palgrave
macmillan

QUERY FORM

Journal	**BIOSOC**	
Manuscript ID	**[Art. Id: biosoc201522]**	

PLEASE NOTE:

UNLESS WE RECEIVE A SIGNED COPYRIGHT FORM FROM THE AUTHOR, WE WILL BE UNABLE TO PUBLISH THIS ARTICLE. IF YOU HAVE NOT ALREADY SUPPLIED A SIGNED COPYRIGHT FORM, PLEASE ENSURE YOU INCLUDE THIS WHEN YOU RETURN THIS PROOF

IF YOU HAVE INCLUDED ANY IMAGES THAT ARE NOT YOUR OWN IN THIS ARTICLE, PLEASE ENSURE YOU SUPPLY THE PERMISSION/CREDIT FOR EACH OF THEM WHEN YOU RETURN THIS PROOF

Papers published via advance online publication (AOP) are fully citable using the Digital Object Identifier (DOI) system and the publication date. For example, per the biosoc style guide:

Crane, J.T. ,(2011) Viral cartographies: Mapping the molecular politics of global HIV. BioSocieties, advance online publication January 31, doi: 10.1057/biosoc.2010.37

Author :- The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections at the appropriate positions in the text and acknowledge that a response has been provided in 'Response' column.

Query No.	Description	Response
GQ	Your manuscript has been copy-edited in keeping with journal style. Please pay attention to the changes to the use of double and single quote marks. Double quote marks (" … ") are to be used for direct quotes (directly quoting from articles/ interviews etc), and single quote marks are for 'scare' quotes. If you have any concerns, please flag them when returning your proof.	
Q1	Do you wish us to remove your email address from the correspondence details that appear on this proof? If you would like it to be retained, please be aware that it cannot be removed at a later stage. Please note that your paper will publish online and your email address will be visible to anyone accessing it on the journal website. Do you wish us to remove your email address? Yes/No.	yes fine to have it.
Q2	If the answer to Query 1 is 'No', then please provide email addresses for the authors Ruth McNally, Richard Mills and Stuart Sharples.	

palgrave
macmillan

QUERY FORM

Journal	**BIOSOC**	
Manuscript ID	**[Art. Id: biosoc201522]**	

Author :- The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections at the appropriate positions in the text and acknowledge that a response has been provided in 'Response' column.

Query No.	Description	Response
Q3	Please provide complete affiliation details for the authors Ruth McNally, Richard Mills and Stuart Sharples.	
Q4	Please provide biography for the authors Ruth McNally, Richard Mills and Stuart Sharples, not exceeding 80 words.	Richard Mills is a computational social scientist who studies peer production and behaviour through online data traces. Richard's PhD thesis concerned distributed moderation voting platforms like Reddit and the Stack Exchange - and he has also studied subjects like Github and the Sequence Read Archive. Richard currently works on the "WikiRate" project, based at Cambridge University. The aim of this project is to develop a peer production platform for the collection and analysis of information about companies' ethical and sustainability-
Q5	There are four SRAs mentioned here. Please check.	
Q6	Please provide the expansion of 'STS'.	
Q7	Please confirm change of (Mackenzie, 2015) to (Mackenzie, 2014) as per the reference list.	
Q8	Please provide volume number, issue number and page range details in the reference Goldman et al (2013).	
Q9	Reference Kelty and Landecker (2009) not cited in the text. Please cite in the text, else delete from the reference list.	
Q10	Please provide volume number, issue number and page range in the reference Leonelli and Ankeny (2011).	
Q11	Please provide issue number and page range in the reference Leonelli (2014).	
Q12	Please provide captions for Figures 1–6.	

Original Article

Post-archival genomics and the bulk logistics of DNA sequences

Adrian Mackenzie*, Ruth McNally, Richard Mills and Stuart Sharples
Q1 Sociology Department, Lancaster University, Bailrigg, LA14YL, UK.
E-mail: a.mackenzie@lancaster.ac.uk
Q2 Q3 *Corresponding author.

Abstract DNA sequence data are currently viewed as a 'bedrock' or 'backbone' of modern biological science. This article traces DNA sequence data produced by so-called 'next generation sequencing' (NGS) platforms as it moves into a biological data infrastructure called the Sequence Read Archive (SRA). Since 2007, the SRA has been the leading repository for NGS-produced nucleotide (DNA and RNA) sequences. The way sequence data move into the SRA, we suggest, is symptomatic of a decisive shift towards *post-archival genomics*. This term refers to the increasing importance of the *logistics* rather than the biology of sequence data. In the SRA, logistical concerns with the bulk movements of sequence data somewhat supplant the emphasis in previous genomic and biological databases on contextualising particular sequences and cross-linking between different forms of biological data. At the same time, post-archival logistics do not necessarily flatten genomic research into global genomic homogeneity. Rather, the SRA provides evidence of an increasingly polymorphous flow of sequence data deriving from an expansion and diversification of sequencing techniques and instruments. The patterns of movement of data in and around the SRA suggest that sequence data are proliferating in various overlapping and sometimes disparate forms. By mapping differences in content across the SRA, by tracking patterns of absence or 'missingness' in metadata, and by following how changes in file formats highlight uncertainties in the definitions of seemingly obvious DNA-related artefacts such as a sequencer 'run', we highlight the growing lability of nucleotide sequence data. The movements of data in the SRA attest to a decisive mutation in sequences from biological bedrock to an increasingly expandable material whose epistemic and technological value remains open to reinvention.

BioSocieties (2015) **0,** 1–24. doi:10.1057/biosoc.2015.22

Keywords: DNA sequencing; genomics; data infrastructures; scale; value; archive

Introduction

> There's no real relation between the physical organisation of the sequence and the submission to the sequence archive.
>
> (Interview with SRA data curator)

Recent ethnographic work has highlighted the Toyotarisation or post-Fordist re-organisation of DNA sequencing in large biological research laboratories such as the Broad Institute in Cambridge MA, the Sanger Centre in Cambridge UK, or the J. Craig Venter Institute (JCVI) in San Diego (Helmreich, 2009; Chow-White and Garcia-Sancho, 2011; Stevens, 2011, 2012; Hilgartner, 2013). We can well imagine similar developments at other large sequencing centres such the BGI (formerly Beijing Genomics Institute), Shenzhen, China. The 'consumption' of contemporary DNA sequence data as information to be searched, mined or modelled to produce knowledge and economic value has also been widely discussed in terms of processes of globalisation (Thacker, 2005) and financialisation (Sunder Rajan, 2006). As Stefan Helmreich (Helmreich, 2008) and Hallam Stevens in turn point out, sociological and anthropological work explores "how living things have become a form of property or a commodity, how they have become involved in regimes of speculation and profit generation" (Stevens, 2011, p. 218). Finally, between sequencing and the consumption of sequence data stand data sequence databases. It is usually thought that the growth of data infrastructures for DNA sequences over the last two decades responds to the ongoing epistemic and economic investment in DNA sequencing as a fundamental technique in the biosciences. In science and business, databases epitomise the organisational practices and logics that knit aggregates of people, things and transactions together in vast accumulations. It is no surprise that the growth in biological data infrastructures such as sequence databases has been analysed in terms of the tensions between the economically loaded ambitions of large-scale genomics research and the ongoing dispersed localised practices of genomic scientists, many of whom only partially conform to the demands and rhythms of global technoscience (Leonelli, 2013; Leonelli, 2014). If the global ambitions of genomic data infrastructures exemplify the encounter between capitalist economies and biology, the dispersed localised practices suggest at least some ongoing irreducibility or resistance to the patterns of circulation typically associated with biocapital.

In many settings, how something is stored and accessed, how something is remembered or retrieved, constitutively affects what that thing is (see Bowker, 2005 for an extended development of this point). This has long been the case in molecular biology and genomics (Hilgartner, 1995). Currently, it may be, however, that the data infrastructures themselves are in certain respects taking on a more generative role in genomic research. It could be that the contemporary explosion of whole genome sequencing of large cohorts and population strata, metagenomic studies of whole ecosystems, or the multiplying varieties of targeted sequencing of individual cells or tissues respond to the existence of archives and data infrastructures that promise to accommodate and ameliorate all the troubling subtle biological complexities that interest biosciences. If that is the case, the practical arrangements for moving, storing, copying and cataloguing sequences would have a rather different significance than usually understood. They would take on an importance rather like the logistics and supply chain management systems that Amazon uses or that arrange the delivery of goods in abundance to the shelves of a large contemporary supermarket (Busch, 2007; Neilson, 2012). The logistics of sequences would, on this account, constitute a key material practice or a dynamic that reaches upstream into their Toyotarised production and downstream into their often speculative 'consumption'. From this perspective, we might understand the relentless investment and intensification of DNA sequencing in many different venues as an effect of the power of logistics to configure extremely far-reaching patterns of movement (Tsing, 2009).

The primary aim of this article is to explore some recent genomic sequence data infrastructures with an eye on how they move sequences around. We suggest that the logistics of sequence data in these infrastructures gives rise to what we term *post-archival* genomics. In order to empirically ground this claim, we turn to three leading contemporary DNA sequence archives (in Europe, USA and Japan) known collectively as the sequence read archive (SRA). We argue that the way the SRA responds to, and even incites, the bulkiness of next generation sequencing (NGS) data sets on the one hand, and the ever-expanding diversity of its applications and users on the other, distinguishes it from earlier archives. In contrast to earlier archives, the SRA is organised in relation to the diverse and changing modes by which sequences are produced and processed rather than their biological features and functions. Rather than a showroom where users can browse and interact with and experience the data, the SRA functions like a supply chain management system for receiving, storing and dispatching the data in a containerised form. Methodologically speaking, the SRA can be understood not only as a distribution system for bulky DNA sequence datasets, but also as a kind of informant about post-archival genomics based on its unique perspective on the bulk movement of NGS data. As we shall see, by interrogating the SRA we can learn something about the range of bioscience projects, the many different domain-specific practices, the shifts in sequencing techniques and the varying scales of investment in DNA sequences in recent genomics. The SRA as informant adds invaluable insights into post-archival genomics, in particular that the constituent repositories of the SRA do not exactly mirror each other. While the sequence data may be fully shared and mirrored, other metadata objects (submissions, studies) that are present in the SRA in Europe are not present in the SRA in the USA and *vice versa*. However, rather than failures or aberrations, we suggest that lack of alignment and chronic 'missingness' of metadata in the SRA are characteristic features of post-archival genomics that derive from the primary function of the repository, which is to move, rather than to archive, data deriving from the proliferation and diversification of sequencing practices.

We begin our article with an overview of the accelerating rate of genome data production and the diversification of genomics applications, and how these have given rise to the inclusion within the International Nucleotide Sequence Database Collaboration (INSDC) of a new type of repository specifically for NGS genomics, namely, the SRA. In the next section we begin our interrogation of the SRA as post-archival informant. Key to the movement of sequences through the SRA are logistical infrastructures known as application programme interfaces (APIs). Designed to allow users to search, deposit and retrieve data from the SRA, we repurpose the APIs to allow us to interrogate the SRA as informant for mapping global concentrations and dispersal of DNA sequencing. The responses of this informant, however, are often quite hard to decipher and need to be interpreted and mapped in dialogue with genomic scientists who use it and data curators who design and manage it. To that end, we draw on interviews, workshops and scientific publications to help make sense of the query results and data we obtained from the SRA. We start with a comparison of the SRA and post-archival genomics with the archival repositories and associated practices of community databases and encyclopaedic biomolecular databases. We graphically illustrate how, in post-archival genomics, the topology of different study types and their modes of producing sequence data are highly varied. In the next two sections we document two phenomena, namely, the lack of alignment between the various instances of the SRA, and significant gaps or 'missingness' across the SRA metadata fields, both of which, we suggest, are manifestations of the dispersal and discontinuities that define post-archival genomics. In our final investigation we turn to 'runs', the SRA metadata object that contains NGS sequence data, only to discover an orientation towards the logic of logistics even at this bedrock level in post-archival genomics.

Sequences: From Base Pairs to Databases

The work of associating small differences in DNA with biological processes has long entailed comparison of DNA sequences and sometimes whole genomes. In the 1980s, the sequence databases GenBank in the USA, European Molecular Biology Laboratory (EMBL) Bank in Europe and the DNA Data Bank of Japan (DDBJ) started exchanging their sequence data in order to facilitate those comparisons. For almost three decades now, the INSDC has represented an international commitment to 'free and unrestricted access' to sequence data. The three institutions comprising INSDC – the DDBJ, the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL–EBI) and the US National Library of Medicine's National Center for Biotechnology Information (NCBI) – sought to make all sequence data, along with the crucial metadata, available not only because of the widely recognised need to facilitate re-analysis or replication. That now long-standing commitment to shared sequence data was a crucial component of the advent of genomics as a deeply databased knowledge project. The advent of genome sequencing projects in the early 1990s more or less confirmed this databased state of affairs. As genomic scientists sequenced whole and part genomes (humans, microbes, plants, animals, fungi and so on), genomics and associated 'omic' sciences could only maximise the epistemic and economic harvest of sequencing work if sequences were available for comparative analysis. Like molecular biology before it, the epistemic power of genomics was predicated on comparing sequences in various ways, and to this end, sequences were of necessity shared. DNA sequences were the shared base of genomics, and genomics was predicated on shared base pairs.

In the decade after the sequencing of the human genome (Mardis, 2011), genomics developed new forms of sequencing. Both the sequencing instruments and their application diversified. Whole genomes, transcriptomes, exomes, epigenomes, CHIP-seq and RNA-seq are just some of the roughly 20 different kinds of sequencing recognised at the end of 2012 (Shendure and Aiden, 2012). Genomic research is currently sequencing increasingly large cohorts of humans (1000 Genomes, UK10K, 100 K Genome, Genomics England 100 K Genomes Project) and non-humans (Snyder *et al*, 2009). Sequencing is pursuing the so-called non-coding aspects of genomes. Prominently, for instance, the ENCODE project (the Encyclopedia of DNA Elements in the Human Genome (Encode Consortium, 2012)) is, despite its name, cataloguing non-coding DNA elements. Finding 'missing heritability' and tracking down elusively rare variations associated with disease or population differences are typical goals of large sequencing projects (as in the "1000 Genomes Project" (Siva, 2008)). At the same time, sequencing has begun to figure in increasingly fluid and diverse problems such as environmental management, public health and biosecurity (for instance, the US Government's Defense Threat Reduction Agency in 2012 was funding work to "Identify Organisms from a Stream of DNA Sequences"). The social and economic value of sequencing practice has remained somewhat open to contestation. At times, sequencing has been linked with the democratisation of science. For instance, in high-profile 'open source' sequencing events, such as the *E. coli* Shiga outbreak in Europe in 2011, public health scientists claimed there was a "propitious confluence of high-throughput genomics, crowd-sourced analyses, and a liberal approach to data release" (Rohde *et al*, 2011, p. 723). Amidst all this sequencing, genomics researchers could justifiably claim that DNA sequencing is "emerging as a ubiquitous, digital 'readout' for the deep, comprehensive exploration of genetics, molecular biology and cellular biophysics" (Shendure and Aiden, 2012, p. 1092). But while "DNA sequences form one of the bedrocks of modern biological science" (Cochrane *et al*, 2012, p. 1), this bedrock of archived and shared base pairs is becoming, as we will see, surprisingly fluid.

The proliferation of sequence data poses some often-mentioned problems in making sequences available. Genomics has very frequently, indeed relentlessly, been described in terms of the increasing quantity of data, from the gigabytes of the 1990s Human Genome Project to the petabytes of contemporary genomics (Cochrane *et al*, 2009). In many discussions, graphics and plots of the falling costs and increasing speed of sequencers loom large. These trends are usually attributed to advances in sequencing technology. Commercial 'next generation sequencers' (NGS) dating from roughly 2006, such as the Roche 454, the Illumina Genome Analyzer or HiSeq 2000 and the Applied Biosystems SOLiD(tm), are prominent in these discussions (see Mardis, 2011 for an often-referred to graph of increasing sequencer output), and they follow on a previous generation of 'high-throughput' sequencers, the so-called 'capillary sequencers'.

With some important exceptions, much of the proliferating DNA and RNA sequence data are publicly available from sequence databases. How does data flow from the high-throughput NGS sequencing platforms to contemporary sequence repositories? It is common, in discussing the expected volumes of sequence data, to compare the increasing rate at which NGS machines sequence DNA to the rate at which computing capacity and digital data storage increases according to Moore's Law (Stevens, 2012). NGS sequencers are accelerating the rate of data production faster than the semi-conductor miniaturisation that underpins Moore's Law. Sequence archives and repositories cannot simply keep up with the flow of data by buying more disk storage. As high-throughput sequencing machines proliferate, augmented by fast desktop sequencers (Loman *et al*, 2012) and soon, perhaps, highly portable 'realtime' sequencers such as Oxford Nanopore Technologies' strangely named 'minION' (Oxford Nanopore MinION sequencing technology), sequence data are very likely to exponentially multiply. The potential mismatch between sequencing and ~~storing sequences~~ was so pronounced in the wake of the advent of NGS machines that new public databases and forms of online repository were established to store publicly accessible NGS data. Most prominently, three 'SRAs' were established to store NGS sequence data: the USA NCBI SRA; EMBL's European Bioinformatics Institute (EBI)~~;~~ European Read Archive (ERA~~,~~ now stabled in the European Nucleotide Archive (ENA)); and the DNA Database of Japan Sequence Read Q5 Archive (DDBJ SRA).[1] These three repositories collaborate as part of the INSDC to constitute a shared NGS sequence archive known collectively as the SRA.

[1] Closely affiliated databases for biomedical and clinical research sequence data include NCBI's dbGAP ('Genotype and Phenotype') and EBI's EGAP ('European Genotype and Phenotype') databases.

The establishment of the SRA does not tell us much about the circulation of the data, its diversity or how it relates to diverse scientific projects. The flat summaries of quantities of data, or speed and cost, that accompany many accounts of NGS platform sequencing say little about how sequences move in and out of this archive, and indeed these aggregate numbers do not differentiate the very different origins and destinations of the sequence data. As Sabina Leonelli has argued, ostensibly vexing problems of an increasing rate of production of sequence data obscure important underlying issues of scale concerned with how widely data infrastructures can be used by different scientific groups (Leonelli, 2013). The analytical question then is whether the SRA itself, as it accepts NGS data from a great variety of different projects, disciplines and styles of scientific work, can divulge anything of how sequence data are produced or how they are used. Can we see in the SRA, as it accepts data from heterogeneous scientific projects and quite different disciplines (agricultural biotechnology, cancer treatment, environmental monitoring, infectious disease control, the evolution of human populations and so on), any signs of the ways in which the archive itself gives rise to the production of sequence data?

Like many data curators, the scientists, bioinformaticians and administrators who design, maintain and curate the SRA face ongoing problems of not only how to cope with the anticipated growth in sequences, but of knowing where they come from and where they will go. ~~T~~he *diversity* of NGS data means that the SRA data curators themselves cannot always comprehend what data their archives house, how they could be used, how much they are used and for what purposes. On the other hand, they have interest in *discoverability* – in how one can locate relevant data – and in making this diversity visible, in as many ways as possible. One important way to do that is through sequence metadata, which acts as a surrogate form of order for the implicit biological order contained in DNA.

The SRA upholds the INSDC policy of accepting all DNA sequences. The collaborating repositories "accept all sequences that submitting scientists present as being relevant and publicly available" (Cochrane *et al*, 2012, p. 2). The SRA, as Table 1 shows, tries to marshal this diversity by using metadata to code submissions according to study types. The variety of sequence-based study types is quite large, and the range of submitting scientists expands all the time. For instance, reporting on recent developments at the EBI ENA, Cochrane *et al* write, "the ever broadening adoption of sequencing as a discovery and assay platform brings the

Table 1: Study types at the SRA

Type	*Number*	*Samples*	*Runs*
Cancer genomics	53	2456	4576
Epigenetics	1582	18150	31136
Exome sequencing	96	6250	8910
Metagenomics	2754	77626	120846
Other	6355	175603	254520
Pooled clone sequencing	32	2892	4047
Population genomics	469	24989	30175
Rnaseq	15	162	162
Synthetic genomics	8	861	1056
Transcriptome analysis	4858	47677	73392
Transcriptome sequencing	2	2	4
Whole genome sequencing	22416	137884	287158

challenge of a user base that is both growing and diversifying” (Cochrane *et al*, 2013, p. 32). As NGS platforms drive a seemingly ineluctable turn to sequencing as the ‘ubiquitous digital readout’ for biological processes, SRA submissions display a diversity that increasingly derives not only from biological differences (between species, for instance) but from the different uses researchers have found for sequencers. Furthermore, genomic science ranges across public, private, government and commercial settings, from small laboratories to global sequencing consortia, and the SRA must countenance a corresponding organisational diversity in the rhythms and rates of sequence submissions. Scientists working at the repositories that constitute the SRA highlight the problems that free and unrestricted deposition of any sequence data poses both to the archives themselves (Leinonen *et al*, 2010; Cochrane *et al*, 2012; Cochrane *et al*, 2013) and to science that depends on the ‘backbone of DNA sequences’ (Nekrutenko and Taylor, 2012).

Post-archival Environments and Sequences on the Move

If we query the EBI’s ENA for all the data relating to a particular NGS study, the resulting list begins something like Table 2, which shows the first few lines of data returned for `ERP000108` (European Read (Archive) Project number 108 might be a rough translation). A project/study typically includes samples (biological materials), experiments (the actual configuration of instruments and assays applied to the sample) and runs (the automatic processes carried out by an NGS machine on the sample). However, as we will see, even a ‘run’, the unit of sequence data that should in principle directly refer to some biological material, is riven in various ways.

This short extract in the table shows that, like other biological databases, metadata saturates the SRA. ~~In an SRA~~ study deposited at the EBI, for instance, there are 129 available fields for metadata (but only 121 at the NCBI SRA), a small sample of which is shown above. Some of these fields point to other databases (such as Entrez, PubMed and so on), some to the locations of the actual sequence data files (stored at ftp – File Transfer Protocol – sites), and others provide details about sequencing platforms, biological samples or the institutions carrying out the research.

In any given study or submission to the SRA, some of these metadata fields are empty, and some are filled. The variable patterns of these accession numbers and the wide ~~span~~ of metadata about sequences not only suggest that the SRA is ~~not~~ evidence of the problems of curating biological information (taxonomic details, sample details and so on), but also that post-archival genomic databases are more like a terminal traversed by many trajectories and

Table 2: Sample SRA study results

study_accession	*secondary_study_accession*	*experiment_accession*	*scientific_name*
PRJEB2054	ERP000108	ERX004046	human gut metagenome
PRJEB2054	ERP000108	ERX004047	human gut metagenome
PRJEB2054	ERP000108	ERX004048	human gut metagenome
PRJEB2054	ERP000108	ERX004049	human gut metagenome
PRJEB2054	ERP000108	ERX004050	human gut metagenome

itineraries in which route numbers (Studies, Projects, Experiments, Runs) matter most. Databases on their way to more or less prominent features on a landscape are shaped by the flow of sequences from NGS machines, by software systems that assemble and tag sequence data, and by scientists, laboratories, institutions and corporations that do things with sequence data in the interests of a wide range of biological questions from biofuels to leukemia. The SRA is clearly not the only sequence repository, and it does not stand in isolation or stand still. It overlaps with earlier iterations of sequence databases (such as NCBI–GenBank, EMBL Bank, DNA Bank; or the DNA Trace Archives, which, starting in 2001, stored the raw sequence data produced by the previous generation of high-throughput capillary sequencing machines), with linked databases such as the database of Single Nucleotide Polymorphisms (dbSNP), ArrayExpress and Gene Expression Omnibus (GEO),[2] meta-databases such as BioProject (that collect information relating to a single research project or consortium from various databases), scientific publication databases such as PubMed, and the many model organism community databases.

The SRA differs markedly from existing biological databases in important respects. First, it is no longer anchored in a single scientific community. Community databases such as WormBase (for *C. elegans*), The Arabidopsis Information Resource (TAIR), *Saccharomyces* Genome Database (SGD) or EcoliWiki have been crucial to biology in recent decades. As Sabina Leonelli and Rachel Ankeny point out,

> [b]y bringing results, people and specimens together using infrastructure, community databases have come to play a crucial role in defining what counts as knowledge of organisms in the post-genomic era. Thus, we argue, they are an integral part of what defines what counts as a 'model organism'.
>
> (Leonelli and Ankeny, 2011, p. 8)

The community databases, especially in their efforts to connect genomic data with other biological information (for example on biochemical pathways, phenotypes, interactions), have become integral to the life of various scientific communities. As Leonelli and Ankeny observe, certain model organisms – yeast, mouse, worm and so on – exist as models only in relation to the accretion of data in such databases. Importantly, the primary database-related practice associated with the community databases continues to be information retrieval based on similarity searches (using the all-import BLAST – Basic Local Alignment Search Tool) and record-linkage, a practice that is simply unwieldy on the huge sequence data in the SRA. Second, the SRA can also be distinguished from what we might call the encyclopaedic biological databases. Since the early 1990s, the bioinformatic mode of practice has relentlessly linked DNA sequences to other data forms using automated retrieval techniques or manual curation of annotations and linkages. The NCBI's GenBank would be a prime exemplar of such encyclopaedic practices. More widely, the extensive, varied and cross-linked informational retrieval systems of contemporary biology – approximately 1900 databases are listed in the annual *Nucleic Acids Research* Database Issue (Galperin *et al,* 2014) – generate increasingly detailed knowledge of genotypes, haplotypes, gene functions, gene networks, signalling, metabolic pathways and the various diseases and traits using record-matching and

[2] GEO and ArrayExpress were meant to store microarray data, but now accept NGS sequence submissions and 'broker' them for the SRA.

linking techniques. Knowledge practice in these settings is, however, largely based on similarity searching and annotation on the basis of similarity. The tremendous accretions of annotated sequence data in GenBank (NCBI), ENA (EBI) and the DDBJ come from approximately 700 000 organisms, and in that respect vastly outnumber the community databases. By and large, these encyclopaedic databases do not store whole genome sequences, only shorter fragments (for instance, a sequence that codes for a particular protein).

Compared to both the purposeful particularity of the community-centred databases, and the expansive universality of the encyclopaedic databases, the SRA metadata is surprisingly generic. The forms of data in, the design of and the modes of access to the SRA differ from ~~the community~~ databases in that the sequencing practices themselves rather than organisms or particular biological entities such as genes (in GenBank) or proteins (in UniProt) begin to organise the archive. In the SRA, sequencing practices "begin to interweave themselves with elements of the formal infrastructure to create a unique and evolving hybrid" (Star and Ruhleder, 1996, p. 132), in which sequence data no longer appear as something to be accessed, read, browsed or compared but as data to be shifted *en masse*, constituting a shift to what we are calling post-archival genomics. At the same time, as we will see, distance, dispersal and discontinuities also appear in the SRA. As different communities, contexts and projects intersect in the SRA, the resulting tensions cannot be resolved in the archives.[3]

In comparison to the relatively short annotated gene sequences found in GenBank or EMBL Bank, whole genome sequence data produced by NGS machines are difficult to browse, and whole genome alignments are almost undisplayable. Post-archival databases, and the SRA in particular, store data multiples that cannot be easily displayed in a web browser. Despite ongoing visualisation efforts, displaying a whole genome sequence rarely shows much of biological interest. Slight variations in sequence data are frustratingly subtle and hard to see, so that the potential for DNA sequences to be the 'ubiquitous digital readout' for biology remains tantalizingly close yet not fully actual. Rather than a public repository of scientific data, the SRA functions as a 'post-archival' environment in which negotiations and encounters between NGS platforms, rapid transformations in database and network devices largely driven by commerce, and the burgeoning investments in sequencing and sequence data, play out. The new problem faced by the SRA is that the actual bedrock scientific data – the DNA sequences – are not themselves easily stored in databases. While earlier DNA sequence databases such as GenBank could accommodate the sequence data alongside metadata about species, biological samples, associated scientific publications and so on, in the SRA the very large sequence data files produced by NGS machines reside outside the database in

[3] How would one research the dispersed, multiple, potential interactions of a post-archival database? Understanding post-archival database dynamics entails suspending the notion of community that underpins scientific community databases. The latter have often been studied using ethnographic techniques as well as interviews with scientists and database managers (Leonelli and Ankeny, 2011; Nadim, 2012). We propose that treating the database itself as an informant allows more of the scale-varying interactions associated with contemporary sequence circulation to come to light. In order to communicate with databases as informants, we re-purpose the tools for accessing data as instruments for reading databases. This entails some technical work in the form of programming in order to run queries against the databases. Data in the many EBI databases, which include EMBL Bank, the SRA, the Trace Archive and ArrayExpress among others, can be browsed using standard web browsers. Web browser interfaces are familiar in genomics, as in almost any other scientific field, as ways of finding data, and more specifically, of looking at alignments between different sequences.

compressed files. The files are characteristically large (sometimes hundreds of gigabytes) and putatively monolithic (that is, comprising 'G', 'A', 'T' and 'C' encoded in various ways; but as we show below, shifts in sequence file formats are a source of instability). The SRA is more like a logistics or supply chain management system. It largely manages movements of containerised datasets that it itself does not look into (or indeed, afford others to look into either). At the same time, there would be obviously little point in assiduously cramming sequence data into warehouses that were impossible to search ~~for~~ or difficult to retrieve. To this end, the SRA has gradually elaborated a hierarchy of metadata objects and submission protocols that accommodate the different scales and aggregations of sequence data being produced. At the base of this hierarchy stands the 'run' that represents the product of an NGS platform, and points to a single sequence data file. At the peak of the hierarchy stands the 'study' or 'project' (the terminology is not uniform across the United States and EU SRA partners). The SRA presents a fairly hierarchical view of genomic research, in which an NGS project is characterised by metadata that describe ~~study,~~ experiments and samples (tissues, organisms and so on), and then by a series of 'runs' on sequencers (from one to tens of thousands of runs) submitted as sequence data files in various formats.

The SRA, like nearly all other biological databases today, offers 'programmatic access' modes that allow database users to write software or scripts that search, retrieve or submit multiple datasets that can then be analysed by more sophisticated bioinformatic, statistical and, increasingly, machine-learning techniques (Mackenzie, 2014). APIs allow programmatic access to a meta-database or a stable of databases containing sequence data and metadata. APIs depend on various generic software protocols (REST – representational state transfer; SOAP – simple object access protocol) and data formats (XML, JSON, csv and so on), themselves largely developed in the course of the last decade as Web 2.0 has developed. [4] APIs like EBI's 'ENA-Browser' or the NCBI's 'e-utilities' make available a huge range of data through a single point of access that responds to many different commands or invocations. As in other domains of contemporary communication and media, APIs profoundly affect the ways in which databases, data, instruments and people relate to each other. They open new pathways and connections between widely dispersed patches of practice, and enable the coordinated, individualised, targeted, real-time and predictive behaviours we increasingly, for better or worse, expect of contemporary media. Importantly for our purposes, the material practices of post-archival genomics can be mapped through these APIs because the APIs are gateways through which nearly all contemporary DNA sequence data increasingly move. The APIs, put bluntly, make post-archival genomics possible.[5] Through the APIs, the coherence and apparent unity of databases as repositories begin to change. They begin to appear differently, as much more layered and sedimented architectures aligned to promote the movement of sequences. In many instances, the lines of code that retrieve sequences or

[4] In the life sciences themselves, there are many examples of APIs; see BioCatalog – the Life Sciences Web Registry for a list of these.

[5] We worked with NCBI's 'e-utilities' and EBI's 'ENA-Browser' in querying the SRA. Neither visually displays annotated genomes like the UCSC Genome Browser or ENSEMBL. Methodologically speaking, APIs open some novel paths for social scientists to explore. Through them, claims and concerns about the production and use of sequence data can be explored empirically. Although ethnographies of genomic research, interviews with scientists, bioinformaticians and archive managers, attending meetings, workshops and conferences, and reading scientific literature and online discussions remain critically important, we focus here on the APIs as ways to explore the topography of DNA sequences as represented in the SRA.

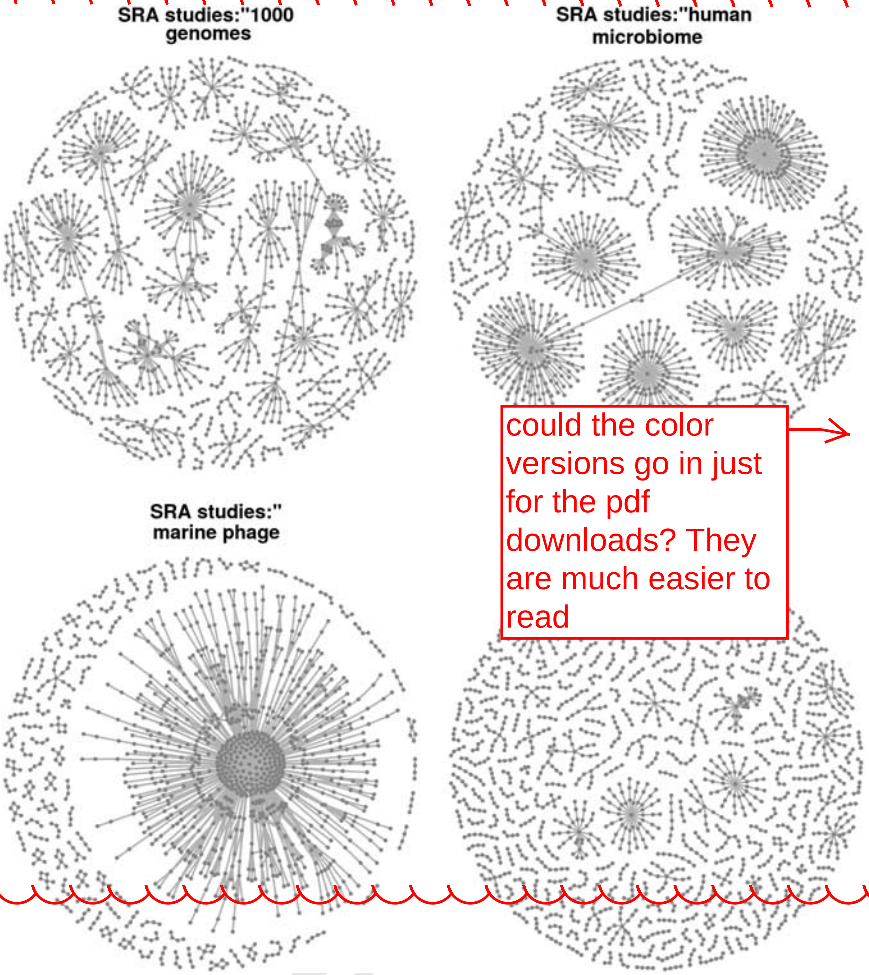
metadata about sequences from the databases are relatively brief, but the APIs open onto the dozens of databases curated by the NCBI, EBI and DDBJ. Like so many other databases today, the SRA invites programmatic access because the data are shaped by format, by size or by a multiplicity of components that are difficult to render or work with in human readable forms. For our purposes, both the character of the APIs and the flows of sequence-related data they emit also attest to the forms of dispersal that link local communities of practice into a broader post-archival genomics.

To illustrate the shift to an API-based logistics of sequence data associated with the SRA, we could compare the now classic molecular biology-style Genbank/EMBL Bank records that can be viewed on the NCBI GenBank website with the data programmatically accessed via APIs in a programming language such Python, Perl or, in this case, R:

```
  library(‘RCurl’)
  #retrieve EMBL Bank record with accession number A00145
  embl_url <-‘http://www.ebi.ac.uk/ena/data/view/A00145&display=fasta’
  embl_fasta <-RCurl::getURLContent(embl_url)
  print(embl_fasta)
  [1] “>ENA|A00145|A00145.1 B.taurus BoIFN-alpha A mRNA \nCTGAAC
TTCAGAGAACCTAGAGAGCAGGTTCACAGAGTCACCCACCTCACCAG\
GCATCTGCAAGGTCCCCGATGGCCCCAGCCTGGTCCTTCCTGCTATCC
\nCTGCTCAGCTGCAACGCCATCTGCTCTCTGGGTTGCCACCTGCCTCACACCCA-
CAGCCTG\nGCCAACAGGAGGGTCCTGATGCTCCTGCAACAACTGAGAAGGGTC
TCCCCTTCCTCCTGC\nCTGCAGGACAGAAATGACTTCGAATTCCTCCAGGAGGC
TCTGGGTGGCAGCCAGTTGCAG\nAAGGCTCAAGCCATCTCTGTGCTCCACGAGG
TGACCCAGCACACCTTCCAGCTCTTCAGC\nACAGAGGGCTCGCCCGCCACGTGG-
GACAAGAGCCTCCTGGACAAGCTACGCGCTGCGCTG\nGATCAGCAGCTCACT-
GACCTGCAAGCCTGTCTGACGCAGGAGGAGGGGCTGCGAGGGGCT\nCCCCTGC
TCAAGGAGGACTCCAGCCTGGCTGTGAGGAAATACTTCCACAGACTCACTCTC
\nTATCTGCAAGAGAAGAGACACAGCCCTTGTGCCTGGGAGGTTGTCAGAGCAGA
AGTCATG\nAGAGCCTTCTCTTCCTCAACAAACTTGCAGGAGAGTTTCAGGAGAAA
GGACTGACACACA\nCCTGGTCCAACACGGAAA\n”
  attr(, “Content-Type”)
  charset
  “text/plain” “UTF-8”
```

I'd expect more spacing around this code excerpt. Also the font should be san serif

In molecular biology databases dating from 1980 to 1990, each accession number identifies, at least in principle, a unique biological entity, such as a molecule or a unique DNA sequence. In this example, we see the proximity of the sequence data, their accession number, indications of biological species, and biochemical and biomolecular classifications. This data format reflects the somewhat ‘flat world’ character of molecular biology (Rose, 2006, p. 15) in the sense that the biological data and their metadata (the species *B. taurus*, and the GenBank accession number) stand next to each other. As we saw in Table 2, in the SRA and post-archival genomics, matters are somewhat different. In the SRA, accession numbers point to biological materials such as samples, to particular platforms and instruments for sequencing, to runs containing sequence data files, to epistemic processes such as experiments and studies, and to a range of other biological databases. More importantly, much of the metadata, as represented by accession numbers, connect sequence data to *particular* NGS machines,


SRA studies:"1000 genomes
SRA studies:"human microbiome
SRA studies:" marine phage

could the color versions go in just for the pdf downloads? They are much easier to read

Q12 **Figure 1:**

genomics centres and research consortia scattered across different fields in the biomedical and life sciences. In other words, the SRA is less an archive of sequences and associated biological knowledge than a multi-scale map of the concentrated and dispersed locations of DNA sequencing.

In post-archival genomics, for instance, experiments, biological samples, individual NGS machines and organisations group together differently in different genomic fields.

Figure 1 uses accession numbers for SRA studies, experiments, samples and runs for the 1000 Genomes project, human microbiome, marine phage and *E. coli*.[6] Each of these genomic localities has its own forms of sequencing practice that can be glimpsed by following the links between accession numbers in that locality. For instance, a study accession such as ERP000123 will be associated with some experiment accessions – ERX000456 – and some

[6] All data graphics in this article were generated by the authors using either the EBI ENA Browser http://www.ebi.ac.uk/ena/browse/programmatic-access or the NCBI SRAdb data release (Zhu *et al*, 2013).

run accessions – ERR000890. The network graphs in the figure display the accession numbers as nodes in order to indicate something of the diverse network structures comprising post-archival genomics. The different study types vary greatly in their topology and in their modes of producing sequence data. The ‘1000 Genomes’ network, a large international consortial project to characterise genetic variation in the global human population, is clustered around a number of interconnected studies that display what we term *verticality* because they deposit huge sets of sequence data in the archive. The bulk of ‘marine phage’ runs are associated with a single sequencing centre, the JCVI, that sequenced DNA found during the Craig Venter’s ocean sampling expeditions. By contrast, sequences stemming from the workhorse model organisms of contemporary biology, such as the bacteria *E. coli*, are dispersed across many small disparate sequencing studies or projects of greatly varying size.

The Same Data in Different Places: Dispersal

Despite the differences in architecture and institutional ethos at the NCBI, EBI and DDBJ sequence archives, genomics research is predicated on the principle that the same accession number works in all three instances of the SRA. It is this understood commonality or universality that allows genomics researchers to just invoke ‘the SRA’, rather than specifying exactly which database they mean. The sharing of accession numbers is a key mechanism in making genomic data available, and enables the circulation of sequence data. From its inception, INSDC set out to make the same sequence data available globally. The official doctrine is that sequence data are mirrored or synchronised constantly between the three partner databases that constitute the SRA. As the Japanese DDBJ puts it,

> Since we exchange the collected data with EMBL-Bank/EBI; European Bioinformatics Institute and GenBank/NCBI~~;~~ National Center for Biotechnology Information on a daily basis, the three data banks share virtually the same data at any given time. The virtually unified database is called “INSD; International Nucleotide Sequence Database”.
>
> (DDBJ, 2013)

What does it mean to ‘share virtually the same data at any given time’? Does it mean that the three instances of the SRA will be exactly the same? In principle that would be difficult to achieve, as the EBI, DDBJ and NCBI run separate websites, servers, databases and network infrastructures. Although the same data might be in all of them, separating out what belongs to the sequence data and what belongs to the actual database itself might not be so simple. ‘Sharing virtually the same data’, especially when the doubling time is faster than the 18 months of Moore’s Law, might entail certain complications. We asked a coordinator at the EBI’s ENA about the relation between the different read archives:

> *Interviewer:* Are ERA [EBI European Read Archive] and [NCBI] SRA the same or different things?
> *Interviewee:* A brand name was introduced before the Sequence Read Archive collaboration was formally set up. That name [‘European Read Archive’] is no longer in active use. We just call ourselves the ‘Sequence Read Archive’ jointly together with

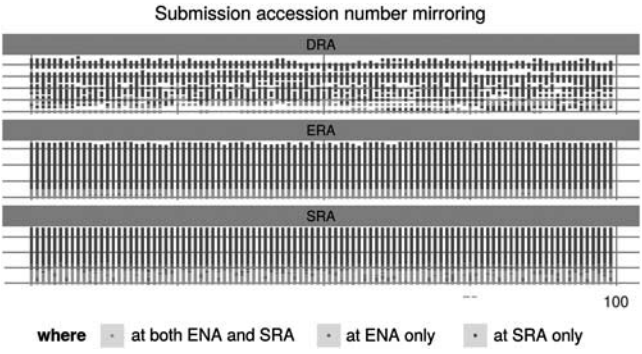
> NCBI and DDBJ. And what we do every day, we have a data and metadata exchange running constantly. We have a multi-terabyte daily data exchange flow in both directions. Other metadata is being exchanged as well.

The statement “We just call ourselves ‘the Sequence Read Archive’” suggests that differences between the three major instances of the read archives are immaterial. The claims about ‘multi-terabyte daily data exchange’ attest to the work done to synchronise the archives.

In February 2011, genomics blogs, Twitter and science media were alive with discussion about the imminent demise of the NCBI’s SRA. Although massive amounts of sequence data were being regularly produced by scientists equipped with the new sequencing machines, the US National Library of Medicine-funded NCBI announced that its SRA would be closing. A literally ‘post-archival’ time seemed imminent. On various blogs scattered across the life sciences, scientists reacted to this news with a mixture of incredulity, satisfaction and concern. On the one hand, the NCBI SRA is a pillar of INSDC, and ostensibly crucial to the ongoing public availability of sequence data. On the other hand, as many scientists observed in the lengthy discussion that followed on forums and blogs such as SEQanswers (a popular sequencing discussion site), the NCBI SRA is difficult to submit data to and difficult to retrieve data from (SEQanswers, 2011). *Nature News Blog* headlined the announcement “Unpopular genomics database faces budget axe” (Callaway, 2011). (Ever ready to organise the world’s information, Google Corporation offered to host NGS sequence data and began to act as a new, *de facto* INSDC partner.)

Nine months later, in October 2011, the NCBI announced that the US SRA would remain open (NCBI, 2011). What had changed? US Federal Government budget measures are not the topic of this article, so we leave aside the budgetary crisis politics attached to these events, and read the NCBI SRA crisis of 2011 in terms of a broader transformation in the global circulation of sequence data. As we saw in Table 1, the variety of NGS study types in the SRA reflects some of the manifold ways in which sequencing is used. Although all the sequence data in the SRA come from one of just a few different kinds of sequencing platforms, different sequencing practices shape the flow of sequence data very differently. By virtue of the great variety of biological questions channelled through DNA sequencing, genomics research projects produce widely disparate datasets. Numbers of runs, numbers of machines, numbers of samples, numbers of experiments, numbers of centres involved and so on fluctuate widely from study to study, from biological community to community (as illustrated in Figure 1). The abundant whole genome sequencing projects, for instance, tend to have a small number of samples or machine runs compared to cancer genomics or population genomics studies. But small numbers of samples might be subject to much more extensive sequencing in a ‘whole genome sequencing’ compared to an ‘exome’ study. These differences trouble the neatly nested hierarchical record structure that runs from the overarching study down to the runs of an NGS machine. Further organisational complications such as the number of centres involved, the number of submissions comprising a study, the timing of submissions, and the different types of runs, experiments and biological samples present in the studies only add to the very uneven profile of sequence data. Even submitting sequence data to and retrieving sequence data from the SRA is not an ‘atomic’ action. That is, it may be a many-staged, multi-part process.

These kinds of practical difficulties made the NCBI’s SRA archive unpopular in 2011. More importantly, they suggest that the archival ambition to ensure that all data are universally and



again, could the colour versions go in for the online versions of the article? they are much more legible.

Figure 2:

identically available from all INSDC partners may encounter great difficulties. We can empirically observe signs of these difficulties in the patterns of accession numbers in different instances of the SRA. Every accession number more or less explicitly encodes something about time and place. Regarding place, INSDC accession numbers indicate to which archive data were submitted. Accession numbers for submissions to the NCBI, EBI and DDBJ begin with ‘S’, ‘E’ and ‘D’, respectively. Furthermore, as accession numbers are assigned consecutively, we know that the study ERP000345 was submitted before the study ERP000777. In terms of the post-archival sequence data flow, we can readily see that despite its unpopularity, the NCBI SRA attracted much of the flow of NGS data. Of 42 333 studies in the SRA, only 1605 were submitted to the DDBJ, 3763 to the EBI, and 36 965 to the NCBI. This quite uneven global distribution of submissions perhaps suggests that most sequencing is done in North America (but then this does not take into account sequences that are not deposited). We know from James Hadfield and Nick Loman’s “Next Generation Genomics: World Map of High-Throughput Sequencers” that many of the several thousand NGS machines are in the USA.[7] Of the 733 centres shown on this map, 264 are in the USA, and only 12 in China. No doubt, centres vary greatly in size. On this map, which relies on crowd-sourced data, most of the centres list one or two NGS machines, but several report many more (for instance, the BGI, formerly the Beijing Genomics Institute, lists more than 100).

But these geographical and institutional variations, which we would expect given the pre-eminence of the USA in the life sciences, should be irrelevant to the data actually held in each of the instances of the SRA. Despite the very different levels of submissions (and presumably downloads) of data, the contents of the three archives should mirror each other. One should see the same data in the three different instances.

It seems not. Submission accession numbers are generated every time data are submitted to the SRA. A submission might refer to a single run or a whole research project. Although this variation in reference is somewhat confusing, submission accession numbers allow different patterns of deposition to be traced. We generated a list of all the *submission* accession numbers of the form DRA000001, DRA0000002, ERA000001, ERA000002, SRA0000001, SRA000002 held by the NCBI SRA. We then wrote a script that queried the EBI ENA Browser API for all the submissions

[7] http://omicsmaps.com/

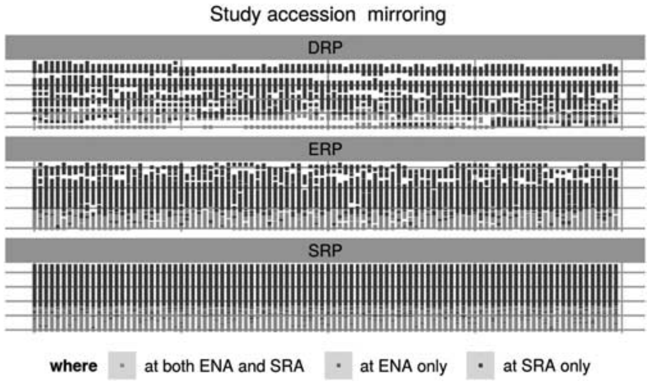


Figure 3:

with those accession numbers (see Figure 2). (This involved incrementing the accession numbers until the API returned no results.) The ENA returned 70 795 submission accession numbers. We compared all the submission metadata gathered in this way from ENA to the list of submission accession numbers published by the NCBI in their periodic release of SRA metadata in the form of the 'SRAdb' R Bioconductor package (Zhu *et al*, 2013). Figure 2 shows that there were substantial discrepancies between the submissions stored at EBI ENA and those at NCBI SRA. Data about submissions are clearly not the same. While there were 70 538 at both NCBI SRA and EBI ENA, there were 257 and 193 678 that were only at EBI ENA and NCBI SRA, respectively. This disparity between the pattern of submission at EBI ENA and NCBI SRA does not mean that the INSDC vision of a common sequence data archive is broken, only that the vision of the same data in all places no longer comprehensively organises the circulation of sequence data.

Perhaps this disparity is peculiar to submission practices? Submissions, especially for large consortial projects, occur according to different rhythms that are difficult to synchronise between the different instances of the SRA. Yet similar discrepancies can be seen in the accession numbers for other kinds of SRA metadata objects. One might expect the accession numbers for studies to be more stable, as studies are key anchor points for the collections of sequences, samples and experiments in a given genomic research project. Using the same procedure as before, we generated a list of *study* accession numbers from the NCBI SRA, queried the EBI ENA using its API and checked how the two lists aligned.

As Figure 3 shows, this revealed a similar pattern of differences in the alignments. Instead of a shared pattern of accession numbers running across the three instances of the SRA, many study accession numbers are only found at the NCBI instance and some are only found at the EBI instance. The lists of study accession numbers at the NCBI SRA and the EBI ENA are different. Of the 42 348 total number of studies, only 14 448 are at both archives. This is a striking finding as it goes to the heart of the post-archival genomics. While all the sequence data may be fully shared and mirrored between the archives, if the organisation of those data as part of studies differs, then local genomic communities effectively inhabit different worlds. Post-archival genomics is more dispersed than it believes. It is as if a given currency had different values in different places, despite the existence of international exchange rate mechanisms meant to establish universal exchange rates.

Post-archival Frictions

What occasions these misalignments? We see them less as organisational errors, infrastructural failure or ongoing epistemological tensions in biology (Leonelli, 2014) and more as signs Q6 of the divergent patterns of circulation of sequence data in post-archival genomics. Recent ~~STS work~~ has described how 'social distance' affects the movement of data, and occasions the production of metadata. Metadata such as accession numbers have been a key focus of attention in this analysis. In their description of metadata frictions, Edwards *et al* write:

> Metadata products are supposed to substitute for direct contact with data producers – and they can do that, to a greater or lesser degree, in many contexts. Yet in very many cases, metadata products remain incomplete, ambiguous, or corrupted … . When this happens, the conversation about data cannot continue without repair. Such repair can, and often does, include direct communication with the data creators: metadata-as-process. As with ordinary conversations, the greater the social distance between the disciplines of data creators, the more metadata-as-process is likely to be needed.
>
> (Edwards *et al*, 2011, p. 684)

In the archival genomics of GenBank or the many genome browsing databases developed during the 1990s and early 2000s, metadata such as the study and submission accession numbers were vital to 'conversation about' DNA (or RNA). The metadata problems Edwards describes – incompleteness, ambiguity and brokenness – were common in data infrastructures where distances, dispersal, patchiness and heterogeneity occur – but they were also the object of much attention and interest. These problems sometimes occasioned repair work, and sometimes whole new standards, operating practices, infrastructures and institutions. Edwards suggests that "social distance between data creators" necessitates the development of other forms of metadata, "metadata-as-process", that address the mismatch or gaps in the metadata.

We would suggest, by contrast, that post-archival genomics is less concerned with these frictions. Indeed, these frictions might even be tolerated or ignored as long as they do not impede the movement of sequences in and out of the databases. This might explain the relatively high proportions of missing metadata in the SRA. Patterns of metadata missingness attest to different forms of sequence data movement rather than epistemic negligence or professional frictions in scientific communities. We counted the proportion of missing data in all the metadata fields relating to studies, experiments, samples and runs in the periodic NCBI SRA metadata release (Zhu *et al*, 2013). Of the 121 fields, only 32 are close to fully populated (see Figure 4). (Figure 4: "Missing metadata in the SRA" does not include metadata fields that are close to 100 per cent empty. We assume that those fields are not relevant to the table in question.) We would expect that every NGS study has at least one 'run' associated with it, and in fact they do. But we might also expect every NGS study to be accompanied by some kind of description of what the study is about, and yet a significant proportion (78 per cent) do not. Or we might think that it would be very important to know what platform (Illumina, Roche, Pacific Biosciences, Life Technologies and so on) produced the sequences, but around 5 per cent of the submissions in the archives show no data in that field. While the SRA is at no risk of collapse owing to incomplete metadata, endemically missing or ambiguous metadata again points to different movements

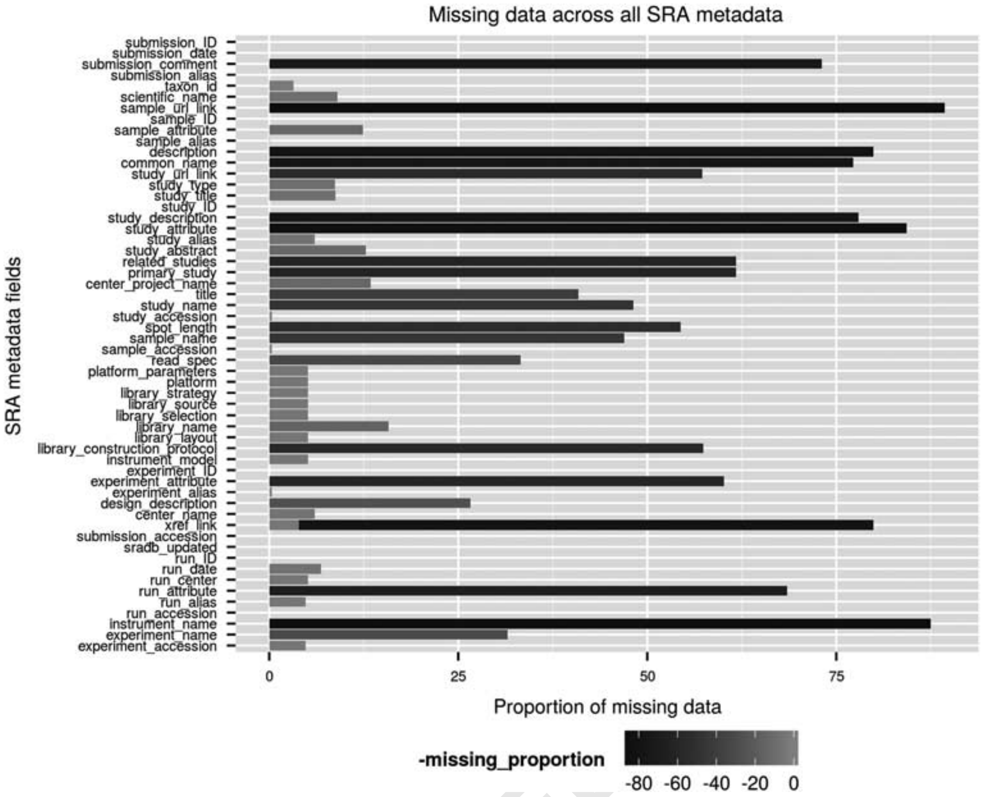


Figure 4:

and routes followed by sequence data. We could ask whether certain fields of metadata tend to be missing in association with certain study types, or in relation to the size of projects, or is dependent on the route through which sequence data come into the SRA (of around 30 000 submissions, several thousand come via the GEO portal, for instance). In other words, we might regard missing metadata as evidence of sequence supply chain logistics. The missingness of metadata is a pattern created as data move into the archives following some paths or channels, and not others.

The Slipperiness of the Run: From Machine to Functions

Even if different patterns of metadata incompleteness point to different ecologies of genomic research practice and different modes of archiving sequences within the same database, do the sequence data themselves not still remain as the stable reference to biological samples? Is this level of the SRA not stable, even if all the organisational coordination is much more fluid? Sequence data in the SRA are contained in runs. That is, the only way to access the actual sequence data, themselves stored in sequence files, is through runs and run accession numbers (such as ERR000585). Run accession numbers, with their associated metadata ('run date',

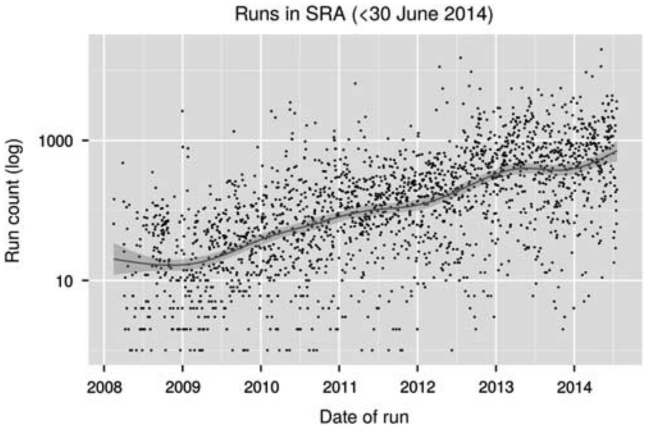


Figure 5:

'platform', 'model', and so on), imply that a run is a coherent or meaningful unit of sequencing practice. That is, a run will contain the DNA sequence produced when a machine sequences a biological sample. But viewed in terms of how they change over time, runs start to lead a more complicated life in the SRA. While, from the perspective of the reported SRA metadata, runs endure as a category over time, it may be that run practices actually change as different kinds of sequencing practice come and go. One indication of this comes from the patterns of runs deposited over time in the SRA. Although they showed massive growth in the years 2008–2011, during 2012 there is a decline in the run rate. Figure 5 plots the number of runs submitted to the SRA each week starting from 2005. In some ways this curve reflects the increasing use of NGS sequencers in genomic research. In other respects, it looks anomalous. In 2012, for instance, the overall rate of runs seems to be declining, even though there are many more NGS machines, and the newer models complete runs in a shorter time.

What could account for these apparent ups and downs in the number of sequencer runs appearing in the SRA? One possible explanation is a substantial transformation in the meaning of a 'run'. This change can be seen in the files associated with the runs. We downloaded a sample of 38 114 sequence files from the EBI ENA to examine how file formats manipulated the actual sequence data in relation to runs. As Figure 6, "NGS file formats", shows, the Binary Sequence Alignment (BAM) file format prevails in recent years (24 015 of the files are BAM format.) The BAM Map file is now the main file format used to store NGS sequence data. If this file format dominates at least at the EBI ENA, what does it tell us about how sequence data are changing shape? Some of the formats shown here, such as csfasta or sff, are proprietary sequence formats associated with particular sequencing platforms (Applied Biosystems' SOLiD(tm) and Roche 454 sequencers, respectively); others, such as srf (Sequence Read Format), fastq and BAM , have developed out of community efforts to standardise sequence file formats (Li *et al*, 2009; Cock *et al*, 2010). While sequence file formats are no more diverse than, say, digital image file formats (jpeg, png, tiff, bmp, and so on), their shifting patterns of usage, like the missing metadata, suggest that runs are changing in post-archival genomics.

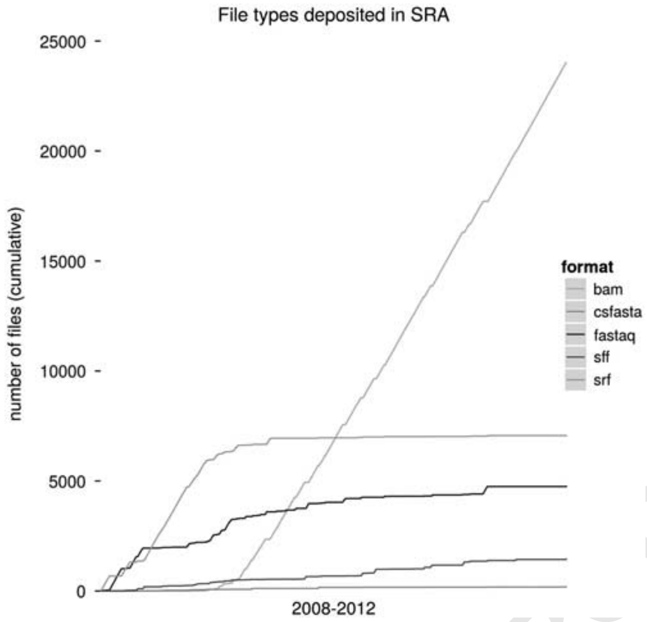


Figure 6:

Given the steep growth of the BAM file format, and trying to find out how their proliferation related to NGS instrument practice, we asked an EBI ENA database administrator how BAM files relate to runs:

> Interviewer: Can you say one run equals one BAM file?
> Interviewee: That's exactly the case for us. You cannot talk of more than one BAM file in a single run. So one possible functional definition of a run is a set of data which can be easily processed together. For BAM it means a single BAM file.

Note that the interviewee does not simply equate a run to an NGS machine cycle. Although a 'run' means a single BAM file, a run is a "set of data which can be easily processed together". Runs, he says, have a "functional definition". The atomicity of a run as the bedrock practice of contemporary genomics comes into question here. This functional openness of a run means that the bedrock of DNA sequences becomes much more fluid in the SRA. File formats, for instance, reflect compromises between sequencing platforms, storage systems, network and bandwidth costs, proprietary instrument-related software, and various analytical pipelines. The BAM file, our interviewer tells us, is "a file format that can act in different roles". If this is the case, then, again, DNA sequence data inhabit a fluid environment where different species of sequencing practice coexist.

Conclusion

> Most scientific techniques, it can be argued, are in fact nothing more than methods for moving things around and changing the relations among objects.
>
> (Canguilhem, 2000, p. 319)

George Canguilhem's suggestion that scientific techniques be seen as methods of movement aptly describes many of the changes associated with recent genome sequencing. Many genomic techniques and data infrastructures move things around, and focus on how to move things around more. This logistic imperative reaches deep into genomics. In December 2012, a new submission to the SRA appeared entitled "ERP002040: storage of 5 computer files (739 kB) by coding into synthetic DNA oligos, and recovery of original information via high-throughput sequencing". A Letter in *Nature* in January 2013 (Goldman *et al*, 2013) describes how the authors, including Ewan Birney, Associate Director of the EBI, along with co-authors from the scientific instrument company Agilent Technologies, used DNA to encode computer files (Shakespeare sonnets and so on), and then retrieved that information with 100 per cent accuracy by sequencing it. They describe the implications of their work for digital archiving:

> theoretical analysis indicates that our DNA-based storage scheme could be scaled far beyond current global information volumes and offers a realistic technology for large-scale, long-term and infrequently accessed digital archiving. In fact, current trends in technological advances are reducing DNA synthesis costs at a pace that should make our scheme cost-effective for sub-50-year archiving within a decade.
>
> (Goldman *et al*, 2013, p. 1)

In ERP002040, DNA sequences appear in a DNA sequence archive as a solution to the problem of "global information volumes". As we have seen, the individual instances of the SRA embrace a rather brutal, sometimes faster-than-Moore's Law growth in sequence size, and explicitly position themselves to receive, organise and render available sequence data for many different purposes. DNA storage, with its potential to scale "far beyond current global information volumes", encapsulates the perfectly recursive solution for archiving all the sequence data being produced by NGS machines and their inevitable successors. Ironically, then, post-archival genomics would not only treat DNA as the biological bedrock of their investigations, but would use DNA to deal with the problems of archiving. DNA sequences would store DNA sequences, and accessing DNA sequences would involve sequencing DNA in order to extract sequences. The patterns of the circulating sequences would materialise as DNA sequences. DNA sequences would not only be a 'ubiquitous readout for biology', but a store of all 'global information'.

Whether or not such loopy recursiveness – storing DNA sequences in DNA – becomes common practice or not, this development highlights the increasingly bulk mode of existence of DNA sequences in general. Although they may be characterised as bedrock or backbones for genomic research, DNA sequences are more like an expanding foam that encompasses a widening array of biological projects ranging from medicine to the environment, from biosecurity to information storage. The SRA attempts to manage the burgeoning expectations associated with sequencing by accepting sequence data and "moving it around", as Canguilhem puts it. As we have seen in our traversals of the SRA, however, its attempts to open sequence data to access and discovery differ from existing biological data infrastructures in that it no longer presents sequence data themselves for exploration. In post-archival genomics, the mobility of large aggregates of sequence data supplants the exploration of the data themselves.

The post-archival tendencies of the SRA show themselves in several ways. The existence of the SRA as a collective entity rests on the movement of sequence data between Europe, North

America and Asia. This copying or mirroring is unstable at every level apart from the sequence data files themselves. The SRA exists as a repository for NGS platform sequence data, but the proliferating variety of sequencing techniques, and their expanding domains of application, means that connecting those data to specific localities and fields of research becomes increasingly complicated. Genomics predicates sequence comparison and biological contextualisation via metadata as core practices, yet, as we saw, the mass of missing metadata is less a defect of the SRA (as it might be seen in a more conventional database) and more an artefact of the multiple trajectories passing through the SRA. Finally, we saw that even the apparent stability of DNA sequence data begins to mutate in the SRA. The 'run', for instance, has taken on functional meanings that can be seen in the changing composition of sequence file formats and the sometimes surprising changes in the run rates.

Reflecting on the prominence of discussions of scale in data infrastructures for biology, Sabina Leonelli writes:

> the scale of data infrastructures can be measured through the range and scope of biological questions that data stored therein can be used to address – where range indicates the number of research areas and specific queries potentially served by the database and scope indicates the types of organisms whose study can thus be fostered.
>
> (Leonelli, 2013, p. 461)

We would agree that discussion of scale in genomics often draws attention away from the localised frictions involved in doing research. While many of the features and dynamics of the SRA resonate with this re-definition of scale in terms of range of research or scope of biological problems, sequence data in the SRA also seem to overflow this description in various ways. The range of organisms and the range of organisations present in the SRA is huge. The range and scope of biological questions that might be addressed by the large sequencing projects or the many small ones is very open-ended in the SRA. Yet the SRA largely eschews 'specific queries' in favour of a verticality or massive depth of sequence data whose real value comes from the logistical power to mobilise units of data such as BAM files, and to marshal those units into analytical pipelines (for instance, via the APIs we used to query the SRA). Rather than scale, scaleability matters here.

How we understand DNA as an economic, epistemic or ontological form today depends on how we make sense of the techniques that make it and move it around as sequence data. The altered movements of DNA sequence data attest to a reorganisation of biological knowledge whose ramifications for biotechnology, medicine, health, the environment, agriculture and energy are still in development, but will rely on logistically scaled Q7 management of sequence data (see Mackenzie, 2014) for further exploration of these movements). Emerging forms of data structuring such as the SRA can be useful here. As well as reading scientific literature, and interviewing scientists and technical professionals, we suggest that querying such databases via their various programmatic interfaces moves DNA differently. This logistical mode of exploration depends heavily on equipment, tools and affordances developed by genomics researchers themselves, moderately repurposed and tweaked in various ways to construct views of the flow of sequence data during the last half decade or so.

About the Author

Adrian Mackenzie (Professor of Technological Cultures, Department of Sociology, Lancaster University) has published work on technology: *Transductions: Bodies and Machines at Speed*, (2002/2006), *Cutting Code: Software and Sociality* (2006) and *Wirelessness: Radical Empiricism in Network Cultures* (2010). He is currently working on the circulation of data-intensive methods across science, government and business in network media. He co-directs the Centre for Science Studies, Lancaster University, UK.

Q4

References

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

Busch, L. (2007) Performing the economy, performing science: From neoclassical to supply chain models in the agrifood sector.. *Economy and Society* 36(3): 437–466.

Callaway, E. (2011) Unpopular genomic database faces budget axe: Nature News Blog. http://blogs.nature.com/news/2011/02/database_cuts.html, accessed 29 January 2013.

Canguilhem, G. (2000) In: Delaporte (ed.) *A Vital Rationalist: Selected Writings from Georges Canguilhem*. New York: Zone Books.

Chow-White, P.A. and Garcia-Sancho, M. (2011) Bidirectional shaping and spaces of convergence: Interactions between biology and computing from the first DNA sequencers to global genome databases', science, technology & human values. http://sth.sagepub.com/content/early/2011/02/26/0162243910397969.abstract , accessed 1 September 2011.

Cochrane, G. *et al* (2009) Petabyte-scale innovations at the European nucleotide archive. *Nucleic Acids Research* 37(1): D19–D25.

Cochrane, G. *et al* (2013) Facing growth in the European nucleotide archive. *Nucleic Acids Research* 41(D1): D30–D35.

Cochrane, G., Cook, C. E. and Birney, E. (2012) The future of DNA sequence archiving. *GigaScience* 1(1): 2.

Cock, P.J.A. *et al* (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* 38(6): 1767–1771.

DDBJ (2013) Introduction of DDBJ | DDBJ. Available at <http://www.ddbj.nig.ac.jp/intro-e.html>, accessed 25 January 2013.

Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C. and Borgman, C.L. (2011) Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41(5): 667–690.

Encode Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414): 57–74, Available at <http://www.nature.com/nature/journal/v489/n7414/full/nature11247.html>, accessed 10 September 2012.

Galperin, M.Y., Rigden, D.J. and Fernandez-Suarez, X.M. (2014) The 2015 nucleic acids research database issue and molecular biology database collection. *Nucleic Acids Research* 43(D1): D1–D5.

Q8 Goldman, N. *et al* (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, <http://www.nature.com.ezproxy.lancs.ac.uk/nature/journal/vaop/ncurrent/full/nature11875.html>, accessed 29 January 2013.

Helmreich, S. (2009) *Alien Ocean: Anthropological Voyages in Microbial Seas*. Berkeley, CA: University of California Press.

Helmreich, S. (2008) Species of biocapital. *Science as Culture* 17(4): 463–478.

Hilgartner, S. (1995) Biomolecular databases new communication regimes for biology? *Science Communication* 17(2): 240–263.

Hilgartner, S. (2013) Constituting large-scale biology: Building a regime of governance in the early years of the Human Genome Project. *BioSocieties* 8(4): 397–416.

Q9 ~~Kelty, C. and Landecker, H. (2009) Ten thousand journal articles later: Ethnography of "The literature" in science. *Empiria: Revista de metodología de ciencias sociales* (18): 173–192.~~

Leinonen, R. *et al* (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Research* 38(Database issue): D39–D45.

Q10 Leonelli, S. and Ankeny, R.A. (2011) Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences.*

Leonelli, S. (2013) Global 43(1):29-36 *es* 8(4): 449–465.

Q11 Leonelli, S. (2014) What difference does quantity make? On the epistemology of Big Data in biology. *Big Data and Society* 1, 2053951714534395.

Li, H. *et al* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16): 2078–2079.

Loman, N.J. *et al* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* 30(5): 434–439.

Mackenzie, A. (2014) Machine learning and genomic dimensionality: From features to landscapes. In: H. Stevens and S. Richardson (eds.) *Post-genomic Sciences*. Durham, NC: Duke University Press.

Mardis, E.R. (2011) A decade/'s perspective on DNA sequencing technology. *Nature* 470(7333): 198–203.

Nadim, T. (2012) Inside the sequence universe: The amazing life of data and the people who look after them. PhD thesis. http://eprints.gold.ac.uk/8012/, accessed 30 October 2013.

NCBI (2011) Status of the NCBI sequence read archive (SRA). http://www.ncbi.nlm.nih.gov/About/news/13Oct2011.html, accessed 29 January 2013.

Neilson, B. (2012) Five theses on understanding logistics as power. *Distinktion: Scandinavian Journal of Social Theory* 13(3): 322–339. http://www.tandfonline.com/doi/abs/10.1080/1600910X.2012.728533, accessed 30 October 2013.

Nekrutenko, A. and Taylor, J. (2012) Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility. *Nature Reviews Genetics* 13(9): 667–672.

Rohde, H. *et al* (2011) Open-source genomic analysis of Shiga-Toxin–producing *E. coli* O104:H4. *New England Journal of Medicine* 365(8): 718–724.

Rose, N. (2006) *The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-first Century*. Princeton, NJ: Princeton University Press.

SEQanswers (2011) Short Read archive canned – SEQanswers. http://seqanswers.com/forums/showthread.php?t=9431, accessed 29 January 2013.

Shendure, J. and Aiden, E.L. (2012) The expanding scope of DNA sequencing. *Nature Biotechnology* 30(11): 1084–1094.

Siva, N. (2008) 1000 Genomes project. *Nature Biotechnology* 26(3): 256–256.

Snyder, L.A.S., Loman, N., Pallen, Mark, J and Penn, C.W. (2009) Next-generation sequencing-the promise and perils of charting the great microbial unknown. *Microbial Ecology* 57(1): 1–3.

Star, S.L. and Ruhleder, K. (1996) Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 7(1): 111–134.

Stevens, H. (2011) Coding sequences: A history of sequence comparison algorithms as a scientific instrument. *Perspectives on Science* 19(3): 263–299.

Stevens, H. (2012) Dr. Sanger, meet Mr. Moore. *BioEssays* 34(2): 103–105.

Sunder Rajan, K. (2006) *Biocapital: The Constitution of Postgenomic Life*. Durham, NC: Duke University Press.

Thacker, E. (2005) *The Global Genome: Biotechnology, Politics, and Culture*. Cambridge, MA: MIT Press.

Tsing, A. (2009) Supply chains and the human condition. *Rethinking Marxism* 21(2): 148–176.

Zhu, Y., Stephens, Robert, M., Meltzer, Paul, S. and Davis, S.R. (2013) SRAdb: Query and use public next-generation sequencing data from within R. *BMC Bioinformatics* 14(1): 19.