openSNP - a crowdsourced web resource for personal genomics

Bastian Greshake[1,*], Philipp E. Bayer[2], Helge Rausch[3], Fabian Zimmer[4], Julia Reda[5]
1 Molecular Ecology Group, Goethe University Frankfurt am Main, Germany
2 Applied Bioinformatics Group, School of Agriculture and Food Sciences, The University of Queensland, Australia
3 Hochschule für Technik und Wirtschaft Berlin, Germany
4 Evolutionary Genomics and Transcriptomics Lab, Department of Genetics, Evolution and Environment, University College London, England
5 Johannes Gutenberg University Mayence, Germany
∗ **E-mail: info@opensnp.org**

Abstract

Genome-wide association studies are widely used to correlate phenotypic traits with genetic variants. These studies usually compare the genetic variation between two groups to single out certain Single Nucleotide Polymorphisms (SNPs) that are linked to a phenotypic variation in one of the groups. However, it is necessary to have a large enough sample size to find statistically significant correlations. Direct-To-Consumer (DTC) genetic testing can supply additional data: DTC-companies offer the analysis of a large amount 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 companies. However, this data is not public for a variety of reasons and thus cannot be used in research. 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 the web platform openSNP, an open database which allows participants of Direct-To-Consumer genetic testing to publish at no cost their genetic data along with phenotypic information. Through this crowdsourced effort of collecting genetic and phenotypic information openSNP has become a resource, which can be used in a wide area of studies, including Genome-wide association studies. openSNP is hosted at http://www.opensnp.org, and the code is released under MIT-license at http://github.com/gedankenstuecke/snpr

Author Summary

Introduction

The availability of new DNA sequencing techniques has shifted the focus of biological data acquisition towards new biomedical applications. Many diseases - for example Alzheimer's [1], Parkinson's [2] or different types of cancers [3, 4] - 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 [5].

If genetic information on 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 Association 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 [6]. Since the beginning, the number of participants in such studies has been rising. To date, over 1200 GWAS have been performed [7] and over 5000 SNPs have been linked to different diseases and traits [8].

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 [9] - 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 [10, 11]. 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 [12], liability and consent [13]. 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 [14], 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* [15], still do not allow for an easy re-use of the data, as they currently 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*.

Here, we present openSNP, an online platform which enables DTC customers to share genotypic and phenoytypic information, as well as receive additional information on their genotypes. The genotypes are made available to researchers via the open Creative Commons Zero license.

Results

Sharing genotypic information

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 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) [16].

Sharing genotypic information

Currently users can upload their genotyping results from the companies *23andMe*, *deCODEme* and *FamilyTreeDNA* via a web interface to the openSNP project. There is experimental support for uploading exomes in the VCF format [17], 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 [18] - allows a complete re-use of the data without any constraints. Between the start of openSNP on 09/27/2011 and 10/27/2012, 633 people have signed up with openSNP, and 270 genetic datasets have been made available. The openSNP database lists 215,546,685 genotypes

which are distributed over 2,140,643 unique SNPs. Figures 1 and 2 depict the increase in users and genotyping files since September 2011.

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 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 badges that are shown on their profile pages.

In the same timeframe mentioned above, all users combined have entered a total of 4743 variations on 130 different phenotypes with those variations being the different values on a given trait or phenotype. The mean number of users that have entered their variations for a single phenotype is 36.48. The distribution of how many users have entered their data per phenotype, compared to the amount of unique phenotypes, can be seen in figure 3. The phenotype provided by the most users is "eye color", for which 207 users entered their phenotype (retrieved 10/27/2012).

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 21,134 documents relevant to the SNPs listed in openSNP could be found in the publication and annotation databases of Mendeley, the Public Library of Science, in the *GET Evidence System* [15] and the *NHGRI GWAS Catalog* [8] and in the crowdsourced SNPedia (Figure 4). Of the primary literature listed on Mendeley, the *NHGRI GWAS Catalog* & the Public Library of Science, about 20 % are released in open access journals and can be accessed free of charge (Figure 5), although probably not all publications on Mendeley are correctly flagged and the *NHGRI GWAS Catalog* does not give details on whether a publication is Open Access or not. So the total number of Open Access publications might be higher.

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 how available this information is to the public. 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 and the curated databases of the *GET Evidence System* and the *NHGRI GWAS Catalog*. 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 or entries in *GET* or the *GWAS Catalog* and 5 times as high as a Mendeley entry.

Users are also able to link their Fitbit [19] accounts to their user-accounts. Fitbit is a commercial service which lets its customers track their BMI, movement and sleep data. This data can be linked to openSNP to give interested researchers an automatically maintained dataset of body and sleep developments over time.

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 [16, 20], which

offers all data for specific chromosomes and specific positions on single chromosomes. An example of how the DAS can be used is implemented 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 criticism of DTC genetic testing focuses on the practice of delivering medical information without consulting a physician or genetic counselor to help patients/customers make sense of the information and to put the new knowledge to good use [21–23].

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 [13, 24].

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. Further versions of openSNP may optionally include further consent processes.

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 [25–27].

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 results, as those 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. 21 people, mainly from underrepresented demographics, have been offered free genotyping using funding provided by the Wikimedia Germany association in order to mitigate this bias.

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. The *NHGRI GWAS Catalog* and the *GET Evidence System* provide complete dumps in plain text formats. Those are regularly downloaded and parsed. SNPs that are described as 'Insufficiently evaluated' in the *GET Evidence System* are not stored. All databases are queried or parsed 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 [28] 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 key-value storage 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: firstly *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 6. 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 and Prof. David Edwards for constructive advice in grammar, spelling and structure of this study. Further thanks go to Samantha Clark and Dan Bolser for providing valuable feedback and feature ideas during the development and Thomas Down, author of *BioDalliance*, and Rafael Jimenez, author of *MyKaryoView*, for their support on implementing the Distributed Annotation System and the genome browser. For help with their APIs we are grateful to Mike Cariaso of *SNPedia* and the PLOS & Mendeley API teams. We would especially like to thank the users of openSNP.org for their participation, their constructive criticism and bug-finding abilities and especially for sharing their genotyping and phenotype data.

References

1. Tanzi RE (2012) The genetics of alzheimer disease. Cold Spring Harbor Perspectives in Medicine 2.

2. Lill CM, Roehr JT, McQueen MB, Kavvoura FK, Bagade S, et al. (2012) Comprehensive research synopsis and systematic meta-analyses in parkinson's disease genetics: The pdgene database. PLoS Genet 8: e1002548.

3. Njiaju UO, Olopade OI (2012) Genetic determinants of breast cancer risk: A review of current literature and issues pertaining to clinical application. The Breast Journal 18: 436–442.

4. Abate-Shen C, Shen MM (2000) Molecular genetics of prostate cancer. Genes & Development 14: 2410-2434.

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

6. 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.

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

8. 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.

9. 23andMe (2011) 23andMe 2011 State of the Database Address. The Spittoon http://spittoon23andmecom/2011/06/15/23andme-2011-state-of-the-database-address/ .

10. 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.

11. 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.

12. 23andMe (2012) 23andMe Privacy Statement http://www.23andme.com/about/privacy/. 23andMe Homepage .

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

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

15. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, et al. (2012) A public resource facilitating clinical use of genomes. Proceedings of the National Academy of Sciences 109: 11920-11927.

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

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

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

19. FitbitInc (2012) Fitbit http://www.fitbit.com. Fitbit Homepage .

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

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

22. 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.

23. 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.

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

25. 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.

26. 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.

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

28. 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

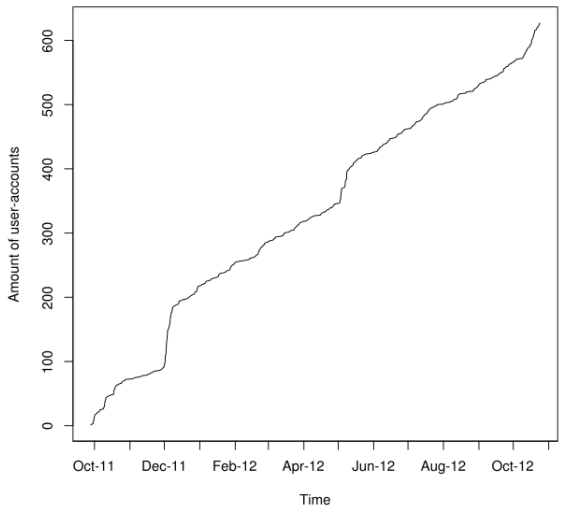

Figure 1. Growth of openSNP-user-accounts. The increase in numbers for users from 27.09.2011 to 27.10.2012 is shown.

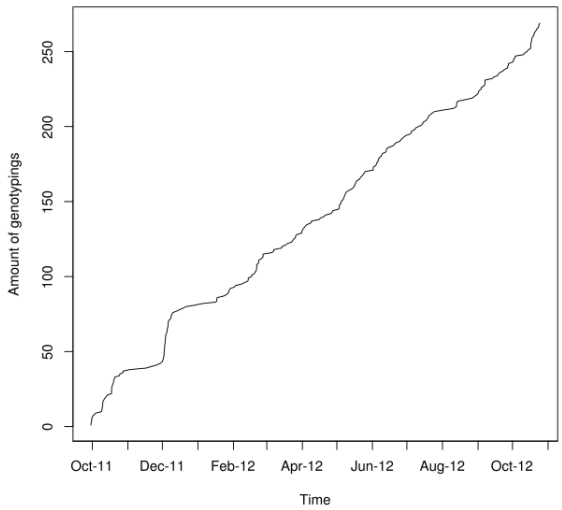

Figure 2. Growth of available genotypings. The increase in numbers for genotyping-files from 27.09.2011 to 27.10.2012 is shown.

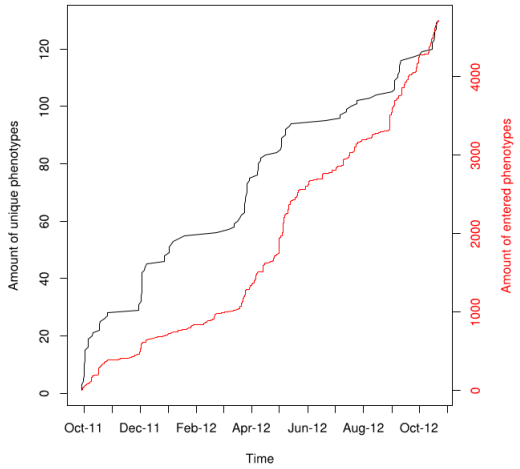

Figure 3. Development of unique phenotypes and phenotypic information over time. The x-axis shows the time-frame from start of the project until October 2012, the left y-axis shows how many unique phenotypes have been entered, and the right y-axis shows the amount of phenotypes users entered.

Tables

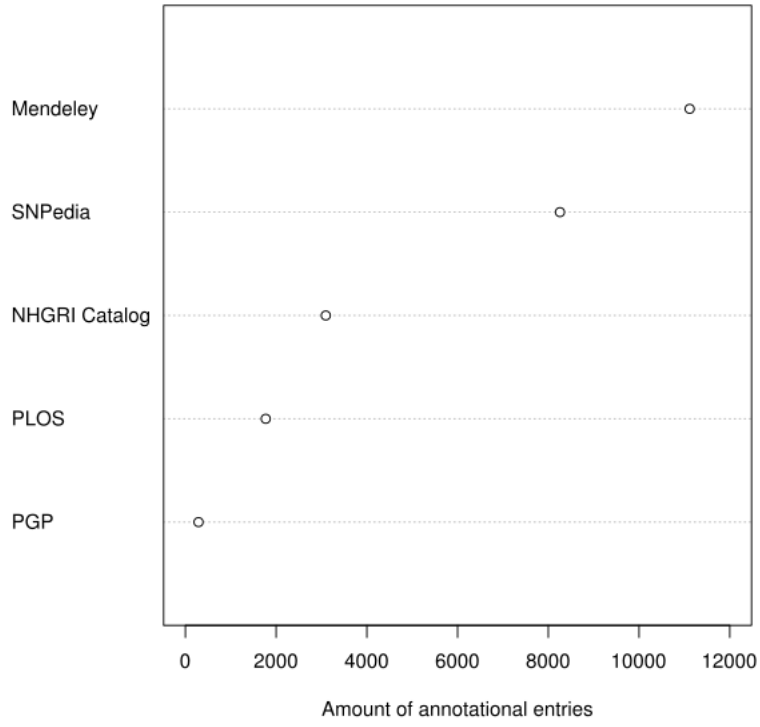

Figure 4. Distribution of annotation-sources at openSNP. Currently, SNP-annotations from SNPedia, PLOS, Mendeley, the *GET Evidence System* and the *NHGRI GWAS Catalog* are being collected.

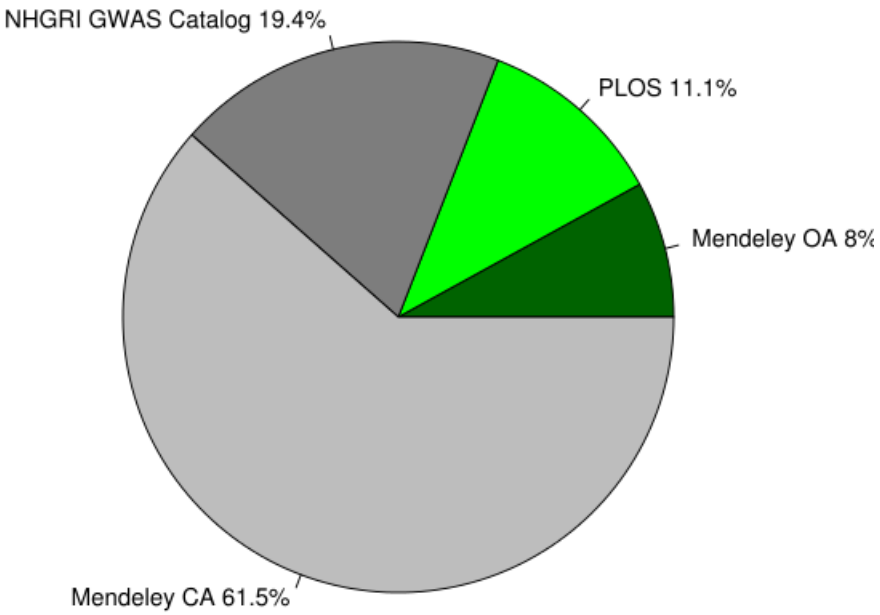

Figure 5. Ratio of Open Access Publications. Green pieces are Open Access. The *NHGRI GWAS Catalog* doesn't give information about the Open Access status.

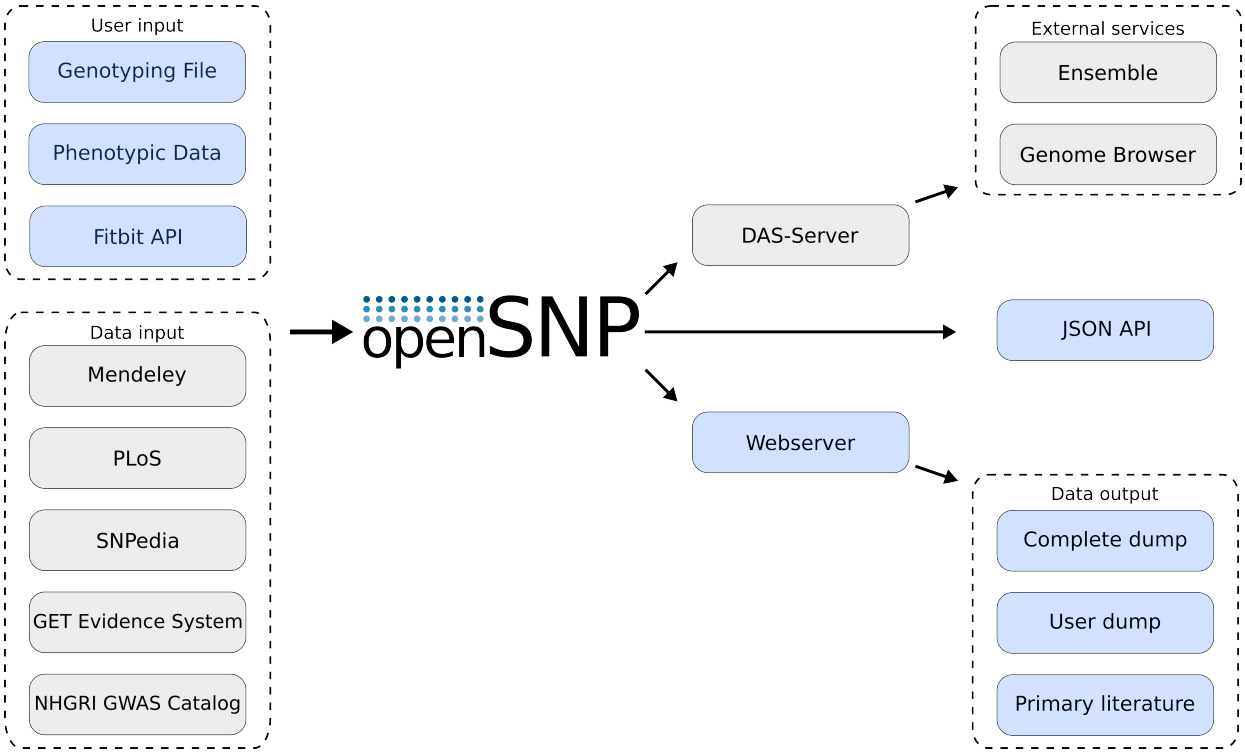

Figure 6. Flow of data inside openSNP. External databases and user-provided data are used as input. Output of data is done using the website, the *Distributed Annotation System* and a JSON-API.