BITS :: Call for Abstracts 2021 - Oral communication

Type	Oral communication
Session	Sequencing and genotyping technologies
Title	A novel computational framework for the identification of the TD-plus phenotype in high grade serous ovarian cancer
All Authors	Sergi A(1), Beltrame L(2), Paracchini L(2), Venturini L(2), D'Incalci M(3), Marchini S(2), Masseroli M(1)

Affiliation

(1) Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, via Ponzio 34/5, 20133, Milan, Italy
(2) IRCSS Humanitas Research Hospital -, Via Manzoni 56, 20089 Rozzano, Milan, Italy
(3) Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20090 Pieve Emanuele, Milan, Italy

Motivation

The Tandem Duplicator Phenotype (TDP) is a genomic instability configuration present in several tumors. Popova et al. [1] reported a particular TDP, called TDP-plus, in some ovarian cancers, associated with a large number of somatic TDs (200-800 TDs up to 10 Mbp) and with a dysfunctional (when the CDK12 gene is mutated) or functional (when the CCNE1 gene is amplified) Homologous Recombination (HR) pathway. Moreover, Menghi et al. [2] classified TDP into three intervals, depending on TDs length: TDs with ~11 kbp are associated with loss of TP53 and BRCA1 genes, while TDs with ~231 kbp and ~1.7 Mbp are associated with CCNE1 gene amplification and CDK12 gene inactivation.
Deficiency of the homologous recombination mechanism (HRD) is associated with hypersensitivity to DNA-damaging agents and Poly(ADP-ribose) polymerase inhibitors (PARPi); taking advantage of the synthetic lethality mechanism, PARPi are used as single-agent therapeutics and as maintenance treatment, following platinum-based chemotherapy, in High Grade Serous Epithelial Ovarian Cancers (HGS-EOC).
Thus, the TDP-plus should be identified whenever possible, in order to choose the correct treatment. However, actual methods do not take into account the presence of this genomic alteration. Moreover, no formalized approaches to identify the TDP-plus currently exist.
To overcome these limitations, we developed a novel integrated framework, able to identify, filter and select putative somatic variants associated with the TDP and calculate purity and ploidy of the tumor sample. Alongside these mutational information, our framework is able to work on the structural side: TDs are identified, filtered and quantified based on their length, and the output is merged with the previous mutational information to correctly identify the TDP-plus.

Methods

-Sequencing and pre-processing of the samples
109 HGS-EOC samples were sequenced with NextSeq 500 and a custom library covering 387 genes of clinical interest for HGS-EOC and 12 Mbp of structural regions spanning the whole genome ('backbone') was constructed from genomic DNA. Raw FASTQ files are then aligned to the reference genome (hg19) with the BWA [3].

-Variant calling
Calling of somatic variants was performed with vardict [4]. VCF files are assembled together, annotated (with vcfanno [5] and VEP [6]) and used to build a GEMINI-compatible database [7] using vcf2db [8]. Database preparation and queries are done with a modified version of the variantdb[9] Python package, that queries the database using the SQLalchemy framework [10].
Variants arising from the genes involved in the TDP are filtered using an in house Python algorithm that removes variants according to the following filters:
-coverage < 60X;
-variant allele fraction (VAF) lower than 10%
-low or no impact on the protein
-variants with a known population frequency < 1% (combined data from GNOMAD [11] and 1000 genomes [12])
-variants clinically validated as 'benign' or 'likely benign' in ClinVar [13]
-suspected artefacts (frequency > 60% above all samples).
Remaining variants are then validated through the CGI database [14], using another in-house algorithm. Only variants marked as cancer ‘drivers’ are kept. These include known driver mutations from the literature, "tier 1" (strong probability of being drivers) and "tier 2" (medium probability of being drivers) grade

mutations.

-Purity and ploidy estimation
Estimation of purity and ploidy for each tumor sample and copy-number variation along the genome is performed with PureCN [15]. Data thus obtained is merged with the previous mutational dataset, generating a combined dataset containing purity, copy number and loss of heterozygosity (LOH) information. Only variants flagged as LOH (meaning the loss of the other, wild-type, allele) are kept, considering that the HR pathway deficiency occurs with the loss of one (or more) protein involved in the pathway.

-Identification of structural variations
Structural variations are extracted using Manta [16] and GRIDSS [17]. Output VCF files are processed with our algorithm, which is based on the cyvcf2 [18] Python package. TDs are extracted and filtered based on their length (up to 10 Mbp, according to [1]) and results are merged with the mutational information previously obtained.

Results

We reported a large number (dozens) of TDs in CDK12-mutated samples (3 out of 109 samples) with a length of ~1Mb, compared to wild type samples, which presented only a few TDs in the interval, in accordance with the results obtained from [1] and [2]. This led us to a first confirmation of the reliability of our framework. We are currently working on the algorithm to be able to differentiate the three different TD intervals reported by [2] more accurately, and to further merge outputs with the other TDP-related genes. Our gold aim is to create a single, integrated framework capable of clustering (given a set of samples) the three TDs intervals alongside the related mutational information.

Info

Bibliography
[1] Popova T, Manié E, Boeva V, Battistella A, Goundiam O, Smith NK, Mueller CR, Raynal V, Mariani O, Sastre-Garau X, Stern MH. Ovarian Cancers Harboring Inactivating Mutations in CDK12 Display a Distinct Genomic Instability Pattern Characterized by Large Tandem Duplications. Cancer Res. 2016; 76(7): 1882-1891. doi: 10.1158/0008-5472.CAN-15-2128. PMID: 26787835.
[2] Menghi F, Barthel FP, Yadav V, Tang M, Ji B, Tang Z, Carter GW, Ruan Y, Scully R, Verhaak RGW, Jonkers J, Liu ET. The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. Cancer Cell. 2018; 34(2): 197-210.e5. doi: 10.1016/j.ccell.2018.06.008. PMID: 30017478.
[3] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17(1): 122.
[4] Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016; 44(11): e108. doi: 10.1093/nar/gkw227. PMID: 27060149.
[5] Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. 2016; 17(1): 118. doi: 10.1186/s13059-016-0973-5. PMID: 27250555.
[6] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17(1): 122. doi: 10.1186/s13059-016-0974-4. PMID: 27268795.
[7] Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. PLOS Comput. Biol. 2013; 9(7): e1003153.
[8] Paila U, Chapman BA, Kirchner R, Quinlan AR. vcf2db. https://github.com/quinlan-lab/vcf2db
[9] Vandeweyer, G., Van Laer, L., Loeys, B. et al. VariantDB: a flexible annotation and filtering portal for next generation sequencing data. Genome Med 6, 74 (2014). https://doi.org/10.1186/s13073-014-0074-6
[10] Bayer M. SQLAlchemy. In Brown A, Wilson G, editors. The Architecture of Open Source Applications. Volume II: Structure, Scale, and a Few More Fearless Hacks. 2012. http://aosabook.org
[11] Karczewski, K.J., Francioli, L.C., Tiao, G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7
[12] The 1000 Genomes Project Consortium., Corresponding authors., Auton, A. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). https://doi.org/10.1038/nature15393
[13] Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D1067. doi:10.1093/nar/gkx1153
[14] Tamborero D, Rubio-Perez C, Deu-Pons J. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018; 10, 25. https://doi.org/10.1186/s13073-018-0531-8
[15] Riester M, Singh AP, Brannon AR, Yu K, Campbell CD, Chiang DY, Morrissey MP. PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med. 2016; 11: 13. doi: 10.1186/s13029-016-0060-z. PMID: 27999612.
[16] Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S,

Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016; 32(8): 1220-1222. doi: 10.1093/bioinformatics/btv710. PMID: 26647377.
[17] Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 2017;27(12):2050-2060. doi: 10.1101/gr.222109.117. PMID: 29097403.
[18] Pedersen BS, Quinlan AR. cyvcf2: fast, flexible variant analysis with Python, Bioinformatics, Volume 33, Issue 12, 2017, Pages 1867–1869. https://doi.org/10.1093/bioinformatics/btx057

Figure	
-	
Availability	-

Corresponding Author

Name, Surname	Aldo, Sergi
Email	aldo.sergi@polimi.it

Submitted on	09.05.2021