SOFTWARE TOOL ARTICLE

seqCAT: a Bioconductor R-package for variant analysis of high throughput sequencing data [version 1; peer review: 1 approved with reservations]

Erik Fasterius [1], Cristina Al-Khalili Szigyarto [1,2]

[1]School of Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, 10691, Sweden
[2]Science for Life Laboratory, KTH - Royal Institute of Technology, Solna, Sweden

v1 **First published:** 14 Sep 2018, **7**:1466 (https://doi.org/10.12688/f1000research.16083.1)
Latest published: 14 Sep 2018, **7**:1466 (https://doi.org/10.12688/f1000research.16083.1)

Abstract

High throughput sequencing technologies are flourishing in the biological sciences, enabling unprecedented insights into *e.g.* genetic variation, but require extensive bioinformatic expertise for the analysis. There is thus a need for simple yet effective software that can analyse both existing and novel data, providing interpretable biological results with little bioinformatic prowess. We present *seqCAT*, a Bioconductor toolkit for analysing genetic variation in high throughput sequencing data. It is a highly accessible, easy-to-use and well-documented R-package that enables a wide range of researchers to analyse their own and publicly available data, providing biologically relevant conclusions and publication-ready figures. SeqCAT can provide information regarding genetic similarities between an arbitrary number of samples, validate specific variants as well as define functionally similar variant groups for further downstream analyses. Its ease of use, installation, complete data-to-conclusions functionality and the inherent flexibility of the R programming language make seqCAT a powerful tool for variant analyses compared to already existing solutions. A publicly available dataset of liver cancer-derived organoids is analysed herein using the seqCAT package, demonstrating that the organoids are genetically stable. A previously known liver cancer-related mutation is additionally shown to be present in a sample though it was not listed in the original publication. Differences between DNA- and RNA-based variant calls in this dataset are also analysed revealing a high median concordance of 97.5%.

Keywords

High throughput sequencing, whole exome sequencing, RNA sequencing, variant analysis, single nucleotide variant, R, Bioconductor

Open Peer Review

Referee Status: ?

	Invited Referees 1
version 1 published 14 Sep 2018	? report

1 **Matej Lexa**, Masaryk University, Czech Republic

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the RPackage gateway.

This article is included in the Bioconductor gateway.

Corresponding author: Cristina Al-Khalili Szigyarto (caks@kth.se)

Author roles: Fasterius E: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Al-Khalili Szigyarto C**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the European Community 7th Framework Program under grant agreement no. 278 568 ``PRIMES''.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Fasterius E and Al-Khalili Szigyarto C. **seqCAT: a Bioconductor R-package for variant analysis of high throughput sequencing data [version 1; peer review: 1 approved with reservations]** F1000Research 2018, **7**:1466 (https://doi.org/10.12688/f1000research.16083.1)

First published: 14 Sep 2018, **7**:1466 (https://doi.org/10.12688/f1000research.16083.1)

Introduction

High throughput sequencing (HTS) technologies such as genome, exome and RNA sequencing (RNA-seq) have become some of the most powerful and widely used tools in biological research worldwide, and an increasing amount of such data is being stored in online data repositories (*e.g.* the Gene Expression Omnibus, GEO, and the Sequence Read Archive, SRA)[1–3]. While decreasing experimental costs and optimised protocols enable a broad range of researchers to apply HTS to their respective scientific questions, the analysis of the resulting data is not a trivial matter, often requiring a high level of bioinformatic expertise[4,5]. Two examples of such software include the command line tool *vcftools*[6] and the R-package *VariantAnnotation*[7]. There are also several software tools with a more easy-to-use graphical interface (such as the *Integrative Genomics Viewer*[8] or the web-based *Ensembl Genome Browser*),[9], but these are limited in their functionality. Web-based applications are limited in the amount of data that can be uploaded, and also come with the added issue of data security[10]. Proprietary software (such as the *Ingenuity Variant Analysis*)[11] not only require a licence to use, but also constitute a "black box" where the underlying methods are not available for direct inspection or scrutiny.

There is thus a need for transparent, user-friendly and powerful bioinformatic tools to enable as many researchers as possible to analyse and interpret their own and publicly available HTS data. Two important aspects of such analyses is the true identity of cells analysed and comparability of both the biological samples and the data sets. Validation and evaluation of cell line authenticity, for example, is an increasingly widespread issue, as is the question of biological equivalence for any sample in general[12]. Here we present an open source R-package, the *High Throughput Sequencing Cell Authentication Toolkit* (seqCAT), which uses data from HTS experiments (whether it be of DNA or RNA origin) to investigate these matters.

One of the common outputs from HTS experiments is that of *sequence variation*. Single nucleotide variants (SNVs), for example, are sequence variations at the nucleotide level. Such data is the output of many variant calling programs and algorithms, which is usually stored in *variant call format* (VCF) files. These files are used by seqCAT in order to analyse genetic differences between samples. We have previously demonstrated the usefulness and general applicability of such analyses for both cell line authentication[13] and genetic heterogeneity in public cell line datasets[14]. The capabilities of seqCAT include creation of SNV profiles from VCF files, comparisons of the overall genetic similarity between profiles, investigations of SNV impact distributions (*i.e.* variants' predicted impact on protein function) as well as interrogations of the genotypes of previously known or user-specified variants across samples. Each individual profile can represent SNVs from a HTS experiment or from an external variant database.

In the present study, we use seqCAT to explore genetic differences within a public dataset containing both whole exome sequencing (WES) and RNA-seq data for long-term organoid cultures. We show that the organoids are genetically stable over a culture-period of several months, corroborating the original authors' conclusions. We also demonstrate how seqCAT can be used to compare DNA- and RNA-based variant calls using the same dataset. The results highlight potential uses of variant analyses and demonstrate how seqCAT may be utilised to interrogate genetic differences at both the global and gene-specific level.

Methods

SeqCAT was developed for the *Bioconductor*[15] repository for R-packages. It follows existing best coding practises, including a clean, modular and robust design. The basis of all seqCAT analyses are *SNV profiles*: collections of filtered, high-quality SNVs for any given sample. The creation of these SNV profiles is performed by filtering an input VCF file based on the available variant calling quality metrics as well as an optional sequencing depth threshold (set to ten by default)[13]. These profiles are saved as simple text files on the user's hard-drive, in order to facilitate re-use and to reduce the run-time of downstream analyses. While profiles for individual samples may be created as needed by the user, several convenience-functions for working with multiple VCFs and profiles in aggregate are also available. SeqCAT can analyse VCF files with or without annotations from *e.g.* snpEff[16] and also includes a Python-implementation of profile creation, which reduces the run-time of this step five to ten times.

The SNV profiles are subsequently read and compared to each other in a pairwise manner, yielding information on *e.g.* the *overlap* (SNVs that are present in both samples being compared), the *concordance* (the proportion of SNVs with identical genotypes for both samples) and the *similarity score* (a previously defined weighted measure of the concordance)[14]. Comparisons may be performed individually or in aggregate, depending on what type of analysis the user is interested in. Comparisons with external databases is also possible; seqCAT currently contains functionality to read and compare variants present in the *Catalogue of somatic mutations in cancer* (COSMIC) database[17]. Only overlapping variants are analysed by default, but non-overlaps can optionally be included as well. Examining specific chromosomes, genes or genomic regions is also possible, as are analyses of variant functionality through their predicted impact on protein-function.

Installation of both seqCAT and its dependencies is simple, and its use is described in-depth in its vignette since a major design goal of seqCAT was ease-of-use for a broad range of researchers, regardless of expertise in R. While existing data structures and objects from Bioconductor are used internally, none of these are required learning for the user; results are given as standard R-objects[7,18]. This makes exploration of the data as simple and easy as possible for the user. SeqCAT allows for re-analysis of already created SNV profiles, facilitating comparisons of samples across any number of datasets and includes several functions for creating publication-ready figures. All these capabilities make seqCAT a useful, simple and intuitive tool for a wide range of researchers.

Operation

The seqCAT package is designed to work with Bioconductor version 3.7 and R version 3.5.

Results

Using seqCAT to investigate genetic heterogeneity in liver cancer-derived organoids

To demonstrate the capabilities of seqCAT, we analysed a recently published dataset from Broutier *et al.*[19]. The authors created liver cancer-derived organoids for modelling disease and performed both whole exome sequencing and RNA-seq on the original tissues and the organoid cultures. We used seqCAT to analyse the raw VCF files available at GEO under accession GSE84073 (see the Supplementary Code for details and Supplementary Data 1 for the study metadata). The overall genetic similarities between tissues and organoids are clearly grouped according to their respective patient of origin, as can be seen in Figure 1A. We also investigated if this holds true for SNV profile subsets containing only coding and missense variants. The original VCF files were thus annotated using *snpEff*[16], followed by creation, reading and sub-setting of SNV profiles. Figure 1B shows the pairwise comparisons of these variant subsets, indicating that groupings based on genetic similarities of missense variants also separate the dataset in a per-patient manner. This data covers upwards of hundreds of thousands of overlapping variants for each pairwise comparison (Table 1).

We sought to investigate the genetic stability of the organoids both in terms of their transition from primary tissue to organoid culture, as well as long-term culturing. Figure 2A shows a boxplot of genetic similarities for both of these comparisons, indicating that the long-term cultures seem to be more genetically similar than the transition from tissue to organoid. This conclusion is not statistically significant, however, with p-values of 0.36 and 0.41 for all and subset variants, respectively (Supplementary Code). A larger cohort may thus be needed to fully explore the difference between tissue-to-organoid and long-term-culturing stability. The overall high genetic similarities of all the organoids are clear, however: the lowest median similarity score across all patients and all variants is 93.9 (patient CHC2), while reaching as high as 97.9 (healthy patient 1); see Table 1. The similarity scores across coding and non-subset profiles are roughly equivalent.

The original publication[19] lists a number of previously known liver cancer variants (Supplementary Data 2), which we analysed with seqCAT. This analysis reveals that some of the known variants are present in the organoids but absent in their corresponding tissue (Figure 2B). SeqCAT indicates that these specific variants would need to be investigated further, which the original authors have done in most cases. However, it revealed that the GPRIN1 variant is present in the CC1 samples, something not mentioned in the original publication.

Annotations with *snpEff* include variant *impacts*, which are the predicted effects on protein functions and range from HIGH, MODERATE, LOW through MODIFIER, in decreasing order of importance. An example of a HIGH impact is a variant leading to protein truncation, while a MODIFIER variants is predicted to a little to no effect on their resulting protein (such as intronic variants). SeqCAT can summarise and visualise

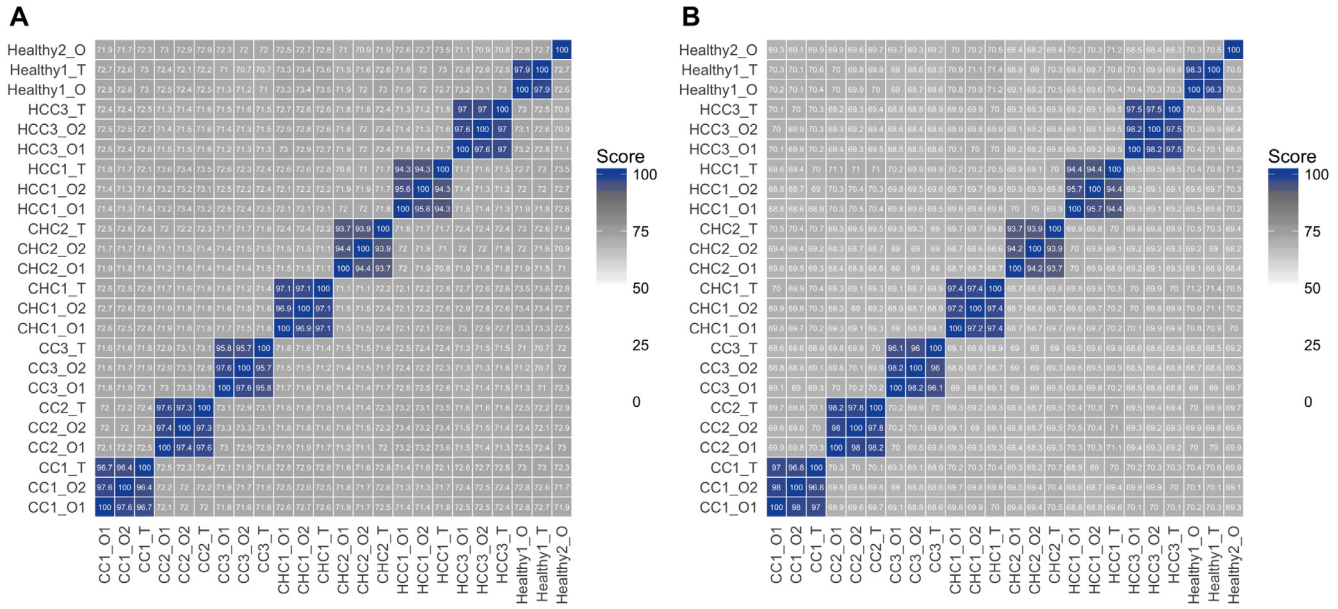

Figure 1. Pairwise comparisons of all WES SNV profiles, showing the genetic similarity of all individual samples for either no variant sub-setting (**A**) or sub-setting for coding variants only (**B**). The colour gradient is defined for ranges of the similarity score: scores between 0 and 50 are shown as white, scores between 50 and 90 as a white-to-grey gradient and, finally, a grey-to-blue gradient for 90 to 100. Samples are named according to their type: original tissues (T), established organoids (O1) and long-term cultured organoids (O2). These figures were created using the `plot_heatmap` seqCAT function.

Table 1. Summary statistics of whole exome sequencing SNV profile comparisons.

Patient	median overlaps (all)	median overlaps (coding)	median similarity score (all)	median similarity score (coding)
CC1	153815	111977	96.7	97.0
CC2	137261	97344	97.4	98.0
CC3	122577	87604	95.8	96.1
CHC1	153589	112011	97.1	97.4
CHC2	132805	95203	93.9	93.9
HCC1	142389	104087	94.3	94.4
HCC3	130186	92613	97.0	97.5
Healthy1	155949	113592	97.9	98.3

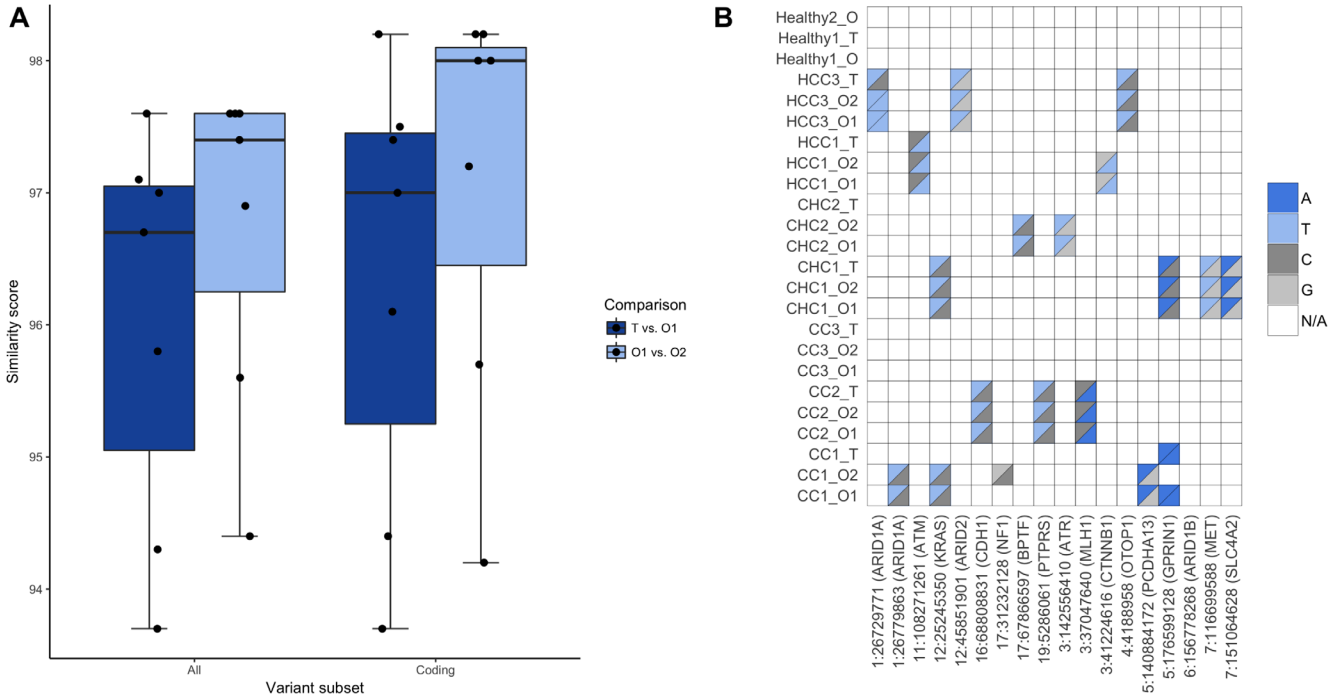

Figure 2. (**A**) Comparisons of genetic similarities between original tissue, derived organoids and long-term cultured organoids. Results are shown for both non-subset variant comparisons and for subsets including coding variants only. The differences between T vs. O1 and O1 vs. O2 for each subset are not statistically significant ($\alpha = 0.01$). (**B**) Analysis of previously known liver cancer SNVs as listed in the original publication, where the genotype of each individual variant is visualised by different colours. White squares indicate that no confident variant was called for that position in that particular sample. This figure was created using the `plot_variant_list` seqCAT function.

these impacts across profile comparisons. Figure 3 shows the impact distributions of matching and mismatching variants for an aggregation of all comparisons between samples in the tissue-to-organoid transition as well as through the long-term culturing process. There is a higher proportion of mismatching MODIFIER variants, and there are only a limited number of mismatching HIGH variants.

In order to investigate if any of these mismatching variants are biologically relevant, we performed GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment using DAVID[20] on genes affected by mismatching variants in the HIGH and MODERATE impact categories. While no terms were significantly enriched for the tissue-to-organoid transition ($\alpha = 0.01$), three olfactory-related terms and one related to protein de-ubiquitination were significantly enriched for long-term culturing comparisons (see Supplementary Data 3 and Supplementary Data 4).

In summary, these results corroborate the original authors' conclusion that the organoids are accurate and genetically stable *in vitro* models of liver cancer and demonstrate how seqCAT can be used to analyse genetic variation in HTS data.

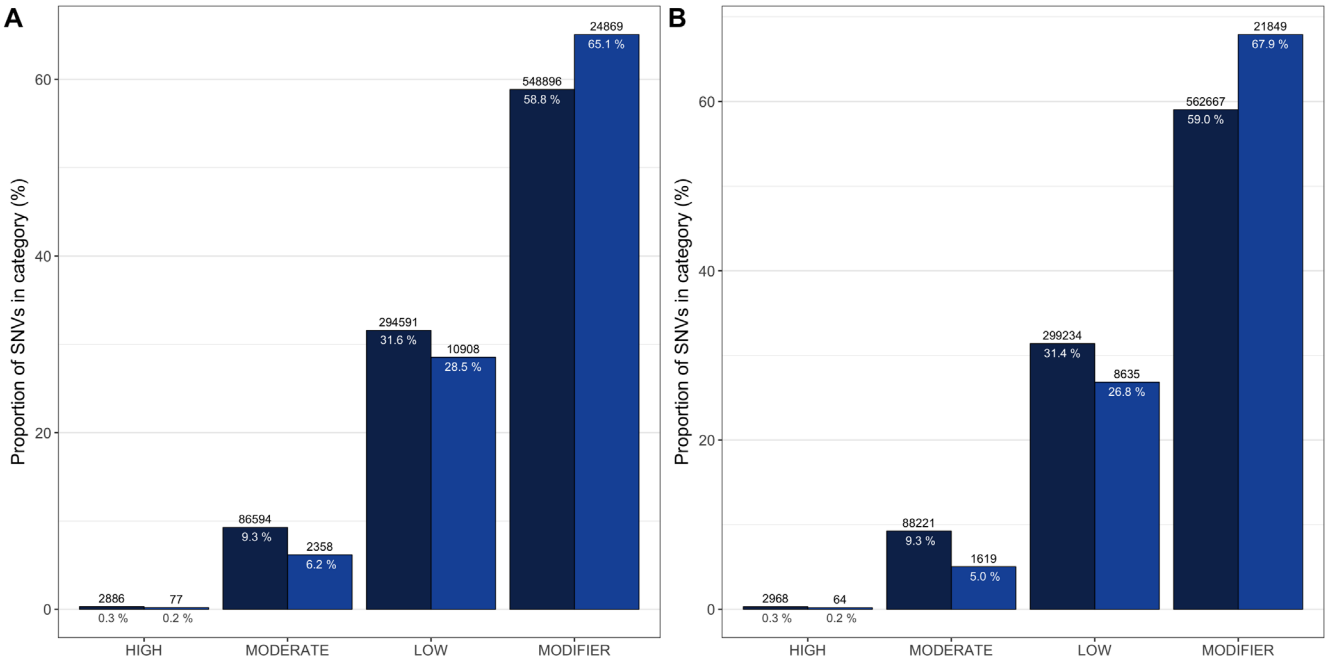

Figure 3. Distribution of variant impacts for the an aggregate of all pairwise comparisons between tissue and early organoid cultures (**A**), and early versus late organoid cultures (**B**). Matching variants (i.e. variants with identical genotypes for both samples being compared) are dark blue, while mismatching variants are a lighter shade of blue. These figures were created using the `plot_impacts` seqCAT function.

Using seqCAT to examine differences between DNA and RNA variants

The Broutier dataset contains not only WES data but also RNA-seq data on the same samples, enabling comparison of RNA-seq data to the already performed WES analyses. We thus downloaded the publicly available raw FASTQ files, performed read alignment with the 2-pass mode of STAR[21], variant calling using GATK[22] and annotation using snpEff[16], as previously described[13]. We subsequently used seqCAT to create SNV profiles for each RNA-seq sample and performed pairwise comparisons across all WES and RNA-seq SNV profiles. This resulted in a grouping with high similarities between WES and RNA-seq samples for the same patient (Figure 4).

There are several previously published studies that show discrepancy between DNA and RNA variants with varying extent and proposed causes[23,24]. In order to quantify the differences between DNA- and RNA-based variants in the organoid dataset, the median concordance for all same-sample comparisons was calculated to be 97.5%; the concordance was used in lieu of the similarity score in order to increase comparability with previously published results. This was also performed for sample type-specific comparisons, where the concordance for tissue versus tissue comparisons was 96.5% and 97.7% for organoid versus organoid. Per-patient (*e.g.* CC1 vs. CC1) calculations were also performed, shown in Table 2. The minimum per-patient concordance was 94.8% and the maximum 98.9%, while the minimum for any individual comparison was 81.1% and a maximum of 99.0% (see the Supplementary Code for the calculations). The minimum value of 81.1% (tissue versus tissue for patient CC1) is the only DNA/RNA comparison with a concordance lower than 90%. These concordances are generally higher than the 80 to 90% that have previously been shown[24].

In summary, results from seqCAT demonstrate an overall high level of concordance between DNA and RNA variant calls, but highlight that there is some variation between sample types and patients.

Discussion

HTS experiments are becoming increasingly more common and the need for simple and powerful bioinformatic software is as great as ever. Analyses of genetic variation through *e.g.* SNVs represents a common endeavour for many scientific studies, but the methods and data analysis pipelines used vary. In this study we present seqCAT, an easy-to-use and well-documented Bioconductor[15] R-package that performs variant analyses of HTS data. The capabilities of seqCAT include the creation of SNV profiles (including a five to ten times faster implementation in Python), comparisons of global genetic similarities for all variants common between samples and analyses of single variants or genes of special interest. While the seqCAT package itself is new, the underlying theory and general methodology have previously been used for investigations into cell line authenticity[13] and genetic heterogeneity in public cell line datasets[14].

SeqCAT may be used to analyse both novel sequencing data as well as publicly available data in repositories (such as the GEO)[1], but may also be utilised to define genetic profiles for any sample of interest. Such profiles are of great interest for researchers using model systems (such as cell lines or

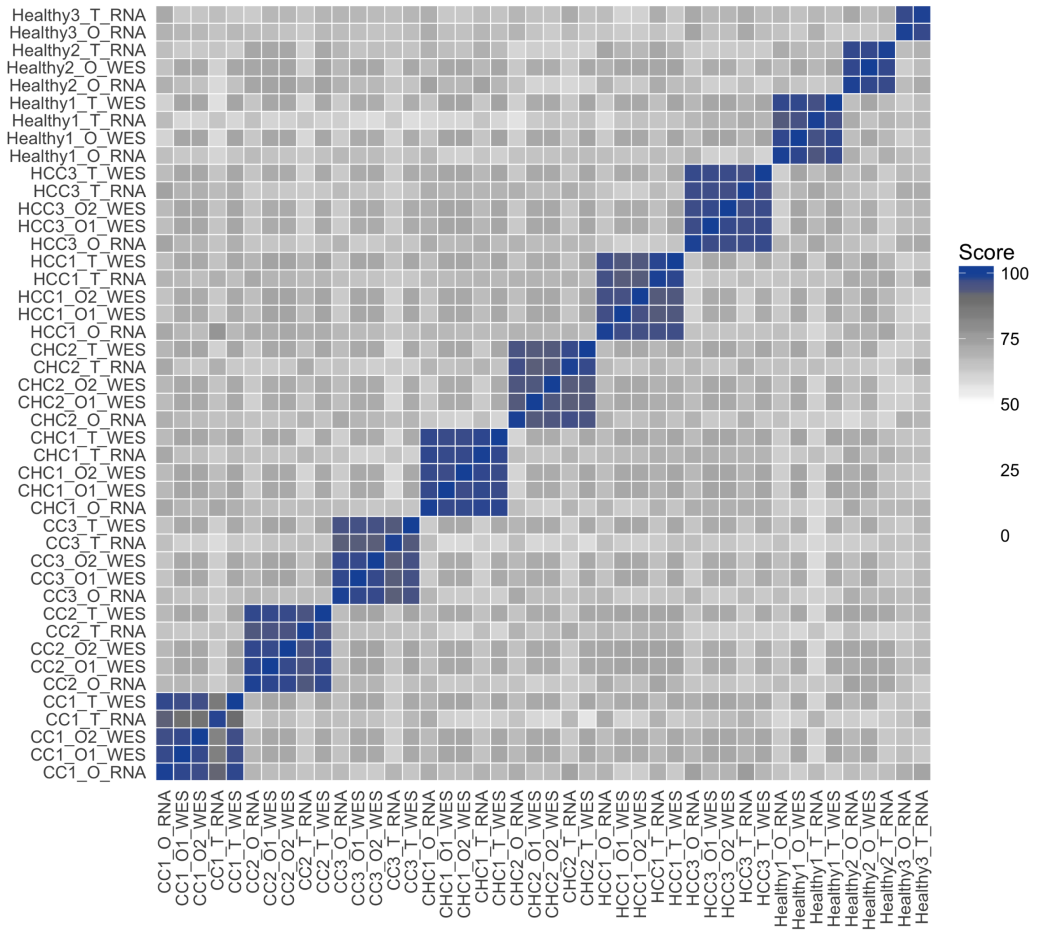

Figure 4. Pairwise comparisons of all WES and RNA-seq SNV profiles, demonstrating the high similarity between DNA/RNA-based variant callings. The colour gradient is the same one used for Figure 1: scores between 0 and 50 are white, scores between 50 and 90 are shown with a white-to-grey gradient, and a grey-to-blue gradient for scores between 90 and 100. This figure was created using the `plot_heatmap` seqCAT function.

Table 2. Median concordance for WES versus RNA-seq SNV profile comparisons across all patients.

Patient	Median concordance	Median overlaps
CC1	96.9%	3744
CC2	98.9%	609
CC3	98.2%	718
CHC1	97.9%	3164
CHC2	95.1%	920
HCC1	96.5%	1872
HCC3	97.3%	745
Healthy1	94.8%	1606
Healthy2	98.7%	1367

organoids), as it allows for a clear definition of the genetic background of the model itself. This could then be referred back to at a later time, to make sure that genetic drift (that obscure interpretation of biological results) has not occurred. SeqCAT is both easy to install and to use, and includes in-depth documentation on its functionality and underlying theory.

In the present study, we have used seqCAT to analyse a publicly available dataset containing WES and RNA-seq data from organoid cultures and their tissues-of-origin[19]. The global analysis of WES SNVs demonstrate the overall high genetic similarities between the organoids and their respective tissues, with equivalent results for comparisons covering all variants or only missense variants. The seqCAT-analysis of known variants indicate that a GPRIN1 variant is present for the CC1 patient; this variant is only listed as present in a CHC-type patient in the original study. Given the importance of these previously

known variants it is likely that the GPRIN1 mutation may be of significance not only for the originally listed CHC1 patient, but also for CC1 patients. The results presented herein corroborate the original authors' conclusions that organoids are genetically stable over time, but higher level of genetic similarity between early and long-term cultured organoids as compared to the tissue-to-organoid transition is statistically non-significant.

The analyses of genes affected by mismatching HIGH and MODERATE impact variants show that none of the differences between tissue and initial organoid cultures are significantly enriched for specific biological functions, indicating that these differences likely are random. The transition from primary tissue to organoid can thus be viewed as a highly stable transition, especially given the high overall similarity previously discussed. The long-term culturing results do, however, present four significantly enriched terms. Three of these are related to ectopic expression of olfactory receptors, which have previously been shown to be present in both healthy and cancerous tissues[25,26]. The single GO-term related to protein de-ubiquitination may be important for studies investigating ubiquitination in liver cancer. Both of these points should thus be accounted for when performing a study with these organoids. The overall results yielded by the seqCAT-analyses corroborate the conclusions from the original study, *i.e.* that these organoids are genetically stable and suitable models for studying liver cancer.

There have been several studies comparing variant calls from DNA and RNA of the same samples, but they have come to differing conclusions as to both the extent and causes of the DNA/RNA discrepancies. Li *et al.* performed both DNA/RNA-seq across 27 individuals in addition to analyses of protein expression using mass spectrometry, where peptides corresponding to variants found in both DNA and RNA were present[23]. They argue that their results indicate biological significance of RNA variants, given that they are translated to proteins, and that the differences between DNA and RNA variants can be biologically meaningful. Indeed, there have been several studies analysing RNA-seq variants that yielded novel biological insights, demonstrating the utility of such endeavours[27–31]. A study by Guo *et al.* analysed DNA/RNA-seq data for 10 breast cancer patients from the TCGA and calculated DNA/RNA concordances to range between from 80 to 90%[24]. They argue that these differences are mostly technical rather than biological.

The results of the present study indicate that the extent of DNA/RNA differences may not be as large as previously shown: the median concordance for DNA/RNA pairs was 97.5% overall, with a range of 90 to 99% (plus a single comparison with 81.1%), while Guo *et al.* reported a range of 80 to 90% concordance. Both studies thus find a discrepancy between DNA- and RNA-based variant calls, but disagree on its extent. The RNA-seq pipeline utilised in this study is based on the current best practices of GATK, which uses the STAR software for read alignment that has proven to be highly accurate for RNA-seq data[21,32]. The latest assembly of the human genome (GRCh38) was also used, as the choice of assembly has been highlighted as an important parameter that can yield higher accuracy[24]. Guo *et al.* used an earlier assembly from 2009 (GRch37), which might partly explain the discrepancy between the results. While technical issues will always exist even for DNA/DNA or RNA/RNA comparisons, the results of the present study may represent a closer estimate of the biological relevance of DNA/RNA differences first noted by Li *et al.*

It is clear is that there is a discrepancy between DNA- and RNA-based variant calls, but the exact extent of this difference remains to be determined, as well as whether it is a consequence of technical artefacts or biological variation. A full evaluation of these matters likely require a larger study than what has previously been attempted, including using the latest technologies as well as protein-level validation. The analyses performed herein demonstrate how seqCAT may be utilised as a part of such an endeavour.

Conclusions

The seqCAT Bioconductor R-package provides an effective and easy-to-use toolkit for analysing HTS variant data, enabling researchers to investigate genetic differences and potential variation within and between their samples or publicly available data from other laboratories. Little R expertise is required to use seqCAT, and its use is extensively documented. We have used seqCAT to analyse genetic variation in a publicly available dataset of liver cancer organoids, corroborating the conclusions drawn by its original authors, as well as demonstrate high levels of DNA/RNA SNV concordance in this dataset. These results serve as a case study in how to utilise the capabilities of seqCAT, which make it a valuable and intuitive tool for a wide range of researchers.

Software and data availability

Software is available from: https://bioconductor.org/packages/release/bioc/html/seqCAT.html

Source code available from: https://github.com/fasterius/seqCAT

Archived source code as at time of publication: https://doi.org/10.5281/zenodo.1404027[33]

Software license: MIT

The data used in this article is publicly available at the GEO through the accession number GSE84073.

Grant information

This work was supported by the European Community 7th Framework Program under grant agreement no. 278 568 "PRIMES".

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

Acknowledgements

We would like to acknowledge support from Science for Life Laboratory (SciLifeLab), the National Genomics Infrastructure (NGI) and Uppmax for providing assistance in computational infrastructure.

Supplementary material

Supplementary Code: A RMarkdown document for reproducing the analyses and figures of the study using the seqCAT package.

Click here to access the data.

Supplementary Data 1: Metadata for the Broutier *et al.* study[19].

Click here to access the data.

Supplementary Data 2: List of the previously known SNVs used in the Broutier *et al.* study[19].

Click here to access the data.

Supplementary Data 3: Full results of the enrichment analysis of tissue versus established organoids.

Click here to access the data.

Supplementary Data 4: Full results of the enrichment analysis of established organoids versus long-term cultured organoids.

Click here to access the data.

References

1. Edgar R, Domrachev RM, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res.* 2002; **30**(1): 207–210.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Zhu Y, Stephens RM, Meltzer PS, *et al.*: **SRAdb: query and use public next-generation sequencing data from within R.** *BMC Bioinformatics.* 2013; **14**: 19.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Heather JM, Chain B: **The sequence of sequencers: The history of sequencing DNA.** *Genomics.* 2016; **107**(1): 1–8.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Sboner A, Mu XJ, Greenbaum D, *et al.*: **The real cost of sequencing: higher than you think!** *Genome Biol.* 2011; **12**(8): 125.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Muir P, Li S, Lou S, *et al.*: **The real cost of sequencing: scaling computation to keep pace with data generation.** *Genome Biol.* 2016; **17**: 53.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Danecek P, Auton A, Abecasis G, *et al.*: **The variant call format and VCFtools.** *Bioinformatics.* 2011; **27**(15): 2156–2158.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Obenchain V, Lawrence M, Carey V, *et al.*: **VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants.** *Bioinformatics.* 2014; **30**(14): 2076–2078.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.*: **Integrative genomics viewer.** *Nat Biotechnol.* 2011; **29**(1): 24–26.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Zerbino DR, Achuthan P, Akanni W, *et al.*: **Ensembl 2018.** *Nucleic Acids Res.* 2018; **46**(D1): D754–D761.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Pabinger S, Dander A, Fischer M, *et al.*: **A survey of tools for variant analysis of next-generation genome sequencing data.** *Brief Bioinform.* 2014; **15**(2): 256–278.
PubMed Abstract | Publisher Full Text | Free Full Text
11. QIAGEN: **Ingenuity Variant Analysis**. 2018; accessed 2018-05-30.
Reference Source
12. Capes-Davis A, Neve RM: **Authentication: A Standard Problem or a Problem of Standards?** *PLoS Biol.* 2016; **14**(6): e1002477.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Fasterius E, Raso C, Kennedy S, *et al.*: **A novel RNA sequencing data analysis method for cell line authentication.** *PLoS One.* 2017; **12**(2): e0171435.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Fasterius E, Al-Khalili Szigyarto C: **Analysis of public RNA-sequencing data reveals biological consequences of genetic heterogeneity in cell line populations.** *Sci Rep.* 2018; **8**(1): 11226.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Huber W, Carey VJ, Gentleman R, *et al.*: **Orchestrating high-throughput genomic analysis with Bioconductor.** *Nat Methods.* 2015; **12**(2): 115–121.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Cingolani P, Platts A, Wang le L, *et al.*: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3.** *Fly (Austin).* 2012; **6**(2): 80–92.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Forbes SA, Beare D, Gunasekaran P, *et al.*: **COSMIC: exploring the world's knowledge of somatic mutations in human cancer.** *Nucleic Acids Res.* 2015; **43**(Database issue): D805–11.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Lawrence M, Huber W, Pagès H, *et al.*: **Software for computing and annotating genomic ranges.** *PLoS Comput Biol.* 2013; **9**(8): e1003118.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Broutier L, Mastrogiovanni G, Verstegen MM, *et al.*: **Human primary liver cancer-derived organoid cultures for disease modeling and drug screening.** *Nat Med.* 2017; **23**(12): 1424–1435.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc.* 2008; **4**(1): 44–57.
PubMed Abstract | Publisher Full Text
21. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text
22. 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): 1297–1303.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Li M, Wang IX, Li Y, *et al.*: **Widespread RNA and DNA sequence differences in the human transcriptome.** *Science.* 2011; **333**(6038): 53–58.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Guo Y, Zhao S, Sheng Q, *et al.*: **The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data.** *BMC Genomics.* 2017; **18**(Suppl 6): 690.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Flegel C, Manteniotis S, Osthold S, *et al.*: **Expression profile of ectopic olfactory receptors determined by deep sequencing.** *PLoS One.* 2013; **8**(2): e55368.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Abaffy T: **Human olfactory receptors expression and their role in non-olfactory tissues-a mini-review.** *J Pharmacogenomics Pharmacoproteomics.* 2015; **6**: 152.
Publisher Full Text
27. Miller AC, Obholzer ND, Shah AN, *et al.*: **RNA-seq-based mapping and candidate identification of mutations from forward genetic screens.** *Genome Res.* 2013; **23**(4): 679–686.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Piskol R, Ramaswami G, Li JB: **Reliable identification of genomic variants from RNA-seq data.** *Am J Hum Genet.* 2013; **93**(4): 641–651.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Lee MC, Lopez-Diaz FJ, Khan SY, *et al.*: **Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing.** *Proc Natl Acad Sci U S A.* 2014; **111**(44): E4726–E4735.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Deelen P, Zhernakova DV, de Haan M, *et al.*: **Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels.** *Genome Med.* 2015; **7**(1): 30.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Kang HM, Subramaniam M, Targ S, *et al.*: **Multiplexed droplet single-cell RNA-sequencing using natural genetic variation.** *Nat Biotechnol.* 2018; **36**(1): 89–94.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Engström PG, Steijger T, Sipos B, *et al.*: **Systematic evaluation of spliced alignment programs for RNA-seq data.** *Nat Methods.* 2013; **10**(12): 1185–1191.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Fasterius E, vobencha, hpages: **fasterius/seqCAT: seqCAT version 1.2.1 (Version 1.2.1).** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1404027

Open Peer Review

Current Referee Status: ?

Version 1

Referee Report 09 April 2019

https://doi.org/10.5256/f1000research.17563.r46167

? **Matej Lexa**
Department of Machine Learning and Data Processing, Faculty of Informatics, Masaryk University, Botanicka, Czech Republic

The manuscript describes a novel computational tool for genotype analysis and comparison called seqCAT.

The tool has been created as a package for R/Bioconductor and has already been accepted into the Bioconductor repository. I was able to install it and follow the example code in the package as well as study its use in the manuscript. In all my tests the package and its functions worked as designed/described in the accompanying materials.

Although the code is fully functional, both the code and the submitted manuscript leave much to be deserved. The most important issues in this respect are i) the absense of critical comparison with existing tools, ii) better description of some of the available functionality and last but not least, iii) better integration into the existing data and code structure.

- As far as other tools are regarded, the authors cite the need for a tool like seqCAT by referring to vcftools, VariantAnnotation R package, IGV and Ensembl Genome Browser and some proprietary software. However, today there are dozens of tools that may come close to the functionality presented here and deserve to be mentioned and compared critically. Just a quick browsing of several sources yielded software, such as adegenet (https://cran.r-project.org/web/packages/adegenet/index.html), anvi'o (http://merenlab.org/2015/07/20/analyzing-variability/), SomaticSniper (http://gmt.genome.wustl.edu/packages/somatic-sniper/), PhyloSNP (https://hive.biochemistry.gwu.edu/dna.cgi?cmd=phylosnp), GATK or BEDOPS that has a vcf2bed function (https://bedops.readthedocs.io/en/latest/content/reference/file-management/conversion/vcf2bed.ht that can lead to comparison based on interval sets. Would PLINK and its SNP profiling abilities be powerful enough (http://zzz.bwh.harvard.edu/plink/profile.shtml)? Are methods typically used for small and medium-sized SNP samples, such as the MATLAB code here (https://jamanetwork.com/journals/jamaoncology/fullarticle/2598491) different from methods that must be applied to whole-genome data? I don't know the answers to some of these questions but I feel the authors should look wider to show the advantages of seqCAT, if any. One advantage, also mentioned by the authors is simplicity of use. However, it should be clear what the trade-offs are.
- The manuscript mentions SNVs are filtered based on quality and other criteria but doesn't give enough details about what is happening under the hood. The software is open source, however the manuscript should lay out basic principles of data manipulation done by their package in plain

English. Also, reading a profile into a package and comparing it to others create different GenomicRanges/data frame data objects in R that should also be described briefly.

- Loosely connected to the data frame data structures mentioned above, I see the way seqCAT manipulates data as a weak point. First of all, it calculate profiles and saves them into a file, effectively outside R, only to read the files in the next step. It would seem much more natural, to use some internal data structure, maybe even the same data frame created later, to keep the data in R and offer appropriate writing/reading/conversion functions to create files outside R. As for conversions, data formats for some of the data calculated by seqCAT already exist and would make the software much more powerful, if the users could write to them (or even read from them). Although the profiles can be exported into BED/GFF3 with some third party libraries (e.g. rtracklayer), perhaps it would be useful to go to BAM/SAM, back to VCF after some manipulation (right now only filtration, presumably), or hapmap and others for transfer of data into other software (e.g. PLINK, VarDict)?

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.