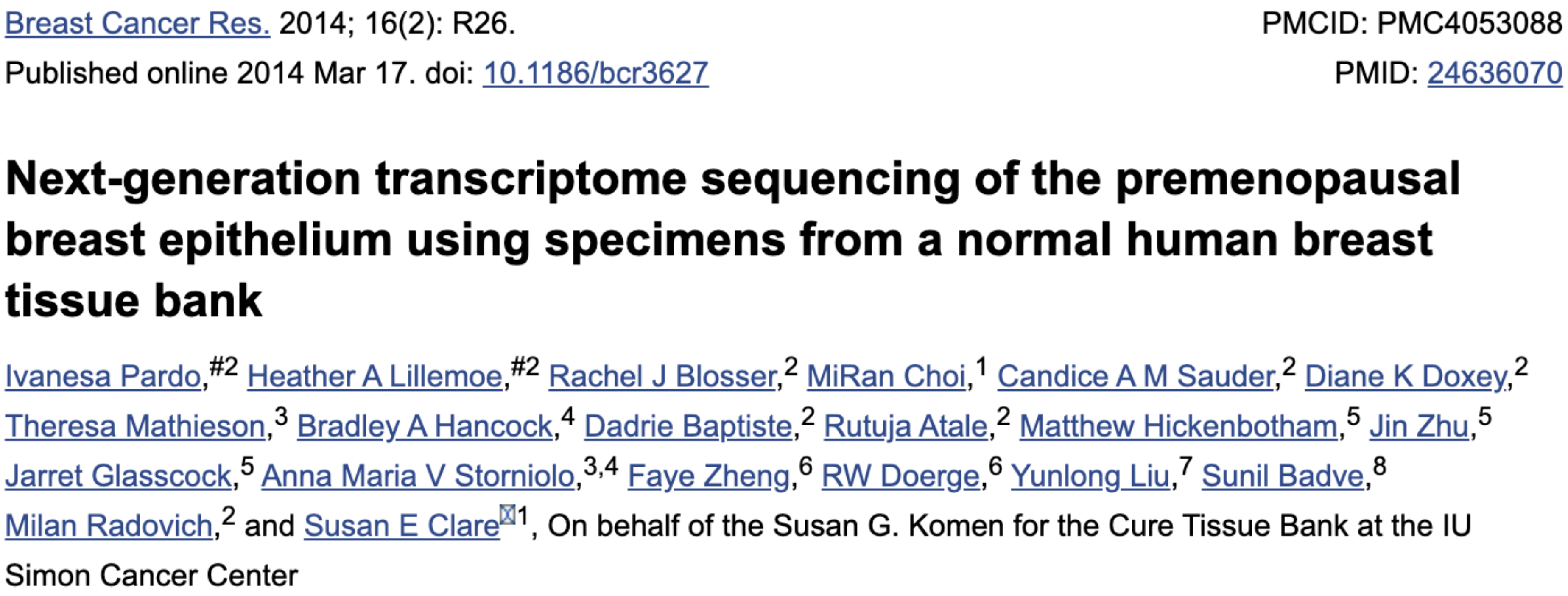
Menstrual cycle dataset

Dataset number 5

Background

- We will work with the dataset from this link https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4053088/



Breast Cancer Res. 2014; 16(2): R26.

Published online 2014 Mar 17. doi: 10.1186/bcr3627

PMCID: PMC4053088

PMID: 24636070

Next-generation transcriptome sequencing of the premenopausal breast epithelium using specimens from a normal human breast tissue bank

Ivanesa Pardo,[#2] Heather A Lillemoe,[#2] Rachel J Blosser,[2] MiRan Choi,[1] Candice A M Sauder,[2] Diane K Doxey,[2] Theresa Mathieson,[3] Bradley A Hancock,[4] Dadrie Baptiste,[2] Rutuja Atale,[2] Matthew Hickenbotham,[5] Jin Zhu,[5] Jarret Glasscock,[5] Anna Maria V Storniolo,[3,4] Faye Zheng,[6] RW Doerge,[6] Yunlong Liu,[7] Sunil Badve,[8] Milan Radovich,[2] and Susan E Clare☒[1], On behalf of the Susan G. Komen for the Cure Tissue Bank at the IU Simon Cancer Center

Background

- The normal human breast is under the influence of many endogenous (from the body) hormones as well as exogenous hormones (for instance from contraception)
- Using normal (healthy) breast tissue from 20 premenopausal donors, the changes in the mRNA of the normal breast epithelium was studied.
- Women answered a full questionary at the day of the biopsy, such as menstrual cycle day and hormonal contraception with information of the type of contraception.
- Next-generation whole transcriptome sequencing (RNA-Seq) was used to study mRNA expression in those biopsies.

Dataset and exercise

- 20 biopsies from pre-menopausal women with indication of phase of the menstrual cycle (follicular or luteal phase) or hormonal contraception (corresponding to 9, 5 and 6 samples respectively)
- Download the data from the website (or load the files from the course page) (EXCEL sheet with tabs corresponding to meta-data and tabs corresponding to raw expression)
- Observe the data, is it numeric ?
- Prepare the table, remove genes with sum of the row less or equal to 20.

Exercise

- Select only patients that are not taking contraceptives (L=Luteal and F=Follicular, not HC=Hormone Contraceptives)
- Perform a dimension reduction using principal component analysis.
- What do you observe?
- Sequencing depth might be different from sample to sample. Convert the data to log counts per million (this is recommanded for sequencing data)
- Do a loop of t.test to find which genes are significant between Luteal and Follicular phase.
- How many significant genes ? Did you adjust for multiple testing ?

Exercise

- Again perform a dimension reduction using principal component analysis.
- What do you observe now ? Are there any clear groups?
- Check inside the paper, in the analysis of the sequencing. Were there any batches?
- Include batches into a linear model to infer differential expression between menstrual phase groups (L=Luteal, F= Follicular)
- What do you observe?

Exercise

- Bonus: Find out which genes are correlated with Estradiol concentration in blood.