Computational characterisation of cancer molecular profiles derived using next generation sequencing

Urszula Oleksiewicz1,2*, Katarzyna Tomczak1,2,*, Jakub Woropaj4*, Monika Markowska1, Piotr Stępniak5, Parantu K Shah6

1Laboratory of Gene Therapy, Department of Cancer Immunology, The Greater Poland Cancer Centre, Poznan, Poland
2Department of Cancer Immunology and Diagnostics, Chair of Medical Biotechnology, Poznan University of Medical Sciences, Poznan, Poland
3Postgraduate School of Molecular Medicine, Medical University of Warsaw, Warsaw
4Poznan University of Economics, Poznań, Poland
5Transition Technologies S.A., Warsaw, Poland
6Institute for Applied Cancer Science, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

*These authors contributed equally to this paper.

Molecular profiling of cancer genomes

Over the years, individual laboratories and large-scale projects such as TCGA and ICGC have discovered that cancer is a heterogeneous disease with lots of variability within a single tumour type or even within a single tumour [1–4]. Nonetheless, much of our current understanding of cancer genetics is grounded on the principle that cancer arises from a clone that has accumulated the requisite somatically acquired genetic aberrations leading to the malignant transformation [5]. Characterising individual tumours or cohorts at the molecular level has helped in identifying common and type specific cancer vulnerabilities as well as recording the individual history of tumours [4, 6–8]. This has enabled the creation of drugs that target these molecular vulnerabilities and provide tailored treatments for patients, improving therapy efficacy and minimising its side effects [9, 10]. For example, Imatinib specifically targets the BCR-Abl fusion tyrosine kinase that exists only in the cells of chronic myelogenous leukaemia and other tumours but not in healthy cells [11]. Similarly, Herceptin is a monoclonal antibody that is used to target HER2 positive breast tumours [12].

At present, the TCGA and other large-scale projects characterise tumours with microarray and next generation sequencing (NGS) platforms to obtain a different type of genetic information at the whole genome level [6–8, 13–15]. The microarray platform had been, and is currently being, used to identify gene and microRNA expression, alternative splicing, copy number alterations, DNA methylation, and identification of protein-DNA and protein-RNA interactions [16]. Next generation sequencing platforms are now replacing the microarray platforms for obtaining these data. Moreover, sequence reads from whole exome sequencing, along with DNA and RNA sequencing, also allow detection of mutations and gene fusions for coding and non-coding regions of the genome [17]. While conceptually similar in experiment design, the sequence read information generated using NGS platforms has very different statistical properties to intensity-based information acquired from microarray platforms [18].

Multiple articles have reviewed protocols to generate microarray profiles and their statistical analysis to extract meaningful information [19–22]. While relatively new, a vast amount literature describing statistical methodologies to analyse NGS data already exists [23–28]. So, in this article we...
focus only on the methodologies to extract meaningful information from NGS reads. Moreover, we will discuss only those data types that are generated and analysed by TCGA. Furthermore, for each data type, we only describe the main steps to obtain this information and point the readers to an exhaustive list of methodologies/software and articles for deeper insight. Also, we do not provide any comments on comparison of these methods but rather point to articles comparing different methodologies and identifying their strengths and errors [29-32].

**Pre-processing and Quality Control of NGS data**

To date, several next-generation sequencing platforms are available, including the Illumina Genome Analyser, which is being used extensively TCGA by consortium for tumour profiling. Each platform has its own method for generating sequencing reads from samples. But in every case, the sequence reads obtained using these platforms are short – typically from 36 to several hundred nucleotides. Furthermore the sequencing run can be single-end or paired-end, meaning the reads are sequenced in one or two directions (from 3’ and 5’ ends). The first tasks in any NGS computational pipeline are: performing primary data acquisition, determining base calls and confidence scores from the fluorescent signals of the sequencer, and converting them to FASTQ files containing the raw sequence reads and per base quality scores. When multiple samples are pooled in one lane using sample-specific index/barcode adapters, the FASTQ should be demultiplexed and reorganised based on index information, and the adapters ought to be trimmed [33].

Quality control is a very important part of the data preparation (Table 1). There are several kinds of sequencing artefacts that could have a serious negative impact on downstream analyses. The artefacts commonly exist in raw reads, regardless of the sequencing platform. Firstly, sequences may be contaminated with adapters on their 5’- or 3’- ends that were added as part of the sequencing protocol. Secondly, base quality and sequence complexity vary both within and between reads. The qualities of bases on most sequencing platforms will degrade as the run progresses, so it is common to see the quality of base calls falling towards the end of the read. It is desirable to remove or trim such sequences with appropriate thresholds. Additionally, NGS reads can be highly redundant with the same sequence being represented in large numbers, so it is important to reduce these PCR amplification artefacts. The contamination in the sequencing dataset can also be caused by laboratory factors such as sample preparation, library construction, and other steps of the experiment. Moreover, samples may contain DNA/RNA from other sources including viruses, which are hard to avoid during the sample preparation process. Finally, general statistical methods like sample clustering and principal component analysis (PCA), and outlier detection can be used for assessment of overall quality and sample comparison according to experiment design.

### Aligning short reads to the reference genome

Accurate alignment of short sequence reads generated using NGS platforms to a repeat masked reference genome is the first step in obtaining biological information from NGS data. Since, the numbers of reads generated in any given NGS experiment are very large (typically in millions), many efficient algorithms have been developed to deal with the alignment process. It is important to note that different read mapping procedures are necessary depending on the needs of downstream analysis, and alignment accuracy has a high impact on the interpretation of the data. We comment on that in sections to follow. Most applications aim to identify uniquely mapped reads - matching to a single “best” genomic position. The non-uniquely

| Method name | Year published | PMID | Data type | Platform | Statistical method | Input requirements |
|-------------|----------------|------|-----------|----------|--------------------|--------------------|
| BIGpre      | 2011           | 22289480 | Illumina, 454 | Perl     | Correlation between forward and reverse reads, trimming low-quality reads | FASTQ |
| FastQC      | 2010           | –     | any       | Java     | Sequence length, quality, k-mers presence reports | FASTQ, SAM/BAM |
| HTQC        | 2013           | 23363224 | Illumina | C++      | Tail trimming, filter by quality/length/tile | FASTQ |
| QC-Chain    | 2013           | 23565205 | any       | C++      | Quality assessment, trimming, filtering unknown contamination | FASTQ |
| Qualimap    | 2012           | 22914218 | any       | Java, R  | Alignment biases detection, sample comparison | SAM/BAM |
| PRINSEQ     | 2011           | 21278185 | any       | Perl     | Sequence complexity, duplicates, occurrence of Ns and poly-A/T tails, tag sequences reports | FASTQ, Fasta |
| PIQA        | 2009           | 19602525 | Illumina | R        | Assess the clusters density per tile, base-calls proportions per tile/cycle | FASTQ |
| FastUniq    | 2012           | 23284954 | any       | C++      | De novo PCR duplicates removal for paired short reads | FASTQ |
mapped reads are filtered using an upper boundary for the number of reported mappings.

Most short read alignment algorithms use auxiliary data structures (also called indexes) for the reads or reference sequence. The main indexing methods are based on hash tables, prefix/suffix trees, or merge sorting methods (Table 2). Such representation of the entire human genome takes only a few GB of memory and enables exact matches to be found in a short time. Burrow-Wheeler transform and FM-index-based algorithms give better results for reads from repeated regions, but there is no efficient general method for handling errors in the reads for this category. Some hybrid solutions have been proposed, e.g. Stampy (see Table 2). These enhancements result in higher sensi-

| Method name         | Year published | PMID        | Data type | Platform            | Statistical method                                                                 | Input requirements |
|---------------------|----------------|-------------|-----------|---------------------|-------------------------------------------------------------------------------------|--------------------|
| BFAST               | 2009           | 19907642    | RNA       | C                   | Based on creating flexible, efficient whole genome indexes to rapidly map reads to candidate alignment locations, with arbitrary multiple independent indexes allowed to achieve robustness against read errors and sequence variants. Final local alignment uses a Smith-Waterman method, with gaps to support the detection of small INDELs | FASTQ, FASTA       |
| Bowtie              | 2009           | 19261174    | RNA       | C++ (SeqAn library) | Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches | FASTQ, FASTA       |
| BWA                 | 2009           | 19451168    | RNA       | C                   | Backward search with Burrows-Wheeler Transform (BWT), allowing mismatches and gaps. | FASTQ              |
| BWA-PSSM            | 2014           | 24717095    | RNA       | C                   | Probabilistic adaptable alignment based on the use of position specific scoring matrices (PSSM) and BWT | FASTQ              |
| CUSHAW2             | 2013           | 22576173    | RNA       | C++                 | Uses Burrows-Wheeler transform (BWT), the Ferragina-Manzini index and CUDA parallel programming model for GPUs. Supports only ungapped alignment | FASTQ              |
| DistMap             | 2013           | 24009693    | RNA       | Perl, Java          | Wrapper for many aligners, based on MapReduce API for parallel processing. Currently not handling spliced alignments | FASTQ              |
| MAQ                 | 2008           | 18714091    | RNA       | C++, Perl           | Based on Smith-Waterman gapped alignment and Bayesian statistical model that incorporates the mapping qualities and error probabilities | FASTQ              |
| MOSAIK              | 2014           | 24599324    | RNA       | C++                 | Uses hash clustering strategy coupled with the Smith-Waterman algorithm. Detects mismatches, short insertions and deletions | FASTQ              |
| PASS                | 2009           | 19218350    | RNA       | C++                 | Based on precomputed score tables (PST) calculated with the Needleman and Wunsch algorithm | FASTQ              |
| RMAP                | 2009           | 19736251    | RNA       | C++                 | Uses multiple filtration (Pevzner and Waterman) and approximate pattern matching. Incorporates the use of quality scores directly into the mapping process | FASTQ, FASTA       |
| SOAPaligner/ SOAP2 | 2009           | 19497933    | RNA       | C                   | Based on Burrows Wheeler Transformation (BWT) compression index | FASTQ              |
| Stampy              | 2011           | 20980556    | RNA       | Python              | Hybrid probabilistic model for mapping quality (measured by Phred score) | FASTQ              |
| ZOOM                | 2008           | 18684737    | RNA       | C                   | Custom filtering model | FASTQ              |
Computational characterisation of cancer molecular profiles derived using next generation sequencing

A deficiency of the RPKM/FPKM approach is that the proportional representation of each gene is dependent on the expression levels of all other genes. Often a small fraction of genes account for large proportions of the sequenced reads, and small expression changes in these highly expressed genes will skew the counts of lowly expressed genes under this scheme. This can result in deduction of erroneous differential expression. Therefore, methods for calculating differential expression require counts to begin with. Thus RNA-Seq non-negative counts follow discrete distribution as opposed to the intensities recorded from microarrays, which are treated as continuous measurements and commonly assumed to follow a log-normal distribution. For RNA-Seq data Poisson distribution and Negative Binomial (NB) distribution are the two most commonly used models [38–42] (Table 3). Other distributions, such as beta-binomial [43], have also been proposed.

The Poisson distribution has the advantage of simplicity and has only one parameter, but it constrains the variance of the modelled variable to be equal to the mean. The Negative Binomial distribution has two parameters, encoding the mean and the dispersion, and hence allows modelling of more general mean-variance relationships. For RNA-Seq, it has been suggested that the Poisson distribution is well suited for analysis of technical replicates, whereas the higher variability between biological replicates necessitates a distribution incorporating overdispersion, such as Negative Binomial [28, 44, 45]. Analogous to microarray data analysis, it is clear that borrowing the variance from other genes help to better estimate the variation in read counts for a gene and condition. This overcomes a common problem with an underestimation of variance when based on a low number of observations. The most commonly used parametric methods include EdgeR, DESeq, and baySeq and use negative binomial distribution. Other methods such as Cuffdiff2 uses a beta-negative binomial distribution, which is a combination of beta and negative binomial distribution. Non-parametric methods like SAM-Seq also work relatively well on the count data.

Identification of alternative splicing from transcriptomic reads

A widely recognised source of proteome diversity in eukaryotic species is expression of multiple distinct mRNA transcripts from a single gene locus by alternative transcript initiation, alternative splicing [47] (Table 4), and alternative polyadenylation [48]. If RNA-Seq reads span exon junctions, parts of reads will map to two different exons. This allows inference of alternative splicing. However, such a read structure will pose problems to standard aligners that map reads contiguously to the reference. Splice sites can be detected initially by identifying reads that span exon junctions. Split-read aligners such as TopHat, methods that identify minimal match on either side of exon junction, and genomic short-read nucleotide alignment are used to identify alternative splicing. Most methods utilise a database of expression and alternative expression sequence ‘features’. These and other strategies that perform de-novo assemblies present a number of computational...
Table 3. Software for RNA-Seq data analysis

| Method name | Year published | PMID | Data type | Platform | Statistical method | Input requirements |
|-------------|----------------|------|-----------|----------|-------------------|--------------------|
| Aldex       | 2013           | 23843979 | RNA-seq   | R        | ANOVA, Dirichlet distribution | Raw counts |
| ALDEx2      | 2014           | 24910773 | RNA-seq   | R        | Dirichlet distribution, glm  | Raw counts |
| ASC         | 2010           | 21080965 | RNA-seq   | R        | Empirical Bayes, Dirichlet distribution | Raw counts |
| baySeq      | 2010           | 20698981 | RNA-seq   | R        | Empirical Bayes | Raw counts |
| BBSeq       | 2011           | 21810900 | RNA-seq   | R        | Beta-Binomial, linear model | Raw counts |
| CEDER       | 2012           | 22641709 | RNA-seq   | R        | Negative binomial | Raw counts |
| campcodeR   | 2014           | 24813215 | RNA-seq   | R        | Empirical Bayes | Raw counts |
| COV2HTML    | 2014           | 24512253 | RNA-seq   | Web site | Long-read sequence w/ annotation, short-read sequence tag // fasta, fastq |
| CPTRA       | 2009           | 19811681 | RNA-seq   | Python   | ??? |
| CQN         | 2012           | 22285995 | RNA-seq   | R        | Conditional quantile normalisation, robust generalised regression | Raw counts |
| Cuffdiff    | 2013           | 23222703 | RNA-seq   | Standalone | Beta negative binomial distribution |
| DegPack     | 2013           | 24981075 | RNA-seq   | Web site | Non-parametric (ranks) | Raw counts |
| DEGseq      | 2010           | 19855105 | RNA-seq   | R        | MA-plot-based (?) | Raw counts |
| DER Finder  | 2014           | 24398039 | RNA-seq   | R        | Hidden Markov Model | Raw counts |
| DESeq       | 2010           | 20979621 | RNA-seq   | R        | Negative binomial distribution | Raw counts |
| DEXUS       | 2013           | 24049071 | RNA-seq   | R        | Expectation-maximisation algorithm, Bayes | Raw counts |
| EBSeq       | 2013           | 23428641 | RNA-seq   | R        | Empirical Bayes | Raw counts |
| EDASEq      | 2011           | 22177264 | RNA-seq   | R        | Empirical Bayes | Raw counts |
| edgeR       | 2012           | 22287627 | RNA-seq   | R        | Empirical Bayes, glm | Raw counts |
| edgeR-robust| 2014           | 24753412 | RNA-seq   | R        | Weights, empirical Bayes | Raw counts |
| GExposer    | 2012-2013      |        | Machine learning algorithm |
| iFad        | 2012           | 22581178 | RNA-seq   | R        | Bayesian sparse factor model | Raw counts |
| MRFSQ (uses DESeq) | 2013 | 23793751 | RNA-seq   | Standalone | Markov random field model | Raw counts, co-expression database |
| Myrna       | 2010           | 20701754 | RNA-seq | Cloudcomputing, Bowtie, R |
| NOISeq      | 2011           | 21903743 | RNA-seq   | R        | Non-parametric | Raw counts |
| NPEBseq     | 2013           | 23981227 | RNA-seq   | R        | Non-parametric Bayesian | Raw counts |
| pairedBayes | 2012           | 24049071 | RNA-seq   | R        | Empirical Bayes | Raw counts |
| PoissonSeq  | 2012           | 22287627 | RNA-seq   | R        | Poisson goodness-of-fit | Raw counts |
| QuasiSeq    | 2012           | 23104842 | RNA-seq   | R        | Quasi-Poisson, quasi-negative binomial | Raw counts |
| RNASeqGUI (uses edgeR, DESeq, NOIseq, BaySeq) | 2014 | 24812338 | RNA-seq | R | Empirical Bayes, negative binomial | Raw counts |
| SAMSeq      | 2013           | 22127579 | RNA-seq   | Standalone | Non-parametric | Raw counts |
challenges because of computation time and the depth of sequencing, which results in few junction-spanning reads. This quantitation of alternatively spliced transcripts needs to be followed by identification of differential expression of these transcripts between samples.

Apart from detection and quantitation of expression and alternative splicing, RNAseq has the capacity to identify RNA editing events [49–51], allele-specific expression (ASE) [52, 53], quantify noncoding RNAs [54, 55], and detect exogenous RNA [56, 57], single-nucleotide polymorphisms (SNPs), somatic mutations [8, 58], and structural variations.

**Detection of somatic copy number alterations and structural variants**

Genomic alterations accumulate in tumours during cancer development [59]. In addition to point-mutations, inversions, and translocations, somatic copy-number al-

### Table 3. Alternative splicing algorithm

| Method name  | Year published | PMID    | Data type | Platform | Statistical method                                      | Input requirements                  |
|--------------|----------------|---------|-----------|----------|--------------------------------------------------------|-------------------------------------|
| ABMapper     | 2011           | 21169577| RNA       | C++, Perl| Fast suffix-array algorithm and a dual-seed strategy for spliced alignment | FASTA, FASTQ                        |
| ERANGE       | 2008           | 18516045| RNA       | Python   | Splice junctions identification relies on reference genome exon positions | FASTQ                               |
| GEM Mapper   | 2012           | 23103880| RNA       | C, Objective Caml | Based on Burrows-Wheeler Transform and custom mapping algorithms. Uses custom mappability concept | FASTQ                               |
| MapSplice    | 2010           | 20802226| RNA       | C++, Python| Algorithm not dependent on splice site features or intron length; consequently, it can detect novel canonical as well as non-canonical splices. This method has tag alignment phase and splice inference phase | FASTQ                               |
| PALMapper    | 2010           | 21154708| RNA       | Web/ Galaxy | Combines GenomeMapper (based on BWT and k-mer indexes) read mapper with the spliced aligner QPALMA | FASTQ                               |
| QPALMA       | 2008           | 18689821| RNA       | C++, Python| SVM-based splice site predictor with the so-called ‘weighted degree’ kernel. Alignment based on extended BWT | FASTQ                               |
| SpliceMap    | 2010           | 20371516| RNA       | C++, Python| Canonical GT-AG splice sites identification using half-read mapping | FASTQ                               |
| SplitSeek    | 2010           | 20236510| RNA       |          | Candidate junction reads generation in intermediate BEDPE format feasible for paired-end sequences | FASTQ                               |
| Subread      | 2013           | 23558742| RNA       | R        | Seed-and-vote - new multi-seed alignment strategy for overlapping seeds from each read (subreads) | FASTQ                               |
| TopHat       | 2009           | 19289445| RNA       | C++      | Canonical GT-AG splice sites identification            | FASTQ                               |
tions (SCNAs) are ubiquitous in cancer [60, 61], and several recurrent SCNAs are associated with particular cancer types [62], tumour aggressiveness [63], and patient prognosis [64]. Reliable detection of SCNAs can lead to the identification of cancer driver genes [65] and the development of new therapeutic approaches [66–68]. Deep sequencing [69] and exome-based sequencing efforts are now replacing microarray-based array CGH (aCGH) [70] and single nucleotide polymorphism arrays (SNP arrays) [23, 71]. Moreover, similarly to microarray-based studies, the inference about cancer related copy number alterations requires comparison of paired samples – normal and tumour.

The general workflow to detect SCNAs from sequencing data consists of three main steps: (i) raw copy-number inference in the local genome region by either calculating read counts or depth of coverage ratio between tumour and control samples, (ii) the raw copy-number profiles segmentation to find change-points in the raw copy-number signal and divide the chromosomes accordingly into segments with similar copy-numbers, and (iii) classification of different segments into gains or losses. The first step is essentially based on understanding variations in depth of coverage (DOC) of aligned sequence reads against the reference genome [69, 72, 73]. Deviation from the background in DOC may signify the presence of a copy number variation (CNV). The last two steps for obtaining copy number alteration are not specific to the sequencing data.

Multiple methods exist for identification of structural variations using whole genome sequencing. Methods include identification of atypical alignment patterns of sequence reads against the reference genome, which reflect gaps in the sequence alignment [74]. In paired-end read mapping, the sequenced ends of a short DNA fragment are aligned against the reference genome. The mean insert size of the fragment is compared with the reference genome distance between aligned fragment ends to deduce the presence of deletions or insertions [74–76]. This detection requires high alignment accuracy and underscores its importance.

Although coding regions comprise only ~1% of the genome, they are enriched for causal variation, making exome-based studies valuable, manageable, and cost-effective. Whole exome sequencing (WES) data have been used effectively for the identification of small INDELs, usually of a size < 50 bp, within exon targets that are typically sized between 200 and 300 bp. The approaches discussed above, while appropriate for (deeply sequenced) DNA-sequence data, are less effective for exome sequencing and detecting CNV, as the CNV's breakpoints are likely to lie outside the targeted exons [77]. Detecting structural and copy number variations from RNA-Seq data presents similar challenges.

Identification of cancer driver mutations and their functional impact

Cancer is abundantly composed of somatic mutations accumulating in the genome over an individual’s lifetime, only a fraction of which drive cancer progression. Mutations can be identified from DNA-seq, RNA-seq, and Exome sequencing data [78–80] (Table 5). The most basic way of detecting somatic mutations from NGS reads is to identify mismatch/gaps in the alignment of the read with the reference. However, large datasets possess sequencing errors: random mutations that occur during cell division and single nucleotide polymorphisms that differ from reference assembly. This makes identification of cancer driver mutations a challenging issue [81]. Moreover, intra-tumour heterogeneity also hinders the identification of all types of somatic mutations [82]. Several methods for detecting somatic mutations are currently in use, such as MuTect [83], Strelka [84], and VarScan 2 [85] for SNV detection or BIC-Seq [86], APOLO [87], CoNIEFER [88], BreakDancer [89], and Meerkat [87] for CNA or SV detection. Most methods for somatic mutation detection take into account only part of the possible source of errors; therefore, running different methods simultaneously is advisable.

The most basic task for mutation analysis in cancer is the distinction between driver and passenger mutations. To help filter a subset of driver mutations from the long list of detected somatic and passenger mutations, three major computational predictive approaches utilising different statistical tests can be applied [90–92]:

1. **Identification of recurrent somatic mutations** is based on the idea of clonal evolution of tumour cell populations. To predict genes with recurrent single-mutations in a cohort of cancer patients, several statistical methodologies including MutSigCV [3], MuSiC [93], and DrGap [94] are available. These methods are based on the determination of the probability of the observed number of mutations in a gene to the expected background mutation rate, the BMR (probability of observed passenger mutation) across a cohort of patients. As opposed to mutations, there is no accurate model established to identify genes with recurrent copy number aberrations (CNAs); therefore, methods are based on a non-parametric approach, e.g. GISTIC2 [95], CMDIS [96], and ADMIRE [97].

2. **Prediction of the functional impact of individual mutations** is based on the utilisation of additional information about protein sequence and/or structure and evolutionary conservation of the protein encoded by the mutated gene. Methods like SIFT [98], Polyphen-2 [99], and MutationAssesor [100] predict the functional impact (deleteriousness) of missense mutations. CHASM utilises random forest classification to identify driver and passenger somatic missense mutations, based on a training set of labelled positive (driver) and negative (passenger) examples [101]. Furthermore, clusters of non-synonymous mutations across patients, typically to detect ‘activating’ mutations, NMC [102] and the Invex [103] method can be applied. Moreover, the iPAC method is able to search for clusters of mutations, but in the context of crystal structures of proteins [104].

3. **Identification of recurrent combinations of mutations** is based on assessment of combinations of mutations enriched in known pathways (e.g. GSEA [105], PathScan [106], Patient-oriented gene sets [107]), interaction networks (NetBox [108], HotNet [109], MEMo [110]), or de novo defined sets (Endrix [109], Mut-Dendrix [111] or RME [112]), enabling the discovery of novel combinations of mutated genes in cancer.
### Table 5. Methods for finding mutations

| Method name          | Year published | PMID       | Data type | Platform   | Statistical method                                      | Input requirements                             |
|----------------------|----------------|------------|-----------|------------|--------------------------------------------------------|-----------------------------------------------|
| ActiveDriver         | 2013           | 23340843   | DNA-seq   | R          | Generalized linear regression                          | FASTA, TAB                                    |
| CancerMutationAnalysis | 2014      | 24233780   | DNA-seq   | R          | Empirical Bayes                                        | Non-standard tables                           |
| CanDrA               | 2013           | 24205039   | DNA-seq   | Perl       | U-Mann Whitney, AUC (area under curve)                 | Non-standard tables                           |
| CanPredict           | 2007           | 17537827   | DNA-seq   | Web site   | SIFT, Pfam-based logRE-value metric, GOSS              | FASTA                                         |
| CAROL                | 2012           | 22261837   | DNA-seq   | R          | SIFT, PolyPhen-2                                       | Tab-delimited, FASTA                          |
| CHASM/SNV-Box        | 2009           | 19654296   | DNA-seq   | Standalone | CHASM, SNV-Box                                         | dnSNP r#, Pubmed ID, VCF, bed                  |
| CRAVAT               | 2013           | 23325621   | DNA-seq   | Web site   | CHASM, SnvGet                                          | Non-standard tables                           |
| DDIG-in              | 2013           | 23497682   | DNA-seq   | Web site   | Support vector machine-based method                    | Non-standard tables                           |
| DMI                  | 2012           | 23044540   | DNA-seq   | Standalone | Machine learning, discrimination index                 | Text file                                     |
| DrGaP                | 2013           | 23954162   | DNA-seq   | Standalone | Chi-square distribution                                | Non-standard tables                           |
| e-Driver             | 2014           | 25064568   | DNA-seq   | Perl       | Binomial distribution                                  | Non-standard tables                           |
| eXtasy               | 2013           | 24076761   | DNA-seq   | Web site   | Variant impact prediction, haploinsufficiency prediction, phenotype-specific gene prioritisation | VCF                                           |
| FATHMM               | 2013           | 23620363   | DNA-seq   | Web site   | Hidden Markov Model                                    | Annotated VCF                                 |
| InVEx                | 2012           | 22817889   | DNA-seq   | Python     | Permutation-based                                      | Non-standard tables, power FASTA              |
| MuSIC                | 2012           | 22759861   | DNA-seq   | Standalone | Fisher p-value, likelihood ratio test, convolution test (summarised log statistic of joint binomial point probability) | BAM, SNV, MAF                                 |
| MutSig               | 2013           | 23770567   | DNA-seq   | Standalone | MutSigCV (Background mutation rate)                    |                                              |
| nsSNPAnalyzer        | 2005           | 15980516   | DNA-seq   | Web site   | Machine learning (random forest)                       | FASTA, SNP                                    |
| Oncodrive-fm         | 2012           | 22904074   | DNA-seq   | Standalone | SIFT, PolyPhen2, MutationAssessor                      | TDM, TSV                                      |
| OncodriveCLUST       | 2013           | 23884480   | DNA-seq   | Python     | Clustering                                             | Non-standard tables                           |
| PANTHER              | 2013           | 23193289   | DNA-seq   | Web site   | subPSEC                                                | FASTA                                         |
| PhD-SNP              | 2006           | 16895930   | DNA-seq   | Web site   | Sequence and Profile-Based                             | FASTA                                         |
| PROVEAN              | 2012           | 23056405   | DNA-seq   | Standalone | Alignment-Based                                        | Non-standard tables                           |
| transFIC             | 2012           | 23181723   | DNA-seq   | Web site   | SIFT, PolyPhen2, MutationAssessor                      | File w/ chromosome/protein coordinates (hg19) |
Identification of gene fusions

Gene fusions appear as a result of chromosomal rearrangements, such as deletion, insertion, inversion, or translocation. A fused gene is expressed as a hybrid entity encoding sequences of two distinct genes. Tumorigenic fusions successfully evade the gene regulation that its constituents are subjected to. Multiple cancer-related gene fusions have been identified, including prototypic BCR-ABL [113], EML4-ALK [114], TMPRSS2-ERG [115], KIF5B-RET [116], and others [7, 13, 14, 117]. Such alterations may serve as a good cancer biomarker or therapeutic target [120, 121]. Whole genome [75, 119] and transcriptome sequencing [120, 121] profiles can be used to identify fusions.

Transcriptome sequencing is proven to be superior over WGS and therefore is more commonly used. This is due to the fact that RNA-seq covers only transcribed sequences, which constitute a small percentage of the genome, thus reducing the cost, time, and resources needed for full analysis. Furthermore, RNA-seq provides information on the transcriptionally active fusion genes and their splicing variants. However, it also harbours certain limitations, including lack of information regarding non-transcribed regions and dependence on the heterogeneity in the expression levels between various cell types [122]. Over the last few years several software packages (Table 6) have been developed for the detection of gene fusions and/or structural variants (SV) that cause gene fusions. The majority of the software utilises RNA-seq data as an input. However, other tools use WGS data or both to increase the likelihood of detection of true fusion.

The most common analysis steps for identifying gene fusions are as follows: (i) alignment and filtering, (ii) detection of fusion junctions in candidate genes, and (iii) fragment assembly and selection of putative fusions [122]. Apart from mapping to the current reference genome available, RNA-seq reads are additionally mapped to annotated transcriptome libraries (e.g. RefSeq). The most commonly used mapping tool is Bowtie, due to its speed and high efficiency. The reads that map appropriately to the reference genome are filtered out from further analysis. The unmapped or discordantly mapped reads are fusion candidates that might be further passed through additional filters (e.g. ribosomal fil-

Table 6. Identifying gene fusions

| Method name        | Year published | PMID             | Data type          | Platform       | Statistical method                                                                 | Input requirements       |
|--------------------|----------------|------------------|--------------------|----------------|------------------------------------------------------------------------------------|--------------------------|
| BreakFusion        | 2012           | 22563071         | RNA-seq            | C++, Perl      | A computational pipeline for identifying gene fusions from RNA-seq data            | BAM                      |
| BreakTrans         | 2013           | 23972288         | RNA-seq            | Perl           | Uncovering the genomic architecture of gene fusions                               | Tab-delimited text files |
| chimerascan        | 2011           | 21840877         | RNA-seq            | Python         | Identifying chimeric transcription in sequencing data                               | FASTQ                    |
| conrad             | 2011           | 21478487         | RNA-seq, DNA-seq   | C++            | Discovery of gene fusions using paired end RNA-Seq and WCSS.                       | FASTQ                    |
| deFuse             | 2011           | 21625565         | RNA-seq            | C++            | Detecting gene fusions from paired-end RNA-Seq                                    | FASTQ                    |
| FusionAnalyser     | 2012           | 22570408         | RNA-seq            | C#             | Detecting gene fusions from paired-end RNA-Seq data                               | SAM/BAM                  |
| FusionHunter       | 2011           | 21546395         | RNA-seq            | Perl           | Detecting gene fusions from paired-end RNA-Seq data                               | FASTQ                    |
| FusionMap          | 2011           | 21593131         | RNA-seq, DNA-seq   | C#             | Detecting gene fusions from single- and paired-end RNA-Seq and DNA-seq data       | FASTQ/BAM with unmapped reads |
| FusionSeq          | 2010           | 20964841         | RNA-seq            | C              | A modular framework for finding gene fusions by analysing Paired-End RNA-Sequencing data | MRE, SAM |
| ShortFuse          | 2011           | 21330288         | RNA-seq            | C++, Python    | Detecting gene fusions from paired-end RNA-Seq data                               | FASTQ                    |
| SnowShoes-FTD      | 2011           | 21622959         | RNA-seq            | Perl           | Detecting gene fusions from paired-end RNA-Seq data                               | FASTQ                    |
| SOAPfusion         | 2013           | 24123671         | RNA-seq            | Perl           | Detecting gene fusions from paired-end RNA-Seq data                               | FASTQ                    |
| TopHat-Fusion      | 2011           | 21835007         | RNA-seq            | C++            | Detecting gene fusions from single- and paired-end RNA-Seq data                   | FASTQ                    |
Table 7. Available software for purity estimation

| Method name         | Year published | PMID          | Data type (technique) | Platform          | Statistical method                                      | Input requirements                                                                 |
|---------------------|----------------|---------------|-----------------------|-------------------|--------------------------------------------------------|--------------------------------------------------------------------------------------|
| Dsection            | 2010           | 20631160      | RNA (Microarray)      | Web-based and MATLAB | Bayesian model                                         | Expression and proportion data required                                               |
| csSAM               | 2010           | 20208531      | RNA (Microarray)      | -                 | Linear regression-based model                          | Expression profile of mixed tissue samples                                            |
| mixture_estimation.R| 2010           | 20202973      | RNA (Microarray)      | R based           | Variation of electronic subtraction method             | Expression profile of mixed tissue samples                                            |
| ASCAT               | 2010           | 20837533      | DNA (Microarray)      | R based           | Analytical optimisation method                         | SNP array data with Log R and B-allele frequency information                         |
| PERT                | 2012           | 23284283      | RNA (Microarray)      | Octave            | Perturbation model                                     | Expression data from mixed cell type and expression profile of each homogeneous cell type |
| ABSOLUTE            | 2012           | 22544022      | DNA (Microarray and HTS) | R based | Gaussian mixture model                                 | Copy number data in segmentation file                                                |
| JointSNVMix1, JointSNVMix2 | 2012 | 22285562 | DNA (HTS) | Python | Probabilistic graphical model | Sequence data from tumour/normal pairs                                                                 |
| CNAnorm             | 2012           | 22039209      | DNA (HTS)             | R based           | Analytical optimization method                         | Sequencing data of tumour and normal samples in bam format                           |
| DeconRNASeq         | 2013           | 23428642      | RNA (Microarray and HTS) | R based | Globally optimised nonnegative decomposition algorithm | Expression data from multiple tissue, signature of individual tissue and proportion data required |
| TEMT                | 2013           | 23735186      | RNA (HTS)             | Python            | Probabilistic model including position and sequence-specific biases | Required RNA-seq sequencing data from pure tissue and mixed tissue                   |
| ESTIMATE            | 2013           | 24113773      | RNA (HTS)             | R based           | Gene signature (ssGSEA) based model                   | Expression data in Gene Set Enrichment Analysis (GSEA) gct format                    |
| THetA               | 2013           | 23895164      | DNA (HTS)             | Python            | Explicit probabilistic model                           | Copy number data in interval count file format                                        |
| ExPANdS             | 2013           | 24177718      | DNA (HTS)             | R based and MATLAB | Probability distributions model                       | Somatic mutations and copy number data required                                        |
| Virmid              | 2013           | 23987214      | DNA (HTS)             | Java based        | Probabilistic model and maximum likelihood estimator   | Disease and normal sequencing data in bam format                                      |
| MuTect              | 2013           | 23396013      | DNA (HTS)             | Java based        | Bayesian model                                         | Tumour and normal sequencing data                                                   |
| TrAp                | 2013           | 23892400      | DNA (HTS)             | Java based        | Linear mixture model with evolutionary framework       | Tumour karyotypes and somatic hypermutation datasets                                 |
| Seo et al.          | 2013           | 23650637      | RNA (Microarray)      | –                 | Linear mixture model                                    | Disease-associated variants and expression of heterogeneous normal tissue            |
ter, repetitive region filter, short distance filter, etc.) to eliminate potential false negatives. Next, the reads remaining after filtration are divided into smaller fragments (so-called “split reads”) with even or pre-defined length, and both terminal parts are independently aligned to the reference genome. If they map to two different genes, they are further subjected to detection of fusion junction. The sequences of both genes are put together according to the fusion boundary, and the whole read is re-aligned to the candidate fusion gene to call supporting reads essential for the final selection of fusion. Another approach for the detection of fusion intersection is grouping discordantly mapped reads (“spanning reads”) according to the same breakpoints. Detection of fusion junction from such groups fuels prediction of the putative fusion transcript. Subsequently, the reads are re-aligned to predicted sequences and the predictions with the highest mapping scores aid identification of candidate fusion genes. The final selection of the fusion genes depends on several parameters including the number of supporting reads, quality of the alignment, and sequencing coverage [122, 123].

**Estimating sample purity**

Most genomics and expression-profiling studies including TCGA use a mixture of different clonal populations of tumour cells, which is often contaminated with stromal and immune cells. Indeed, many common tumours, such as pancreatic tumours, are intensively infiltrated by stroma [124], making it difficult to obtain homogenous material for genomic studies. Furthermore, epithelial cells are also often found in tumour samples, as they are at the interior surface of blood vessels necessary for providing nutrients for cancer cells. Methods like laser capture micro-dissection are rarely used in RNA-studies requiring stable material [125]. Estimating purity and clonality of a tumour sample containing a mixed population of cells requires accurate measurement of the proportion of tumour and stromal cell samples. Over the years several different methods (Table 7) have been developed to deconvolute genomic and transcriptomic data obtained from mixed-cell populations. Software packages based on these methods provide powerful tools for estimation of tumour heterogeneity and purity and in consequence identification of likely early driver events during tumorigenesis [126].

**Comparative and integrative analysis of tumour samples**

One of the major achievements of the TCGA project is the generation of different types of data from the same sample for a large number of tumours. This data generation is followed by uniform data processing and correlation with clinical information by Firehose and other analysis pipelines at various genome data analysis centres. The availability of such paired data allows detection of functional impact on genomic lesions (e.g., mutations, copy numbers, and gene fusions) on gene/miRNA (using RNA-Seq) and protein expression (using RPPA arrays) and pathway levels while reducing errors due to individual patient variation. Another example is the utilisation of multiple data types to identify integrated subtypes for a given tumour type using the iCLUSTER method [13, 14, 127]. Other approaches include integration of pathway information (e.g., PARADIGM & Paradigm-shift) and regulatory network information (e.g., GEMINI – [128]).

Comparative analysis of multiple tumour types increases the statistical power to detect common events that drive tumorigenesis and repurpose the therapy. For example, ERBB2-HER2 is mutated and/or amplified in subsets of glioblastoma, gastric, serous endometrial, bladder, and lung cancers. The result, at least in some cases, is responsiveness to HER2-targeted therapy, analogous to that previously observed for HER2-amplified breast cancer. There are more examples that underscore the importance of such comparative analysis [4].

**Future of cancer profile analysis**

As we are entering the era of $1000$ genome sequencing, tumour profiles are being sequenced routinely. Moreover, tumour catalogues and pre-clinical models [129, 130] have similar types of information available, with or without drug treatments. Integration of such datasets can speed up pre-clinical drug development and repurposing of available drugs. Tumour profiling by sequencing is also expected to enter both the pre-clinical and clinical setting for standardised testing as well as personalisation of medicine. However, the sequencing data fits the definition of “big data”, and a reliable computational infrastructure for storage, processing, analysis, and visualisation [131, 132] is required to make most of this avalanche of information [133]. Indeed, ambitious efforts like the cancer moonshot program and APOLO launched by the UT MD Anderson Cancer Centre, aim to combine big data warehousing with IBM WATSON based cognitive and adaptive learning to reduce cancer mortality for several tumour types, will fully realise the power of tumour profiling.

---

Authors declare no conflict of interest.

The authors thank Hubert Świerczyński, Wojciech Pieklik, Juliusz Pukacki, and Dr Cezary Mazurek from the Poznan Supercomputing and Networking Centre affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences for their help in preparation of the tables. This work was supported by the Foundation for Polish Science Welcome program grant No: 2010-3/3 to Maciej Wiznerowicz and UT MD Anderson Cancer Center intramural grants.

---

**References**

1. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012; 366: 883-92.
2. Ju YS, Alexandrov LB, Gerstung M, et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife 2014; 3.
3. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013; 499: 214-8.
Computational characterisation of cancer molecular profiles derived using next generation sequencing

4. Weinstein JN, Colisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013; 45: 1113-20.

5. Stratton MR. Exploring the genomes of cancer cells: progress and promise. Science 2011; 331: 1533-58.

6. Network TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008; 455: 1061-8.

7. Network TCGA. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012; 489: 519-25.

8. Network TCGA. Comprehensive molecular portraits of human breast tumours. Nature 2012; 490: 61-70.

9. Jackson SE, Chester JD. Personalised cancer medicine. Int J Cancer 2014 [Epub].

10. Vanneman M, Dranoff G. Combining immunotherapy and targeted therapies in cancer treatment. Nat Rev Cancer 2012; 12: 237-51.

11. Capdeville R, Buchdunger E, Zimmermann J, Matter A. Glivec (Gleevec) drug. Nat Rev Drug Discov 2002; 1: 493-502.

12. Nagayama A, Hayashida T, Jinno H, et al. Comparative effectiveness of neoadjuvant therapy for HER2-positive breast cancer: a network meta-analysis. J Natl Cancer Inst 2014; 106.

13. Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet 2006; 7: 335-41.

14. Network TCGA. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2011; 511: 543-50.

15. Network TCGA. Comprehensive genomic characterization defines human colon and rectal cancer. Nature 2012; 487: 330-7.

16. Grant GM, Fortney A, Gorreta E, et al. Microarrays in cancer research. Anticancer Res 2004; 24: 441-8.

17. Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet 2006; 7: 335-41.

18. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of methods for differential gene expression from RNA-seq data. Stat Appl Gen Mol Biol 2011; 10: Article 24.

19. Irizarry RA, Warren D, Spencer F, et al. Multiple-laboratory comparison of RNA-seq. Nat Biotechnol 2009; 27: 246-50.

20. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001; 2: 183-201.

21. Quackenbush J. Microarray data normalization and transformation. Nat Rev Genet 2002; 3: 256-65.

22. Yang YH, Speed T. Design issues for cDNA microarray experiments. Nat Biotechnol 2000; 18: 1121-6.

23. Alkodsi A, Louhimo R, Hautaniemi S. Comparative analysis of methods for evaluating read trimming effects on illumina NGS data analysis. Am J Hum Genet 2013; 93: 641-51.

24. Flicek P, Birney E. Sense from sequence reads: methods for aligning from next-generation sequencing data. Nat Meth 2009; 6: S22-32.

25. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Rev Genet 2011; 12: 443-51.

26. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics 2010; 95: 315-327.

27. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. PLoS One 2014; 9: e87864.

28. Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing illumina next-generation sequencing short read sequences. Source Code Biol Med 2014; 9: 8.

29. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics 2010; 95: 315-327.

30. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009; 10: R25.

31. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25: 1105-11.

32. Liu G-Z, Yang W, Ma Y-K, Wang X-I. IRNA: an integrative online toolkit for short reads from high-throughput sequencing data. Bioinformatics 2014; 30: 434-6.

33. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010; 11: R106.

34. Auer PL, DR. A two-stage poisson model for testing RNA-seq data. Stat Appl Gen Mol Biol 2011; 10: Article 26.

35. Di Y, SD, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-seq data. Stat Appl Gen Mol Biol 2011; 10: Article 24.

36. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 2010; 11: 422.

37. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 2008; 9: 321-32.

38. Zhou YH, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA-seq sequence count data. Bioinformatics 2011; 27: 2672-8.

39. Bullard J, Purdom E, Hansen K, Dudoit S. Evaluation of statistical methods for normalization and differential expression in RNA-Seq experiments. BMC Bioinformatics 2010; 11: 94.

40. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008; 18: 1509-17.

41. Demouëd F, Kapranov P, Udca C, et al. Prominent use of distal 5′ transcript start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res 2007; 17: 746-59.

42. Tess ML, Martelli PL, Frankish A, et al. The implications of alternative splicing in the ENCODE protein complement. Proc Natl Acad Sci U S A 2007; 104: 5495-500.

43. Lee JY, Yeh I, Park JY, Tian B. PoLyA_DB 2: mRNA polyadenylation sites in vertebrate genes. Nucleic Acids Res 2007; 35: D165-8.

44. Zhang R, Cheng Y, Tan BC-M, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol 2012; 30: 253-60.

45. Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. Accurate identification of human Alu and non-Alu RNA editing sites. Nat Meth 2012; 9: 579-81.

46. Ramaswami G, Zhang R, Piskol R, Keegan LR, Deng P, O’Connell MA, Li JB. Identifying RNA editing sites using RNA sequencing data alone. Nat Meth 2013; 10: 128-32.

47. Gregg C, Zhang J, Weissbourd B, Luo S, Schnroth GR Haig D, Dulac C. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science 2010; 329: 643-8.

48. Zang R, Li X, Ramaswami G, Smith KS, Turecki G, Montgomery SB, Li JB. Quantifying RNA allelic ratios by microfluidic multiplex PCR and sequencing. Nat Methods 2014; 11: 51-4.

49. Cabili MN, Trapnell C, Goff L, Kozkol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011; 25: 1915-27.

50. Djebali S, Davis CA, Mortazavi A, et al. Landscape of transcription in human cells. Nature 2012; 489: 101-8.

51. Khoury JD, Tannir NM, Williams MD, et al. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. J Virol 2013; 87: 8916-26.

52. Tang K-W, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. The landscape of viral expression and host gene fusion and adaptation in human cancer. Nat Commun 2013; 4: 2513.

53. Piskol R, Ramaswami G, Li J, & B. Reliable identification of genomic variants from RNA-Seq data. Am J Hum Genet. 2013; 93: 641-51.
59. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer 2004; 4: 177-83.
60. Kim TM, Xi R, Luquette LJ, Park RW, Johnson MD, Park PJ. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. Genome Res 2013; 23: 217-27.
61. Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability— an evolving hallmark of cancer. Nat Rev Mol Cell Biol 2010; 11: 220-8.
62. Baudis M. Genomic imbalances in 1918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. BMC Cancer 2007; 7: 226.
63. Xing M. Molecular pathogenesis and mechanisms of thyroid cancer. Nat Rev Cancer 2013; 13: 184-99.
64. Döhner H, Stilgenbauer S, Benner A, et al. Genomic Aberrations and Survival in Chronic Lymphocytic Leukemia. N Engl J Med 2000; 343: 1910-6.
65. Louhimo R, Lepikhova T, Monni O, Hautaniemi S. Comparative analysis of algorithms for integration of copy number and expression data. Nat Meth 2012; 9: 351-5.
66. Kallioniemi A, Kallioniemi O, Sadar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992; 258: 818-21.
67. Mullighan CG, Downing JR. Genome-wide profiling of genetic alterations in acute lymphoblastic leukemia: recent insights and future directions. Leukemia 2009; 23: 1209-18.
68. Weir BA, Woo MS, Getz G, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007; 450: 893-8.
69. Chiang DY, Getz G, Jaffe DB, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Meth 2009; 6: 99-103.
70. Pinkel D, Segraves R, Sudar D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 1998; 20: 207-11.
71. Bignell GR, Huang J, Greshock J, et al. High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res 2004; 14: 287-95.
72. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genomic structural polymorphism by sequencing on a population scale. Nat Genet 2011; 43: 269-76.
73. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 2009; 19: 1586-92.
74. Tuzun E, Sharp AJ, Bailey JA, et al. Fine-scale structural variation of the human genome. Nat Genet 2005; 37: 727-32.
75. Campbell PL, Stephens PJ, Pleasance ED, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008; 40: 722-9.
76. Karakoc E, Alkanc O, O’Roak BJ, Dennis MY, Vives L, Mark K, Rieder MJ, Nickerson DA, Eichler EE. Detection of structural variants and in-del events within exome data. Nat Meth 2012; 9: 176-8.
77. Fromer M, Moran JL, Chambert K, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing. Am J Hum Genet 2012; 91: 597-607.
78. Soon WW, Harirhan M, Snyder MP. High-throughput sequencing for biology and medicine. Mol Syst Biol 2013; 9: 640.
79. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science 2013; 339: 1546-58.
80. Wang Q, Jia P, Li F, Chen H, Liu J, Liu P, Lu Y, DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. Am J Hum Genet 2013; 93: 439-51.
81. Merid SK, Goranskaya D, Alexeyenko A. Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. BMC Bioinformatics 2014; 15: 308.
82. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med 2014; 6: 5.
83. Dees ND, Zhang Q, Kandoth C, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res 2012; 22: 1589-98.
84. Hua X, Xu H, Yang Y, Zhu J, Liu P, Lu Y. DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. Am J Hum Genet 2013; 93: 439-51.
85. Merid SK, Goranskaya D, Alexeyenko A. Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. BMC Bioinformatics 2014; 15: 308.
86. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med 2014; 6: 5.
87. Zhang Q, Ding L, Larson DE, et al. CORDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. Bioinformatics 2010; 26: 464-9.
88. van Dyk E, Reinders MJ, Wessels LF. A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control. Nucleic Acids Res 2013; 41: e100.
89. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009; 4: 1073-81.
90. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 2010; 7: 248-9.
91. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011; 39: e118.
92. Carter H, Shen S, Isik L, Tyekucahe S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 2009; 69: 6660-7.
93. Ye J, Pavlicek A, Lunney EA, Rejto PA, Teng CH. Statistical methods on nonrandom clustering with application to somatic mutations in cancer. BMC Bioinformatics 2010; 11: 11.
94. Hodiis E, Watson IR, Kryukov GV, et al. A landscape of driver and passenger mutations in individual cancer genomes by network enrichment analysis. BMC Bioinformatics 2014; 15: 308.
95. Merid SK, Goranskaya D, Alexeyenko A. Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. BMC Bioinformatics 2014; 15: 308.
96. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011; 39: e118.
97. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005; 102: 15545-50.
106. Wendl MC, Wallis JW, Lin L, Kandoth C, Mardis ER, Wilson RK, Ding L. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics 2011; 27: 1595-602.

107. Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene set analysis for cancer mutation data. Genome Biol 2010; 11: R112.

108. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS One 2010; 5: e8918.

109. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol 2011; 18: 507-22.

110. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res 2012; 22: 398-406.

111. Leiserson MD, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol 2013; 9: e1003054.

112. Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Med Genomics 2011; 4: 34.

113. Bartram CR, de Klein A, Hagemeijer A, et al. Translocation of c-ab1 oncogene correlates with the presence of a Philadelphia chromosome in chronic myelocytic leukaemia. Nature 1983; 306: 277-80.

114. Soda M, Choi YL, Enomoto M, et al. Identification of the trans-forming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007; 448: 561-6.

115. Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005; 310: 644-8.

116. Kohno T, Ichikawa H, Totoki Y, et al. KIF5B-RET fusions in lung adenocarcinoma. Nat Med 2012; 18: 375-7.

117. Edgren H, Murumagi A, Kangaspeska S et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 2011; 12: R6.

118. Gridelli C, Peters S, Sgambato A, Casaluce F, Adjei AA, Ciardiello F. ALK inhibitors in the treatment of advanced NSCLC. Cancer Treat Rev 2014; 40: 300-6.

119. Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 2008; 456: 66-72.

120. Maher CA, Kumar-Sinha C, Cao X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009; 458: 97-101.

121. Zhao Q, Caballero OL, Levy S, et al. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. Proc Natl Acad Sci U S A 2009; 106: 1886-91.

122. Wang Q, Xia J, Jia F, Pao W, Zhao Z. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. Brief Bioinform 2013; 14: 506-19.

123. Supper J, Gugenmus C, Wollnik J, et al. Detecting and visualizing gene fusions. Methods 2013; 59: S24-8.

124. Biankin AV, Waddell N, Kassahn KS, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 2012; 491: 399-405.

125. Paweletz CP, Liotta LA, Petricoin EF, 3rd. New technologies for biomarker analysis of prostate cancer progression: Laser capture microdissection and tissue proteomics. Urology 2001; 57: 160-3.

126. Yadav VK, De S. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. Brief Bioinform 2014 [Epub].

127. Shen R, Mo Q, Schultz N, Seshan VE, Oshen AB, Huse J, Ladanyi M, Sander C. Integrative subtype discovery in glioblastoma using iCluster. PLoS One 2012; 7: e55236.

128. Yan Z, Shah PK, Amin SB, et al. Integrative analysis of gene and miRNA expression profiles with transcription factor-miRNA feed-forward loops identifies regulators in human cancers. Nucleic Acids Res 2012; 40: e135.

129. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012; 483: 603-7.

130. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012; 483: 570-5.

131. Cerami E, Gao J, Dognosz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012; 2: 401-4.

132. Samur MK, Yan Z, Wang X, Cao Q, Munshi NC, Li C, Shah PK. canEvolve: a web portal for integrative oncogenomics. PLoS One 2013; 8: e56228.

133. Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. Genes Dev 2011; 25: 534-55.

Address for correspondence
Parantu K. Shah, PhD
Institute for Applied Cancer Science
The UT MD Anderson Cancer Centre
Houston, TX 77030
phone: 713 794 4727
fax: 713 792 6882
e-mail: parantu.shah@gmail.com