sol genomics network

Bioinformatic Command-line tools

sol genomics network
Boyce Thompson Institute for Plant Research
Ithaca, New York 14853
U.S.A.

presented by

Naama Menda

Command-line

- Linux based
- More options running programs
- Run programs locally
- Run jobs on bigger server

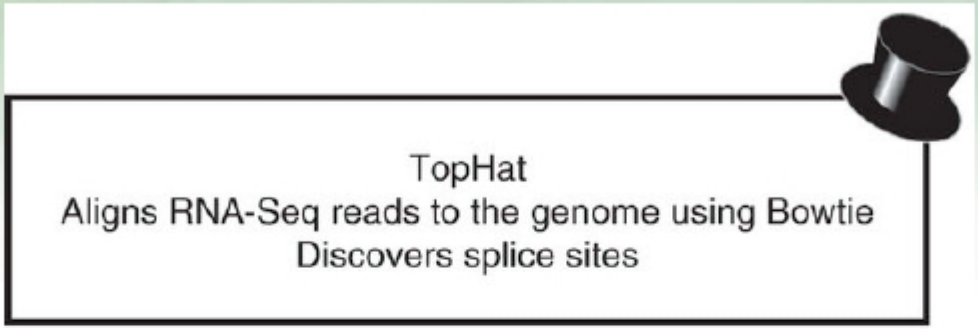
sol genomics network

Command-line tools

- Local BLAST
- RNAseq analysis
  - Pre-processing (FastQC, fastq-mcf)
  - Assembly (Bowtie, Tophat, Trinity)
  - View output (samtools tview)
  - Expression (Cufflinks)

Tophat

Reference-guided Assembly



- http://tophat.cbcb.umd.edu/
- deals with splice junctions
- Illumina RNA-Seq reads (paired or single end)

Exercise 1

Two RNA-seq datasets used in the tomato genome project were downloaded from the SRA in .sra format and extracted using the SRA toolkit (NCBI).

http://www.ncbi.nlm.nih.gov/sra

They were cleaned using fastq-mcf in last week's class. All data is from *S. pimpinellifolium*.

Datasets :

- breaker fruit (two files)
- immature fruit (two files)

In this exercise, we will map the reads to tomato chromosome 4 using reference-guided assembly

- Use bowtie2-build to index the reference

- Index tomato reference file

  - *bowtie2-build SL2.40ch04.fa SL2.40ch04*

* Download files from
  ftp.solgenomics.net/bioinfo_class/interns/data/

- Use tophat2 to map each .fq read set to the reference

```
tophat2 --no-coverage-search
--no-novel-juncs -o SRR404334_tophat_out/
~/bwt2_index/SL2.40ch04 SRR404334_ch4.fq
```

Repeat tophat2 command for remaining 3 data sets
(1 more breaker, 2 immature fruit).

Tophat output

sol genomics network

Tophat output:

- ✦ bam file of mapped reads (bam = compressed .sam file)
- ✦ bed files: junctions, insertions, deletions

Convert accepted_hits.bam to accepted_hits.sam
samtools view -o accepted_hits.sam accepted_hits.bam

SAM format

sol genomics network

Sam file sample

```
@SQ     SN:SL2.40ch00  LN:21805821
@SQ     SN:SL2.40ch01  LN:90304244
@SQ     SN:SL2.40ch02  LN:49918294
@SQ     SN:SL2.40ch03  LN:64840714
@SQ     SN:SL2.40ch04  LN:64064312
@SQ     SN:SL2.40ch05  LN:65021438
@SQ     SN:SL2.40ch06  LN:46041636
@SQ     SN:SL2.40ch07  LN:65268621
@SQ     SN:SL2.40ch08  LN:63032657
@SQ     SN:SL2.40ch09  LN:67662091
@SQ     SN:SL2.40ch10  LN:64834305
@SQ     SN:SL2.40ch11  LN:53386025
@SQ     SN:SL2.40ch12  LN:65486253
HWI-EAS339_0011:2:92:9381:20526#0     99     SL2.30ch00     78     20     7S77M2S =
    182    187    AAAATGAAAATAAAAATACAAATAGTAATAATAATAATAATAATAATAATAATCATAATAATAATAATAATTACAATAATATT
DDD?DDDDDDD6:32-18:6<BCC>DDDDD>>A?A=CACCAC=CCC6@>@;>C@C;>>6@?CCACDD=D=>7BC8B?BAABACCCA
PG:Z:novoalign  ZO:Z:+-  AS:i:249       UQ:i:249       NM:i:7  MD:Z:7T1A1T5A31A17A1T7  PQ:i:459       SM:i:150       AM:i:21
HWI-EAS339_0011:3:88:16221:9366#0     99     SL2.30ch00     90     5      10S76M  =
    127    121    TGACTAAGCAAAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAAT
GGGGGGGGGFGGGGGGFGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGFGEGGBGGFGGFFGGGGGFGGGGDGGDGGEGFD
FF  PG:Z:novoalign  ZO:Z:+-  AS:i:88  UQ:i:88  NM:i:0  MD:Z:76  PQ:i:88  SM:i:0
  AM:i:0
HWI-EAS339_0011:2:3:16931:15264#0     163    SL2.30ch00     92     11     3S73M1D9M1H     =      139    132
ATTATAATAATAATAATAATAAGAATAATAATAAGAATAATAATAATAATAATAATAATAATAATAATAATAAAATAATAAT
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGFGGGGGGGGGGGEGGGGGFGGFGGGGGFGGGGGEGGGFGGGDGGD
GF   PG:Z:novoalign  ZO:Z:+-  AS:i:143       UQ:i:143       NM:i:3  MD:Z:19T11T41^T9       PQ:i:203       SM:i:0  AM:i:0
```

* reference-guided assembly output

SAM format

sol genomics network

Sam format column definitions

Index	Field Name	Description
1	QNAME	Query pair NAME if paired; or Query NAME if unpaired
2	FLAG	Bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of the clipped sequence
5	MAPQ	MAPping Quality
6	CIGAR	Extended CIGAR string
7	MRNM	Mate Reference sequence NaMe; "=" if the same as RNAME
8	MPOS	1-based leftmost Mate POSition of the clipped sequence
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence
11	QUAL	Query QUALity

Table 1.1: Brief summary of the SAM format

SAM format

sol genomics network

Sam flag field definitions

Field	Hex Code	Description
f_0	0x0001	the read is paired in sequencing, no matter whether it is mapped in a pair
f_1	0x0002	the read is mapped in a proper pair (depends on the protocol, normally inferred during alignment)
f_2	0x0004	the query sequence itself is unmapped
f_3	0x0008	the mate is unmapped
f_4	0x0010	strand of the query (0 for forward; 1 for reverse strand)
f_5	0x0020	strand of the mate
f_6	0x0040	the read is the first read in a pair
f_7	0x0080	the read is the second read in a pair
f_8	0x0100	the alignment is not primary (a read having split hits may have multiple primary alignment records)
f_9	0x0200	the read fails platform/vendor quality checks
f_10	0x0400	the read is either a PCR duplicate or an optical duplicate

Table 1.2: Brief summary of the SAM format

Excellent explanation for bitwise flags found here: http://seqanswers.com/forums/showthread.php?t=2301

Tool to translate meaning of bitwise flag: http://picard.sourceforge.net/explain-flags.html

SAM format

sol genomics network

Cigar String

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Manipulating Sam files with linux

sol genomics network

1. How many reads are in the tophat output file?

```
grep -v "^@SQ" accepted_hits.sam |wc
```

2. How many reads map to chromosome 4?

```
awk '$3=="SL2.40ch04" {print $0}' accepted_hits.sam |wc
```

3. How many reads map to each chromosome?

```
cut -f3 accepted_hits.sam |sort |uniq -c
```

Bed format

sol genomics network

- zero-based coordinates
- http://genome.ucsc.edu/FAQ/FAQformat#format1

```
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

GFF3 format

sol genomics network

Popular file format for storing annotation information

- Column 1: "seqid"
- Column 2: "source" ex: software used
- Column 3: "type"
- Columns 4 & 5: "start" and "end"
- Column 6: "score" ex: e-values
- Column 7: "strand"
- Column 8: "phase" deals with where feature begins in ref to the reading frame
- Column 9: "attributes"

1-based coordinates

```
##gff-version 3
ctg123  .  exon  1300  1500  .  +  .  ID=exon00001
ctg123  .  exon  1050  1500  .  +  .  ID=exon00002
ctg123  .  exon  3000  3902  .  +  .  ID=exon00003
ctg123  .  exon  5000  5500  .  +  .  ID=exon00004
ctg123  .  exon  7000  9000  .  +  .  ID=exon00005
```

Viewing output

sol genomics network

- Samtools tview

Assembly stats

sol genomics network

How good is the assembly?

- Check for contaminants by using BLAST to search appropriate databases (SeqClean)
- How many contigs are there vs how many genes expected - total sequence length?
- Length of contigs - compare to known long transcripts?
- How many reads map (reference-guided)?

Expression analysis

sol genomics network

Illumina

- Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks (Trapnell et al., 2012)

454

- The comparison of gene expression from multiple cDNA libraries. (Stekel and Falciani, 2000).

Expression analysis

sol genomics network

- Expression experimental design - replicates, treatments, tissue, etc

Statistical Design and Analysis of RNA Sequencing Data
Paul L. Auer and R. W. Doerge 1

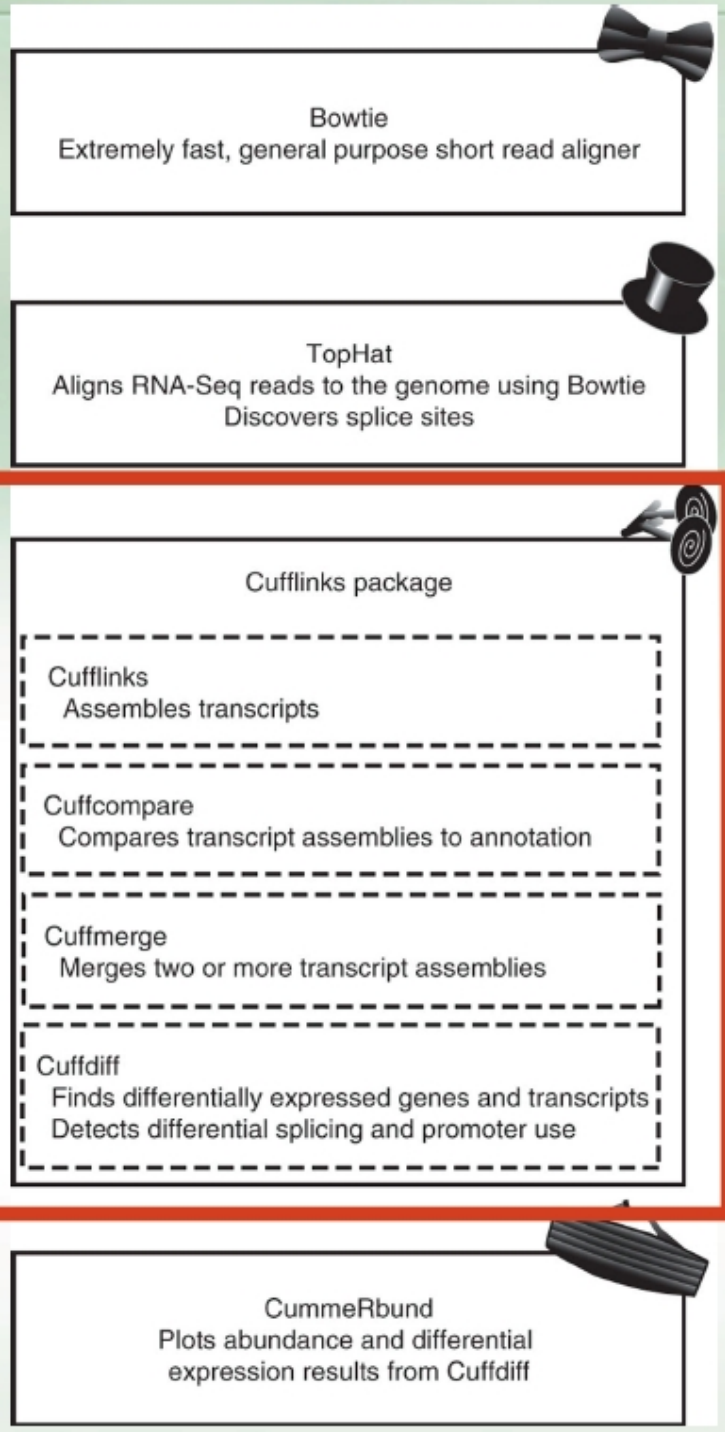
- map to genome or transcriptome, is a reference genome available?

3. Analysis

Expression

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

- Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter



RPKM

Reads Per Kilobase of exon model per Million mapped reads (Mortazavi et al., 2008)

* Cufflinks uses the analogous FPKM

Computational methods for transcriptome annotation and quantification using RNA-seq . Garber et al., Nature Methods 2011

Exercise 2

Using the known tomato gene models for chromosome 4, detect differentially expressed gene in immature and breaker fruit.

STEP 1: Use cuffdiff for detection of differential expression.

```
cuffdiff -o cuffdiff_out -b ~/Desktop/ch4_demo_dataset/bwt2_index/SL2.40ch04.fa -u
annotation/ITAG2.3_gene_models_ch4.gtf  breaker/SRR404334/SRR404334_tophat_out/
accepted_hits.bam,breaker/SRR404336/SRR404336_tophat_out/accepted_hits.bam
immature_fruit/SRR404331/SRR404331_tophat_out/accepted_hits.bam,immature_fruit/
SRR404333/SRR404333_tophat_out/accepted_hits.bam
```

~12 minutes

Expression

Cufflinks Output:

```
-rw-r--r-- 1 bioinfo bioinfo   12 Apr 16 11:17 cds.count_tracking
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 cds.diff
-rw-r--r-- 1 bioinfo bioinfo  124 Apr 16 11:17 cds_exp.diff
-rw-r--r-- 1 bioinfo bioinfo   91 Apr 16 11:17 cds.fpkm_tracking
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 cds.read_group_tracking
-rw-r--r-- 1 bioinfo bioinfo 343K Apr 16 11:17 gene_exp.diff
-rw-r--r-- 1 bioinfo bioinfo 183K Apr 16 11:17 genes.count_tracking
-rw-r--r-- 1 bioinfo bioinfo 333K Apr 16 11:17 genes.fpkm_tracking
-rw-r--r-- 1 bioinfo bioinfo 564K Apr 16 11:17 genes.read_group_tracking
-rw-r--r-- 1 bioinfo bioinfo 348K Apr 16 11:17 isoform_exp.diff
-rw-r--r-- 1 bioinfo bioinfo 189K Apr 16 11:17 isoforms.count_tracking
-rw-r--r-- 1 bioinfo bioinfo 346K Apr 16 11:17 isoforms.fpkm_tracking
-rw-r--r-- 1 bioinfo bioinfo 586K Apr 16 11:17 isoforms.read_group_tracking
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 promoters.diff
-rw-r--r-- 1 bioinfo bioinfo  466 Apr 16 11:17 read_groups.info
-rw-r--r-- 1 bioinfo bioinfo  451 Apr 16 11:17 run.info
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 splicing.diff
-rw-r--r-- 1 bioinfo bioinfo  124 Apr 16 11:17 tss_group_exp.diff
-rw-r--r-- 1 bioinfo bioinfo   12 Apr 16 11:17 tss_groups.count_tracking
-rw-r--r-- 1 bioinfo bioinfo   91 Apr 16 11:17 tss_groups.fpkm_tracking
-rw-r--r-- 1 bioinfo bioinfo  115 Apr 16 11:17 tss_groups.read_group_tracking
```

extract genes that have a significant value for differential expression

```
awk -F "\t" 'BEGIN {OFS = "\t"} $14 = "yes" {print $0}' gene_exp.diff >
significant_genes.txt
```

Expression

Cufflinks can also be used for:

- ✦ strand-specific RNA-seq
- ✦ novel transcript discovery in annotated genomes
- ✦ identification of novel splice variants
- ✦ detecting transcripts in genomes without annotation

For protocols see:

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks (Trapnell et al., 2012).