Speeding up the Progress of Biomedicine by Making Data Reusable through Crowdsourcing

Bastian Greshake* and Philipp Bayer

Bastian Greshake,
Zehnthofstr. 36, 55252 Mainz-Kastel, Germany
{bgreshake,philippbay}@googlemail.com
http://opensnp.org

Abstract. Thanks to the advent of new biomedical technology it is now possible to produce large amounts of data about patients and research participants at a comparatively small cost at a high speed. This revolutionizes the way biomedical studies are performed and how physicians can diagnose genetic diseases. But the progress also leads to new bioethical issues which have to be kept in mind. Those bioethical concerns are one of the reasons why biomedical data most often isn't publicly available. We suggest how one could solve many of the bioethical problems, while at the same time creating open data resources by crowdsourcing the publication of biomedical information through patients and research participants. We also show where community efforts are already doing the first steps in this endeavor.

Keywords: personal genetics, personal genomics, genetics, open data, direct-to-consumer genetic testing, quantified self

If biomedical science remains an us and them proposition, it bodes well neither for us nor for them - Misha Angrist

1 The State of Data in Biomedical Sciences

The quantity of data which is collected for biomedical research and diagnostics has exponentionally increased over the last years. At the same time, the focus on which data to collect has shifted drastically. Traditionally, biomedical research as well as health records focussed on the history of diseases in an individual, accompanied with medical imaging and diagnostics by means of tissue samples, blood tests etc. There is a growing trend to supplement this data with a wide array of genetic information, which is likely to increase over the next years. Over ten years have passed since the first human genome was fully sequenced. While some of the claims on how this will revolutionize medicine may have been exaggerated, the availability of the human genome definitely has changed the perception and research on genetic disorders [1]. The progress in sequencing

* Coordinating Author

techniques has led to the adoption of large-scale sequencing projects to detect variations in the genomes of patients, which help to diagnose genetic disorders and to find the causes of previously unknown diseases. In 2008 clinical whole genome sequencing was applied to diagnose an 11-month old girl which showed xanthomas and very high cholesterol levels. Using this technique, a known disease with an atypical presentation in this patient could be diagnosed [2]. In 2011 the complete exome – the parts of the human genome which encode proteins – were sequenced of a 5-year old who showed a congenital disorder of glycosylation (CDG) with no mutations in the genes known to contribute to the disease. Whole exome sequencing led to the discovery of previously unknown mutations, this knowledge can now be applied to diagnostics in all patients with CDG [3].

While those are still isolated cases, these examples show how new sequencing techniques can be used in a clinical setting. This progress is made possible by the increasingly dropping price of sequencing. The price of sequencing a human genome – which cost over 3 billion USD in 2000 – has dropped to less than 10.000 USD recently and is already available through companies on the mass-market [4]. Other suppliers recently presented sequencers for less than 1.000 USD [5]. Customers can already get around 1 million so-called Single Nucleotide Polymorphisms (SNPs), of their genome tested for about 200 USD, or their complete exome sequenced for less than 1.000 USD, without the need to enroll in scientific studies or visiting their physician, through Direct-To-Consumer testing companies like *23andMe* [6]. These technical advances in the clinical and private setting catapult biomedical research into the age of "Big Data", with all the upcoming challenges of storing data in a meaningful way, making it accessible, putting it to use and making sense of it [7].

2 Why isn't Biomedical Data Open?

Although some research projects, like the annotation of the EHEC-causing *Escherichia coli*-strain, apply open data and crowdsourcing principles [8], this sharing of data unfortunately still isn't a widespread practice in biomedical research especially in studies which enlist human participants. A number of reasons can be identified, including ethical and legal constraints as well as the prevailing rather closed culture of this field [9]. On the cultural side, researchers still fear losing control over usage of the data and thus won't be able to publish more findings based on the data – especially as there are no standard mechanisms for data citation – while those publications are much-needed to secure their funding. Additionally, filing of patents is a common practice in biomedical research. Some example ethical and legal issues that come with the creation of large quantities of DNA sequence data are whether researchers should share the raw data with the participants of their studies, let alone the rest of the world, and whether they should share incidental findings about diseases or non-paternity with research participants.

The amount of data sharing with third parties in a traditional research setting is usually not only limited directly through laws but also by institutional review

boards (IRBs) or ethics committees which monitor and approve research involving humans. Advocates of strict rules on data sharing often focus on patients' and participants' privacy, as opening up the data could easily lead to abuse of the data through law enforcement agencies, insurance companies and employers. Policy-makers are working on minimizing those impacts by passing laws like the *Genetic Information Non-Discrimination Act* (GINA) in the United States or the *Gendiagnosikgesetz* (GenDG) in Germany. With those policy changes in mind, others argue on the side of the benefits for researchers of sharing data with third parties, noting that participants are open to share de-identified data [10].

A similar case can be made for sharing data with participants themselves: While some people argue that the data in the hands of individuals could do more harm than good (as they lack the necessary training in genetics to put the data to use), others make the argument that participants should have access and own their genetic information, or that ownership of this data should run in the family, as family members share most of their genetic information [11].

Sharing incidental findings with research participants is another ethics conflict in human genetics research right now: Opponents of sharing those incidental findings over the course of a study argue that the research settings are often not as rigorous in handling the samples. Thus, samples could be accidentally swapped and the wrong person receives the diagnosis. Additionally, findings made in a bleeding edge research setting have a higher chance of being preliminary, thus creating false positives which are to be falsified by more research [12]. In a nutshell, sharing incidental findings with the research participants might cause unnecessary harm. Proponents of sharing incidental findings see a moral responsibility to share those findings with participants, as they can have a profound impact on participants' lives [13]. One example is the case of a woman who participated in research on the deadly Ogden-syndrome. As she was pregnant, scientists knew due to her genome sequencing that she had a 50% chance of passing the syndrome on to her unborn son. She passed the disease mutations to her son, and the researchers did not inform either of the inheritance [14].

The practice of Direct-To-Consumer genetic testing is also critized in terms of bioethics. Some criticize that giving people the results of genetic tests without appropriate genetic counseling could do more harm than good. Participants would need counselors to analyze the accuracy of those results and make sense of them, to put them in perspective and to recommend appropriate actions after doing the test [15]. Since these concerns arose, some studies have looked into the matter and found out that most customers of such tests are able to adapt to the information provided and many show no sustained behavioral change as a result of taking DTC genetic testing [16].

3 Drawbacks of Closed Data & Benefits of Open Data

All those concerns about bioethics and privacy are important points, which have to be kept in mind in any approach that tries to open up data in biomedical

sciences. The number of complete genome sequences is rising as is the number of people who have undergone DTC genetic testing (*23andMe* already reported 100.000 customers in mid 2011 [17]). Aside from the few reference genomes, much of this data will be kept behind lock and key on the hard drives of researchers and physicians, as well as the genotyping data of companies like *23andMe*. We feel that one should keep in mind that the lack of access comes at a cost: Reusing this data isn't easily if at all possible. Recognizing disease-causing mutations often needs data of multiple individuals and especially for rare conditions it is hard to get enough participants enrolled in studies.

In extreme cases, this leads to an odd ratio of researchers to participants in the corresponding study: A study published in 2011 with about 180 authors on the genetic foundation of being underweight and eating disorders had a sample size of 138 participants. The reason for this strange researcher/participant ratio lies in the low percentage of carriers of the mutations in question [18]. Another example are the exome sequencing results of the 5-year old with CDG. In this case, the raw data could be put to use by other researchers working on the same disease [3]. Opening up those resources would enable more people to participate in the research and would allow them to add their own data to integrate data sets in order to get better results. Similarly, biomedical data which has been created in a clinical setting for diagnostics could be put to broader use if there was a way to share genetic and more traditional biomedical information, such as results of blood tests.

The same is true for the results of DTC genetic testing. SNP data as produced by DTC companies gets frequently used in Genome Wide Association Studies, a method used for finding genetic associations where large sample size is crucial in order to find significant associations between different traits and genetic variations, due to the multiple testing one performs [19]. These tests could be drastically improved by opening up already collected genetic and biomedical information to incrementally increase sample size.

A problem resulting from bioethics concerns and the closed access policies applied up to now is that patients or research participants and researchers nowadays are largely alienated from each other. For the most part, participants in research are reduced to a completely passive role. Former genetic counselor Misha Angrist critizes this lack of exchange between reseacher and researchee as being against research's own interest [20], as it fuels the lack of public support for science that leads to cuts in funding. Opening up data to participants (and possibly a wider audience) together with increasing the exchange between those groups helps to counter these effects.

4 Solutions and Examples for Open Data

An elegant way of promoting open data in the biomedical community, while keeping the ethical problems in mind, is to crowdsource the data collection to the people who have to deal with the consequences of making the data available: The participants and patients themselves. This approach stops treating them as

passive participants, but instead engages them as central agents in promoting scientific progress. We believe that individuals who enroll as participants in studies should get access to the raw data DNA sequences, results of lab diagnostics and tissue samples etc. which has been created through their participation. Participants and patients should not only have their privacy protected, but should also be allowed to make an informed decision on whether they want to receive and/or share their data with third parties. There is already a small number of projects which embrace this style of open data for biomedical applications.

One of this projects is *PatientsLikeMe*, which works mainly with information found in traditional health records. Patients can sign up to the website and share with the community details about their diseases, the symptoms they have and how well different drugs work for them. Over 130,000 patients have signed up so far. This not only allows patients to empower themselves by learning from each other, but the data can also be used for research purposes. *PatientsLikeMe* has published several articles on research done with their data so far. For example, *PatientsLikeMe* did a study on the off-label use (using drugs for purposes for which they have not been officially approved yet) of Modafinil [21]. They also published a study on the effects of taking lithium carbonate in order to slow down the progress of amyotrophic lateral sclerosis (ALS): After an independent study of 16 treated patients and 28 controls found that lithium carbonate could slow down ALS, *PatientsLikeMe* did a study with 149 treated patients and 447 matched controls that could not replicate the effect [22]. As these studies are not double-blind randomized control trials they can't substitute the traditional research. However, this type of study can be a quick and cheap additional tool to evaluate drugs. A similar approach is pursued by *Genomera* [23], a recent start-up which works on small-scale studies where not only the data collection is crowdsourced to its users, but also the concepts of the different trials are collaboratively designed. Additionally, *Genomera* allows utilization of DTC testing results in studies.

There are also projects working more extensively with genetic information. One of these projects is the *Personal Genome Project* (PGP) [24], which tries to enroll 100.000 participants who are willing to share their medical records and further phenotypic information along with their complete genomes, which will be sequenced through the PGP.

We are working on a similar project called *openSNP* [25] which tries to use genetic data. The project enables customers of DTC genetic testing to upload the raw results, which they received from their DTC provider along with self-reported phenotypic information about traits and diseases. Instead of relying on predefined categories for diseases, clinical lab results etc. with *openSNP* we crowdsource these categories as well, in order to make it easy for users (patients/DTC customers as well as researchers) to add new phenotypes. The goal of this approach is to collect enough genetic information, together with self-reported phenotypic data, to enable Genome Wide Association Studies (GWAS) at basically no cost: The users themselves paid for the genotypings and are providing information about their diseases on their own. A similar approach is em-

ployed by *23andMe*. Their customers can opt in to become part of their research and take surveys on different phenotypic traits. *23andMe* performs Association Studies with the genetic and phenotypic information and already published new findings on Parkinson's Disease using this approach [26].

5 Challenges of crowdsourcing data collection to open up biomedical data

The results of *PatientsLikeMe* and *23andMe* show that it is possible to perform studies using self-reported and crowdsourced data. Still some challenges remain in making this approach feasible for more scientists. First of all, there is the question of **how open the current approaches really are**. While the data of the PGP and *openSNP* is released as Creative Commons Zero to enable a complete reuse of the data, this isn't the case right now for *Genomera*, *23andMe* and *PatientsLikeMe*. While *23andMe* offers customers a download of their own genetic information, there is currently no way of making the data and survey answers open for everybody on their platform. *PatientsLikeMe* is a for-profit company which sells access to its data to paying researchers, along with bonus features like creating one's own, IRB-approved surveys. *Genomera* has not made any statements on their business model yet.

Another point that needs work in order to make crowdsourcing-based models feasible is **the process of informed consent**. Opening up data needs a sound way of making sure that users who opt to share their biomedical and genetical information with a broader audience understand the risks attached to doing so. This is equally important for scientists who want to make sure they only use data of participants who gave consent in this kind of research. To solve this problem, the PGP offers its own IRB-approved consent form, but traditionally, there is no portable standard for informed consent. A project that currently tries to create a portable consent process is *Consent To Research* [27] in collaboration with *Sage Bionetworks* [28].

To facilitate those changes, the research community should help in creating repositories where patients and participants in research can upload their data using open licenses. Similarly, the community needs to address the problem of unportable ethics approvals. Additionally, policy-makers need to address the challenges of making biomedical data available by protecting patients and researchees who opt to make data available.

Last but not least, there has to be a change in scientific culture in Biomedicine: away from the traditional secrecy and passivity and towards a more open attitude which embraces a relationship with research participants. We are confident that this change in culture will come over the next years, not out of pure altruism, but because every researcher will benefit from open data in the long run. On the side of patients, the first few groups working on this have already emerged, like the *Society for Participatory Medicine* [29]. The first steps to transform biomedicine from the closed "us and them" to a more open "us" are done, let's step further on this road.

References

1. Collins, F.: Has the revolution arrived? Nature 464, 674–675 (2010)
2. Rios, J., Stein, E., Shendure, J., Hobbs, H.H. and Cohen, J.C.: Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. Hum. Mol. Genet. 19 (22), 4313–4318 (2010)
3. Jones, M.A., Ng, B.G., Bhide, S. et al.: *DDOST* Mutations Identified by Whole-Exome Sequencing Are Implicated in Congenital Disorders of Glycosylation. Am. J. Hum. Genet. 90(2), 363-368 (2012)
4. Illumina everygenome, `http://www.everygenome.com/`
5. Oxford Nanopore Technologies `http://www.nanoporetech.com/news/press-releases/view/39`
6. 23andMe Exome 80x `https://www.23andme.com/exome/`
7. Howe, D., Costanzo, M., Fey, P. et al.: Big data: The future of biocuration. Nature 455, 47–50 (2008)
8. *E. coli* O104:H4 Genome Analysis Crowdsourcing `https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki`
9. Nelson, B.: Data sharing: Empty archives. Nature 461, 160–163 (2009)
10. Brown Trinidad, S., Fullerton, S.M., Bares, J.M. et al.: Genomic research and wide data sharing: Views of prospective participants. GENET MED 12, 486495 (2010)
11. Lucassen, A.: Should families own genetic information? Yes. BMJ 2007;335:22 (2007)
12. Middelton, A.: Ethics and Genomic Research: Genomethics. Genomes Unzipped `http://www.genomesunzipped.org/2012/01/genomethics.php`
13. Miller, F.G., Mello, M.M., Joffe, S.: Incidental Findings in Human Subjects Research: What Do Investigators Owe Research Participants? J Law Med Ethics 36(2), 271-279 (2008)
14. Lyon, G.J.: Personalized medicine: Bring clinical standards to human-genetics research. Nature 482, 300301 (2012)
15. Norgard, K.: DTC Genetic Testing for Diabetes, Breast Cancer, Heart Disease and Paternity. Nature Education 1(1) `http://www.nature.com/scitable/topicpage/dtc-genetic-testing-for-diabetes-breast-cancer-698`
16. Caulfield, T., McGuire, A.L.: Direct-to-Consumer Genetic Testing: Perceptions, Problems, and Policy Responses. Annu. Rev. Med. 63, 23–33 (2012)
17. 23andMe 2011 State of the Database Address `http://spittoon.23andme.com/2011/06/15/23andme-2011-state-of-the-database-address/`
18. Jacquemont, S., Reymond, A., Zufferey, F. et al.: Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478, 97-102 (2011)
19. Hindorff, L.A., Sethupathy, P., Junkins, H.A. et al.: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. PNAS 106(23), 9362-9367 (2009)
20. Angrist, M.: Open Season. soapbox science `http://blogs.nature.com/soapboxscience/2012/02/08/open-season`
21. Frost, J., Okun, S., Vaughan, T. et al.: Patient-reported Outcomes as a Source of Evidence in Off-Label Prescribing: Analysis of Data From PatientsLikeMe. J. Med. Internet. Res. 2011;13(1):e6
22. Wicks, P., Vaughan, T.E., Massagli, M.P. et al.: Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. Nat. Biotechnol. 29, 411-414 (2011)

23. genomera: heal the world http://genomera.com/
24. Personal Genome Project http://www.personalgenomes.org/
25. openSNP: Crowdsourcing Genome-Wide Association Studies http://opensnp.org
26. Do, C.B., Tung, J.Y., Dorfman, E. et al.: Web-Based Genome-Wide Association Study Identifies Two Novel Loci and a Substantial Genetic Component for Parkinson's Disease. PLoS Genet 7(6): e1002141 (2011)
27. Consent For Research http://weconsent.us/
28. Sage Bionetworks http://sagebase.org/
29. Society for Participatory Medicine http://participatorymedicine.org/