Removing batch effects in gene expression array study

Chung-Hsing Chen

National Cancer Institute, National Health Research Institutes

Data background (1)

- Hedenfalk et al. measured **3226** genes in seven BRCA1 and eight BRCA2 mutation-positive tumor samples.
- The goal of the study was to identify genes that showed differential expression across breast cancer tumor subtypes defined by these germline mutations.
- Several genes with apparent outliers were removed. This left **3170** genes.

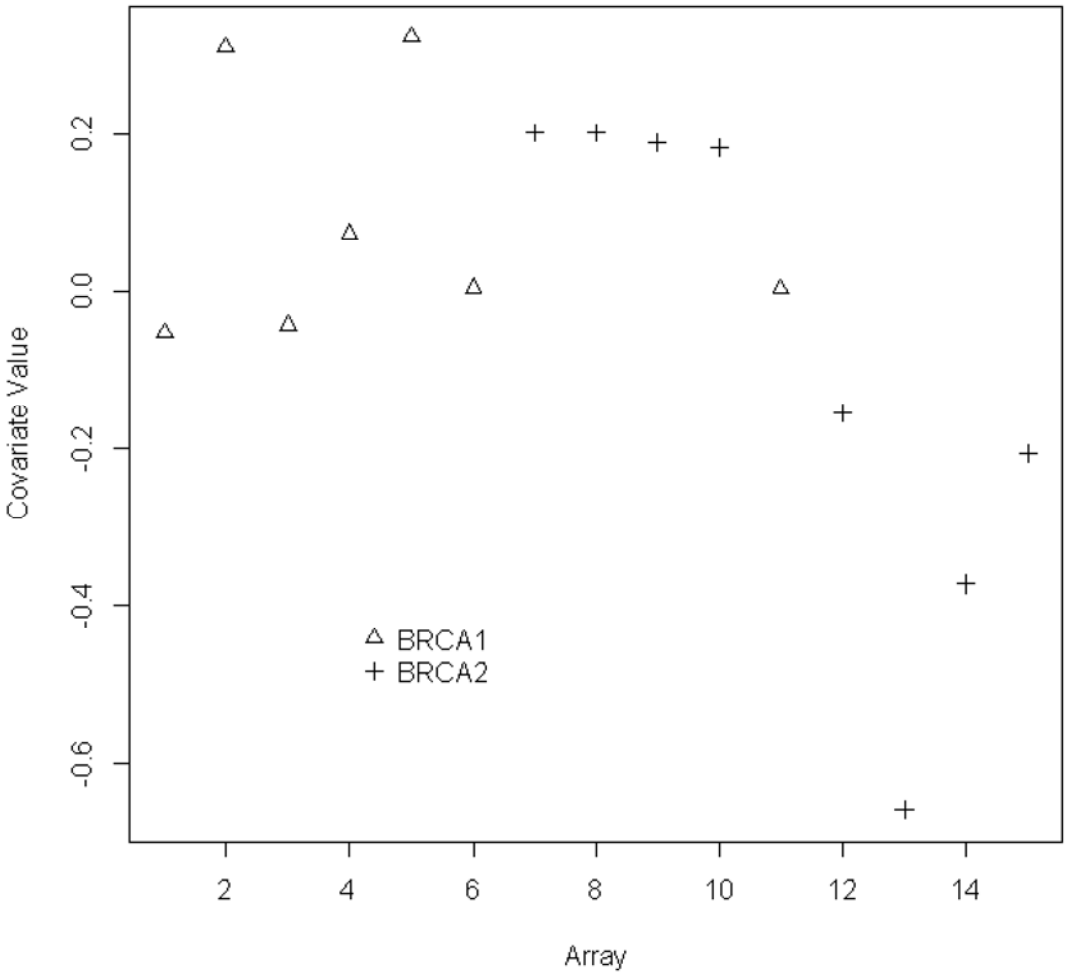
Data background (2)

```
> SVA.pheno
    NEJM-PatientID Mutation
V7               1    BRCA1
V8               5    BRCA1
V9               3    BRCA1
V10              7    BRCA1
V11              2    BRCA1
V12              4    BRCA1
V13             10    BRCA2
V14              9    BRCA2
V15              8    BRCA2
V16             10    BRCA2
V17              6    BRCA1
V18             13    BRCA2
V19             14    BRCA2
V20             11    BRCA2
V21             12    BRCA2
> head(SVA.exp.preprocessed)
      q-value    p-value fold-change (log base 2) s1996 s1822 s1714 s1224 s1252 s1510 s1900 s1787 s1721 s1486 s1905
[1,] 0.089529 0.01223344                    1.203  0.15  0.22  0.30  0.26  1.22  0.44  0.35  1.10  1.07  1.46  0.38
[2,] 0.213752 0.07611987                   -0.521  1.54  1.27  0.76  0.85  1.27  0.64  0.90  0.64  0.78  0.55  0.61
[3,] 0.672438 0.99530284                    0.002  1.72  1.57  2.13  1.09  1.98  0.74  1.71  1.16  1.33  1.46  2.43
[4,] 0.163987 0.04212934                    0.690  0.71  1.24  1.69  2.23  1.16  0.82  1.44  2.03  3.60  1.20  2.08
[5,] 0.637670 0.84745741                    0.053  0.94  1.53  1.87  1.19  1.16  1.54  1.05  0.91  0.85  1.22  1.01
[6,] 0.377451 0.25437224                   -0.227  0.80  0.95  1.53  1.37  1.02  1.22  0.78  0.96  0.65  1.02  1.09
     s1816 s1616 s1063 s1936
[1,]  0.73  0.63  0.77  0.66
[2,]  0.71  0.30  0.62  1.00
[3,]  1.71  1.26  1.41  3.00
[4,]  3.24  2.41  1.56  2.56
[5,]  3.25  2.20  1.09  1.29
[6,]  0.66  1.40  1.32  1.13
> head(exp.annotation)
  PlatePosition ImageCloneID                                                                                Title
4         HK1A1        21652                                        catenin (cadherin-associated protein), alpha 1 (102kD)
5         HK1A2        22012                                                                 ADP-ribosylation factor 3
6         HK1A4        22293                              uroporphyrinogen III synthase (congenital erythropoietic porphyria)
7         HK1A5        22493                                                                        ribosomal protein L26
8         HK1A6        23019 guanine nucleotide binding protein (G protein), alpha stimulating activity polypeptide 1
9         HK1A7        23132                   pre-mRNA splicing factor SF3a (120 kDa subunit), similar to S. cerevisiae PRP21
```

Substructure detected

- Hierarchical clustering of the data reveals notable substructure within the BRCA2 samples.
- Applied **SVA** (Surrogate Variable Analysis) to identified a single surrogate variable that appears to capture this trend.

Substructure detected



Significance analysis

- Included this surrogate variable in a significance analysis comparing BRCA1 and BRCA2 tumors.
- The number of genes differentially expressed between BRCA1 and BRCA2 before and after adjusting for surrogate variables.

	q-Value Threshold			
Analysis Type	0.01	0.025	0.05	0.10
Unadjusted	1	19	96	275
SVA adjusted	0	10	48	190

R package: sva (1)

- Create the full model matrix - including both the adjustment variables and the variable of interest (BRCA1/BRCA2).
- The null model contains only the adjustment variables.

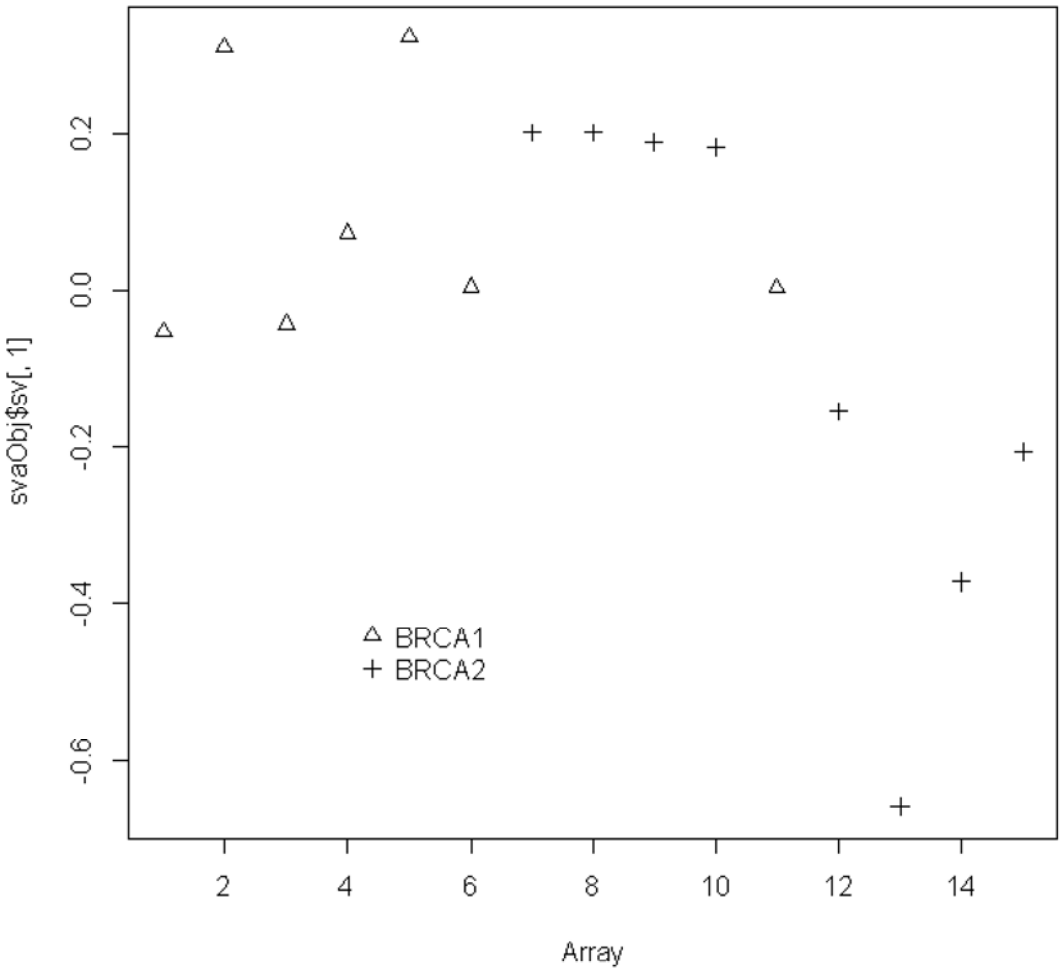
```
> mod=model.matrix(~as.factor(Mutation),data=SVA.pheno)
> mod0=model.matrix(~1,data=SVA.pheno)
> svaObj=sva(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
Number of significant surrogate variables is:  4
Iteration (out of 5 ):1  2  3  4  5  >
```

R package: sva (2)

- The f.pvalue function can be used to calculate parametric F-test p-values for each gene. The F-test compares the models mod and mod0.

```
> pValues.before=f.pvalue(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
> qValuesObj.before<-qvalue(pValues.before)
> qsummary(qValuesObj.before)

Call:
qvalue(p = pValues.before)

pi0:    0.6786508

Cumulative number of significant calls:

        <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1   <1
p-value      9     62   228    392   565  832 3170
q-value      0      0     1     19    96  275 3170
```

R package: sva (3)



R package: sva (4)

- Now we can perform the same analysis, but adjusting for surrogate variables. The first step is to include the surrogate variables in both the null and full models.

```
> svaObj=sva(log2(SVA.exp.preprocessed[,-1:-3]),mod,mod0)
Number of significant surrogate variables is:  4
Iteration (out of 5 ):1  2  3  4  5  >
> modSv=cbind(mod,svaObj$sv[,1])
> mod0Sv=cbind(mod0,svaObj$sv[,1])
```

- Then P-values and Q-values can be computed as before.