

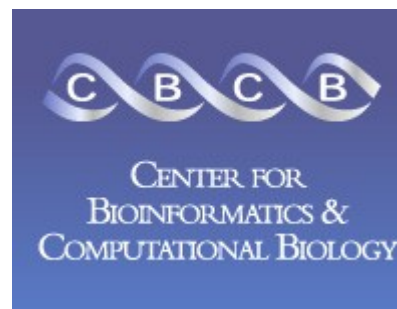
Keeping up with DNA technologies

Mihai Pop

Department of Computer Science

Center for Bioinformatics and Computational Biology

University of Maryland, College Park



The evolution of DNA sequencing

Since	Technology	Read length	Throughput/run	Throughput/hour	cost/run
1977-	Sanger sequencing	> 1000bp	4hr 400-500 kbp	100 kbp	\$200
2005-	454 pyrosequencing	250-400bp	4hr 100-500 Mbp	25-100 Mbp	\$13,000
2006-	Illumina/Solexa	50-100bp	3 days 2-3 Gbp	25-40 Mbp	\$3,000
2007-	ABI SOLiD	35-50bp	3 days 6-20 Gbp	75-250 Mbp	est. \$3-5,000
2008-	Helicos single molecule	25-50 bp	8 days 10 Gbp	~50 Mbp est. 1Gbp/hour	~\$18,000
TBA (2010)	Pacific Biosciences single molecule	100-200 kbp	?	?	?

From DNA to images

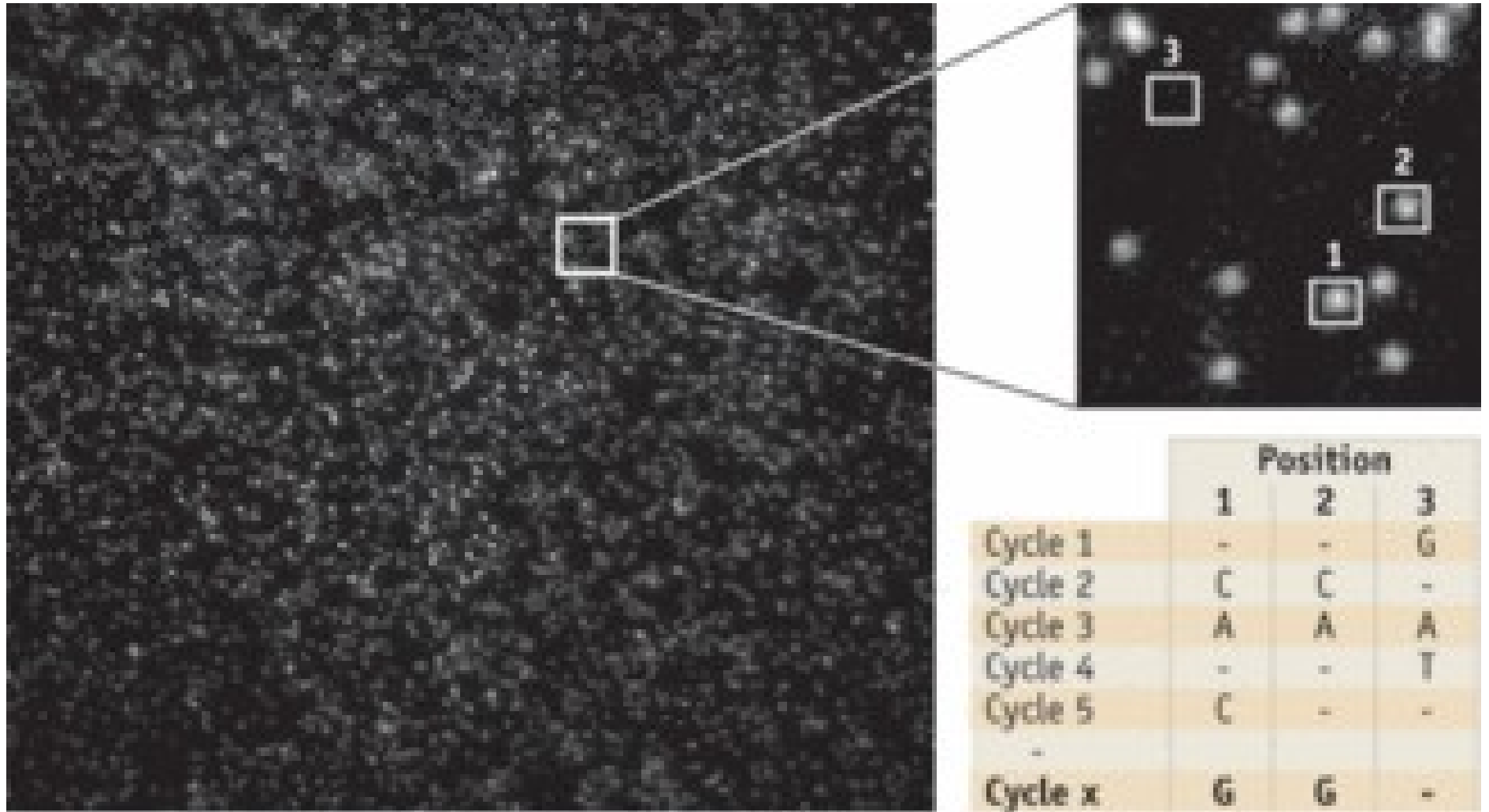
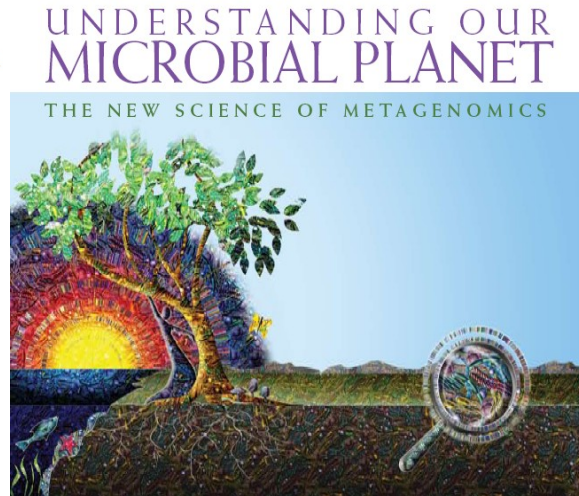


Image from Helicos through <http://www.laserfocusworld.com>

What?



Image from Tree of Life Web Project
www.tolweb.org



<http://dels.nas.edu/metagenomics/>



X 1000

www.1000genomes.org

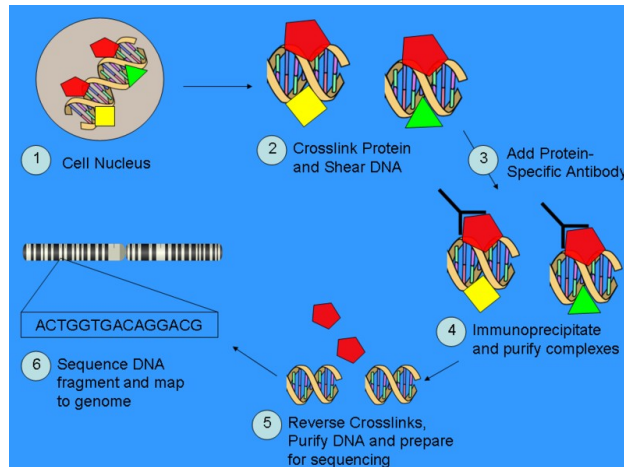


Image from wikipedia

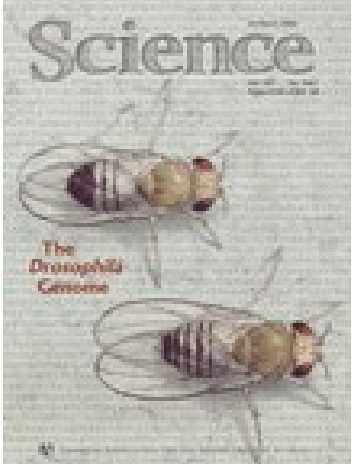


Sage/RNAseq

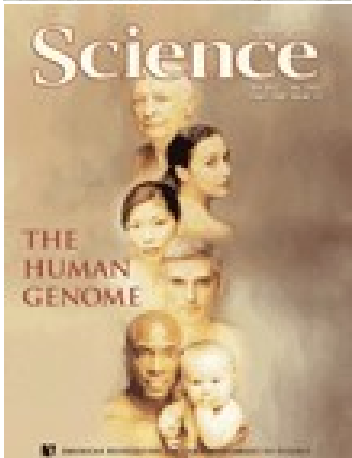
Not too long ago



1995 *Haemophilus influenzae*



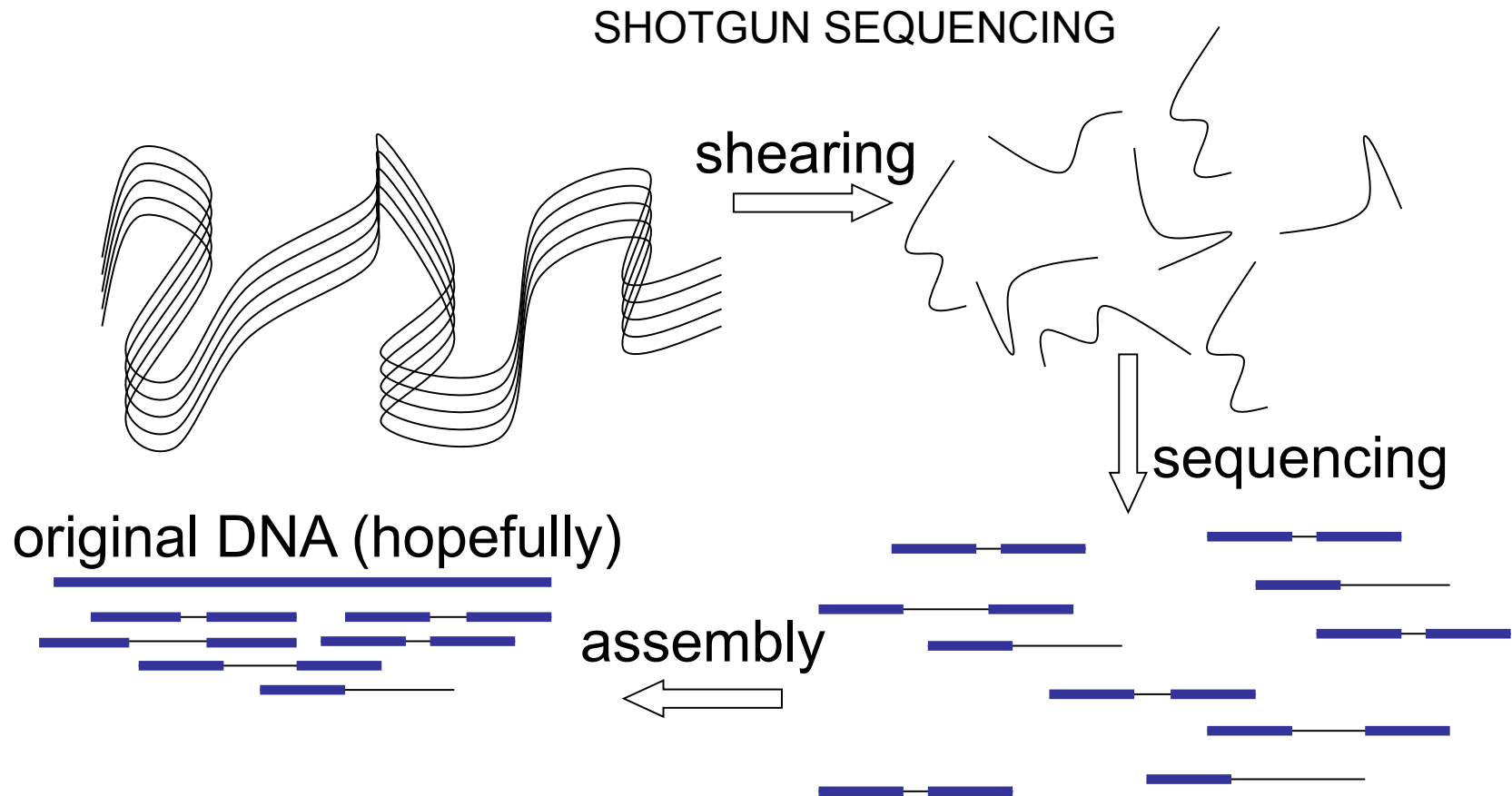
2000 *Drosophila melanogaster*



2001 *Homo sapiens*

Assembly

- Given a collection of short DNA pieces from a genome
- Reconstruct the original genome sequence
- Related to Shortest Common Superstring – given a set of strings S , find the shortest string that contains all the strings in S as a substring. NP-hard



Challenges

Note: Research on alignment/assembly algorithms is quite “old” and the basic concepts are unchanged by technology

- Large amounts of data – even data storage/transfer is a problem
- Special error characteristics
 - most – very short reads (25-35 bp instead of 1000s)
 - 454 – homopolymer stutter

```
ACAAGAAATGT  
ACAAGA--TGT  
ACAAGAA-TGT
```

- Helicos – high error rates, same DNA fragment sequenced multiple times
 - Pacific Biosciences – very long reads, possibly high error rate
- Special data
 - ABI/SOLiD – color space sequencing

Short-read assembly at the CBCB

- *De novo* assembly

Minimus-SR (w/ Dan Sommer)

- all-pair alignments performed fast (~2 minutes for 8.6 million reads)
- key – use bit-wise operations (e.g. XOR to test equality)
- very good results (better than most short read assemblers)

- Comparative assembly

AMOScmp (w/ Adam Phillippy, Daniela Puiu, S. Salzberg)

- originally developed for Sanger data – works for short reads too

ABBA (Dan Sommer, Steven Salzberg)

- comparative assembly with a protein reference (resequencing of genes)

- Short read alignment

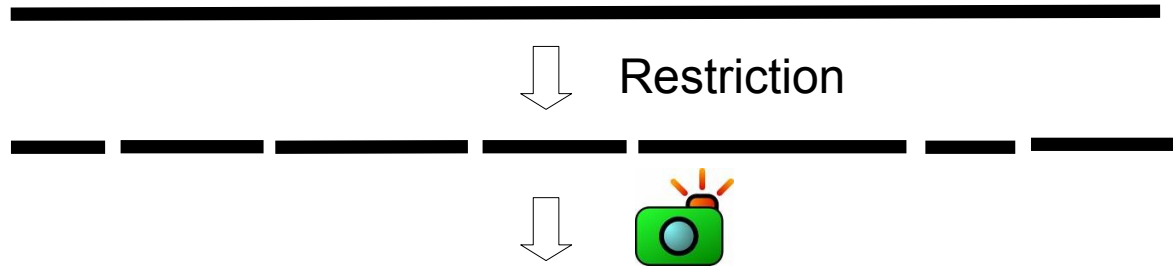
Bowtie (w/ Ben Langmead and Steven Salzberg)

- uses Burrows-Wheeler index of a genome (only 1.1GB for human)
- up to 30-40-fold faster than other aligners

Optical mapping

- The better restriction mapping

output: ordered list of sizes



Fragment	Size(kb)	Standard Deviation (kb)
0	13.802	0.146
1	16.607	0.150
2	11.719	0.119

- Map-map alignment and contig-map alignment easier – relatively simple dynamic programming

SOMA – scaffolding with optical maps

- Dynamic programming $O(m^2n^2)$

$$S[i, j] = \max_{0 \leq k \leq i, 0 \leq l \leq j} -C_r \times (i - k + l - j) - \frac{(\sum_{s=k}^i c_s - \sum_{t=l}^j o_t)^2}{\sum_{t=l}^j \sigma_t^2} + S[k-1, l-1]$$

penalty for mismatched sites

chi-squared score of
alignment btwn fragments

- Bootstrapped computation of p-values (how likely is a match)
- “Scheduling” algorithm for near-optimal placement of contigs with multiple good matches.

$$\frac{\chi_{\text{best}}^2 / n_{\text{best}}}{\chi_{\text{other}}^2 / n_{\text{other}}} \quad \text{F-test to assess “equally good” matches}$$

- Approx. 80-90% of a genome can be placed in one scaffold

Opportunities for collaborations

- We are driven by data and technology
 - Let us know if you have interesting data-sets
 - ... interesting DNA technologies
 - ... useful computational technologies (in particular parallel/grid infrastructure and high performance storage)
- Examples of collaborations
 - with UM School of Medicine – study of diarrhea in 3rd world countries
 - with USDA – sequencing of cow and turkey
 - with Google/IBM/NSF – applications of Cloud Computing to bioinformatics
 - with Opgen Inc. - new ways to analyze optical mapping data
 - ... and many more

Acknowledgments

CBCB

- Steven Salzberg
- Art Delcher
- Carl Kingsford
- Daniela Puiu

Pop group

- Niranjan Nagarajan
- Mohammad Ghodsi
- Sergey Koren
- Ben Langmead
- Bo Liu
- Dan Sommer
- James White

Optical mapping

- Tim Read
- Zheng Wang
- Dave Schwartz
- Opgen, Inc.

Funding

- NSF
- NIH
- Bill & Melinda Gates Foundation
- Henry Jackson Foundation