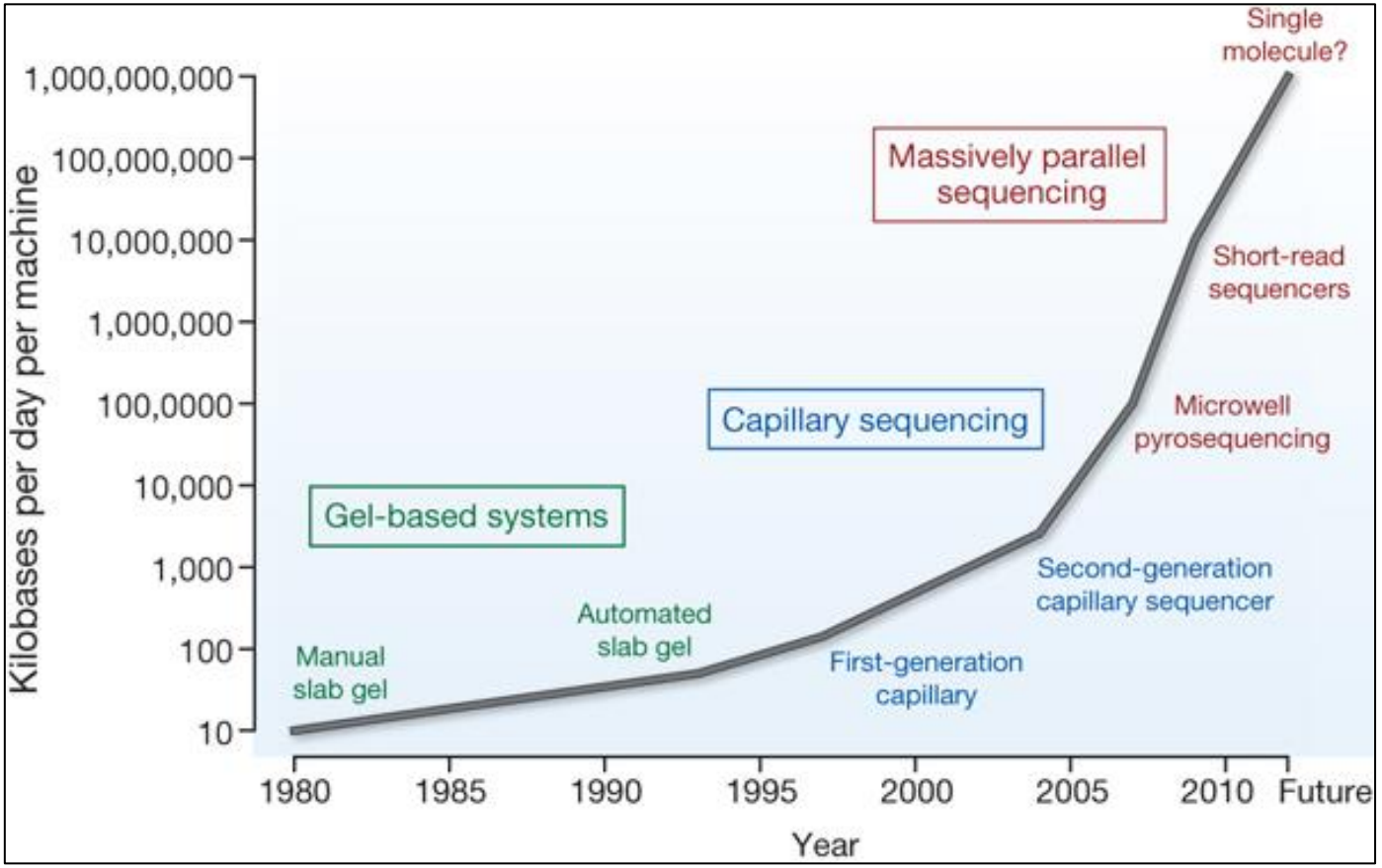
Botany 2014 Next Generation Sequencing Workshop

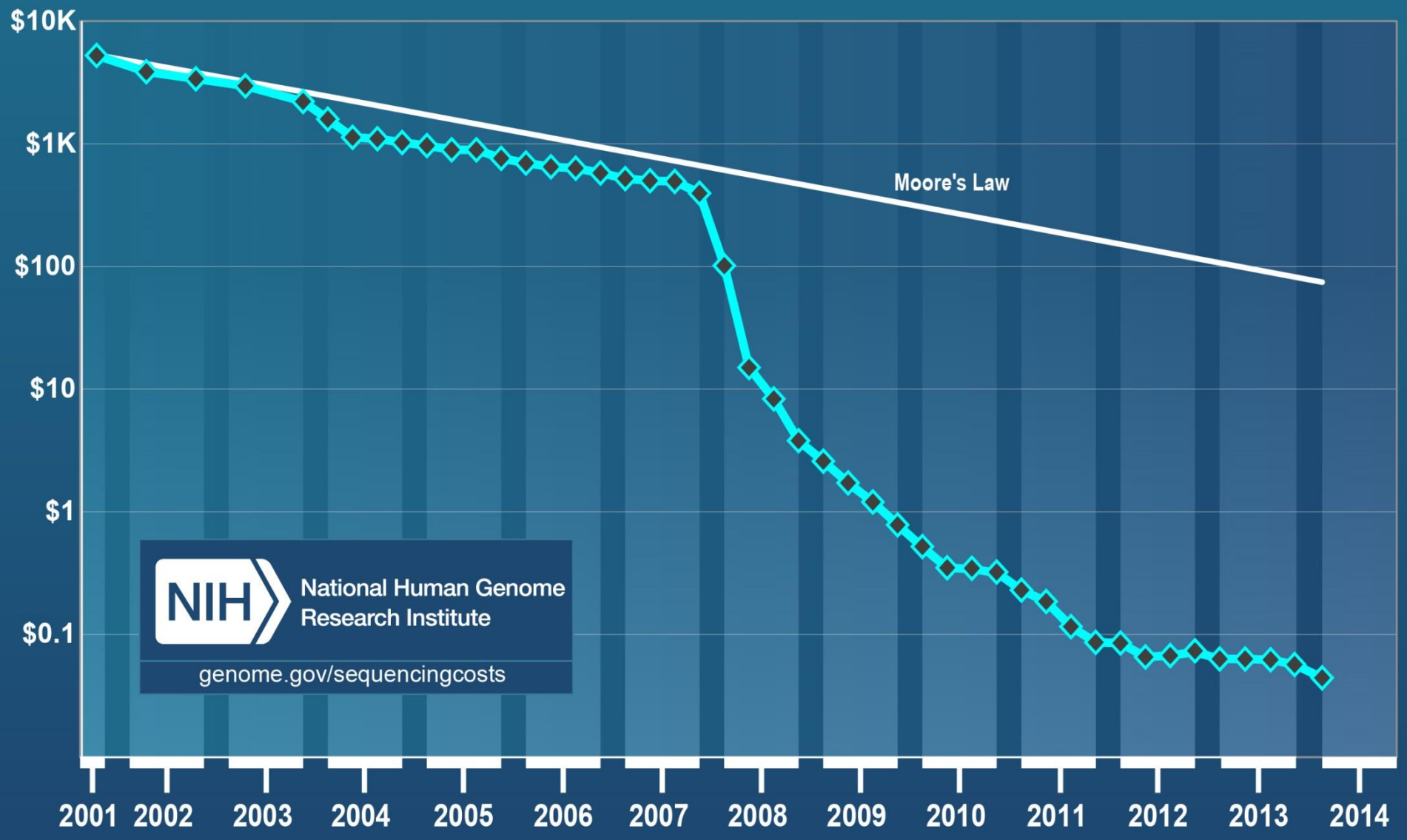
1. Sequencing Technology
2. Genome Reduction
3. Sequence Processing
4. Bionformatics
5. What you really need to know.

Improvements in the rate of DNA sequencing over the past 30 years

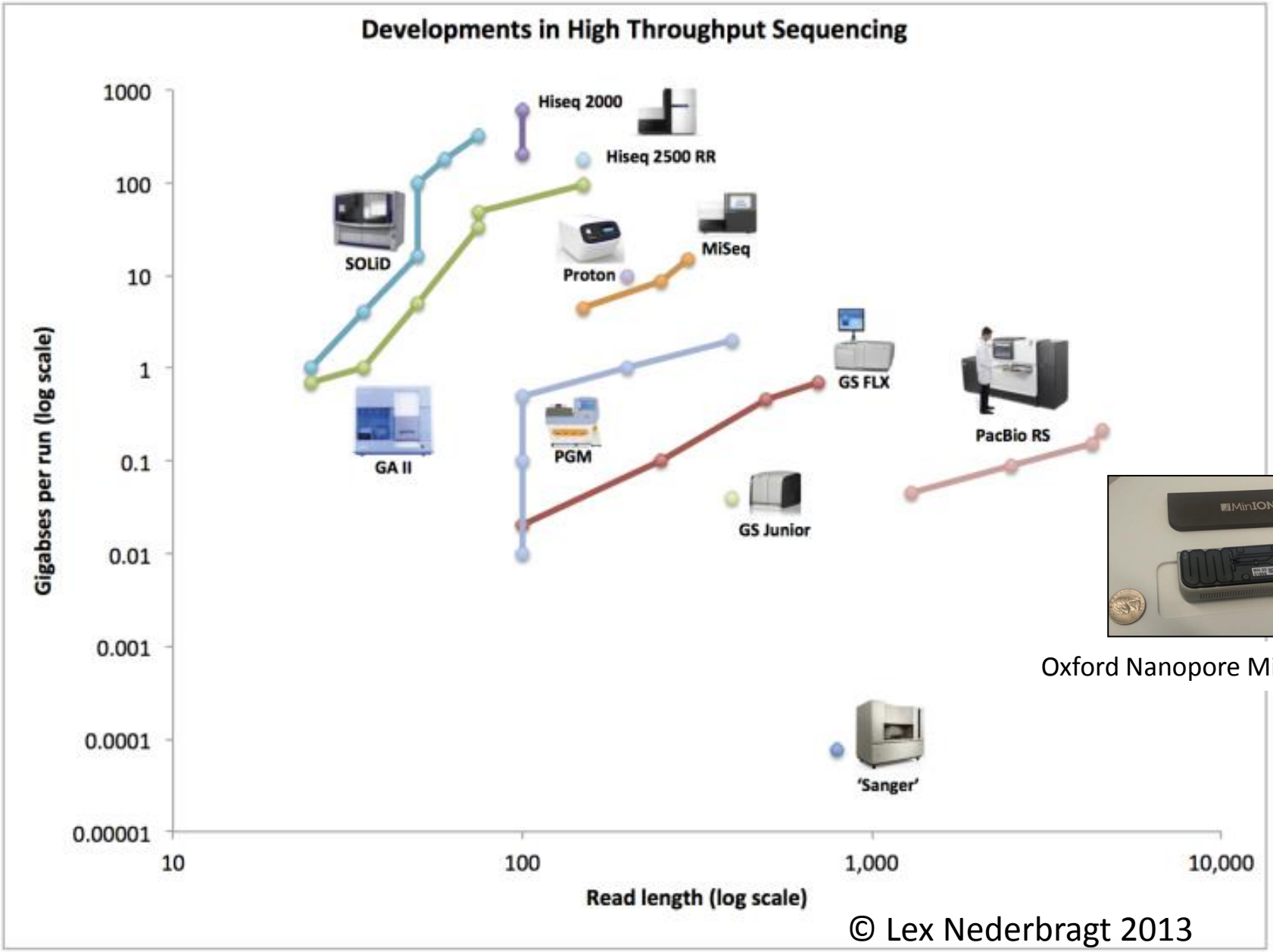


MR Stratton et al. (2009) The cancer genome. Nature **458**, 719-724

Cost per Raw Megabase of DNA Sequence



January 2014



Oxford Nanopore MinION

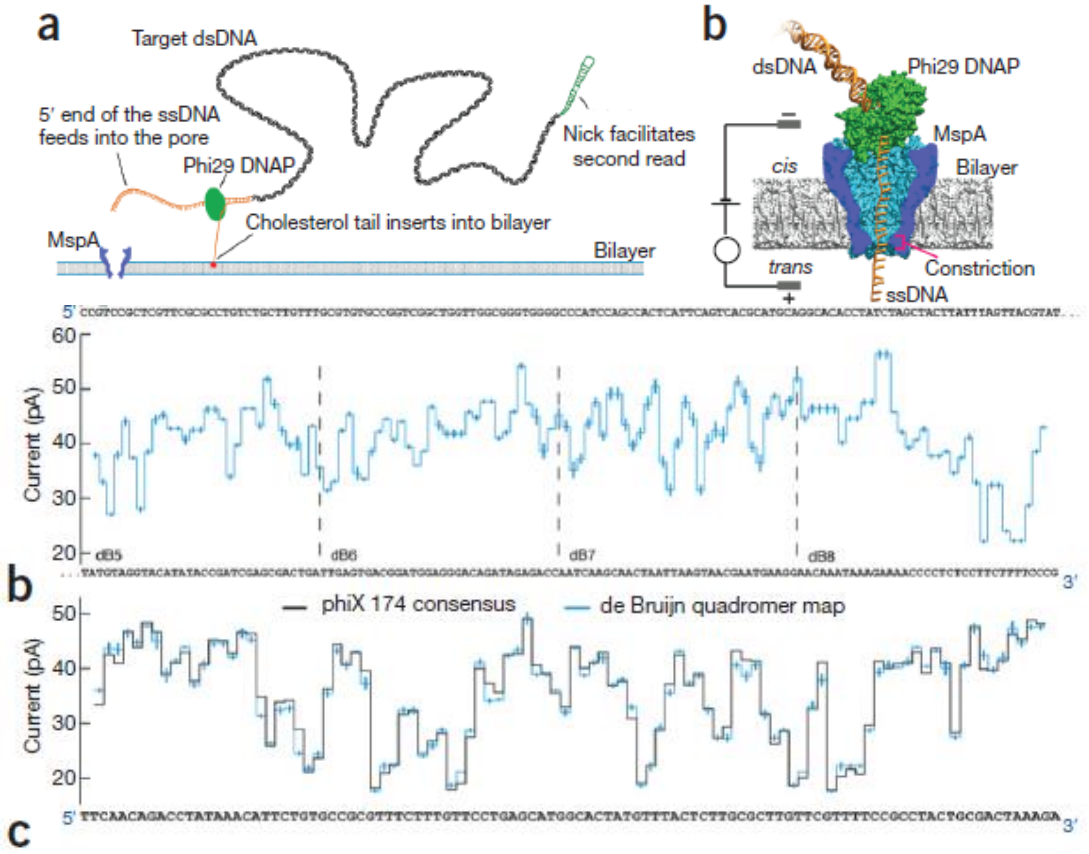
http://flxlexblog.wordpress.com/2013/10/01/developments-in-next-generation-sequencing-october-2013-edition/

Received 18 April; accepted 6 June; published online 25 June 2014; doi:10.1038/nbt.2950

NATURE BIOTECHNOLOGY ADVANCE ONLINE PUBLICATION

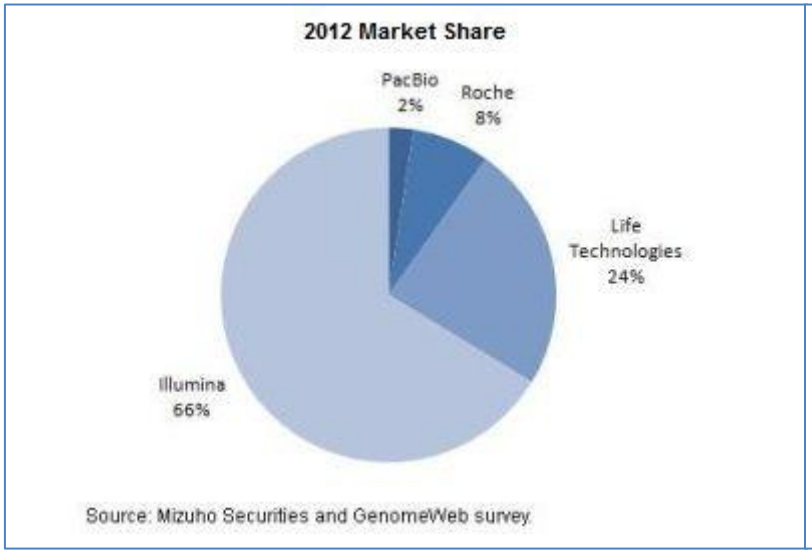
Decoding long nanopore sequencing reads of natural DNA

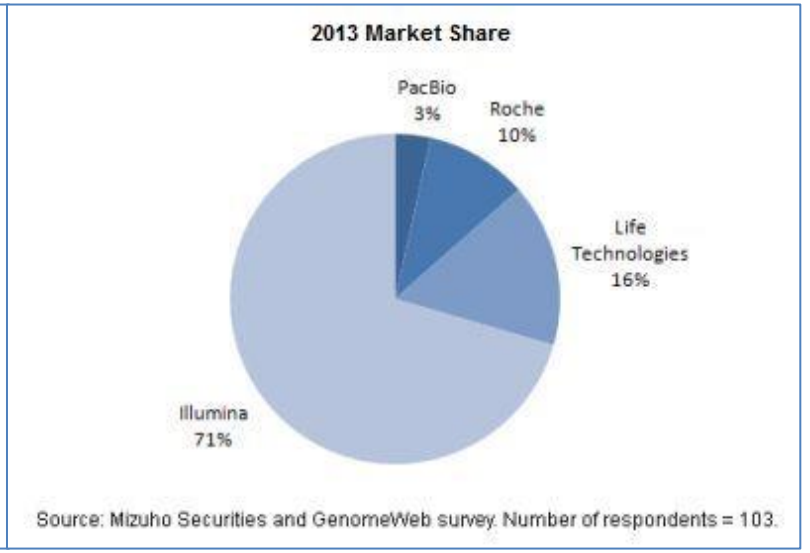
Andrew H Laszlo[1], Ian M Derrington[1], Brian C Ross[1], Henry Brinkerhoff[1], Andrew Adey[2], Ian C Nova[1], Jonathan M Craig[1], Kyle W Langford[1], Jenny Mae Samson[1], Riza Daza[2], Kenji Doering[1], Jay Shendure[2] & Jens H Gundlach[1]



Each of the 256 possible quadromers has a unique current value

In Sequence 2013 Survey: Illumina Pulls Further Ahead



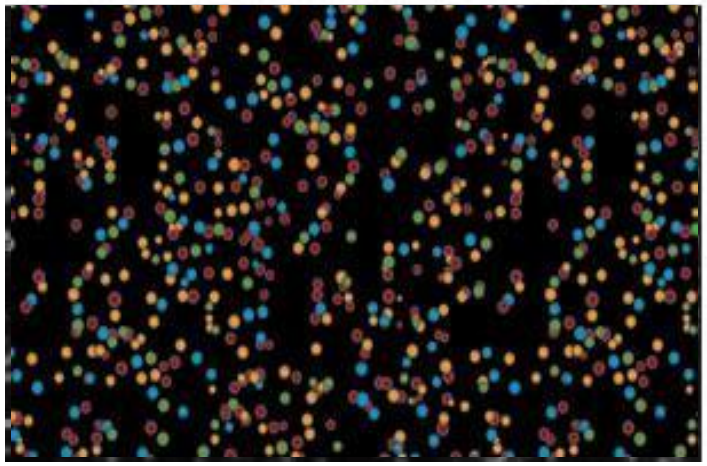


http://www.genomeweb.com/sequencing/sequence-2013-survey-illumina-pulls-further-ahead-interest-oxford-nanopore-remai

Illumina Sequencing

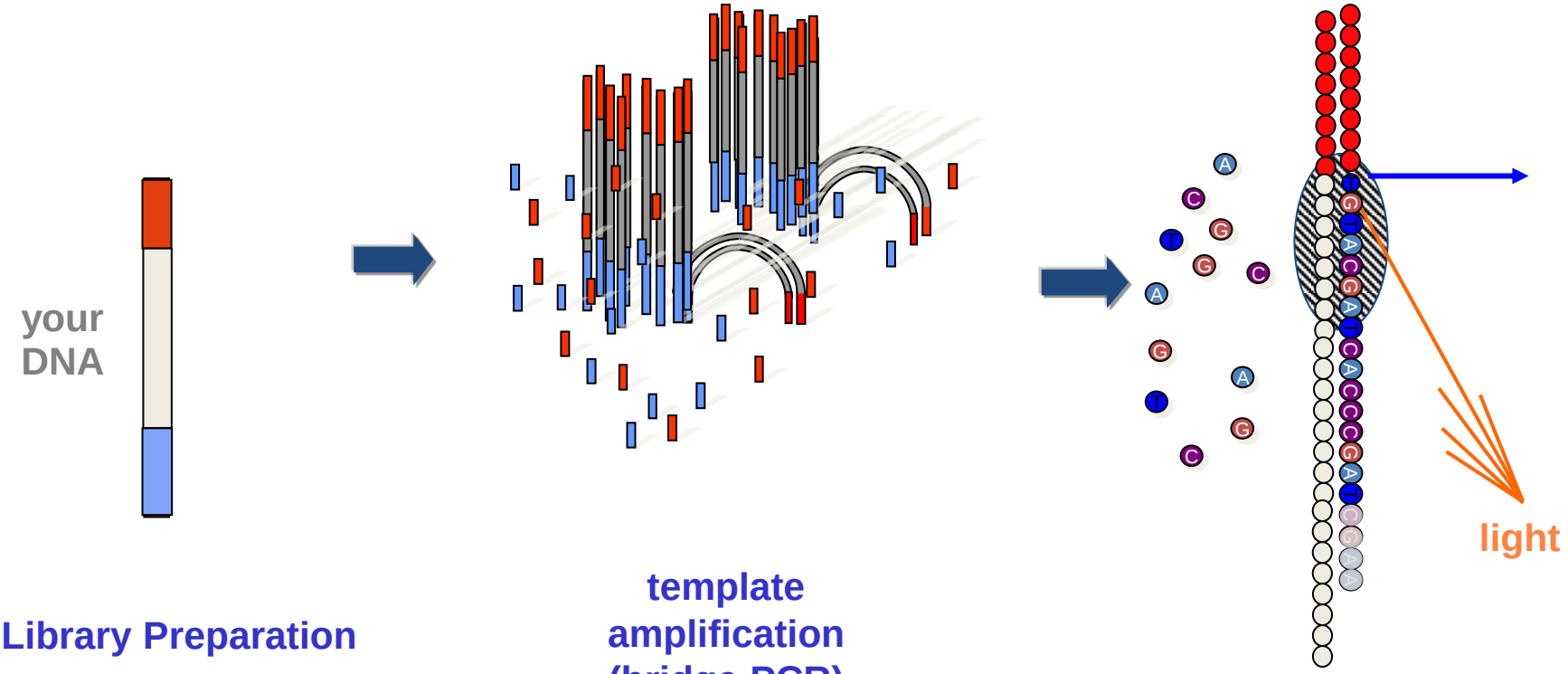
released in 2007 by Solexa

Template Type	Sequencing Method	Imaging Method
Clonally amplified by solid phase amplification	Sequencing by synthesis with cyclic reversible termination	Four color imaging of single events using fluorescence



http://www.illumina.com/systems/hiseq_2000.ilmn

Illumina Sequencing



Library Preparation

template
amplification
(bridge PCR)

seq-by-synthesis

input:

- **1 ng – 1ug DNA**
- **1 ug total RNA**

“Classic” Illumina Library Prep

genomic DNA, RNA
BACs, amplicons ...

DNA fragmentation

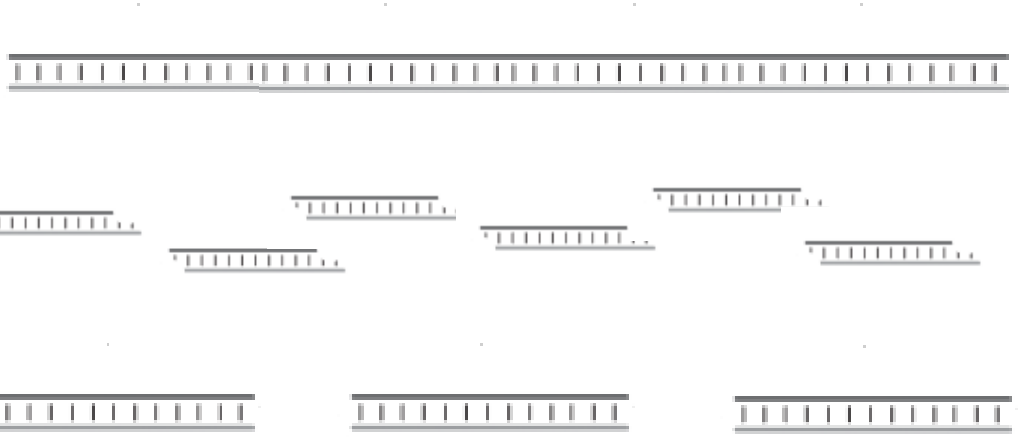
end-repair

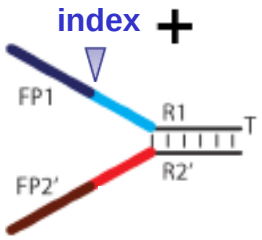
A-tailing

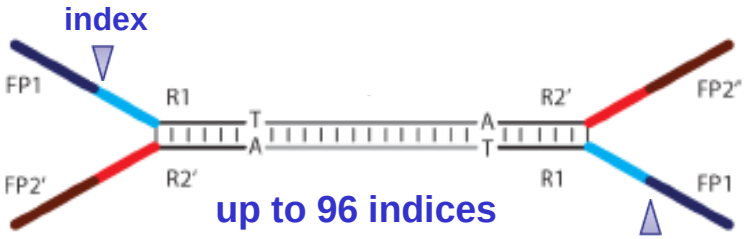
adapter-ligation

size selection

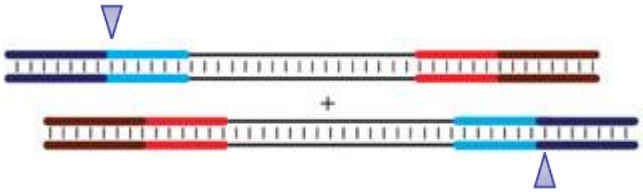
enrichment

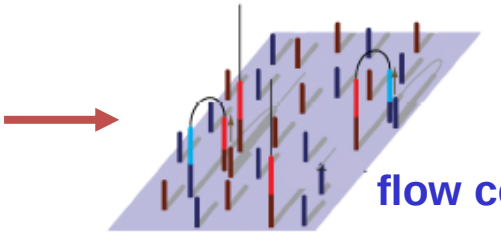






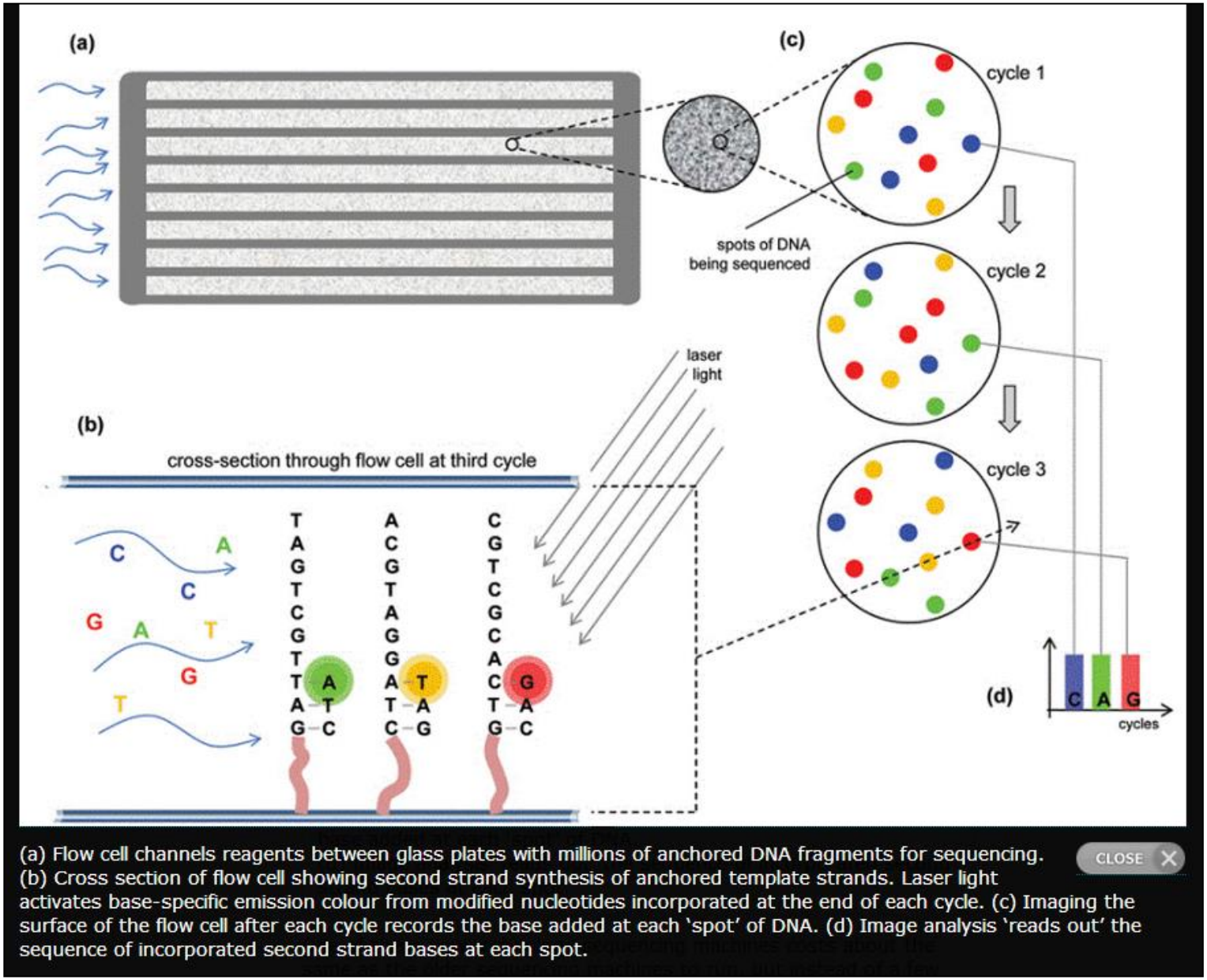
e.g. 200-400 bp





Illumina Sequencing

1 lane = 200 million reads



(a) Flow cell channels reagents between glass plates with millions of anchored DNA fragments for sequencing. (b) Cross section of flow cell showing second strand synthesis of anchored template strands. Laser light activates base-specific emission colour from modified nucleotides incorporated at the end of each cycle. (c) Imaging the surface of the flow cell after each cycle records the base added at each 'spot' of DNA. (d) Image analysis 'reads out' the sequence of incorporated second strand bases at each spot.

CLOSE

Illumina Sequencing

HiSeq	
Run time	6-12 days
Read length	125+125 bp
Yield/lane	60 Gbp

MiSeq	
Run time	1-3 days
Read length	300+300 bp
Yield/lane	10-15 Gbp

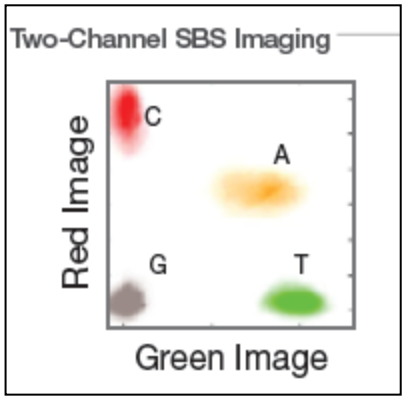
New Illumina Sequencing Platforms Announced January, 2014

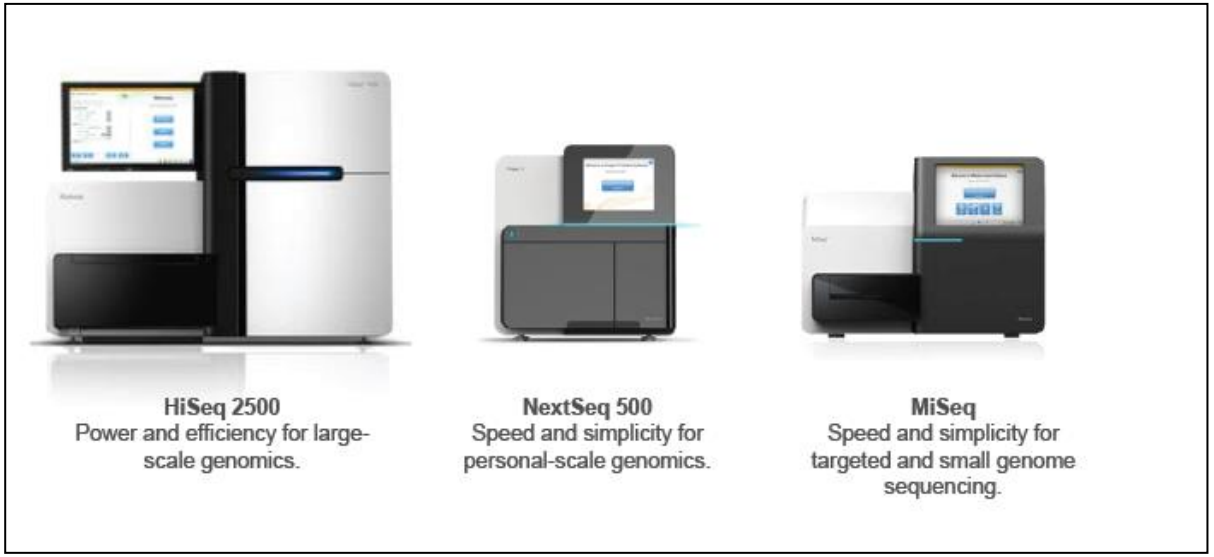
“Factory” Scale

HiSeq X — Patterned Flow Cells (billions of nanowells)
1.8 Tb/ 3 days — Only Human Genome Sequencing Supported!
150 bp paired ends — Sold in Sets of 10...

Desktop Model

NextSeq 500 — Two-Channel Base Detection (faster run time)
up to 400 million reads — 75 or 150 bp single or paired ends





NGS Error Rates

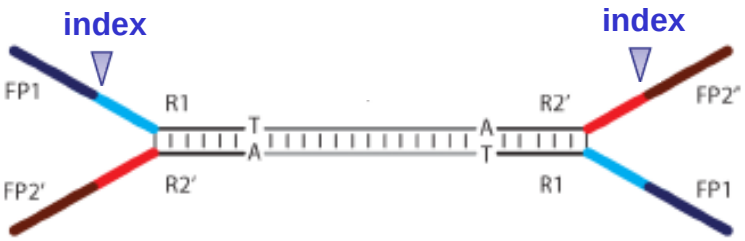
2014 NGS Field Guide. Travis Glenn. www.molecularecologist.com

Platform	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)	Notes
3730xl (capillary)	Substitution	0.1-1	0.1-1	
454	Indel	1	1	
Illumina	Substitution	~0.1	~0.1	≥ 75-85% of reads
SOLiD	A-T bias	~5	≤0.1	2x-3x sequencing
Ion Torrent	Indel	~1	~1	0.46-2.4%
PacBio RS	Indel	~13	≤1	consensus of 3 reads
Oxford Nanopore	Deletion	≥4	4	press release only

How to Deal with Overcapacity

Multiplexing

Addition of a unique sequence identifier (barcode or index) allowing multiple samples to be run together on a single flow cell lane.



Internal indexes (Cronn et al 2008)

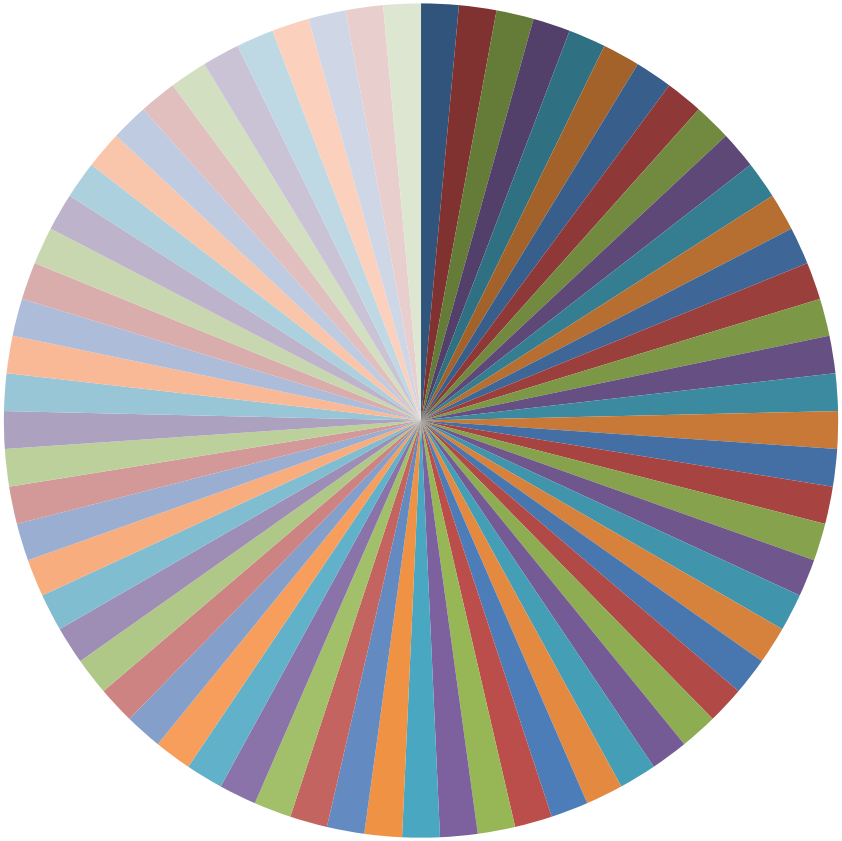
External indexes (extra round of sequencing)

- Illumina 24 single; 96 dual
- NEB 24 single, 96 dual
- Nextflex 48 single, 192 dual
- NuGen 384 dual

Multiplexing Example

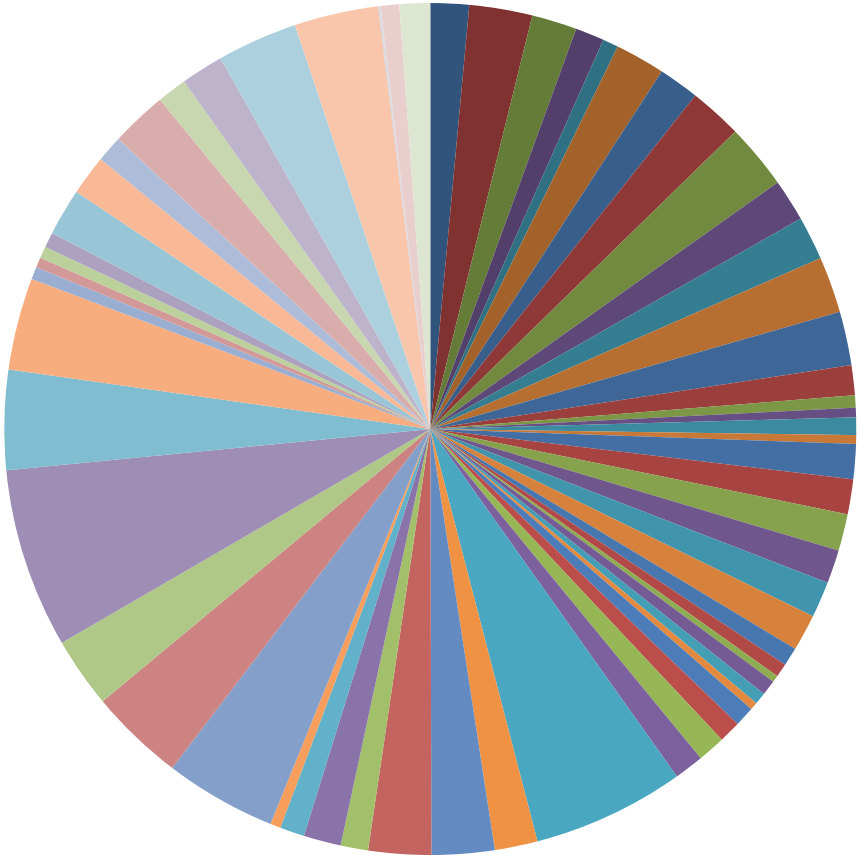
MiSeq v.3 November, 2013
20.3 million paired end reads
69 *Fragaria virginiana* BACs

Ideal Equimolar Results



average = 295,000

Actual Results



1200-13.8 million reads
median = 236,000

Outsourcing Library Prep

1. Fairly standardized (although diverse library options exist)
2. Expensive equipment (sonication, quantification)
3. Automation leading to lower prices

Botany 2014 Next Generation Sequencing Workshop

1. Sequencing Technology
2. Genome Reduction
3. Sequence Processing
4. Bionformatics
5. What you really need to know.

Whole Genome Sequencing (not covered here)

Genome Reduction

A. PCR, fosmids, BACs
B. Low Coverage (=Genome Skimming)
C. Restriction Digest Methods (GBS, RAD, etc.)
D. Target Capture (Hyb-Seq)
E. Transcriptome (RNA-Seq)

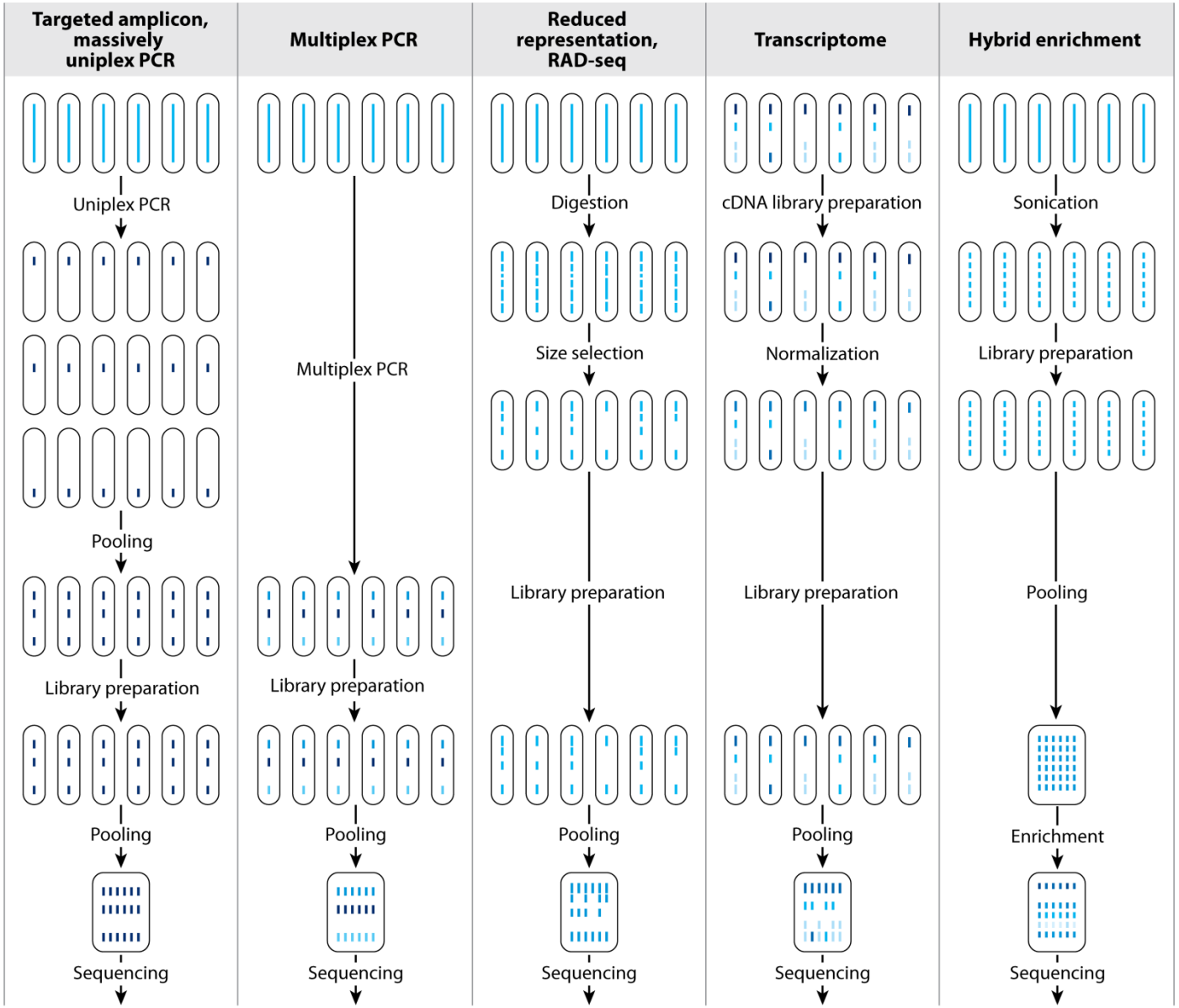
Cronn et al. 2012. Targeted enrichment strategies for next-generation plant biology. American Journal of Botany 99:291-311.

Straub et al. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. American Journal of Botany 99:349-364.

Lemmon & Lemmon. 2013. High-throughput genomic data in systematics and phylogenetics. Annual Review of Ecology, Evolution, and Systematics 44:99-121.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. Trends Ecol. Evol. 29:51–63.

Genome Reduction Approaches



Lemmon EM, Lemmon AR. 2013.
Annu. Rev. Ecol. Evol. Syst. 44:99–121

A. Amplicons and BACs

= genome reduction prior to NGS library preparation

1. Amplicons **Fluidigm automation**

Uribe-Convers et al 2014. A Long PCR–Based Approach...
Appl. Plant Sci. 2:1300063.

48 amplicons X 48 samples = $600

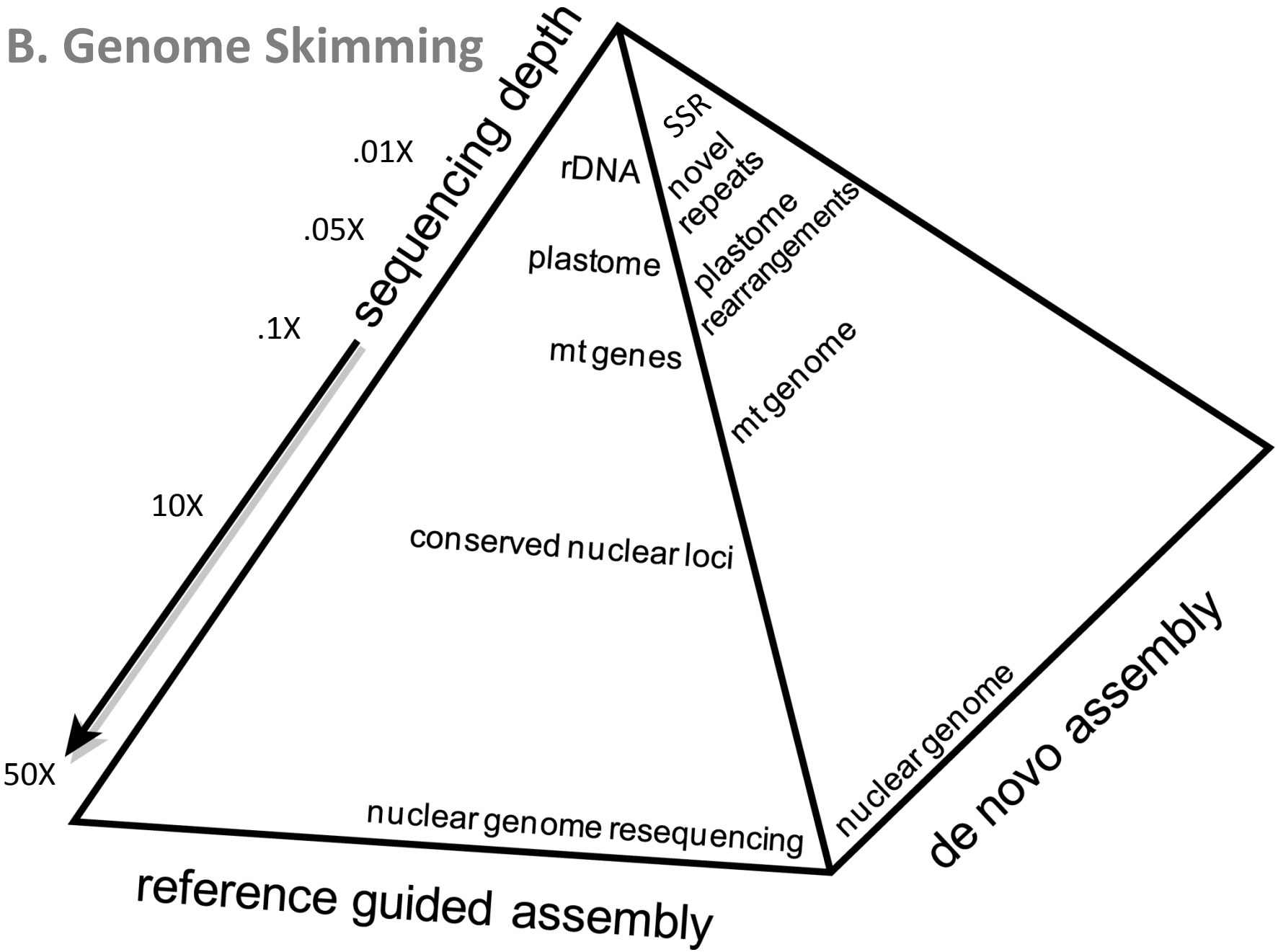
1000+ can be sequenced on on MiSeq run = $1500

2. **Fosmids (10-40 kbp) and BACs (100-150 kbp)**

Relatively expensive ($10,000 for a BAC library)

Standard approach in large (well-funded) genome projects.

B. Genome Skimming

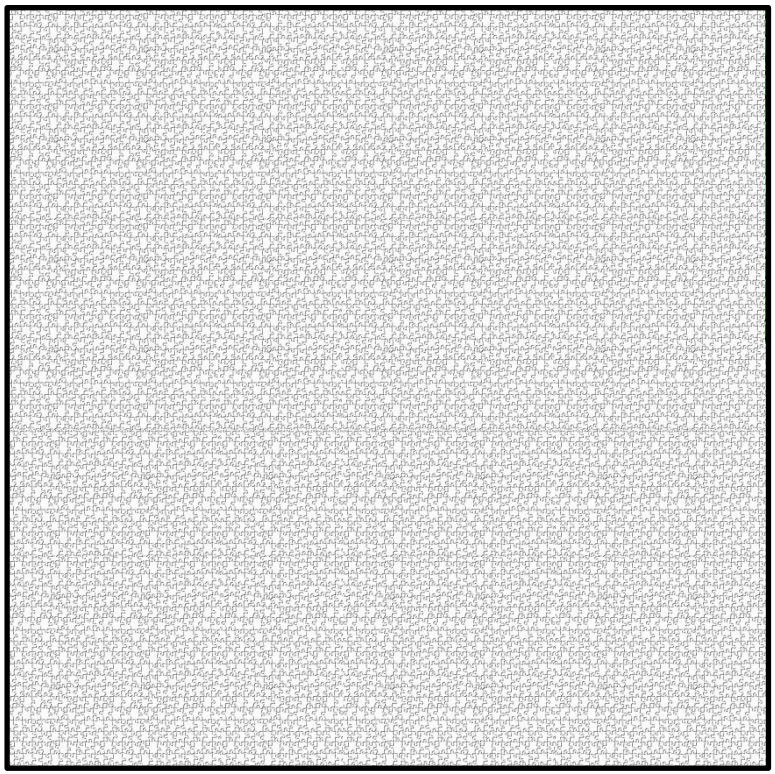


Straub et al. 2012. American Journal of Botany 99:349-364.

Genome Assembly

Reference guided assembly

De novo assembly

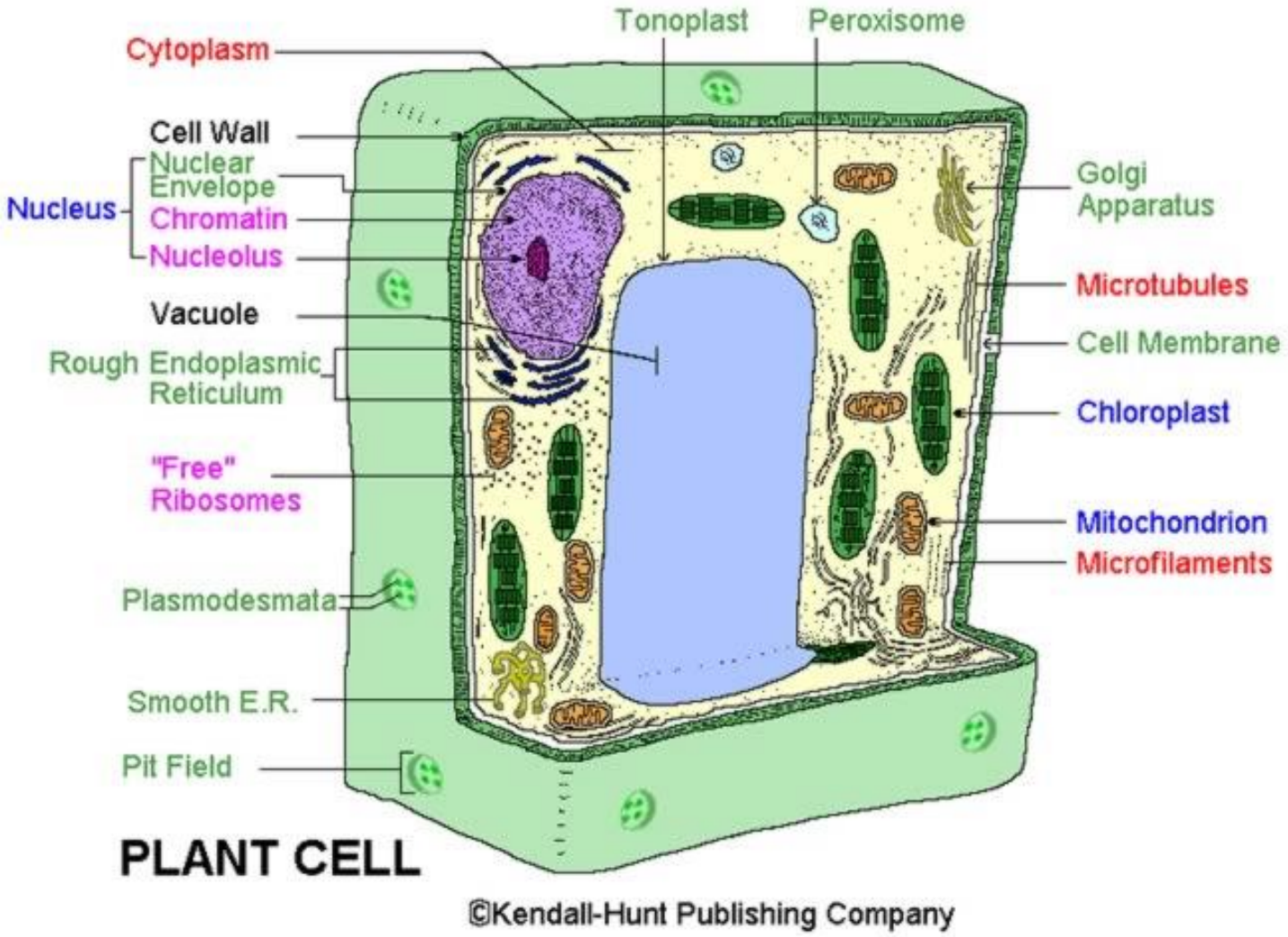


Genome Coverage

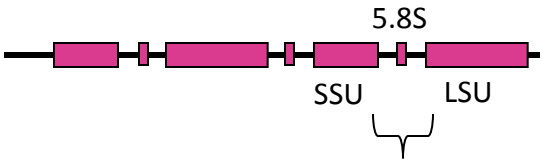
Low Coverage

High Coverage

Plant Cells


Tonoplast
Peroxisome
Cytoplasm
Cell Wall
Nucleus
Nuclear Envelope
Chromatin
Nucleolus
Vacuole
Rough Endoplasmic Reticulum
"Free" Ribosomes
Plasmodesmata
Smooth E.R.
Pit Field
Golgi Apparatus
Microtubules
Cell Membrane
Chloroplast
Mitochondrion
Microfilaments
PLANT CELL
©Kendall-Hunt Publishing Company

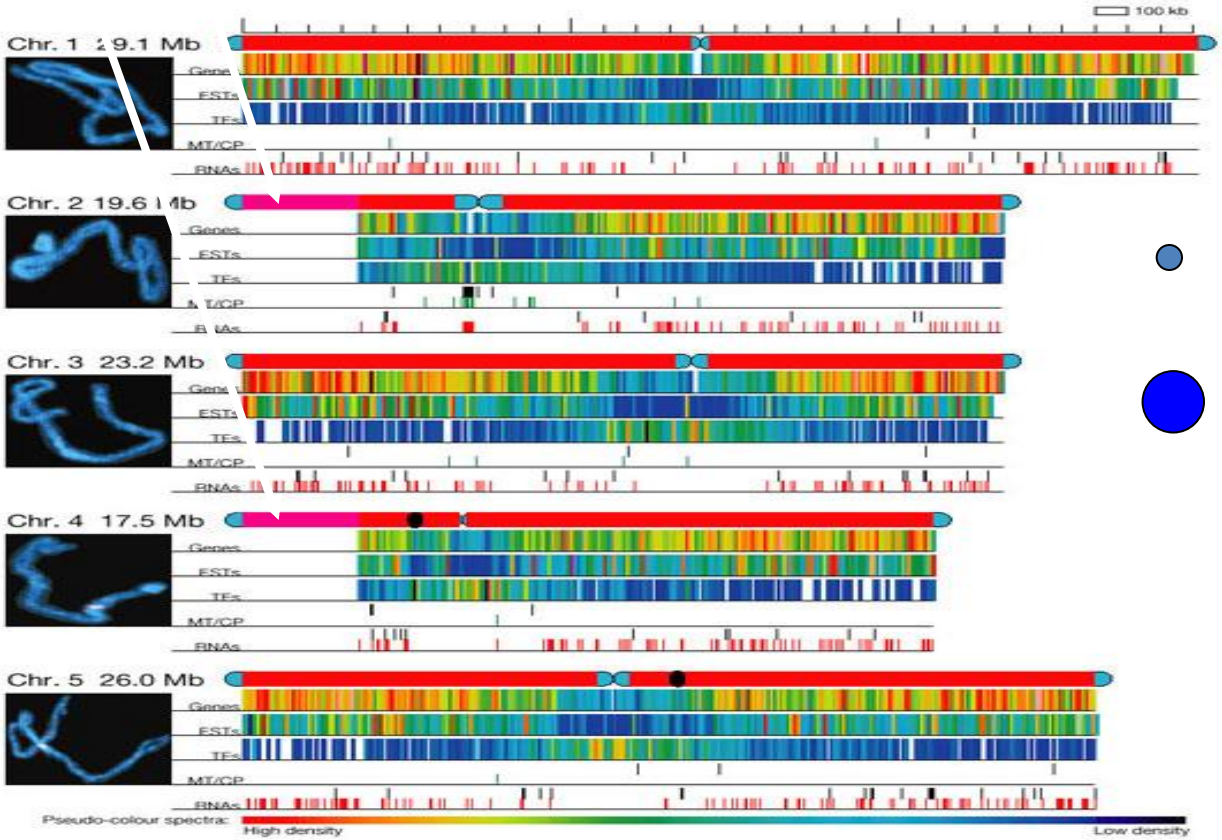
Plant Genomes



Ribosomal DNA

Internal transcribed spacer = nrDNA ITS
651 bp

Arabidopsis thaliana

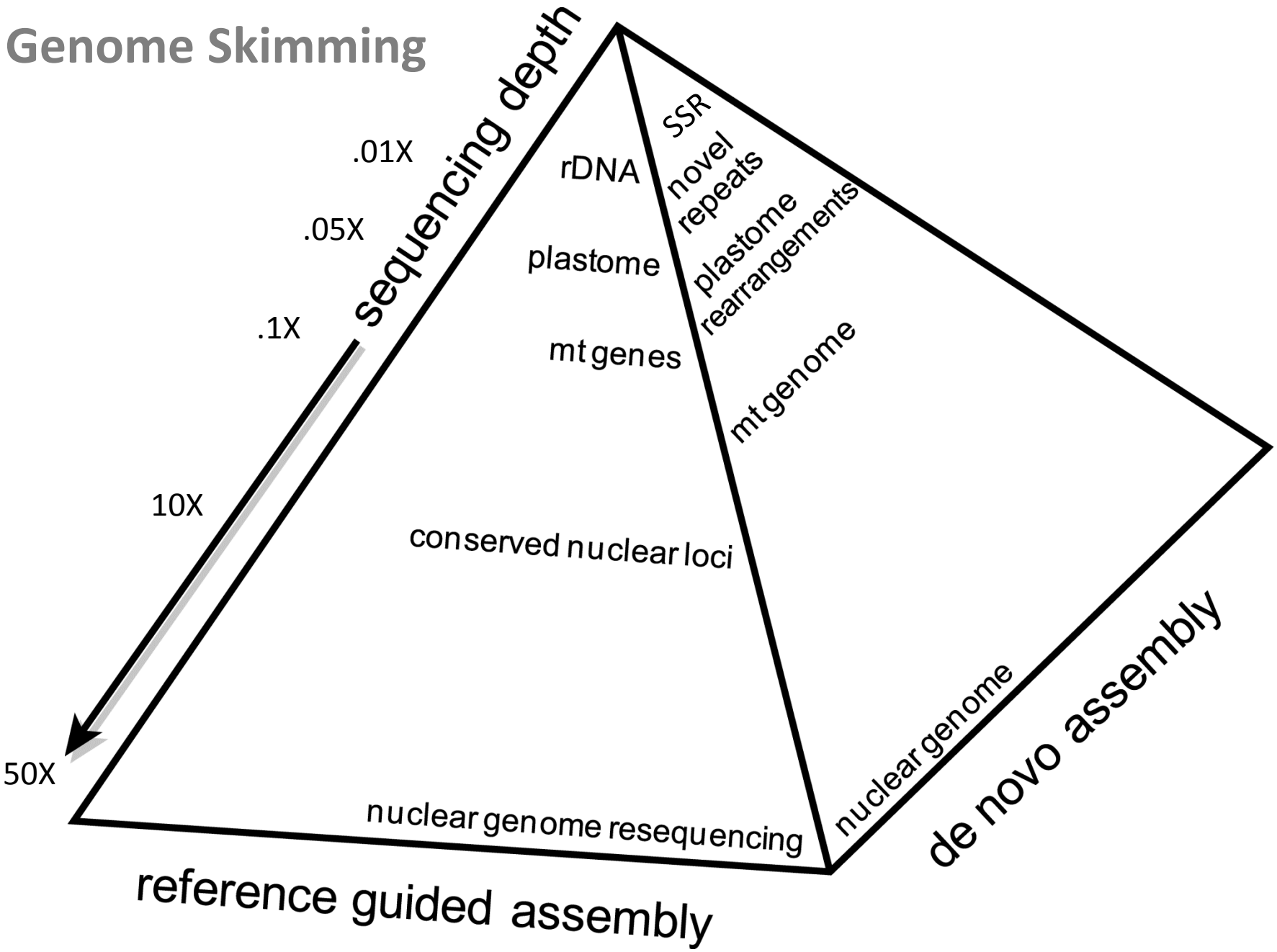


Chloroplast genome 154 kbp

Mitochondrial genome 367 kbp

Arabidopsis Genome Initiative. 2000. *Nature* 408:796-815.

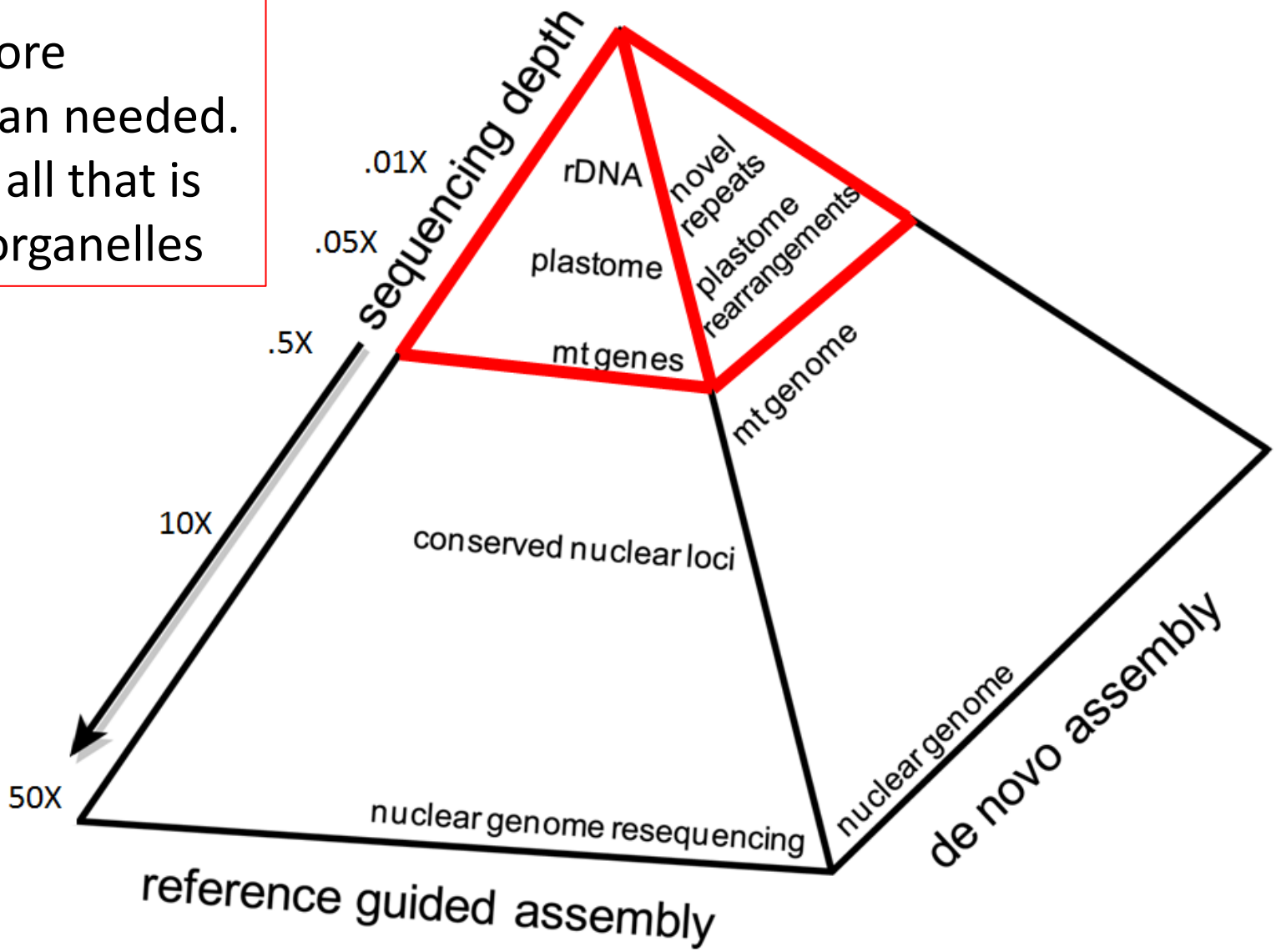
Genome Skimming



Straub et al. 2012. American Journal of Botany 99:349-364.

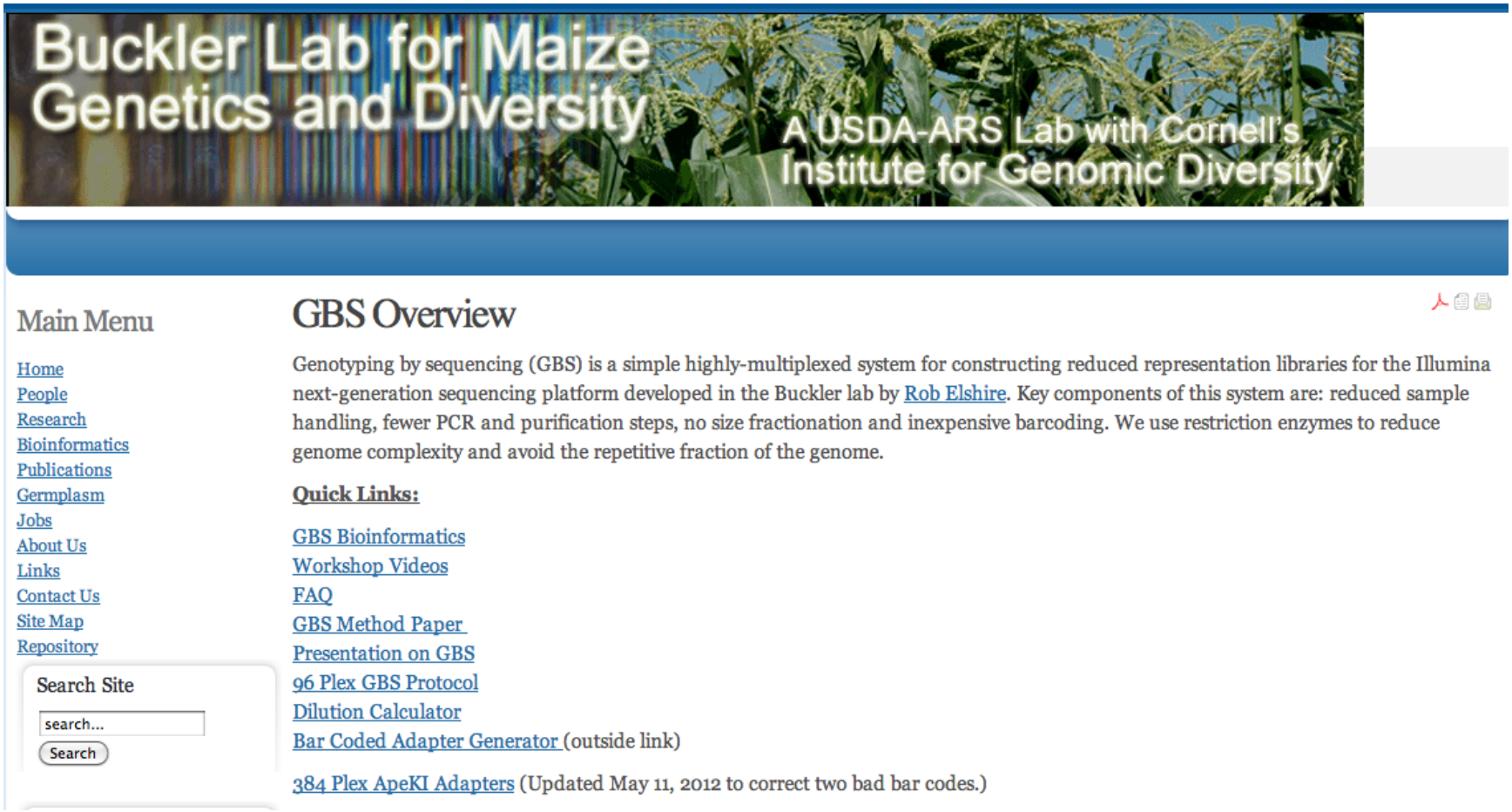
Genome Skimming

Important:
Don't use more sequence than needed.
100-125× is all that is needed for organelles



C. Restriction Digest Approaches (RAD-Seq, GBS, 2b-RAD)

http://www.maizegenetics.net/



Advantages of Restriction Digest approaches

1. Can obtain thousands of SNPs without a reference genome

2. Inexpensive library prep = hundreds to thousands of individuals

3. Discovery of candidate loci associated with phenotypic traits (quantitative trait loci = QTL) in natural populations (genome wide association studies = GWAS)

4. Analytical pipelines available

5. Hundreds of studies published to date.

Why I Don't Like Restriction Digest approaches

1. Generally requires at least 1 µg of good quality DNA.

2. Restriction digestion adds another variable to the library prep.
   a. Restriction site polymorphism results in missing data.
   b. Short fragments are cut less efficiently.

3. Does not target specific genes or SNPs.

4. References are easily obtained, and soon to be widely available.

5. Loci are not transferrable among species.

6. High potential for ascertainment bias.

Restriction Site Bias

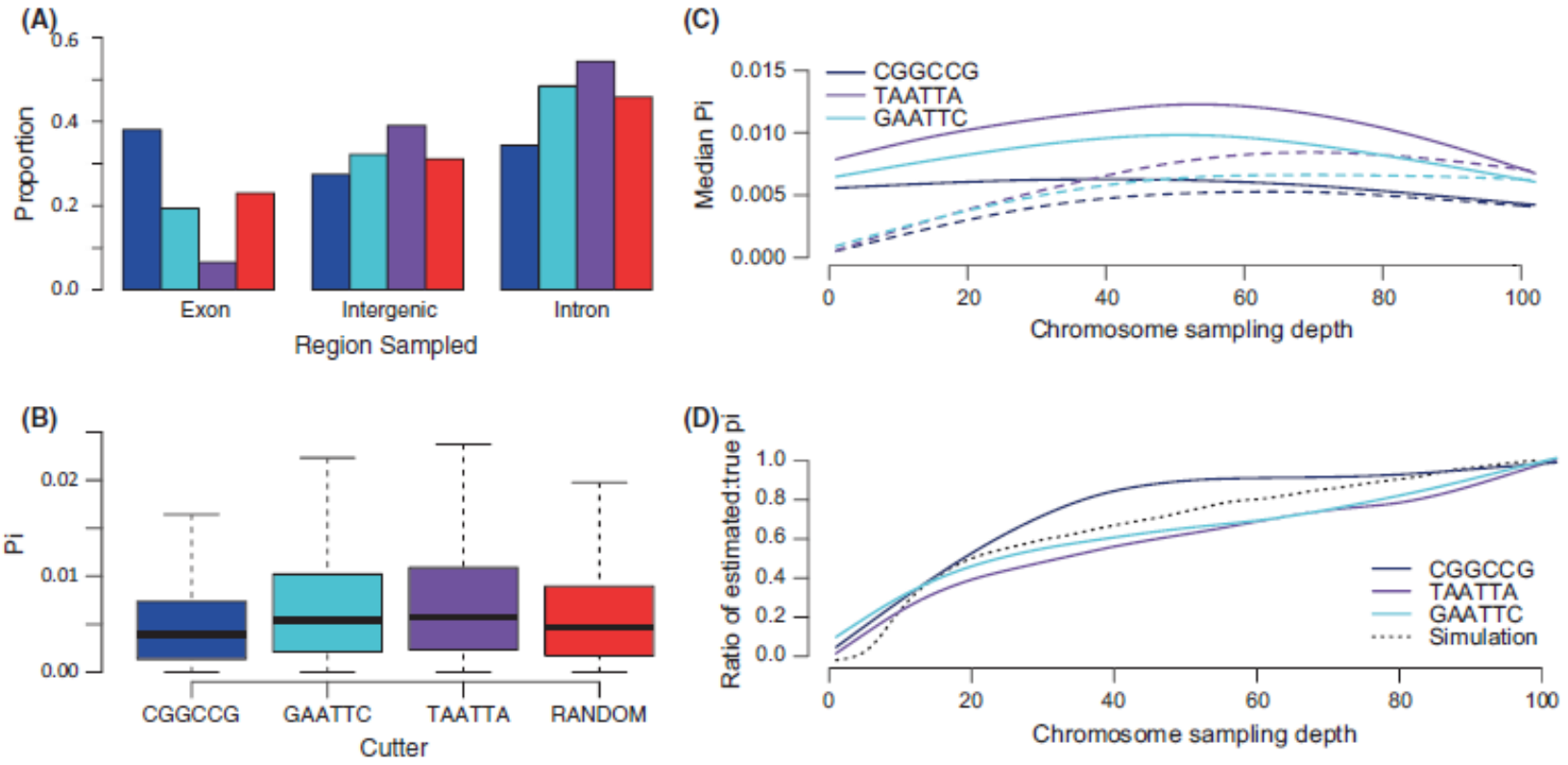
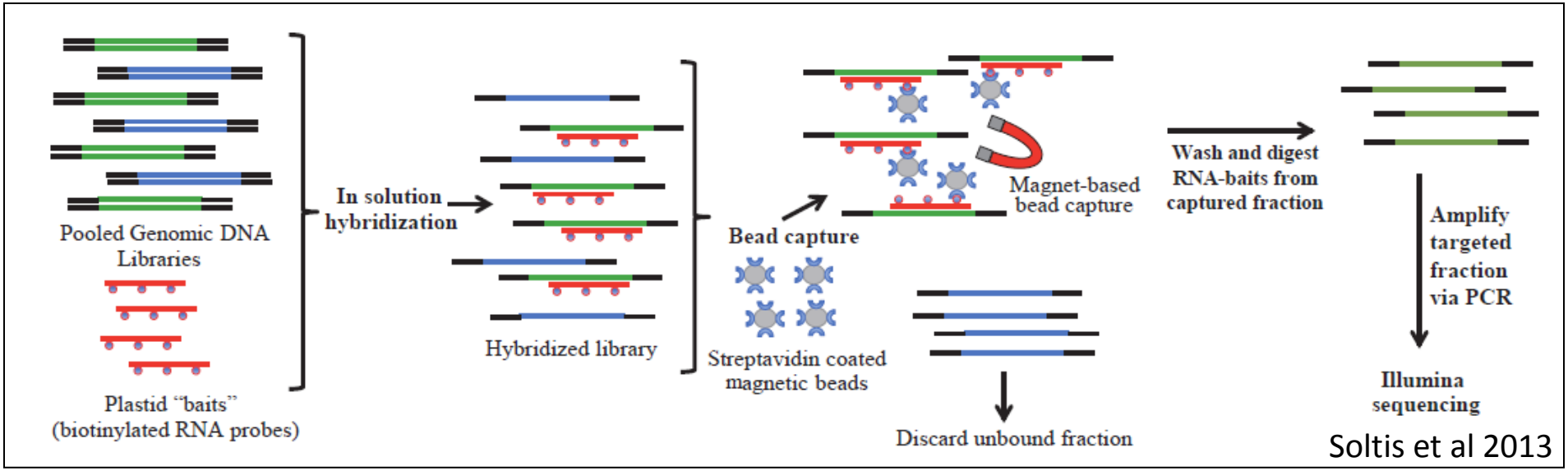


Fig. 8 Results for the *in silico* digests 102 *Drosophila melanogaster* genomes. (A) Proportion of sites located in distinct regions of the genome when *in silico* digests are performed with different enzyme recognition sequences. GC-rich recognition sequences sample more exons, whereas AT-rich recognition sequences sample comparatively more introns and intergenic regions. 'Random' values are calculated from fragments selected at random throughout the genome. (B) Box plots of true π for regions sampled by enzymes with different recognition sequences. (C) The median true π (solid line) and estimated π (dashed line) as a function of chromosome sampling depth for three different recognition sequences. (D) Median of the ratio of estimated π to true π as a function of the number of sampled chromosomes. Dark blue, purple and cyan lines represent the three different restriction enzymes used in the *in silico* digest of the *D. melanogaster* genomes, and the dotted black line is from simulations with $\rho = 0.1$ per bp, $\theta = 0.01$ per bp

Arnold et al. 2013 Molecular Ecology 22:3179-3190

D. Targeted Sequence Capture = Hyb-Seq



Soltis et al 2013

12-24 individuals per hybridization

48 individuals per MiSeq lane

384 individuals per HiSeq lane

$50-$75 per individual

0.45-2 Mbp / individual

=250-1000 nuclear genes
or 6500-20,000 SNPs

~50% efficiency
= 50% off-target genome skimming

Weitemier et al. 2014 APPS, in press.

Advantages of Hyb-Seq

1. A single laboratory procedure and bioinformatics pipeline can be used for phylogenetics (deep and shallow), population genetics and genetic linkage mapping.

2. A relatively distant (e.g. plant family) reference can be used.

3. Candidate genes can be targeted.

4. Can be scaled from 250 genes to an entire exome (25,000 - 30,000 genes).

5. Data sets can be easily combined and extended (in contrast to GBS, SNP chips).

6. Minor potential for ascertainment bias.

Why not sequence the same genes?

BMC Genomics

BioMed Central
The Open Access Publisher

Identification, characterization, and utilization of single copy genes in 29 angiosperm genomes

BMC Genomics 2014, **15**:504 doi:10.1186/1471-2164-15-504

Fengming Han (hanfengming2004@163.com)
Yong Peng (ypeng@implad.ac.cn)
Lijia Xu (ljxu@implad.ac.cn)
Peigen Xiao (xiaopg@public.bta.net.cn)

3000-11000 (8-35%) single copy genes in 29 angiosperms

12 single copy genes shared across all 29 species

Genome Reduction

A. PCR, fosmids, BACs
B. Low Coverage (=Genome Skimming)
C. Restriction Digest Methods (GBS, RAD, etc.)
D. Target Capture (Hyb-Seq)
E. Transcriptome (RNA-Seq)

RNA-SEQ For Everyone

http://rnaseq.uoregon.edu/index.html

Design Experiment
1. Carefully design the experiment.

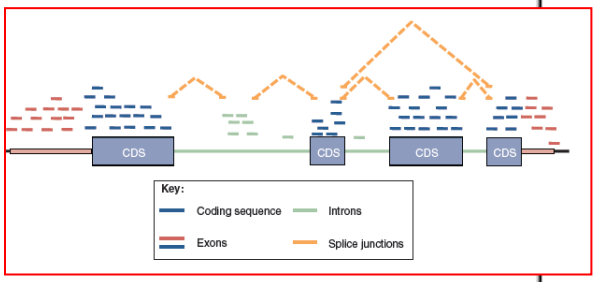
Purify RNA
2. Isolate and purify input RNA.

Prepare Libraries
3. Convert the RNA to cDNA and add sequencing adapters.

Sequence
4. Sequence cDNAs using one of the available NGS platforms.

Analysis
5. Analyze the resulting short-read sequences.

quantify gene expression

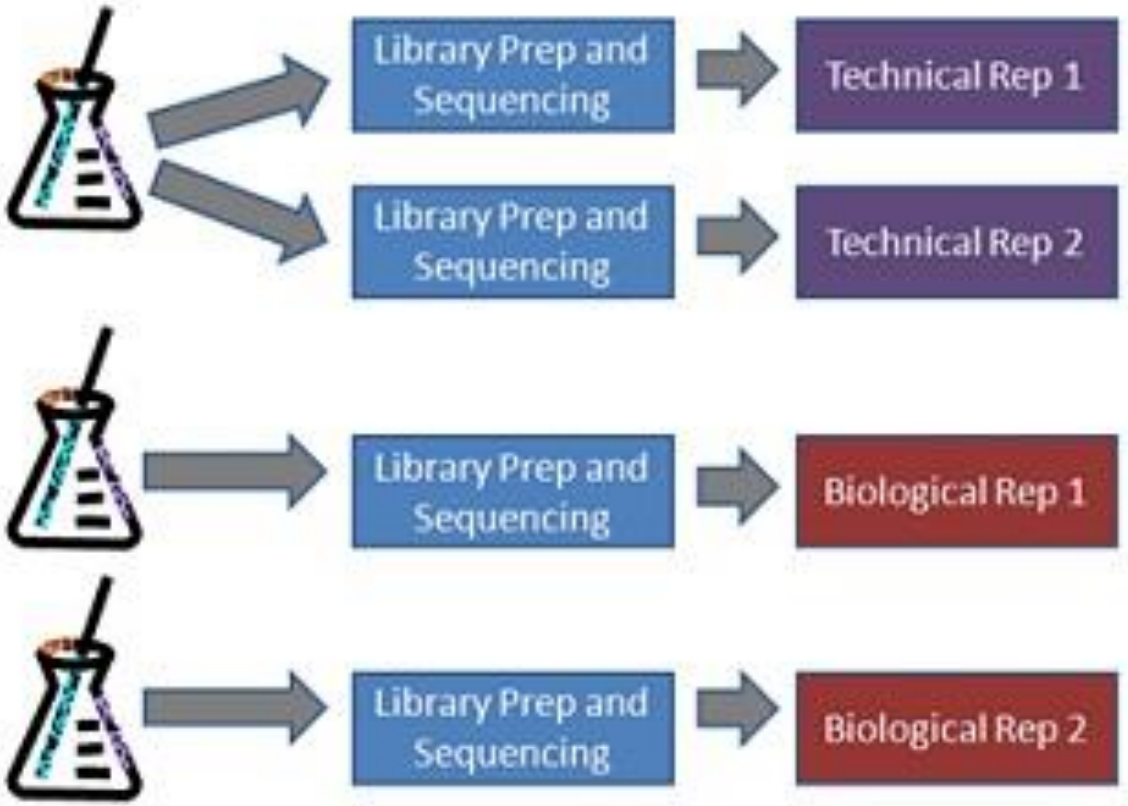


read mapping

transcriptome reference genome reduction

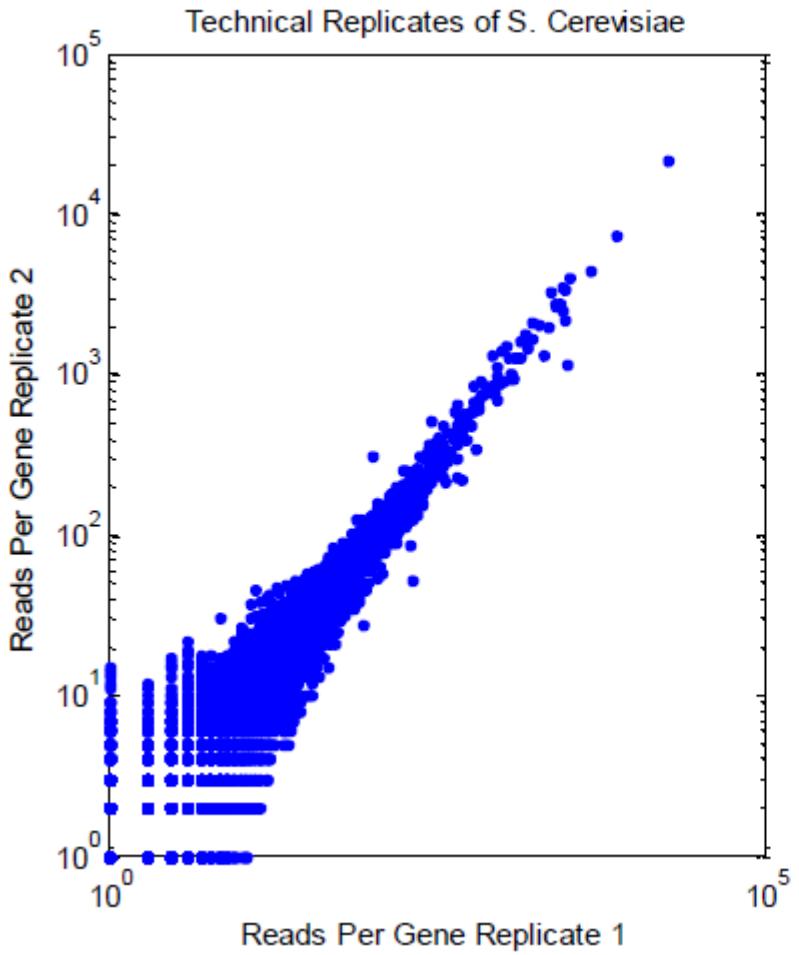
de novo assembly

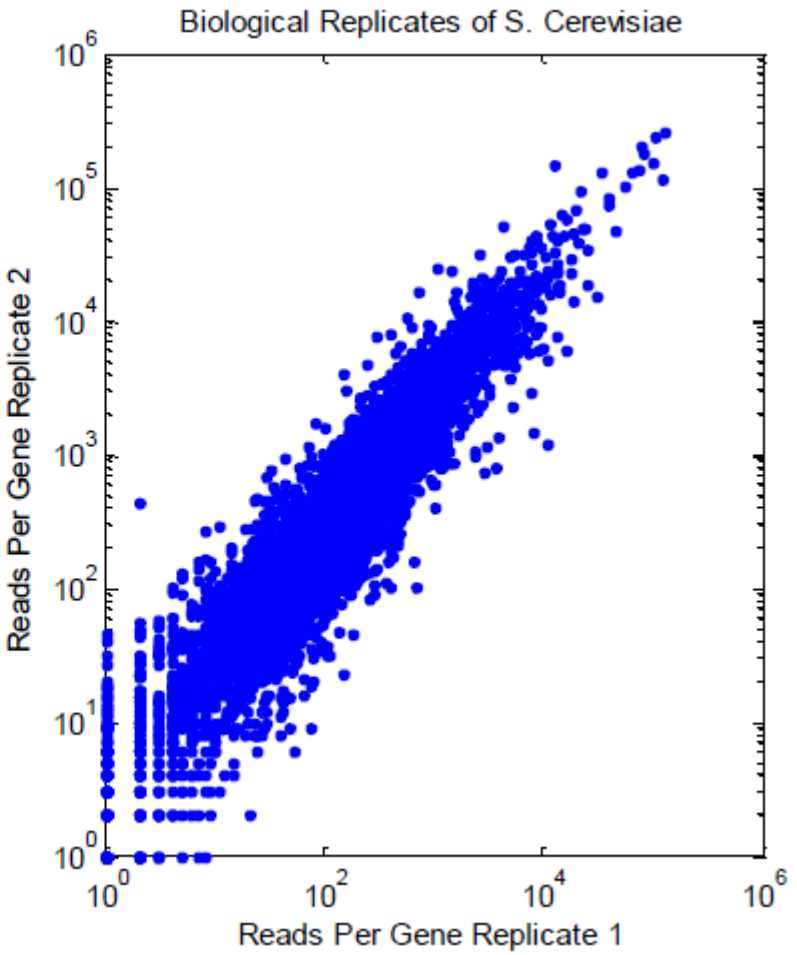
Experimental Design



gkno

RNA-Seq Technical vs. Biological Replicates



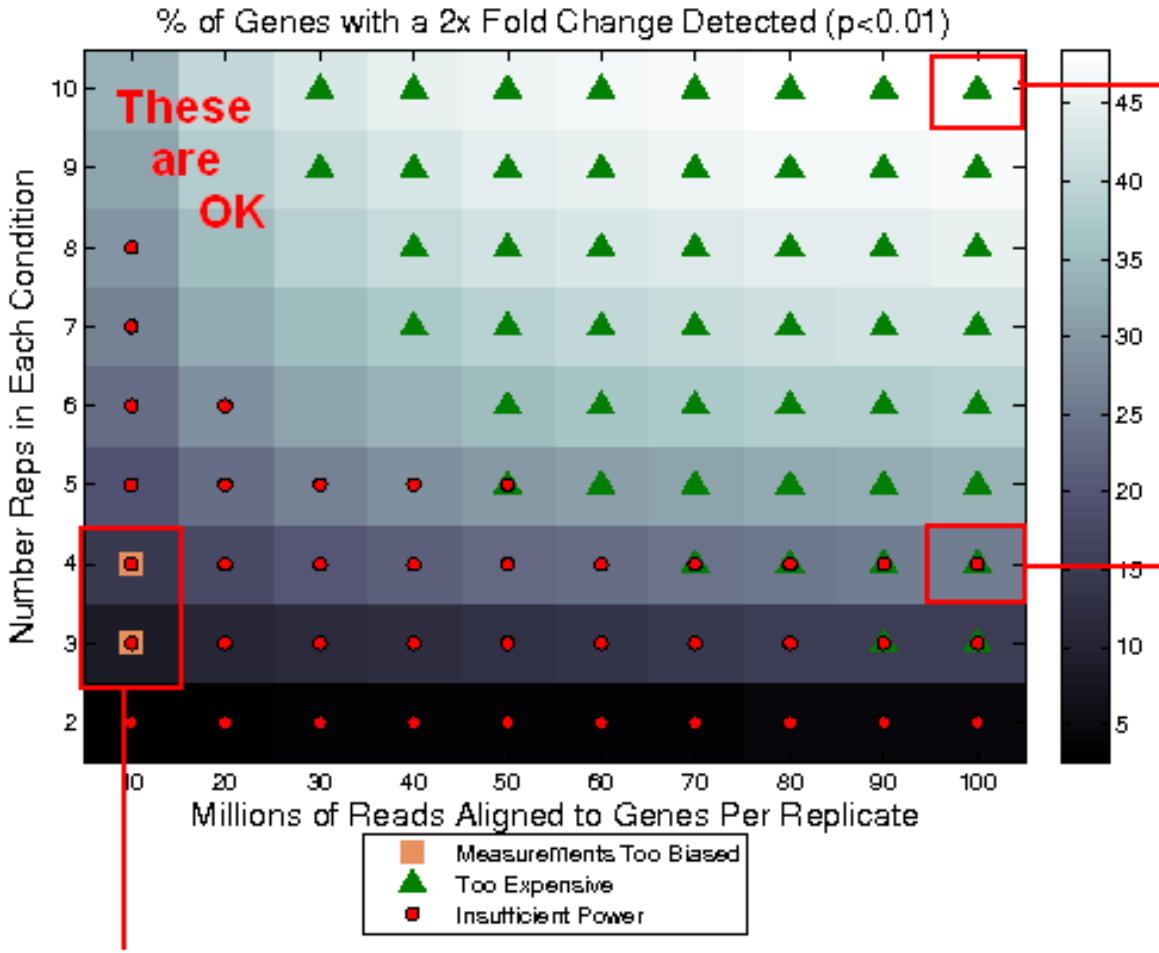


Busby et al. 2011. Expression divergence measured by transcriptome sequencing of four yeast species. BMC Genomics 12: 635.

Scotty - Power Analysis for RNA Seq Experiments

Scotty is a tool to assist in the designing of RNA Seq experiments that have adequate power to detect differential expression at the level required to achieve experimental aims.

http://bioinformatics.bc.edu/marthlab/scotty/scotty.php



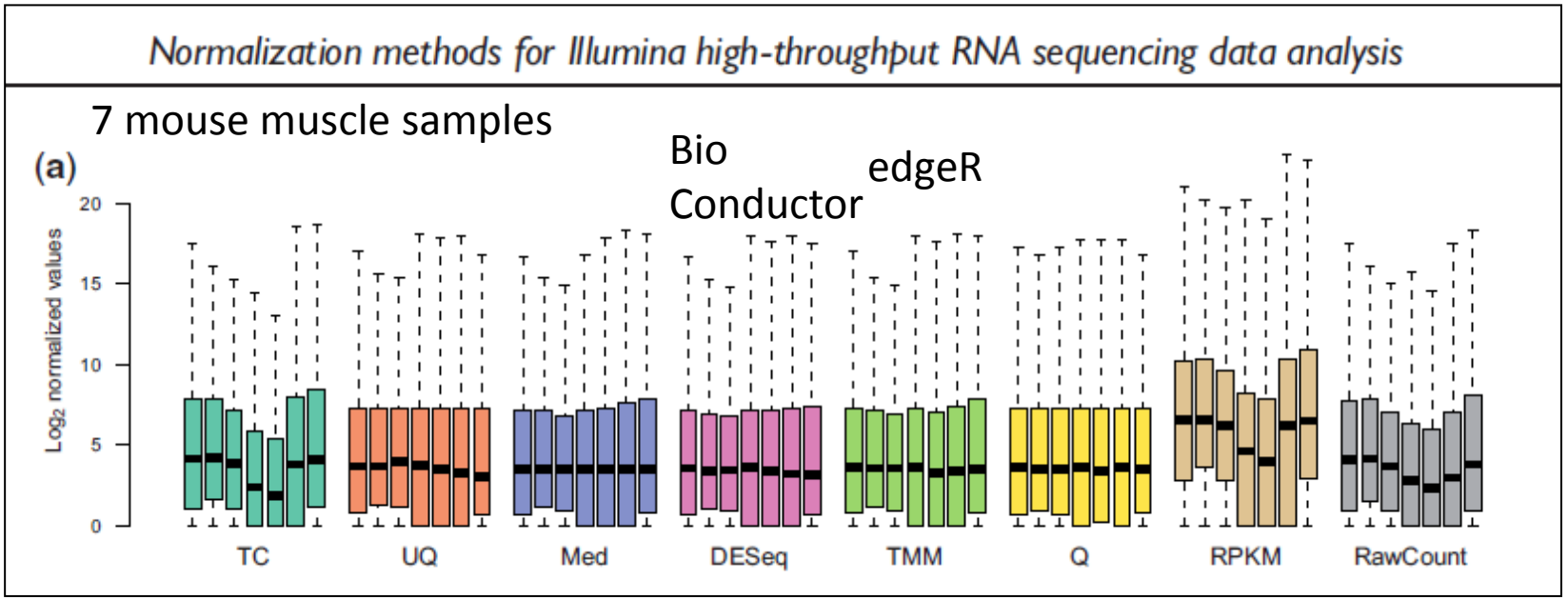
These are OK

This configuration costs too much

This configuration costs too much and has too little power

These have too little power and too much measurement bias

Count Normalization



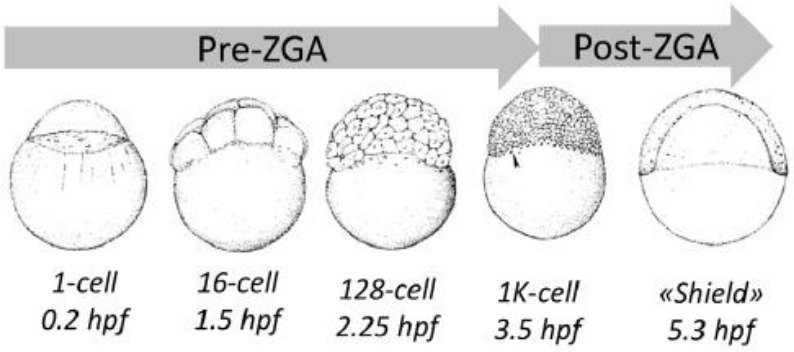
RPKM = reads per kilobase per million reads (most popular method, performs the worst)

All of these methods assume a constant denominator (total RNA per cell)

Dillies et al. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics 14: 671-683.

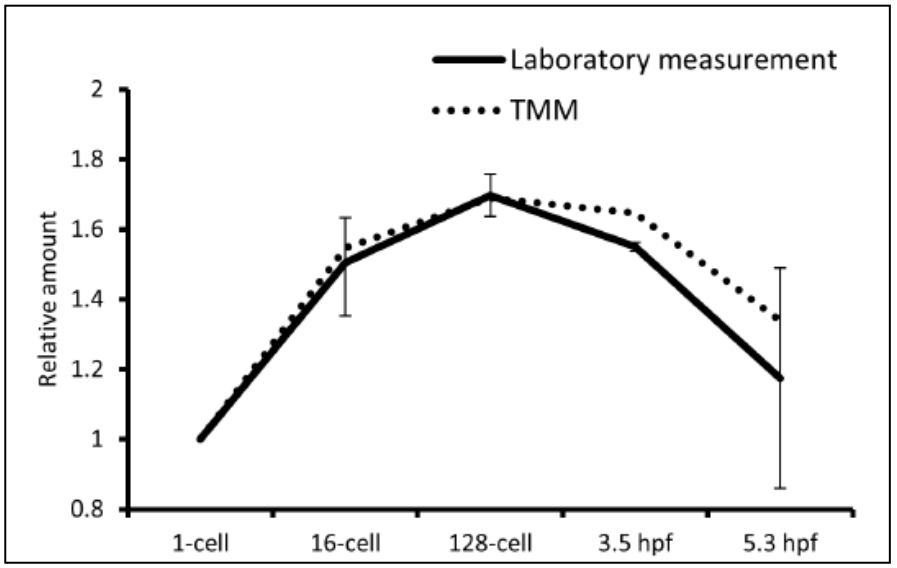
Total RNA Amount Varies Among Samples

different developmental stages
different ploidy levels
(Ilut et al, 2012)



<u>Potential Solutions:</u>
Scale by Total RNA / Cell

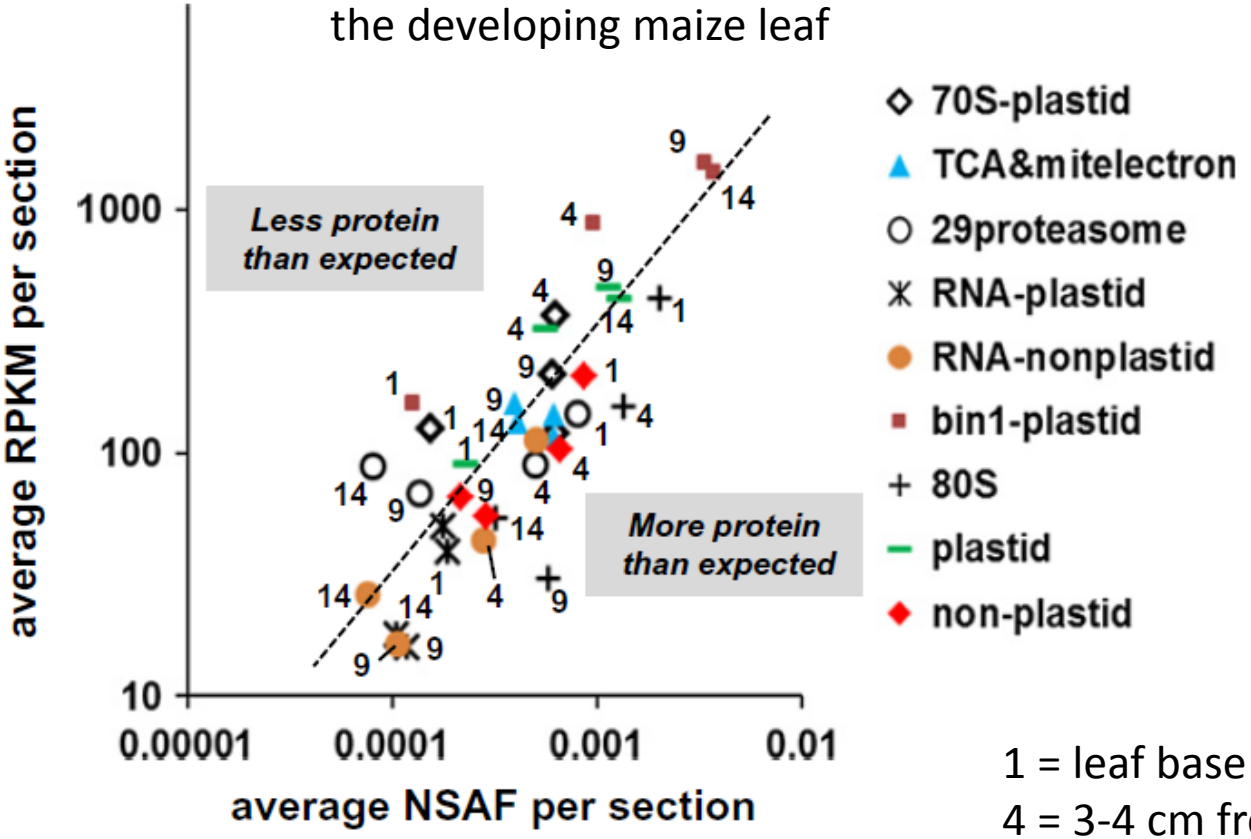
Include Internal RNA Standards



Aanes et al. 2014. Normalization of RNA-Sequencing Data from Samples with Varying mRNA Levels. PLOS ONE 9: e89158.

RNA and proteins not perfectly correlated

Correlation of mRNA and protein abundance in the developing maize leaf



1 = leaf base
4 = 3-4 cm from base
9 = 9-10 cm from base
14 = leaf tip (13-14 cm)

Ponnala et al 2014. The Plant Journal 78:424-440.

RNA-Seq

Pros:

Powerful method with multiple applications:

SNPs, differential expression, genome annotation, phylogenomics, alternative splicing, RNA editing

Cons:

Requires living tissue

RNA extraction is more involved than DNA

RNA expression biases (biological and analytical) are far from understood

Analysis of isoforms and paralogs is challenging

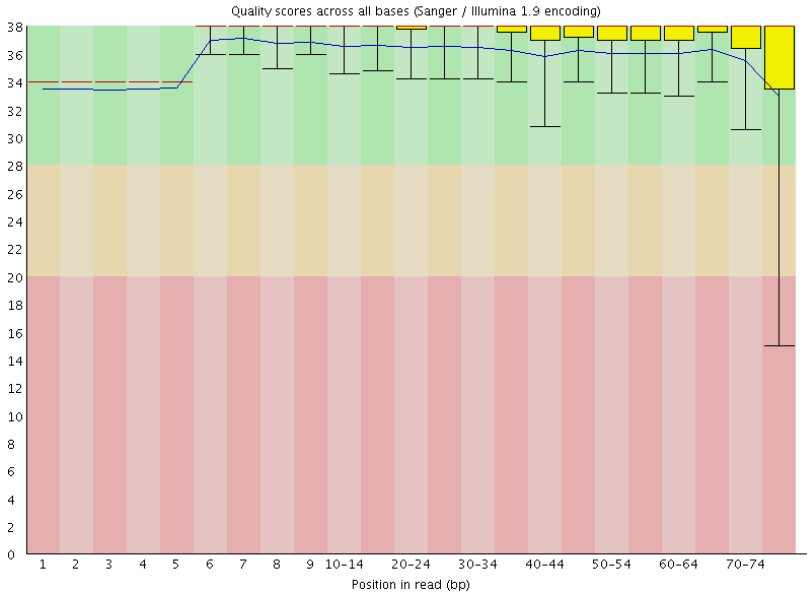
Botany 2014 Next Generation Sequencing Workshop

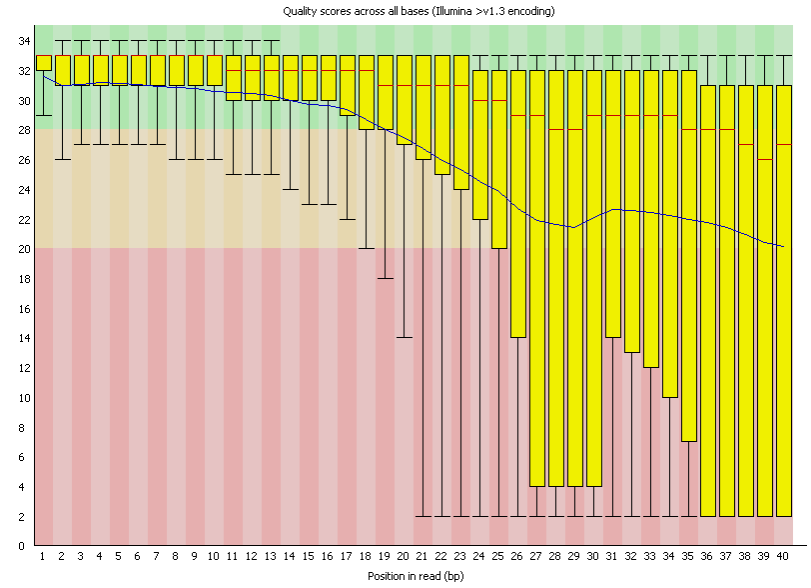
1. Sequencing Technology
2. Genome Reduction
3. Sequence Processing
4. Bionformatics
5. What you really need to know.

Sequence Processing

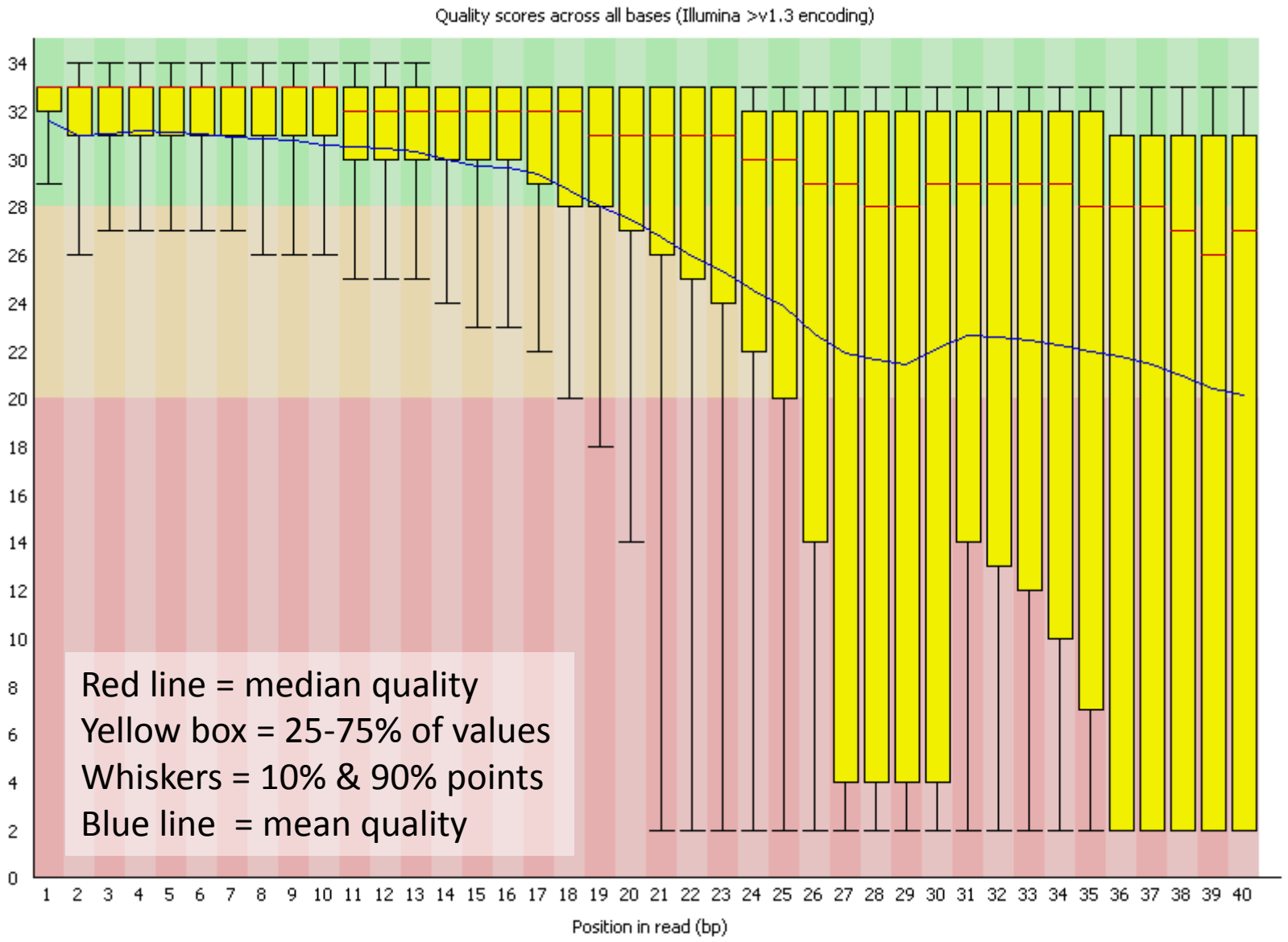
Potential Contaminants and Biases

1. Low Quality Bases
2. Adapters
3. Duplicate Reads
4. Uneven Sequencing Depth
5. Fragment Size Variation
6. Read Overlap
7. Biological Contamination





FASTQC output



Sequence Processing

Quality Assessment Tools:

Illumina BaseSpace https://basespace.illumina.com/runs

FastQC
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

SGA Simpson, 2014. Bioinformatics
uses kmers to estimate:
paired end insert sizes
heterozygosity
repeat content
genome size

A Popular Read Trimming & Adapter Removal Tool:

Trimmomatic (Bolger et al 2014)
http://www.usadellab.org/cms/?page=trimmomatic

- trims bases at start & end of reads, using a threshold quality score
- also can use a sliding window average of base quality to trim
- filters reads below a minimum length after trimming
- keeps track of orphaned pairs
- short adapters can be trimmed in palindrome mode

Phred = 20 (1% error rate) is commonly used
But is this too stringent? (MacManes 2014)

Duplicate Removal

Optical duplicates = split clusters

PCR duplicates from final “enrichment” step of library prep
Best to limit to absolute minimum of cycles (6-12)

Deep sequencing (>500X) will result in “natural duplication”

Should these be removed? Differences of opinion exist.
See Zhou et al. Bioinformatics January 2, 2014.
Yes for SNPs; No for counting [but incorporate into model]

Tools for Duplicate Read Removal:
Picard (picard.sourceforge.net; requires mapping)
FastUniq (Xu et al 2012; requires paired sequences)
fastq_collapse.py (Weitemier, unpublished)

Sequence Processing

k-mer = substrings of length *k* in DNA sequence data

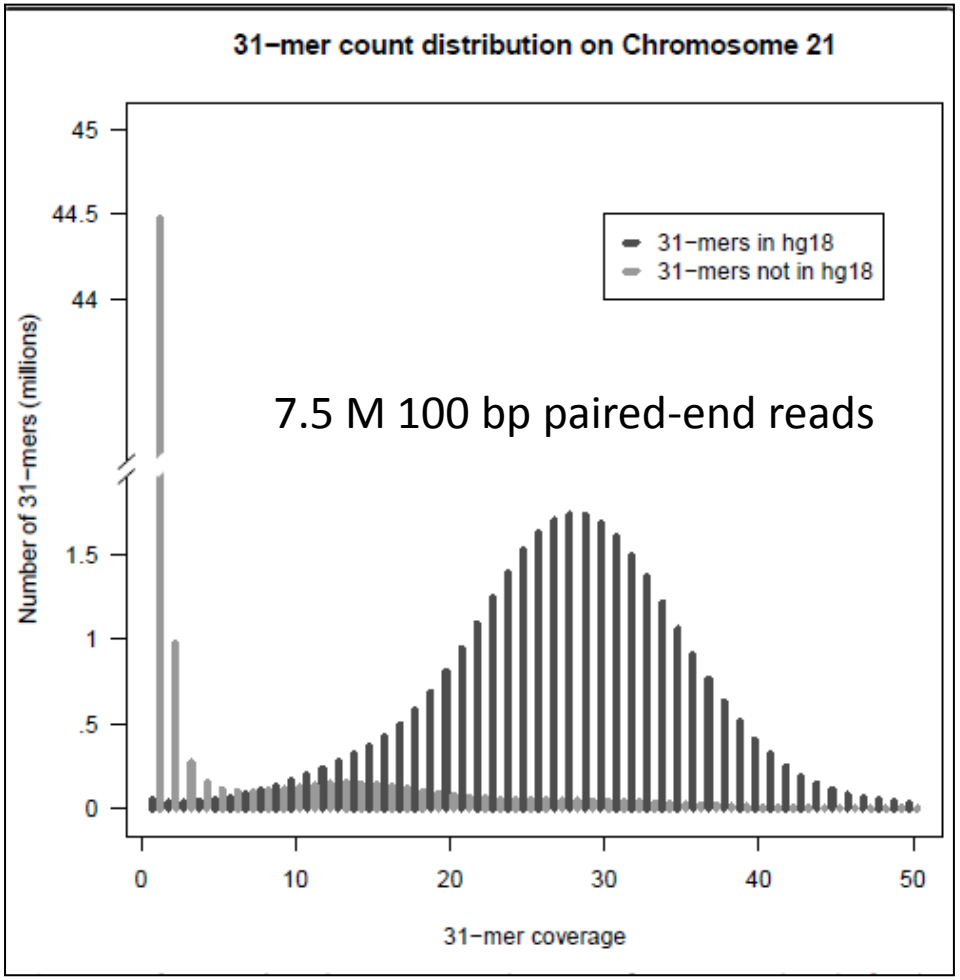
ACGTC**ACG**

acg tca
cgt cac
gtc **acg**

five 3-mers

Counting the occurrences of all such substrings in a sequence generates a k-mer frequency distribution:

(reverse complements counted as a single k-mer)



Gray = errors and SNPs

Melsted & Pritchard (2011) Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinformatics 12:333.

Sequence Processing

Digital Normalization

Brown et al (2014)

1 error => up to *k* erroneous *k*-mers

5-mer example

Table 1. Digital normalization to C=20 removes many erroneous k-mers from sequencing data sets. Numbers in parentheses indicate number of true k-mers lost at each step, based on reference.

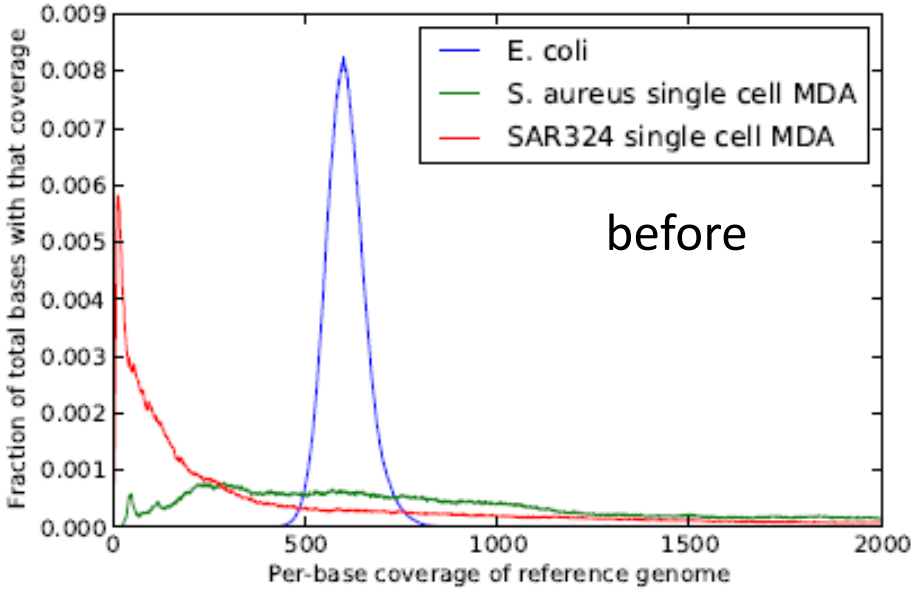
Data set	True 20-mers	20-mers in reads	20-mers at C=20	% reads kept
Simulated genome	399,981	8,162,813	3,052,007 (-2)	19%
Simulated mRNAseq	48,100	2,466,638 (-88)	1,087,916 (-9)	4.1%
E. coli genome	4,542,150	175,627,381 (-152)	90,844,428 (-5)	11%
Yeast mRNAseq	10,631,882	224,847,659 (-683)	10,625,416 (-6,469)	9.3%
Mouse mRNAseq	43,830,642	709,662,624 (-23,196)	43,820,319 (-13,400)	26.4%

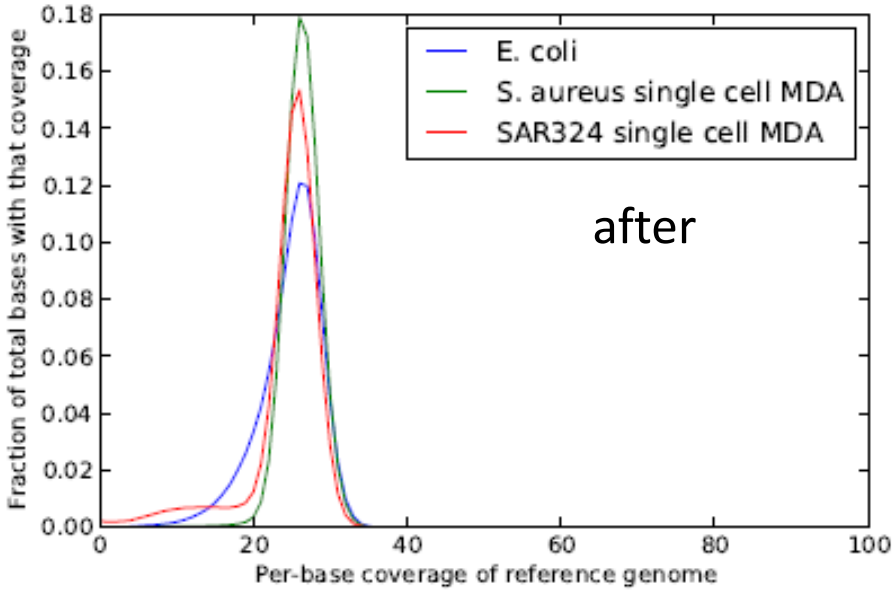
Digital Normalization

Brown et al (2014)

Retains nearly all real *k*-mers while discarding the majority of erroneous *k*-mers.

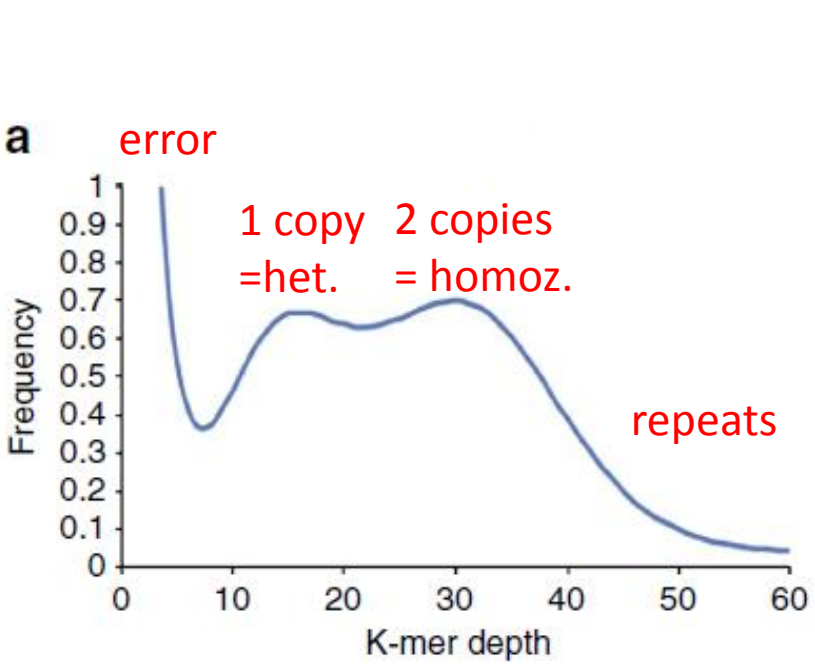
Reduces the number of over-represented *k*-mers (duplicates, repeats)

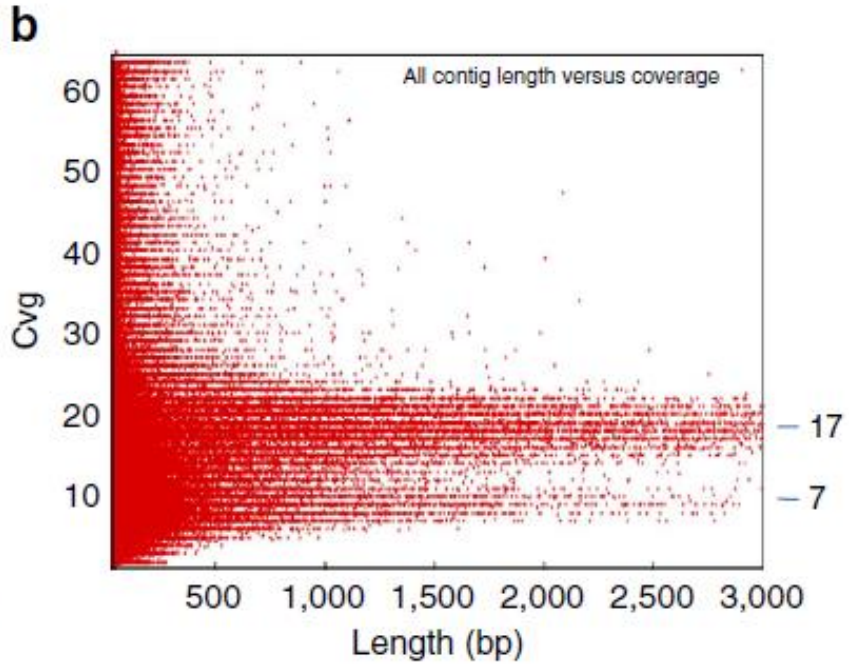




Uneven Sequencing Depth:

- Heterozygosity
- Repeated Nuclear Sequences
- Organelles
- Biological Contamination





Whole Genome Illumina Reads

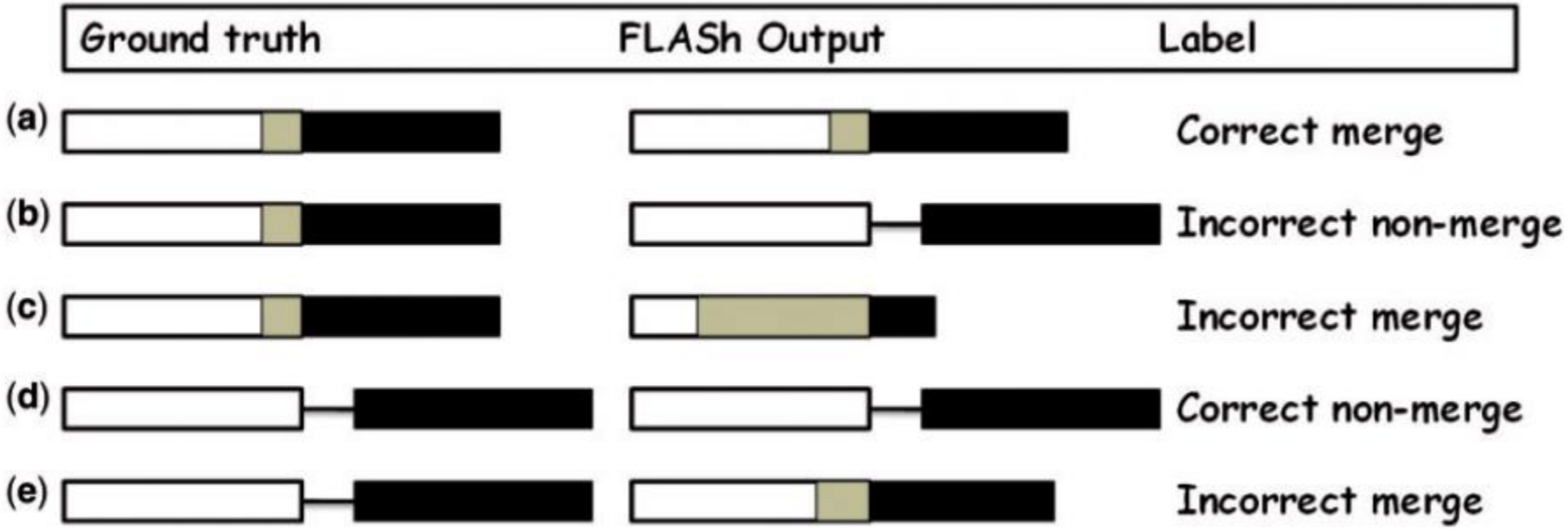
Assembled Contigs

Zheng et al. 2013. Wheat stripe rust fungus. Nature Communications 4:2673.

Read Overlap

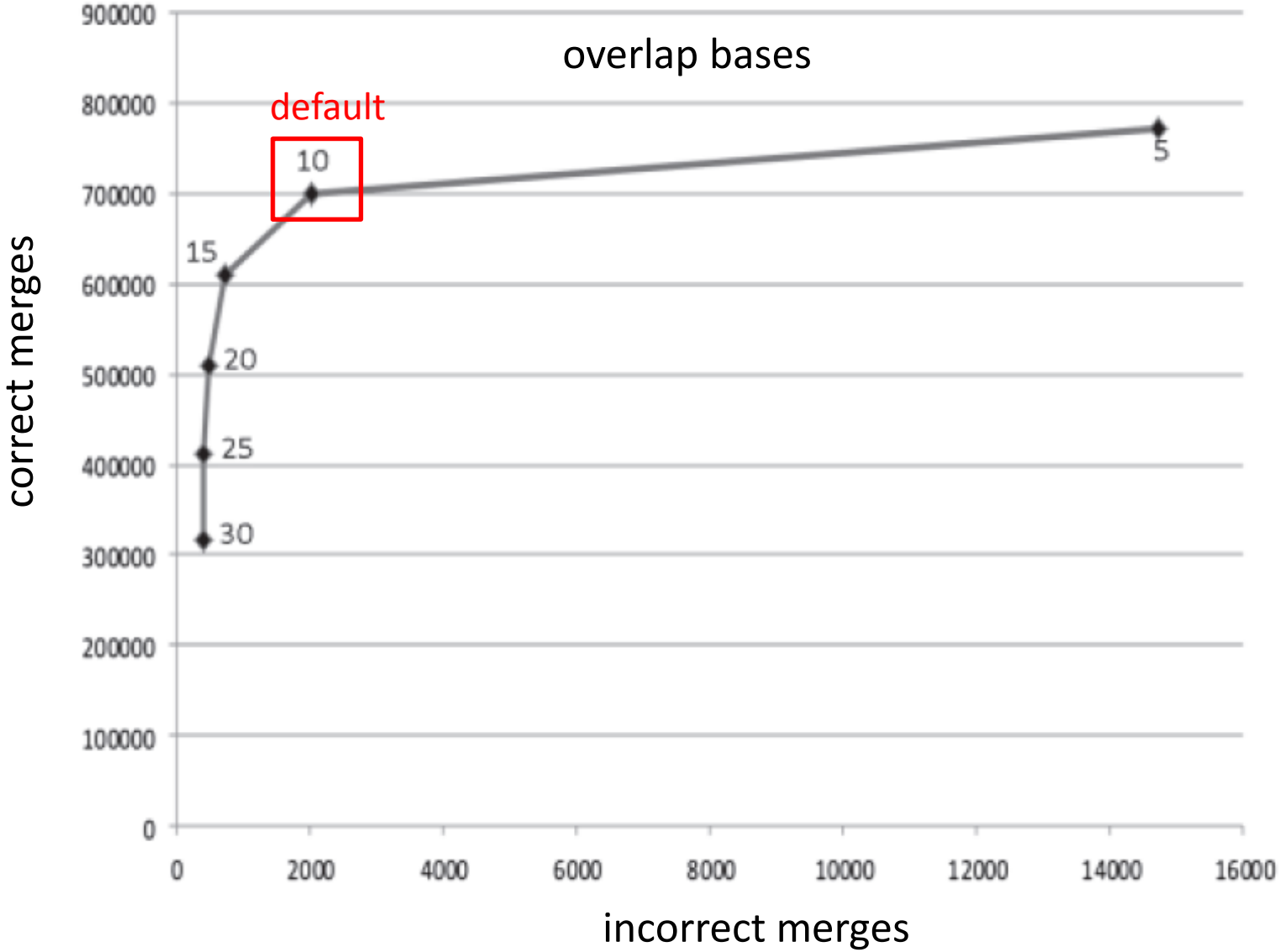
FLASH Majoc & Salzberg 2011 Bioinformatics 27:2957-2963.

Possible Outcomes of Paired Read Merge Algorithm



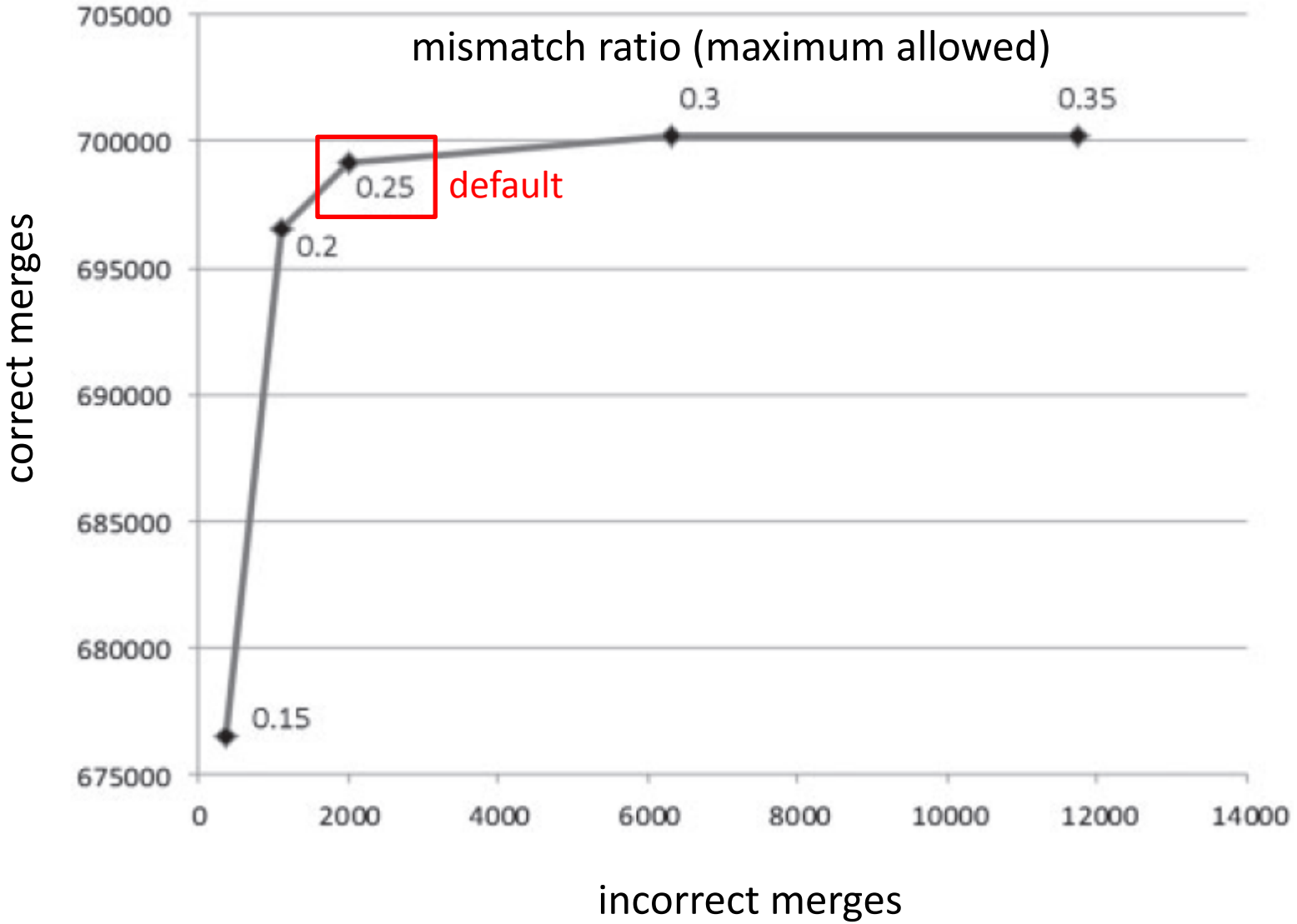
Read Overlap

FLASH results with 1 million simulated reads, 1% error



Read Overlap

FLASH results with 1 million simulated reads, 1% error



bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data

Richard W Lusk

bioRxiv posted online January 30, 2014
Access the most recent version at doi:10.

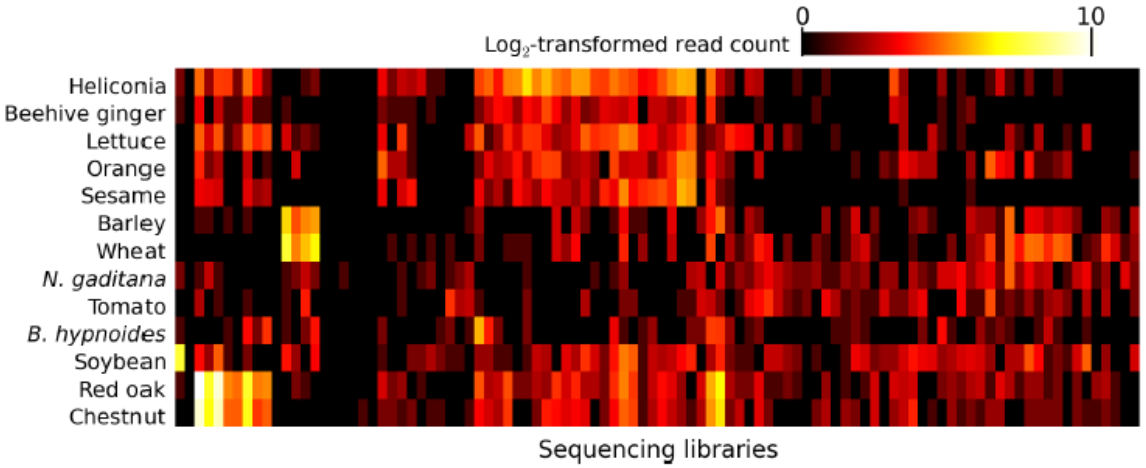


Figure 4. Heterogeneous species appear to contaminant samples from the same tissue and experiment. The "Tumor" [18] experiment dissociated 100 individual cells from a sample of a single tumor and sequenced libraries from each. Following the analysis pipeline of a study that claimed to find different plant species in different blood plasma samples from a single experiment, I used bowtie to screen each read in each library against

Bioinformatics

Illumina data from 1 lane:

200 million paired end reads

40 gigabase

800 million lines

60 gigabyte file size

Data Transfer

FTP

HTTP

portable hard drives

Bioinformatics Computing Hardware

CENTER FOR GENOME RESEARCH & BIOCOMPUTING

OSU Oregon State University

CORE LABS NEWS & EVENTS ETA

Main Menu

- HOME
- CGRB BIOINFO
- WEB ACCESS (ETA)
- DOCUMENTATION
- DONATIONS
- POLICIES
- ACCOUNT REQUEST
- ANNOUNCEMENTS
- SERVICES
- WORKSHOP ACCOUNT
- CLOUD STATUS
- CONTACT

CGRB Bioinformatics

Resources

- > 1800 AMD and Intel processors
- >900TB shared disk space
- Hardware:
  - AMD
  - Intel
  - Dell
  - Advanced HPC
  - Hewlett Packard
  - Sun Microsystem
  - APC and Eaton
- 10G main network connection
- Gigabit private network
- Secure, climate controlled

Projects

- Assembling the Fungal Tree of Life
- Marine Microbial Genomics
- CGRB Bioinformatics
- Plantontology.org
- Phytophthora-id.org

Computing Hardware

Desktop Computer with 16 GB RAM, 1 TB HD	$1,500
48 CPU Server with 96 GB RAM	$9,000
24 TB Storage Array	$10,000

Bioinformatics

Data Processing Options

Commercial Packages (Geneious, CLC Bio)

$400 and up

Bioinformaticist or Undergraduate Programmer

NGS experience or not

Web-Services

iPlant, Galaxy

Do It Yourself

Programming (R, Python, Perl, Ruby, Java …)

Linux + Google

Bioinformatics

Open Source Software

Blat (Kent et al 2002) — Sequence Similarity Searching
Trimmomatic (Bolger et al 2014) — Short Read Quality Trimming
khmer (Brown et al 2014) — Digital Normalization

Velvet (Zerbino & Birney 2008)
YASRA (Ratan 2009)
BWA (Li & Durbin 2009)
Abyss (Simpson et al 2009)
Trinity (Grabherr et al 2011)

Short Read Assemblers

(RNA)

SAMTools (Li et al 2009) — Analyzing Mapped Reads
Tablet (Milne et al 2010) — Visualizing Mapped Reads

Scripting Languages:
Python, R swirlstats.com

IPython
Interactive Computing

{swirl}
statistics with interactive R learning: an R package

Linux
My 6 favorite commands:
sed, grep, sort, uniq, join, awk

For questions that we didn't answer:

SEQanswers
the next generation sequencing community

SEQanswers > Bioinformatics
Bioinformatics

Welcome, liston.
You last visited: 06-27-2014 at 02:57 PM

User CP | FAQ | Community | Calendar | New Posts: 1,470 | New Posts In | Search | Quick Links | Log Out

New Thread

Page 1 of 521 1 2 3 11 51 101 501 > Last »

Threads in Forum : Bioinformatics — Forum Tools | Search this Forum

Thread / Thread Starter	Last Post	Replies	Views
Sticky: ICGC-TCGA DREAM Somatic Mutation Calling Challenge khoulahan	05-27-2014 10:47 PM by karencmartin	7	28,838
Sticky: New Resources for 1000 Genomes (1 2) laura	12-05-2013 02:37 AM by laura	35	19,914
Sticky: SEQwiki dan	08-09-2010 11:56 PM by scofield_gao	3	20,421
Sticky: Software packages for next gen sequence analysis (1 2 3 ... Last Page) sci_guy	12-25-2009 06:45 PM by ECO	236	357,003
How to blast a transposon sequence to a MiSeq transposon library helelein	Today 05:06 PM by Brian Bushnell	3	84
KEGG Pathway menegidio	Today 04:21 PM by menegidio	0	27
Help with Bowtie2 aakriti	Today 04:14 PM by GenoMax	9	122
Remote blast+ skipping query sequences arundurvasula	Today 04:07 PM by GenoMax	1	69
unmark duplicates SWP	Today 03:04 PM by mebbert	2	126
Mitochondria genome-denovo assembly bioman1	Today 02:24 PM by francicco	12	895
RNA-Seq Pathway and Gene-set Analysis Workflows in R/Bioconductor with GAGE/Pathview (1 2 3 ... Last Page) bigmw	Today 01:49 PM by tigerxu	82	7,528
BioGPS, command line? sindrle	Today 01:44 PM by sindrle	4	238
bootstrapping gene-content trees with discrete binary data using fseqboot (EBASSY) someperson	Today 01:20 PM by someperson	0	63
HMMSearch on 6-frame translated genome vs protein sequences yields different results Loddi	Today 12:51 PM by Loddi	0	54