FunSVPT: Function-based somatic variants prioritization tool for cancer whole-genome sequencing

Yao Fu[1], Zhu Liu[2], Shaoke Lou[3], Jason Bedford[1], Xinmeng Jasmine Mu[1,4], Kevin Yip[3], Ekta Khurana[1,5,*], Mark Gerstein[1,5,6,*]

[1] Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, 06520,United States of America
[2] School of Life Science, Fudan University, Shanghai, 200433, P.R. China
[3] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
[4] Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[5] Molecular Biophysics and Biochemistry Department, Yale University, New Haven, Connecticut, 06520, United States of America
[6] Department of Computer Science, Yale University, New Haven, Connecticut, 06520, United States of America

[*] To whom correspondence should be addressed

Abstract

Summary: Cancer drivers cause tumorigenesis. Separating them from acquired passenger mutations is a huge challenge, especially the noncoding ones. Here we report automated software FunSVPT to prioritize deleterious somatic variants from cancer whole genome sequencing. The tool first builds integrated data context from large-scale data resources, including population-scale polymorphisms, functional annotations and biological networks. For example, it systematically associates regulatory elements with coding genes and creates a recurrent database from publicly available cancer whole genome sequencing data. It then annotates and scores variants using various features, such as conservation, mutational impact in transcription factor binding sites, target gene and network properties. Variants are scored and ranked with a weighted-sum scoring scheme, which is based on patterns of human polymorphisms. FunSVPT is a downloadable tool and also implemented as a user-friendly web server. It can be directly used by researchers to prioritize potentially damaging cancer mutations.

Availability and Implementation: FunSVPT is implemented in Perl, R, C and Python. The FunSVPT web interface, source code, and detailed documentation are available at funsvpt.gersteinlab.org.

Contact: ekta.khurana@yale.edu or pi@gersteinlab.org

Supplementary Information: Supplementary data are available.

Introduction

Noncoding mutations are known to be important in human disease (Grossman, et al., 2013; Maurano, et al., 2012; Sakabe, et al., 2012; Ward and Kellis, 2012). In particular, many recent studies have implicated noncoding mutations as cancer drivers (Horn, et al., 2013; Huang, et al., 2013; Killela, et al., 2013; Vinagre, et al., 2013). While some methods exist for identification of cancer driver genes (Dees, et al., 2012; Reimand and Bader, 2013; Tamborero, et al., 2013; Tamborero, et al., 2013), this is not the case for noncoding drivers. Through analyzing variation patterns of inherited polymorphisms, we developed a prototype approach for identification of deleterious noncoding variants (Khurana, et al., 2013). Here, we report automated software - FunSVPT, a specialized tool for somatic variants from cancer whole-genome sequencing. Although a number of tools are available for analysis of noncoding variants - Haploreg (Ward and Kellis, 2012), RegulomeDB (Boyle, et al., 2012), ANNOVAR (Wang, et al., 2010), GEMINI (Paila, et al., 2013), FunciSNP (Coetzee, et al., 2012) and VEP (McLaren, et al., 2010) - FunSVPT offers a different framework for identifying noncoding drivers among somatic cancer variants (Supplementary Table 1). For example, it - analyzes patterns of inherited polymorphisms among humans and evolutionary conservation across species to identify regions that are less likely to tolerate mutations; uses functional annotations from ENCODE and systems-level information from various biological networks; uses

YAO FU 1/5/14 7:54 AM
Deleted: It

YAO FU 1/5/14 7:54 AM
Deleted: function

YAO FU 1/5/14 7:54 AM
Deleted: FunSVPT

YAO FU 1/5/14 8:00 AM
Deleted: more comprehensive

functional essentiality and knowledge of known cancer genes; predicts loss-of- and gain-of- function mutations for transcription-factor (TF) -binding and estimates recurrence of somatic mutations in publicly available whole-genome sequenced cancer data; consists of a weighted-sum scoring scheme to prioritize variants. In addition, FunSVPT offers a flexible framework to integrate user-specific data (e.g. gene-prioritization list) to both rebuild the underlying data context and prioritize case-specific variants. FunSVPT can be directly used by researchers and clinicians to identify potential driver mutations, especially noncoding ones.

Design and implementation

FunSVPT is an integrative tool to first build an organized data context from various resources and then prioritize case-specific variants from cancer whole-genome sequencing. The workflow of FunSVPT is depicted in Figure 1 and the detailed description is in the supplement. The various modules are described below.

YAO FU 1/5/14 7:54 AM
Deleted: of FunSVPT

Data resources

FunSVPT integrates large-scale publicly available data resources to build the data context. It collects polymorphisms from 1000 Genomes project (Genomes Project, et al., 2012), conservation data from (Bejerano, et al., 2004; Cooper, et al., 2005), functional genomics data from ENCODE (Consortium, et al., 2012) and REMC (Bernstein, et al., 2010), gene function (Futreal, et al., 2004; Wagle, et al., 2012) and networks data, as well as somatic variants from cancer whole- genome sequencing.

Variants in functional annotations and conserved regions

FunSVPT utilizes functional annotations from ENCODE (transcription factor binding sites and the high-resolution motifs within them, enhancers, ncRNAs and DNase I hypersensitive sites) and conservation data from different resources – across-species conservation from GERP scores (Cooper, et al., 2005 and ultra- conserved elements (Bejerano, et al., 2004) as well as population-level conservation from 1000 Genomes (Genomes Project, et al., 2012; Khurana, et al., 2013) to detect likely deleterious variants. FunSVPT also provides a pipeline to find novel population-level conserved regions with user input polymorphism or annotation data (Supplement).

Correlating histone marks with gene-expression data to identify likely target genes of distal regulatory elements

To interpret likely functional consequences of noncoding variants, we comprehensively define associations between regulatory elements and genes through correlating various epigenetic modifications with expression levels of genes. We considered the enhancer marks H3K4me1 and H3K27ac as two types of activity signals, and DNA methylation as an inactivity signal. Using ChIP-seq and RNA-seq data from the Roadmap Epigenomics project, for each regulatory element-candidate target pair, we computed the correlations of their activity/inactivity and expression levels across different tissue types (Supplement). We identified ~769K distal regulatory elements significantly associated with ~17K genes. All non-coding variants in these regulatory elements could be associated with potential target genes. To incorporate the ever-increasing amounts of

YAO FU 1/5/14 7:54 AM
Deleted: (e.g. H3K27ac and DNA methylation)

YAO FU 1/5/14 7:54 AM
Deleted: (Supplement). Using

genomic data, FunSVPT offers a flexible framework for users to extend the data context with their own data, for example, users can input annotation regions and chromatin marks to find novel associations between regulatory elements and coding genes (Supplement).

Gene prioritization: using expression and prior knowledge of target genes

Interpretation of the functional impact of variants can be greatly enhanced if the function of its target protein-coding genes is known. We provide a "differential gene expression analysis" module to detect differentially expressed genes in cancer samples (relative to matched normal) from RNA-Seq data. Commonly used tools – DESeq (Anders and Huber, 2010) and NOISeq (Tarazona, et al., 2011) are utilized to analyze raw read counts and RPKM values, respectively. Differentially expressed gene list will be generated and used to prioritize variants, as differential expression of target genes in cancer samples is indicative of potential effect of noncoding variants. FunSVPT also incorporates prior knowledge of genes, such as known cancer-driver genes (Futreal, et al., 2004) and actionable genes ('druggable' genes) (Wagle, et al., 2012) into our annotation scheme. In addition, user-specific gene lists can be easily input (Supplement).

Network analysis of variants associated with genes

FunSVPT uses the target genes to connect the non-coding elements into a variety of networks. For both coding and non-coding variants, we examine the network properties of the associated genes, since disruption of highly connected genes or their regulatory elements is more likely to be deleterious (Khurana, et al., 2013; Kim, et al., 2007). We make the scheme flexible so it can integrate user networks in addition to the pre-collected networks such as protein-protein interaction, regulatory and phosphorylation networks.

Recurrence across multiple cancer samples

Publicly available cancer whole-genome sequencing data provides a rich source of cancer variants. Similar to the cancer recurrent gene database in cBio (Cerami, et al., 2012; Gao, et al., 2013), we developed the recurrence database (coding genes and noncoding elements with recurrent mutations including same- site mutations) for whole-genome sequencing. Currently we have collected somatic mutations from 570 samples of 10 cancer types (Alexandrov, et al., 2013; Baca, et al., 2013; Berger, et al., 2011). For each cancer type, loci or sites with recurrent mutations in at least two samples are identified. Variants in user-input tumor genome are compared to the recurrence database and the results in different cancer types are reported in the output. Number of somatic mutations varies from a few hundreds to tens of thousands in different cancer types. Summary data about these cancer types is available for users' reference (Supplementary Table 2). FunSVPT can also detect recurrence in user-input set of multiple cancer

High-impact variants in motifs: Nucleotide resolution effects

Loss-of-function variants occurred in transcription factor binding motifs are more likely to cause deleterious impact (Kheradpour, et al., 2013; Khurana, et al., 2013; Mu, et al., 2011). Variants decreasing the position weight matrix (PWM), scores could potentially alter the binding strength of transcription factors, or even eliminate the binding. When variants hit transcription factor binding motifs, we calculate the changes in PWMs and define those decreasing PWMs as motif-breaking events (Supplement). Many studies

YAO FU 1/5/14 7:54 AM
Deleted: data for

YAO FU 1/5/14 7:54 AM
Deleted: and expression

YAO FU 1/5/14 7:54 AM
Deleted: Interpretation of the functional impact of variants can be greatly enhanced if the function of its target protein-coding gene is known. FunSVPT incorporates prior knowledge of genes, such as known cancer-driver genes (Futreal, et al., 2004) and actionable genes ('druggable' genes) (Wagle, et al., 2012) into our annotation scheme. Differential expression of target genes in cancer samples (relative to matched normal) is indicative of potential effect of noncoding variants. FunSVPT provides a "differential gene expression analysis" module to detect differentially expressed genes in cancer samples from RNA-Seq data. Commonly used tools – DESeq (Anders and Huber, 2010) and NOISeq (Tarazona, et al., 2011) are utilized to analyze raw read counts and RPKM values, respectively. In addition, user-specific gene lists can be easily input (Supplement).

YAO FU 1/5/14 7:54 AM
Deleted: noncoding

YAO FU 1/5/14 7:54 AM
Deleted: centralities

YAO FU 1/5/14 7:54 AM
Deleted: their

YAO FU 1/5/14 7:54 AM
Deleted: in various networks

YAO FU 1/5/14 7:54 AM
Deleted: FunSVPT can easily

YAO FU 1/5/14 7:54 AM
Deleted: analyzed

YAO FU 1/5/14 7:54 AM
Deleted: cancers

YAO FU 1/5/14 7:54 AM
Deleted: (Kheradpour, et al., 2013; Khurana, et al., 2013; Mu, et al., 2011). When variants occur in transcription factor binding motifs, the change in position-weight matrix (PWM) is calculated and reported in the output (Supplement).

YAO FU 1/5/14 7:54 AM
Deleted: cause loss-of-

YAO FU 1/5/14 7:54 AM
Deleted: .

have shown that gain of new binding sites caused by somatic mutations can constitute driver events (Horn, et al., 2013; Huang, et al., 2013; Killela, et al., 2013; Vinagre, et al., 2013). However, an automated tool to detect such events in whole tumor genomes is not available. FunSVPT contains a gain-of-motif scheme to scan and statistically evaluate (Touzet and Varre, 2007) all possible motifs created by variants compared to reference alleles. For each variant, we concatenate it with +/- 29bp reference sequences and calculate sequence scores against the PWMs. Gain-of-motif events are identified when sequence score with mutated allele is significantly higher than the background ($p < 4e-8$), whereas that with reference allele is not. Our scheme is validated by the detection of ETS motifs created by the two cancer driver mutations in TERT promoter (Supplementary Table 3).

Weighted-sum scoring scheme to prioritize variants

All of the above features are used to annotate and score variants. In general, features used to score variants can be classified into two classes - discrete and continuous (Figure 2). Discrete features are binary, such as in ultra-conserved regions or not. For continuous features, taking 'motif-breaking' as an example, the values are the changes in PWMs (details description is in the Supplement).

We developed a weighted-sum scoring scheme, based on the mutation pattern observed in 1000 Genomes polymorphisms, to integrate all features (Supplement). The probabilities of observing each feature in polymorphisms are calculated as p_i. Features with higher probabilities are assigned with lower weights, whereas those with lower probabilities are assigned with higher weights. For discrete features, we calculated 1-Shannon entropy *(1)* as weights. Taking 'motif-breaking' as an example for continuous features, for each motif-breaking value v observed in polymorphisms, we calculated the probability of observing values $\geq v$ then used *(1)* as weight for v. We fitted the curve of motif-breaking values and 1- Shannon entropy to obtain the smooth function (Figure 2). For each cancer variant, we score it by summing up the weights of all its features *(2)*. In addition, we considered the dependency structure of features when calculating the sum-up scores (Supplementary Table 4). Variants ranked on top of the output are those with higher scores and are most likely to be deleterious.

$$w_i = 1 + p_i log_2 p_i + (1 - p_i) log_2 (1 - p_i) \quad (1)$$

$$S = \sum_i w_i \quad (2)$$

For discrete features, p_i is the probability of observing feature i;
For continuous features, p_i is the probability of observing values $\geq v$ for feature i.

Output formats and performance

FunSVPT is a downloadable tool and also implemented as a user-friendly web server with pre-built data context (Supplementary Figures 2 and 3). It takes VCF or BED

YAO FU 1/5/14 7:54 AM
Deleted: Supplementary Table XX).. ... [1]

YAO FU 1/5/14 7:54 AM
Deleted: prioritize variants. The ... [2]

YAO FU 1/5/14 7:54 AM
Deleted: entropy …robabilities are as ... [3]

YAO FU 1/5/14 7:54 AM
Deleted: .

YAO FU 1/5/14 7:54 AM
Deleted: incorporated…onsidered the ... [4]

YAO FU 1/5/14 7:54 AM
Deleted: $e_i = -p_i log_i p_i - (1 - p_i) log_i (1 - p_i)$ . ... [5]

Unknown
Formatted ... [6]

YAO FU 1/5/14 7:54 AM
Deleted: interface (FunSVPT.gersteinlab.org)…erver with ... [7]

formatted cancer variants and generates results in either BED or VCF format (refer to Supplement for examples), which can be visualized in the UCSC genome browser. Users can either retrieve or visualize results in concise tables through the web interface.

FunSVPT runs in a tiered fashion. Building data context from bulk of data resources is time-consuming. Currently FunSVPT takes about one week (on ~20 4-core 1955 nodes) to rebuild the data context based on pre-processed functional genomics data, such as ENCODE peak calls. The data context will be updated regularly to keep it up-to-date. Users can input additional data to customize the data context upon the existing one. Variant prioritization step is quite efficient. It takes ~2-3 mins to prioritize one genome with thousands of variants on 4-core 3.00 Ghz 1955 node with 16GB RAM. FunSVPT uses parallel processing fork manager for efficient memory utilization to tackle multiple genomes in a single run. With a flexible and modularized structure, researchers can restructure the pipeline to incorporate more data and new features.

YAO FU 1/5/14 7:54 AM
Deleted: ~ 3 days

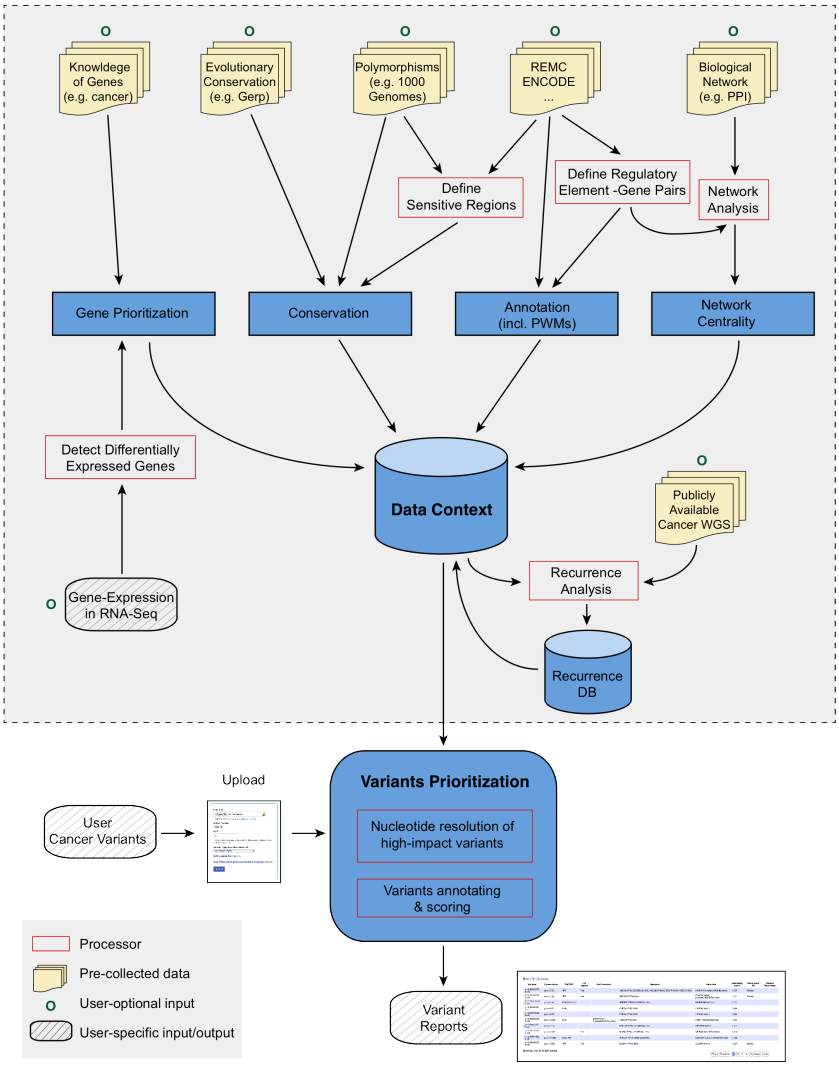

Figure 1. Schematic workflow of FunSVPT.

Unknown

Formatted: Font:Times New Roman

Features	Class
Functional annotations	Discrete
Sensitive	Discrete
Ultra-sensitive	Discrete
Motif-breaking	Continuous
Motif-gaining	Continuous
Network centrality	Continuous
Gerp score	Continuous
Ultra-conserved	Discrete
HOT Regions	Discrete
Regulatory regions associated to genes	Discrete
Recurrent	Discrete

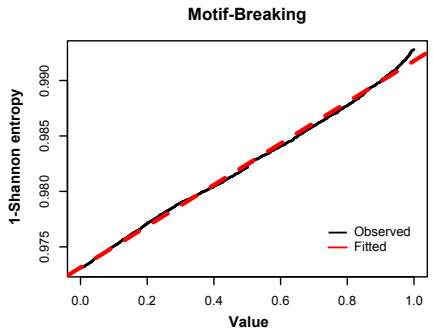

Figure 2. Weighted scoring scheme. Upper panel: Features used to score variants; Lower panel: Motif-breaking values and 1-Shannon entropy observed in polymorphisms.

Conclusions

In summary, FunSVPT is a flexible tool, which first builds data context from various data resources and then annotates and prioritizes candidate driver mutations from cancer whole-genome sequencing data. It has a weighted-sum scoring scheme to rank variants based on their potentially deleterious effect. We believe that it would be useful for researchers to identify a few somatic events among thousands for further in-depth analysis to understand the mechanisms underlying oncogenesis.

Supplementary Material

Supplementary data can be found here.

Acknowledgement

We thank NIH and A L Williams Professorship for funding.
Conflict of Interest: none declared.

Footnotes

Unknown
Formatted: Font:Times New Roman

YAO FU 1/5/14 7:54 AM
Deleted: consists of

YAO FU 1/5/14 7:54 AM
Deleted: potential

YAO FU 1/5/14 7:54 AM
Deleted: impact

References

Alexandrov, L.B.*, et al.* (2013) Signatures of mutational processes in human cancer, *Nature*, **500**, 415-421. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data, *Genome biology*, **11**, R106.

Baca, S.C.*, et al.* (2013) Punctuated evolution of prostate cancer genomes, *Cell*, **153**, 666-677. Bejerano, G.*, et al.* (2004) Ultraconserved elements in the human genome, *Science*, **304**, 1321-1325.

Berger, M.F.*, et al.* (2011) The genomic complexity of primary human prostate cancer, *Nature*, **470**, 214-220. Bernstein, B.E.*, et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium, *Nature biotechnology*, **28**, 1045-1048.

Boyle, A.P.*, et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB, *Genome research*, **22**, 1790-1797. Cerami, E.*, et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, *Cancer discovery*, **2**, 401-404. Coetzee, S.G.*, et al.* (2012) FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs, *Nucleic acids research*, **40**, e139.

Consortium, E.P.*, et al.* (2012) An integrated encyclopedia of DNA elements in the human genome, *Nature*, **489**, 57-74. Cooper, G.M.*, et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence, *Genome research*, **15**, 901-913.

Dees, N.D.*, et al.* (2012) MuSiC: identifying mutational significance in cancer genomes, *Genome research*, **22**, 1589-1598. Futreal, P.A.*, et al.* (2004) A census of human cancer genes, *Nature reviews. Cancer*, **4**, 177-183.

Gao, J.*, et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, *Science signaling*, **6**, pl1. Genomes Project, C.*, et al.* (2012) An integrated map of genetic variation from 1,092 human genomes, *Nature*, **491**, 56-65.

Grossman, S.R.*, et al.* (2013) Identifying recent adaptations in large-scale genomic data, *Cell*, **152**, 703-713. Horn, S.*, et al.* (2013) TERT promoter mutations in familial and sporadic melanoma, *Science*, **339**, 959-961.

Huang, F.W.*, et al.* (2013) Highly recurrent TERT promoter mutations in human melanoma, *Science*, **339**, 957-959. Kheradpour, P.*, et al.* (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay, *Genome research*, **23**, 800-811.

Khurana, E.*, et al.* (2013) Interpretation of genomic variants using a unified biological network approach, *PLoS computational biology*, **9**, e1002886. Khurana, E.*, et al.* (2013) Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics, *Science*, **342**, 1235587.

Killela, P.J.*, et al.* (2013) TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal, *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 6021-6026.

Kim, P.M., Korbel, J.O. and Gerstein, M.B. (2007) Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 20274-20279.

Maurano, M.T.*, et al.* (2012) Systematic localization of common disease- associated variation in regulatory DNA, *Science*, **337**, 1190-1195. McLaren, W.*, et al.* (2010)

Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor, *Bioinformatics*, **26**, 2069-2070.
Mu, X.J*., et al.* (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project, *Nucleic acids research*, **39**, 7058-7076. Paila, U*., et al.* (2013) GEMINI: integrative exploration of genetic variation and genome annotations, *PLoS computational biology*, **9**, e1003153. Reimand, J. and Bader, G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers, *Molecular systems biology*, **9**, 637. Sakabe, N.J., Savic, D. and Nobrega, M.A. (2012) Transcriptional enhancers in development and disease, *Genome biology*, **13**, 238. Tamborero, D., Gonzalez-Perez, A. and Lopez-Bigas, N. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes, *Bioinformatics*, **29**, 2238-2244. Tamborero, D*., et al.* (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types, *Scientific reports*, **3**, 2650.
Tarazona, S*., et al.* (2011) Differential expression in RNA-seq: a matter of depth, *Genome research*, **21**, 2213-2223. Touzet, H. and Varre, J.S. (2007) Efficient and accurate P-value computation for Position Weight Matrices, *Algorithms for molecular biology : AMB*, **2**, 15. Vinagre, J*., et al.* (2013) Frequency of TERT promoter mutations in human cancers, *Nature communications*, **4**, 2185.
Wagle, N*., et al.* (2012) High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing, *Cancer discovery*, **2**, 82-93. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, *Nucleic acids research*, **38**, e164. Ward, L.D. and Kellis, M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants, *Nucleic acids research*, **40**, D930-934. Ward, L.D. and Kellis, M. (2012) Interpreting noncoding genetic variation in complex traits and human disease, *Nature biotechnology*, **30**, 1095-1106.