UNIVERSITEIT GENT

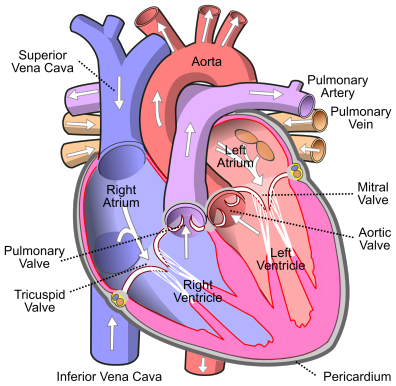
BIG N2N
Bioinformatics Institute Ghent
from nucleotides to networks

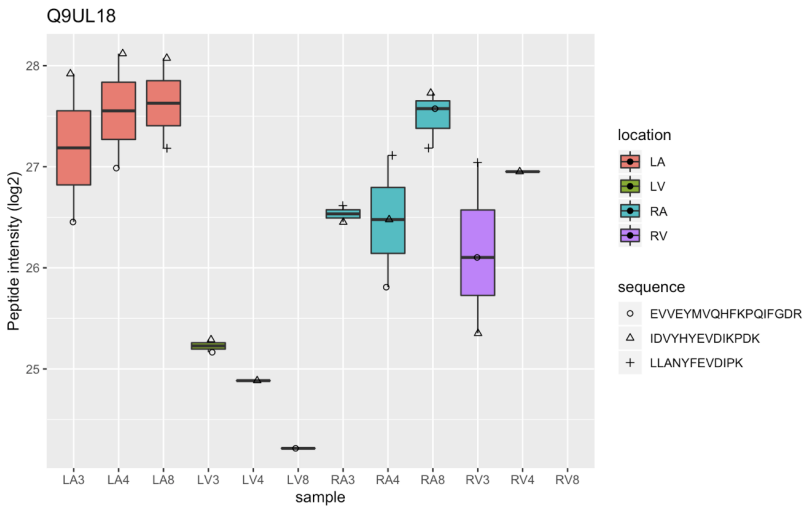
Omnibus testing and post-hoc tests for high throughput experiments

Lieven Clement

Proteomics and transcriptomics data analysis

(R)evolution in high throughput experiments



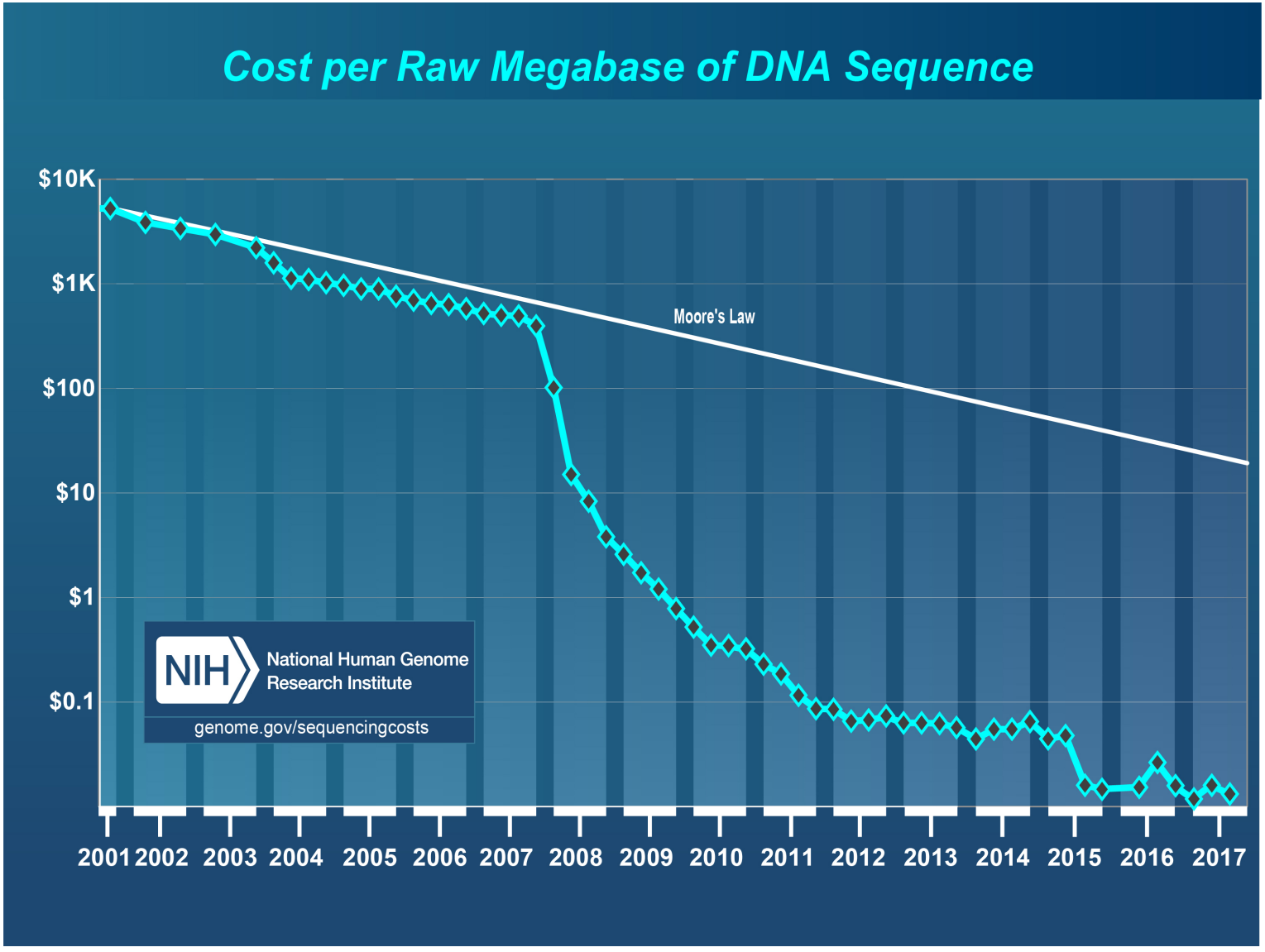


- Higher throughput and Declining costs $\rightarrow$ experiments with complex designs

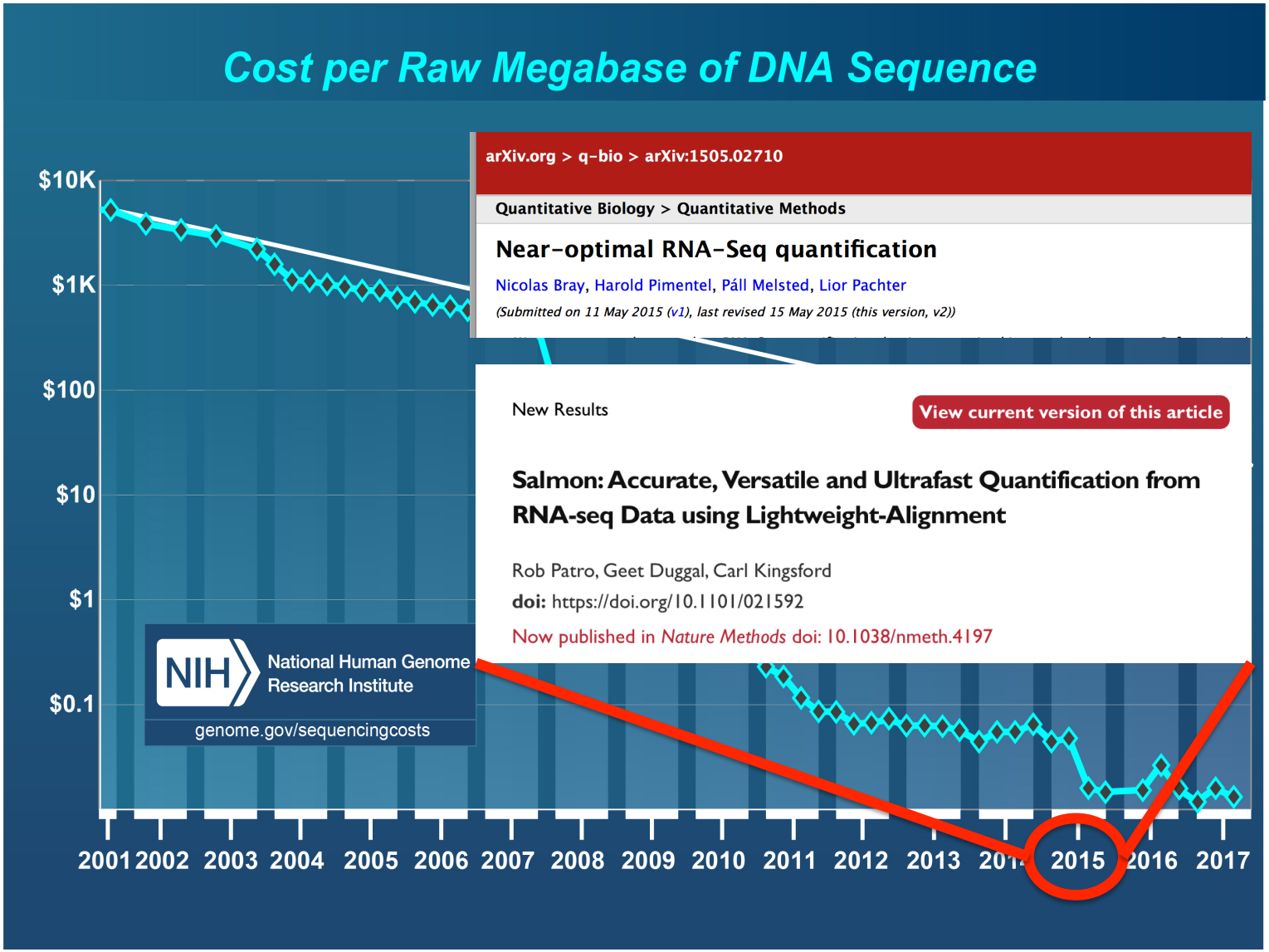
- Complex designs $\rightarrow$ multiple hypotheses of interest:
  1. Is protein DA in different heart regions?
  2. Does the DA pattern changes left to right?

  $\rightarrow$ **To be assessed for thousands of proteins!**

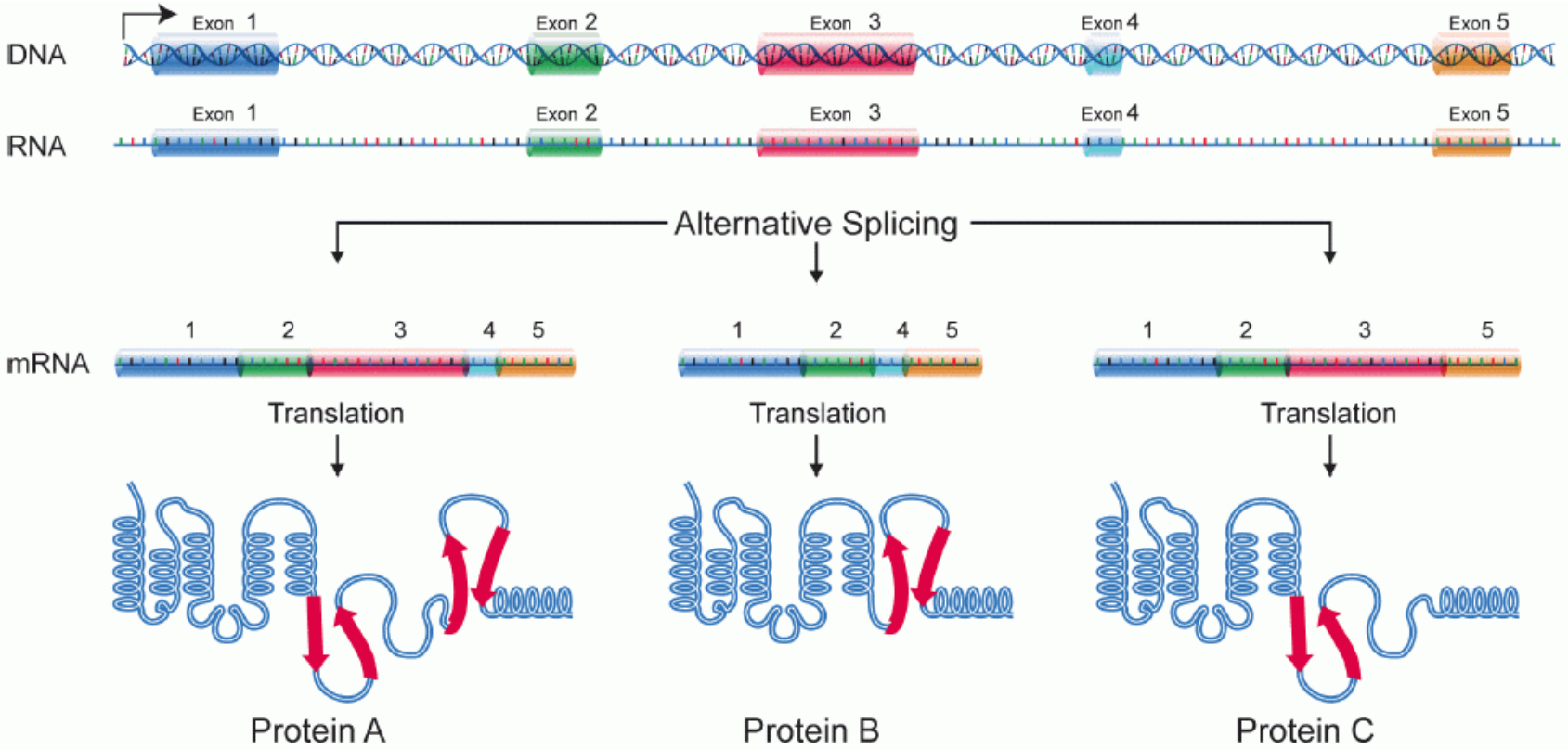
(R)evolution in high throughput experiments



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

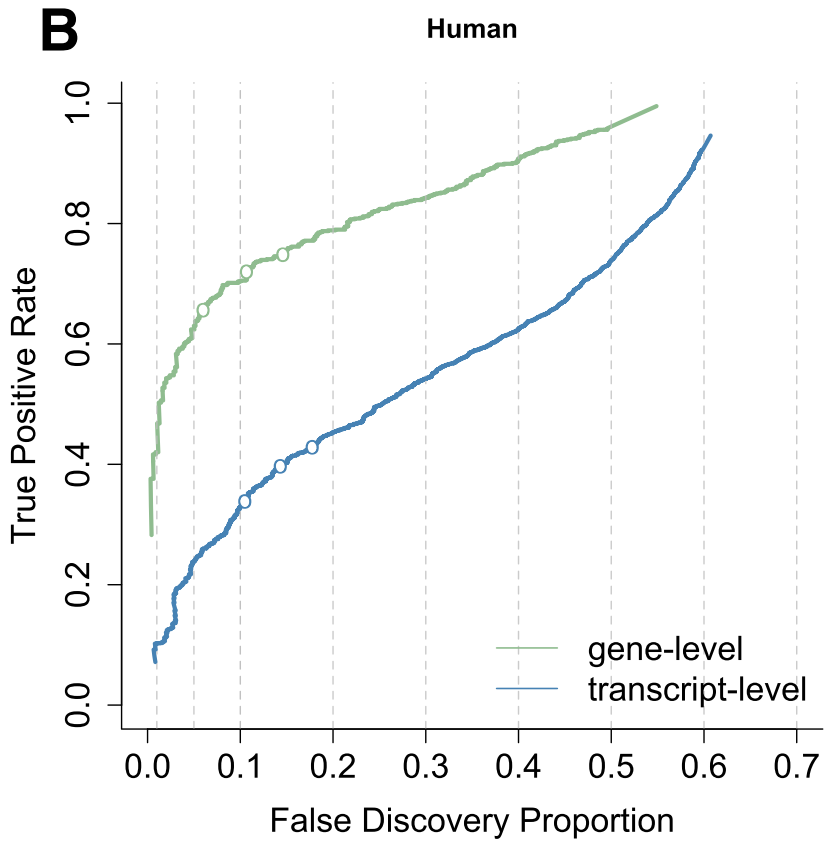


State-of-the-art RNA-seq tools allow transcript-level analysis



https://en.wikibooks.org/wiki/Proteomics/Protein_Primary_Structure/Alternative_Splicing

Power Issue Transcript Level Analysis



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

Human: $>$ 38000 genes and $>$ 173000 transcripts

Single cell transcriptomics

nature COMMUNICATIONS

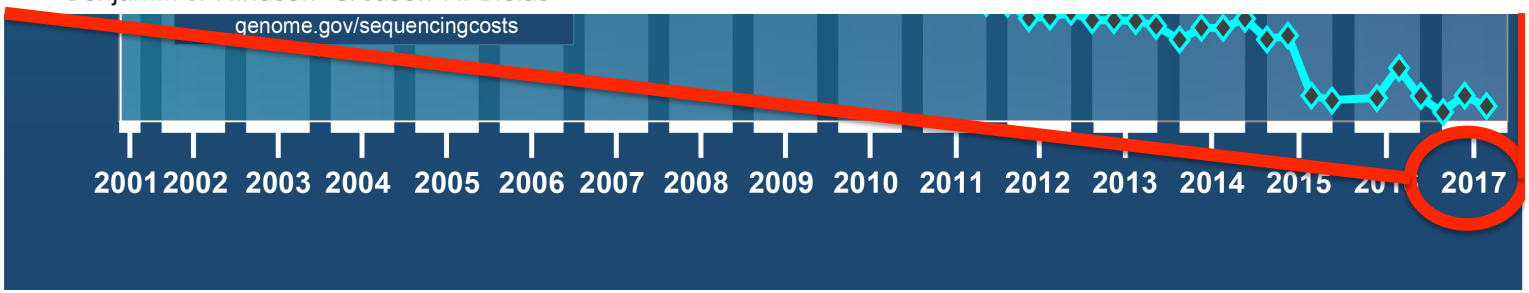
ARTICLE

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

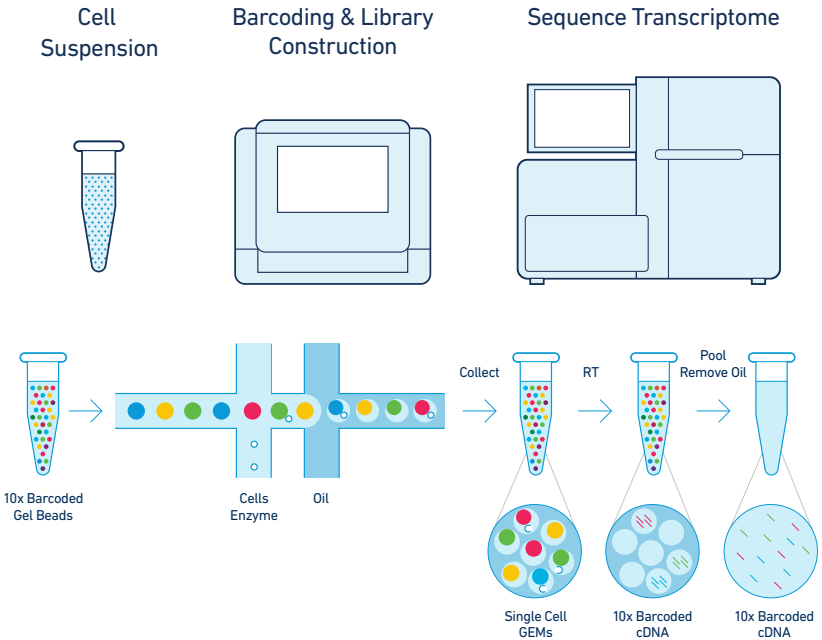
DOI: 10.1038/ncomms14049 OPEN

Massively parallel digital transcriptional profiling of single cells

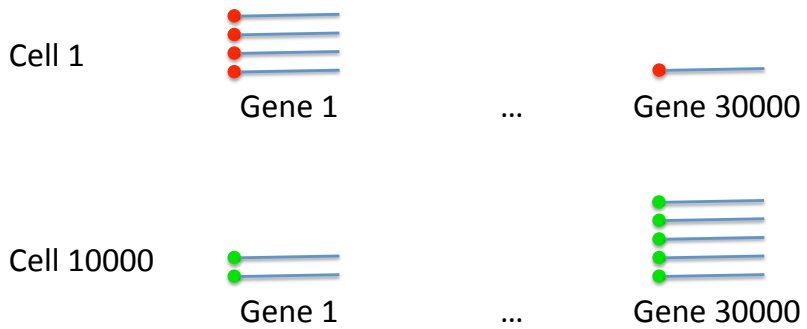
Grace X.Y. Zheng[1], Jessica M. Terry[1], Phillip Belgrader[1], Paul Ryvkin[1], Zachary W. Bent[1], Ryan Wilson[1], Solongo B. Ziraldo[1], Tobias D. Wheeler[1], Geoff P. McDermott[1], Junjie Zhu[1], Mark T. Gregory[2], Joe Shuga[1], Luz Montesclaros[1], Jason G. Underwood[1,3], Donald A. Masquelier[1],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

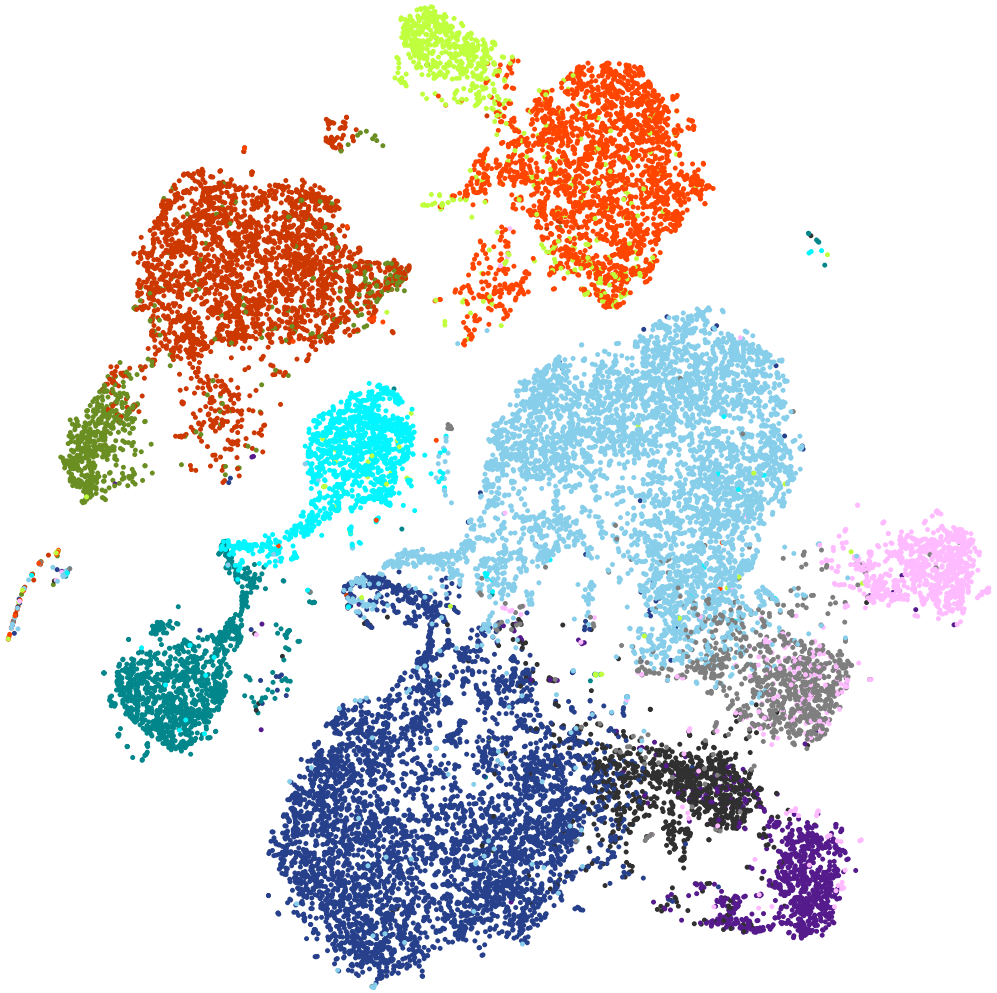


Single cell transcriptomics



Transcriptome profile for each individual cell

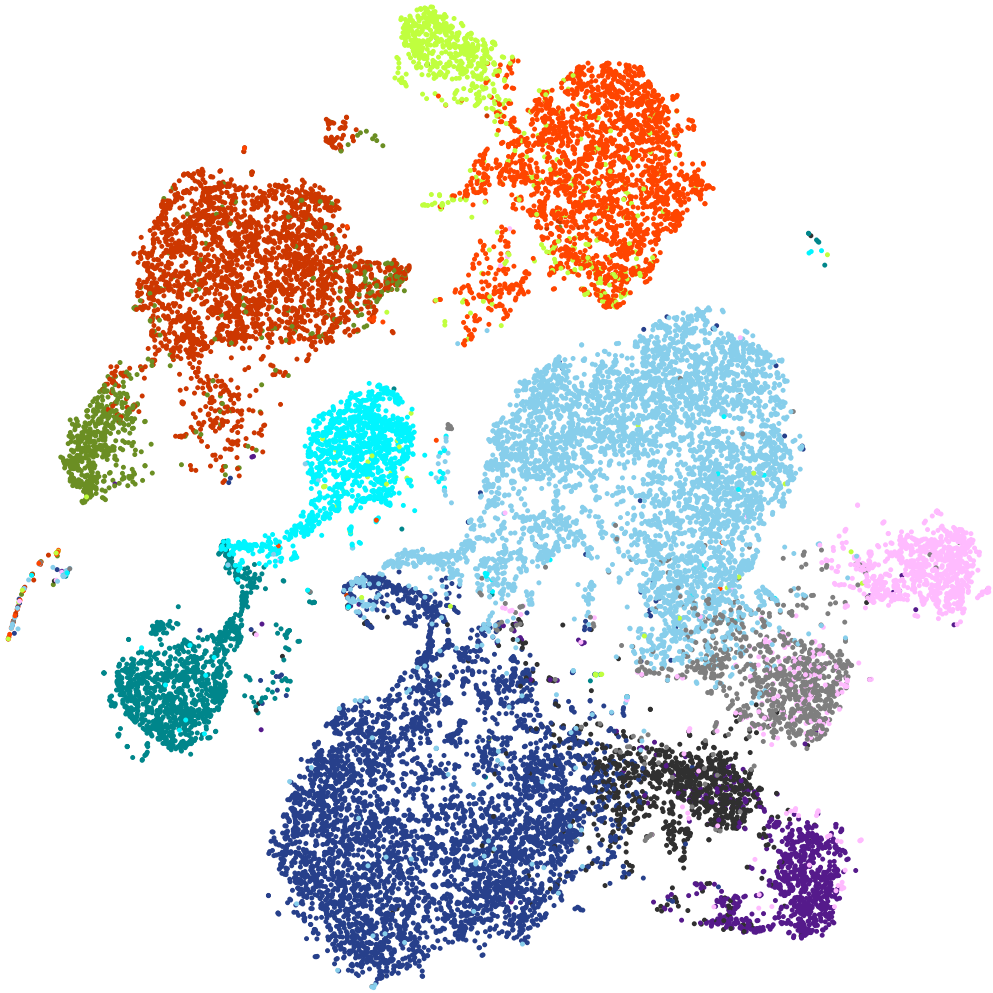




Kang et al. Nat. Biotechnol. 2018 36(1):89-94

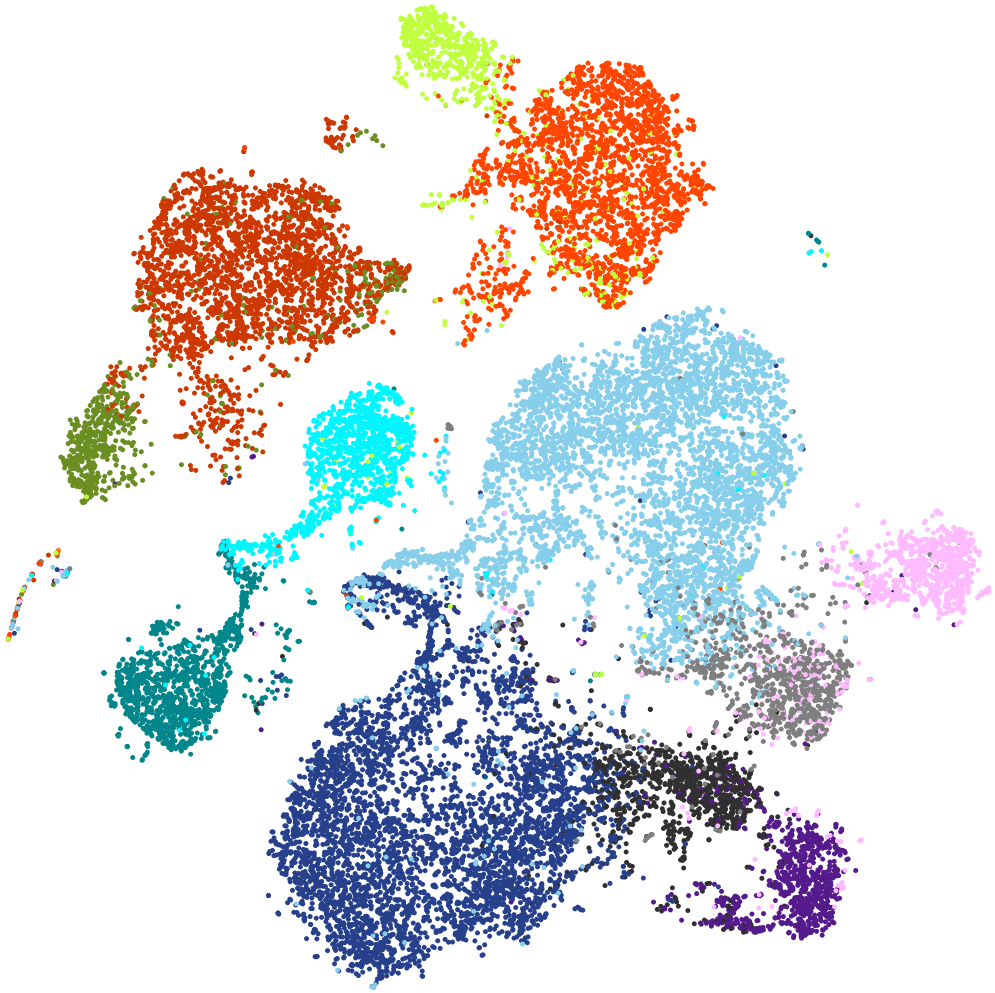
- peripheral blood mononuclear cells
- from 8 individuals
- Stimulated vs control
- $>$ 29000 cells

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/protein in contemporary high throughput studies

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

The conventional strategy

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

Many hypotheses per gene/protein in contemporary high throughput studies

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

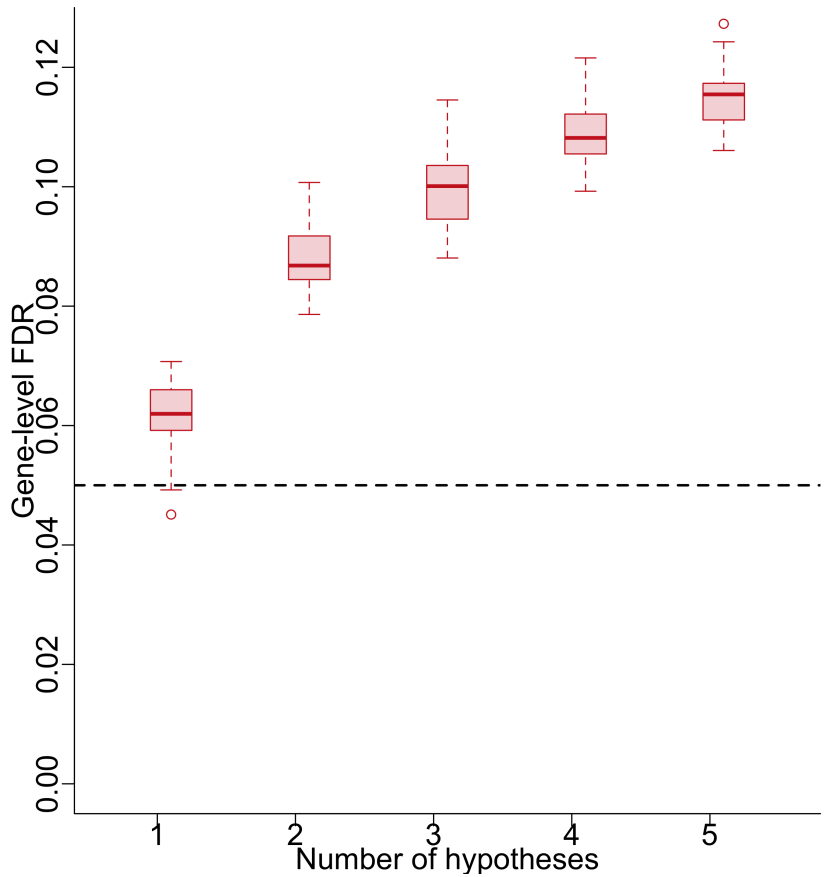
The conventional strategy

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

However,

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

Simulation study conventional analysis in sequencing applications

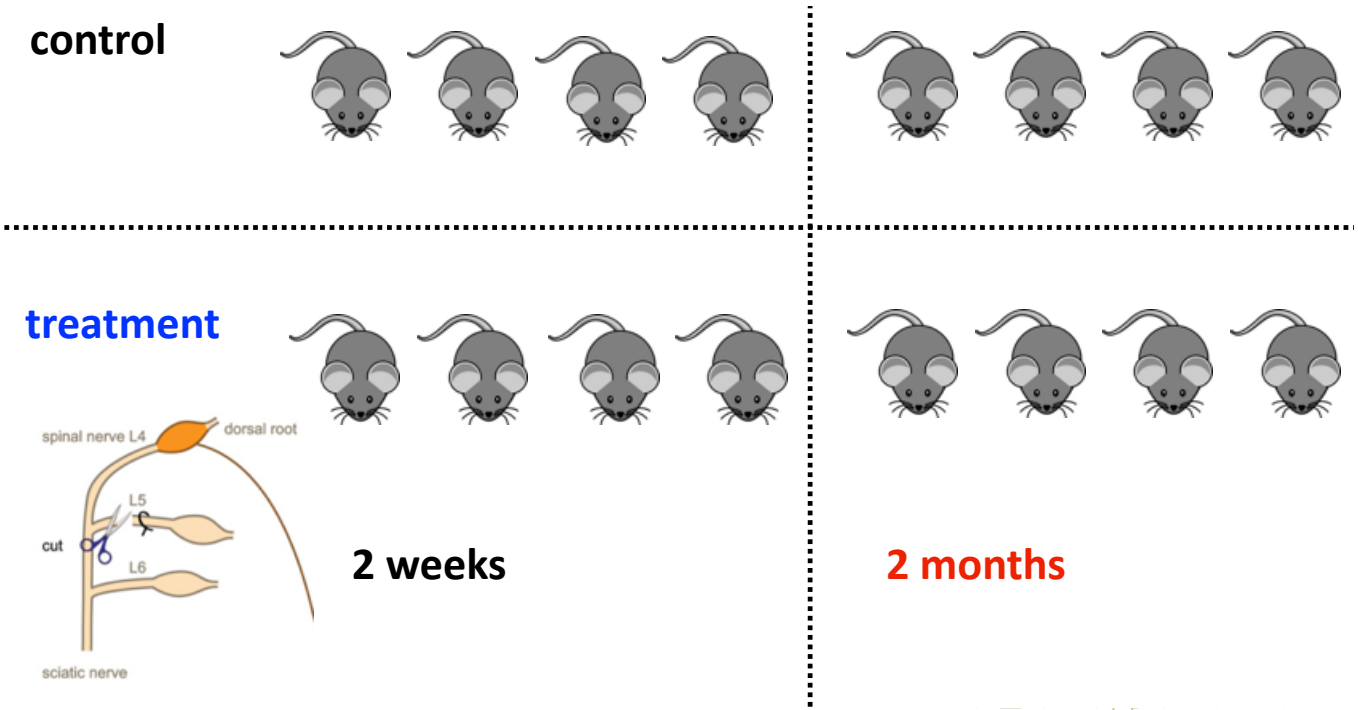


Example

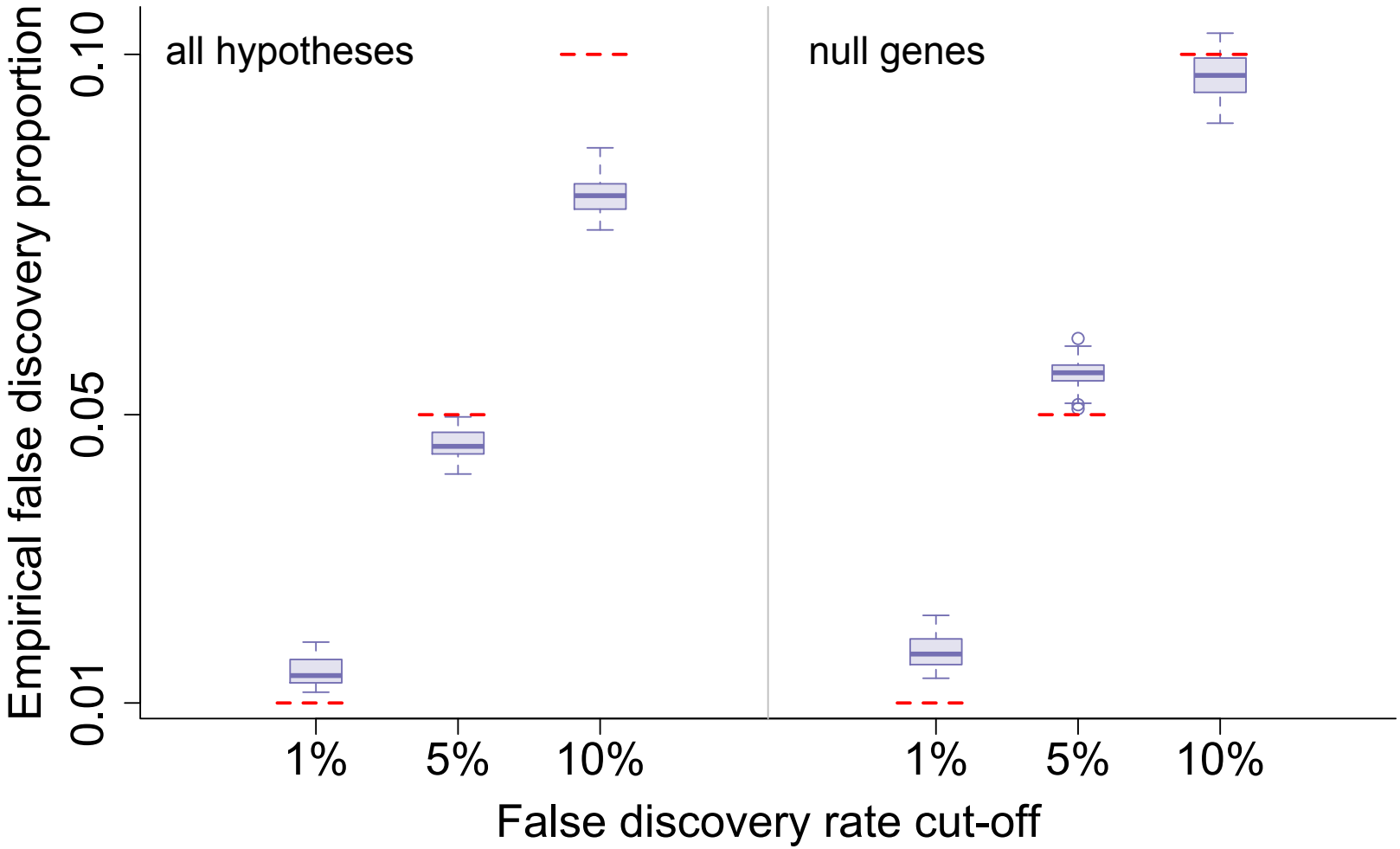
- Based on Hammer et al. (2010), Genome Research
- Two conditions (control - SNL)
- Two timepoints (2 weeks - 2 months)

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



Example transcript level analysis: control FDR on gene level by aggregated testing

A simple strategy would be to

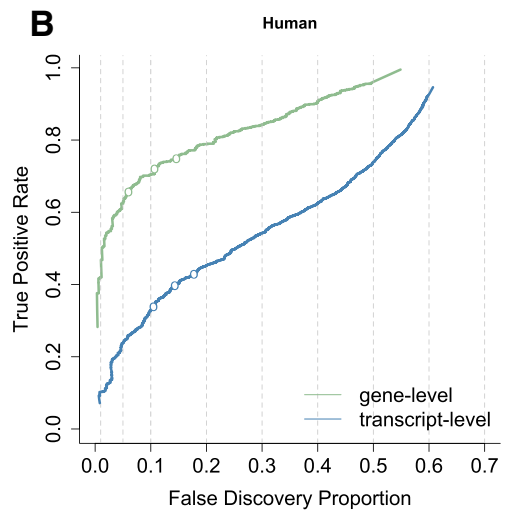
1. Aggregate p-values across hypotheses (i.e. omnibus test)
2. Control FDR on level α_I on the aggregated p-values

Example transcript level analysis: control FDR on gene level by aggregated testing

A simple strategy would be to

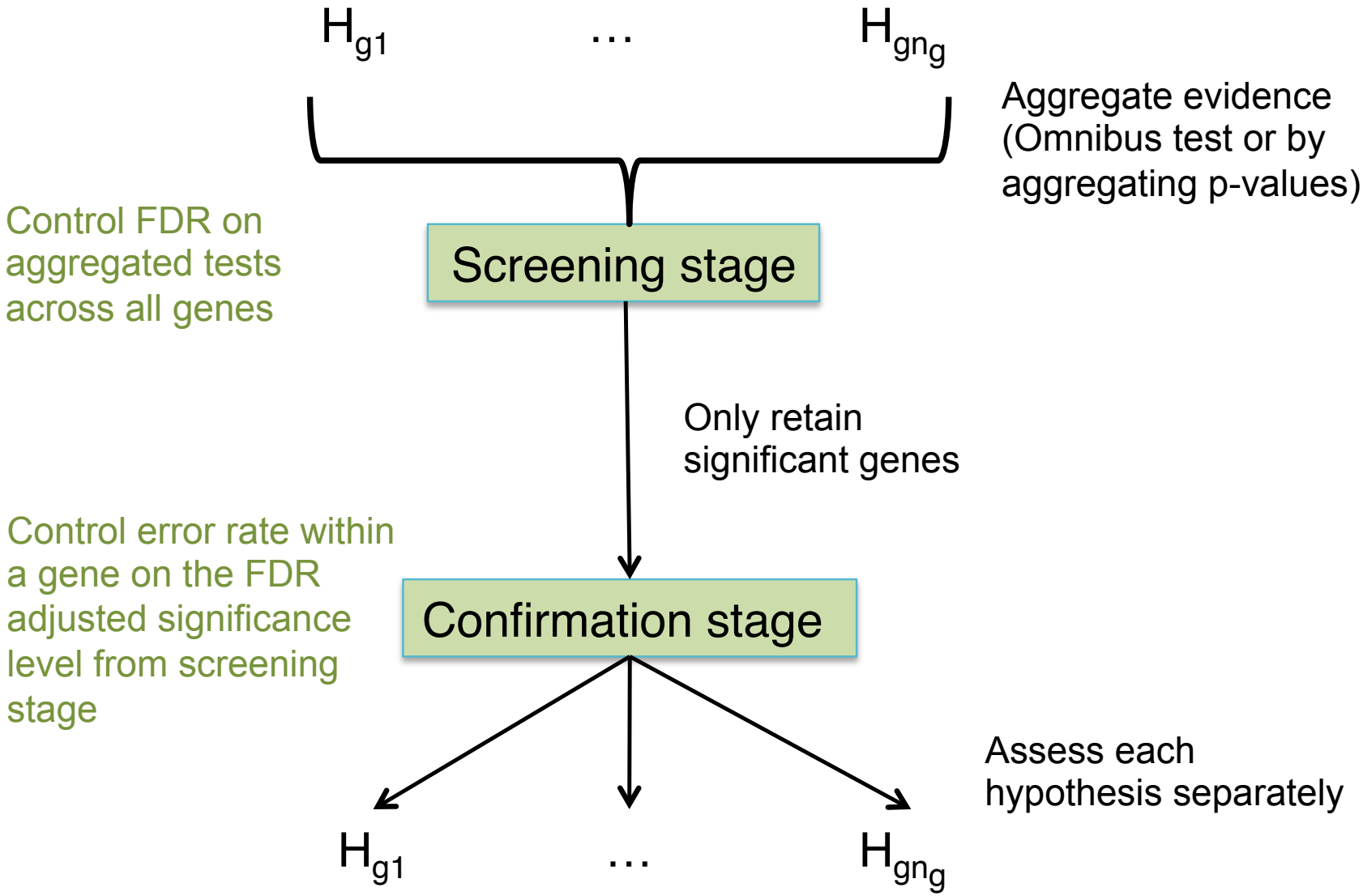
1. Aggregate p-values across hypotheses (i.e. omnibus test)
2. Control FDR on level α_I on the aggregated p-values

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[1]

1. **Screening Stage:**
   - Assess the screening hypothesis H_g^S/ global null hypothesis for all genes/proteins in the set G.
   - Apply the Benjamini Hochberg (BH) FDR procedure to the screening p-values at FDR level α. Let R be the number of rejected screening hypotheses.

[1] Heller et al. 2009, Bioinformatics.

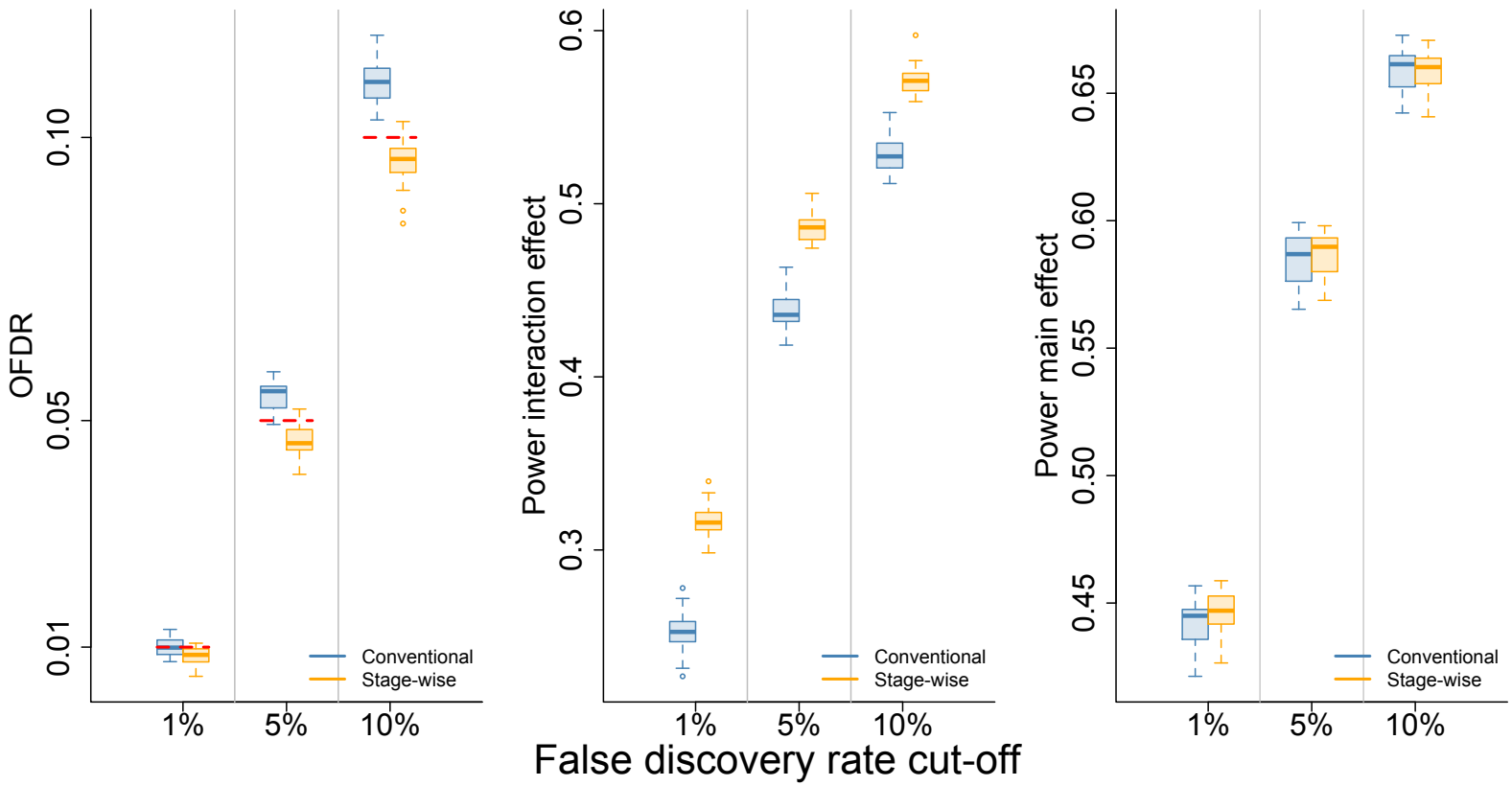
Stage-wise testing procedure[1]

1. **Screening Stage:**
   - Assess the screening hypothesis H_g^S/ global null hypothesis for all genes/proteins in the set G.
   - Apply the Benjamini Hochberg (BH) FDR procedure to the screening p-values at FDR level α. Let R be the number of rejected screening hypotheses.
2. **Confirmation Stage:** For all R genes/proteins 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 α_{II}.

[1] Heller et al. 2009, Bioinformatics.

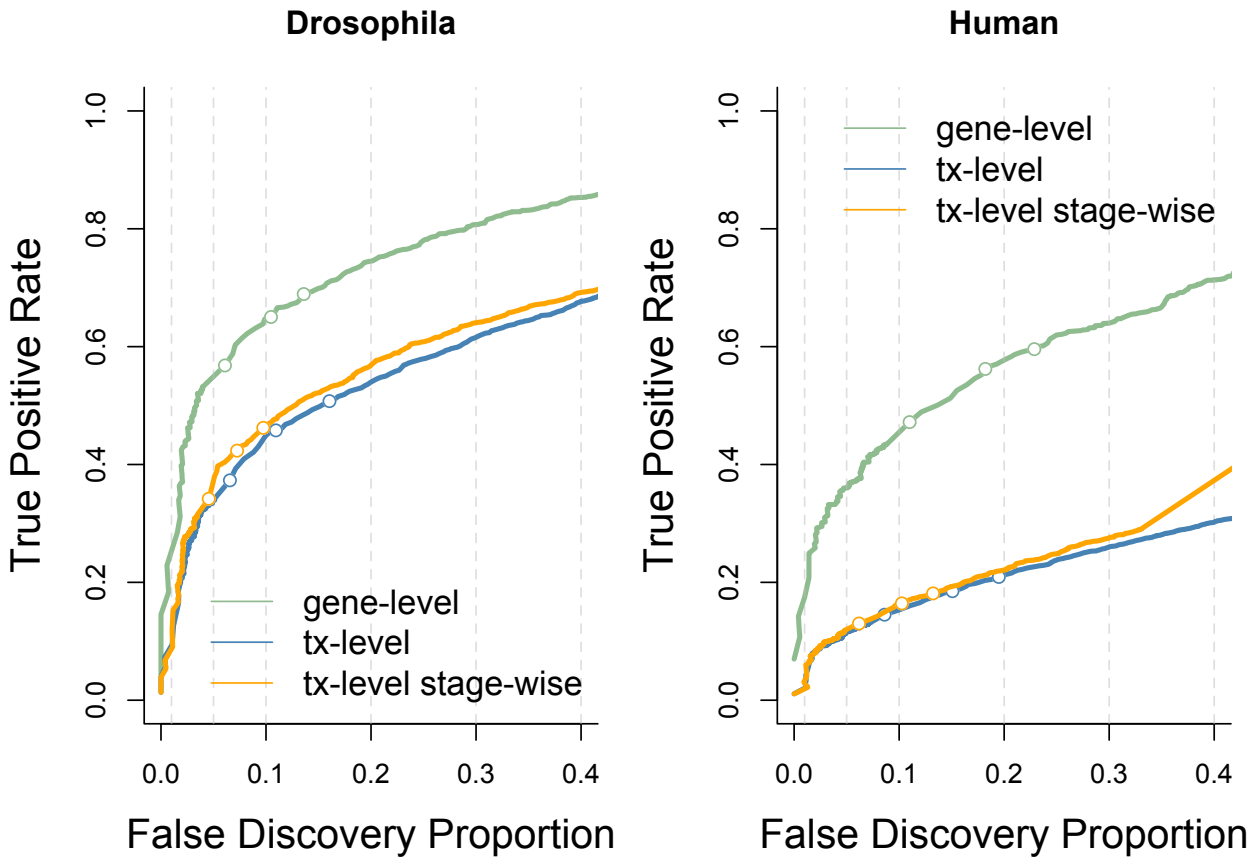
DGE experiments with complex designs

- 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



Van den Berge *et al. Genome Biology* (2017) 18:151
DOI 10.1186/s13059-017-1277-0

Genome Biology

METHOD Open Access

stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage

Koen Van den Berge[1,2], Charlotte Soneson[3,4], Mark D. Robinson[3,4] and Lieven Clement[1,2*]

Abstract

RNA sequencing studies with complex designs and transcript-resolution analyses involve multiple hypotheses per gene; however, conventional approaches fail to control the false discovery rate (FDR) at gene level. We propose stageR, a two-stage testing paradigm that leverages the increased power of aggregated gene-level tests and allows post hoc assessment for significant genes. This method provides gene-level FDR control and boosts power for testing interaction effects. In transcript-level analysis, it provides a framework that performs powerful gene-level tests while maintaining biological interpretation at transcript-level resolution. The procedure is applicable whenever individual hypotheses can be aggregated, providing a unified framework for complex high-throughput experiments.

Keywords: RNA-sequencing, Stage-wise testing, Differential transcript usage, Differential expression