Too Much Data -
Cautionary Tales of Next-generation
and Next-next Generation Sequencing

Dave Ussery
“special talk”
King Mongkut’s University of Technology Thonburi
Bangkok, Thailand

4 March, 2010
Researchers Sequence Genome of Ancient Human

Researchers have reconstructed an ancient human genome, obtained from a 4000-year-old preserved clump of hair in Greenland permafrost. Experts say that similar techniques could be employed in many other ways, such as analysing the DNA of South American mummies or crime victims.

Publications

- Analysis of high-throughput sequencing and annotation strategies for phage genomes
- BAC-HAPPY Mapping (BAP Mapping): A new and efficient protocol for physical mapping
- Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries
- Monitoring genomic sequences during SELEX using high-throughput sequencing

News

- Scientists sequence genomes of Southern African Bushmen and Bantu individuals
- Agricultural scientists sequence genome of grass that can be a biofuel model crop
- Scientists spot genetic 'fingerprints' of individual cancers
- Genome analysis of marine microbe reveals a metabolic minimalist
A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin1*, Ruiqiang Li1*, Jeroen Raes2,3, Manimozhiyan Arumugam2, Kristoffer Solvsten Burgdorf4, Chayavanh Manichanh5, Trine Nielsen6, Nicolas Pons6, Florence Levenez7, Takui Yamada8, Daniel R. Mende9, Junhua Li10, Junming Xu10, Shaohuan Li10, Dongfang Li10, Jianjun Cao10, Bo Wang10, Huiqiang Liang4, Huisong Zheng10, Yilong Xie10, Julian Tap9, Patricia Lepage6, Marcelo Bertalan5, Jean-Michel Batto5, Torben Hansen9, Denis Le Paslier11, Allan Linneberg11, H. Bjørn Nielsen2, Eric Pelletier11, Pierre Renault6, Thomas Sicheritz-Ponten2, Keith Turner11, Hongmei Zhu11, Chang Yu11, Shengting Li11, Min Jian11, Yan Zhou12, Yingrui Li11, Xiuxing Zhang12, Songgang Li11, Nan Qin11, Huanming Yang11, Jian Wang11, Søren Brunak11, Joel Doré1, Francisco Guarner3, Karsten Kristiansen4, Oluf Pedersen6,12, Julian Parkhill3, Jean Weissenbach10, MetaHIT Consortium10, Peer Bork2, S. Dusko Ehrlich2 & Jun Wang1

To understand the impact of gut microbes on human health and well-being, it is crucial to assess their genetic potential. Here, we describe the Illumina-based metagenomic sequencing, assembly, and characterization of 3.3 million non-redundant microbial genes derived from 576.7 gigabases of sequence from fecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genome is bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species each with at least 16 such species, which may also be largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of the functional presence in all individuals and most bacteria, respectively.

It has been estimated that the microbes in our bodies collectively make up to 10.0 trillion cells, tenfold the number of human cells, and they encode 100-fold more unique genes than our own genome. The majority of microbes reside in the gut, have a profound influence on human physiology and nutrition, and are crucial to human life. Furthermore, the gut microbes contribute to energy harvest from food, and changes in gut microbiome may be associated with bowel diseases or obesity. To get a broader overview of the human gut microbial genes, we used Illumina Genome Analyser (GA) technology to carry out deep sequencing of total DNA from fecal samples of 124 European adults. We generated 376.7 Gb of sequence, almost 200 times more than in all previous studies, assembled it into contigs and predicted 3.3 million unique open reading frames (ORFs). This gene catalogue contains virtually all of the prevalent gut microbial genes in our cohort, provides a

The international MetaHIT (Metagenomics of the Human Intestinal Tract) project has published a gene catalogue of the human gut microbiome derived from 124 healthy, overweight and obese human adults, as well as inflammatory disease patients, from Denmark and Spain. The data provide the first insights into this gene set - over 150 times larger than the human gene complement - and permit the definition of both a minimal gut metagenome and a minimal gut bacterial genome. Credit: Roger Harris /Science Photo Library.
Outline

• The problem - too much data!
• A brief history - The speed of sequencing
• Cautionary tales
• Some approaches to handle this....
1. The problem - too much data!

Technology

The data deluge

Businesses, governments and society are only starting to tap its vast potential

Feb 25th 2010 | From The Economist print edition

EIGHTEEN months ago, Li & Fung, a firm that manages supply chains for retailers, saw 100 gigabytes of information flow through its network each day. Now the amount has increased tenfold. During 2009, American drone aircraft flying over Iraq and Afghanistan sent back around 24 years' worth of video footage. New models being deployed this year will produce ten times as many data streams as their predecessors, and those in 2011 will produce 30 times as many.

Everywhere you look, the quantity of information in the world is soaring. According to one estimate, mankind created 150 exabytes (billion gigabytes) of data in 2005. This year, it will create 1,200 exabytes. Merely keeping up with this flood, and storing the bits that might be useful, is difficult enough. Analysing it, to spot patterns and extract useful information, is harder still. Even so, the data deluge is already starting to transform business, government, science and everyday life (see our special report in this issue). It has great potential for good—as long as consumers, companies and governments make the right choices about when to restrict the flow of data, and when to encourage it.
1. The problem - too much data!

Overload
Global information created and available storage
Exabytes

Data inflation

<table>
<thead>
<tr>
<th>Unit</th>
<th>Size</th>
<th>What it means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bit (b)</td>
<td>1 or 0</td>
<td>Short for &quot;binary digit&quot;, after the binary code (1 or 0) computers use to store and process data</td>
</tr>
<tr>
<td>Byte (B)</td>
<td>8 bits</td>
<td>Enough information to create an English letter or number in computer code. It is the basic unit of computing</td>
</tr>
<tr>
<td>Kilobyte (KB)</td>
<td>1,000, or $2^{10}$ bytes</td>
<td>From &quot;kilo&quot; in Greek. A typical pop song is about 4MB</td>
</tr>
<tr>
<td>Megabyte (MB)</td>
<td>1,000KB; $2^{20}$ bytes</td>
<td>From &quot;large&quot; in Greek. The complete works of Shakespeare total 5MB. A typical movie is 1GB</td>
</tr>
<tr>
<td>Gigabyte (GB)</td>
<td>1,000MB; $2^{30}$ bytes</td>
<td>From &quot;giant&quot; in Greek. A two-hour film can be compressed into 1-2GB</td>
</tr>
<tr>
<td>Terabyte (TB)</td>
<td>1,000GB; $2^{40}$ bytes</td>
<td>From &quot;monster&quot; in Greek. All the catalogued books in America's Library of Congress total 15TB</td>
</tr>
<tr>
<td>Petabyte (PB)</td>
<td>1,000TB; $2^{50}$ bytes</td>
<td>All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour</td>
</tr>
<tr>
<td>Exabyte (EB)</td>
<td>1,000PB; $2^{60}$ bytes</td>
<td>Equivalent to 10 billion copies of The Economist</td>
</tr>
<tr>
<td>Zettabyte (ZB)</td>
<td>1,000EB; $2^{70}$ bytes</td>
<td>The total amount of information in existence this year is forecast to be around 1.2ZB</td>
</tr>
<tr>
<td>Yottabyte (YB)</td>
<td>1,000ZB; $2^{80}$ bytes</td>
<td>Currently too big to imagine</td>
</tr>
</tbody>
</table>

Source: The Economist

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Too much to read
Scientific articles published in peer-reviewed journals by authors' origin, 1990-2008

27 February, 2010 | From The Economist print edition
1. The problem - too much data!

Is this everybody’s future? Probably not. But as the torrent of information increases, it is not surprising that people feel overwhelmed. “There is an immense risk of cognitive overload,” explains Carl Pabo, a molecular biologist who studies cognition. The mind can handle seven pieces of information in its short-term memory and can generally deal with only four concepts or relationships at once. If there is more information to process, or it is especially complex, people become confused.

Moreover, knowledge has become so specialised that it is impossible for any individual to grasp the whole picture. A true understanding of climate change, for instance, requires a knowledge of meteorology, chemistry, economics and law, among many other things. And whereas doctors a century ago were expected to keep up with the entire field of medicine, now they would need to be familiar with about 10,000 diseases, 3,000 drugs and more than 1,000 lab tests. A study in 2004 suggested that in epidemiology alone it would take 21 hours of work a day just to stay current. And as more people around the world become more educated, the flow of knowledge will increase even further. The number of peer-reviewed scientific papers in China alone has increased 14-fold since 1990 (see chart 3).

“What information consumes is rather obvious: it consumes the attention of its recipients,” wrote Herbert Simon, an economist, in 1971. “Hence a wealth of information creates a poverty of attention.” But just as it is machines that are generating most of the data deluge, so they can also be put to work to deal with it. That highlights the role of “information intermediaries”. People rarely deal with raw data but consume them in processed form, once they have been aggregated or winnowed by computers. Indeed, many of the technologies described in this report, from business analytics to recursive machine-learning to visualisation software, exist to make data more digestible for humans.
1. The problem - too much data!

How to visualize lots of data....

In Nature this week, features and opinion pieces on one of the most daunting challenges facing modern science: how to cope with the flood of data now being generated. A petabyte is a lot of memory, however you say it - a quadrillion, $10^{15}$, or tens of thousands of trillions of bytes. But that is the currency of 'big data'. We visited the Sanger Institute's supercomputing centre, and its petabyte of capacity. [News Feature p. 16]
1. The problem - too much data!

Three Current “next-generation” technologies:

1. illumina (aka “Solexa”) - 500 million reads (100 bp)

Genome Analyzer IIx

Applications: DNA Sequencing, Gene Regulation Analysis, Sequencing-Based Transcriptome Analysis, SNP Discovery and Structural Variation Analysis, Cytogenetic Analysis, DNA-Protein Interaction Analysis (ChIP-Seq), Sequencing-Based Methylation Analysis

The Genome Analyzer IIx offers a unique combination of 2 x 100 bp read length and up to 500 million reads per flow cell with the simplest and fastest workflow. The highest raw accuracy and the largest number of perfect reads enables a broad range of high-throughput sequencing applications. Power your discoveries and generate highly accurate results in a week with the Genome Analyzer IIx. More...
1. The problem - too much data!

Three Current “next-generation” technologies:

1. illumina (aka “Solexa”) - 500 million reads (100 bp)
2. Roche 454
1. The problem - too much data!

Three Current “next-generation” technologies:

1. illumina (aka “Solexa”) - 500 million reads (100 bp)
2. Roche 454 - > 1 million reads (1000 bp)
3. ABI SOLiD

~100 Gbp per run!

35 bp reads
1. The problem - too much data!

Next-Generation DNA Sequencing/Review

Genome Research, Jan 2009

The new paradigm of flow cell sequencing
Robert A. Holt and Steven J.M. Jones

"Indeed, any of these new machines running at full capacity for a year will generate more sequence than existed in the whole of NCBI at the beginning of 2008. Analysis of the sequence data has rapidly become the limiting step and will likely become the most expensive part. The sheer volume of data will provide challenges in processing, networking, storage, and analysis of the flow-cell images just to provide the initial base calling." after Holt & Jones, 2009

---

Sanger Center has 28 Solexa machines, 8 ABI Solids, 2 Roche 454 machines

>1000 teraBytes per month!
1. The problem - too much data!

- **WGS section**
  - 104,000,000,000 bp

- **GenBank release**
  - 15 December 2007
  - 83,874,179,730 bp

- **Number of base pairs**

  - **GenBank**
  - **Moore's law**

  - **Y-axis**: Number of base pairs
  - **X-axis**: Years (1982 to 2006)
DNA: CLIMBING THE LADDER OF DISCOVERY

IMPACT

FORENSICS

CONFIDENTIALITY WITH RELIGION

DNA TIMELINE

CONFIDING ON WORLD EVENTS

MEDICAL TECHNOLOGIES

CHANGE

INNOCENCE PROJECT

PRIVATE ISSUES

GINA
2. A brief history - The speed of sequencing

What is a genome?

**genome** d3i.noun. *Biol.* Formerly also genom -nom. [a. *G. genom* (H. Winkler *Verbreitung u. Ursache d. Parthenogenesis* (1920) iv. 165), irreg. f. *gen* gene¹ + *chromosom* chromosome.] A haploid set of chromosomes; the sum-total of the genes in such a set.

- 1930 *Cytologia* I. 14 Chromosomes from different sets (or genoms) of *Triticum vulgarum* show affinity toward each other.
- 1930 [see allopolyploidy].
- 1932 *Proc. 6th Int. Congr. Genetics* I. 275 The inviability of deficient genomes in the haploid generation serves to some extent as an alternative distinction between mutation and deficiency.
- 1932 *Proc. 6th Int. Congr. Genetics* II. 5 There are two species having genoms resembling *C. neglecta*.
- 1952 *C. P. Blacker* *Eugenics* x. 243 The appearance of such terms as gene-complex and genome (denoting a set of chromosomes as a working unity) testify to the movement towards holism in genetics.
- 1965 *A. M. Srb et al. Gen. Genetics* (ed. 2) vii. 190 Among organisms with chromosomes, each species has a characteristic set of genes, or genome. In diplods a genome is found in each normal gamete. It consists of a full set of the different kinds of chromosomes.
- 1970 *Sci. Amer.* Oct. 19/1 The human genome..consists of perhaps as many as 10 million genes.
2. A brief history - The speed of sequencing

The Human Genome Project

Started more than 20 years ago (~1985)

The U.S. government agreed to invest
$200,000,000 U.S. per year for 20 years.

\(~3,400,000,000\) bp per haploid genome
\(~6,800,000,000\) bp per diploid genome

One base per second = 216 years!
2. A brief history - The speed of sequencing

1. "First Human Genome"
   $3,000,000,000 + 15 years

2. Celera genome (a.k.a. J. Craig Venter)
   $100,000,000 + 0.75 years (9 months)

3. Jim Watson's genome
   $900,000 + 0.17 years (2 months)

4. John Doe's genome
   $1,000 + 0.0002 years (0.1 day)

5. "next next-generation" machines
   - Helicos Biosystems machine can sequence human genome in 1 hour (2009).
   - Pacific Biosciences machine can sequence human genome in 4 minutes (2010).
   - Omni Molecular Recognizer Application - human genome less than $1, <1 minute.
2. A brief history - The speed of sequencing

36 hours, from purified DNA to GenBank file, ready to submit...

- Library preparation
- Emulsion set-up
- Emulsion breaking
- Bead enrichment
- Sequencer set-up
- Sequencer run
- Assembly
- Gene finding
- Gene annotation
- Genome analysis, visualization

Purified genomic DNA

Genome Atlas

NCBI
- PubMed
- Entrez
- Blast
- OMIM
- Books
- TaxBr

Submitting Sequence Data to GenBank
GENE SEQUENCING

The Race for the $1000 Genome

Fast, cheap genetic analyses will soon become a reality, and the consequences—good and bad—will affect everybody.

Marco Island, Florida—Computers aren’t the only things getting better and cheaper every time you turn around. Genome-sequencing prices are in free fall, too. The initial draft of the first human genome sequence, released just 5 years ago, cost an estimated $300 million. (The final draft and all the technology that made it possible came in near $3 billion.) Last month, genome scientists completed a draft of the genome sequence of the rhesus macaque for $22 million. And by the end of the year, at least one company expects to turn out a full mammalian genome sequence for about $100,000, a 3000-fold cost reduction in just 6 years.

It’s not likely to stop there. Researchers are closing in on a new generation of technology that they hope will slash the cost of a genome sequence to $1000. “Advances in this field are happening fast,” says Kevin McKernan, chief scientist at Agencourt Bioscience in Beverly, Massachusetts. “And they are coming more quickly than I think anyone was anticipating.” Jeffrey Schloss, who heads the sequencing-technologies grant program at the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland, agrees. “People are roundly encouraged and nervous,” Schloss says—encouraged because their own technologies are working, and nervous because their competitors’ are too.

A host of these novel sequencing technologies were on display last month at a meeting here. Although no one at the meeting claimed to have cracked the $1000 genome sequence yet, researchers are getting more confident that it’s a real possibility. “From what I’ve listened to the last few days, there is no physical principle that says we shouldn’t be able to do a $1000 genome,” says Harvard University sequencing pioneer George Church.

Even today, the declining cost of genome sequencing is triggering a flowering of basic research, looking at broad-ranging topics such as how the activation of genes is regulated and understanding genetic links to cancer. And as prices continue to drop, sequencing will revolutionize both the way biologists hunt for disease genes and the way medical professionals diagnose and treat diseases. In fact, some researchers say cheap sequencing could finally usher in personalized medicine in a major way. “The promise of cheap sequencing is in the understanding of disease and biology, such as cancer, where the genome changes over time,” says Dennis Gilbert, chief scientist of Applied Biosystems, the leading gene-sequencing-technology company based in Foster City, California. “It will enable different kinds of science to be done.” Of course, as with other forms

Free fall. As computer technology, the plunging cost of DNA sequencing has opened new applications in science and medicine.

*Advances in Genome Biology and Technology Conference, Marco Island, Florida, 8–11 February 2006.
<table>
<thead>
<tr>
<th>GPID</th>
<th>Organism</th>
<th>King</th>
<th>Size</th>
<th>Stain</th>
<th>Genome</th>
<th>Arrangement</th>
<th>Spore</th>
<th>Arranges</th>
<th>Motility</th>
<th>Salinity</th>
<th>Oxygen Req.</th>
<th>Habitat</th>
<th>Temp. range</th>
<th>Pathogenicity</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>30807</td>
<td>Nostoc azollae '0708</td>
<td>B</td>
<td>8.36</td>
<td>Filaments</td>
<td>Filaments</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td>Aerobic</td>
<td>Multiple</td>
<td>Mesophilic</td>
<td></td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>12997</td>
<td>Acaryochloris marina MBIC11017</td>
<td>B</td>
<td>8.36</td>
<td>Sphere, Ellipse</td>
<td>Singles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Aerobic</td>
<td>Aquatic</td>
<td>Mesophilic</td>
<td></td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>16707</td>
<td>Acaryochloris sp. CCME75410</td>
<td>B</td>
<td>8.36</td>
<td>Sphere, Ellipse</td>
<td>Singles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Aerobic</td>
<td>Aquatic</td>
<td>Mesophilic</td>
<td></td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>32769</td>
<td>Acetohalobium arabaticum DSM 5501</td>
<td>B</td>
<td>8.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Moderate halophilic</td>
<td>Anaerobic</td>
<td>Specialized</td>
<td></td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>19259</td>
<td>Acholeplasma laidlawii PG-8A</td>
<td>B</td>
<td>1.5</td>
<td>Pleomorphic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Facultative</td>
<td>Specialized</td>
<td>Mesophilic</td>
<td></td>
<td>Opportunistic infections</td>
<td>None</td>
</tr>
<tr>
<td>32215</td>
<td>Acidaminococcus intestinalis RYC-MR95</td>
<td>A</td>
<td>1.8</td>
<td>Coccolith</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Anaerobic</td>
<td>Host-associated</td>
<td>Mesophilic</td>
<td></td>
<td>Yes</td>
<td>None</td>
</tr>
<tr>
<td>29525</td>
<td>Acidimicrobium ferrooxidans DSM 10331</td>
<td>B</td>
<td>67.69</td>
<td>Coccus</td>
<td>Singles</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td>Anaerobic</td>
<td>Specialized</td>
<td>Thermophilic</td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>15753</td>
<td>Acidithiobacillus ferrooxidans JF-5</td>
<td>B</td>
<td>3.97</td>
<td>Rod</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Aerobic</td>
<td>Multiple</td>
<td>Mesophilic</td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>53</td>
<td>Acidithiobacillus ferrooxidans ATCC 23270</td>
<td>B</td>
<td>3</td>
<td>Spiral</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td>Aerobic</td>
<td>Specialized</td>
<td>Mesophilic</td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>16698</td>
<td>Acidithiobacillus ferrooxidans ATCC 53993</td>
<td>B</td>
<td>2.9</td>
<td>Spiral</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td>Aerobic</td>
<td>Specialized</td>
<td>Mesophilic</td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>15771</td>
<td>Acidobacteria bacterium Ellin345</td>
<td>B</td>
<td>5.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Aerobic</td>
<td>Multiple</td>
<td>Mesophilic</td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>1</td>
<td>Acidobacterium capsulatum ATCC 51196</td>
<td>B</td>
<td>4.15</td>
<td>Rod</td>
<td>Singles</td>
<td>No</td>
<td>Yes</td>
<td>Non-halophilic</td>
<td></td>
<td>Aerobic</td>
<td>Multiple</td>
<td>Mesophilic</td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>28085</td>
<td>Acidobacterium capsulatum ATCC 51196</td>
<td>B</td>
<td>2.4</td>
<td>Rod</td>
<td>Singles</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Non-halophilic</td>
<td>Aerobic</td>
<td>Multiple</td>
<td>Mesophilic</td>
<td>No</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>16097</td>
<td>Acidothermus cellulosilyticus 11B</td>
<td>B</td>
<td>2.4</td>
<td>Rod</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Aerobic</td>
<td>Aquatic</td>
<td>Thermophilic</td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>15708</td>
<td>Acidovorax avenue subsp. citulinii AAC00-1</td>
<td>B</td>
<td>5.4</td>
<td>Rod</td>
<td>No</td>
<td>Yes</td>
<td>Non-halophilic</td>
<td></td>
<td>Aerobic</td>
<td>Multiple</td>
<td>Mesophilic</td>
<td>Fruit</td>
<td>Bacterial fruit blotch</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>32605</td>
<td>Acidovorax delafeldii 2AN</td>
<td>B</td>
<td>6.66</td>
<td>Rod</td>
<td>No</td>
<td>Yes</td>
<td>Non-halophilic</td>
<td></td>
<td>Aerobic</td>
<td>Aquatic</td>
<td>Mesophilic</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>15685</td>
<td>Acidovorax sp. JS42</td>
<td>B</td>
<td>4.54</td>
<td>Rod</td>
<td>No</td>
<td>Yes</td>
<td>Non-halophilic</td>
<td></td>
<td>Aerobic</td>
<td>Terrestrial</td>
<td>Mesophilic</td>
<td>No</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>19277</td>
<td>Acidiphilum profundum Boonei</td>
<td>B</td>
<td>39.1</td>
<td>Coccus</td>
<td>Yes</td>
<td>Mesophilic</td>
<td>Anaerobic</td>
<td>Specialized</td>
<td>Thermophilic</td>
<td>No</td>
<td></td>
<td></td>
<td>No</td>
<td>None</td>
<td></td>
</tr>
</tbody>
</table>
as of 4 March, 2010

<table>
<thead>
<tr>
<th>GPID</th>
<th>Organism</th>
<th>Group</th>
<th>* size</th>
<th>GC</th>
<th>Chr</th>
<th>#plan</th>
<th>GenBank</th>
<th>RefSeq</th>
<th>Released</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>12997</td>
<td>Acaryochinos marina MBIC11017</td>
<td>Cyanobacteria</td>
<td>8.36</td>
<td>47.8</td>
<td>1</td>
<td>9</td>
<td>CP0008281</td>
<td>NC_009295.1</td>
<td>10/16/07</td>
<td>04/27/07</td>
</tr>
<tr>
<td>31129</td>
<td>Acetobacter pasteurianus JFQ 3263-01</td>
<td>Alphaproteobacteria</td>
<td>3.33</td>
<td></td>
<td>6</td>
<td>AP011211</td>
<td>NC_012309.1</td>
<td>09/26/09</td>
<td>09/14/09</td>
<td></td>
</tr>
<tr>
<td>31141</td>
<td>Acetobacter pasteurianus JFQ 3263-01-42C</td>
<td>Alphaproteobacteria</td>
<td>3.23</td>
<td></td>
<td>6</td>
<td>AP011128</td>
<td>NC_012308.1</td>
<td>09/26/09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31133</td>
<td>Acetobacter pasteurianus JFQ 3263-03</td>
<td>Alphaproteobacteria</td>
<td>3.33</td>
<td></td>
<td>6</td>
<td>AP011135</td>
<td>NC_012307.1</td>
<td>09/26/09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32203</td>
<td>Acetobacter pasteurianus JFQ 3263-12</td>
<td>Alphaproteobacteria</td>
<td>3.33</td>
<td></td>
<td>6</td>
<td>AP011170</td>
<td>NC_012306.1</td>
<td>09/26/09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31137</td>
<td>Acetobacter pasteurianus JFQ 3263-22</td>
<td>Alphaproteobacteria</td>
<td>3.33</td>
<td></td>
<td>6</td>
<td>AP011149</td>
<td>NC_012305.1</td>
<td>09/26/09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31137</td>
<td>Acetobacter pasteurianus JFQ 3263-26</td>
<td>Alphaproteobacteria</td>
<td>3.33</td>
<td></td>
<td>6</td>
<td>AP011156</td>
<td>NC_012304.1</td>
<td>09/26/09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31137</td>
<td>Acetobacter pasteurianus JFQ 3263-32</td>
<td>Alphaproteobacteria</td>
<td>3.33</td>
<td></td>
<td>6</td>
<td>AP011156</td>
<td>NC_012303.1</td>
<td>09/26/09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19259</td>
<td>Acholeplasma laidlawii PG 8A</td>
<td>Firmicutes</td>
<td>1.5</td>
<td>31.9</td>
<td>1</td>
<td></td>
<td>CP0008961</td>
<td>NC_010163.1</td>
<td>12/07/07</td>
<td>04/29/09</td>
</tr>
<tr>
<td>26655</td>
<td>Acetamidobacter fermentans DSM 20731</td>
<td>Firmicutes</td>
<td>2.3</td>
<td>68.3</td>
<td>1</td>
<td>CP0018592</td>
<td>NC_013740.1</td>
<td>01/15/10</td>
<td>01/19/10</td>
<td></td>
</tr>
<tr>
<td>20525</td>
<td>Acidimicrobium fermentans DSM 10331</td>
<td>Firmicutes</td>
<td>4.4</td>
<td>68.3</td>
<td>1</td>
<td>CP0018361</td>
<td>NC_013124.1</td>
<td>04/30/09</td>
<td>09/01/09</td>
<td></td>
</tr>
<tr>
<td>15735</td>
<td>Acidithiobacillus ferrooxidans JFQ 3263-01</td>
<td>Alphaproteobacteria</td>
<td>3.97</td>
<td>67.1</td>
<td>1</td>
<td>CP0006972</td>
<td>NC_009484.1</td>
<td>05/11/07</td>
<td>04/28/09</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>Acidithiobacillus ferrooxidans ATCC 23270</td>
<td>Gammaproteobacteria</td>
<td>0.3</td>
<td>58.8</td>
<td>1</td>
<td>CP0012198</td>
<td>NC_011761.1</td>
<td>12/19/98</td>
<td>04/25/09</td>
<td></td>
</tr>
<tr>
<td>16689</td>
<td>Acidithiobacillus ferrooxidans ATCC 53993</td>
<td>Comamonadetes</td>
<td>2.9</td>
<td>58.9</td>
<td>1</td>
<td>CP0013222</td>
<td>NC_011206.1</td>
<td>09/05/08</td>
<td>06/08/08</td>
<td></td>
</tr>
<tr>
<td>20808</td>
<td>Acidithiobacillus ferrooxidans ATCC 53993</td>
<td>Comamonadetes</td>
<td>2.9</td>
<td>58.9</td>
<td>1</td>
<td>CP0013222</td>
<td>NC_011206.1</td>
<td>09/05/08</td>
<td>06/08/08</td>
<td></td>
</tr>
<tr>
<td>16699</td>
<td>Acidithiobacillus ferrooxidans ATCC 53993</td>
<td>Comamonadetes</td>
<td>2.9</td>
<td>58.9</td>
<td>1</td>
<td>CP0013222</td>
<td>NC_011206.1</td>
<td>09/05/08</td>
<td>06/08/08</td>
<td></td>
</tr>
<tr>
<td>16699</td>
<td>Acidithiobacillus ferrooxidans ATCC 53993</td>
<td>Comamonadetes</td>
<td>2.9</td>
<td>58.9</td>
<td>1</td>
<td>CP0013222</td>
<td>NC_011206.1</td>
<td>09/05/08</td>
<td>06/08/08</td>
<td></td>
</tr>
<tr>
<td>20808</td>
<td>Acidithiobacillus ferrooxidans ATCC 53993</td>
<td>Comamonadetes</td>
<td>2.9</td>
<td>58.9</td>
<td>1</td>
<td>CP0013222</td>
<td>NC_011206.1</td>
<td>09/05/08</td>
<td>06/08/08</td>
<td></td>
</tr>
<tr>
<td>16699</td>
<td>Acidithiobacillus ferrooxidans ATCC 53993</td>
<td>Comamonadetes</td>
<td>2.9</td>
<td>58.9</td>
<td>1</td>
<td>CP0013222</td>
<td>NC_011206.1</td>
<td>09/05/08</td>
<td>06/08/08</td>
<td></td>
</tr>
<tr>
<td>16699</td>
<td>Acidithiobacillus ferrooxidans ATCC 53993</td>
<td>Comamonadetes</td>
<td>2.9</td>
<td>58.9</td>
<td>1</td>
<td>CP0013222</td>
<td>NC_011206.1</td>
<td>09/05/08</td>
<td>06/08/08</td>
<td></td>
</tr>
</tbody>
</table>
Staphylococcus aureus subsp. aureus TW20, complete genome

Features

LOCUS FN433596
DEFINITION Staphylococcus aureus subsp. aureus TW20, complete genome.
ACCESSION FN433596
VERSION FN433596.1 GI:269939526
DBLINK Project:36647
KEYWORDS complete genome.
SOURCE Staphylococcus aureus subsp. aureus TW20
ORGANISM Staphylococcus aureus subsp. aureus TW20
Bacteria; Firmicutes; Bacillales; Staphylococcaceae.
REFERENCE 1
AUTHORS Holden,M.T., Lindsay,J.A., Corton,C., Quail,M.A., Cockfield,J.D., Pathak,S., Batra,R., Parkhill,J., Bentley,S.D. and Edgeworth,J.D.
TITLE Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant Staphylococcus aureus, sequence type 239 (TW)
PUBMED 19948800
REFERENCE 2 (bases 1 to 3043210)
AUTHORS Holden,M.T.G.
TITLE Direct Submission
JOURNAL Submitted (30-JUL-2009) Holden M.T.G., The Wellcome Trust Sanger Institute, Pathogen Sequencing Unit, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UNITED KINGDOM
FEATURES
source
Location/Qualifiers
1..3043210
/organism="Staphylococcus aureus subsp. aureus TW20"
/mol_types="genomic DNA"
3. Cautionary tales

As of 9 March, 2009
3. Cautionary tales
3. Cautionary tales
3. Cautionary tales

**human**
- 23,621 genes
- 19,568 orthologs

**chimp**
- 19,829 genes
- 14,000 orthologs

**chicken**
- 18,529 genes
- 10,000 orthologs

**worm**
- 18,000 genes
- 1,700 orthologs

**yeast**
- 6,000 genes
- 1,700 orthologs

**C. botulinum**
- type A, strain ATCC 3502
  - 3606 genes
  - 3202 orthologs
- type A, strain ATCC 19397
  - 3553 genes
  - 2974 orthologs
- type A, strain Kyoto
  - 3874 genes
  - 1126 orthologs
- type C, strain Ecklund
  - 2801 genes
  - 1092 orthologs
- type E1, strain BoNT E Beluga
  - 3760 genes
  - 1,092 orthologs
3. Cautionary tales

**BACTERIA**
- Proteobacteria
- Actinobacteria
- Planctomycetes
- Flavobacteria
- Cyanobacteria
- Chlamydiae
- Chlorobi
- Bacteroidetes
- Spirochaetes
- Clostridium
- Firmicutes
- Bacillus
- Chloroflexi
- Acidobacteria
- Thermus
- Deinococcus
- Thermotogae
- Armatimonadetes

**ARCHAEA**
- Euryarchaeota
- Crenarchaeota
- Methanobacterium
- Methanococcus
- Thermoplasma
- Archaea

**EUCARYA**
- Giardia
- Saccharomyces
- Trypanosoma

**Unicellular eukaryotes**
- Slime mold
- Babesia

**Animals**
- Macro-organisms

**Plants**
- Protozoans
4. Approaches to handle lots of data

Statistics
4. Approaches to handle lots of data

**Statistics**

Size distribution of Prokaryotic genomes (n=490)

- A_Crenarchaeota (n=12)
- A_Euryarchaeota (n=26)
- A_Nanoarchaeota (n=1)
- B_Acidobacteria (n=2)
- B_Actinobacteria (n=39)
  - B_Aquificae (n=1)
- B_BacteroidetesChlorobi (n=12)
- B_Chlamydiae (n=11)
- B_Chloroflexi (n=2)
- B_Cyanobacteria (n=23)
- B_DeinococcusThermus (n=4)
- B_Firmicutes (n=106)
- B_Fusobacteria (n=1)
- B_Planctomycetes (n=1)
- B_Proteobacteria_Alpha (n=56)
- B_Proteobacteria_Beta (n=43)
- B_Proteobacteria_Delta (n=14)
- B_Proteobacteria_Epsilon (n=11)
- B_Proteobacteria_Gamma (n=115)
- B_Spirochaetes (n=9)
- B_Thermotogae (n=1)
4. Approaches to handle lots of data
   - Visualization

```plaintext
418 base pairs of DNA sequence.
```
coliphage phiX174
5,386 bp
Comparative Microbial Genomics group

coliphage phiX174
5,386 bp

G Content
dev 0.07
avg 0.39

A Content
dev 0.01
avg 0.47

T Content
dev 0.10
avg 0.53

C Content
dev 0.04
avg 0.39

Annotations:
- CDS +

AT Skew
dev -0.33
avg 0.18

GC Skew
dev -0.10
avg 0.14

Percent AT
dev 0.47
avg 0.63

Resolution: 3

BASE ATLAS
H. influenzae Rd KW20
1,830,138 bp

Intrinsic Curvature
0.19
0.26

Stacking Energy
-7.91
-7.09

Position Preference
0.14
0.17

Annotations:
- CDS +
- CDS -
- rRNA
- tRNA

Global Direct Repeats
5.00
7.50

Global Inverted Repeats
5.00
7.50

GC Skew
-0.07
0.07

Percent AT
0.20
0.80

Resolution: 733

GENOME ATLAS
<table>
<thead>
<tr>
<th>Strain/Strain Type</th>
<th>Homology within genomes</th>
<th>Homology between genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>3625 / 5379</td>
<td>67.4%</td>
<td>9.12 - 13.67</td>
</tr>
<tr>
<td>3637 / 5379</td>
<td>67.6%</td>
<td>65.11 - 97.19</td>
</tr>
<tr>
<td>3502 / 5379</td>
<td>65.1%</td>
<td></td>
</tr>
<tr>
<td>680 / 5379</td>
<td>12.6%</td>
<td></td>
</tr>
</tbody>
</table>

**Escherichia coli**

- **strain K-12 isolate MG1655**
  - 4289 genes, 4,639,221 bp

- **strain K-12 isolate W3110**
  - 4387 genes, 4,641,433 bp

- **strain K-12 substr. DH10B**
  - 4126 genes, 4,686,137 bp

- **strain CFT073**
  - 5379 genes, 5,231,428 bp
Vibrio cholerae M66-2

52% AT
Vibrio fischeri MJII

62% AT