bioRxiv preprint doi: https://doi.org/10.1101/127241; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A Collection of 2,280 Public Domain (CC0) Curated Genotypes

Richard J. Shaw, Manuel Corpas*

Repositive Ltd, Future Business Centre, Cambridge, UK

*Contact email: manuel@repositive.io

Abstract

Cheap sequencing has driven the proliferation of big human genome data aggregation consortiums, providing extensive reference datasets for genome research. These datasets, however, may come with restrictive terms of use, conditioned by the consent frameworks with which individuals donate their data. Having an aggregated genome dataset with unrestricted use analogous to public domain licensing is therefore unusually rare. Yet public domain data is tremendously useful because it allows freedom to perform research with it. This comes with the price of donors surrendering their privacy and accepting the associated risks derived from publishing personal data. Using the Repositive platform (https://repositive.io/?23andMe), an indexing service for human genome datasets, we aggregated all deposited files in public data sources under a CC0 license from 23andMe, a leading Direct-to-Consumer genetic testing service. After downloading 3,137 genotypes, we filtered out those that were incomplete, corrupt or duplicated, ending up with a dataset of 2,280 curated files, each one corresponding to a unique individual. Although the size of this dataset is modest compared to current major genome data aggregation projects, its full access and licensing terms, which allows free reuse without attribution, make it a useful reference pool for validation purposes and control experiments.

Background & Summary

The availability of personal genome Direct-to-Consumer (DTC) tests has been fuelled by companies like 23andMe, which makes it easy for users to access their personal genotype data for an affordable price. Despite recent improvements in cost and availability of Next Generation Sequencing (NGS) technology for DTC use, array genotyping is still a viable option [1], particularly when users want to have a whole genome screen of known Single Nucleotide Variations (SNVs) for a cheap price. For about $100 one can buy a DTC test and have a satisfying state-of-the-art analysis of ancestry and genetic health risk [2]. 23andMe has been able to capitalise on the curiosity of many customers from around the world, generating one of the most extensive private genome data ecosystems in the world [3]. The genotype data aggregated from 23andMe customers, combined with phenotype information from questionnaires, has already been proven to be an effective way to discover new markers, e.g. [4–6].

bioRxiv preprint doi: https://doi.org/10.1101/127241; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Using the genotype data from 23andMe as a gateway for personal/recreational genomics has also other advantages. The format of the different SNP array versions (although the number of tested SNV may vary) remains constant and relatively easy to handle: the size of the genotype files is in the order of tens of MB, which are more manageable than the prohibitively heavy GB sizes of Next Generation Sequencing (NGS) files. 23andMe genotype files are also a useful proxy for understanding the individual's main genetic features.

In this study we use the Repositive platform [7] (https://repositive.io/?23andMe), which indexes human genomic datasets from all major and known genome data repositories, to gather the greatest possible open access 23andMe set of individual genotypes. We only took 23andMe genotypes that have been deposited in public archives and have no restriction of use, i.e., CC0 license (Table 1). These archives include (in order of greatest contribution): openSNP [8], the Personal Genome Project [9], Open Humans [10], Genomes Unzipped [11], the Corpasome [12] and Stephen Keating's data source [13]. Any genotype from these data sources can be copied, modified, distributed and used, even for commercial purposes, all without having to ask permission.

23andMe Data Source	License	Terms of Use
openSNP	CC0	https://www.iubenda.com/privacy-policy/641811
PGP	CC0	http://www.personalgenomes.org/tos
Open Humans	CC0	https://www.openhumans.org/community-guidelines/#public-data
Genomes Unzipped	CC0	http://genomesunzipped.org/data
Corpasome GRCh37	CC0	https://figshare.com/articles/23andMe_hg37/4491215
Stephen Keating	CC0	http://stevenkeating.info/main.html

Table 1: Summary of licenses and their provenance for all data used for this study. All of them are in the public domain (CC0) [14]. This means that the data referenced in this study can be used, distributed or modified, even for commercial purposes, without needing to ask permission to the individuals from whom this data originated.

We downloaded every open access public domain 23andMe genotype available in every public repository currently indexed by Repositive, curating and bundling them into a single dataset. This amounted to 3,137 genotype files that, after curation, were reduced to 2,280 genotypes.

These 2,280 genotypes constitute an open access dataset of high utility in the personal genomics field. Although a number of resources are available that provide genome-based SNV data at a population level (e.g., 1000K genomes [15], ExAC [16], GnomAD [17], UK10K [18], PGP [9]), to there is no a single repository that meets all the stated characteristics our proposed aggregation dataset provides: full genotype

bioRxiv preprint doi: https://doi.org/10.1101/127241; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

data, open access, unrestricted use, derived from the same source at the scale of thousands. Hence we believe this dataset will prove a valuable resource for the genome research community, both academic and industry, who will be able to benefit from a curated cohort of genotypes.

Methods

In order to aggregate all open access 23andMe genotypes in the public domain we used the Repositive platform [7] (https://repositive.io/?23andMe). Repositive is an indexing service that catalogues all known human genome datasets in public sources and repositories throughout the internet. Repositive stores both the metadata which describes the deposited dataset, and the link through which the dataset is accessed. A total of 3,137 links from the Repositive platform matched a potential 23andMe entry (complete list of links available in Suppl. File 1). These links pointed to 6 sources: 2,318 from openSNP, 514 from the Personal Genome Project (PGP), 286 from Open Humans, 13 from Genomes Unzipped, 5 from the Corpasome [19] and 1 from Stephen Keating's source [20]. We then downloaded all the files and proceeded with their curation:

1. Among the 3,137 downloaded files we found 37 that were non-text files: e.g., PDFs or image files.
2. For the remaining 3,100 files we discarded 210 files (168 VCFs and 42 of unknown format)
3. This left us a total of 2,890 files:
    a. 2,484 for build GRCh37
    b. 378 for build 36
    c. 28 for unknown build
4. From the 2,484 that mapped to the GRCh37 build, we discarded 87 that had less than 500k SNP rows, giving us a total of 2,397 23andMe text files. The Corpasome 23andMe files (corresponding to 5 individuals from the Corpas family) were mapped to build 36. Since we had direct access to their 23andMe account, we downloaded them directly from 23andMe, this time mapped to build 37. This produced a total of 2,402 GRCh37 23andMe text files.
5. We further refined the 2,402 GRCh37 text files, extracting the consensus SNP positions from all autosomes (chromosome 1-22). This yielded a total of 445,734 SNPs. Using the R packages gdsfmt and SNPRelate[21], the 445,734 SNPs were further reduced to 63,486 by linkage pruning.
6. We then created a Principal Component Analysis (PCA) to cluster the 23andMe data from collected individuals into populations. The resulting PCA showed 141 very clear duplications.
7. Note also that the 141 duplications do not necessarily imply 2402 - 141 unique samples. Since some samples were multiply duplicated, there were actually 2280 unique genotype (Corpasome is included).

Code availability

A python script that takes the 23andMe URLs and downloads them is available.

bioRxiv preprint doi: https://doi.org/10.1101/127241; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data Records

For version 1 and 2 of their SNP chip 23andMe used a customised Illumina Hap550+ array. Version 3 is based on a customised Illumina OmniExpress+ array and version 4 uses a custom Illumina HumanOmniExpress-24 format chip. The downloaded raw data file is a zipped text file 5 MB to 30 MB in size in bed format [22], with a header describing the date in which the file was downloaded and the human reference assembly build used as coordinates for the mapping of the SNVs tested. The file itself provides the SNPdb ID for each SNV [23], the chromosome and position in the mapped assembly and the genotype, including both alleles. Figure 2 shows a snippet of the 23andMe genotype bed file from one of us downloaded on “Tue Dec 13 09:12:17 2016”.

```
# This data file generated by 23andMe at: Tue Dec 13 09:12:17 2016
#
# This file contains raw genotype data, including data that is not used in 23andMe reports.
# This data has undergone a general quality review however only a subset of markers have been
# individually validated for accuracy. As such, this data is suitable only for research,
# educational, and informational use and not for medical or other use.
#
# Below is a text version of your data.  Fields are TAB-separated
# Each line corresponds to a single SNP.  For each SNP, we provide its identifier
# (an rsid or an internal id), its location on the reference human genome, and the
# genotype call oriented with respect to the plus strand on the human reference sequence.
# We are using reference human assembly build 37 (also known as Annotation Release 104).
# Note that it is possible that data downloaded at different times may be different due to ongoing
# improvements in our ability to call genotypes. More information about these changes can be found at:
# https://www.23andme.com/you/download/revisions/
#
# More information on reference human assembly build 37 (aka Annotation Release 104):
# http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606
#
# rsid  chromosome      position        genotype
rs3094315       1       752566  AA
rs12562034      1       768448  GG
rs3934834       1       1005806 CT
rs9442372       1       1018704 AA
rs3737728       1       1021415 AG
rs11260588      1       1021658 GG
rs6687776       1       1030565 CC
rs9651273       1       1031540 AA
rs4970405       1       1048955 AA
rs12726255      1       1049950 AA
rs11807848      1       1061166 CC
rs9442373       1       1062638 CC
rs2298217       1       1064979 CC
rs12145826      1       1066029 AG
rs9442380       1       1087683 CC
rs7553429       1       1090557 AA
```

Figure 1: A screenshot of the top of a 23andMe raw data genotype showing the date in which it was downloaded, the assembly (human assembly build 37) and for each SNV analysed its SNV ID (RSID), its location (chromosome and position) as well as the genotype for both alleles in that position.

Table 2 summarises the number of links for each data source, the number of text files downloaded and the number of 23andMe files mapped to the build 37 with more than 500k SNPs each.

bioRxiv preprint doi: https://doi.org/10.1101/127241; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Source	Total # Links	Downloaded Text Files	GRCh37 23andMe Text Files with >500k SNP
openSNP	2,318	2,296	1,947 (81%)
PGP	514	499	316 (13%)
Open Humans	286	286	133 (5.5%)
Genomes Unzipped	13	13	0
Corpasome	5	5	5 (0.2%)
Stephen Keating	1	1	1 (0.04%)
Total	3,137	3,100	2,402

Table 2: Summary of the total number of links obtained through the Repositive platform and their sources. The data source from which we get the greatest number of links for 23andMe data files is openSNP. A process of curation was carried out to select only 23andMe files that map to build GRCh37 and have >500k SNPs.

Data Availability

The 2,280 curated set of links is available. Using the python script from above the data can be downloaded.

Technical Validation

One common problem when aggregating genome data from different data sources is duplication. In this case, users wishing to share their genotype data may have submitted their 23andMe file to more than one open access repository.

We created a Principal Component Analysis (PCA) to cluster the 23andMe data from the 2,402 collected 23andMe individuals with more than 500k SNPs mapped to GRCh37 (Table 2**)**. The resulting PCA clusters are shown in Figure 2. From the PCA, we were able to identify 141 very clear duplications. There were some further cases of very similar (but not identical) values but there were no obvious cases of siblings or other relatives from the metadata associated to these 23andMe files. We performed a second PCA including both our curated 23andMe and 1000 Genomes Project individuals (1000G) (data not shown). The 1000G individuals were classified as Asian, European and African clusters and they overlapping the rectangles shown in Figure 2. The Asian cluster (Figure 2, left) has 58 individuals, 46 of which are sourced from OpenSNP. The African cluster (Figure 2, right) included 50 individuals, 43 from OpenSNP and the European cluster (Figure 2, middle) had 2098 individuals. OpenSNP allows users to self-report their ethnicity. To further validate our ethnicity clusters we found that of the 46 OpenSNP individuals in the Asian cluster 5 had

bioRxiv preprint doi: https://doi.org/10.1101/127241; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

reported ethinicity: 4 East Asian and 1 Korean. Surnames for these users were also suggestive of Asian ethnicity in 13 other cases where we found the individuals' names. We also looked at the self-reported ethnicity of the OpenSNP individuals in the African cluster, 11 in total. 10 of them reported African/Black and 1 reported many ethnicities, including African.

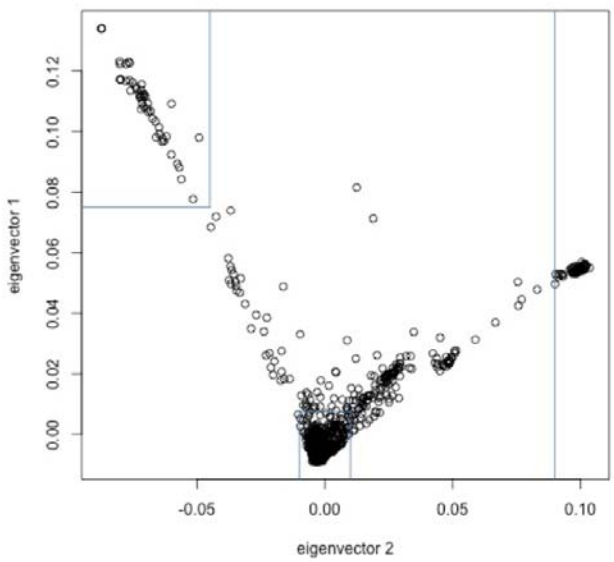

Figure 2: Principal Component Analysis (PCA) of 2,402 23andMe data files from individuals who have openly shared their genotypes in public repositories. We find that by looking at their PCA values, there are 141 that are duplicated. The individuals contained within the top left rectangle overlap with 1000 genomes individuals for Asian ethnicity, in the bottom centre European and right side African.

Acknowledgements
We are grateful to the public repositories that make it possible for users to upload their genomic datasets at no cost.

Author contributions
MC conceived the study and wrote the paper.
RS performed the curation and validation of the data.

Competing interests
We have read the journal's policy and we have the following conflict: At the time of writing RS and MC are employees of Repositive Ltd.

bioRxiv preprint doi: https://doi.org/10.1101/127241; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Shahandeh A, Johnstone DM, Atkins JR, Sontag J-M, Heidari M, Daneshi N, et al. Advantages of Array-Based Technologies for Pre-Emptive Pharmacogenomics Testing. Microarrays (Basel). 2016;5. doi:10.3390/microarrays5020012

2. Corpas M, Valdivia-Granda W, Torres N, Greshake B, Coletta A, Knaus A, et al. Crowdsourced direct-to-consumer genomic analysis of a family quartet. BMC Genomics. 2015;16: 910.

3. AnneW. Power of One Million [Internet]. [cited 22 Dec 2016]. Available: https://blog.23andme.com/news/one-in-a-million/

4. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016;48: 709–717.

5. Hinds DA, Barnholt KE, Mesa RA, Kiefer AK, Do CB, Eriksson N, et al. Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. Blood. 2016;128: 1121–1128.

6. Chahal HS, Lin Y, Ransohoff KJ, Hinds DA, Wu W, Dai H-J, et al. Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. Nat Commun. 2016;7: 12048.

7. Kovalevskaya NV, Whicher C, Richardson TD, Smith C, Grajciarova J, Cardama X, et al. DNAdigest and Repositive: Connecting the World of Genomic Data. PLoS Biol. 2016;14: e1002418.

8. Greshake B, Bayer PE, Rausch H, Reda J. openSNP--a crowdsourced web resource for personal genomics. PLoS One. 2014;9: e89204.

9. Church GM. The personal genome project. Mol Syst Biol. 2005;1: 2005.0030.

10. Home - Open Humans [Internet]. [cited 19 Oct 2016]. Available: https://www.openhumans.org/

11. Author G, MacArthur D, Wright C, Pickrell J. Genomes Unzipped [Internet]. [cited 19 Oct 2016]. Available: http://genomesunzipped.org/

12. Corpas M. Crowdsourcing the corpasome. Source Code Biol Med. 2013;8: 13.

13. Steven Keating's Homepage [Internet]. [cited 22 Dec 2016]. Available: http://stevenkeating.info/main.html

14. Creative Commons — CC0 1.0 Universal [Internet]. [cited 22 Dec 2016]. Available: https://creativecommons.org/publicdomain/zero/1.0/

15. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74.

bioRxiv preprint doi: https://doi.org/10.1101/127241; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

16. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536: 285–291.

17. gnomAD browser [Internet]. [cited 14 Dec 2016]. Available: http://gnomad.broadinstitute.org/about

18. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526: 82–90.

19. Manuel C. 23andMe hg37. 2016; doi:10.6084/m9.figshare.4491215.v1

20. Kovalevskaya N. DNAdigest interviews Steven Keating: scientist and patient - DNAdigest.org. In: DNAdigest.org [Internet]. 4 Mar 2016 [cited 19 Oct 2016]. Available: http://dnadigest.org/dnadigest-interviews-steven-keating-scientist-patient/

21. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28: 3326–3328.

22. UCSC Genome Bioinformatics: FAQ [Internet]. [cited 14 Dec 2016]. Available: https://genome.ucsc.edu/FAQ/FAQformat#format1

23. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29: 308–311.