Reference Genomes

A Molecular Foundation for Modern Biology

Applications and Impacts of Reference Genomes

Once upon a time, researchers believed that sequencing the human genome—or any genome, really—would unlock all of our biological secrets. Today we know better. "The genome is only the beginning," goes the familiar refrain. And in most cases, obtaining a reference genome *is* the beginning.

As its name suggests, a reference genome is essentially the working genome sequence of a given organism—a molecular yardstick against which other sequences can be measured (though it may in fact represent the contributions of several distinct individuals). From resequencing to trait mapping, phylogenenetics to microbial forensics, reference genomes provide the foundation upon which future genomics datasets can build. After all, how can you discern variation without a reference against which to compare it?

Applications run the gamut from basic science to medicine to agriculture. Researchers at the University of Michigan and in China used the draft genome of the giant panda to probe its switch from a carnivorous to herbivorous lifestyle, while researchers in the United States and Mexico are using the maize reference genome to study the plant family's genetic diversity and domestication. At the University of California, Berkeley, researchers used 27 reference genomes to study the evolution of social behavior in the honeybee, and an international team of scientists sequencing the domestic turkey (*Meleagris gallopavo*) used the chicken and zebra finch genomes to address the evolution of birds.

From St. Louis to Boston, Cambridgeshire to Shenzhen, sequencing factories armed with dozens of state-of-the-art next generation instruments are cranking out sequences to the tune of terabases—the equivalent of several hundred human genomes—every day. Much of that work revolves around reference genomes, whether de novo sequencing of an organism for the first time, or resequencing to identify diversity and map loci of interest.

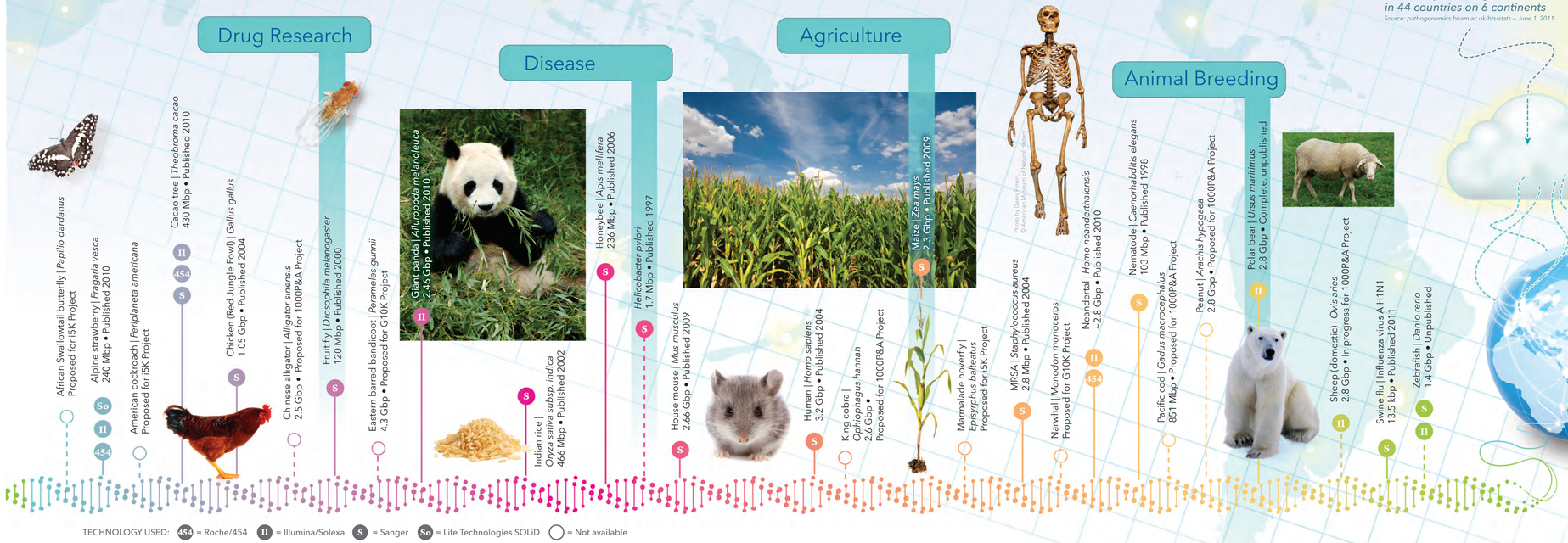
To date, researchers have compiled complete reference genomes for well over 1,000 organisms (see graphic below). The NCBI's Genome Project lists 1,613 complete microbial sequences, plus 40 eukaryotes. Another 1,121 eukaryotic projects are listed as either draft assemblies or works-in-progress.

Their potential impact could be wide reaching. Consider the human reference genome, which has provided the molecular foundation for studies of human variation (for instance, the 1000 Genomes Project), disease gene mapping (Miller syndrome), and evolution (comparative genomics of the Neandertal genome).

In agriculture, reference genomes could drive the development of drought- or pest-resistant plants as well as more robust and nutritious crops. Animal breeders can mine reference genomes to better understand inter- and intra-species diversity, while drug developers can better stratify patient populations in clinical trials and, ultimately, for the delivery of more personalized therapies.

First, though, researchers need to sequence the genomes. Driven by technology improvements and falling costs, large-scale efforts to do just that are under way. For the broader scientific community, such efforts will be invaluable. And, like the human genome itself, they are only the beginning.

1,608 instruments worldwide in
529 sequencing centers
in 44 countries on 6 continents
Source: pathogenomics.bham.ac.uk/hts/stats – June 1, 2011



Drowning in Data

Between falling per-base costs and ever-more-widespread adoption of next generation sequencing instrumentation, sequencing a DNA genome is no longer as daunting. As the Genome 10K Community of Scientists wrote in 2009, "The bold insight behind the success of the human genome project was that, although vast, the roughly 3 billion letters of digital information specifying the total genetic heritage of an individual is finite and might, with dedicated resolve, be brought within the reach of our technology."

All those As, Cs, Gs, and Ts have to be quality controlled, stored, analyzed, copied, annotated, and compared

That doesn't mean the challenges are gone. Suppose your organism's genome contains three billion base pairs. That's six billion nucleotides when you factor in both maternal and paternal chromosomes. Yet it's not enough to read each of those bases only once; between error rates, experimental biases, and sequence polymorphisms, the sequence must be read again and again to make sure it's correct, like a copyeditor proofreading a document. A complete sequencing effort might require reading each base an average of, say, 20 to 40 times (that is, to 20x to 40x read depth). To saturate the genome—to read every base confidently—you're looking at sequencing on average at least 120 billion bases per individual. All those As, Cs, Gs, and Ts have to be quality controlled, stored, analyzed, copied, annotated, and compared—all of which takes information technology infrastructure, storage space, and considerable human and financial resources. How many resources? There's no hard and fast rule, but a good rule of thumb is: For every dollar spent on sequencing hardware, expect to spend at least that amount on informatics.

Base Count

Center	Output	Number of Sequencers: 3730	HS	GA2	SOLiD	454	PacBio	IT
BGI	~5.6 T/day	15	137		27			
Wellcome Trust Sanger Institute	~0.3 T/day	10	20	9		2	1	2
The Genome Institute at Washington University	~4.9 T/day	15	30	26		8	1	1

KEY: T = terabases; 3730 = Applied Biosystems/Life Technologies 3730 capillary sequencer; HS = Illumina HiSeq 2000; GA2 = Illumina Genome Analyzer II; SOLiD = Life Technologies SOLiD; 454 = Roche/454 Life Sciences Genome Sequencer; PacBio = Pacific Biosciences PacBio RS; IT = Ion Torrent/Life Technologies Personal Genome Machine

It Takes a Village

The typical data workflow in reference genome construction involves three steps: assembly (aligning the short sequencing reads into "contigs"), annotation (gene and regulatory element identification and function prediction), and analysis (such as phylogenetic analysis or whole-genome alignment).

Dedicated sequencing centers have sizable staffs and resources dedicated to these informatics tasks. Yet between new sequencing technologies and evolving analytical tools, even these experts can have trouble staying current. For instance, the SEQanswers software wiki (SEQanswers.org/wiki/SEQAnswers) lists some 55 programs dedicated to the task of genome assembly alone (out of 409 applications overall). The problem is such that in late 2010 the bioinformatics community held the first "Assemblathon," a sort of genome assembly competition whose goal is to encourage scientists to produce novel computational solutions that address informatics challenges. A second Assemblathon is slated for mid-2011.

For the rest of the scientific community, there's no shortage of free tools, from single-purpose software like BGI's SOAPdenovo assembler, to integrated suites like Penn State University's Galaxy, to do-it-yourself virtual bioinformatics machines like CloudBioLinux. Or, for a more polished approach, there are fee-based cloud services like DNAnexus. Either way, the research community can help fill the knowledge gaps.

Indeed, in the genomics era, it takes a village. Efforts such as SEQanswers.com provide a forum for debate and answering questions, while organism-specific databases like WormBase, FlyBase, SilkDB, and The Arabidopsis Information Resource (TAIR) provide a central online meeting hall for an organism's research community to collectively probe, annotate, and aggregate everything from SNPs and RNAi data to gene expression and literature resources. The Ensembl project has built genome databases for over 50 organisms, from alpaca to zebrafish. Now BGI is getting in on the act, with a plan to automatically generate and seed databases for every organism it sequences.

Online Resources

1000 Genomes Project – 1000genomes.org
NCBI Genome Project – ncbi.nih.gov/genomeprj
Wormbase – wormbase.org
Flybase – flybase.org
SilkDB – silkworm.genomics.org.cn
TAIR – arabidopsis.org
SEQanswers – seqanswers.com
Ensembl – ensembl.org

Selected Genome Sequencing Projects

Many one-off genome projects are under way, but a few particularly ambitious ones aim to sequence multiple organisms. Here we highlight four such efforts.

Genome 10K Project (G10K)

The G10K Project will sequence some 16,203 vertebrate genomes, a "genomic zoo" representing at least one member of each vertebrate genus. The first 101 species have already been announced.

genome10k.soe.ucsc.edu

1000 Plant & Animal (1000P&A) Reference Genomes Project

The BGI-initiated project focuses on "1,000 economically and scientifically important plant/animal species." Fifty have been completed, another 100 are in progress.

ldl.genomics.cn/page/pa-research.jsp

5,000 Insect (i5K) and Other Arthropod Genome Initiative

The Arthropod Genomic Consortium will target 5,000 insects with agricultural, medical, or research significance. So far, 76 have been proposed.

arthropodgenomes.org/wiki/i5K

Ten Thousand Microbial (10K M) Genomes Project

This BGI-led project is sequencing microbes from habitats as diverse as earth, air, glaciers, and hot springs. Their goal is the development of a genomic encyclopedia of microorganisms in China. Over 1,200 have been completed to date.

ldl.genomics.cn/page/M-research.jsp

Sponsored by BGI

Produced by the *Science*/AAAS Business Office

华大基因 BGI Premier Scientific Partner

- de novo
- Whole Genome Resequencing
- Exome Sequencing
- Target Region Sequencing
- RNA-Seq
- Epigenomics
- Metagenomics
- Proteomics

de novo

de novo sequencing aims to sequence a species afresh from the beginning without referencing any previous sequencing data of the species. Based on the genome characteristics, two different strategies are used to obtain the whole genome map.

Genome Characteristics	Common Genome*	Complex Genome**
Chromosomes ploidy	Haploid or homozygous diploid	Heterozygous diploid or polyploid
Heterozygosity rate	< 0.5%	> 0.5%
GC content	35% ~ 65%	< 35% or > 65%
Repeats content	< 50%	> 50%

★ All conditions need to be fulfilled. ★★ At least one of the conditions need to be fulfilled.

Strategy for Common Genomes

Strategy for Complex Genomes

Bioinformatics Analysis

- Genome analysis: GC content, sequence depth distributions, evaluation of assembled genome.
- Genome annotation: repeat sequence, ncRNA annotation, gene structure prediction, gene function annotation.
- Comparative genomics and evolution analysis: orthologous gene clusters, phylogenetic analysis, whole genome alignment, segmental duplication, conserved element.

1. Xia QY, Guo YR, *et al*. Science 2009, 326: 433-436. 2. Li RQ, Fan W, *et al*. Nature 2009, 463: 311-317. 3. Huang SW, Li RQ, *et al*. Nature Genetics 2009, 41: 1275-1281.

Sequencing Solutions Using

137 Illumina HiSeq 2000 27 ABI SOLiD 4 System

Super Computing Centers

102T Flops, 20TB Memory, 10PB Storage

Relying on the powerful high-performance computing systems, efficient software and reliable computing services, BGI can provide a full-scale solution for data analysis to help you achieve your research goals.

www.bgisequence.com **tech@genomics.cn**

Reference Genome
A Molecular Foundation for Modern Biology

"If I have seen a little further it is by standing on the shoulders of giants."
– ISAAC NEWTON TO ROBERT HOOKE
February 5, 1676.

Scientific discoveries made completely de novo—out of thin air—are very rare. Almost all research breakthroughs, both big and small, are based on and build upon the work of others. Indeed, the creation of a firm foundation of knowledge is often imperative for scientific inquiry to thrive.

This holds true in the field of genomics, where one strives to tease out how a genome—a relatively simple string of A, G, C, and T bases—can encode all the instructions necessary to create life in all its forms, take the human genome as an example. The sequencing of all 3 billion base pairs of the first human genome was a mammoth undertaking; some were skeptical that it was even possible. Many, however, believed that it was worth pursuing, perhaps because they saw the potential that this knowledge would open up. It's unlikely that even the most optimistic proponents imagined that 20 years later, through advances in sequencing technology and data analysis, what previously took 13 years to complete now takes only a few days.

Completing a draft of the human genome provided the research community with a valuable prize. Not only did analysis of this genome clarify some previously erroneous guesswork, for example, regarding the number of genes encoded by our DNA (closer to 30,000 than the predicted 100,000), but it also provided a much-needed template on which to build future research. The power of having this reference genome became clear as scientists began using it as a baseline to compare and contrast their own sequencing data collected from different populations, or segments of populations (like those with a predisposition to certain diseases). It also allowed us to investigate what makes us human, at least at a genetic level, through comparisons with our closest ape relatives and even with fragments of Neandertal DNA.

Today, as hundreds, perhaps thousands, of human genomes are being sequenced, reference genomes are also being pursued for numerous other creatures with whom we share the planet. The motivators for such research are multifold. Often the drivers are economic, such as understanding pest or pesticide resistance in plants, or medical, as with virulence factors carried by microbes. For others, the motivation is sustainability, for example, enabling the breeding of endangered species, like pandas and polar bears, in captivity.

Whatever the reason, the sequencing of reference genomes provides us with a new way to see life and, through comparisons between and amongst species, offers a deeper and richer understanding of the world we live in.

Sean Sanders, Ph.D.
Commercial Editor, *Science*

Publication date: July 8, 2011

Writer: Jeffrey Perkel, Ph.D. • Illustrator/Designer: Mica Duran, M.S., C.M.I. • Editor: Sean Sanders, Ph.D.

Science AAAS
A Science/AAAS Business Office Publication

Sponsored by 华大基因 BGI