Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

COMS W4761 Project Report

Systematic Analysis of Tumor Purity

Kaiyi Zhu[1,*]

[1]Department of Electrical Engineering, Columbia University, New York, NY 10027.

*To whom correspondence should be addressed.

Instructor: Itsik Pe'er

Received on 9 May 2016

Abstract

Motivation: Tumors have high heterogeneity, resulting in a mixture of different cell types, which have diverse response characteristics of clinical therapy. Recent advances in high-throughput sequencing technology make genomic information of thousands of tumors available. Consequently, many computational methods were reported with the aim for estimating tumor purity, i.e., the fraction of cancerous cells in the admixture. Therefore, it is likely that a systematic analysis of tumor purity estimating methods across multiple cancer types may lead to a bunch of meaningful and interesting discoveries.
Results: This study presents a systematic analysis of four tumor-purity-estimating methods on about 7500 patients across 15 cancer types from the Cancer Genome Atlas. Concordance and correlation of different methods were observed but with diverse patterns in different cancer types. Clinical and prognosis analysis were conducted in order to find prominent features associated with purity level. A new way for evaluating estimating performance was proposed, which was able to select the optimal method for each specific cancer type.
Availability: https://github.com/zky0708/COMSW4761_project.git
Contact: kz2232@columbia.edu
Supplementary information: Supplementary data are available in the same submission folder.

1 Introduction

Tumors are not uniform, but consist of many different cell types, including normal tissue microenvironment, infiltrating immune cells, stromal cells, tumor cells, (Albini and Sporn, 2007) etc. The heterogeneity of cancer can be explained by the cancer evolutionary theory that tumors evolve from a single cell of origin by acquiring genetic variability along with sequential selection (Nowell, 1976). These different types of cells are also called subclones (Alizadeh et al., 2015), the unique characteristics of which underlie different response to clinical therapy and correspondingly pose a problem for diagnosis and therapy. Therefore, understanding tumor heterogeneity becomes increasingly critical for designing and selecting effective and long-term treatment, especially for personalized precision medicine.

Tumor purity is one measurement of this heterogeneity property, which is defined to be the proportion of cancerous cells in the tumor admixture. Traditionally, it was estimated by pathological review of biospecimen slides extracted from tumor samples. Due to rapid advances in high-throughput sequencing technologies in recent years, large amounts of genomic data have become available, and many computational approaches have been devised to estimate tumor purity. These methods are based on different data types of genomic information, such as gene expression profiles (Yoshihara et al., 2013), copy number abbreviation (Carter et al., 2012; Oesper et al., 2013), somatic mutations (Nik-Zainal et al., 2012; Roth et al., 2014; Jiao et al., 2014), and so on.

The DREAM Challenges are series of crowdsourcing competitions, powered by Sage Bionetworks, aimed to examine questions in biological science and human health in open science community. ICGC-TCGA DREAM Somatic Mutation Calling Tumor Heterogeneity Challenge (ICGC, 2010; TCGA, 2016; ICGC-TCGA-DREAM SMC-Het Challenge, 2016) is one of the open challenges, the first subchallenge of which also includes the task to identify the purity of tumor sample.

The Cancer Genome Atlas (TCGA) is currently the largest tumor profiles collection to be analyzed for the key genomic and molecular characteristics of cancer. To date, this project includes 11,000 patients across 33 cancer types. PanCanAtlas was launched in order to expand the original Pan-Cancer project by analyzing data for all TCGA tumor types, providing additional data for researchers to discover more complex and interesting relationships across cancer types (TCGA: The next stage, 2015).

This project compares and evaluates four methods that estimate tumor purity using different data types of inputs on pan-cancer datasets of TCGA. Clinical associations with purity levels were examined, the results of which were then used for choosing the best purity-estimating method for each cancer type.

2 Methods

2.1 Data Sets

All the datasets except for biospecimen slides information come from the TCGA PancanAtlas Data Freeze 1.3.1 (last updated on 21 March 2016) with synapse ID: syn3241074. I downloaded the level 3 normalized RNA-Seq expression profiles (IlluminaHiSeq_RNASeqV2), clinical data and merged mutation annotation format (MAF) files. The slides files were obtained directly from TCGA data portal (contained in data type "Complete Clinical Set", downloaded on April 8th, 2016, https://tcga-data.nci.nih.gov/tcga/findArchives.htm). The tumor types selected in this project were those with more than 100 samples in all data types (i.e., gene expression, copy number abbreviation, somatic mutations and biospecimen slides review). Only solid cancers were considered, and matched normal samples were excluded. In total, this project analyzed 7454 samples across 15 cancer types (Table 1).

Table 1. Summary of TCGA samples: sample size of each data type in each cancer type that were analyzed in the project.

Cancer Types	ESTIMATE (geneExpr)	ABSOLUTE (CNA)	DPC (mutation)	Pathological Review
BLCA	408	144	395	412
BRCA	1085	940	976	1101
COAD	278	396	217	457
HNSC	515	501	508	526
KIRP	285	155	161	284
LGG	514	418	513	512
LIHC	368	187	198	377
LUAD	510	429	542	505
LUSC	487	435	178	474
PCPG	178	138	178	179
PRAD	494	353	425	483
SKCM	469	257	367	471
STAD	401	330	428	471
THCA	501	184	402	506
UCEC	174	415	248	548

BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; COAD, colon adenocarcinoma; HNSC, head and neck squamous cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma.

2.2 Purity Estimation

ESTIMATE purity values were calculated using ESTIMATE R package. The inputs of ESTIMATE are the gene expression data after Log2(· +1) and quantile normalization (using R Bioconductor 'preprocessCore' package). Since the segmented allelic data are not publicly available, ABSOLUTE purity levels were extracted from 'ABSOLUTE purity/ploidy data' (syn3242754) provided by TCGA PanCanAtlas Data Freeze 1.3.1. DPC tool was downloaded from the GitHub page of Sage-Bionetworks DREAM SMC-Het Challenge (SMC-Het-Challenge GitHub, 2016). Tumor purity via pathological review was determined as the average percent of tumor cells of the top and bottom biospecimen slides. Consensus purity level was defined as the median of the measurements from the four methods.

2.3 Clinical Association Test

One-way analysis of variance (ANOVA) and Spearman correlation were applied on categorical and continuous clinical features, respectively. For survival analysis, the Cox proportional hazard model and Kaplan-Meier analyses were conducted with the R Bioconductor 'survival' package. If the effective sample size was no less than 100, corresponding P values were computed subsequently, and multiple hypothesis correction was performed. However, no sample-size estimates were performed to ensure adequate power, the threshold of sample size mentioned above was chosen arbitrarily. Each test was repeated on each of the five purity levels.

3 Results

3.1 Comparison of purity estimates from the four methods

In this project, nearly 7500 tumor samples across 15 cancer types were analyzed. I obtained gene expression (RNA-SeqV2) and somatic mutation profiles from PanCanAtlas Data Freeze 1.3.1, while downloaded biospecimen slides information from TCGA data portal. The four tumor purity estimating methods selected for comparison and evaluation are: ESTIMATE, which uses gene expression signatures of infiltrating immune cells and stromal cells (Yoshihara et al., 2013); ABSOLUTE, whose input is somatic copy-number alterations (Carter et al., 2011); DPC, one of example tools in SMC-Het Challenge inferring the percentage of cancer cells by clustering somatic mutations (SMC-Het-Challenge GitHub, 2016), as well as the pathological review.

I used the median of all methods as a consensus measurement for each sample. Purity levels differ by cancer types. For instance, low purity (< 70%) shows in adenocarcinoma, which indicates a malignancy in the epithelial cells lining of glandular tissues, such as lung (LUAD), prostate (PRAD) and stomach (STAD). All the five estimates of each sample are listed in Supplementary Data 1.

I calculated pairwise Pearson correlations of estimates from the four methods (Fig. 1). I found that the gene expression, copy number abbreviation, and somatic mutation-based methods showed high concordance; however, the correlation of the three genomic-based methods with the results of pathological review were lower in all cancer types.

In order to exploit more detailed relationship, I also drew the scatterplots of each pair of comparisons for each cancer type (Supplementary Fig. 1), which presented a universal pattern in all cancer types that method ESTIMATE always overestimates the tumor purity when compared with the other three methods. This limitation results from the fact that ESTIMATE only considers the levels of stromal and immune cells in tumor tissue when inferring tumor purity. Nevertheless, there are many other kinds of non-cancerous cells consisting of the tumor microenvironment, such as fibroblasts, endothelial cells and epithelial cells (Joyce and Pollard, 2009). The varying extents of bias across cancer types also illustrate the diverse patterns of the presence of stromal and immune cells. In the case of prostate adenocarcinoma (or PRAD), ESTIMATE infers a much higher tumor purity than the other three estimates due to the increased fractions of normal epithelial cells.

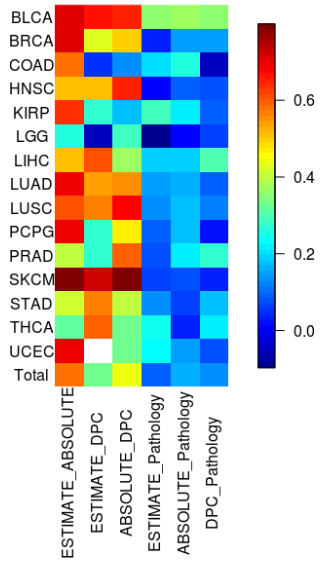

Fig. 1. Pairwise correlation of four purity estimates. Pairwise Pearson correlations between tumor purity estimates of the four methods across 15 cancer types. One white cell on the row of cancer type UCEC means unavailable because there are less than 10 common samples between the datasets for methods ESTIMATE and DPC. Correlations with pathological review are lower compared with the ones between genomic-based methods.

3.2 Associations with clinical features

TCGA PanCanAtlas provides 746 clinical features though not applicable for all cancer types. For the 15 cancer types in this project, I examined the association between tumor purity with 237 unique features (Supplementary Data 2). In general, most clinical features including age, sex, smoking behavior, etc. are not significantly associated with tumor purity level.

However, I detected strong associations (false positive rate < 1%) with two histological features, histological subtypes and neoplasm grade, in at least three cancer types (Fig. 2). Histological subtypes are specific for each cancer type, classifying cancers on the basis of cell morphology, growth and architecture patterns. The histological diversity may have relevant prognosis implications (Dieci et al., 2014). Different tumor purity levels were observed in different subtypes of BRCA, STAD and THCA. Tumor grade is a description indicating how quickly a tumor is likely to grow and spread, assigned by a pathologist based on how abnormal the tumor cells and tumor tissue look under a microscope. Prominent associations between tumor purity and grade were found in cancer types BLCA, STAD and UCEC.

In addition to histological clinical features, this project also detected all the other significant features for each cancer type (Supplementary Fig. 2). For example, in brain lower grade glioma (LGG), samples have lower purity if no IDH1 mutation is tested, which may be related to some research findings about the relationship between IDH mutation and glioma prognosis (Houillier et al., 2010).

As for survival analysis, I applied a Cox proportional hazard model to test the association between tumor purity and patient's survival time. Significant association was only observed in skin cutaneous melanoma (SKCM) that lower tumor purity (in other words, smaller fraction of cancerous cells) indicates longer survival time, which is apparently consistent with common sense.

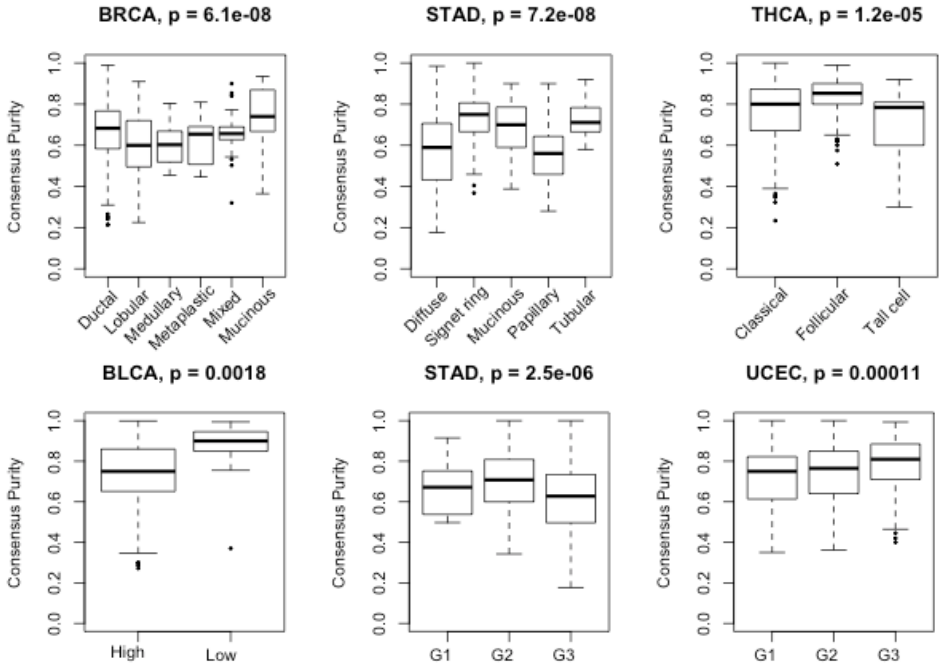

Fig. 2. Tumor purity with clinical features. The upper three boxplots are consensus tumor purity in different histological types of breast invasive cancer (BRCA), stomach adenocarcinoma (STAD) and thyroid carcinoma (THCA); the lower three boxplots shows consensus tumor purity levels in different tumor grades of bladder urothelial carcinoma (BLCA), STAD and uterine corpus endometrial carcinoma (UCEC). One-way ANOVA P values are presented next to the corresponding cancer type names.

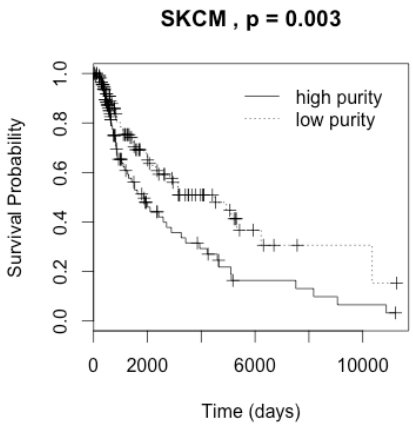

Fig. 3. Kaplan Meier survival curve of SKCM. High and low correspond to the 4th and 1st quantile of tumor purity level, respectively. P value is presented.

3.3 Evaluation estimating performance of the four methods

Due to the diverse characteristics of different cancer types, I hoped to find the best purity-estimating way for each cancer. However, there is no ground truth yet of tumor purity, and hence regular evaluation approach, such as computing root-mean-square error (RMSE), is not applicable to measure estimation error in this case. Taking advantage of the clinical associations examined above, I evaluated the effectiveness of each method in an opposite way.

I defined that one method was "effective" for a specific cancer type if the associations of its estimate with the significant features (prognosis was viewed as one clinical feature), which were identified by consensus purity, were also prominent. Suppose there are *k* clinical features detected to have significant associations with consensus tumor purity level for cancer type *X*, we then test the associations of all the four methods with these *k* features separately and record the times when method *i* (=1, 2, 3, 4) has a P value no larger than the one of consensus measurement as a "score" of method *i*. The method with highest score is considered as the most effective one for cancer type *X* (Table 2).

As a result, the gene-expression-based method ESTIMATE is the optimal choice for most cancer types. At the same time, there are many other computational methods reported that could predict accurate fractions of different cell types in tumor admixtures using gene expression profiles (Newman et al., 2015; Qiao et al., 2012; Shen-Orr and Gaujoux, 2013). It could be caused by the more direct data acquisition step of gene expression profiles while copy-number abbreviations or somatic mutations require an additional calling step, which may probably introduce man-made, non-negligible error.

Table 2. Optimal estimating method for each cancer type

Methods	Cancer Types
ESTIMATE	BLCA, PRAD, SKCM, STAD, THCA
ABSOLUTE	SKCM
DPC	BRCA, PRAD
Pathology review	LGG

One cancer type may have more than one optimal method because of equal scores.

4 Discussion

This project compares four tumor-purity-estimating methods with different data types as inputs on TCGA tumor samples. Strong concordance was observed among genomic-based methods, while the correlation with pathological method is poor (Figure 1). However, the correlation coefficients of pathological estimate of tumor purity are positive in almost all cancer types, suggesting that it is a qualitative estimating approach. Moreover, regardless of the method used, tumor purity levels vary across different cancer types, suggesting tumor purity may be a property of cancer type rather than individual patient.

Through examination of clinical associations, I did not find many major clinical features that have statistically significant associations with tumor purity. The most prominent feature is histological subtypes that patients in different subtype groups usually have diverse purity levels. Based on the detected significantly related features, I evaluated the performance of all methods and attempted to select an optimal one for each cancer type. Nevertheless, this evaluation approach has a limitation that not all cancer types have significantly-associated clinical features with the purity estimates, thus it provides answers to only half of cancer types (Table 2). One possible reason for this limitation could be that none of the four estimating methods are effective for some cancer type. One evidence is that although the dependency on infiltrating immune cells and stromal cells shows an obvious limitation of method ESTIMATE, it is still the best choice for most cancer types. It indicates that the other three methods may have even more severe limitations on purity estimation.

Single-cell technology has developed a lot in recent years (Shapiro et al., 2013), which provides another possibility of exploiting tumor heterogeneity (Buettner et al., 2015) on a single-cell level rather than on tumor admixture. However, single-cell sequencing is still in its infancy and has its own complicated challenges to resolve, such as the prohibitive costs. Therefore, in near future, a more robust, more accurate and generally applicable method for estimating tumor purity based on information of a mixture sample still has to be developed.

Acknowledgements

This project is a course project of class COMS W4761: Computational Genomics. The advisory from Professor Itsik Pe'er and TA Shuo Yang helped me a lot in both the project preparation and course study.

Conflict of Interest: none declared.

References

Albini,A. and Sporn,M.B. (2007) The tumour microenvironment as a target for chemoprevention. *Nat.Rev.Cancer*, **7**, 139-147.

Alizadeh,A.A. *et al.* (2015) Toward understanding and exploiting tumor heterogeneity. *Nat. Med.*, **21**, 846-853.

Buettner,F., *et al.* (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155-160.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413-421.

Dieci,M.V., *et al.* (2014) Rare breast cancer types: histological, molecular and clinical peculiarities. *The Oncologist*, **19**, 805-813.

Houillier,C., *et al.* (2010) IDH1 or IDH2 mutations predict longer survival and response to temozolomide in low-grade gliomas. *Neurology*, **75**, 1560-1566.

ICGC-TCGA-DREAM Somatic Mutation Calling Tumor Heterogeneity (SMC-Het) Challenge. https://www.synapse.org/#!Synapse:syn2813581/wiki/303137 (Assessed on May 2016).

International Cancer Genome Consortium (ICGC) *et al.* (2010) International network of cancer genome projects. *Nature* **464**, 993–998.

Isella,C., *et al*. (2015) Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.*, **47**, 312-319.

Jiao,W., *et al*. (2014) Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinform*., **15**, 35.

Joyce,J.A. and Pollard,J.W. (2009) Microenvironmental regulation of metastasis. *Nat. Rev. Cancer*, **9**, 239-252.

Newman,A.M., *et al*. (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 453-457.

Nik-Zainal, S., *et al*. (2012) The life history of 21 breast cancers. *Cell*, **149**, 994-1007.

Nowell,P.C. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23-28.

Oesper,L., *et al*. (2013) THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.*, **14**, R80.

Qiao,W, *et al*. (2012) PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol*., **8**, e1002838.

Roth,A., *et al*. (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, doi: 10.1038/nmeth.2883.

Shapiro,E., *et al*. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, **14**, 618-630.

Shen-Orr,S.S. and Gaujoux, R. (2013) Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol*., **25**, 571-578.

SMC-Het-Challenge GitHub page: https://github.com/Sage-Bionetworks/SMC-Het-Challenge-Examples/tree/master/dpc (Assessed in April 2016)

The Cancer Genome Atlas (TCGA). http://cancergenome.nih.gov/ (Assessed in May 2016).

The Cancer Genome Atlas (TCGA): The next stage (2015). http://cancergenome.nih.gov/newsevents/newsannouncements/TCGA_The_Next_Stage (Assessed in May 2016).

Yoshihara,K., *et al*. (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**, 2612.