

OV_Histotypes: Report of Statistical Findings

Derek Chiu

2019-10-08

Contents

Preface	4
1 Introduction	5
2 Methods	6
2.1 Data Processing	6
2.2 Housekeeping Genes	6
2.3 Common Samples and Genes	7
3 Results	8
References	8

List of Tables

List of Figures

Preface

This report of statistical findings describes the classification of ovarian cancer histotypes using data from NanoString CodeSets.

Marina Pavanello conducted the initial exploratory data analysis, Cathy Tang implemented class imbalance techniques, Derek Chiu conducted the normalization and statistical analysis, and Aline Talhouk lead the project.

1. Introduction

Ovarian cancer has five major histotypes: high-grade serous carcinoma (HGSC), low-grade serous carcinoma (LGSC), endometrioid carcinoma (ENOC), mucinous carcinoma (MUC), and clear cell carcinoma (CCOC). A common problem with classifying these histotypes is that there is a class imbalance issue. HGSC dominates the distribution, commonly accounting for 70% of cases in many patient cohorts, while the other four histotypes are spread over the rest.

In the NanoString CodeSets, we also run into a problem with trying to find suitable control pools to normalize the gene expression. For prospective NanoString runs, the pools can be specifically chosen, but for retrospective runs, we have to utilize a combination of common samples and common genes as references for normalization.

The supervised learning is performed under a consensus framework: we consider various classification algorithms and use evaluation metrics to help making decisions of which methods to carry forward.

2. Methods

2.1 Data Processing

There are 3 NanoString CodeSets:

- CS 1: OvCa2103_C953
 - Samples = 412
 - Genes = 275
- CS 2: PrOTYPE2_v2_C1645
 - Samples = 1223
 - Genes = 384
- CS 3: OTTA2014_C2822
 - Samples = 5424
 - Genes = 532

These datasets contain raw counts extracted straight from NanoString RCC files.

2.2 Housekeeping Genes

The first normalization step is to normalize all endogenous genes to housekeeping genes (POLR1B, SDHA, PGK1, ACTB, RPL19; reference genes expressed in all cells). We normalize by subtracting the average log2 housekeeping gene expression from the log2 endogenous gene expression. The updated CodeSet dimensions are now:

- CS 1: OvCa2103_C953
 - Samples = 412
 - Genes = 256
- CS 2: PrOTYPE2_v2_C1645
 - Samples = 1223
 - Genes = 365
- CS 3: OTTA2014_C2822
 - Samples = 5424
 - Genes = 513

The number of genes are reduced by 19: 5 housekeeping, 8 negative, 6 positive (the latter 2 types are not used).

2.3 Common Samples and Genes

Since the reference pool samples only exist in CS2 and CS3, we need to find an alternative method to normalize all three CodeSets. One method is to select common samples and common genes that exist in all three. We found 72 common genes. Using the `summaryID` identifier, we found 78 common summary IDs, which translated to 320 samples. The number of samples that were found in each CodeSet differed:

- CS 1: OvCa2103__C953
 - Samples = 93
 - Genes = 72
- CS 2: PrOTYPE2__v2__C1645
 - Samples = 87
 - Genes = 72
- CS 3: OTTA2014__C2822
 - Samples = 140
 - Genes = 72

3. Results