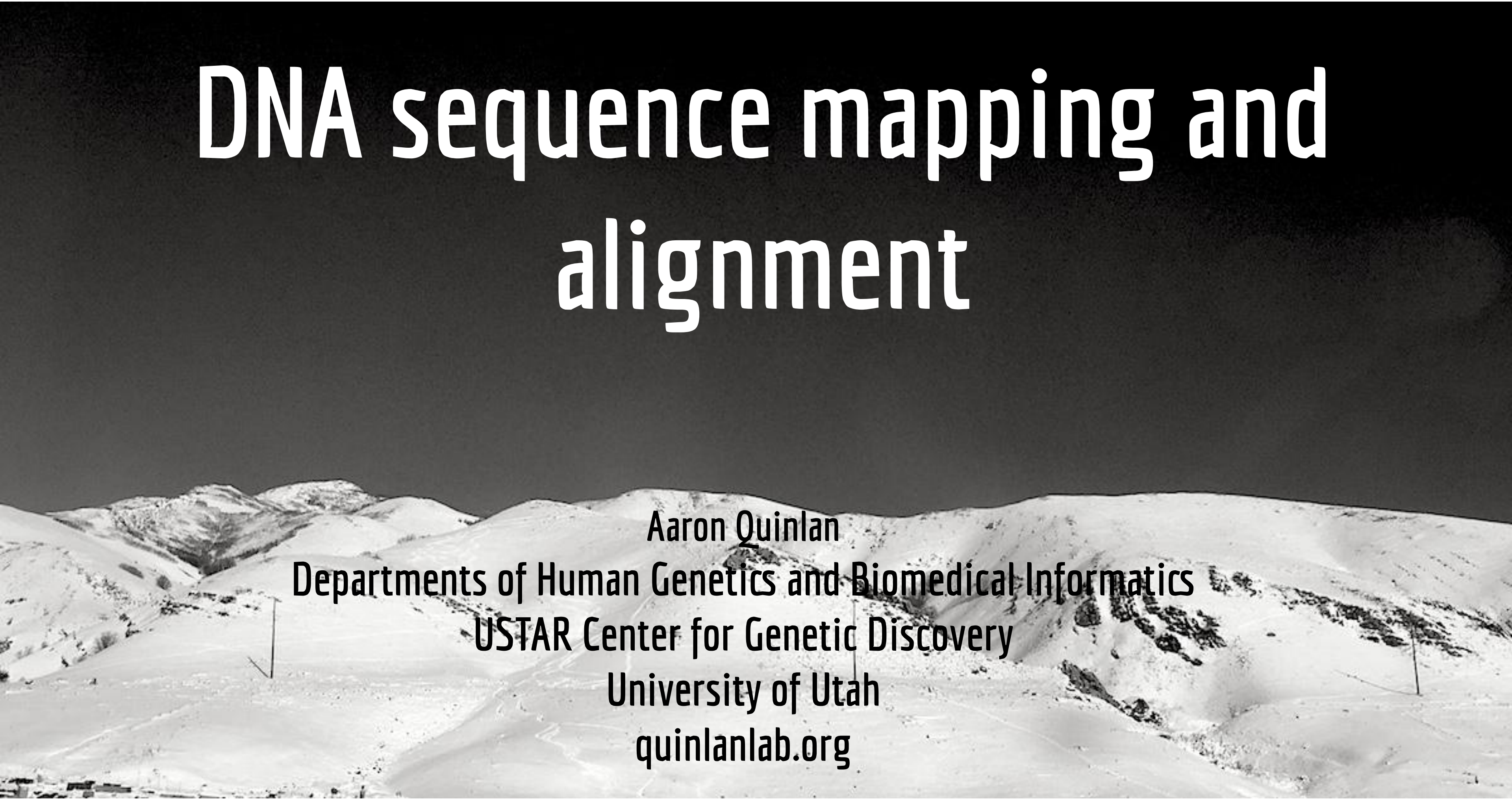
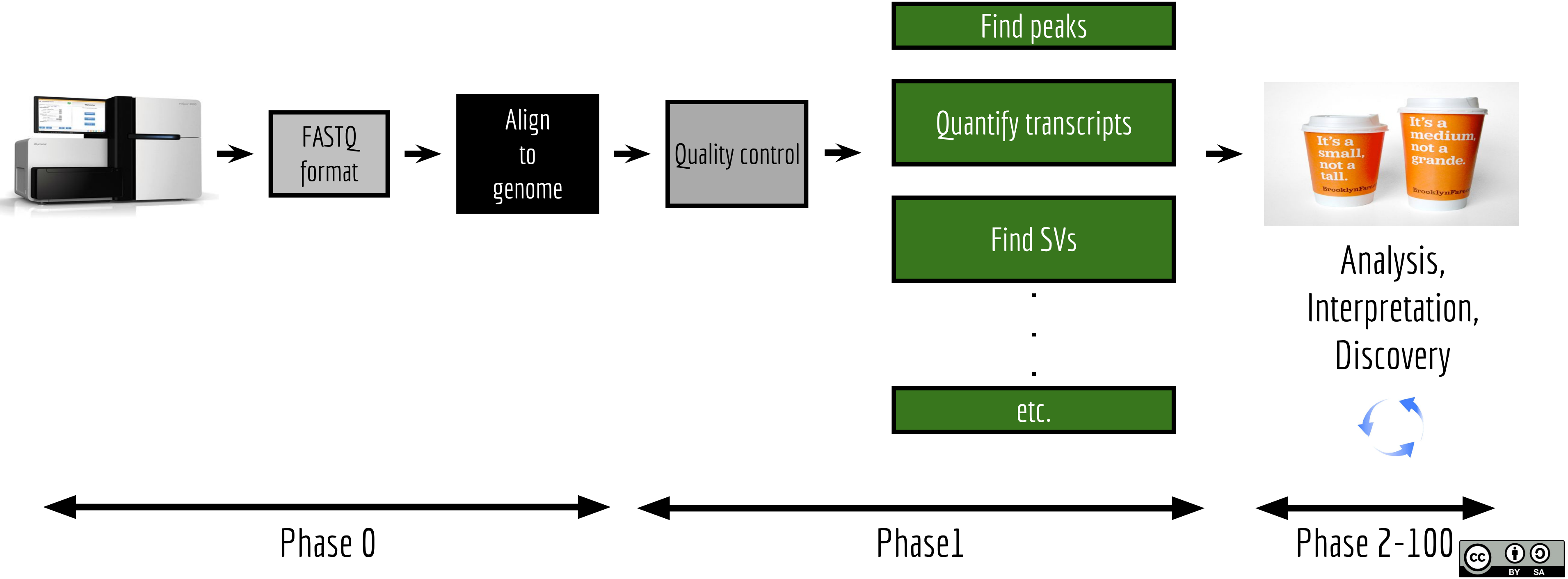
DNA sequence mapping and alignment



Aaron Quinlan
Departments of Human Genetics and Biomedical Informatics
USTAR Center for Genetic Discovery
University of Utah
quinlanlab.org

Alignment is central to most genomic research



FASTQ format

Align to genome

Quality control

Find peaks

Quantify transcripts

Find SVs

.
.
.

etc.

Analysis, Interpretation, Discovery

Phase 0

Phase1

Phase 2-100

The problems

- The human genome is big. Oh yeah, it’s complex too.
- Sequencers can produce 1 billion reads / run.
- But they make mistakes. Frequently.
- **Accurate alignment takes time, but it’s worth it.**
    - Shortcuts lead to artifacts
- Alignment strategy is highly nuanced, depending on experimental context

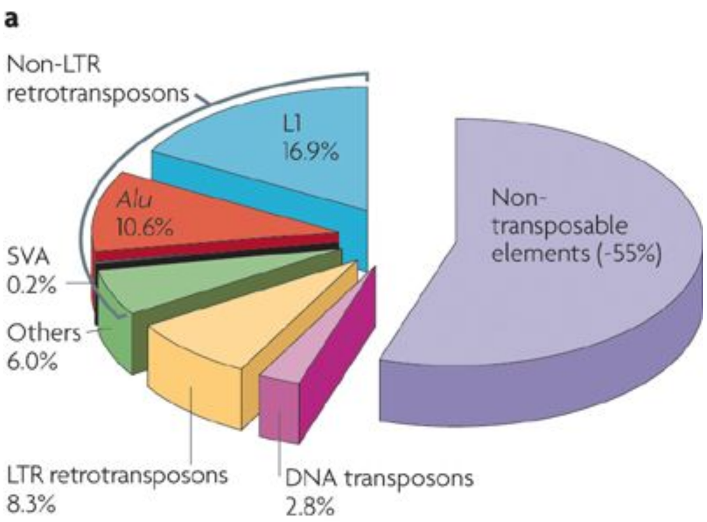
We have FASTQ files. Now what?

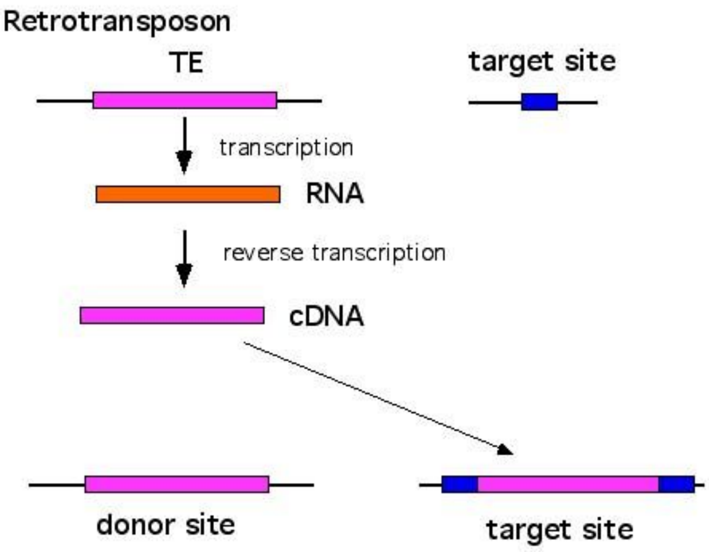
- Need to find a home for every read in the file.

- Must get the alignment just right. Else problems.

- Must choose the right tool for the experiment.

Problem: Half of the human genome is comprised of repeats





McClintock's "jumping genes" in maize

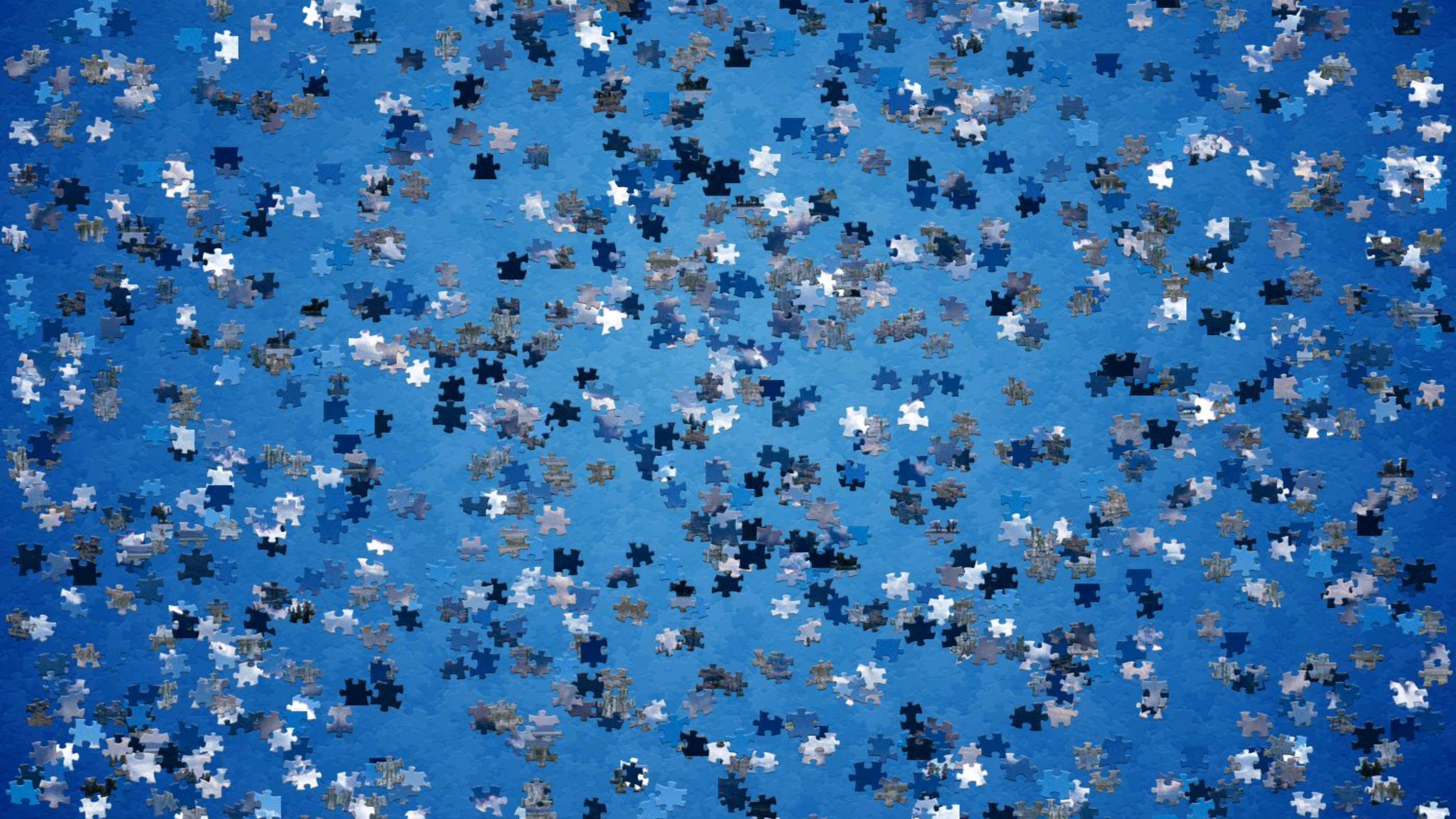
Retrotransposons use a "copy/paste" mechanism
DNA transposons use a "cut/paste" mechanism

http://www.nature.com/nrg/journal/v10/n10/pdf/nrg2640.pdf

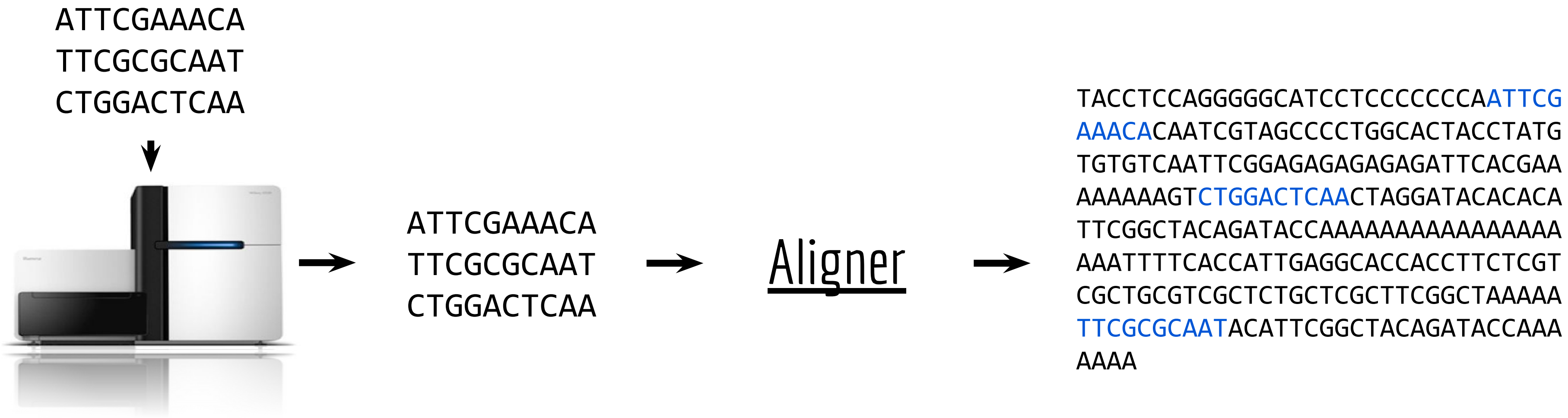
Problem: Half of the human genome is comprised of repeats

```
taaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac
cctaacccaaccctaaccctaaccctaaccctaaccctaaccctaacccc
taaccctaaccctaaccctaaccctaacctaaccctaaccctaaccctaa
cccccctaaccctaaccctaaccctaacccctaaccctaaccctaaaccc
ccctaaaccctaaccctaaccctaaccctaaccctaaccccaaccccaac
cccaaccccaaccccaaccccaaccctaacccctaaccctaaccctaacc
ctaccctaaccctaaccctaaccctaaccctaaccctaacccctaacccc
taaccctaaccctaaccctaaccctaaccctaaccctaacccctaaccct
aaccctaaccctaaccctcgcggtaccctcagccggcccgcccgcccggg
tctgacctgaggagaactgtgctccgccttcagagtaccaccgaaatctg
tgcagaggacaacgcagctccgccctcgcggtgctctccgggtctgtgct
gaggagaacgcaactccgccggcgcaggcgcagagaggcgcgccgcgccg
gcgcaggcgcagacacatgctagcgcgtcggggtggaggcgtggcgcagg
cgcagagaggcgcgccgcgccggcgcaggcgcagagacacatgctaccgc
gtccaggggtggaggcgtggcgcaggcgcagagaggcgcaccgcgccggc
gcaggcgcagagacacatgctagcgcgtccaggggtggaggcgtggcgca
ggcgcagagacgcaagcctacgggcgggggttgggggggcgtgtgttgca
ggagcaaagtcgcacggcgccgggctggggcgggggggagggtggcgccgt
gcacgcgcagaaactcacgtcacggtggcgcggcgcagagacgggtagaa
```

(first bit of human chromosome 1)



Best case scenario: an error-free sequencing technology



ATTCGAAACA
TTCGCGCAAT
CTGGACTCAA

ATTCGAAACA
TTCGCGCAAT
CTGGACTCAA

Aligner

```
TACCTCCAGGGGGCATCCTCCCCCCCAATTCG
AAACACAATCGTAGCCCCTGGCACTACCTATG
TGTGTCAATTCGGAGAGAGAGAGATTCACGAA
AAAAAAGTCTGGACTCAACTAGGATACACACA
TTCGGCTACAGATACCAAAAAAAAAAAAAAAA
AAATTTTCACCATTGAGGCACCACCTTCTCGT
CGCTGCGTCGCTCTGCTCGCTTCGGCTAAAAA
TTCGCGCAATACATTCGGCTACAGATACCAAA
AAAA
```

Computers are rather good at finding ***exact*** matches.

Think Google.

Reality check. Errors happen. Frequently.

ATTCGAAACA

→ ATTTGAAACA → Aligner →

```
TACCTCCAGGGGGCATCCTCCCCCCCAATTCG
AAACACAATCGTAGCCCCTGGCACTACCTATG
TGTGTCAATTCGGAGAGAGAGAGATTGGAAAC
AAAAAAGTGCTACAGATACCACTAGGATACAC
ACATTCGGCTACAGATACCAAAAAAAAAAAAA
AAAAAATTTTCACCATTGAGGCACCACCTTCT
CGTCGCTGCGTCGCTCTGCTCGCGGCTAAAAA
ATTAGAAACAACATTCGGCTACAGATACCAAA
ATTT
```

"Fuzzy" matching is much more computationally expensive.

Think Google's "Did you mean..."

Sequence ***mapping*** versus ***alignment***

Mapping: (quickly) find the best possible loci to which a sequence could be aligned

Alignment: for each locus to which a sequence can be mapped, determine the optimal base by base alignment of the query sequence to the reference sequence

Hash-based mapping:

Step1: hash/index the genome

Toy genome
(16 bp)

CATGGTCATTGGTTCC

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3
GGT	4

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3
GGT	4
GTC	5

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1
ATG	2
TGG	3
GGT	4
GTC	5
TCA	6

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3
GGT	4
GTC	5
TCA	6

Hash-based mapping:

Step1: hash/index the genome

CATGGTCATTGGTTCC

k = 3

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

Complete hash/kmer index of our toy genome (forward strand only)

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

→ Read TGGTCA

kmer index is used to quickly find candidate alignment locations in genome.

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

→

Read TGGTCA

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

→

Read TGGTCA

Hash match

Hash matches 3,10

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

→

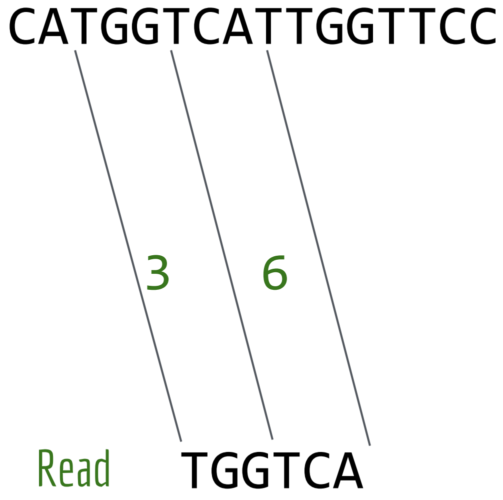
Read TGGTCA

Hash matches 3,10,6

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC



Read TGGTCA

Hash matches 3,10,6

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

Okay, that was a bit easy because the read and the reference exactly matched. What about if there is a sequencing error or a genetic variant in the read?

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

→ Read TGGTCT

kmer index is used to quickly find candidate alignment locations in genome.

Hash-based mapping:

Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

Hash match

→

Read TGGTCT

Hash matches 3,10

Hash-based mapping:

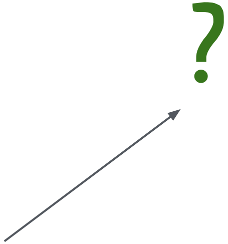
Step2: use the index to map (i.e., find alignment locations) reads

Toy genome CATGGTCATTGGTTCC

Kmer/Hash	Genome Positions
CAT	1,7
ATG	2
TGG	3,10
GGT	4,11
GTC	5
TCA	6
ATT	8
TTG	9
GTT	12
TTC	13
TCC	14

→

Read TGGTCT



Hash matches 3,10

Mapping quality (MAPQ)

What is the probability that the sequence should be mapped here and only here?

MAPQ also uses the Phred (log) scale:

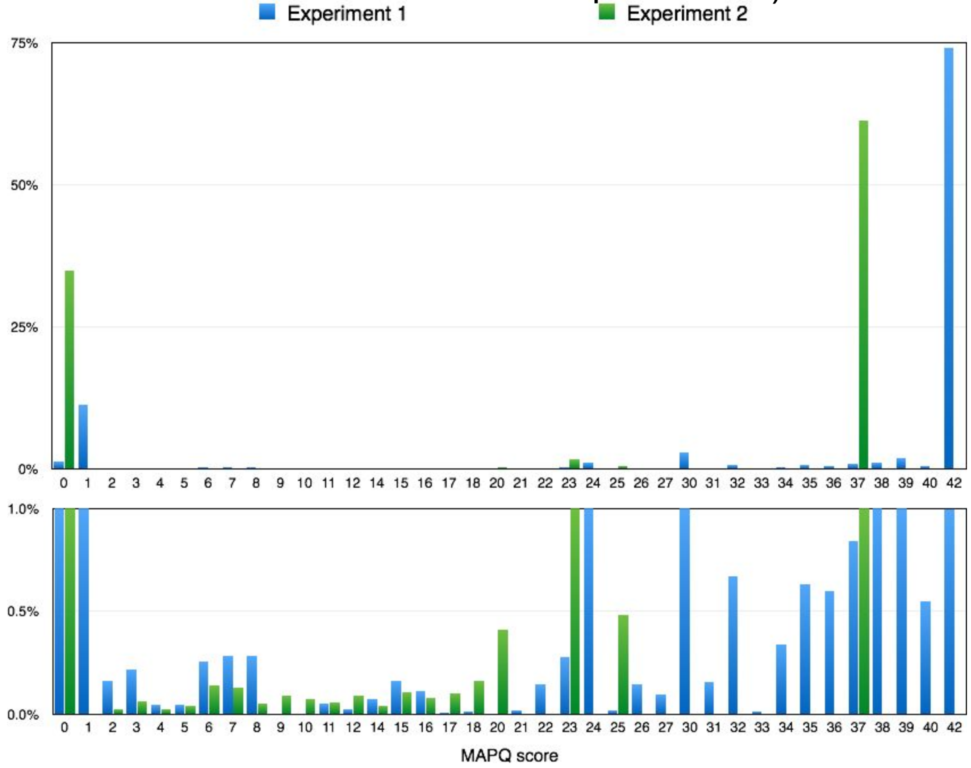
$$MAPQ = -10 * \log_{10}(P_{map\_loc\_wrong})$$

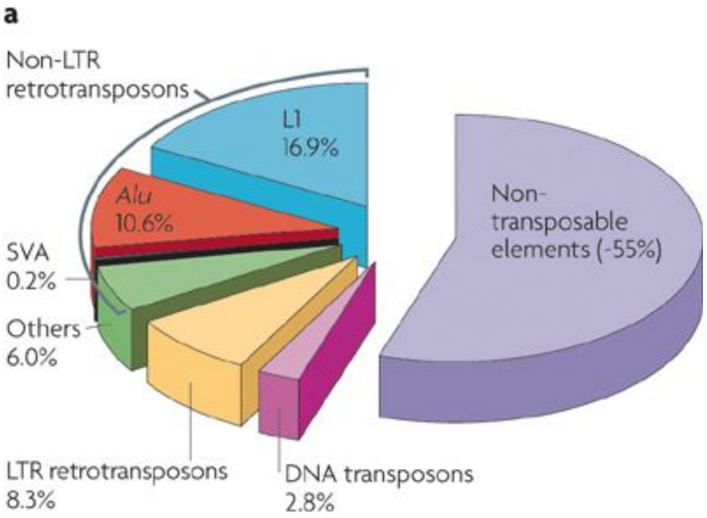
$(P_{map\_loc\_wrong})$	$\log_{10}(P_{map\_loc\_wrong})$	MAPQ
1	0	0
0.1	-1	10
0.01	-2	20
0.001	-3	30
0.0001	-4	40

Mapping quality (MAPQ)

(Bowtie, single-end)

(BWA, paired-end)





http://www.acgt.me/blog/2014/12/16/understanding-mapq-scores-in-sam-files-does-37-42

Edit distance

How many edits (changes) must be made to a word or kmer to make it match (align) to another word or kmer?

```
CURLED
HURLED
```

→ Edit distance = 1. Substitute C for H

```
SHORT
SHO-T
```

→ Edit distance = 1. Delete R

```
TGTTACGG
GGTTGACTA
```

?

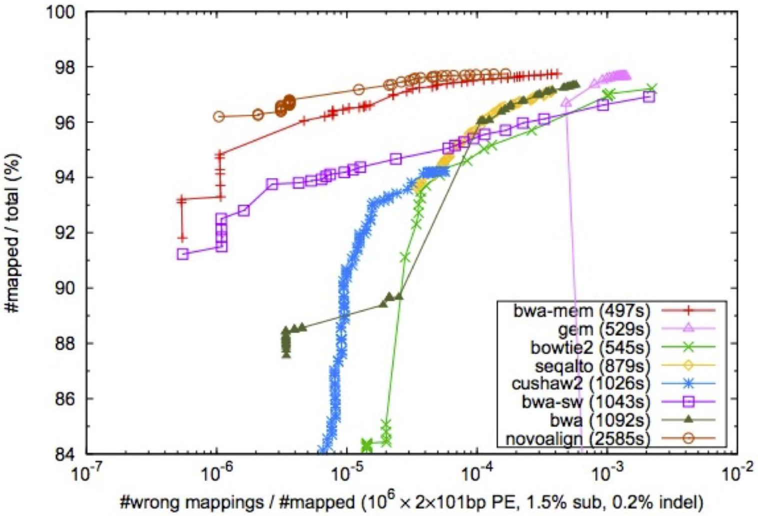
```
TG-TT-ACGG
-GGTTGACTA
```

Edit distance = 5

```
TGTT-ACGG
GGTTGACTA
```

Edit distance = 4

BWA-MEM: never "published" ; widely used.



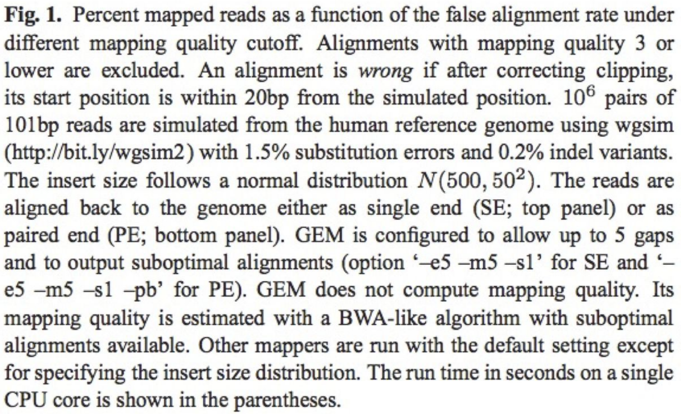
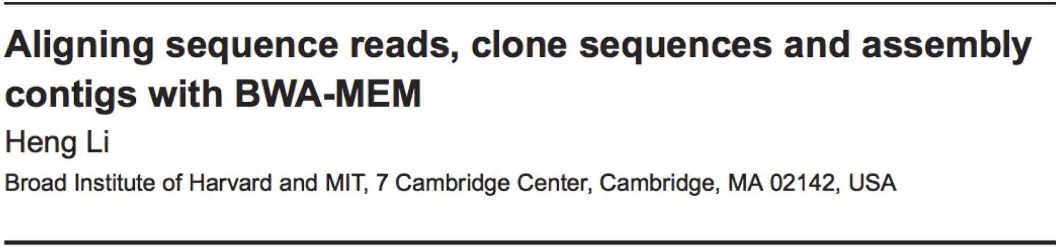


Fig. 1. Percent mapped reads as a function of the false alignment rate under different mapping quality cutoff. Alignments with mapping quality 3 or lower are excluded. An alignment is *wrong* if after correcting clipping, its start position is within 20bp from the simulated position. 10^6 pairs of 101bp reads are simulated from the human reference genome using wgsim (http://bit.ly/wgsim2) with 1.5% substitution errors and 0.2% indel variants. The insert size follows a normal distribution $N(500, 50^2)$. The reads are aligned back to the genome either as single end (SE; top panel) or as paired end (PE; bottom panel). GEM is configured to allow up to 5 gaps and to output suboptimal alignments (option '–e5 –m5 –s1' for SE and '–e5 –m5 –s1 –pb' for PE). GEM does not compute mapping quality. Its mapping quality is estimated with a BWA-like algorithm with suboptimal alignments available. Other mappers are run with the default setting except for specifying the insert size distribution. The run time in seconds on a single CPU core is shown in the parentheses.



Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

Heng Li

Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA

https://arxiv.org/pdf/1303.3997v2.pdf

BWA-MEM

Unaligned
Sample Data
In FASTQ (SE or PE)

```
@seq1
ATTCGAAACA...
+
DDED88(999...
@seq2
CCCCGTTTCA...
+
AAC887BBAC...
```

Reference genome (FASTA)

```
>chr1
TACCTCCAGGGGGCATCCTCCCCCCCAATTCG
AAACACAATCGTAGCCCCTGGCACTACCTATG
TGTGTCAATTCGGAGAGAGAGAGATTCACGAA
AAAAAAGTCTGGACTCAACTAGGATACACACA
TTCGGCTACAGATACCAAAAAAAAAAAAAAAA
AAATTTTCACCATTGAGGCACCACCTTCTCGT
CGCTGCGTCGCTCTGCTCGCTTCGGCTAAAAA
TTCGCGCAATACATTCGGCTACAGATACCAAA
```

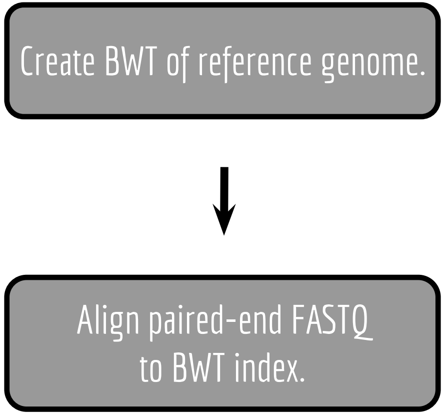
BWA MEM

Aligned
Sample Data in
SAM format

```
seq1    99      1       3666901       60
149M    =       3666935       185
ATTCGAAACA...DDED88(999   MC:Z:151M
MD:Z:149      RG:Z:15-0017315_1   NM:i:0
MQ:i:60       AS:i:149      XS:i:44
seq2    147     1       3666935       60
151M    =       3666901       -185
CCCCGTTTCA...AAC887BBAC...MC:Z:149M
MD:Z:151      RG:Z:15-0017315_1   NM:i:0
MQ:i:60       AS:i:151      XS:i:59
```

BWA-MEM workflow

This takes a long time, but you do it once



Create BWT of reference genome.

Align paired-end FASTQ to BWT index.

```
$ bwa index grch38.fa
```

Output is in SAM format. Use multiple threads if you have a computer with multiple CPUs.

```
$ bwa mem -t 16 grch38.fa 1.fq 2.fq > sample.sam
```

Let’s get our hands dirty

https://gist.github.com/arq5x/4716b710f967998e9feaeb134e0ebe2b#file-alignment-md