Emerging Topics in Life Sciences (2021) **5** 747–756
https://doi.org/10.1042/ETLS20210212

Review Article

Artificial intelligence, molecular subtyping, biomarkers, and precision oncology

John Paul Shen

Department of Gastrointestinal Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, U.S.A.

Correspondence: John Paul Shen (Jshen8@mdanderson.org)

A targeted cancer therapy is only useful if there is a way to accurately identify the tumors that are susceptible to that therapy. Thus rapid expansion in the number of available targeted cancer treatments has been accompanied by a robust effort to subdivide the traditional histological and anatomical tumor classifications into molecularly defined subtypes. This review highlights the history of the paired evolution of targeted therapies and biomarkers, reviews currently used methods for subtype identification, and discusses challenges to the implementation of precision oncology as well as possible solutions.

The molecular heterogeneity of cancer and the importance of molecular subtyping

One of the most important lessons learned from the sequencing of tens of thousands of tumors through early efforts including The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) is that cancer is tremendously diverse at the molecular level [1]. Although there is considerable variation in the degree of inter-tumor heterogeneity in different tumor types, with hematological tumors generally showing less diversity than solid tumors, when looked at closely no two tumors share exactly the same somatic mutation profile, rather like snowflakes they are all unique [2]. The fact that tumors that are similar in terms of both anatomic origin and histologic appearance generally have limited overlap in terms of somatic mutations is highly problematic as this is thought to be one of the main reasons that response to chemotherapy is so variable in solid tumors. Cancer is not just one disease, but hundreds or thousands of different diseases each with different oncogenic drivers [2]. Thus, unlike other common diseases, such as coronary artery disease, where platelet inhibition with aspirin is effective to some degree for essentially all patients, there will be no one unifying treatment for all of cancer. To address this inter-tumor heterogeneity in cancer it is critical to subdivide the current tumor classifications, which are based on anatomic location and microscopic appearance, into smaller, more homogenous subtypes. While this remains a challenging task, there is a growing abundance of molecular data, including multiplexed IHC, DNAseq, RNAseq, proteomic, and epigenetic profiling, as well as emerging spatially resolved and single-cell sequencing data upon which clinically effective subtypes can be built [3]. This review will focus on colorectal cancer (CRC) but challenges and methods to address these challenges are common to other tumor types.

Implications for precision oncology

The need for greater subclassification of tumors is being driven by the rapidly increasing number of available targeted therapies. In 2020 alone, of all 53 new FDA approvals 18 (34%) were targeted anti-cancer therapies; this trend towards the rapid increase in new cancer therapies is expected to continue [4]. The current wealth of targeted therapies is a new development in cancer treatment. The first effective cancer therapies, developed in the 1940's and 1950's were compounds such as nitrogen mustards and antifolates which were generally effective in all rapidly dividing tumors but also toxic to rapidly dividing normal cells such as those in the bone marrow [5]. The next three decades of chemotherapy development produced agents that targeted DNA synthesis, DNA repair, and cell division;

Received: 19 August 2021
Revised: 23 November 2021
Accepted: 24 November 2021

Version of Record published:
9 December 2021

Emerging Topics in Life Sciences (2021) **5** 747–756
https://doi.org/10.1042/ETLS20210212

all targeting general features of cancer (later described as the Hallmark features of cancer), and not specific to the particular driver alterations (mutation, CNV, fusion, etc) of each individual tumor [6]. The revolutionary development of tyrosine kinase inhibitors such as imatinib mesylate for Chronic Myeloid Leukemia (CML) and monoclonal antibodies such as trastuzumab for HER2 expressing breast cancer revealed the tremendous potential of directly inhibiting the oncogenic signaling of a tumor [7]. The early success of these pioneering molecules has led to the explosive growth of targeted cancer therapies, aided by the comprehensive identification of oncogenes [2] and advances in chemistry changing the old definition of what is 'druggable' [8–10]. However, one of the first lessons in the era of targeted therapy was learned from the development of EGFR inhibitors in non-Small Cell Lung Cancer (NSCLC), where thousands of unselected patients were treated with the drugs gefitinib and erlotinib with poor response rates before it was learned that these drugs are most effective in EGFR mutant NSCLC [11,12]. The EGFR experience in NSCLC, and similar in other tumor types, has led to the current paradigm where a predictive biomarker is required to accompany a targeted therapy. In particular, it has been difficult to develop targeted therapies in tumor types such as CRC, that display a high degree of intertumor heterogeneity and lack predictive biomarkers [13].

Predictive biomarkers: single gene approaches

When the age of targeted therapy began in the early 2000s, tumor molecular profiling was quite limited. The early success stories of imatinib and trastuzumab benefitted from unique situations; in CML the BCR-Abl translocation targeted by imatinib is universal [14], in breast cancer IHC testing for HER2 was already performed as a prognostic biomarker [15]. In other tumor types the identification of predictive biomarkers did not occur until after widespread use of the targeted therapy, such as the case of the anti-EGFR antibody cetuximab in CRC. Initially, it was thought that amplification of *EGFR* would predict response to cetuximab, but it was later learned that activating *KRAS* mutations predict complete lack of response [16]. One of the major limitations in the development of clinical biomarkers is the so-called chicken and egg problem — without cohorts with matched tumor molecular data and clinical outcome data it is not possible to discover or validate a biomarker. However, if a molecular test has not been validated as a biomarker, there is little incentive or ability to routinely measure it on tumor specimens. Now that large scale sequencing efforts have mapped out the landscape of somatic mutations and CNV for the common tumor types [2], predictive biomarkers are generally developed in tandem with their associated targeted therapies. Frequently this development is now done in a pan-cancer fashion, meaning that any tumor type, regardless of histology, positive for the biomarker could be treated with the drug. An example of this model is the TRK inhibitor larotrectinib, now approved for any TRK fusion-positive solid tumor [17]. However, oncogenic drivers do not exist in a vacuum, and chemo-genetic relationships are in many cases dependent on factors including cell lineage and the presence of other genomic aberrations [18]. The success of single-agent BRAF inhibition in $BRAF^{V600E}$ melanoma and NSCLC, but subsequent failure in $BRAF^{V600E}$ CRC is a cautionary tale highlighting the importance of considering tumor lineage in biomarker and drug development.

Moving beyond single gene biomarkers: gene expression based molecular subtypes

The initial successes of targeted therapy were primarily limited to situations where an activating mutation or fusion in Gene X served as the biomarker for an inhibitor of X. This approach has been expanded to look at pairs of mutations, particularly in NSCLC but so far this co-mutation approach has yielded only prognostic, not predictive biomarkers [19]. While there are certainly many more Gene X's to be drugged, and the caveats of tissue specificity remain important, in many ways the one gene approach represents the low hanging fruit of precision oncology. Unfortunately, there are many tumors without any classically druggable targets, let alone targets of an FDA approved drug [20]. An orthologous approach to using the mutation or expression status of a single gene is to instead to consider the entire tumor transcriptome [21]. Transcriptomic approaches became popular after technological advances allowed for rapid and cost-effective whole-transcriptome measurement, first with microarrays and later with RNAseq [22]. Initially supervised methods were favored, where patients with known outcomes were separated into groups with good vs. poor outcome and the most differentially expressed genes between the groups identified. This approach has been successful in some cases, such as the OncotypeDx test in breast cancer [23], however supervised approaches are prone to overfitting and thus generated many classifiers that failed validation [24]. In contrast, OncotypeDx Colon did not provide meaningful

stratification and has not been adopted into clinical use [25]. Many factors influence why these supervised approaches have worked in some cases but not others including size, quality, and comprehensiveness of training data, the specific algorithm used, as well as the intrinsic inter- and intra-tumor heterogeneity of the tumor in question.

Rather than trying to identify gene expression patterns that predict drug response, transcriptomic data can also be used for unsupervised clustering of patients. Here the goal is to reduce the patient-to-patient heterogeneity by breaking a large tumor classification into multiple subgroups, each of which is more homogenous than the group as a whole [26,27]. The term 'unsupervised' refers to the fact that no information regarding the patient outcome is used in the clustering, in contrast with supervised where outcomes are known and used to define groups (i.e. responder vs. non-responder). Many different algorithms have been developed to perform unsupervised clustering (reviewed in [26]), but the commonality between them is that they first identify a distance metric, such as the correlation of gene expression profiles, to quantitatively score how similar each tumor is to every other tumor. The algorithm will then assign tumors to subgroups in a way that minimizes the distance within subgroups, essentially putting like and like together. The number of subgroups can either be predefined or separately optimized to achieve the best separation between groups [28].

In CRC in an international collaboration pooled over 3000 samples from six different CRC classification systems and used a Markov cluster algorithm to detect recurring subtype patterns, identifying four robust subtypes. A random forest approach was then used to build a classifier capable of identifying each of the four subtypes [29]. The subtyping method, known as the Consensus Molecular Subtypes (CMS) as it was a consensus from the six prior classification systems, has become a standard in the CRC research community and now serves as a platform to identify effective targeted therapies for each subgroup [30]. The CMS classifier has now been optimized by several different groups using a smaller number of genes, being performed in a CLIA environment and allowing for compatibility with RNAseq generated from FFPE tissues, an important practical consideration to allow for widespread clinical use [31]. As an unsupervised approach, there were no labels used is assigning tumors to a CMS, however the CMS recovered important biological differences between subtypes: CMS1 associated with immune activation, microsatellite instability (MSI-H) as well as *BRAF* mutation, CMS2 associated with up-regulation of the canonical CRC genes WNT and MYC, CMS3 associated with metabolic dysregulation as well as *KRAS* mutation, CMS4 associated with epithelial-to-mesenchymal transition (EMT) and transforming growth factor (TGF)-β signaling [29]. The CMS have also been shown to have significant association with survival in multiple cohorts (prognostic biomarker), with CMS4 tumors having the worst survival in non-metastatic CRC [29]. However, while it is thought that the CMS will also prove useful to predict response to targeted therapies, the use of CMS as a predictive biomarker for anti-EGFR, anti-VEGF or other targeted therapies has not been robustly demonstrated [32,33]. That targeted therapies in CRC are usually given in combination with one or two different cytotoxic chemotherapy combinations has complicated this analysis.

Cancer as network-based disease: utilizing knowledge of cancer networks for molecular subtyping

The cancer phenotype has been succinctly described as the result of dysregulation of several hallmark cancer pathways [6,34]. Although any particular mutation or mutated gene may be a rare event when viewed independently, these rare events converge on a smaller number of protein complexes, signaling cascades, or transcriptional regulatory circuits [35]. Given the complex nature of solid tumors, which often contain more than hundreds of CNVs and ten or more driver mutations [2,36], which act in concert with one another, accounting for all possible mutation combinations is not a tractable problem as it would require far too many subgroups with too few patients in each. One promising method to address the complex and heterogenous nature of cancer genomes is to map oncogenic mutations onto molecular networks [37–39]. Rather than associating genotype with phenotype directly, variations or mutations in genotype are first mapped onto the knowledge of molecular networks; affected subnetworks are then associated with phenotype. Prerequisite for this approach is accurate knowledge of relevant molecular networks, which are known to be tissue and tumor type specific [40]. Application of these network aggregation methods are currently quite limited in CRC; grouping loss-of-function of the genes MLH1, MSH2, MSH6, and PMS2 as MSI-H, and *KRAS*, *NRAS*, *HRAS*, and *BRAF* mutation as non-responsive to anti-EGFR therapy. Network aggregation has been further advanced in other tumor types, one notable example being the concept of 'BRCAness,' which describes tumors without *BRCA1/2*

Emerging Topics in Life Sciences (2021) **5** 747–756
https://doi.org/10.1042/ETLS20210212

mutation that display defects in homologous recombination, presumably due to mutations in other related DNA repair genes such as those in the Fanconi Anemia pathway [41–45].

There are now numerous examples where prior knowledge regarding how genes are organized into complexes and pathways has been used to aid the interpretation of cancer 'omics' data [1,35,46–55]. This approach has been quite successful when applied to transcriptomic data, where rather than looking at the change in expression of individual genes, sets of related genes are evaluated together as a group [56]. This method, known as Gene Set Enrichment Analysis (GSEA), is now frequently used, often in combination with other methods, for cancer subtyping [57]. Similarly, knowledge of network relationships between regulatory genes and the genes they regulate has allowed transcriptomic data to be converted into protein activity scores using an algorithm known as VIPER (Virtual Inference of Protein-activity by Enriched Regulon analysis) [47]. The fact that each activity score is derived from the experimental measurement of many mRNA transcripts makes this approach robust to noise in the underlying transcriptomic data. This regulon approach has been widely applied predict drug sensitivity from transcriptomic data, and recognizing that the regulatory relationships between genes are context dependent, has recently been adapted to work in context-specific fashion [58]. Transcriptomic data, which produces non-zero values for over 18 000 genes, generally needs dimensional reduction with a method such as GSEA or VIPER before use in downstream analysis. Somatic mutation data, where a tumor may have ∼50–100 mutations in the exome, has an opposite problem of being to sparse for unsupervised clustering. Network-based stratification (NBS) is one method to overcome this sparsity that works by mapping a tumor's full somatic mutation profile, then propagating through the network to 'smooth' the profile [59]. This propagation allows for meaningful and robust clustering by grouping patients with similar 'active' regions in the gene interaction network (in place of attempting to do this based on individually mutated genes). These 'network-smoothed' profiles are then clustered into a predefined number of subtypes using the unsupervised technique of non-negative matrix factorization. Similar network-based approaches have been used to identify subnetworks of genes enriched in mutation [60], the subnetworks can then be used like an extended single-gene approach to divide tumors into mutated and wild-type subtypes.

Machine learning to predict drug response for an individual patient without tumor subtyping (n of 1 approach)

A new idea to enable precision oncology is to use a tumor molecular profile to directly predict its' drug sensitivity using machine learning without first aggregating similar tumors together into subtypes. Although sometimes used interchangeably with the term Artificial intelligence (AI), which broadly refers to the science of creating intelligent machines that can simulate human thinking [61,62], machine learning is a specific application of AI that allows machines to learn from data without being programmed explicitly. The basic concept behind machine learning is that a machine can be trained on past data where outcomes are known, learn from the past data connections between inputs and outputs, then make predictions on new data [63]. Machine learning methods have previously been used in cancer going back to the 1980's, but these early efforts focused mostly on detection and diagnosis [64]. Now the rapid expansion of tumor molecular profiles, critical training data for any machine learning approach, as well as advances in machine learning algorithms are for the first time raising the possibility that machine learning could be used to predict drug response for individual patients.

It is beyond the scope of this review to highlight all of the machine learning methods currently being applied to predict drug response, but they can be classified into the broad categories of supervised learning, unsupervised learning, or semi-supervised learning which in a multi-step approach combines elements of both supervised and unsupervised learning [65]. The input data for a machine learning algorithm is often a whole-exome sequence (WES) or whole-transcriptome sequence (WTS) with ∼20 000 genes (called features in machine learning terminology); this presents a problem called the 'curse of dimensionality' which describes the issue that as the number of dimensions increases, the amount of data required to achieve statistical significance also increases [66]. To combat this curse most machine learning methods first perform dimensionality reduction, feature selection, and feature extraction to limit the number of input features (usually genes) to just the most informative subset. In terms of actual learning algorithms, the most commonly used methods are variations of one of these three techniques: Artificial Neural Network (ANN), Support Vector Machines (SVM), and Bayesian Networks [65]. ANN are called such as these networks consist of layers nodes (artificial neurons) linked by edges which form connections similar to synapses; this structure where nodes receive and process

information, then signal other nodes, is similar to the organization of neurons in the human brain. The weights of the edges (connections between nodes) in the ANN are learned on training data, and then the trained network is used to make predictions on new inputs. ANN methods are sometimes referred to as 'deep' learning when the ANN includes many hidden layers of neurons, these layers can be iteratively tuned on training data to set the weights that are best predictive of outcome [67]. Although they can be computationally intense, ANN are rapidly becoming the most commonly used machine learning method for cancer prediction. In addition to the supervised task of predicting drug response, ANN can also be used for unsupervised tasks such learning informative features (which could then be inputs for a supervised task) [68]. SVM, which are similar to elastic nets [69,70], work by mapping input data into a multi-dimensional space, then identifying hyperplanes that separate the inputs into different clusters [71]. Bayesian Networks are built on Bayes theorem, which produces a probability estimation (as opposed to a definitive classification) based on a prior probability which is updated when new information is available [72]. Of note many supervised classifiers will combine multiple different learning algorithms, this ensemble approach can sometimes produce a better predictive performance than any individual algorithm, but at a cost of being more computationally intense [73]. Random Forest classifiers are one such ensemble method; compatible with both regression (predicting a continuous variable, such as survival time) and classification (such as responder vs. non-responder) they have several advantages including being relatively simple to implement, fast to train, and easy to implement in parallel [74–76].

Hierarchical models: a flexible approach to tumor subtyping

The primary need for subtyping of cancer is to facilitate the effective delivery of targeted therapy by identifying subsets of patients who will respond to a given therapy from a larger cohort. Thus a subtyping method should be flexible so that it can work with many different drugs and the incorporation of new information. One way to build in such flexibility the use of a hierarchical model, groups of tumors can progressively subdivided until a group exists that shows nearly universal response to a given targeted agent (Figure 1). Hierarchical models have previously been proposed as a way to model the subsystems of the cell, which are known to consist of proteins which form together in complexes, which further aggregate into pathways and organelles, etc [77].

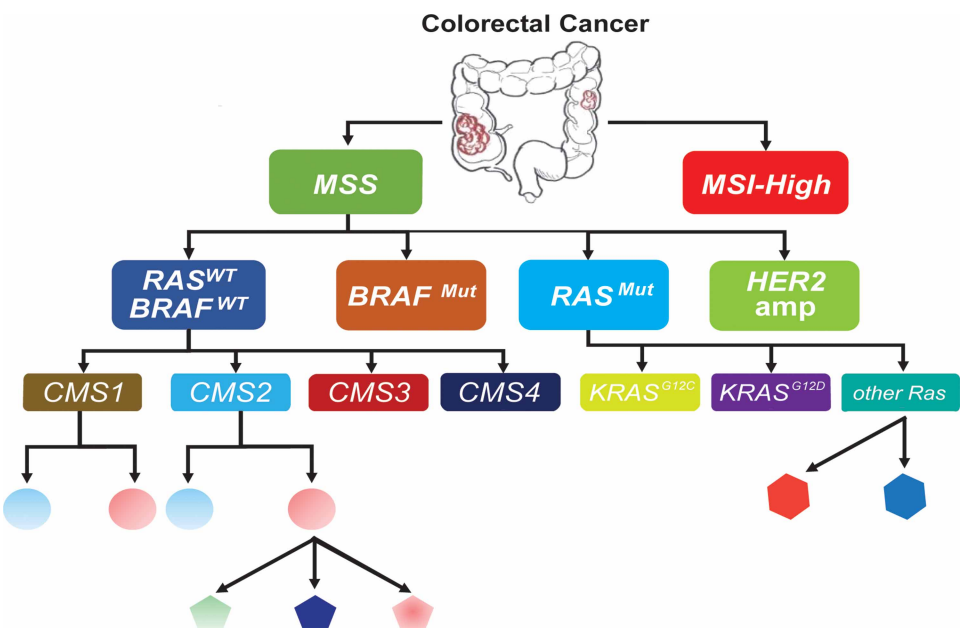

Figure 1. Example hierarchical model of CRC subtypes.

Example of a hierarchical characterization, tumors are first split into microsatellite stable (MSS) and microsatellite unstable (MSI-H), MSS tumor are further split by Ras mutational status, Ras wild-type tumors are further split by transcriptional subtype (CMS), KRAS mutant tumors are split by specific mutant allele. Future subdivisions are yet to be discovered, indicated by unlabeled nodes.

Emerging Topics in Life Sciences (2021) **5** 747–756
https://doi.org/10.1042/ETLS20210212

In biological systems these hierarchical relationships have been captured in the Gene Ontology, a manually curated framework cataloging cellular components and their relationships with each other [78]. More recently it has been shown that gene ontologies can build directly from genetic and protein interaction data [79], that that these data driven ontologies can be used to aggregate multiple somatic mutations into 'ontotypes,' which can then be used to predict phenotypes such as drug response [51].

Gene ontologies, either manually curated or generated from interaction data, can also be used in combination with neural network machine learning approaches to provide mechanistic insight to the AI predictions [49]. By fusing the nodes of a multilayered neural network with a gene ontology, each node in the neural network takes on a biological meaning (gene, pathway, etc). In the model system *S. cerevisiae* this Deep Cell method has been shown to robustly predict lethal genetic interactions as well as identifying the key pathways mediating the interaction. This 'white box' machine learning method has been recently been used to predict drug response in cell lines from the chemical structure of a drug, importantly because nodes in the neural network correspond to genes, synergistic drug combinations were identified using the network weights [80].

Performance assessment and validation

Just as proper positive and negative controls are critical for interpretation of wet lab experiments, controls are similarly important in computational experiments. Common controls for network-based experiments include scrambling the gene labels but preserving the network structure (which should degrade any signal boost provided by the network) or first performing the analysis on simulated data [1]. Performance assessment of a classifier is generally performed by determining the sensitivity and specificity of the classifier against a known gold standard; the overall performance can be best summarized generating an Area Under the Curve (AUC) plot. Ideally, external dataset(s) are available for validation, external validation protects against discovery of features specific to one cohort. If an independent dataset is not available methods to partition the data into training and test cohorts include holdout, random sampling, cross-validation, and bootstrapping [65]. It is important to note that the accuracy of a predictive or classification algorithm may not translate into clinical utility. For clinical applications, assessment of clinical impact as measured by patient outcome such as overall survival, progression free survival, or objective response rate remain the most important metric by which to judge any biomarker or classification algorithm [81]. Ideally this validation of clinical impact is performed prospectively, however, given the difficulty of executing prospective clinical trials another attractive approach is prospective–retrospective analysis, in which new molecular data is generated on archival samples from large clinical trials where the patient outcomes are known [82].

Current challenges and potential solutions for AI in molecular subtyping and precision oncology

Although the promise of using molecular subtyping to achieve precision oncology is currently more achievable today than at any time in the past, significant obstacles still remain. While the idea of molecular subtyping was created to address the problem of inter-tumor heterogeneity, advances in single cell sequencing have revealed that intra-tumor heterogeneity is another major issue to be addressed [83–85]. Intra-tumor heterogeneity is particularly challenging with transcriptomic measurements, where contribution from non-tumor stroma and immune cells can confound analysis of tumor intrinsic transcription and also increase risk of spatial sampling bias [86]. Fortunately computational methods are being developed to deconvolute bulk transcriptomic data, including algorithms that can go beyond simply estimating the percentage contribution of tumor and non-tumor elements [87–92], to actually generate separate profiles for tumor, immune and stromal cells [93].

Although the cost of generating tumor molecular profiles has decreased, the number of tumors that undergo molecular profiling is still quite small relative to ∼1.9 million new cancer diagnoses each year in the United States [94]. Recent technical improvements in RNA sequencing now allow for high-quality, Clinical Laboratory Improvement Amendments (CLIA) certified transcriptomic measurements from formalin-fixed, paraffin embedded (FFPE) tissue [95], which will greatly expand the number of tumors that undergo transcriptional profiling. Additionally collaborative efforts to aggregate molecular data, most notably Project GENIE, are gaining traction and now include more than 120 000 sequenced tumors [96]. Patient driven efforts to share molecular data have also been successful and are expanding [97]. The increasing utility of tumor molecular profiling should also drive oncologists to order profiling on a greater number of patients, fueling an exponential increase in the amount of available data. For the purpose of tumor subtyping tumor molecular is most valuable

when it is paired with clinical outcomes data, however historically molecularly profiled patient cohorts have had limited clinical annotation [98]. Even when clinical data is available, it is not always easy to compress a complex clinical history into a machine-readable outcome measurement, and important prerequisite for machine learning approaches.

Fortunately, new subtyping and drug response prediction methods are being developed which will complement the rapidly expanding amount of tumor molecular data. The incorporation of prior knowledge of genetic network relationships, long known to be an effective strategy in computational biology, will improve as the quality of the networks improve [99]. Cellular networks are known to be context (lineage, metabolic state, etc) dependent; the continued generation of interaction data will allow for the generation of networks to fit each context [100,101]. Just as genomic and transcriptomic data is being aggregated in publicly available repositories, NDEx (https://www.ndexbio.org) has been established as a repository specifically for networks [102]. So called 'Few Shot' machine learning methods are designed to make predictions based on very little training data. This approach has been successfully applied in a pilot study, learning the connections in a neural network on cell line data, then rapidly re-weighting the edges when transitioning to a difference context, like PDX [103].

Conclusions

Given the heterogeneity of cancer, methods to subdivide tumors into more homogeneous subgroups are critical for the success of targeted cancer therapies. Currently, still in its infancy, the increasing availability of tumor molecular data and the continuous development of new computations tools will drive tumor subtyping and with it precision oncology in the years to come.

Summary

- The development of targeted anti-cancer therapies drives a need to divide tumors into more homogenous subtypes.
- Many different molecular data types, including mutation, gene expression, methylation, and proteomics can be used for subtyping.
- Both supervised and unsupervised methods have been used with success.
- New methods propose to use machine learning to identify the best therapy for an individual patient using molecular data (n of 1 approach) without subtyping.

Competing Interests

The author declares that there are no competing interests associated with this manuscript.

Acknowledgements

This work was supported by the National Cancer Institute (L30 CA171000 and K22 CA234406 to J.P.S., and The Cancer Center Support Grant P30 CA016672), the Cancer Prevention & Research Institute of Texas (RR180035 to J.P.S., J.P.S. is a CPRIT Scholar in Cancer Research), and the Col. Daniel Connelly Memorial Fund.

Abbreviations

AI, artificial intelligence; ANN, artificial neural network; CLIA, clinical laboratory improvement amendments; CML, chronic myeloid leukemia; CMS, consensus molecular subtypes; CRC, colorectal cancer; FFPE, formalin-fixed, paraffin embedded; GSEA, gene set enrichment analysis; MSS, microsatellite stable; NSCLC, non-small cell lung cancer; SVM, support vector machines.

References

1 Hofree, M., Shen, J.P., Carter, H., Gross, A. and Ideker, T. (2013) Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 https://doi.org/10.1038/nmeth.2651

2 Bailey, M.H. et al. (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e318 https://doi.org/10.1016/j.cell.2018.02.060
3 Ghandi, M. et al. (2019) Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 https://doi.org/10.1038/s41586-019-1186-3
4 Mullard, A. (2021) 2020 FDA drug approvals. *Nat. Rev. Drug Discov.* **20**, 85–90 https://doi.org/10.1038/d41573-021-00002-0
5 Schirrmacher, V. (2019) From chemotherapy to biological therapy: a review of novel concepts to reduce the side effects of systemic cancer treatment (Review). *Int. J. Oncol.* **54**, 407–419 https://doi.org/10.3892/ijo.2018.4661
6 Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 https://doi.org/10.1016/j.cell.2011.02.013
7 Druker, B.J. et al. (1996) Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat. Med.* **2**, 561–566 https://doi.org/10.1038/nm0596-561
8 Henley, M.J. and Koehler, A.N. (2021) Advances in targeting 'undruggable' transcription factors with small molecules. *Nat. Rev. Drug Discov.* **20**, 669–688 https://doi.org/10.1038/s41573-021-00199-0
9 Schapira, M., Calabrese, M.F., Bullock, A.N. and Crews, C.M. (2019) Targeted protein degradation: expanding the toolbox. *Nat. Rev. Drug Discov.* **18**, 949–963 https://doi.org/10.1038/s41573-019-0047-y
10 Dang, C.V., Reddy, E.P., Shokat, K.M. and Soucek, L. (2017) Drugging the 'undruggable' cancer targets. *Nat. Rev. Cancer* **17**, 502–508 https://doi.org/10.1038/nrc.2017.36
11 Armour, A.A. and Watkins, C.L. (2010) The challenge of targeting EGFR: experience with gefitinib in nonsmall cell lung cancer. *Eur. Respir. Rev.* **19**, 186 https://doi.org/10.1183/09059180.00005110
12 Thatcher, N. et al. (2005) Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (Iressa survival evaluation in lung cancer). *Lancet* **366**, 1527–1537 https://doi.org/10.1016/S0140-6736(05)67625-8
13 Muzny, D.M. et al. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 https://doi.org/10.1038/nature11252
14 Thompson, P.A., Kantarjian, H.M. and Cortes, J.E. (2015) Diagnosis and treatment of chronic myeloid leukemia in 2015. *Mayo Clin. Proc.* **90**, 1440–1454 https://doi.org/10.1016/j.mayocp.2015.08.010
15 Delaney, P. (1999) HER-2: the making of herceptin, a revolutionary treatment for breast cancer. *J. Natl Cancer Inst.* **91**, 1329–1330 https://doi.org/10.1093/jnci/91.15.1329
16 Lièvre, A. et al. (2006) KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Res.* **66**, 3992 https://doi.org/10.1158/0008-5472.CAN-06-0191
17 Hong, D.S. et al. (2020) Larotrectinib in patients with TRK fusion-positive solid tumours: a pooled analysis of three phase 1/2 clinical trials. *Lancet Oncol.* **21**, 531–540 https://doi.org/10.1016/S1470-2045(19)30856-3
18 Nagaraja, A.K. and Bass, A.J. (2015) Hitting the target in BRAF-mutant colorectal cancer. *J. Clin. Oncol.* https://doi.org/10.1200/jco.2015.63.7793
19 Arbour, K.C. et al. (2018) Effects of co-occurring genomic alterations on outcomes in patients with KRAS-mutant non-small cell lung cancer. *Clin. Cancer Res.* **24**, 334–340 https://doi.org/10.1158/1078-0432.ccr-17-1841
20 Martínez-Jiménez, F. et al. (2020) A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 https://doi.org/10.1038/s41568-020-0290-x
21 Califano, A. and Alvarez, M.J. (2017) The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat. Rev. Cancer* **17**, 116–130 https://doi.org/10.1038/nrc.2016.124
22 Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 https://doi.org/10.1038/nrg2484
23 Paik, S. et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 https://doi.org/10.1056/NEJMoa041588
24 Koscielny, S. (2010) Why most gene expression signatures of tumors have not been useful in the clinic. *Sci. Transl. Med.* **2**, 14ps12 https://doi.org/10.1126/scitranslmed.3000313
25 Chee, C.E. and Meropol, N.J. (2014) Current status of gene expression profiling to assist decision making in stage II colon cancer. *Oncologist* **19**, 704–711 https://doi.org/10.1634/theoncologist.2013-0471
26 Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **31**, 651–666 https://doi.org/10.1016/j.patrec.2009.09.011
27 Kiselev, V.Y., Andrews, T.S. and Hemberg, M. (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 https://doi.org/10.1038/s41576-018-0088-9
28 Yeang, C.H. et al. (2001) Molecular classification of multiple tumor types. *Bioinformatics (Oxford, England)* **17**, S316–S322 https://doi.org/10.1093/bioinformatics/17.suppl_1.s316
29 Guinney, J. et al. (2015) The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 https://doi.org/10.1038/nm.3967
30 Dienstmann, R. et al. (2017) Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. Cancer* **17**, 79–92 https://doi.org/10.1038/nrc.2016.126
31 Morris, J.S. et al. (2021) Development and validation of a gene signature classifier for consensus molecular subtyping of colorectal carcinoma in a CLIA-certified setting. *Clin. Cancer Res.* **27**, 120–130 https://doi.org/10.1158/1078-0432.ccr-20-2403
32 Stintzing, S. et al. (2019) Consensus molecular subgroups (CMS) of colorectal cancer (CRC) and first-line efficacy of FOLFIRI plus cetuximab or bevacizumab in the FIRE3 (AIO KRK-0306) trial. *Ann. Oncol* **30**, 1796–1803 https://doi.org/10.1093/annonc/mdz387
33 Sveen, A. et al. (2018) Colorectal cancer consensus molecular subtypes translated to preclinical models uncover potentially targetable cancer cell dependencies. *Clin. Cancer Res.* **24**, 794–806 https://doi.org/10.1158/1078-0432.CCR-17-1234
34 Vogelstein, B. and Kinzler, K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.* **10**, 789–799 https://doi.org/10.1038/nm1087
35 Shen, J.P. and Ideker, T. (2018) Synthetic lethal networks for precision oncology: promises and pitfalls. *J. Mol. Biol.* **430**, 2900–2912 https://doi.org/10.1016/j.jmb.2018.06.026
36 Sanchez-Vega, F. et al. (2018) Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173**, 321–337.e310 https://doi.org/10.1016/j.cell.2018.03.035

37 Wang, K., Li, M. and Hakonarson, H. (2010) Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11**, 843–854 https://doi.org/10.1038/nrg2884
38 Kim, Y.A. and Przytycka, T.M. (2012) Bridging the Gap between genotype and phenotype via network approaches. *Front. Genet.* **3**, 227 https://doi.org/10.3389/fgene.2012.00227
39 Califano, A., Butte, A.J., Friend, S., Ideker, T. and Schadt, E. (2012) Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 https://doi.org/10.1038/ng.2355
40 Zhao, J., Cheng, F. and Zhao, Z. (2017) Tissue-specific signaling networks rewired by major somatic mutations in human cancer revealed by proteome-wide discovery. *Cancer Res.* **77**, 2810 https://doi.org/10.1158/0008-5472.CAN-16-2460
41 Turner, N., Tutt, A. and Ashworth, A. (2004) Hallmarks of BRCAness in sporadic cancers. *Nat. Rev. Cancer* **4**, 814 https://doi.org/10.1038/nrc1457
42 Lord, C.J. and Ashworth, A. (2016) BRCAness revisited. *Nat. Rev. Cancer* **16**, 110 https://doi.org/10.1038/nrc.2015.21
43 Konstantinopoulos, P.A. et al. (2010) Gene expression profile of BRCAness that correlates With responsiveness to chemotherapy and With outcome in patients with epithelial ovarian cancer. *J. Clin. Oncol.* **28**, 3555–3561 https://doi.org/10.1200/JCO.2009.27.5719
44 Bast, Jr, R.C. and Mills, G.B. (2010) Personalizing therapy for ovarian cancer: BRCAness and beyond. *J. Clin. Oncol.* **28**, 3545–3548 https://doi.org/10.1200/jco.2010.28.5791
45 Swisher, E.M. et al. (2017) Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma (ARIEL2 Part 1): an international, multicentre, open-label, phase 2 trial. *Lancet Oncol.* **18**, 75–87 https://doi.org/10.1016/s1470-2045(16)30559-9
46 Alvarez, M.J. et al. (2016) Network-based inference of protein activity helps functionalize the genetic landscape of cancer. *Nat. Genet.* **48**, 838–847 https://doi.org/10.1038/ng.3593
47 Alvarez, M.J. et al. (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 https://doi.org/10.1038/ng.3593
48 Wang, S. et al. (2018) Typing tumors using pathways selected by somatic evolution. *Nat. Commun.* **9**, 4159 https://doi.org/10.1038/s41467-018-06464-y
49 Ma, J. et al. (2018) Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **15**, 290–298 https://doi.org/10.1038/nmeth.4627
50 Kramer, M., Dutkowski, J., Yu, M., Bafna, V. and Ideker, T. (2014) Inferring gene ontologies from pairwise similarity data. *Bioinformatics (Oxford, England)* **30**, i34–i42 https://doi.org/10.1093/bioinformatics/btu282
51 Yu, M.K. et al. (2016) Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell Syst.* **2**, 77–88 https://doi.org/10.1016/j.cels.2016.02.003
52 Yu, M.K. et al. (2018) Visible machine learning for biomedicine. *Cell* **173**, 1562–1565 https://doi.org/10.1016/j.cell.2018.05.056
53 Kuenzi, B.M. and Ideker, T. (2020) A census of pathway maps in cancer systems biology. *Nat. Rev. Cancer* **20**, 233–246 https://doi.org/10.1038/s41568-020-0240-7
54 Bouhaddou, M. et al. (2019) Mapping the protein-protein and genetic interactions of cancer to guide precision medicine. *Curr. Opin. Genet Dev.* **54**, 110–117 https://doi.org/10.1016/j.gde.2019.04.005
55 Ideker, T., Dutkowski, J. and Hood, L. (2011) Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* **144**, 860–863 https://doi.org/10.1016/j.cell.2011.03.007
56 Subramanian, A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. U.S.A.* **102**, 15545 https://doi.org/10.1073/pnas.0506580102
57 Gao, F. et al. (2019) DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis* **8**, 44 https://doi.org/10.1038/s41389-019-0157-8
58 Broyde, J. et al. (2021) Oncoprotein-specific molecular interaction maps (SigMaps) for cancer network analyses. *Nat. Biotechnol.* **39**, 215–224 https://doi.org/10.1038/s41587-020-0652-7
59 Raphael, B.J. (2013) Making connections: using networks to stratify human tumors. *Nat. Methods* **10**, 1077–1078 https://doi.org/10.1038/nmeth.2704
60 Reyna, M.A., Leiserson, M.D.M. and Raphael, B.J. (2018) Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics (Oxford, England)* **34**, i972–i980 https://doi.org/10.1093/bioinformatics/bty613
61 McCarthy, J. (1998) What is artificial intelligence?
62 Turing, A.M. and Haugeland, J. (1950) *Computing Machinery and Intelligence*, MIT Press Cambridge, MA.
63 Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E. (2006) Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.* **26**, 159–190 https://doi.org/10.1007/s10462-007-9052-3
64 Cruz, J.A. and Wishart, D.S. (2006) Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2**, 117693510600200030 https://doi.org/10.1177/117693510600200030
65 Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I. (2015) Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17 https://doi.org/10.1016/j.csbj.2014.11.005
66 Bellman, R. and Rand Corporation. (1957) *Dynamic Programming*, Princeton University Press.
67 Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C. and Collins, J.J. (2018) Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 https://doi.org/10.1016/j.cell.2018.05.015
68 Yuan, B., Yang, D., Rothberg, B.E.G. Chang, H. and Xu, T. (2020) Unsupervised and supervised learning with neural network for human transcriptome analysis and cancer diagnosis. *Sci. Rep.* **10**, 19106 https://doi.org/10.1038/s41598-020-75715-0
69 Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**, 301–320 https://doi.org/10.1111/j.1467-9868.2005.00503.x
70 Zhou, Q. et al. (2015) *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence 3210–3216*, AAAI Press, Austin, Texas.
71 Huang, S. et al. (2018) Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics* **15**, 41–51
72 Snoek, J., Larochelle, H. and Adams, R.P. (2012) Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inform. Process. Syst.* **25**
73 Rokach, L. (2009) Ensemble-based classifiers. *Artif. Intell. Rev.* **33**, 1–39 https://doi.org/10.1007/s10462-009-9124-7

74 Lind, A.P. and Anderson, P.C. (2019) Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PLoS ONE* **14**, e0219774 https://doi.org/10.1371/journal.pone.0219774
75 Fawagreh, K., Gaber, M.M. and Elyan, E. (2014) Random forests: from early developments to recent advancements. *Syst. Sci. Control Eng.* **2**, 602–609 https://doi.org/10.1080/21642583.2014.956265
76 Zhang, C. and Ma, Y. (2012) *Ensemble Machine Learning*, Springer.
77 Carvunis, A.-R. and Ideker, T. (2014) Siri of the cell: what biology could learn from the iPhone. *Cell* **157**, 534–538 https://doi.org/10.1016/j.cell.2014.03.009
78 Ashburner, M. et al. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**, 25–29 https://doi.org/10.1038/75556
79 Dutkowski, J. et al. (2013) A gene ontology inferred from molecular networks. *Nat. Biotechnol.* **31**, 38–45 https://doi.org/10.1038/nbt.2463
80 Kuenzi, B.M. et al. (2020) Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684.e676 https://doi.org/10.1016/j.ccell.2020.09.014
81 Mandrekar, S.J. and Sargent, D.J. (2010) Predictive biomarker validation in practice: lessons from real trials. *Clin. Trials* **7**, 567–573 https://doi.org/10.1177/1740774510368574
82 Patterson, S.D. et al. (2011) Prospective–retrospective biomarker analysis for regulatory consideration: white paper from the industry pharmacogenomics working group. *Pharmacogenomics* **12**, 939–951 https://doi.org/10.2217/pgs.11.52
83 Vitale, I., Shema, E., Loi, S. and Galluzzi, L. (2021) Intratumoral heterogeneity in cancer progression and response to immunotherapy. *Nat. Med.* **27**, 212–224 https://doi.org/10.1038/s41591-021-01233-9
84 Chowdhury, S. et al. (2021) Implications of intratumor heterogeneity on consensus molecular subtype (CMS) in colorectal cancer. *Cancers* **13**, 4923 https://doi.org/10.3390/cancers13194923
85 Sun, X.X. and Yu, Q. (2015) Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacol. Sin.* **36**, 1219–1227 https://doi.org/10.1038/aps.2015.92
86 Wang, Q. et al. (2017) Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell* **32**, 42–56.e46 https://doi.org/10.1016/j.ccell.2017.06.003
87 Newman, A.M. et al. (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 https://doi.org/10.1038/nmeth.3337
88 Becht, E. et al. (2016) Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 https://doi.org/10.1186/s13059-016-1070-5
89 Li, T. et al. (2020) TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* **48**, W509–W514 https://doi.org/10.1093/nar/gkaa407
90 Plattner, C., Finotello, F. and Rieder, D. (2020) Deconvoluting tumor-infiltrating immune cells from RNA-seq data using quanTIseq. *Methods Enzymol.* **636**, 261–285 https://doi.org/10.1016/bs.mie.2019.05.056
91 Aran, D., Hu, Z. and Butte, A.J. (2017) Xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 https://doi.org/10.1186/s13059-017-1349-1
92 Racle, J. and Gfeller, D. (2020) EPIC: a tool to estimate the proportions of different cell types from bulk gene expression data. *Methods Mol. Biol. (Clifton, N.J.)* **2120**, 233–248 https://doi.org/10.1007/978-1-0716-0327-7_17
93 Wang, Z. et al. (2018) Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *iScience* **9**, 451–460 https://doi.org/10.1016/j.isci.2018.10.028
94 Siegel, R.L., Miller, K.D., Fuchs, H.E. and Jemal, A. (2021) Cancer statistics, 2021. *CA Cancer J. Clin.* **71**, 7–33 https://doi.org/10.3322/caac.21654
95 Pennock, N.D. et al. (2019) RNA-seq from archival FFPE breast cancer samples: molecular pathway fidelity and novel discovery. *BMC Med. Genomics* **12**, 195 https://doi.org/10.1186/s12920-019-0643-z
96 The, A.P.G.C. (2017) AACR project GENIE: powering precision medicine through an international consortium. *Cancer Discov.* **7**, 818 https://doi.org/10.1158/2159-8290.CD-17-0151
97 Wagle, N. et al. (2018) Count me in: a patient-driven research initiative to accelerate cancer research. *J. Clin. Oncol.* **36**, e13501 https://doi.org/10.1200/JCO.2018.36.15_suppl.e13501
98 Liu, J. et al. (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.e411 https://doi.org/10.1016/j.cell.2018.02.052
99 Zhang, W., Ma, J. and Ideker, T. (2018) Classifying tumors by supervised network propagation. *Bioinformatics (Oxford, England)* **34**, i484–i493 https://doi.org/10.1093/bioinformatics/bty247
100 Huang, J.K. et al. (2018) Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.* **6**, 484–495.e485 https://doi.org/10.1016/j.cels.2018.03.001
101 Krogan, N.J., Lippman, S., Agard, D.A., Ashworth, A. and Ideker, T. (2015) The cancer cell map initiative: defining the hallmark networks of cancer. *Mol. Cell* **58**, 690–698 https://doi.org/10.1016/j.molcel.2015.05.008
102 Pratt, D. et al. (2015) NDEx, the network data exchange. *Cell Syst.* **1**, 302–305 https://doi.org/10.1016/j.cels.2015.10.001
103 Ma, J. et al. (2021) Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer* **2**, 233–244 https://doi.org/10.1038/s43018-020-00169-2