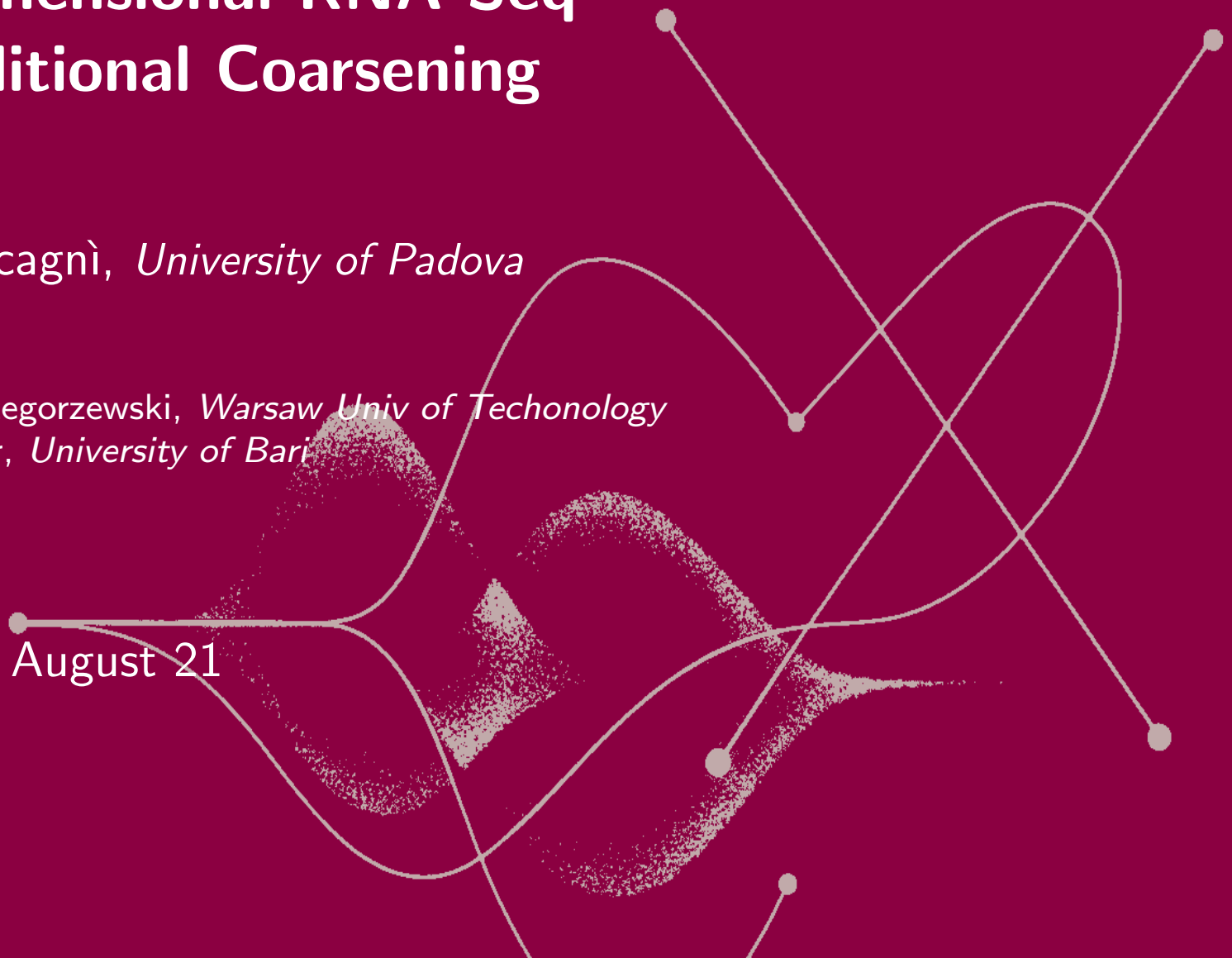
Statistical Modeling of Fuzzy Counts in High-Dimensional RNA-Seq via Conditional Coarsening

Antonio Calcagnì, *University of Padova*

jointly with
Przemyslaw Grzegorzewski, *Warsaw Univ of Techonology*
Corrado Mencar, *University of Bari*

HDDA-XIV, August 21



Introduction

RNA sequencing (**RNA-seq**) is a key technology in computational biology, enabling comprehensive **gene expression** measurement and advancing our understanding of genetic regulation.

Introduction

Typically, RNA-seq involves three main steps:

- Generating short sequencing reads from RNA molecules (e.g., Illumina)
- Aligning them to a reference transcriptome (e.g., HISAT2)
- Quantifying gene expression levels and perform statistical analyses (e.g., differential expression)

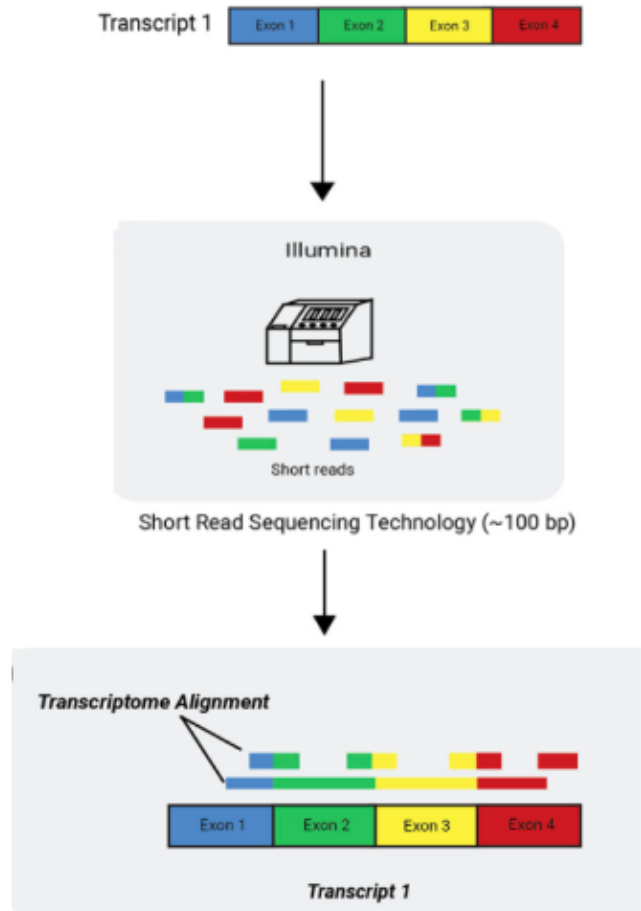


Figure adapted from [Deshpande et al., 2023]

Introduction

RNA-seq and statistical modeling

Due to their **high-throughput nature**, RNA-seq data pose a major challenge for statistical modeling (tens of thousands of gene expression measurements for a relatively small number of samples).

This raises inference issues such as:

- controlling for multiple testing in differential expression
- modeling severe overdispersion inherent in RNA-seq count data
- accounting for non-independence in gene expression from individual cells

Introduction

RNA-seq and statistical modeling

14TH HIGH DIMENSIONAL DATA ANALYSIS WORKSHOP (HDDA-XIV)

An additional problem arises with non-integer gene expression counts:

Multireads - reads that align to multiple genomic locations simultaneously.

Multicovers - reads that align to overlapping regions of the transcriptome.

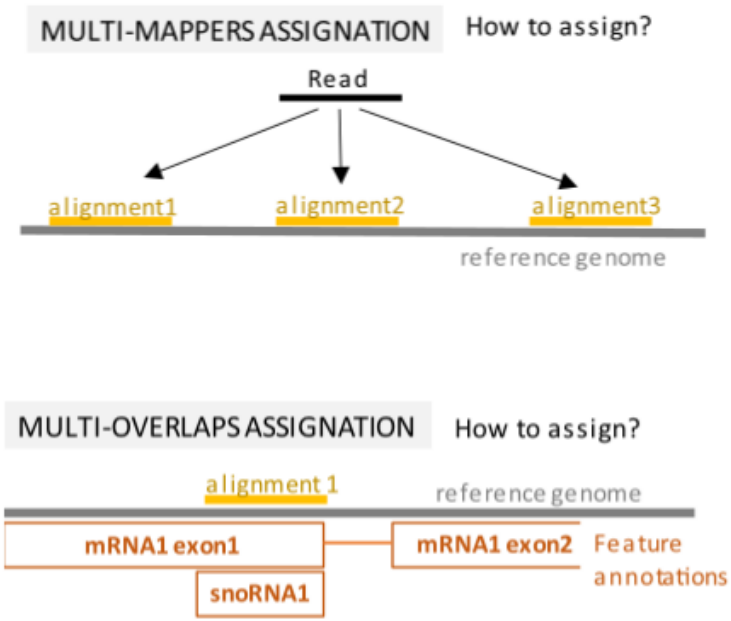


Figure adapted from [Hita et al., 2022]

Introduction

RNA-seq and statistical modeling

MULTIREADS

read_id	chr	MD	NM	gene_id
SRR069840.5	chr2L	28T0C2T0A0G0T0	6	FBti0019163
SRR069840.5	chr2L	28T0C2T0A0G0T0	6	FBti0019210
SRR069840.5	chr2L	28T0C2T0A0G0T0	6	FBti0019145
SRR069840.5	chr2L	28T0C2T0A0G0T0	6	FBgn0265002
SRR069840.5	chr3L	28T0C2T0A0G0T0	6	FBti0020070
SRR069840.5	chr3L	28T0C2T0A0G0T0	6	FBgn0262719
⋮	⋮	⋮	⋮	⋮

MULTICOVERS

read_id	chr	MD	NM	gene_id
SRR069840.5	chr2L	28T0C2T0A0G0T0	6	FBti0019163
SRR069840.5160	chr2L	26C1T0C2T0A0G0T0	7	FBti0019163
SRR069840.6831	chr2L	27T0A0A1C0C2C0	6	FBti0019163
SRR069840.8804	chr2L	26C0T0A1A1T0T0T0G0	8	FBti0019163
SRR069840.10148	chr2L	25T3T1A0T0T0G0A0	7	FBti0019163
SRR069840.11245	chr2L	27C1C0G1T0A0G1	6	FBti0019163
⋮	⋮	⋮	⋮	⋮

[An excerpt of Drosophila melanogaster RNA-seq alignment (with Bowtie)]

Introduction

RNA-seq and statistical modeling

Some ways of handling multireads are [Deschamps-Francoeur et al., 2020]:

- ignoring them entirely
- distributing them equally across all possible alignments (fractional counts)
- allocating them probabilistically via EM-based solutions (e.g. Kallisto, Salmon), producing normalized expected counts (e.g., TPMs)

Introduction

Read-to-gene alignment problem

This many-to-one counting process introduces an additional layer of uncertainty that arise from **imperfect knowledge** of the underlying genome [Ji et al., 2011].

Note that such uncertainty occurs *after* data collection (**post-sampling epistemic uncertainty**).

In this context, many-to-one counting can be naturally represented using granular computing or **fuzzy counts** [Consiglio et al., 2016, Mencar and Pedrycz, 2020].

Read-to-gene alignment problem

Fuzzy counts

A **fuzzy count** $\tilde{n}$ is characterized by its characteristic function

$$\xi_{\tilde{n}} : \mathbb{N}_0 \to [0, 1]$$

where the quantity $\xi_{\tilde{n}}(n)$ is usually interpreted as the possibility that the crisp count $n \in \mathbb{N}_0$ has to occur, with $\xi_{\tilde{n}}(n) = 1$ indicating that n is fully possible.

Read-to-gene alignment problem

Fuzzy counts

A **fuzzy count** $\tilde{n}$ is characterized by its characteristic function

$$\xi_{\tilde{n}} : \mathbb{N}_0 \to [0, 1]$$

where the quantity $\xi_{\tilde{n}}(n)$ is usually interpreted as the possibility that the crisp count $n \in \mathbb{N}_0$ has to occur, with $\xi_{\tilde{n}}(n) = 1$ indicating that n is fully possible.

Note that $\xi_{\tilde{n}}$ is not a probability distribution:

- it is not linked to any random experiment
- it encodes epistemic uncertainty: the true but hidden realization n randomly occurrs, yet it is imprecisely measured
- it is a non-additive measure, namely a **possibility measure**.

Read-to-gene alignment problem

Statement of the problem

Let

$$\mathcal{R} = \{r_1, \ldots, r_i, \ldots, r_I\} \quad \text{and} \quad \mathcal{G} = \{g_1, \ldots, r_j, \ldots, g_J\}$$

be the sets of reads and genes, respectively, and

$$(r_i, g_j) \in \mathcal{M} \subseteq \mathcal{R} \times \mathcal{G}$$

be the read-to-gene association. Here,

$$\mathcal{G}_{r_i} = \{g_j \in \mathcal{G} : (r_i, g_j) \in \mathcal{M}\} \quad \text{or} \quad \mathcal{R}_{g_j} = \{r_i \in \mathcal{R} : (r_i, g_j) \in \mathcal{M}\}$$

are the subsets induced by fixing a particular read or gene.

Read-to-gene alignment problem

Statement of the problem

Aim: represent the output of the read-to-gene alignment as a fuzzy count.

Read-to-gene alignment problem

Statement of the problem

Particularly, the **expression level** of gene g_j is the fuzzy count

$$\tilde{n}_{g_j} = \{(n, \xi_{\tilde{n}_{g_j}}(n))\}_{n=0}^{N_j},$$

where the possibility distribution $\boldsymbol{\xi}_{\tilde{n}_g}$ needs to be determined from the data.

Several solutions can be adopted here, which either emphasize theoretical coherence [Mencar and Pedrycz, 2020] or adopt a more computational-oriented approach [Consiglio et al., 2016].

Read-to-gene alignment problem

A computational-oriented solution

Particularly, for a given gene g_j:

$$\xi_{\tilde{n}_{g_j}} = f\left(\texttt{alignment\_quality}(\mathcal{R}_{g_j}), \texttt{multicover\_quality}(\mathcal{R}_{g_j}), \texttt{penality}(\mathcal{G}_{r_i})\right)$$

Read-to-gene alignment problem

A computational-oriented solution

Particularly, for a given gene g_j:

$$\begin{aligned}\xi_{\tilde{n}_{g_j}} &= f(\texttt{alignment\_quality}(\mathcal{R}_{g_j}), \texttt{multicover\_quality}(\mathcal{R}_{g_j}), \texttt{penality}(\mathcal{G}_{r_i})),\\ &= f\left(\frac{|\mathcal{M}_{g_j}|}{|\mathcal{M}_{g_j}| + |\mathcal{NM}_{g_j}|} \times \frac{|\mathcal{R}_{g_j}|}{\max_k |\mathcal{R}_{g_k}|} \times \frac{1}{|\mathcal{G}_{r_i}|}\right),\\ &= f(\mathbf{u}),\end{aligned}$$

with

$\mathcal{M}_{g_j}$ the set of matches in the collection of `MD` strings,

$\mathcal{NM}$ the set of `NM` numbers.

Read-to-gene alignment problem

A computational-oriented solution

By computing the above triples over $\mathcal{R} \times \mathcal{G}$, we obtain a matrix $\mathbf{U} \in [0,1]^{I \times J}$ containing the masses of evidence for the read-gene associations.

Note that if $\mathbf{U}$ was row-normalized so that $\sum_j u_{ij} = 1$ for each $i \in \{1, \ldots, I\}$, it would resemble the so-called membership matrix in fuzzy clustering.

However, while fuzzy clustering focuses on assigning each read to a gene, fuzzy counting focuses on the *accumulation of evidence* for each gene.

Read-to-gene alignment problem

A computational-oriented solution

gene_id	read_id	multicover_qual	align_qual	penal	u
NM_001142431	11710154	0.027	1.000	0.058	0.001
NM_001142431	21832111	0.027	1.000	0.058	0.001
NM_001142431	21779325	0.027	1.000	0.058	0.001
NM_001142431	26638916	0.027	0.986	0.058	0.001
NM_001142431	5995783	0.027	1.000	0.058	0.001
NM_001142431	2978344	0.027	1.000	0.058	0.001
⋮	⋮	⋮	⋮	⋮	⋮

[An excerpt of Human chrX RNA-seq u-calculus]

Read-to-gene alignment problem

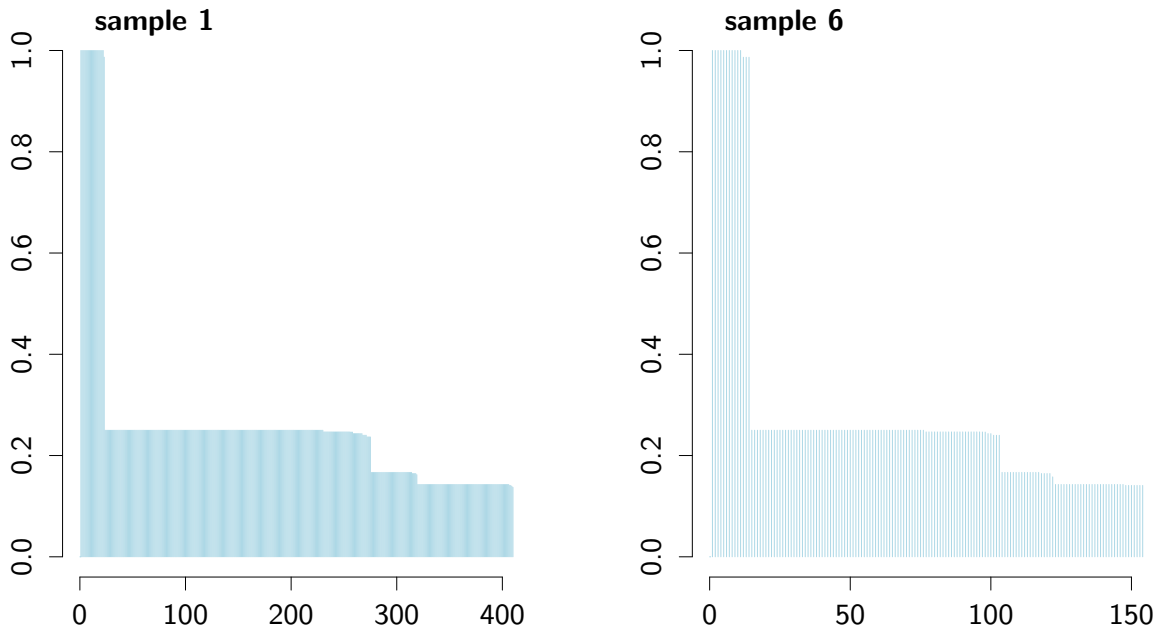
A computational-oriented solution

Finally, the fuzzy count $\tilde{n}_{g_j}$ is computed retaining the *strongest minimal evidence* obtainable across all possible subsets of k reads supporting that gene:

$$\mathbf{u}^{(j)} = (u_{1j}, \ldots, u_{ij}, \ldots, u_{Ij}),$$
$$N_j = |\{i : u_{ij} > 0\}|,$$
$$\xi_k^{(j)} = \max_{\substack{\mathcal{S} \subseteq \{1,\ldots,I\} \\ |\mathcal{S}|=k}} \min_{s \in \mathcal{S}} \mathbf{u}_s^{(j)}.$$

Read-to-gene alignment problem

A computational-oriented solution



[An excerpt of Human chrX RNA-seq fuzzy counts]

Fuzzy counts as coarsened data

Let $N : (\Omega, \mathcal{A}, \mathbb{P}) \to (S, \mathcal{S})$ be a $\mathcal{A}$-$\mathcal{S}$-measurable function. The induced distribution $\mathbb{P}_N$ on $(S, \mathcal{S})$ is assumed to belong to a *parametric* family $\{\mathbb{P}_\theta : \theta \in \Theta\}$.

The sample $N_1, \ldots, N_I$ is assumed to be blurred into the **fuzzy sample**

$$\tilde{\mathbf{n}} = (\tilde{n}_1, \ldots, \tilde{n}_I),$$

with $\tilde{n}_i$ being a fuzzy subset of S characterized by a Borel-measurable membership function $\xi_{\tilde{n}_i} : S \to [0, 1]$.

The statistical problem here is to identify $\hat{\theta} \in \Theta$ such that $\mathbb{P}_{\hat{\theta}}$ describes the distribution of $\mathbf{n}$ based on $\tilde{\mathbf{n}}$. This is a type of **filtering** or **de-blurring** problem.

Fuzzy counts as coarsened data

Fuzziness requires CNAR

Under the **Tanaka-Okuda approach** to fuzzy data analysis [Tanaka et al., 1977, Gebhardt et al., 1998], **fuzziness is not ignorable**.

Fuzzy counts as coarsened data

Fuzziness requires CNAR

Under the **Tanaka-Okuda approach** to fuzzy data analysis [Tanaka et al., 1977, Gebhardt et al., 1998], **fuzziness is not ignorable**.

> [Gill and Grünwald, 2008]
>
> A *coarsening mechanism* is a mapping $\phi : S \to \mathcal{S}_{\setminus\{\emptyset\}}$ such that for any realization $n \in S$ of N, the observers measures a coarsened version of it, namely the set $A \in \mathcal{S}_{\setminus\{\emptyset\}}$ containing n.
>
> ϕ is coarsening-at-random (CAR) iff
>
> $$\mathbb{P}[A \mid N = n] = \mathbb{P}[A \mid N = n'], \quad \forall n, n' \in A.$$
>
> Hint: n and n' are *exchangeable* within A.

Fuzzy counts as coarsened data

Fuzziness requires CNAR

Under the **Tanaka-Okuda approach** to fuzzy data analysis [Tanaka et al., 1977, Gebhardt et al., 1998], **fuzziness is not ignorable**.

If $\tilde{\mathcal{S}}$ constitutes a collection of fuzzy subsets of S (i.e., a fuzzy cover) then ϕ **is no longer CAR**:

$$\underbrace{\mathbb{P}[\tilde{A} \mid N = n]}_{\propto\ \xi_{\tilde{A}}(n)} \neq \underbrace{\mathbb{P}[\tilde{A} \mid N = n']}_{\propto\ \xi_{\tilde{A}}(n')}$$

Hint: $\xi_{\tilde{A}}(n)$ varies over $n \in \tilde{A}$, realizations are *no longer exchangeable* in A.

Fuzzy counts as coarsened data

Fuzziness requires CNAR

This argument leads to coarsening-not-at-random (**CNAR**).

As in MNAR problems [Molenberghs and Verbeke, 2005], a similar factorization arises in this context:

$$\mathbb{P}_\theta(\mathbf{n}, \tilde{\mathbf{n}} \mid \ldots) = \underbrace{\mathbb{P}_\theta(\tilde{\mathbf{n}} \mid \mathbf{n}, \ldots)}_{\substack{\text{coarsening} \\ \text{mechanism}}} \underbrace{\mathbb{P}_\theta(\mathbf{n} \mid \ldots)}_{\substack{\text{measurement} \\ \text{distribution}}}.$$

Fuzzy counts as coarsened data

Fuzziness requires CNAR

This argument leads to coarsening-not-at-random (**CNAR**).

As in MNAR problems [Molenberghs and Verbeke, 2005], a similar factorization arises in this context:

$$\mathbb{P}_\theta(\mathbf{n}, \tilde{\mathbf{n}} \mid \ldots) = \underbrace{\mathbb{P}_\theta(\tilde{\mathbf{n}} \mid \mathbf{n}, \ldots)}_{\substack{\text{coarsening} \\ \text{mechanism}}} \underbrace{\mathbb{P}_\theta(\mathbf{n} \mid \ldots)}_{\substack{\text{measurement} \\ \text{distribution}}}.$$

- Fuzziness affecting RNA-seq counts can be properly modelled here.

Analysing RNA-seq fuzzy counts

Toy Case: Chromosome X (Homo sapiens)

Data

- 2.71×10^7 reads (over twenty-seven million) from chromosome X (*Homo sapiens*), aligned to the GRCh38 reference genome using HISAT2 [Pertea et al., 2016]
- 12 samples (6 females)
- 18,866 fuzzy counts, corresponding to the number of annotated genes

Analysing RNA-seq fuzzy counts

Toy Case: Chromosome X (Homo sapiens)

Model

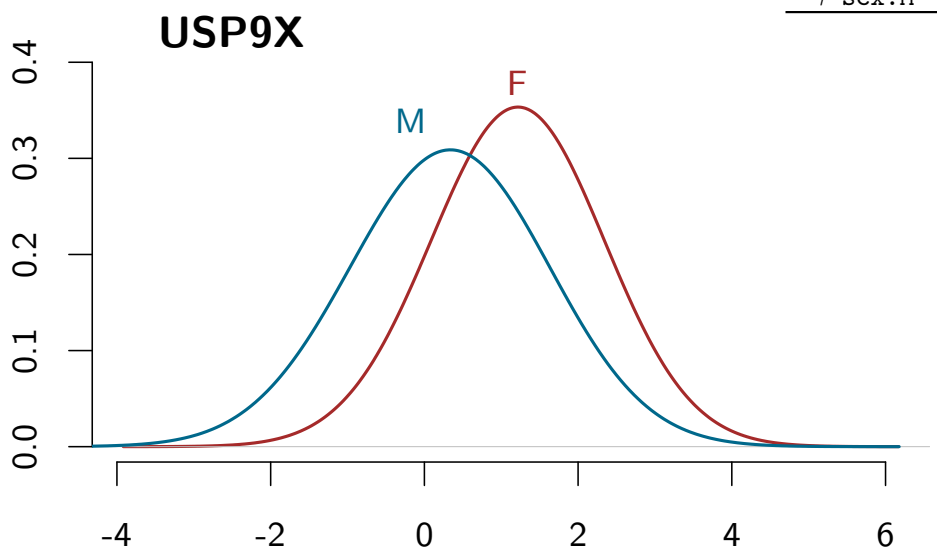
- Gene selected: *USP9X* playing roles in protein degradation, cell signaling, and neural development
- $n \sim \mathcal{P}oi(n; \lambda)$, with $\lambda = \exp\left(\beta_0 + \mathtt{sex}\beta_{\mathtt{sex}}\right)$
- Estimate $\{\beta_0, \beta_{\mathtt{sex}}\}$ using fuzzy counts $\tilde{\mathbf{n}}$
  via MCMC (MH adaptive) with $4 \times 9e3$ samples (burnin: $2e3$)

Analysing RNA-seq fuzzy counts

Toy Case: Chromosome X (Homo sapiens)

Results

	mean	sd	HPDI	
			lower	upper
$\beta_{\texttt{sex:F}}$	1.207	0.525	0.165	2.208
$\beta_{\texttt{sex:M}}$	0.321	0.810	-1.245	1.884



Conclusions

- Read-to-gene alignment induces **epistemic** (not purely stochastic) uncertainty
  - represent RNA-seq counts as **fuzzy numbers** $\tilde{n}$
- Fuzziness is a form of coarsening, but standard CAR assumptions imply n-independent coarsening probabilities
  - $\xi_{\tilde{n}}$ introduce n-dependence violating CAR
  - fuzziness needs to be treated as CNAR
- Hierarchical models can be used to explicitly specify the CNAR mechanism
  - tailor-made to handle multiple genes at the same time

[Consiglio et al., 2016] Consiglio, A., Mencar, C., Grillo, G., Marzano, F., Caratozzolo, M. F., and Liuni, S. (2016).
A fuzzy method for RNA-Seq differential expression analysis in presence of multireads.
BMC bioinformatics, 17:95–110.

[Deschamps-Francoeur et al., 2020] Deschamps-Francoeur, G., Simoneau, J., and Scott, M. S. (2020).
Handling multi-mapped reads in rna-seq.
Computational and structural biotechnology journal, 18:1569–1576.

[Gebhardt et al., 1998] Gebhardt, J., Gil, M. A., and Kruse, R. (1998).
Fuzzy set-theoretic methods in statistics.
In *Fuzzy sets in decision analysis, operations research and statistics*, pages 311–347. Springer.

[Gill and Grünwald, 2008] Gill, R. D. and Grünwald, P. D. (2008).
An algorithmic and a geometric characterization of coarsening at random.
The Annals of Statistics, 36(5):2409–2422.

[Ji et al., 2011] Ji, Y., Xu, Y., Zhang, Q., Tsui, K.-W., Yuan, Y., Norris Jr, C., Liang, S., and Liang, H. (2011).
Bm-map: Bayesian mapping of multireads for next-generation sequencing data.
Biometrics, 67(4):1215–1224.

[Mencar and Pedrycz, 2020] Mencar, C. and Pedrycz, W. (2020).
Granular counting of uncertain data.
Fuzzy Sets and Systems, 387:108–126.

[Molenberghs and Verbeke, 2005] Molenberghs, G. and Verbeke, G. (2005).
Models for discrete longitudinal data.
Springer.

[Pertea et al., 2016] Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016).
Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown.
Nature protocols, 11(9):1650–1667.

[Tanaka et al., 1977] Tanaka, H., Okuda, T., and Asai, K. (1977).
On decision-making in fuzzy environment fuzzy information and decision making.
The international journal of production research, 15(6):623–635.

antonio.calcagni@unipd.it
https://unipd.link/acalcagni