Genome Medicine

SOFTWARE

Open Access

OncoGEMINI: software for investigating tumor variants from multiple biopsies with integrated cancer annotations

Thomas J. Nicholas[1,2], Michael J. Cormier[1,2], Xiaomeng Huang[1,2], Yi Qiao[1,2], Gabor T. Marth[1,2] and Aaron R. Quinlan[1,2,3*]

Abstract

Background: DNA sequencing has unveiled extensive tumor heterogeneity in several different cancer types, with many exhibiting diverse subclonal populations. Identifying and tracing mutations throughout the expansion and progression of a tumor represents a significant challenge. Furthermore, prioritizing the subset of such mutations most likely to contribute to tumor evolution or that could serve as potential therapeutic targets represents an ongoing problem.

Results: Here, we describe OncoGEMINI, a new tool designed for exploring the complex patterns and trajectory of somatic and inherited variation observed in heterogeneous tumors biopsied over the course of treatment. This is accomplished by creating a searchable database of variants that includes tumor sampling time points and allows for filtering methods that reflect specific changes in variant allele frequencies over time. Additionally, by incorporating existing annotations and resources that facilitate the interpretation of cancer mutations (e.g., CIViC, DGIdb), OncoGEMINI enables rapid searches for, and potential identification of, mutations that may be driving subclonal evolution.

Conclusions: By combining relevant genomic annotations alongside specific filtering tools, OncoGEMINI provides powerful and customizable approaches that enable the quick identification of individual tumor variants that meet specified criteria. It can be applied to a wide range of tumor-derived sequence data, but is especially designed for studies with multiple samples, including longitudinal datasets. It is available under an MIT license at github.com/fakedrtom/oncogemini.

Background

Cancers arise from a variety of genetic alterations and, over time, often accumulate a substantial mutational load leading to subclonal diversity [1, 2]. As a result, DNA sequencing of tumors has revealed extensive heterogeneity within primary tumors or between primary and subsequent occurrences [3] and that the degree of heterogeneity is correlated with patient outcomes [4]. Prioritizing mutations in the face of this heterogeneity relies upon accurate variant discovery and being able to differentiate variants with potential relevance from those that are less likely to contribute to the proliferation of a given tumor. Nevertheless, properly deciphering which identified variants, if any, contribute to tumor origin, survival, and proliferation is crucial to understanding tumor biology and determining potential treatments.

Recognizing this need, we introduce OncoGEMINI [5] as a new software to explore genetic variation observed

* Correspondence: aaronquinlan@gmail.com
[1]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA
[2]Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT 84112, USA
Full list of author information is available at the end of the article

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

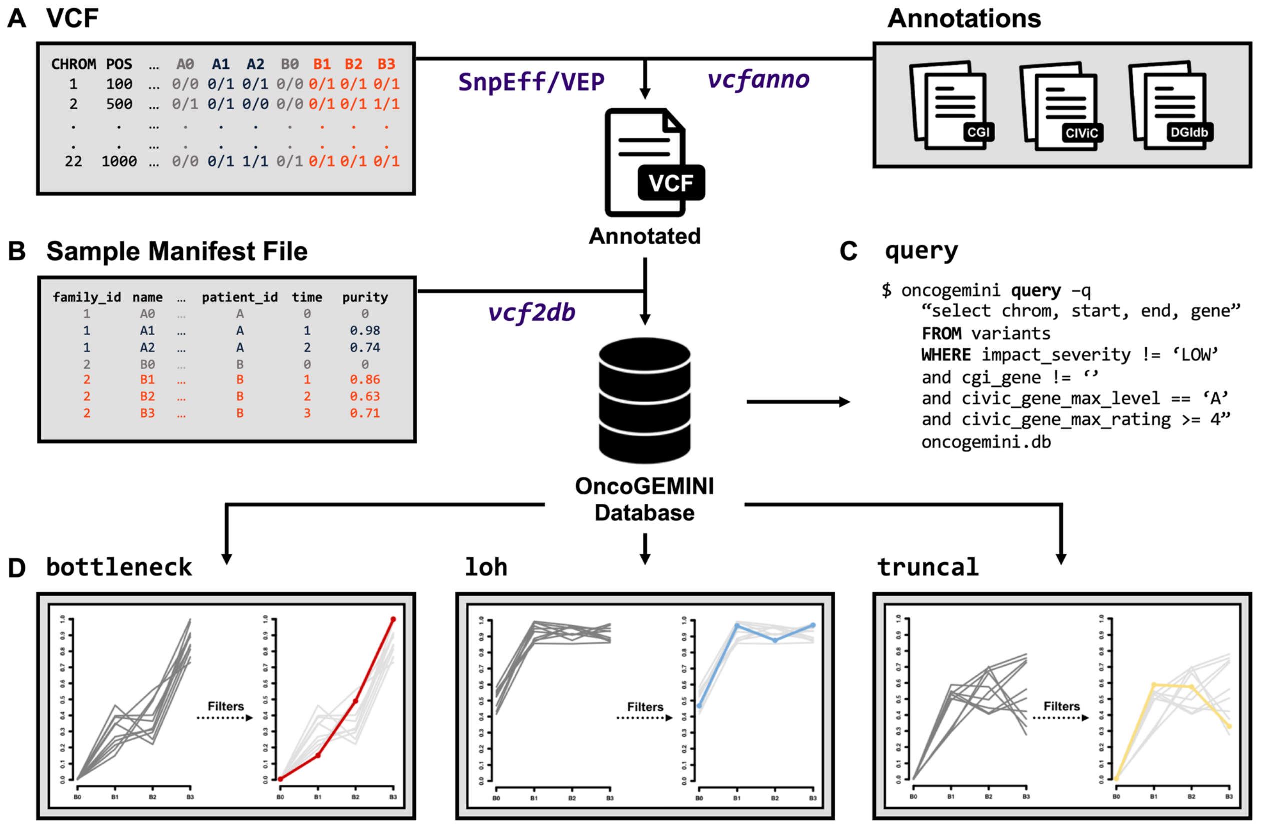
across multiple tumor biopsies and facilitate the identification of both inherited and somatic mutations that may be involved in tumor progression or resistance. OncoGEMINI is uniquely effective in the analysis of variation observed in either a single or multiple cancer biopsies from one or more patients. OncoGEMINI builds upon the GEMINI framework [6], which creates a database from a VCF [7] file and extensively annotates genetic variants in an effort to facilitate analysis. GEMINI was designed for the analysis of inherited variants in studies of rare disease [8–10] and is poorly suited to the analysis of somatic mutations, tumor heterogeneity, and the analysis of multiple biopsies that vary over time and location in the patient's body.

OncoGEMINI addresses these limitations and enables rapid variant exploration in tumor sequencing studies, especially those featuring longitudinal data across multiple time points from a single patient. To better select or prioritize tumor variants, OncoGEMINI integrates several cancer-relevant genomic annotations that serve as searchable terms to differentiate variants from one another. Additionally, OncoGEMINI provides multiple filtering tools that search for various signatures of tumor heterogeneity, including observable allele frequency changes across multiple samples. These include the `bottleneck`, `loh`, `truncal`, and `unique` tools. By using these filtering tools in combination with specific cancer annotations, OncoGEMINI can effectively prioritize tumor variants that may drive tumor progression or be potential treatment targets.

Implementation

OncoGEMINI framework

OncoGEMINI employs the same general functionality as GEMINI and imports variant information, including sample genotypes, from a VCF file into a searchable SQLite database. OncoGEMINI is intended to be used alongside the VCF annotation tool, vcfanno [11], and the database creation tool, vcf2db [12]. Together, these allow for efficient loading of user-specified (or created) annotations to be included in the resulting database (Fig. 1). Database creation times vary depending on the number



```
CHROM POS  …  A0  A1  A2  B0  B1  B2  B3
  1   100  … 0/0 0/1 0/1 0/0 0/1 0/1 0/1
  2   500  … 0/1 0/1 0/0 0/0 0/1 0/1 1/1
  .    .   …  .   .   .   .   .   .   .
  .    .   …  .   .   .   .   .   .   .
 22   1000 … 0/0 0/1 1/1 0/1 0/1 0/1 0/1
```

```
family_id name  …  patient_id time  purity
    1      A0   …      A       0      0
    1      A1   …      A       1     0.98
    1      A2   …      A       2     0.74
    2      B0   …      B       0      0
    2      B1   …      B       1     0.86
    2      B2   …      B       2     0.63
    2      B3   …      B       3     0.71
```

```
$ oncogemini query -q
    “select chrom, start, end, gene”
    FROM variants
    WHERE impact_severity != ‘LOW’
    and cgi_gene != ‘’
    and civic_gene_max_level == ‘A’
    and civic_gene_max_rating >= 4”
    oncogemini.db
```

Fig. 1 Overview of OncoGEMINI database creation and usage. **a** Genetic variants from a VCF file are first annotated by the user with either SnpEff or VEP and then additional user-determined annotations can be added to the SnpEff or VEP annotated VCF file using the vcfanno tool. **b** The fully annotated VCF, along with a sample manifest file, is then used to create an OncoGEMINI database via vcf2db. Once the OncoGEMINI database is created, users are able to **c** create their own customized queries with the **`query`** command or **d** select from built-in tools (such as `bottleneck`, `loh`, or `truncal`) to filter variants. To better select and focus on specific variants, users may combine tools and queries with included annotations to further filter the number of variants

of variants and annotations included in the annotated VCF file, but over a thousand variants per second can typically be inserted into the database using vcf2db.

An important change to the existing GEMINI framework that is necessary for some of OncoGEMINI's functionality is the inclusion of a tumor-specific sample manifest file used during database loading. OncoGEMINI utilizes a user-generated sample manifest file to describe the distinctive relationships that may exist between tumor samples. This manifest retains the same basic structure of common pedigree files with additional columns describing patient_id, time, and lastly, purity. Given that multiple samples may be derived from the same patient, the patient_id column ensures that sample relatedness is properly cataloged. Since these samples may be obtained through a series of biopsies at different time points, the time column describes the temporal relationship between samples for a given patient. While OncoGEMINI does not compute tumor purity estimates, if such estimates are known, they may be included in the purity column and invoked as a command line parameter to alter allele frequencies accordingly. Altogether, these extra columns give OncoGEMINI greater flexibility and specificity when performing queries or employing different filtering schemes.

Cancer-specific annotations

Diverse databases and annotations have been developed to help interpret identified tumor variants with their specific relevance to various aspects of tumor origin, growth, and treatment [13–17]. Such annotations can be essential in determining which identified tumor variants are actionable or otherwise merit further investigation from those that are less relevant to tumor progression. Many of these annotation sources exist independently of one another and not always in formats that lend themselves readily to bioinformatic applications.

We have integrated a number of cancer-relevant annotation resources including the Cancer Genome Interpreter (CGI) [18], Clinical Interpretations of Variants in Cancer (CIViC) [19], and the Drug Gene Interaction Database (DGIdb) [20]. Each annotation provides pertinent information relating to various aspects of tumor biology and, in some cases, has also summarized various genome annotations from other resources. We have made all of these annotation files, and guides to their creation, available via the Cancer Relevant Annotations Bundle, CRAB, resource (https://github.com/fakedrtom/crab). While the relevance of each annotation varies, especially with regard to particular cancer types, they provide valuable insights into the pathogenicity and frequency of genetic variants, the propensity for certain genes to harbor mutations in specific cancer types, or known drug susceptibilities and interactions with genes or individual variants. From each of these, we have converted selected information into a series of summarized files that can then be used to annotate a VCF file in preparation for loading into an OncoGEMINI database. Given the inherent flexibility of vcfanno, any combination of these annotations or any additional ones not included here may be selected and added to a VCF. We expect that this will be a valuable resource even outside of use within OncoGEMINI and are eager to add more annotations, references, and data to improve its utility.

OncoGEMINI functionality

Combining tumor variants with genome annotations into an OncoGEMINI database enables the identification and prioritization of tumor-relevant genetic variants. Once loaded, the database is populated with the information contained within the selected VCF. Variants from the VCF become rows in the database, and annotations are stored as columns within the OncoGEMINI database that can be useful for isolating variants of interest.

The query tool

Via OncoGEMINI's `query` tool, users are able to impose specific search parameters to identify variants meeting custom search criteria. For example, if one wanted to filter all variants in an OncoGEMINI database to only those with the highest validated clinical association as documented in CIViC (i.e., an evidence level of "A"), the following query would restrict the output to the genomic coordinates, reference and alternate alleles, and gene (if applicable), for all variants that match those in the CIViC database with an evidence level of "A":

```
oncogemini query -q "select chrom, start, end,
ref, alt, gene from variants where civic_evi_
level = 'A'" oncogemini.db
```

The `query` tool can take advantage of the flexibility of OncoGEMINI's SQL language to build sophisticated search commands. Anticipating common searches pertaining to tumor growth and propagation, we have also developed a suite of tools that identify variants that follow patterns of particular interest in tumor evolution and alleviate the need to repeatedly craft lengthy search queries. Each of the tools described in the following sections is capable of further customization with additional tool-specific parameters and via the annotations included in the database. Together, these options provide the ability to design explicit search queries that can effectively reduce the list of variants, thereby narrowing in on potential variants of interest.

The bottleneck tool

As individual tumor cells replicate, they may acquire private mutations that give rise to subclonal structures and heterogeneity within tumors. Over time, owing to drift or the selective pressure of drug and other treatments, the prevalence of certain tumor variants can be reduced or increased. Thus, the ability to determine the frequency at which any given tumor variant exists is both a function of sampling time and efficiency. Variants which increase in frequency over time, especially across different interventions, may represent potential contributors to the proliferation and survivability of a given tumor. We have developed the `bottleneck` tool to identify variants exhibiting this increased allele frequency over time. This tool relies upon having information regarding the sequential order that tumor samples arose; therefore, this tool is most applicable to longitudinal data, but may be useful with other studies, including metastatic studies, provided temporal information is otherwise appropriately known or ascertained. The `bottleneck` tool scans each variant in the database for samples that exhibit increasing allele frequencies over the sampling time indicated in the sample manifest file (Fig. 2). To accomplish this, the `bottleneck` tool uses the allele frequencies of specified samples across the sample time points and calculates the slope of these values. Variants that exceed a given slope (default 0.05) are reported in the output.

The loh tool

Loss of heterozygosity (LOH) is a common genomic alteration in cancer genomes that leads to the loss of one allele. LOH events can result in functional consequences including reduced gene expression, haploinsufficiency, or acting as the second “hit” in a tumor suppressor gene. LOH mutations can indicate causative variants or otherwise serve as biomarkers for cancer identification and potential patient care. One method for identifying LOH relies upon observing genetic loci that appear heterozygous in germline DNA but are not heterozygous in tumor DNA. The `loh` tool identifies potential LOH variants by observing allele frequency changes that are consistent with heterozygosity present in the normal tissue samples, but absent in all specified tumor samples (Fig. 2).

The truncal tool

Somatic mutations that arose early in tumor development are likely to be present in all sequenced samples. Such mutations may serve as important therapeutic targets since they are present throughout all samples, rather than localized and private to any specific subclones. For example, this subset of variants may harbor potential neoantigens which represent ideal targets for patient-specific T cell-based cancer immunotherapy [21]. The `truncal` tool was designed to identify genetic variants that are present in all given tumor samples, but absent from any normal samples (Fig. 2). Similar to the `loh` tool, the `truncal` tool also requires a normal sample to be included and accepts a defined maximum allele frequency to be allowed in the normal tissue samples (default is 0). Variants where all the tumor samples have allele frequencies that are higher than the maximum normal tissue allele frequency are included in the output.

The unique tool

While one of the key features of analyses with OncoGEMINI is the inclusion of multiple samples from different spatial and temporal biopsies, the `unique` tool also allows OncoGEMINI to highlight variants that are specific to a single sample or group of samples. This tool requires desired samples to be listed and returns all variants that are present in that samples and absent from all others. Similar to the `truncal` tool, this is done by specifying a minimum allele frequency (default is 0) that must be exceeded in the listed samples, but not be met in all other samples.

Identifying somatic mutations

OncoGEMINI will evaluate all variants within the database and select those that meet specified tool and annotation filter requirements. Thus, if the VCF used to create the database contained both germline and somatic mutations, both mutation types would be considered by OncoGEMINI commands. To focus solely on somatic mutations, it is recommended that the VCF used for the creation of a OncoGEMINI database be pre-filtered to only include somatic mutations. Somatic mutations may also be labeled with a specific identifier in the “INFO” field of the VCF, thereby enabling facile filtering for such mutations within the database. Alternatively, the `set_somatic` tool may be employed to annotate variants as somatic within an existing OncoGEMINI database, based on genotypes or user-defined criteria regarding normal and tumor sample sequencing depths and allele frequencies. The `set_somatic` tool creates an “is_somatic” annotation within the database. OncoGEMINI tools may then take advantage of the `--somatic-only` parameter to restrict variant evaluations to only those variants that have been marked as somatic in the database by the `set_somatic` tool.

Results

OncoGEMINI’s integration of genetic variants identified across one or more biopsies with annotations relevant to cancer enables a wide range of analyses and variant prioritization strategies. Furthermore, the OncoGEMINI framework can be used to study variation observed in diverse study designs, including studies of a single tumor

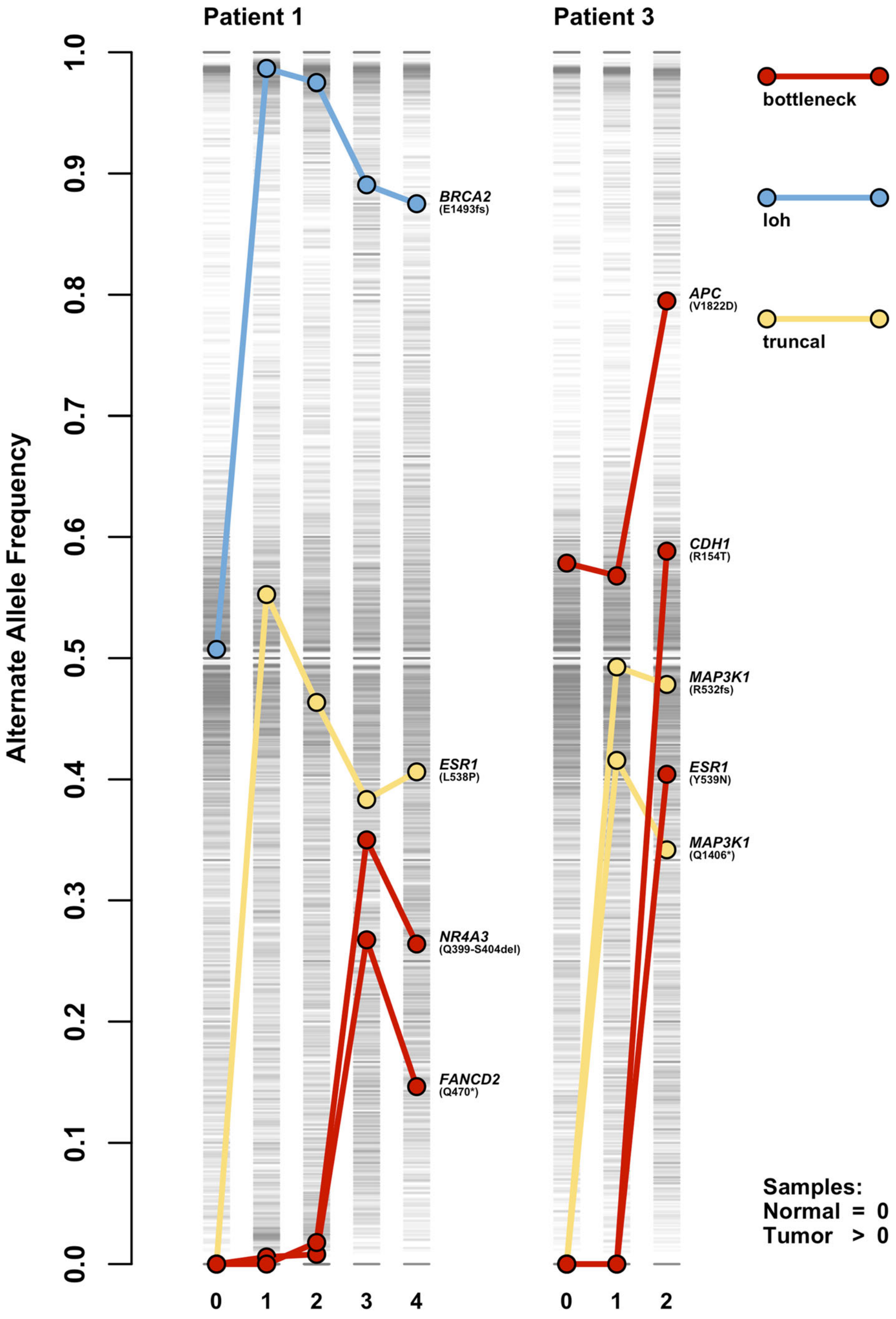


Fig. 2 Using OncoGEMINI filtering tools to identify variants with specific patterns. Genetic variants corresponding to all samples originating from representative patients, #1 and #3 (further details in "Results" section), are indicated with single gray lines at their respective alternate allele frequencies, where darker grays represent a greater accumulation of variants with similar frequencies. Samples are represented by their sampling time as indicated in the sample manifest file, where a time of 0 indicates a normal tissue sample and values greater than 0 correspond to subsequent tumor samples. Individual variants and their alternate allele frequency changes across multiple tumor time points are highlighted with colored points connected by lines and were isolated from all other variants using the OncoGEMINI filtering tools: `bottleneck` (red), `loh` (blue), and `truncal` (yellow). For each of these variants, the gene in which they occur and their specific mutation are listed next to the point in the final sample for each patient

and normal biopsy and multiple tumor and/or ascite biopsies from the same patient over the course of treatment. Existing tools [22–26] are well-suited to the study of mutations found in matched tumor-normal studies and OncoGEMINI's primary innovation is the ability to analyze multiple related samples alongside one another.

We therefore demonstrate OncoGEMINI's functionality by providing example analyses that isolate variants of interest in a recent study of longitudinal biopsies from three breast cancer patients [27], as well as a study of four breast cancer patients that explored somatic mutations revealed from a total of 48 primary tumor and metastatic biopsies [28].

Breast cancer longitudinal data

We first highlight OncoGEMINI's utility with an example analysis that identifies and filters variants of interest from longitudinal biopsies obtained from three breast cancer patients [27]. We specifically focused on patients #1 and #3, each of which was followed for more than 3 years and experienced multiple tumor recurrences (4 total tumor samples in patient #1 and 2 in patient #3), undergoing biopsies with samples taken before, during, or after a variety of distinct drug treatments. Specific details regarding these patients, their tumor history, treatments, data generated, and conducted analyses were previously reported [27]. In summary, the original study reported 5,543,181 and 5,355,078 total unfiltered variants (40,853 and 28,647 were identified as being somatic) in patients #1 and #3, respectively, from whole-genome sequencing that were further filtered through a combination of manual curation, analysis, and validation. Automating similar future analyses as much as possible was a key motivation for the development of OncoGEMINI. Of these previously reported somatic mutations, the authors of the original study identified nine total SNV or small indel variants (8 mutations and 1 inherited variant) in patients #1 and #3 as cancer drivers or otherwise biologically relevant variants (for specific variant details, see Fig. 2). These variants serve as ideal candidates for the types of variants of interest that OncoGEMINI should be able to prioritize using its improved annotations and filtering tools.

For each VCF, variants were annotated in two distinct steps. First, consistent with GEMINI, OncoGEMINI also incorporates standard variant effect predictions made by SnpEff [29] or VEP [30], and in this case, initial annotations were added via SnpEff. Second, we downloaded additional cancer-relevant features from various databases, including CGI, CIViC, and DGIdb, via the CRAB and subsequently added them as variant annotations to the VCFs using vcfanno. A sample manifest file was prepared that listed all of the samples, their names, the patients they corresponded to (either patient #1 or #3), and time points where the normal samples were given a time of 0 and subsequent tumor samples were given times greater than 0 with each consecutive sample time increasing by 1. The optional purity column was not included for either of these patients. With the sample manifest file, OncoGEMINI databases, including all of the added annotations, were then created for each patient's VCF using vcf2db. The resulting databases contained all unfiltered variants for patients #1 and #3, which are reduced to 4,986,519 and 4,928,034 total variants (32,840 and 26,106 that are marked as somatic variants), respectively, if we require a minimum sequencing depth threshold of 10 for all samples.

To further reduce this number and reveal potential variants of interest (i.e., the previously reported variants), we required certain filters via the included annotations to be met and utilized many of the previously described OncoGEMINI tools. As expected, the count of variants that OncoGEMINI returns depends upon the number of annotation filters and tools that are specified where, generally speaking, the more filters alongside tools that are required, the more the list of returned variants is reduced.

Specifying that variants be filtered to only those with a SnpEff impact prediction of "medium" or "high" (`impact_severity != 'LOW'`) drastically reduced the number of returned variants to 20,580 and 19,516, respectively. Also, restricting variants to those found in genes with previous implications towards tumorigenesis in certain cancer types via CGI's Catalog of Cancer Genes (`cgi_gene != ''`) and genes with high CIViC evidence levels or ratings (`civic_gene_max_level == 'A' or civic_gene_max_level == 'B' or civic_gene_max_rating >= 4`) refined the number of variants to 140 and 149. Combining these annotation filters with each of the OncoGEMINI `bottleneck`, `loh`, and `truncal` tools resulted in a total of three variants from patient #1 and 2 variants from patient #3 being returned which accounted for three of the nine previously reported variants. To try and recover the remaining 6 previously reported variants, we relaxed the filter requirements by removing the CIViC evidence level and rating annotation filters. This increased the number of resulting variants to 1245 and 1253, respectively, but by once again passing them through the `bottleneck`, `loh`, and `truncal` tools only 17 and 27 variants remained, including eight of the nine previously reported variants. Altogether with just a few commands that returned results in seconds of time, we were able to filter millions of tumor variants to a much more manageable number in each patient, which included nearly all of the previously reported variants (Fig. 3). We note that distinct annotations from SnpEff, CGI, and CIViC are each capable of substantially reducing the number of variants, but by combining these annotation filters with one another and each of the OncoGEMINI filtering tools the number of variants is refined to a specific few. We also identified in each patient additional mutations that were not previously reported,

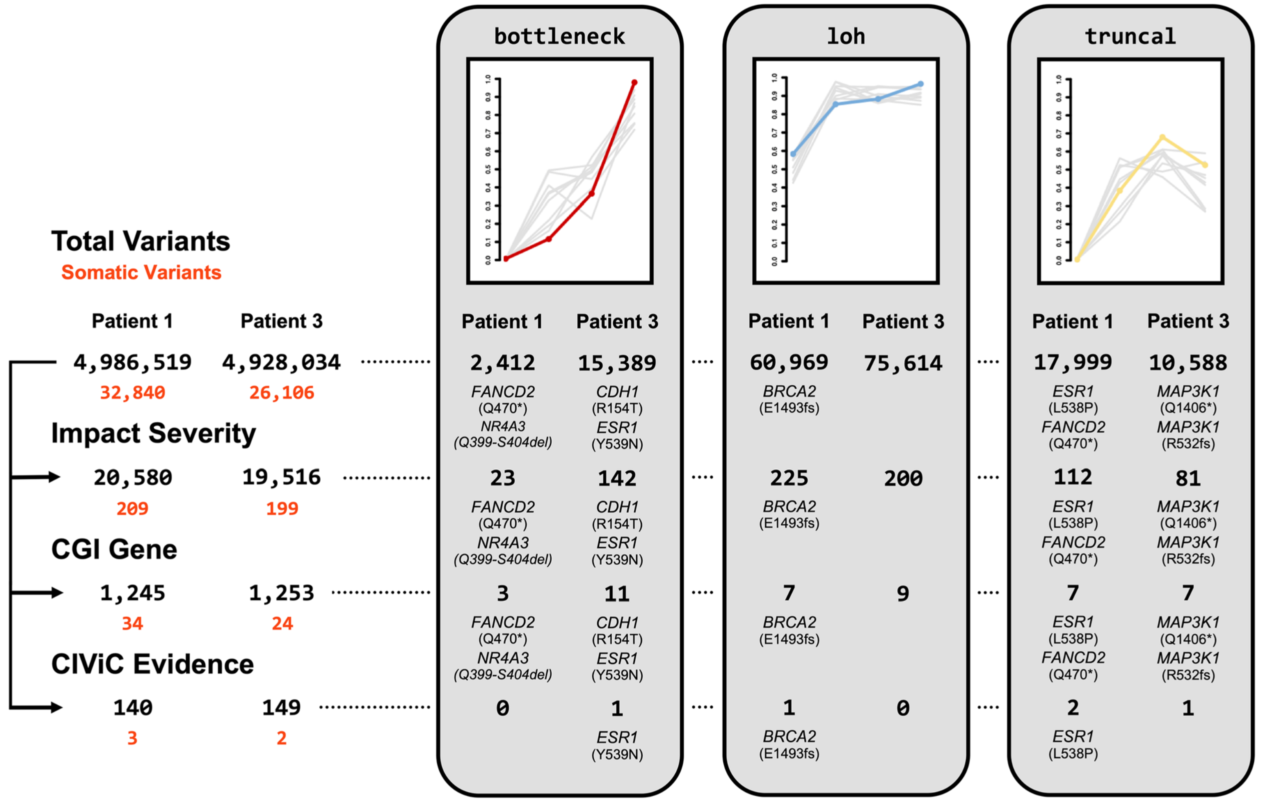


Fig. 3 Schematic of an OncoGEMINI workflow for filtering variants. Variants are filtered using a combination of OncoGEMINI tools and included annotations as filter requirements. For each set of numbers, those listed on the left belong to patient #1, and those on the right are from patient #3. On the leftmost side of the figure, the total number of variants for each patient is listed with an increasing number of cancer annotation filter requirements being added from top to bottom, and the corresponding number of variants that meet those requirements, being specified as individual rows. Below each of these values, colored in orange, are the total number of variants that are flagged as somatic for that given combination of annotation filters. For each row, the number of variants is further restricted by individual OncoGEMINI tools listed towards the right side of the figure, highlighted in light gray boxes. The number of variants returned by different combinations of OncoGEMINI tools and cancer annotation filters are continued as rows within each of the gray boxes corresponding to different tool types. All variant counts listed in the gray boxes correspond to the total number of somatic and inherited variants. However, the majority of returned variants are somatic in the bottleneck and truncal boxes, especially as more annotation filters are used. Given that LOH variants are present in the germline, they are not labeled as somatic mutations in the same manner as bottleneck or truncal variants. Thus, variants identified in the loh box are not necessarily labeled as somatic. Any combinations of OncoGEMINI tools and cancer annotation filters that return any of the previously identified somatic variants are specified with the genes in which the variants were found and their specific mutations

but that meet the same parameters as those that were. While these variants are not validated and their potential role in the subclonal tumorigenesis in each of the patients is unknown, they may be worth further consideration.

It is important to note that while this filtering scheme has been described in a sequential manner, all of the searches, including the application of multiple filters with individual tools, can be accomplished with a single command. For example, using the `truncal` tool alongside all of the previously described filters can be done like so:

```
oncogemini truncal --minDP 10 --columns “chrom,start,end,ref,alt,gene” --filter “impact_severity != ‘LOW’ and cgi_gene != “ and (civic_gene_max_level == ‘A’ or civic_gene_max_level == ‘B’ or civic_gene_max_rating >= 4)” patient1.db
```

This command will return the chromosome, start and end positions, reference and alternate alleles, and the gene for the two truncal variants in patient #1 that meet these criteria. This allows for maximum customization into a single command when devising search specifications.

Breast cancer regional metastatic data

OncoGEMINI also provides a means to evaluate multiple tumor samples as is common for many metastatic and other regional tumor studies. Here, we demonstrate an example spatial analysis that focuses on four individual breast cancer patients, identified as ER1, ER2, ER3, and TN1, that each exhibited extensive metastatic

Table 1 Summary of breast metastatic samples and called variants

Patient	Primary tumor biopsies	Pre-mortem metastatic biopsies	Post-mortem metastatic biopsies	Total variants	Somatic variants
ER1	3	0	5	1,593,541	71,987
ER2	4	1	8	1,363,283	170,695
ER3	2	2	12	2,796,856	67,354
TN1	4	1	6	2,628,149	152,713

expansions from a primary tumor to other organs while receiving various treatments and interventions over the course of several months or years [28]. Multiple biopsies were obtained from each patient, including multiple sampling of the primary tumor and post-mortem metastatic samples. For some patients, pre-mortem metastatic samples were also acquired (Table 1). The previously published study highlights specific mutations identified in each patient, including truncal mutations found in all tumor samples that are reported as metastatic drivers. Additionally, subclonal mutations that are only present in specific metastatic samples are also identified. By focusing on only SNV and indel variants, we established, for each patient, a list of variants that we expect to be present in all tumor samples as well as those that we expect to be present in all metastatic samples for each patient (Tables 2 and 3). We further demonstrate that by using appropriate tools from OncoGEMINI, we can quickly identify these truncal and metastatic mutations from data corresponding to multiregional biopsies.

In total, 48 tumor sample biopsies were obtained from these four patients. Whole-exome sequencing was performed for all samples, including a blood sample from each patient, and patient-specific, joint-called VCFs were generated for all patient biopsies using FreeBayes [31]. FreeBayes parameters were relaxed to maximize sensitivity in variant calling in the multiple samples (see Additional file 1: Supplementary Methods). Each VCF was then annotated as previously described for the longitudinal example using SnpEff and vcfanno with data from the CGI, CIViC, and DGIdb databases. Sample manifests were created for each patient with the blood sample being treated as a normal or germline sample and assigned a time value of 0. Primary tumor samples were treated as the next subsequent time point and given the value of 1 with pre-mortem metastatic samples following with a time value of 2 and post-mortem samples given the value of 3. The ER3 patient was exceptional in that a pre-mortem metastatic sample was actually obtained before any of the primary tumor biopsies. In this case, the

Table 2 Summary of OncoGEMINI truncal filtering of breast metastatic samples

Patient	Expected truncal mutations	OncoGEMINI truncal	
		Tool only	With filters[a]
ER1	3	1,158	27
	CDH1 (indel)	*CDH1* (indel)	*CDH1* (indel)
	PIK3CA (E545K)	*PIK3CA* (E545K)	*PIK3CA* (E545K)
	PIK3CA (E726K)	*PIK3CA* (E726K)	*PIK3CA* (E726K)
ER2	3	1,582	50
	AKT1 (E17K)	*AKT1* (E17K)	*AKT1* (E17K)
	ARID1A (W1844X)	*ARID1A* (W1844X)	*ARID1A* (W1844X)
	PAX6 (splice)		
ER3	2	1,474	32
	SPEN (K1838X)		
	TP53 (splice)		
TN1	4	1,930	29
	DIDO1 (R2008X)	*DIDO1* (R2008X)	*PIK3CA* (H1047R)
	ITPR1 (R170X)	*PIK3CA* (H1047R)	*TP53* (T118fs)
	PIK3CA (H1047R)	*TP53* (T118fs)	
	TP53 (T118fs)		

[a]Requires impact_severity ! = 'LOW' and the cgi_gene ! = "

Table 3 Summary of OncoGEMINI unique filtering of breast metastatic samples

Patient	Expected metastatic mutations	OncoGEMINI unique	
		Tool only	With filters[a]
ER1	3	100	2
	CTPS2 (D153E)	*CTPS2* (D153E)	*FGFR4* (N495K)
	FGFR4 (N495K)	*FGFR4* (N495K)	*MGA* (splice)
	MGA (splice)	*MGA* (splice)	
ER2	1	181	6
	SPEN (E2151K)	*SPEN* (E2151K)	*SPEN* (E2151K)
ER3	0	4	0
TN1	0	17	0

[a]Requires impact_severity ! = ‘LOW’ and the cgi_gene ! = “

pre-mortem sample was given a time point of 1 with the primary tumor, an additional pre-mortem biopsy, and the post-mortem samples then given values of 2, 3, and 4, respectively. Purity values that were previously calculated by the authors using FACETS [32] were also included in the manifests for each patient, thus enabling the use of the optional `--purity` allele frequency correction parameter. Using the annotated VCFs and the sample manifests, an OncoGEMINI database was then created for each patient using vcf2db. The `set_somatic` tool was then applied to each database to differentiate variants as somatic or not. Parameters for `set_somatic` required a minimum sequencing depth of 10 for all samples and allowed an allele frequency of 0.05 in the normal samples while simultaneously requiring that at least a single tumor sample had an allele frequency of 0.2 or higher. The `--purity` parameter was also invoked, resulting in purity-adjusted allele frequencies being used for all determinations by the `set_somatic` tool. All subsequent analyses focused on only the variants marked somatic by the `set_somatic` tool which substantially reduces the number of variants under consideration (Table 1).

The `truncal` tool is ideally suited to recover variants that are present in all tumor samples, but absent from any other samples included. For each patient, the `truncal` tool (with the --maxNorm parameter set to 0.05) recovered nearly all previously reported mutations that fit this expected mutation profile with a few exceptions (Table 2). The four unrecovered mutations are present in the VCF, but did not pass the OncoGEMINI specifications, generally by not meeting either minimum depth or allele frequency thresholds. Some exceptions failed to meet the truncal definition of being present in all tumor samples while being absent from the normal samples, unless certain samples were omitted from the analysis. For example, the *SPEN* (K1838X) mutation in the ER3 patient does appear as truncal, but only if the first metastatic sample, ER3_M1, is omitted from the analysis. As mentioned previously, ER3_M1 was the first biopsy taken from this patient, and it would appear that this *SPEN* mutation was either not sampled in ER3_M1 or it developed later and persisted in all subsequent biopsies. Similarly, an *ITPR1* mutation in TN1 fails to be recovered by the truncal tool, because it appears in only two of the four primary tumor samples that were included, and is also absent from one metastatic sample. It is possible that the lack of evidence for these mutations in these samples can be attributed to variant calling differences employed here versus the original publication. By removing the samples in question using the --samples parameter, these mutations are then recovered by the truncal tool.

Identifying mutations specific to groups of samples is precisely what the unique tool is designed to do and is therefore appropriate for the identification of subclonal mutations that are specific to certain metastatic samples, but absent from all others (including the primary tumor samples). For example, ER1 has five metastatic samples labeled as ER1_A1, ER1_A2, ER3_A3, ER1_A4, and ER1_A5 within the ER1 OncoGEMINI database, and the following command can use the unique tool to identify variants with a minimum sequencing depth of 10 across all samples and that are present in all indicated metastatic biopsies, but absent from all other included samples:

```
oncogemini unique --minDP 10 --specific ER1_A1,ER1_A2,ER3_A3,ER1_A4,ER1_A5 --columns “chrom,start,end,ref,alt,gene” ER1.db
```

This command returns the chromosome, start and end positions, reference and alternate alleles, and the gene name (if the variant is within a gene region) for all variants that are found in the indicated metastatic samples from ER1. By using similarly structured commands (including the --maxOthers parameter set at 0.05) that specify only the metastatic samples for each patient, we expect to find certain alterations in the metastatic samples of ER1 and ER2 and we are able to recover all of these using the unique tool (Table 3). The unique tool can also enable further refinement of subclonal mutations present within subsets of metastatic samples. Previous work in ER1 indicated additional mutational signatures shared by metastatic samples ER1_A1, ER1_A2, and ER1_A4 that suggest a shared origin between them. Indeed, using the unique tool to specify only these samples reveals mutations that appear to be unique to these samples including the previously reported *SPOP* (Y87F) variant.

Consistent with the previous longitudinal data analysis, the tools provided by OncoGEMINI are capable of greatly reducing the number of variants, but when used alongside annotation filters, the number of returned variants is further refined (Tables 2 and 3). We emphasize that the OncoGEMINI tools and filters used to identify these previously reported variants also return additional variants that meet the same parameters and thus may be of interest. In this manner, OncoGEMINI serves as a comprehensive tool to report all variants that meet biologically relevant criteria in a variety of tumor studies.

Conclusions

Building upon the GEMINI framework, OncoGEMINI is ideally suited to the exploration and prioritization of tumor mutations. Cancer research increasingly requires the integration of data from not only multiple biological samples, but also the accumulated genomic information that is found in a variety of distinct databases. We have designed OncoGEMINI to combine information from numerous sources alongside longitudinal tumor sequence data to enable the rapid identification of variants that match specific patterns and criteria. Therefore, OncoGEMINI provides a unique tool that assists in complex cancer analyses.

While much of OncoGEMINI's functionality is applicable to a variety of data from different tumor studies, it is optimized to incorporate and analyze data from multiple tumor samples belonging to the same patient and is thus most powerful when used with longitudinal data that has been collected over various time points and treatments. OncoGEMINI offers a number of specific filtering tools that focus on variant allele frequency changes between samples and each tool can be further adjusted with included options and parameters. Additionally, using the vcfanno tool, user-defined OncoGEMINI databases can be built that enable different genomic annotations to be incorporated and used in identifying relevant tumor mutations, thus empowering powerful cancer-specific queries. We have suggested and prepared specific cancer annotations, but given the inherent flexibility afforded by vcfanno, custom annotations provided or created by users can be included. This allows OncoGEMINI greater versatility and suitability to a wide range of cancer analysis projects. By combining annotation information with OncoGEMINI tools, we demonstrated that individual tumor variants can be identified with simple commands that run quickly in a matter of seconds.

Even though OncoGEMINI is not clinically deterministic, it can help rapidly sort through tumor variants from sequencing data and identify individual variants fitting specific requirements that may be indicative of variants with potential clinical significance. The default settings of the OncoGEMINI tools aim to pinpoint such likely relevant variants, but being primarily an exploratory tool, OncoGEMINI's individual tool settings can be adjusted, as is appropriate, to expand query criteria. For example, the last remaining variant from the previously discussed longitudinal example was not recovered in the described analysis (patient #3, *APC* V1822D) because it was a germline variant that increased in frequency over the time course of the tumor samples (Fig. 2). This variant follows the allele frequency pattern that is similar to what is expected to be found by the bottleneck tool; however, the default settings of the bottleneck tool would ignore this variant because of its high allele frequency in the included germline (normal) tissue. By adjusting the defaults and allowing for a larger initial normal tissue allele frequency using the --maxNorm parameter (or by omitting the normal sample altogether from the analysis using the --samples parameter), we are able to recover that previously reported variant as well, but further filtering would be required to narrow in on this specific mutation. In this manner, OncoGEMINI enjoys further flexibility that enhances its capabilities as a tool for surveying tumor heterogeneity from sequencing data.

OncoGEMINI is an open-source software package, and it is freely available. Source code and further documentation can be found at https://github.com/fakedrtom/oncogemini.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-00854-6.

Additional file 1. : Supplementary Methods for the running of FreeBayes.

Acknowledgements

The authors wish to thank Brady et al. and Savas et al. for their previous work and for making their results available. The authors also acknowledge all those that contributed to the GEMINI tool and other accompanying tools from which OncoGEMINI was developed.

Availability and requirements

Project name: OncoGEMINI
Project home page: https://github.com/fakedrtom/oncogemini
Operating systems: Platform independent
Programming language: Python
License: MIT

Authors' contributions

TJN and ARQ designed and developed OncoGEMINI. TJN and MRC provided code improvement, functional testing, modularization, software packaging, and packaging distribution through conda via the bioconda channel. MRC, XH, YQ, and GTM provided valuable feedback and suggestions for expanded functionality. TJN and ARG wrote the manuscript with support from all authors. All authors read and approved the final manuscript.

Funding

OncoGEMINI is developed and maintained with generous funding from the NIH, National Cancer Institute (U24CA209999), and the NIH, National Institute of General Medical Sciences (R01GM124355).

Availability of data and materials

For the version of OncoGEMINI available at the time of this publication, please refer to the OncoGEMINI citation [5] or use the following link: https://doi.org/10.5281/zenodo.4477434.
The latest developments to OncoGEMINI can be found here https://github.com/fakedrtom/oncogemini.

Ethics approval and consent to participate

This study only utilizes data that has been previously published [26, 27].

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA. [2]Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT 84112, USA. [3]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT 84112, USA.

Received: 23 March 2020 Accepted: 11 February 2021
Published online: 26 March 2021

References

1. Boveri T. The origin of malignant tumors; 1929.
2. Huxley J. Biological aspects of cancer; 1958.
3. Johnson BE, et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. Science. 2014;343:189–93.
4. Greaves M. Evolutionary determinants of cancer. Cancer Discov. 2015;5:806–20.
5. Nicholas TJ. OncoGEMINI. Zenodo. 2021. https://doi.org/10.5281/zenodo.4477434.
6. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol. 2013;9:e1003153.
7. Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
8. Frisk S, et al. Early activating somatic PIK3CA mutations promote ectopic muscle development and upper limb overgrowth. Clin Genet. 2019;96:118–25.
9. Jenkins, M. M. et al. Exome sequencing of family trios from the National Birth Defects Prevention Study: tapping into a rich resource of genetic and environmental data. Birth Defects Res. 2019:1618–32.
10. Cochran JN, et al. Non-coding and loss-of-function coding variants in TET2 are associated with multiple neurodegenerative diseases. bioRxiv. 2019. https://doi.org/10.1101/759621.
11. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. 2016;17:118.
12. Pedersen, B. S. vcf2db. (quinlan-lab, 2019).
13. Landrum MJ, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862–8.
14. Tate JG, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47:D941–7.
15. Douville C, et al. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics. 2013;29:647–8.
16. Ainscough BJ, et al. DoCM: a database of curated mutations in cancer. Nat Methods. 2016;13:806–7.
17. Chakravarty D, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;2017.
18. Tamborero D, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10:25.
19. Griffith M, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49:170–4.
20. Cotto KC, et al. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. Nucleic Acids Res. 2018;46:D1068–73.
21. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015;348:69–74.
22. Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213–9.
23. Kim S, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15:591–4.
24. Koboldt DC, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.
25. Ramos AH, et al. Oncotator: cancer variant annotation tool. Hum Mutat. 2015;36:E2423–9.
26. Nakken S, et al. Personal Cancer Genome Reporter: variant interpretation report for precision oncology. Bioinformatics. 2018;34:1778–80.
27. Brady SW, et al. Combating subclonal evolution of resistant cancer phenotypes. Nat Commun. 2017;8:1231.
28. Savas P, et al. The subclonal architecture of metastatic breast cancer: results from a prospective Community-Based Rapid Autopsy Program "CASCADE". PLoS Med. 2016;13:e1002204.
29. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms. SnpEff Fly (Austin). 2012;6:80–92.
30. McLaren W, et al. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
31. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio (2012).
32. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44:e131.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC