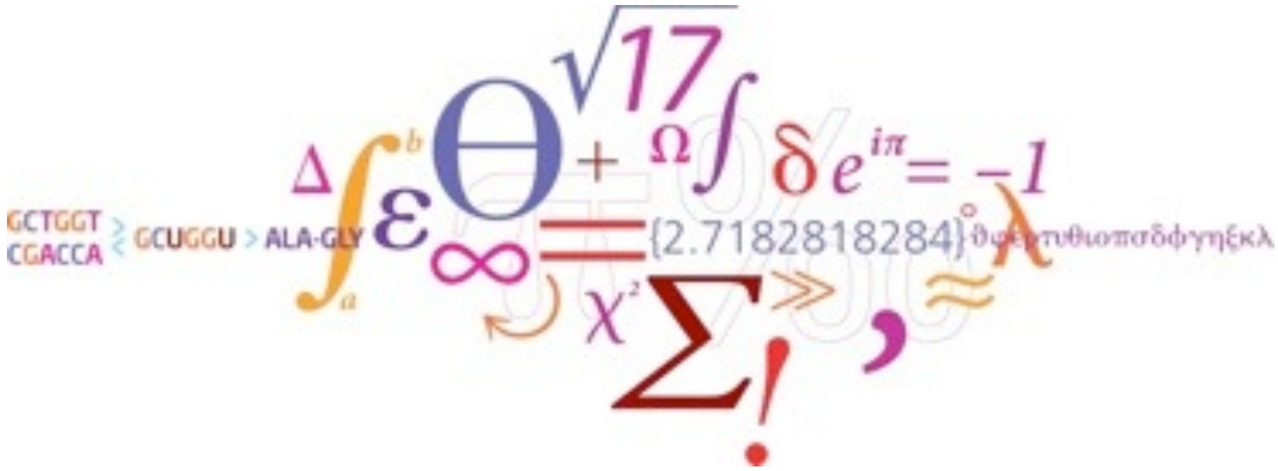
Genome Sequencing

First, ‘next-generation’, and ‘third generation’



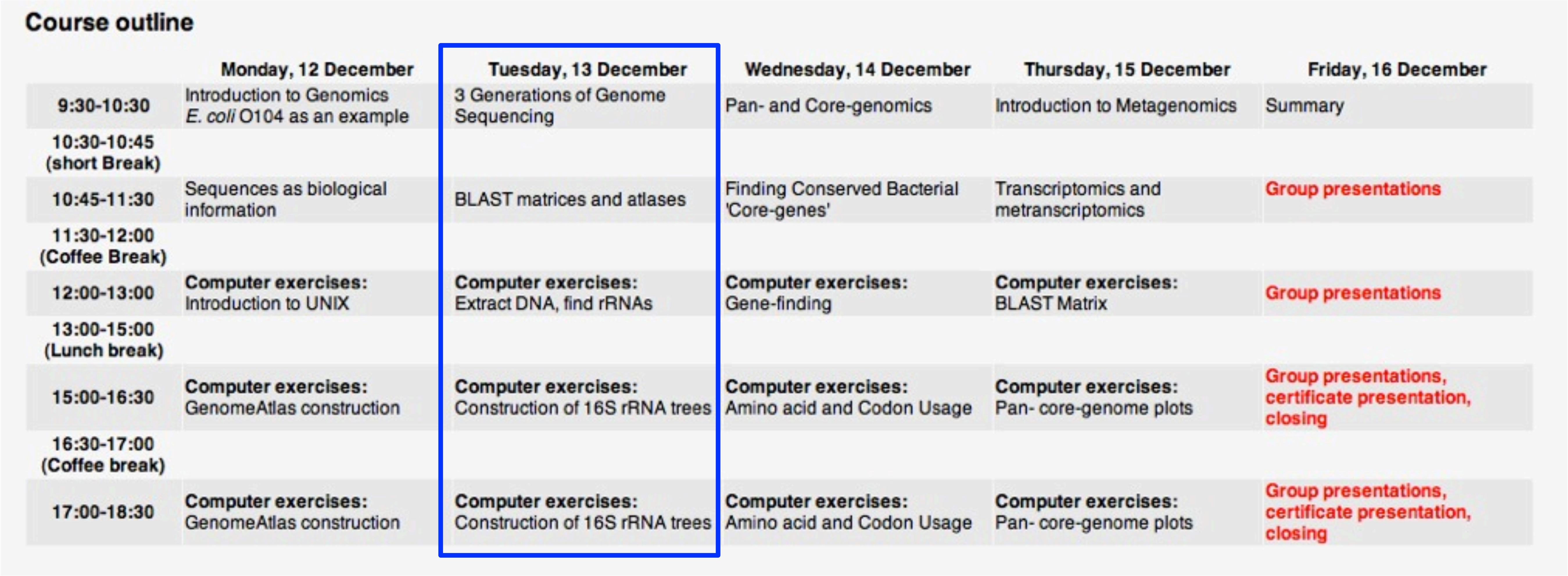
Dave Ussery

Genomics of Prokaryotes Workshop
Universidad Miguel Hernández
Alicante, Spain

Tuesday, 13 December, 2011

Center for Biological Sequence Analysis
Department of Systems Biology

http://egg.umh.es/course.html



Course outline

	Monday, 12 December	Tuesday, 13 December	Wednesday, 14 December	Thursday, 15 December	Friday, 16 December
9:30-10:30	Introduction to Genomics *E. coli* O104 as an example	3 Generations of Genome Sequencing	Pan- and Core-genomics	Introduction to Metagenomics	Summary
10:30-10:45 (short Break)					
10:45-11:30	Sequences as biological information	BLAST matrices and atlases	Finding Conserved Bacterial 'Core-genes'	Transcriptomics and metranscriptomics	Group presentations
11:30-12:00 (Coffee Break)					
12:00-13:00	**Computer exercises:** Introduction to UNIX	**Computer exercises:** Extract DNA, find rRNAs	**Computer exercises:** Gene-finding	**Computer exercises:** BLAST Matrix	Group presentations
13:00-15:00 (Lunch break)					
15:00-16:30	**Computer exercises:** GenomeAtlas construction	**Computer exercises:** Construction of 16S rRNA trees	**Computer exercises:** Amino acid and Codon Usage	**Computer exercises:** Pan- core-genome plots	Group presentations, certificate presentation, closing
16:30-17:00 (Coffee break)					
17:00-18:30	**Computer exercises:** GenomeAtlas construction	**Computer exercises:** Construction of 16S rRNA trees	**Computer exercises:** Amino acid and Codon Usage	**Computer exercises:** Pan- core-genome plots	Group presentations, certificate presentation, closing

http://www.cbs.dtu.dk/staff/dave/genomics_course/Alicante2011.php

‘First Generation’ sequencing

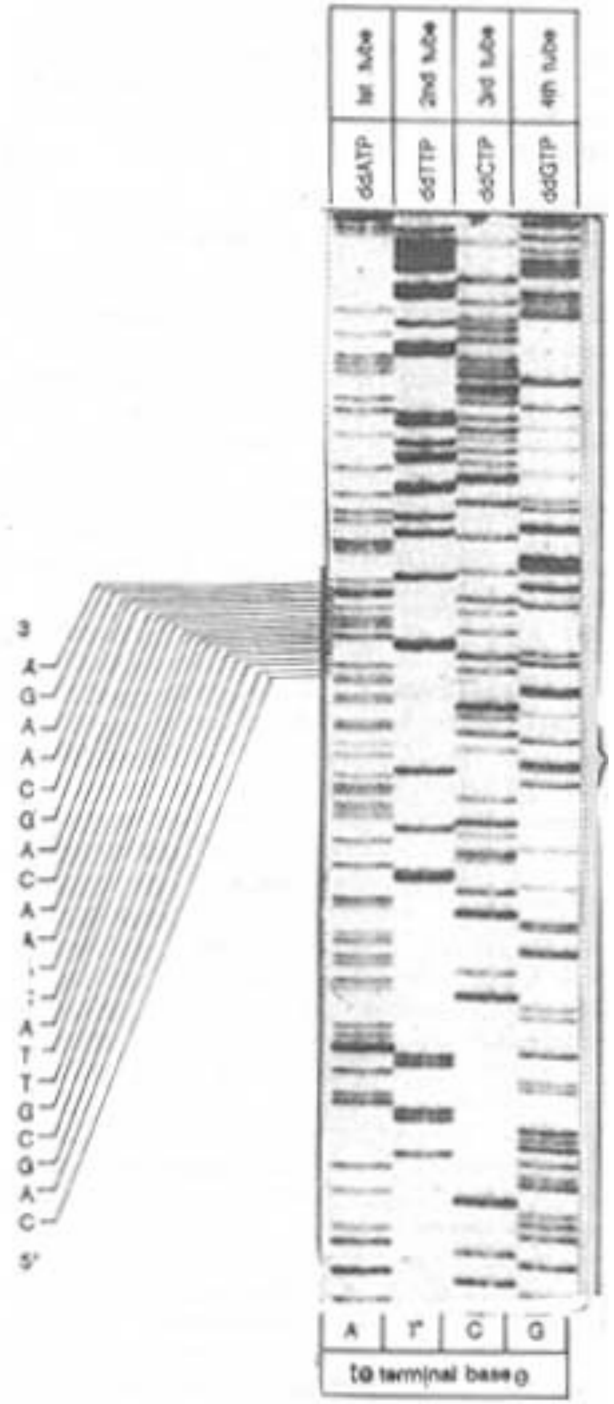
(i) Four reaction tubes are set up, each containing single stranded DNA sample (cloned in M13 phage) to be sequenced, all the four dNTPs, an oligonucleotide sequencing primer (radioactively labelled) and an enzyme for DNA synthesis (DNA polymerase I = sequenase). Each tube also contains a small amount (much smaller amount relative to four dNTPs) of one of the four ddNTP, bringing about termination at a specific base-adenine (A), cytosine (C), guanine (G) and thymine (T).

(ii) The fragments, generated by random incorporation of ddNTP leading to termination of reaction, are then separated by electrophoresis on a high resolution polyacrylamide gel. This is done f r all the four reaction mixtures on adjoining lanes in the gel.

(iii) The gel' is used for autoradiography so that the position of different bands in each lane can be visualized.

(iv) The bands on autoradiogram can be used for getting the DNA sequence as shown in.

An Autiograph of a Sangers sequencing Gel



‘First Generation’ sequencing

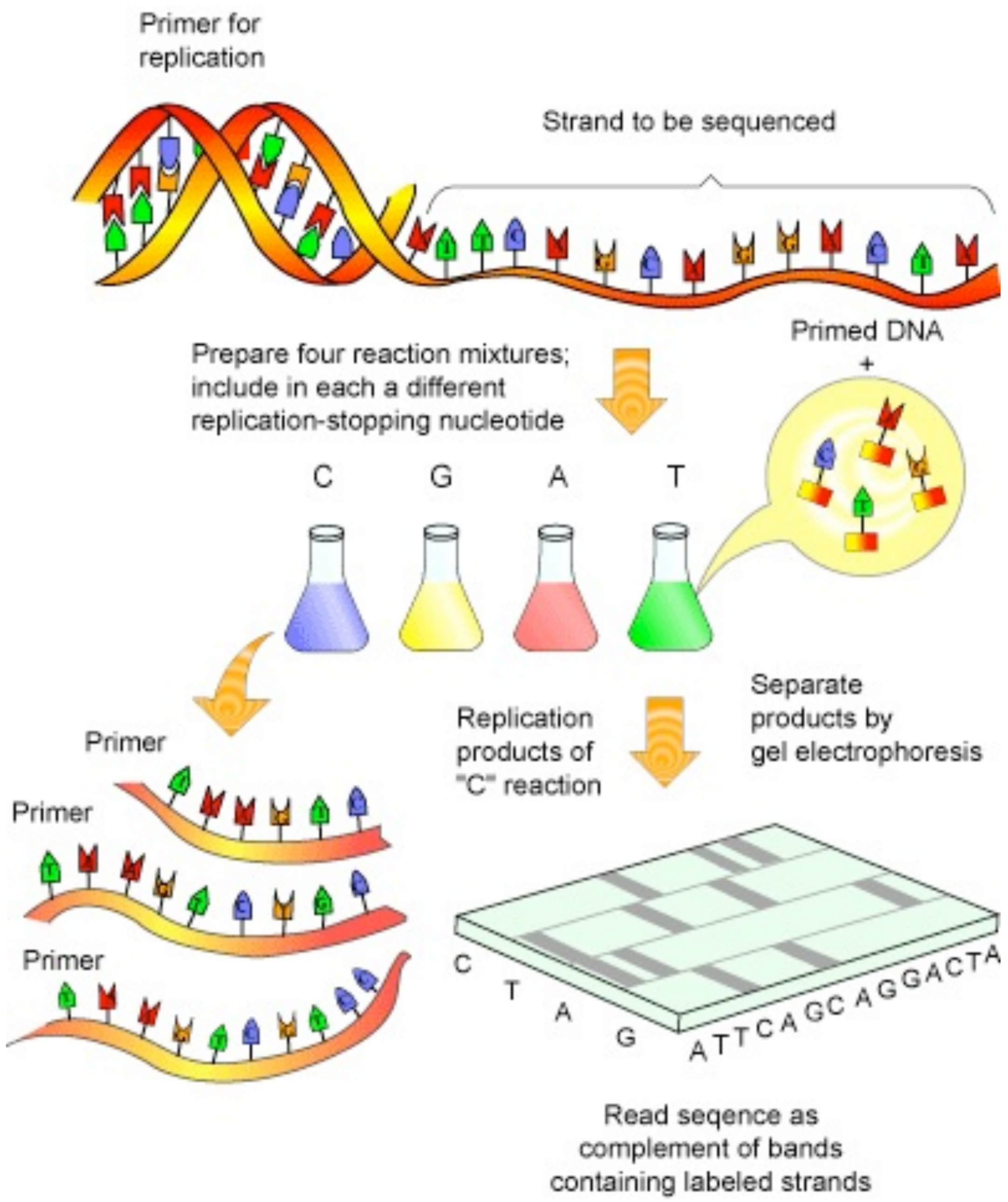

Primer for replication
Strand to be sequenced
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
Primed DNA
+
C
G
A
T
Primer
Primer
Primer
Replication products of "C" reaction
Separate products by gel electrophoresis
C
T
A
G
ATTCAGCAGGACTA
Read seqence as complement of bands containing labeled strands

Figure 1. The Sanger sequencing reaction. *Single stranded DNA is amplified in the presence of fluorescently labelled ddNTPs that serve to terminate the reaction and label all the fragments of DNA produced. The fragments of DNA are then separated via polyacrylamide gel electrophoresis and the sequence read using a laser beam and computer.*

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.

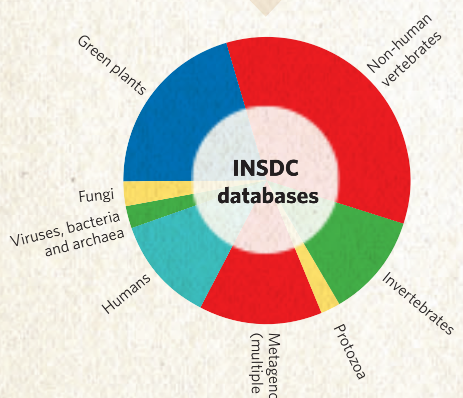
THE SEQUENCE EXPLOSION

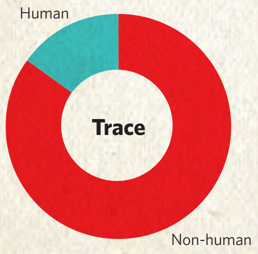
At the time of the announcement of the first drafts of the human genome in 2000, there were 8 billion base pairs of sequence in the three main databases for 'finished' sequence: GenBank, run by the US National Center for Biotechnology Information; the DNA Databank of Japan; and the European Molecular 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).

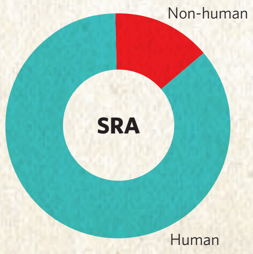
See Editorial, page 649, and human genome special at www.nature.com/humangenome

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.

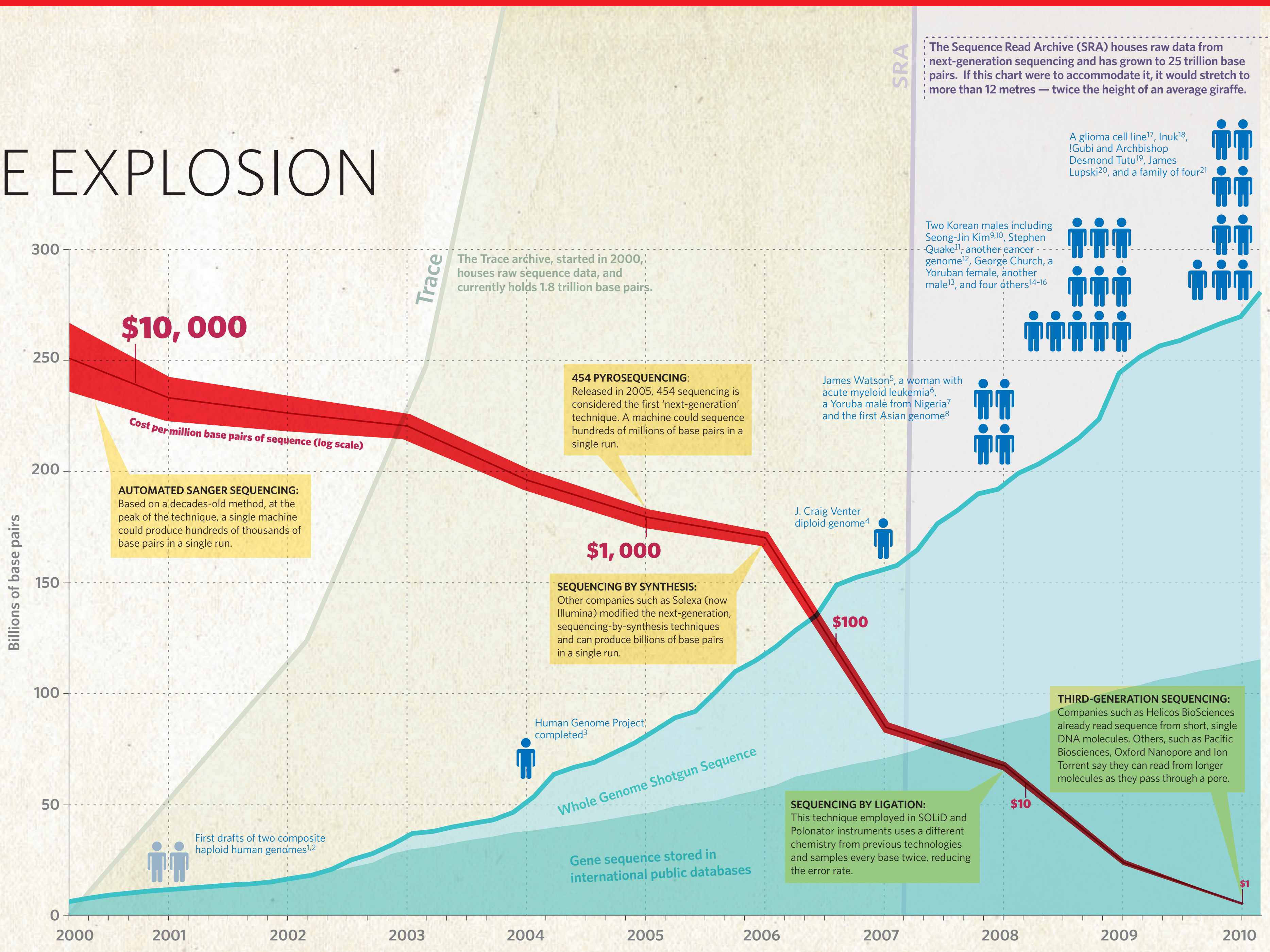






Trace Archive: Developed to house the raw output of high-throughput sequencers built in the late 1990s, the trace archive spans a wide range of taxa.

Sequence Read Archive: Houses raw data from next-generation sequencers. Dominated by human sequence, including multiple coverage for more than 170 people.



The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.

The Sequence Read Archive (SRA) houses raw data from next-generation sequencing and has grown to 25 trillion base pairs. If this chart were to accommodate it, it would stretch to more than 12 metres — twice the height of an average giraffe.

AUTOMATED SANGER SEQUENCING: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

454 PYROSEQUENCING: Released in 2005, 454 sequencing is considered the first 'next-generation' technique. A machine could sequence hundreds of millions of base pairs in a single run.

SEQUENCING BY SYNTHESIS: Other companies such as Solexa (now Illumina) modified the next-generation, sequencing-by-synthesis techniques and can produce billions of base pairs in a single run.

SEQUENCING BY LIGATION: This technique employed in SOLiD and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

THIRD-GENERATION SEQUENCING: Companies such as Helicos BioSciences already read sequence from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.

First drafts of two composite haploid human genomes[1,2]

Human Genome Project completed[3]

J. Craig Venter diploid genome[4]

James Watson[5], a woman with acute myeloid leukemia[6], a Yoruba male from Nigeria[7] and the first Asian genome[8]

Two Korean males including Seong-Jin Kim[9,10], Stephen Quake[11], another cancer genome[12], George Church, a Yoruban female, another male[13], and four others[14–16]

A glioma cell line[17], Inuk[18], !Gubi and Archbishop Desmond Tutu[19], James Lupski[20], and a family of four[21]

HOW MANY HUMAN GENOMES?

The graphic shows all published, fully sequenced human genomes since 2000, including nine from the first quarter of 2010. Some are resequencing efforts on the same person and the list does not include unpublished completed genomes.

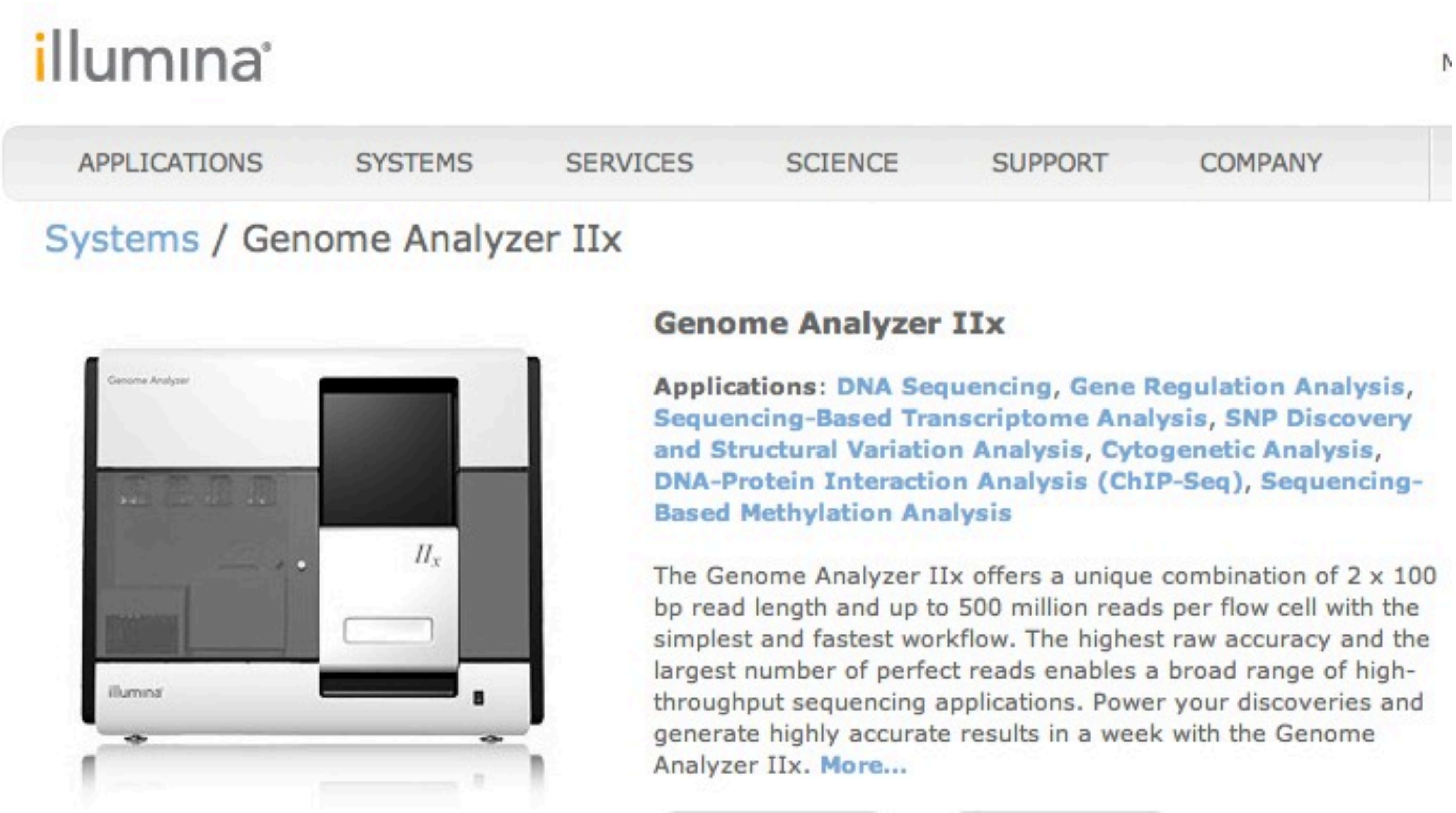
1. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
2. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
3. International Human Genome Sequencing Consortium *Nature* **431,** 931–945 (2004).
4. Levy, S. *et al. PLoS Biol.* **5,** e254 (2007).
5. Wheeler, D. A. *et al. Nature* **452,** 872–876 (2008).
6. Ley, T. J. *et al. Nature* **456,** 66–72 (2008).
7. Bentley, D. R. *et al. Nature* **456,** 53–59 (2008).
8. Wang, J. *et al. Nature* **456,** 60–65 (2008).
9. Ahn, S.-M. *et al. Genome Res.* **19,** 1622–1629 (2009).
10. Kim, J.-I. *et al. Nature* **460,** 1011–1015 (2009).
11. Pushkarev, D., Neff, N. F. & Quake, S. R. *Nature Biotechnol.* **27,** 847–850 (2009).
12. Mardis, E. R. *et al. N. Engl. J. Med.* **10,** 1058–1066 (2009).
13. Drmanac, R. *et al. Science* **327,** 78–81 (2009).
14. McKernan, K. J. *et al. Genome Res.* **19,** 1527–1541 (2009).
15. Pleasance, E. D. *et al. Nature* **463,** 191–196 (2010).
16. Pleasance, E. D. *et al. Nature* **463,** 184–190 (2010).
17. Clark, M. J. *et al. PLoS Genet.* **6,** e1000832 (2010).
18. Rasmussen, M. *et al. Nature* **463,** 757–762 (2010).
19. Schuster, S. C. *et al. Nature* **463,** 943–947 (2010).
20. Lupski, J. R. *et al. N. Engl. J. Med.* doi:10.1056/NEJMoa0908094 (2010).
21. Roach, J. C. *et al. Science* doi:10.1126/science.1186802 (2010).

Page size by comparison

SOURCE: NCBI; GRAPHICS BY N. SPENCER & W. FERNANDES

Three Current “next-generation” technologies:

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



Systems / Genome Analyzer IIx

Genome Analyzer IIx

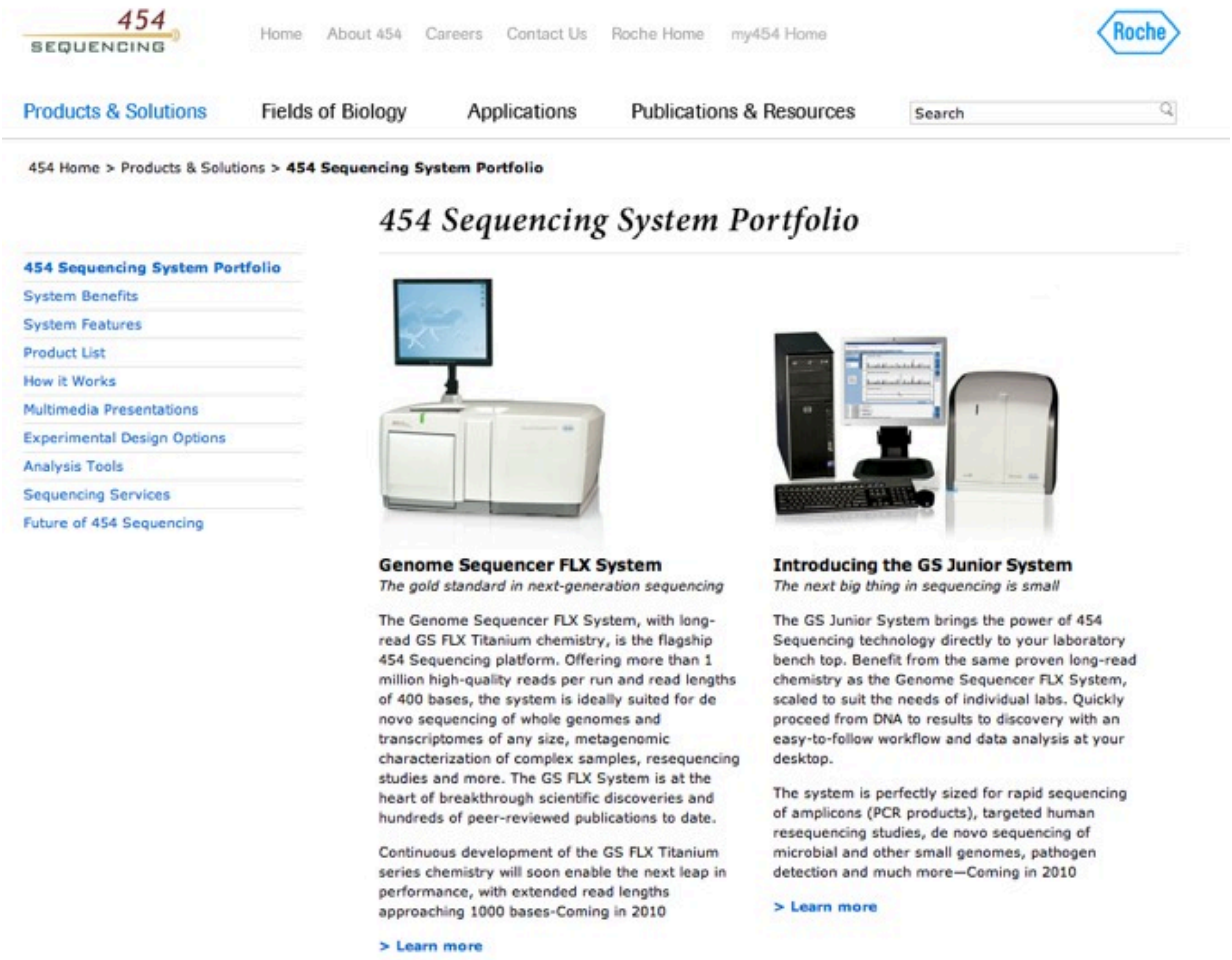
Applications: DNA Sequencing, Gene Regulation Analysis, Sequencing-Based Transcriptome Analysis, SNP Discovery and Structural Variation Analysis, Cytogenetic Analysis, DNA-Protein Interaction Analysis (ChIP-Seq), Sequencing-Based Methylation Analysis

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...

Three Current “next-generation” technologies:

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

2. Roche 454



454 Sequencing System Portfolio

Genome Sequencer FLX System

The gold standard in next-generation sequencing

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

Three Current "next-generation" technologies:

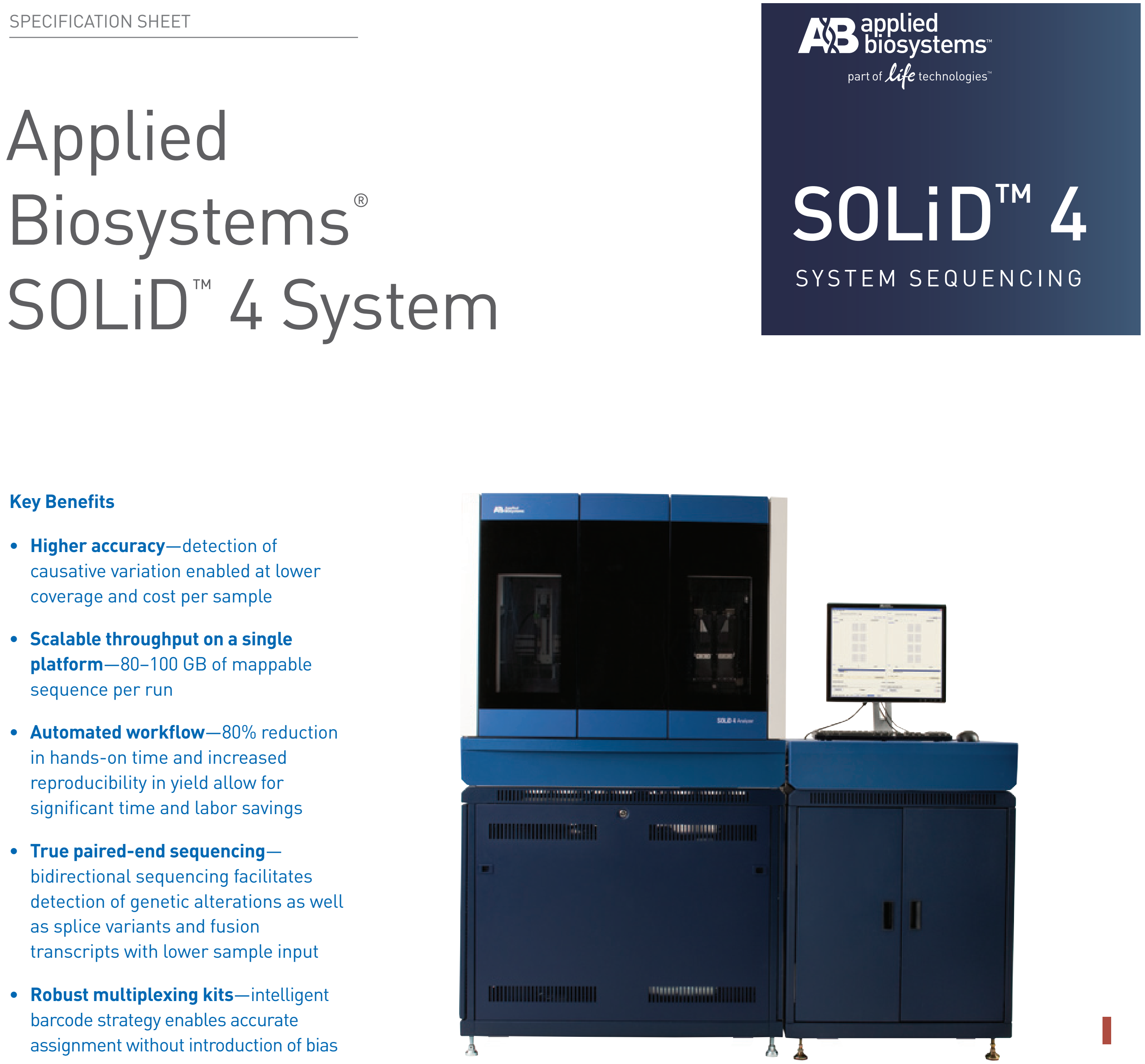
1. Illumina ("Solexa") - 500 million reads (100 bp)

2. Roche 454

3. ABI SOLiD

~100 Gbp per run!

35 bp reads



SPECIFICATION SHEET

Applied Biosystems® SOLiD™ 4 System

applied biosystems™
part of Life technologies™

SOLiD™ 4

SYSTEM SEQUENCING

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

SOLiD™ Sequencing and 2-Base Encoding

Poster Number
2624

Heather E. Peckham[1], Stephen F. McLaughlin[1], Jingwei N. Ni[2], Michael D. Rhodes[2], Joel A. Malek[1], Kevin J. McKernan[1] and Alan P. Blanchard[1]
1. Applied Biosystems, 500 Cummings Center, Beverly, MA 01915
2. Applied Biosystems, 850 Lincoln Centre Dr, Foster City, CA 94404

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.

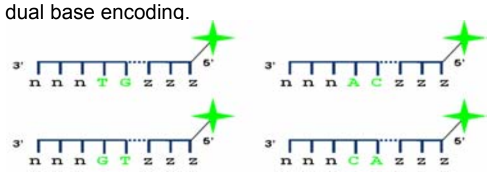


Figure 1. Each probe consists of 8 bases. As shown, the first 3 bases are degenerate (n), and the last 3 are universal (z), with the 4th and 5th bases as the two bases being interrogated. Thus, a single color observation only limits the potential dinucleotide to being four out of the 16 possible dinucleotides. As seen above, a green signal represents a AC, CA, TG or GT.

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.

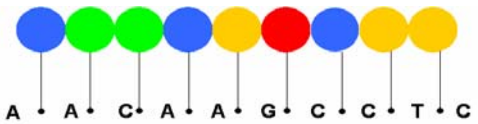


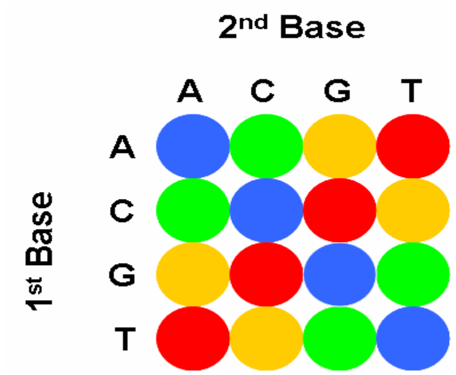
Figure 2 demonstrates the principle of double interrogation. Each color measurement represents four possible dinucleotide combinations. For example, the first measured blue represents 'AA' and the third blue represents 'CC'.

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 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:



	A	C	G	T
A	🔵	🟢	🟡	🔴
C	🟢	🔵	🔴	🟡
G	🟡	🔴	🔵	🟢
T	🔴	🟡	🟢	🔵

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 Base Insertions/Deletions

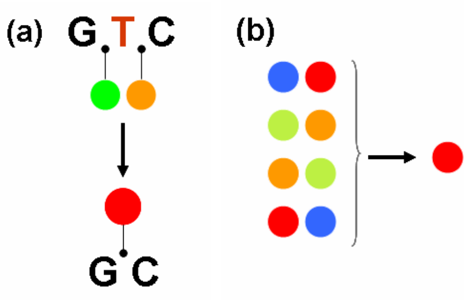


Figure 4. (a) If a deletion occurs in the sequence GTC the result has to be GC. The number of observed transitions will decrease from 2 to1. The single transition must be a G to C thus giving a signature to the event. **(b)** The reverse is true if a single base insertion occurs with the result that only 4 of the potential adjacent transitions can occur for any individual starting transition.

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:

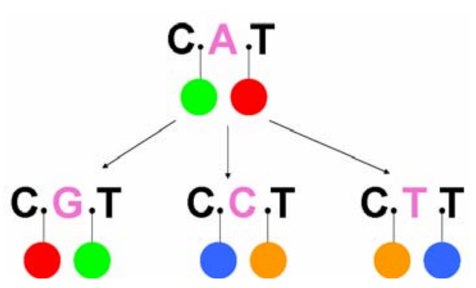


Figure 5. If a SNP occurs in the sequence 'C-A-T' there are only 3 possible results: CGT, CCT and CTT. This means that only 3 dibase combinations are allowed and any other dibase combinations are illegal. Since any base is defined by two nucleotides (e.g., C-A and A-T), then two adjacent changes must 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 dibase 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^{TM} Accuracy

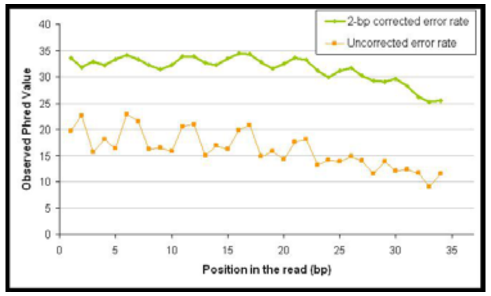
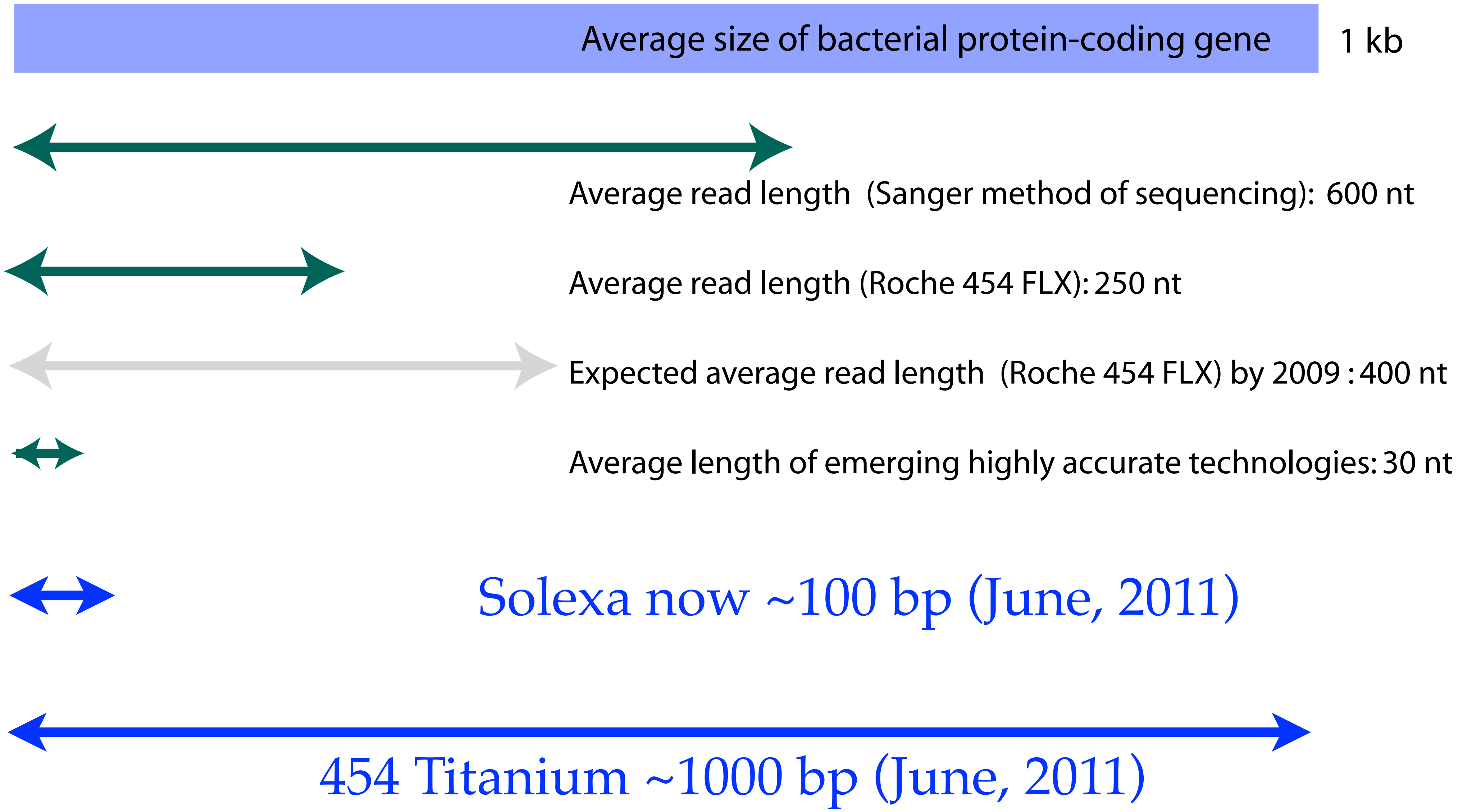


Figure 6: SOLiD^{TM} System's error rate per base position in sequence read.

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.



Average size of bacterial protein-coding gene 1 kb

Average read length (Sanger method of sequencing): 600 nt

Average read length (Roche 454 FLX): 250 nt

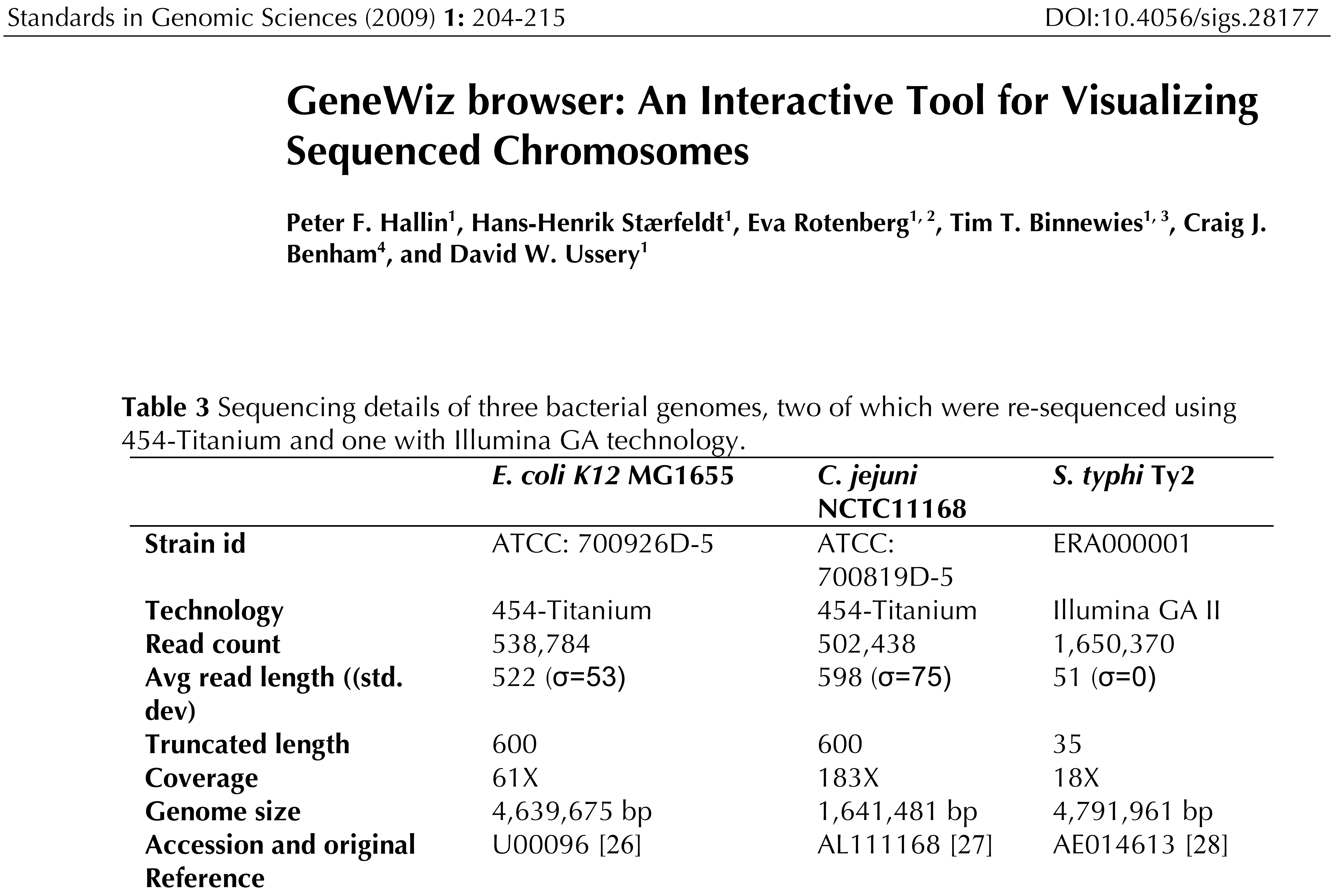
Expected average read length (Roche 454 FLX) by 2009 : 400 nt

Average length of emerging highly accurate technologies: 30 nt

Solexa now ~100 bp (June, 2011)

454 Titanium ~1000 bp (June, 2011)

D.W. Ussery et al., *Computing for Comparative Microbial Genomics*, Computational Biology 8, DOI 10.1007/978-1-84800-255-5_14, © Springer-Verlag London Limited 2009



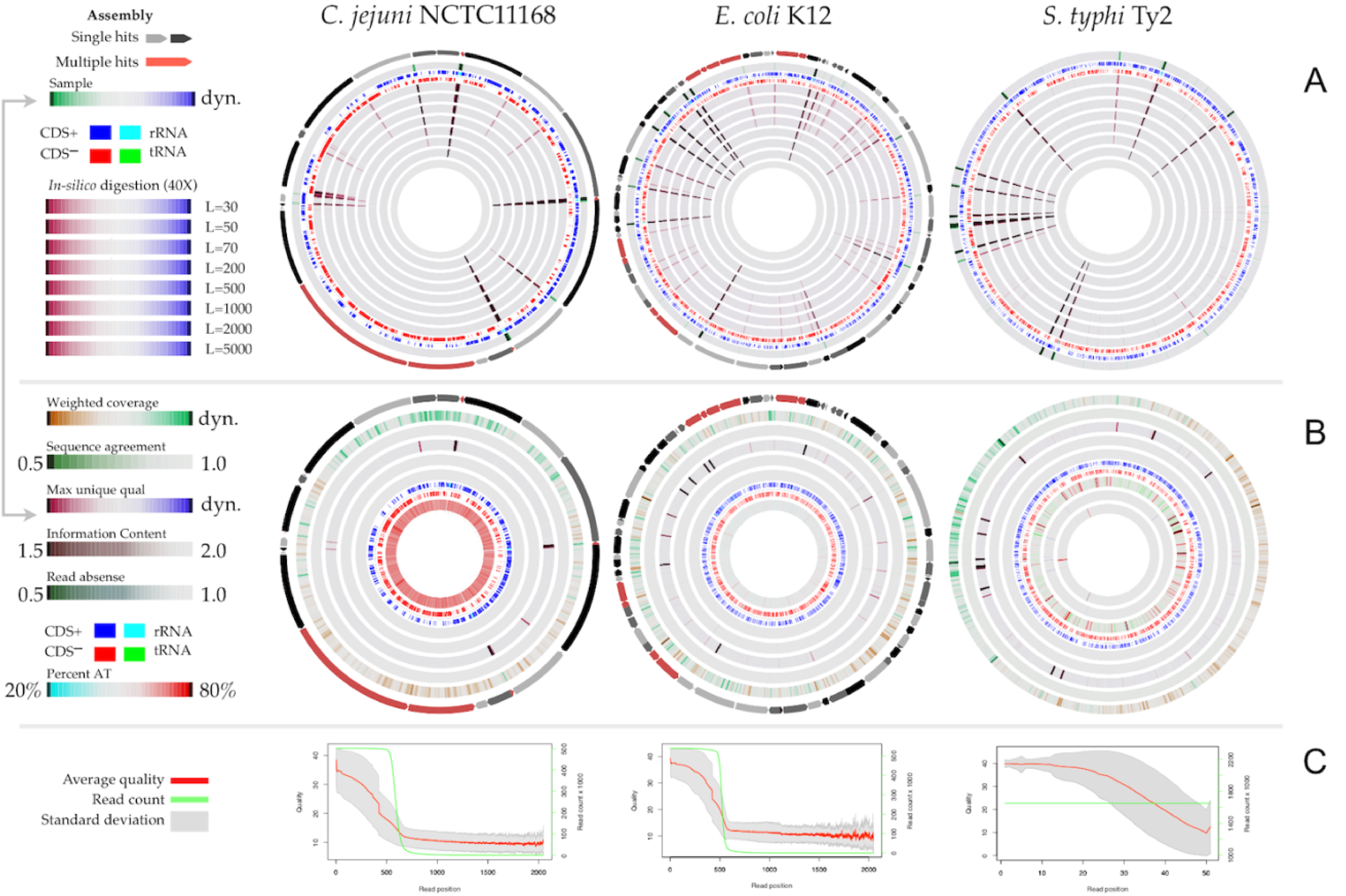
Standards in Genomic Sciences (2009) **1:** 204-215 DOI:10.4056/sigs.28177

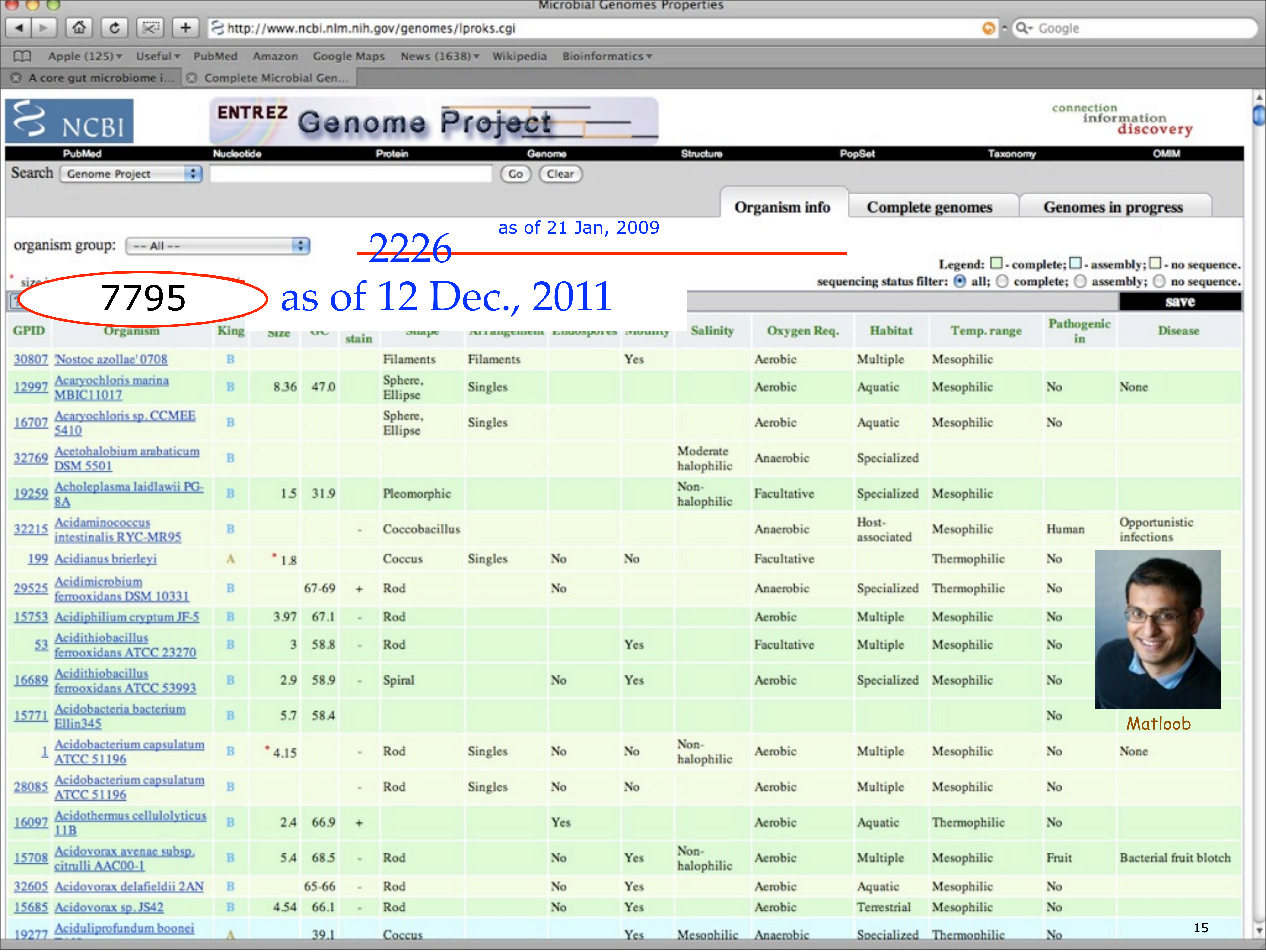
GeneWiz browser: An Interactive Tool for Visualizing Sequenced Chromosomes

Peter F. Hallin[1], Hans-Henrik Stærfeldt[1], Eva Rotenberg[1, 2], Tim T. Binnewies[1, 3], Craig J. Benham[4], and David W. Ussery[1]

Table 3 Sequencing details of three bacterial genomes, two of which were re-sequenced using 454-Titanium and one with Illumina GA technology.

	***E. coli K12* MG1655**	***C. jejuni* NCTC11168**	***S. typhi* Ty2**
Strain id	ATCC: 700926D-5	ATCC: 700819D-5	ERA000001
Technology	454-Titanium	454-Titanium	Illumina GA II
Read count	538,784	502,438	1,650,370
Avg read length ((std. dev)	522 (σ=53)	598 (σ=75)	51 (σ=0)
Truncated length	600	600	35
Coverage	61X	183X	18X
Genome size	4,639,675 bp	1,641,481 bp	4,791,961 bp
Accession and original Reference	U00096 [26]	AL111168 [27]	AE014613 [28]


Assembly
Single hits
Multiple hits
Sample
dyn.
CDS+
rRNA
CDS−
tRNA
In-silico digestion (40X)
L=30
L=50
L=70
L=200
L=500
L=1000
L=2000
L=5000
Weighted coverage
dyn.
Sequence agreement
0.5
1.0
Max unique qual
dyn.
Information Content
1.5
2.0
Read absense
0.5
1.0
CDS+
rRNA
CDS−
tRNA
Percent AT
20%
80%
Average quality
Read count
Standard deviation
C. jejuni NCTC11168
E. coli K12
S. typhi Ty2
A
B
C
Quality
Read count x 1000
Read position
0
500
1000
1500
2000
10
20
30
40
100
200
300
400
500
1000
1400
1800
2200
50



Search: Genome Project

organism group: -- All --

~~2226~~ as of 21 Jan, 2009

7795 as of 12 Dec., 2011

Legend: □ - complete; □ - assembly; □ - no sequence.
sequencing status filter: ● all; ○ complete; ○ assembly; ○ no sequence.

GPID	Organism	King	Size	GC	stain	Shape	Arrangement	Endospores	Motility	Salinity	Oxygen Req.	Habitat	Temp. range	Pathogenic in	Disease
30807	'Nostoc azollae' 0708	B				Filaments	Filaments		Yes		Aerobic	Multiple	Mesophilic		
12997	Acaryochloris marina MBIC11017	B	8.36	47.0		Sphere, Ellipse	Singles				Aerobic	Aquatic	Mesophilic	No	None
16707	Acaryochloris sp. CCMEE 5410	B				Sphere, Ellipse	Singles				Aerobic	Aquatic	Mesophilic	No	
32769	Acetohalobium arabaticum DSM 5501	B								Moderate halophilic	Anaerobic	Specialized			
19259	Acholeplasma laidlawii PG-8A	B	1.5	31.9		Pleomorphic				Non-halophilic	Facultative	Specialized	Mesophilic		
32215	Acidaminococcus intestinalis RYC-MR95	B			-	Coccobacillus					Anaerobic	Host-associated	Mesophilic	Human	Opportunistic infections
199	Acidianus brierleyi	A	* 1.8			Coccus	Singles	No	No		Facultative		Thermophilic	No	
29525	Acidimicrobium ferrooxidans DSM 10331	B		67-69	+	Rod		No			Anaerobic	Specialized	Thermophilic	No	
15753	Acidiphilium cryptum JF-5	B	3.97	67.1	-	Rod					Aerobic	Multiple	Mesophilic	No	
53	Acidithiobacillus ferrooxidans ATCC 23270	B	3	58.8	-	Rod			Yes		Facultative	Multiple	Mesophilic	No	
16689	Acidithiobacillus ferrooxidans ATCC 53993	B	2.9	58.9	-	Spiral		No	Yes		Aerobic	Specialized	Mesophilic	No	
15771	Acidobacteria bacterium Ellin345	B	5.7	58.4										No	
1	Acidobacterium capsulatum ATCC 51196	B	* 4.15		-	Rod	Singles	No	No	Non-halophilic	Aerobic	Multiple	Mesophilic	No	None
28085	Acidobacterium capsulatum ATCC 51196	B			-	Rod	Singles	No	No		Aerobic	Multiple	Mesophilic	No	
16097	Acidothermus cellulolyticus 11B	B	2.4	66.9	+			Yes			Aerobic	Aquatic	Thermophilic	No	
15708	Acidovorax avenae subsp. citrulli AAC00-1	B	5.4	68.5	-	Rod		No	Yes	Non-halophilic	Aerobic	Multiple	Mesophilic	Fruit	Bacterial fruit blotch
32605	Acidovorax delafieldii 2AN	B		65-66	-	Rod		No	Yes		Aerobic	Aquatic	Mesophilic	No	
15685	Acidovorax sp. JS42	B	4.54	66.1	-	Rod		No	Yes		Aerobic	Terrestrial	Mesophilic	No	
19277	Aciduliprofundum boonei	A		39.1		Coccus			Yes	Mesophilic	Anaerobic	Specialized	Thermophilic	No	

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

27 February, 2010 | From *The Economist* print edition

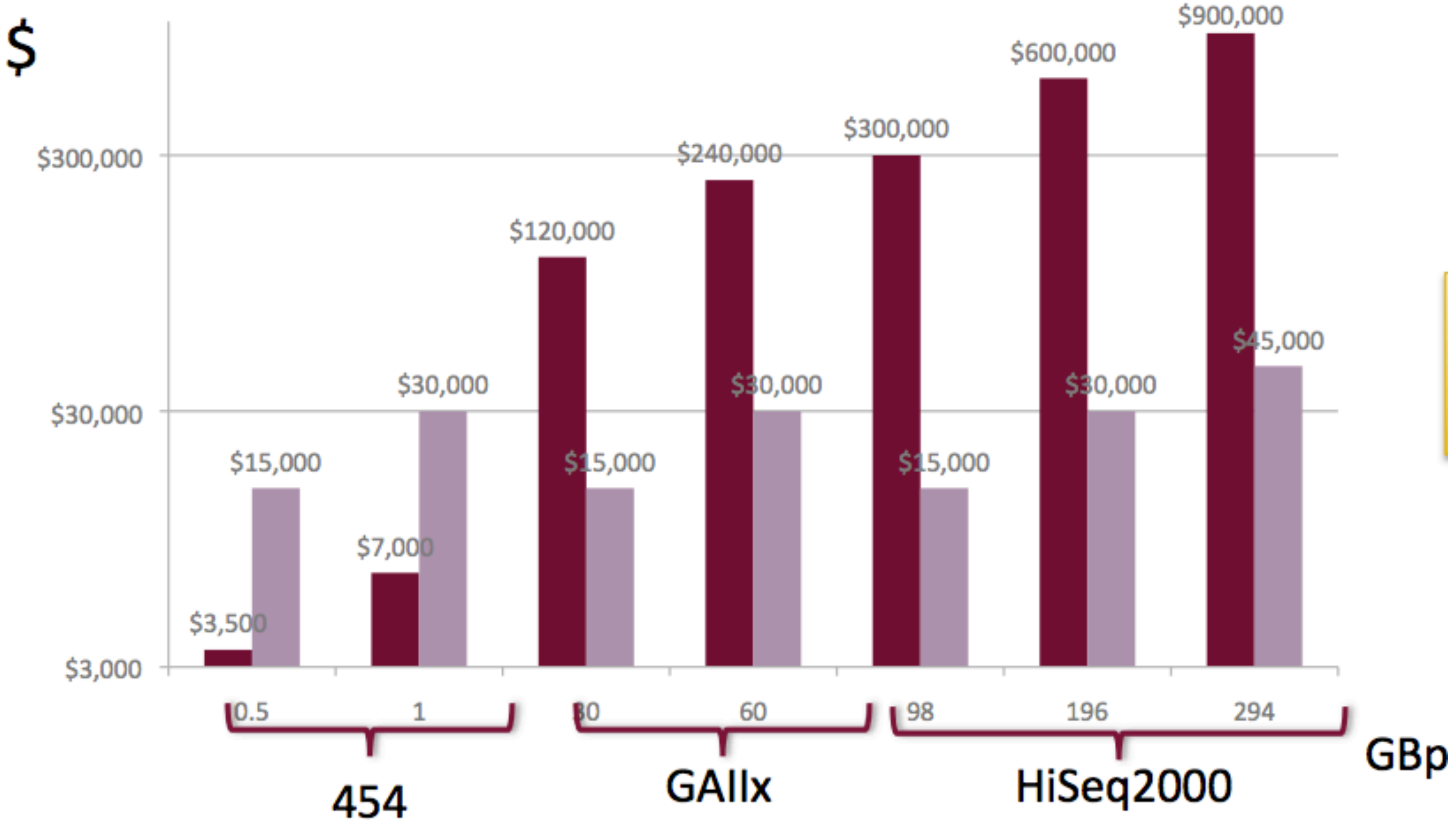
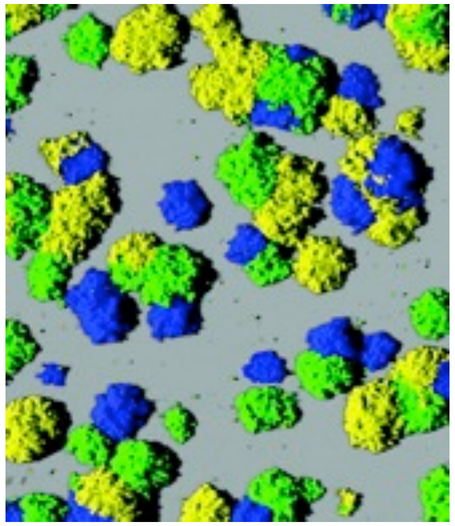


Figure 3: Computing cost dominate sequencing costs. While sequencing costs remain almost identical across platforms, the analysis costs vary with data set sizes. The cost of sequencing compared to the cost of running BLASTX analysis. Data from [11] using the Amazon EC2 cloud machine as a cost model.

Folker Meyer and Nikos Kyrpide, Proposal for open discussion: Informatics challenges for next generation sequencing metagenomics experiments, DOE JGI User Meeting (March 23, 2010).

High-throughput comparisons

Rolf Shinny Marlene Rasmus



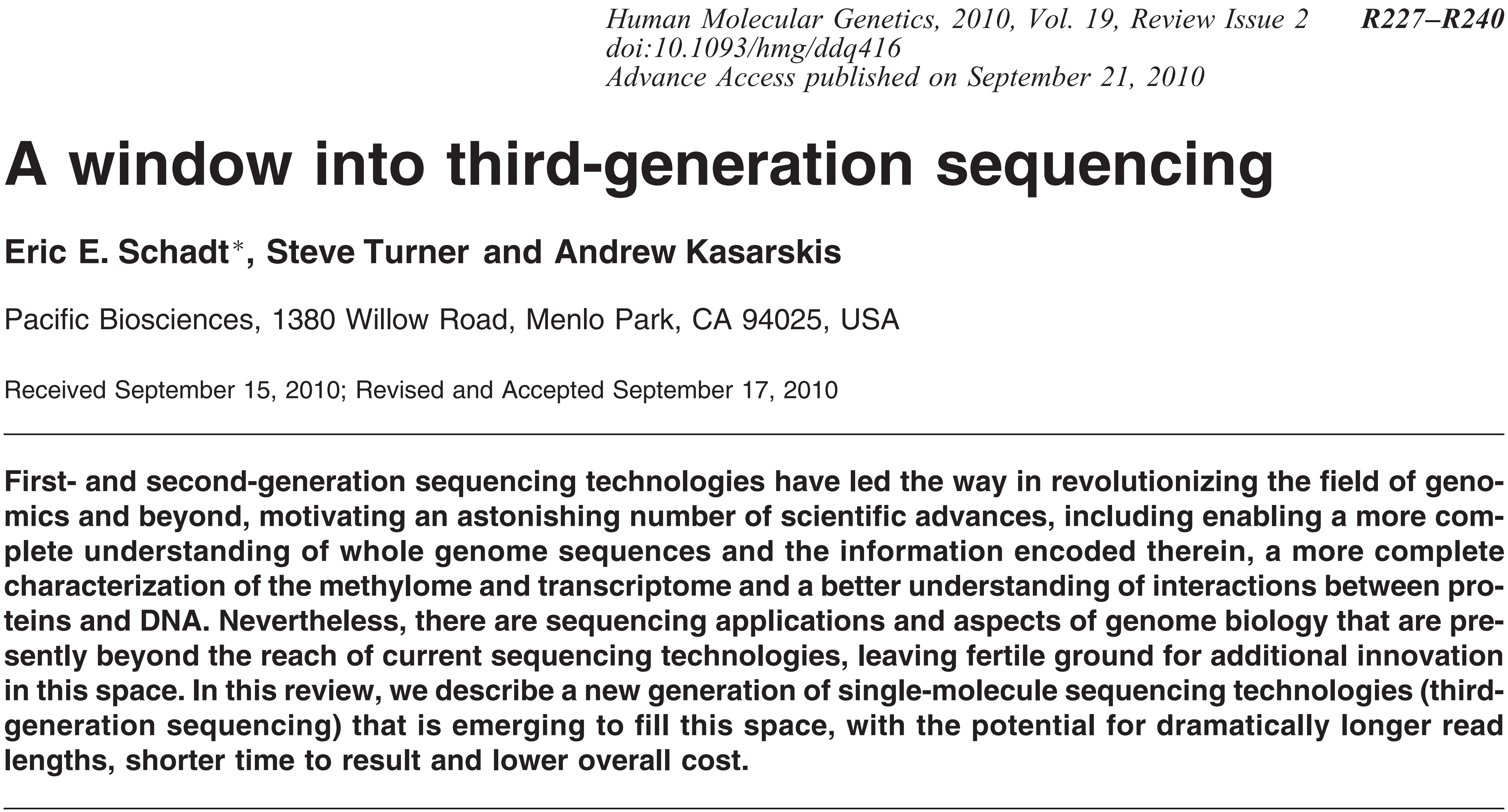
E.coli
Publicly available genomes:
417 genomes

Salmonella
Publicly available genomes:
142 genomes

Campylobacter
Publicly available genomes:
144 genomes

Pseudomonas

72 genomes



Human Molecular Genetics, 2010, Vol. 19, Review Issue 2 ***R227–R240***
doi:10.1093/hmg/ddq416
Advance Access published on September 21, 2010

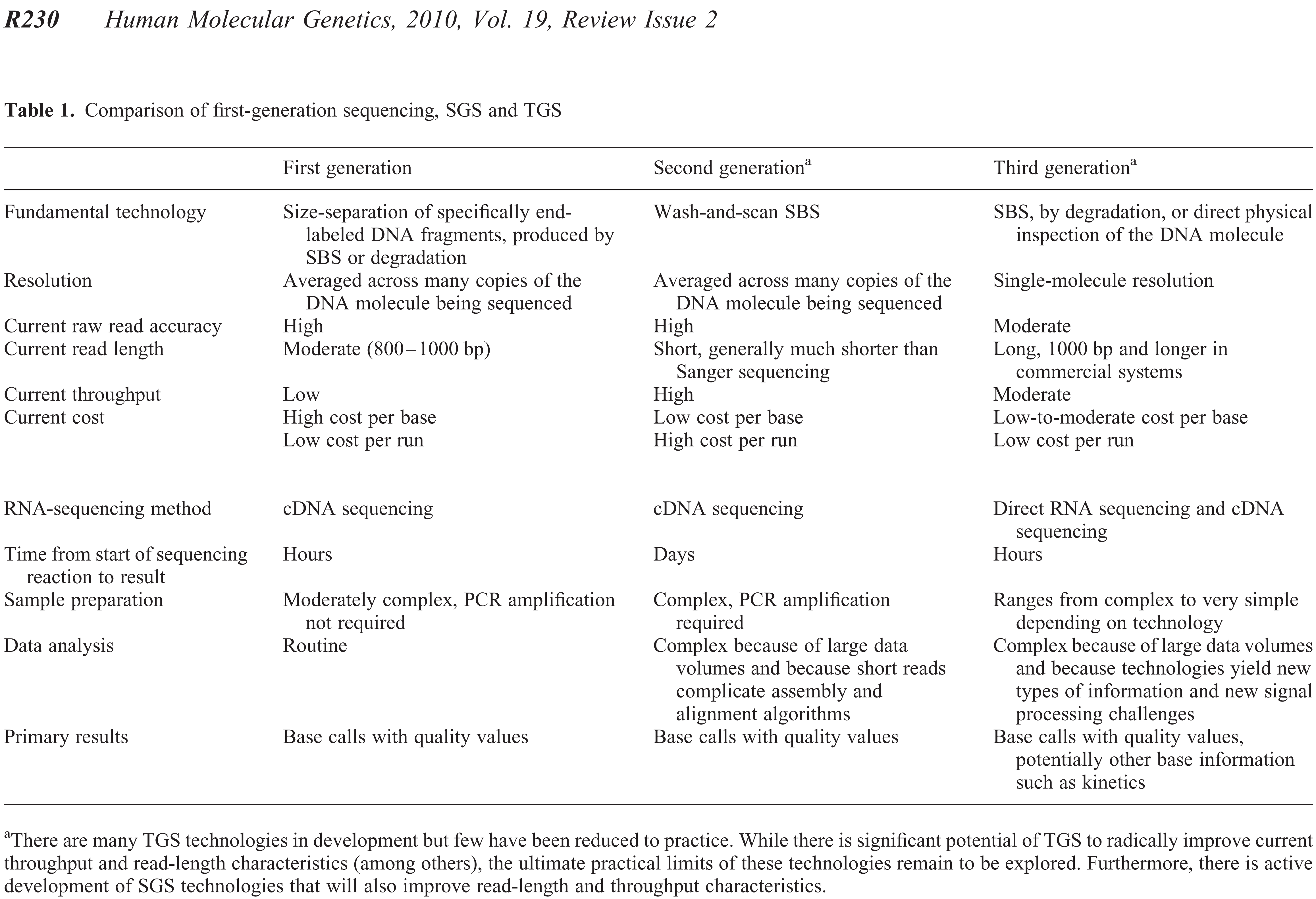
A window into third-generation sequencing

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.



R230 *Human Molecular Genetics, 2010, Vol. 19, Review Issue 2*

Table 1. Comparison of first-generation sequencing, SGS and TGS

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

[a] There 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.

NEWS AND VIEWS

VOLUME 28 NUMBER 5 MAY 2010 NATURE BIOTECHNOLOGY

Third-generation sequencing fireworks at Marco Island

David J Munroe & Timothy J R Harris

NEWS AND VIEWS

Advances in sequencing platforms promise to make this technology more accessible.

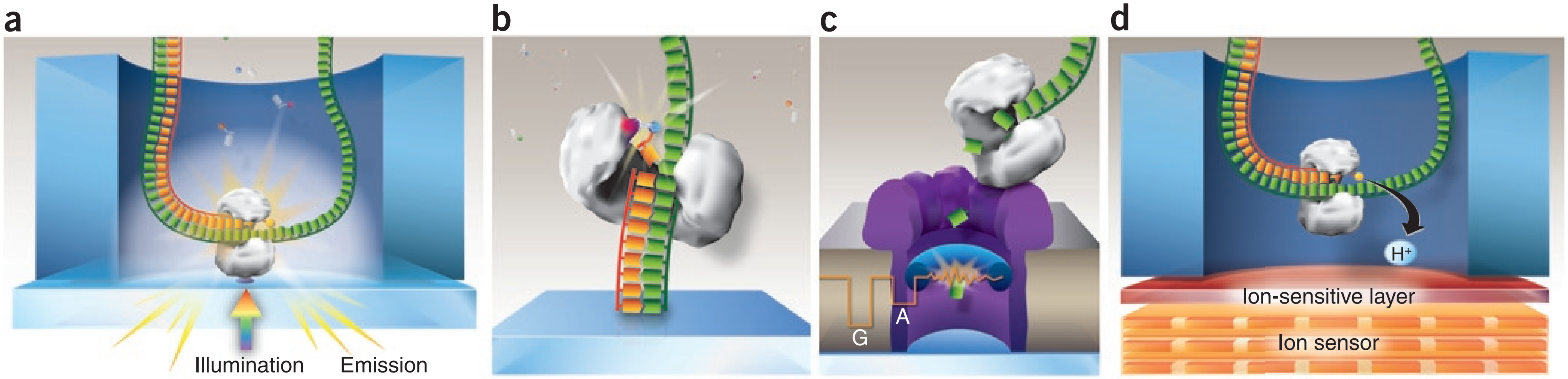


Figure 1 Third-generation sequencing platforms. (**a**) Pacific Biosciences SMRT (single-molecule real-time) DNA sequencing method. The platform uses a DNA polymerase anchored to the bottom surface of a ZMW (pictured in cross section). Differentially labeled nucleotides enter the ZMW via diffusion and occupy the 'detection volume' (white translucent halo area) or microseconds. During an incorporation event, the labeled nucleotide is 'held' within the detection volume by the polymerase for tens of milliseconds. As each nucleotide is incorporated, the label, located on the terminal phosphate, is cleaved off and diffuses out of the ZMW. (**b**) Life Technologies FRET sequencing platform uses base fluorescent labeling technology, a DNA polymerase modified with a quantum dot and DNA template molecules immobilized onto a solid surface. During an incorporation event, energy is transferred from the quantum dot to an acceptor fluorescent moiety on each labeled base. Light emission can only emanate from labeled nucleotides as they are being incorporated. (**c**) The Oxford nanopore sequencing platform uses an exonuclease coupled to a modified α-hemolysin nanopore (purple, pictured in cross section) positioned within a lipid bilayer. As sequentially cleaved bases are directed through the nanopore, they are transiently bound by a cyclodextrin moiety (blue), disturbing current through the nanopore in a manner characteristic for each base. (**d**) The Ion Torrent sequencing platform uses a semiconductor-based high-density array of microwell reaction chambers positioned above an ion-sensitive layer and an ion sensor. Single nucleotides are added sequentially, and incorporation is recorded by measuring hydrogen ions released as a by-product of nucleotide chain elongation.

10 DECEMBER, 2010, 9:08 AM

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.

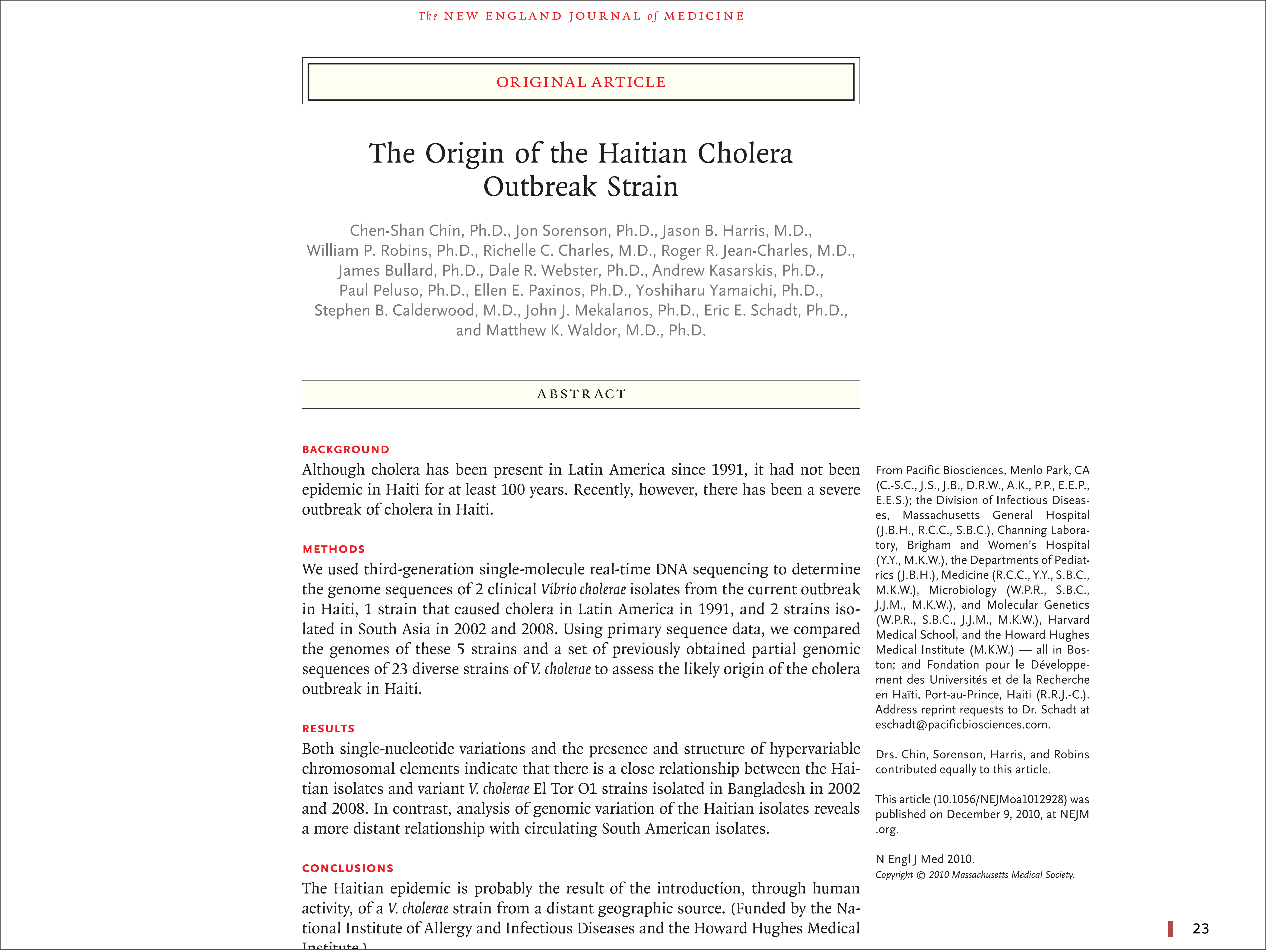
The company received the samples at its headquarters in Menlo Park, Calif., on Nov. 10. By midday the next day, it had some raw sequences done. By Nov. 15, it had sequenced and analyzed not only two strains from Haiti but two from Bangladesh and one from Peru. The draft of the paper was submitted to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



ORIGINAL ARTICLE

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.)

From Pacific Biosciences, Menlo Park, CA (C.-S.C., J.S., J.B., D.R.W., A.K., P.P., E.E.P., E.E.S.); the Division of Infectious Diseases, Massachusetts General Hospital (J.B.H., R.C.C., S.B.C.), Channing Laboratory, Brigham and Women's Hospital (Y.Y., M.K.W.), the Departments of Pediatrics (J.B.H.), Medicine (R.C.C., Y.Y., S.B.C., M.K.W.), Microbiology (W.P.R., S.B.C., J.J.M., M.K.W.), and Molecular Genetics (W.P.R., S.B.C., J.J.M., M.K.W.), Harvard Medical School, and the Howard Hughes Medical Institute (M.K.W.) — all in Boston; and Fondation pour le Développement des Universités et de la Recherche en Haïti, Port-au-Prince, Haiti (R.R.J.-C.). Address reprint requests to Dr. Schadt at eschadt@pacificbiosciences.com.

Drs. Chin, Sorenson, Harris, and Robins contributed equally to this article.

This article (10.1056/NEJMoa1012928) was published on December 9, 2010, at NEJM.org.

N Engl J Med 2010.

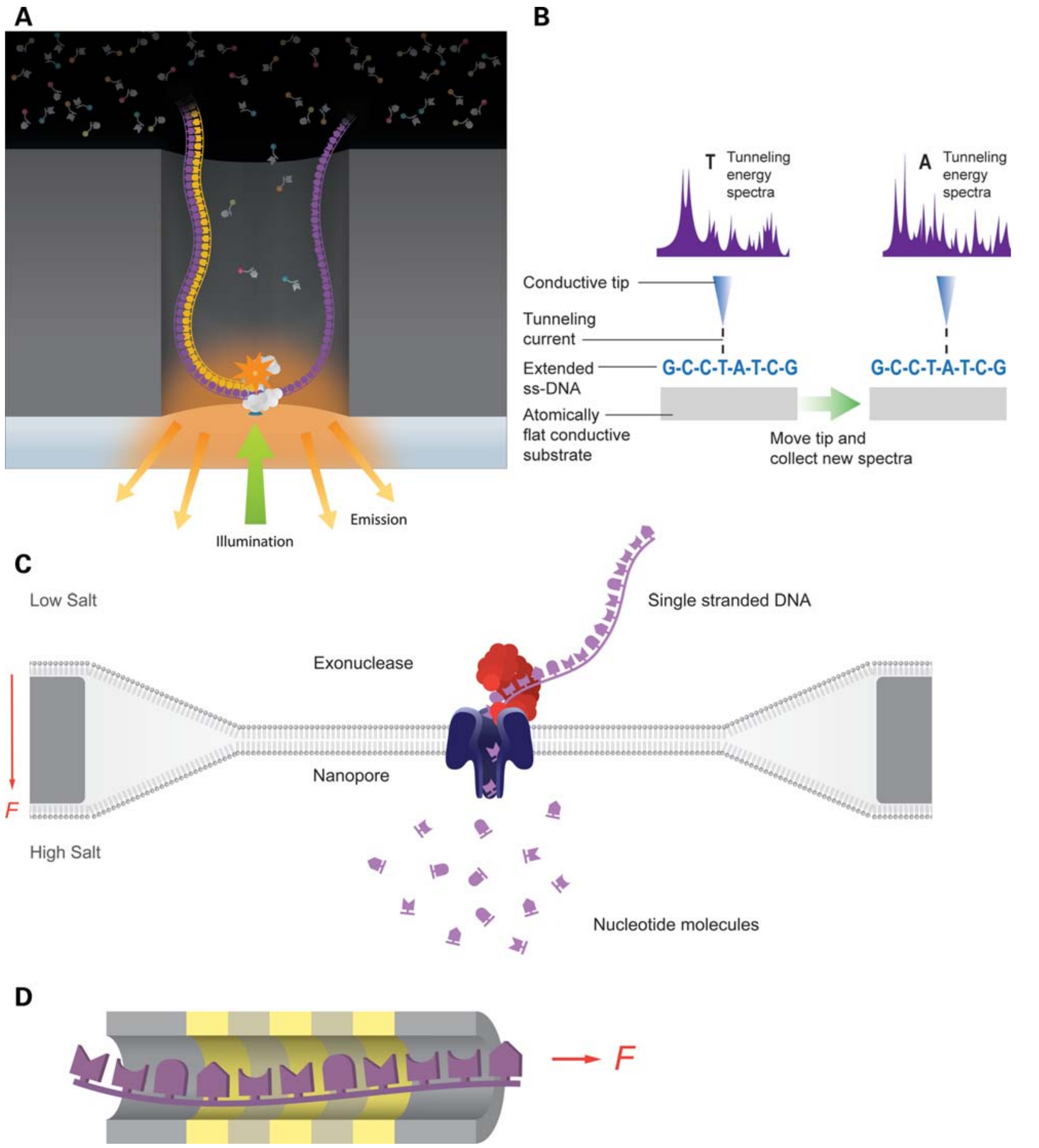
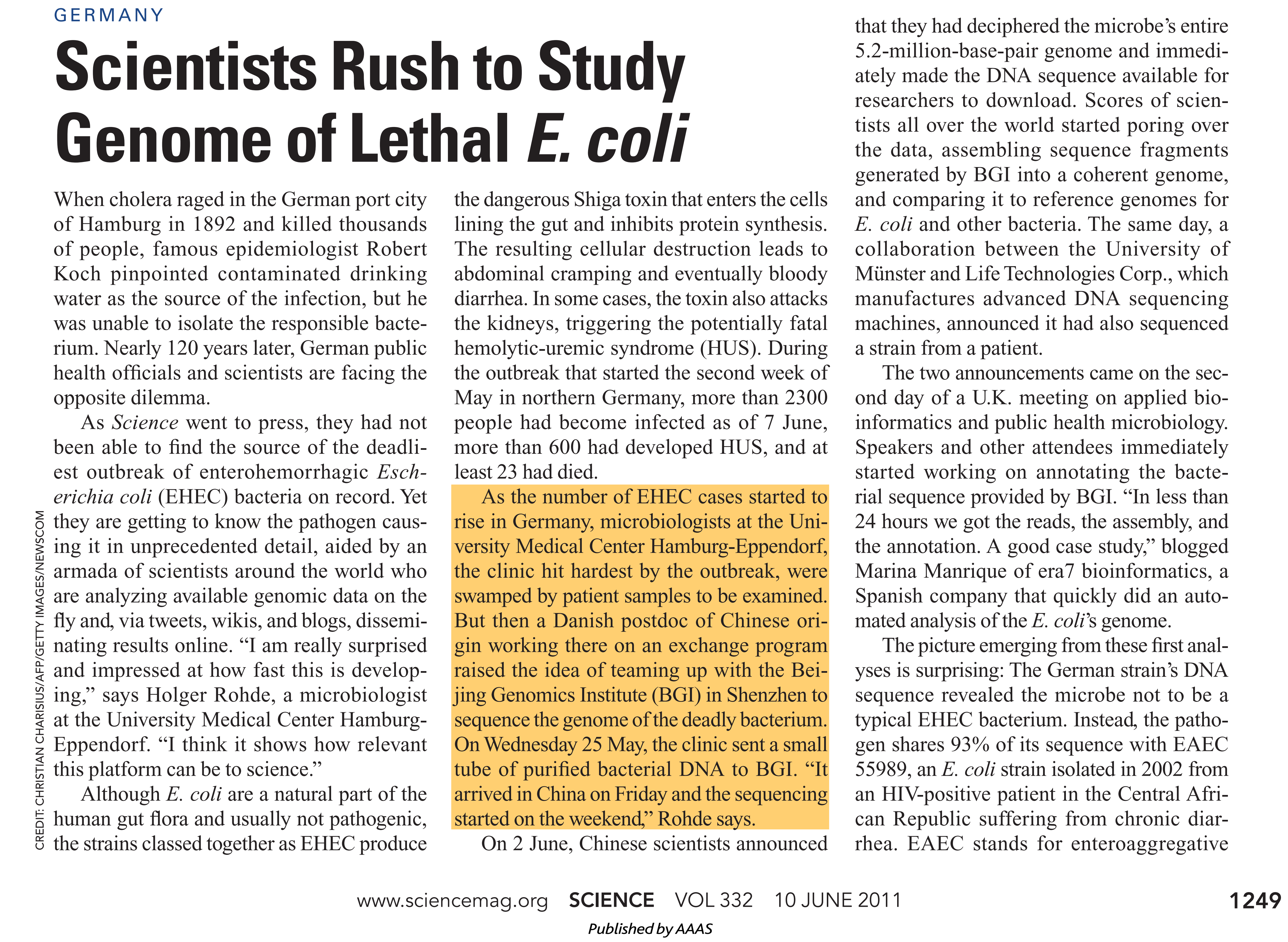


Figure 2. How third-generation DNA-sequencing technologies work. Third-generation DNA-sequencing technologies are distinguished by direct inspection of single molecules with methods that do not require wash steps during DNA synthesis. (**A**) Pacific Biosciences technology for direct observation of DNA synthesis on single DNA molecules in real time. A DNA polymerase is confined in a zero-mode waveguide and base additions 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.



GERMANY

Scientists Rush to Study Genome of Lethal *E. coli*

When cholera raged in the German port city of Hamburg in 1892 and killed thousands of people, famous epidemiologist Robert Koch pinpointed contaminated drinking water as the source of the infection, but he was unable to isolate the responsible bacterium. Nearly 120 years later, German public health officials and scientists are facing the opposite dilemma.

As *Science* went to press, they had not been able to find the source of the deadliest outbreak of enterohemorrhagic *Escherichia coli* (EHEC) bacteria on record. Yet they are getting to know the pathogen causing it in unprecedented detail, aided by an armada of scientists around the world who are analyzing available genomic data on the fly and, via tweets, wikis, and blogs, disseminating results online. "I am really surprised and impressed at how fast this is developing," says Holger Rohde, a microbiologist at the University Medical Center Hamburg-Eppendorf. "I think it shows how relevant this platform can be to science."

Although *E. coli* are a natural part of the human gut flora and usually not pathogenic, the strains classed together as EHEC produce the dangerous Shiga toxin that enters the cells lining the gut and inhibits protein synthesis. The resulting cellular destruction leads to abdominal cramping and eventually bloody diarrhea. In some cases, the toxin also attacks the kidneys, triggering the potentially fatal hemolytic-uremic syndrome (HUS). During the outbreak that started the second week of May in northern Germany, more than 2300 people had become infected as of 7 June, more than 600 had developed HUS, and at least 23 had died.

As the number of EHEC cases started to rise in Germany, microbiologists at the University Medical Center Hamburg-Eppendorf, the clinic hit hardest by the outbreak, were swamped by patient samples to be examined. But then a Danish postdoc of Chinese origin working there on an exchange program raised the idea of teaming up with the Beijing Genomics Institute (BGI) in Shenzhen to sequence the genome of the deadly bacterium. On Wednesday 25 May, the clinic sent a small tube of purified bacterial DNA to BGI. "It arrived in China on Friday and the sequencing started on the weekend," Rohde says.

On 2 June, Chinese scientists announced that they had deciphered the microbe's entire 5.2-million-base-pair genome and immediately made the DNA sequence available for researchers to download. Scores of scientists all over the world started poring over the data, assembling sequence fragments generated by BGI into a coherent genome, and comparing it to reference genomes for *E. coli* and other bacteria. The same day, a collaboration between the University of Münster and Life Technologies Corp., which manufactures advanced DNA sequencing machines, announced it had also sequenced a strain from a patient.

The two announcements came on the second day of a U.K. meeting on applied bioinformatics and public health microbiology. Speakers and other attendees immediately started working on annotating the bacterial sequence provided by BGI. "In less than 24 hours we got the reads, the assembly, and the annotation. A good case study," blogged Marina Manrique of era7 bioinformatics, a Spanish company that quickly did an automated analysis of the *E. coli*'s genome.

The picture emerging from these first analyses is surprising: The German strain's DNA sequence revealed the microbe not to be a typical EHEC bacterium. Instead, the pathogen shares 93% of its sequence with EAEC 55989, an *E. coli* strain isolated in 2002 from an HIV-positive patient in the Central African Republic suffering from chronic diarrhea. EAEC stands for enteroaggregative

CREDIT: CHRISTIAN CHARISIUS/AFP/GETTY IMAGES/NEWSCOM

Questions:

- Will 1000x coverage, with 25 nt reads, be enough to assemble an *E. coli* genome completely to one piece?
- What is ‘third generation’ sequencing technology?
- What is the lower range of quality that is still ‘good enough’ to use in comparing genomes?