RESEARCH ARTICLE

The GenABEL Project for statistical genomics [version 1; referees: 2 approved]

Lennart C. Karssen[1,2], Cornelia M. van Duijn[2], Yurii S. Aulchenko[1,3-5]

[1]PolyOmica, Groningen, 9722 HC, Netherlands
[2]Department of Epidemiology, Erasmus Medical Center, Rotterdam, 3000 CA, Netherlands
[3]Institute of Cytology and Genetics, Siberian Division of the Russian Academy of Sciences, Novosibirsk, 630090, Russian Federation
[4]Novosibirsk State University, Novosibirsk, 630090, Russian Federation
[5]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, UK

v1 **First published:** 19 May 2016, **5**:914 (doi: 10.12688/f1000research.8733.1)
Latest published: 19 May 2016, **5**:914 (doi: 10.12688/f1000research.8733.1)

Abstract

Development of free/libre open source software is usually done by a community of people with an interest in the tool. For scientific software, however, this is less often the case. Most scientific software is written by only a few authors, often a student working on a thesis. Once the paper describing the tool has been published, the tool is no longer developed further and is left to its own device. Here we describe the broad, multidisciplinary community we formed around a set of tools for statistical genomics. The GenABEL project for statistical omics actively promotes open interdisciplinary development of statistical methodology and its implementation in efficient and user-friendly software under an open source licence. The software tools developed withing the project collectively make up the GenABEL suite, which currently consists of eleven tools. The open framework of the project actively encourages involvement of the community in all stages, from formulation of methodological ideas to application of software to specific data sets. A web forum is used to channel user questions and discussions, further promoting the use of the GenABEL suite. Developer discussions take place on a dedicated mailing list, and development is further supported by robust development practices including use of public version control, code review and continuous integration. Use of this open science model attracts contributions from users and developers outside the “core team”, facilitating agile statistical omics methodology development and fast dissemination.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 19 May 2016	✓ report	✓ report

1 **Giulietta Minozzi**, University of Milan Italy

2 **Bjarni J. Vilhjálmsson**, Aarhus University Denmark

Discuss this article

Comments (0)

Corresponding authors: Lennart C. Karssen (l.c.karssen@polyomica.com), Yurii S. Aulchenko (y.s.aulchenko@polyomica.com)

How to cite this article: Karssen LC, van Duijn CM and Aulchenko YS. **The GenABEL Project for statistical genomics [version 1; referees: 2 approved]** *F1000Research* 2016, **5**:914 (doi: 10.12688/f1000research.8733.1)

Copyright: © 2016 Karssen LC *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: Funding from the following sources enabled work on specific packages in the GenABEL suite: PolyOmica, Groningen, The Netherlands; Centre for Medical Systems Biology (CMSB), The Netherlands; the Netherlands Genomics Initiative (NGI); the Netherlands Organisation for Scientific Research (NWO); the Department for Health Evidence, Radboud University Medical Centre, Nijmegen, The Netherlands; Deutsche Forschungsgemeinschaft (German Research Association, grant GSC 111); Russian Foundation for Basic Research (RFBR, 12-04-33182, 15-34-20763, 15-04-07874); the RFBR-Helmholtz society Joint Research Groups programme (12-04-91322); the European Union FP7 framework projects MIMOmics (grant agreement nr. 305280) and Pain-Omics (grant agreement nr. 602736). The work of YSA was supported by a grant from the Russian Science Foundation (RSCF 14-14-00313).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare no conflicts of interest in the authorship or publication of this contribution.

First published: 19 May 2016, **5**:914 (doi: 10.12688/f1000research.8733.1)

Introduction

The field of statistical (gen-)omics lies at the heart of current research into the genetic aetiology of (human) disease and personalized or precision medicine[1]. Genome-wide association studies (GWAS), genotype imputation and next-generation sequencing (NGS) are just a few of the techniques used in this field that is driven by increasingly larger data sets[2,3]. With the advent of polyphenotype analysis as is now customary in e.g. lipidomics and metabolomics, the issues of dealing with big data have become imminent[4,5]. In recent years, scientists and funding organizations alike have come to realize that in order to successfully tackle the challenges of the field, close collaboration between various disciplines, e.g. statistics, molecular biology, genetics, and computer science, is of paramount importance[2,6,7].

Many software tools developed by scientists are distributed as free/libre open source software (FLOSS). FLOSS tools are often developed by groups of people with different backgrounds, working from different geographical locations, either under central guidance or in a loose cooperation, sometimes as part of their employment, sometimes "just for fun"[8]. The key to successful, sustainable open source software is an active community of both developers/contributors and end users[9]. Unfortunately, creators of scientific software are usually not funded to actively build such a community. Moreover, our experience shows that once the peer-reviewed article describing a tool has been published, funding and time to continue development and support of that tool are usually limited or non-existent, and consequently, the tool often slowly fades into oblivion. It needs no explanation that this amounts to a waste of effort and money.

It was with these premises in mind that the *GenABEL Project for Statistical Genomics* was started as an extension of the original community of users and developers around the `GenABEL` package[10].

The GenABEL project

The GenABEL project aims to provide a framework for collaborative, sustainable, robust, transparent, opensource based development of statistical genomics methodology. Within the project, statisticians devoted to method development work together with statistical geneticists and biologists to refine existing statistical methods as well as develop new ones and make them applicable to genomic analysis. With the help of computer scientists and scientific software developers these mathematical models are then implemented into efficient and user-friendly software. This flow of work and information is not linear, but rather more circular in nature, with information and feedback being continuously transferred between the various layers as depicted in Figure 1. In short, it is a form of agile community-driven development[11,12].

Openness is an important aspect of the GenABEL project[13]. It enables a free flow of information between the layers in the project resulting in rapid feedback between the various levels. Not only do we require that all tools are released under an open source or free software licence like the GNU Public Licence (GPL), we also try to create an atmosphere of open communication using public mailing lists and web forums (see the sections Interaction with the user community and Development infrastructure below). Moreover, because of this openness results of the project (i.e. statistical methods as well as software packages) are easily disseminated among the end users, be they epidemiologists, bioinformaticians or others.

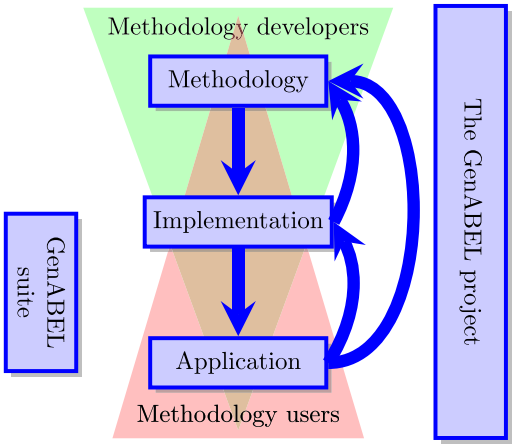

Figure 1. The structure of the GenABEL project and the information flow within it.

The GenABEL suite

The software tools developed within the GenABEL project collectively make up the GenABEL suite. Many tools are R packages, however, this is not a requirement for inclusion in the suite. Any software that is related to the field of statistical (gen-)omics is welcome (technical requirements are discussed in section Development infrastructure). Currently, the suite consists of 11 officially released tools (cf. Table 1) and two that are in beta stage.

The `GenABEL` R package (not to be confused with the GenABEL project or the GenABEL suite), provides an efficient file format for storing genotype data and facilitates pre-GWAS quality control as well as running GWAS of continuous and binary phenotypes, and time-to-event data. The collaborative nature of the project is demonstrated in the `GenABEL` package as it implements several statistical methods developed within the framework, including approximate mixed models[21–23] and various methods for genomic control[24,25]. This shows that the project is really a platform for implementation of (statistical) methods which removes the burden of thinking about data formats etc. allowing method developers to focus on what they do best. The `GenABEL` package is the most popular package in the GenABEL suite with more than 809 citations of its paper (according to Google Scholar)[10].

`ProbABEL` is a tool for running GWAS on imputed genotype data. Like the `GenABEL` package it allows running linear or logistic regression, as well as Cox proportional hazards model, however, `ProbABEL` is tailored to the large file sizes that are inherent to current data sets with approximately 30 million imputed genotypes per individual. It is the second most-used tool from the suite with more than 267 citations (according to Google Scholar)[14].

Table 1. The tools included in the GenABEL suite. Currently, all tools are licensed under the GNU Public Licence (GPL).

Tool	Year of first release	Year of latest release	Latest version
GenABEL[10]	2007	2014	1.8-0
ProbABEL[14]	2009	2016	0.5.0
MetABEL	2009	2014	0.2-0
DatABEL	2010	2015	0.9-5
MixABEL	2010	2015	0.1-3
ParallABEL[15]	2010	2015	0.2-0
VariABEL[16]	2011	2014	0.9-2
PredictABEL[17]	2011	2014	1.2-2
OmicABEL[18]	2013	2015	0.8.0
RepeatABEL[19]	2015	2015	1.0
CollapsABEL[20]	2016	2016	0.10.8

As indicated by its name, MixABEL is an R package for running genome-wide association analyses using mixed models in quantitative traits.

GWAS usually involves meta-analysis of the regression results of various cohorts. The R package MetABEL provides simple meta-analysis functions including generation of forest plots.

The R package VariABEL can be used to look for variance heterogeneity in genetic studies. Such heterogeneity is an indication of interaction between a genetic marker and either another marker or an unknown factor[16,26].

In 2013 OmicABEL was added to the suite. It contains a high-performance computing based approach facilitating extremely fast mixed-model based regression of multiple omics traits like metabolomics or lipidomics on imputed genotype data[18]. OmicABEL aims to increase computational throughput while reducing memory usage and energy consumption. This was achieved by using optimal (hardware-tailored) algorithms using state-of-the-art linear algebra kernels, incorporating optimizations and avoiding redundant computations.

PredictABEL is an R package for the assessment of genetic risk prediction models. It includes functions to compute univariate and multivariate odds ratios of the predictors, the area under the receiver operating characteristic (ROC) curve (AUC), Hosmer-Lemeshow goodness of fit test, reclassification table, net reclassification improvement and integrated discrimination improvement[17].

RepeatABEL allows one to run a GWAS for multiple observations on related individuals[19]. Like ParallABEL, this package is a great example of contributions by the community since its development was not initiated by the core GenABEL developers.

CollapsABEL is the most recent addition to the GenABEL suite. It is an R library for detecting compound heterozygote (CH) alleles in GWAS. It is a computationally efficient solution for screening general forms of CH alleles in densely imputed microarray or whole genome sequencing datasets[20].

Apart from the aforementioned packages which directly address certain types of analysis and/or data management, several packages in the suite have a supportive role. DatABEL is an R interface to our filevector library which provides a file format that is optimised for fast access to data in matrix form, e.g. imputed genotype data. ParallABEL is an R library for parallel execution of GWAS in R.

The latest stable version of the R packages are available on CRAN (http://cran.r-project.org), the Comprehensive R Archive Network. The source code for the other packages can be downloaded from our website at http://www.genabel.org, from the project's version control server or on GitHub (see section Development infrastructure).

Interaction with the user community

The GenABEL project website is the central hub that points to package descriptions, tutorials, the development website, and other information for potential and existing users and developers. Usage statistics such as number of visits and country of origin of visitors are monitored using Google Analytics (http://www.google.com/analytics/) in order to get an estimate of the number of users of the tools and their origins. As an example of the information that can be obtained from this data, Figure 2 shows the top 20 cities of origin of the visitors of the GenABEL website in the period of 28 April 2015 till 28 April 2016. Only visits lasting more than 60 seconds and cities with more than 15 visits were taken into account in an attempt to filter out "accidental" visits. The website was visited 16319 times in that period, of which 696 visits were from an unknown city.

Collecting visitor data like this helps getting an insight in the institutes that use software from the GenABEL suite, which can then be used to show the impact the tools have, e.g. when applying for funding.

Interaction with the user community is done via social media like Twitter (https://twitter.com/GenAproj) and Facebook (https://www.facebook.com/pages/GenABEL-project/329281857167394), as well as a dedicated mailing list for announcements of new package releases, making it easy for both users and system administrators to keep up to date with new releases and developments in the GenABEL project and the GenABEL suite.

Each tool in the GenABEL suite has its own documentation and the GenABEL Tutorial[27] with more than 260 pages takes the user from learning basic R to performing more complicated analyses, showing how the various packages interconnect. Moreover, several video tutorials are available online.

Interactive user support is mostly done through our forum (http://forum.genabel.org). Having an open forum serves various purposes. First of all it is a central, easy to point to reference. Moreover,

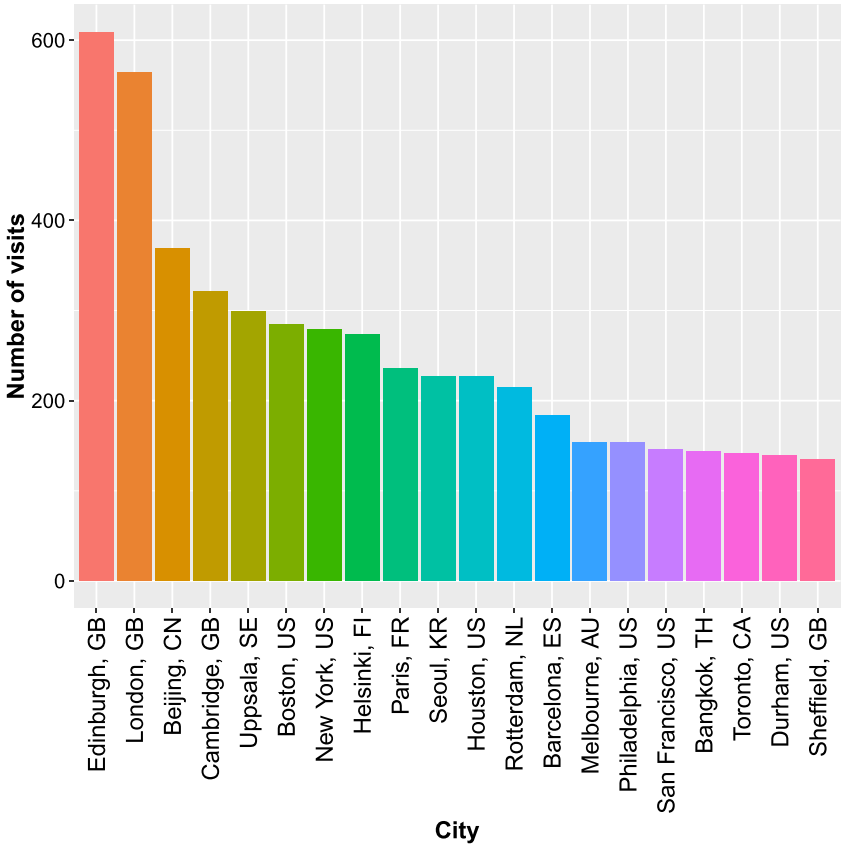

Figure 2. The top 20 cities of origin of visitors to the GenABEL website in the period 28 April 2015 – 28 April 2016. Only visits lasting more than 60 seconds and from cities from which more than 15 visits originated were taken into account. The total number of visits in that period was 16319, of which 696 came from unknown cities. Each city name is followed by the two-letter ISO code of the country in which it is located.

compared to having individual users e-mailing a package author, who may be on holiday or otherwise unavailable, an open forum where users and developers collaborate helps in shortening the time-to-answer. Furthermore, having an active forum where users can help each other allows the developers to focus on fixing bugs and implementing new features. As of March 1st, 2015, the GenABEL forum has 538 activated user accounts, with an average 2.92 new registrations per week since the start of the forum in January 2011. These users have contributed 1422 posts in 427 topics, with an average 7.15 posts per week.

The first hurdle many users of (scientific) software encounter is the installation process. Within the GenABEL project we aim to make installation as simple as possible. Using CRAN for the R packages makes installation and upgrading as simple as typing a single command. For the tools that don't use R, we aim to provide up-to-date packages for various Linux distributions. Currently, `ProbABEL` is packaged in the Stable, Testing and Unstable repositories of Debian with the help of the Debian Med team[28]. Other packages are planned to be added before the end of 2016. For Ubuntu Linux a Personal Package Archive is available (https://launchpad.net/~l.c.karssen/+archive/genabel-ppa). Packages for Red Hat Enterprise Linux and CentOS are on the road map, but haven't been released yet.

Development infrastructure

The GenABEL project welcomes contributions of all sorts, from new tools to fixing spelling errors in the documentation, to bug reports and feature requests. To this end all program code and documentation are either stored in a publicly readable instance of the Subversion version control system, with write access limited to a group of core contributors, or on GitHub (https://github.com/GenABEL-Project), which is one of the leading platforms for what is termed "social coding", which perfectly fits the project's goals. These version control systems record any change to the files so they can easily be reviewed and reverted if necessary[7,29,30].

In November 2010 a mailing list was created as a central place for development discussions. As of April 2016 this list has 34 subscribers. The GenABEL development website (https://r-forge.r-project.org/projects/genabel/) including the Subversion server,

the mailing lists, and the trackers for bugs and feature requests are kindly provided by the R-Forge project[31]. Currently, a total of 94 bugs have been submitted to the bug trackers on R-forge and GitHub since their opening in 2010 and 2015, respectively. Of these 94, 12 were directly contributed by people outside of the core team of developers. Another 42 bug reports were filed by regular contributors based on user reports on the forum, for example, which means that 56% of the bugs have been reported by people in our community that are not core contributors.

In order to be able to maintain the quality of both old and new software in the GenABEL suite prospective tools go through a review process in which both the functional quality of the code is evaluated (does the tool do what it intends to do?), as well as the actual quality of the code (is the code clearly written, including developer documentation in the form of e.g. comments; does the code conform to the GenABEL coding style guidelines; etc.). Moreover, as set out in the GenABEL developer guidelines, we expect commitment of the person or team submitting a tool to the suite to maintain and support it, otherwise the maintenance burden would end up with the core team and it would be too easy to create a tool, write a paper and then 'dump' it in the GenABEL project hoping "the community" will take care of it. Therefore, the community has the option to mark a tool as obsolete, warning the user that bugs will no longer be fixed and support is limited or non-existent.

In 2013 we have started to use a Jenkins Continuous Integration server. Using Jenkins various tests (e.g. regression tests, build tests and tests for memory leaks) are automatically run on each commit to the version control systems. Consequently, changes that break existing functionality are detected at an early stage, thus leading to more stable software releases.

Conclusion

The original publication of the `GenABEL` package for statistical analysis of genotype data[10] has led to the evolution of a community which we now call the GenABEL project, which brings together scientists, software developers and end users with the central goal of making statistical genomics work by openly developing and subsequently implementing statistical models into user-friendly software.

The project has benefited from an open development model, facilitating communication and code sharing between the parties involved. The use of a free software licence for the tools in the GenABEL suite promotes quick uptake and widespread dissemination of new methodologies and tools. Moreover, public access to the source code is an important ingredient for active participation by people from outside the core development team and is paramount for reproducible research. Feedback from end users is actively encouraged through a web forum, which steadily grows into a knowledge base with a multitude of answered questions. Furthermore, our open development process has resulted in transparent development of methods and software, including public code review, a large fraction of bugs being submitted by members of the community, and quick incorporation of bug fixes.

Data and software availability

The file `tracker_report-2016-04-16.csv` contains the data exported from the GenABEL R-forge bug tracker as it was on the date listed in the file name. Because of the recent move of some of the tools from R-forge to Github, the number of issues on the Github pages of the GenABEL project was still low. Therefore, these were counted manually.

The file `Analytics www.genabel.org Locatie Lennart 20150428-20160428.csv` contains the data extracted from the Google Analytics page for the GenABEL website for the period listed in the file name. The columns contain the ISO code of the country, city, number of sessions, number of new viewers, bounce percentage, pages per session and average session duration, respectively.

The file `analysis_GenABELpaper.org` contains the source code used for the automated data extraction for this paper in Emacs Org mode literate programming format (http://orgmode.org)[32].

The code contained in the Org mode file and the data in the csv files listed above are in the public domain (Creative Commons CC0 license) and can be used without restriction.

The data related to the GenABEL forum were extracted manually from the forum control panel.

The tools currently in the GenABEL suite are all Free Software, licensed under the GNU Public License. An up-to-date list of the packages in the suite can be found on http://www.genabel.org/packages, which also contains pointers to the source code of the latest stable versions and the version control repositories on R-forge and GitHub (see the section Development infrastructure above for the URLs).

Archived source code at the time of publication https://zenodo.org/record/51008[33]

Author contributions

CMvD and YSA jointly conceived the GenABEL suite. YSA conceived the idea the GenABEL project and formulated its initial guidelines. LCK and YSA are co-authors and maintainers of various packages in the GenABEL suite and act as maintainers of parts of the project's infrastructure (e.g. forum and mailing lists). LCK drafted the initial version of the manuscript and analyzed the data. All authors contributed to the review of the manuscript and agreed to the final content.

Competing interests

The authors declare no conflicts of interest in the authorship or publication of this contribution.

Grant information

Funding from the following sources enabled work on specific packages in the GenABEL suite: PolyOmica, Groningen, The

Netherlands; Centre for Medical Systems Biology (CMSB), The Netherlands; the Netherlands Genomics Initiative (NGI); the Netherlands Organisation for Scientific Research (NWO); the Department for Health Evidence, Radboud University Medical Centre, Nijmegen, The Netherlands; Deutsche Forschungsgemeinschaft (German Research Association, grant GSC 111); Russian Foundation for Basic Research (RFBR, 12-04-33182, 15-34-20763, 15-04-07874); the RFBR-Helmholtz society Joint Research Groups programme (12-04-91322); the European Union FP7 framework projects MIMOmics (grant agreement nr. 305280) and Pain-Omics (grant agreement nr. 602736). The work of YSA was supported by a grant from the Russian Science Foundation (RSCF 14-14-00313).

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 thank the GenABEL community, i.e. the people who have used the software, filed bug reports, posted on the forum, or contributed tools, patches or ideas, for their involvement.

References

1. Collins FS, Varmus H: **A new initiative on precision medicine.** *N Engl J Med.* 2015; **372**(9): 793–795.
PubMed Abstract | Publisher Full Text

2. Margolis R, Derr L, Dunn M, *et al.*: **The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data.** *J Am Med Inform Assoc.* 2014; **21**(6): 957–958.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Marx V: **Biology: The big challenges of big data.** *Nature.* 2013; **498**(7453): 255–260.
PubMed Abstract | Publisher Full Text

4. Demirkan A, Henneman P, Verhoeven A, *et al.*: **Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses.** *PLoS Genet.* 2015; **11**(1): e1004835.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Demirkan A, van Duijn CM, Ugocsai P, *et al.*: **Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations.** *PLoS Genet.* 2012; **8**(2): e1002490.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Knapp B, Bardenet R, Bernabeu MO, *et al.*: **Ten simple rules for a successful cross-disciplinary collaboration.** *PLoS Comput Biol.* 2015; **11**(4): e1004214.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Merali Z: **Computational science: ...Error.** *Nature.* 2010; **467**(7317): 775–777.
PubMed Abstract | Publisher Full Text

8. Torvalds L, Diamond D: **Just for Fun: The Story of an Accidental Revolutionary.** HarperBusiness, 2002.
Reference Source

9. Fogel K: **Producing Open Source Software: How to Run a Successful Free Software Project.** O'Reilly Media, first edition, 2005; ISBN: 978–0–596–00759–1.
Reference Source

10. Aulchenko YS, Ripke S, Isaacs A, *et al.*: **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics.* 2007; **23**(10): 1294–1296.
PubMed Abstract | Publisher Full Text

11. Kane DW, Hohman MM, Cerami EG, *et al.*: **Agile methods in biomedical software development: a multi-site experience report.** *BMC Bioinformatics.* 2006; **7**: 273.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Budd A, Corpas M, Brazas MD, *et al.*: **A quick guide for building a successful bioinformatics community.** *PLoS Comput Biol.* 2015; **11**(2): e1003972.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Prli´c A, Procter JB: **Ten simple rules for the open development of scientific software.** *PLoS Comput Biol.* 2012; **8**(12): e1002802.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Aulchenko YS, Struchalin MV, van Duijn CM: **ProbABEL package for genome-wide association analysis of imputed data.** *BMC Bioinformatics.* 2010; **11**: 134.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Sangket U, Mahasirimongkol S, Chantratita W, *et al.*: **ParallABEL: an R library for generalized parallelization of genome-wide association studies.** *BMC Bioinformatics.* 2010; **11**: 217.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Struchalin MV, Amin N, Eilers PH, *et al.*: **An R package "VariABEL" for genome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity.** *BMC Genet.* 2012; **13**: 4.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Kundu S, Aulchenko YS, van Duijn CM, *et al.*: **PredictABEL: an R package for the assessment of risk prediction models.** *Eur J Epidemiol.* 2011; **26**(4): 261–264.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Fabregat-Traver D, Sharapov SZh, Hayward C, *et al.*: **High-performance mixed models based genome-wide association analysis with omicABEL software [version 1; referees: 2 approved, 1 approved with reservations].** *F1000Res.* 2014; **3**: 200.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Rönnegård L, McFarlane ES, Husby A, *et al.*: **Increasing the power of genome wide association studies in natural populations using repeated measures-evaluation and implementation.** *Methods Ecol Evol.* 2016.
Publisher Full Text

20. Zhong K, Karssen LC, Kayser M, *et al.*: **CollapsABEL: an R library for detecting compound heterozygote alleles in genome-wide association studies.** *BMC Bioinformatics.* 2016; **17**(1): 156.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Aulchenko YS, de Koning D, Haley C: **Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis.** *Genetics.* 2007; **177**(1): 577–585.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Svishcheva GR, Axenovich TI, Belonogova NM, *et al.*: **Rapid variance components-based method for whole-genome association analysis.** *Nat Genet.* 2012; **44**(10): 1166–1170.
PubMed Abstract | Publisher Full Text

23. Belonogova NM, Svishcheva GR, van Duijn CM, *et al.*: **Region-based association analysis of human quantitative traits in related individuals.** *PLoS One.* 2013; **8**(6): e65395.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Amin N, van Duijn CM, Aulchenko YS: **A genomic background based method for association analysis in related individuals.** *PLoS One.* 2007; **2**(12): e1274.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Tsepilov YA, Ried JS, Strauch K, *et al.*: **Development and application of genomic control methods for genome-wide association studies using non-additive models.** *PLoS One.* 2013; **8**(12): e81431.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Struchalin MV, Dehghan A, Witteman JC, *et al.*: **Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations.** *BMC Genet.* 2010; **11**: 92.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Aulchenko YS, Karssen LC, The GenABEL project developers: **The GenABEL Tutorial.** *Zenodo.* 2015.
Publisher Full Text

28. Möller S, Krabbenhöft HN, Tille A, *et al.*: **Community-driven computational biology with Debian Linux.** *BMC Bioinformatics.* 2010; **11**(Suppl 12): S5.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Wilson G, Aruliah DA, Brown CT, *et al.*: **Best practices for scientific computing.** *PLoS Biol.* 2014; **12**(1): e1001745.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Blischak JD, Davenport ER, Wilson G: **A Quick Introduction to Version Control with Git and GitHub.** *PLoS Comput Biol.* 2016; **12**(1): e1004668.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Theußl S, Zeileis A: **Collaborative software development using R-forge.** *R J.* 2009; **1**(1): 9–14.
Reference Source

32. Schulte E, Davison D, Dye T, *et al.*: **A multi-language computing environment for literate programming and reproducible research.** *J Stat Softw.* 2012; **46**(3): 1–24.
Reference Source

33. Karssen LC, van Duijn CM, Aulchenko YS: **Data of GenABEL Project for Statistical Genomics.** 2016.
Data Source

Open Peer Review

Current Referee Status:

Version 1

Referee Report 17 June 2016

doi:10.5256/f1000research.9397.r14125

Bjarni J. Vilhjálmsson
Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark

Karsen *et al.* presents a paper on the GenABEL project, which is really a software suite that contains 11 different packages. This is a large and impressive project that contains packages that are routinely used by researchers studying genetics. Indeed, this is reflected by more than 800 citations (according to google scholar) for the original GenABEL paper published in 2006. However, the 11 packages presented in this paper have been published previously in some shape or form, albeit (presumably) not in their most recent version. It is therefore tempting to question the novelty of the current paper that summarizes the GenABEL project and describes its user and developer community. Nevertheless, despite limited amount of novel scientific ideas or scientific results in the current manuscript, the authors have clearly put a lot of work into creating a very impressive interactive user and developer community. Furthermore, the paper is well written and it will undoubtedly be highly cited by future researchers. Lastly, to reiterate, the GenABEL is a very impressive large scale project that is heavily used by the community! I therefore think this is overall a nice publication that is certainly suitable for indexation.

Minor comments:

- Abstract: "…developed withing.." -> "…developed within.."
- I think it would be magnanimous if you acknowledged contributors as authors, e.g. under an umbrella term "The GenABEL community".
- Leave citations counts out of manuscript, it looks awkward. They will also change.
- On p. 3 "polyphenotype", I think multivariate would be a better word.

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.

Competing Interests: No competing interests were disclosed.

Referee Report 27 May 2016

doi:10.5256/f1000research.9397.r14057

Giulietta Minozzi
Department of Veterinary Science and Public Health, University of Milan, Milan, Italy

This very well written article describes the GenABEL project for statistical genomics and high lightens the great success of the project, that in the years has lead to the creation of an actual scientific community that is spread in several countries worldwide.

It is my understanding that the manuscript is aimed at reaching potential new users, but even to old users, unaware of the possible new tools included in the GenABEL project.

The methodologies underlining the different tools and their potential applications are well described. Figure 2 describes the extent of the use of the GenABEL website/suite and Table 1 gives a glance of the tools available and on the new versions implemented.

It is clear, from the texts, the effort that has taken place and is taking place aiming at bringing together top scientists, software developers and end users with the central goal of making statistical genomics work by openly developing and subsequently implementing statistical models into a user-friendly software.

I retain this article, the tools and the information provided of extreme importance to the scientific community.

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.

Competing Interests: No competing interests were disclosed.