Dealing with batch effects with high-throughput genomic data

Chung-Hsing Chen

National Cancer Institute, National Health Research Institutes

Batch effects

- Accuracy of measurements depend on reagents, hardware, highly trained personnel.
- In high-throughput experiments, many of quantities being measured are **simultaneously** affected by both biological and non-biological factors.
- Batch effects are **subgroups** of measurements that have **qualitatively** different behavior across conditions and are **unrelated** to the biological variables in a study.

Relief and Confidence

- Although batch effects are difficult to detect in low-dimensional assay, high-throughput technologies provide the opportunity to detect and remove them.

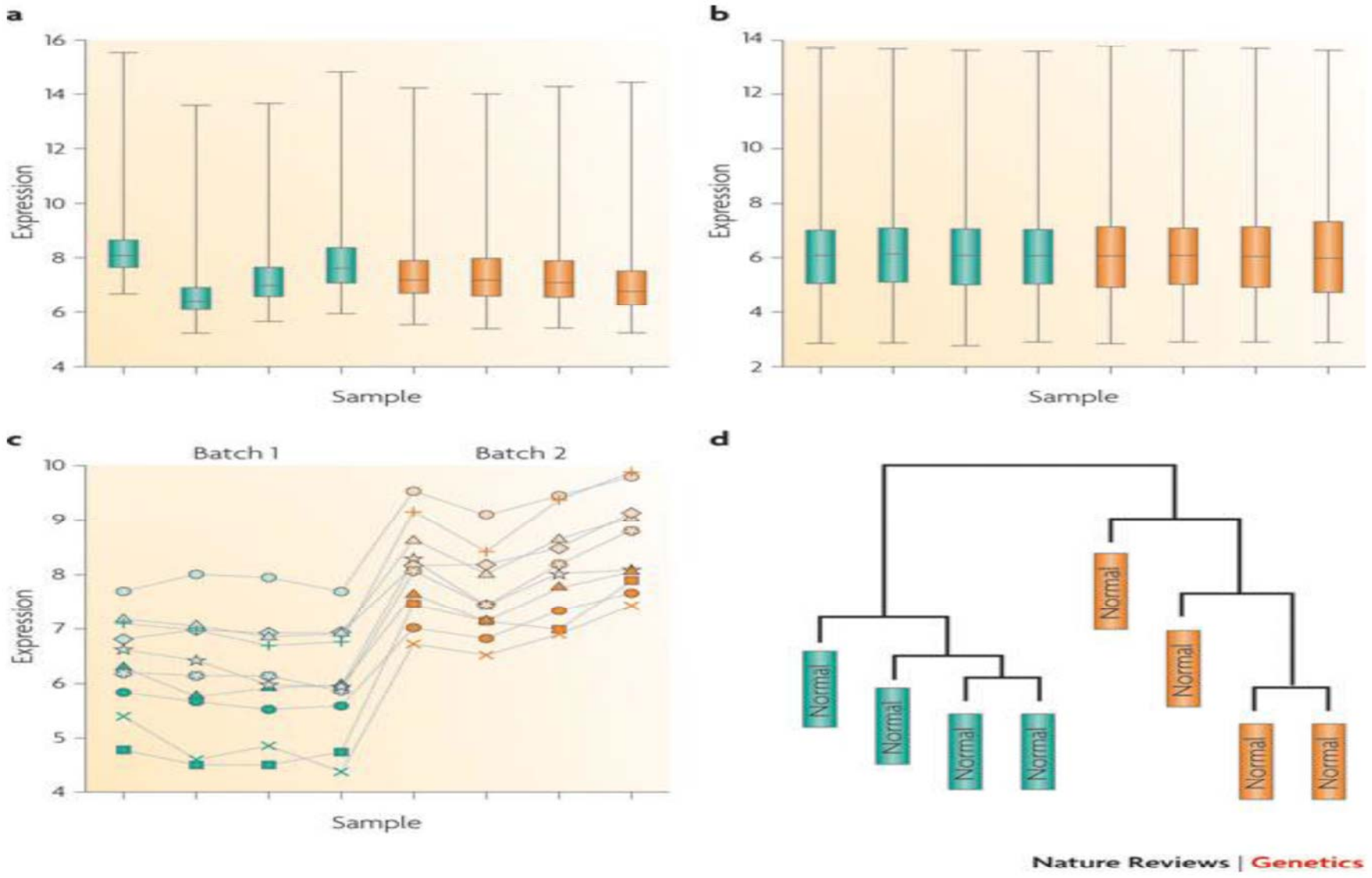
Systematic bias in microarray

- Sample preparation, hybridization, measurement of expression;
- Batch to batch variation in array manufacture;
- Day to day variation in laboratory conditions

Gene expressions correlated with processing date (I)

- Dyrskjot, L. *et al*. Gene expression in the urinary bladder: a common carcinoma *in situ* gene expression signature exists disregarding histopathological classification. *Cancer Res.* **64**, 4040-4048 (2004).
- Zilliox, M. J. & Irizarry, R. A. A gene expression bar code for microarray data. *Nature Methods* **4**, 911-913 (2007).

- Microarray expression profiling was used to examine the gene expression patterns in superficial transitional cell carcinoma(sTCC) with or without surrounding carcinoma *in situ* (CIS).
- **Cluster analysis** based on microarray expression data separated the sTCC samples according to the presence or absence of CIS.
- However, the presence or absence of CIS was strongly ***confounded*** with processing date (Zilliox & Irizarry, 2007).



Nature Reviews Genetics 2010 (11), 733-739.

- For a published bladder cancer microarray data set obtained using an Affymetrix platform, we obtained the raw data for **only** the normal samples. Here, green and orange represent two different processing dates. **a** | Box plot of raw gene expression data (log base 2). **b** | Box plot of data processed with RMA, a widely used preprocessing algorithm for Affymetrix data. RMA applies quantile normalization — a technique that forces the distribution of the raw signal intensities from the microarray data to be the same in all samples. **c** | Example of ten genes that are susceptible to batch effects even after normalization. Hundreds of genes show similar behaviour but, for clarity, are not shown. **d** | Clustering of samples after normalization. Note that the samples perfectly cluster by processing date.

Nature Reviews Genetics 2010 (11), 733-739.

- Normalization helps reduce global differences among arrays, does not address batch effects.
- In gene expression studies, the greatest source of differential expression is nearly always across batches rather than across biological groups.

Gene expressions correlated with processing date (II)

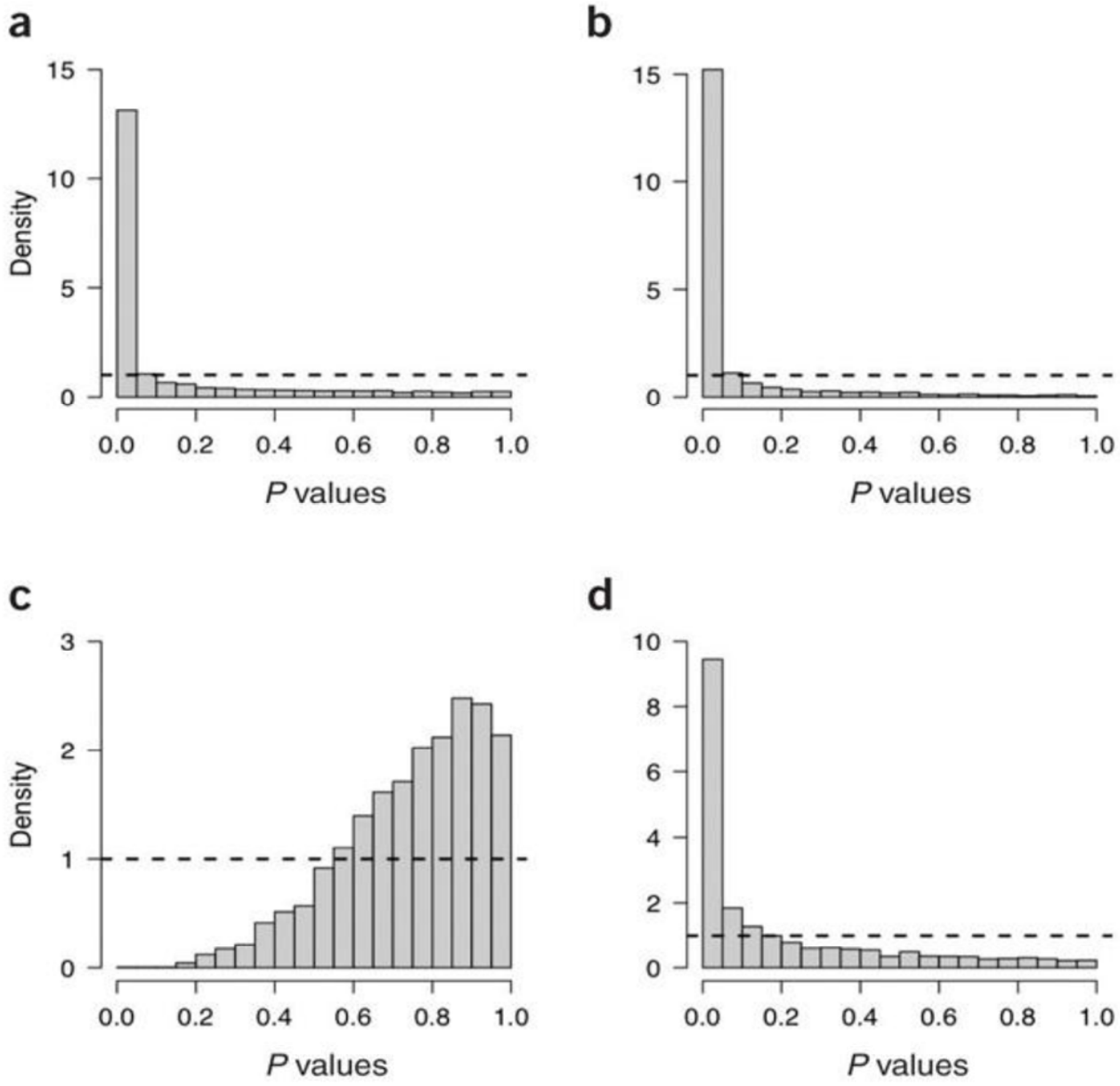
- Spielman, R. S. et al. Common genetic variants account for differences in gene expression among ethnic groups, *Nature Genetics* **39**, 226-231 (2007).
- Akey, J. M. et al. On the design and analysis of gene expression studies in human populations, *Nature Genetics* **39**, 807-809 (2007).

Gene expression, genetic variant, and ethnic group

- Allele frequency differences between populations often have highly significant phenotypic consequences.
- The proportion of gene expression phenotypes differs significantly between populations and to what extent the phenotypic differences are attributable to specific genetic polymorphism.

- Between European-derived and Asian-derived populations, expression phenotypes differs significantly for 25% of 4197 genes at **p-value** less than 10^{-5}, based on cell lines from 60 CEU and 41 CHB and 41 JPT of the **HapMap Project**. (Spielman et al., 2007)
- Storey and coworkers think this is a too stringent criterion. Using the **complete distribution of P-values**, they found the proportion is 78%.

- A possible explanation for the pervasive signature of differential expression observed in Spielman et al. is a systematic bias introduced during microarray expression measurements.
- CEU individuals were primarily processed from 2003 to 2004 and ASN individuals were all in 2005-2006.

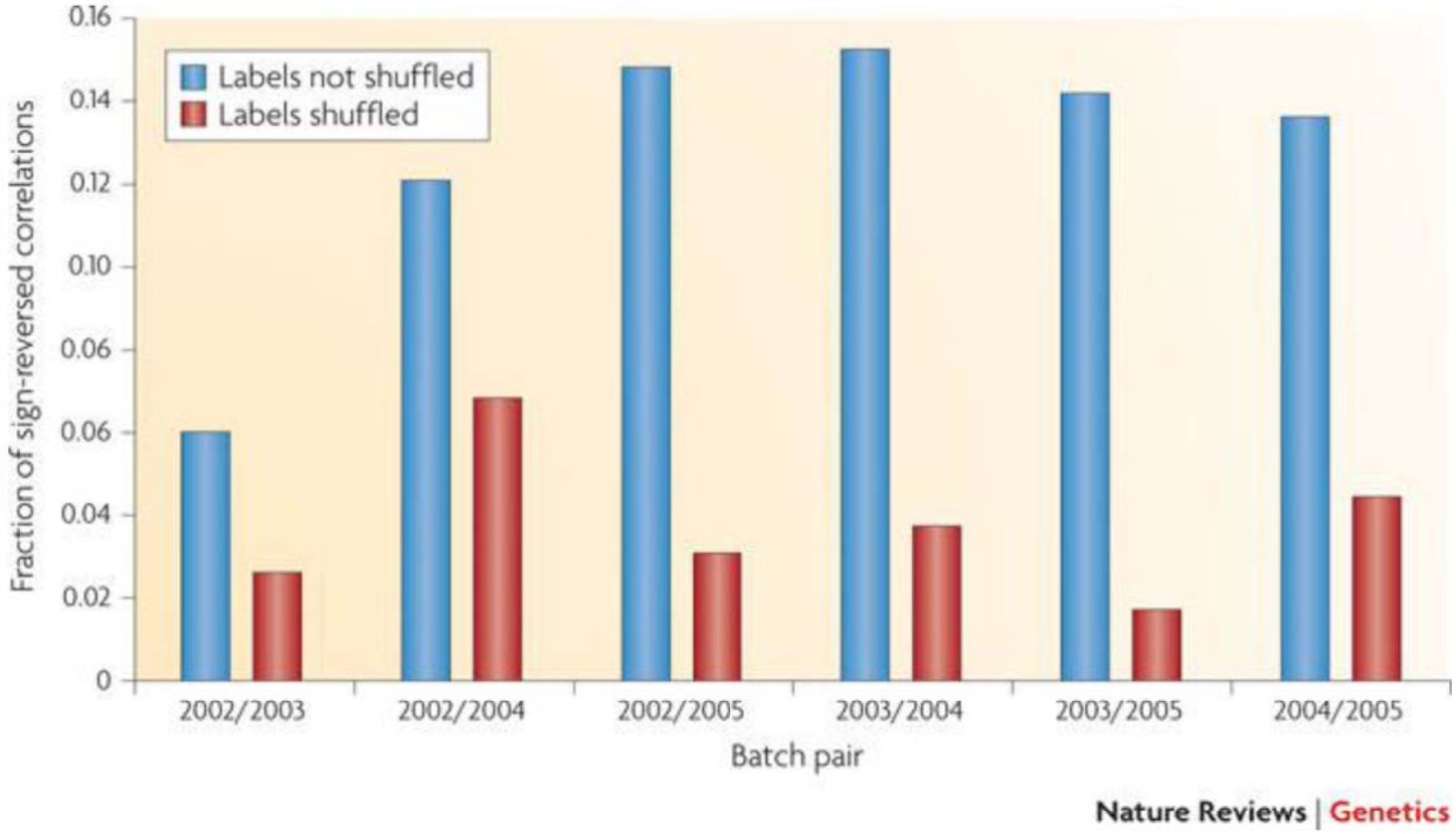


Nature Genetics 2007 (7), 807-809.

- (**a**) *P* values comparing CEU and ASN samples. (**b**) *P* values comparing samples having different microarrays processing year. (**c**) *P* values comparing CEU and ASN samples, controlling for the sample processing year. (**d**) *P* values comparing samples having different microarrays processing year among the CEU individuals. Under the null hypothesis of no differential expression, we expect the *P* values to be uniformly distributed between 0 and 1, forming a histogram with frequencies following the dashed black line. We estimate the proportion of differentially expressed genes in **a**–**d** to be 78%, 94%, 0% and 79%, respectively. The odd shape of the histogram in **c** is attributable to the almost complete confounding of year of processing and population, illustrating the underlying problem with the study design.

Nature Genetics 2007 (7), 807-809.

Batch effects and correlations between genes



Nature Reviews Genetics 2010 (11), 733-739.

- We identified all significant correlations ($p < 0.05$) between pairs of genes within each batch using a linear model. We looked at genes that showed a significant correlation in two batches and counted the fraction of times that the correlation changed between the two batches. A large percentage of significant correlations reversed signs across batches, suggesting that the correlation structure between genes changes substantially across batches. To confirm this phenomenon is due to batch, we repeated the process but with the batch labels randomly permuted. With random batches, a much smaller fraction of significant correlations change signs. This suggests that correlation patterns differ by batch, which would affect rank-based prediction methods as well as system biology approaches that rely on between-gene correlation to estimate pathways.

Nature Reviews Genetics 2010 (11), 733-739.

Some remarks

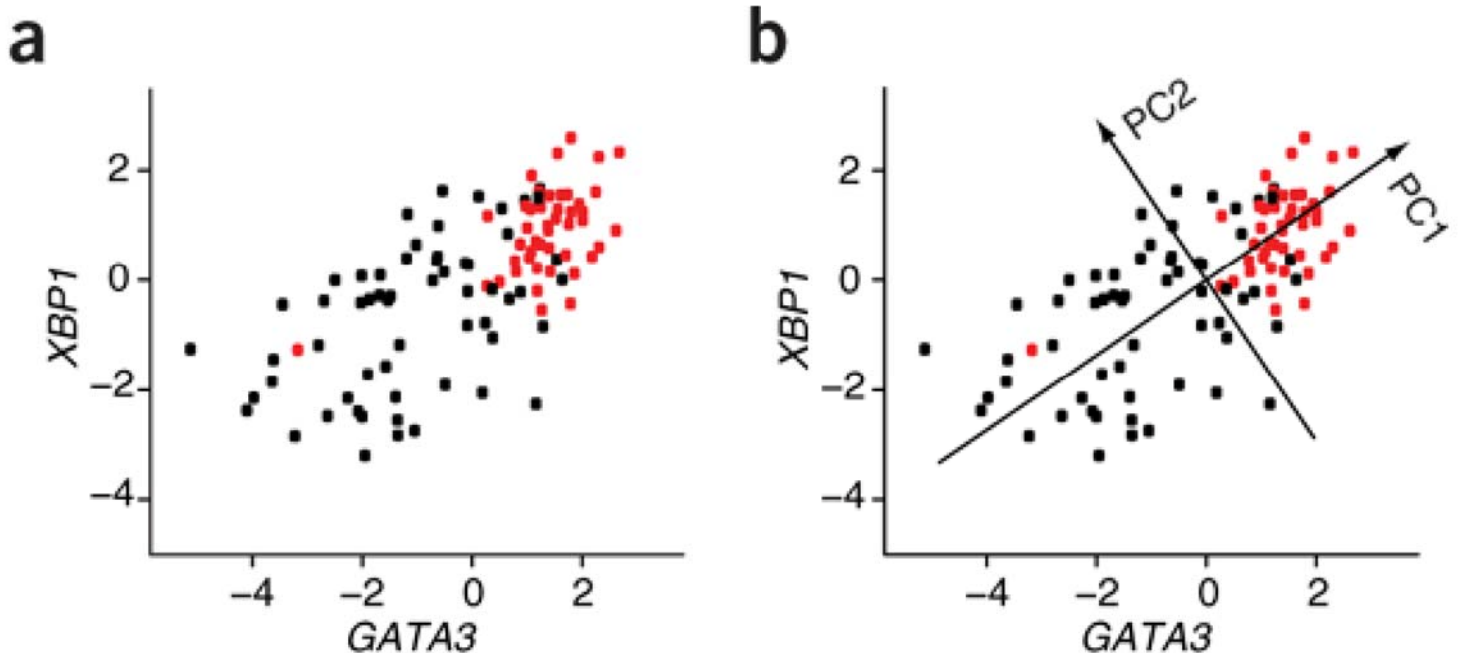
- Batch effects appear quite frequently.
- Try best to avoid batch effects in the first place.

Principal component analysis

- Leek, J. T. and Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by **Surrogate Variable Analysis**, *PLoS genetics* **3**, e161 (2007).
- Reich, D. et al. Principal component analysis of genetic data, *Nature Genetics* **40**, 491-492 (2008).
- Ringnér, M. What is principal component analysis?, *Nature Biotechnology* **26**, 303-304 (2008).

Principal component analysis

- Change the coordinates to best represent the data, singular value decomposition, analytic geometry.
- An important topic in linear algebra and multivariate analysis for **data reduction**.
- Detect the **hidden** population substructure in genetic studies.
- Detect the **unmeasured** batch effects in expression array.



(**a**) Each dot represents a breast cancer sample plotted against its expression levels for two genes. (samples are colored according to estrogen receptor (ER) status: ER+, red; ER-, black.) (**b**) PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread.

Nature Biotechnology 2008 (26), 303-304.

Surrogate Variable Analysis

- SVA accurately estimates the unobserved factor even when there is strong dependence between the primary and unobserved factors, with a subset of genes affected by both.
- SVA results in a more powerful and reproducible ranking of gene for differential expression.
- Improves estimation of the false discovery rate.

Data pre-processing in our lab

- Quality assessment

  Data quality control/check according to wet lab/manufacture guidelines.

- Within-array normalization

  Background correction.

- Between-array normalization

  Adjustment by observed confounding variables, SVA and PCA to check any unobserved confounding variables.

Acknowledgements

- National Institute of Cancer Research, NHRI
  I-Shou Chang, Shih-Sheng Jiang
- Institute of Population Health Sciences, NHRI
  Chao Hsiung, Ying-Hsiang Chen
- Institute for Translational Medicine, TMU
  Wen-Chang Wang
- Capable RAs
- TBI Bioinformatics core supported by NSC

Thank you for your attention