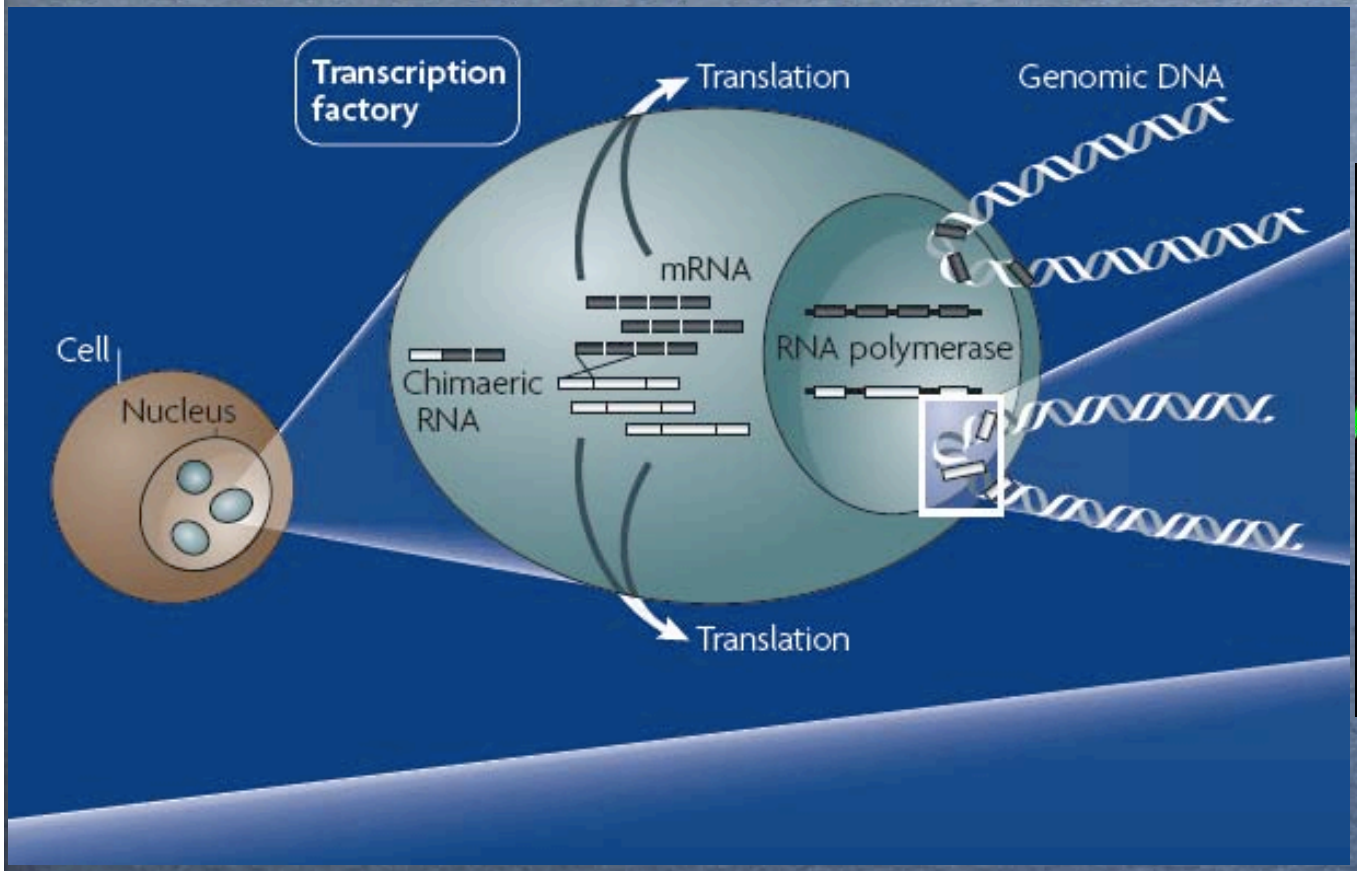
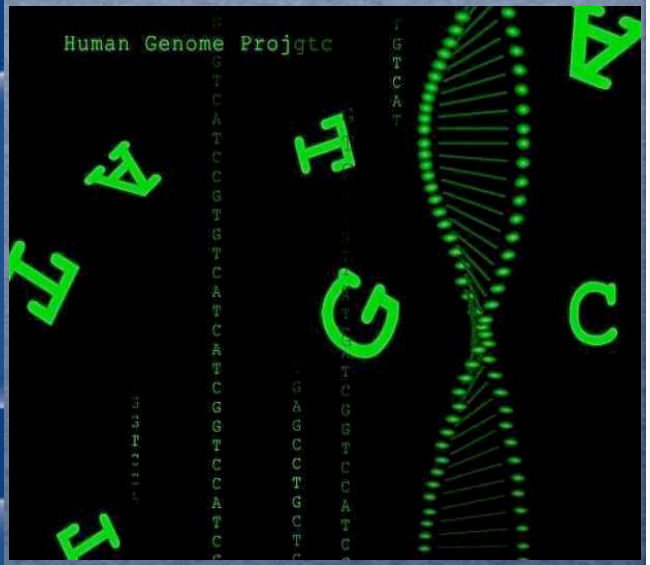




"RNA-Seq: from computational challenges to biological insights"

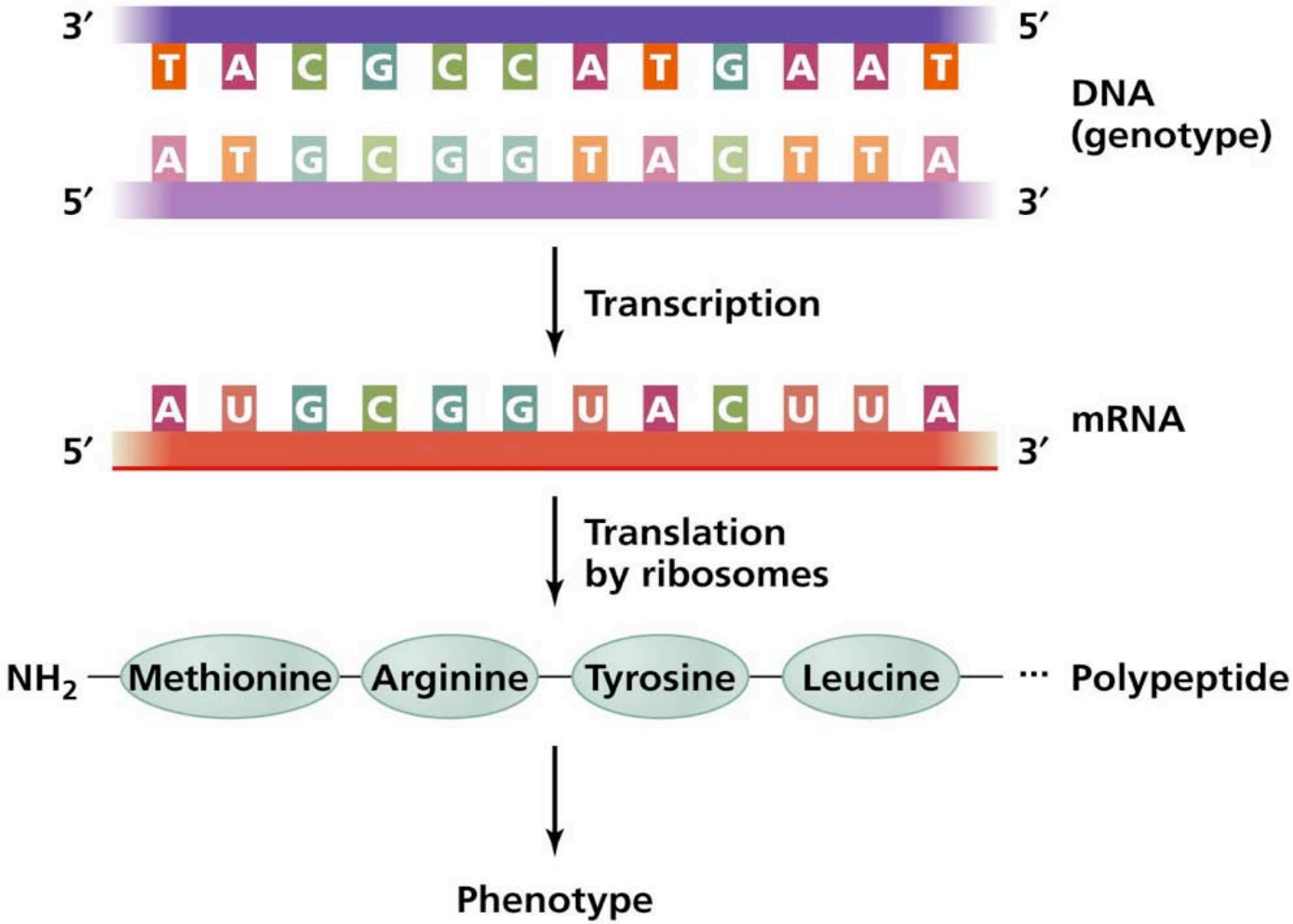
Valerio Costa, PhD

Laboratory of Human Genetics Diseases

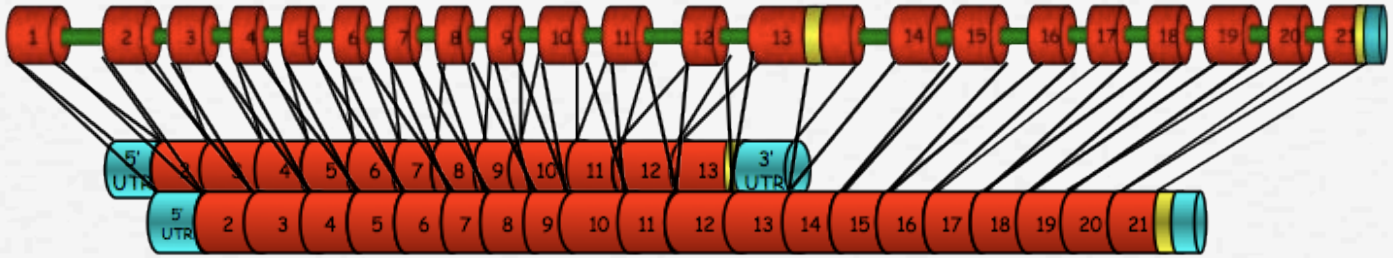
at the Institute of Genetics and Biophysics "A.Buzzati-Traverso" IGB-CNR, Naples

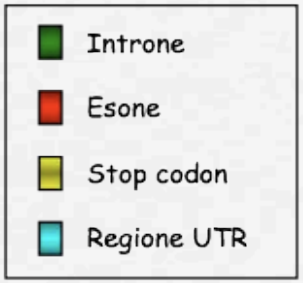
Joint NETTAB 2010 and BBCC 2010 workshops, Naples, Italy

The central dogma of genetics: one gene, one protein

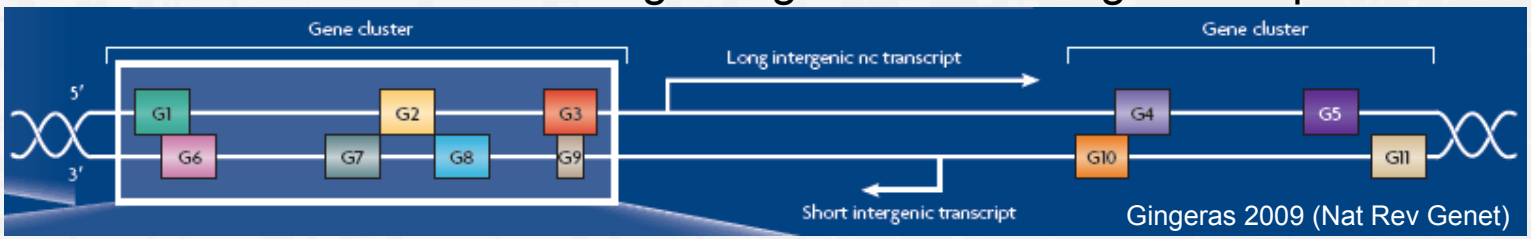


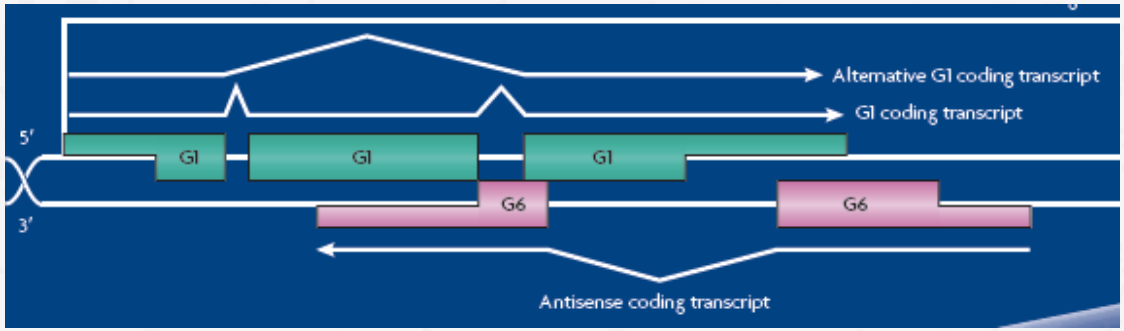
Pervasive transcription and genome complexity



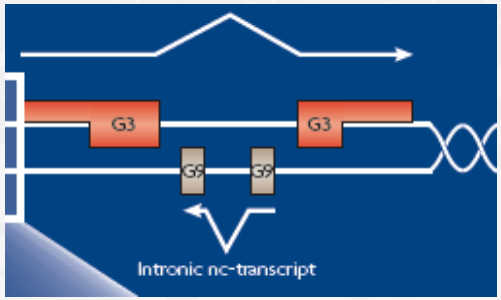


Short and long intergenic non-coding transcripts





Antisense transcripts



Intronic non-coding transcripts

The "next-generation" sequencing era

RNA-Seq allows to:

Characterize organisms' full set of genes

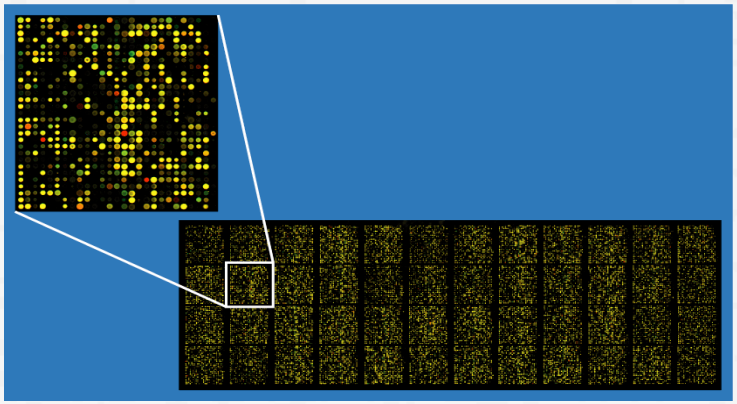
- Detect and quantifiy expression from known genes;
- Find both new coding and non-coding genes;
- Compare genes among organisms (evolution of genomes);

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 "large" sample production

- Low computational complexity

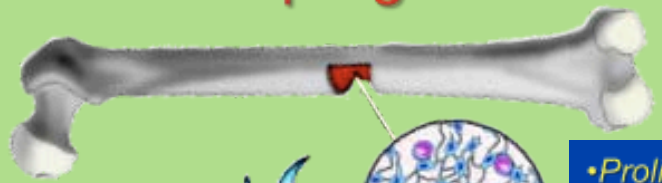
RNA-Seq:- Low "background signal"- Identification of novel transcribed regions and splice isoforms;- Determination of correct gene boundaries- No upper limit for gene quantification

- Still expensive (sample preparation and sequencing)

- Much more computationally demanding- Still limited amount of software available

RNA source	Characteristics
PolyA RNA	• Polyadenylated fraction of the transcriptome • Represents 2 – 5% of all transcripts • Coding mRNA only
rRNA-depleted RNA	• Coding & non coding RNA minus the abundant rRNA
Total RNA	• See entire transcriptome • Includes abundant rRNAs

Endothelial progenitor cells (EPCs)



Bone marrow stromal cells

Haemangioblast

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

EPC

CD133+
CD34+
VEGFR2+

Hematopoietic stem cells (HSC)

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

CD133 -
CD34 +/-

CD14+
CD34-

Early circulat

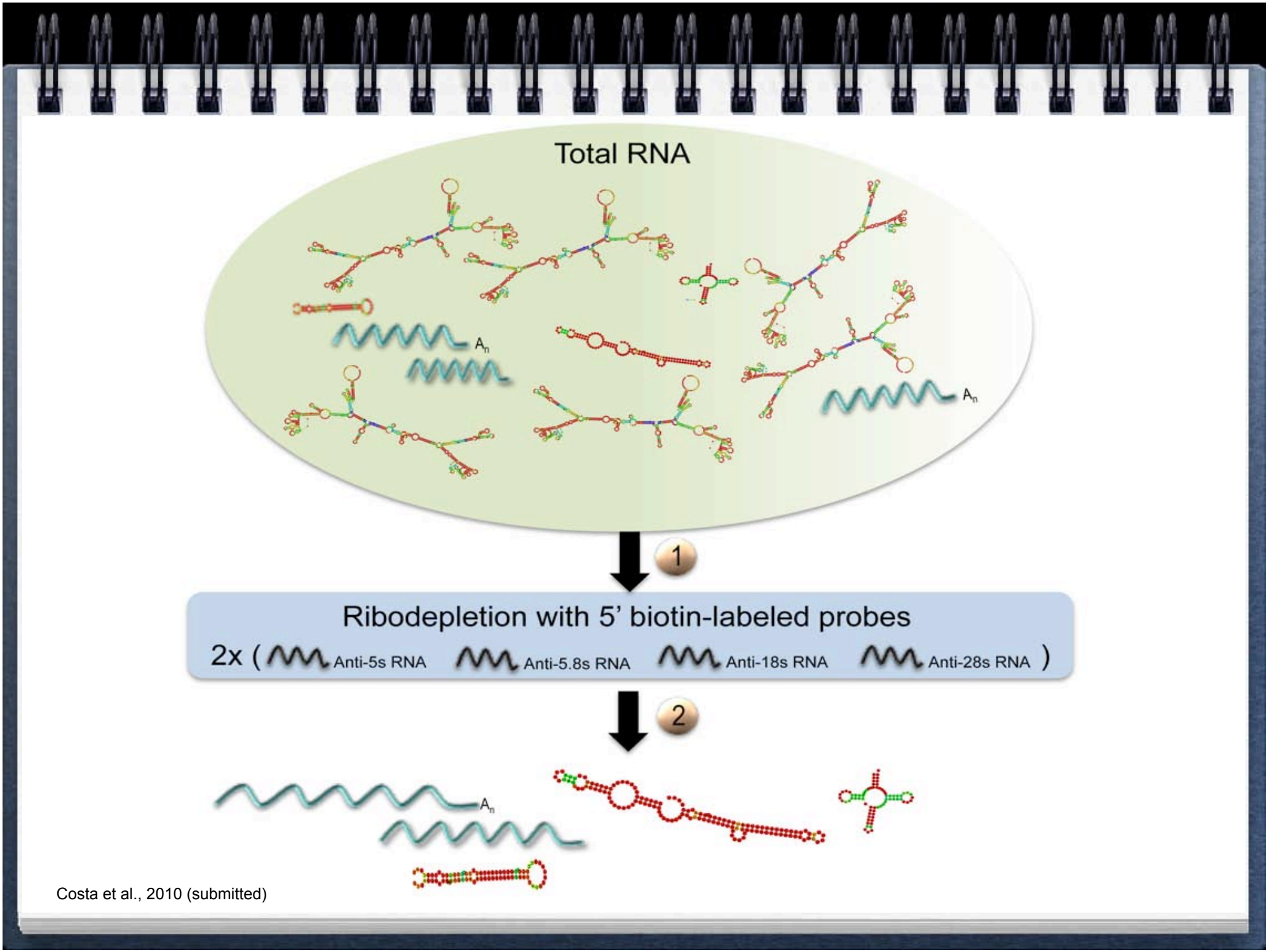
Costa *et al. BMC Medical Genomics* 2010, **3**:40
http://www.biomedcentral.com/1755-8794/3/40

BMC Medical Genomics

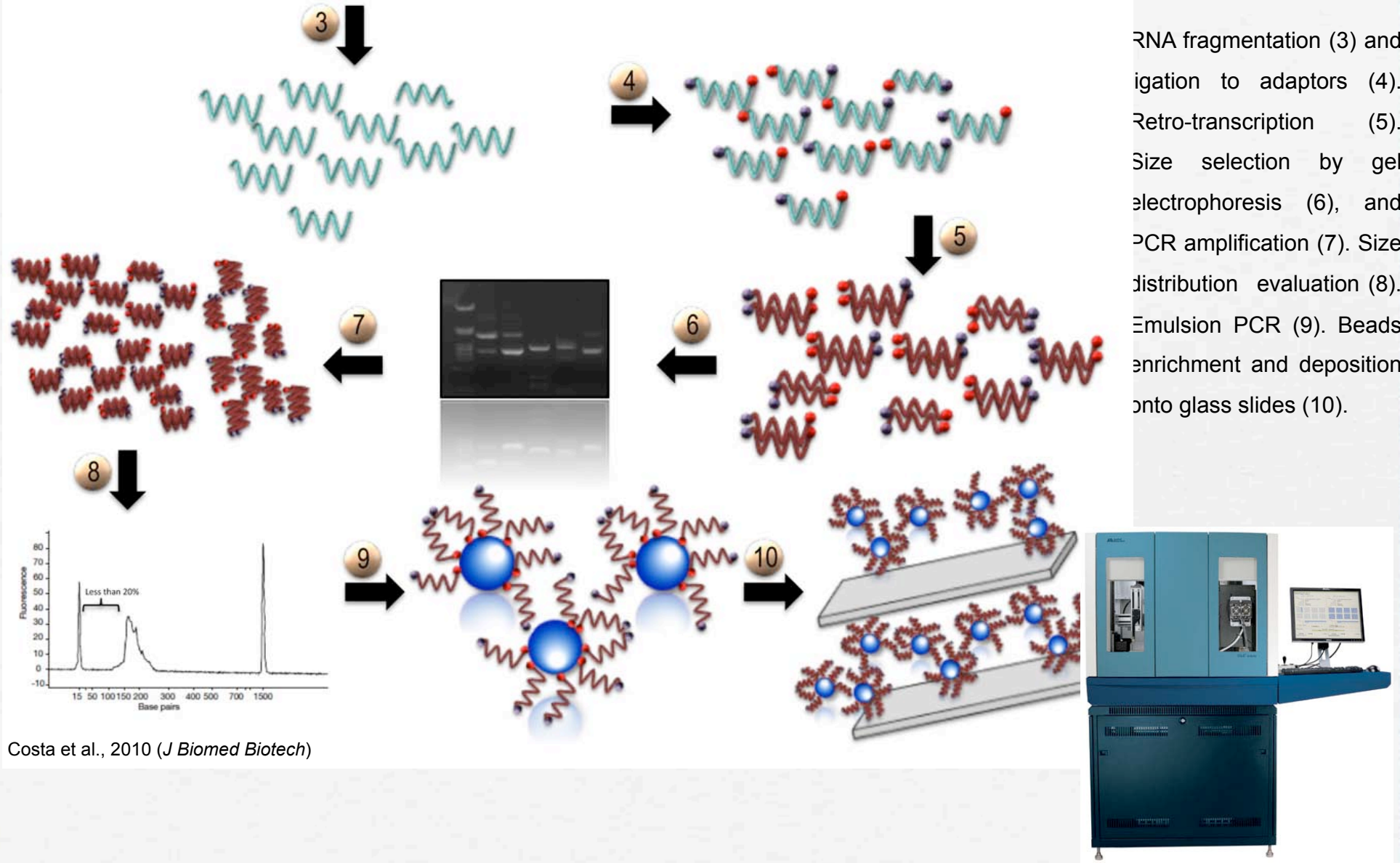
RESEARCH ARTICLE Open Access

Impairment of circulating endothelial progenitors in Down syndrome

Valerio Costa[1], Linda Sommese[2], Amelia Casamassimi[3], Roberta Colicchio[4], Claudia Angelini[5], Valentina Marchesano[6], Lara Milone[7], Bartolomeo Farzati[3], Alfonso Giovane[7], Carmela Fiorito[8], Monica Rienzo[3], Marco Picardi[9], Bice Avallone[6], Massimiliano Marco Corsi[10], Berardo Sarubbi[11], Raffaele Calabrò[11], Paola Salvatore[12], Alfredo Ciccodicola[1*†], Claudio Napoli[3†]



Costa et al., 2010 (submitted)



RNA fragmentation (3) and igation 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

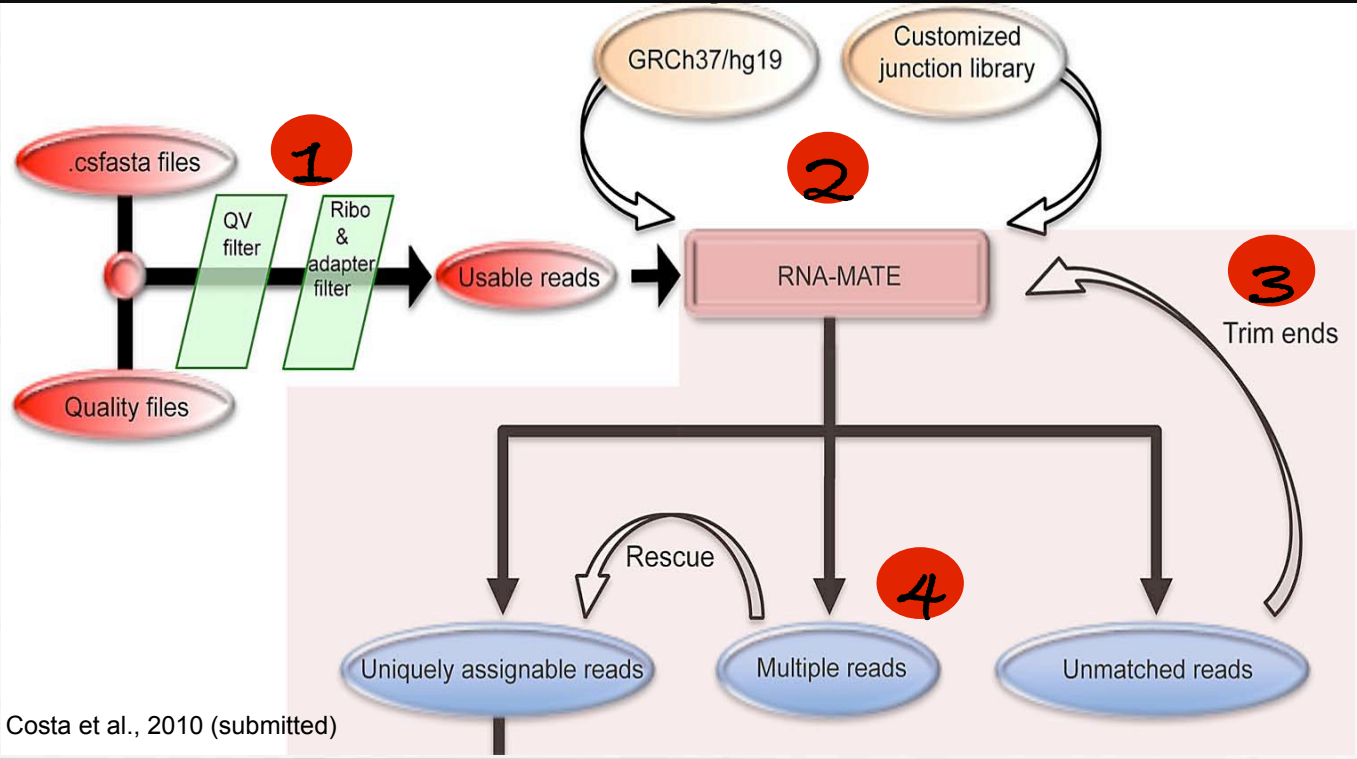
.csfasta

>852_2042_1999_F3T320112011230222013321101020110311320130233210023 03

.qual

>852_2042_1999_F319 20 14 14 8 5 9 16 11 11 6 14 21 14 11 21 -1 20 11 21 12 22 14 18 14 6 11 16 14 16 5 11 23 13 18 4 6 20 13 15 21 17 18 15 11 4 8 7 5 11

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

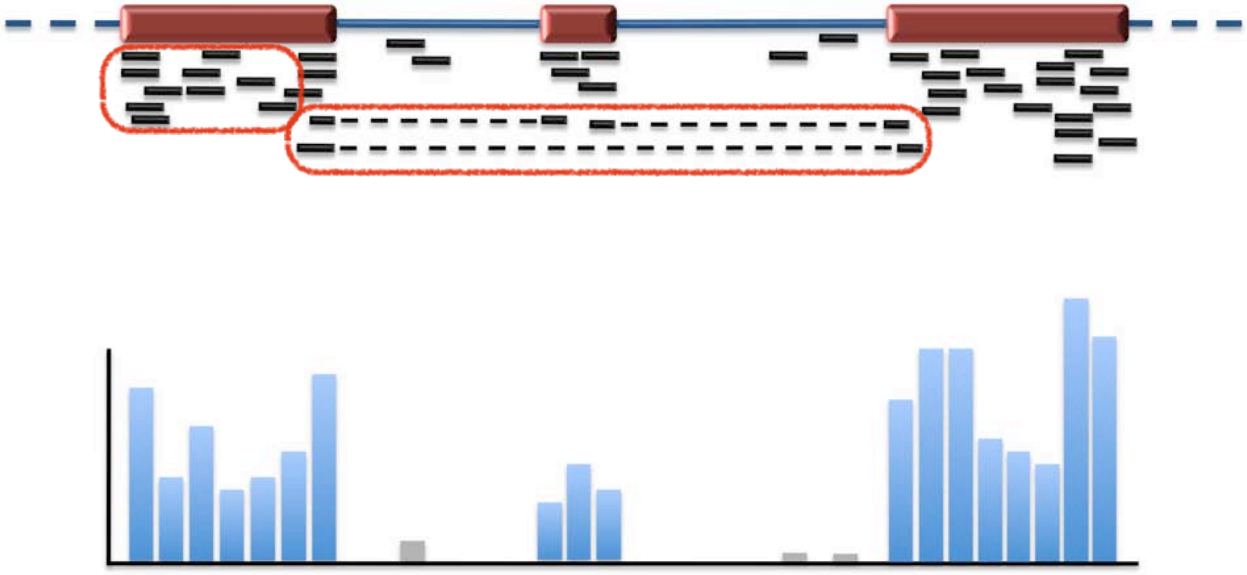


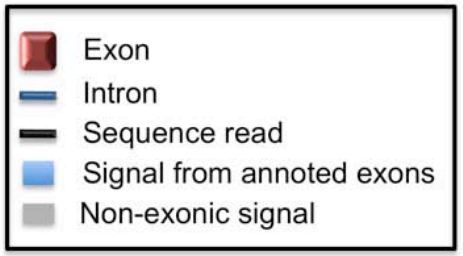
Costa et al., 2010 (submitted)

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 rigth-side of the reads and cyclically repeats the step;
4. Handle “multiple” reads;

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



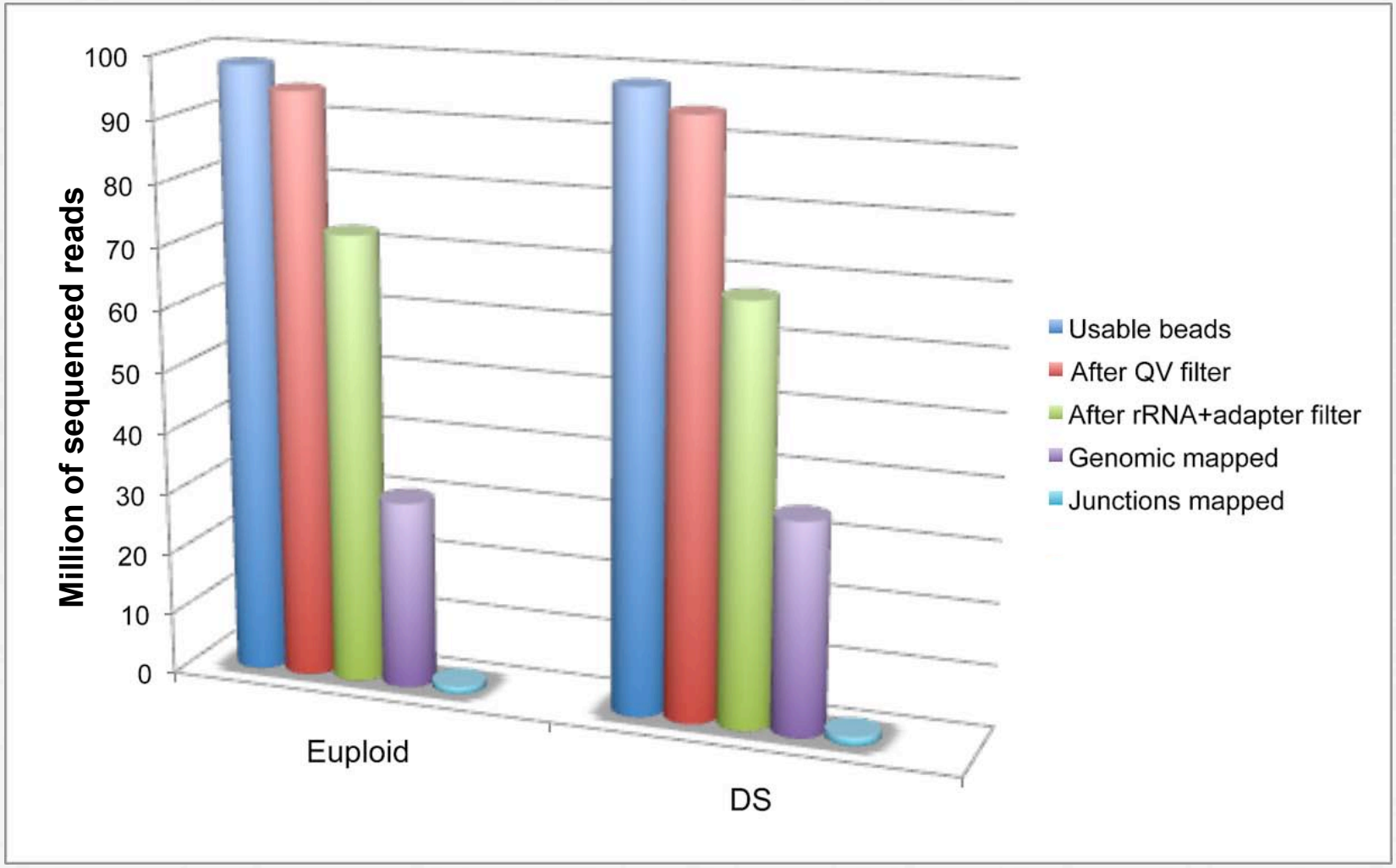


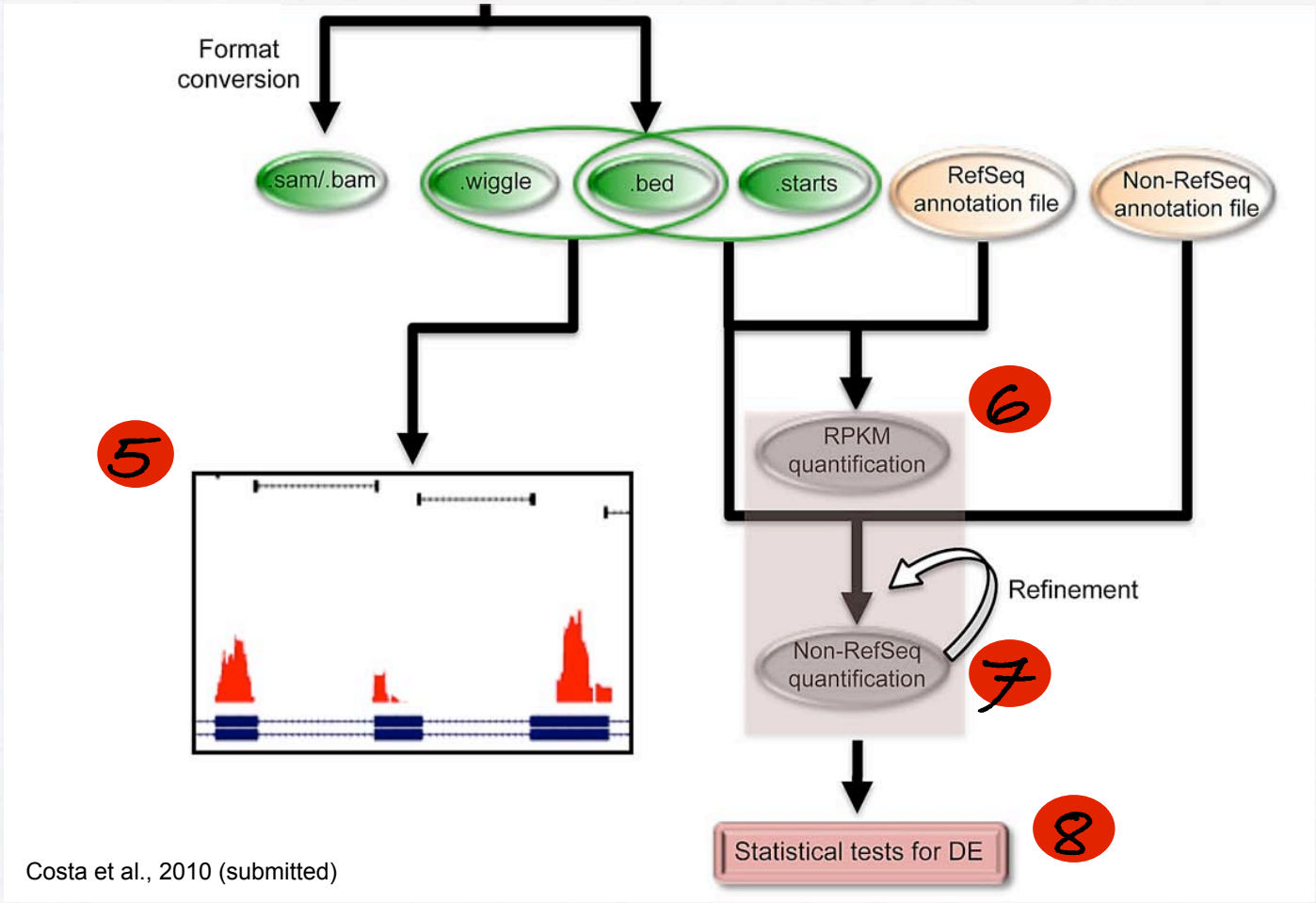
RNA-seq alignment to a reference genome

- **Model based junction detection**
- **Data driven junction detection**
- ***"De-novo"* transcript assembly**

Costa et al., 2010 (*J Biomed Biotech*)

"Numbers" of DS and euploid RNA-Seq





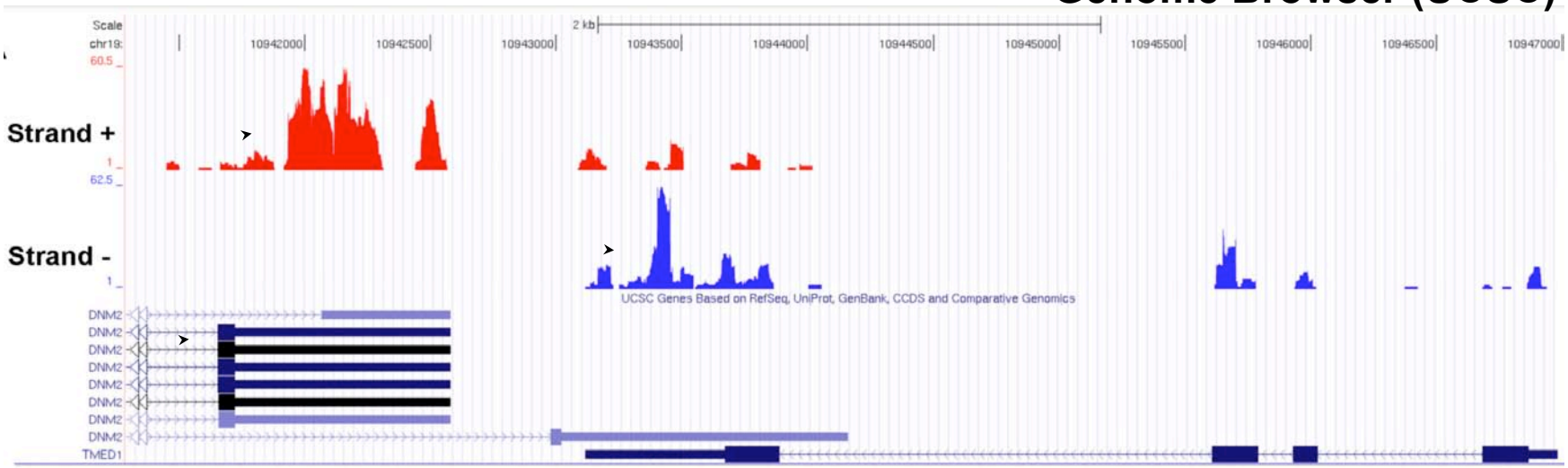
Costa et al., 2010 (submitted)

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.

Genome Browser (UCSC)

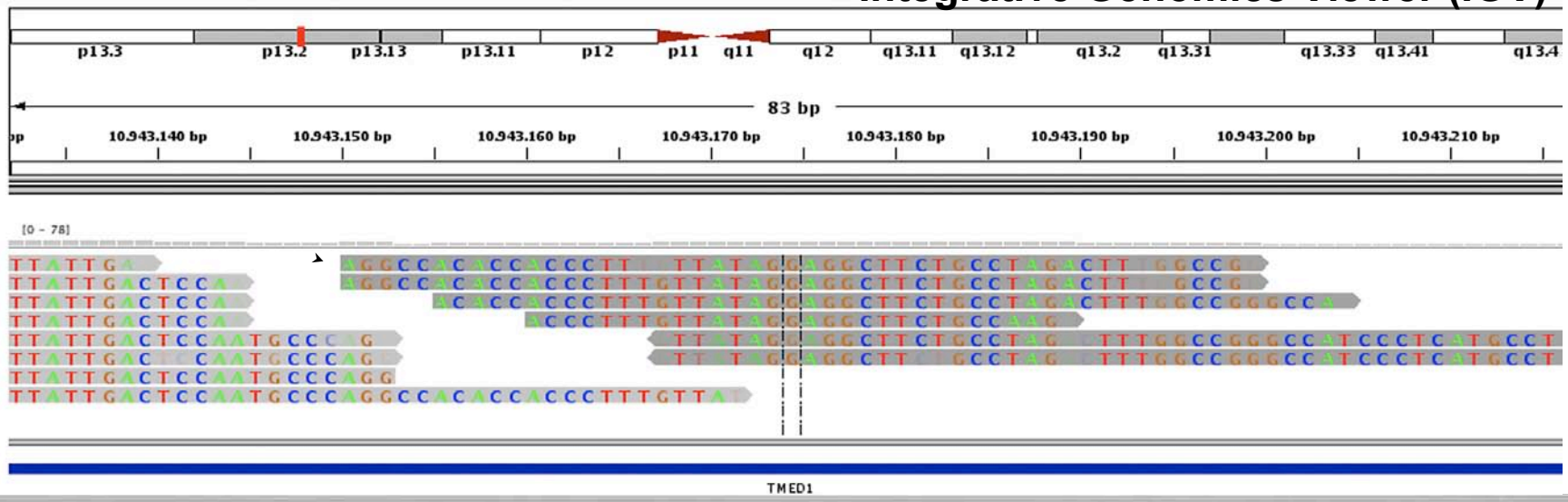
.WIG

.BED

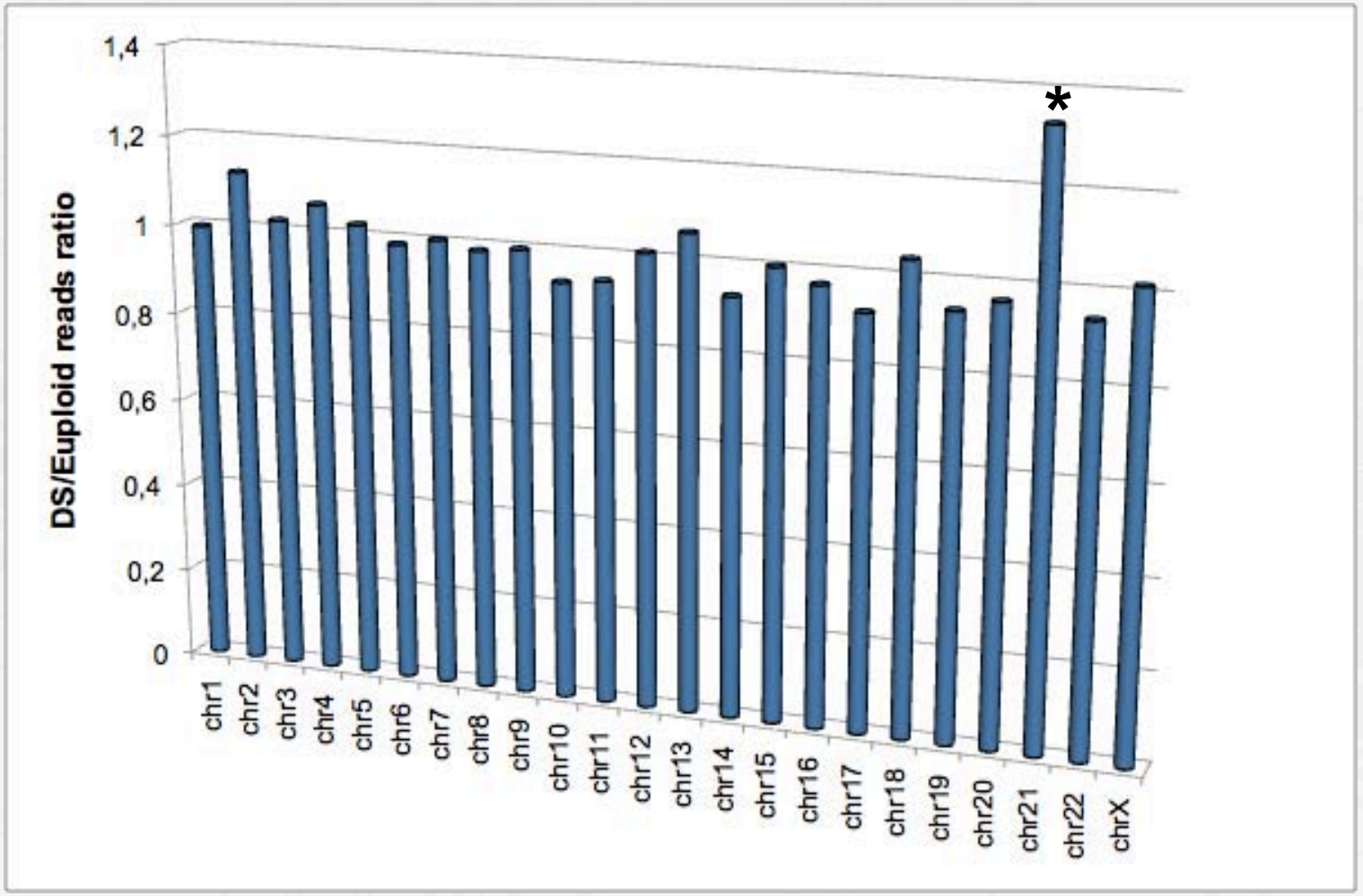


Integrative Genomics Viewer (IGV)

.BAM



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;

Within sample analysis

Between samples analysis

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:

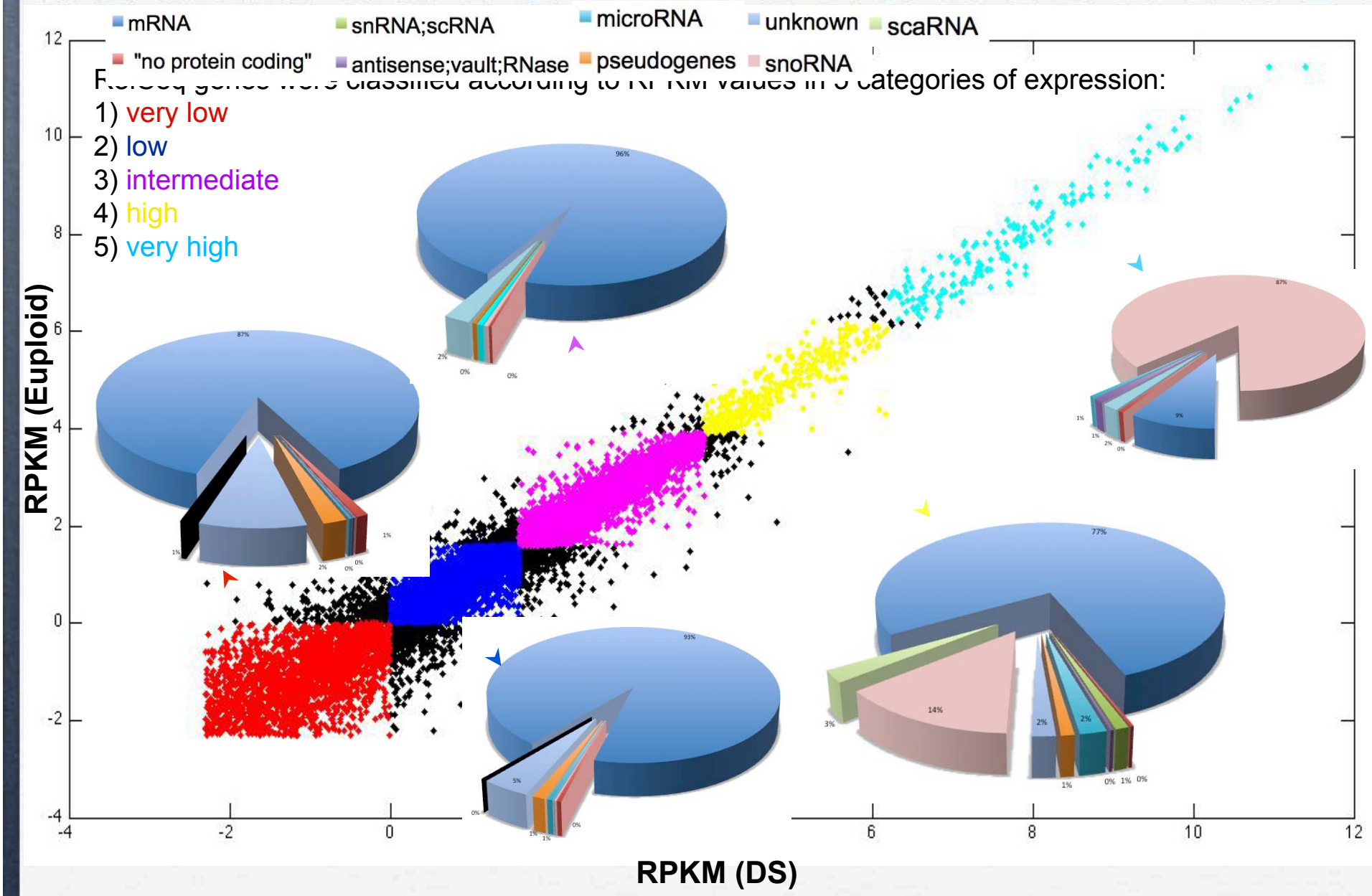
- Remove ambiguities due to genes overlapping by strand;
- Use either “exon reads” and “junction reads”;
- Use unique reads + "uniquely assigned" reads after the “*rescue*” step.

Expression was measured as the Number of Reads Mapped on the feature *i*
or as
Reads **P**er **K**ilobase of transcript per **M**illion of mapped reads (**RPKM**)

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

N = Total number of mapped read

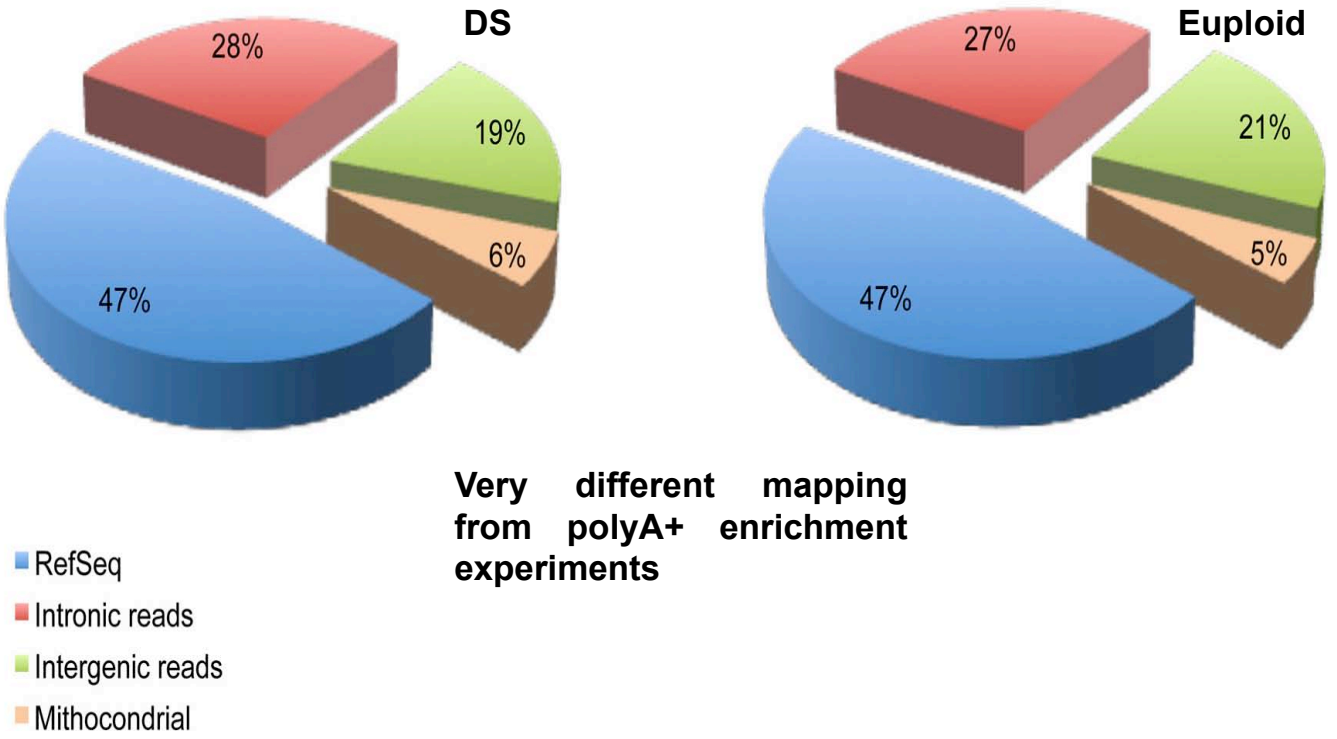
L_i = Length (bp) of the feature i



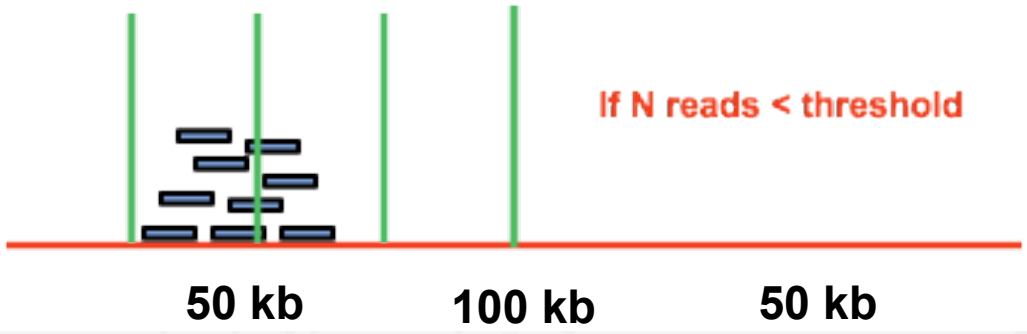
Refseq genes were classified according to RPKM values in 5 categories of expression:

1) very low
2) low
3) intermediate
4) high
5) very high

Analysis of "extra-genic" transcription

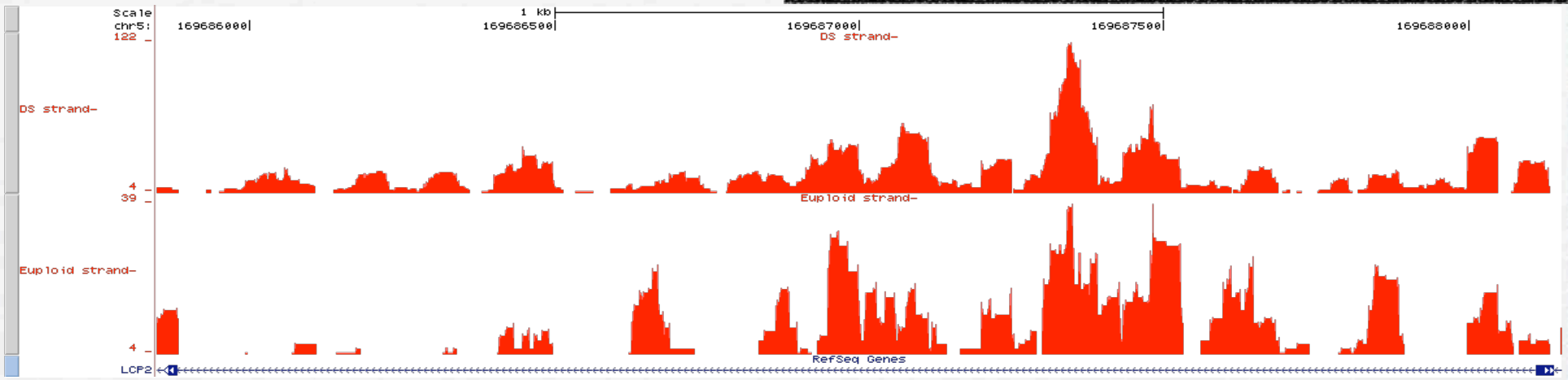


Very different mapping from polyA+ enrichment experiments

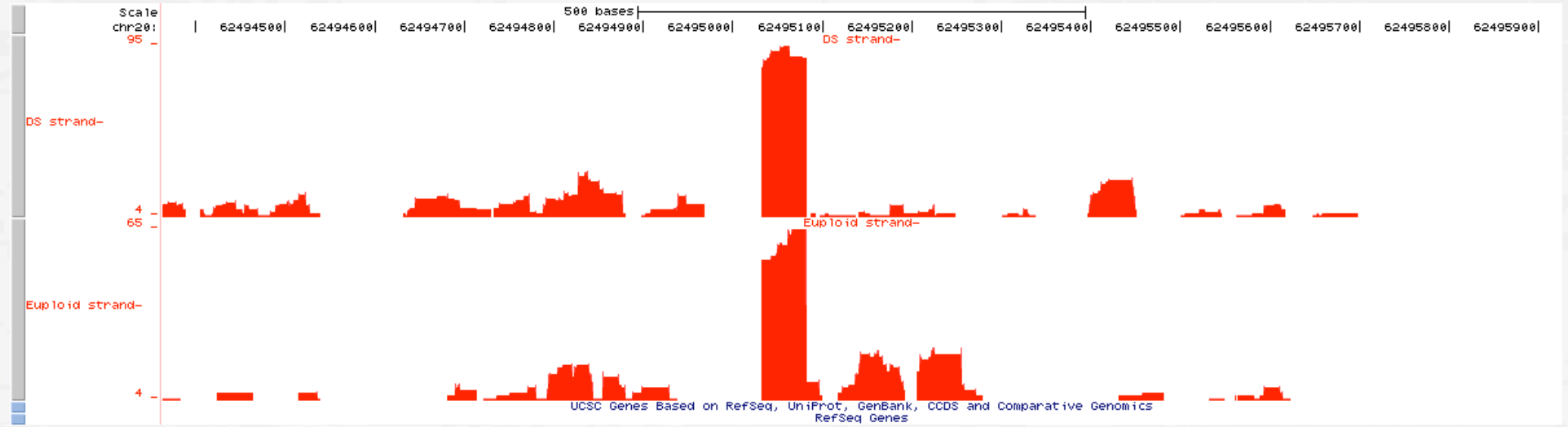


Analysis of "extra-genic" transcription

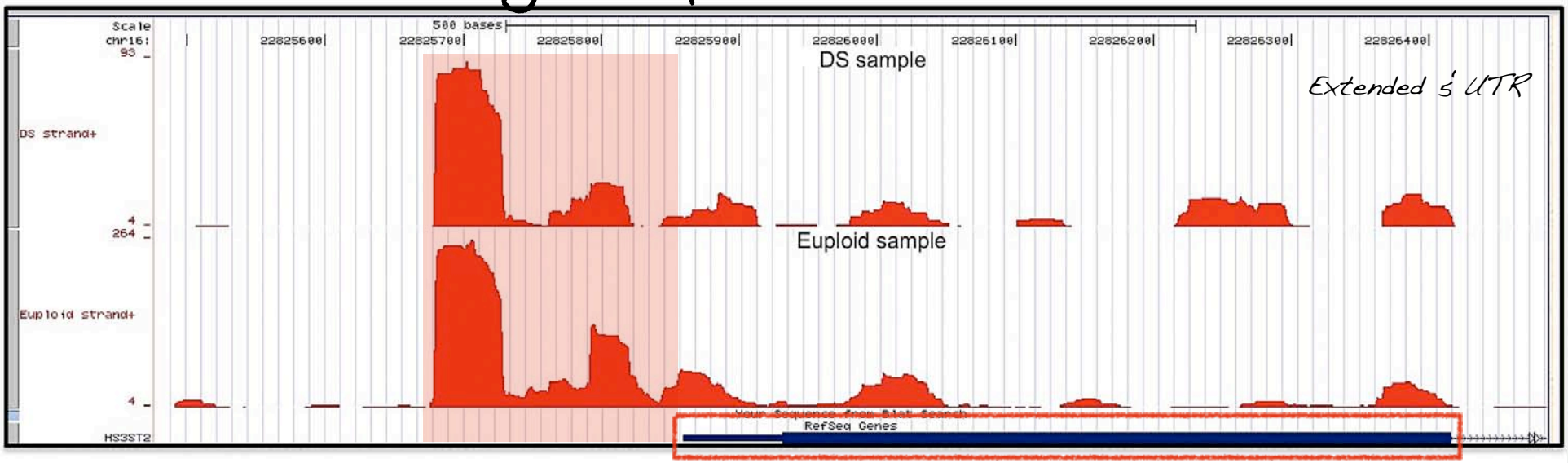
InTARS (Intronic Transcriptionally Active Regions)

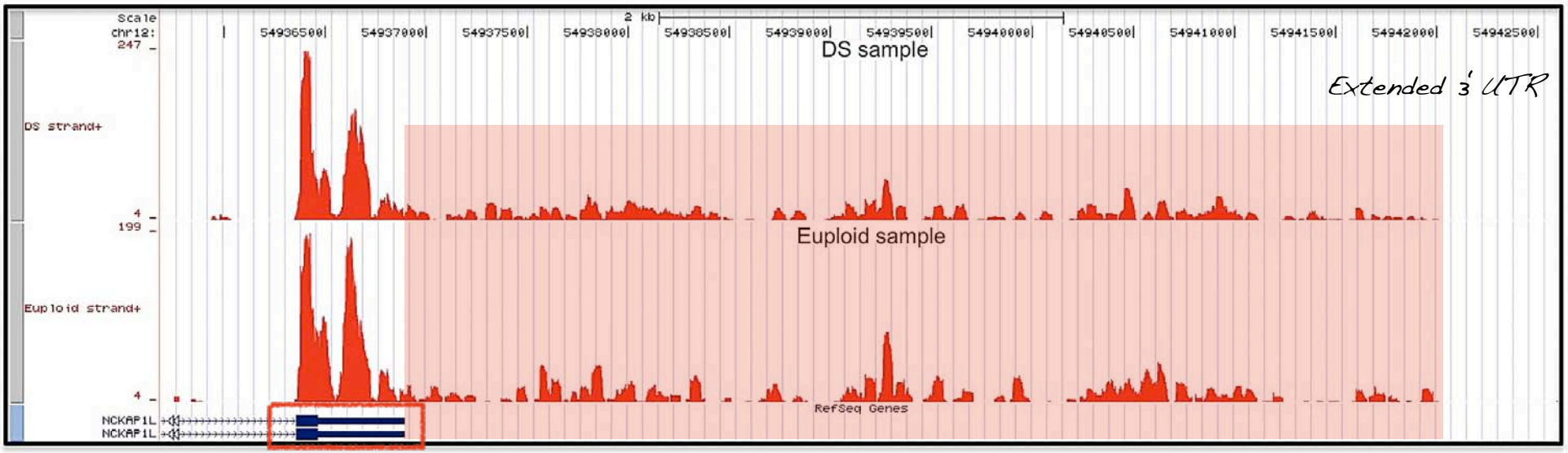


IgTARS (Intergenic Transcriptionally Active Regions)



Analysis of 5' and 3' 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:

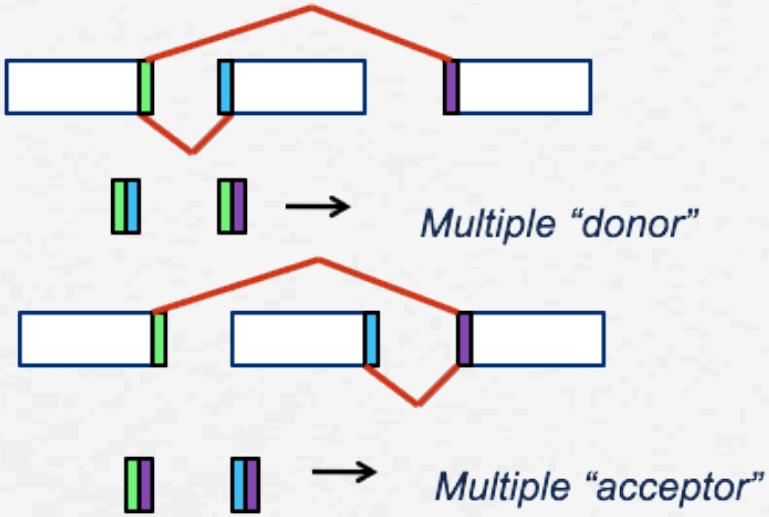
- "known" transcript isoforms;- detection of new alternative splice isoforms and their quantification (in progress)

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;

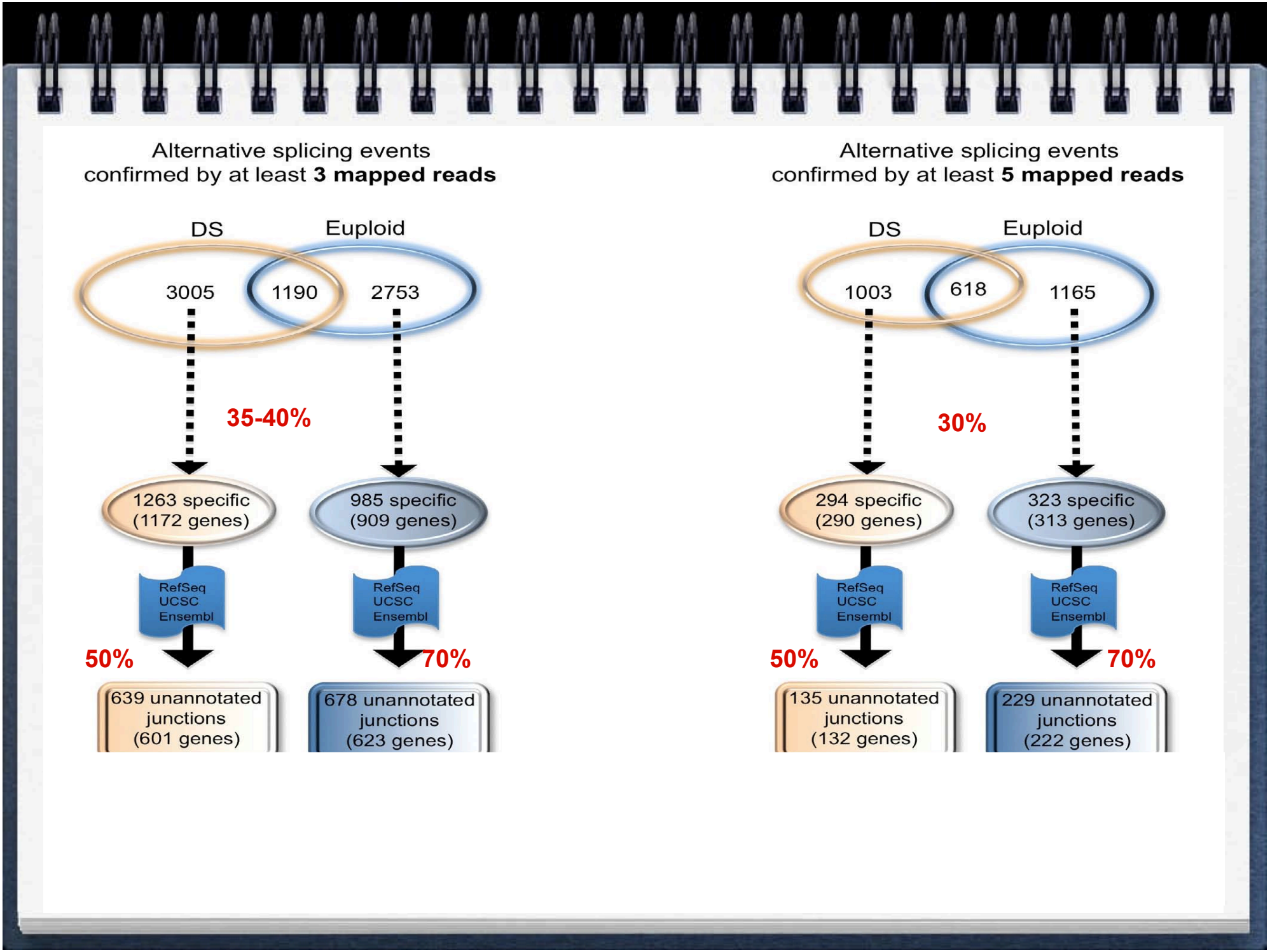


"*Guilty by evidence*"

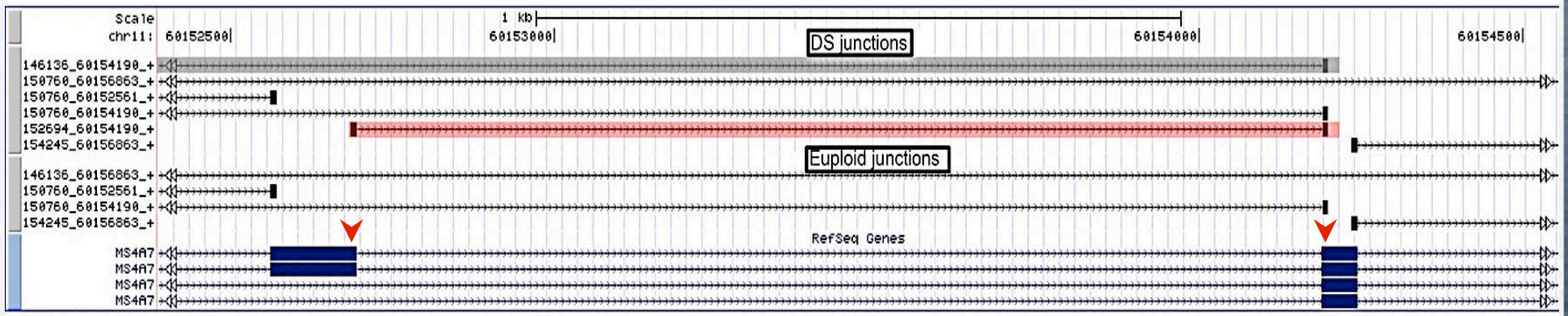
The presence of multiple isoforms is inferred by reads mapping to *multiple donor /acceptor*" splice junctions.

Known RefSeq junctions

New "combinatorial" RefSeq junctions


Alternative splicing events
confirmed by at least 3 mapped reads
DS
Euploid
3005
1190
2753
35-40%
1263 specific
(1172 genes)
985 specific
(909 genes)
RefSeq
UCSC
Ensembl
RefSeq
UCSC
Ensembl
50%
70%
639 unannotated
junctions
(601 genes)
678 unannotated
junctions
(623 genes)
Alternative splicing events
confirmed by at least 5 mapped reads
DS
Euploid
1003
618
1165
30%
294 specific
(290 genes)
323 specific
(313 genes)
RefSeq
UCSC
Ensembl
RefSeq
UCSC
Ensembl
50%
70%
135 unannotated
junctions
(132 genes)
229 unannotated
junctions
(222 genes)

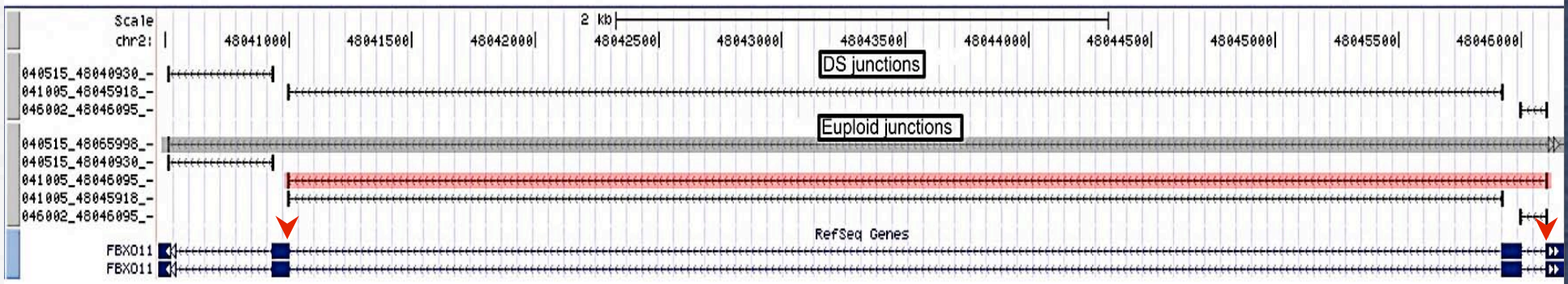
DS-specific junction



Candidate state-specific splice junctions (>5 mapped reads)

Discarded state-specific splice junctions (<5 mapped reads)

Euploid-specific junction



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)

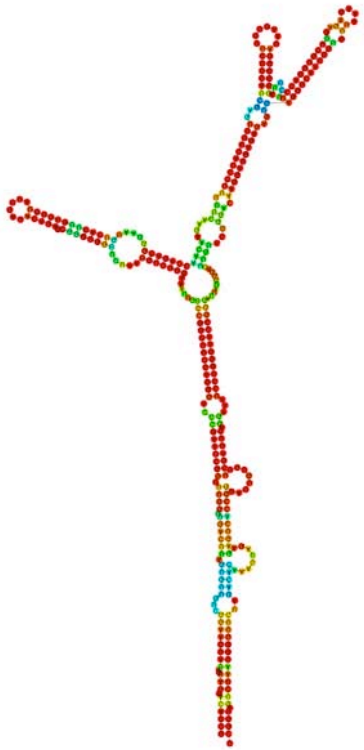
4) Detection of differentially expressed (DE) "features":

- detection of DE "known" genes;
- " " of DE newly identified genes;- " " of sample/condition-specific isoforms;- " " of DE alternative splicing isoforms
- " " of DE ncRNAs;

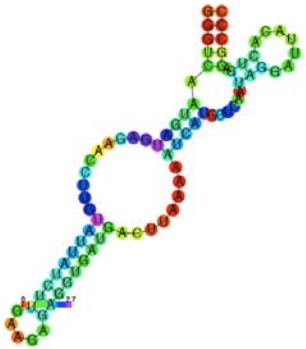
The transcription beyond rRNA

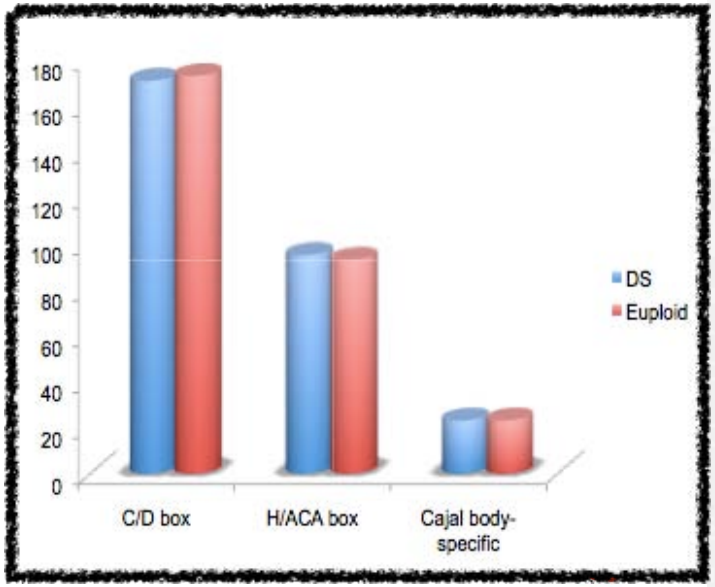
Small nucleolar RNA (snoRNA)

Cajal-body scaRNA

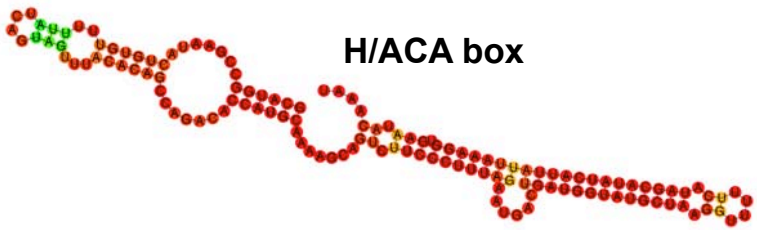


C/D box



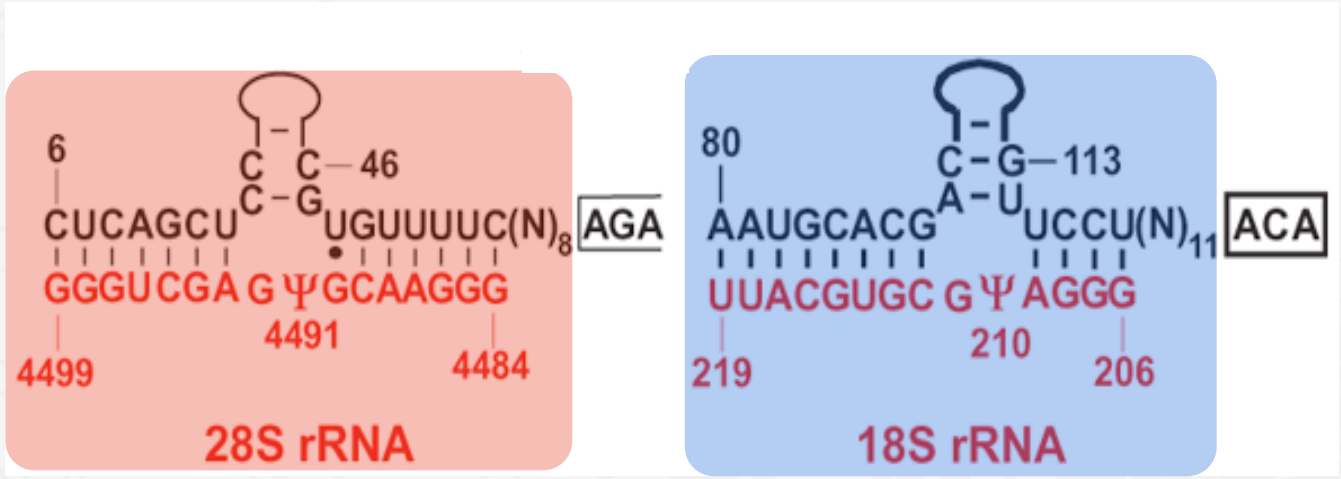


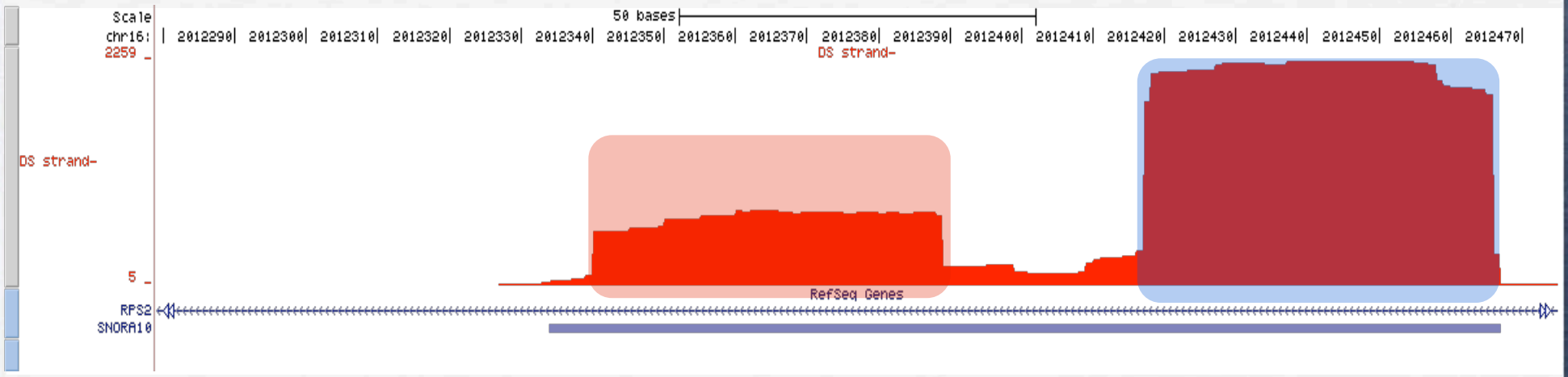
H/ACA box



- 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;

Statistical analysis: Differential Expression

Significant changes in the expression of genes are usually identified by using a **statistical Test** and the results are then corrected for **multiple testing**

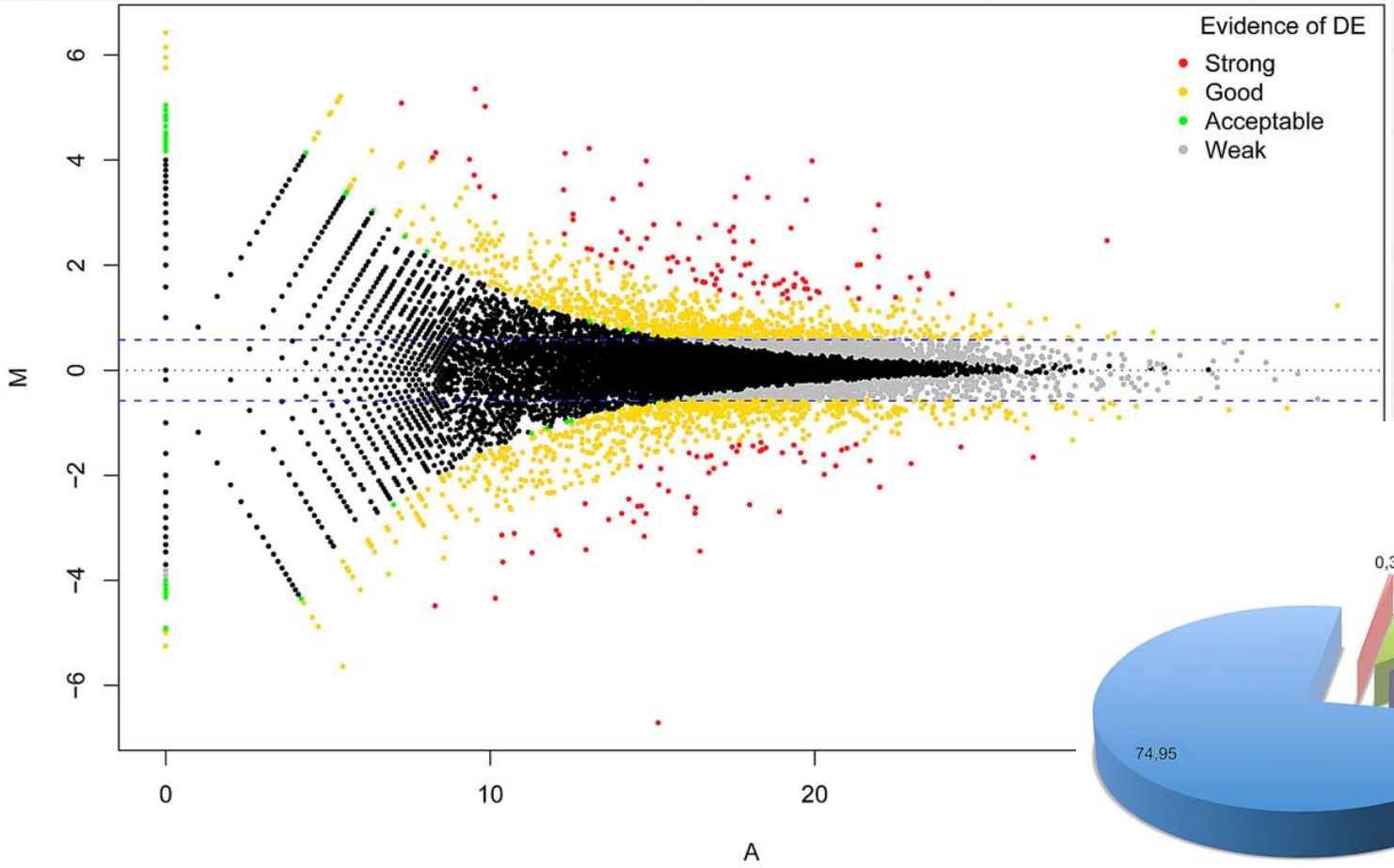
Unfortunately one cannot use ordinary tests developed for microarray since **RNA-Seq data** are **count data, and they are heteroscedastic (have no the same finite variance).**

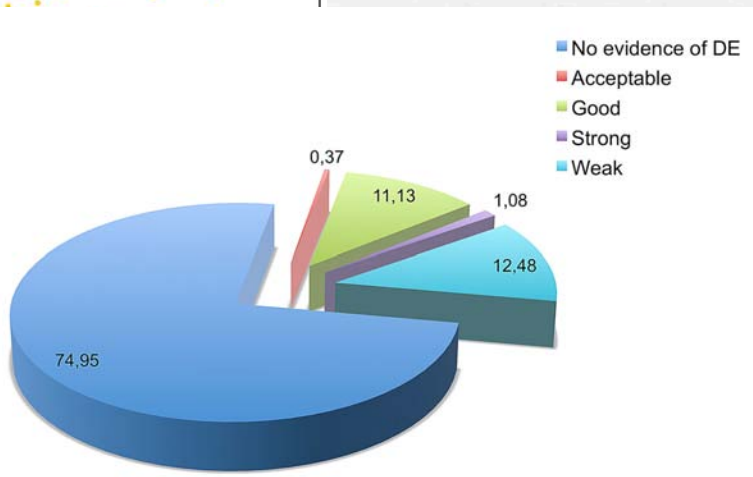
- Statistical significance has been inferred from total reads count for each RefSeq gene combining 3 tests:
DEGseq (based on Poisson distribution)
DESeq and edgeR (based on negative binomial).

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

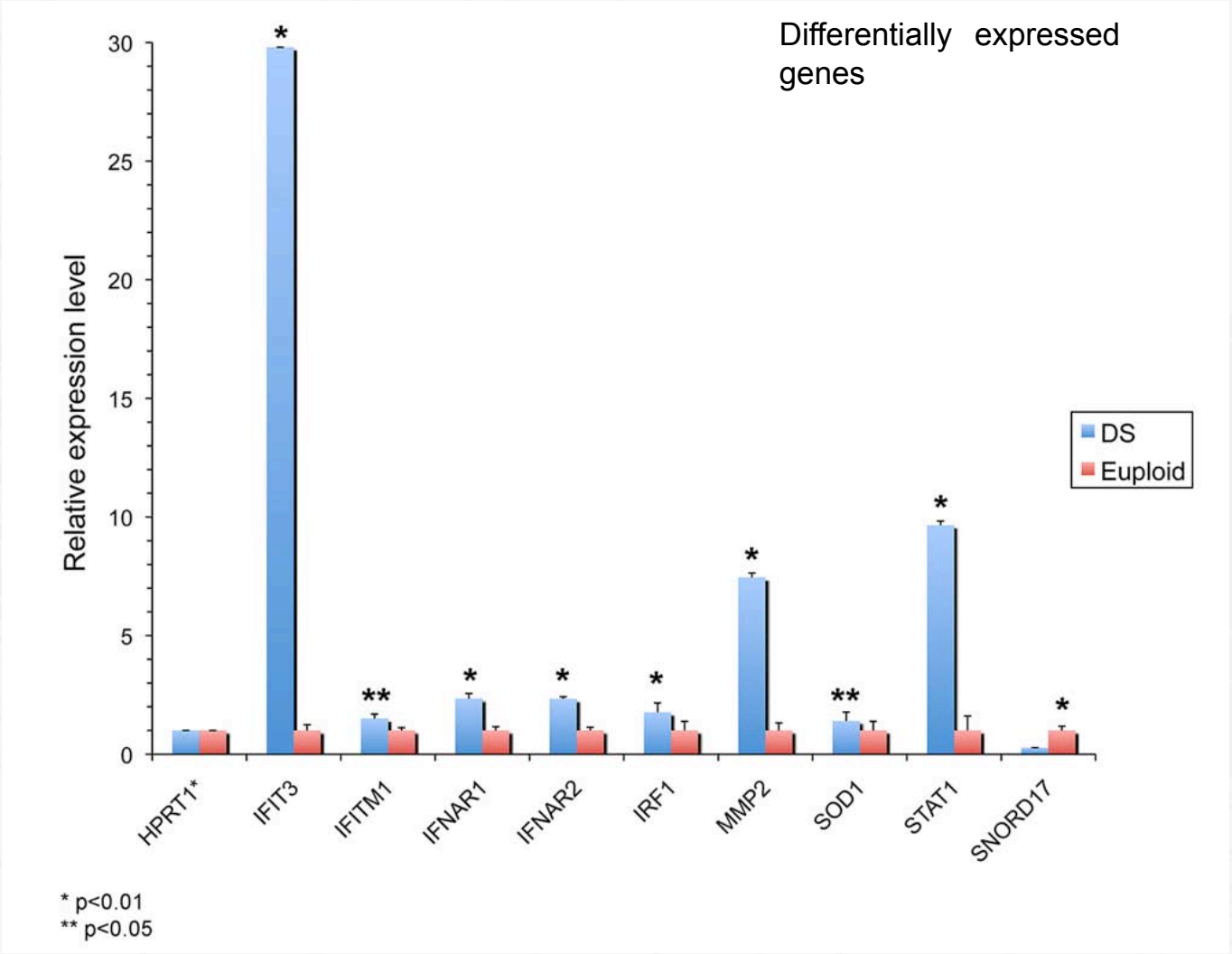
Differential expression of RefSeq genes

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

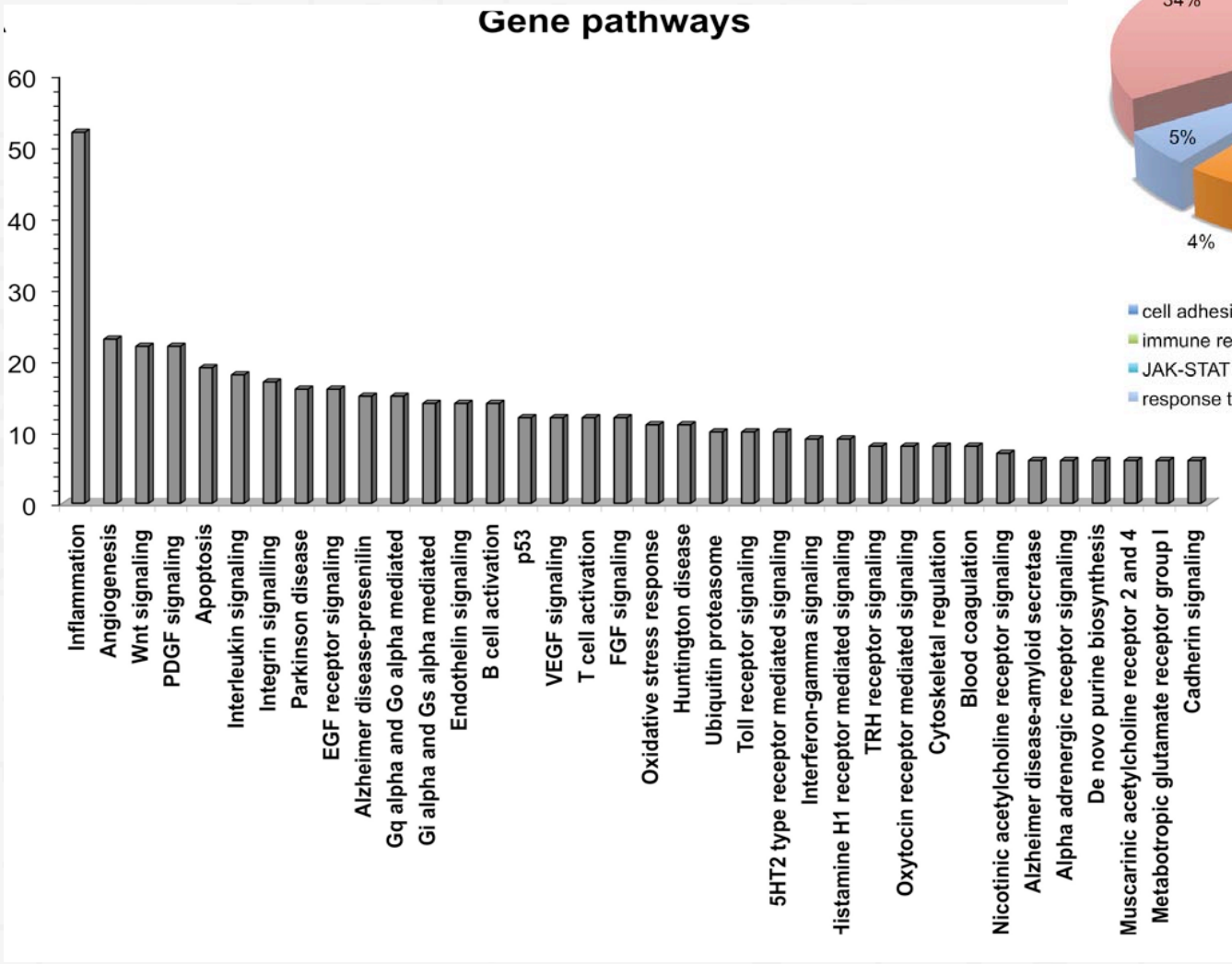


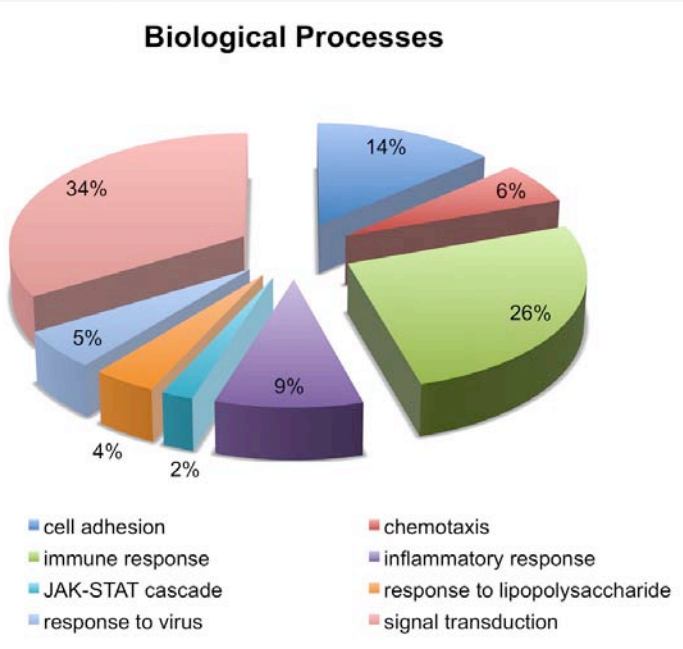


quantitative RT validation

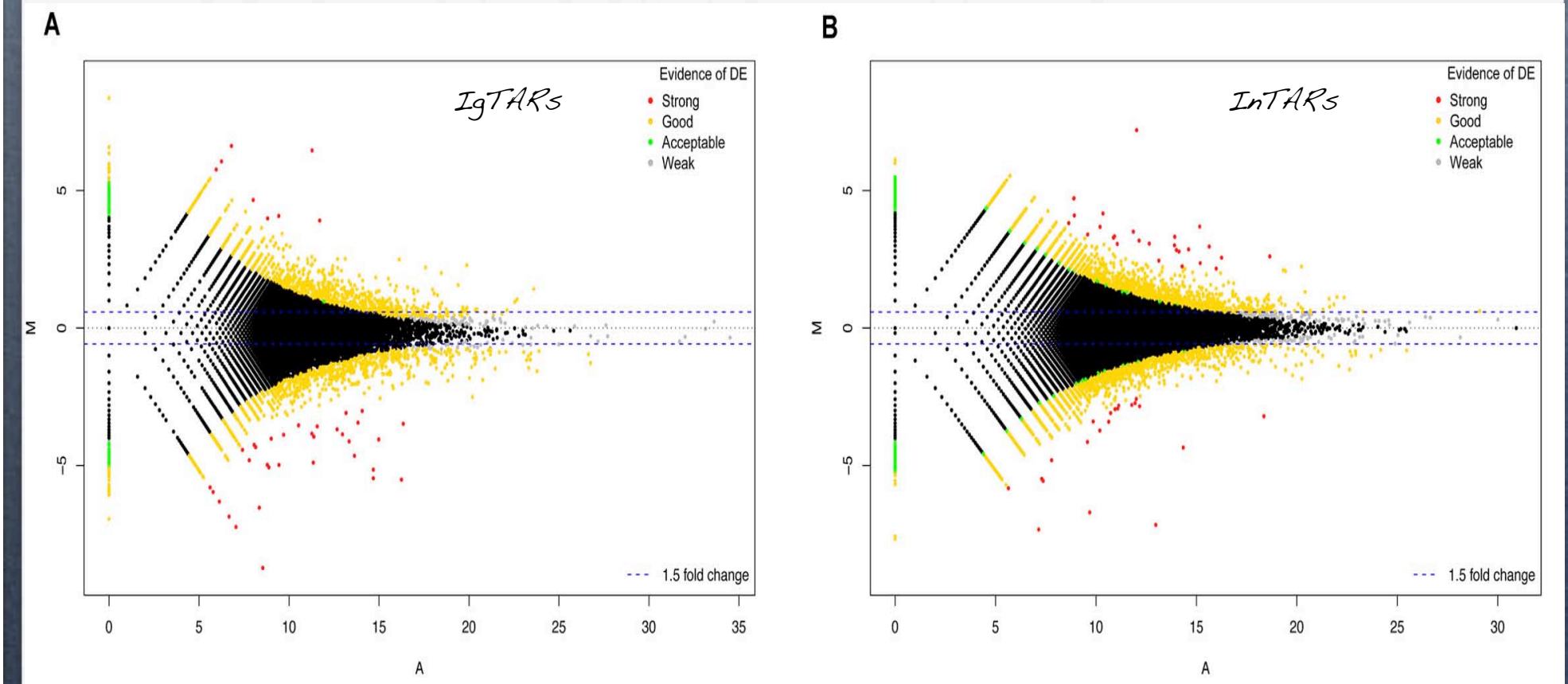


Differentially expressed genes



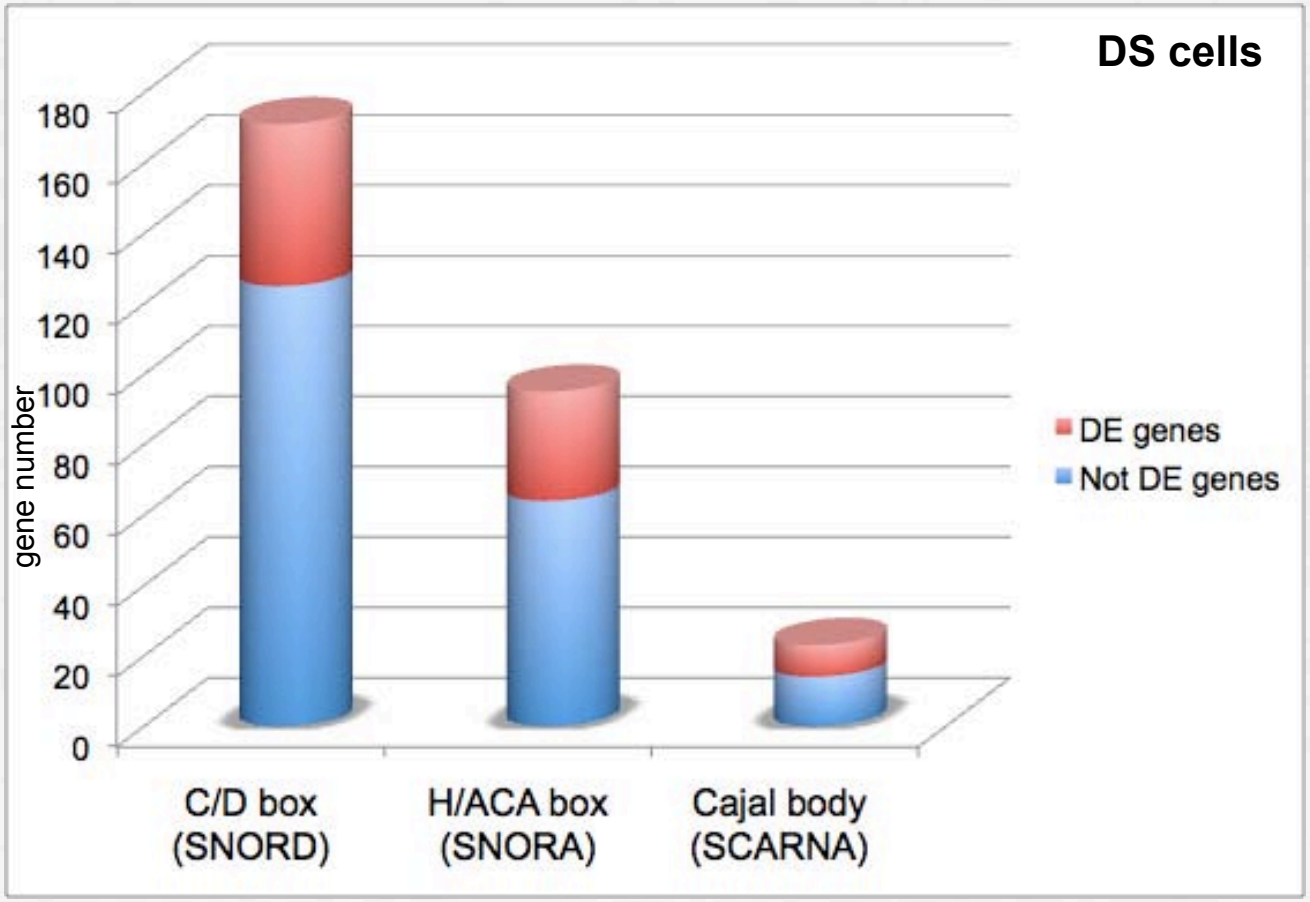


Differential expression of extra-genic regions



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;

→ **Both polyA+ and rRNA depleted samples**

- 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;

→ **Only rRNA depleted samples**

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.

Future perspectives (2/2)

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.

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

- Athough 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

IGB

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