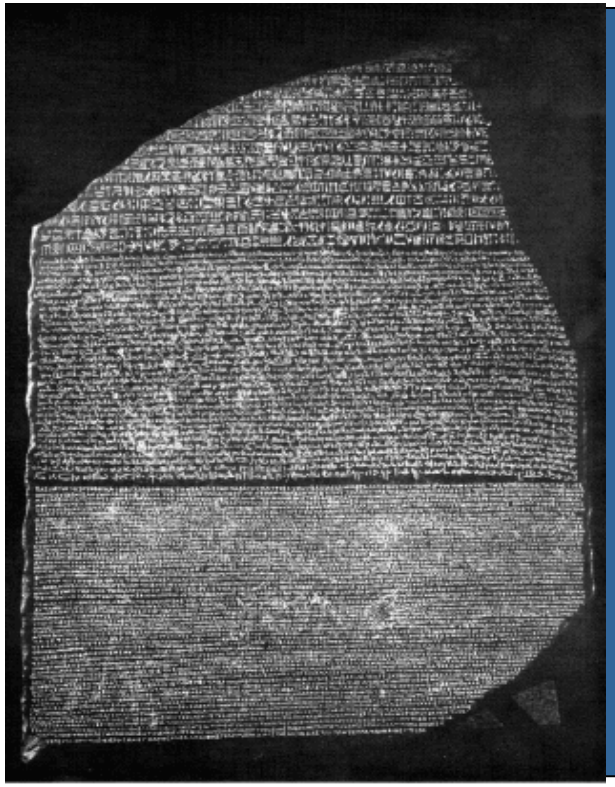


SEQUENCE ANALYSIS

Josep F. Abril, PhD

jabril@ub.edu

Computational Genomics Lab

https://compgen.bio.ub.edu/

ESCI upf. School of International Studies

UNIVERSITAT DE BARCELONA

I→B=U→B

Summary

Where do the sequences come from?

- Sequencing Technologies.
  - Classic sequencing vs NGS approaches.
  - Sequence quality.
  - Sequencing issues.
- Properties of genomic sequences.
  - Indexing sequences & *k*-mer analyses.
  - Sequence complexity.
  - Repetitive sequences.

SEQUENCING OVERVIEW

Getting Sequences

BIOLOGICAL SAMPLE

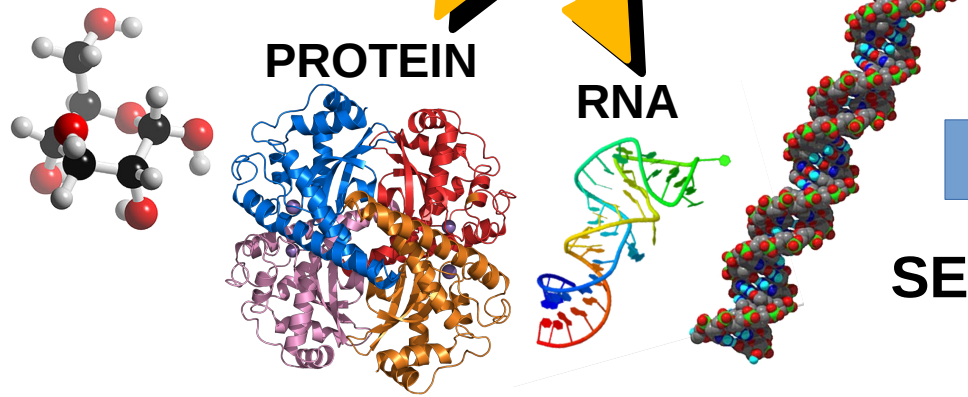
EXTRACTION & PURIFICATION

METABOLITES

PROTEIN

RNA

DNA



SEQUENCING

SANGER (nt) / EDMAN (aa) SEQUENCING

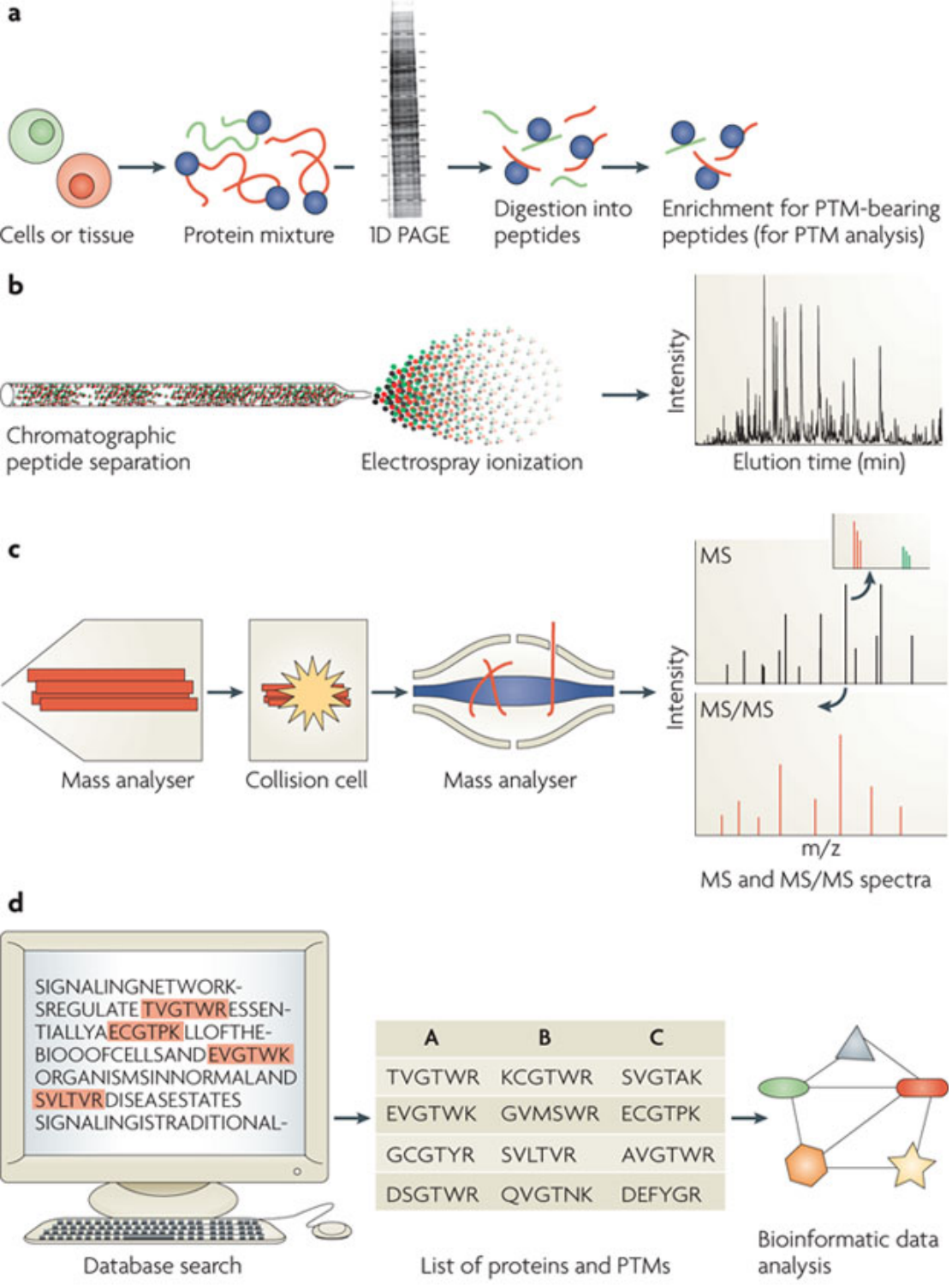
MASS SPECTROMETRY

MICROARRAYS

NEXT GENERATION SEQUENCING (NGS)

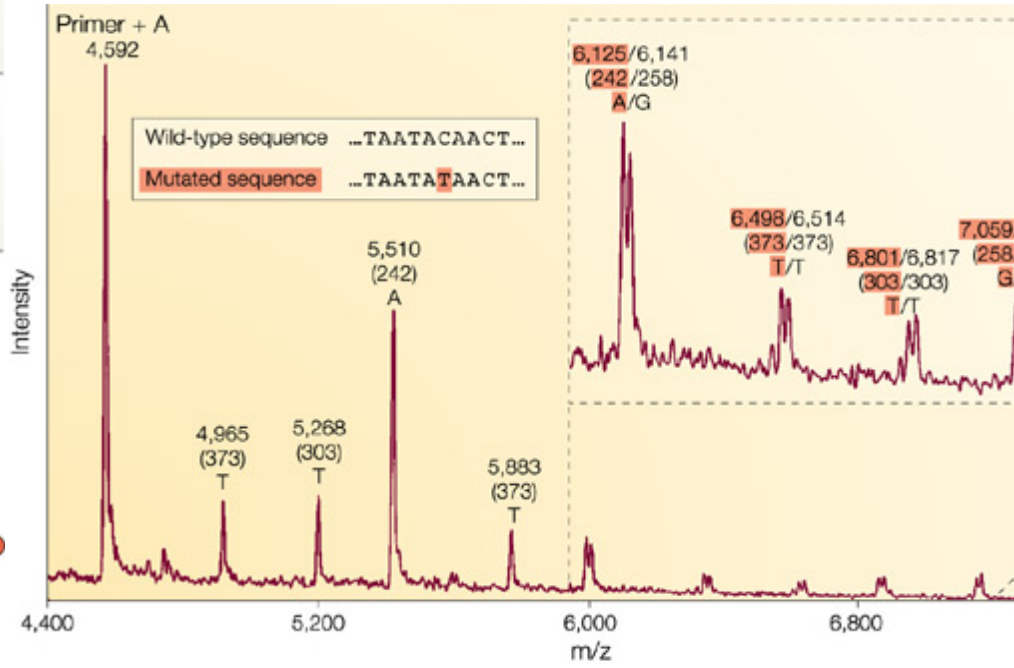
- PYROSEQUENCING (454 Roche)
- BY LIGATION (SOLID)
- BY SYNTHESIS (Illumina HiSeq/MiSeq)
- SINGLE MOLECULE (PacBio SMRT/Oxford NanoPore)

MS-Sequencing



Choudhary & Mann,
Nature Reviews Molecular Cell Biology,
11, 427-439, 2010

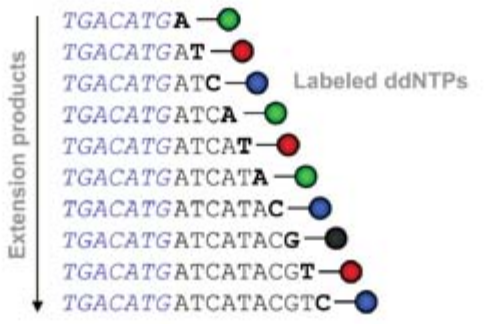
Kim, Ruparel, Gilliam & Ju,
Nature Reviews Genetics,
4, 1001-1008, 2003



DNA Sequencing Technologies

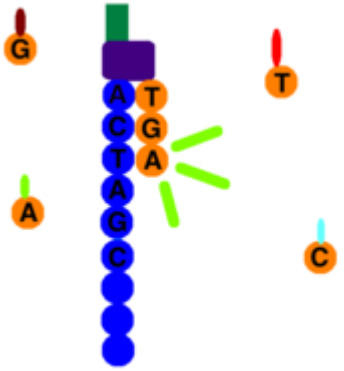
1st Generation

- Tagged Sequences
- Identification by Length
- Long reads / Low Coverage



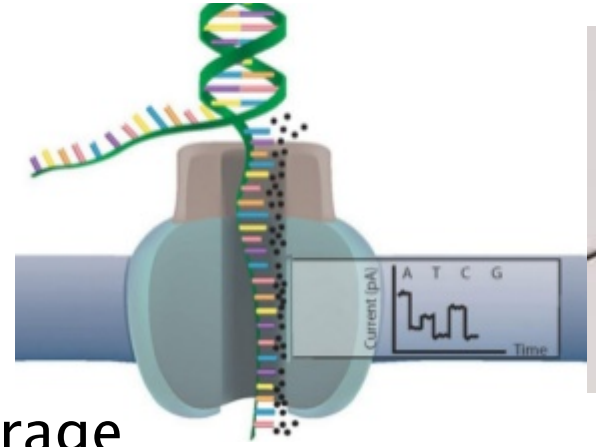
2nd Generation

- Tagged Nucleotides
- Indentification by Synthesis
- Short reads / High Coverage

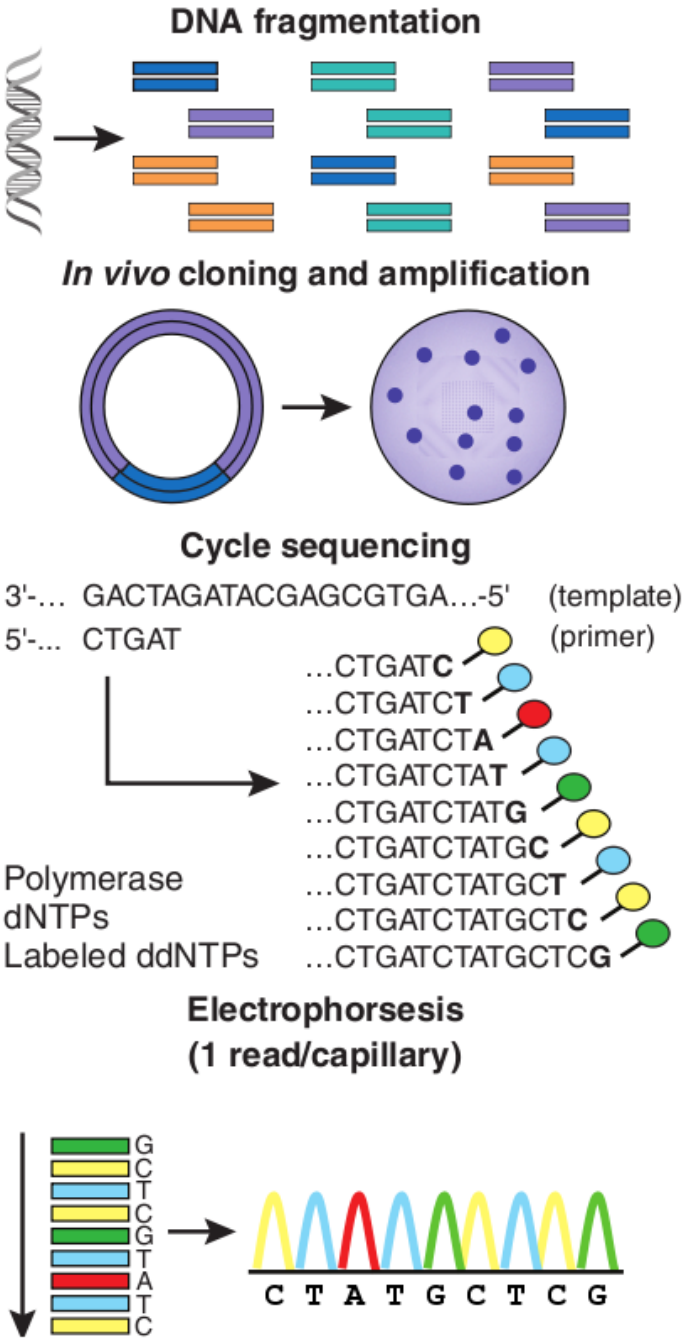


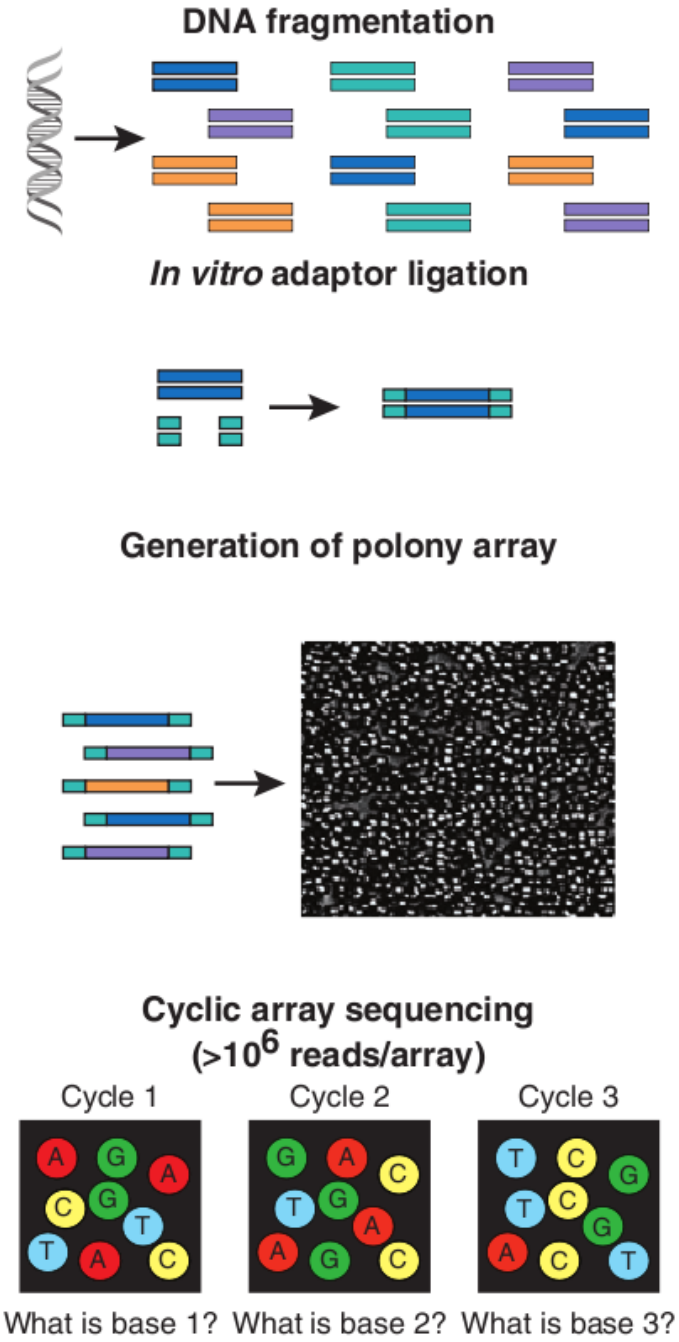
3rd Generation

- Direct Identification
- Very Long reads / Med Coverage



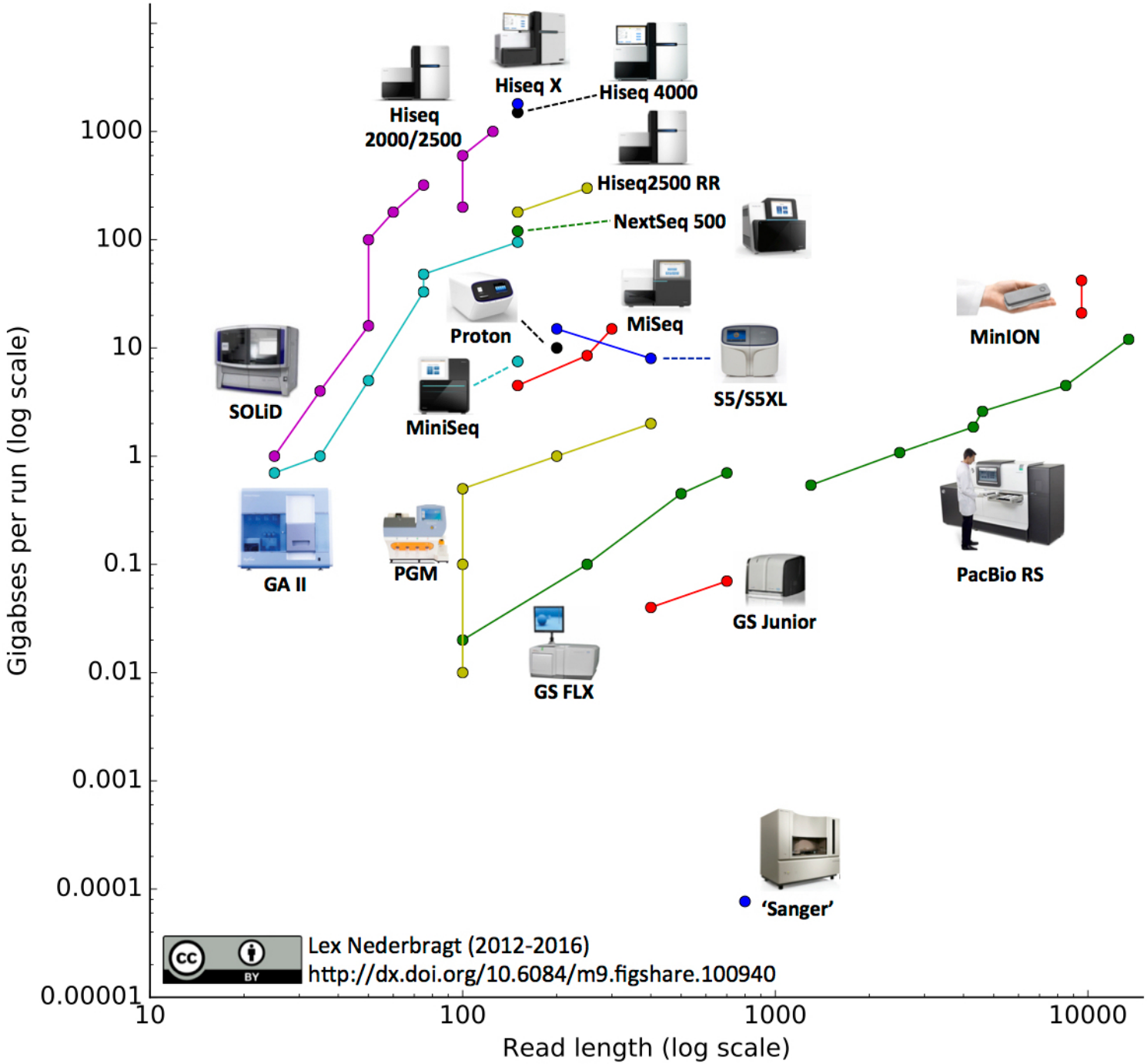
Sequencing Technologies: Sanger vs NGS



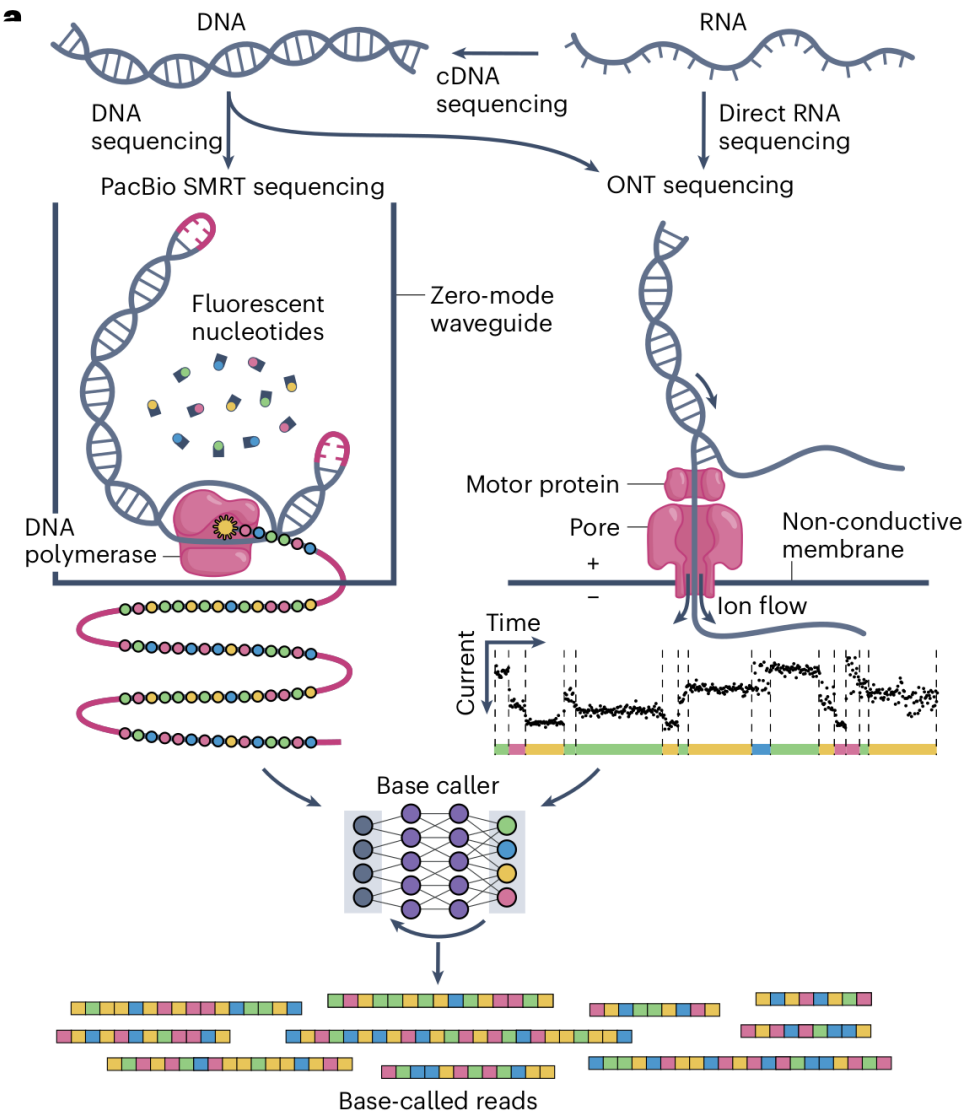


J Shendure & H Ji, *Nature Biotechnology*, 26(10):1135, 2008.

Yield of Sequencing Technologies



Nanopore Single Molecule Sequencing

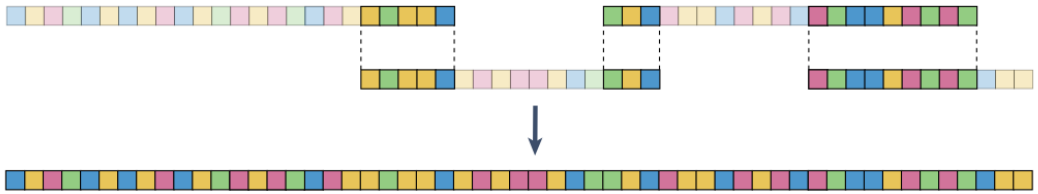


Kovaka et al, *Nat.Methods*, 20(41592): 12-16, 2023

Signal analysis and modification detection

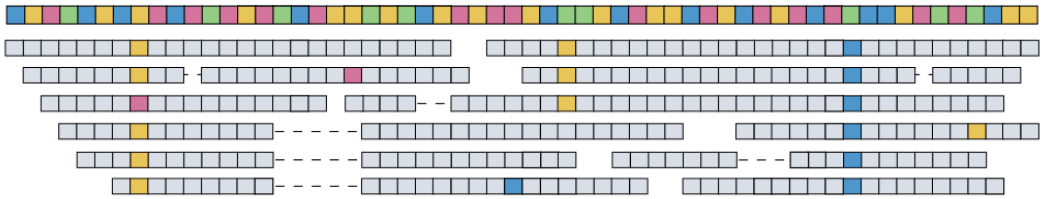
Base calling, epigenomics, epitranscriptomics, labeling, RNA structure

De novo genome assembly



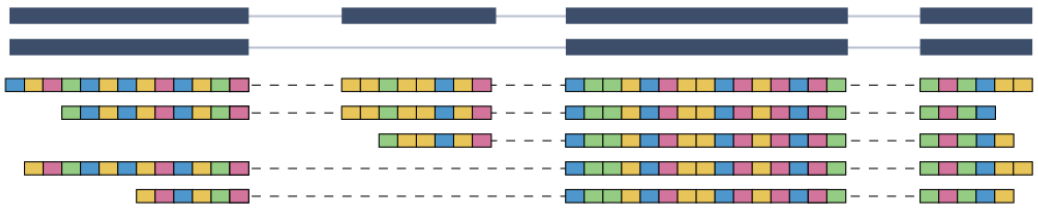
Telomere-to-telomere, haplotype resolved, metagenomic, pan-genomic

Read alignment and variant analysis



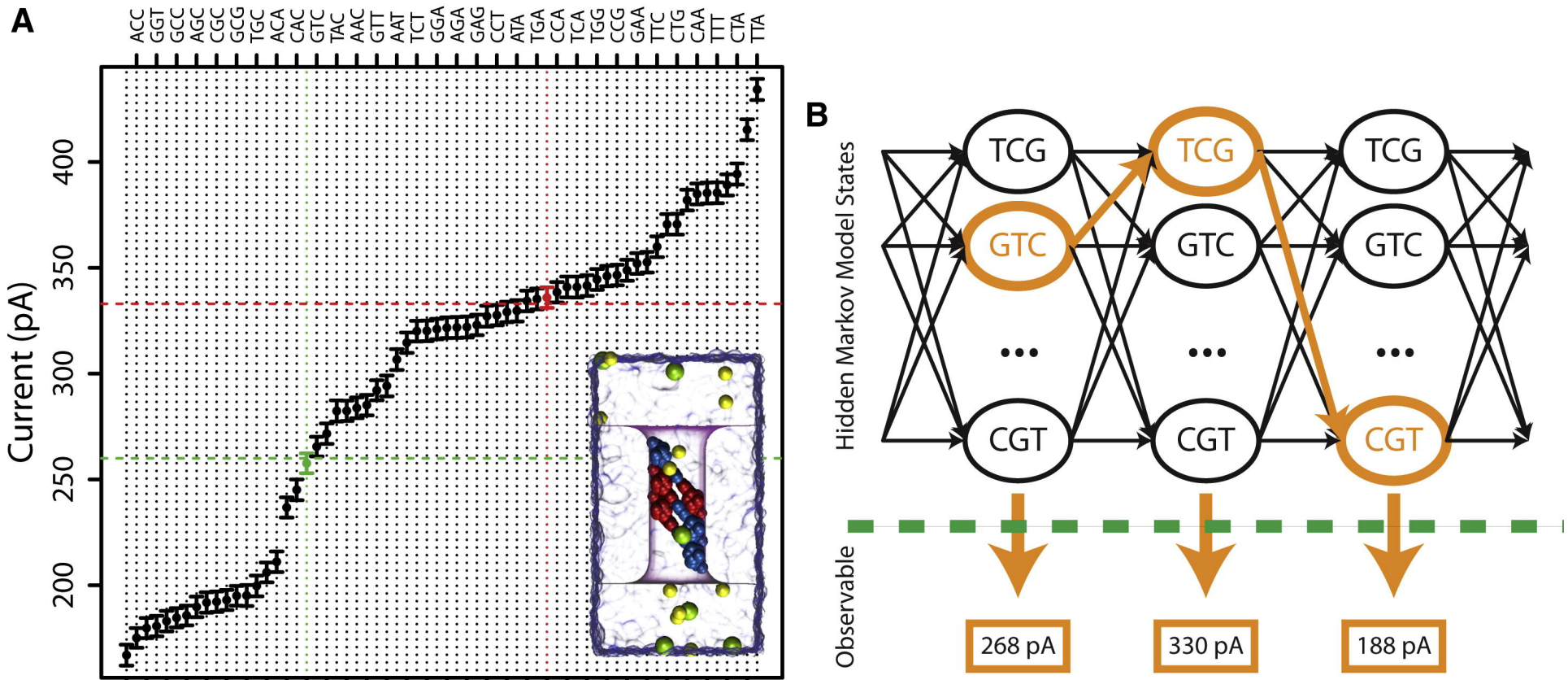
Variant calling (SNP, indel, SV), haplotype phasing, adaptive sampling

Transcriptomics and annotation



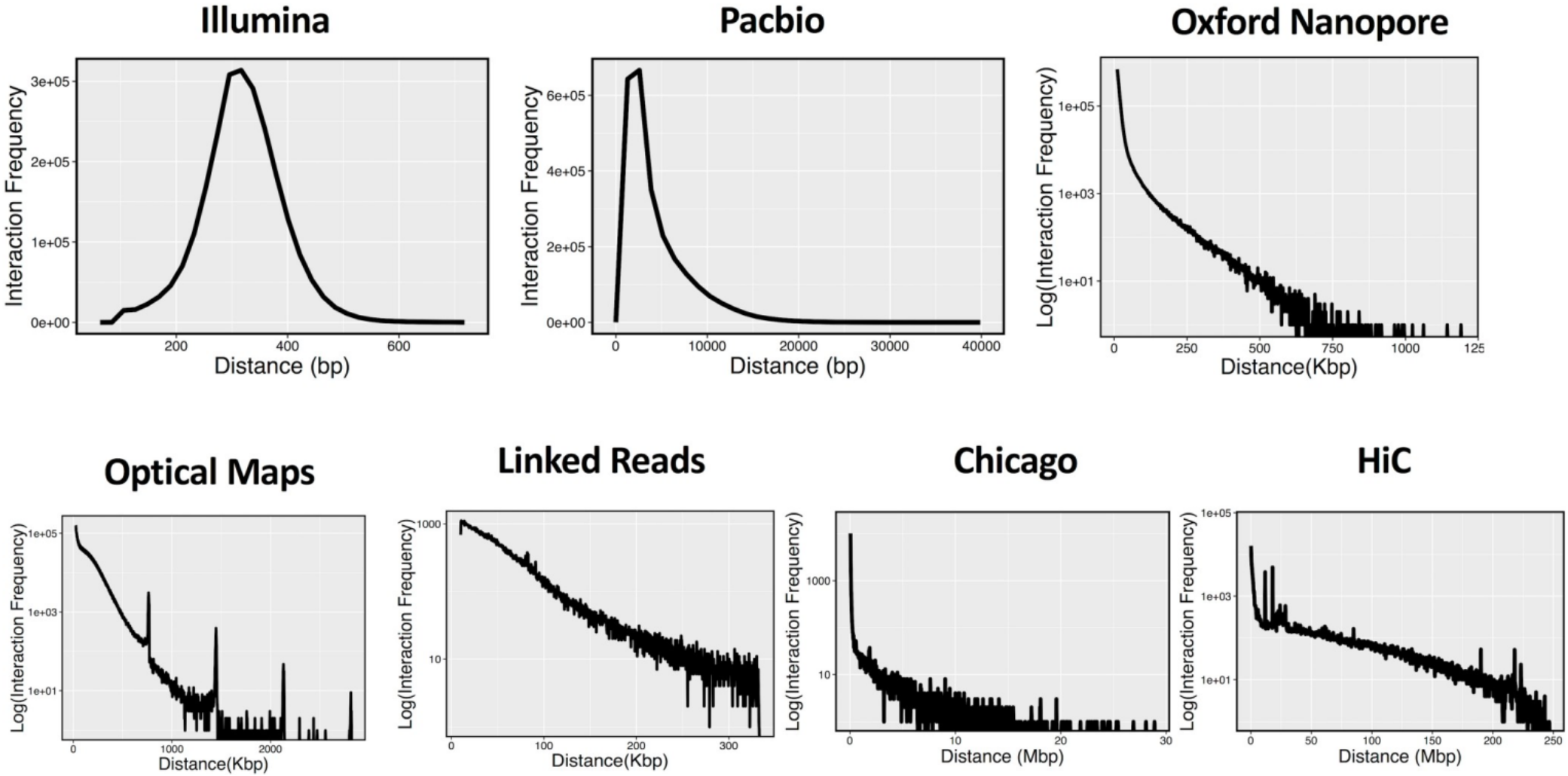
Full-length reads, assembly, transcript- or allele-specific quantification

Decyphering Changes in Current



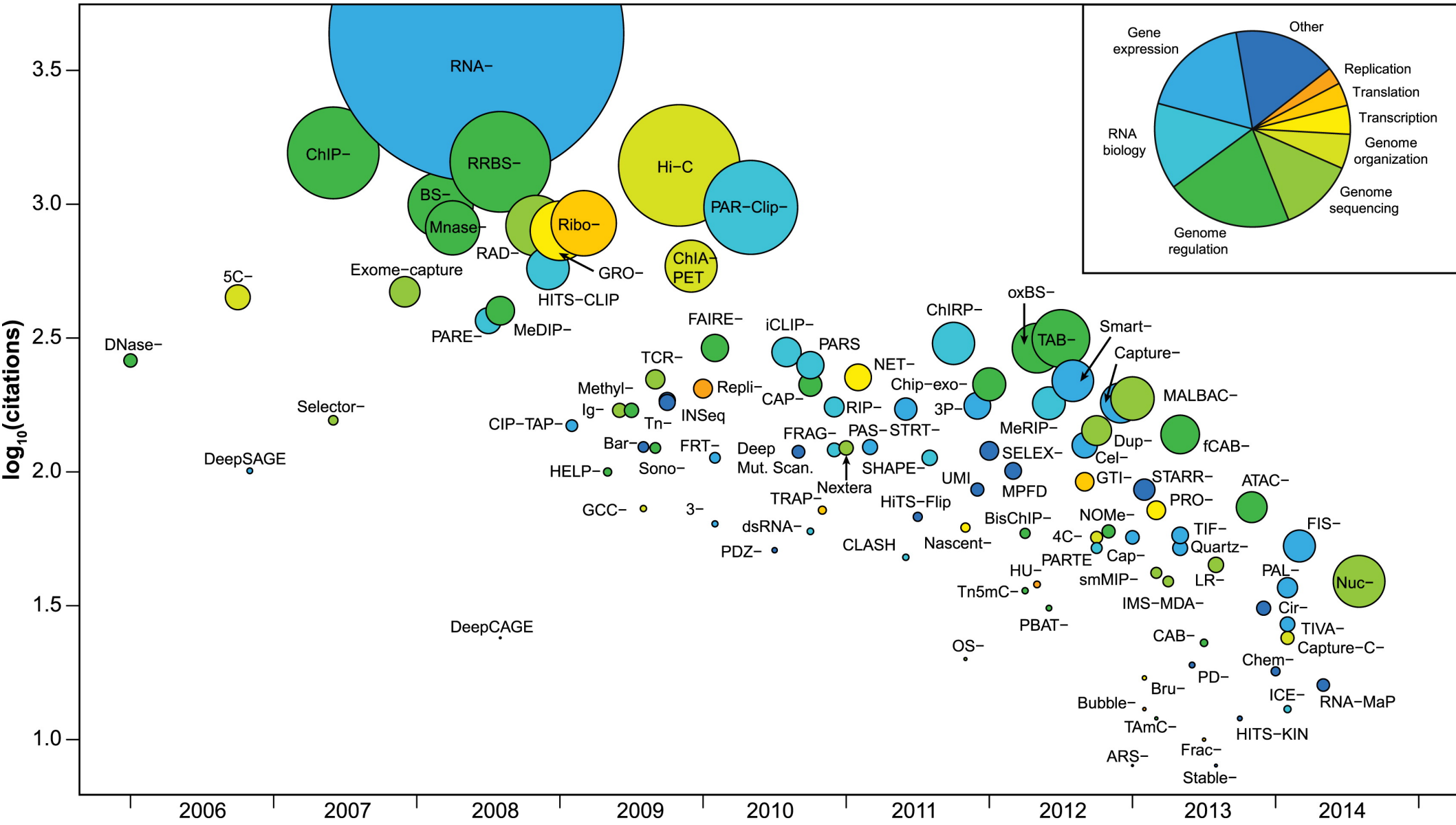
Timp et al, *Biophysical Journal*, 2012

“Reads” Length Distribution



Ghurye & Pop, *PLoS Comput Biol*, 2019

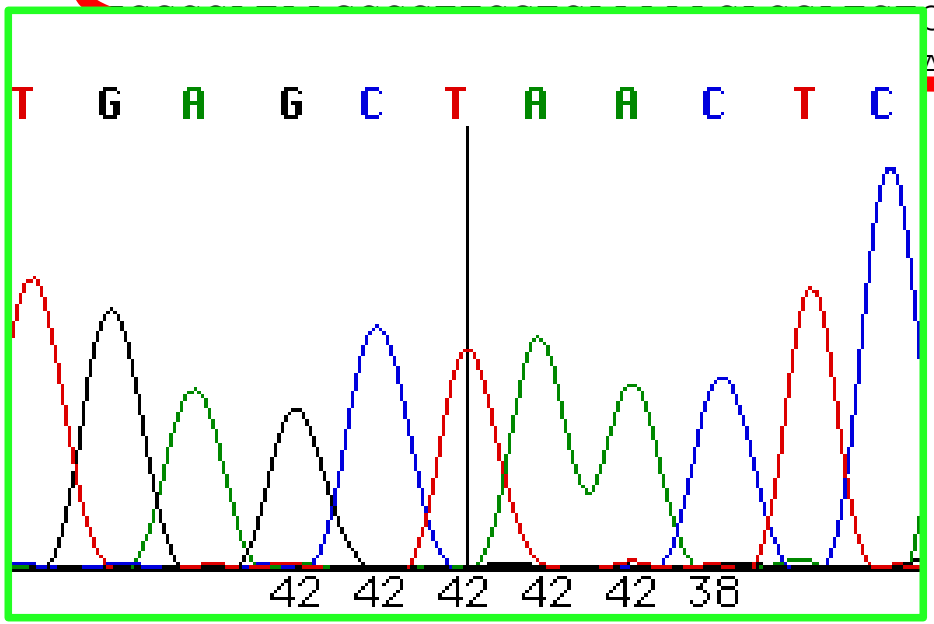
HTS Applications

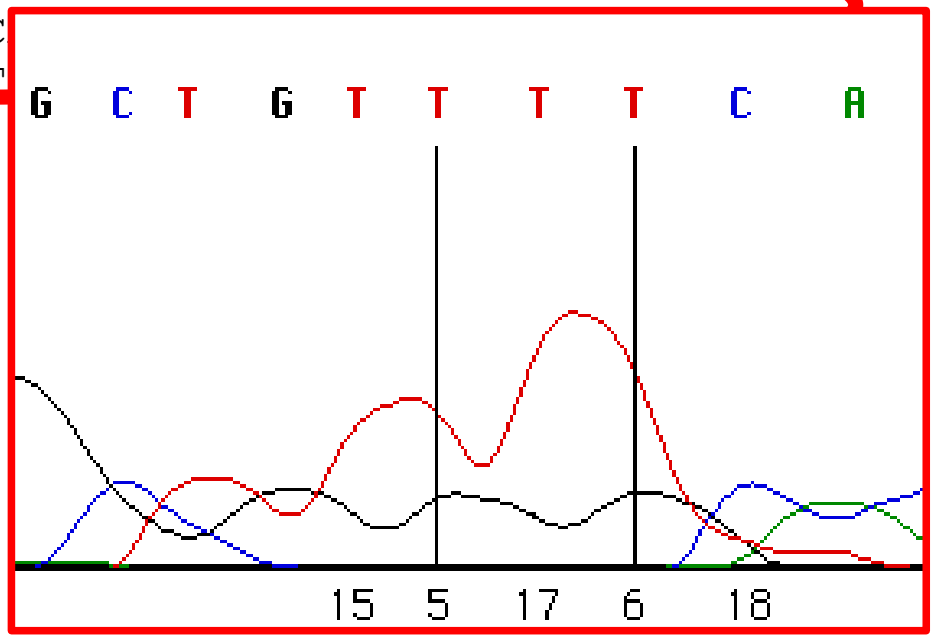


Reuter, Spacek & Snyder, *Molecular Cell*, 2015

Errors in Sanger Trace Sequences

```
>gnl|ti|312701263 name:ufx55a02.b1 mate:312701356
GGGCCCGACACGTAAACATCGCCACTGGTGGTGGGATTCAAGTACTAAAAGAATACCATCAACATAACA
TTTATGTATATTAGACCTAACTAGAAGAAAATCAGAAATATTTTTCTCTTATTGTAAAAGAAAAATTGAT
GCAACAGTTCTAACTAAAGGTGATTCCATAGCCACTCGATCAATTTGTTGATACCAATTTACTTGGGAA
AAACTCAACATTAAAAACTCCCAAACATCAACAATTCACGAAATGTAACTTCATCTATTGTTTGACCCTT
CAACATATTCTTTTGTCACAAATCATAACAAAGATCAATAGTCTCAATTCCTAGAACATCTGTAAACAAA
TTAGTTACATCAAAAGTCGCCATATAATCAGAAGAATCAGTAACTTGTTTTGTTATAATCAACAAATTCA
AATGAATTCCGACACTCATTCCTTAAAAATCAAAGATTGTAATAATGAATCAAGAAAATGAACAATCTTA
TATTGTATTGAATCTATCATGGAAAGTACTTGGTGCATTCAAACACCCTTTTTTAGTTTTTGGTCGACCA
TATAATTTTGCAGATTGTGACCCAACTGTGCATATTATTGCTTACCAAGATCACAATCTAAATGCTTATC
TTTCATCACACGGAAAACAAGTTTTTTTAAACAATTTCTTAATTTTTTTTGAGGGAGAAATTTGAGTAAA
TTATTCATCTTAACAACGAATTAATCAACTTCATTCAAAACTACAAACTTACAAACTTTCACAGACTTTTT
AGAGAGTGTCTTTAAAACAGCAAAATCGCTAATATTTTGTAATTTATTATTTTTGAGATAATCAACAGCA
CTTTGATTATAAGCTTTAATACATTTTTTATATTACGAATTTTTTGATTAACTACGATATAATTCAATCT
                                                    CTATTACC
                                                    ACAACTGT
```



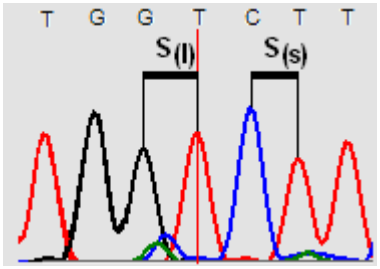


PHRED Score

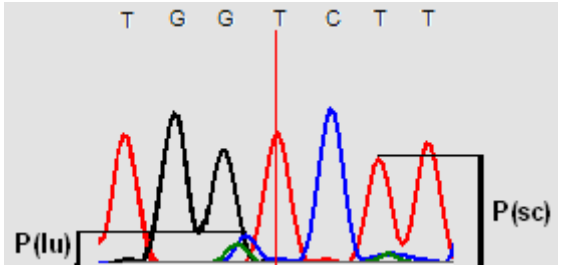
$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

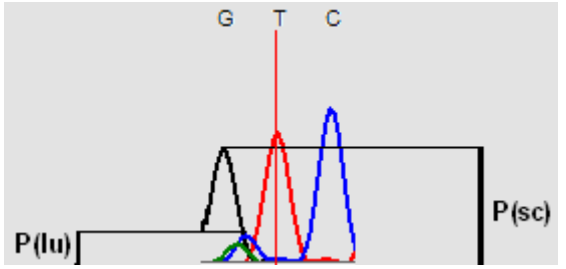
Ewing et al, *Genome Research*, 8: 175-185, 1998
Ewing & Green, *Genome Research*, 8: 186-194, 1998



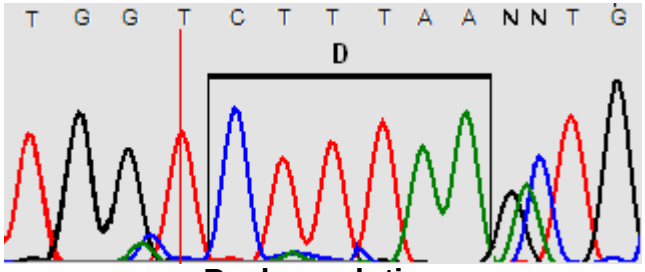
Peak spacing



Uncalled to called peak height ratio



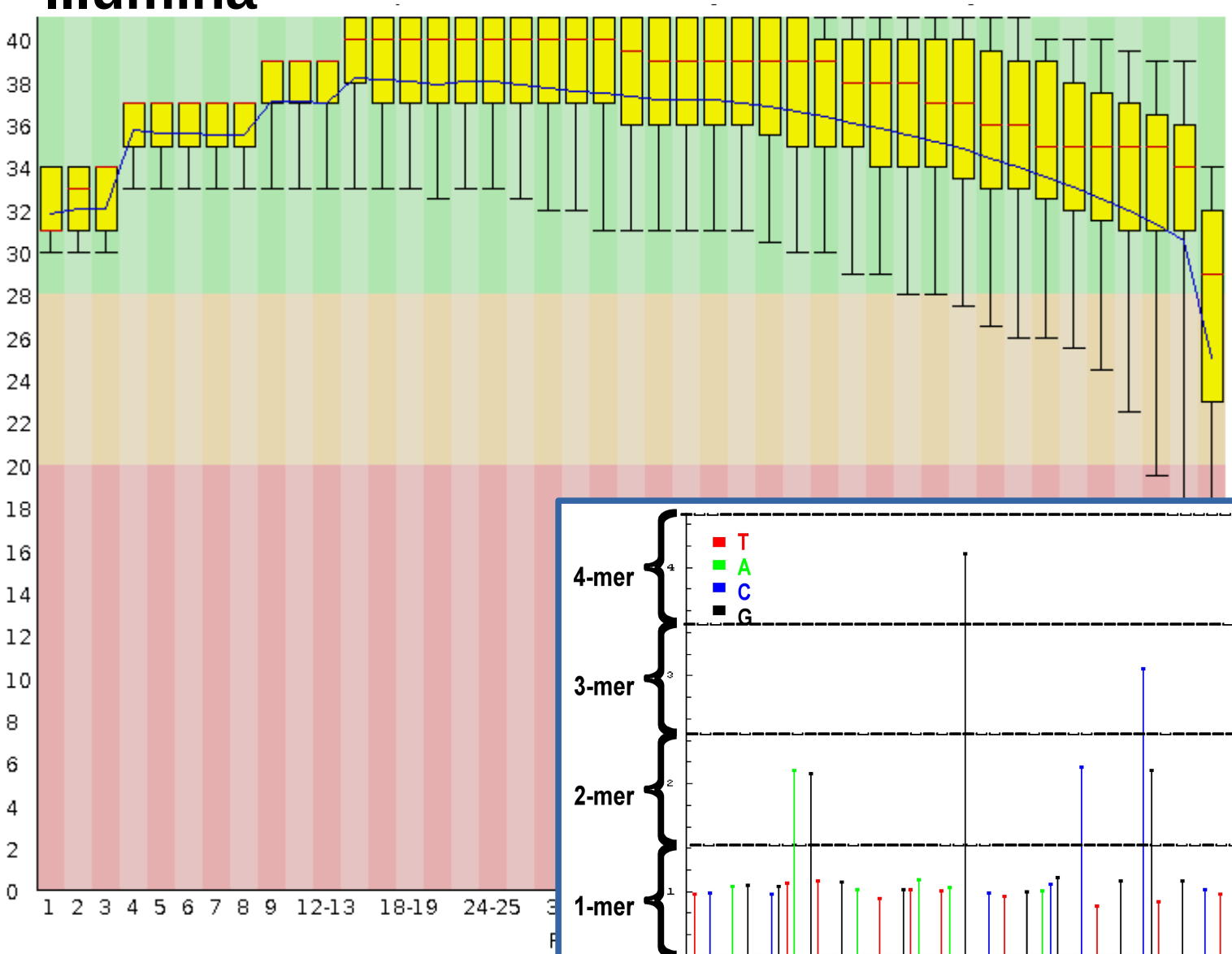
Uncalled to called peak height ratio in 3bp window



Peak resolution

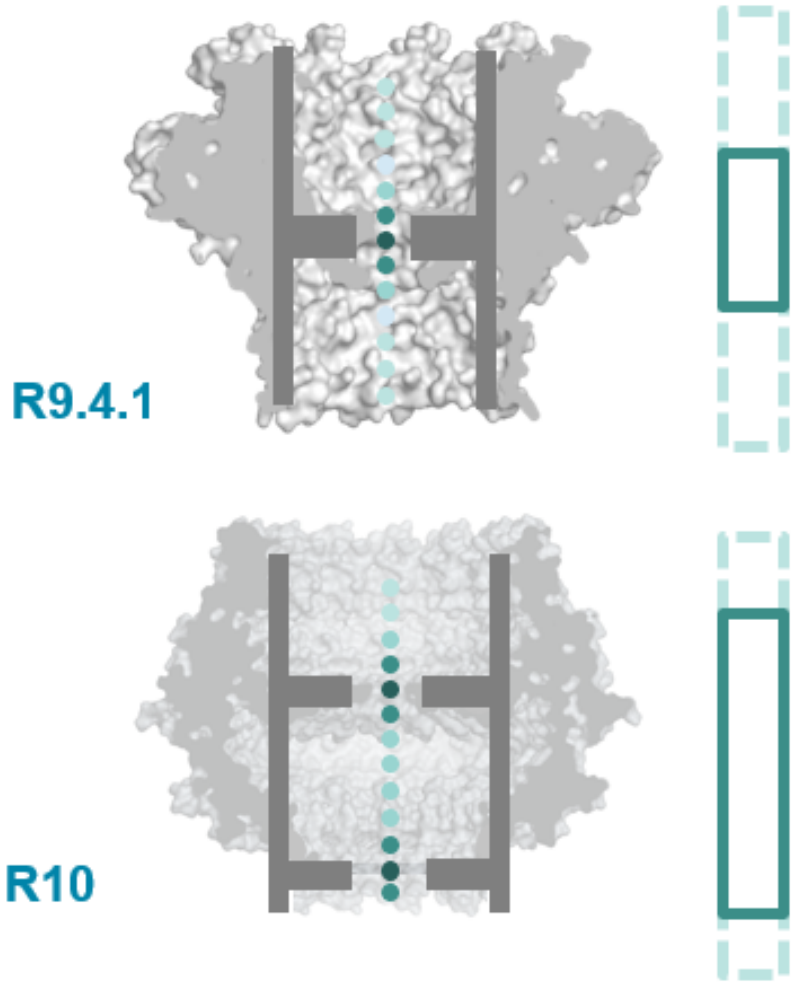
Errors in NGS Read Sequences

Illumina



Roche 454 (flowgram)

Single Molecule Stochastic Errors

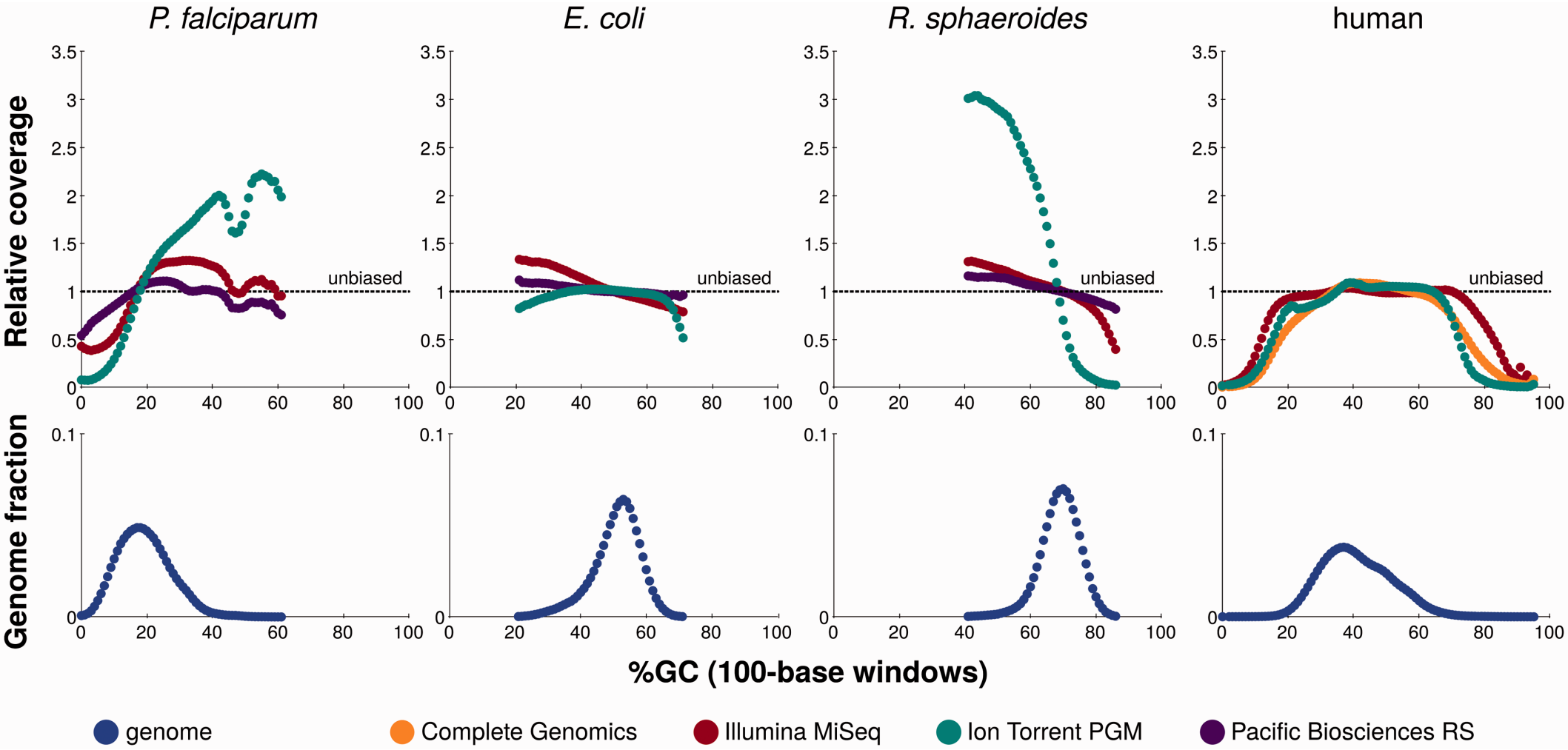


A FASTQ RECORD

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTAATAAATAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGATTTGTTGGGGGAGA
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffeefffcfddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`ddd^dddadd^BBB
```

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
  0........................26...31........41

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

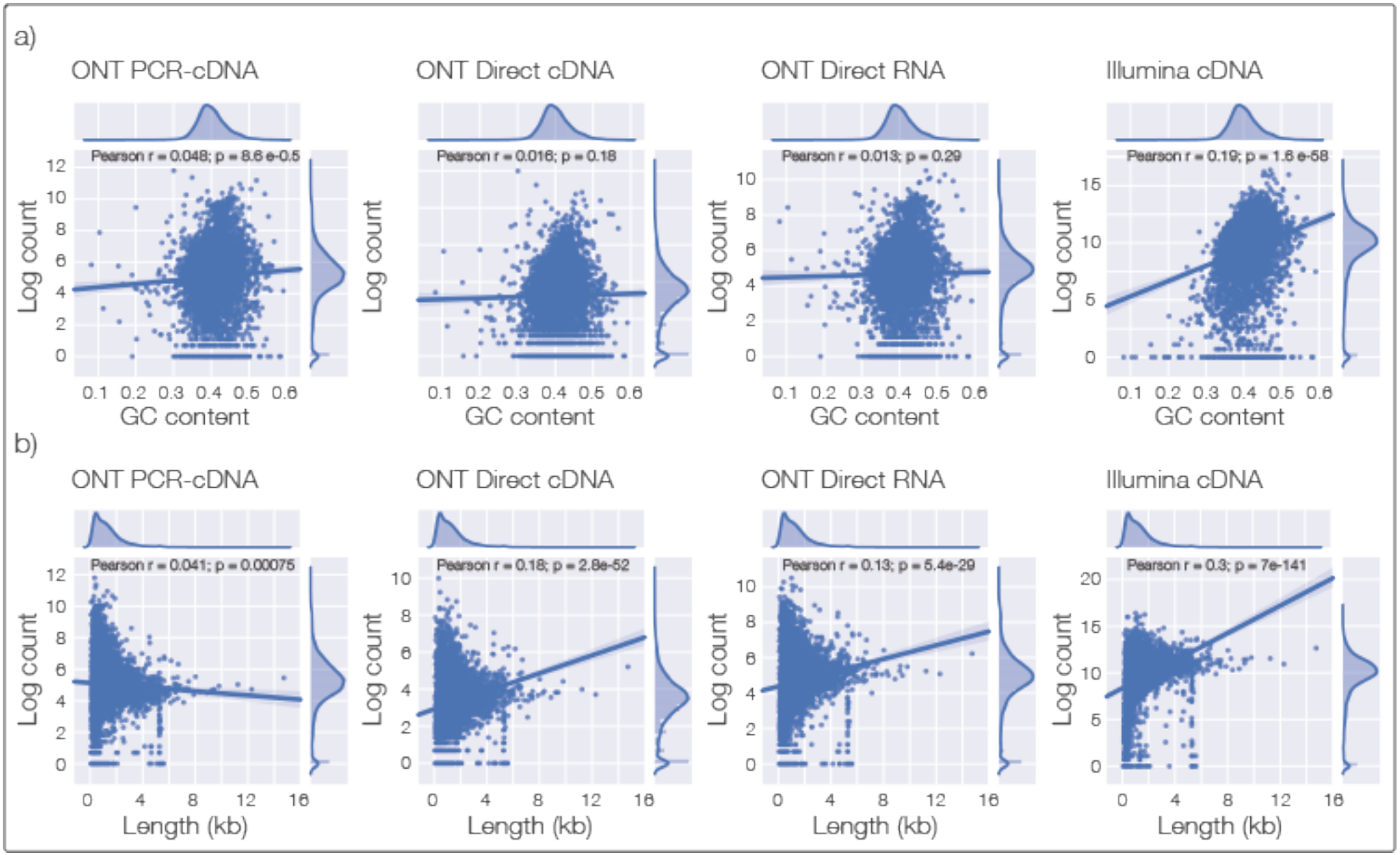
NGS Sequencing Biases



Laehnemann, Borkhardt & McHardy, *Briefings in Bioinformatics*, 2016

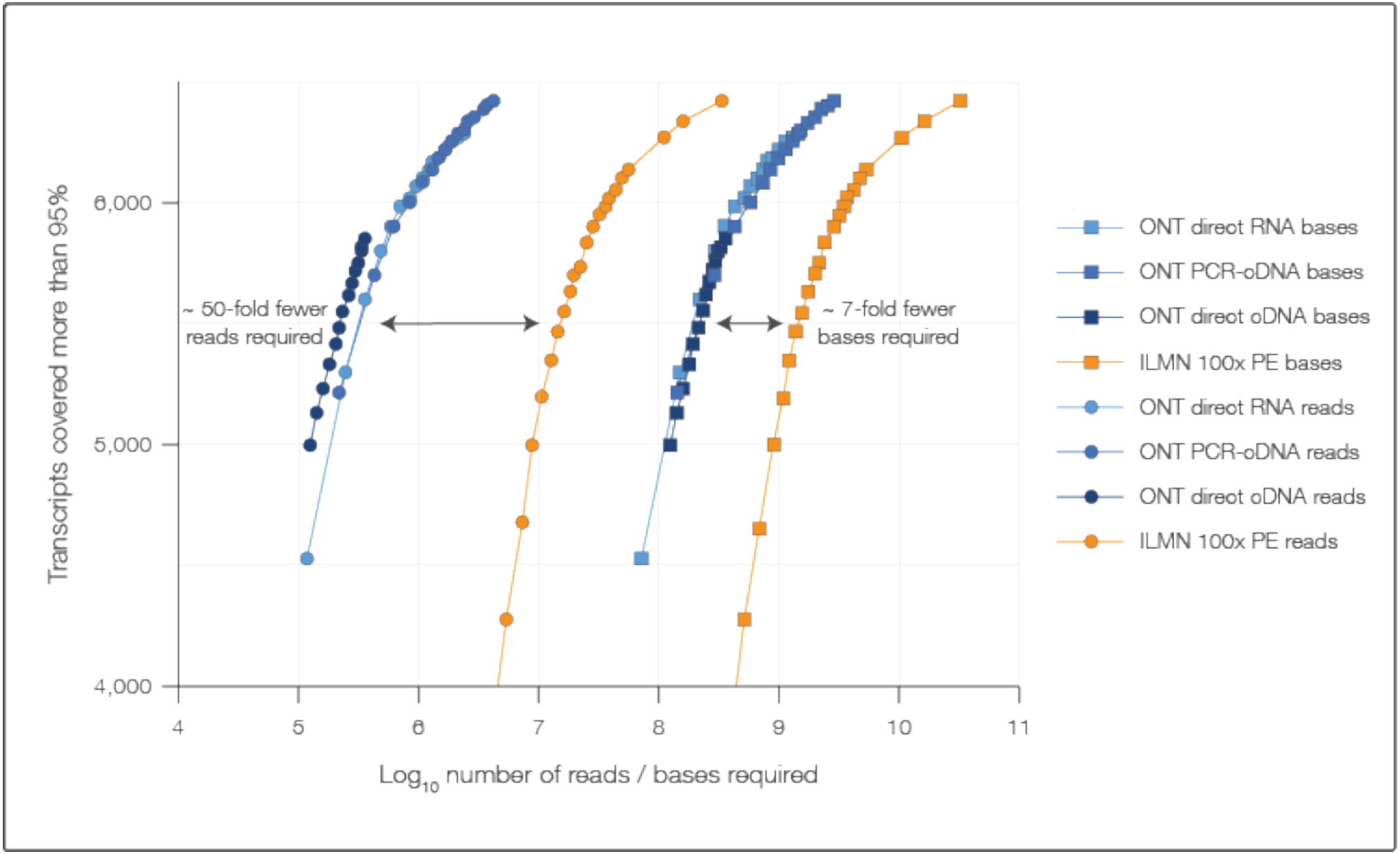
RNA-seq Sequencing Bias

Quantitative assessment of bias a) GC bias b) length bias



Less is More...

Data required to cover transcripts along 95% of their length, in terms of bases and reads



NGS Error Rates

Error rates of high-throughput sequencing platforms (per 100 sequenced bases)

Platform	Substitutions	SD Subs	Indels	SD Indels
454 GS FLX	0.090	N/A	0.900	N/A
454 GS Junior	0.054	N/A	0.391	N/A
Complete Genomics	2.300	N/A	0.019	N/A
Illumina HiSeq	0.264	0.112	0.026	0.024
Illumina MiSeq	0.246	0.111	0.009	0.014
Ion Torrent PGM	0.170	0.173	1.458	1.219
Pacific Biosciences RS	1.103	0.448	15.566	3.294

Laehnemann, Borkhardt & McHardy, *Briefings in Bioinformatics*, 2016

High-throughput Sequencing Biases

Company (former companies)	Platforms	Library amplification	Carrier of library molecules or beads during sequencing	Sequencing principle	Nucleotide modifications	Signal detection method	Dominant type of sequencing error
Roche (454 until 2006)	454 FLX Titanium 454 FLX+ 454 GS Junior Titanium	emPCR on microbeads	Picotiterplate	Pyrosequencing	None (except for dATP, which is added as thiol derivative dATPαS)	Optical detection of light, emitted in secondary reactions initiated by release of PP_i upon nucleotide incorporation	Indels in homopolymeric regions
Illumina (Solexa until 2007)	Illumina GAIIx Illumina HiSeq1000 Illumina HiSeq1500 Illumina HiSeq2000 Illumina HiSeq2500 Illumina MiSeq Illumina NextSeq 500 Illumina HiSeq X ten	Bridge-PCR on flow cell surface	Flow cell	Reversible terminator sequencing by synthesis	End-blocked fluorescent nucleotides	Optical detection of fluorescent emission from incorporated dye-labeled nucleotides	Substitutions, in particular at the end of the read
Life Technologies (Agencourt until 2006, Applied Biosystems until 2008)	SOLiD 4 SOLiD 5500 SOLiD 5500xl SOLiD 5500 W SOLiD 5500xl W	emPCR on microbeads; PCR on FlowChip surface for the 5500 W models	FlowChip	Sequencing by ligation	2-base encoded fluorescent oligonucleotides	Optical detection of fluorescent emission from ligated dye-labeled oligonucleotides	Substitutions, in particular at the end of the read
Life Technologies (Ion Torrent until 2010)	Ion PGM Ion Proton	emPCR on microbeads	Ion Chip, a semiconductor chip	Semiconductor-based sequencing by synthesis	None	Transistor-based detection of H^+ shift upon nucleotide incorporation	Indels
Pacific biosciences	PacBio RS	Not applied	SMRT cell	Single-molecule, real-time DNA sequencing by synthesis	Phosphor-linked fluorescent nucleotides	Real-time optical detection of fluorescent dye in polymerase active site during incorporation	Indels

Knief C, *Front. Plant Sci.*, 2014

Sboner *et al. Genome Biology* 2011, **12**:125
http://genomebiology.com/2011/12/8/125

Genome Biology

OPINION

The real cost of sequencing: higher than you think!

Andrea S

Abstra
Advanc
sharp d
this suf
'knowle

Keywo
data an
sequen

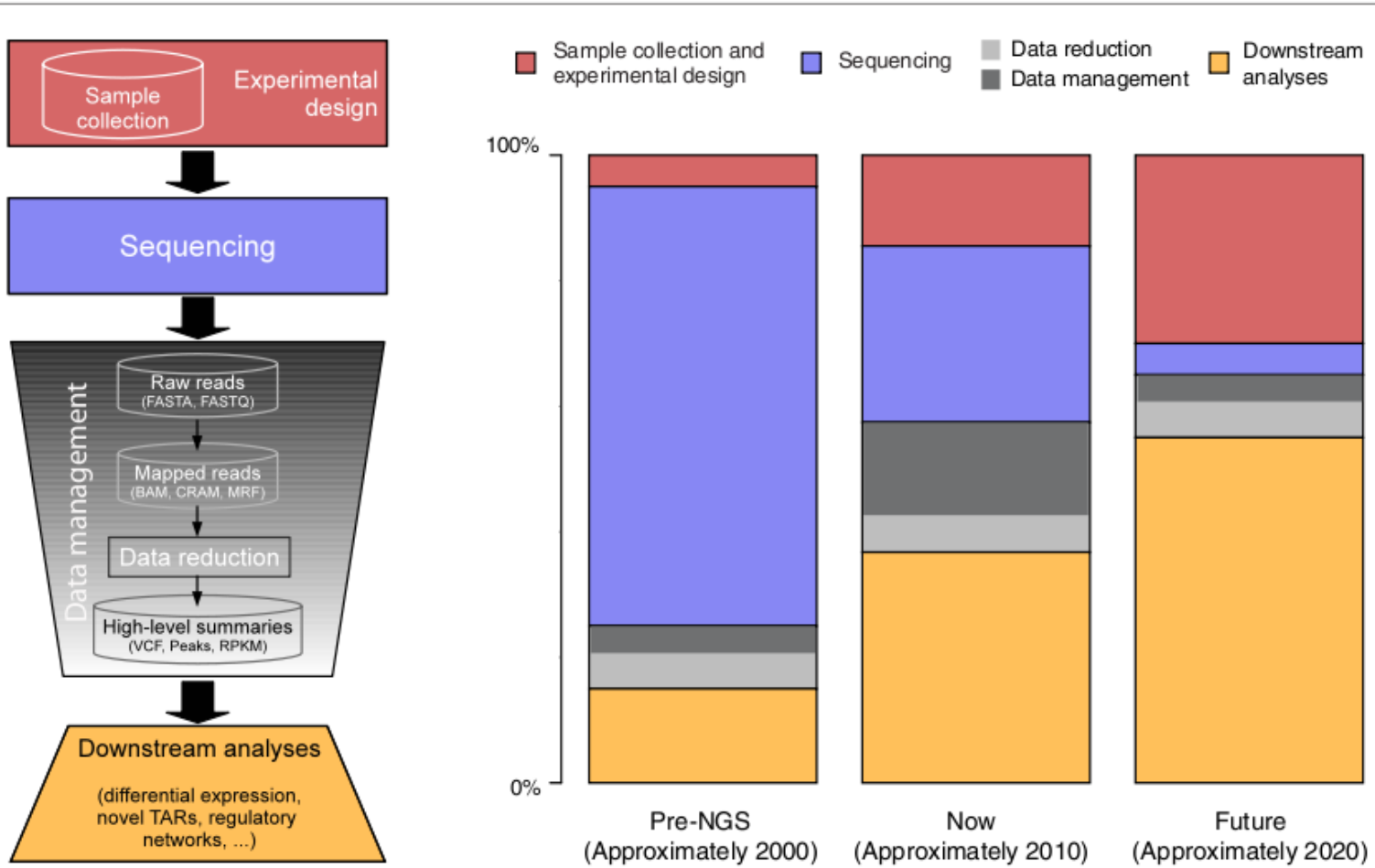
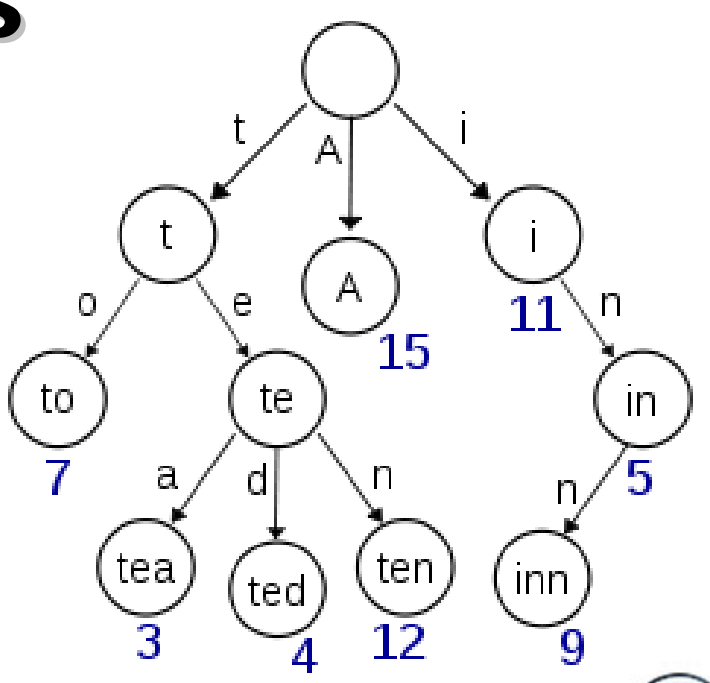


Figure 1. Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.

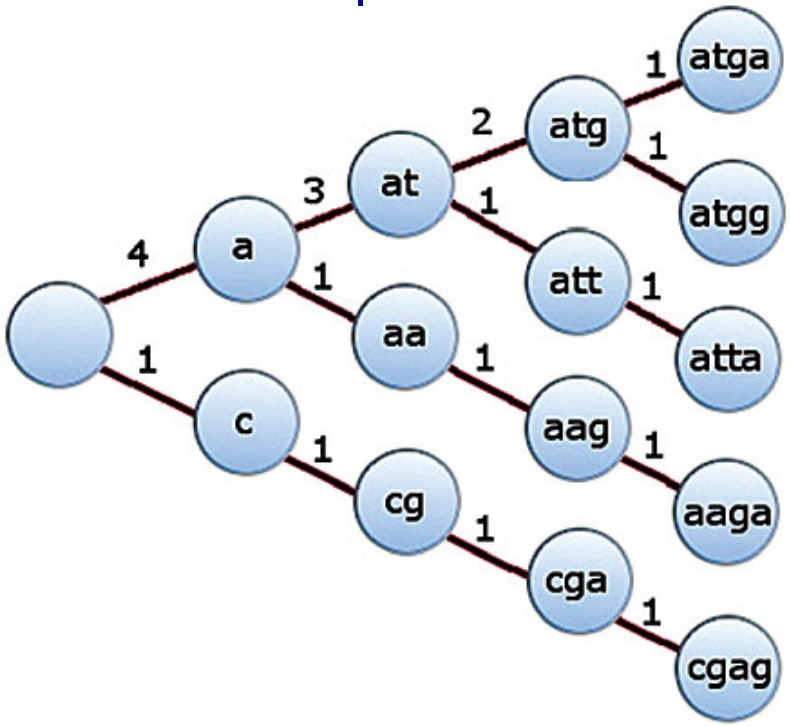
SEQUENCE ANALYSIS

Prefix Trees

A **trie**, or **prefix tree**, is an ordered tree data stucture used to store an associative array, where the keys are usually strings. It can be used to replace a hash table, over which it has several advantages.



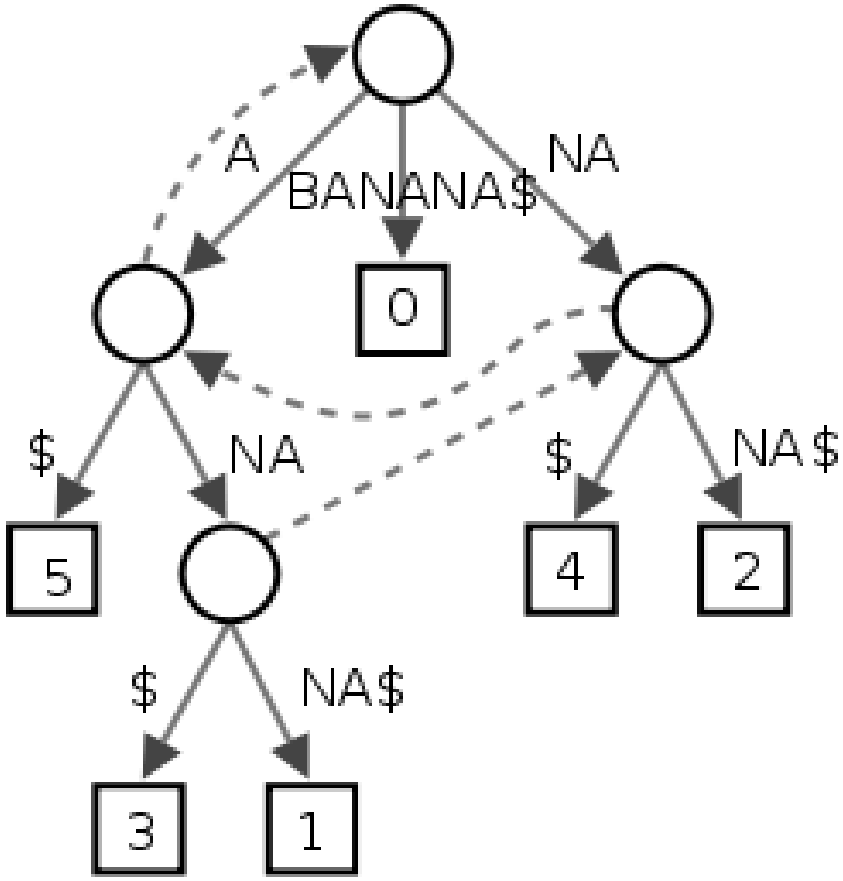
- Looking up data in a trie is faster in the worst case, O(m) time, compared to an imperfect hash table. An imperfect hash table can have key collisions. A key collision is the hash function mapping of different keys to the same position in a hash table. The worst-case lookup speed in an imperfect hash table is O(N) time, but far more typically is O(1), with O(m) time spent evaluating the hash.
- There are no collisions of different keys in a trie.
- Buckets in a trie which are analogous to hash table buckets that store key collisions are only necessary if a single key is associated with more than one value.
- There is no need to provide a hash function or to change hash functions as more keys are added to a trie.
- A trie can provide an alphabetical ordering of the entries by key, which can be used to store dictionaries.



Suffix Trees

A **suffix tree**, or **position tree**, is a data structure presenting the suffixes of a given string in a way that allows for a particularly fast implementation of many important string operations. Used to solve a large number of string problems (text-editing, free-text search, computational biology, etc):

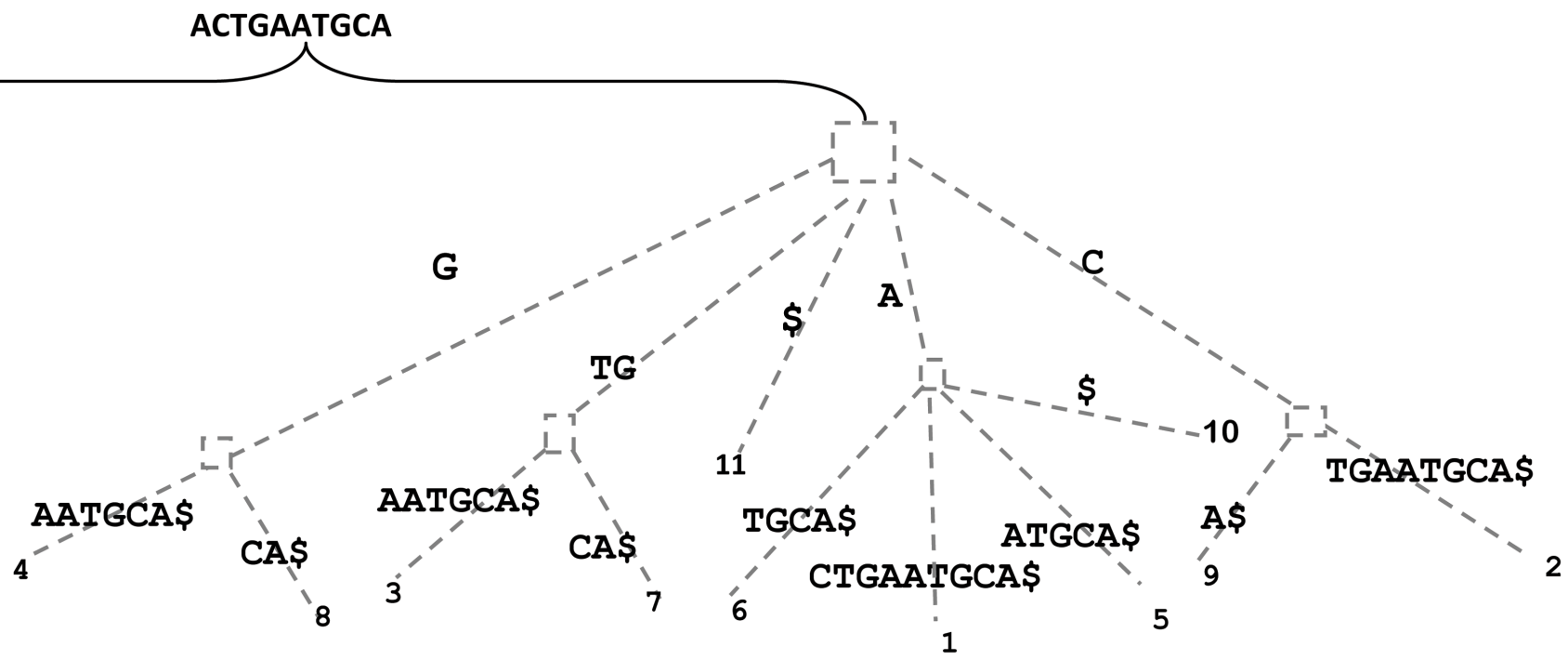
- String search, in O(m) complexity, where m is the length of the sub-string (but with initial O(n) time required to build the suffix tree for the string).
- Finding the longest repeated substring.
- Finding the longest common substring.
- Finding the longest palindrome in a string.
- Bioinformatics applications: searching for patterns in DNA or protein sequences. Its greatest strength is the ability to search efficiently with mismatches.
- Suffix trees are also used in data compression.



Indexing Sequences

ACTGAATGCA$	1
CTGAATGCA$	2
TGAATGCA$	3
GAATGCA$	4
AATGCA$	5
ATGCA$	6
TGCA$	7
GCA$	8
CA$	9
A$	10
$	11

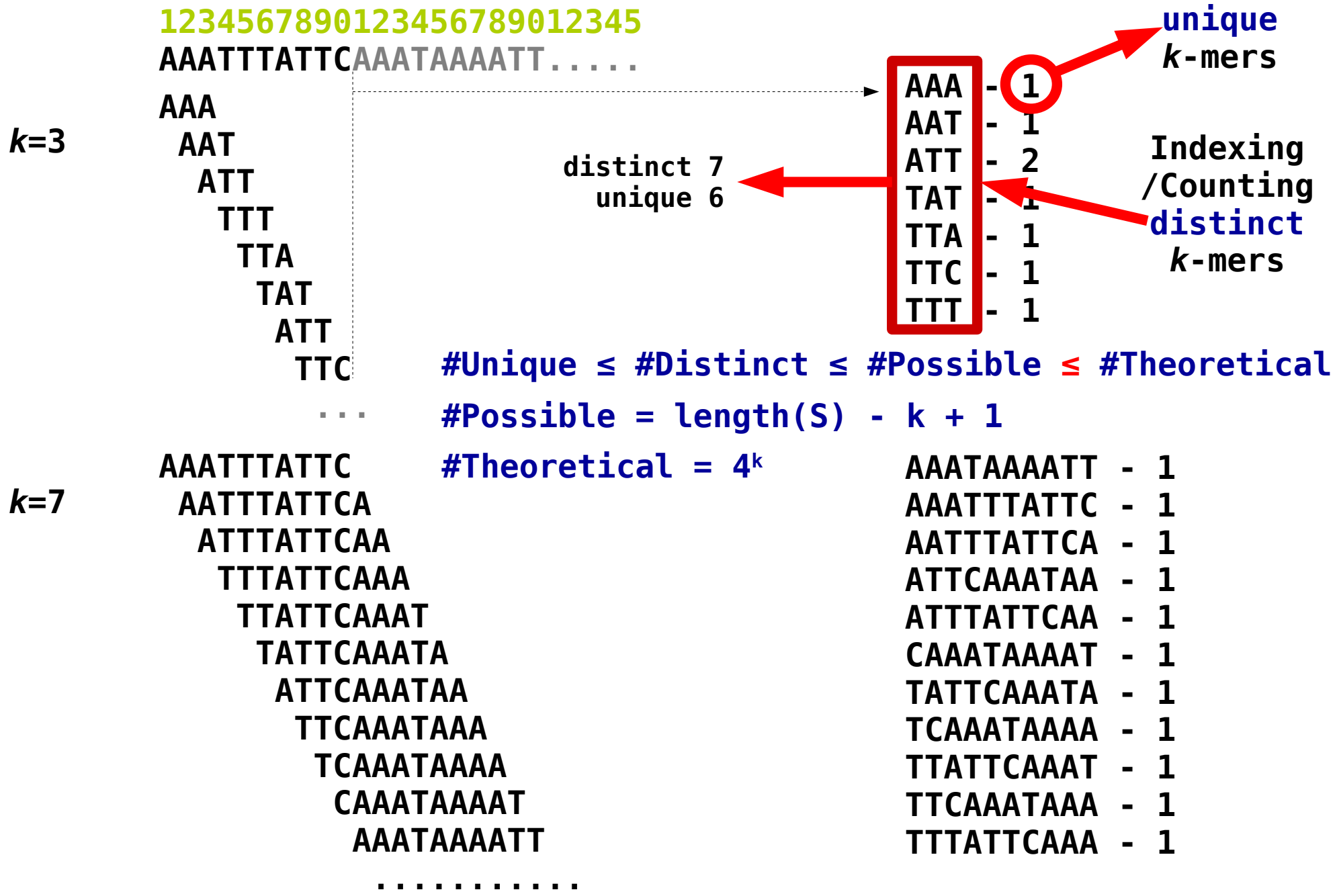
ARRAY APPROACH



SUFFIX TREE APPROACH

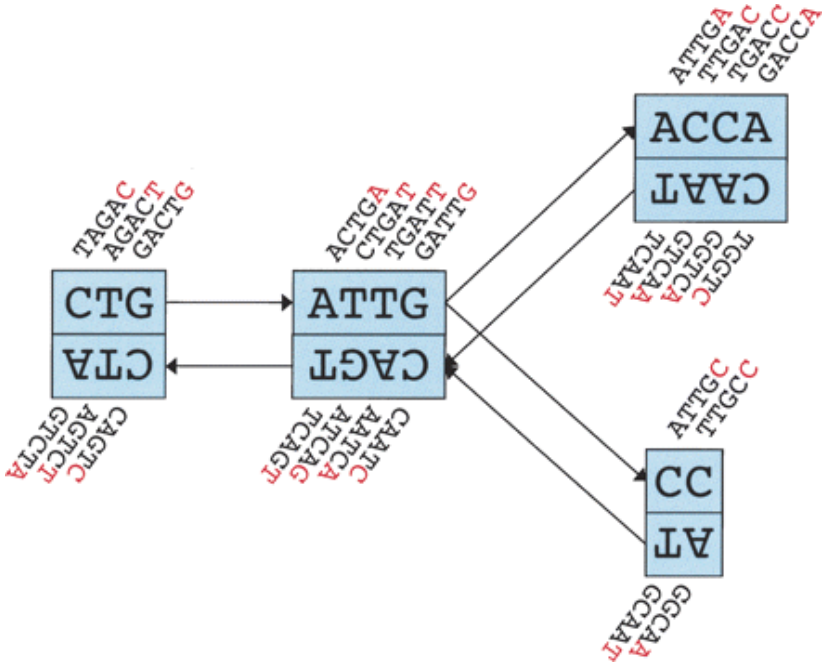
El-Metwally et al, *PLOS Comp. Biol.*, 2013

Counting *k*-mer frequencies



k-mer Counting and Strands

Sequencing is often not strand specific,
so we can consider *k*-mers in both strands.



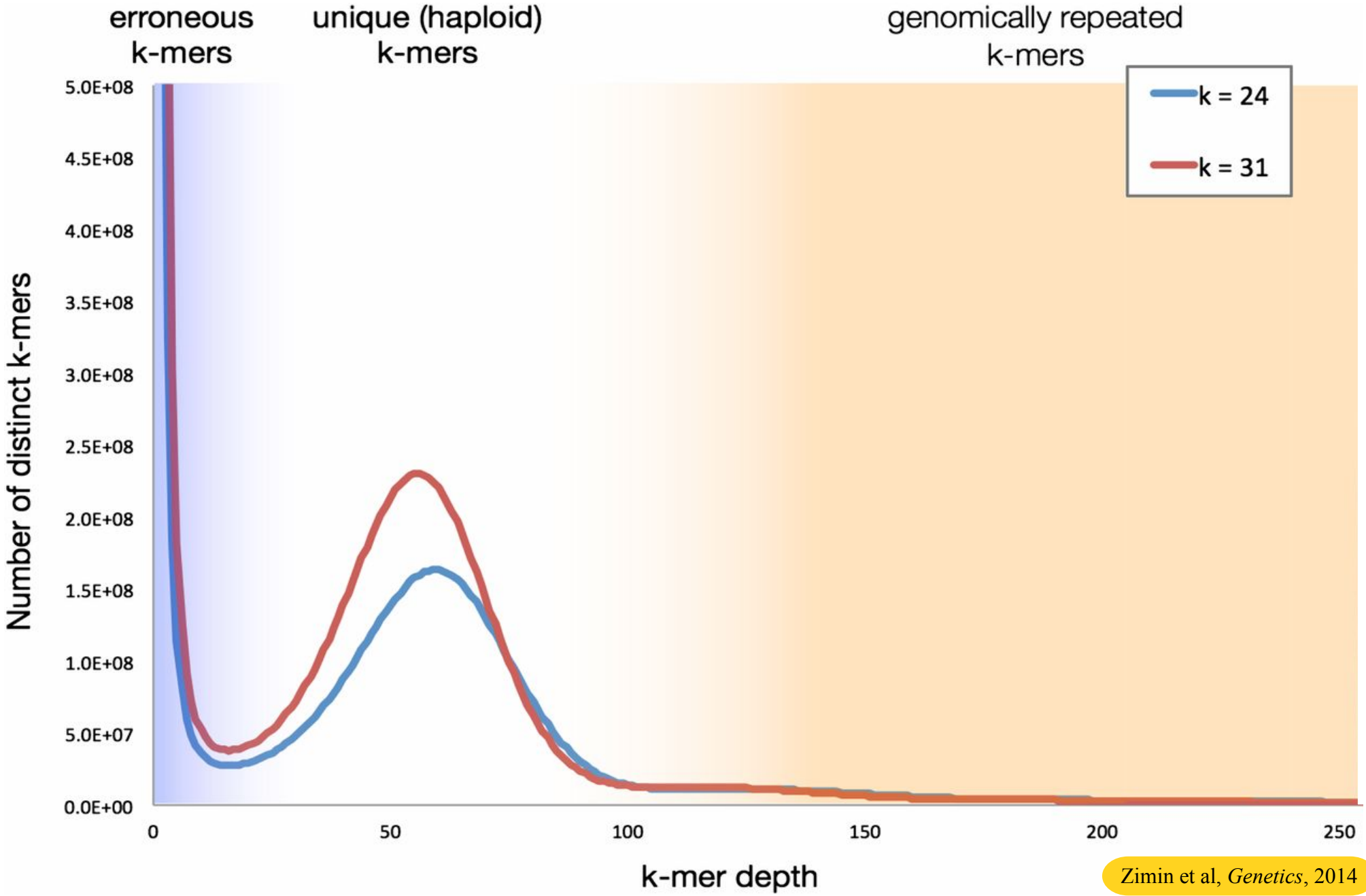
Computing canonical *k*-mers,
take smaller alphabetically
ordered (A<C<G<T) from *k*-mer
and its reverse-complement.

k-mers set	*k*-mers revcomp	canonical *k*-mers
AAATAAAATT	AATTTTATTT	AAATAAAATT
AAATTTATTC	GAATAAATTT	AAATTTATTC
AATTTATTCA	TGAATAAATT	AATTTATTCA
ATTCAAATAA	TTATTTGAAT	ATTCAAATAA
ATTTATTCAA	TTGAATAAAT	ATTTATTCAA
CAAATAAAAT	ATTTTATTTG	ATTTTATTTG
TATTCAAATA	TATTTGAATA	TATTCAAATA
TCAAATAAAA	TTTTATTTGA	TCAAATAAAA
TTATTCAAAT	ATTTGAATAA	ATTTGAATAA
TTCAAATAAA	TTTATTTGAA	TTCAAATAAA
TTTATTCAAA	TTTGAATAAA	TTTATTCAAA

k-mer Counters

Algorithm	$k = 28$ RAM	$k = 28$ Disk	$k = 28$ Time	$k = 55$ RAM	$k = 55$ Disk	$k = 55$ Time
			SSD			
Jellyfish 2	62	0	3,212	*out of memory*		
KAnalyze	*out of disk (> 650 GB)*			*unsupported k*		
DSK	6	263	5,487	6	256	7,732
Turtle	*out of memory*			*out of memory*		
MSPKC	*out of time (> 10 hours)*			*out of time (> 10 hours)*		
KMC 1	17	396	2,998	*out of disk (> 650 GB)*		
KMC 2 (12GB)	12	101	1,615	13	70	2,038
KMC 2 (6GB)	6	101	1,706	13	70	2,446
			HDD			
Jellyfish 2	62	0	3,231	*out of memory*		
DSK	6	263	18,493	6	256	22,432
KMC 1	17	396	4,898	*out of disk (> 650 GB)*		
KMC 2	12	101	2,259	13	70	2,640

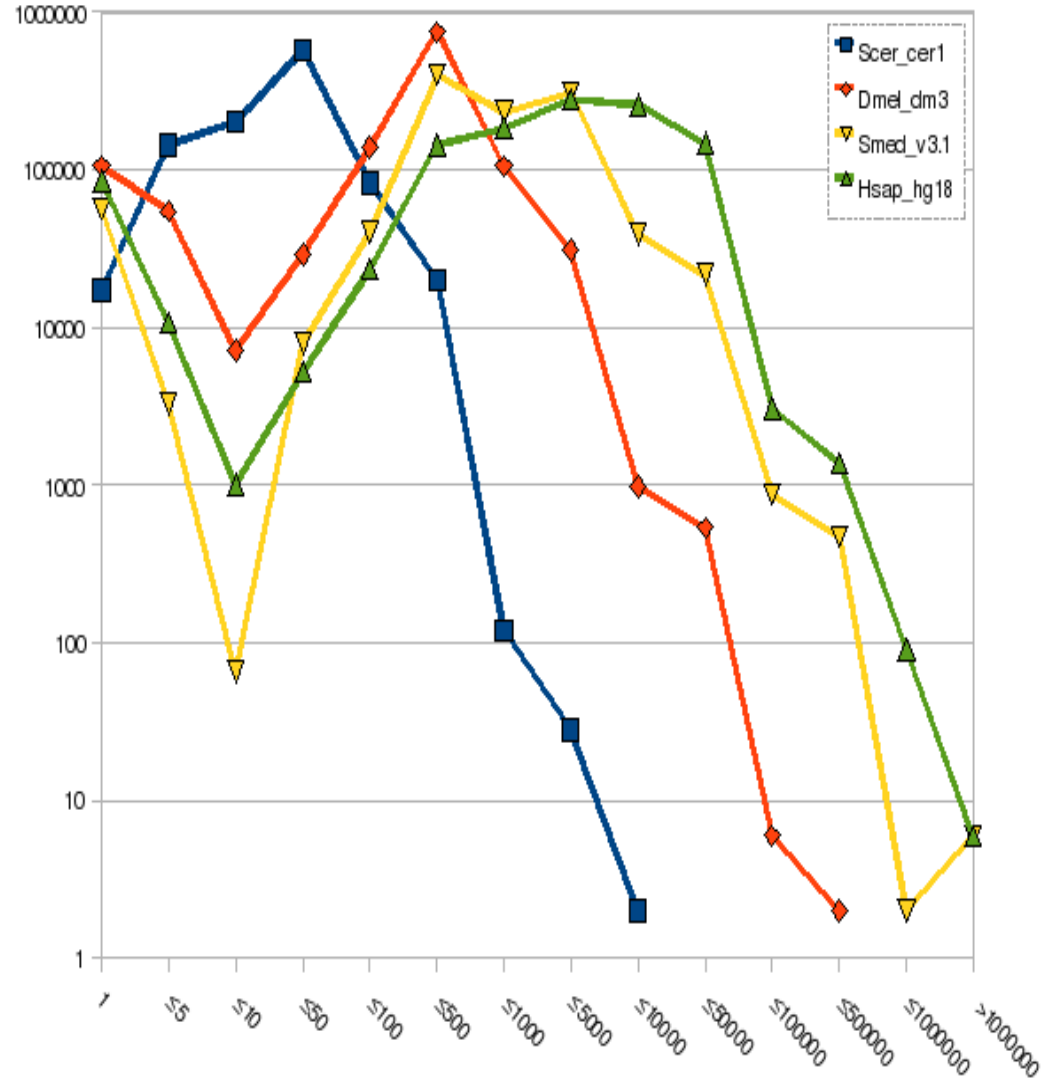
Distinc *k*-mers Depth



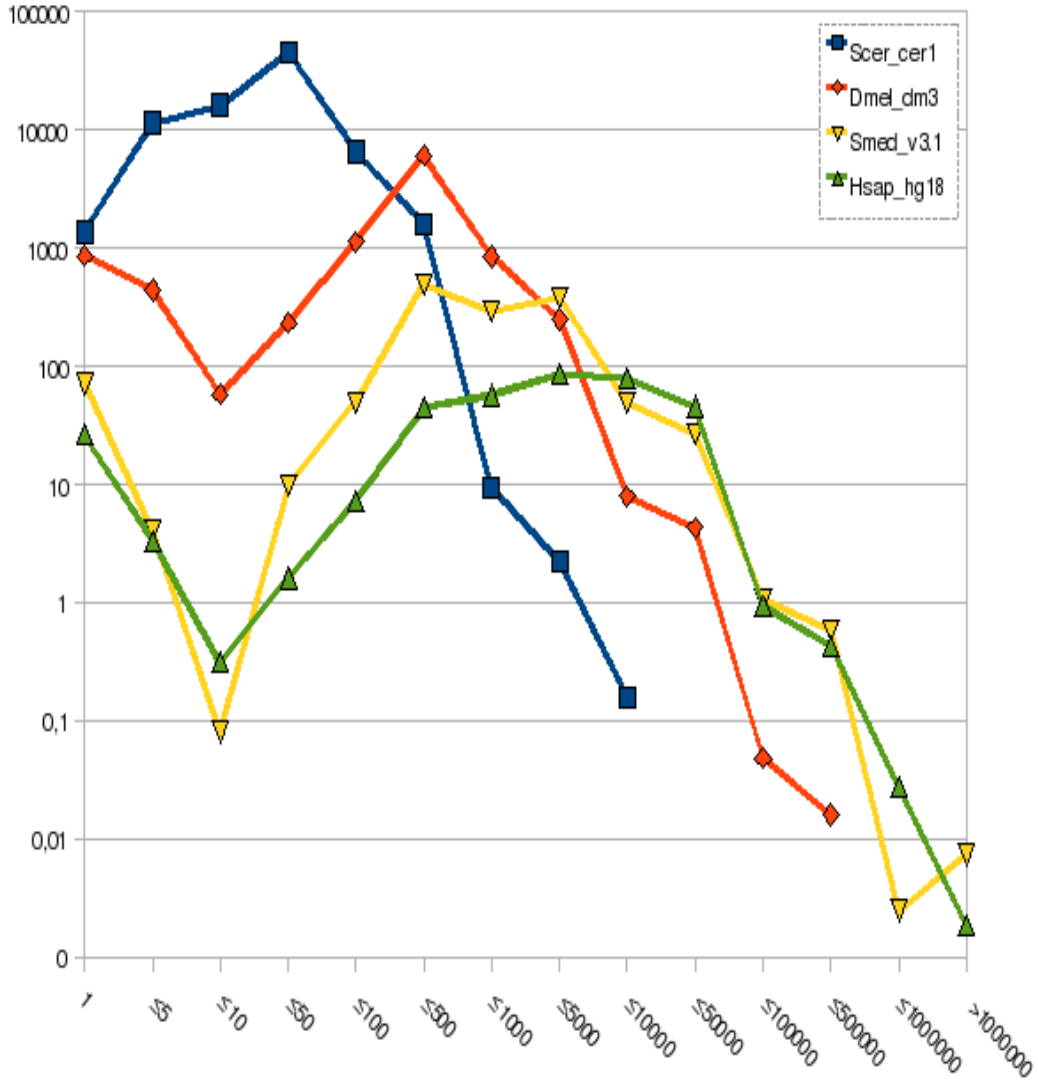
Zimin et al, *Genetics*, 2014

K-mers Depth on Different Genomes

k=10bp



k=10bp, counts adjusted to genome size



Ultra frequent *k*-mers in Human Genome

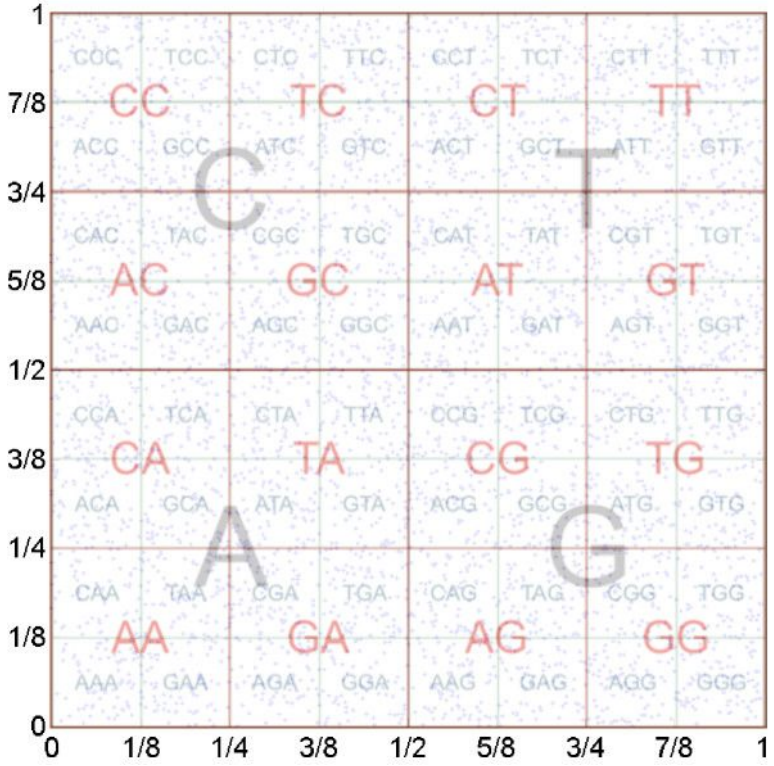
5-mer	*Freq*	*6-mer*	*Freq*	*7-mer*	*Freq*	*8-mer*	*Freq*
AAAAA	38,658,471	AAAAAA	19,638,479	AAAAAAA	12,559,969	AAAAAAAA	9,155,123
ATTTT	23,349,997	AAAAAT	9,299,025	AAATAAA	3,521,836	ATATATAT	1,863,212
TATTT	19,344,297	TATTTT	8,283,181	AAAAAAT	3,335,157	TGTGTGTG	1,701,426
AGAAA	18,271,461	AAATAA	7,271,302	AAAGAAA	3,255,464	TTTTAAAA	1,574,448
AAATT	16,119,174	AGAAAA	7,027,241	TTTTAAA	3,218,950	AAAATAAA	1,562,386

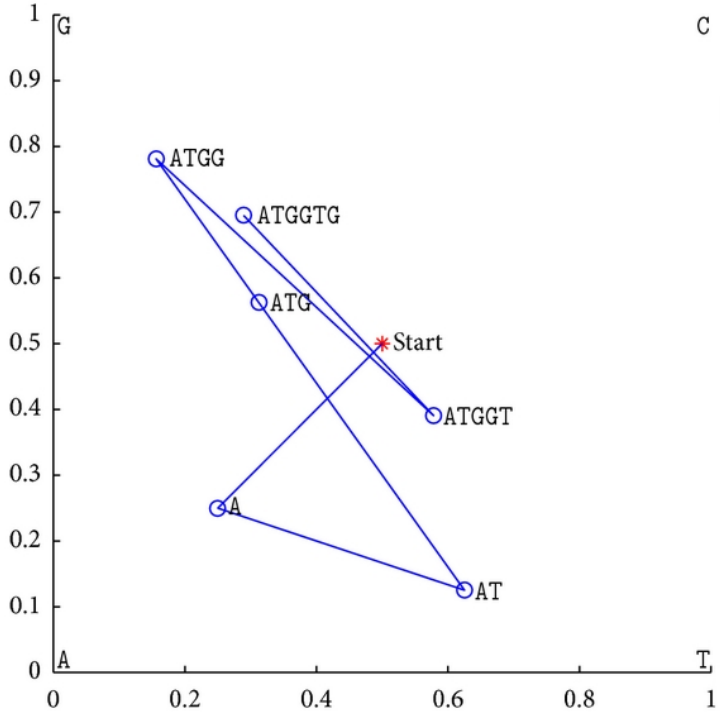
9-mer	*Freq*	*10-mer*	*Freq*	*11-mer*	*Freq*
AAAAAAAAA	7,276,886	AAAAAAAAAA	5,952,617	AAAAAAAAAAA	4,945,619
TGTGTGTGT	1,424,846	TGTGTGTGTG	1,168,929	TGTGTGTGTGT	1,067,659
ATATATATA	1,253,711	ATATATATAT	967,302	CTGTAATCCCA	804,201

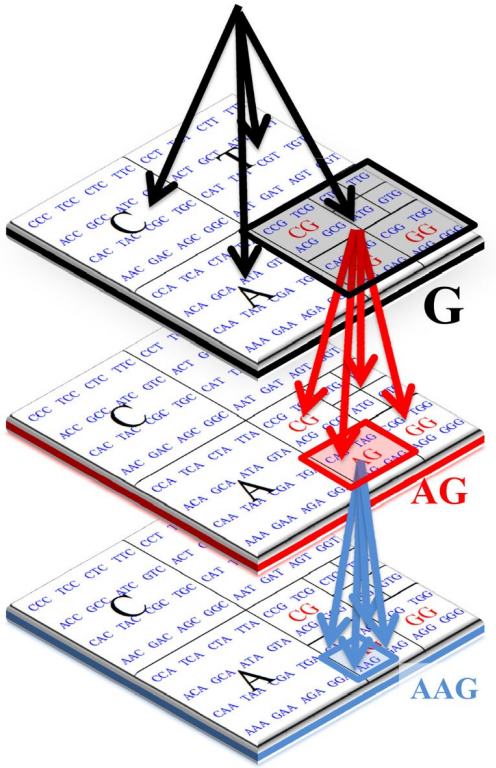
12-mer	*Freq*	*13-mer*	*Freq*	*14-mer*	*Freq*
AAAAAAAAAAAA	4,144,156	AAAAAAAAAAAAA	3,468,084	AAAAAAAAAAAAAA	2,889,704
TGTGTGTGTGTG	928,266	TGTGTGTGTGTGT	867,556	GTGTGTGTGTGTGT	775,928
TGGGATTACAGG	744,980	CTGTAATCCCAGC	687,709	CCTGTAATCCCAGC	639,010
CTGGGATTACAG	737,944	CTGGGATTACAGG	684,576	ATATATATATATAT	503,574
GCTGGGATTACA	727,608	ATATATATATATA	571,291	CTGGGATTACAGGC	478,945
ATATATATATAT	664,692	GCCTGTAATCCCA	520,611	AGCACTTTGGGAGG	459,948
GGAGGCTGAGGC	562,149	GCCTCCCAAAGTG	493,408	GCACTTTGGGAGGC	448,768
GCCTGTAATCCC	532,230	GGAGGCTGAGGCA	489,998	AAGTGCTGGGATTA	448,388
GCCTCCCAAAGT	527,552	CTCCCAAAGTGCT	486,025	AAAGTGCTGGGATT	445,915
TGCCTCAGCCTC	523,666	GCACTTTGGGAGG	474,289	CTCCCAAAGTGCTG	443,681

Liu et al, *BMC Genomics*, 2008

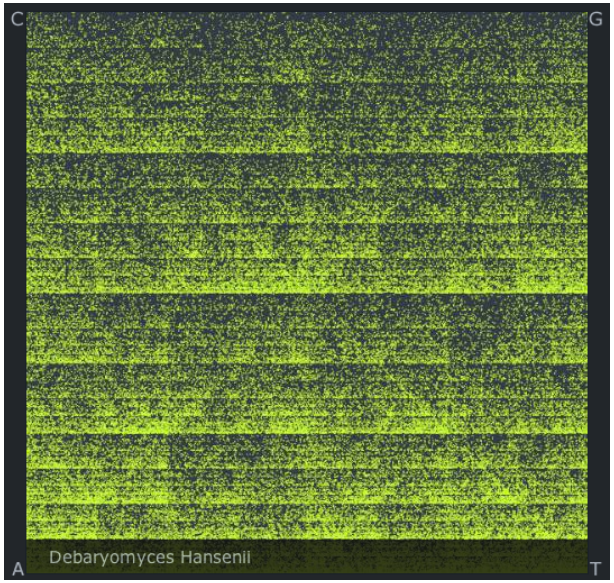
DNA Chaos Plot

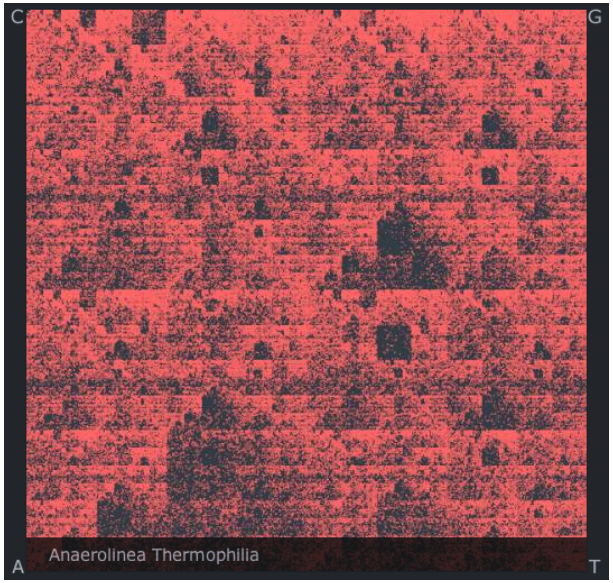


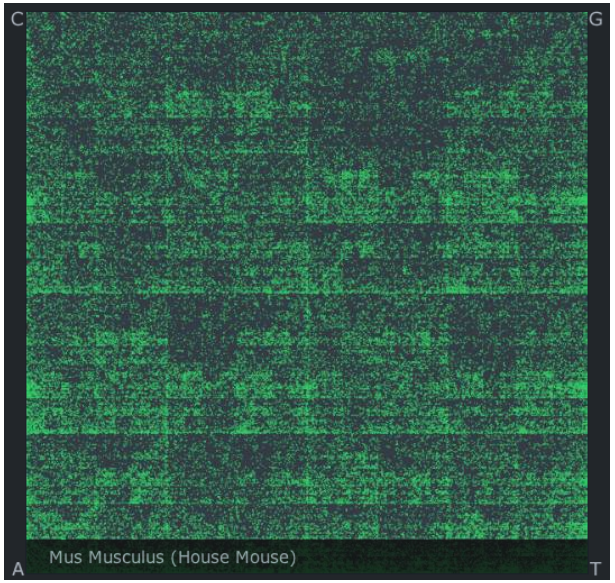




Vinga et al, *Algorithms for Molecular Biology,* 2012







Shannon Entropy

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

$$X = \{x_1, x_2, \ldots, x_n\}$$
$$p_i = P(x_i)$$

DNA sequence alphabet = 4 symbols/states $H_{DNA} = -4 \cdot (1/4) \times \log_2(1/4) = 2\ bits$

Protein sequence alphabet = 20 symbols/states $H_{AA} = -20 \cdot (1/20) \times \log_2(1/20) = 4.322\ bits$

$$I(X) = \sum_{i=1}^{n} p_i \log_2(p_i / q_i)$$

$$X = \{x_1, x_2, \ldots, x_n\}$$
$$p_i = P_{observed}(x_i)$$
$$q_i = P_{expected}(x_i)$$

Complexity

$$CE_S = -\sum_{i=1}^{k} \left(\frac{n_i}{N}\right) \cdot \log_k \left(\frac{n_i}{N}\right)$$

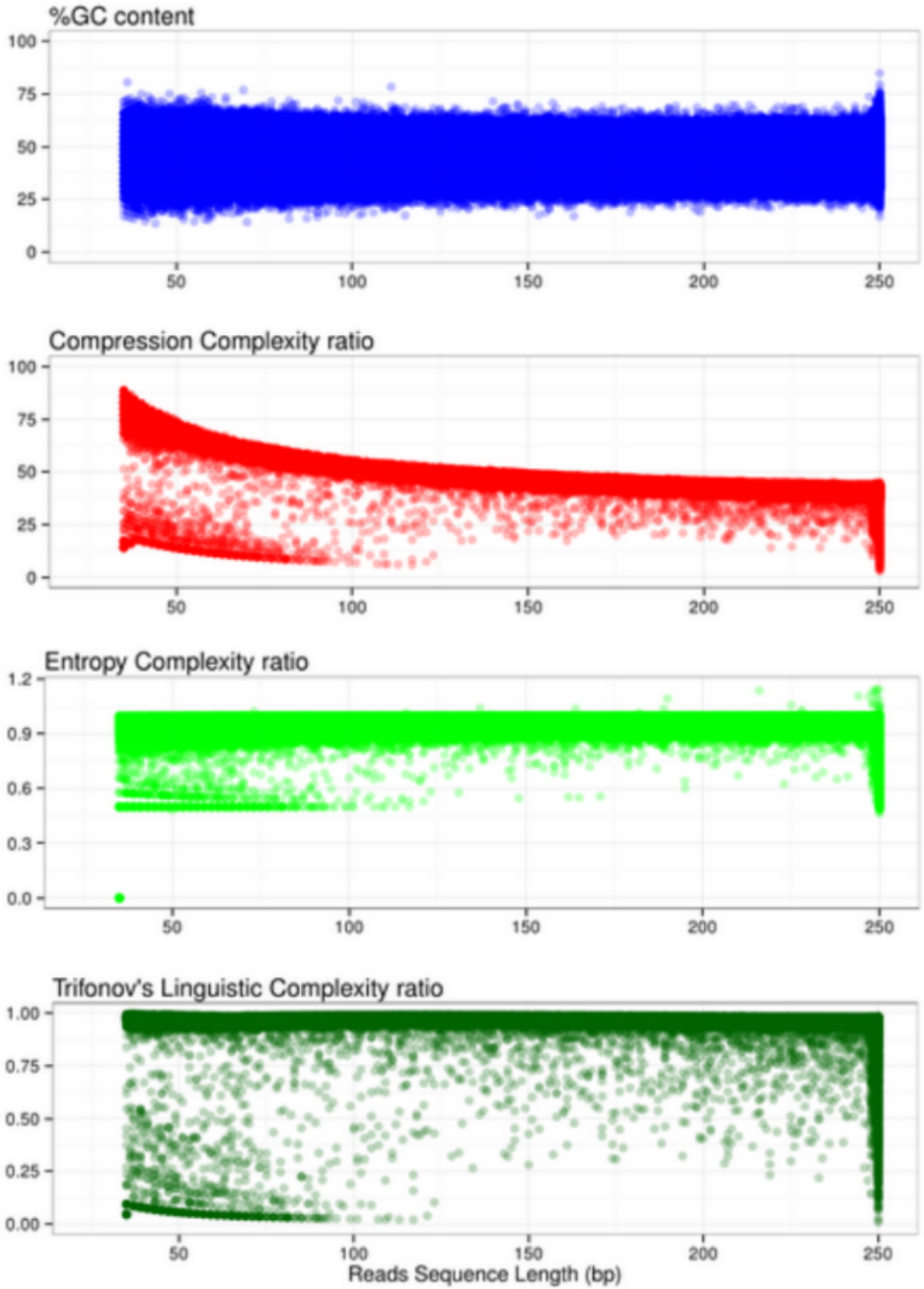
$$x_i \in \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$$
$$S = \{x_1, x_2, \ldots, x_N\}$$

Orlov i Potapov, 2004

$$CL_{(S,w)} = \prod_{j=1}^{w} U_j \quad \forall\, w \in (1,\ \omega \leqslant length(S))$$

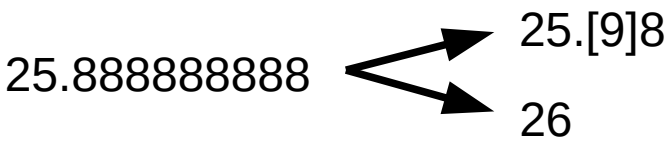
Gabrielian i Bolshoy, 1999

ACGGGAAGCTGATTCCA					
U_1	A	C	G	T	4/4
U_2	CA CC CG CT	AA AC AG AT	GA GC GG ~~GT~~	~~TA~~ TC TG TT	14/16
$CL_2 = U_1 \cdot U_2 = 4/4 \cdot 14/16 = 14/16$					



Compression Algorithms

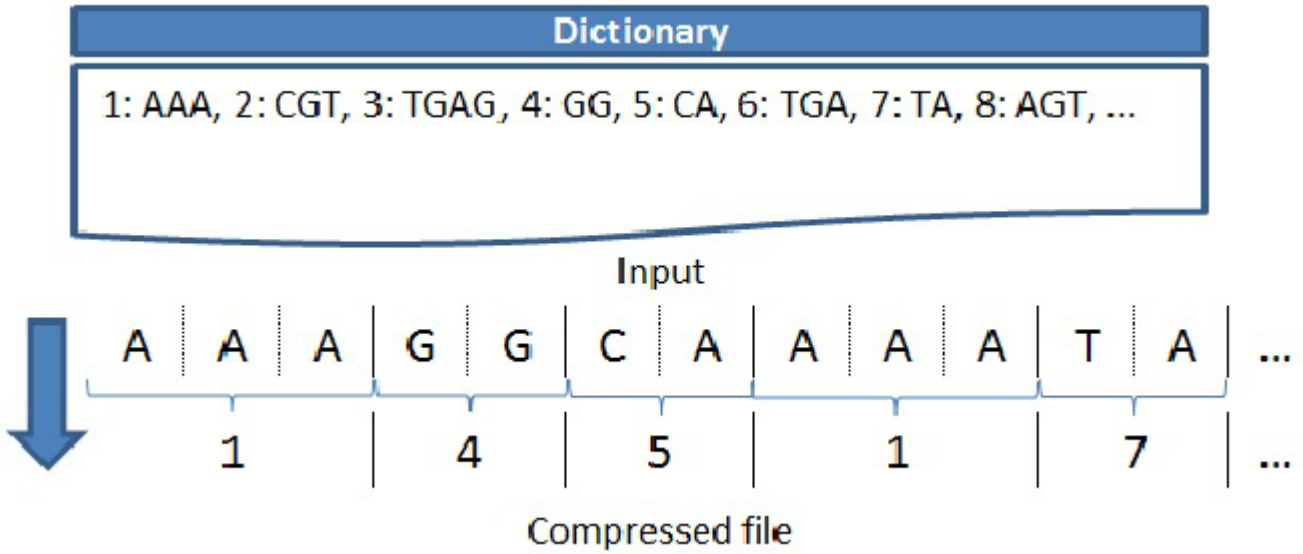
Lossless compression algorithms usually exploit statistical redundancy in such a way as to represent the sender's data more concisely without error; i.e. PNG.
Lossy compression algorithms or perceptual coding, is possible if some loss of fidelity is acceptable; i.e. JPG. Generally, a lossy data compression will be guided by research on how people perceive the data in question.

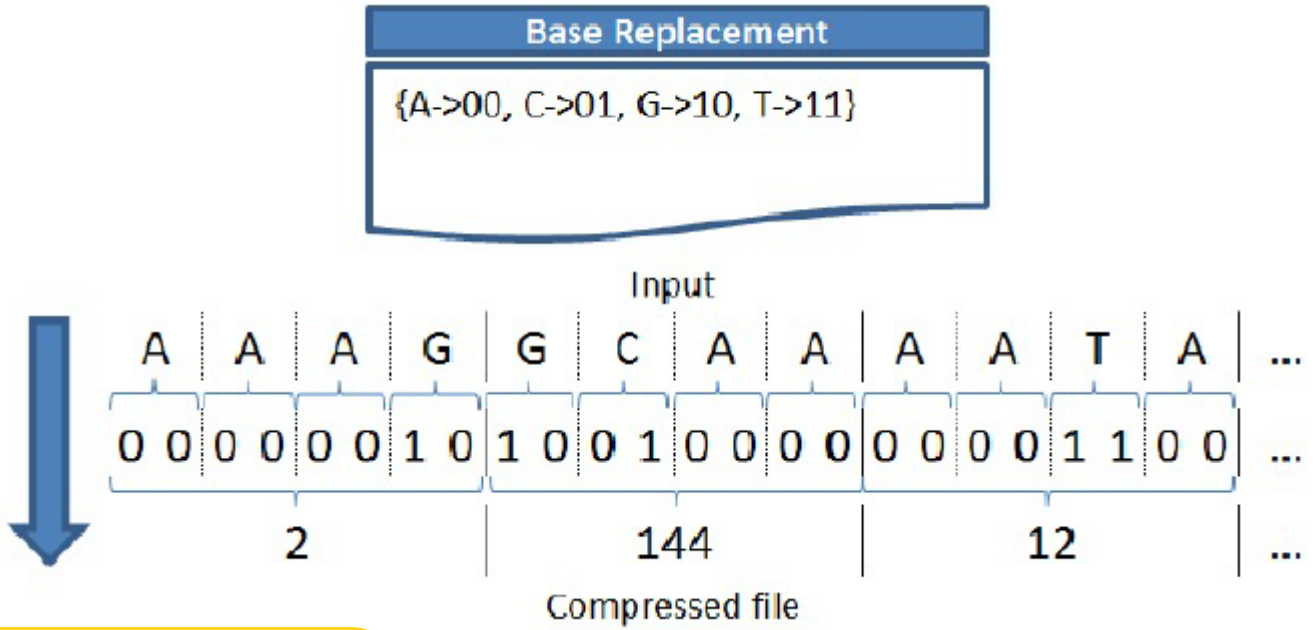


LZ77 [Lempel-Ziv 1977]

```
AGTTCGGTTCGGTTCGGTTCGGTTCGA

 vvvvv
AGTTCGGTTCGGTTCGGTTCGGTTCGA
      ^^^^^

AGTTCGGTTCG ──► AGTTCG[D=5,L=5]

AGTTCG[D=5,L=25]A
```

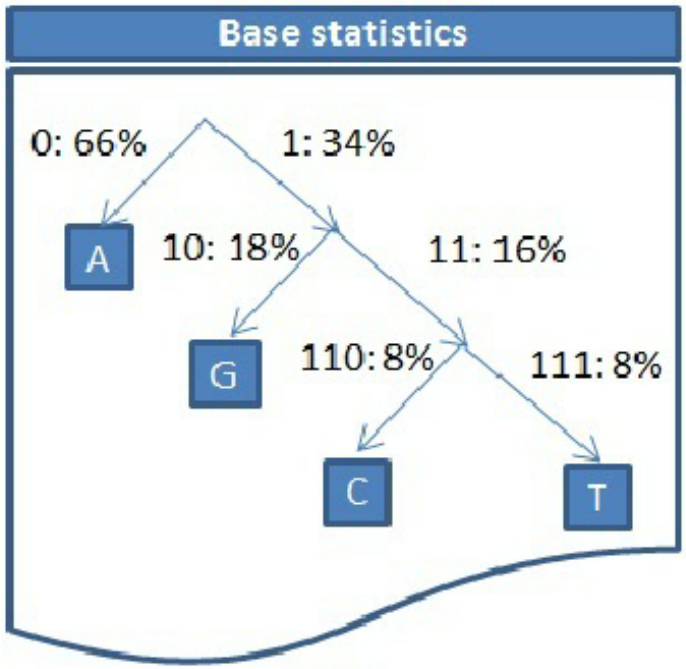
Different Sequence Compression Approaches

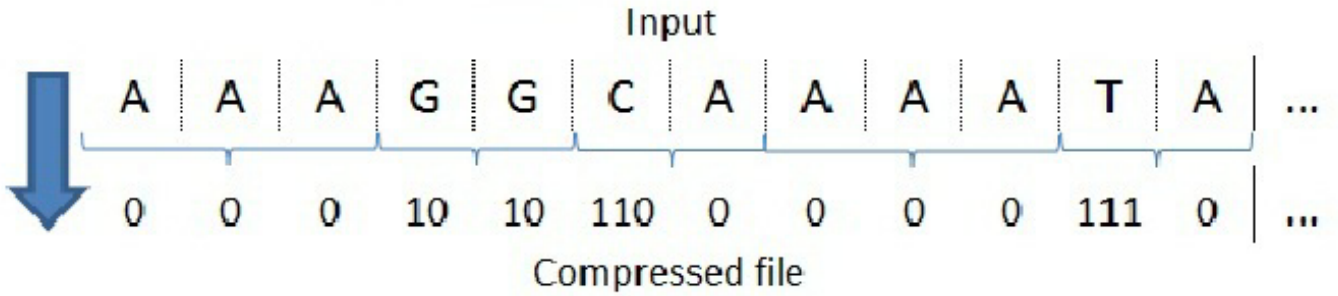


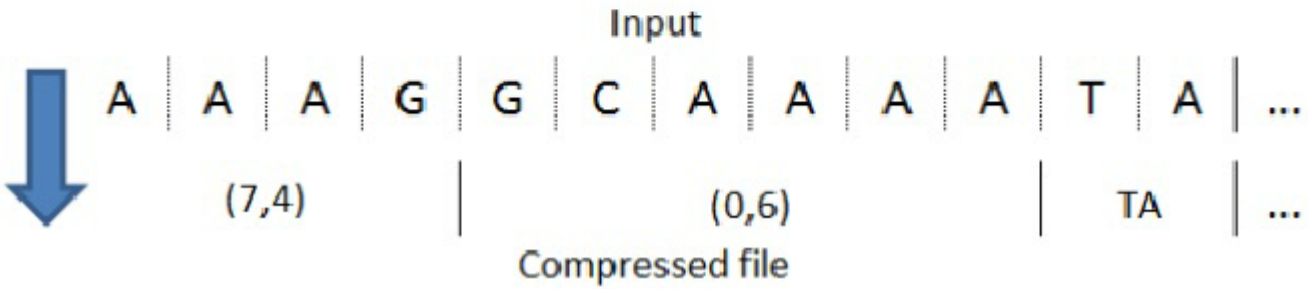


Wandelt, Bux & Leser, *Current Bioinformatics,* 2014

Different Sequence Compression Approaches







Wandelt, Bux & Leser, *Current Bioinformatics,* 2014

Compressive genomics

Po-Ru Loh, Michael Baym & Bonnie Berger

Nature Biotechnology 30(7): 627-630, 2012

Algorithms that compute directly on compressed genomic data allow analyses to keep pace with data generation.

In the past two decades, genomic sequencing capabilities have increased exponentially[1–3], outstripping advances in computing power[4–8]. Extracting new insights from the data sets currently being generated will require not only faster computers, but also smarter algorithms. However, most genomes currently sequenced are highly similar to ones already collected[9]; thus, the amount of new sequence information is growing much more slowly.

Here we show that this redundancy can be exploited by compressing sequence data in such a way as to allow direct computation on the compressed data using methods we term 'compressive' algorithms. This approach reduces the task of computing on many similar genomes to only slightly more than that of operating on just one. Moreover, its relative advantage over existing algorithms will grow with the accumulation of genomic data. We demonstrate this approach by implementing compressive versions of both the Basic Local Alignment Search Tool (BLAST)[10] and the BLAST-Like Alignment Tool (BLAT)[11], and we emphasize how compressive genomics will enable biologists to keep pace with current data.

(a 10-year, $400-million effort[1,2]), technologies[3] have been developed that can be used to sequence a human genome in 1 week for less than $10,000, and the 1000 Genomes Project is well on its way to building a library of over 2,500 human genomes[8].

These leaps in sequencing technology promise to enable corresponding advances in biology and medicine, but this will require more efficient ways to store, access and analyze large genomic data sets. Indeed, the scientific community is becoming aware of the fundamental challenges in analyzing such data[4–7]. Difficulties with large data sets arise in settings in which one analyzes genomic sequence libraries, including finding sequences similar to a given query (e.g., from environmental or medical samples) or finding signatures of selection in large sets of closely related genomes.

Currently, the total amount of available genomic data is increasing approximately tenfold every year, a rate much faster than Moore's Law for computational processing power (Fig. 1). Any computational analysis, such as sequence search, that runs on the full genomic library—or even a constant fraction

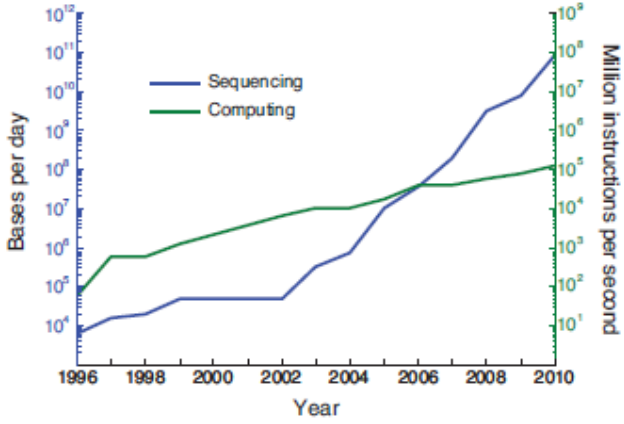


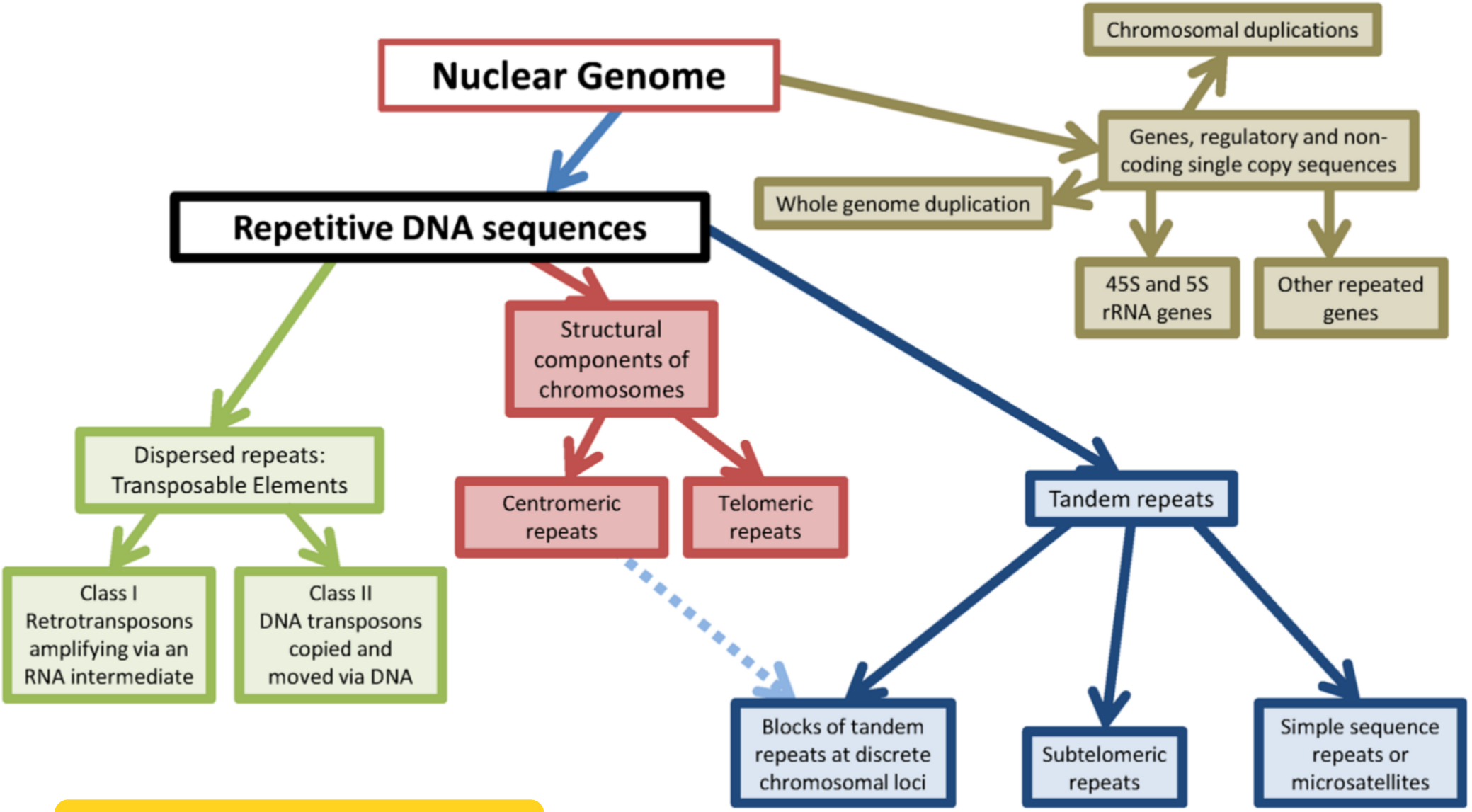
Figure 1 Sequencing capabilities versus computational power from 1996–2010. Sequencing capabilities are doubling approximately every four months, whereas processing capabilities are doubling approximately every eighteen. (Data adapted with permission from Kahn[4].)

analysis on the other. We note that although efficient algorithms, such as BLAST[10], have been developed for individual genomes, large genomic libraries have additional structure: they are highly redundant. For example, as human genomes differ on average by only 0.1% (ref. 2), 1,000 human genomes contain less than twice the unique information

DNA Sequence Periodicities

Period (bp)	Meaning	References
3	Protein-coding genes	Jimenez-Montano *et al.* (2002)
5–6	Telomeric/subtelomeric repeats	Kim and Wu (1997)
10–11	DNA bendability (helical repeat structure)	Fukushima *et al.* (2001)
		Herzel *et al.* (1999)
48–50	Centromeric repeats	Guy *et al.* (2003)
68	β satellite DNA	Waye and Willard (1989)
102	Nucleosomal structure in eukaryotes	Holste *et al.* (2003)
105–106	Isochores (DNA regions with low G + C content)	Buldyrev *et al.* (1995)
~135	Dimeric *Alu* repeat structure	Holste *et al.* (2003)
~165	Homopolymeric A-rich sequences within *Alu* repeats	Holste *et al.* (2003)
171	α satellite DNA	Haaf and Willard (1997)
~300	*Alu* repeats	Holste *et al.* (2003)
~680	DNA bend sites	Wada-Kiyama and Kiyama (1996)

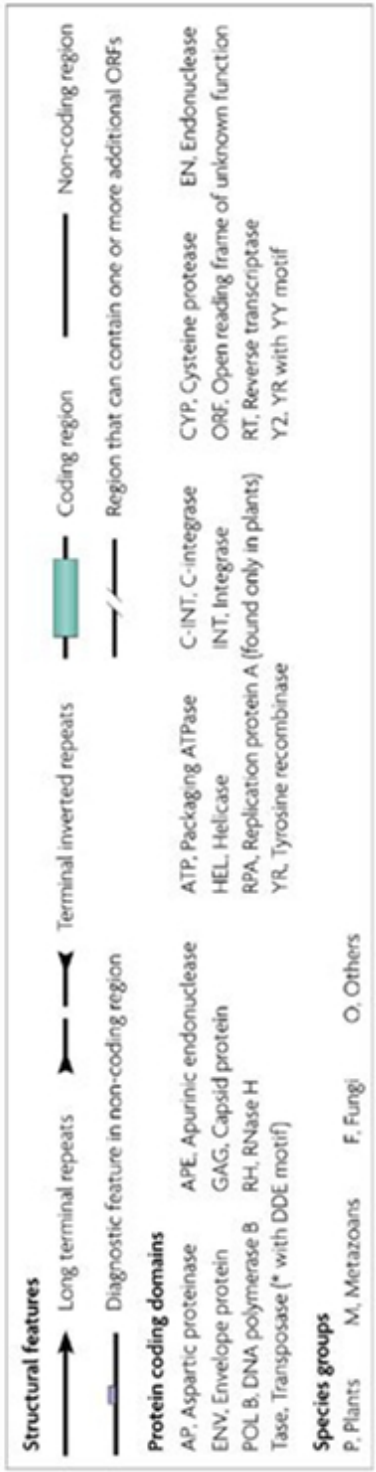
Repetitive Sequences



Biscotti et al, *Chromosome Research*, 2015

Transposable Elements

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O



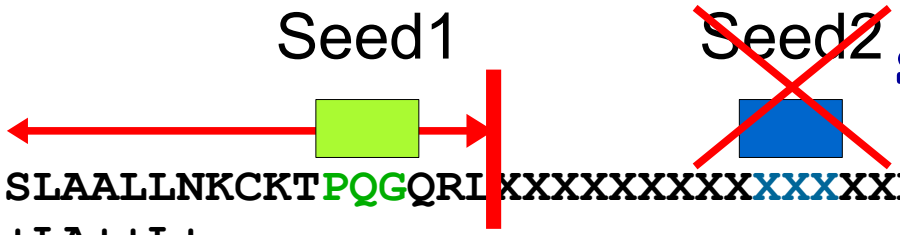
Wicker et al, *Nat Rev Genetics*, 2007

Masking Repetitive Sequences

- **RepeatMasker:** `http://www.repeatmasker.org/`
  - Uses a previously compiled library of repeat families.
  - Users can configure an external sequence search program
  - Computationally intensive, yet web-site also provides pre-masked genomic data for many completed genomes altogether with their statistical characterization.
- **De novo identification and classification**
  - **RECON**: `http://www.genetics.wustl.edu/eddy/recon`
  - **RepeatGluer**: `http://nbcr.sdsc.edu/euler/`
  - **PILER**: `http://www.drive5.com/piler`
- **Repeat databases**
  - **RepBase**: `http://www.girinst.org/repbase/index.html`
  - Plant Repeats: `https://plantrepeats.plantbiology.msu.edu/`
- **Masking Low complexity sequences** (NCBI):
  - **SEG**, **PSEG**: amino acid sequences.
  - **DUST**, **XNU**: DNA sequences.

BLAST Algorithm: Extension


```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA   365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA   330
```



STANDARD "N"-MASKING

```
Query:  325 SLAALLNKCKTPQGQRLXXXXXXXXXXXXXXXXXXXXXXXXX   365
            +LA++L+
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA   330
```

LOWERCASE-MASKING

```
Query:  325 SLAALLNKCKTPQGQRLVnqwikqplmdknrieerlnlvea   365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA   330
```