SOFTWARE TOOL ARTICLE

CanVar: A resource for sharing germline variation in cancer patients [version 1; peer review: 4 approved]

Daniel Chubb¹, Peter Broderick¹, Sara E. Dobbins¹, Richard S. Houlston^{1,2}

¹Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK

²Division of Molecular Pathology, The Institute of Cancer Research, London, UK

V1 First published: 05 Dec 2016, 5:2813
<https://doi.org/10.12688/f1000research.10058.1>
 Latest published: 05 Dec 2016, 5:2813
<https://doi.org/10.12688/f1000research.10058.1>

Abstract

The advent of high-throughput sequencing has accelerated our ability to discover genes predisposing to disease and is transforming clinical genomic sequencing. In both contexts knowledge of the spectrum and frequency of genetic variation in the general population and in disease cohorts is vital to the interpretation of sequencing data. While population level data is becoming increasingly available from publicly accessible sources, as exemplified by The Exome Aggregation Consortium (ExAC), the availability of large-scale disease-specific frequency information is limited. These data are of particular importance to contextualise findings from clinical mutation screens and small gene discovery projects. This is especially true for cancer, which is typified by a number of hereditary predisposition syndromes. Although mutation frequencies in tumours are available from resources such as Cosmic and The Cancer Genome Atlas, a similar facility for germline variation is lacking. Here we present the Cancer Variation Resource (CanVar) an online database which has been developed using the ExAC framework to provide open access to germline variant frequency data from the sequenced exomes of cancer patients. In its first release, CanVar catalogues the exomes of 1,006 familial early-onset colorectal cancer (CRC) patients sequenced at The Institute of Cancer Research. It is anticipated that CanVar will host data for additional cancers, providing a resource for others studying cancer predisposition and an example of how the research community can utilise the ExAC framework to share sequencing data.

Keywords

exome sequencing , ExAC , CanVar , cancer , colorectal cancer , NGS , Germline , database

Open Peer Review

Approval Status

	1	2	3	4
version 1				
05 Dec 2016	view	view	view	view

1. **Laura Valle** , L'Hospitalet de Llobregat, Barcelona, Spain
Victor Moreno , Hospitalet de Llobregat, Barcelona, Spain
 University of Barcelona, Hospitalet de Llobregat, Barcelona, Spain
2. **Pavel Vodička**, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, Prague, Czech Republic
3. **Wolfgang Huber** , European Molecular Biology Laboratory, Heidelberg, Germany
Vladislav Kim, European Molecular Biology Laboratory, Heidelberg, Germany
4. **James Ware**, Imperial College London, London, UK
 MRC London Institute of Medical Sciences, London, UK
Nicola Whiffin , Imperial College London, London, UK

Any reports and responses or comments on the

article can be found at the end of the article.

Corresponding author: Daniel Chubb (daniel.chubb@icr.ac.uk)

Competing interests: The authors declare no competing interests.

Grant information: This work was supported by grant funding from Cancer Research UK (C1298/A8362), the European Union Seventh Framework Programme (FP7/2007–2013) under grant 258236, FP7 collaborative project SYSCOL and BLOODWISE (LRF05001). All grants assigned to Richard S Houlston.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2016 Chubb D *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Chubb D, Broderick P, Dobbins SE and Houlston RS. **CanVar: A resource for sharing germline variation in cancer patients [version 1; peer review: 4 approved]** F1000Research 2016, 5:2813 <https://doi.org/10.12688/f1000research.10058.1>

First published: 05 Dec 2016, 5:2813 <https://doi.org/10.12688/f1000research.10058.1>

Introduction

With the widespread adoption of high-throughput sequencing as a tool for disease gene discovery and clinical diagnostics there is a need to evaluate candidate disease predisposition genes through defining the spectrum and frequency of genetic variation in the general population and in specific disease cohorts. For this to be meaningful, large sample sizes are required in order that variant frequencies are accurately defined. Such data is often only acquired through combining multiple datasets. Although these data are being rapidly produced by both large consortia and individual research groups, their acquisition and integration are subject to logistical, computational and ethical challenges. When undertaken by multiple agencies, this results in considerable duplication of effort, the products of which may not be widely shared. It is therefore desirable for large, processed sequencing datasets to be made easily accessible to the community. Recently, a paradigm for sharing has been provided by the Exome Aggregation Consortium^{1,2} (ExAC). ExAC have aggregated and analysed a set of 60,706 exomes from over twenty different studies, providing this information as an intuitive online resource. The ExAC website presents these data as variant frequencies stratified by different ethnic groups alongside additional sequencing quality metrics and transcript based annotations.

Similar resources providing frequencies of variants in specific disease associated cohorts are not widely available. Such datasets are of particular importance for small-scale studies, where the confirmation of rare variant frequencies in genes of interest is critical to determine the importance of candidate genes. Furthermore, in the case of clinical genetic testing, they aid in the interpretation of variants of unknown significance. This is especially true for cancer, where it is estimated that 5–10% of cases have a strong heritable basis³. The identification of genes involved in hereditary cancers not only provide valuable biological insight but can allow for screening of at risk individuals, providing an opportunity for early diagnosis, which is key to long-term survival. To address the deficiency of germline frequency data in the realm of cancer research, we have produced CanVar, an online resource derived from cancer patient germline exome sequencing data. CanVar has been produced by

adapting the ExAC framework² to provide cancer type specific variant frequencies, presenting them as a familiar and intuitive online interface modelled after the ExAC browser.

CanVar datasets

CanVar currently catalogues frequency data for 1,006 early-onset familial colorectal cancer cases⁴. In total, 1,096,907 variant sites are catalogued in CanVar: specifically 981,491 single nucleotide variants (SNVs) and 115,416 insertion deletions (indels). As previous studies have observed, rare variation is itself common, indeed 52% of these variants are only observed in one sample.

It is beneficial to be able to compare cancer variant frequency in cases with that observed in population frequency control data. We have therefore annotated each cancer variant with ExAC allele frequency data excluding samples from The Cancer Genome Atlas (TCGA, n=53,105, henceforth referred to as non-TCGA ExAC). Links are also provided to the relevant ExAC browser entries at the gene and variant levels in order to assess loss of function tolerance and overall gene burden.

CanVar website

CanVar utilises an adapted ExAC framework, providing SNV and INDEL frequency data and can be accessed via <http://canvar.icr.ac.uk>. The interface mirrors the ExAC browser available at <http://exac.broadinstitute.org/>² and is divided into three main parts: the front page (Figure 1), the gene page (Figure 2) and the variant page (Figure 3).

Front page

The front page (Figure 1) contains a search bar where either genes or individuals variants can be queried. Genes are queried either by entering an HGNC gene name or ensemble gene ID. Individual transcripts within a gene can also be queried through entering an Ensembl transcript ID. Variants are queried either by dbSNP rsid or entering the chromosome, position, reference and alternate alleles. Additionally, whole regions can be queried, which opens a page similar to the gene view, providing coverage data and variants present in the queried region.

CanVar Browser Beta

About Downloads Terms Contact FAQ

CanVar Browser (Beta) | The Cancer Variation Resource

Examples - Gene: [PCSK9](#), Transcript: [ENST00000407236](#), Variant: [22-46615880-T-C](#), Multi-allelic variant: [rs1800234](#), Region: [22:46615715-46615880](#)

About CanVar

CanVar is a resource of variant level frequency data from cancer germline sequencing studies.

CanVar Currently contains 1,006 Familial early onset CRC patients

Recent News

November 14th, 2016

- Public release of CanVar Browser (beta)

Figure 1. The CanVar front page features a search bar, example queries and additional news and updates.

Gene page

The Gene page (Figure 2) first provides metadata and external links followed by a per base resolution coverage plot on top of the exon-intron structure of the gene of interest. These features default to the Ensembl canonical transcript but different transcripts can either be searched from the front page or selected from a drop down menu. A table provides frequency information and annotations for each

variant identified within the gene assuming the worst effect in any transcript. The quality of a variant in the gene table is assessed by its filter status, obtained from the variant recalibration step of the GATK pipeline (Methods). To simplify the table display, users can select the cancers of interest. Non-TCGA ExAC frequencies are also displayed for each variant. Selecting a variant will open up the appropriate variant page.

Figure 2. The Gene page is divided in to three parts. A) metadata and external links, including the ExAC page for a given gene; **B)** coverage plot and exon/intron structure **C)** table containing annotations and variant frequencies for each variant identified within a gene. The ExAC_AF column refers to the frequency from non-TCGA ExAC. The variant table has a menu C.1) which is used to select which cancer frequencies are displayed. Currently only NSCCG CRC samples are available.

Figure 3. The Variant page can be divided in to five parts. **A)** Call rate of a given variant **B)** Metadata and external links, including equivalent ExAC page; **C)** Quality metrics **D)** Transcript annotations **E)** Frequency information in different studies.

Variant page

More detailed quality and frequency information is provided in the variant page (Figure 3). Links are provided to external resources such as the equivalent ExAC page and users can explore genotype, depth and site quality metrics. The call rate of each variant according to the QC thresholds (Methods) is provided at the top of the page. Care should be taken when interpreting variants with lower call-rates as they are typically more likely to be false positives. Annotation particular to different transcript can be browsed along with an assessment of loss of function variant quality according to the Loss-Of-Function Transcript Effect Estimator (LOFTEE - <https://github.com/konradjk/loftee>). The frequency of the variant across studies included within CanVar is also provided in a sortable table.

Discussion

ExAC, the most comprehensive attempt at a large-scale aggregation of sequencing data, has been a great success, proving the usefulness of providing open-access population level genetic data for the research community. Here we present an adaptation of the ExAC framework to create CanVar, a cancer specific online resource for germline sequencing data.

CanVar currently provides SNV and INDEL frequency data, with associated annotations. As ExAC introduce new features it is anticipated that these will be merged in to future versions of CanVar.

The data currently catalogued in CanVar will provide a valuable resource for researchers investigating genetic predisposition to colorectal cancer and those engaged in delivery of clinical cancer genetic testing programs. It is expected that the utility of CanVar will increase as additional sequencing data is integrated through a number of different mechanisms: firstly, in-house sequencing of ongoing projects at the Institute of Cancer Research; secondly, applications for publically available data e.g. samples deposited in the [Ensembl EGA archive](#) and [dbGap](#); and thirdly, collaborations with others engaged in the germline sequencing of cancer patients.

Only when the community fully embraces a policy of data sharing will resources such as ExAC and CanVar fulfil their potential. We therefore encourage all researchers engaged in cancer germline sequencing projects to consider sharing their data (email canvar@icr.ac.uk). Where consent or other factors preclude the sharing of the individual level data, we encourage others to adopt the ExAC framework to make their data available. To facilitate this we have made our adapted ExAC code available.

Methods Implementation ExAC framework

CanVar is built upon the Python-based framework designed to accommodate the ExAC database downloaded from https://github.com/konradjk/exac_browser. A full description of the framework's

construction and optimisation is available from the ExAC browser publication².

Briefly, custom python scripts parse input data into a mongoDB database. These data consist of variant calls with VEP annotations (from VCF files) and sample coverage metrics (derived from BAM files) in addition to other annotation data in the form of downloaded flat files from dbSNP (for rsids), Gencode v19 (for transcript and gene structure), dbNSFP (for gene names and aliases) and OMIM (to link to the relevant OMIM entry).

The python Flask framework is then used to serve variant frequencies and associated annotations from mongoDB to webpages based upon HTML templates.

Hardcoded paths contained within the original code were altered and additional changes were made to the provided HTML templates to remove ExAC specific references and to make specific changes in the interface. For example, the gene results page was altered to annotate CanVar frequencies with ExAC frequency data and to allow for multiple studies to be viewed on the same table.

Full installation instructions with all software dependencies are provided at <https://github.com/danchubb/CanVar/blob/master/readme.txt>. The required python modules, installed using the pip package management system are described in <https://github.com/danchubb/CanVar/blob/master/requirements.txt>.

Hardware

CanVar runs on a Dell PowerEdge R310 with 1x Intel i3-540 CPU and 4 GB DDR3 RAM using Apache version 2.4.6. The variant and associated annotation mongoDB files are 55GB in size.

Website

The CanVar website itself can be accessed using any modern internet browser.

Curation of colorectal cancer exome data within CanVar

CanVar currently contains summary level exome sequencing data from 1,006 early-onset familial CRC cases⁴ from the National Study of Colorectal Cancer Genetics (NSCCG)⁵. All samples had previously undergone quality control, ensuring the removal of those with: non-northern European ancestry, high levels of heterozygosity, sex discrepancy, poor call rate and contamination. The full sequencing and analysis pipeline is described in detail in the dataset's publication⁴. Briefly: all samples underwent exome capture utilising Illumina's Truseq 62 Mb expanded exome enrichment kit followed by sequencing using Illumina Hi-seq 2500 technology. Alignment to build 37 (hg19) of the human reference genome was performed using Stampy(v1.0.17)⁶ and BWA(v0.5.9)⁷ software. Alignments were processed using the Genome Analysis Tool Kit (GATKv3) pipeline according to best practices^{8,9}. Analysis was restricted to capture regions defined in the Truseq 62Mb bed file plus 100bp padding. Combined individual level VCF files generated using the GATK 3 pipeline were assessed using variant quality recalibration (VQSR). In this step a variant is assigned a tranche which

represents the sensitivity threshold required to call a given variant, the higher the tranche, the less confidence is given to a call. Variants are assigned a PASS value if they fall below the 99.0 tranche for SNVs and the 95.0 tranche for indels. Above these values, the sensitivity required for a given variant is reported in increments of 0.1 to provide users with the most accurate assessment of variant quality. The CRC cases were jointly called and subjected to VQSR alongside a larger set of exomes therefore calls may differ from those reported in previous publications. Finally, each variant was annotated using the Ensembl Variant Effect Predictor(VEP v78)¹⁰ before being converted to the summary level site format required by the ExAC framework using custom python scripts.

Data conversion to ExAC format

The ExAC framework requires individual level variant and coverage files to be converted into specific summary formats before they can be parsed into mongoDB.

Variant frequency and annotation. Individual level vcf files are converted into a summary site format, providing allele count and frequency data for different groups in addition to depth and genotype quality data. For ExAC these groups correspond to ethnic groups whereas CanVar utilises this facility to instead group samples in to separate phenotypic classes, allowing the expansion of the database to contain data from a variety of malignancies. This process is accomplished using a custom python script https://github.com/danchubb/CanVar/blob/master/vcf_to_site_canvar.py which takes as input a VCF file and a list of which populations (or phenotypes) each contained sample belongs to. Variant frequencies and VEP annotations are then output according to QC parameters. In order to provide maximum sensitivity for users, minimum variant QC is imposed: requiring a site to be called in > 50% of samples and for an individual sample call to have a depth of > 2 reads with a GQ>20. All female Y chromosome calls are removed, as are male heterozygous Y and X calls.

Coverage data. Per base coverage files are generated for each sample using the GATK DepthOfCoverage command. Individuals coverage files are then indexed using the tabix tool and average coverage across all captured bases is calculated across all samples using a custom python script: https://github.com/danchubb/CanVar/blob/master/average_coverage_calculate.py.

Data and software availability

The CanVar website is available at: <https://canvar.icr.ac.uk>

Latest source code: <https://github.com/danchubb/CanVar>

Archived source code as at the time of publication: [10.5281/zenodo.168019](https://zenodo.org/record/168019)¹¹

License: The source code is licensed using the same MIT open source license as ExAC (<https://github.com/danchubb/CanVar/blob/master/LICENSE>).

Raw data

Raw alignment (BAM files) data on the 1,006 CRC samples have been deposited at the European Genome-phenome Archive with

accession number [EGAS00001001666](#). The availability of individual level data for future datasets included within CanVar will be specific to each study.

Author contributions

Conception and design: Daniel Chubb, Sara E. Dobbins, Peter Broderick, Richard S. Houlston, Collection and assembly of data: Daniel Chubb, Peter Broderick, Sara E. Dobbins. Implementation: Daniel Chubb. Manuscript writing: All authors. Final approval of manuscript: All authors.

Competing interests

The authors declare no competing interests.

Grant information

This work was supported by grant funding from Cancer Research UK (C1298/A8362), the European Union Seventh Framework Programme (FP7/2007–2013) under grant 258236, FP7 collaborative project SYSCOL and BLOODWISE (LRF05001). All grants assigned to Richard S Houlston.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

Thanks to Nikolas Pontikos (https://github.com/pontikos/uclex_browser) for his assistance with the ExAC framework and data parsing.

References

- Lek M, Karczewski KJ, Minikel EV, *et al.*: **Analysis of protein-coding genetic variation in 60,706 humans.** *Nature*. 2016; **536**(7616): 285–91.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Karczewski KJ, Weisburd B, Thomas B, *et al.*: **The ExAC Browser: Displaying reference data information from over 60,000 exomes.** *bioRxiv*. 2016.
[Publisher Full Text](#)
- Nagy R, Sweet K, Eng C: **Highly penetrant hereditary cancer syndromes.** *Oncogene*. 2004; **23**(38): 6445–6470.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Chubb D, Broderick P, Dobbins SE, *et al.*: **Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer.** *Nat Commun*. 2016; **7**: 11883.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Penegar S, Wood W, Lubbe S, *et al.*: **National study of colorectal cancer genetics.** *Br J Cancer*. 2007; **97**(9): 1305–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lunter G, Goodson M: **Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads.** *Genome Res*. 2011; **21**(6): 936–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics*. 2009; **25**(14): 1754–60.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McKenna A, Hanna M, Banks E, *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res*. 2010; **20**(9): 1297–303.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- DePristo MA, Banks E, Poplin R, *et al.*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet*. 2011; **43**(5): 491–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McLaren W, Pritchard B, Rios D, *et al.*: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.** *Bioinformatics*. 2010; **26**(16): 2069–70.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- danchubb: **danchubb/CanVar: Canvar code beta 0.1 F1000.** *Zenodo*. 2016.
[Data Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 06 January 2017

<https://doi.org/10.5256/f1000research.10838.r18218>

© 2017 Ware J et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

James Ware

¹ National Heart & Lung Institute, Imperial College London, London, UK

² MRC London Institute of Medical Sciences, London, UK

Nicola Whiffin

National Heart & Lung Institute, Imperial College London, London, UK

Summary & Impression

The authors describe an online resource for exploring germline cancer variants. The resource currently contains data from 1,006 samples representing a single cancer type, a single ethnicity, and a single centre. The authors have invited collaborators to help expand this to other cancer types which will add further value to what is already an excellent tool over time.

We share the authors' enthusiasm for intuitive data sharing, and agree that presenting variant frequencies from disease cases, as well as reference samples, is hugely valuable. Overall, the manuscript is very clear, and we anticipate that the intuitive web resource will be well received.

We have a couple of high level comments, and some minor suggestions for the authors to consider.

Comments

- The authors describe 2 uses for this data: variant-level analyses (i.e. interpreting individual variants, primarily in established disease genes), and gene-level analyses (assessing candidate disease predisposition genes). The resource in its current form is primarily suitable for the first. The case data and control data are unlikely to be technically matched sufficiently for case/control association testing at the gene level (burden tests), and the authors wisely do not provide this sort of comparative data on the gene page. So, while variant frequency data is important in interpreting genes, in our opinion the present resource is primarily valuable for variant interpretation.
- A critical strength of the ExAC project was the joint and unified analysis of aggregated data. The authors describe adopting the "ExAC framework", but at present this represents only

data from only a single source. As well as adopting the ExAC web architecture, it would be interesting to hear the authors' plans for data analysis going forwards - will CanVar seek to harmonise variant calling and analysis on data from disparate sources as it grows?

Suggestions for consideration

- Readers are likely to know ClinVar as the go-to resource for germ-line variants in inherited diseases. It may be worth highlighting the complementary value of ClinVar & CanVar - i.e. the addition of consistent frequency data.
- I would provide a little more detail on the case series in the section on "CanVar datasets" (introduction). In particular, are cases all unrelated probands? I would add here that cases are all European (given in methods). Drs Huber & Kim note that variants were limited to "germline variants that were identified as risk-associated" - I did not appreciate this from the manuscript, and think this is important to note if the dataset does not include all variants in protein-coding regions.

Gene view

- Are variants annotated with ExAC frequencies even if they do not 'PASS' filters in ExAC? May be helpful to display the ExAC filter status as well as CanVar filter status.
- Would be helpful to indicate whether variants absent in ExAC were well covered - i.e. give some summary measure of ExAC coverage for all variant sites, since capture platforms and coverage profiles may be very different
- We understand that ExAC_AF on this page is non-TCGA ExAC frequency. Is this ethnicity matched too, since samples are all European)?

Variant view

- It would be invaluable to incorporate ExAC frequencies into the frequency table (Fig 3e) - especially since the non-TCGA data is not accessible via click through to the web browser (only by download). Users may be misled by a link to the full ExAC dataset (with TCGA included).
- As data is added would be desirable to stratify by cancer type AND ethnicity
- The call rate is reported with 12 decimal places of precision

Methods

- "Curation of colorectal cancer exome data within CanVar"
May be helpful to indicate which of the technical parameters differ from ExAC where relevant.
- It would be interesting to hear about any challenges encountered in reconciling the data sets that may be relevant to others attempting something similar - Were there any problems with multi-allelic sites? e.g. GATK filters by site, rather than variant. Are multi-nucleotide polymorphisms phased and jointly interpreted? Any other technical challenges?

Data availability

- Is the sites-only vcf available for download? This may be useful in addition to raw data available via application to EGA.
- Could add source code link to final sentence of the discussion.

Competing Interests: Dr Whiffin has previously worked directly with the authors (PhD supervised by Prof Houlston -2014). Dr Ware has no competing interests, and takes responsibility for the content of the review.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 03 January 2017

<https://doi.org/10.5256/f1000research.10838.r18217>

© 2017 Huber W et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Wolfgang Huber

EMBL Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

Vladislav Kim

European Molecular Biology Laboratory, Heidelberg, Germany

Short synopsis. The authors describe an online resource for exploring disease-associated germline variants. The Cancer Variation Resource (CanVar) browser is inspired by the ExAC browser at the Broad Institute. Currently CanVar is limited to those germline variants that were identified as risk-associated in a study of 1,006 familial early-onset colorectal cancer (CRC) patients published by the authors in Nature Communications in 2016.

Overall impression. CanVar is a useful tool for mining variants implicated in CRC, and the authors do a good job at explaining how to use the resource. The authors also provide some background information on the underlying data, methodology and technologies. We have only a few minor suggestions as to how its presentation could be improved.

Suggestions. In the Introduction, the sentence “CanVar has been produced by adapting the ExAC framework” is a bit vague. You could be clearer what is meant by framework: software? APIs? some or all of the concepts, and which? datasets?

Abstract and introduction can be confusing (at least to the rushed reader, of which there are many) as to whether CanVar also contains or interfaces to the ExAC 60,000 exome data on top of the CRC data. This is clarified in the “CanVar datasets” subsection, but in our view this should be clarified earlier. For instance, abstract and introduction talk more, and earlier, about ExAC than about the data that are actually contained in CanVar; in the “CanVar website” section, links to the ExAC browser and to the CanVar website are provided side by side, which might lead to further

confusion. Since both websites are almost look-alikes, readers might even be led to expect that both sites might also mirror each other. Perhaps, it would be better to provide the links in a more asymmetric manner.

In “CanVar datasets” it is mentioned that each variant is annotated “with ExAC allele frequency data excluding samples from the TCGA”. Perhaps, a short explanation of why this has been done could be provided, as not all readers may be familiar with the source of non-TCGA samples in the ExAC dataset.

It should be time to remove the “Beta” state from the resource. Bring it to a good enough state to warrant release, and then be not afraid to update it later with new releases.

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 21 December 2016

<https://doi.org/10.5256/f1000research.10838.r18681>

© 2016 Vodička P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Pavel Vodička

Department of the Molecular Biology of Cancer, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, Prague, Czech Republic

The study of Chubb *et al.* presents an online database and software tool Cancer Variation Resource (CanVar), developed on the basis of Exome Aggregation Consortium (ExAc) framework (sequenced exomes of cancer patients). The main aim of the database is to enable an open access to germline variant frequency. CanVar focuses on colorectal cancer, as it summarizes exome sequencing data from more than 1000 familial early-onset patients with this disease. Strikingly, the CanVar database catalogues almost 1.1 million variants and more than 100000 insertions/deletions. An additional advantage for the user are the data on associated annotations of variants and insertions/deletions.

The information, which may be acquired from on the basis of the published database, may provide valuable information on gene variants and indels in populations assuming disease-specific context. The data that could be mined with the help of the present database may also find utilization in clinics, particularly in the context with mutational screening in cancer, which becomes to be routine.

A sentence in Introduction (When undertaken by multiple agencies...) would benefit from re-phrasing into a more reader-friendly form.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 20 December 2016

<https://doi.org/10.5256/f1000research.10838.r18219>

© 2016 Valle L et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Laura Valle

Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain

Victor Moreno

¹ Unit of Biomarkers and Susceptibility, Catalan Institute of Oncology, IDIBELL and CIBERESP, Hospitalet de Llobregat, Barcelona, Spain

² Department of Clinical Sciences, School of Medicine, University of Barcelona, Hospitalet de Llobregat, Barcelona, Spain

In this article Chubb *et al.*¹ describe a software tool (CanVar) that the authors have developed to make publically available germline variant frequency information from sequenced exomes of cancer patients. Most importantly, they have used the Exome Aggregation Consortium (ExAc) framework, a tool the scientific community is currently very familiar to, in order to facilitate its access and use. The incalculable value of the open accessibility to the germline variation data obtained from >60,000 exomes provided by the Exome Aggregation Consortium (ExAc) seem to finally begin to reach disease-specific cohorts, as it is the case of CanVar. Hopefully this will soon become a reality not only for cancer but also for other common diseases. This information, together with the variation frequencies observed in the general population, is key when trying to evaluate the pathogenic relevance of disease-predisposing genes and/or variants, not only for novel candidate genes but also for well-known susceptibility genes.

So far, the data available through CanVar correspond to the 1,006 exomes of early onset familial colorectal cancer cases recently studied by the same group². Being this a very insightful cohort in the field of colorectal cancer predisposition, much needs yet to be done to make CanVar a relevant routine tool for the scientific community, and it is the responsibility of all of us to make this possible. The tool is already available, so I encourage all researchers with germline exome sequencing data in cancer patients to submit their data to CanVar, as larger representation of tumor types, populations and patients in general is required. I also would like to encourage researchers to use these extremely useful data in their cancer predisposition studies and to increase the visibility of CanVar among their colleagues and peers.

Despite the so far limited availability of germline exome sequencing data from cancer patients, a

huge amount of data has been gathered in the last years from genome-wide association studies and exome SNP arrays. This information would be of value if added to CanVar, at least the variants included in exome arrays and rare exonic variants included in genotyping arrays.

Another issue that needs to be contemplated is the implementation of filters for ethnicities/studies, anticipating the inclusion of data from other groups. Alternatively, as occurs in ExAC, data could be itemized by ethnicity/geographic origin and study.

References

1. Chubb D, Broderick P, Dobbins SE, Houlston RS: CanVar: A resource for sharing germline variation in cancer patients [version 1; referees: awaiting peer review]. *F1000Research*. 2015; **5** (2813). [Publisher Full Text](#)
2. Chubb D, Broderick P, Dobbins SE, Frampton M, et al.: Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nat Commun*. 2016; **7**: 11883 [PubMed Abstract](#) | [Publisher Full Text](#)

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research