3,131 cancer specimens, which represented 26 different histological types of cancer, and identified 158 regions with focal CNAs that were significantly altered across all samples. Interestingly, 122 of these CNAs did not harbour a known cancer gene. Each of these papers used the approach in which an analysis of each data set is made independently of the others and produces a list of interesting entities, which

Table 1 | **Tools and algorithms for the detection of activated and altered pathways**

Method	Summary	Refs
Sequential analysis		
MCD	Identification of subsets of genes that are affected on multiple levels by some condition	44,132
CNAmet	Identification of genes that show simultaneous methylation, copy number and expression alterations	45
iPAC	Integration of copy number and gene expression to detect genes and associated pathways or processes that are influenced in *trans* by copy number	42
Consensus clustering	Starting from multiple clusterings (each can represent a data type), obtaining a single integrated cluster assignment	133–136
CHESS	Determining the effect of copy number on gene expression	137
Latent variable		
iCluster	Starting from multiple data types, obtaining a single integrated cluster assignment	48,49,138
PSDF	Integrating copy number and gene expression data to discover prognostic patient subtypes	50
IntegrOmics	Identification of relationships between two 'omics' data sets	72
Penalized likelihood		
Lasso	Identification of omics features with predictive ability for a given response (such as survival), using all data as covariates or using some data to decide the penalty of others	52–54
Elastic Net	Identification of omics features with predictive ability for a given response (such as survival), using all data as covariates or using some data to decide the penalty of others	55
PLRS	Studying relationships between copy number and mRNA expression; detection of copy number-induced sample subgroup-specific effects	139
Camelot	Outputs a linear regression model that uses genotype and expression to predict phenotype; powered by regularized linear regression	140
Lol (Lots of Lasso)	Integration of copy number and gene expression to detect in-*cis* and in-*trans* regulation of gene expression	141
Gene set analysis		
GeneXPress	Extraction of modules and characterization of gene expression profiles in tumours as a combination of activated and deactivated modules	56
GSEA	Gene set annotation of differentially expressed genes	59,142,143
MAPPFinder	Gene ontology term annotation of differentially expressed genes	64, 67,144
SPIA	Pathway annotation of differentially expressed genes	65,145,146
Pathologist	A consistency score and an activity score is calculated for each pathway	66
KOBAS	Pathway and disease annotation of gene sets	147,148
SubpathwayMiner	Pathway annotation of gene sets	149,150
MGSA	Identification of active gene sets	82
Pair-wise correlation		
WGCNA	Finding modules of highly correlated genes using eigengene network methodology	73
Oncodrive-CIS	Ranking genes according to the effect of copy number on gene expression	151
Network-based analysis		
jActiveModules	Identification of expression-activated sub-networks	78,152,153
GiGA	Identification of the gene subgraphs showing the most significant gene expression pattern	79

Table 1 (cont.) | **Tools and algorithms for the detection of activated and altered pathways**

Method	Summary	Refs
Network based analysis (cont.)		
PARADIGM	Prediction of the degree to which the activities of a pathway are altered in an individual	86,87,125
PathExpress	Determining if there is enrichment of genes around each enzyme, on the basis of gene–metabolic relations in KEGG	154
AMBIENT	Discovery of metabolic sub-networks that are significantly changed by some condition	155
Bayesian		
CONNEXIC	Integration of copy-number variation and gene expression to identify driving cancer mutations and the processes that they influence	91
COALESCE	Using gene expression and DNA sequence data as inputs, this method produces putative co-regulated modules as outputs	68
MDI	Identify groups of genes that tend to be allocated to the same components in multiple data sets or molecular levels	156
Other		
RegMOD	Identification of active modules or dysfunctional pathways	61

AMBIENT, Active Modules for BIpartitE NeTworks; CHESS, CgHExpreSS; CNAmet, Copy Number Alteration and methylation; COALESCE, Combinatorial ALgorithm for Expression and Sequence-based Cluster Extraction; CONNEXIC, COpy Number and EXpression In Cancer; GiGA, Graph-based iterative Group Analysis; GSEA, gene set enrichment analysis; iPAC, in-*trans* Process Associated and *Cis*-correlated; KEGG, Kyoto Encyclopedia of Genes and Genomes; KOBAS, KEGG Orthology Based Annotation System; MAPPFinder, MicroArray Pathway Profile Finder; MCD, Multiple Concerted Disruption; MDI Multiple Dataset Integration; MGSA, Model-based Gene Set Analysis; PARADIGM, PAthway Recognition Algorithm using Data Integration on Genomic Models; PLRS, Piecewise Linear Regression Splines; PSDF, Patient-Specific Data Fusion; RegMOD, Regression MODel with diffusion kernel; SPIA, Signalling Pathway Impact Analysis; WGCNA, weighted gene correlation network analysis.

are then linked to each other. For example, differentially expressed genes in one list are compared with each other and then with different CNAs that have been matched to the closest gene in a second list. Usually, the lists are intersected to find the genes that are confirmed in the analysis of each data type[40]. Comparing ranks of each gene in each list leads to measures of concurrence. If each entity in each list has a value (for example, a *t*-test statistic) then these values are combined. Although each list often contains significantly selected entities after multiple testing corrections, it is not obvious how to assign a *P* value to the intersection. Permutation testing of each individual analysis before intersection could, in principle, be used. One such flow chart-based data integration framework is Anduril, in which the ultimate goal is to elucidate the impact of various omics data on patient survival[41]. Occasionally, the various analyses are not performed in parallel but as a sequence of filtering steps, each functioning on a single data type. In this approach, the order of the filtering steps matters. In the in-*trans* Process Associated and *Cis*-correlated (iPAC) algorithm[42], which was designed to detect cancer drivers, whole-genome gene expression measurements are correlated with segmented copy-number data to obtain a list of genes with strong in-*cis* correlation. Using each of these in turn as a pivot, all other genes in the genome are ranked according to their correlation to the pivot, and enrichment of gene ontology terms for genes at the top of the ranked list is investigated. As a result, iPAC identifies CNAs with a phenotypic effect in the sense that they have an impact on expression in *cis*, as well as on processes in *trans*. In another study that used data from sarcomas, Chibon *et al.*[43] derived a gene signature, known as Complexity INdex in SARComas (CINSARC), by combining known genes of importance with genes whose expression correlated to CNAs and were members of over-represented pathways, including those that effect chromosomal instability or histological grade. Interestingly, CINSARC predicted the likelihood of metastasis development (a surrogate for survival) in patients with sarcomas and also in patients with gastrointestinal stromal tumours (GISTs), breast cancer and lymphoma, which points to the ability of integrative approaches to identify universal features of aggressive cancer.

Several methods regard the expression level of any transcript as a function of copy number and DNA methylation. An example is a tool called Multiple Concerted Disruption (MCD), which aims to integrate DNA copy number and methylation to explain variation in mRNA expression data in *cis*[44]. The MCD method searches for deviation from the normal at several levels: a differential expression, a change in gene copy number or a change in the degree of DNA methylation[44] (hypomethylation or hypermethylation). The procedure involves several sequential steps and can be carried out either per sample or across a set of samples. By sequentially examining more genomic dimensions at the DNA level (that is, copy number, allelic status and DNA methylation) one can explain a higher proportion of the observed changes in gene expression. Notably, this varies to a great degree from sample to sample, which indicates intrinsically distinct mechanisms leading to deregulation[44]. The MCD method was followed by a similar method[45] (Copy Number Alteration and methylation (CNAmet)), which was implemented in open-source R[28,46].

T-test statistic
T-tests are used to determine whether the mean of a continuous variable is different in two groups of individuals. It is based on a quantity called a *t*-test statistic, which is computed from the data and reflects the signal-to-noise ratio.

Expectation-maximization algorithm
(EM algorithm). An iterative algorithm for the estimation of parameters in statistical models depending on unobserved variables. A limitation with EM is that it requires specification of initial values for the iteration, and the estimated parameters may depend on these.

Lasso
A shrinkage and variable selection method for linear regression, used in particular when there are many covariates (for example, genes).

Latent variable analysis: using common factor labels derived from multiple omics levels. Unsupervised clustering of omics data can be used to partition individuals or samples into subgroups of potential clinical relevance[47]. In the iCluster package[48], for example, the clustering of individual samples is carried out by applying metrics (or noise structures) that are specific to each data type but using common latent labels among all data types, employing an expectation-maximization algorithm (EM algorithm). This method can be extended to supervised clustering when the data are continuous, such as for expression data or CNAs[49], and it can accommodate any number of data types. The number of clusters is difficult to determine and is estimated by cross-validation methods. Using iCluster, Curtis *et al.*[49] found that genome variation influenced gene expression and identified putative cancer genes, and it defined novel subgroups of patients with breast cancer who had distinct outcomes[48]. Furthermore, *trans*-acting aberrant DNA hot-spots that modulated subtype-specific gene networks were shown. A further development has been suggested by Yuan *et al.*[50]. Their Patient-Specific Data Fusion (PSDF) algorithm exploits the fact that the data to be integrated in individual samples might seem to be contradictory within the data pool; for example, a high copy number of a gene could be associated with a high expression of the same gene in *cis* in most, but not all, samples. Such a contradiction can be seen as a measurement error or biological variation due to the cell composition of a biopsy or patient characteristics. PSDF estimates a latent variable per patient, which helps to exclude (or minimize) contradictory samples. This idea could potentially be used beyond clustering, for other tasks of integrative analysis.

Penalized likelihood analysis: using regularization to handle high-dimensional multi-omics data. The aim of integrative regression is to determine the genes (or entities) — using at least two different omics data types — that allow the best prediction of the outcome. Since the number of covariates mostly supersedes the number of samples, some form of variable selection or penalized regression is necessary[51]. When sparsity can be assumed (that is, when only a few entities are expected to actually be relevant for the outcome), Lasso[52,53] is a very useful penalization method, as it carries out variable selection. Cross-validation is used to determine an optimal level of penalization, which influences the sparsity of the solution. A straightforward way to use Lasso with two different data types is to use all data as covariates[54] (after appropriate standardization): in this case, the algorithm chooses the optimal set of predictors from either omics source. Adaptive Lasso works in two steps and, like Elastic Net[55], is more parsimonious than Lasso. A different analysis is known as Weighted Lasso, in which the Lasso uses only one covariate type (such as the mRNA expression level) while the other covariate modifies the penalization so that genes are individually penalized. For example, the penalization of a gene expression can depend on the correlation between the CNA and the outcome, so genes with an important CNA will be penalized less in the expression analysis[38]. Regression-based integration is mostly in *cis*, but it can easily be extended to more data types.

Gene set analysis: discovering novel or using known groups of related molecules. One of the earliest reported examples of an integrative approach for gene expression data was the use of GeneXPress to identify modules of genes that affect the activity of a tumour[56]. Segal *et al.*[56] analysed data from 22 cancer types and found that distinct shared modules of gene activity, which probably represented common tumour progression mechanisms, characterized distinct tumour types. A different strategy involves initially defining a collection of gene sets (for example, gene ontology terms or pathways). This step typically involves the use of publically available databases that collect extensive annotation and knowledge (for example, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome and WikiPathways[57]; see Further information). A score is calculated for each gene[58] (for example, a *P* value that reflects the degree of differential expression), and all gene sets that are 'enriched' or over-represented with high or low scores are identified. The scores can also be binary (0 or 1), thereby indicating, for example, membership in a group of differentially expressed genes. By combining gene ontology, gene expression and clinical data, Subramanian *et al.*[59] used gene set enrichment analysis (GSEA) to identify genes consistently associated with poor outcome in two independent cohorts of patients with lung cancer. Information based on known protein–protein interactions has been used to identify gene modules expressed in non-malignant bystander cells[60], associated with metastatic disease[61] or associated with aggressive disease in lymphoma[60–63,157]. Several alternative ways of scoring the abnormal presence of specific pathways have also emerged, including Gene Microarray Pathway Profiler (also known as MicroArray Pathway Profile Finder (MAPPFinder))[64]. These methods describe the functional profile of a list of genes/proteins by comparing with known (*a priori*) interactions, scoring the over-representation of a given pathway, ignoring any knowledge about the network structure. Other tools, such as Signalling Pathway Impact Analysis (SPIA)[65] and Pathologist[66], exploit pathway topology by taking into account the position of a gene in a pathway. SPIA uses the number of neighbours for every gene (the 'degree'), so that a gene with a higher degree is more likely to have a master role than a more isolated gene and is then favoured in the analysis of the original data[67]. Pathologist assumes that every gene is either active or inactive in a network, and this method models this as a mixture of two gamma distributions, using the EM algorithm to compute both a gene activity score and an overall pathway score[68,69].

Pairwise correlation analysis: inferring molecular network interactions from strengths of associations. In this type of analysis, for each pair of co-measured omics data, a correlation matrix is estimated[70], with *P* values that are corrected for multiple testing and that therefore reflect the strength of association. This approach includes associations in *trans*. The structure in the matrix can be used

to identify master regulators[71]. Correlation analysis does not directly facilitate the study of how entities (such as expression levels and CNAs) regulate outcomes of interest, but highly correlated entries can be used in further studies, such as in canonical correlation analysis[72]. There are multiple ways to extend the correlation analysis to more than two data types. For example, weighted gene co-expression network analysis describes the correlation patterns among genes across microarray samples. Weighted gene correlation network analysis (WGCNA) is a method for finding clusters (or modules) of highly correlated genes using matrix calculus[73]. Correlation networks facilitate network-based gene screening methods that can be used to identify candidate biomarkers or therapeutic targets. In order to identify higher order interactions, for example those in which highly cooperative processes involve many subunits of a protein, dependence among multiple variables can be established using maximum entropy techniques[74] or information theory approaches. These methods are distinct from correlation methods and in some ways might be more powerful[75].

Network analysis: using molecular network interactions to identify active or aberrant subgraphs. Networks are a representation of how genes or other entities collaborate in certain biological systems[76,77]. A graph 'sums up' these effects over time, and two genes will be linked by an edge if they seem to interact in a specific process. Graphical algorithms that capture the interaction between differentially expressed genes by correlation include jActiveModules[78] and Graph-based iterative Group Analysis[79] (GiGA). jActiveModules integrates knowledge from protein–protein and protein–DNA interaction databases into mRNA expression data by assigning a Z-score for differentially expressed genes, and it searches for connected sub-networks by simulated annealing and greedy search algorithms[60,80]. Both simulated annealing and greedy search identify differentially expressed sub-networks; in the first case, in an optimal but computationally very intensive way; in the second case, more rapidly but less accurately. GiGA also ranks genes on the basis of differential expression levels and searches for sub-networks. In 2011, Stingo *et al.*[81] proposed a Bayesian approach that selects both the actual pathways (out of a large set of possible ones) and the key genes that allow the best prediction of an outcome. Additional Bayesian methods to infer the functional content of a list are presented in REFS 82,83.

Statistical graphical models with feedbacks have been successful in summarizing data and representing pathways. The study of such networks can lead to an important understanding of biological mechanisms[84,85]. A few, such as PAthway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM), have been extended to integrative analysis[86]. Activity levels of each gene are considered as latent variables, which are estimated and then used in subsequent analyses. Given a set of genes of interest, the first step is to assemble from public databases (see Further information for examples) a large enough network of genes, with their activating and inhibitory interactions, and to transform this into a directed graph, for which the key biological assumptions are followed: that is, for each gene, CNA affects expression, which affects protein levels, which affect the latent protein activity. These activity nodes are then connected to other gene-specific nodes in ways that are predicted from existing knowledge. This graph represents the 'normal' or reference state. When data are attached to some of the nodes (for example, all expression levels and CNAs for a sample of individuals with a specific disease), a joint posterior distribution is then computed for all latent activity nodes, and this approach is called integrated pathway activities (IPAs). By comparing pre- and post-activity levels, it is possible to obtain a quantitative description of the alteration that is induced by the disease, with respect to normality (or between different groups). In order to make the computations feasible, all measurements are categorized in three discrete states (inhibited, normal and activated). Despite this, the computational burden is so large that the EM algorithm is locally applied in each node, and this leads to an approximation that reduces computational time. PARADIGM was tested using copy number and mRNA expression data[86], as well as with the addition of methylation and miRNA expression data[87]. The methodology is, in principle, open to incorporate further levels of complexity, such as different progression levels (from normal tissue to pre-invasive and invasive cancer), that add an additional level to the analysis. An example from breast cancer that combined both different progression levels, as well as multiple levels of molecular data, clinical data and pathway information, used the properties of PARADIGM to define groups of patients with distinct biological signatures and different prognoses[87].

Bayesian analysis: imposing realistic assumptions to coherently integrate multiple omics data. Bayesian methods naturally facilitate the integration of biological knowledge through the design of appropriate prior distributions. In a Bayesian multiple testing setup, one can use a second type of omic data (for example, CNAs) to modulate the *a priori* probability that each test for a first data set (for example, expression levels) is likely to be rejected[88]. Bayesian networks are not new, and they were used in the early 2000s to incorporate various data[89]. As is true for every statistical method, Bayesian analysis is based on assumptions (both probabilities and prior assumptions) and models based on these have to be realistic and well designed so that they can be trusted. Usually, in a Bayesian setting, one carries out sensitivity analysis to assess informative prior assumptions, but, in most cases, prior assumptions are non-informative. Nevertheless, Bayesian approaches have a natural and important role in data integration, and they differ by the fact that prior distribution can represent knowledge, and conditional independence facilitates the integration of data in a coherent way. Bayesian variable selection has been successfully applied to situations that comprise one data type[90], but it could be extended to multiple data types using similar fundamental biological assumptions as in Huttenhower *et al.*[69]. Other computational frameworks use integrative Bayesian approaches to identify

Maximum entropy techniques
An alternative to maximum likelihood, maximum entropy techniques are a way to estimate models from data, by finding the most random probability distribution that fits the data.

Simulated annealing
A global optimization algorithm that seeks a good approximation to the point of absolute maximum of a function.

Greedy search algorithms
In optimization, a greedy algorithm is an iterative algorithm that takes an optimal (or semi-optimal) choice at every step, in the hope of obtaining the global solution at convergence. These algorithms do not generally result in optimal solutions and are used when the determination of a global solution would require an unacceptable amount of computing time.

Bayesian approach
An approach to statistics that involves starting from our current (*a priori*) level of knowledge, collecting data and then using both to infer our (*a posteriori*) knowledge. Bayesian inference allows the incorporation of additional external knowledge into the estimation process.

Latent variables
In statistics, latent variables (as opposed to observable data) are not measured but must be estimated from data, similar to parameters. However, contrary to parameters, latent variables are random and have a distribution. Latent models are inherently Bayesian.

Support vector machines
In machine learning, support vector machines are supervised learning models that are used for classification and regression analysis.

candidate drivers from copy number and expression data (for example, COpy Number and EXpression In Cancer (CONNEXIC)[91]), in order to cluster samples, while simultaneously estimating the number of clusters[50], or to perform regulatory module predictions from co-expressed biclusters (Combinatorial ALgorithm for Expression and Sequence-based Cluster Extraction (COALESCE)[68]).

There are additional methods that do not naturally fit into the defined classifications above. Models that are based on ordinary differential equations can integrate various types of data but, as they model complex chemical reactions at a molecular level, they require a large number of input parameters that are usually unknown[92]. In some cases, support vector machines have been used to evaluate the active score of each gene and to identify nonlinear dependencies in active networks in a computationally efficient way (for example, Regression MODel with diffusion kernel (RegMOD)[61]).

The novel statistical and computational approaches described above are bringing us to a level at which we can analyse molecular data at all studied molecular levels (FIG. 1), in an integrated manner. For instance, by using the matrix of IPAs generated by PARADIGM, from the summarization of copy number, expression level and known interactions among the genes, one could identify a better prognostic signature than the one derived from expression clusters alone[87] (FIGS 2–5). By layering on CNAs and mutation data it has become possible to deduce how an individual tumour evolved[93,94]. Furthermore, Chari *et al.*[44] showed that by examining samples using more genomic dimensions, including copy number, allelic status and DNA methylation, they were able to explain a higher proportion of the variation in gene expression compared with studying each genomic level separately, using only one genomic level. Remarkably, the proportion of variation in gene expression widely varied from patient to patient, which indicates different regulatory mechanisms and complex individual gene–gene interactions in *trans* that are specific for every tumour. This inter-individual variation might be a limiting factor in the identification of molecular markers that are associated with tumour aggressiveness, response to therapy and patient outcome.

Integrative analyses across tumour types

Over the past decade, the accumulation of high-throughput molecular data from various cancer types has revealed an enormous range of alterations. Although subgroups of tumours with similarities in biological properties or clinical behaviour can be defined, the initial studies mainly analysed one type of molecular data at a time. The access to large data sets that have been made available by the ICGC and TCGA has made it possible to compare the performance of some of the tools described above, on the same data set, as well as to compare the identified deregulated pathways between different cancer types. A pilot project from TCGA integrated DNA copy number, gene expression and DNA methylation, as well as nucleotide sequence aberrations from glioblastoma samples[95]. Enrichment analysis revealed new roles for known cancer genes, as well as network activity. Later, the same data set was interrogated by Anduril[41] and by PARADIGM[86]. Both approaches suggested that amplification of the epidermal growth factor receptor (*EGFR*) was important in glioblastoma. Anduril, which can make use of DNA methylation data, also indicated DNA hypomethylation as a significant change that was evident in glioblastoma.

Data from the first pan-cancer analyses aim to identify drivers of tumorigenesis that are common to multiple tumour types[96,97]. For example, the aim of TCGA is to generate genomic data at all molecular layers in 10,000 tumours from 20 tumour types and to make these data available for the community[98]. A recent endeavour to integrate somatic mutations, CNAs and DNA methylation was carried out in 3,299 tumours of 12 different cancer types[96]. After integration with mRNA expression, a total of 479 candidate functional alterations were predicted, including 116 copy-number gains, 151 copy-number losses, 199 recurrently mutated genes and 13 epigenetically silenced genes. A hierarchical stratification was built using principles from network modularity[99]. Interestingly, on the basis of these analyses, tumours seemed to be driven either by somatic mutations or by CNAs — a phenomenon that the authors named 'the cancer genome hyperbola', owing to the inverse relationship between these events. However, some genes, such as *TP53* and phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit-α (*PIK3CA*), can be subjected to both aberration modes, thereby leading to the deregulation of common pathways such as p53-mediated apoptosis, PI3K–AKT signalling and cell cycle control.

Studying the relationship between the different genomic levels (FIGS 2–5) opens a debate over their explanatory weight and potential to discover drivers of cancer[100]. Ovaska *et al.*[41] found unexpectedly poor concordance between gene amplification, overexpression of the genes from the amplicons and survival in patients with glioblastoma. Akavia *et al.*[91] showed that the expression of a driver (not its copy number per se) drives a phenotype. The authors draw our attention to the fact that many of the current studies attempt to identify drivers only in genomic loci for which there is a good correlation between copy number and mRNA expression. So far, many current approaches have been based on linear correlation analysis. On the basis of knowledge of the enzyme kinetics and gene regulation, we expect nonlinear dependencies to occur in addition to linear effects. We recently proposed a statistical approach to investigate linear and nonlinear dependencies between CNA and mRNA expression[101].

Clinical application. The discussion above has addressed the problem of inferring biological networks of relevance for translation into the clinic, based on a simple map of genes, transcripts and proteins. A paradigm shift is needed, from searching for single strong clinical markers to searching for a combined effect of multiple markers, as, in general, genes

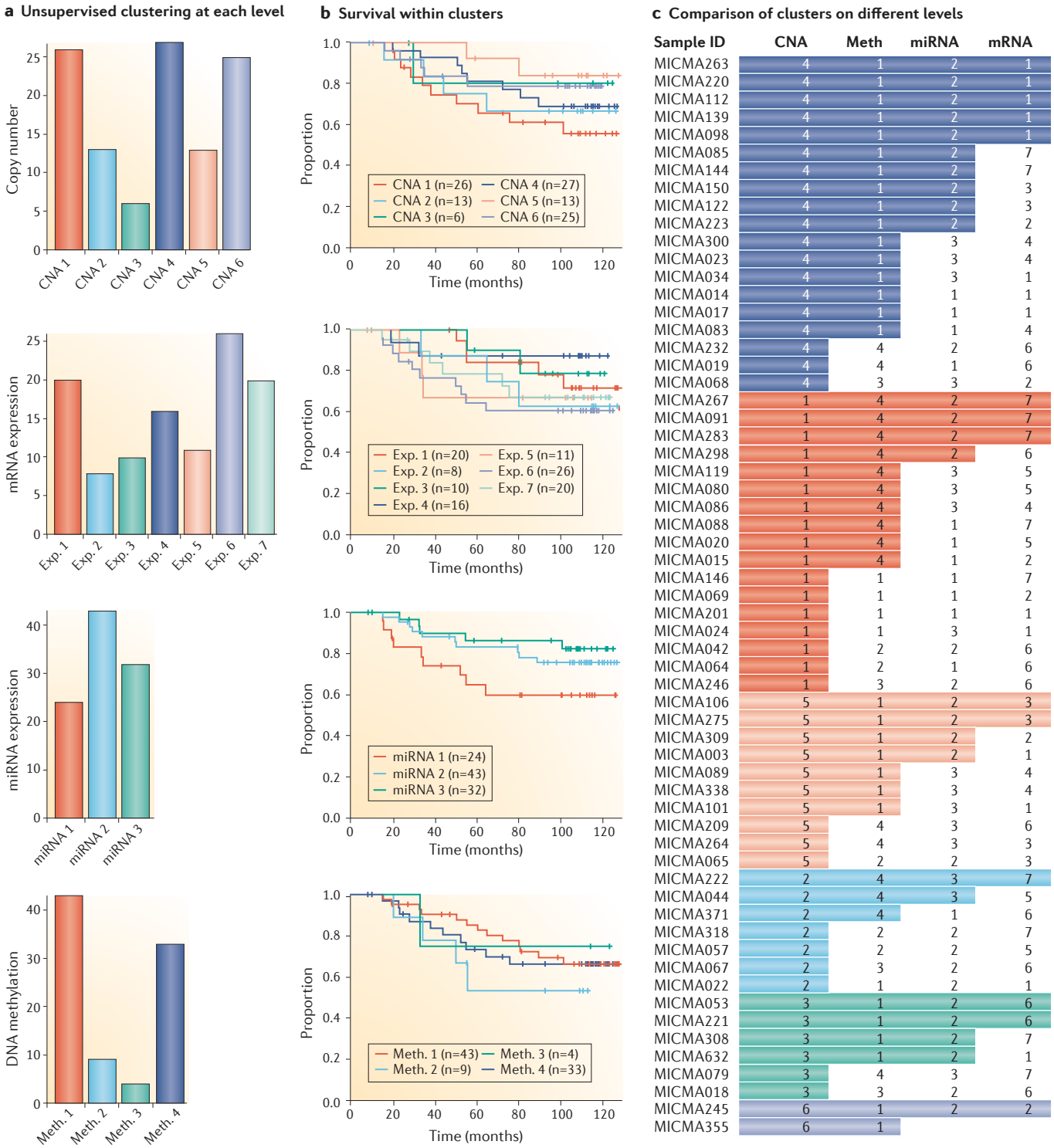

Figure 2 | **Classifying breast cancer using unsupervised clustering.** The first solid tumour to be profiled by expression arrays was carcinoma of the breast[119]. The most reproducible classification by mRNA expression is based on the biological entities referred to as the intrinsic subtypes — luminal A, luminal B, basal-like, human epidermal growth factor receptor 2 (HER2)-enriched and the normal-like groups[120,121]. In the past decade, several molecular studies to classify breast cancer have added one or two molecular levels — most frequently, DNA copy number[42,49,122,123] and gene sequencing[124]. However, few of the studies have integrated more than two levels of information from the same patients[87,125]. In our laboratory, we have collected several layers of high-throughput molecular data from patients with breast cancer, including DNA methylation, DNA copy number alterations, mRNA expression and microRNA (miRNA) expression[93,126–131]. Clustering according to each molecular level reveals a variable number of clusters (part **a**). Kaplan–Meier plots are shown for each patient cluster within each molecular level (part **b**). Comparison of clusters on different molecular levels reveals that some breast cancer samples cluster together at all the molecular levels, while others cluster in different groups according to the particular molecular endpoint (part **c**). Figure parts **a**,**b** are reproduced, with permission, from REF. 87. Exp, expression; Meth, methylation.

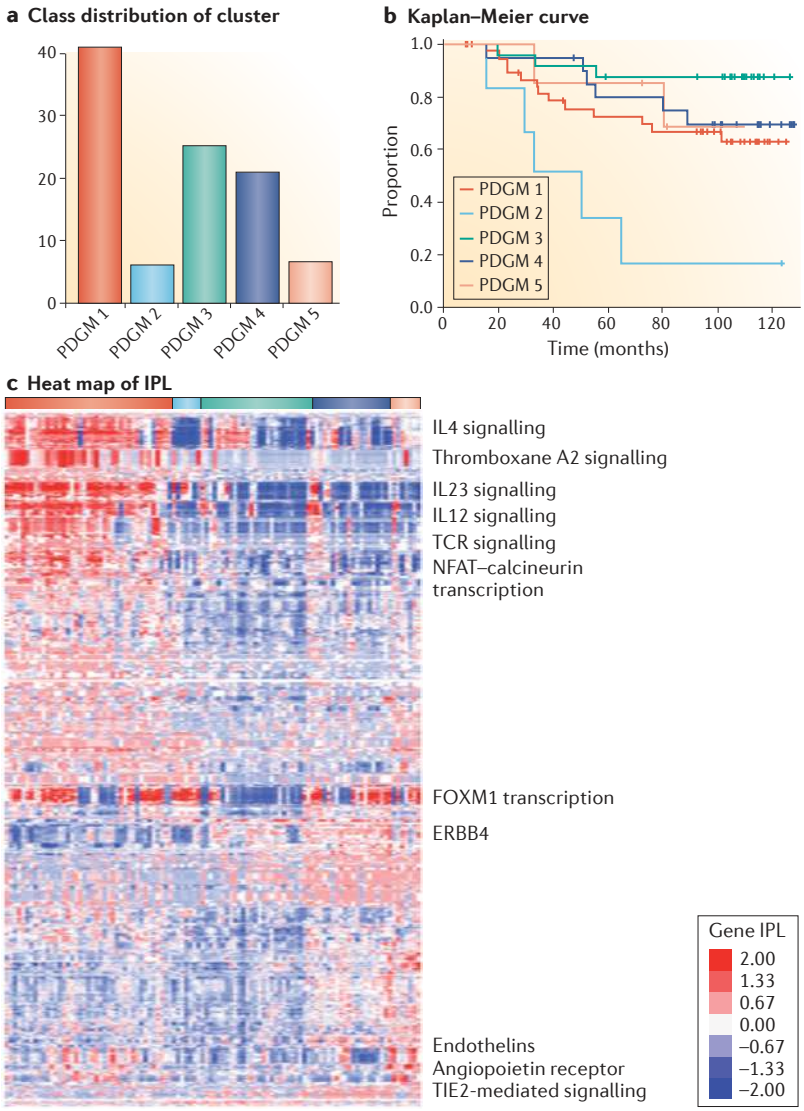

Figure 3 | **Classifying breast cancer using PARADIGM.** All multiple layers of high-throughput molecular data described in FIG. 2, including DNA methylation, DNA copy number alterations, mRNA expression, microRNA (miRNA) expression as well as *TP53*-mutation status, were subjected to integrated analysis using the PAthway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM). This resulted in five clusters (part **a**) with survival differences (part **b**) and this was validated in multiple other datasets[87]. A heat map of integrated pathway levels (IPLs) is shown in part **c**. FOXM1, forkhead box M1; IL, interleukin; PDGM, PARADIGM cluster; TCR, T cell receptor; TIE2, tyrosine kinase, endothelial. Figure is reproduced, with permission, from REF. 87.

and proteins function by interacting with DNA, RNA and proteins, and these interactions might be specific for a given disease subclass[102]. Many of the current targeted therapies focus on proteins that are involved in cell signalling pathways, which form a complex cellular communication system that governs basic cellular functions[103,104]. Established examples of targeted cancer treatment include EGFR-mutated non-small-cell lung cancer that can be treated with tyrosine kinase inhibitors (gefitinib or erlotinib)[105,106], ERBB2 (also known as HER2)-directed therapy in breast cancer[107,108], and melanomas with $BRAF^{V600E}$ mutations that can be targeted with vemurafenib[109].

A major challenge in drug development is to precisely define the subset of cancer patients that are likely to respond. Within each pathway, a range of drugs may be available, and the optimal target (and, hence, the optimal drug) will be determined by the rate-limiting protein and the individual perturbations in the pathway. In colorectal cancer, EGFR-directed therapy with monoclonal antibodies has proven to be effective[110]. However, in the presence of a downstream activating KRAS mutation, the inhibition of EGFR is ineffective[111]. It seems likely that similar mechanisms are present in cases with resistance to other cancer treatments (both targeted and more traditional chemotherapeutic agents). Iadevaia *et al.*[112] have proposed a computational procedure to generate experimentally testable intervention strategies for the optimal use of available drugs in a cocktail. They used reverse phase protein array to evaluate the changes in the phosphorylation status of proteins after stimulation of the MDA-MB 231 breast cancer cell line with insulin-like growth factor, and they were able to conclude that the simultaneous inhibition of MAPK and PI3K–AKT pathways was sufficient to significantly halt cell proliferation[112]. Future methods will require adding methylation and expression data to such integrative approaches. Introducing systematic clinical screenings for mutations that perturb these pathways is of great importance to identify the targets for targeted therapies and the patients that will respond to each treatment.

Outcome prediction that is based on genomic data is another central area of genomic research, and it has proven to be promising in breast cancer. One of the crucial issues in retrospective studies is that treatment selection is mostly based on the predicted risk of recurrence. Thus, treatment might be confounded by prognosis. This challenges the identification of pure prognostic markers, as the treatment interaction is not known. Even though the results from prospective validation trials, such as the Microarray In Node-negative and 1–3 positive lymph node Disease may Avoid Chemotherapy (MINDACT) trial and the Trial assigning individualized options for treatment (TailorX), are still pending, prediction tools based on gene expression are included in some clinical guidelines[113,114]. Optimal strategies for risk prediction are, however, not settled and remain controversial. Crowdsourcing strategies for problem solving, which were previously successfully applied to biology in areas such as the prediction of protein folding and function[115,116], have been applied to this problem. In the DREAM BCC competition[33], participants competed to create an algorithm that could predict — more accurately than current benchmarks — the prognosis of patients with breast cancer from clinical information (age, tumour size and histological grade), genome-scale tumour mRNA expression data and DNA copy-number data from 1,980 patients[33]. Integration of data was encouraged, and more than 1,400 models were submitted. The winners used a mathematical approach that was based on co-expression gene networks associated with tumour phenotype and

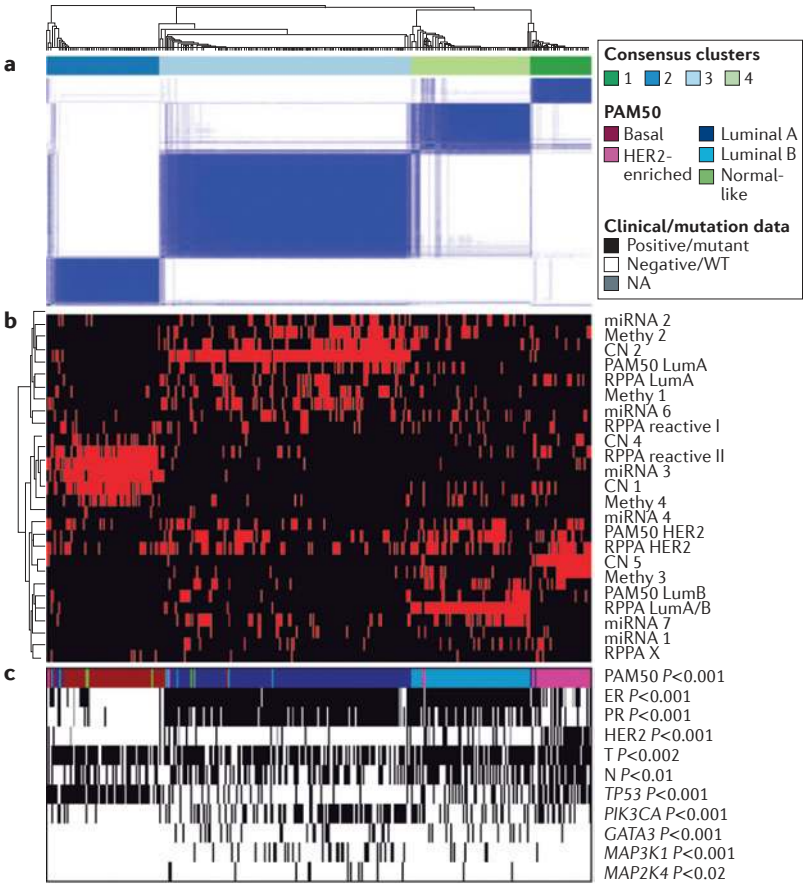

Figure 4 | **Classifying breast cancer using clustering of clusters.** Consensus clustering (or 'cluster of clusters') of 348 breast cancer cases, based on data from five different genomic and proteomic platforms. Consensus clustering analyses of the subtypes identifies four major groups; the blue and white heat map displays sample consensus (part **a**). A heatmap display of the subtypes defined independently by microRNAs (miRNAs), DNA methylation, copy number, PAM50 mRNA expression, and reverse phase protein array (RPPA) expression; the red bar indicates membership of a cluster type (part **b**). Associations with molecular and clinical features, with *P* values from a chi-squared test are shown in part **c**. CN, copy number; ER, oestrogen receptor; *GATA3*, GATA binding protein 3; HER2, human epidermal growth factor receptor 2; LumA, luminal A; LumB, luminal B; *MAP2K4*, mitogen-activated protein kinase kinase 4; *MAP3K1*, mitogen-activated protein kinase kinase kinase 1, E3 ubiquitin protein ligase; Methy, methylation; N, node status; NA, not available; PAM50, gene expression subtyping based on the PAM50 gene signature; *PIK3CA*, phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha; PR, progesterone receptor; T, tumour size; WT, wild type. Figure is reproduced, with permission, from REF. 125 © (2012) Macmillan Publishers Ltd. All rights reserved.

functional characteristics to identify signature 'attractor' meta-genes, and this approach outperformed other models to predict outcome[117,118]. These examples support the notion that using the expertise of participants outside of traditional biological disciplines could be a powerful way to accelerate the translation of biomedical science into the clinic.

Limitations of integrative analyses

Integrative analyses are likely to become ever more important as computational strategies and tools are further improved and multilevel omics data sets become more abundant. The quest to understand the interplay within and between different molecular levels in cancer is no longer beyond our reach. It is important, however, to be aware of the limitations of the current methodologies. From a statistical perspective, the most fundamental challenge in integrative analyses is dimensionality: taking more levels into account in the analysis tends to increase the dimensionality of the problem. Adding more layers of data or increasing the resolution of measurements increases the dimension of unknown parameters, which are often difficult to estimate, thereby making the overall inference weaker. This might seem paradoxical, as the purpose of taking multiple levels into account is precisely the opposite — to use more observations to obtain a more accurate picture of the biological system under study. The way out of this apparent paradox is to realize that, first, one is able to infer more properties of a system with integrative approaches and, second, statistically efficient integrative methodologies can be constructed by actively using known properties of the relationships between the molecular levels. The second point ensures that additional variables in the analysis are not, in effect, increasing the degrees of freedom of the underlying model but rather lending information to existing variables. In addition, at every step, there will be checkpoints of compatibility of the data, such as normalization to the same scale, sample selection from representative cohorts, adequate correction for technical batch effects and use of different platforms. Although numerous methods and tools are introduced to address these obstacles, it is still, so far, the case that large-scale true integration is possible within only a few projects worldwide, which have sufficient funding that allows all analyses to be carried out simultaneously and on the entire data set. Intuitively, it seems that as a 'gold standard', integration attempts are best carried out in supervised settings that are based on some priming biological knowledge or within the frame of defined biological hypotheses. Combining additional layers in unsupervised analyses might fail to contribute new information, as multiple use of the same data might artificially reduce variance or will increase the false discovery rate.

Conclusions

A more fundamental understanding of the biological dynamics of cancer will enable us to better identify risk factors, refine cancer diagnosis, predict therapeutic effects and prognosis, and identify new targets for therapy. We are seeing a paradigm shift from large randomized clinical trials towards treatment modalities that are tailored for stratified patient groups, down to N-of-1 trials, in which data from a single patient represents an entire trial. This will fundamentally alter the way that we statistically model and evaluate treatment strategies, from identifying patient groups that have a response to treatment that is above random to identifying pathways and biological entities that are druggable and altered above random; and from evaluating the response in randomized arms, using the other arm as a control, to evaluating the response of experimental and control interventions in each individual, using the same individual as a control. The real challenge

would be to develop statistical models to identify crucial, rate-limiting molecular targets for intervention, out of the wealth of information that next-generation sequencing uncovers, on the background of great redundancy of pathways and heterogeneity of tumours. As we are moving towards an era in which the amount of data produced every year is increasing exponentially, the biomedical community needs to embrace this complexity and find new methods of shared analysis. We need to learn from physicists

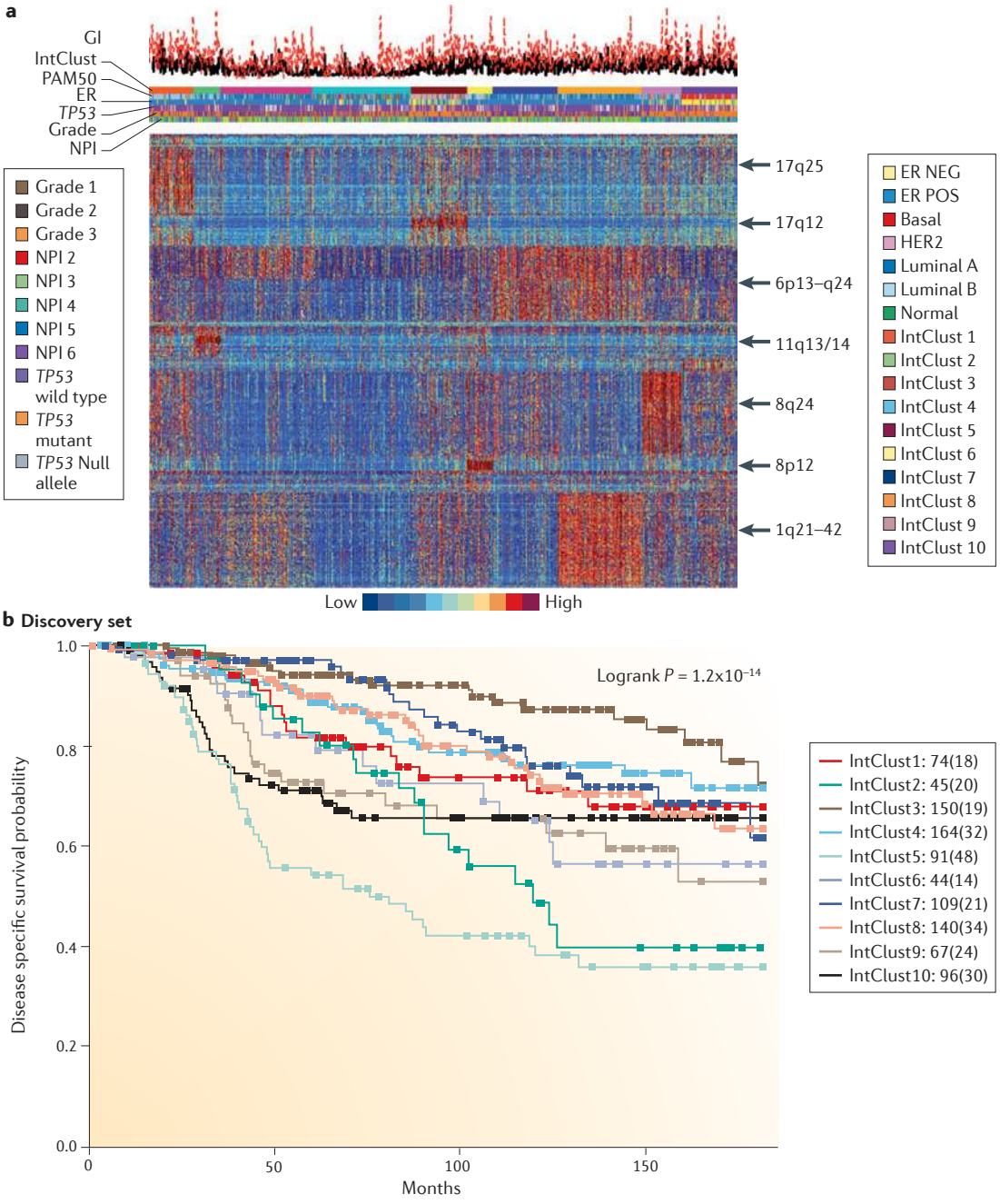

Figure 5 | **Classifying breast cancer using integrative clustering.** Integrative clustering of 997 breast cancer cases from the METABRIC cohort, based on segmented copy number and gene expression for the top 1,000 *cis*-acting copy number-expression associations. Heatmap showing the product of scaled gene expression and copy number values for the selected features and for k = 10 clusters; columns represent breast cancer cases and rows represent features (part **a**). Kaplan–Meier plot of disease-specific survival (truncated at 15 years) for the integrative subgroups. For each cluster, the number of samples at risk is indicated as well as the total number of deaths in parentheses (part **b**). ER, oestrogen receptor; ER NEG, ER negative; ER POS, ER positive; GI, genomic instability based on the proportion of genome altered (black line) and jump measure (red line); grade, genomic grade; IntClust, groups found using integrative clustering with k = 10 clusters; NPI, Nottingham prognostic index; PAM50, gene expression subtyping based on the PAM50 gene signature. Figure is reproduced, with permission, from REF. 49

and mathematicians and transform our way of working, thereby making data available on a hub so that everyone who is interested in it can work on it. New ideas can then be instantly picked up by anyone, rather than waiting for their publication. This was essential in the success of the DREAM BCC and is an example of the one of the many computational challenges that have been set by DREAM, with the goal of catalysing the interaction between theory and experiment, specifically in the area of cellular network inference and quantitative model building in systems biology.

An enormous challenge is also the functional validation of the *in silico* findings in relevant living biological systems, as well as the development of adequate *in vitro* functional studies (such as small interfering RNA screens, knock-in systems and knockout systems) to keep up with the increasing throughput by which candidates for validation are generated. We still need to explore functions of thousands of candidate cancer genes and proteins to ascertain their value as risk factors, as predictive factors for therapy response and as therapeutic targets.

1. Hood, L., Heath, J. R., Phelps, M. E. & Lin, B. Systems biology and new technologies enable predictive and preventative medicine. *Science* **306**, 640–643 (2004).
2. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
3. Auffray, C. & Hood, L. Editorial: Systems biology and personalized medicine - the future is now. *Biotechnol. J.* **7**, 938–939 (2012).
This paper outlines the definitions and state of the art methodology in systems biology.
4. Tian, Q., Price, N. D. & Hood, L. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. *J. Intern. Med.* **271**, 111–121 (2012).
5. Schadt, E. Eric Schadt. Interview by H. Craig Mak. *Nature Biotech.* **30**, 769–770 (2012).
6. Joyce, A. R. & Palsson, B. Ø. The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell. Biol.* **7**, 198–210 (2006).
7. Martin, M. Semantic Web may be cancer information's next step forward. *J. Natl. Cancer Inst.* **103**, 1215–1218 (2011).
8. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
9. Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl Acad. Sci. USA* **108**, 12372–12377 (2011).
10. Martin, M. Rewriting the mathematics of tumor growth. *J. Natl Cancer Inst.* **103**, 1564–1565 (2011).
11. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.* Chapter 10, Unit 10.11 (2008).
12. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
13. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
This is a description and the first results of the ICGC, a worldwide endeavour to characterize a wide range of tumours by next-generation sequencing.
14. The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genet.* **45**, 1113–1120 (2013).
15. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
This is a genome-wide encyclopaedia of structural and regulatory elements in the genome.
16. Quigley, D. A. *et al.* The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. *Mol. Oncol.* **8**, 273–284 (2013).
17. Fletcher, M. N. C. *et al.* Master regulators of FGFR2 signalling and breast cancer risk. *Nature Commun.* **4**, 2464 (2013).
18. Brower, V. Epigenetics: Unravelling the cancer code. *Nature* **471**, S12–13 (2011).
19. Chin, L., Andersen, J. N. & Futreal, P. A. Cancer genomics: from discovery science to personalized medicine. *Nature Med.* **17**, 297–303 (2011).
20. Yuan, Y. *et al.* Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.* **4**, 157ra143–157ra143 (2012).
21. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30**, 1234–1248 (2012).
22. Kilpinen, S. *et al.* Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol.* **9**, R139 (2008).
23. Wong, A. K. *et al.* IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res.* **40**, W484–W490 (2012).
24. Engreitz, J. M., Daigle, B. J., Marshall, J. J. & Altman, R. B. Independent component analysis: mining microarray data for fundamental human gene expression modules. *J. Biomed. Inform.* **43**, 932–944 (2010).
25. Engreitz, J. M. *et al.* ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression. *Bioinformatics* **27**, 3317–3318 (2011).
26. Rhodes, D. R. *et al.* ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1–6 (2004).
27. Madhavan, S. *et al.* Rembrandt: helping personalized medicine become a reality through integrative translational research. *Mol. Cancer Res.* **7**, 157–167 (2009).
This paper describes integrated genomic analyses in medicine.
28. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
29. Saito, R. *et al.* A travel guide to Cytoscape plugins. *Nature Methods* **9**, 1069–1076 (2012).
30. Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nature Protocol.* **2**, 2366–2382 (2007).
This paper describes a widely used space for genomic analysis and visualization.
31. Gundem, G. *et al.* IntOGen: integration and data mining of multidimensional oncogenomic data. *Nature Methods* **7**, 92–93 (2010).
32. Gonzalez-Perez, A. & Lőpez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).
33. Margolin, A. A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5**, 181re1–181re1 (2013).
34. Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. Computational solutions to large-scale data management and analysis. *Nature Rev. Genet.* **11**, 647–657 (2010).
35. Quigley, D. & Balmain, A. Systems genetics analysis of cancer susceptibility: from mouse models to humans. *Nature Rev. Genet.* **10**, 651–657 (2009).
36. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
This paper describes an integration of next-generation sequencing data from DNA and RNA levels that reveals the structure of many regulatory elements.
37. Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541 (2006).
38. Lando, M. *et al.* Gene dosage, expression, and ontology analysis identifies driver genes in the carcinogenesis and chemoradioresistance of cervical cancer. *PLoS Genet.* **5**, e1000719 (2009).
39. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
40. Sun, Z. *et al.* Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS ONE* **6**, e17490 (2011).
41. Ovaska, K. *et al.* Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med.* **2**, 65 (2010).
42. Aure, M. R. *et al.* Identifying in-*trans* process associated genes in breast cancer by integrated analysis of copy number and expression data. *PLoS ONE* **8**, e53014 (2013).
43. Chibon, F. *et al.* Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nature Med.* **16**, 781–787 (2010).
44. Chari, R., Coe, B. P., Vucic, E. A., Lockwood, W. W. & Lam, W. L. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Syst. Biol.* **4**, 67 (2010).
45. Louhimo, R. & Hautaniemi, S. CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics* **27**, 887–888 (2011).
46. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
47. Shen, Y., Sun, W. & Li, K.-C. Dynamically weighted clustering with noise set. *Bioinformatics* **26**, 341–347 (2010).
48. Shen, R. *et al.* Integrative subtype discovery in glioblastoma using iCluster. *PLoS ONE* **7**, e35236 (2012).
49. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
50. Yuan, Y., Savage, R. S. & Markowetz, F. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.* **7**, e1002227 (2011).
51. Bøvelstad, H. M. *et al.* Predicting survival from microarray data—a comparative study. *Bioinformatics* **23**, 2080–2087 (2007).
52. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. Series B.* **58**, 267–288 (1996).
53. Nowak, G., Hastie, T., Pollack, J. R. & Tibshirani, R. A fused lasso latent feature model for analyzing multi-sample aCGH data. *Biostatistics* **12**, 776–791 (2011).
54. Mankoo, P. K., Shen, R., Schultz, N., Levine, D. A. & Sander, C. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS ONE* **6**, e24709 (2011).
55. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc.: Series B (Statist. Methodol.)* **67**, 301–320 (2005).
56. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nature Genet.* **36** 1090–1098 (2004).
This landmark publication establishes the principles of identification of regulatory modules.
57. Kelder, T. *et al.* WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* **40**, D1301–D1307 (2012).

58. Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nature Rev. Genet.* **9**, 509–515 (2008).
59. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
60. Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. & Müller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* **24**, i223–i231 (2008).
61. Qiu, Y.-Q., Zhang, S., Zhang, X.-S. & Chen, L. Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics* **11**, 26 (2010).
62. Guo, Z. *et al.* Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics* **23**, 2121–2128 (2007).
63. Chuang, H.-Y. *et al.* Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. *Blood* **120**, 2639–2649 (2012).
64. Doniger, S. W. *et al.* MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* **4**, R7 (2003).
65. Tarca, A. L. *et al.* A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75–82 (2009).
66. Efroni, S., Schaefer, C. F. & Buetow, K. H. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS ONE* **2**, e425 (2007).
67. Drier, Y., Sheffer, M. & Domany, E. Pathway-based personalized analysis of cancer. *Proc. Natl Acad. Sci. USA* **110**, 6388–6393 (2013).
68. Huttenhower, C. *et al.* Detailing regulatory networks through large scale data integration. *Bioinformatics* **25**, 3267–3274 (2009).
69. Huttenhower, C. *et al.* Exploring the human genome with functional maps. *Genome Res.* **19**, 1093–1106 (2009).
70. Mayer, C.-D., Lorent, J. & Horgan, G. W. Exploratory analysis of multiple omics datasets using the adjusted RV coefficient. *Stat. Appl. Genet. Mol. Biol.* **10**, Article 14 (2011).
71. Quigley, D. A. *et al.* Genetic architecture of mouse skin inflammation and tumour susceptibility. *Nature* **458**, 505–508 (2009).
72. Lê Cao, K.-A., González, I. & Déjean, S. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* **25**, 2855–2856 (2009).
73. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
74. Margolin, A. A., Wang, K., Califano, A. & Nemenman, I. Multivariate dependence and genetic networks inference. *IET Syst. Biol.* **4**, 428–440 (2010).
75. Margolin, A. A. & Califano, A. Theory and limitations of genetic network inference from microarray data. *Ann. NY Acad. Sci.* **1115**, 51–72 (2007).
76. Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques.* (Massachusetts Institute of Technology, 2009).
This study describes one of the basic approaches for studying gene–gene dependencies.
77. Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nature Genet.* **44**, 841–847 (2012).
This paper describes a fundamental attempt to identify genotype–phenotype interactions.
78. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18** (Suppl. 1), S233–240 (2002).
79. Breitling, R., Amtmann, A. & Herzyk, P. Graph-based iterative Group Analysis enhances microarray interpretation. *BMC Bioinformatics* **5**, 100 (2004).
80. Ideker, T. & Krogan, N. J. Differential network biology. *Mol. Syst. Biol.* **8**, 565 (2012).
81. Stingo, F. C. & Vannucci, M. Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics* **27**, 495–501 (2011).
82. Bauer, S., Gagneur, J. & Robinson, P. N. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.* **38**, 3523–3532 (2010).
83. Newton, M. A., He, Q. & Kendziorski, C. A model-based analysis to infer the functional content of a gene list. *Stat. Appl. Genet. Mol. Biol.* **11**, http://dx.doi.org/10.2202/1544-6115.1716 (2012).
84. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genet.* **34**, 166–176 (2003).
85. Segal, E., Friedman, N., Kaminski, N., Regev, A. & Koller, D. From signatures to models: understanding cancer using microarrays. *Nature Genet.* **37** S38–S45 (2005).
86. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).
This paper describes an application of approaches from the probabilistic graphical models in the identification of pathways or dependencies deviating from a given norm.
87. Kristensen, V. N. *et al.* Integrated molecular profiles of invasive breast tumors and ductal carcinoma *in situ* (DCIS) reveal differential vascular and interleukin signaling. *Proc. Natl Acad. Sci. USA* **109**, 2802–2807 (2012).
88. Ferkingstad, E., Frigessi, A. & Lyng, H. Indirect genomic effects on survival from gene expression data. *Genome Biol.* **9**, R58 (2008).
89. Imoto, S. *et al.* Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *J. Bioinform. Comput. Biol.* **2**, 77–98 (2004).
90. Bottolo, L. *et al.* Bayesian detection of expression quantitative trait loci hot spots. *Genetics* **189**, 1449–1459 (2011).
91. Akavia, U. D. *et al.* An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).
92. Birtwistle, M. R. *et al.* Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. *Mol. Syst. Biol.* **3**, 144 (2007).
93. Nik-Zainal, S. A. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
94. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
95. Cancer, Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
96. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nature Genet.* **45**, 1127–1133 (2013).
97. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature Genet.* **45**, 1134–1140 (2013).
98. Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genet.* **45**, 1113–1120 (2013).
99. Newman, M. E. J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* **69**, 066133 (2004).
100. Louhimo, R., Lepikhova, T., Monni, O. & Hautaniemi, S. Comparative analysis of algorithms for integration of copy number and expression data. *Nature Methods* **9**, 351–355 (2012).
101. Solvang, H. K., Lingjærde, O. C., Frigessi, A., Børresen-Dale, A.-L. & Kristensen, V. N. Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer. *BMC Bioinformatics* **12**, 197 (2011).
102. Heiser, L. M. *et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl Acad. Sci. USA* **109**, 2724–2729 (2012).
103. Hoshino, D. *et al.* Network analysis of the focal adhesion to invadopodia transition identifies a PI3K-PKCα invasive signaling axis. *Sci. Signal.* **5**, ra66 (2012).
104. Stronach, E. A. *et al.* DNA-PK mediates AKT activation and apoptosis inhibition in clinically acquired platinum resistance. *Neoplasia* **13**, 1069–1080 (2011).
105. Mok, T. S. *et al.* Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N. Engl. J. Med.* **361**, 947–957 (2009).
106. Shepherd, F. A. *et al.* Erlotinib in previously treated non-small-cell lung cancer. *N. Engl. J. Med.* **353**, 123–132 (2005).
107. Piccart-Gebhart, M. J. *et al.* Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N. Engl. J. Med.* **353**, 1659–1672 (2005).
108. Romond, E. H. *et al.* Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N. Engl. J. Med.* **353**, 1673–1684 (2005).
109. Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).
110. Jonker, D. J. *et al.* Cetuximab for the treatment of colorectal cancer. *N. Engl. J. Med.* **357**, 2040–2048 (2007).
111. Karapetis, C. S. *et al.* K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N. Engl. J. Med.* **359**, 1757–1765 (2008).
112. Iadevaia, S., Lu, Y., Morales, F. C., Mills, G. B. & Ram, P. T. Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis. *Cancer Res.* **70**, 6704–6714 (2010).
113. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
114. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
115. Cooper, S. *et al.* Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760 (2010).
116. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nature Methods* **10**, 221–227 (2013).
117. Cheng, W.-Y., Ou Yang, T.-H. & Anastassiou, D. Biomolecular events in cancer revealed by attractor metagenes. *PLoS Comput. Biol.* **9**, e1002920 (2013).
118. Cheng, W.-Y., Ou Yang, T.-H. & Anastassiou, D. Development of a prognostic model for breast cancer survival in an open challenge environment. *Sci. Transl. Med.* **5**, 181ra50 –181ra50 (2013).
119. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
120. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
121. Sørlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA* **100**, 8418–8423 (2003).
122. Russnes, H. G. *et al.* Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci. Transl. Med.* **2**, 38ra47–38ra47 (2010).
123. Chin, S.-F. *et al.* Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. *Oncogene* **26**, 1959–1970 (2007).
124. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
125. Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
126. Naume, B. *et al.* Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. *Mol. Oncol.* **1**, 160–171 (2007).
127. Nordgard, S. H. *et al.* Genome-wide analysis identifies 16q deletion associated with survival, molecular subtypes, mRNA expression, and germline haplotypes in breast cancer patients. *Genes Chromosomes Cancer* **47**, 680–696 (2008).
128. Rønneberg, J. A. *et al.* Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer. *Mol. Oncol.* **5**, 61–76 (2011).
129. Enerly, E. *et al.* miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. *PLoS ONE* **6**, e16915 (2011).
130. Joshi, H., Bhanot, G., Børresen-Dale, A.-L. & Kristensen, V. N. Potential tumorigenic programs associated with TP53 mutation status reveal role of VEGF pathway. *Br. J. Cancer* **107**, 1722–1728 (2012).

131. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
132. Sun, Z. *et al.* Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Med. Genom.* **4**, 84 (2011).
133. Strehl, A. & Ghosh, J. Cluster ensembles — a knowledge reuse framework for combining partitionings. *Journal of Machine Learning* **3**, 583–617 (2002).
134. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learn.* **52**, 91–118 (2003).
135. Collisson, E. A. *et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature Med.* **17**, 500–503 (2011).
136. Lancichinetti, A. & Fortunato, S. Consensus clustering in complex networks. *Sci. Rep.* **2**, 336 (2012).
137. Lee, M. & Kim, Y. CHESS (CgHExpreSS): a comprehensive analysis tool for the analysis of genomic alterations and their effects on the expression profile of the genome. *BMC Bioinformatics* **10**, 424 (2009).
138. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).
139. Leday, G. G. R. & van de Wiel, M. A. PLRS: a flexible tool for the joint analysis of DNA copy number and mRNA expression data. *Bioinformatics* **29**, 1081–1082 (2013).
140. Chen, B.-J. *et al.* Harnessing gene expression to identify the genetic basis of drug resistance. *Mol. Syst. Biol.* **5**, 310 (2009).
141. Yuan, Y., Curtis, C., Caldas, C. & Markowetz, F. A. Sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **9**, 947–954 (2012).
142. Carro, M. S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).
143. Saadi, A. *et al.* Stromal genes discriminate preinvasive from invasive disease, predict outcome, and highlight inflammatory pathways in digestive cancers. *Proc. Natl Acad. Sci. USA* **107**, 2177–2182 (2010).
144. Hamatani, T. *et al.* Global gene expression analysis identifies molecular pathways distinguishing blastocyst dormancy and activation. *Proc. Natl Acad. Sci. USA* **101**, 10326–10331 (2004).
145. Draghici, S. *et al.* A systems biology approach for pathway level analysis. *Genome Res.* **17**, 1537–1545 (2007).
146. Engström, P. G. *et al.* Digital transcriptome profiling of normal and glioblastoma-derived neural stem cells identifies genes associated with patient survival. *Genome Med.* **4**, 76 (2012).
147. Wu, J., Mao, X., Cai, T., Luo, J. & Wei, L. KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.* **34**, W720–W724 (2006).
148. Xie, C. *et al.* KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **39**, W316–W322 (2011).
149. Li, C. *et al.* SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res.* **37**, e131–e131 (2009).
150. Chang, H.-T. *et al.* Comprehensive analysis of microRNAs in breast cancer. *BMCGenomics* **13**, S18 (2012).
151. Tamborero, D., Lopez-Bigas, N. & Gonzalez-Perez, A. Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PLoS ONE* **8**, e55489 (2013).
152. Warsow, G. *et al.* ExprEssence—revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Syst. Biol.* **4**, 164 (2010).
153. Deshpande, R., Sharma, S., Verfaillie, C. M., Hu, W.-S. & Myers, C. L. A scalable approach for discovering conserved active subnetworks across species. *PLoS Comput. Biol.* **6**, e1001028 (2010).
154. Goffard, N., Frickey, T. & Weiller, G. PathExpress update: the enzyme neighbourhood method of associating gene-expression data with metabolic pathways. *Nucleic Acids Res.* **37**, W335–W339 (2009).
155. Bryant, W. A., Sternberg, M. J. E. & Pinney, J. W. AMBIENT: Active Modules for Bipartite Networks—using high-throughput transcriptomic data to dissect metabolic response. *BMC Syst. Biol.* **7**, 26 (2013).
156. Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z. & Wild, D. L. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**, 3290–3297 (2012).
157. Brodtkorb, M. *et al.* Whole-genome integrative analysis reveals expression signatures predicting transformation in follicular lymphoma. *Blood*, **123**, 1051–1054 (2014).

Acknowledgements
The authors thank numerous collaborators, most notably D. Quigley, R. Sachidanandam, S. Hautaniemi, P. van Loo and C. Vaske for the critical reading of the manuscript and for sharing their overview of the field and valuable discussions. Special thanks to C. Perou and C. Creighton of The Cancer Genome Atlas (TCGA) and O. Rueda and C. Caldas of the METABRIC study, as well as M. M. Holmen, from Oslo University Hospital for providing original images. The authors also thank the Norwegian Cancer Society, the K.G. Jebsen Foundation, the Norwegian Research Council, Health Region South East, and the Norwegian Radium Hospital's Foundation for financial support over many years.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
Databases and sites for integrating tools:
Cancer Genome Project: www.sanger.ac.uk/genetics/CGP/
Catalogue of Somatic Mutations in Cancer (COSMIC) database: http://www.sanger.ac.uk/genetics/CGP/cosmic/
ENCyclopedia Of DNA Elements (ENCODE): http://genome.ucsc.edu/ENCODE/
International Cancer Genome Consortium (ICGC): www.icgc.org/
NCI/TCGA: http://cancergenome.nih.gov
The Cancer Genome Atlas (TCGA): www.cancergenome.nih.gov

Storage and compute spaces:
Bioconductor: http://www.bioconductor.org/
Bionimbus: http://www.bionimbus.org/
CytoScape: http://www.cytoscape.org/
Federation of SAGE: http://sagebase.org/
Synapse: https://synapse.prod.sagebase.org/

Protein–protein interactions:
HPRD: www.hprd.org/
Kyoto Encyclopedia of Genes and Genomes (KEGG): www.genome.jp/kegg
MIPS (Mammalian protein–protein interaction): http://mips.helmholtz-muenchen.de/proj/ppi/
PID Pathway Interaction Database (NCI): www.pid.nci.nih.gov
Reactome: www.reactome.org
WikiPathways: http://wikipathways.org/

Annotation, visualization and integrated discovery:
Biowaver: http://sonorus.princeton.edu/bioweaver/
DAVID: http://david.abcc.ncifcrf.gov
GOLEM: http://reducio.princeton.edu/GOLEM/
GRIFn: http://reducio.princeton.edu/GRIFn/
HEFalMp: http://hefalmp.princeton.edu/
Mefit: http://avis.princeton.edu/mefit
MsigDB Molecular Signatures Database: www.broadinstitute.org/gsea/msigdb/index.jsp
Oncomine: https://www.oncomine.org/resource/login.html
Rembrandt: http://cabig.cancer.gov/action/collaborations/rembrandt/
Search-Based Exploration of Expression Compendium (SEEK): http://seek.princeton.edu
Sleipnir: http://libsleipnir.bitbucket.org/
Summary of gene ontology tools: http://www.geneontology.org/GO.tools.microarray.shtml

Omics integration:
Combinatorial ALgorithm for Expression and Sequence-based Cluster Extraction (COALESCE): http://reducio.princeton.edu/cm/coalesce
COpy Number and EXpression In Cancer (CONNEXIC): http://www.c2b2.columbia.edu/danapeerlab/html/software.html
DR-Integrator: http://pollacklab.stanford.edu/
IntOGen: http://bg.upf.edu/group/tools.php#intogen
Magellan: http://cabig.nci.nih.gov/
OncoDrive: http://bg.upf.edu/blog/tag/oncodrive/
PAthway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM): http://sbenz.github.com/Paradigm

ALL LINKS ARE ACTIVE IN THE ONLINE PDF