B@G 2018

Theory and methods of RNA-seq studies

Part III
differential splicing: DTE, DTU, DEU and sQTL, p.value correction, assessing methods' performance and recent developments

Simone Tiberi, University of Zurich

16/02/2018

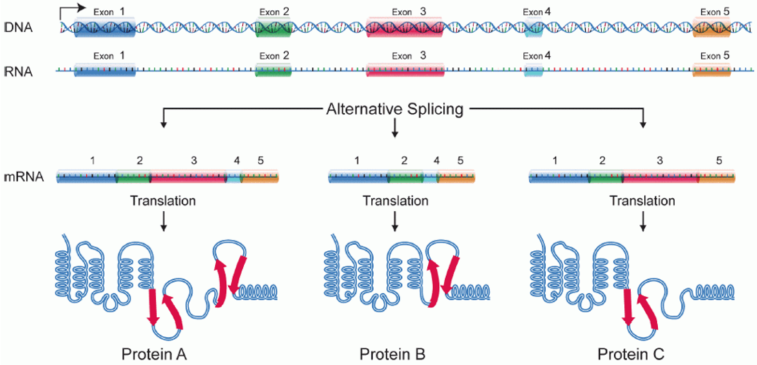
1. Alternative splicing
2. Modelling approaches
3. p.value correction (for all methods)
4. Assessing the performance of methods (for all methods)
5. Extra topics

References

1. Alternative splicing
2. Modelling approaches
3. p.value correction (for all methods)
4. Assessing the performance of methods (for all methods)
5. Extra topics

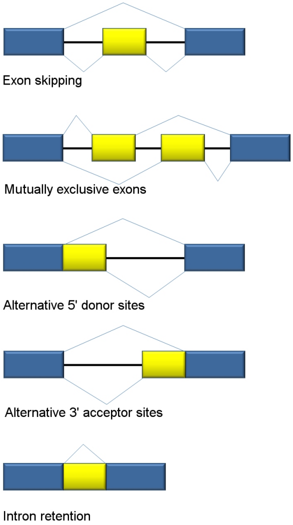
References

Alternative splicing



Wikipedia

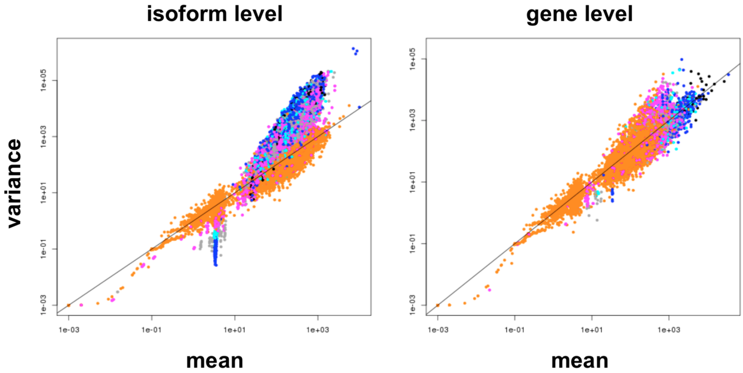
Alternative splicing



Wikipedia

More variable counts

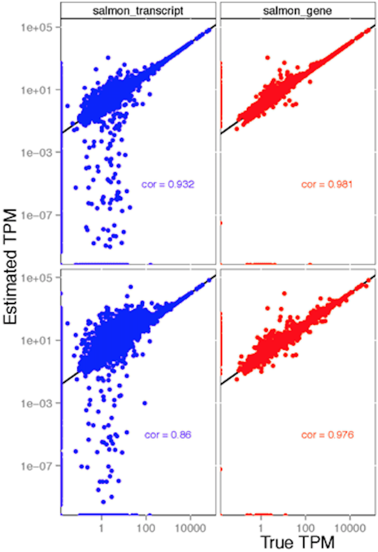
- Transcript level counts have higher variability than gene level counts: more biological variability.



Hubert Rehrauer, ETH Zurich

Less accurate estimates

- Transcript level estimates are less accurate than gene level estimates: higher measurement error.

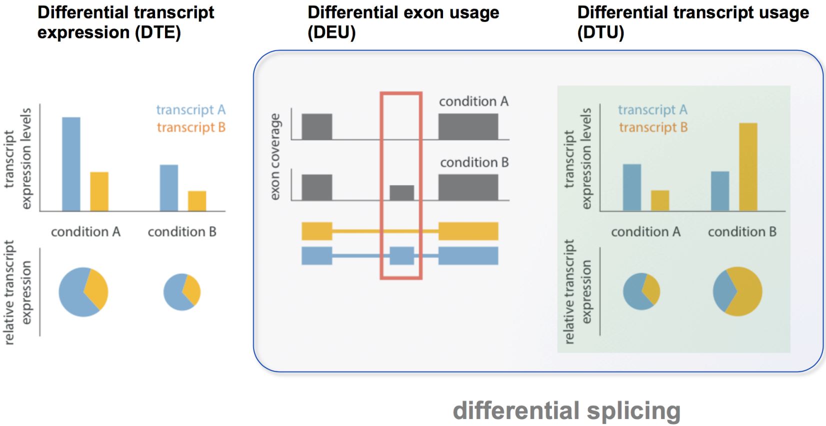


Soneson et al. F1000Research, 2015

DTE, DTU & DEU

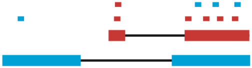
- Similarly to DGE for total gene expression, we can study if splicing patterns change between conditions.
- The main branches for differential splicing are differential transcript expression (DTE), differential transcript usage (DTU) and differential exon usage (DEU).
- **DTE** happens when the overall expressions of transcripts change between conditions (similar to DGE on transcripts instead of genes).
- **DTU** happens when the relative abundances of transcripts within a gene change between conditions, i.e. when the proportions of transcripts change. DTU implied DTE, not viceversa.
- **DEU** happens when the relative abundance of exons within a gene change between conditions, i.e. when the proportions of exons change.
- DEU is similar to DTU but it focuses on exons instead of transcripts. DTU is more appropriate because it looks at transcripts, which we are interested in: DEU is a proxy for DTU which focuses on exons for convenience.

DTE, DTU & DEU


Differential transcript expression (DTE)
Differential exon usage (DEU)
Differential transcript usage (DTU)
transcript expression levels
transcript A
transcript B
condition A
condition B
relative transcript expression
exon coverage
condition A
condition B
transcript expression levels
transcript A
transcript B
condition A
condition B
relative transcript expression
differential splicing

Slide adapted from Ma González-Porta's talk at ECCB 2014
http://radiant-project.eu/ECCB/gonzalez-porta-140907065638-phpapp01.pdf

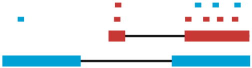
Transcript mapping uncertainty



Slide adapted from Trapnell et al. (2013), Nat Botech

- A big challenge in differential splicing analyses is that counts at the transcript level are not observed because many/most reads/fragments map to multiple transcripts.
- Three ways to overcome this issue are presented next.

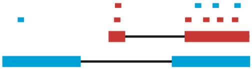
Input data (1): estimated transcript level counts



Slide adapted from Trapnell et al. (2013), Nat Botech

- Many DTU & DTE methods use transcript level estimated counts, obtained via EM algorithms, e.g. Salmon and Kallisto, and use these counts as input (*plug-in* like approach).
- The drawback is that transcript level estimated counts are treated as true counts, hence neglecting the uncertainty in their estimate.
- Approach used by **BayesDRIMSeq**, **DRIMSeq**, **limma**, **edgeR** (spliceVariants function) and **sleuth** (sleuth uses bootstrap replicates to account for the uncertainty in the estimates).
- Input data:
  - $\hat{\theta}_{blue}$: 4.4 and $\hat{\theta}_{red}$: 5.6.

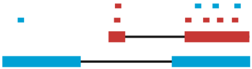
Input data (2): alignment of reads to the genome



Slide adapted from Trapnell et al. (2013), Nat Botech

- Other methods for DTU & DTE, such as **cjBitSeq** and **DEIsoM**, take as input the actual alignment of fragments in the genome and consider what transcripts each fragment maps to, e.g. via TopHat or STAR.
- The information about all fragments is typically summarized in clusters: all fragments mapping to the same transcripts are grouped together in one class by counting the total number of fragments in the class.
- Input data:
  - Clusters: $C_1 = \{blue\}$, $C_2 = \{red\}$ and $C_3 = \{blue, red\}$;
  - Cluster counts: $f_1 = 1$, $f_2 = 2$ and $f_3 = 7$.

Input data (3): disjoint bin counts



Slide adapted from Trapnell et al. (2013), Nat Botech

- **DEXSeq**, instead of considering transcript level counts, focuses on exon-level counts (where there is no mapping uncertainty) and studies DEU.
- DEXSeq divides the genome into disjoint bins (4 bins in the previous example) and takes as input the counts in every bin, which (unlike transcript level counts) are observed.
- Input data:
  - Exon bin counts: $e_1 = 1$, $e_2 = 2$, $e_3 = 0$ and $e_4 = 7$.

1. Alternative splicing

2. Modelling approaches

3. p.value correction (for all methods)

4. Assessing the performance of methods (for all methods)

5. Extra topics

References

Overview

- DTE is very similar to DGE, the main differences are that transcript level counts are not observed directly (they are latent states) and that fewer counts are available (hence the variability is higher).
    - sleuth and cjBitSeq are popular methods for DTE.
    - Alternatively, we can use the standard methods for DGE (edgeR, DESeq, DESeq2, etc...) on transcript level estimated counts (input (1)) and test each transcript separately.
- In the next slides we will illustrate a popular model for DTU: the Dirichlet-Multinomial (used by DRIMSeq, BayesDRIMSeq and DEIsoM).
- Then, we will show how DEXSeq performs DEU analyses.

DTU: Multinomial for one sample

- We consider one gene with K transcripts, where N samples are available.
- The transcript level counts for each sample of that condition are assumed, *a priori*, to have been generated from a Multinomial distribution:

$$X^{(i)}|\pi^{(i)} \sim Multinom(n^{(i)}, \pi^{(i)}), i = 1, ..., N, \tag{1}$$

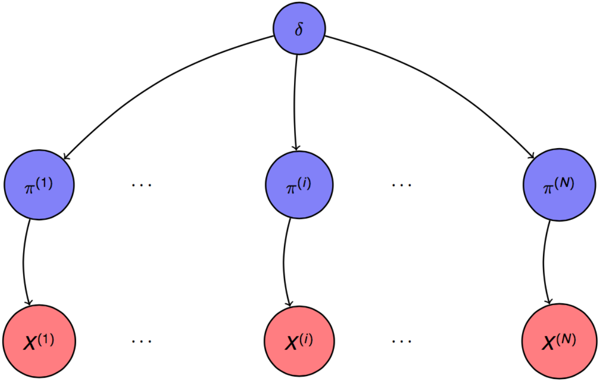
  where $\pi^{(i)} = (\pi_1^{(i)}, ..., \pi_K^{(i)})$ indicates the relative expression of transcripts $1, ..., K$ within the gene and $n^{(i)}$ represents the total number of counts aligning to the gene of interest in the i-th sample.
- When modelling gene expression, the negative binomial (NB) extends the Poisson distribution by allowing for over-dispersed data.
- Similarly, we extend the multinomial distribution allowing for extra variability, via random effects (frequentist statistics) or hierarchical modelling (Bayesian statistics).
- $\pi^{(i)}$ is assumed to vary between samples (under the same condition) due to biological variation.

Hierarchical structure/random effect model

- A hierarchical structure or random effect model is assumed where $\pi^{(i)}$ is assumed, *a priori*, to have been generated from a common distribution for all samples:

$$\pi^{(i)} \sim Dirichlet(\delta), i = 1, ..., N, \tag{2}$$

with $\delta = (\delta_1, ..., \delta_K)$.



DTU: Dirichlet-multinomial for multiple samples

- The marginal probability of the transcript level counts conditional on the hyper-parameter δ is

$$f(X^{(i)} = x|\delta) = \int f(x, \pi|\delta)\, d\pi = \int f_{Multin}(x|\pi)\, f_{Dir}(\pi|\delta)\, d\pi. \quad (3)$$

- This integral can be solved in closed form integrating out π, where (3) is the density of a Dirichlet-multinomial (DM) distribution.
- The distribution of $X^{(i)}$ given δ turns out to be a Dirichlet-multinomial (DM):

$$X^{(i)}|\delta \sim \mathcal{DM}(n^{(i)}, \delta), i = 1, ..., N. \quad (4)$$

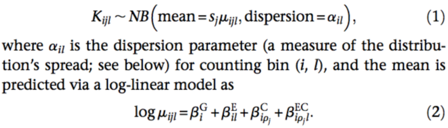
- δ can be decomposed in:
  - $\delta_+ = \sum_{k=1}^{K} \delta_k$, the dispersion parameter indicating how proportions $\pi^{(i)}$ vary between samples;
  - $\bar{\pi} = \dfrac{\delta}{\delta_+}$, representing the mean relative abundance of transcripts (mean across samples).

DTU: Dirichlet-multinomial for multiple samples

- When comparing two conditions, interest lies primarily in testing whether $\bar{\pi}$ varies between conditions.
- DRIMSeq, BayesDRIMSeq and DEIsoM all provide a global p.value at the gene level.

DEU: DEXSeq

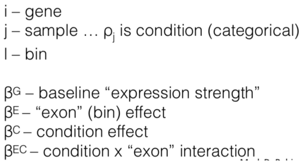
- DEXSeq works similarly to DGE methods but at the exon level: for every gene, it separately models the counts of every exon via a negative binomial (NB) distribution.
- DEXSeq includes an interaction term for the group/condition, β^C, and tests if this term is significant.
- Therefore it tests individual exons, not the entire gene: the p.value for the gene is obtained as the minimum p.value across the exons, which is not very elegant...
- Nevertheless, it works well and is quite popular.



$$K_{ijl} \sim NB\left(\text{mean} = s_j \mu_{ijl}, \text{dispersion} = \alpha_{il}\right), \qquad (1)$$

where α_{il} is the dispersion parameter (a measure of the distribution's spread; see below) for counting bin (i, l), and the mean is predicted via a log-linear model as

$$\log \mu_{ijl} = \beta_i^{G} + \beta_{il}^{E} + \beta_{i\rho_j}^{C} + \beta_{i\rho_j l}^{EC}. \qquad (2)$$



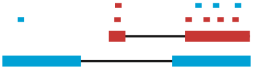
i – gene
j – sample … ρ_j is condition (categorical)
l – bin

β^G – baseline “expression strength”
β^E – “exon” (bin) effect
β^C – condition effect
β^{EC} – condition x “exon” interaction

Mark D Robinson, UZH

Our DTU model

- Me and Mark D Robinson (UZH) are working on a Bayesian model for DTU that combines the key two features of the available methods: i.e. the Dirichlet-Multinomial hierarchical approach and the modelling of the allocation of the fragments to the transcripts.



Slide adapted from Trapnell et al. (2013), Nat Botech

- Input data (2):
    - Clusters: $C_1 = \{blue\}$, $C_2 = \{red\}$ and $C_3 = \{blue, red\}$;
    - Cluster counts: $f_1 = 1$, $f_2 = 2$ and $f_3 = 7$.
- We treat the transcript allocation each fragment as a latent variable, i.e. as an unobserved random variable.
- In other words, the 7 allocations (blue or red) of the fragments in f_3 are treated as unknown parameters we also infer.

Our DTU model

- The advantage over inputing estimated counts is that the uncertainty in the transcript allocations is not neglected and properly modelled.
- We embed the model in a Bayesian hierarchical framework where, via a data augmentation procedure, we alternately sample the models' parameters and the allocation of fragments to the transcripts:
    - $X|\delta, \pi, \mathit{Data}$;
    - $\delta, \pi|X$.

Our DTU model

- Given two conditions, A and B, we denote by $\bar{\pi}^A$ and $\bar{\pi}^B$ the respective mean relative abundance of transcripts.
- When comparing conditions A and B, we test

$$\begin{cases} \mathcal{H}_0 & : \quad \omega_k = 0, \text{ for } k = 1, \dots, K, \\ \mathcal{H}_1 & : \quad \text{otherwise}, \end{cases} \tag{5}$$

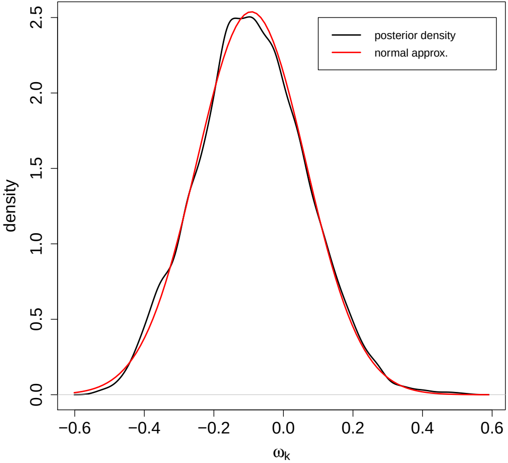
where $\omega_k = \bar{\pi}_k^A - \bar{\pi}_k^B$, $k = 1, \dots, K$.

- We approximate the posterior of $\omega = (\omega_1, \dots, \omega_K)$ with a normal distribution:

$$\omega | D \dot{\sim} \mathcal{N}\left(\hat{\omega}, \hat{\Sigma}_{\hat{\omega}}\right), \tag{6}$$

where $\hat{\omega}$ and $\hat{\Sigma}_{\hat{\omega}}$ are the posterior mode and variance-covariance of ω, inferred from the posterior chains.

Our DTU model



Example of normal approximation to a univariate posterior density.

Our DTU model

- Under $\mathcal{H}_0$, $\omega_k = 0 \, \forall k$,

  $$\hat{\omega}_{-K} \hat{\Sigma}^{-1}_{\omega_{-K}} \hat{\omega}^T_{-K} \dot{\sim} \chi^2_{K-1}, \tag{7}$$

  where the subscript $_{-K}$ indicates that the elements associated to the $K - th$ transcript have been removed.
  Clearly $K - 1$ parameters are involved because $\sum_{k=1}^K \omega_k = \sum_{k=1}^K \bar{\pi}^A_k - \sum_{k=1}^K \bar{\pi}^B_k = 0$.
- The result in (7) can be used to build a multivariate Wald test.

Splicing quantitative trait loci (sQTL)

- In Part II we introduced eQTL: if we have information about phenotypes, via SNP data, we can test if a gene has different expression levels between phenotypes.
- Similarly, splicing quantitative trait loci (sQTL) tests if a phenotype is linked to differential splicing.
- In other words, we can perform DTE, DTU and DEU analyses (with the same methods described above) where the separation in groups is defined by the SNPs.
- As for eQTL, also in sQTL we perform many more tests than for DTE, DTU and DEU: for every gene we need to test many SNPs; we typically only test the SNPs in the neighbourhood of the gene we are considering.

Transcript pre-filtering

- Most methods for differential splicing, in particular all the ones mentioned here, rely on a transcriptome reference.
- Transcript pre-filtering: Soneson et al. (2016), Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage, Genome Biology.
- Filtering the reference transcriptome (i.e. the cdna fasta file) by removing lowly expressed transcripts improves the performance of differential splicing methods:
    - run transcriptome aligner (e.g. Salmon or kallisto) with standard reference transcriptome;
    - select the transcripts with estimated proportion < threshold (5% seems an ideal choice) and filter the reference transcriptome by removing the lowly abundant transcripts and make a new reference transcriptome;
    - re-run the transcriptome aligner (e.g. Salmon or kallisto) with the new filtered reference.

1. Alternative splicing

2. Modelling approaches

3. p.value correction (for all methods)

4. Assessing the performance of methods (for all methods)

5. Extra topics

References

The p.value

- First of all, for a test we need a system of hypothesis, H_0 vs H_1, and a test statistic.
- A p.value is the probability of generating, under H_0, a value of the statistic as extreme (or more) as the one observed in our sample.
- A p.value of 0.05 means that, under H_0, there is a 5% probability of generating a value as extreme (or more) than the observed one: so, under H_0, every 100 tests, on average, 5 will be false positives (FPs) at the 5% significance threshold.
- If we perform 20,000 tests (one per gene) and H_0 is true for all of them, on average 1,000 will be FPs.
- Alternative ways of selecting the significant genes:
    - family wise error rate (FWER), the probability of obtaining at least 1 FP: very strict when testing thousands of genes;
    - false discovery rate (**FDR**), the proportion of FP among the significant genes: the most used method for selecting genes.

Controlling the FDR

- Instead of selecting the significant genes according to the p.value, we can correct the p.value such that:
  - the ordering of genes is preserved;
  - the FDR is controlled at the α threshold.
- The Benjamini-Hochberg (BH) correction is the most popular method for controlling the FDR.
- We select the genes whose corrected p.value (often called q.value) is $< \alpha$: therefore for $\alpha = 0.05$, we expect that 5% of the selected genes will be FPs and 95% will be TPs.
- This is typically performed in both differential expression and differential splicing methods.

Independent filtering

- We often filter out genes *a priori*, i.e. before testing.
- Typically we filter genes with very low expression, below a specified threshold.

1. Alternative splicing

2. Modelling approaches

3. p.value correction (for all methods)

4. Assessing the performance of methods (for all methods)

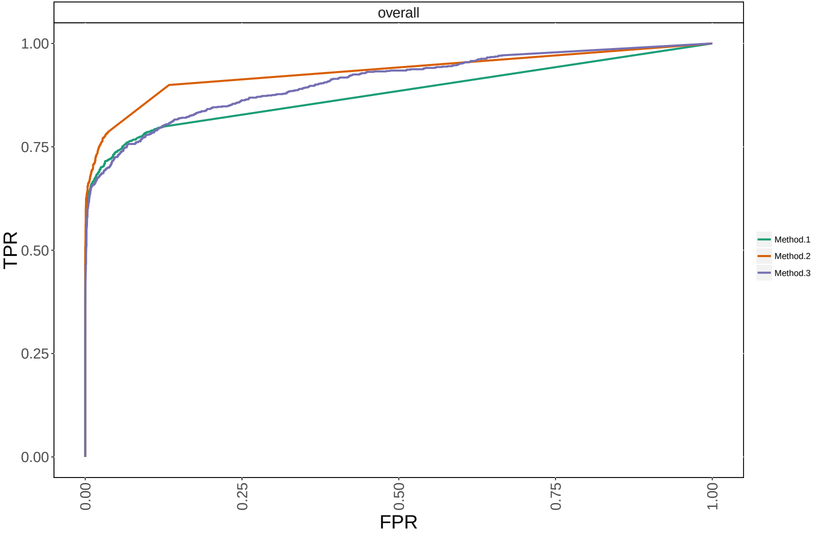
5. Extra topics

References

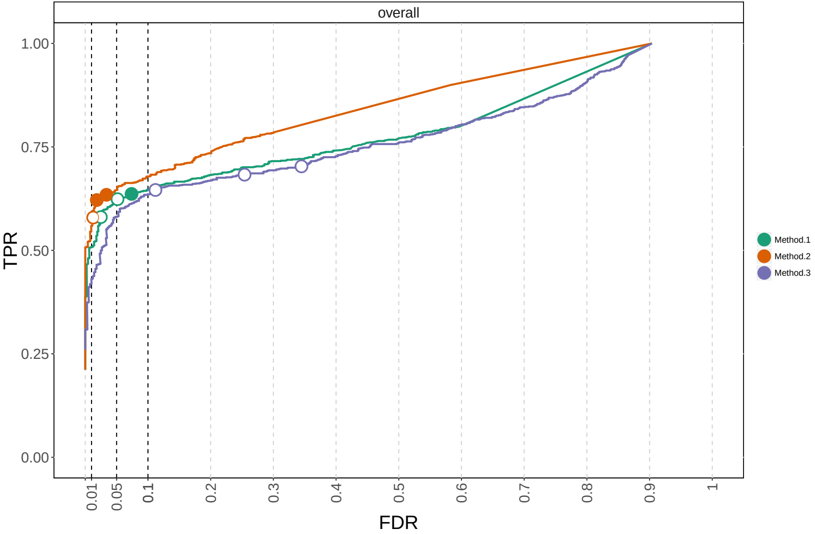
Assessing the performance of method

- When assessing the performance of methods (DGE, eQTL, DTU, etc...) and comparing them, we mostly investigate several plots.
- This is always done in simulations where we know the **truth**.
- The main plots (seen in the 4th tutorial) are:
  - ROC: true positive rate (TPR) vs false positive rate (FPR);
  - **FDR**: TPR vs FDR (more interesting than the ROC in this field), mostly interested in seeing if FDR is controlled at the chosen threshold;
  - **FPs** vs top significant genes: it tells us how well a model does in the very top genes, which are those we are mostly interested in;
  - precision vs recall plot, where:
    - precision = TP/(TP + FP), i.e. TP over all positive,
    - recall = TP/(TP + FN), i.e. TP over signifiant genes.
- **iCOBRA** is a very useful R package for plotting the performance of a methods.

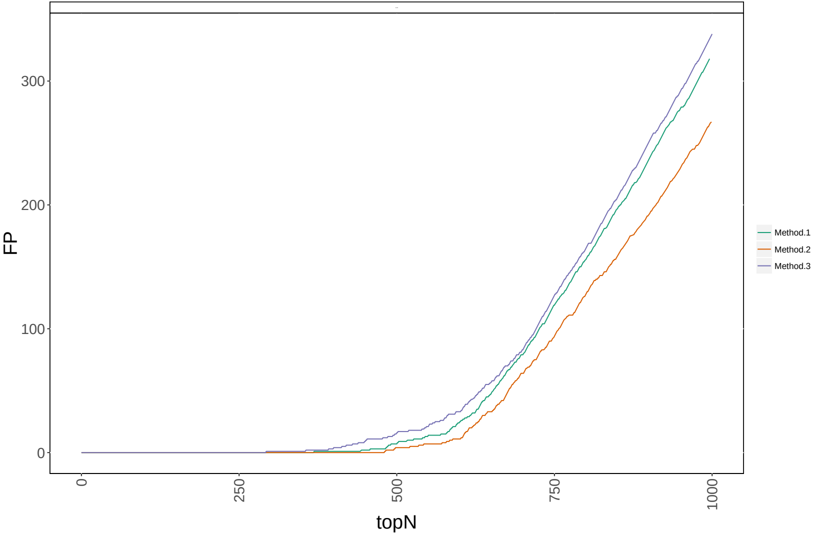
ROC curve


overall
TPR
1.00
0.75
0.50
0.25
0.00
FPR
0.00
0.25
0.50
0.75
1.00
Method.1
Method.2
Method.3

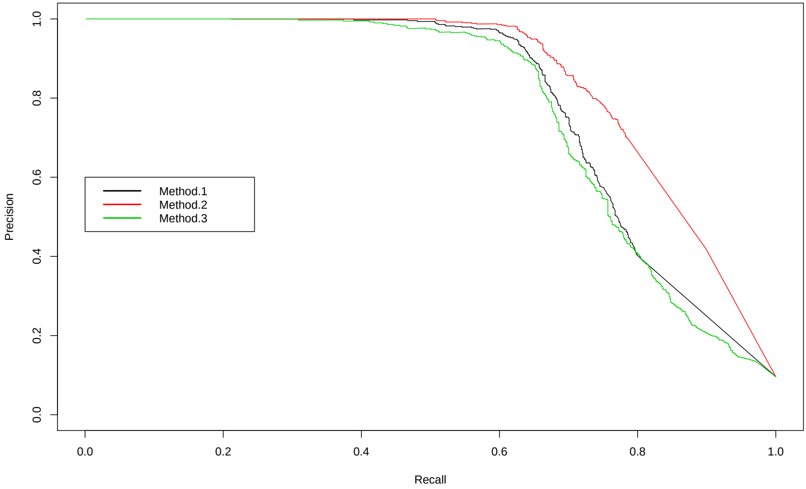
FDR plot


overall
TPR
1.00
0.75
0.50
0.25
0.00
FDR
0.01
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
● Method.1
● Method.2
● Method.3

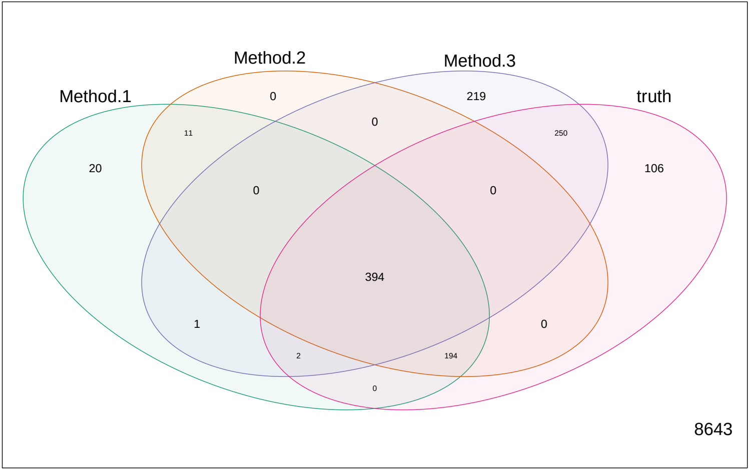
FPs vs top selected genes



Precision recall plot



Venn diagram



What comes after RNA-seq studies?

- The main aim of RNA-seq studies is to identify genes (but also transcripts or pathways) which show interesting characteristics (DGE, differential splicing, etc...): these genes will typically be studied more in depth in further biological analyses.
- This is why we are mostly interested in selecting correctly the top genes, because these genes might undergo further studies.

1. Alternative splicing

2. Modelling approaches

3. p.value correction (for all methods)

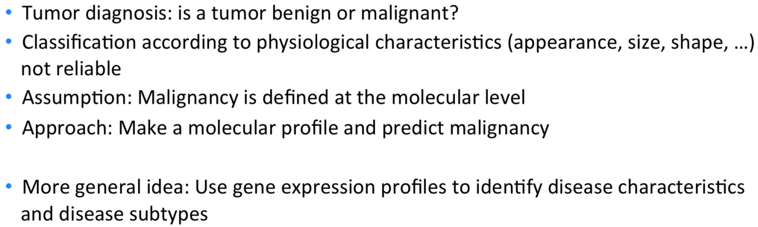
4. Assessing the performance of methods (for all methods)

5. Extra topics

References

Classification

- We can use RNA-seq profiles first to classify diseases, e.g. tumor sub-types, and then eventually to predict them.
- We can also classify species or populations.



- Tumor diagnosis: is a tumor benign or malignant?
- Classification according to physiological characteristics (appearance, size, shape, …) not reliable
- Assumption: Malignancy is defined at the molecular level
- Approach: Make a molecular profile and predict malignancy

- More general idea: Use gene expression profiles to identify disease characteristics and disease subtypes

Hubert Rehrauer, ETH Zurich

Gene set, gene ontology and pathway analyses

- Once we identify differential genes, usually differential expression but also differential splicing, we can study if they group together in terms of function and pathway.
- In other words, stating from gene-level results (p.values), we study the significance of entire pathways/groups of genes, involved in similar processes or having related functions.
- A common way to proceed is to set a significance threshold and split the genes in two groups, differential and non-differential genes, and to see if in each pathway there are more differential genes than we would expect at random (null distribution: hypergeometric distribution). However this relies on the threshold we choose.
- Alternative, more sophisticated, approaches focus on the ranking of genes based on the p.value.

Gene set, gene ontology and pathway analyses

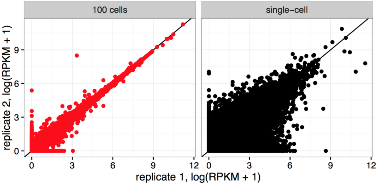
- R library goseq to perform gene ontology analyses.
- Overview on the topic:
  - Khatri et al. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges, Plos Computational Biology.
  - Goeman et al. (2007). Analyzing gene expression data in terms of gene sets: methodological issues, Bioinformatics.

Recent developments: long reads

- Thirds generation sequencing, long reads, mostly PacBio and Oxford Nanopore.
- Recent technology able to sequence longer strings of mRNA, i.e. thousands of base pairs (bp).
- Pros: it can sequence full transcripts, hence it is very useful for *de novo* transcript detection and for alternative splicing studies.
- Cons: higher per base error rate and more expensive than Illumina RNA-seq.

Recent developments: single-cell RNA-seq

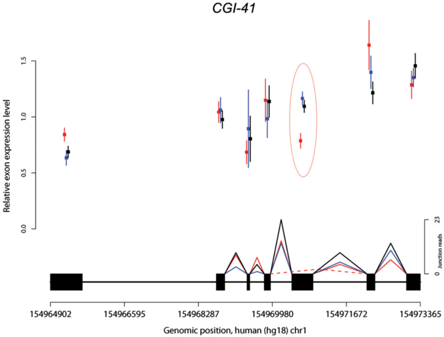
- Standard RNA-seq, also called “bulk RNA-seq”, report expression levels for a population of cells.
- Recent technologies allow to observe mRNA molecules in single-cells: single-cell RNA-seq (scRNA-seq).
- Pros: higher resolution, it allows to study biological variability more in depth and to study how gene expression varies during the life cycle of the cell.
- Cons: higher variability (more difficult to model and to detect signals), excess of 0 counts (cells with no expression) and more expensive.



Mark D Robinson, ETH Zurich

And the adaptation part?

- Changes in expression and splicing patterns are mostly studied related to diseases, but they are also ways of adapting and can be studied in the context of adaptation, e.g. by comparing species, populations in different environments, etc...
- Below: an example of human-specific change in exon usage, between human (red), chimpanzee (blue) and rhesus macaque (black).



Blekhman et al. (2009). Sex-specific and lineage-specific alternative splicing in primates, Genome Res.

1. Alternative splicing

2. Modelling approaches

3. p.value correction (for all methods)

4. Assessing the performance of methods (for all methods)

5. Extra topics

References

References I

- **DEU: DEXSeq**: Anders et al. (2012). Detecting differential usage of exons from RNA-seq data, Genome Research.
- **DTU: DRIMSeq**: Nowicka et al. (2016). DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics, F1000Research.
- **DTU: cjBitSeq & BayesDRIMSeq**: Papastamoulis et al. (2017). Bayesian estimation of differential transcript usage from RNA-seq data, Statistical Applications in Genetics and Molecular Biology.
- **DTE: cjBitSeq**: Papastamoulis et al. (2017). A Bayesian model selection approach for identifying differentially expressed transcripts from RNA sequencing data, J. Royal Statistical Society, Series C.
- **DTE: sleuth**: Pimentel et al. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty, Nature Methods.
- **DTE & DTU: DEIsoM**: Peng et al. (2017). DEIsoM: a hierarchical Bayesian model for identifying differentially expressed isoforms using biological replicates, Bioinformatics.

References II

- Transcript pre-filtering: Soneson et al. (2016). Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage, Genome Biology.
- **iCOBRA**: Soneson et al. (2016). iCOBRA: open, reproducible, standardized and live method benchmarking, Nature Methods.
- Pathway anaysis: Khatri et al. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges, Plos Computational Biology.
- Gene sets: Goeman et al. (2007). Analyzing gene expression data in terms of gene sets: methodological issues, Bioinformatics.

Questions?