Cancer classification in the genomic era: five contemporary problems

Qingxuan Song[1], Sofia D. Merajver[2], Jun Z. Li[1,*].

[1]Department of Human Genetics, [2]Department of Internal Medicine and Epidemiology, University of Michigan, Ann Arbor.

Authors' email address:

Qingxuan Song: qsong@umich.edu

Sofia D. Merajver: smerajve@med.umich.edu

*Corresponding author

Jun Z. Li, Ph.D.

Department of Human Genetics

University of Michigan

5789A Medical Science II, Box 5618

Ann Arbor, MI, 48109-5618

Phone: 734-615-5754

Email: junzli@med.umich.edu

Keywords (3-10 keywords):

cancer, classification, genomics, integration, evolution, uncertainty, precision,

Abstract

Classification is an everyday instinct as well as a full-fledged scientific discipline. Throughout the history of medicine, disease classification is central to how we organize knowledge, obtain diagnosis, and assign treatment. Here we discuss the classification of cancer, the process of categorizing cancers based on their observed clinical and biological features. Traditionally, cancer nomenclature is primarily based on organ location, e.g., "lung cancer" designates a tumor originating in lung structures. Within each organ-specific major type, further subgroups can be defined based on patient age, cell type, histological grades, and sometimes molecular markers, e.g., hormonal receptor status in breast cancer, or microsatellite instability in colorectal cancer. In the past 15+ years, high-throughput technologies have generated rich new data for somatic variations in DNA, RNA, protein, or epigenomic features for many cancers. These data, representing increasingly large tumor collections, have provided not only new insights into the biological diversity of human cancers, but also exciting opportunities for discovery of new cancer subtypes. Meanwhile, the unprecedented volume and complexity of these data pose significant challenges for biostatisticians, cancer biologists, and clinicians alike. Here we review five related issues that represent long-standing problems in cancer taxonomy and interpretation. 1. How many cancer types are there? 2. How can we evaluate the robustness of a new classification system? 3. How are classification systems affected by intratumor heterogeneity and tumor evolution? 4. How should we interpret cancer subtypes? 5. Can multiple classification systems coexist? While these problems are not new, we will focusing on aspects that were magnified by the recent influx of complex multi-omics data. Ongoing exploration of these problems is essential for developing data-driven cancer classification and the successful application of these concepts in precision medicine.

Introduction

Classification and labeling represent the most intuitive forms of learning. According to the Confucius, "If the name is not right then speech will not be in order, and if speech is not in order then nothing will be accomplished". Instances of classification can be found in every aspect of life: the Linnaean system in biology is a seven-level framework for cataloging living organisms; E-commerce companies must organize their holdings in an easy-to-search system. Whether the goal is to recommend movies, screen job candidates, or rank colleges, a system of dividing and summarizing is involved, and uncertainties inherent to this task are common. In a quote attributed to Albert Einstein [1], classification is "an attempt to make the chaotic diversity of our sense experience correspond to a logically uniform system of thought". The tension between "chaotic diversity" on one hand and a "logically uniform system" on the other makes classification a long-standing subject in scientific research. In practice, it was regarded more often as an art than a science, as there is no axiomatic classification theory that applies to all problems. This review will discuss the topic of cancer classification. Instead of offering a systematic review of facts for specific cancers we will focus on the implicit assumptions and common caveats of classification, especially how it has been confronted with new challenges in the genomic era. We organize the many strands of this topic under the heading of Five Contemporary Problems, although in many ways they are new faces of old problems.

A few notes of clarification before we begin. In the field of statistical learning the term "classification" refers to *supervised sample assignment*, using features identified in a training set of samples with known class labels. Here we adopt its other, more commonly understood meaning, referring to *ab initio* pattern recognition, also known as *unsupervised class discovery*. We will use the terms (sub)class and (sub)type interchangeably and will only discuss sample classification, not gene clustering. We sometimes use "genomics" to refer to all high-throughput -omics data.

1. How many cancer types are there?

The short answer is about 200: the National Cancer Institute keeps an A-Z list of nearly 200 cancer types [2], organized by organ location, although with some exceptions, such as "HIV-related" or "unknown primary". This organ-centric system is further stratified most often by the known cell type of origin within the organ, e.g., "astrocytomas", as a subtype of brain tumors [3], or by patient age, e.g., "childhood leukemia". If each of these have four subtypes, there will be 800 subtypes of cancer. This review will focus on how to identify subtypes for a broadly recognized organ-specific cancer.

"Endless forms most beautiful"

While the organ-derived naming system has been used for over a century by doctors, patients and cancer registries, there has always been the drive to recognize finer subtypes; and this trend has accelerated dramatically with genomic data. For example, primary breast cancers have traditionally been classified in the clinic by the expression status of cell surface receptors into three subtypes: estrogen receptor (ER)-positive or progesterone receptor (PR)-positive, human epidermal growth factor receptor 2 (HER2)-positive, and triple negative [4]. The result of the immunohistochemical (IHC) assays of these receptors can thus guide the selection of patients for different hormonal therapies or targeted therapies [5, 6]. The arrival of microarray-based gene expression data led to the division of breast cancers into five intrinsic molecular subtypes: luminal A, luminal B, HER2-overexpression, normal-like, and basal-like, and they showed notable differences in clinical outcome [7, 8]. More recently, analyses of copy number and gene expression data for several thousand malignant breast tumor samples revealed 10 molecular subtypes [9, 10]. Meanwhile, the triple negative subtype was further divided [11, 12]; and with a sample size of >10,000 and additional basal markers Blow et al. [13] identified six IHC subtypes for breast cancer by splitting the basal group.

At least three factors contribute to the continuing fracturing of cancer subtypes. First, larger sample cohorts are being used—they contain not only more phenotypic heterogeneity but also greater power to define and replicate finer subtypes. Second, simultaneous collection of diverse -omics data types, such as

aberrations in DNA, gene expression, and epigenetic features, tend to recognize additional "structures" in a given tumor series. Third, intrinsic gene-gene correlations in the data, often unrelated to the disease itself, may create illusions of stable clusters with the use of some methods [14] even in cases where the structure is weak.

An ever finer classification system has many potential benefits. It is needed to capture the full spectrum of biological diversity—the "endless forms" that Darwin spoke of. It could lead to a better recognition of patient-specific disease mechanisms, and importantly, could suggest treatment options that are more accurately matched to the patient's tumor. Precision medicine, at its very foundation, relies on valid and continuously optimized disease classification that reflect the underlying mechanisms. However, a fine-grained classification system also has many potential drawbacks. The newly proposed splits may not be technically robust (see #2 below). Even when the finer categories are robustly supported by statistical significance and by replication, they may still lack a clear biological meaning, or have little impact on treatment options (#3 below) if it turns out that some subtypes share the same clinical endpoint, or if treatment options are limited.

2. Is the classification system robust?

The merit of a classification system depends on what it will be used for. Often, a new map is drawn, but its intended function is unclear. Only by explicitly laying down the purpose can we evaluate the merit of a proposed system according to what it aims to achieve. The potential measures of merit include: statistical robustness, connections with existing standards, prognostic value, insights into biological mechanisms, and predictive power of treatment outcomes. As we discuss later, sometimes these merits cannot be achieved at once; and multiple classification systems can be simultaneously "correct" depending on which purpose they are optimized to serve. Here we delve mainly into the issue of robustness: how can we tell if a classification system is more reliable compared to alternative systems?

This question is worth considering as it is regularly overlooked in practice: many reports of new classifications are filled with biological interpretation but are light on evaluating the strength of evidence.

Is there a p value?

In epidemiological or genetic association studies, evidence of credible association is measured by effect size and statistical significance, the latter being expressed by a p-value and a hypothesis-testing procedure used to calculate it. For example, a DNA variant's additive effect on a continuous trait can be evaluated by linear regression. However, the task of classification cannot be easily cast into a hypothesis-testing framework: when declaring K clusters for a sample, is the null hypothesis "no cluster" or "K-1 and K+1 clusters"? While the confidence of *class assignment* can be assessed by cross-validation in test samples for which the class labels are already known, there is no well-established statistics to compare the performance of *class discovery*. In theory, one can quantify the degree of clustering by indices such as compactness, connectedness, separation [15], or silhouette width [16], however there is no universally accepted p value-like measure to report how likely the observed clusters could arise merely due to naturally occurring data "structure". Two types of structure are frequently encountered in high-dimensional molecular profiling studies: that due to separations between groups, i.e., stratification, and that due to locally tight clusters, i.e., cryptic relatedness. These structures are routinely seen in human population genetics studies and ultimately they came from shared ancestry of sampled individuals at different time depths. Their impact on association tests can be monitored and corrected by well-established procedures [17, 18]. However, for gene expression or other functional genomics data (such as proteomic, metabolomic, epigenomic data), the information used in classification is sample-sample similarity in high-dimensional feature space, and the basis of co-ancestry is lacking (at least not self-evident). Indeed, how to evaluate competing algorithms or alternative results is an active topic of research [19]. Many groups have studied the issue of cluster validation and have proposed the use of either internal or external standards [20-22]. More often however, there is no real dataset that can serve as a reliable external standard. Our recent analyses have shown that even the datasets that are said to

contain well separated clusters can have an uncertain number of clusters (i.e., the true K), thus making it difficult to use them as benchmarks for comparing class discovery methods [14]. We and others have also noted that as sample size increases, the number of clusters will increase and exhibit increasing apparent stability [14, 23]. Given this, we recommend the use of simulations to create internal standards, in which the number of clusters and the degree of separation are known, and the gene-gene correlations can be introduced based on empirical data.

Quantitative reporting of the robustness of classification result is often lacking in publications proposing new classification systems. Sometimes the data structure was *illustrated* by pre-selecting the best discriminating genes and showing how they could visually separate the reported clusters crisply. Although this form of presentation is well suited for annotation: which genes appeared in which group, it is not appropriate as a demonstration of cluster strength, because with many more genes than samples (i.e., the p >> n situation) seemingly informative discriminators can always be found for any random partition, even for samples without clear groupings. When classification strength is not properly assessed, visual display of clusters using the best genes can inadvertently turn into an exaggerated inference, even if subsequent interpretations seem appealing [14].

3. Can classification capture intratumor heterogeneity and evolutionary progression?

Every living cancer inevitably change its character in time and every solid tumor is spatially heterogeneous, yet samples used in research are bulk tissue blocks collected as a single time point. Thus most cancer genomics data, by the very nature of sampling, provide a one-time view of a mixed pool of cells. This limitation can only be overcome with single-cell analysis or longitudinal sampling. For a given cohort of patients, classifying their bulk tumor data is an attempt to find natural groupings among many mixed cell populations, while ignoring the within-population diversity and its variation in space and time. Standard cancer classifications are aimed at capturing *inter*-tumoral heterogeneity, since they treat

each tumor as homogeneous and unvarying. Not surprisingly, sometimes the cancer subtypes reported are actually driven by, and therefore reflect, intratumor heterogeneity and tumor evolution.

Spatial heterogeneity

With bulk-tissue data, spatial heterogeneity is an unobserved property. If the sample consists of a limited number of clonal populations, the number and molecular features of these component populations can be potentially "deconvolved" computationally. As sequencing costs drop, spatial heterogeneity can be analyzed with increasing resolution: first by multi-region analysis of smaller sectors of the same tumor [24-26], ultimately by DNA or RNA sequencing of single cells [27, 28]. In multi-region analyses, a typical assumption is that individual regions are clonally "pure", and the results can be presented as such. However, it remains the rule rather than the exception that each region still contains a mixture of many cell types, albeit with presumed lower heterogeneity compared to the entire tumor [29]. Single-cell analysis provides the ultimate solution, as it describes the smallest unit of cancer heterogeneity and provides truly clonal data for use in classification. Single-cell studies have identified many more cell types than previously seen with bulk tissue analysis. For example, a single-cell RNAseq study of cortical tissues [30] has found nine major classes and 47 "molecularly distinct" subclasses of brain cells, significantly expanding known cellular repertoire of the mammalian cortex. For the foreseeable future, bulk tissue sampling will remain the predominant source of cancer genomics data, both for basic research and for real-time clinical testing. While spatial heterogeneity can be reduced by sampling smaller and smaller "core" regions it cannot be fully removed. In this regard the traditional, *hard* classification into disjoint categories is a poor fit for admixed samples, as they contain (1) cancer cells carrying somatic mutations or aneuploid segments and (2) surrounding normal-like cells that are euploid and carrying only germline mutations. Partial membership modeling has been proposed to address this scenario [31-33], reminiscent of similar methods for ancestry inference using genotype data for individuals with mixed ancestry [34, 35].

Cancer life history and impact on classification

Classification of cancers can also be viewed as a problem of cataloging evolutionary trajectories of complex genomes [36, 37]. Each cancer genome carries many variations, with a distribution of fitness effects as seen in a changeable environment. Each tumor thus undergoes its own Darwinian evolution, with many intricate details that make it distinct from all other tumors. However our effort to classify them is predicated on the notion that convergent evolution does happen, such that a limited number of evolutionary paths are traveled repeatedly by tumors from different patients, leading to recognizable major hallmarks recurring in different tumors. The discovery of N subtypes of breast cancer, for example, reflects N destinations of convergent evolution in this cancer type. But N needs not be a fixed parameter. For example, if the tissue has K cell types that could eventually turn to cancer cells, the pre-cancerous cells have K starting points to explore the initial evolutionary paths, and multiple paths may merge or split during the life history of a cancer before congregating to one of N major types at the time of sampling. Metastasis and treatment response would further extend, diversify, or reshuffle the evolutionary trajectories [38, 39].

Current cancer classification systems can only slightly portray this complex succession of events, as different tumors may be sampled at different stages along their own life histories. In a typical lifespan of a tumor, some cells start acquiring oncogenic potential, showing increased proliferation, escaping apoptosis and immunosuppression, competing successfully for resources with other cells in the tissue niche, harnessing the right combination of driver mutations with enough fitness gain to overcome the fitness burden incurred by the larger number of passenger mutations [40]. The cancer then needs to grow sufficiently large and diverse before tissue turnover, and further, to acquire migration and invasion properties that are essential for metastasis. It may also carry treatment-resistance subclones that can thrive after therapy. Viewed in this light, cancers are chronic diseases with changeable character and rare episodes of acceleration, essentially slow-evolving populations that occasionally rush to a "successful" endpoint. A given sample collection may have captured the tumors at different "stations" of their life

history. For example, we and others have noted that the mesenchymal subtype of glioblastoma multiforme (GBM) possesses the signature of macrophages/microglial infiltration [41, 42], and has a greater degree of mixing of aneuploid and euploid cells. More recently, another group reported that the mesenchymal subtype may have evolved from another known subtype, a proneural-like precursor of GBM [43].

4. How to interpret cancer subtypes?

In most cases, the investigator's task after class discovery is to explain the meaning of the found classes. Typical actions include: describe the appearance of known cancer-related genes as a way to report specific signaling pathways active in different subtypes; map the new class nomenclature to some of the previously established systems; assess differences in clinical outcome, e.g., survival time or treatment responsiveness; dissect tissue-specific signatures or enrichment of previously curated gene clusters. In the following sections we discuss three caveats in such interpretations. Both marker selection and analysis method have a strong impact on the classification results and how the resulting classes can be explained. Further, the third factor affecting interpretation is the strength of clustering inherent in the dataset, which is not known until a specific collection of tumors has been assembled and analyzed.

Feature selection bias

In population genetics, DNA variation data can be used to (1) infer historical demography or (2) detect natural selection. These two tasks are related, but are also inherently different, even conflicting with each other. Demographic inference seeks to describe historical changes of population size and distribution, including migration, self-isolation, expansion, and admixture. The best genetic markers for this task are "neutral" variations, those not under natural selection, or more accurately, those not likely to affect the Darwinian fitness of the individuals (to our best knowledge at the present). Typical examples of such neutral markers are those found in intergenic regions of the genome. In contrast, the second task, inference of natural selection, relies on markers with a fitness effect and must be built on a null model of

genetic drift, because demographic forces could have generated DNA variation patterns similar to those due to natural selection.

In cancer evolution, the concept of *driver* and *passenger* mutations are almost exact parallels of the *adaptive* (*positively selected*) and *neutral* variants defined in population genetics [44, 45]. However, we rarely ask whether the goal of cancer classification is for understanding the tumors' past, or their future. The story of the cancer's past is shaped by the drivers but may be best recorded by the passengers. And the drivers in the past may not be the forces of selection acting in the future. The type of features selected for use in classification will therefore directly affect the downstream interpretation. For example, if the analysis of gene expression data pre-selects transcripts that correlate with survival time, these markers, by being most informative for future outcome, are likely to reveal tumor subtypes that differ in outcome, and this will lead directly to the interpretation that the subtypes thus discovered have a stronger prognostic/predictive value than classes discovered by other means. Alternatively, if the selected transcripts correspond to cell type or pathway-specific genes, the resulting classes will likely exhibit differential loading of pathway or source cell signals, and make it easy to map the discovered classes to biological mechanisms. Even the presumably "safe" choice of selecting the most variable genes may have inadvertently loaded the interpretation towards the most dynamic or most strongly co-expressed pathways, such as those for stress response or immune cell infiltration.

Hidden assumptions in choosing a method

Unsupervised clustering is the basic tool for *ab initio* class discovery. The term "unsupervised" refers to statistical inference of data structure without relying on existing knowledge of sample labels. However, "unsupervised" does not mean assumption-free. Some important assumptions have always been made, sometimes unknowingly, in marker selection, data processing (such as how to treat outliers and how to deal with batch effects), and the choice of the clustering method [46-49]. This methodological choice is already based on an implied data-generating model, i.e., what type of biological heterogeneity could have

produced the observed data structure. If we assume that the objects - different tumor samples - arise from distinct, non-overlapping causes, a method for finding disjoint taxonomy, such as the k-means clustering method, is appropriate. If, alternatively, we see each tumor sample as a mixed population of cells comprising a small number of canonical clones, a mixed membership model is appropriate [31, 32, 50], as has been routinely applied in studies of human population diversity involving individuals of mixed ancestry [34]. The number of co-existing clones and the rate of clonal replacement depend on population size, mutation rate, and the distribution of fitness effects of the new mutations. If the task is to classify cells in a single evolving population with major branches, it would be best to capture the lineage relationship by using hierarchical clustering - akin to the coalescence analysis in classic population genetics [51] - aiming to identify hierarchical classes to reflect their shared ancestry as the historical truth. When using simulated data to compare methods such as k-means clustering or hierarchical clustering, the data-generating model will usually dictate the winner: the method that matches the model will fit the simulated data best.

Inherent intensity of data structure

Perhaps it is worth reiterating that when the clustering signal is strong, most methods will perform well and they will be in strong agreement. However, when the data structure is subtle, slight differences in sampling, data processing, feature selection, or the choice of method will yield highly discordant results, and it is difficult to tell which is better. A recent example of strong clustering signal is from a joint analysis of 3,527 specimens of 12 cancer types by the Cancer Genome Atlas (TCGA) Research Network [52]. Most identified classes follow the cancer's known organ of origin, undoubtedly due to the distinct cell types found in different tissues. This result is as expected, because it affirms the earliest insight that cancer cells are transformed normal cells rather than entirely "foreign" cells (as is the case in bacterial infection). As cells are "canalized" during differentiation, a tumor occurs in one of the Waddington's valleys and bears its local hallmark. Speaking in more modern terms, the malignant transformation, although a radical step in the adaptive evolution of the cancer cells, usually could not have erased their

inherited tissue-of-origin signature of prior differentiation. This identity, perhaps coded in the epigenomes of the fate-committed cells, remain the most noticeable molecular character in matured organs despite subsequent oncogenesis. Meanwhile, remarkable exceptions to the tissue-centric classification are also found. The study revealed instances of cross-tissue subtypes [52], e.g., squamous-like lung, head and neck, and bladder cancers that are more similar to each other than other cancers from the same organ, and they appear to have overcome the ontological divergence of the source tissues. Is this because convergent evolution—newly acquired oncogenic characters over-writing history, or shared lineage—the same group of differentiated cells "seeded" into two different body sites? Either scenario would be immensely interesting. In principle, it is entirely conceivable that the same pathways are activated in different cell types as the oncogenic driver. Such cross-tissue classification is at the forefront of pan-cancer analysis today, made possible by the availability of multi-cancer datasets collected under uniform technical conditions. We are at the beginning of the best time to study population genetics of somatic cells. It will stimulate new theoretic work on the evolution of non-recombining populations, produce patterns that can be contrasted with experimental evolution of microbial systems, and will enable the increasingly promising practice of precision oncology. Convergent evolution means that our children's tumors will be about the same as ours; and here lies the hope of eventually defeating this disease by knowing it better.

5. Can multiple classification schemes coexist?

Traditional classification methods rely on organ type, appearance, and histological markers. Multiple systems, such as tumor grade and stage, have coexisted for decades. In recent years, the influx of high-throughput molecular studies have rapidly increased the number of competing systems. For example, the analyses of human breast tumors by the TCGA [53] produced multiple answers to the "how many subtypes" question for the same tumor cohort. It found that breast tumors' gene expression data supported 13 classes with the use of a consensus cluster-based method, 12 classes with a second method, and five classes with the semi-supervised PAM50 method [54]. The concordance rate among the three results was

modest, as the best-matched classes between any two methods only account for 50-60% of the samples (our unpublished observation). Further, the study found seven breast tumor subtypes from microRNA data, five subtypes from methylation data, and five subtypes from copy number alterations, again with poor to modest agreement (per our analysis). While the classification solutions were described as "correlated" across data types, their differences impacted a large fraction of the samples, making it difficult to give a straight answer to the simple question: how many clinically relevant subtypes are there for breast cancer? To integrate such complex data requires quantitative assessment of clustering strength within each data type, and a system to truly integrate the solutions rather than recounting them side-by-side.

When the same data type lead to two or more different answers by the use of different methods, it becomes a methodological imperative to work out a consensus: there is no sound reason for multiple "truths" within the same raw data. However, across different genomic data types it is less clear that there must be a single unifying classification. If we assume that biological information flows from DNA to epigenetic marks, then RNA, and then to proteins, if the cells commit to its differentiated fate primarily using epigenetic codes yet react to short-term needs by gene expression adaptation, further, if the cells interact with their microenvironment chiefly through variations of metabolites, shouldn't it be possible that different levels of biology coalesce to different patterns of grouping, and the same sample truly belongs to different groups depending on the level of inquiry? We think it remains an open question whether different layers of genomic data signify different archetypes of cellular states and could lead to different but equally valid classification systems. To test such a multi-layered classification would require very large datasets that can validate the class-to-class mapping across layers.

Conclusions

Cancer classification is both a scientific technique and a living art, to be performed for each dataset with individualized care. Classification results are widely used, and form the foundational knowledge for both

basic and translational oncology. In this review we outlined five contemporary challenges at the interface of computational data mining, biological understanding, and clinical utility. To classify clinically observed late-stage tumors is to engage in probabilistic cataloging and reverse story-telling, not unlike other observational sciences such as archeology or anthropology. The arrival of genomic data has dramatically increased the power to peer into the past, but even now, amidst the excitement of many new opportunities, it is useful to keep in mind that sometimes the sample series at hand may not be sufficient to support the full ambition of fine-grained classification or tracing the full evolutionary trajectories. Assessing and communicating the strength of data, in quantitative terms whenever possible, is essential for the long-term management of predictive uncertainty, and for the successful application of genomics in patient care.

Competing Interests

There is no competing interest.

Authors' contributions

QS and JZL conceived the article and took the lead in writing the initial draft. All authors contributed to the research discussed in the article.

Acknowledgements

QS is supported by funding from the Joint Institute for Translational and Clinical Research of the University of Michigan Health System and Peking University Health Sciences Center. SDM is supported by the Breast Cancer Research Foundation and the Metavivor Foundation. JZL is supported by general funds of the Department of Human Genetics and the Department of Computational Medicine and Bioinformatics of the University of Michigan.

References

1. McKusick VA: **On lumpers and splitters, or the nosology of genetic disease**. *Perspect Biol Med* 1969, **12**(2):298-312.
2. [http://www.cancer.gov/types/by-body-location].
3. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW, Kleihues P: **The 2007 WHO classification of tumours of the central nervous system**. *Acta Neuropathol* 2007, **114**(2):97-109.
4. Prat A, Perou CM: **Deconstructing the molecular portraits of breast cancer**. *Molecular oncology* 2011, **5**(1):5-23.
5. Morrison DH, Rahardja D, King E, Peng Y, Sarode VR: **Tumour biomarker expression relative to age and molecular subtypes of invasive breast cancer**. *British journal of cancer* 2012, **107**(2):382-387.
6. Caudle AS, Yu TK, Tucker SL, Bedrosian I, Litton JK, Gonzalez-Angulo AM, Hoffman K, Meric-Bernstam F, Hunt KK, Buchholz TA *et al*: **Local-regional control according to surrogate markers of breast cancer subtypes and response to neoadjuvant chemotherapy in breast cancer patients undergoing breast conserving therapy**. *Breast cancer research : BCR* 2012, **14**(3):R83.
7. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al*: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747-752.
8. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S *et al*: **Repeated observation of breast tumor subtypes in independent gene expression data sets**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8418-8423.
9. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y *et al*: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups**. *Nature* 2012, **486**(7403):346-352.
10. Ali HR, Rueda OM, Chin SF, Curtis C, Dunning MJ, Aparicio SA, Caldas C: **Genome-driven integrated classification of breast cancer validated in over 7,500 samples**. *Genome Biol* 2014, **15**(8):431.
11. Prat A, Parker J, Karginova O, Fan C, Livasy C, Herschkowitz J, He X, Perou C: **Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer**. *Breast Cancer Research* 2010, **12**(5):R68.
12. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA: **Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies**. *The Journal of Clinical Investigation* 2011, **121**(7):2750-2767.
13. Blows FM, Driver KE, Schmidt MK, Broeks A, van Leeuwen FE, Wesseling J, Cheang MC, Gelmon K, Nielsen TO, Blomqvist C *et al*: **Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies**. *PLoS Med* 2010, **7**(5):e1000279.
14. Şenbabaoğlu Y, Michailidis G, Li JZ: **Critical limitations of consensus clustering in class discovery**. *Sci Rep* 2014, **4**.
15. Handl J, Knowles J, Kell DB: **Computational cluster validation in post-genomic data analysis**. *Bioinformatics* 2005, **21**(15):3201-3212.
16. Rousseeuw PJ: **Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis**. *J Comput Appl Math* 1987, **20**:53-65.
17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nat Genet* 2006, **38**(8):904-909.

18. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies**. *Nat Genet* 2010, **42**(4):348-354.
19. Hand DJ: **Classifier technology and the illusion of progress**. *Statistical Science* 2006, **21**(1):1-15.
20. Kleinberg J: **An Impossibility Theorem for Clustering**. In: *Adv Neural Inf Process Syst.* 2002.
21. Lange T, Roth V, Braun ML, Buhmann JM: **Stability-based validation of clustering solutions**. *Neural Comput* 2004, **16**(6):1299-1323.
22. de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A: **Clustering cancer gene expression data: a comparative study**. *BMC Bioinformatics* 2008, **9**:497.
23. Ben-David S, von Luxburg U, Pal D: **A Sober Look at Clustering Stability**. In: *Learning Theory: Lecture Notes in Computer Science.* vol. 4005. Berlin Springer; 2006: 5-19.
24. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, Jamal-Hanjani M, Shafi S, Murugaesu N, Rowan AJ *et al*: **Spatial and temporal diversity in genomic instability processes defines lung cancer evolution**. *Science* 2014, **346**(6206):251-256.
25. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, Fisher R, McGranahan N, Matthews N, Santos CR *et al*: **Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing**. *Nat Genet* 2014, **46**(3):225-233.
26. Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, Seth S, Chow CW, Cao Y, Gumbs C *et al*: **Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing**. *Science* 2014, **346**(6206):256-259.
27. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H *et al*: **Clonal evolution in breast cancer revealed by single nucleus genome sequencing**. *Nature* 2014, **512**(7513):155-160.
28. Ting DT, Wittner BS, Ligorio M, Vincent Jordan N, Shah AM, Miyamoto DT, Aceto N, Bersani F, Brannigan BW, Xega K *et al*: **Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells**. *Cell reports* 2014, **8**(6):1905-1918.
29. Polyak K: **Heterogeneity in breast cancer**. *J Clin Invest* 2011, **121**(10):3786-3788.
30. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C *et al*: **Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq**. *Science* 2015, **347**(6226):1138-1142.
31. Oesper L, Mahmoody A, Raphael BJ: **THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data**. *Genome Biol* 2013, **14**(7):R80.
32. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Cote A, Shah SP: **PyClone: statistical inference of clonal population structure in cancer**. *Nature methods* 2014, **11**(4):396-398.
33. Niknafs N, Guthrie VB, Naiman DQ, Karchin R: **SubClonal Hierarchy Inference from Somatic Mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing**. In: *Biorxiv.* 2015.
34. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data**. *Genetics* 2000, **155**(2):945-959.
35. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA *et al*: **Genome-wide patterns of population structure and admixture in West Africans and African Americans**. *Proc Natl Acad Sci U S A* 2010, **107**(2):786-791.
36. Greaves M, Maley CC: **Clonal evolution in cancer**. *Nature* 2012, **481**(7381):306-313.
37. Burrell RA, McGranahan N, Bartek J, Swanton C: **The causes and consequences of genetic heterogeneity in cancer evolution**. *Nature* 2013, **501**(7467):338-345.

38. Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL *et al*: **Genome remodelling in a basal-like breast cancer metastasis and xenograft**. *Nature* 2010, **464**(7291):999-1005.
39. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD *et al*: **Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing**. *Nature* 2012, **481**(7382):506-510.
40. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5):646-674.
41. Li B, Senbabaoglu Y, Peng W, Yang M-l, Xu J, Li JZ: **Genomic estimates of aneuploid content in Glioblastoma Multiforme and improved classification**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2012, **18**(20):5595-5605.
42. Bhat KP, Balasubramaniyan V, Vaillant B, Ezhilarasan R, Hummelink K, Hollingsworth F, Wani K, Heathcock L, James JD, Goodman LD *et al*: **Mesenchymal differentiation mediated by NF-kappaB promotes radiation resistance in glioblastoma**. *Cancer Cell* 2013, **24**(3):331-346.
43. Ozawa T, Riester M, Cheng YK, Huse JT, Squatrito M, Helmy K, Charles N, Michor F, Holland EC: **Most human non-GCIMP glioblastoma subtypes evolve from a common proneural-like precursor glioma**. *Cancer Cell* 2014, **26**(2):288-300.
44. Holderegger R, Kamm U, Gugerli F: **Adaptive vs. neutral genetic diversity: implications for landscape genetics**. *Landscape Ecol* 2006, **21**(6):797-807.
45. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW: **Cancer genome landscapes**. *Science* 2013, **339**(6127):1546-1558.
46. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL: **Bayesian correlated clustering to integrate multiple datasets**. *Bioinformatics* 2012, **28**(24):3290-3297.
47. Lock EF, Hoadley KA, Marron JS, Nobel AB: **Joint and Individual Variation Explained (Jive) for Integrated Analysis of Multiple Data Types**. *The annals of applied statistics* 2013, **7**(1):523-542.
48. Shen R, Wang S, Mo Q: **Sparse Integrative Clustering of Multiple Omics Data Sets**. *The annals of applied statistics* 2013, **7**(1):269-294.
49. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A: **Similarity network fusion for aggregating data types on a genomic scale**. *Nature methods* 2014, **11**(3):333-337.
50. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, Vij R, Tomasson MH, Graubert TA, Walter MJ *et al*: **SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution**. *PLoS computational biology* 2014, **10**(8):e1003665.
51. Bouaziz M, Paccard C, Guedj M, Ambroise C: **SHIPS: Spectral Hierarchical clustering for the Inference of Population Structure in genetic studies**. *PloS one* 2012, **7**(10):e45685.
52. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V *et al*: **Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin**. *Cell* 2014, **158**(4):929-944.
53. TCGA: **Comprehensive molecular portraits of human breast tumours**. *Nature* 2012, **490**(7418):61-70.
54. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z *et al*: **Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes**. *Journal of Clinical Oncology* 2009, **27**(8):1160-1167.