Application of Accelerators and Parallel Computing to Sequencing Data

Wen-mei W. Hwu
University of Illinois at Urbana-Champaign

With
Deming Chen and Jian Ma

Blue Waters Computing System

Fully Operational at Illinois since 3/2013

11.1 PF
1.5 PB DRAM

10/40/100 Gb Ethernet Switch

IB Switch

>1 TB/sec

120+ Gb/sec

100 GB/sec

WAN

Spectra Logic: 300 PBs

Sonexion: 26 PBs

ILLINOIS
1867
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Blue Waters and Titan Computing Systems

System Attribute	NCSA Blue Waters	ORNL Titan
Vendors	Cray/AMD/NVIDIA	Cray/AMD/NVIDIA
Processors	Interlagos/Kepler	Interlagos/Kepler
Total Peak Performance (PF)	11.1	27.1
Total Peak Performance (CPU/GPU)	7.1/4	2.6/24.5
Number of CPU Chips	48,352	18,688
Number of GPU Chips	3,072	18,688
Amount of CPU Memory (TB)	1511	584
Interconnect	3D Torus	3D Torus
Amount of On-line Disk Storage (PB)	26	13.6
Sustained Disk Transfer (TB/sec)	>1	0.4-0.7
Amount of Archival Storage	300	15-30
Sustained Tape Transfer (GB/sec)	100	7

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Science Area	Number of Teams	Codes	Struct Grids	Unstruct Grids	Dense Matrix	Sparse Matrix	N-Body	Monte Carlo	FFT	PIC	Sig I/O
Climate and Weather	3	CESM, GCRM, CM1/WRF, HOMME	X	X		X		X			X
Plasmas/ Magnetosphere	2	H3D(M),VPIC, OSIRIS, Magtail/UPIC	X				X		X		X
Stellar Atmospheres and Supernovae	5	PPM, MAESTRO, CASTRO, SEDONA, ChaNGa, MS-FLUKSS	X			X	X	X		X	X
Cosmology	2	Enzo, pGADGET	X			X	X				
Combustion/ Turbulence	2	PSDNS, DISTUF	X						X		
General Relativity	2	Cactus, Harm3D, LazEV	X			X					
Molecular Dynamics	4	AMBER, Gromacs, NAMD, LAMMPS				X	X		X		
Quantum Chemistry	2	SIAL, GAMESS, NWChem			X	X	X	X			X
Material Science	3	NEMOS, OMEN, GW, QMCPACK			X	X	X	X			
Earthquakes/ Seismology	2	AWP-ODC, HERCULES, PLSQR, SPECFEM3D	X	X			X				X
Quantum Chromo Dynamics	1	Chroma, MILC, USQCD	X		X	X					
Social Networks	1	EPISIMDEMICS									
Evolution	1	Eve									
Engineering/System of Systems	1	GRIPS,Revisit						X			
Computer Science	1			X	X	X			X		X

Initial Production Use Results

- NAMD
  - 100 million atom benchmark with Langevin dynamics and PME once every 4 steps, from launch to finish, all I/O included
  - 768 nodes, Kepler+Interlagos is 1.8X faster over Interlagos+Interlogs
- Chroma
  - Lattice QCD parameters: grid size of 48^3 x 512 running at the physical values of the quark masses
  - 768 nodes, Kepler+Interlagos is 2.4X faster over Interlagos+Interlogos
- QMCPACK
  - Full run Graphite 4x4x1 (256 electrons), QMC followed by VMC
  - 700 nodes, Kepler+Interlagos is 2.7X faster over Interlagos+Interlogos

Cost per Raw Megabase of DNA Sequence

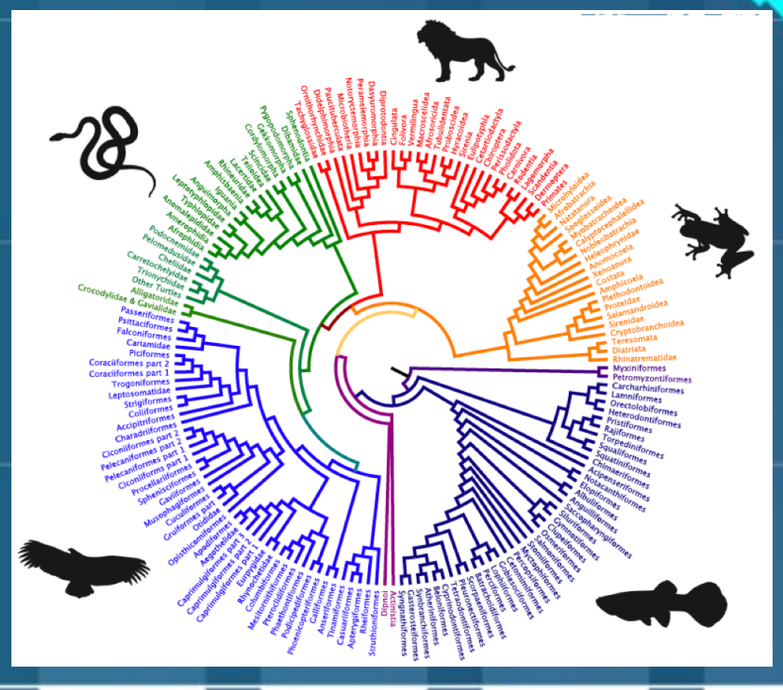
$10K
$1K
$100
$10
$1
$0.1

Moore's Law

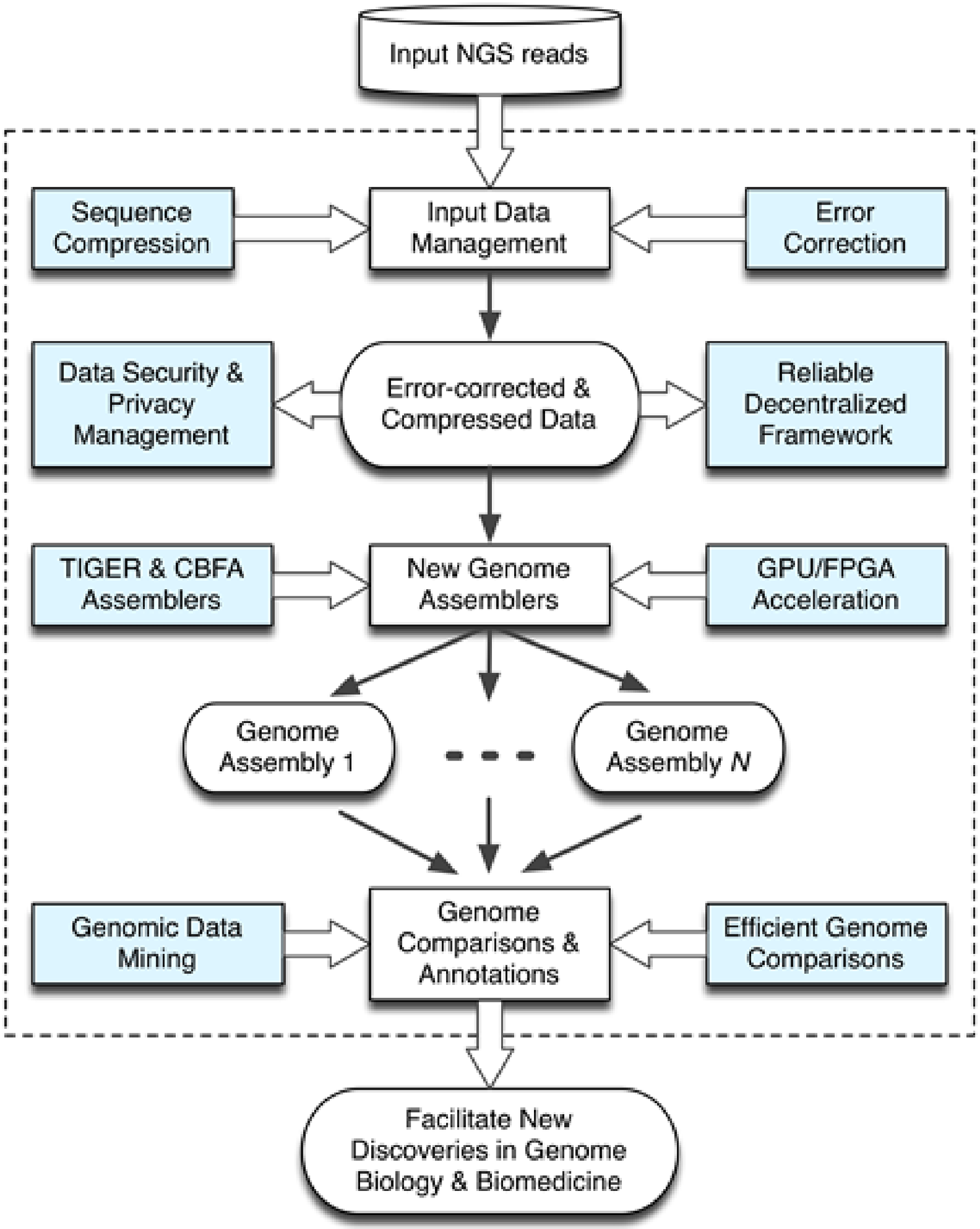
GENOME 10K©

5k

genome.gov/sequencingcosts



2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012



CompGen NGS Sequence Data Workflow

Scalability for all critical steps of the workflow

Error Correction

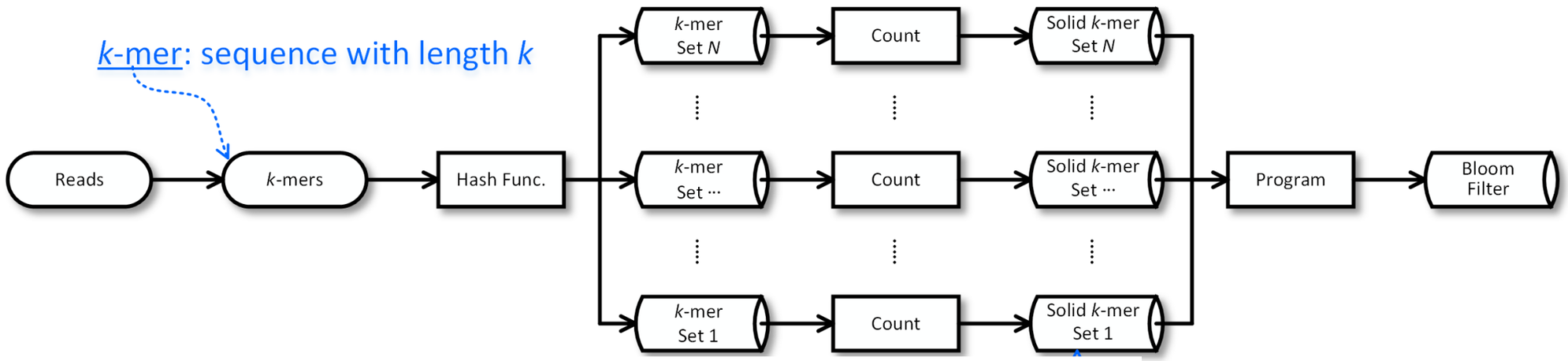
Challenges in Error Correction for Genome Data

- **Large memory usage**
- **Hard-to-handle repeats in genomes**

Our Approach in a Nutshell

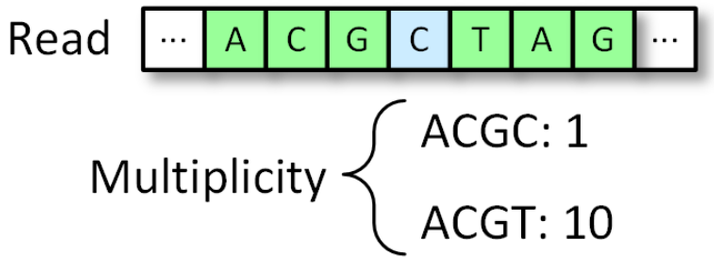
- **Use a Bloom filter for memory reduction**

  All the solid *k*-mers are programmed into a Bloom filter



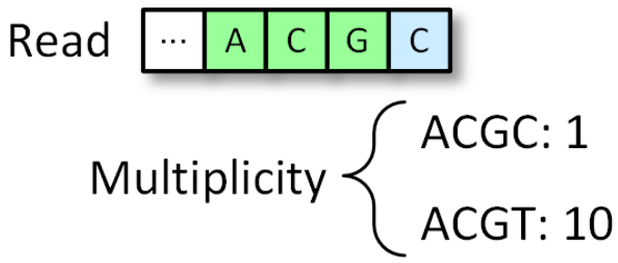
• Determine whether a k-mer is solid without saving k-mers and their multiplicities
• Increase *k*-mer length without increasing memory usage for better repeats handling
• Use adjacent *k*-mers and quality scores to make better corrections in repeats

Correcting Errors at the Ends of Reads



Is changing ACGC to ACGT always right? → No, ACGT may be
1) an erroneous pattern with a high multiplicity.
2) an false positive from a Bloom filter.

Solution:
Check the multiplicity of CGCT, GCTA, and CTAG for lowering the probability of false correction.



What if an error exists at the end of a read?

Solution:
Extend the read to the right.
Check the multiplicity of CGTA, CGTC, CGTG, and CGTT.
Repeat this a certain number of times.

Experimental Results: Memory Usage

Data Set	Memory Usage (MB)						
	BLESS	DecGPU	ECHO	HiTEC	Musket	Quake	Reptile
S. Aureus	11	1,556	6,063	2,127	362	644	2,184
E. Coli	14	2,171	N/A	14,096	347	8,339	1,008
Human Chr. 14	150	2,223	N/A	N/A	3,763	2,126	12,928
Human Chr. 1	372	2,473	N/A	N/A	7,815	8,863	20,041

Average : 2.7%

Worst case: 14.9%

Experimental Results: Accuracy

$$Gain = \frac{(\#\ of\ corrected\ errors)\ -(\#\ of\ newly\ introduced\ errors)}{(\#\ of\ errors\ in\ the\ original\ reads)}$$

Data Set	Gain						
	BLESS	DecGPU	ECHO	HiTEC	Musket	Quake	Reptile
S. Aureus	0.894	0.002	0.707	0.838	0.703	0.144	0.175
E. Coli	0.967	-0.028	N/A	0.880	0.926	0.837	0.724
Human Chr. 14	0.644	-0.058	N/A	N/A	0.537	0.126	0.379
Human Chr. 1	0.870	-0.017	N/A	N/A	0.866	0.539	0.560

Best accuracy for all the inputs

Experimental Results: Alignment

$$Aligned\ Ratio = \frac{\#\ of\ reads\ that\ can\ be\ exatly\ aligned\ to\ a\ unique\ point\ of\ the\ reference\ sequence}{Total\ \#\ of\ reads}$$

Data Set	Aligned Ratio (%)								
	Uncorr.	Error-Free	BLESS	DecGPU	ECHO	HiTEC	Musket	Quake	Reptile
S. Aureus	19.5	N/A	**75.1**	36.6	53.6	70.1	66.9	58.1	31.6
E. Coli	73.5	N/A	**96.5**	90.7	N/A	95.0	95.3	94.3	90.7
Human Chr. 14	42.8	N/A	**74.1**	55.7	N/A	N/A	69.6	72.0	62.4
Human Chr. 1	36.4	80.3	**77.2**	63.8	N/A	N/A	74.3	65.9	61.3

Largest improvement in alignment

Experimental Results: Assembly

- Input: 10Mbp region of human chr.1
- Assembler: Velvet 1.2.09

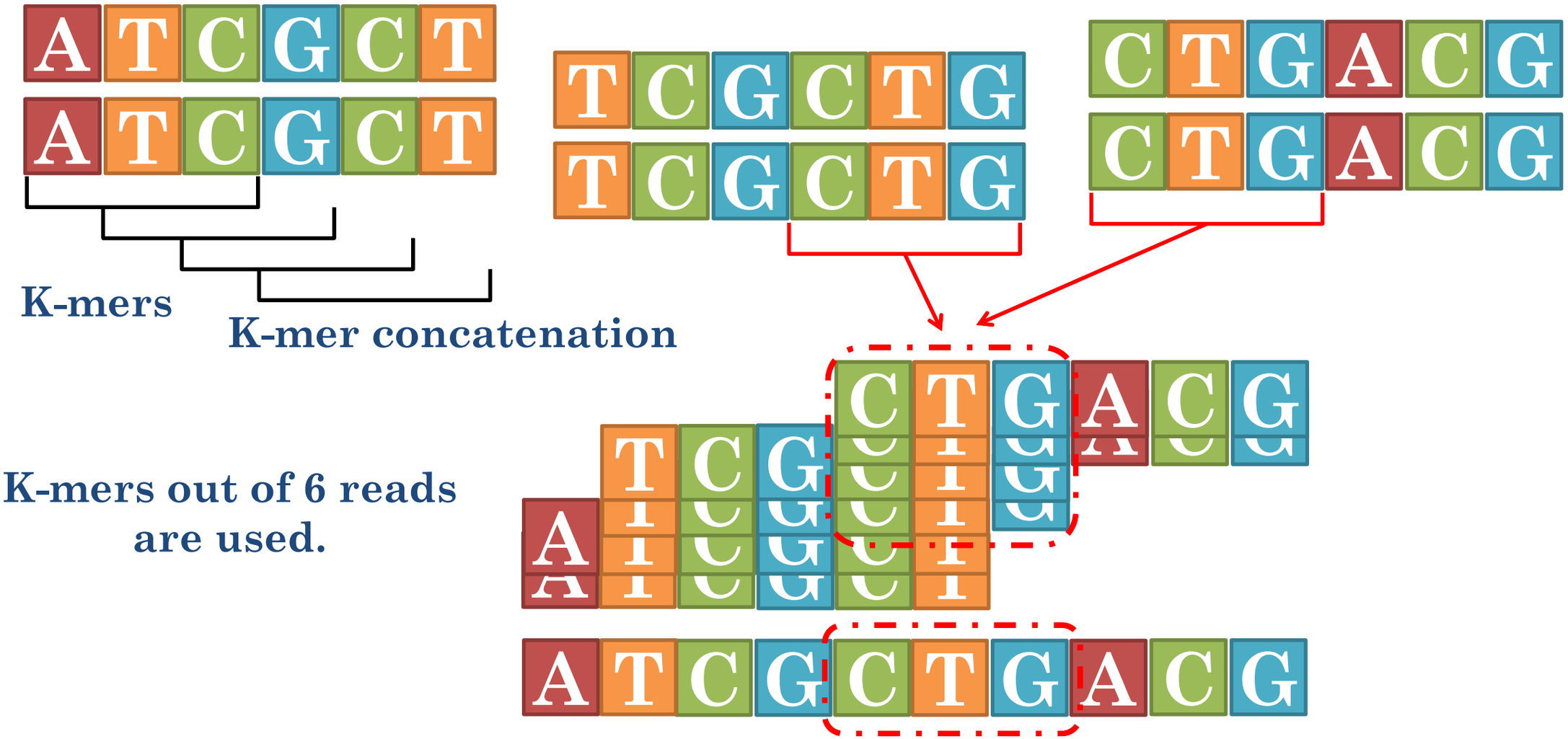
Metric	Uncorr.	Error-Free	BLESS	DecGPU	ECHO	HiTEC	Musket	Quake	Reptile
Corrected NG50	671	1,239.1	**1,004.1**	751.6	665.4	805.2	1,004.1	850.4	750.9
# of Errors	1,321	550	**449**	568	835	813	479	555	554
Genome Coverage	99.5	99.8	**99.8**	99.8	99.8	99.7	99.8	99.8	99.8

Largest improvement in assembly

DE NOVO ASSEMBLY

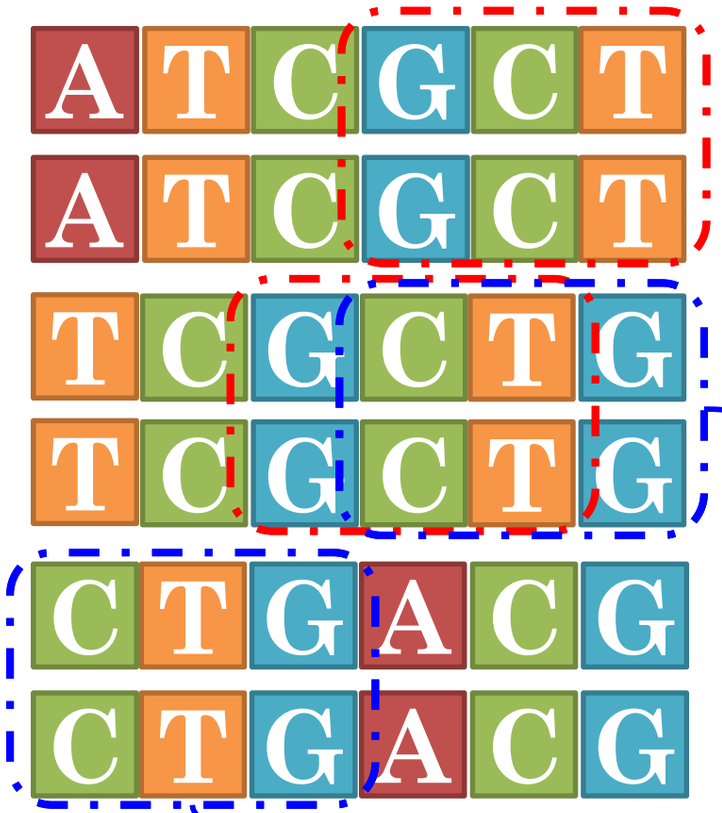
De Novo Assembly Based on K-mers

Reads with 2x coverage

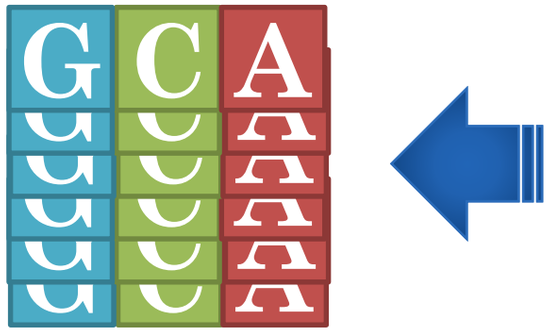


Too Much Information can Hurt

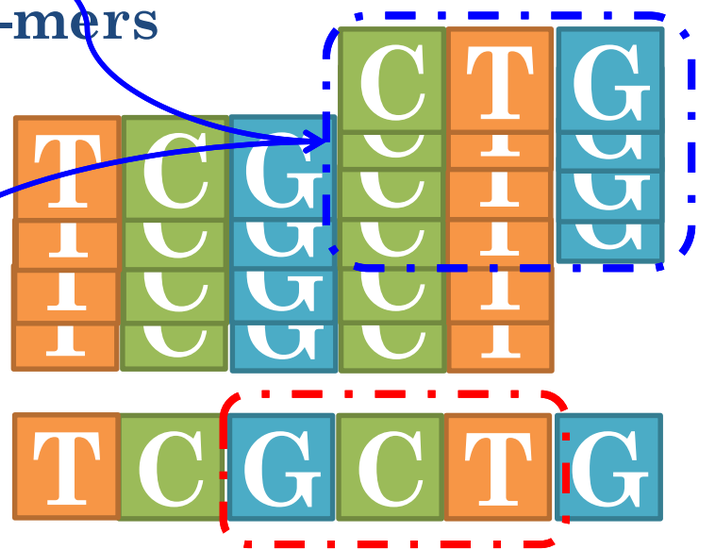
From Chromosome 1



If consider another 6 reads with similar k-mers.



6 k-mers override 4 k-mers

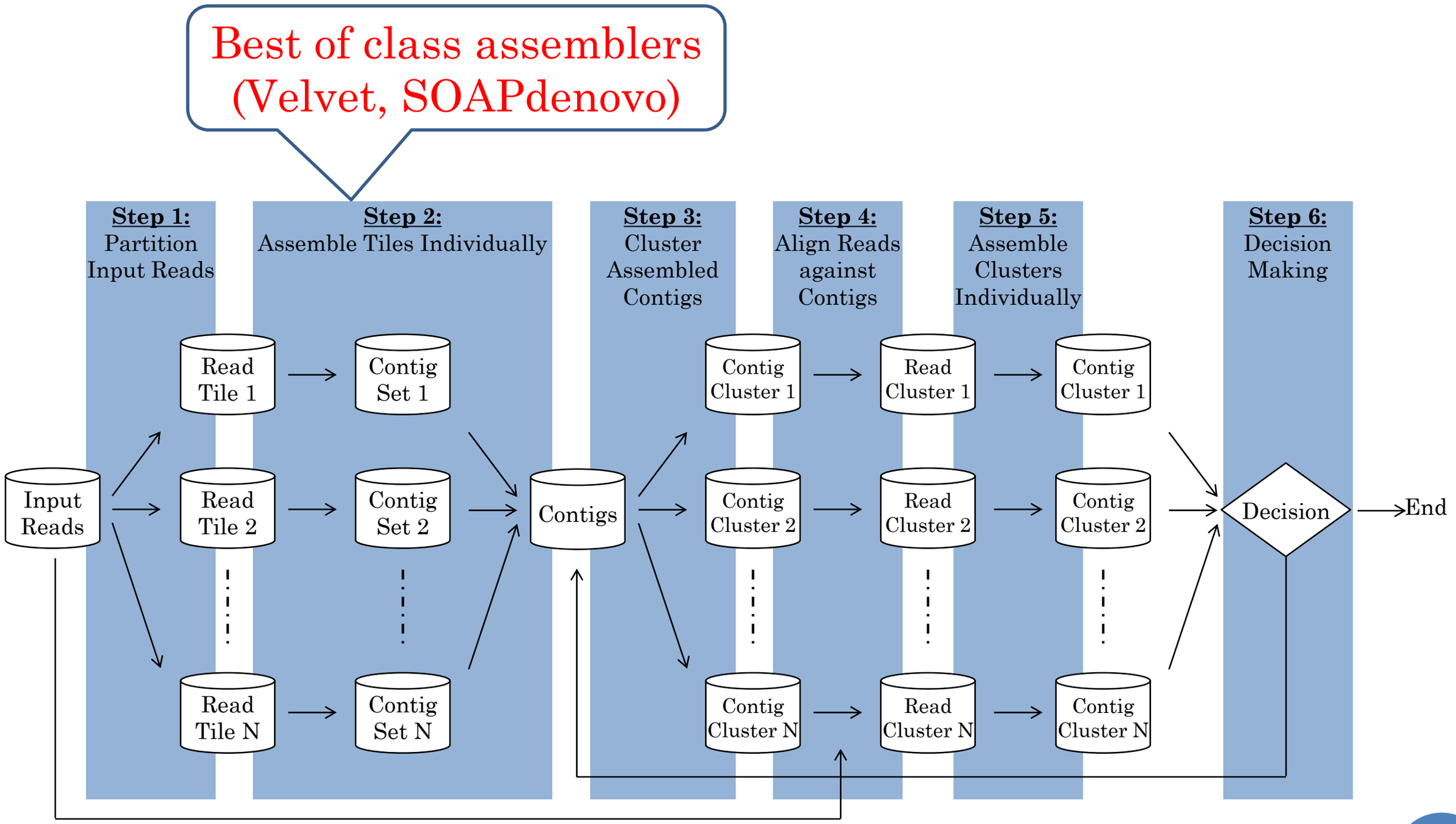


Won't exist because of the 6 k-mers

From Chromosome 2

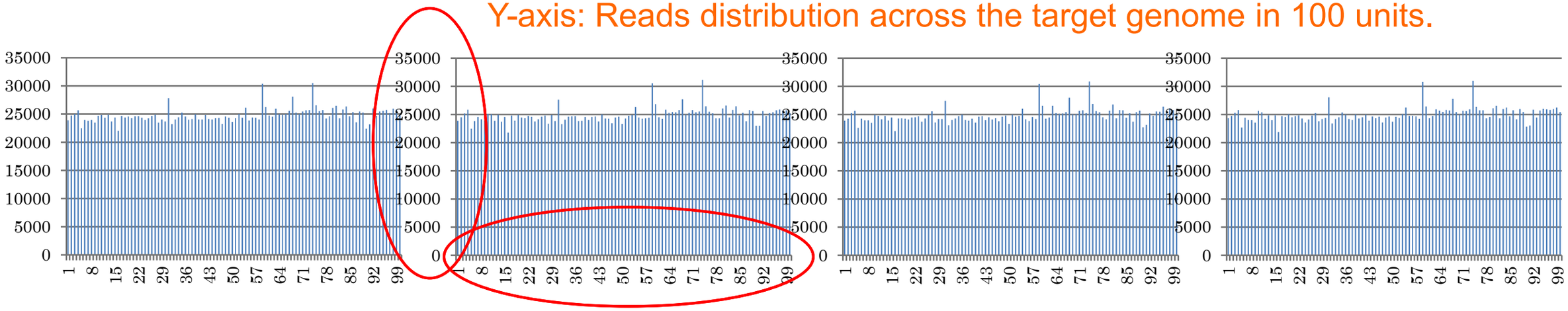
Ambiguous k-mers

Schematic View of Tiger



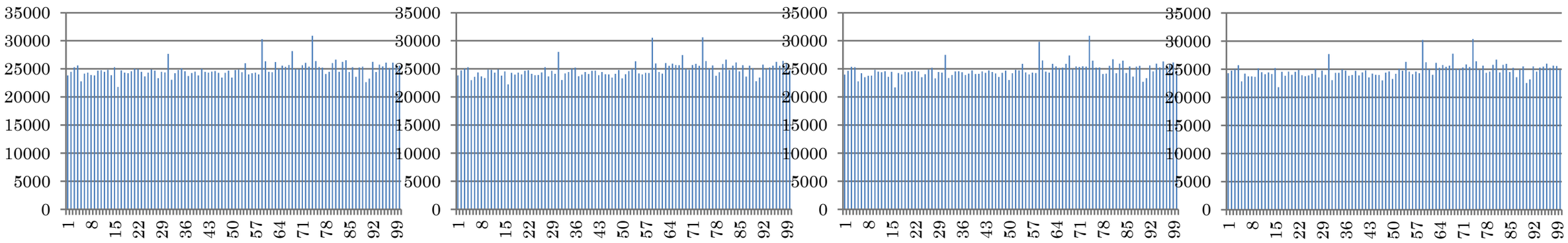
E. Coli RANDOM Read Distribution:

X-axis: Number of reads
Y-axis: Reads distribution across the target genome in 100 units.



Read set 1 Read set 2 Read set 3 Read set 4

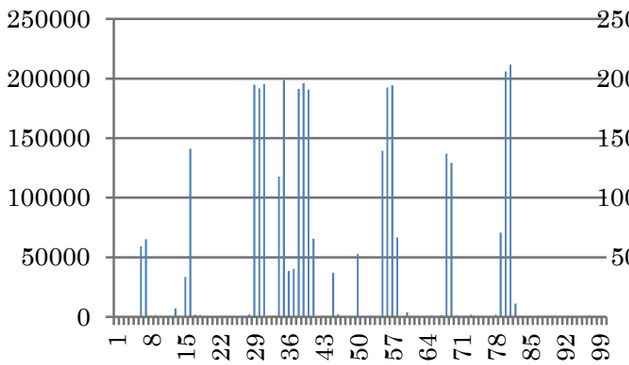
All reads are well distributed across the target genome.



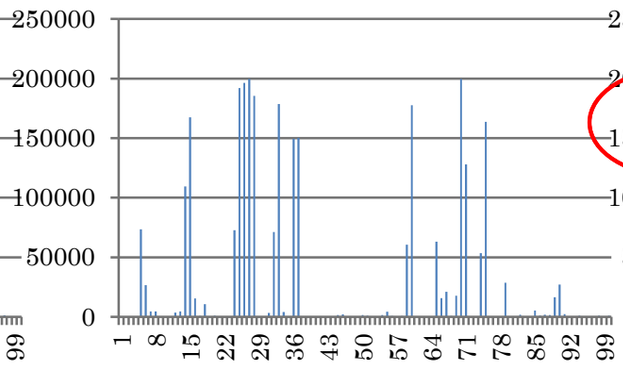
Read set 5 Read set 6 Read set 7 Read set 8

E. Coli Clustered Read Distribution:

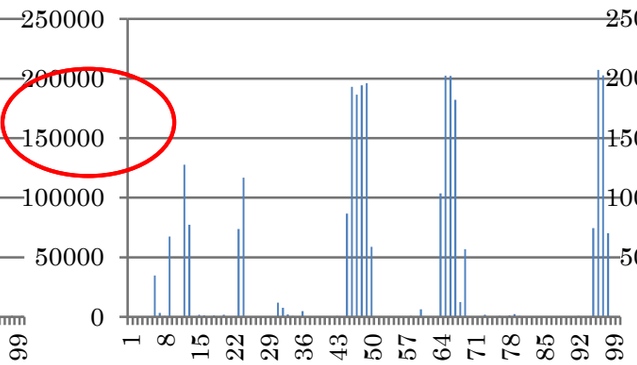
Y-axis increases by 10x.



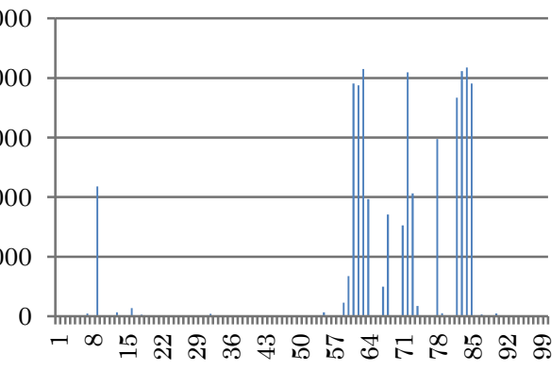
Read set 1



Read set 2

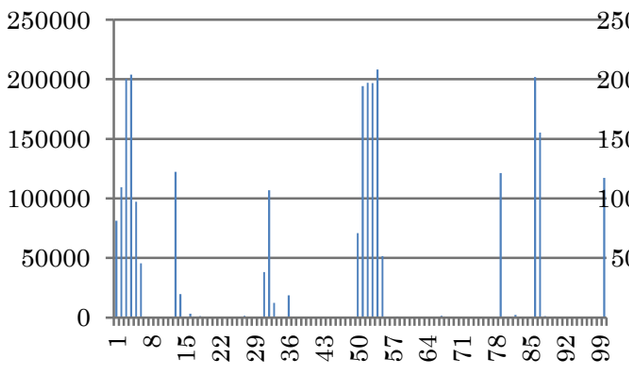


Read set 3

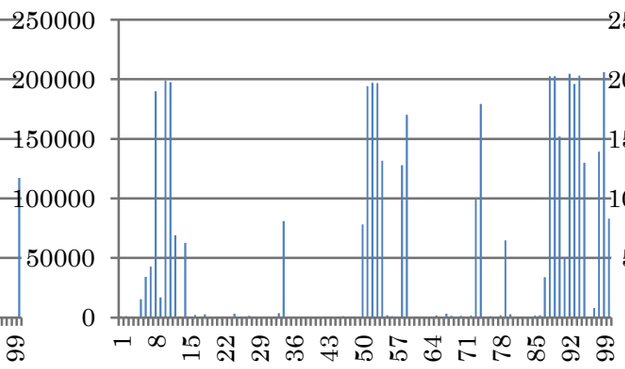


Read set 4

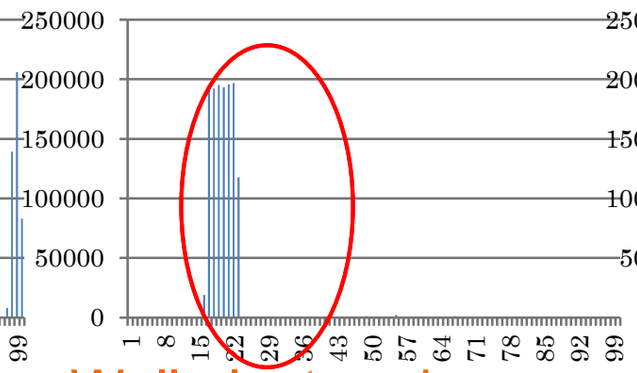
Reads are clustered across the target genome.



Read set 5

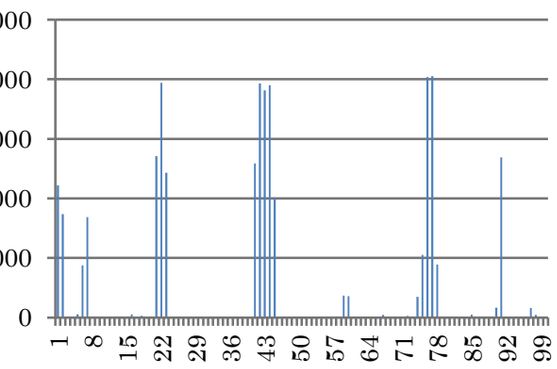


Read set 6



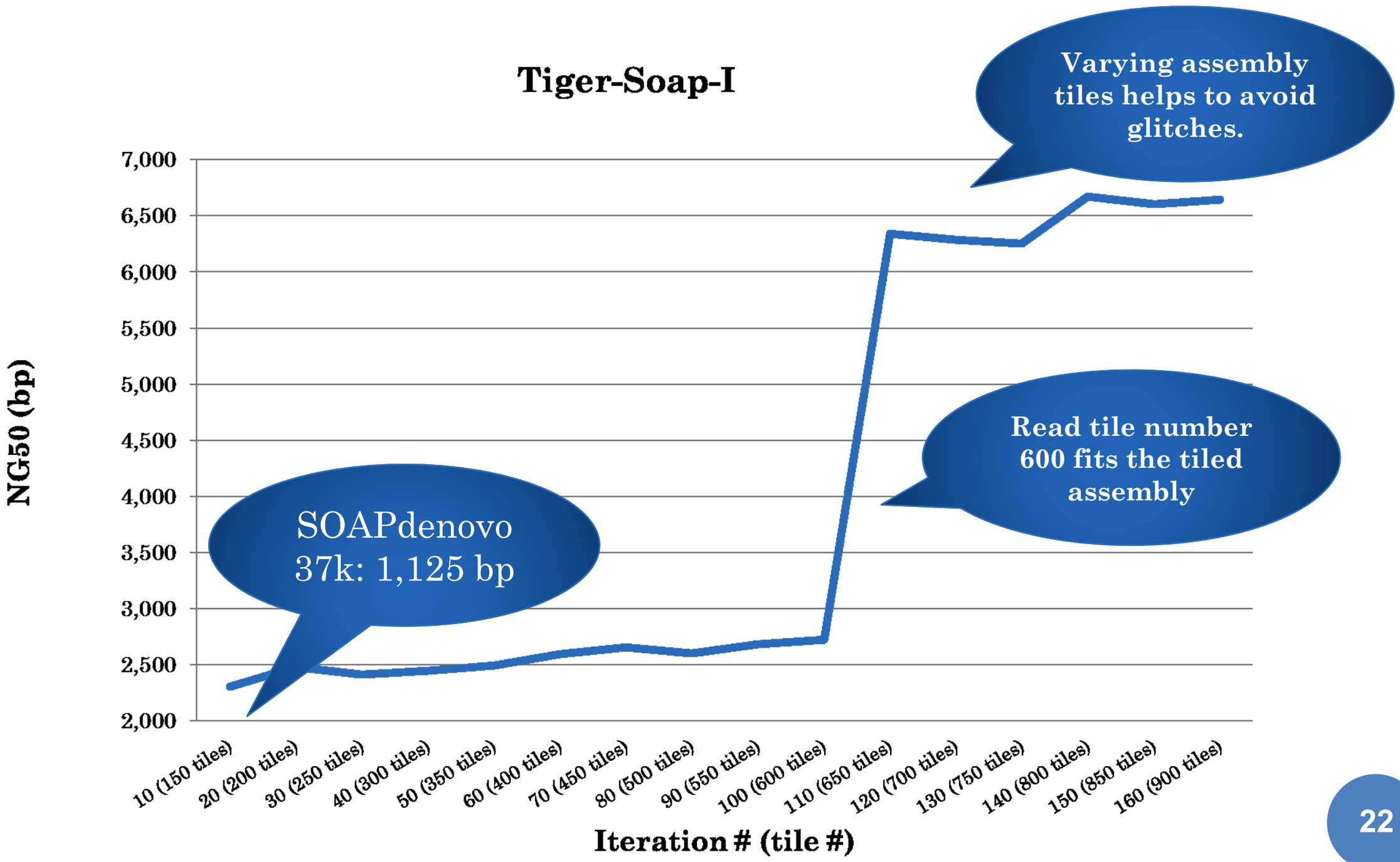
Well-clustered

Read set 7

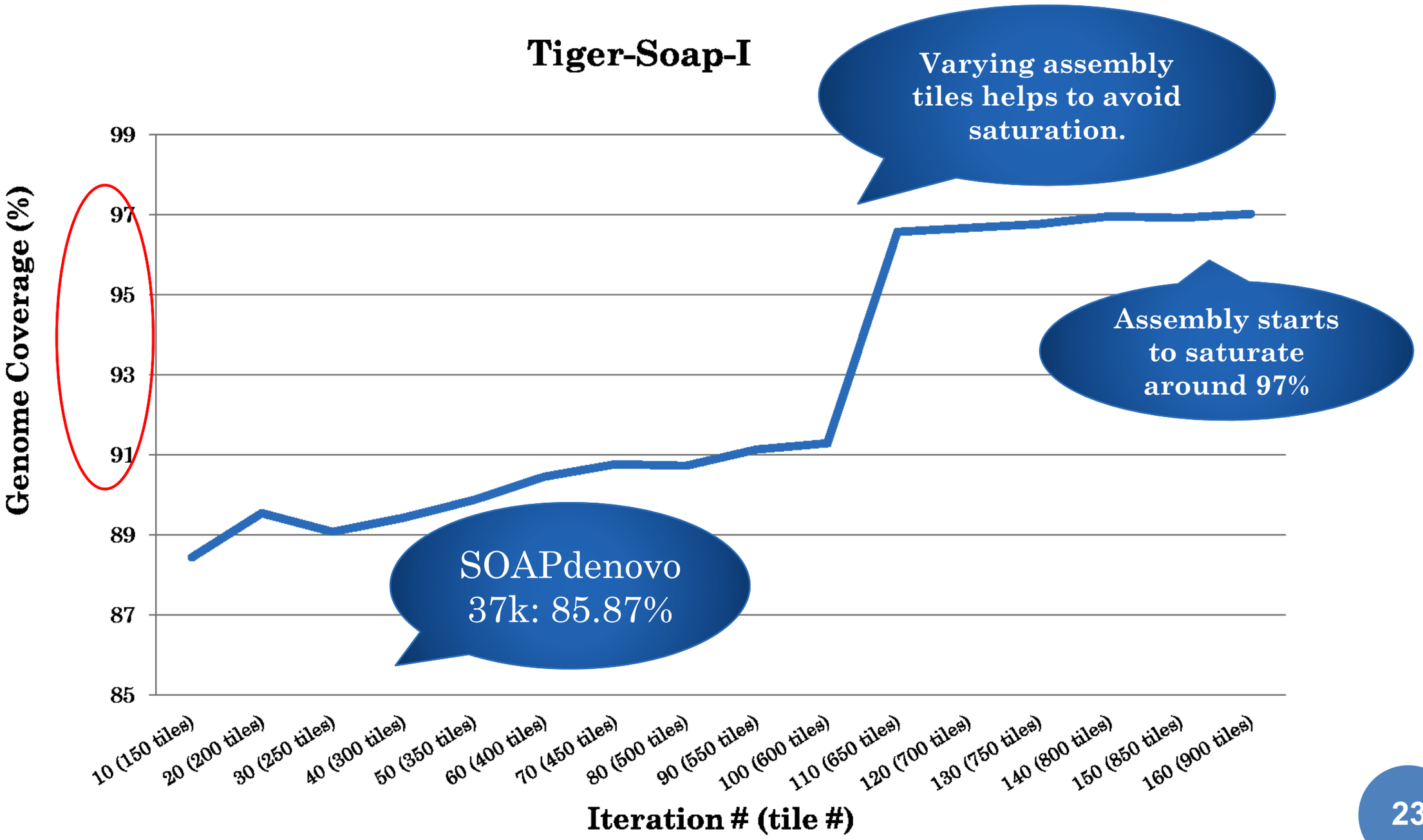


Read set 8

Assembly Results: Chromosome 22



Assembly Results: Chromosome 22



Achievements: Better Quality Results

- Results compared to two widely used assemblers, SOAPdenovo and Velvet.

	Chr. 14	Chr. 22	Orchid Bee
NG50/N50	1.3x ~ 2.2x	1.3x ~ 6.0x	7x/57x
Coverage Improvement	98.02% (+0.1%)	97.02% (+11.2%)	N/A
Errors Improvement	533 bp (-68 bp)	229 bp (+224 bp)	N/A

Longer

Higher

Similar

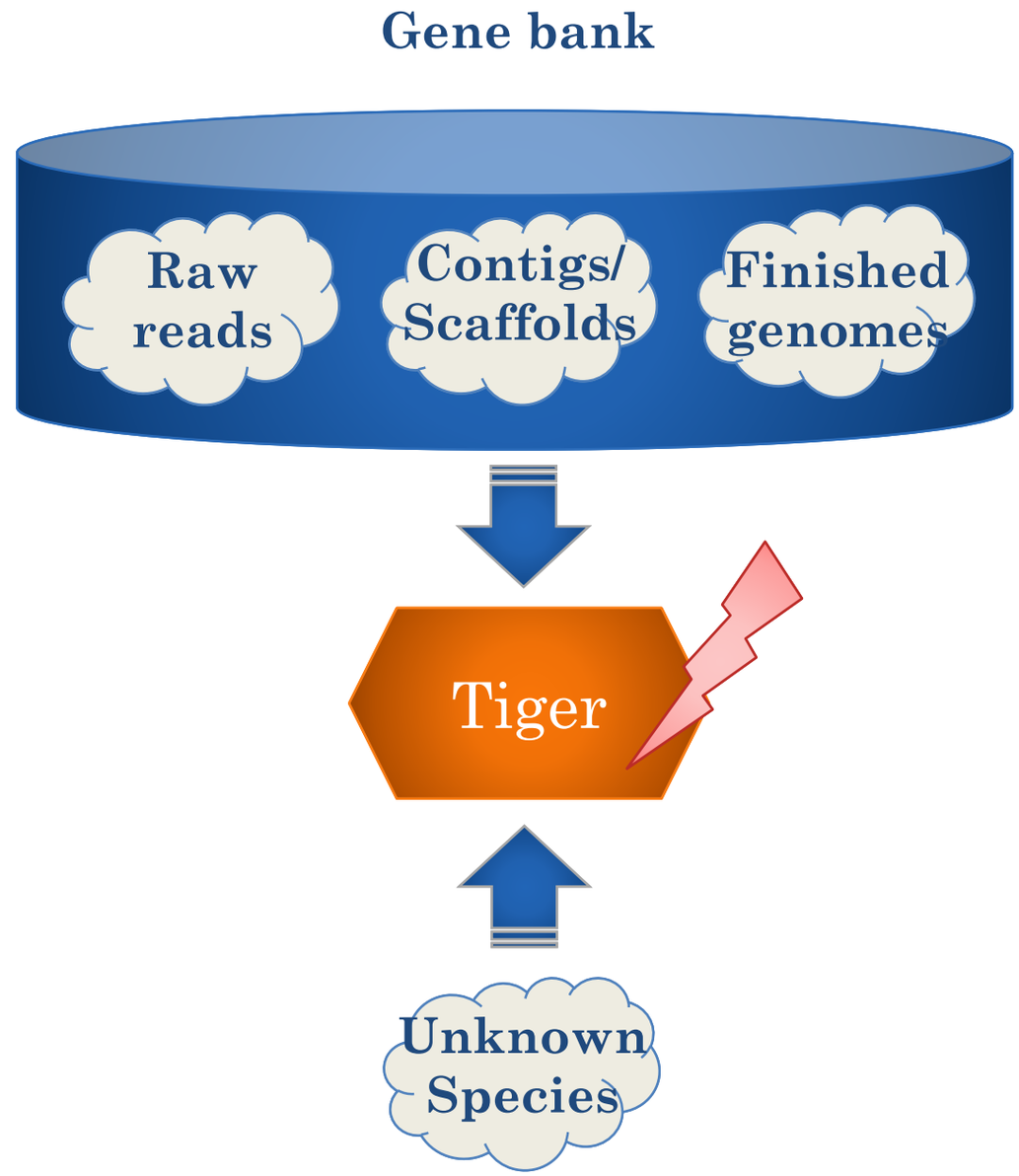
Software downloadable with a research license signed first:
http://impact.crhc.illinois.edu/Tiger/tiger.aspx

Xiao-Long Wu, et al. "TIGER: Tiled iterative genome assembler," BMC Bioinformatics, 2012.

ROBUST ALIGNMENT

Overview of Tiger Aligner

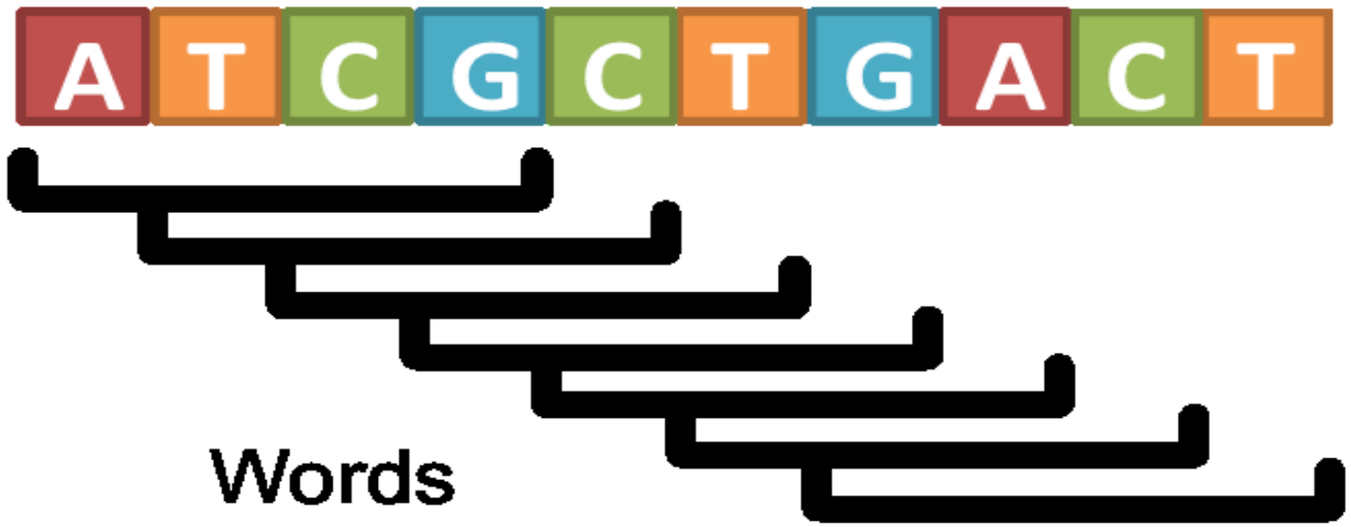
- Approximate multi-genome sequence aligner
  - Genomes can be in various forms in terms of:
    - Evolutionary relationships
    - Sequence completeness
  - Two approaches: Bowtie-based & Word-based
    - Low-memory, fast alignment
    - Better mismatch tolerance, more information

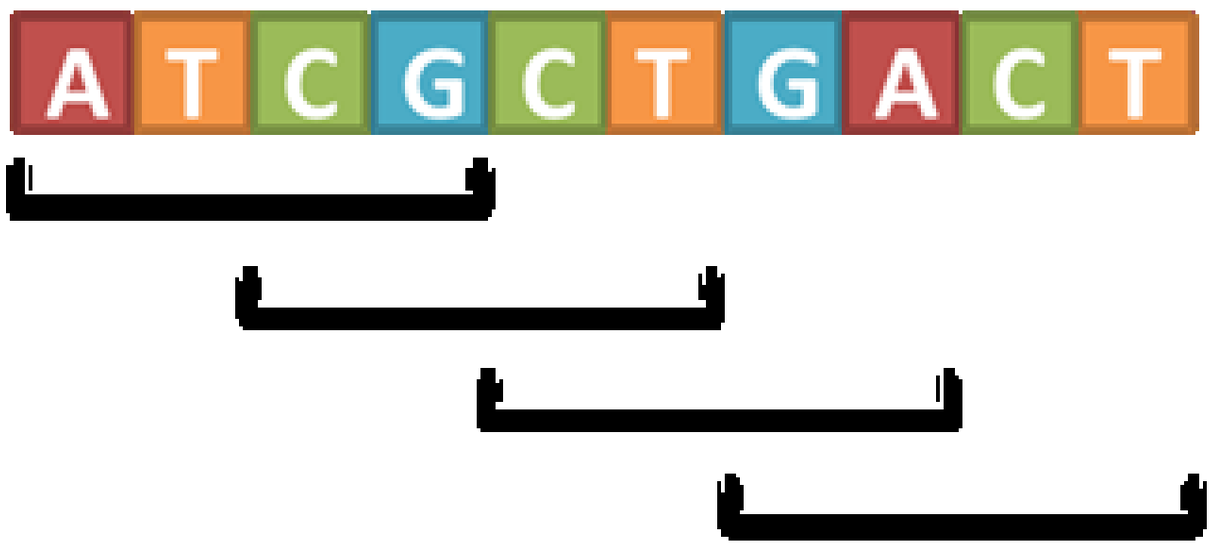


- **Software downloadable with a research license signed first:**
  http://impact.crhc.illinois.edu/Tiger/tiger.aspx

Xiao-Long Wu, et al. "TIGER: An Approximate Multi-Genome Sequence Aligner," to be submitted, 2013.

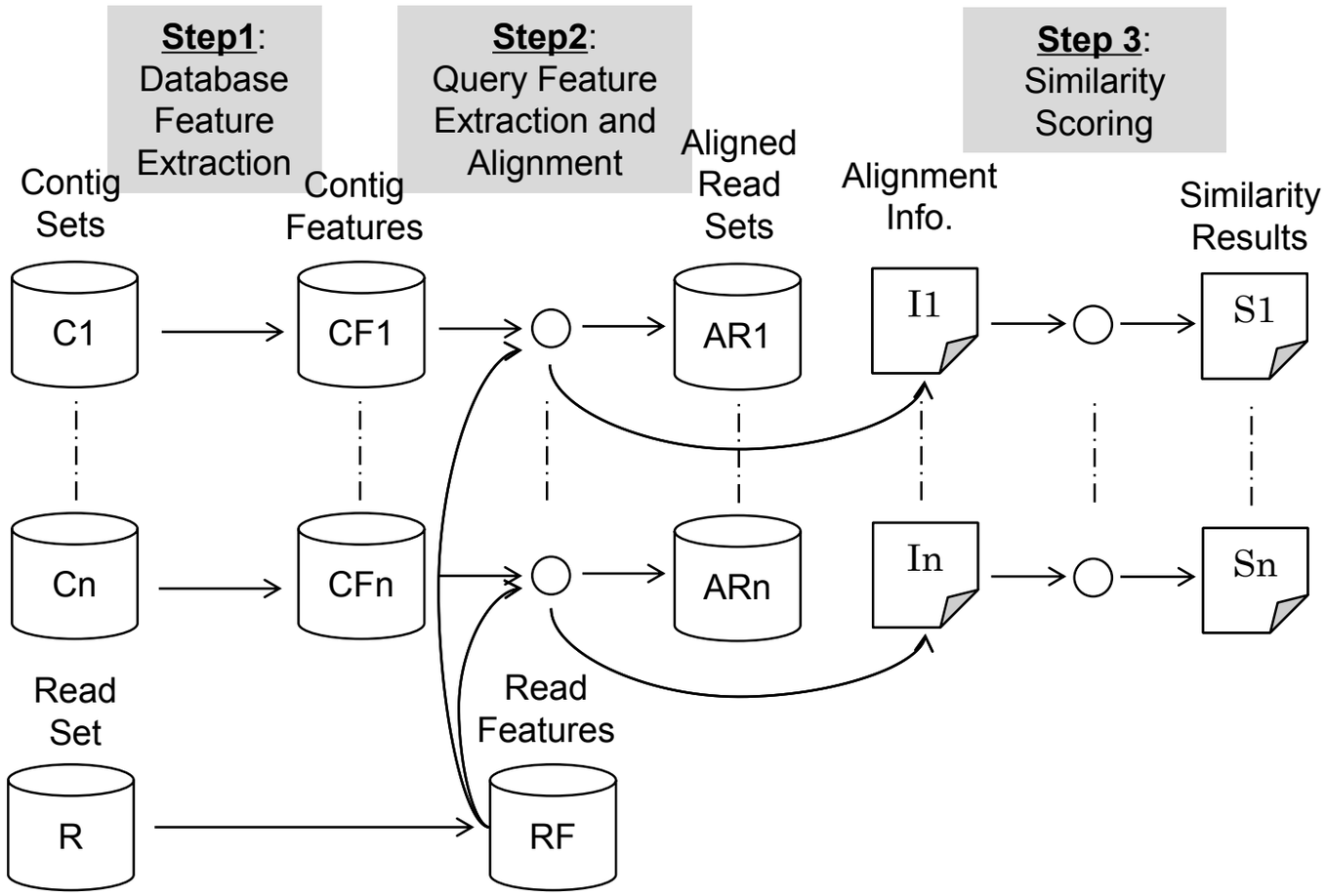
Word-Based Feature Extraction





Extracted words with stride 2

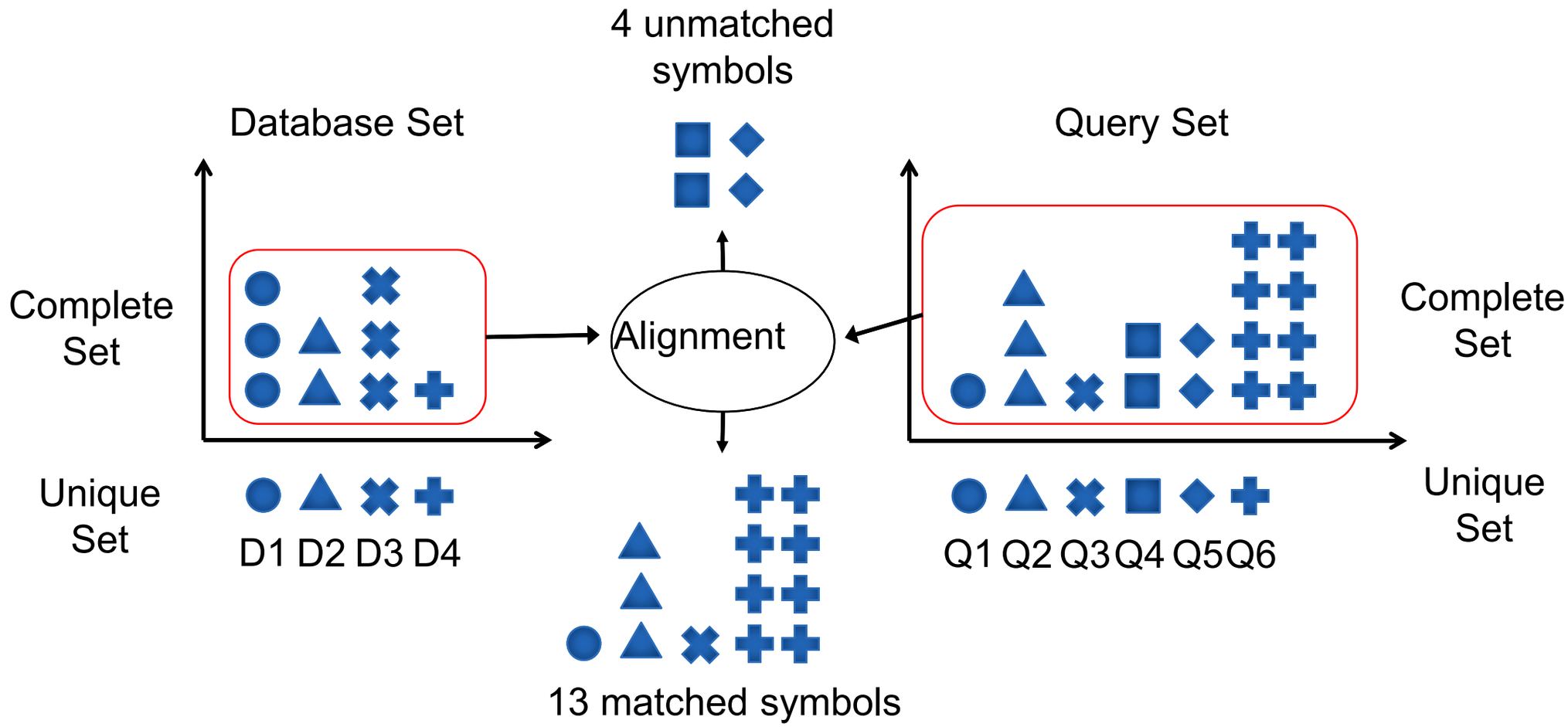
Word-Based Approach



- Tree-based
- Good mismatch tolerance
- More information

- Rough alignment
- Higher memory requirement
- Slower than BWT-based methods

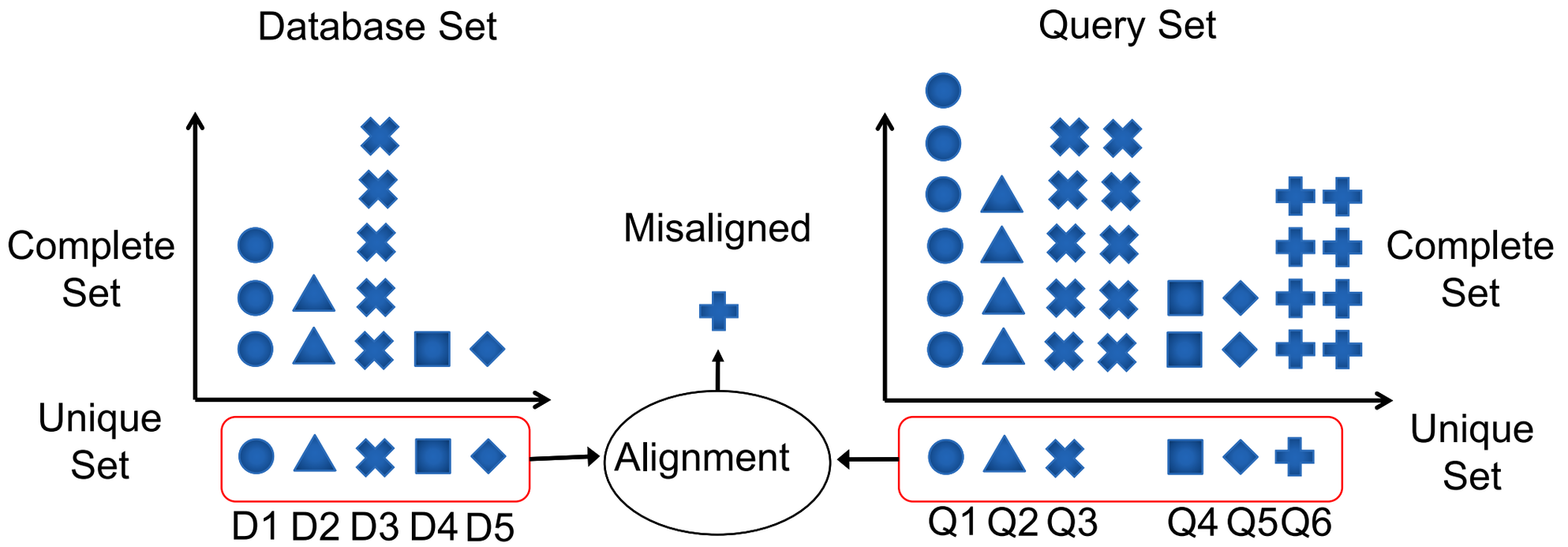
Similarity Evaluation Example: Coverage Checking (CC)

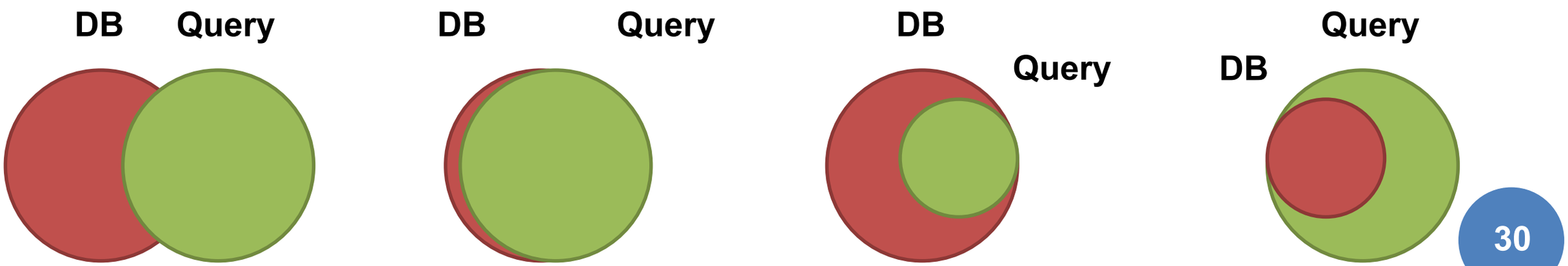


$$Similarity\ Scores_{CC} = (S + P \times (1+\alpha)) / (T \times (1+\alpha))$$

, where S is the number of aligned single reads, P is the number of aligned paired reads, T is the total number of (paired) reads, and α is the weight to add for paired alignment.

Similarity Evaluation Example: Existence Checking (EC)



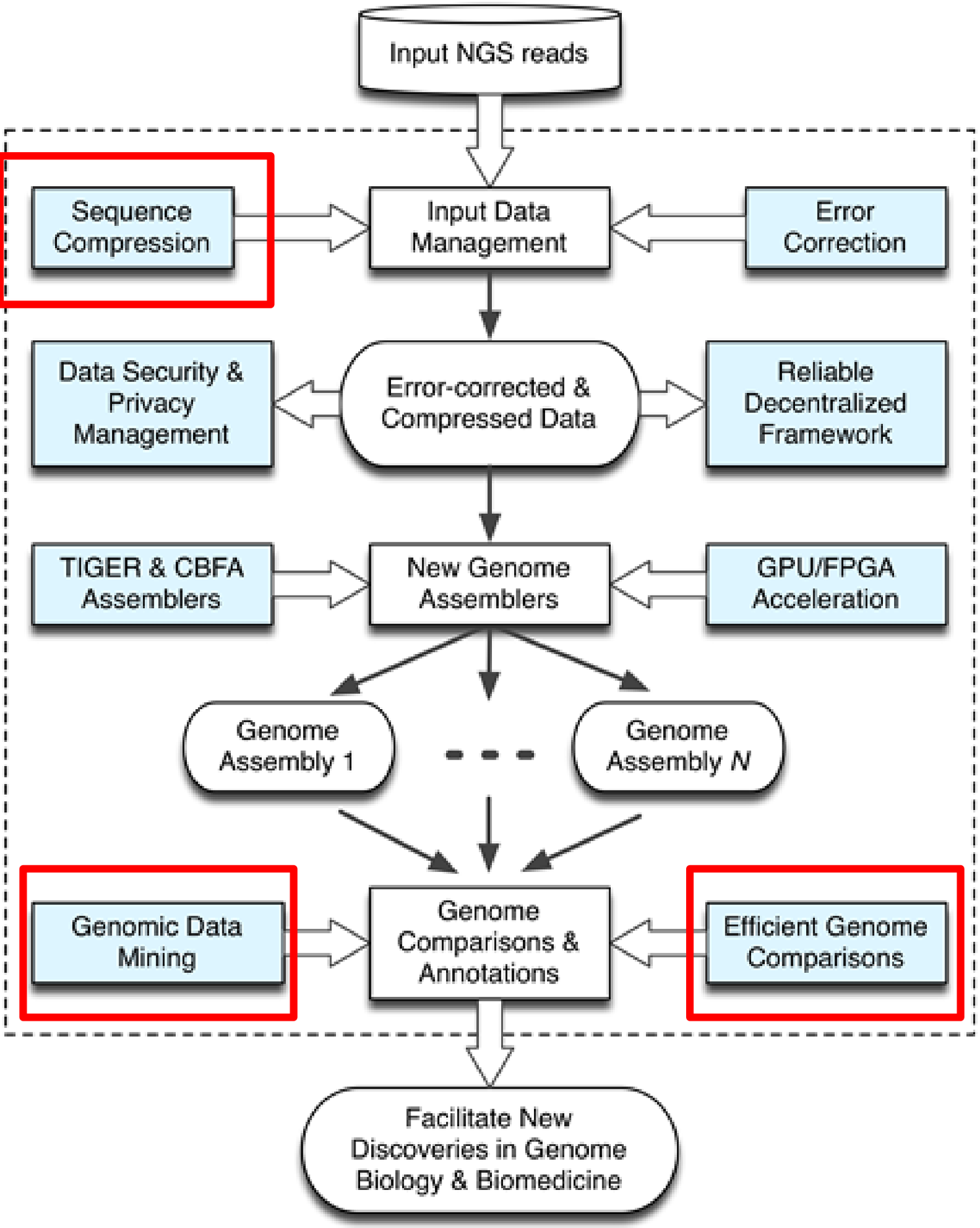


Simulated E.coli Reads w/ Errors as DB and Query Sets

		DB	Query			
	Read Set ID	1	2	3	4	5
	Error rate (%)	0	5	20	40	80
CC	Similarity scores	1.000	0.974	0.867	0.687	0.354
CC	% of aligned reads regardless single/paired	100.00	93.10	73.71	51.99	23.45
CC	% of aligned single reads	0.00	12.86	38.79	49.97	35.88
CC	% of aligned paired reads	100.00	86.67	54.32	27.00	5.51
EC	# of unique words in db.	4,530,123	Same	Same	Same	Same
EC	% of unique words aligned	100.00	100.00	100.00	100.00	99.88
WC	Mean word freq.	28.45	26.13	20.19	14.24	6.93
WC	SD word freq.	12.54	11.57	9.22	6.95	3.78
WC	Peak mean word freq.	43.53	40.16	32.07	23.65	12.99
WC	Peak SD word freq.	32.87	30.41	23.01	16.00	7.79

Runtime Comparison on Chr. 14

Assembler	Velvet 61k	SOAPdenovo 55k	Tiger-Velvet-R 69i	Tiger-Soap-R 53i
Tiger aligner (Bowtie-based) CC scheme (4 tx)	0.11 hr	0.10 hr	0.09 hr	0.11 hr
Tiger aligner (Word-based) CC, EC, WC schemes (4 tx)	0.87 hr	1.03 hr	0.88 hr	0.87 hr
MUMmer (1 tx)	5.27 hr	1.42 hr	9.52 hr	10.09 hr
SOAPaligner (4 tx)	0.53 hr	0.44 hr (Seg fault on library 2)	0.34 hr (Seg fault on library 2)	0.32 hr (Seg fault on library 2)



CompGen NGS Sequence Data Workflow

Recap

THANK YOU!