ZygosityPredictor

Marco Rheinnecker[1]*, Martina Fröhlich[1,2]*, Marc Rübsam[1]*, Nagarajan Paramasivam[1], Christoph E. Heilig[2,3], Stefan Fröhling[2,3], Richard F. Schlenk[2,4,5,6], Barbara Hutter[1,2], Daniel Hübschmann[1,2,7]

[1] Computational Oncology Group, Molecular Precision Oncology Program, National Center for Tumor Diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ) Heidelberg
[2] German Cancer Consortium (DKTK), Heidelberg
[3] Department of Translational Medical Oncology, NCT Heidelberg and DKFZ Heidelberg
[4] Department of Medical Oncology, NCT Heidelberg, Heidelberg University Hospital, Heidelberg, Germany;
[5] Department of Hematology, Oncology and Rheumatology, Heidelberg University Hospital, Heidelberg, Germany;
[6] NCT Trial Center, NCT Heidelberg, Heidelberg University Hospital and DKFZ, Heidelberg, Germany.
[7] Pattern Recognition and Digital Medicine Group, Heidelberg Institute for Stem cell Technology and Experimental Medicine (HI-STEM), Heidelberg
* Shared first authorship.

Correspondence should be addressed to:
Daniel Hübschmann
d.huebschmann@dkfz-heidelberg.de

Keywords

zygosity, next generation sequencing, variant allele frequency, variant calling

Abstract

Summary

ZygosityPredictor provides functionality to evaluate how many copies of a gene are affected by mutations in next generation sequencing data. In cancer samples, both somatic and germline mutations are accurately processed. In particular, ZygosityPredictor computes the number of affected copies for single nucleotide variants and small insertions and deletions (Indels). In addition, the tool integrates information at gene level by performing haplotype phasing and subsequent logic to derive how strongly a gene is affected by mutations. This

information is of particular interest in precision oncology, e.g. when assessing whether unmutated copies of tumor-suppressor genes remain.

Availability and implementation

ZygosityPredictor was implemented as an R-package and is available via Bioconductor at https://bioconductor.org/packages/ZygosityPredictor. Detailed documentation is provided in the vignette including application to an example genome.

Introduction

Variant calling is the process of detecting mutations in data obtained from sequencing nucleic acids in a biological sample (Hübschmann and Schlesner, 2019). When working with short read next generation sequencing (NGS), variant calling is performed downstream of alignment of the reads to the reference genome of the respective organism. Variant calling can be performed for both germline variants, which are present in all cells of an organism, and for somatic variants, which occurred during the life span of the organism and which are present only in clonally expanded cells such as those found in neoplasms. Variant calling of different mutation types requires different algorithms and pipelines or workflows, including single nucleotide polymorphisms (SNPs) / single nucleotide variants (SNVs), small insertions and deletions (Indels), copy number variations (CNVs), somatic copy number aberrations (sCNAs) and structural variants (SVs) (Hübschmann and Schlesner, 2019; Reisinger *et al.*, 2017). In the following, we refer to SNVs and Indels together by the technical term *small variants*.

Especially when analyzing cancer samples, variant calling is strongly affected by the purity, i.e., the tumor cell content, and the ploidy of the sample. For all mutation types, variant calling is strongly influenced by the choice of algorithm and/or pipeline used for calling the variant. If applied to patients in a precision oncology setting, this choice has strong consequences on downstream processing, including identification of driver gene aberrations and targetable lesions, and on clinical decision making (Garcia-Prieto *et al.*, 2022).

Small variant calling is almost always based on criteria of read counts or allele frequency for alternative alleles. It therefore is modulated by the copy number state of the genomic segment the called small variant is located in, as well as many other factors, e.g., purity. However, many strategies for small variant calling don't make use of the local copy number information.

For correct interpretation of the functional state of a mutated gene, it is important to determine how many copies of that gene are affected by a mutation ("affected copies", "*ac*"). In particular, it is of high relevance if all copies are affected (bi-allelic mutations in the diploid case) or if unmutated copies remain (mono-allelic mutations in the diploid case). Furthermore, if a gene is affected by more than one mutation, e.g. in the case of compound heterozygosity, it is important to phase the mutations, i.e., to determine whether they affect the same allele or different ones.

Here, we present ZygosityPredictor, an R/Bioconductor package with functionality for the computation of the number of affected copies (downstream of CNV and sCNA calling) for both germline and somatic mutations as well as for phasing of small variants. Phasing is

performed on DNA sequencing data and, if available, also on RNA-seq data, which enables the creation of larger haplo-blocks if reads span exon boundaries and splice sites. These data are integrated in order to provide information if all copies are affected (bi-allelic mutations in the diploid case) or if unmutated copies remain (mono-allelic mutations in the diploid case) for all genes.

Methods and Results

ZygosityPredictor is to be executed after small variant calling (germline and somatic) (Horak *et al.*, 2021; Campbell *et al.*, 2020) and after sCNA calling (Kleinheinz *et al.*, 2017). It takes the variant calls as input. By integrating this information at gene level, ZygosityPredictor computes if all copies are affected (bi-allelic mutations in the diploid case) or if unmutated copies remain (mono-allelic mutations in the diploid case). For the integration at gene level, the variants need to be annotated with gene names. This is often performed downstream of variant calling. The functionality used by ZygosityPredictor needs to employ the same gene model as the one used for annotation of the variant calls (input `geneModel` to the function `predict_zygosity()`).
Of note, the functionality of ZygosityPredictor is also applicable in cases with higher ploidy than 2, which is often the case in tumor samples. The analysis is performed in two steps: (i) assessment at variant level including the calculation of affected copies for each input variant, and (ii) aggregation at gene level achieved by integrating all data layers and phasing several variants if necessary. Ultimately, every gene that carries mutations is classified in one of the following categories:

- “all copies affected”
- “wt copies left”

Assessment per variant: calculation of affected copies

To determine the contribution of each variant to the affection of a gene, it is necessary to estimate how many copies are affected by the mutation. The different input data layers are handled differently.
Of note, we make the assumption that the variant calls and copy number profiles are provided for a clonal tumor, i.e. we assume that the cancer cell fraction (CCF), i.e., the fraction of the tumor cells that carry the variant, is 1 (Satas *et al.*, 2021).
For small variants, the number of affected copies is computed according to formulae (4) and (8) below. Let ac_{tumor} define the number of affected copies in the tumor ($ac_{tumor}^{somatic}$ for somatic variants and $ac_{tumor}^{germline}$ for germline variants), c_{tumor} the copy number in the tumor of the genomic segment the variant is located in and $VAF_{tumor}^{somatic}$ the allele-frequency of the somatic variant in the tumor sample. Let n_{tumor} and n_{normal} designate the number of tumor and normal cells in the sample, respectively. Then, in case of a 100% pure tumor, the allele frequency of a somatic variant is defined as:

$$VAF_{tumor}^{somatic} = \frac{ac_{tumor}^{somatic} \times n_{tumor}}{c_{tumor} \times n_{tumor}} = \frac{ac_{tumor}^{somatic}}{c_{tumor}} \quad (1)$$

In case the purity is not 100%, the formula needs to be extended (Figure 1a). Let c_{normal} define the expected copy number in normal tissue of the segment the variant is located in. This will be 2 for most genes located on autosomes and for genes on the X chromosome in female samples, and 1 for genes on the X and Y chromosomes in male samples (without the pseudoautosomal regions).
Formula (1) can then be extended as follows:

$$VAF_{tumor}^{somatic} = \frac{ac_{tumor}^{somatic} \times n_{tumor}}{c_{tumor} \times n_{tumor} + c_{normal} \times n_{normal}} \quad (2)$$

With the definition of $p = \frac{n_{tumor}}{n_{tumor} + n_{normal}}$ as the purity of the tumor, we can deduce $n_{normal} = n_{tumor} \times \frac{1-p}{p}$ and:

$$VAF_{tumor}^{somatic} = \frac{ac_{tumor}^{somatic}}{c_{tumor} + c_{normal} \times (\frac{1}{p} - 1)} \quad (3)$$

and thus

$$ac_{tumor}^{somatic} = VAF_{tumor}^{somatic} \times (c_{tumor} + c_{normal} \times (\frac{1}{p} - 1)) \quad (4)$$

In the tumor sample, the allele frequency of germline variants can be defined as (Figure 1b):

$$VAF_{tumor}^{germline} = \frac{ac_{tumor}^{germline} \times n_{tumor} + ac_{normal}^{germline} \times n_{normal}}{c_{tumor} \times n_{tumor} + c_{normal} \times n_{normal}} \quad (5)$$

Of note, $ac_{tumor}^{germline}$ and $ac_{normal}^{germline}$ define numbers of affected copies for the germline variant in tumor and normal tissue, respectively. Using $VAF_{normal}^{germline}$, the allele-frequency of the germline variant in the normal control, we can substitute $ac_{normal}^{germline} = VAF_{normal}^{germline} \times c_{normal}$ and obtain:

$$VAF_{tumor}^{germline} = \frac{ac_{tumor}^{germline} \times n_{tumor} + VAF_{normal}^{germline} \times c_{normal} \times n_{normal}}{c_{tumor} \times n_{tumor} + c_{normal} \times n_{normal}} \quad (6)$$

When applying the same arithmetics as above, we get

$$VAF_{tumor}^{germline} = \frac{ac_{tumor}^{germline} + VAF_{normal}^{germline} \times c_{normal} \times (\frac{1}{p} - 1)}{c_{tumor} + c_{normal} \times (\frac{1}{p} - 1)} \quad (7)$$

and thus

$$ac_{tumor}^{germline} = VAF_{tumor}^{germline} \times c_{tumor} + (VAF_{tumor}^{germline} - VAF_{normal}^{germline}) \times c_{normal} \times (\frac{1}{p} - 1) \quad (8)$$

In case $VAF_{normal}^{germline}$ is unknown, it can be assumed to be 0.5 for heterozygous variants in most genes located on autosomes or on the X chromosome in female samples, and 1 for

homozygous variants and variants in genes on the X and Y chromosomes in male samples (without the pseudoautosomal regions), for a normal chromosome set.

Subtracting the affected copy number $ac_{tumor}^{somatic}$ or $ac_{tumor}^{germline}$, obtained through the functions `aff_som_copies()` and `aff_germ_copies()`, from c_{tumor} yields the number of wt copies at this position. ZygosityPredictor uses this information to evaluate the status of a variant in the sample as follows. For classification of somatic small variants into the categories “all copies affected” and “wt copies left”, we check whether LOH is present at the genomic position of the variant. Only if that is the case, the number of affected copies calculated by Formula (4), implemented in the function `aff_som_copies()` in ZygosityPredictor, is used to assign the final status. If more than 0.5 wt-copies remain, the variant is attributed the status “wt copies left”, otherwise “all copies affected”. If no LOH is detected at the position of the variant, we always assign the status “wt copies left”. This is motivated by the fact that pipelines used for detection of sCNAs already apply sophisticated logic for the assignment of the copy number and/or LOH state of every segment.

For germline variants, we also first check whether they are affected by LOH. If LOH is present, it must be checked if it is the mutated or unmutated allele which is exclusively present in the tumor. This check is performed by assessing the allele frequency of the variant in the tumor. If the latter is greater than 0.5, it is assumed to be present in the tumor and, since the other allele is completely deleted, it affects all other copies as well and will be termed “all copies affected”. If the allele frequency is below 0.5, the variant is removed from all subsequent steps, as it is assumed that it is no longer present in the tumor. Germline variants in which no LOH was detected are labeled "wt copies left". Beside small variants, ZygosityPredictor can also take into consideration large deletions (i.e., deletions that exceed the size of Indels, i.e., > 50 base pairs). They need to be provided by the sCNA calling. A homozygous deletion always leads to “all copies affected”.

Until this point, we have estimated for all variants how many copies they affect. Variants that do not affect all copies, i.e., which leave intact wildtype copies, can, in the case of compound heterozygosity with another variant in the same gene, still contribute to a constellation in which all copies of the gene are affected by mutations. For such an evaluation at gene level, the affected copies are also needed for the heterozygous germline variants, which are calculated according to formula (8) and the function `aff_germ_copies()`. For more detailed information on usage and functionality we refer to the vignette of the package available at https://bioconductor.org/packages/ZygosityPredictor.

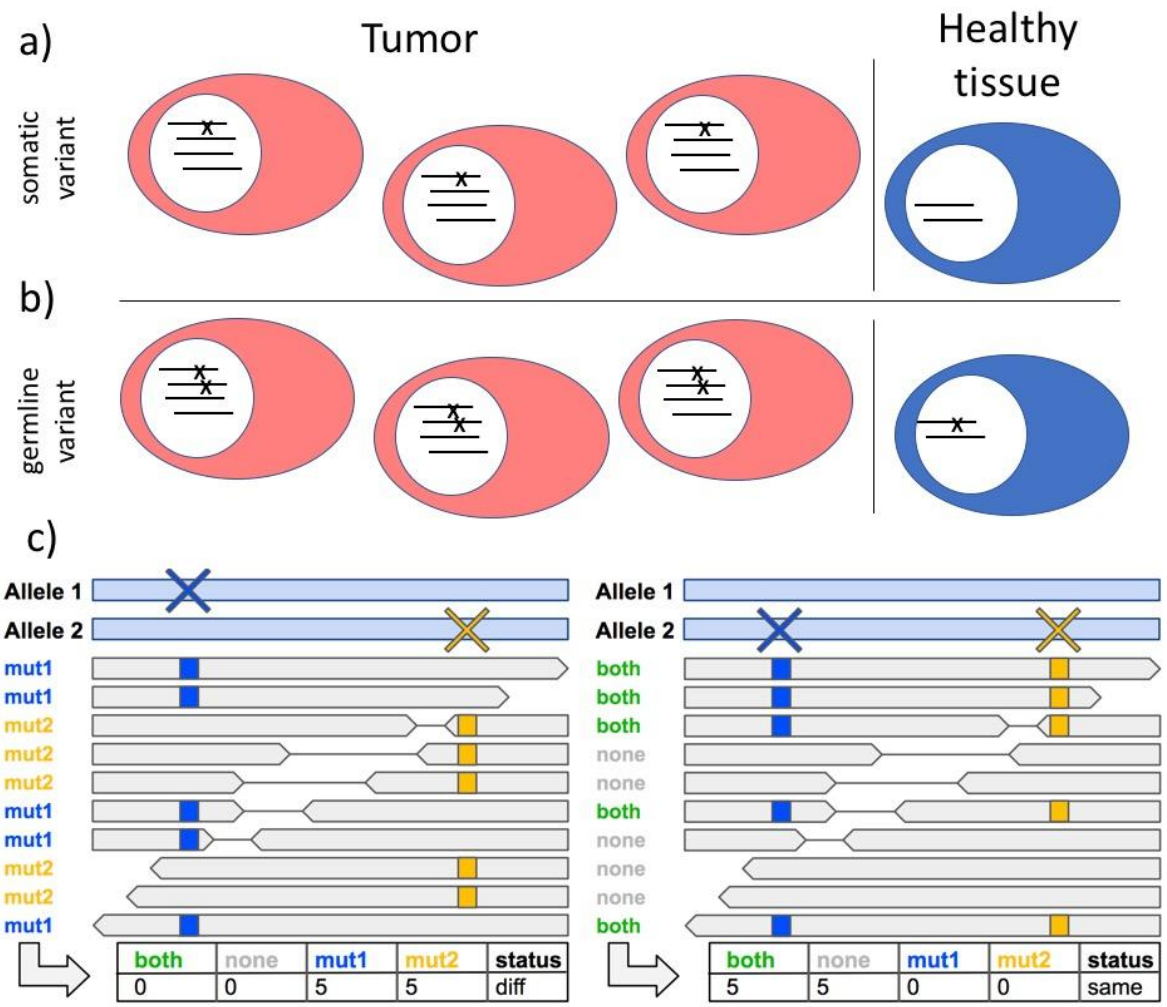

Figure 1: Schematic for determination of the number of affected copies for a somatic or germline mutation in a tumor sample. Tumor cells are depicted in red, admixed healthy cells are depicted in blue. (a) Example of a somatic mutation in a sample with a tumor purity of $p = \frac{3}{4} = 75\%$. The tumor cells carry a somatic mutation, depicted as *x*, which is located in a genomic segment with a copy number of $c_{tumor} = 4$ in the tumor and $c_{normal} = 2$ in the normal tissue. The number of tumor cells carrying the variant is $n_{tumor} = 3$ and the number of admixed healthy cells is $n_{normal} = 1$. As can be counted in this example, the total number of copies present for this genomic segment is $4 \times 3 + 2 \times 1 = 14$ and the observed allele-frequency of the variant in the sample would be $VAF_{tumor}^{somatic} = \frac{3}{14}$, i.e. 3 of the 14 total copies are affected in the sample. Using formula (4) it can be calculated that $ac_{tumor}^{somatic} = \frac{3}{14} \times (4 + 2 \times (\frac{4}{3} - 1)) = 1$, which aligns with the observed number in the example. (b) Analogous example of a germline mutation, taking into consideration the number of affected copies at the genomic position of the variant in tumor $ac_{tumor}^{germline} = 2$ and normal tissue $ac_{normal}^{germline} = 1$. The observed allele-frequency of the variant in the normal control would be $VAF_{normal}^{germline} = \frac{1}{2}$, i.e. 1 of the 2 are affected in a pure normal control, and in the tumor sample would be $VAF_{tumor}^{germline} = \frac{7}{14}$, i.e. 7 of the 14 total copies are affected in the sample.

Using formula (8) it can be calculated that $ac_{tumor}^{germline} = \frac{7}{14} \times 4 + (\frac{7}{14} - \frac{1}{2}) \times 2 \times (\frac{4}{3} - 1) = 2$, which aligns with the observed number in the example. (c) Schematic for haplotype phasing in a diploid genome with two variants in a region of interest. On the left, the variants are located on different alleles; on the right variants are located on the same allele and therefore leave one wt allele intact. Below, the repercussions on NGS reads / read-pairs are shown. For the left example, reads overlapping both variant positions carry exclusively only one of the two variants. In the right example they carry both or none of them. Based on this, each read / read-pair can be classified into one of the four categories: *both*, *none*, *mut1* or *mut2*. By summing them up a final status label (*diff* or *same*) can be attributed to the variant pair.

Assessment per gene: haplotype phasing

The second step, carried out by the function `predict_zygosity()`, is the integration at gene level to assess whether, considering all variants affecting a certain gene, any wildtype copies are left. To this end, all variants of a given gene are first checked if any one affects all copies. In such a case the final status of “all copies affected” is assigned for that gene. If no such variant is found, the next step is to check how many variants are in the gene. If there is only one variant, the status “wt copies left” will be assigned to the gene. In the case of several variants, none of which individually leads to "all copies affected", a phasing of these variants has to be carried out in order to find out whether they affect different alleles. For this purpose, we use reads / read-pairs from DNA-sequencing and, if available, RNA-sequencing, which map to the positions of both variants. If there are reads/read-pairs which simultaneously carry both variants, it can be concluded that they originate from the same copy, which in the software is labeled by the phasing status “same”. If no reads / read-pairs carrying both variants are present, the variants most likely originate from different copies (Figure 1 II). As there is still the possibility of present wildtype copies even if the variants are located on different reads / read pairs, e.g. due to early duplications occurring before the somatic small variant, the phasing status can not yet be finally assigned. In this case, the number of copies affected by variant 1, ac^1, and the number of copies affected by variant 2, ac^2, are added ($ac^1 + ac^2$). Only if this leads to a number of left wt copies below 0.5, it is concluded that the two variants affect all copies. With this type of phasing, the influence of maximum two variants can be checked together. In the case of more than two variants in one gene, all pairwise combinations need to be tested. If at least one combination leads to “all copies affected”, this status is also attributed to the gene, otherwise the gene is assigned “wt copies left”. In haplotype phasing, it is possible that no conclusion can be drawn because, e.g., the variants may be too far apart, so that no overlapping reads are present, or RNA reads do not cover the relevant exon due to alternative splicing. In such cases, ZygosityPredictor can use SNPs to try an “extended” phasing. In this process, an attempt is made to establish a mapped connection on the basis of SNPs located between the two mutations. An attempt is made to shimmy along the SNPs, since the probability of finding overlapping reads at smaller intervals is increased. However, since there is still the possibility of not finding SNPs between the mutations, especially in exonic regions, or not being able to phase them, cases in which no conclusion can be drawn remain. Then, the more conservative statement “wt copies left” is made for the gene and the respective information is annotated in the info column of the output of `predict_zygosity()`. Detailed information about the steps performed by ZygosityPredictor during the evaluation of the

genes can be taken from the optional command prompt output. For more detailed information on usage and functionality we refer to the vignette of the package available at https://bioconductor.org/packages/ZygosityPredictor.

Conclusion

We developed the R/Bioconductor package ZygosityPredictor which downstream of variant calling computes the number of affected copies for every gene, taking into consideration both somatic and germline variants. In particular, it computes the number of affected copies for so-called small variants (SNVs and Indels). Then, in a second step the tool performs haplotype phasing based on DNA and, if available also RNA sequencing data. ZygosityPredictor then finally assigns categories of whether functional copies of a gene remain, which is of particular use, e.g. in precision oncology for the assessment of tumor-suppressor genes.

Funding

This work was supported by the National Center for Tumor Diseases (NCT) Molecular Precision Oncology Program. MRu was funded by the DKTK project MARRIAGE. MRh was funded by allowances for scientific assistant from the graduate school HIDSS4health.

Bibliography

Campbell,P.J. *et al.* (2020) Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.

Garcia-Prieto,C.A. *et al.* (2022) Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools. *Bioinformatics*, **38**, 3181–3191.

Horak,P. *et al.* (2021) Comprehensive Genomic and Transcriptomic Analysis for Guiding Therapeutic Decisions in Patients with Rare Cancers. *Cancer Discov.*, candisc.0126.2021.

Hübschmann,D. and Schlesner,M. (2019) Evaluation of whole genome sequencing data. In, *Methods in Molecular Biology*. Humana Press Inc., pp. 321–336.

Kleinheinz,K. *et al.* (2017) ACEseq - allele specific copy number estimation from whole genome sequencing. *bioRxiv*.

Reisinger,E. *et al.* (2017) OTP: An automatized system for managing and processing NGS data. *J. Biotechnol.*, **261**, 53–62.

Satas,G. *et al.* (2021) DeCiFering the elusive cancer cell fraction in tumor heterogeneity and evolution. *Cell Syst.*, **12**, 1004-1018.e10.