Short read sequence analysis

Manuel Garber
Krumlov 2015
Overview of the session

- Explaining diversity: Transcriptional regulation
  - A short story from our recent work

- RNA Sequencing
  - The different BLA-Seq libraries. A common theme
  - Read mapping (alignment): Placing short reads in the genome
  - Quantification:
    - Assigning scores to regions
    - Finding regions that are differentially represented between two or more samples.
    - How much depth?
    - Reconstruction: Finding the regions that originated the reads

- RNA-Seq Vignette: non-coding RNA evolution
Why do organisms look the way that they do?
Why do different cell types do what they do!

However, all this diversity arises from the same genome sequence! Proteins are very conserved across vertebrates, what is the driving force of variability?
Cell identity is determined by gene regulation

Positive Feedback Between PU.1 and the Cell Cycle Controls Myeloid Differentiation

Hao Yuan Kueh,1* Ameya Champhekar,1 Stephen L. Nutt,2 Michael B. Elowitz,1,3 Ellen V. Rothenberg1**

Diagram showing the cell cycle control of PU.1 in different cell populations:

- Mac: PU.1 high
- Pro: PU.1 intermediate
- B: PU.1 low
Dissecting neural differentiation regulatory networks through epigenetic footprinting

Michael J. Ziller1,2,3*, Reuven Edri4*, Yakey Yaffe5, Julie Donaghy1,2,3, Ramona Pop1,2,3, William Mallard1,2, Robbyn Issner4, Casey A. Giftord1,2,3, Alon Goren1,2,3, Jeffrey Xing4, Hongcang Gu5, Davide Cacchiarelli5, Alexander M. Tsankov4,2,3, Charles Epstein1, John L. Rin1,2,3, Tarjei S. Mikkelson1, Oliver Kohlbacher2, Andreas Gnirke3, Bradley E. Bernstein1,5,6, Yechiel Elkabetz4 & Alexander Meissner1,5,6*

Transcription factors regulate gene programs. Epigenome informs (determines?) potential for expression
Multicellular development requires complex regulation

Proximal TFBS

TSS

CRM

Enhancer

Insulator

Chromatin

Genic region

Proximal promoter

Figure 1

Regulation of transcription.

a | A summary of promoter elements and regulatory signals. Chromatin is comprised of DNA wrapped around histones to form nucleosomes. The structure of chromatin can be tightly wrapped or accessible to proteins. Boundaries between these states may be marked by insulators. The region around the transcription start site (TSS) is often divided into a larger proximal promoter upstream of the TSS and a smaller core promoter just around the TSS. The exact boundaries vary between studies. To recruit RNA polymerase II (RNAPII) and to activate transcription of the gene, sequence-specific regulatory proteins (transcription factors) bind to specific sequence patterns (namely, transcription factor binding sites (TFBSs)) that are near to the TSS (proximal elements) or that are far away from it (enhancers). TFBSs can also occur in clusters, forming cis-regulatory modules (CRMs).

b | Sequence patterns in core promoters. The region around the TSS has several over-represented sequence patterns; the TATA box and initiator (Inr) are the most studied and most prevalent. The location of patterns relative to the TSS and their sequence properties are shown as boxes and as associated sequence logos based on the JASPAR database. The Inr pattern is not shown as it varies considerably between studies, ranging from a TCA(G/T)TC(C/T) to a single dinucleotide (pyrimidine (C/T)–purine (A/G)). Importantly, most promoters only have one or a few of these patterns, and some patterns are mostly found in certain species. BRE, B recognition element; DCE, downstream core element; DRE, DNA recognition element; MTE, motif ten element. Figure modified, with permission, from REF. 91 © (2004) Macmillan Publishers Ltd. All rights reserved.
Indeed Enhancers are both species and cell type specific.

Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants

Stephen C. J. Parker, Michael L. Stitzel, D. Leland Taylor, Jose Miguel Orozco, Michael R. Erdos, Jennifer A. Akiyama, Kelly Lammerts van Bueren, Peter S. Chines, Narisu Narisu, NISC Comparative Sequencing Program, Brian L. Black, Axel Visel, Len A. Pennacchio, and Francis S. Collins

Histone modifications at human enhancers reflect global cell-type-specific gene expression


Enhancer elements are poorly conserved, are cell type specific, How do we find them?
DNA is not naked
Nucleosomes interact with nuclear factors through tails

Histone tails

146 bp
Cell identity is determined by its epigenetic state

Catherine Dulac, Nature 2010
Which controls the genome functional elements

Active promoter
- H3K4me2
- H3K4me3
- Acetylation H2A.Z

Active transcription
- KNAPII

Active enhancer
- CTCF
- Boundary
- H3K4me1
- H3K4me2
- H3K27ac

Repressive marks
- H3K9me2
- H3K9me3
- H3K27me3

Dissecting a gene regulatory network

We want to ultimately understand the cell circuits of the cell

Comparative genomics – measure constraint

- new methods
- models of transcriptional regulation
- models of epigenetic interactions
- perturbations

ChIP

RNA

RNA-Protein interactions
Understanding innate immunity

Figure 1 | Dendritic cells and immunoregulation.

Non-lymphoid tissues
- Immature dendritic cell
- Steady-state migration
- Danger-signal-induced migration

Draining lymph nodes
- Danger signal A
- Danger signal B
- Foreign antigens

Activated DC
- T_H1 bias
- T_H1

Immunity
- MHC II presentation
- CD4 T cell
- CD8 T cell
- Deletion Regulatory cells

Tolerance
- MHC I cross-presentation
- Self-antigens
- CD4 T cell
- CD8 T cell
- Deletion Regulatory cells

Steady-state DC
- MHC I presentation
- Self-antigens
- CD4 T cell

Quiescent DC
- MHC I cross-presentation
- Self-antigens
- CD8 T cell

Danger
- Foreign antigens
- CD4 T cell
- CD8 T cell

Activated DC
- T_H1 bias
- T_H1

Danger-signal-induced migration
- Danger
- Foreign antigens
- CD4 T cell
- CD8 T cell
LPS (TLR4 receptor) stimulation as it elicits the most broad gene expression response.
Chip-Seq + RNA-Seq to map and relate components

Sequencing libraries allow us to map output, state and the circuit of the cell
Transcription factors control specific pathways

Inflammation:
- B, T activation
- Anti-viral
- Inflammation
- Anti-apoptotic
- Cell cycle

RNA Pol II
TF fold change (log2)
4–4
Binding score at t = 0
950
log2
3–3

Late induced
Immediate early
Intermediate
Repressed

PU.1
Cebpb
Atf3
Junb
Irf4
Irf1
Egr2
Runx1
Rela
Relb
Nfkb1
E2f1
E2f4
Ets2
Egr1
Stat1
Stat2
Cd274
Cd38
Tnfrsf14
Stat1
Irf7
Mx2
TNF
Cxcl2
Nfkbia

Cell cycle
Specific factors control amplitude of expression

500 late induced genes

How different is the regulation of different expression patterns?
Different control of early vs. late induced genes

Late induced genes:
- Regulated by few factors

Early induced genes:
- Highly pre-bound
Factors that control early induced genes are more redundant
Conclusions and considerations

- A large fraction of binding exist prior to stimulus
- Immediate vs. late regulation is drastically different:
  - Early induced genes regulators are more redundant
  - Late induced regulators are less redundant
  - Are the early inflammation pathways evolutionary more malleable?
- Factors act in layers, consistent with previous reports
- Genomic approaches like this are applicable to many systems
  - Protocols can handle smaller input material (Alon Goren, Oren Ram, Amit)
- Test models using a genome wide genetic screens
- Map TFs with no available antibodies
Sequencing: applications

Counting applications
- Profiling
  - microRNAs
  - Immunogenomics
  - Transcriptomics
- Epigenomics
  - Map histone modifications
  - Map DNA methylation
  - 3D genome conformation
- Nucleic acid Interactions

Polymorphism/mutation discovery
- Bacteria
- Genome dynamics
- Exon (and other target) sequencing
- Disease gene sequencing

Variation and association studies
- Genetics and gene discovery
- Cancer genomics
  - Map translocations, CNVs, structural changes
  - Profile somatic mutations
- Genome assembly
- Ancient DNA (Neanderthal)
- Pathogen discovery
- Metagenomics
Sequencing libraries to probe the genome

- **RNA-Seq**
  - Transcriptional output
  - Annotation
  - miRNA
  - Ribosomal profiling

- **ChIP-Seq**
  - Nucleosome positioning
  - Open/closed chromatin
  - Transcription factor binding

- **CLIP-Seq**
  - Protein-RNA interactions

- **Hi-C**
  - 3D genome conformation
RNA-Seq libraries I: “Standard” full-length

- “Source: intact, **high qual.** RNA (polyA selected or ribosomal depleted)
- RNA → cDNA → sequence
- Uses:
  - Annotation. Requires high depth, paired-end sequencing. ~50 mill
  - Gene expression. Requires low depth, single end sequence, ~5-10 mill
  - Differential Gene expression. Requires ~ 5-10 mill, at least 3 replicates, single end
RNA-Seq libraries II: End-sequence libraries

- Target the start or end of transcripts.
- Source: End-enriched RNA
  - Fragmented then selected
  - Fragmented then enzymatically purified
- Uses:
  - Annotation of transcriptional start sites
  - Annotation of 3’ UTRs
  - Quantification and gene expression
  - Depth required 3-8 mill reads
  - Low quality RNA samples
  - Single cell RNA sequencing
RNA-Seq libraries III: Small RNA libraries

- **Source:** size selected RNA
- **Uses:** miRNA, piRNA annotation and quantification
  - Short single end 30-50 bp reads
  - Depth: 5-10 mill reads

Malonne et al. CSHL protocols, 2011
When you need both annotation and quantification

• Attempt three replicates per condition
• Pool libraries to obtain ~15 mill reads per replicate
• Sequence using paired ends
• Analysis:
  – Merge replicate alignments for annotation
  – Split alignments for differential expression analysis
RNA-Seq libraries: Summary

Poly-A selected RNA

Zn-based fragmentation

Exo-CAGE

3'-end-Seq

PolyA selection

Full-length RNA-Seq

Library Construction
ChIP-Seq libraries:

- Crosslinked, immunoprecipitated DNA
- DNA → sequence
- Uses:
  - Mapping nucleosomes (huge depth required)
  - Mapping histones with specific tails
  - Mapping transcription factor sites
  - Requires ~ 5-10 mill, at least 2-3 replicates, single end
ChIP-Seq protocol

Purify DNA

End repair and adapter ligation

Cluster generation

Sequencing on NGS platforms

Kidder et al. Nature Immunology, 2011
CLIP-Seq libraries and ribosome footprinting:

- Crosslinked, immunoprecipitated RNA
- RNA → cDNA → sequence
- Uses:
  - Mapping RNA/protein interactions
  - Find miRNA regulated transcripts
  - Mapping translation rates
  - Annotate ORFs
Analysis of counting data requires 3 broad tasks

• Read mapping (alignment): Placing short reads in the genome

• Quantification:
  • Assigning scores to genes/transcripts
  • Determining whether a gene is expressed
  • Normalization
  • Finding genes/transcripts that are differentially represented between two or more samples.

• Reconstruction: Finding the regions that originated the reads
Once sequenced the problem becomes computational
Analysis of counting data requires 3 broad tasks

• Read mapping (alignment): Placing short reads in the genome

• Quantification:
  • Assigning scores to genes/transcripts
  • Determining whether a gene is expressed
  • Normalization
  • Finding genes/transcripts that are differentially represented between two or more samples.

• Reconstruction: Finding the regions that originated the reads
The sequencing era alignment problem

• Finding 100,000s of small (30-500 bp) sequence in a 10 - 10000 million bp genome.
• Sequences are error prone (~1% error rate)
• Reference and sequence may not be the same haplotype
• **Many techniques are great at finding perfect matches**
Short read alignment strategies

Breaks reads into “seeds” that can be perfectly matched

• Create an easily searchable genome (*index*)
  – Hash table: address map of small words (**k-mers**)
  – Suffix Arrays: Efficient way to look up words
  – FA indices (i.e. Burrows Wheelers)

• Seed search using the index:
  – Matching of smaller portions (seeds) of the read
  – Grouping and prioritizing seeds

• Extending seed alignments
  – Use algorithms that handle mismatches and gaps
### Spaced seed alignment – Hashing the genome

**G:**

```
acccattgaactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg.....
```

#### Store spaced seed positions

<table>
<thead>
<tr>
<th>Seed</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>accg attg **** ****</td>
<td>0</td>
</tr>
<tr>
<td>accg **** actg ****</td>
<td>0</td>
</tr>
<tr>
<td>accg **** **** aatg</td>
<td>0,45</td>
</tr>
<tr>
<td>**** attg actg ****</td>
<td>0</td>
</tr>
<tr>
<td>**** attg **** aatg</td>
<td>0</td>
</tr>
<tr>
<td>**** **** actg aatg</td>
<td>0</td>
</tr>
<tr>
<td>ccga ttga **** ****</td>
<td>1</td>
</tr>
<tr>
<td>ccga **** ctga ****</td>
<td>1</td>
</tr>
<tr>
<td>ccga **** **** atgg</td>
<td>1</td>
</tr>
<tr>
<td>**** ttga ctga ****</td>
<td>1</td>
</tr>
<tr>
<td>**** ttga **** atgg</td>
<td>1</td>
</tr>
<tr>
<td>**** **** ctga atgg</td>
<td>1</td>
</tr>
</tbody>
</table>
Spaced seed alignment – Mapping reads

G: accgattgacctgaatggccttaagggtcctagttgcaaacatgctgaccgtgggattgaatg

q: accq atag accg aatg

2 mismatches

Report position 0

But, how confidence are we in the placement?

$q_{MS} = -10 \log_{10} P(\text{read is wrongly mapped})$
Mapping quality

What does \( q_{MS} = -10 \log_{10} P(\text{read is wrongly mapped}) \) mean?

Let's compute the probability the read originated at genome position i

\( q \): accg atag accg aatg

\( q_s \): 30 40 25 30 30 20 10 20 40 30 20 30 40 40 30 25

\( q_s[k] = -10 \log_{10} P(\text{sequencing error at base } k), \) the PHRED score. Equivalently:

\[
P(\text{sequencing error at base } k) = 10^{-\frac{q_s[k]}{10}}
\]

So the probability that a read originates from a given genome position i is:

\[
P(q \mid G,i) = \prod_{j \text{ match}} P(q_j \text{ good call}) \prod_{j \text{ mismatch}} P(q_j \text{ bad call}) \approx \prod_{j \text{ mismatch}} P(q_j \text{ bad call})
\]

In our example

\[
P(q \mid G,0) = \left[ (1 - 10^{-3})^6 (1 - 10^{-4})^4 (1 - 10^{-2.5})^2 (1 - 10^{-2})^2 \right] 10^{-1} 10^{-2} = [0.97] * [0.001] \approx 0.001
\]
Mapping quality

What we want to estimate is $q_{MS} = -10 \log_{10} P(\text{read is wrongly mapped})$

That is, the posterior probability, the probability that the region starting at $i$ was sequenced \textit{given} that we observed the read $q$:

$$P(i | q) = \frac{P(q | i)P(i)}{P(q)} = \frac{P(q | i)P(i)}{\sum_j P(q | j)}$$

Fortunately, there are efficient ways to approximate this probability (see Li, H \textit{genome Research} 2008, for example)

$$q_{MS} = -10 \log_{10} (1 - P(i | q))$$
Considerations

• Trade-off between sensitivity, speed and memory
  – Smaller seeds allow for greater mismatches at the cost of more tries
  – Smaller seeds result in a smaller tables (table size is at most $4^k$), larger seeds increase speed (less tries, but more seeds)
Considerations

- BWT-based algorithms rely on perfect matches for speed.
- When dealing with mismatches, algorithms “backtrack” when the alignment extension fails.
- Backtracking is expensive.
- As read length increases novel algorithms are required.
# Short read mapping software for ChIP-Seq

<table>
<thead>
<tr>
<th>Short read mapping software</th>
<th>Seed-extend</th>
<th>BWT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short indels</strong></td>
<td>Use base qual</td>
<td>Use Base qual</td>
</tr>
<tr>
<td>Maq</td>
<td>No</td>
<td>YES</td>
</tr>
<tr>
<td>RMAP</td>
<td>Yes</td>
<td>YES</td>
</tr>
<tr>
<td>SeqMap</td>
<td>Yes</td>
<td>NO</td>
</tr>
<tr>
<td>SHRiMP</td>
<td>Yes</td>
<td>NO</td>
</tr>
<tr>
<td>BWA</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>Bowtie</td>
<td>NO</td>
<td></td>
</tr>
<tr>
<td>Stampy*</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>Bowtie2*</td>
<td>(NO)</td>
<td></td>
</tr>
</tbody>
</table>

*Stampy is a hybrid approach which first uses BWA to map reads then uses seed-extend only to reads not mapped by BWA
*Bowtie2 breaks reads into smaller pieces and maps these “seeds” using a BWT genome.
The RNA-Seq alignment problem

Challenges:

• Genes exist at many different expression levels, spanning several orders of magnitude.
• Reads originate from both mature mRNA (exons) and immature mRNA (introns) and it can be problematic to distinguish between them.
• Reads are short and genes can have many isoforms making it challenging to determine which isoform produced each read.
Mapping RNA-Seq reads: Exon-first spliced alignment (e.g. TopHat2)
Mapping RNA-Seq reads: Maximal Mapping Prefix (STAR)
RNA-Seq specific problems

Pseudo gene attraction problem

Intron invasion

Current aligners deal directly with these problems
### Seed-extend

<table>
<thead>
<tr>
<th></th>
<th>Short indels</th>
<th>Use base qual</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAR</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>QPALMA</td>
<td>Yes</td>
<td>NO</td>
</tr>
<tr>
<td>BLAT</td>
<td>Yes</td>
<td>NO</td>
</tr>
</tbody>
</table>

### Exon-first

<table>
<thead>
<tr>
<th>Use base qual</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopHat2</td>
</tr>
<tr>
<td>NO</td>
</tr>
</tbody>
</table>

Exon-first alignments will map contiguous first at the expense of spliced hits
IGV: Integrative Genomics Viewer

A desktop application for the visualization and interactive exploration of genomic data
Visualizing read alignments with IGV — RNASEq

Strand specific library!

Gap between reads spanning exons
Visualizing read alignments with IGV — zooming out
Analysis of counting data requires 3 broad tasks

• Read mapping (alignment): Placing short reads in the genome

• Quantification:
  • Assigning scores to genes/transcripts
  • Determining whether a gene is expressed
  • Normalization
  • Finding genes/transcripts that are differentially represented between two or more samples.

• Reconstruction: Finding the regions that originated the reads
What does significance mean?

• RNA-Seq: The gene is expressed
• ChIP-Seq: Factor binds the region
• CLIP-Seq: Protein binds RNA region
• Ribosomal footprinting:
  – Transcript is translated
  – Ribosomes stalling at region
How do we find peaks?

Scripture is a method to solve this general question.
We have an efficient way to compute read count p-values...
We need to correct for multiple hypothesis testing

The genome is large, many things happen by chance

Expected ~150,000,000 bases
Bonferroni correction is way too conservative

Bonferroni corrects the number of hits but misses many true hits because it's too conservative – How do we get more power?
Given a region of size $w$ and an observed read count $n$. What is the probability that one or more of the $3 \times 10^9$ regions of size $w$ has read count $\geq n$ under the null distribution?

We could go back to our permutations and compute an FWER: \textbf{max of the genome-wide distributions of same sized region)} → but really really really slow!!!
Scan distribution, an old problem

- Is the observed number of read counts over our region of interest high?
- Given a set of Geiger counts across a region find clusters of high radioactivity
- Are there time intervals where assembly line errors are high?

Scan distribution

\[ \alpha = 0.05 \quad \alpha_{FWER} = 0.05 \]

Thankfully, the Scan Distribution computes a closed form for this distribution.

ACCOUNTS for dependency of overlapping windows thus more powerful!
By utilizing the dependency of overlapping windows we have greater power, while still controlling the same genome-wide false positive rate.
Segmentation method for contiguous regions

Example: PolII ChIP

Significant windows using the FWER corrected p-value

Merge

Trim

But, which window?
We use multiple windows

- Small windows detect small punctuate regions.
- Longer windows can detect regions of moderate enrichment over long spans.
- In practice we scan different windows, finding significant ones in each scan.
- In practice, it helps to use some prior information in picking the windows although globally it might be ok.
Applying Scripture to a variety of ChIP-Seq data

200, 500 & 1000 bp windows

100 bp windows
Can we identify enriched regions across different libraries?

Using chromatin signatures we discovered hundreds of putative genes. **What is their structure?**

Discontinuous data: RNA-Seq to find gene structures for this gene-like regions
Analysis of counting data requires 3 broad tasks

• Read mapping (alignment): Placing short reads in the genome

• Quantification:
  • Assigning scores to genes/transcripts
  • Determining whether a gene is expressed
  • Normalization
  • Finding genes/transcripts that are differentially represented between two or more samples.

• Reconstruction: Finding the regions that originated the reads
RNA-Seq quantification

- Is a given gene (or isoform) expressed?
- Is expression gene A > gene B?
- Is expression of gene A isoform $a_1$ > gene A isoform $a_2$?
- Given two samples is
  - expression of gene A in sample 1 different from gene A in sample 2?
  - Is the expression of one isoform changing?
RNA-Seq measures relative abundance

RNA-Sequences: Infer fraction of molecules in sample
RNA-Seq quantification units

\[ RPKM = 10^9 \frac{\# reads}{\text{length} \times \text{TotalReads}} \]

- Fragmentation of transcripts results in length bias: longer transcripts have higher counts
- Different experiments have different yields. Normalization is key for cross lane comparisons

Garber et al. Nat. Methods 2011
RNA-Seq quantification “units”

- To compare within a sequence run (lane), RPKM accounts for length bias.

- RPKM (Mortazavi et al 2008) is not optimal for cross experiment comparisons.
  - Different samples may have different compositions.

- FPKM (Trapnell et al. 2011) superseded RPKM to deal with paired end data
  - Paired end reads originate from the same Fragment

- And later $\text{TPM} = 10^6 \times \text{Fraction of transcript in sample}$ (Li et al 2009)
  - More robust to changes in sample composition

Complexity increases when multiple isoforms exist
But, how to compute counts for complex gene structures?

Three popular options:

Exon *intersection* model: Score constituent exons

Exon *union* model: Score the “merged” transcript

Transcript expression model: Assign reads uniquely to different isoforms. *Not a trivial problem!*
Read assignment involves probabilistic assignment.

Figure 4. Illustration of the EM algorithm. The gene has three isoforms (red, green, blue) of the same length. There are five reads (a, b, c, d, e) mapping to the gene. One maps to all three isoforms, one only to red, and the other three to each of the three pairs of isoforms. Initially, every isoform is assigned the same abundance (0.33, 0.33, 0.33). During the expectation (E) step, reads are proportionately assigned to transcripts according to the isoform abundances. Next, during the maximization (M) step, isoform abundances are recalculated from the proportionately assigned read counts. For example, the abundance of red after the first M step is estimated by $0.47 = (0.33 \times 1 + 0.55 \times 1 + 0.23 \times 1) / (0.33 + 0.33 + 0.33)$.

Current quantification models are complex

• In its simplest form we assume that reads can be unequivocally mapped. This allows:
  – Read counts distribute multinomial with rate estimated from the observed counts

• When this assumption breaks, multinomial is no longer appropriate.

• In general models use:
  – Fragments as inferred from paired-end data
  – Base quality scores
  – Sequence mapability
  – Protocol biases (e.g. 3’ bias)

• Handling each of these involves a more complex model where reads are assigned probabilistically not only to an isoform but to a different loci
Why paired end matters for isoform quantification?

How do we define the gene expression?
How do we compute the expression of each isoform?
Computing gene expression

Idea 1: RPKM of the constitutive reads (Neuma, Alexa-Seq, Scripture)
Computing gene expression — isoform deconvolution
Computing gene expression — isoform deconvolution

If we knew the origin of the reads we could compute each isoform’s expression. The gene’s expression would be the sum of the expression of all its isoforms.

\[ E = \text{RPKM}_1 + \text{RPKM}_2 + \text{RPKM}_3 \]
Paired-end reads are easier to associate to isoforms

Paired ends increase isoform deconvolution confidence
- $P_1$ originates from isoform 1 or 2 but not 3.
- $P_2$ and $P_3$ originate from isoform 1

Do paired-end reads also help identifying reads originating in isoform 3?
We can estimate the insert size distribution

Splice and compute insert distance

Estimate insert size empirical distribution

Get all single isoform reconstructions
… and use it for probabilistic read assignment

For methods such as MISO, Cufflinks and RSEM, it is critical to have paired-end data
Other considerations

• Duplicates – What to do with PCR artifacts
• Multimapper reads – What to do with reads that map to multiple places in the genome
RNA-Seq quantification summary

• Counts must be estimated from ambiguous read/transcript assignment.
  – Using simplified gene models (intersection)
  – Probabilistic read assignment

• Counts must be normalized
  – RPKM/FPKM/TPM are designed for intra-library comparisons:
    • Is gene A more highly expressed than gene B

• How do we normalize More sophisticated normalization to account for differences in library composition for inter-library comparisons.
### Programs to measure transcript expression

<table>
<thead>
<tr>
<th>Implemented method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cufflinks2</td>
<td>Transcript deconvolution by solving the maximum likelihood problem</td>
</tr>
<tr>
<td>RSEM</td>
<td>Transcript deconvolution by solving the maximum likelihood problem</td>
</tr>
<tr>
<td>eXpress</td>
<td>Incorporated biases into model</td>
</tr>
</tbody>
</table>
Analysis of counting data requires 3 broad tasks

- Read mapping (alignment): Placing short reads in the genome

- Quantification:
  - Assigning scores to genes/transcripts
  - Determining whether a gene is expressed
  - **Normalization**
  - Finding genes/transcripts that are differentially represented between two or more samples.

- Reconstruction: Finding the regions that originated the reads
Sample composition impacts transcript *relative* abundance

Normalizing by total reads does not work well for samples with very different RNA composition.
Example normalization techniques

\[ s_j = \text{median}_{i} \left( \frac{k_{ij}}{\left( \prod_{v=1}^{m} k_{iv} \right)^{1/m}} \right). \]

Counts for gene \( i \) in experiment \( j \)

Geometric mean for that gene over ALL experiments

\( i \) runs through all \( n \) genes
\( j \) through all \( m \) samples
\( k_{ij} \) is the observed counts for gene \( i \) in sample \( j \)
\( s_j \) is the normalization constant

Alders and Huber, 2010
Let's do an experiment (and do a short R practice)

```r
> s1 = c(100, 200, 300, 400, 10)
> s2 = c(50, 100, 150, 200, 500)
> norm = sum(s2)/sum(s1)
> plot(s2, s1*norm, log="xy")
> abline(a = 0, b = 1)

> g = sqrt(s1 * s2)
> s1n = s1/median(s1/g); s2n = s2/median(s2/g)
> plot(s2n, s1n, log="xy")
> abline(a = 0, b = 1)
```

Similar read number, one transcript many fold changed

Size normalization results in 2-fold changes in all transcripts
When everything changes: Spike-ins

A) Limited transcriptional response

B) mRNA levels per cell

-normalized signal

log-fold change

Observed fold-change

Perceived response

Counts

Genes

A B C D E F G H I

A B C D E F G H I

A B C D E F G H I

A B C D E F G H I

C) Transcriptional amplification

D) mRNA levels per cell

-normalized signal

log-fold change

Observed fold-change

Perceived response

Counts

Genes

A B C D E F G H I

A B C D E F G H I

A B C D E F G H I

A B C D E F G H I

Figure 1. Normalization and Interpretation of Expression Data

(A) Schematic representation of microarray normalization when the overall levels of mRNA per cell are not changing in two conditions. Relative mRNA levels for a set of genes (B) are indicated along the y axis for condition 1 (black) and condition 2 (orange). The panels, from left to right, depict the actual relationship between mRNA levels for the two conditions; the effect of median normalization; the calculated fold changes based on median normalization, assuming that the absolute amount of total mRNA in the other sample is the same or average value or that the distributions of samples from different experiments should be scaled to have the same median or average value.

(B) Heatmap showing the results of different normalization methods on the interpretation of microarray data. The data represent fold change of expression in high-Myc versus low-Myc cells. Each line represents data for individual probes on the microarray. Red indicates increased expression in high-Myc versus low-Myc cells. Black indicates no change in expression. Left: data using a standard normalization method (MAS5). Right: the same data, now renormalized by using spike-in standards.

(C) Heatmap showing the results of different normalization methods on the interpretation of RNA-sequencing data. The data represent fold change of expression above the midline; and the perceived transcriptional response following transcriptional amplification of gene expression normalized with spike-in RNAs. Where most genes are expressed at higher levels. The square box represents a perturbation such as increased expression of a gene regulator or a change in environment or cell state. Red arrows point to target genes affected by the perturbation, which are represented as circles. Red shading of circles indicates relative transcriptional increase.

(D) Heatmap showing the results of different sample preparation methods on the interpretation of digital quantification data. The data represent fold change of reads per kilobase of exon model per million mapped reads. Left: data using a standard sequencing normalization indication that the expression levels of some genes are unchanged, whereas others increase or decrease. Similar amounts of RNA from the low- and high-Myc cell RNA populations (Digital Bio) and equivalent numbers of high- and low-Myc cells from equal numbers of high-Myc and low-Myc cells were introduced into the Affymetrix DNA microarray assay (available online). Similar amounts of RNA from the low- and high-Myc cell populations were used as standards for normalization. mRNA levels are indicated along the y axis for condition 1 (black) and condition 2 (orange); individual genes are represented by orange triangles (S1–S3). The panels, from left to right, depict the actual relationship between mRNA levels for the two conditions; the effect of median normalization; the calculated fold changes based on median normalization, assuming that the absolute amount of total mRNA in the other sample is the same or average value or that the distributions of samples from different experiments should be scaled to have the same median or average value.

Lin et al., 2012

Nie et al., 2012

477

2012 Elsevier Inc.

Li and Wong, 2001

Irizarry et al., 2003

Reimers, 2010

Mortazavi et al., 2008

Wu et al., 2004

Schuhmacher et al., 1999

Benes and Muckelroy, 2012

Jiang et al., 2012

Lovén et al, Cell 2012
Analysis of counting data requires 3 broad tasks

- **Read mapping (alignment):** Placing short reads in the genome

- **Quantification:**
  - Assigning scores to genes/transcripts
  - Determining whether a gene is expressed
  - Normalization
  - **Finding genes/transcripts that are differentially represented between two or more samples.**

- **Reconstruction:** Finding the regions that originated the reads
Differential Gene Expression Questions

- Finding genes that have different expression between two or more conditions.
- Find gene with isoforms expressed at different levels between two or more conditions.
  - Find differentially used slicing events
  - Find alternatively used transcription start sites
  - Find alternatively used 3’ UTRs
General strategy for differential gene expression

• Normalize count data
  • Key: We only compare each gene across samples NOT one gene to another.

• Estimate normalized mean gene counts

• Estimate gene variance
  – Assume variance is similar for similarly expressed transcripts
  – Model variance as a function of expression
  – Use model to estimate variance for a transcript given its mean count

• Define a test
  – DESeq: Generalization of a fisher exact test
  – Cufflinks: Log transformed of counts divided by its variance (~ normally distribute).
    • Null hypothesis: log ratio = 0
Differential analysis strategies

- Use read counts and Standard Fisher exact test

<table>
<thead>
<tr>
<th></th>
<th>Condition A</th>
<th>Condition B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A reads</td>
<td>$n_a$</td>
<td>$n_b$</td>
</tr>
<tr>
<td>Rest of reads</td>
<td>$N_a$</td>
<td>$N_b$</td>
</tr>
</tbody>
</table>

- Not naturally extendable to experiments with replicates
Why not just simple models?

<table>
<thead>
<tr>
<th>Condition A</th>
<th>Condition B</th>
<th>Log fold-change (union count)</th>
<th>Log fold-change (intersect count)</th>
<th>Log fold-change (true expression)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\log_2 \left( \frac{10}{10} \right) = 0$</td>
<td>$\log_2 \left( \frac{8}{7} \right) = 0.19$</td>
<td>$\log_2 \left( \frac{10}{6/L + \frac{4}{2L}} \right) = 0.32$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\log_2 \left( \frac{6}{8} \right) = -0.41$</td>
<td>$\log_2 \left( \frac{5}{5} \right) = 0$</td>
<td>$\log_2 \left( \frac{6/L}{8/2L} \right) = 0.58$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\log_2 \left( \frac{5}{10} \right) = -1$</td>
<td>$\log \left( \frac{4}{5} \right) = -0.1$</td>
<td>$\log_2 \left( \frac{5}{10/2L} \right) = 0$</td>
</tr>
</tbody>
</table>
## RNA-Seq differential expression software

<table>
<thead>
<tr>
<th>Software</th>
<th>Underlying model</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>EdgeR</td>
<td>Negative Bionomial</td>
<td>Gene read counts table</td>
</tr>
<tr>
<td>DESeq2</td>
<td>Negative Bionomial</td>
<td>Gene read counts table</td>
</tr>
<tr>
<td>Cufflinks2</td>
<td>Poisson Negative Bionomial</td>
<td>Works directly from the alignments</td>
</tr>
<tr>
<td>Myrna</td>
<td>Empirical</td>
<td>Sequence reads and reference transcriptome</td>
</tr>
<tr>
<td>Miso</td>
<td>Multinomial</td>
<td>Specifically to test exon cassette inclusion/exclusion.</td>
</tr>
</tbody>
</table>
MISO: Specifically testing exon inclusion

Fragmentation and amplification
mRNA fragments sampled in proportion to isoform length and expression level $\psi_k$

Sequencing
Short reads generated from fragments

Alignment
Reads aligned to genome and splice junctions

Incorporation of paired ends
Assign reads to isoforms using insert length distribution

--- Introns
☐ Skipped exon
☐ Constitutive exons

Bayesian inference

Single-end estimate, $\hat{\psi}_{SJ}$

Insert variability:
$\sigma = d \sqrt{\Gamma}$

Insert length (nt)

Paired-end estimate, $\hat{\psi}_{MISO}$

Inclusion reads
Constitutive reads
Exclusion reads

Katz et al. Nat. Methods 2010
Our typical pipeline (e.g. RNA-Seq)

1. Upload your sequence data (fastq)
2. Align to the ribosome (Bowtie)
3. Align remaining reads to genome (TopHat)
4. Quantify transcriptome
5. Call differentially expressed genes (if multiple samples)
6. Make report of quality metrics
7. Output ribosomal contamination metrics report
8. Produce RNA-Seq report % aligned, % intergenic, % exonic, % UTR
9. Produce IGV/UCSC friendly files
10. Produce a table with normalized expression values
11. Report pairwise significant genes that are differentially expressed
The quest for inexpensive expression assays

- **Goal:** Routinely profile hundreds of samples
- **Why?**
  - Human variability in health and disease
  - Perturbation studies
  - Clinical applications of expression profiling
  - Single cell sequencing
- **Current costs**
  - Affy ~$300-$400/sample
  - Illumina bead arrays $150/sample
  - RNA-Seq (20 mill reads) ~$400-$500/sample ($350 in sequencing)
- **RNA-Seq disadvantages**
  - Complex analysis
  - Length bias
Reading molecules: end-sequencing and molecular barcodes

Poly-A selected RNA

fragmentation

Exo-CAGE

Full-length RNA-Seq

Library Construction

3’-end-Seq

PolyA selection

Maxim Artyomov
Molecule counting – Unique Molecular Identifiers (UMI)

Figure S2. Experimental procedure. Schematic diagram presenting the process of converting single-cell RNA samples to sequencing-ready DNA libraries. Shown are ten experimental steps describing how RNA is tagged, pooled, amplified, fragmented, and how library construction is being performed. Colored lines represent RNA (blue) or DNA (black) molecules, or oligos and primers (see methods for a detailed description).

Step 1: Reverse transcription

Step 2: Exonuclease I

Step 3: Sample pooling

Step 4: Second strand synthesis

Step 5: In Vitro Transcription

Step 6: DNaseI

Legend:
- RNA
- cDNA
- 2nd strand

Step 8: RNA/ssDNA ligation

Step 9: Reverse transcription

* Optional addition of pool barcode

Step 10: Amplification + Illumina primers addition by nested PCR

Legend:
- NT<sub>20</sub>-UMI-barcode
- partial rd2<sup>-1</sup>-T7 promoter
- Un-UMI-barcode
- partial rd2<sup>-rev</sup>-T7 promoter
- Un-UMI-barcode
- partial rd2<sup>-rev</sup>
- P5<sub>rd1</sub> forward primer
- P7<sub>rd2</sub> reverse primer

Library ready for Illumina sequencing

NT<sub>20</sub>XXXXXXXX-SSSSSS-adapter

XXXXXXXXXXXX: UMI

SSSSSSS: Sample Barcode

Jaitin et al. Science 2014
End-sequencing solution

---

**Exo-CAGE**

**Full length RNA-Seq**

**3' RNA-Seq**

---

**Il23a**

Normalized gene length (to 100 bins)

Normalized coverage (X)

- 100
- 80
- 60
- 40
- 20
- 0

1
2
3
4

196
228
216
228
216
228
241
258
228
258
228
228
216

Exo-CAGE

3' RNA-Seq

95% confidence interval
Although annotated ends far from perfect
While annotated starts are much more conserved

- **Tlr1**
  - 2hrR1
  - 2hrR2
  - 4hrR1
  - 4hrR2
  - 920bp longer

- **Nod1**
  - 2hrR1
  - 2hrR2
  - 4hrR1
  - 4hrR2
  - 140 bp shorter

Graph showing:
- 374 genes
- 46 genes

Distance from annotated 5' end

# Genes
We take full advantage of the data

1. Slide a window and identify major 3’ end
2. Identify all other significant windows (using a local background)
3. Repeat for each sample
4. Take all significant windows across samples
5.1 Report gene level counts: Sum across all sig. windows
5.2 Report isoform level counts: Each sig. window
Reproducibility is as good as with full length
With 8.5 Million reads similar yet somewhat reduced power

Having established a robust analysis pipeline => Single cell RNA-Seq
Why Single-cell analysis?

qPCR analysis of \textit{CXCR5} vs \textit{CCL5} expression in ‘bulk’ 100-cell T cell populations and single T cells:

Type 1 Diabetes study

- It is unclear what triggers T1D
- The mechanism(s) of β-cell death are not well understood.
- Rat model with inducible T1D within 10 (± 1) days.
- Bulk RNA-Seq can’t reveal tissue composition
Cell sorting

- Pancreatic islets are composed of:
  - $\alpha$-cells: primarily produce glucagon
  - $\beta$-cells: primarily produce insulin
  - $\delta$-cells, PPY producing cells, and others

- Issues with sorting cells by FACS:
  - Only known cell types can be selected
  - Preprocessing may affect the observed cell state
  - Islet cells are very difficult to isolate, and FACS discards “other” cells in the sorting process (wasteful for rare cells)

- In addition, “bulk” RNA-Seq can mask underlying heterogeneity of even a sorted cell population…
Islet single cell sequencing

Dissociate Islets
Flow Sort Single Cells

SCRB-Seq Library Construction

<table>
<thead>
<tr>
<th>No extension</th>
<th>5kb extension</th>
<th>% increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>8574</td>
<td>9648</td>
<td>12.5%</td>
</tr>
</tbody>
</table>

After filtering:

<table>
<thead>
<tr>
<th>cells with &gt;200 total UMIs:</th>
<th>263</th>
<th>283</th>
<th>7.6%</th>
</tr>
</thead>
<tbody>
<tr>
<td>genes with &gt;50 total UMIs:</td>
<td>296</td>
<td>367</td>
<td>24.0%</td>
</tr>
</tbody>
</table>
Single cell RNA-Seq cell sorting

PC3 (9% of variance)

PC4 (<5% of variance)

PC3 (9% of variance) vs PC4 (<5% of variance)

Ins1

Ppy

Gcg

Ins2

Sst

lapp
Which allow us to recover the known islet composition
More in depth exploration of depth

- Very deep (30 million reads) dataset with triplicates.
  - Mouse WT vs double Jnk1/2 KO (Roger Davis)
  - Worm diet changes (Marian Walhout)
- Call DE with full dataset, then in-silico downsample data
Is the loss qualitatively significant?

15 Million reads
Is the loss qualitatively significant?

12.5 Million reads
Is the loss qualitatively significant?

10 Million reads
Is the loss qualitatively significant?

7.5 Million reads
Is the loss qualitatively significant?

5 Million reads
Is the loss qualitatively significant?

2.5 Million reads
Is the loss qualitatively significant?

1 Million reads
The loss is qualitatively small
The loss is qualitatively small
Final considerations: The steps of Sequencing analysis

- Filter reads (fastq file) by removing adapter, splitting barcodes.
  - Evaluate overall quality, look for drop in quality at ends. Trim reads if ends are of low quality
- Alignment to the genome
  - Use transcriptome if available
  - Filter out likely PCR duplicates (reads that align to the same place in the genome)
  - Evaluate ribosomal contamination
  - What percent of reads aligned
- Reconstruct (?)
- Quantify
  - Normalize according to application
Thanks

Sebastian Kadener
(HUJ)

Maxim Artyomov
(WashU)

Diabetes Center
David Blodget
Chaoxing Yang
Rita Bortel
Dale Greiner
David Harlan

Broad Technology Labs
Tarjei Mikelsen
Magali Soumillon

Garber lab & Bioinformatics Core

Alper Kucukural
Sabah Kadri
(Broad)

Jenny Chen
(MIT)

Alan Derr

http://garberlab.umassmed.edu/