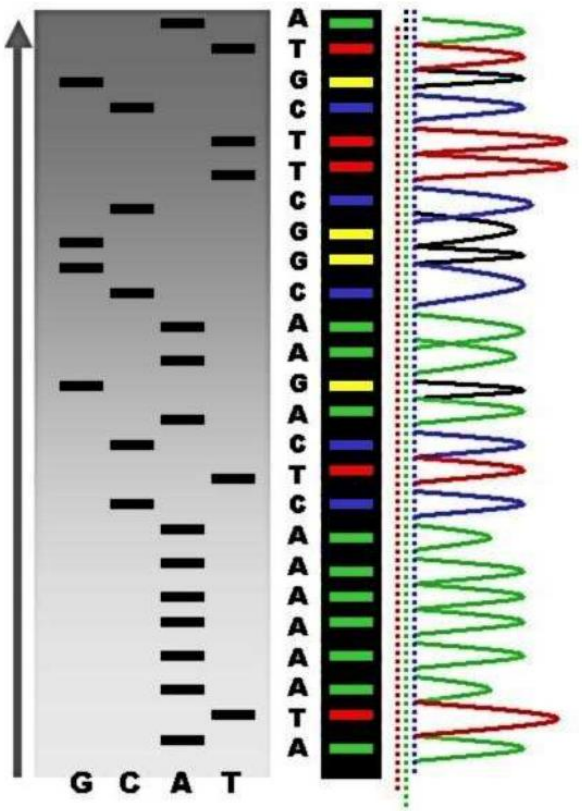
BIO306: Bioinformatics

Lecture 2

NGS and Reads mapping

Wenfei JIN PhD
jinwf@sustc.edu.cn
Department of Biology, SUSTech

Sanger Sequencing



Progression of Sequencing Reaction

dideoxynucleotides (ddNTPs)

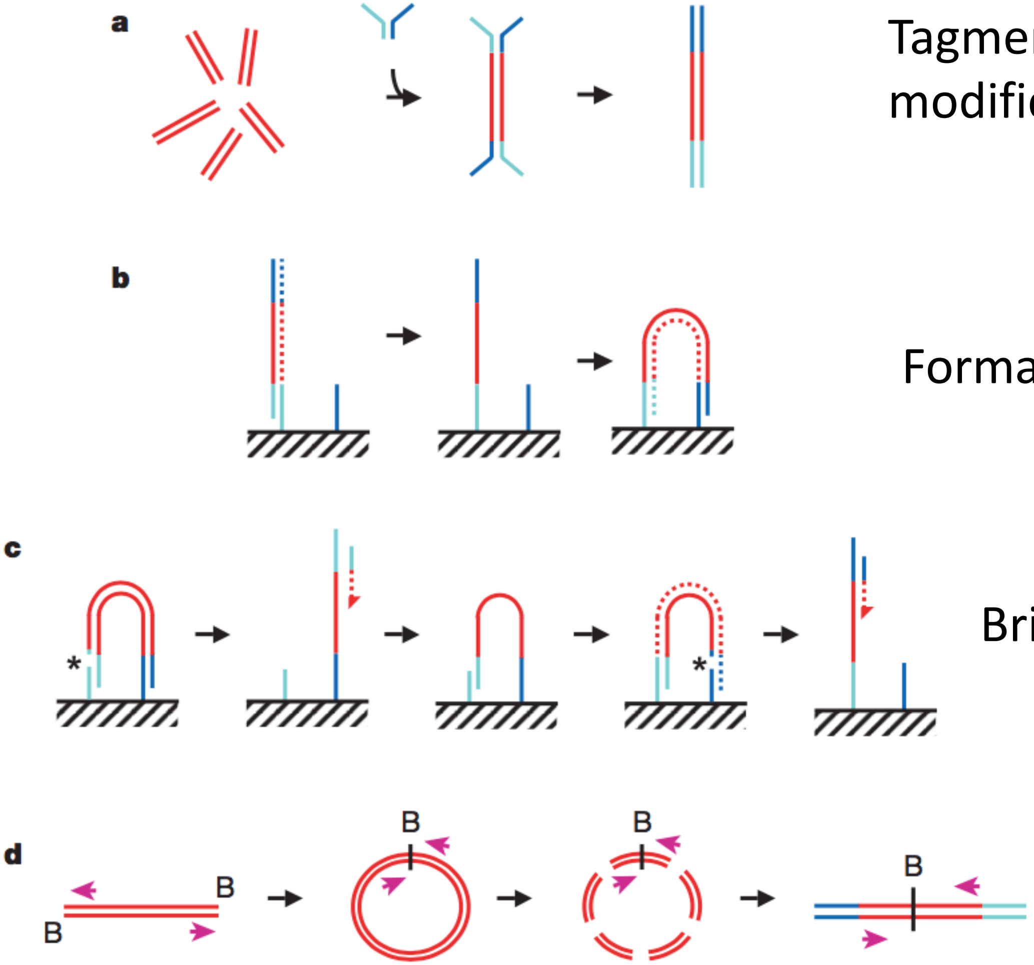
What is Next generation sequencing (NGS)?

- High-throughput sequencing

- Massively parallel sequencing

- Illumina dye sequencing as example

Illumina sequencing showed in original nature paper



Tagmentation and molecular modifications

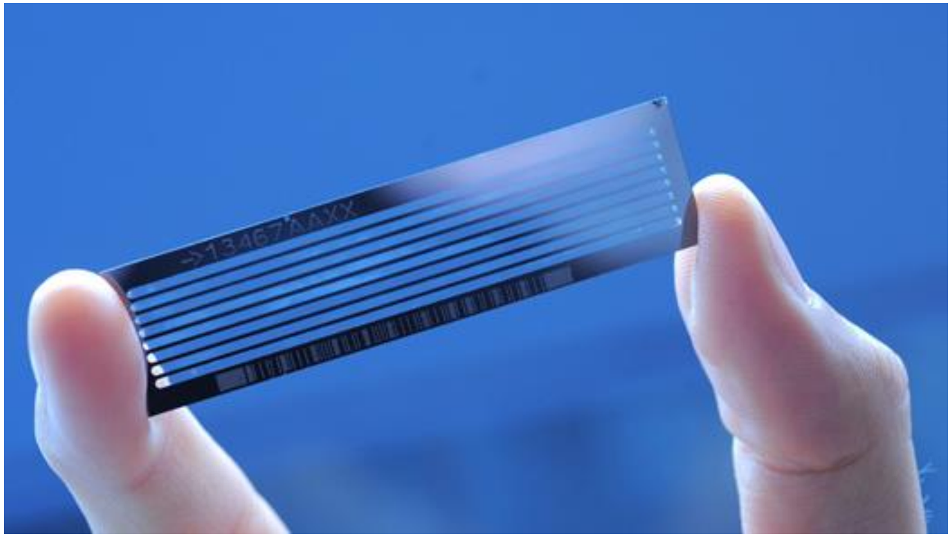
Formation of bridge

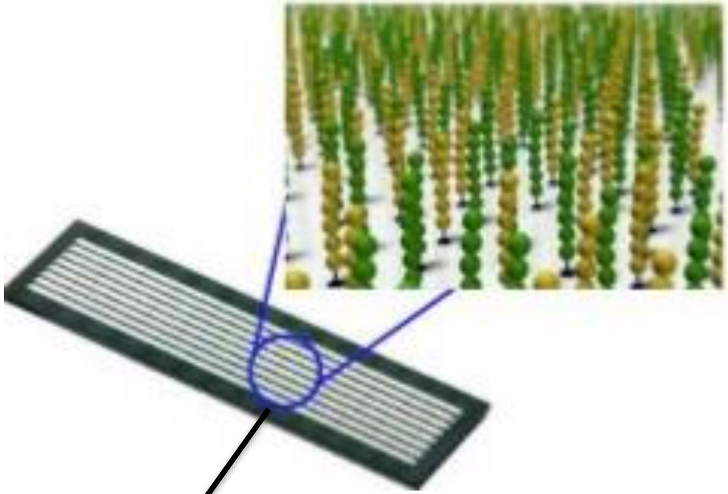
Bridge amplification

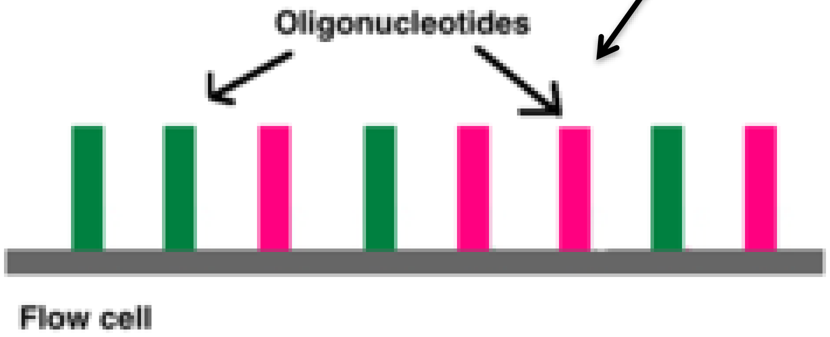
Paired-end sequencing

2008 Bentley, Nature

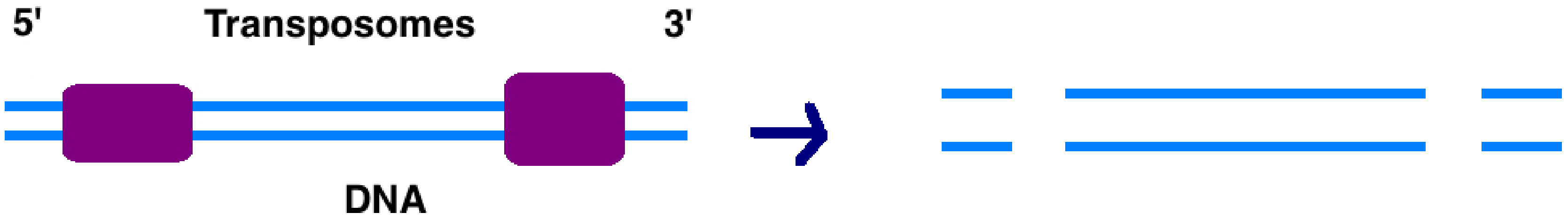
Flow cell

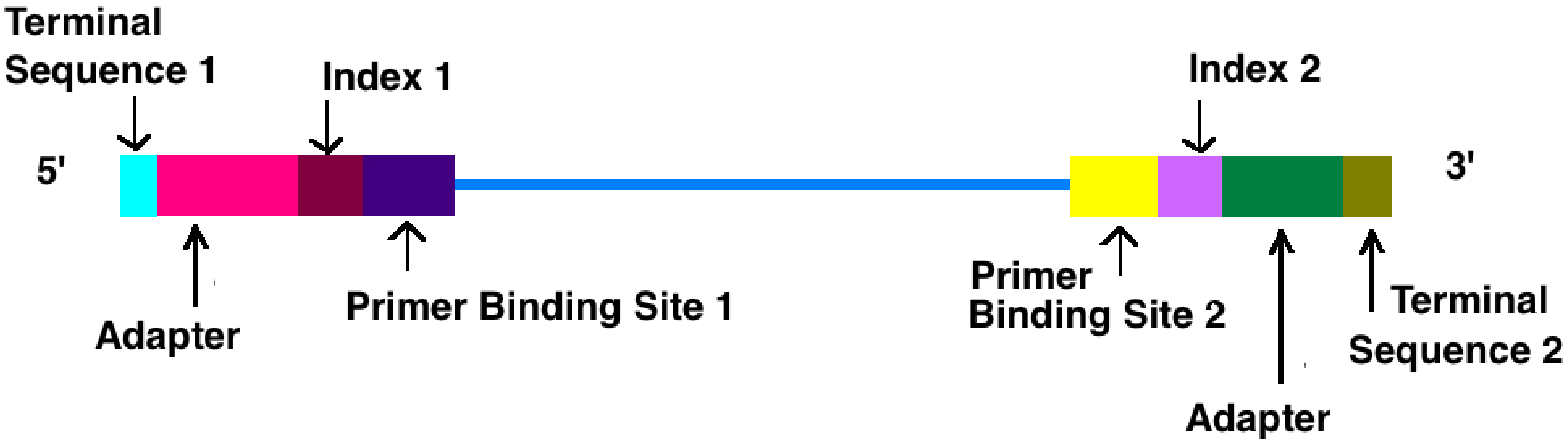




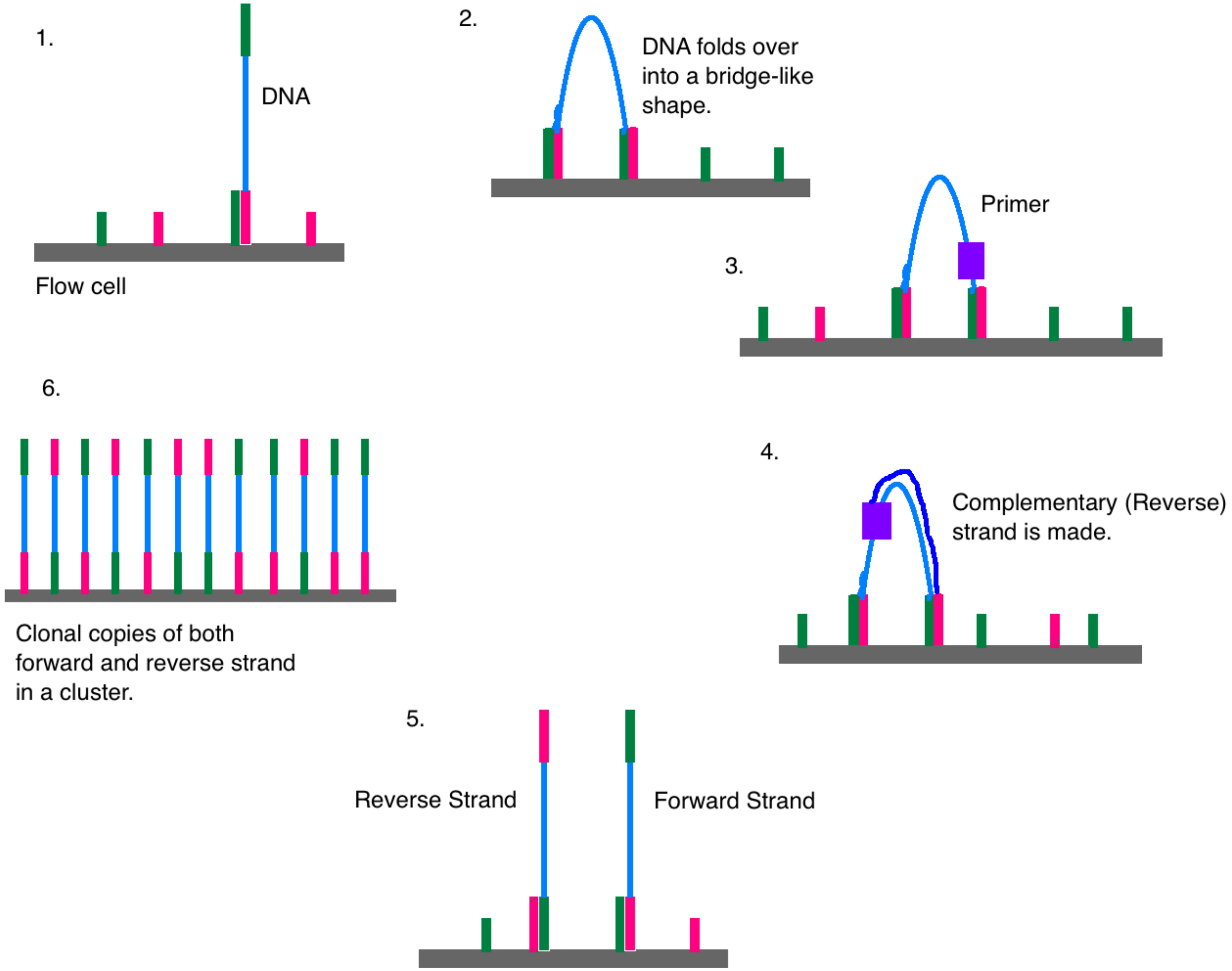


Library preparation

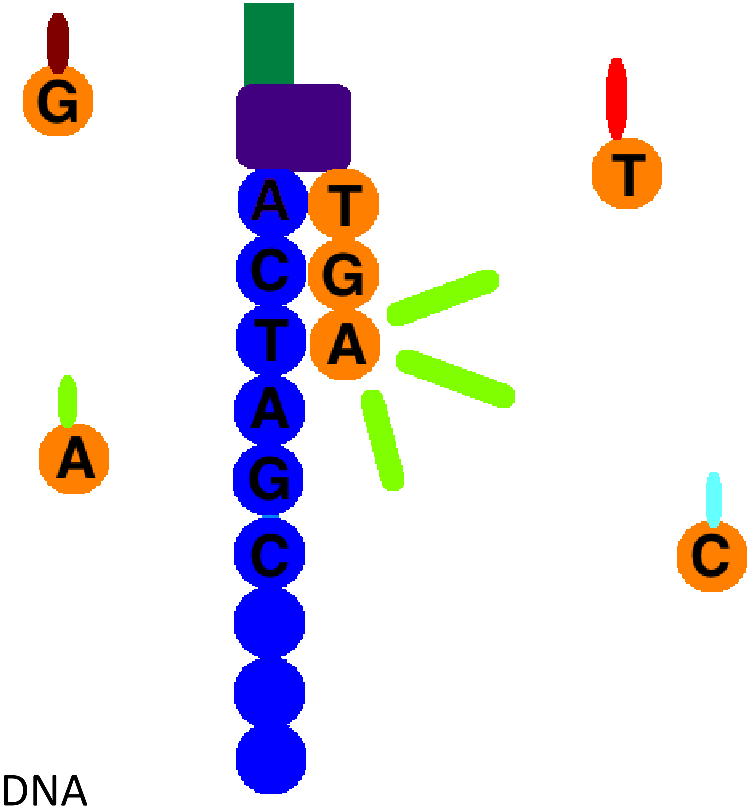




Amplification/Cluster generation

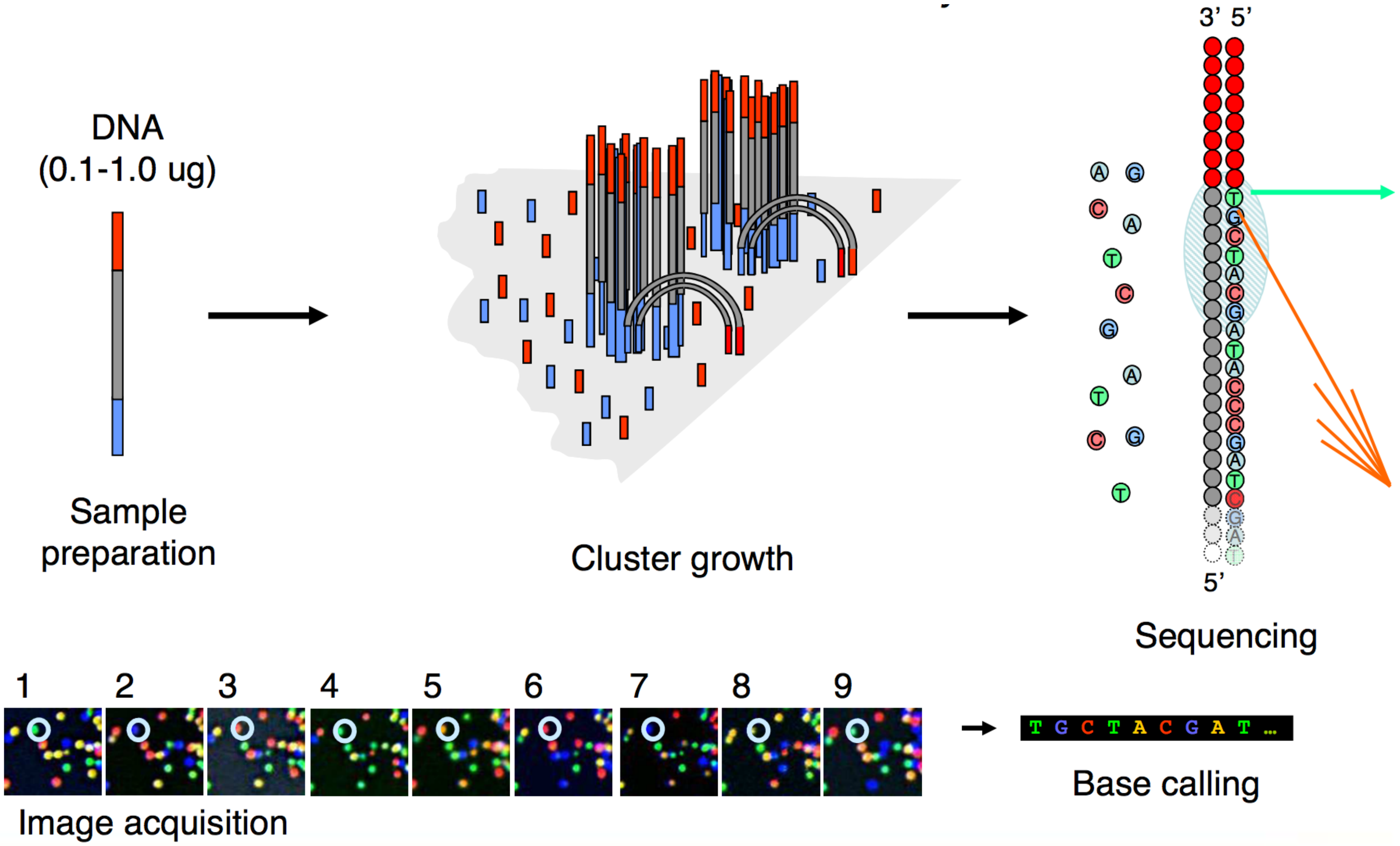

1.
DNA
Flow cell
2.
DNA folds over into a bridge-like shape.
3.
Primer
4.
Complementary (Reverse) strand is made.
5.
Reverse Strand
Forward Strand
6.
Clonal copies of both forward and reverse strand in a cluster.

Sequence by Synthesis (1)



fluorescently tagged nucleotides to the DNA strand

Sequence by Synthesis (2)



DNA
(0.1-1.0 ug)

Sample preparation

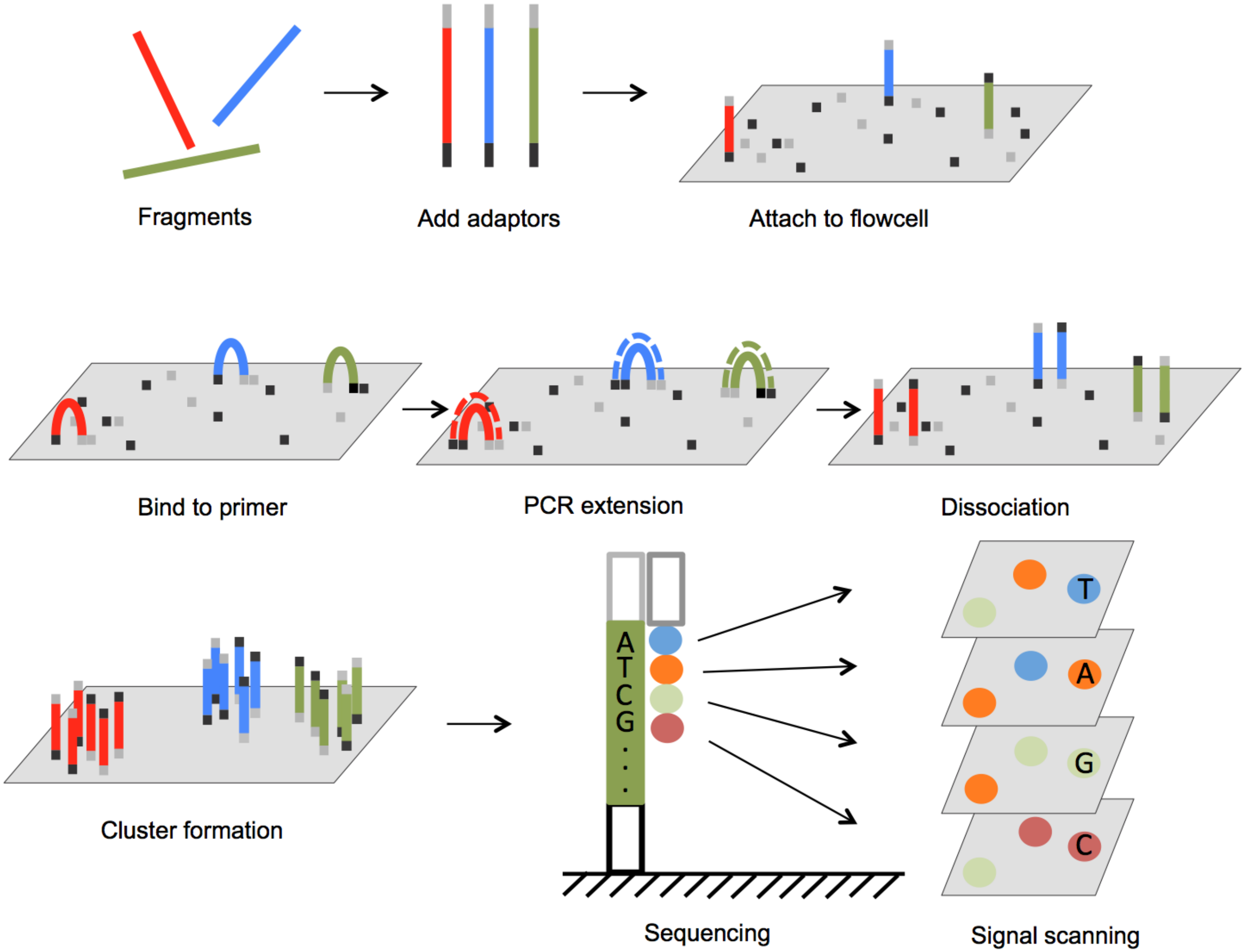
Cluster growth

Sequencing

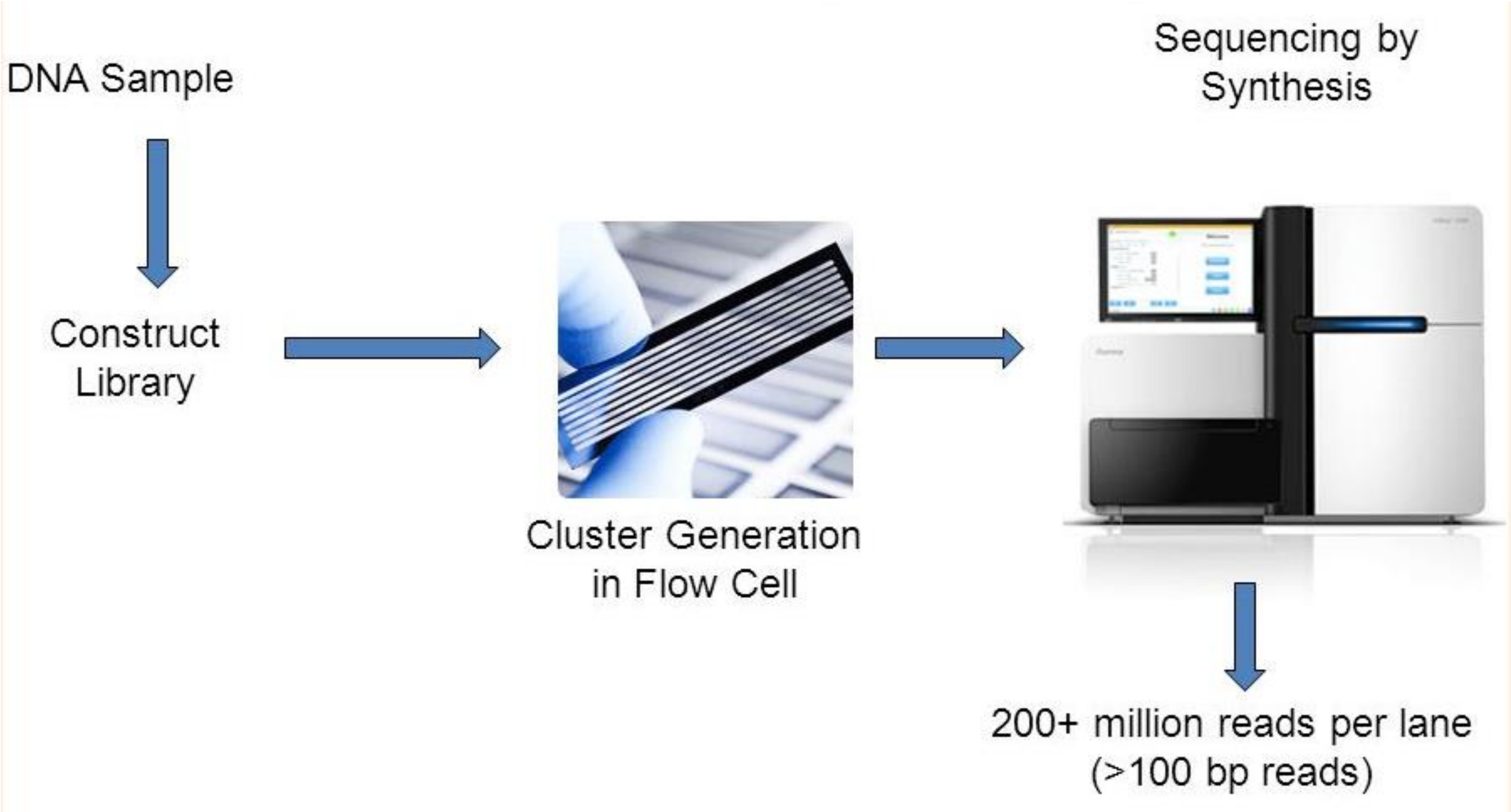
Image acquisition

Base calling

Schematic of Illumina dye sequencing



Illumina sequencing



DNA Sample

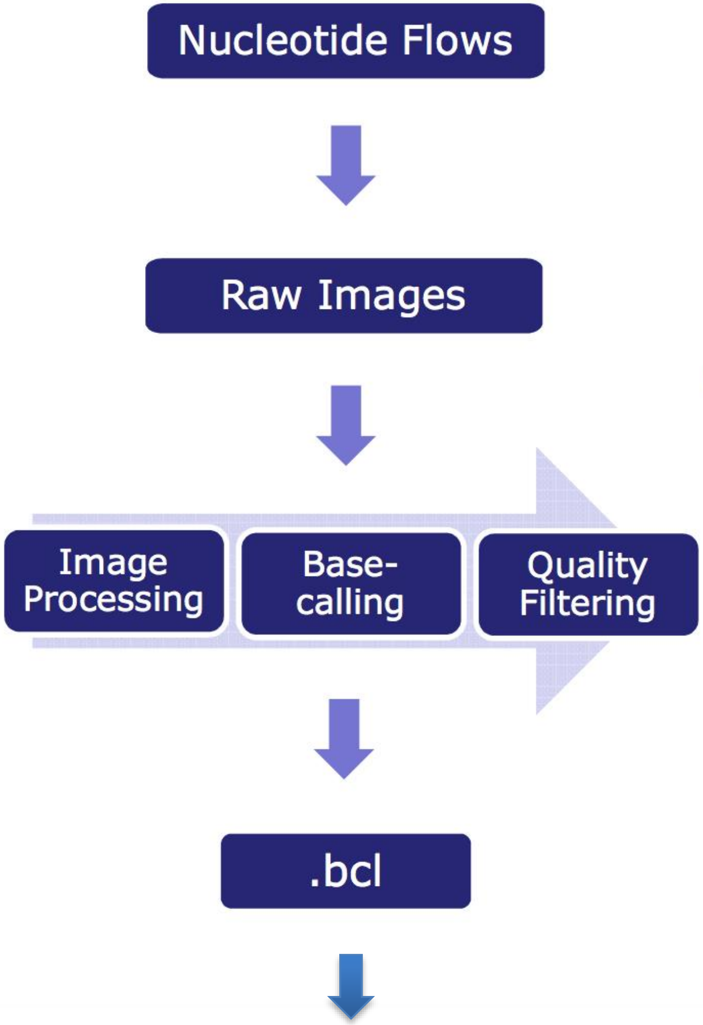
Construct Library

Cluster Generation in Flow Cell

Sequencing by Synthesis

200+ million reads per lane (>100 bp reads)

Data Processing



Nucleotide Flows

Raw Images

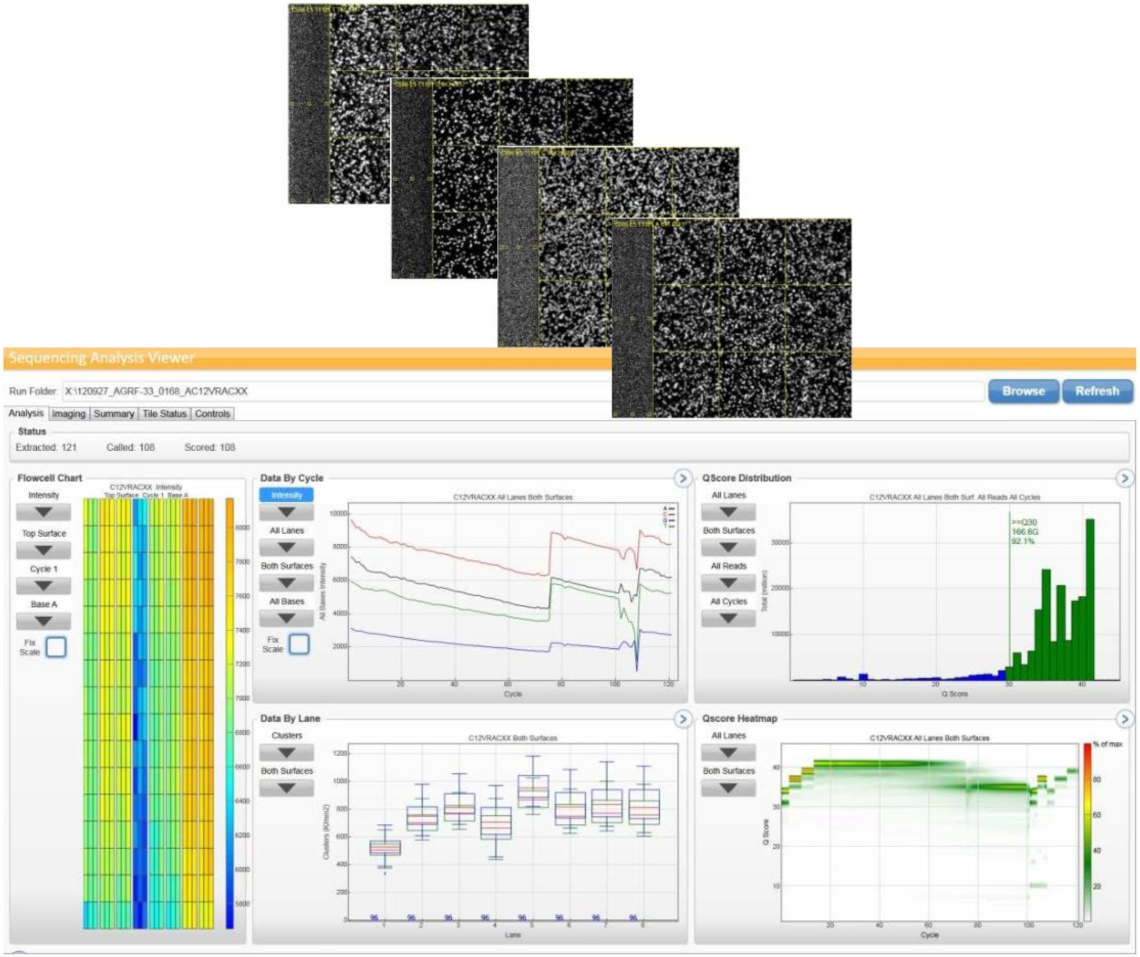
Image Processing

Base-calling

Quality Filtering

.bcl

Fastq



Sequencing Analysis Viewer

Run Folder: X:\120927_AGRF-33_0168_AC12VRACXX

Status

Extracted: 121 Called: 108 Scored: 108

Fastq format

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAA
+
!''*((((***+))%%%++)(%%%%).1***+*''))**55CCF>>>>>>
```

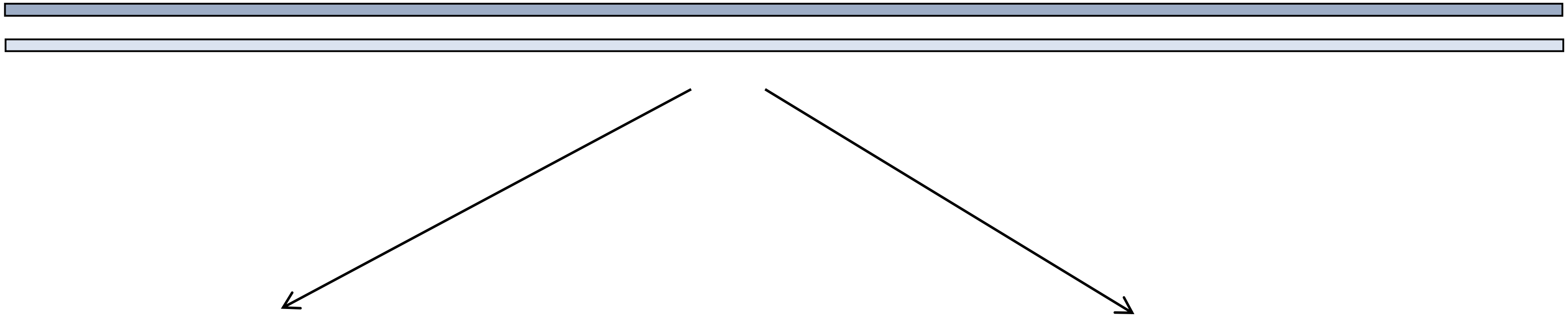
4 lines per sequence/read
Line 1 begins with a @ and is followed by ID
Line 2 sequence letters.
Line 3 begins with a ‘+’ is optionally followed by any characters
Line 4 quality values for the sequence in Line 2,
must contain the same number of symbols as letters in the sequence.
quality score are each encoded with a single ASCII character for brevity.

Characters of NGS data

- Short reads
  - Illumina (36 – 300bp)
  - SoLID (75bp max)
  - Ion Torrent (200-300bp max – currently...)
  - Roche 454 – 400-800bp
- Data set is large (multiple coverage)
  - Millions or even billions reads

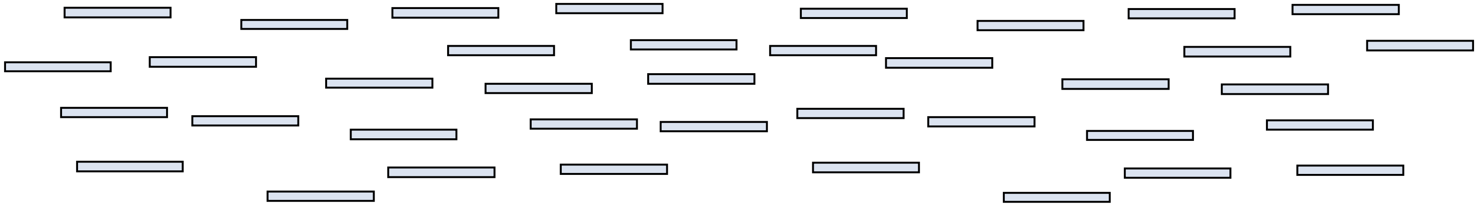
Reads alignment

Alignment of reads to a reference



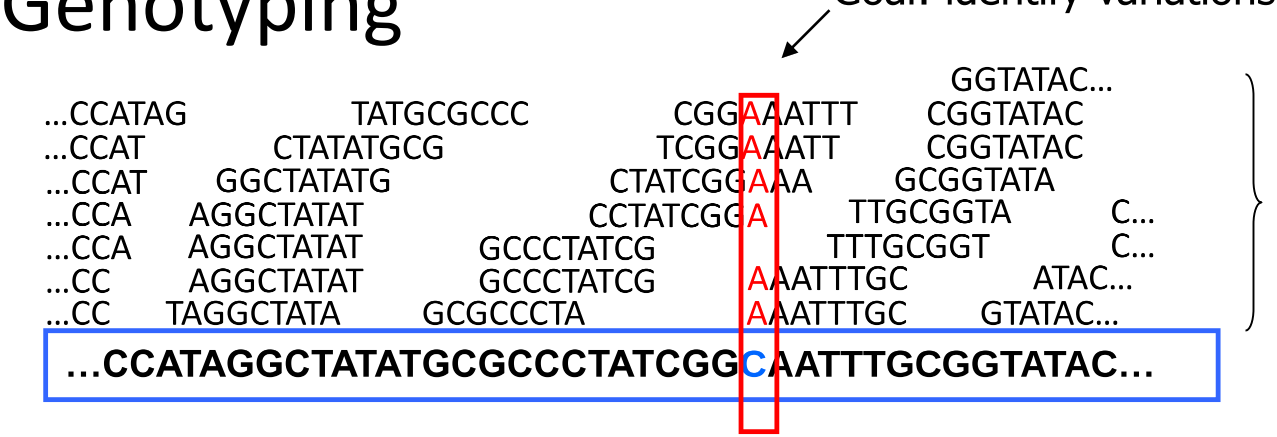
..ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA.. Reference

..ACTGGGTCATCGTACGATCGATAGATCGATCGATCGCTAGCTAGCTA.. Sample



Short Read Applications

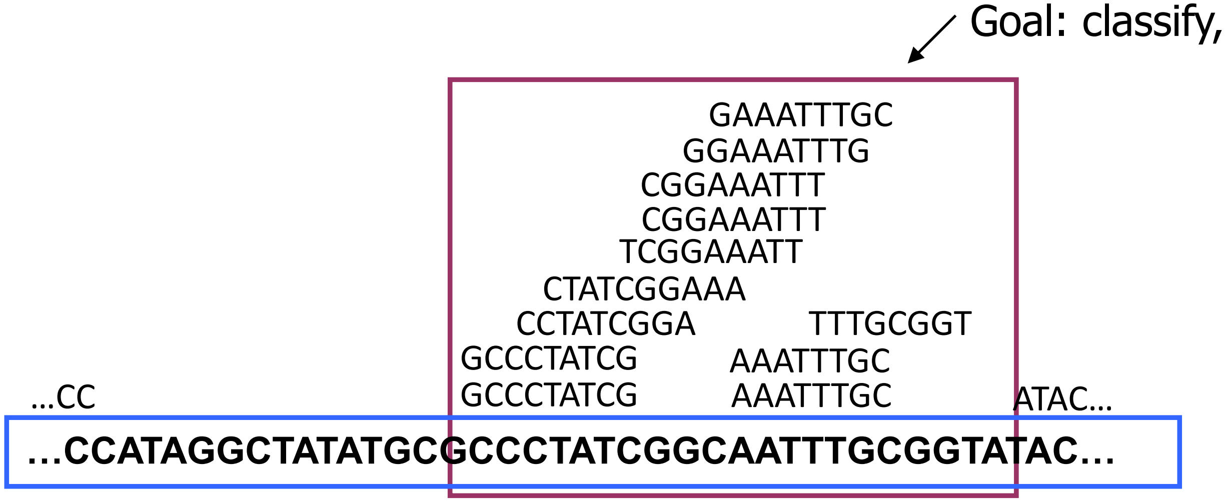
- Genotyping



Goal: identify variations

```
                                                  GGTATAC...
...CCATAG        TATGCGCCC        CGGAAATTT     CGGTATAC
...CCAT        CTATATGCG         TCGGAAATT     CGGTATAC
...CCAT     GGCTATATG          CTATCGGAAA    GCGGTATA
...CCA     AGGCTATAT          CCTATCGGA    TTGCGGTA       C...
...CCA     AGGCTATAT      GCCCTATCG       TTTGCGGT        C...
...CC      AGGCTATAT      GCCCTATCG       AAATTTGC      ATAC...
...CC     TAGGCTATA     GCGCCCTA          AAATTTGC     GTATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

- RNA-seq, ChIP-seq, Methyl-seq



Goal: classify, measure significant peaks

```
                         GAAATTTGC
                        GGAAATTTG
                      CGGAAATTT
                      CGGAAATTT
                     TCGGAAATT
                 CTATCGGAAA
                CCTATCGGA        TTTGCGGT
             GCCCTATCG        AAATTTGC
             GCCCTATCG        AAATTTGC
...CC                                      ATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

Why is short read alignment hard?

The shorter a read, the less likely it is to have a unique match to a reference sequence

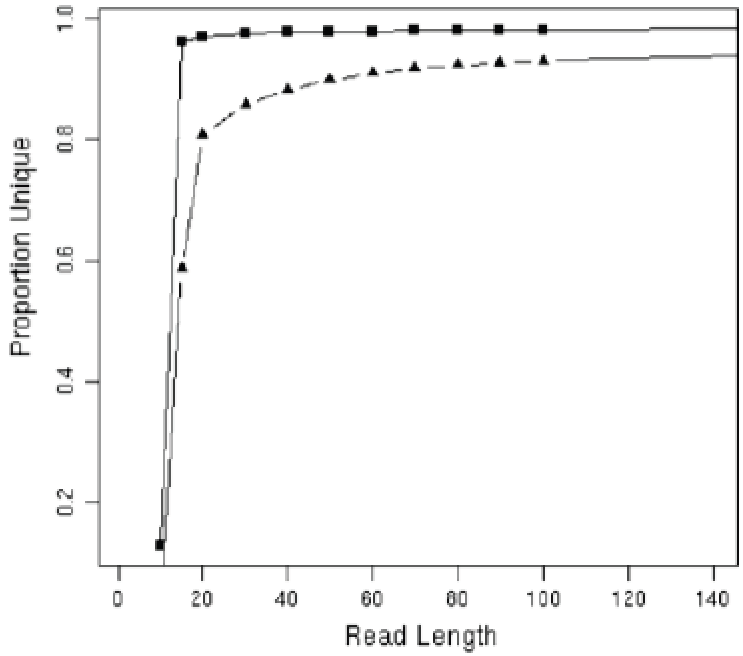


Fig. 1 The proportion of unique sequence in the *Streptococcus suis* (squares) and *Mus musculus* (triangles) genomes for varying read lengths. This graph indicates that read length has a critical affect on the ability to place reads uniquely to the genome

Why do we generate short reads?

- Sanger reads lengths ~ 800-2000bp

- Generally we define short reads as anything below 200bp
    - Illumina (100bp – 250bp)
    - SoLID (75bp max)
    - Ion Torrent (200-300bp max – currently...)
    - Roche 454 – 400-800bp

- Even with these platforms it is cheaper to produce short reads (e.g. 50bp) rather than 100 or 200bp reads

- Diminishing returns:
    - For some applications 50bp is more than sufficient
        - Resequencing of smaller organisms
        - Bacterial de-novo assembly
        - ChIP-Seq
        - Digital Gene Expression profiling
        - Bacterial RNA-seq

Contents

- **Alignment algorithms for short-reads**
  - Adapting hashed seed-extend algorithms to work with shorter reads
  - Indel detection
  - Suffix/Prefix Tries
  - Other alignment considerations
  - Typical alignment pipeline
- **Assembly algorithms for short reads**
  - Effect of repeats
  - Overlap-Consensus
  - de Bruijn graphs
  - Assembly evaluation metrics
  - Typical assembly pipeline

Adapting hashed seed-extend algorithms to work with shorter reads

- Improve seed matching sensitivity
  - Allow mismatches within seed
    - BLAST
  - Allow mismatches + Adopt spaced-seed approach
    - ELAND, SOAP, MAQ, RMAP, ZOOM
  - Allow mismatches + Spaced-seeds + Multi-seeds
    - SSAHA2, BLAT, ELAND2
- Above and/or Improve speed of local alignment for seed extension
  - Single Instruction Multiple Data
    - Shrimp2, CLCBio
  - Reduce search space to region around seed

Hashed seed-extend algorithms

- **2 step process**
  - Identify a match to the seed sequence in the reference
  - Extend match using sensitive (but slow) Smith-Waterman algorithm (dynamic programming)

Seed-extend algorithm

Reference sequence:

...ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA...

Short read:

GTCATCGTACGATCGATAGATCGATCGATCGGCTA

Note that the short read has 1 difference wrt to reference

Seed-extend algorithm

Reference sequence:

...ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA...

Short read:

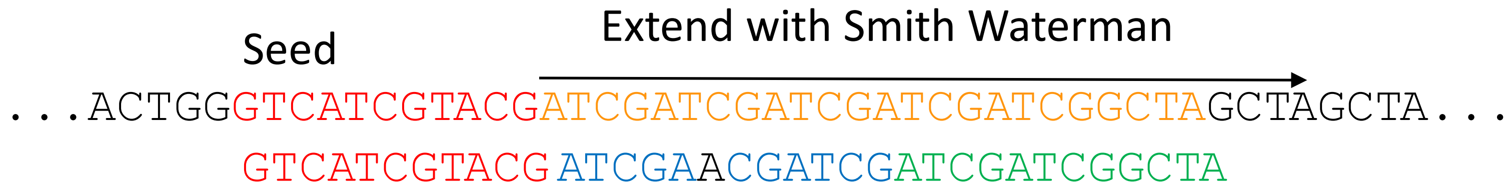
GTCATCGTACG ATCGATAGATCG ATCGATCGGCTA

11bp word 11bp word 11bp word

The algorithm will try to match each word to the reference. If there is a match at with any single word it will perform a local alignment to extend the match

Seed-extend algorithm

Reference sequence:



Extend with Smith Waterman

Seed

```
...ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA...
        GTCATCGTACG ATCGAACGATCGATCGATCGGCTA
```

Short read:

GTCATCGTACG   ATCGATAGATCG   ATCGATCGGCTA

Here the algorithm is able to match the short read with a word length of 11bp

Seed-extend algorithm

Reference sequence:

...ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA...

Short read:

GTCATCGTACGATCGATCGATCGATCGATCGGCAA

Note that the short read has 3 differences
Possibly sequencing errors, possibly SNPs

Seed-extend algorithm

Reference sequence:

...ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA...

Short read:

GTCATCGTACG	ATCGATCGATCG	ATCGATCGGCAA
11bp word	11bp word	11bp word

Note that the short read has 3 differences

Seed-extend algorithm

Reference sequence:

...ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA...

Short read:

GTCATCGTACG ATCGATCGATCG ATCGATCGGCAA

No seeds match

Therefore the algorithm would find no hits at all!

Adapting hashed seed-extend algorithms to work with shorter reads

- Improve seed matching sensitivity
  - **Allow mismatches within seed**
    - **BLAST**
  - Allow mismatches + Adopt spaced-seed approach
    - ELAND, SOAP, MAQ, RMAP, ZOOM
  - Allow mismatches + Spaced-seeds + Multi-seeds
    - SSAHA2, BLAT, ELAND2
- Above and/or Improve speed of local alignment for seed extension
  - Single Instruction Multiple Data
    - Shrimp2, CLCBio
  - Reduce search space to region around seed

Adapting hashed seed-extend algorithms to work with shorter reads

- Improve seed matching sensitivity
  - **Allow mismatches within seed**
    - **BLAST**
  - **Allow mismatches + Adopt spaced-seed approach**
    - **ELAND, SOAP, MAQ, RMAP, ZOOM**
  - Allow mismatches + Spaced-seeds + Multi-seeds
    - SSAHA2, BLAT, ELAND2
- Above and/or Improve speed of local alignment for seed extension
  - Single Instruction Multiple Data
    - Shrimp2, CLCBio
  - Reduce search space to region around seed

Contents

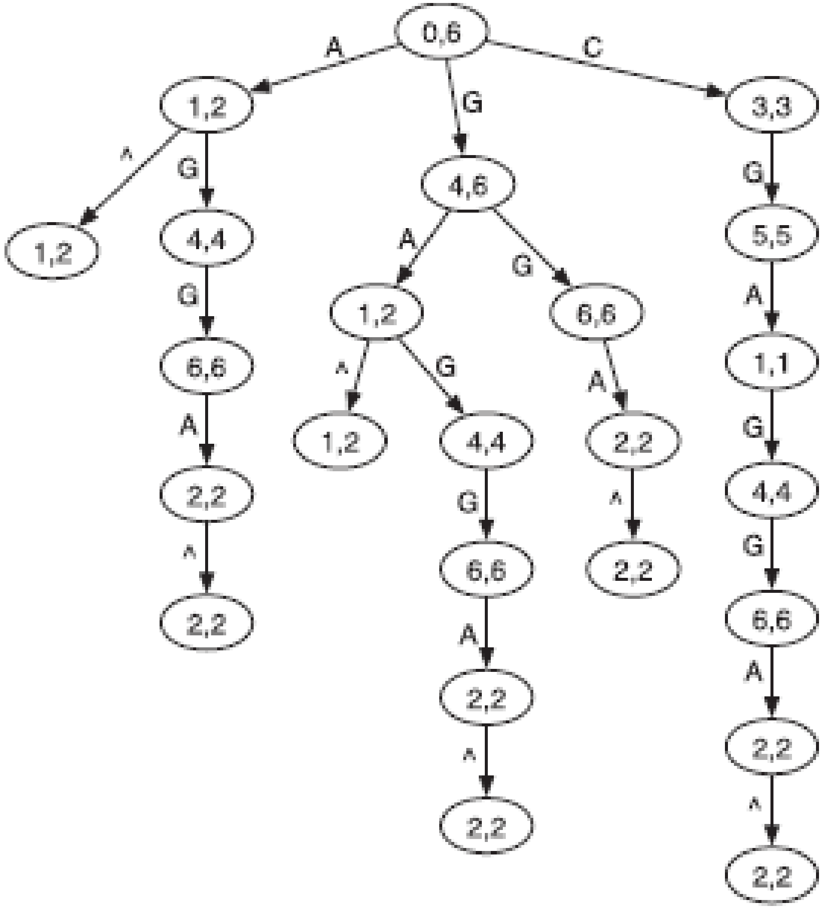
- **Alignment algorithms for short-reads**
  - Background – Blast (why can't we use it?)
  - Adapting hashed seed-extend algorthims to work with shorter reads
  - **Suffix/Prefix Tries**
  - Other alignment considerations
  - Typical alignment pipeline
- **Assembly algorithms for short reads**
  - Effect of repeats
  - Overlap-Consensus
  - de Bruijn graphs
  - Assembly evaluation metrics
  - Typical assembly pipeline

Suffix-Prefix Trie

- A family of methods which uses a Trie structure to search a reference sequence
  - Bowtie
  - BWA
  - SOAP version 2
- Trie – data structure which stores the suffixes (i.e. ends of a sequence)
- Key advantage over hashed algorithms:
  - Alignment of multiple copies of an identical sequence in the reference only needs to be done once
  - Use of an FM-Index to store Trie can drastically reduce memory requirements (e.g. Human genome can be stored in 2Gb of RAM)
  - Burrows Wheeler Transform to perform fast lookups

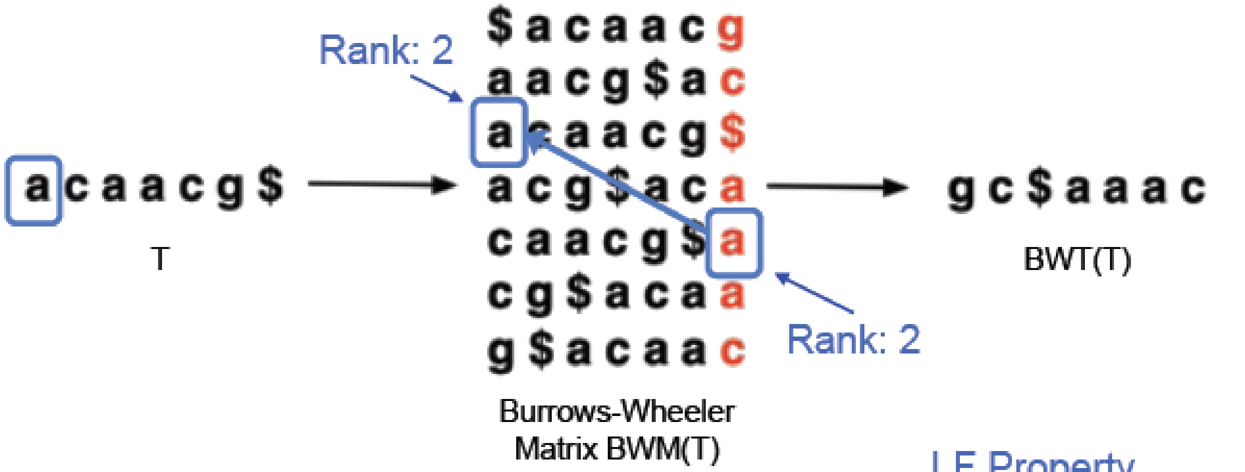
Suffix Trie

AGGAGC



Heng Li & Nils Homer. Sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics. Vol 11. No 5. 473 483, 2010

Suffix Trie



acaacg$

T

$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac

Burrows-Wheeler
Matrix BWM(T)

gc$aaac

BWT(T)

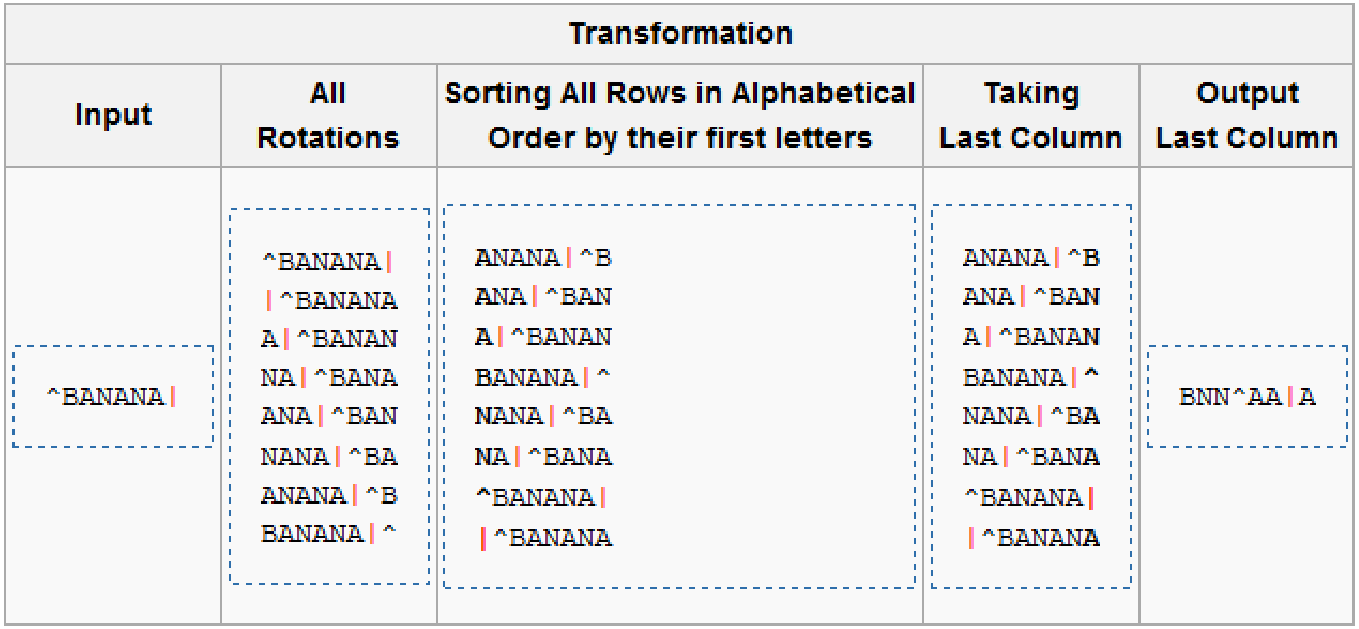
LF Property implicitly encodes Suffix Array

- BWT(T) is the index for T

A block sorting lossless data compression algorithm.
Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation.* Technical Report 124

Burrows-Wheeler Algorithm

- Encodes data so that it is easier to compress
- Burrows-Wheeler transform of the word BANANA
- Can later be reversed to recover the original word



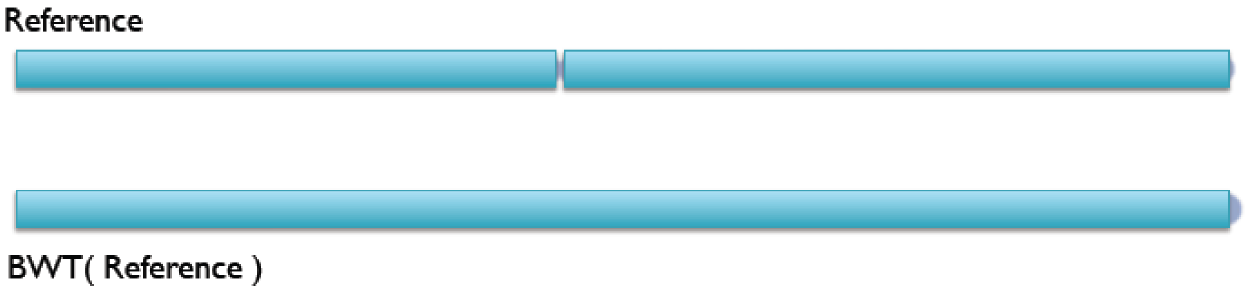
<table>
<tr><th colspan="5">Transformation</th></tr>
<tr><th>Input</th><th>All Rotations</th><th>Sorting All Rows in Alphabetical Order by their first letters</th><th>Taking Last Column</th><th>Output Last Column</th></tr>
<tr><td>^BANANA|</td><td>^BANANA|
|^BANANA
A|^BANAN
NA|^BANA
ANA|^BAN
NANA|^BA
ANANA|^B
BANANA|^</td><td>ANANA|^B
ANA|^BAN
A|^BANAN
BANANA|^
NANA|^BA
NA|^BANA
^BANANA|
|^BANANA</td><td>ANANA|^B
ANA|^BAN
A|^BANAN
BANANA|^
NANA|^BA
NA|^BANA
^BANANA|
|^BANANA</td><td>BNN^AA|A</td></tr>
</table>

More Burrows-Wheeler

Input	SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES
Burrows-Wheeler Output	TEXYDST.E.IXIXIXXSSMPPS.B..E.S.EUSFXDIIOIIIT

Repeated characters mean that it is easier to compress

Bowtie/Soap2 example

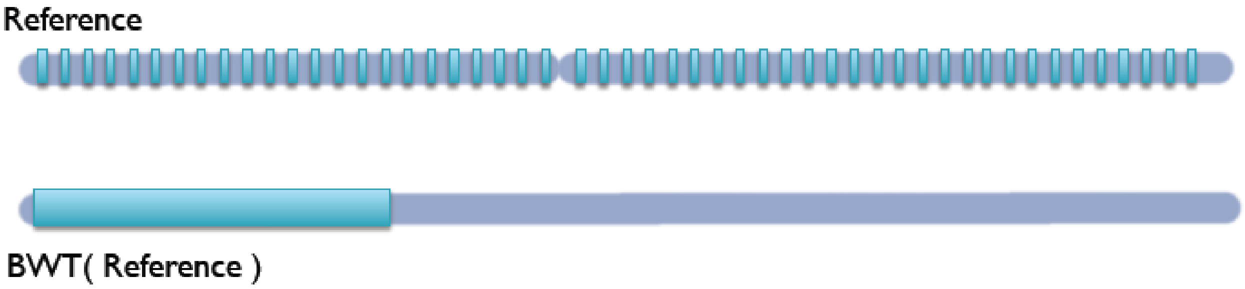


Reference

BWT(Reference)

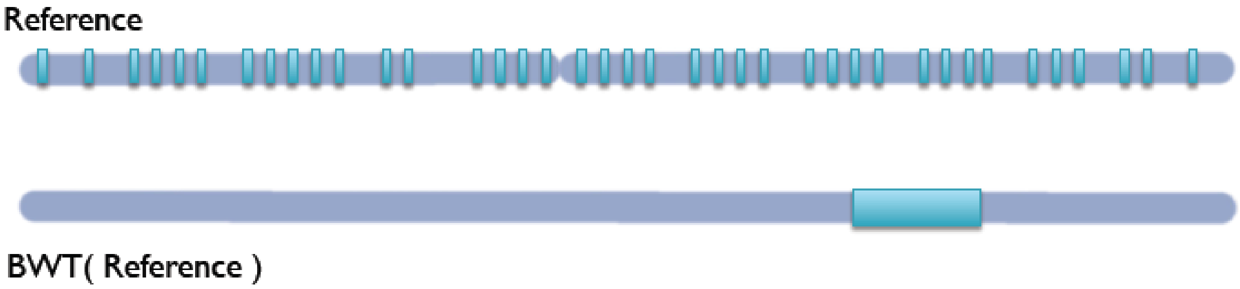
Query:
AATGATACGGCGACCACCGAGATCTA

Bowtie/Soap2 example



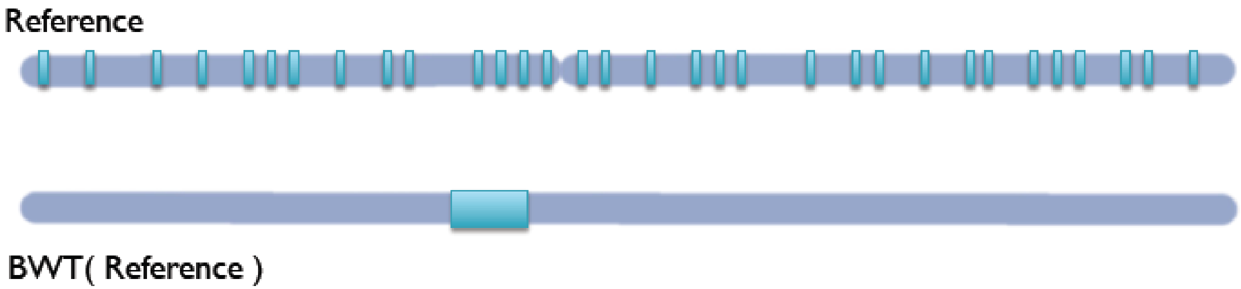
Query:
AATGATACGGCGACCACCGAGATCTA

Bowtie/Soap2 example



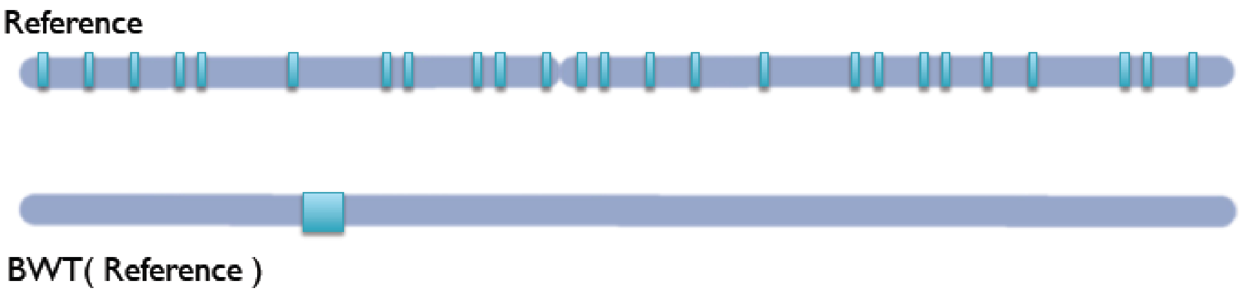
Query:
AATGATACGGCGACCACCGAGATCTA

Bowtie/Soap2 example



Query:
AATGATACGGCGACCACCGAGATCTA

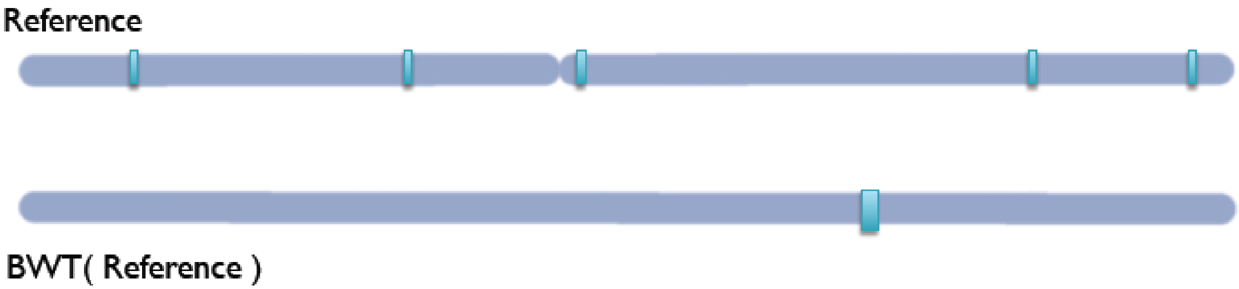
Bowtie/Soap2 example



Query:

AATGATACGGCGACCACCGAGATCTA

Bowtie/Soap2 example



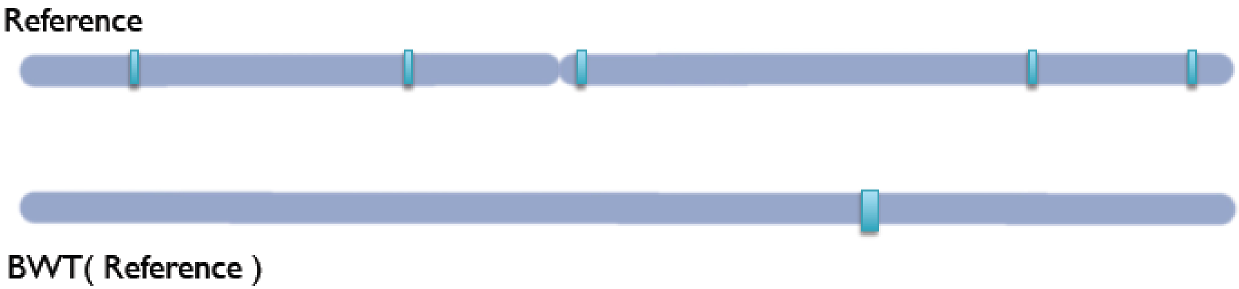
Query:
AATGATACGGCGACCACCGAGATCTA

Bowtie/Soap2 example

Reference
BWT(Reference)

Query:
AATGATACGGCGACCACCGAGATCTA

Bowtie/Soap2 example



Query:
AATGTTACGGCGACCACCGAGATCTA

Bowtie/Soap2 example

Reference
BWT(Reference)

Query:

AATGTTACGGCGACCACCGAGATCTA

Bowtie/Soap2 vs. BWA

- Bowtie and Soap2 cannot handle gapped alignments
    - No indel detection => Many false SNP calls

Bowtie/Soap2:

```
ACTCCCATTGTCATCGTACTTGGGATCGTAACA Reference
     CCATTGTCATCGTACTTGGGATCTA
           TCATCGTACTTGGGATCTA

      TTGGGATCTA
```

False SNPs

N.B. Bowtie2 can handle gapped alignments

Bowtie/Soap2 vs. BWA

- Bowtie and Soap2 cannot handle gapped alignments
  - No indel detection => Many false SNP calls

BWA:

```
ACTCCCATTGTCATCGTACTTGGGATCGTAACA Reference
     CCATTGTCATCGTACTTGGGATC-TA
           TCATCGTACTTGGGATC-TA

      TTGGGATC-TA
```

N.B. Bowtie2 can handle gapped alignments

Comparison

Hash referenced spaced seeds

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

Suffix/Prefix Trie

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

Comparison

- Bowtie's reported 30-fold speed increase over hash-based MAQ with small loss in sensitivity
- Limitations to Trie-based approaches:
    - Only able to find alignments within a certain 'edit distance’
    - Bowtie does not do gapped alignments – no indels!
    - Important to quality clip reads (-q in BWA)
    - Non-A/C/G/T bases on reads are simply treated as mismatches
    - Make sure Ns are removed!

Hash based approaches are more suitable for divergent alignments

- Rule of thumb:
    - <2% divergence -> Trie-based
        - E.g. human alignments
    - >2% divergence -> seed-extend based approach
        - E.g. wild mouse strains alignments

Thank you for your attention!