SOFTWARE

Open Access

Acorn: an R package for de novo variant analysis

Tychele N. Turner[1*]

*Correspondence:
tychele@wustl.edu

[1] Department of Genetics, Washington University School of Medicine, 4523 Clayton Avenue, Campus Box 8232, St. Louis, MO 63110, USA

Abstract

Background: The study of de novo variation is important for assessing biological characteristics of new variation and for studies related to human phenotypes. Software programs exist to call de novo variants and programs also exist to test the burden of these variants in genomic regions; however, I am unaware of a program that fits in between these two aspects of de novo variant assessment. This intermediate space is important for assessing the quality of de novo variants and to understand the characteristics of the callsets. For this reason, I developed an R package called acorn.

Results: Acorn is an R package that examines various features of de novo variants including subsetting the data by individual(s), variant type, or genomic region; calculating features including variant change counts, variant lengths, and presence/absence at CpG sites; and characteristics of parental age in relation to de novo variant counts.

Conclusions: Acorn is an R package that fills a critical gap in assessing de novo variants and will be of benefit to many investigators studying de novo variation.

Keywords: De novo variants, De novo variation, Genomics, Genetics

Background

The study of de novo variants (DNVs) consists of (1) calling of DNVs from parent–child sequencing data, (2) quality control and DNV dataset characterization, and (3) assessment of DNVs for statistical interpretation (e.g., assessment of burden of DNVs in genomic regions in human phenotypes). Several tools exist for step 1 [1–4] and step 3 [5–9]. However, I am unaware of a tool that sits in between these steps at step 2. The analysis at step 2 is critical for inputting high quality data into step 3. I developed acorn as a strategy to centralize code I use for performing these quality checks for public distribution to help others also analyzing DNVs. The implementation of acorn as an R package is key for its use on multiple platforms. Functions are built so investigators can use them in several ways to assess their DNV data.

Implementation

Acorn (version 0.99.9) is an R package with minimal prerequisites and is publicly available at https://github.com/TNTurnerLab/acorn. It requires baseline R packages graphics, stats, and utils making it a quick package to install by the user. Both source code and a pre-built version of the package are available allowing users to download and build the package themselves or quickly install using `R CMD INSTALL acorn_0.99.9.tar.gz`, respectively.

Results and discussion

Acorn functions

Acorn consists of several functions to analyze DNVs. One function is `readDNV`, which reads in DNV data and turns it into an R object for use with other functions within acorn. The next set of functions are used to extract different subsets of DNV data. These include `extractIndividual`, `extractSNVs`, `extractIN-DELs`, `extractMNVs`, `extractAutosomes`, `extractX`, and `extractY`. The `extractIndividual` function can extract an individual or set of individuals from the dataset. The `extractSNVs`, `extractINDELs`, and `extractMNVs` functions are used to extract variant types including single-nucleotide variants (SNVs), small insertions/deletions (INDELs), or multi-nucleotide variants (MNVs) from the dataset, respectively. Finally, the `extractAutosomes`, `extractX`, and `extractY` functions are all used to extract DNVs by genomic regions. The `extractAutosomes` function extracts the autosomes (chromosomes 1 to 22) from the dataset. The `extractX`, and `extractY` functions extract DNVs on the X and Y chromosomes, respectively.

The next set of functions are used for examining summary characteristics of DNV data. These include the `calculateTiTvRatio`, `calculateDeletionInsertionRatio`, `calculateDeletionLengths`, `calculateInsertionLengths`, and the `calculateMNVLengths` functions. The `calculateTiTvRatio` function automatically grabs only the SNVs from the DNV object for the calculation of the transition/transversion (Ti/Tv) ratio. It returns the counts of transitions, the counts of transversions, the Ti/Tv ratio, and a barplot of the different types of SNV changes observed in the DNV object. The `calculateDeletionInsertionRatio` automatically grabs only the INDELs from the DNV object for the calculation of the deletion/insertion ratio. It returns the counts of deletions, the counts of insertions, and the deletion/insertion ratio. The `calculateDeletionLengths`, `calculateInsertionLengths`, and `calculateMNVLengths` functions determine the lengths of the deletions, insertions, and MNVS, respectively. For each, it also returns a barplot of the lengths.

Another key feature of DNV data, is the percent at CpG sites. The `annotateCpG` function annotates and summarizes CpG information for the DNV dataset. It extracts single-nucleotide variants (SNVs) and assigns whether they are at a CpG site or not. This function also requires a pre-computed rda file for the CpG sites in the genome

of interest. I have made this available for b38 of the human genome at: https://data.cyverse.org/dav-anon/iplant/home/tycheleturner/genomic_annotations/cpg_b38.rda. The function returns a DNV object containing only SNVs and includes a column with a note on whether the DNV is at a CpG or not. This function also prints out the number of CpG and the percent of DNV SNVs at CpG.

The `countsPerIndividual` function returns the mean of the DNV counts per individual, the standard deviation of the DNV counts per individual, a plot of the density of the DNV counts per individual, and an object consisting of the sample name and the counts of their DNVs that can be assigned to another object.

The final set of functions focus on the parental age characteristics of DNVs. The first is `parentalAgeObject`, which takes in a counts object that is either the result of the `countsPerIndividual` function or is already read into an object from a file. This function returns back an object with the de novo counts and parental age data together. The fields in this file are sample, dnm_counts, fatherAge, and motherAge. The `parentalAge` function calculates the correlation between father's and mother's age at birth and DNV counts per individual, the results of the linear model taking the form: `lm(formula = dnm_counts ~ fatherAge + motherAge, data = parentalAgeObject)` or the results of the exponential model taking the form: `lm(log(dnm_counts) ~ fatherAge + motherAge, data = parentalAgeObject)`. The linear and exponential models are both classical models for use in assessing DNVs [10]. The input required is output from the `parentalAgeObject` function in this package. This function returns the results of the linear or exponential model and a plot of father's and mother's age at birth and DNV counts. The `fatherAge` and `motherAge` functions are similar to the `parentalAge` function except that they only consider the father's age or mother's age, respectively.

DNV assessment with acorn

We provide test data with acorn, on our acorn GitHub site, consisting of a subset of our DNV dataset from our recent paper [11] (Additional file 1: Table S1), simulated count data (Additional file 2: Table S2), and simulated parental age data (Additional file 3: Table S3). We detail the use of these in the Vignette included with the acorn package. Briefly, the DNV data consists of 9,741 DNVs of which 8,558 are SNVs and 1,183 are INDELs. For SNV changes (Fig. 1A), the number of transitions are 5,708 and the number of transversions are 2850 with a Ti/Tv ratio of 2.0. The number of deletions is 540 and the number of insertions is 643 with a deletion/insertion ratio of 0.8. Acorn also shows the length of deletions (Fig. 1B) and insertions (Fig. 1C). These are just a few ways that acorn can be used to examine DNV data.

For the parental age analysis on the simulated data, acorn presents a plot of DNV counts versus parental age at birth (Figure 1D). It also shows the result of using a linear model to examine DNV counts versus the parental age at birth reporting back the coefficients, r-squared values, and confidence intervals. In the test set, the number of DNVs rises by 1.0 [0.5, 1.6] each increasing year of father's age at birth and rises by 0.7 [0.1, 1.3] for each increasing year of mother's age at birth.
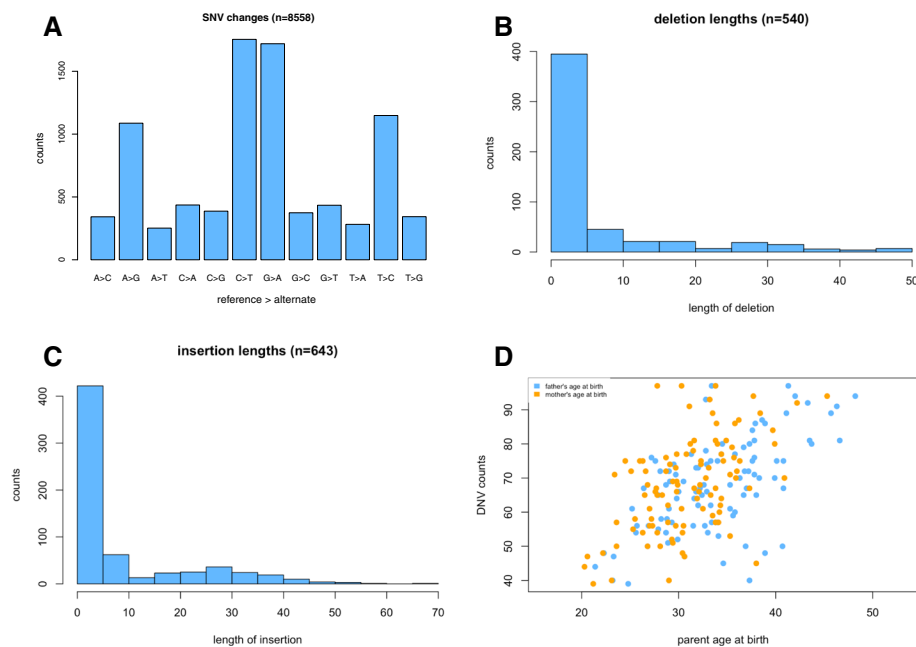
Fig. 1 Example results from acorn. **A** SNV changes as shown using the `calculateTiTvRatio` function, **B** Deletion lengths from the `calculateDeletionLengths` function, **C** Insertion lengths from the `calculateInsertionLengths` function, and **D** DNV counts versus parental age at birth from the `parentalAge` function

Conclusions

The summary statistics and metrics generated using acorn will help users to determine the quality of their DNV callset(s) and whether they should move forward with advanced statistical tests. Acorn fills a gap in genomic DNV analyses between the calling of DNVs and ultimate downstream statistical assessment (i.e., DNV association with phenotypes).

Availability and requirements

Project name: acorn.

Project home page: https://github.com/TNTurnerLab/acorn

Operating system(s): Platform independent.

Programming language: R

Other requirements: R 3.5.0

License: MIT.

Any restrictions to use by non-academics: none.

Abbreviations
DNV	De novo Variant
INDEL	Insertion/deletion
MNV	Multi-nucleotide variant
SNV	Single-nucleotide variant
Ti	Transition
Tv	Transversion

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05457-z.

Additional file 1. de novo variants from Ng et al. 2022, Human Mutation [11]. These de novo variants are a subset of the published variants and are used in this paper to show case uses of acorn for assessing de novo variants.

Additional file 2. Simulated de novo variant counts in a sample dataset. These samples and de novo variant counts are simulated based on real data. They are used to demonstrate the parental age functions in acorn.

Additional file 3. Simulated parental ages in a sample dataset. These samples and parental ages are simulated based on real data. They are used to demonstrate the parental age functions in acorn.

Author contributions
TNT wrote the acorn R package, performed analyses, and wrote the paper.

Availability of data and materials
All data generated or analyzed during this study are included in this published article in the supplementary information files and/or at the acorn GitHub site at https://github.com/TNTurnerLab/acorn.

Declarations

Ethics approval and consent to participate
Not applicable as the data in this study is already published, publicly available, de-identified human de novo variants.

Consent for publication
Not applicable.

Competing interests
Not applicable.

References
1. Wei Q, Zhan X, Zhong X, Liu Y, Han Y, Chen W, Li B: A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics (Oxford, England)* 2014.
2. Ng JK, Turner TN: HAT: de novo variant calling for highly accurate short-read and long-read sequencing data. *bioRxiv* 2023:2023.2001.2027.525940.
3. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012;151(7):1431–42.
4. Lian A, Guevara J, Xia K, Sebat J. Customized de novo mutation detection for any variant calling pipeline: Synth-DNM. Bioinformatics (Oxford, England). 2021;37(20):3640–1.
5. Coe BP, Stessman HAF, Sulovari A, Geisheker MR, Bakken TE, Lake AM, Dougherty JD, Lein ES, Hormozdiari F, Bernier RA, et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat Genet. 2019;51(1):106–16.
6. Ware JS, Samocha KE, Homsy J, Daly MJ. Interpreting de novo variation in human disease using denovolyzeR. Curr Protoc Hum Genet. 2015;87:7–25.
7. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013;9(8): e1003671.
8. Kaplanis J, Samocha KE, Wiel L, Zhang Z, Arvai KJ, Eberhardt RY, Gallone G, Lelieveld SH, Martin HC, McRae JF, et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature. 2020;586(7831):757–62.
9. Jiang Y, Han Y, Petrovski S, Owzar K, Goldstein DB, Allen AS. Incorporating functional information in tests of excess de novo mutational load. Am J Hum Genet. 2015;97(2):272–83.
10. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488(7412):471–5.

11. Ng JK, Vats P, Fritz-Waters E, Sarkar S, Sams EI, Padhi EM, Payne ZL, Leonard S, West MA, Prince C, et al. de novo variant calling identifies cancer mutation signatures in the 1000 Genomes Project. Hum Mutat. 2022;43(12):1979–93.

Publisher's Note

Dr. Tychele N. Turner   is a human genomicist who studies the role of genomic variation in neurodevelopmental disorders and has published several papers on de novo variation.