Genomics 106 (2015) 96–106

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Personalized diagnosis of medulloblastoma subtypes across patients and model systems

Deena Mohamad Ameen Gendoo [a,b,*], Petr Smirnov [a], Mathieu Lupien [a,b,c], Benjamin Haibe-Kains [a,b,*]

[a] Bioinformatics and Computational Genomics Laboratory, Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada
[b] Medical Biophysics Department, University of Toronto, Toronto, Ontario, Canada
[c] Ontario Institute for Cancer Research, Toronto, Ontario M5G 1L7, Canada

ARTICLE INFO

Article history:
Received 4 May 2015
Accepted 6 May 2015
Available online 12 May 2015

Keywords:
Cancer
Medulloblastoma
Classification
Subtypes
Human
Mouse model

ABSTRACT

Molecular subtyping is instrumental towards selection of model systems for fundamental research in tumor pathogenesis, and clinical patient assessment. Medulloblastoma (MB) is a highly heterogeneous, malignant brain tumor that is the most common cause of cancer-related deaths in children. Current MB classification schemes require large sample sizes, and standard reference samples, for subtype predictions. Such approaches are impractical in clinical settings with limited tumor biopsies, and unsuitable for model system predictions where standard reference samples are unavailable. Our developed *Medullo-Model To Subtype* (MM2S) classifier stratifies single MB gene expression profiles without reference samples or replicates. Our pathway-centric approach facilitates subtype predictions of patient samples, and model systems including cell lines and mouse models. MM2S demonstrates >96% accuracy for patients of well-characterized normal cerebellum, WNT, or SHH subtypes, and the less-characterized Group 4 (86%) and Group 3 (78.2%). MM2S also enables classification of MB cell lines and mouse models into their human counterparts.

1. Introduction

Medulloblastoma (MB) subtype stratification has evolved with the increased availability of genomic data and improved understanding of MB inter-tumor heterogeneity, signalling pathways, and molecular pathogenesis mechanisms [1–5]. Four MB subtypes, referred to as the WNT, SHH, Group 3, and Group 4, are recognized. These differ based on histopathology, epidemiology, prognosis, and genomic profiles [4,5]. Of these, Group 3 and Group 4 exhibit poor prognosis and are poorly characterized at the molecular level, thus presenting ongoing clinical challenges [1–7]. Ongoing studies to determine subtype-specific biological mechanisms currently involve extensive gene expression profiling and sequencing of MB subtypes, to detect recurrent mutations, novel SNVs, CNAs, and CpG methylation sites [7–9]. These efforts were instrumental to improving MB subtype classification but relied on large data sets, a requirement not easily attainable in future classification studies where limited sample sizes are available. This concern instigated development of several MB prediction assays in a clinical setting [4,10,11]. These include identification of a set of 22 subtype-specific signature genes using nanoString nCounter technology to measure mRNA expression from FFPE [11], as well as development of a 13-gene multiplex mRNA expression signature specific for the WNT and SHH subtypes [10]. Unfortunately, these signatures have not been further developed into automated classifiers for use by researchers and clinicians. To use these signatures, new MB samples would need to be interrogated by hierarchical cluster analysis or other orthogonal methods, which requires additional external samples to generate comparisons. Accordingly, the proposed classification practices have provided a systems-wide view of subtype tumorigenesis, but remain deficient in providing individualized predictions of particular MB samples.

The need for personalized predictors of MB samples is not only necessary for prediction of patient samples, but also for predictions of samples pertaining to model systems. These latter predictions are important for future research that utilizes MB cell lines and mouse models to study MB subtype disease origins and signaling pathways [12]. Applicability of current classification schemes towards model predictions is unclear, especially as the reference 'gold standard' to compare model systems to human subtypes remains highly ambiguous. The developed Agreement of Differential Expression (AGDEX) algorithm attempted to address this problem by comparing expression of orthologous genes to determine transcriptomic similarities between tissues from different experiments [13–15]. However, AGDEX relies on *a priori* assumptions regarding

Abbreviations: MB, medulloblastoma; MM2S, Medullo-Model To Subtype; MH2H, medullo-human to human; MM2H, medullo-mouse to human; IHC, immunohistochemistry; FFPE, formalin-fixed paraffin embedded; GTML, Glt1-tTA (glutamate transporter 1-tetracycline transactivator) and tetracycline response element (TRE)-MYCN/luciferase (Luc) mouse medulloblastoma; MCC, Matthews correlation coefficient; GEO, Gene Expression Omnibus; KNN, *k*-nearest neighbor; PCA, principal component analysis.

* Corresponding authors at: Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada.
E-mail addresses: deena.gendoo@utoronto.ca (D.M.A. Gendoo), bhaibeka@uhnresearch.ca (B. Haibe-Kains).

http://dx.doi.org/10.1016/j.ygeno.2015.05.002
0888-7543/

which samples can be compared in subtype differential expression analyses, which introduces user bias and ultimately prevents robust classification of single tumor samples. Additionally, the algorithm also necessitates a user-selected reference model, perhaps human or mouse normal cerebellum, to generate differential analysis comparisons. Any *a priori* assumptions of reference samples may yield imprecise comparisons with the tested data sets and is doubly problematic when matched reference samples to the tested data are unavailable.

Collectively, current classification methods remain impractical for both research and clinical settings, where reference samples are unattainable, or limited samples or tumor materials are available. They fail to address the critical requirement of generating *personalized, single-sample* predictions for both MB patients and samples pertaining to model systems. To address this focal and much-needed research direction in MB, we developed a novel, *Medullo-Model To Subtype* (MM2S) classifier that matches individual MB samples against human medulloblastoma subtypes. To the best of our knowledge, MM2S is the first *single-sample* classifier of MB samples, which does not rely on a reference sample or multiple sample replicates to generate predictions. We developed a systems-based methodology that facilitates application of the MM2S algorithm to MB subtype prediction of both patient samples and model systems, including cell lines and mouse models. We demonstrate the efficacy and versatility of MM2S via the largest MB stratification analysis of 23 publicly gene expression data sets, spanning 754 patients, 26 cell lines, and 261 mouse samples. We discuss the implications of MM2S towards narrowing the gap between MB subtype classification methods and the development of singular, subtype-specific diagnosis of patients, cell lines, and mouse models.

2. Results

2.1. MM2S accuracy on human medulloblastoma training samples

We trained the MM2S classifier on the 347 human samples from three data sets and pre-validated its accuracy in correctly predicting

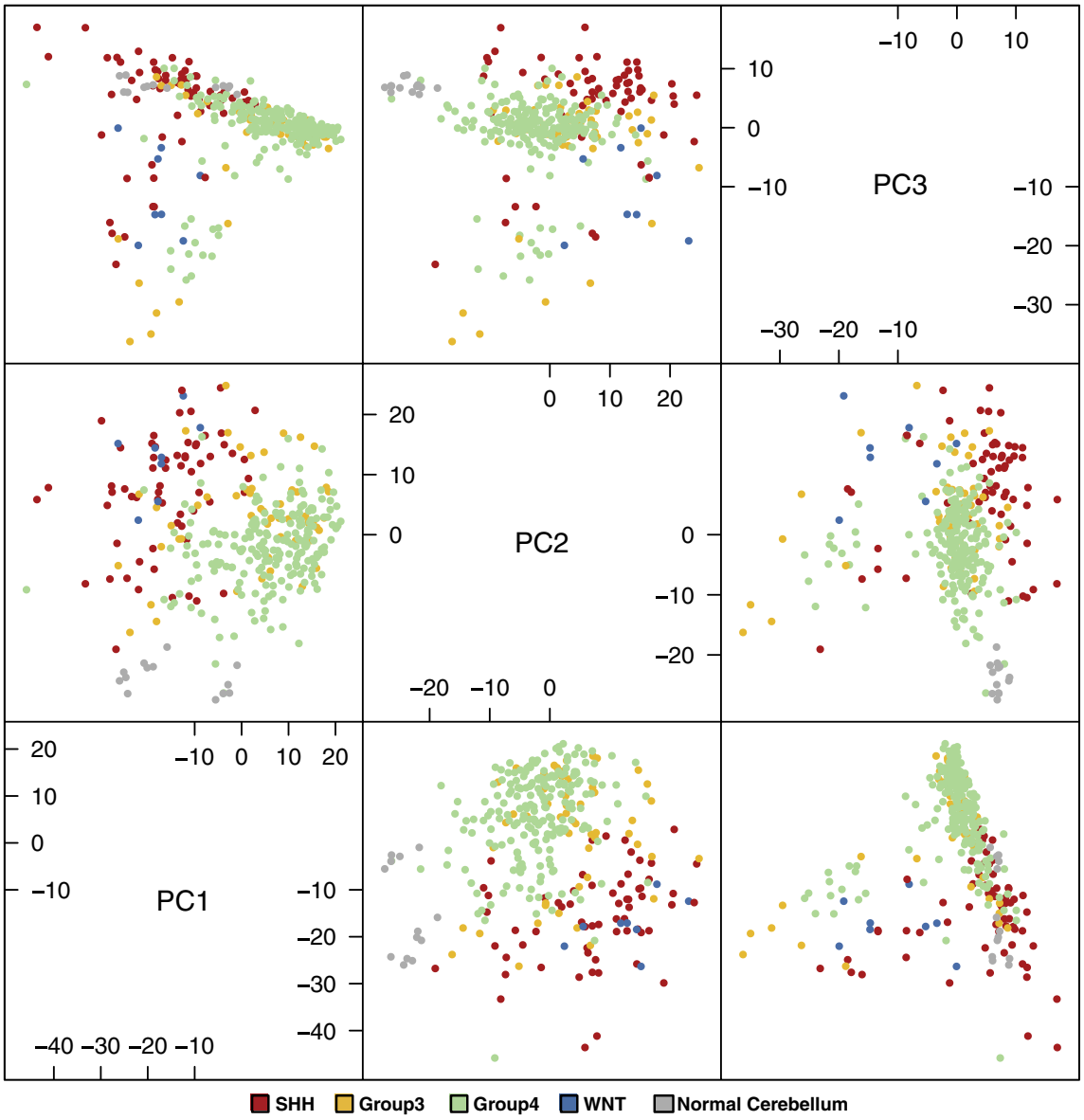

Fig. 1. Principal component analysis of 674 ssGSEA-ranked gene sets (rank matrix) for the human training data set. This is a principal component analysis (PCA) of the human training set, prior to feature selection. Shown is a lattice plot of the first three principal components, with principal components axes rendered across the diagonal. PC1-PC2, PC1-PC3, and PC2-PC3 represent the three plots above the diagonal and are mirrored in the three plots below the diagonal. Samples are colored by subtype.