“RNA-Seq: from computational challenges to biological insights”
Valerio Costa, PhD
Laboratory of Human Genetics Diseases
at the Institute of Genetics and Biophysics “A.Buzzati-Traverso” IGB-CNR, Naples
Joint NETTAB 2010 and BBCC 2010 workshops, Naples, Italy
The central dogma of genetics: one gene, one protein

DNA (genotype) → Transcription → mRNA → Translation by ribosomes → Phenotype

- DNA: TACGCCCATGAAAT → ATGCGGGTACCTTA
- mRNA: AUGCGGUACCUUA
- Protein: NH₂-Methionine-Arginine-Tyrosine-Leucine-... Polypeptide

Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.
Pervasive transcription and genome complexity

Short and long intergenic non-coding transcripts

Intronic non-coding transcripts
Antisense transcripts

Gingeras 2009 (Nat Rev Genet)
The “next-generation” sequencing era

**RNA-Seq allows to:**

**Characterize organisms’ full set of genes**
- Detect and quantify expression from known genes;
- Find both new coding and non-coding genes;
  - Compare genes among organisms (evolution of genomes);

**Characterize transcript isoforms**
- Identify and quantify known splice events;
- Find novel alternative splice isoforms and/or transcript ends (5’-3’ UTRs);

**Monitor gene expression changes between cells/tissues/organisms or conditions**
- Identify differential expression between 2 conditions;
- Understand the basis of gene expression regulation in a disease;
- Identify gene regulatory regions (e.g. coupled with ChIP-Seq);
**RNA-Seq and microarrays**

**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)
- Nowadays still cheaper
- Low computational complexity

**RNA-Seq:**
- Low "background signal"
- Identification of novel transcribed regions and splice isoforms
- Determination of correct gene boundaries
- No upper limit for gene quantification
- Still expensive (sample preparation and sequencing)
- Much more computationally demanding
- Still limited amount of software available
<table>
<thead>
<tr>
<th>RNA source</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>PolyA RNA</td>
<td>• Polyadenylated fraction of the transcriptome</td>
</tr>
<tr>
<td></td>
<td>• Represents 2 – 5% of all transcripts</td>
</tr>
<tr>
<td></td>
<td>• Coding mRNA only</td>
</tr>
<tr>
<td>rRNA-depleted RNA</td>
<td>• Coding &amp; non-coding RNA minus the abundant rRNA</td>
</tr>
<tr>
<td>Total RNA</td>
<td>• See entire transcriptome</td>
</tr>
<tr>
<td></td>
<td>• Includes abundant rRNAs</td>
</tr>
</tbody>
</table>
Endothelial progenitor cells (EPCs)

- Proliferate and differentiate in vitro to endothelial cells
- Maintenance of endothelium integrity and repair
- Postnatal neovascularization

Bone marrow stromal cells

Haemangioblast

CD133+
CD34+
VEGFR2+

EPC

CD133 +/
CD34 +
VEGFR2 +
CD31 +
VE-cadherin -
vWF -

Hematopoietic stem cells (HSC)

CD133 -
CD34 +/

CD14 +
CD34 -

Early circulat

RESEARCH ARTICLE

Impairment of circulating endothelial progenitors in Down syndrome

Valerio Costa1, Linda Sommese2, Amelia Casamassimi3, Roberta Colicchio4, Claudia Angelini4, Valentina Marchesano5, Lara Milone6, Bartolomeo Farzati7, Alfonso Giovane8, Carmela Florito9, Monica Rienzo9, Marco Picardi9, Bice Avallone8, Massimiliano Marco Corsi10, Berardo Sarubbi11, Raffaele Calabro11, Paola Salvatore12, Alfredo Ciccodicola13, Claudio Napoli3
Ribodepletion with 5’ biotin-labeled probes

2x (Anti-5s RNA, Anti-5.8s RNA, Anti-18s RNA, Anti-28s RNA)
RNA fragmentation (3) and ligation to adaptors (4).

Retro-transcription (5).

Size selection by gel electrophoresis (6), and PCR amplification (7). Size distribution evaluation (8).

Emulsion PCR (9). Beads enrichment and deposition onto glass slides (10).

Costa et al., 2010 (J Biomed Biotech)
Mapping strategy

1. Quality assessment and filters (quality plot, remove low quality reads, ribosomal RNA reads, sequencing adapters);
2. Alignment to a reference genome (genome+junction library);
3. "Trim" the right-side of the reads and cyclically repeats the step;
4. Handle "multiple" reads;

A suitable treatment of the multiple matched reads is fundamental to reduce the bias.

Costa et al., 2010 (submitted)
Since the huge number - and the short size - of reads (50 nt in length), using conventional alignment algorithms is not feasible; In addition, not all developed aligners support all `.csfasta` formats from SOLiD platform;

*The alignment of reads in RNA-seq is particularly challenging due to the reads spanning across splice-junctions*

Costa et al., 2010 (*J Biomed Biotech*)
"Numbers" of DS and euploid RNA-Seq

[Bar chart showing Million of sequenced reads for Euploid and DS, with categories including Usable beads, After QV filter, After rRNA+adapter filter, Genomic mapped, and Junctions mapped.]
5. Visualize output data;
6. Quantify known “features” (at different level of resolution) or
7. Identify and quantify novel ones;
8. Perform between samples comparisons.
RNA-Seq sensitivity
1) **Identification and quantification of transcriptional regions:**
- "known" regions (i.e. RefSeq, UCSC annotated genes, transcripts);
- novel transcriptionally active regions (TARs);
- gene boundaries (5' and 3' UTRs analysis);

2) **Identification and quantification of splicing isoforms:**
- "known" transcript isoforms; detection of new alternative splice isoforms and their quantification

3) **Analysis of non-coding RNAs (ncRNAs):**
- detection and quantification of "known" ncRNAs;
- "" of new ncRNAs and their quantification;

4) **Detection of differentially expressed (DE) "features":**
- detection of DE "known" genes;
- "" of DE newly identified genes;
- "" of sample/condition-specific isoforms;
- "" of DE alternative splicing isoforms
- "" of DE ncRNAs;
1) Identification and quantification of transcriptional regions:
- "known" regions (i.e. RefSeq, UCSC annotated genes, transcripts);
- novel transcriptionally active regions (TARs);
- gene boundaries (5' and 3' UTRs analysis);

Quantification based on RefSeq Annotation:
- Remove ambiguities due to genes overlapping by strand;
- Use either “exon reads” and “junction reads”;
- Use unique reads + "uniquely assigned" reads after the “rescue” step.

Expression was measured as the Number of Reads Mapped on the feature $i$
or as

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

$N =$ Total number of mapped reads

$L_i =$ Length (bp) of the feature $i$

Reads Per Kilobase of transcript per Million of mapped reads (RPKM)
RefSeq genes were classified according to RPKM values into 5 categories of expression:

1) very low
2) low
3) intermediate
4) high
5) very high
Analysis of "extra-genic" transcription

Very different mapping from polyA+ enrichment experiments

If N reads < threshold

50 kb 100 kb 50 kb
Analysis of "extra-genic" transcription

InTARS (Intronic Transcriptionally Active Regions)

IgTARS (Intergenic Transcriptionally Active Regions)
Analysis of 5’ and 3’ UTRs

- Extended 5’ UTR
- Extended 3’ UTR

[Graph showing analysis of UTRs]
1) Identification and quantification of transcriptional regions:
- "known" regions (i.e. RefSeq, UCSC annotated genes, transcripts);
- novel transcriptionally active regions (TARs);
- gene boundaries (5' and 3' UTRs analysis);

2) Identification and quantification of splicing isoforms:
- "known" transcript isoforms;
- detection of new alternative splice isoforms and their quantification (in progress);

3) Analysis of non-coding RNAs (ncRNAs):
- detection and quantification of "known" ncRNAs;
- "" of new ncRNAs and their quantification;

4) Detection of differentially expressed (DE) "features":
- detection of DE "known" genes;
- "" of DE newly identified genes;
- "" of sample/condition-specific isoforms;
- "" of DE alternative splicing isoforms;
- "" of DE ncRNAs;
Guilty by evidence

The presence of multiple isoforms is inferred by reads mapping to *multiple donor / acceptor* splice junctions.

Known RefSeq junctions

New "combinatorial" RefSeq junctions
Alternative splicing events confirmed by at least **3 mapped reads**

- DS: 3005
- Euploid: 1190
- 35-40%
- 1263 specific (1172 genes)
- RefSeq, UCSC Ensembl
- 639 unannotated junctions (601 genes)
- 50%

Alternative splicing events confirmed by at least **5 mapped reads**

- DS: 1003
- Euploid: 618
- 30%
- 294 specific (290 genes)
- RefSeq, UCSC Ensembl
- 135 unannotated junctions (132 genes)
- 50%

- DS: 2753
- Euploid: 1165
- 70%
- 985 specific (909 genes)
- RefSeq, UCSC Ensembl
- 678 unannotated junctions (623 genes)
- 70%
DS-specific junction

Euploid-specific junction

- Candidate state-specific splice junctions (>5 mapped reads)
- Discarded state-specific splice junctions (<5 mapped reads)
1) **Identification and quantification of transcriptional regions:**
- "known" regions (i.e. RefSeq, UCSC annotated genes, transcripts);
- novel transcriptionally active regions (TARs);
- gene boundaries (5' and 3' UTRs analysis);

2) **Identification and quantification of splicing isoforms:**
- “known” transcript isoforms;
- detection of new alternative splice isoforms and their quantification

3) **Analysis of non-coding RNAs (ncRNAs):**
- detection and quantification of "known" ncRNAs;
- " " of new ncRNAs and their quantification (in progress)

4) **Detection of differentially expressed (DE) "features":**
- detection of DE "known" genes;
- " " of DE newly identified genes;
- " " of sample/condition-specific isoforms;
- " " of DE alternative splicing isoforms
- " " of DE ncRNAs;
The transcription beyond rRNA

- A significant increase (170-fold) of mean RPKM values of snoRNAs vs mRNAs;

- About 95-98% of snoRNAs belong to "Very High RPKM" category and almost exclusively map within introns of genes (host) belonging to "Very High" & "High RPKM" categories.
- A strong correlation between reads' distribution and functional snoRNA sites, suggesting these short RNA fragments may derive from the processing of snoRNAs;

- Some of them may have miRNA-like activities (very recently termed sno-miRNAs).
1) Identification and quantification of transcriptional regions:
- "known" regions (i.e. RefSeq, UCSC annotated genes, transcripts);
- novel transcriptionally active regions (TARs);
- gene boundaries (5' and 3' UTRs analysis);

2) Identification and quantification of splicing isoforms:
- "known" transcript isoforms;
- detection of new alternative splice isoforms and their quantification

3) Analysis of non-coding RNAs (ncRNAs):
- detection and quantification of "known" ncRNAs;
- "" of new ncRNAs and their quantification;

4) Detection of differentially expressed (DE) "features":
- detection of DE "known" genes;
- "" of DE newly identified genes;
- "" of sample/condition-specific isoforms (in progress);
- "" of DE alternative splicing isoforms (in progress);
- "" of DE ncRNAs;
Significant changes in the expression of genes are usually identified by using a statistical Test and the results are then corrected for multiple testing.

Unfortunately, one cannot use ordinary tests developed for microarray since RNA-Seq data are count data, and they are heteroscedastic (have no the same finite variance).

- Statistical significance has been inferred from total reads count for each RefSeq gene combining 3 tests:
  - DEGseq (based on Poisson distribution)
  - DESeq and edgeR (based on negative binomial).

- Such tests are based on slightly different assumptions that usually produce a different level of stringency.
Differential expression of RefSeq genes

STRONG = detected with all 3 methods
GOOD = detected with 2 methods
ACCEPTABLE = detected with only 1 method
WEAK = below the FC threshold (1,5)
quantitative RT validation

Differentially expressed genes

Relative expression level

* p<0.01
** p<0.05
Differentially expressed genes

Gene pathways

Biological Processes

- cell adhesion
- immune response
- inflammatory response
- JAK-STAT cascade
- response to lipopolysaccharide
- response to virus
- signal transduction
Differential expression of extra-genic regions

A

IgTARs

B

InTARs

Evidence of DE
- Strong
- Good
- Acceptable
- Weak

1.5 fold change
Differential expression of snoRNAs

46 SNORD (3 up- and 43 down-regulated) on a total of 171 expressed (27%);
31 SNORA (9 up- and 22 down-regulated) on a total of 95 expressed (32.6%);
9 SCARNA (2 up- and 7 down-regulated) on a total of 23 expressed (39%);

The gene with the highest expression on HSA21 was a member of H/ACA box, SNORA80, DE in the trisomic cells.
In summary

- Massive transcriptome sequencing of DS endothelial progenitors, and differential expression vs euploid cells;
- Splice isoforms (known and novel) of crucial genes, even those specifically expressed in cells with trisomy;
- Detection and quantification of snoRNAs, miRNAs and ncRNAs, emerging as candidates to the pathogenesis of human diseases;
- Correlation between reads' distribution and snoRNA processing, and identification of candidate sno-miRNAs;
- Differential expression of ncRNAs;
- Differential expression of newly identified "extra-genic" regions actively transcribed in DS cells vs euploid;
Future perspectives (1/2)

- RNA-Seq experiments are a powerful tool for addressing biological questions, although they still require the setup of “sophisticated” computational methods and the development of novel computational/statistical tools;

To develop a probabilistic model which takes into account the uncertainty due to the mapping

To build appropriate gene models to better define & quantify the high level of transcription within yet unannotated extra-genic regions

To reconstruct, and thus further quantifying, multiple isoforms of a transcript (isoform abundance).

*TopHat* aligns reads to the genomes using *Bowtie* - an ultrafast short reads aligner - and then analyzes the mapping results to identify splice junctions between exons (both known & newly identified).

*Cufflinks* assembles transcripts, estimates their abundances, and tests DE and regulation.
Biological conclusions inferred from the direct comparisons of two samples are however limited;

RNA-Seq experiments can (optimistically) reduce the technical variability, but they do not affect the biological variability.

Future perspectives (2/2)

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

- Although our results are very promising, all the capability and information have not been fully extracted from data;
- Further steps of "biological validations" (Real-Time PCR, WesternBlot, RNA interference, etc.) are also required;
Acknowledgements

Experimental design & Sample preparation

Human genetics disease Lab
Alfredo Ciccodicola
Marianna Aprile
Roberta Esposito

SOLiD Sequencing Core Facility
Stefania Crispi
Luigi Leone
Aldo Donizetti

Patients’ enrolling & Sample preparation

Claudio Napoli
Raffaele Calabrò
Berardo Sarubbi
Linda Sommese
Amelia Casamassimi
Paola Salvatore

Data analysis

Claudia Angelini

Margherita Mutarelli

Piergiuseppe De Berardinis
Luciana D’Apice

Data validation