TCGA data repository and treatment selection

Francesco C. Stingo

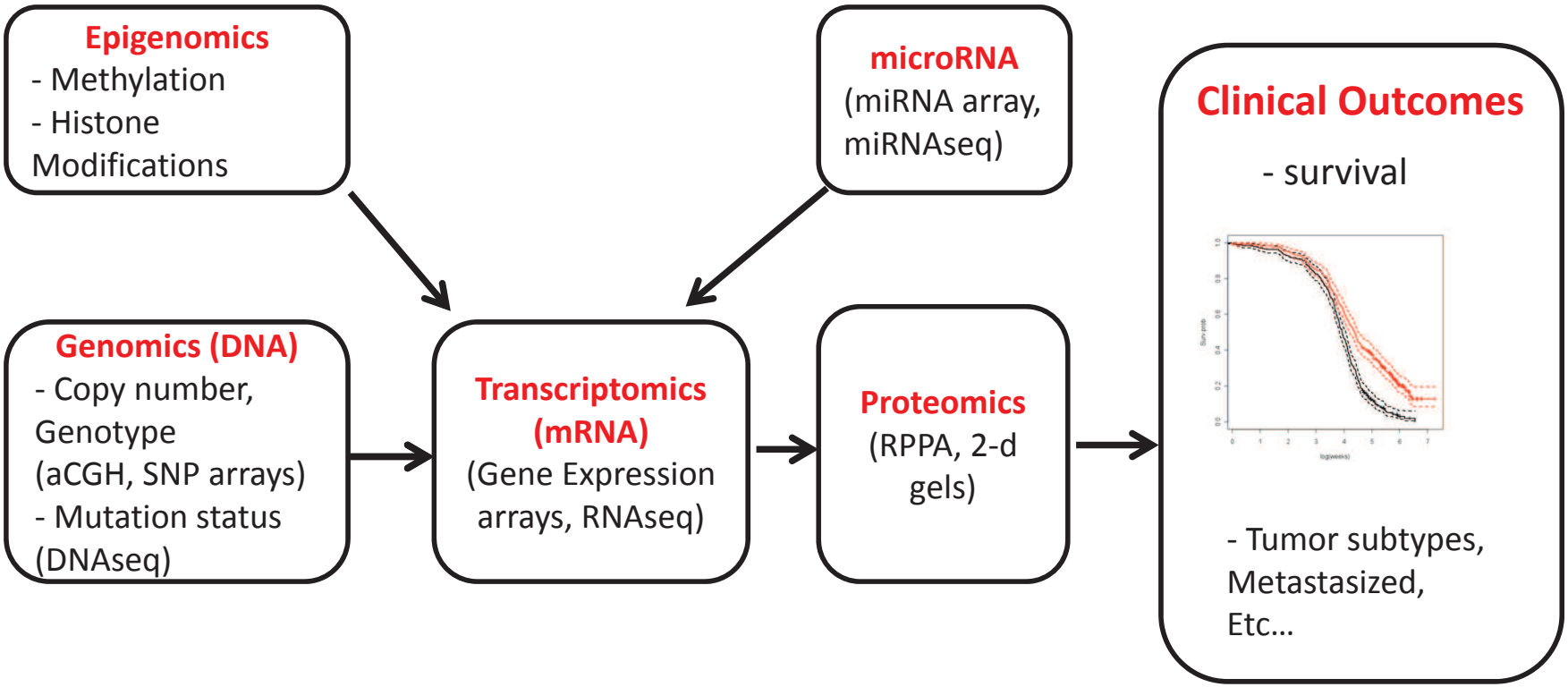
Dipartimento di Statistica, Informatica, Applicazioni “G. Parenti”

July, 15th 2024

Intro to TCGA

- The Cancer Genome Atlas (TCGA) is a landmark collection that maps the genomic profiles of 33 cancer types and subtypes,
- These atlases reveal the molecular features associated with cancer and help to inform everything from basic research to drug development and precision medicine.
- All TCGA data are all in the public domain, allowing any researcher with an interest to access this information.

TCGA overview



Courtesy of Veera Baladandayuthapani

Data science for TCGA

- TCGA data portal gives you a full suite of web-based tools for studying TCGA data, including building and comparing cohorts, examining mutation frequencies, visualizing gene expression clusters, and more. These analysis tools can help you **access and use genomic data**.
- And you don't have to start from scratch. Scientists use TCGA data to develop numerous pipelines and other methodologies for studying cancer. You can also find tools with shortcuts for processing and analyzing data from start to finish.
- TCGA's tumor profiles provide a rich resource for exploring any number of topics of interest: from new drugs or biomarkers to new ways of diagnosing, preventing, and treating cancer.

Personalized Medicine in Oncology

Cancer

- is a set of diseases characterized by cellular alterations the complexity of which is defined at multiple levels of cellular organization (Reimand, 2013; Hanahan et al. 2011).

Personalized medicine

- attempts to combine a patient's genomic and clinical characteristics to devise a treatment strategy that exploits current understanding of the biological mechanisms of the disease

Prognostic versus Predictive Biomarkers

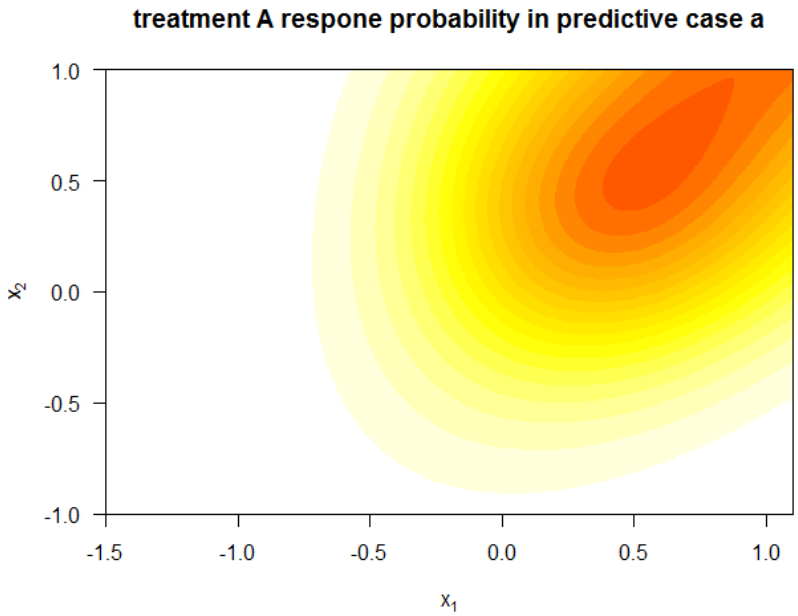
Prognostic

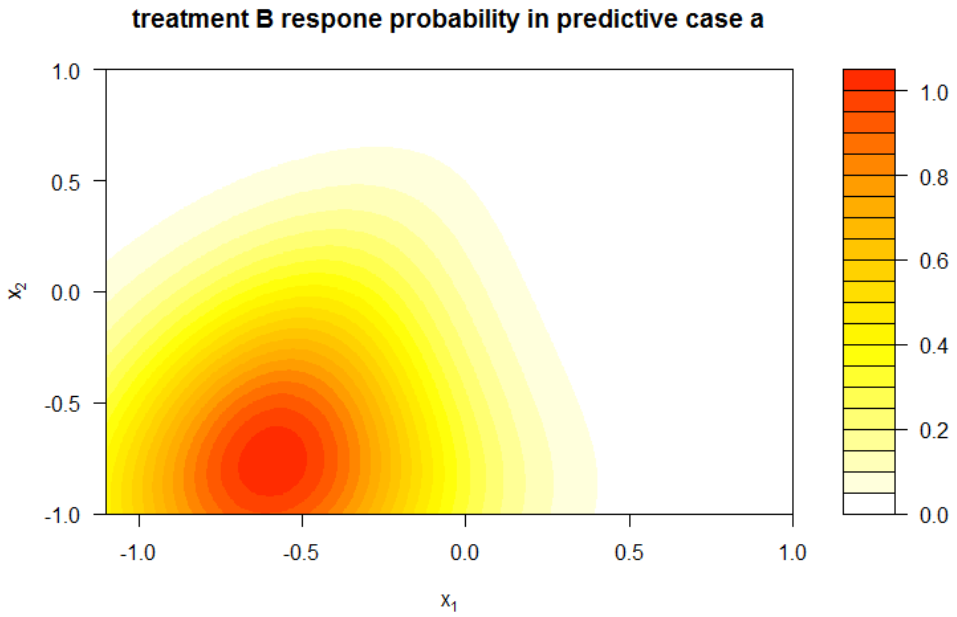
- biomarkers are correlates for the *extent of disease* or *extent to which the disease is curable*
- impact the likelihood of achieving a therapeutic response *regardless* of the type of treatment

Predictive

- distinguish patients who are likely or unlikely to benefit from a *particular class of therapies*
- used to guide treatment selection for individualized therapy based on the specific attributes of a patient's disease
- statistically an interaction between a candidate biomarker and targeted therapy

Predictive Features





Existing approaches

1. Subgroups based on disparate biomarker/mutation subtypes
   - assume patients identified to have the same marker status are statistically exchangeable and thereby ignore heterogeneities within the marker subgroups
   - may not adequately characterize variations in the alterations within cell signaling pathways that enable cancer cells to evade routine cell death and to proliferate and migrate

Existing approaches

Study of individual candidate genes/mutations is limiting

- from 2003 to 2011, 71.7% of new agents failed in phase II (Hay et al. 2014, Nature biotechnology)
- only 10.5% were approved by the FDA (Hay et al. 2014, Nature biotechnology)
- intrinsically complex biological phenomena not well described by markers based on a few features (Knox, 2010)

Future advances in PM will rely on molecular signatures that derive from synthesis of multifarious interdependent molecular quantities requiring more robust quantitative methods

Existing approaches

2. Methods that characterize treatment-by-feature interactions
   - Y denote observeable patient outcome
   - A denote the treatment assignment
   - $\boldsymbol{X} = X_1, X_2, \ldots, X_p$, represents a vector of values for p features
   - $E(Y|A, \boldsymbol{X}) = \mu(A, \boldsymbol{X})$ denote the expected value of Y given A and $\boldsymbol{X}$
   - Optimal treatment rule assigns $A = 1$ versus $A = 0$

$$g^{opt}(\boldsymbol{X}) = I\{\mu(A = 1, \boldsymbol{X}) - \mu(A = 0, \boldsymbol{X}) > 0\},$$

Existing approaches

2. Methods that characterize treatment-by-feature interactions (cont)

E.g., binary endpoints:

- $\mu(A, \boldsymbol{X}) = P(Y = 1 | A, \boldsymbol{X})$
- X_1 represents prognostic feature
- X_2 denotes a predictive feature

$$\log\left\{\frac{\mu(A, \boldsymbol{X})}{1 - \mu(A, \boldsymbol{X})}\right\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + A(\beta_3 + \beta_4 X_2)$$

- Optimal treatment rule assigns $A = 1$ versus $A = 0$ if

$$g^{opt}(\boldsymbol{X}) = I\{(\beta_3 + \beta_4 X_2) > 0\}$$

Protenomics

Proteomics:

- Proteins represent the downstream summation of changes $\rightarrow$ directly related to the phenotype
- (Usually) poor concordance between mRNA and protein abundance

Limitations:

- Mass spectrometry assays are expensive and require plenty of sample material.
- Reverse phase protein arrays (RPPAs) are cheaper and require less material, but limited to a relatively small number of known proteins

TCGA protein data set

Lower Grade Glioma (LGG)

- Glioma is a type of cancer that develops in the glial cells of the brain.
- Tumors are classified into grades I, II, III or IV based on standards set by the World Health Organization
- For this study, TCGA studied lower grade glioma, which consists of grades II and III.
- Regardless of grade, as a glioma tumor grows, it compresses the normal brain tissue, frequently causing disabling or fatal effects.

TCGA protein LGG data set

Data processing: `LGGdata.rda`

- Patients matched on the basis of the baseline covariates of tumor grade, gender, age and initial year of pathological diagnosis (IYPD).
- 79 pairs of patients obtained using the *R* package of *MatchIt* (default settings)
- 173 protein expressions (level 3 data) were measured in the LGG data, 23 proteins were selected as potential predictive features and 5 as potential prognostic features based on univariate screening

Beyond treatment selection

Many possible analyses of interest

- Differential expression analysis
- Combine clinical characteristics with genomics data to inform treatment selection;
- Study similarity and differences in the biology (e.g., in terms of protein networks) of subset of patients (respondent Vs non-respondent);
- Integration with additional TCGA data (but it needs to download additional data, e.g., mutations, from TCGA);
- Perform dimension reduction;
- Perform variable/protein selection with respect of a response of interest (e.g., treatment response).

Beyond LGGdata.rda dataset

Additional data from TCGA R code to directly download the most recent protein and clinical dataset from TCGA

- Expression levels of many more proteins
- More samples
- Additional patient-level information
- Can be seen as an example on how to retrieve data from TCGA

Beyond LGGdata.rda dataset

Data stored in `clinical_data.csv`, `clinical_drug.csv`, and `protein_data.csv`

- **Clinical Data**: detailed clinical information about the patients, including demographics, diagnosis, treatment history, and follow-up data. It has 515 records and 75 features.
- **Treatment Data**: treatments administered to the patients, e.g., chemotherapy, radiotherapy, and molecularly targeted treatments. It has 695 records and 24 features.
- **Protein Expression Data**: protein expression levels in primary tumor samples. It has 487 records and 434 features.

The `bcr_patient_barcode` serves as the primary key. These datasets provide raw data. Therefore, no data were excluded or imputed, and no variables of interest were derived for analysis.