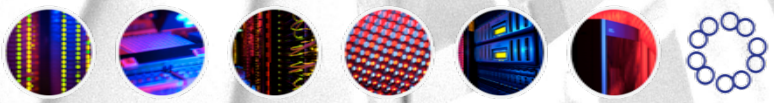
Photo: R. Warren

Scaffolding Genome Assemblies with Nanopore Reads

René Warren

ISMB, July 12th 2015



CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE
WWW.BCGSC.CA

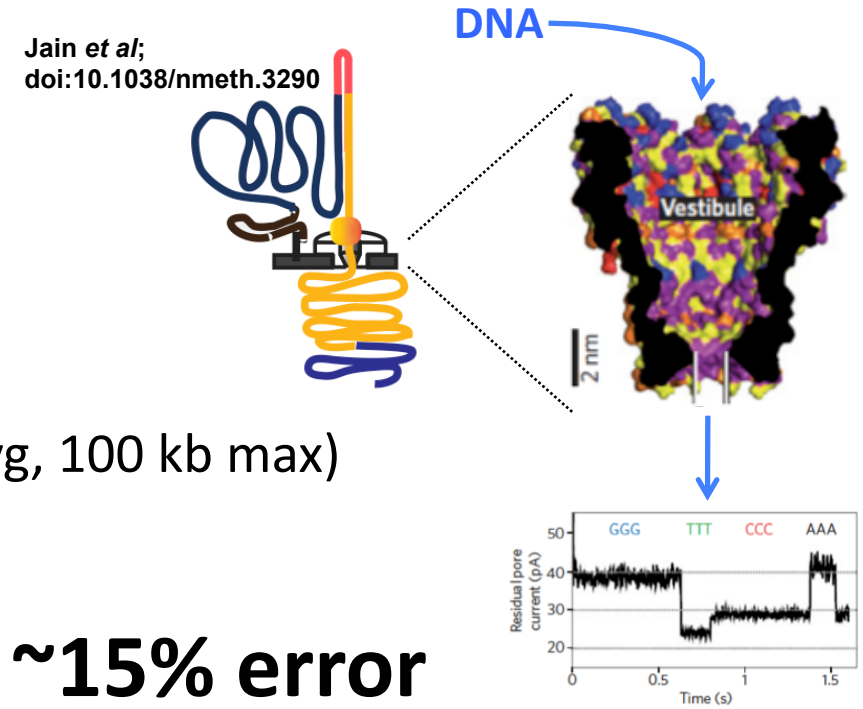
Genome Scaffolding

- **Ordered & oriented contigs**
  - Separated by gaps/overlaps based on distance info
- **Goal: one genetic element per scaffold**
  - Recovery of complete genes
- **Retrospective scaffolding**
  - Many high-quality fragmented assemblies
  - Advantageous while long read of low quality

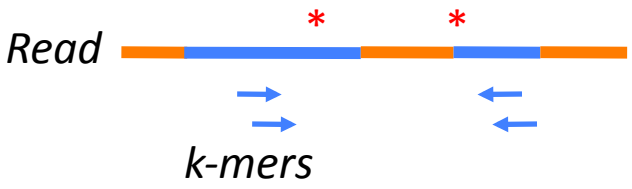
Existing Methods

- **Technologies**
  - SSPACE-LR, AHA
  - Bambus, SOPRA, Celera Assembler (CA), Others
- **Alignment-based**
  - Fast, memory-efficient
  - Scalability, error tolerance : depends on aligner
  - ***Variable success on error-rich long reads***
- **Limiting pairing info**
  - Standard “paired” sequence libraries
  - Specialized (eg. MPET) long fragment libraries
  - ***Limited by fragment size of sequence library***

Oxford Nanopore Technologies (ONT)

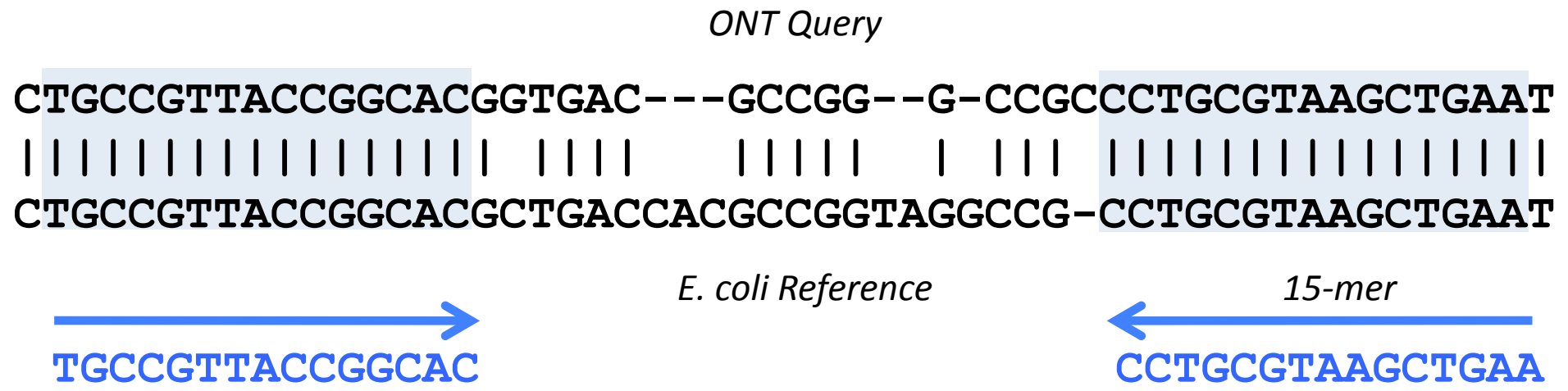


- **Single molecule**
- **Low cost** (< $1K per run)
- **Long sequences** (~5 kb avg, 100 kb max)
  - Template, Complement, 2D
- **Vendor R7 chemistry: ~15% error**
  - Stretches of indels + correct bases
  - Errors tend to cluster



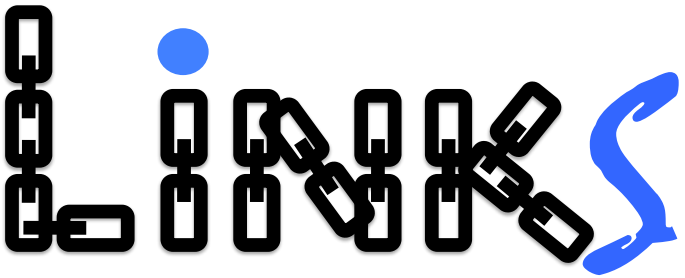
k-mers as Pairing Information

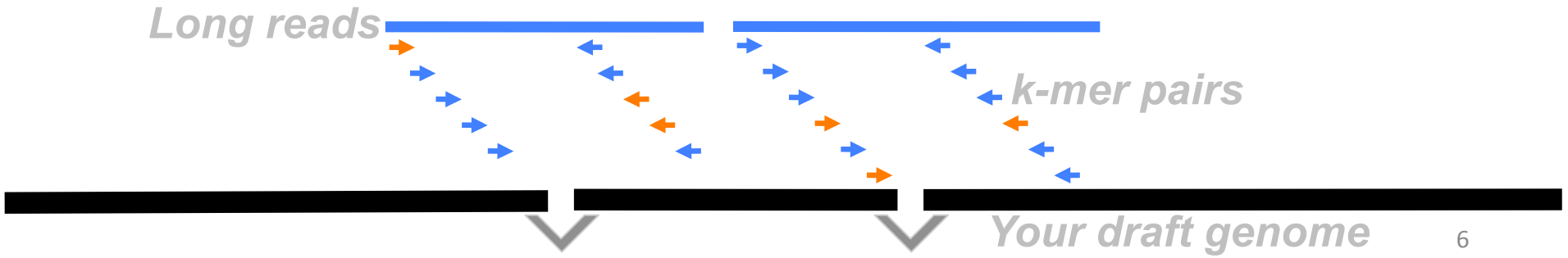
vast resource: *k*-mers ≅ bases in genome



Long Interval Nuc. K-mer Scaffolder

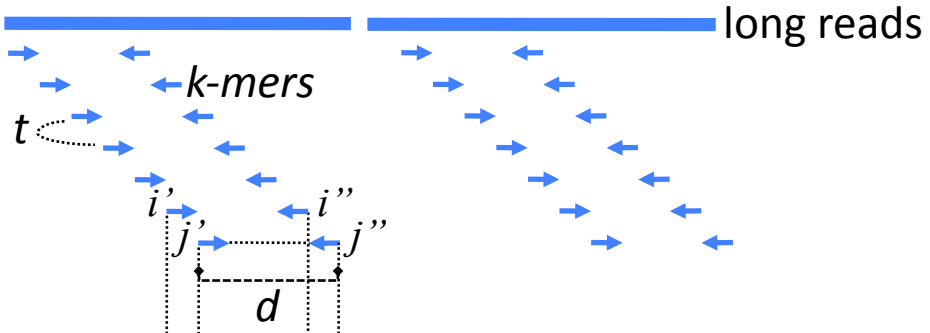
- **Uses SSAKE scaffolding engine (our lab)**
  - Predecessor to popular SSPACE:
    - SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension
- ***k*-mer based**
  - No read alignments
  - Explore vast *k*-mer space
    - not limited by fragment length
  - Works on long-reads, draft sequences
    - any length, almost any error profile, no base correction

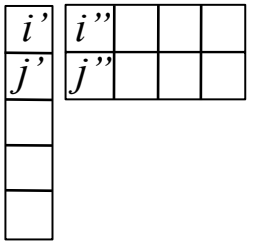




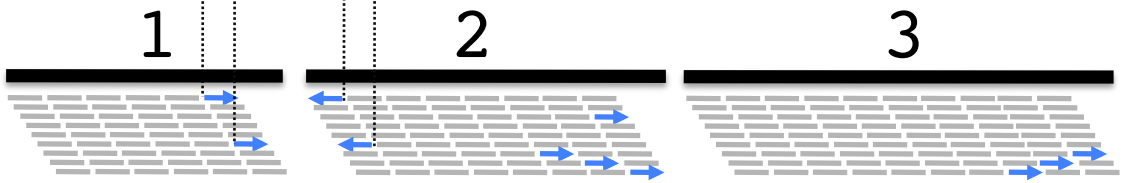
LINKS Algorithm

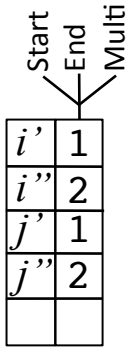
1
Extract *k*-mer pairs
from reads
(*d* intervals, *t* steps)



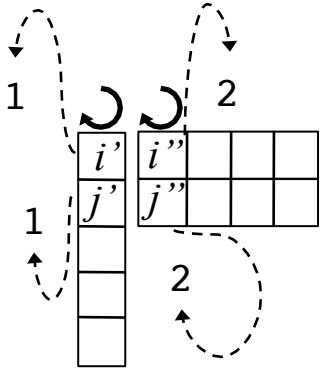


2
Extract *k*-mers
from contigs

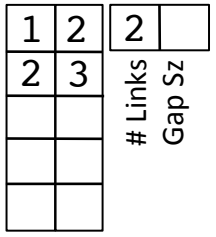




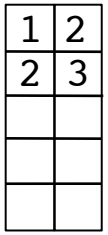
3
Pair Contigs
track orientation
& distance




```
1+2+ :2 links est. gap
2+3- :3 links est. overlap
```



4
Produce Layout

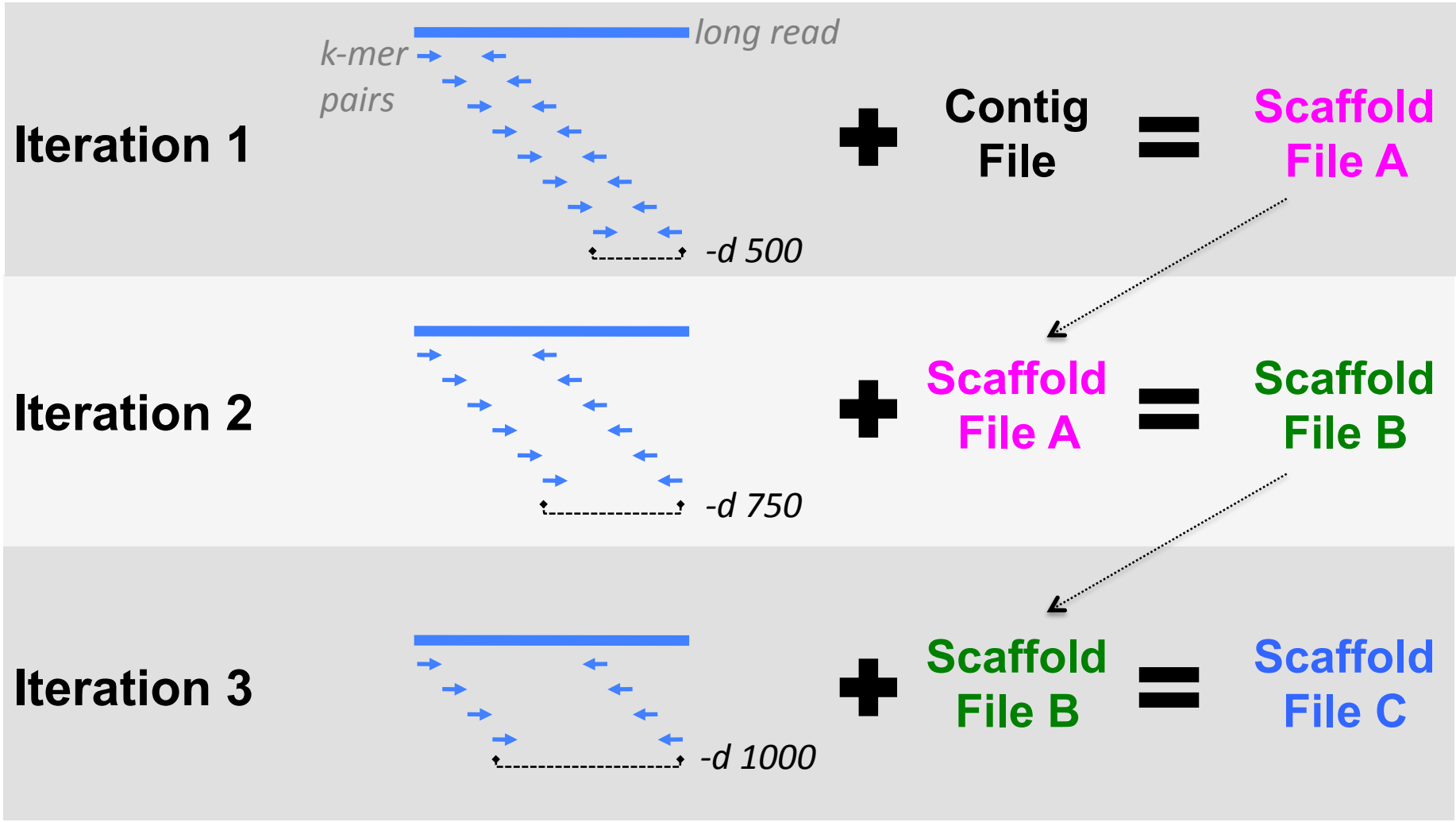


Iterative Extension
***All contigs incorporated into scaffold**
Influenced by heuristics:
–*l* minimum links
–*a* maximum links ratio

1+ ➔ 2+ ➔ 3–

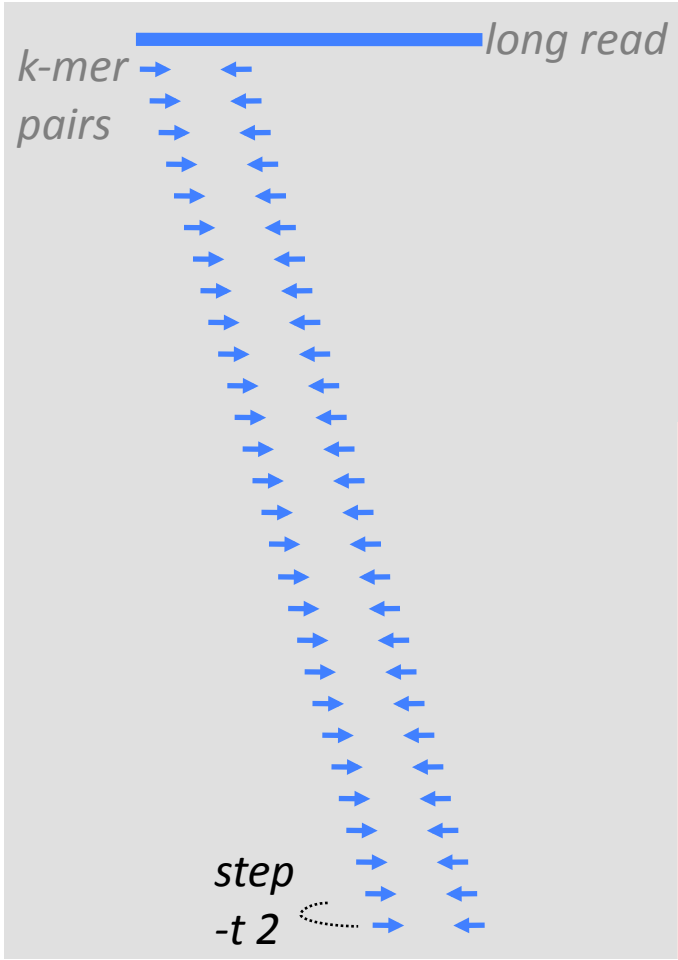
Practical Considerations

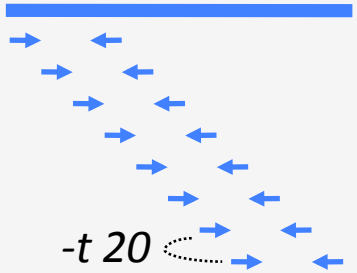
Iterative Scaffolding

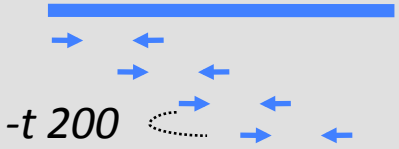


Practical Considerations

Minimizing Memory Usage

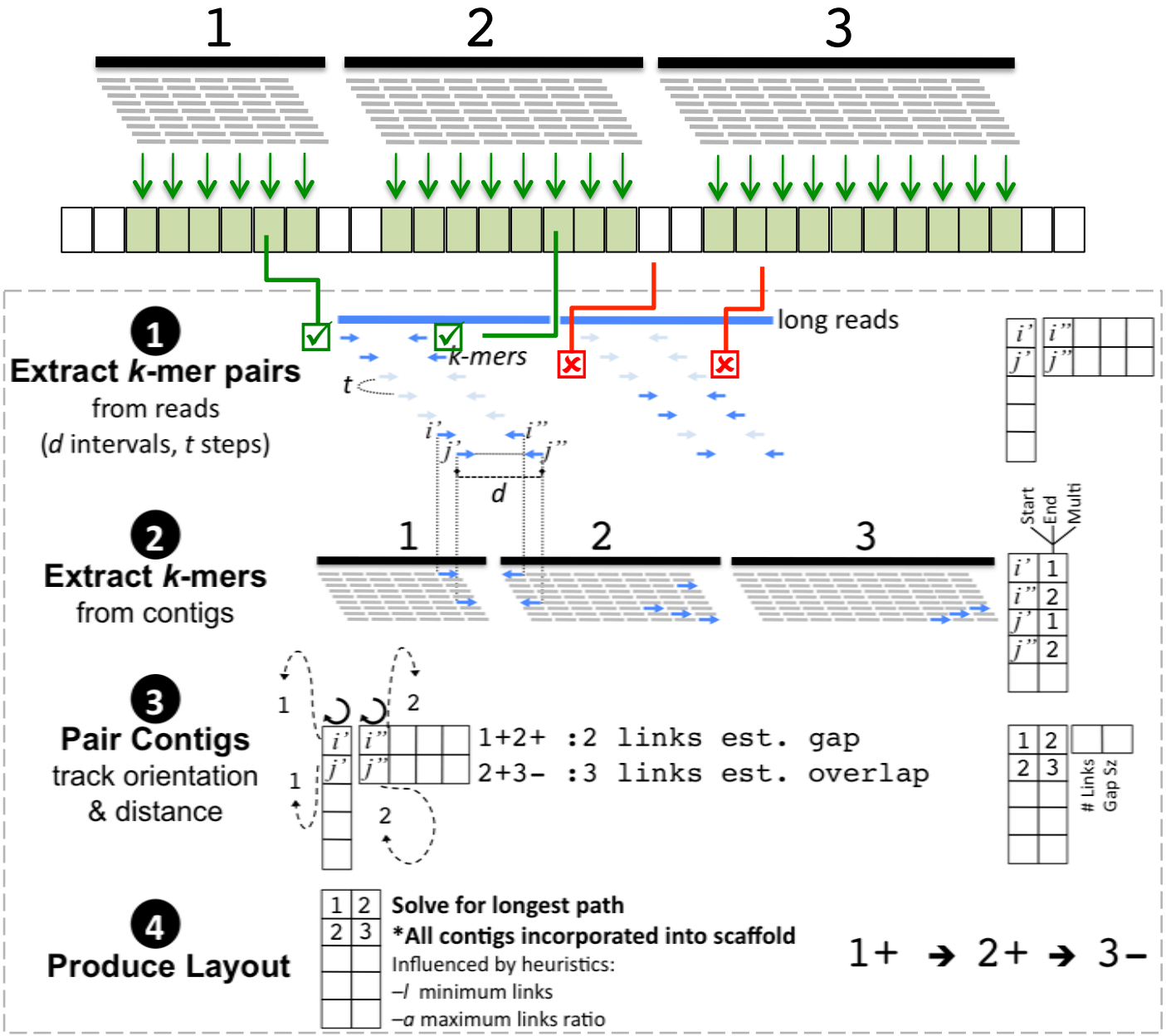






- Larger values of *t* will decrease the *k*-mer pair space
- Drawback: potential loss of information
- Adjust minimum required links *-l* to compensate
- Remediate with Bloom filter

Bloom Filter Implementation

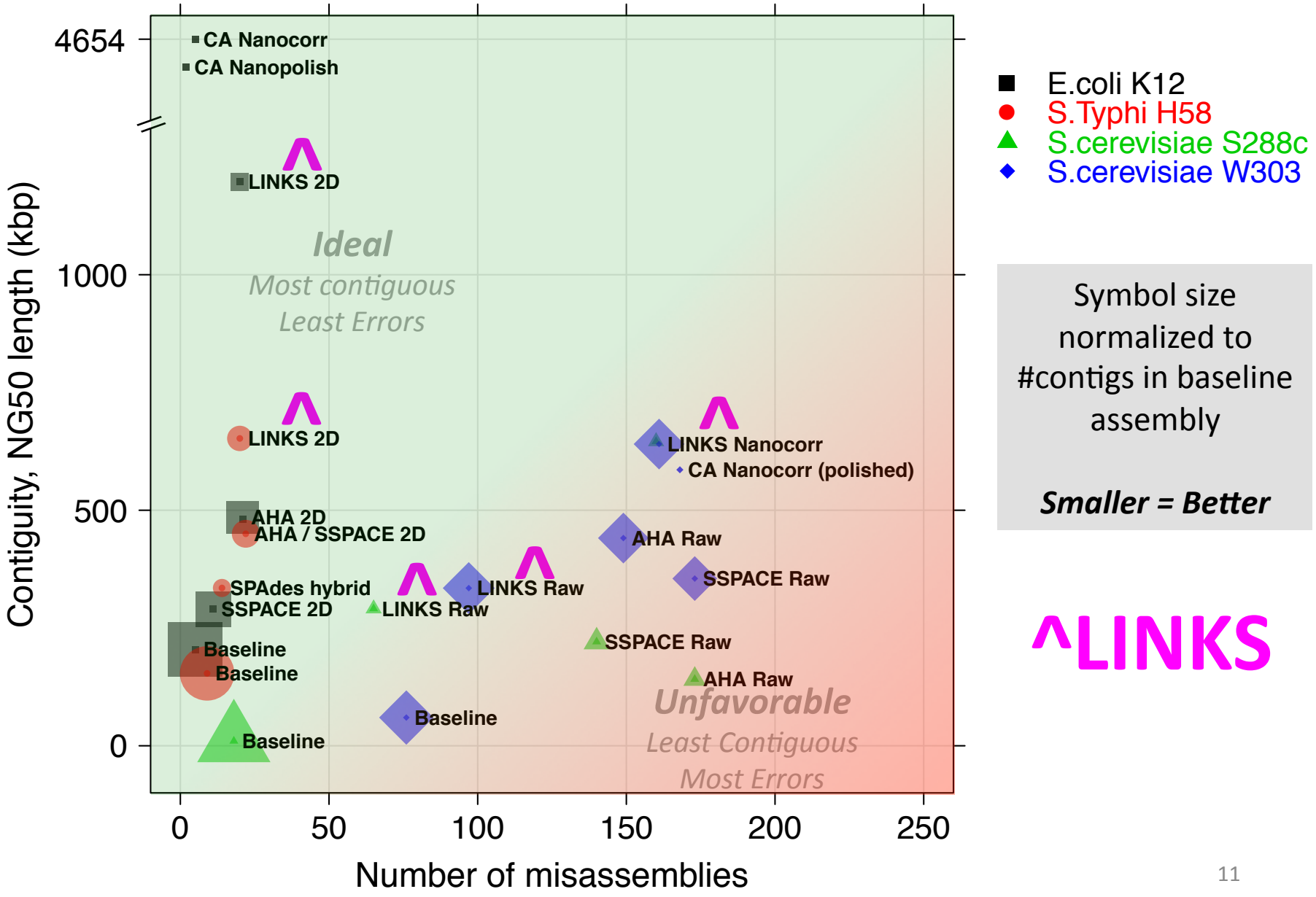


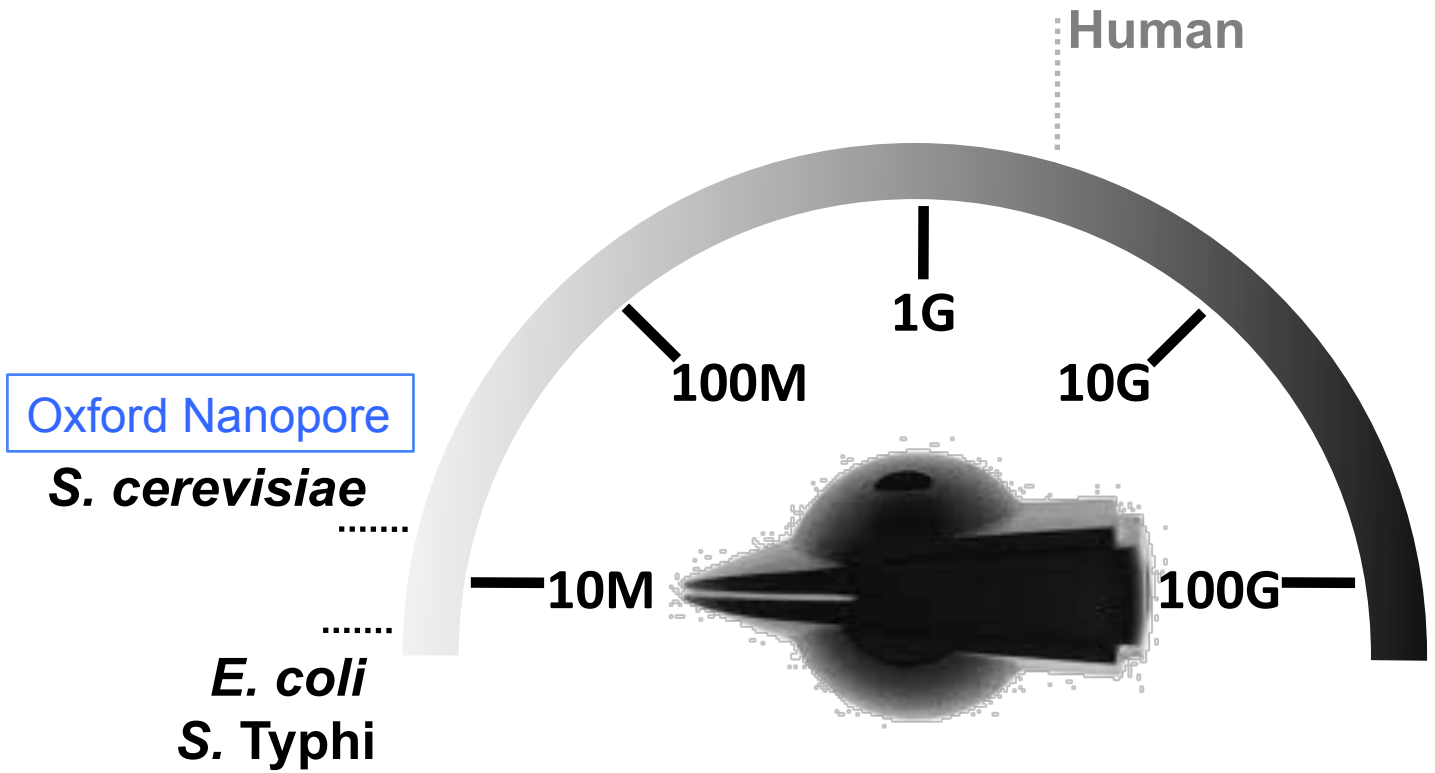
Contigs

Bloom filter

Succinct
Space-efficient
Probabilistic
data structure

Scaffolding with Public ONT Data

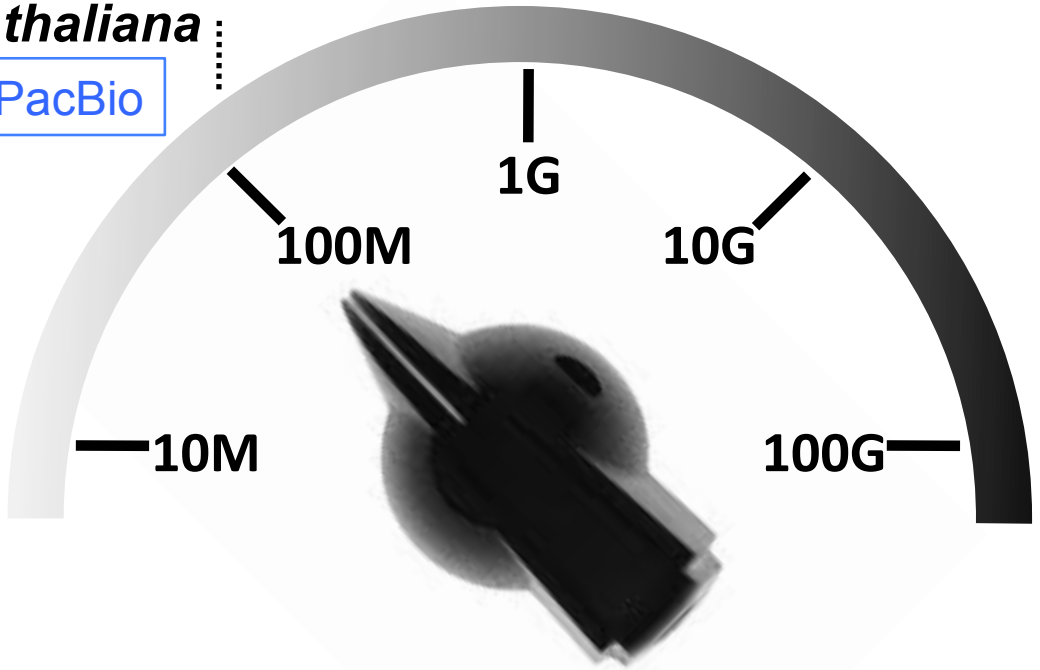



Human
1G
100M
10G
Oxford Nanopore
S. cerevisiae
10M
100G
E. coli
S. Typhi

~GENOME SIZE~

A. thaliana

PacBio



~GENOME SIZE~

Re-Scaffolding the 120-Mbp *Arabidopsis* Genome♮

Using raw or ECtools-corrected PacBio reads

1001 Genomes Data Center
A Catalog of *Arabidopsis thaliana* Genetic Variation

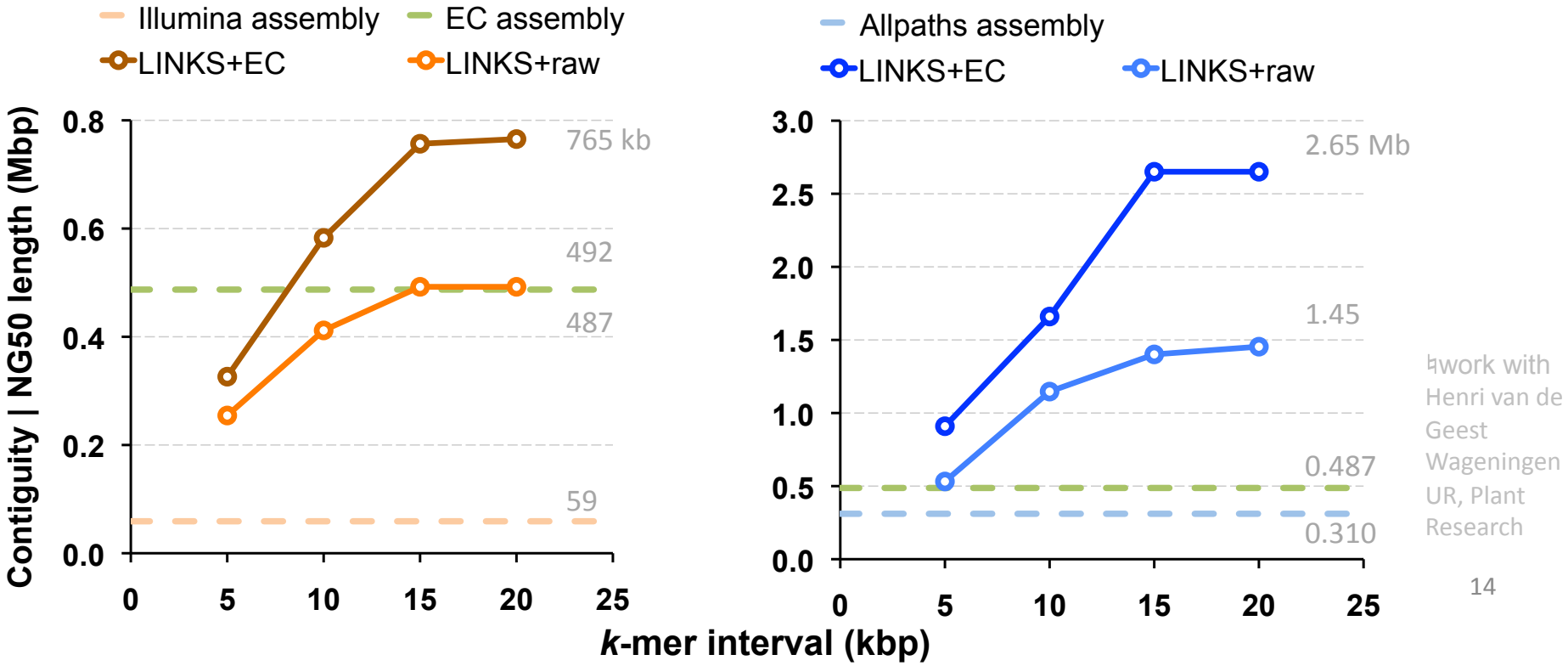
ECTools: Hybrid Error Correction Pipeline for long reads
http://schatzlab.cshl.edu/data/ectools/

http://1001genomes.org/data/MPI/MPISchneeberger2011/releases/current/Ler-1/

Illumina-only | **Raw pacbio reads (5Gb,40X)** | **EC reads**

Allpaths-LG

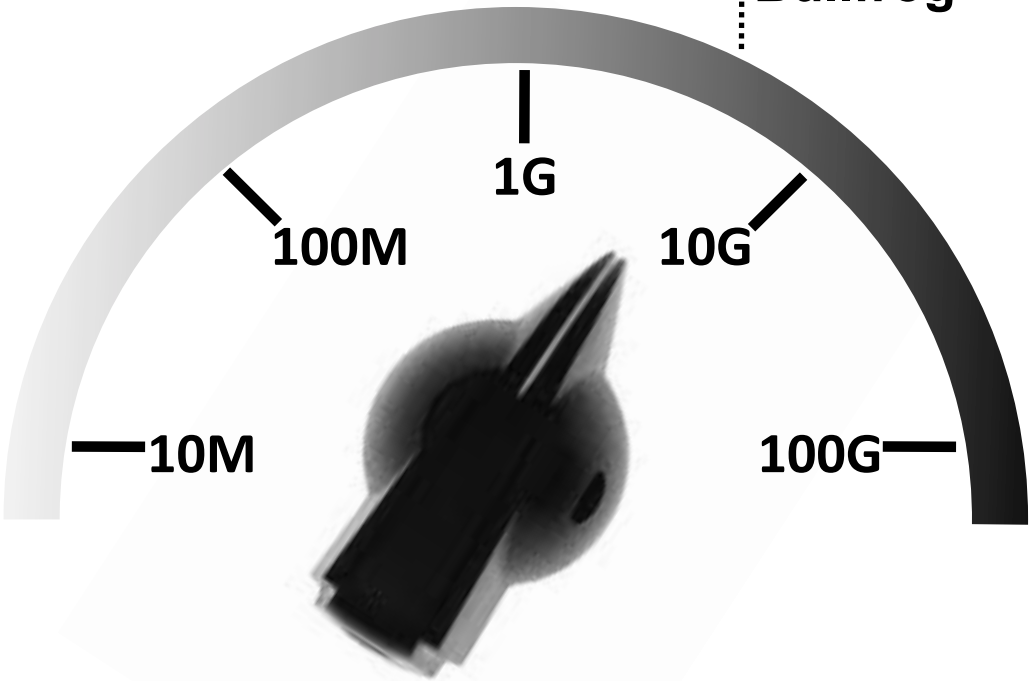
LINKS [v1.5] –k 21 –t 5 –l 5 –a 0.3



♮work with Henri van de Geest Wageningen UR, Plant Research

Moleculo
MPET

Bullfrog

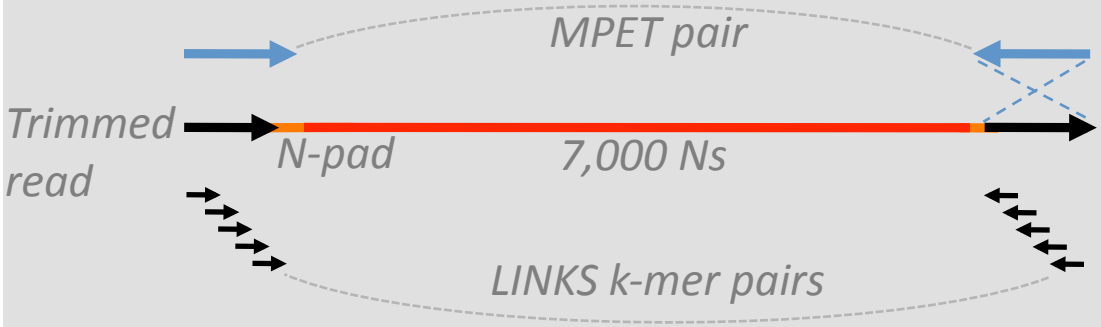


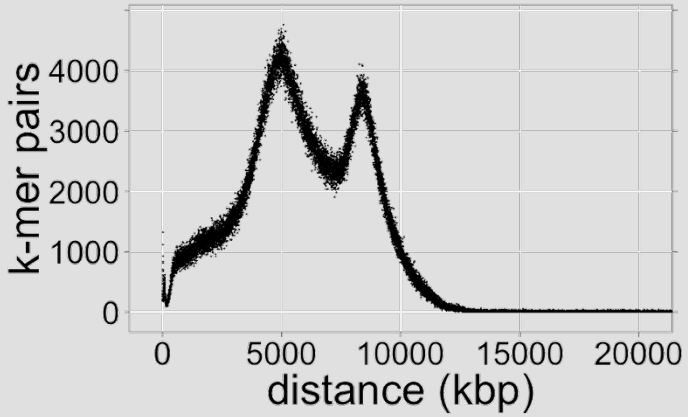
~GENOME SIZE~

Re-Scaffolding the 5-Gbp Bullfrog Genome♮

I. *Moleculo (TruSeq) – 0.4X coverage*
II. *MPET library (613.5M reads, 61Gb)*
III. *Another draft assembly (k128)*

Creating "long reads" from MPET

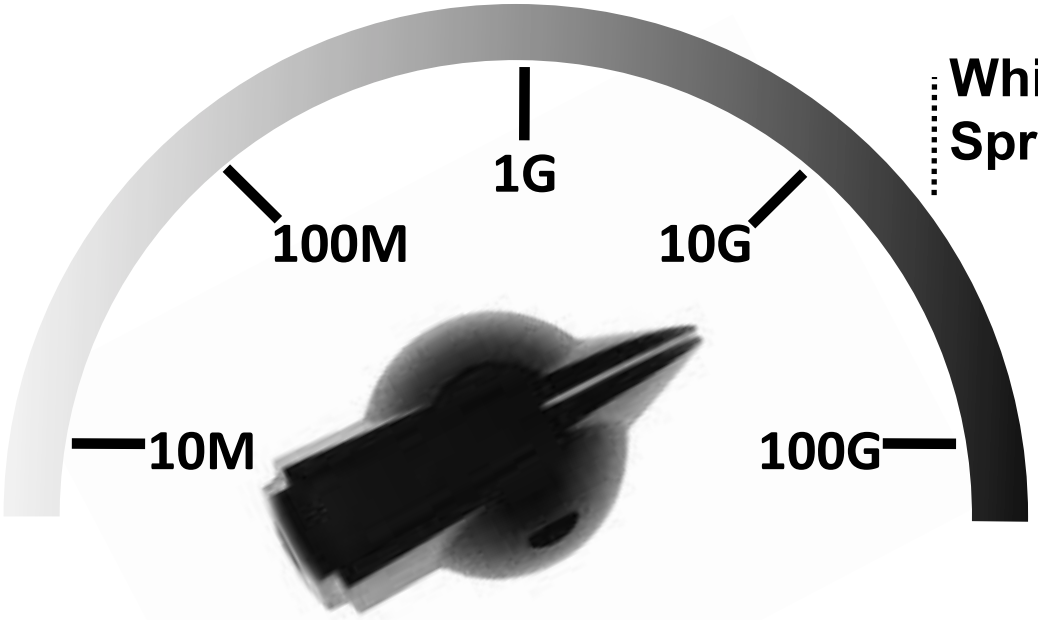




Bullfrog Assemblies	Original	1. Moleculo -l 5 -a 0.3 x14	2. MPET -l 10* -a 0.5	3. Draft *k*128 -l 5 –a 0.3 x14
#merges	NA	7,811	121,591	58,592
NG50 length (bp)	13,925	14,389	27,040	43,138

*≈2 MPET pairs support the main linkage

♮work with Caren Helbing, UVic



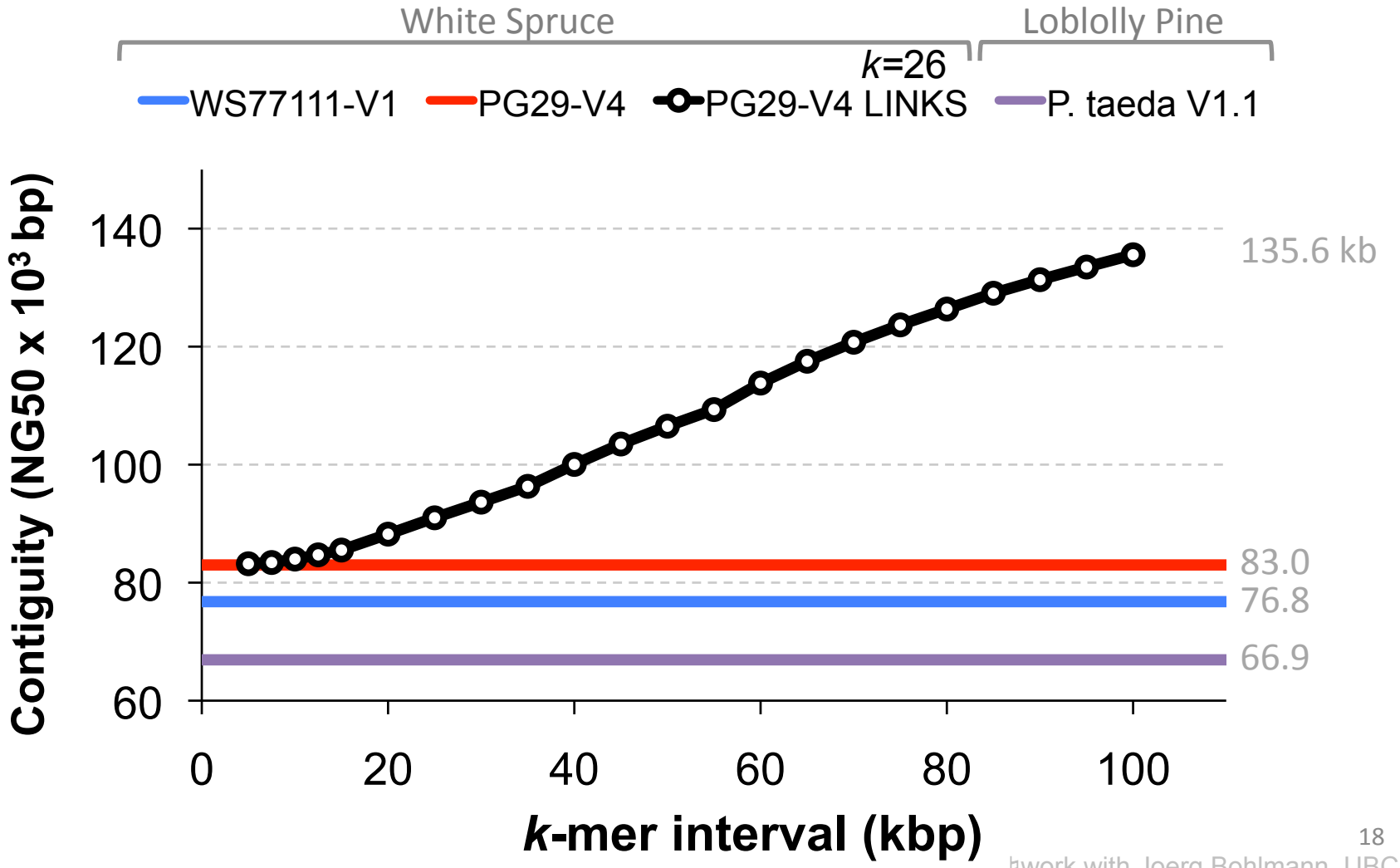
White Spruce

Genotype WGA

~GENOME SIZE~

Re-Scaffolding the 20-Gbp White Spruce Genome♮

Using another genotype draft genome as long reads



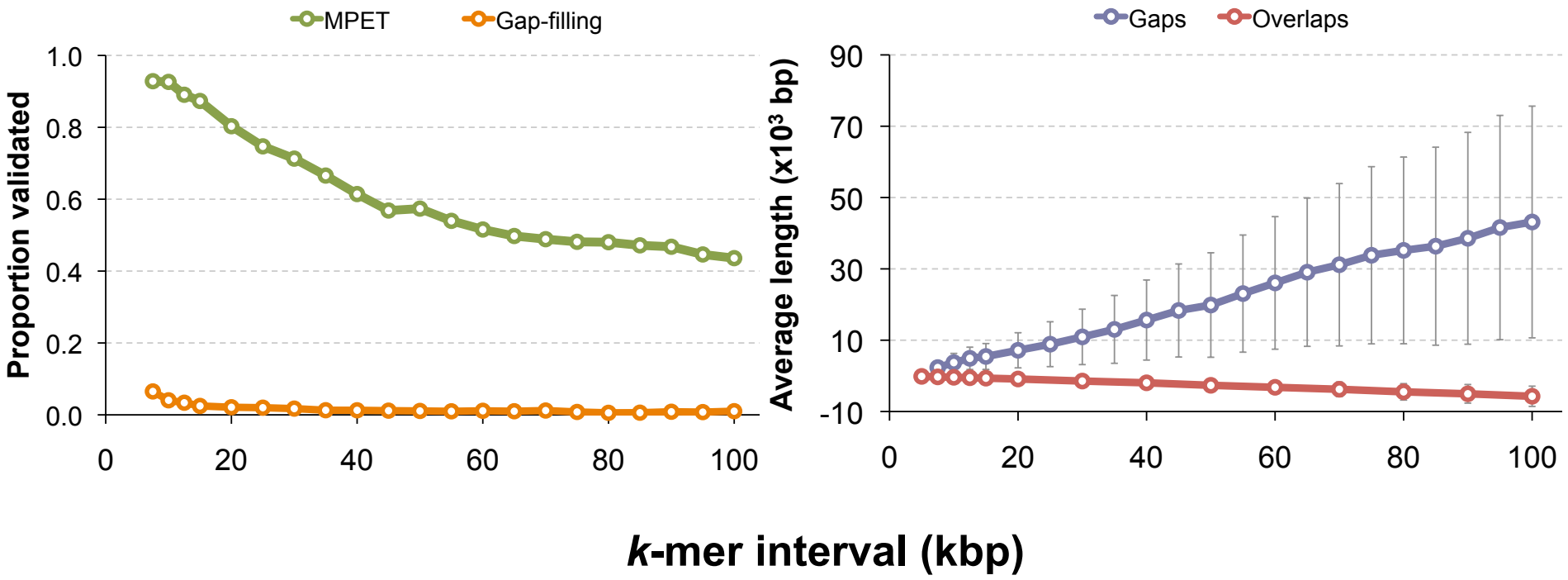
♮work with Joerg Bohlmann, UBC

Re-Scaffolding the 20-Gbp White Spruce Genome

Using another genotype draft genome as long reads

Validation

5, 8, 12 kbp MPET libs

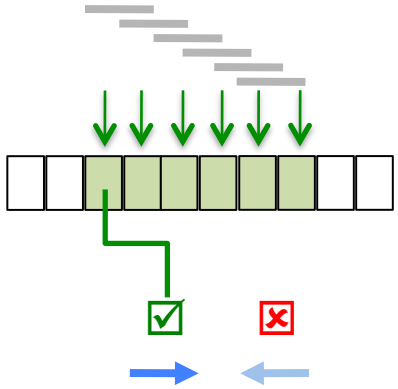


***k*-mer interval (kbp)**

Useful Tips

Run iteratively

- Increasing distances (-d↗)
- Decreasing steps (-t↘)



Create a Bloom filter

- Genome draft *k*-mer Bloom filter
- Built at first iteration / pre-computed

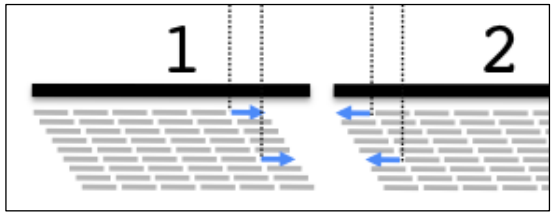
.bloom

Re-use filter

- Re-utilize the Bloom filter each iteration

Summary

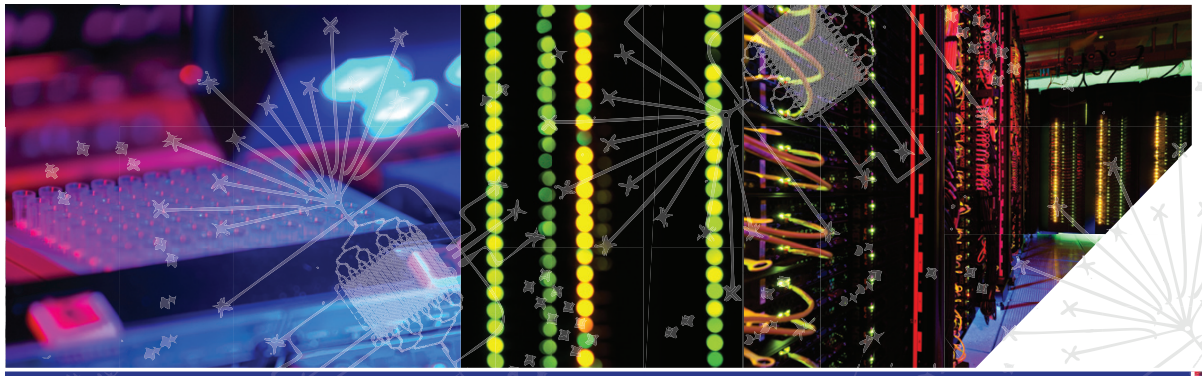
- **Alignment-free | Varied Read Source | Scalable**



Pacbio
ONT
Drafts WGA
MPET

Future Work

- **New Bloom filter implementation**
  - Faster run time, esp. on large genomes
- **Port codebase to C++**
  - Benefit from extensive parallel libraries
- **Native iterative scaffolding**
  - Determine *-t*, *-d*, Bloom filter specs based on data size, resources available



CANADA'S MICHAEL SMITH

GENOME SCIENCES CENTRE

A leading international centre for genomics and bioinformatics research committed to advancing knowledge of cancer-related diseases, improving human health through disease prevention, diagnosis and therapeutic approaches, and realizing social and economic benefits of genomics research.

54 TERABASES SEQUENCED • A HUMAN GENOME EVERY 17 MINUTES • HIGH-PERFORMANCE COMPUTING

AFFILIATIONS BC Cancer Research Center • BC Cancer Agency • BC Cancer Foundation • Genome BC • Simon Fraser University • University of Britsh Columbia • Genome Sciences Institute

Thank You!

http://www.bcgsc.ca/platform/bioinfo/software/links

Chen Yang | Ben Vandervalk | Albert Lagman | Bahar Behsaz | Steve Jones | Inanc Birol

NIH

BC CANCER FOUNDATION
partners in discovery

GenomeBritishColumbia

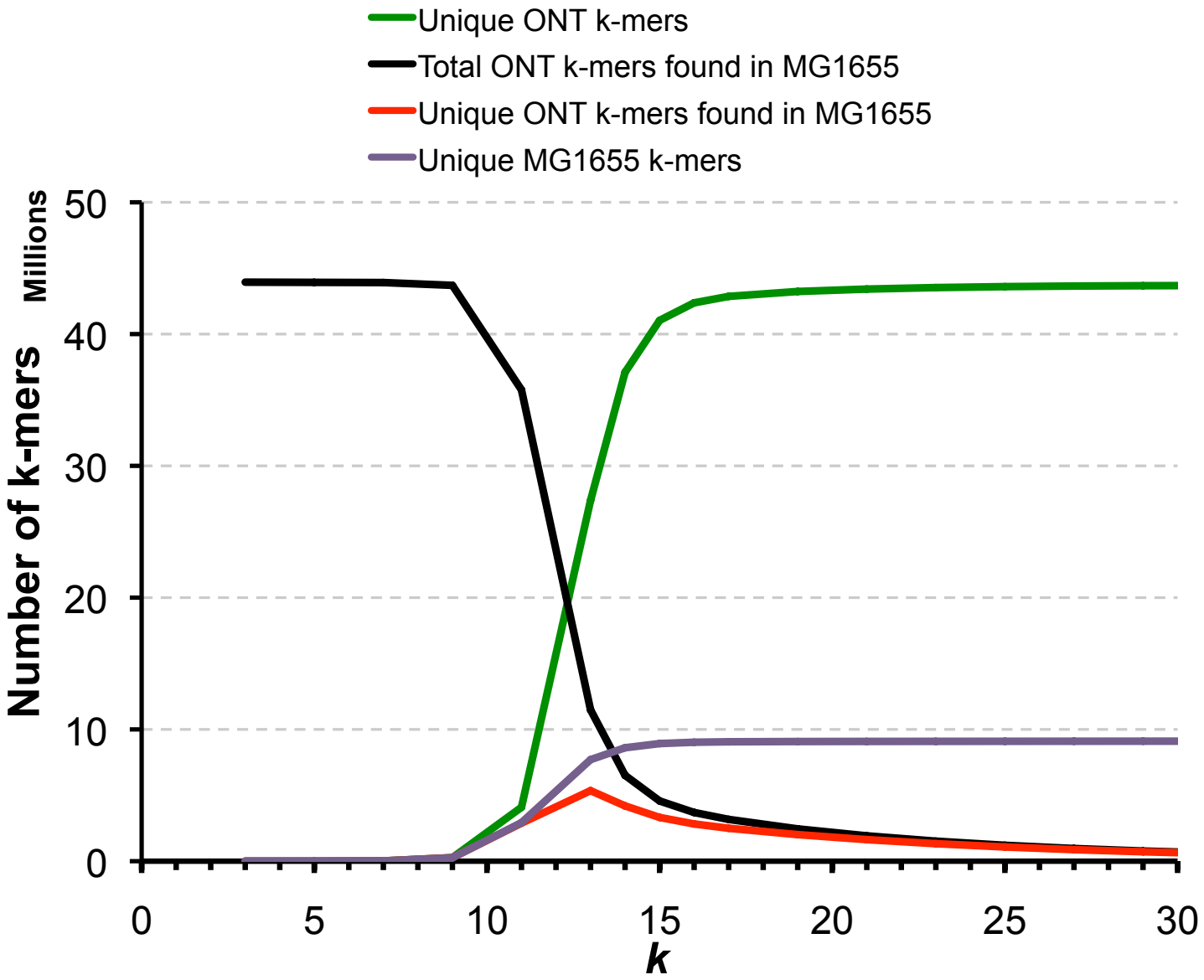
GenomeCanada

intel

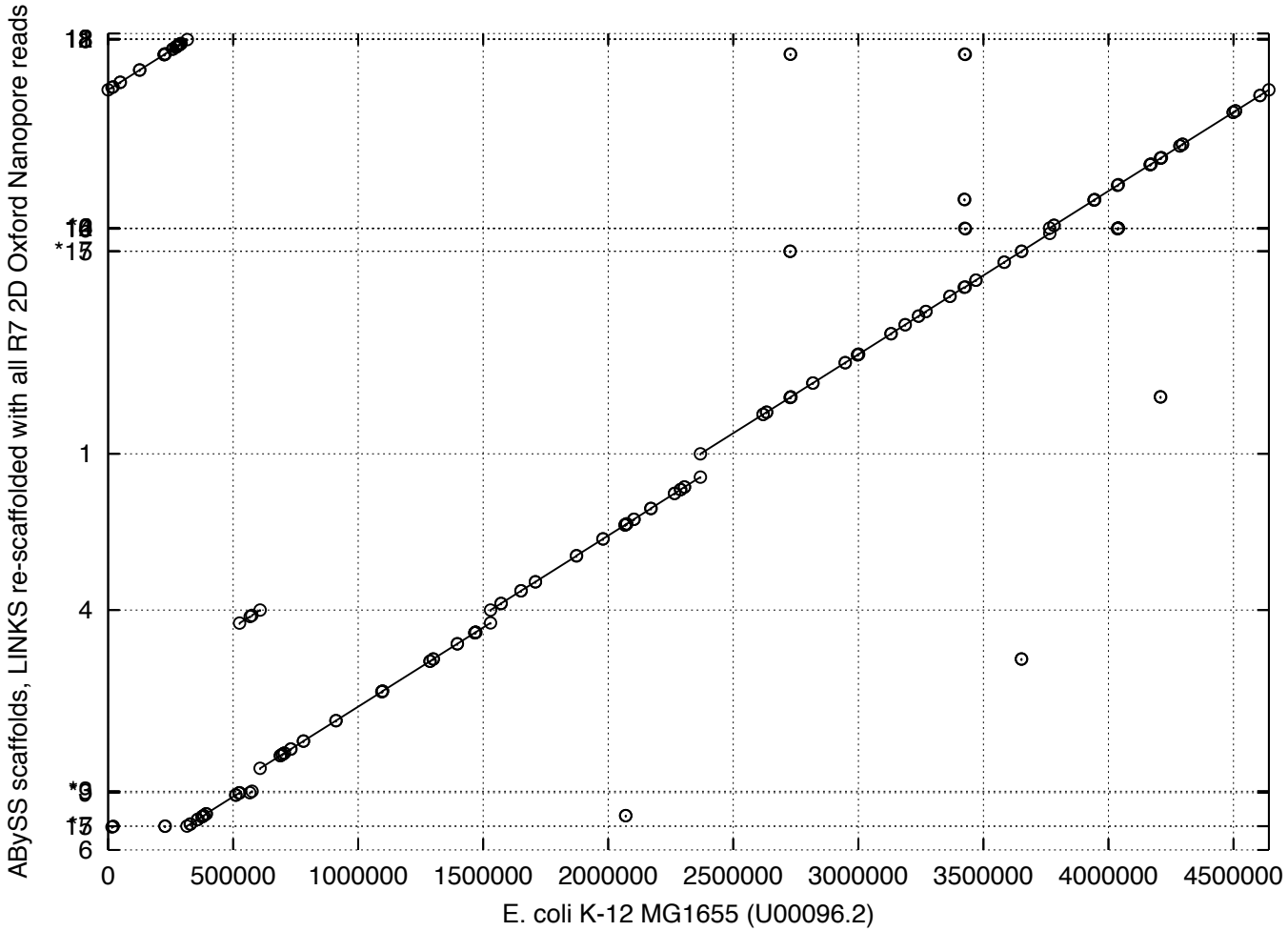
John Jambor Knowledge fund

Photo: R. Warren

Choosing *k*

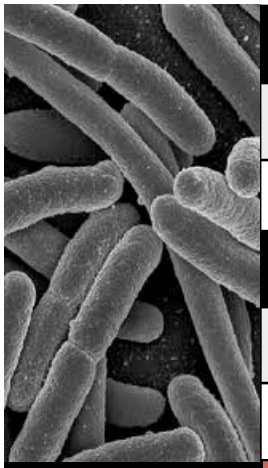

Unique ONT k-mers
Total ONT k-mers found in MG1655
Unique ONT k-mers found in MG1655
Unique MG1655 k-mers
Number of k-mers
Millions
50
40
30
20
10
0
0
5
10
15
20
25
30
k

LINKS Scaffolds Co-linearity



LINKS scaffolds using all available R7 2D ONT reads compared to the reference *E. coli* K-12 genome (30 *k-mer* pair interval runs). Iterative LINKS scaffolding rounds (*k*=15, *d*=500 to 16000 bp, 30 iterations) were performed on ABySS assembly sequence scaffolds (Table 1**G** in manuscript), bringing the number of scaffolds further down to 16 from 61. MUMmer co-linear analysis indicates that six large scaffolds comprise *E. coli* K-12 MG1655 re-scaffolded sequences in the correct order and orientation.

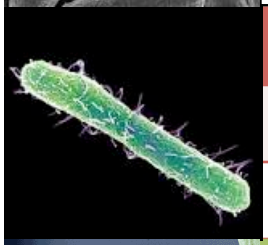
Computational Efficiency



E. coli	v1.3 (x1)	v1.5 (x1)
time (mm:ss)	**1:10**	1:33
RAM (GB)	6.4	**2.1**

BLASR

E. coli	v1.3 (x30)	v1.5 (x30)	AHA	SSPACE-LR
time (mm:ss)	22:02	17:43	48:34	**7:15**
RAM (GB)	13.3	4.3	1.57	**0.3**



S. Typhi	v1.3 (x11)	v1.5 (x11)	AHA	SSPACE-LR
time (mm:ss)	28:43	21:28	10:11	**1:35**
RAM (GB)	26.4	8.2	0.7	**0.5**

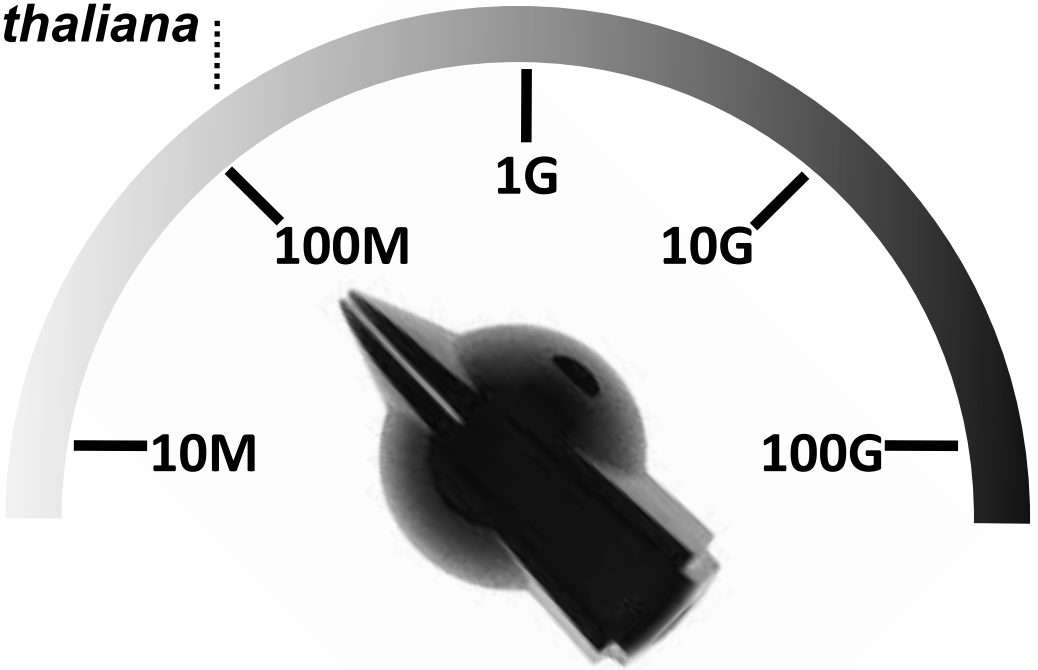
S288c	v1.3 (x29)	v1.5 (x29)	AHA	SSPACE-LR
time (h:mm:ss)	6:28:33	3:26:36	9:06:07	**1:53:46**
RAM (GB)	117.9	35.1	115.6	**6.4**

W303	v1.3 (x27)	v1.5 (x27)	AHA	SSPACE-LR
time (h:mm:ss)	6:18:54	3:15:20	10:02:21	**2:04:49**
RAM (GB)	118.0	35.1	115.7	**6.8**

Benchmark: Intel(R) Xeon(R) CPU E5-2699 v3 @ 2.30GHz, 72 CPUs, 132GB RAM

BF size (elements) estimated using file size and FPR 0.01%

A. thaliana



~GENOME SIZE~

Re-Scaffolding the 120-Mbp *Arabidopsis* Genome♮

Using raw or ECtools-corrected PacBio reads

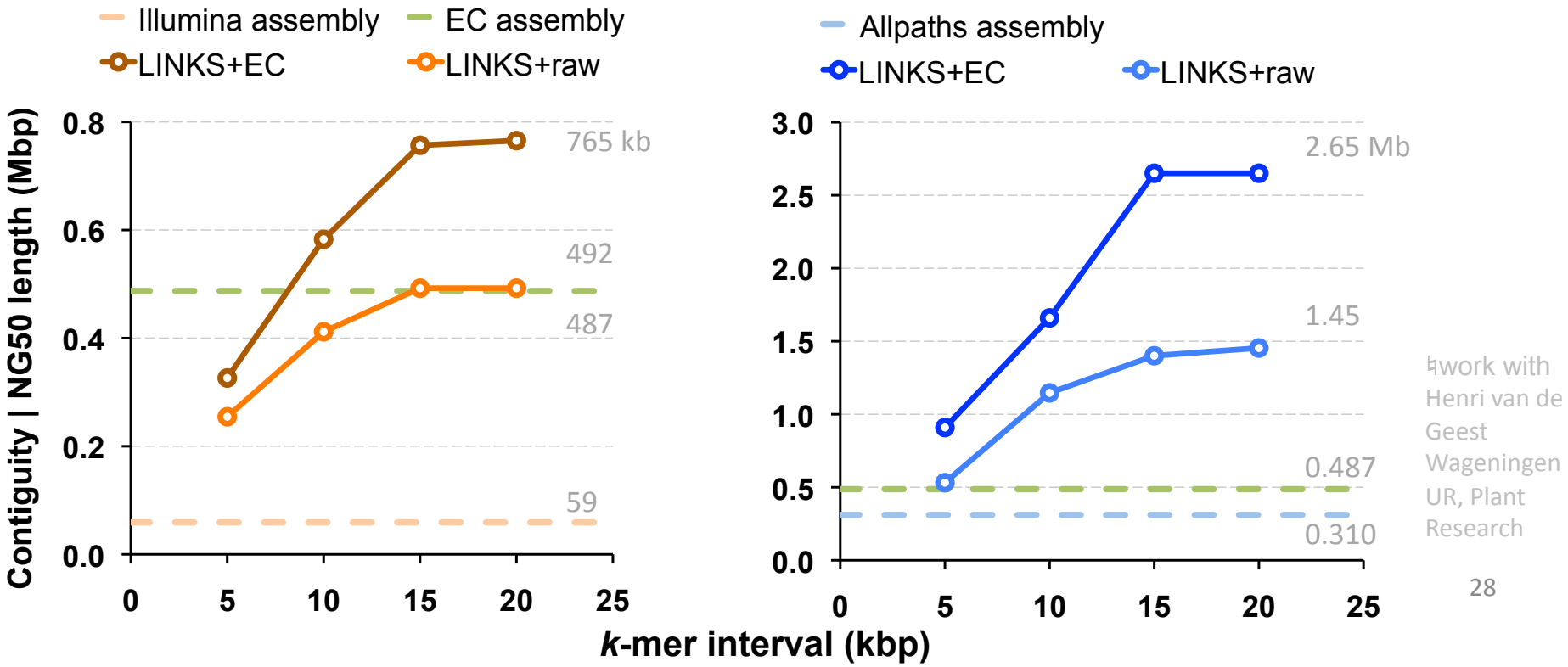
1001 Genomes Data Center
A Catalog of *Arabidopsis thaliana* Genetic Variation

ECTools: Hybrid Error Correction Pipeline for long reads
http://schatzlab.cshl.edu/data/ectools/

http://1001genomes.org/data/MPI/MPISchneeberger2011/releases/current/Ler-1/

Illumina-only | **Raw pacbio reads (5Gb,40X)** | **EC reads** — **Allpaths-LG**

LINKS [v1.5] –k 21 –t 5 –l 5 –a 0.3



♮work with Henri van de Geest Wageningen UR, Plant Research

	Input Libraries	Total Input Bases (Genome coverage)	Corrected Mean Read Lengths (Corrected Coverage)	Max Contig	N50 Size (Assembly Performance)	N50 Cnt	Percent Identity
A. thaliana (Ler-0)							
Genome	5 Chromo. + chloroplast + MT	120MB	-	30,427,671	23,459,830 (100.0)	3	
Illumina	MiSeq 2x300bp @ 450bp PE	13.8GB (115x)	-	282,909	54,525 (0.2)	649	99.97
HGAP	93 SMRTcells	14.2GB (118x, 38x over 10kb)	9719 +/- 4489 (21.28x, 15.44x > 10kb)	12,431,823	8,429,818 (35.9)	6	96.20
ECTools	19 SMRTcells	4.8GB (40x, 6x over 10kb)	2479 +/- 2323 (31.56x, 3.97x > 10kb)	3,841,500	616,869 (2.6)	54	99.91
PacBioToCA	19 SMRTcells	4.8GB (40x, 6x over 10kb)	2079 +/- 1989 (25.41x, 2.26x > 10kb)	1,618,669	365,151 (1.6)	90	99.81