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;
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 inter-tumor 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*<sup>V600E</sup> melanoma and NSCLC, but subsequent failure in *BRAF*<sup>V600E</sup> 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 pre-defined 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 ‘BRCAAness,’ which describes tumors without BRCA1/2
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 consists of proteins which form together in complexes, which further aggregate into pathways and organelles, etc [77].

![Figure 1. Example hierarchical model of CRC subtypes.](image-url)

Example of a hierarchical characterization, tumors are first split into microsatellite stable (MSS) and microsatellite unstable (MSI-H), MSS tumors 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.
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 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
https://doi.org/10.1038/nmeth.2651


12 Thatcher, N. et al. (2005) Ge


17 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


