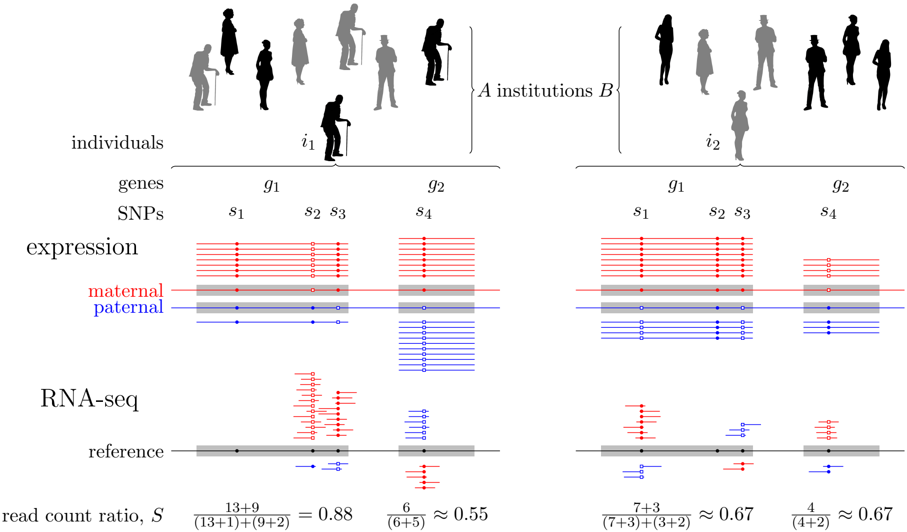
The “Imprinting Manuscript”

Normal Expression Bias of Imprinted Genes in Schizophrenics

Attila Gulyas-Kovacs

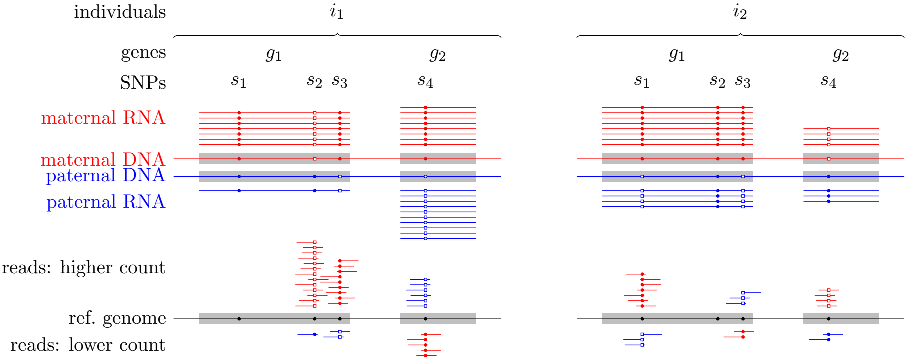
Chess lab meeting 12/12/17

The CommonMind data



- questions
  1. schizophrenia and imprinting (15q11-q13 microduplications)
  2. imprinted genes in adult human DLPFC
  3. determinants of imprinting (age, ancestry, gender)
- key studies
  1. Fromer et al 2016 Nat Neurosci
  2. Gregg et al 2010 Science
  3. Baran et al 2015 Genome Res
  4. Perez et al 2015 eLife

Read count ratio gauges *allelic bias* and thus *imprinting*



individuals i_1 i_2

genes g_1 g_2 g_1 g_2

SNPs s_1 s_2 s_3 s_4 s_1 s_2 s_3 s_4

maternal RNA

maternal DNA

paternal DNA

paternal RNA

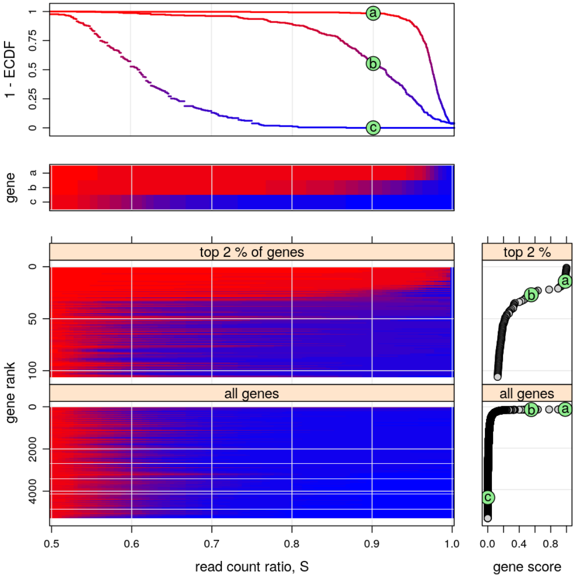
reads: higher count

ref. genome

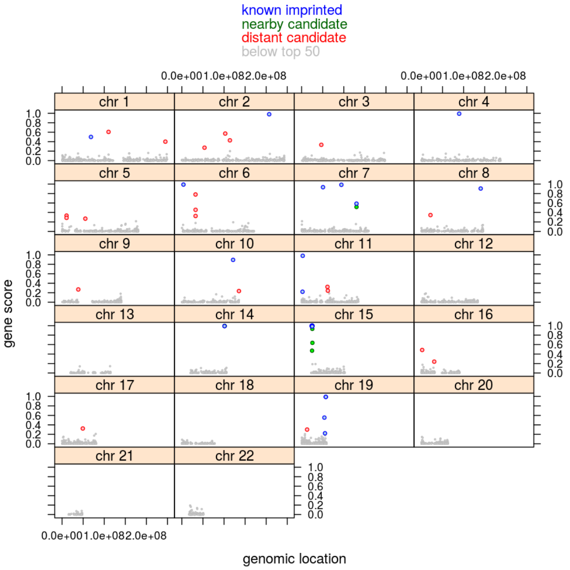
reads: lower count

read count ratio, S $\frac{13+9}{(13+1)+(9+2)} = 0.88$ $\frac{6}{(6+5)} \approx 0.55$ $\frac{7+3}{(7+3)+(3+2)} \approx 0.67$ $\frac{4}{(4+2)} \approx 0.67$

Ranking genes based on variation across individuals

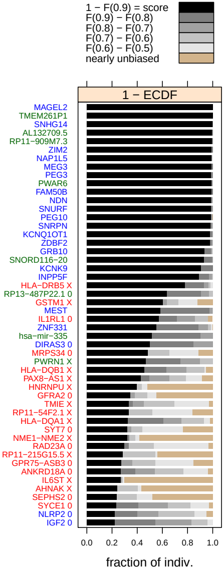


Gene score and previous imprinted gene clusters

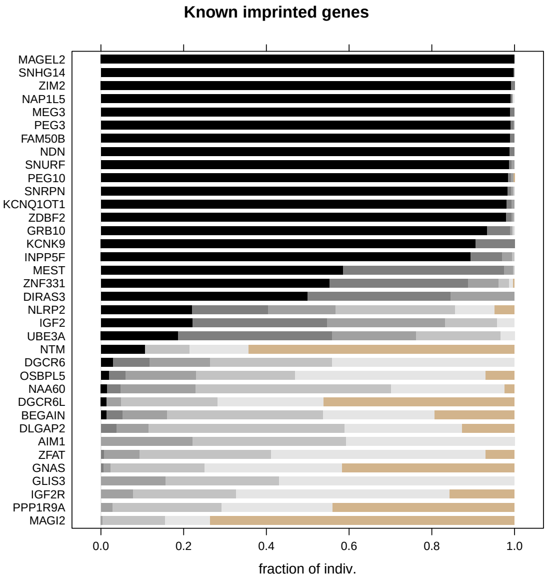


Establishing imprinting status in the human DLPFC

- prior expectation: near cluster
- alternative causes of high read count ratio
  1. mapping bias
  2. eQTL

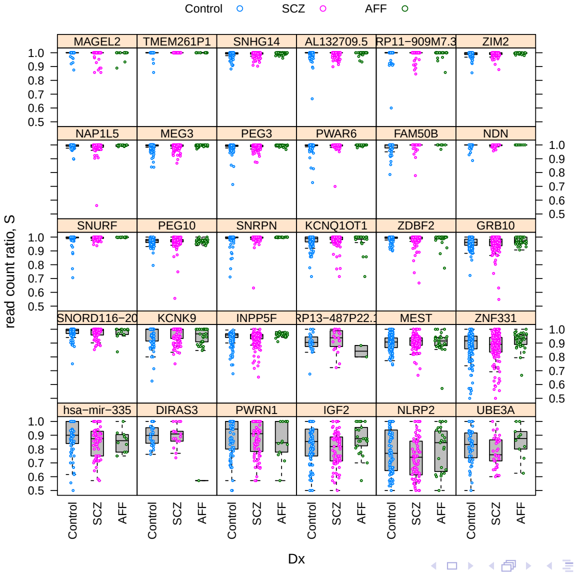


Including 3 slightly lower scoring genes



Explaining variation with psychiatric diagnosis, Dx

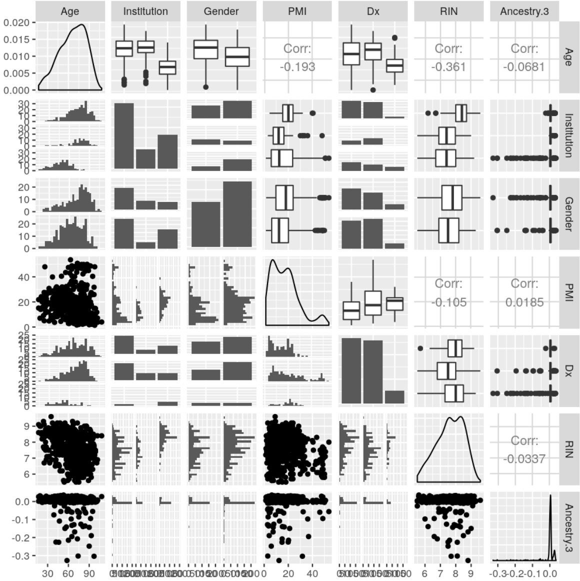
The simple but “confounded” approach



More information with more explanatory variables

explanatory variable	levels
Age	
Institution	[MSSM], Penn, Pitt
Gender	[Female], Male
PMI	
Dx	[Control], SCZ, AFF
RIN	
RNA_batch	[A], B, C, D, E, F, G, H, 0
Ancestry.1	
⋮	
Ancestry.5	

Dependencies: the source of confounding



Several regression models of read count ratio Y_g

Quantities

- observed variables
  - $Y_g = S_g$: response = read count ratio for gene g
  - $Y_g = Q_g$ (or $Y_g = R_g$): response = transformed read count ratio
  - X_j: the j-th column of design matrix X
- model parameters
  - β_{jg} (or b_{jg}): regression coefficient for Y_g and X_j
  - σ_g (or m_{ig}): parameters for noise

Properties

- given g, the structure of dependencies among $Y_g, X_1, ..., X_p$
- parametric family (normal or logistic)
  - link function
  - noise distribution

Several regression models of read count ratio Y_g

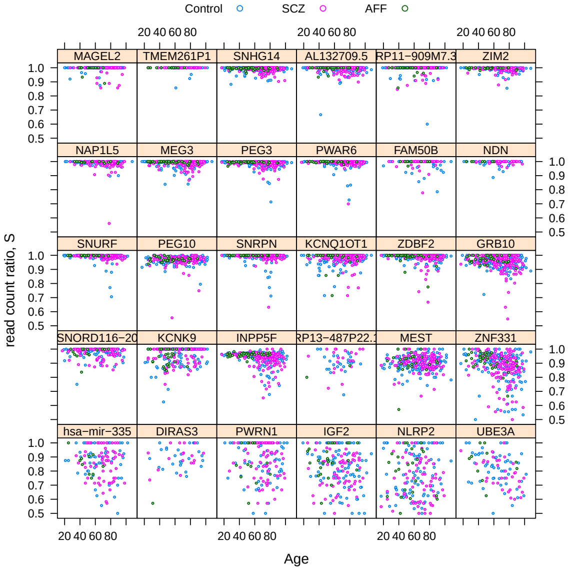
Quantities

- observed variables
  - $Y_g = S_g$: response = read count ratio for gene g
  - $Y_g = Q_g$ (or $Y_g = R_g$): response = transformed read count ratio
  - X_j: the j-th column of design matrix X
- model parameters
  - β_{jg} (or b_{jg}): regression coefficient for Y_g and X_j
  - σ_g (or m_{ig}): parameters for noise

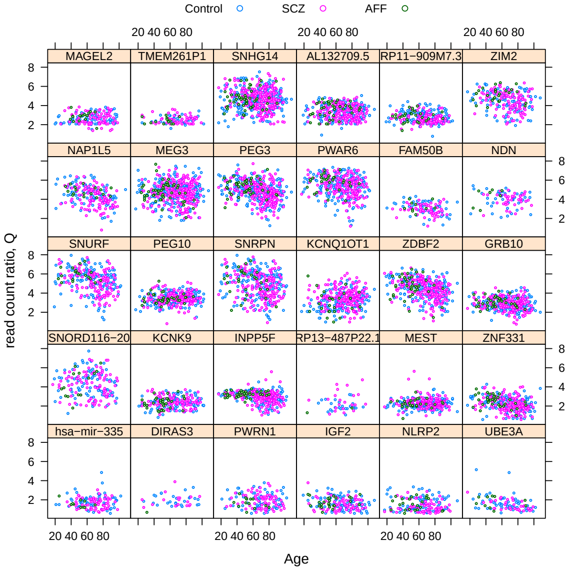
Properties

- given g, the structure of dependencies among $Y_g, X_1, ..., X_p$
- parametric family (normal or logistic)
  - link function
  - noise distribution

Untransformed read count ratio S_g



Transformed read count ratio Q_g



Several regression models of read count ratio Y_g

Quantities

- observed variables
  - $Y_g = S_g$: response = read count ratio for gene g
  - $Y_g = Q_g$ (or $Y_g = R_g$): response = transformed read count ratio
  - X_j: the j-th column of design matrix X
- model parameters
  - β_{jg} (or b_{jg}): regression coefficient for Y_g and X_j
  - σ_g (or m_{ig}): parameters for noise

Properties

- given g, the structure of dependencies among $Y_g, X_1, ..., X_p$
- parametric family (normal or logistic)
  - link function
  - noise distribution

Several regression models of read count ratio Y_g

Quantities

- observed variables
  - $Y_g = S_g$: response = read count ratio for gene g
  - $Y_g = Q_g$ (or $Y_g = R_g$): response = transformed read count ratio
  - X_j: the j-th column of design matrix X
- model parameters
  - β_{jg} (or b_{jg}): regression coefficient for Y_g and X_j
  - σ_g (or m_{ig}): parameters for noise

Properties

- given g, the structure of dependencies among $Y_g, X_1, ..., X_p$
- parametric family (normal or logistic)
  - link function
  - noise distribution

Several regression models of read count ratio Y_g

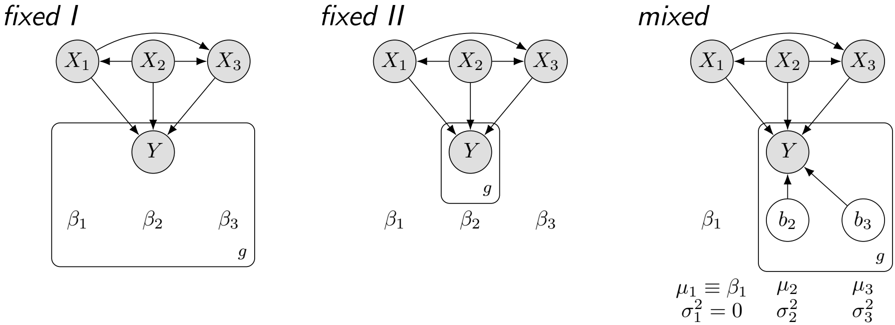
Quantities

- observed variables
  - $Y_g = S_g$: response = read count ratio for gene g
  - $Y_g = Q_g$ (or $Y_g = R_g$): response = transformed read count ratio
  - X_j: the j-th column of design matrix X
- model parameters
  - β_{jg} (or b_{jg}): regression coefficient for Y_g and X_j
  - σ_g (or m_{ig}): parameters for noise

Properties

- given g, the structure of dependencies among $Y_g, X_1, ..., X_p$
- parametric family (normal or logistic)
  - link function
  - noise distribution

Three classes of dependency structure



- fixed I: too complex ⇒ low power
- fixed II: too simplistic ⇒ bias
- mixed: powerful middle ground—even with interactions

Several regression models of read count ratio Y_g

Quantities

- observed variables
  - $Y_g = S_g$: response = read count ratio for gene g
  - $Y_g = Q_g$ (or $Y_g = R_g$): response = transformed read count ratio
  - X_j: the j-th column of design matrix X
- model parameters
  - β_{jg} (or b_{jg}): regression coefficient for Y_g and X_j
  - σ_g (or m_{ig}): parameters for noise

Properties

- given g, the structure of dependencies among $Y_g, X_1, ..., X_p$
- parametric family (normal or logistic)
  - link function
  - noise distribution

Several regression models of read count ratio Y_g

Quantities

- observed variables
  - $Y_g = S_g$: response = read count ratio for gene g
  - $Y_g = Q_g$ (or $Y_g = R_g$): response = transformed read count ratio
  - X_j: the j-th column of design matrix X
- model parameters
  - β_{jg} (or b_{jg}): regression coefficient for Y_g and X_j
  - σ_g (or m_{ig}): parameters for noise

Properties

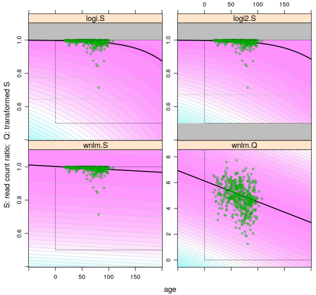
- given g, the structure of dependencies among $Y_g, X_1, ..., X_p$
- parametric family (normal or logistic)
  - link function
  - noise distribution

Parametric families

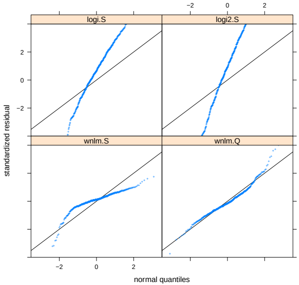
model family	abbrev.	response var.
*u*nweighted *n*ormal *l*inear	unlm	S, Q, or R
*w*eighted *n*ormal *l*inear	wnlm	S, Q, or R
*logi*stic	logi	S
*logi*stic, $\frac{1}{2}\times$ down-scaled link fun.	logi2	S

Fit of fixed I models for PEG3

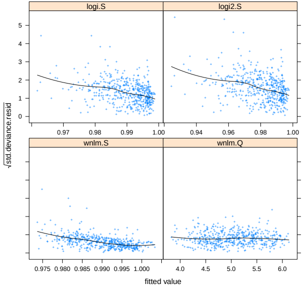
fixed simple regression model, PEG3



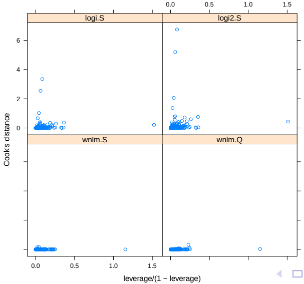
Normality of residuals.
Fixed multiple regression model, PEG3



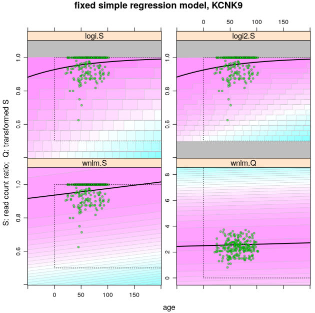
Homogeneity of error variance.
Fixed multiple regression model, PEG3



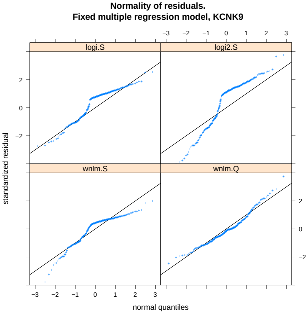
Influence of outliers.
Fixed multiple regression model, PEG3



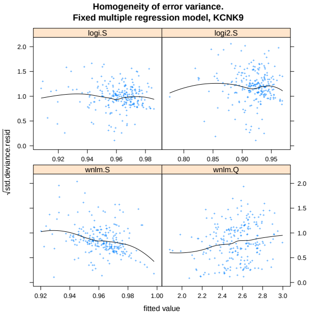
Fit of fixed I models for KCNK9



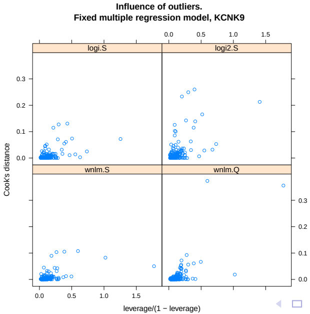
fixed simple regression model, KCNK9



Normality of residuals.
Fixed multiple regression model, KCNK9



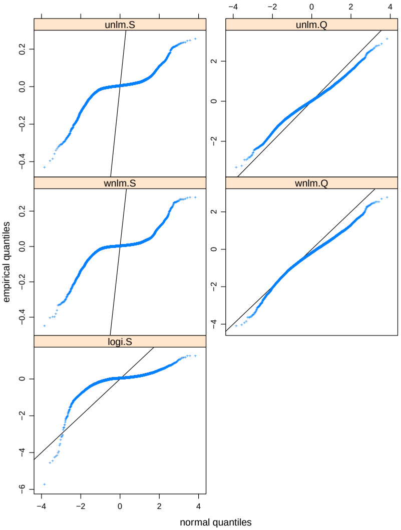
Homogeneity of error variance.
Fixed multiple regression model, KCNK9



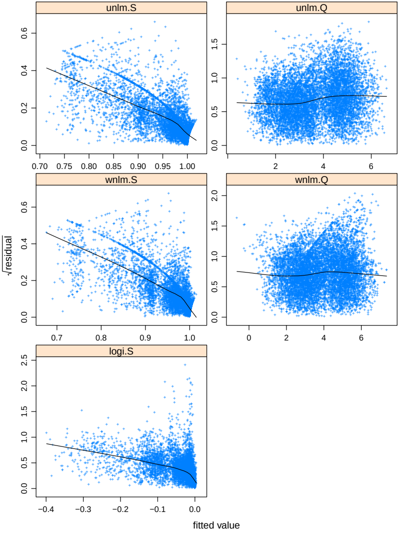
Influence of outliers.
Fixed multiple regression model, KCNK9

Fit of mixed models (all genes jointly)

Normality of residuals.
Mixed multiple regression model, all genes

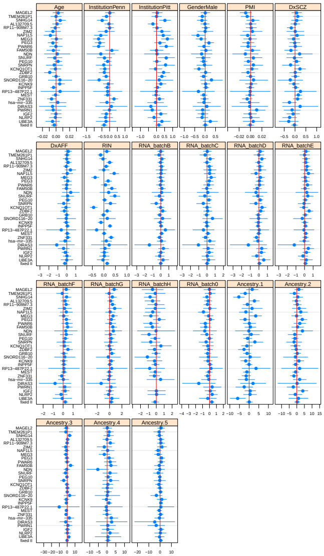


Homogeneity of error variance.
Mixed multiple regression model, all genes

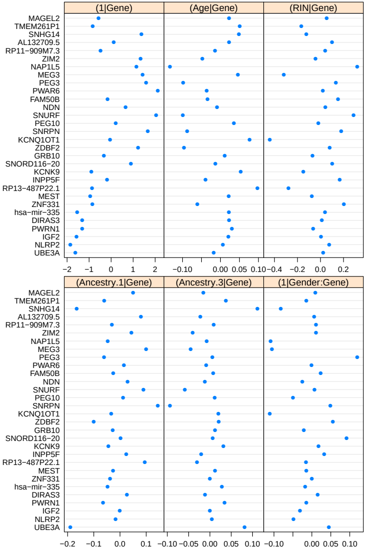


Regression coefficients

Estimate and 99 % CI for β_{jg}. Fixed effects, unlm.Q



Predicted random coefficient b_{jg}. Mixed model unlm.Q

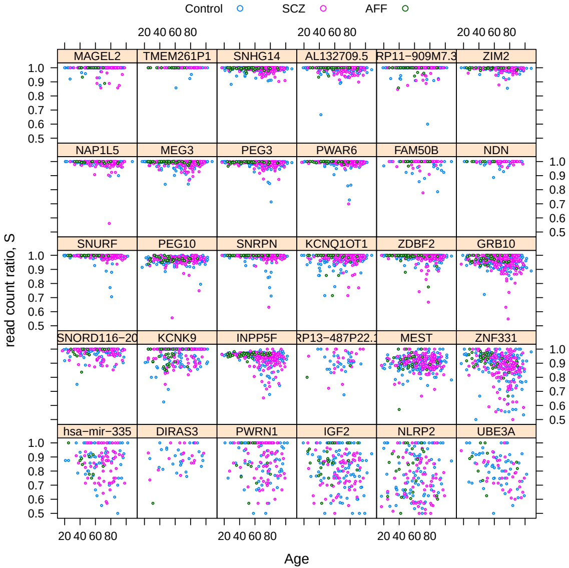


Testing independence of read count ratio

Based on unlmQ mixed model

predictor term	interpretation	ΔAIC	p-value
$(1 \mid \mathrm{Gene})$	variability among genes	-126.8	8.5×10^{-28}
$(1 \mid \mathrm{Dx})$	variability among Control, SCZ, AFF	2.0	1.0
$(1 \mid \mathrm{Dx} : \mathrm{Gene})$	Gene specific variability among Ctrl, SCZ, AFF	0.4	0.21
Age	effect of Age	1.3	0.39
$(\mathrm{Age} \mid \mathrm{Gene})$	Gene specific effect of Age	-18.9	2.5×10^{-5}
$\mathrm{Ancestry.1}$	effect of Ancestry.1	0.6	0.24
$(\mathrm{Ancestry.1} \mid \mathrm{Gene})$	Gene specific effect of Ancestry.1	-71.2	4.6×10^{-16}
$\mathrm{Ancestry.3}$	effect of Ancestry.3	1.6	0.54
$(\mathrm{Ancestry.3} \mid \mathrm{Gene})$	Gene specific effect of Ancestry.3	-17.9	3.8×10^{-5}
$(1 \mid \mathrm{Gender})$	difference between Male and Female	2.0	1.0
$(1 \mid \mathrm{Gender} : \mathrm{Gene})$	Gene specific difference between M and F	-5.7	5.5×10^{-3}

Untransformed read count ratio S_g



Summary

1. CommonMind RNA-seq read count ratio gauging allelic bias
2. ≈ 30 imprinted genes in human DLPFC
   - in agreement with more recent estimates
3. normal allelic bias of imprinted genes in schizophrenics
   - subtle effect + noise and bias?
   - complex genetic architecture
4. gene-specific effect of ancestry, gender, and age
   - aging: "imprinting and the social brain"
   - "DNA methylation age"

Summary

1. CommonMind RNA-seq read count ratio gauging allelic bias
2. ≈ 30 imprinted genes in human DLPFC
   - in agreement with more recent estimates
3. normal allelic bias of imprinted genes in schizophrenics
   - subtle effect + noise and bias?
   - complex genetic architecture
4. gene-specific effect of ancestry, gender, and age
   - aging: “imprinting and the social brain”
   - “DNA methylation age”

Summary

1. CommonMind RNA-seq read count ratio gauging allelic bias
2. ≈ 30 imprinted genes in human DLPFC
   - in agreement with more recent estimates
3. normal allelic bias of imprinted genes in schizophrenics
   - subtle effect + noise and bias?
   - complex genetic architecture
4. gene-specific effect of ancestry, gender, and age
   - aging: "imprinting and the social brain"
   - "DNA methylation age"

Summary

1. CommonMind RNA-seq read count ratio gauging allelic bias
2. $\approx$ 30 imprinted genes in human DLPFC
    - in agreement with more recent estimates
3. normal allelic bias of imprinted genes in schizophrenics
    - subtle effect + noise and bias?
    - complex genetic architecture
4. gene-specific effect of ancestry, gender, and age
    - aging: “imprinting and the social brain”
    - “DNA methylation age”