--- layout: page title: Monoall-ms --- [Download](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/monoall-ms.docx) this page in docx format (without figures).

## Aging and genomic imprinting in the human brain

### Preliminaries

#### Notation

**Color key**

* normal color: text related to main figure
* first color: ...supplementary figure
* second color: ...figure still missing

**refname:** this should be used as unique identifier for a figure until the final figure numbers may be established

#### Definitions

* monoallelic expression: only one allele is expressed to a certain degree
* (genomic) imprinting: a subtype of monoallelic expression where the same allele is expressed in some tissue across all individuals
* known imprinted and candidate genes: with and without prior evidence for imprinting, respectively
* candidate, \lt 1 MB gene: a candidate gene in the 1 MB neighborhood of some known imprinted gene
* imprinted gene cluster: a sequence of neighboring genes that fall in either the "known" or the "candidate, \lt 1 MB gene" category
* biallelic expression: both alleles are expressed to a certain degree
* unbiased, or balanced, expression: the two alleles are expressed at equally high level
* parental expression bias or simply parental bias: one allele is expressed at a higher level
* mean transcript ratio, \mu\_{ig} of a given individual i and gene g:
  1. the mean ratio of the number of maternal transcripts in all transcripts over a suitable period of time and set of cells
  2. hence a natural but not directly unobservable measure of expression bias

### Parental expression bias: variation across genes and individuals

ranking-genes

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-07-19-genome-wide-S/complex-plot-1.pdf)

**Motivation:** Genome-wide and population-wide quantification parental bias

1. variability across genes and the fraction of monoallelically expressed genes
2. variability across individuals

We defineread count ratio statistic, S\_{ig} of a given individual i and gene g as the greater of the observed read count of the two alleles at each heterozygous SNP summed over all such SNPs and normalized to the corresponding total read count. We use S\_{ig} to quantify parental bias. This is justified by the reasonable assumptions that

1. the number of paternal and maternal transcripts are Poisson distributed with rate \lambda and \lambda \mu, respectively
2. the observed paternal and maternal read count are both Poisson distributed with rate proportional to the number of paternal and maternal transcripts, respectively.

We also define the gene g's score as the fraction of individuals i for which S\_{ig}\gt 0.9. This is the same as 1 - ECDF(0.9), where ECDF is the empirical cumulative distribution function based on observations S\_{1g},...,S\_{Ig} on I individuals or, equivalently, the same as the corresponding survival function.

**Result:** Three typical genes are shown in the **upper 3 panels**: PEG10 and ZNF331 are known imprinted genes, whereas AFAP1 is a candidate gene. The each gene has a characteristic survival function 1 - ECDF(s) based on which receive different scores **green circles a, b, c**. The survival functions are color-coded for terse representation. All ca. 5K genes that passed our quality filters are shown in this representation**2 main lower panels**; they are ranked according to their score, shown in the auxiliary **right lower panels**.

**Interpretation:** Given some reasonable assumptions (see definition of read count ratio statistic S) only about ca. 1 % of the genes are expressed monoallelically

**See also:** [2016-07-19-genome-wide-S](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-07-19-genome-wide-S%20%%7D)

### Parental bias and imprinted gene clusters

clusters

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-08-08-imprinted-gene-clusters/score-genomic-location-1.pdf)

**Motivation:** Given our definition of clusters (above) identify them along with the "known" and "candidate, \lt 1MB" genes. These gene sets will be combined with the observed ranking for more detailed analysis.

**Result:** Altogether 60 "known" and 701 "candidate, \lt 1MB" genes before quality filtering. These form 36 clusters, each with at least a few genes and a median of 15 genes. After filtering 36 "known" and 266 "candidate, \lt 1MB" genes.

**Interpretation:** None.

**See also:** [2016-08-08-imprinted-gene-clusters](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-08-08-imprinted-gene-clusters%20%%7D), [2016-07-14-imprinting-resources](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-07-14-imprinting-resources%20%%7D)

### Top ranking genes

top-genes

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-08-01-ifats-filters/top-ranking-genes-1.pdf)

**Motivation:** Based on gene score and clusters

1. detect novel imprinted genes
2. define gene set for detailed subsequent analysis

**Result:** The figure presents the top ranking genes colored according to their imprinting category ("known",...).

**Interpretation:** Evidence for some novel imprinted genes, e.g. TMEM261P1. Several "known" imprinted genes are not among these top ranking genes suggesting these may not be imprinted in most human brains. Limitation: we did not estimate error rates.

**See also:** [2016-08-01-ifats-filters](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-08-01-ifats-filters%20%%7D), [Ifat's version](https://docs.google.com/presentation/d/1YvpA1AJ-zzir1Iw0F25tO9x8gkSAzqaO4fjB7K3zBhE/edit#slide=id.p4)

### Variation of parental bias with age

S-age-institution

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-06-26-trellis-display-of-data/S-age-smooth-1.pdf)

**Motivation:** The observed variability across individuals calls for explanation. Several variables were recorded including age, gender and disease status. Imprinting: clearly important for perinatal developement but how about later stages and aging?

**Result:** Age seems to affect both the mean and variance of S for some genes either by decreasing (ZNF331) or increasing (DIRAS3), but not others (MEST). Smoothing with local regression (LOESS) supports this.

**Interpretation:** The result appears to suggest dependence of parental bias on age. However, age seems to be associated with the brain-collecting(?) institution. Possible association with other variables may further complicate interpretation.

**See also:** [2016-06-26-trellis-display-of-data](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-06-26-trellis-display-of-data%20%%7D)

### Pairwise associations among predictors

predictor-associations

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-06-26-trellis-display-of-data/evar-scatterplot-matrix-2.pdf)

**Motivation:** Uncover association among putative predictors of parental bias. In theory higher order associations are of interest but here only pairwise ones are studied, in particular those involving age of death.

**Result:** Association of age is strong to institution, disease, RNA quality (RIN) and ancestry and is also substantial to gender. Institution is also strongly associated to most variables.

**Interpretation:** The latter finding suggests institution-specific technical procedures and sampling bias.

**See also:** [2016-06-26-trellis-display-of-data](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-06-26-trellis-display-of-data%20%%7D)

### Weight of evidence

weight-of-evidence

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-06-26-trellis-display-of-data/S-age-tot-read-count-1.pdf)

**Motivation:** The total read count is informative for statistical inference on the effect of age and other predictors since the latter may be considered as a weigh of evidence on the read count ratio for each data point that corresponds to some individual--gene pair.

**Result:** The total read count varies both across genes and, for a given gene, across individuals. Summing the higher and total read count over genes yields a weighted average (WA) of the read count ratio.

**Interpretation:** The across-genes variation of read count may chiefly reflect across-genes variation in expression, whereas the across-individuals variation of read count likely arises from variation both in expression and in technological effects. Modeling these levels of variation is challenging but desirable since that affords higher sensitivity and specificity.

**See also:** [2016-06-26-trellis-display-of-data](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-06-26-trellis-display-of-data%20%%7D)

### Regression models

predicted-curves

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-04-22-glm-for-s-statistic/s-stat-cmp-simple-multiple-regr-peg3-1.pdf)

**Motivation:** Fitting a suitable regression model is the standard technique to quantitatively infer putative predictors' effect on a response variable, in this case the read count ratio. We need a regression model that best accounts for our observation of associations among predictors and of varying weight of evidence. We tested several models that differ in the following characteristics:

1. data transformation (none or rank tr.)
2. weighting according to total read count (weighted or unweighted)
3. link function and error distribution (logistic and normal linear)
4. relationship among predictors (simple or multiple regression)
5. relationship among genes (independent or identically distributed).

**Result:** In summary fitting various models revealed the similar qualitative trends. The figure shows that normal linear models are limited to a narrow age interval otherwise they predict read count ratios greater than one while logistic models suffer in another weakness: the need of massive extrapolation. Multiple regression yields typically shallower dependence on age.

**Interpretation:** The steeper dependence for simple regression alludes to a bias from the association of age with other predictors. Multiple regression accounts for that. The logistic models have other theoretical advantages over the normal linear models besides keeping predicted ratios below 1

* they model the higher read count as binomial variables, which is not only more reasonable than normality but provides a natural way of using total read counts as weights of evidence
* they model the strongly unequal error variance observed in the data.

**See also:** [2016-04-22-glm-for-s-statistic](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-04-22-glm-for-s-statistic%20%%7D)

### Normal linear model fits well only to transformed data

non-constant-variance

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-05-07-comparing-regression-models/s-r-stat-nlm-check-peg3-1.pdf)

**Motivation:** Normal linear models become strongly biased when fitted to data with non-constant error variance.

**Result:** The error variance is strongly varying: both systematically (with age, for instance) as seen in the **top left** panel and non-systematically (several outlier data points) as shown by the **top right** panel. The **bottom** panels were obtained with rank transformation and show much improvement due to reduced systematic and non-systematic variation.

**Interpretation:** The result is in line with the theoretical disadvantages of the normal linear model and suggests that inference from this model should not be taken into account because the inferred parameters are strongly biased. Rank transformation is likely to remove most of the bias at the expense of reduced sensitivity and interpretability since it places read count ratio on relative scale.

**See also:** [2016-05-07-comparing-regression-models](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-05-07-comparing-regression-models%20%%7D)

### Age effect and gene clusters

age-effect

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-08-08-imprinted-gene-clusters/segplot-1.pdf)

**Motivation:** Given the results and theoretical considerations the logistic multiple regression model is used to infer \beta\_\mathrm{age}, the regression coefficient that mediates the effect of age on read count ratio.

**Result:** When genes are taken independent, a separate \beta\_\mathrm{age} is inferred for each of them. This shows great variability across genes: at 99 % confidence some have significantly negative, others significantly positive coefficient and yet others a coefficient not significantly different from zero. The variability is also seen within imprinted gene clusters. Compared to the logistic model (shown) the normal linear model with rank transformation appears much less powerful (significance only for 1 gene: ZDBF2). Decreasing confidence to 95 % improves power (significance for 5 genes).

**Interpretation:** Some genes loose and others gain parental expression bias with age, while a third set of genes are not affected by age. An alternative, not necessary exclusive, interpretation is that the extrapolation under the logistic model leads to biased coefficients. The qualitative agreement between the two models supports the biological interpretation. If the age effect is indeed true then age operates on genes within the same cluster independently rather than in concert.

**See also:** [2016-08-08-imprinted-gene-clusters](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-08-08-imprinted-gene-clusters%20%%7D)

### All regression coefficients

all-effects

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-06-22-extending-anova/reg-coef-logi.S-1.pdf)

**Motivation:** Apart from age, we have data on other biological explanatory variables, specifically disease status, gender and ancestry. Do these also modulate parental bias like age seems to do?

**Result:** Gender and some principal components of the ancestry have significant effects for several genes. In contrast, disease status (Dx) does not, except for a single gene, NDN.

**Interpretation:** TODO

**See also:** [2016-06-22-extending-anova](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-06-22-extending-anova%20%%7D)

### All regression coefficients (wnlm.R model)

all-effects-wnlm.R

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-06-22-extending-anova/reg-coef-wnlm.R-1.pdf)

**Motivation:** Does the qualitative agreement between the logistic model logi.S (fitted on read count ratios without transformation) and normal linear model wnlm.R (fitted after rank-transformation) observed for age's effect also hold for other predictors.

**Result:** Similar qualitative patterns are found under wnlm.R as under logi.S. In general, the confidence intervals are relatively broader.

**Interpretation:** These results provide further support for the significant effects of some of the biological variables explaining parental bias. wnlm.R is less powerful (sensitive) than logi.S.

**See also:** [2016-06-22-extending-anova](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-06-22-extending-anova%20%%7D)

### Comparison of age effect under different regression models

age-effect-two-models

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-06-17-extending-regression-analysis/beta-age-wnlm.R-vs-logi.S-1.pdf)

**Motivation:** A more direct comparison of the logistic (logi.S) and normal linear models (wnlm.R) may help more firmly establish the preceding results.

**Result:** Clearly, the two models agree qualitatively on age's effect.

**Interpretation:** The two models quite different in several statistical properties. Thus, relatively few shared sources of bias are possible making the biological interpretation of the significant effects more plausible. In what follows we turn to an obvious shared property of the two models: (generalized) linearity.

**See also:** [2016-06-17-extending-regression-analysis](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-06-17-extending-regression-analysis%20%%7D)

### Associations obscure the size of the age effect

anova

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-06-22-extending-anova/anova-effects-fw-rv-logi.S-1.pdf)

**Motivation:** Before more directly inspecting linearity we study the possibility of near collinearity of sets of predictors. Orthogonality (lack of collinearity) is desirable to make statements on the relative impact of predictors and therefore the biological-"signal"-to-technical-"noise" ratio. In particular we ask: how much change in parental bias can aging induce? This can be studied with ANOVA as well as with the decomposition of the response response vector along regression coefficients (observed mediators of effects) and unobserved mediators of effects.

**Result:** When the age component is taken first into account the corresponding deviance and effect is roughly as high as the highest technical variables (Institution and RNA\_batch). But in the reverse direction the age component greatly shrinks.

**Interpretation:** The shrinkage is due to age's association to institution and other predictors, as we found above. This does not weaken the previously observed significance of age's effect but it does preclude determination of age's effect size.

**See also:** [2016-06-22-extending-anova](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-06-22-extending-anova%20%%7D)

### Interaction of age with other predictors

interaction

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-07-08-conditional-inference/beta-age-cond-logi.S-1.pdf)

**Motivation:** Linearity is a shared assumption by all models used here, whose violation might bias inference on the effect of age and other predictors. Linearity is equivalent to the conditional independence of predictors given the response (i.e. read count ratio).

**Result:** Conditioning the inference of \beta\_\mathrm{age} on institution and gender seems to more strongly affect inferences under the logistic model logi.S (shown) than those under the normal linear model wnlm.R. The pattern of change in the inferred \beta\_\mathrm{age} is more often similar than different.

**Interpretation:** The result suggests substantial nonlinearity; in other words age interacts with institution and gender. This interaction is not accounted for by our models, which may potentially lead to quit strong bias. The normal linear model (fitted to rank-transformed read counts) is more robust to departures from linearity. This points to the importance of comparing our observational study with factorially designed experiments.

**See also:** [2016-07-08-conditional-inference](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-07-08-conditional-inference%20%%7D)

### Comparison to previous mouse study: genome-wide parental bias

TODO

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-08-08-imprinted-gene-clusters/beta-age-logi.S-braim.pdf)

**Motivation:** TODO

**Result:** TODO

**Interpretation:** TODO

**See also:** [2016-08-08-imprinted-gene-clusters](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-08-08-imprinted-gene-clusters%20%%7D)

### Comparison to previous mouse study: age effect

TODO

[pdf](%7B%7B%20site.baseurl%20%7D%7D/monoall-ms/2016-08-08-imprinted-gene-clusters/beta-age-logi.S-braim.pdf)

**Motivation:** TODO

**Result:** TODO

**Interpretation:** TODO

**See also:** [2016-08-08-imprinted-gene-clusters](%7B%7B%20site.baseurl%20%7D%7D%7B%%20post_url%202016-08-08-imprinted-gene-clusters%20%%7D)