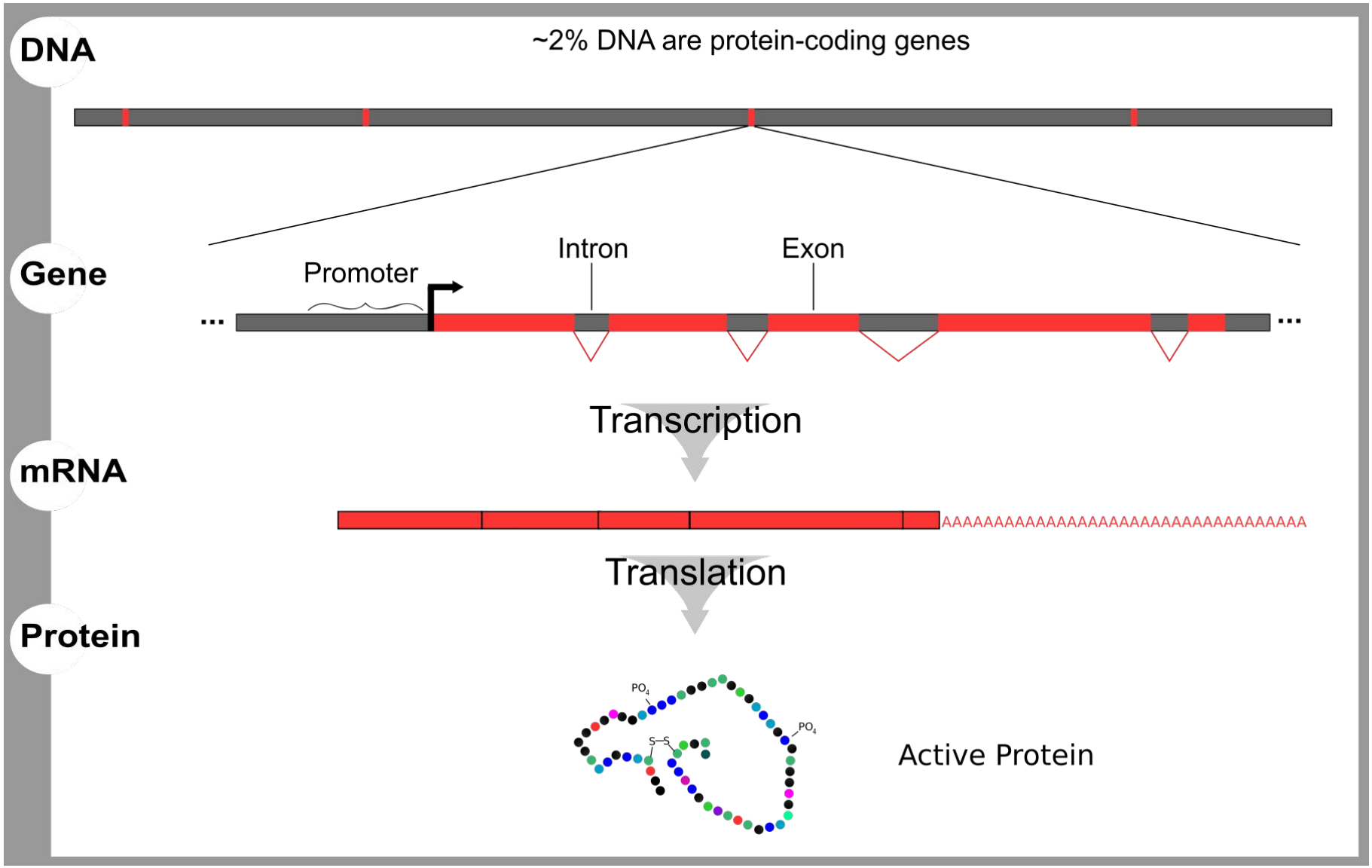
Consequences of DNA variation on gene regulation and human disease via RNA sequencing

Daniel M Bader
Gagneur lab - Computational biology

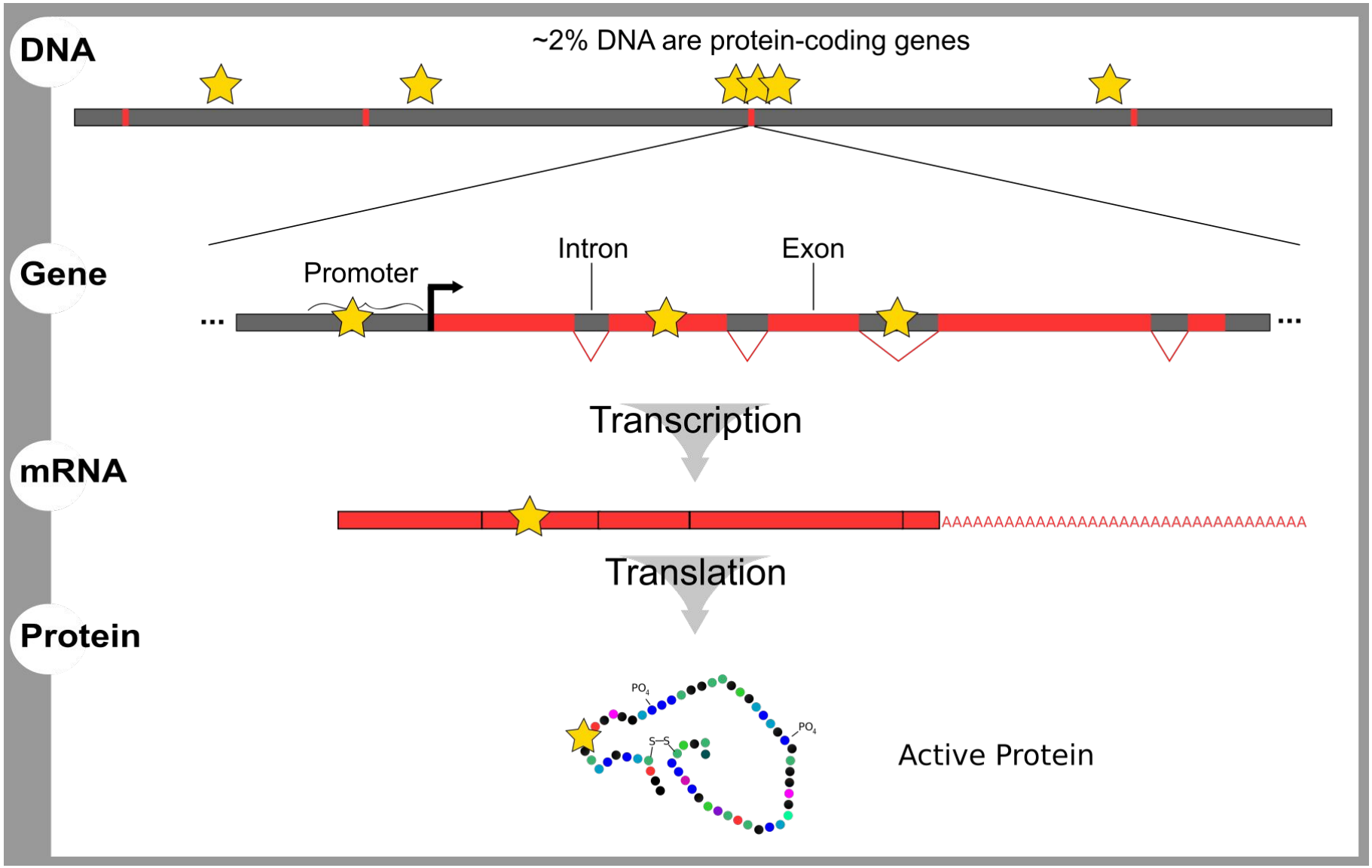
July 5, 2017

Information flow from DNA to active protein



Adapted from Mike Jones 2006 en.wikipedia

Study RNA for consequences of DNA variation



Adapted from Mike Jones 2006 en.wikipedia

Consequences of DNA variation in 2 studies

- Negative feedback buffers effects of regulatory variants on gene expression
    - Bader et al. 2015 Molecular Systems Biology
- Genetic diagnosis of Mendelian disorders via RNA sequencing
    - Kremer, Bader et al. 2017 Nature communications

Negative feedback buffers effects of regulatory variants on gene expression

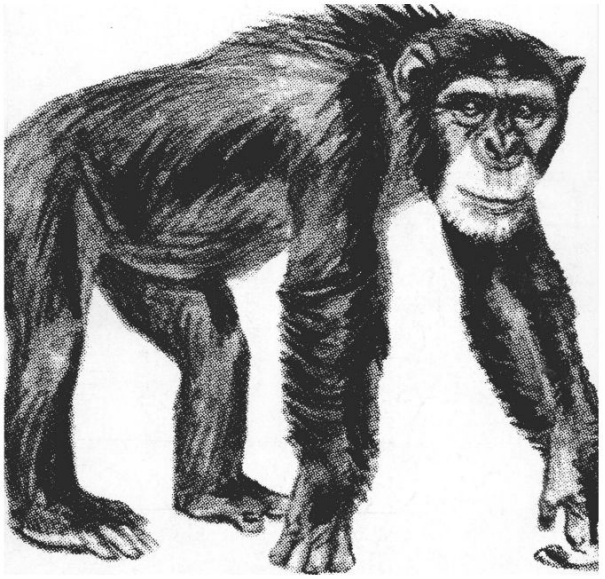
Bader et al. 2015 Molecular Systems Biology

Regulatory variants influence phenotypic diversity and evolution

- Regulatory variants as major drivers of speciation and adaptation,
  e.g. between human and chimpanzee

- Majority of genetic associations
  with common diseases are non-coding

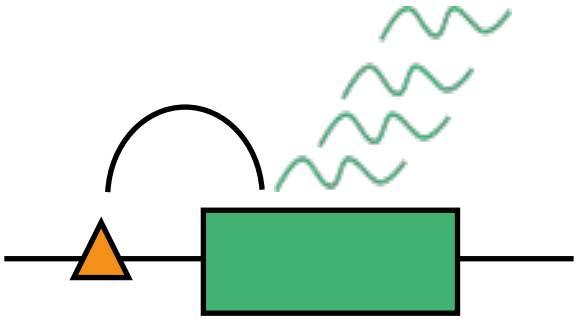
- Regulatory variants influencing
  RNA levels are frequent
  (~30% of genes affected across species)



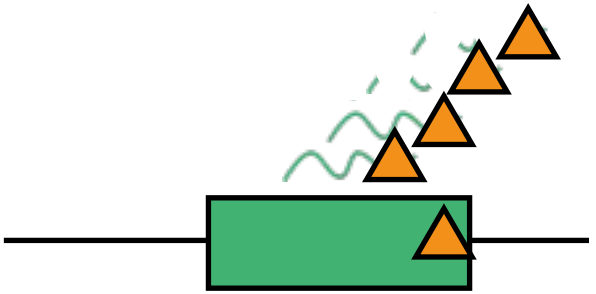
King & Wilson 1975 Science
Fraser 2013 Genome res
Gibson 2009 Nat Rev Gen
Manolio 2010 NEJM

Regulatory variants: What are cis effects?

Cis-regulation: the molecule itself



Transcription factor binding site

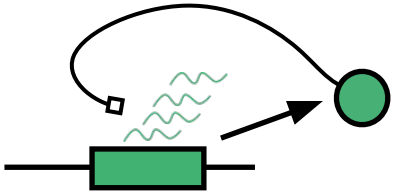


3’UTR element (RNA stability)

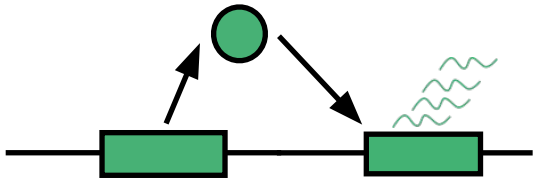
What are local and distant trans effects?

Trans-regulation: via another molecule

Local trans: other molecule encoded in the genetic vicinity

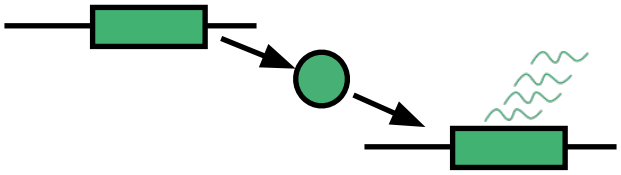


Feedback



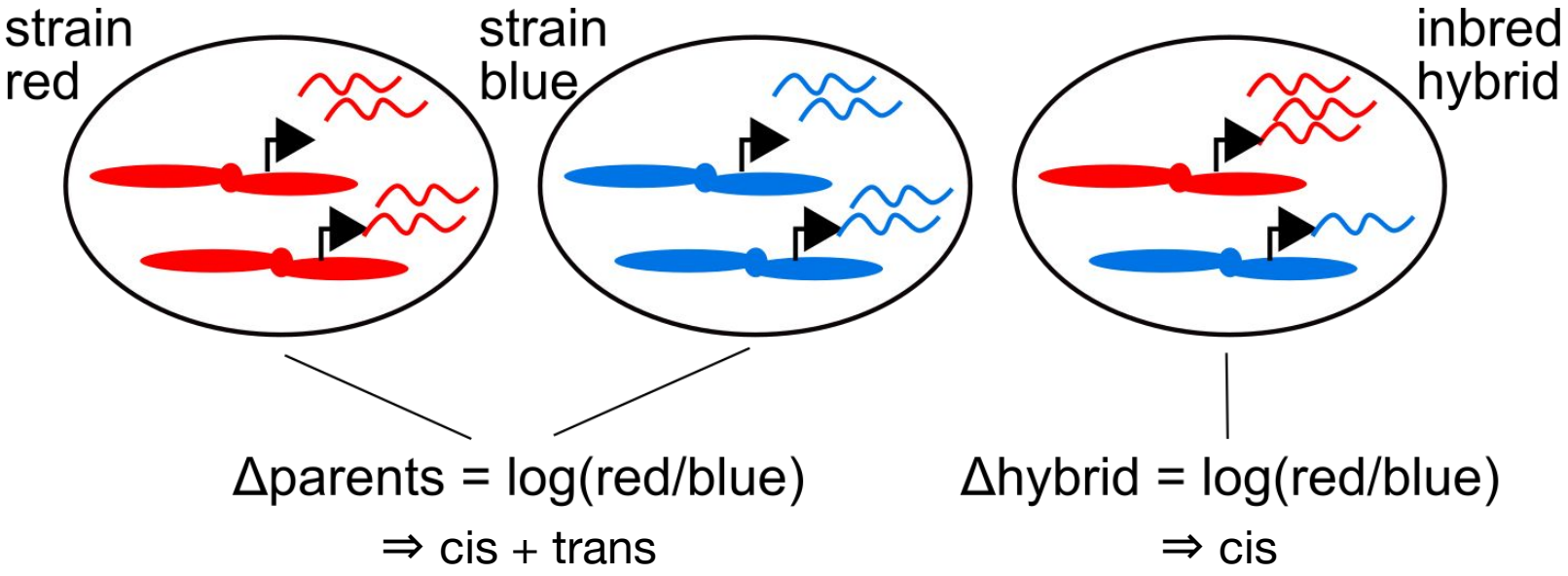
Nearby encoded transcription factor

Distant trans: other molecule encoded elsewhere in the genome

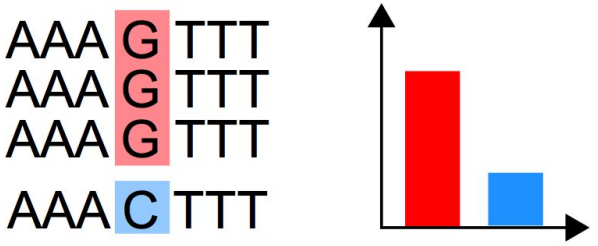


TF or microRNA
from another chromosome

How to quantify cis and trans?



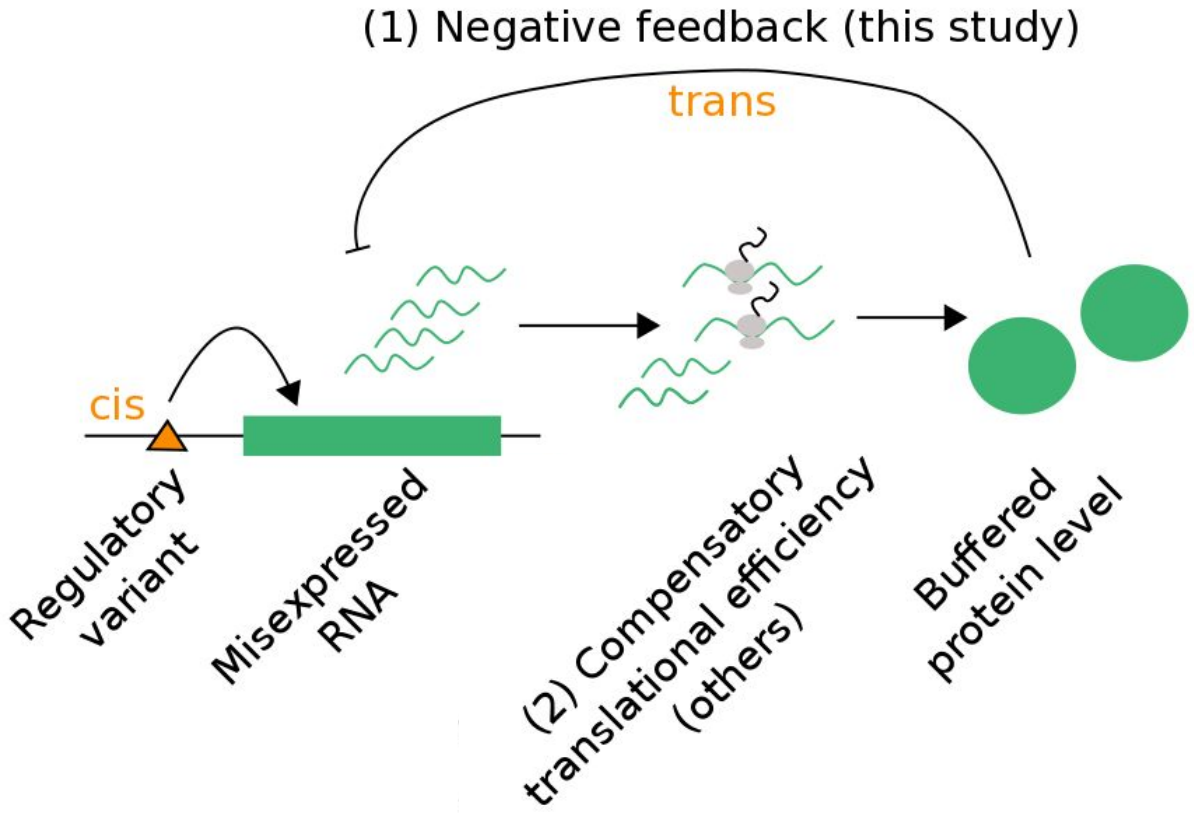
Compare ADE:
allelic differential expression



Yeast: Cowles 2002 Nat gen, Brem 2002 Science, Tirosh 2009 Science
Drosophila: McManus 2010 Genome res
Mouse: Goncalves 2012 Genome res

How to buffer effects of regulatory variants on RNA levels?

Negative feedback as buffering mechanism



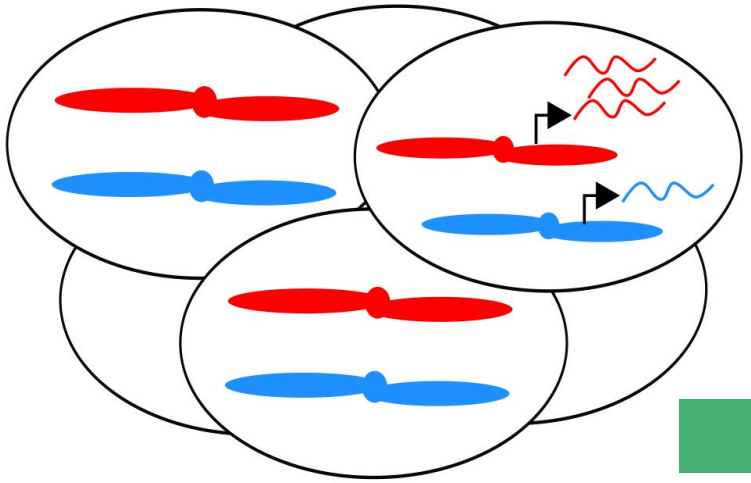
Denby 2012 PNAS

How widespread is feedback as buffering mechanism?

Artieri 2014, McManus 2014, Muzzey 2014 Genome Res, Albert 2014 PLoS Genet

Genome-wide allelic differential expression in a hybrid and its pool of spores

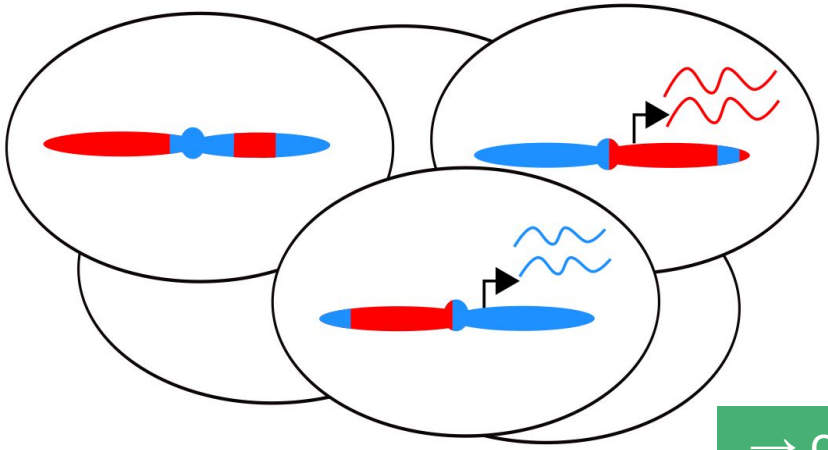
Diploid Hybrid
(F1 Generation)



S96: Lab Strain
SK1: Wild isolate
~1 SNP / 70bp

→ cis effects

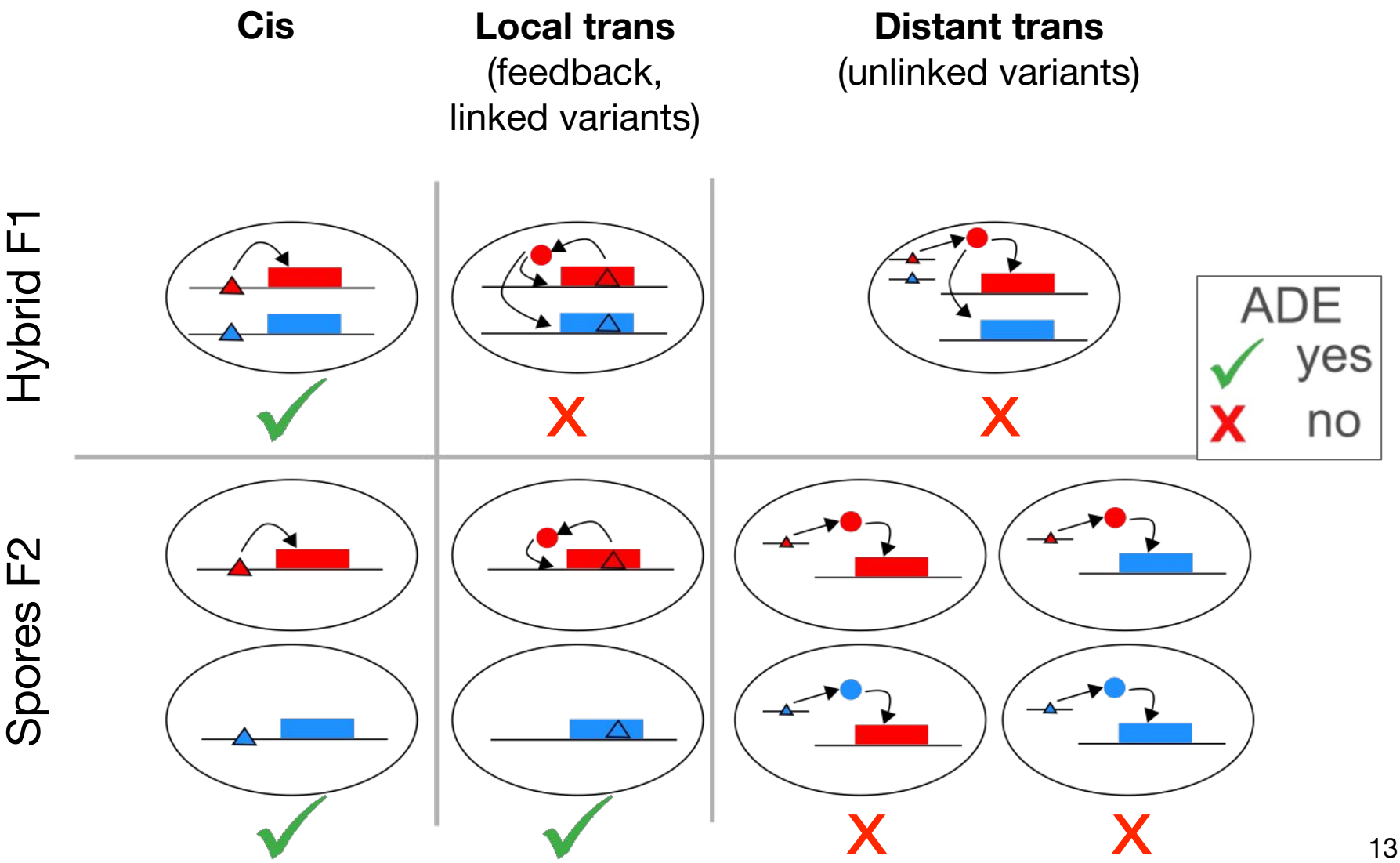
Haploid spore pool
(F2 Generation)



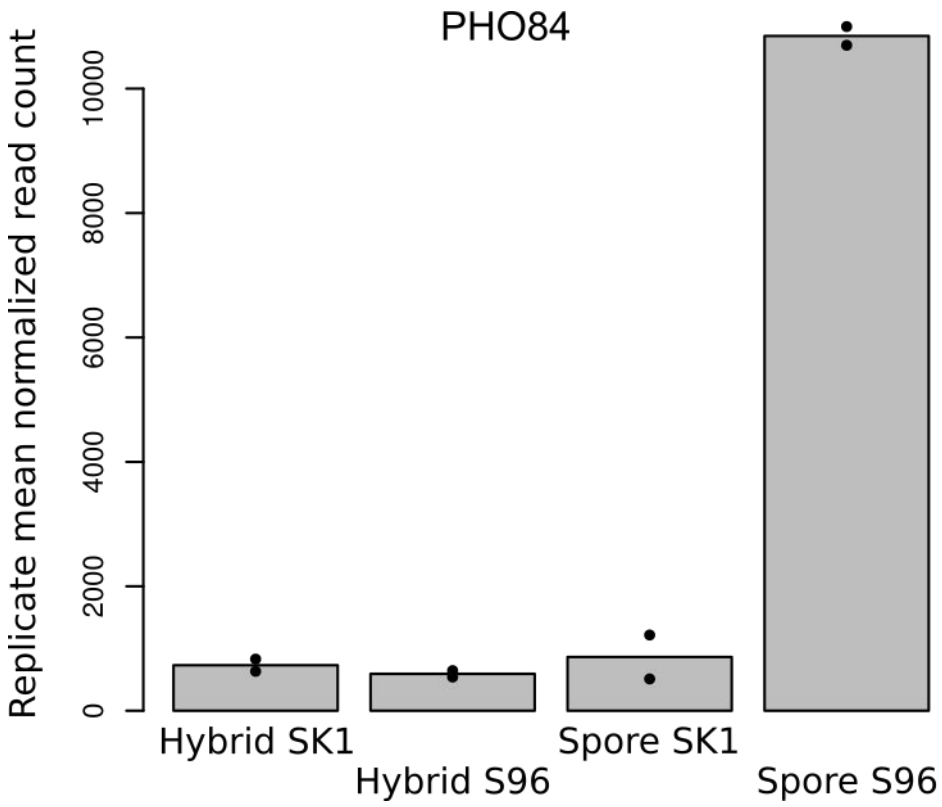
→ cis + local trans effects

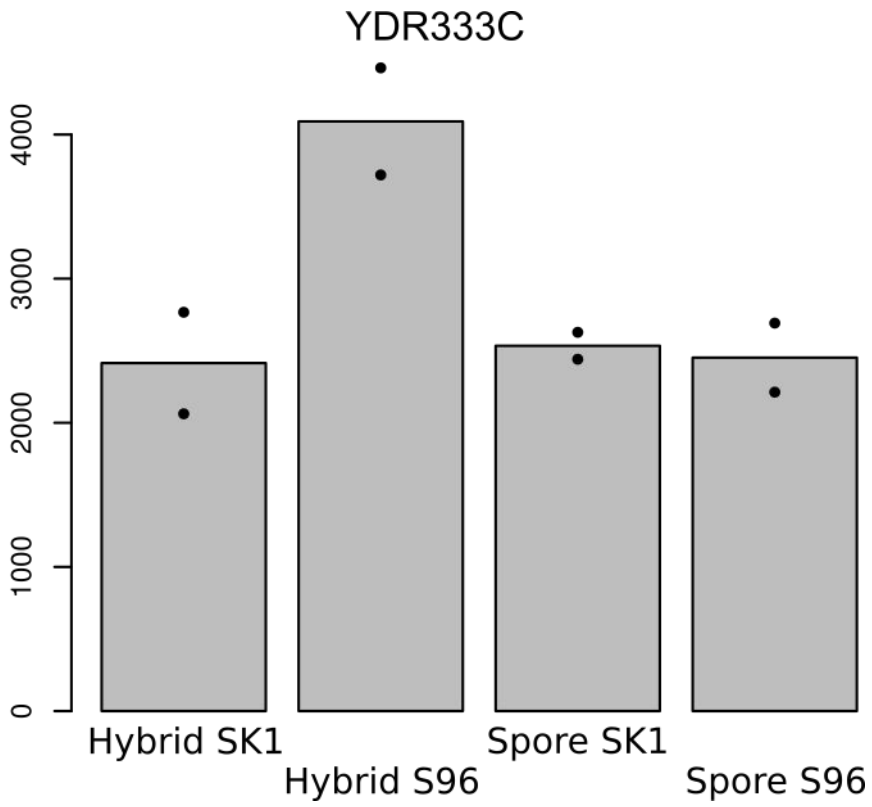
Ronald 2005 PLoS genet

Dissecting cis from local trans by comparing ADE in hybrid and spores



Exemplary local trans effects



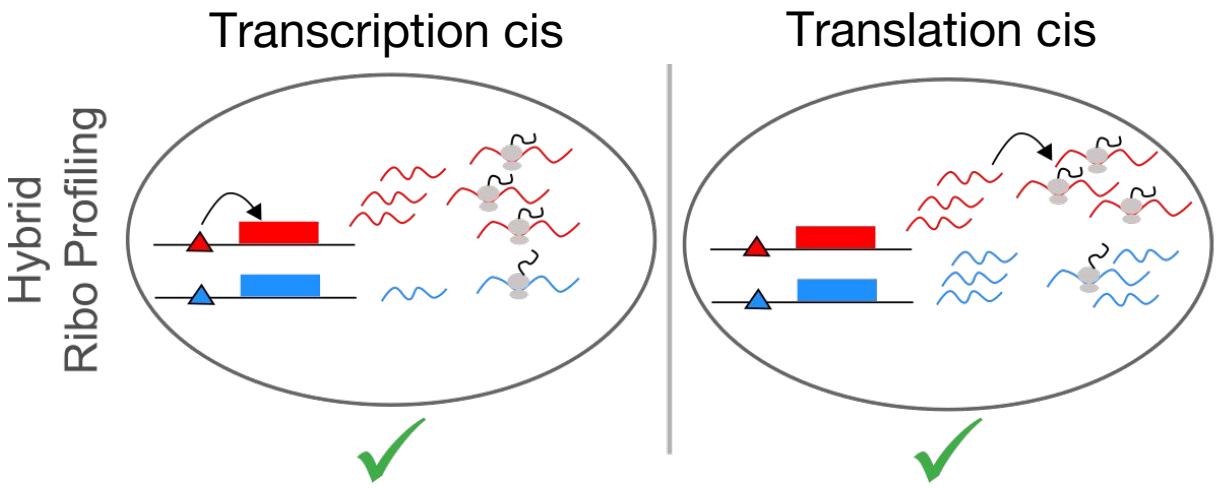


⇒ no cis, only local trans differences

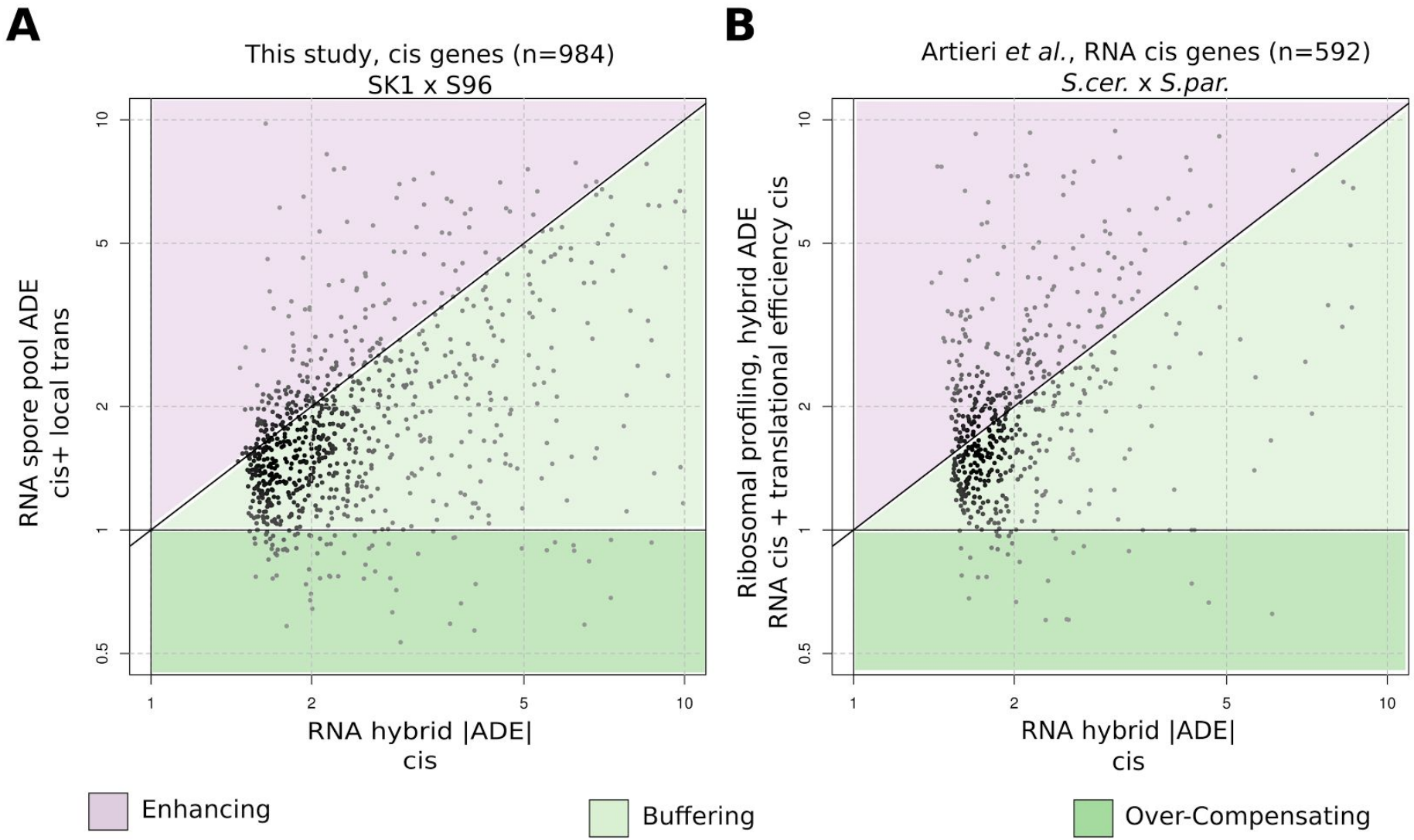
⇒ cis differences buffered by local trans

Two hypotheses explaining buffering

Negative feedback	vs	Adapted translation efficiency
Measured with RNAseq (our study)		Measured with RNAseq + Ribosome Profiling



RNA cis effects are buffered by local trans and not by translational cis

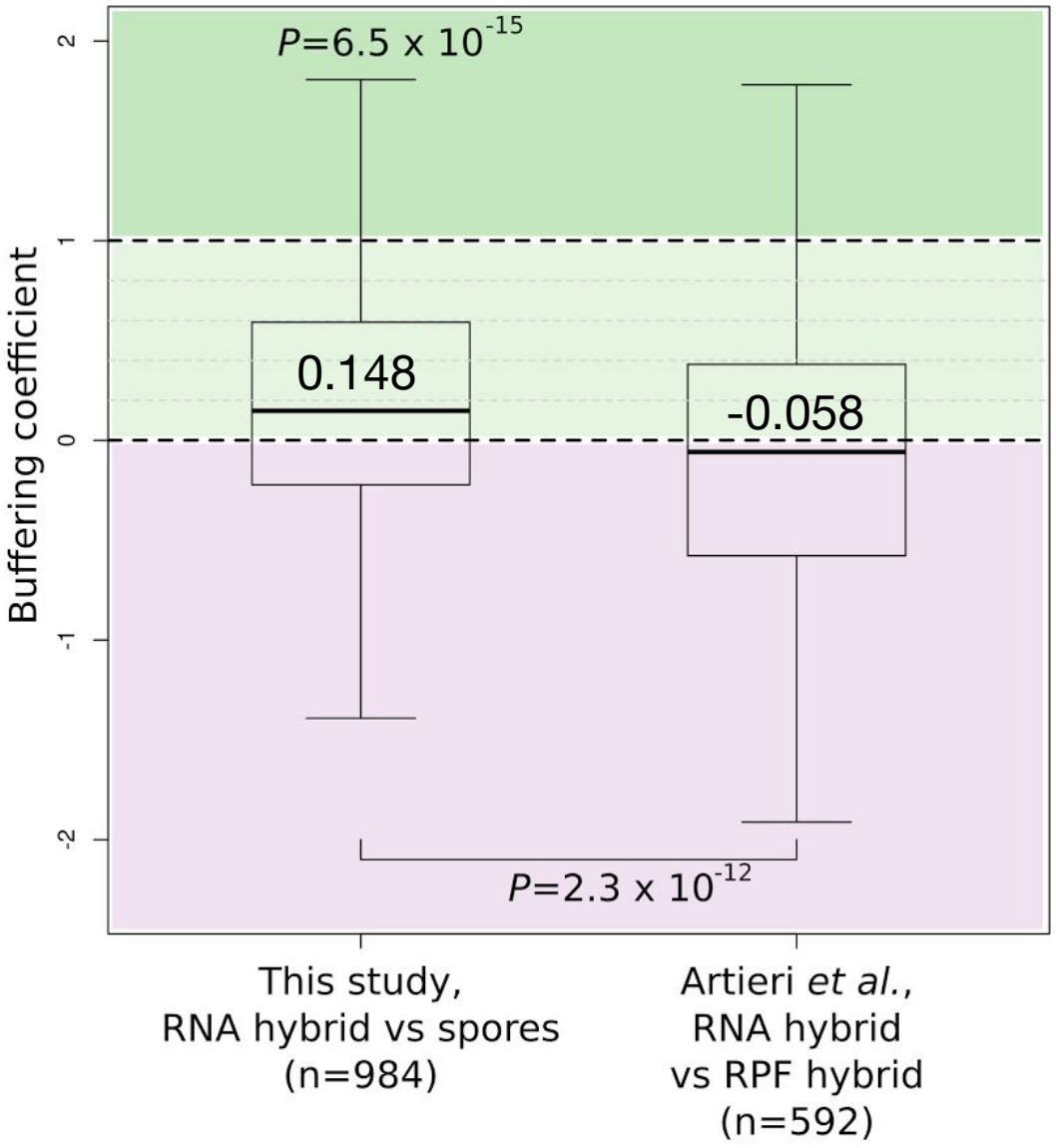


The buffering coefficient

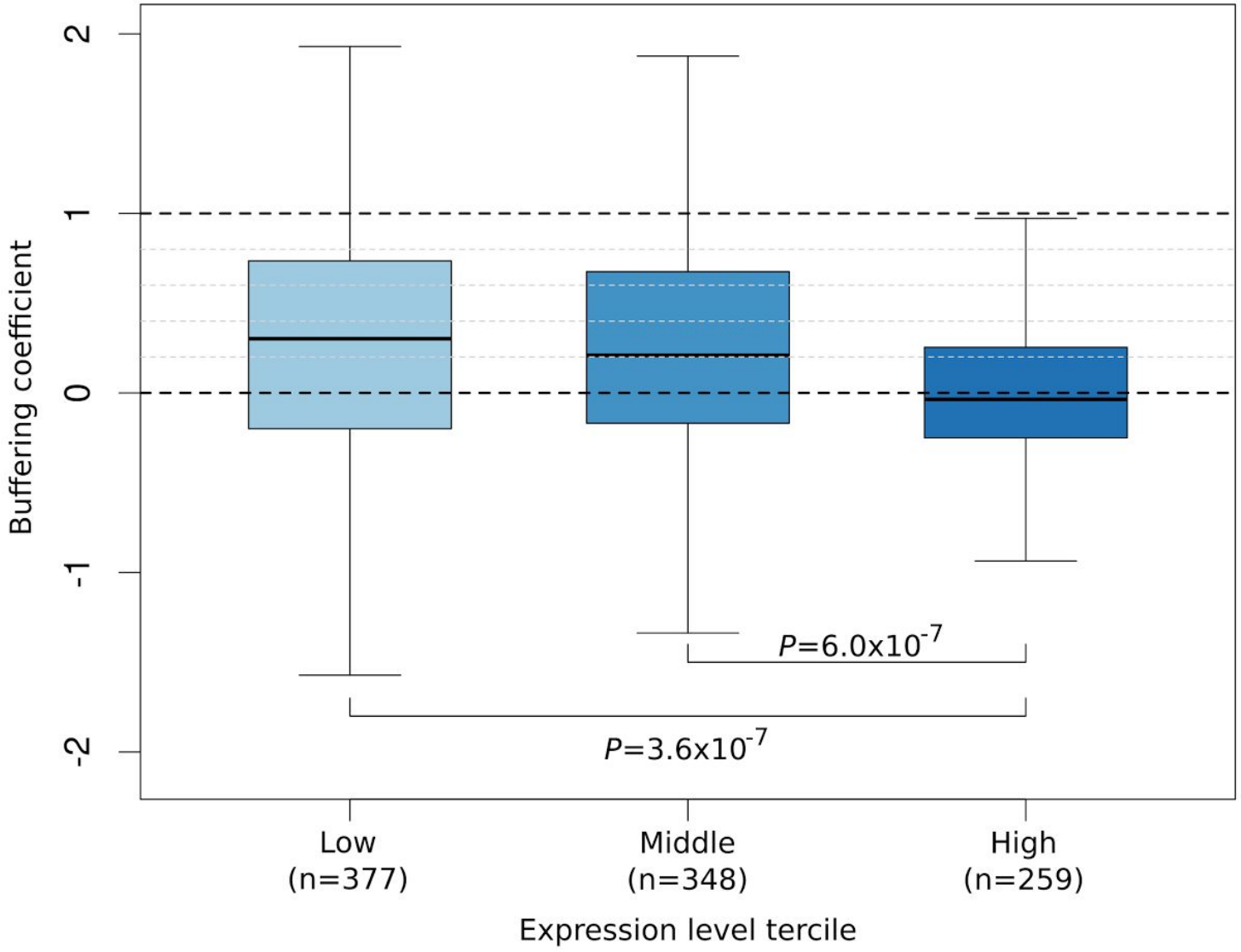
$$C = 1 - \frac{\log(\text{spore SK1/S96})}{\log(\text{hybrid SK1/S96})}$$

Cis effects are buffered by 15% through local trans effects

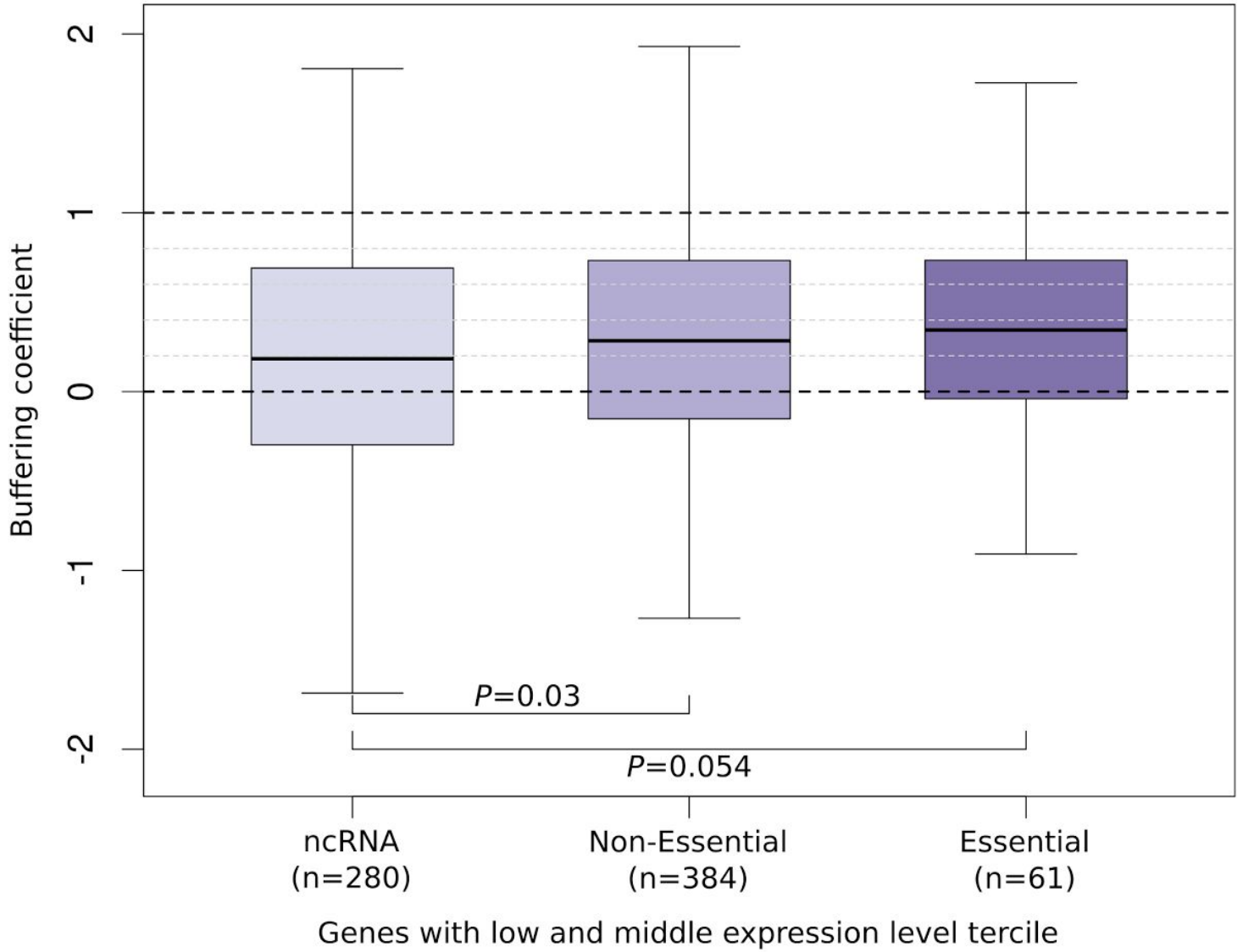
$$C = 1 - \frac{\log(\text{spore } SK1/S96)}{\log(\text{hybrid } SK1/S96)}$$



Buffering is stronger for cis genes in low to middle expression levels



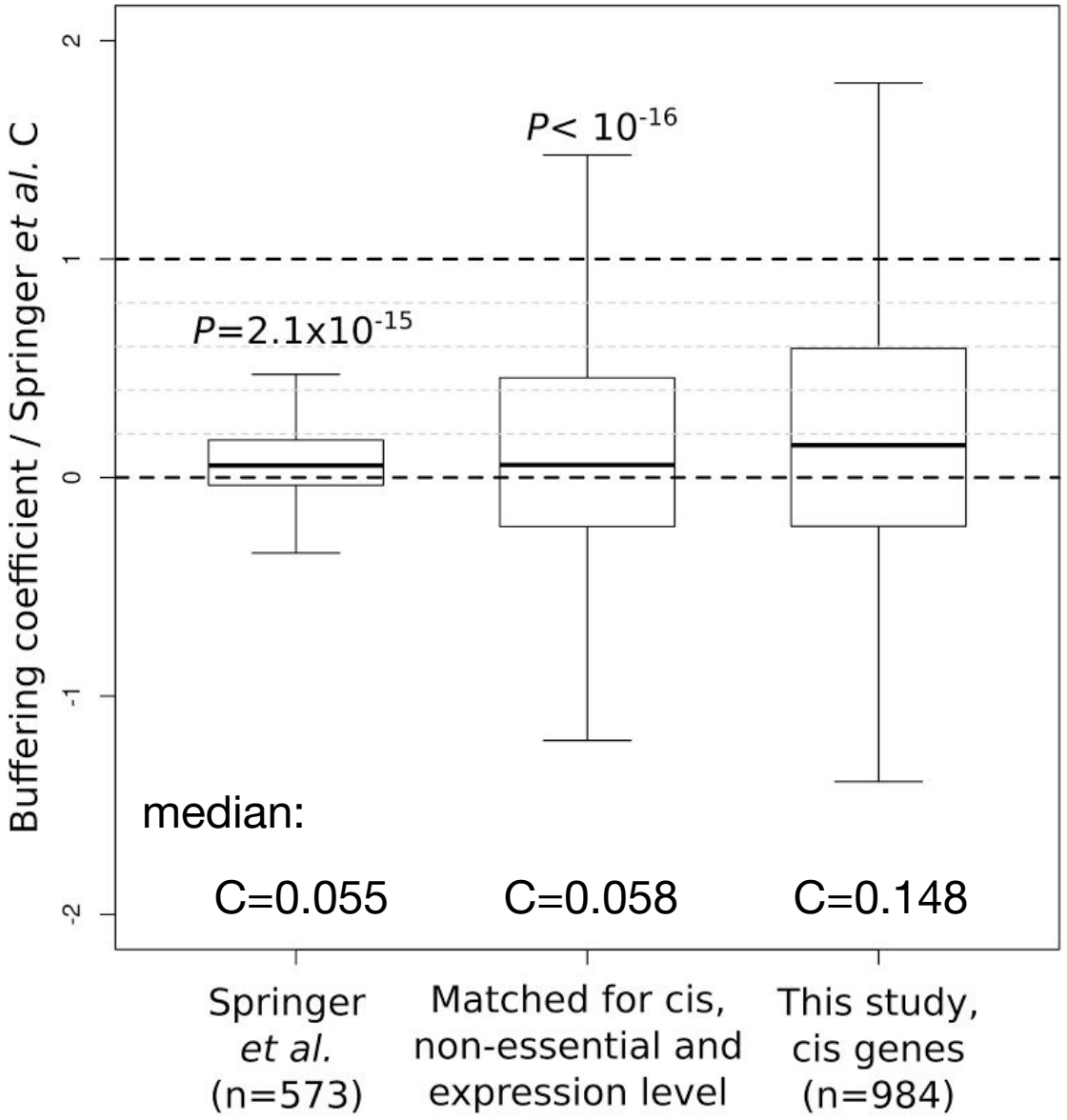
Buffering is stronger for essential cis genes



Local trans buffering is primarily due to negative feedback

Wild type (WT)

Heterozygote (HET)



Springer 2010 Mol Syst Biol

Conclusion

- Local trans effects buffer effects of cis-regulatory variants in *S. cerevisiae* by about 15%

- Two robustness strategies:
  - high level of expression
  - negative feedback

- No evidence for a widespread buffering through translation

Implications

- Waddington's canalization
  - Redundancy
  - Chaperon proteins
  - Negative feedback on RNA

- Negative feedback could explain partial dosage compensation of autosomal genes in higher eukaryotes (fly)

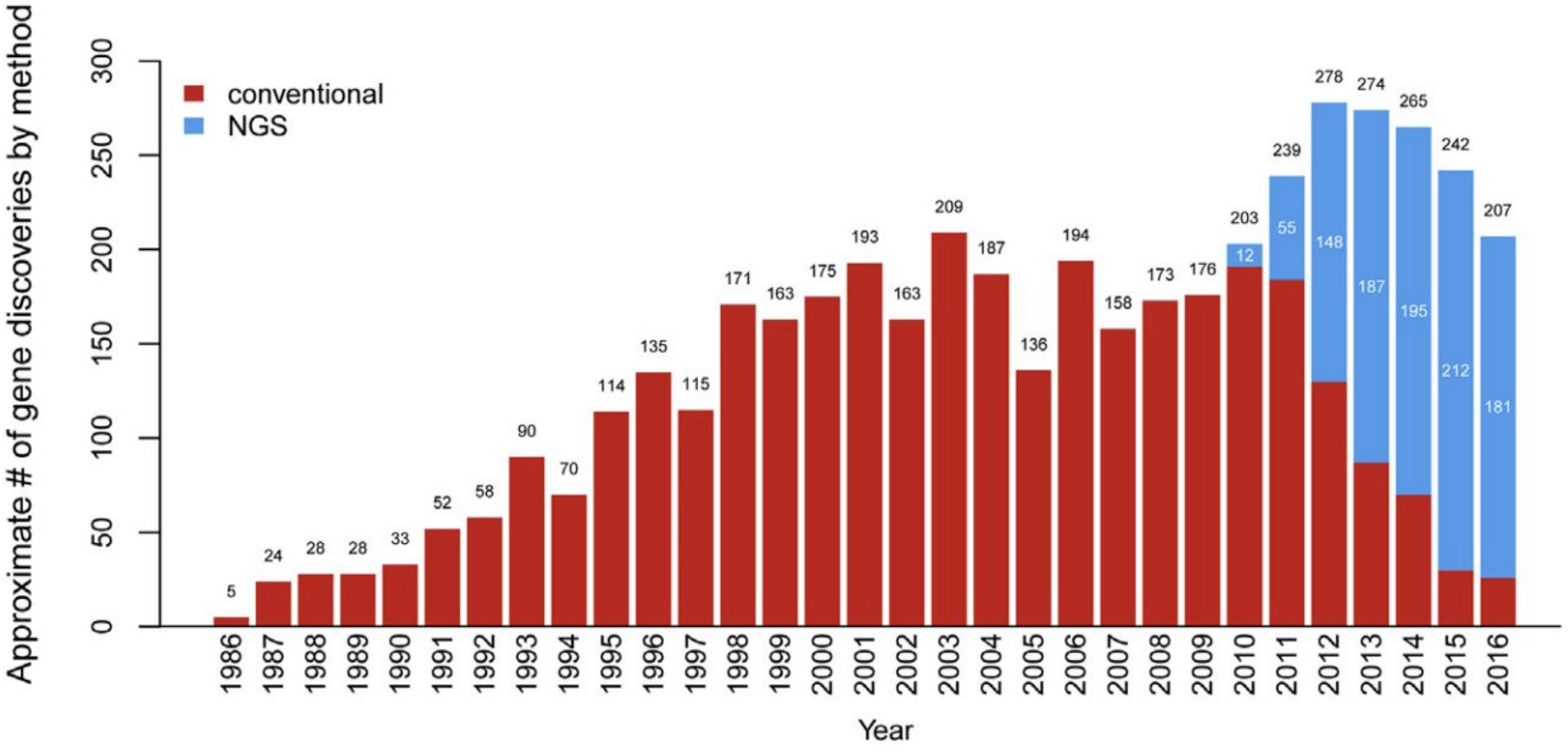
- Feedback could also act at the protein level

Waddington 1942 Nature
Malone 2012 Genome biol

Genetic diagnosis of Mendelian disorders via RNA sequencing

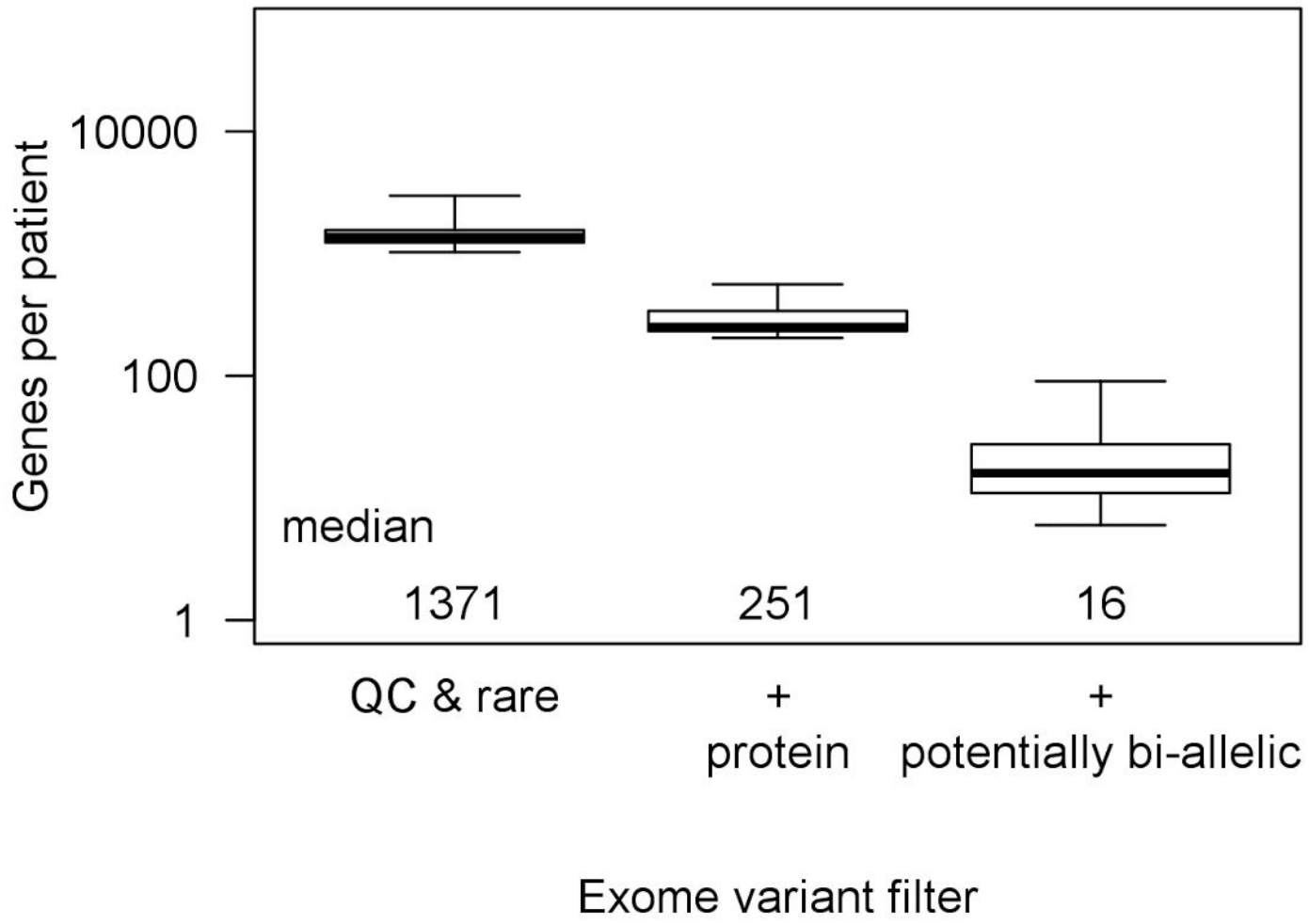
Kremer, Bader et al. July 2016 bioRxiv
Bader talk at ESHG May 2017
Kremer, Bader et al. July 2017 Nature communications

Sequencing as standard diagnosis tool



Boycott 2017 AJHG

Variant prioritization based on whole exome sequencing



Rare: minor allele frequency < 0.001

Limitations of genome sequencing

- Exome sequencing
  - ~2% of genome covered
  - ~50% patients not diagnosed

Chong 2015 AJHG review
Wortmann 2015 J Inherit Metab Dis
Retterer 2016 Genetics in Medicine

- Genome sequencing
  - detection of all variants
  - difficult prioritization and interpretation

Taylor 2015 Nature genetics

- Many variants of unknown significance
  → synonymous or non-coding

- Knowledge gap between coding and regulatory sequence

⇒ RNA sequencing!

Using the power of RNA-seq to investigate Mendelian disorders of mitochondria

Mendelian disorders

- EU: less than 5 in 10,000
- US: less than 200,000 affected

EUR-Lex 32000R0141, 1999
USA Orphan Drug Act, 1983

Mitochondrial disorders

- Birth prevalence 2 in 10,000

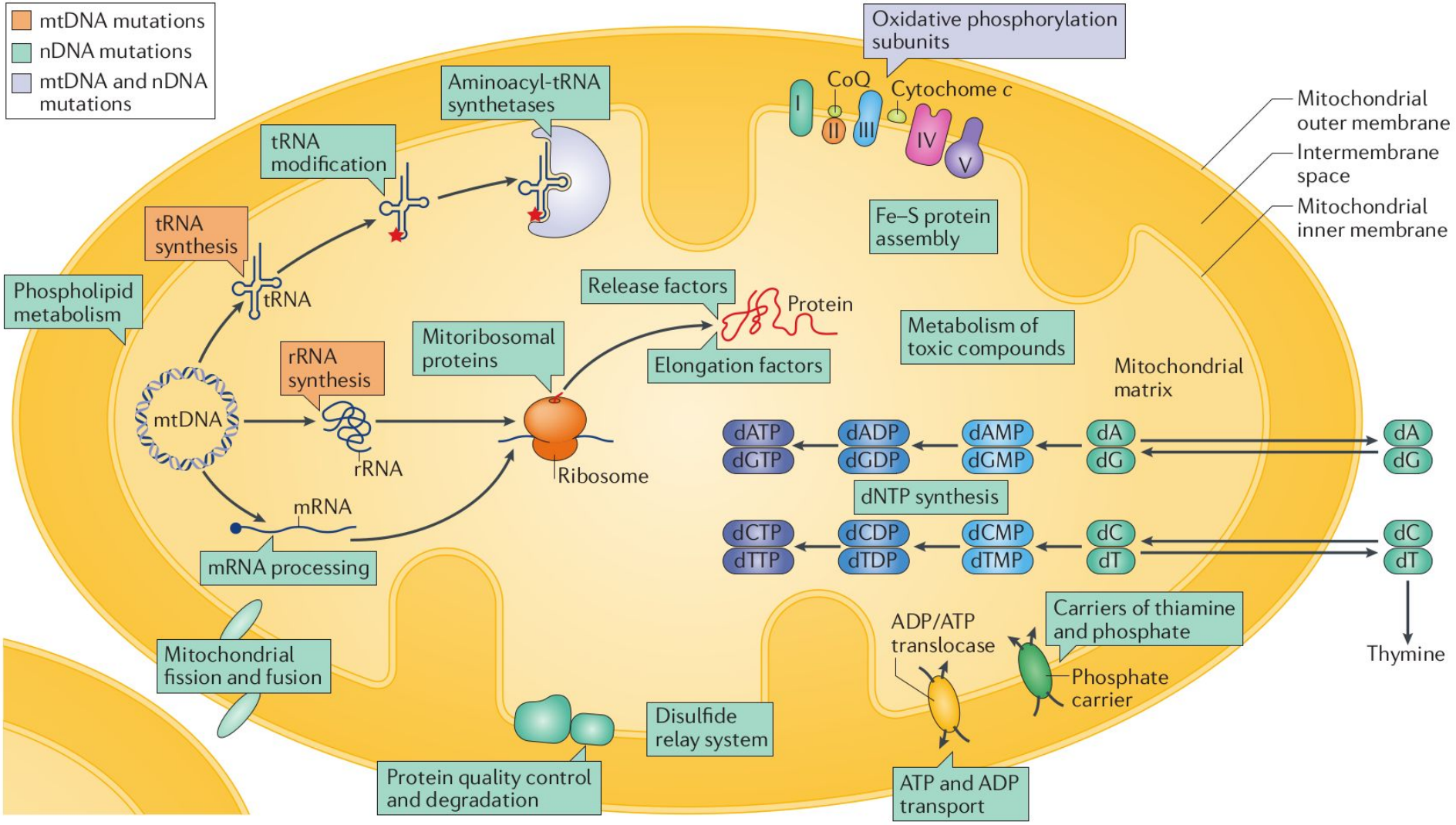
Gorman 2016 Nat Rev Dis Primers

- Large genetic basis with causative defects identified in more than 250 genes
  - mtDNA
  - Nuclear DNA

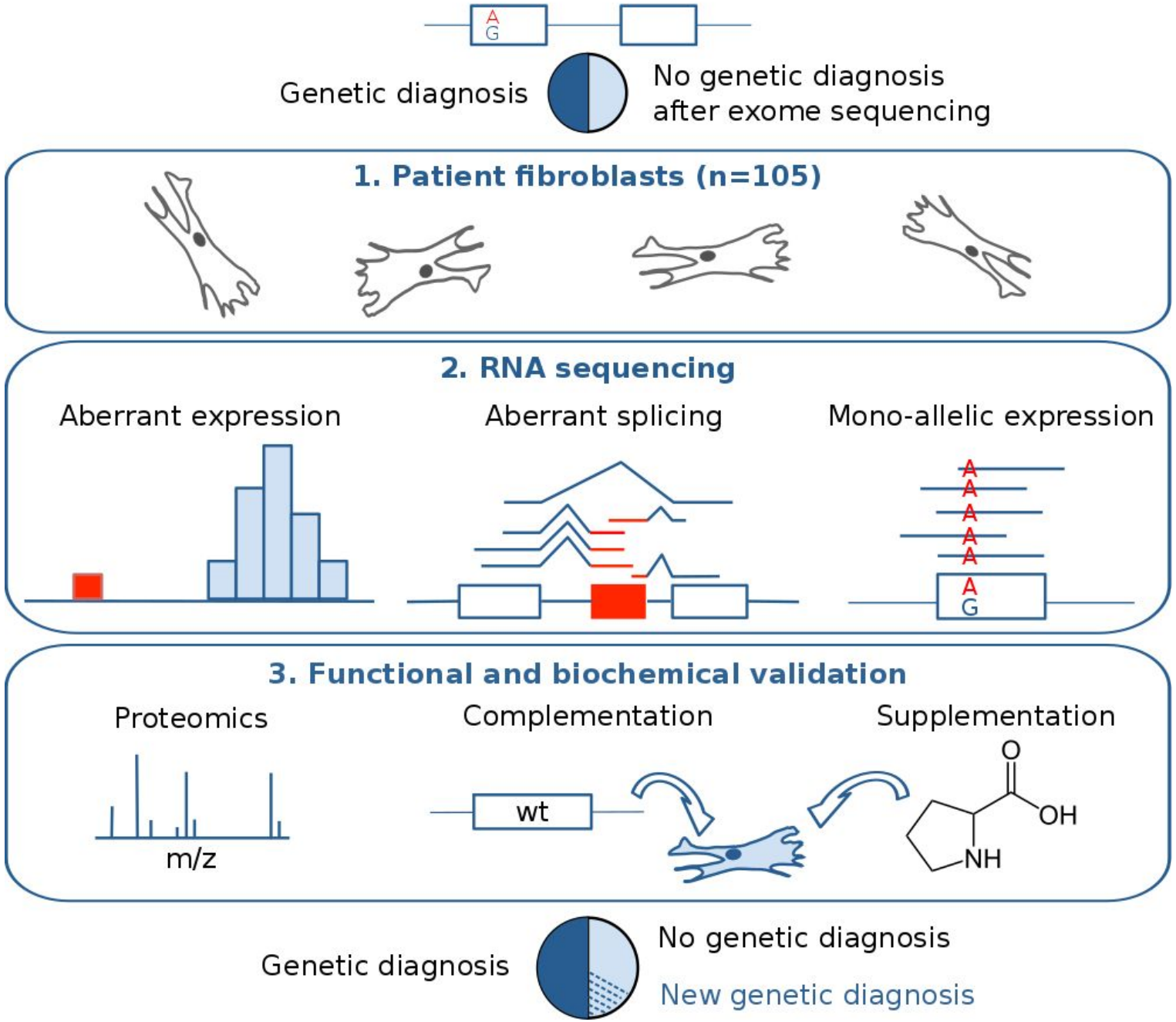
Mayr 2015 J Inherit Metab Dis

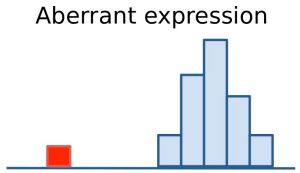
- Test for respiratory activity and its rescue using fibroblast cell lines

Encoding of mitochondrial disorder genes



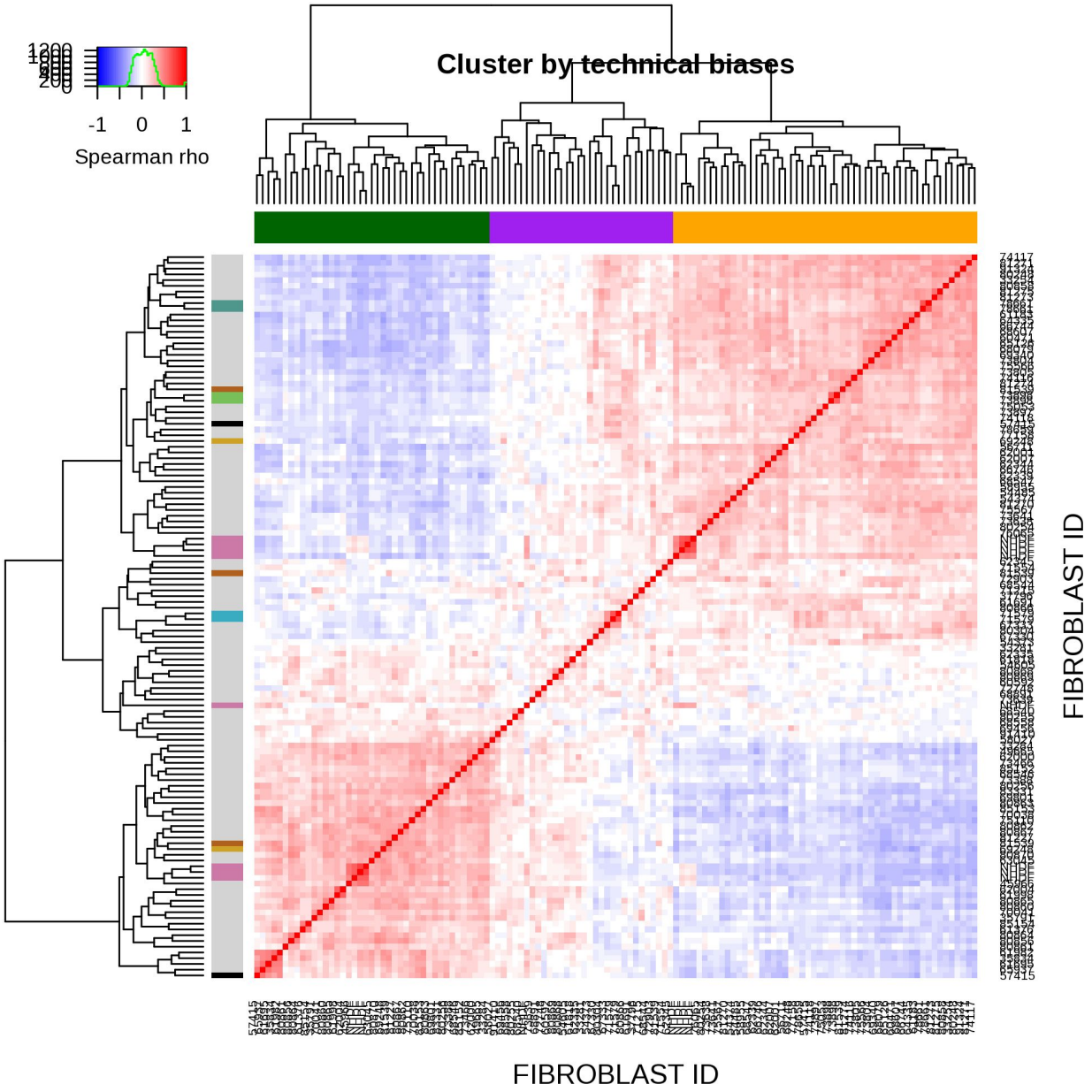
Gorman 2016 Nat Rev Dis Primers


A
G
Genetic diagnosis
No genetic diagnosis
after exome sequencing
1. Patient fibroblasts (n=105)
2. RNA sequencing
Aberrant expression
Aberrant splicing
Mono-allelic expression
A
A
A
A
A
A
G
3. Functional and biochemical validation
Proteomics
m/z
Complementation
wt
Supplementation
O
OH
NH
Genetic diagnosis
No genetic diagnosis
New genetic diagnosis

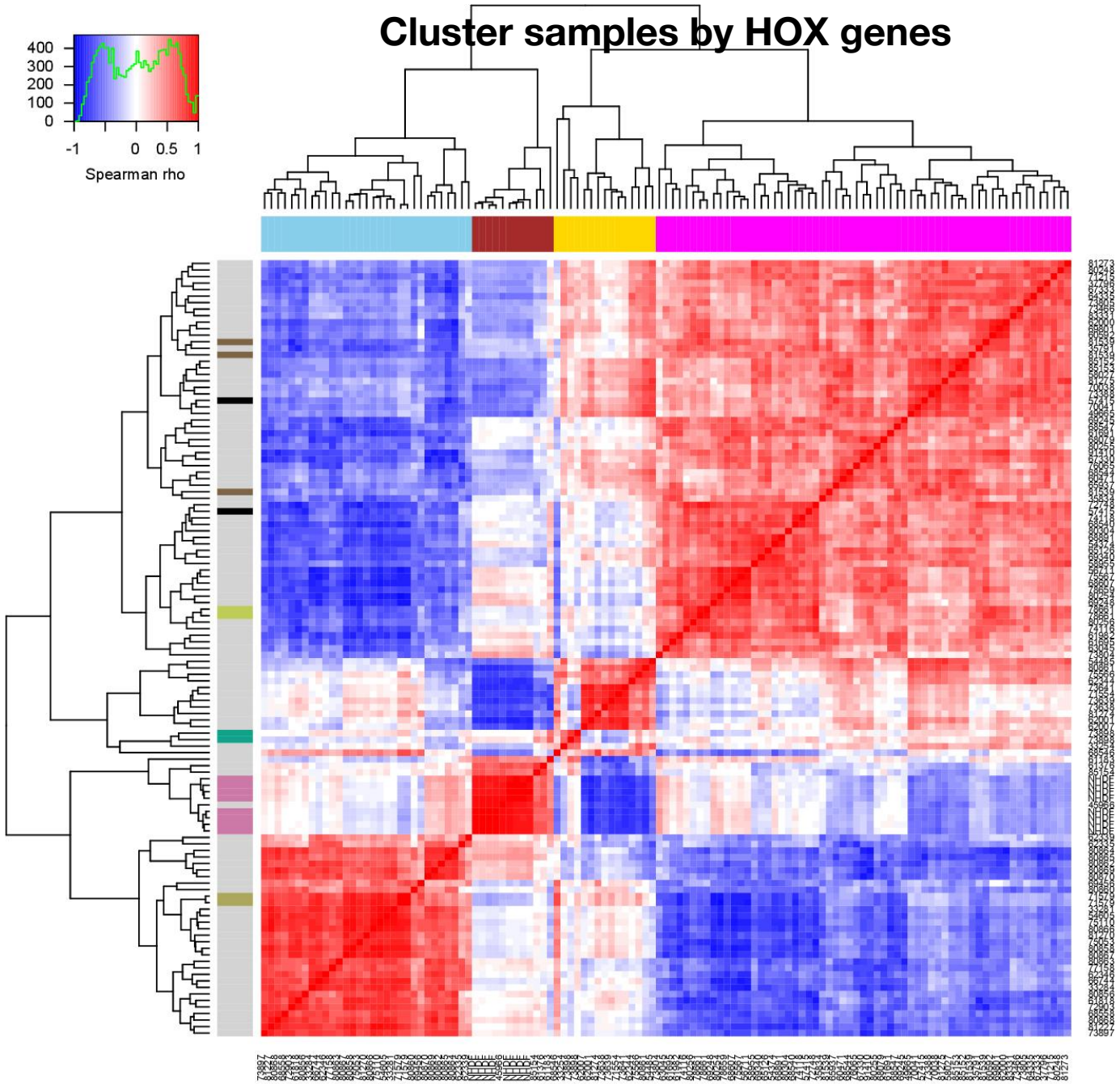


Aberrant expression

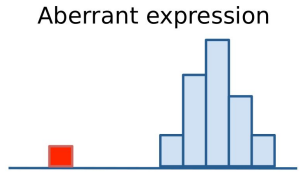
RNA normalization of confounding effects

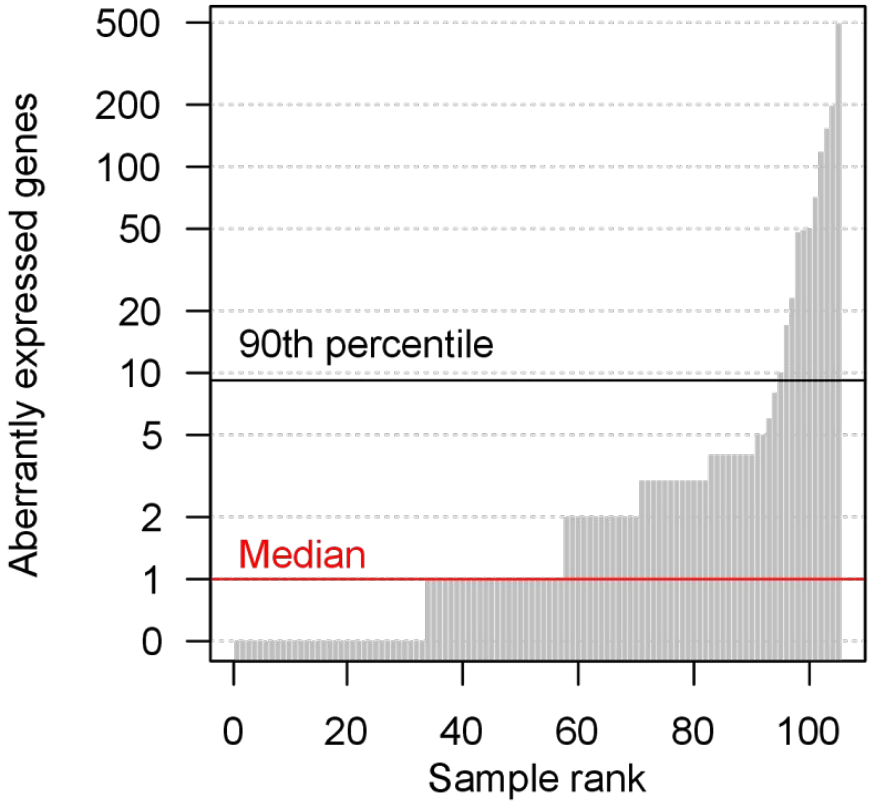


Cluster samples by HOX genes



Detection of aberrant expression with strict statistics

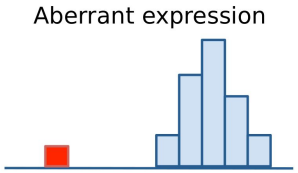


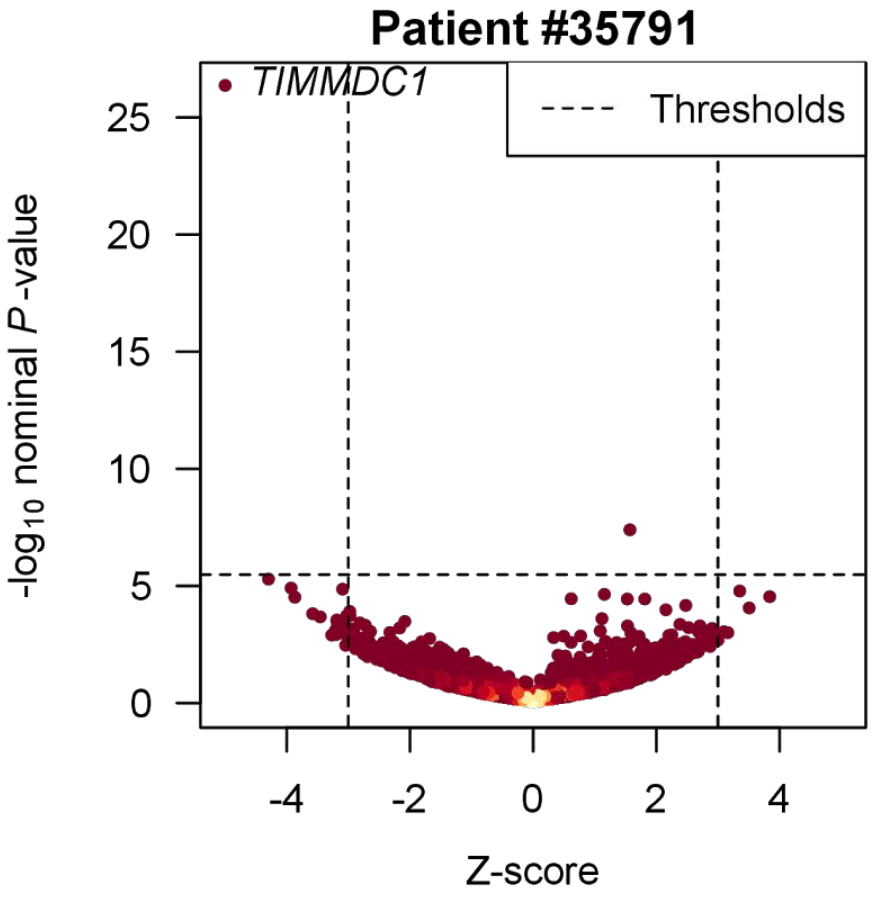


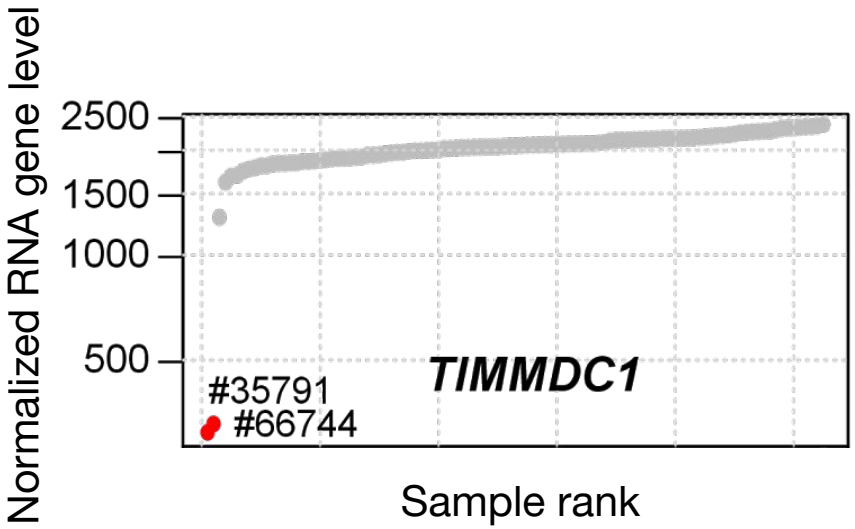
- Gene-wise read counts
- Normalize for sample effects
- Test 1 sample versus rest
- Z-score = difference to mean / standard deviation
- Adjusted P-value < 0.05 & |Z-score| > 3

Love 2014 Genome Biology
Li, ..., Montgomery 2016 bioRxiv

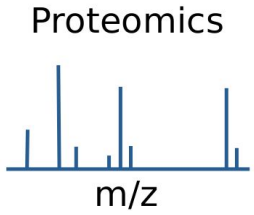
TIMMDC1 as example for aberrant expression

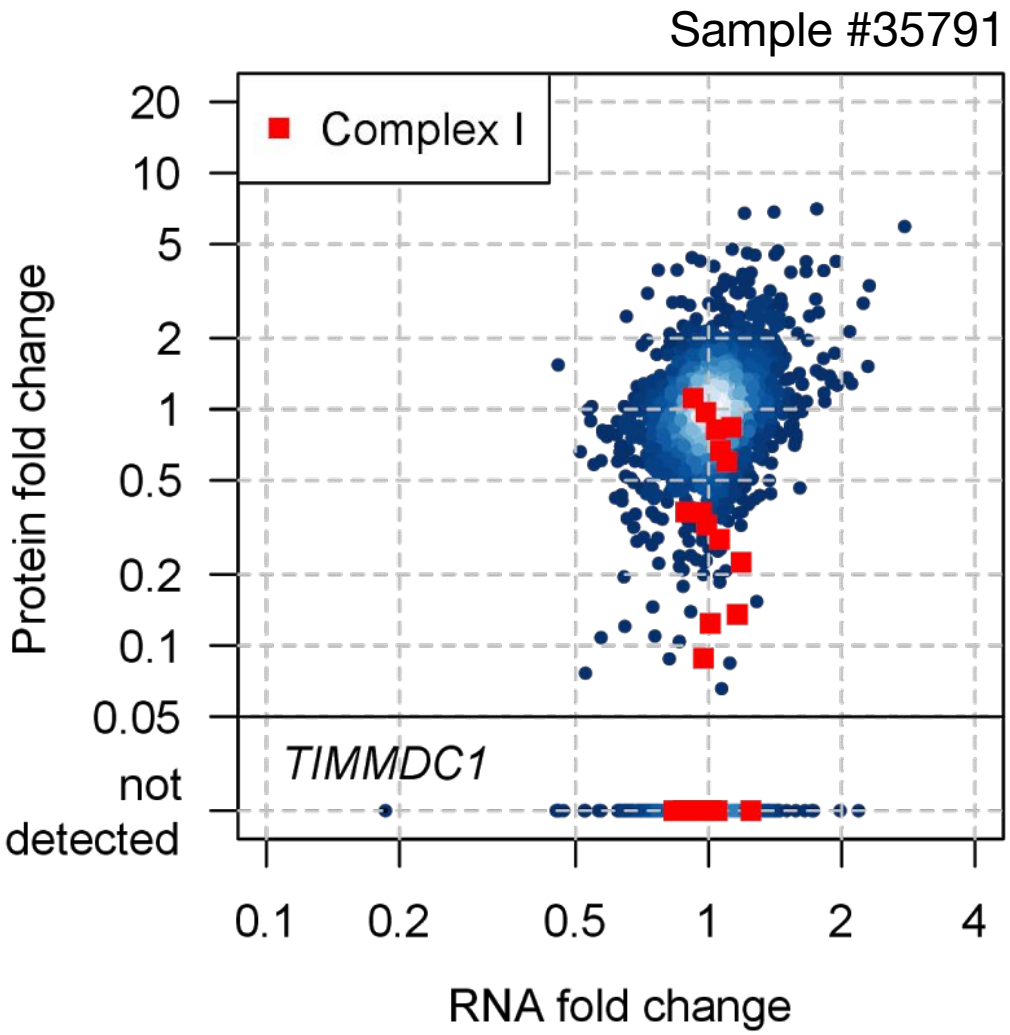






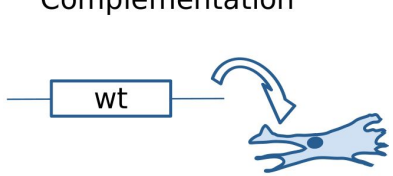
Proteomic validation

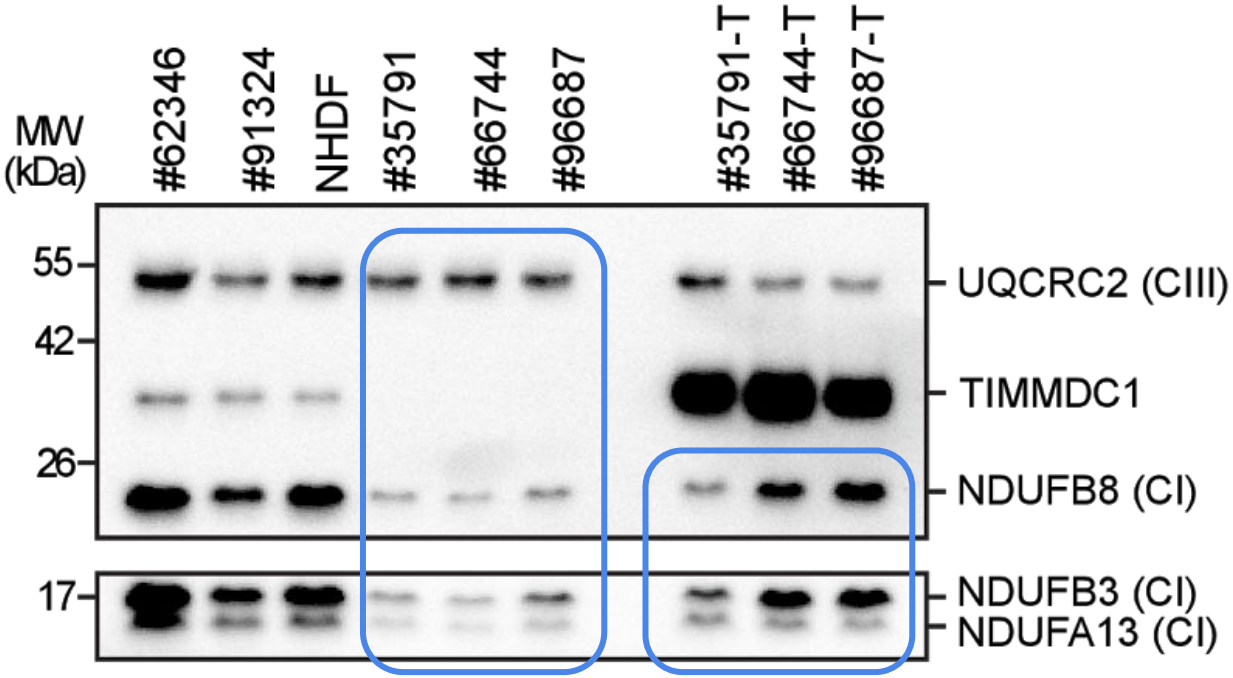




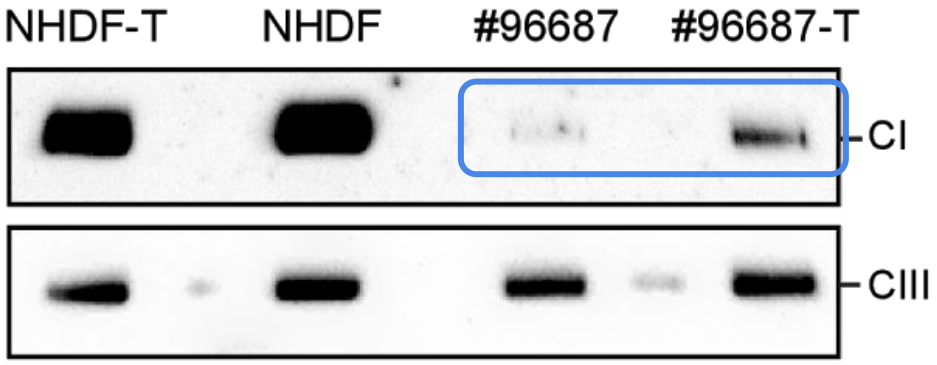
TIMMDC1: Translocase of Inner Mitochondrial Membrane Domain Containing 1; Complex I assembly factor

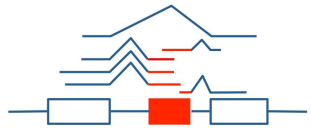
Rescue complex I assembly by wildtype complementation





Laura S Kremer

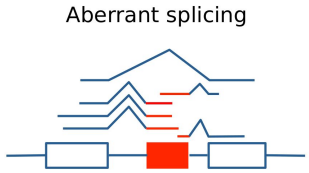


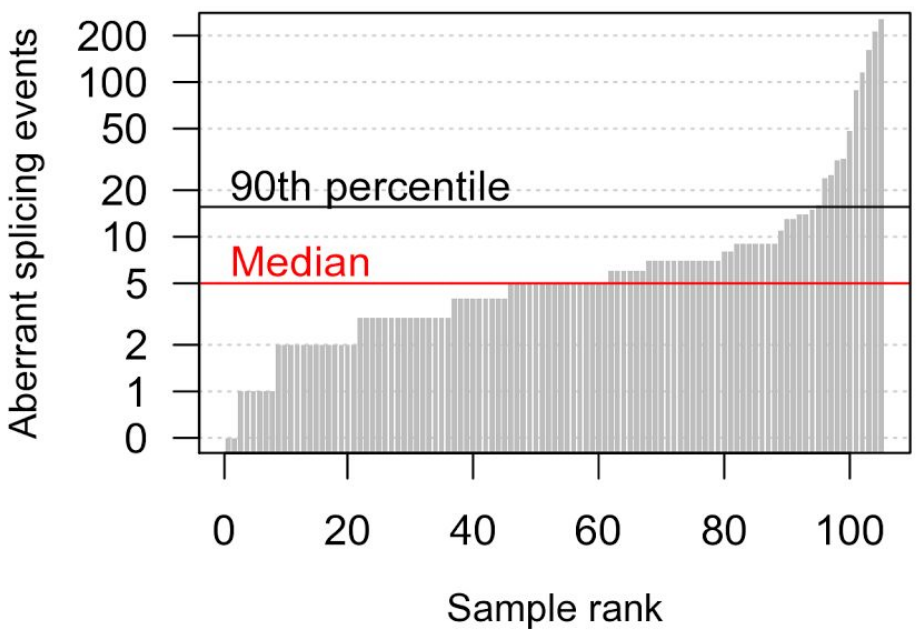


Aberrant splicing

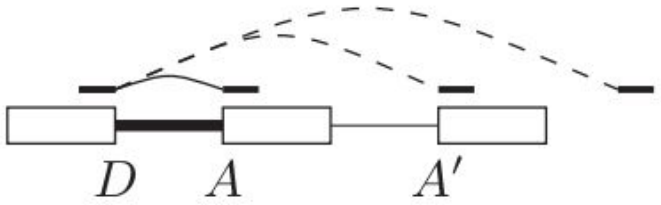
Developed by Christian Mertes

Detection of aberrant splicing





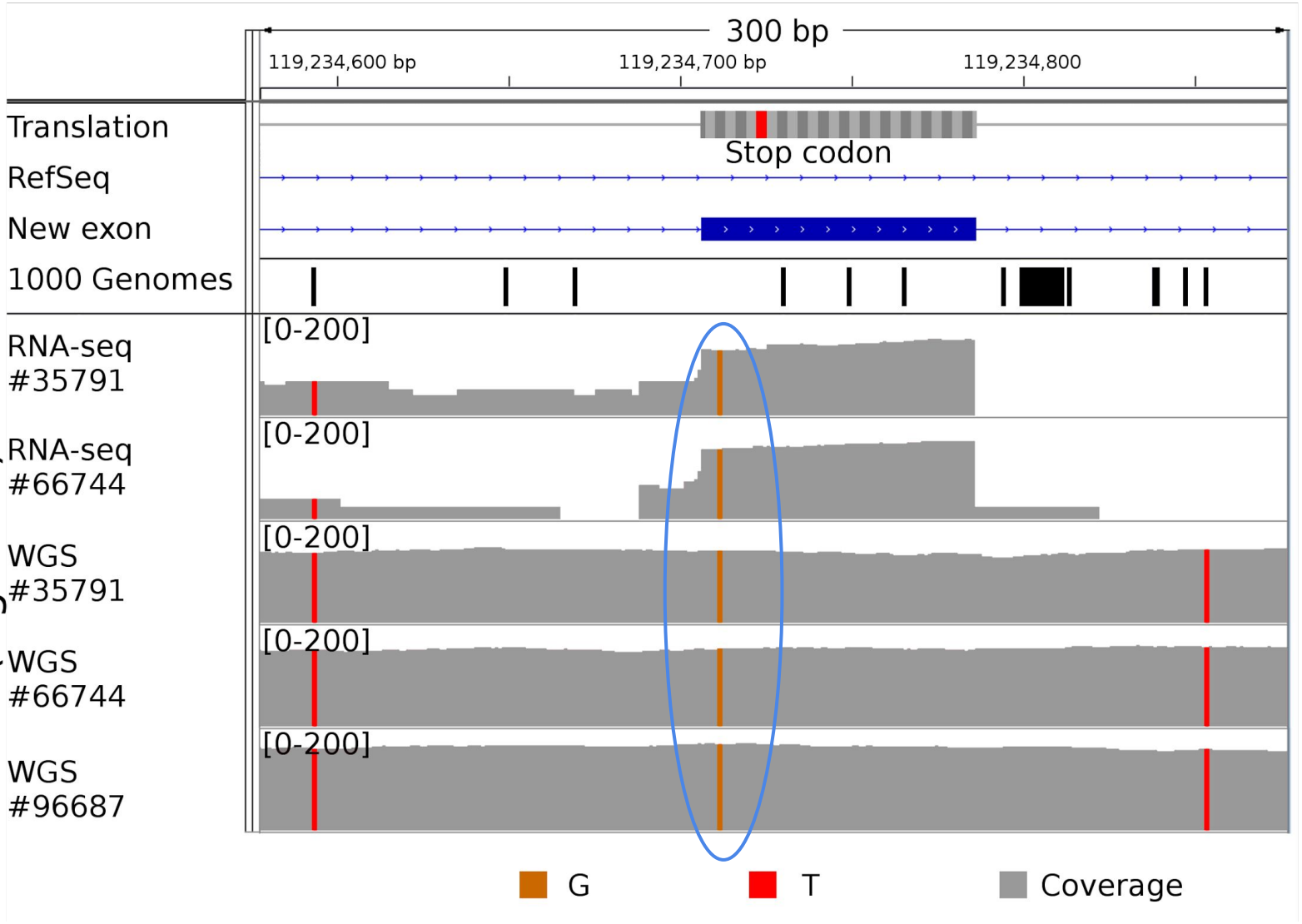
- Based on Leafcutter
  - Exon-junction read counts
  - Statistics on percent spliced in (PSI **Ψ**) values

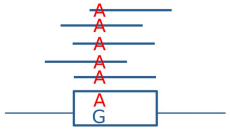


- Method adapted for testing 1 vs rest by pseudo counts
- Adjusted P-value < 0.05

Pervouchine 2013 Bioinformatics
Li, Knowles, Pritchard 2016 bioRxiv
Cummings 2017 Sci Transl Med

Rare deep intronic mutation in TIMMDC1 inside the new exon

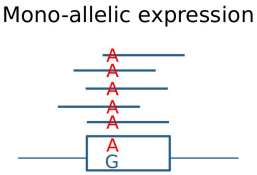


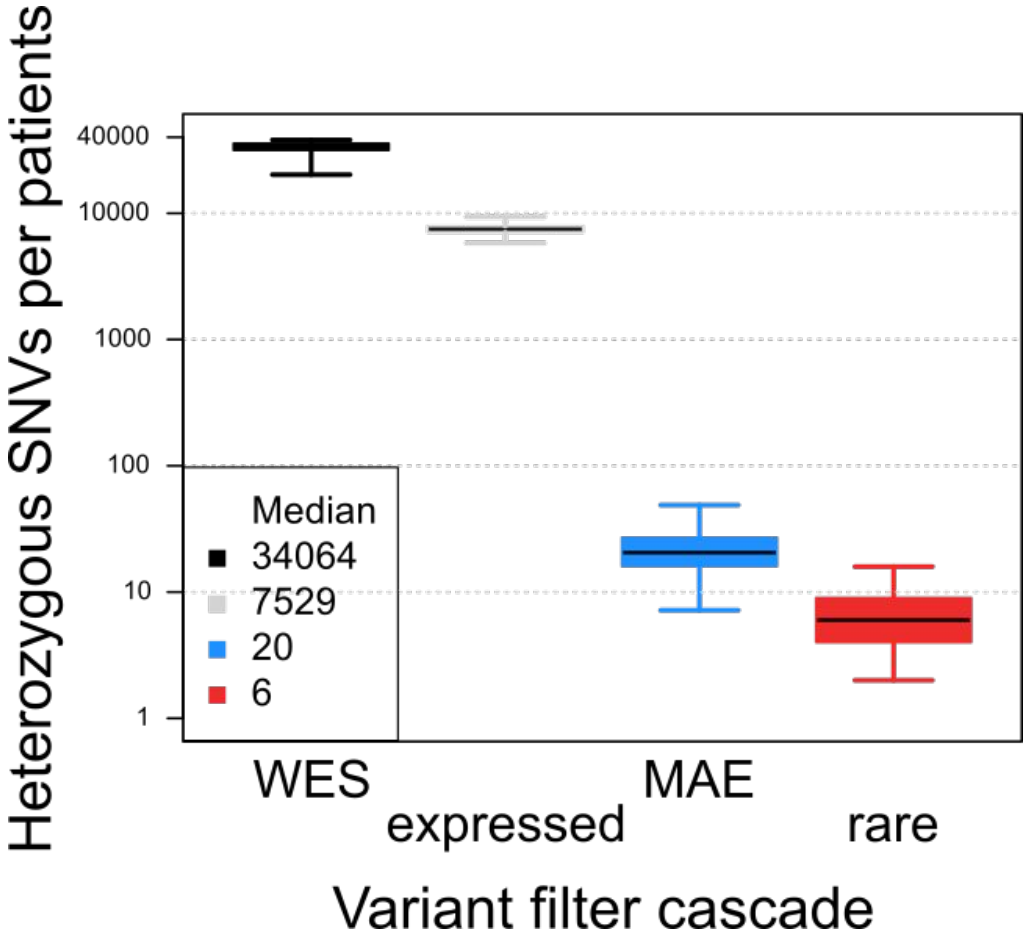


Mono-allelic expression

Developed together with Christian Mertes

Detection summary mono-allelic expression

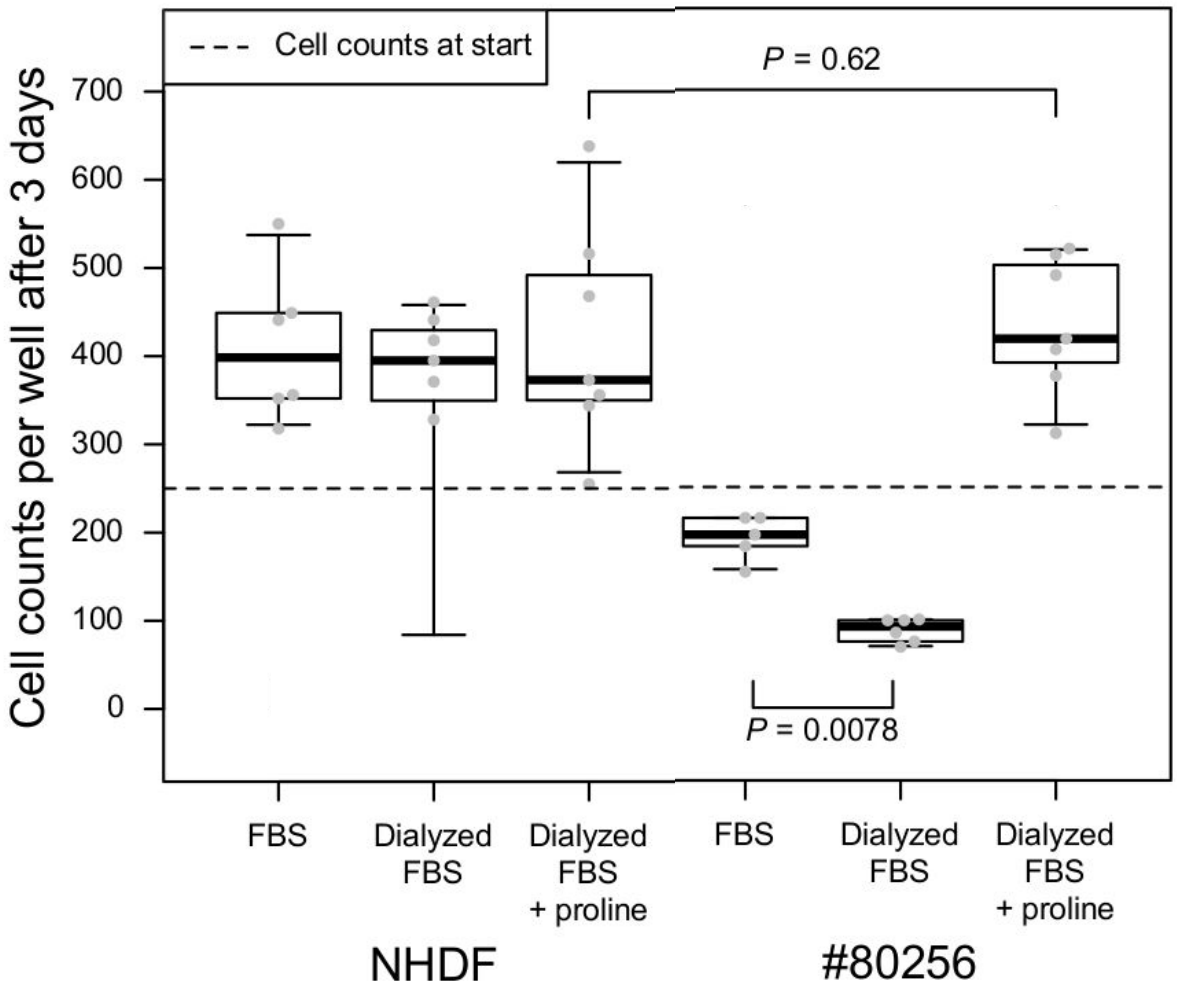


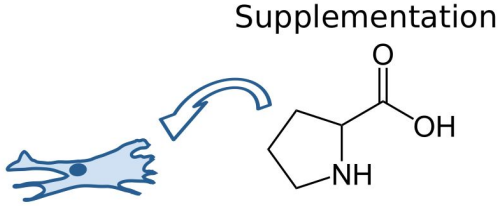


- MAE :=
  Mono-allelic expression
- Read counts per variant
- Test for MAE
  within each sample
- MAE := P-value < 0.05
  & Allele-frequency >= 0.8

⇒ rare MAE in ALDH18A1 for sample #80256

Rescue through proline supplementation



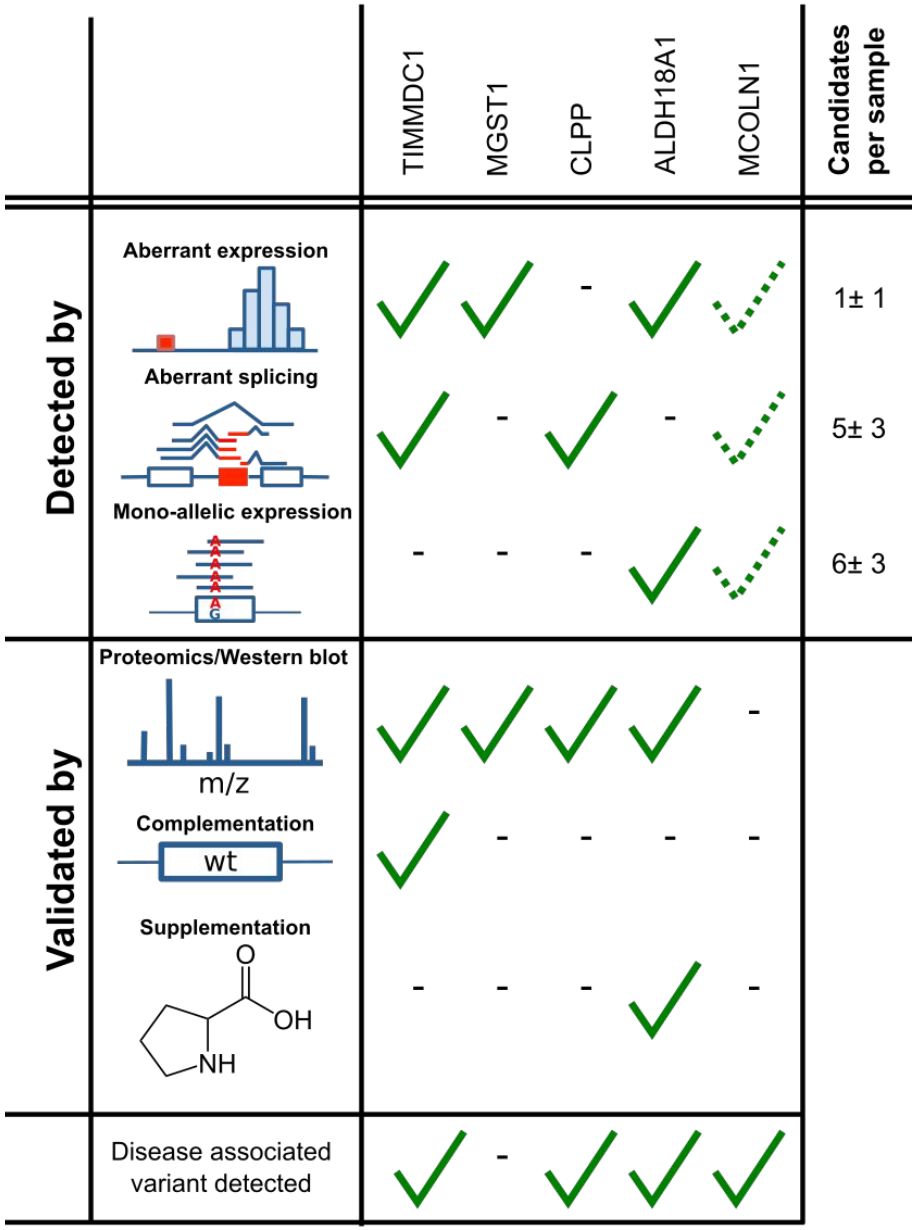


Robert Kopajtich

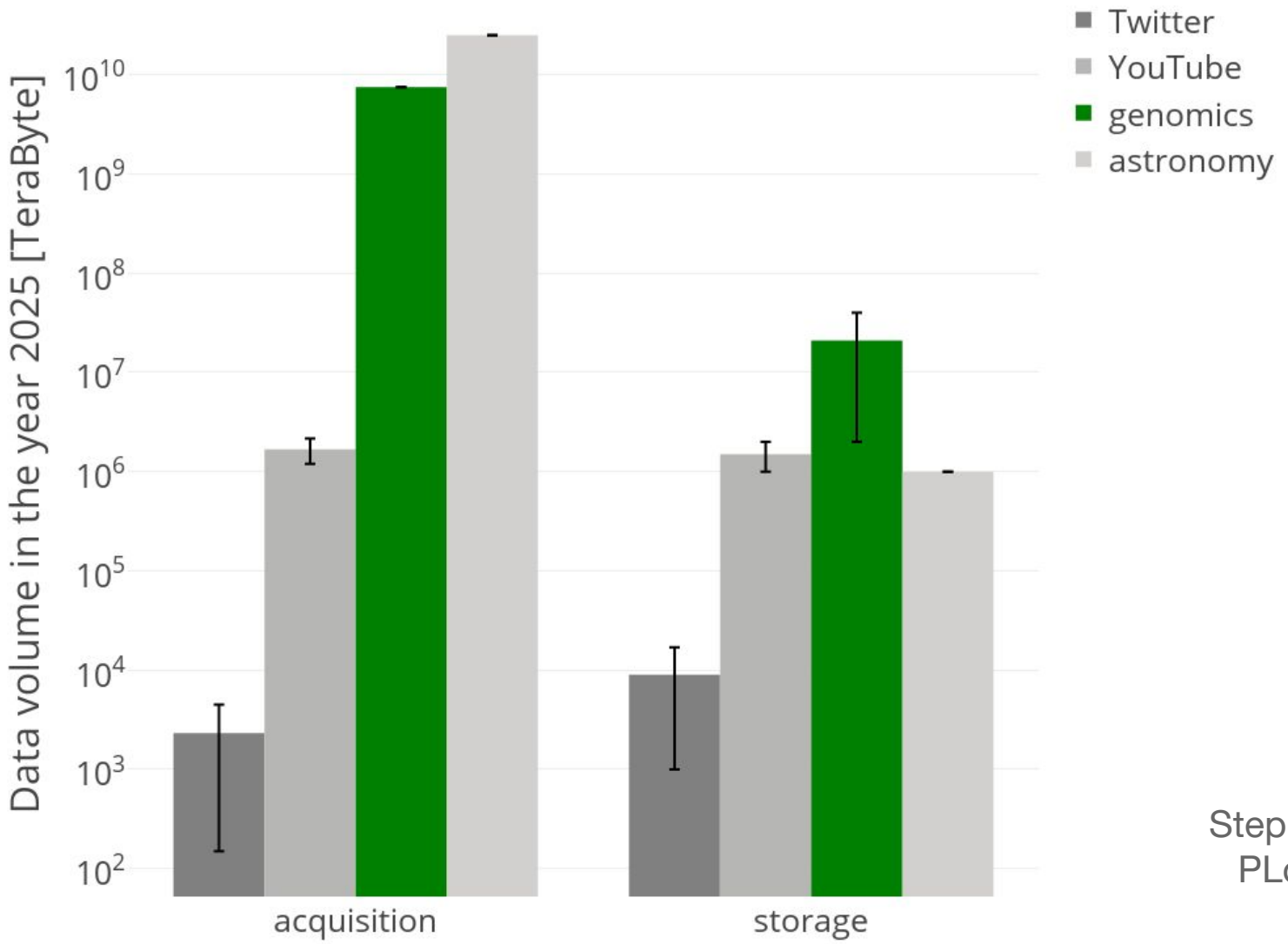
ALDH18A1: Mitochondrial; involved in the biosynthesis of proline

Conclusion

- RNA-Seq improves the diagnosis rate of WES by at least 10% (5/48)
- Strong candidates for 75% (36/48) of unsolved
- RNA-Seq can detect
  - Aberrant expression
  - Aberrant splicing
  - Mono-allelic expression
- Weak (“cryptic”) splice sites are susceptible to variation



Year 2025: sequenced genomes for 25% of people in developed nations

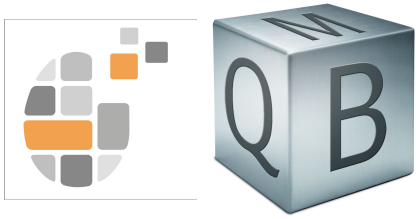


Stephens 2015
PLoS biology

Acknowledgements

Julien Gagneur and his lab:

- gagneurlab.in.tum.de
- Technical University of Munich
- Graduate School
  Quantitative Biosciences Munich



TAC Members

Ulrike Gaul, Klaus Foerstemann, Christoph Klein, Eckhard Wolf, Dietmar Martin, Veit Hornung, Franz Herzog

mitOmics

Bioinformatische und statistische Analyse genomischer Daten von Patienten mit mitochondrialen Krankheiten zur Identifizierung kausaler Mutationen und Pathways

Statistical Multi-Omics Understanding
Grant Agreement no. 633974

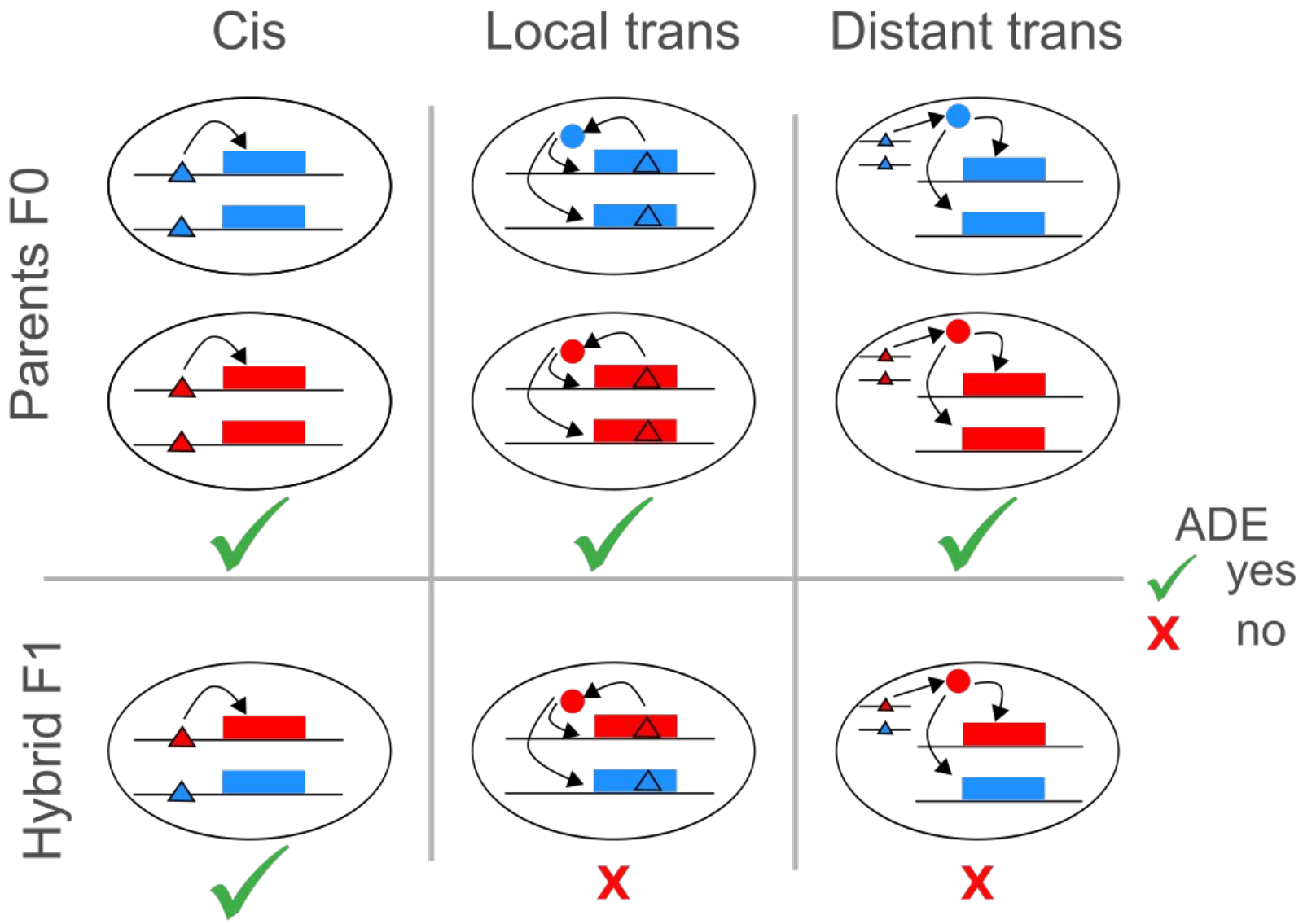
The groups of Holger Prokisch and Tim Strom

The group of Lars Steinmetz

Supplement

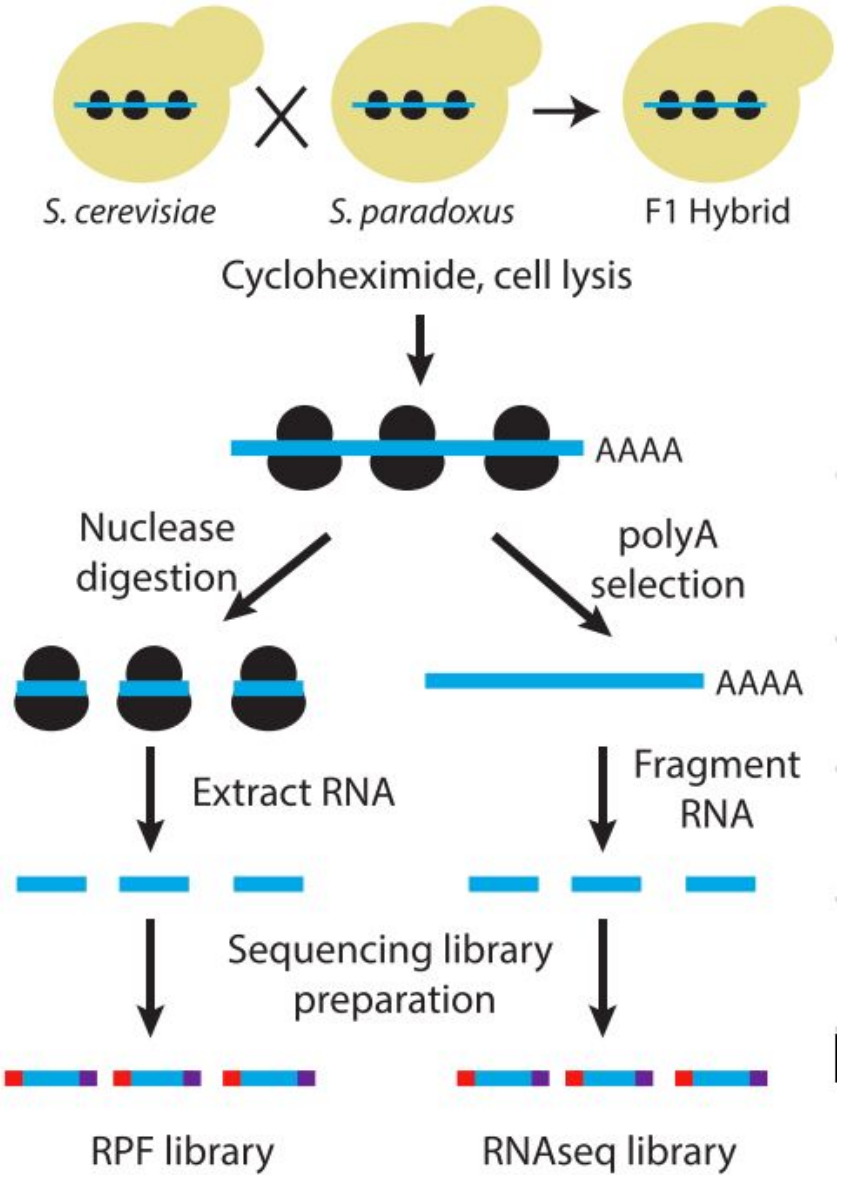
Negative feedback

Dissecting cis and trans by comparing ADE in parents and hybrid

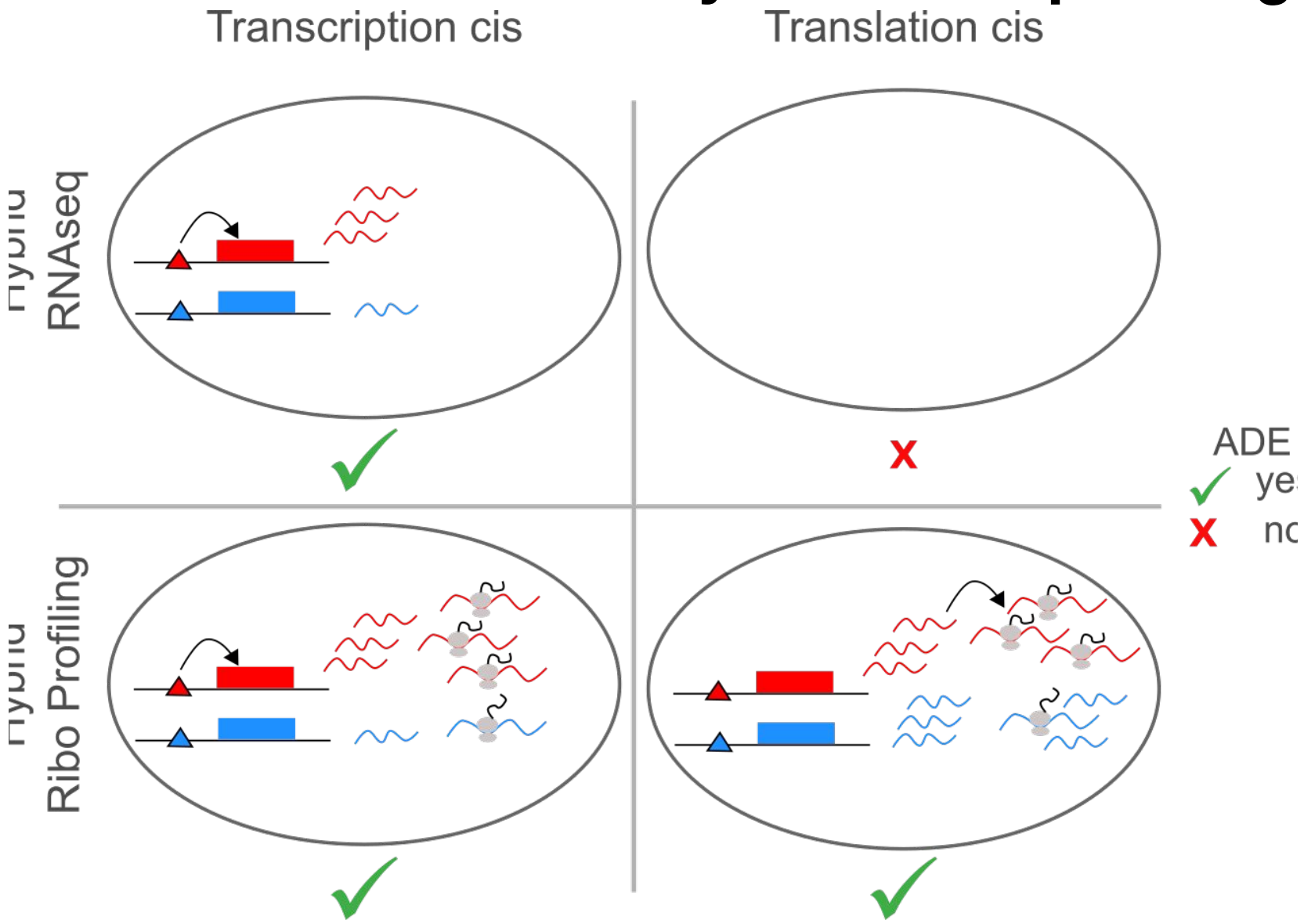


Ribosome profiling protocol

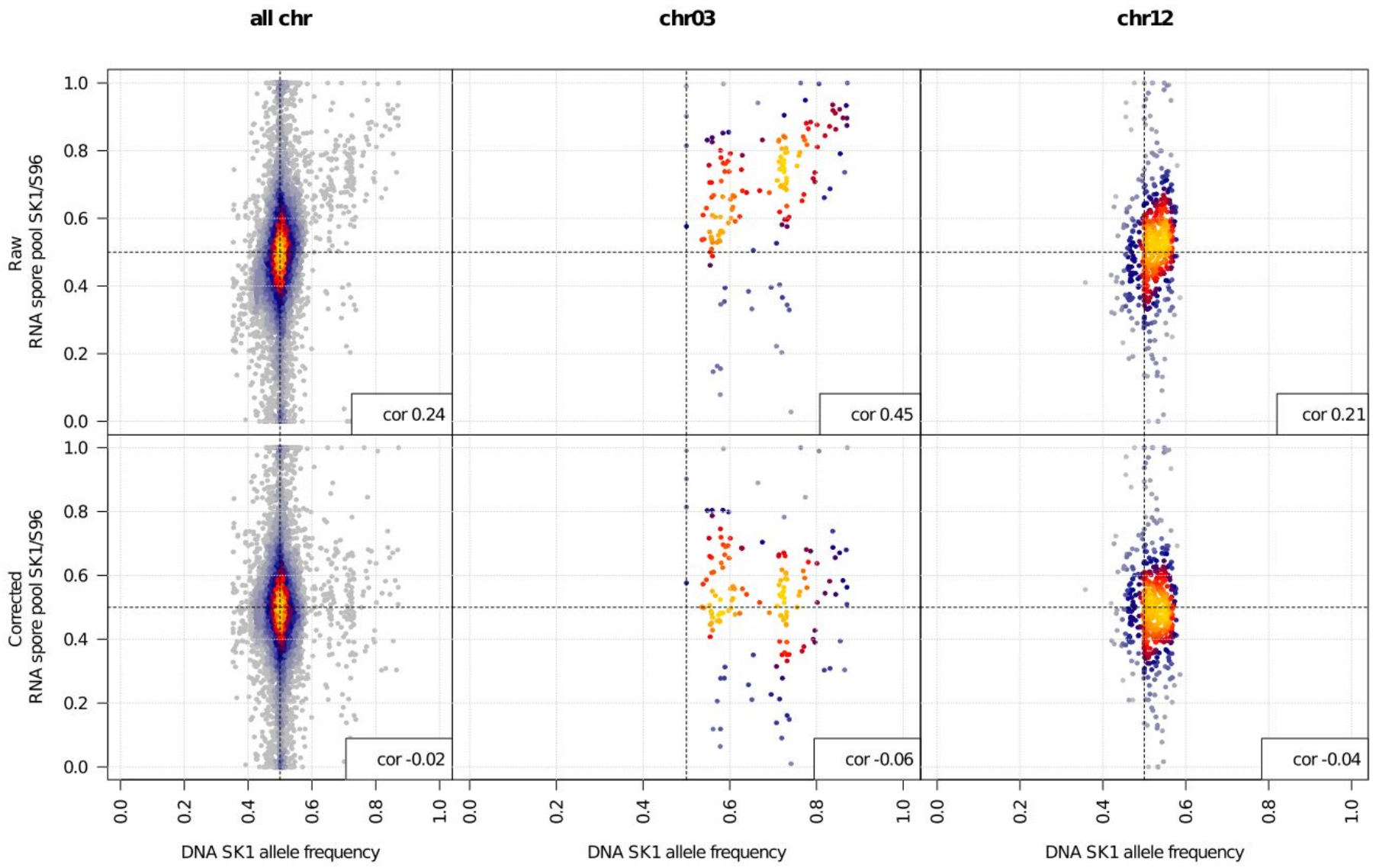
McManus 2014
Genome research



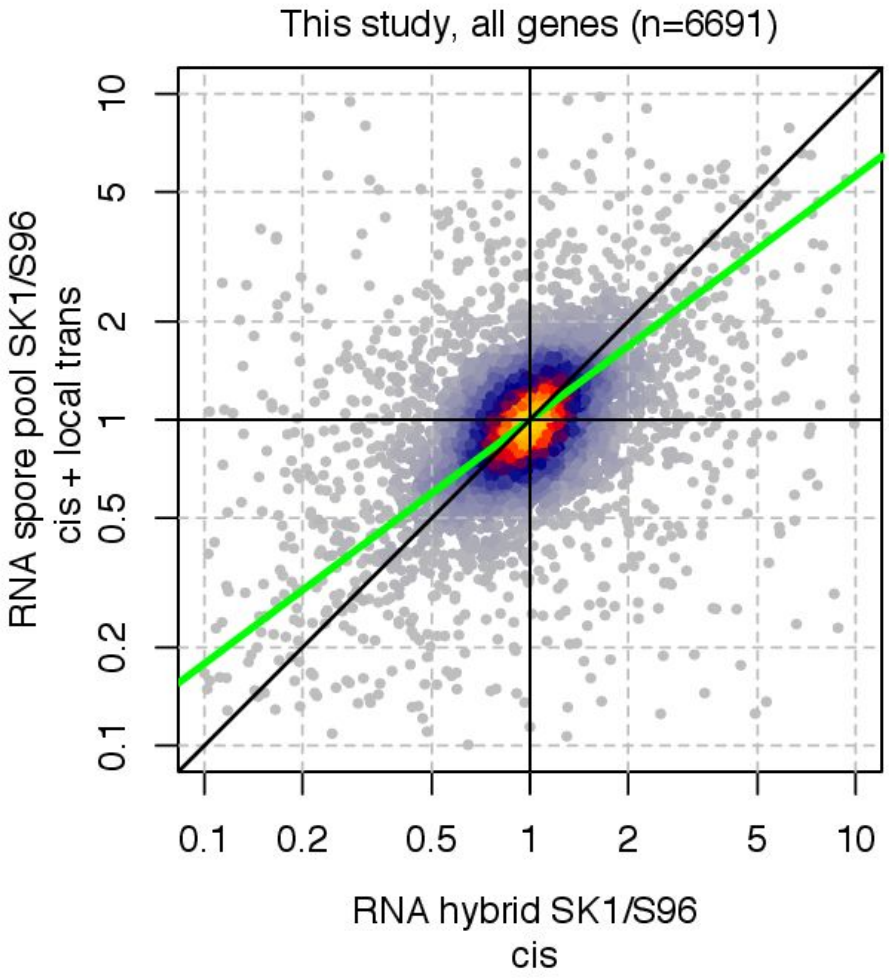
Dissecting transcriptional cis from translational cis by ribosome profiling

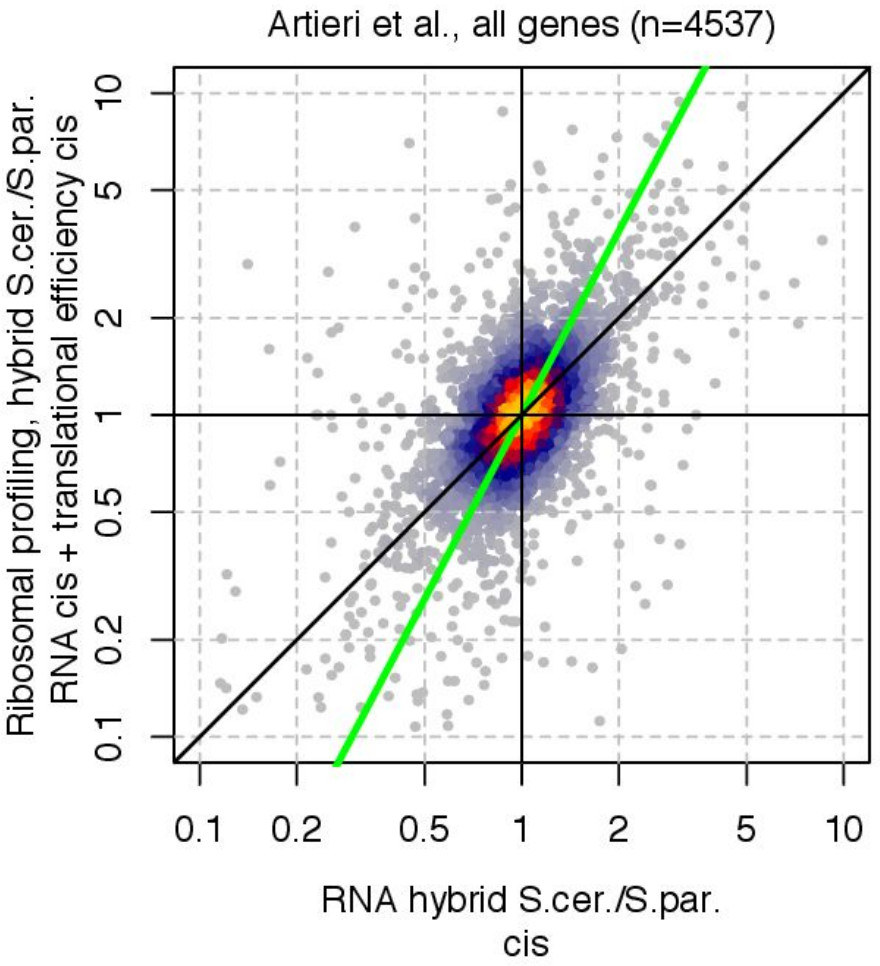


Normalization of DNA allelic imbalances

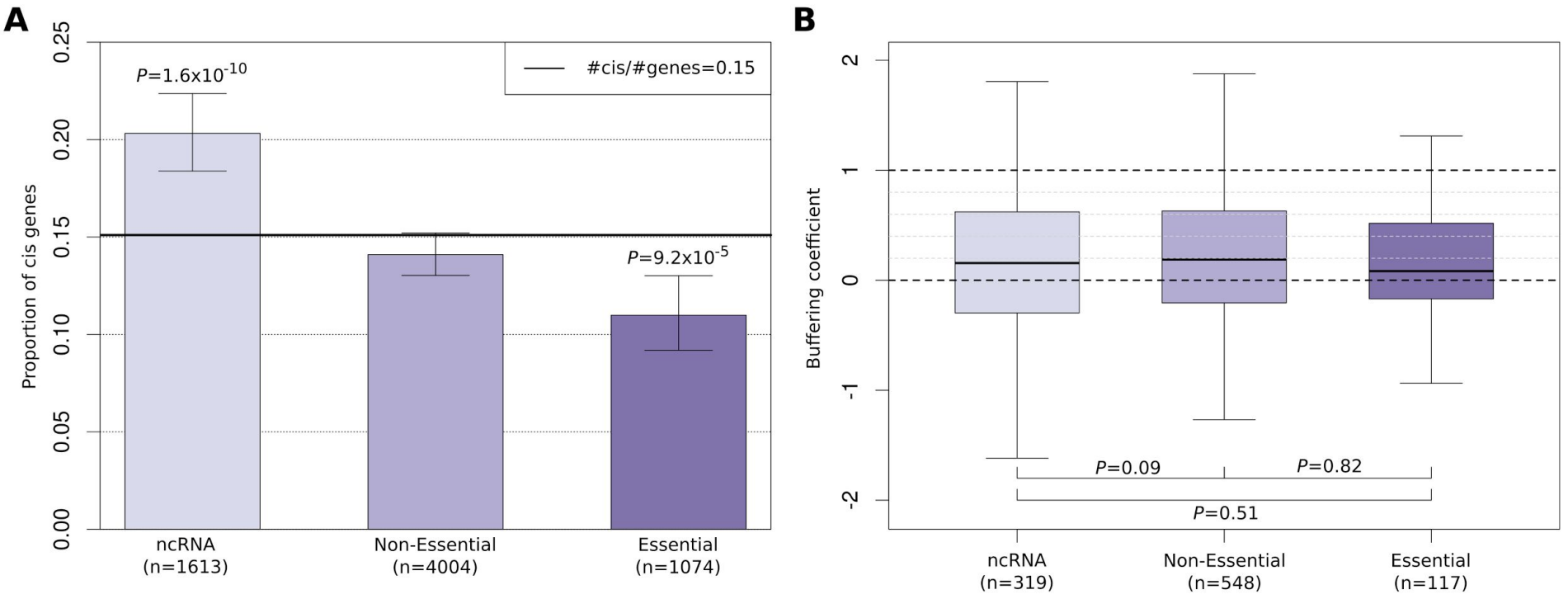


Raw data cis vs local trans

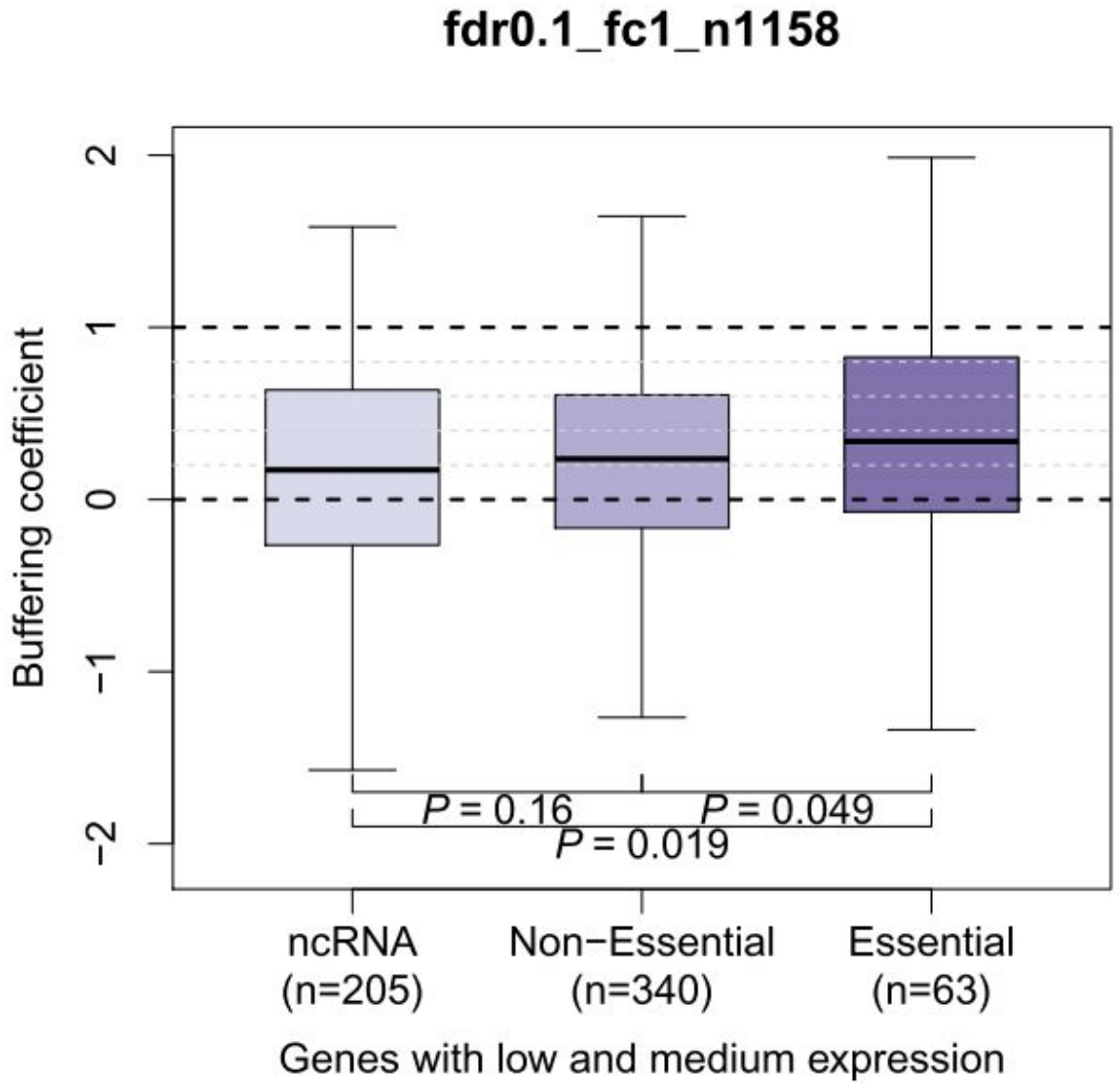




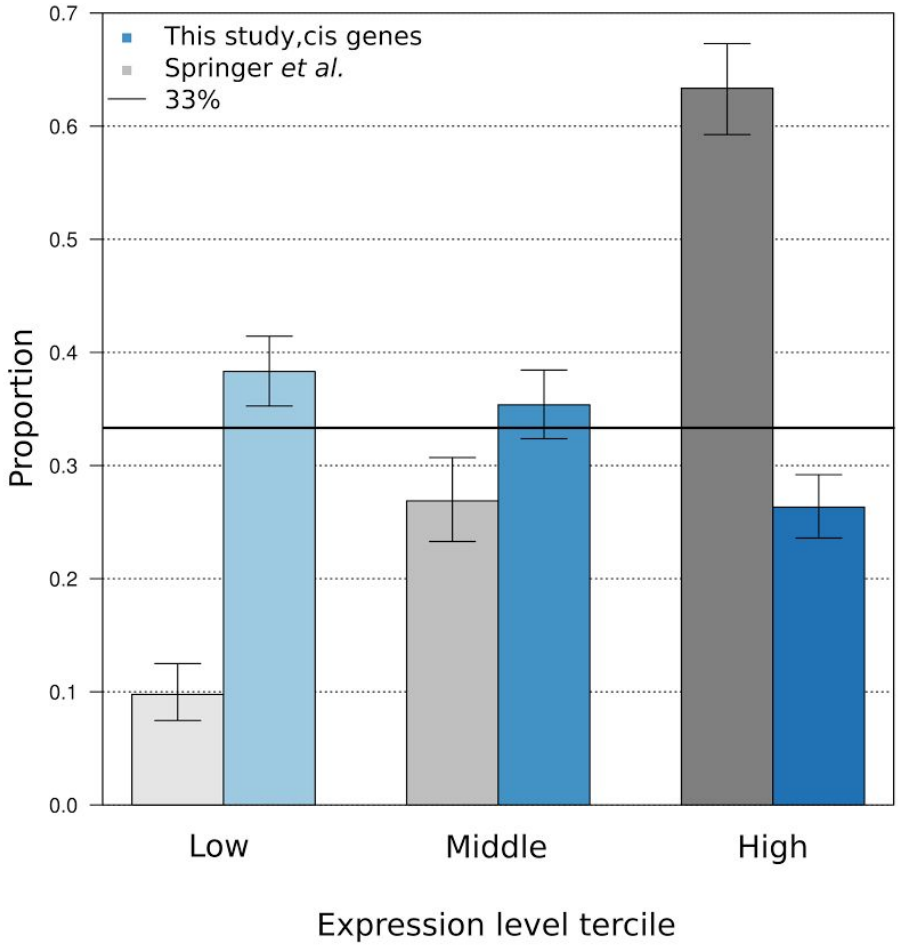
Lack of association between buffering and functional relevance genome-wide

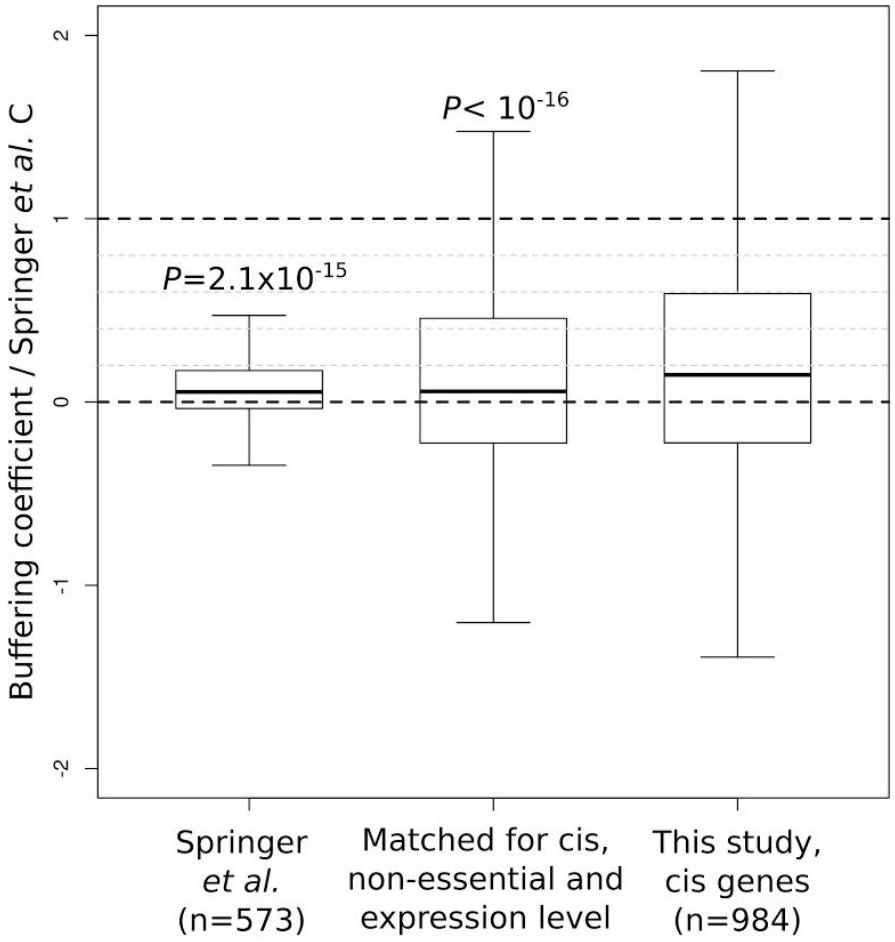


Different cis cutoff: buffering vs gene type



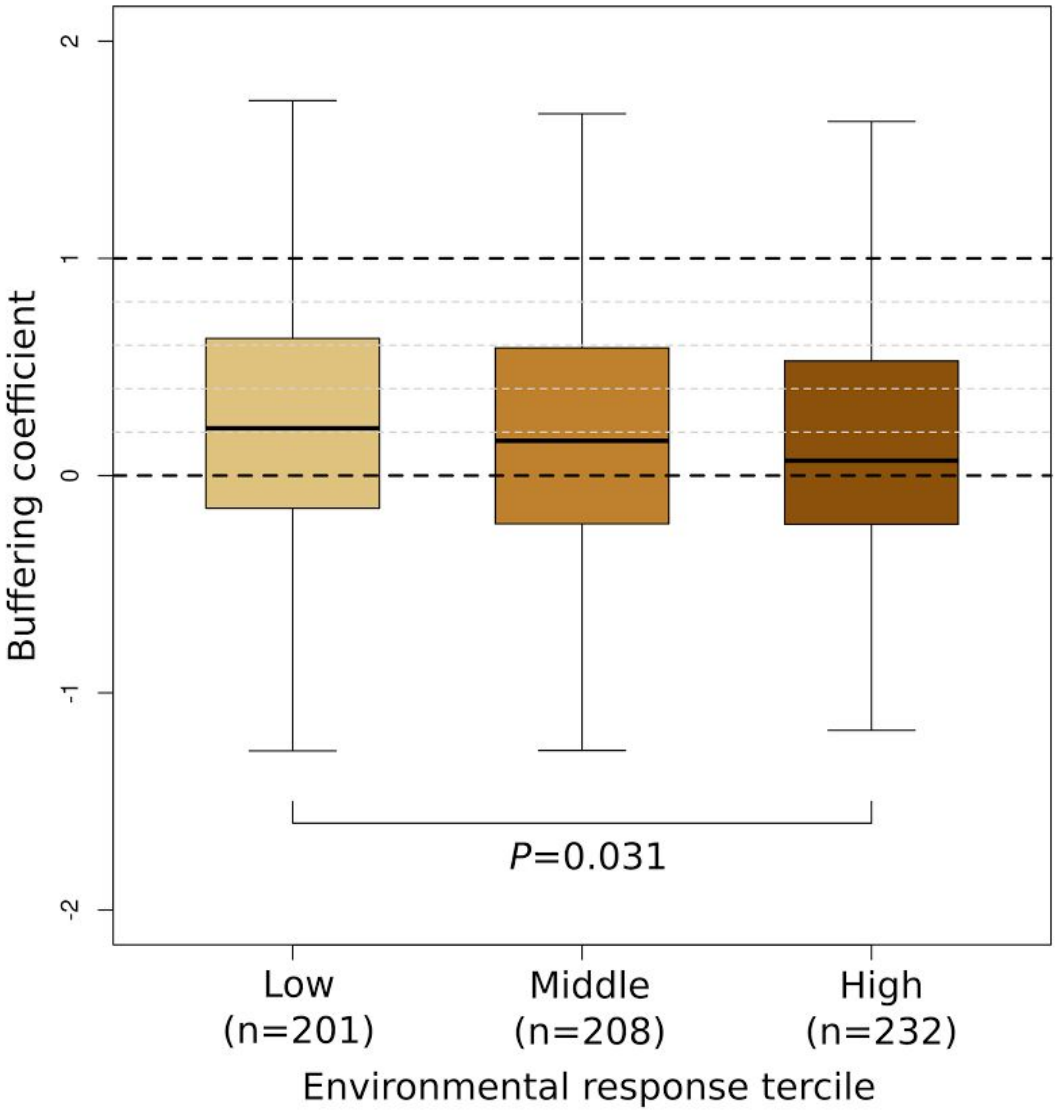
Local trans buffering is primarily due to negative feedback





Springer 2010 Mol Syst Biol

Negative feedback confers robustness to environmental variation



Tirosh 2009 Science

Modeling local trans

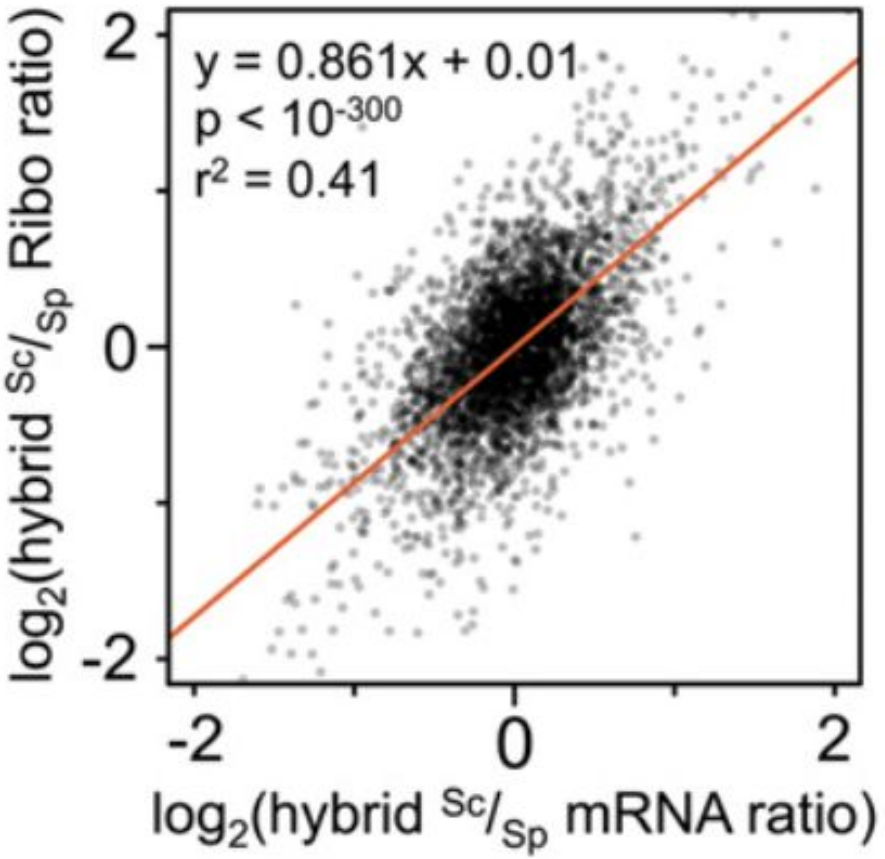
$$K_{i,j} \sim \mathrm{NB}(\mu_{i,j}, \alpha_i)$$

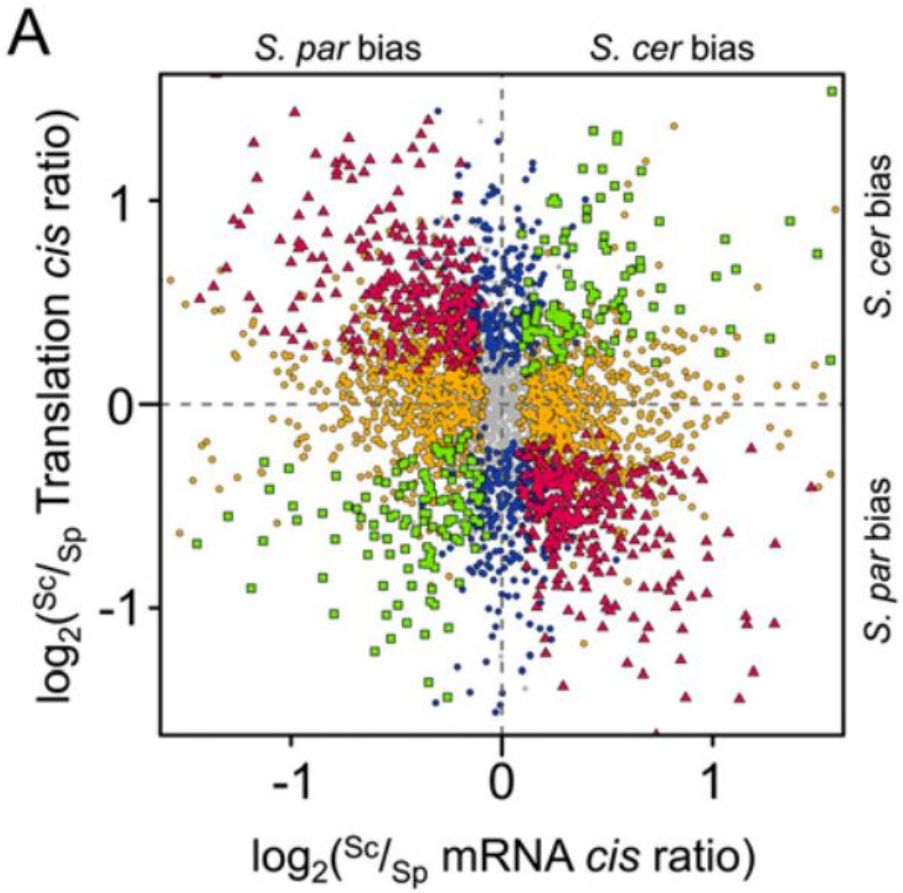
$$\mu_{i,j} = s_j \times f_{i,j} \times q_{i,j} \times l_i$$

$$\log_2(q_{i,j}) = \boldsymbol{\beta}_i^0 + \boldsymbol{\beta}_i^{cis}\mathbf{x}_{i,j}^{cis} + \boldsymbol{\beta}_i^{localtrans}\mathbf{x}_{i,j}^{localtrans} + \boldsymbol{\beta}_i^{nuis\,T}\mathbf{x}_{i,j}^{nuis}$$

SAMPLE \ FACTOR	cis	local trans	diploid	hybrid B	spore B
hybrid A only SK1	1	1	1	0	0
hybrid A only S96	0	1	1	0	0
hybrid B only SK1	1	1	1	1	0
hybrid B only S96	0	1	1	1	0
spore A only SK1	1	1	0	0	0
spore A only S96	0	0	0	0	0
spore B only SK1	1	1	0	0	1
spore B only S96	0	0	0	0	1

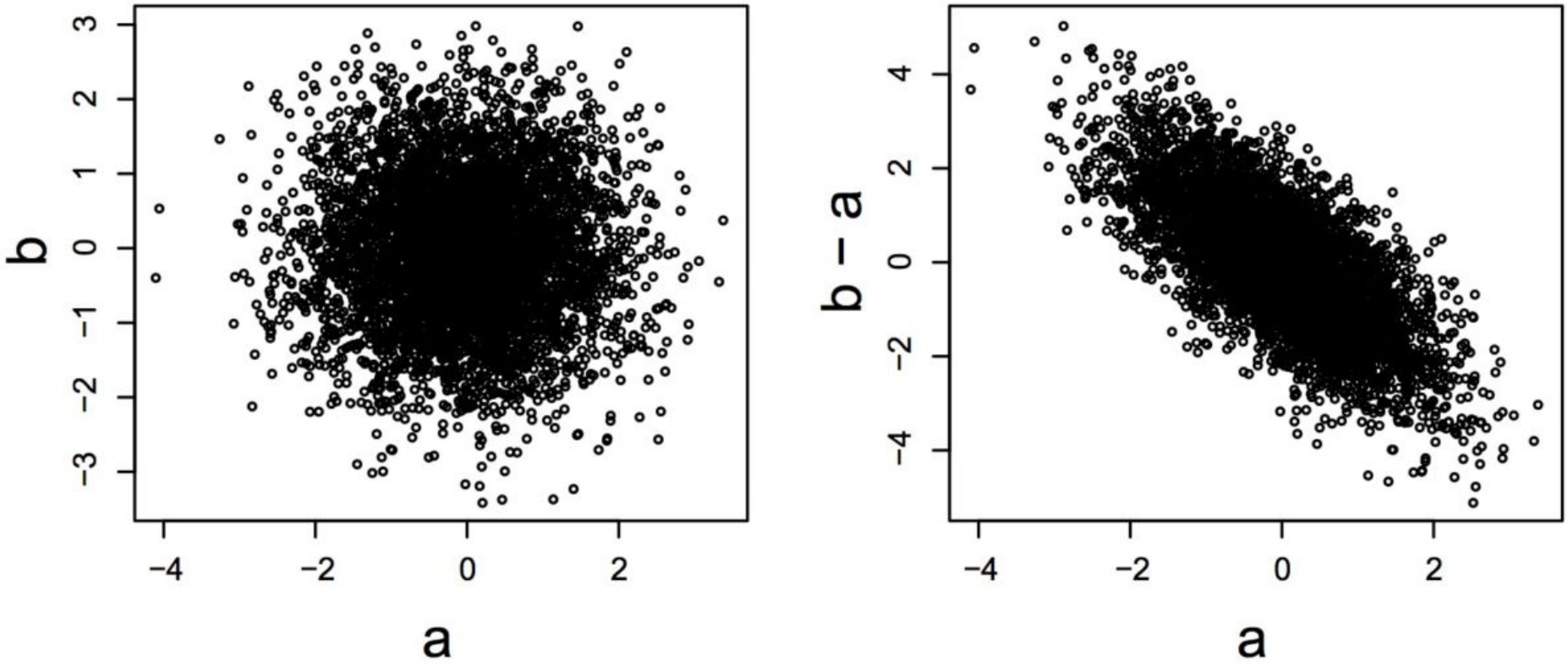
Artieri2014: Misleading anti-correlation





Artieri and Fraser 2014 Genome research

Spurious anti-correlation



- two random samples a and b of size 5,000
- standard normal distribution with mean = 0 and standard deviation = 1
- correlations between a log ratio and its denominator ⇒ Spurious correlations

Fig S4, Albert et al. 2014 PLoS genetics

Modeling translational efficiency

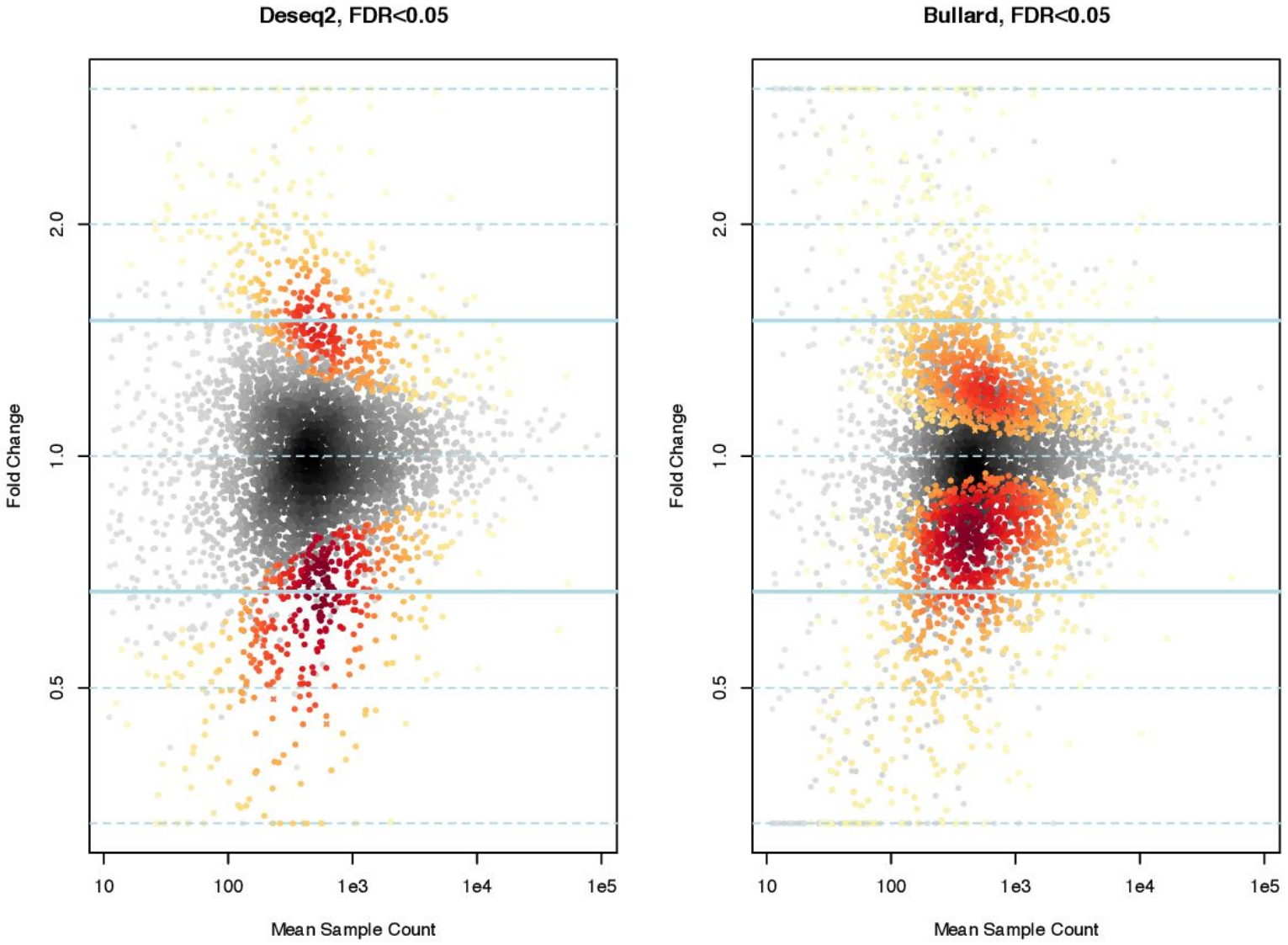
$$K_{i,j} \sim \mathrm{NB}(\mu_{i,j}, \alpha_i)$$

$$\mu_{i,j} = s_j \times q_{i,j}$$

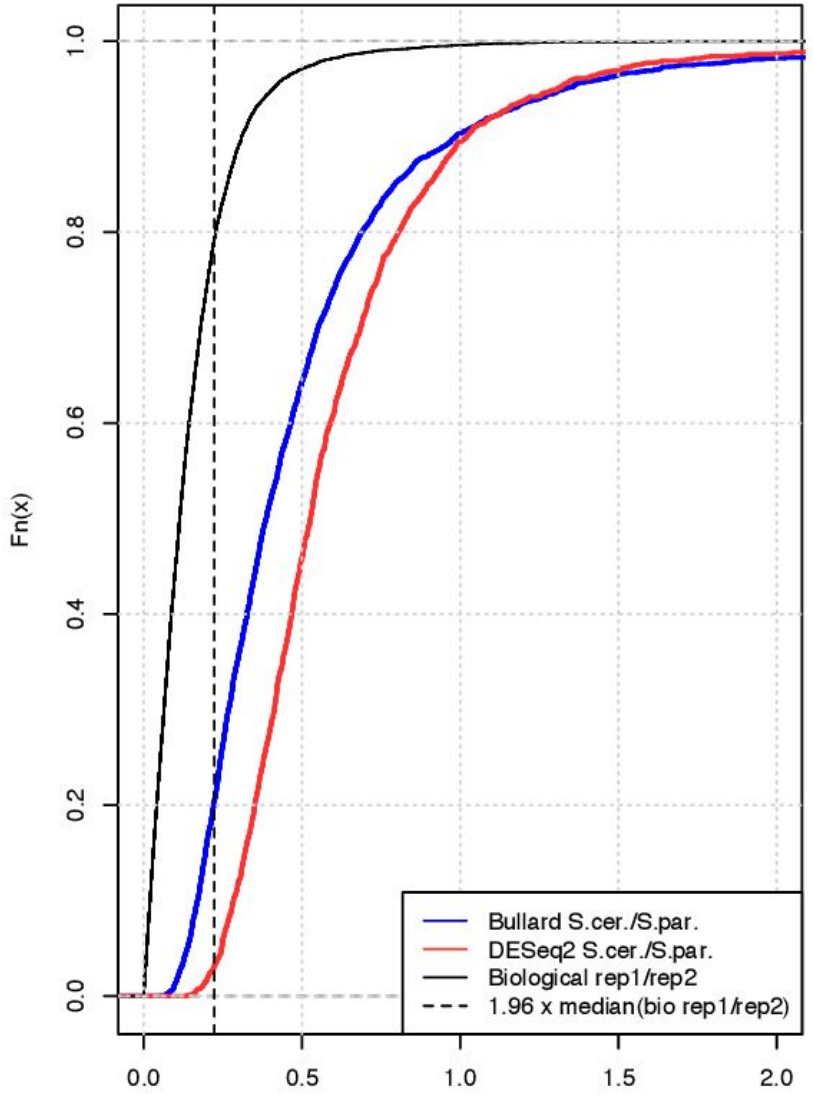
$$\log_2(q_{i,j}) = \boldsymbol{\beta}_i^0 + \boldsymbol{\beta}_i^{cisRNA}\mathbf{x}_{i,j}^{cisRNA} + \boldsymbol{\beta}_i^{cisTE}\mathbf{x}_{i,j}^{cisTE} + \boldsymbol{\beta}_i^{nuis^T}\mathbf{x}_{i,j}^{nuis}$$

SAMPLE	RNA cis	TE cis	RNA bias	hybrid rep2
hybrid RNA 1 SCER	1	0	1	0
hybrid RNA 2 SCER	1	0	1	1
hybrid RNA 1 SPAR	0	0	1	0
hybrid RNA 2 SPAR	0	0	1	1
hybrid RIBO 1 SCER	1	1	0	0
hybrid RIBO 2 SCER	1	1	0	1
hybrid RIBO 1 SPAR	0	0	0	0
hybrid RIBO 2 SPAR	0	0	0	1

DESeq2 vs Bullard et al 2010 by MA-plot



DESeq2 vs Bullard et al 2010 by cdf



Absolute log2 fold change at nominal P-value 0.05

Supplement

Genetic diagnosis

Sample numbers by measurement technique

Method	Diagnosed	Not diagnosed	Total
RNA	57	48	105
RNA & WES	40	48	88
RNA & proteomics	11	20	31

Unaffected tissue: fibroblasts

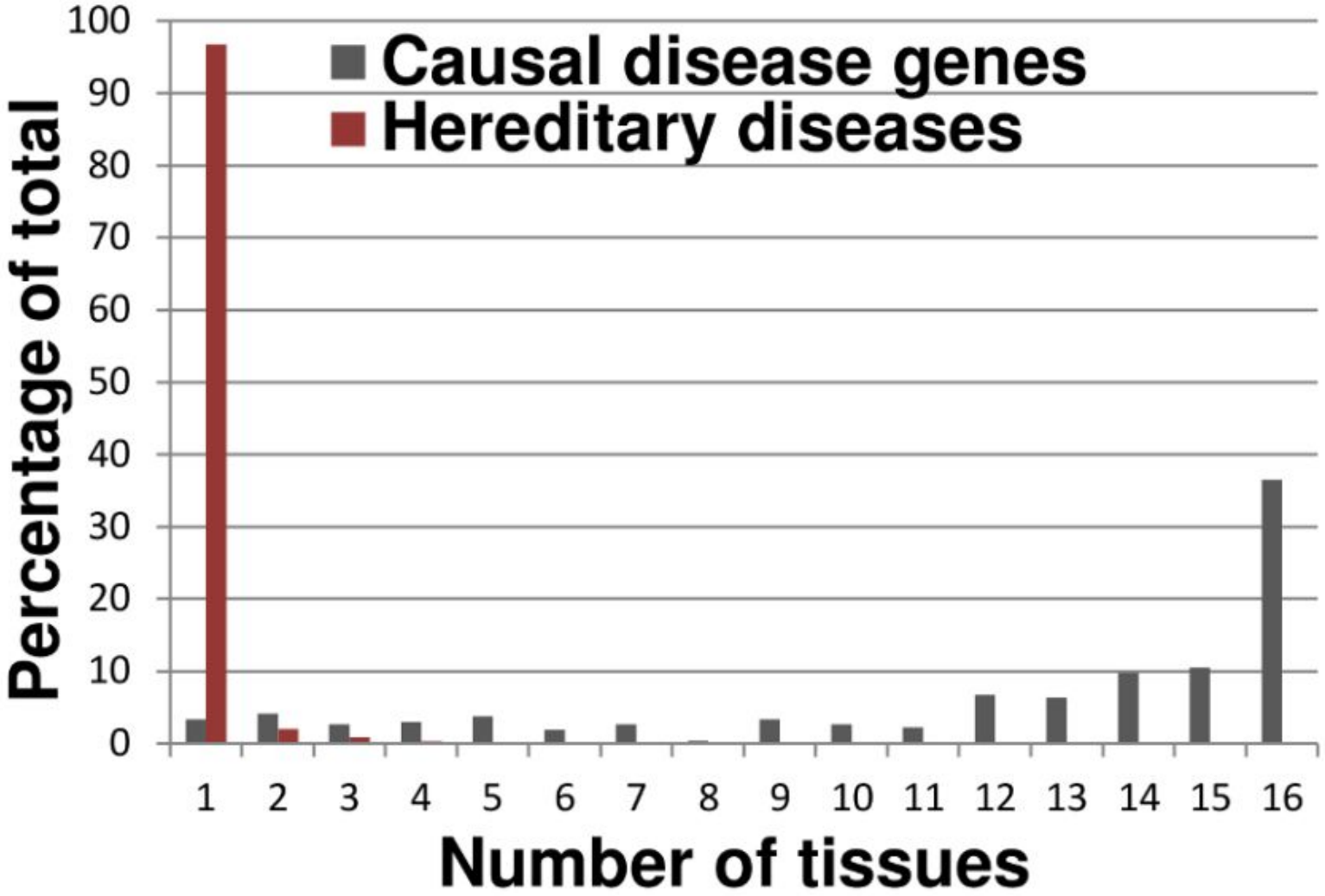
- Byproducts of muscle biopsies (routine in clinic)
  → biochemical diagnosis of mitochondrial disorders with enzymatic assays

- Limited accessibility of affected tissue,
  e.g. brain, heart, skeletal muscle or liver

- RNA defects detectable
  - physiological consequences on fibroblasts might be negligible
  - regulatory consequences on other genes might be limited

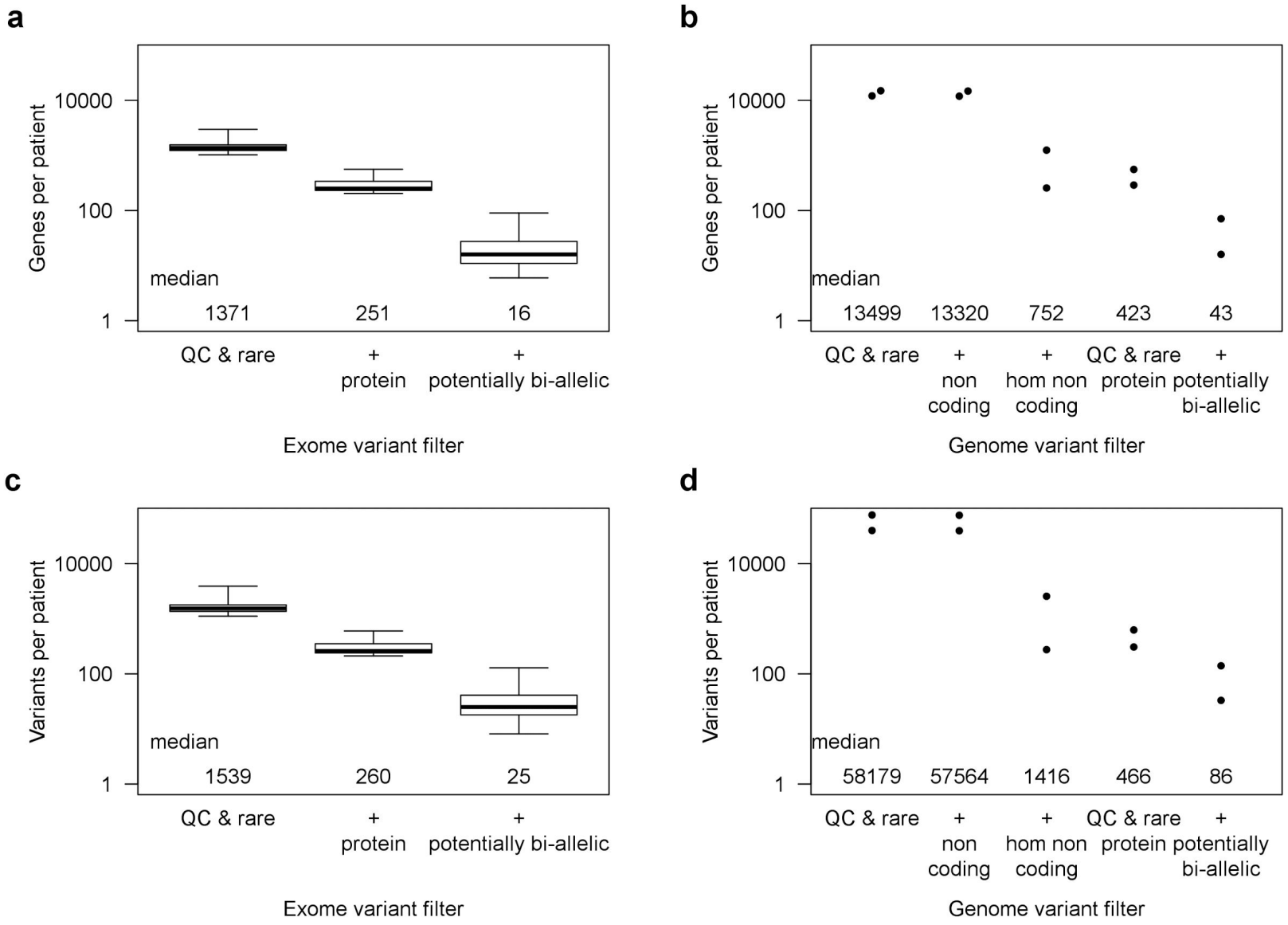
- Perform perturbation and complementation tests in cell lines
  → rapid demonstration of candidate variant's role

No tissue-specific expression of disease causing genes



Barshir 2014 PLoS genetics

Standard variant prioritization



Recovery rate of diagnosed patients

- 40 patients diagnosed before our study
  with WES & RNAseq available

- 3/4 homozygous stop by aberrant expression
- 7/8 splice variants detected by aberrant splicing
- 3/6 missense + stop,frame-shift by MAE

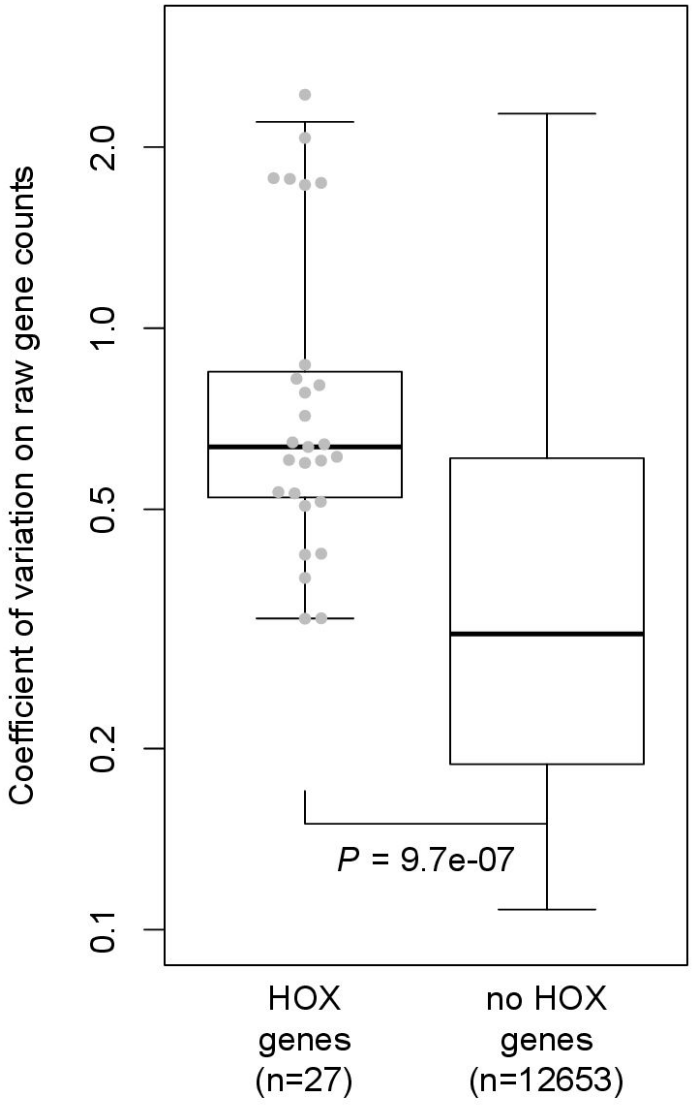
- 0/14 homozygous missense variants
  with detected RNA defect

Modeling aberrant expression

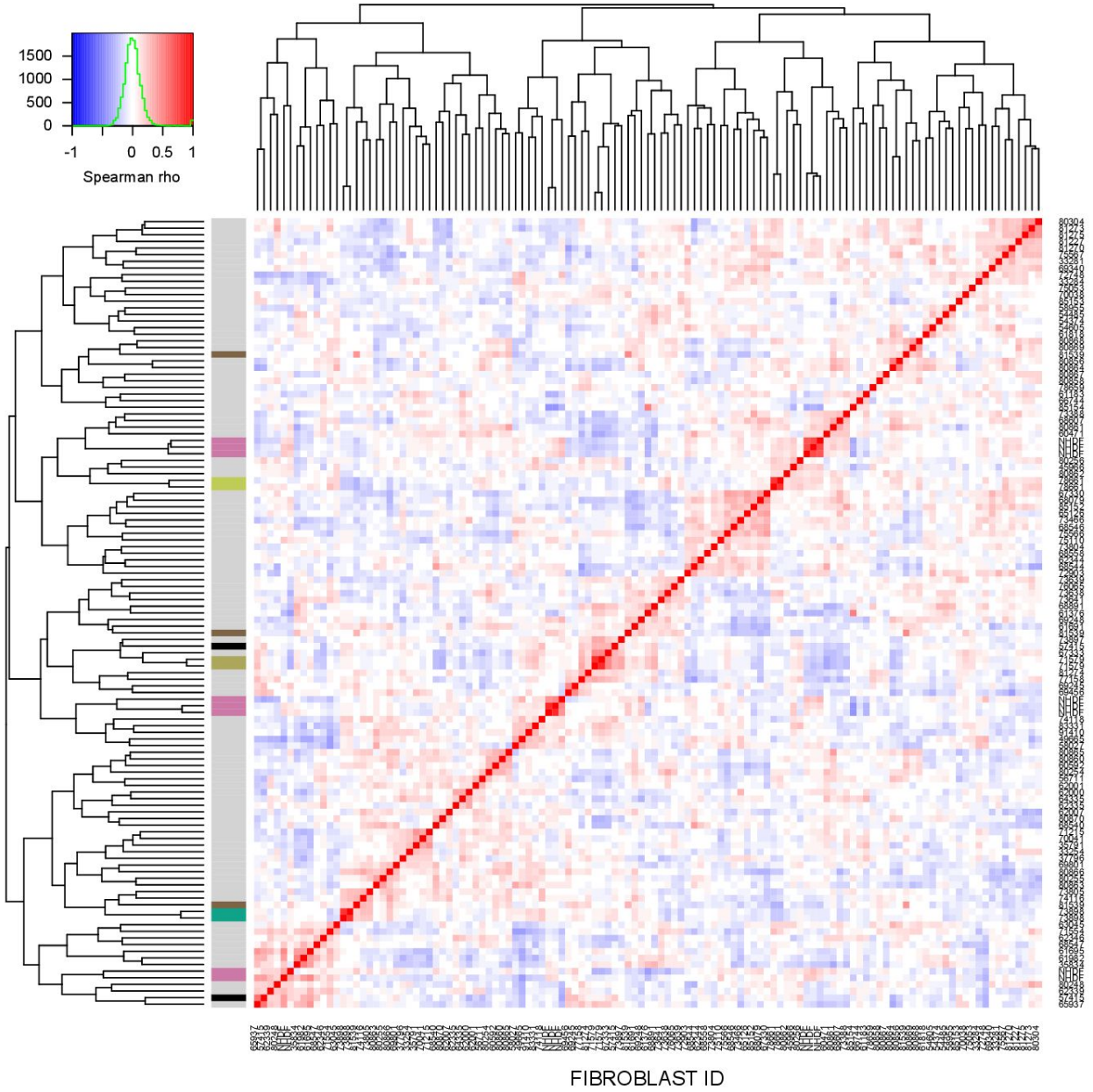
$$K_{i,j} \sim NB(s_j \times q_{i,j}, \alpha_i)$$

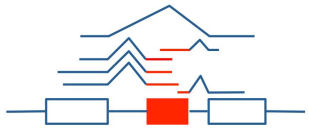
$$\log_2(q_{i,j}) = \beta_i^0 + \beta_i^{condition}\mathbf{x}_{i,j}^{condition} + \beta_i^{batch}\mathbf{x}_{i,j}^{batch} + \beta_i^{sex}\mathbf{x}_{i,j}^{sex} + \beta_i^{hox}\mathbf{x}_{i,j}^{hox}$$

High variation in HOX genes



Fully normalized sample correlation



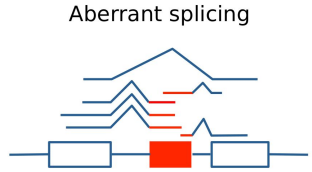


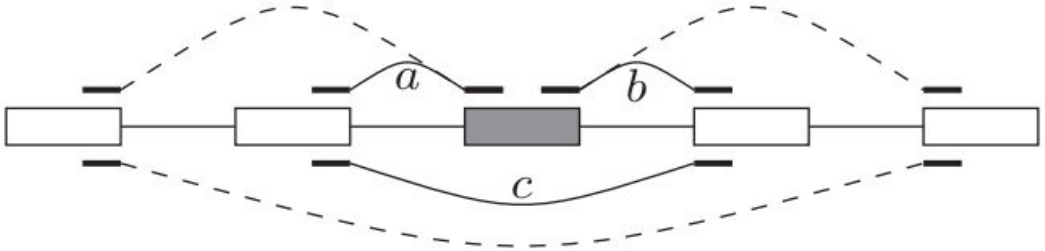
Leafcutter adaptation

- Extract split reads per sample
- Cluster junctions over all samples ⇒ Genes
  - Keep low expressed junctions ($\Psi > 0.0001$)
  - Keep junctions supported by ≥ 10 reads by ≥ 1 sample
- Test for differential splicing
  - 1 versus rest (all other samples)
  - Include a pseudo sample in control group to have all junctions represented (conservative)

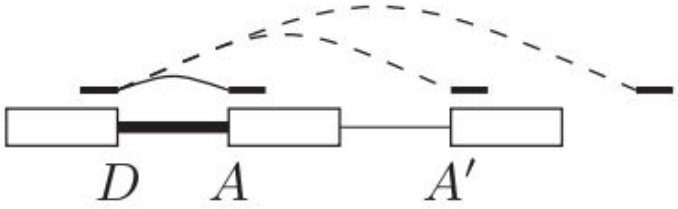
Li, Knowles, Pritchard bioRxiv 2016

How to measure splicing: Percentage Spliced In (PSI - Ψ)

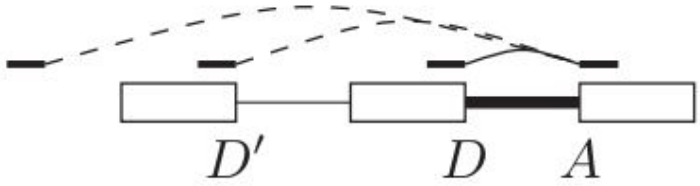




$$\Psi = \frac{a+b}{a+b+2c}$$



$$\psi_5(D,A) = \frac{\mathrm{n}(D,A)}{\sum_{A'} \mathrm{n}(D,A')}$$



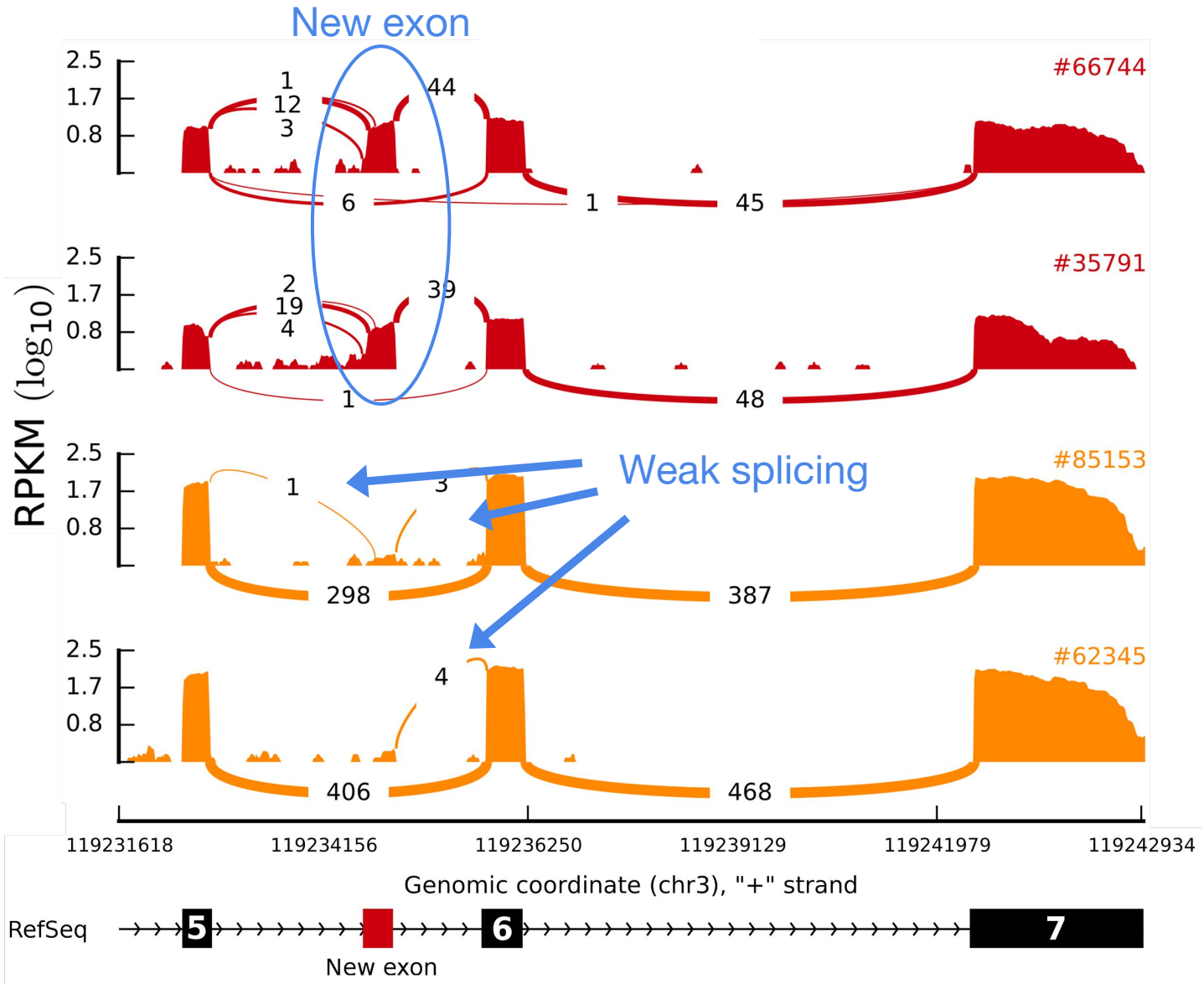
$$\psi_3(D,A) = \frac{\mathrm{n}(D,A)}{\sum_{D'} \mathrm{n}(D',A)}$$

Modeling aberrant splicing

$$P\left(n(D,A)|N(D,A)\right) = \sum_{c\in\{bg,wk,st\}} \sum_{s\in\{0,1,2\}} \pi_{s,c} BetaBin(n(D,A)|N(D,A),\alpha_c,\beta_c)$$

- n(D,A) split reads
- N(D,A) total number of reads
- s: the number of annotated sites
- π_{s,c}: mixing proportions to be group-specific
- beta-binomial distribution
  - binomial distribution with probability of success follows the beta distribution
  - used to capture overdispersion in binomial type distributed data
- alpha, beta: shape parameters for Beta distribution

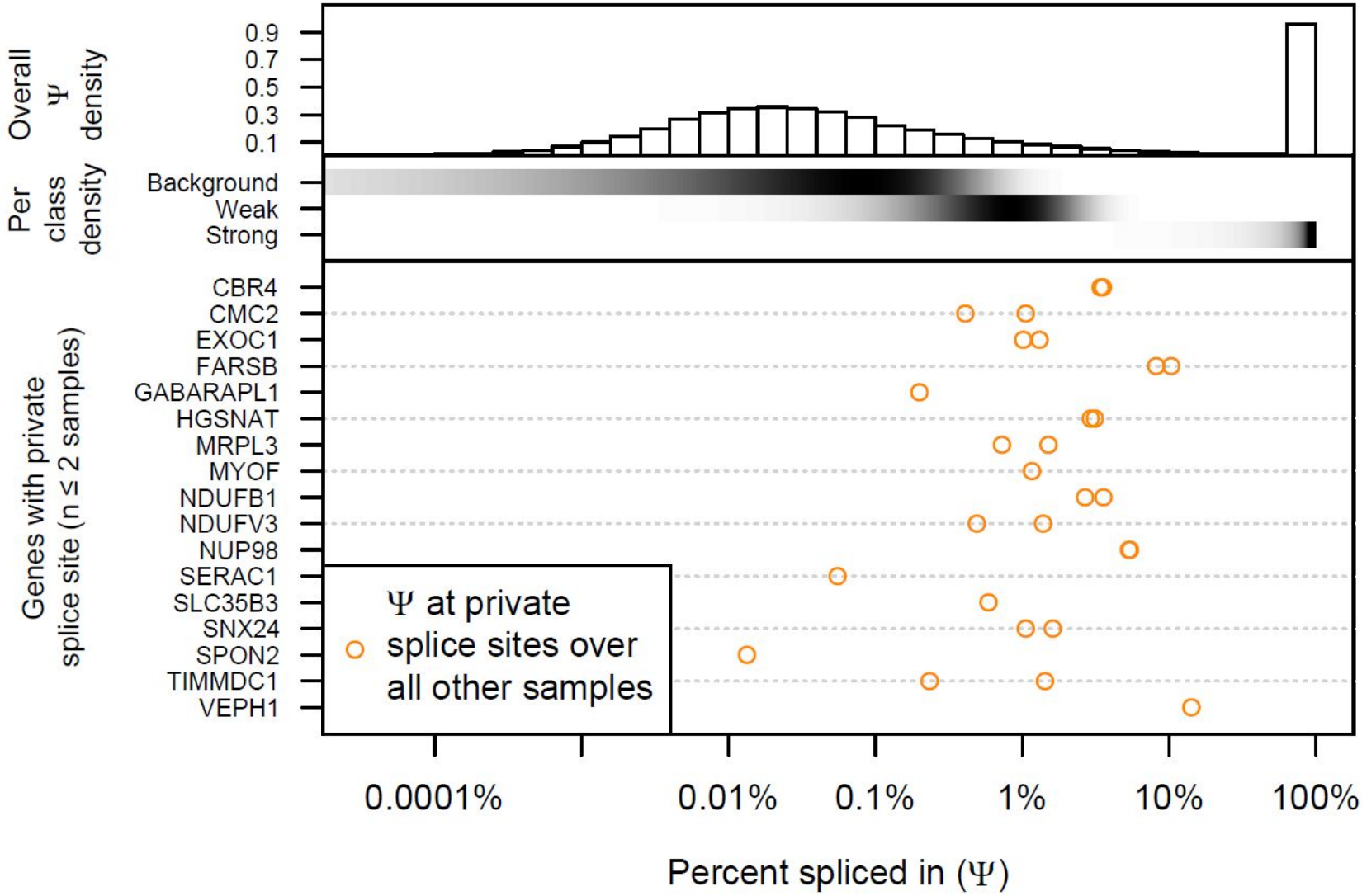
TIMMDC1 - new exon creation



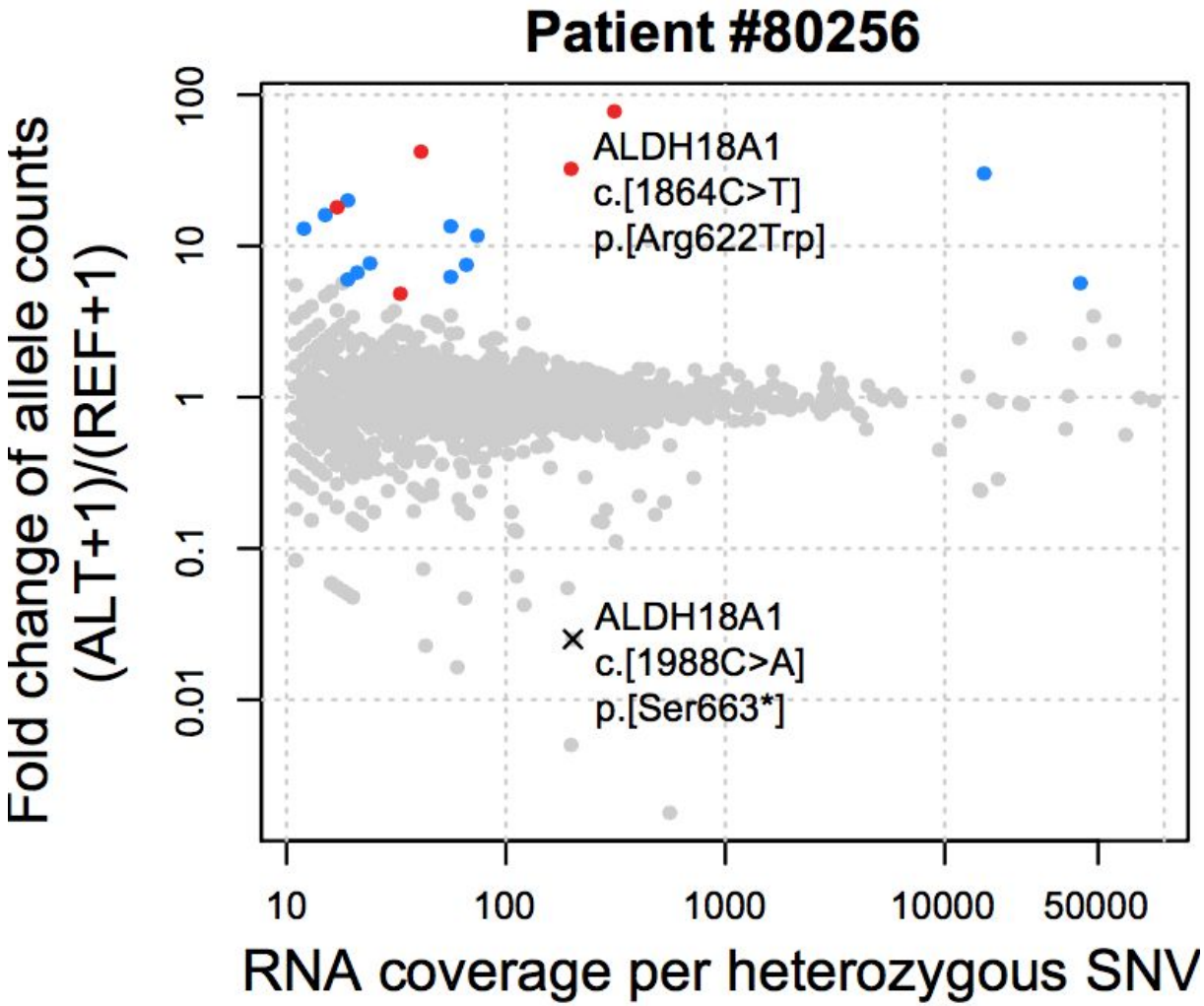
Affected patients

Other samples

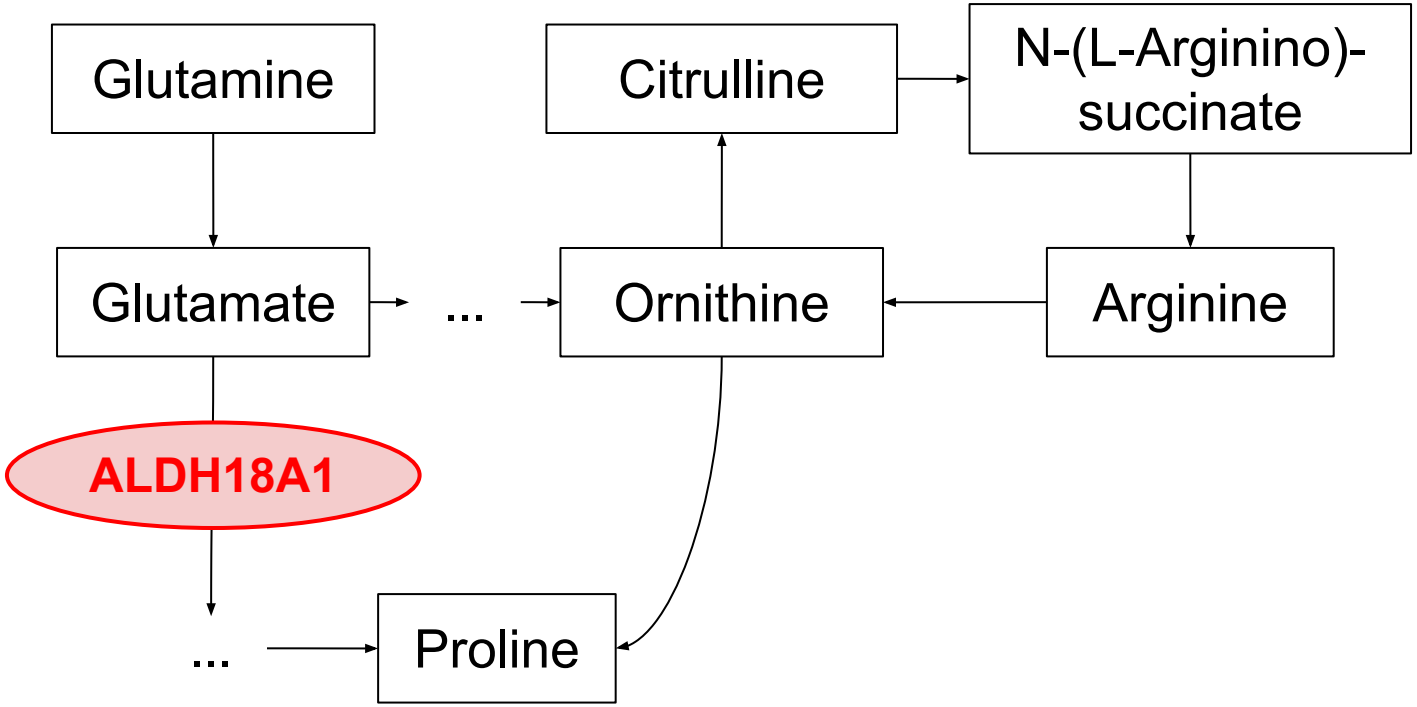
Weak splice sites as origin for new exons



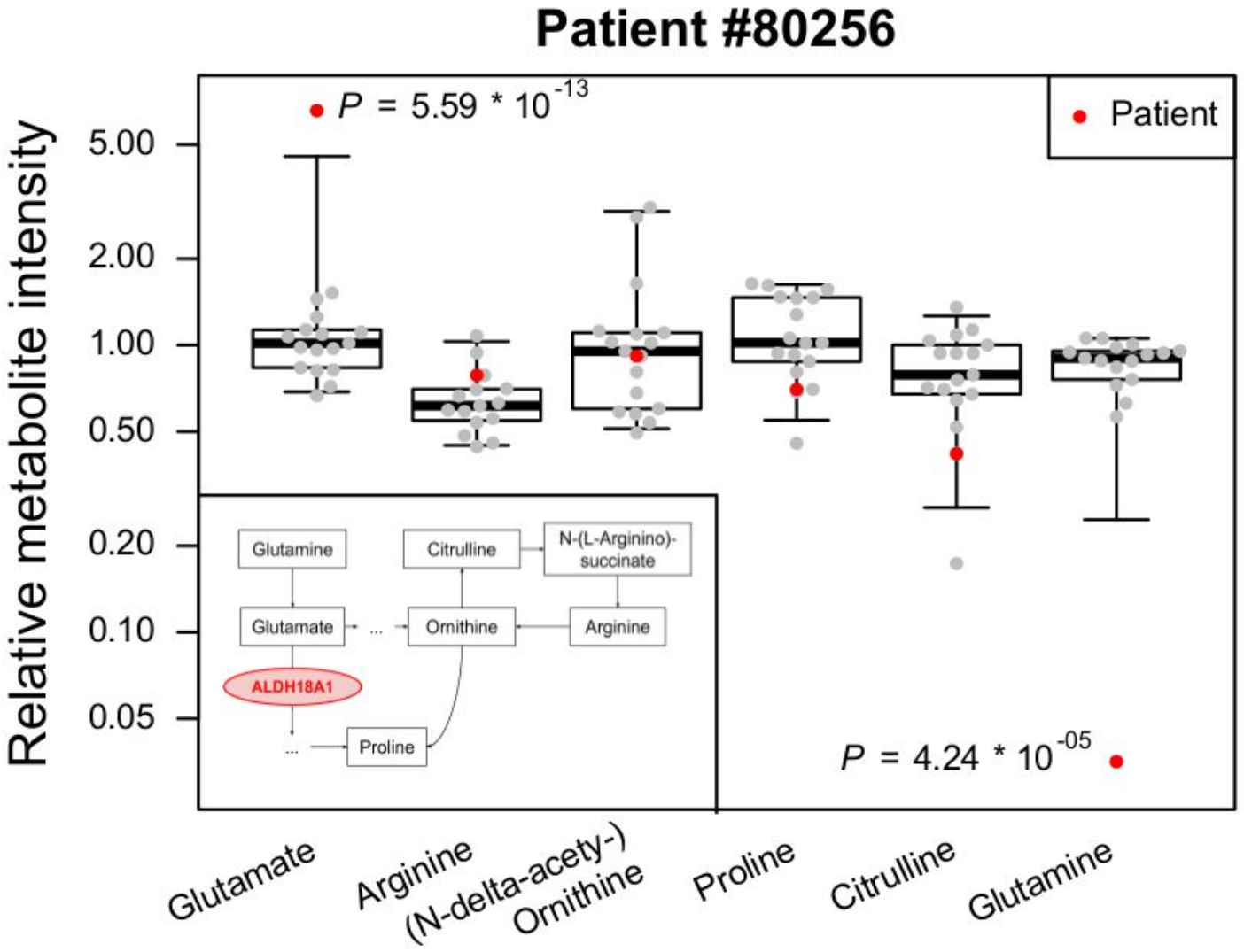
ALDH18A1 for mono-allelic expression



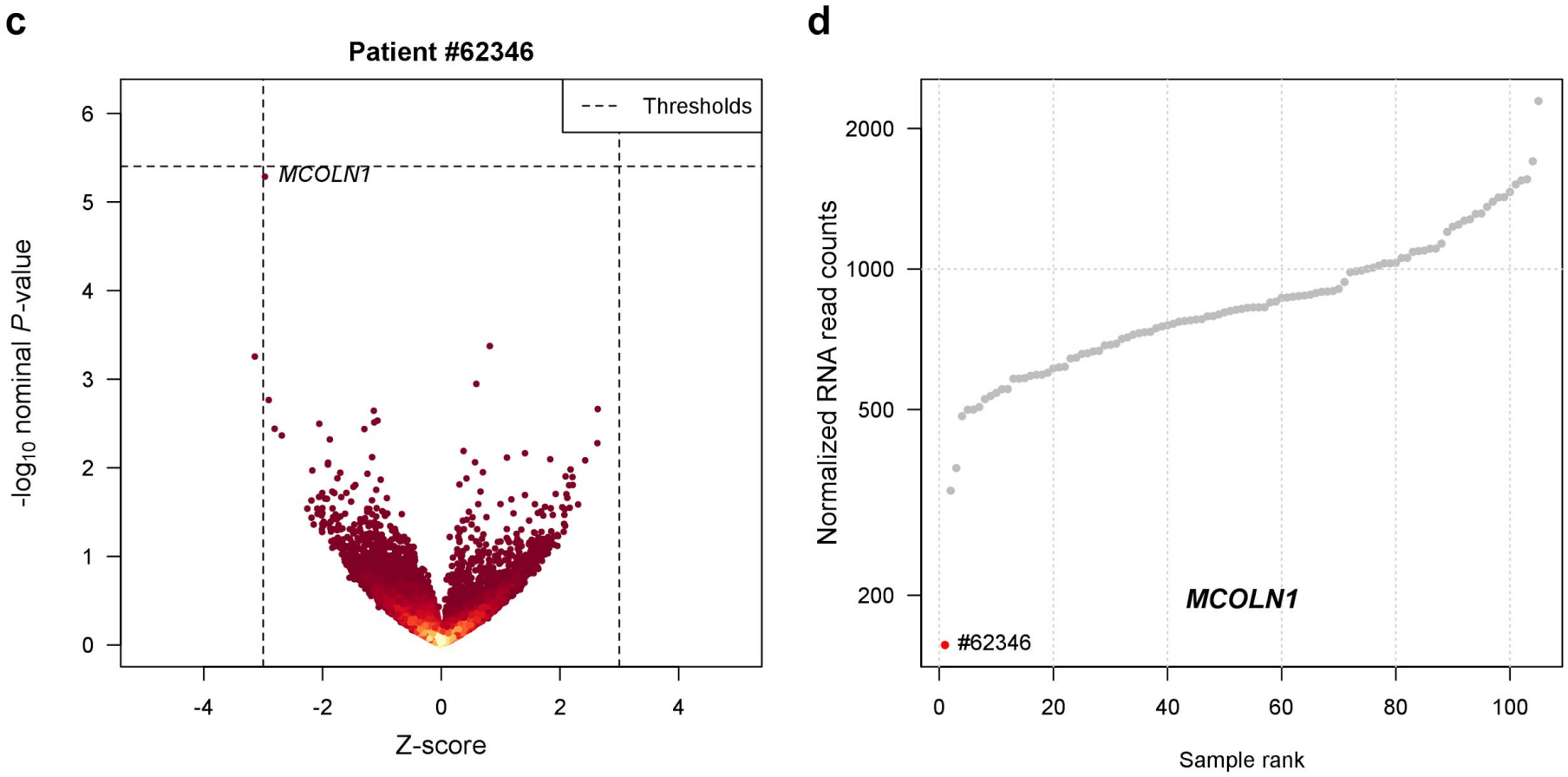
Proline synthesis pathway



Metabolites associated with ALDH18A1

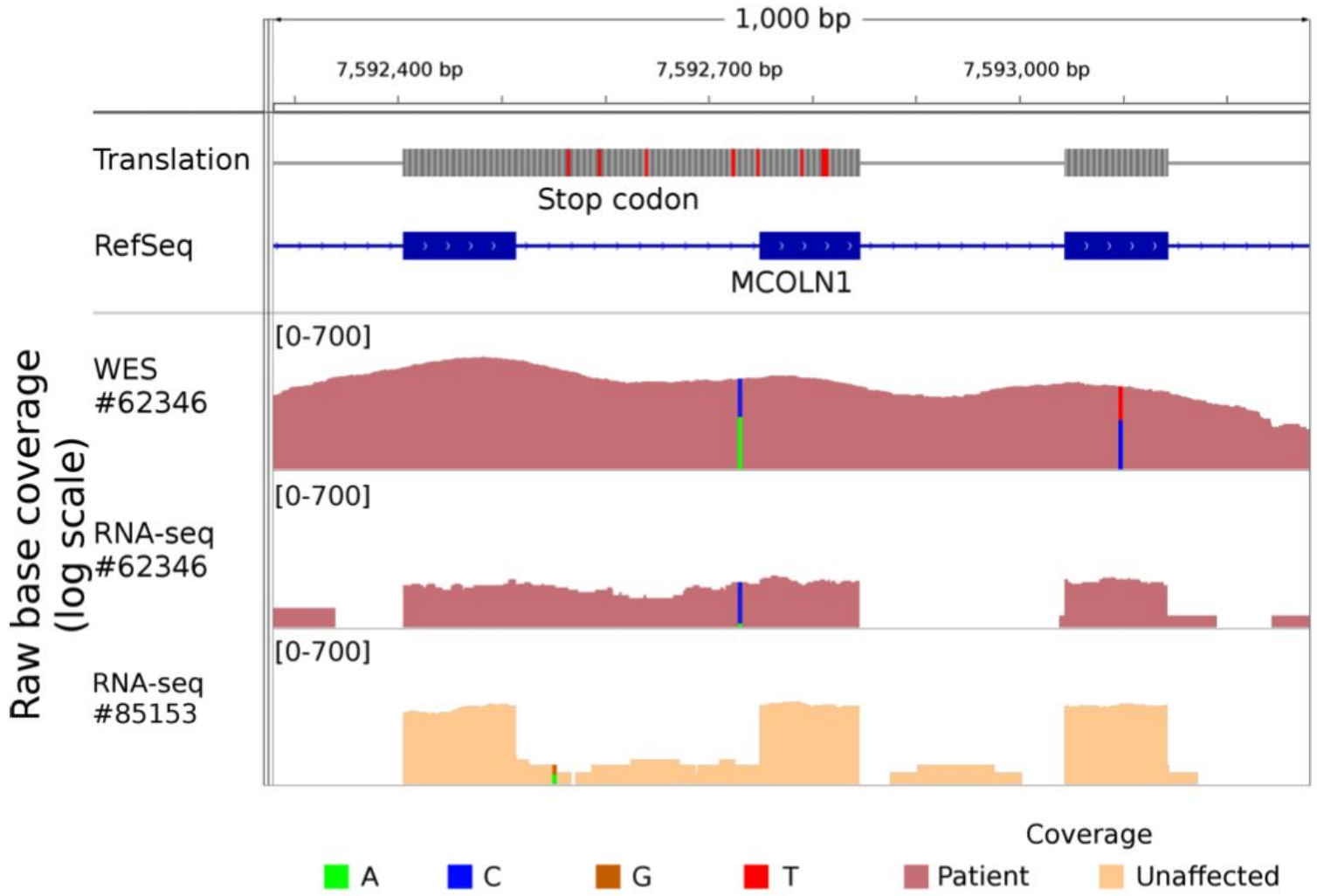


MCOLN1 quasi aberrant expression



MCOLN1: Diseases associated with Mucolipidosis IV and Cerebral Palsy, Ataxic.

MCOLN1 intron retention



MCOLN1: Diseases associated with Mucolipidosis IV and Cerebral Palsy, Ataxic.

Modeling mono-allelic expression

$$K_{i,j} \sim NB(s_j \times q_{i,j}, \alpha)$$

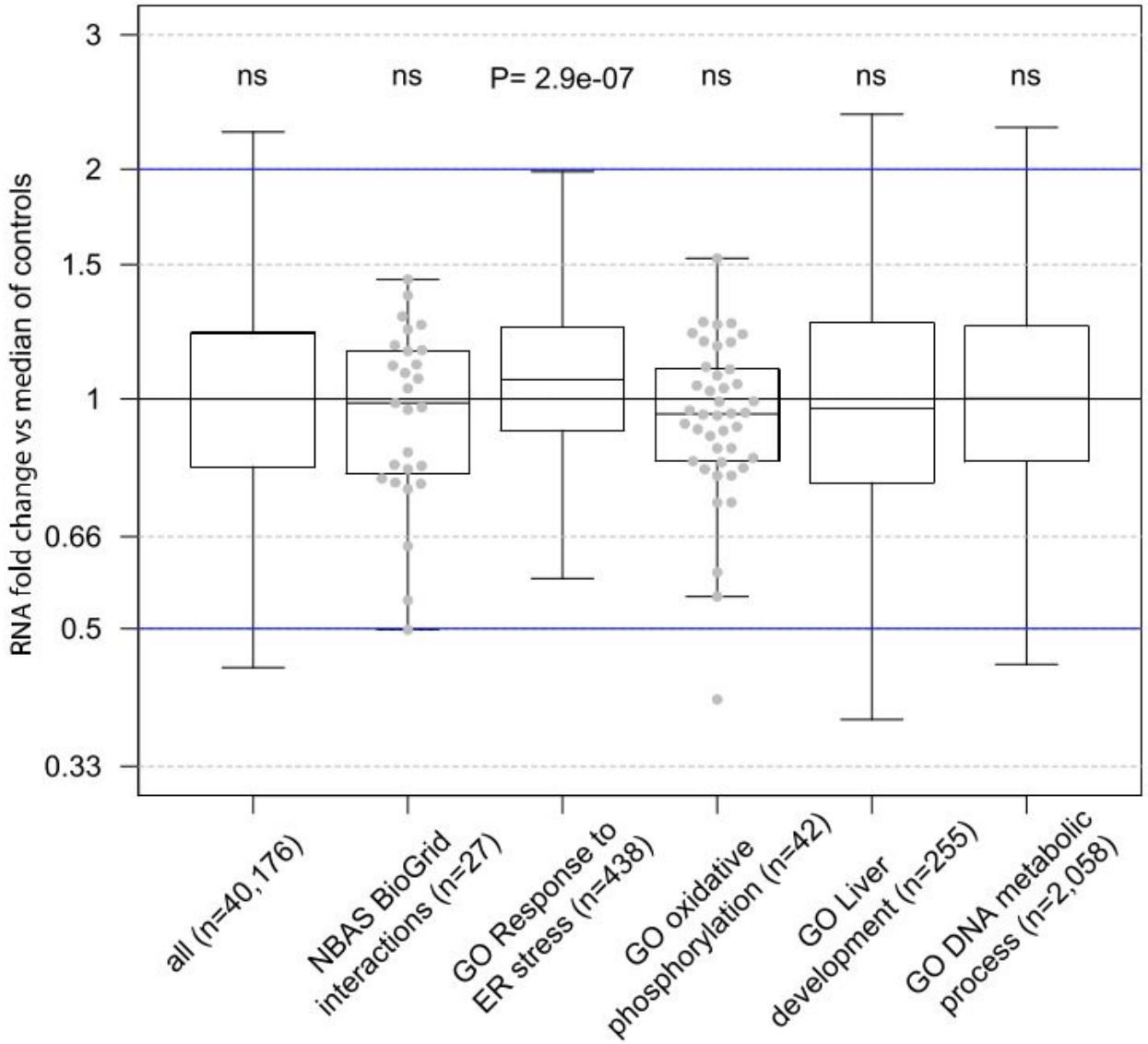
$$\log_2(q_{i,j}) = \beta_i^0 + \beta_i^{condition} \mathbf{x}_{i,j}^{condition}$$

- K_{i,j} the number of reads of variant i in sample j as
- s_j: sample specific size factor
- alpha: dispersion parameter fixed for all variants to 0.05 (~avg. dispersion of gene-wise analysis)
- x: 1 for alternative allele, 0 for reference

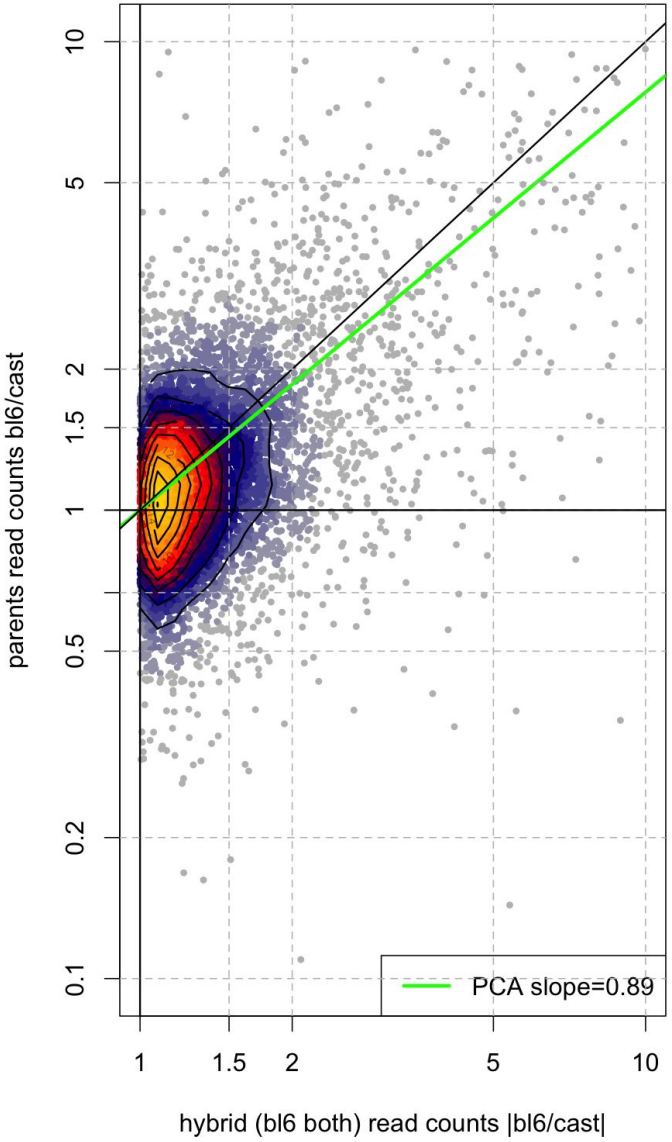
What else?

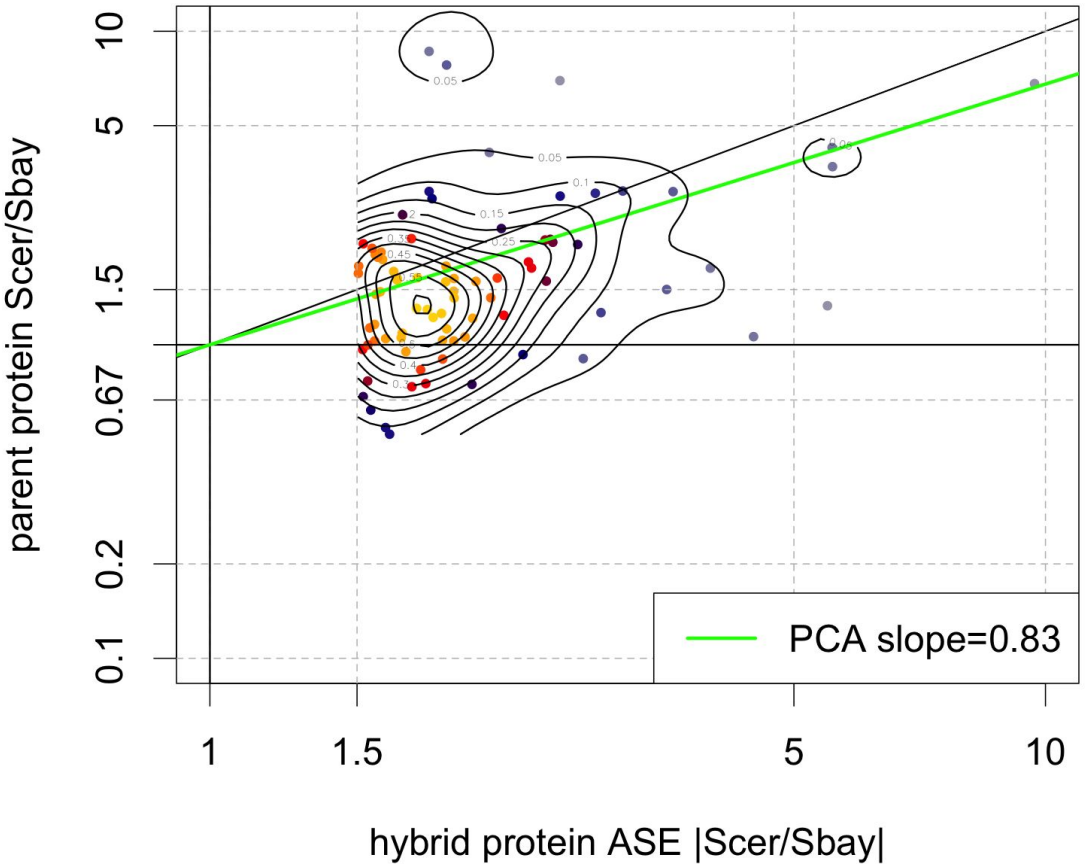
The unpublished stuff

Haack et al. 2015 NBAS



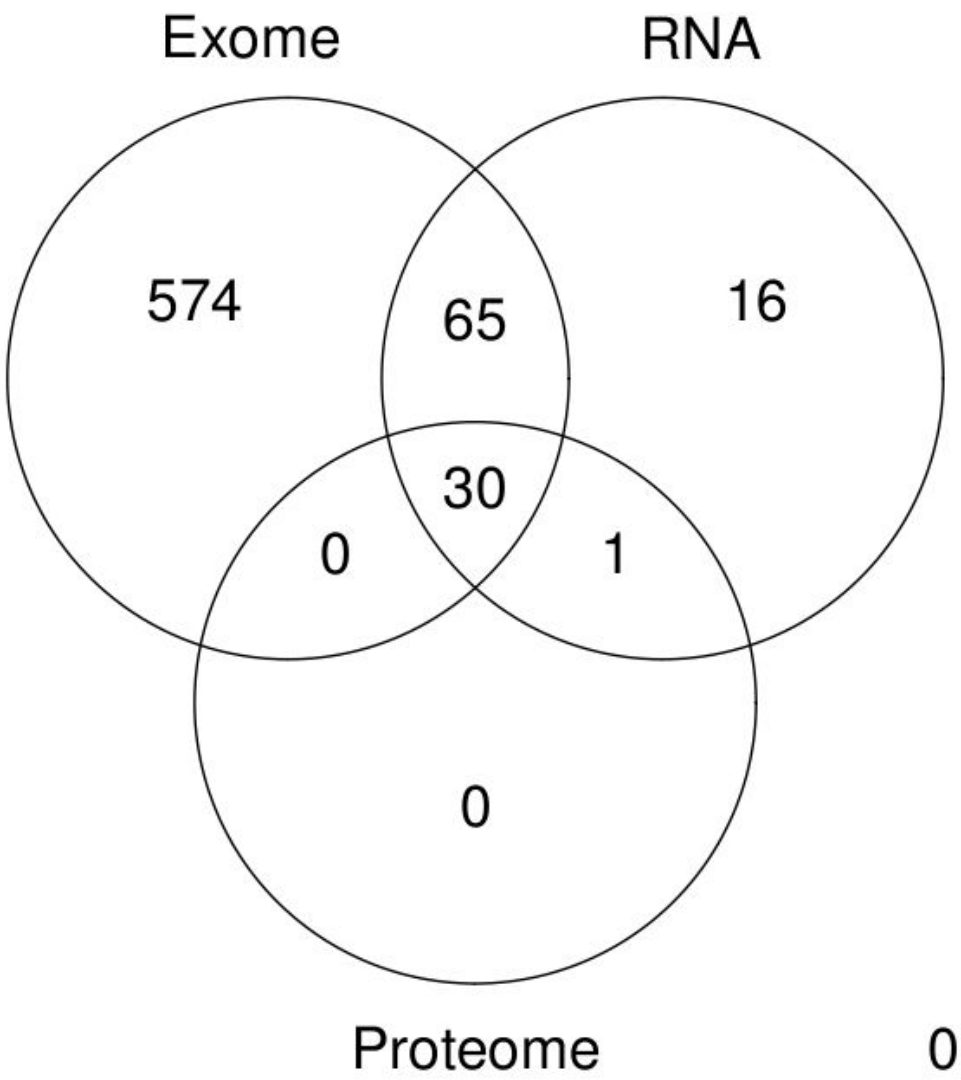
Re-analysis of other hybrid expression data





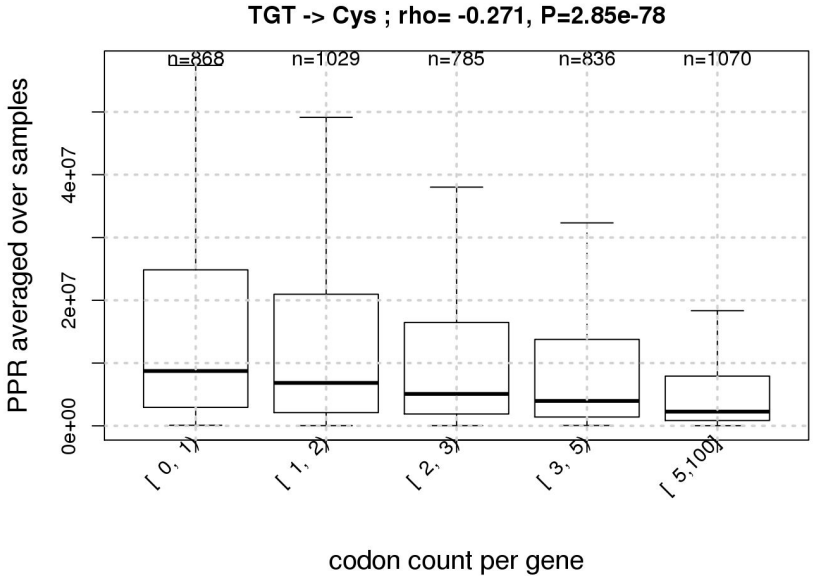
Goncalves 2012 GenRes,
Khan2012 MSB, ...

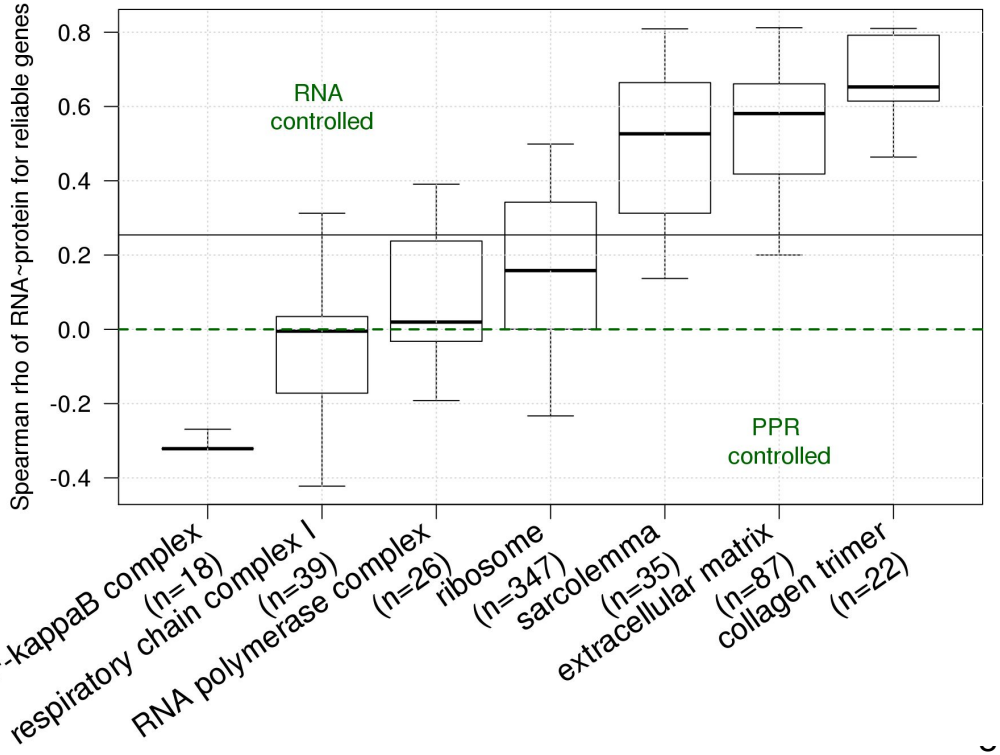
Building and maintaining a multi-omics sample annotation



- mitomap 6992
  - addons 6880
    - 00_create_mitomap_from_excel.html 6880
    - 00_create_mitomap_from_excel.R 6880
    - 20150507_merge_sample_info_rna_seqdate.html 6880
    - 20150507_merge_sample_info_rna_seqdate.R 6880
    - 20150601_add_mito_sample_pedigree.html 6880
    - 20150601_add_mito_sample_pedigree.R 6880
    - 20150602_read_laura_sample_info.html 6880
    - 20150602_read_laura_sample_info.R 6880
    - 20150617_add_pichler_proteome_samples.html 6880
    - 20150617_add_pichler_proteome_samples.R 6880
    - 20150625_merge_sample_info_pellet_date.html 6880
    - 20150625_merge_sample_info_pellet_date.R 6880
    - 20150727_add_seahorse_column.R 6880
    - 20150729_add_klein_sinfo_table.html 6880
    - 20150729_add_klein_sinfo_table.R 6880
    - 20150729_add_prokisch_enrichment_kits_mapping.html 6880
    - 20150729_add_prokisch_enrichment_kits_mapping.R 6880
    - 20150820_add_sex_prediction.html 6880
    - 20150820_add_sex_prediction.R 6880
    - 20150902_prokisch_sample_known_mutations.html 6880
    - 20150902_prokisch_sample_known_mutations.R 6880
    - 20151008_add_dpyd_samples.html 6880
    - 20151008_add_dpyd_samples.R 6880
    - 20151008_add_ids_2nd_batch_rna_proteome.html 6880
    - 20151008_add_ids_2nd_batch_rna_proteome.R 6880
    - 20160205_add_exomeid_to_expression.html 6880
    - 20160205_add_exomeid_to_expression.R 6880
    - 20160222_add_clinical_diagnosis.html 6880
    - 20160222_add_clinical_diagnosis.R 6880
  - 000_create_mitomap_from_ihg_database.html 5758
  - 000_create_mitomap_from_ihg_database.R 5758
  - add_meta_ibd_vcfs.R 4099
  - build_all_mitomap.R 6880
  - check_sanity_mito_id_map.html 6002
  - check_sanity_mito_id_map.R 6001
  - expert_knowledge_mitomap.R 6992
  - functions_mitomap.R 6882
  - get_batch_info_from_bam.R 4099
  - get_mitomap_summary.html 6926
  - get_mitomap_summary.R 6926
  - match_patientID.html 6112
  - match_patientID.R 6112
  - missing_exomes_by_expression.tsv 6532
  - mito_id_map.tsv 6350
  - parse_mito_id_map.R 6881
  - prokisch_mitomap.tsv 6926
  - README 4162

RNA-protein regulation studies





The end