

Alignment

Alignment

- The process of determining the most likely source within the genome sequence for the observed sequencing read

At each base, extend alignment;
is total score still above threshold?

```
GCGGAGatggac
|||||
GCGGAGgcggac
```

```
GCGGAGatggac
|||||| xx....
GCGGAGgcggac
```

each mismatch has a *cost*

Insertions/deletions introduce *lots* more ambiguity:

GCGGAGagaccaacc
| | | | |
GCGGAGggaaccacc =

GCGGAGagaccaacc GCGGAGaga-ccaacc
| | | | | | | | | |
GCGGAGggaaccacc GCGGAGggaacca-cc =>

Different error profiles

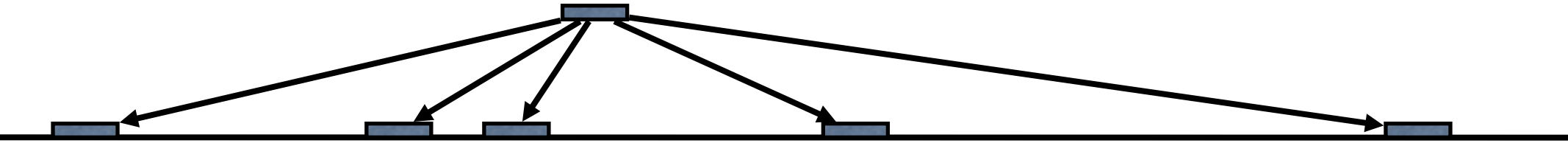
- Roche 454: insertion or deletion errors at homopolymers
- Illumina: low quality calls can occur anywhere in a read
- ABI: Increasing likelihood of sequence errors toward the end of the read
- SOLiD: a series of colors representing two nucleotides

- DNA alignment
 - Very small evolutionary distances (human-human, strains of the reference genome)
 - Assumptions about the number of expected mismatches
 - Allow for much faster processing

Short read mapping

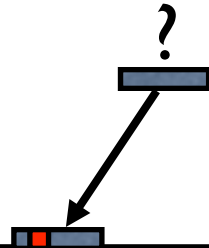
- Input:
 - A reference genome
 - A collection of many 25-100bp tags (reads)
 - User-specified parameters
- Output:
 - One or more genomic coordinates for each tag
- In practice, only 70-75% of tags successfully map to the reference genome.

Multiple mapping



- A single tag may occur more than once in the reference genome.
- The user may choose to ignore tags that appear more than n times.
- As n gets large, you get more data, but also more noise in the data.

Inexact matching



- An observed tag may not exactly match any position in the reference genome.
- Sometimes, the tag *almost* matches one or more positions.
- Such mismatches may represent a SNP (single-nucleotide polymorphism, see [wikipedia](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism)) or a bad read-out.
- The user can specify the maximum number of mismatches, or a phred-style quality score threshold.
- As the number of allowed mismatches goes up, the number of mapped tags increases, but so does the number of incorrectly mapped tags.

- Amount of data several orders of magnitude higher
=> memory + speed
- Ever growing number of implementations for short-read sequence alignment

Sense from sequence reads: methods for alignment and assembly

Paul Flicek & Ewan Birney

The most important first step in understanding next-generation sequencing data is the initial alignment or assembly that determines whether an experiment has succeeded and provides a first glimpse into the results. In parallel with the growth of new sequencing technologies, several algorithms that align or assemble the large data output of today's sequencing machines have been developed. We discuss the current algorithmic approaches and future directions of these fundamental tools and provide specific examples for some commonly used tools.

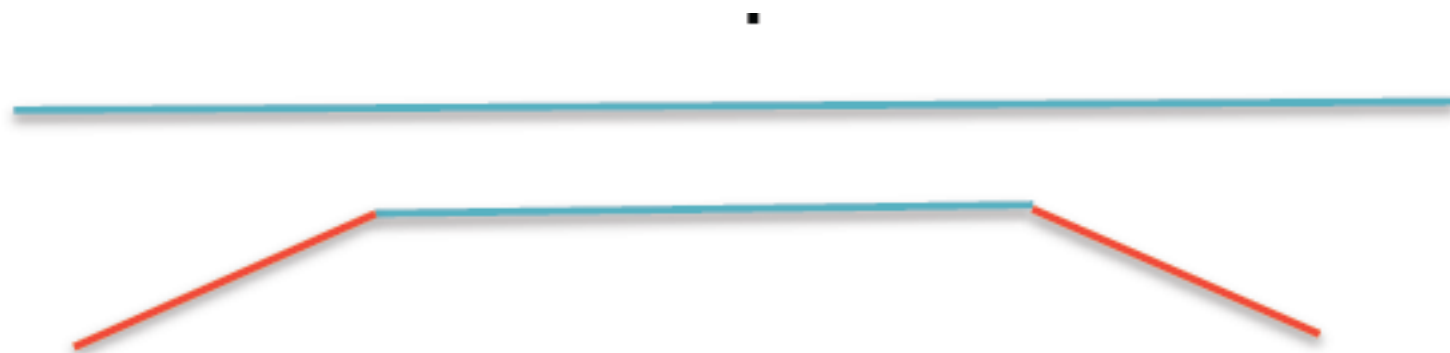
The advent of ultra-high-throughput sequencing technology has captured the imagination of the biological sciences, and with good reason. Ten years ago, on 23 November 1999, the publicly funded human genome project held a massive, worldwide celebration to mark the completion of 1 billion base pairs (bp), one-third of the way to the full sequence of the human genome (<http://www.genome.gov/10002105/>). The amount

a separate lane for each base) range from approximately 800 bp, using the older technology used by the human genome project, to approximately 30 bp for the introductory versions of the second-generation sequencing machines so popular today. Current output ranges from 50 to 400 bp, depending on the technology and the specific biological application. Although uninformative by themselves, once analyzed collectively DNA sequencing

Steps

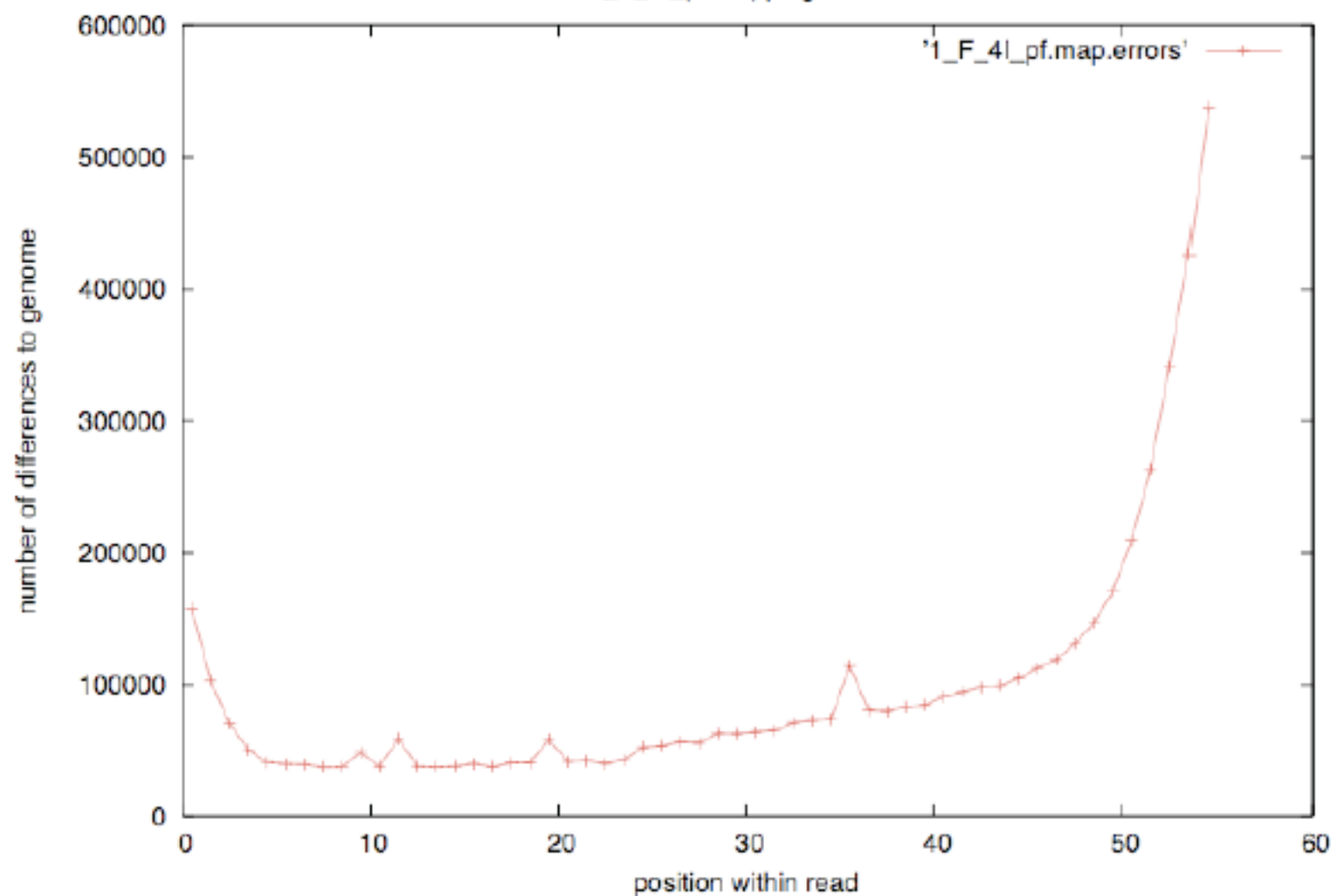
- quickly identify a small set of places in the reference sequence where the location of the best mapping is most likely to be found.





When mapping, a mismatch is not necessarily “real”.

1_F_4l_pf mapping x chick



- ‘mapping policy’ that governs key performance aspects of the specific implementation.

Mapping Reads Back

- Hash Table (Lookup table)
 - FAST, but requires perfect matches. [$O(mn + N)$]
- Array Scanning
 - Can handle mismatches, but not gaps. [$O(mN)$]
- Dynamic Programming (Smith Waterman)
 - Indels
 - Mathematically optimal solution
 - Slow (most programs use Hash Mapping as a prefilter) [$O(mnN)$]

- Main approaches
 - Hash table based implementations
 - Burrows-Wheeler transform (BWT)
- Both approaches apply to sequence and colour space and all technologies

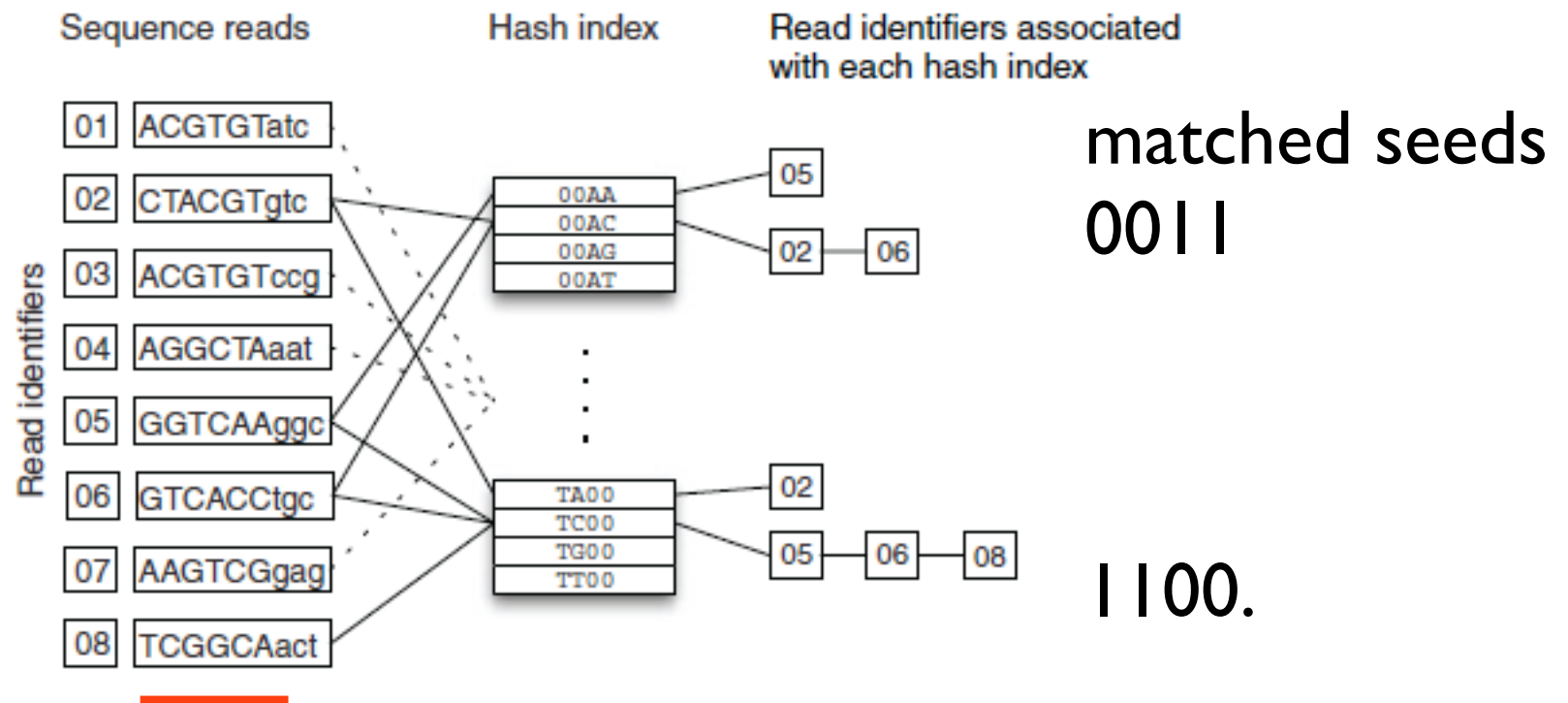
Challenge

- 20m+ reads from genome/transcriptome.
- reference genome/transcriptome

- Heuristics to find potential locations on the genome
 - Slower more accurate alignment run on a subset of potential locations
 - Similar strategy to traditional read alignment algorithms: Blat, Blast, SSAHA2
- Constant trade-off: speed vs. sensitivity
- Guaranteed high accuracy always takes longer

Hash tables

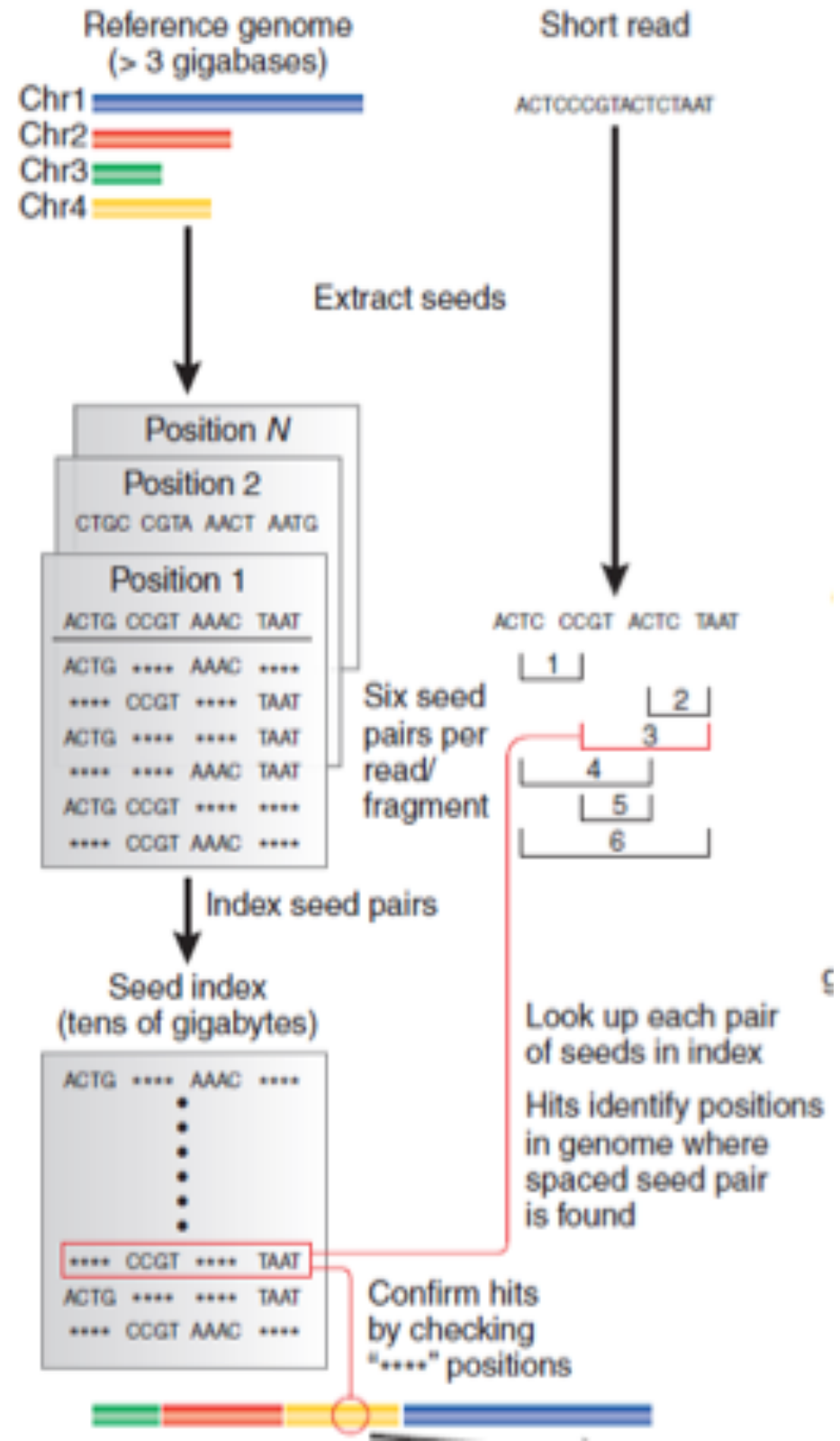
- a common data structure that is able to index complex and nonsequential data in a way that facilitates rapid searching



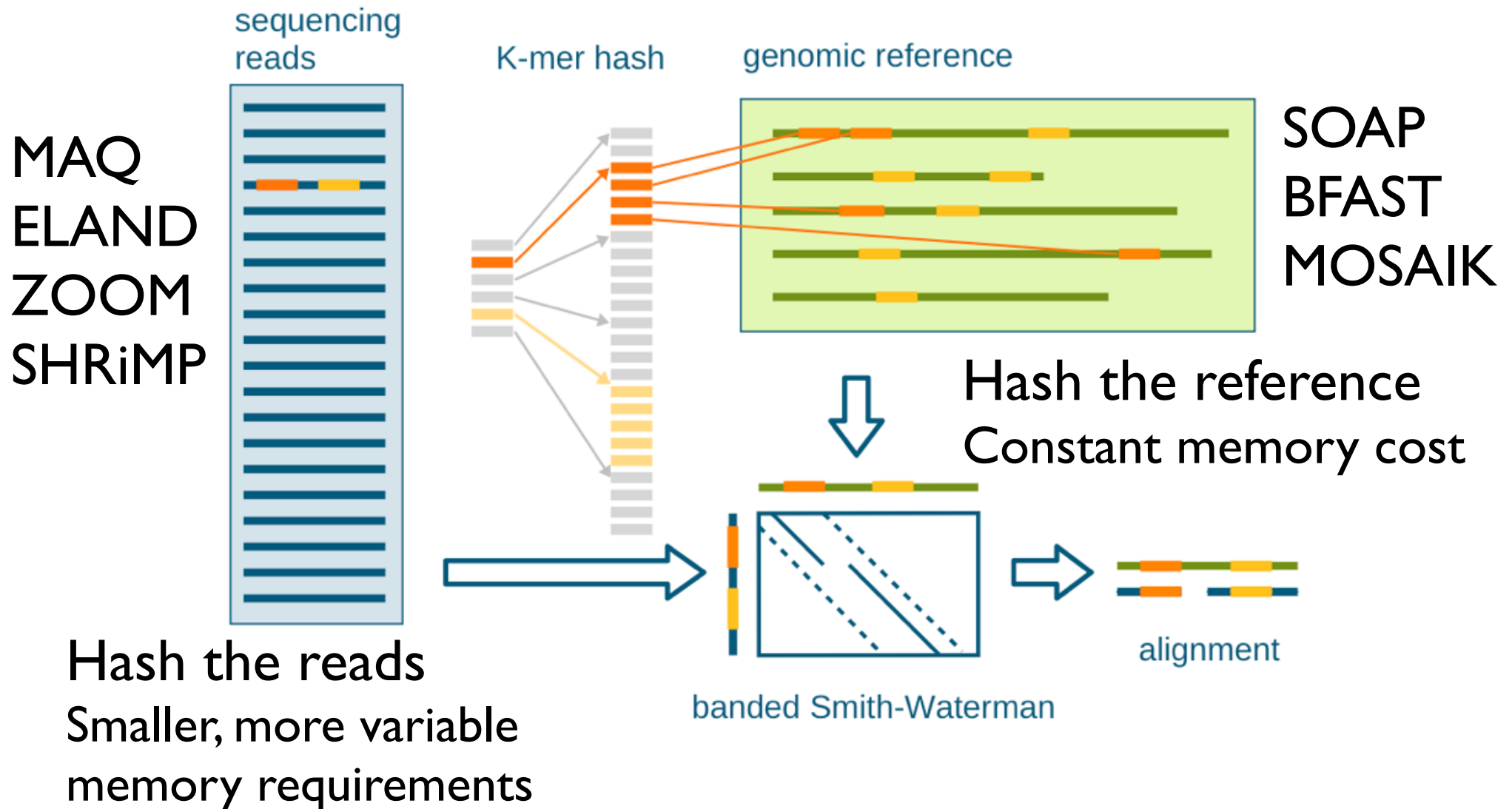
regions that will
be used for seed selection

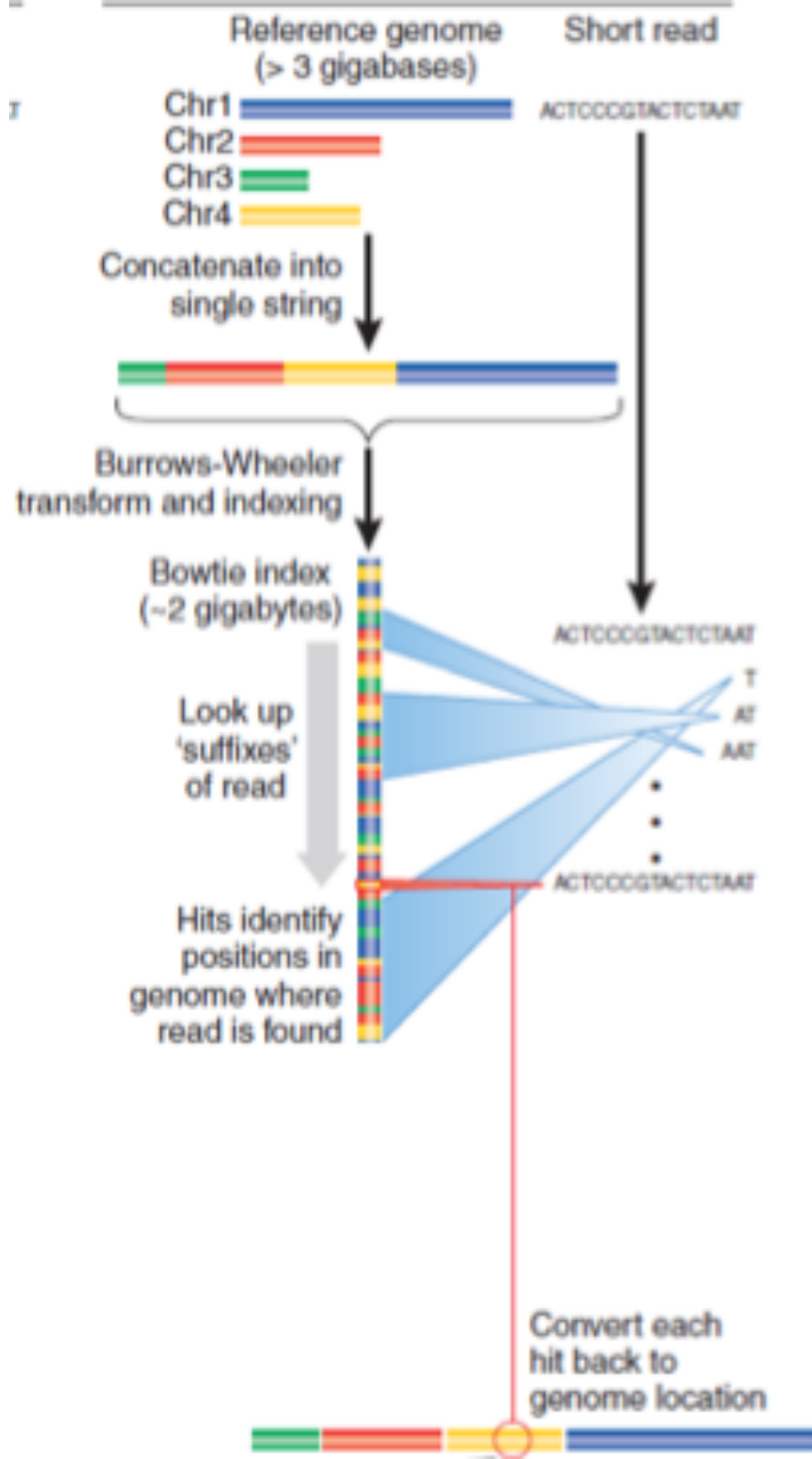
Spaced seed alignment

- Tags and tag-sized pieces of reference are cut into small “seeds.”
- Pairs of spaced seeds are stored in an index.
- Look up spaced seeds for each tag.
- For each “hit,” confirm the remaining positions.
- Report results to the user.



Hash table: common data structure
indexes complex and non-sequential data to
facilitate rapid searching





Burrows-Wheeler

- Store entire reference genome.
- Align tag base by base from the end.
- When tag is traversed, all active locations are reported.
- If no match is found, then back up and try a substitution.

Why Burrows-Wheeler?

BWT very compact:

Approximately $\frac{1}{2}$ byte per base

As large as the original text, plus a few “extras”

Can fit onto a standard computer with 2GB of memory

- Linear-time search algorithm
 - proportional to length of query for exact matches

Burrows-wheeler

Ferragina and Manzini



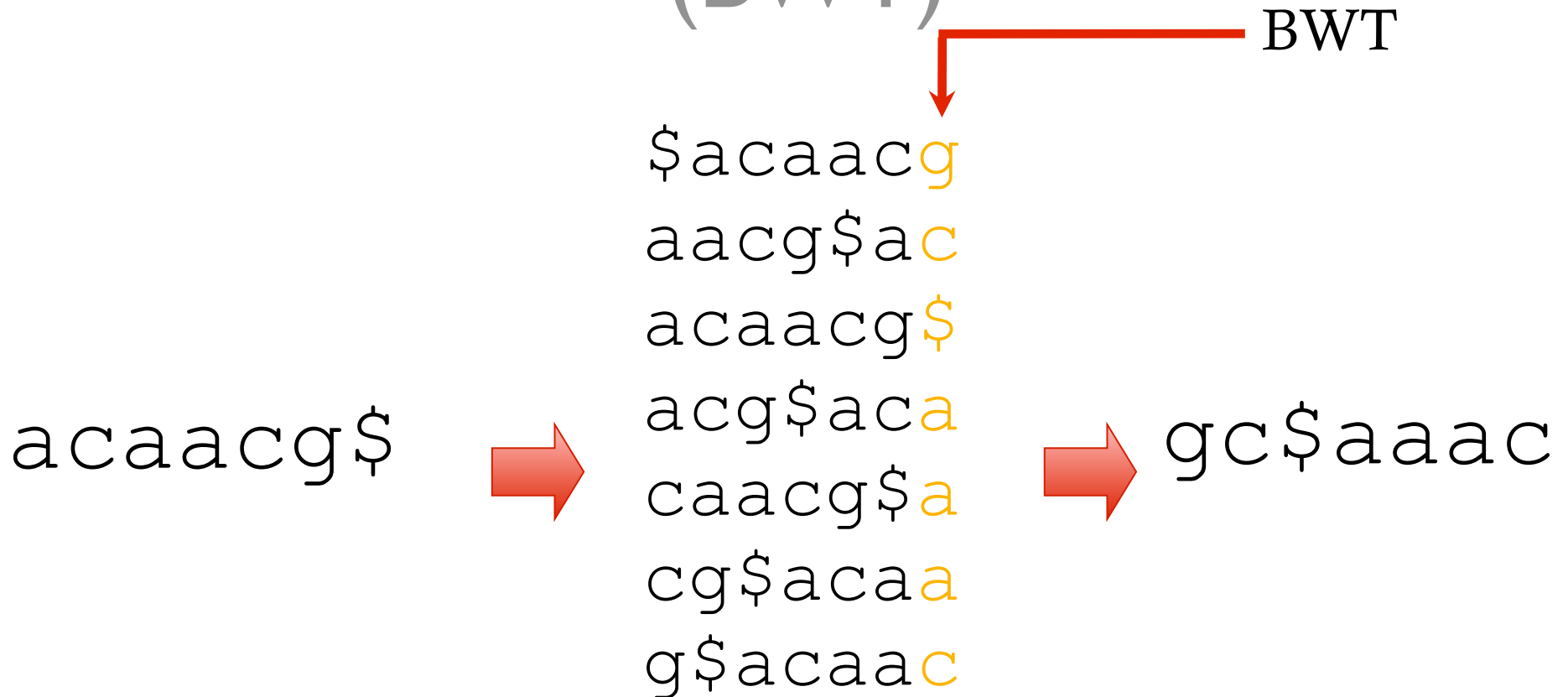
`^TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG$`

Genomic sequence

`GGTTGGTCGGATTCGGAATCACGGAAAATT^AGATTCC$G`

Transform

Burrows-Wheeler Transform (BWT)



Burrows-Wheeler Matrix (BWM)

Exact match

BWT(agcagcagact) = tgcc\$ggaaaac Search for pattern: gca

gca			
\$agcagcagact	\$agcagcagact	\$agcagcagact	\$agcagcagact
act\$agcagcag	act\$agcagcag	act\$agcagcag	act\$agcagcag
agact\$agcagc	agact\$agcagc	agact\$agcagc	agact\$agcagc
agcagact\$agc	agcagact\$agc	agcagact\$agc	agcagact\$agc
agcagcagact\$	agcagcagact\$	agcagcagact\$	agcagcagact\$
cagact\$agcag	cagact\$agcag	cagact\$agcag	cagact\$agcag
cagcagact\$ag	cagcagact\$ag	cagcagact\$ag	cagcagact\$ag
ct\$agcagcaga	ct\$agcagcaga	ct\$agcagcaga	ct\$agcagcaga
gact\$agcagca	gact\$agcagca	gact\$agcagca	gact\$agcagca
gcagact\$agca	gcagact\$agca	gcagact\$agca	gcagact\$agca
gcagcagact\$a	gcagcagact\$a	gcagcagact\$a	gcagcagact\$a
t\$agcagcagac	t\$agcagcagac	t\$agcagcagac	t\$agcagcagac

- final index for the human genome smaller than the genome
- BWA and BOWTIE : 2.3 GB
- SOAP2 uses a different routine index 5.4 GB

Bowtie

- cannot handle indels
 - Many reads with one good valid alignment
 - lots of high quality reads
 - Small number of alignments/read

BWA

- equivalent strategy to Bowtie but in addition performs gapped alignment (will allow indels)
- 10x faster than maq
- 'hot'

Comparison

Spaced seeds

- Requires ~50Gb of memory.
- Runs 30-fold slower.
- Is much simpler to program.

MAQ

Burrows-Wheeler

- Requires <2Gb of memory.
- Runs 30-fold faster.
- Is much more complicated to program.

Bowtie

Short-read mapping software

Software	Technique	Developer	License
Eland	Hashing reads	Illumina	?
SOAP	Hashing refs	BGI	Academic
Maq	Hashing reads	Sanger (Li, Heng)	GNUPL
Bowtie	BWT	Salzberg/UMD	GNUPL
BWA	BWT	Sanger (Li, Heng)	GNUPL
SOAP2	BWT & hashing	BGI	Academic

[http://www.oxfordjournals.org/our_journals/bioinformatics/
nextgenerationsequencing.html](http://www.oxfordjournals.org/our_journals/bioinformatics/nextgenerationsequencing.html)

MassGenomics

Medical genomics in the post-genome era

[Home](#) [About](#) [Aligners](#) [Genomes](#) [Journals](#) [VarScan](#)

Aligners






At AGBT 2009 I presented a poster comparing 10 short read aligners on multiple sets of Illumina/Solexa sequencing data. This was an idea I'd conceived six months earlier, when it seemed that each new issue of *Bioinformatics* or *Genome Research* had an article about another short read alignment algorithm. We built Maq into the pipeline here over a year ago. It was a bit of a gamble to discard the Illumina-provided aligner (ELAND) and go with something else, but I think that the success of the AML cancer genome paper shows that our bet paid off. So as the list of alternative aligners grew, I came up with a set of questions to ask when considering any aligner for our pipeline here.

Ten Short Read Aligners I've Evaluated

For my AGBT poster I chose 10 aligners (including Maq) to install and evaluate on multiple Illumina/Solexa data sets:

Aligner	Version	Developer	License
BFAST	0.3.1	UCLA	Academic
Bowtie	0.9.8	Salzberg/UMD	GNUPL
cross_match	1.080721	U. Wash.	Academic
CELL	2.0	CLC bio	Commercial
Maq	0.7.1	WTSI	GNUPL
Novoalign	2.00.12	Novocraft	Academic
RMAP	0.41	CSHL	Public
SeqMap	1.0.12	Stanford	Academic
Shrimp	1.1.0	U. Toronto	Public
SOAP	2.01	BGI	Academic










RECENT POSTS

-  [First Look: Data from IonTorrent's 316 Chip](#)
-  [Recurrent mutations in chronic lymphocytic leukemia](#)
-  [A virus-associated liver cancer genome](#)
-  [Inflammation, Genetic Instability, and Cancer](#)
-  [Genome sequencing of multiple myeloma](#)

DISCLAIMER

The views expressed on this site, including blog posts and static pages, do not necessarily reflect the opinions of the Genome Institute at Washington University, the Washington University School of Medicine, or Washington University in St. Louis.

BLOGROLL

-  [Adaptive Complexity](#)
-  [Bio-Inf@Becker](#)
-  [Evolgen](#)
-  [Fejes.ca](#)
-  [Genetic Inference](#)
-  [GeneticFuture](#)
-  [GT Daily Scan](#)
-  [Nick Loman](#)
-  [Politigenomics](#)

References

- (Bowtie) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Langmead et al, Genome Biology 2009, 10:R25
- SOAP: short oligonucleotide alignment, Ruiqiang Li et al. Bioinformatics (2008) 24: 713-4
- (BWA) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform, Li Heng and Richard Durbin, (2009) 25:1754–1760
- SOAP2: an improved ultrafast tool for short read alignment, Ruiqiang Li, (2009) 25: 1966–1967
- (MAQ) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Li H, Ruan J, Durbin R. Genome Res. (2008) 18:1851-8.
- Sense from sequence reads: methods for alignment and assembly, Paul Flicek & Ewan Birney, Nature Methods 6, S6 - S12 (2009)
- <http://www.allisons.org/II/AlgDS/Strings/BWT/>

- How many mismatches?
- multiple matches?
- Best matches or any matches?
- Indels?
- What are you searching for?