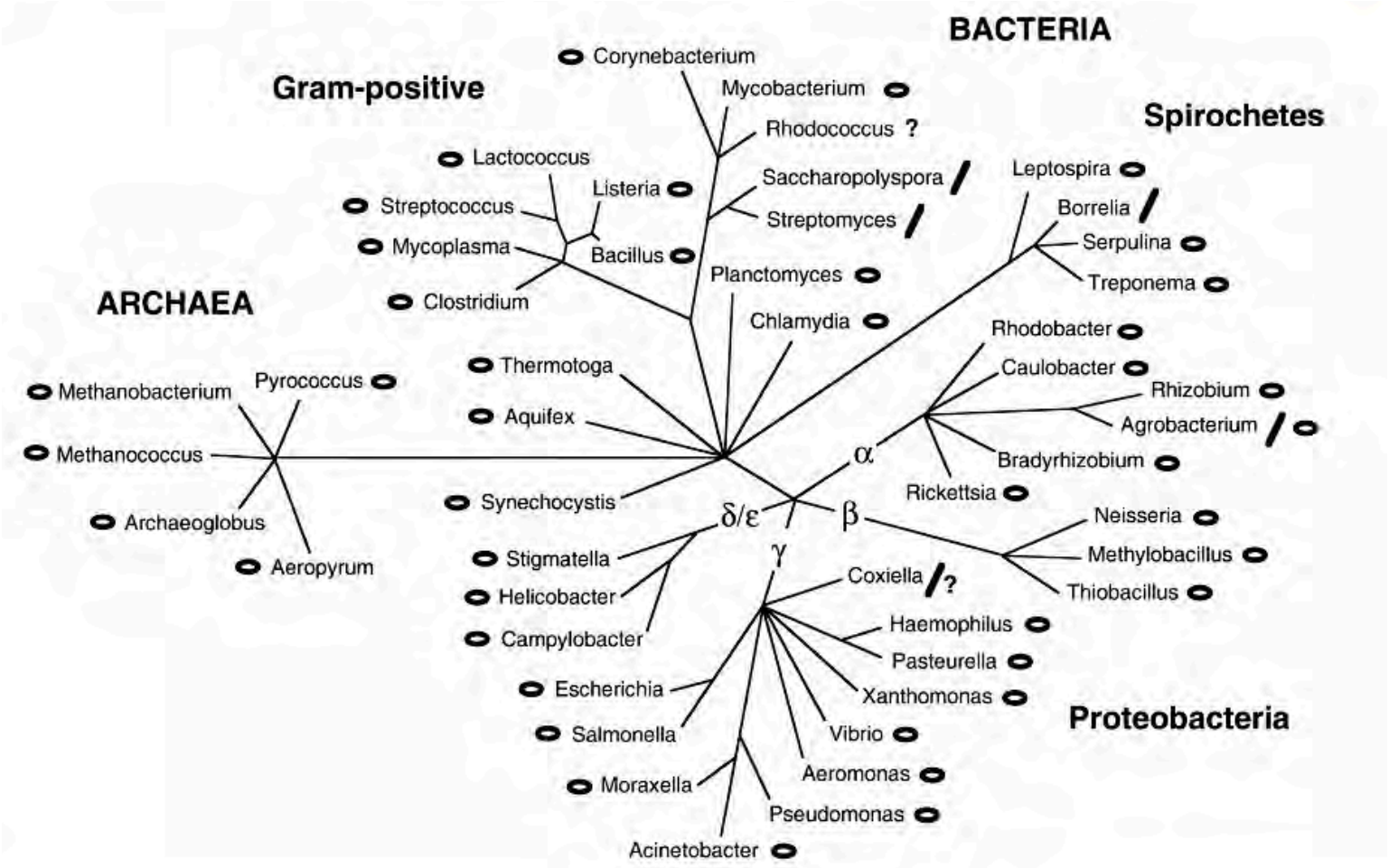
Bacterial genomics

Dmitri Mavrodi

WSU Plant Pathology, dmavrodi@wsu.edu

PLP 514 • December 06, 2012

Bacterial genome features

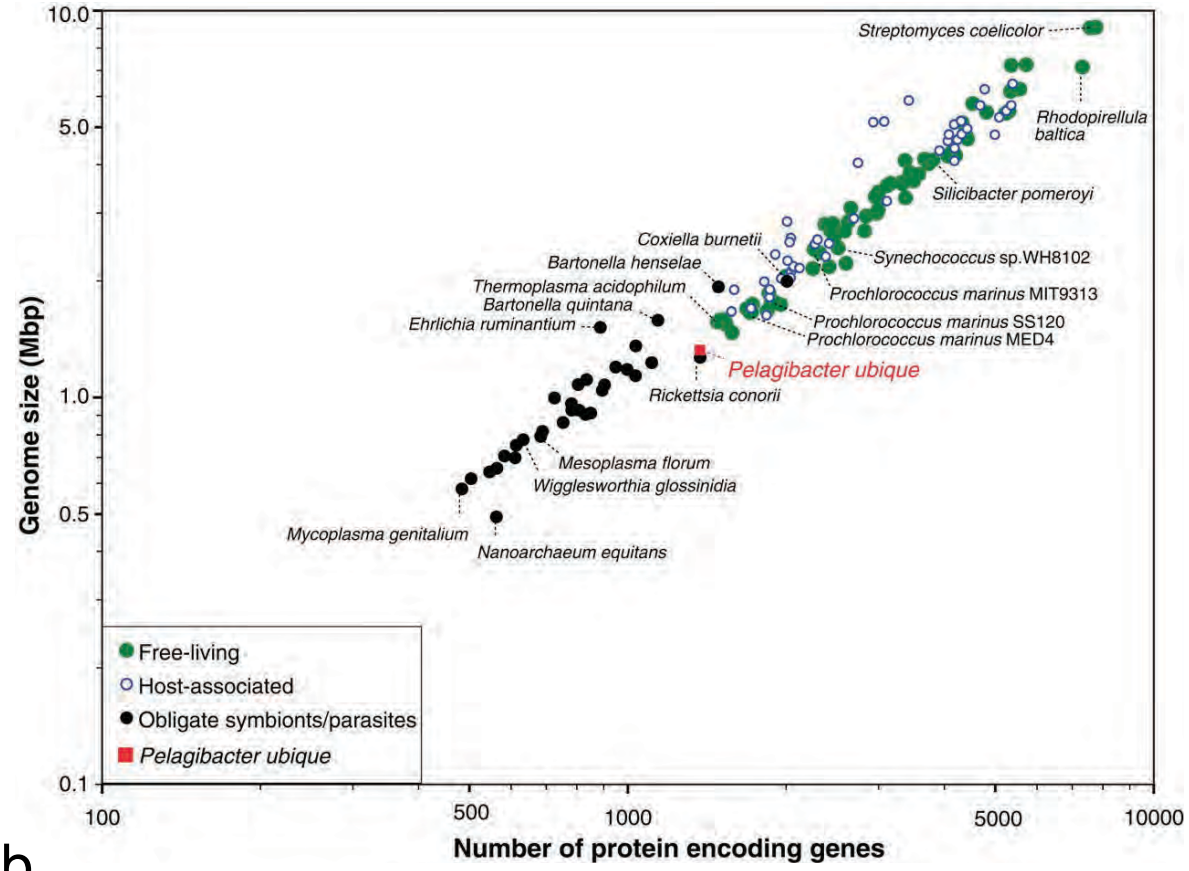


Volff & Altenbuchner (2000) FEMS Lett

Bacterial genome features

They are small and typically expressed in megabases (1 Mb = 10^6 bp)

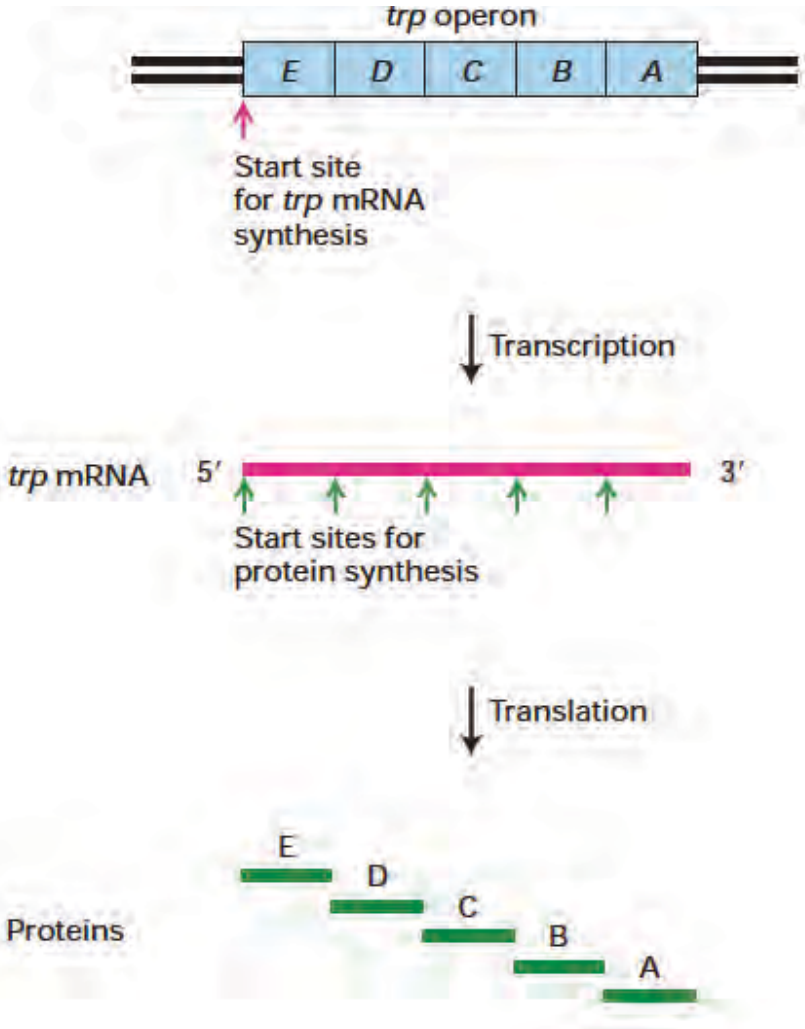
- Bacteria 0.8 - 9 Mb
- Yeast 18 Mb
- *C. elegans* 100 million
- Rice 420 Mb
- *Arabidopsis* 125 Mb
- Fugu 450 Mb
- Mouse 3,000 Mb
- Corn 2,500 Mb
- Human 3,200 Mb
- Wheat 16,000 Mb
- Loblolly pine 20,000+ Mb



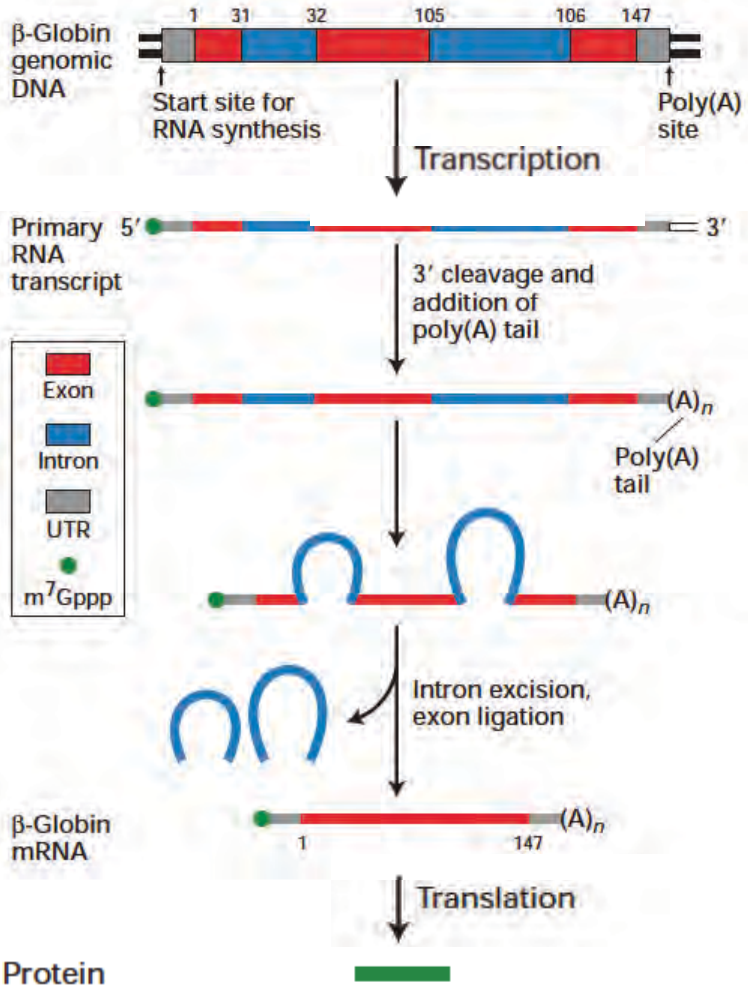
Giovannoni et al (2005) Science

Gene structure in prokaryotes and eukaryotes

Prokaryotes

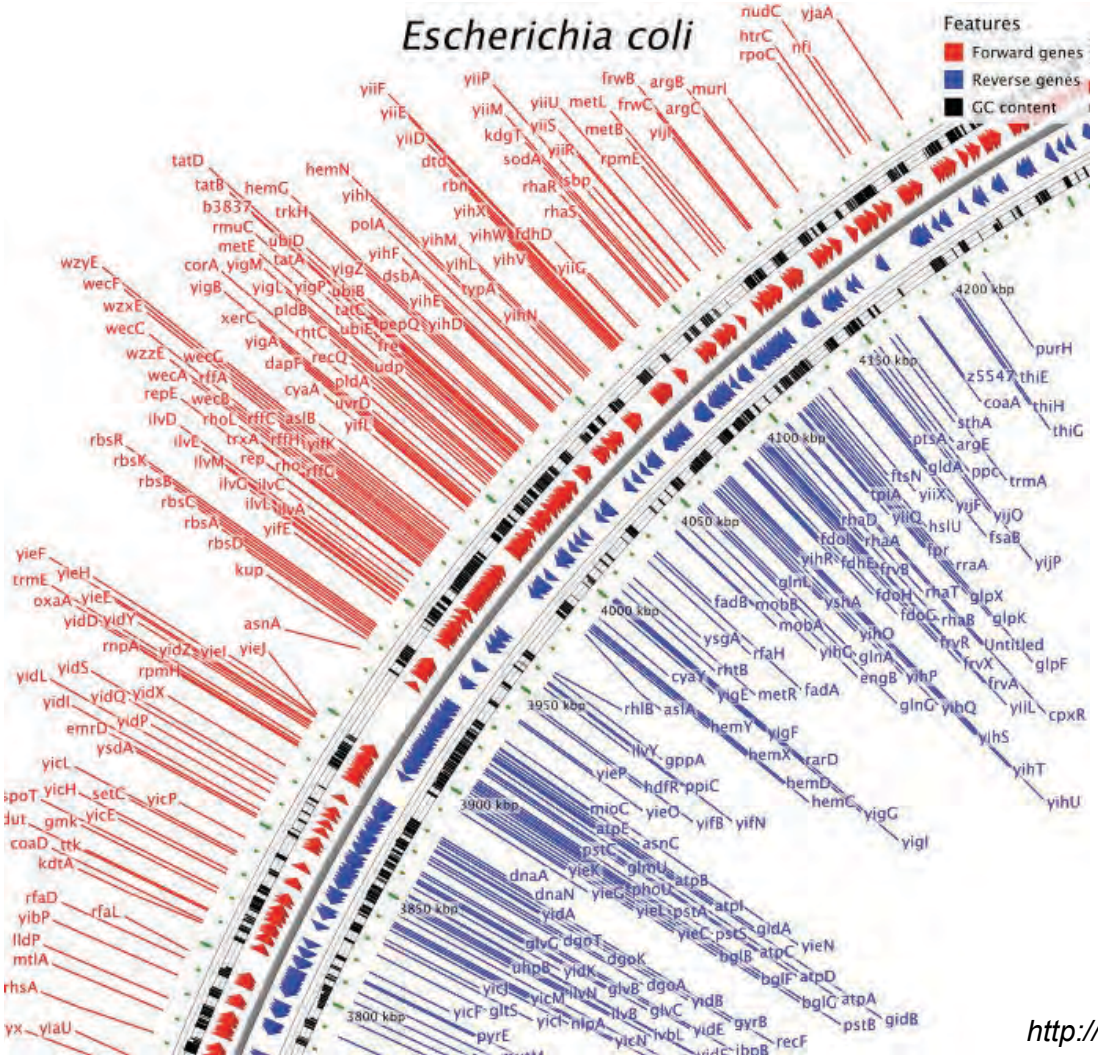


Eukaryotes



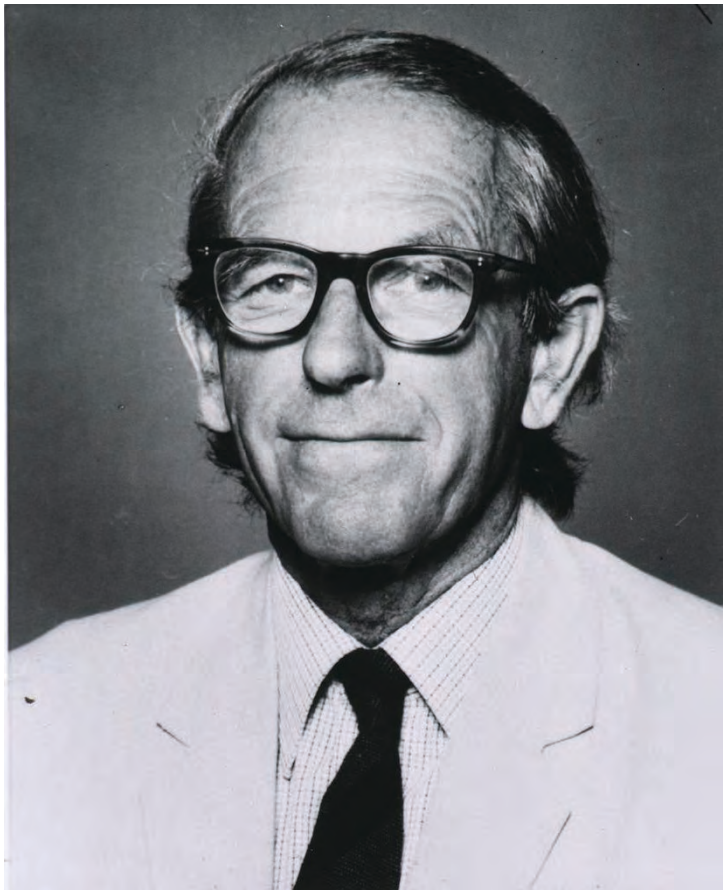
Bacterial genome features

Bacterial genomes are gene-dense



http://wishart.biology.ualberta.ca

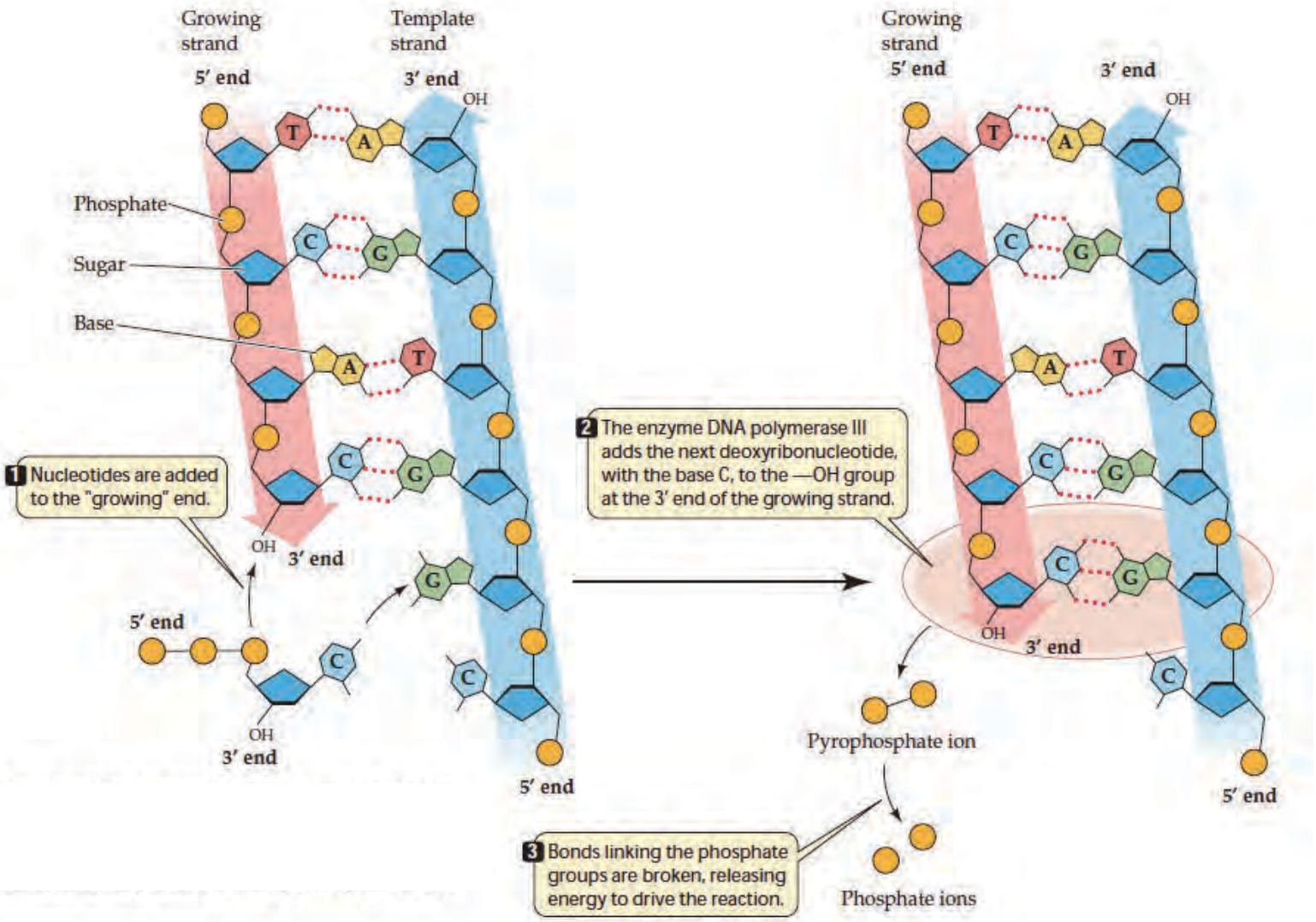
1st generation sequencing technologies



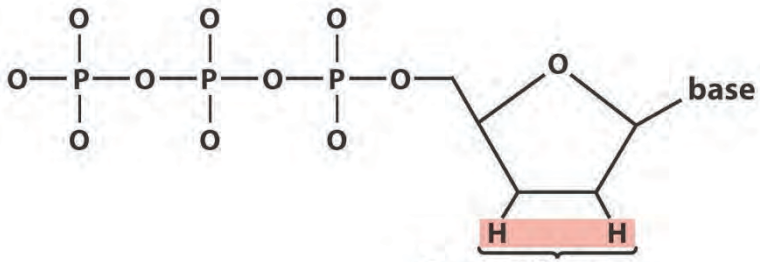
Fred Sanger

- In the 1960s and 1970s, British scientists Frederick Sanger and Alan Coulson, and Alan Maxam and Walter Gilbert in the US, developed DNA sequencing techniques. Gilbert and Sanger won the 1980 Nobel Prize in Chemistry for their

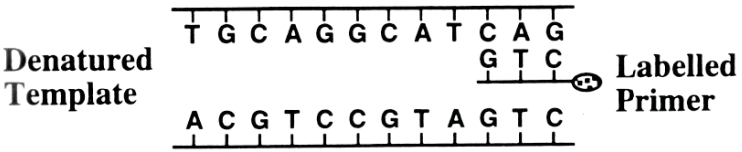
DNA replication



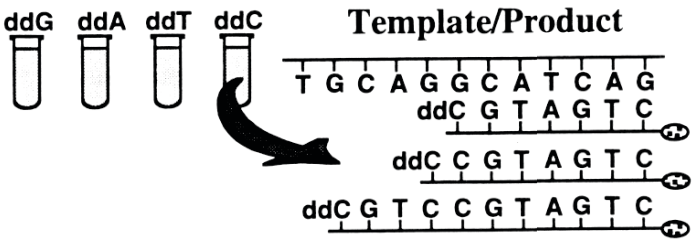
Sanger's dideoxy sequencing technique

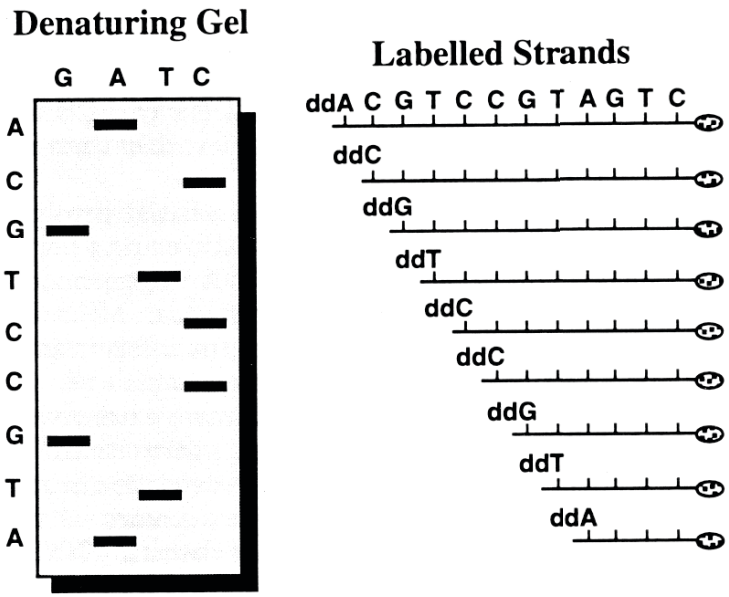


Cannot form a phosphodiester bond with next incoming dNTP

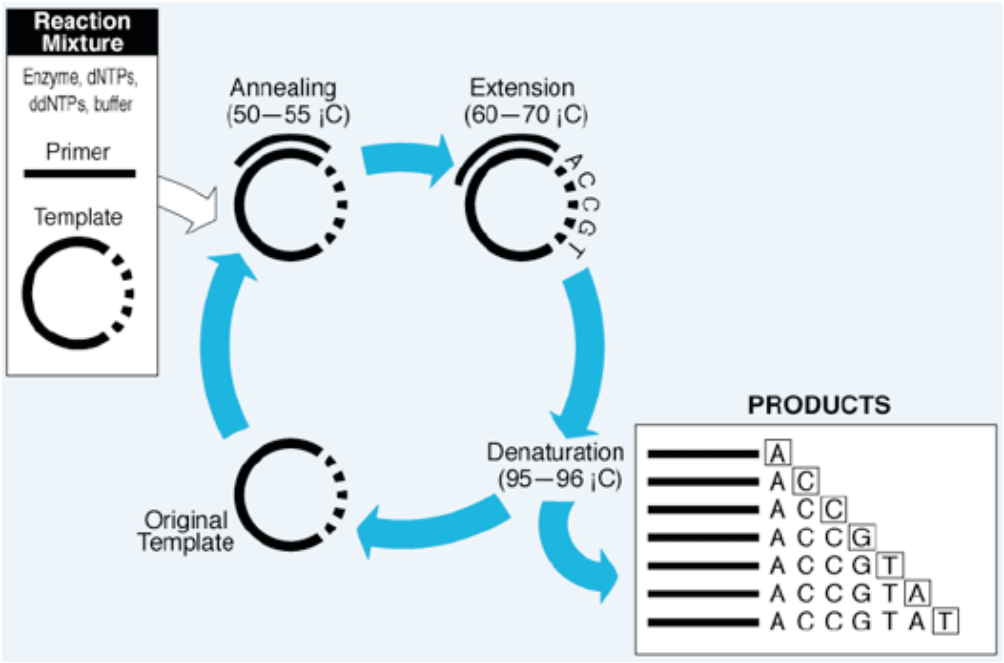


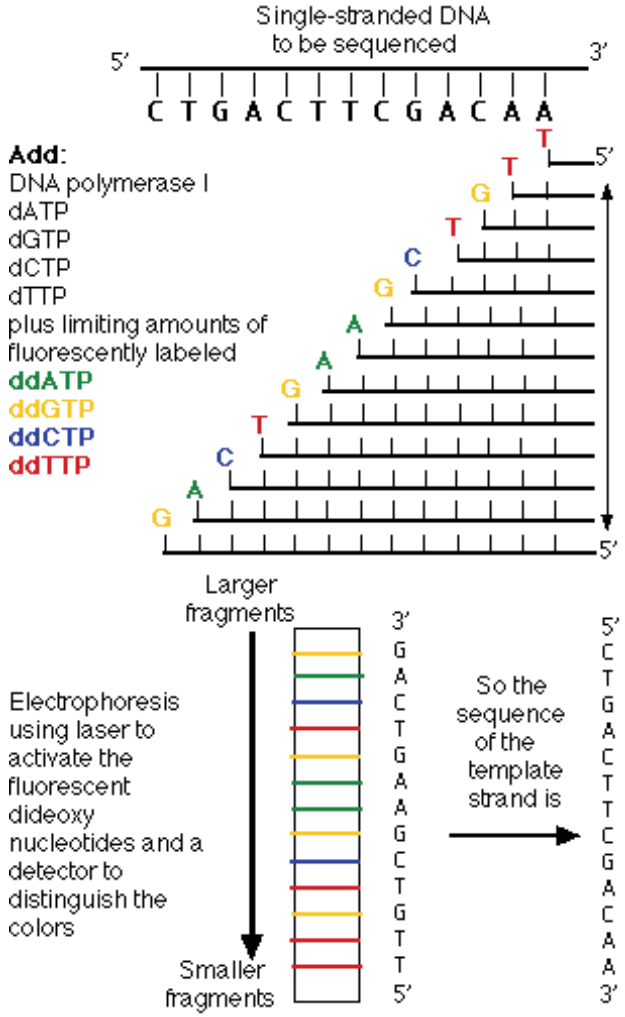
Add dNTPs and Polymerase



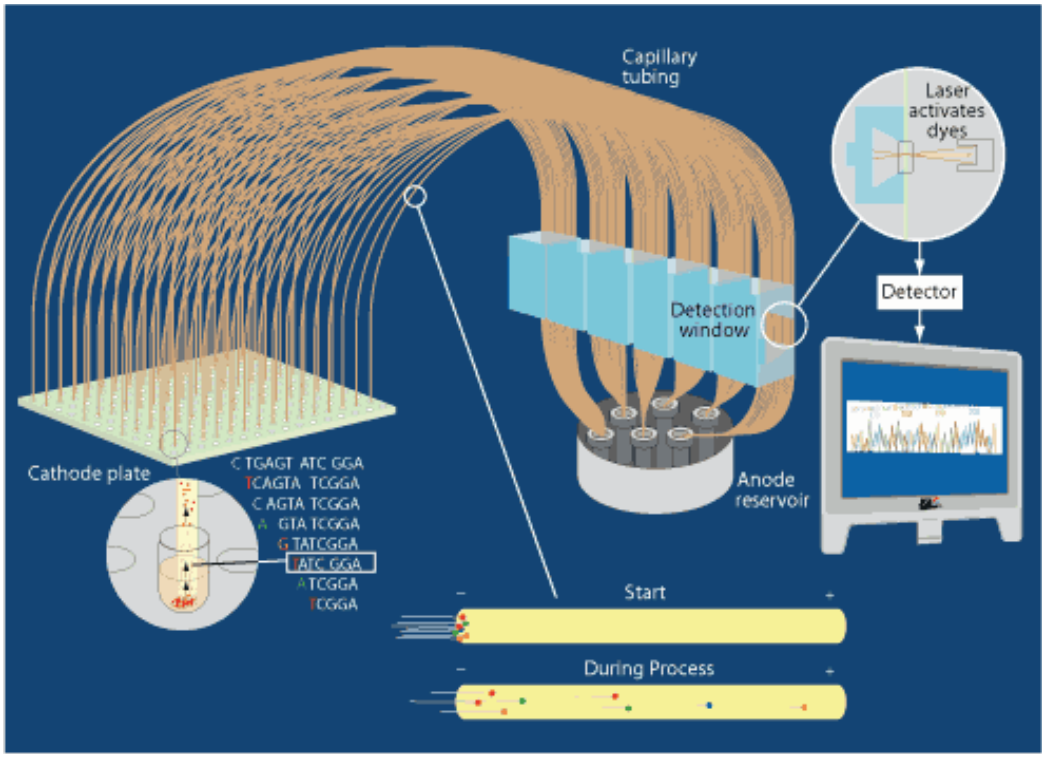


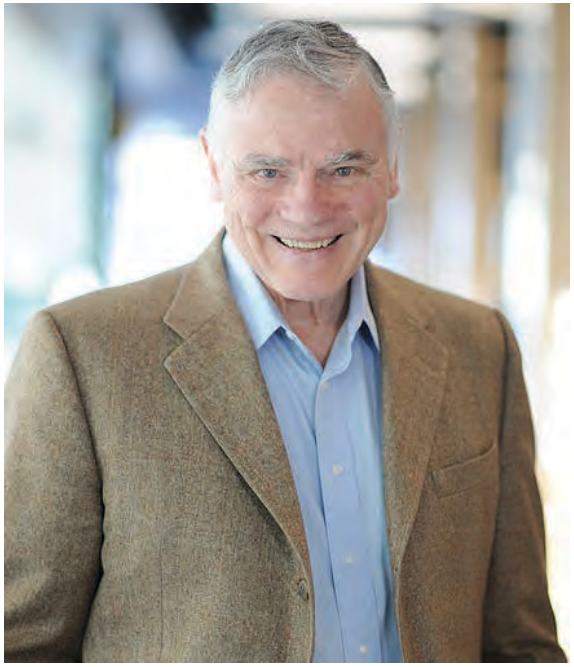
Cycle sequencing and fluorescent labels



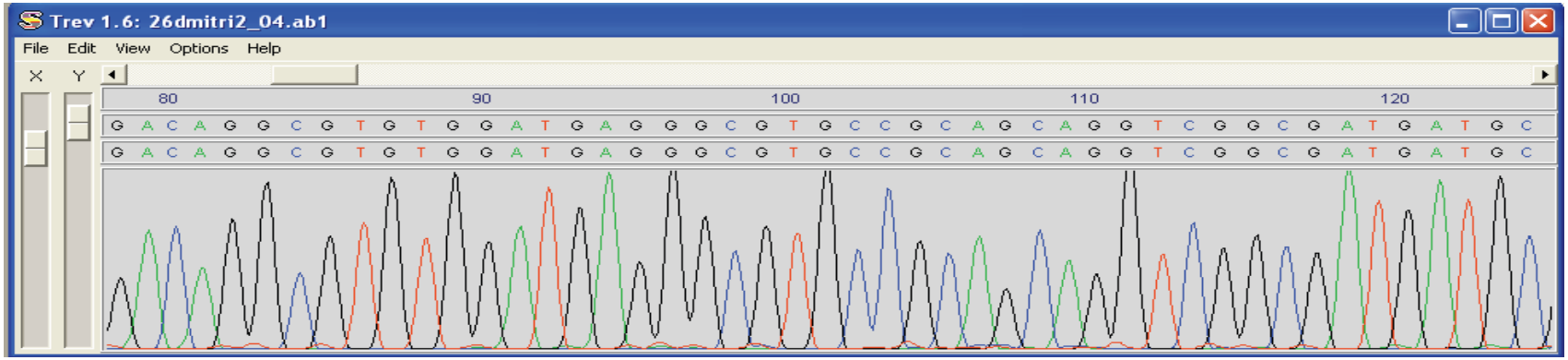


Capillary sequencing





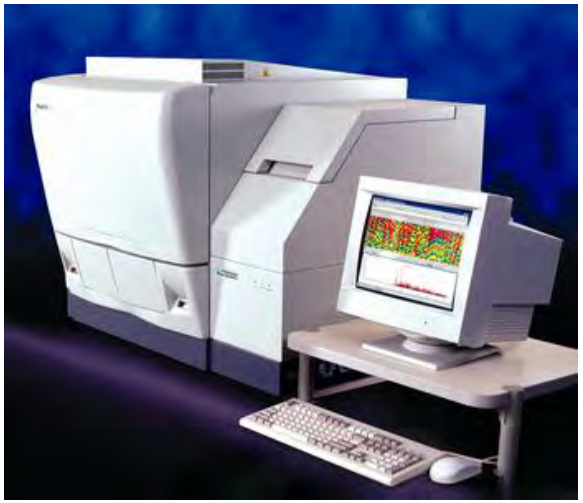
Leroy Hood



Capillary sequencers

ABI PRISM 3100
Genetic Analyzer
(16 capillaries)

ABI PRISM 3730xl
DNA Analyzer
(96 capillaries)



MegaBACE 4000 DNA
Analysis System
(384 capillaries)

The greatest achievement...

Comparison of 1st, 2nd and 3rd generation sequencing technologies

	1st generation	2nd generation	3rd generation
Approach	Separation of labeled fragments produced by SBS	Wash-and-scan SBS	SBS or by degradation
Resolution	Many molecules are sequenced	Many molecules are sequenced	Single-molecule sequencing
Read accuracy	High	High	Moderate
Read length	Moderate (800–1000 bp)	Short	Long, 1 kb and longer
Throughput	Low	High	Moderate
Cost/base	High	Low	Low-to-moderate
Cost/run	Low	High	Low
Time	Hours	Days	Hours
Sample preparation	Moderately complex, no amplification	Complex, PCR amplification required	Ranges from complex to simple
Data analysis	Routine	Complex: large data volumes, short reads	Complex: large data volumes, signal processing

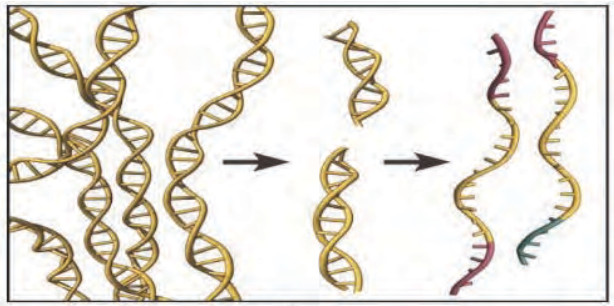
2nd generation sequencing technologies

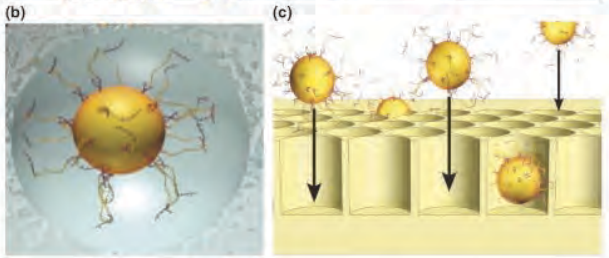
2nd generation sequencing includes **wash-and-scan** techniques

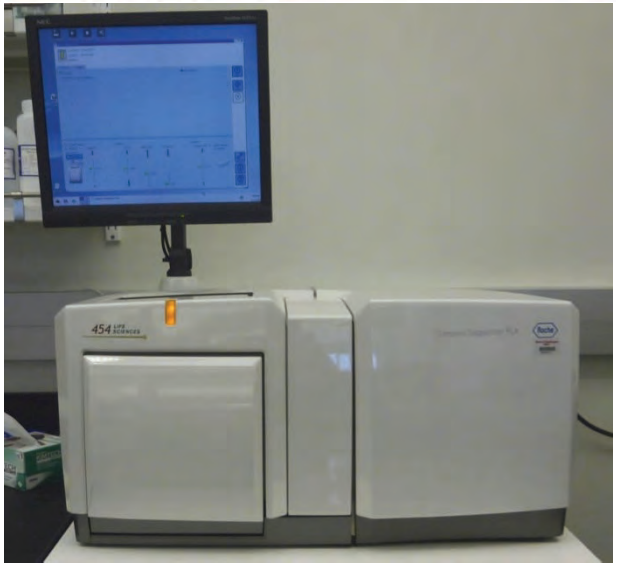
The DNA polymerase and other reagents are washed off after adding every base or an oligonucleotide; these steps are repeated many times, which consumes a lot of reagents and time

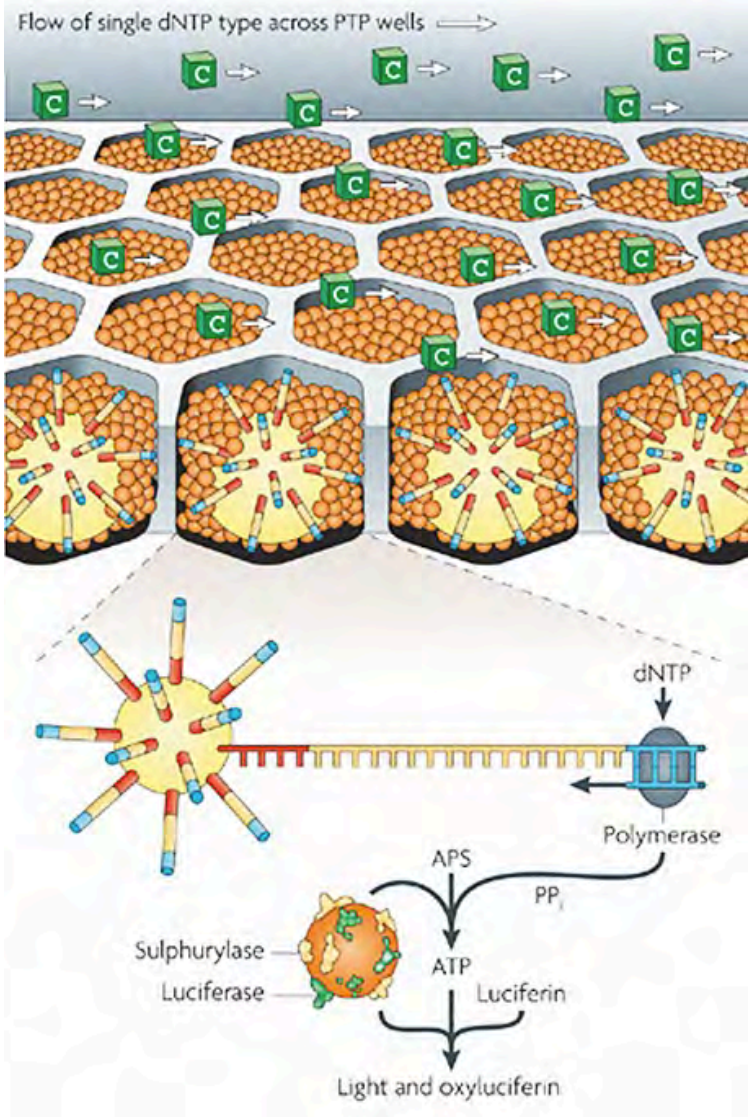
2nd generation sequencing technologies

454 SEQUENCING





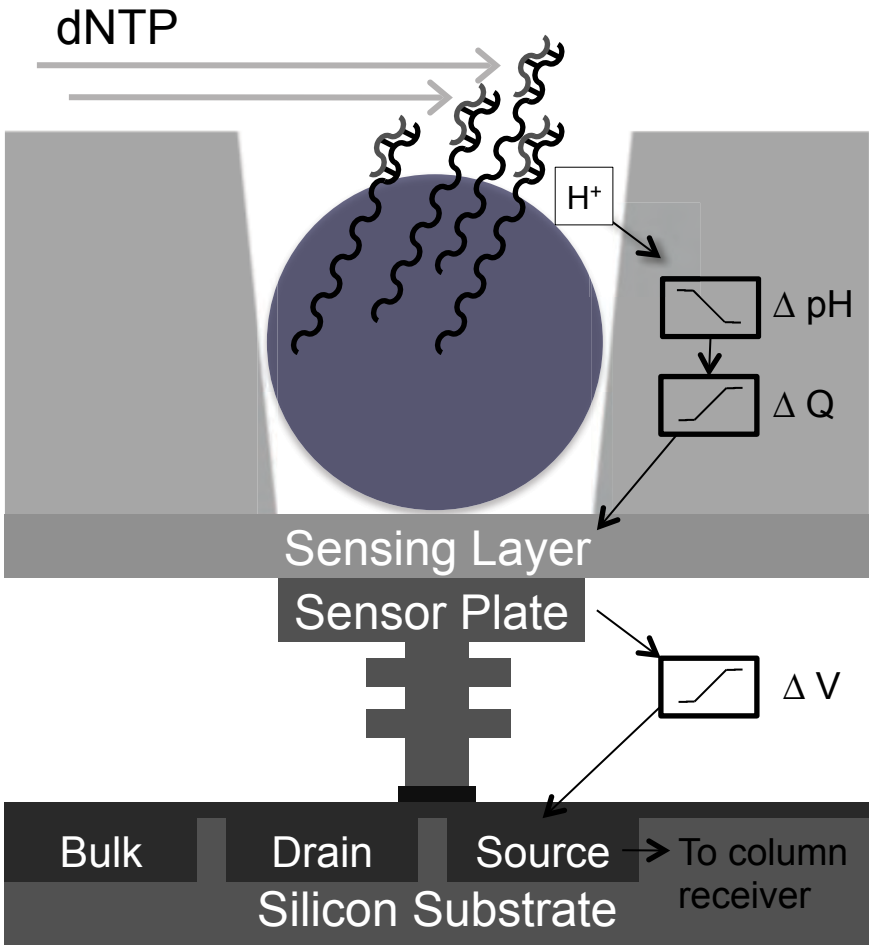


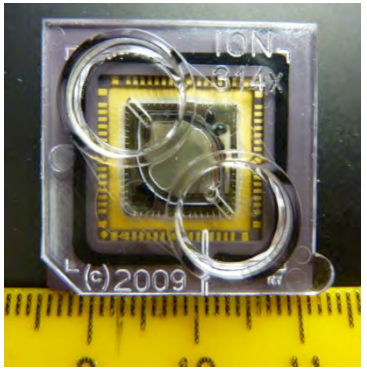


2nd generation sequencing technologies

ion torrent

by life technologies™





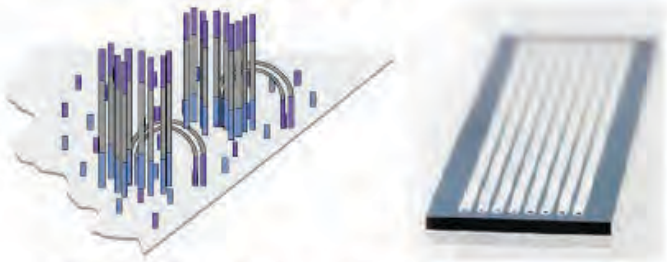
2nd generation sequencing technologies

illumina®

1. Sample Prep
(1-5 days)

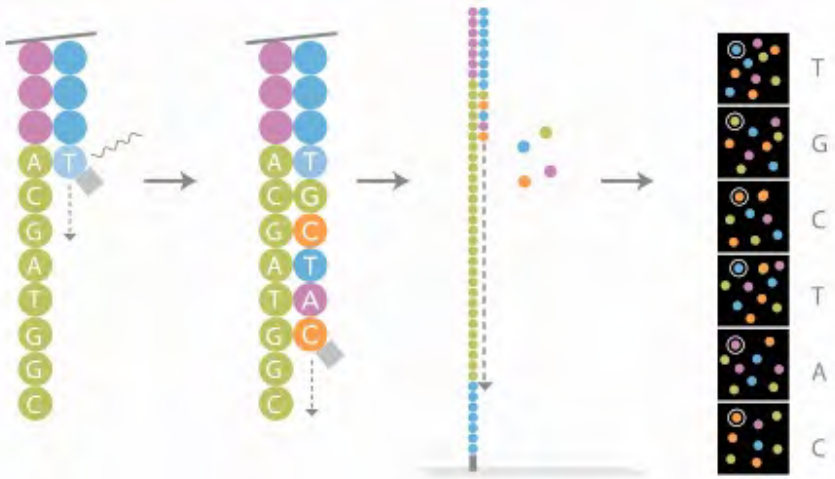
Ligate adapters

2. Cluster generation on flow cell
(1.5 day)



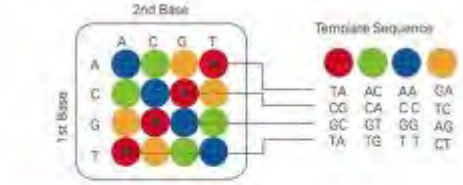
Clonal Single molecular Array

3. Sequencing and imaging



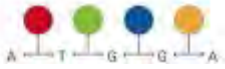
Applied Biosystems

Possible Dinucleotides Encoded By Each Color



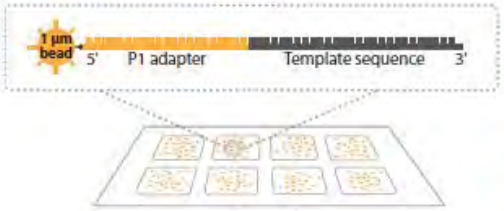
Double Interrogation

With 2 base encoding each base is defined twice

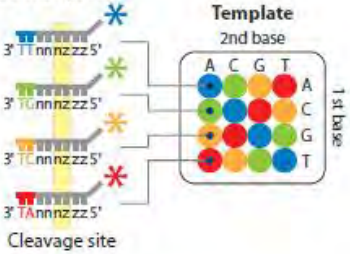


a

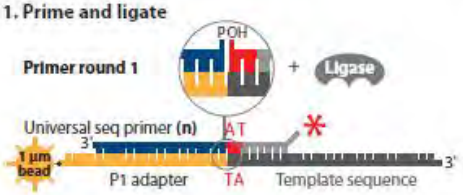
SOLiD™ substrate



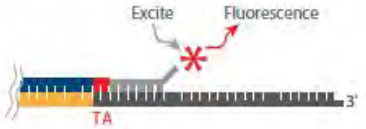
Di base probes



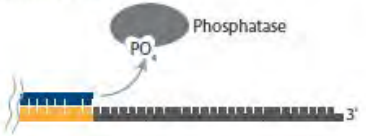
1. Prime and ligate



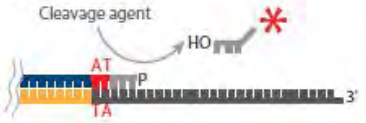
2. Image



3. Cap unextended strands

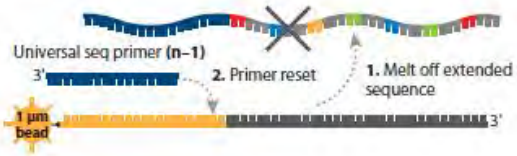


4. Cleave off fluor

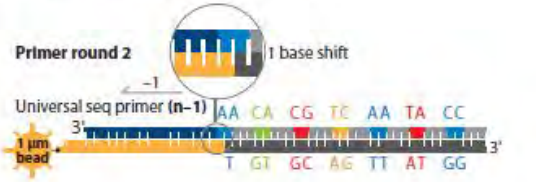


5. Repeat steps 1–4 to extend sequence

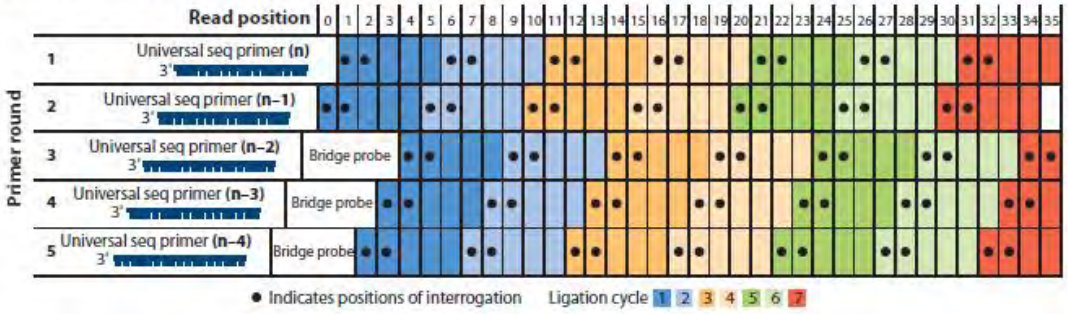
6. Primer reset



7. Repeat steps 1–5 with new primer



8. Repeat Reset with , n–2, n–3, n–4 primers

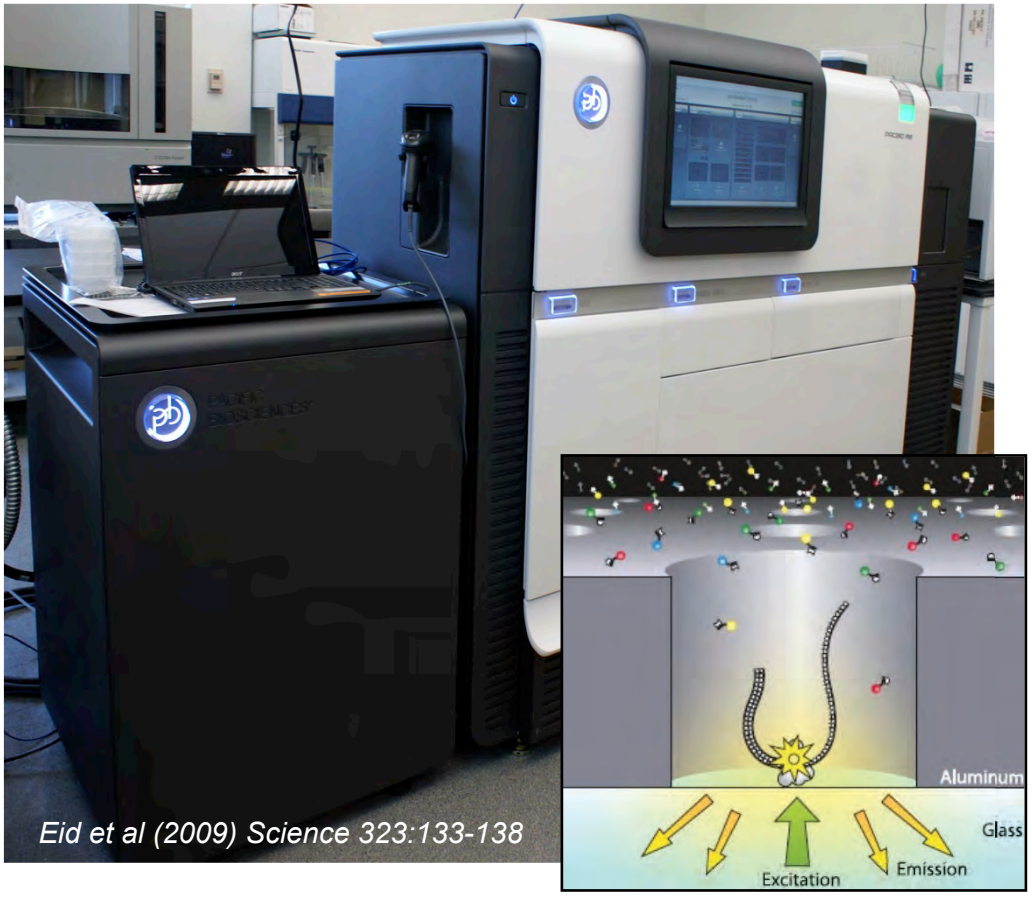


Comparison of 1st, 2nd and 3rd generation sequencing technologies

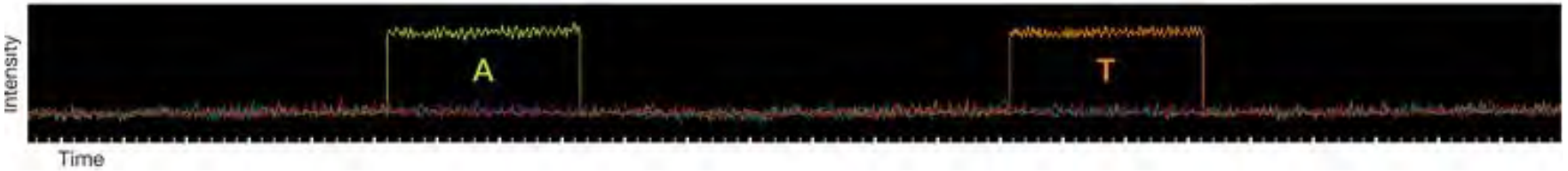
	1st generation	2nd generation	3rd generation
Approach	Separation of labeled fragments produced by SBS	Wash-and-scan SBS	SBS or by degradation
Resolution	Many molecules are sequenced	Many molecules are sequenced	Single-molecule sequencing
Read accuracy	High	High	Moderate
Read length	Moderate (800–1000 bp)	Short	Long, 1 kb and longer
Throughput	Low	High	Moderate
Cost/base	High	Low	Low-to-moderate
Cost/run	Low	High	Low
Time	Hours	Days	Hours
Sample preparation	Moderately complex, no amplification	Complex, PCR amplification required	Ranges from complex to simple
Data analysis	Routine	Complex: large data volumes, short reads	Complex: large data volumes, signal processing

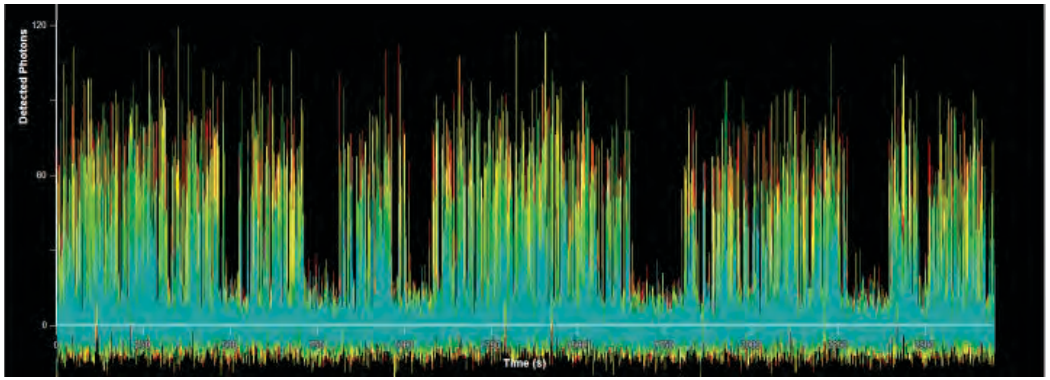
3nd generation sequencing technologies

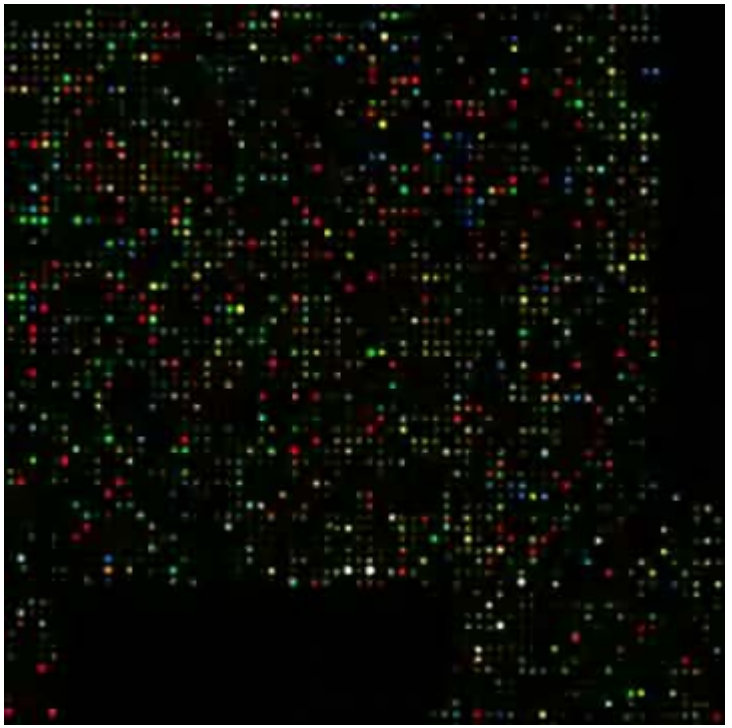
- PacBio – a third generation platform that relies on real-time sequencing from a single DNA polymerase molecule



Eid et al (2009) Science 323:133-138

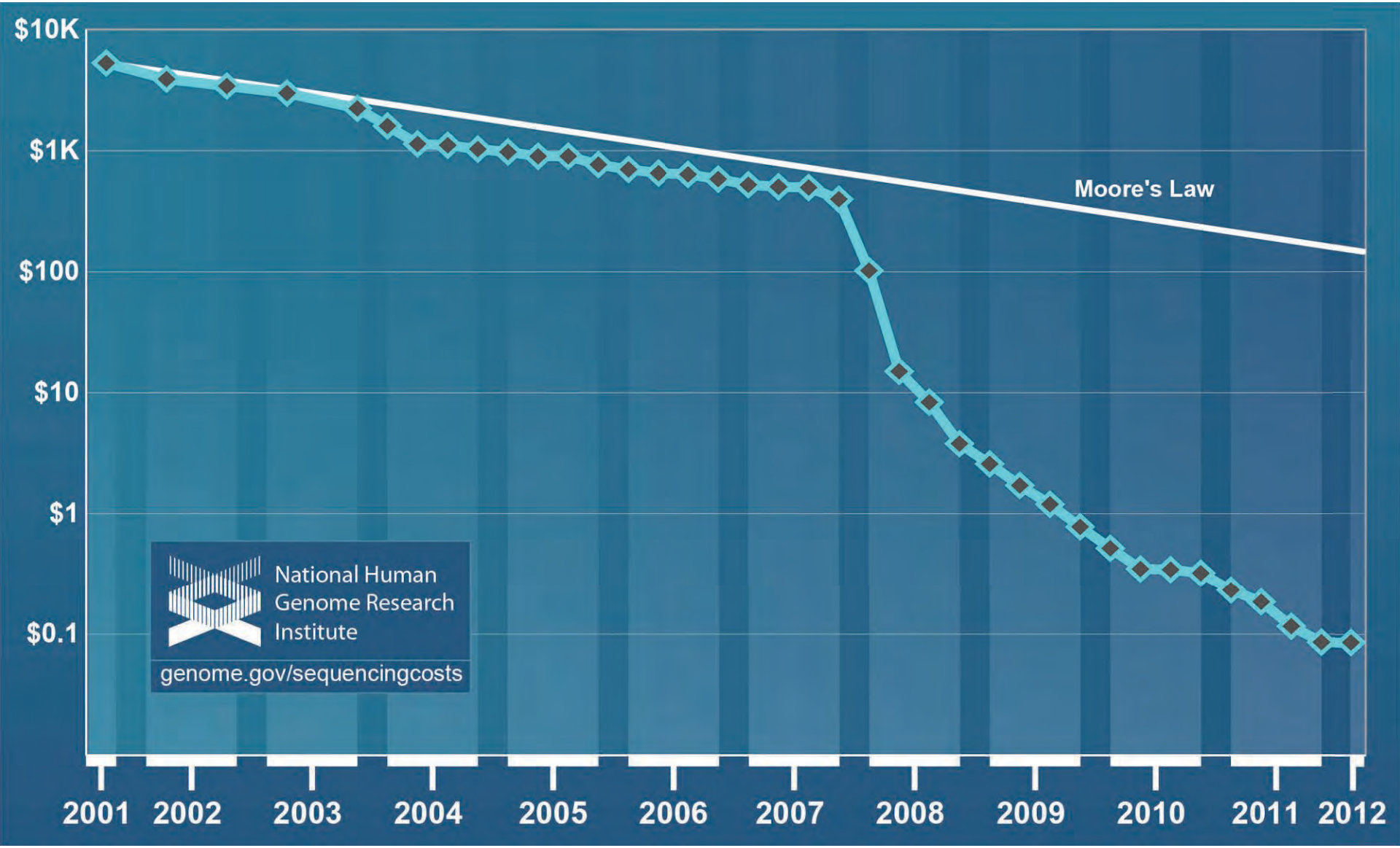






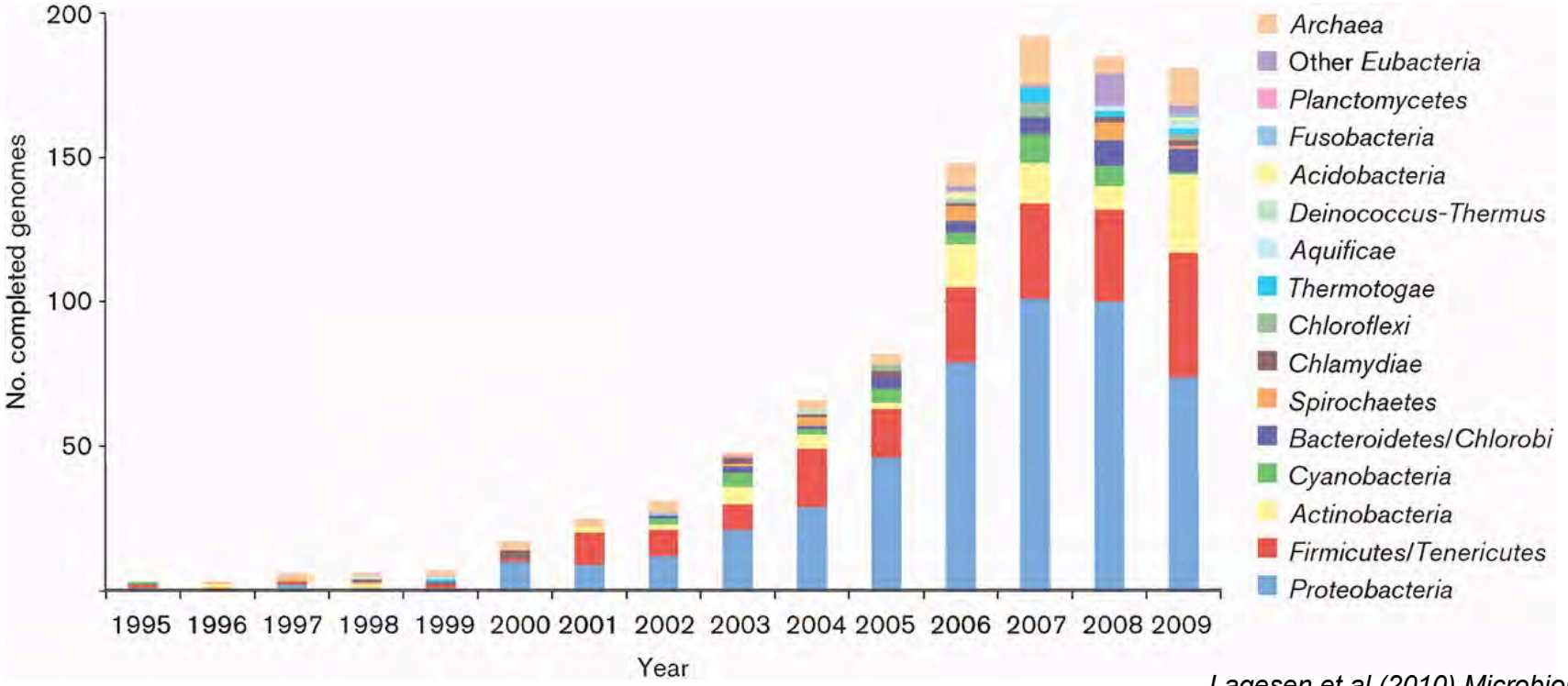
Ritz et al (2010) Bioinformatics

Cost of per raw Mb of DNA sequencing



Date	Cost per Mb	Cost per Genome	Date	Cost per Mb	Cost per Genome
Sep-01	$5,292.39	$95,263,072	Jul-07	$495.96	$8,927,342
Mar-02	$3,898.64	$70,175,437	Oct-07	$397.09	$7,147,571
Sep-02	$3,413.80	$61,448,422	Jan-08	$102.13	$3,063,820
Mar-03	$2,986.20	$53,751,684	Apr-08	$15.03	$1,352,982
Oct-03	$2,230.98	$40,157,554	Jul-08	$8.36	$752,080
Jan-04	$1,598.91	$28,780,376	Oct-08	$3.81	$342,502
Apr-04	$1,135.70	$20,442,576	Jan-09	$2.59	$232,735
Jul-04	$1,107.46	$19,934,346	Apr-09	$1.72	$154,714
Oct-04	$1,028.85	$18,519,312	Jul-09	$1.20	$108,065
Jan-05	$974.16	$17,534,970	Oct-09	$0.78	$70,333
Apr-05	$897.76	$16,159,699	Jan-10	$0.52	$46,774
Jul-05	$898.90	$16,180,224	Apr-10	$0.35	$31,512
Oct-05	$766.73	$13,801,124	Jul-10	$0.35	$31,125
Jan-06	$699.20	$12,585,659	Oct-10	$0.32	$29,092
Apr-06	$651.81	$11,732,535	Jan-11	$0.23	$20,963
Jul-06	$636.41	$11,455,315	Apr-11	$0.19	$16,712
Oct-06	$581.92	$10,474,556	Jul-11	$0.12	$10,497
Jan-07	$522.71	$9,408,739	Oct-11	$0.09	$7,743
Apr-07	$502.61	$9,047,003	Jan-12	$0.09	$7,666

Increase in the number of genomes sequenced per year (1000 genomes in 2009)



Lagesen et al (2010) Microbiology

华大基因 BGI

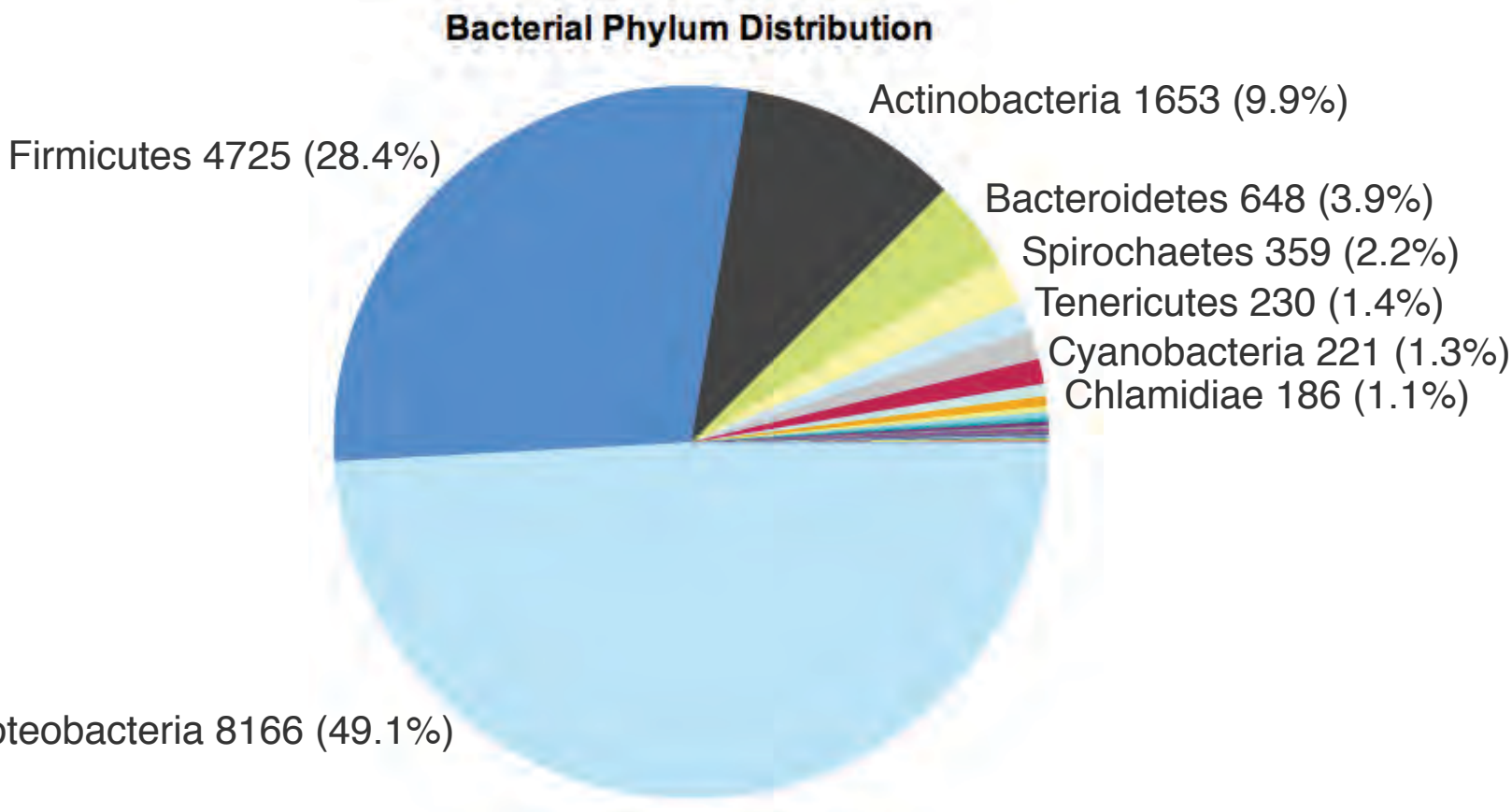
Beijing Genomics Institute

→

The Ten Thousand Microbial Genomes Project!

Today...

As of December 2012, 17009 bacterial genome projects are listed in the Genomes Online Database



What do you do with all these data?

The main aim is to correlate the genome sequence of an organism with its phenotype

Genome sequences need to be annotated first - searched for putative genes and analyzed for the function of the encoded proteins

The annotated genomes can be searched for genes responsible for a phenotype of interest (in the case of plant pathogens, any aspect of their interactions with plants)

What do you do with all these data?

Genome sequences can also be compared with those of closely related organisms

Transcriptomic, proteomic and molecular genetic screens can further elucidate gene function

Genome comparison data and experimental data can be dowloades in the internet databases to maximize utility to the research community

P. fluorescens group genome sequencing project

OPEN ACCESS Freely available online

PLoS GENETICS

Comparative Genomics of Plant-Associated *Pseudomonas* spp.: Insights into Diversity and Inheritance of Traits Involved in Multitrophic Interactions

Joyce E. Loper[1,2*], Karl A. Hassan[3], Dmitri V. Mavrodi[4], Edward W. Davis II[1], Chee Kent Lim[3], Brenda T. Shaffer[1], Liam D. H. Elbourne[3], Virginia O. Stockwell[2], Sierra L. Hartney[2], Katy Breakwell[3], Marcella D. Henkels[1], Sasha G. Tetu[3], Lorena I. Rangel[2], Teresa A. Kidarsa[1], Neil L. Wilson[3], Judith E. van de Mortel[5], Chunxu Song[5], Rachel Blumhagen[1], Diana Radune[6], Jessica B. Hostetler[6], Lauren M. Brinkac[6], A. Scott Durkin[6], Daniel A. Kluepfel[7], W. Patrick Wechter[8], Anne J. Anderson[9], Young Cheol Kim[10], Leland S. Pierson III[11], Elizabeth A. Pierson[12], Steven E. Lindow[13], Donald Y. Kobayashi[14], Jos M. Raaijmakers[5], David M. Weller[15], Linda S. Thomashow[15], Andrew E. Allen[16], Ian T. Paulsen[3]

1 Agricultural Research Service, U.S. Department of Agriculture, Corvallis, Oregon, United States of America, 2 Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, United States of America, 3 Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia, 4 Department of Plant Pathology, Washington State University, Pullman, Washington, United States of America, 5 Laboratory of Phytopathology, Wageningen University, Wageningen, The Netherlands, 6 The J. Craig Venter Institute, Rockville, Maryland, United States of America, 7 Agricultural Research Service, U.S. Department of Agriculture, Davis, California, United States of America, 8 Agricultural Research Service, U.S. Department of Agriculture, Charleston, South Carolina, United States of America, 9 Department of Biology, Utah State University, Logan, Utah, United States of America, 10 Institute of Environmentally-Friendly Agriculture, Chonnam National University, Gwangju, Korea, 11 Department of Plant Pathology and Microbiology, Texas A&M University, College Station, Texas, United States of America, 12 Department of Horticultural Sciences, Texas A&M University, College Station, Texas, United States of America, 13 Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, California, United States of America, 14 Department of Plant Biology and Pathology, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, United States of America, 15 Agricultural Research Service, U.S. Department of Agriculture, Pullman, Washington, United States of America, 16 The J. Craig Venter Institute, San Diego, California, United States of America

Loper et al (2012) PLoS Genetics 8: e1002784

P. fluorescens group genome sequencing project

- Sequenced the genomes of 7 strains of the *P. fluorescens* group that colonize plant surfaces and function as biological control agents
- Compared strains to each other and to three other strains that were sequenced previously
- Looked for known and new genes that are likely to confer biocontrol traits and/or contribute to plant-microbe interactions

Loper et al (2012) PLoS Genetics 8: e1002784

P. fluorescens group genome sequencing project

Strain	Source	Target disease(s) for biological control
***P. chlororaphis* subsp. *aureofaciens*:**		
30-84	Wheat rhizosphere, Washington, USA	Take-all of wheat
O6	Soil, Utah, USA	Wildfire of tobacco target spot of cucumber
***P. protegens*:**		
Pf-5	Soil, Texas, USA	Seedling emergence
***P. brassicacearum*:**		
Q8r1-96	Wheat rhizosphere, Washington, USA	Take-all of wheat
***P. fluorescens*:**		
Pf0-1	Soil, Massachusetts, USA	
Q2-87	Wheat rhizosphere, Washington, USA	Take-all of wheat
SBW25	Sugar beet phyllosphere, Oxfordshire, UK	Seedling emergence
A506	Pear phyllosphere, California, USA	Fire blight of pear and apple, frost injury, fruit russeting
SS101	Wheat rhizosphere, The Netherlands	Diseases caused by *Pythium* spp. and *Phytophthora* spp.
***Pseudomonas* sp.:**		
BG33R	Peach rhizosphere, South Carolina, USA	The plant-parasitic nematode *Mesocriconema xenoplax*

Loper et al (2012) PLoS Genetics 8: e1002784

Steps in a typical genome sequencing project

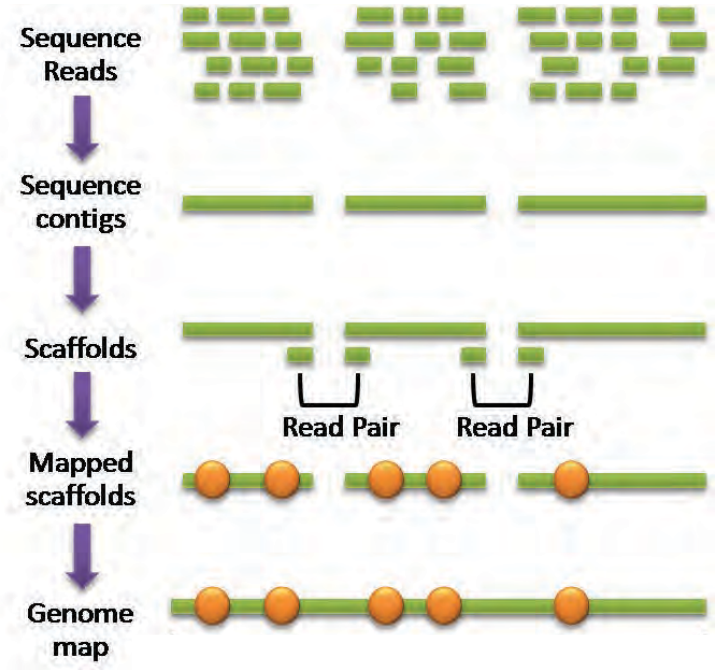
Sample preparation and sequencing

Finishing

- Assembly
- Gap filling
- Conflict resolution
- Verification

Analysis

- Gene predictions
- Homology searches
- Annotation



Base calling: reading the chromatograms

RESEARCH

Base-Calling of Automated Sequencer Traces Using *Phred.* I. Accuracy Assessment

Brent Ewing,[1] LaDeana Hillier,[2] Michael C. Wendl,[2] and Phil Green[1,3]

[1]Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA;
[2]Genome Sequencing Center, Washington University School of Medicine, Saint Louis, Missouri 63108 USA

Genome Research (1998) 8:175-185

- Developed by Phil Green (University of Washington, Seattle)
- De facto industry standard for base calling

Phred assigns probability value to each base

Phred q	Probability that the base is called wrong	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Draft vs finished genomes

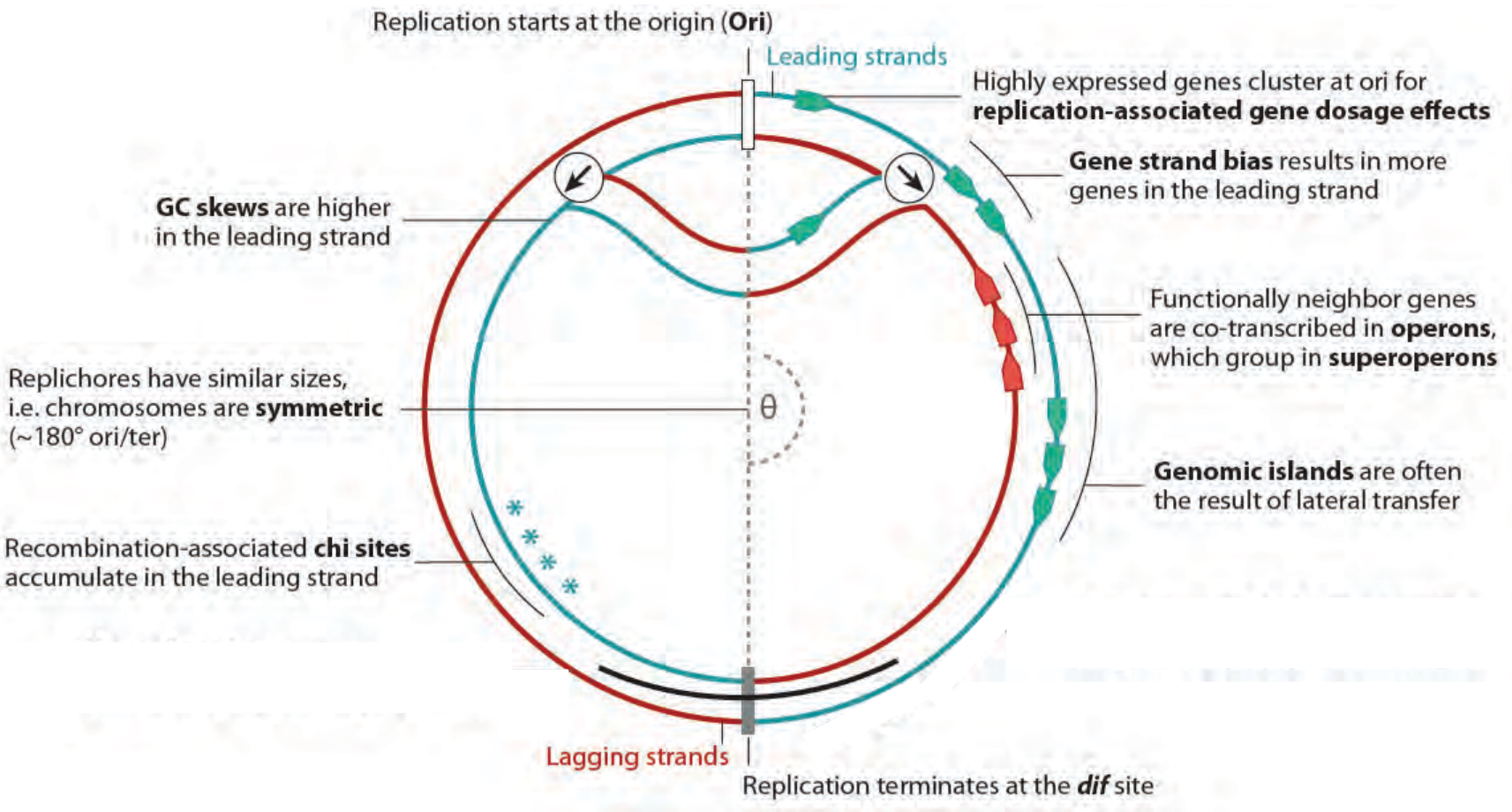
Finishing is the process of assembling and refining raw sequence data into a highly accurate final genomic sequence

- Automated sequence editing
- Manual, interactive sequence inspection
- Directed sequencing
- Assembly verification

Disadvantages: cost and time

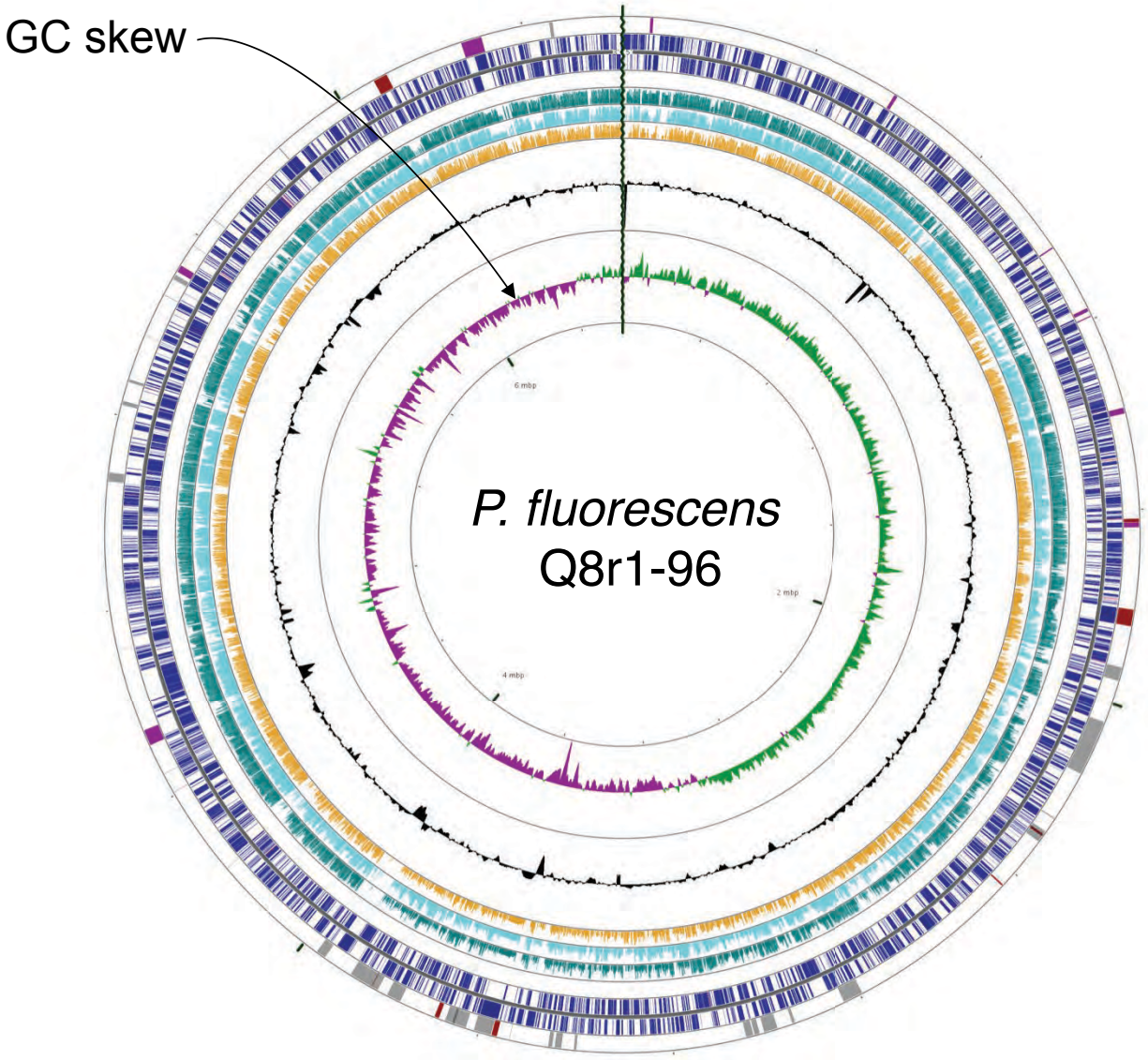
GC skew and scaffold order

%G+C composition (reflecting horizontal gene transfer) and GC skew (reflects strand-specific mutational bias)



Rocha (2008) Annu Rev Genet

GC skew and scaffold order



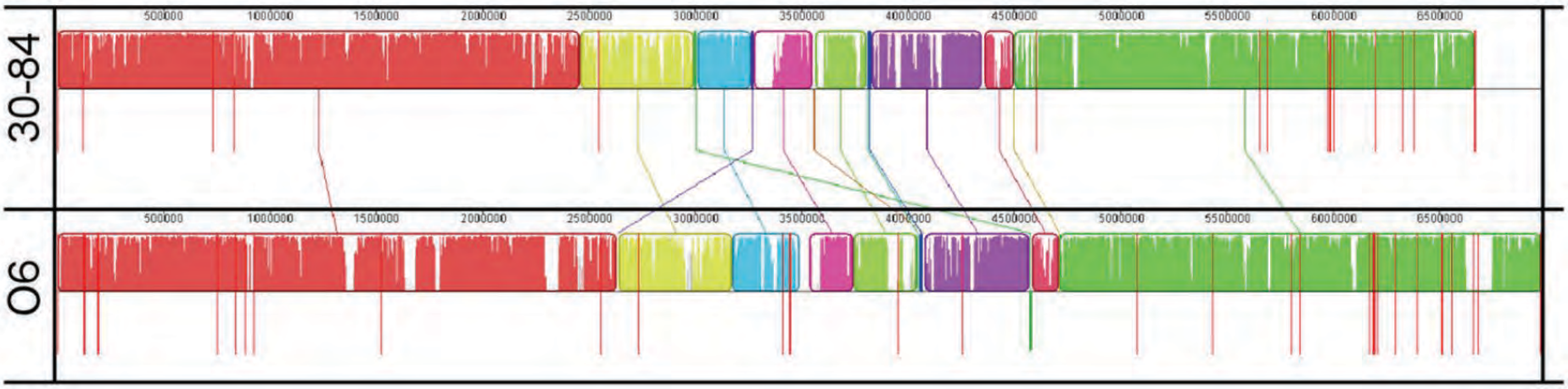
Gene synteny analyses

Finished genomes are useful for studying the preservation of the gene order (establishment of the orthology of genomic regions in different species; important functional relationships between genes; evolution)

Some whole-genome alignment tools:

MUAVE (http://gel.ahabs.wisc.edu/mauve/)

ACT (http://www.sanger.ac.uk/Software/ACT)



Plasmid gene synteny analysis

Draft vs finished genomes

Sequence is experimental data, subject to experimental error

Virtually any accuracy can be obtained, but the cost of a project and the accuracy determined are directly related

It makes sense to use an accuracy appropriate for the results to be obtained

Finding the genes in a genome and their relative location does not require extremely high accuracy

The goal of annotation is to locate genes (aka coding sequences or CDSs) and assign their putative functions

Open reading frames are predicted by length of by using interpolated Markov models (Glimmer)

The proteins encoded by each putative gene are compared by BLAST searches against closely related proteomes and the NCBI databases

To improve gene function prediction, protein motifs are assigned (TIGRFAM, Pfam and COG databases)

BLAST scores and other evidence for function are compared to choose CDSs most likely to represent functional genes

Automated annotation

All analysis step can be assembled into annotation "pipeline" and performed without human intervention

Some free online automatic annotation engines:

RAST (http://rast.nmpdr.org)

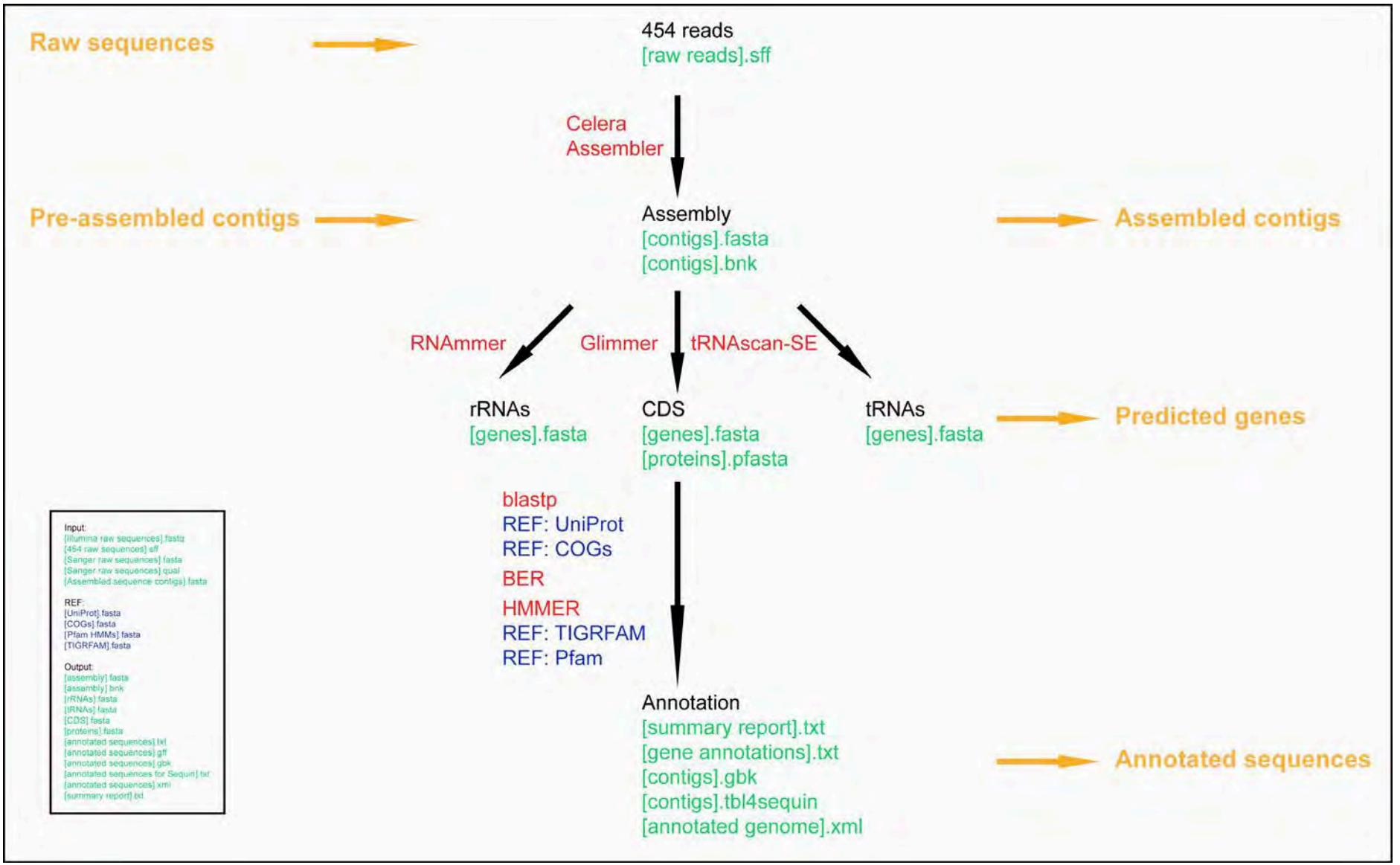
BASYS (http://basys.ca)

IGS Annotation Engine (http://ae.igs.umaryland.edu)

The Prokaryotic Genome Annotation Pipeline (PGAAP) (http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html)

DOE-JGI Microbial Annotation Pipeline (https://img.jgi.doe.gov)

CloVR virtual machines



Community annotation

Automated annotation often followed by “community annotation”, when a community of researchers integrates computer predictions with experimentally validated data

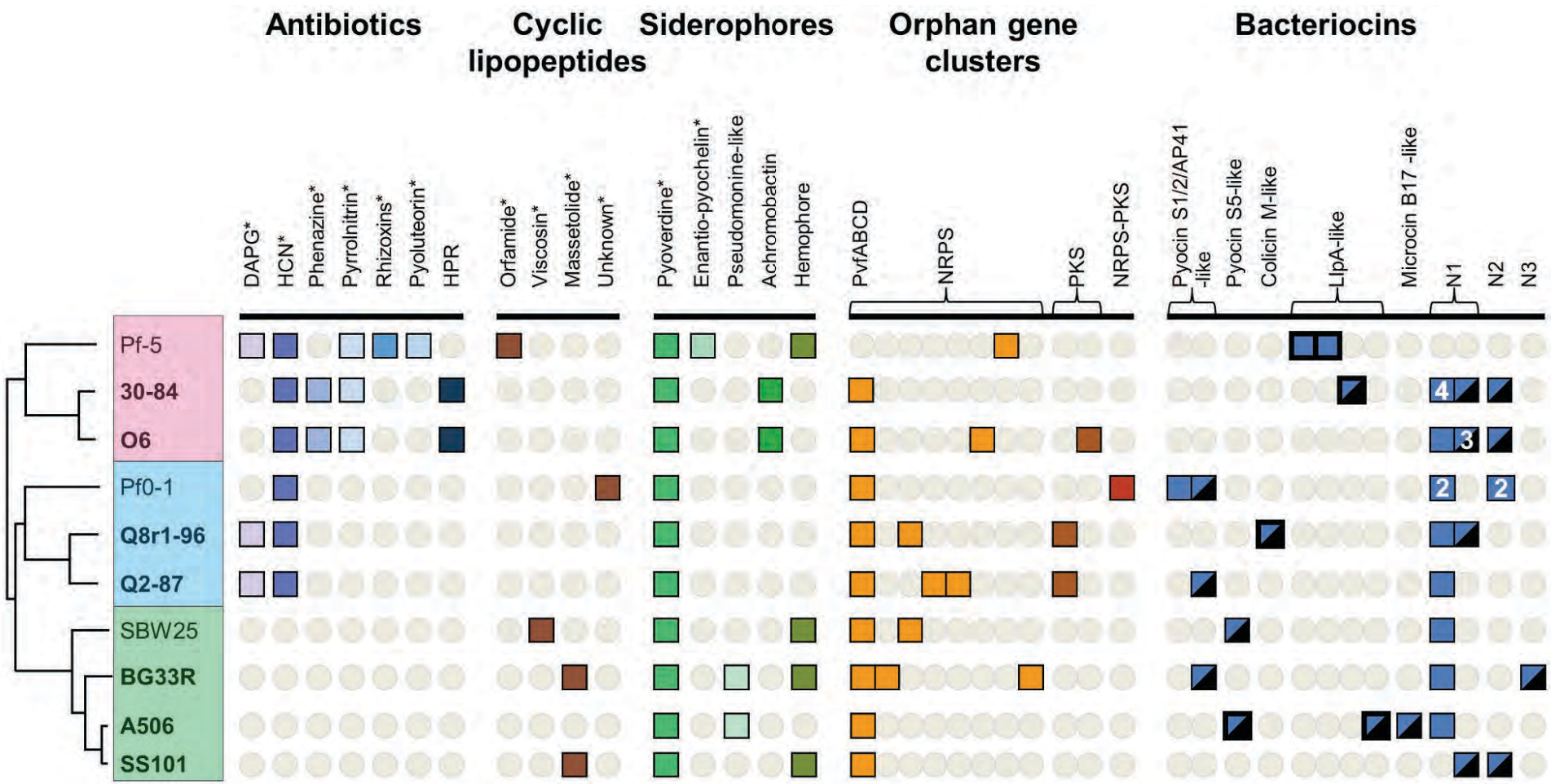
Some tools for integrating manual and automated annotations:

Artemis (http://www.sanger.ac.uk/resources/software/artemis)

Manatee (http://manatee.sourceforge.net/)

Apollo (http://apollo.berkeleybop.org)

Traits of potential importance for plant-microbe interactions and biological control



Loper et al (2012) PLoS Genetics 8: e1002784

Mining individual genomes

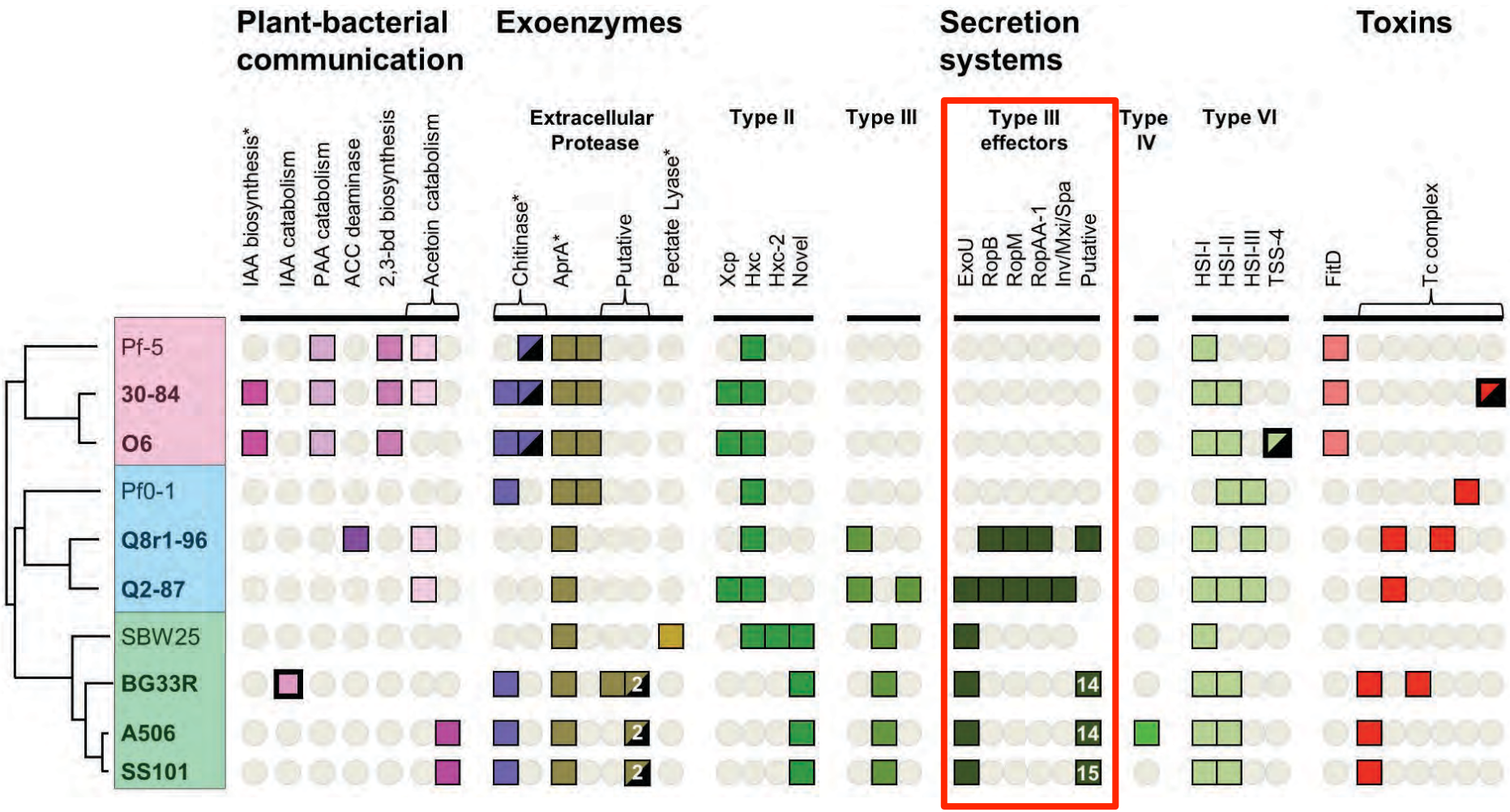
Keyword searches

BLAST searches is the most common way of mining a genome (online vs local; nr vs specialized databases)

BLAST searches alone are often not sufficient

T3 effectors is a good example – we need to look at motifs (a stretch of protein with conserved amino acids in conserved positions), patterns (certain amino acids occur more frequently), regulatory elements (hrp boxes) and neighboring genes

Traits of potential importance for plant-microbe interactions and biological control (contd.)



Loper et al (2012) PLoS Genetics 8: e1002784

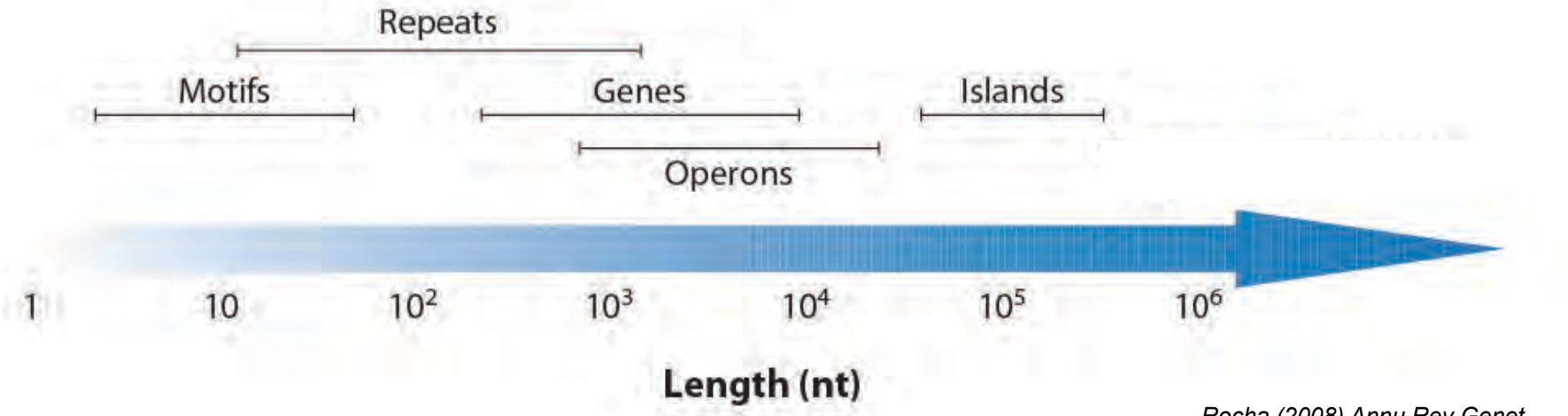
Community annotation

Other interesting features: repeats, operons, islands

Some free online automatic annotation engines:

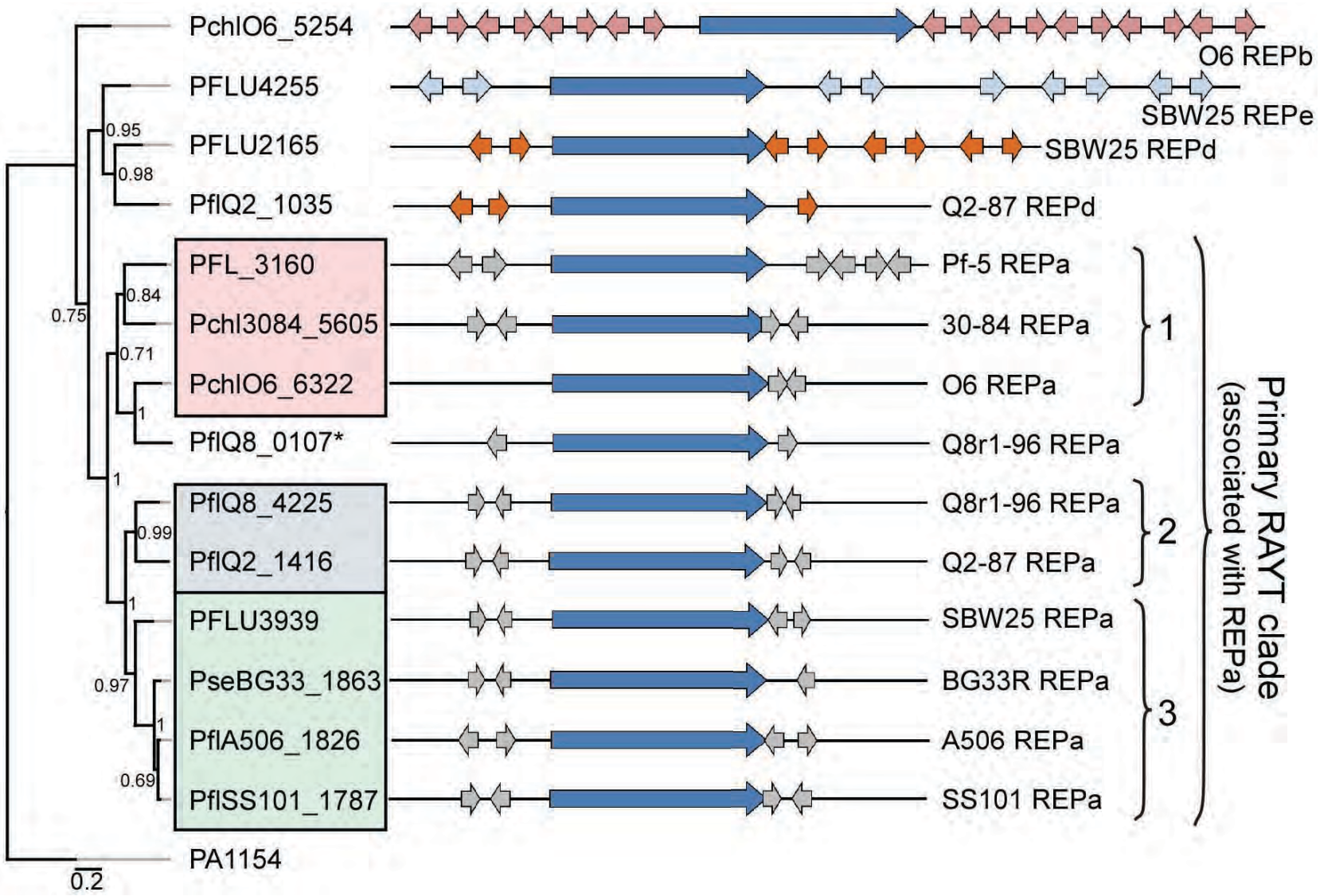
RepeatScout (http://bix.ucsd.edu/repeatscout/)

RepeatFinder (http://cbcb.umd.edu/software/RepeatFinder/)

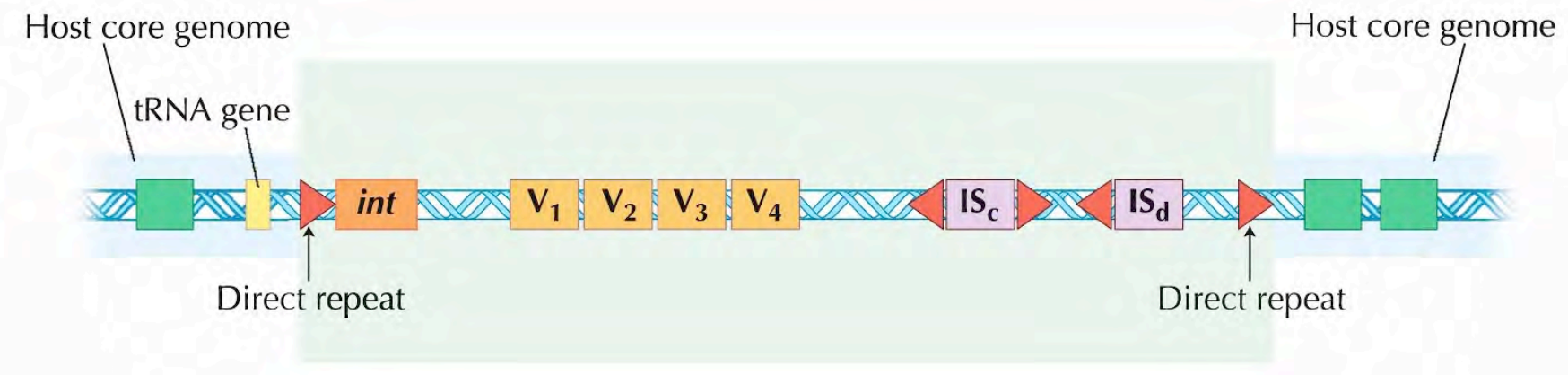


Rocha (2008) Annu Rev Genet

REP elements



Anomalous regions: genomic islands and mobile elements

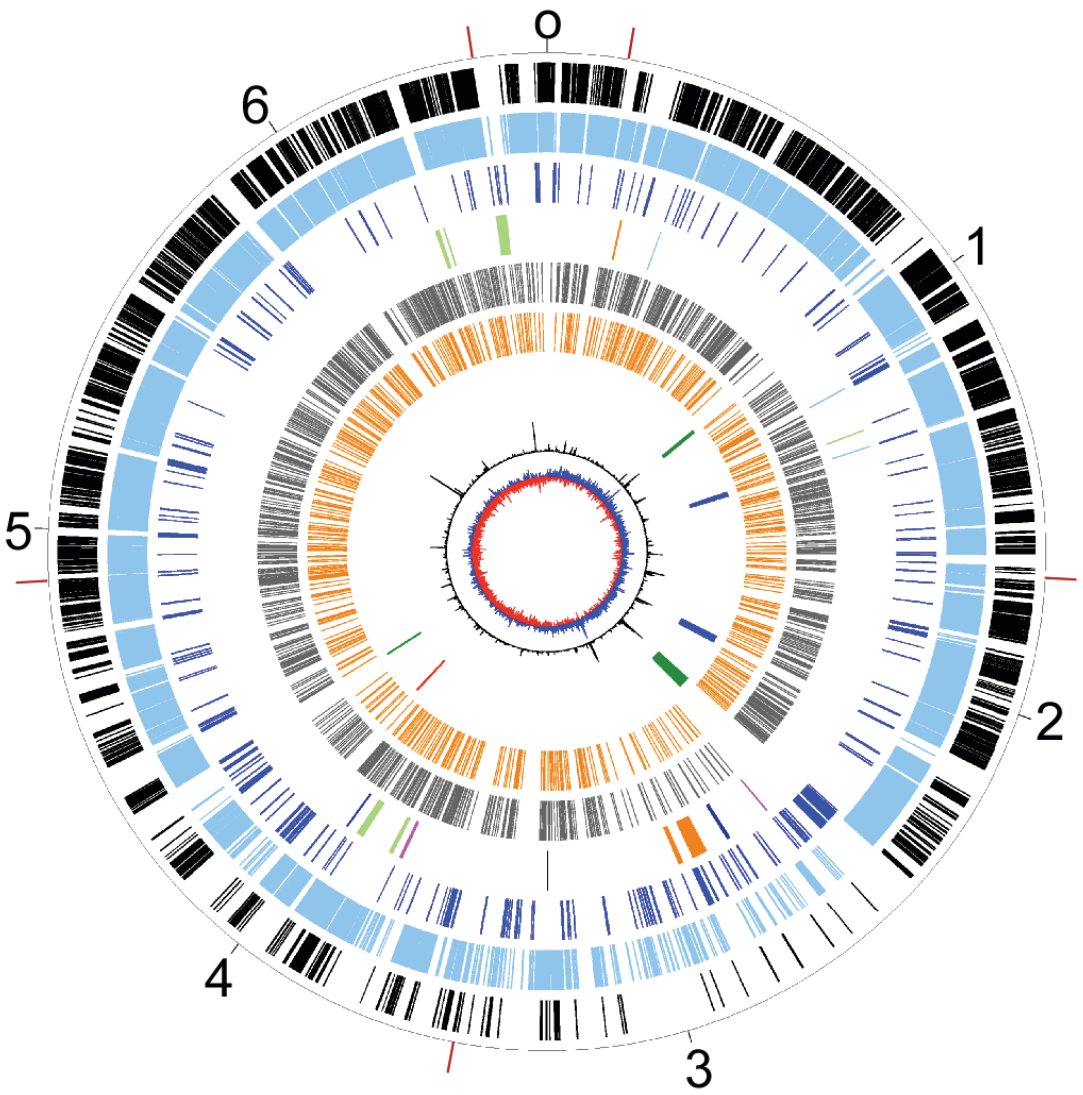


Genomic islands differentially distributed in the genomes of related strains

Often integrate into tRNAs and, in addition to cargo genes, carry mobility genes (integrases, insertion elements) and have atypical G+C content

Core and lineage-specific regions in the sequenced genomes

Phl^{+} strain *Pseudomonas fluorescens* Q8r1-96



- core *Pseudomonas* genes
- lineage-specific genes
- strain-specific genes
- REPa elements
- REPb elements

Cross-genome comparisons

All-against-all BLAST analyses to look at the presence/absence of genes

Helps to define core and pan genomes

Helps to find differences in gene content between organisms and identify candidate genes that may determine the observed phenotypic variations

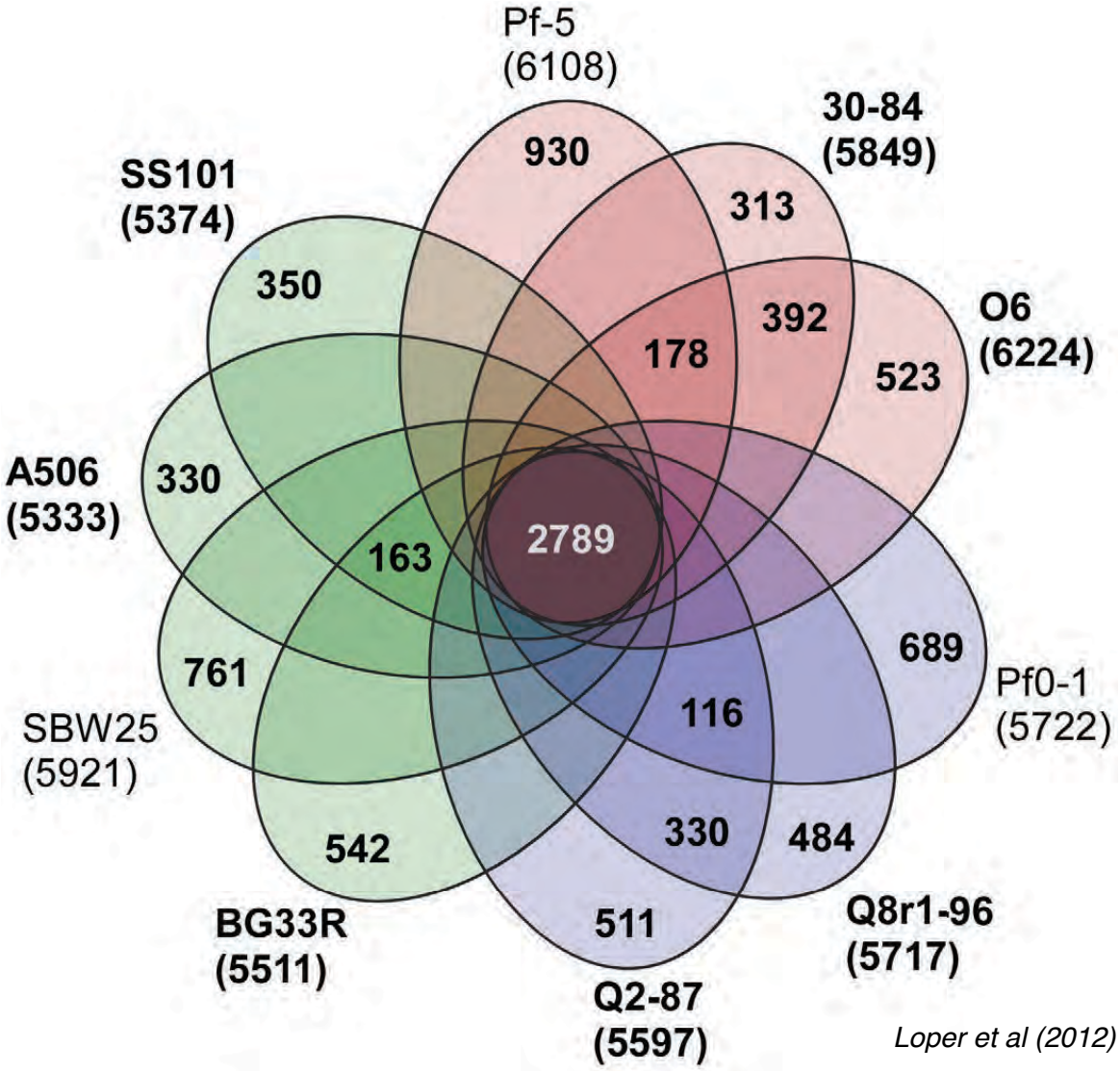
Presence/absence of genomic regions:

MUAVE (http://gel.ahabs.wisc.edu/mauve/)

ACT (http://www.sanger.ac.uk/Software/ACT)

Proteome conservation between strains from the *P. fluorescens* group

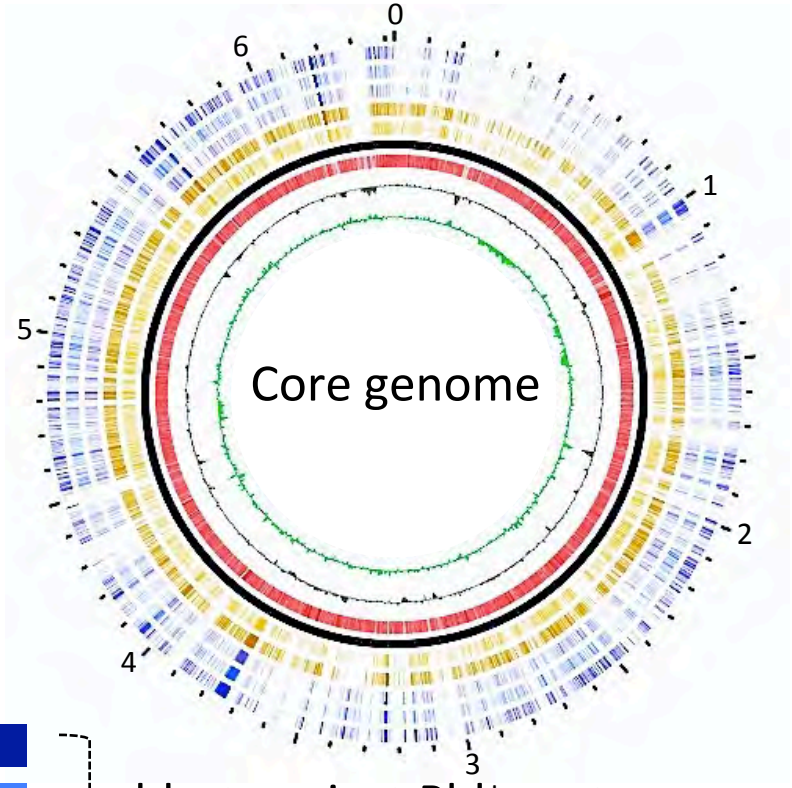
(blastp e-value cutoff = 10^{-30})

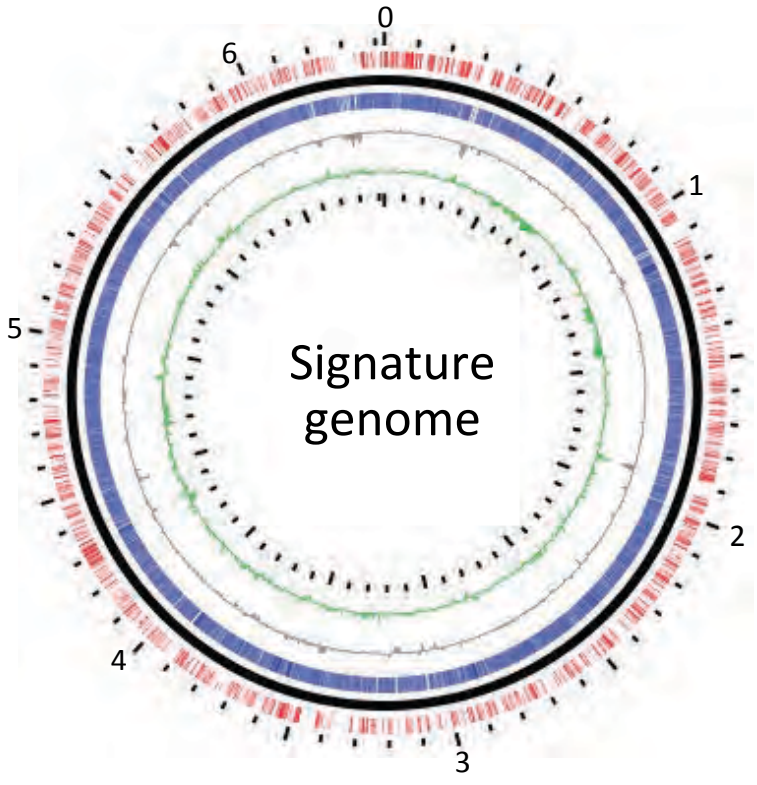


Loper et al (2012) PLoS Genetics 8: e1002784

Core and signature genome analyses

Core and signature genome comparisons of three Phz$^+$ and three Phl$^+$ *Pseudomonas* spp.





Conclusions I

- Thousands of bacterial genomes including those of plant pathogenic and biocontrol bacteria have been sequenced
- These genomes represent an invaluable resource for molecular plant pathologists and plant microbiologists
- Many different approaches can be used for mining bacterial genomes - one by one or through powerful comparative analyses
- There are numerous free and easy-to-use web databases and tools that can be utilized for genome mining