Cancer Predisposition Sequencing Reporter (CPSR): a flexible variant report engine for germline screening in cancer

Authors: Sigve Nakken[1,2], Vladislav Saveliev[3], Oliver Hofmann[3], Pål Møller[1], Ola Myklebost[1,4,5], and Eivind Hovig[1,5]

Affiliations: [1]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Norway

[2]Centre for Cancer Cell Reprogramming, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Norway

[3]Centre for Cancer Research, University of Melbourne, Australia

[3]Department of Clinical Science, University of Bergen, Norway

[4]Western Norway Familial Cancer Center, Haukeland University Hospital, Bergen, Norway

[5]Centre for Bioinformatics, Department of Informatics, University of Oslo, Norway

Abstract

Motivation: While somatic mutagenesis is the driving force of most human cancers, the germline genome is of significant clinical value in several tumor types. Cancer predisposition variants are important for risk management and surveillance, and can also have major implications for treatment strategies since many are in DNA repair genes. Following the incorporation of high-throughput DNA sequencing in cancer clinics and research, there is thus a need to provide clinically oriented sequencing reports for risk-associated germline variants and their potential therapeutic relevance on a per patient basis.

Results: We have developed the Cancer Predisposition Sequencing Reporter (CPSR), an open-source computational workflow that provides a structured report of germline variants identified in known cancer predisposition genes. Building upon existing knowledge sources and variant databases relevant for cancer susceptibility, CPSR combines a transparent and cancer-dedicated scoring scheme for variant pathogenicity (American College of Medical Genetics and Genomics, ACMG) with existing variant classifications from ClinVar in order to derive a structured and prioritised list of variant findings. The workflow outputs a comprehensive and interactive HTML report that highlights putative markers of therapeutic, prognostic and diagnostic relevance. Importantly, the set of cancer predisposition genes profiled in the report can be flexibly chosen from nearly 40 virtual gene panels established by scientific experts, enabling a customization of the report for different screening purposes. The report can be configured to also list potential incidental variant findings as recommended by ACMG, as well as the status of low-risk variants from genome-wide association studies in cancer.

Availability and Implementation: The software is implemented in Python/R, and is freely available through Docker technology. Documentation, example reports, and installation instructions are accessible via the project GitHub page: https://github.com/sigven/cpsr

Contact: sigven@ifi.uio.no

1 Introduction

A considerable fraction of human cancers are rooted in rare pathogenic germline mutations in cancer predisposition genes (Huang *et al.*, 2018). Screening of cancer patients for predisposing germline alterations may yield valuable decision support for risk-reducing interventions and surveillance, and has also proven its significance for the application of platinum-based chemotherapy and targeted drugs (Thavaneswaran *et al.*, 2019).

High-throughput screening for a broad collection of cancer predisposition genes is currently feasible due to technological advances in genome-wide DNA sequencing. However, there is to our knowledge no bioinformatics tool that can transform raw sequencing results to structured and interactive reports for clinical interpretation on a per patient basis. Previous efforts have mostly focused on the implementation of variant pathogenicity predictions according to criteria defined by the American College of Medical Genetics and Genomics (ACMG). Thus, general-purpose command-line tools for variant classification according to ACMG guidelines are offered by both InterVar and CharGer (Scott et al., 2019; Li and Wang, 2017). A comprehensive refinement of the ACMG criteria was outlined in SherLoc, which captured important edge cases and exceptions in clinical genetics (Nykamp et al., 2017). With the understanding that different pathogenicity criteria may need novel or adjusted implementations for a given disease phenotype, a cancer-focused solution was recently made available through PathoMAN (Ravichandran *et al.*, 2019). The limited web-based service offered by PathoMAN is, however, inconvenient for integration in high-throughput analysis environments.

Here, we extend previous approaches with a flexible cancer predisposition interpretation workflow, coined the *Cancer Predisposition Sequencing Reporter* (CPSR). Technically, CPSR builds upon the framework developed for the Personal Cancer Genome Reporter (Nakken *et al.*, 2018). It is intended for integration with standard variant calling output from whole-genome, exome or targeted sequencing, accepting germline variant calls encoded in the VCF format as input. CPSR can furthermore target the analysis and report towards risk genes associated with a particular cancer type or syndrome. The workflow implements a cancer-dedicated refinement of ACMG criteria in order to classify variants according to pathogenicity, and produces a structured and interactive predisposition report that highlights variants with therapeutic implications.

2 *CPSR implementation and functionality*

We have previously developed the Personal Cancer Genome Reporter (PCGR) for the analysis and clinical interpretation of acquired aberrations in a given tumor. Expanding upon this framework, we have now implemented a dedicated workflow for the interpretation of germline variants related to cancer susceptibility and inherited cancer syndromes. Central tools and knowledge resources that are built into the CPSR workflow are depicted in Figure 1.

In order to serve a wide range of clinical cases, CPSR can produce variant reports that are dedicated towards predisposition genes for specific tumor types or cancer syndromes. Here, we exploit virtual gene panels as available from the Genomics England PanelApp, a crowdsourcing initiative in which scientific experts are evaluating risk genes for nearly 40 different cancer phenotypes on a continuous basis (Martin *et al.*, 2019). In the second step of the workflow variant consequences are determined by Variant Effect Predictor (VEP), using GENCODE as the gene reference model. Notably, variants with a putative loss-of-function consequence (i.e. stopgain, frameshift and splice site disruption) are subject to careful evaluation and filtering through the LOFTEE plugin in VEP. Specifically, LOFTEE assigns confidence to a loss-of-function variant based on multiple features, such as transcript location, ancestral allele state, and intron size and donor site nature (for splice site mutations). Through the use of *vcfanno*, the second workflow step will also query the variant set against up-to-date knowledge resources of relevance for cancer predisposition and functional variant effect. These resources include pre-classified variants in ClinVar, population allele frequencies (gnomAD), known mutational hotspots in cancer, precomputed *insilico* deleteriousness predictions (dbNSFP), low-risk risk alleles identified from genome-wide association studies of cancer phenotypes, and most importantly, biomarkers of relevance for prognosis, diagnosis or therapeutic regimens (CIViC). Notably, the CPSR data bundle relies upon open-access resources that can be freely distributed, with the implication that locus-specific databases from InSiGHT (MMR) and IARC (TP53) are not included.

Based on the set of annotations retrieved in step two, CPSR conducts a standard five-level variant pathogenicity classification to aid the interpretation of disease-causing variants (Plon *et al.*, 2008). The classification algorithm performed in step three constitutes an open-source implementation of a comprehensive and refined list of ACMG criteria, most of which were outlined in SherLoc (Nykamp *et al.*, 2017). In short, evidence scores that support a pathogenic nature are accumulated alongside evidence scores that support a benign nature, ultimately producing a final score that falls within five predefined ranges or levels of pathogenicity (Supplementary Materials and Supplementary Table S1).

The final step of the workflow exploits the R Markdown framework to display all variants findings in a structured and interactive HTML report (Allaire *et al.*, 2019). Additional output formats are also available to the user, i.e annotated VCF, JSON, and TSV. An important asset of the predisposition report is the ability to filter the result sets for various types of annotations, e.g. population frequency, consequence type, or existing phenotype associations. For variants of uncertain significance, which frequently makes up the largest group of variants, the report specifically enables the use of the quantitative pathogenicity score to prioritise potential borderline cases. To cater for reproducibility and transparency, the report is also populated with a complete documentation section, indicating the configuration settings for how the report was run, and the versions of all tools and databases that are being used (an example report can be found as Supplementary File 1).

3 CPSR classification performance

In order to evaluate the performance of our automated variant classification procedure, we utilized a dataset of variants reported in TCGA's PanCancer germline study (Huang *et al.*, 2018). Specifically, we assessed the degree to which CharGer and CPSR classifications agreed with those provided with high confidence in ClinVar (n = 329). Although CPSR showed lower concordance than CharGer for pathogenic/likely pathogenic variants in this testset (91.4% versus 98.6%, n = 290), the concordance for variants of uncertain significance (VUS) was significantly better with CPSR (94.7% versus 0%, n = 39). Compared with CharGer, these results reflect a more conservative classification algorithm in CPSR that is likely predicting fewer false positive P/LP variants and with an improved balance when it comes to the separation of P/LP and VUS (Supplementary Materials, Supplementary Figure S1, and Supplementary Table S2).

4 Conclusion

We believe that joint consideration of the germline and somatic mutation landscape of each cancer patient will make up a central component of cancer precision medicine. The workflow offered by the Cancer Predisposition Sequencing Reporter provides in this respect an efficient way to navigate and explore the clinical utility of the germline genome. A future version may also include important pharma- and radiogenomic risk variants.

Acknowledgements

This work was supported by grant #221580 from the Norwegian Research Council to the Norwegian Cancer Genomics Consortium.

References

Allaire,J.J. *et al.* (2019) Rmarkdown: Dynamic Documents for R.

Huang,K.-L. *et al.* (2018) Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*, **173**, 355–370.e14.

Li,Q. and Wang,K. (2017) InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am. J. Hum. Genet.*, **100**, 267–280.

Martin,A.R. *et al.* (2019) PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.*, **51**, 1560–1565.

Nakken,S. *et al.* (2018) Personal Cancer Genome Reporter: variant interpretation report for precision oncology. *Bioinformatics*, **34**, 1778–1780.

Nykamp,K. *et al.* (2017) Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria. *Genet. Med.*, **19**, 1105.

Plon,S.E. *et al.* (2008) Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum. Mutat.*, **29**, 1282–1291.

Ravichandran,V. *et al.* (2019) Toward automation of germline variant curation in clinical cancer genetics. *Genet. Med.*, **21**, 2116–2125.

Scott,A.D. *et al.* (2019) CharGer: clinical Characterization of Germline variants. *Bioinformatics*, **35**, 865–867.

Thavaneswaran,S. *et al.* (2019) Therapeutic implications of germline genetic findings in cancer. *Nat. Rev. Clin. Oncol.*, **16**, 386–396.

Figures

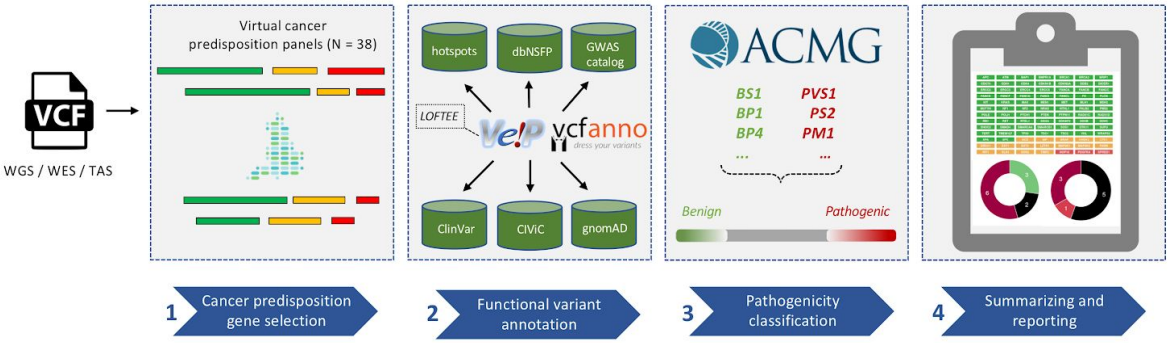

Figure 1: CPSR workflow with key databases and tools, illustrating how the query variant set from germline variant calling (formatted as VCF) is subject to four main steps for predisposition interpretation. Locus filtering against a selected cancer predisposition gene panel from the Genomics England PanelApp, where colors indicate confidence of association to phenotype, from diagnostic-grade in green to low-level confidence genes in red (**step 1**). Annotation through VEP and *vcfanno* with functional variant annotations: variant consequences by VEP, mutation hotspots from cancerhotspots.org, *in silico* deleteriousness predictions from dbNSFP, loss-of-function predictions through VEP's LOFTEE plugin, population allele frequencies from gnomAD, germline biomarkers from CIViC, and low-risk alleles from NHGRI-EBI GWAS Catalog (**step 2**). Pathogenicity classification of novel variants according to a cancer-dedicated implementation of refined ACMG criteria (**step 3**). Aggregation and structuring of the results in a tiered cancer predisposition report (**step 4**). Abbreviations: VEP = Variant Effect Predictor; LOFTEE = Loss-Of-Function Transcript Effect Estimator; VCF = Variant Call Format; dbNSFP = database of non-synonymous functional predictions; CIViC = Clinical Interpretations of Variants in Cancer, ACMG = American College of Medical Genetics and Genomics; gnomAD = Genome Aggregation Database; WGS = Whole-Genome Sequencing; WES = Whole-Exome Sequencing; TAS = Targeted Amplicon Sequencing

Supplementary Materials

CPSR pathogenicity classification

As shown in Supplementary Table S1, CPSR implements a series of ACMG criteria for pathogenicity classification, most of which were outlined in SherLoc (Nykamp *et al.*, 2017). For each variant, scores associated with applicable criteria are accumulated, producing a final pathogenicity score that can be either positive or negative. Currently, the five different levels of pathogenicity are determined through the following thresholds/ranges of pathogenicity scores:

- Pathogenic: [5,]
- Likely Pathogenic: [3.5, 4.5]
- VUS: [3,-2.5]
- Likely Benign: [-3, -4.5]
- Benign: [-5,]

Evaluation of CPSR variant pathogenicity classification

In order to assess the ability of CPSR to classify variants according to pathogenicity, we used the list of putative cancer predisposition variants identified in TCGA's PanCancer study which were classified as *Pathogenic*, *Likely Pathogenic* or *Uncertain Significance* by CharGer (Huang *et al.*, 2018). The total variant set (n = 858) was run through the CPSR workflow (v0.5.2, cancer predisposition gene set 0 (exploratory), otherwise default options). CPSR classifications were compared with CharGer classifications for variants that had existing classifications in ClinVar, release *20191102* (n = 515). Variants that were assigned a different variant consequence by CharGer and CPSR were ignored, producing a final test set of n = 495 variants (Supplementary Table S2). Figure S1 illustrates the concordance (percentage of variants with similar classifications as ClinVar) for CharGer and CPSR, considering both high-confidence ClinVar classifications (A), and variants with any confidence level (B). High-confidence ClinVar classifications are here defined as those variants assigned with at least two gold stars when it comes to review status, denoting variants that are part of practice guidelines, reviewed by expert panels, or submitted multiple times with assertion criteria and evidence. While CharGer achieves higher sensitivity than CPSR for the identification of pathogenic/likely pathogenic (P/LP) variants, CPSR performs considerably better when considering the joint classification of VUS and P/LP variants. The latter reflects a more conservative approach for pathogenicity classification in CPSR, and also suggests that the false positive rate for variants classified as P/LP with CPSR is significantly lower compared to that obtained with CharGer.

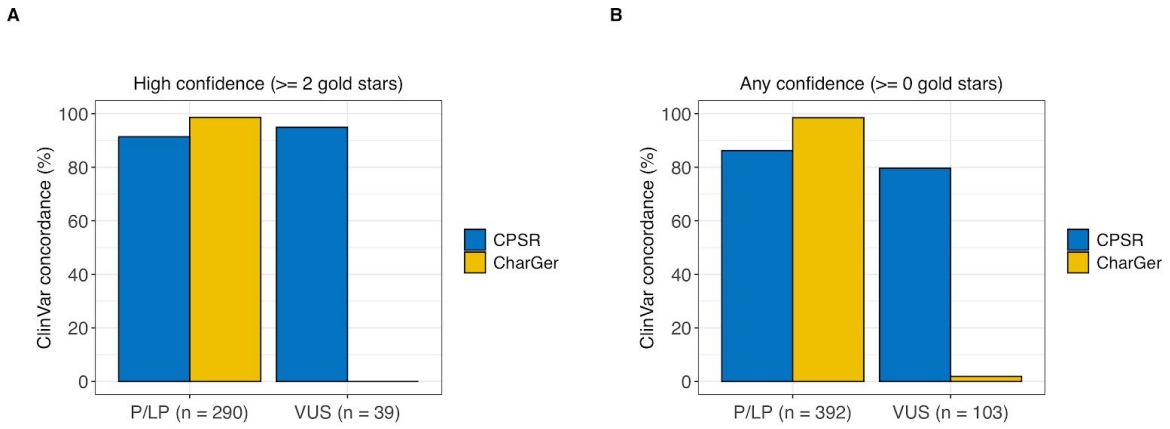

Figure S1: Concordance of CPSR and CharGer classifications against cancer predisposition variants from TCGA's PanCancer germline study (Huang et al., 2018) with **A)** high-confident ClinVar classifications, and **B)** ClinVar classifications with any level of confidence. P/LP = Pathogenic/Likely Pathogenic; VUS = Variant of Uncertain Significance

References

Huang,K.-L. *et al.* (2018) Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*, **173**, 355–370.e14.

Nykamp,K. *et al.* (2017) Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria. *Genet. Med.*, **19**, 1105.