Analyzing the Whole Transcriptome by RNA-Seq data
Statistical and Computational Challenges

Claudia Angelini
Istituto per le Applicazioni del Calcolo

c.angelini@iac.cnr.it
Our Research Activities:

Statistical and Computational methods for

✔ Microarray data analysis (since 2007)
✔ NGS data analysis (since 2010)
Microarray data analysis at IAC

http://www.na.iac.cnr.it/bats/

Bayesian model

\[ Z_{i,k} = S_{i}(t_j) + \varepsilon_{i,j,k} \]

“One (statistical) sample”

One Class

Two Class (Paired)

Two Class (Unpaired)

“Two (statistical) samples”

Gene Expression Control

transcription factor

gene expression

upstream DNA sequence

gene
“Next-generation” sequencing is “Today”

The Human whole genome project started in 1995 and finished in 2003. It required the collaboration of hundreds of researchers from 20 institutions in 6 countries. The estimated cost is about 270,000,000$. It took about 8 years to arrive at the first draft of the human genome.

“Today” generation sequencers such as 454, Illumina, SOLiD can sequence a whole human genome in less than one week at a cost of less than 50,000$. Costs are decreasing at each new release, speed, accuracy and resolution improves, but the amount of data explodes!!

NGS platforms are capable of producing hundreds of millions of DNA/RNA short sequences in a single run.

In January 2008 Nature Methods declared that:

2007 was “The year of sequencing”
The main idea of the NGS sequencing consists of producing hundreds of millions of “short reads” or sequences with their quality values (although each platform uses different chemistry and experimental protocols and, hence, has different features)
What would you do if you could sequencing everything?

We are mainly focusing our attention in **RNA-Seq** and in **ChIP-Seq** studies (and possibly also in their connection).

It is true for all NGS applications

“**RNA-Seq** [...] is expected to revolutionize the manner in which eukaryotic transcriptomes are analysed.”

Wang et al, Nature Genetics 2009

---

Figure and title from Nature Biotechnology 2008
Biological Background (1)

Chromosome

“Gene”

DNA (positive strand)

Gene

mRNA

Alternative Splicing

mRNA

Protein A

Protein B

Exon 1

Exon 2

Exon 3

Exon 4

Introns

Exons

Stop codon

UTR

DNA (positive strand)
Different isoforms can produce different proteins:

Note that the “Central Dogma” is violated.

Current annotations is contained in databases such as RefSeq, UCSC, Ensembl,…

Note that there can be ambiguities due to gene’s overlap, uncorrect or partial annotations.
Transcriptome Analysis: Key Research Questions

Using quantitative genome-wide experiments we want to:

Characterize an organism’s full complement of genes
- Find new (possibly also non-coding) genes
- Compare genes among organisms

Characterize transcript isoforms
- Find new alternative splice forms or transcript boundaries

Monitor transcriptome changes between tissues or in response to environmental changes (e.g. stress)
- Identify significant expression changes between conditions

The aim is try to understand:
- Differences of active components between conditions/organisms
- What changes when perturbing the biological system
- How the active components are regulated
Hybridization-based technologies:
• Background and cross-hybridization issues
• Only transcripts included in the array design
• Specific studies requires specific array types
• Limited dynamic range
• Nowadays much easier to analyze (several software available)
• Low computational complexity
• Nowadays much cheaper and easy to handle → “large” sample production

RNA-Seq:
• Low “background signal”;
• Identification of novel transcribed regions;
• Identification, characterization and quantification of new splice isoforms;
• Determination of correct gene annotations 5'/3'-UTRs;
• No upper limit for gene quantification
• Much more computationally demanding
• “Limited” amount of software available
• Still expensive either for sample preparation and sequencing → “limited” sample production

How to obtain high-throughput quantitative measures of the transcriptome?
RNA-Seq experiments

**Wet Lab**

- RNA isolation
  - rRNA depletion
  - Poly(A) RNA

- RNA fragmentation
- Ligation to adaptors
- Retro-transcription
- Size selection and PCR amplification
- Emulsion PCR
- Beads enrichment and deposition

**Sequencing**

**Transcripts**

**Short Reads**

The output of a RNA-Seq experiment consists of **hundred millions of short reads**

**DATA ANALYSIS**

- File_reads
- File_quality
Computational pipeline

- Data analysis is carried out using pipelines that involve several computational steps and combine several software packages.
- The amount of data involved in each experiment is of the order of several GB.
- Computational processes require High Performance Computing and specific resources.
- Many software packages are continuously coming out at a speed that is difficult to trace of the changes.
- Most of the time is spent in installing and testing them or in formatting input/output and inspecting the differences.
To align or to assembly? That is only the first question

- Alignment can be used if a reference genome is available, otherwise one need to de-novo assembly the “reads”
- Reference genomes are not the sequenced genome, difference can be due introduced due to individual variability (i.e., SNPs, Indels, rearrangement)
- Genomes are usually highly repetitive
- Observed “reads” are short (50-100bp) and can be affected by noise
- Genomes of eukaryotes are huge (human about 3*10^9 bp)

It represents a compromise between fidelity of the alignment and amount of data aligned
Alignment to a reference genome (1)

- Owing to the large number of produced short reads, the use of conventional alignment algorithms is not feasible → A new generation of aligners has been developed, and more are expected soon.
- It represents the most computational demanding step. While several algorithms exist for alignment, assembly is still an open area of research especially for complex organisms.

- Data Input/output format and compatibility is often a serius issue

- The type of DNA library should be taken into account (Fragment vs. Paired-end)
- Every aligner is a balance between accuracy, speed, memory and flexibility, and no aligner can be best suited for all applications

- Not all aligners are suited for RNA-seq mapping

RNA-seq alignment is particularly challenging due to the splicing!

Exon-exon splice junction
Alignment to a reference genome (2)

“Reads” are “short” and can be affected by sequencing error, “biological variability” i.e., SNPs, indels, structural variations and rearrangement etc, reference genomes are often incomplete and usually very repetitive.

Three types of results can be distinguished:

→ Uniquely assigned reads
→ Multiple Matching reads
→ Unmatched reads

RNA-seq alignment strategies to a reference genome

Model based junction detection
Data driven junction detection
“De-novo” transcript assembly
What are the “**basic**” biological questions?

1. Identification and quantification of transcriptional regions

- “**Known**” regions (i.e., already annotated regions in classical database such as RefSeq, UCSC, Ensembl…)
- Correct boundary annotations (exon boundary, splicing sites, UTR’s)
- “**Novel**” transcribed regions → gene model

**Within sample/condition analysis**

Compute number of reads that cover base /

Number of Reads aligned per base

**Genome Browser (UCSC)**
Expressed or not expressed? What is the value?

Quantification on known regions

Detection of correct boundaries in the annotation

Identification of novel transcribed regions

Annotation... it can contain ambiguities or be not precise

Searching for a needle in a haystack

Annotation... it can contain ambiguities or be not precise
From “aligned” reads to quantitative measures

- What is the “expression”?
- At what resolution level?

- Gene level
- Isoform level
- Exon/transcribed region...
- Novel transcribed regions

<table>
<thead>
<tr>
<th>Features</th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td>S11</td>
<td>S12</td>
</tr>
<tr>
<td>Feature 2</td>
<td>S21</td>
<td>S22</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature N</td>
<td>SN1</td>
<td>SN2</td>
</tr>
</tbody>
</table>

“Expression” is often measured as $S_i$:

Number of Reads mapped on features i

Or as Reads Per Kilobase of transcript per Million mapped reads (RPKM)

$$\frac{S_i}{N} \frac{10^3}{L_i} 10^6$$

$$\frac{N}{L_i}$$

Total number of mapped reads

Length (bp) of features i
The first step of any “Seq-data analysis” is the alignment or the assembly of the reads. The observations are “the reads”.

All subsequent results and inference are based on the aligned reads. The alignment is an inferential process, uncertainty should be taken into account.

“Reads” do not have the same quality, the mapping can produce uniquely aligned reads, multiple matching and unmatched reads. How to account for?

Poisson and Negative Binomial model do not naturally account for such uncertainty.

Probabilistic model for RNA-Seq (as well as for other “Seq” protocols) are still needed.
"Signal + Noise" functional approach

Segmentation and estimation of Poisson contaminated process

\[ \hat{n}_l = n_l + \varepsilon_l, \quad l = 1, \ldots, L, \]

L = genome length

Number of reads that cover base \( \Rightarrow \) Poisson process with intensity \( \lambda_l \)

Random error (due to the alignment process) \( \Rightarrow \) correlated (approx band-matrix);

We assume: independence among reads, large coverage \( (N>>1) \),

Either almost identical to zero or piece-wise linear (approximately constant)

We set-up a "biologically inspired" simulator aimed to evaluate the performance of any computational approach
What are the “basic” biological questions? (2)

2. Identification and quantification of isoforms.

- Identification and quantification of “known” isoforms
- Detection of “novel” isoforms and their quantification

Many types of splicing events can occur, moreover one gene can be constituted of several “blocks” (i.e., exons), and can be expressed in several isoforms.
Isoforms detection

- "Guilty by evidence" → Presence of multiple isoforms is inferred by "multiple donor / acceptor" junctions.

- Isoform quantification as a "deconvolution" problem

- (Novel) isoform detection and estimation as "local" de-novo assembly
“Guilty by evidence” for alternative splicing

- Almost an empirical approach
- Thresholds have to be set to distinguished “true” junctions from noise (due to the mapping)
- Quality of the mapped junctions or the exon usage have to be considered
- It allows to detect presence of isoforms but not to quantify their abundance
Isoform quantification as a “deconvolution” problem

- Genes may contain up to hundreds of exons, several tens of isoforms.
- Most of the isoforms (in particular the one less abundant or condition-specific) are not annotated → but it is important to discover their presence and estimate their abundance.

Different algorithms but similar “ideas”

- Work on annotated isoforms
- Not always the software is available or the input/output format can be easily handled
- Attention to the type of RNA-seq library that can be handled
**Isoform detection and estimation as “local” de-novo assembly**

Software **CUFFLINKS**

- Graph theory to locally assembly genes/transcripts
- Genes are treated independently
- It requires high coverage experiments
- Several parameters have to be manually tuned

Software **SCRIPTURE**

*Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs

Mitchell Guttman1,2,6, Manuel Garber1,6, Joshua Z Levin1, Julie Donaghey1, James Robinson1, Xian Adiconis1, Lin Fan1, Magdalena J Koziol1,3, Andreas Gnirke1, Chad Nusbaum1, John I. Rinn1,3, Eric S Lander1,2,4 & Aviv Regev1,2,5

These two methods have been voted as the most innovative approaches in 2010 by a selection of 15 top leading scientists
3. Detection of differential expressed events between two or more experimental conditions

- Detection of “differentially expressed” genes
- Detection of sample/condition specific isoforms
- Detection of “differentially expressed” isoforms
- ….

Between samples/conditions analysis (at different resolution level)

Sample specific junction
Significant expression changes are typically identified with a **Statistical Test** and results have to be corrected for **multiple hypotheses testing**

Unfortunately one cannot use ordinary tests developed for microarray straight away because **RNA-Seq data** are **count data**, they are neither homoscedastic nor can the residuals be expected to be normally distributed.

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Condition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>Sample 2</td>
<td>.....</td>
</tr>
<tr>
<td>Feature 1</td>
<td>S11</td>
<td>S12</td>
</tr>
<tr>
<td>Feature 2</td>
<td>S21</td>
<td>S22</td>
</tr>
<tr>
<td>Feature n</td>
<td>Sn1</td>
<td>Sn2</td>
</tr>
<tr>
<td>TOTAL MAPPED READS</td>
<td>N1</td>
<td>N2</td>
</tr>
</tbody>
</table>

- Data normalization is still an open issue
- Due to the costs of the NGS platforms the number of samples is still very small.
- n × p with n >> p
Statistical analysis: Differential Expression (2)

Available in Bioconductor

- edgeR
- DEGseq
- DeSeq
- BaySeq

Differ wrt to the test statistics, parametric model and the way in which they estimate the parameters

Tends to underestimate the variability

\[ \lambda_{i,j,k} = \mu_{i,j} \times N_{i,k} \]

For two conditions we have:

\[ H_0 : \mu_{i,1} = \mu_{i,2} \quad vs \quad H_0 : \mu_{i,1} \neq \mu_{i,2} \]

One can obtain p-values, compute FDR and so on

"extra parameter" to model overdispersion

\[ Y_{i,j,k} \sim \text{Poiss}(\mu_{i,j}; N_{j,k}) \quad i = 1,...,G \quad (# \text{genes or features}) \]
\[ j = 1,...,M \quad (# \text{conditions}) \]
\[ k = 1,...,K \quad (# \text{replicates}) \]

\[ Y_{i,j,k} \sim \text{NB}(\mu_{i,j}; N_{j,k}, \phi_{i,j}) \quad i = 1,...,G \quad (# \text{genes or features}) \]
\[ j = 1,...,M \quad (# \text{conditions}) \]
\[ k = 1,...,K \quad (# \text{replicates}) \]
### Our Pilot study with RNA-Seq

<table>
<thead>
<tr>
<th>Experiment Type</th>
<th>Number of Samples</th>
<th>Number of reads produced (50bp)</th>
<th>Number of reads Aligned</th>
<th>Software</th>
<th>Comments</th>
<th>Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq</td>
<td>2</td>
<td>196,371,166</td>
<td>67,215,796</td>
<td>RNA-MATE, EdgeR, DESeq, DEGseq, GoSeq, Matlab and R custom packages</td>
<td>Human hg19, Down syndrome</td>
<td>rRNA depletion, Fragment Library</td>
</tr>
</tbody>
</table>

The study was aimed to characterize and investigate the difference between Down Syndrome and Euploid samples from EPCs cells.

Data are available at GEO accession numbers: GSE27443

The experiments have been conducted on the SOLiD 3.0 platform and using the core facilities available at the IGB-CNR.
Set-up of the computational pipeline

1. Quality assessment and filters (quality plot, remove low quality reads, ribosomal reads, adapters, etc...) 
2. Alignment to a reference genome (genome+junctions) + “Trim” the right-side of the reads and repeats the step
3. Handle “multiple” reads
4. Quantify known “features” at RefSeq (gene) level→ handle “ambiguities”
5. Identify novel transcribed regions (using customized bisection strategy)
6. Identify alternative splicing isoforms
7. Detect differential expression (i.e., statistical analysis)
Results on our Pilot Study

Results of DE were validated (are under validation) by using qRT-PCR + literature knowledge. Pathway analysis was performed using GOSEQ (and other classical tools).

Detection score

RPKM>0.1 \(\rightarrow\) 14642

“RefSeq genes” were investigated

Additionally, we identified extended UTRs, intergenic and intronic transcription which turned out to be very relevant when using total RNA (which includes non-coding RNA); we also provide evidence of sample specific splicing events involving either known and combinatorial junctions.

\(\rightarrow\) Results are going to appear (hopefully) in PLOSONE
Intronic and intergenic transcription

The size of the human genome is about $3 \times 10^9$ bp (for each strand).

Given the RefSeq annotation we can define strand specific “intronic” and “intergenic” regions (that span approximately $4.9 \times 10^9$ and $1.0 \times 10^9$ bp).

We implemented a fast preliminary searching procedure based on idea the bisection algorithm.

99030 Intronic transcriptionally active regions (inTARs) spanning across about 80 Mb and accounting for more than 55% of mapped “intronic” unannotated reads.

21804 Intergenic transcriptionally active regions (igTARs) spanning across about 17 Mb and accounting for more than 45% of mapped “intergenic” unannotated reads.

Searching for a needle in a haystack

If $N_{\text{reads}} < \text{threshold}$

$\rightarrow$ Discard
Conclusions (1)

Results are very promising, however biological conclusions that can be inferred from the direct comparisons of only two (or very few) samples are limited → RNA-Seq experiments can (optimistically) reduce the technical variability, but they do not affect the biological variability. Several other sources of variability should be better considered (at the experimental design level)

✓ From statistical significance to biological significance.

Extend the analysis to a larger number of samples/conditions to increase the detection power for identifying disease-associated genes/features → Generalized linear model of the Poisson family or generalized linear models tailored for the negative binomial case (or a quasi-Poisson GLM with proper variance estimation) → To account for other sources of variability
What are the “other” biological questions?

4. More sophisticated statistical analysis (clustering, classification, gene networks, etc …)

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>......</th>
<th>......</th>
<th>Sample p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td>S11</td>
<td>S12</td>
<td>......</td>
<td>......</td>
<td>S1p</td>
</tr>
<tr>
<td>Feature 2</td>
<td>S21</td>
<td>S22</td>
<td>......</td>
<td>......</td>
<td>S2p</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>......</td>
<td>......</td>
<td></td>
</tr>
<tr>
<td>Feature n</td>
<td>Sn1</td>
<td>Sn2</td>
<td>......</td>
<td>......</td>
<td>Snp</td>
</tr>
<tr>
<td>TOTAL MAPPED READS</td>
<td>N1</td>
<td>N2</td>
<td>......</td>
<td>......</td>
<td>Np</td>
</tr>
</tbody>
</table>

Similar to what has been done with microarray (but with different type of data)

\[ nxp \text{ with } n \gg p \]

however RNA-Seq data are also “functional” in mathematical sense, hence some different strategies could be considered

At the moment not yet practicable due to the limited number of samples available, but of great interest in a short term period
What about more “complex” biological questions?

5. How are the active components (i.e., genes) regulated?

The analysis of RNA-Seq is only one side of the coin. The road toward the fully understanding of differences of active components between conditions/states and changes occurring when perturbing the biological system requires the integration of RNA-Seq with other sources of data such as Chip-seq.

6. How do structural variants affect the expression?

Another interesting issue from biological point of view is the detection of SNPs in RNA-Seq data and their association to pathological conditions. A fully understanding of such aspects requires the integration of RNA-Seq with other sources of data such as Exome-seq.
**Hot-topic:** There are still very few statistical approaches.

Regression model:

\[
\log Y_i = \mu + \sum_{j=1}^{M} \beta_j X_{ij} + \epsilon_i,
\]

- Statistical test (i.e., significance \( \rightarrow \) regression coefficients different from zero)
- Model selection (prediction of gene expression given chip data)
- Sample Classification (i.e., use of ChIP-data for discriminating samples with differential expression profiles)
- etc
Conclusions (2)

- RNA-Seq experiments are a powerful tool for investigating the whole transcriptome, however they require the set up of “sophisticated” computational methods.
- There is no a single software that allows the user to perform all the main steps of the analysis → several tools need to be combined in pipeline, most of the time is spent in input/output formatting, no integrated tool available.
- Improvements are very fast. Novel software are going to be proposed almost every week, but still tools are often very badly documented and not of friendly usage.
- Most of the analysis also requires very advanced hardware resources.
- Results are very promising, but still all the capability and information have not been fully extracted → There is a need of develop novel computational/statistical tools.
- There is a strong need to combine multidisciplinary approaches → COST ACTION BM1006 on Next Generation Sequencing Data Analysis Network.

**Nature Methods** November 2009

Next-generation gap

John D McPherson
Software

About 50 of which are related to RNA-seq

http://seqanswers.com/wiki/Software

358→400 in 2 months
Other resources


http://rna-seqblog.com
References

153 → 216 papers in 2 months

**Computation for ChIP-seq and RNA-seq studies**

Shirley Pepke¹, Barbara Wold² & Ali Mortazavi²

**RNA-Seq: a revolutionary tool for transcriptomics**

Zhong Wang, Mark Gerstein and Michael Snyder

**Review Article**

RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing

Brian T Wilhelm, Josette-Renée Landry


Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing.

Horner DS, Pesesi C, Castrignanò T, De Mee PD, Liuni S, Sammeth M, Picard E, Pesole G.
Questions?