openSNP - Crowdsourcing Genome-Wide Association Studies

Bastian Greshake[1,*], Philipp Bayer[2], Fabian Zimmer[3], Julia Reda[4]
1 Goethe University, Frankfurt am Main, Germany
2 University of Queensland, Brisbane, Australia
3 Westfälische Wilhelms Universität, Münster, Germany
4 Johannes-Gutenberg University, Mainz, Germany
∗ E-mail: info@opensnp.org

Abstract

Genome-wide association studies are used to correlate phenotypic traits with genetic variants like Single Nucleotide Polymorphisms (SNPs). Genome-wide association studies compare the genetic variation between two groups to find correlations between membership in a group and the genetic variation. To find significant correlations it is necessary to have a large enough sample size. The advent of Direct-To-Consumer genetic testing offers the analysis of SNPs for an individual at low cost without the need to consult a physician or geneticist. Over 100,000 people have already been genotyped through Direct-To-Consumer genetic testing. This data could be a valuable source for association studies. However, this data is not public for various reasons and thus cannot be used by other scientists. It seems reasonable to create a central open data-repository for such data, but it was previously unknown if and how people would submit their data to such a repository. Here we present a survey which evaluates whether people are willing to publicly share their genetic information. In the light of those results we also present the web-platform openSNP, an open database which allows participants of Direct-To-Consumer genetic testing to publish their genetic data along with phenotypic information into the public domain. Through this ”crowdsourced” effort of collecting genetic and phenotypic information we create a valuable resource which can be used in a wide array of studies, included Genome-wide association studies. OpenSNP is hosted at www.opensnp.org, and the code is released under MIT-license at github.com/gedankenstuecke/snpr.

Author Summary

Missing for now.

Introduction

The availability of new DNA sequencing techniques has shifted the focus of biological data acquisition towards new biomedical applications. Many diseases for example Alzheimers, Parkinsons or different types of cancers (needs citations) are at least partially heritable so the genome of patients can be used for diagnostic purposes. Using the genetic information of patients for diagnostics is made possible through the sharp decrease in costs for analysing genetic information [1].

If genetic information for more than one individual is known, the analysis of allele frequencies of Single Nucleotide Polymorphisms (SNPs) can be used to associate such SNPs with diseases and other inheritable traits. Genome-Wide Assocation Studies (GWAS) make use of statistics to compare the allele frequencies in patients to the alleles in healthy controls. This enables GWAS to find SNPs which are significantly overrepresented in patients and associates those SNPs with a trait or disease. This method does not allow inference of causal differences but merely identifies correlations. The first GWAS was published in 2005 and compared age-related macular degeneration in contrast to a healthy control group [2]. Since the beginning, the number of participants in such studies has been rising. To date, over 1200 GWAS have been performed [3] and over 5000 SNPs have been linked to different diseases and traits [4].

GWAS are not only performed inside the traditional scientific community. Since 2006, companies like 23andMe, deCODEme or FamilyTreeDNA have been offering Direct-To-Consumer (DTC) genetic testing. These companies use DNA microarrays to screen for around 0.5 to 1 million SNPs spread over the human genome. In return, customers receive an analysis of the results, as well as a file that includes the customer's raw individual genotypes. In 2011, 23andMe alone had over 100,000 customers [5] - the company realizes the potential to perform GWAS with this amount of data by using surveys to ask their customers about traits and diseases. With the consent of the customer the data is used for association studies. 23andMe has published several articles in which known findings are replicated together with new associations disorders like Parkinson's Disease [6, 7]. So far, over 30,000 23andme-customers have participated in 23andme's association studies, which proves that this data-source has a lot of potential for other researchers.

The generation of biomedical data by private companies raises concerns about privacy [8], liability and consent [9]. Nevertheless, in some instances individual customers are willingly sharing their data. Most do so by uploading their data to their personal website or to open software repositories like *GitHub*. This data is scattered and unorganized, making it hard to use in studies. While projects like SNPedia try to keep track of all the publicly available genotyping files [10], they usually do not provide the information necessary to perform GWAS, as the phenotypic information is often not attached to the genetic information. Projects that attach the phenotype to the genetic information, like the *Personal Genome Project*, still do not allow for an easy re-use of the data, as they lack an application programming interface (API) or other methods by which researchers could download the data. Additionally, not every customer of DTC genetic testing can participate in the *Personal Genome Project*.

As a follow up of a survey, which evaluated whether customers of DTC genetic testing would be willing to share their genotypic and phenotypic information (see Supplementary Materials), we designed openSNP, an online platform which enables DTC customers to freely share their genotypic and phenotypic data under a Creative Commons-Zero license. The phenotypic and genotypic data can be retrieved from openSNP using different methods, which also include different API standards to allow automated processing of the data. Additionally openSNP also enables users of DTC genetic testing to receive further information on their genotypes, through the crowdsourced resource SNPedia and primary publications, like those of the Public Library of Science journals.

Results

We created the openSNP project (http://opensnp.org) as an open, crowdsourced online platform for DTC customers interested in sharing their raw data and for researchers interested in performing GWAS or perform other types of analysis with the data. Customers of DTC testing are encouraged to share their genotyping results along with their phenotypic traits to enable easy access for researchers. Users of openSNP can create a personal profile, discuss SNPs and phenotypes on the platform using a simple commenting system, or send each other private messages.

People interested in using the data of openSNP can download complete dumps of the genotypic and phenotypic information or use query API endpoints utilizing JavaScript Object Notation (JSON) objects or the Distributed Annotation System (DAS) [11].

Sharing genotypic information

Currently users can upload their genotyping results from the companies *23andMe*, *deCODEme* and *FamilyTreeDNA* via a webinterface to the openSNP project. There is experimental support for uploading exomes in the VCF format [12], as *23andMe* recently started exome sequencing for its customers. So far only the SNPs of the exome data sets are visualized on openSNP, but the downloads include all variation found in the exome. The uploaded data is published under the Creative Commons Zero-license, which - in accordance with the Panton Principles [13] - allows a complete reuse of the data without any constraints. Between the start of openSNP on 09/27/2011 and 12/18/2012, 214 people have signed up with openSNP, and 79 genotyping files were made available. The openSNP database lists 69,486,471 genotypes which are distributed over 1,938,603 unique SNPs. Figure 1 depicts the increase of users and genotyping files over time.◇

BG: update all numbers

Crowdsourcing phenotypes

Users are able to create new phenotypes that are not yet listed by openSNP. The specification of these phenotypes is open and not limited to pre-defined categories. To reduce the amount of manual data curation, openSNP tries to harmonize the expression and spelling of the same phenotype or variation. We implemented an autocompletion-feature, which helps users to reuse already entered phenotypes. Users are encouraged to list as many phenotypes as possible through a simple achievement system, rewarding users that upload their data and enter phenotypic information with small badges that are shown on their profile pages.

In the same timeframe as above, all users combined have entered a total of 675 variations on 47 different phenotypes with those variations being the different values on a given trait or phenotype. See figure 1 for the increase of phenotypic information over time.

The mean number of users that have entered their variations for a single phenotype is 14.36 (SD 12.65), the median is 10. The distribution of how many users have entered their data per phenotype can be seen in figure 2. The phenotype provided by the most users is the eye color, which has been entered by 54 users. There are two phenotypes which have so far only been provided by a single user: the SAT Writing score and triglyceride-levels.◇

BG: update all numbers and graphs

Connection to external services

In order to provide users with relevant information on their respective genotypes, openSNP scans databases of the scientific literature for specific SNPs. A total number of 15,229 documents◇relevant to the SNPs listed in openSNP could be found in the publication databases of Mendeley, the Public Library of Science and in the crowdsourced SNPedia. Of the primary literature, 25 % are released in open access journals and can be accessed free of charge (Figure 3). For usability reasons, SNPs are ranked by the amount of information gathered through the external services. The external services themselves are ranked by how easily non-scientists can understand information from these sources and available this information. The SNPedia entries are given the highest impact, as those are already manually curated and summarized in plain English, followed by open access publications out of the Public Library of Science. Lowest values are given to the Mendeley results, as the publications listed there are for the most part not freely available without subscriptions or one-time payments. An entry on SNPedia is valued 2.5 times as high as a PLoS publication and 5 times as high as a Mendeley entry.

BG: number needs to ge updated

Data access

OpenSNP offers extensive access to the data uploaded by users. Anyone can download single genotyping files for specific users, get archives of multiple genotyping files grouped by phenotypic variation, or access a single download that includes all genotyping files and all phenotypic variation in a comma-separated table. The genetic data is also accessible through the Distributed Annotation System [11, 14], which offers all data for specific chromosomes and specific positions on single chromosomes. An example of how the DAS can be used can be found on openSNP where users genotypes are visualized inside a genome browser. So far all chromosomal positions are based on the human reference genome NCBI36, as this is the standard reference used by DTC providers right now.◇

BG: should this maybe include an image as well?

The data is additionally available over a JSON-API, which allows users to directly access data in the JSON-format. The methods allow users to programmatically look for the genotypes and annotations at a given SNP as well as for phenotypes for a given user and phenotypic variation for a given phenotype.

Discussion

Privacy, health implications and ethical considerations

Much of the critique on DTC genetic testing focusses on the practice of delivering medical information without consulting a physician or genetic counsellor to help patients/customers make sense of the information and to put the new knowledge to good use [15–17].

As we have found in our survey on sharing such results (see supplementary methods), many DTC customers are willing to share their results with the public to help scientific progress, without forgetting about the privacy implications that come with openly sharing genetic information. There is a variety of ethical and privacy implications when it comes to DTC genetic testing [9, 18].

Our survey has shown that people are concerned about their privacy and fear that stakeholders like employers, insurance companies, governments or advertisers might misuse the information. Policy makers start to react to those changes by introducing laws like the *Genetic Information Non-Discrimination Act* in the United States or the *Gendiagnostikgesetz* in Germany to minimize the impact of widely available genetic information. DTC genetic testing companies themselves also try to educate their customers about the risks of releasing genetic data.

OpenSNP openly addresses the problem of privacy implications that come with releasing genetic data twice, once during registration for openSNP and once during the upload of the DTC genetic testing results. Users have to confirm that they have read and understood the disclaimer about possible side-effects of publishing their data. To further improve this process we are looking forward to implement an informed consent-processes by a provider like *Consent for Research* (http://www.weconsent.us).

GWAS and Open Data

Although prices of exome or even full genome sequencing are dropping rapidly, GWAS are still considerably cheaper. However, GWAS can only detect correlations of SNPs with those traits and do not allow inference on the cause for any correlation. Furthermore, for a statistically sound analysis GWAS need a large enough sample size. Nevertheless, GWAS are still frequently used and new associations are found [19–21].

One way of bringing down costs for GWAS even further is to make use of already available genotyping results and datasets. Data produced by DTC genetic testing companies is a promising source for such data, as such companies already have high numbers of customers which are willing to pay for the genotyping by themselves.

By crowdsourcing the acquisition of genetic and phenotypic data, openSNP faces the same problems as any other open platform on the Internet, namely the need to trust users regarding the data they upload and enter on openSNP. Additionally, the quality of the data varies, especially in terms of accuracy on the phenotypic variation, with users entering data in different measurement systems. Another problem with user-entered data is the frequent switching between categorical and continuous phenotypes - for example, some users entered the specific value of their height, while other users entered their height according to a category like ”150cm to 160cm”.

While we try to suggest similar entries to the users, there are some cases where users will not follow those suggestions, so duplicates or similar phenotypes or variations in traits may arise. There are two possible solutions to this problem: The first one would be to only allow a trusted subset of users to enter new phenotypes. The other one would be to make users enter all possible variations of a phenotype while creating a new phenotype, so that later users cannot add variations that have not been available from the start.

In both cases it makes it harder for users to enter their data which raises the bar for participation. We decided to keep data entry as easy as possible, at the cost of forcing users who want to perform GWAS with the data to perform additional quality control.

Another risk regarding data quality that should be kept in mind is a possible bias in data availability on openSNP: only a subset of people buy DTC genetic testing, from which an even smaller subset is willing to publish the results, which can potentially lead to skewed GWAS-results.

With openSNP, we have built a platform that can be used by customers of DTC genetic testing to easily share their genetic and phenotypic data with a wide audience, as well as by scientists and interested citizens who are looking for datasets to freely use in their studies. Customers of DTC genetic testing also benefit from an easy access to primary literature on SNPs and genetic variations they carry. While there is not enough data uploaded to perform a statistically sound GWAS yet, this will be possible in the future, as user numbers continue to rise. By including the option of uploading exome data sets the platform already is capable of adjusting for changes in the type of data generated by DTC genetic testing.

Materials and Methods

Technical implementation of the platform

The main platform is implemented using the web framework Ruby on Rails 3.0.10. Postgres 9.2 is used as the main database backend for Rails. The database stores genotyping results, users' phenotypic information, literature results from Mendeley and the Public Library of Science as well as summaries on SNPs which can be found in SNPedia. The literature database of Mendeley is queried using the REST API, which delivers results in JSON. The literature database of the Public Library of Science is queried using the respective REST API, which delivers results in an XML-format. Summaries on SNPs are provided by SNPedia, through querying the content via the MediaWiki API. All databases are queried using the unique identifier of each SNP as the search term.

SNPs are catalogued by their unique identifier, which consists of a prefix (mostly *rs*, rarely *i*) and a unique number. This is a common format, which is employed by the NCBI dbSNP database [22] and is also widely used and easily parsed from different literature sources. Publications from the different databases as well as the users' genotypes are associated with individual SNPs by the Rs-ID. Allele and genotype frequencies are updated regularly, based on the data present in openSNP.

Processes with a longer runtime, such as parsing the genotyping results, creating archives of results which are to be mailed to users and queries to external resources are handled using the ruby gem Resque and the standalone server Redis. Search features on the platform itself are implemented using Solr and the ruby gem Sunspot. Additionally, data can be requested from openSNP using the Distributed Annotation System. The required data is stored in a PostgreSQL database. Requested data is delivered in XML-format to facilitate parsing. Additionally, users can request data in the JSON-format, using a system not specified in any standard.

OpenSNP only serves as a platform for SNPs, so methods for the delivery of nucleotide sequences as described in the DAS-standard are not implemented. Currently, two methods are implemented: *features*, which is used to deliver SNPs located on specific chromosomes or between specific nucleotide positions, based on the user's query. The second method is *sources*, which advertises all DAS-sources for all genotypes present in openSNP.

A flowchart of all services incorporated in openSNP and of all the ways users can upload or access the data is given in figure 4. The source code of openSNP is published under the MIT-license and can be downloaded at http://github.com/gedankenstuecke/snpr. The genetical and phenotypical data is licensed under Creative Commons Zero.

Acknowledgments

We thank Dr. Manuel Corpas for constructive advice in grammar, spelling and structure of this study.

References

1. Brown PO, Botstein D (1999) Exploring the new world of the genomewith DNA microarray. nature genetics supplement 21: 33–37.

2. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor h polymorphism in age-related macular degeneration. Science 308: 385-389.

3. Johnson A, O'Donnell C (2009) An open access database of genome-wide association results. BMC Medical Genetics 10: 6.

4. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences 106: 9362-9367.

5. 23andMe (2011) 23andMe 2011 State of the Database Address. The Spittoon .

6. Eriksson N, Macpherson JM, Tung JY, Hon LS, Naughton B, et al. (2010) Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet 6: e1000993.

7. Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, et al. (2011) Web-based genome-wide association study identifies two novel loci and a substantial genetic component for parkinson's disease. PLoS Genet 7: e1002141.

8. 23andMe (2012) 23andMe Privacy Statement. 23andMe Homepage .

9. Caulfield T, McGuire AL (2011) Direct-to-consumer genetic testing: Perceptions, problems and policy responses. Annual Review of Medicine 63: 1.1-1.11.

10. Cariaso M, Lennon G (2011) Snpedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Research .

11. Dowell R, Jokerst R, Day A, Eddy S, Stein L (2001) The distributed annotation system. BMC Bioinformatics 2: 7.

12. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and vcftools. Bioinformatics 27: 2156-2158.

13. Molloy JC (2011) The open knowledge foundation: Open data means better science. PLoS Biol 9: e1001195.

14. Jenkinson A, Albrecht M, Birney E, Blankenburg H, Down T, et al. (2008) Integrating biological data - the distributed annotation system. BMC Bioinformatics 9: S3.

15. Hauskeller C (2011) Direct to consumer genetic testing. Bmj 342: d2317–d2317.

16. Hogarth S, Javitt G, Melzer D (2008) The current landscape for direct-to-consumer genetic testing: legal, ethical, and policy issues. Annual review of genomics and human genetics 9: 161–82.

17. Wasson K (2009) Direct-to-consumer genomics and research ethics: should a more robust informed consent process be included? The American Journal of Bioethics 9: 56–58.

18. Joh EE (2011) Ethics watch: DNA theft: your genetic information at risk. Nature reviews Genetics 12: 3113.

19. Mei H, Chen W, Jiang F, He J, Srinivasan S, et al. (2012) Longitudinal replication studies of gwas risk snps influencing body mass index over the course of childhood and adulthood. PLoS ONE 7: e31470.

20. Xu B, Tong N, Chen SQ, Yang Y, Zhang XW, et al. (2012) Contribution of hogg1 ser326cys polymorphism to the development of prostate cancer in smokers: Meta-analysis of 2779 cases and 3484 controls. PLoS ONE 7: e30309.

21. Sebastiani P, Solovieff N, DeWan AT, Walsh KM, Puca A, et al. (2012) Genetic signatures of exceptional longevity in humans. PLoS ONE 7: e29848.

22. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbsnp: the ncbi database of genetic variation. Nucleic Acids Res 29: 308–311.

Figure Legends

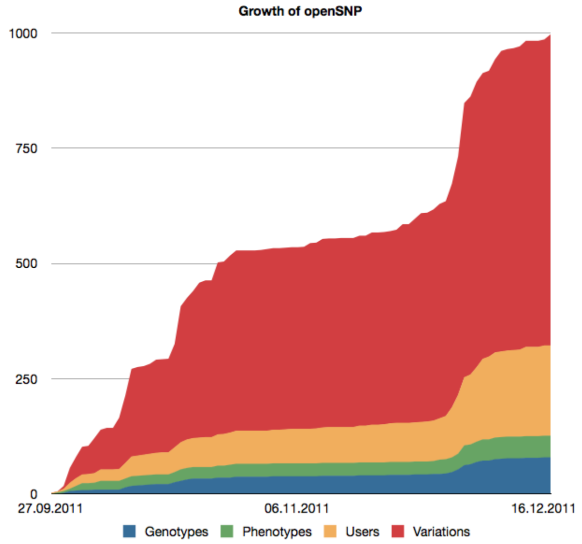

Figure 1. Growth of openSNP. The increase in numbers for users, genotyping-files, phenotypes and their variation from 27.09.2011 to 16.12.2011 is shown.

Tables

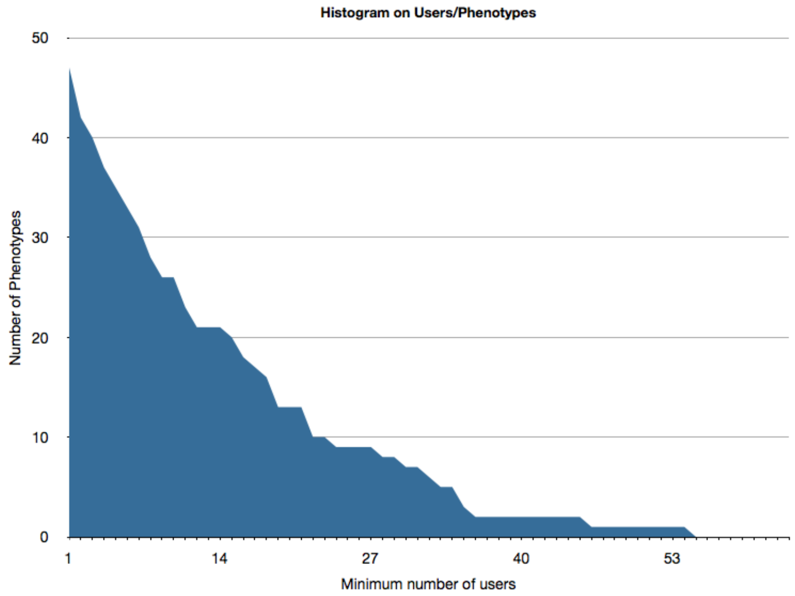

Figure 2. Histogram of users/phenotype-distribution. The x-axis shows the minimum number of users who provide information for a phenotype, the y-axis shows how many phenotypes have at least that many users.

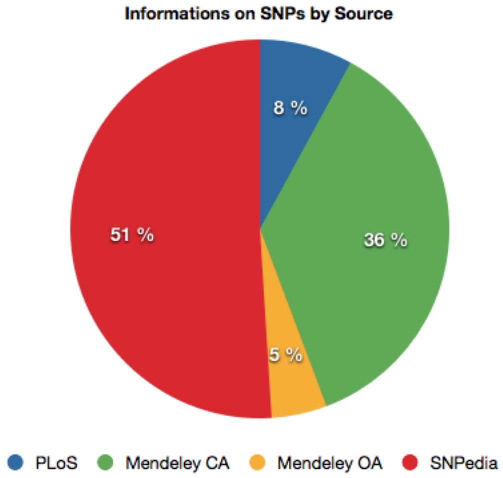

Figure 3. Distribution of external information gathered for SNPs in the openSNP-database. Data on PLoS and SNPedia is openly available for every user. Publications on Mendeley are either Open Access (OA) or Closed Access (CA).

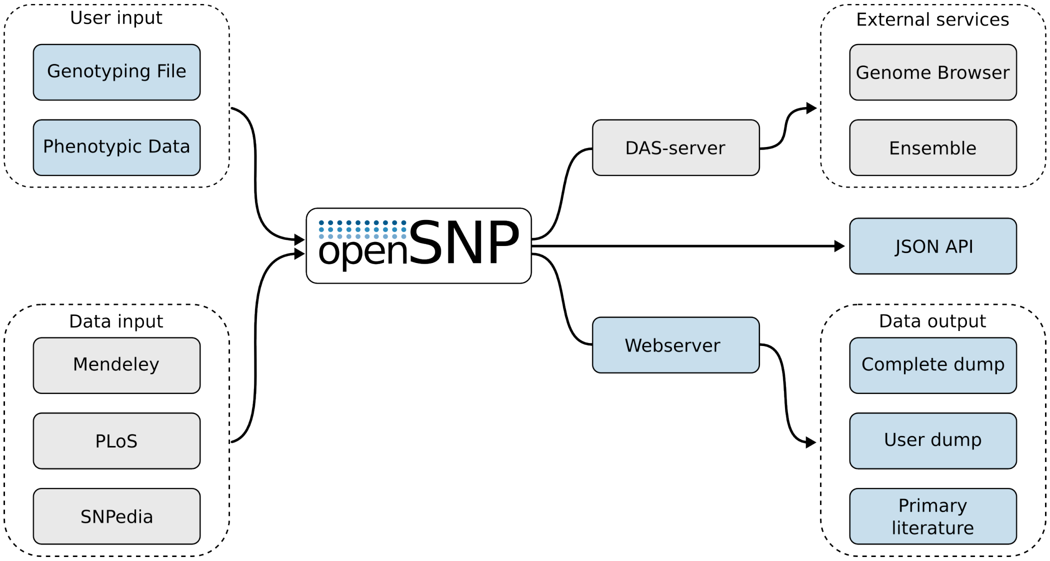

Figure 4. Flow of data inside openSNP. External databases and user-provided data are used as input. Output of data is done using the website, the *Distributed Annotation System* and a JSON-API.