Genome Biology

OPEN LETTER

Dissemination of scientific software with Galaxy ToolShed

Daniel Blankenberg[1,4], Gregory Von Kuster[1,4], Emil Bouvier[1,4], Dannon Baker[2,4], Enis Afgan[4,5], Nicholas Stoler[3], the Galaxy Team[4], James Taylor[2,4*] and Anton Nekrutenko[1,4*]

Abstract

The proliferation of web-based integrative analysis frameworks has enabled users to perform complex analyses directly through the web. Unfortunately, it also revoked the freedom to easily select the most appropriate tools. To address this, we have developed Galaxy ToolShed.

Previously, our group has investigated the persistence of mitochondrial variants (heteroplasmies) through mother-child transmissions [1]. Many disease-causing mitochondrial variants are heteroplasmic and their clinical manifestations depend on the relative proportion of normal to mutant alleles [2-4]. Because almost all of the mitochondrial genome is transcribed [5], the next important question is whether the relative frequencies of heteroplasmic alleles are maintained in transcripts. We turned to published studies to find the appropriate dataset that would include matched genomic and transcriptomic data. The initial analysis of DNA/RNA differences by Li *et al.* [6] omitted the mitochondrial transcriptome and a much more comprehensive dataset by Chen *et al.* [7] has since become available. The latter contains both whole genome and RNA sequencing data from a single individual and is therefore ideally suited for our purpose. To perform this analysis, we started with a 'clean' Galaxy Amazon EC2 instance [8-10], mapped the reads against the latest version of the human genome, retained properly mapped pairs, removed reads mapping to multiple locations, added readgroup information, and combined all results into a single binary version of the sequence alignment/map format (BAM) dataset for further analysis (Additional file 1) [11].

At this point in the analysis, we ran into the first roadblock: the Galaxy instance we were using did not contain any tools for detecting sequence variants. This is exactly the type of situation where the ToolShed is the most useful, as it already contains a collection of utilities for variant detection such as FreeBayes [12]. Installing the FreeBayes tool along with the required dependencies into Galaxy using the ToolShed is accomplished through the web-based graphical user interface [11]. Behind the scenes, the ToolShed fetches source code from the FreeBayes GitHub repository, compiles it, and registers all necessary components with the Galaxy instance, making it accessible to the user [13]. Application of FreeBayes to our dataset has identified two potential heteroplasmic sites with minor allele frequencies >2% (a heteroplasmy detection threshold derived from empirical and simulation data [1,14]): 2,619 and 13,636 (Figure 1a,b). Site 13,363 is a textbook example of a heteroplasmy - it is biallelic (T/C) with an average minor allele frequency of 22% across the 21 samples in our study. However, the other site, 2,619, is different and represents a potential RNA modification reported recently by our group [15]. Within genomic DNA it is represented by an invariable A, while in all RNA-seq datasets it is scored by FreeBayes as a heterozygous locus with the major allele being a T. Moreover, while the total coverage at this site across all samples was 40,132, the numbers of reference and alternative observations were 11,086 and 20,584, respectively (summing to a total of 31,670), suggesting that the site is multiallelic. FreeBayes used here only reports two possibilities: reference and alternative. However, in many cases, such as genotyping of pooled, bacterial or viral samples, it is necessary to report exact counts for all variants. In a typical sequence analysis experiment this is the point where custom scripts are often being developed. While we did exactly that - developed two custom Python-based tools, 'Naïve Variant Caller' (NVC) and

* Correspondence: james@jamestaylor.org; anton@bx.psu.edu
[2]Departments of Biology and Computer Sciences, Johns Hopkins University, Baltimore, MD 21218, USA
[1]Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA 16802, USA
Full list of author information is available at the end of the article

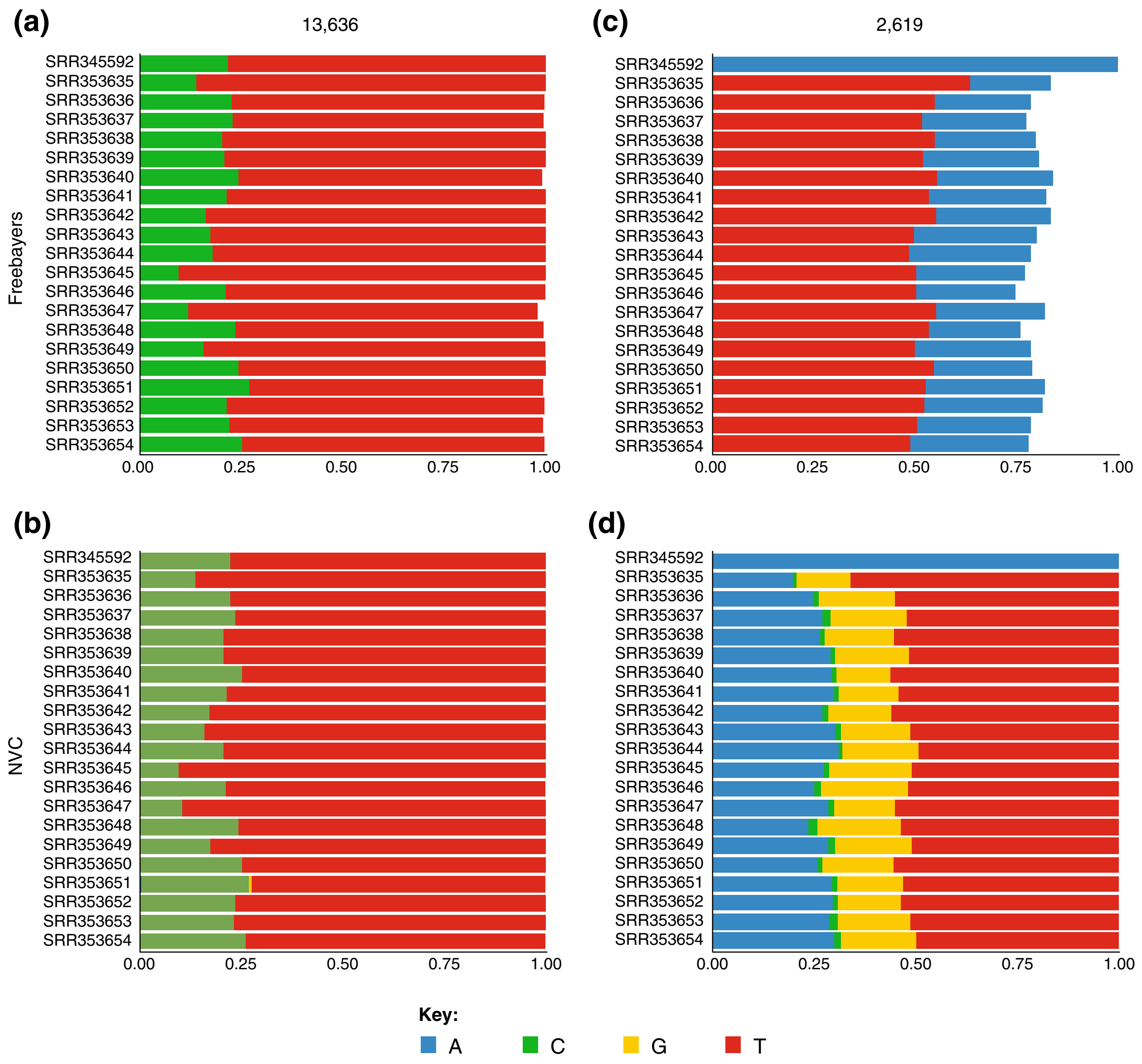

Figure 1 Frequency of the four possible nucleotides across genomic DNA (accession number SRR345592) and RNA-seq (accession numbers SRR353635-SRR353654) samples for sites 13,636 and 2,619. NVC, Naïve Variant Caller. Data is deposited in the Short Read Archive at the National Center for Biotechnology Information (NCBI).

'Variant Annotator' - we went a step further and deposited these tools into the ToolShed. By doing so, we not only made it accessible to any Galaxy instance, but also ensured reproducibility of our experiment, which is almost universally lacking in studies utilizing custom scripts [16]. The NVC produces Variant Call Format (VCF) output [17] containing counts for all observed variants from multisample BAM datasets (Additional file 1), while Variant Annotator converts VCF data into allele counts stratified by samples. To deposit the tools into the ToolShed, we have created a version-controlled repository and uploaded all software components, including the tool configuration file, NVC Python script, information about necessary software dependencies, and a set of functional tests. At this point, the tool becomes 'visible' to any Galaxy installation, including the cloud-based instance we use in this study. After installing the NVC from the ToolShed [18], we have applied it to the original BAM dataset to obtain counts shown in Figure 1c,d. Here the multiallelic nature of site 2,619 is clearly seen as well as the fact that this variation only appears in transcriptome data.

This short example has illustrated that the ToolShed behaves as a *de facto* AppStore: when users need an analysis tool that is not present in a given Galaxy instance, it can be easily fetched and installed. Just like a brand new iPad, Galaxy comes with a small number of preinstalled applications providing basic functionality. Additional tools may subsequently be installed from the ToolShed to create a 'flavor' of Galaxy suitable for a particular analysis. An expanded discussion of the ToolShed can be found in the online supplement.

Additional file

Additional file 1: Contains examples of tools deposited to ToolShed and discusses implications of this system for improving the reproducibility of biomedical research.

Abbreviations

BAM: Binary version of the sequence alignment/map format; NVC: Naïve Variant Caller; VCF: Variant call format.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The efforts of the Galaxy Team (Dave Clements, Nate Coraor, Carl Eberhard, Dorine Francheteau, Jeremy Goecks, Sam Guerler and Jennifer Jackson) were instrumental for making this work happen. Special thanks to the members of the ToolShed oversight committee (Ira Cooke, Jim Johnson, Ed Kirton, Peter Cock, Brad Chapman, Björn Grüning, Ross Lazarus) for continuing their review of tools being submitted to the ToolShed. This project was supported through grant number HG005542 from the National Human Genome Research Institute, National Institutes of Health as well as grants HG005133, HG004909 and HG006620 and NSF grant DBI 0543285. Additional funding is provided by Huck Institutes for the Life Sciences at Penn State and, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Author details

[1]Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA 16802, USA. [2]Departments of Biology and Computer Sciences, Johns Hopkins University, Baltimore, MD 21218, USA. [3]Interdisciplinary Graduate Program in BioSciences, Penn State University, University Park, PA 16802, USA. [4]Galaxyproject.org, University Park, PA 16802, USA and Baltimore, MD 21218, USA. [5]Center for Informatics and Computing, Ruđer Bošković Institute, Zagreb 1000, Croatia.

Published: 20 February 2014

References

1. Goto H, Dickins B, Afgan E, Paul IM, Taylor J, Makova KD, Nekrutenko A: Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biol* 2011, **12:**R59.
2. Chinnery PF, Thorburn DR, Samuels DC, White SL, Dahl HM, Turnbull DM, Lightowlers RN, Howell N: The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet* 2000, **16:**500–505.
3. Jacobs HT: Making mitochondrial mutants. *Trends Genet* 2001, **17:**653–660.
4. DiMauro S: Mitochondrial diseases. *Biochim Biophys Acta* 2004, **1658:**80–88.
5. Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood A-MJ, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA, Filipovska A, Mattick JS: The human mitochondrial transcriptome. *Cell* 2011, **146:**645–658.
6. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG: Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 2011, **333:**53–58.
7. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, *et al*: Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012, **148:**1293–1307.
8. Blankenberg D, Taylor J, Schenck I, He J, Zhang Y, Ghent M, Veeraraghavan N, Albert I, Miller W, Makova KD, Hardison RC, Nekrutenko A: A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res* 2007, **17:**960–964.
9. Goecks J, Nekrutenko A, Taylor J: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010, **11:**R86.
10. Afgan E, Baker D, Coraor N, Goto H, Paul IM, Makova KD, Nekrutenko A, Taylor J: Harnessing cloud computing with Galaxy Cloud. *Nat Biotechnol* 2011, **29:**972–974.
11. Introduction to Galaxy ToolShed 1 [http://vimeo.com/73458993].
12. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR: A general approach to single-nucleotide polymorphism discovery. *Nat Genet* 1999, **23:**452–456.
13. Introduction to Galaxy ToolShed 2 [http://vimeo.com/73460697].
14. Li M, Schonberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M: Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am J Hum Genet* 2010, **87:**237–249.
15. Bar-Yaacov D, Avital G, Levin L, Richards A, Hachen N, Rebolledo Jaramillo B, Nekrutenko A, Zarivach R, Mishmar D: RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. *Genome Res* 2013, **23:**1789–1796.
16. Nekrutenko A, Taylor J: Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet* 2012, **13:**667–672.
17. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, Depristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group: The variant call format and VCFtools. *Bioinformatics* 2011, **27:**2156–2158.
18. Introduction to Galaxy ToolShed 3 [https://vimeo.com/73462389].

doi:10.1186/gb4161
Cite this article as: Blankenberg *et al.*: Dissemination of scientific software with Galaxy ToolShed. *Genome Biology* 2014 **15**:403.