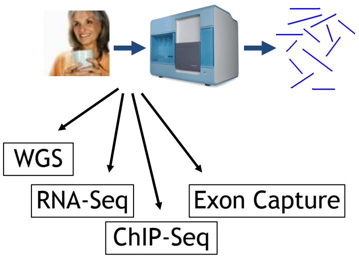
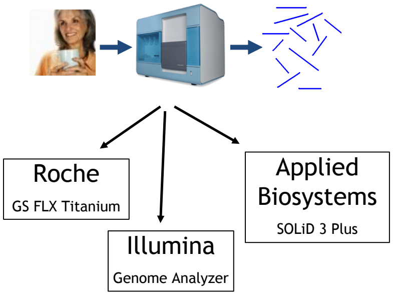
Short Read Alignment

Tobias Rausch
7th June 2010

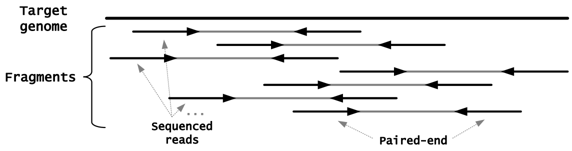
Sequencing



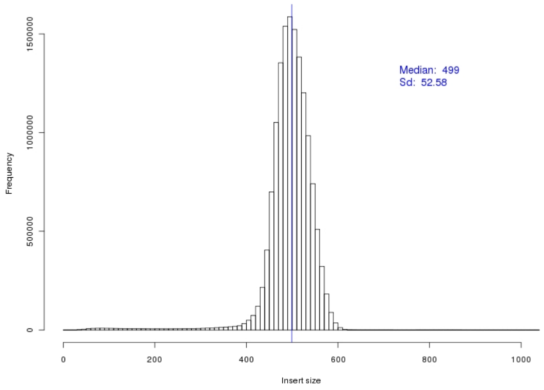
Sequencing



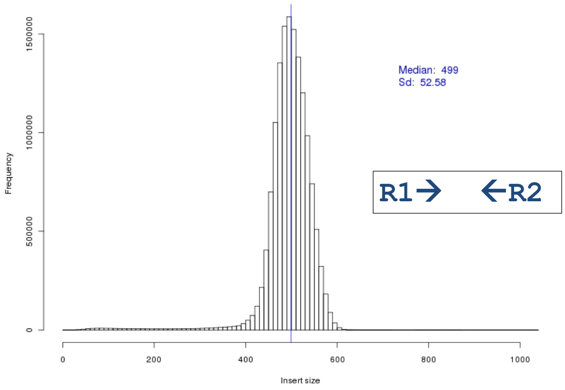
Paired-End Sequencing



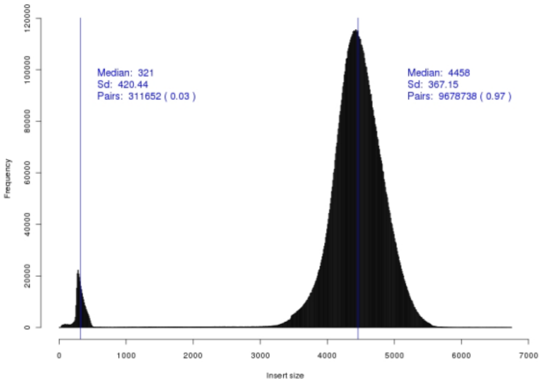
Paired-End Libraries



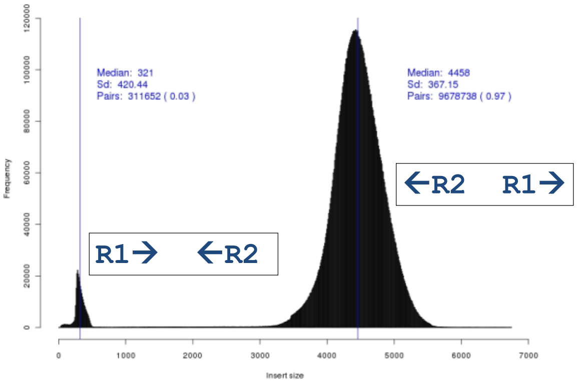
Paired-End Libraries



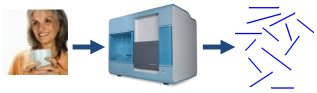
Mate-Pair Libraries



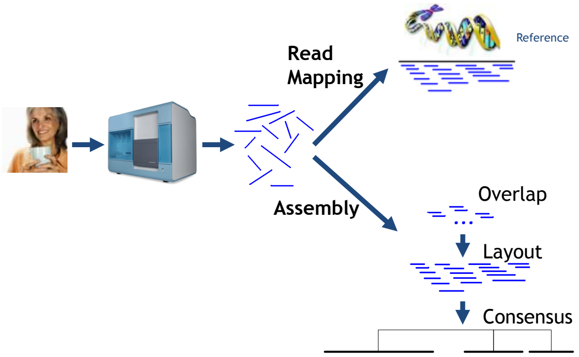
Mate-Pair Libraries



Data Analysis



Data Analysis

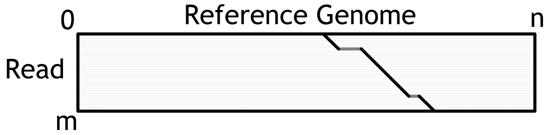


Assembly

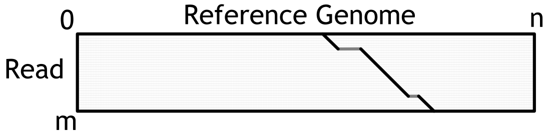
- String Graph Assembler
  - Overlap - Layout - Consensus assemblers
  - Examples
    - *Celera Assembler, Arachne, Atlas*
- De-Bruijn Graph Assembler
  - Short-read assemblers
  - Examples:
    - *Velvet, Abyss, SOAPdenovo*
  - Transcriptome assembly: *Oases*

Read Mapping

Read Mapping

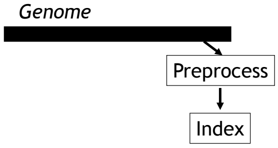


Read Mapping

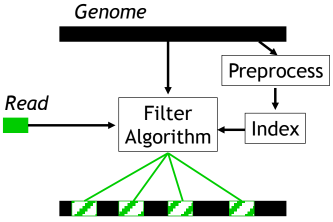


- Quadratic algorithm
  - Requires O(m*n) time and space
- Infeasible for millions of short reads

Filtering



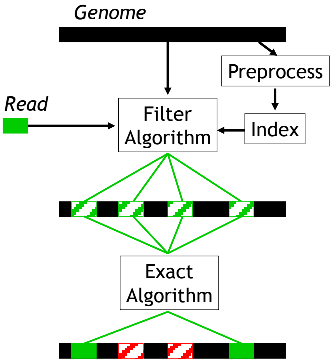
Filtering



Filtration Phase

Potential Matches ▨

Filtering



Filtration Phase

Potential Matches ▨

Verification Phase

True Matches 🟩

False Matches 🟥

Simple k-mer Index, k=3

S = ACGAAAACTCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	
AAC		ACG		…	
AAG		ACT		GAA	
AAT		AGA		…	
ACA		…		TTT	

- Size of that table: 4^3 = 64 entries = $|\Sigma|^k$

Simple k-mer Index, k=3

S = ACGAAAACTCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	
AAC		ACG	0	…	
AAG		ACT		GAA	
AAT		AGA		…	
ACA		…		TTT	

- Size of that table: 4^3 = 64 entries = $|\Sigma|^k$

Simple k-mer Index, k=3

S = ACGAAAACTCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	1
AAC		ACG	0	…	
AAG		ACT		GAA	
AAT		AGA		…	
ACA		…		TTT	

- Size of that table: 4^3 = 64 entries = $|\Sigma|^k$

Simple k-mer Index, k=3

S = ACGAAAACTCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	1
AAC		ACG	0	…	
AAG		ACT		GAA	2
AAT		AGA		…	
ACA		…		TTT	

- Size of that table: 4^3 = 64 entries = $|\Sigma|^k$

Simple k-mer Index, k=3

S = ACGAAAACTCGATTACTCGACC

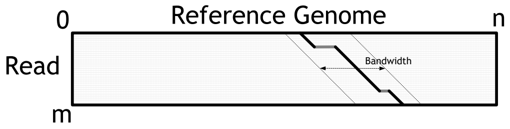
	Hitlist		Hitlist		Hitlist
AAA	3,4	ACC	19	CGA	1
AAC	5	ACG	0	…	…
AAG	Empty	ACT	6,14	GAA	2
AAT	Empty	AGA	…	…	…
ACA	Empty	…	…	TTT	Empty

Searching a Read

	Hitlist		Hitlist		Hitlist
AAA	3,4	ACC	19	CGA	1
AAC	5	ACG	0	…	…
AAG	Empty	ACT	6,14	GAA	2
AAT	Empty	AGA	…	…	…
ACA	Empty	…	…	TTT	Empty

- Read Sequence: ACTG
  - Potential match at position 6 and 14

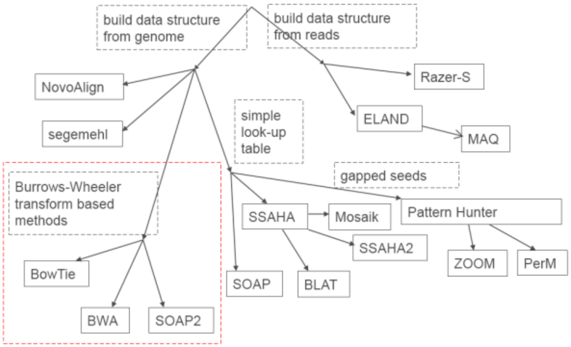
Verification Algorithm Banded Dynamic Programming



Techniques

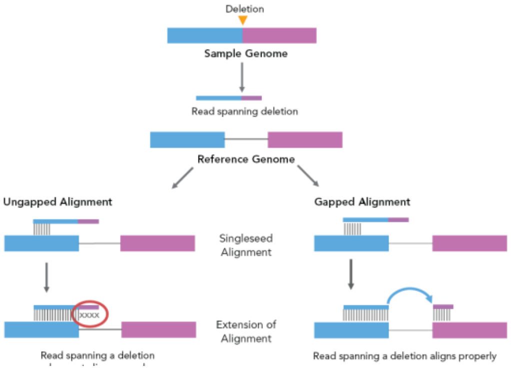
- Index
  - Hash tables, k-mer Index
  - Suffix arrays
  - Burrows-Wheeler-Transformation (BWT) of a suffix array
- Filtering Algorithms
  - Single or multiple seeds
  - Pigeonhole principle
  - Q-gram filtering
- Verification
  - Simple seed-and-extend
  - Banded dynamic programming
  - Quality-based dynamic programming

Read Mappers



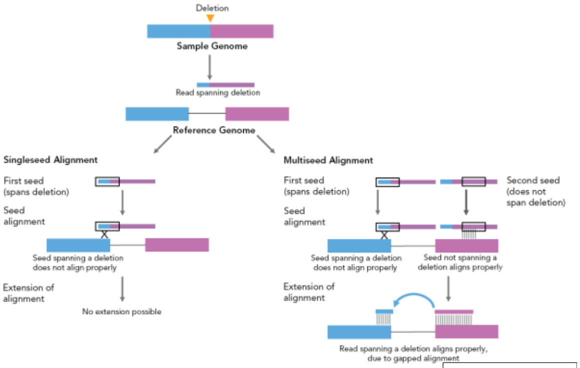
Source: illumina

ELANDv2 – Gapped Banded Alignment (20bp)



Source: illumina

ELANDv2 – Multiseed Alignment (Seed max. 2 errors)


Deletion
Sample Genome
Read spanning deletion
Reference Genome
Singleseed Alignment
First seed (spans deletion)
Seed alignment
Seed spanning a deletion does not align properly
Extension of alignment
No extension possible
Multiseed Alignment
First seed (spans deletion)
Second seed (does not span deletion)
Seed alignment
Seed spanning a deletion does not align properly
Seed not spanning a deletion aligns properly
Extension of alignment
Read spanning a deletion aligns properly, due to gapped alignment

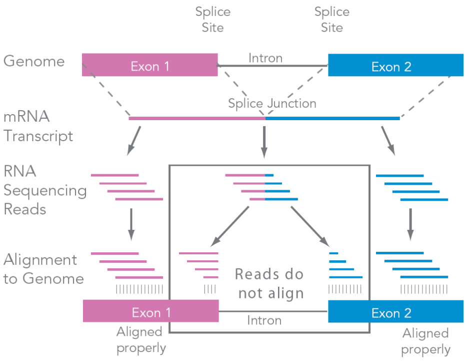
Source: illumina

Parallelization

- Data Decomposition
  - Split the reads
  - Examples: Bowtie, Eland
- Functional Decomposition
  - Separate filtering and verification processes

RNA-Seq

RNA-Seq

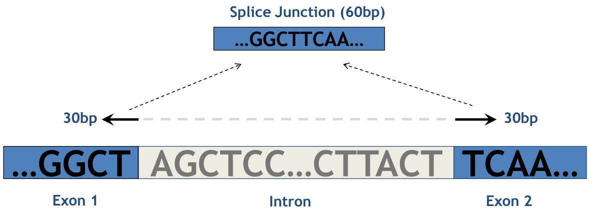


RNA-Seq

- Read-Mapping Protocol
  - Alignment against contaminants (rRNA)
  - Alignment against splice-junctions
  - Alignment against genome

RNA-Seq

- Read-Mapping Protocol
  - Alignment against contaminants (rRNA)
  - Alignment against splice-junctions
  - Alignment against genome


Splice Junction (60bp)

...GGCTTCAA...

30bp

30bp

...GGCT AGCTCC...CTTACT TCAA...

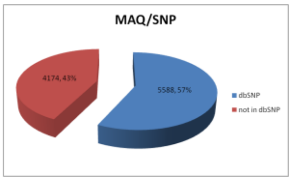
Exon 1

Intron

Exon 2

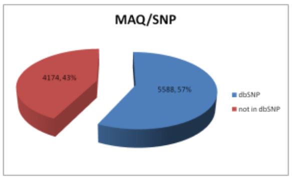
Calling SNPs

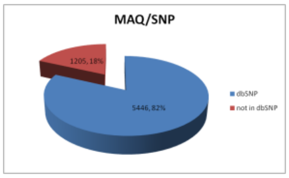
- Direct Alignment against hg18



Calling SNPs

- Direct Alignment against hg18





- Alignment against rRNA (1%) + Alignment against splice junctions (11%)

SAM/BAM

- Generic format for storing large nucleotide sequence alignments
- SAM Tools
  - Sorting alignments
  - Merging alignments
  - Indexing alignments
  - Viewing alignments

SAM record

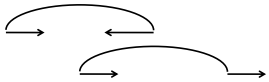
□ Tab-delimited format

- □ Field 1: Query name
- □ Field 2: Flag
- □ Field 3: Reference sequence name
- □ Field 4: 1-based leftmost coordinate of the clipped sequence
- □ Field 5: Mapping quality
- □ Field 6: CIGAR strings
- □ Field 7: Mate reference sequence name
- □ Field 8: 1-based leftmost coordinate of the clipped sequence
- □ Field 9: Insert size (5’ to 5’)
- □ Field 10: Query sequence
- □ Field 11: Sequence qualities

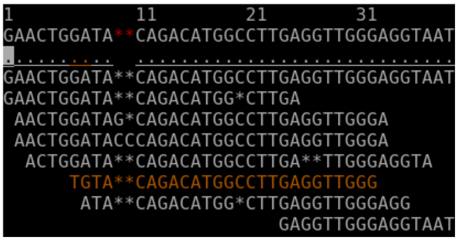
SAM record

□ Tab-delimited format

- □ Field 1: Query name
- □ Field 2: Flag
- □ Field 3: Reference sequence name
- □ Field 4: 1-based leftmost coordinate of the clipped sequence
- □ Field 5: Mapping quality
- □ Field 6: CIGAR strings
- □ Field 7: Mate reference sequence name
- □ Field 8: 1-based leftmost coordinate of the clipped sequence
- □ Field 9: Insert size (5’ to 5’)
- □ Field 10: Query sequence
- □ Field 11: Sequence qualities

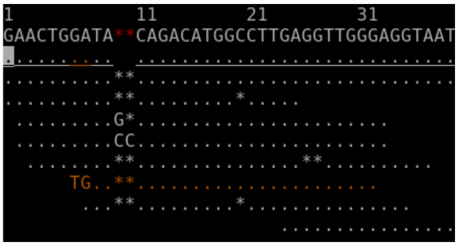


Sam / Bam Format



```
1           11        21        31
GAACTGGATA**CAGACATGGCCTTGAGGTTGGGAGGTAAT
..........  .............................
GAACTGGATA**CAGACATGGCCTTGAGGTTGGGAGGTAAT
GAACTGGATA**CAGACATGG*CTTGA
 AACTGGATAG*CAGACATGGCCTTGAGGTTGGGA
 AACTGGATACCCAGACATGGCCTTGAGGTTGGGA
  ACTGGATA**CAGACATGGCCTTGA**TTGGGAGGTA
      TGTA**CAGACATGGCCTTGAGGTTGGG
       ATA**CAGACATGG*CTTGAGGTTGGGAGG
                         GAGGTTGGGAGGTAAT
```

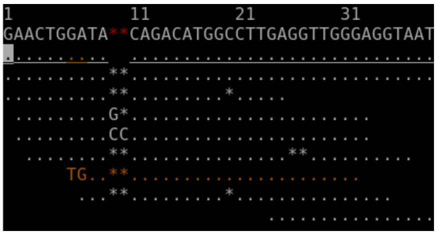
Sam / Bam Format



```
1           11        21        31
GAACTGGATA**CAGACATGGCCTTGAGGTTGGGAGGTAAT
..........  .............................
..........**.............................
..........**.........*.....
 .........G*.......................
 .........CC.......................
  ........**...............**..........
      TG..**......................
       ...**.........*...............
                         ................
```

- Sequence characters agreeing with the reference are set to “ . “ or “ , “ for reads aligned to the forward or reverse strand.

Sam / Bam Format



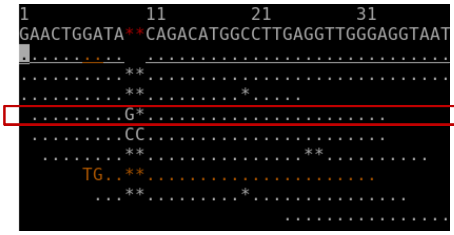
```
1           11        21        31
GAACTGGATA**CAGACATGGCCTTGAGGTTGGGAGGTAAT
..........  .............................
..........**.............................
..........**.........*.....
 .........G*.......................
 .........CC.......................
  ........**...............**..........
      TG..**......................
       ...**.........*...............
                         ................
```

CIGAR Strings

- 39M
- 19M1D5M
- 9M1I23M
- 9M2I23M
- 23M2D10M
- 26M
- 12M1D15M
- 16M

- M: Alignment match or mismatch
- I: Insertion to the reference
- D: Deletion from the reference

Sam / Bam Format



```
1           11        21        31
GAACTGGATA**CAGACATGGCCTTGAGGTTGGGAGGTAAT
..........  .............................
..........**.............................
..........**.........*.....
 .........G*.......................
 .........CC.......................
  ........**...............**..........
      TG..**......................
       ...**.........*...............
                         ................
```

CIGAR Strings

- 39M
- 19M1D5M
- 9M1I1P23M
- 9M2I23M
- 23M2D10M
- 26M
- 12M1D15M
- 16M

- P: Padding (silent deletion)
- This is not even implemented by BWA
  - Because it would require a *de novo local assembler*!

Sam / Bam Format

- N: Skipped region from the reference
  - For spliced reads:
    - ACATGATA……………………………………….GAGCTTTA (Cigar: 8M56N8M)
- Two more CIGAR characters
  - S: Soft clip on the read
  - H: Hard clip on the read

Flags

Bitwise FLAG: $f_{15}f_{14}f_{13}f_{12}f_{11}f_{10}f_9f_8f_7f_6f_5f_4f_3f_2f_1f_0$ with f_i= {0,1}

f_0: 0 = Read is not paired in sequencing, 1 = Read is paired in seq.
f_1: 1 = The read is mapped in a proper pair
f_2: 1 = The query sequence itself is unmapped
f_3: 1 = The mate is unmapped
f_4: 0 = forward strand, 1 = reverse strand
…