Memorial Sloan Kettering
Cancer Center

Validation and Refinement of Breast Cancer Subtype Classification based on Multi-Modal Genomic Data

July 26, 2017

Margaret L. Hannum

MS Candidate, Department of Biostatistics, Columbia University

Mentors: Ronglai Shen, PhD (MSK) & Shuang Wang, PhD (Columbia)

Presentation Overview

- Background
  - Cancer Subtypes
  - Genomic Data Types
  - METABRIC study
- Goals of project
- Results
- Further analysis

Background

Cancer Subtypes

- Cancer is a heterogeneous disease
  - Example: Breast cancer (Estrogen Receptor (ER), Progesterone (PR), HER2, Basal)
- Subtypes have clinical implications (treatment and survival)
  - Luminal A: ER+, HER2-
  - Luminal B: ER+, either HER2+ or –
  - HER2 type: most are HER2+, ER-, PR-, Lymph-node+
  - Triple Negative Breast Cancer (TNBC)/Basal-like: ER-, PR-, HER2-

Genomic Data Types

- Copy Number Alteration (CNA): on DNA level
  - Affymetrix 6.0 Single Nucleotide Polymorphism (SNP) arrays
- Gene Expression: on mRNA level
  - High-dimensional platforms such as Expression arrays, Next Generation sequencing (NGS)
- DNA Methylation
  - Illumina methylation arrays 27K, 450K, 850K platforms
  - Differentially Methylated Loci (DML) vs. Regions (DMRs)
- CN gains lead to increased expression, hypermethylation of promoter regions leads to under expression
- There are many other genomic data types but these are the three used in this project.

Breast Cancer Subtype Classification

- METABRIC Study (Curtis et al., Nature, 2012)
  - Identified 10 BRCA subtypes by integrating copy number and expression data from 2,000 breast tumor samples
  - Identified new therapeutic targets including 11q13 copy number amplification
  - Identified new subtypes including a copy number quiescent subgroup that show favorable prognosis
  - Validated in a combined external patient cohort of over 7,500 tumor samples (Ali et al., Genome Biology, 2014)

Our Project: Validate and Refine using the TCGA data set

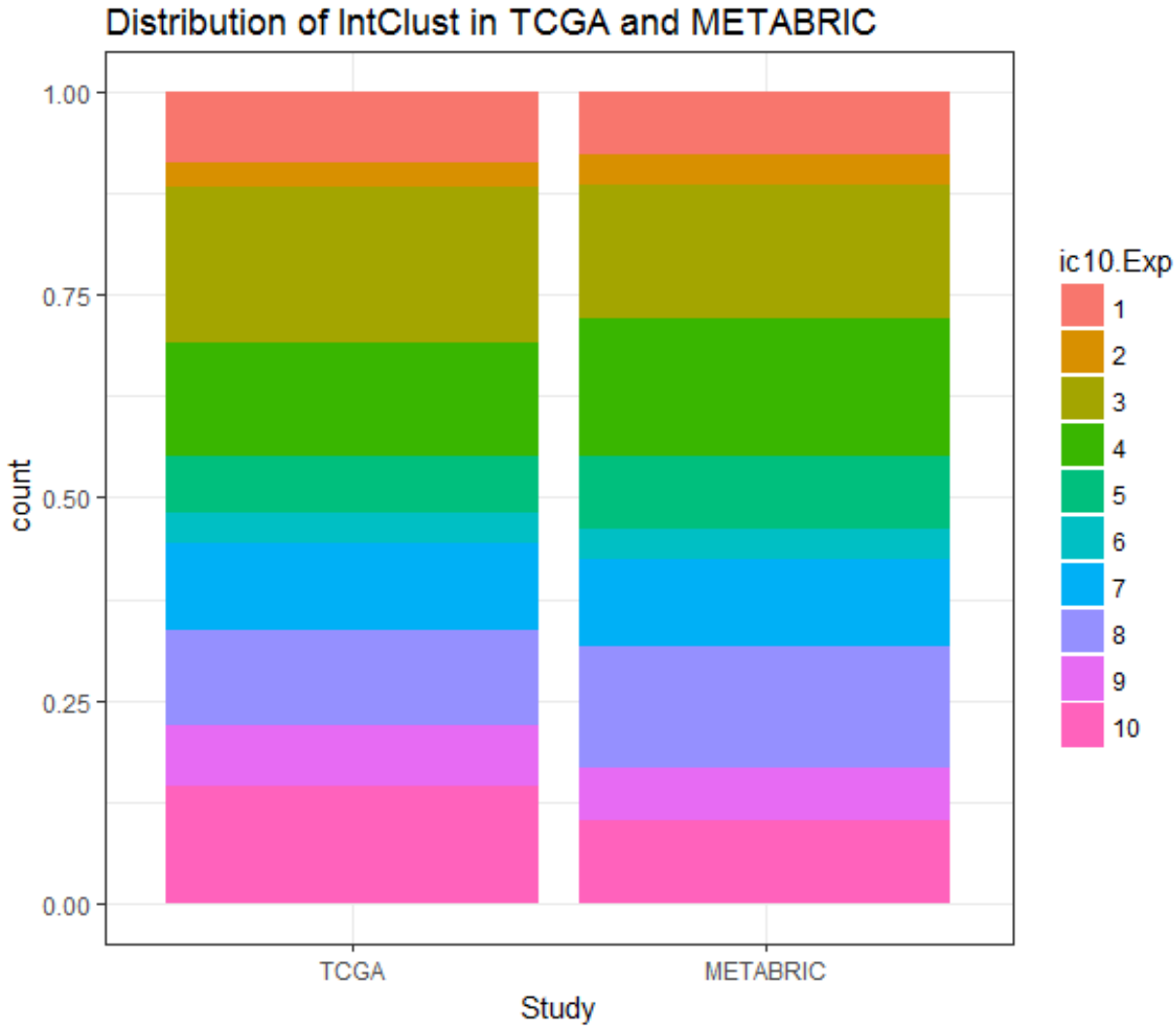
1. *METABRIC Subtype Validation*
2. *DMR Identification from DNA methylation data*
3. *Subtype Refinement by incorporating DMRs*

1. Validation

- *Goal: Validate 10 subtypes from METABRIC study using an independent cohort from The Cancer Genome Atlas (TCGA)*

METABRIC IntClust Validation

- All 10 subtypes were identified. Although frequency is different between two cohorts, this is not unexpected since they are drawn from different patient populations.



Memorial Sloan Kettering Cancer Center

METABRIC IntClust Validation

- Goodness of Fit within TCGA clusters was very promising (further analysis with silhouette and In-group proportions is needed to confirm)

Correlation between Centroids and Predicted Profiles

Cluster	Correlation
ic1	0.924
ic2	0.947
ic3	0.923
ic4	0.699
ic5	0.957
ic6	0.972
ic7	0.94
ic8	0.984
ic9	0.090
ic10	0.915
Overall Correlation	0.955

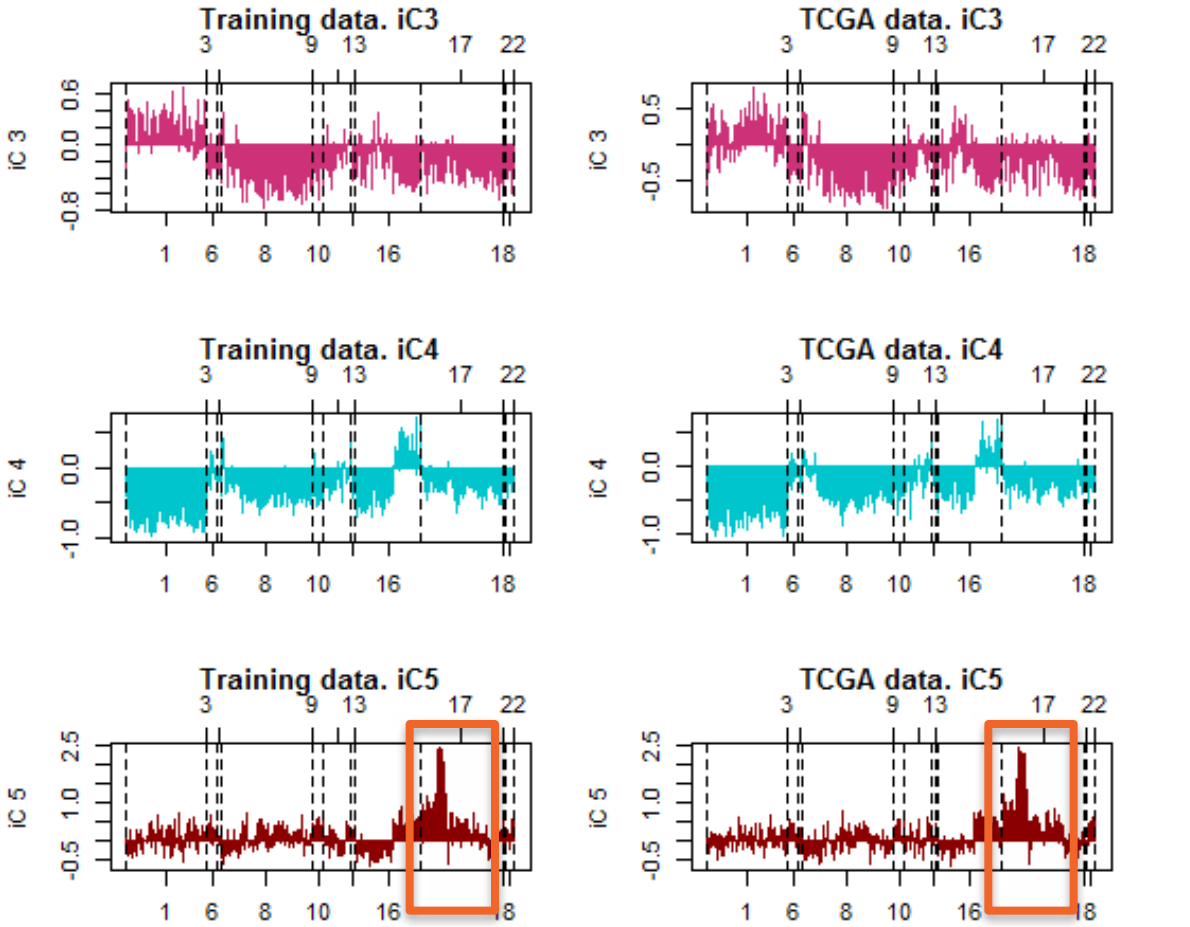
METABRIC IntClust Validation

- Comparison of copy number profiles was also a nice confirmation of consistent clusters
- For example, 17 Q amplification in both METABRIC and TCGA in IntClust5, which was a cluster highly defined by HER2 amplified intrinsic subtypes.

Copy Number Profiles

METABRIC Discovery Cohort

TCGA Validation Cohort



2. DMR Identification

- *Goal: Identify Differentially Methylated Regions (DMRs) using TCGA methylation data, to reduce dimensionality of methylation data, and signals maybe more reliable than DML*

DMR Identification Method

- Use thresholds to identify sites of methylation, then identify regions using DMRcate algorithm (Peters, et al. 2015)
- After cleaning and quality control, used algorithm on 326K probes in 90 matched tumor samples
- Identified over 6000 significant DMRs (FDR <0.05) in the TCGA sample for use in next step of subtype refinement

3. Data Integration and Subtype Refinement

- *Goal: Identify novel breast cancer subtypes on top of 10 METABRIC clusters by integrating copy number, gene expression, and methylation data*

Integrative Clustering with multiple genomic datatypes

- Use iCluster method (Shen, et al. 2012), a joint latent variable model
  - In this case, subtypes are considered latent (unknown)
  - We will use measurable data (multimodal genomic: copy number, gene expression, and methylation) to uncover subtypes
  - Analysis in progress

Next Steps

- Complete iCluster analysis
- Perform further statistical tests to assess goodness of fit of IntClust classification validation (silhouette and in-group proportion)
- Compare demographics b/w METABRIC and TCGA along with PAM50 classification

Questions?

- Thank you:
  - Ronglai Shen, MSK
  - Shuang Wang, Columbia
  - Ya Wang, Columbia
  - Arshi Arora, MSK
  - Esther Drill, MSK
  - Ariel Chernofsky, MSK
  - …and Shireen Lewis and Cynthia Berry for organizing the intern program

References

Ali HR, Rueda OM, Chin SF, Curtis C, Dunning MJ, Aparicio SA, et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biol. 2014;15(8):431. pmid:25164602

Aryee MJ, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics (Oxford, Engl). 2014;30(10):1363-9.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B. 1995;57(1):289-300.

Brenton JD, Tavare S, Caldas C, et al: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012, 486: 346-352.

Butcher LM, Beck S. Probe Lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data. Methods. 2015;72:21-8.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL,

Gao Y, Jones A, Fasching PA, et al. The integrative epigenomic-transcriptomic landscape of ER positive breast cancer *Clinical Epigenetics*. 2015;7:126. doi:10.1186/s13148-015-0159-0.

Holm K, Staaf J, Lauss M, et al. An integrated genomics analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to chromatin states in normal mammary cells *Breast Cancer Research : BCR*. 2016;18:27. doi:10.1186/s13058-016-0685-5.

iC10: a copy number and expression-based classifier for breast tumors. Version 1.1.3. Comprehensive R Archive Network; 2015 Sep 23 [cited 2016 Nov 8]. Available from: https://CRAN.R-project.org/package=iC10.

Jaffe AE, Murakami P, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012;41(1):200-9.

Kapp, A. V. & Tibshirani, R. [Are clusters found in one dataset present in another dataset?](https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxj029) Biostatistics 8, 9-31 (2007).

Michaut M, Chin S-F, Majewski I, et al. Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Scientific Reports*. 2016;6:18517. doi:10.1038/srep18517.

Morris TJ, et al. ChAMP: 450k chip analysis methylation pipeline. Bioinformatics. 2014;30:428-30.

Peters TJ, Buckley MJ, Statham AL, et al. *De novo* identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin*. 2015;8:6. doi:10.1186/1756-8935-8-6.

Teschendorff AE, Marabita F, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. Bioinformatics (Oxford, Engl.). 2013a;29(2):189-96.

TCGA Network: Comprehensive molecular portraits of human breast tumors. Nature 2012, 490:61-70.