A Data Journey Through Dataset-Centric Population Genomics

James Griesemer

Abstract I describe a data journey drawn from a case study of research in human population genomics. The case is framed in dialogue with a project on what has been called the "re-situation" of scientific knowledge (Morgan 2014). The kind of journey described elicits a missing concept—"data*set*-centric" biology—in the conversation around the emergence of "big data" and data-centric biology (Leonelli 2016) and its contrast, "traditional" or "small data" biology. I distinguish data*point*-centric from data*set*-centric practices. The case study is about the development, use, and amendment of data sets in one lab's pursuit of human genome diversity studies. I offer a model of data journeys to interpret the case. The model is comprised of three kinds of components: scientific data structures, data representations, and data journey narratives. The case study illustrates two visualizations that frame the data-set journey.

1 Traveling Findings and Data Journeys in Human Population Genomics

In this chapter, I make a case for a "middle ground" landscape of data *set*-centric biology as an important setting for data journeys in twenty-first century science, adding "middle sized" facts to the big and the small (Howlett and Morgan 2011, Leonelli 2016). Communities of specialists in fields practicing dataset-centric biology are organized around data *sets* rather than dissociable, individually retrievable data *points*, even though the dissociability of the latter is key to the data journeys of dataset-centric biology. For dataset-centric biology, if datapoints are disaggregated from their context in a dataset, datapoints may lose value or meaning as datasets add value and change meaning. Scientific focus on datasets prods dataset-centric sciences down toward a "craft" scale of operation rather than up to an "industrial" scale: in dataset-centric biology, datapoints are not interchangeable parts, nor independently valuable "widgets" in a datapoint-as-product economy of science. At

J. Griesemer (✉)
Department of Philosophy, University of California, Davis, Davis, CA, USA
e-mail: jrgriesemer@ucdavis.edu

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_8

craft scale, datapoints are more like individualized parts of whole dataset products and less like anonymized members of possibly arbitrary or merely conventional sets.

In a broad sense, the data journeys in human population genomics of interest in this chapter begin with tissue specimen collection, proceed to extraction of DNA from specimens, and eventually result in sequencing, production of digital sequence records, and archiving of the records. My focus here, however, is on the journey *after* digital data is produced: how these records are collected into datasets that can travel (or not), just as Leonelli (2016) has documented how genomics datapoints can travel. These journeys must of course be planned, including developing protocols for subject sampling and specimen collection, but here I focus on journeys of datapoints and datasets derived from DNA already extracted and archived. After tissue collection and curation, extracted DNA specimens are allowed to circulate in a limited fashion to qualified research labs. The labs then conduct or arrange sequencing so as to use the digital data in a range of biomedical and ancestry studies. Once the data gets into digital form, the datasets can have a life of their own. This "workflow" can be summarized by distinguishing: (1) a "field" setting in which a study design is put into action to produce "data," (2) a lab setting in which specimens or data are put in motion to produce findings and reports, and (3) a community setting in which findings are put into circulation in various social worlds that become evaluated as "facts" or sent back into scientific workflows to be reworked, reinterpreted, reevaluated (Fig. 1). My case study focuses on the latter: the use of genomic DNA data to infer ancestry relations among human populations.

The case is part of a project on what has been called the "re-situation" of scientific knowledge (Morgan 2014). The kind of journey described elicits a missing concept—"data*set*-centric" biology—in the conversation around the emergence of "big data" and data-centric biology (Leonelli 2016) and its contrast, "traditional" or "small data" biology. I distinguish data*point*-centric from data*set*-centric practices. The case study is about the development, use, and amendment of datasets in one lab's pursuit of human genome diversity studies.

The data journey I re-trace here begins with sequence data analyzed in a paper by Noah Rosenberg et al. (2002) in *Science* magazine: "Genetic Structure of Human Populations." This paper reports "big findings," that is, findings about worldwide ancestry relationships derived from analysis of a substantial collection of datapoints in a dataset using advanced analytical methods and theoretical models. The paper also reports (or refers to) "small findings," e.g. findings of particular sequences detected in particular DNA samples. Some of the small findings are presented simply by citation of the datasets used in the analysis leading to the big findings, based on sequencing cell line panel DNA collected for the Human Genome Diversity Project (HGDP).

Data for the HGDP that supplied the Rosenberg lab came from 1064 lymphoblastoid cell lines (LCLs) cultured from blood samples collected from people of different localities or regions around the world by a variety of laboratories interested in participating in the shared effort (Cann et al. 2002). These collection efforts were heterogeneous. Specimens were eventually deposited and archived at the Center for the Study of Human Polymorphism (CEPH), in Paris, which provides samples of

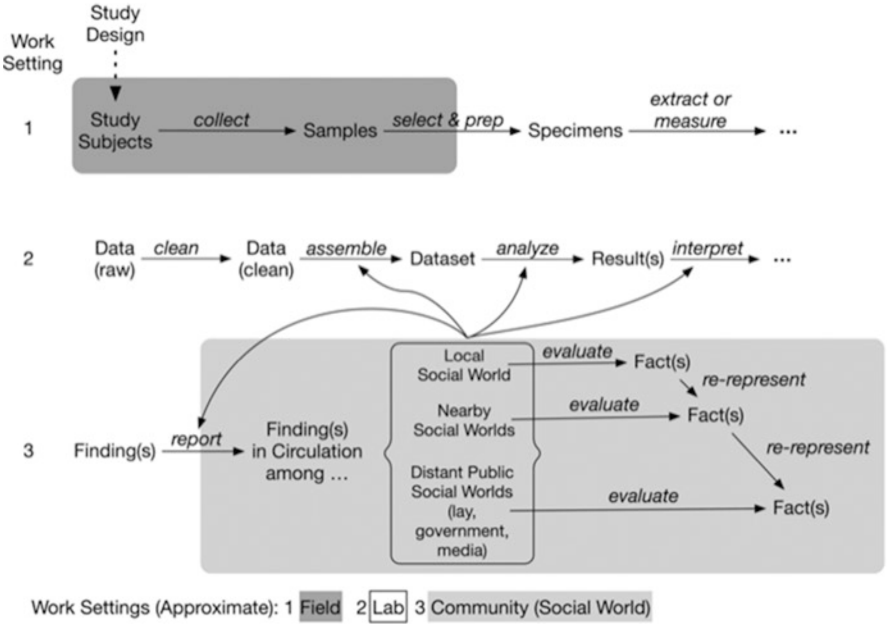


Fig. 1 Diagram illustrating the kind of work flow from a study design, to field work (*stage 1*) producing specimens or raw data, which is then assembled into datasets, analyzed and interpreted as yielding findings in the lab (stage 2), that are then circulated via talks, publications and online media in various social worlds (stage 3) that evaluate findings, elevating some of them to the status of facts and returning others for reconsideration, reinterpretation, and reevaluation. Many points in such processes feed into future study designs or the modifications of ongoing studies

extracted DNA to qualified researchers. These users of HGDP-CEPH specimens then generated data by sequencing the DNA (or in some cases RNA) or by arranging for third parties to do the sequencing.[1] Attention to data in the HGDP, like data in the Human Genome Project (HGP) more broadly, reflects emerging sensibilities of data-centric biology. DNA sequences—"digital" data derived from DNA samples—are the main form of data used to reconstruct ancestry in population genomics. Over the course of the 1990s and 2000s, this data—the data *points*—became increasingly archived in online databases of the kinds Leonelli (2016) describes.

That said, the kind of data *journey* of the sequence data in the HGDP data *sets*, is quite different, in mode of travel, in the organization and standardization of data practices, and in the institutionalization of the data packaging practices that govern the work. It is a data *set* journey—of datapoints between datasets and datasets within and among projects—as much or more than a journey of datapoints into and out of a centralized database.

[1] E.g. the Mammalian Genotyping Service of the Marshfield Clinic Research Institute (Marshfield Clinic Research Institute 2014).

One of the big findings reported concerns relationships between clusters of similar genetic sequence markers and continent-scale geographic distribution of humans. The finding is big enough to be reported in the abstract of the paper (Rosenberg et al. 2002, 2381):

> … without using prior information about the origins of individuals, we identified six main genetic clusters, five of which correspond to major geographic regions, and subclusters that often correspond to individual populations.

The 2002 publication was a landmark and its findings, methods, and conceptual presuppositions widely debated (Horton 2003). The model-based clustering algorithm implemented in analytical software the authors and their collaborators built, program STRUCTURE (Pritchard et al. 2000), assumes a pre-defined number of clusters and then allocates datapoints to clusters based on patterns of genetic similarity. The methodology is to allocate sample individuals to clusters by similarity across a collection of loci—sequences that are either shared or not shared between individual samples. The particular clusters to which individuals are assigned emerge in the clustering process and can then be compared to the "pre-defined" population labels from which the samples came.[2] The "big fact" of continental geographic patterns of human ancestral groups in circulation since the eighteenth century was affirmed by Rosenberg et al. (2002) in a novel way: based on genotype sequence distributions without reference to the pre-defined population labels. The paper is easy to read, contra the authors' intentions, as endorsing a presupposed biological concept of race by conflating a geographic interpretation of genetic classification with race on the grounds that the pre-defined populations (either from the sampling design or in the analysis) somehow biased the results.[3] The analysis is subtle and interpretation tricky.

The analysis leading to the big finding was also contested for its theoretical presuppositions (e.g. by Serre and Pääbo 2004) and defended (e.g. Rosenberg et al. 2005). Some challenges to the results questioned the sampling methodology that produced the HGDP-CEPH cell lines. Others challenged that the analysis was flawed mainly due to theoretical presuppositions regarding whether human genetic variation can be assumed to be organized in more or less discrete "clusters," perhaps with some admixture, or rather in more or less continuous "clines," perhaps with some clumping and isolation. There has been discussion of the analytical methodology as well, including examination of the models and algorithms used by STRUCTURE, alternative cluster algorithms, and alternative multivariate statistical approaches (see Sect. 4 below).

[2] Part of the methodological controversy about this research concerns the sampling methods used to collect samples in the first place and part with whether and how DNA donors "self-identify" with population labels assigned as "meta-data" to the DNA sequence data. Our larger project will address the latter topic in detail (Griesemer and Barragán 2019).

[3] See Wills 2017 for an analysis of "rhetorical appropriations" of the article; see Wade 2014 for a journalist's reading of the paper as supporting a concept of race as "clusters of variation."

It is not my purpose to characterize how well this big "fact" of continental differences (variously as a story of race, ethnicity, or genetic variation) has traveled through the centuries or spread among disciplines or societies, nor to assess the critical charges by post-colonialist thinkers, even while I fully agree that issues of race and ethnicity are far more important in the grand schemes of human cultures and societies than is reconstruction of the data journeys of the datapoints, their uptake in datasets, or interpretations of narratives of facts related to the journeys of the constructed datasets. Nevertheless, my interest *here* is to understand scientific practices involved in using the kinds of data that fuel the work of producing big findings, rather than the findings themselves.

2 Scientific Data Structures

In contrast to the big findings—the stuff of "results" and "discussion" sections of published scientific papers—key small findings mentioned or referred to in Rosenberg et al. (2002) concern the genotypes at the particular loci of the particular sample subjects used to assemble the genome diversity dataset for the analysis. These small findings are, in effect, "asserted" by reference, via the computer files in which the data are represented and recorded, to "scientific data structures." These data structures are displayed in the files and described in "materials and methods" sections, figures, tables, information supplementary to main publications, and software manuals. The data structures and files link sample subject identifiers to sequence data, e.g. diversitydata.stru, which is described in another file, diversityreadme.txt.[4]

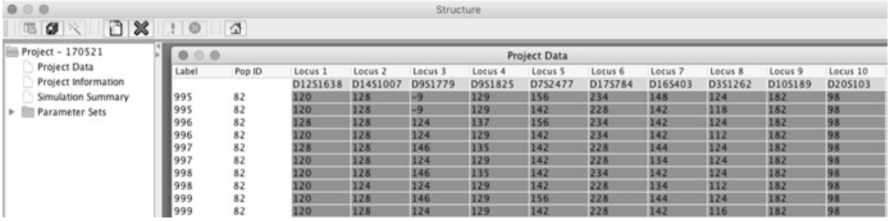
The representation of genotypes in the diversitydata.stru file is clear but indirect, involving pointers (labels) to sequence data records stored in centralized databases such as GenBank. GenBank labels for DNA sequences appear as names of loci in the data file.[5]

Genotypes for each sample individual are coded in labels for the two alleles at each locus represented in the file: 377 loci in this dataset × 2 alleles for each diploid sample individual, with two rows in the data table for each sample subject, one row for each of the paired chromosomes. The allele at the first sequenced locus for sample individual 995, for example, from "Karitiana Brazil AMERICA," (Pop ID 82) is an allele coded as "120" (Fig. 2).

Allele encodings report "genotypes (measured in base pairs)" (Rosenberg et al. 2002), that is, by integer labels: "Each allele at a given locus should be coded by a unique integer" (Pritchard et al. 2010, p. 6). "120" encodes a unique allele at locus

[4] Rosenberg maintains downloadable copies of the exact data used in the original paper at the Rosenberg Lab website (Rosenberg Lab 2018).

[5] Another downloadable file, diversityreadme.txt, contains "meta-data" information about how diversitydata.stru is organized. The reference to "the structure program" is to the software, called "STRUCTURE," authored by some of the authors of Rosenberg et al. (2002).



Label	Pop ID	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5	Locus 6	Locus 7	Locus 8	Locus 9	Locus 10
		D12S1638	D14S1007	D9S1779	D9S1825	D7S2477	D17S784	D16S403	D3S1262	D10S189	D20S103
995	82	120	128	-9	129	156	234	148	124	182	98
995	82	120	128	-9	129	142	228	142	118	182	98
996	82	128	128	124	137	156	234	142	124	182	98
996	82	120	128	124	129	142	234	142	112	182	98
997	82	128	128	146	135	142	228	144	124	182	98
997	82	120	128	124	129	142	228	134	124	182	98
998	82	120	128	146	135	142	234	142	124	182	98
998	82	120	124	124	129	142	228	134	112	182	98
999	82	120	128	146	129	156	228	144	124	182	98
999	82	120	128	124	129	142	228	142	116	182	98

Fig. 2 Screen shot of records in a dataset visualization in program STRUCTURE, after I cleaned (pruned) out meta-data from the file downloaded from the Rosenberg Lab's dataset web page, so the software could read the data file. STRUCTURE is a free software package described by Prichard et al. (2000) and downloadable at http://web.stanford.edu/group/pritchardlab/structure.html. The dataset used by Rosenberg et al. (2002) is downloadable from the Rosenberg Lab "Data sets" webpage: https://rosenberglab.stanford.edu/datasets.html

D12S1638. Sample subject 995 happens to have the same allele, "120," on both chromosomes and is thus homozygous for that locus.

A different data file, diversityloci.txt, associates GenBank sequence identifiers such as D12S1638 with Marshfield Screening Set labels (AFMB002VD5) linking the sequence to the tissue sample from which it was sequenced. This link represents and visualizes an early part of an "omics"-like datapoint journey from samples to sequences in the workflow of population genomicists. In turn, the GenBank identifier points to a record in NCBI's Nucleotide Database, "a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB," (NCBI 2019). reflecting a datapoint journey from a HGP reference sequence contributed to this centralized, online-accessible database. The GenBank sequence label D12S1638 is itself a reference to an actual sequence of 233 nucleotides, reported at the NCBI web site.[6]

These several files, maintained at the Rosenberg lab website as "the data" (and meta-data), correspond to a simple relational data structure that points in one direction to the tissue sample sources in the Marshfield Screening set of CEPH-curated cell lines and points in the other direction to the DNA sequences generated from those cell lines that are eventually encoded in datasets in the Rosenberg lab (and potentially uploadable to GenBank's NCBI Nucleotide Sequence Database).

For the text-based cluster analysis methods implemented in program STRUCTURE, which are used to analyze the dataset in Rosenberg et al. (2002), and for the project of studying allele polymorphisms in these sequences, all that matters is that the text used to label the sequences, e.g. "120," be unique.[7] Whether the

[6] The complete reference sequence for locus D12S1638 can be retrieved from a NCBI Nucleotide Sequence Database Fasta search report https://www.ncbi.nlm.nih.gov/nuccore/Z53031.1?report=fasta. Accessed 5 June, 2018.

[7] The mathematical method at the heart of the software's algorithm, latent Dirichlet allocation, is also used for topic modeling in digital humanities (see Blei and Lafferty 2009). There are journeys of models and software within and among fields to be tracked alongside the data journeys described here.

software actually compares sequence “data” or rather encodings of genotypic differences in text labels for these similarities or differences is irrelevant to the form of analysis and findings presented in the publication, although quite relevant to how we might interpret their datapoint journeys and what other uses or “re-situations” might be made of the datapoints and datasets.

The software, program STRUCTURE, is also downloadable from the laboratory of Jonathan Pritchard, one of its authors, now at Stanford University (Pritchard Lab 2019). The downloadability of the data *set*, which visualizes a scientific data structure, and analytical software from a *local* but accessible website, i.e. a lab web site rather than a community- or government-maintained online database, is a feature of the kind of dataset-centric practice I suggest is now widespread in contemporary biology. This dataset archiving practice occupies a middle ground between the non- or poorly-circulating datasets of hypothesis-centric traditional practices and the highly accessible datapoints archived in centralized databases of the datapoint-centric sciences. It is notable that while web links for this kind of local hosting of datasets and software tend to break as researchers move from one research organization (typically, a university) to another, links to the datasets, software, and references do mostly get reestablished and are relatively speaking “findable” (by internet search) if not by archiving in stable, centrally located internet resources of a federal government (e.g., NCBI, CEPH) or major NGO (e.g., Coriell, Marshfield, Simons).

3 Dataset Journey Representations: Two Visualizations

Datapoint and dataset structure representations for the Rosenberg et al. (2002) paper were already introduced in Fig. 2. What I am *not* talking about is the widely noted and discussed figures in Rosenberg et al. (2002, Figures 1 and 2) and other publications using program STRUCTURE (and in its early versions, the separate visualization software, DISTRUCT). These are visualizations of the *output* of the dataset analysis which are interpreted to produce “big findings.”

The description of this dataset in the supplemental information to the paper already narrates a dataset journey by relating the dataset constructed and analyzed for the publication from its source material in DNA extracted from one of the Marshfield screening sets of tissue samples used as sources of DNA. I describe that narrative in the next section. Here, I describe two data visualizations that are central to dataset journey narratives.

Figure 2 displayed a fragment of the Rosenberg et al. (2002) dataset in the form it takes when the dataset file is opened with the Apple MacOS graphical interface implementation of program STRUCTURE, version 2.3.4, after I did some “cleaning” or “pruning” of the “raw” data file. There was a data journey even from the “raw-raw” data—that is, the downloadable data file as archived on the Rosenberg lab's dataset web page.[8] The "raw-raw" data file contains redundant "meta-data," i.e. data that is not used by program STRUCTURE for data analysis, but which makes the data file more human-readable without following cross-references to other data files, as described above. This meta-data about "pre-defined" populations embedded in the dataset is also used to interpret what genotype similarity clusters *mean* so as to formulate big findings.

Indeed, this meta-data added to the data file is redundant because it is also linked by a data field in each data record to the "population code," e.g. "82" standing for "Karitiana Brazil AMERICA," which also appears in a separate "meta-data" file called diversitycodes.txt.[9] This meta-data must be removed from the data file in order for STRUCTURE to read it.

So far, I have considered datapoint and dataset representations in data tables (stored in computer data files). I turn now to visualized representations of datapoint and dataset *journeys*. These journey visualizations are not narratives themselves, i.e. stories of the travels of points and sets through and to various sets, publications and research projects. Rather, visualizations of scientific data structure representations can *facilitate* data journeys as "chronicles" promoting certain sorts of dataset "travel narratives" in a research community. These visualizations "chart the territory" or "map the waters" in which dataset "ships" can travel from research project to research project.

Thus far, I have mentioned the journeys of samples to specimens to datapoints in dataset assembly, visualized by the kinds of data files discussed above. Next, I describe two kinds of visualizations of data *set* journeys linking different datasets into sequences or chronologies.

3.1 Example: Lab Web Page Dataset Journey Visualization

Rosenberg's lab "diversity" web page links to a "Data sets" web page with a link titled: "HGDP-CEPH human genome diversity cell line panel" (Rosenberg Lab 2018). The main "Data sets" page shows that the Rosenberg lab maintains data sets mostly on humans, but includes non-humans (chickens) and also links to datasets "hosted by collaborating labs."[10]

This diversity web page provides links to many of the maintained datasets for human data. It *also* visualizes a kind of data journey itself. The web page does this as a structured framework of boxes/panels—a vertical, textual "triptych"—in the

[8] I discovered the raw data file was not in a format program STRUCTURE could process directly by trial and error, as have many other naïve users. For evidence, see the Google Groups FAQ: https://groups.google.com/forum/#!forum/structure-software. Accessed 13 August, 2019.

[9] Additional figures can be viewed in an expanded version of this chapter at: http://philsci-archive.pitt.edu. For diversitycodes.txt see Rosenberg Lab (2018).

[10] Chicken breeds with known population structure are used to test "the utility of genetic cluster analysis in ascertaining population structure," see Rosenberg et al. 2001.

web page. Each panel includes a descriptive title, summary dataset description, references to sources, and links to downloadable dataset files. The panels start with the HGDP 2002 dataset from Rosenberg et al. (2002) at the bottom of the page (reading up to the top of the page to follow the journey chronologically) or start with the most recently archived dataset of exome data from 2013 (reading down the page from top to bottom to retrace the lineage of current work back to source datasets). The triptych is headed (at the top) by a summary of the "lineage" of datasets from 2013 back to 2002: "[2013] [2011] [2009] [2008] [2006] [2005] [2002]."

Each panel title indicates the character of the dataset as a modification from HGDP 2002, e.g. "HGDP+other 2013 microsatellites", indicating that 645 autosomal microsatellite loci were added to the original 377 of the HGDP 2002 study in the study published by Pemberton et al. (2013). The web page overall visualizes the journey of the HGDP 2002 datapoints in the 2002 dataset in summary form as each new dataset (or version) is assembled from previous ones, sometimes noting variation from other, related or similar datasets referenced in the literature.[11]

3.2 Example: Excel Spreadsheet Dataset Journey Visualization

In 2006, Rosenberg published a paper attempting to frame the story of a dataset journey in terms of a different kind of visualization than the vertical triptych in his Lab's "Data sets" webpage. Interestingly, because this was also a project concerning the HGDP 2002 dataset, the 2006 project also appears as a place in the dataset journey in that triptych visualization, titled "HGDP 2006 relatives" (Rosenberg Lab 2018).

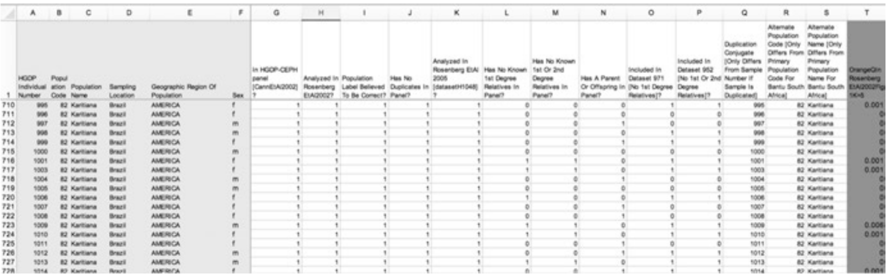
Rosenberg (2006) seeks to put some order into the proliferation of datasets serving human population genomics ancestry reconstructions by offering a naming convention for datasets and an assessment of which of the datasets that his lab assembled are appropriate for what kinds of work, based on their characteristics *as* datasets.

Rosenberg's dataset visualization is in the form of an Excel Spreadsheet (Fig. 3) that offers a different kind of triptych than the one previously discussed.

The spreadsheet lists individual HGDP sample donors by sample number (e.g., sample donor 995 discussed above). The population codes and "meta-data" of population names, sample locations (usually nation-states) and large scale regions follow. Meta-data information on the sex of the donor is also included. Then, a series of columns are used to indicate whether each donor's sample (in the form of DNA sequence datapoints) is included in datasets that figured in the research projects marked by publications cited in the column headings.

Wherever a "1" appears in the rows of these columns, the individual's DNA sequence data is included among the records of the dataset used in that column's publication. By scanning across the columns from left to right, one can see when a

[11] See additional figures in the expanded version of this chapter at: http://philsci-archive.pitt.edu



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	HGDP Individual Number	Population Code	Population Name	Sampling Location	Geographic Region Of Population	Sex	In HGDP-CEPH panel [CannEtAl2002] ?	Analyzed In Rosenberg EtAl2002?	Population Label Believed To Be Correct?	Has No Duplicates In Panel?	Analyzed In Rosenberg EtAl 2005 [datasetH1048] ?	Has No Known 1st Degree Relatives In Panel?	Has No Known 1st Or 2nd Degree Relatives In Panel?	Has A Parent Or Offspring In Panel?	Included In Dataset 971 [No 1st Degree Relatives]?	Included In Dataset 952 [No 1st Or 2nd Degree Relatives]?	Duplication Conjugate [Only Differs From Sample Number If Sample Is Duplicated]	Alternate Population Code [Only Differs From Primary Population Code For Bantu South Africa]	Alternate Population Name [Only Differs From Primary Population Name For Bantu South Africa]	OrangeQin Rosenberg EtAl2002Fig 1K=5
710	995	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	0	0	0	1	1	995	82	Karitiana	0.001
711	996	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	0	0	0	0	0	996	82	Karitiana	0
712	997	82	Karitiana	Brazil	AMERICA	m	1	1	1	1	1	0	0	1	0	0	997	82	Karitiana	0
713	998	82	Karitiana	Brazil	AMERICA	m	1	1	1	1	1	0	0	0	1	1	998	82	Karitiana	0
714	999	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	0	0	1	1	1	999	82	Karitiana	0
715	1000	82	Karitiana	Brazil	AMERICA	m	1	1	1	1	1	0	0	0	0	0	1000	82	Karitiana	0
716	1001	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	1	1	0	1	1	1001	82	Karitiana	0.001
717	1003	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	1	1	0	1	1	1003	82	Karitiana	0.001
718	1004	82	Karitiana	Brazil	AMERICA	m	1	1	1	1	1	0	0	1	0	0	1004	82	Karitiana	0
719	1005	82	Karitiana	Brazil	AMERICA	m	1	1	1	1	1	0	0	0	0	0	1005	82	Karitiana	0
720	1006	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	1	0	0	1	1	1006	82	Karitiana	0
721	1007	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	0	0	1	0	0	1007	82	Karitiana	0
722	1008	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	0	0	1	0	0	1008	82	Karitiana	0
723	1009	82	Karitiana	Brazil	AMERICA	m	1	1	1	1	1	1	1	0	1	1	1009	82	Karitiana	0.006
724	1010	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	1	1	0	1	1	1010	82	Karitiana	0.001
725	1011	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	0	0	1	0	0	1011	82	Karitiana	0
726	1012	82	Karitiana	Brazil	AMERICA	m	1	1	1	1	1	0	0	1	1	1	1012	82	Karitiana	0
727	1013	82	Karitiana	Brazil	AMERICA	m	1	1	1	1	1	1	1	0	1	1	1013	82	Karitiana	0
728	1014	82	Karitiana	Brazil	AMERICA	f	1	1	1	1	1	0	0	1	1	1	1014	82	Karitiana	0.001

Fig. 3 Screen shot of a fragment of the Rosenberg (2006) spreadsheet "SampleInformation.xls". The figure displays a "triptych" or rather 10-ptych (columns G-P) of points of embarkment/disembarkment of datapoints originating in the HGDP-CEPH LCL cell line panel and ending in dataset H952, which has dropped all data (and records) that include close (1st or 2nd degree) relatives. The spreadsheet is downloadable from Rosenberg Lab (2018). It is not included as supplemental information to the published paper. https://rosenberglab.stanford.edu/data/rosenberg2006ahg/SampleInformation.xls. Accessed 26 August 2019

particular datapoint embarked or disembarked the research program (sequence of research projects) in the Rosenberg Lab. The stops along the journey are from the HGDP-CEPH sample set, to the dataset analyzed in Rosenberg et al. (2002) to the dataset analyzed in Rosenberg et al. (2005), to the dataset called H971 to the dataset called H952.

4 Data Journey Narratives: Datapoints and Datasets

A data journey narrative appears in a particular research publication to tell the story of the dataset that arrived at the research project reported in the publication and is analyzed *there*. Such narratives have the form of stories about "how the dataset got to its destination," after a perhaps circuitous route through other research projects, labs, programs, or specialties.

Dataset journey narratives support a form of narrative explanation (Currie 2018). However, because they are narratives of *dataset* journeys, the target of explanation is not some phenomenon in nature, but rather an explanation of the use of a particular dataset in a particular research project.

The aim is to explain how and why a particular dataset "arrived" at this particular destination, given a particular research project. Dataset journey narratives are needed to persuade an audience to accept the dataset as appropriate for data analysis and thus to accept the results as findings worthy of circulation.

4.1 Dataset Assembly Narrative

Rosenberg et al. (2002) describe a dataset derived from 1056 individuals from 52 "pre-defined" populations, sequenced at 377 autosomal microsatellite loci. The 1056 individual DNA samples are a different set than the samples delivered to the lab from CEPH because not all of those samples could be used for Rosenberg et al.'s purposes. As they write (Rosenberg et al. 2002 supplemental, 1):

> The data set that we analyzed differs from the HGDP-CEPH Human Genome Diversity Cell Line Panel of 1064 individuals in its inclusion of Japanese individual #1026, whose cell line could not be produced owing to technical problems, and its exclusions of She #1331, who was not genotyped, and 8 individuals whose populations had samples of size 1 or 2 (#993, #994, #1028, #1030, #1031, #1033, #1034, #1035). Individual #1410, who is not included in the Cell Line Panel, was genotyped, but as the only representative of his population, was not analyzed. The loci studied, from Marshfield Screening Set #10 (http://research.marshfieldclinic.org/ genetics/sets/combo.html), include a mixture of 377 polymorphic di-, tri-, and tetra-nucleotide repeat loci spread across all 22 autosomes (2, 19), with 3.8% missing data. Genotyping was performed by the Mammalian Genotyping Service (19).

This kind of attention to precisely what dataset is being assembled for a particular investigation is central to the kind of data journey of interest here. Consideration is given to why individual datapoints may or may not embark on the journey. The goal is to use as much of the HGDP-CEPH world-wide sample tissue collection as possible to reflect as much of the world-wide genetic diversity sampled and to provide the most robust inferences of ancestry relations possible, given the available data and background knowledge at the time.

Datasets assembled for specific projects seek to answer questions or test hypotheses. In the case of Rosenberg et al. (2002), the question is whether STRUCTURE can reveal population diversity through study of genetic diversity data without appeal to "self-identified" population membership of sample donors. The datapoints and dataset are described, their provenance and relations to previously assembled datasets are also described, and the reasoning behind the beginnings and endings of journeys *of particular datapoints* (or *specimens*, in the early stages of these data journeys) is given.

The reasons the data journey takes particular twists and turns are a mix of kinds, starting from the usual kinds of "cleaning" of "raw" data familiar from other contexts and discussed above. "Japanese individual #1026" was included in the Rosenberg study even though the extracted DNA was not derived from the CEPH cell line diversity panel due to technical problems with the CEPH cell line. Other tissue samples were not sequenced and hence could not supply data. Samples that were included in the Rosenberg study collectively have 3.8% missing data, i.e. sequences missing for particular loci within the 377 loci sequenced for each individual. Missing data reduces the resolution and precision of the analysis, but not so much that the whole data record for those individuals must be excluded from the analysis. Some data, in other words, fails to be generated from specimens while other data is dropped when the records in which they are coded are eliminated from consideration for various reasons. These are typical kinds of "missing data."

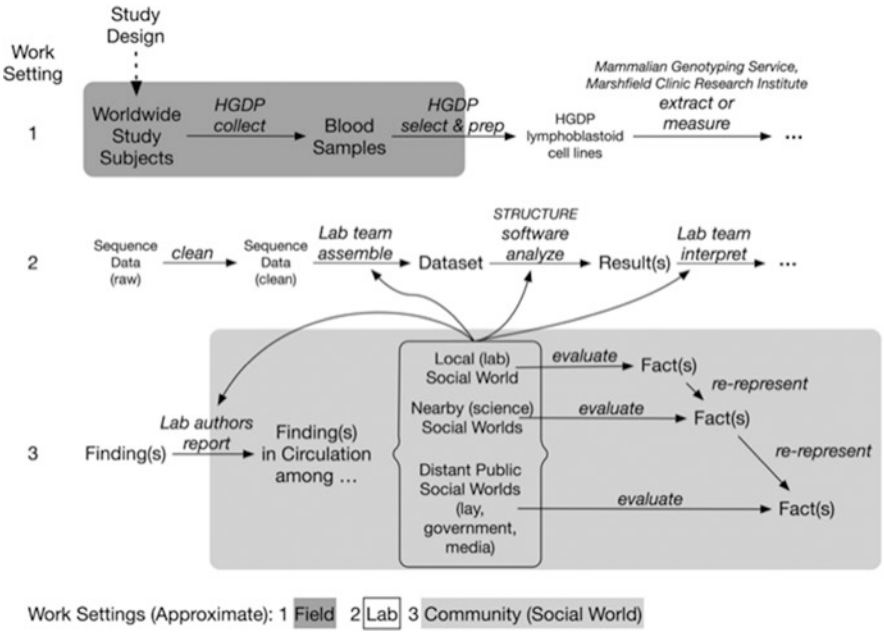


Fig. 4 Workflow diagram following the format of Fig. 1, illustrating specific elements of the dataset assembly and use of data in the Rosenberg et al. 2002 study

Of more interest is when researchers drop DNA sequences in the transition from specimens to data because samples don't meet *theoretical* requirements of their "model-driven" analysis tools. Population genetics theory (and statistical sampling theory) says inferences will be poor for populations represented by only one or two specimens (i.e. sample size n = 1 or 2), so they are not included in the dataset, although they are included in the HGDP-CEPH donor blood tissue specimens, lymphoblastoid cell lines, and DNA sample "screening sets."[12] This kind of hiatus or end to a datapoint and sub-dataset journey is the tip of an iceberg of ways in which data may be "cleaned" or "pruned" in the processing steps leading from material samples to "raw raw" data to "raw" data to "cooked" or processed data.[13] Figure 4 illustrates a workflow for dataset assembly in the work of the Rosenberg Lab following the outline of Fig. 1.

[12] Different investigators and labs set different local sample size thresholds based on varying theoretical requirements for their specific research purposes, so whether a given datapoint can continue on a dataset journey depends on the lab and the project.

[13] The cleaning metaphor supports a useful contrast between "raw" and "cooked" data, even if Bowker (2005, p. 184) is right that "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care."

4.2 Dataset Journey Narrative

Of still more interest are beginnings and endings of the journeys of data *points* that result from further analyses inspired by working with the data *set*. These further practices support stories of data journeys of datapoints from dataset to dataset and journeys of datasets from research project to research project. They are *dataset* journeys: a voyage of the *Beagle* rather than Darwin's voyage or FitzRoy's voyage; voyages of the starship *Enterprise* rather than Kirk's voyage or Spock's voyage.

Samples are gathered together; information is collected from samples and assembled into a dataset; the data journey begins with a scientific study of the dataset. Small and big findings arise and emerge from this traditional kind of scientific work. In addition, medium-sized facts arise about the dataset itself, where a medium-sized fact is a relational fact over the group of datapoints, or a fact derived from the set, but not extending or applying beyond the sample specimens that led to the group of datapoints. Medium-sized facts contrast with Leonelli's (2016) small facts or findings corresponding to individual datapoints and with big facts or findings derived from the analysis of the whole dataset in the light of a theory, question or hypothesis.

Because of the technical nature of the work of comparing genetic sequences, results of model-driven analysis in hypothesis-centric research often reveal salient features *of the dataset*, e.g. features that identify particular datapoints or small groups of datapoints as exceptional.[14] These are "medium-sized" facts or findings about the dataset itself, and thus about the sample set or sample sub-sets. These medium-sized facts can drive dataset journeys less visible than the big fact journeys in which scientists *use* data and whose reports grab the headlines when the science is perceived to have important scientific implications, societal impact or is otherwise controversial.

One of these less visible data journeys concerns individual 995 from the Karitiana in Brazil. The challenge in her journey was due to her traveling companion, individual 996. Individual/datapoint 995 from the Karitiana remained on the dataset journey from 2002 to 2006 at least, but when it was inferred that individual 996 was probably 995's sister (due to the level of genetic similarity), one of them had to get off the ship (dataset). Rosenberg (2006) introduced the convention to drop whichever among pairs of such datapoints had arbitrarily been given the higher-numbered label, so Ms. 996's journey ended while Ms. 995's continued. In other cases, whole families had to exit the journey for analogous reasons. This is not how the data journeys would go if socio-cultural anthropologists rather than geneticists were arranging the journeys, given the fundamentally different orientation of the two disciplines to family-level data. For anthropologists, families represent important units in the organization of cultures, but in the context of population-level genomics, they

[14] Compare Tempini, this volume a, b, on analogous discoveries of middle-sized facts about environmental public health datasets, and Hoeppe, this volume, on discovery of "artifacts" in radio telescope datasets.

represent complications to sampling assumptions needed to apply theory to data and thus are to be avoided.

The character of the journeys of the datapoints in the Rosenberg et al. (2002) dataset does not become apparent until one looks at some of the destinations to which the *dataset* traveled. Here, I focus more on dataset journeys *within* the practices of the Rosenberg Lab and its collaborations and less on data journeys out into the wider specialty and beyond where others can download Rosenberg et al.'s data and software and try to repeat the analysis reported in the publication or construct new datasets from old. My goal in this chapter is modest: to formulate the idea of dataset-centric biology, display some of its narrative forms and visualizations, and underscore its potential value for understanding the organization of contemporary sciences, using an illustrative case, not to establish its generality or reach.

In 2005, Rosenberg et al. (2005) published a defense of their methods and findings in the 2002 paper. They "expanded their earlier dataset" from "377 to 993 markers" so they could evaluate critical responses (e.g. Serre and Pääbo 2004) that human populations are ordered in clines, not clusters. Since this was mostly an expansion, with new datapoints joining the journey, including datapoints of kinds other than microsatellite data, I will not further discuss this paper. I note simply that in 2005 a bunch of new travelers joined on, so we can think of datasets as both structures serving as vehicles for the travel of datapoints and as destinations: datapoints travel from dataset to dataset, getting on or getting off different ships at various "stops."

A different paper, by Ramachandran et al. in 2005, is more interesting for present purposes. Certain features of some of the datapoints in the 2002 study were noted, causing some of them to be dropped and others to be added for this study. The account of the dataset structure in the "Materials and Methods" section (p. 15942) is instructive. In this quotation, note that reference (11) is to Rosenberg et al. (2002).

> **Data.** The data set that we analyzed consists of 1,027 individuals from the HGDP-CEPH Human Genome Diversity Cell Line Panel (10). Several individuals from the collection of 1,056 individuals studied by Rosenberg et al. (11) were excluded from the present analysis. These included the following: (i) no. 1026, who was studied by Rosenberg et al. (11) but who was not in the HGDP-CEPH panel; (ii) nos. 770 and 980, who were identified by Rosenberg et al. (11) as likely labeling errors; (iii) nos. 589, 652, 659, 826, 979, 981, 1022, 1025, 1087, 1092, 1154, and 1235, each of whom was identified by Mountain and Ramakrishnan (12) as a duplicate sample of another individual included in the panel; (iv) nos. 111 and 220, who were identified by Mountain and Ramakrishnan (12) as duplicates of each other but whose population labels differed; and (v) 21 individuals from the Surui population, an extreme outlier in a variety of previous analyses (11, 13, 14). Individuals not studied by Rosenberg et al. (11) but analyzed here included the following: (i) no. 1331, whose genotypes had been unavailable at the time of the Rosenberg et al. (11) study; (ii) nos. 993, 994, 1028, 1030, 1031, 1033, 1034, and 1035, who were previously excluded as members of populations with small sample sizes but who were grouped for the present analysis into Southwestern Bantu (individuals no. 1028, 1031, and 1035) and Southeastern Bantu (individuals no. 993, 994, 1030, 1033, and 1034) populations. Thus, the present data set includes two additional populations along with all populations studied by Rosenberg et al. (11) except Surui for a total of 53 populations.

In addition to the kinds of data "cleaning" mentioned previously, this paper dropped a whole population, the 21 individuals sampled from the Surui in Brazil, who live near the Karitiana by the way, as an "extreme outlier." 21 individual data points were dropped from the journey because of a characteristic of that population as a whole—bad traveling companions one might say. This points to the dataset as itself a "fact" or finding produced by the analyses cited. I describe such facts as "medium" sized because they form the basis for the analyses leading to big facts, but are facts about the datasets themselves, analogous to the way the small facts of interest here are facts about individual sample subjects.

Equally interesting is the continuation on the dataset journey of datapoints 1028, 1031, 1034 and 993, 994, 1030, 1033, and 1034 who didn't make the earlier segment of the journey from HGDP-CEPH sample set to the dataset of Rosenberg et al. (2002), but who were allowed to get back into the research program and the overall dataset journey at a different research project and publication "stop" due to the small sample size threshold set by Rosenberg's project. Ramachandran et al. regrouped them into Southwestern and Southeastern Bantu, in effect defining new populations by means of a statistical procedure and adding population labels ("metadata") in the lab rather than as a result of "self-reporting" or "data collection" in the field. In effect, they were interpreted as coming from different places than their original "relevance labels" (place of origin) designated, so they in effect, got new "visas" to travel by Ramachandran et al. (see Leonelli 2011 and 2016 on relevance and reliability labels).[15]

These and other papers appearing between 2002 and 2005 prompted Rosenberg to publish the 2006 paper described above (Sect. 4.2). It visualizes datapoint journeys to and among datasets in a spreadsheet format. Although this paper can be read as part of the other visualization of dataset journeys in the Rosenberg lab (on the datasets web page), this paper can alternatively be read as a new kind of publication in this specialty: a data "curator" paper, signaling a kind of work analogous to that of the specialized data curators in the bio-ontology projects Leonelli (2016) discusses. Instead of tracking changes to datasets within the "materials & methods" or "supplementary" sections of publications of a research project, Rosenberg (2006) is a publication aimed at tracking datasets and, more importantly, proposing standards for naming and using these datasets. This implies a new level of attention to the

[15] M'charek 2005 writes about the "passports" DNA samples needed to pass from one part of the forensics lab she studied to another. I use the related metaphor of "visa." The difference of metaphors is that the passport is a license to travel. The visa is a license to travel in a specific place for a specific period of time. To continue the metaphor, DNA sequences or their tissue samples get "passports" when they are enrolled as samples in the CEPH bio-repository. To get a visa to be included in a particular dataset, the "receiving" country—research group in this case—has to approve. Approval can turn on questions of "desirability" (un-sequence-able tissue samples are undesirable; duplicates are undesirable) or for "theoretical" reasons (sample size too small). Barragán, on the other hand, writes about dataset curating practices in terms of data noise and data silencing as life scientists confront genomic datasets with archaeological, ethnographic, ethnohistorical and linguistic datasets about pre-Columbian and contemporary indigenous groups in northern South America (Barragán 2016, 2017).

ways in which data visualizations (and narratives) set data in motion and contribute to data travel among research projects.

The curation of HGDP-derived datasets in Rosenberg (2006) is not for the sake of online database management and curation of sequence *datapoints*, accessible in the way the “omics” databases are. Rather, it attempts to curate, by documenting in a publication, both the dataset that was initially assembled for the 2002 study *and the journeys* of the datapoints among datasets as a widening circle of researchers used and tinkered with the 2002 dataset to produce new datasets. Differently put, researchers such as Rosenberg (and perhaps those involved in the HGDP more broadly) seem to be taking a new and active interest in conceptualizing and representing the “middle-ground” dataset landscape in which many of their data-centric practices are enacted.

5 A Model of Dataset Journeys and Conclusions

I don’t pretend to have done more than scratch the surface of a case study of dataset-centric human population genomics. What I hope to have illustrated is that there is a “middle ground” data landscape between the traditional hypothesis-driven use of data as familiarly described by philosophies of “scientific method” and the new ground of data-centric science described so well by Leonelli. I have gestured at ways in which individual datapoints in datasets, at least in human population genomic diversity studies, make data journeys that are of neither of Leonelli’s two kinds, but which resemble them in some respects and to some degree and differ in other respects. Perhaps other question-driven scientific specialties are also influenced by what is newly afforded in the rapidly changing landscape of computational and online digital methods, so there may be many forms of dataset-centric scientific practices waiting to be described. Morgan’s study (this volume) of two kinds of data journeys in economics regarding national income accounts and indicator series also concern humans and population data, though with a very different subject matter and principles for dataset formation and use than the biological genomics studies considered here.

In this chapter, I have characterized data journeys in terms of a model comprised of three kinds of components: data structures, data visualizations and data journey narratives. The details of specific scientific practices involved in producing and using these components do matter, if we are to understand these data journeys in middle-ground landscapes of datasets and how they might inform big findings and facts. This is particularly true of genomic ancestry projects like HGDP and biomedical projects like personalized genomic medicine. A further result of this case study is important for present purposes to signal a connection of dataset-centric biology to characteristic features of emerging data-centric “omics” research practices: the emergence of a “bioinformatics” practice alongside the basic, craft research process of asking and answering questions, posing and testing hypotheses.

A distinct and notable line of investigation emerged in population genomics in roughly the time frame 2002–2006 around detection of close relationships among individuals with sequence data in genetic datasets of this kind, both for ancestry and biomedical studies (e.g. Boehnke and Cox 1997; Epstein et al. 2000). This literature, reviewing both datasets and software and modeling approaches, flourished to the point that there are now review articles "benchmarking" different relatedness inference methods (e.g. Porras-Hurtado et al. 2013; Ramstetter et al. 2017). This is evidence of a "standards" specialization emerging within dataset-centric population genomics analogous to the kind of "infrastructure" supporting a bioinformatics specialization that Leonelli (2011, 2016) discusses for data-centric "omics" biology (see also Tempini 2017, this volume a, b).

Moreover, Rosenberg's efforts in (2006) are, I suggest, aimed at supporting a narrative that *steers* the dataset *journeys* of particular datapoints. This is not quite like the curation that goes on in the world of "omics," because the target is *datasets* that are purpose-built and question-driven. The corresponding findings reported in this emerging dataset curation literature are medium-sized, regarding these datasets themselves. The normative directions derive from the standards concerning what sorts of findings or "big" facts can or should be derived from datasets of particular kinds or with particular characteristics.[16]

The data journey discussed here is not quite like the ones Leonelli describes, nor like many of those detailed in Howlett and Morgan (2011) on traveling facts. The journey of the *dataset* is driven in part by the conventional publication system in which peer-reviewed publications of findings using these datasets (together with ancillary visualizations in web pages, spreadsheets and supplementary material) draw attention to the datasets themselves and provoke scrutiny of the datapoints. This scrutiny may extend, moreover, to science studies analysts tracing dataset and datapoint journeys in terms of the components of a model in which data structures, data visualizations and data journey narratives mobilize datapoints in dataset journeys. These journeys may encourage re-use of the dataset or construction of related or alternative datasets, adding and dropping datapoints, thus driving the data journey(s) forward. A different story will be needed for the drivers of "sample sets" such as blood donor samples, cell lines, and extracted DNA sample sets because the differences in materiality matter. The contingency of such sample sets being *available* to feed the production of datasets is critical to dataset journeys.[17]

Dataset journeys, classification schemes and data visualizations designed to maintain and manage them in contemporary biology are driven by a hybrid system of formal, institutionalized, community-sanctioned publishing and quasi-"samizdat" or "self-publishing" systems of personal, individual, laboratory, and university-sponsored websites for distributing datasets and software as well as publications. Unsurprisingly, there is also an emerging effort to institutionalize these kinds of

[16] On the links between data, classification systems and standards, see Bowker and Star (1999).

[17] It remains to be seen whether the model described here applies to sample journeys as well as to data journeys. Thanks to Carlos Andrés Barragán for emphasizing this point.

publication as well, in data journals and dataset archiving services. There is, nevertheless, less standardization of data formats in *dataset* curation and publication as displayed in this case study, even if there is substantial standardization of some of the data content of datapoints due to the rise of data-centric biology and centralized, shared databases for datapoints.[18]

The lower degree of standardization is no doubt partly due to the fact that nearly every population geneticist running a lab today is (or is becoming) a coder who writes their own software in their own way, typically built to read and analyze data formatted anachronistically for their own lab's purposes. It is a relatively manageable problem for others to gain access to such data and tools: if the software and the dataset can be downloaded and the provenance and versioning meta-data for the software is curated along with the dataset, one can (with effort) get the original software to analyze the original dataset. Nevertheless, it *is* a problem. And it entails different kinds of practices and workflows than biological research had required before the data and software coding revolutions of the last few decades.[19]

It means that data journeys may require *software journeys:* particular software versions (and perhaps operating systems or whole virtual machine execution environments) may have to chaperone datasets in order for scientific analyses to be repeated and re-evaluated. Indeed, software versioning is a form of software journey in this middle-ground landscape between the small landscapes of datapoints and small facts on the one hand, and the big landscapes of research findings and big facts on the other.[20]

One more comparison of *dataset*-centric biology with the bioinformatics dimensions of *datapoint*-centric biology will display some similarities and highlight differences. Rosenberg also engages in dataset packaging practices which parallel Leonelli's (2011, 2016) labeling story. Relevance labels, which signal the value of datapoints for particular kinds of journeys and analyses, are included in the dataset (or linked to it) by coding what are called "pre-defined" populations as part of the data records. These are names like *Karitiana*, for the name of the people/place of a certain culturally specific, geographically localized group of people; like *Brazil*, for the name of the nation-state in which the Karitiana are (largely) thought to reside at present; and like *AMERICA*, for the name of the "region" or "continent" of which the relevant nation-state is considered part (see Barragán 2016). As we saw, these "pre-defined" populations played no role in the cluster based *inference* of ancestry

[18] See Tempini, this volume a, b, for a case where infrastructures are built to systematize, institutionalize and standardize the sourcing, hosting, manipulation and generation of datasets. See also Tempini (2017). Morgan's two cases (this volume)—national income accounts and UN indicators of national "health"—also suggest different subject matters and principles may require or lead to different respects and degrees of both standards and infrastructure.

[19] A recent trend in bioinformatics is to solve this problem by making the entire "execution environment" of a whole computational "scientific workflow" the basic unit to be prepared for data journeys. Rather than just data, or software or both, this workflow-centric biology involves creating whole execution environments of data, software and computer operating system as the "basic units" (Meng and Thain 2017).

[20] Thanks to Jason Oakes for pressing this point.

relations in Rosenberg et al. (2002) directly, though they surely did play a role in attracting the attention of those who conducted the initial *sampling* effort because the collectors were interested in sampling human genetic diversity, especially among small groups that might soon disappear. It is no accident that the HGDP-CEPH samples are not (all) drawn from nation-state capital cities, for example, nor from a conventional grid of equally spaced sample locations defined by the geometry of the Earth (constrained by availability of time, money, skill, and interest of collectors in sampling at a particular geographic "scale"). The HGDP-CEPH sample panel was made after several years of inconclusive internal battle over what would be an appropriate sampling protocol for the HGDP (see NAS 1997, for example), but it is not the focus of interest and concern here.

Leonelli's "reliability" labeling practices are also included in Rosenberg's dataset curation practices, though the latter do not appear in "evidence codes" stored in an online accessible "bio-ontology" or "database." Rather, they appear in the "Materials and Methods" sections of "ordinary" scientific papers or coded in archived, downloadable "data" (i.e. meta-data) files devoted to answering a research question or testing a model-driven hypothesis. Cross-referencing a DNA sequence dataset via joining ID field, "Pop ID," is perhaps assurance of both reliability and readability of the data file.

It is common to describe the sources and methods used to generate a dataset in any scientific paper worthy of the name. In the case of human population genomics diversity papers, this extends to discussion of individual datapoints and, increasingly, to a methods literature of papers like Rosenberg (2006) devoted to curation of datasets apart from the research papers devoted to reporting the "big"-fact findings of question-driven research projects. Interestingly, unlike the methods sections of ordinary "omics" papers from molecular biology labs, precious little, if any, space in the Materials and Methods sections is devoted to reporting on the protocols and technologies used to actually generate the sequence data. This may seem surprising, but the data curation tasks for these dataset-centric research programs are less concerned with reporting on *sequence* data reliability than on sequence *dataset* reliability for the question at hand.[21]

In the illustrative case of dataset-centric research discussed here, there are two aspects of the case that may require recalibrating the concept for use beyond my case study of a human population genomics data journey. First, the research is in the population sciences. Population sciences by their nature deal with collections of "individuals" (members of populations). There is a sense of compositionality of the relevant data that is integral to this kind of research. The very idea of a population is that it be composed of members (or parts, depending on one's metaphysics). Surely attention in such contexts is focused on datasets since *collections* of datapoints tend to be used to represent data about populations, e.g. through statistical reasoning that treats the collected data as a sample from a population whose

[21] Studies of ancient human DNA are something of an exception, since the quality of sequence data deriving from ancient, even fossil, specimens is a special problem. See e.g. Veeramah and Hammer (2014) for a relatively recent overview of whole genome sequence data.

unknown properties are subjects of theoretical inquiry, or through some other mode of aggregation, extrapolation or inference from information about members to a set or population. Inquiry may even focus on properties of individuals *qua* members of a population, in a form of research known in some fields as "downward causation," whereby properties of the group cause (or determine) properties of the members. So perhaps the notion that the case discussed here illustrates dataset-centric biology may not generalize beyond population sciences.

A second kind of particularity of the case study is the way it focuses on humans. Data in human biology can be difficult to collect for familiar reasons of ethical or legal restraint or constraint, difficulty of access, expense, entanglement with political, social or cultural differences between researchers, sponsors and potential "subjects," and for many other reasons (Barragán 2012). The constraints may be quite different than for social science data collection about humans (e.g. Morgan, this volume). Biological datasets collected from human subjects thus tend to be more "precious" to researchers than data collected from non-humans (though not always of course—natural history is often pursued in out of the way places that can be hard, expensive, or unpleasant to get to and work in). Human genome diversity data on members of the Karitiana in South America, for example, are critical for the story of human diversity in ways that make these people much more than mere "sample subjects" (see Barragán 2016).

The virtues of "model organisms" include features that tend to make data collection easy, cheap, and fast, and the data, in consequence, relatively disposable. As the unit cost of DNA sequencing falls with advances in technology, on top of scaling and standardizing effects of commercialization, researchers may find it easier to collect new fruit fly specimens, extract new DNA samples, and generate new collections of sequence data for their project-specific uses, than to rely on data already generated by other labs (that may have used doubtful or out-of-date methods, or with questionable expertise, or based on samples less specifically suited to a different project's questions and purposes).

I conclude by noting that the case study analysis and model of datapoint and dataset journeys sketched here indicates not only that new modes of data-centric science are emerging, but that old ones are transforming—particularly around the packaging, vehicles, conveyances, and infrastructure that gets organized or reorganized to put research subjects, specimen samples, extracted materials, and data *points* and *sets* into motion on new kinds of journeys to new kinds of destinations.

Acknowledgments I thank Sabina Leonelli and Niccolò Tempini for inviting me to contribute to their project on "Varieties of Data Journeys" and to Mary Morgan in particular for an interesting collaborative writing experience which helped shape the way I think about data points and data sets. I thank Sabina and Mary for encouraging my larger collaborative project on the re-situation of scientific knowledge with Elihu Gerson, Jason Oakes, Carlos Andrés Barragán, and Alok Shrivastava. I thank Rasmus Winther for prompting my initial interest in this topic. I thank Mary

Morgan for conversations that led to the paper and the conference audience in Exeter for comments. Carlos Andrés Barragán and Sabina Leonelli made many very helpful comments, clarifications and corrections that greatly improved the manuscript. Members of Bill Bechtel's lab group and Bob Westman at UC San Diego provided helpful feedback at a critical stage. Members of my joint lab group with Roberta Millstein, especially Roberta, Michael Hunter, and Tiernan Armstrong-Ingram, provided important input and discussion. I thank Elihu Gerson for feedback, encouragement and general support. I thank the people of California and NSF Grant SES 1849307 for financial support.

References

Barragán, C.A. 2012. Molecular Vignettes of the Colombian Nation: The Place(s) of Race and Ethnicity in Networks of Biocapital. In *Racial Identities, Genetic Ancestry and Health in South America*, ed. Sahra Gibbon, Ricardo Ventura Santos, and Mónica Sans, 41–68. New York: Palgrave Macmillan.

———. 2016. *Lineages Within Genomes: Situating Human Genetics Research and Contentious Bio-Identities In Northern South America*. PhD Dissertation, Department of Anthropology/ Science & Technology Studies Program (STS), University of California, Davis.

———. 2017. *Substantiating Genetic and Cultural Continuity: Partial Connections Between Genomic, Archaeological and Linguistic Datasets*. Paper presented at the Annual Meeting of the International Society for the History, Philosophy and Social Studies of Biology (ISHPSSB) and the Associação Brasileira de Filosofia e História da Biologia (ABFHIB). Panel: The Re-situation of Scientific Knowledge, Organized by J. R. Griesemer. July 17, 2018. São Paulo: ISHPSSB/ABFHIB.

Blei, D. and J. Lafferty. 2009. Topic Models.. http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf. Accessed 14 June 2018.

Boehnke, M., and N. Cox. 1997. Accurate Inference of Relationships in Sib-Pair Linkage Studies. *American Journal of Human Genetics* 61: 423–429.

Bowker, G. 2005. *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker, G., and S. Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Cann, H., C. de Toma, L. Cazes, M. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. Ferrara, J. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. Herrera, X. Huang, J. Kidd, K. Kidd, A. Langaney, A. Lin, S. Mehdi, P. Parham, A. Piazza, M. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. Weber, H. Greely, M. Feldman, G. Thomas, J. Dausset, and L. Cavalli-Sforza. 2002. A Human Genome Diversity Cell Line Panel. *Science* 296 (5566): 261–262.

Currie, A. 2018. *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences*. Cambridge, MA: MIT Press.

Epstein, M., W. Duren, and M. Boehnke. 2000. Improved Inference of Relationship for Pairs of Individuals. *American Journal of Human Genetics* 67: 1219–1231.

Griesemer, J., and C. A. Barragán. 2019. *Standard Grant: A Case Study of How Re-Situation of Scientific Knowledge from Human Population Genomics Works*. NSF grant SES-1849307, 2019–2021.

Hoeppe, Götz. this volume. Sharing Data, Repairing Practices: On the Reflexivity of Astronomical Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Horton, R. 2003. Paper of the Year. *Lancet* 362: 2101–2103.

Howlett, P., and M.S. Morgan, eds. 2011. *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*. Cambridge: Cambridge University Press.

Leonelli, S. 2011. Packaging Small Facts for Re-Use: Databases in Model Organism Biology. In *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, ed. Peter Howlett and Mary S. Morgan, 325–348. Cambridge: Cambridge University Press.

———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

M'charek, Amade. 2005. *The Human Genome Diversity Project: An Ethnography of Scientific Practice*. Cambridge: Cambridge University Press.

Marshfield Clinic Research Institute. 2014. http://www.marshfieldresearch.org/about/welcome, http://www.marshfieldresearch.org/irdl/research-support. Accessed 26 Aug 2019.

Meng, H., and D. Thain. 2017. Facilitating the Reproducibility of Scientific Workflows with Execution Environment Specifications. *Procedia Computer Science* 108C: 705–714.

Morgan, M.S. 2014. Resituating Knowledge: Generic Strategies and Case Studies. *Philosophy of Science* 81 (5): 1012–1024.

———. this volume. The Datum in Context: Measuring Frameworks, Data Series and the Journeys of Individual Datums. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

NCBI (National Center for Biological Information). 2019. https://www.ncbi.nlm.nih.gov/nucleotide/. Accessed 26 Aug 2019.

Pemberton, T., M. DeGiorgio, and N. Rosenberg. 2013. Population Structure in a Comprehensive Data Set on Human Microsatellite Variation. *Genes, Genomes, Genetics* 3: 891–907.

Porras-Hurtado, L., Y. Ruiz, C. Santos, C. Phillips, A. Carracedo, and M. Lareu. 2013. An Overview of STRUCTURE: Applications, Parameter Settings, and Supporting Software. *Frontiers in Genetics* 4: 1–13. https://doi.org/10.3389/fgene.2013.00098.

Pritchard Lab. 2019. *Structure Software*. http://web.stanford.edu/group/pritchardlab/structure.html. Accessed 26 Aug 2019.

Pritchard, J., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155: 945–959.

Pritchard, J., X. Wen, and D. Falush. 2010. Documentation for *Structure* Software: Version 2.3. http://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/structure_doc.pdf. Accessed 6 June 2018.

Ramachandran, S.O., C. Deshpande, N. Roseman, M. Feldman Rosenberg, and L. Cavalli- Sforza. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *PNAS* 102 (44): 15942–15947. www.pnas.org, 10.1073 pnas.0507611102.

Ramstetter, M., T. Dyer, D. Lehman, J. Curran, R. Duggirala, J. Blangero, J. Mezey, and A. Williams. 2017. Benchmarking Relatedness Inference Methods with Genome-wide Data from Thousands of Relatives. *Genetics* 207: 75–82. https://doi.org/10.1534/genetics.117.1122.

Rosenberg, N. 2006. Standardized Subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, Accounting for Atypical and Duplicated Samples and Pairs of Close Relatives. *Annals of Human Genetics* 70: 841–847.

Rosenberg Lab. 2018. https://rosenberglab.stanford.edu/diversity.html. Accessed 26 Aug 2019.

Rosenberg, N., T. Burke, M.W. Feldman, P. Friedlin, M.A.M. Groenen, J. Hillel, A. Mäki-Tanila, M. Tixier-Boichard, A. Vignal, K. Wimmers, and S. Weigend. 2001. Empirical Evaluation of Genetic Clustering Methods Using Multilocus Genotypes from 20 Chicken Breeds. *Genetics* 159: 699–713.

Rosenberg, N., J. Pritchard, J. Weber, H. Cann, K. Kidd, L. Zhivotovsky, and M. Feldman. 2002. Genetic Structure of Human Populations. *Science* 298 (5602): 2381–2385.

Rosenberg, N., S. Mahajan, S. Ramachandran, C. Zhao, J. Pritchard, and M. Feldman. 2005. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genetics* 1 (6): e70. https://doi.org/10.1371/journal.pgen.0010070.

Serre, D., and S. Pääbo. 2004. Evidence for gradients of human genetic diversity within and among continents. *Genome Research* 14: 1670–1685. http://www.genome.org/cgi/doi/10.1101/gr.2529604.

Tempini, Niccolò. 2017. Till Data Do us Part: Understanding Data-Based Value Creation in Data-Intensive Infrastructures. *Information and Organization* 27: 191–2010.

Tempini, Niccolò. this volume-a. The Reuse of Digital Computer Data: Transformation, Recombination and Generation of *Data Mixes* in Big Data Science. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Tempini, Niccolò. this volume-b. Visual Metaphors: Howardena Pindell, Video Drawings, 1975. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Veeramah, K., and M. Hammer. 2014. The Impact of Whole-Genome Sequencing on the Reconstruction of Human Population History. *Nature Reviews Genetics* 15: 149–162.

Wade, Nicholas. 2014. *A Troublesome Inheritance: Genes, Race and Human History*. New York: Penguin Books.

Wills, Melissa. 2017. Are Clusters Races? A Discussion of the Rhetorical Appropriation of Rosenberg et al.'s "Genetic Structure of Human Populations.". *Philosophy Theory, and Practice in Biology* 9 (12): 1–24.

James Griesemer is a Distinguished Professor and Chair of the Department of Philosophy at the University of California, Davis, and Member of the UC Davis Science and Technology Studies Program, the Center for Science and Innovation Studies, the Cultural Studies Graduate Group, the Population Biology Graduate Group and the Center for Population Biology. He is also Past President of the International Society for History, Philosophy and Social Studies of Biology and a Member of the KLI in Klosterneuburg, Austria. His primary interests are philosophical, historical and social understanding of the biological sciences, especially evolutionary biology, genetics, developmental biology, ecology and systematics. He has written on a wide variety of topics in history, philosophy and social studies of biology, including models and practices in museum-based natural history, laboratory-based ecology, units and levels of inheritance and selection in evolutionary biology and visual representation in embryology and genetics. He is currently writing a book, *Reproduction in the Evolutionary Process*, which develops a theory of reproduction more comprehensive than current philosophical accounts of inheritance, with applications to theoretical problems ranging from the nature and origin of living systems, evolutionary transitions, eco-evo-devo and cultural change.