

Genomics - Technology to Decipher Biology

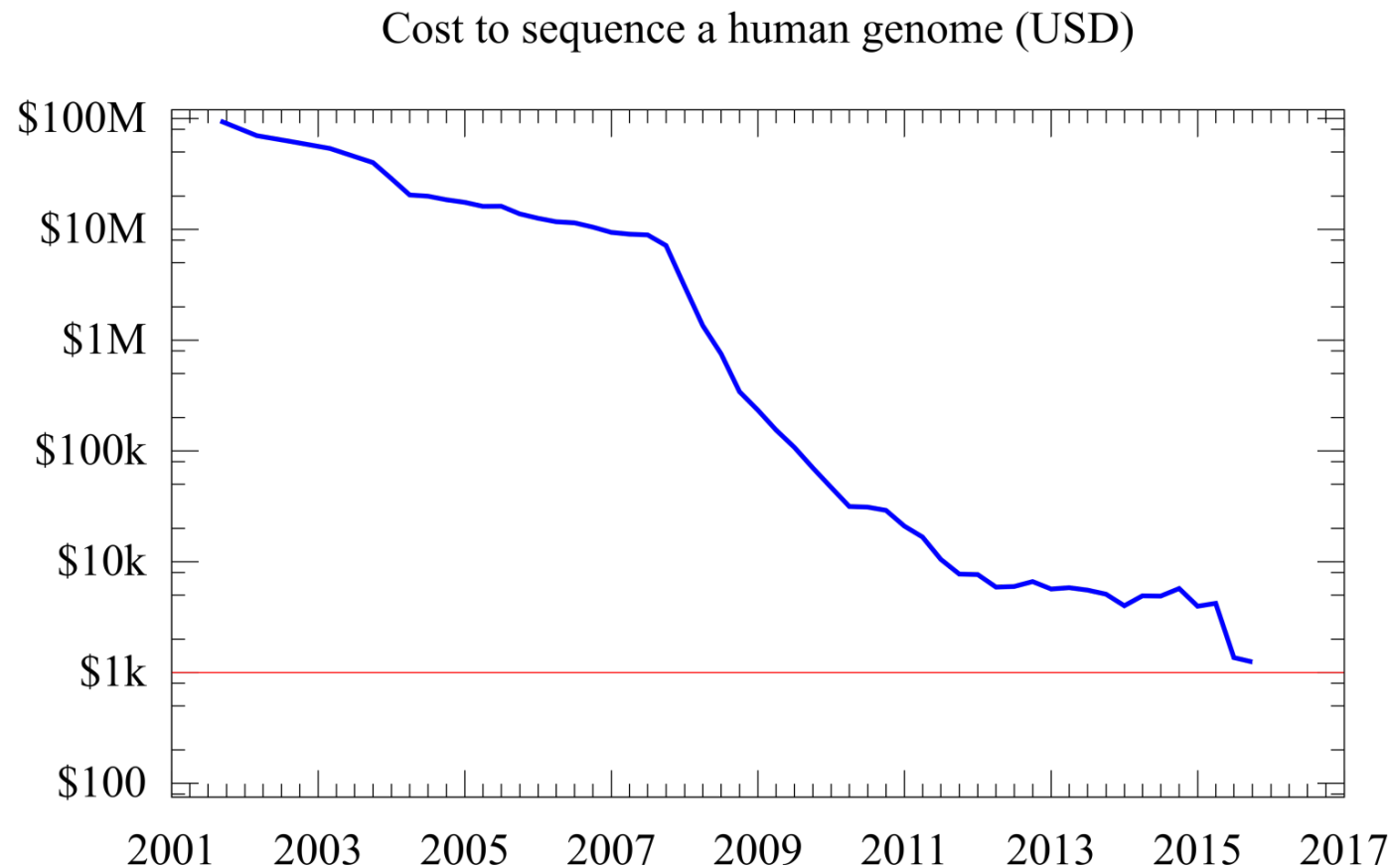
Klaus Mayer



HelmholtzZentrum münchen
German Research Center for Environmental Health

- Genomes - their coming of age
- technology leaps
- Assemblies and the struggle to cope with cheap and large data (buzz word: big data...)
- Evaluations and a bit of genome biology

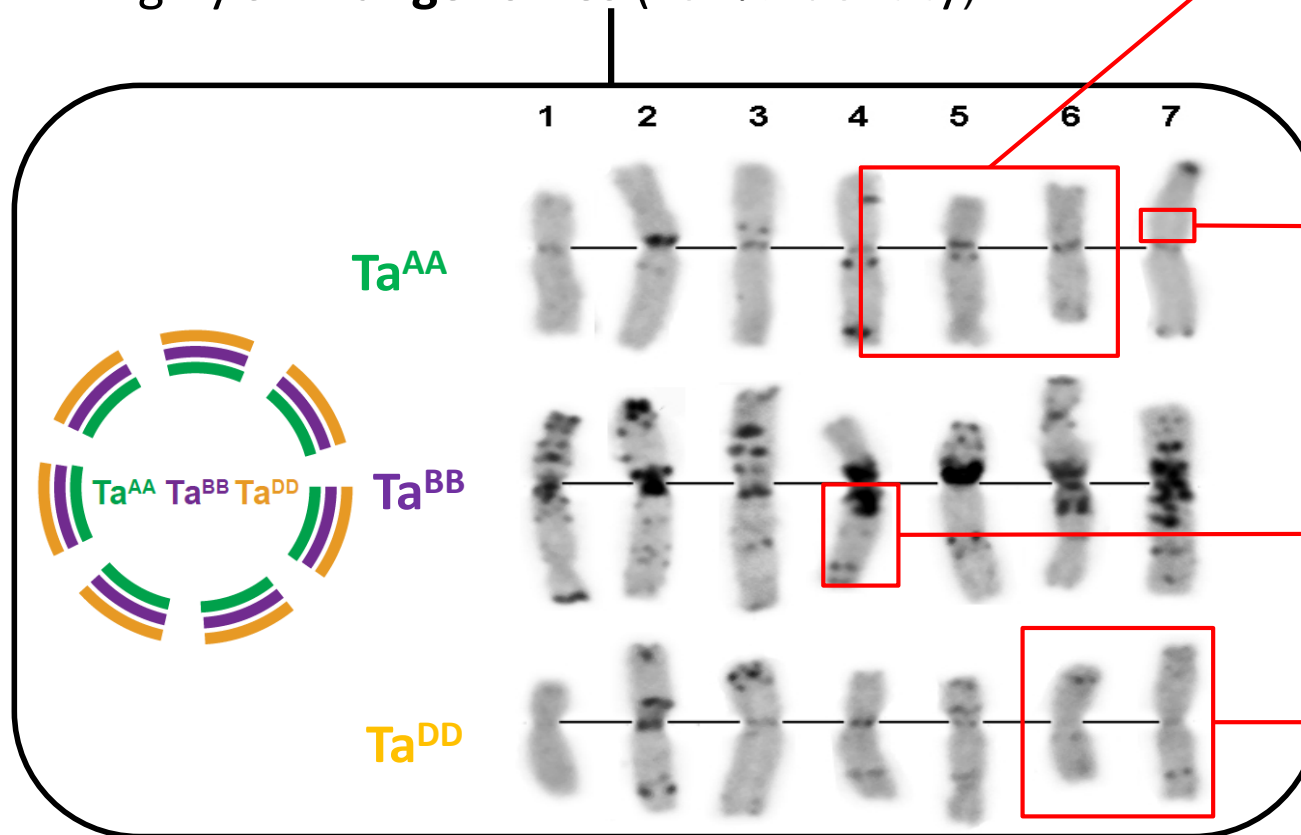
The \$1000 Genome



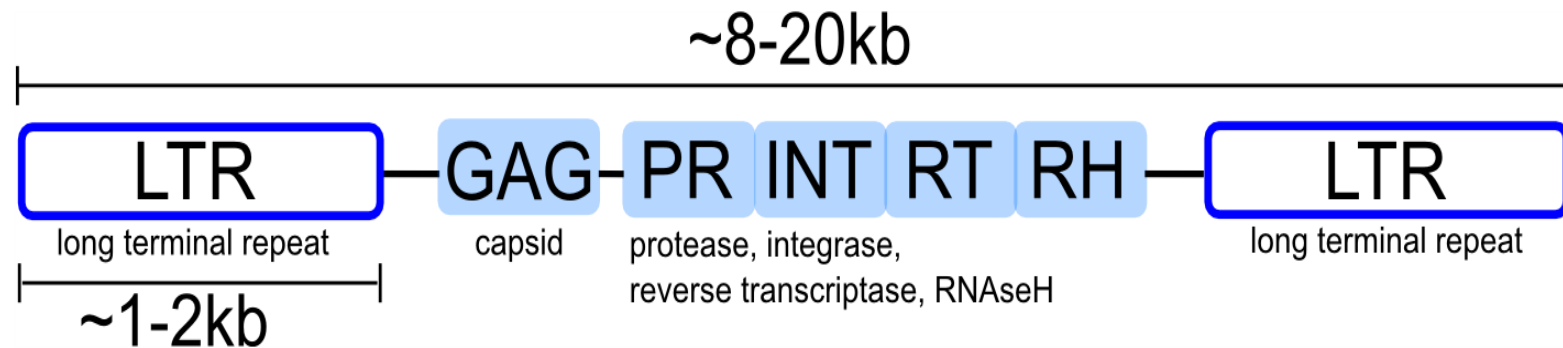
PGSB

The genome of bread wheat (*T. aestivum*) is ...

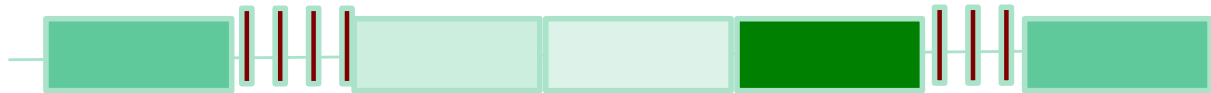
- allohexaploid: three “diploid-like” ($2n = 6x = 42$)
- ~17Gb genome size; >80% repeat content
- Highly similar genomes (>97% identity)



Why is the assembly so complicated?



Repeat Structures in Vertebrates and (many) Plants



Plants

Repeat elements up to 20kb
Often intercalated;
predominantly outside of
genes



Vertebrates

Dominated by Alu repeats;
300bp units;
Can be embraced and
assembled using
conventional sequencing
libraries

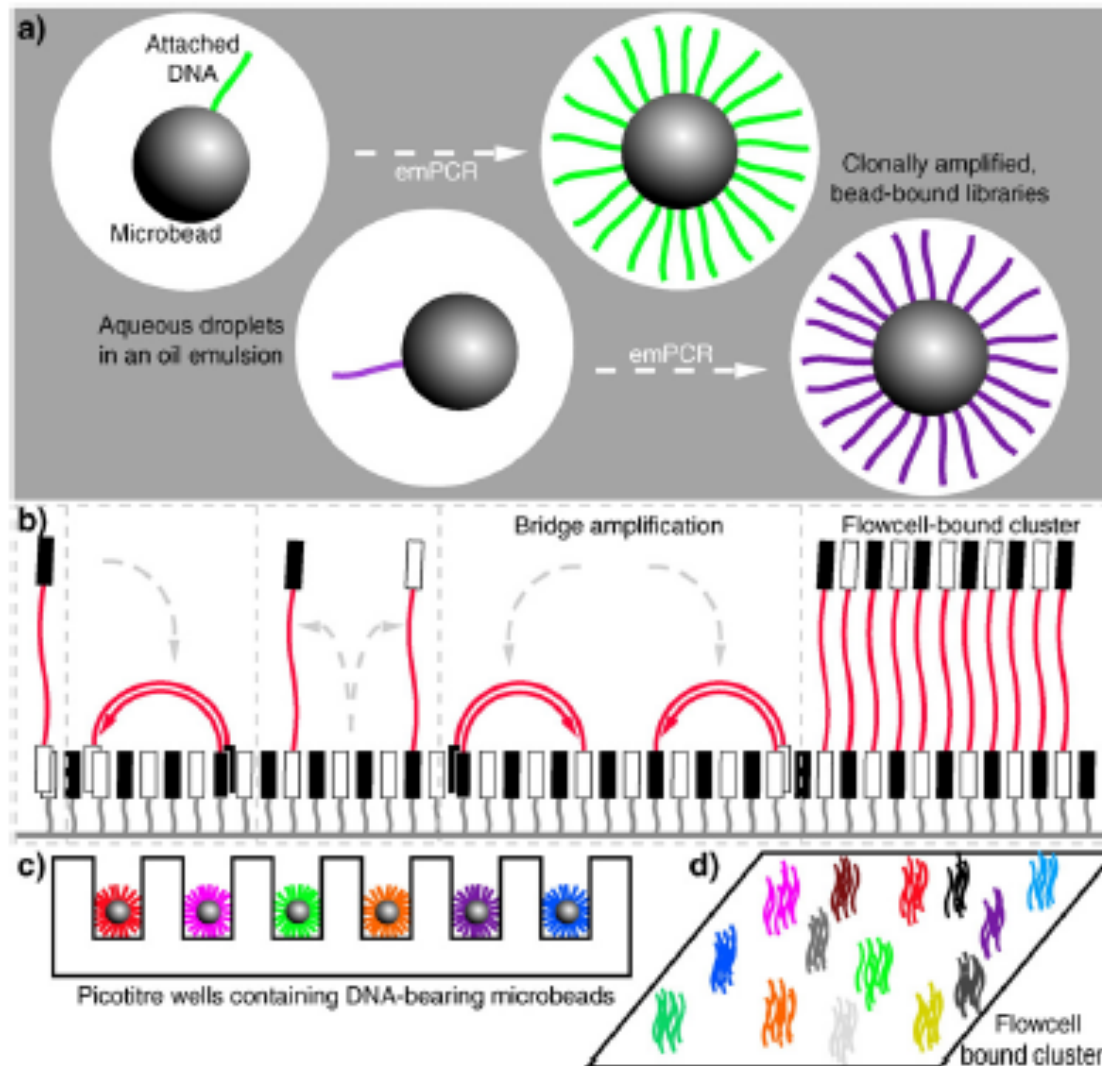


Fig. 2. Second-generation DNA sequencing parallelized amplification. (a): DNA molecules being clonally amplified in an emulsion PCR (emPCR). Adapter ligation and PCR produces DNA libraries with appropriate 5' and 3' ends, which can then be made single stranded and immobilized onto individual suitably oligonucleotide-tagged microbeads. Bead-DNA conjugates can then be emulsified using aqueous amplification reagents in oil, ideally producing emulsion droplets containing only one bead (illustrated in the two leftmost droplets, with different molecules indicated in different colours). Clonal amplification then occurs during the emPCR as each template DNA is physically separate from all others, with daughter molecules remaining bound to the microbeads. This is the conceptual basis underlying sequencing in 454, Ion Torrent and polony sequencing protocols. (b): Bridge amplification to produce clusters of clonal DNA populations in a planar solid-phase PCR reaction, as occurs in Solexa/Illumina sequencing. Single-stranded DNA with terminating sequences complementary to the two lawn-oligos will anneal when washed over the flow-cell, and during isothermal PCR will replicate in a confined area, bending over to prime at neighbouring sites, producing a local cluster of identical molecules. (c) and (d) demonstrate how these two different forms of clonally-amplified sequences can then be read in a highly parallelized manner: emPCR-produced microbeads can be washed over a picotitre plate, containing wells large enough to fit only one bead (c). DNA polymerase can then be added to the wells, and each nucleotide can be washed over in turn, and dNTP incorporation monitored (e.g. via pyrophosphate or hydrogen ion release). Flow-cell bound clusters produced via bridge amplification (d) can be visualized by detecting fluorescent reversible-terminator nucleotides at the ends of a proceeding extension reaction, requiring cycle-by-cycle measurements and removal of terminators.



Contents lists available at ScienceDirect

Genomics

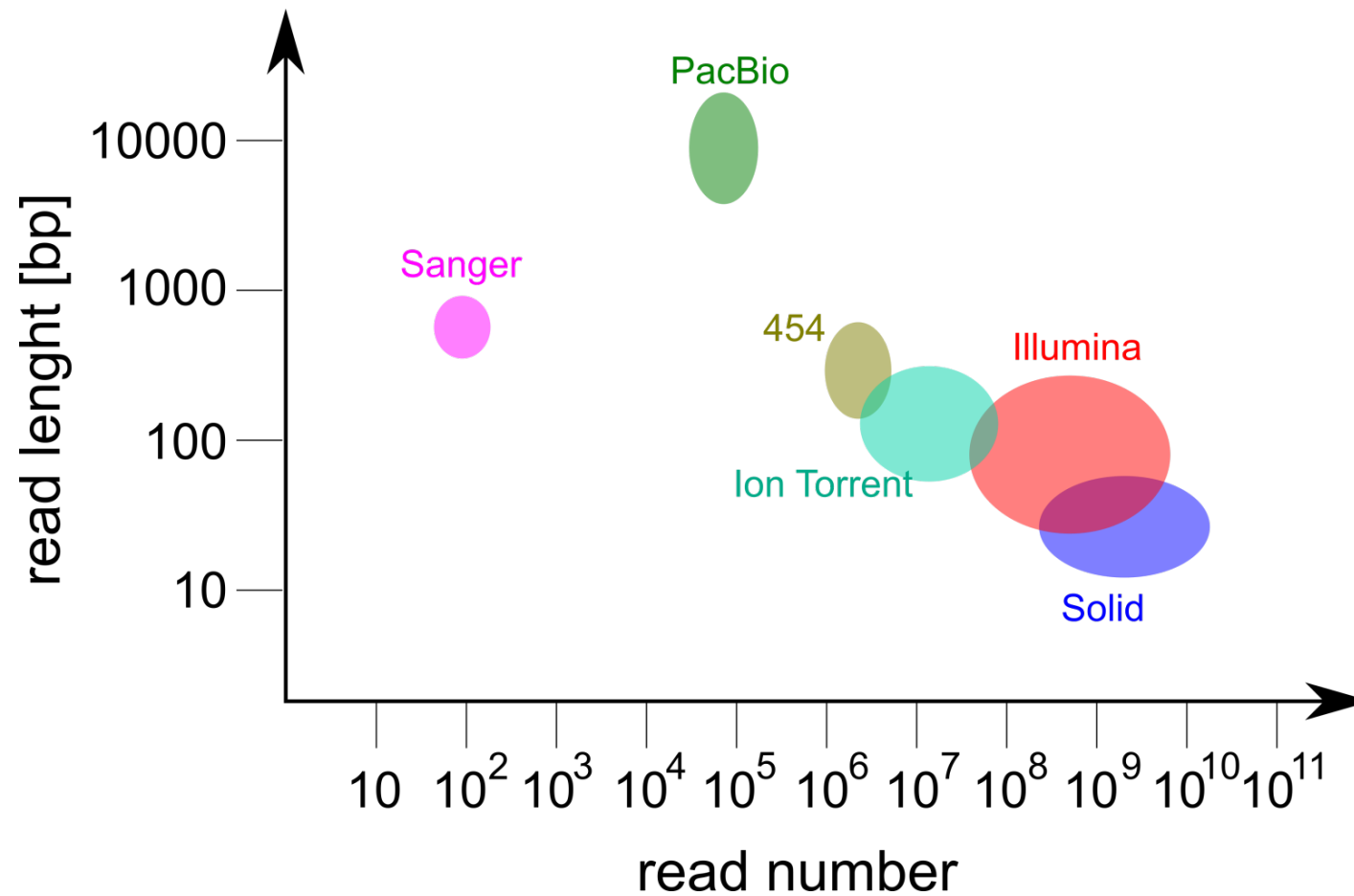
journal homepage: www.elsevier.com/locate/ygeno

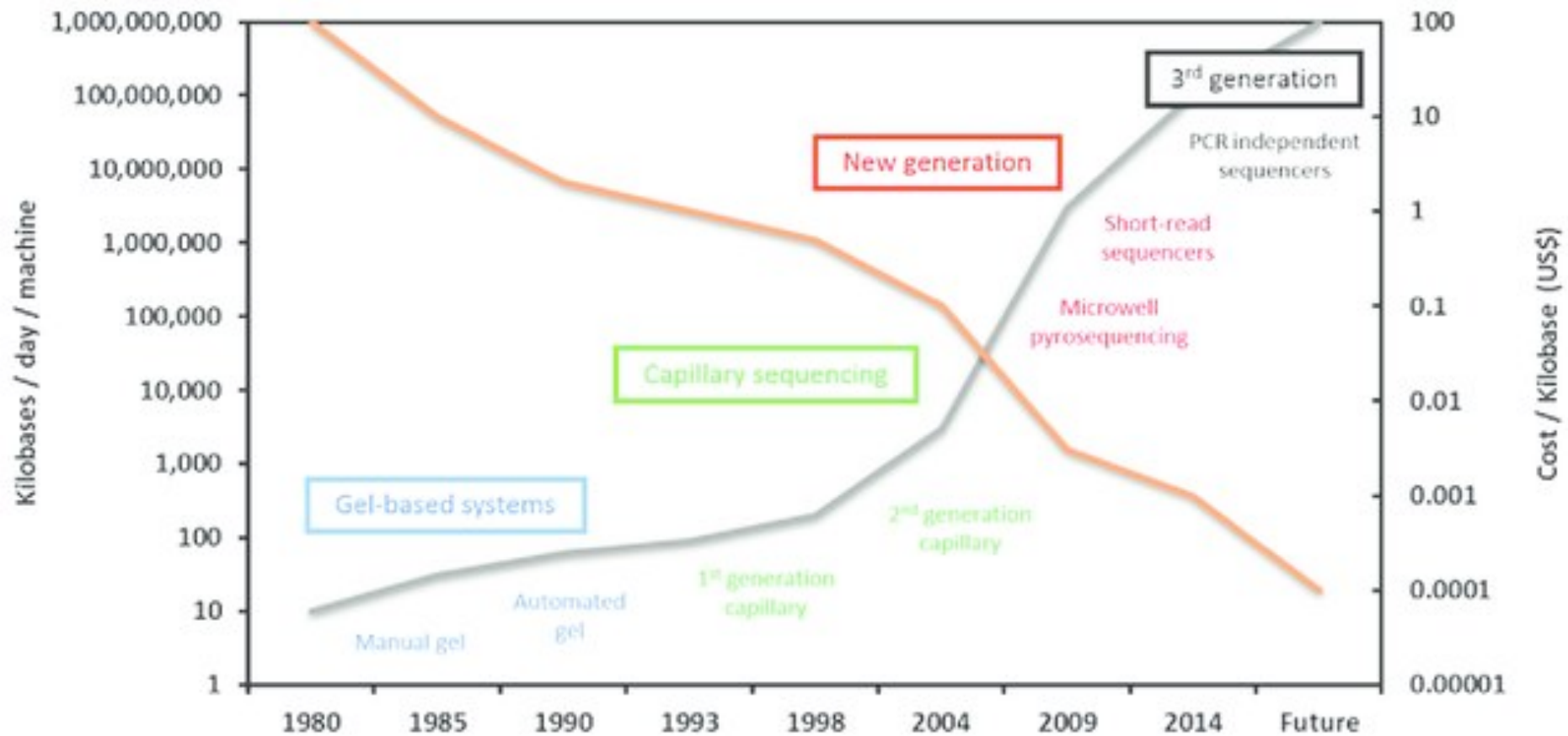
Review

The sequence of sequencers: The history of sequencing DNA

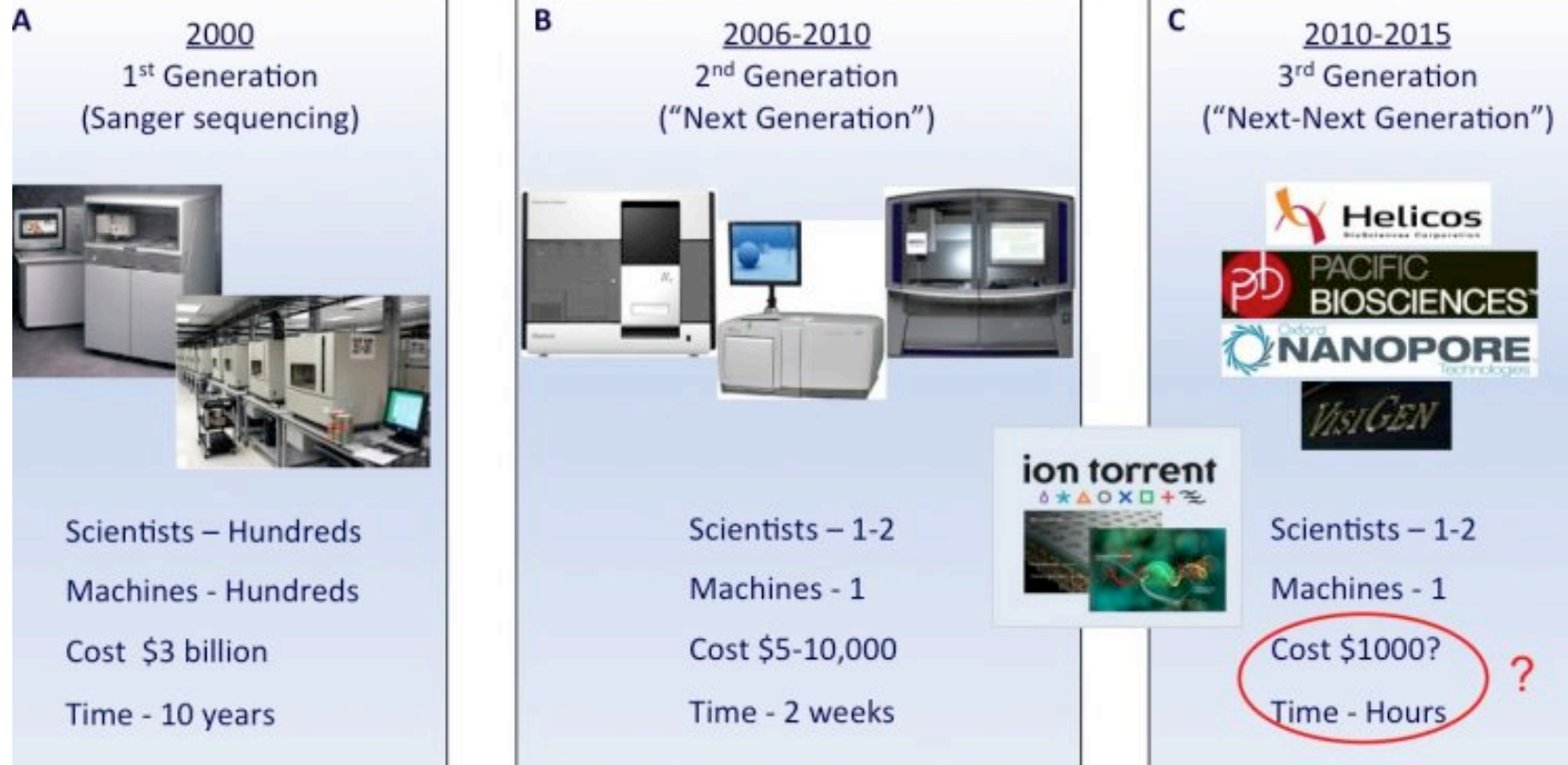
James M. Heather ^a, Benjamin Chain

Output Dimensions



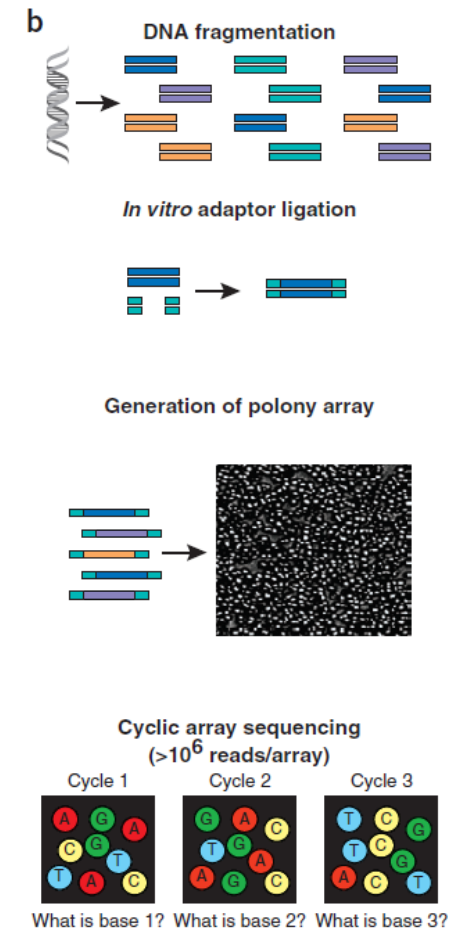


Effort required to sequence a human genome



Overview NGS Features

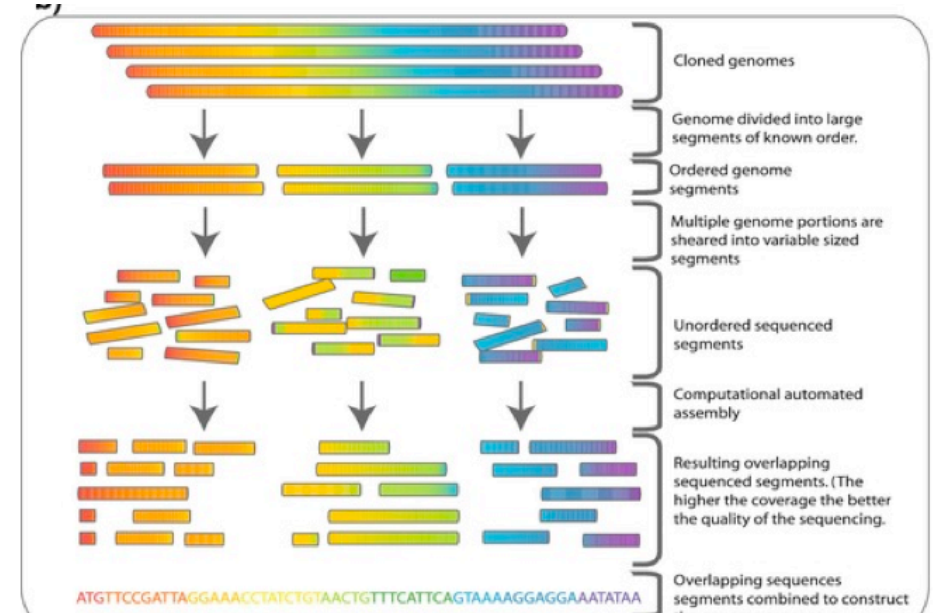
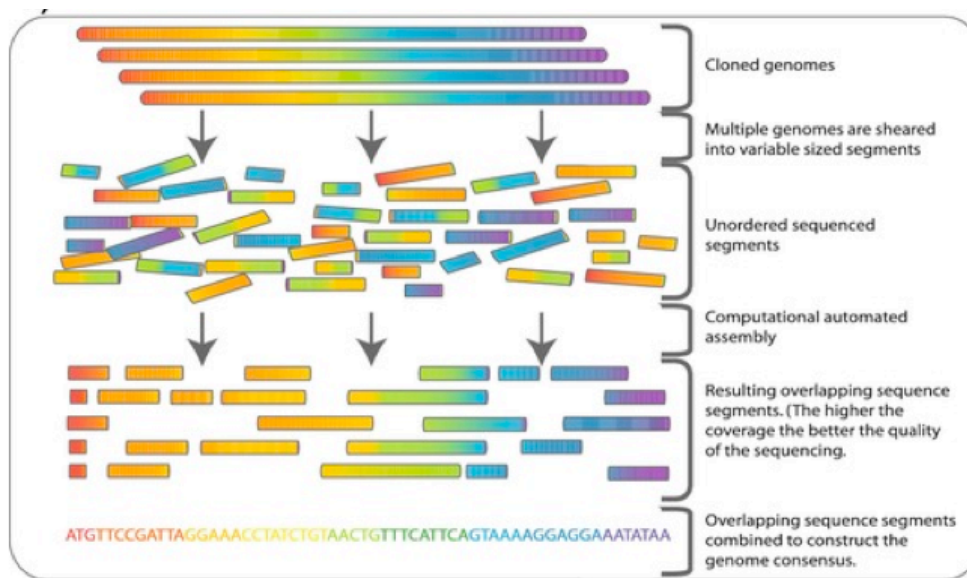
- Tiny reaction volumes
 - High parallelization
 - Very low cost per bp
 - Huge amount of sequence information (Tbp)
-
- Small reads
 - High error rates (~1% illumina, up to 15% PacBio)
 - In part complex error models
 - Expensive equipment



Applications of NGS

- RNA-Seq
 - ➡ DEG, Transcriptome
- DNA-methylation studies
 - ➡ Methylome, Epigenetics
- ChIP-Seq
 - ➡ TF-binding sites, regulation
- Exome Sequencing
 - ➡ Variation analysis
- De novo assembly
 - ➡ SVs, reference genomes
- Genome re-sequencing
 - ➡ Large scale variation analysis

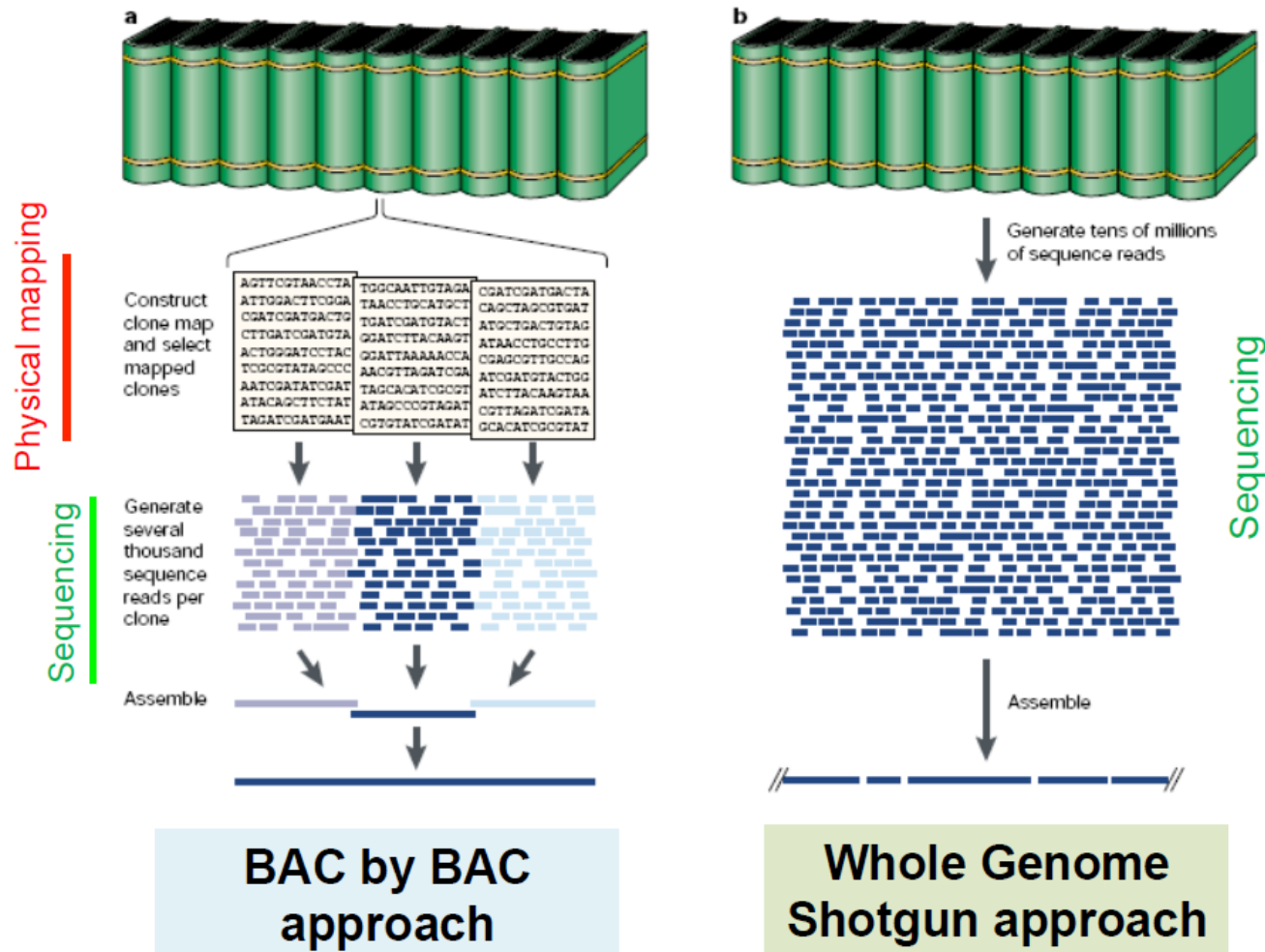
Oriented vs. Shotgun Sequencing and Assembly



Source: Wikimedia commons

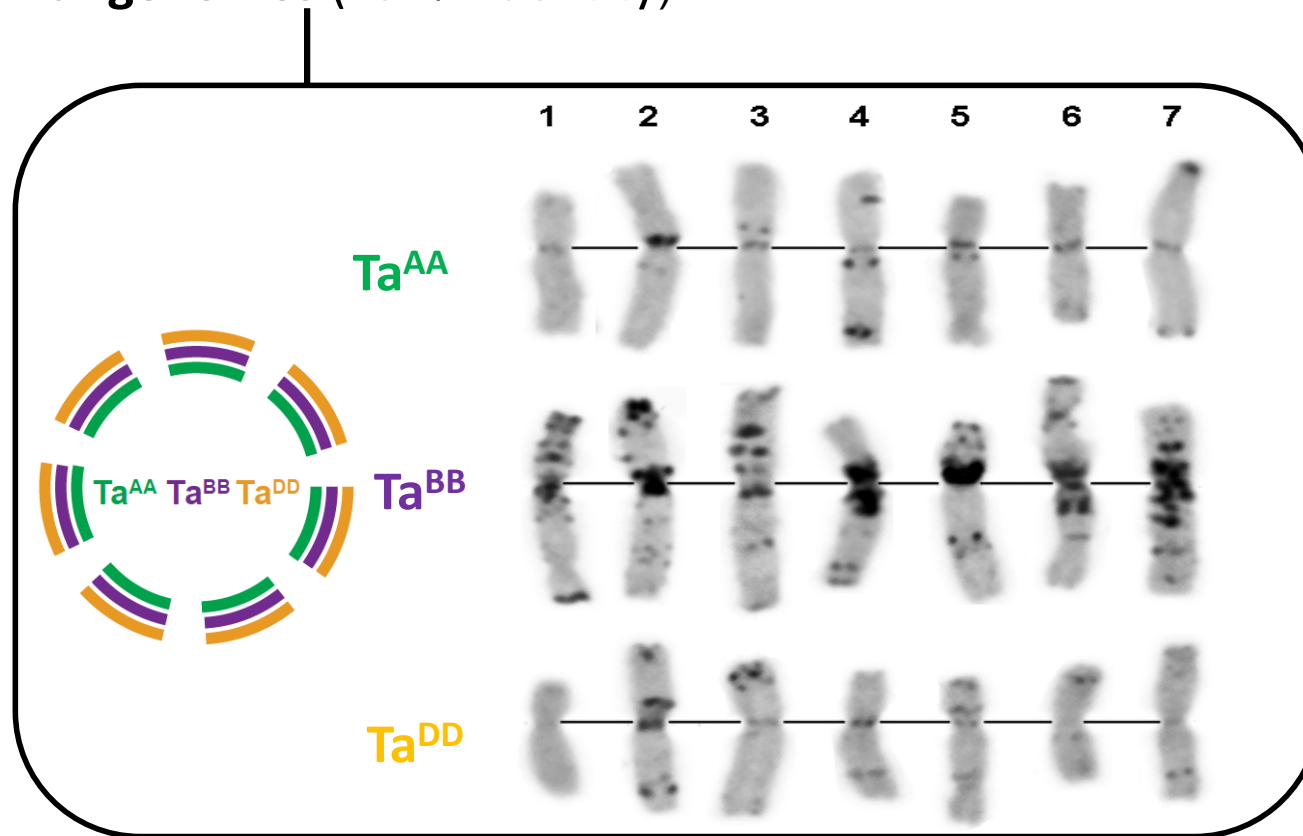


Wheat genome sequencing strategies

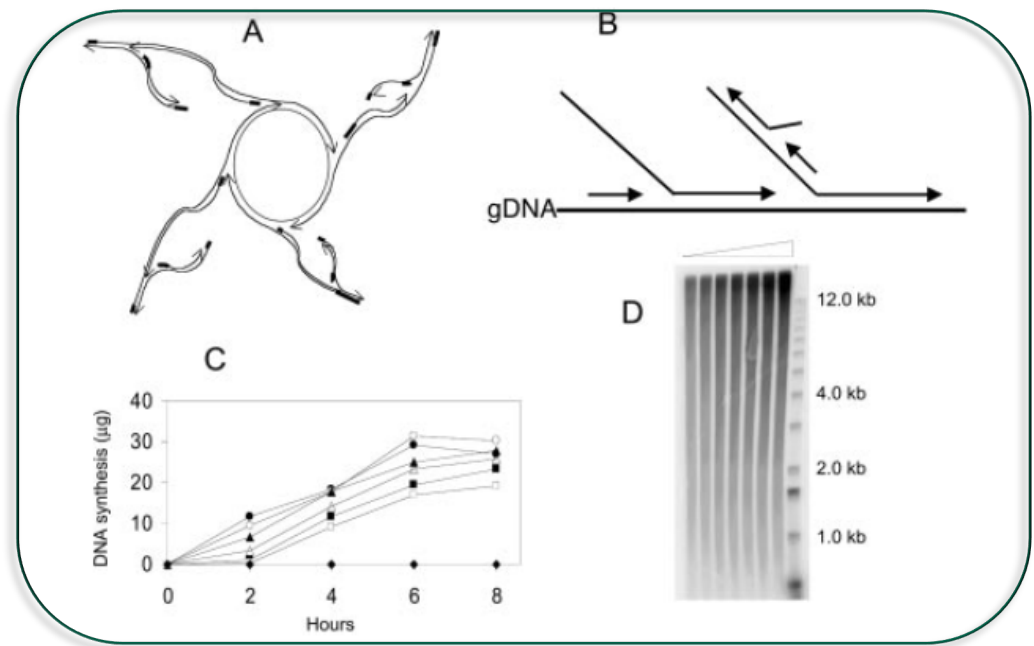
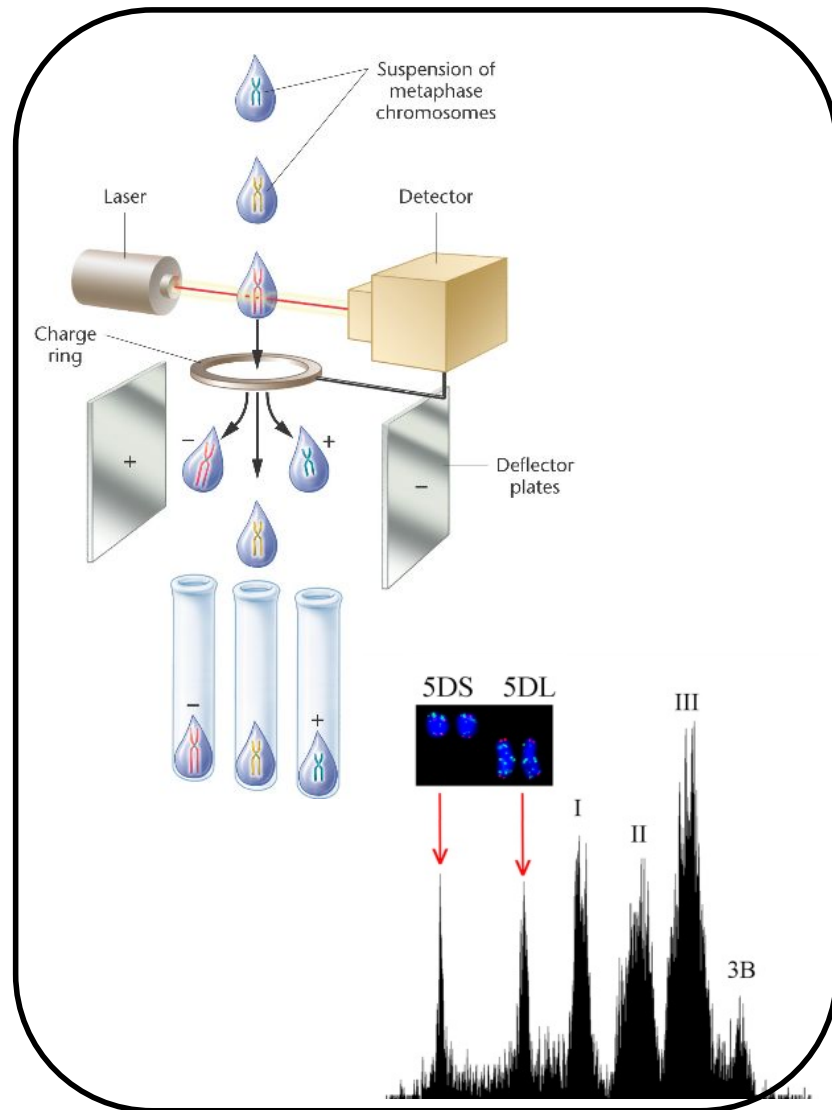


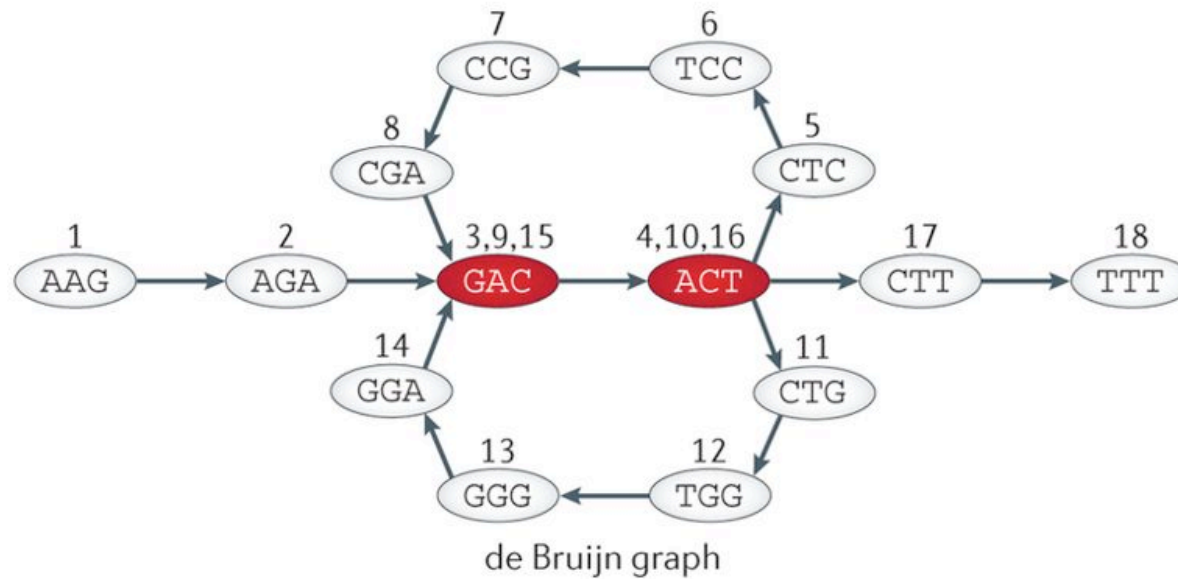
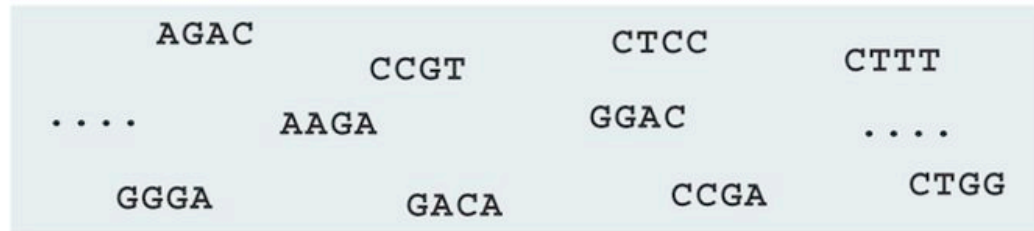
The genome of bread wheat (*T. aestivum*) is ...

- allohexaploid: three “diploid-like” ($2n = 6x = 42$)
- ~17Gb genome size; >80% repeat content
- Highly similar genomes (>97% identity)



Reduction of complexity by chromosome sorting





AA**GACT**CC**GACT**GG**GACT**TT

Reads

Eulerian path

Assembly

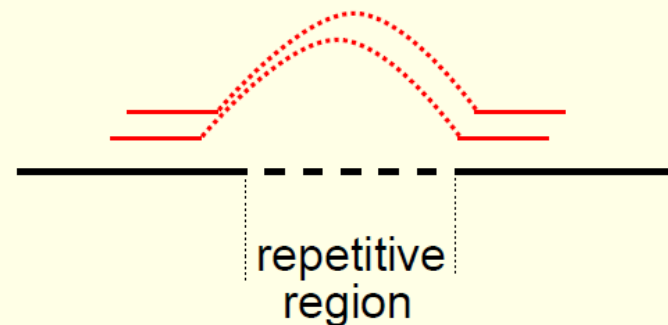


Usage of paired end reads and mate pairs

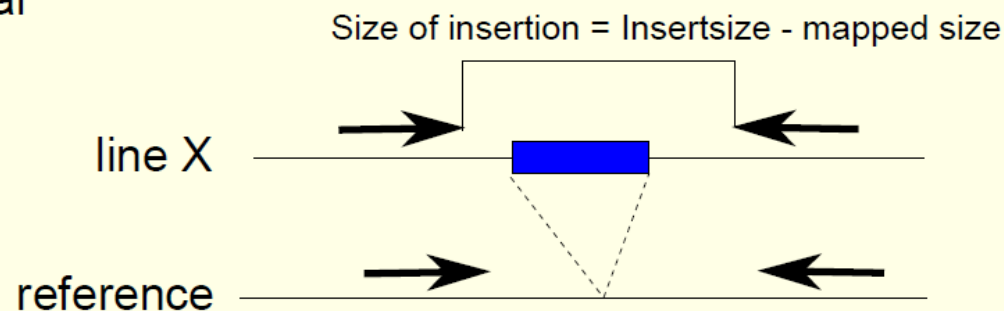
Paired-end libraries: term used for insert sizes ~ up to 1 kb

Mate pairs: term used for libraries with larger insert sizes
eg. fosmid libraries ~40-50 kb inserts

Connect contigs in genome assemblies over repetitive regions → scaffolding



Detection of structural variants



repeat repeat

GTACTGTTACTGTCTATCGATTCCCTATCTATAGGGACTCTAGATTCACG**GTACTGTTACT**

TCTATAGGGACT
 GACTCTAGAT
 GATTCACGGTA
 TCACG**GTACTG**
 ACTGTTACT
 GTACTGTTAC
 TTACTGTCT
 CTGTCTATCGA
 CGATTCCTATC
 TATCTATAGGGA

Repeat collapsed into one sequence

TCTATAGGGACTCTAGATTCACG**GTACTGTTACT**GTCTATCGATTCCCTATCTATAGGGA

The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*

Aleksey V. Zimin^{1,2}, Daniela Puiu¹, Richard Hall³, Sarah Kingan³, Bernardo J. Clavijo⁴, and Steven L. Salzberg^{1,5,*}

The most effective way to resolve repeats is to generate individual reads that contain them. If a single read is longer than a repeat, and if both ends of the read contain unique sequences, then genome assemblers can unambiguously place the repeat in the correct location. Without such reads, every long repeat creates a breakpoint in the assembly. Recent advances in sequencing, particularly the long read, single-molecule sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore (MinION), can produce reads in excess of 10,000 bp, although with a high error rate. By combining these very long reads with highly accurate shorter reads, we have been able to produce an assembly of the wheat genome with contigs that are more than ten times longer than those produced in any previous attempt. Ours is the first assembly that contains nearly the entire length of the genome, with more than 15.3 billion bases.

Construction of super-reads and mega-reads required approximately 100,000 CPU hours, of which 95% was spent in the mega-reads step. By using large multi-core computers to run these steps in parallel, these steps took 1.5 months of elapsed (wall clock) time. The peak memory (RAM) usage was 1.2 terabytes.

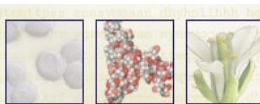
We then assembled the mega-reads and the synthetic pairs using the Celera Assembler [8] (v8.3), which was modified to work with our parallel job scheduling system. (The modified software is available at <ftp://ftp.ccb.jhu.edu/pub/dpuu/OTHER/SLURM/runCA>.) The CA assembly process required many iterations of the overlapping, error correction, and contig construction steps, and it was extremely time

releases. The total CPU time was ~470,000 CPU hours (53.7 years), which was only made feasible by running it on a grid with thousands of jobs running in parallel (the maximum number was 3,320) for some of the major steps. The total elapsed time was just over 5 months. When combined with the earlier steps, the entire assembly process took 6.5 months. The resulting assembly, labelled Triticum 1.0,

with an N50 size of 16 Kb. Error correction and assembly of the corrected reads was completed using ~150,000 CPU hours, which took ~3 weeks on a 16-node cluster. The contigs output from FALCON require further polishing, which involves realignment of raw reads and calculation of a new consensus [11]. For the polishing step, we used Pacbio's resequencing pipeline from the SMRT Analysis package

possible without the availability of very large parallel computing grids. All together, the various assembly

steps took 880,000 CPU hours, or just over 100 CPU years. An important technical note is that the



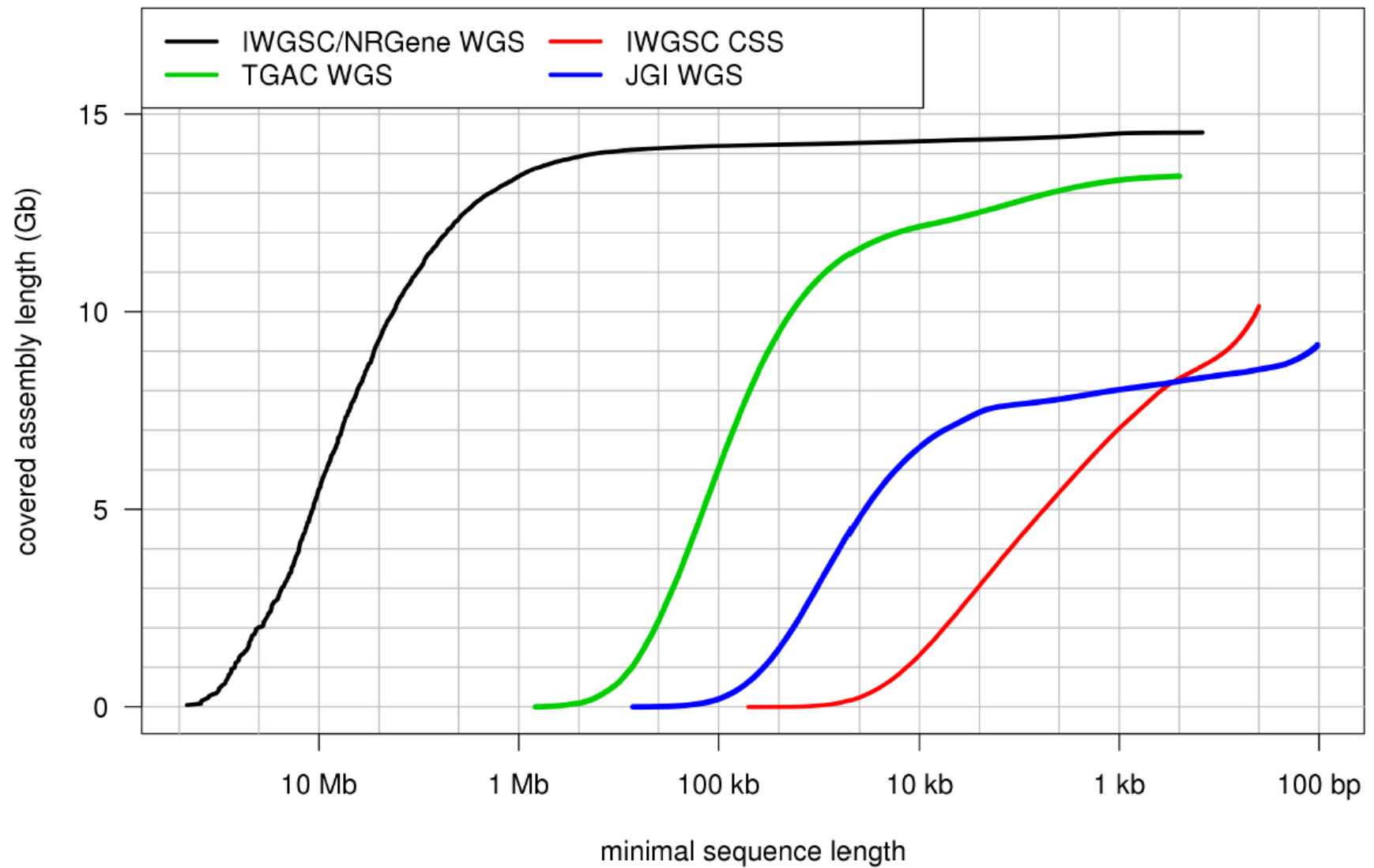
A Breakthrough in Sequence Assembly

NRGene

DeNovoMAGIC™ 2.0

- **Novel algorithm for (genome) sequence assembly**
- Uses standard **Illumina short reads** (at high coverage)
- Some specific insert library sizes needed
- **Long scaffolds** even in presence of high repetitivity and large genome sizes
- **Fast:** 14 days for hexaploid wheat genome (17 Gb!)



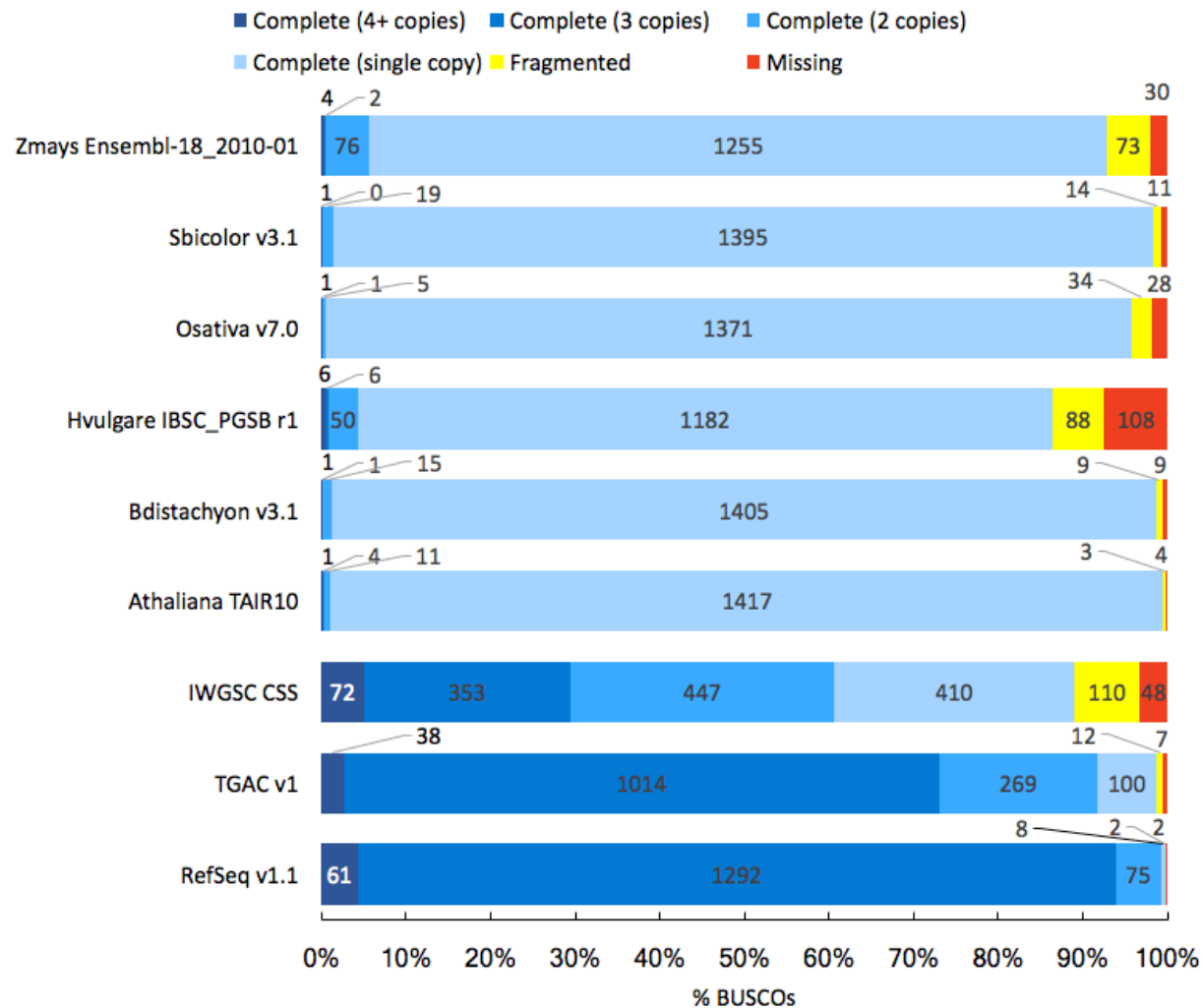


Martin Mascher, IPK



HelmholtzZentrum münchen
German Research Center for Environmental Health

Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs (BUSCO)



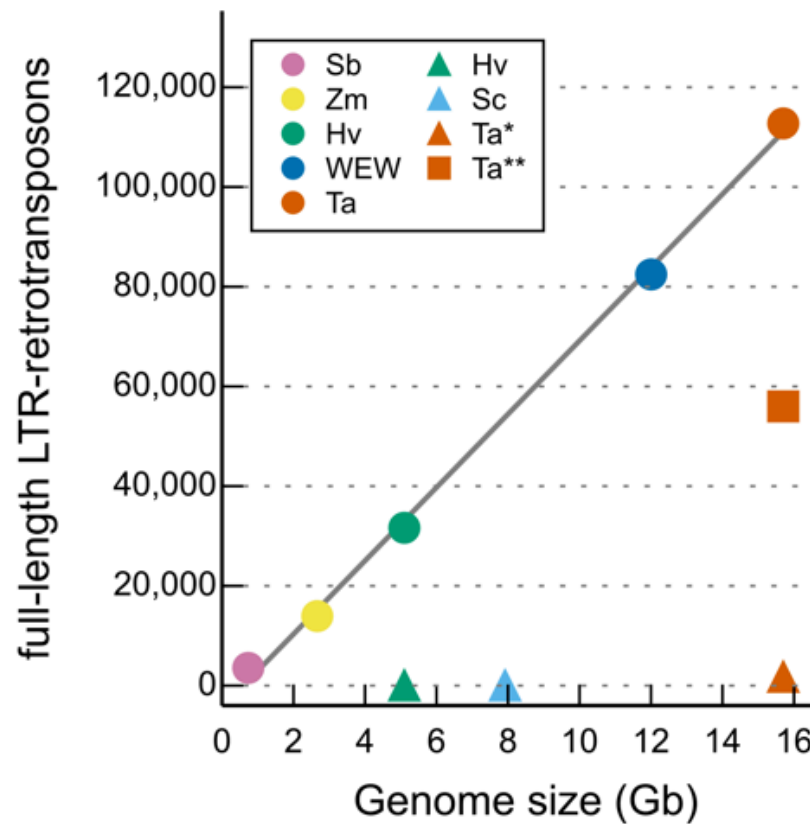
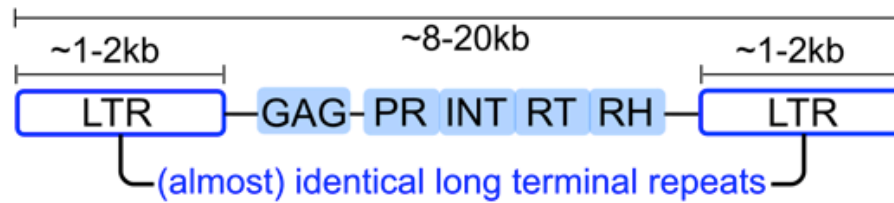
gene annotation and results for the IWGSC Refseq v1.0 genome assembly. (A) BUSCO gene model evaluation comparing previous bread wheat whole genome annotations with the current annotation and selected grass reference species.

BUSCO provides a measure for the recall of highly conserved gene models.

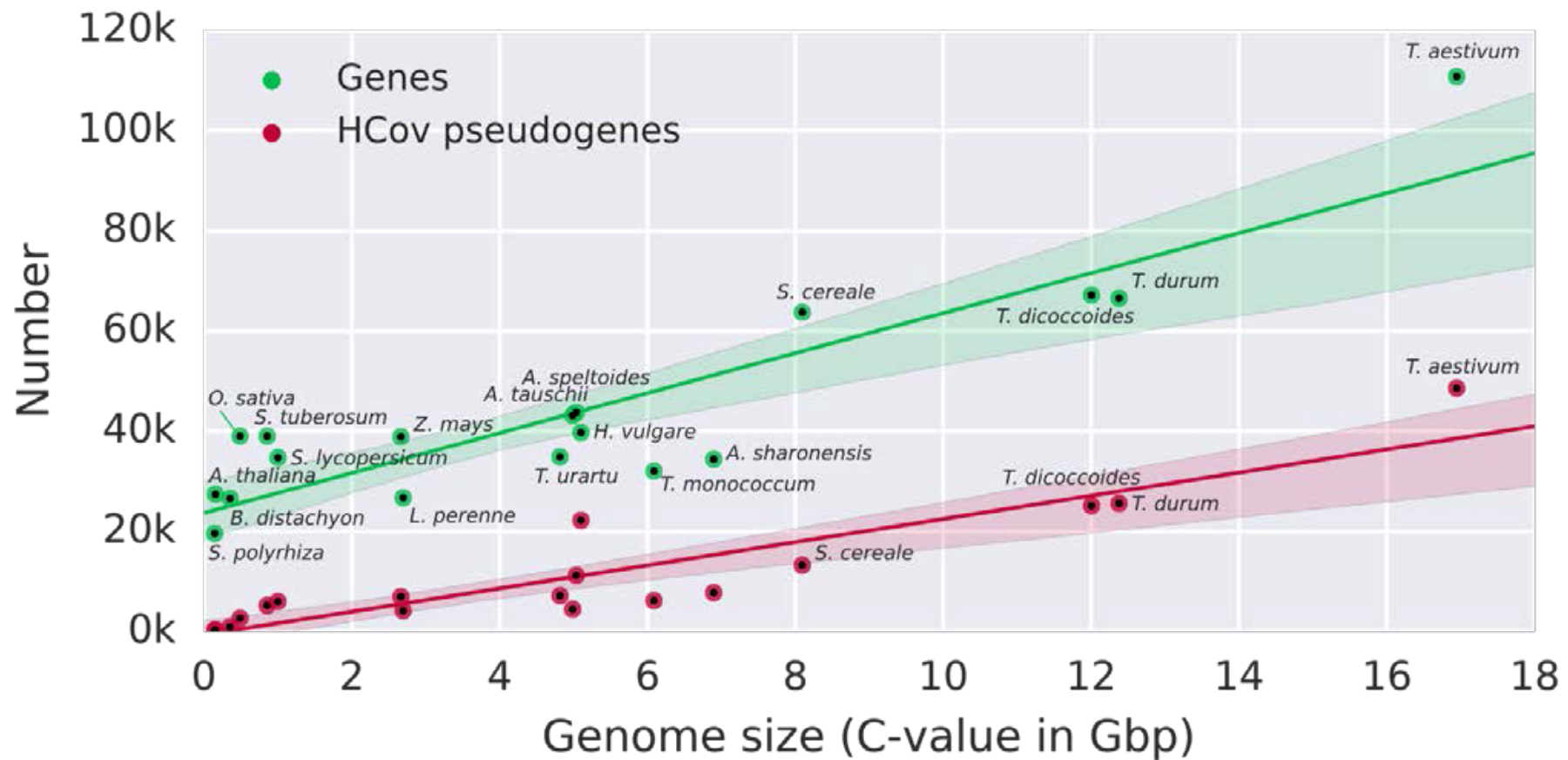


Triticeae Genomes

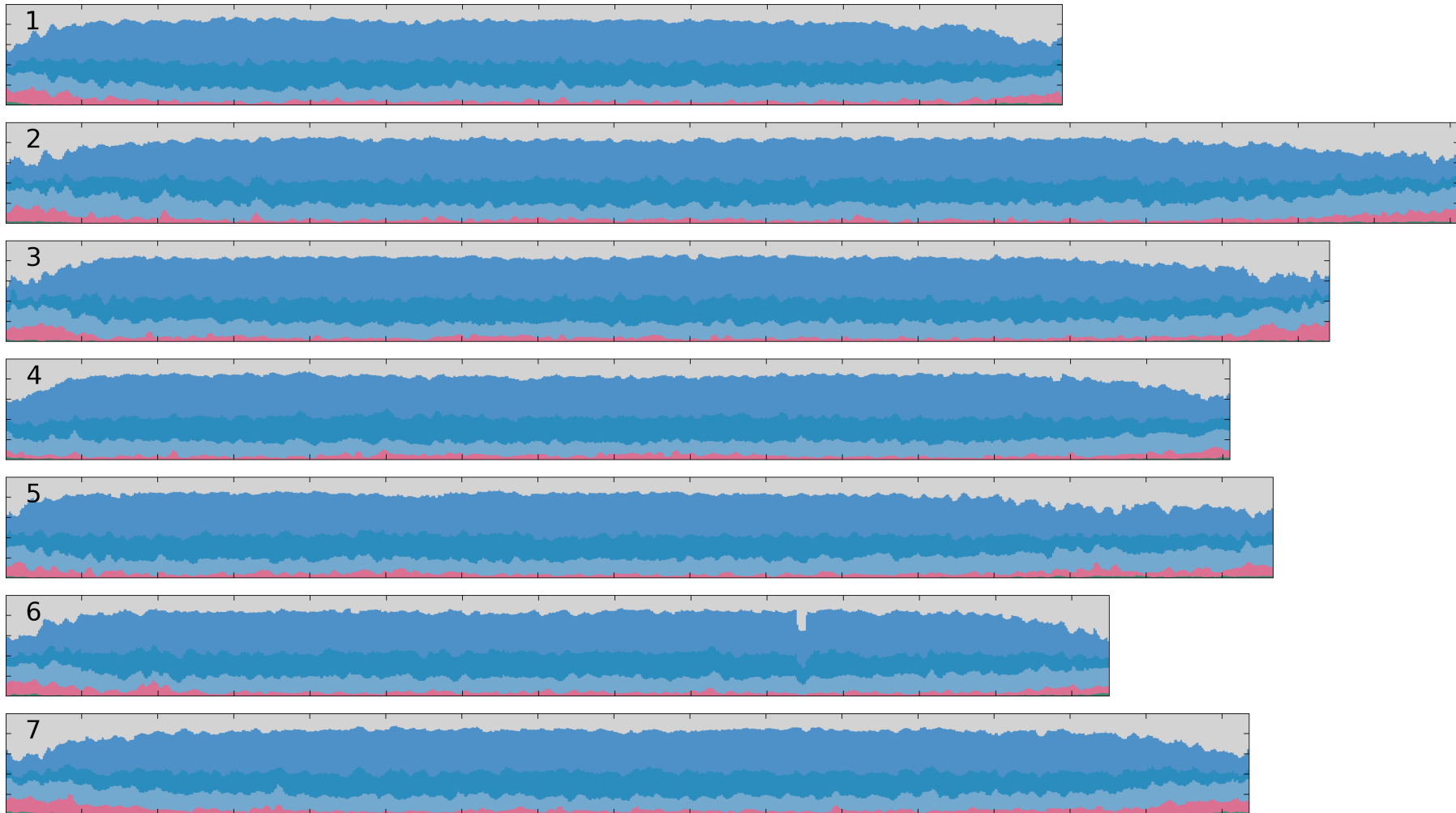
-Their Coming of Age-

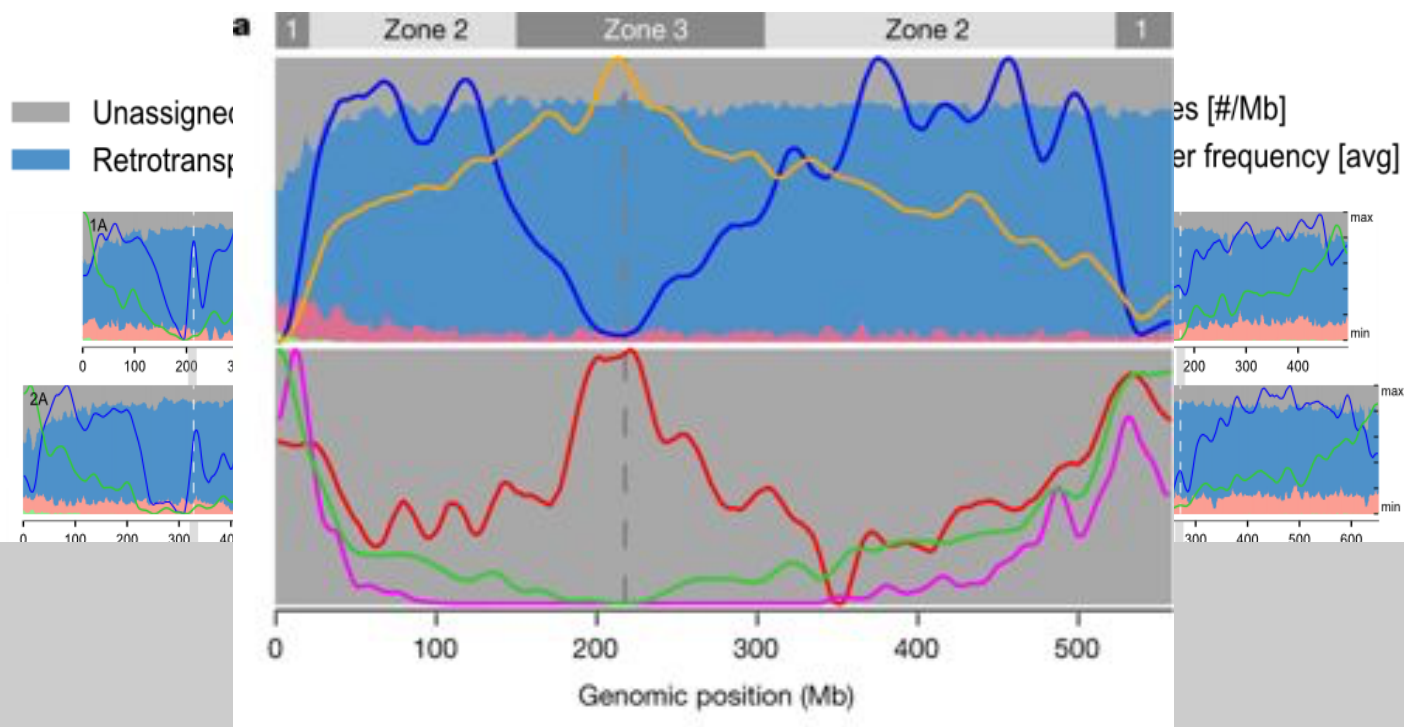


Gene and Pseudogene Number correlates with Genome Size



Genes = "needles in the haystack"





Distinct chromosomal compartments:

Zone 3: the „quite“ neighborhood:
centromer/housekeeping

Zone 2: the „industrial area“: middle parts

Zone 1: the „innovation quarter“: chromosome
ends/high recombination