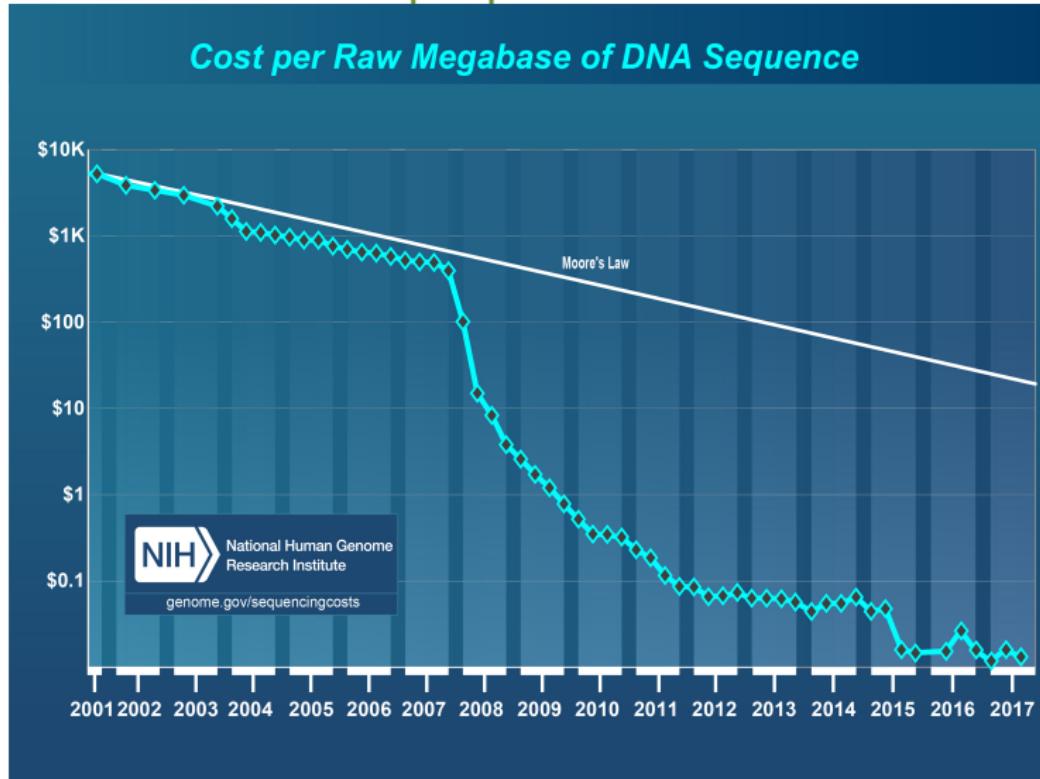


# Omnibus testing and post-hoc tests for RNA-seq

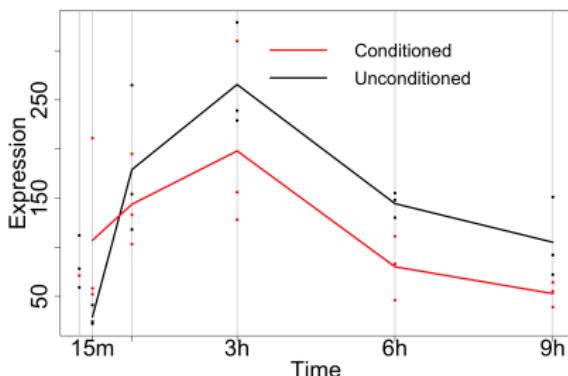
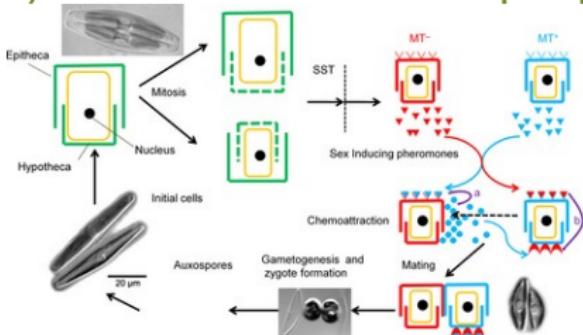
Lieven Clement

Statistical Genomics: Master of Science in Bioinformatics

# (R)evolution in RNA-seq experiments

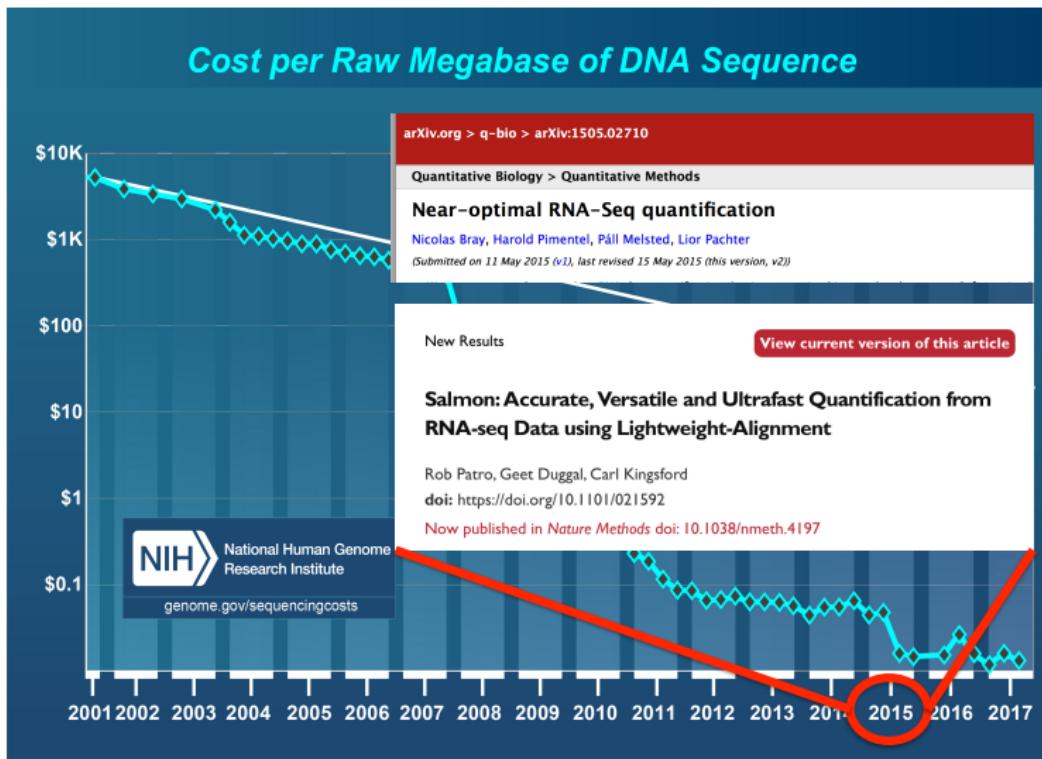


# (R)evolution in RNA-seq experiments

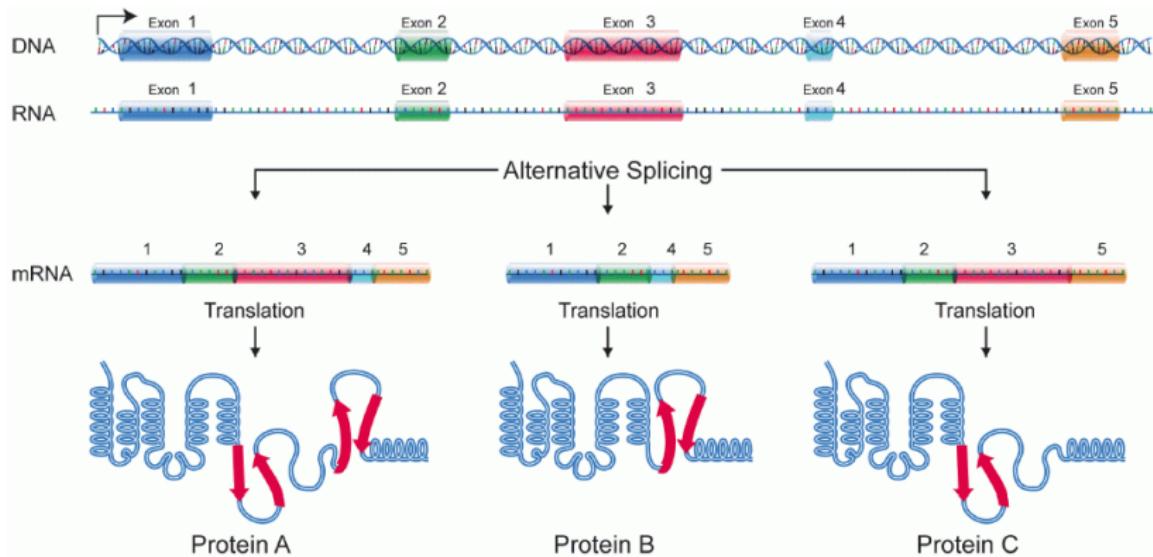


- Declining sequencing cost → experiments with complex designs
- Complex designs → multiple hypotheses of interest:
  - ① Is gene DE in different conditions?
  - ② Does the DE pattern change over time?  
→ **To be assessed for thousands of genes!**

# State-of-the-art RNA-seq tools allow transcript-level analysis

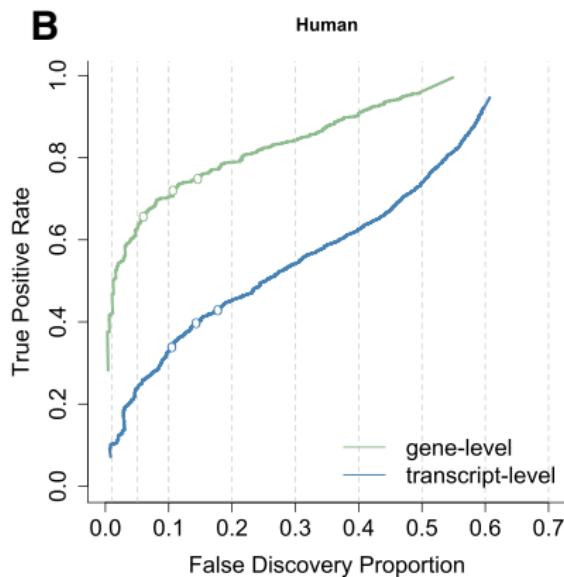


# State-of-the-art RNA-seq tools allow transcript-level analysis



[https://en.wikibooks.org/wiki/Proteomics/Protein\\_Primary\\_Structure/Alternative\\_Splicing](https://en.wikibooks.org/wiki/Proteomics/Protein_Primary_Structure/Alternative_Splicing)

# Power Issue Transcript Level Analysis



Van den berge et al. 2017 Genome Biology 18:151

Human: > 38000 genes and > 173000 transcripts



# Single cell transcriptomics



## ARTICLE

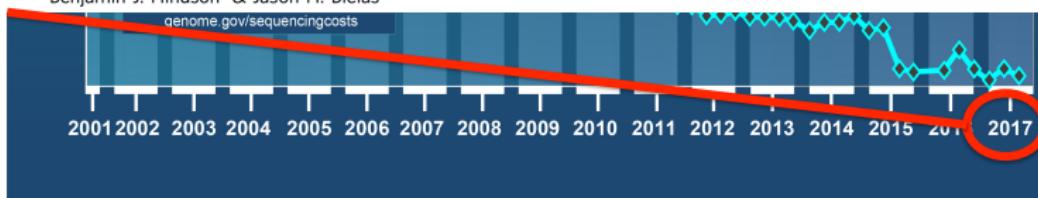
Received 20 Sep 2016 | Accepted 23 Nov 2016 | Published 16 Jan 2017

DOI: 10.1038/ncomms14049

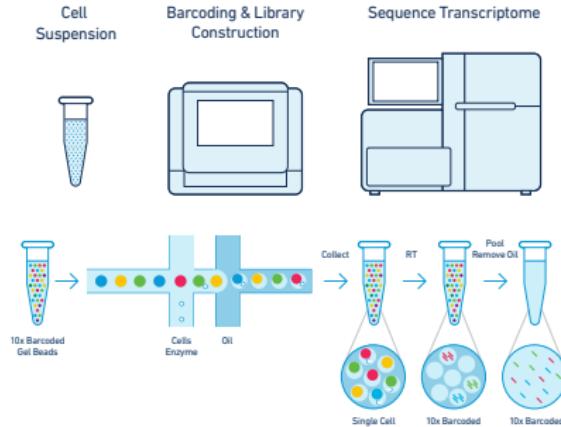
OPEN

## Massively parallel digital transcriptional profiling of single cells

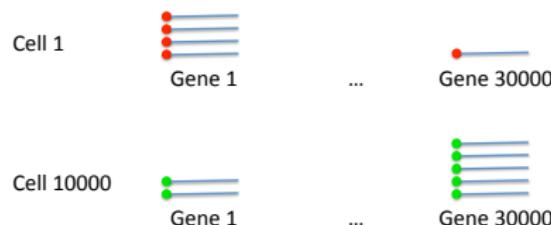
Grace X.Y. Zheng<sup>1</sup>, Jessica M. Terry<sup>1</sup>, Phillip Belgrader<sup>1</sup>, Paul Ryvkin<sup>1</sup>, Zachary W. Bent<sup>1</sup>, Ryan Wilson<sup>1</sup>, Solongo B. Ziraldo<sup>1</sup>, Tobias D. Wheeler<sup>1</sup>, Geoff P. McDermott<sup>1</sup>, Junjie Zhu<sup>1</sup>, Mark T. Gregory<sup>2</sup>, Joe Shuga<sup>1</sup>, Luz Montesclaros<sup>1</sup>, Jason G. Underwood<sup>1,3</sup>, Donald A. Masquelier<sup>1</sup>, Stefanie Y. Nishimura<sup>1</sup>, Michael Schnall-Levin<sup>1</sup>, Paul W. Wyatt<sup>1</sup>, Christopher M. Hindson<sup>1</sup>, Rajiv Bharadwaj<sup>1</sup>, Alexander Wong<sup>1</sup>, Kevin D. Ness<sup>1</sup>, Lan W. Beppu<sup>4</sup>, H. Joachim Deeg<sup>4</sup>, Christopher McFarland<sup>5</sup>, Keith R. Loeb<sup>4,6</sup>, William J. Valente<sup>2,7,8</sup>, Nolan G. Ericson<sup>2</sup>, Emily A. Stevens<sup>4</sup>, Jerald P. Radich<sup>4</sup>, Tarjei S. Mikkelsen<sup>1</sup>, Benjamin J. Hindson<sup>1</sup> & Jason H. Bielas<sup>2,6,8,9</sup>

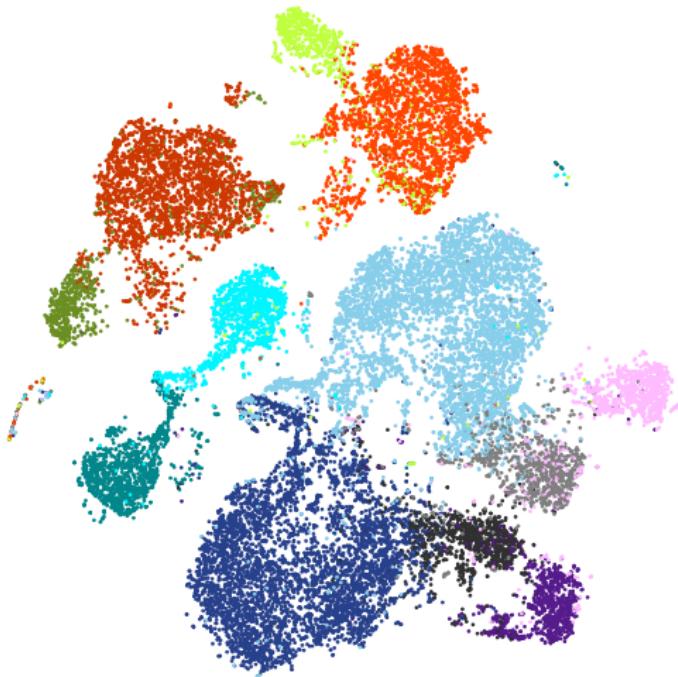


# Single cell transcriptomics



Transcriptome profile for each individual cell





Kang et al. Nat. Biotechnol. 2018 36(1):89-94

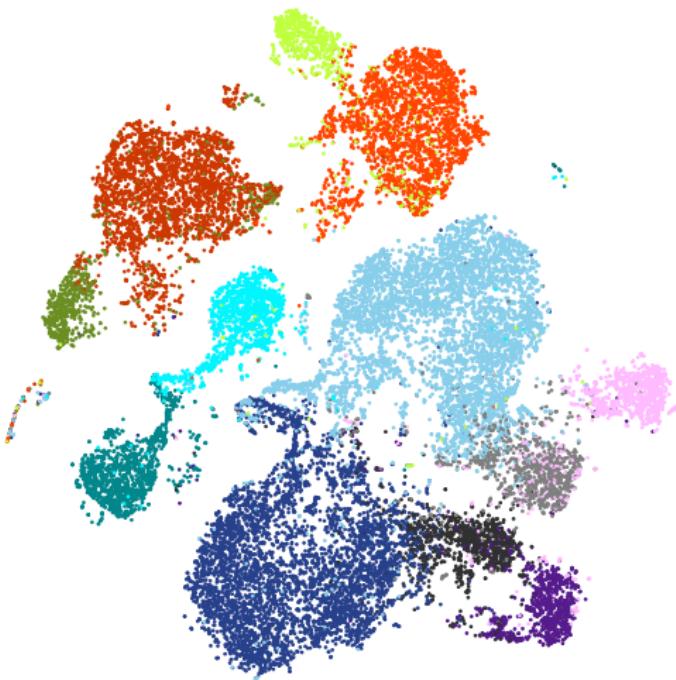
- peripheral blood mononuclear cells
- from 8 individuals
- Stimulated vs control
- > 29000 cells

Stimulated	Control
● NK cells	● NK cells
● FCGR3A+ Monocytes	● FCGR3A+ Monocytes
● CD8 T cells	● CD8 T cells
● CD4 T cells	● CD4 T cells
● CD14+ Monocytes	● CD14+ Monocytes
● B cells	● B cells



Kang et al. Nat. Biotechnol. 2018 36(1):89-94

- peripheral blood mononuclear cells
- from 8 individuals
- Stimulated vs control
- > 29000 cells
- Two channels of 10x genomics chip
- Two lanes of hiseq run
- Demultiplexing individuals via SNPs



Kang et al. Nat. Biotechnol. 2018 36(1):89-94

- DE stimulated vs control in each cell type (6 tests/gene)
- Different stimulus effect across cell types (15 tests/gene)

# Many hypotheses per gene in contemporary RNA-seq

Transcript-level analysis, single cell experiments and complex designs result in multiple hypotheses of interest per gene.

The conventional strategy

- ① assess each hypothesis separately
- ② on FDR level  $\alpha$
- ③ provide the biologist with list of top-genes for every contrast

# Many hypotheses per gene in contemporary RNA-seq

Transcript-level analysis, single cell experiments and complex designs result in multiple hypotheses of interest per gene.

The conventional strategy

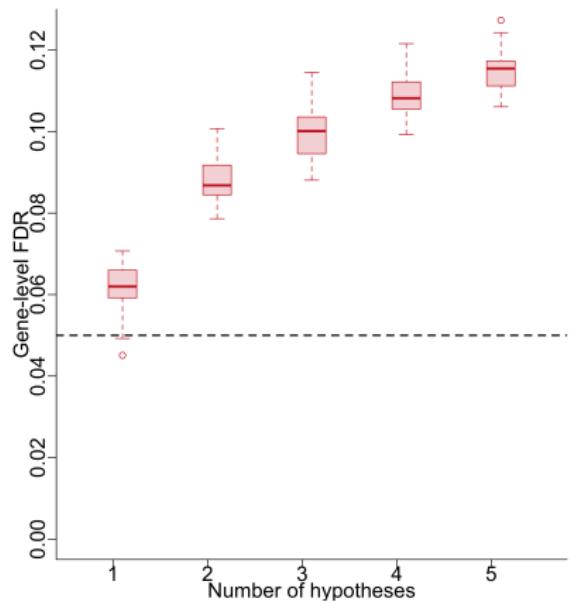
- ① assess each hypothesis separately
- ② on FDR level  $\alpha$
- ③ provide the biologist with list of top-genes for every contrast

However,

- Shortlist of interesting genes when we assess multiple hypotheses per gene?
- Post-hoc tests for each hypothesis within a gene if omnibus null hypothesis is rejected?
- Gene-level FDR control required because downstream analysis and validation is done at the gene-level.



# Simulation study conventional analysis

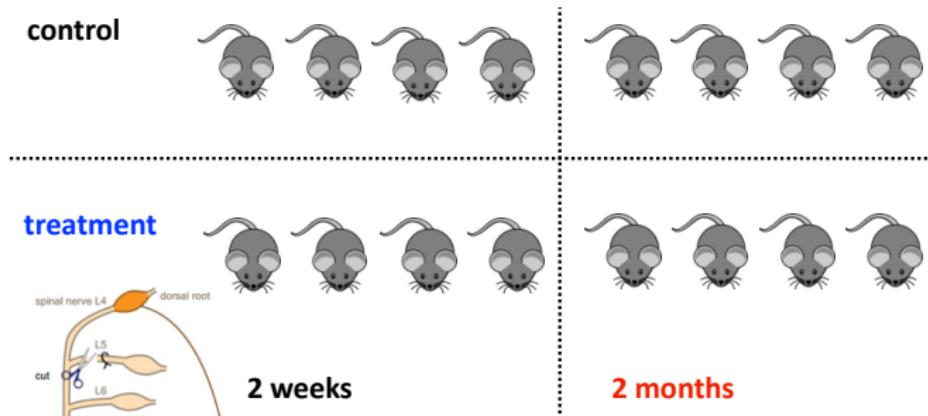


# Example

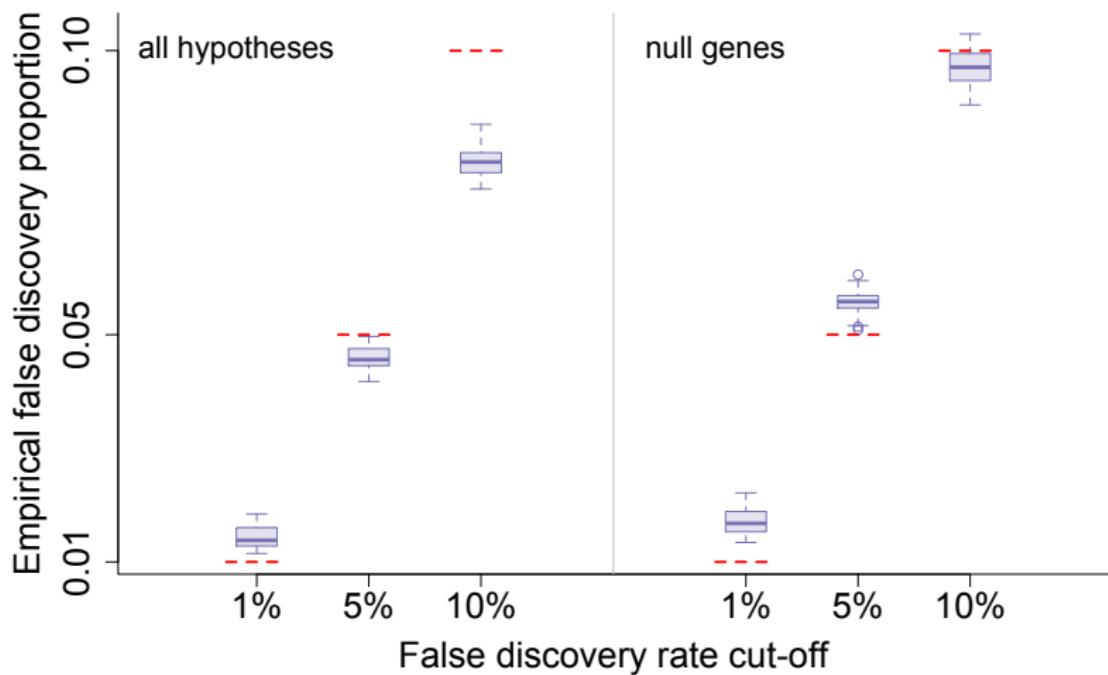
- Based on Hammer et al. (2010), Genome Research
- Two conditions (control - SNL)
- Two timepoints (2 weeks - 2 months)

Interested in:

- ① DE between conditions at 2 weeks (> 7000 DE genes)
- ② DE between conditions at 2 months (> 6500 DE genes)
- ③ Different FC between timepoints (interaction, 0 ΔFC genes)



## Example: Gene-level tests



## Control FDR on gene level by aggregated testing

A simple strategy would be to

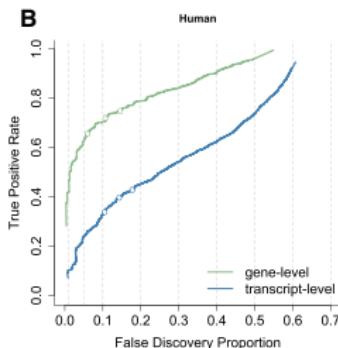
- ① Aggregate p-values across hypotheses (i.e. omnibus test)
- ② Control FDR on level  $\alpha_I$  on the aggregated p-values

# Control FDR on gene level by aggregated testing

A simple strategy would be to

- ① Aggregate p-values across hypotheses (i.e. omnibus test)
- ② Control FDR on level  $\alpha_I$  on the aggregated p-values

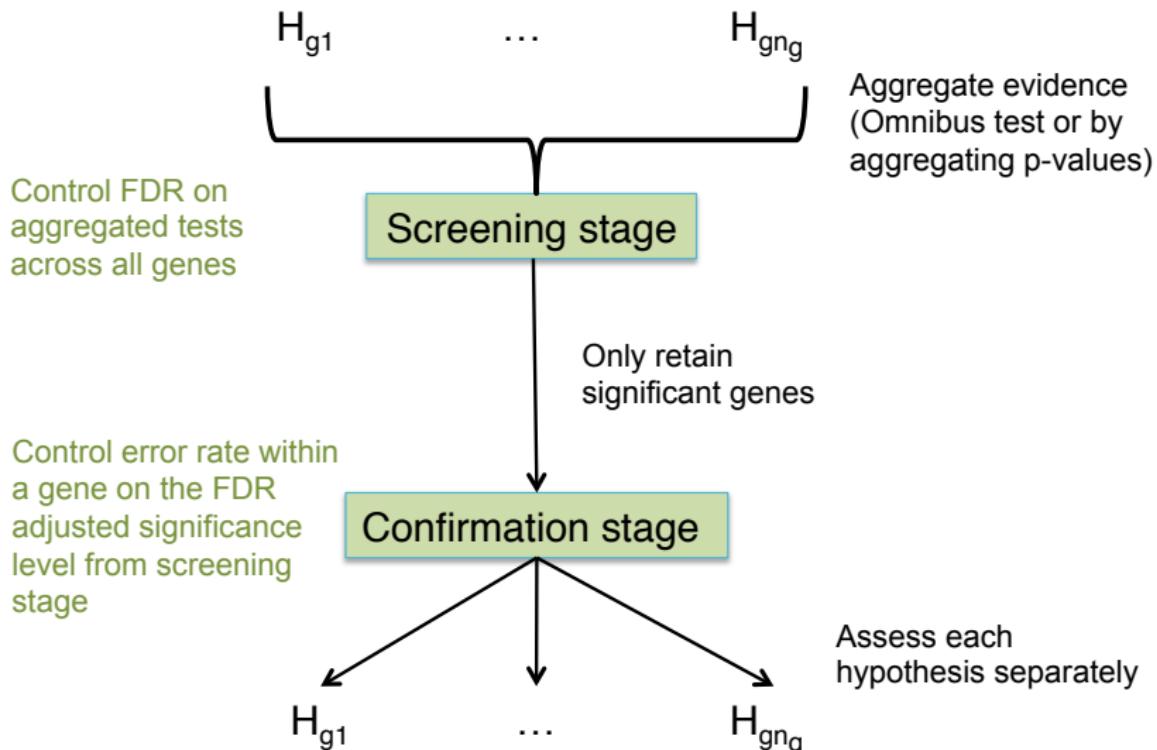
Additionally takes advantage of aggregated tests with higher sensitivity



However, we lose resolution on the biology



## Solution: Stage-wise testing procedure: aggregate and split evidence



# Stage-wise testing procedure<sup>1</sup>

## ① Screening Stage:

- Assess the screening hypothesis  $H_g^S$  / global null hypothesis for all genes in the set  $G$ .
- Apply the Benjamini Hochberg (BH) FDR procedure to the screening p-values at FDR level  $\alpha$ . Let  $R$  be the number of rejected screening hypotheses.

---

<sup>1</sup>Heller et al. 2009, Bioinformatics.



# Stage-wise testing procedure<sup>1</sup>

## ① Screening Stage:

- Assess the screening hypothesis  $H_g^S$  / global null hypothesis for all genes in the set  $G$ .
- Apply the Benjamini Hochberg (BH) FDR procedure to the screening p-values at FDR level  $\alpha$ . Let  $R$  be the number of rejected screening hypotheses.

## ② Confirmation Stage: For all $R$ genes that pass the screening stage.

- Let  $\alpha_{II} = R\alpha/G$  be FDR-adjusted significance level from the first stage.
- Adopt a multiple testing procedure to assess all  $n_g$  hypotheses while controlling the within gene error rate at the adjusted level  $\alpha_{II}$ .

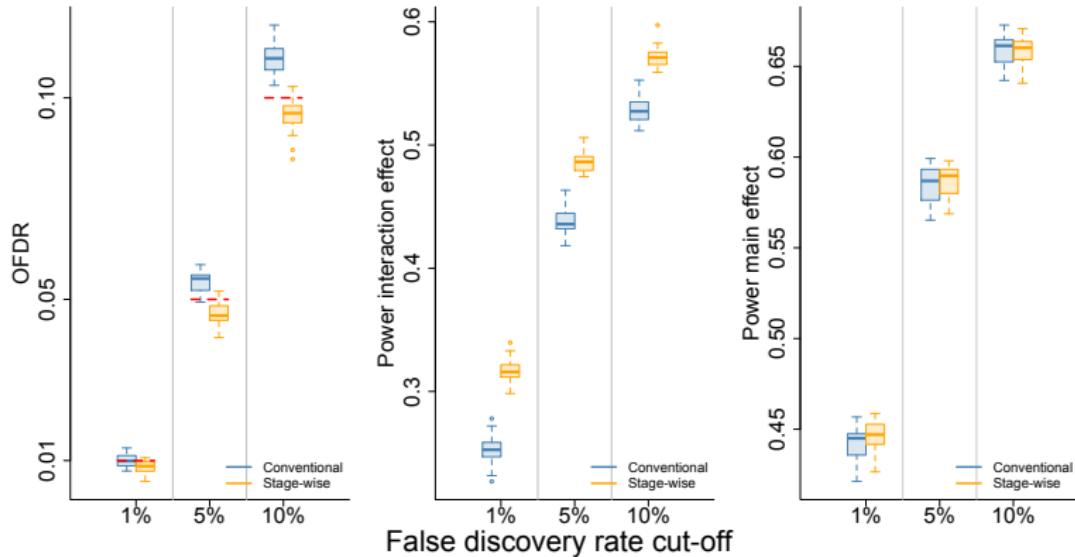
---

<sup>1</sup>Heller et al. 2009, Bioinformatics.



# DGE experiments with complex designs

- Our procedure correctly controls the FDR at gene-level
- The omnibus test enriches for genes with interaction effects
- While maintaining equivalent power for main effects



# Stage-wise testing unlocks powerful transcript-level analysis

- Naturally unites high gene-level power with transcript-level resolution of the results
- Equal or better power at transcript level
- Better FDR control

