RNA-seq analysis pipeline

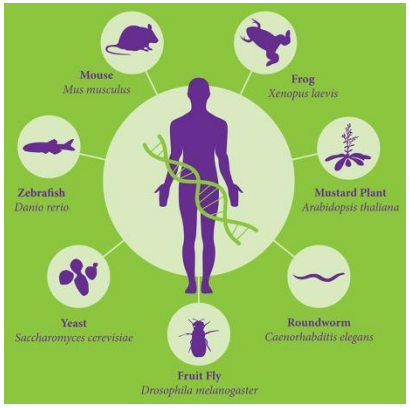
«EASY» PATH

Sequencing

↓

Mapping on a annotated reference genome

Can be applied to model organisms



«HARD» PATH

Sequencing

↓

De novo assembly to create a reference transcriptome

↓

Several quality checks and filtering steps

↓

Back-mapping on the reference transcriptome

Has to be applied to non-model species

RNA-seq: not as straightforward as it seems

First we need to choose **THE STRATEGY** best suited to our aims and to understand **HOW MUCH** sequence data do we need.

Single-end or paired end sequencing? How many lanes do we need?

In any RNA-seq experiment there are a series of biases which we need to take into account, detect and/or correct

- *De novo* assembly limitations
- Alternative splicing management
- Inter-individual sequence variability
- Calculating gene expression data
- Etc. Etc. Etc...

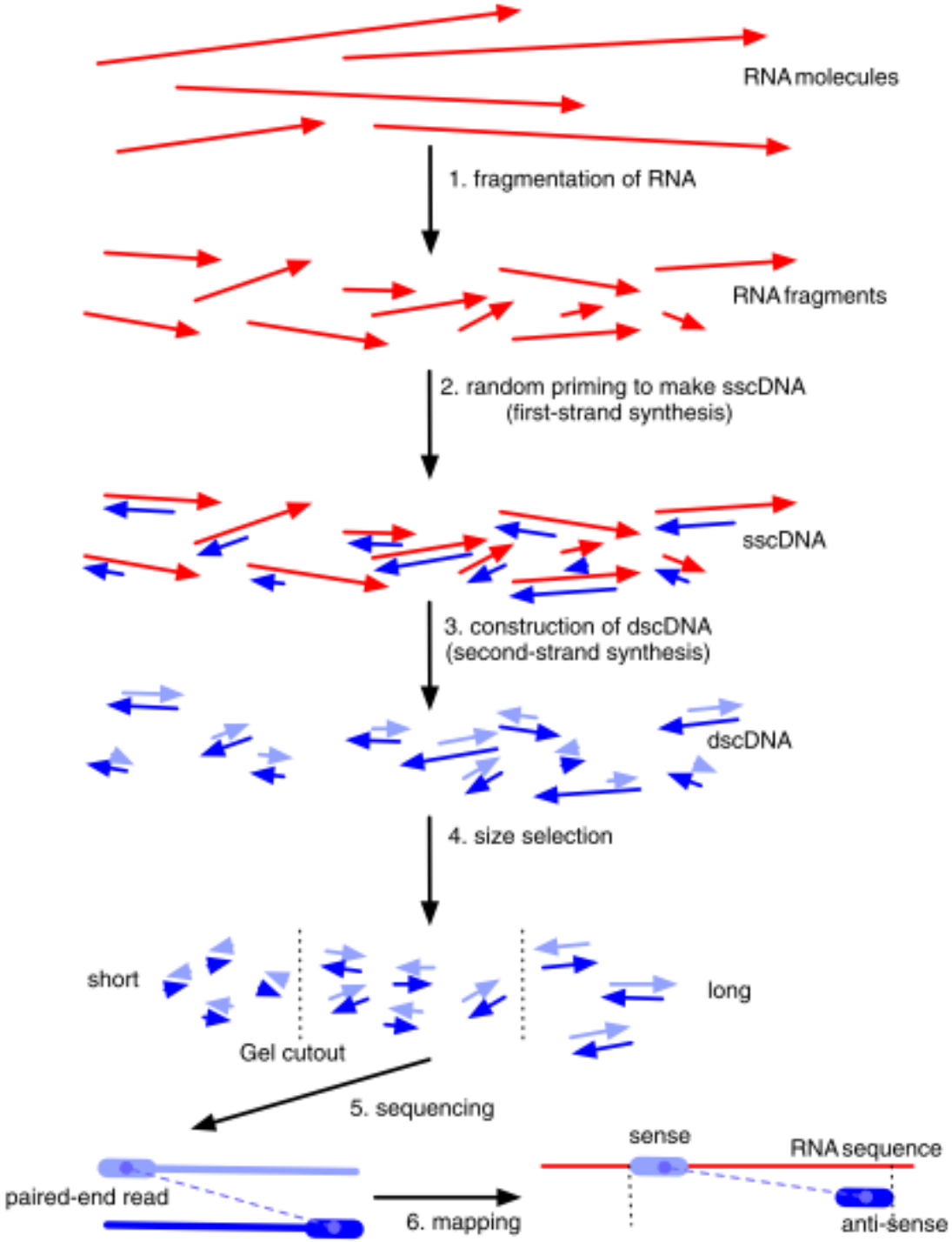
Experimental design

- **Single-end sequencing** with short reads (1x50) is fine when you already have a reference genome or transcriptome available
- **Paired-end sequencing** with longer reads (2x100) is advisable when you need to build your own reference transcriptome

How deep is enough? It depends on what you are looking for... As a rule of thumb 30M reads is usually enough for gathering information concerning all importantly regulated genes, but you may loose detection power for poorly expressed genes with a lower coverage.

The deeper the better, but **it is BETTER to sequence more replicates** than to go ultra-deep in a single sample. **No need for technical replicates!**

How many replicates should I use? The more you can, the better it is, but money is always an issue. In order to save sequencing costs, you can choose to pool replicates before sequencing. This will reduce statistical power and mask inter-individual variability, but it is a reasonable compromise between keeping sequencing costs down and getting reliable results.


RNA molecules
1. fragmentation of RNA
RNA fragments
2. random priming to make sscDNA
(first-strand synthesis)
sscDNA
3. construction of dscDNA
(second-strand synthesis)
dscDNA
4. size selection
short
long
Gel cutout
5. sequencing
sense
RNA sequence
paired-end read
6. mapping
anti-sense

Challenges in *de novo* transcriptome assembly

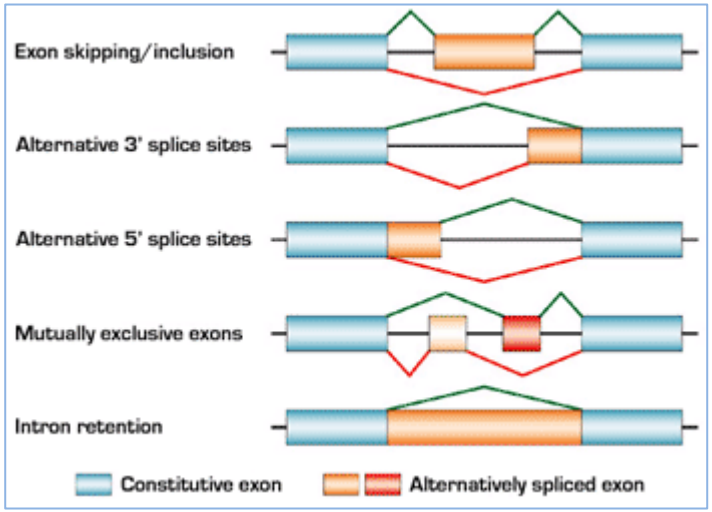
- ✓ Assemby size much smaller than those of genomes: is it easier?
- ✓ Transcriptome assembly presents challenges of its own
- ✓ Non-uniform coverage of transcripts (some mRNAs are expressed much more than others)
- ✓ Alternative splicings, paralog genes, repetitive sequences, ovelapping genes, antisense transcription, etc.

SOAP

Trinity (Inchworm, Chrysalis, Butterfly)

CLC bio — a result of science

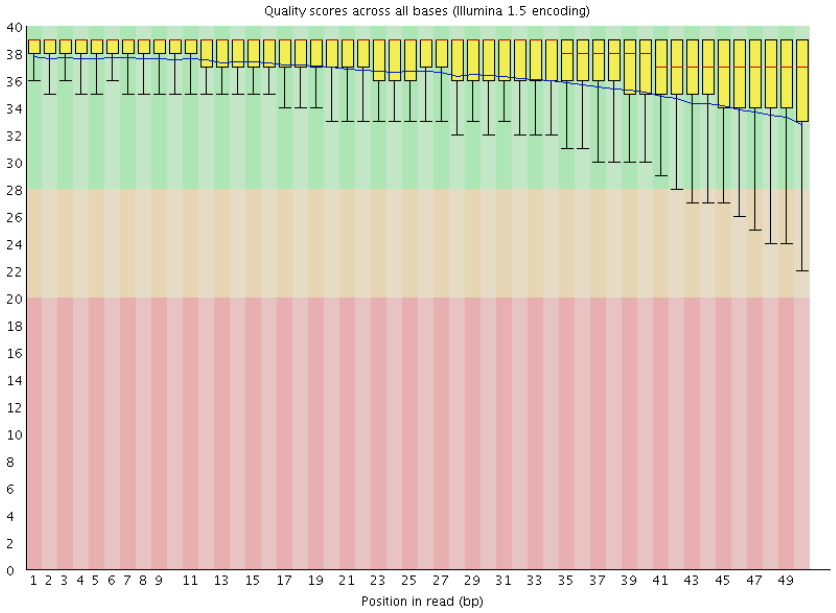
DNASTAR SeqMan NGen®

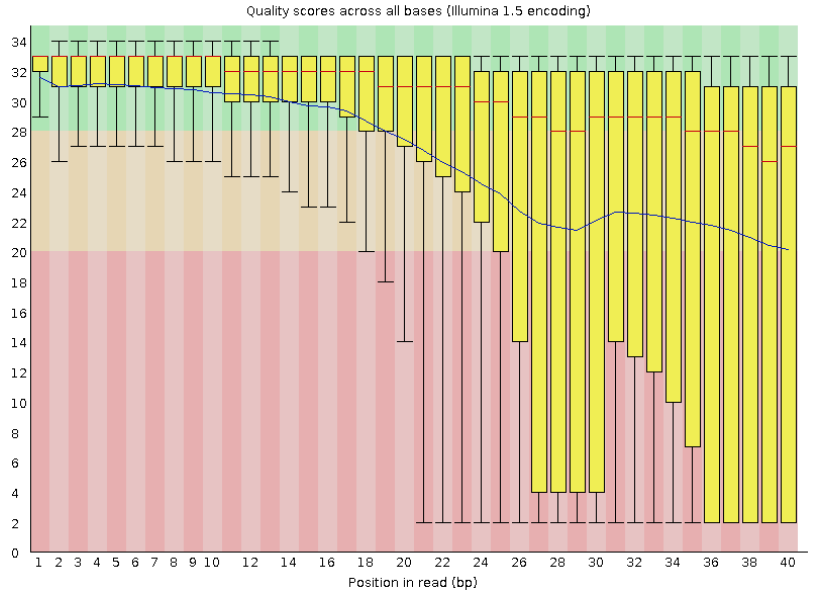


Not all reads are perfect!

- ✓ We need to trim data based on quality scores
- ✓ We need to remove Illumina adaptors and ambiguous nucleotides
- ✓ We need to discard reads too short to be informative

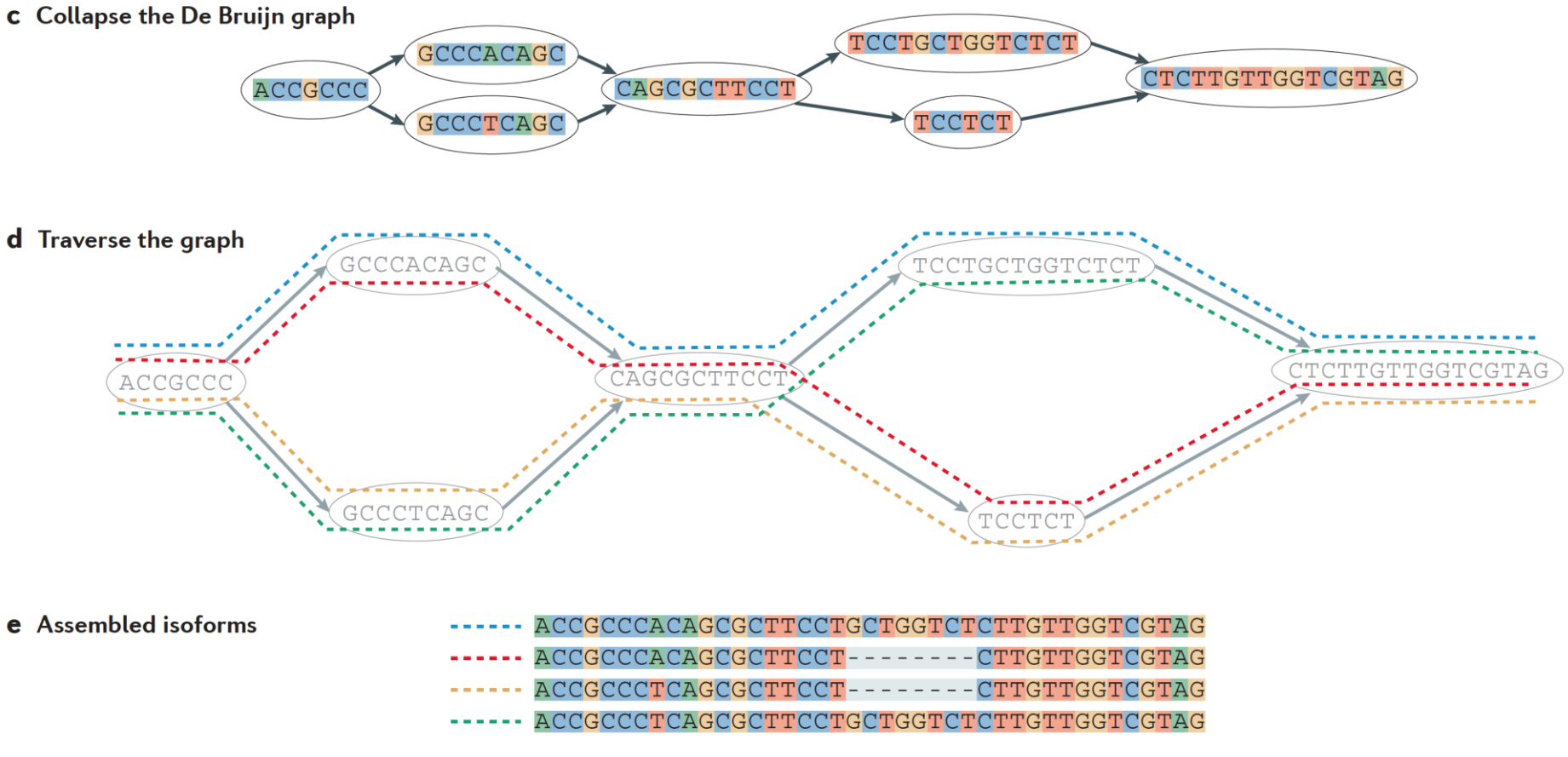
This process is known as TRIMMING

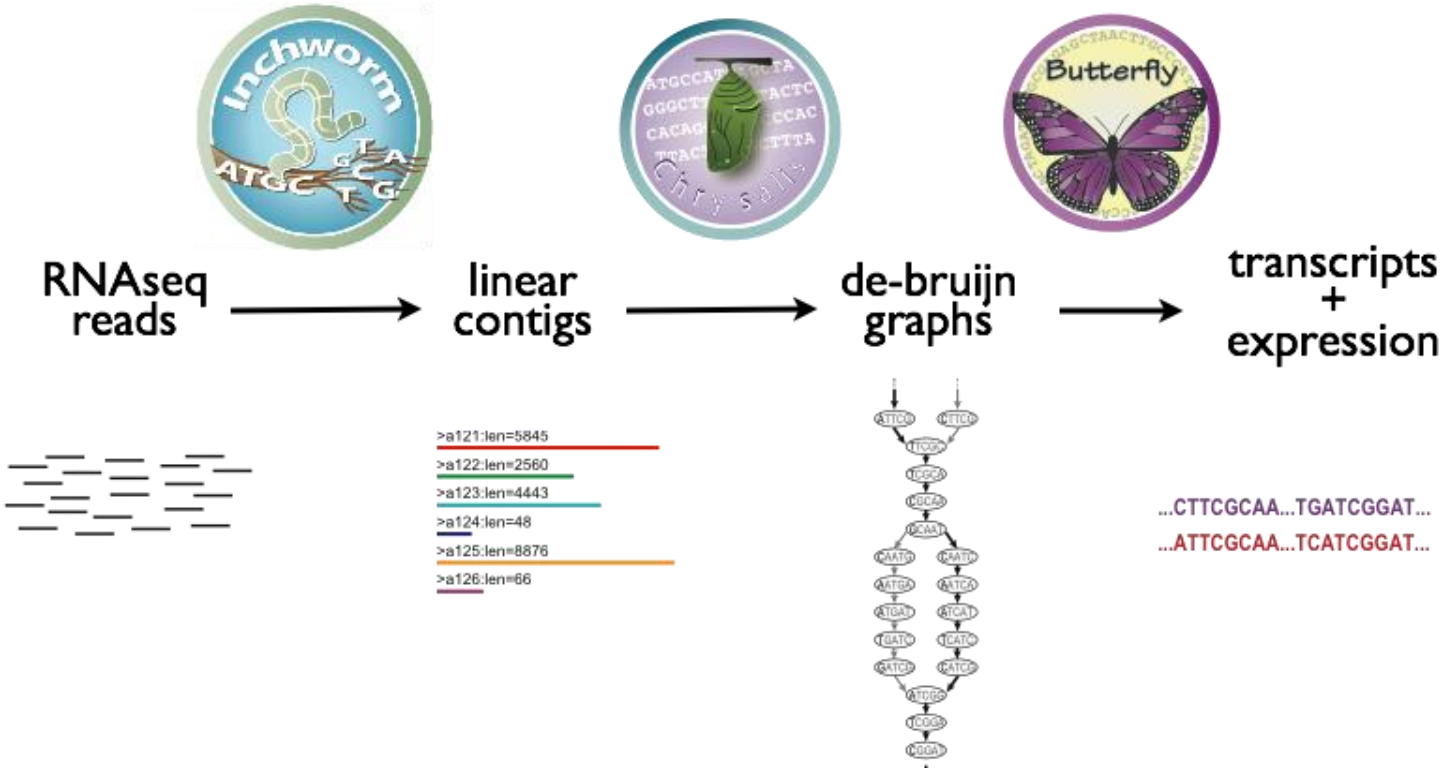




good qualiy read VS bad quality read

De novo assembly (De Bruijn graph construction)



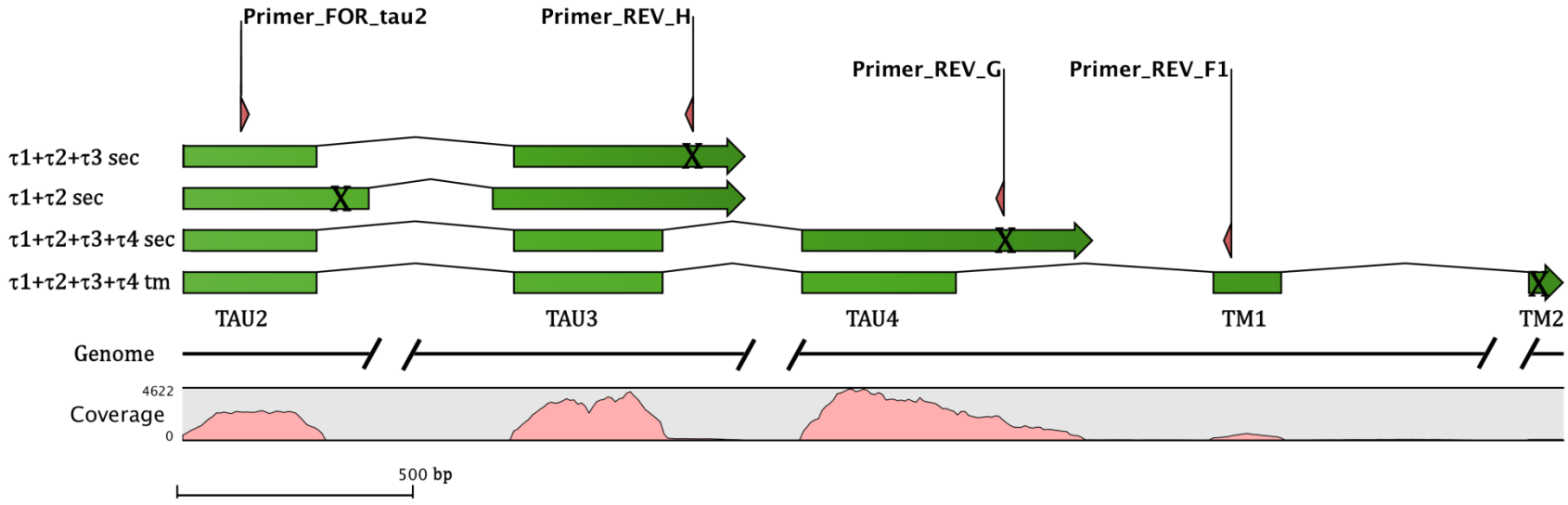


This is the **Trinity** pipeline: highly popular de novo assembler, its output comprises genes and putative alternatively spliced variants

Dealing with alternatively spliced variants is potentially interesting, but you have **high sequence redundancy.** Sometimes it is useful to filter out contigs by only keeping one transcript per gene (usually the longest one).

Other softwares for *de novo* transcriptome assembly aim to reuce the complexity of a transcriptome by only providing one representative transcript (the most supported by sequence data) per gene

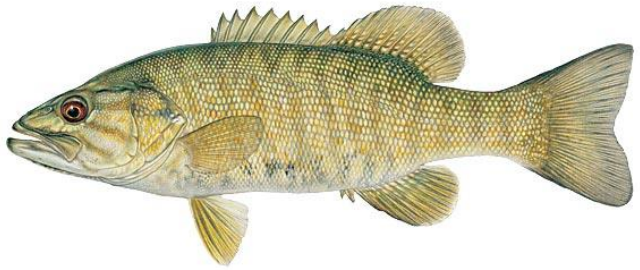
How do we deal with alternative splicing?



- ✓ They can be a problem for the calculation of gene expression
- ✓ Are all these splicing variants real or are they artifacts?
- ✓ Usually, you do not really need all of them to get the general picture of a transcriptomic response
- ✓ One representative transcript per gene is usually ine for this aim
- ✓ In any case remember that alternatively spliced isoforms may have very important functional differences!

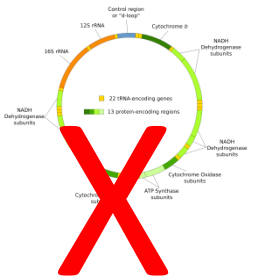
RNA-seq of non-model species: contamination

Let’s suppose to plan a RNA-seq experiment on a fish species...

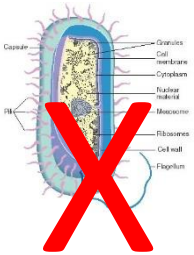


De novo assembled and annotated reference transcriptome

Residual rRNA

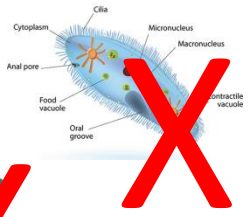


Mitochondrial RNA



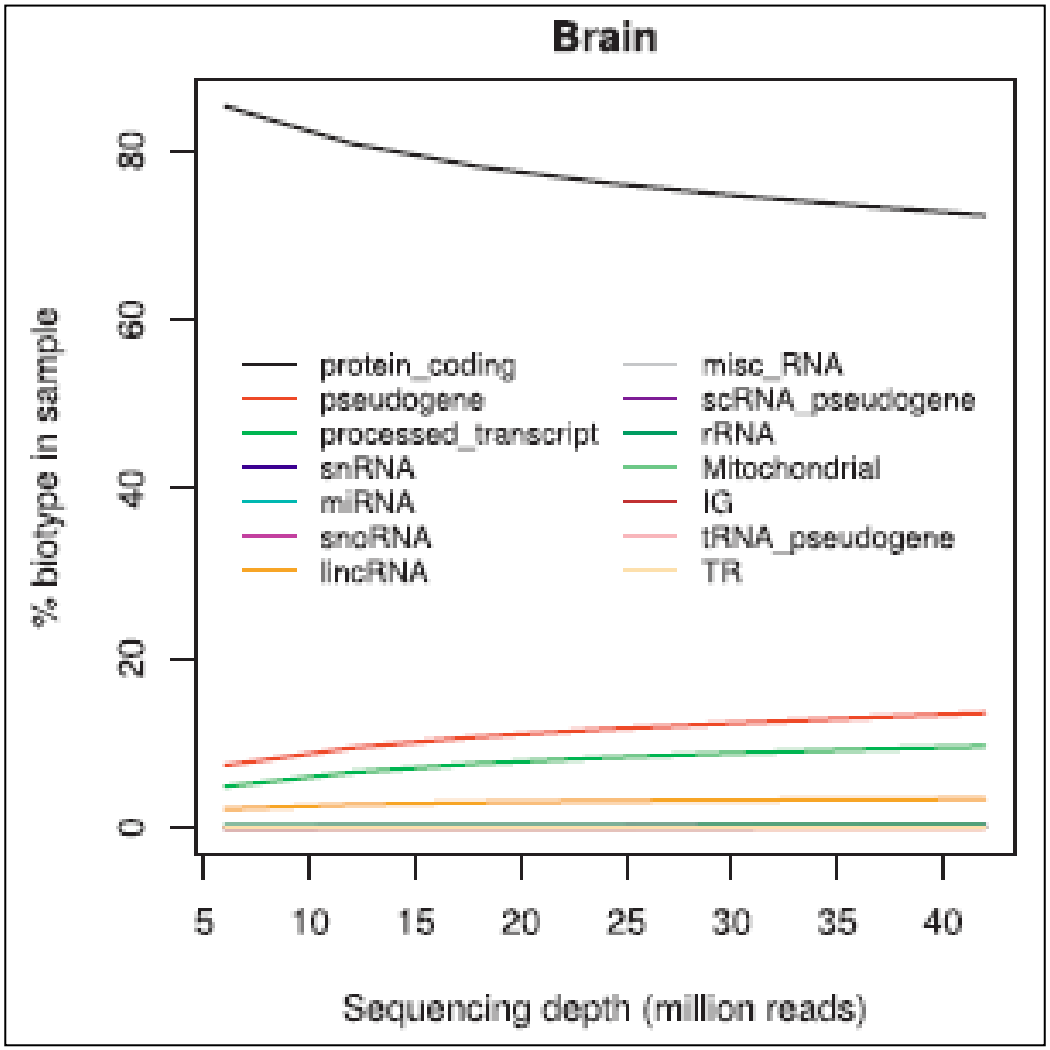
bacteria

virus



protozoa

microalgae



Tarazona et al (2011) Genome research, 21(12), 2213–23

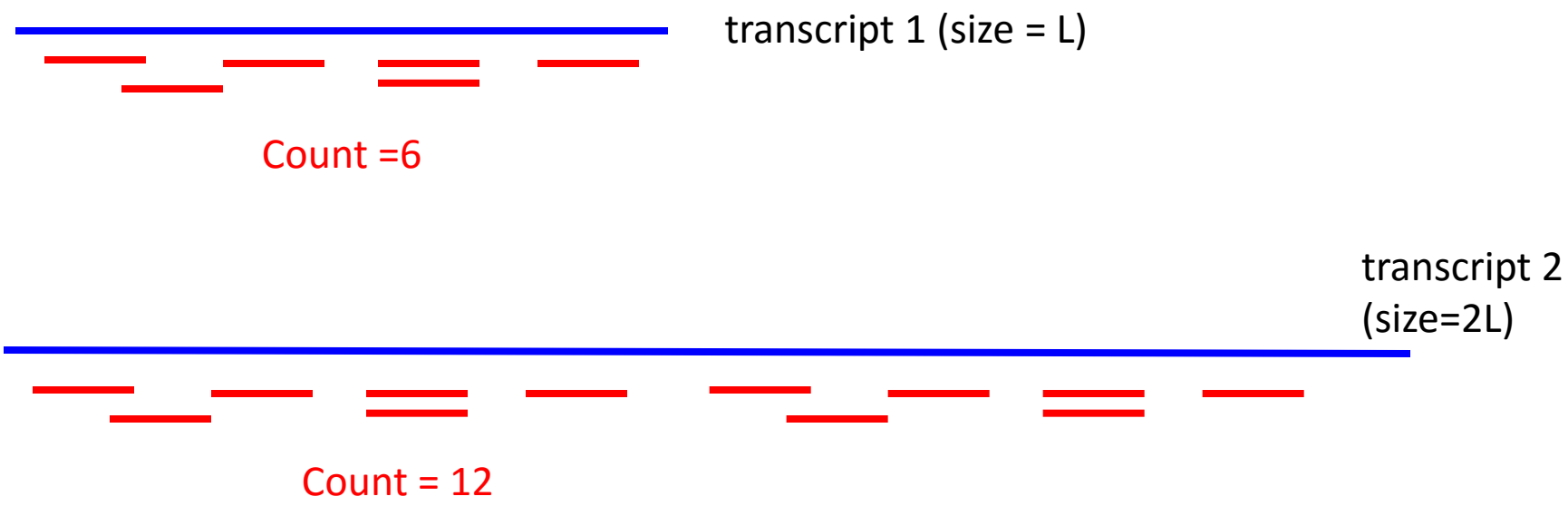
- ✓ In any RNA-seq you will get A LOT of assembled contigs with no apparent function
- ✓ Are they contaminants, lncRNAs or what?
- ✓ Importance to adopt a **pre-filtering step based on sequencing coverage**

Anything with a sequencing coverage lower than a certain threshold can be considered as NOT USEFUL for downstream analysis

=

Not necessarily a contaminant, but in most cases you will remove just small fragments with no annotation

Calculating gene expression: not just read counting!



Question: Which transcript is the most expressed one?

Answer: They are equally expressed: the number of reads mapped are directly proportional both to gene expression level and to the transcript length.

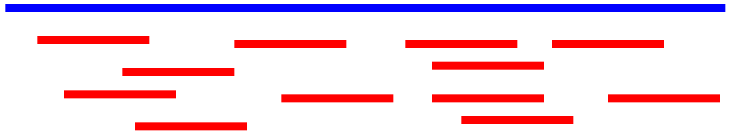
BUT

This just applies to a comparison between genes within the same sample

What about a comparison between two samples?

transcript 1 (sample 1)

Count =6, **library size = 600**



transcript 1 (sample 2)

Count =12, **library size = 1200**

Question: is transcript 1 more expressed in sample 1 or in sample 2?

Answer: transcript 1 is equally expressed in both samples, since read count is also directly proportional to sequencing depth!

Normalizing RNA-seq gene expression data

- ✓ Keep in mind these three factors: **gene length**, **read count** and **sequencing depth**
- ✓ Also keep in mind that we might want to be able to make comparisons **within a sample** and **between samples**
- ✓ Various methods have been developed for normalizing RNA-seq gene expression data over the years

- ✓ **Quantile** normalization (developed for microarrays)
- ✓ **Upper quartile** (based on read counts only)
- ✓ **Normalization "by totals"** (based on read counts only)
- ✓ **RPKM/FPKM** (all three factors are taken into account)
- ✓ **TPM** (evolution of the RPKM concept)

Normalizing RNA-seq gene expression data

$$RPKM = \frac{number\ of\ reads\ of\ the\ region}{\frac{total\ reads}{1{,}000{,}000} \ x \ \frac{region\ length}{1{,}000}}$$

- ✓ Proposed by Mortazavi et al (2008) Nature Methods, 5(7), 621
- ✓ Highly popular method, as it considers both sequence length and sequencing depth

- ✓ Permits comparison of expression **WITHIN** samples (i.e. to compare the expression of gene A and gene B in the same tissue or experimental condition) and, allegedly, **BETWEEN** samples (i.e. to compare the expression of gene A in two different samples)

The second assumption is not correct!!!

Normalizing RNA-seq gene expression data

- ✓ **TPM = Transcripts Per Million**
- ✓ Introduced by Wagner et al. 2012
- ✓ Measure relative to the molar concentration of each mRNA species

- ✓ RPKM of all transcripts in a sample are summed up
- ✓ They are normalized so that the sum of the expression values of all transcripts = 1 million

- ✓ **Allows a reliable comparison between samples, regardless to the sequencing depth**

Short Communication
Theory in Biosciences
December 2012, Volume 131, Issue 4, pp 281-285

First online: 08 August 2012

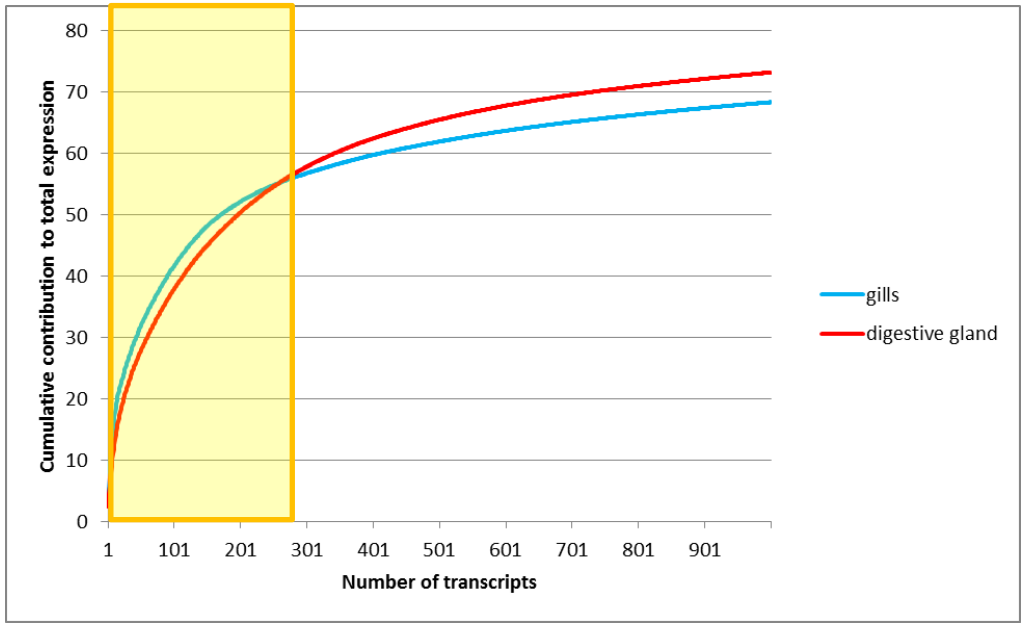
Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner ✉ , Koryu Kin, Vincent J. Lynch

RNA-seq can only provide changes in relative quantities

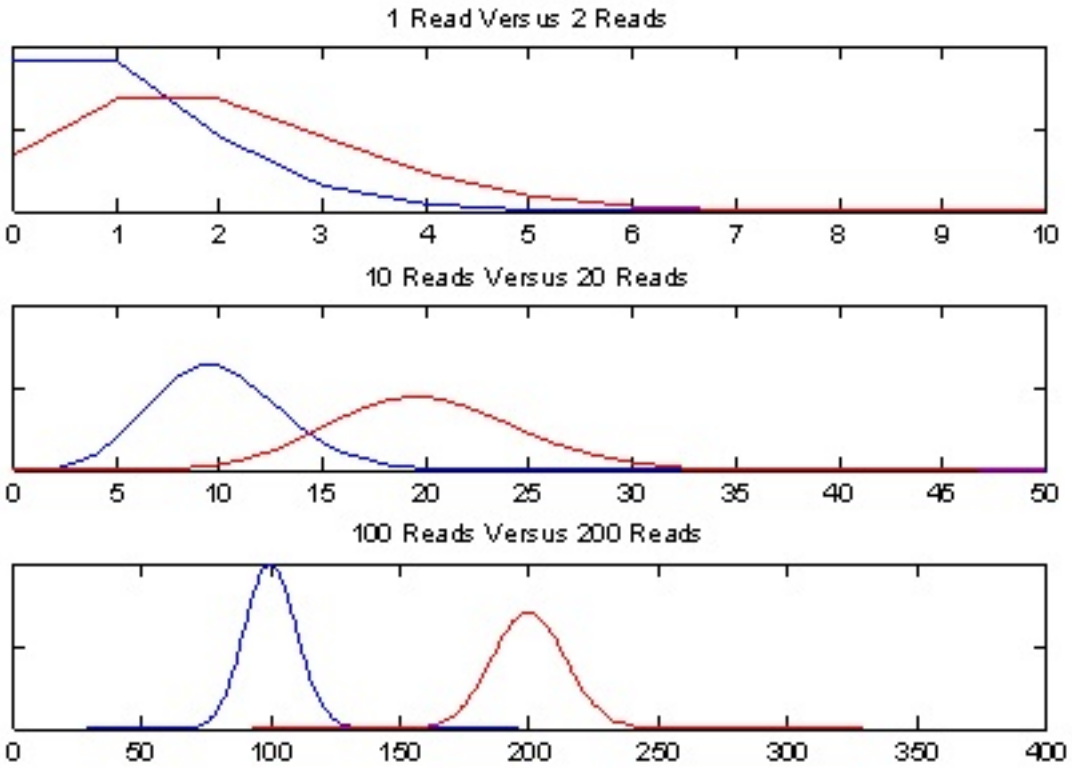
- ✓ Since RNA sequencing works by random sampling, a small fraction of highly expressed genes may consume the majority of reads
- ✓ TPM, like any other measure of gene expression, assumes that the global amount of mRNA is the same in all cells, which is obviously not guaranteed in reality!
- ✓ Example: when the transcription of ALL transcripts equally drops by 50% no changes in gene expression can be detected by RNA-seq

Ribosomal proteins or transcripts with important tissue specific functions



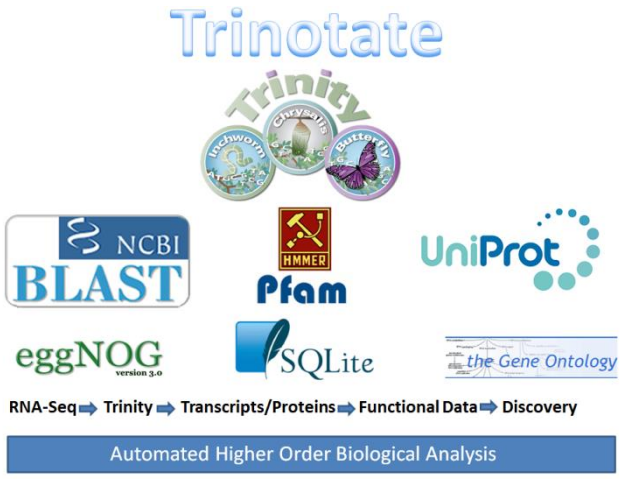
How is differential expression calculated?

- ✓ Based on gene expression values (normalized), assuming a Poisson distribution
- ✓ Based on **statistical significance testing**
- ✓ **False Discovery Rate** or **Bonferroni** corrected p-values
- ✓ This takes into account the fact that changes in genes covered by a low number of counts are not significant
- ✓ Each transcript will have a **p-value** and a **Fold Change**
- ✓ We shoould establish **thresholds** to detect differential expression



How to interpret data?

- We need to **annotate** sequences based on BLAST similarity to a target database
- Automated pipelines: Blast2GO and Trinotate
- BLAST vs UniprotKB and Gene Ontology assignment
- Search for conserved protein domains with HMMER (PFAM or Interpro databases
- **Hypergeometric tests** or **Gene Set Enrichment** analyses to detect processes, functions and domains significantly over or under-represented in a subset of differentially expressed genes



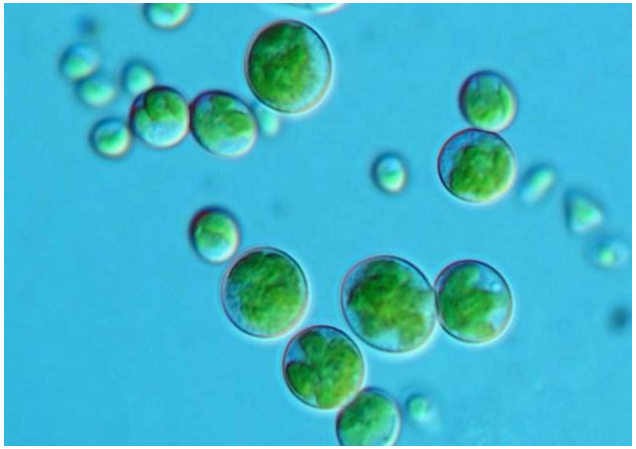
Why do my annotation statistics look so bad?

- ✓ **Keep in mind that lncRNA and non-coding transcripts in general may have important functions**
- ✓ **Coding transcripts without an annotation are NOT garbage! They may represent innovations and they may cover VERY IMPORTANT functions**
- ✓ **Even if you can't identify their function they might be used as molecular markers of a certain condition**

More than 50% of the assembled transcripts in a non model species are usually not coding

About 40% of the protein-coding transcripts of a non-model invertebrate have an unknwon function and lack any kind of annotation

A case study: *Trebouxia gelatinosa*



- ✓ Lichens are extremely resistant to dehydration
- ✓ They can live in a desiccated state for a very long time and quickly fully recover rehydration
- ✓ Lichens are the result of a symbiotic relationship between fungi (mycobiont) and unicellular algae or cyanobacteria (photobiont)
- ✓ *Trebouxia* spp. Is one of the algal species most commonly associated with lichens

Different organisms use different strategies to cope with dehydration (and rehydration)

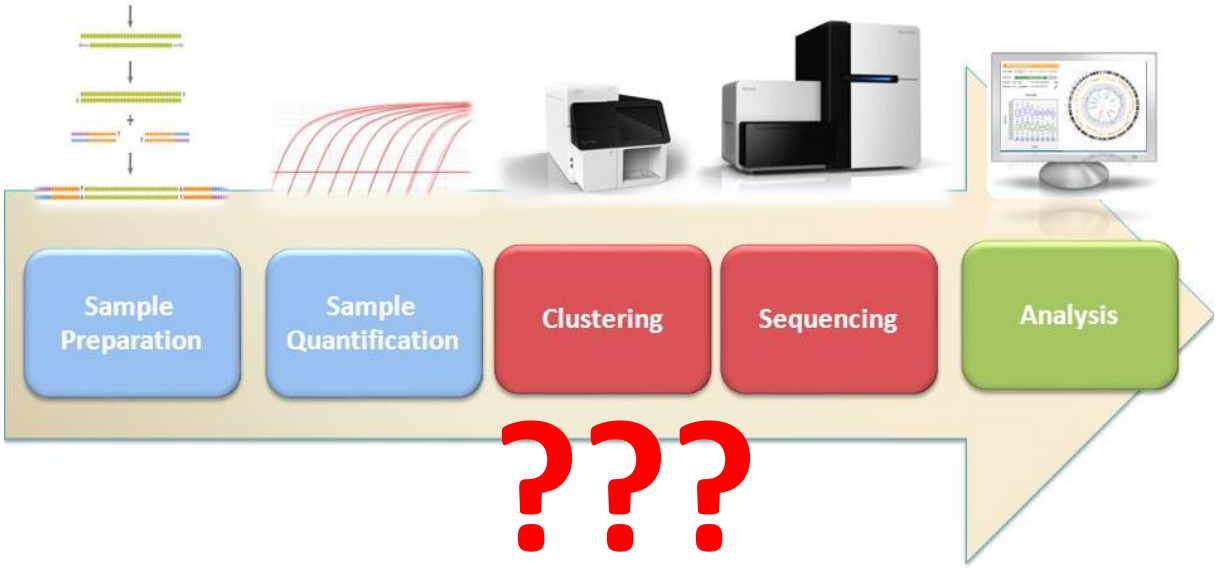
Before

After

- ✓ Resurrection plants
- ✓ Morphological, physiological and biochemical adaptations
- ✓ Can we trace back these changes to alteration of gene expression?
- ✓ Can RNA-seq reveal the secrets behind this remarkable tolerance?

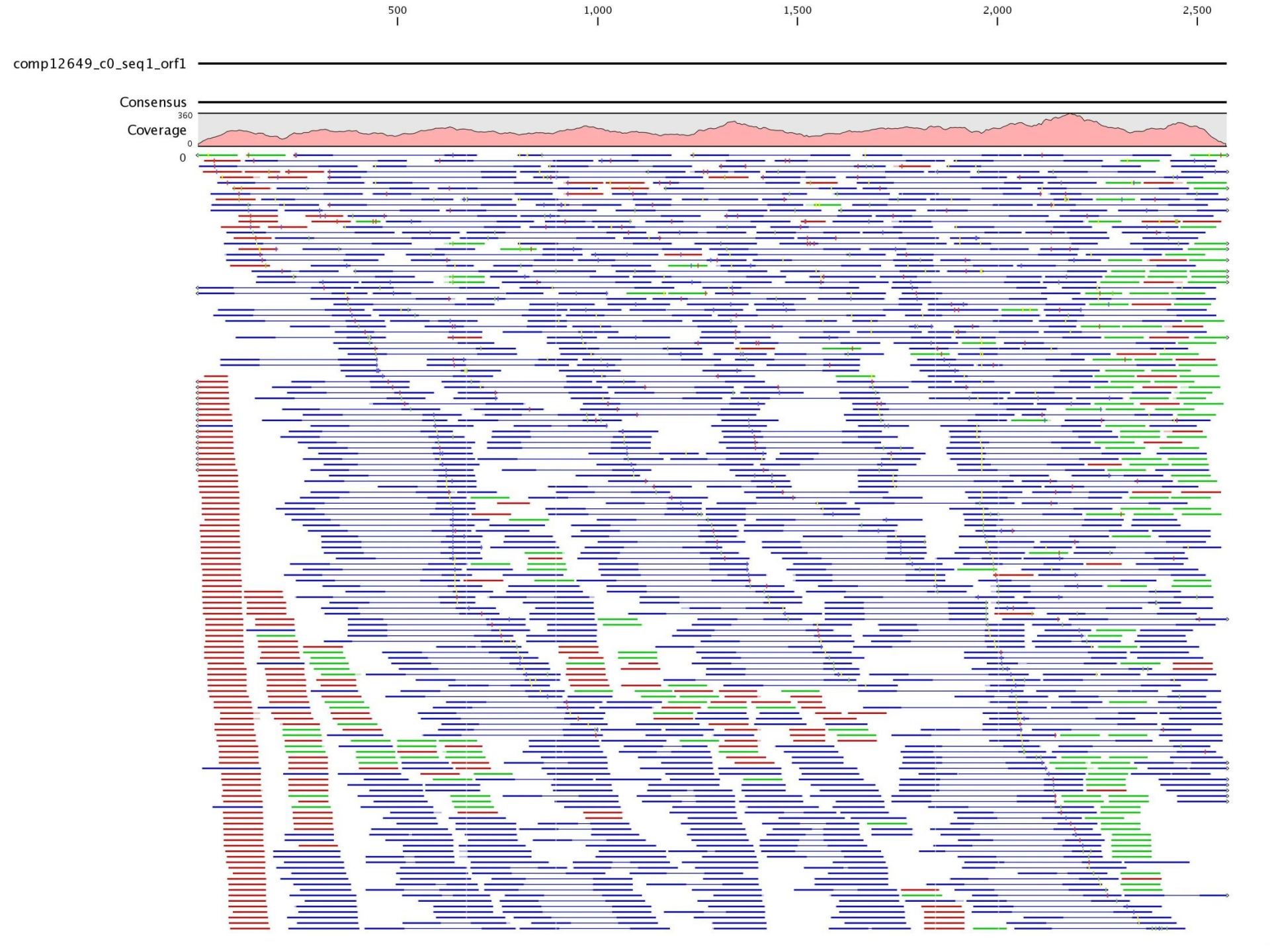
Experimental planning

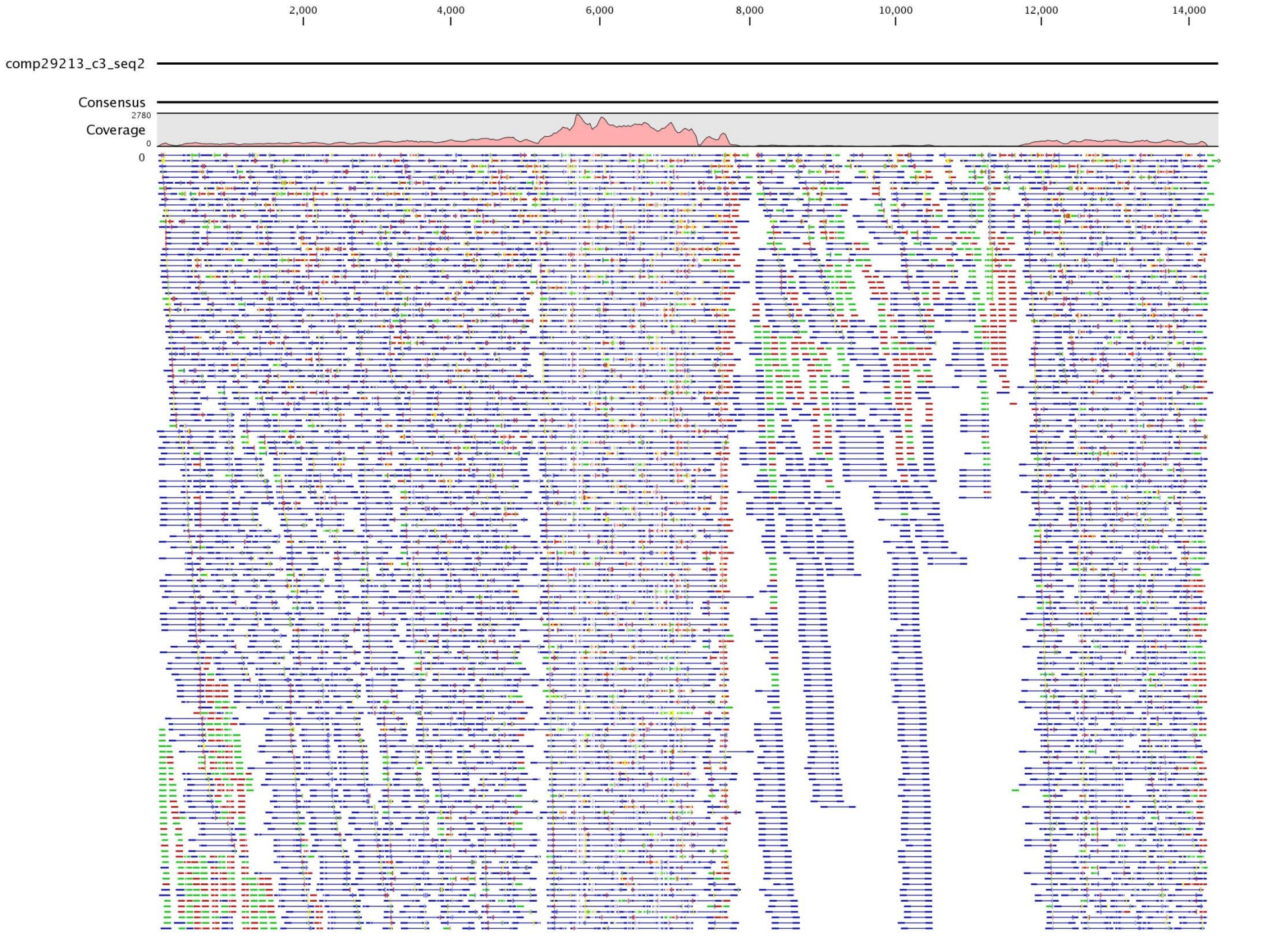
- ✓ Budget available: 1 lane with 3 libraries is fine
- ✓ **We need to build a reference transcriptome**: 2x100 paired-end sequencing is needed
- ✓ Pooling of 3 biological samples before sequencing
- ✓ Three experimental samples: fully hydrated control (C), dehydrated for 10 hours (D) and rehydrated for 12 hours (R)
- ✓ We expect to have **60-90 million reads per sample**
- ✓ **We plan to use Trinity and reduce transcriptome complexity by removing redundant contigs**



From RNA-seq to data interpretation

- We expect to observe adaptations somewhat similar to that of other dessication-tolerant organisms
- We expect to have some novel/alternative mechanisms which have not been described before
- Keep in mind that not all adaptations are expected to be reflected by changes in gene expression
- Most importantly, keep in mind that you need an expert advice to correctly interpret your result (if you're not an expert about the topic yourself)


500
1,000
1,500
2,000
2,500
comp12649_c0_seq1_orf1
Consensus
360
Coverage
0
0


2,000
4,000
6,000
8,000
10,000
12,000
14,000
comp29213_c3_seq2
Consensus
2780
Coverage
0
0


5,000
10,000
15,000
20,000
mRNA1
mRNA2
mRNA3
comp29563_c1_seq26
Consensus
529
Coverage
0
0
We can try to figure out where a transcript starts and where another one ends based on shifts in sequening coverage

Solution to the problem: using ORFs

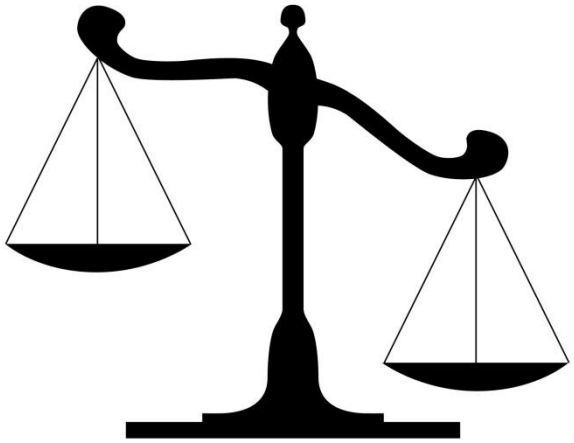
ADVANTAGES

- ✓ Each ORF corresponds to a single protein-coding transcript
- ✓ It's automated and unbiased (TransDecoder)
- ✓ Can be performed even if ORFs are overlapping
- ✓ No uncertainties about 5' and 3' UTR boundaries

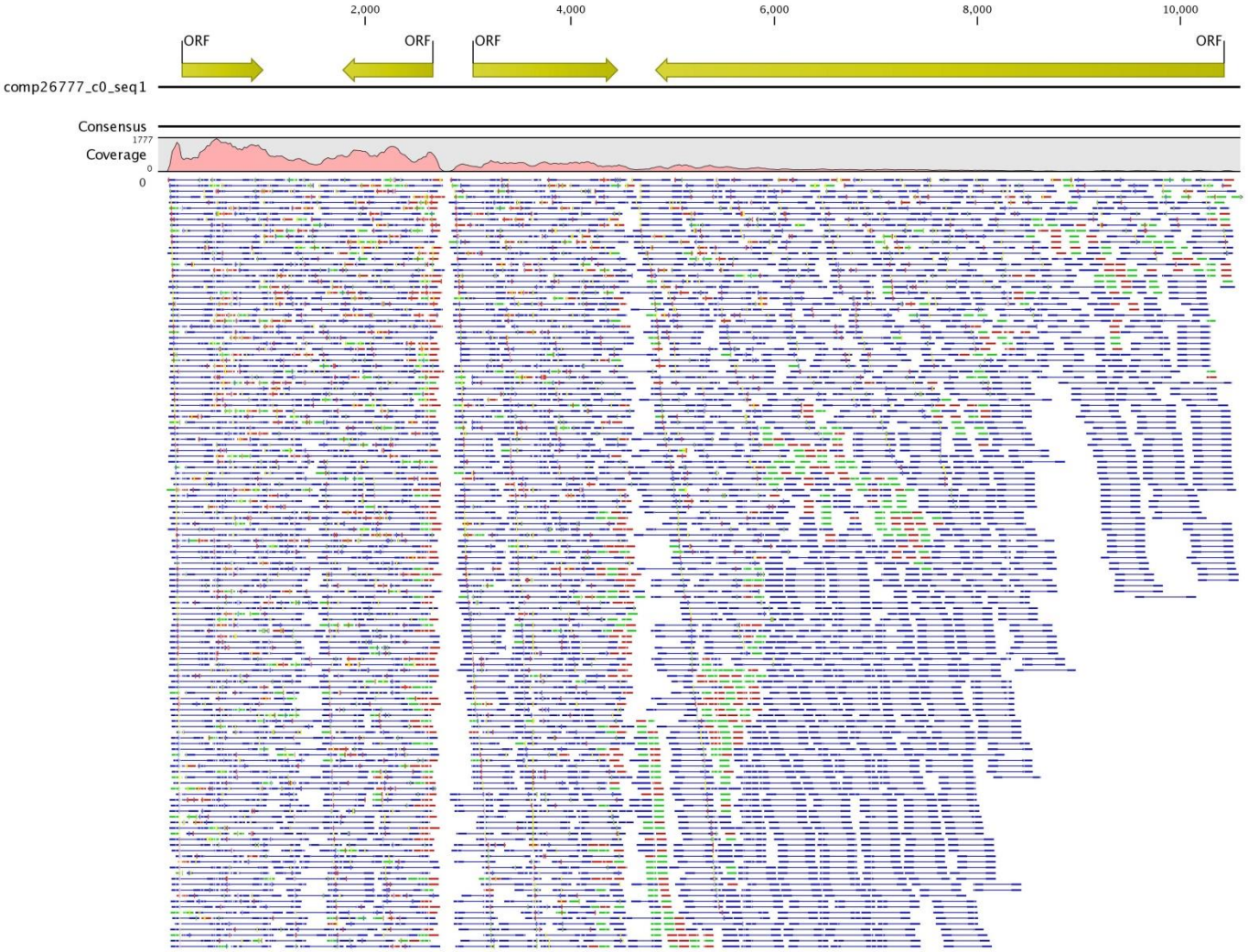
DISADVANTAGES

- ✓ The information about lncRNAs will be lost
- ✓ You have to set a minimum ORF length (i.e. Transcripts encoding very short proteins will be discarded)
- ✓ Some ORFs will be erroneously predicted as protein coding

Sometimes the only possible way to solve a problem is an heuristic compromise



In this case, the contig is a chimaera which likely comprises 4 partially overlapping mRNAs



However...

We need to futher filter CDS to avoid potential errors (i.e. ORFs «casually» present in the transcriptome, which are not protein coding)

This is done by:

1: **BLAST** vs a closely related proteome of a species with a fully sequenced genome (*Chlorella* spp, *Asterochloris* spp., etc.) -> positive HIT means high probability of coding potential

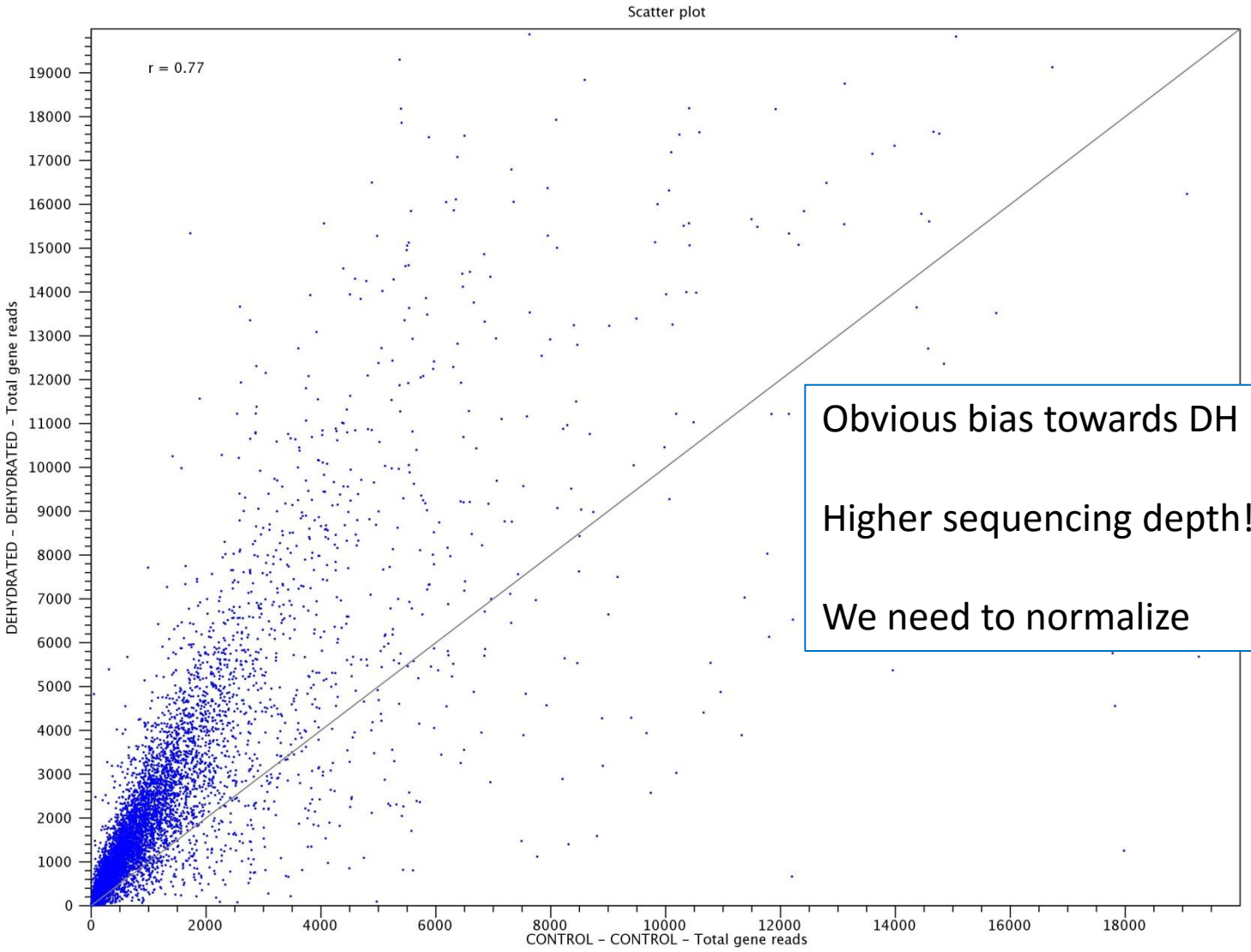
2. Presence of conserved protein domains (**PFAM or InterPro**) -> high probability of coding potential

3. What if you have highly divergent, species-specific genes? We can include long ORFs (> 300 codons, even if they don't meet the 2 above mentioned criteria

Final reference set: **13,648 ORFs**

- Once we have a good reference library, we can proceed to read mapping of each of the 3 samples to **calculate gene expression values**
- 0,75 length fraction and 0,95 similarity fraction **mapping parameters**: this means that at least 75% of a read needs to be 95% identical to the reference in order to be mapped
- Calculation of read counts are computed for each ORF
- We also need to annoate our transcriptome, and we will do that using the **Trinotate** pipeline
- We are now ready to compare and interpret data

Comparison of raw read counts

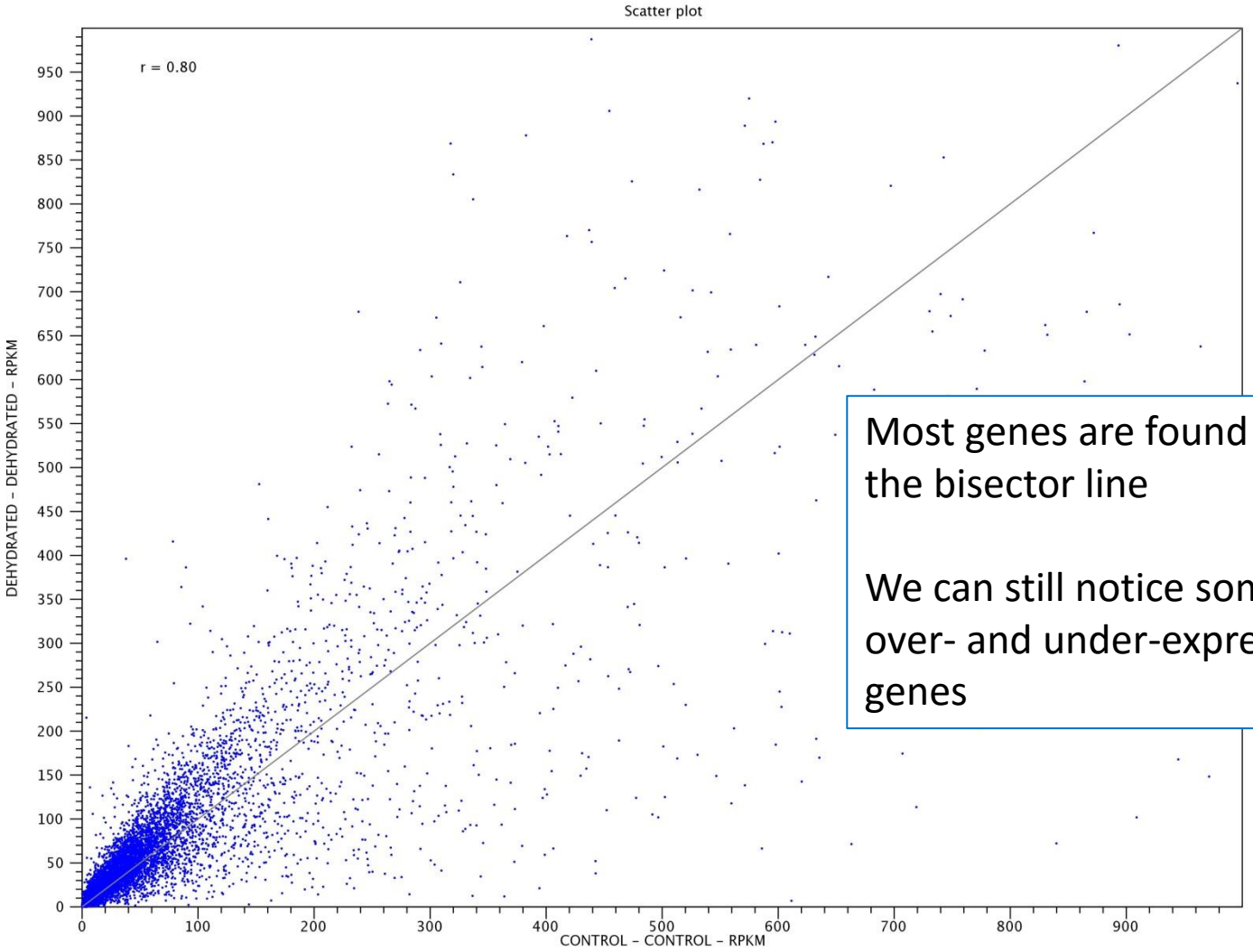


Obvious bias towards DH

Higher sequencing depth!!!

We need to normalize

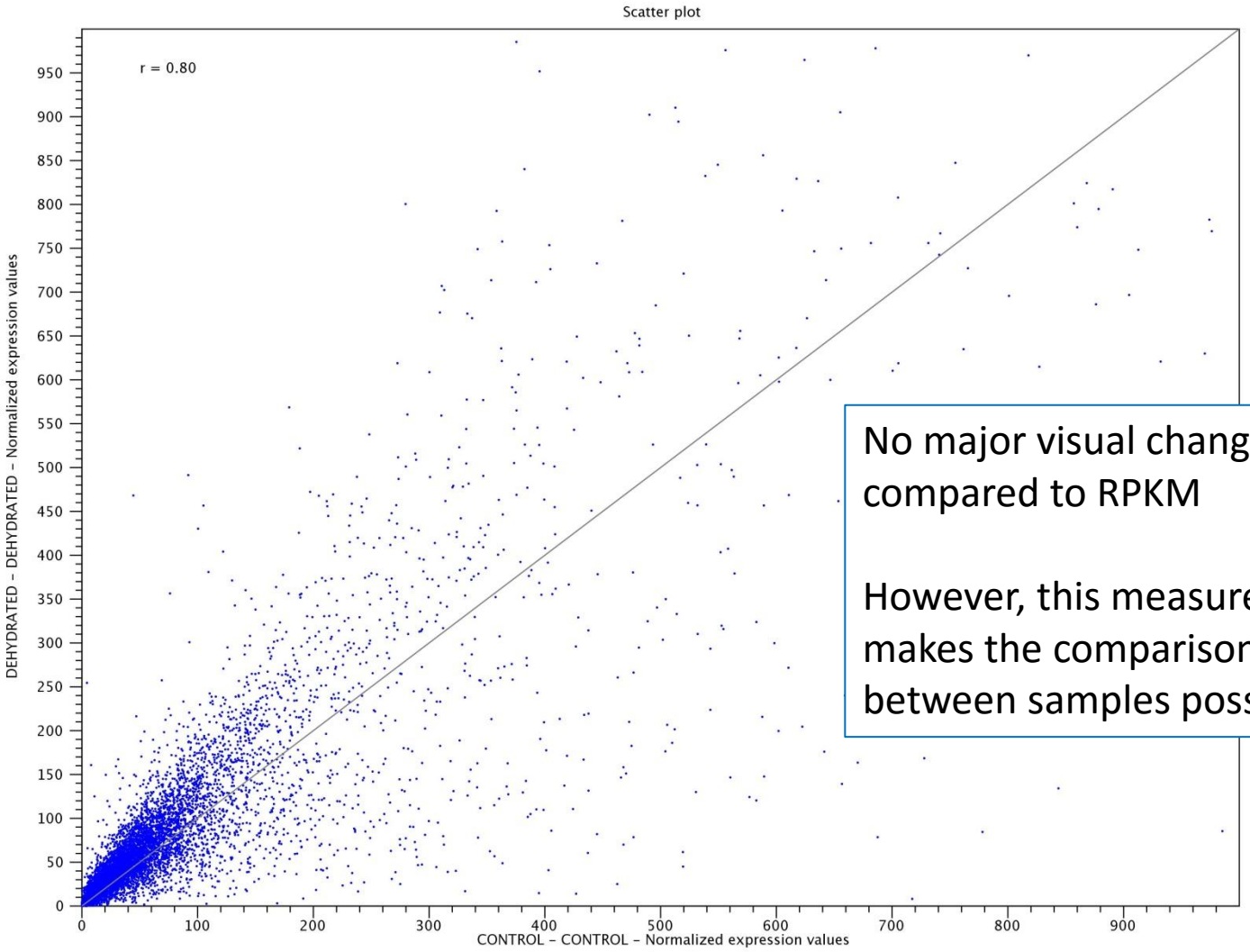
Comparison of RPKM values



Most genes are found along the bisector line

We can still notice some over- and under-expressed genes

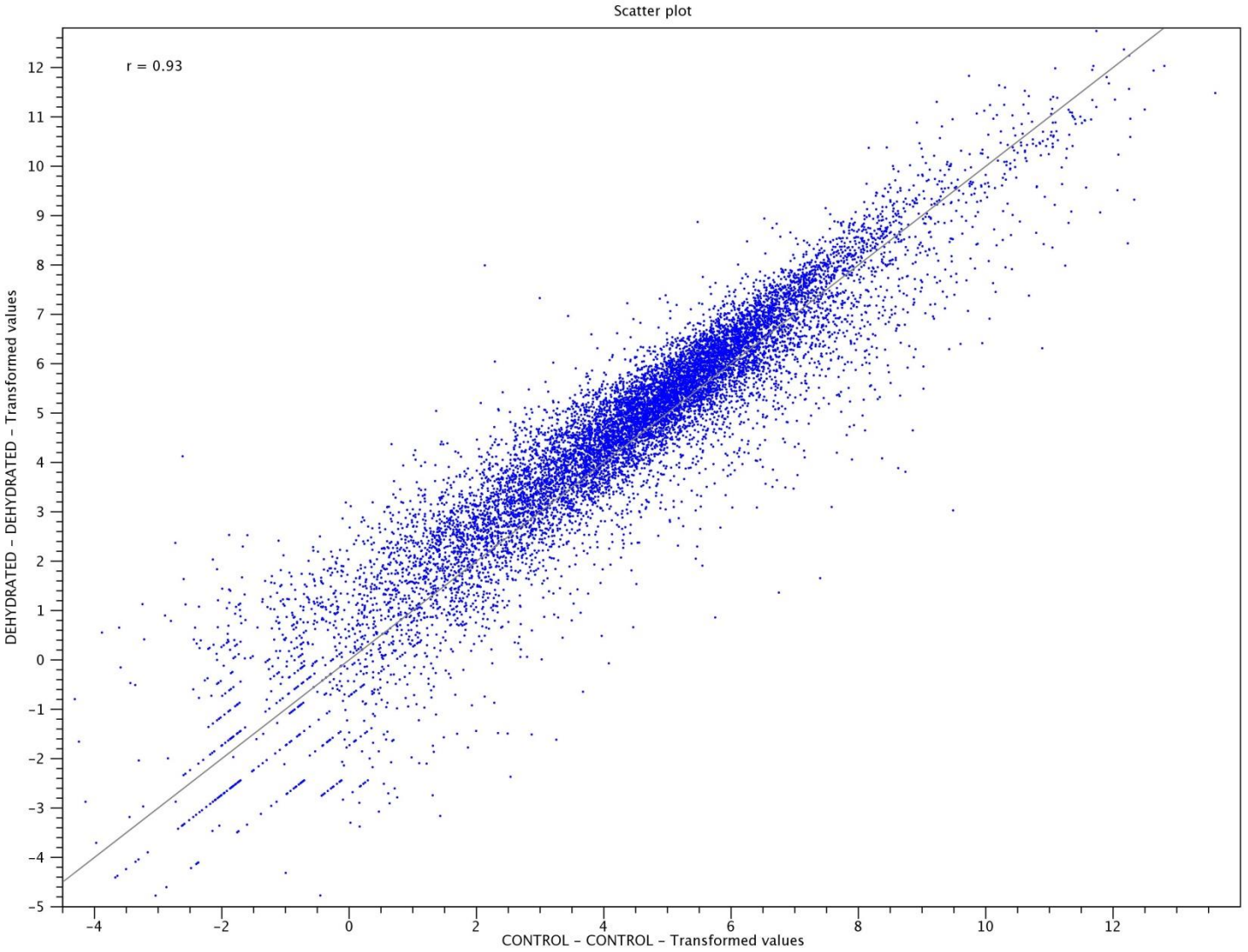
Comparison of TPM values



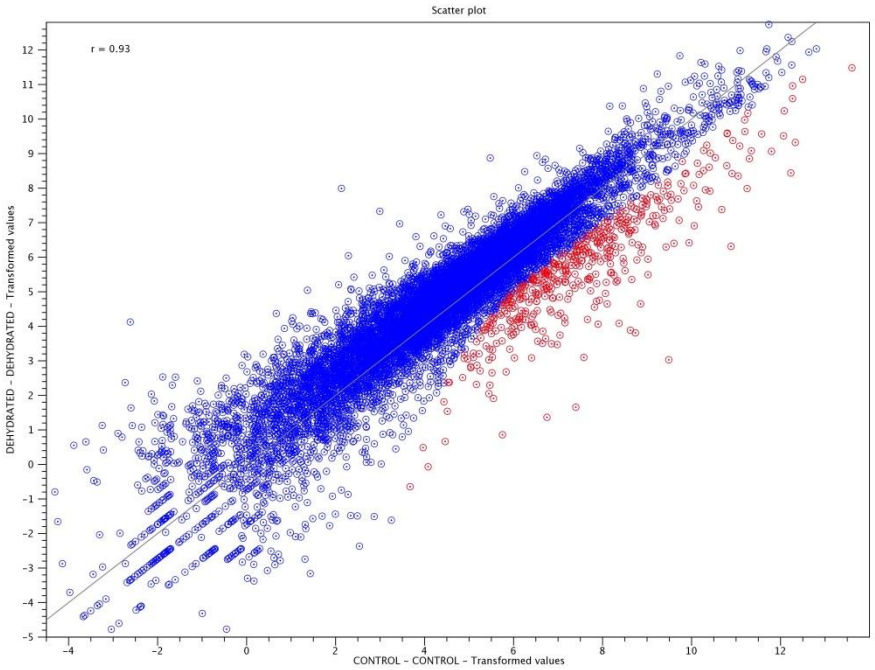
No major visual changes compared to RPKM

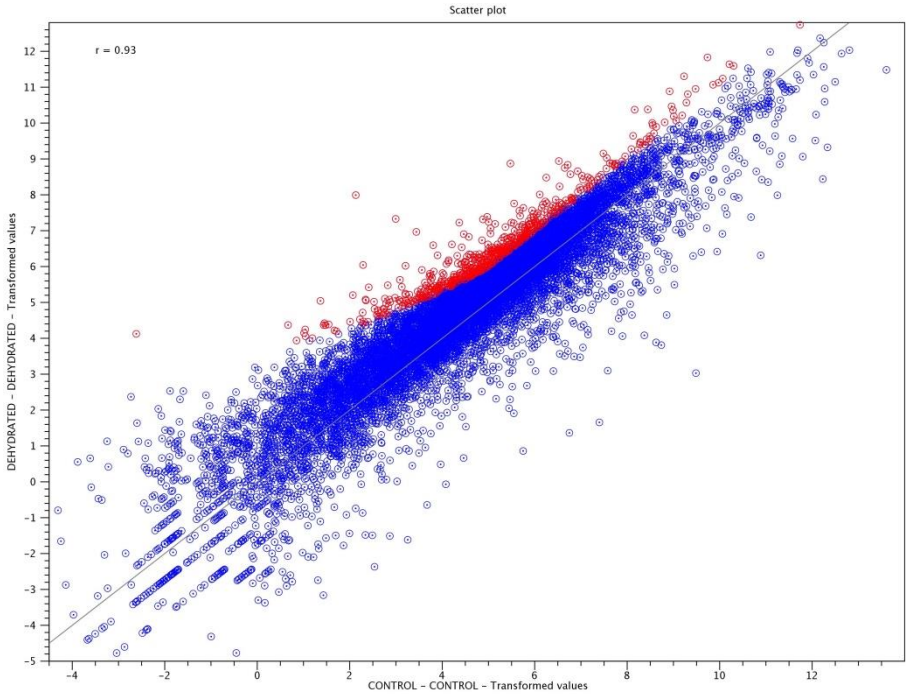
However, this measure makes the comparison between samples possible

Let’s switch to a better representation (Log2 transformation)



Statistical analysis of differential expression





We can easily spot up- and down-regulated genes.
Log2 transformation gives a «teardrop» effect in the graph
Higher expression changes are needed to achieve the same statistical significance for poorly expressed genes.

Dehydrated vs control comparison -> 530 up-regulated transcripts, 427 down-regulated transcripts

Results interpretation

- We have way to many **DEGs** to analyze them one by one
- Is there any class of genes which is up-regulated or down-regulated?
- We can use annotations (Gene Ontology, PFAM, eggNOG, etc.) to try figuring that out
- Hypergeometric tests on annotations can detect these alterations

Globally...

Groupp	Dehydrated vs Control	Rehydrated vs Dehydrated	Number of genes	% of the total
1	↑	↑	6	0.04
2	↑	=	500	3.66
3	↑	↓	24	0.18
4	=	↑	138	1.01
5	=	=	12479	91.43
6	=	↓	74	0.54
7	↓	↑	54	0.40
8	↓	=	366	2.68
9	↓	↓	7	0.05

91.43% of Trebouxia genes are NOT affected by the treatment. Yet, as you will see, the relatively low number of DEGs can tell us quite a lot...

Hypergeometric test results: an example – D vs C comparison (up-regulated genes)

Category	ID	Description	P-value	Proportion
Dehydrated vs Control				
Up-regulated				
eggNOG	COG0580	Glycerol uptake facilitator and related permeases	3.91E-4	4/9
eggNOG	COG1028	Dehydrogenases with different specificities	1.38E-3	8/50
GO_BP	GO:0009765	photosynthesis, light harvesting	2.72E-14	14/21
GO_BP	GO:0018298	protein-chromophore linkage	3.67E-12	14/27
GO_BP	GO:0015979	photosynthesis	1.13E-5	11/47
GO_BP	GO:0009736	cytokinin mediated signaling pathway	9.32E-4	4/10
GO_BP	GO:0005975	carbohydrate metabolic process	1.09E-3	10/65
GO_BP	GO:0006950	response to stress	4.39E-3	10/78
GO_CC	GO:0009522	photosystem I	2.22E-16	16/23
GO_CC	GO:0009523	photosystem II	3.87E-12	14/27
GO_CC	GO:0009535	chloroplast thylakoid membrane	2.66E-10	28/146
GO_CC	GO:0009538	photosystem I reaction center	2.75E-7	5/5
GO_CC	GO:0016021	integral to membrane	8.95E-6	96/1319
GO_CC	GO:0009543	chloroplast thylakoid lumen	5.31E-5	8/29
GO_CC	GO:0009505	plant-type cell wall	9.95E-5	5/11
GO_CC	GO:0005615	extracellular space	4.46E-3	5/23
GO_MF	GO:0016168	chlorophyll binding	1.55E-15	15/23
GO_MF	GO:0008242	omega peptidase activity	2.51E-4	4/8
GO_MF	GO:0043169	cation binding	4.27E-3	7/46
PFAM	PF00504	Chlorophyll A-B binding protein	7.48E-11	14/32
PFAM	PF13668	Ferritin-like domain	1.61E-6	7/14
PFAM	PF01124	MAPEG family	5.63E-6	4/4
PFAM	PF00134	Cyclin, N-terminal domain	5.60E-5	5/10
PFAM	PF00230	Major intrinsic protein	1.62E-4	5/12
PFAM	PF01370	NAD dependent epimerase/dehydratase family	3.20E-4	9/46
PFAM	PF00106	short chain dehydrogenase	3.55E-4	13/89
PFAM	PF00168	C2 domain	3.81E-4	6/21
PFAM	PF13561	Enoyl-(Acyl carrier protein) reductase	4.41E-4	10/58
PFAM	PF08659	KR domain	2.47E-3	10/72

Photosynthesis seems to be predominantly present in this table

How can we explain this in desiccation???

Hypergeometric test results: an example – D vs C comparison (down-regulated genes)

Down-regulated				
GO_BP	GO:0006950	response to stress	1.14E-5	12/78
GO_BP	GO:0009408	response to heat	1.38E-5	10/55
GO_BP	GO:0016485	protein processing	2.42E-4	4/10
GO_CC	GO:0005886	plasma membrane	6.14E-4	33/575
GO_CC	GO:0000502	proteasome complex	2.41E-3	5/30
GO_CC	GO:0009706	chloroplast inner membrane	7.70E-3	5/39
GO_MF	GO:0017111	nucleoside-triphosphatase activity	3.35E-3	7/61
GO_MF	GO:0043565	sequence-specific DNA binding	6.40E-3	5/37
PFAM	PF00012	Hsp70 protein	4.56E-6	6/14
PFAM	PF00320	GATA zinc finger	6.21E-4	4/12
PFAM	PF00004	ATPase associated with various cellular activities (AAA)	3.88E-3	8/72

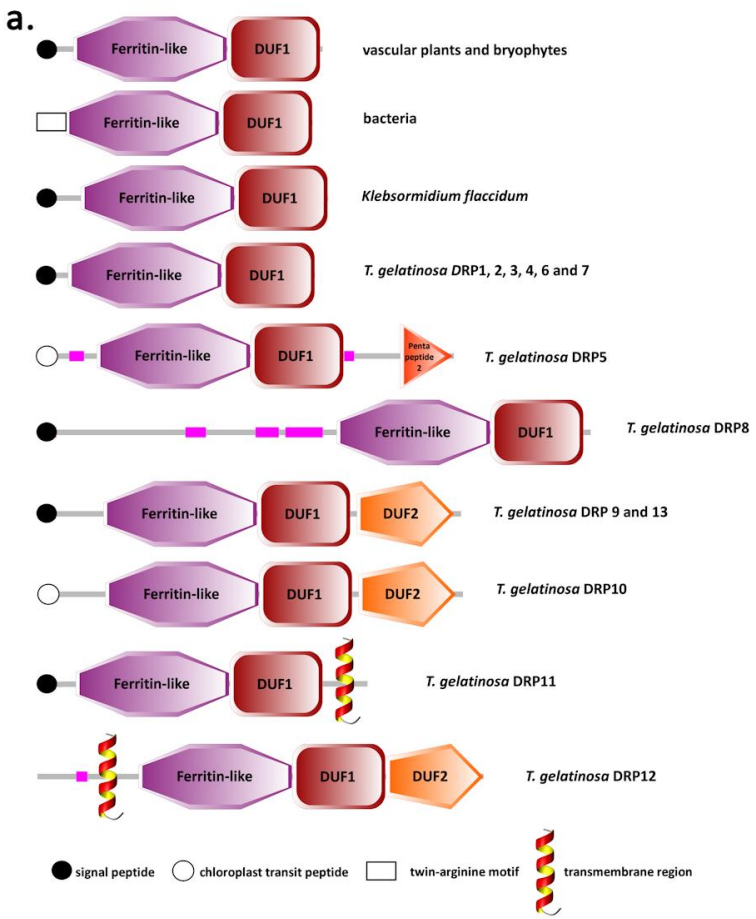
Heath Shock Proteins are down-regulated.

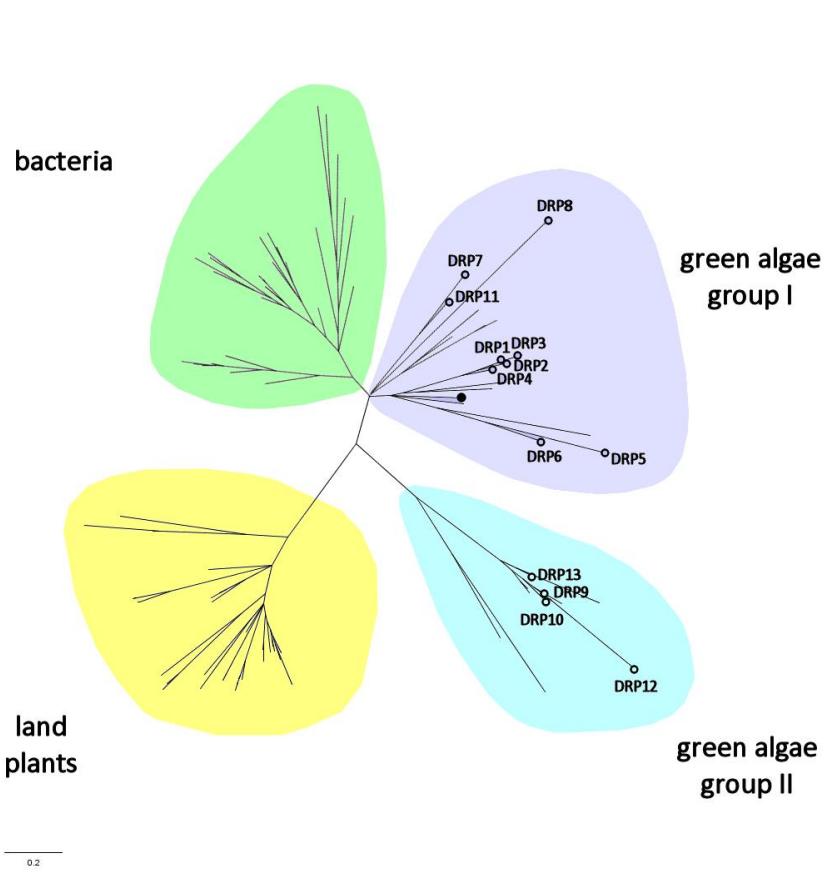
How can we explain this???

How do other desiccation-related plants deal with dehydration?

	Dehydration									Rehydration				
	Tg	Kc	Sr	Sl	Cp	Mf	Hr	Xh		Tg	Sr	Cp	Mf	Hr
Cell wall modifications	=	=	=	=	↓↑	=	↓	=		↑	=	↑	=	↑
HSPs and other chaperones	↓	=	↑	↑	=	=	=	=		↓	=	=	=	=
Late Embryogenesis Abundant proteins	=	↑	↑	↑	↑	↑	↑	↑		=	↑	↓	↓	↓
Aquaporins	↑	=	=	↑	↑↓	=	=	↓		=	↑	↑	=	=
Oxidative stress response	↑*	↑	↑	↑	↓	↑	↑↓	↑		=	↑	↑	↓	↑↓
Photosynthetic apparatus	↑	↑	↓	↑	↓	↑	↓	↓		=	↑	↑	=	↑

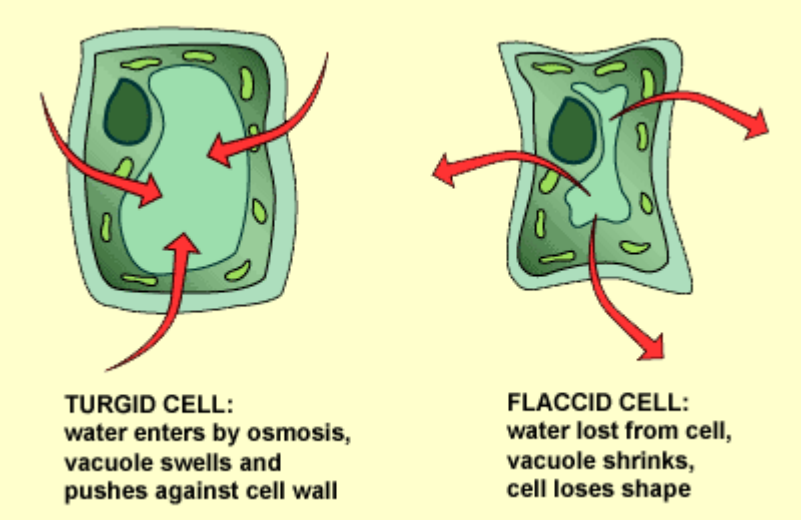
Structure and phylogeny of DRPs

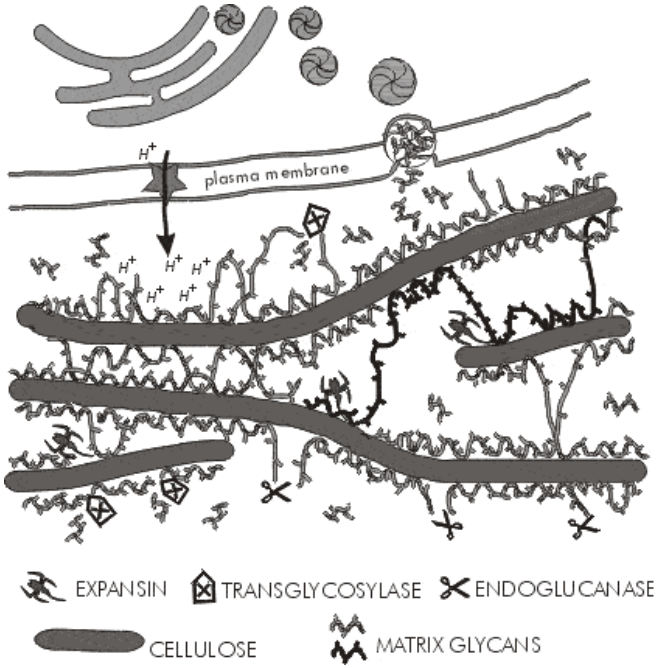




Cell wall modifications

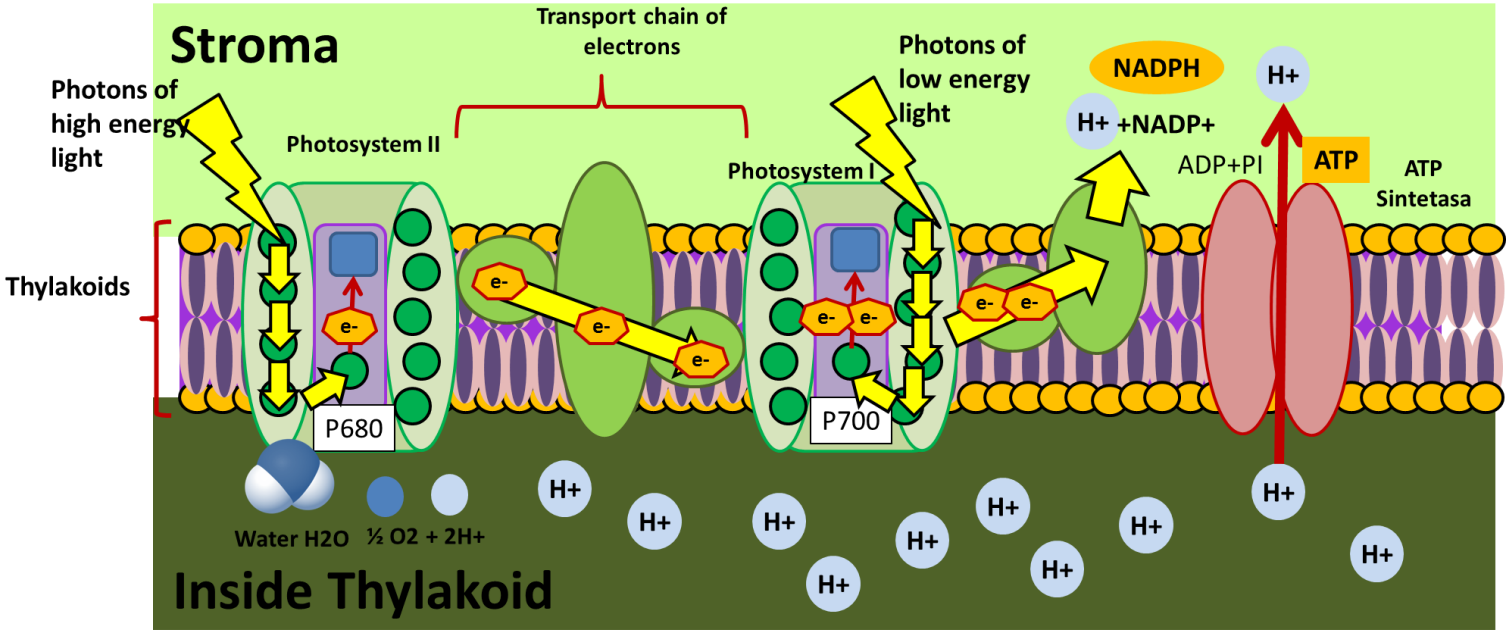
- Upon dehydration, cells are not turgid anymore and the cell wall needs to be remodeled
- Aquaporins (which regulate solute traficking across membranes) are also upregulated
- Consistent with literature!





Photoshynthesys

- Homoiochlorophyllic vs poikylochlorophyllic plants
- Destroy and rebuild vs protect and save energy
- If the poikylochlorophyllic strategy is chosen, then there is the need to deal with photoxidative damage!
- We find some genes involved in oxidative stress response upregulated
- We find some genes involved in the synthesis of anthocyanins upregulated



Confirmation by RT-PCR

- ✓ This is often asked by referees
- ✓ Need to confirm RNA-seq data with a differenet technique
- ✓ Assessment of inter-individual response variabilty