Next Generation Sequencing (NGS) Data Analysis for the Oral Microbiome

Keeping an Eye on Tomorrow's Genomics Technology: Oralgen 2.0

www.lanl.gov/bioscience
www.oralgen.lanl.gov

Patrick Chain
Metagenomics Applications Team
Los Alamos National Lab (LANL)

IADR, Barcelona, Spain; July 2010

Genomics and Bioinformatics at LANL

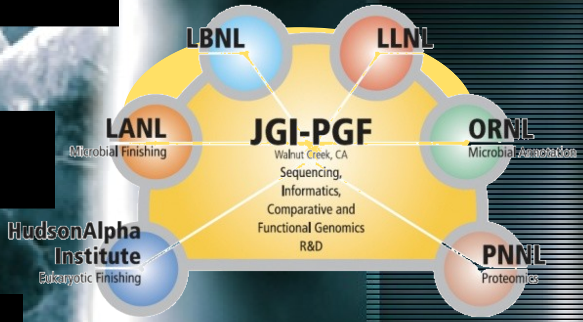
1982 GenBank

1988 The Los Alamos Center for Human Genome Studies

1990 HIV Sequence Database

1995 Influenza Sequence Database

1997 Joint Genome Institute

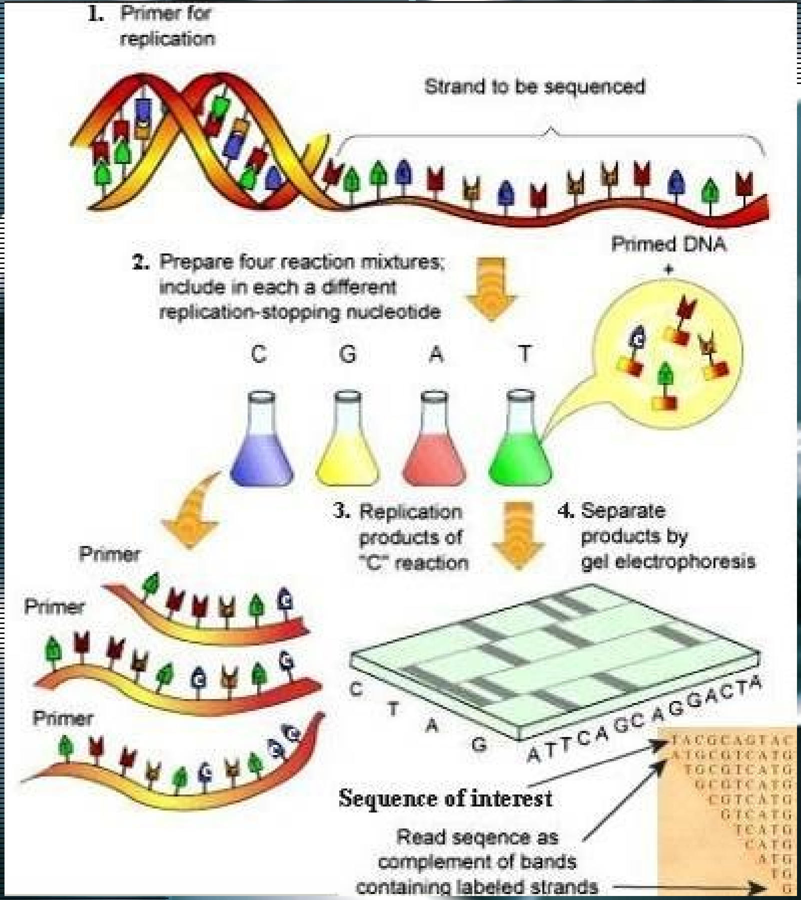


1998 STD Sequence Database

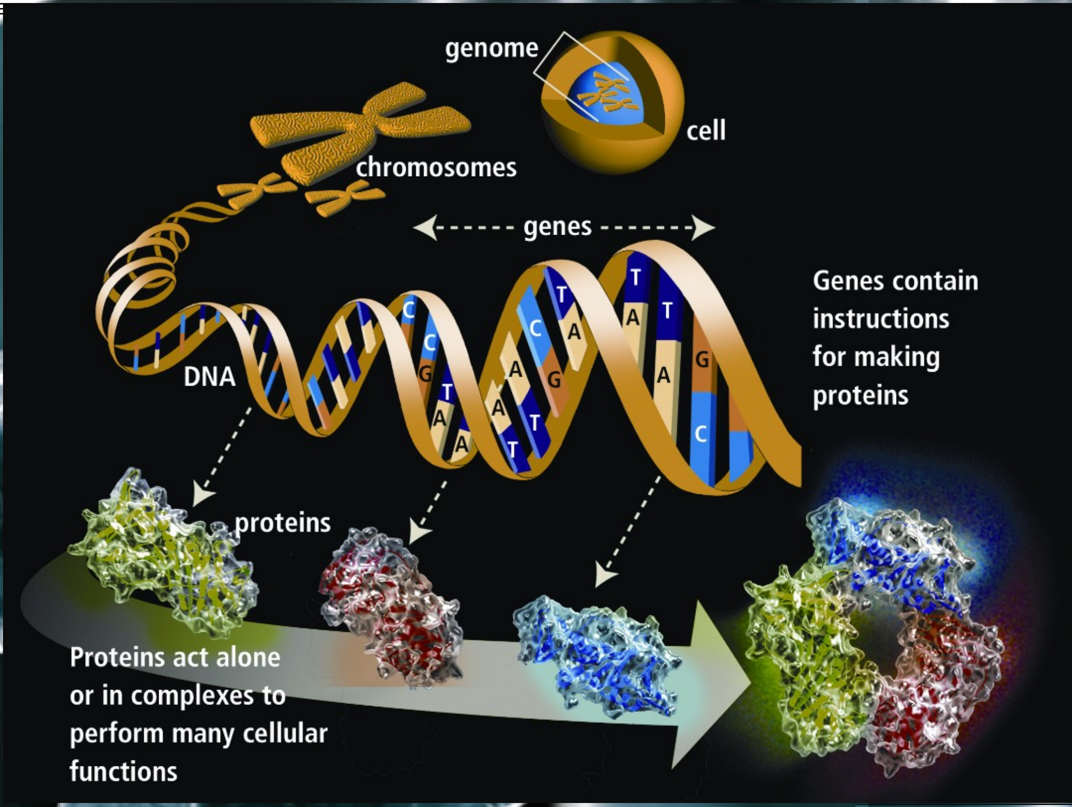
1999 Oral Genomics and Metagenomics Resource

www.oralgen.lanl.gov

DNA sequencing and the Birth of Genomics



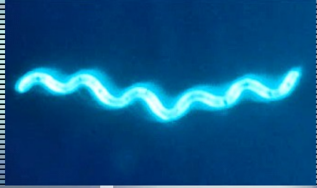
1. Primer for replication
Strand to be sequenced
2. Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
Primed DNA
+
C
G
A
T
Primer
Primer
Primer
3. Replication products of "C" reaction
4. Separate products by gel electrophoresis
C
T
A
G
ATTCAGCAGGACTA
TACGCAGTAC
ATGCGTCATG
TGCGTCATG
GCGTCATG
CGTCATG
GTCATG
TCATG
CATG
ATG
TG
G
Sequence of interest
Read seqence as complement of bands containing labeled strands

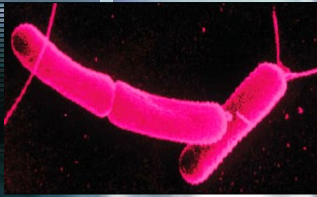


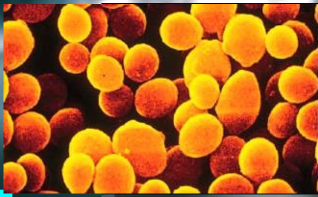
genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions

The revolution: obtain the genetic basis for all functions of the organism

Initial Targets



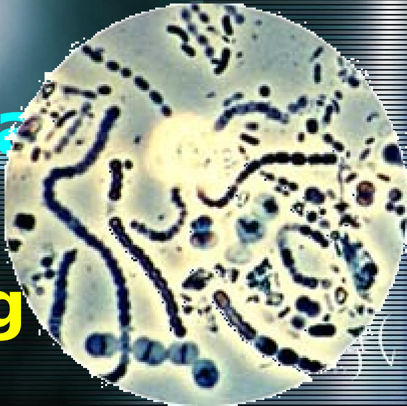




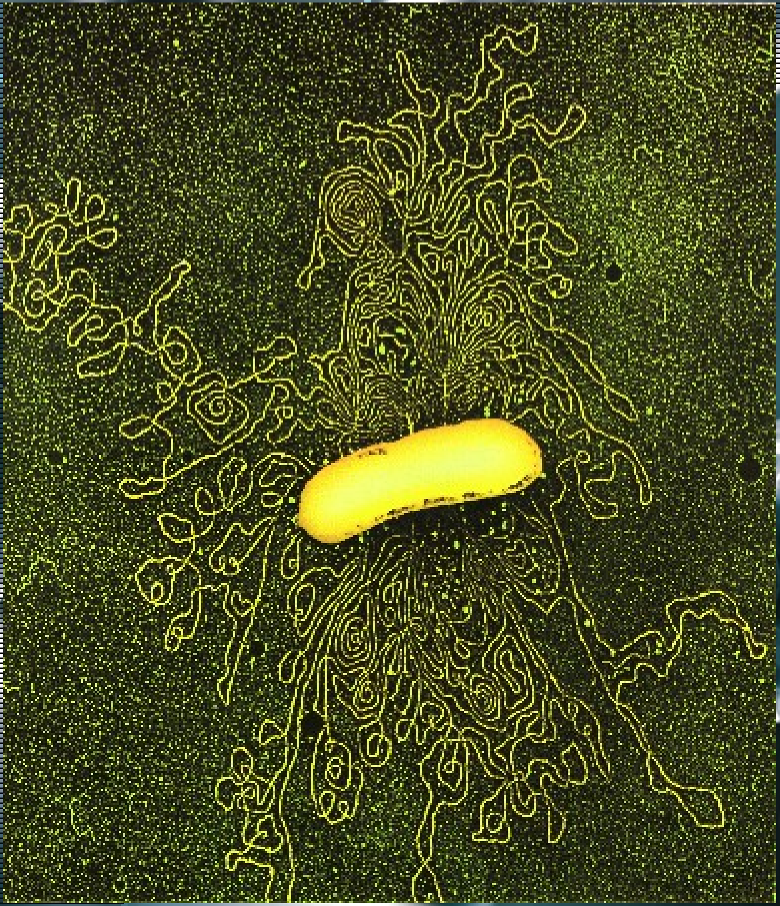
- Bacterial Human Pathogens
  - Public Health
  - Biodefense

- Important Environmental Bacteria and Archaea
  - Carbon sequestration and climate change
  - Bioremediation
  - Bioenergy

- Important Agricultural Bacteria
  - Crop pathogens
  - Important plant growth promoting

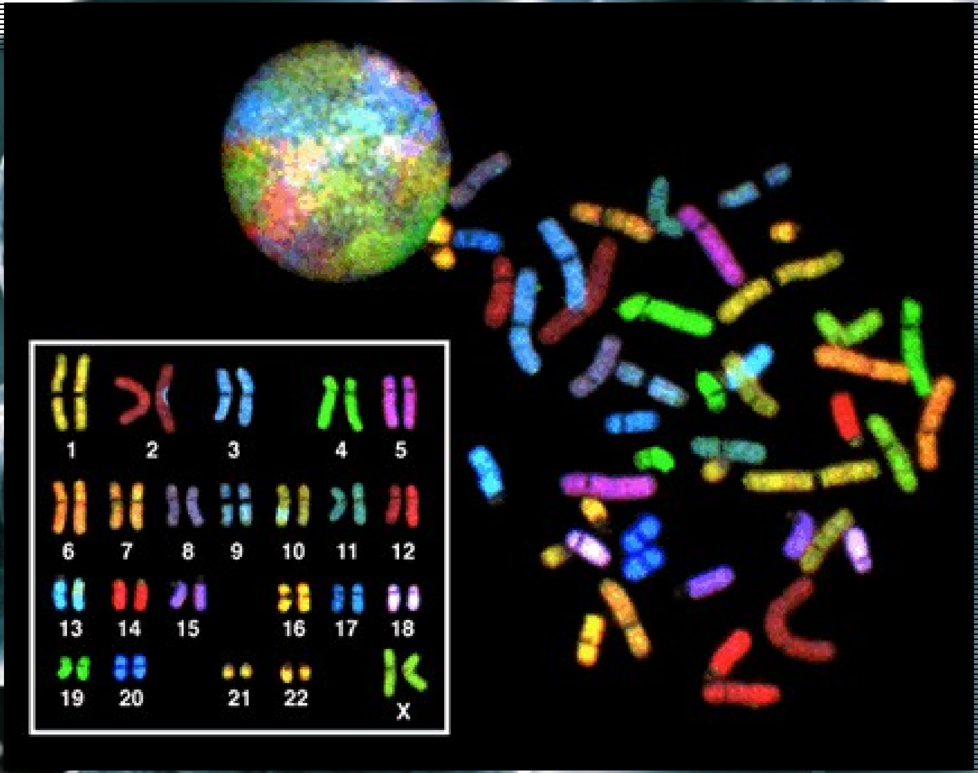


The Genomics Revolution



E. coli: 4.5 million base pairs

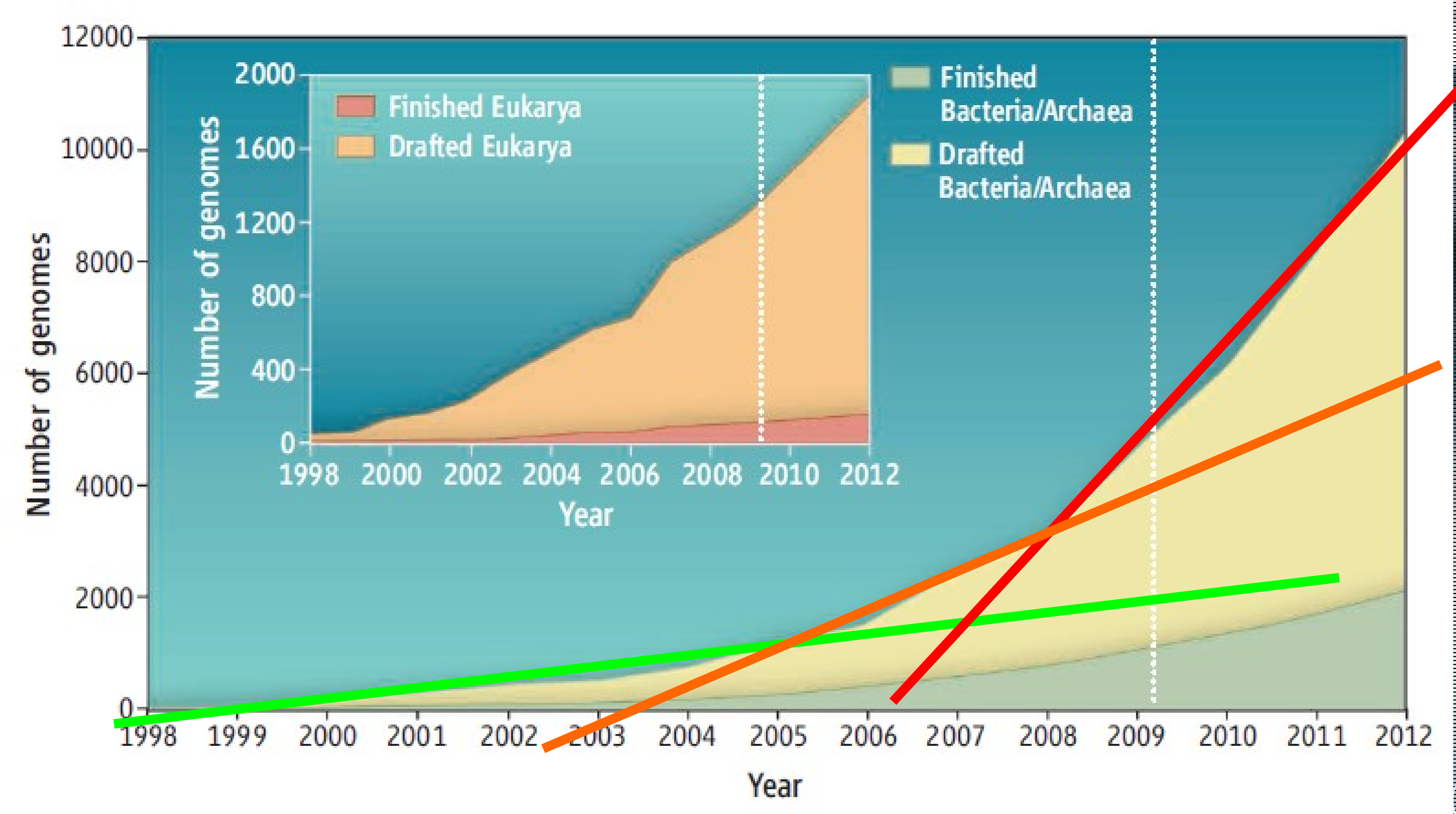
Cost: ~$3M
(1997)



H. sapiens: 3000 million base pairs

Cost: > $300M
(2003)

Behind the Surge?: NGS Technologies



When "more" isn't just more!

- Sanger - 1975
- ABI gel "automated" - 1986
- ABI Capillary 1999 – current

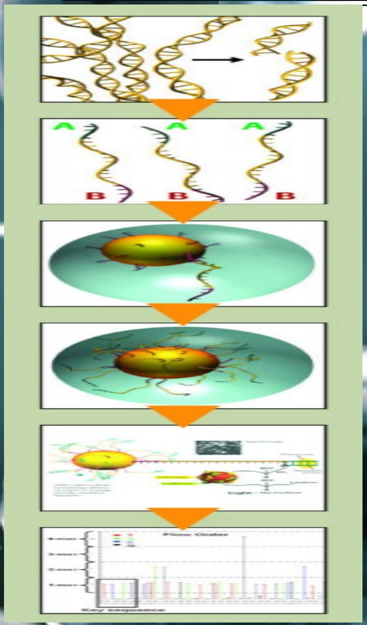
Capillary Based Sequencer, 70 kb / run

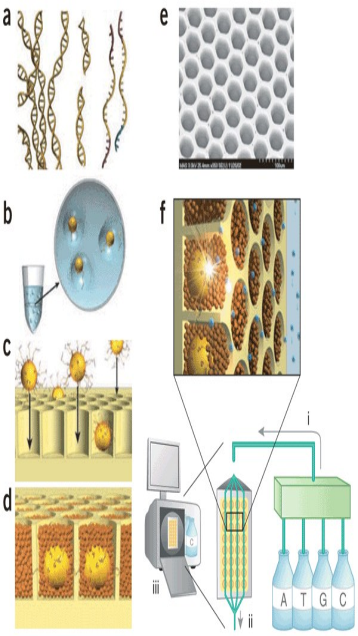
2005

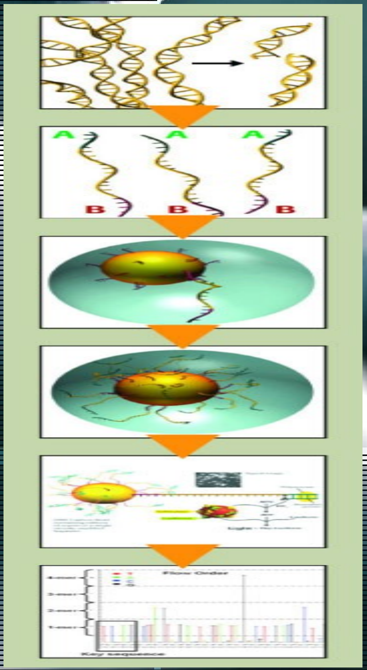
½ day

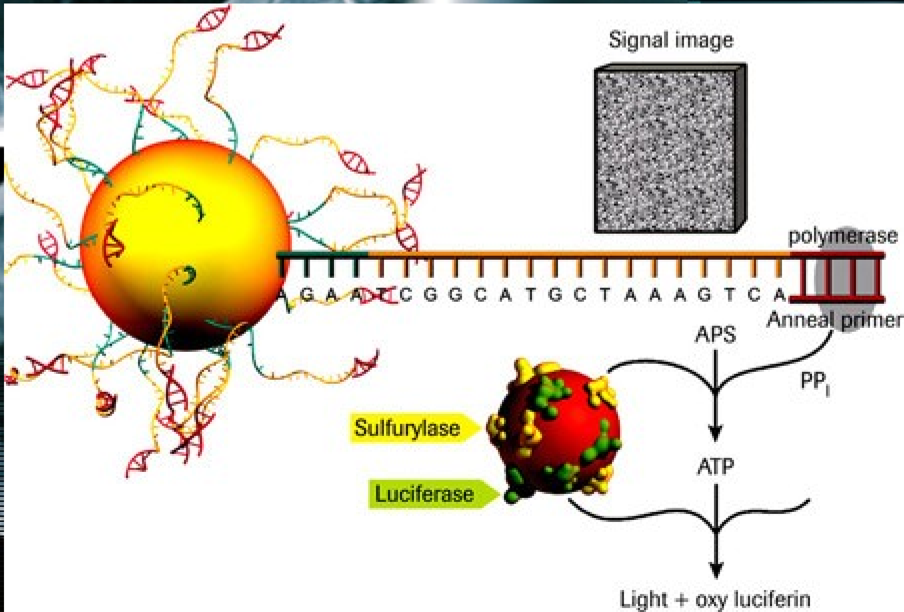
30 > 100 > 400 mb / run
100bp > 250bp > 400bp
Pyrosequencing

What is 454 pyrosequencing?

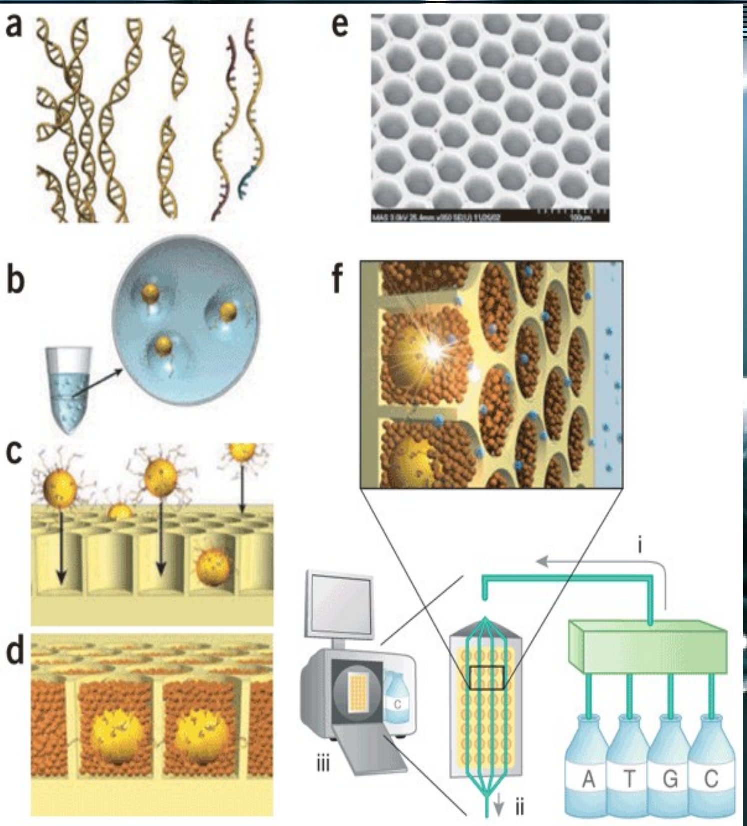


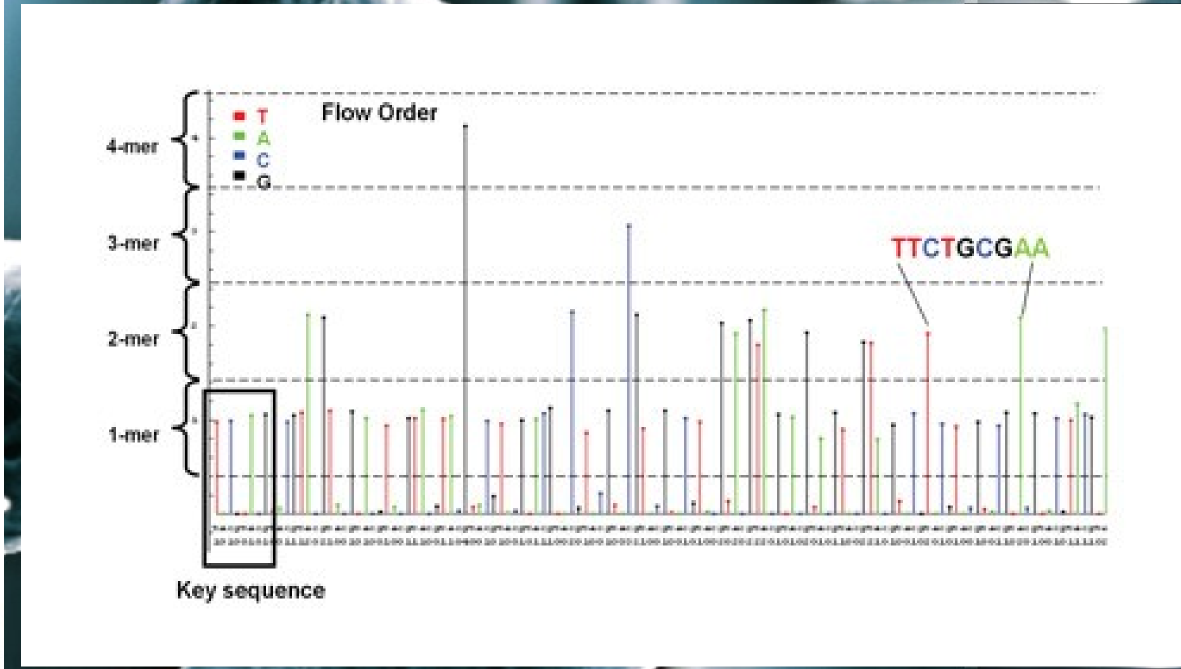






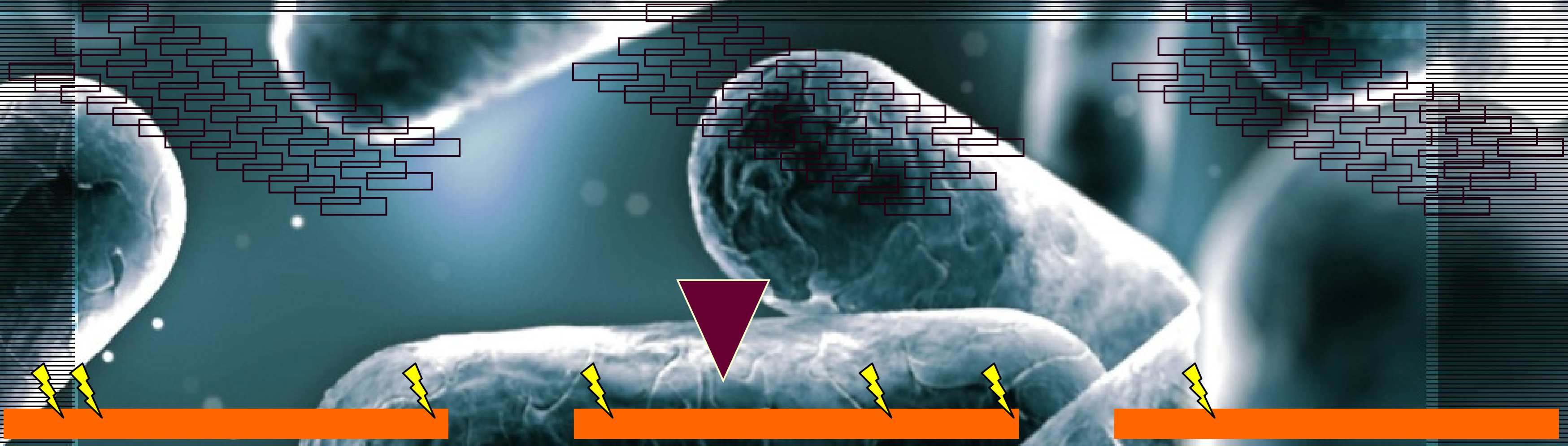
Reading light





Genome Assembly with Newbler

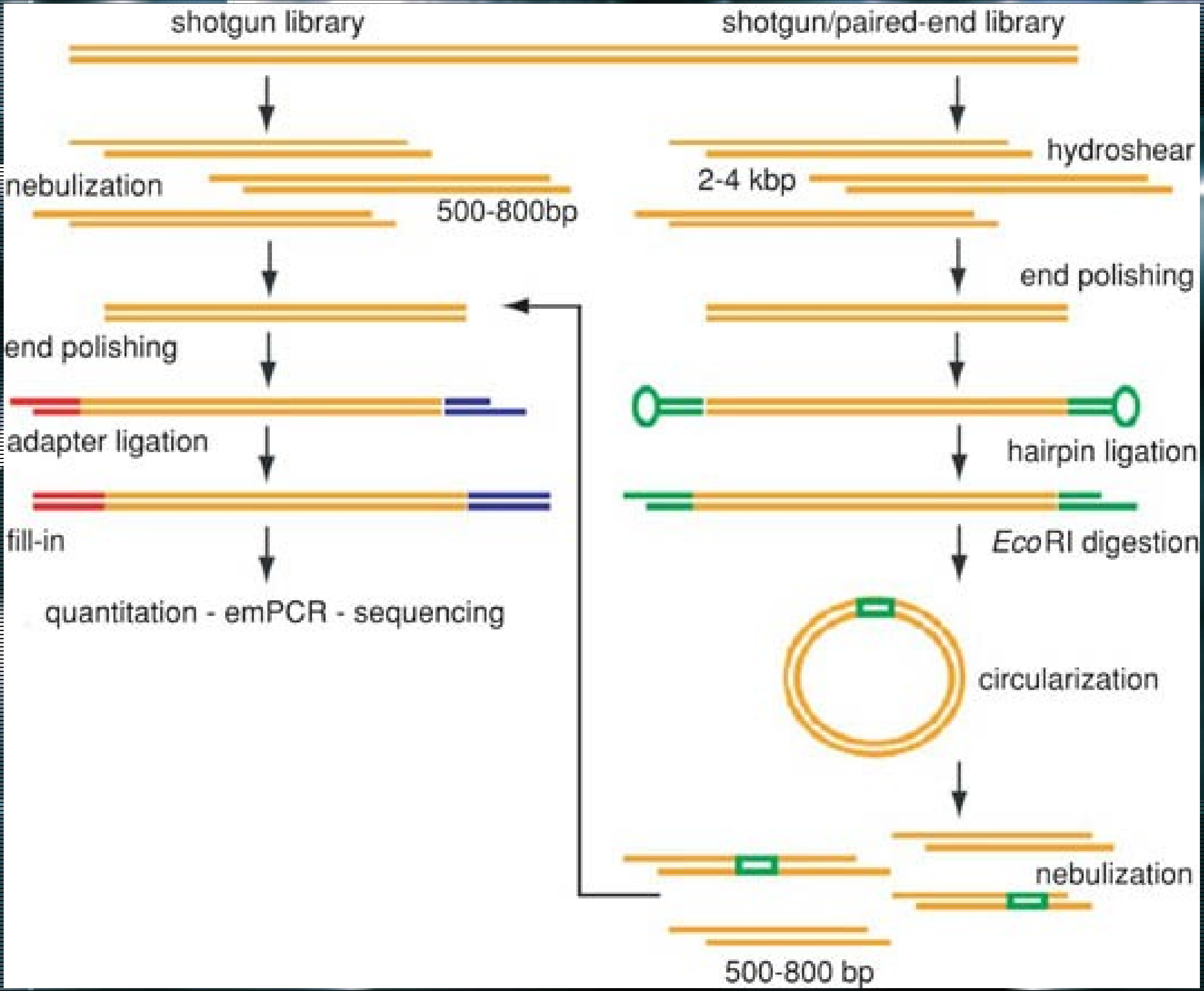
Random shotgun short reads – but many of them!



Contigs (consensus sequences)

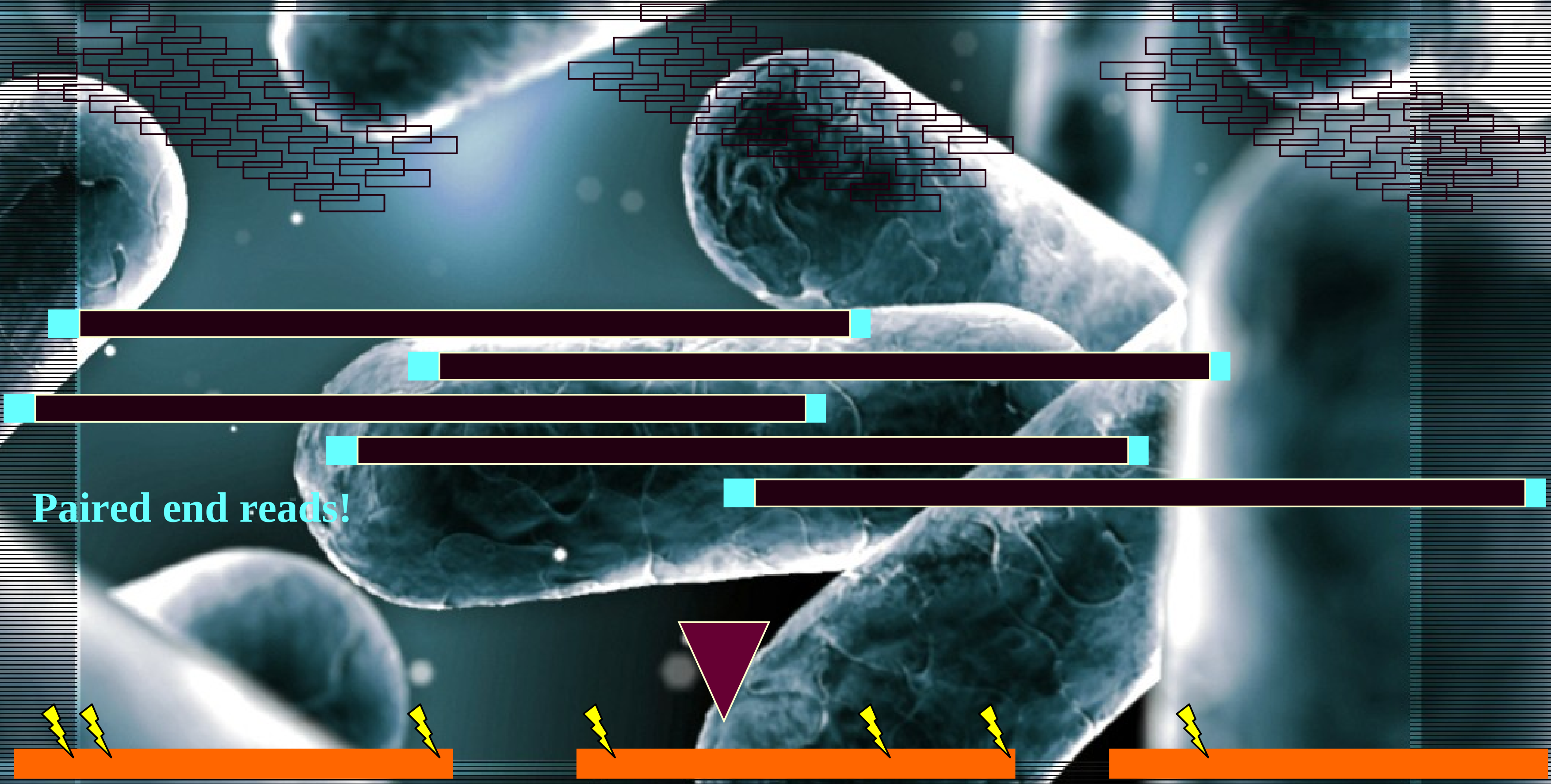
Need scaffolding information!!

Getting "paired-end reads"



Genome PE Assembly with Newbler

Random shotgun short reads – but many of them!



Paired end reads!

Contigs (consensus sequences)

When “more” isn’t just more!

- Sanger - 1975
- ABI gel “automated” - 1986
- ABI Capillary 1999 – current

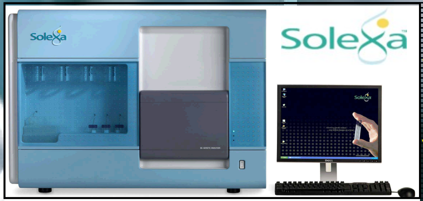
Capillary Based Sequencer, 70 kb / run

2005

½ day

30 > 100 > 400 mb / run
100bp > 250bp > 400bp
Pyrosequencing

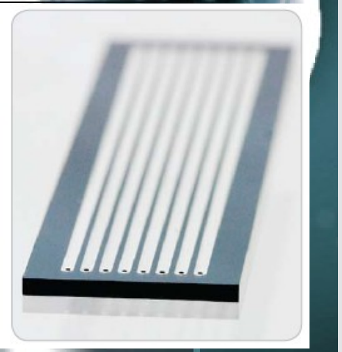
2007

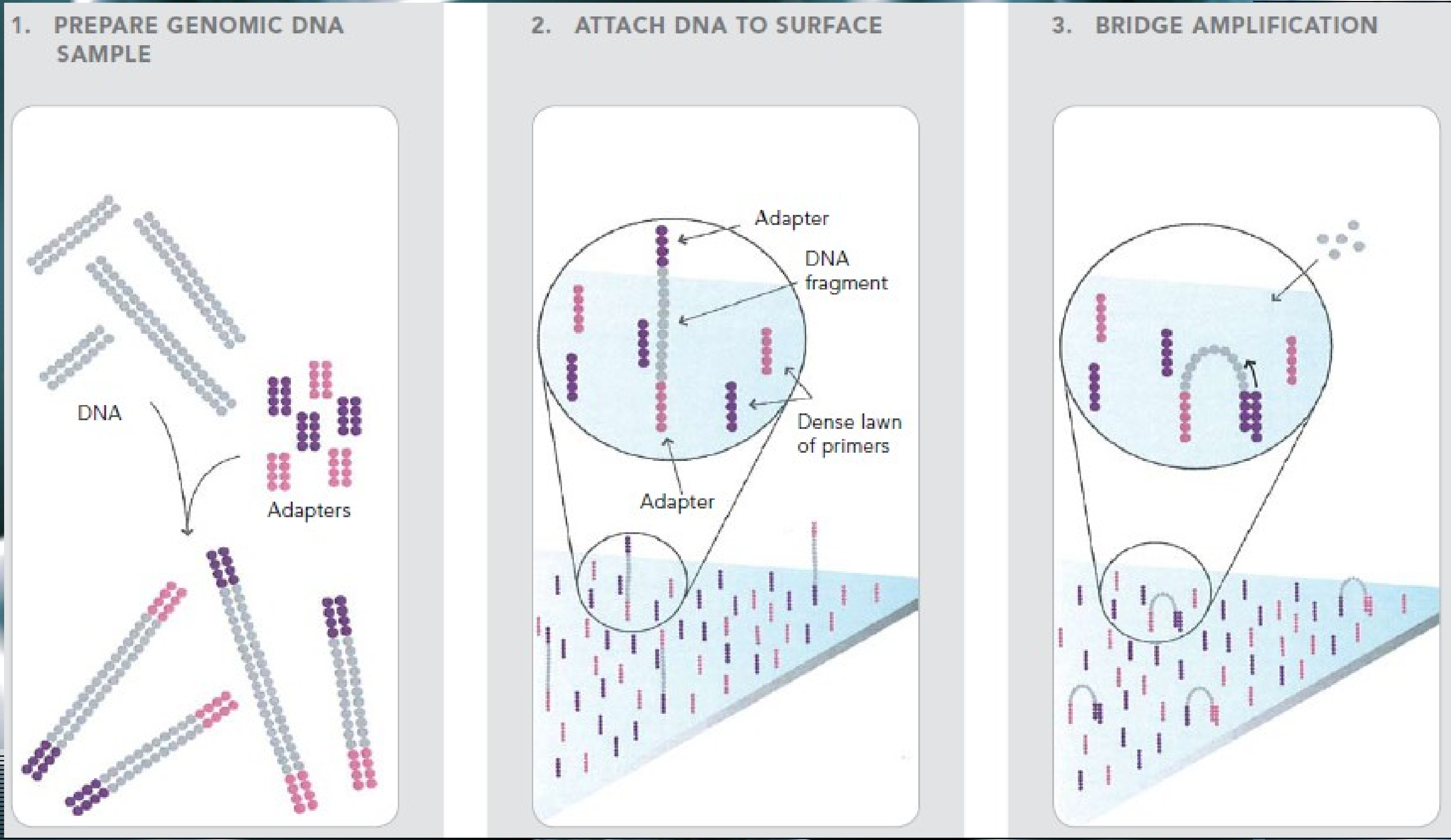


3 days

1.0 > 3 > 10 gb / run
25bp > 50bp > 75bp
Seq. by Synthesis

What is Solexa/Illumina sequencing?



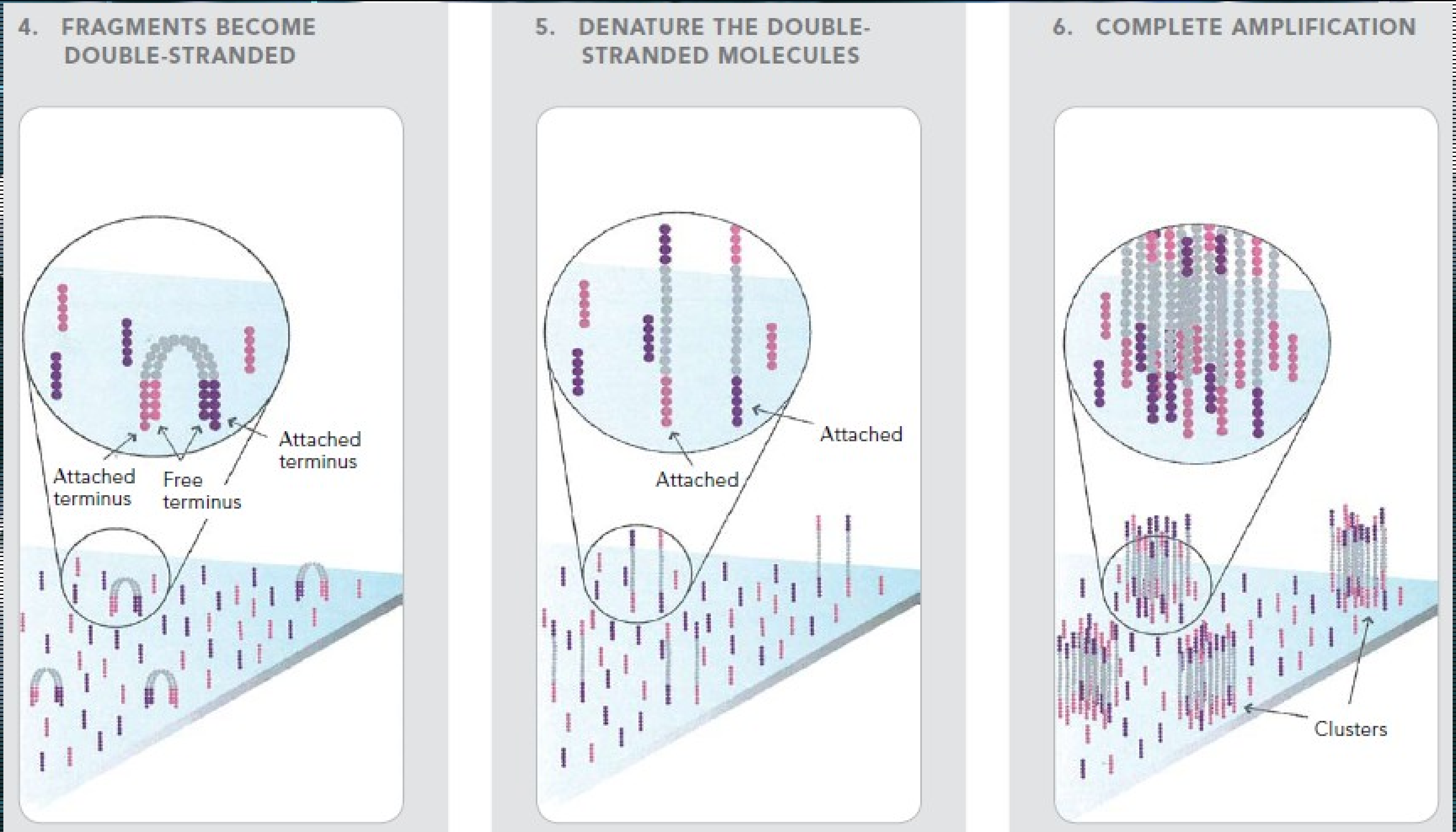


1. PREPARE GENOMIC DNA SAMPLE

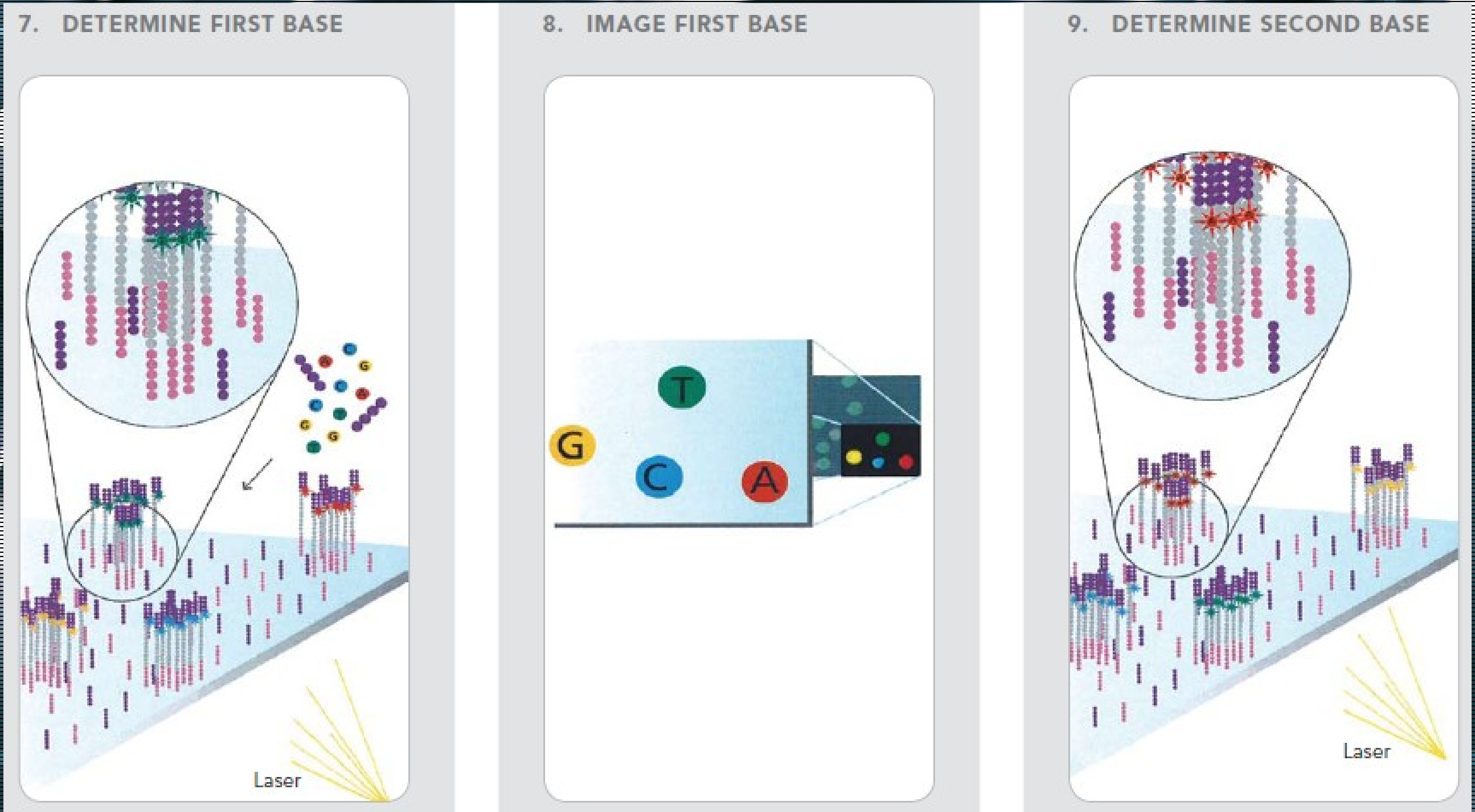
2. ATTACH DNA TO SURFACE

3. BRIDGE AMPLIFICATION

What is Solexa/Illumina sequencing?

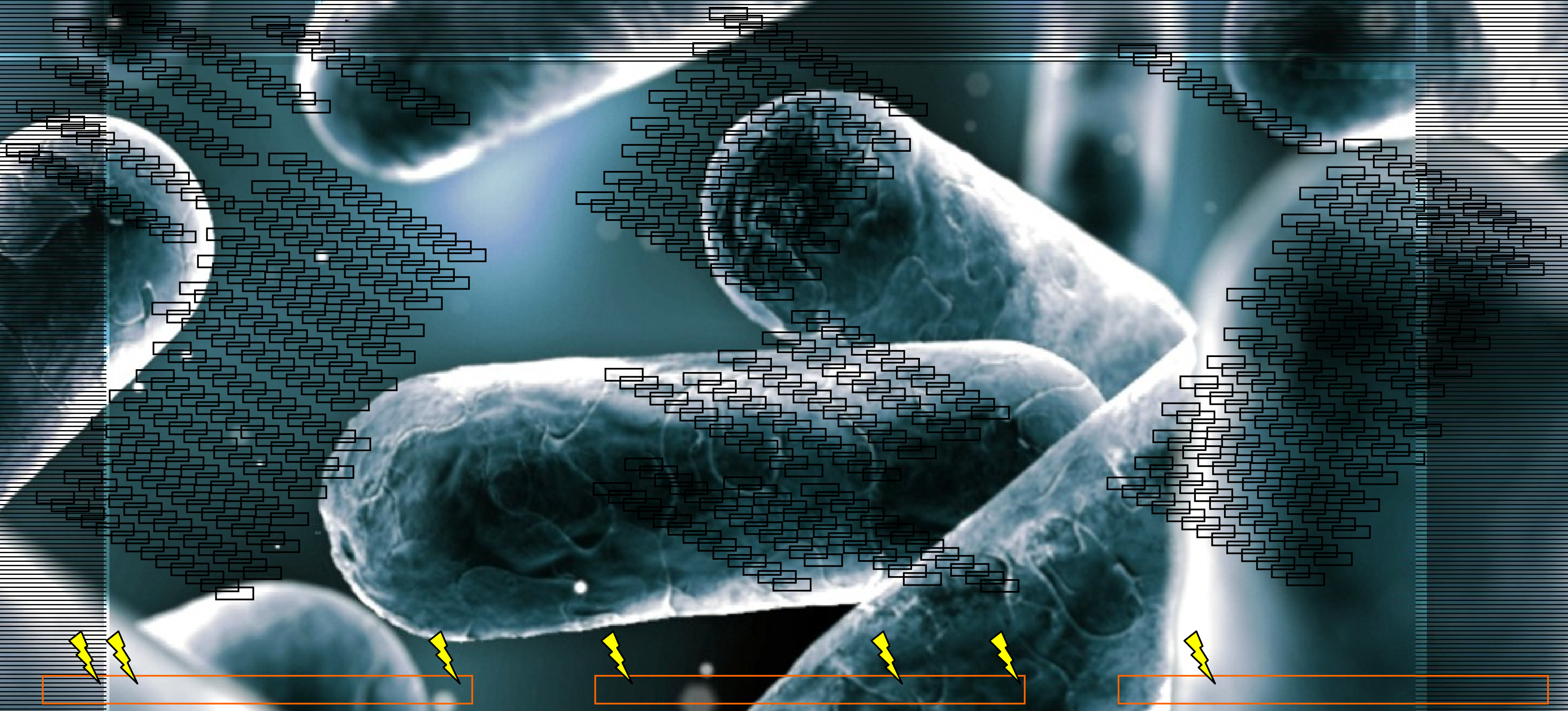

4. FRAGMENTS BECOME DOUBLE-STRANDED
Attached terminus
Free terminus
Attached terminus
5. DENATURE THE DOUBLE-STRANDED MOLECULES
Attached
Attached
6. COMPLETE AMPLIFICATION
Clusters

What is Solexa/Illumina sequencing?

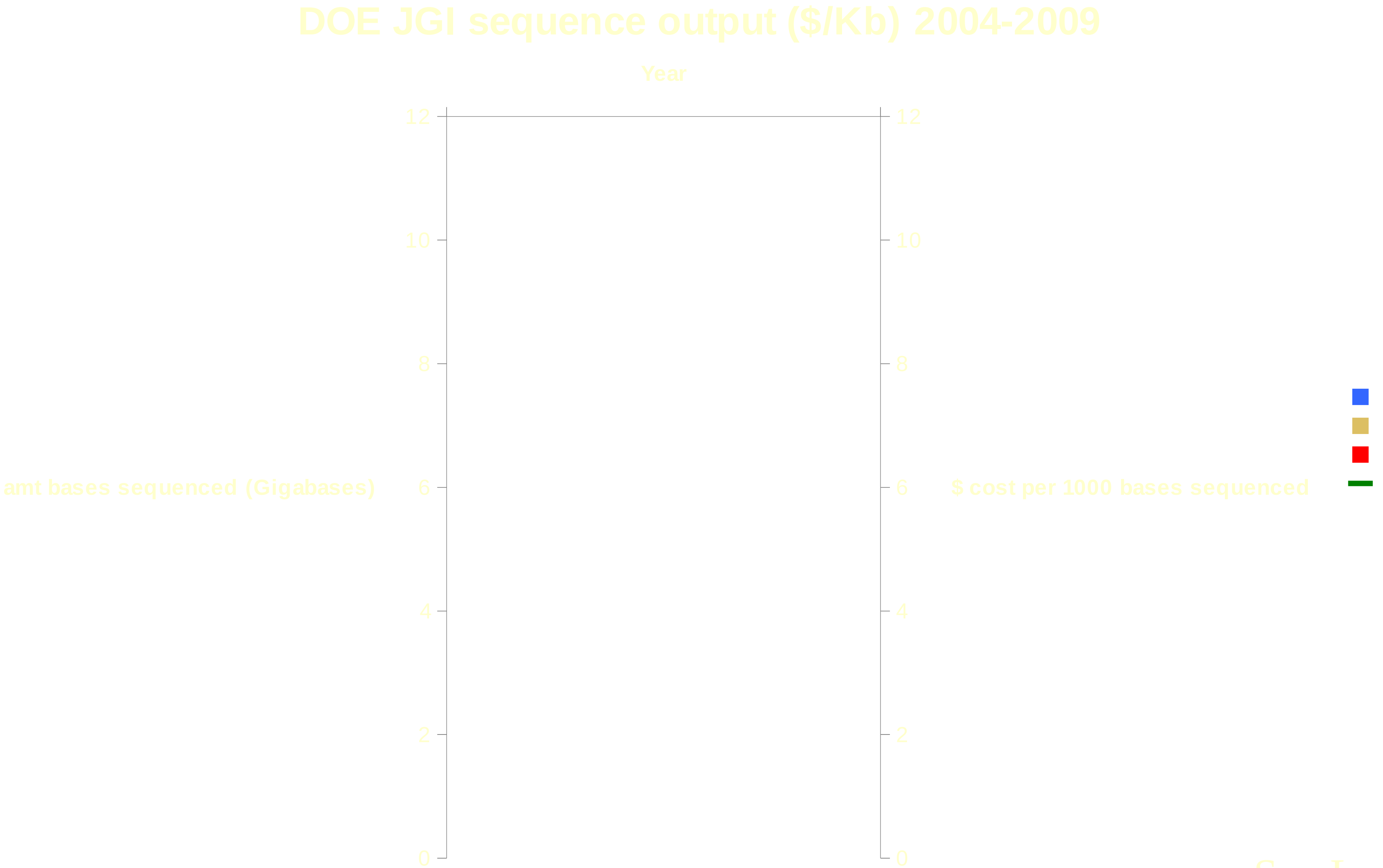


Get millions of clusters per lane/channel

Genome improvement/polishing



**Corrected sequences, but still many gaps generally


DOE JGI sequence output ($/Kb) 2004-2009
Year
12
10
8
6
4
2
0
12
10
8
6
4
2
0
amt bases sequenced (Gigabases)
$ cost per 1000 bases sequenced
🟦
🟨
🟥
▬

Sue Lucas

When “more” isn’t just more!

- Sanger - 1975
- ABI gel “automated” - 1986
- ABI Capillary 1999 – current

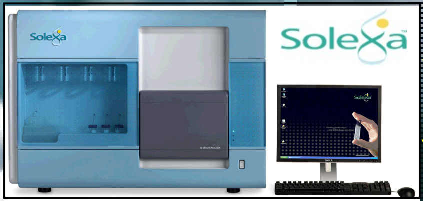
Capillary Based Sequencer, 70 kb / run

2005

½ day

30 > 100 > 400 mb / run
100bp > 250bp > 400bp
Pyrosequencing

2007



3 days

1.0 > 3 > 10 gb / run
25bp > 50bp > 75bp
Seq. by Synthesis

95G kits
HiSeq2000 : 200G

2008

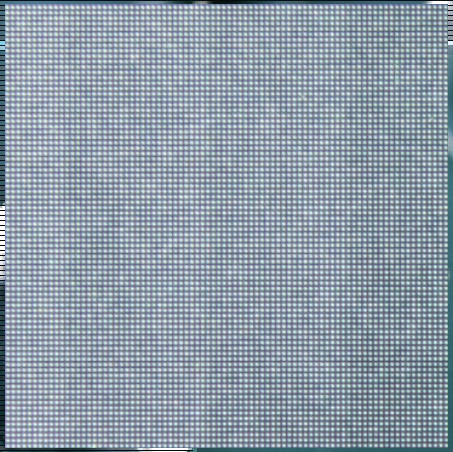
7 days

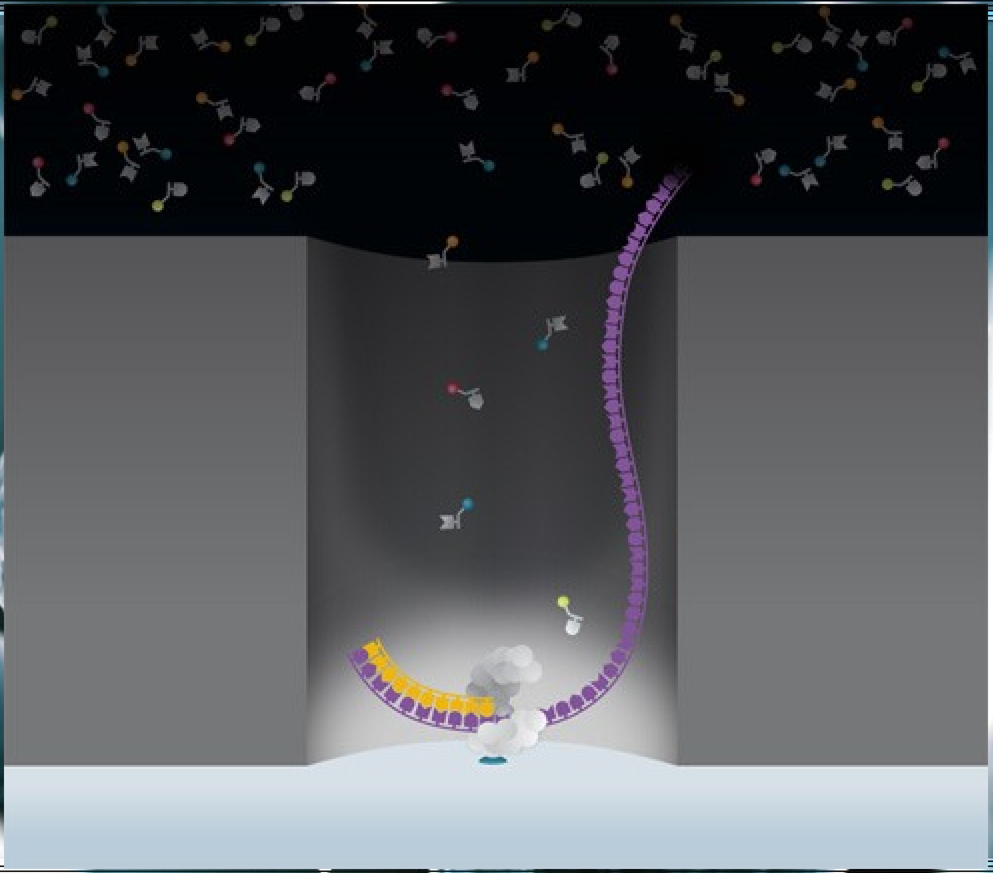
1.0 > 10 > 20 gb / run
25bp > 35bp > 50bp
Seq. by Ligation

2010

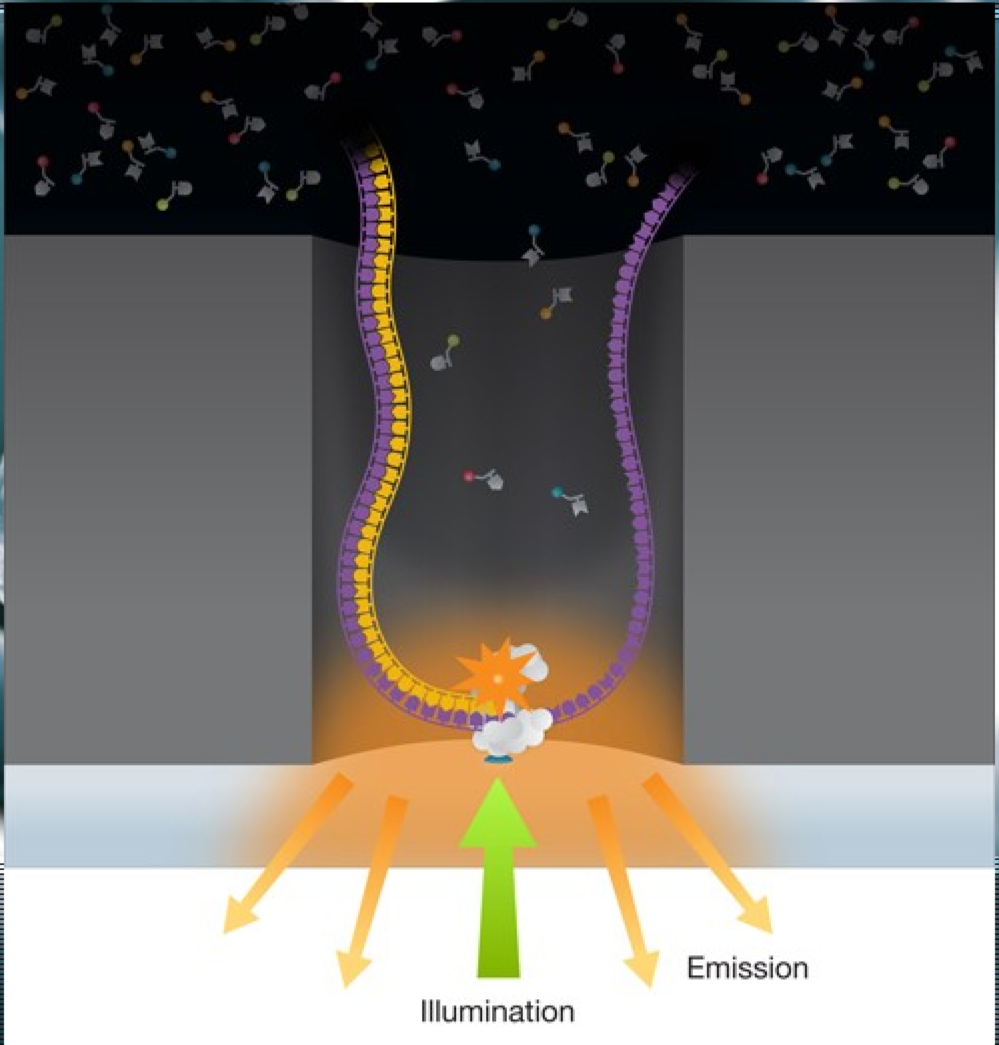
What’s next???

Moving beyond amplification: Pacific Biosciences

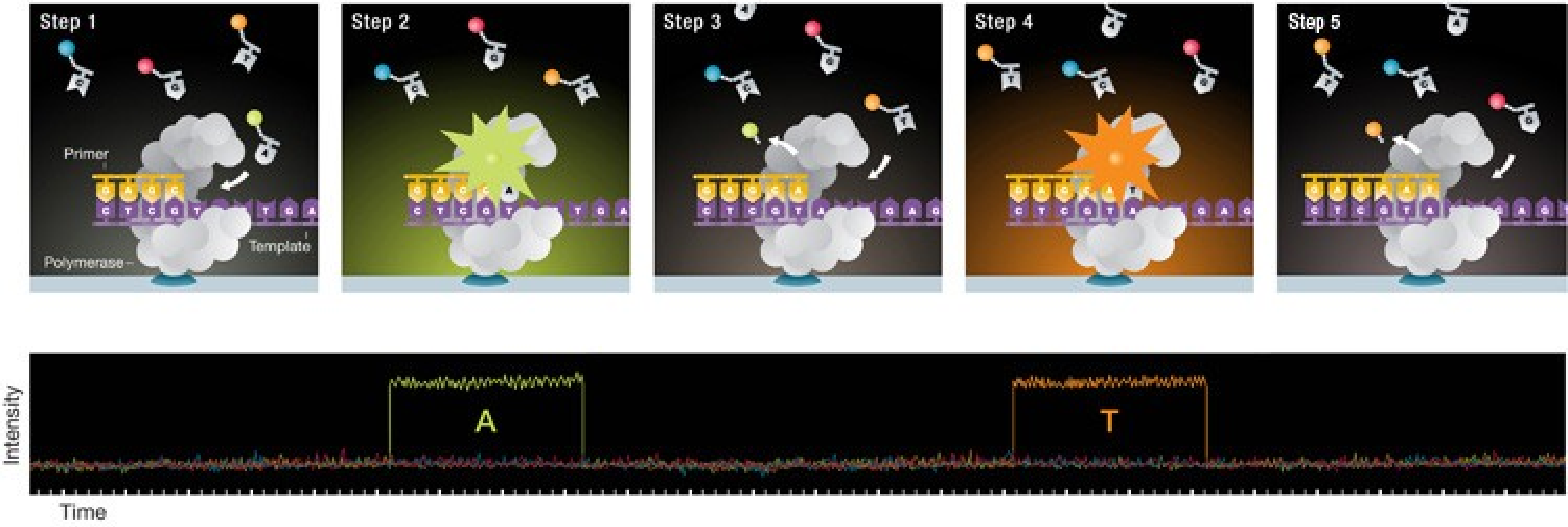




Moving beyond amplification: Pacific Biosciences



Moving beyond amplification: Pacific Biosciences



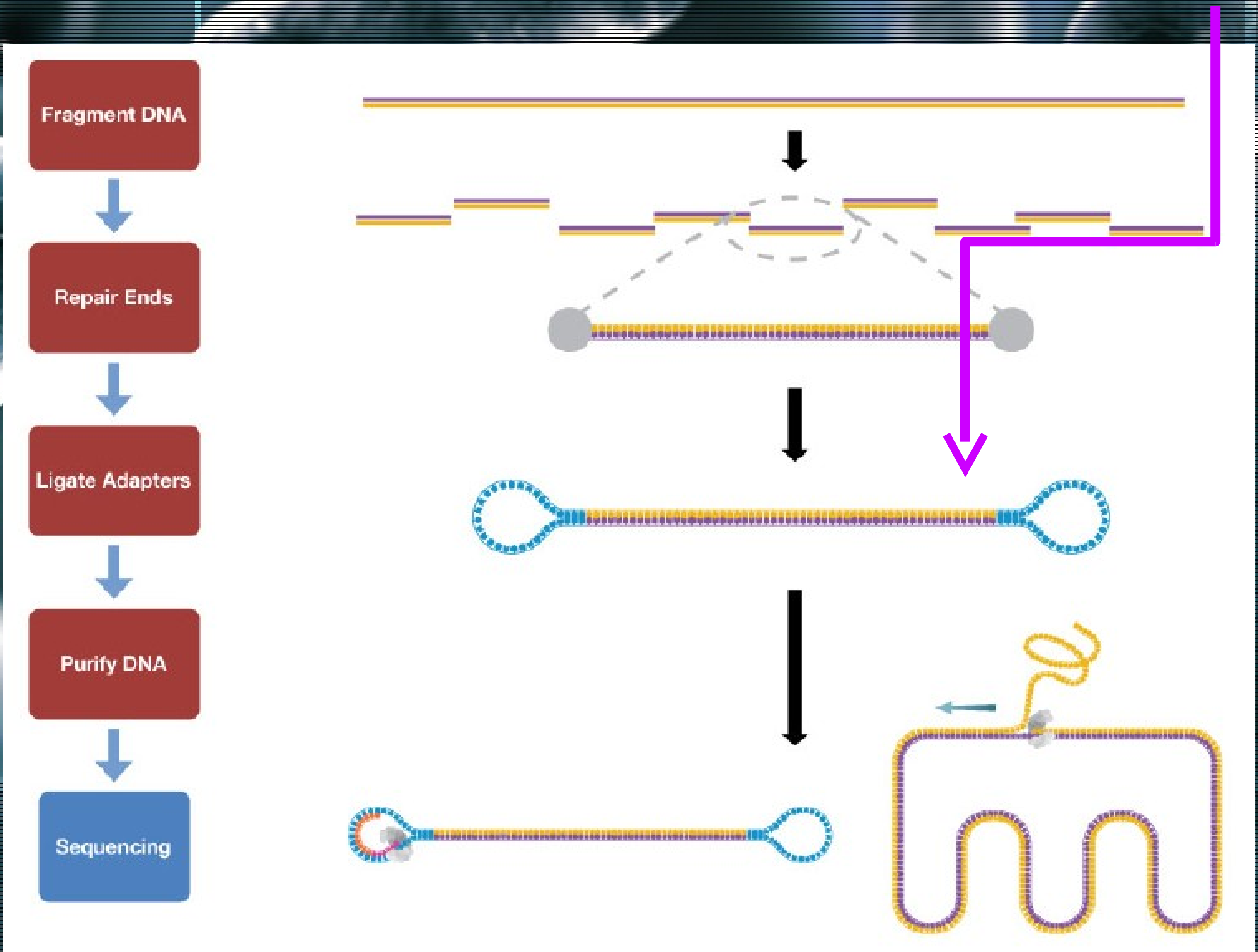
Can get an average of 1kb reads!

Takes 15 min. for a run!

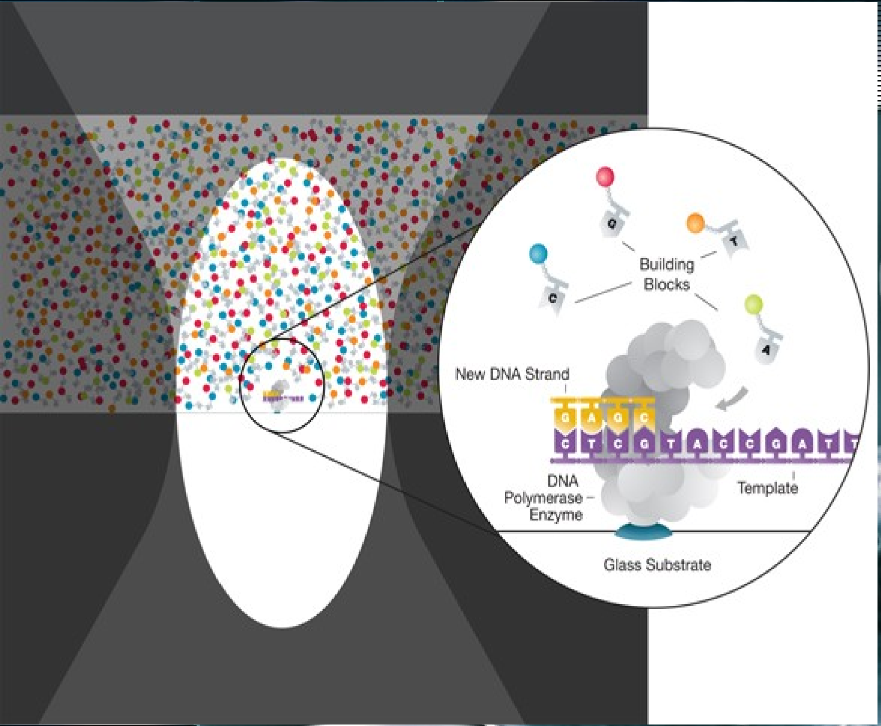
SMRT!!

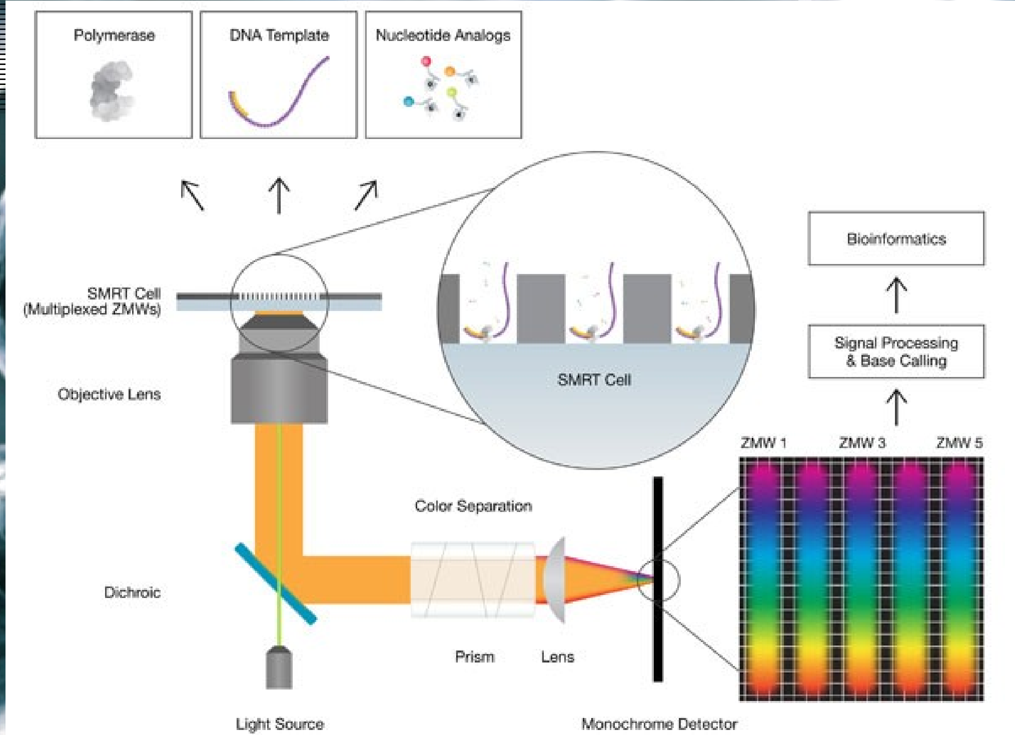
Still need a priming site

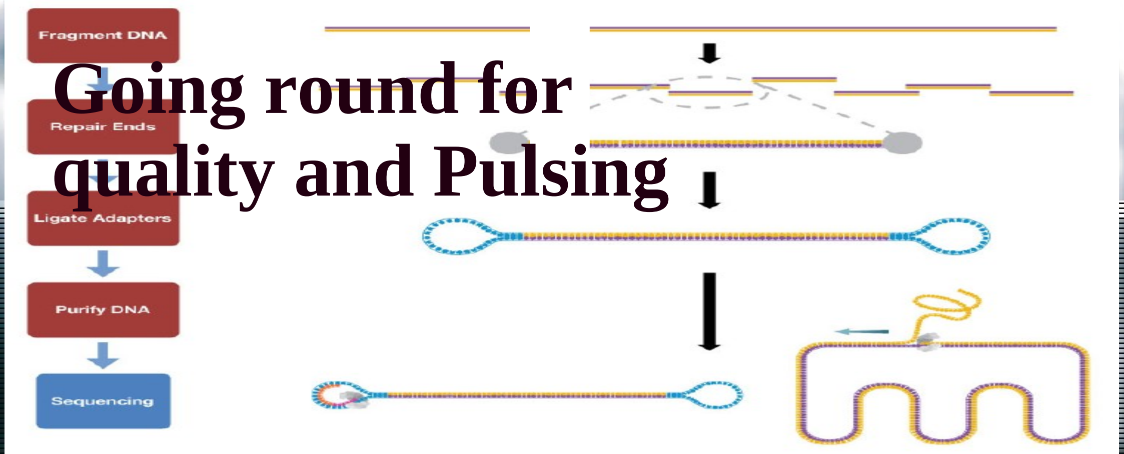
SMRTbell!



Moving beyond amplification: Pacific Biosciences







Going round for quality and Pulsing

“Starlite” – FRET chemistry

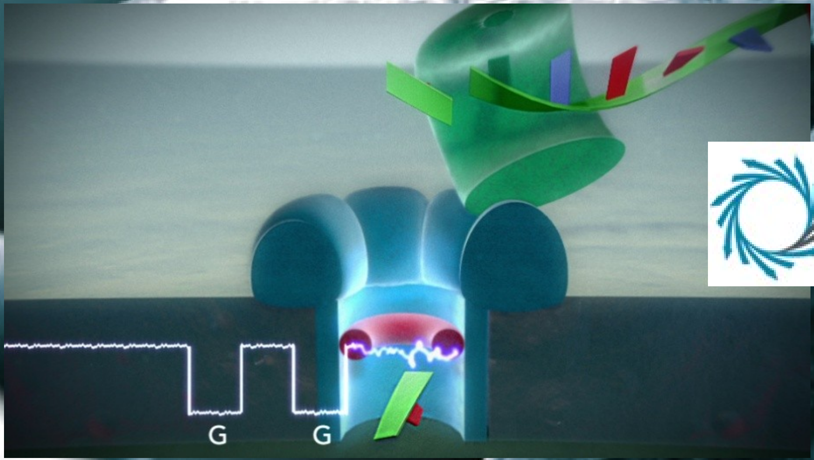
Ion Torrent's technology is based on a semiconductor chip that includes 1.55 million electronic sensors.

Ion Torrent presented its desktop-sized sequencer at the 2010 Advances in Genome Biology and Technology conference.

ion torrent

life technologies™

The technology uses a quantum dot tethered to a DNA polymerase and measures fluorescence in real time as bases get incorporated by the polymerase.



Oxford NANOPORE Technologies®

HALCYON MOLECULAR

Nano-robotic DNA manipulation technology??

The Economist

Obama the warrior

Misgoverning Argentina

The economic shift from West to East

Genetically modified crops blossom

The right to eat cats and dogs

FEBRUARY 27TH–MARCH 5TH 2010 Economist.com

The data deluge

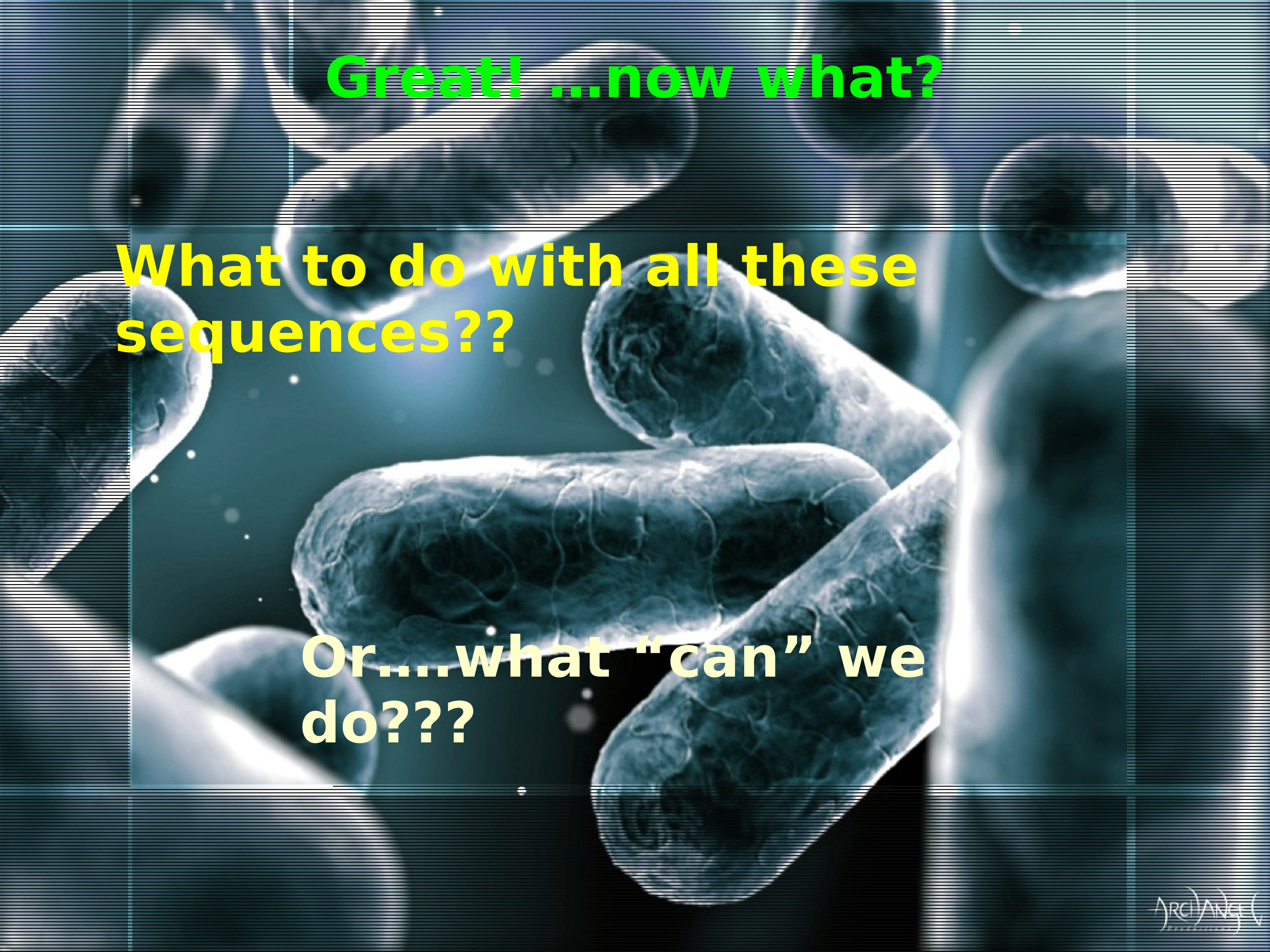
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

US$6.99 · C$7.99

Argentina	$7.00	Canada	C$7.99	Jamaica	J$510	Trinidad & Tobago	TD$43
Bahamas	$9.95	Chile	Ch$5,000	Mexico	Mex$70	Turks & Caicos	$9.50
Barbados	Bds$16.50	Colombia	Col$22,000	Peru	S/.38.00	UK	£4.00
Bermuda	Bd$ 7.00	Costa Rica	₡6,900	Spain	€5.50	USA	US$6.99
Brazil	R$24.90	Guyana	GYD 1,650	St.Maarten	$9.25	Venezuela	Bs27

The technology uses a quantum dot tethered to a DNA polymerase and measures fluorescence in real time as bases get incorporated by the polymerase.

Great! …now what?



What to do with all these sequences??

Or….what “can” we do???

A next-gen challenge!: Metagenomics

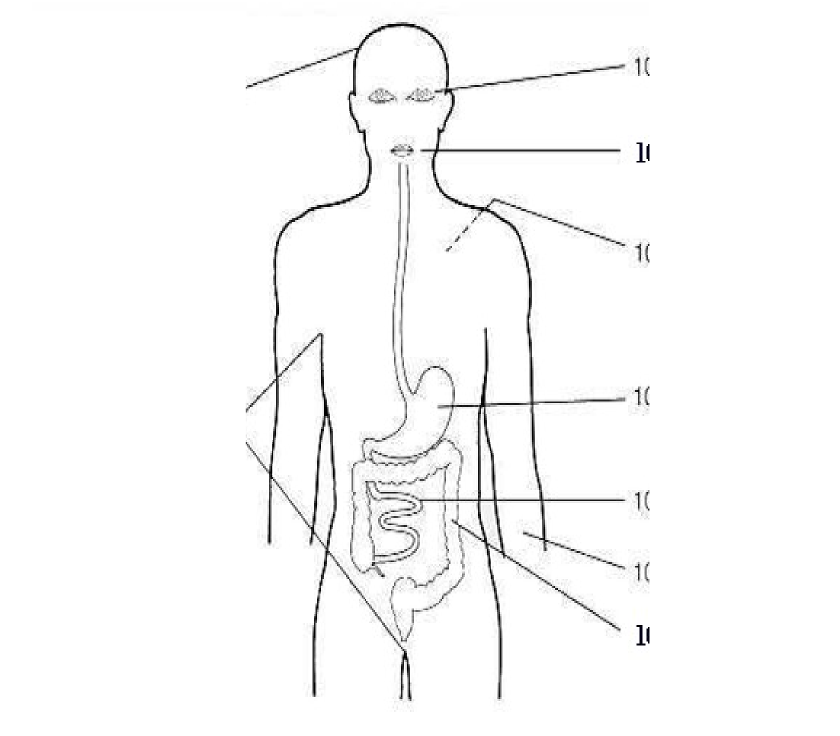
THE NEW SCIENCE OF METAGENOMICS

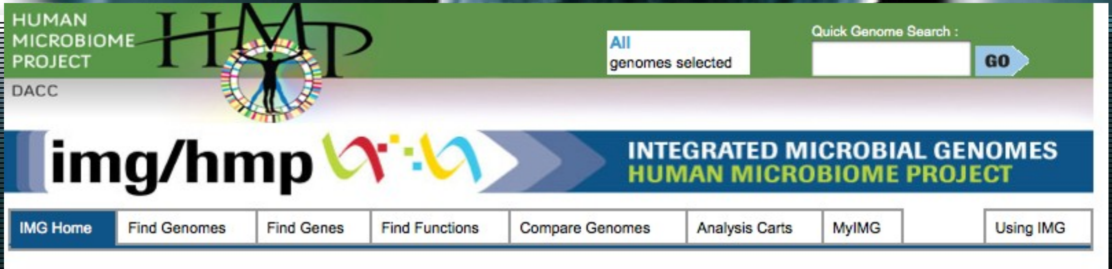
Revealing the Secrets of Our Microbial Planet

Of Microbes and Men

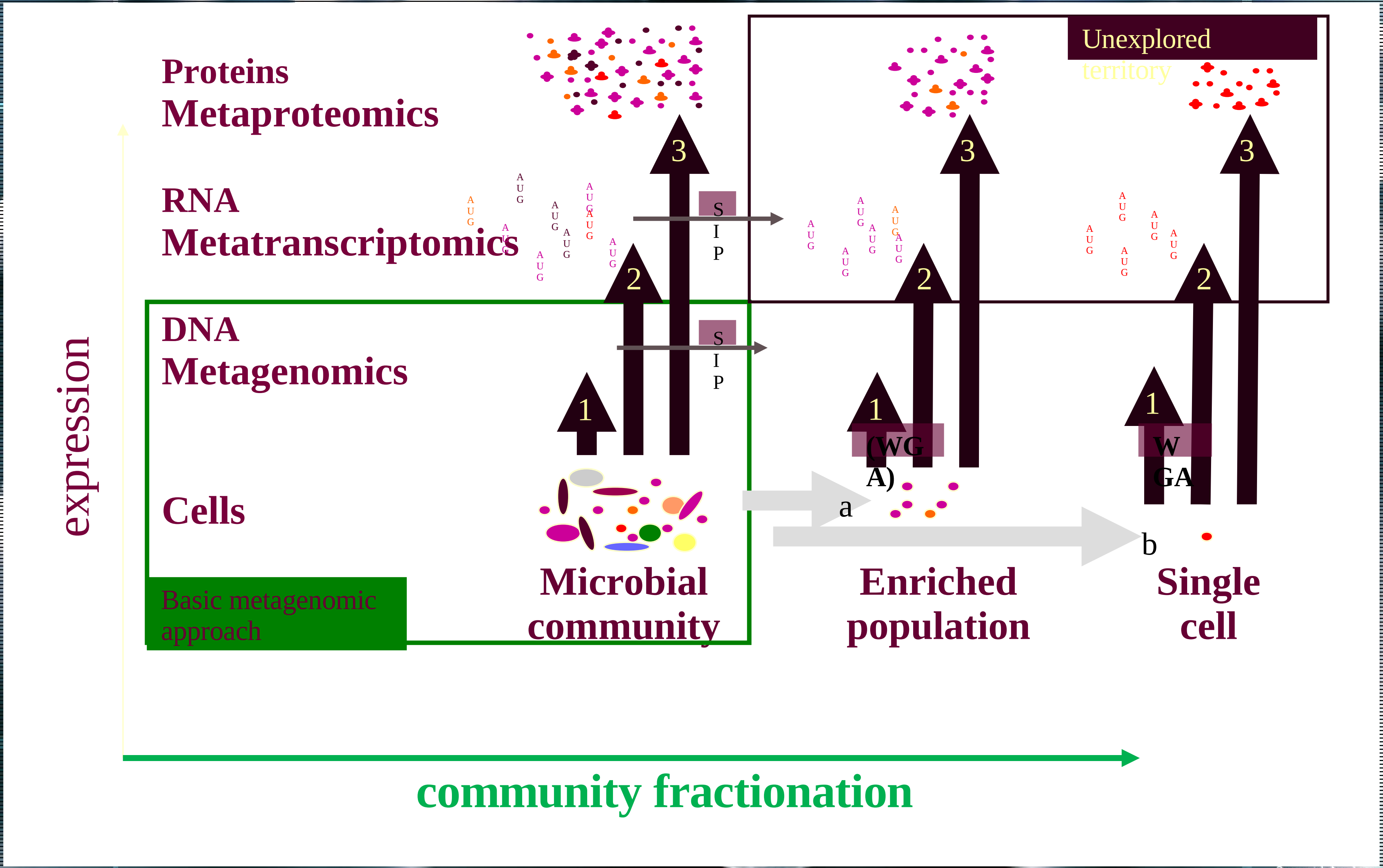
- The human body contains about 1013 cells
- From shortly after birth until death, the human body, routinely harbors 1014 bacteria (100 trillion!!!)
- You are roughly 10% hu

1 ml saliva: 40 million cells





Complementary technologies



Proteins
Metaproteomics

RNA
Metatranscriptomics

DNA
Metagenomics

Cells

Basic metagenomic approach

Unexplored territory

Microbial community

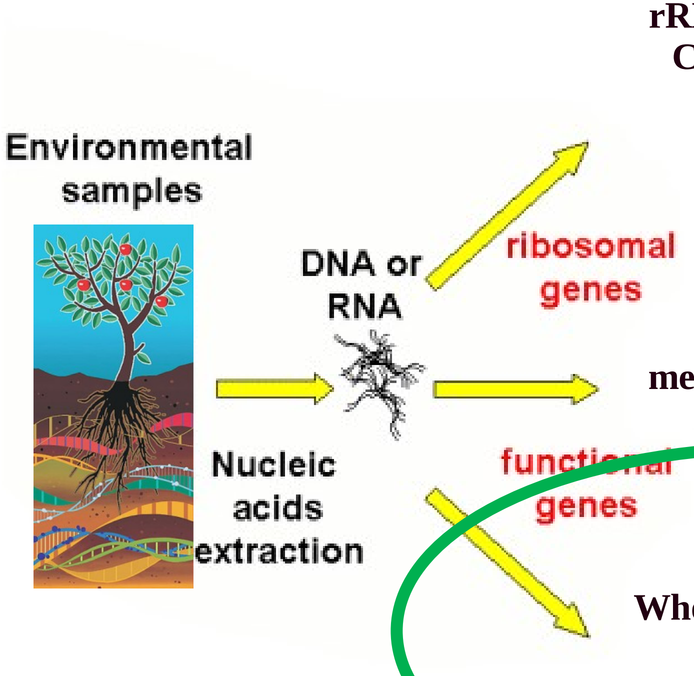
Enriched population

Single cell

expression

community fractionation

Different Metagenomic Flavours



rRNA (16S) gene survey
Community profiling

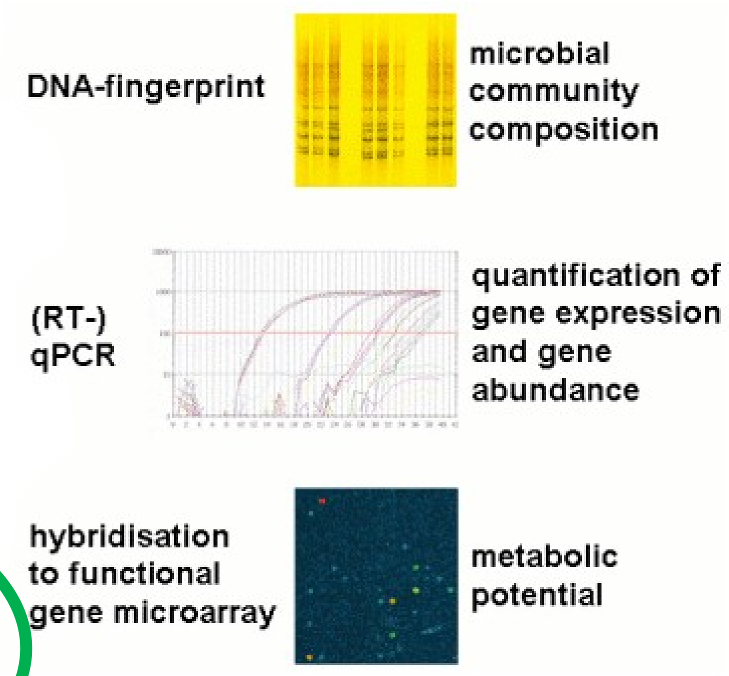
PCR-based

RNAseq
metatranscriptomics

RNA-based

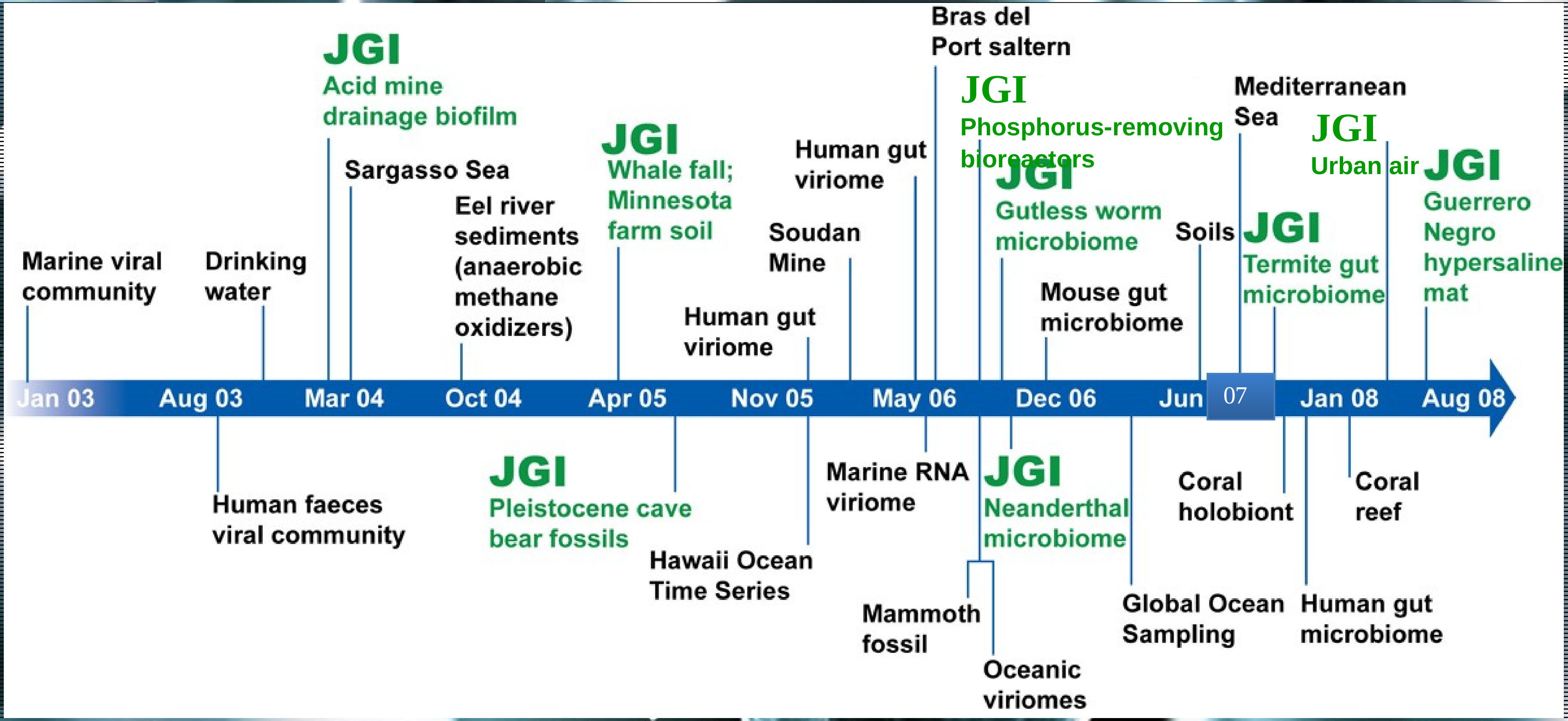
Whole genome shotgun metagenomics

DNA-based


DNA-fingerprint
microbial community composition
(RT-) qPCR
quantification of gene expression and gene abundance
hybridisation to functional gene microarray
metabolic potential

JGI: A metagenomics hub

published shotgun metagenomic studies



Jan 03 – Aug 03 – Mar 04 – Oct 04 – Apr 05 – Nov 05 – May 06 – Dec 06 – Jun 07 – Jan 08 – Aug 08

Marine viral community

Drinking water

Human faeces viral community

JGI Acid mine drainage biofilm

Sargasso Sea

Eel river sediments (anaerobic methane oxidizers)

JGI Whale fall; Minnesota farm soil

JGI Pleistocene cave bear fossils

Human gut viriome

Hawaii Ocean Time Series

Soudan Mine

Human gut viriome

Marine RNA viriome

Bras del Port saltern

JGI Phosphorus-removing bioreactors

Mammoth fossil

Oceanic viriomes

JGI Gutless worm microbiome

JGI Neanderthal microbiome

Mouse gut microbiome

Global Ocean Sampling

Soils

Mediterranean Sea

JGI Termite gut microbiome

Coral holobiont

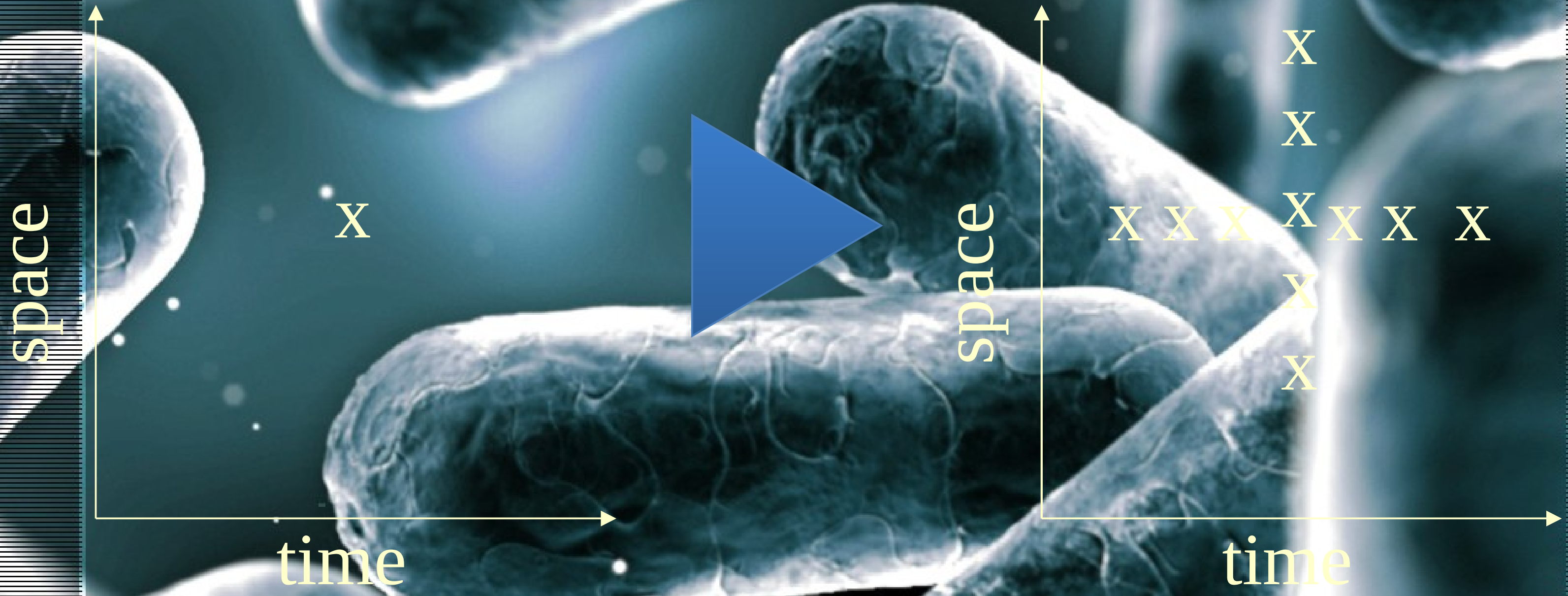
Human gut microbiome

JGI Urban air

Coral reef

JGI Guerrero Negro hypersaline mat

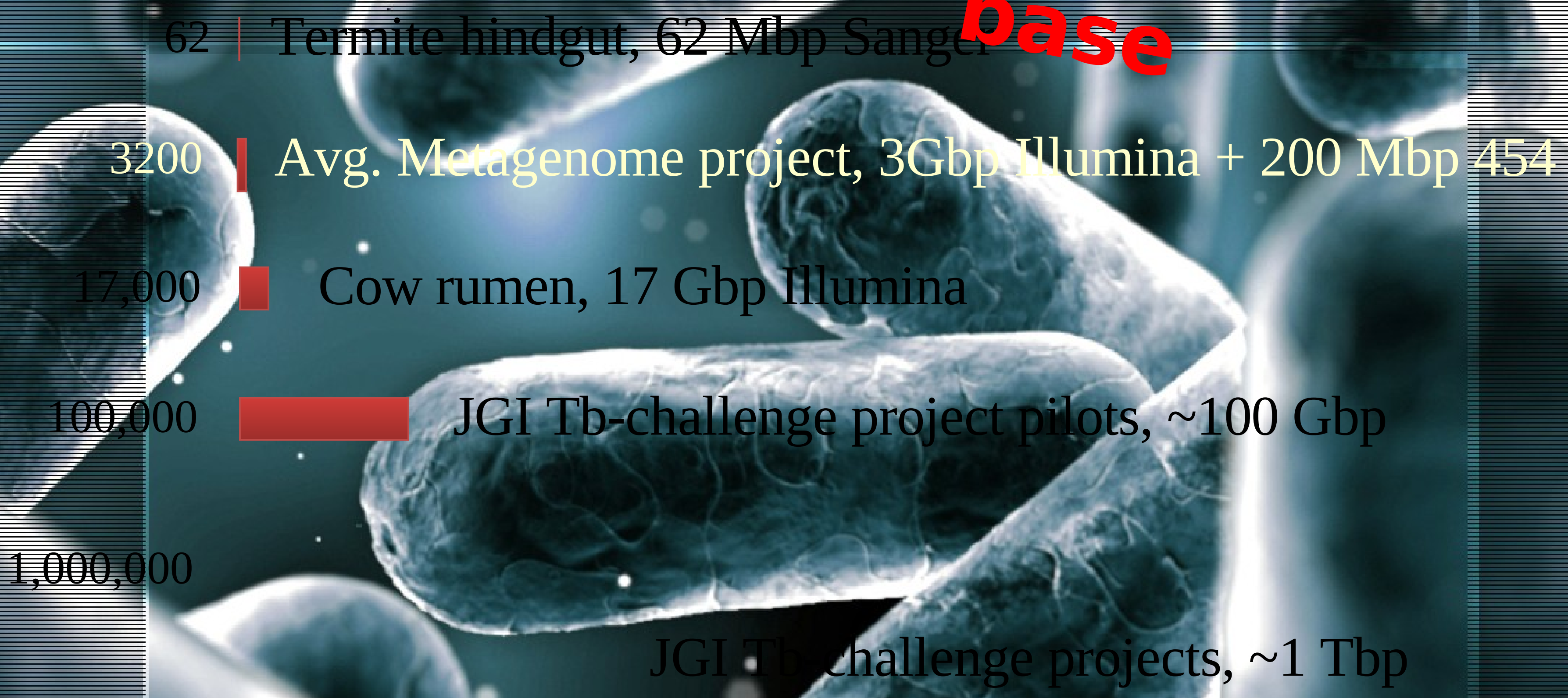
From a snapshot to a series



NOTE: Sample metadata is critical!

The next challenge: Terabase scale

Terror-base



62 Termite hindgut, 62 Mbp Sanger

3200 Avg. Metagenome project, 3Gbp Illumina + 200 Mbp 454

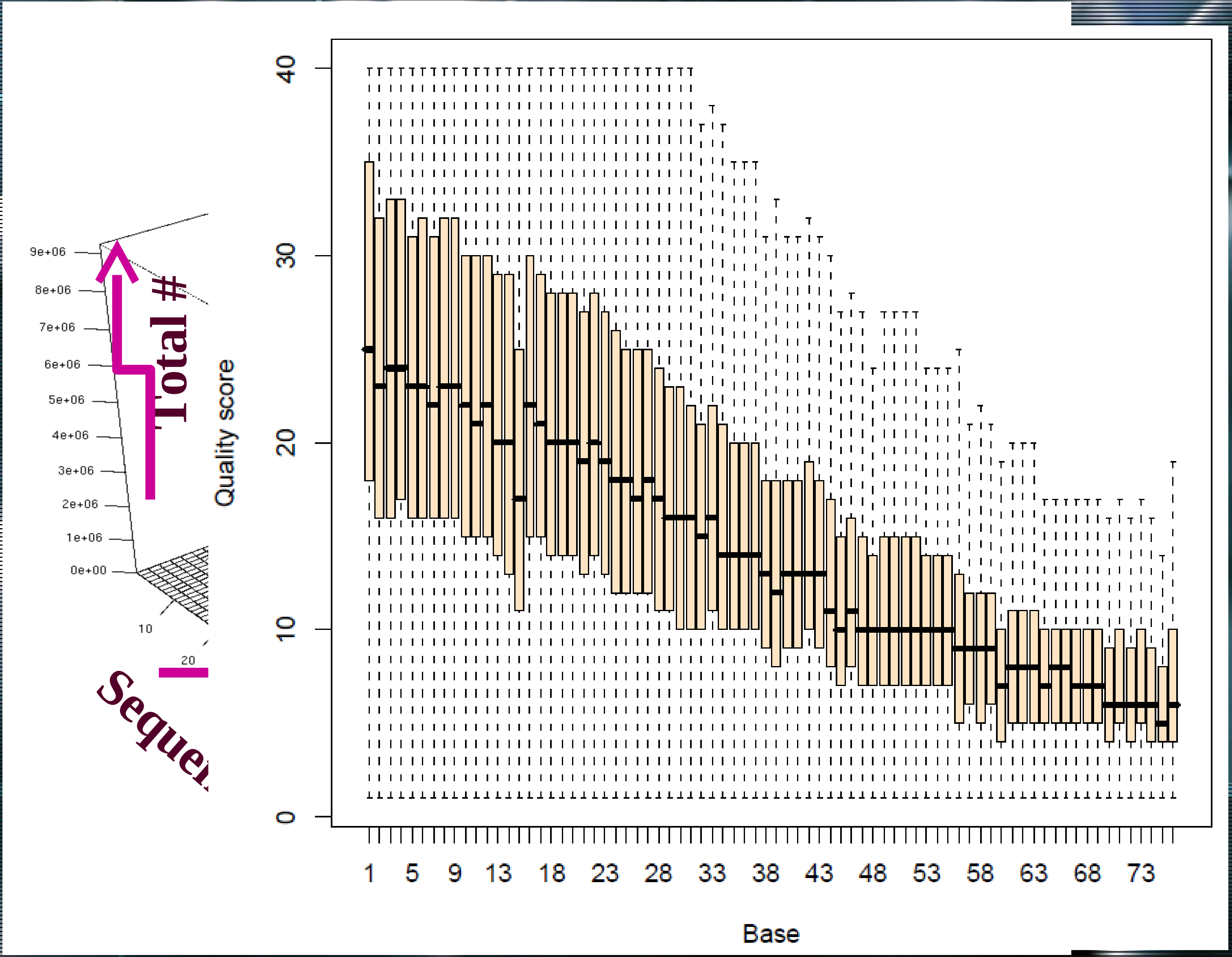
17,000 Cow rumen, 17 Gbp Illumina

100,000 JGI Tb-challenge project pilots, ~100 Gbp

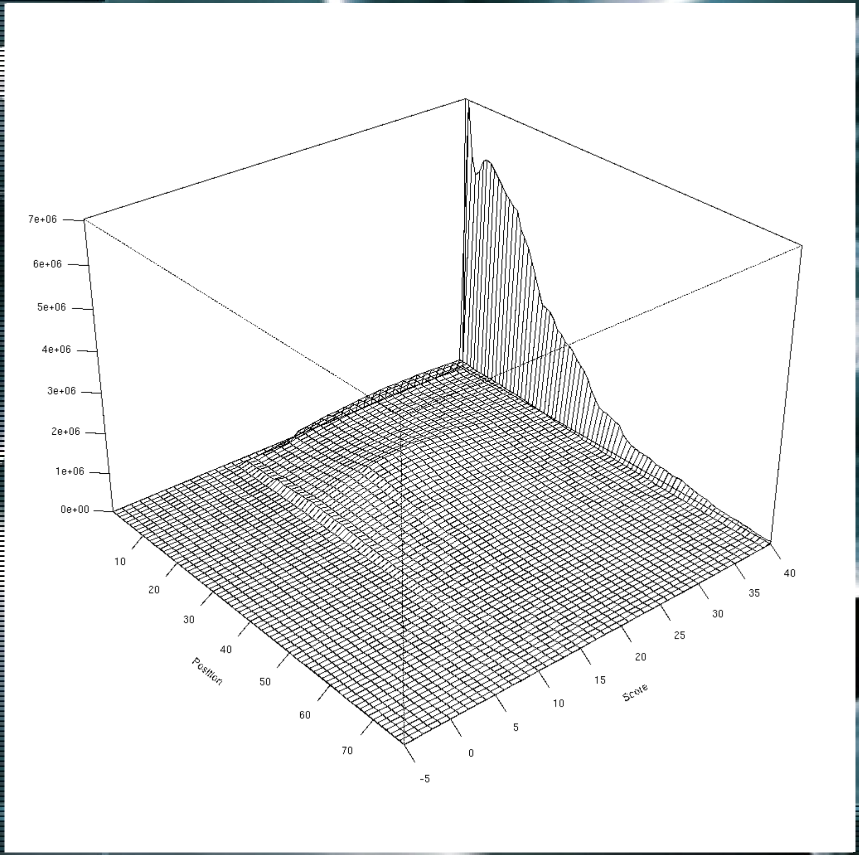
1,000,000 JGI Tb-challenge projects, ~1 Tbp

High probability of computational bottlenecks
all vs all will NOT scale!
New approaches needed…

Quality still a possible pitfall...

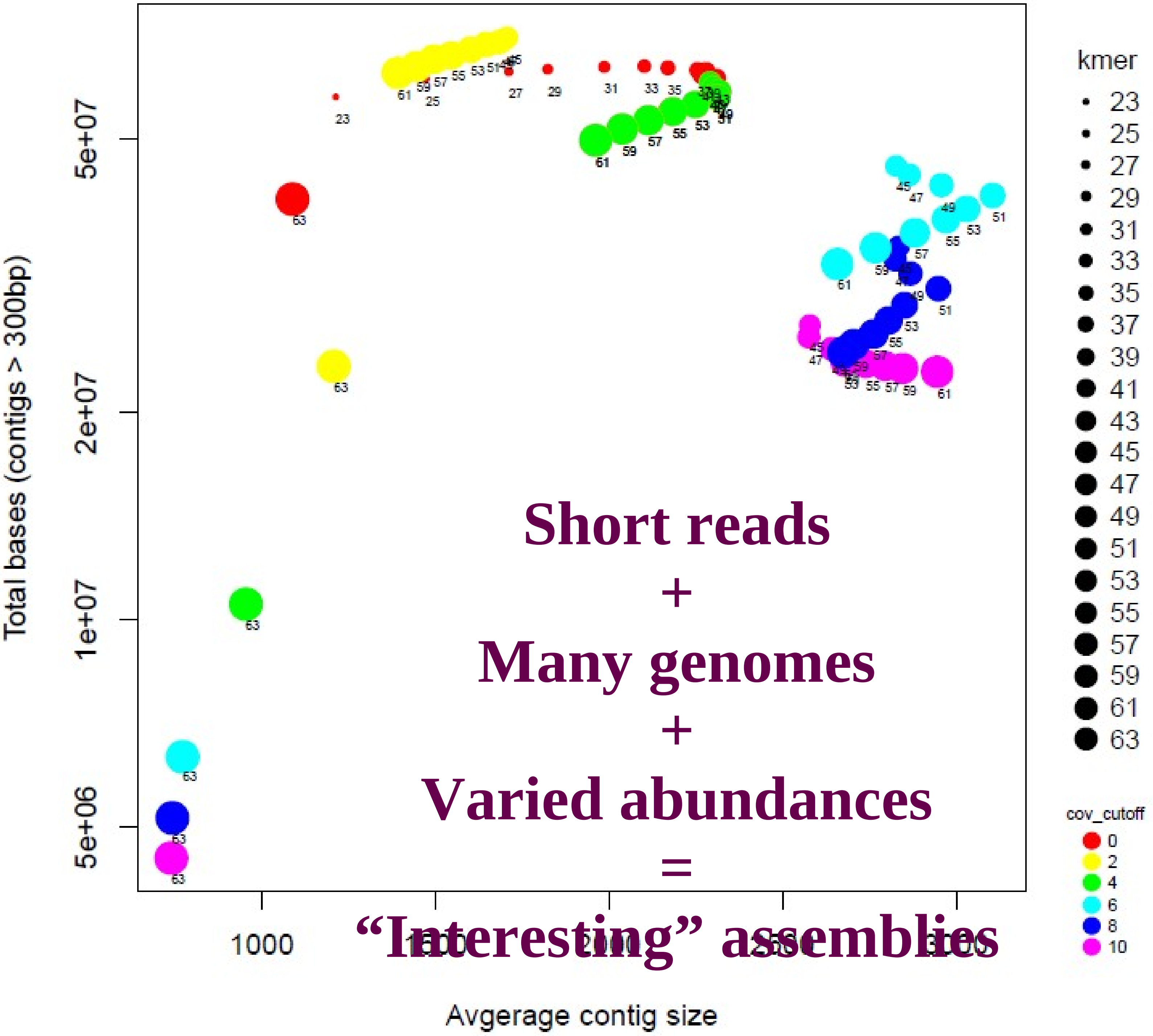


Illumina data not always so good…



- Raw data
  - 29’753’554 reads
  - 76 bp
  - 2’261’270’104 bp
- Trimming
  - 14’866’717 reads, 50%
  - ~36.5 bp, 48%
  - 542’160’999 bp, 24%

On to assembly…



Short reads

+

Many genomes

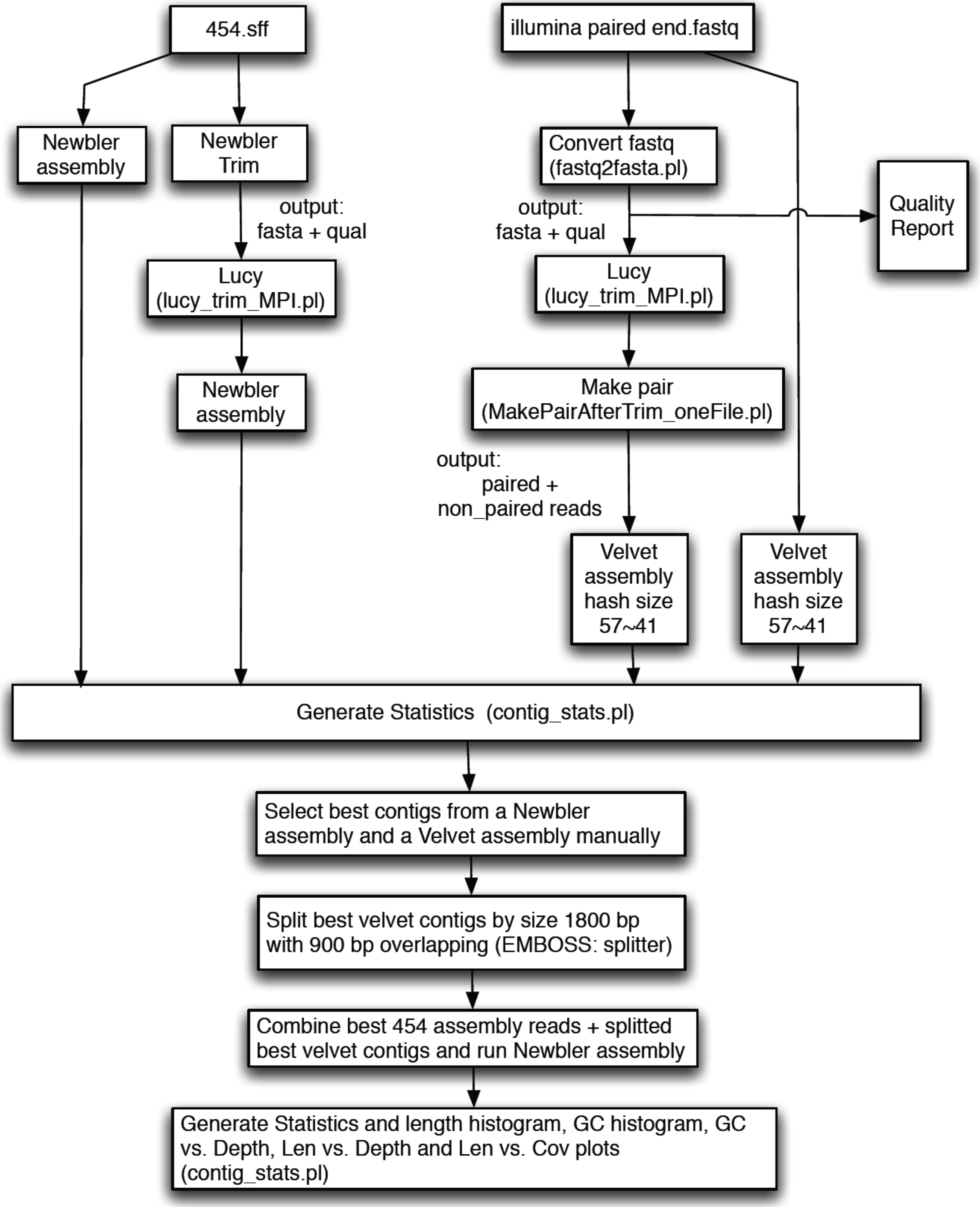
+

Varied abundances

=

“Interesting” assemblies

Combining 454 with Illumina data: A metagenomics nightmare



454.sff

Newbler assembly

Newbler Trim

output: fasta + qual

Lucy (lucy_trim_MPI.pl)

Newbler assembly

illumina paired end.fastq

Convert fastq (fastq2fasta.pl)

output: fasta + qual

Quality Report

Lucy (lucy_trim_MPI.pl)

Make pair (MakePairAfterTrim_oneFile.pl)

output: paired + non_paired reads

Velvet assembly hash size 57~41

Velvet assembly hash size 57~41

Generate Statistics (contig_stats.pl)

Select best contigs from a Newbler assembly and a Velvet assembly manually

Split best velvet contigs by size 1800 bp with 900 bp overlapping (EMBOSS: splitter)

Combine best 454 assembly reads + splitted best velvet contigs and run Newbler assembly

Generate Statistics and length histogram, GC histogram, GC vs. Depth, Len vs. Depth and Len vs. Cov plots (contig_stats.pl)

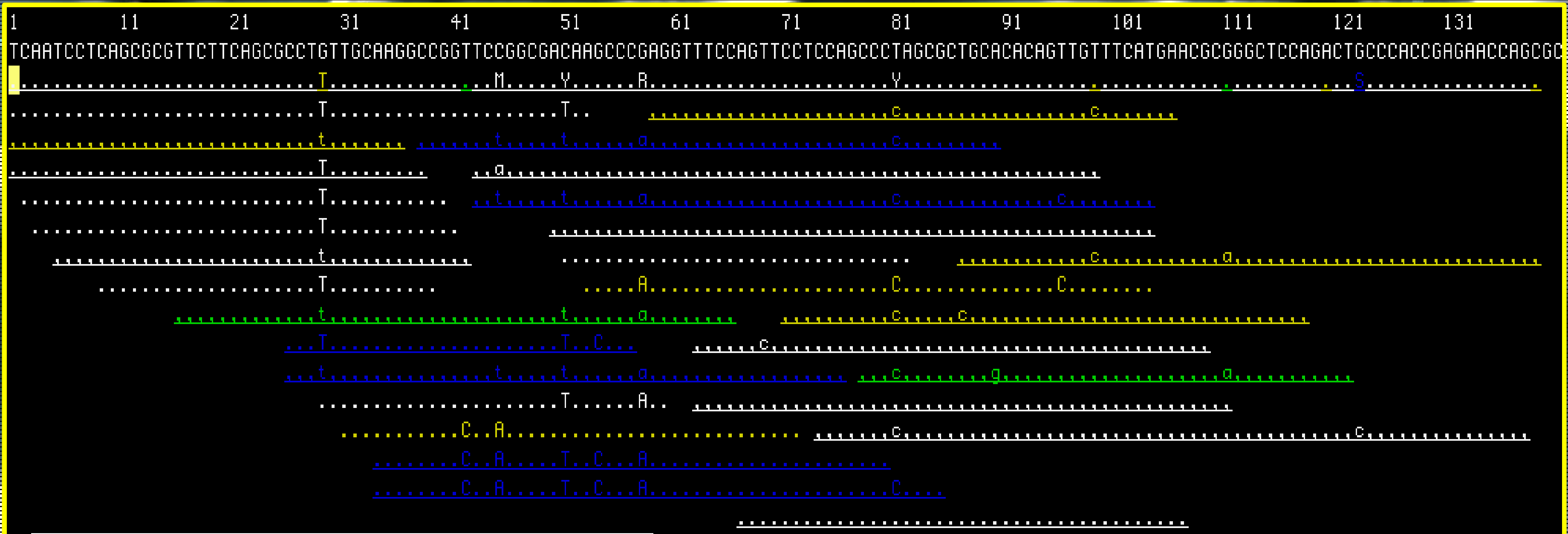
Successful Assemblies!

			454+Illumina
	454.sff	Illumina.hash37	300contigsplit
Tool:	Newbler	Velvet	Newbler
Assembled_reads:	86.43%	88.19%	89.75%
Total_bases:	22347052	30510493	29448133
Singleton:	63419	3992352	48360
Contigs_number:	13072	41749	9239
N50:	3217	6144	10761
Max:	71063	183750	120185
Top10_bases:	508436	1176640	823422
Top20_bases:	859965	1824883	1456953
Top40_bases:	1434355	2836397	2539934
Top100_bases:	2624547	4876231	4883464
>100kb_bases:	0	745385	120185
>50kb_bases:	242880	2270979	2539934
>25kb_bases:	1511575	5809841	7026589
>10kb_bases:	4339620	11955915	15222002
>5kb_bases:	8098882	16469453	19767917
>3kb_bases:	11636377	19039014	22587424
>2kb_bases:	14093342	20783840	24498444
>1kb_bases:	17718317	22937756	26762015

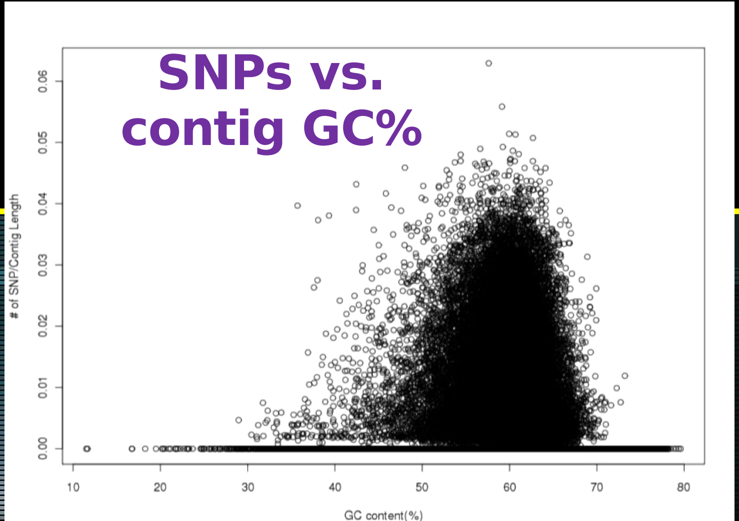
Successful Assemblies?

	454	illumina.hash57	454+illumina 300 contigsplit
Tool:	Newbler	Velvet	Newbler
Assembled_reads:	19.38%	3.51%	22.69%
Total_bases:	4321654	3314719	3850267
Singleton:	367455	46684681	314974
Contigs number:	9399	16845	7781
N50:	531	196	573
Max:	10804	15298	16110
Top10_bases:	62001	41192	73005
Top20_bases:	98823	63034	114008
Top40_bases:	157762	100884	184488
Top100_bases:	292812	185144	341570
>100kb_bases:	0	0	0
>50kb_bases:	0	0	0
>25kb_bases:	0	0	0
>10kb_bases:	10804	15298	16110
>5kb_bases:	48045	15298	63869
>3kb_bases:	124047	22609	206016
>2kb_bases:	269515	63034	385995
>1kb_bases:	832447	227448	1011106

Can see population sequence heterogeneity



```
1         11        21        31        41        51        61        71        81        91        101       111       121       131
TCAATCCTCAGCGCGTTCTTCAGCGCCTGTTGCAAGGCCGGTTCCGGCGACAAGCCCGAGGTTTCCAGTTCCTCCAGCCCTAGCGCTGCACACAGTTGTTTCATGAACGCGGGCTCCAGACTGCCCACCGAGAACCAGCGC
............................T...............M.....Y......R......................Y.........................................S................
............................T.....................T..     ,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,c,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,t,,,,,,, ,,,,,,,t,,,,,t,,,,,,a,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,T,,,,,,,,,    ,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
 ...........................T...........  ,,t,,,,,t,,,,,,a,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,,,c,,,,,,,,
  ..........................T............        ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
    ,,,,,,,,,,,,,,,,,,,,,,,,t,,,,,,,,,,,,,        ................................    ,,,,,,,,,,,,c,,,,,,,,,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,
        ....................T..........             .....A......................C..............C........
               ,,,,,,,,,,,,,t,,,,,,,,,,,,,,,,,,,,,t,,,,,,a,,,,,,,,    ,,,,,,,,,,c,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                         ...T.....................T..C...     ,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                         ,,,t,,,,,,,,,,,,,,,t,,,,,t,,,,,,a,,,,,,,,,,,,,,,,,, ,,,c,,,,,,,,g,,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,
                            ......................T......A..  ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                              ...........C..A........................... ,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,,,,
                                 ........C..A.....T..C...A......................
                                 ........C..A.....T..C...A......................C....
                                                                  .........................................
                                                                  ...............................................
                                                                       .........C................................
                                                                          ,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,c,,,
                                                                          ,,,,,,c,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,
```



SNPs vs. contig GC%

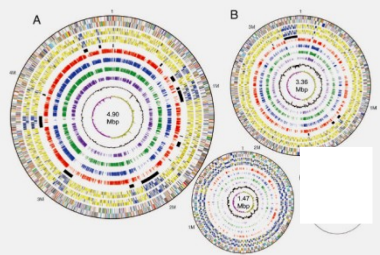
Re-map reads to contigs

Toward a comprehensive bioinformatic workbench for the genomics community

Sequence Data

MG-DNA

Assembly profile by phylogeny and function; Profile rRNA hits; Mapping to genomes; Characterize protein hits



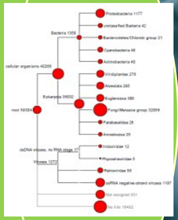
RNAseq

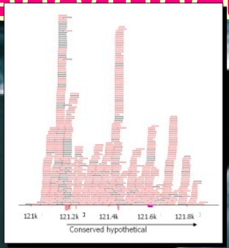
rRNA hits; map to genomes or to functions; count hits to gene/function;

Survey

rRNA profiling

MEGAN or NBC, Trees

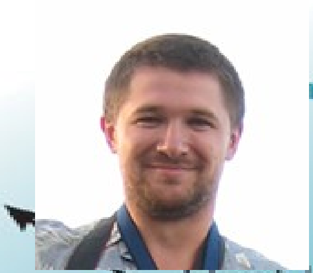




Compare datasets (communities and/or treatments)

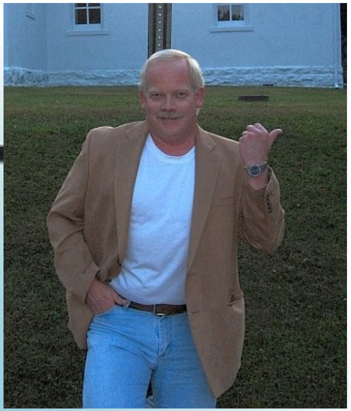
Acknowledgements

Chien-Chi Lo

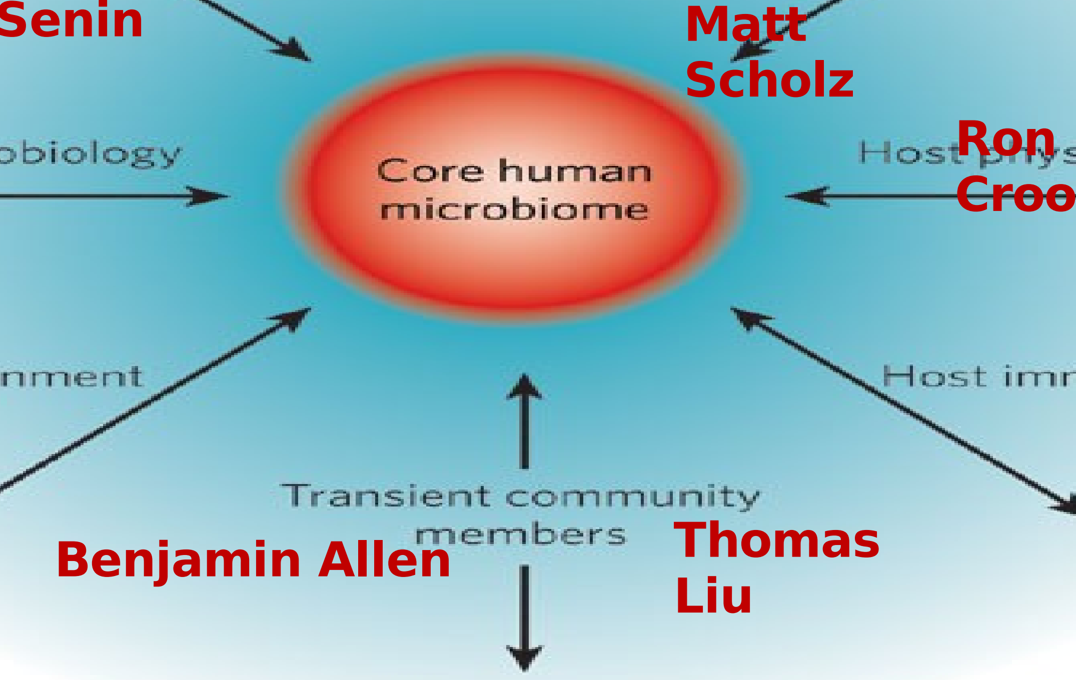


Pavel Senin

Matt Scholz



Ron Croonenberg



Andrew Liu

Benjamin Allen

Thomas Liu

Gary Xie

NIDCR National Institute of Dental and Craniofacial Research
National Institutes of Health