基礎生物学研究所　ゲノムインフォマティクス・トレーニングコース　2015秋

RNA-seq解析パイプライン：
Transcript-based pipeline

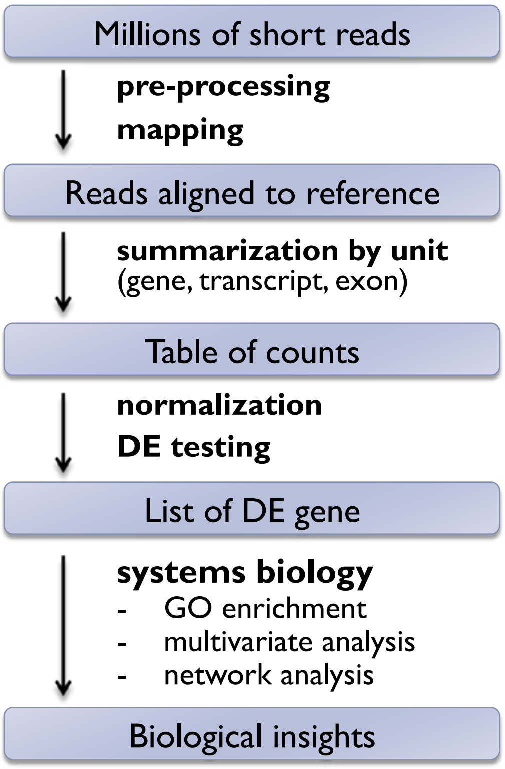
Shuji Shigenobu
重信　秀治

基礎生物学研究所
生物機能解析センター

Two Basic Pipelines

- Choice of reference
  - **Genome** – standard for genome-known species
  - **Transcript** – the only way for genome-unknown species
    -- can be used for genome-known species

RNA-seq workflow with reference genome



Millions of short reads

pre-processing

mapping

Reads aligned to reference

summarization by unit
(gene, transcript, exon)

Table of counts

normalization

DE testing

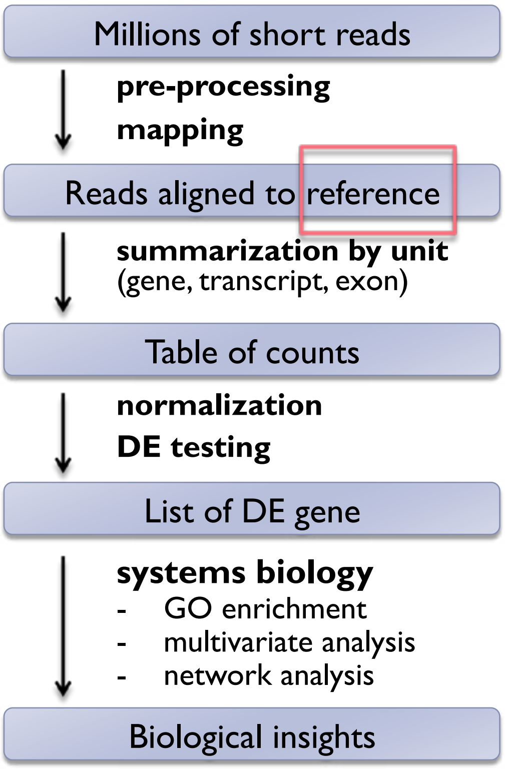
List of DE gene

systems biology

- GO enrichment
- multivariate analysis
- network analysis

Biological insights

RNA-seq workflow **without** reference genome



Millions of short reads

pre-processing

mapping

Reads aligned to reference

summarization by unit
(gene, transcript, exon)

Table of counts

normalization

DE testing

List of DE gene

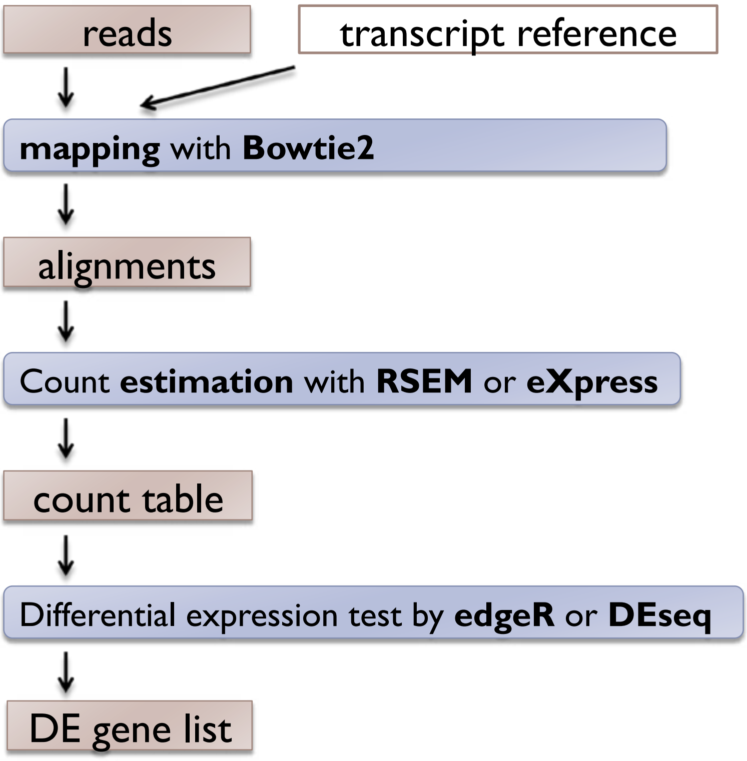
systems biology

- GO enrichment
- multivariate analysis
- network analysis

Biological insights

1. **Build** reference
2. **Characterize** reference

A Pipeline: Transcript-based



reads

transcript reference

mapping with **Bowtie2**

alignments

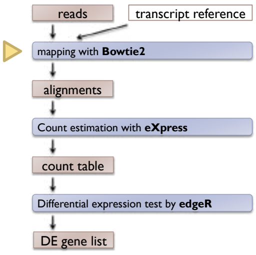
Count **estimation** with **RSEM** or **eXpress**

count table

Differential expression test by **edgeR** or **DEseq**

DE gene list

Mapping – alignment software



reads

transcript reference

mapping with **Bowtie2**

alignments

Count estimation with **eXpress**

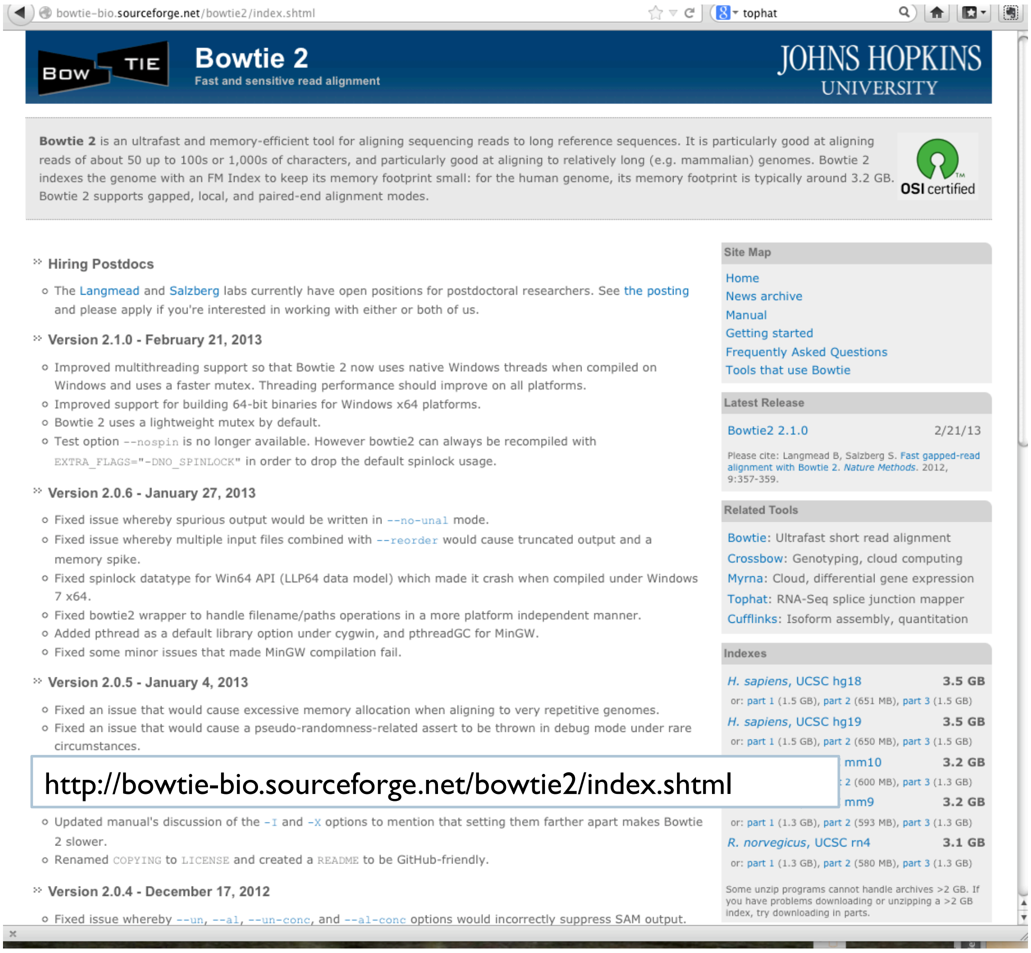
count table

Differential expression test by **edgeR**

DE gene list

- For mapping reads onto transcript reference *short read mapper (unspliced read aligner)* is used
  - **Bowtie2** – basic mapping to reference sequence

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml



Bowtie 2

Fast and sensitive read alignment

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

Hiring Postdocs

- The Langmead and Salzberg labs currently have open positions for postdoctoral researchers. See the posting and please apply if you're interested in working with either or both of us.

Version 2.1.0 - February 21, 2013

- Improved multithreading support so that Bowtie 2 now uses native Windows threads when compiled on Windows and uses a faster mutex. Threading performance should improve on all platforms.
- Improved support for building 64-bit binaries for Windows x64 platforms.
- Bowtie 2 uses a lightweight mutex by default.
- Test option `--nospin` is no longer available. However bowtie2 can always be recompiled with `EXTRA_FLAGS="-DNO_SPINLOCK"` in order to drop the default spinlock usage.

Version 2.0.6 - January 27, 2013

- Fixed issue whereby spurious output would be written in `--no-unal` mode.
- Fixed issue whereby multiple input files combined with `--reorder` would cause truncated output and a memory spike.
- Fixed spinlock datatype for Win64 API (LLP64 data model) which made it crash when compiled under Windows 7 x64.
- Fixed bowtie2 wrapper to handle filename/paths operations in a more platform independent manner.
- Added pthread as a default library option under cygwin, and pthreadGC for MinGW.
- Fixed some minor issues that made MinGW compilation fail.

Version 2.0.5 - January 4, 2013

- Fixed an issue that would cause excessive memory allocation when aligning to very repetitive genomes.
- Fixed an issue that would cause a pseudo-randomness-related assert to be thrown in debug mode under rare circumstances.

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

- Updated manual's discussion of the `-I` and `-X` options to mention that setting them farther apart makes Bowtie 2 slower.
- Renamed `COPYING` to `LICENSE` and created a `README` to be GitHub-friendly.

Version 2.0.4 - December 17, 2012

- Fixed issue whereby `--un`, `--al`, `--un-conc`, and `--al-conc` options would incorrectly suppress SAM output.

Site Map

Home
News archive
Manual
Getting started
Frequently Asked Questions
Tools that use Bowtie

Latest Release

Bowtie2 2.1.0 2/21/13

Please cite: Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

Related Tools

Bowtie: Ultrafast short read alignment
Crossbow: Genotyping, cloud computing
Myrna: Cloud, differential gene expression
Tophat: RNA-Seq splice junction mapper
Cufflinks: Isoform assembly, quantitation

Indexes

H. sapiens, UCSC hg18 **3.5 GB**
or: part 1 (1.5 GB), part 2 (651 MB), part 3 (1.5 GB)
H. sapiens, UCSC hg19 **3.5 GB**
or: part 1 (1.5 GB), part 2 (650 MB), part 3 (1.5 GB)
mm10 **3.2 GB**
2 (600 MB), part 3 (1.3 GB)
mm9 **3.2 GB**
or: part 1 (1.3 GB), part 2 (593 MB), part 3 (1.3 GB)
R. norvegicus, UCSC rn4 **3.1 GB**
or: part 1 (1.3 GB), part 2 (580 MB), part 3 (1.3 GB)

Some unzip programs cannot handle archives >2 GB. If you have problems downloading or unzipping a >2 GB index, try downloading in parts.

bowtie2

Bowtie is an ultrafast, memory-efficient short read aligner.

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

```
(example)

$ bowtie2 –x transcript.fa –U reads.fq –a –S out.sam
```

- Output
  - Alignment in SAM format : **`out.sam`**

(ex1) Let’s Try Bowtie2

Align 75-bp Illumina reads with a transcript reference using Bowtie2.

Prepare reads and reference genome

Sequences for this exercise are stored in **`~/data/SS/`**.

```
IlluminaReads1.fq – Illumina reads in fastq format
minimouse_mRNA.fa – a set of transcript sequnences
```

Build index of reference sequence

```
$bowtie2-build minimouse_mRNA.fa myref
```

Align reads with reference

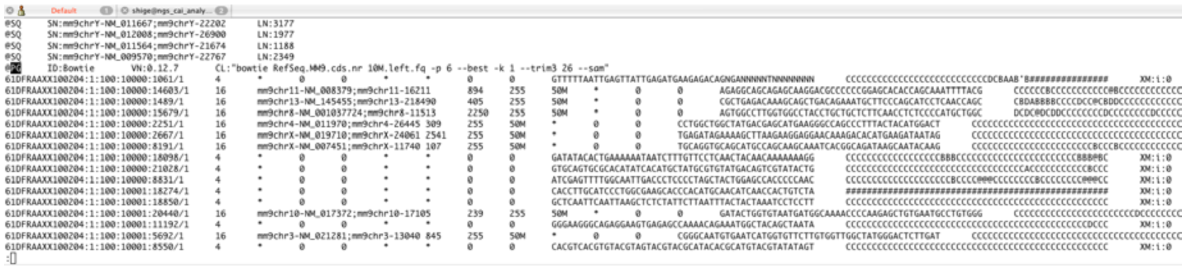
```
$bowtie2 –x myref –U IlluminaReads1.fq –a –S out.sam
```

Evaluation of mapping results

- Evaluation of SAM/BAM file
- Check statistics
- Visualization

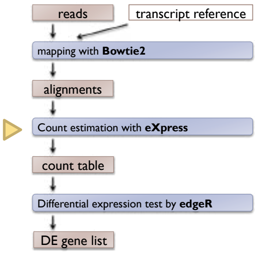
```
(example)

$ samtools view bowtieout.bam
```



```
@SQ     SN:mm9chrY-NM_011667;mm9chrY-22202      LN:3177
@SQ     SN:mm9chrY-NM_012008;mm9chrY-26900      LN:1977
@SQ     SN:mm9chrY-NM_011564;mm9chrY-21674      LN:1188
@SQ     SN:mm9chrY-NM_009570;mm9chrY-22767      LN:2349
@PG     ID:Bowtie       VN:0.12.7       CL:"bowtie RefSeq.MM9.cds.nr 10M.left.fq -p 6 --best -k 1 --trim3 26 --sam"
61DFRAAXX100204:1:100:10000:1061/1      4       *       0       0       *       *       0       0       GTTTTTAATTGAGTTATTGAGATGAAGAGACAGNGANNNNNTNNNNNNNN      CCCCCCCCCCCCCCCCCCCCCCCCCCDCBAAB'B###############       XM:i:0
61DFRAAXX100204:1:100:10000:14603/1     16      mm9chr11-NM_008379;mm9chr11-16211       894     255     50M     *       0       0       AGAGGCAGCAGAGCAAGGACGCCCCCGGAGCACACCAGCAAATTTTACG       CCCCCCBCCCCCCCCCC@BCCCCCCCCCCC
61DFRAAXX100204:1:100:10000:1489/1      16      mm9chr13-NM_145455;mm9chr13-218490      405     255     50M     *       0       0       CGCTGAGACAAAGCAGCTGACAGAAATGCTTCCCAGCATCCTCAACCAGC       CBDABBBBCCCCDCC@CBDDCCCCCCCCCCC
61DFRAAXX100204:1:100:10000:15679/1     16      mm9chr8-NM_001037724;mm9chr8-11513      2250    255     50M     *       0       0       AGTGGCCTTGGTGGCCTACCTGCTGCTCTTCAACCTCTCCCCATGCTGGC       DCDC@DCDDCCCCCCCCDCCCCCCCCDCCCCC
61DFRAAXX100204:1:100:10000:2251/1      16      mm9chr4-NM_011970;mm9chr4-26445 309     255     50M     *       0       0       CCTGGCTGGCTATGACGAGCATGAAGGGCCAGCCCTTTACTACATGGACT       CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
61DFRAAXX100204:1:100:10000:2667/1      16      mm9chrX-NM_019710;mm9chrX-24061 2541    255     50M     *       0       0       TGAGATAGAAAAGCTTAAGAAGGAGGAACAAAGACACATGAAGATAATAG       CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
61DFRAAXX100204:1:100:10000:8191/1      16      mm9chrX-NM_007451;mm9chrX-11740 107     255     50M     *       0       0       TGCAGGTGCAGCATGCCAGCAAGCAAATCACGGCAGATAAGCAATACAAG       CCCCCCCCCCCCCCCCCCCCCCCBCCCBCCCCCCCCCCCCC
61DFRAAXX100204:1:100:10000:18098/1     4       *       0       0       *       *       0       0       GATATACACTGAAAAAATAATCTTTGTTCCTCAACTACAACAAAAAAAGG      CCCCCCCCCCCCCCCCCCBBBCCCCCCCCCCCCCCCCCCCCCCBBB@BC       XM:i:0
61DFRAAXX100204:1:100:10000:21028/1     4       *       0       0       *       *       0       0       GTGCAGTGCGCACATATCACATGCTATGCGTGTATGACAGTCGTATACTG      CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCCCCCBCCC       XM:i:0
61DFRAAXX100204:1:100:10000:8831/1      4       *       0       0       *       *       0       0       ATCGAGTTTTGGCAATTGACCCTCCCCTAGCTACTGGAGCCACCCCCAAC      CCCCCCCCCCCCCCCCCCCCBCCCC@@@CCCCCCCCBCCCCCCCC@@@CC      XM:i:0
61DFRAAXX100204:1:100:10001:18274/1     4       *       0       0       *       *       0       0       CACCTTGCATCCCTGGCGAAGCACCCACATGCAACATCAACCACTGTCTA      ##################################################      XM:i:0
61DFRAAXX100204:1:100:10001:18850/1     4       *       0       0       *       *       0       0       GCTCAATTCAATTAAGCTCTCTATTCTTAATTTACTACTAAATCCTCCTT      CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC      XM:i:0
61DFRAAXX100204:1:100:10001:20440/1     16      mm9chr10-NM_017372;mm9chr10-17105       239     255     50M     *       0       0       GATACTGGTGTAATGATGGCAAAACCCCAAGAGCTGTGAATGCCTGTGGG       CCCCCCCCCCCCCCCCCCCCCCCCCDCCCCCCCC
61DFRAAXX100204:1:100:10001:11192/1     4       *       0       0       *       *       0       0       GGGAAGGGCAGAGGAAGTGAGAGCCAAAACAGAAATGGCTACAGCTAATA      CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCDCCC       XM:i:0
61DFRAAXX100204:1:100:10001:5692/1      16      mm9chr3-NM_021281;mm9chr3-13040 845     255     50M     *       0       0       CGGGCAATGTGAATCATGGTGTTCTTGTGGTTGGCTATGGGACTCTTGAT       CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
61DFRAAXX100204:1:100:10001:8550/1      4       *       0       0       *       *       0       0       CACGTCACGTGTACGTAGTACGTACGCATACACGCATGTACGTATATAGT      CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC      XM:i:0
:
```

Count Reads by Transcript/gene


reads

transcript reference

mapping with **Bowtie2**

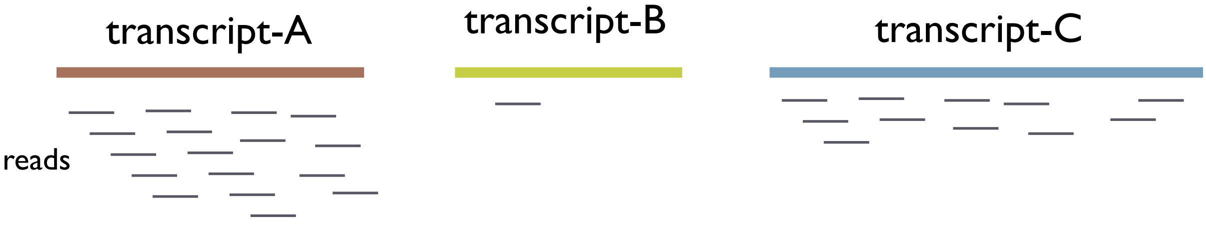
alignments

Count estimation with **eXpress**

count table

Differential expression test by **edgeR**

DE gene list



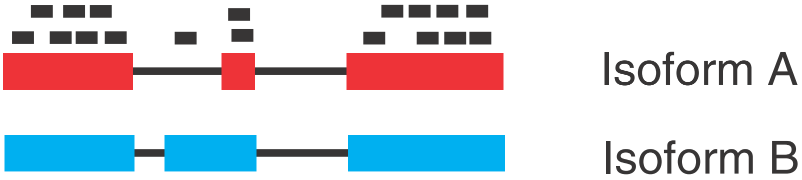
- The simplest way: just count reads by contig.

But…

- Mapping ambiguity should be taken into consideration.

Estimate Abundance

- **Multimapping issues**
  - Isoforms
  - Very similar paralogs
  - Repetitive sequences
  - => cannot align reads uniquely
- Mapping ambiguity should be taken into consideration.



- Critical for RNA-seq de novo analysis
- Software: RSEM and eXpress (EM algorithm)

eXpress

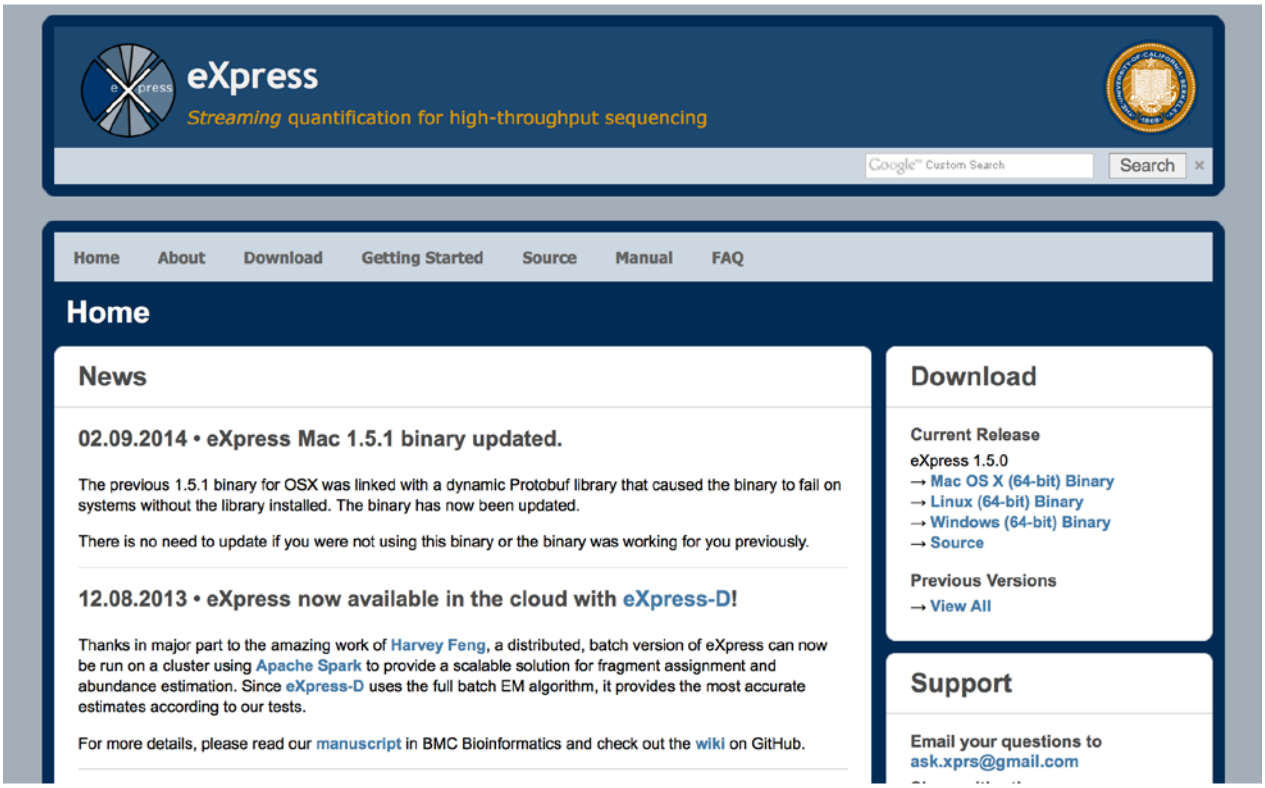
eXpress is a streaming tool for quantifying the abundances of a set of target sequences from sampled subsequences.

http://bio.math.berkeley.edu/eXpress/

```
(example)

$ express transcripts.fasta hits.bam
```

- Output
  - Count estimation table: **`results.xprs`**



Home

News

02.09.2014 • eXpress Mac 1.5.1 binary updated.

The previous 1.5.1 binary for OSX was linked with a dynamic Protobuf library that caused the binary to fail on systems without the library installed. The binary has now been updated.

There is no need to update if you were not using this binary or the binary was working for you previously.

12.08.2013 • eXpress now available in the cloud with eXpress-D!

Thanks in major part to the amazing work of Harvey Feng, a distributed, batch version of eXpress can now be run on a cluster using Apache Spark to provide a scalable solution for fragment assignment and abundance estimation. Since eXpress-D uses the full batch EM algorithm, it provides the most accurate estimates according to our tests.

For more details, please read our manuscript in BMC Bioinformatics and check out the wiki on GitHub.

Download

Current Release

eXpress 1.5.0
→ Mac OS X (64-bit) Binary
→ Linux (64-bit) Binary
→ Windows (64-bit) Binary
→ Source

Previous Versions
→ View All

Support

Email your questions to
ask.xprs@gmail.com

http://bio.math.berkeley.edu/eXpress/index.html

(ex1) Let’s Try eXpress

Align 75-bp Illumina reads with a transcript reference using Bowtie2.

Prepare alignments and reference genome

Sequences for this exercise are stored in **`~/data/SS/`**.

```
IlluminaReads1.fq – Illumina reads in fastq format
out.sam – this file should be generated in the previous bowtie practice
```

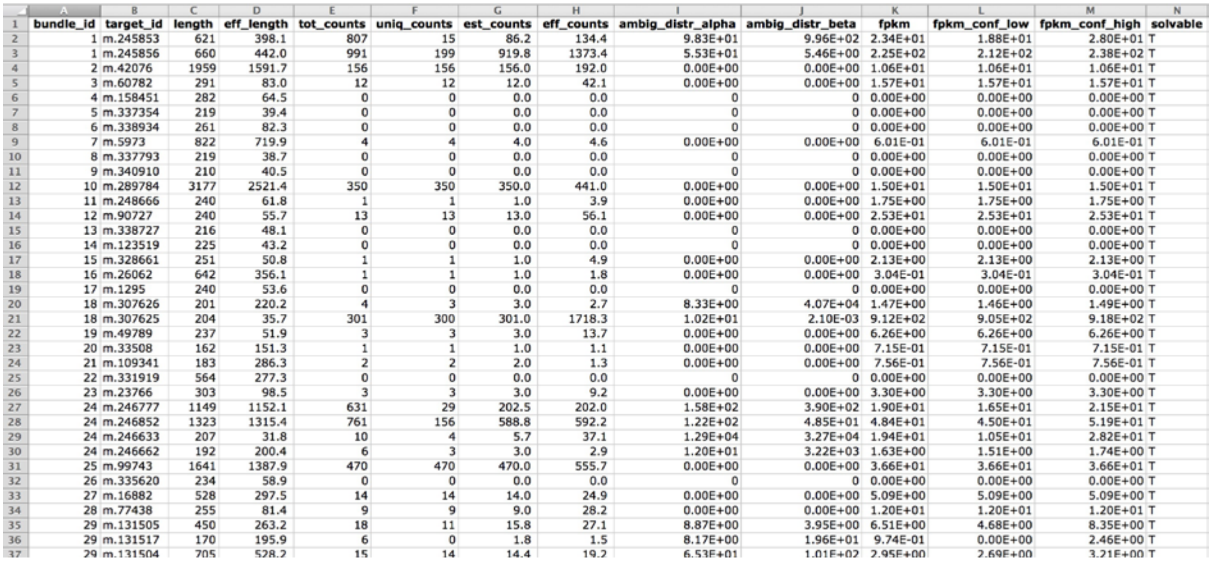
Run eXpress

```
$ express minimouse_mRNA.fa out.sam
```

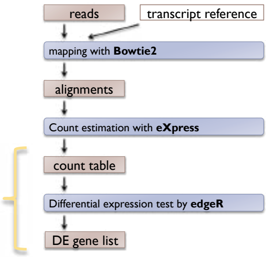
```
Output : results.xprs, params.xprs
```

eXpress: output

`results.xprs`



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	bundle_id	target_id	length	eff_length	tot_counts	uniq_counts	est_counts	eff_counts	ambig_distr_alpha	ambig_distr_beta	fpkm	fpkm_conf_low	fpkm_conf_high	solvable
2	1	m.245853	621	398.1	807	15	86.2	134.4	9.83E+01	9.96E+02	2.34E+01	1.88E+01	2.80E+01	T
3	1	m.245856	660	442.0	991	199	919.8	1373.4	5.53E+01	5.46E+00	2.25E+02	2.12E+02	2.38E+02	T
4	2	m.42076	1959	1591.7	156	156	156.0	192.0	0.00E+00	0.00E+00	1.06E+01	1.06E+01	1.06E+01	T
5	3	m.60782	291	83.0	12	12	12.0	42.1	0.00E+00	0.00E+00	1.57E+01	1.57E+01	1.57E+01	T
6	4	m.158451	282	64.5	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
7	5	m.337354	219	39.4	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
8	6	m.338934	261	82.3	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
9	7	m.5973	822	719.9	4	4	4.0	4.6	0.00E+00	0.00E+00	6.01E-01	6.01E-01	6.01E-01	T
10	8	m.337793	219	38.7	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
11	9	m.340910	210	40.5	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
12	10	m.289784	3177	2521.4	350	350	350.0	441.0	0.00E+00	0.00E+00	1.50E+01	1.50E+01	1.50E+01	T
13	11	m.248666	240	61.8	1	1	1.0	3.9	0.00E+00	0.00E+00	1.75E+00	1.75E+00	1.75E+00	T
14	12	m.90727	240	55.7	13	13	13.0	56.1	0.00E+00	0.00E+00	2.53E+01	2.53E+01	2.53E+01	T
15	13	m.338727	216	48.1	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
16	14	m.123519	225	43.2	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
17	15	m.328661	251	50.8	1	1	1.0	4.9	0.00E+00	0.00E+00	2.13E+00	2.13E+00	2.13E+00	T
18	16	m.26062	642	356.1	1	1	1.0	1.8	0.00E+00	0.00E+00	3.04E-01	3.04E-01	3.04E-01	T
19	17	m.1295	240	53.6	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
20	18	m.307626	201	220.2	4	3	3.0	2.7	8.33E+00	4.07E+04	1.47E+00	1.46E+00	1.49E+00	T
21	18	m.307625	204	35.7	301	300	301.0	1718.3	1.02E+01	2.10E-03	9.12E+02	9.05E+02	9.18E+02	T
22	19	m.49789	237	51.9	3	3	3.0	13.7	0.00E+00	0.00E+00	6.26E+00	6.26E+00	6.26E+00	T
23	20	m.33508	162	151.3	1	1	1.0	1.1	0.00E+00	0.00E+00	7.15E-01	7.15E-01	7.15E-01	T
24	21	m.109341	183	286.3	2	2	2.0	1.3	0.00E+00	0.00E+00	7.56E-01	7.56E-01	7.56E-01	T
25	22	m.331919	564	277.3	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
26	23	m.23766	303	98.5	3	3	3.0	9.2	0.00E+00	0.00E+00	3.30E+00	3.30E+00	3.30E+00	T
27	24	m.246777	1149	1152.1	631	29	202.5	202.0	1.58E+02	3.90E+02	1.90E+01	1.65E+01	2.15E+01	T
28	24	m.246852	1323	1315.4	761	156	588.8	592.2	1.22E+02	4.85E+01	4.84E+01	4.50E+01	5.19E+01	T
29	24	m.246633	207	31.8	10	4	5.7	37.1	1.29E+04	3.27E+04	1.94E+01	1.05E+01	2.82E+01	T
30	24	m.246662	192	200.4	6	3	3.0	2.9	1.20E+01	3.22E+03	1.63E+00	1.51E+00	1.74E+00	T
31	25	m.99743	1641	1387.9	470	470	470.0	555.7	0.00E+00	0.00E+00	3.66E+01	3.66E+01	3.66E+01	T
32	26	m.335620	234	58.9	0	0	0.0	0.0	0	0	0.00E+00	0.00E+00	0.00E+00	T
33	27	m.16882	528	297.5	14	14	14.0	24.9	0.00E+00	0.00E+00	5.09E+00	5.09E+00	5.09E+00	T
34	28	m.77438	255	81.4	9	9	9.0	28.2	0.00E+00	0.00E+00	1.20E+01	1.20E+01	1.20E+01	T
35	29	m.131505	450	263.2	18	11	15.8	27.1	8.87E+00	3.95E+00	6.51E+00	4.68E+00	8.35E+00	T
36	29	m.131517	170	195.9	6	0	1.8	1.5	8.17E+00	1.96E+01	9.74E-01	0.00E+00	2.46E+00	T
37	29	m.131504	705	528.2	15	14	14.4	19.2	6.53E+01	1.01E+02	2.95E+00	2.69E+00	3.21E+00	T



reads

transcript reference

mapping with **Bowtie2**

alignments

Count estimation with **eXpress**

count table

Differential expression test by **edgeR**

DE gene list

Table of counts

data import

diagnostics

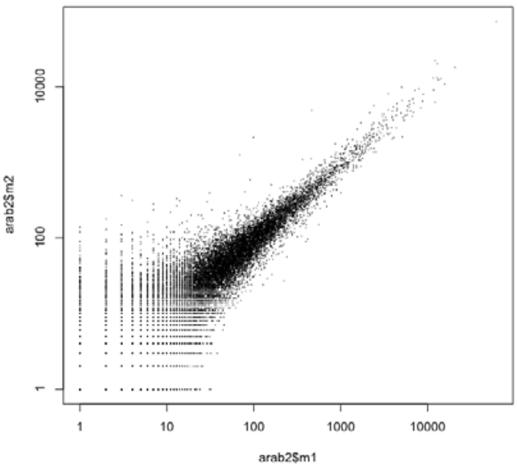
normalization

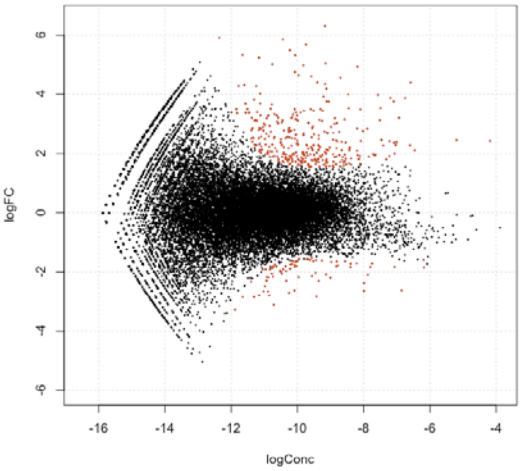
DE testing

evaluation

List of DE gene

Diagnostics: Scatter plot & MA plot





edgeR

- A Bioconductor package for differential expression analysis of digital gene expression data
- **Model**: An over dispersed Poisson model, negative binomial (NB) model, is used
- **Normalization**: TMM method (trimmed mean of M values) to deal with composition effects
- **DE test**: exact test and generalized linear models (GLM)

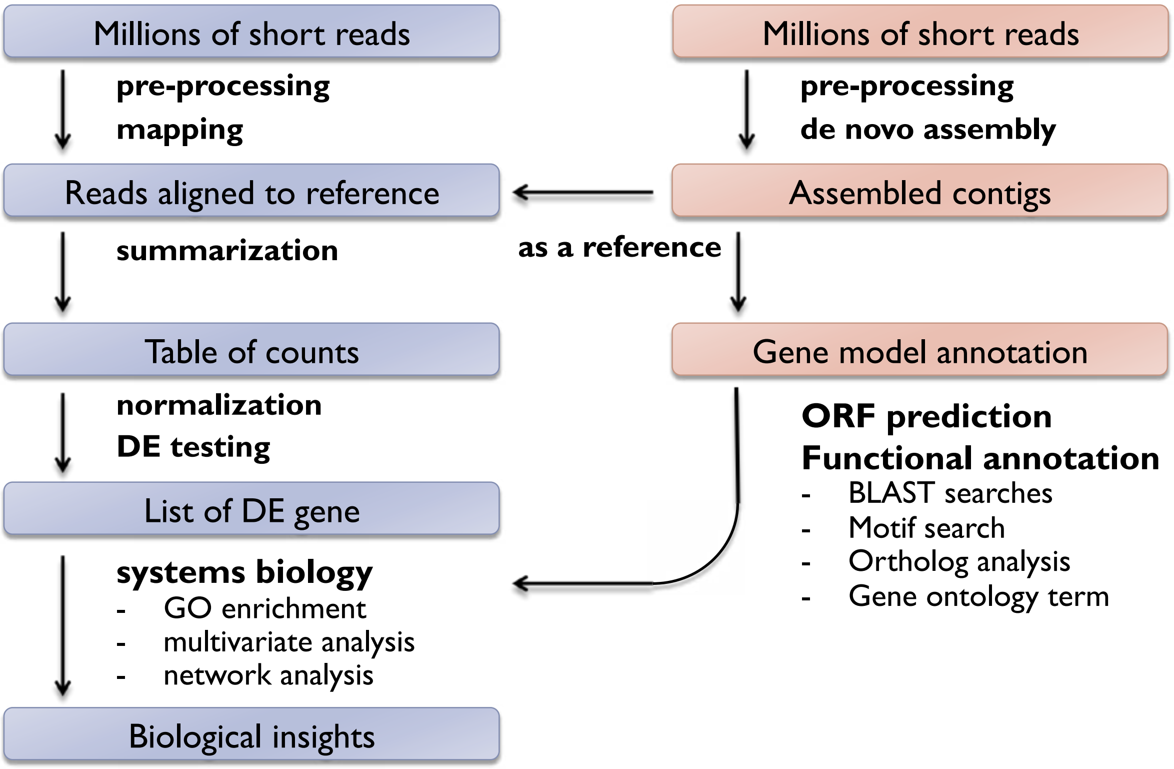
edgeR

- input: count data (not RPKM)
- output: gene table with DE significance statistics (FDR)

```
(example)

$ R
> library(edgeR)                        #load edgeR library
> dat <- read.delim(“count_data.txt”, …) #import count table to R
> group <- c(rep("M", 3), rep("H", 3))   #assign groups
> D <- DGEList(dat, group=group)        #import data to edgeR
> D <- calcNormFactors(D)               #normalization (TMM)
> D <- estimateCommonDisp(D)            #estimate common dispersion
> D <- estimateTagwiseDisp(D, …)        #estimate tagwise dispersion
> de <- exactTest(D, pair=c("M", "H")) #DE test
> topTags(de)
Comparison of groups: H-M
            logConc    logFC      P.Value          FDR
AT5G48430 -15.36821 6.255498 9.919041e-12 2.600872e-07
AT5G31702 -15.88641 5.662522 3.637593e-10 4.083773e-06
AT3G55150 -17.01537 5.870635 4.672331e-10 4.083773e-06
…
```

RNA-seq analysis pipeline (***de novo*** strategy)



Millions of short reads

pre-processing

mapping

Reads aligned to reference

summarization

Table of counts

normalization

DE testing

List of DE gene

systems biology

- GO enrichment
- multivariate analysis
- network analysis

Biological insights

Millions of short reads

pre-processing

de novo assembly

Assembled contigs

as a reference

Gene model annotation

ORF prediction

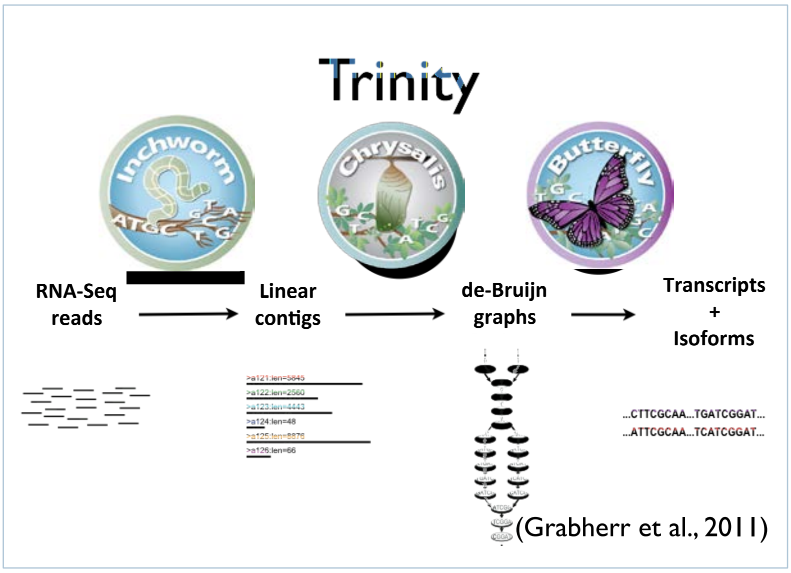
Functional annotation

- BLAST searches
- Motif search
- Ortholog analysis
- Gene ontology term

de novo assemblers of RNA-seq

De novo assemblers use reads to assemble transcripts directly, which does not depend on a reference gnome.

- Trinity
- Oases
- TransAbyss
- EBARDenovo
- …



(Grabherr et al., 2011)

http://trinityrnaseq.sourceforge.net/

Trinity example

- Input: Illumina short reads in FASTQ | FASTA format
- Output: assembled contigs in FASTA format

```
# prepare input reads
$ cat *.R1.fastq > left_all.fq
$ cat *.R2.fastq > right_all.fq

# Run Trinity
$ Trinity --seqType fq --left left_all.fq --right right_all.fq --
CPU 8 --max_memory 20G
```

(Trinity is supported on only Linux)

optional

ORF prediction

- Special consideration in ORF prediction after de novo RNA-seq assembly
  - Sometimes partial: Start Met or terminal codon may be missing.
  - Ideally one ORF is present per contig, but erroneously joined contigs may include multiple ORFs.
  - Possible frame shifts.
    - Frame shifts do not occur so often in Illumina, while it happens very frequently in 454 and IonProton.

optional

Functional Annotation of Predicted ORFs

- BLAST
  - NCBI NR (or UniProt)
  - species of interest (model organisms, close relatives etc)
  - specific DB (SwissProt, rRNA DB, CEGMA etc)
  - self (assembly v.s. assembly)
- Motif search
  - Pfam, SignalP etc.
- Ortholog analysis
  - vs model organism
  - ortholog database (OrthoDB, eggNOG, OrthoMCL etc)
  - close relatives
- Gene Ontology term assignment

optional

Quick annotation by BLASTX

- Query: assembled contigs

  (nucleotide sequences in multi-fasta format)
- DB: Protein sequences of a model organism

Format DB

```
$ makeblastdb –in protein.fa -dbtype prot
```

Search

```
$ blastx -query trinity_contigs –db protein.fa \
  -num_threads 8 -evalue 1.0e-8 –outfmt 0 > blastxout.txt
```

optional

Let’s try BLASTX

- Query: minimouse_mRNA.fa
- DB: human.protein.faa (human RefSeq protein)

1. Format DB

```
$ makeblastdb –in human.protein.faa -dbtype prot –parse_seqids
```

2. Search

```
$ blastx -query minimouse_mRNA.fa –db human.protein.faa \
  -num_threads 8 -evalue 1.0e-8 –outfmt 0 > blastxout.txt
```

```
$ blastx -query minimouse_mRNA.fa –db human.protein.faa \
  -num_threads 8 -evalue 1.0e-8 –outfmt 7 > blastxout.tab
```