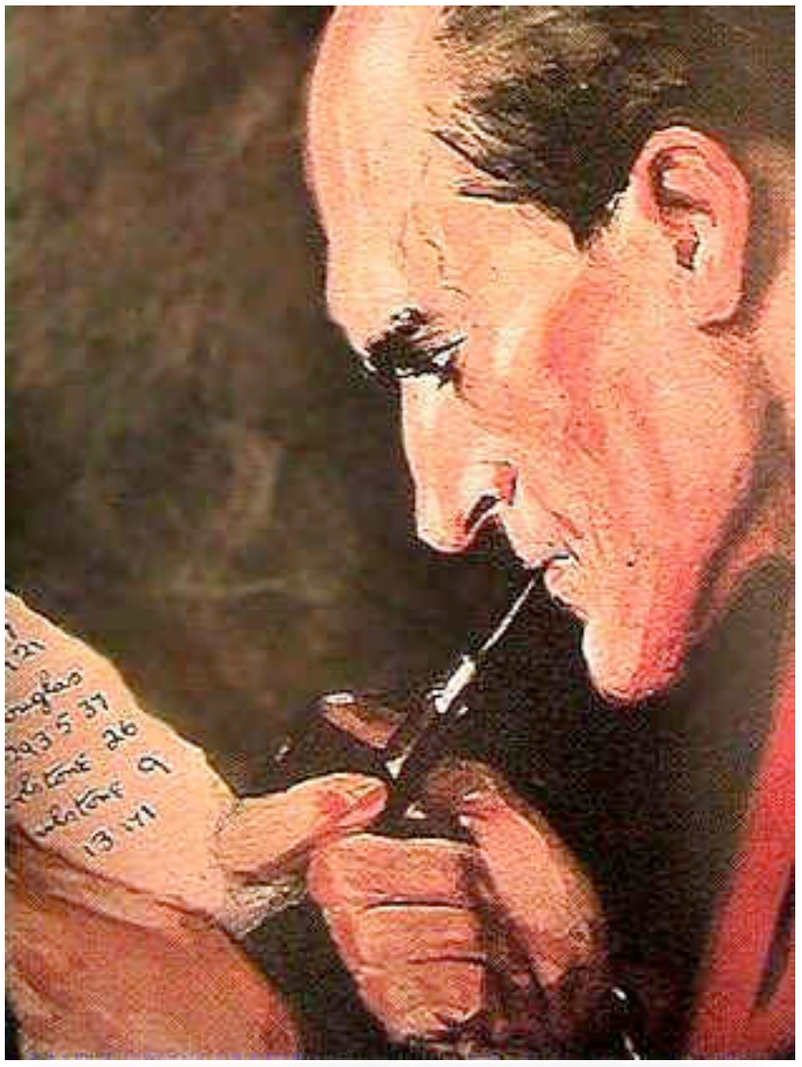
Genomics

Malay K Basu (malay@uab.edu)

"All science is either physics or stamp collecting..."

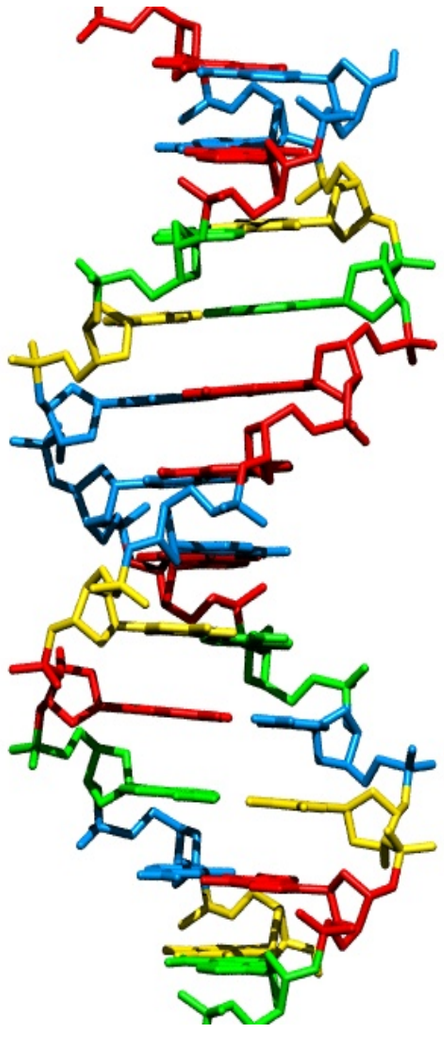
- Ernest Rutherford
"As quoted in Rutherford at Manchester"



http://www.perkydesigns.com

"Data! Data! Data! ...
I can't make bricks
without the clay"

- Sherlock Holmes
"Adventures of Copper Beeches"



...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT...

Evolution of sequencing

Archaic sequencing methods

Early 70s: chromatography

First nucleotide sequencing

Article

Nature **237**, 82-88 (12 May 1972) | doi:10.1038/237082a0

Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein

W. MIN JOU, G. HAEGEMAN, M. YSEBAERT & W. FIERS

1. Laboratory of Molecular Biology and Laboratory of Physiological Chemistry, State University of Ghent, Belgium

▲ Top

By characterization of fragments, isolated from a nuclease digest of MS2 RNA, the entire nucleotide sequence of the coat gene was established. A "flower"-like model is proposed for the secondary structure. The genetic code makes use of 49 different codons to specify the sequence of the 129 amino-acids long coat polypeptide.

First DNA sequencing

Proc. Nat. Acad. Sci. USA
Vol. 70, No. 12, Part I, pp. 3581–3584, December 1973

The Nucleotide Sequence of the *lac* Operator

(regulation/protein–nucleic acid interaction/DNA–RNA sequencing/oligonucleotide priming)

WALTER GILBERT AND ALLAN MAXAM

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Communicated by J. D. Watson, August 9, 1973

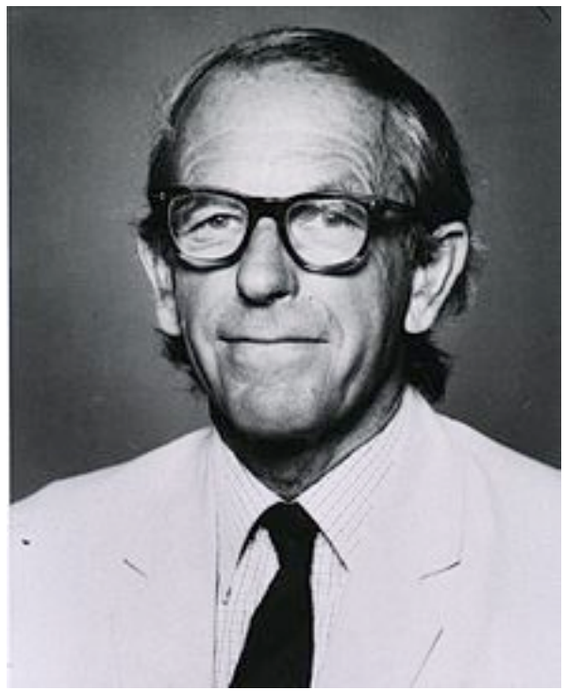
First Genome Sequence

(Reprinted from Nature, Vol. 260, No. 5551, pp. 500–507, April 8, 1976)

Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene

W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert & M. Ysebaert

Laboratory of Molecular Biology, University of Ghent, 9000 Ghent, Belgium



Sanger dideoxy sequencging

First DNA genome sequenced in 1977: φX174.

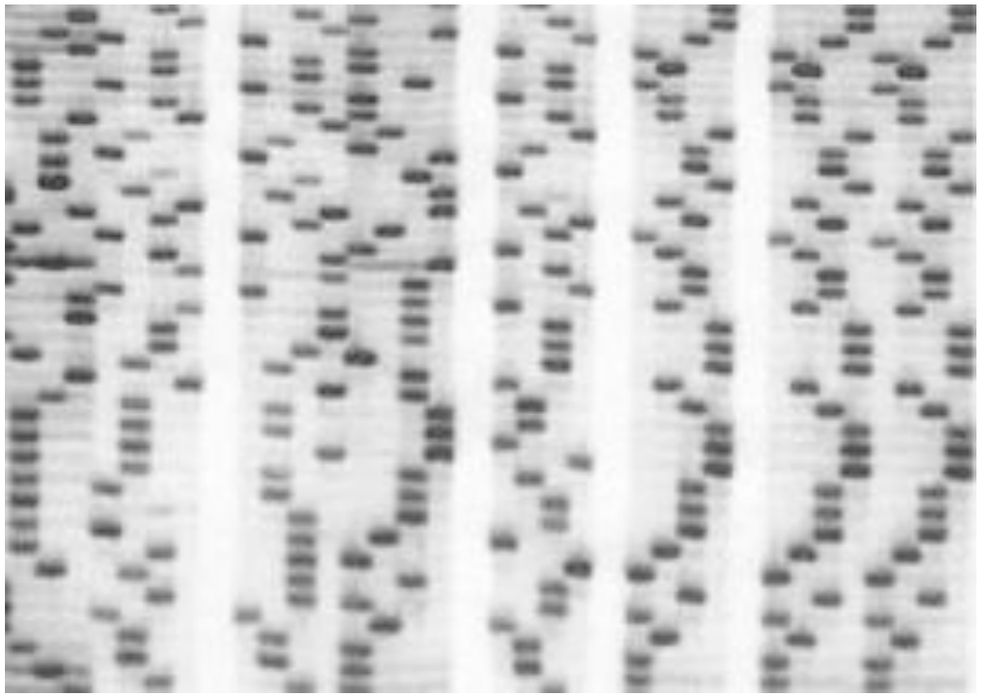
Nature Vol. 265 February 24 1977

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III‡, P. M. Slocombe§ & M. Smith‘

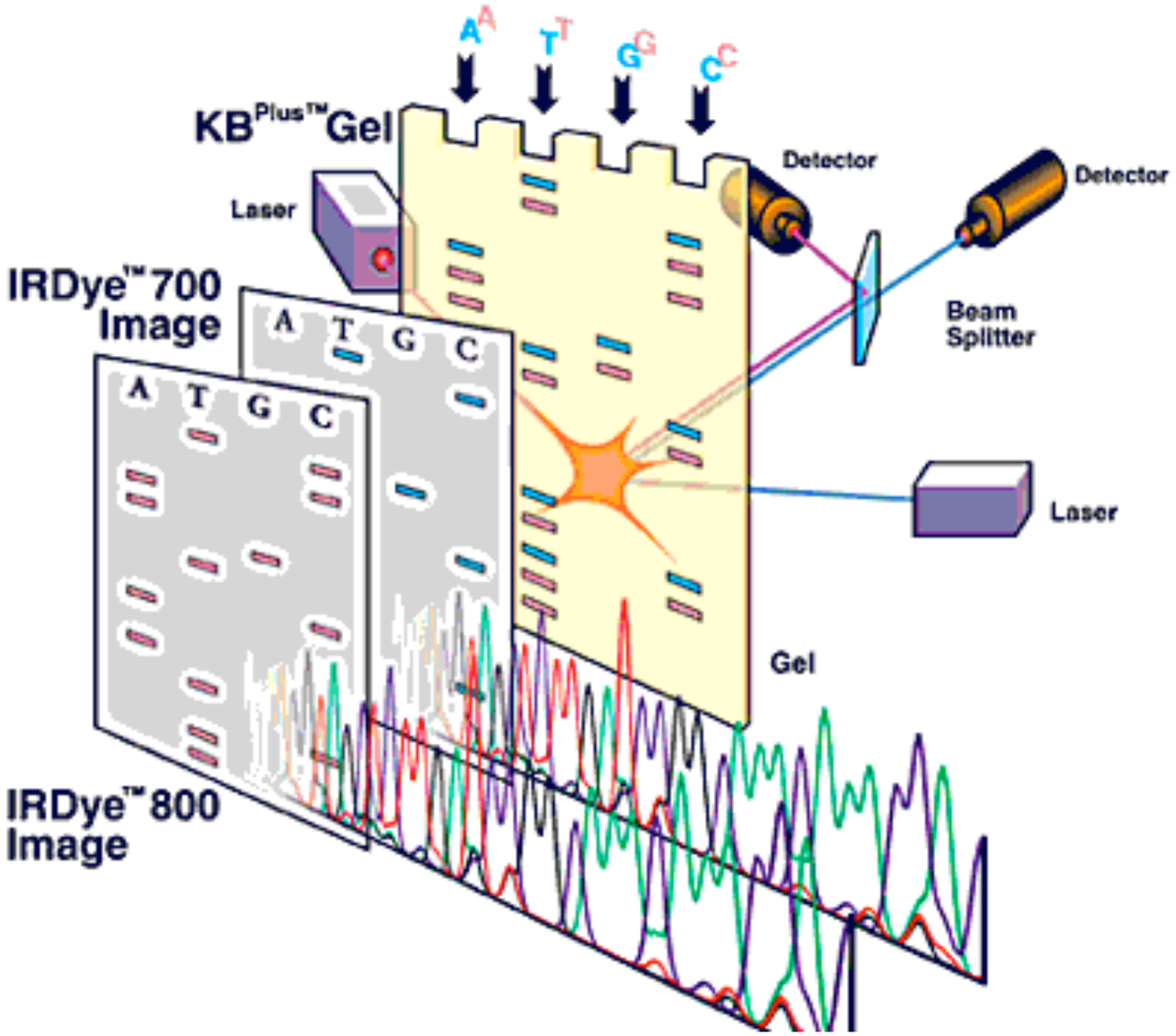
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK



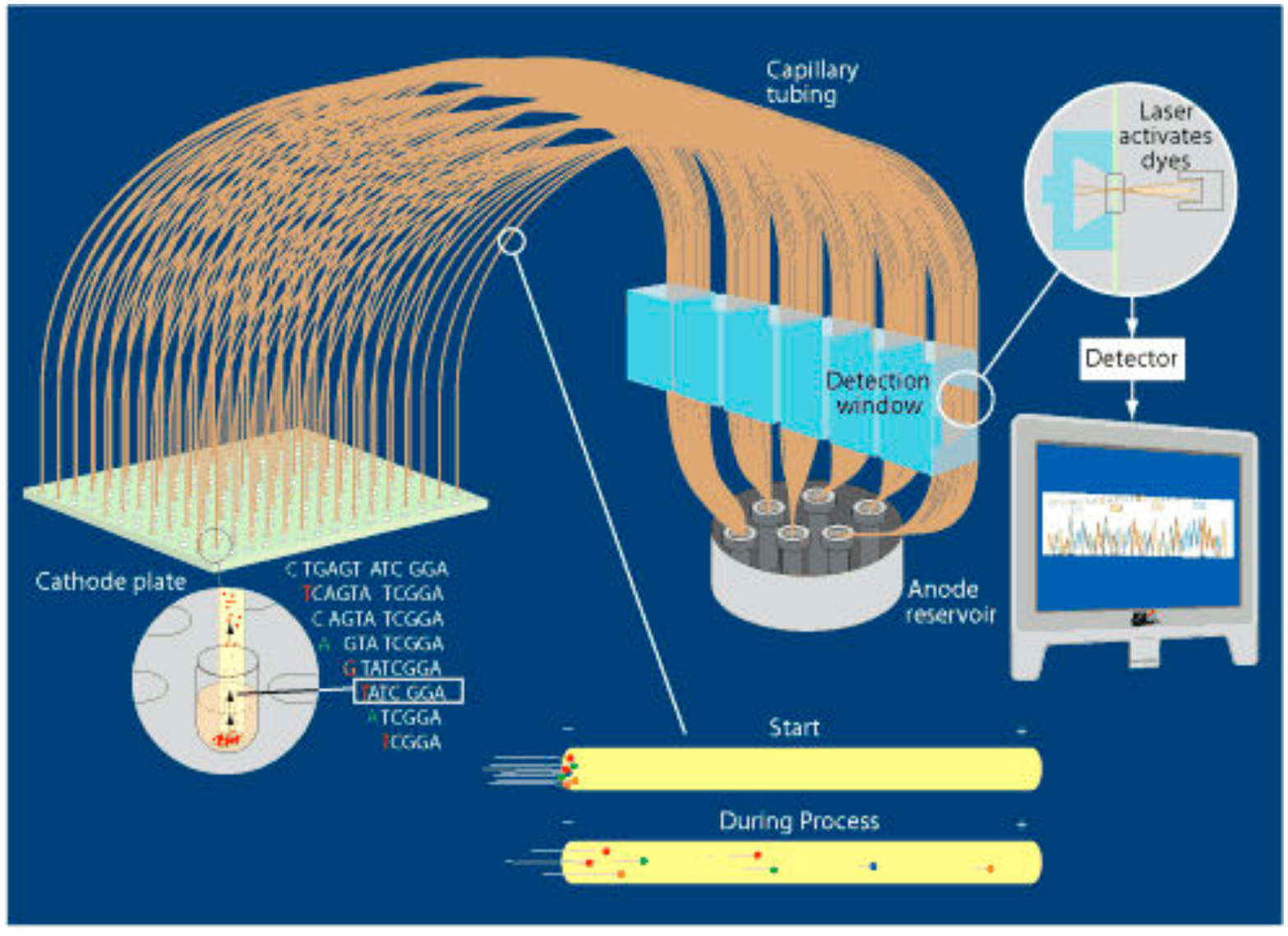
1990s: Large scale automated Sequencing

Generation 1: Gel based or capillary

First automated sequencing



Capillary Sequencing



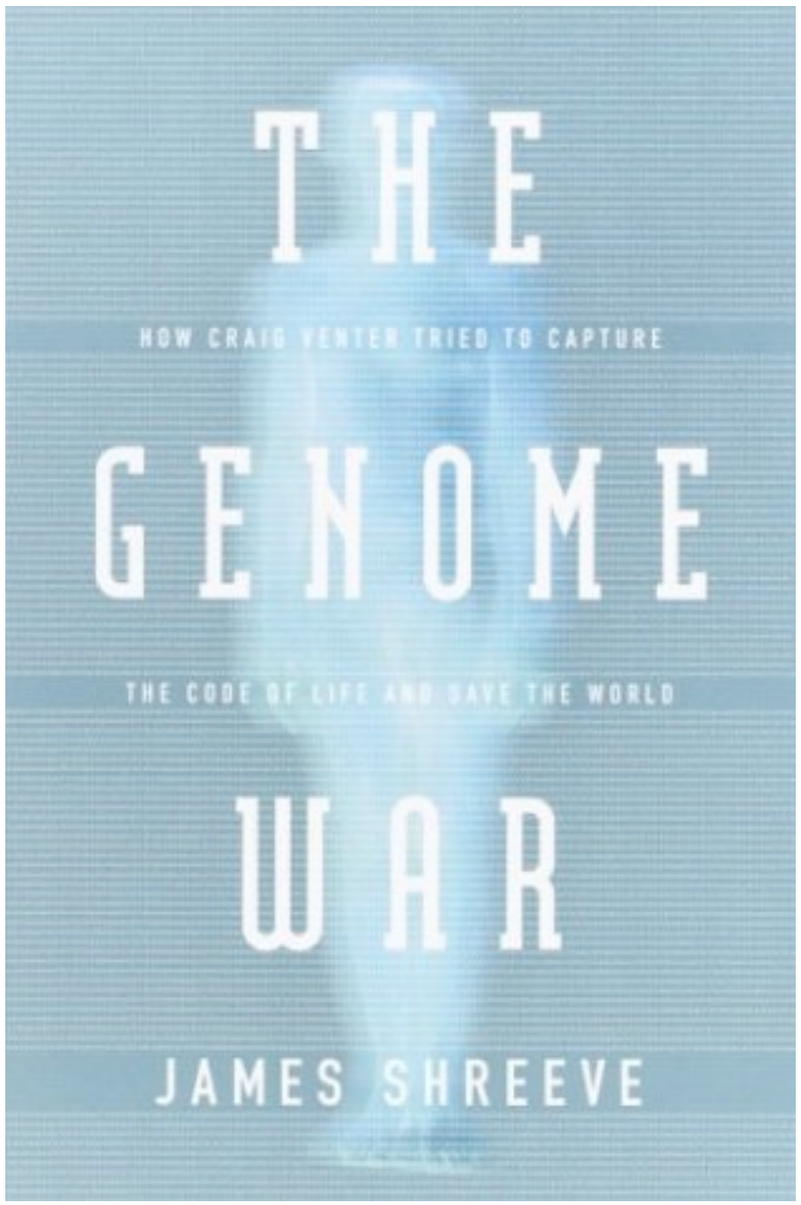
1995: *Haemophilus influenza*

RESEARCH ARTICLE

Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

An approach for genome analysis based on sequencing and assembly of unselecte pieces of DNA from the whole chromosome has been applied to obtain the complet nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium *Hae mophilus influenzae* Rd. This approach eliminates the need for initial mapping efforts an is therefore applicable to the vast array of microbial species for which genome maps ar unavailable. The *H. influenza*e Rd genome sequence (Genome Sequence DataBase ac cession number L42023) represents the only complete genome sequence from a free living organism.


THE
HOW CRAIG VENTER TRIED TO CAPTURE
GENOME
THE CODE OF LIFE AND SAVE THE WORLD
WAR
JAMES SHREEVE

2001 Human Genome

Human Genome

Not a single individual

Was a hack job

Refined over the next 5 yrs

Human Genome Assembly Information

Metrics for the current genome assembly.

Statistics for the current assembly are available below. Information on tiling path files (TPFs) for the human assembly is available at TPF Overview.

Assembly Statistics for GRCh37.p12

Choose another assembly GRCh37.p1

Chromosome Lengths | Total Lengths | Ungapped Lengths | N50s | Gaps | Counts

Spanned gaps are found within scaffolds and there is some evidence suggesting linkage between the two sequences flanking the gap. Unspanned gaps are found between scaffolds and there is no evidence of linkage.

Primary Assembly | Information By Region

	Spanned Gaps			Unspanned Gaps		
chr	All Scaffolds	Placed Scaffolds	Unplaced Scaffolds	All Scaffolds	Placed Scaffolds	Unplaced Scaffolds
1	19	19	0	22	22	0
2	3	3	0	15	15	0
3	0	0	0	7	7	0
4	1	1	0	12	12	0
5	1	1	0	6	6	0
6	6	6	0	8	8	0
7	9	9	0	8	8	0
8	1	1	0	9	9	0
9	15	15	0	29	9	0
10	8	8	0	12	8	0
11	4	4	0	11	11	0
12	1	1	0	8	8	0
13	0	0	0	0	1	0
14	0	0	0	5	5	0
15	2	2	0	10	10	0
16	1	1	0	10	10	0
17	2	2	0	5	5	0
18	2	2	0	7	7	0
19	1	1	0	8	8	0
20	2	2	0	9	9	0

Global stats for GRCh37.p12

General Info	
Assembly Type	haploid with alt loci
Release Type	patch
Number of Assembly Units	12
Total Bases in Assembly	3,230,373,980
Total Non-N Bases in Assembly	2,987,105,853
Primary Assembly N50	46,395,641
Region Information	
Total number of defined regions	172
Number of Regions with Alternate Loci	3
Number of Regions with Fix Patches	110
Number of Regions with Novel Patches	60
Number of Regions as PAR	4
Alternate Loci/PATCH Information	
Total Number of Alternate Loci scaffolds	9
Number of Alternate Loci scaffolds aligned to the Primary Assembly	9
Number of FIX Patch scaffolds	121
Number of FIX Patch scaffolds aligned to the Primary Assembly	121
Number of NOVEL Patch scaffolds	73
Number of NOVEL Patch scaffolds aligned to the Primary Assembly	73

Reference assembly

Next generation sequencing

Massively Parallel Signature Sequencing (MPSS)

Early 1990s: created by Lynx technologies, purchased by Solexa/Illumina

Illumina Video

https://www.youtube.com/watch?v=womKfikWlxM

Early next-gen sequencers

Table 1. Comparison of different sequencing technologies, taken from [34].

Sequencer	ABI 3730	Roche 454	Solexa[a]	SOLiD (mp, frag)[b]	HeliScope[c]
Read length	600–900	400–500	75–100	50	25–35
Run time	6–10 h	10 h	2–10 d	(4–7 d,8–14 d)	h
Yield (Mbp)	0.01	1	2,300–3,500/d	(500, 1,000)	105–140/h
Cloning bias	Yes	No	No	No	No
Mate pair information	Yes	No	Yes	Yes	No

[a]Based on the GA IIx. See full specifications at: http://www.illumina.com/systems/genome_analyzer.ilmn.
[b]mp, mate pair; frag, fragment. See https://products.appliedbiosystems.com/ SOLiD 3 Plus System.
[c]See: http://www.helicosbio.com/Products/HelicosregGeneticAnalysisSystem/HeliScopetradeSequencer/tabid/87/Default.aspx.
doi:10.1371/journal.pcbi.1000667.t001

Next-gen sequencers

Current fashion:

Illumina

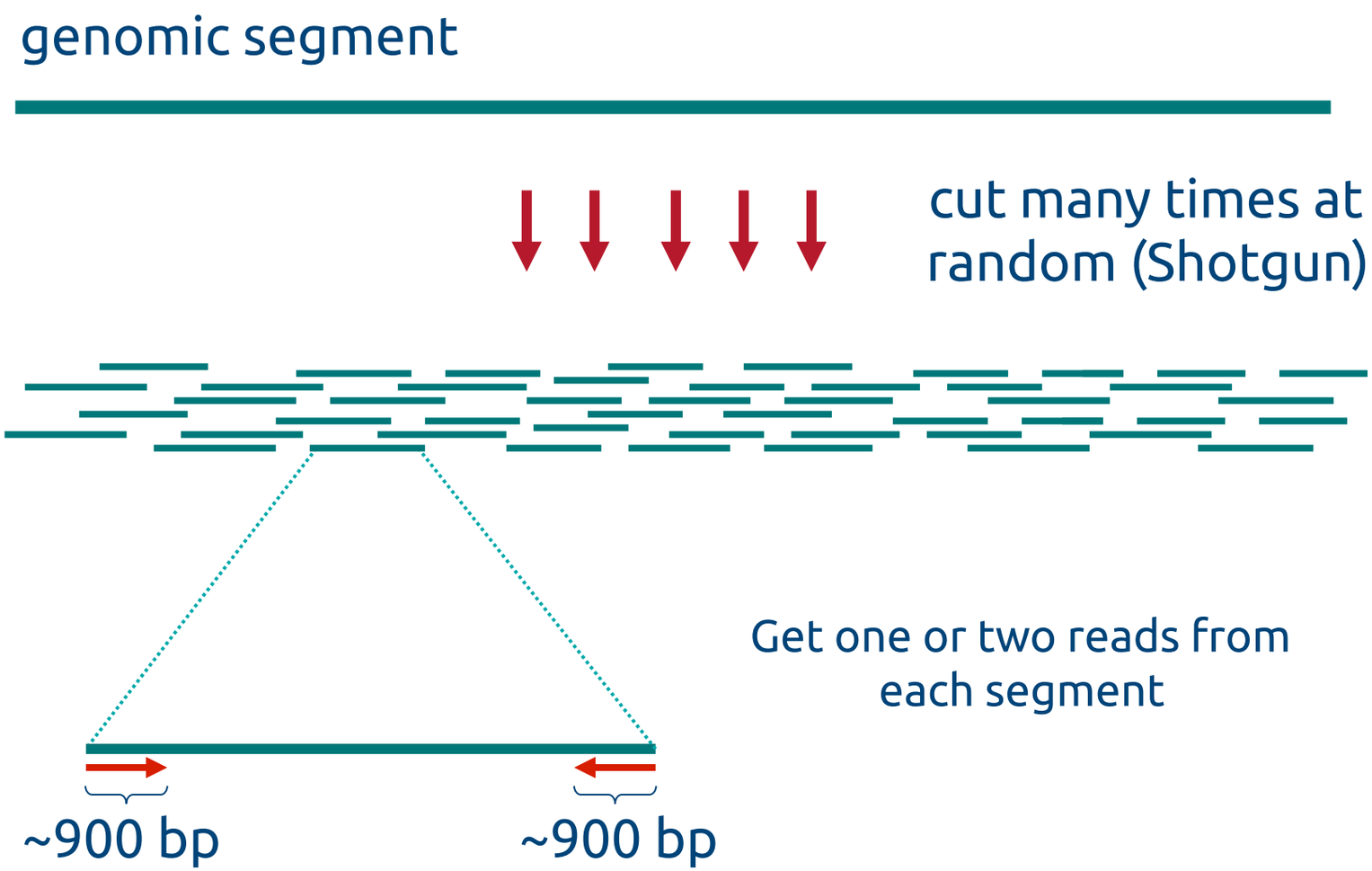
IonTorrent

Around the corner

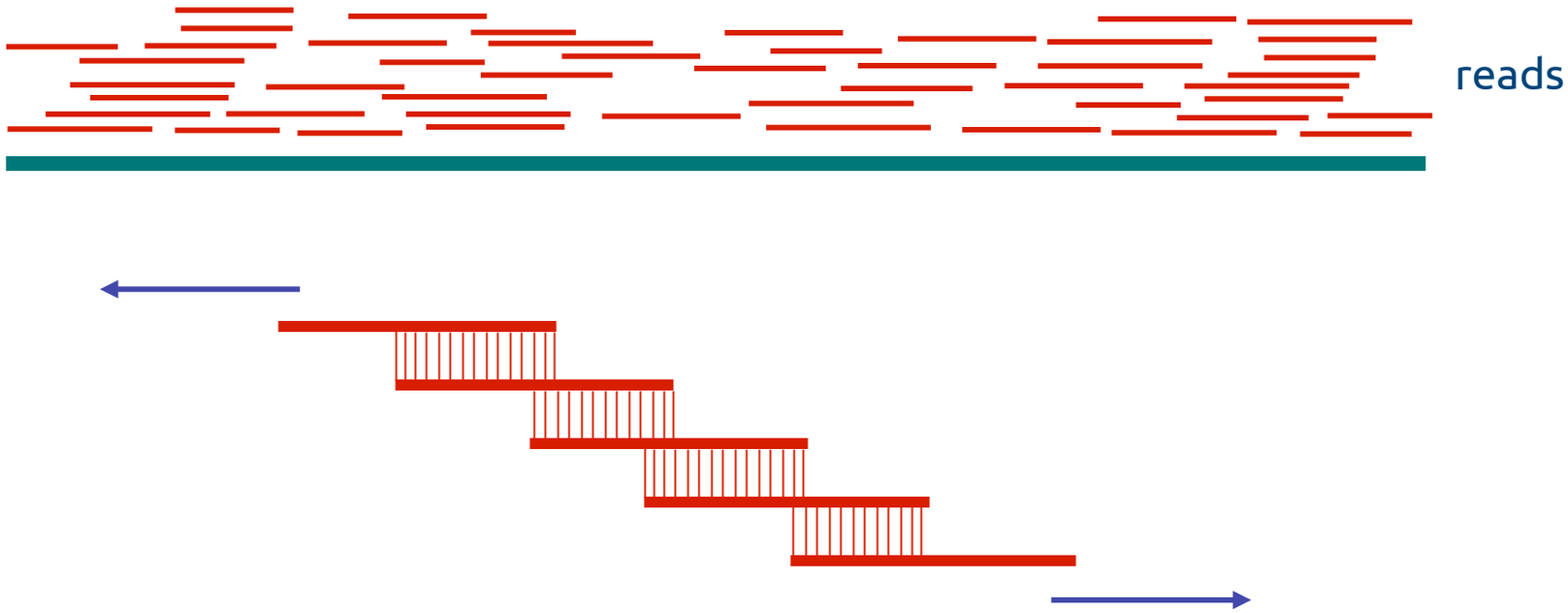
Real Time (PacBio)

Nanopore (Oxford)

Sequencing Overview



Reconstructing the Sequence (Fragment Assembly)



Cover region with high redundancy

Overlap & extend reads to reconstruct the original genomic region

Steps to Assemble a Genome

Some Terminology

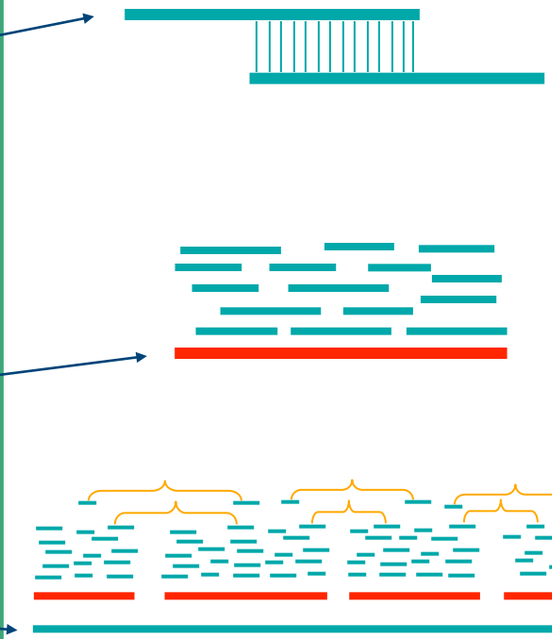
read a 500-900 long word that comes out of sequencer

mate pair a pair of reads from two ends of the same insert fragment

contig a contiguous sequence formed by several overlapping reads with no gaps

supercontig (scaffold) an ordered and oriented set of contigs, usually by mate pairs

consensus sequene sequence derived from the multiple alignment of reads in a contig



..ACGATTACAATAGGTT..

Definition of Coverage

Length of genomic segment: **G**
Number of reads: **N**
Length of each read: **L**

Definition: Coverage **C = N L / G**

How much coverage is enough?

Lander-Waterman model: **Prob[not covered bp] = e^{-C}**
Assuming uniform distribution of reads, C=10 results in 1 gapped region /1,000,000 nucleotides

Draft sequencing of full genome

6 to 8X coverage

SNP finding

>= 20x coverage

Assembly

Join reads to larger sequence: "contigs".

Reference based assembly

De Novo assembly

Publicly available de novo assemblers

Phrap (www.phrap.org)

Celera (wgs-assembler.sf.net)

Paracel (www.paracel.com)

Arachne (ftp://ftp.broadinstitute.org/pub/crd/ARACHNE/)

CAP3 (http://seq.cs.iastate.edu/)

Gene prediction

Evidence based gene calling: BLAST

Ab initio gene calling; no homolog required:
GeneMark, Glimmer, MetaGene.

ORFans

Open Reading Frame (ORFs) with no similarity to any sequence in the database.

Annotation

Finding function of a gene

Next-gen sequencing

Whole Genome Sequencing

RNA-Seq

Exome

Chip-Seq

Methylation (Bisulfite sequencing)

Lior Pachter's list

https://liorpachter.wordpress.com/seq/

Personal Genomes

Table 1 Comparison of sequenced personal human genomes

Individual	Ploidy	Technology	Av Depth	Total SNPs [M]	Known SNPs [M] (%)	Novel SNPs [M] (%)	Heterozygous SNPs [M] (%)	Homozygous SNPs [M] (%)	cSNPs	nsSNPs	InDels	CNVs (≥100 bp)
Venter	2n	Sanger	7.5×	3.21	2.80 (87.22%)	0.41 (12.77%)	1.76 (54.85%)	1.45 (45.15%)	21,152	6,114	214,691	6,485
Watson	2n	Roche 454	7.4×	3.32	2.71 (81.73%)	0.61 (18.27%)	1.67 (50.53%)	1.64 (49.47%)	22,041	10,659	222,718	1,674
Chinese (YH)	2n	Illumina	36.0×	3.07	2.65 (87.13%)	0.41 (12.87%)	1.72 (56.03%)	1.35 (43.97%)	15,759	7,062	135,262	2,682
African (NA18507)*	2n	Illumina	40.6×	3.61	2.72 (75.50%)	0.88 (24.50%)	2.28 (63.21%)	1.32 (36.79%)	26,140	5,361	404,416	8,470
African (NA18507)*	2n	AB SOLiD	17.9×	3.86	3.13 (81.00%)	0.73 (19.00%)	2.33 (60.30%)	1.53 (39.70%)	68,624	9,902	226,529	6,714
Korean (SJK)	2n	Illumina	28.9×	3.43	3.01 (87.79%)	0.42 (12.21%)	2.00 (58.21%)	1.43 (41.79%)	27,118	9,334	342,965	3,303
Korean (AK1)	2n	Illumina	27.8×	3.45	2.86 (83.30%)	0.59 (16.70%)	2.11 (61.11%)	1.34 (38.89%)	21,606	10,162	170,202	414
Khoisan (KB1)	2n	Roche 454	10.2×	4.05	3.31 (81.65%)	0.74 (18.35%)	2.39 (59.00%)	1.66 (41.00%)	22,119	na	463,788	na
D. Tutu (ABT)	2n	AB SOLiD	30.0×	3.62	3.21 (88.61%)	0.41 (11.39%)	2.17 (60.00%)	1.44 (40.00%)	17,342	na	3,395	na
Lupski	2n	AB SOLiD	29.6×	3.42	2.85 (83.58%)	0.56 (16.42%)	2.00 (58.72%)	1.41 (41.28%)	18,406	9,069	na	530

*Same HapMap sample was independently sequenced and reported using two different technologies.

Abbreviations: cSNPs, coding SNPs; nsSNPs, nonsynonymous SNPs; CNVs, copy-number variants; na, data not available.

Gonzaga-Jauregui 2012

Personal Genomes

~14.6 mil non-redundant SNPs

Each genome V reference assembly ~3.5mil SNPs and ~1000 CNVs

The Burrows-Wheeler Transform is a reversible representation with handy properties

- Sort all the possible rotations of original string

- Once BWT(T) is built, *all else shown here is discarded*
  - Matrix will be shown for illustration only

Burrows M, Wheeler DJ: **A block sorting lossless data compression algorithm**. *Digital Equipment Corporation, Palo Alto, CA* 1994, Technical Report 124; 1994

A text occurrence has the same rank in the first and last columns

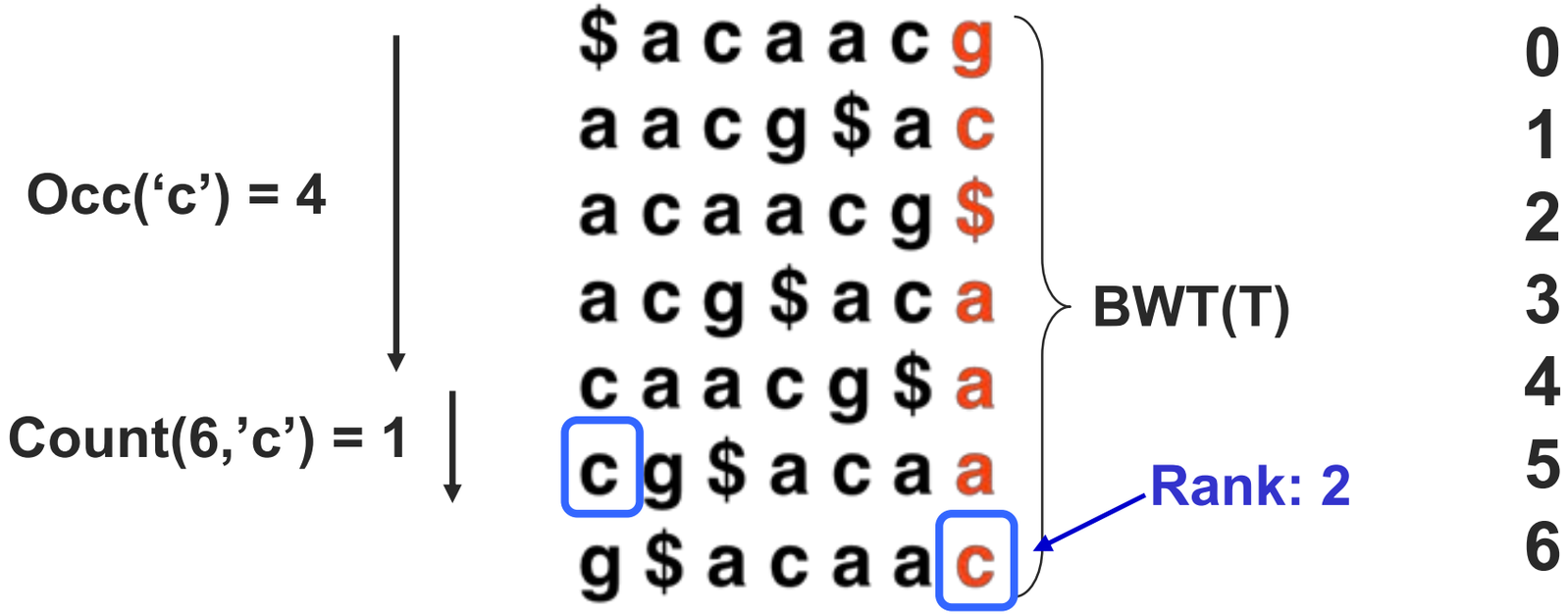
- When we rotate left and sort, the first character retains its rank. Thus the same text occurrence of a character has the same rank in the **L**ast and **F**irst columns.

Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

The Last to First (LF) function matches character and rank

LF(6, ‘c’) = Occ(‘c’) + Count(6,’c’) = 5



Occ(qc) – Number of characters lexically smaller than qc in BWT(T)

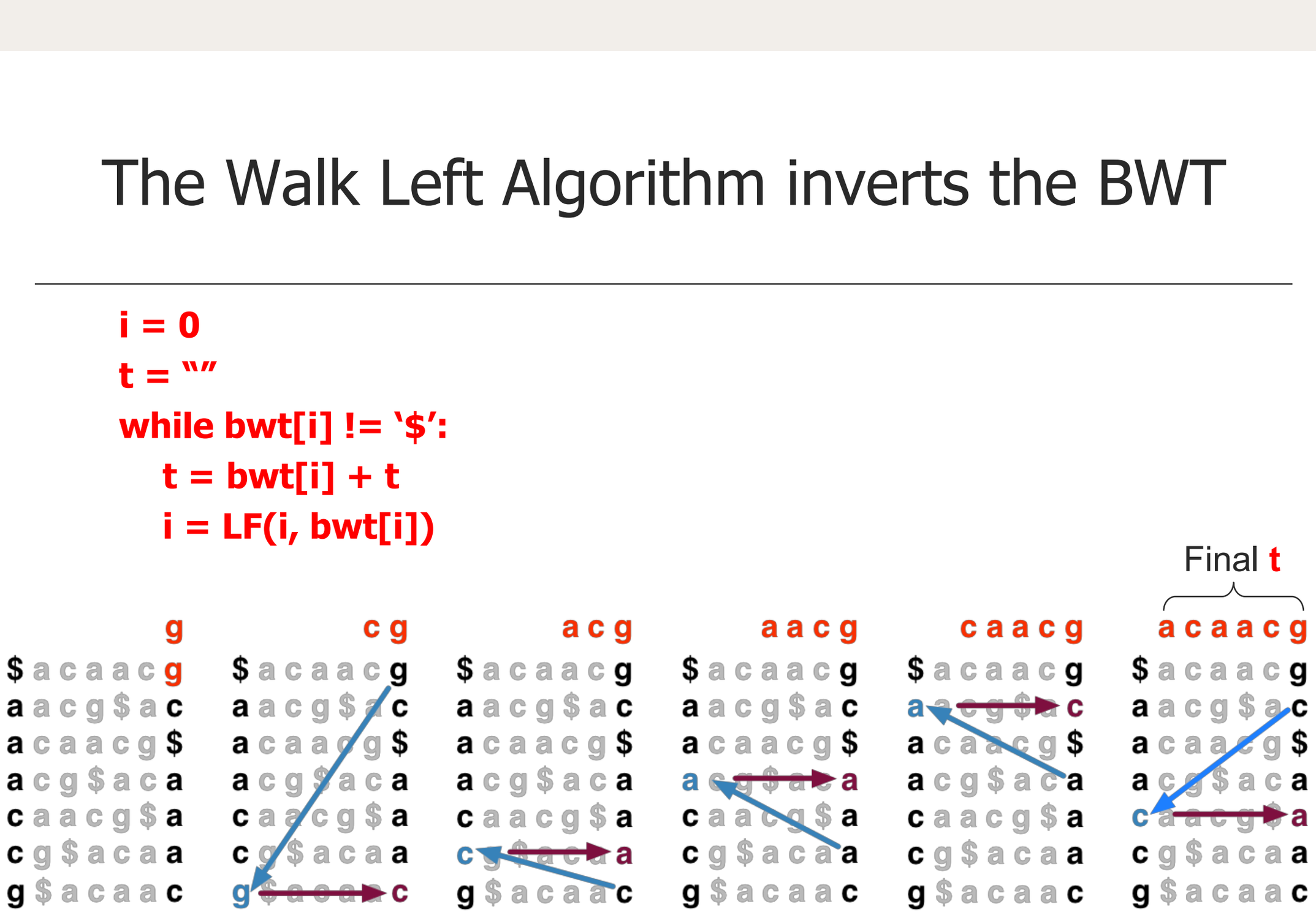
Count(idx, qc) – Number of qc characters before position idx in BWT(T)

Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

The Walk Left Algorithm inverts the BWT

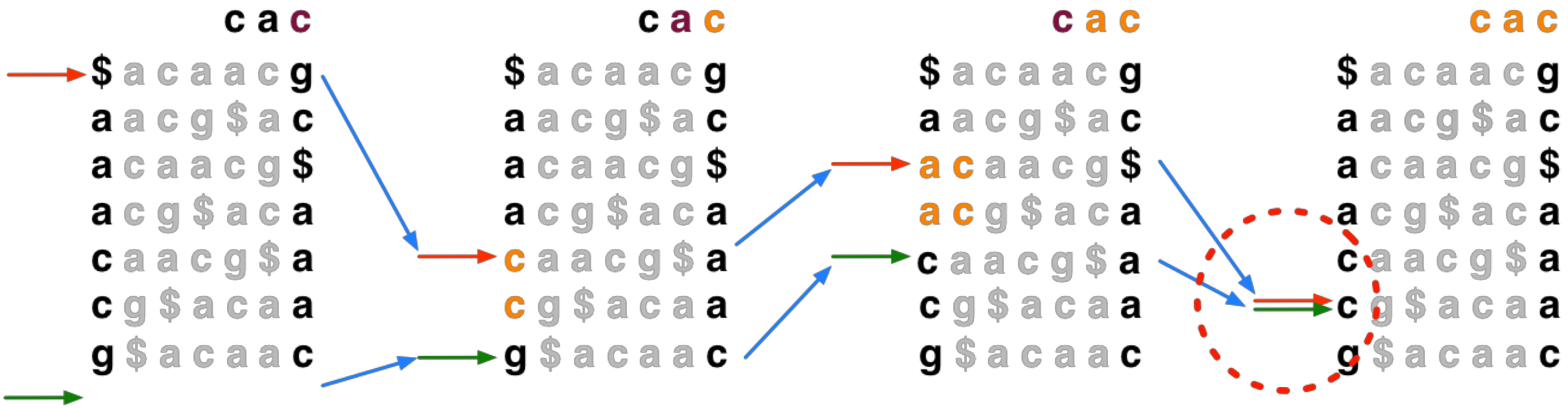
```
i = 0
t = “”
while bwt[i] != ‘$’:
    t = bwt[i] + t
    i = LF(i, bwt[i])
```



Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

Exact Matching with FM Index



- If range becomes empty (**top** = **bot**) the query suffix (and therefore the query) does not occur in the text

Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt