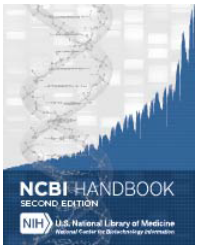
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Ostell J. What's in a Genome at NCBI?. 2013 Nov 8. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



What's in a Genome at NCBI?

James Ostell, PhD[1]

Created: November 8, 2013.

Scope

It seems like a simple request to a bioinformatics center like NCBI—"Download the human genome", "Display the *HIV-1* genome"—and yet this is a complicated question in terms of biology, experimental data, the current state of knowledge, and the use to which a particular scientist may wish to put the data. This chapter provides an introduction to those questions, a brief history of genome representations at NCBI as the state of the science evolved over the last few decades, and a summary of some of the many resources and tools that are relevant to Genomes at NCBI.

What is a Genome?

Biologically speaking, of course, a "genome" of an organism is the complement of genes that make up the organism. However, this is already an abstraction. Traditionally "genes" are heritable units that manifest some observable trait over related generations of an organism. It was only later we discovered they are pieces of DNA contained in a larger DNA strand comprising a chromosome, leading to the current assumption that a genome is the set of DNA strands of the chromosomes. However, we already know this isn't quite right, because any particular individual from whom we measured the DNA may not be typical in every gene, in fact, may be completely missing some genes that are found in other individuals of the same species. In addition, there may be DNA sequences that are found in NO real organism (because they are chimeras from many individuals or because they contain errors), which are nonetheless the standard reference for a particular community, and thus the sequence they expect to find when they ask for e.g., "the human mitochondrial genome" (also known as "the revised Cambridge Reference Sequence") (1, 2).

In addition to these biological and historical issues, science is a moving target. At various points in time, the genome data for a particular organism may be incomplete to varying degrees, in which case the best "genome" available may still have unique difficulties. Over time as technology changes, a different set of data may become a better approximation for the biological reality, then the community may suddenly find itself getting a different answer for "download the genome" than they got last time.

Finally, there are many contexts for use of "the genome." For many types of research, "the genome" doesn't mean DNA at all but instead means the set of transcripts or protein sequences coded for by that organism's genome. In other cases, for example medical genetics, it means not only DNA, but a very specific piece of reference standard DNA that is used by the medical community to record single base changes at specific locations (3). This important piece of DNA may only cover one gene, and for that length, it may not be identical to the commonly used complete "human genome" sequence in that region. Still other contexts involve investigations of large and

Author Affiliation: 1 NCBI.

small sequence variation across a population of individuals, by browsing or downloading data for the human 1000 Genomes project (4), monitoring sequence variation during an influenza outbreak (5), or comparative analysis datasets as provided in HomoloGene (6), Protein Clusters (7), or the viral Pairwise Sequence Comparison (PASC) tool (8).

So, NCBI's efforts to provide a response for "give me the genome of my research organism" has necessarily had to change over time, may be different for different organisms, and may bring complications and nuances that surprise users who may only be familiar with their particular view of what "the genome" is.

History

The history of genomes at NCBI is intimately tied to the history of GenBank and RefSeq. GenBank is part of the International Nucleotide Sequence Database Collaboration (INSDC) (9) of which NCBI is the US member, which collects annotated nucleotide sequences from contributing scientists, typically when they publish the sequence in a journal article. As such, GenBank is like the primary research literature. Each "article" or sequence entry represents the view of the contributing author at the time. The database staff does not curate the sequence records for correctness other than conformance to standards and internal consistency. The biological validity of the record is reflected in the peer review process of the published article. Just as with the primary literature, articles can go out of date or later articles may contradict, or even invalidate, a previous article. There may be legitimate disagreement about which view is more correct. But this evolution of scientific knowledge over time is normal for any experienced scientist reading the literature, and it is normal and healthy for GenBank.

NCBI's initial expectation was that we would simply identify the GenBank record which was currently the most widely accepted genome sequence for a particular organism when someone requested to view or download "the genome." However, as early as the 1990's it was clear there were problems with this. At that time the *HIV-1* "genome" had been published (in fact several times) in GenBank. The problem was that these were the first sequences, and they were partial in different ways, in particular missing the long terminal repeat (LTR). So, scientists working on *HIV-1* had developed and were sharing a "standard" *HIV-1* genome by editing and combining the partial GenBank sequences into one complete genome, but it had never been deposited into GenBank. So, it was not possible to provide any single GenBank record in response to the request "download the *HIV-1* genome." In an attempt to remedy this problem, NCBI staff worked with the authors of an authoritative book on retroviruses, to both deposit the retrovirus genome sequences into GenBank (e.g., AF033819.3), and to cite those sequences when referring to the genomes in the RETROVIRUSES book, which is available on the NCBI Bookshelf. While this solution continued to use GenBank as the vehicle for authoritative genomes, it was no longer a pure model of the data coming voluntarily from the scientific community. NCBI had a guiding and active role to make it happen.

Shortly after that the yeast genome was being completed chromosome by chromosome. Each chromosome was sequenced by a different US or European group, and published as a submission of a chromosome, a chromosome arm, or a collection of contigs and scaffolds. It was only at the time of publication that the chromosome sequence was deposited and shared through GenBank. This was a long slow process, especially when the work on one chromosome might be the result of the work by many separate labs working at different rates. During this time, it was only possible for NCBI to provide a partial answer from GenBank to "download the yeast genome." Once the whole genome was completed, however, much work, especially in annotating the genes, but also in correcting sequence problems, continued. The scientists in the US who did the initial work on half of the yeast chromosomes "ceded" their GenBank records to a single database group, the Saccharomyces Genome Database (SGD) (10). This gave SGD the right to update "their" GenBank records as annotation improvements were made and sequence problems corrected. SGD worked with NCBI to make this happen, and that half of the yeast genome was keep current and up to date in GenBank. Unfortunately, the European half of the genome was not held under this model, so even though annotation improvements and sequence updates were being made to

those chromosomes, they were not being deposited into GenBank. The result, again, was the NCBI was unable to respond with a complete up-to-date set of records when someone wanted to "download the yeast genome." After much negotiation to make this approach work through GenBank, it became clear that half of the yeast genome in GenBank would always be out of date. In this same time-frame, the human genome project was actively underway and again sequencing was completed and submitted chromosome by chromosome but by this time the "Bermuda Principles" (11) had been established and data was now starting to be submitted in advance of peer-reviewed publication. The international community was actively discussing the human gene count and planning for genome annotation; during this time several groups attempted to define the gene content before the sequencing was completed (12, 13).

Thus, the climate at that time with regards to both data submission and scientific discussion were instrumental to the decision that NCBI needed a new database, called RefSeq (14). Just as you may think of GenBank as equivalent to the primary research literature written by many authors, RefSeq could be considered the "NCBI review article" on genomes. Unlike GenBank, it reflects the judgment of a single group, NCBI, about what the most useful sequences are to represent a particular genome and its products, typically in collaboration or consultation with experts wherever possible. Like a review article it is drawn from the primary literature and archival sequence databases, but it may be aggregated, edited, or reorganized by NCBI to represent a better summary overview in light of current knowledge. RefSeq included the *HIV-1* genome from GenBank with no annotation or sequence changes compared to the GenBank record. But it could also now include human transcripts and proteins (organized by genes in LocusLink (14), the precursor to NCBI's Gene resource), in preparation for annotating the human genome. As for the *Saccaromyces cerevisiae* genome, for many years the RefSeq annotated genome—which was taken entirely from SGD so that it would be complete, consistent, and up to date—was the only way that NCBI could answer the request to view or download all chromosomes of the yeast genome. Much more recently, the complete yeast genome was also made available, by SGD, as a Third Party Annotation submission so that the current up-to-date version is also available at all member databases of the INSDC.

The RefSeq initiative allowed NCBI to start building representations of genomes that might vary considerably in how they were built or the sources they came from. For example, when the first bacterial genomes were sequenced, NCBI took a chromosome-centric view of the organism because we had the whole DNA sequence, but not much information on transcripts or proteins. In contrast, for human, at the time there were many cDNA sequences of transcripts from individual genes, but only relatively short stretches of chromosomal DNA for a few regions. So "the human genome" in RefSeq was actually the most comprehensive collection of cDNAs we could aggregate and curate at the time. During the early and middle phases of sequencing and assembling the human genome, the cDNA sequences were still the highest quality data for the genes, but the assembled chromosome fragments started filling in the intragenic regions and providing long range order. In this version, "download the human genome" would get chromosomal sequence and aligned cDNAs, where the sequence of the cDNA may not match exactly the sequence of chromosomal DNA. Because the cDNA is more reliable, the protein from the coding region is derived from the cDNA, not the chromosome. Today, the chromosomal sequence of human is very good quality (although not without flaws), and NCBI has gone to considerable lengths to ensure that the sequence of chromosome and the cDNA do match, sometimes correcting the chromosome, sometimes correcting the cDNA, in collaboration with sequencing laboratories and other scientists.

While "the human genome" is now very consistent across chromosomal DNA and cDNA, there remain cases where there are exceptions, such as when the human genome represents a rare allele or a base that is suspected of being an error. NCBI is one of the collaborators in the Genome Reference Consortium (15) which is actively involved in ongoing maintenance and improvement of the reference human genome sequence. NCBI's RefSeqGene project (16) is another case where there may be differing definitions of the reference genomic sequence. RefSeqGene provides a collection of chromosomal DNA regions, in chunks covering single genes, which are intended for use in clinical genetics. Since the traditional sequence for some genes predates the

complete human genome sequence, or where the gene in the complete human genome may not be a common allele or traditional allele, the RefSeqGene record may not be identical to "the human genome." RefSeqGene genomic records are aligned, as with the cDNAs in the middle period of human genome sequencing, to the human genome to support (using the NCBI Genome Remapping Service) mapping coordinates and sequence back and forth between the two. NCBI has made every effort to make the RefSeqGene identical to the human genome when it can support the needs of the medical reporting community, but it is not totally consistent. So the "complete human genome" in the research world may not be exactly the same thing as "the complete human genome" in the clinical world.

The evolution continues as we start to accumulate many genomes for a single organism. If one asks for the "*Salmonella* Genome" today, the question is which strain of *Salmonella* do you really want? In some cases you want the well annotated typical genome, but in other cases you may specifically want the one from "the Montevideo outbreak." Even with humans, we know that some humans contain genes not contained in other humans and vice versa. So by "the human genome" do you want a single example of a real human genome? Or do you want that single example, plus the additional genes that are found in other humans, to get the full complement of possible human genes?

Finally, in some aspects, RefSeq is coming full circle back to GenBank. NCBI has gone to considerable efforts to persuade the scientific community to keep GenBank up to date directly themselves, so that RefSeq can again be simply a selection of particular records. One case in point is Drosophila. The *Drosophila melanogaster* genome records were "ceded" to the FlyBase database by the original sequencing team. Unlike yeast, in this case FlyBase "owns" all the chromosomes. NCBI has worked closely with FlyBase both to provide computational support and evidence for building and validating the gene models, and for facilitating the update of the GenBank records from the FlyBase database, and maintaining the RefSeq records from those GenBank records. FlyBase provides valuable communication within the Drosophila community and manual curation of gene models by experts in that organism.

RefSeq necessarily continues to contain "genomes" for different organisms done in a variety of possible ways, with a variety of possible consequences. But for the goal of comprehensiveness, one necessarily pays the price of complexity. NCBI makes every effort to provide simple, intuitive views of genomes where possible, while still hinting at the additional layers and nuances to the genome concept so users who may need that more sophisticated view are aware it exists.

Resources, Tools, and Access

NCBI's early commitment to reliably and robustly support the simple request to download, view, or analyze a genome resulted in a large suite of resources, tools, and shareable code-base (via NCBI toolkit libraries) that range from broadly scoped multi-kingdom resources such as RefSeq, Gene, and Genome, and eukaryotic and prokaryotic genome annotation pipelines—to niche resources such as Viral Variation or CloneDB. Some resources reflect natural organizing principles of genomic data and support access from a gene- or organism-centric perspective, whereas others were developed in response to a particular disease outbreak (NCBI FLU resource) or based on collaboration, community feedback or requests (e.g., International Standards for Cytogenomics Array, HIV-1:human protein interactions, Conserved CDS database). Along the way we also developed a suite of viewing platforms including Map Viewer, the Graphical sequence viewer (for example, the RefSeq *Escherichia coli* genome record NC_000913.3), and a standalone, multiplatform, downloadable graphical user interface (GUI) Genome Workbench. NCBI continues to be fully committed to supporting access to genome data and several of our newer resources—BioProject, BioSample, and Assembly—describe aspects of the research project, the biological sample, and how the genome assembly is organized.

References

1. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290(5806):457–65. PubMed PMID: 7219534.
2. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature genetics. 1999;23(2):147. PubMed PMID: 10508508.
3. Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, et al. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med. 2010;2(4):24. PubMed PMID: 20398331.
4. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73. PubMed PMID: 20981092.
5. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the National Center for Biotechnology Information. Journal of virology. 2008;82(2):596–601. PubMed PMID: 17942553.
6. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2013;41(Database issue):D8–D20. PubMed PMID: 23193264.
7. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, et al. The National Center for Biotechnology Information's Protein Clusters Database. Nucleic acids research. 2009;37(Database issue):D216–23. PubMed PMID: 18940865.
8. Bao Y, Chetvernin V, Tatusova T. PAirwise Sequence Comparison (PASC) and its application in the classification of filoviruses. Viruses. 2012;4(8):1318–27. PubMed PMID: 23012628.
9. Nakamura Y, Cochrane G, Karsch-Mizrachi I.; The International Nucleotide Sequence Database Collaboration. Nucleic acids research. 2013;41(Database issue):D21–4. PubMed PMID: 23180798.
10. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research. 2012;40(Database issue):D700–5. PubMed PMID: 22110037.
11. Marshall E. Bermuda rules: community spirit, with teeth. Science (New York, NY. 2001;291(5507):1192.
12. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, et al. A physical map of 30,000 human genes. Science (New York, NY. 1998;282(5389):744-6.
13. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. Nature genetics. 2000;25(2):239–40. PubMed PMID: 10835646.
14. Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. 2000;16(1):44–7. PubMed PMID: 10637631.
15. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011;9(7):e1001091. PubMed PMID: 21750661.
16. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic acids research. 2012;40(Database issue):D130–5. PubMed PMID: 22121212.