CHAPTER 6

Making Reference Genomes Useful: Annotation

This chapter looks at the processes of annotation: the identification and adding of biologically-relevant information to the reference genome, which can then be visualised in genome browsers, with the annotations aligned against the reference sequence itself. Annotation is both a key part of the creation of a reference genome and a definitional criterion of being a designated reference genome in the RefSeq database. It is the way in which the data produced in genomics are linked with the concerns and interests of the empirical life sciences and particular problems that motivate the work of specific communities: what historian Jon Agar (2012, 2020) has termed "working worlds".

This chapter demonstrates that the establishment of ever-more automated and refined pipelines incorporating multi-dimensional data—including cross-species comparative and 'beyond the genome' data such as protein sequences—was only part of the story of the development of genome annotation. We show that the manner in which annotation has developed was affected by: the ways in which the algorithms, protocols and operations of these pipelines were configured and improved; how they related to practices of *manually* annotating genomes; and the role played by the interactions of specialist genomicists with particular research communities. These factors were also pertinent to shaping what got annotated, how, and what use was made of the resulting enriched reference resources.

M. García-Sancho, J. Lowe, *A History of Genomics across Species, Communities and Projects*, Medicine and Biomedical Sciences in Modern History, https://doi.org/10.1007/978-3-031-06130-1_6

We show that different models of annotation are shaped by the relationship between reference sequence production efforts and the nature of the involvement of different communities converging around the genomes of particular species. Pig genome annotation, as a collaboration between the community of pig genomicists outlined in Chap. 5 and a well-developed annotation infrastructure at the Sanger Institute, differed in its nature and outcomes from yeast and human genomics. In yeast genomics, the community of yeast biologists was intimately involved in the reference genome production, while the initial annotation of the genome was orchestrated by the central bioinformatics coordinator, the Martinsried Institute for Protein Sequences.[1] In human genomics, two models existed: one involved the creation of high-throughput annotation pipelines at the institutions participating in the International Human Genome Sequencing Consortium (IHGSC), while the other—developed by their rival, the company Celera Genomics—was more open to input from prospective sequence users. In the former case, as with the reference genome sequencing, the medical genetics community was largely uninvolved. In the latter case, a subset of this medical genetics community was brought into the fold and contributed towards the realisation of a product—an annotated genome—distinct from that emanating from the large-scale sequencing centres leading the IHGSC effort.

One key commonality between the multiple species we have examined is the involvement of the Sanger Institute. In the previous chapter, we saw how the Sanger Institute's relationships with the different species communities varied in important and consequential respects. In this chapter, we show how the relationship of the Sanger Institute to the existing pig genetics community, already particularly close during the production of the *Sus scrofa* reference genome, was even more entangled for the annotation of the resulting sequence. This annotation used data from prior annotation and sequencing (in particular of the human genome) and availed itself of the Sanger Institute's infrastructures and procedures (pipelines) developed through human (and pre-human) sequencing projects. However, this annotation effort also had crucial input from the pig genomics community, whose members played a significant role in manually annotating the genome, confirming the automated annotations of the Sanger Institute, and contributing to an already-established panoply of comparative resources, empirical data and theoretical insights. Rather than

[1] Yeast genomicists often refer to annotation as *sequence analysis* or *functional analysis* of the genome. The term *annotation* is more uniformly preferred in human and pig genomics.

just being a large-scale data producer, the Sanger Institute features here as a collaborator, facilitator, trainer and provider of quality assurance, as well as the manager of various data infrastructures.

This changing role exhibited by the Sanger Institute enables us to show that the story of increasingly automated and data-intensive annotation pipelines merely corresponds to *some* of the ways in which the IHGSC institutions operated. We demonstrate that a broader multi-species approach to examining the history of annotation practices helps us to notice strategies that connect to the working worlds of the communities using the sequence data. This allows us to disclose the activities of communities that had long been generating and interpreting sequences, and to incorporate their trajectories into the history of the production of reference genomes.

6.1 Annotation: Pipelines and Jamborees

6.1.1 What Is Annotation and How Does It Contribute to the Production of a Usable Reference Genome?

Broadly speaking, annotation is the marking of features of interest in the abstract landscape of the sequences of nucleotides. Typically, representations of the genome accessible to researchers and the lay public are in the form of a browser, a window in which the user can select or deselect different features and modes of presentation of the genome to be conveyed to them (Fig. 6.1). The different selected features are aligned vertically next to a horizontal representation of the strands of the chromosome, which depicts the order of nucleotides along it if the user zooms in sufficiently. The browsers are based on database resources, perhaps incorporating several nested layers of data drawn from different sources.

The features that can be annotated include:

- Open Reading Frames (ORFs; segments between start and stop codons—specific sequences that may indicate the presence of transcribable DNA such as a gene);
- Genes (and their structure, organisation and variants);
- Repeat sequence regions, including those constituting telomeres at the ends of chromosomes and centromeres that perform a key role in the chromosome dynamics of cell division;

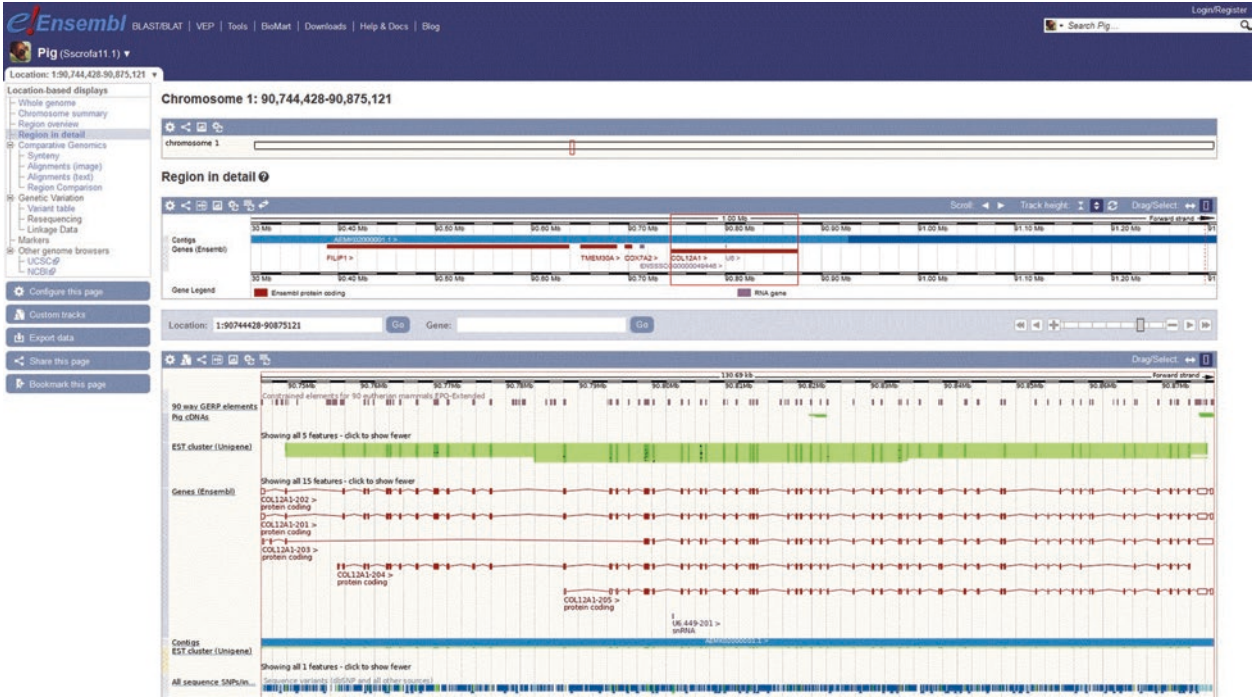

Fig. 6.1 Example of a display of a reference genome—Sscrofa11.1 for *Sus scrofa*, the pig—on a genome browser: Ensembl. Taken from Sscrofa11.1 Chromosome 1: 90,744,428-90,875,121, Ensembl (Howe et al., 2020), release 105 (https://www.ensembl.org/Sus_scrofa/Location/View?r=1:90744428-90875121;db=core, last accessed 18th December 2022)

- Pseudogenes (which appear similar to genes but do not function as such, due to mutations—these may have originally been copies of functioning genes);
- Regulatory regions that are not themselves expressed, but that affect the expression of genes.

Beyond these, many different kinds of sequence variants can also be identified and annotated, including structural variants in which stretches of nucleotides have been deleted, inserted, added, moved and inverted (Mahmoud et al., 2019).[2] Genomic variation comes in many forms, from differences in individual nucleotides, through variation in the sequence of individual coding regions, variation in the number of copies of repeat sequence in particular regions, to differences in sequence at a more gross level such as structural variants.

Two key distinctions have emerged to describe the processes and objects of annotation: manual and automated annotation; and structural and functional annotation.

Manual annotation involves the marking of genomic features using biological knowledge, such as the known sequence and location of a given gene. This way, the sequence is interpreted and contextualised using evidence from a variety of sources that may include earlier automated annotations. In automated processes, the genome assembly is first computationally analysed to identify key features such as repeat sequences and ORFs, and then existing datasets are interrogated to make predictions as to the annotation of more complex features such as protein-coding genes. These predictions are then examined further using a variety of algorithms embedded in different software to synthesise different forms of data and thus establish consensus models of the gene, which may include its structure and the existence of different forms. The data used in these automated processes include Expressed Sequence Tags (ESTs), known protein sequences and RNA sequences. These data can concern the species being annotated, as well as other species known—through prior comparative work—to be genomically close enough to the target species in order that cross-species inferences between parts of the genomes known to be equivalent can be made (Lowe, 2022).

Typically, generic pipelines have been designed and continually developed to annotate genomes, similar to the way that ones have evolved to produce and assemble sequence data (Stevens, 2013). These pipelines

[2] https://www.ncbi.nlm.nih.gov/dbvar/content/overview/ (last accessed 18th December 2022).

involve the specification of a series of sequential tasks and associated protocols, though typically different options for routes along the pipeline may exist to enable projects with differing levels of resources to navigate it. While some projects may have the resources to, for example, pay for additional manual annotation to refine the automated annotation, others may not. The existence of generic pipelines, together with the use of cross-species data, shows how genomic endeavours for different species interact. The infrastructures are built to accommodate difference, but also to channel it to ensure that the products of the pipelines are commensurate, even though they may serve—and be used by—different communities.

Alongside the selection of the source of DNA and the planning of the project, it is in annotation that the reference genome as a creative product of a particular configuration of actors is most manifest. The ways and the extent to which the annotation process enables new forms of genomics and genome-related research and resource development, however, depends on the details of the construction of that reference genome. Such details include the libraries used and how the genomic variation of the species was abstracted into the reference sequence. Also crucial are the relationships of particular research communities to various aspects of the process from pre-reference genomics through to annotation, as we show below.

The distinction between structural and functional annotation appears to map onto the distinction between (reference) genomics and post-(reference) genomics, which is explored further in the following chapter (Chap. 7). Structural annotation is the identification of particular features of the genome such as genes and their organisation, but also other functional and non-functional elements. Functional annotation is the connection of this structural data to other forms of data that help to make sense of the products and role of particular genomic elements. Broadly, we discuss structural annotation more in this chapter and functional annotation in the following chapter, but in doing so, we reveal that the distinction and apparent temporal succession from structural to functional is not clear cut.

6.1.2 Creation of Annotation Infrastructures

Annotation practices pre-date the annotation of reference genomes, and even the invention of DNA sequencing: for instance, the annotations in Margaret Dayhoff's early DNA sequence database were modelled on those in her previously-established protein sequence database (Strasser, 2019, p. 209). In the generation, collection and curation of annotations in databases such as GenBank, Stephen Hilgartner (2017) and Bruno Strasser

(2019) have identified two broad periods. There was an earlier period in which database staff themselves had to collect and annotate individual sequences, by trawling the literature. Then there was the period that succeeded this, in which the producer of the sequence data was able to submit it—with pertinent annotation—directly to databases with the help of specially-designed software tools.[3] In alliance with funders and journal editors, the databases helped to increasingly transform this practice into a duty.

In the first period, in the 1980s, annotation was essentially in the form of metadata; curators would read journal articles reporting a new nucleotide sequence and annotate the sequence by indicating the source of DNA and key features within the string of nucleotides. This process was advanced by agreements forged from 1982 onwards between GenBank and the Nucleotide Sequence Data Library hosted at the European Molecular Biology Laboratory (EMBL), and later (1987) between these and the DNA Data Bank of Japan. This tripartite alliance later became formalised as the International Nucleotide Sequence Database Collaboration. They divided up the laborious tasks of going through the literature and extracting and annotating sequences between themselves. Furthermore, to get around existing compatibility problems, they strove to harmonise the format that the data was recorded in.

In spite of this, and the use of supercomputers at the US Department of Energy's National Laboratory at Los Alamos to try to automate data processing and annotation, the rapidly-increasing production of sequences led to a backlog. This encouraged GenBank to streamline the process, in part by skipping the annotation or making it more cursory (Hilgartner, 2017, pp. 157–161; Strasser, 2019, pp. 228–230). As annotation was meant to be about making the data useful and "biologically meaningful", enabling it to be picked up and re-used by researchers using the database, this was problematic (EMBL Director General Lennart Philipson, as quoted by Strasser, 2019, p. 232). The EMBL, closer to bench biology than the physicist-led GenBank (which was based at Los Alamos from 1982 to 1992), was less keen on short-cuts around or through the annotation process (Strasser, 2019). The inadequacy of the initial algorithms designed for annotating sequences at the EMBL led to the conscription of biology students and clerical staff to contribute to the effort. When this also proved insufficient, more senior biologists were cultivated, which involved

[3] Though, as we detail in Chap. 7, this transition does not occur so neatly for species-specific databases targeted at particular organismal communities such as the *Saccharomyces* Genome Database, or for functional annotation rather than structural annotation.

informing them about some of the basics of the operation of the database, as well as circulating new sequences that may have been of interest to them. Biological researchers at the EMBL could then work with the database staff to refine the sequences stored on the database—and their annotations—as well as helping to improve the algorithms used in automated annotation (García-Sancho, 2012, pp. 111–114).

From 1987 onwards, there was a strategic shift towards securing agreements with journals, by which they would only publish articles including DNA sequence data if they were accompanied with accession numbers, indicating that they had been submitted to a publicly-accessible database such as the DNA Data Bank of Japan, the EMBL one or GenBank. Even though these agreements and rules were variably enforced, they succeeded in encouraging more direct submission, especially when software tools making data submission easier for researchers spread. Further changes in rules and norms of submission followed in the 1990s and improvements in the way data were submitted and accessed also occurred. There was increasing adoption and ease of internet access, additional tools to interrogate the databases were developed (such as the Basic Local Alignment Search Tool—BLAST—sequence comparison software), additional databases beyond the basic sequence ones were launched, and ongoing improvements were made to the fundamental DNA sequence databases.

In 1992, GenBank came under the umbrella of the National Center for Biotechnology Information, which maintains a panoply of other reference and software resources, including the RefSeq database (Chap. 1), and ClinVar, which is explored in Chap. 7. As we showed earlier (Chap. 4), in 1994, the EMBL database moved from Heidelberg—where the EMBL headquarters are—to what is now known as the Wellcome Genome Campus in Hinxton, Cambridgeshire, to form the EMBL's European Bioinformatics Institute (EBI). The Wellcome Genome Campus is also where the Sanger Institute is based, a co-location of significance to the story of the development of annotation infrastructures, and the specific examples of annotation we detail in the following section.

For now, the relationship between the Sanger Institute and the EBI is pertinent, because of the role of these institutions in the creation of means by which the data in well-stocked nucleotide databases could be brought together and presented in a useable form for researchers. These resources, the database system AceDB and the genome browser Ensembl, were forged in the exigencies of reference genome sequencing: of the nematode worm *Caenorhabditis elegans* and the human, respectively.

AceDB, which stands for '*A C. elegans Data Base*', was originally founded in 1989 by Jean Thierry-Mieg and Richard Durbin. The former was a Centre National de la Recherche Scientifique (CNRS) researcher in France, and Durbin was in a spell at Stanford University in-between doctoral and postdoctoral work based at the Laboratory of Molecular Biology in Cambridge; he moved to the Sanger Institute in 1992 and stayed there full-time until 2017. As it developed, AceDB allowed users to access and relate different kinds of representations of the genome of *C. elegans* in an internet browser, to move between representations of the DNA sequence, and the genetic linkage and physical maps. In her historical investigation of *C. elegans* genomics and the nature of the AceDB enterprise, Soraya de Chadarevian has highlighted the infrastructuring work that is required to make maps—that have been produced in very different ways and constitute distinct representations—commensurable in databases and visualisations generated using them. The production of new kinds of maps, including the full genome sequence, was driven by specific concrete demands (e.g. of particular communities) that were often independent of those that drove the construction of preceding maps. In making different kinds of maps interoperable through this work of commensuration, the specificities of the objectives, communities, practices and historical trajectories involved in forming these resources are flattened (de Chadarevian, 2004). This eases visualisation and navigation by users, but at the cost of abstracting the underlying specificities and lineages. As we show below, this double-edged sword—of easing inter-operability at the expense of flattening specificities—persisted in other infrastructures produced at the Wellcome Genome Campus.[4]

In 1999, the same institution at which AceDB was developed—the Sanger Institute—collaborated with the EBI to launch a key platform to accelerate the IHGSC human reference sequence effort: Ensembl. The Ensembl team devised a pipeline to help assemble the reference sequence

[4] The adoption of the map-based approach for human genomics was due to the success of the whole-genome sequencing of *C. elegans* using a prior physical map. Maps have historically informed sequencing, and small-scale sequencing was often a key part of genome mapping. Genomics research is also inextricably entangled across species, with practices, resources and tools developed by communities working on one species regularly used and adapted for different species (de Chadarevian, 2004; Lowe, 2022; Stevens, 2013, Ch. 7). All this involves the construction of infrastructures to enable commensurability, or at least interoperability across different representations and resources. Star and Bowker (2002) is a foundational text concerning infrastructuring, while Baker and Millerand (2010) examine infrastructuring concerning data in the life sciences. For examinations of analogous trade-offs involved in the creation and mobilisation of data itself, see Leonelli (2016); Leonelli and Tempini (2020).

and present it online through a genome browser.[5] The Ensembl browser presents an abstracted view of any part of the genome one chooses to zoom-in to. It offers a variety of 'tracks' representing different annotated features of the genome that can be selected and lined up alongside the reference nucleotide sequence, which is itself arrayed horizontally (Fig. 6.1). Ensembl does not only generate these visualisations but, for vertebrate species, also produces the annotations that are included in them, through its own automated annotation pipelines. It augments this with downloaded annotation data for other key non-vertebrate species. Ensembl, therefore, exhibits the clear will in the late-1990s to automate the annotation process and to bring it 'in-house' into the small number of institutions producing sequence data.

The manual annotation of select species was conducted by the Human And Vertebrate Analysis and Annotation group (HAVANA) at the Sanger Institute. HAVANA had its origins in the Human Sequence Analysis team led by Tim Hubbard within 'Team 71', the Informatics division that was led by Durbin at the Sanger Institute. The Sanger Institute component of Ensembl led by Michele Clamp was also part of Hubbard's team. Jennifer Ashurst (later Harrow) joined this team in April 2000 and led a distinct HAVANA group within the team from 2002. At the time she joined, there were two people working on manual annotation. However, it became apparent that Ensembl's automated annotation generated too many false positives due to the quality of sequence data then available to them.[6] It did predict approximately 70% of human genes accurately, good enough for a rough-and-ready annotation of the draft genome, but not of the required quality for biomedical research or diagnostic purposes. To improve the quality of the annotation, manual annotation was required that would make use of data coming in from the automated pipelines, but also involve curatorial decisions based on biological knowledge.[7]

[5] Other key general genome browsers include the UCSC Genome Browser hosted by the University of California Santa Cruz and Genome Data Viewer hosted by the NCBI. There are also more specialist species or taxon-specific browsers, such as the *Saccharomyces* Genome Database.

[6] This was despite it being expressly designed to produce gene predictions of high specificity at the cost of high sensitivity, in other words, to try to avoid false positives even if that meant missing true positives. This is a reflection of how difficult it was to generate effective automated procedures, and from early on the Ensembl team recognised that subsequent manual curation—and evaluation and refinement of the gene structures that were the outputs of automated annotation—was vital (Birney, Andrews, et al., 2004).

[7] Jennifer Harrow, interview conducted in Cambridge by James Lowe, October 2017.

HAVANA developed the curated Vertebrate Genome Annotation (VEGA) database and browser, which was built on Ensembl. VEGA was operational for human manual annotations from 2002, and mouse and zebrafish from 2003.[8] The browser was curated using both manual annotations conducted by the HAVANA group itself (such as for human chromosome 20) and by other groups and institutions (such as Ian Dunham's for human chromosome 22, and Genoscope and the CNRS for human chromosome 14). From early on, this annotation and curation work was accompanied by the development of protocols for manual annotation. At two 'Human Annotation Workshops' (HAWK1 and HAWK2) hosted by HAVANA in March and September 2002, participants from multiple institutions involved in manual genome annotation discussed possible standards and guidelines. A test sequence was annotated using different manual and automated methods at HAWK1, and the results of this were compared.[9] These workshops formed the basis for the manual annotation standards used in VEGA and were intended to aid commensurability across other resources and genome browsers developed at the NCBI and University of California Santa Cruz (see note 5). The Otter manual annotation system that was developed for HAVANA by Ensembl and used in VEGA was designed in accordance with the standards formulated in the HAWK workshops (Searle et al., 2004).

From 2014 to 2017, Ensembl became solely part of the EBI. HAVANA became part of Ensembl at the EBI in 2017. By then, HAVANA had branched out to work directly with some species communities on manual annotation; the pig was one of these, as we see later in this chapter.

6.2 Annotating the Yeast, Human and Pig Genomes

When we consider the annotation process across the main three species we look at, we find that the nature of it depended on: the generation and use of existing genomic resources such as maps and genome libraries; the existence of data such as that on Expressed Sequence Tags (ESTs), complementary DNA (cDNA) sequences, RNA sequences, and protein sequences; the nature of the inferential apparatus available for intra-specific and

[8] Sequencing the reference genome of zebrafish (*Danio rerio*), a model organism, was an initiative begun at the Sanger Institute in 2001.

[9] https://web.archive.org/web/20020825133038/http://www.sanger.ac.uk/HGP/havana/hawk.shtml (last accessed 18th December 2022).

inter-specific data analysis; the kind of community and actors involved and their interests; and the mode of organisation of genomic projects. The available data sources and inferences were marshalled to find and elucidate the fine structure of genes and other elements of the genome. A closer look at the specifics of annotation practices across these three different species allows us to complicate the relationships between automated and manual processes, as well as between structural and functional annotation.

Four basic models of annotation were identified by bioinformatician Lincoln Stein in an article published in the summer of 2001 (Stein, 2001). He associated these models with particular stages of the annotation process, in terms of its increasing complexity and recontextualisation through forming connections to other kinds of biological data and knowledge. Two of his terms—factory and cottage industry—are familiar from earlier debates concerning the proper organisation of genomics (Chaps. 2 and 3). We have interpreted his designations in the scheme displayed in Fig. 6.2.

Stein's scheme, as interpreted, highlights the importance of the establishment of mechanisms by which existing datasets and resources can be accessed and used in annotation, as well as the significance of the role of annotation itself in enabling the creation of new links to other datasets and standard references, such as the Gene Ontology. This enables the

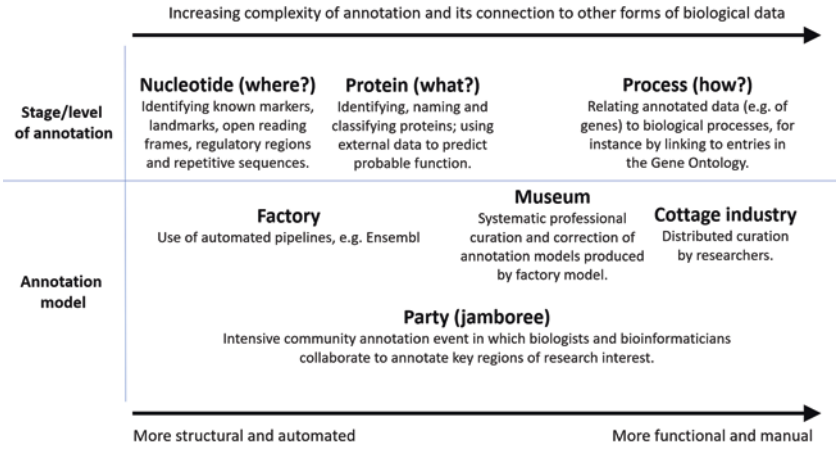

Fig. 6.2 Diagrammatic depiction of models of annotation and how they relate to different stages or levels of annotation. (Produced by both authors, based on Stein, 2001)

decontextualised reference sequence to be progressively connected to other forms of biological data and therefore recontextualised.[10] This process makes use of software and algorithms to search external databases. Crucially, it also uses maps and libraries employed in the construction of the reference genome to initially annotate the sequence. This seeds further annotation by providing reference points to aid the searching of external data, and also aids the later contextualisation of the annotated data. Stein's conception, while consisting of stages, does break down firm distinctions between manual and automated annotation, and also structural and functional annotation, as entanglements of each are implicated in any one point. Key here is that the weights of the different modes (automated/manual; structural/functional) change as the annotation process proceeds. The schematic we have drawn from Stein is a useful overview of the general trends in the annotation process, and it constitutes a helpful reference point with which to consider examples that depart from the sequential and separable stages implied by it. For instance, we may observe that the main genome browsers such as Ensembl moved towards a hybrid factory-museum model (Loveland et al., 2012).

Quite apart from the particular manifestations of sequencing, and the extent to which they may depart from Stein's ideal types, the ways in which particular communities and genomic endeavours undertake annotation is constrained by multiple factors. These include the histories, motives and resources of particular communities of genomicists. Furthermore, groups such as HAVANA developed forms of community annotation, in which they acted as facilitators—rather than the sole conductors—of the annotation process. As we detail below, these forms of community annotation involved the creation of software tools such as Otterlace/Zmap for manual annotation on the cottage industry model, and also more direct interactions with research communities, such as the one that had been working on pig genetics and genomics (Loveland et al., 2012).

6.2.1 Yeast Genome Annotation

For yeast genome sequencing, as previously noted, one finds a community of geneticists, cell biologists, biochemists and molecular biologists, often dedicated to working with standardised strains of *Saccharomyces cerevisiae*.

[10] Stein makes explicit mention of the entries in the Gene Ontology; such data resources also involve processes of annotation, albeit featuring different models and kinds of curatorial roles; see Leonelli (2016).

The ease of working on this unicellular eukaryote was what made it a model organism, and this engendered the virtuous cycle by which the existing weight of scientific capital—in the form of mounting knowledge, resources, tools, and mechanisms of dissemination and sharing—justified new investment in its further augmentation. When the perception began to grow that "[t]he yeast genome was becoming overstudied, and yet…, largely unexplored!"—that different research groups were working on the same genes while much of the genome was *terra incognita*—multiple laboratories across Europe, Japan, Canada and the USA rallied to participate in an unprecedented collaboration to sequence the first full eukaryotic genome (Dujon, 1996, p. 263; Chap. 2).

The structural annotation of the yeast genome reflected the hierarchical, top-down and distributed approach of the sequencing effort in Europe. Within the initiative funded by the European Commission, the centralised bioinformatics function located at the Martinsried Institute for Protein Sequences (MIPS) was married with the specific expertise of the laboratories performing sequencing, and seeking to make use of the data so generated.

MIPS, on assuring the quality of the sequences it received and assembling contiguous tracts of sequence (contigs) on the basis of them, screened the data for ORFs by identifying stretches of minimum numbers of nucleotides (from about 50 to 300, a lower number risking more false positives and a higher one more false negatives) with no stop codon. They also sought contigs with sizes below the threshold by searching for sequences that were homologous (showed sufficient similarity) to known protein sequences, based on the knowledge of the genetic code and processes of transcription and translation. Already, this analysis relied upon existing experimental knowledge of this well-studied organism, as well as the prior delineation of protein sequences and elucidation of their functions. Using sequence homologies, the MIPS team was able to classify the ORFs in terms of their putative functions (Mewes et al., 1998). Once the data had been passed on to the sequencing laboratories, the initial identification of the ORFs could be built on with a deeper analysis of these sequences. This was done either using existing biological data or materials (for example, concerning centromeric and telomeric DNA, tRNA and Ty elements for chromosome II) or by performing a variety of experiments to characterise their functional role. Following the conclusion of the

reference genome sequencing, such experiments were organised and conducted in a concerted way in a successor project on functional analysis and annotation called EUROFAN, which is discussed in Chap. 7. Due to the limitations of homology analysis, with about 40% of putative genes being "orphans" either having no discovered homologues or homologues with no known function, such functional analysis would also enable the verification of the structural annotation.

Once the presumed coding regions were separated from the non-coding, the non-coding regions could be further analysed to detect sequence motifs (including promoter regions of genes) and other features such as transposable elements (Ty elements). Many of these non-coding elements were of interest to participants in the network, who could use the genomic data that they generated—and MIPS processed—to further their research. For example, Horst Feldmann at Ludwig-Maximilian University of Munich was particularly interested in Ty elements (Chap. 2) and advanced his research using the structurally annotated sequences he now had access to. These sequences had themselves been augmented using the data he had previously collected (Feldmann et al., 1994; Heumann et al., 1996; Mewes et al., 1998). While the centralised parts of this process, such as the role of MIPS, will seem analogous to some of the informatics pipelines and groups of the IHGSC discussed in the next section, the yeast biology laboratories played an important role in refining and developing the initial annotations that were made by MIPS. Unlike in human reference sequencing, in which prospective users were not involved in the processes of data production, in the yeast genome effort there was a set of users incorporated in those processes (García-Sancho, Lowe, et al., 2022).

The completion of the sequencing and sequence analysis of the different chromosomes at different times enabled innovations developed for one chromosome to be taken up by groups working on other parts of the yeast genome. For example, the methods that yeast geneticist Bernard Dujon developed for the evaluation of ORFs to identify which ones were indeed "functional genes" in the chromosome XI paper published in June 1994 were then used in the chromosome VIII paper published in September that year (Dujon et al., 1994). Chromosome XI was Europe-led, while VIII was coordinated from Washington University by Mark Johnston. While they exhibited different organisational models, as we saw in Chap. 2, there was enough of a connection for each to build on the advances of the other.

Washington University's model of annotation was also different, though in practice they used searches of public nucleotide and protein databases to identify cross-species homologies with known genes and protein sequences, as well as examining other elements such as tRNAs, much as MIPS did. For assembly and annotation, they (along with some European-led groups) used a version of AceDB: AScDB, with 'Sc' standing for *S. cerevisiae* rather than the 'ce' of *C. elegans*. AScDB had been specially adapted for yeast by Richard Durbin, young EMBL bioinformatician Erik Sonnhammer and LaDeana Hillier, the director of informatics at the Washington University Genome Sequencing Center (Johnston et al., 1994). Hillier collaborated closely with Johnston, and also worked on *C. elegans* and human genomics. With the benefit of a comparative perspective gained from interaction with the yeast, human and *C. elegans* efforts, she observed that a significant problem with "smaller numbers of groups doing the sequencing" was that "user education" could be "an issue". However, for "yeast the user education was taken care of because the sequencing was done at so many different places that everybody [...] understood the limits of the data" (Hillier, 2012, p. 7).

Dujon and Johnston gave assistance to the chromosome I team that mainly operated at McGill University. They were the next to publish—in April 1995—with Dujon helping with sequence analysis and Johnston providing the chromosome VIII sequence, which enabled some genome duplications to be identified. Later papers indicate a continuation of this cooperation around sequence analysis. These publications document a refinement of the processes, datasets and software used from the early published chromosomes onwards (Bussey et al., 1995; see also Galibert et al., 1996). This stands in contrast to the development of novel tools and the infrastructural transformations associated with human genome annotation or the adaptation of established infrastructures and processes to the particular demands of pig genomics.

For the Europe-led sub-projects, MIPS continued its role in sequence analysis. It did not see its task as restricted to identifying individual genomic elements, but also as aiding the global characterisation of the genome, by using their initial structural annotation to partition the genome into units. As a consequence, sequence comparisons could be made between these units, in order to identify gene duplications to aid future functional analysis and provide data that could be used in tracking

the evolution of the *S. cerevisiae* genome. These twin approaches of targeting function and diversity that arose out of the initial work to structurally characterise the genome form an important part of the narrative of Chap. 7.

For the purposes of sequencing and annotation, yeast had clear advantages over the bulkier organisms that we consider next: humans and pigs. The yeast genome is considerably smaller in size, but also more economical, in that it contains comparatively little non-coding DNA and complex gene structures, compared with multicellular eukaryotes. As a model organism, it also had a panoply of available experimental evidence that could be used and built on to inform both automated and manual approaches to annotation. Additionally, the range and extent of functional analysis conducted by the yeast genomics community that we discuss in Chap. 7 was not possible for human and pig. This meant that distinct strategies for annotation needed to be developed for these species. For the human genome, this involved making use of the abundant ESTs and protein sequence data that had been gathered, the creation of automated and manual sequencing pipelines, and advancing the means with which to conduct analyses of homology by harnessing and further developing comparative genomic approaches.

6.2.2 Human Genome Annotation

In the three major papers describing the sequence of the entire human genome (authored by the IHGSC in 2001 and 2004, and by Celera in 2001), only the Celera paper includes details of the annotation process. For the IHGSC, the details of annotation are dealt with only in the subsequent individual papers describing the sequence of each chromosome. This reflects, we suggest, the IHGSC primary concern of getting assembled sequence out in the public domain to prevent its enclosure by some form of intellectual property. On the part of Celera, the inclusion of information about annotation evinces their commercial strategy of building the foundations for the exploitation of the genome for biomedical purposes. Even though they described aspects of their annotation process, users would still have to pay to access Celera's full annotated sequence. In this way, Celera sought to make itself an obligatory passage point for those seeking the richly-annotated data that they produced.

The first chromosome that the IHGSC sequenced was chromosome 22, by a team led by Ian Dunham at the Sanger Institute. The paper announcing this appeared in December 1999, before Ensembl and HAVANA were up and running. Tim Hubbard's sequence analysis team were involved, though, and they integrated existing data on nucleotides and protein sequences, using similarity searches (through programmes implementing the 'BLAST' algorithm developed at the NIH by Gene Myers and colleagues) and prediction programmes (Dunham et al., 1999). Like the annotation of subsequent chromosomes, an early stage was identifying repetitive sequences and 'masking' them. This meant filtering them from view so that they were not incorporated in automated analyses of the sequence data. To do this, the annotators used 'RepeatMasker', a piece of software developed and (then) hosted by the Genome Sequencing Center at Washington University. The remaining unmasked sequence was then analysed for the presence of various genomic features, such as spotting areas of the genome with a relatively high proportion of guanine and cytosine bases in order to discern the presence and location of CpG islands, in which cytosine is next to guanine. These are frequently located in the promoter regions of genes and are therefore a good indicator of the presence of genes.

At this point, the automated aspects of searches and the use of prediction programmes were interweaved with manual approaches. In large part, this was because of the calibration and verification required for each method, and the overall need to evaluate and refine the annotation process. A re-evaluation of the chromosome 22 annotation in 2003 reaffirmed the value of combining automated prediction, sequence similarity and comparative methods in annotation, but observed that the optimum configuration of them with respect to each other had not yet been found. Furthermore, at this time the ideal comparator species for similarity analysis was unclear. The authors acknowledged that while annotation processes would be improved, at that point automated approaches had significant limitations. As well as refining data categories and making use of new sources of data (e.g. new human ESTs and various kinds of data on related species), overcoming these limitations would involve manual analysis and experimentation (Collins et al., 2003).

The only other chromosome sequence published before the announcement of the completed draft of the whole genome in February 2001 was

for chromosome 21, conducted by a consortium led by RIKEN (**Ri**kagaku **Ken**kyūjo, the Institute of Physical and Chemical Research) in Japan.[11] This team also conducted gene predictions and sequence similarity searches. They additionally defined criteria by which putative gene classifications were assigned to one of five categories, depending on the strength of the evidence for them being protein-coding genes. They, therefore, placed the discernment of functional elements of the genome such as protein-coding genes at the heart of their annotation effort, an orientation appropriate to the biomedical interests of many of the institutions that worked on chromosome 21. That emphasis—and the function-centred annotation—motivated and aided the paper's substantial analysis of the medical implications of their results (Hattori et al., 2000).

The biomedical interests of RIKEN's collaborators were the exception rather than the rule for most institutions involved in sequencing subsequent chromosomes within the IHGSC effort. This was reflected in the way that the sequence data was analysed in the publications announcing their completion. Advances in the analysis of sequence data were heralded, but in so doing, the potential biomedical users of the data were a secondary concern. As we now detail, these analytical advances constituted refinements and additions that augmented the annotation pipelines for each successive chromosome. The augmentations that these specialist genomicists introduced were directed towards improving the capabilities of genomics *qua* genomics, as an enterprise in itself with its own internal goals and motivations. They sought to improve their assemblies and annotations according to internal generic metrics of quality, contiguity and coverage, guided by an overall ideal of completeness. In other words, they did not primarily shape the annotation process and its products in such a way as to fulfil the requirements of any specific external community or set of users.

The first chromosome sequence published after the announcement of the draft whole sequence was chromosome 20 in December 2001; after

[11] Other members of the consortium were: Keio University School of Medicine in Japan and from Germany the Max Planck Institute for Molecular Genetics in Berlin, Institute for Molecular Biotechnology in Jena, and German Research Centre for Biotechnology in Braunschweig. Collaborating institutions were the National Cancer Center Research Institute and University of Tokyo (both Japan), UMR 8602 CNRS at UFR Necker Enfants-Malades and CNRS UPR 1142 at the Institute of Biology (both France), Eleanor Roosevelt Institute (USA), University of Geneva Medical School (Switzerland) and School of Pharmacy, University of London (UK): Hattori et al. (2000).

this, there was a gap in 2002 before a flurry were published across 2003 to 2006.[12] What did the progressive accretion of methods and sources of data consist of, across the five years since the completion of chromosome 20?

The chromosome 20 paper, signed only by authors from the Sanger Institute, was the first to use the Ensembl database in the analysis of the sequence; this sequence was, though, still assembled and visualised in AceDB. The genomicists were able to make use of sequence data from two vertebrates (the mouse *Mus musculus* and the pufferfish *Tetraodon nigroviridis*) in their comparative analyses rather than merely the mouse maps that the previous chromosomes had relied on (Deloukas et al., 2001).

For chromosome 14 (February 2003), a two-step annotation approach was employed by the collaboration between Genoscope, the Institute of Systems Biology in Seattle and the Washington University Genome Sequencing Center. In this, automated methods using computational predictions to formulate provisional models of the structure of genes, were refined by sequence similarity analysis. This was complemented by experimental data on gene expression using microarrays, a tool containing potentially many thousands of DNA probes that can indicate the presence or absence of specific complementary sequences. In the "manual curation" that followed, the genomicists used additional data to refine the gene models produced in the first stage and remove "suspicious data" such as partial matches that were not found to contain any significant coding sequences (Heilig et al., 2003, p. 607).

Washington University Genome Sequencing Center was also heavily involved in the completion of chromosome 7 (July 2003), as well as the Y chromosome (June 2003). These featured a significant focus on methods for the identification of pseudogenes, including K_A/K_S analysis to identify the kind and extent of selection operating on putative pseudogenes and known genes. In this type of analysis, the scientists generated reconstructed ancestral sequences to detect signatures of neutral evolution (and therefore an absence of positive or purifying selection) which would indicate the presence of a pseudogene. They then checked these inferences

[12] Chromosome numbers were assigned according to the observed size of the chromosomes in karyotypes. Generally speaking, this is reflected in their length, with the longest nuclear chromosome being 1, the second-longest being 2, and so forth. There are some exceptions at the shorter end: 21 is longer than 22, and 20 is longer than 19, for instance. It is easy to see why, therefore, the higher-numbered (and therefore shorter) chromosomes tended to be sequenced earlier, and the lower-numbered ones tended to be sequenced later (1 was the last to be published), though this was only a general trend.

using the available mouse sequence data (Skaletsky et al., 2003; Hillier et al., 2003).[13]

Like chromosome 20, the paper heralding the completion of chromosome 6 (in October 2003) was wholly authored by people at the Sanger Institute. Since 2001, there had been considerable developments in their annotation process. Ensembl was now more refined, and the HAVANA team was established and embarking on their extensive manual annotation. VEGA was now up and running and hosting the annotated sequence data. Built into the heart of Ensembl's automated annotated process were two sequence-matching tools: GeneWise for exploiting protein sequence data and Genomewise for using EST and cDNA data indicative of the presence of transcribed genes (Curwen et al., 2004; Birney, Clamp and Durbin, 2004). In its design, the Ensembl pipeline had been configured to integrate and more effectively deploy existing annotation methods. In addition, it was now able to make use of sequence data on the rat (*Rattus norvegicus*; an animal model), another pufferfish (*Fugu rubripes*; with a far more economical genome than other vertebrates) and zebrafish (*Danio rerio*; a model organism) as well as the mouse and *Tetraodon nigroviridis*. Using the protocols and standards forged in the HAWK meetings in 2002, the HAVANA group manually curated the gene structures generated through the Ensembl pipeline. Given their later role in facilitating community annotation of immune response genes in the pig, it is appropriate that HAVANA's first formal role in human genomics concerned chromosome 6, which contains the Major Histocompatibility Complex implicated in immune response.

[13] The theory behind this approach is that compared with a reconstruction of the ancestral version of the gene, a functional gene will exhibit *either* a high ratio of nonsynonymous substitutions to synonymous substitutions—reflecting positive (directional) selection—or it will show a low ratio resulting from stabilising selection. Synonymous substitutions mean that observed mutations—when compared with the ancestral version of the gene—will result in no change in the amino acid that is specified by the codon (the triplet of bases read during DNA transcription); there will therefore be no change in the function of any gene products as a result of such substitutions. A gene that has undergone positive selection has had its sequence altered in a manner that increases the fitness of its holders. Stabilising selection, by contrast, ensures that the sequence does not change—as changes would be disadvantageous to the organism. These evolutionary mechanisms can therefore be identified using this analysis. Pseudogenes can also be detected. They should exhibit a ratio of about 1, indicating that there has been no selection either way. This absence of selection is expected for non-functional parts of the genome such as pseudogenes.

We will return shortly to the annotation of the remaining chromosomes, focusing on the development of Ensembl and HAVANA at the Sanger Institute. For now, with the expansion of the number of creatures for which informative sequence data was available in mind, we make a brief excursion into the development of comparative genomic resources and approaches.

As we noted in earlier chapters, a comparative genomic perspective was present in genomics from its inception. Genome sequencing projects on other species were used as pilots to aid the planning of the Human Genome Project. Furthermore, the map and sequence data of those other species were used to help construct human genome maps and sequences, by applying knowledge about comparative regions between the species. Finally, it was also envisaged that establishing a rich understanding of comparative connections between human and non-human genomes would enable the more fruitful exploitation of the human resource. In one respect, this was because experimental interventions on organisms such as yeast and animal models could then be connected to and inform human biology through genomic and other omics data. In another respect, this was because of the mooted contribution of data on other species towards enriching the annotation of the human genome.

To aid human genome annotation in this way, in December 2003, the Large-Scale Sequencing Program of the US National Human Genome Research Institute (NHGRI) established two Working Groups: one on 'Annotating the Human Genome' chaired by Robert Waterston and the other on 'Comparative Genome Evolution' chaired by Laura Landweber and John Gerhart. Both groups were tasked with identifying what new sequencing could be conducted in large-scale sequencing centres to advance human genome annotation and functional analysis. The Comparative Genome Evolution group also had to identify which organisms to sequence to shed new light on human evolution and genome evolution across eukaryotes in general. Each of the groups identified three components of research, a range of organisms and appropriate sequencing strategies (including coverage to be obtained) to contribute to these components, and indicated percentages of total sequencing capacity to be allotted to each task.

The Annotation Working Group recommended that 15 non-primate mammalian genomes be shotgun sequenced at relatively low coverage in two successive sets (known as 'Bins'). They further indicated that other genome efforts already in progress, including for non-mammals such as the chicken, should proceed further so that complete high-quality sequences be produced to aid the identification of conserved sequences across

mammals. The second component suggested by the Annotation Working Group was the high-quality sequencing of two primate genomes and relatively high-coverage shotgun sequencing of three others, to enable differences to be identified between these and the human genome. The third component was a recommendation to survey human genomic variation by sequencing 1000 people at very low coverage. The group additionally suggested that "a modest cDNA effort be included as a component of all genomic sequencing projects" to aid assembly and gene prediction.[14]

The Comparative Genome Evolution working group's recommendations ranged more deeply and widely across the tree of life, further extending the selection criterion employed by the Annotation Group by which some species would be preferentially sequenced due to representing key phylogenetic positions. Both groups also deployed other criteria to recommend particular organisms as candidates for sequencing, including the quality of the submissions ('white papers') sent in by the relevant communities; the role of the organism as a model; its potential biomedical significance; its economic importance; the possibility that a genome sequence for it would enable the construction of reference sequences for closely-related organisms of biological significance and the size and heterogeneity of the genome.[15]

A Coordinating Committee (chaired by William Gelbart) then evaluated the proposals, presenting a modified set of recommendations to the NHGRI's Advisory Council for approval in May 2004.[16] We consider this further in the following chapter when addressing different aspects of post-reference genome work on the human. For now, it is pertinent to note that in the documented assessment of species proposals by the Working Group on Comparative Genome Evolution, their conception of the communities working on these organisms and submitting white papers to the NHGRI was very much as groups of *users*. The evaluations that the NHGRI made of the white papers were based on the readiness of these

[14] "New Sequencing Targets for Genomic Sequencing: Recommendations by the Coordinating Committee", part of the documents for the Meeting of the NHGRI Research Network for Large-scale Sequencing and the NHGRI Sequencing Advisory Panel, May 16, 2004 (NHGRI History Archive 7036–021).

[15] The community of pig genomicists submitted one of these white papers (Chap. 5).

[16] "New Sequencing Targets for Genomic Sequencing: Recommendations by the Coordinating Committee", part of the documents for the Meeting of the NHGRI Research Network for Large-scale Sequencing and the NHGRI Sequencing Advisory Panel, May 16, 2004 (NHGRI History Archive 7036–021).

communities for receiving the genome. Their role was envisaged as developers of proposals for the NHGRI to judge, and as groups that needed to corral the appropriate resources to make use of what the NHGRI would end up providing for them.[17] New research goals were added for subsequent rounds of sequencing additional species, such as identifying the mammalian "core genome". The increasing apparatus and empirical basis of comparative analysis guided the number and selection of sequencing targets and the methods deployed on them.[18]

Returning to the annotation of the individual chromosomes, the remaining ones that the Sanger Institute was involved with were: 13, 9, 10, X, 17 and 1. For chromosome 13, published in April 2004, the availability of a new database for non-coding RNAs, Rfam, advanced the annotation of these, which had been deemed extremely tricky as recently as in the chromosome 6 paper published in October 2003. For chromosome 13, modifications had been made to the Ensembl pipeline to aid manual curation. With the chromosome 9 paper, published in May 2004, there was a special focus on duplications of segments of the chromosome, which were assessed using K_A/K_S analysis (see note 13). Having previously mapped Single Nucleotide Polymorphisms (single base changes; SNPs) against their sequence using data from the dbSNP database, for chromosome 9 the genomicists identified their own bank of SNPs by analysing the sequence data from overlapping portions of DNA fragments (clones). In May 2004's chromosome 10 paper, the authors continued their identification of SNPs and extended this focus at the single nucleotide level by comparing 617,071 single nucleotide sequence differences between human and chimpanzee, conducting K_A/K_S analysis on the results to ascertain the presence of sites of selection. From this paper on, there was an increasing focus on annotating alternative splice variants, which result from transcription processes that generate multiple different messenger RNA sequences from a single gene.

In the X chromosome paper published in March 2005, there was a particular focus on the evolution of the X chromosome and comparisons were made between it and the Y chromosome. The chicken (*Gallus gallus*) genome assembly was used for this analysis in addition to previously mentioned comparator species, many of which now had newer versions of their

[17] "Report of the Annotation of the Human Genome Working Group", dated January 3, 2005 (NHGRI History Archive 7039–005).

[18] E.g., https://www.genome.gov/Pages/Research/Sequencing/SeqProposals/2x-7x_promotion_seq.pdf (last accessed 18th December 2022).

assemblies that were used. For the April 2006 paper on chromosome 17, human sequencing was conducted at the Broad Institute; the Sanger Institute's role focused more on the sequencing of mouse chromosome 11 as part of the Mouse Genome Sequencing Project.[19] The paper was mostly dedicated to a comparative analysis of the two chromosomes and a reconstructed ancestral chromosome, with the authors focusing on an assessment of the different changes to the chromosomes that occurred in the distinct evolutionary lineages.

The final chromosome to be published, in May 2006, was 1. In the paper, the genomicists aligned the chromosomal sequence to the now-standard array of comparator species (minus the chicken) to identify regions of evolutionary conservation. This paper also represented a culmination of the increasing focus on SNPs from 2004 onwards. These SNPs were used to identify and map genomic diversity within species, identify recombination at a higher resolution than previously possible, detect signals of selection, and as a resource to augment the utility of the reference genome (Dunham et al., 2004; Humphray et al., 2004; Deloukas et al., 2004; Ross et al., 2005; Zody et al., 2006; Gregory et al., 2006).[20] The comparative approaches and cataloguing of diversity were conducted to ease the process of developing genomic resources, by feeding into and augmenting the pipelines of the IHGSC participants. The intended use of the resources so produced, however, was generic rather than tailored to specific user communities.

Compared to the IHGSC effort discussed above, Celera's approach was quite distinct, giving potential communities of users of genomic data a more active and participatory role than in the IHGSC and NHGRI's annotation strategies. As noted above, Celera's 2001 paper discussed annotation far more than the contemporary IHGSC one. It was an automated annotation that it chronicled, though, in a discussion of their Otto gene prediction system. This software was designed to weigh different forms of data constituting evidence for particular annotations, namely cDNAs and ESTs. The weighting was based on Celera's previous

[19] The Broad Institute was opened in 2004, the result of collaboration between the Whitehead Institute, Harvard University and hospitals affiliated with Harvard.

[20] The other chromosomes were handled by the Stanford Human Genome Center and the US Department of Energy (19, 5, 16), Washington University (2 and 4), the Broad Institute (18, 8, 15, 11, 17; 18 with RIKEN, 11 primarily RIKEN with the Broad Institute, and 17 with the Sanger Institute), and Baylor College of Medicine (12 and 3; 3 with BGI, formerly known as the Beijing Genomics Institute).

experience of the manual annotation of the *Drosophila* genome. This approach therefore reaffirmed and reflected the process of genomic discovery promoted by Venter in the early-1990s, especially the crucial importance it conferred to protein-coding regions of the genome, as revealed by EST and cDNA sequence data. While the paper reported some computational validation of Otto's results, it acknowledged that the "[e]xtensive manual annotation to establish precise characterization of gene structure" that was still deemed necessary lay in the future (Venter et al., 2001, p. 1317).

As their automated annotation took inspiration from prior work on *Drosophila,* so did their manual annotation, by using the jamboree model. *Drosophila* genomics was not the only inspiration, however. A challenge that Celera faced was the absence of information about the means and decision-making procedures by which the public project's annotations were made. Therefore, to develop their own annotation capabilities, they needed to obtain institutional knowledge of how the sausage was made. To that end, they recruited Peter Li from Johns Hopkins University, who had worked on the GDB Human Genome Database and the Online Mendelian Inheritance in Man (OMIM) catalogue while there, and as a result was acutely aware of the details of the annotation process. The OMIM connection, deepened by the use of data from it in the annotation of Celera's gene sets, was just as significant as the model of *Drosophila* genomics to the way that Celera manually annotated the human genome. OMIM used curators who were experts on particular diseases, with their knowledge of the relevant genetics feeding into the published data. The need for biological expertise to contribute towards the annotation—and more broadly, the contextualisation of the data that Celera was generating—was keenly felt by the company. Due to its particular sequencing strategy, it had invested considerably in computational infrastructure and expertise for the purposes of assembly rather than in acquiring biological knowledge. But because of the need to generate rich and translationally-relevant data to be incorporated into proprietary databases (such as The Celera Discovery System™), drawing on this kind of expertise was essential.

A variety of academics were therefore invited to participate in a human annotation jamboree that took place in April 2001, two months after the publication of the draft reference sequence. This jamboree built on the previous one that Celera had held on the *Drosophila* genome and involved some of the OMIM curators (García-Sancho, Leng, et al., 2022). The human genome jamboree presented an opportunity for participation on the part of medical geneticists who had been largely uninvolved in the

IHGSC effort. They would contribute their expertise, in concert with the computational experts at Celera, and in turn were given access to the latest proprietary data on their area of interest, as well as the fruits of their collaboration with Celera. Following the publication of their sequence in *Science* in 2001, Celera kept further improvements to their assembly behind a paywall for their clients, who were primarily pharmaceutical and biotechnology companies rather than academics. At the jamboree, though, the academics could assess the sequence assemblies in regions on which they had expertise, contributing information that would not just refine the gene structures predicted by Otto, but also inform improvements to the overall automated annotation pipeline.

The involvement with medical geneticists did not end there. A further Chromosome 7 Annotation Project was initiated, prompted by a suggestion by medical geneticist Stephen Scherer to Richard Mural, the head of the Annotation Team at Celera. The result was a higher quality re-sequenced chromosome 7 that better connected to biomedical and clinical research due to the expertise and physical mapping data provided by medical geneticists. This provided the medical genetics community with a useful resource, as well as aiding Celera in its strategic reorientation towards identifying diagnostic and therapeutic targets.[21]

The ways in which genomes are improved and connected to other forms of data are explored further in the next chapter. For now, we note that the institutional imperatives of the IHGSC and Celera shaped the design of their respective annotation processes. Annotation, therefore, emerged in ways that reflected the trajectories, networks and goals of practitioners; Celera was more open to the medical genetics community, while the IHGSC was more self-contained.

In the following section, we consider the annotation of the pig genome, an effort in which existing pig genomicists interacted closely with teams at different stages of the sequencing and analysis pipeline established at the Sanger Institute. This reflected the model of interaction between medical geneticists and Celera more than the way that annotation unfolded within the IHGSC human reference genome sequencing. Furthermore, the relationship between the existing community of researchers working on the pig and the Sanger Institute helped to shift

[21] Peter Li, interview conducted over Skype by both authors, September 2020. See also Kerlavage et al. (2002).

some of the Sanger Institute's operations towards a model closer to the community annotation advanced by Celera.

6.2.3 Pig Genome Annotation

As it came after the sequencing of other genomes at the Sanger Institute, by the time the pig genome was sequenced, the annotation process used an established pipeline derived from procedures that had been deployed and refined in previous initiatives, in particular the sequencing and annotation of *Homo sapiens*. Like in sequencing and assembly, the pig project adopted and used repertoires established through the experience of projects on other species, while adding distinctive twists on these.

For the sequencing itself, the community of pig genomicists through the Swine Genome Sequencing Consortium (SGSC) had contracted with the Sanger Institute rather than the project being initiated from within the IHGSC (Chap. 5). This contractual relationship did not, however, imply a hands-off approach by the community; it was intimately involved in guiding the strategic—and in some cases operational—direction of the project. Part of this direction meant indicating to the Sanger Institute where they should target sequencing efforts, so they could focus on particular areas associated with genes of interest to individual research groups. This was reflective of a desire to make genome data useable as promptly as possible. As a result, even while the sequencing was still underway the community pursued annotation, the identification of SNPs and the creation of a SNP chip that captured agriculturally-relevant genetic variation.

We discuss the creation of the SNP chip in the following chapter. Here we detail the annotation effort. Just over £1.1 million of funding was secured from the UK Biotechnology and Biological Sciences Research Council (BBSRC) for 2007–2010 by the Roslin Institute (with Alan Archibald as Principal Investigator and Andrew Law as co-investigator), the EBI (Ewan Birney as Principal Investigator) and the Sanger Institute (Tim Hubbard as Principal Investigator and Jane Rogers as co-investigator).[22] These grants funded four posts, one each in Hubbard and

[22] https://gtr.ukri.org/projects?ref=BB%2FE010520%2F1#/tabOverview (last accessed 18th December 2022); https://gtr.ukri.org/projects?ref=BB%2FE010520%2F2#/tabOverview (last accessed 18th December 2022); https://gtr.ukri.org/projects?ref=BB%2FE010768%2F1#/tabOverview (last accessed 18th December 2022); https://gtr.ukri.org/project/6AB44634-8225-4645-8935-CC9977F581BD#/tabOverview (last accessed 18th December 2022).

Rogers' teams at the Sanger Institute, one in Archibald's group at the Roslin Institute and one supervised by Birney at the EBI. Two of these positions (with Hubbard and Birney) were in the Ensembl teams at the EBI and Sanger Institute. As noted above, the annotation effort began while the sequencing itself was still being conducted. Like in human genome sequencing, the pig genome was scanned using algorithms to predict the presence of genomic features. Pig protein and RNA sequence data were obtained from specific databases, and data on pig cDNA and ESTs were also downloaded from GenBank. Many of the cDNAs and ESTs had been generated by the Animal Genome Research Program at the National Institute of Agrobiological Sciences in Japan, and the Japan Institute of Association for Techno-innovation in Agriculture, Forestry and Fisheries (Groenen et al., 2012 and Supplementary Information; Lowe, 2018). These resources were generated in part using samples from cloned offspring of TJ Tabasco (Schook et al., 2005; Uenishi et al., 2012).[23]

A key feature of the automated annotation in the Swine Genome Sequencing Project (SGSP) was the integration into the Ensembl pipeline of multiple forms of data already generated by the community from prior projects. These data concerned maps, Quantitative Trait Loci, and clones, in addition to the cDNA and ESTs mentioned above. The community provided Ensembl with these rich resources to enable the annotated reference sequence to be connected with—and immediately contextualised by—other forms of data and information produced by pig geneticists. This enabled functional inferences to be made concerning parts of the genome, but also inferential pathways to be constructed between the pig genome and other porcine biological data, and also between the pig genome and the genomes of other species. With the means to generate comparisons with other mammalian genomes being a key product of the grant work, this connectivity was intended to boost the pig as a comparative model, with data and the results of experiments intended to travel along the connections forged within the species, but also then to be able to travel beyond the species. Crucially, this wider horizon was accompanied by a desire to embrace the varied research needs of the community of pig researchers in the annotation, through the addition of tracks comprising other forms of data to the Ensembl browser. This was

[23] This Japanese effort also used tissues from crossbred pigs derived from Landrace, Large White and Duroc breeds, and ones from a Chinese Meishan pig, two Landrace pigs, a Berkshire pig and a miniature pig (Uenishi et al., 2012).

effected through Ensembl's Distributed Annotation System, and pig geneticists who were interested in adding these tracks for the forms of data valuable to them were invited to contact Archibald, who was in regular liaison with teams at the Sanger Institute and the EBI.[24]

There were therefore multiple kinds of community involvement in even the automated annotation of the pig genome. The community helped to define the nature of the annotation, taking advantage of the clone-based sequencing to squeeze as much use out of the products of sequencing and assembly as possible, through integrating assembly and annotation as well as incorporating data and resources already developed by the community into the pipeline, or through the Distributed Annotation System. This was particularly important, as the resource limitations of the overall genome project entailed a trade-off between comprehensiveness and utility, with the community opting for a more rough-and-ready but more immediately exploitable resource, above aspirations for completeness.

This meant that the drawbacks of automated annotation, well-appreciated by the Ensembl team for the more refined human genome, were even greater for the pig genome. As Jennifer Harrow reported to us, the algorithms at the heart of Ensembl were only as good as the assemblies they were working on, and for the pig these were incomplete and of lower quality than for the human. Manual curation of the data by the biologically-trained members of the HAVANA team was therefore more critical for improving and developing the initial assemblies of the pig genome produced by the Ensembl pipeline, than it was for human or mouse.[25]

As with human genome sequencing, the annotated sequences produced through the Ensembl pipeline were published in the Ensembl database, while additional manual annotation was published on the HAVANA-led VEGA database, built on the Ensembl database.[26] HAVANA worked closely with some of the members of the pig genomics community, such as Christopher Tuggle at Iowa State University. James Reecy, an animal geneticist in Tuggle's group, spent his faculty leave (equivalent to a sabbatical) with them from September 2007 to August 2008. Like many pig geneticists, Reecy worked on multiple livestock species, in his case

[24] "PIG TALES: Newsletter of the International Swine Genome Sequencing Consortium (SGSC) Pig Genome Sequence Project", 2nd Quarter 2007—Volume 1 Issue 3. On the Distributed Annotation System, see: Dowell et al. (2001).

[25] Jennifer Harrow, interview conducted in Cambridge by James Lowe, October 2017.

[26] For more on VEGA, see Harrow et al. (2014).

primarily cattle. Reecy was interested in developing skills in manual annotation and areas of programming, and HAVANA had put together the most comprehensive approach to manual annotation in the world at the time. He was able to pursue this because of the close interactions between the pig genomics community and leading figures at the Sanger Institute, which we saw in Chap. 5. During his visit, Reecy met with Jane Rogers, Tim Hubbard and Richard Durbin, as well as Jennifer Harrow and Jane Loveland of HAVANA, discussing what he could offer in situ at the Sanger Institute. Aided by his demonstration that an animal geneticist could pick up the techniques of manual annotation, Reecy's advocacy of community involvement in annotation met a receptive audience in the HAVANA team.

As a result, HAVANA decided to dedicate more attention to manual annotation than they had been contracted to do and in so doing developed new means of manually annotating a genome.[27] This new model took two forms. HAVANA consulted with the SGSC members on an informal basis for guidance on what precise parts of the genome they wanted special attention paid to. This was a continuation of the targeted approach to sequencing and meant that the annotation could be preferentially refined in particular regions of interest to researchers. In the process, information was fed back to the assembly team if a problem was detected in the course of the manual curation.[28] As the annotation started while the reference genome was being assembled,[29] this allowed it to feed into the assembly (and even inform the amendment of algorithms in automated assembly pipelines), as well as adding value to the eventual sequence.

Additionally, HAVANA shifted its mode of operation, developing new capabilities in education, training and engagement to increasingly function as community annotation facilitators, providing the pig geneticists with the tools, training and assistance so that they could annotate the genome themselves. This began with a training programme hosted at the

[27] This illustrates the importance of the initial choice of the Sanger Institute to host the sequencing of the pig genome, even if this was not made with the eventual model of annotation in mind. The Human Genome Sequencing Center at Baylor College of Medicine, the other candidate to sequence the pig genome, as it had the cattle, was comparatively quite small. The kind of manual annotation employed for the pig and the development of community annotation would therefore have been less likely to occur there.

[28] Jennifer Harrow, interview conducted in Cambridge by James Lowe, October 2017; Kerstin Howe, interview conducted at Wellcome Genome Campus (Hinxton, Cambridgeshire) by James Lowe, October 2017.

[29] Craig Beattie, interview conducted over Skype by James Lowe, March 2017.

Sanger Institute in July 2008. While this event was labelled as a "jamboree", it differed from the *Drosophila* and human jamborees organised by Celera. Rather than just annotating the genomes in situ, the Sanger Institute event was intended to equip the researchers to go back to their own institutions and conduct annotation on regions of the genome pertinent to their existing research projects there. Abridged guidelines were created for pig annotation, due to the need to do the annotation quickly because of resource constraints, but also to economically document the key processes and procedures for these amateur annotators scattered around the world. Conference calls were used to share problems, observations and advice, but a manual was still needed for the HAVANA facilitators to refer to, and for the manual annotators to consult in their own offices and labs between meetings (see Fig. 6.3).

This community annotation effort was aided by the availability of the Otterlace/ZMap system combining a relational database and graphical interface for the manual annotators to use (Loveland et al., 2012; Dawson et al., 2013). In turn, HAVANA used their close working relationship with the pig genomicists to develop their tools and annotation processes.

The initial step in the manual annotation process was the computational alignment of multiple forms of data from the pig—and other species such as human and mouse—onto the *S. scrofa* genome assembly. A crucial feature of the Otterlace/ZMap manual annotation system used by HAVANA and VEGA was that it enabled annotation of an ongoing assembly rather than just individual clones, which was all that previous curation tools had allowed users to annotate (Searle et al., 2004). This functionality was helpful to pig genomicists, who wanted to promptly exploit and further augment the sequences so assembled. It meshed with the more significant role that manual procedures had in the annotation of the *S. scrofa* reference genome. The combination of the automated pipeline with the bespoke manual sequencing distributed in laboratories across the world constituted a combination of Stein's factory and cottage industry models, and was therefore different to the case of Ensembl discussed above (Lowe, 2018).[30]

This initial curation created a visualisation that displayed the sequence data along with another layer of information indicating evidence for the possible presence of genes. With this, anyone with an account could log in

[30] As we discuss later, the manual annotation of the X and Y chromosomes was performed by the Sanger Institute itself.

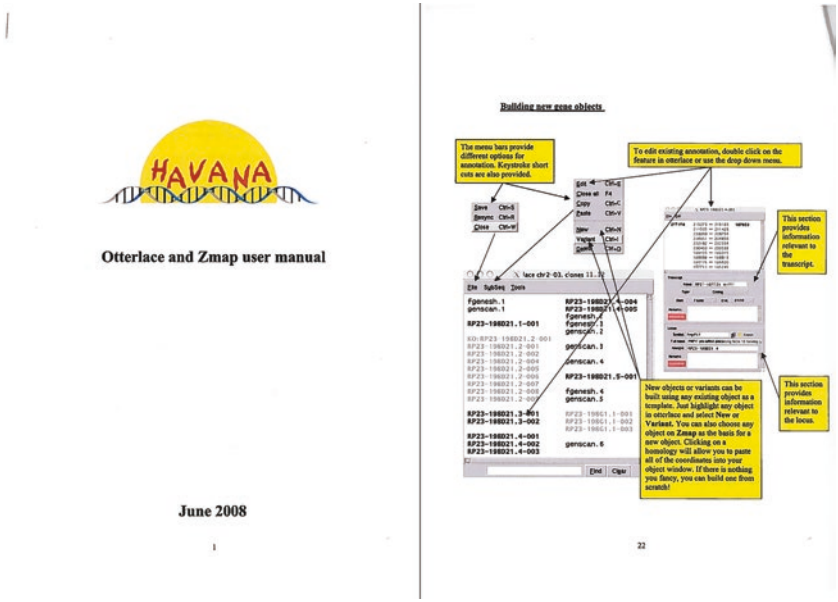

Fig. 6.3 Cover and selected page of a manual produced by the HAVANA team for use by manual annotators of the pig genome community. From personal papers of Alan Archibald, "Pig Sequencing" folder, obtained 17th May 2017. Reproduced with permission, courtesy of Alan Archibald and the Human and Vertebrate Annotation group at the Wellcome Trust Sanger Institute. For a larger version of this figure that can be zoomed in and out, see https://www.pure.ed.ac.uk/ws/portalfiles/portal/314800096/higheres_fig_6_3.pdf

to the Otterlace/Zmap system and start to annotate a chosen gene. The annotator could weigh the different forms of evidence presented to them, and amend the model of the gene according to that evidence and any specific knowledge of the gene that they have. They would then be able to submit it for inspection by a HAVANA team member, who could then work on it further to finish off the annotations to the required standard.[31]

In the earlier annotation of the human genome, as well as for well-funded model organisms such as the mouse, HAVANA had generally performed manual annotation wholly in-house. Its role was quite different for the pig, instead conducting education and training to enable researchers

[31] Jennifer Harrow, interview conducted in Cambridge by James Lowe, October 2017.

to themselves manually annotate genes, with the HAVANA team then performing quality control on the results. The only other species that HAVANA was providing community annotation support for at the time was cattle (*Bos taurus*). There were, though, weaker interactions between HAVANA and the cattle genomics community, partly because its greater funding meant that a close relationship was less necessary, but also because of the less-established links that the Sanger Institute had enjoyed with members of this community compared to pig genomicists (Chap. 5).

Parallel to HAVANA's tasks, the pig genomics community itself helped to organise the manual annotation activity. As bioinformatics coordinator, Reecy led the community side of the work and provided training on manual annotation in the USA and China. In Scotland, training was also provided by the Roslin Institute. With Reecy, Iowa State University colleague Zhi-Liang Hu set up a website listing the genes and gene families that were candidates for manual annotation, and individual researchers were invited to indicate which they intended to annotate. This has been described as an "adopt-a-gene type approach" by Reecy, building on the targeting strategy in the sequencing phase.[32] The community did not have the resources to manually curate the whole genome to a high standard. They needed to maximise the utility of the genome for their particular research purposes, and for this, selectivity and distribution of the sequencing were appropriate. The value of the genome was therefore not primarily assessed in terms of generic metrics, even if data on the number of genes annotated still constituted a useful barometer of progress. The key was the utility of what had been done, not the extent of it; such concerns with completeness were more of a priority for the IHGSC. The pig community assessed the *S. scrofa* genome in terms of its use as a research tool for their own purposes. They were themselves deeply imbued with an awareness of what was required in the domains of agricultural or other forms of translation that they worked towards.[33]

[32] The website (still live as of 18th December 2022) is: https://www.animalgenome.org/cgi-bin/host/ssc/gene2bacs. It was actively updated from November 2009 to September 2010. We thank Zhi-Liang Hu for kindly providing us with the information on this, following an initial lead provided to us by James Reecy. Most of the adopted genes were taken by the Immune Response Annotation Group (see below) and Cathy Ernst's research group. Zhi-Liang Hu defines himself as a bioinformaticist: someone who programmes new tools as well as using them (personal communication with James Lowe, January 2022).

[33] James Reecy, interview conducted over Zoom by both authors, May 2021.

For this manual annotation, particular groups were established based on common research and translation interests. Some of these focused on resolutely structural elements such as repetitive sequences, while others operated in areas where the line between structural and functional was blurred. Examples of the latter were the groups that aimed to annotate genes and analyse genomic regions relating to olfaction, immune response, and retroviral insertions into pig DNA such as Porcine Endogenous Retroviruses (PERVs). The range of interests of the pig genome community was reflected in these groups. In addition to the interests listed above, genomicists working on domestication and the relationships between the sequenced domesticated pig and European and Asian wild boar contributed analyses to the publication heralding the reference sequence (Groenen et al., 2012).

It was the involvement of the pig genomics community in annotation processes that helped to blur the line between structural and functional annotation. This is illustrated by the most developed of the annotation groups, which became the Immune Response Annotation Group (IRAG) and continued its activities well beyond the initial analysis of the reference genome. IRAG comprised 51 researchers based in thirteen institutions in China, France, India, Italy, Japan, UK and USA. There had been considerable work on immune response prior to genomic research, as we showed in Chap. 5. Further, a high-quality manually annotated sequence of the pig's MHC (the Swine Leucocyte Antigen complex, or SLA) was published in 2006, as a result of work by Laboratoire Mixte CEA-INRA de Radiobiologie Appliquée (CEA-INRA), Genoscope, Tokai University in Japan, and the Sanger Institute. The HAVANA team and the CEA-INRA group (in particular, Christine Renard) performed the manual annotation of the SLA region (Renard et al., 2006). It did not therefore need to be developed further in the subsequent 'immunome' project.

This ambitious 'immunome' project and group arose out of discussions between researchers at CEA-INRA and Iowa State University, in particular Claire Rogel-Gaillard at the former and Christopher Tuggle at the latter. They each had straightforward motivations for establishing this effort, since they both worked on the immunogenetics of the pig. We have already encountered Rogel-Gaillard, part of the team at CEA-INRA (and later, just INRA) that had adopted genomic approaches to investigating immune response. This had involved studying the dense polymorphic regions containing genes implicated in it from the 1980s, as well as investigating PERVs in the late-1990s, which had implications for the prospective

xenotransplantation of pig organs and tissues into humans. Together with Patrick Chardon, she had led the development of the YAC and BAC libraries of pig DNA to aid those research efforts (Chap. 5). Her research interests had increasingly been directed towards studying the genetics of immune response variability in terms of pig health and resilience against disease. Tuggle's research had trended in a similar direction, though from a different origin: his work in the 1990s was at the heart of the mapping endeavour to try to identify (and then exploit) genes and Quantitative Trait Loci primarily involved in livestock production traits in pigs.[34]

From this nucleus, a call for interested parties was issued, and once the participants were confirmed, the group set about seeking data from databases and the literature to identify a list of genes to annotate.[35] Once this list was agreed and the rules for annotation established, particular sets of genes were assigned to individual teams. The approach embodied the advantages and disadvantages of distributed, targeted community annotation, as while expertise could be applied to particular regions by researchers, this meant that some regions went unadopted, for instance those with lower sequence quality that were difficult to annotate as a result or ones that simply did not contain genes of interest.[36]

Reecy provided training for the group's annotators in a workshop, but beyond that people worked in their own offices and labs, using Otterlace. Annotators would be able to see the analysis for their particular region, with the data tracks (for example the RNAs aligned to it) depicted. They would also be able to use the software tools to tweak the predictions made at the Sanger Institute.[37] The work was coordinated, and credit negotiated, in regular conference calls, using the Webex videoconferencing application to share screens. Jennifer Harrow had overall oversight at the HAVANA end, which included making the decisions about which annotations to exclude. She guided Jane Loveland in the day-to-day management, coordinating annotation between different groups, showing

[34] Tuggle took over from Max Rothschild, his frequent co-author and superior in the Iowa State University Department of Animal Science in the 1990s, as the National Swine Genome Coordinator for the US Department of Agriculture.

[35] Claire Rogel-Gaillard, interviews conducted over Skype by James Lowe, May 2017. See also Dawson et al. (2013). In particular, the group searched for annotations in the Gene Ontology, using "immune system process", GO:0002376, as the inclusion criterion.

[36] Jane Loveland, interview conducted at Wellcome Genome Campus (Hinxton, Cambridgeshire) by James Lowe, October 2017.

[37] Christopher Tuggle, interview conducted over Skype by James Lowe, March 2017.

annotators how to use tools and access data, conducting quality control on the annotations and giving feedback. The motivation for HAVANA was to enable communities to take on as much of the task of annotation themselves as possible, both as a general aim and a particular solution for the resource-poor pig genomics community.[38] While the HAVANA team primarily supplied support for the informatics aspects of the manual annotation, on the community side a trio of coordinators—Rogel-Gaillard, Tuggle and Harry Dawson—guided the effort with a view to making the resulting annotated sequence as valuable as possible for those who would make use of it. Dawson, based at the USDA's Beltsville facility in Maryland, monitored which genes were being annotated, following up on any genes that remained unannotated. He also conducted cross-species comparative analyses based on the annotation data he compiled from the whole project.[39] Dawson had led the development of the Porcine Immunology and Nutrition (PIN) Database at Beltsville, which was launched in 2005 containing data on 2600 annotated pig genes, with gene expression data linked to information on gene function. The database (now known as the Porcine Translational Research Database) was configured to enable users to identify genetic pathways related to genes of interest and to connect to human and mouse databases for comparative purposes, as well as to other pig genomic databases (Dawson et al., 2007).[40]

Because the annotation began with a panel of genes, rather than simply annotating the assembly that was there, genes missing from the assembly could be identified, and therefore areas of the assembly that needed further work could be pinpointed. Indeed, having conducted the annotation using version (build) 9 of the swine genome, the results of the annotation fed into the newer and improved version 10.2. The annotators refined the models of 1369 genes and elucidated 3472 transcripts from these, around a third of which were inferred using only data from other species. They extended the analysis concerning genes under positive selection undertaken in the 2012 *Nature* paper announcing the reference sequence. And finally, the group used transcriptomic data derived from experiments to

[38] Jennifer Harrow, interview conducted in Cambridge by James Lowe, October 2017; Jane Loveland, interview conducted at Wellcome Genome Campus (Hinxton, Cambridgeshire) by James Lowe, October 2017.

[39] Claire Rogel-Gaillard, interviews conducted over Skype by James Lowe, May 2017. See also Dawson et al. (2013).

[40] https://web.archive.org/web/20220928072749/https://www.ars.usda.gov/news-events/news/research-news/2005/pig-gene-database-supports-human-nutrition-immunity-studies/ (last accessed 18th December 2022).

discern the role of some of the genes involved in immune response, identify networks of co-expression of genes and to annotate accordingly (Dawson et al., 2013).

This work had direct translational impact motivating it, and this gave the group clear indications on how to target their focus and structure the division of labour within the project. To achieve the translational ends of the researchers involved, the methods and approaches employed in the project were comparative, and explorations of function were knitted together with examinations of diversity and evolution.[41] For example, inferences that the researchers made about the evolution of genes accompanied functionally-oriented transcriptomic studies. Genes identified for their putative function enabled both the functional and structural annotation of the genome to be improved. And these in turn fed into the refined assembly of the genome itself.

Concerning the improvement of the reference genome as a community-generated resource, we close with an account of the sequencing and annotation of the pig's X and Y chromosomes. This project filled the gap left by the SGSP, which had excluded the sex chromosomes due to the complexities involved in their sequencing. The sequencing of the sex chromosomes therefore finally completed a reference sequence for the whole of the nuclear genome of *S. scrofa*. This project also shows how the existing community of pig genomicists were able to broker and contribute to a collaboration between the Sanger Institute and an external group of researchers who had been working on these sex chromosomes for both biomedical and agriculturally-oriented purposes.

This project involved the EBI and the Sanger Institute, was funded with a BBSRC grant, and used infrastructure and work that was supported by the European Commission and the Wellcome Trust, much like previous work we have described. It did not involve any of the 'usual suspects' from the pig community as a collaborative partner, however, but a group based in the Department of Pathology at the University of Cambridge who had been consistently investigating the sex chromosomes of the pig since the turn of the century.[42] Their research had a dual aspect, being

[41] In Chap. 7, we term such research on diversity and evolution as 'systematic' and examine the different ways in which explorations of these topics relate to functional studies across yeast, human and pig.

[42] https://gtr.ukri.org/projects?ref=BB%2FF021372%2F1#/tabOverview (last accessed 18th December 2022).

motivated by biomedical objectives, as well as being supported by a major pig breeding firm, the Pig Improvement Company (PIC), due to the implications of the genetics of sperm development and male fertility for breeding purposes.[43] The Cambridge University-led arm of the sequencing and annotation of the pig X and Y chromosomes was also conducted in collaboration with PIC. A key figure in the mapping of individual genes relating to sperm fertility was Andy Day. His funding came from PIC, who he had worked for since leaving university in 1995 and continued to be employed by until 2006. Day's research at the University of Cambridge used comparative approaches to exploit the more plentiful and refined data and resources concerning the human genome to aid in the mapping of specific genes in the pig (Day et al., 2003; Kollers et al., 2006). One of his collaborators, Claire Quilter, approached human–pig comparative genomics from a medical genetic angle: she worked on the role of the Y chromosome in male infertility and Turner syndrome, a condition that affects women and involves the lack of all or part of an X chromosome.[44]

In the early-2000s, Quilter had been the lead author of a paper that surveyed porcine sex chromosomes, identifying and mapping 19 genes onto them. For this, she made use of the PigEBAC library developed by the Roslin Institute and the UK Human Genome Mapping Project Resource Centre. This work explored the evolutionary consequences of this mapping data, in part by comparing the order of genes determined on the porcine Y chromosome with the corresponding order of those genes on the human and mouse Y chromosomes (Quilter et al., 2002). As well as representing a convergence of biomedical and agriculturally-inclined research, it also presaged the entanglement of comparative, evolutionary and functional studies that would be further realised in the work conducted with the Sanger Institute, and also the relationship between systematic and functional genomics explored in Chap. 7.

The X and Y chromosomes were an interesting challenge for the HAVANA team, due to the high level of conservation in X chromosomes and the tricky genomics of the Y chromosome. Y chromosomes contain a

[43] As with many of the institutions mentioned in this book, we have affixed one name for an institution that changed names and did not have a straightforward institutional history. The Pig Improvement Company was founded in 1962, was bought by Dalgety plc in 1970, which became the PIC International Group in 1998, and then Sygen International Group in 2001. Genus, a cattle breeder, bought Sygen in 2005. 'PIC' remains a brand for the pig breeding side of their business; for more, see Bruce and Lowe (2022).

[44] https://www.researchgate.net/profile/Claire-Quilter (last accessed 18th December 2022).

lot of repetitive sequences and degenerated genes due to its near-complete isolation from recombination with the X chromosome during meiosis.[45] In the original reference genome operation by the SGSP, some limited sequencing of the Y chromosome had been conducted using clones from the DNA libraries derived from males. However, only 11 clones were sequenced—in a draft rather than finished condition—and a limited number of scaffolds containing positioned contigs were placed on the chromosome: hardly an assembly (Groenen et al., 2012, Supplementary Information).

On the sequencing side, the X and Y chromosomes project began under the leadership of Jane Rogers. When she left the Sanger Institute, it was taken over by Chris Tyler-Smith, a human evolutionary geneticist. The sex chromosome sequencing project began in 2009. Both sides of the project were funded by the BBSRC for three years, with the Sanger Institute being alloted £1,369,161 to Cambridge's £349,639.[46] The endeavour would contribute an improved assembly and annotation of the X chromosome and the first assembly and annotation of the Y chromosome.

Beyond the original pig genome sequencing, the X and Y work benefited from a change in mapping techniques and improvements to sequencing techniques.[47] Optical mapping was used to build a new assembly of X. To conduct this, Kerstin Howe—who led the team that analysed, validated and improved genome assemblies such as the pig one—worked alongside David C. Schwartz, who pioneered the method for eukaryotes.[48] Optical mapping does not require the use of library clones and the technique obviates the need for reconstruction of the order of the clones. It was therefore useful in correcting problematic repetitive regions that are difficult to resolve using clone-based mapping. The new optical-based map enabled the corrected assembly to be produced, which was then improved further, for example with targeted sequencing to close gaps and resolve assembly problems. This improved assembly in turn enabled an improved annotation, with 690 protein-coding genes annotated, a

[45] Jane Loveland, interview conducted at Wellcome Genome Campus (Hinxton, Cambridgeshire) by James Lowe, October 2017. See also Skinner et al. (2016).

[46] https://gtr.ukri.org/projects?ref=BB%2FF02195X%2F1#/tabOverview (last accessed 18th December 2022); https://gtr.ukri.org/projects?ref=BB%2FF021372%2F1#/tabOverview (last accessed 18th December 2022).

[47] Jennifer Harrow, interview conducted in Cambridge by James Lowe, October 2017.

[48] Kerstin Howe, interview conducted at Wellcome Genome Campus (Hinxton, Cambridgeshire) by James Lowe, October 2017.

considerable advance over the 422 in the original (for Sscrofa10.2), with increased numbers of non-coding genes and pseudogenes identified as well. As with the SGSP, there was close interaction between the annotation and assembly teams at the Sanger Institute.

For the Y chromosome, a bespoke library was created using DNA from a Duroc boar (the same breed as the originator of the CHORI-242 clones from which the bulk of the reference sequence was derived) donated by Genus, the company that incorporated PIC. At the Sanger Institute, a fingerprint contig map was produced using this library to create a map of overlapping clones which formed the basis of a minimum tiling path to guide the sequencing and assembly. They used and combined the outputs of multiple sequencing platforms, and then improved it further as with the X chromosome, to bring the sequence towards 'Finished' standard. This updated assembly was validated using PacBio long-read technology, which affirmed the high quality of the new assembly, using the same clone library as the original sequencing conducted by the SGSP.

For both the X and Y chromosomes, annotation involved the alignment of various EST, messenger RNA, and protein sequence data against the sequence. This was performed through the Otter annotation pipeline, and it then underwent manual curation by the HAVANA team, using the Otterlace/Zmap tools according to the procedures developed for both human genome annotation through GENCODE (Chap. 7) and the immunome project (Skinner et al., 2016 and Supplementary Information).[49] The Y chromosome assembly subsequently became incorporated into the updated Sscrofa11.1 assembly, which became the reference genome (at 'representative genome' level in RefSeq) for the pig in 2017 (Warr et al., 2020).

Cambridge University's side of the project involved identifying shared regions between the two chromosomes to aid in the sequencing of them and in tracing their evolutionary history, identifying functional genes and non-coding sequences on the Y chromosome, and locating and analysing a gene—*HSFY*—found in cows to study chromosomal evolution across pigs and closely-related species. The insights gained from this project were explicitly designed to inform the sequencing and assembly of the chromosomes using the knowledge gained about their structure and the location

[49] On the PacBio validation: Jane Loveland, interview conducted at Wellcome Genome Campus (Hinxton, Cambridgeshire) by James Lowe, October 2017.

of repetitive sequences, but also to guide the exploitation of the data.[50] This research was therefore a good example of the functional and systematic synergies that are explored further in the following chapter.

It also shows how the specific genetic expertise of a group of researchers newly admitted to the community of pig genomicists, fed into and informed the highly-developed pipelines and expertise at the Sanger Institute. Here, the Sanger Institute did not conduct this work merely at its own initiative or at the behest of the Wellcome Trust or an international collaboration like the IHGSC. It also was not merely contracted to perform the work, as per the original relationship with the pig genomicists. Instead, building on the relationships developed through pig genome sequencing, which intensified as attention was directed towards annotation and the development of a new community-oriented model of it, the X and Y project constituted a more horizontal peer-to-peer collaboration from the start. This collaboration involved the highly-refined infrastructures and personnel of a large-scale genome centre. It incorporated a community of pig genomicists with a core of operators such as Alan Archibald who married a drive towards the development of genomic resources intended for wide use with a sensitivity to particular uses to which they could be put. And finally, it included an existing set of researchers seeking to conduct sequencing and annotation pertaining directly to their ongoing interests.

The X and Y project instantiates deep entanglements between different models of sequencing and annotation. It challenges strict demarcations and distinctions, and also the linearities indicated by presumed separations between stages, whether in particular projects or pertaining to the wider development of genomics. Who would dare reduce this X and Y project—or any part of it—to a singular form of annotation along the lines of Stein's ideal types, or even to any of the strategies pursued in prior genomics projects such as the genome centre model of the IHGSC, or the distributed model of the European Commission-funded Yeast Genome Sequencing Project? Instead, as the progression of pig genomics illustrates, aspects of these models were mobilised and combined, mediated by the historical trajectories of the actors coming together to form particular projects.

[50] https://gtr.ukri.org/projects?ref=BB%2FF021372%2F1#/tabOverview (last accessed 18th December 2022).

6.3 Annotation Strategies and Lineages of Genomics

In examining the different models of reference genome annotation for yeast, human and the pig, this chapter has begun to explore the development and use of genomic resources beyond the determination of the nucleotide sequence of the reference genome. This broader perspective expands the range of narratives that historians can mobilise to capture genomics as an ongoing and multifaceted endeavour, moulded in distinct ways by different communities.

The yeast genome annotation followed the distributed-but-hierarchical model of the European Commission's sequencing project, with a key role for MIPS as the bioinformatics coordinator. The centralisation through MIPS reflected the division of labour of the sequencing across multiple, often small, laboratories and the need for a genome-wide perspective for some forms of genome analysis that the consortium wanted to perform. In this model, we see a strict separation of structural from functional annotation.

The human reference genome, on the IHGSC side, involved the development of the Ensembl pipeline and HAVANA to automatically and then manually annotate the sequence data. IHGSC institutions progressively added new sources of data and methods for the annotation of various elements in the human genome, such as protein-coding genes. Compared with Celera's approach, this involved far less interaction with wider communities of researchers, and instead a concentration on developing pipelines and repertoires to improve the quality and extent of annotation, without directing or targeting it towards particular users. The aims and operations were therefore internal to a community of specialist genomicists, institutions and operatives, who sought to improve the output as measured by general metrics and guided by an ideal of completeness.

This, as we have seen, was not a fixed or essential characteristic of the genome centres, the key institution in the IHGSC model. In the case of the Sanger Institute, for example, the relationship of some of its departments and key personnel to a well-coordinated pig genome community effected a change in the way this institution worked. As a result, the model and results of the annotation of the pig genome were quite distinct from the human annotation that preceded it.

Some of this was driven by resource constraints that limited the quality of the pig genome assembly in some respects, making manual curation more crucial in correcting the automated predictions. As funding would

only go so far in paying for in-house manual curation, the community would need to take up the slack. The extent they were able to do this owed much to the community's own history of coming together to coordinate the work of identifying genetic markers, compiling and integrating genetic, cytogenetic and physical maps, and creating databases and materials (such as genome libraries and radiation hybrid panels). They pursued the creation of genomic resources because they knew what kinds of data they needed to advance their own research. Together, they advanced their overall endeavour of improving the genomic reference resources concerning the pig, secured pots of money from various sources to do so, and then worked out how to stretch what they had as far as they could. This accommodated but also drew upon the heterogeneous but often overlapping interests held across the pig genome community. For their members, like those forming the yeast genomics community, genomics has constituted a nexus around which multiple different interests could draw upon the resources generated through it, with those interests and motivations also shaping the creation of those resources in distinctive ways.

Indeed, a reference genome is a creative and dynamic product. The selection of the materials that are used in its creation and the decisions made in sequencing and assembly reaffirm that. It matters what libraries are used, what methods are used in sequencing and assembly, and what is or is not targeted for special treatment to refine sequence quality. This is even more the case for annotation. Annotation is affected by the prior steps, but in turn, what is annotated can feed back to further develop the assembly. It will also affect what the genome can be used for. The model of distributed community annotation—involving individuals, laboratories and groupings of researchers interested in genes with particular hypothesised functions—guided the annotation of the pig genome towards those regions deemed useful for proximate research purposes. In terms of the allotting of work, there was a similarity with the yeast genome sequencing network, though for the pig it was less hierarchical and comprehensive, and more discretionary.

The activities of the SGSP more generally, and IRAG and the X and Y chromosome sequencing more specifically, involved a wider set of actors, approaches and interests than the IHGSC. IRAG involved members of an existing community of pig genomicists that dated back to at least the 1990s. The project to sequence the X and Y chromosomes, though, showed how that community still had the ability to form new connections.

While the scale, speed and automation of sequencing operations had all increased at the Sanger Institute, this did not intensify the tendency we observed in the IHGSC effort: the narrowing of participation and the concentration of operations in-house (Chap. 4). Indeed, the Sanger Institute, and in particular the HAVANA group, opened out to and engaged with a specific external community to develop new genomic resources, tools and expertise through the assembly and annotation activities of the SGSP, IRAG and the X-Y project. That community shaped the direction of various aspects of the sequencing process, in so doing affecting the nature of the product. In turn, the Sanger Institute, at a time in which it was adjusting to the period following the 'completion' of the human reference sequence and each chromosome in turn, itself changed the way it worked.

In considering how the Sanger Institute and the pig genomics community shaped their emerging community annotation strategy and practices, we observe that the cottage industry model (Stein, 2001) needed to be implemented and combined with factory-style approaches. These genomicists, therefore, deployed modes of annotation regarded as characteristic of earlier 'pre-genomic' stages, in conjunction with the concentrated factory style that came to dominate the sequencing of the human reference genome. This challenge to the idea of progression through distinct and separate models and stages of activity, is an important historiographical consequence of our account of pig genome annotation.

As well as helping to re-shape the way that HAVANA operated, the work of pig genome annotation fed into the processes of assembly, automated annotation and indeed manual annotation itself. This was enabled by the temporality of annotation that existed in the pig genome project, with manual annotation occurring alongside ongoing assembly. The manual annotation was therefore able to help correct the assembly as well as contributing to the improvement of automated prediction algorithms. The pig genome community conceived the genome they were helping to produce as provisional and incomplete; their attitude was one of satisficing (on satisficing, see Wimsatt, 2007).

Of course, as we see at the outset of the following chapter, reference genomes are never complete; they are always subject to changes intended to improve their quality and utility. But the pig genome community did not hold an ideal of completeness or comprehensiveness to be paramount in the creation of the first reference assemblies. In one respect, they shared this attitude with Celera. For Celera, the very provisionality of their human

sequence was its selling point; it was important that the publicly-available data it had released in 2001 quickly became outmoded, and that it was widely known to be so. This was to make access to the continually-improved genome and associated data that they held behind a paywall more valuable to potential subscribers. It was this commercial strategy, along with the model of OMIM and their experiences with *Drosophila* sequencing and annotation, that encouraged Celera to forge collaborations with medical geneticists who had been peripheral to the IHGSC.

We have shown that distinctions between manual and automated annotation, annotation and assembly, and functional and structural annotation should all be qualified. In the next chapter, we demonstrate something analogous as we explore the changing relationship between the functional and systematic genomic research that followed the initial sequencing and annotation of the reference genomes of our three species.

References

Agar, J. (2012). *Science in the twentieth century and beyond*. Polity Press.

Agar, J. (2020). What is science for? The Lighthill report on artificial intelligence reinterpreted. *The British Journal for the History of Science, 53*(3), 289–310.

Baker, K. S., & Millerand, F. (2010). Infrastructuring ecology: Challenges in achieving data sharing. In J. N. Parker, N. Vermeulen, & B. Penders (Eds.), *Collaboration in the new life sciences* (pp. 111–138). Routledge.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., et al. (2004). An overview of Ensembl. *Genome Research, 14*, 925–928.

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise. *Genome Research, 14*, 988–995.

Bruce, A., & Lowe, J. W. E. (2022). Pigs and Chips: The making of a biotechnology innovation ecosystem. *Science & Technology Studies*. https://doi.org/10.23987/sts.111111

Bussey, H., Kaback, D. B., Zhong, W.-W., Vo, D. T., Clark, M. W., Fortin, N., et al. (1995). The nucleotide sequence of chromosome I from *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America, 92*(9), 3809–3813.

Collins, J. E., Goward, M. E., Cole, C. G., Smink, L. J., Huckle, E. J., Knowles, S., et al. (2003). Reevaluating human gene annotation: A second-generation analysis of chromosome 22. *Genome Research, 13*, 27–36.

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J., & Clamp, M. (2004). The Ensembl automatic gene annotation system. *Genome Research, 14*, 942–950.

Dawson, H. D., Guidry, C. A., Vangimalla, V., & Urban, J. F., Jr. (2007). The Beltsville Human Nutrition Research Center's porcine immunology and nutrition resource database. *The FASEB Journal, 21*(5), A377.

Dawson, H. D., Loveland, J. E., Pascal, G., Gilbert, J. G. R., Uenishi, H., Mann, K. M., et al. (2013). Structural and functional annotation of the porcine immunome. *BMC Genomics, 14*, 332.

Day, A. E., Quilter, C. R., Sargent, C. A., & Mileham, A. J. (2003). Chromosomal mapping, sequence and transcription analysis of the porcine fertilin beta gene (*ADAM2*). *Animal Genetics, 34*, 375–378.

de Chadarevian, S. (2004). Mapping the worm's genome. Tools, networks, patronage. In J.-P. Gaudillière & H.-J. Rheinberger (Eds.), *From molecular genetics to genomics: The mapping cultures of twentieth-century genetics* (pp. 95–110). Routledge.

Deloukas, P., Matthews, L. H., Ashurst, J., Burton, J., Gilbert, J. G., Jones, M., et al. (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature, 414*, 865–871.

Deloukas, P., Earthrowl, M. E., Grafham, D. V., Rubenfield, M., French, L., Steward, C. A., et al. (2004). The DNA sequence and comparative analysis of human chromosome 10. *Nature, 429*, 375–381.

Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R., & Stein, L. (2001). The Distributed Annotation System. *BMC Bioinformatics, 2*, 7.

Dujon, B. (1996). The yeast genome project: What did we learn? *Trends in Genetics, 12*(7), 263–270.

Dujon, B., Alexandraki, D., André, B., Ansorge, W., Baladron, V., Ballesta, J. P., et al. (1994). Complete DNA sequence of yeast chromosome XI. *Nature, 369*, 371–378.

Dunham, A., Matthews, L. H., Burton, J., Ashurst, J. L., Howe, K. L., Ashcroft, K. J., et al. (2004). The DNA sequence and analysis of human chromosome 13. *Nature, 428*, 522–528.

Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E., et al. (1999). The DNA sequence of human chromosome 22. *Nature, 402*, 489–495.

Feldmann, H., Aigle, M., Aljinovic, G., André, B., Baclet, M. C., Barthe, C., et al. (1994). Complete DNA sequence of yeast chromosome II. *The EMBO Journal, 13*(24), 5795–5809.

Galibert, F., Alexandraki, D., Baur, A., Boles, E., Chalwatzis, N., Chuat, J. C., et al. (1996). Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome X. *The EMBO Journal, 15*(9), 2031–2049.

García-Sancho, M. (2012). *Biology, computing, and the history of molecular sequencing: From proteins to DNA, 1945–2000.* Palgrave Macmillan.

García-Sancho, M., Leng, R., Viry, G., Wong, M., Vermeulen, N., & Lowe, J. W. E. (2022). The Human Genome Project as a singular episode in the history of genomics. *Historical Studies in the Natural Sciences, 52*(3), 320–360.

García-Sancho, M., Lowe, J. W. E., Viry, G., Leng, R., Wong, M., & Vermeulen, N. (2022). Yeast sequencing: 'Network' genomics and institutional bridges. *Historical Studies in the Natural Sciences*, *52*(3), 361–400.

Gregory, S. G., Barlow, K. F., McLay, K. E., Kaul, R., Swarbreck, D., Dunham, A., et al. (2006). The DNA sequence and biological annotation of human chromosome 1. *Nature, 441*, 315–321.

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature, 491*, 393–398.

Harrow, J. L., Steward, C. A., Frankish, A., Gilbert, J. G., Gonzalez, J. M., Loveland, J. E., et al. (2014). The Vertebrate Genome Annotation browser: 10 years on. *Nucleic Acids Research, 42*, D771–D779.

Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H. S., et al. (2000). Chromosome 21 mapping and sequencing consortium. The DNA sequence of human chromosome 21. *Nature, 405*, 311–319.

Heilig, R., Eckenberg, R., Petit, J. L., Fonknechten, N., Da Silva, C., Cattolico, L., et al. (2003). The DNA sequence and analysis of human chromosome 14. *Nature, 421*, 601–607.

Heumann, K., Harris, C., & Mewes, H. W. (1996). A top-down approach to whole genome visualization. *ISMB-96 proceedings*, 98–108. Retrieved December 18, 2022, from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=21e9d845b94393d0d8371452cd8c2e61cb6a9581

Hilgartner, S. (2017). *Reordering life: Knowledge and control in the genomics revolution*. The MIT Press.

Hillier, L. (2012). Interview conducted over telephone by Kathryn Maxson, Robert Cook-Deegan, 5 April 2012. Retrieved December 18, 2022, from https://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/7701/2012%2005%20April%20LaDeana%20Hillier%20interview.pdf?sequence=1&isAllowed=y

Hillier, L. W., Fulton, R. S., Fulton, L. A., Graves, T. A., Pepin, K. H., Wagner-McPherson, C., et al. (2003). The DNA sequence of human chromosome 7. *Nature, 424*, 157–164.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., et al. (2020). Ensembl 2021. *Nucleic Acids Research, 49*(D1), D884–D891.

Humphray, S. J., Oliver, K., Hunt, A. R., Plumb, R. W., Loveland, J. E., Howe, K. L., et al. (2004). DNA sequence and analysis of human chromosome 9. *Nature, 429*, 369–374.

Johnston, M., Andrews, S., Brinkman, R., Cooper, J., Ding, H., Dover, J., et al. (1994). Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. *Science, 265*, 2077–2082.

Kerlavage, A., Bonazzi, V., di Tommaso, M., Lawrence, C., Li, P., Mayberry, F., et al. (2002). The Celera Discovery System™. *Nucleic Acids Research, 30*(1), 129–136.

Kollers, S., Day, A., & Rocha, D. (2006). Characterization of the porcine *FSCN3* gene: cDNA cloning, genomic structure, mapping and polymorphisms. *Cytogenetic and Genome Research, 115*, 189–192.

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. The University of Chicago Press.

Leonelli, S., & Tempini, N. (Eds.). (2020). *Data journeys in the sciences*. Springer Nature.

Loveland, J. E., Gilbert, J. G. R., Griffiths, E., & Harrow, J. L. (2012). Community gene annotation in practice. *Database*, 2012, bas009.

Lowe, J. W. E. (2018). Sequencing through thick and thin: Historiographical and philosophical implications. *Studies in History and Philosophy of Biological and Biomedical Sciences, 72*, 10–27.

Lowe, J. W. E. (2022). Humanising and dehumanising pigs in genomic and transplantation research. *History and Philosophy of the Life Sciences*, *44*, 66.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biology, 20*, 246.

Mewes, H.-W., Frishman, D., Zollner, A., & Heumann, K. (1998). The bioinformatics of the yeast genome. In A. J. P. Brown & M. Tuite (Eds.), *Methods in microbiology. Volume 26: Yeast gene analysis* (pp. 33–51). Academic Press.

Quilter, C. R., Blott, S. C., Mileham, A. J., Affara, N. A., Sargent, C. A., & Griffin, D. K. (2002). A mapping and evolutionary study of porcine sex chromosome genes. *Mammalian Genome, 13*, 588–594.

Renard, C., Hart, E., Sehra, H., Beasley, H., Coggill, P., Howe, K., et al. (2006). The genomic sequence and analysis of the swine major histocompatibility complex. *Genomics, 88*, 96–110.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., et al. (2005). The DNA sequence of the human X chromosome. *Nature, 434*, 325–337.

Schook, L. B., Beever, J. E., Rogers, J., Humphray, S., Archibald, A., Chardon, P., et al. (2005). Swine Genome Sequencing Consortium (SGSC): A strategic roadmap for sequencing the pig genome. *Comparative and Functional Genomics, 6*(4), 251–255.

Searle, S. M. J., Gilbert, J., Iyer, V., & Clamp, M. (2004). The Otter annotation system. *Genome Research, 14*, 963–970.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature, 423*, 825–837.

Skinner, B. M., Sargent, C. A., Churcher, C., Hunt, T., Herrero, J., Loveland, J. E., et al. (2016). The pig X and Y chromosomes: Structure, sequence, and evolution. *Genome Research, 26*, 130–139.

Star, S. L., & Bowker, G. C. (2002). How to infrastructure? In L. A. Lievrouw & S. Livingstone (Eds.), *The handbook of new media: Social shaping and consequences of ICTs* (pp. 151–162). Sage.

Stein, L. (2001). Genome annotation: From sequence to biology. *Nature Reviews Genetics, 2*, 493–503.

Stevens, H. (2013). *Life out of sequence: A data-driven history of bioinformatics.* The University of Chicago Press.

Strasser, B. J. (2019). *Collecting experiments: Making big data biology.* The University of Chicago Press.

Uenishi, H., Morozumi, T., Toki, D., Eguchi-Ogawa, T., Rund, L. A., & Schook, L. B. (2012). Large-scale sequencing based on full-length-enriched cDNA libraries in pigs: Contribution to annotation of the pig genome draft sequence. *BMC Genomics, 13*, 581.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science, 291*(5507), 1304–1351.

Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., et al. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. *GigaScience, 9*(6), giaa051.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality.* Harvard University Press.

Zody, M. C., Garber, M., Adams, D. J., Sharpe, T., Harrow, J., Lupski, J. R., et al. (2006). DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature, 440*, 1045–1049.