Human Molecular Genetics, 2010, Vol. 19, Review Issue 2 ***R188–R196***
doi:10.1093/hmg/ddq391
Advance Access published on September 15, 2010

Analysis of next-generation genomic data in cancer: accomplishments and challenges

Li Ding, Michael C. Wendl, Daniel C. Koboldt and Elaine R. Mardis*

Department of Genetics, The Genome Center at Washington University School of Medicine, 4444 Forest Park Blvd, Box 8501, St Louis, MO 63108, USA

Received August 12, 2010; Revised and Accepted September 6, 2010

The application of next-generation sequencing technology has produced a transformation in cancer genomics, generating large data sets that can be analyzed in different ways to answer a multitude of questions about the genomic alterations associated with the disease. Analytical approaches can discover focused mutations such as substitutions and small insertion/deletions, large structural alterations and copy number events. As our capacity to produce such data for multiple cancers of the same type is improving, so are the demands to analyze multiple tumor genomes simultaneously growing. For example, pathway-based analyses that provide the full mutational impact on cellular protein networks and correlation analyses aimed at revealing causal relationships between genomic alterations and clinical presentations are both enabled. As the repertoire of data grows to include mRNA-seq, non-coding RNA-seq and methylation for multiple genomes, our challenge will be to intelligently integrate data types and genomes to produce a coherent picture of the genetic basis of cancer.

INTRODUCTION

Next-generation sequencing (NGS) platforms are revolutionizing cancer genomics research. Their ever-increasing data generation capabilities now enable fast, high-depth sequencing of human cancer genomes. These technologies hold enormous promise for the study of focused mutations (point mutations and small indels), copy number alterations (CNAs) and structural variants, including fusion genes in cancer genomes. At the same time, data volume and relatively short-read lengths have also presented difficulties for data analysis. These challenges have stimulated the development of new computational tools for every NGS data analysis task from variation detection and assembly to downstream biological and functional analyses. In this review, we will discuss some of these tools and their application in cancer genomics studies.

POST-ALIGNMENT CONSIDERATIONS OF SHORT-READ, PAIRED-END DATA

NGS platforms generate hundreds of millions of sequence reads per instrument run. Following each run, standardized instrument manufacturer-defined pipelines process the signal-based data into sequence reads. These pipelines include routine quality control on a per-lane or per-region basis to provide metrics of success for each data set. One secondary aspect of quality control that must be addressed prior to downstream analysis is that of read duplication, in which the same DNA fragment begets multiple reads or read pairs. This artifact has been attributed to the initial PCR-based library amplification steps and can affect as many as 10% of read pairs (1). Removal of duplicate reads is advantageous to most downstream analytical approaches, since these reads may contain PCR-introduced errors that masquerade as variant nucleotides, for example. The Picard suite (http://picard.sourceforge.net/) includes tools for the de-duplication process that operate on both single-end and paired-end data. In addition to de-duplication, data sets containing reads with insufficient read length, base quality, mapping quality or paired-end reads having an atypical distribution of insert sizes also should be flagged/soft-trimmed and discarded when necessary.

POINT MUTATION DISCOVERY IN PAIRED TUMOR AND NORMAL GENOMES

One of the predominant applications of NGS has been the comparison of tumor genomes with their matched constitutional genomes, for the purpose of identifying tumor-unique

*To whom correspondence should be addressed. Tel: +314 2861805; Fax: +314 2861810; Email: emardis@wustl.edu

Table 1. Selected analysis tools for NGS of cancer genomes

Software	Description	URL
SNP detection		
SAMtools	Bayesian SNP calling	http://samtools.sourceforge.net
SOAPsnp	Bayesian SNP calling	http://soap.genomics.org.cn
SNVmix	SNP calling by probabilistic binomial mixture model	http://www.bcgsc.ca/platform/bioinfo/software/SNVMix
Somatic mutation detection		
VarScan	Heuristic germline and somatic variant calling	http://varscan.sourceforge.net
SomaticSniper	Bayesian somatic variant calling	http://genome.wustl.edu/tools/cancer-genomics
Small insertions/deletions (indels)		
Pindel	Indel prediction with paired-end data	http://www.ebi.ac.uk/~kye/pindel
GATK	Heuristic germline and somatic indel calling	ftp://ftp.broadinstitute.org/pub/gsa/GenomeAnalysisTK
CNV detection		
EWT	CNV calling with EWT	http://genome.cshlp.org/content/19/9/1586
SegSeq	CNV calling with local change-point analysis and merging	http://www.broad.mit.edu/cancer/pub/solexa_copy_numbers
CMDS	RCNA calling in sample populations	https://dsgweb.wustl.edu/qunyuan/software/cmds
Structural variation (SV) detection and assembly		
Geometric analysis of structural variants	Geometric method for SV detection	http://cs.brown.edu/people/braphael/software.html
BreakDancer	Prediction of SVs with paired-end Illumina data	http://genome.wustl.edu/tools/cancer-genomics
TIGRA	De novo assembly of SV breakpoints	http://genome.wustl.edu/tools/cancer-genomics
Data visualization		
IGV	Integrative genomics viewer for next-gen sequencing data	http://www.broadinstitute.org/igv/StartIGV
Pairoscope	BAM-driven visualization of predicted SVs	http://pairoscope.sourceforge.net
Pathway analysis		
PathScan	Pathway analysis with convolution and Fisher-Lancaster theory	http://genome.wustl.edu/tools/cancer-genomics
HotNet	Pathway analysis with diffusion and permutation testing	http://cs.brown.edu/people/braphael/software.html
Netbox	Pathway analysis with hypergeometric test and edge algorithm	http://cbio.mskcc.org/netbox

(somatic) variation in an unbiased, genome-wide fashion. Numerous single nucleotide variant (SNV) detection algorithms for NGS data have been developed in recent years (2–5). SAMtools (4) and SOAPsnp (5) utilize Bayesian statistics to compute probabilities of all possible genotypes. In principle, these tools could be adapted for somatic mutation calling in cancer studies where both tumor and matched normal are sequenced. However, both expect a heterozygous variant allele frequency of 50%. Although valid for germline sites, this figure does not hold for somatic sites in most tumors due to normal contamination and/or tumor heterogeneity. Development is now focusing on callers designed specifically for somatic mutations. One example is SNVmix (6), which utilizes a probabilistic Binomial mixture model and adjusts to deviation of allelic frequencies using an expectation maximization algorithm. SNVmix has been applied to genomic and RNA sequencing data from a lobular breast tumor, leading to the discovery of a set of novel mutations in breast cancer (7). However, SNVmix remains limited, in that it does not attempt the next step of deriving the probability of a given site as somatic by utilizing tumor and normal data at the same time (Table 1).

We developed two somatic point mutation discovery algorithms: VarScan (2) and SomaticSniper (D. Larson *et al.*, manuscript in preparation). VarScan determines overall genome coverage, as well as the average base quality and number of strands observed for each allele. Read counts are used to infer variant allele frequency and in calculating somatic status using Fisher's exact test. VarScan is well suited for somatic mutation detection in data sets having varying coverage depths, such as from targeted capture. SomaticSniper uses Bayesian theory to calculate the probability of differing genotypes in the tumor and normal samples, assuming independence of both genotypes and data in the samples. It reports a phred-scaled probability that the tumor and normal were identical as the 'somatic' score. SomaticSniper has been used in several recent studies (8,9), leading to the discovery of *IDH1* R132C mutations in acute myeloid leukemia (AML), as well as numerous other novel somatic mutations in genes not previously reported to harbor genetic mutations (8). These findings are consistent with earlier large-scale Sanger-based studies, suggesting that cancer is characterized by a small number of frequently mutated genes and a long tail of infrequent mutations in a large number of genes (10,11). VarScan and SomaticSniper have been applied to the analysis of hundreds of tumor and normal pairs for various projects such as The Cancer Genome Atlas Project (http://cancergenome.nih.gov/) and the Pediatric Cancer Genome Project (http://www.pediatriccancergenomeproject.org/site/).

FOCUSED INSERTION/DELETION DETECTION

Although existing alignment tools are regarded as adequate for mapping reads that contain SNVs, they generally lack the necessary accuracy and sensitivity for reads that overlap indels or structural variants. Most tools by default allow only two mismatches and no gaps in the 'seeded' regions (e.g. the first 28 bp in a read), which prohibits indel-containing reads from aligning to the reference. Paired-end mapping is tremendously helpful in identifying larger indels, when read pair alignment occurs in flanking regions and allows the inference of altered intervening sequences.

Pindel (12) takes a pattern growth approach borrowed from protein data analysis (13), to detect breakpoints of indels from paired-end reads. Our experience suggests that Pindel achieves

high specificity but that it suffers from lower sensitivity, primarily due to not allowing mismatches during the pattern matching process. SAMtools summarizes short indel information by correcting the effect of flanking tandem repeats and it tends to produce a large number of indel calls. Local *de novo* assembly or multiple alignments around the candidate indel sites reduces the number of false-positive indels. This process was used in the analysis of whole-genome data from a basal-like breast cancer (9) and is currently one of the methods utilized in our pipeline for indel detection.

Like VarScan, the GATK Indel Genotyper (http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit) is based on a set of heuristic cutoffs for indel calling. It collects raw statistics, such as coverage, numbers of indel-supporting reads, read mapping qualities and mismatch counts, which are useful for post-filtering of the initial calls. Currently, somatic indel identification is generally achieved by simple subtraction of indels also found present in the normal. A probabilistic model for somatic indel detection is urgently needed.

IDENTIFYING CNAS IN TUMOR GENOMES

CNAs, such as large amplifications or deletions of chromosomal segments, represent an important type of somatic alteration in cancer. SNP genotyping data have long been utilized for studying CNAs in cancer and the CNA landscape across multiple cancer types has been recently reported (14,15). However, whole-genome sequencing of tumor and matched normal samples enables the identification of CNAs at a scale and precision unmatched by traditional array-based approaches (16–20). Accurate inference of the copy number from sequence data requires normalization procedures to address certain biases inherent in NGS data. For example, GC content bias arises from mechanistic differences between NGS platforms (1,21,22), whereas read mapping bias originates from the computational difficulties of assigning relatively short sequences (25–450 bp) to their correct locations in a large, complex reference genome. Approaches have been developed for both GC-based coverage normalization (23) and mapping bias (24). Following these corrections, the unique (non-redundant) read depth can serve as a quantitative measure of genome copy number (17,19,20). Several methods have been developed for identifying significant copy number changes within a single genome or between a tumor sample and its matched normal. Campbell *et al.* (24) adapted the circular binary segmentation algorithm, originally developed for SNP array data, to identify statistically significant copy number changes (24). Event-wise testing (EWT) (23) was developed as a three-step approach to identify copy number variants (CNVs) from read depth data. One drawback of all window-based segmentation methods is that structural variation (SV) breakpoints cannot be localized more finely than the boundaries of the windows. The SegSeq algorithm (25) uses an alternative approach that combines local change-point analysis with a subsequent merging procedure to join adjacent chromosomal regions. The merged segments are then tested for significance between tumor and normal samples.

Identifying regions of recurrent CNA (RCNA) within or across tumor types offers a powerful system for localizing cancer-causing genes. In principle, such regions could be identified with a two-step process: (i) call significant CNAs in each tumor genome and then (ii) perform cross-sample analysis. This approach, however, carries a heavy computational burden and also loses statistical power due to the segmentation and normalization performed at the level of individual samples. To address this issue, we developed the population-based correlation matrix diagonal segmentation (CMDS) method that identifies RCNAs based on a between-chromosomal-site correlation analysis (26). CMDS adopts a diagonal transformation strategy to reduce compute time and is well suited to high-resolution studies of large sample populations.

SOMATIC STRUCTURAL VARIANT DISCOVERY

Beyond CNAs, structural changes in chromosomes such as inversions and translocations represent a major source of somatic variation in cancer genomes. The majority of known cancer genes are altered by rearrangement, resulting either in a fusion transcript (e.g. BCR-ABL in leukemia) or in transcriptional dysregulation (27). Yet, the discovery and characterization of somatic SV in cancer genomes, using NGS reads, remains challenging. Cytogenetics, spectral karyotyping and fluorescent *in situ* hybridization have previously identified large chromosomal events. SNP and comparative genome hybridization microarrays provide limited resolution and also miss copy-number-neutral events as well as most translocations. End-sequencing profiling (28) of BAC libraries has revealed the complex architectures of several human cancers (29–32), but remains costly and laborious.

Conversely, tumor genome sequencing by NGS platforms offers the power to detect somatic rearrangements and to characterize their breakpoints with unprecedented resolution. The identification and analysis of read pairs that do not align as anticipated to the genome point to a wide range of SV events—deletions, tandem or inverted duplications, inversions, insertions and translocations—and this approach already has been applied to studies of glioblastoma (33), breast cancer (9), melanoma (34) and lung cancer (24,35,36) genomes. While sensitive, the paired-end strategy does tend to yield many false positives due either to sequencing errors or to read mis-alignments, especially to repetitive sequences.

Several programs are now available for SV analyses. For example, geometric analysis of structural variants attempts to precisely define breakpoints using a geometric bounding algorithm (37). We have developed a comprehensive pipeline for Illumina paired-end data centered around three components. The BreakDancer program (38) identifies candidate SVs, after which *de novo* assembly of both tumor- and normal-supporting reads is performed with the TIGRA package (K. Chen *et al*., submitted for publication) to remove false positives, precisely define breakpoints and determine the somatic status of each prediction. Finally, we review the evidence supporting each putative SV with the Pairoscope visualizer (http://pairoscope.sourceforge.net/) and the Integrative Genomic Viewer (http://www.broadinstitute.org/igv/).

VALIDATION OF GENETIC CHANGES IDENTIFIED IN CANCER GENOME STUDIES

A variety of factors affect the actual distribution of read depth in whole-genome sequencing which can often diverge

Table 2. Published tumor or tumor cell line genomes sequenced on next-generation platforms

Study	Tumor type	Tumor	Normal	SNVs	Indels	SVs	Notable altered genes
Ley *et al.*	Acute myeloid leukemia	32.7×	13.9×	8	2	—	NPM1, FLT3
Mardis *et al.*	Acute myeloid leukemia	23.3×	21.3×	10	2	—	NRAS, NPM1, IDH1
Shah *et al.*	Lobular breast cancer	43.1×	—	32	0	—	ERBB2, HAUS3
Ding *et al.*	Basal-like breast cancer	29.0×	38.8×	43	7	41	MYCBP2, TGFBI, CTNNA1
Pleasance *et al.*	Small-cell lung cancer cell line	39×	31×	134	2	58	TP53, RB1, CHD7
Pleasance *et al.*	Malignant melanoma cell line	40×	32×	292	0	37	BRAF, PTEN, SPDEF
Lee *et al.*	Non-small-cell lung cancer	60×	46×	392	—	43	KRAS, LRP1B, NF1, NEK9
Clark *et al.*	Glioblastoma cell line	29.5×	—	23	7	4	CDKN2A/B, PTEN, MLLT3

Summary of somatic SNVs, insertion/deletion variants (indels) and structural variants (SVs) identified from sequencing tumor genomes.

significantly from the idealized Poisson distribution (39). The issue is further exacerbated in targeted exome capture due to the differential efficiencies of capture oligos (40). Non-uniform coverage coupled with both systematic and non-systematic errors in NGS data present a challenge for mutation detection in regions with extremely high or low coverage. Many existing tools trade sensitivity for specificity and vice versa, rendering experimental validation, a necessary second step for maximizing both of these metrics. Two main validation/confirmation approaches have been used. One approach uses PCR-based amplification of suspect variant sites, followed by either Sanger sequencing or NGS. This approach has been used in many large-scale cancer studies (10,11,41–43), but is giving way to custom-targeted capture-based validation due to its reduced cost and increased throughput. Targeted capture can validate the majority of genetic changes identified through whole-genome sequencing (including SNVs, indels and SVs), as well as the focused mutations identified in exome sequencing data. For regions of CNA identified via NGS data analysis, the coincident analysis of SNP array data to identify regions of CNA typically is considered an appropriate means of orthogonal validation.

MUTATION RATE AND SPECTRUM

Somatic mutations and mutation signatures vary according to tumor type, as well as across individual tumors of the same tumor type (7,8,11,34–36,44–46) (Table 2). Even though the biological basis of these differences remains largely unknown, some studies suggest that they may be due to defects in DNA repair or to specific mutagenic exposures (34,35,43). Recent whole-genome sequencing and analysis of tumors from adult patients with AML, breast cancer and lung cancer suggest that mutation rates could range from 0.1 to over 10 mutations per megabase. These studies also revealed tumor type-specific signatures (7–9,34–36,41). Our recent analyses of pediatric cancer genomes have indicated a significantly lower mutation rate than that observed in adult cancers (data not shown). The accurate assessment of background mutation rate and mutation signature in different tumor types is critical and is a prerequisite for the analysis of significantly mutated genes and pathways. Mutation rate and spectrum have been estimated either using mutations identified in a selected genomic region or in a set of selected genes (e.g. a few hundred) in large-scale studies of colorectal cancers, breast cancers, lung cancers and glioblastomas (10,11,42,47). Clearly, these estimates are limited by both the breadth and depth of sequencing and mutation data, meaning that various statistical and computational corrections have to be applied to overcome sampling zeros and data sparseness. With the recent switch to NGS instruments, mutation data are now being obtained from sequencing whole genomes, exomes and transcriptomes. The whole-genome approach provides the most uniform coverage, making it the most suitable for estimating background mutation rate and revealing the mutation spectrum in individual tumors. Deep sequencing of tumor and matched normal genomes (e.g. 30×) usually captures >99% of SNVs, allowing the identification of nearly all somatic changes in the genome. Somatic mutations that likely are not under selection (e.g. mutations in non-exonic and non-conserved regions) provide the most accurate estimation of background rate.

Since the first report of whole-genome sequencing and analysis of a tumor/normal pair in 2008 (45), the rapidity with which primary data sets can be produced for whole-genome analysis has grown exponentially. This poses the interesting dilemma of being able to pace through primary sequencing of a cohort in relatively short order, only to be challenged by the lack of tools that enable a higher level look across multiple genomes. Ideally, this rich data resource will be mined to address key questions such as more rapidly identifying frequently mutated genes and common SV events, and then more comprehensively identifying the cellular pathways most commonly impacted by somatic variation.

SIGNIFICANTLY MUTATED GENE ANALYSIS

Three statistical methods have been described for identifying genes that are mutated above background levels (11,41,42). The first approach uses a standard one-tailed binomial test and treats all mutations equally with a uniform probability for each position in a gene. The second approach is a gene-specific variation of the binomial test, the advantage being that it uses gene-specific background rates instead of a single, global rate. The third approach is inspired by the analysis of mutation patterns in a variety of cancer types (44,48), demonstrating that (i) transitions generally occur at a higher rate than transversions, (ii) substitution rates are influenced by flanking sequences, a clear example being that cytosines in CpG dinucleotides have a significantly higher mutation

rate than cytosines in other sequence contexts and (iii) the rate of indel events is roughly an order of magnitude lower than the rate of substitutions. These observations suggest scoring mutations according to their prevalence. In this approach, independent tests are performed for different sequence mutation categories (e.g. A/T transitions, A/T transversions, C/G transitions, C/G transversions, CpG transitions, CpG transversions and indels). Then, methods like convolution, Fisher's test and likelihood test, can be used on the category-specific binomials to obtain an overall *P*-value. In all three approaches, the false discovery rate (FDR) for multiple-gene testing can be controlled using the standard Benjamini and Hochberg FDR procedure.

Comparative mutational analysis of significant genes can illustrate the common and unique players in different tumor types. For example, the analysis across five tumor types (glioblastoma, pancreatic cancer, breast cancer, lung adenocarcinoma and colon cancer) using Sanger-based data from large-scale cancer studies (10,11,41–43,49) revealed a high frequency of *KRAS* mutation in lung, pancreatic and colon cancers, but low frequency in breast cancer and glioblastoma. These results suggest that the latter two types might not be strongly related to smoking. Furthermore, low frequencies of *EGFR* mutations were found in colon, pancreatic and breast cancers. This may be due to high *KRAS* mutation rates in colon and pancreatic cancers and high rate of *ERBB2* alterations in breast cancer, releasing the selection pressure on *EGFR* alterations in these three cancer types. Notably, *STK11* is highly mutated in lung cancer from smokers, but not in other cancer types. Moreover, high frequencies of *PTEN*, *PIK3CA* and *PIK3R1* mutations have been found in glioblastoma, but not in lung and pancreatic cancers. *TP53* is the only uniformly mutated gene with high frequency in all five cancer types analyzed. Undoubtedly, novel significant genes in various cancer types will be discovered at a rapid pace using NGS methods, allowing more comprehensive cross-cancer type comparisons.

PATHWAY AND NETWORK ANALYSIS

Sequencing of many tumors allows an evaluation beyond individual genes to the consideration of pathways. Whole-genome sequencing of AML, breast, melanoma and lung tumors identified a large number of affected genes from diverse functional categories and biological processes (7–9,34–36,41) (Fig. 1). Two cancer samples often have little in common beyond a few mutated genes (41), but such heterogeneity may collapse at the pathway level when considering multiple tumor genomes. The presumptive mechanism here is mutation or other somatic alterations in a key subset of genes that alter a pathway such that it then results in tumor initiation or progression. Pathway analysis methods can be divided broadly into two groups: annotation-based approaches that use predefined pathway architectures and *de novo* procedures that do not presume such information.

Annotation-based methods are the somewhat simpler category in that they obtain pathway structures directly from a database like KEGG (50) and focus exclusively on the statistical testing problem. Various implementations simply pool mutation counts, for instance Fisher's test, on 2×2 tables of mutated and total bases within the pathway versus without, or binomial tests of the number of mutations per megabase versus an estimated background mutation rate, e.g. Group-CaMP (51). Unfortunately, the pooling procedure discards much important information, tending generally to make results appear more significant than what they actually are. Recent work concentrates on re-incorporating relevant information. For example, variation in gene lengths and even in background mutation rates means that genes have different Bernoulli's probabilities of mutation under the null hypothesis of random mutation. Moreover, properly combining sample-specific *P*-values into a study-wide composite value is equally critical. We recently proposed the PathScan algorithm (M. Wendl *et al.*, submitted for publication) based precisely on these two improvements. The former aspect is managed using mathematical *convolution*, whereas the latter relies on the Fisher–Lancaster theory. PathScan has reproduced the significantly mutated pathway list reported by the Tumor Sequencing Project (TSP) for lung adenocarcinoma data (11), but also has determined that several pathways identified as 'inconclusive' in the original study actually are not significantly mutated.

De novo methods by contrast begin with some catalog of interaction data, e.g. the Human Protein Reference Database (HPRD), to supply candidate edges and nodes to an initial 'seed node' upon which a network 'graph' will be built. For example, the NetBox algorithm (52) was used to study glioblastoma by first selecting known altered genes as seeds, then applying the hypergeometric test to check the likelihood that higher-degree nodes called 'linker genes' would connect to the observed number of altered genes solely by chance. Relevant subnetworks were then identified using a well-known edge relationship algorithm (53). In this study, NetBox was able to identify modules centered around RB, PI3K and TP53. A slightly different approach is taken by the HotNet algorithm (54), which uses the physical model of *diffusion* to identify neighborhoods of influence based on node distances and numbers of connections, with permutation testing for judging significance. This algorithm also identified a TP53-based pathway from glioblastoma data (42), as well as Notch and cell signaling pathways.

Both types of pathway analysis approaches will continue integrating broader genetic and epigenetic information as it is discovered. The whole pathway analysis enterprise depends on having sufficient data and consequent power of discovery. This type of analysis also will increase our ability to discern how tumor subtypes differ from one another, suggesting different clinical aspects of testing, classification and treatment. Even more importantly, it will help reveal existing commonalities across different cancers at the pathway level. These common denominators will have incredible clinical implications.

CONCLUSIONS AND CHALLENGES

Various tools for analyzing genetic changes in cancer based on heuristic and probabilistic models have been developed and applied in recent NGS-based studies. Although some

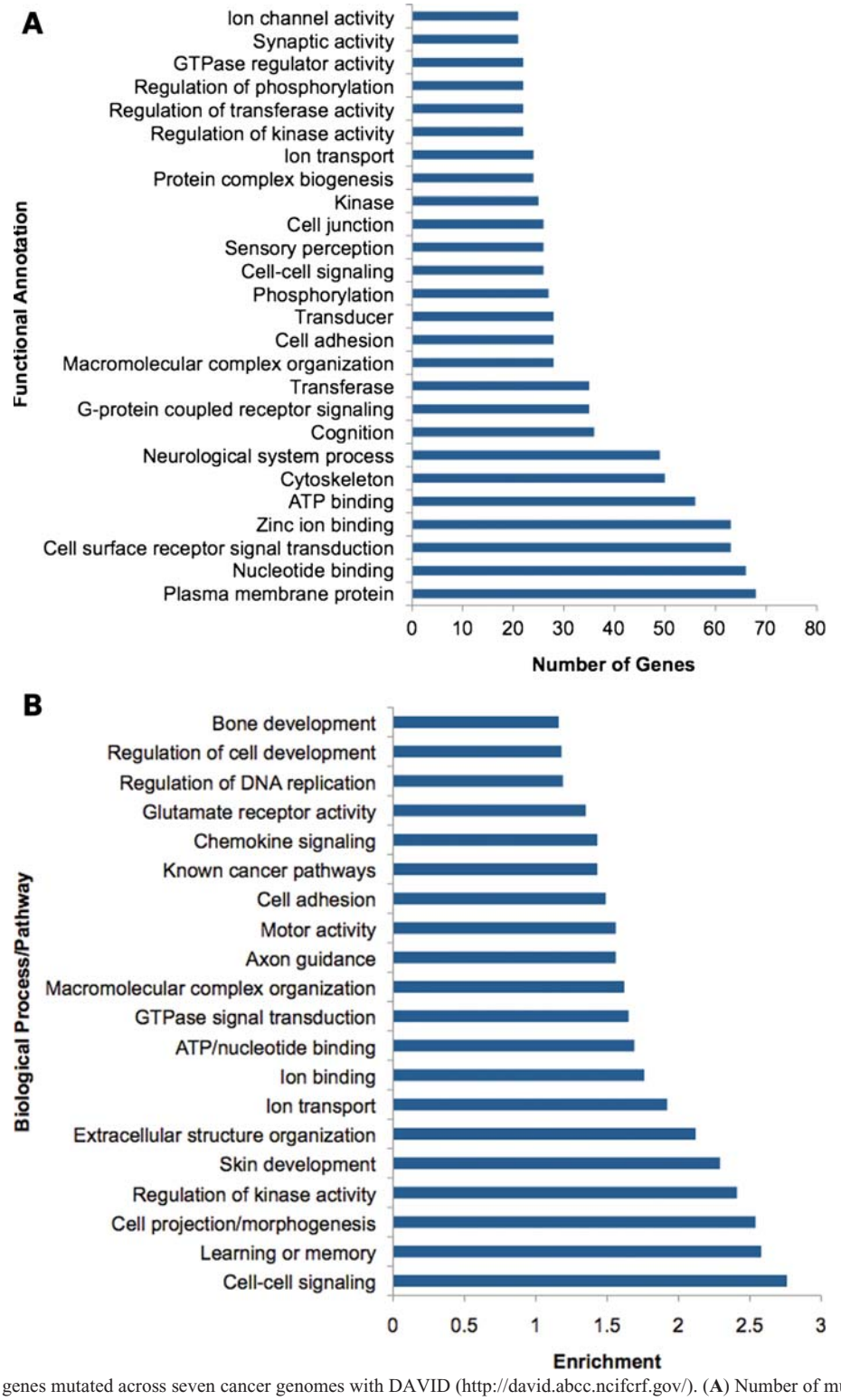

Figure 1. Analysis of 430 genes mutated across seven cancer genomes with DAVID (http://david.abcc.ncifcrf.gov/). (**A**) Number of mutated genes by GO terms of gene function. (**B**) Enriched biological processes and/or pathways among mutated genes according to DAVID clustering.

significant findings have been made, most of these tools are still under active development. Detection and evaluation of genetic events, including somatic mutations, CNAs and SVs, still is not a solved problem. Consequently, the use of several algorithms followed by validation often represents the best combination of sensitivity and specificity for mutation discovery in cancer studies.

We feel that the advancement of NGS-based cancer genomics will develop along two dimensions. In the first, NGS data will grow rapidly from whole genomes of a few tumors

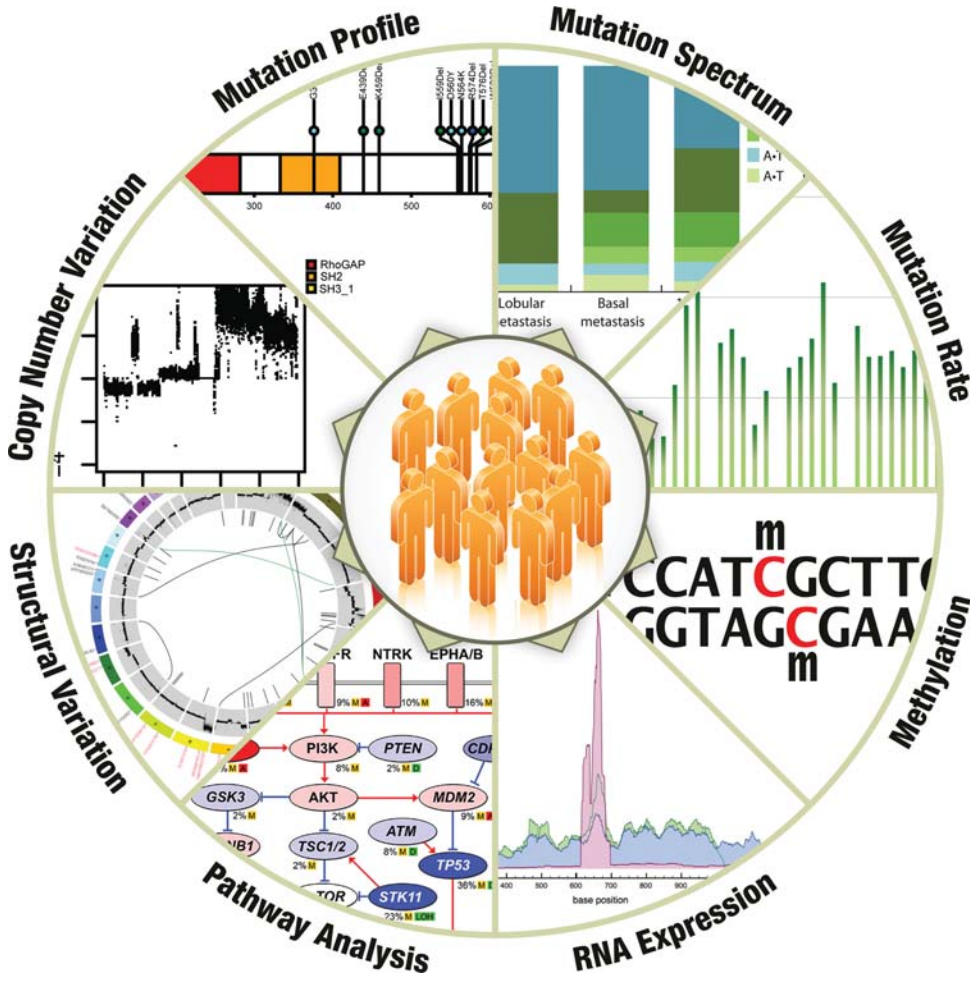

Figure 2. Landscape of cancer genomics analyses. NGS data will be generated for hundreds of tumors from all major cancer types in the near future. The integrated analysis of DNA, RNA and methylation sequencing data will help elucidate all relevant genetic changes in cancers.

to hundreds of tumors and from a few tumor types to almost all major types. This development will address the all-important issue of statistical power needed to detect the full range of somatic variations important in the disease development and progression. In the second dimension, NGS technologies will be broadly used for mRNA and non-coding RNA sequencing, as well as DNA methylation sequencing to help elucidate all the relevant biological contributions from these entities toward genomic dysregulation (Fig. 2). Although this explosion provides an unprecedented opportunity to systematically explore the entire spectrum of DNA and RNA changes in all major human cancers, it also represents a great challenge for cross-cancer type and cross-data type integration analysis. In our view, successful integration requires at least two key elements. First, the development of robust statistical models to integrate diverse data types and to produce summary statistics that reflect the level and direction of genetic alterations for genes and samples will be needed. This summary information then will be utilized to prioritize genes both for further studies and for analyses targeting clinical correlations, downstream pathways and interactive networks. Second, the establishment of a data framework that can be readily incorporated into any automated computational pipeline should be developed. This framework will be sufficiently flexible to allow the bidirectional flow of information. Under such a framework, clinicians and cancer biologists will be able to conveniently obtain relevant data from the system and to direct further analyses based on their knowledge and hypotheses. This latter element is particularly important because we expect that a key strategy for integrating gene information across data types will be both unbiased and hypothesis-driven and that no single score or statistic will capture all relevant information about a gene for all applications.

In addition to the identification of significant molecular changes in cancer, integration efforts will reveal the intrinsic interactions of such molecular changes, as well as their associations with specific clinical features and subtypes (Fig. 2). These data will contribute to the formulation of a reference platform for the development of prominent drug candidates as well as diagnostic and prognostic markers. We fully anticipate that large-scale functional validation of the identified candidates using high-throughput RNAi and cell culture-based screening, transgenic and xenograft animal models, and other high-throughput systems will then be positioned at the forefront of cancer research.

Conflict of Interest statement. None declared.

FUNDING

The authors wish to acknowledge funding from the National Institutes of Health, National Human Genome Research Institute U54 HG003079 (PI: Richard K. Wilson).

REFERENCES

1. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
2. Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K. and Ding, L. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.
3. Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E.A., Liu, Y., Weinstock, G.M., Wheeler, D.A., Gibbs, R.A. *et al.* (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res.*, **20**, 273–280.
4. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
5. Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.
6. Goya, R., Sun, M.G., Morin, R.D., Leung, G., Ha, G., Wiegand, K.C., Senz, J., Crisan, A., Marra, M.A., Hirst, M. *et al.* (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.
7. Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J. *et al.* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.
8. Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D. *et al.* (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.*, **361**, 1058–1066.
9. Ding, L., Ellis, M.J., Li, S., Larson, D.E., Chen, K., Wallis, J.W., Harris, C.C., McLellan, M.D., Fulton, R.S., Fulton, L.L. *et al.* (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, **464**, 999–1005.
10. Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.
11. Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B. *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.
12. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. and Ning, Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
13. Ye, K., Kosters, W.A. and Ijzerman, A.P. (2007) An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences. *Bioinformatics*, **23**, 687–693.
14. Bignell, G.R., Greenman, C.D., Davies, H., Butler, A.P., Edkins, S., Andrews, J.M., Buck, G., Chen, L., Beare, D., Latimer, C. *et al.* (2010) Signatures of mutation and selection in the cancer genome. *Nature*, **463**, 893–898.
15. Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.
16. Shih Ie, M., Sheu, J.J., Santillan, A., Nakayama, K., Yen, M.J., Bristow, R.E., Vang, R., Parmigiani, G., Kurman, R.J., Trope, C.G. *et al.* (2005) Amplification of a chromatin remodeling gene, Rsf-1/HBXAP, in ovarian carcinoma. *Proc. Natl Acad. Sci. USA*, **102**, 14004–14009.
17. Leary, R.J., Cummins, J., Wang, T.L. and Velculescu, V.E. (2007) Digital karyotyping. *Nat. Protoc.*, **2**, 1973–1986.
18. Morozova, O. and Marra, M.A. (2008) From cytogenetics to next-generation sequencing technologies: advances in the detection of genome rearrangements in tumors. *Biochem. Cell Biol.*, **86**, 81–91.
19. Salani, R., Chang, C.L., Cope, L. and Wang, T.L. (2006) Digital karyotyping: an update of its applications in cancer. *Mol. Diagn. Ther.*, **10**, 231–237.
20. Wang, T.L., Maierhofer, C., Speicher, M.R., Lengauer, C., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002) Digital karyotyping. *Proc. Natl Acad. Sci. USA*, **99**, 16156–16161.
21. Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S. *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**, R32.
22. McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527–1541.
23. Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.
24. Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.
25. Chiang, D.Y., Getz, G., Jaffe, D.B., O'Kelly, M.J., Zhao, X., Carter, S.L., Russ, C., Nusbaum, C., Meyerson, M. and Lander, E.S. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.
26. Zhang, Q., Ding, L., Larson, D.E., Koboldt, D.C., McLellan, M.D., Chen, K., Shi, X., Kraja, A., Mardis, E.R., Wilson, R.K. *et al.* (2010) CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics*, **26**, 464–469.
27. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
28. Raphael, B.J., Volik, S., Collins, C. and Pevzner, P.A. (2003) Reconstructing tumor genome architectures. *Bioinformatics*, **19**(Suppl. 2), ii162–ii171.
29. Raphael, B.J., Volik, S., Yu, P., Wu, C., Huang, G., Linardopoulou, E.V., Trask, B.J., Waldman, F., Costello, J., Pienta, K.J. *et al.* (2008) A sequence-based survey of the complex structural organization of tumor genomes. *Genome Biol.*, **9**, R59.
30. Volik, S., Raphael, B.J., Huang, G., Stratton, M.R., Bignel, G., Murnane, J., Brebner, J.H., Bajsarowicz, K., Paris, P.L., Tao, Q. *et al.* (2006) Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res.*, **16**, 394–404.
31. Volik, S., Zhao, S., Chin, K., Brebner, J.H., Herndon, D.R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W.L. *et al.* (2003) End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc. Natl Acad. Sci. USA*, **100**, 7696–7701.
32. Bignell, G.R., Santarius, T., Pole, J.C., Butler, A.P., Perry, J., Pleasance, E., Greenman, C., Menzies, A., Taylor, S., Edkins, S. *et al.* (2007) Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.*, **17**, 1296–1303.
33. Clark, M.J., Homer, N., O'Connor, B.D., Chen, Z., Eskin, A., Lee, H., Merriman, B. and Nelson, S.F. (2010) U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet.*, **6**, e1000832.
34. Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordonez, G.R., Bignell, G.R. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.
35. Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C. *et al.* (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, **463**, 184–190.
36. Lee, W., Jiang, Z., Liu, J., Haverty, P.M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K.P., Bhatt, D. *et al.* (2010) The mutation spectrum

revealed by paired genome sequences from a lung cancer patient. *Nature*, **465**, 473–477.
37. Sindi, S., Helman, E., Bashir, A. and Raphael, B.J. (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**, i222–i230.
38. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.
39. Smith, D.R., Quinlan, A.R., Peckham, H.E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W.F., Tusneem, N., Stromberg, M.P. *et al.* (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.*, **18**, 1638–1642.
40. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.
41. Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N. *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.
42. The Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
43. Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.M., Gallia, G.L. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.
44. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.
45. Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M. *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**, 66–72.
46. Clark, M.J., Homer, N., O'Connor, B.D., Chen, Z., Eskin, A., Lee, H., Merriman, B. and Nelson, S.F. (2010) U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet.*, **6**, e1000832.
47. Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., Bignell, G., Teague, J., Smith, R., Stevens, C. *et al.* (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.*, **37**, 590–592.
48. Rubin, A.F. and Green, P. (2009) Mutation patterns in cancer genomes. *Proc. Natl Acad. Sci. USA*, **106**, 21766–21770.
49. Jones, S., Zhang, X., Parsons, D.W., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.
50. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic. Acids Res.*, **36**, D480–D484.
51. Lin, J., Gan, C.M., Zhang, X., Jones, S., Sjoblom, T., Wood, L.D., Parsons, D.W., Papadopoulos, N., Kinzler, K.W., Vogelstein, B. *et al.* (2007) A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res.*, **17**, 1304–1318.
52. Cerami, E., Demir, E., Schultz, N., Taylor, B.S. and Sander, C. (2010) Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE*, **5**, e8918.
53. Girvan, M. and Newman, M.E. (2002) Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, **99**, 7821–7826.
54. Vandin, F., Upfal, F. and Raphael, B.J. (2010) Algorithms for detecting significantly mutated pathways in cancer. *Lect. Notes Comput. Sci.*, **6044**, 506–521.