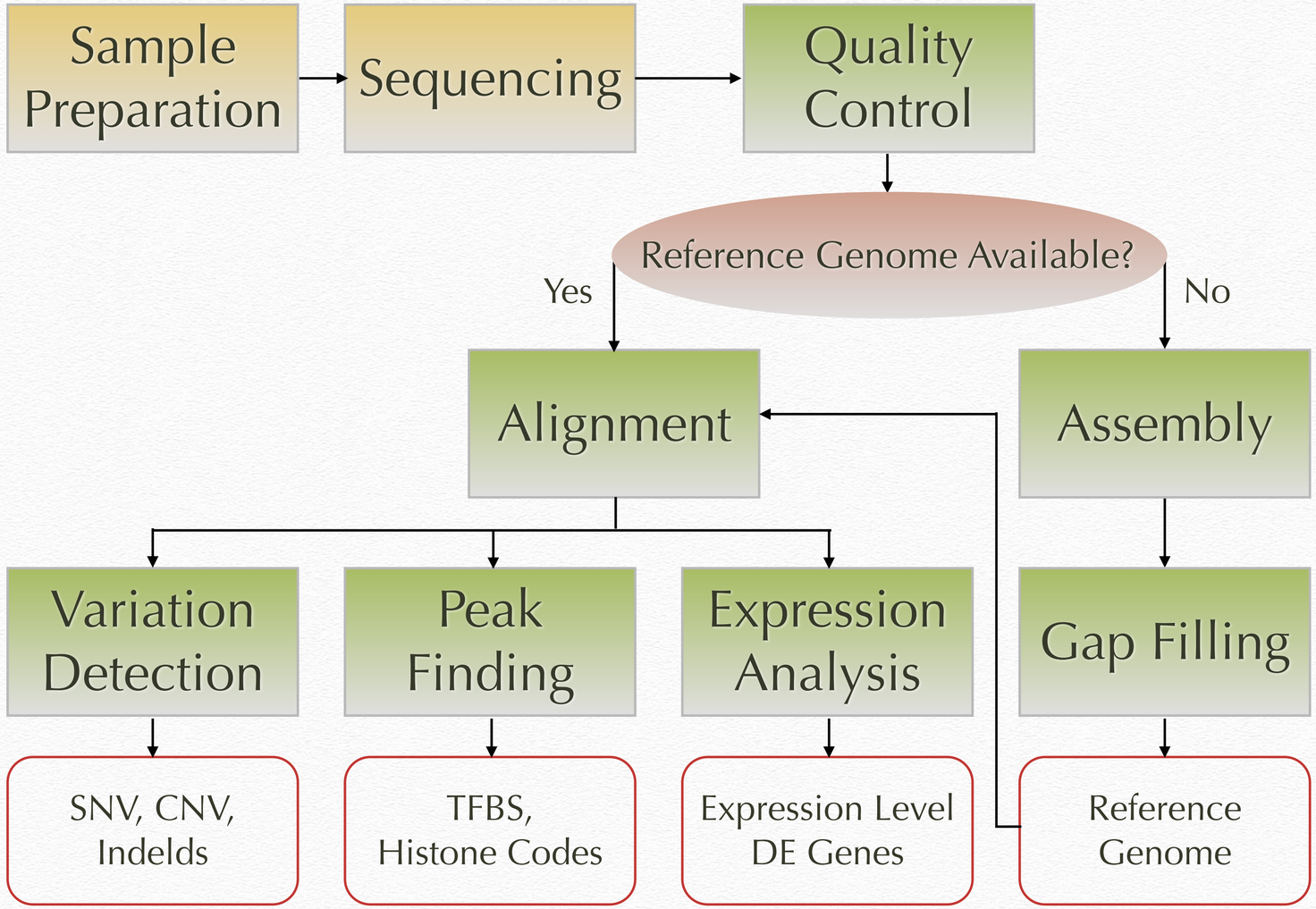
In the name of the most high

Introduction to Bioinformatics

Short Read Alignment

Ali Sharifi-Zarchi
Department of Computer Engineering, Sharif University of Technology

Analysis Pipeline

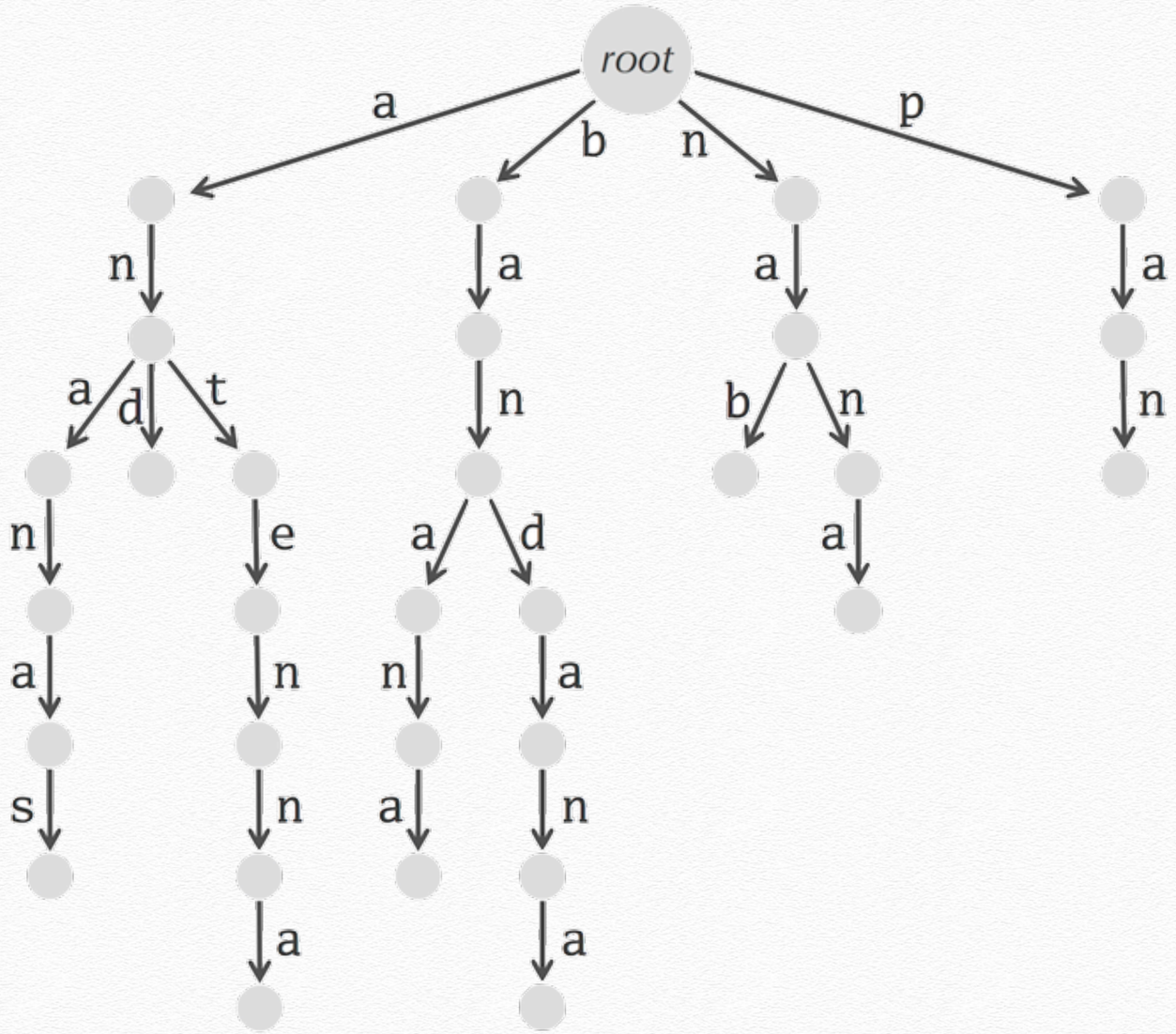


Aligning Short Reads to the Reference Genome

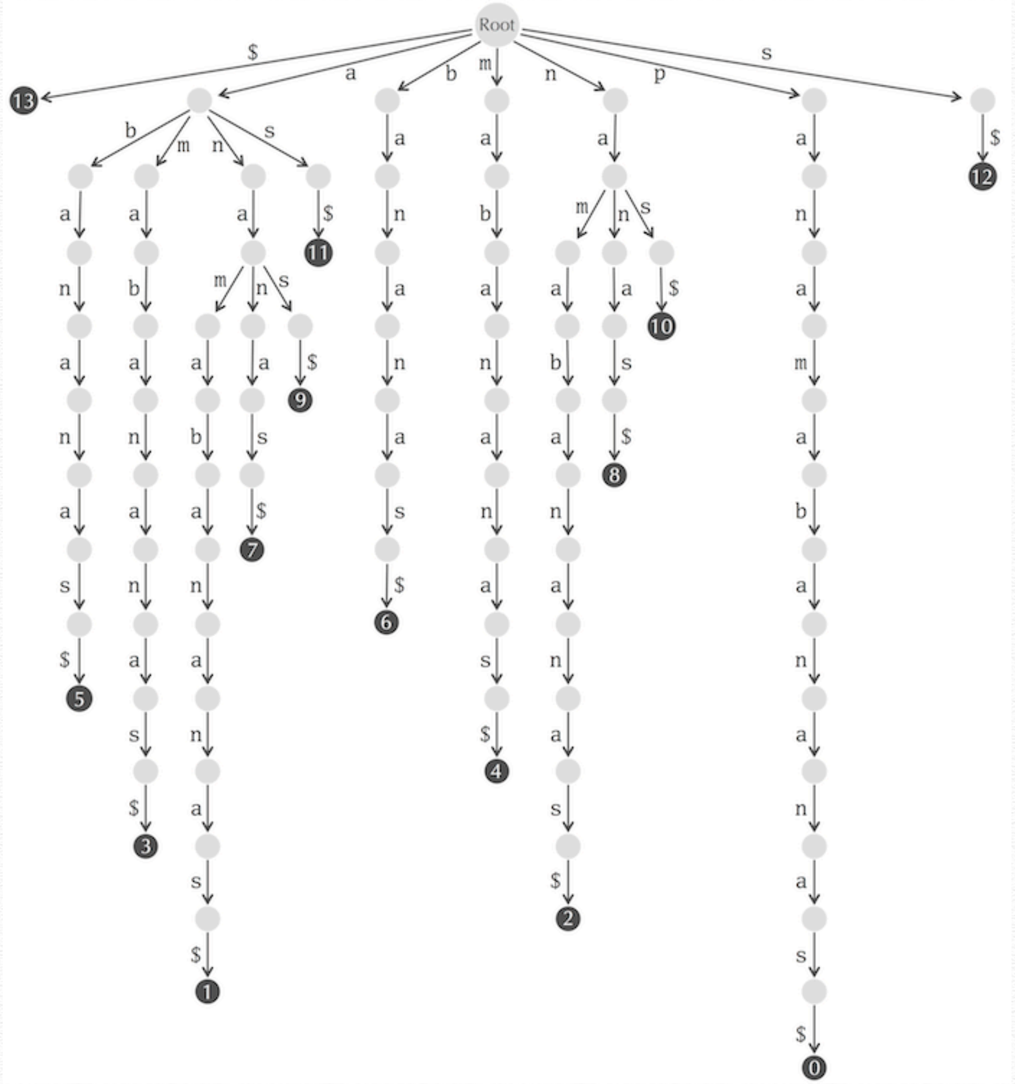
Possible Strategies

- Hashing
- Trie
- Burrows Wheeler Transform

Trie



Suffix Tree



Burrows Wheeler Transform (BWT)

babaabdad$ ⟶

$babaabdad
aabdad$bab
abaabdad$b
abdad$baba
ad$babaabd
baabdad$ba
babaabdad$
bdad$babaa
d$babaabda
dad$babaab

BWA

BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Advance Access publication May 18, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The enormous amount of short reads generated by the new DNA sequencing technologies call for the development of fast and accurate read alignment programs. A first generation of hash table-based methods has been developed, including MAQ, which is accurate, feature rich and fast enough to align short reads from a single individual. However, MAQ does not support gapped alignment for single-end reads, which makes it unsuitable for alignment of longer reads where indels may occur frequently. The speed of MAQ is also a concern when the alignment is scaled up to the resequencing of hundreds of individuals.

Results: We implemented Burrows-Wheeler Alignment tool (BWA), a new read alignment package that is based on backward search with Burrows–Wheeler Transform (BWT), to efficiently align short of scanning the whole genome when few reads are aligned. The second category of software, including SOAPv1 (Li *et al.*, 2008b), PASS (Campagna *et al.*, 2009), MOM (Eaves and Gao, 2009), ProbeMatch (Jung Kim *et al.*, 2009), NovoAlign (http://www.novocraft.com), ReSEQ (http://code.google.com/p/re-seq), Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) and BFAST (http://genome.ucla.edu/bfast), hash the genome. These programs can be easily parallelized with multi-threading, but they usually require large memory to build an index for the human genome. In addition, the iterative strategy frequently introduced by these software may make their speed sensitive to the sequencing error rate. The third category includes slider (Malhis *et al.*, 2009) which does alignment by merge-sorting the reference subsequences and read sequences.

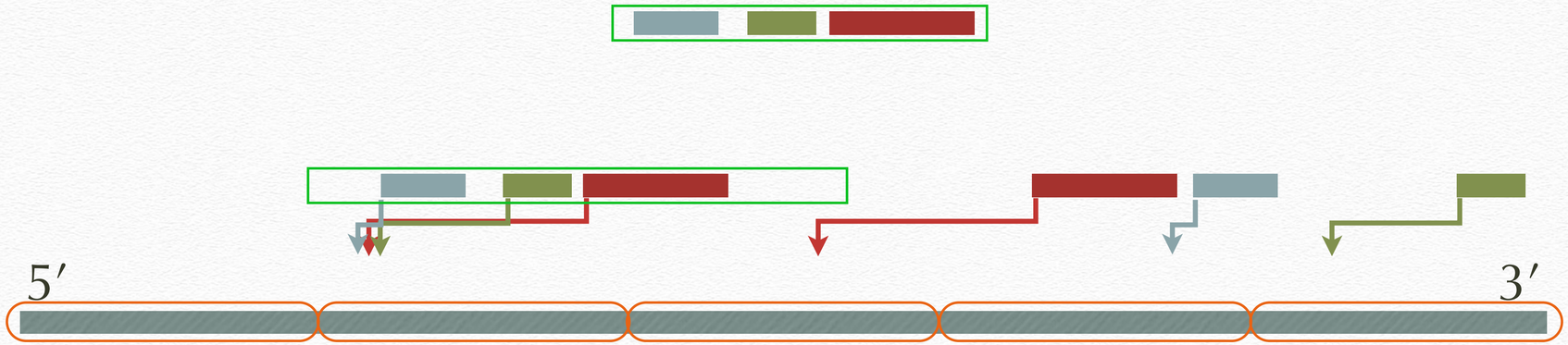
Bowtie


Exact
266, 266
g
278, 290
t
1, 104
a
Inexact
88, 88
177, 177
266, 266
396, 396
g
1, 30
104, 124
184, 184
278, 290
t
1, 104
104, 184
184, 278
278, 401
a
60, 60
145, 145
203, 203
390, 390
g
1, 30
104, 124
184, 184
278, 290
c
1, 104
104, 184
184, 278
278, 401
a
51, 51
140, 140
184, 184
278, 278
g
1, 30
104, 124
184, 184
278, 290
a
1, 104
104, 184
184, 278
278, 401
a
75, 75
160, 160
261, 262
396, 396
g
102, 103
349, 349
270, 273
396, 399
g
80, 88
167, 174
240, 266
390, 396
t
1, 104
104, 184
184, 278
278, 401
g

Aryana

- Our short and long read aligner
- The same platform aligns both gDNA and Bis-seq data
- Uses BWA implementation of the Burrows-Wheeler Transform, but with completely different approach

Strategy

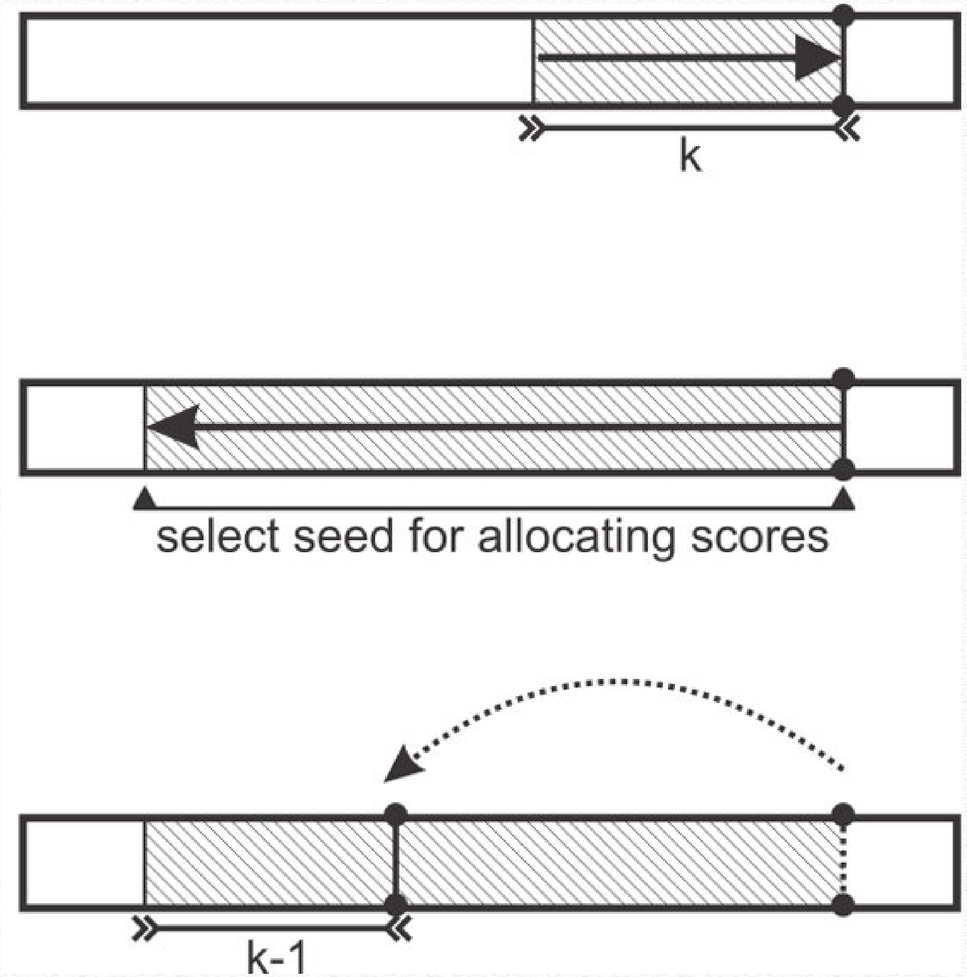


Strategy

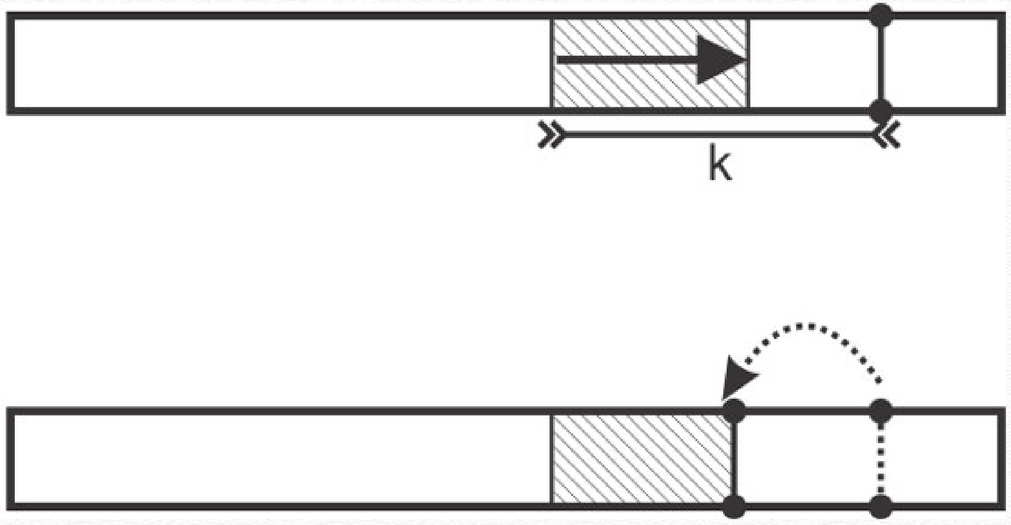
Bidirectional Matching

- The reverse-complement of the reference genome is concatenated at the end
- Hence we can extend the seed sequence in both directions

Jumping With Match



Jumping Without Match



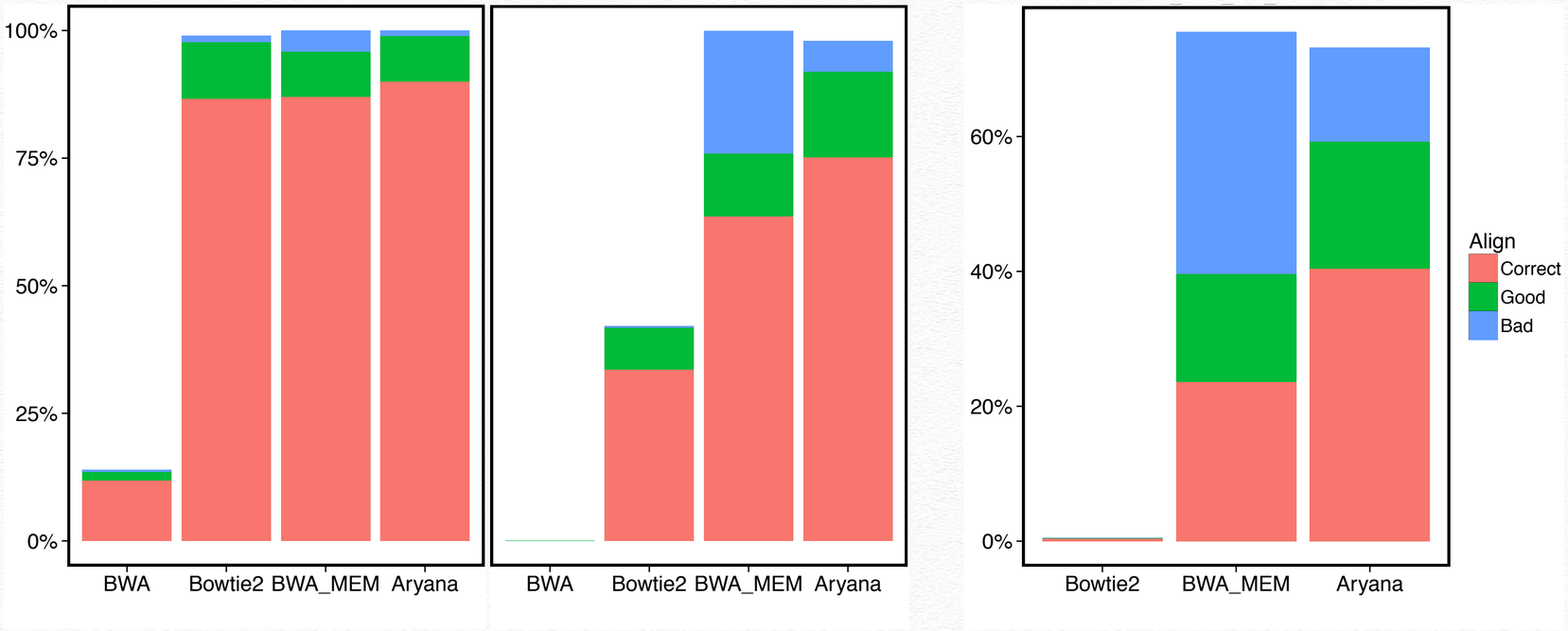
Seed Search Algorithm

Algorithm 1 extracting seeds

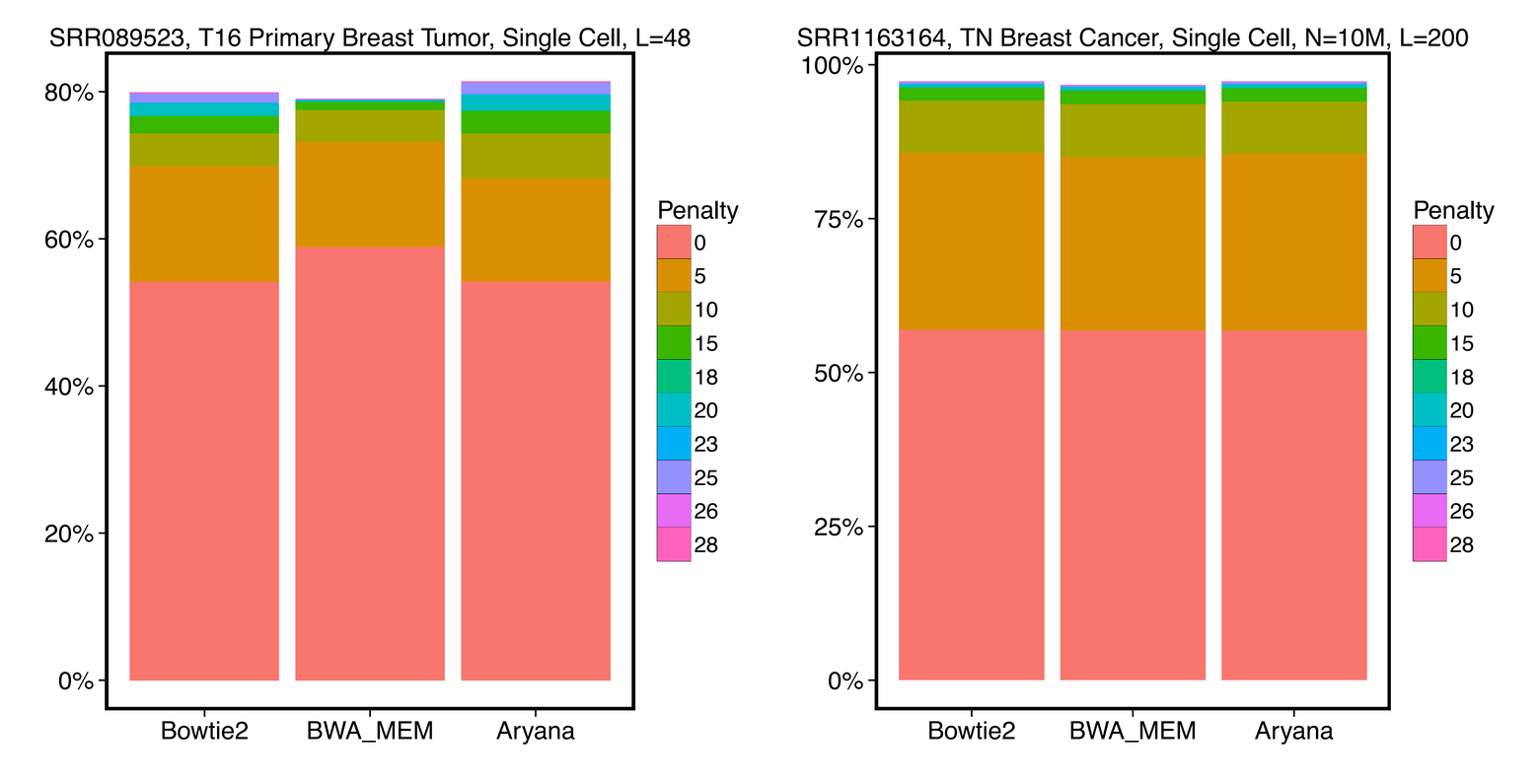
```
function MAXIMALLYSEED(seq, k)
  right ← LENGTH(seq)
  while right ≥ k do
    matched ← MATCHLEFTTORIGHT(seq, right − k + 1, k)
    if matched < k then
      right ← right − k + matched
      continue
    end if
    begin, end, matched ← MATCHRIGHTTOLEFT(seq, right, INF)
    for index from begin to end do
      pos ← BWTPOSITION(index)
       GRANTSCORE(pos − (right − matched + 1), matched)
    end for
    right ← right − matched + k − 1
  end while
end function
```

Mismatch & Indel Rates

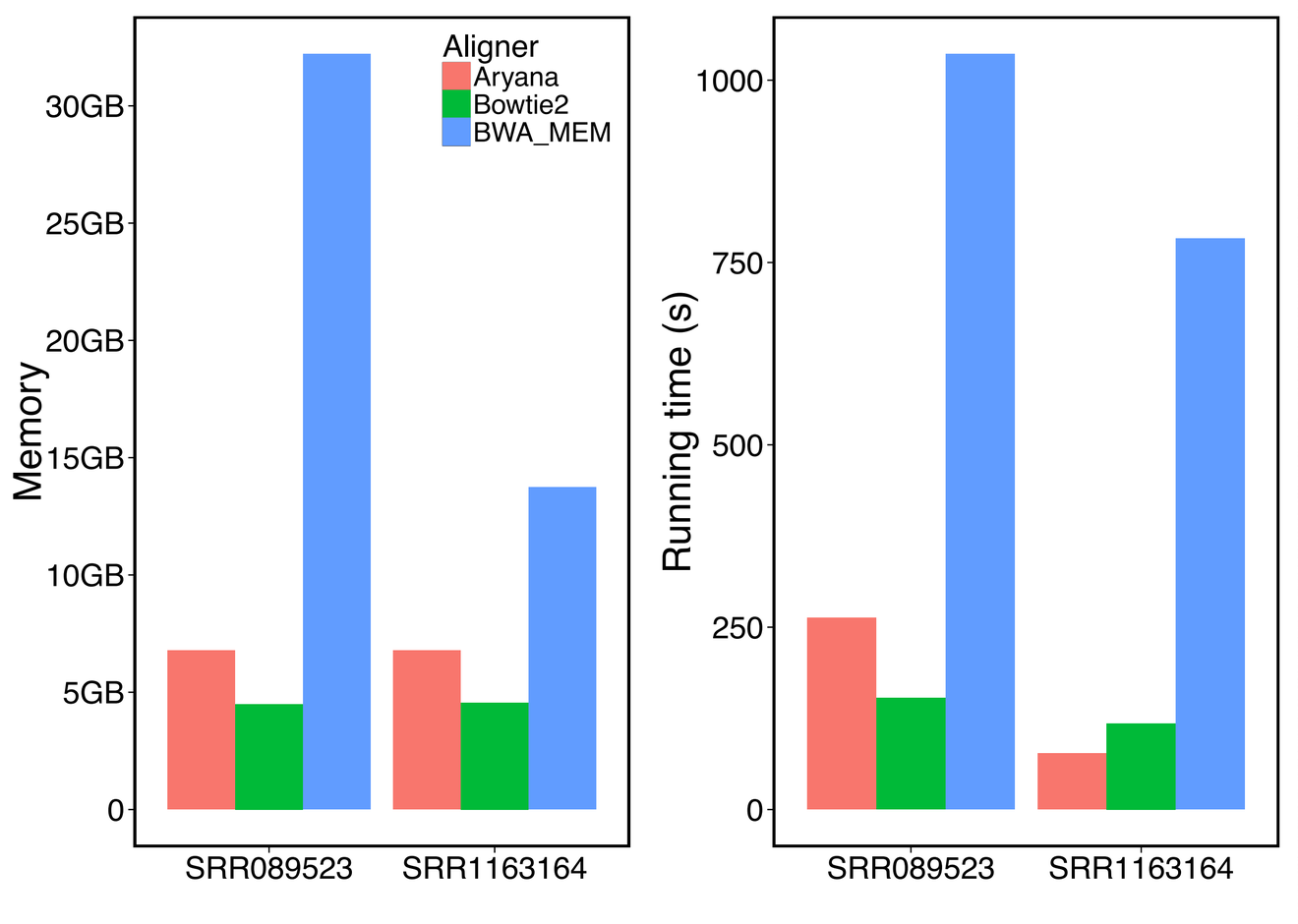
0.01 0.03 0.05



Single Cell Cancer Samples



Performance Analysis



Gholami *et al. BMC Bioinformatics* 2014, **15**(Suppl 9):S12
http://www.biomedcentral.com/1471-2105/15/S9/S12

BMC Bioinformatics

PROCEEDINGS Open Access

ARYANA: Aligning Reads by Yet Another Approach

Milad Gholami[1†], Aryan Arbabi[2†], Ali Sharifi-Zarchi[3,4], Hamidreza Chitsaz[5], Mehdi Sadeghi[6*]

From RECOMB-Seq: Fourth Annual RECOMB Satellite Workshop on Massively Parallel Sequencing
Pittsburgh, PA, USA. 31 March - 05 April 2014

Abstract

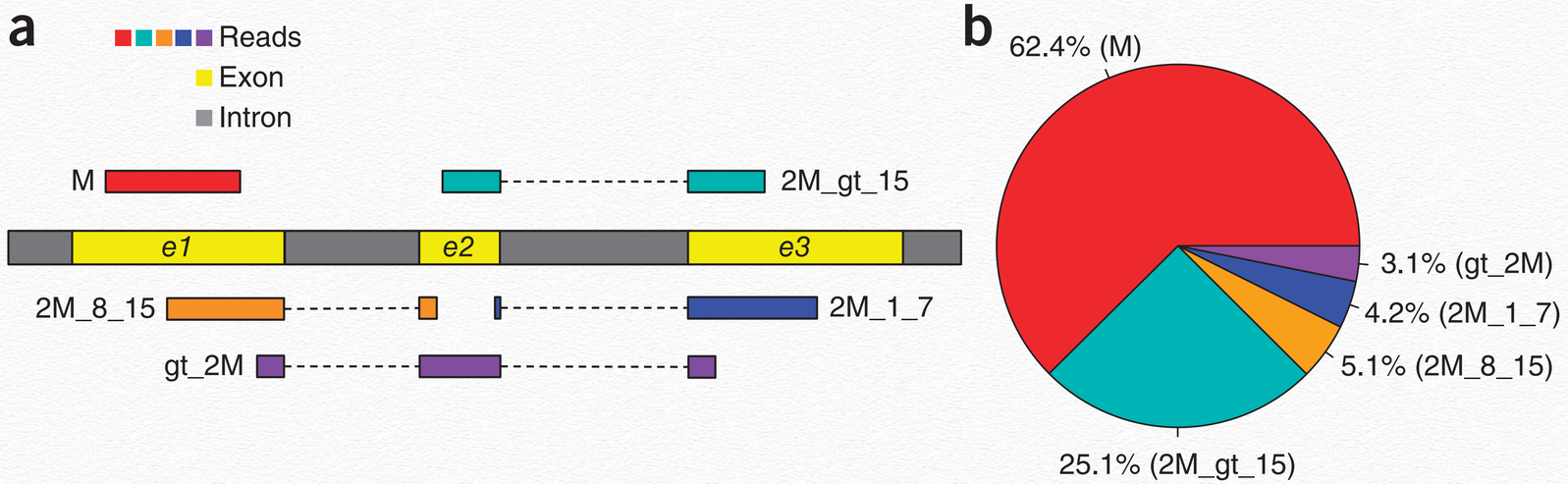
Motivation: Although there are many different algorithms and software tools for aligning sequencing reads, fast gapped sequence search is far from solved. Strong interest in fast alignment is best reflected in the 10^6 prize for the Innocentive competition on aligning a collection of reads to a given database of reference genomes. In addition, *de novo* assembly of next-generation sequencing long reads requires fast overlap-layout-concensus algorithms which depend on fast and accurate alignment.

Contribution: We introduce ARYANA, a fast gapped read aligner, developed on the base of BWA indexing infrastructure with a completely new alignment engine that makes it significantly faster than three other aligners: Bowtie2, BWA and SeqAlto, with comparable generality and accuracy. Instead of the time-consuming backtracking procedures for handling mismatches, ARYANA comes with the seed-and-extend algorithmic framework and a significantly improved efficiency by integrating novel algorithmic techniques including dynamic seed selection, bidirectional seed extension, reset-free hash tables, and gap-filling dynamic programming. As the read length increases ARYANA's superiority in terms of speed and alignment rate becomes more evident. This is in perfect harmony with the read length trend as the sequencing technologies evolve. The algorithmic platform of ARYANA makes it easy to develop mission-specific aligners for other applications using ARYANA engine.

Availability: ARYANA with complete source code can be obtained from http://github.com/aryana-aligner

RNASeq Alignment

Different Types of Reads



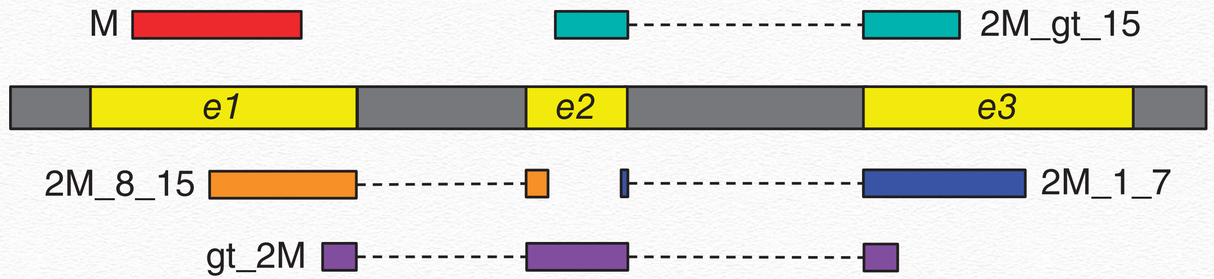
HISAT Indexing Strategy

- A global FM index for entire genome: Tophat, STAR, etc.
- The novel idea of HISAT: small FM indexes, for 64 kbp regions that collectively cover the genome
- For human genome: ~48,000 local FM indexes, each overlapping its neighbor by 1,024 bp
- Overlapping boundaries: to align reads that would otherwise span two indexes

Implementation Details

- HISAT uses the Bowtie2 implementation to construct and search an FM index
- It stores large number of local indexes in a small set of files

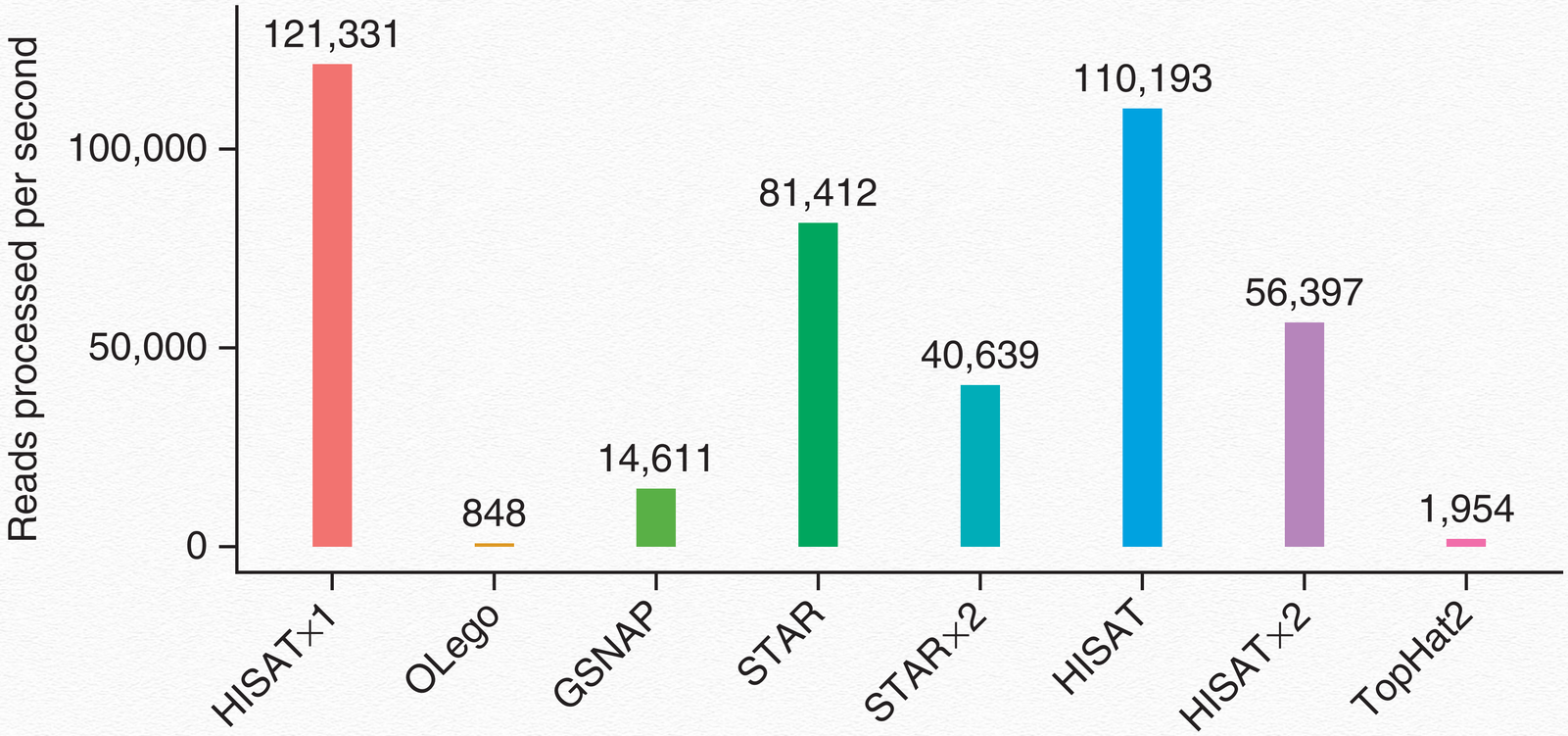
Alignment Strategies



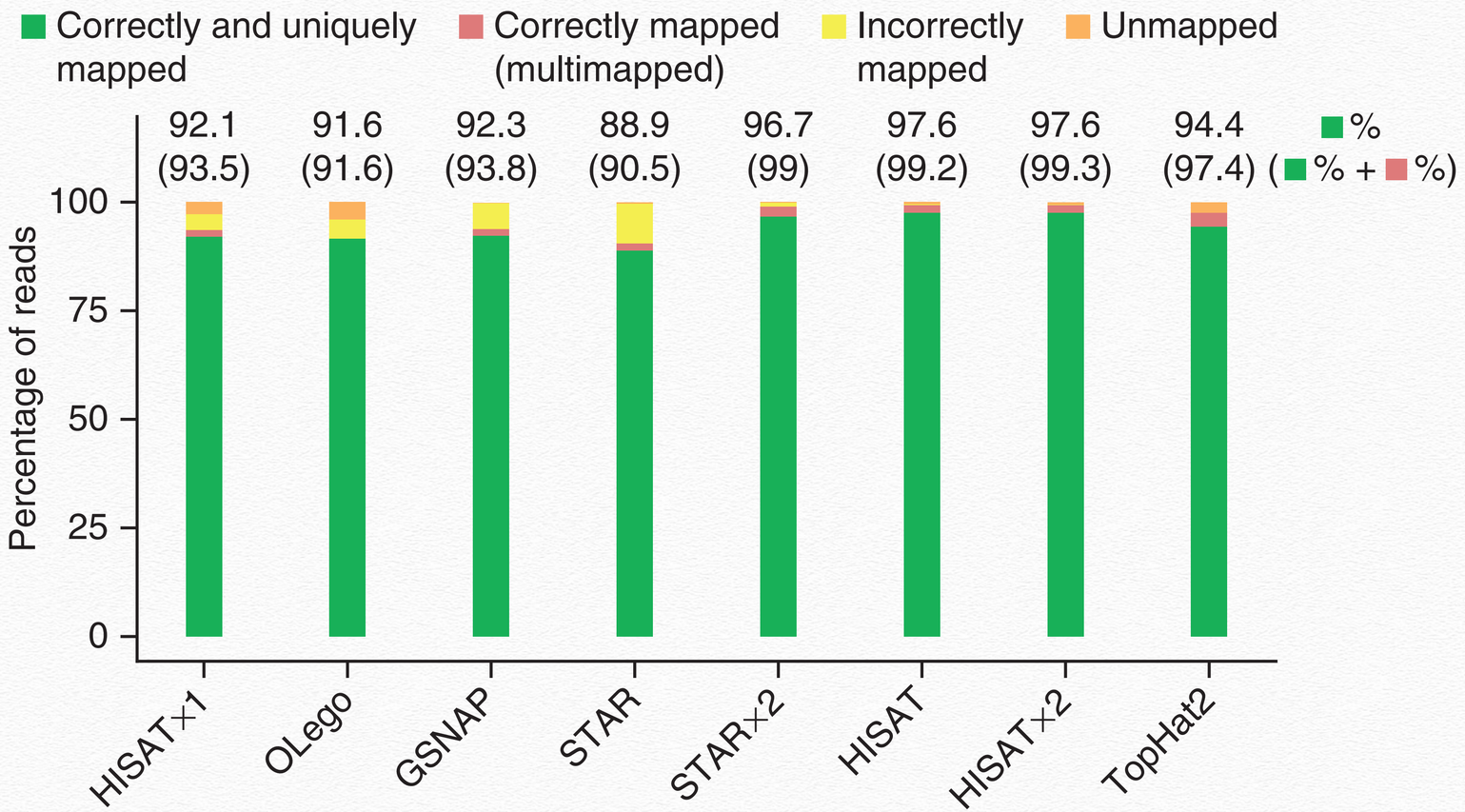
- Anchors of > 7bp: local indexes. An 8 bp anchor expected to occur ~48,000 times in human genome and once in 64 kbp region

- For anchors of 1-7 bp length: Using splice-site information found by aligning other reads (HISATx2)

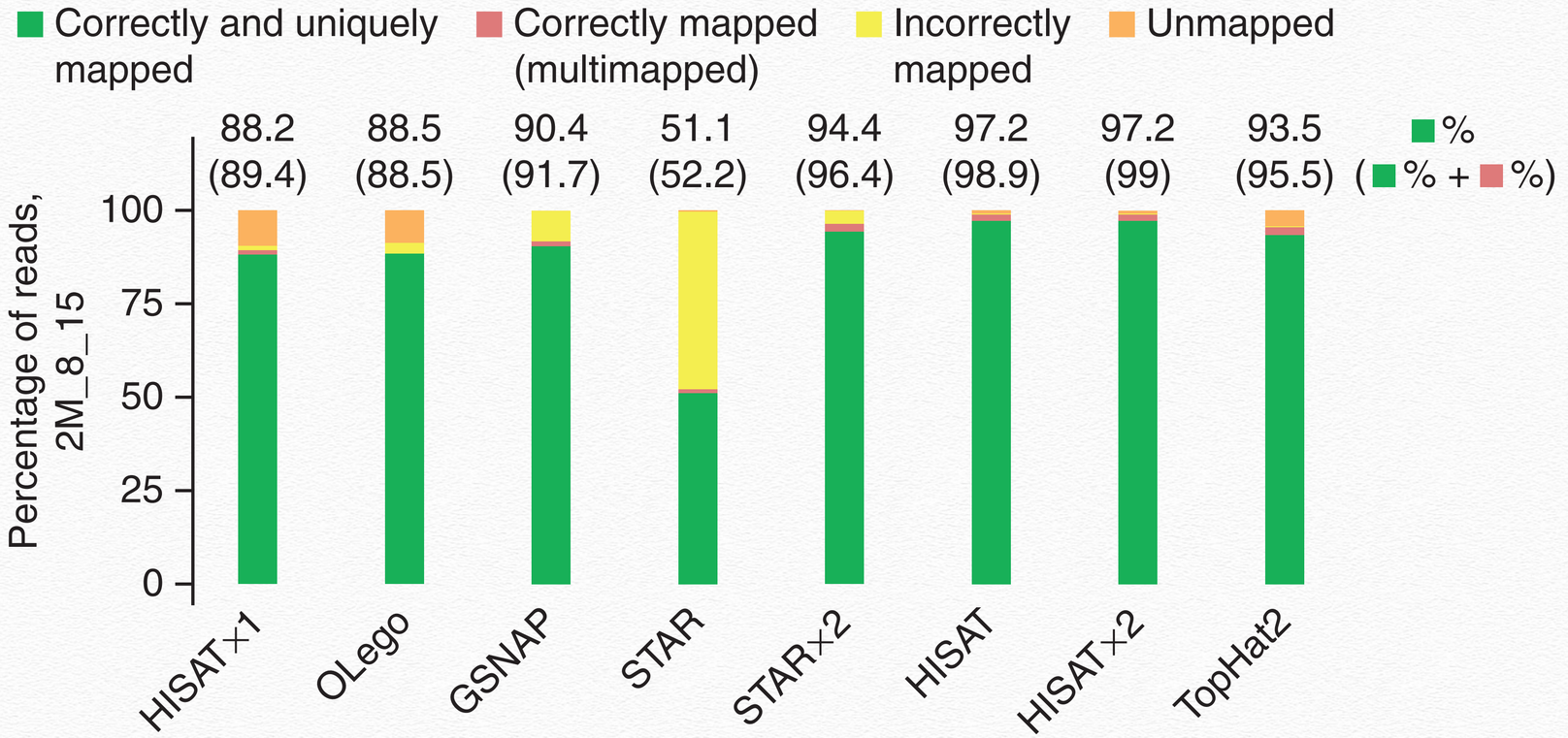
Speed Comparison



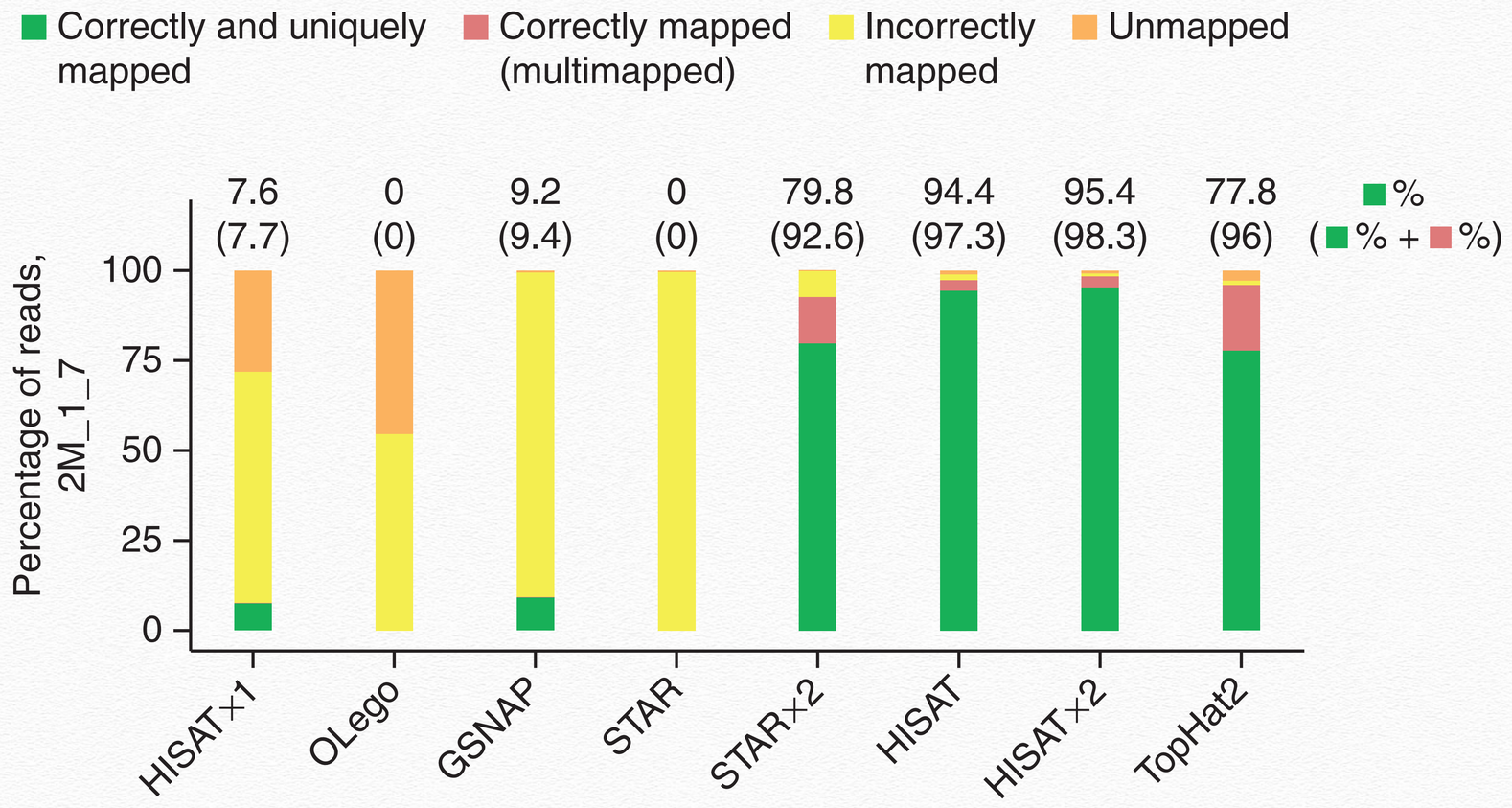
Accuracy Comparison: All Types of Reads



Accuracy Comparison: 8-15bp Anchors



Accuracy Comparison: 1-7bp Anchors



HISAT: a fast spliced aligner with low memory requirements

Daehwan Kim[1,2], Ben Langmead[1–3] & Steven L Salzberg[1–3]

HISAT (hierarchical indexing for spliced alignment of transcripts) is a highly efficient system for aligning reads from RNA sequencing experiments. HISAT uses an indexing scheme based on the Burrows-Wheeler transform and the Ferragina-Manzini (FM) index, employing two types of indexes for alignment: a whole-genome FM index to anchor each alignment and numerous local FM indexes for very rapid extensions of these alignments. HISAT's hierarchical index for the human genome contains 48,000 local FM indexes, each representing a genomic region of ~64,000 bp. Tests on real and simulated data sets showed that HISAT is the fastest system currently available, with equal or better accuracy than any other method. Despite its large number of indexes, HISAT requires only 4.3 gigabytes of memory. HISAT supports genomes of any size, including those larger than 4 billion bases.

Since its introduction in 2008, RNA-seq[1] has become ubiquitous as a tool for the study of gene expression, transcript structure and the identification of long noncoding RNAs and fusion transcripts[2–5]. As RNA-seq has matured, sequencing throughput and read lengths have increased dramatically to 100–500 million reads per run with lengths of 100 bp or longer. These large and ever-increasing data volumes necessitate fast and scalable computational analysis systems.

As a result of HISAT's greatly reduced memory requirements, users can shift these computations from special-purpose servers to a single conventional desktop computer, on which it is possible to run multiple samples at the same time. As developers of TopHat, we intend to make HISAT the core alignment engine for the next major version of that program, TopHat3. HISAT is open-source software freely available at http://www.ccb.jhu.edu/software/hisat/.

RESULTS

Design principles of HISAT

HISAT uses the Bowtie2 (ref. 14) implementation to handle many of the low-level operations required to construct and search an FM index. In contrast to most other aligners, our algorithm employs two different types of indexes: (i) a global FM index that represents the entire genome and (ii) numerous small FM indexes for regions that collectively cover the genome, where each index represents 64,000 bp. For the human genome, we create ~48,000 local FM indexes, each overlapping its neighbor by 1,024 bp, to cover the entire 3 billion bases. The overlapping boundaries make it easier to align reads that would otherwise span the regions covered by two indexes.

The program stores the large number of local indexes in a small set of files and implements other optimizations to minimize the

PROTOCOL

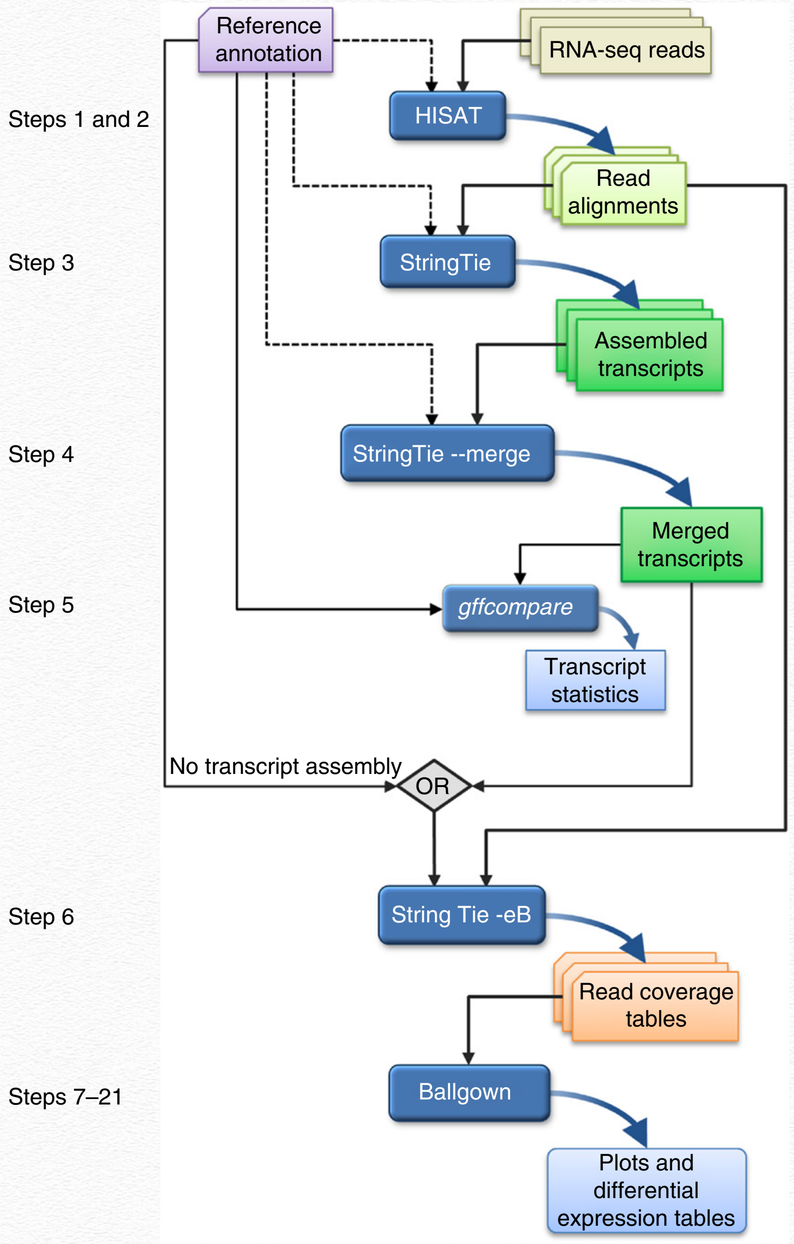
Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea[1,2], Daehwan Kim[1], Geo M Pertea[1], Jeffrey T Leek[3] & Steven L Salzberg[1–4]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [2]Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA. [3]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. Correspondence should be addressed to S.L.S. (salzberg@jhu.edu).

Published online 11 August 2016; doi:10.1038/nprot.2016.095

High-throughput sequencing of mRNA (RNA-seq) has become the standard method for measuring and comparing the levels of gene expression in a wide variety of species and conditions. RNA-seq experiments generate very large, complex data sets that demand fast, accurate and flexible software to reduce the raw read data to comprehensible results. HISAT (hierarchical indexing for spliced alignment of transcripts), StringTie and Ballgown are free, open-source software tools for comprehensive analysis of RNA-seq experiments. Together, they allow scientists to align reads to a genome, assemble transcripts including novel splice variants, compute the abundance of these transcripts in each sample and compare experiments to identify differentially expressed genes and transcripts. This protocol describes all the steps necessary to process a large set of raw sequencing reads and create lists of gene transcripts, expression levels, and differentially expressed genes and transcripts. The protocol's execution time depends on the computing resources, but it typically takes under 45 min of computer time. HISAT, StringTie and Ballgown are available from http://ccb.jhu.edu/software.shtml.

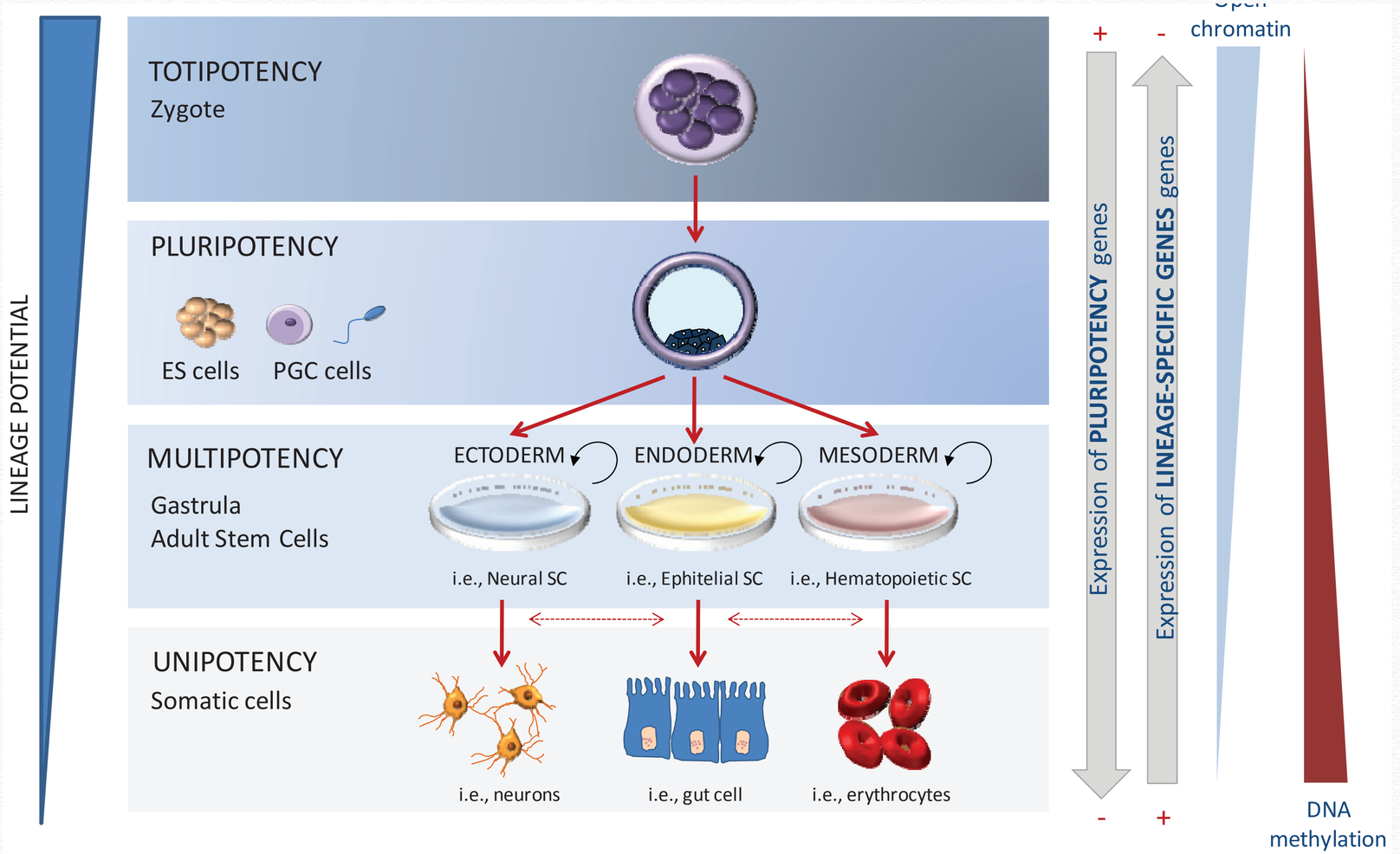

Steps 1 and 2
Step 3
Step 4
Step 5
Step 6
Steps 7–21
Reference annotation
RNA-seq reads
HISAT
Read alignments
StringTie
Assembled transcripts
StringTie --merge
Merged transcripts
gffcompare
Transcript statistics
No transcript assembly
OR
String Tie -eB
Read coverage tables
Ballgown
Plots and differential expression tables

Bisulfite-Seq Alignment

DNA Methylation

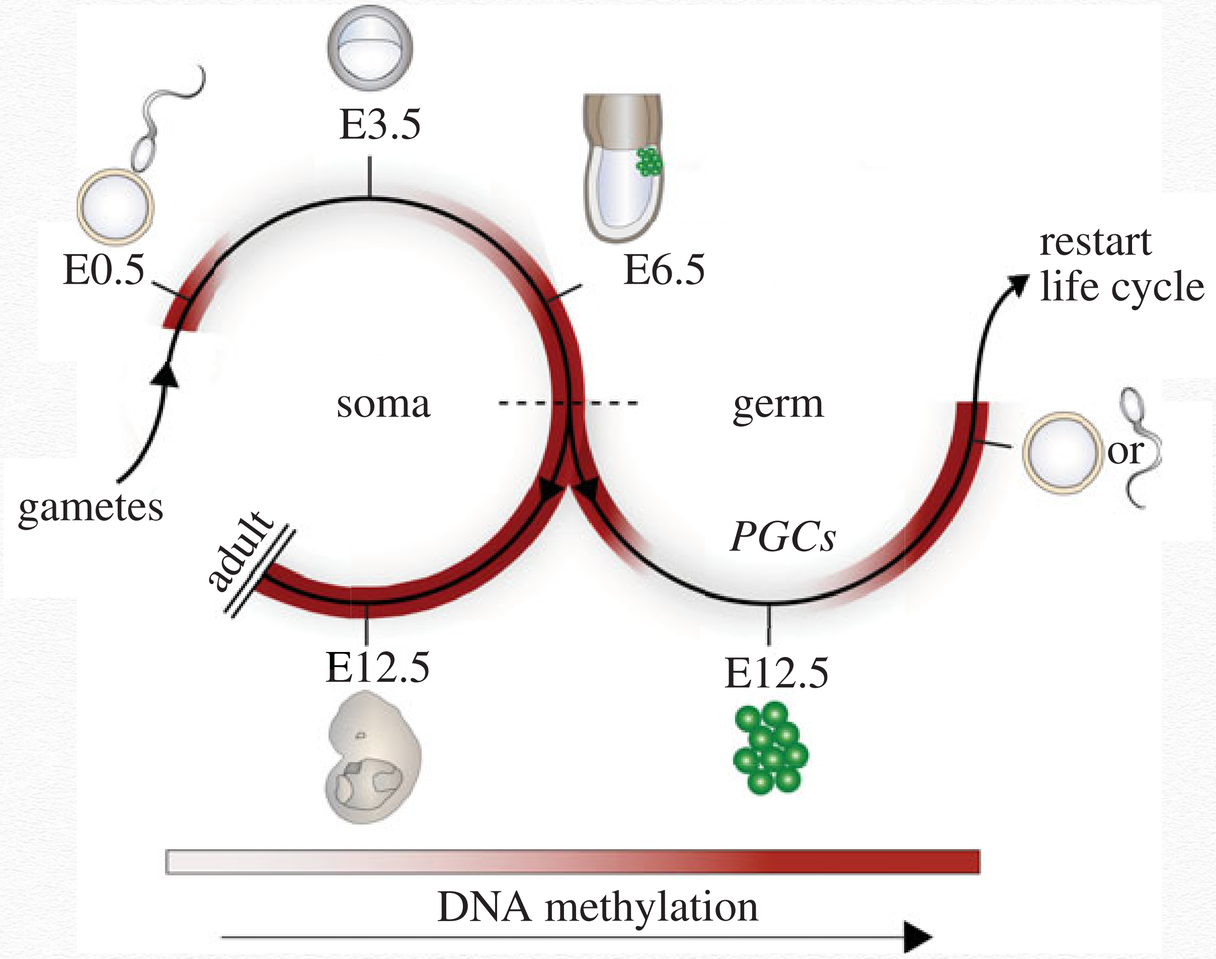
- The only epigenetic mark with a clearly identified mechanism of inheritance
- First reported in 1925 and confirmed in 1950's
- Plays a role in many biological processes including imprinting, transcription silencing, development, differentiation and inflammation
- Linked to cancer and several other diseases

DNA Methylation Changes During Differentiation



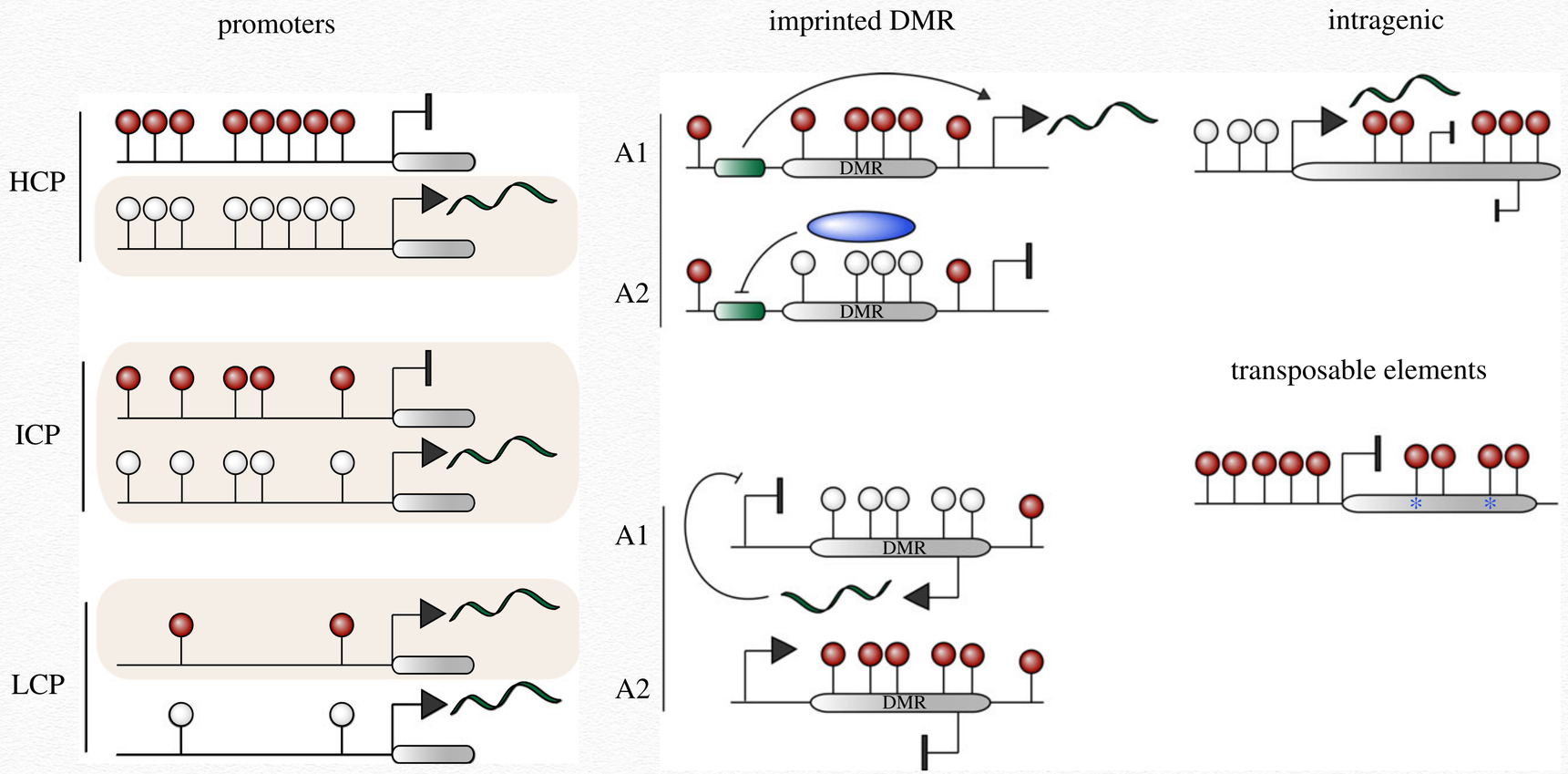
Berdasco, M., & Esteller, M. Stem Cell Research & Therapy, 2011

DNA Methylation Changes During Differentiation



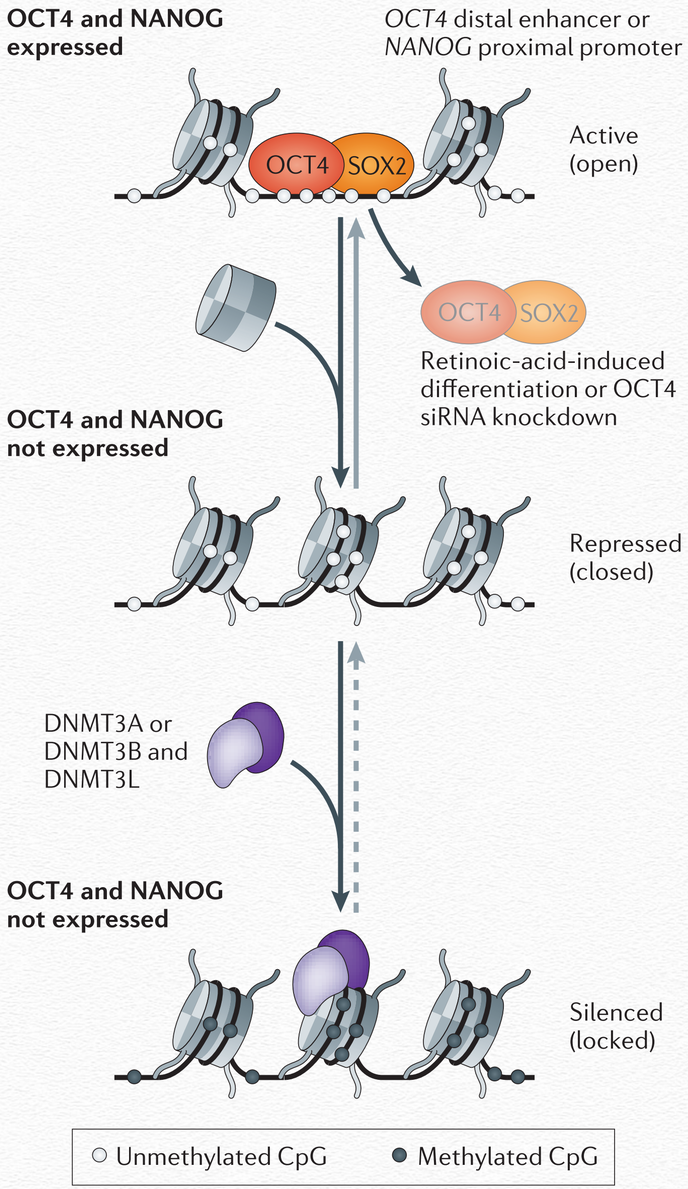
Hackett, J. A., & Surani, M. A., Phil. Trans. R. Soc., 2012

Different Roles of DNA Methylation



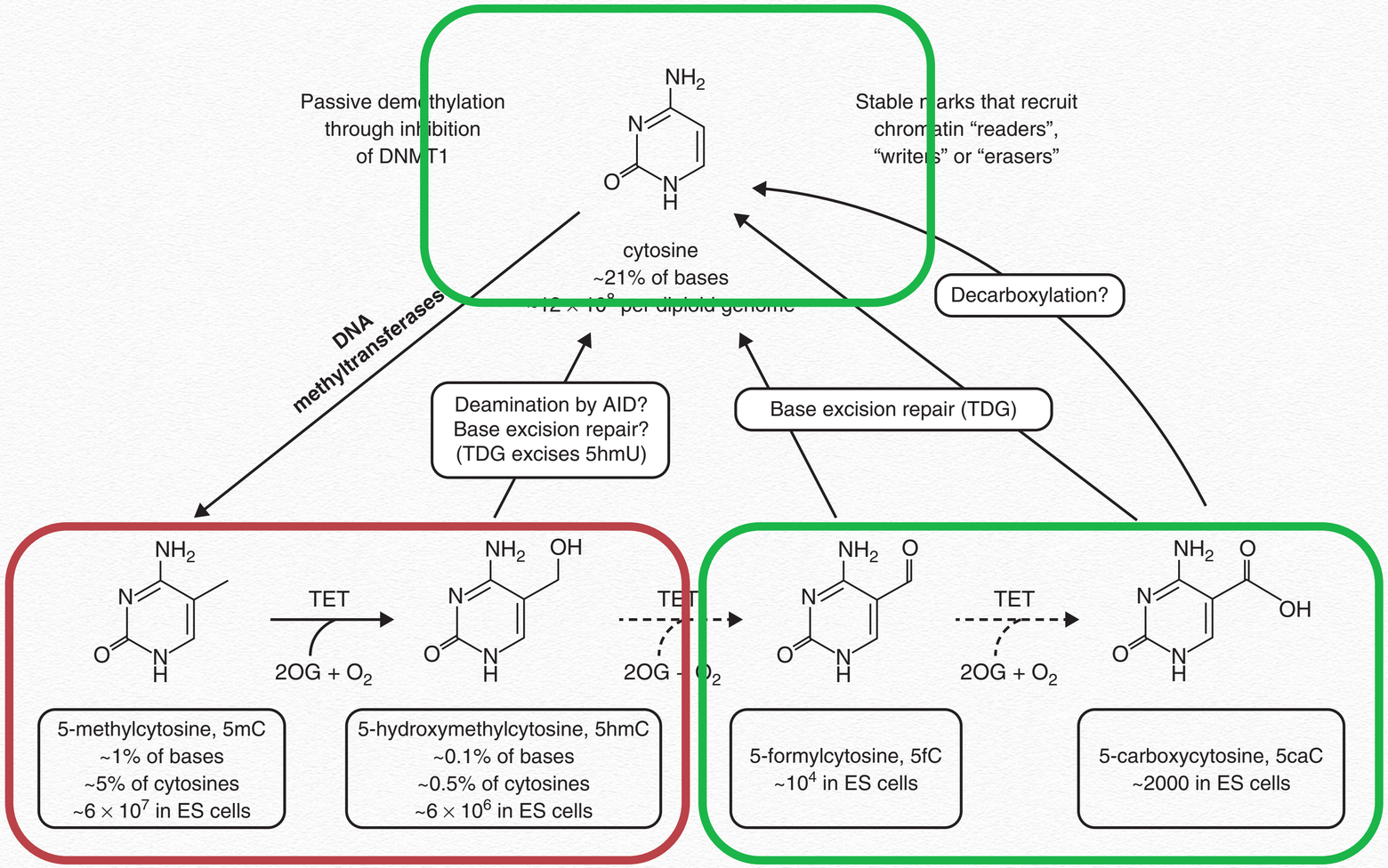
Hackett, J. A., & Surani, M. A., Phil. Trans. R. Soc., 2012

Silencing Precedes DNA Methylation



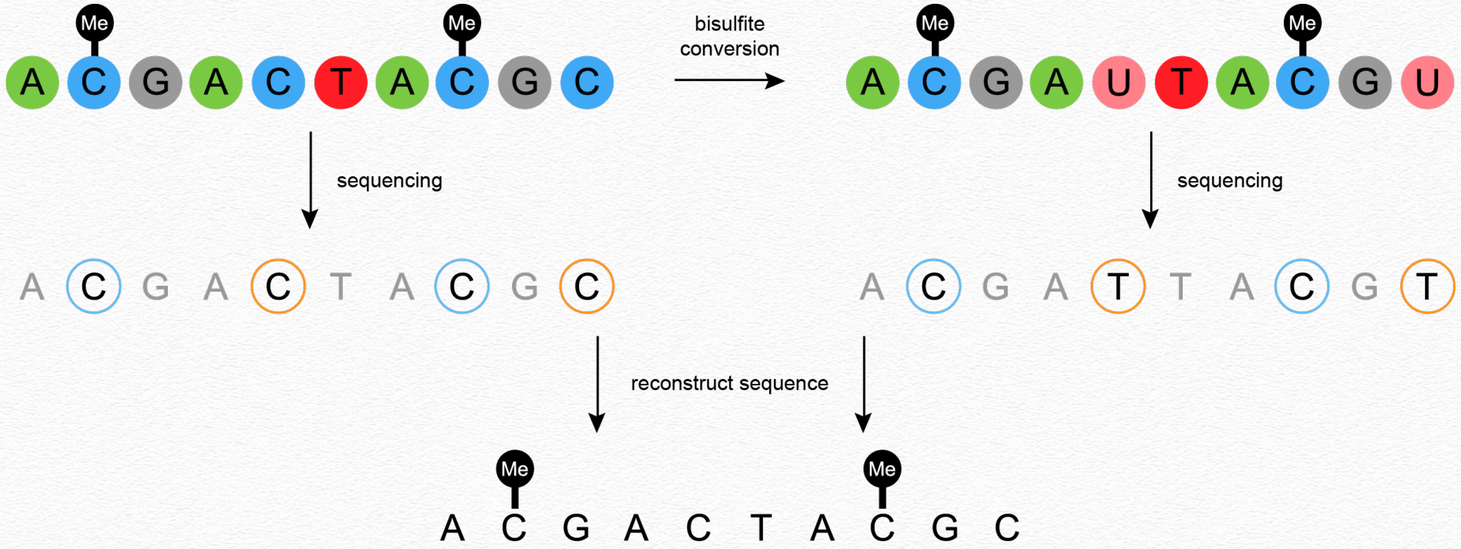
Jones, P. A. Nature Reviews Genetics (2012).

Different Forms of Methylation



Koh, K. *et al.* (2013). Current Opinion in Cell Biology, 25(2), 152–161.

Bisulfite Sequencing



http://www.atdbio.com

Aligning Bis-seq reads

Genomic DNA sequence `CCGATGATGTCGCTGACGCACGA`

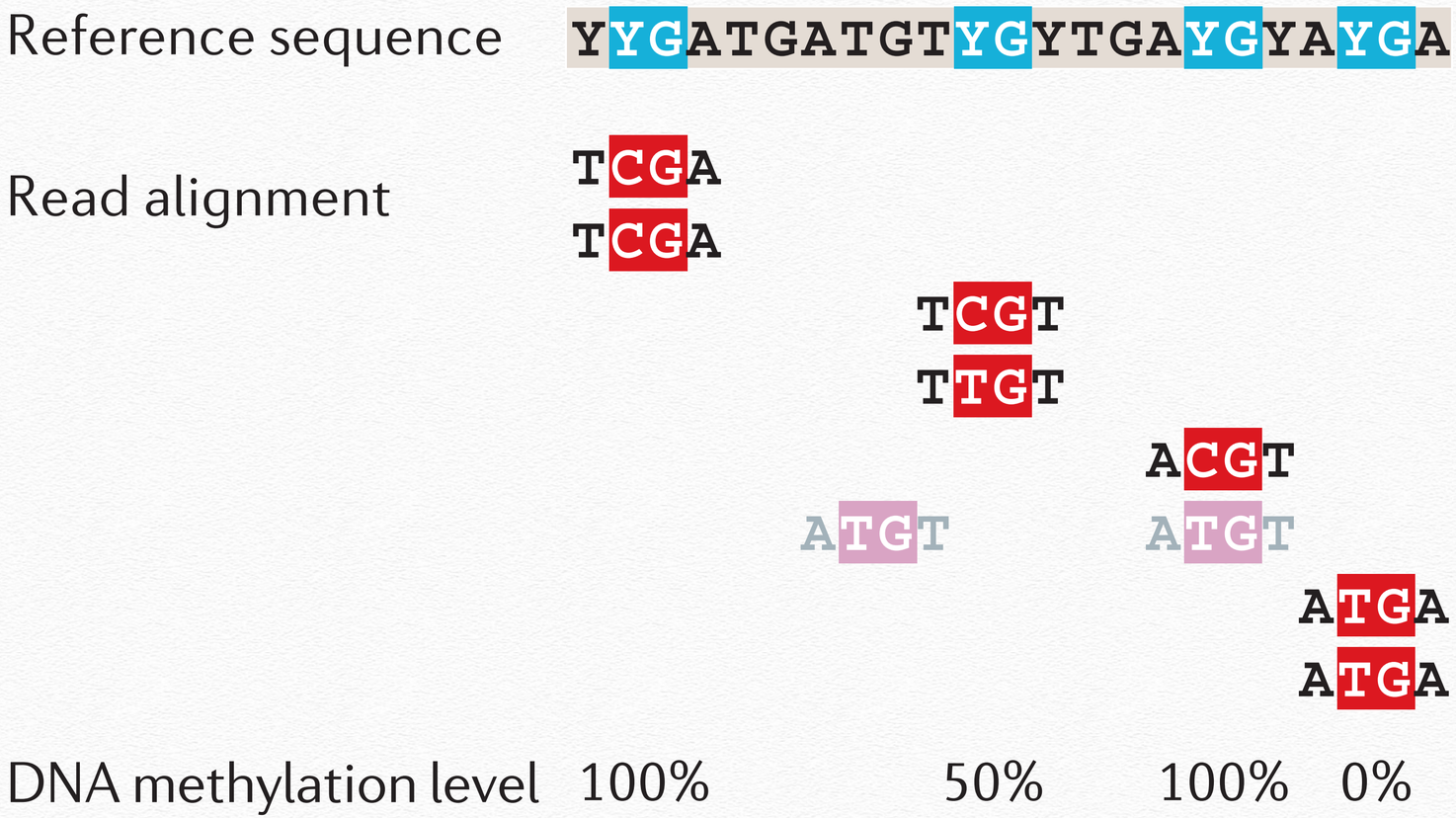
DNA methylation level 100% 50% 50% 0%

DNA fragmentation, selective conversion of unmethylated Cs into Ts, DNA sequencing

Bisulphite-sequencing reads `ACGT,ATGA,ATGA,ATGT,`
`TCGA,TCGA,TCGT,TTGT`

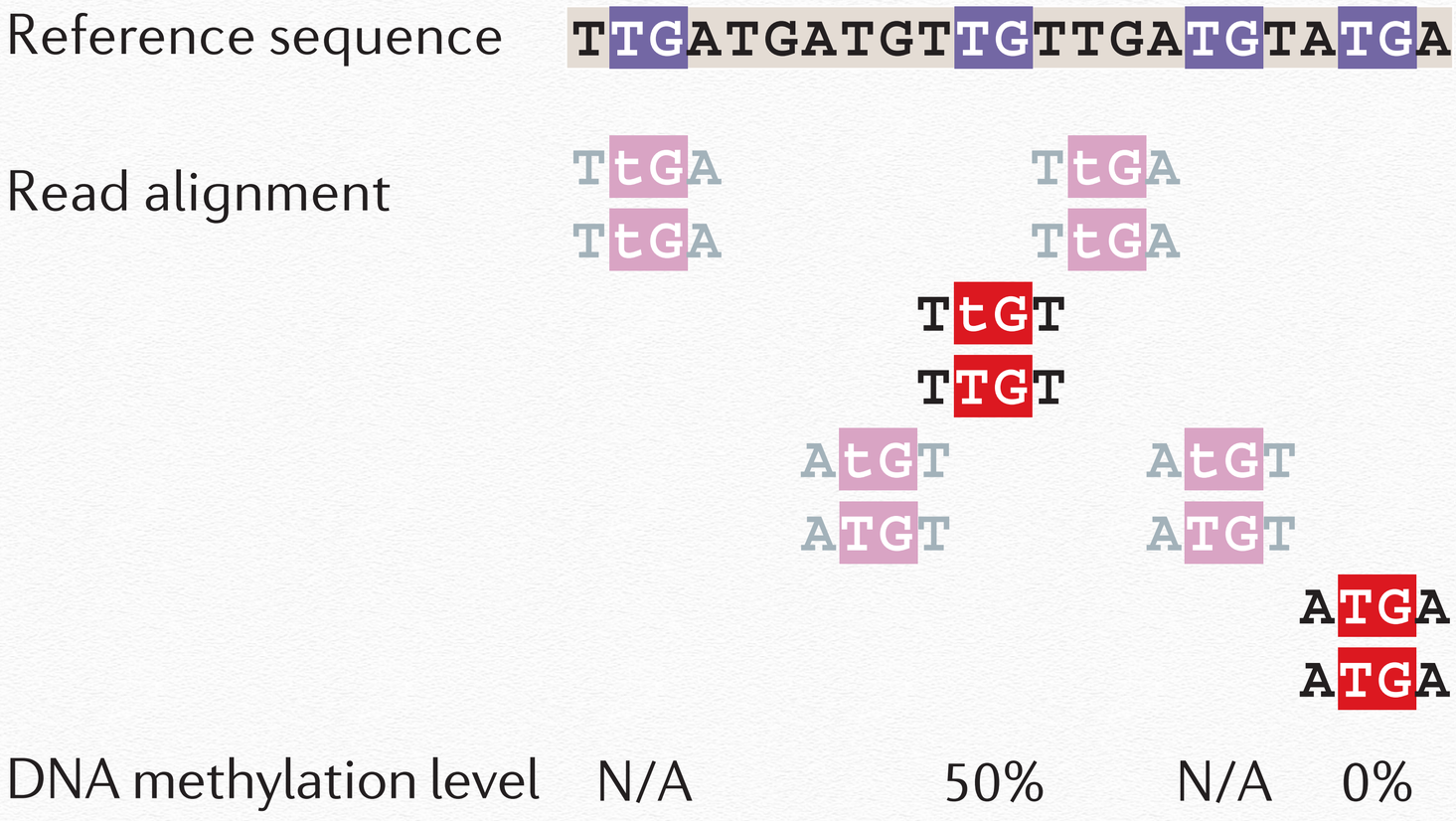
Bock, C. (2012). *Nature Reviews Genetics*, *13*(10), 705–719.

Wild-card Alignment



Bock, C. (2012). *Nature Reviews Genetics*, *13*(10), 705–719.

Three-letter Alignment



Bock, C. (2012). *Nature Reviews Genetics*, *13*(10), 705–719.