Mapping

CTB

6/12/2013

BLAST heuristics:
exact word matching

BLASTN filters sequences for exact matches between "words" of length 11:

```
GAGGGTATGACGATATGGCGATGGAC
||x|||||x|||||||||||x|x||x
GAcGGTATcACGATATGGCGgT-Gag
```

BLAST

…but what about pathological situations?

```
GAGGGTATGACGATATGGCGATGGAC
||x|||||x||||||x||||||x|x||x
GAcGGTATcACGATGTGGCGgT-Gag
```

This will not be scored as a match, because BLAST *only* scores matches with a core "seed" match of 11 bases.

Long reads: BLAST vs ‘blat’

- BLAST requires that a query sequence contains the same 11-mer as a database sequence before it attempts further alignment.
- Any given 11-mer occurs only once in 2m sequences, so this filters out many database sequences quickly.
- You can also store the list of all possible 11-mers in memory easily (~2mb), making it possible to keep track of everything quickly.

- ‘blat’ does the same thing as BLAST, but is faster because it uses longer k-mers.

Mapping

- Goal: assign all reads to location(s) within a reference sequence database.

- Inference: at least one of the locations to which the reads map is the actual location from which the read came.

- Req'd for variant detection, ChIP-seq, and mRNAseq/differential expression/counting.

Mapping challenges

- Incomplete reference.
- Volume of data.

- Errors in reads
- Variation in reference
- Multicopy sequences (e.g. repeats)

Errors in reads

```
Ref:   ATGGACGGACCGATGGACCAGTGCA
              .........X..............
Read:  ATGGACGGAGCGATGGACCAGTGCA
```

Effect of errors on mapping

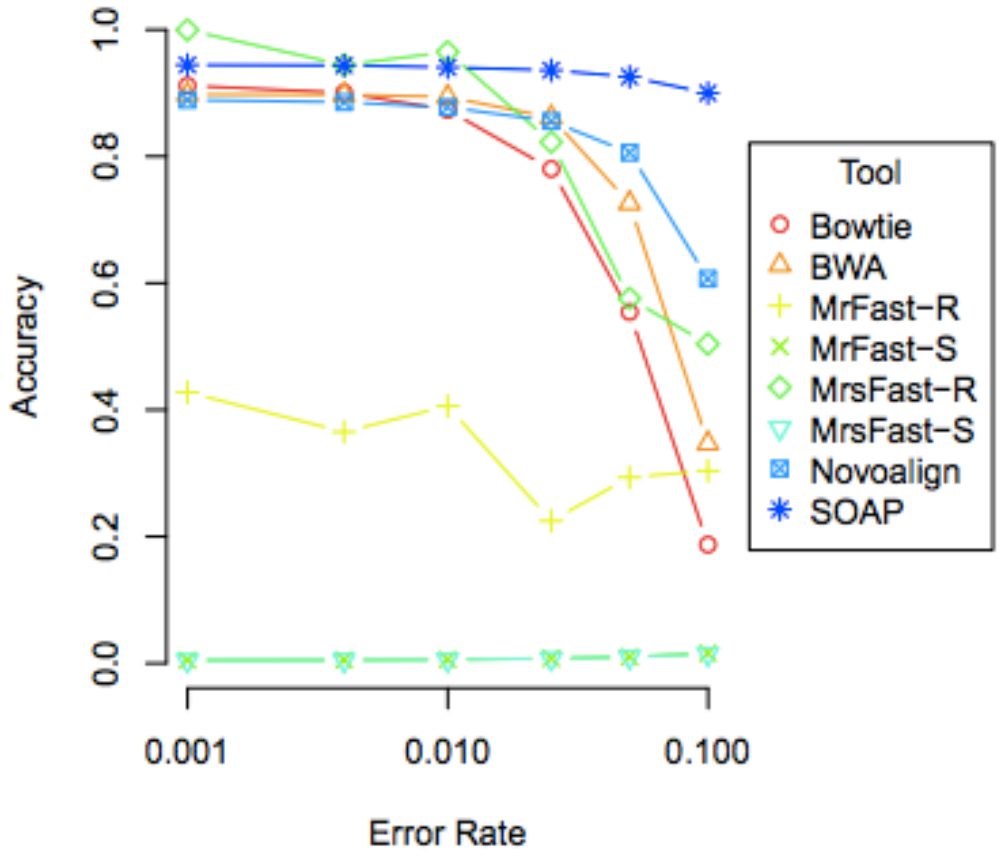


Fig 1a of Ruffalo et al. – mapping against human genome w/errors
PMID 21856737, Bioinformatics 2011.

Variation in reference

```
Ref:   ATGGACGGACCGATGGCCCAGTGCA
       .........X......X........
Read:  ATGGACGGAGCGATGGACCAGTGCA
```

Variation in reference (equivalently, variation in what you're sequencing!) manifests as *errors.*

Variation => allelic mapping bias

```
H2     :   ATGGACGGACCGATGGCCCAGTGCA
H1/ref:   ATGGACGGACCGATGGACCAGTGCA

                .........X...............
H1 read: ATGGACGGAGCGATGGACCAGTGCA

H1/ref:   ATGGACGGACCGATGGACCAGTGCA
                .........X......X........
H2 read: ATGGACGGAGCGATGGCCCAGTGCA
```

If H1 and H2 are both real haplotypes, and H1 is your reference, then reads from H1 are more likely to map correctly. This is because H2 contains extra variation, from the viewpoint of the mapper.

Multi-copy sequence/repeats

Ref 1: ATGGACGGACCGATGGCCCAGTGCA

Ref 2: ATGGACTGACCGATGGTCCAGTGCA

Ref 3: ATGGACGGAGCGATGGTCCAGTGCA

Read : ATGGACGGAGCGATGTCCCAGTGCA

Do you map the read to one, all, or none?

(Depends on your goal :)

Substitutions vs indels

Ref : ATGGACGGAGCGATGG-TCCAGTGCA

Read: ATGGACGGA-CGATGGATCCAGTGCA

More complicated *alignment* of some sort is needed.

How *alignment* works, and why indels are the devil

There are many alignment strategies, but most work like this:

```
GCGGAGatggac      GCGGAGatggac
||||||...... =>   ||||||x.....
GCGGAGgcggac      GCGGAGgcggac
```

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

How *alignment* works, and why indels are the devil

There are many alignment strategies, but most work like this:

```
GCGGAGatggac      GCGGAGatggac
||||||.....  =>   ||||||xx....
GCGGAGgcggac      GCGGAGgcggac
```

Each mismatch *costs*.

How *alignment* works, and why indels are the devil

Insertions/deletions introduce *lots* more ambiguity:

```
GCGGAGagaccaacc      GCGGAGag-accaacc
||||||          =>  ||||||
GCGGAGggaaccacc      GCGGAGggaacc-acc

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

Practical effect of indels --

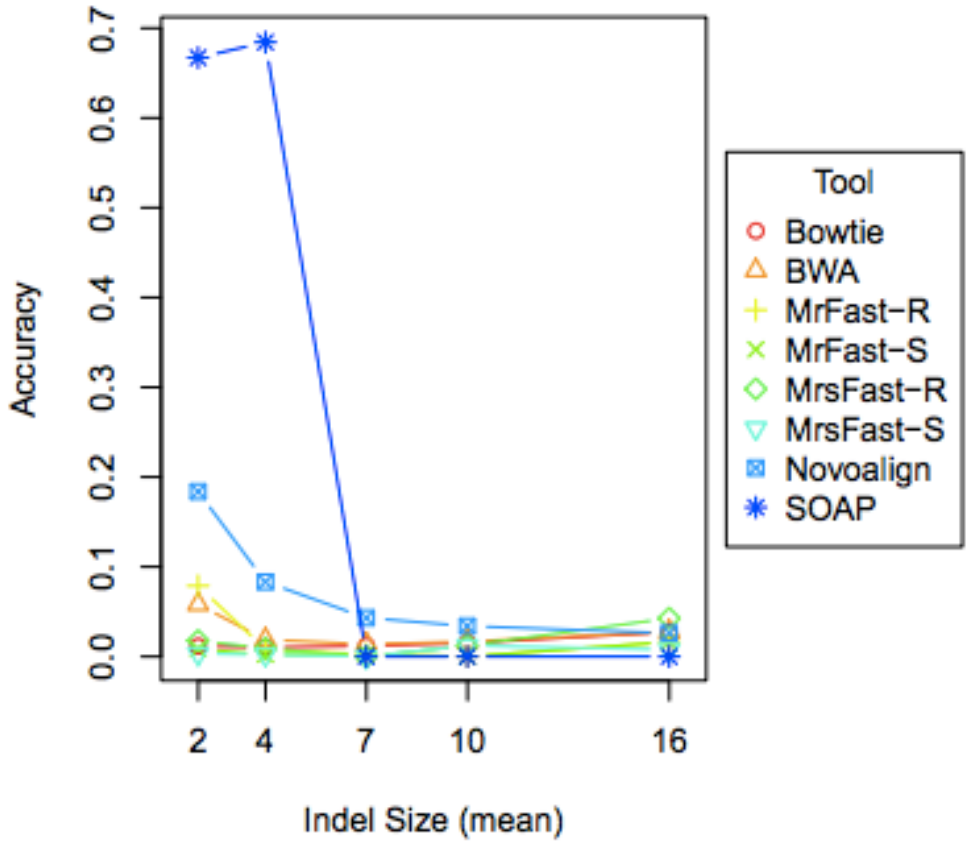
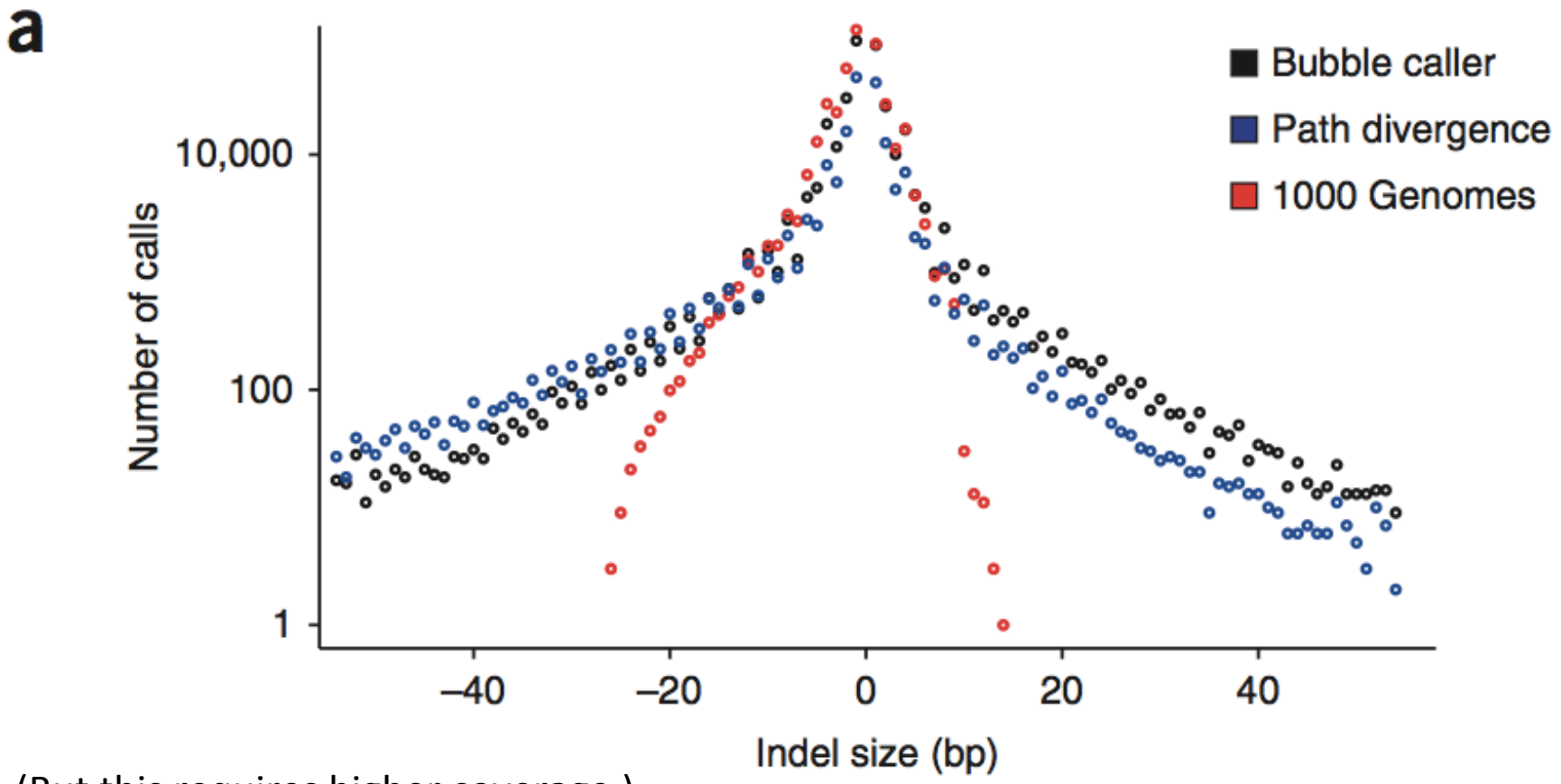


Fig 3a of Ruffalo et al. – mapping against human genome w/indels
PMID 21856737, Bioinformatics 2011.

Variant calling with *assembly* approaches – "Cortex" – good at indels.



(But this requires higher coverage.)

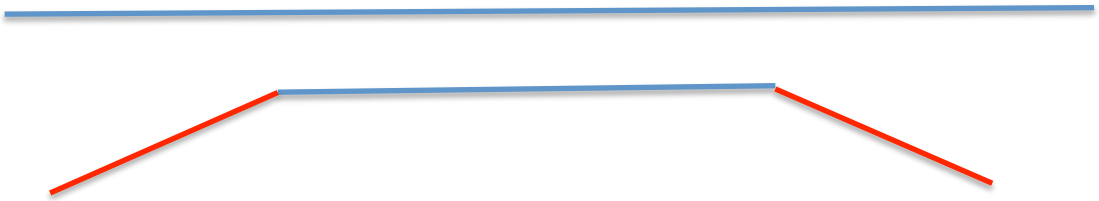
Iqbal et al., Nat Genet. 2012

Mapping miscellany

- Don’t use BLAST.
- Tech-specific bias.
- Splice sites.
- Indexing.
- Which mapper?

Want *global*, not *local*, alignment

- You do not want matches *within* the read, like BLAST would produce.



- Do not use BLAST!
- It's not tuned to kinds of errors, and it's actually quite slow.

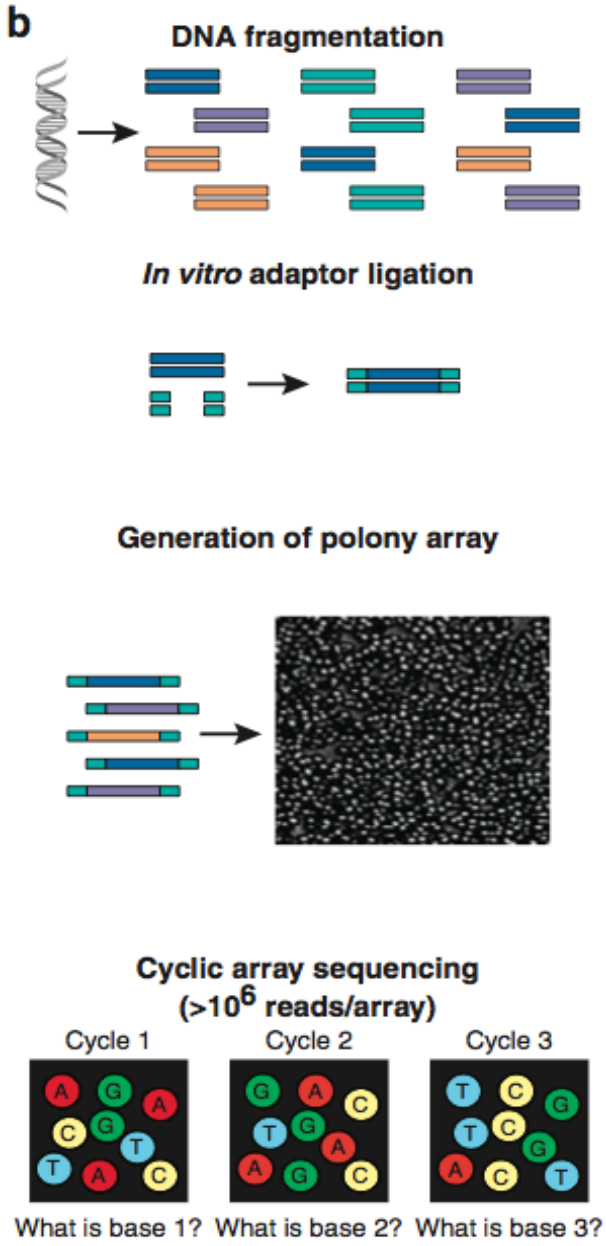
Garbage reads

Overlapping polonies result in mixed signals.

These reads will not map to anything!

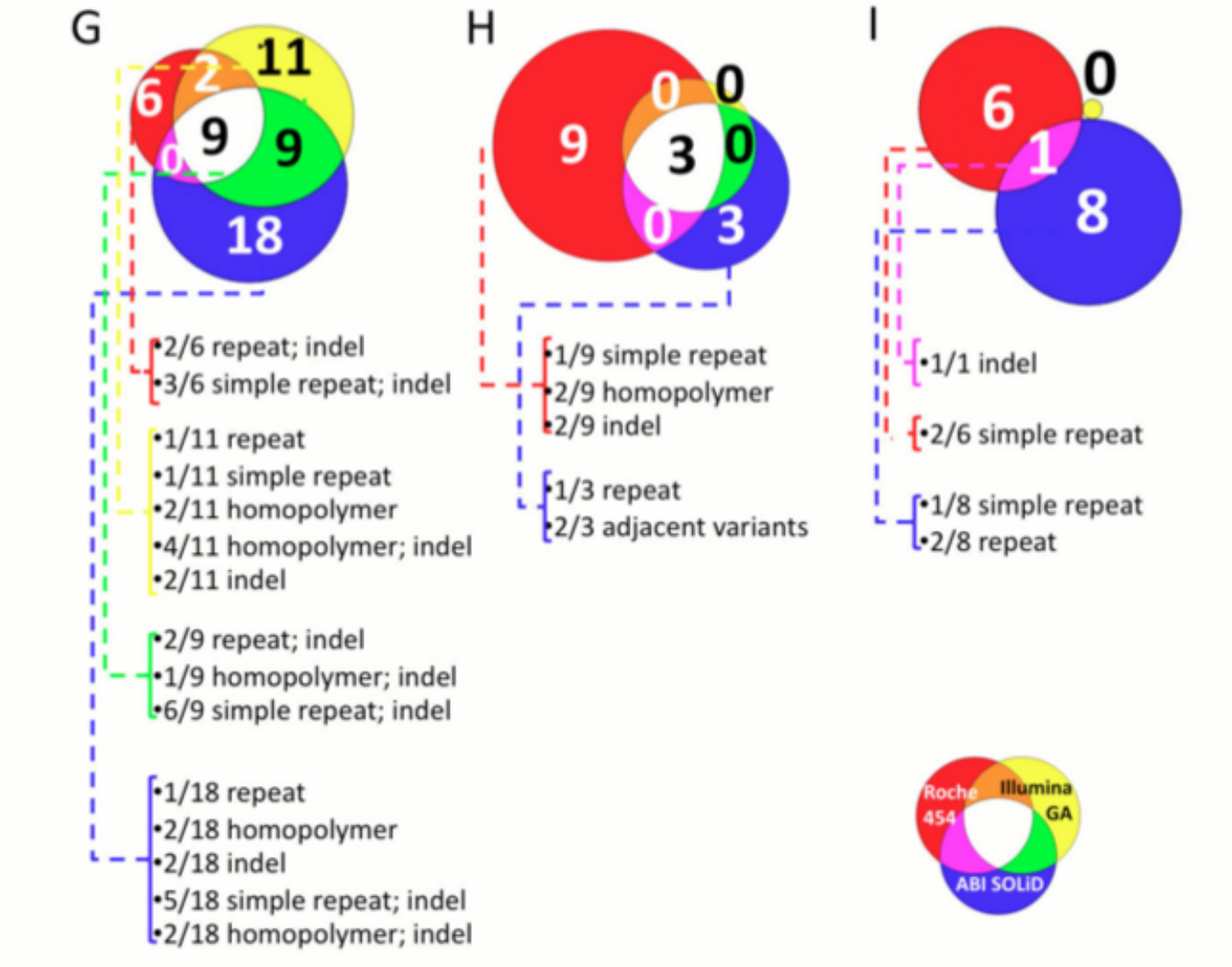
Used to be ~40% of data.

Increasingly, filtered out by sequencing software.



Shendure and Ji, Nat Biotech, 2008

Technology-specific bias



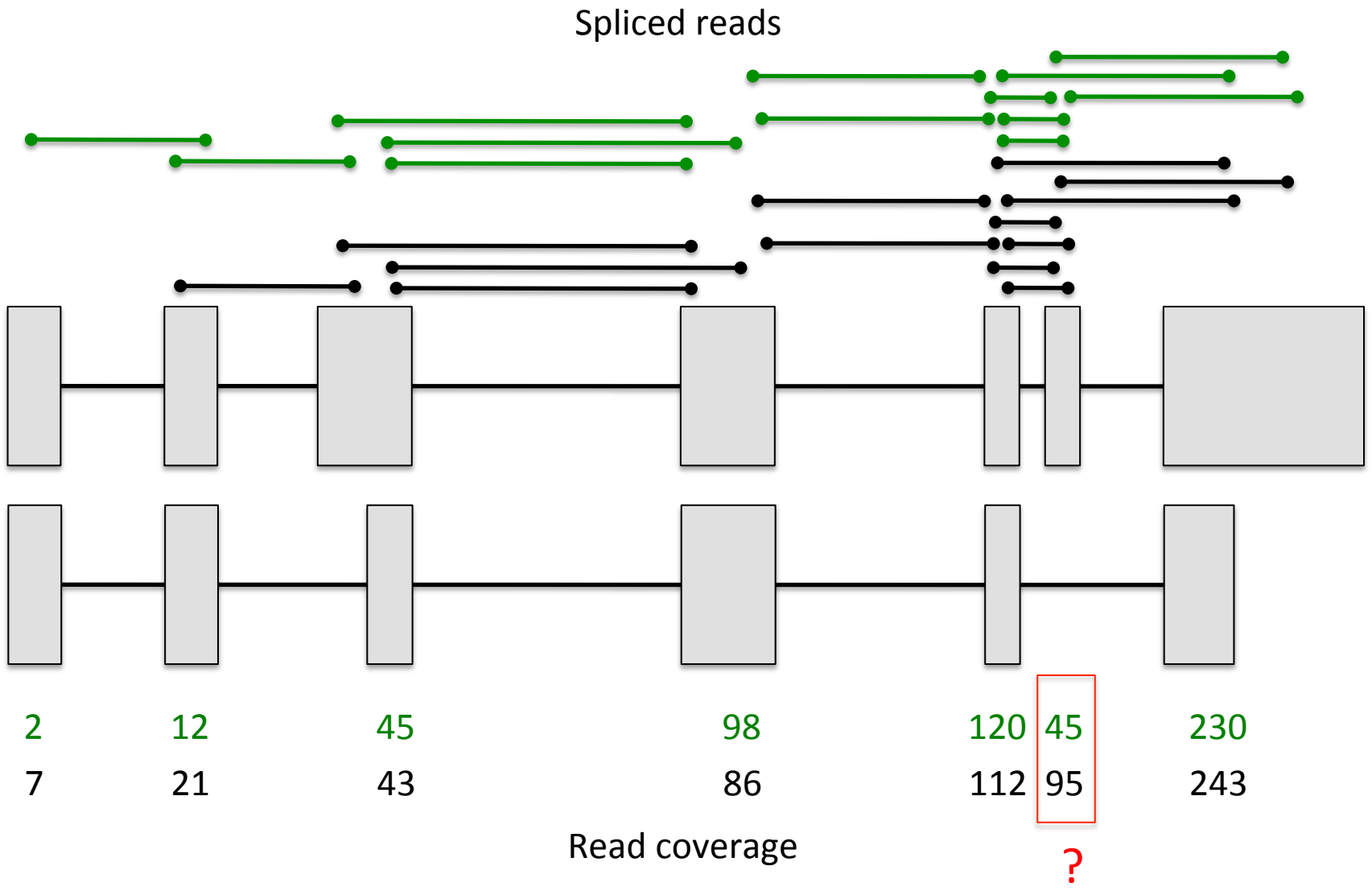
Harismendy et al., Genome Biol. 2009, pmid: 19327155

Splice sites

- If you are mapping transcriptome reads to the genome, your reference sequence is different from your source sequence because of splicing.
- This is a problem if you don't have a really good annotation.

- Main technique: try to map across splice sites, build new exon models (Tophat/Cufflinks)
- Another technique: assembly.

How to detect differential splicing



Indexing – e.g. BLAST

BLASTN filters sequences for exact matches between "words" of length 11:

```
GAGGGTATGACGATATGGCGATGGAC
||x|||||x|||||||||||x|x||x
GAcGGTATcACGATATGGCGgT-Gag
```

What the 'formatdb' command does (see Tuesday's first tutorial) is *build an index* ("index") sequences by their 11-base word content – a "reverse index" of sorts.

Indexing – e.g. BLAST

What the 'formatdb' command does (see Tuesday's first tutorial) is *build an index* ("index") sequences by their 11-base word content – a "reverse index" of sorts.

Since this index only needs to be built once for each reference, it can be slower to build – what matters to most people is *mapping speed.*

All short-read mappers have an indexing step.

Speed of indexing & mapping.

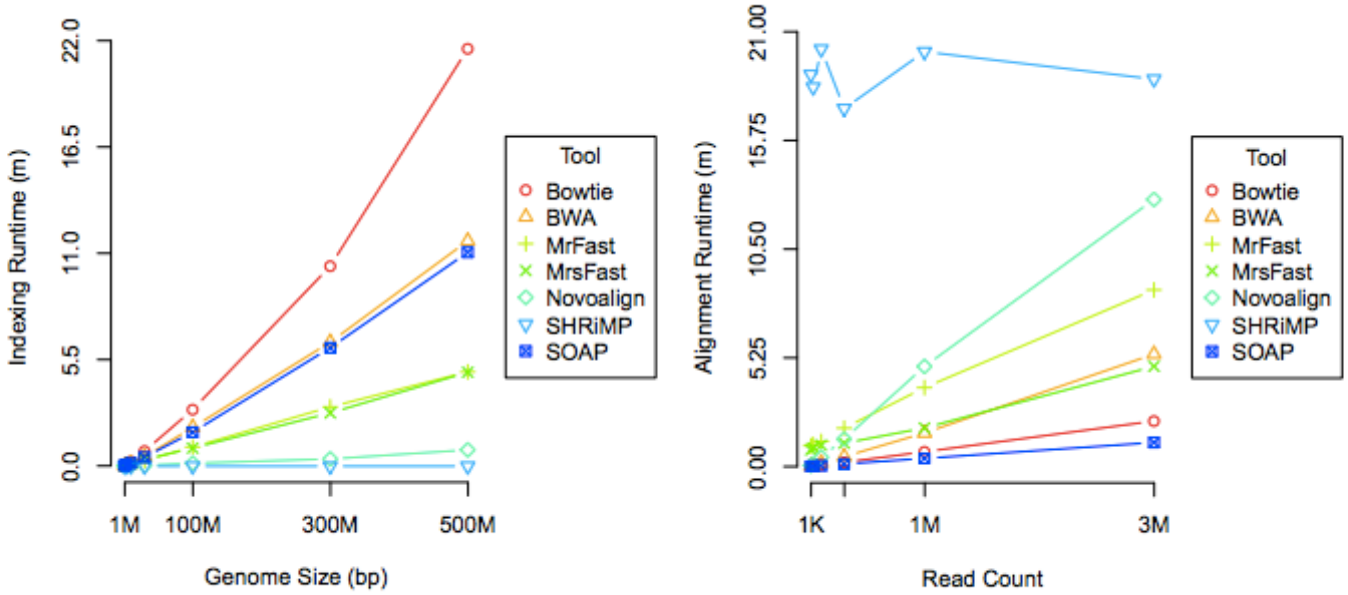


Fig. 5: Runtime measurements: (a) shows indexing time vs. genome size, (b) shows alignment time vs. read count on a 500Mbp genome.

Fig 5 of Ruffalo et al. PMID 21856737, Bioinformatics 2011.

Simulations => understanding mappers

Table 1. Read mapping errors for single (SE) and paired end (PE) reads from random (simulated) and real transcriptomes

Organism	Num Trans	Error	TP (d)	FP (d)	TP (u)	FP (u)	TP (m)	FP (m)
Random (SE)	5000	1%	92%	0%	92%	0%	92%	0%
Mouse (SE)	5000	1%	87%	5%	81%	0%	92%	12%
Random (PE)	5000	1%	85%	0%	85%	0%	85%	0%
Mouse (PE)	5000	1%	81%	4%	77%	0%	85%	9%

Mapping parameters are default (d), unique (u), and multimap (m). True positives are reads that were successfully mapped to their originating transcript. False positives are reads that were mapped to other transcripts (even if the read was an exact match to the alternate transcript).

Mappers will ignore some fraction of reads due to errors.

Pyrkosz et al., unpub.; http://arxiv.org/abs/1303.2411

Does choice of mapper matter?

Not in our experience.

Table 3. Comparison of Three Common Mapping Programs on the Same Chicken Data Sets

Organism	Num Trans	Bowtie TP (d)	FP (d)	BWA TP (d)	FP (d)	SOAP2 TP (d)	FP (d)
Chicken	100%	78%	22%	78%	20%	78%	22%
Chicken	90%	72%	21%	72%	20%	72%	21%
Chicken	80%	65%	22%	65%	21%	65%	22%
Chicken	70%	58%	22%	58%	21%	58%	22%
Chicken	60%	51%	20%	50%	19%	51%	20%
Chicken	50%	44%	19%	44%	18%	44%	19%
Chicken	40%	36%	16%	37%	16%	36%	17%
Chicken	30%	27%	13%	27%	13%	27%	12%
Chicken	20%	19%	11%	19%	11%	19%	11%
Chicken	10%	9%	5%	9%	6%	9%	5%

Comparison of Bowtie, BWA, and SOAP2 mapping programs on the same simulated reads for error-free chicken read sets (triplicate and averaged) with decreasing completeness of the reference transcriptome, showing equivalent results.

Reference completeness/quality matters more!

Pyrkosz et al., unpub.; http://arxiv.org/abs/1303.2411

Misc points

- Transcriptomes and bacterial genomes have very few repeats…
- …but for transcriptomes, you need to think about shared exons.

- For genotyping/association studies/ASE, you may not care about indels too much.

- Variant calling is less sensitive to coverage than assembly (20x vs 100x)

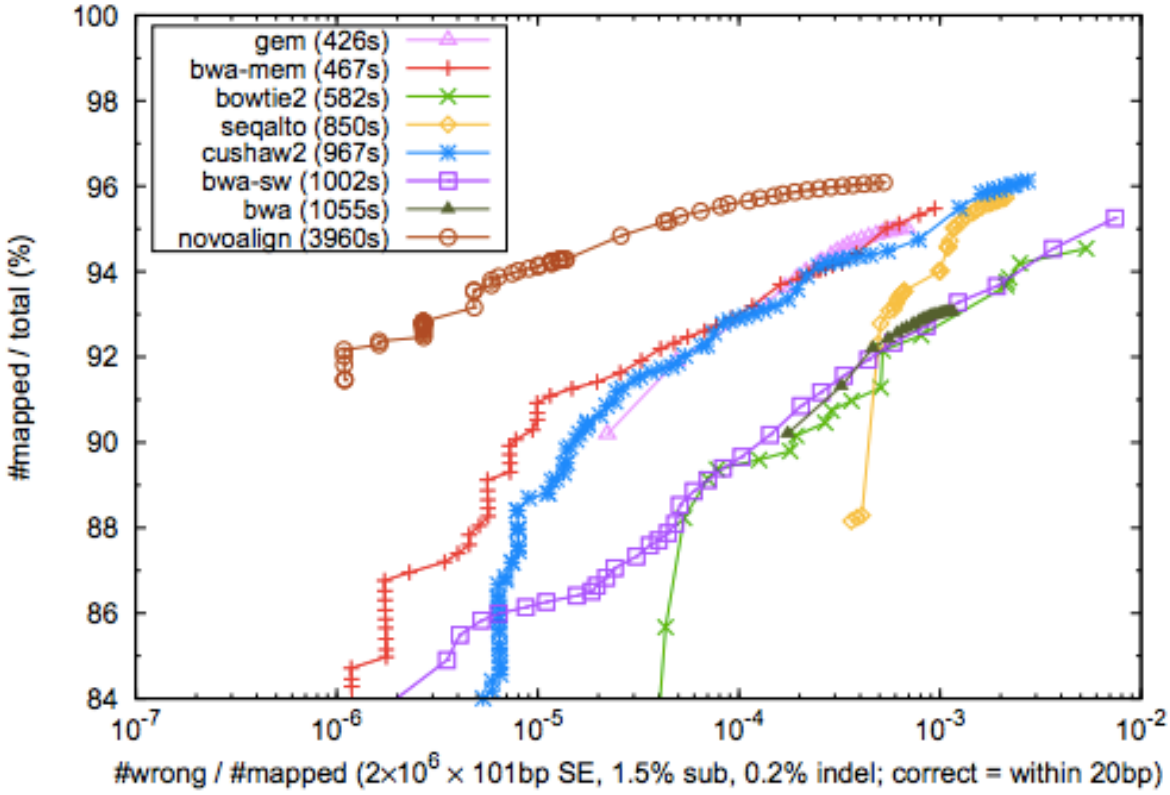
Using quality scores?

- Bowtie uses quality scores; bwa does not.

- This means that bowtie can align some things in FASTQ that cannot be aligned in FASTA.

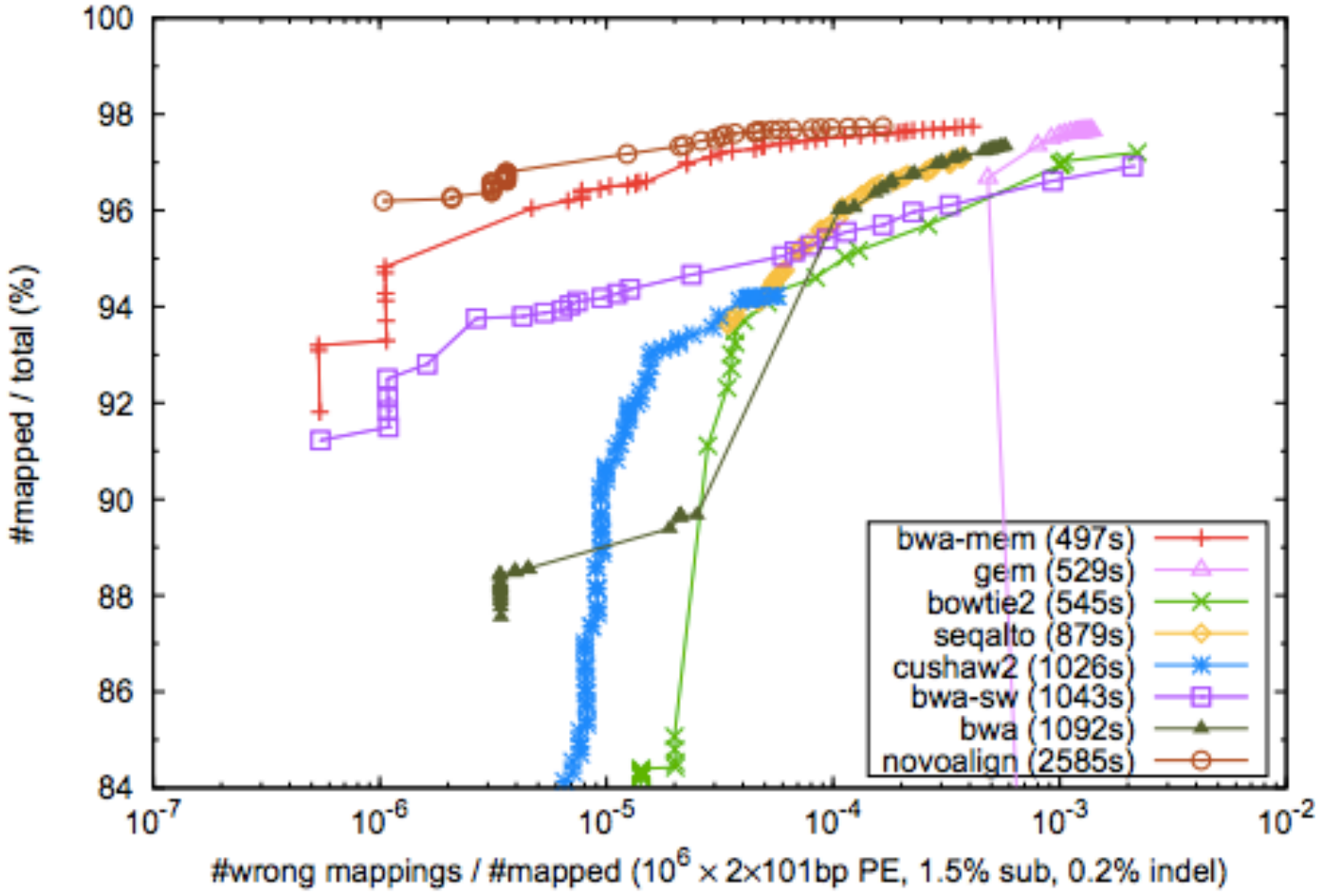
  See: http://www.homolog.us/blogs/blog/2012/02/28/bowtie-alignment-with-and-without-quality-score/

Comparative performance/SE



Heng Li, BWA-MEM: http://arxiv.org/pdf/1303.3997v2.pdf

Comparative performance/PE



Heng Li, BWA-MEM: http://arxiv.org/pdf/1303.3997v2.pdf