Brian Knaus, US Forest Service

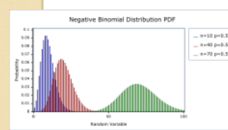
Botany 2012 Intro to Next Generation Sequencing Workshop

Computation - Essentials

- Data formats (fastq, SAM/BAM).
- De novo assembly (DNA/RNA).
- Read mapping (bowtie, bwa).
- Reference guided assembly.

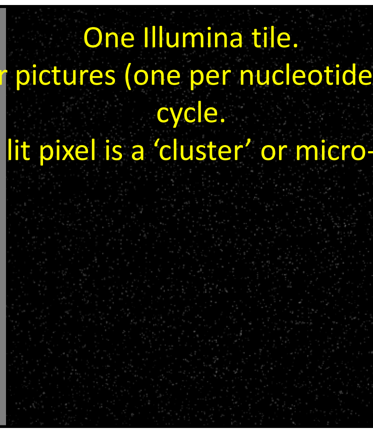
Library preparation

Sequencing



Hypothesis testing

Bioinformatics



One Illumina tile.
Four pictures (one per nucleotide) per cycle.
Each lit pixel is a 'cluster' or micro-read.

Fastq format

Casava v1.8

```
@ILLUMINA-20A1B2:45:FC62N85AAXX:1:115:7298:18600 1:N:0:CGATGT
TGGAGTTTTTGCCTGTATATATCTTCATGCACTTCACTGGCTTTCTGTGGATCAAAACGCAGCTCATTGCAAA
+
GGBDGGGGGEGGGEGEEGGGHHBHHHDGHHDGGGGHHEHFDDHHHHHHDHBHHHGHHH@HGHFHHFDGBGHHH
@ILLUMINA-20A1B2:45:FC62N85AAXX:1:115:7362:18603 1:N:0:CGATGT
GTTTATTCCCCTCTGCTGCCCGCTTGCAGCTCATAGAATGGCCGTCTACAATTTTATTTGGCTCCTCCAGAGC
+
EIIIIIIIIIIIIIIIIIIHIIIIIIIIIHHFIIIIHIIIIIII3HEEEEEHHIHIGHIHCEGIGHGGG;BBBD
@ILLUMINA-20A1B2:45:FC62N85AAXX:1:115:7675:18603 1:Y:0:CGATGT
CGTTTAGGTTCAACGTGGCTTCCACCTCAGATTGCCCACCAGACAAAAACATGATACCGGGAACAGACGGTGG
+
CBBBBBBB7BA@A@?9=/@8?A?##################################################
```


SAM format

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```


SAM file format

```
@HD VN:1.0 SO:sorted
@PG TopHat VN:1.0.13 CL:/local/cluster/bin/tophat -p 4 --solexa1.3-quals
../indexes/psme_ref ../psme_seqs.fq
ILLUMINA-3AB384_0001:6:24:19059:8781#GATT 0 0_54_255 1 255 80M *
0 0
TCTTCTTCATGTTTGGCACGTGTATTCGGGCCTACTTCGCCTTTCCTTCACAGTAGGCGCCTTATCATTATTGGTCAGTT
CCCCCCCCCCCCCCCCDCCCCCCCC@CBCBBCCBCCCCCCCCCCCCCCCCCDCD@C@CCCC4=CCBCCCCAC>B>BBC
NM:i:1
HWI-EAS121_0024_FC61F8DAAXX:7:101:7452:15154#CTGT 0 0_54_255 17 255
76M * 0 0
CACGTGTATTCGGGCCTACTTCGCCTTTCCTTCACAGTAGGCGCCTTGTCATTATTGGTCAGTTATGACCTTAATT
GGGGGGGGGGFEGFFGFEEFFBEECEFFFFFGGDGFDDGE:FBBFEGFFD?DEDEFB=DDD=ECCC=EAACDEDC=
NM:i:0
```

```
@header line1 - file format version
@header line2 - program which created the file

1 read name
2 flag
3 reference name

ILLUMINA-3AB384_0001:6:24:19059:8781#GATT 0 0_54_255 1 255 80M *
0 0
TCTTCTTCATGTTTGGCACGTGTATTCGGGCCTACTTCGCCTTTCCTTCACAGTAGGCGCCTTATCATTATTGGTCAGTT
CCCCCCCCCCCCCCCCDCCCCCCCC@CBCBBCCBCCCCCCCCCCCCCCCCCDCD@C@CCCC4=CCBCCCCAC>B>BBC
NM:i:1
HWI-EAS121_0024_FC61F8DAAXX:7:101:7452:15154#CTGT 0 0_54_255 17 255
76M * 0 0
CACGTGTATTCGGGCCTACTTCGCCTTTCCTTCACAGTAGGCGCCTTGTCATTATTGGTCAGTTATGACCTTAATT
GGGGGGGGGGFEGFFGFEEFFBEECEFFFFFGGDGFDDGE:FBBFEGFFD?DEDEFB=DDD=ECCC=EAACDEDC=
NM:i:0
```

De novo assembly

DNA assemblers
- Velvet
- ABySS
- SOAPdenovo
- ALLPATHS-LG

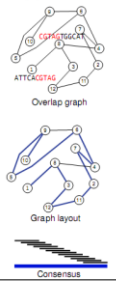
RNA assemblers
- Trinity
- Velvet/oases
- TransABySS
- SOAPdenovo-Trans
- Rnnotator

7

The Overlap-layout-consensus (OLC) approach

1. Pairwise alignments and overlap graph
2. Graph Layout: search of a single path in the graph (i.e. the Hamiltonian path)
3. Multiple sequence alignments and consensus

Examples: Newbler, Celera, Arachne , Edena, YASRA



Tristan Lefébure, Cornell University

ACTGATTG

9

ACTGATTG

k-mer = 5

10

ACTGA

ACTGATTG

k-mer = 5

11

ACTGA

CTGAT

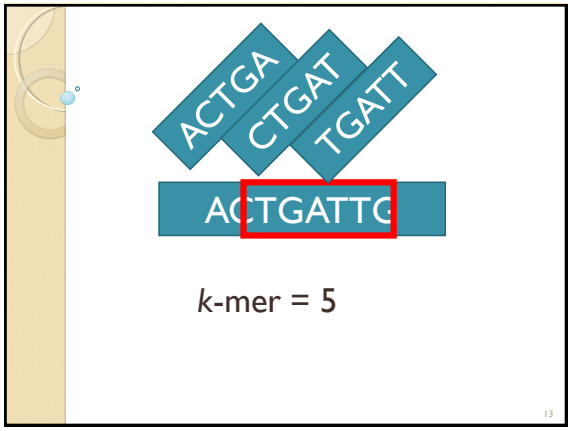
ACTGATTG

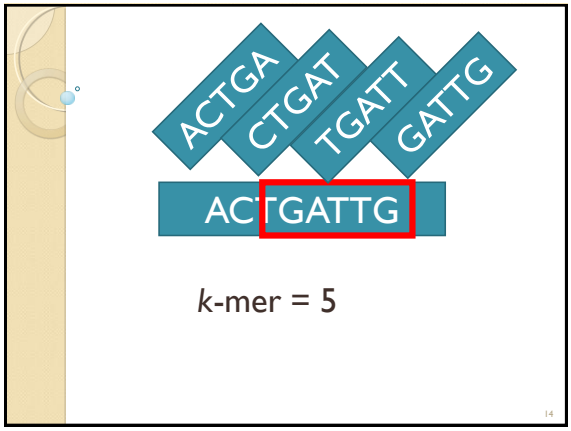
k-mer = 5

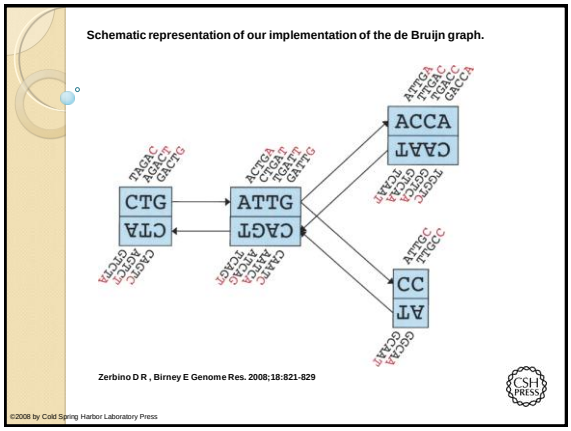
12

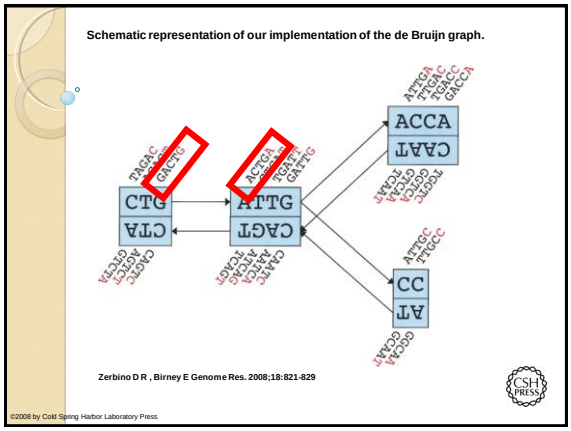
Brian Knaus, US Forest Service

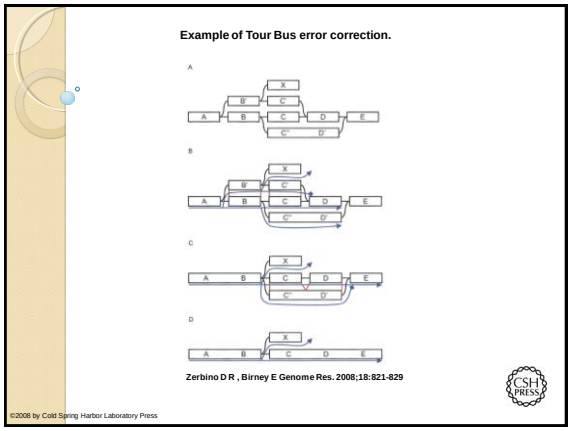
Botany 2012 Intro to Next Generation Sequencing Workshop

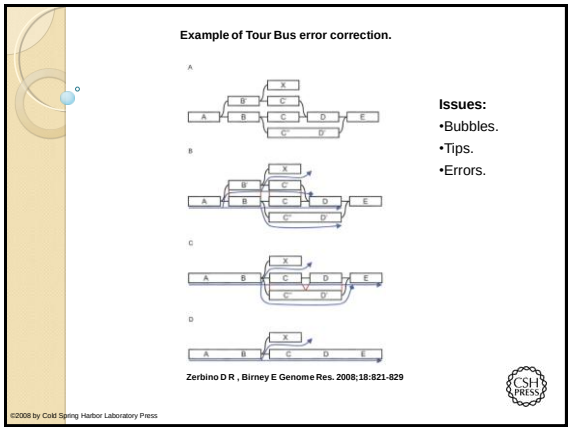


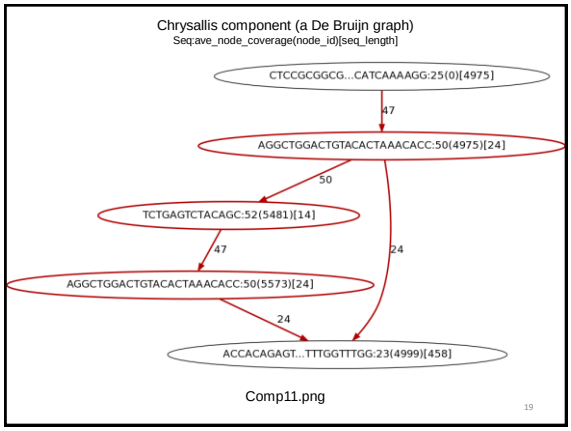






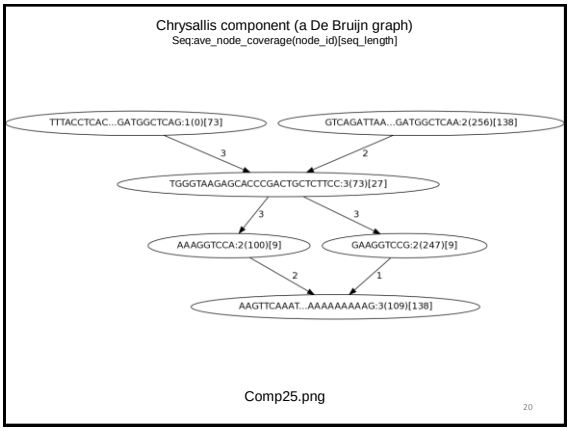






Comp11.png

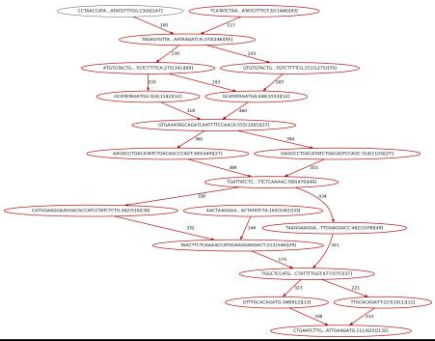
19



Comp25.png

20

Visualizing de Bruijn graphs



21

Read Mapping

- Bowtie/bowtie2
- BWA
- MAQ
- Stampy
- RUM (RNA)

22

MAQ: Mapping and Assembly with Qualities

- Hash based assembler.
- Compiled code (C).
- Seeded alignment.
  - Uses first 28 bp (highest quality region).
  - Allows for up to 2 mismatches in this region (~7% sequence divergence).
  - After finding these regions it extends through the rest of the read length.
- Currently limited to 63 bp reads.
- Ungapped alignment.

23

MAQ vs. Bowtie

- Maq = 0.49 million reads mapped per hour.
- Bowtie = 29.5 million reads mapped per hour.

24

Bowtie

- Reference:
  - Burrows-Wheeler transform.
  - FM index (Feragina and Manzini).
- Reads from 4 to 1,024 bp long.
- Greedily finds matches.

25

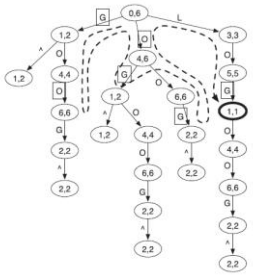
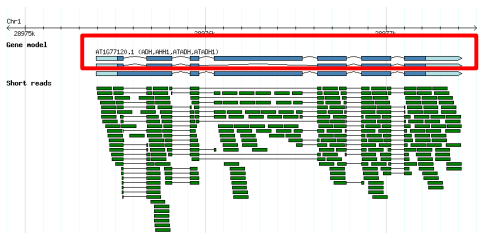


Fig. 1. Prefix trie of string 'GOOGOL'. Symbol ∧ marks the start of the string. The two numbers in a node give the SA interval of the string represented by the node (see Section 2.3). The dashed line shows the route of the brute-force search for a query string 'LOL', allowing at most one mismatch. Edge labels in squares mark the mismatches to the query in searching. The only hit is the bold node [1, 1] which represents string 'GOL'.

Li and Durbin. 2009. Bioinformatics 25: 1754-1760.

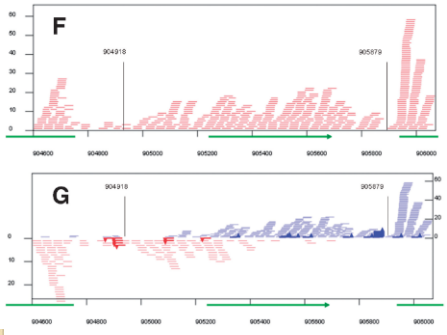
26

Genomic references.
Splice junctions?



27

Strand specificity



Parkhomchuck et al. 2009. Transcriptome analysis by strand-specific sequencing of complimentary DNA. Nucleic Acids Research 37(18):e123

28

Graphical options

- Galaxy
- Easy terminal alternative (ETA)
- CLC Genomics Workbench (commercial)
- Geneious (commercial)