The Human Genome Project

Part 2 - Next-gen & ‘third-generation’ sequencing

Dave Ussery
Biological Sequence Analysis (DTU course 27803)
Thursday, 5 May, 2011
Outline

• The problem - too much data!
• Next-generation sequencing machines
• Cautionary tales
• What can be done with all this data?
• Third-generation sequencing methods
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!

**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 “binary digit”, 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 “thousand” in Greek. One page of typed text is 2KB</td>
</tr>
<tr>
<td>Megabyte (MB)</td>
<td>1,000KB; $2^{20}$ bytes</td>
<td>From “large” in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB</td>
</tr>
<tr>
<td>Gigabyte (GB)</td>
<td>1,000MB; $2^{30}$ bytes</td>
<td>From “giant” 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 “monster” 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 <em>The Economist</em></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.
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]
the time to result, reducing the overall footprint of the instrument, and lowering cost to make DNA sequencing more generally accessible to all. However, this technology is still a 'wash-and-scan' system like all current SGS technologies, requiring PCR amplification of the DNA template in each well, as well as termination events typically halting sequencing after each nucleotide incorporation, in order to monitor in succession the incorporation of each of the four bases across all DNA templates. As a result of this process, the overall read length is limited to that of current SGS systems.

Figure 1. How previous generation DNA-sequencing systems work. (A) A modern implementation of Sanger sequencing is shown to illustrate differential labeling and use of terminator chemistry followed by size separation to resolve the sequence. (B) The Illumina sequencing process is shown to illustrate the wash-and-scan paradigm common to second-generation DNA-sequencing technologies.
1. The problem - too much data!

Three Current “next-generation” technologies:

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

Three Current “next-generation” technologies:

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

2. Roche 454

Genome Sequencer FLX System

The Genome Sequencer FLX System, with long-read GS FLX Titanium chemistry, is the flagship 454 Sequencing platform. Offering more than 1 million high-quality reads per run and read lengths of 400 bases, the system is ideally suited for de novo sequencing of whole genomes and transcriptomes of any size, metagenomic characterization of complex samples, resequencing studies and more. The GS FLX System is at the heart of breakthrough scientific discoveries and hundreds of peer-reviewed publications to date.

Continuous development of the GS FLX Titanium series chemistry will soon enable the next leap in performance, with extended read lengths approaching 1000 bases—Coming in 2010

> Learn more

Introducing the GS Junior System

The next big thing in sequencing is small

The GS Junior System brings the power of 454 Sequencing technology directly to your laboratory bench top. Benefit from the same proven long-read chemistry as the Genome Sequencer FLX System, scaled to suit the needs of individual labs. Quickly proceed from DNA to results to discovery with an easy-to-follow workflow and data analysis at your desktop.

The system is perfectly sized for rapid sequencing of amplicons (PCR products), targeted human resequencing studies, de novo sequencing of microbial and other small genomes, pathogen detection and much more—Coming in 2010

> Learn 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 million reads (1000 bp)
3. ABI SOLiD

~100 Gbp per run!

35 bp reads

Applied Biosystems®
SOLiD™ 4 System

Key Benefits
- **Higher accuracy**—detection of causative variation enabled at lower coverage and cost per sample
- **Scalable throughput on a single platform**—80–100 GB of mappable sequence per run
- **Automated workflow**—80% reduction in hands-on time and increased reproducibility in yield allow for significant time and labor savings
- **True paired-end sequencing**—bidirectional sequencing facilitates detection of genetic alterations as well as splice variants and fusion transcripts with lower sample input
- **Robust multiplexing kits**—intelligent barcode strategy enables accurate assignment without introduction of bias

Experience Peace of Mind

The SOLiD™ System's open slide format and flexible bead densities continue to yield increases in throughput on the same platform with minor upgrades. The SOLiD™ 4 System can generate up to 100 Gb of mappable sequence or greater than 1.4 billion reads per run. Discover the peace of mind provided by the confidence that you will benefit from future technology advances without the purchase of a new system.
ABSTRACT

The next generation of DNA sequencing platforms produces sequencing reads with increased depth of coverage but reduced read length and lower per-base accuracy than data from Sanger-based DNA sequencing. New approaches are needed to overcome these issues and provide accurate mutation discovery and consensus sequences. 2-Base encoding is uniquely enabled by the ligation-based sequencing protocol used in the SOLiD™ system (a massively parallel sequencing technology based on ligation of oligonucleotides). Sequencing is carried out via sequential rounds of ligation with high fidelity and high read quality. In this system there are 16 dinucleotide combinations with 4 fluorescent dyes, each dye corresponding to a probe pool of 4 dinucleotides per pool. Using this dinucleotide, 4-dye encoding scheme in conjunction with a sequencing assay that samples every base, each base is effectively probed in two different reactions. The double interrogation of each base causes a SNP to result in a two-color change while a measurement error results in a single color change. In addition, only one-third of all possible two-color combinations are considered valid and result in a base change. 2-Base encoding rules (a single mismatch is a measurement error, only one-third of adjacent mismatches are valid) significantly reduce the raw error rate (30 bp reads have a 45x reduction in raw measurement errors) and this benefit increases 3/2 as the read length is increased. The reduction in raw error rate enabled by 2-Base encoding translates into more accurate alignment of short reads, polymorphism discovery and consensus calling.

What is 2-Base Encoding?

The SOLiD Sequencing System uses probes with dual base encoding.

![Image](image1)

Figure 1. Each probe consists of 8 bases. As shown, the first 3 bases are degenerate (n), and the last 3 are universal (u), with the 4th and 5th bases as the two bases being interrogated. Thus, a single color observation only limits the potential dinucleotide to being four out of the 16 possible dinucleotides. As seen above, a green signal represents a AC, CA, TG or GT. Using this dinucleotide, 4-dye encoding scheme in conjunction with a sequencing assay that samples every base, each base is effectively probed in two different reactions.

Double Interrogation

Using this dinucleotide, 4-dye encoding scheme in conjunction with a sequencing assay that samples every base, each base is effectively probed in two different reactions.

![Image](image2)

Figure 2 demonstrates the principle of double interrogation. Each color measurement represents four possible dinucleotide combinations. For example, the first measured blue represents ‘AA’ and the third blue represents ‘GC’.

Color Space

In order to use 2-base encoding the concept of color space must be used. Instead of using a nucleotide-based reference sequence, a color space reference sequence is used. As color space and base space both consist of four possible elements (four colors represented as 0, 1, 2, or 3 and A, C, G or T, respectively) existing algorithms can be used for alignment and consensus calling of color space. As will be demonstrated, the properties of 2-base encoding allow significantly enhanced results if 2-base encoding is taken into account and expanded algorithms used.

Decoding

To decode a sequence the decoding matrix in figure 3 is used:

![Image](image3)

Figure 3. The decoding matrix allows a sequence of dinucleotides to be converted to a base sequence, as long as one of two bases is known. The design of encoding probes has been carefully made, as can be seen by the reversed transition (e.g., A → T and T → A is the same color as is the complement A → G and T → C).

Single Nucleotide Polymorphisms (SNPs)

In many resequencing projects one of the most important objectives is to measure Single Nucleotide Polymorphisms (SNPs) that may be responsible for differences in phenotype. In 2-base encoding most measure errors can be distinguished from potential SNPs as demonstrated below in figure 5:

![Image](image4)

Figure 5. If a SNP occurs in the sequence ‘C-T’ there are only 3 possible results: CGT, CCT and CTT. This means that only 3 disease combinations are allowed and any other disease combinations are ignored. Since any base is defined by two nucleotides (e.g., C-A and A-T), then two adjacent changes must be observed for any SNP. Thus, measurement errors are represented by single changes. As there are only 3 alternative bases that can occur when a SNP is observed (i.e., an A can go to C, G or T), there are only three allowed disease combinations for any starting adjacent transition. The other six possible adjacent combinations are therefore by definition invalid. Thus, when two adjacent measurement errors are seen, only 1/3 of them could be mistaken for a real SNP, prior to applying any consensus rules. Since the two surrounding combinations contain information about the incorrect combination it is possible to have support for the hypothesis that the reference sequence is unchanged even if a single changed combination is seen and discarded.

SOLiD System™ Accuracy

![Image](image5)

Figure 6: SOLiD™ System’s error rate per base position in sequence read.

Conclusion

The ability to use 2-base encoding to recognize and eliminate measurement errors from subsequent analysis has been demonstrated. In numerous experiments, a minimum error reduction of 20-fold has been seen. Only sequencing by ligation offers the ability to use 2-base encoding. Thus, SOLID sequencing systems offer the best solution to many applications.
1. The problem - too much data!

"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 37 Solexa machines, 8 ABI Solids, 2 Roche 454 machines

>10,000 terabytes per month!
Figure 3: Computing cost dominate sequencing costs. While sequencing costs remain almost identical across platforms, the analysis costs vary with data set sizes. The cost of sequencing compared to the cost of running BLASTX analysis. Data from [11] using the Amazon EC2 cloud machine as a cost model.

THE SEQUENCE EXPLOSION

At the time of the announcement of the first drafts of the human genome in 2000, there were 8 billion base pairs of sequence in the three main databases for ‘finished’ sequence: GenBank, run by the US National Center for Biotechnology Information, the DNA Databank of Japan, and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database. The databases share their data regularly as part of the International Nucleotide Sequence Database Collaboration (INSDC). In the subsequent first-post-genome decade, they have added another 270 billion bases to the collection of finished sequence, doubling the size of the database roughly every 18 months. But this number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace archive and Sequence Read Archive (SRA).

DNA SEQUENCES BY TAXONOMY

International Nucleotide Sequence Database Collaboration: The main repositories of ‘finished’ sequence span a wide range of organisms, representing the many priorities of scientists worldwide.

TRACE ARCHIVE

- Finished sequences
- Preliminary sequences
- Sequence from partial assemblies
- Sequence from older projects
- Non-human organisms
- Human organisms

SEQUENCE READ ARCHIVE

- Human genomes
- Non-human genomes
- Preliminary sequences
- Whole genome shotgun sequences
- Completed genomes

DNA SEQUENCES BY TAXONOMY

GenBank, run by the US National Center for Biotechnology Information, is the DNA Databank of Japan, and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database. The databases share their data regularly as part of the International Nucleotide Sequence Database Collaboration (INSDC). In the subsequent first-post-genome decade, they have added another 270 billion bases to the collection of finished sequence, doubling the size of the database roughly every 18 months. But this number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace archive and Sequence Read Archive (SRA).

THE SEQUENCE EXPLOSION

At the time of the announcement of the first drafts of the human genome in 2000, there were 8 billion base pairs of sequence in the three main databases for ‘finished’ sequence: GenBank, run by the US National Center for Biotechnology Information, the DNA Databank of Japan, and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database. The databases share their data regularly as part of the International Nucleotide Sequence Database Collaboration (INSDC). In the subsequent first-post-genome decade, they have added another 270 billion bases to the collection of finished sequence, doubling the size of the database roughly every 18 months. But this number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace archive and Sequence Read Archive (SRA).
Acetobacter species have been used historically for making vinegar and are known to be quite variable in terms of their genome content, due to a large number of transposons (Azuma et al., 2009). Although the eight A. pasteurianus strains are also present in DDBJ and GenBank, only one sequence is found in the 'complete genomes' list on the NCBI web pages, and the other seven genomes, although finished to one contiguous piece, are listed as 'in progress' more than 6 months after being deposited in the International Nucleotide Sequence Database.

Finally, we have listed (Table 1) a set of seven genomes deposited in GenBank around the same time. At one stage, Vibrio sp. isolate Ex25, was listed as the 1000th genome on the NCBI pages, although with time, as some older genomes have been removed and the lists have been updated, the 1000th position has moved a bit, so a range of genomes is given. Of the seven genomes in the list, five have at least one publication associated with them at the time of writing; Vibrio sp. Ex25 (CP001805) and Staphylococcus aureus ED98 (CP001781) genomes have not been described in a publication. Three of the genomes are from the GEBA project (Wu et al., 2009): Rhodothermus marinus strain R-10T is a member of the phylum Bacteroidetes and was isolated from hot springs off the coast of Iceland (Nolan et al., 2009); Gordonia bronchialis DSM 43247 is a member of the phylum Actinobacteria, isolated from the sputum of a woman with diseased lungs; and Haliangium ochraceum DSM 14365 is a halophilic deltaproteobacterium isolated from coastal sand in Japan. The Blattabacterium genome sequenced, which is from a cockroach endosymbiotic strain, belongs to the class Flavobacteria in the phylum Bacteroidetes; the newly sequenced genome shows evolutionary convergence with gammaproteobacterium endosymbionts (Lo´pez-S´anchez et al., 2009). Finally, Comamonas testosteroni strain CNB-2 was isolated from soil contaminated with 4-chloronitrobenzene and can grow on this pollutant using it as its sole carbon and nitrogen source (Ma et al., 2009). This betaproteobacterium is a member of the order Burkholderiales, and was given its name from its ability to metabolize testosterone.

Of the 1000 prokaryotic genomes sequenced so far, only 7 % are archaea, with the rest (93 %) being bacteria. Whether this ratio is truly reflective of the relative proportions in the environment is doubtful. Within the bacteria, members of the phyla Proteobacteria and Firmicutes make up the majority of the completed genome sequences, and many of the new genomes sequenced each year are found within these two phyla. Together, these account for 72 % (of 930) of bacterial genomes, with 489 (52 %) proteobacteria and 187 (20 %) firmicute genomes (GOLD data). A similar skewed distribution is observed for the archaeal genomes, where the majority are from the phylum Euryarchaeota (63 %); with nearly all the rest coming from the phylum Crenarchaeota (31 %). The phylum Nanoarchaeota has only a single sequenced genome. Such an uneven distribution in available sequences has consequences for the chance that a particular query sequence will identify similarities in the database: when a query sequence from a proteobacterial genome is used to search the microbial database using BLAST, the chance is much higher that a hit will be found than, for instance, searching with a nanoarchaeota sequence. Since the E-value reported by BLAST is based on the expected background noise, and is based on the assumption that the sequences within the database are random (which obviously is not true in this case), the results should be interpreted with caution when the 'best hit' of a query sequence is found.

**Fig. 1.** Increase in the number of genomes completed per year separated by bacterial phylum. Data source: NCBI, complete genomes (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi).

*Microbiology, 156:603-608, (2010).*
High-throughput comparisons

- **E.coli**
  - Publicly available genomes:
    - 203 genomes

- **Salmonella**
  - Publicly available genomes:
    - 42 genomes

- **Campylobacter**
  - Publicly available genomes:
    - 44 genomes

- **Pseudomonas**
  - 72 genomes
What is a core-genome? What is a pan-genome?

A minimal gene set for cellular life derived by comparison of complete bacterial genomes

ARCADY R. MUSHEGIAN AND EUGENE V. KOONIN*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Communicated by Clyde Hutchinson, University of North Carolina, Chapel Hill, NC, May 17, 1996 (received for review March 11, 1996)

1743 genes

256 genes = ‘core’

1743 genes + 468 genes - 256 ‘core’

= 1995 ‘pan-genes’

Mycoplasma genitalium

Haemophilus influenza
What is a core-genome?

Genomic Sequence of an Otitis Media Isolate of Nontypeable Haemophilus influenzae: Comparative Study with H. influenzae Serotype d, Strain KW20

Alistair Harrison,1 David W. Dyer,2 Allison Gillaspy,2 William C. Ray,1 Rachna Mungur,1 Matthew B. Carson,2 Huachun Zhong,1 Jenny Gipson,2 Mandy Gipson,2 Linda S. Johnson,1 Lisa Lewis,2 Lauren O. Bakaletz,1 and Robert S. Munson, Jr.1*

Center for Microbial Pathogenesis, Columbus Children’s Research Institute, and Department of Pediatrics, Ohio State University College of Medicine and Public Health, Columbus, Ohio 43205,1 and Laboratory for Genomics and Bioinformatics, and Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 731042

Received 11 January 2005/Accepted 29 March 2005

In 1995, the Institute for Genomic Research completed the genome sequence of a rough derivative of Haemophilus influenzae serotype d, strain KW20. Although extremely useful in understanding the basic biology of H. influenzae, these data have not provided significant insight into disease caused by nontypeable H. influenzae, as serotype d strains are not pathogens. In contrast, strains of nontypeable H. influenzae are the primary pathogens of otitis media in children. In addition, these organisms have an important role in acute infections in children as well as other respiratory diseases. Such strains must therefore contain a gene repertoire that differs from that of strain Rd. Elucidation of the differences between these genomes will thus provide insight into the pathogenic mechanisms of nontypeable H. influenzae. The genome of a representative nontypeable H. influenzae strain, 86-028NP, isolated from a patient with chronic otitis media was therefore sequenced and annotated. Despite large regions of synteny with the strain Rd genome, there are large rearrangements in strain 86-028NP’s genome architecture relative to the strain Rd genome. A genomic island similar to an island originally identified in H. influenzae type b is present in the strain 86-028NP genome, while the mu-like phage present in the strain Rd genome is absent from the strain 86-028NP genome. Two hundred eighty open reading frames were identified in the strain 86-028NP genome that were absent from the strain Rd genome. These data provide new insight that complements and extends the ongoing analysis of nontypeable H. influenzae virulence determinants.
E. coli pan-genome
~45,000 gene families

Escherichia coli K-12 MG1655
4144 proteins

~800 E. coli gene families

BMC Genomics 2009, 10:385

CBS, Department of Systems Biology
Biological Sequence Analysis (DTU course #27803)
How is the core- and pan-genome calculated?

Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”


Fig. 2. GBS core genome. The number of shared genes is plotted as a function of the number n of strains sequentially added (see Materials and Methods). For each n, circles are the 8!/[n−1]!(8−n)! values obtained for the different strain combinations. Squares are the averages of such values. The continuous curve represents the least-squares fit of the function \( F_c = \kappa_c \exp\left[-n/\tau_c\right] + \Omega \) (see Eq. 1 in Supporting Text) to data. The best fit was obtained with correlation \( r^2 = 0.990 \) for \( \kappa_c = 610 \pm 38, \tau_c = 2.16 \pm 0.28, \) and \( \Omega = 1,806 \pm 16. \) The extrapolated GBS core genome size \( \Omega \) is shown as a dashed line.

Fig. 3. GBS pan-genome. The number of specific genes is plotted as a function of the number n of strains sequentially added (see Materials and Methods). For each n, circles are the 8!/[n−1]!(8−n)! values obtained for the different strain combinations; squares are the averages of such values. The blue curve is the least-squares fit of the function \( F_p(n) = \kappa_p \exp\left[-n/\tau_p\right] + tg(\delta) \) (see Eq. 2 in Supporting Text) to the data. The best fit was obtained with correlation \( r^2 = 0.995 \) for \( \kappa_p = 476 \pm 62, \tau_p = 1.51 \pm 0.15, \) and \( tg(\delta) = 33 \pm 3.5. \) The extrapolated average number \( tg(\delta) \) of strain-specific genes is shown as a dashed line.
E. coli “core genes” in 32 genomes

32! = 2.6 \times 10^{35}
possible permutations

~1560 “core genes”


Highly accessed Open Access

CBS, Department of Systems Biology

Biological Sequence Analysis (DTU course #27803)
<table>
<thead>
<tr>
<th>Rank</th>
<th>Organism</th>
<th>#genes</th>
<th># new genes</th>
<th># new gene families</th>
<th>pan-core-</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E. coli K-12 MG1655</td>
<td>4289</td>
<td>4289</td>
<td>4057</td>
<td>4057</td>
</tr>
<tr>
<td>2</td>
<td>E. coli K-12 W3110</td>
<td>4387</td>
<td>174</td>
<td>140</td>
<td>4197</td>
</tr>
<tr>
<td>3</td>
<td>E. coli K-12 DH10B</td>
<td>4126</td>
<td>86</td>
<td>81</td>
<td>4278</td>
</tr>
<tr>
<td>4</td>
<td>E. coli CFT073</td>
<td>5379</td>
<td>1710</td>
<td>1524</td>
<td>5802</td>
</tr>
</tbody>
</table>
The *Burkholderia* Pan- and Core Genome

### Fig. 12.5

The pan-genome and core genome for five different Proteobacterial genera. The *Salmonella* graph represents one species (*Salmonella enterica*), whereas the *E. coli/Shigella* figure contains both *E. coli* and four different *Shigella* species. The other graphs represent multiple species per genus. All graphs are drawn on the same scale.
15,993 “pan-gene families”

63! = $2.0 \times 10^{87}$ possible permutations

960 “core gene families”

Number of gene families

- Core genome
- Pan genome
- New gene families

E. coli O157

E. coli K-12

Salmonella

E. fergusonii

E. albertii


1: Escherichia coli 0157:H7 str. EC4196
2: Escherichia coli 0157:H7 str. EC4113
3: Escherichia coli 0157:H7 str. EC508
4: Escherichia coli 0157:H7 str. EC4501
5: Escherichia coli 0157:H7 str. EC4076
6: Escherichia coli 0157:H7 str. EC4115
7: Escherichia coli 0157:H7 str. EC4042
8: Escherichia coli 0157:H7 str. EC4486
9: Escherichia coli 0157:H7 str. EC869
10: Escherichia coli 0157:H7 str. EC4206
11: Escherichia coli 0157:H7 str. EC4401
12: Escherichia coli 0157:H7 str. EDL933
13: Escherichia coli 0157:H7 str. TW14588
14: Escherichia coli 0157:H7 str. Sakai
15: Escherichia coli 0157:H7 EC4045
16: Escherichia coli 0157:H7 str. LANL ECF
17: Escherichia coli 0157:H7 str. LANL ECA
18: Escherichia coli K12 str. DH10B
19: Escherichia coli K12 str. MG1655
20: Escherichia coli K12 str. W3110
21: Escherichia coli K12 str. DH1
22: Escherichia coli BW2952
23: Escherichia coli ATCC8739
24: Escherichia coli B REL606
25: Escherichia coli BL21 (DE3 Korea)
26: Escherichia coli BL21 (DE3 AU)
27: Escherichia coli BL21 (DE3 DOE)
28: Escherichia coli HS
29: Escherichia coli SE11
30: Escherichia coli IA1
31: Escherichia coli 55989
32: Escherichia coli E24377A
33: Escherichia coli O26:H11 str. 11368
34: Escherichia coli 0127:H6 str. E2348/69
35: Escherichia coli O103:H2 str. 12009
36: Escherichia coli O111:H- str. 11128
37: Escherichia coli O103 Oslo
38: Escherichia coli SMS-3-5
39: Escherichia coli UMN026
40: Escherichia coli 53638
41: Escherichia coli IA39
42: Escherichia coli UTI89
43: Escherichia coli SB6
44: Escherichia coli CFT073
45: Escherichia coli SE15
46: Escherichia coli 536
47: Escherichia coli ED1a
48: Escherichia coli F11
49: Escherichia coli APEC01
50: Escherichia coli E110019
51: Escherichia coli E22
52: Escherichia coli BTA
53: Escherichia coli 101-1
54: Shigella flexneri 2a 2457T
55: Shigella flexneri 2a 301
56: Shigella flexneri 5 8401
57: Shigella boydii CDC 3083–94
58: Shigella boydii Sb227
59: Shigella sonneti Ss046
60: Escherichia fergusonii ATCC 35469
61: Escherichia albertii TW07627
62: Salmonella enterica Typhimurium LT2
63: Shigella dysenteriae Sd107
64: Shigella dysenteriae 1012
Who cares? What are pan-genomes good for?

Standard operating procedure for computing pangenome trees

Standards in Genomic Sciences (2010) 2:135-141

Figure 3 Pan-genome clustering of E. coli (black) and related species (colored), based on the alignment of their variable gene content. The genomes now cluster according to species and a relatedness between E. coli K12 derivatives (green block) and group B isolates (orange block) is visible.

CBS, Department of Systems Biology

Biological Sequence Analysis (DTU course #27803)


5 May, 2011
Where do the ‘accessory genes’ in the pan-genome live?

Comparison of 61 Sequenced *Escherichia coli* Genomes

**Figure 5** BLAST atlas. In the middle, a genome atlas of *E. coli* O157:H7 strain EC4115 is shown, around which BLAST lanes are shown. Every lane corresponds to a genome, with the following colors (going outwards): green *E. coli* O157: H7 (15 lanes); light blue *E. coli* LANL strains (two lanes); dark blue *Shigella* spp. (eight lanes); red *E. coli* K12 and derivatives (six lanes); orange *E. coli* strain B phylogroup (four lanes); followed by all other *E. coli* genomes in different colors. The outermost three lanes represent *E. fergusonii*, *E. albertii*, and *S. enterica* Typhimurium LT2. Lack of color indicates that the genes at that position in strain EC4115 were not found in the genome of that lane. The position of replication origin and terminus is indicated.


CBS, Department of Systems Biology

Biological Sequence Analysis (DTU course #27803)
‘Third Generation’ sequencing...

1. Pacific Biosciences

2. Oxford Nanopore
Sequencing Machine Helped Trace Cholera in Haiti

By ANDREW POLLACK

Nothing like a public health crisis to put a new technology through its paces.

Scientists on Thursday said they used new technology to rapidly sequence the genome of the cholera bacterium that has killed more than 2,000 people in Haiti and sickened nearly 100,000.

The analysis confirmed one from the Centers of Disease Control and Prevention that the strain appears to have come from South Asia, not from Latin America. The study was published online late Thursday by The New England Journal of Medicine.

Dr. Matthew K. Waldor, an infectious disease specialist at Harvard Medical School and a senior author of the paper, said one implication was that the disease was transmitted by people, not carried to Haiti on water currents. In the future, he said, people entering countries that might be vulnerable to cholera should be screened for the disease.

“If that was the policy, we could have prevented the Haitian epidemic,” he said.

Yet another implication, he said, is the growing role that fast genome sequencing can play in tracking infectious disease outbreaks.

Dr. Waldor turned to Pacific Biosciences, which next year is planning to begin selling a so-called third generation sequencing machine for $695,000. It has the potential to vastly increase the pace and lower the cost of DNA sequencing. While other companies also make such claims, PacBio has been successful in raising hundreds of millions of dollars from investors, including $200 million in its initial public stock offering in October.

Seeing a chance to prove and publicize its technology, PacBio worked night and day on the cholera project. The bacterium, known as Vibrio cholerae, has about 4.5 million bases of DNA in its genome, compared to 3 billion for the human genome.

The company received the samples at its headquarters in Menlo Park, Calif., on Nov. 10. By midday the next day, it had some raw sequences done. By Nov. 15, it had sequenced and analyzed not only two strains from Haiti but two from Bangladesh and one from Peru. The draft of the paper was submitted to the journal on Nov. 19.

“It was definitely intense,” said Eric Schadt, the company’s chief scientific officer and an author of the paper.

Dr. Schadt said the sequencing might be quick enough to allow construction and updating of a “disease weather map,” something the company is trying to develop for San Francisco’s Bay Area. Samples can be taken every day from various spots and analyzed and mapped to see how different germs are spreading, perhaps providing early warning of an impending outbreak.

The conclusion that the Haitian disease came from South Asia is sensitive. There have already been riots directed at United Nations peacekeeping forces there because of suspicions that Nepalese soldiers brought the disease.

The new study did not have a Nepalese strain for comparison so it is impossible to say if that country was the source, only that the Haitian strains closely resembled a 2008 strain from Bangladesh.
The Origin of the Haitian Cholera Outbreak Strain

Chen-Shan Chin, Ph.D., Jon Sorenson, Ph.D., Jason B. Harris, M.D., William P. Robins, Ph.D., Richelle C. Charles, M.D., Roger R. Jean-Charles, M.D., James Bullard, Ph.D., Dale R. Webster, Ph.D., Andrew Kasarskis, Ph.D., Paul Peluso, Ph.D., Ellen E. Paxinos, Ph.D., Yoshiharu Yamaichi, Ph.D., Stephen B. Calderwood, M.D., John J. Mekalanos, Ph.D., Eric E. Schadt, Ph.D., and Matthew K. Waldor, M.D., Ph.D.

ABSTRACT

BACKGROUND
Although cholera has been present in Latin America since 1991, it had not been epidemic in Haiti for at least 100 years. Recently, however, there has been a severe outbreak of cholera in Haiti.

METHODS
We used third-generation single-molecule real-time DNA sequencing to determine the genome sequences of 2 clinical Vibrio cholerae isolates from the current outbreak in Haiti, 1 strain that caused cholera in Latin America in 1991, and 2 strains isolated in South Asia in 2002 and 2008. Using primary sequence data, we compared the genomes of these 5 strains and a set of previously obtained partial genomic sequences of 23 diverse strains of V. cholerae to assess the likely origin of the cholera outbreak in Haiti.

RESULTS
Both single-nucleotide variations and the presence and structure of hypervariable chromosomal elements indicate that there is a close relationship between the Haitian isolates and variant V. cholerae El Tor O1 strains isolated in Bangladesh in 2002 and 2008. In contrast, analysis of genomic variation of the Haitian isolates reveals a more distant relationship with circulating South American isolates.

CONCLUSIONS
The Haitian epidemic is probably the result of the introduction, through human activity, of a V. cholerae strain from a distant geographic source. (Funded by the National Institute of Allergy and Infectious Diseases and the Howard Hughes Medical Institute.)
A window into third-generation sequencing

Eric E. Schadt*, Steve Turner and Andrew Kasarskis

Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, USA

Received September 15, 2010; Revised and Accepted September 17, 2010

First- and second-generation sequencing technologies have led the way in revolutionizing the field of genomics and beyond, motivating an astonishing number of scientific advances, including enabling a more complete understanding of whole genome sequences and the information encoded therein, a more complete characterization of the methylome and transcriptome and a better understanding of interactions between proteins and DNA. Nevertheless, there are sequencing applications and aspects of genome biology that are presently beyond the reach of current sequencing technologies, leaving fertile ground for additional innovation in this space. In this review, we describe a new generation of single-molecule sequencing technologies (third-generation sequencing) that is emerging to fill this space, with the potential for dramatically longer read lengths, shorter time to result and lower overall cost.
Table 1. Comparison of first-generation sequencing, SGS and TGS

<table>
<thead>
<tr>
<th></th>
<th>First generation</th>
<th>Second generationa</th>
<th>Third generationa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental technology</td>
<td>Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation</td>
<td>Wash-and-scan SBS</td>
<td>SBS, by degradation, or direct physical inspection of the DNA molecule</td>
</tr>
<tr>
<td>Resolution</td>
<td>Averaged across many copies of the DNA molecule being sequenced</td>
<td>Averaged across many copies of the DNA molecule being sequenced</td>
<td>Single-molecule resolution</td>
</tr>
<tr>
<td>Current raw read accuracy</td>
<td>High</td>
<td>High</td>
<td>Moderate</td>
</tr>
<tr>
<td>Current read length</td>
<td>Moderate (800–1000 bp)</td>
<td>Short, generally much shorter than Sanger sequencing</td>
<td>Long, 1000 bp and longer in commercial systems</td>
</tr>
<tr>
<td>Current throughput</td>
<td>Low</td>
<td>High</td>
<td>Moderate</td>
</tr>
<tr>
<td>Current cost</td>
<td>High cost per base</td>
<td>Low cost per base</td>
<td>Low-to-moderate cost per base</td>
</tr>
<tr>
<td></td>
<td>Low cost per run</td>
<td>High cost per run</td>
<td>Low cost per run</td>
</tr>
<tr>
<td>RNA-sequencing method</td>
<td>cDNA sequencing</td>
<td>cDNA sequencing</td>
<td>Direct RNA sequencing and cDNA sequencing</td>
</tr>
<tr>
<td>Time from start of sequencing reaction to result</td>
<td>Hours</td>
<td>Days</td>
<td>Hours</td>
</tr>
<tr>
<td>Sample preparation</td>
<td>Moderately complex, PCR amplification not required</td>
<td>Complex, PCR amplification required</td>
<td>Ranges from complex to very simple depending on technology</td>
</tr>
<tr>
<td>Data analysis</td>
<td>Routine</td>
<td>Complex because of large data volumes and because short reads complicate assembly and alignment algorithms</td>
<td>Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges</td>
</tr>
<tr>
<td>Primary results</td>
<td>Base calls with quality values</td>
<td>Base calls with quality values</td>
<td>Base calls with quality values, potentially other base information such as kinetics</td>
</tr>
</tbody>
</table>

aThere are many TGS technologies in development but few have been reduced to practice. While there is significant potential of TGS to radically improve current throughput and read-length characteristics (among others), the ultimate practical limits of these technologies remain to be explored. Furthermore, there is active development of SGS technologies that will also improve read-length and throughput characteristics.
ZMWs overcome the first obstacle, but not the second. All SGS technologies directly attach the dye to the base, which is incorporated into the DNA strand. This is problematic for any system attempting to observe DNA synthesis in real time because the dye's large size relative to the DNA can interfere with the activity of the DNA polymerase. Typically, a DNA polymerase is confined in a zero-mode waveguide and base additions are measured with florescence detection of gamma-labeled phosphonucleotides. (B) Several companies seek to sequence DNA by direct inspection using electron microscopy similar to the Reveo technology pictured here, in which an ssDNA molecule is first stretched and then examined by STM. (C) Oxford Nanopore technology for measuring translocation of nucleotides cleaved from a DNA molecule across a pore, driven by the force of differential ion concentrations across the membrane. (D) IBM's DNA transistor technology reads individual bases of ssDNA molecules as they pass through a narrow aperture based on the unique electronic signature of each individual nucleotide. Gold bands represent metal and gray bands dielectric layers of the transistor.

Figure 2. How third-generation DNA-sequencing technologies work. Third-generation DNA-sequencing technologies are distinguished by direct inspection of single molecules with methods that do not require wash steps during DNA synthesis. (A) Pacific Biosciences technology for direct observation of DNA synthesis on single DNA molecules in real time. A DNA polymerase is confined in a zero-mode waveguide and base additions are measured with florescence detection of gamma-labeled phosphonucleotides. (B) Several companies seek to sequence DNA by direct inspection using electron microscopy similar to the Reveo technology pictured here, in which an ssDNA molecule is first stretched and then examined by STM. (C) Oxford Nanopore technology for measuring translocation of nucleotides cleaved from a DNA molecule across a pore, driven by the force of differential ion concentrations across the membrane. (D) IBM's DNA transistor technology reads individual bases of ssDNA molecules as they pass through a narrow aperture based on the unique electronic signature of each individual nucleotide. Gold bands represent metal and gray bands dielectric layers of the transistor.