

REVIEW

Open Access

# Computational methods and resources for the interpretation of genomic variants in cancer

Rui Tian<sup>1</sup>, Malay K Basu<sup>1,2</sup>, Emidio Capriotti<sup>1,2,3\*</sup>

From VarI-SIG 2014: Identification and annotation of genetic variants in the context of structure, function and disease

Boston, MA, USA. 12 July 2014

## Abstract

The recent improvement of the high-throughput sequencing technologies is having a strong impact on the detection of genetic variations associated with cancer. Several institutions worldwide have been sequencing the whole exomes and or genomes of cancer patients in the thousands, thereby providing an invaluable collection of new somatic mutations in different cancer types. These initiatives promoted the development of methods and tools for the analysis of cancer genomes that are aimed at studying the relationship between genotype and phenotype in cancer.

In this article we review the online resources and computational tools for the analysis of cancer genome. First, we describe the available repositories of cancer genome data. Next, we provide an overview of the methods for the detection of genetic variation and computational tools for the prioritization of cancer related genes and causative somatic variations. Finally, we discuss the future perspectives in cancer genomics focusing on the impact of computational methods and quantitative approaches for defining personalized strategies to improve the diagnosis and treatment of cancer.

## Background

The advances in high-throughput sequencing techniques are allowing us to identify a large number of genetic variants in human [1,2] and understand the relationship between genotype and phenotype in many genetic disorders [3]. In contrast to Mendelian disorders, in which a disease is the result of inherited variations present in a single gene or a small set of genes, cancer is mainly driven by accumulated somatic variations in multiple genes. These mutations enable a particular subpopulation of cells to proliferate and survive more efficiently than their neighbors [4-6]. The different types of somatic genetic variations detected in cancer samples vary from single nucleotide variants, short insertion and deletion (indels), large copy number alterations, to structural rearrangements [7]. Thus, the identification of causative genomic variations is key point for understanding the mechanism

of cancer. The solution of this challenging task is limited by the accuracy of sequencing technology and the large number of genetic alterations observed in cancer genome. Although current whole-exome sequencing is performed with a coverage between 100x to 150x, still many false positive arise from sequencing error, inaccurate alignments and admixture of noncancer and different subclonal cells [8].

Even with a perfect sequencing approach the detection of causative variants remains a complex task. Indeed, among somatic variants, a large fraction, defined as “passengers”, have no impact on the neoplastic process. But a smaller set of genetic variants, referred to as “drivers”, provide a selective growth advantage, estimated to be minute (~0.4%), to the cell [9]. The accumulation of those variants and their effect, over the years, can generate a tumor. According to a recent review [9], somatic variations observed in common solid tumors affect on average between 33 and 66 genes. The predominant type of variations is the single nucleotide substitution, which accounts for more than 90% of somatic changes [1]. The

\* Correspondence: emidio@uab.edu

<sup>1</sup>Division of Informatics, Department of Pathology, University of Alabama at Birmingham, 619 19<sup>th</sup> St. South, 35249 Birmingham (AL), USA  
Full list of author information is available at the end of the article

role of these genes in tumorigenesis is determined by the functional impact of somatic variants they harbor. In general, mutations that enhance the function of oncogenes and impair the activity of tumor suppressor genes result in a selective growth advantage for the cancer cell.

The frequency of observed somatic variants draw a mutational landscape made up of few “mountains” and a large number of “hills” that respectively correspond to frequently and rarely mutated genes across tumor samples [9]. Among the frequently mutated genes, the distribution of somatic variants can reflect the differences between oncogenes and tumor suppressor genes. Oncogenes tend to be recurrently mutated at the same amino acid positions, whereas tumor suppressor genes are generally mutated in various positions throughout their length. An example is provided in Figure 1, which shows the frequencies of somatic mutation in APC (a tumor suppressor gene) and KRAS (an oncogene) from 220 samples of colon adenocarcinoma from The Cancer Genome Atlas (TCGA) consortium. In this example, ~60% of the variants in APC are stop-gain which cause loss of function. In contrast, ~50% of the mutations in KRAS are missense variants in position 12 which are expected to increase the protein function.

Given that the cancer is a result of interplay of various types of genetic changes, understanding of the role of somatic mutations in tumorigenesis is a complex problem in which different combinations of somatic variants affect different network of genes and associated pathways. Nevertheless, significant breakthroughs have been made in the development of computational methodologies which allow us to identify new driver mutations and genes by analyzing large sets of patients with different tumors [10]. Currently, the main bioinformatics challenges in the analysis of cancer genome consist of the following:

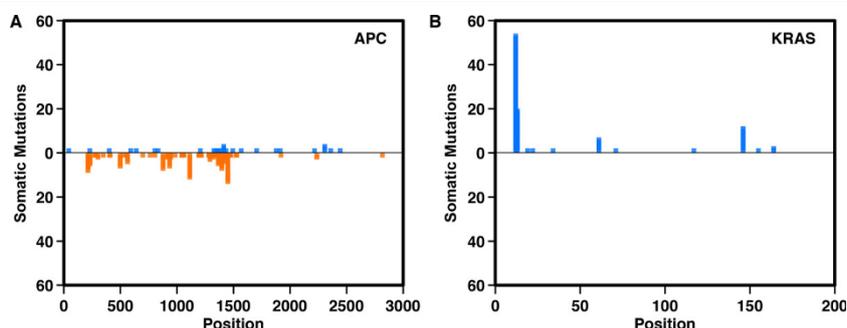
1. Robust pipelines for the detection of genetic variations;

2. Creation of a benchmark dataset of cancer driver mutations and genes;
3. Accurate methods for prioritizing cancer driver mutations and genes;
4. New algorithm for predicting the impact of somatic variants at pathway/ network level;
5. Translational approaches that make the results of computational analysis clinically applicable [11,12].

In this review we focus on the description of computational methods for cancer genome interpretation. First, we describe the basic steps for the detection of genetic variants. We summarize the currently available data sources for implementing and benchmarking new computational tools for cancer genome analysis. Second, we review the available methods for the prioritization of driver mutations and genes. We also include a section describing tools for predicting the impact of genetic variants at network level and methods for estimating the consistence of subclonal populations. Finally, we discuss the future perspective in the field, highlighting the contribution of computational approaches to cancer genomics. In contrast to the recently published cancer genomic reviews [11,13,14], we present an analysis of the somatic mutations in cancer and the data available online and provide a brief description and availability of selected computational tools for the analysis of cancer genomes. This review is targeted towards readers with background in computational biology and bioinformatics, who want to have quick introduction to the available resources and tools for the analysis of cancer genome.

#### Variant calling, filtering and annotation

Accurate variant calling is the prerequisite of any cancer genome analysis, but it is hindered by several limitations. The first limitation stems from the inherent noise and errors in the sequencing technology. Errors are also introduced in the procedures of short reads alignment,



**Figure 1** Number of missense and stop gain somatic mutations in APC and KRAS detected in 220 samples of colon adenocarcinoma from TCGA studies. In blue and orange are reported the nonsynonymous and stop gain somatic variants respectively. For APC we detected a total number of 265 somatic variants 28 of which are nonsynonymous and 150 stop gain. Among the 28 stop-gain variants only 6 are recurrent. For KRAS we detected a total number of 101 somatic variants 97 of which are nonsynonymous. Among the observed nonsynonymous variants 53 are observed in position 12 and only 6 are not recurring.

especially in the low complexity regions of the genome. In addition, a recently published work showed that no single variant calling approach is able to comprehensively capture all genetic variations [15]. Thus, there is still room for improvement in variant calling algorithms.

In general, the variant calling procedure consist of 3 main steps:

1. Short read alignment and mapping to the reference genome.
2. Removing PCR duplicates, realignment and recalibration.
3. Variant calling, filtering and annotation.

This procedure includes the calibration of pre- and post-experimental factors to identify reliable variants from the raw data. Such factors include sequencing coverage, single end or paired end sequencing, short read alignment, PCR duplicates, matched sample sequencing, variant calling algorithm, etc. [11,15,16]. A representation a typical variant calling pipeline is provided in Figure 2.

A plethora of tools have been developed for variant calling purposes. In cancer studies, a standard workflow for variant calling starts with the alignment against the reference genome using BWA [17] or Bowtie [18]. The resulting alignment (SAM file) is generally converted in binary format (BAM) and indexed using SAMtools [19]. The removal of PCR duplicates can be performed using Picard (<http://sourceforge.net/projects/picard/>). This step is followed by realignment and recalibration with GATK [20]. Finally, variant calling is performed using a standard variant caller like GATK or specialized tools such as VarScan 2 [21] and MuTect [22]. This step can be improved by the score recalibration that reduces the number of false positive calls. The final output of this pipeline is a VCF (variant calling format) file.

Although, theoretically the variant calling pipelines are straightforward, the results from different variant callers agree only on a small subset of variants. In a systematic

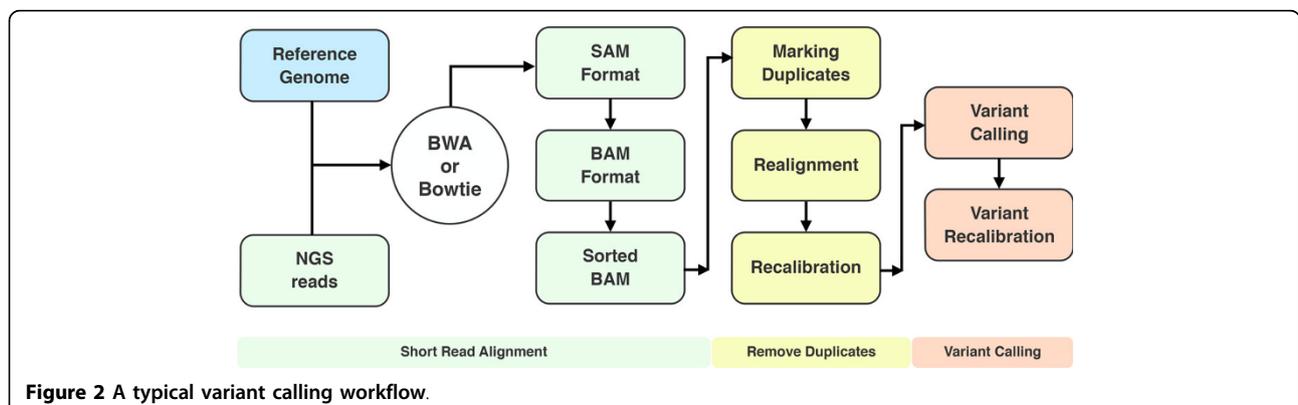
test of several variant callers, namely GRISP [23], GATK, SAMtools, SNVer [24], VarScan 2, only ~50% of the SNPs are shared by all these tools, while the overlap of identified indels is even lower [15]. Recently, a similar study [16] has been performed by comparing the performances of another set of six variant callers, including EBCall [25], JointSNVMix [26], MuTect, SomaticSniper [27], Streika [28] and VarScan 2 [16]. Using experimentally validated SNVs as benchmark, it was reported that VarScan 2 and MuTect are among the best variant callers for analyzing matched normal and tumor samples.

After the variant calling file (VCF) has been obtained, a filtering procedure is often necessary for the downstream analysis. This filtering procedure aims to reduce the number of false positives corresponding either to low quality and common variants. In particular the common variants, which are assumed to have no implications in tumor development and progression, are filtered out by comparison with the germline polymorphisms collected in publicly available databases, such as dbSNP [29] or EVS (Exome Variant Server, <http://evs.gs.washington.edu/EVS/>). The most recent versions of dbSNP (build 142) contains more than 110 million human SNPs, while the current EVS data release (ESP6500SI-V2) include all the exome variant data from 6503 human samples. Some of the common tools used for filtering variants are SnpSift [30], GATK [20], VCFtools (<http://vcftools.sourceforge.net>).

Finally, variants are annotated by mapping each variant to their corresponding gene. This procedure is essential for understanding their functional consequences. Among the most popular tools for variant annotation are ANNOVAR [31], snpEff [32] and VEP [33].

### Cancer variation data and databases

Large-scale cancer genomic experiments, funded by several national and international consortiums, are generating an amount of data in the magnitude of PetaBytes (PB). The space needed to host the data only from The Cancer Genome Atlas (TCGA) is ~1.1PB



(see [https://cghub.ucsc.edu/summary\\_stats.html](https://cghub.ucsc.edu/summary_stats.html)). The analysis of this data enabled the development of several meta databases and resources for the annotation of cancer genomes. In this section, we describe some of the repositories and databases available that collect somatic variants and driver genes putatively involved in the cancer.

#### **Cancer mutation data**

The Cancer Genome Atlas (TCGA) consortium, which began in United States in 2006, is a comprehensive and coordinated effort to understand the molecular basis of cancer using several genomic analysis techniques. The data generated by the experiments are made available through the TCGA Data Portal and the Cancer Genomics Hub (CGHub) [34]. After signing the certification data agreement it is possible to access data about 36 cancer types. The files containing the binary version of the short DNA sequence read alignments (referred as BAM files) can be downloaded using the GeneTorrent application available at the CGHub website (<https://cghub.ucsc.edu/>).

Since the downloading of all the BAM files remains costly in term of time and storage, it is advisable to get all the pre-processed variant files (VCFs) generated by different institutions within the TCGA consortium. Unfortunately, so far the organization of the data in the TCGA repository is not optimized and available data have to be manually selected. In particular, variant data from different platforms, ranging from SNP arrays to SOLID and Illumina sequencers, are reported in different file formats. Out of these files, the most informative ones are the VCF files, which contains the genetic variants for both normal and tumor samples. Currently, VCF file are provided only for a subset of the tumor types (~67%).

The International Cancer Genome Consortium (ICGC) was launched in 2007 to coordinate the efforts of characterizing more than 50 different cancer types from 25,000 patient genomes [35-37]. The results have been published and made publicly available [38-44]. To provide an overview of the variation data, we re-analyzed the somatic mutations publicly available at the ICGC portal, whereas the access to the germline variants requires the approval of a data agreement. The statistics of the release 17 (September 2014) of the ICGC data portal (<https://dcc.icgc.org>) show that in total 12,232 cancer genomes have been sequenced. In the samples from these donors, collected from 18 cancer primary sites, more than 9.8 million simple somatic mutations have been identified. A summary of the mutations detected for each cancer type is provided in Table 1. In this work, we consider samples from 42 sequencing projects corresponding to 33 cancer types. From the available data we excluded the samples from Acute Lymphoblast Leukemia for which only 3 mutations were detected. A brief description of the datasets analyzed in this manuscript as well as the final list of

codes of each cancer project are provided in Section 1 of the Additional file 1 and Supplementary Table 1. In the data collected by the ICGC, breast cancer is the most studied cancer type in which more than 1,100 individuals have been screened.

To study the occurrence of somatic mutations across different donors, we performed a recurrence analysis calculating the *Fraction of Somatic Mutations* and the *Fraction of Donors* corresponding to different subsets of somatic mutations (Section 2 in Additional file 1).

The analysis of whole set of somatic mutations (Pan-Cancer) revealed that a large fraction of them are occurring in a single patient and only ~1.7% are recurring more than once (see Table 1). This percentage decreases to 1.2% if the cancer types are considered separately. The plot in Figure 3A shows the *Fraction of Somatic Mutations* at different levels of *Mutation Recurrence*.

In addition, we calculated the *Fraction of Donors* covered by subsets of recurrent mutations (Figure 3B). Considering only the recurrent mutations, (observed at least in two donors), they are held by 96% of the individuals in all cancer types (Table 1). This percentage decreases to 82% when the cancer types are considered separately. In Figure 3C we plot the *Fraction of Somatic Mutations* and the *Fraction of Donors* of affected at different *Mutation Recurrence* thresholds. This curve allows us to estimate the fraction of donors affected by a subset of recurrent mutations that are more likely to have a functional impact. Although it is well accepted that each cancer sample is different and only a small fraction of variants are recurrent, we show that with mutations recurring more than 30 times (53 variants) explains a comparatively large fraction of patients (~40%). However, this data could be biased toward cancer types with higher number of samples and with higher of mutations detected.

We performed the recurrence analysis on each cancer type fitting the points calculated at different *Mutation Recurrence* threshold. The results in Figure 4 show different trend in 27 tumor types for which the regression curves can be calculated. Although in some cases the fitting is less accurate (THCA and PBCA), we can compare different tumor types estimating the fraction of somatic mutations covering 95% of the donors (Supplementary Table 2, Additional file 1). The smallest value corresponds to the Esophageal Adenocarcinoma (ESAD) for which we estimated somatic mutation rate of ~0.006%. The highest value is reported for the Pediatric Brain Cancer (PBCA) for which 75% the mutations are needed to cover 95% of the samples.

In general, it is expected that heterogeneous cancer types show a large variety of recurrent somatic mutations, which is inversely proportional to the rate of decay of the complementary cumulative distribution

**Table 1**

Cancer Type	Donors	Total Unique Mutations	Total Recurrent Mutations	Donors with Recurrent Mutations
BLCA	233	53,638	737 (1.4%)	219 (94.0%)
BOCA	66	1,422	7 (0.5%)	39 (59.1%)
BRCA	1,071	275,612	1,252 (0.5%)	929 (86.7%)
CLLE	109	5,292	4 (0.1%)	10 (9.2%)
CMDI	129	86	13 (15.1%)	113 (87.6%)
COAD	216	105,786	3,896 (3.7%)	215 (99.5%)
EOPC	11	25,575	2 (0.0%)	4 (36.4%)
ESAD	95	1,780,883	25,425 (1.4%)	95 (100.0%)
ESCA	88	7,256	17 (0.2%)	37 (42.0%)
GACA	9	1,014	0 (0.0%)	0 (0.0%)
GBM	268	19,852	324 (1.6%)	260 (97.0%)
KIRC	404	26,371	688 (2.6%)	372 (92.1%)
KIRP	156	12,932	218 (1.7%)	144 (92.3%)
LAML	75	60,203	7,623 (12.7%)	71 (94.7%)
LGG	279	13,083	432 (3.3%)	278 (99.6%)
LIAD	30	917	11 (1.2%)	19 (63.3%)
LICA	29	747,334	27,107 (3.6%)	6 (20.7%)
LINC	244	437,403	6003 (1.4%)	244 (100.0%)
LIRI	208	2,124,689	4161 (0.2%)	208 (100.0%)
LUSC	289	125,351	1400 (1.1%)	268 (92.7%)
MALY	44	311,297	686 (0.2%)	44 (100.0%)
NBL	41	137	2 (1.5%)	4 (9.8%)
ORCA	50	5,604	35 (0.6%)	38 (76.0%)
OV	181	919,769	869 (0.1%)	119 (65.7%)
PACA	504	1,630,944	6,098 (0.4%)	500 (96.5%)
PAEN	35	112,823	804 (0.7%)	32 (91.4%)
PBCA	248	130,608	1,231 (0.9%)	89 (35.9%)
PRAD	264	90,599	776 (0.9%)	256 (96.7%)
READ	80	23,499	556 (2.4%)	80 (100.0%)
RECA	105	475,986	6,067 (1.3%)	95 (90.5%)
SKCM	323	226,850	11,908 (5.2%)	323 (100.0%)
STAD	289	142,496	3,441 (2.4%)	276 (95.5%)
THCA	411	51,759	4,619 (8.9%)	311 (75.7%)
PanCancer	6,584	9,871,474	171,314 (1.7%)	6,296 (95.6%)

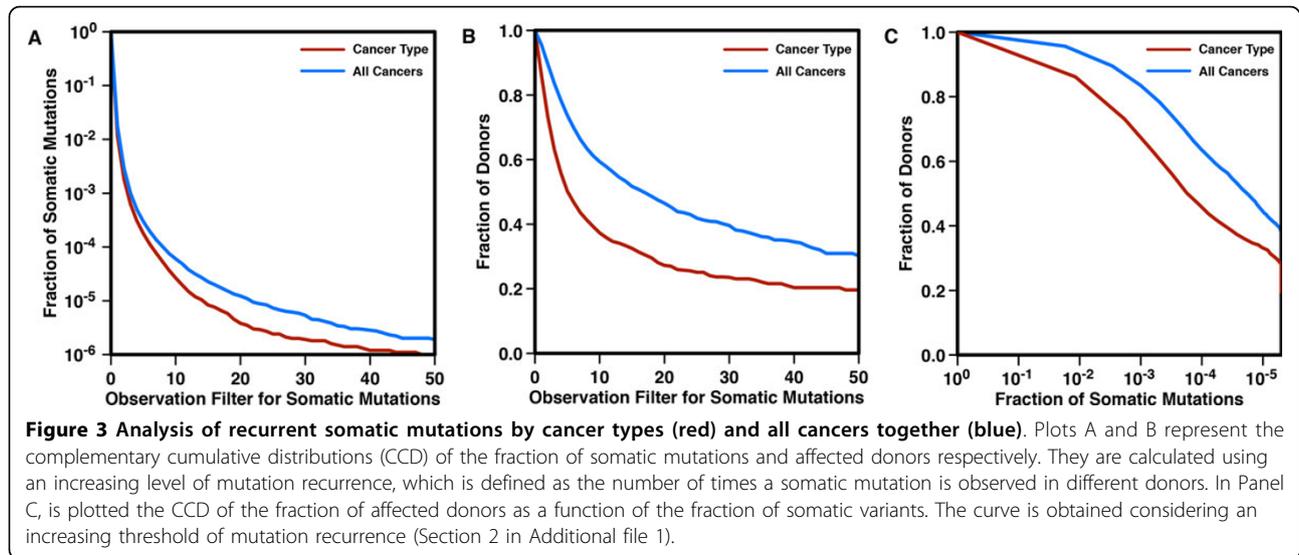
Somatic mutations in 33 tumor types from 42 projects from the ICGC release 17 (September 2014). *Donors*: total number of individual for each tumor type. *Total Unique Mutations*: number of somatic mutations in the whole genome for each cancer type removing repetitions. *Total Recurrent Mutations*: fraction of *Total Unique Mutations* observed at least in two donors for the same tumor type. *Donors with Recurrent Mutations*: fraction of donors with at least one recurrent somatic mutation.

(CCD). Thus, the analysis of recurrent mutations based on the calculation of the CCD can provide an estimation of the mutational heterogeneity of each cancer type (see Figure 4). Furthermore, the CCDs in Figures 3C and 4 provide an estimation of the maximum fraction of patients that can be recovered using a decreasing pool of recurrent variants.

Focusing on the exonic regions for each cancer type, we observed different number of somatic mutations per individual (Figure 5). The statistic shows a difference of ~2.5 orders in magnitude between the number of somatic mutations in Chronic Myeloid Disorders (CMDI) and Skin Cutaneous Melanoma (SKCM) for which their

median values per individual in the exonic regions is 1 and ~400 respectively (Table 1). This result is consistent with the difference in the rate of decay of the CCDs observed for the two cancer types in Figure 4.

In addition, we can use the number recurrent mutations in the exonic regions to estimate the tumor type similarities. Using the subset of mutations repeated more than once in the whole dataset (PanCancer) we built a vector, in which each element represents the number of donors affected by a mutated gene. Only genes harboring the aforementioned recurrent mutations are considered. The gene-based vectors of each cancer type are then used to calculate the cosine similarity (Section 3 in



Additional file 1). In Figure 6 we report the tumor dendrogram obtained from 33 cancer types. Broadly, based just on the similarity between the vectors of recurrent mutated genes we can cluster the tumors in two major subgroups of cancer. The figure shows on the first subgroup on the top-left side Colon Adenocarcinoma (COAD), and Rectum Adenocarcinoma (READ), in which high fractions of donors are mutated in APC, TP53 and KRAS, in the same cluster. In the second subgroup we observe 4 different liver cancers (LINC, LIRI, LICA and LIAD), with recurrent mutation in CTNBB1, in the same cluster.

These results confirm the hypothesis that recurrent somatic mutations, which represent only 1.7% of the whole PanCancer mutations, still retain meaningful information about the mutation profile of each cancer type that allows us to cluster them according to their similarity. Finally, it is noteworthy that filtering out non-recurrent somatic variants we significantly reduce the number putative driver mutations to a relatively small set of variants (~170,000), which are held by ~96% of the patients (see Table 1).

Although the analysis of somatic mutation data from multiple projects provides interesting insights about the genetic cause of tumor, we need to remember that data can have biases due to the use of different sequencing protocols and variant-calling procedures. These problems can be amplified with bias in sampling, significantly different size of the cohorts.

#### **Curated databases and cancer gene lists**

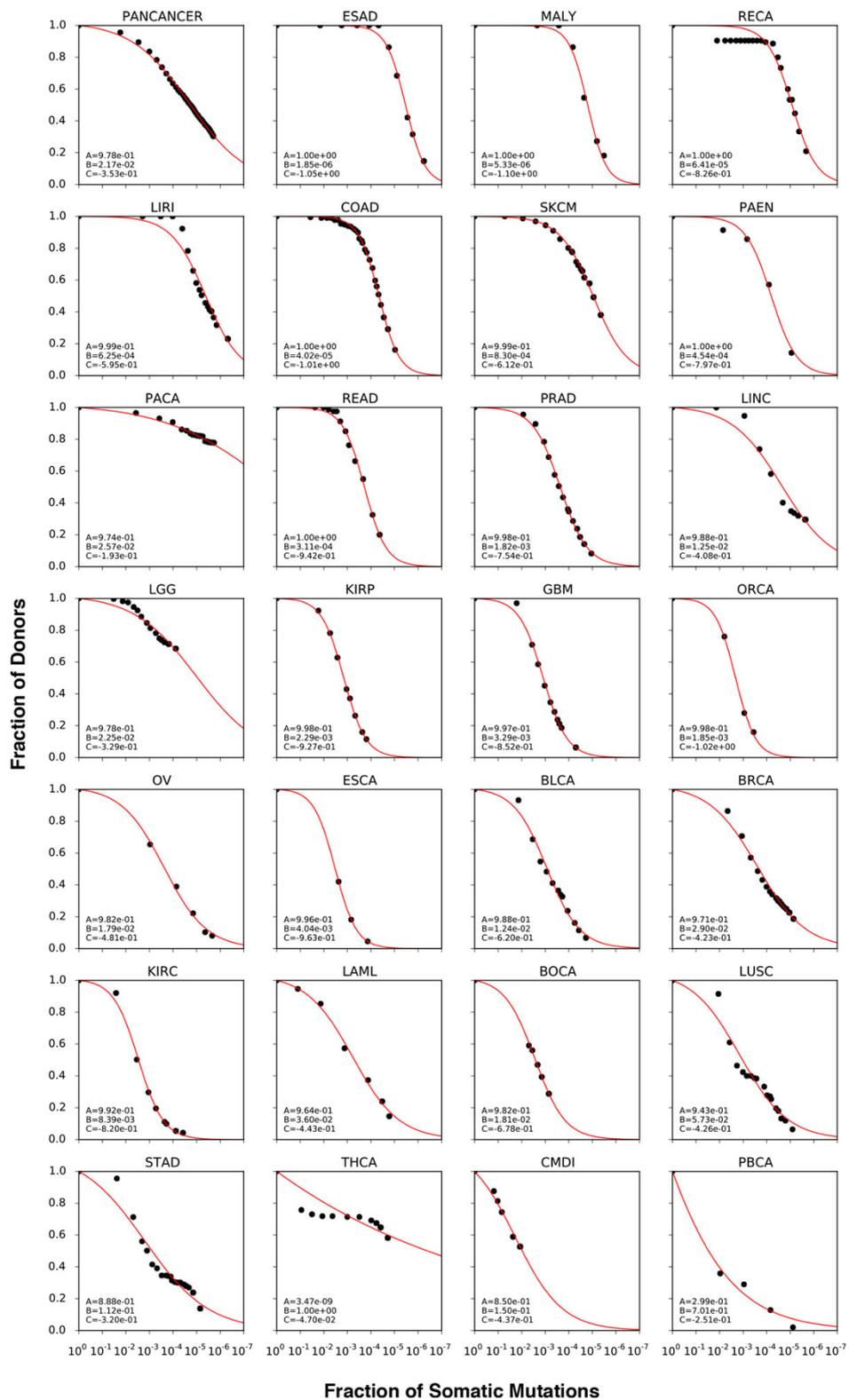
One problem that has been creating a constant roadblock in developing better tools and methods for cancer genome analysis is the lack of reference benchmark datasets of known cancer variants. The datasets discussed in this section are curated known collections of cancer variants

that are used frequently as a substitute of a reference dataset.

The most popular collection of the somatic mutations in cancer is the Catalogue Of Somatic Mutation In Cancer (COSMIC). The last release of COSMIC (version 71, Sep 2014) contains ~2.1 million unique somatic variants detected in ~1 million tumor samples. An important manually curated resource made available through the COSMIC web page is the Cancer Gene Census list. Currently it contains 547 genes for which mutations have been causally implicated in cancer. Out of 547 cancer genes, 85% of them harbor only somatic mutations, 7% harbor only germline variants and the remaining 8% harbor both types.

Beside the Cancer Gene Census there are other lists of putative cancer genes identified using computational or expert-based approaches. One is a list of 125 cancer genes, reported in a publication by Vogelstein and collaborators [9]. The list comprises 54 oncogenes (43%) and 71 tumor suppressor genes (57%). A larger list of 2,125 genes collected by Bushman Lab [45] has been obtained as a union of 8 different datasets. More recently, an analysis of 21 cancer types was published. The authors prioritized the cancer causing genes according to their observed mutation frequency across multiple samples [46]. The final list thus obtained, combining all the cohorts, contains 146 genes, 64 of which are highly significantly mutated (44%), 49 significantly mutated (34%) and 32 near significant (22%). A better description of the available lists of putative cancer genes obtained from the analysis large cancer studies is provided in recent publications [47-49].

Other useful databases for the annotation of cancer genome are NCG [50] and Cancer3D [51]. The NCG 4.0 database contains 2,000 cancer genes, 537 of which



**Figure 4 Cancer-specific complementary cumulative distributions.** The plots shows the decreasing trend of the *Fraction of Donors* (Y axis) as a function of the *Fraction of Somatic Mutations* (X axis) in 27 cancer types for which there are a minimum of 4 points for fitting from at least 50 donors. The PanCancer plot is calculated merging all the cancer types together. The results of the fitting procedure are reported in Supplementary Table 2, Additional file 1.

**Table 2. Cancer genome databases and resources**

Resource	URL	Ref
<i>Data repositories</i>		
CGHub	<a href="https://cghub.ucsc.edu">https://cghub.ucsc.edu</a>	[34]
COSMIC	<a href="http://cancer.sanger.ac.uk">http://cancer.sanger.ac.uk</a>	[132]
ICGC	<a href="https://dcc.icgc.org/">https://dcc.icgc.org/</a>	[133]
TCGA	<a href="https://tcga-data.nci.nih.gov/">https://tcga-data.nci.nih.gov/</a>	[38]
<i>Cancer gene lists</i>		
Bushman Lab	<a href="http://www.bushmanlab.org/links/genelists">http://www.bushmanlab.org/links/genelists</a>	[45]
Cancer Gene Census	<a href="http://cancer.sanger.ac.uk/census">http://cancer.sanger.ac.uk/census</a>	[134]
TumorPortal	<a href="http://cancergenome.broadinstitute.org/">http://cancergenome.broadinstitute.org/</a>	[46]
Vogelstein List	<a href="http://goo.gl/4EmFG6">http://goo.gl/4EmFG6</a> (Table S2A)	[9]
<i>Cancer genome resources</i>		
CaGe	<a href="http://mgrc.kribb.re.kr/cage/">http://mgrc.kribb.re.kr/cage/</a>	[60]
Cancer3D	<a href="http://www.cancer3d.org/">http://www.cancer3d.org/</a>	[51]
Cancer Genomics Browser	<a href="https://genome-cancer.ucsc.edu/">https://genome-cancer.ucsc.edu/</a>	[59]
cBioPortal	<a href="http://www.cbioportal.org/">http://www.cbioportal.org/</a>	[57]
DriverDB	<a href="http://driverdb.ym.edu.tw/DriverDB">http://driverdb.ym.edu.tw/DriverDB</a>	[61]
IntOGen	<a href="http://www.intogen.org/">http://www.intogen.org/</a>	[62]
NCG	<a href="http://ncg.kcl.ac.uk/">http://ncg.kcl.ac.uk/</a>	[50]

have been experimentally validated and 1,463 statistically inferred from the analysis of candidate cancer genes in 3460 human exomes and genomes from 23 different cancer types. For each cancer gene, NCG provides information about duplicability, evolutionary origin, expression, functional annotation, interaction network with human proteins and microRNAs. Finally, protein structure is also a valuable piece of information for predicting the impact genetic variants [52] and effectiveness of drugs [53,54]. Cancer3D database collects ~275,000 somatic mutations mapped to ~15,000 proteins that correspond to more than 24,300 structure from the Protein Data Bank [55]. The idea is to study the impact of missense somatic mutations on protein three-dimensional structure. The Cancer3D database includes predictions from e-Drug and e-Driver, two recently developed tools for predicting sensitivity to 24 anticancer compounds and cancer driver proteins [56].

#### **Resources for the visualization and analysis of cancer genome data**

In the last few years, several web portals have been developed for the visualization and analysis of all kinds of data in cancer genome studies. Among the most used is the cBioPortal maintained by the Memorial Sloan Kettering Cancer Center [57,58]. The cBioPortal can be queried either using the web interface or through programmable application programming interface (API). The cancer data displayed by the cBioPortal are extracted from 18,469 samples from 80 cancer genome studies. To facilitate the

submission of queries, cBioPortal implements the Onco Query Language (OQL) for selecting and defining genetic alterations for any subset of data. The user interface allows searching for somatic mutations or copy number alteration across different samples and cancer studies selecting variations occurring in user-defined gene list or in 22 pre-defined groups cancer-specific genes. When available, expression level from RNA-Seq experiments, methylation and protein phosphorylation data can also be retrieved.

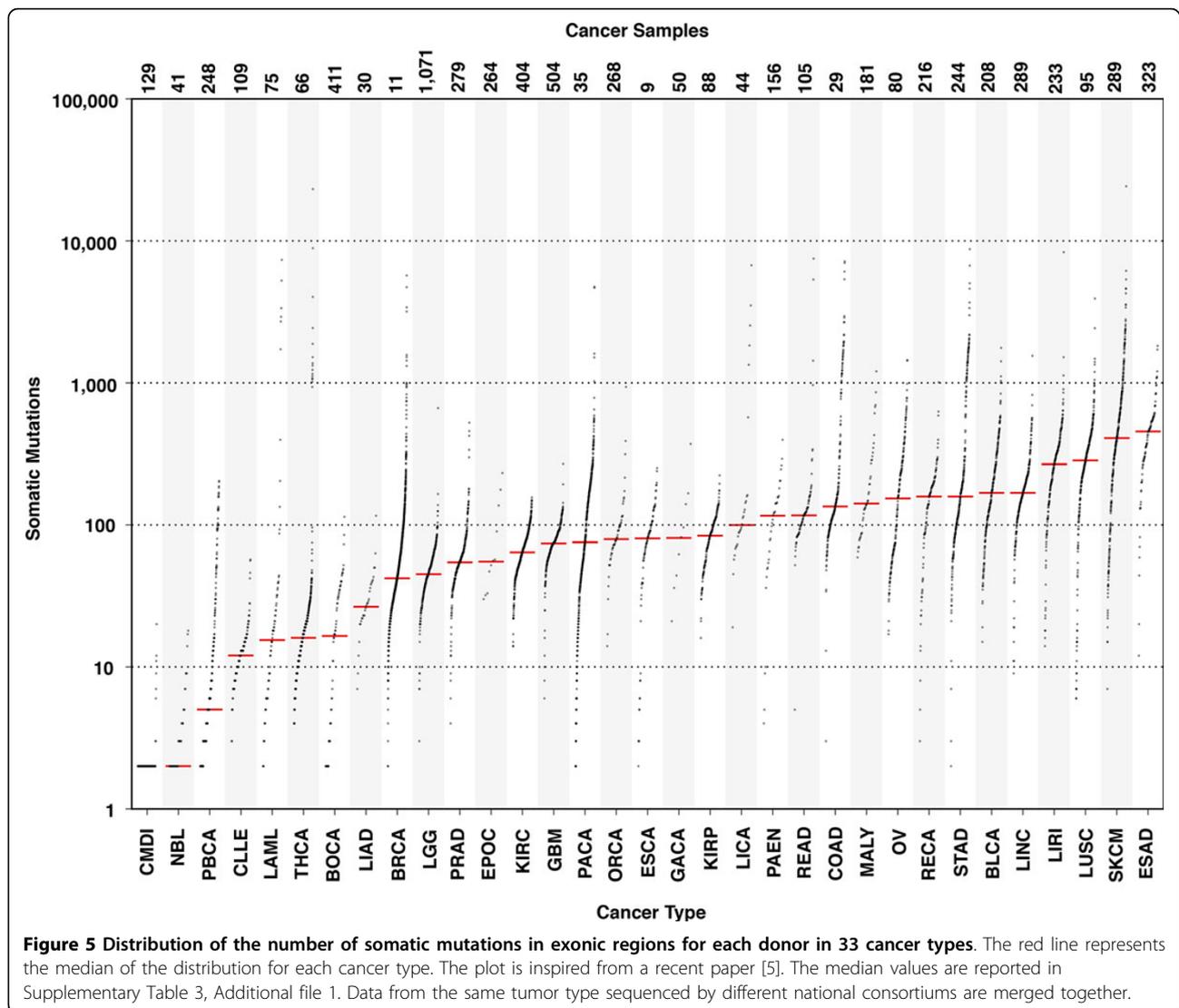
A possible alternative to cBioPortal is the Cancer Genome Browser [59]. The Cancer Genome Browser allows extraction and visualization of information from 574 datasets. Among these two resources, cBioPortal provides better integration with pathways data and the Cancer Genome Browser allows better visualization of clinical data.

Other three interesting resources for the analysis of cancer genome data are the Cancer Gene annotation system for Cancer Genomics (CaGe) [60], the DriverDB [61] and the Integrative Oncogenomics Cancer Browser (IntOGen) [62]. CaGe is a cancer genome annotation system for the classification of candidate genes from cancer genome studies, using either previously reported or novel categories of cancer genes, providing insights about the underlying carcinogenic mechanisms through pathway analysis. DriverDB incorporates data from 6,079 exome-sequencing experiments from 33 cancer studies. It integrates annotation databases and eight bioinformatics algorithms for detecting driver genes and mutations. Finally, Integrative OncoGenomics (IntOGen) is a web available resource integrating data from 6,792 genomes/exome sequencing experiments from 28 cancer types (release 2014.12) [63]. It collects and analyses somatic mutations in a large set of tumor samples to identify putative cancer driver genes. The prediction of putative cancer driver genes is performed by OncodriveFM [64], which has been developed by member of the same group (see next section). The web interface allows browsing data by gene name, cancer type, sites and projects, reporting the frequencies of mutation at gene and genomic location InOGen also allows the study of genomic alterations in cancer in the contest of pathways [65].

The URLs of the cancer genome data repositories, the cancer gene lists and resources for cancer genome analysis are summarized in Table 2.

#### **Computational methods for cancer genome interpretation**

The analysis of the cancer genome is a challenging task from both the experimental and computational point of views. A recent review, in addition to providing an exhaustive overview of the available tools for the detection of somatic mutation from sequencing data, also highlighted these challenges [11]. The lack of a perfect pipeline for the detection of genetic variants and the



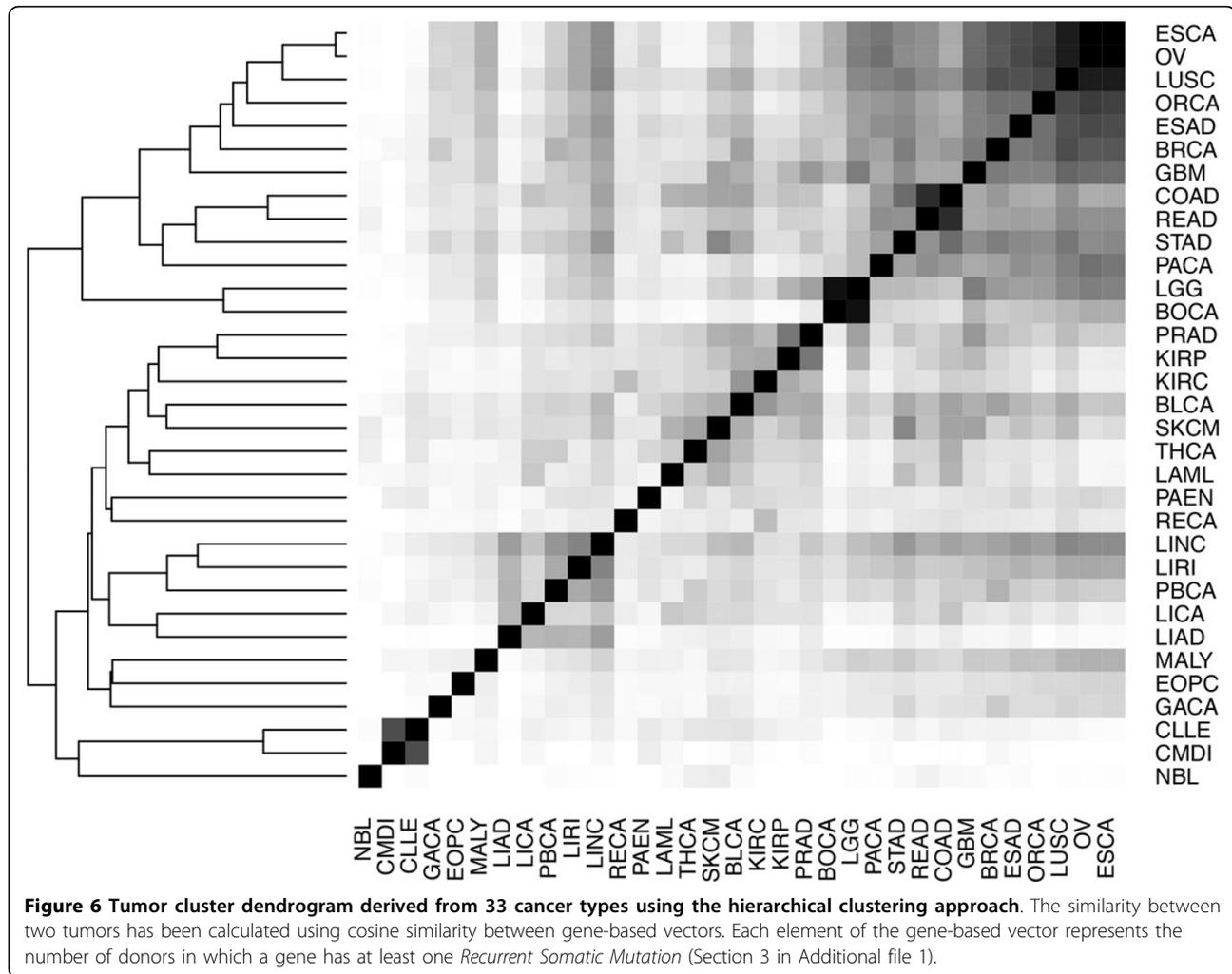
computational analysis of the data remains one of the major bottlenecks in the field. The current computational methods for the interpretation of variants in cancer genome have been developed to address the following issues: i) Detection of recurrent somatic mutations and cancer driver genes; ii) Prediction of driver variants and their functional impact; iii) Estimate the impact of multiple variants at network and pathway level; iv) Differentiate subclonal populations and their variation patterns. In this section we describe a selection of available computational tools for addressing the four issues listed above.

#### Detection of cancer driver genes

In general, the insurgence of cancer is accompanied with an accumulation of somatic mutations. However, this does not imply that all mutations are of equal importance in the oncogenesis and cancer progression. Rather, driver somatic mutations exert a selective advantage to cancer cells. In contrast, passenger mutations are

abundant but do not confer any selective advantage. Driver variants are present in a small fraction that may vary in different cancer types [66,67]. Thus, an obvious approach for the selection of cancer driver genes involves the analysis of recurrent somatic mutations. Following this idea, several methods - such as MuSiC [68], MutSigCV [69] DrGaP [70] and Simon's tool [71] - have been developed to prioritize cancer driver genes using different statistical models for the estimation of a background mutation rate.

The Mutational Significance in Cancer (MuSiC) uses sequence-based input to establish correlations among mutation sites, affected genes and pathways, and to discriminate abundant passenger mutations from significant mutation events. This method aims to identify significantly mutated genes with respect to a background mutation rate. MuSiC has been tested on a set of 316 ovarian cancer samples and it was able to detect 12 significantly



mutated genes [68]. MutSigCV is the newest version of the “Mutation Significance” algorithm that uses gene specific background mutation rates including mutation events in gene covariates. The information from known co-varying genes is important for estimating the background mutation rates in the genomic regions where few mutation events are observed. MutSigCV takes an input list of mutation from different samples and builds a model gene-dependent background mutation rate estimated using clusters of genes. MutSigCV has been used to analyze 3,083 tumor-normal pairs, and it was able to discover strong differences in mutation frequency and spectrum across 27 cancer types, providing interesting insights about the etiology of the disease [69]. DrGap [70] is a statistical framework for identifying driver genes and signaling pathways in cancer genome-sequencing studies. This approach integrates biological knowledge of the mutational process in tumors and uses a heuristic strategy to optimize the mixture proportion of chi-square distribution of likelihood ratio test (LRT) statistics. This

approach improves the accuracy and sensitivity of the prediction of driver genes avoiding zero estimation of the driver mutation rates due to the small probability of observing any mutation in the available samples. Simon’s tool [71] calculates the background mutation rate by accounting for the functional impact of mutations on proteins, variation in background mutation rate among tumors, and the redundancy of the genetic code. Using this algorithm, the authors reanalyzed 623 candidate genes in 188 non-small cell lung tumors and identified 28 driver genes, 6 of which were novel [72].

Alternative approaches have been implemented in ActiveDriver [73], OncodriveFM [64], OncodriveCLUST [74] and ContrastRank [75]. ActiveDriver is able to detect significantly mutated functional sites in cancer genome providing an analysis of the somatic mutations associate to phosphorylation signaling. Indeed, the predictive model has been calculated considering the frequency of somatic variants from ~800 cancer genomes in proximity of ~74,000 phosphorylation sites and 469 kinase domains

[73]. The method was able to identify candidate genes, protein complexes and kinases involved in cancer. OncodriveFM [64] is a method for the detection of putative cancer driver genes or gene modules. It computes the functional impact of variants using three established computational approaches (SIFT [76], PolyPhen2 [77] and MutationAssessor [78]). OncodriveFM prioritizes putative cancer driver genes calculating the distribution of the predicted functional impact scores for several variants across tumor samples and its deviation from the null model. The bias towards the accumulation of variants with high functional impact is used to detect candidate driver genes. Similar approach has been used to prioritize cancer-associated pathways. OncodriveCLUST has been developed to identify significant bias towards somatic mutations clustered within the same protein. The background model for prioritizing genes is calculated using the rate on silent mutations, which are not assumed to be under positive selection. OncodriveCLUST cluster scores have been calculated using ~240,000 mutations in ~9,500 genes from COSMIC database and has been tested on four TCGA datasets. This approach provides variable background mutation rate for each gene and it detects recessive cancer genes not identified using OncodriveFM [74].

Finally, we also highlight ContrastRank [75], which is a new method for the prioritization of putative impaired genes in cancer. With respect to previously developed methods, ContrastRank evaluates the background mutation rate using the maximum value between the mutation rates of each gene in 1000 Genomes [79] and normal TCGA samples. For the calculation of the score the method only considers mutation with allele frequency in 1000 Genomes lower than 0.5% and uses the complement of the cumulative binomial distribution to rank cancer-associated genes. The method has been tested on TCGA from three types of adenocarcinomas. In addition, the Cancer Census, Vogelstein and Bushman cancer gene lists (see Table 2) have been used to assess the quality of the prioritization method. ContrastRank was also used for calculating an exome-based score to discriminate between TCGA normal and tumor TCGA samples resulting in high level of accuracy.

#### **Predicting the impact of non-synonymous variants in cancer**

During the last decades several methods have been developed to predict the impact of non-synonymous single nucleotide variants (nsSNVs) at structural [80] and functional levels [2]. In particular, the algorithms for predicting the functional effect of missense variants estimate the probability that a mutation is disease-associated or functionally deleterious. Although the relationships between molecular state and disease are complex and are far from being completely understood, the pathologic effect

resulting from amino acid substitution is commonly estimated by predicting its functional impact. Most of the available algorithms are based on the evidence that functionally important protein sites are under purifying selection [81]. Therefore, the majority of disease-causing variants should occur in conserved regions that can be detected by evolutionary analysis [82]. Using this approach many classifiers have been implemented to predict whether a nsSNV has any functional impact. The most famous methods for estimating the impact of genetic variants are SIFT [76] and PolyPhen [77] whose predictions are already embedded in many variant annotation pipelines. Recently, more sophisticated methods, which exploit additional structural and functional information, have been developed [19,83-85]. Recent advances in the field are represented by the implementation of consensus-based algorithms [86,87] and a general approach that is able to predict the impact of variants in non-coding regions [81,88].

Although most variant effect predictor reached an adequate level of accuracy, their predictions do not provide information about the possible phenotypic effect. This problem has been partially addressed with the development of a new class of disease-specific predictors trained on a subset of mutations with defined phenotypic effect. In particular, several methods have been developed for discriminating between passenger and cancer driver mutations [78,89-93]. Among them, CanPredict [91] has been the first algorithm for predicting cancer-causing mutations. The method uses a conservation measure from PFAM [94] Hidden Markov Models and functional information encoded through Gene Ontology terms. Similar strategy has been implemented in support vector-based approach (Dr. Cancer) that uses sequence profile and cancer-specific functional terms [89]. Dr. Cancer has been tested on 3,163 cancer-driver mutations from 74 proteins. The results show that cancer-related proteins are enriched for particular Gene Ontology terms that can be used to discriminate between cancer and other phenotypes. The same dataset was previously used to train CHASM a random forest-based algorithm [90]. CHASM was tested on ~600 missense mutations in glioblastoma multiforme achieving better performance than PolyPhen and previously developed cancer specific approaches. More recently, new methods such as MutationAssessor [78], TransFIC [93] and FATHMM [92] have been optimized to detect cancer driver variants. MutationAssessor uses evolutionary information patterns calculating an entropy-based functional impact score from homolog and paralog proteins. The method trained on ~19,000 disease-associated variants has been tested on 10,000 mutations in COSMIC database prioritized according to their recurrence and multiplicity. The authors estimated that ~5% of the cancer-relevant mutation involves a

change in the protein function rather than standard loss and gain of function mechanism events. In TransFIC approach the prediction of cancer drivers mutation has been tuned considering subset of variants associated to the same Gene Ontology term and selecting a variable threshold for discriminating driver from passenger mutations. The performance of the optimized functional impact score has been tested on subset of the COSMIC mutations classified according to their recurrence. The results show that groups of protein with different function, possess distinct baseline tolerance to deleterious mutations. Finally, FATHMM is a Hidden Markov Model based algorithm using protein domain information that has been optimized for predicting cancer causing mutations. The method, tested on previously collected datasets result in improved performances with respect to previously developed methods. Although the algorithms based on functional information (Gene Ontology terms/ protein domains) achieve better results than standard conservation-based approaches, a fair testing procedure is more difficult because the predictions can be biased toward more abundant functional classes.

#### **Analysis of cancer gene pathways and networks**

The accumulation of somatic mutations during the lifespan is the main cause of cancer. Several identified somatic mutations occur in genes involved in many signaling, regulatory and metabolic pathways. Indeed mutated genes such as TP53 and PI3KCA are hubs in pathways and interaction networks which control cell proliferation, growth and apoptosis [95,96]. In addition, recent sequencing studies [4,38] revealed that cancer driver genes tend to cluster within a limited number of essential pathways, and rarely mutations on genes in the same pathway co-occur in the same patient. This mutually exclusive genomic events have been observed, for example, in the case of *BRAF* and *KRAS* (involved in RAS/RAF pathway) in colorectal cancer [97], *APC* and *CTNNB1* (involved in beta catenin pathway) [98], *EGFR* and *KRAS* in lung adenocarcinomas [99], *TP53* and *MDM2* in many different cancer types [100].

From these observations, it is evident that the analysis of genomic variations across gene pathways and networks is important to understand the combinatorial effect of the mutations and explain the disease mechanism. Pathway and network analysis can be performed using previously annotated gene pathways or testing alternative routes from gene interaction networks. Reference databases collecting information about gene pathways are the Kyoto Encyclopedia of Genes and Genomes Kyoto Encyclopedia of Genes and Genomes (KEGG) [65] and the Molecular Signatures Database (MSigDB) [101]. Biological interaction networks mainly consist of protein-protein interaction data, which are collected in databases such as IntAct [102] and iRefIndex [103]. Another important database is the

Reactome, which aggregates data on protein-protein interaction networks and gene pathways [104].

Recently, several methods for the analysis of cancer gene pathway have been developed [105-115]. Few examples are PathScan [114], which is an annotation-based approach, and HotNet2 [115], MDPFinder [107], MEMo [105], and Dendrix [113] which are able to identify driver pathways. PathScan [114] is a probabilistic model that takes into account the length of the genes and differences in their mutation rates under the null hypothesis. The method combines single-sample p-values using an integral approximation that estimates a pathway-specific overall p-value. PathScan represents an alternative to previous approaches, which reduce a gene set into a unique gene simply combining the total mutations and mutable positions of each gene. PathScan was applied to the analysis of data from lung adenocarcinoma sequencing project to find significantly mutated 129 KEGG pathways. The results show that PathScan was able to identify significant pathways detected in previous studies. In addition PathScan found the focal adhesion pathway to be significant in agreement with previous expression studies on prostate and ovarian cancers [114]. HotNet2 [115], has been recently developed for identifying significantly mutated groups of interacting genes from large cancer sequencing studies. It uses a heat diffusion model that encodes both the topology of the network and the significance of the observed frequencies of each mutated gene. HotNet2 has been tested on 3,281 samples from 12 cancer types in TCGA studies. In this analysis it was able to identify 16 significantly mutated subnetworks that comprise well-known cancer signaling pathways. Among them, the well-known TP53, phosphoinositide 3-kinase (PI3K), NOTCH and receptor tyrosine kinases (RTK) signaling pathways.

Although annotation-based methods are able to prioritize important cancer pathways, they present some limitations. The main limitations are the incompleteness of pathway databases and the presence of multiple overlapping gene sets across pathways. Thus, more general methods for detecting significantly mutated gene sets use information from biological interaction networks. The Mutual Exclusivity Modules in Cancer (MEMo) algorithm [105] integrates copy number alteration and mutation data and maps them into biological networks. The method uses correlation analysis and statistical tests to identify network modules of genes recurrently altered across a set of tumor samples, participate in the same biological process, and alteration events are mutually exclusive. MEMo has been tested on a set of ~400 samples from glioblastoma and ovarian cancer. In the first cancer study the method was able to identify important signaling modules such as p53 and PI(3)K. In ovarian cancer, it detected mutually exclusive variation events between BRCA and genes in the Rb module.

The De novo Driver Exclusivity (Dendrix) algorithm [113] is a tool for discovery of mutated driver pathways in cancer using only mutation information from the cancer samples. The method introduces the concepts of coverage and exclusivity to distinguish group of genes with driver mutations from set of genes with passenger mutations. The Dendrix algorithm, which has been applied to the analysis of different cancer types in TCGA study, was able to identify 8 mutation groups mutated in 94% of patients from 17 cancer types and groups of mutually exclusive genes in lung adenocarcinoma and glioblastoma multiforme. Similarly, MDPFinder [107] identifies de novo mutated driver pathways from mutation data solving the maximum weight submatrix problem. The authors implemented an exact method based on binary linear programming and genetic algorithm to combine mutation and expression data. MDPFinder has been tested on a set of ~500 samples from head and neck squamous cell carcinomas glioblastoma and ovarian carcinoma. The results show that the integrative model based on mutation and expression data was able to identify biologically relevant gene sets detected in previous studies [107].

Finally, PARADIGM [116] estimates patient-specific genetic activities incorporating curated gene interactions from the NCI pathway interaction database (PID) [117]. The method, that can use many types of omics data, represents a gene by a factor graph with associated known activity and expression. PARADIGM, tested on breast cancer and glioblastoma samples, was able to identify altered activities in cancer-related pathways with less false-positives than a previous developed method.

An interesting approach, which applies network analysis to the study of the cancer genome, is the Network-Based Stratification (NBS) method [118]. NBS is an algorithm for the classification of cancer subtypes that clusters patients with somatic mutations in similar gene network regions. More in details, the NBS represents each patient with a profile of somatic mutated genes mapped on the human gene interaction network. After a smoothing procedure, the patients are clustered using a non-negative matrix factorization-based approach. The method performs a second clustering procedure to find subgroups of patients frequently co-stratified after random sampling. NBS has been tested on a set of ~1,000 samples from ovarian, uterine endometrial and lung tumors and used to determine relationship between cancer subtypes and patient survival.

#### **Classification of cancer subclonal variants**

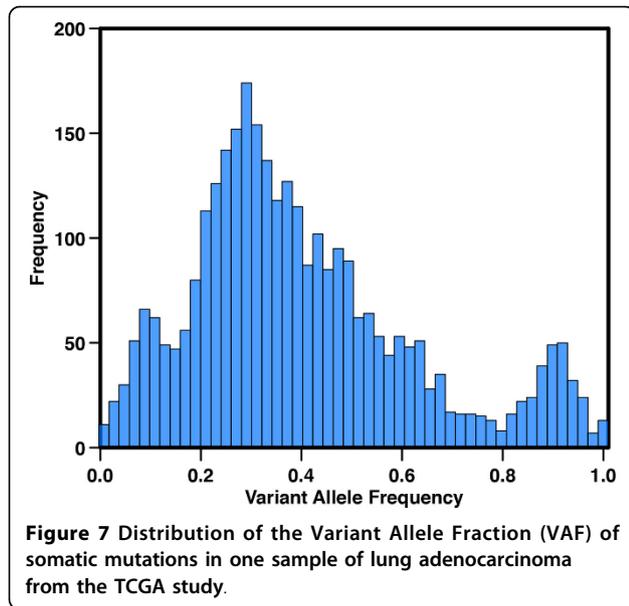
Tumor progression is an evolutionary process which starts from a single cell and results in the selection of more aggressive subclones [119]. The presence of different cell populations affects the accurate detection of somatic mutations in cancer sample. Although all subclones in a sample have the same origin, they are present in different proportions and have heterogeneous patterns of somatic

mutations. Thus, somatic mutations in cancer cells present in small proportion are more difficult to detect because correspond to variations supported by a low number of reads, further reduced by impurity in the tumor sample. In a recent work [13], Ding and colleagues estimated that approximately 340X coverage is needed for detecting (with 99% of chance) at least 3 reads when a heterozygous variant is present in a 5% subclonal population.

A possible solution to this issue lies in single-cell sequencing for revealing diversities in the pattern of somatic mutations within a tumor [120,121]. Although this technique provides a better characterization of the mutations occurring in each subclonal population, the approach is affected by errors introduced during the amplification process. Furthermore, single-cell experiments need to be performed on multiple tumor cells to have a general overview of the mutation pattern. The presence of subclones in the cell population can be detected analyzing the distribution of reads or variant allele fraction for the somatic mutations. If the distribution is multimodal, the presence of subclonal populations is expected. Figure 7 shows the distribution of variant allele fraction (VAF) in a lung adenocarcinoma cancer sample from TCGA. Although the distribution of variant allele fraction provides an indication about the clonal architecture, the analysis of this data is affected by sample impurity and copy number alterations. In the last few years many computational tools have been developed to estimate the presence of subclonal populations [122-127], some of them are discussed below.

ABSOLUTE [122] uses data from copy number variations for optimizing models of recurrent cancer karyotypes and expected allelic fraction values for somatic SNVs. These models are then used to re-extract information about the absolute cellular copy number of local DNA segments and the number of mutated alleles for somatic SNVs. The ABSOLUTE algorithm has been applied to the analysis of copy-number profiles from 3,155 cancer samples, identifying recurrent genome doubling events that influence tumor progression. ABSOLUTE was also used for the analysis of exome sequencing data from 214 ovarian carcinoma tumor-normal pairs. The method was able to identify large subclone populations with predominant somatic mutations and small subset of subclones with heterozygotes mutations in tumor suppressor genes TP53 and NF1 and in a candidate tumor suppressor gene CDK12.

Alternative methods [124,125] use phylogenetic approaches to study the evolution of tumor cells. For example, PyClone [125] assumes that clonal population follows a perfect and persistent phylogeny. Under this assumption, for which each somatic mutation only happen once in the evolutionary history and not reverse mutation are allowed, subclonal cells can be identified and their



prevalence estimated. PyClone uses a Bayesian clustering algorithm to group sets of somatic mutations belonging to the same cluster of subclonal cell accounting for allelic imbalances introduced by copy-number alterations and normal-cell impurity. The method has been tested on a simulated dataset produced from mixtures of DNA extracted from four 1000 Genomes Project sample and four spatially separated samples from a primary, untreated high-grade ovarian tumor. The results show that PyClone outperforms two genotype-naïve methods using binomial and infinite beta-binomial mixture models. Similarly, PhyloSub [124] infers the phylogeny and genotype of the major lineages in the clonal population calculating the Bayesian prior over the trees clustering the SNVs. A sampling procedure is used to find the optimized joint posterior distribution with higher probability to generate the observed frequencies of somatic mutations. PhyloSub has been tested using deep exome sequencing data from acute myeloid leukemia and chronic lymphocytic leukemia. The results show that PhyloSub is able to identify both linear and branching subclonal lineages.

A recent publication [123] presents a novel combinatorial approach based on binary partition tree (BPT) to model the mechanism of clonal expansion in tumor and estimate the subpopulations of tumor using the variant allele frequencies of somatic mutations. The authors demonstrated that finding a BPT is a NP-complete problem, and derived an approximation algorithm for an optimized version of the problem. Finally they implemented a recursive approach that finds the solution of the optimized BPT problem in a polynomial time. The developed algorithm can detect errors in the estimation of the variant allele frequencies of somatic mutations, which cannot be

correctly estimated because of the admixture of normal cells in the tumor sample. The performance of BPT algorithm has been tested on simulated and real cancer data showing it generates more consistent results, and it is faster than previously developed methods.

The methods for the detection of genetic heterogeneity in cancer can be also used to detect subclonal mutation conferring drug resistance. This idea has been investigated using cloneHD [126], a new subclonal reconstruction algorithm optimized using both information about somatic mutations and correlated changes generated by copy-number changes. The method has been applied to the analysis of sequencing data from time-resolved samples from breast cancer and of chronic lymphocytic leukemia. The results demonstrate that cloneHD can be a valuable tool for tracking cancer development and monitoring the response of a patient to therapy regimens.

The URLs of the tools describe above for the analysis of the cancer genome are summarized in Table 3.

#### Concluding remarks and future perspectives

In this review we provide an overview of the challenging topics in the analysis of cancer genome. We mainly focused on the characterization of single nucleotide variant, which is by far the most common type of genetic variation. In particular, we provided a summary of the most important cancer genome data available online and described a selection of the available computational tools for cancer genome interpretation.

Although several algorithms have been developed, the problem of cancer genome interpretation is far from being solved. The progress in the field is limited by many factors mainly associated with (a) the intrinsic complexity of the problem (b) technical limitations, and (c) ethical issues.

The complexity of the problem mainly comes from the huge number of somatic variations present in the in each tumor sample and our inability to select driver and clinically actionable variants. The technical limitations are also affecting the detection of genetic variations present in a smaller fraction of subclones. The third important issue is the restricted access to the data for protecting the privacy of the individual.

It is expected that in the near future most of the limitations will be overcome by the development of more accurate computational tools and experimental approaches which will play an important role in the understanding the relationship between genotype and phenotype in cancer. In particular, brute force sequencing initiatives will result in a better mapping of the functionally important regions in the genome and experimental approaches, such as CRISPR/Cas [128] will provide the opportunity to extensively test the functional impact of genetic

**Table 3. Computational methods for cancer genome interpretation**

Method	URL	Ref
<i>Cancer gene prioritization</i>		
ActiveDriver	<a href="http://individual.utoronto.ca/reimand/ActiveDriver/">http://individual.utoronto.ca/reimand/ActiveDriver/</a>	[73]
ContrastRank	<a href="http://snps.biofold.org/contrastrank/">http://snps.biofold.org/contrastrank/</a>	[75]
DrGaP	<a href="https://code.google.com/p/drgap/">https://code.google.com/p/drgap/</a>	[70]
MuSiC	<a href="http://gmt.genome.wustl.edu/packages/genome-music/">http://gmt.genome.wustl.edu/packages/genome-music/</a>	[68]
MuSigCV	<a href="http://www.broadinstitute.org/cancer/cga/mutsig">http://www.broadinstitute.org/cancer/cga/mutsig</a>	[69]
OncodriveCLUST	<a href="http://bg.upf.edu/oncodriveclust">http://bg.upf.edu/oncodriveclust</a>	[74]
OncodriveFM	<a href="http://bg.upf.edu/oncodrivefm">http://bg.upf.edu/oncodrivefm</a>	[64]
Simon's tool	<a href="http://linus.nci.nih.gov/Data/YounA/software.zip">http://linus.nci.nih.gov/Data/YounA/software.zip</a>	[71]
<i>Cancer variant annotation tools</i>		
CanPredict	<a href="http://goo.gl/UK9lbv">http://goo.gl/UK9lbv</a>	[91]
CHASM	<a href="http://wiki.chasmssoftware.org/">http://wiki.chasmssoftware.org/</a>	[90]
DrCancer	<a href="http://snps.biofold.org/drcancer/">http://snps.biofold.org/drcancer/</a>	[89]
FATHMM	<a href="http://fathmm.biocompute.org.uk/cancer.html">http://fathmm.biocompute.org.uk/cancer.html</a>	[92]
MutationAssessor	<a href="http://mutationassessor.org">http://mutationassessor.org</a>	[78]
TransFIC	<a href="http://bg.upf.edu/transfic/">http://bg.upf.edu/transfic/</a>	[93]
<i>Pathway and network analysis</i>		
Dendrix	<a href="http://compbio.cs.brown.edu/projects/dendrix/">http://compbio.cs.brown.edu/projects/dendrix/</a>	[113]
HotNet2	<a href="http://compbio.cs.brown.edu/projects/hotnet2/">http://compbio.cs.brown.edu/projects/hotnet2/</a>	[115]
MDPFinder	<a href="http://zhangroup.aporc.org/ShiHuaZhang">http://zhangroup.aporc.org/ShiHuaZhang</a>	[107]
MEMo	<a href="http://cbio.mskcc.org/memo">http://cbio.mskcc.org/memo</a>	[105]
PathScan	<a href="http://genome.wustl.edu/software/pathscan">http://genome.wustl.edu/software/pathscan</a>	[114]
PARADIGM	<a href="http://sbenz.github.com/Paradigm">http://sbenz.github.com/Paradigm</a>	[116]
<i>Classification of tumor subclonal variants</i>		
ABSOLUTE	<a href="http://www.broadinstitute.org/cancer/cga/absolute">http://www.broadinstitute.org/cancer/cga/absolute</a>	[122]
BTP	<a href="http://compbio.cs.brown.edu/projects/btp/">http://compbio.cs.brown.edu/projects/btp/</a>	[123]
CloneHD	<a href="https://github.com/andrej-fischer/cloneHD/">https://github.com/andrej-fischer/cloneHD/</a>	[126]
PhyloSub	<a href="https://github.com/morrislab/phylosub/">https://github.com/morrislab/phylosub/</a>	[124]
PyClones	<a href="http://compbio.bccrc.ca/software/pyclone/">http://compbio.bccrc.ca/software/pyclone/</a>	[125]

variants. In addition, an improvement of the single cell sequencing technology will allow better characterization of the progression of tumor and definition of the pattern of mutations in more aggressive subclonal cells.

The integration of more accurate data will have an impact on the development of more accurate computational tools. Indeed, the limited ability to score the performances of currently available algorithms can be addressed by collecting standardized benchmark sets from high-quality experiments. An important component for speeding up this learning process involves the implementation of more effective policies for data sharing. Although major efforts have been made by the TCGA, ICGC and other cancer consortium, the procedures for releasing sequencing data need to be optimized. Standard protocols for extracting and reporting the data are also required for efficient data analysis. Information about the germline variants present in sequenced patients should be made available for validating hypothesis about the tumor predisposition. Furthermore, it will be extremely important to develop

appropriate de-identification procedures [129] and the promotion of informed consent policies for improving the effective usage of genotype/phenotype databases [130,131].

### Additional material

**Additional file 1: Supplementary materials including supplementary tables.**

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

RT and EC carried out the analysis of ICGC data. RT, MKB and EC wrote the manuscript and approved its final version.

### Acknowledgements

The authors acknowledge The Cancer Genome Atlas Consortium for allowing access to the restricted whole-exome sequencing data for Colon and Lung Adenocarcinomas. We also thank the International Cancer Genome Consortium for making available the somatic mutation data from several cancer projects. We would like to thank two anonymous reviewers

for their fruitful comments, which helped to improve the quality of this review.

EC and MKB are supported by start-up funds from the Department of Pathology at the University of Alabama, Birmingham. MKB is also supported by additional funding from Health Science Foundation, University of Alabama, Birmingham. The computational infrastructure for cancer genomics research is provided by the Genome Informatics core at UAB (<http://genfix.ifx.uab.edu/>).

#### Declarations

Publication costs for this article were funded by EC's start-up funds from the Department of Pathology at the University of Alabama, Birmingham. This article has been published as part of *BMC Genomics* Volume 16 Supplement 8, 2015: VarL-SIG 2014: Identification and annotation of genetic variants in the context of structure, function and disease. The full contents of the supplement are available online at <http://www.biomedcentral.com/bmcgenomics/supplements/16/S8>.

#### Authors' details

<sup>1</sup>Division of Informatics, Department of Pathology, University of Alabama at Birmingham, 619 19<sup>th</sup> St. South, 35249 Birmingham (AL), USA. <sup>2</sup>Department of Clinical and Diagnostic Sciences, University of Alabama at Birmingham, 619 19<sup>th</sup> St. South, 35249 Birmingham (AL), USA. <sup>3</sup>Department of Biomedical Engineering, University of Alabama at Birmingham, 619 19<sup>th</sup> St. South, 35249 Birmingham (AL), USA.

Published: 18 June 2015

#### References

- 1000 Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491**(7422):56-65.
- Capriotti E, Nehrt NL, Kann MG, Bromberg Y: **Bioinformatics for personal genome interpretation.** *Brief Bioinform* 2012, **13**(4):495-512.
- Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery.** *Nature reviews Genetics* 2011, **12**(11):745-755.
- Stratton MR, Campbell PJ, Futreal PA: **The cancer genome.** *Nature* 2009, **458**(7239):719-724.
- Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al: **Signatures of mutational processes in human cancer.** *Nature* 2013, **500**(7463):415-421.
- Pleasant ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordóñez GR, Bignell GR, et al: **A comprehensive catalogue of somatic mutations from a human cancer genome.** *Nature* 2010, **463**(7278):191-196.
- Ciriello G, Miller ML, Aksoy BA, Senbabaoğlu Y, Schultz N, Sander C: **Emerging landscape of oncogenic signatures across human cancers.** *Nature genetics* 2013, **45**(10):1127-1133.
- Garraway LA, Lander ES: **Lessons from the cancer genome.** *Cell* 2013, **153**(1):17-37.
- Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW: **Cancer genome landscapes.** *Science* 2013, **339**(6127):1546-1558.
- Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nature genetics* 2013, **45**(10):1113-1120.
- Raphael BJ, Dobson JR, Oesper L, Vandin F: **Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine.** *Genome medicine* 2014, **6**(1):5.
- Fernald GH, Capriotti E, Daneshjouri R, Karczewski KJ, Altman RB: **Bioinformatics challenges for personalized medicine.** *Bioinformatics* 2011, **27**(13):1741-1748.
- Ding L, Wendl MC, McMichael JF, Raphael BJ: **Expanding the computational toolbox for mining cancer genomes.** *Nature reviews Genetics* 2014, **15**(8):556-570.
- Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, et al: **Computational approaches to identify functional genetic variants in cancer genomes.** *Nature methods* 2013, **10**(8):723-729.
- Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efreanova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z: **A survey of tools for variant analysis of next-generation genome sequencing data.** *Briefings in bioinformatics* 2014, **15**(2):256-278.
- Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, Dahlman KB, Pao W, Zhao Z: **Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers.** *Genome medicine* 2013, **5**(10):91.
- Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754-1760.
- Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome biology* 2009, **10**(3):R25.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078-2079.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome research* 2010, **20**(9):1297-1303.
- Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.** *Genome research* 2012, **22**(3):568-576.
- Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: **Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.** *Nature biotechnology* 2013, **31**(3):213-219.
- Bansal V: **A statistical method for the detection of variants from next-generation resequencing of DNA pools.** *Bioinformatics* 2010, **26**(12):i318-324.
- Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H: **SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data.** *Nucleic acids research* 2011, **39**(19):e132.
- Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, et al: **An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data.** *Nucleic acids research* 2013, **41**(7):e89.
- Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A, et al: **JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data.** *Bioinformatics* 2012, **28**(7):907-913.
- Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L: **SomaticSniper: identification of somatic point mutations in whole genome sequencing data.** *Bioinformatics* 2012, **28**(3):311-317.
- Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK: **Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs.** *Bioinformatics* 2012, **28**(14):1811-1817.
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29**(1):308-311.
- Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X: **Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift.** *Frontiers in genetics* 2012, **3**:35.
- Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic acids research* 2010, **38**(16):e164.
- Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3.** *Fly* 2012, **6**(2):80-92.
- McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.** *Bioinformatics* 2010, **26**(16):2069-2070.
- Wilks C, Cline MS, Weiler E, Diehkans M, Craft B, Martin C, Murphy D, Pierce H, Black J, Nelson D, et al: **The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data.** *Database: the journal of biological databases and curation* 2014, **2014**.
- International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, et al:

- International network of cancer genome projects. *Nature* 2010, **464**(7291):993-998.
36. Watson IR, Takahashi K, Futreal PA, Chin L: **Emerging patterns of somatic mutations in cancer.** *Nature reviews Genetics* 2013, **14**(10):703-718.
37. Wheeler DA, Wang L: **From human genome to cancer genome: the first decade.** *Genome research* 2013, **23**(7):1054-1062.
38. Cancer Genome Atlas Research N: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455**(7216):1061-1068.
39. Cancer Genome Atlas N: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487**(7407):330-337.
40. Cancer Genome Atlas N: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490**(7418):61-70.
41. Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, Johns AL, Miller DK, Wilson PJ, Patch AM, Wu J, et al: **Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes.** *Nature* 2012, **491**(7424):399-405.
42. Cancer Genome Atlas Research N: **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474**(7353):609-615.
43. Cancer Genome Atlas Research N: **Comprehensive genomic characterization of squamous cell lung cancers.** *Nature* 2012, **489**(7417):519-525.
44. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR, et al: **The landscape of cancer genes and mutational processes in breast cancer.** *Nature* 2012, **486**(7403):400-404.
45. Sadelain M, Papapetrou EP, Bushman FD: **Safe harbours for the integration of new DNA in the human genome.** *Nature reviews Cancer* 2012, **12**(1):51-58.
46. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G: **Discovery and saturation analysis of cancer genes across 21 tumour types.** *Nature* 2014, **505**(7484):495-501.
47. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, Elledge SJ: **Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome.** *Cell* 2013, **155**(4):948-962.
48. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al: **Mutational landscape and significance across 12 major cancer types.** *Nature* 2013, **502**(7471):333-339.
49. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, et al: **Comprehensive identification of mutational cancer driver genes across 12 tumor types.** *Scientific reports* 2013, **3**:2650.
50. An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD: **NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes.** *Database : the journal of biological databases and curation* 2014, **2014**:bau015.
51. Porta-Pardo E, Hrabe T, Godzik A: **Cancer3D: understanding cancer mutations through protein structures.** *Nucleic acids research* 2014.
52. Capriotti E, Altman RB: **Improving the prediction of disease-related variants using protein three-dimensional structure.** *BMC Bioinformatics* 2011, **12**(Suppl 4):S3.
53. Lahti JL, Tang GW, Capriotti E, Liu T, Altman RB: **Bioinformatics and variability in drug response: a protein structural perspective.** *J R Soc Interface* 2012, **9**(72):1409-1437.
54. Li B, Seligman C, Thusberg J, Miller JL, Auer J, Whirl-Carrillo M, Capriotti E, Klein TE, Mooney SD: **In silico comparative characterization of pharmacogenomic missense variants.** *BMC Genomics* 2014, **15**(Suppl 4):S4.
55. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlic A, Quesada M, et al: **The RCSB Protein Data Bank: new resources for research and education.** *Nucleic Acids Res* 2013, **41**(Database):D475-482.
56. Porta-Pardo E, Godzik A: **e-Driver: a novel method to identify protein regions driving cancer.** *Bioinformatics* 2014, **30**(21):3109-3114.
57. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.** *Cancer discovery* 2012, **2**(5):401-404.
58. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al: **Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.** *Science signaling* 2013, **6**(269):p11.
59. Goldman M, Craft B, Swatloski T, Ellrott K, Cline M, Diekhans M, Ma S, Wilks C, Stuart J, Haussler D, et al: **The UCSC Cancer Genomics Browser: update 2013.** *Nucleic acids research* 2013, **41**(Database):D949-954.
60. Park YK, Kang TW, Baek SJ, Kim KI, Kim SY, Lee D, Kim YS: **CaGe: A Web-Based Cancer Gene Annotation System for Cancer Genomics.** *Genomics & informatics* 2012, **10**(1):33-39.
61. Cheng WC, Chung IF, Chen CY, Sun HJ, Fen JJ, Tang WC, Chang TY, Wong TT, Wang HW: **DriverDB: an exome sequencing database for cancer driver gene identification.** *Nucleic acids research* 2014, **42**(Database):D1048-1054.
62. Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, Furney SJ, Lopez-Bigas N: **IntOGen: integration and data mining of multidimensional oncogenomic data.** *Nature methods* 2010, **7**(2):92-93.
63. Rubio-Perez C, Tamborero D, Schroeder MP, Antolin AA, Deu-Pons J, Perez-Llamas C, Mestres J, Gonzalez-Perez A, Lopez-Bigas N: **In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities.** *Cancer cell* 2015, **27**(3):382-396.
64. Gonzalez-Perez A, Lopez-Bigas N: **Functional impact bias reveals cancer drivers.** *Nucleic acids research* 2012, **40**(21):e169.
65. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic acids research* 2014, **42**(Database):D199-205.
66. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D: **An integrated approach to uncover drivers of cancer.** *Cell* 2010, **143**(6):1005-1017.
67. Greenman C, Stephens P, Smith R, Dalgleish GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446**(7132):153-158.
68. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, et al: **MuSiC: identifying mutational significance in cancer genomes.** *Genome research* 2012, **22**(8):1589-1598.
69. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al: **Mutational heterogeneity in cancer and the search for new cancer-associated genes.** *Nature* 2013, **499**(7457):214-218.
70. 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.** *American journal of human genetics* 2013, **93**(3):439-451.
71. Youn A, Simon R: **Identifying cancer driver genes in tumor genome sequencing studies.** *Bioinformatics* 2011, **27**(2):175-181.
72. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, et al: **Somatic mutations affect key pathways in lung adenocarcinoma.** *Nature* 2008, **455**(7216):1069-1075.
73. Reimand J, Bader GD: **Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.** *Molecular systems biology* 2013, **9**:637.
74. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N: **OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes.** *Bioinformatics* 2013, **29**(18):2238-2244.
75. Tian R, Basu MK, Capriotti E: **ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples.** *Bioinformatics* 2014, **30**(17):i572-i578.
76. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31**(13):3812-3814.
77. 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.** *Nature methods* 2010, **7**(4):248-249.
78. Reva B, Antipin Y, Sander C: **Determinants of protein function revealed by combinatorial entropy optimization.** *Genome biology* 2007, **8**(11):R232.
79. 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**(7319):1061-1073.
80. Compiani M, Capriotti E: **Computational and theoretical methods for protein folding.** *Biochemistry* 2013, **52**(48):8601-8624.
81. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants.** *Nature genetics* 2014, **46**(3):310-315.

82. Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA: **Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans.** *Hum Mutat* 2008, **29**(1):198-204.
83. Capriotti E, Calabrese R, Casadio R: **Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information.** *Bioinformatics* 2006, **22**(22):2729-2734.
84. Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R: **WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation.** *BMC Genomics* 2013, **14**(Suppl 3):S6.
85. Bromberg Y, Yachdav G, Rost B: **SNAP predicts effect of mutations on protein function.** *Bioinformatics* 2008, **24**(20):2397-2398.
86. Capriotti E, Altman RB, Bromberg Y: **Collective judgment predicts disease-associated single nucleotide variants.** *BMC Genomics* 2013, **14**(Suppl 3):S2.
87. Gonzalez-Perez A, Lopez-Bigas N: **Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.** *American journal of human genetics* 2011, **88**(4):440-449.
88. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, Campbell C: **An Integrative Approach to Predicting the Functional Effects of Non-Coding and Coding Sequence Variation.** *Bioinformatics* 2015.
89. Capriotti E, Altman RB: **A new disease-specific machine learning approach for the prediction of cancer-causing missense variants.** *Genomics* 2011, **98**(4):310-317.
90. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R: **Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.** *Cancer research* 2009, **69**(16):6660-6667.
91. Kaminker JS, Zhang Y, Watanabe C, Zhang Z: **CanPredict: a computational tool for predicting cancer-associated missense mutations.** *Nucleic Acids Res* 2007, **35**(Web Server):W595-598.
92. Shihab HA, Gough J, Cooper DN, Day IN, Gaunt TR: **Predicting the functional consequences of cancer-associated amino acid substitutions.** *Bioinformatics* 2013, **29**(12):1504-1510.
93. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N: **Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation.** *Genome medicine* 2012, **4**(11):89.
94. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al: **Pfam: the protein families database.** *Nucleic acids research* 2014, **42**(Database):D222-230.
95. Harris SL, Levine AJ: **The p53 pathway: positive and negative feedback loops.** *Oncogene* 2005, **24**(17):2899-2908.
96. Yuan TL, Cantley LC: **PI3K pathway alterations in cancer: variations on a theme.** *Oncogene* 2008, **27**(41):5497-5510.
97. Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE: **Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status.** *Nature* 2002, **418**(6901):934.
98. Sparks AB, Morin PJ, Vogelstein B, Kinzler KW: **Mutational analysis of the APC/beta-catenin/Tcf pathway in colorectal cancer.** *Cancer research* 1998, **58**(6):1130-1134.
99. Gazdar AF, Shigematsu H, Herz J, Minna JD: **Mutations and addiction to EGFR: the Achilles 'heal' of lung cancers?** *Trends in molecular medicine* 2004, **10**(10):481-486.
100. Momand J, Zambetti GP, Olson DC, George D, Levine AJ: **The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation.** *Cell* 1992, **69**(7):1237-1245.
101. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.
102. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, et al: **The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases.** *Nucleic acids research* 2014, **42**(Database):D358-363.
103. Razick S, Magklaras G, Donaldson IM: **iRefIndex: a consolidated protein interaction database with provenance.** *BMC bioinformatics* 2008, **9**:405.
104. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, et al: **The Reactome pathway knowledgebase.** *Nucleic acids research* 2014, **42**(Database):D472-477.
105. Ciriello G, Cerami E, Sander C, Schultz N: **Mutual exclusivity analysis identifies oncogenic network modules.** *Genome research* 2012, **22**(2):398-406.
106. Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A: **Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors.** *BMC medical genomics* 2011, **4**:34.
107. Zhao J, Zhang S, Wu LY, Zhang XS: **Efficient methods for identifying mutated driver pathways in cancer.** *Bioinformatics* 2012, **28**(22):2940-2947.
108. Szczurek E, Beerenwinkel N: **Modeling mutual exclusivity of cancer mutations.** *PLoS computational biology* 2014, **10**(3):e1003503.
109. Leiserson MD, Blokh D, Sharan R, Raphael BJ: **Simultaneous identification of multiple driver pathways in cancer.** *PLoS computational biology* 2013, **9**(5):e1003054.
110. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP: **DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer.** *Genome biology* 2012, **13**(12):R124.
111. Cerami E, Demir E, Schultz N, Taylor BS, Sander C: **Automated network analysis identifies core pathways in glioblastoma.** *PLoS one* 2010, **5**(2):e8918.
112. Jia P, Zhao Z: **VarWalker: Personalized Mutation Network Analysis of Putative Cancer Genes from Next-Generation Sequencing Data.** *PLoS computational biology* 2014, **10**(2):e1003460.
113. Vandin F, Upfal E, Raphael BJ: **De novo discovery of mutated driver pathways in cancer.** *Genome research* 2012, **22**(2):375-385.
114. Wendl MC, Wallis JW, Lin L, Kandath C, Mardis ER, Wilson RK, Ding L: **PathScan: a tool for discerning mutational significance in groups of putative cancer genes.** *Bioinformatics* 2011, **27**(12):1595-1602.
115. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, et al: **Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes.** *Nature genetics* 2015, **47**(2):106-114.
116. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM: **Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.** *Bioinformatics* 2010, **26**(12):i237-245.
117. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database.** *Nucleic acids research* 2009, **37**(Database):D674-679.
118. Hofree M, Shen JP, Carter H, Gross A, Ideker T: **Network-based stratification of tumor mutations.** *Nature methods* 2013, **10**(11):1108-1115.
119. Nowell PC: **The clonal evolution of tumor cell populations.** *Science* 1976, **194**(4260):23-28.
120. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al: **Tumour evolution inferred by single-cell sequencing.** *Nature* 2011, **472**(7341):90-94.
121. Zong C, Lu S, Chapman AR, Xie XS: **Genome-wide detection of single-nucleotide and copy-number variations of a single human cell.** *Science* 2012, **338**(6114):1622-1626.
122. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, et al: **Absolute quantification of somatic DNA alterations in human cancer.** *Nature biotechnology* 2012, **30**(5):413-421.
123. Hajirasouliha I, Mahmoody A, Raphael BJ: **A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data.** *Bioinformatics* 2014, **30**(12):i78-86.
124. Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q: **Inferring clonal evolution of tumors from single nucleotide somatic mutations.** *BMC bioinformatics* 2014, **15**:35.
125. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Cote A, Shah SP: **PyClone: statistical inference of clonal population structure in cancer.** *Nature methods* 2014, **11**(4):396-398.
126. Fischer A, Vazquez-Garcia I, Illingworth CJ, Mustonen V: **High-definition reconstruction of clonal composition in cancer.** *Cell reports* 2014, **7**(5):1740-1752.
127. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, Vij R, Tomasson MH, Graubert TA, Walter MJ, et al: **SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution.** *PLoS computational biology* 2014, **10**(8):e1003665.
128. Mali P, Esvelt KM, Church GM: **Cas9 as a versatile tool for engineering biology.** *Nature methods* 2013, **10**(10):957-963.

129. El Emam K: **Methods for the de-identification of electronic health records for genomic research.** *Genome medicine* 2011, **3**(4):25.
130. Paltoo DN, Rodriguez LL, Feolo M, Gillanders E, Ramos EM, Rutter JL, Sherry S, Wang VO, Bailey A, Baker R, *et al*: **Data use under the NIH GWAS data sharing policy and future directions.** *Nature genetics* 2014, **46**(9):934-938.
131. Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, Blake JA, Burleigh JG, Chanet B, *et al*: **Finding Our Way through Phenotypes.** *PLoS biology* 2015, **13**(1):e1002033.
132. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, *et al*: **COSMIC: exploring the world's knowledge of somatic mutations in human cancer.** *Nucleic acids research* 2014.
133. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B, *et al*: **International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data.** *Database : the journal of biological databases and curation* 2011, **2011**:bar026.
134. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nature reviews Cancer* 2004, **4**(3):177-183.

doi:10.1186/1471-2164-16-S8-S7

**Cite this article as:** Tian *et al.*: Computational methods and resources for the interpretation of genomic variants in cancer. *BMC Genomics* 2015 **16**(Suppl 8):S7.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

