Tissue purity and cell composition challenges in estimation of association between somatic copy number aberration, DNA methylation, and gene expression

Wei Sun[123], Paul Bunn[2], Chong Jin[2], Paul Little[2], Vasyl Zhabotynsky[2], Charles M. Perou[34], David N. Hayes[4], Mengjie Chen[23], Dan-Yu Lin[24]

[1]Public Health Science Division, Fred Hutchison Cancer Research Center;
[2]Department of Biostatistics, University of North Carolina, Chapel Hill;
[3]Department of Genetics, University of North Carolina, Chapel Hill;
[4]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill

Summary

Tumor purity and cell composition, though often are latent or noisily estimated are important confounders leading to critically reduced ability to distinguish true.
As we show, based on multiple cancer multiple data collected by The Cancer Genome Atlas (TCGA) without accounting for them eQTL analysis (performed using MatrixEQTL[1] software) produces massive 90% of local and even higher proportion of distant associations between gene expression and methylation.
We suggest a method allowing to significantly reduce impact of such latent confounders [2].

Data used

Cancer type:

- Breast (presented)
- Colon
- Glioblastoma
- Leukemia
- Lower-grade glioma
- Prostate

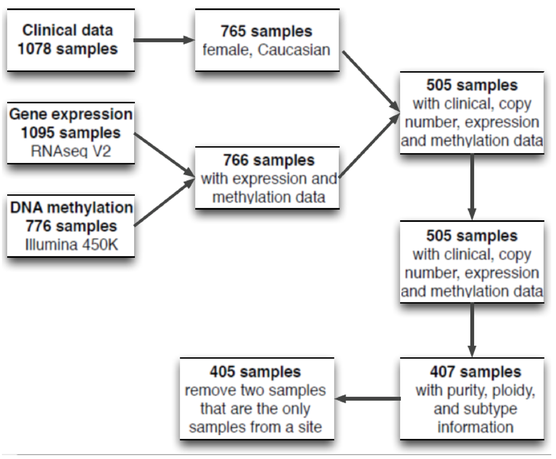
Data type:

- Somatic copy number aberration (SCNA)
- DNA methylation (m-values - logit transformed)
- Gene expression (log transformed gene counts)

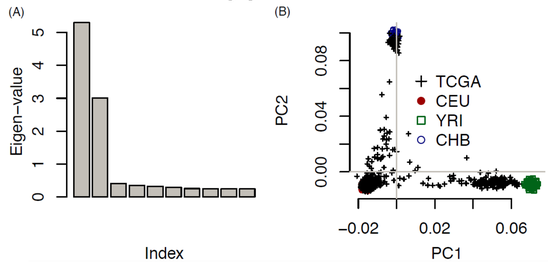
General model considerations

Transformed m-values and log transformed normalized gene counts provide a reasonable distributional assumption for linear model to be used.
Batch effects are quire big, so we routinely incorporate covariates account for the effects of tissue sites, plates as well as demographic covariates.

Sample selection

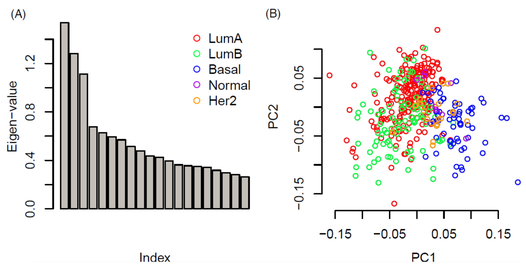


Genotype PCA

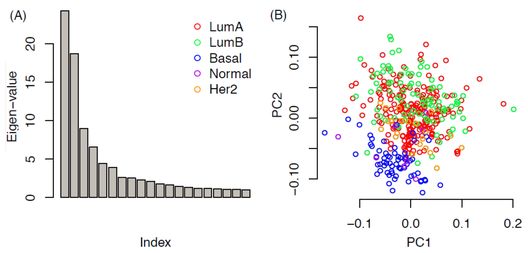


For this analysis we declare Caucasian and select a subset of samples with both PC depicted in figure (B) less than 0.

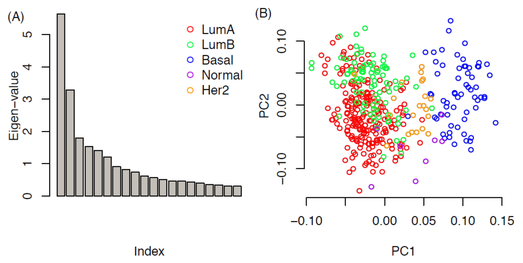
PCA on SCNA



PCA on Methylation



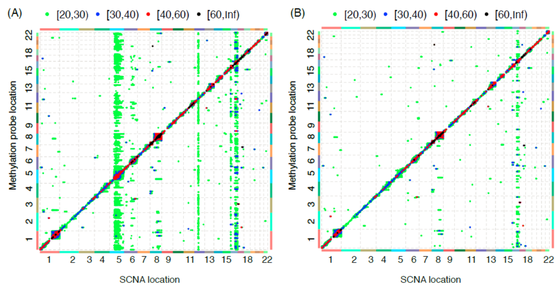
PCA on Gene expression



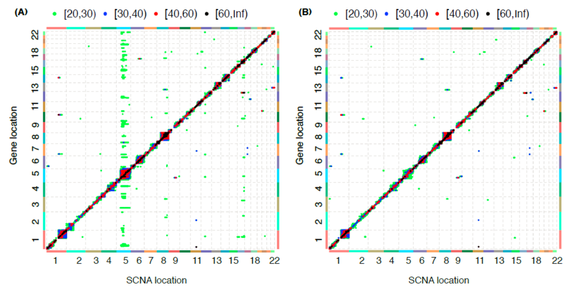
breast cancer subtypes have a strong influence all the above types of data

Methylation vs SCNA

In following slides figure (B) shows results after adding genotype PC and tumor subtypes compared with (A) when those are not included

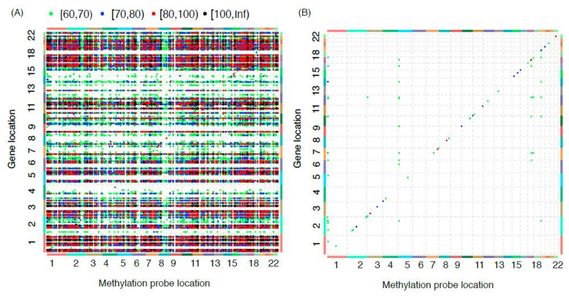


Gene expression vs SCNA



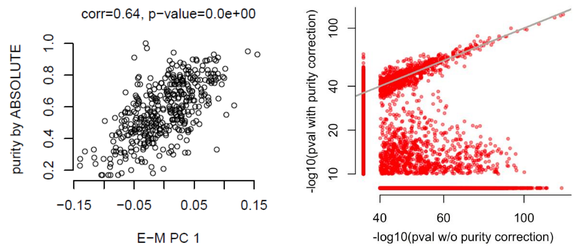
Gene expression vs Methylation

(A) shows results after adding genotype PC and tumor subtypes and
(B) represents the model including SCNA, tumor subtypes, and 1st PC from correlated methylation-expression (ME) pairs



Local false-positive reduction

Thus we need to adjust for purity confounding after which we observe most of significant p-values disappearing completely or much weaker



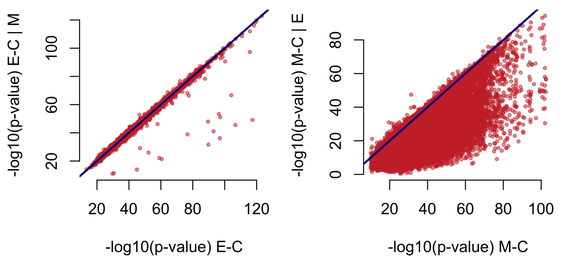
Conditional associations

To distinguish between two possible hypotheses of relationship between SCNA (C), gene expression (E) and methylation (M) we are considering:

(a) C -> E -> M vs (b) C -> M -> E

(a) SCNA is associated with gene expression, gene expression is associated with methylation at CpG islands located around promoter regions, and thus SCNA is often associated with methylation at CpG islands i.e. E is a mediator between C and M
(b) Suggesting that methylation M is a mediator.

To evaluate either hypothesis we add extra covariates produced by PCA on standardized residuals of significantly associated distant Methylation-Gene expression pairs.



Contrasting E-C analysis to M-C analysis in the above plots we see that conditioning on those principal components reduces association strength to much higher degree for M-C model whereas for E-C model for most of the tests association strength stays similar.
We consider these results to be evidence of C->E->M model

Citations

1 Shabalin, Andrey A. "Matrix eQTL: ultra fast eQTL analysis via large matrix operations." Bioinformatics 28.10 (2012): 1353-1358.
2 Wei Sun, Paul Bunn, Chong Jin, Paul Little, Vasyl Zhabotynsky, Charles M. Perou, David N. Hayes, Mengjie Chen, Dan-Yu Lin. " The association between somatic copy number aberration, DNA methylation, and gene expression." PNAS (submitted) (2016)