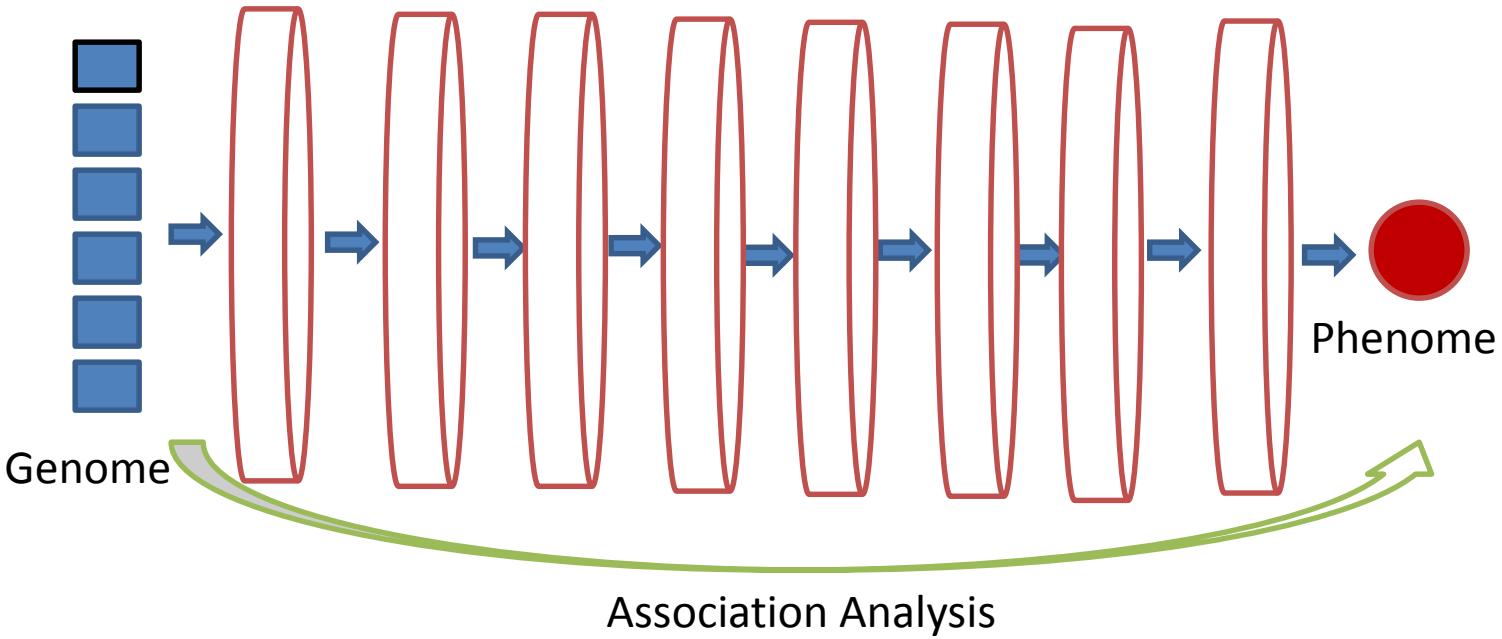
Statistical Considerations in Studying Epigenetic Changes

Integrative Analysis of Metabolomics and Epigenetics using the ELEMENT Cohort

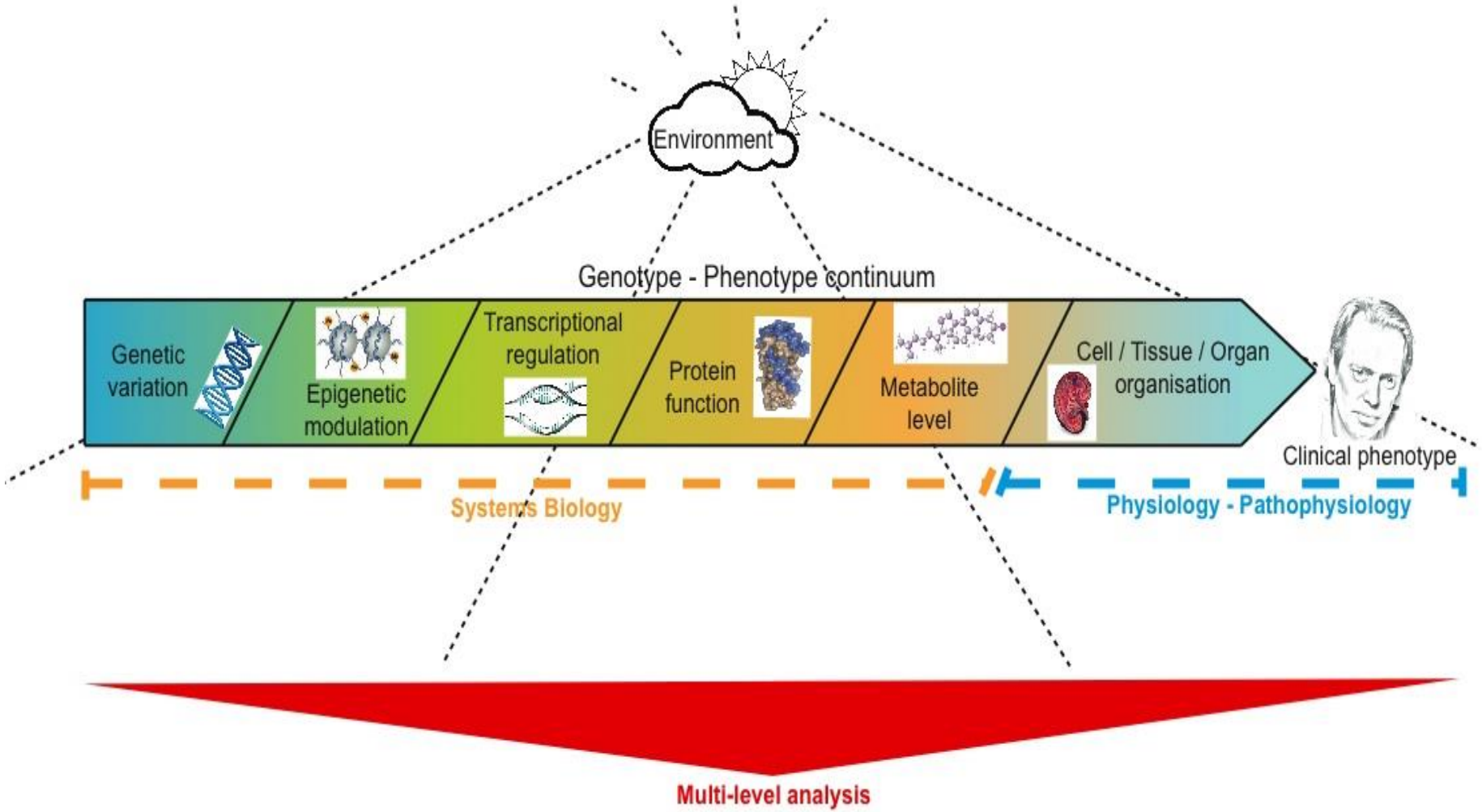
Peter Song
Department of Biostatistics
University of Michigan School of Public Health
September 2, 2015

Funding: P20 ES018171/RD 834800 , P30 ES017885
P01 ES02284401/RD 83543601, P30 DK089503, RO1 ES024732

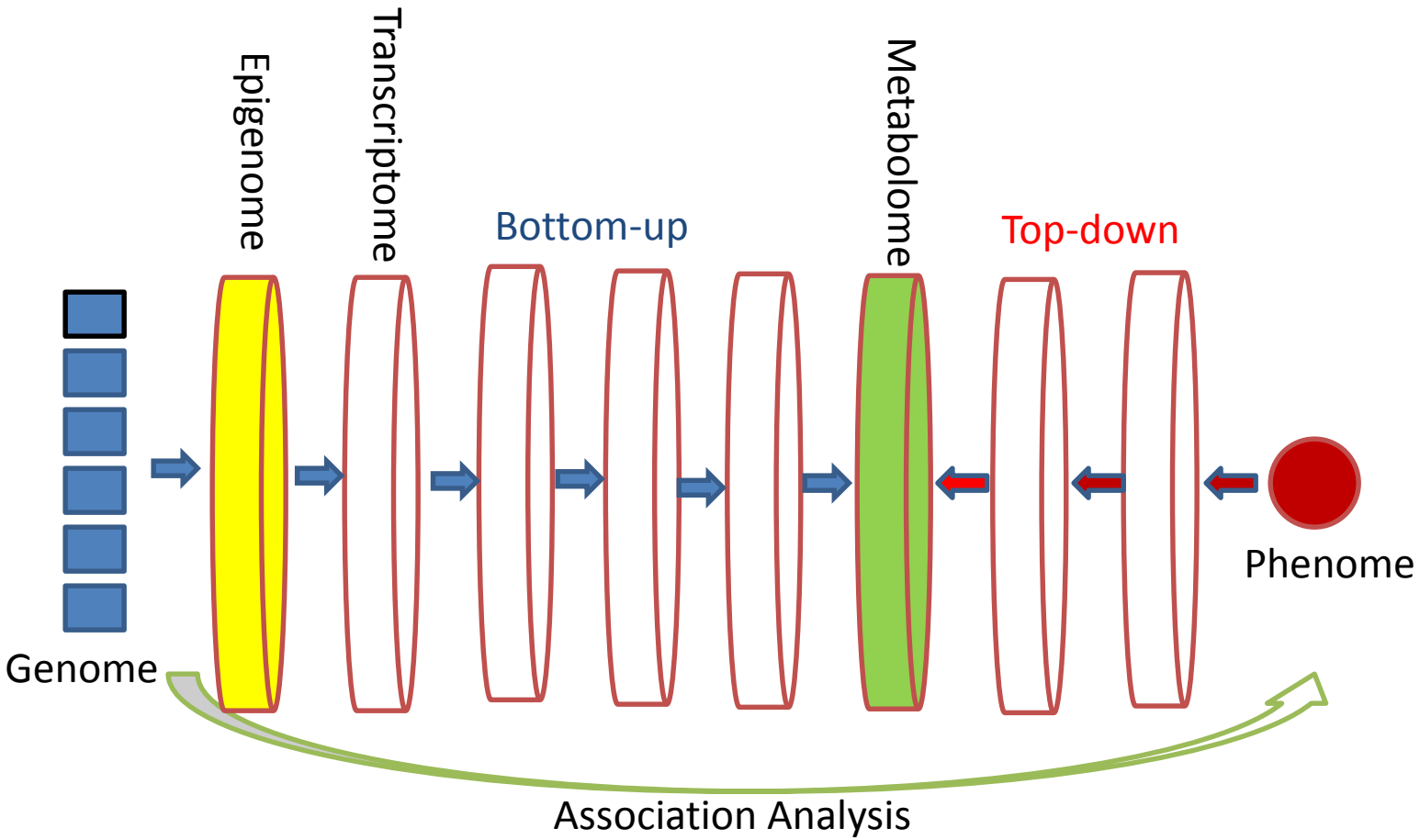
Pitch I:
Biological pathway is the key to success



Translational Biomedical Sciences



Epigenome and Metabolome



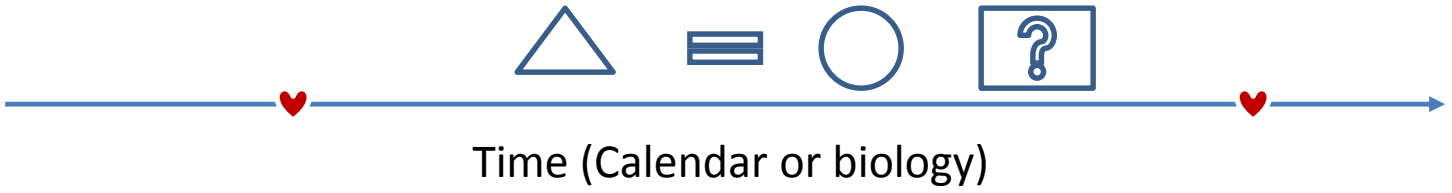
Both "bottom-up" and "top-down" analyses may involve high throughput (high-dimensional) data, and it is an analytic challenge to overcome in order to minimize false discovery.

Integrative Analysis of Metabolomics and Epigenetic Data in Environmental Health Research

- Prenatal, early postnatal, and concurrent exposures influence children's growth and development
- Epigenetic perturbations
  - Biomarker of persistent changes from prenatal exposures and/or the cumulative impact of continued exposures
  - Stable yet potentially modifiable (via pharmacological, dietary, and perhaps lifestyle changes)
  - Mechanistic link between exposures and outcomes
- Metabolite profiles
  - Biomarker of phenotypic changes and possible risk for future chronic disease
  - Mechanistic link between exposures, epigenetics, and outcomes
- Cumulative Risk Assessment
  - Children are exposed to multiple environmental agents throughout development
  - Understanding relationships between exposures, the epigenome, metabolome, and growth will enable better understanding of the cumulative impact of common exposures on life-long health

Pitch II:
Epigenetic change/effect is hard to measure

Numerical (stochastic) change or biological (genetic) change?



- Design of time points for data collection
- Quality control during data collection
- Data pre-processing and pre-treatment
- Types of changes to extract from cleaned data
- Bring biology in and focus on target genes or not
- Others…

250 Subjects from ELEMENT Longitudinal Cohort Study

- Early Life Exposure in Mexico to ENvironmental Toxicants (ELEMENT)
  - 4 longitudinal cohorts: 1994, 1997, 2000, 2005
  - Low-to-moderate income populations in Mexico City
  - Lead exposure is measured in prenatal and concurrent

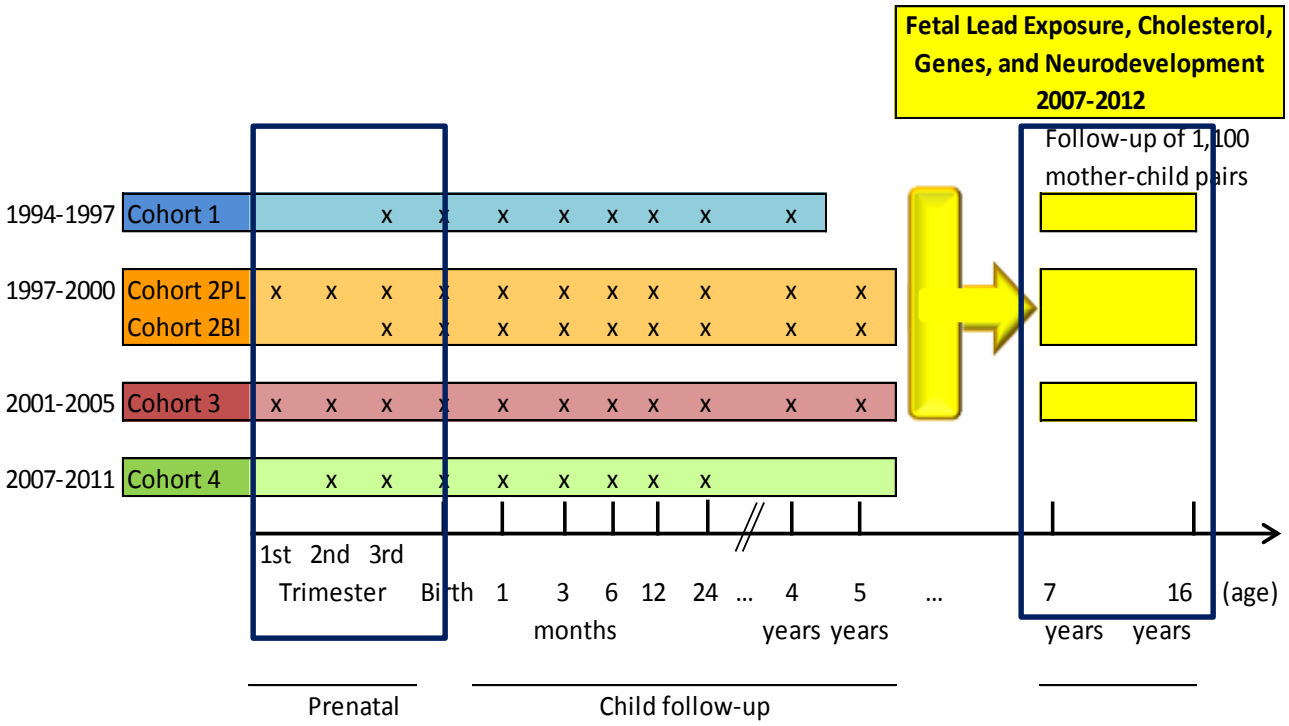
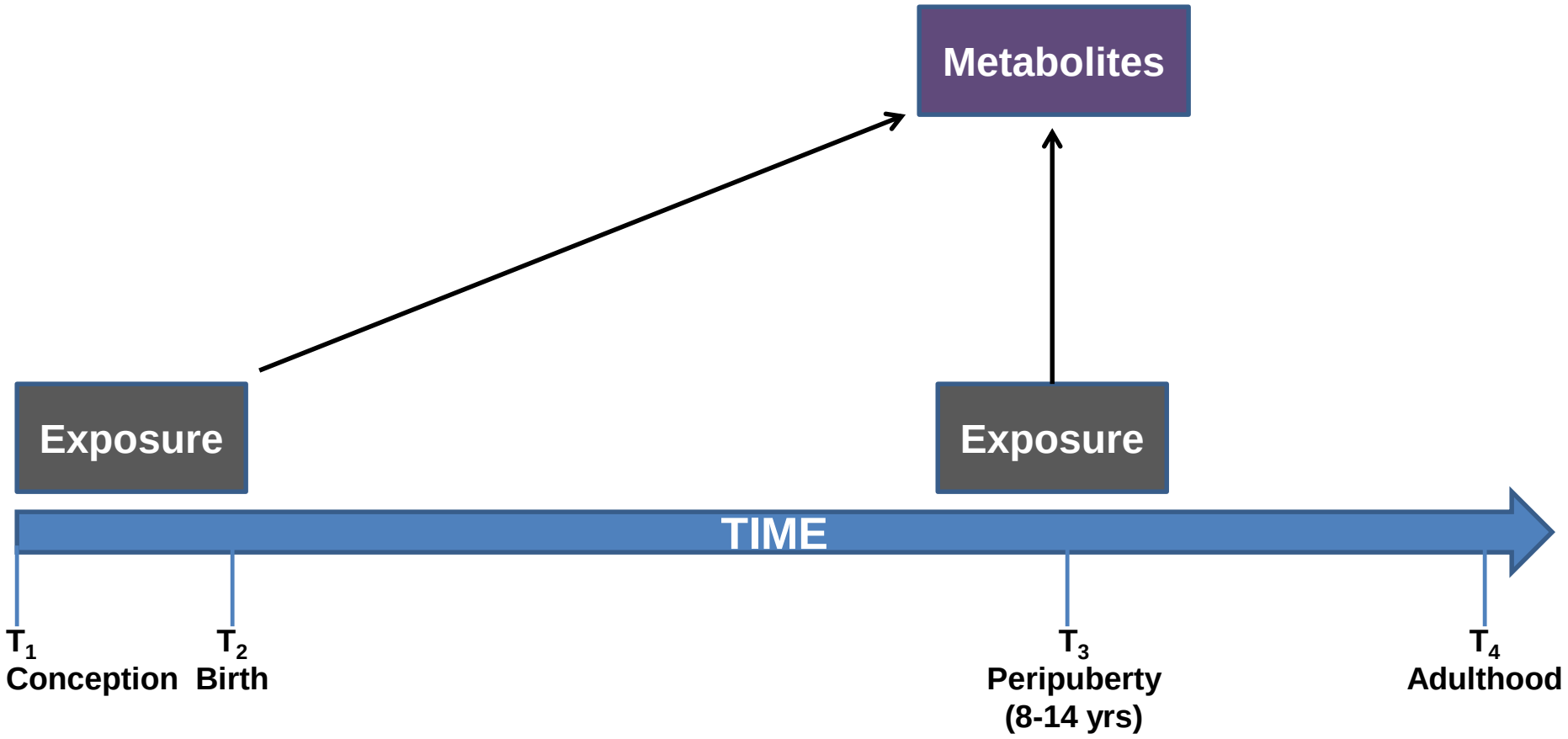


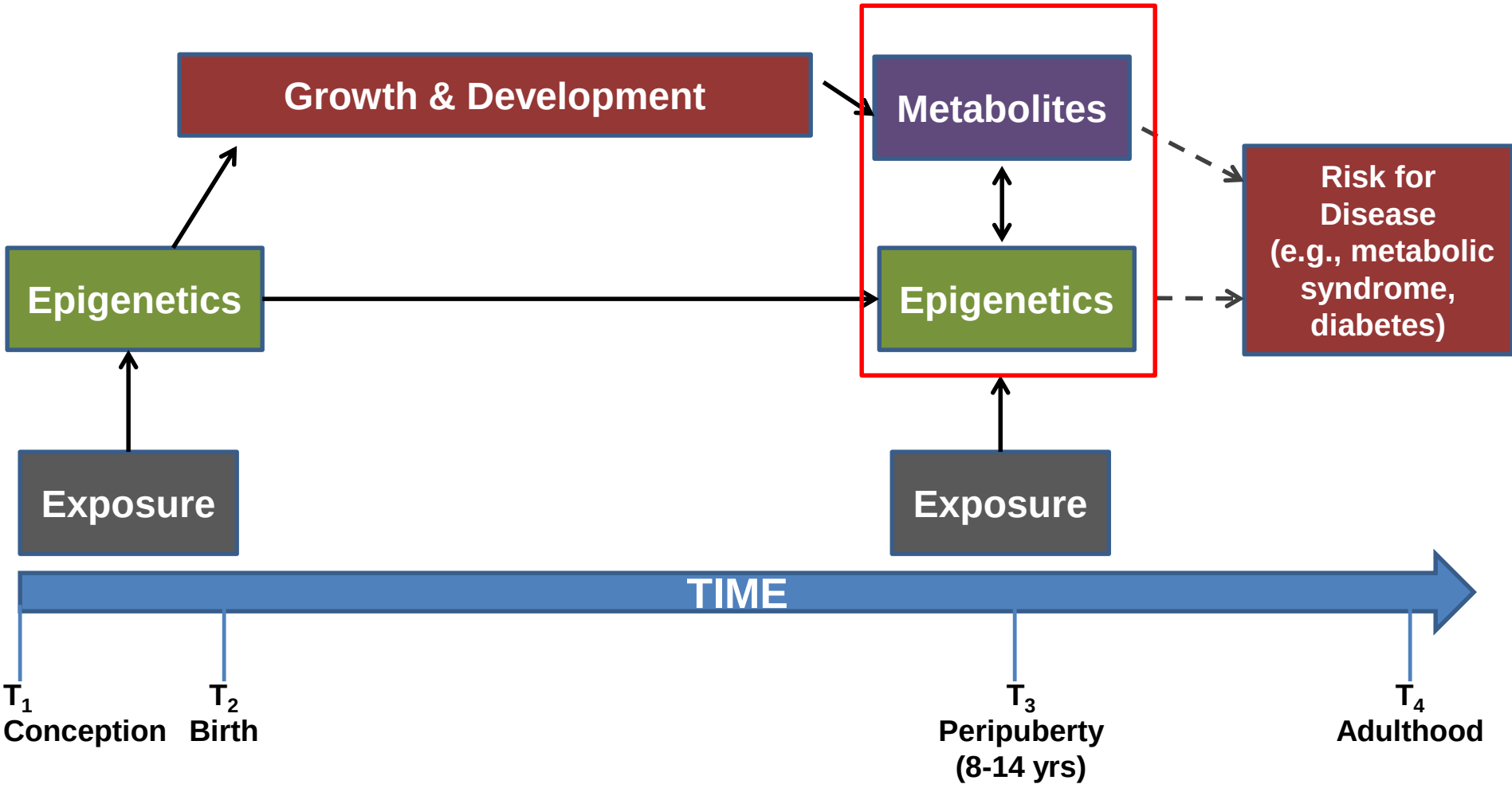
Figure created by Myriam Afeiche

Metabolites may be affected by Prenatal and Concurrent Exposures

To identify metabolites out of 10K features affected by exposures



Some Hypothesized Relationships between Exposures, Epigenetics, Metabolites, & Outcomes

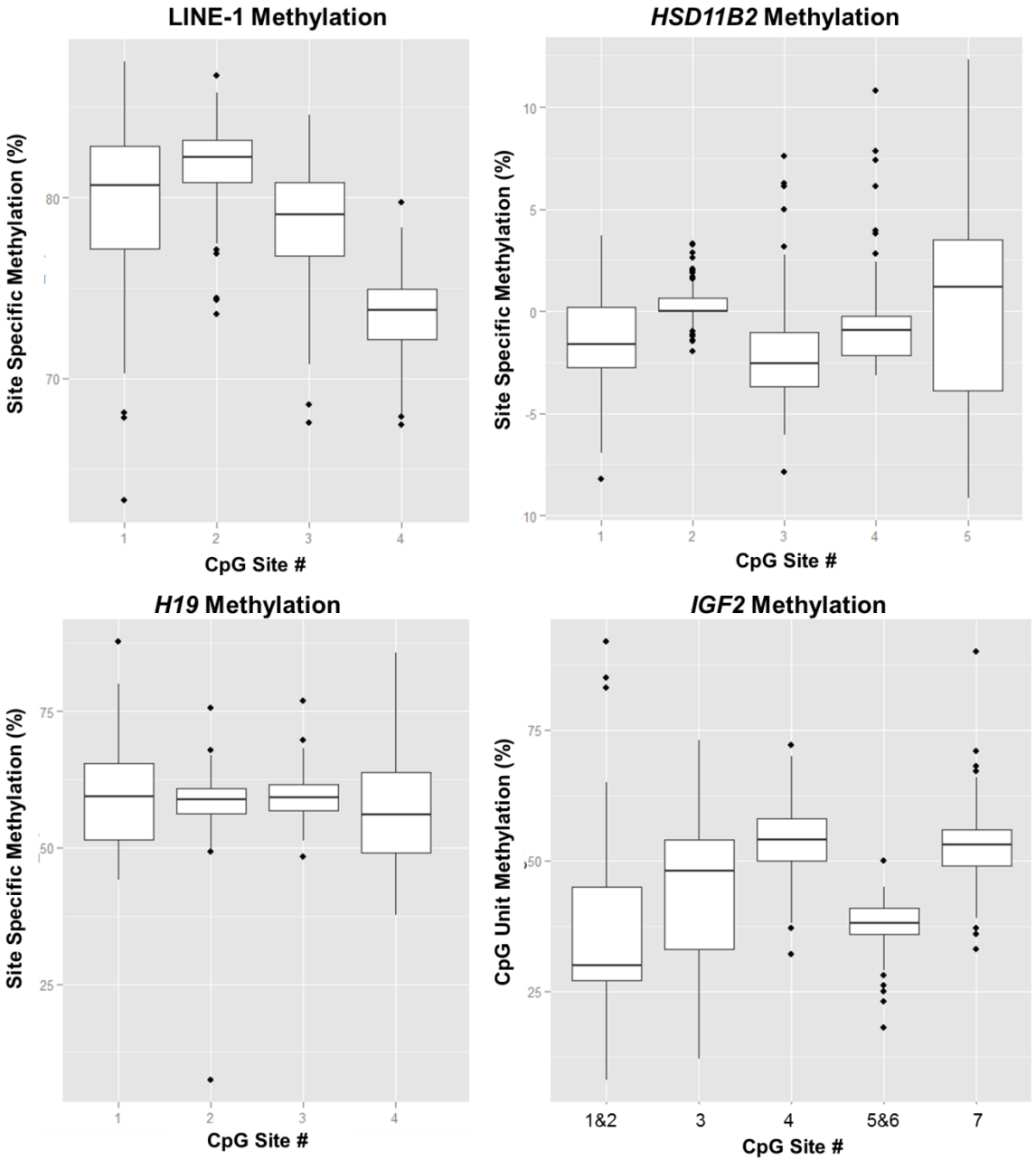


Epigenetic changes may occur from the early life exposure that can propagate through development, and influence metabolite levels in peri-puberty.

DNA Methylation on Candidate Genes

- Methylation quantified in blood leukocyte DNA from peri-puberty of the ELEMENT cohort (n=250)
- Candidate genes/regions selected
  - *LINE-1*
    - Representative of global repetitive element methylation
    - Hypermethylation of repetitive elements necessary to suppress retrotransposition and maintain genomic stability
  - *IGF2* and *H19*
    - Reciprocally imprinted genes important for *in utero* growth
    - *IGF2* promotes growth, *H19* inhibits growth
  - *HSD11B2*
    - Protects against the growth-inhibitor, cortisol, during *in utero* development
- Methylation at these regions associated with *in utero* exposures

Methylation Levels at the Candidate Genes



-Pyrosequencing used for LINE-1, *HSD11B2*, *H19*

-Sequenom EpiTYPER used for *IGF2*

-Data adjusted for experimental batch

-Wide intra-region variability across CpG sites with the exception of *H19*

Metabolomics: Select Candidate Metabolomics Features

Preprocessed metabolomics features (log scale)

Normalization (using QC pooled samples as reference) and standardization

Filtering features of high missing and large batch differences

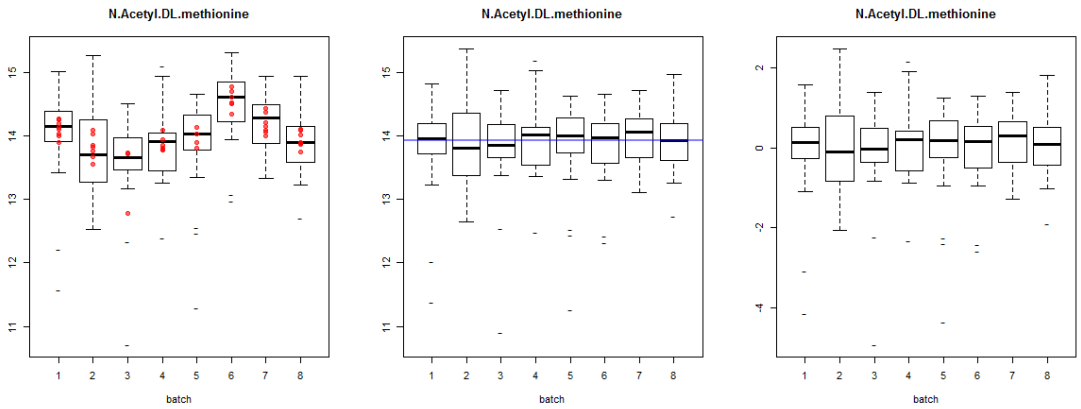
Rank-based inverse normal transformation

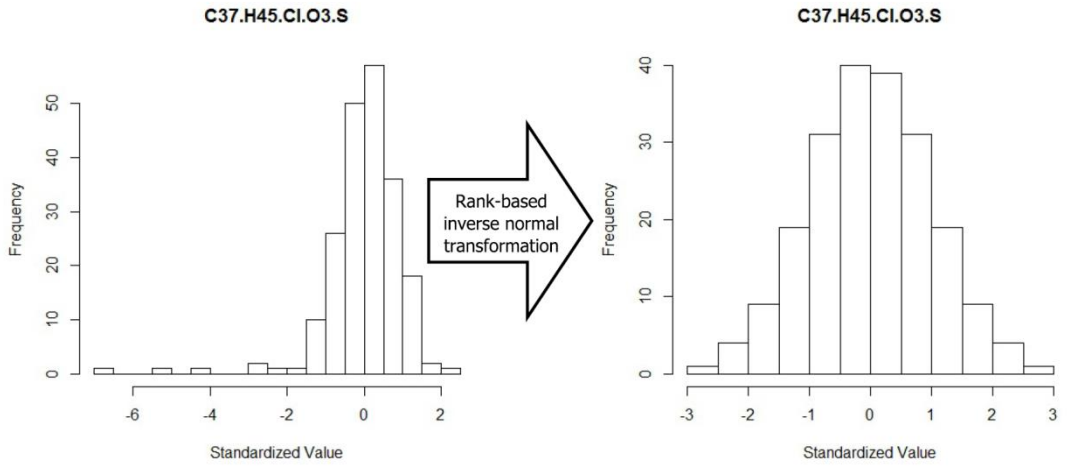
Random batch effect model for each feature

Ranked p-values adjusting for multiple-comparison

Select top-ranked unknown features and known features to build association network

Experimental verification of selected features





High-dimensional Feature Screening

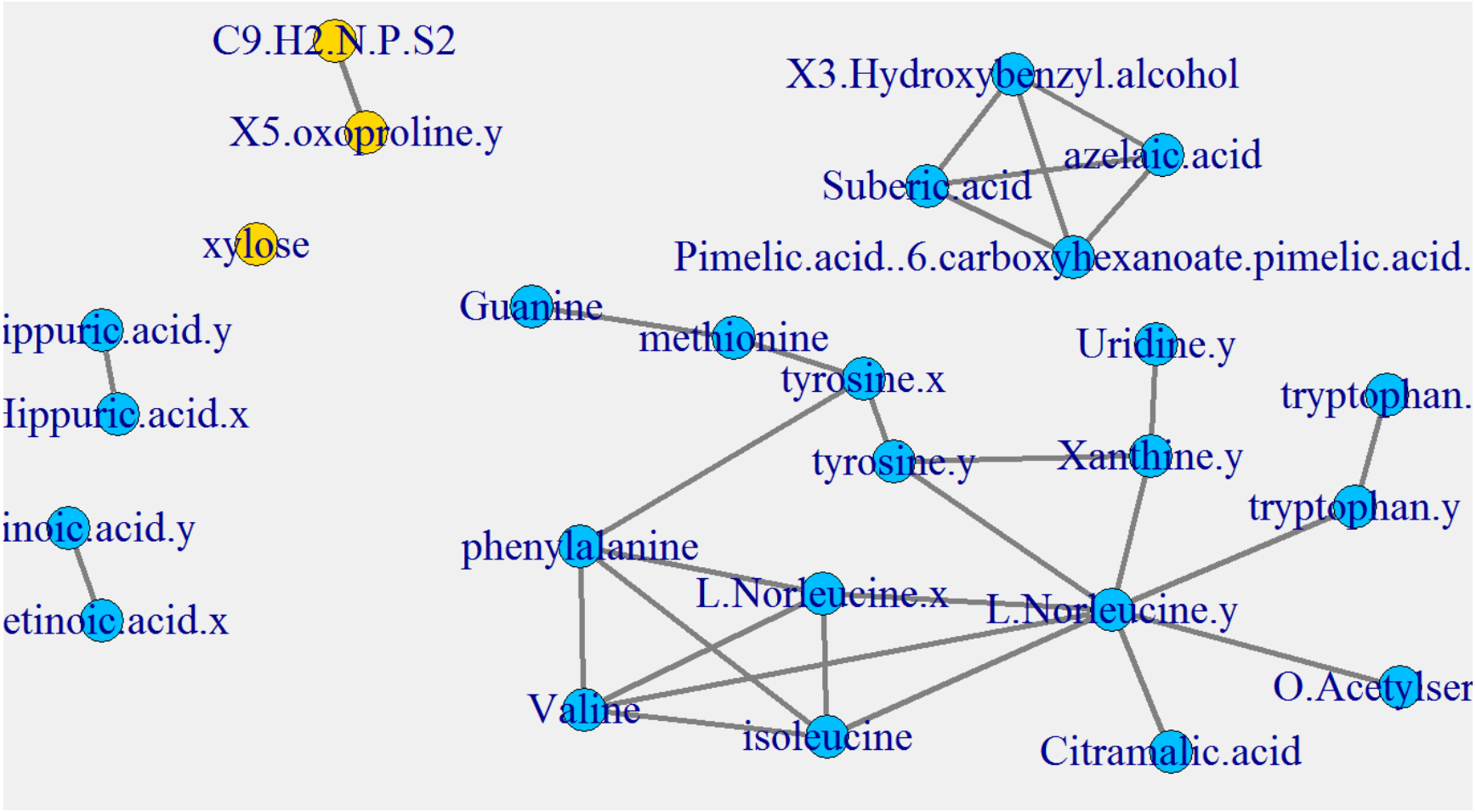
- **Random Batch Effects Model:**

$$\text{Metabolite}_i = \beta_0 + \beta_1 Pb^{prenatal} + \beta_2 sex + \beta_3 age + \beta_4 Pb^{concurrent} + b_i + \epsilon_i$$

i=1,2,…,8 denotes the i-th batch
Prenatal Pb = maternal patella bone Pb
Concurrent Pb = blood Pb

- **Screening Test:** $H_0: \beta_1=0$ (or $H_0: \beta_4=0$)
- **High-dimensional Problem:** ~10,000 features are screened simultaneously
- **Significance level: q value (FDR adjusted p-value) 0.05**
- **Three metabolomics features are detected via the screening on $H_0: \beta_1=0$**

Three Candidates Detected in Network of Known Metabolite Features



Statistical Models for Integrative Analysis of Concurrent Measures

In Sex-Stratified Analyses:

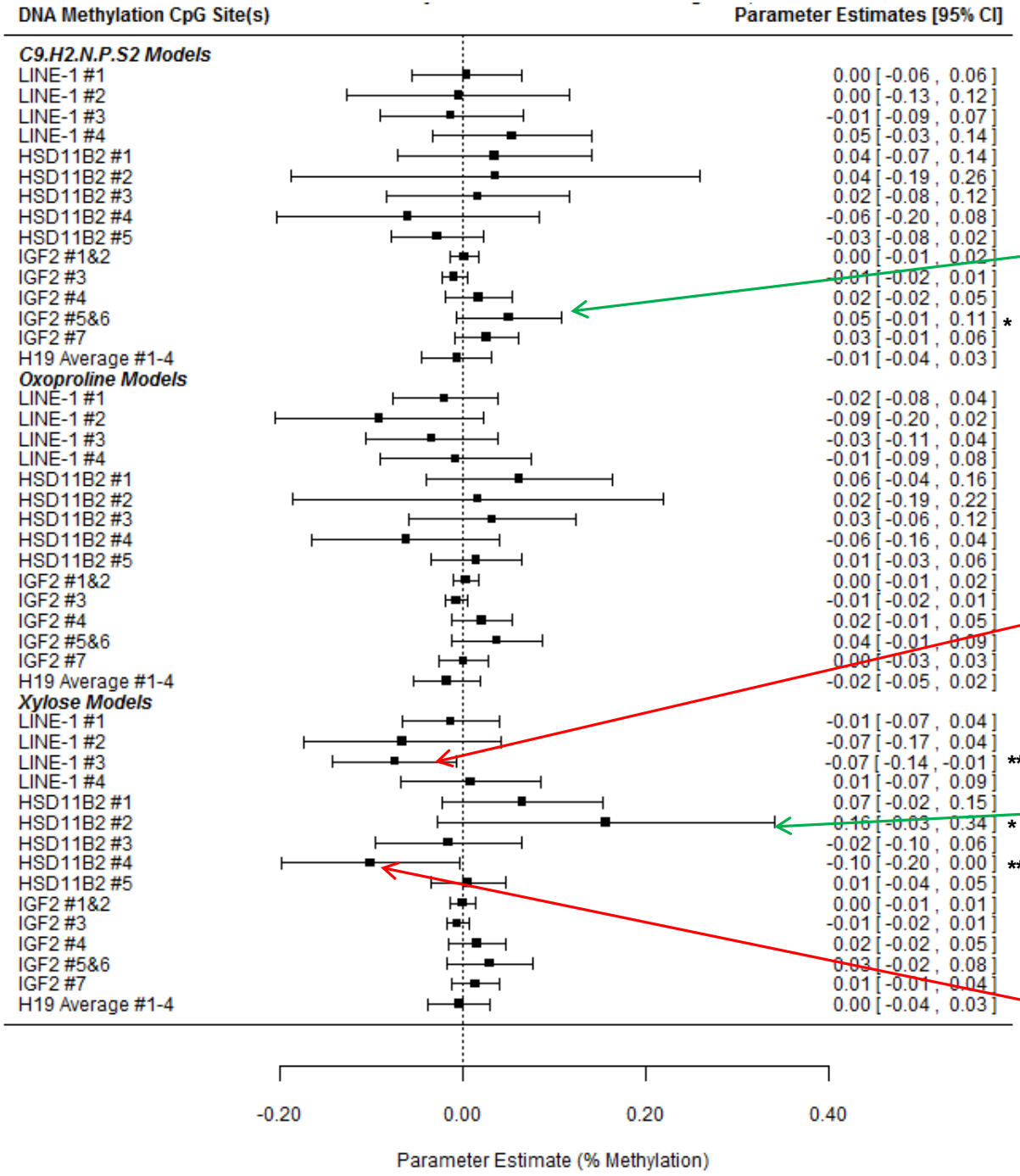
(1) $\text{metabolite}_i = \beta_0 + \beta_1 \text{age} + \beta_2 \text{methylation} + \beta_3 Pb^{prenatal} + b_i + \varepsilon_i$

(2) $\text{metabolite}_i = \beta_0 + \beta_1 \text{age} + \beta_2 \text{methylation} + \beta_3 Pb^{prenatal} + \beta_4 Pb^{concurrent} + b_i + \varepsilon_i$

Metabolites tested: C9.H2.N.P.S2, X5.oxoproline, xylose

Methylation: percent methylation at one of the following:

1) 4 individual CpG sites in LINE-1
2) 5 individual CpG sites in *HSD11B2*
3) Average of all sites in *H19*
4) 5 CpG units in *IGF2* (some units are the average of 2 sites)



DNA Methylation as a Predictor of Metabolites among Girls

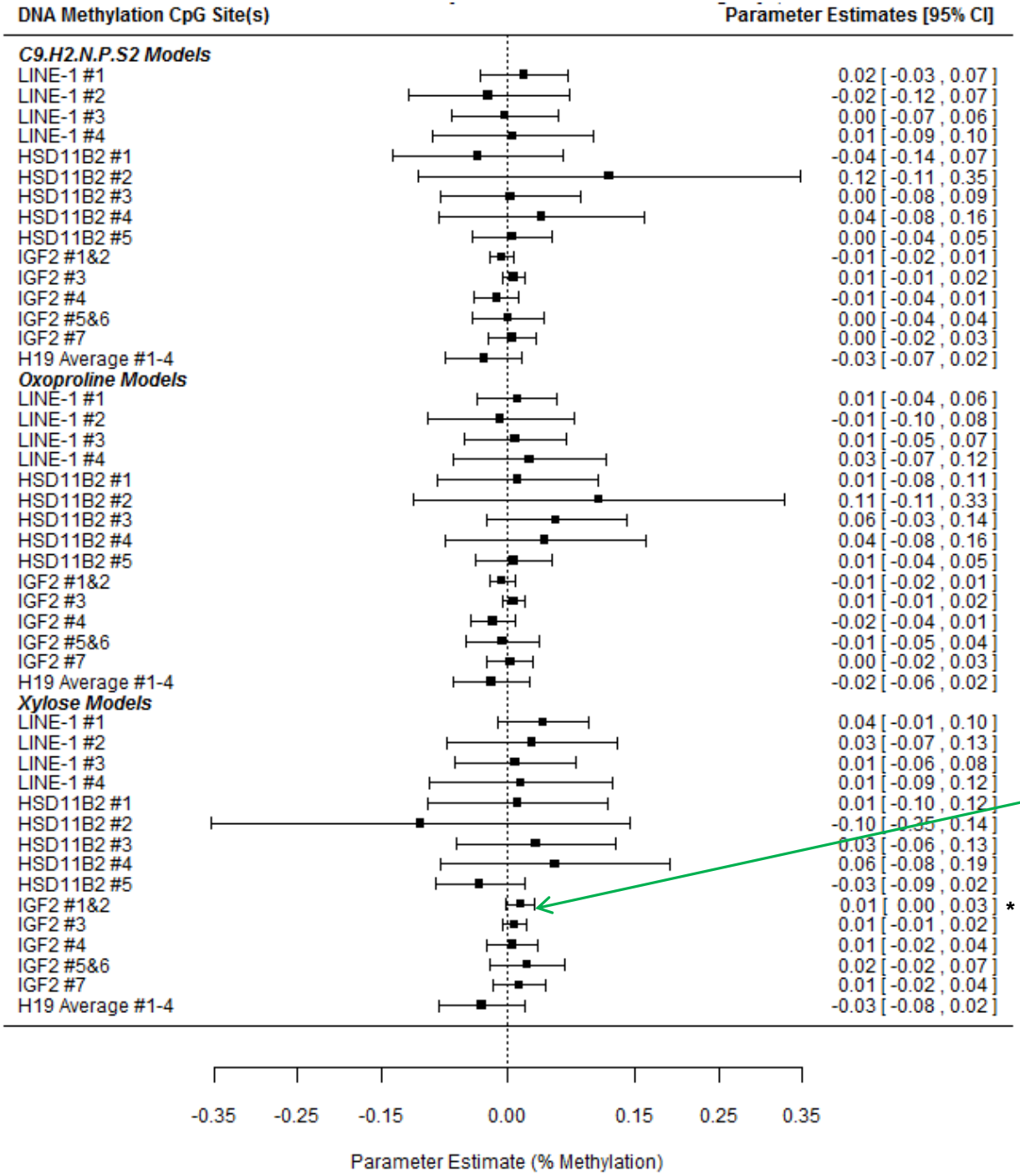
p<0.1 associations:

IGF2 CpGs 5&6 and ↑C9.H2.N.P.S2

LINE-1 CpG 3 and ↓Xylose

HSD11B2 CpG 2 and ↑Xylose

HSD11B2 CpG 4 and ↓Xylose



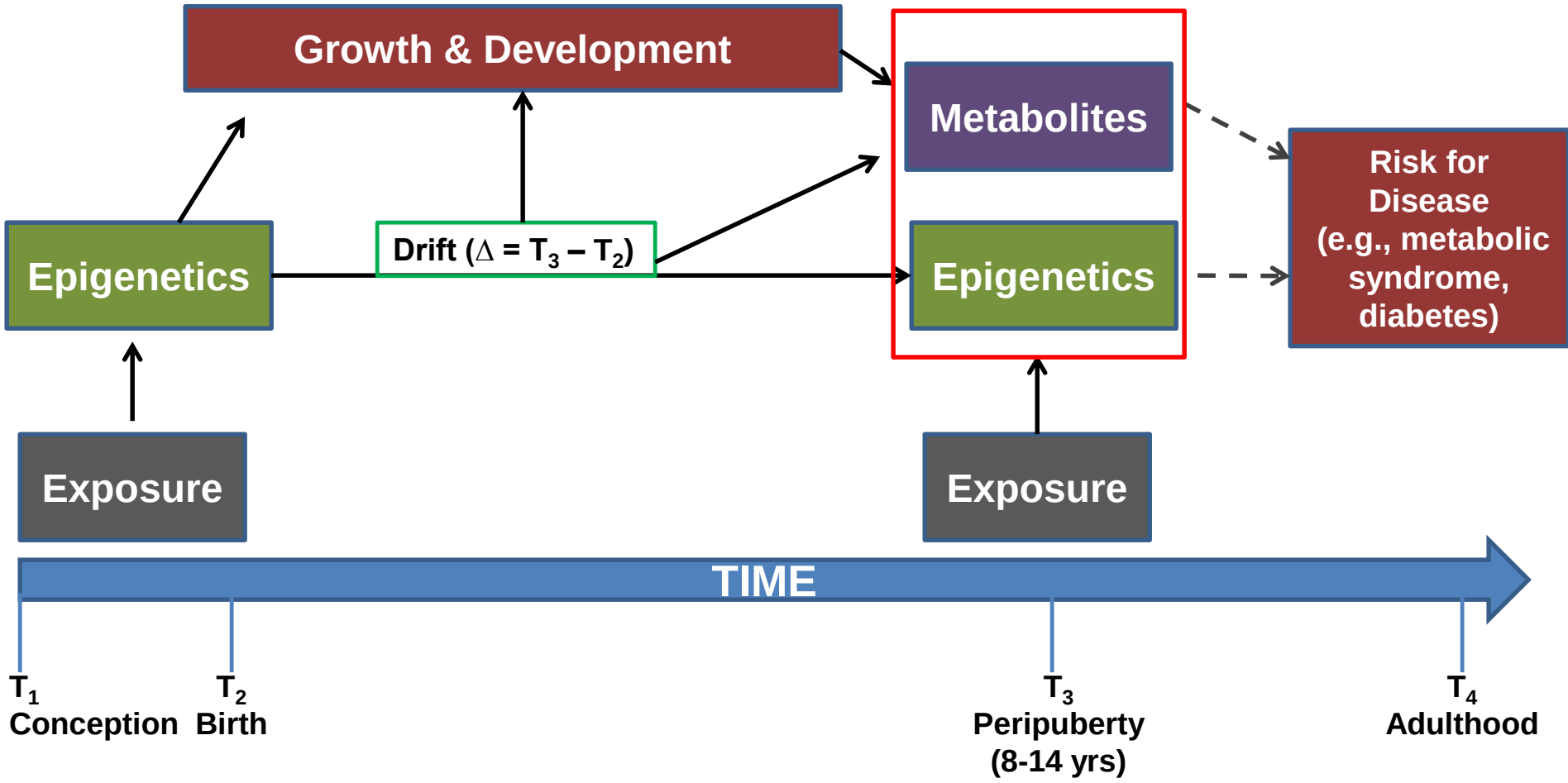
DNA Methylation as a Predictor of Metabolites among Boys

p<0.1 association:

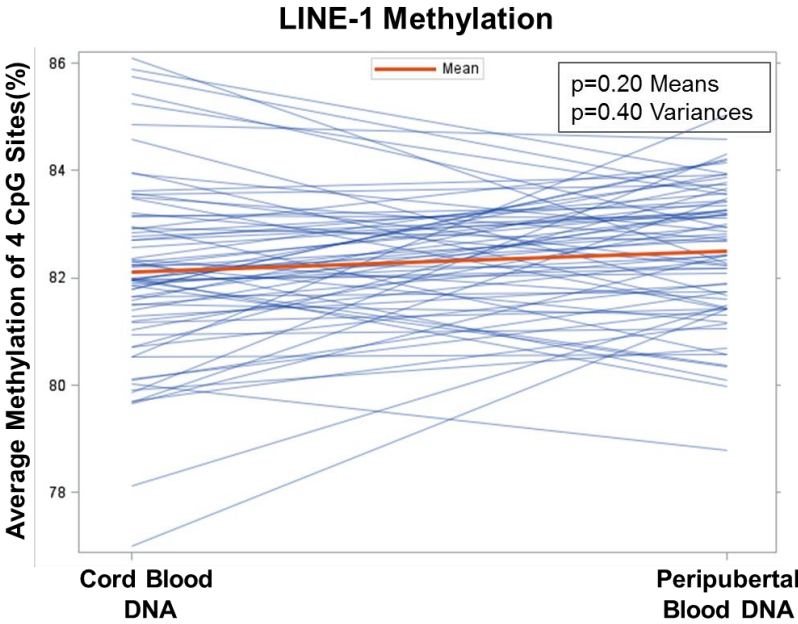
IGF2 CpG 1&2 and ↑Xylose

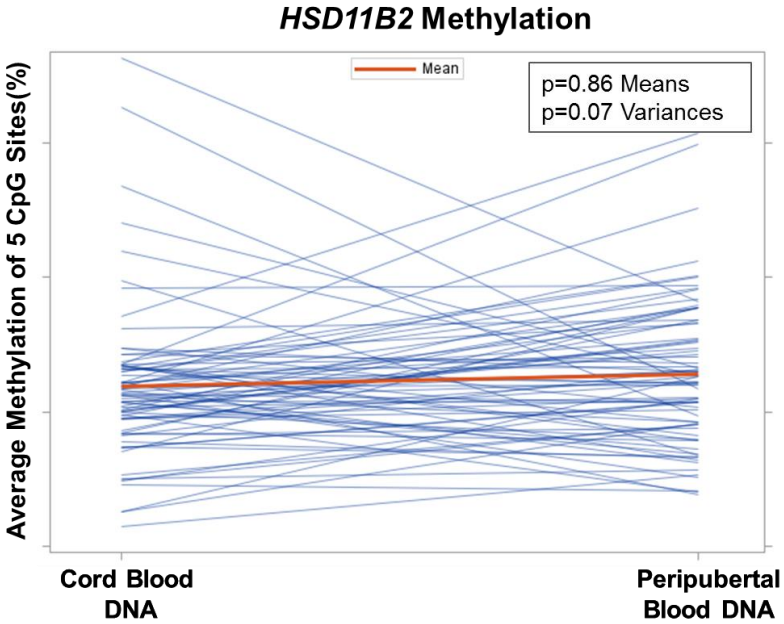
Does Epigenetic Drift Play a Role?

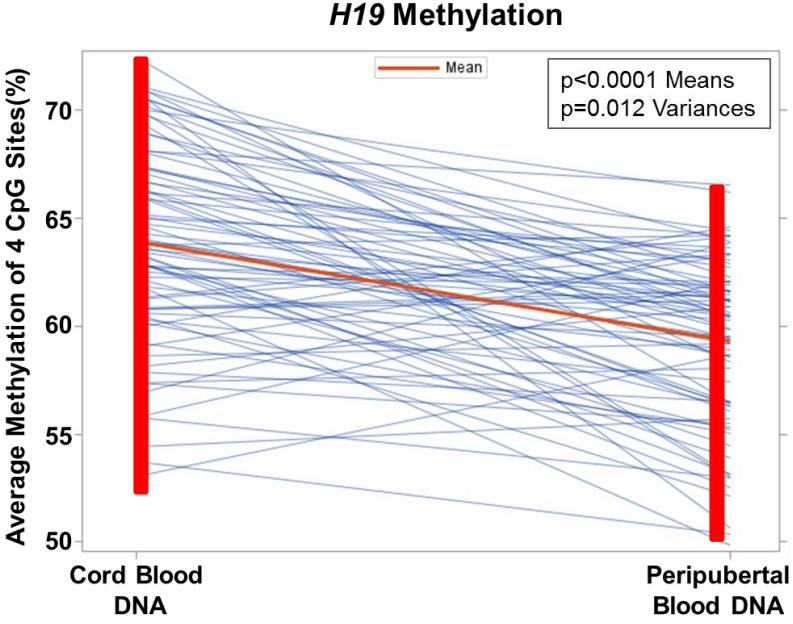
Epigenetic changes from the early life exposure propagate through development, and influence metabolites in peri-puberty

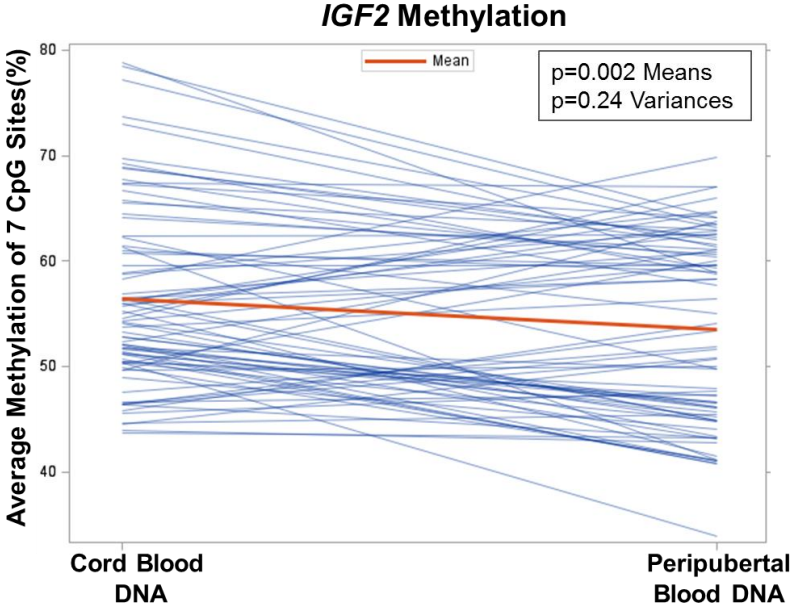


DNA Methylation Drift in the ELEMENT Cohort (n=78)









-H19 methylation mean and variance significantly decreased with time.

-IGF2 methylation significantly increased with time.

Incorporating Epigenetic Drift

(Drift #1) $\text{metabolite}_i = \beta_0 + \beta_1\,\text{age} + \beta_2\Delta\text{methylation} + \beta_3 Pb^{\text{prenatal}} + b_i + \varepsilon_i$

(Drift #2) $\text{metabolite}_i = \beta_0 + \beta_1\,\text{age} + \beta_2\Delta\text{methylation} + \beta_3\Delta\text{variability} + \beta_4 Pb^{\text{prenatal}} + b_i + \varepsilon_i$

Metabolites tested: C9.H2.N.P.S2, X5.oxoproline, xylose

Δ **methylation:** (methylation_j at t_3 – mean methylation at t_3) - (methylation_j at t_2 – mean methylation at t_2)

Δ **variability:** (methylation_j at t_3 – mean methylation at t_3)^2 - (methylation_j at t_2 – mean methylation at t_2)^2

t_2= birth (cord blood leukocyte DNA)
t_3 = peri-pubery (blood leukocyte DNA)

Association of Methylation Drift on Metabolites

- Methylation Drift:

Among those that passed statistical test of a significant mean change between two times, methylation changes of *IGF2#1 &2* and *IFG2#7* are associated with *xylose*.

- Methylation Variability Drift:

Among those that passed statistical test of a significant variance change between two times, the change of variability of *IGF2 5&6* is associated with *xylose*.

Concluding Remarks

- Integrative analysis involves high-dimensional data, in which most of them are noise. To reduce false discoveries, good data is of most importance.
- Statistical design of cohort studies remains a difficult problem, in particular there are many confounding factors involved in such studies.
- Multiple steps are required in data processing and data analysis, which incurs accumulation of human errors along the process, and any findings must be put for validation.
- It remains unknown whether or not, if so and how, to combine site-specific methylations or combine metabolite features.

Future Directions

- Apply structural equation model for high-dimensional mediators to assess mediation effect of biomarkers.
- Utilize data mining techniques to identify biomarkers sensitive to past exposures and predictive to outcomes related to exposures.
- Derive epigenetic or metabolomics change as markers of cumulative exposures or cumulating risk for disease development.
- Analyze genome-wide methylation (being collected by the 450K platform through the P01 Admin Supplement) and full metabolomics data to find new biomarkers or mechanistic pathways.

University of Michigan CEHC Team

- **CEHC Directors**: Karen E. Peterson, Vasantha Padmanabhan
- **Project/Core Leaders**: Dana Dolinoy, John Meeker, Alison Miller, Peter Song
- **Instituto Nacional de Salud Pública PI**: Mara Tellez-Rojo
- **Investigators/Consultants:** Alejandra Cantoral, Jorge Chavarro, Adrienne Ettinger, Howard Hu, Joyce Lee, Sub Pennathur, Lourdes Schnaas, Brisa Sánchez
- **Center Staff**: Lindsey Mitchell, Tamara Jones, Samantha Milewski
- **Postdoctoral Fellows:** Christopher Faulk, Kelly Ferguson, Jaclyn Goodrich, Deborah Watkins
- **Graduate Students:** Emily Hector, Joe Kochmanski, Lisa Marchlewicz, Meghan Moynihan, Lu Tang, Zhenzhen Zhang
- **INSP Field Staff**

Acknowledgements:

Jackie Goodrich, Lu Tang, Emily Hector, Lisa Marchlewicz, Karen Peterson, Dana Dolinoy, **Michigan Regional Comprehensive Metabolomics Resource Core :** Alla Karnovosky, Charles Burant