Patient Similarity Networks for Precision Medicine

Shraddha Pai, Gary D. Bader

PII: S0022-2836(18)30532-1
DOI: doi:10.1016/j.jmb.2018.05.037
Reference: YJMBI 65746

To appear in: *Journal of Molecular Biology*

Received date: 14 April 2018
Revised date: 24 May 2018
Accepted date: 29 May 2018

Please cite this article as: Shraddha Pai, Gary D. Bader , Patient Similarity Networks for Precision Medicine. Yjmbi (2018), doi:10.1016/j.jmb.2018.05.037

Patient similarity networks for precision medicine

Authors:
Shraddha Pai[1], Gary D. Bader*[1,2,3,4]

Affiliations:
1. The Donnelly Centre, University of Toronto, Toronto, Canada
2. Department of Molecular Genetics, University of Toronto, Toronto, Canada
3. Department of Computer Science, University of Toronto, Toronto, Canada
4. The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada
* gary.bader@utoronto.ca

Abstract. Clinical research and practice in the 21st century is poised to be transformed by analysis of computable electronic medical records and population-level genome-scale patient profiles. Genomic data captures genetic and environmental state, providing information about heterogeneity in disease and treatment outcome, but genomic-based clinical risk scores are limited. Achieving the goal of routine precision medicine that takes advantage of this rich genomics data will require computational methods that support heterogeneous data, have excellent predictive performance, and ideally, provide biologically-interpretable results. Traditional machine-learning approaches excel at performance, but often have limited interpretability. Patient similarity networks are an emerging paradigm for precision medicine, in which patients are clustered or classified based on their similarities in various features, including genomic profiles. This strategy is analogous to standard medical diagnosis, has excellent performance, is interpretable, and can preserve patient privacy. We review new methods based on patient similarity networks, including Similarity Network Fusion for patient clustering and netDx for patient classification. While these methods are already useful, much work is required to improve their scalability for contemporary genetic cohorts, optimize parameters, and incorporate a wide range of genomics and clinical data. The coming five years will provide an opportunity to assess the utility of network-based algorithms for precision medicine.

Introduction

Subdividing patients into subgroups homogeneous with respect to biology, disease progression and response to treatment enables “precision medicine”[1]. Although it is a new term, precision medicine is simply traditional medicine that considers more abundant and detailed patient data. It is the idea that an individual patient’s clinical outcome – disease risk, prognosis, and treatment response – is determined by their genetic, genomic, physiological and clinical profile, that corresponds with a known diagnosis. An early example is the use of improved phenotyping in 1971 to recognize that autism spectrum disorder was a different disease than schizophrenia[2]. Achieving the goal of "getting the right treatment to the right person" for a particular clinical outcome thus requires using all available relevant data to identify all possible diagnoses and their disease trajectories, so that appropriate therapy can be recommended.

There are many ways to predict risk for a particular disease (Table 1), though a tool commonly used to achieve this goal is a risk calculator: a mathematical model that converts measures of a set of risk factors into a quantitative estimate that guides clinical monitoring, diagnosis and treatment (Figure 1; Box 1). One of the best-validated risk calculators is the US-based Atherosclerotic Cardiovascular Disease (ASCVD) calculator, which calculates 10-year risk of developing heart disease or stroke for men and women of Caucasian white or African American ancestry aged between 40 and 79 years. Recommended by the American College of Cardiology, this predictor considers 13 pieces of information including a patient's gender, age, blood lipid levels (3 numbers), blood pressure (2 numbers) and basic medical information (e.g. history of diabetes). When used with a clinical assessment, the predictor provides a basis for suggesting lifestyle modifications and initiating clinical interventions. It additionally serves as an objective indicator to both patient and clinician as to the risk of a cardiovascular disease event. The ASCVD is the result of five decades of model development, with its roots in the Framingham Heart Study of the 1950's[3,4]. Despite extensive validation, the model continues to be a work in progress, being updated to accurately estimate risk for other ethnicities common in the US today[5] and to populations in other countries (e.g. the UK[6-8]). Developing a risk model mature enough for clinical decision-making involves several stages, including rounds of internal and external validation and eventually, a clinical trial or the recommended use by a professional body (Fig 1B; e.g. for breast cancer[9]).

Methodologically, risk models are developed using supervised learning algorithms (Box 2, "Key Concepts"). Patient data is encoded as "input features" (e.g. age, gender, genotypes at individual SNPs, metabolite quantities, gene expression levels). A learning or model-fitting algorithm is used to learn a function that accurately maps the features to a predicted outcome. To train the model, patient samples are partitioned into two groups: a training set and a test set. The training set provides examples to the model of what profiles look like for each of the possible outcomes and enables the model to learn predictive patterns. The independent test set is used to measure the classification performance of the model. Feature selection is used to identify the smallest set of the most predictive variables, which can help understand how the prediction is being made and can speed up prediction. A common concern in model-building is overfitting, when the model learns weights based on the particular bias in the training sample and does not generalize to the wider population. Cross-validation is used to estimate model generalizability; here, a portion of the training sample is held out from the learning process and is used to evaluate fitting error on the held out test set, and this is repeated many times. There are several measures for evaluating a model's performance: these include the balance between the specificity and sensitivity of the model (area under the

Receiver Operator Characteristic curve or AUROC; area under the Precision-Recall curve or AUPR); accuracy; F1 score; and Matthews correlation coefficient. Learning methods commonly used in risk models include logistic regression for categorical outcomes (e.g. risk stratification)[10-12] and Cox proportional hazards models for censored data or event-based models (e.g. 10-year risk of developing a disease)[10,13,14]. Evaluating multiple methods to find the best one is an important part of model development because it is known that no method is best for all data sets. The ideal clinical risk model is accurate, generalizable, provides a prediction in a reasonable time frame for clinical decision-making and, we argue, interpretable so that it can be understood by a clinician.

Many risk models are in clinical use, such as for diabetes[12] and prostate cancer[11] (Box 1). Data mining of electronic health records from health care systems is spurring the development of newer models. A notable example is a calculator for neonatal sepsis risk, based on data from ~200,000 infants[15]. While genomics data is now widely available, most risk predictors that do consider genetic information only use it in a general form using family history. The problem with this is that people often do not inherit a risk factor from a parent. As information about a patient continues to grow exponentially due to advances in genomics, medical imaging and other measurement technologies, there is an opportunity to develop accurate predictive risk models for many more diseases and to develop these more quickly, but this requires the development of new analytic methodology that can keep pace with the scale and complexity of the data (Figure 2A).

Integrating genomics into clinical risk models

The predictive value of genomics is driving the generation of multi-omic data and the use of genomics in clinical diagnostics

The past two decades of genomics research has demonstrated the value of genomics data in understanding cellular processes in disease and witnessed the use of ‘omic data in clinical risk prediction models (Figure 2B). The first wave of predictive utility for genomics data was demonstrated at the genetic level by linkage studies and then genome-wide association studies[16]; this research has identified which diseases have a few big risk-effect variants (3X risk increase for each APOe4 allele in Alzheimer’s disease[17,18]; 6.7X risk increase for BRCA1/2 mutations in breast cancer; 6-17X for HLAII alleles in familial type 1 diabetes[19]) and which diseases have smaller, polygenic contributions (1.10-1.20 for schizophrenia[20], coronary artery disease[21], and bipolar disorder[22,23]). The use of exome sequencing technology quadrupled the rate of identification of rare disease-causing genes in two years[24], which enables early diagnosis. The profiling of other ‘omics layers is providing a more complete picture of the relative contribution of different genomic layers to disease risk. Gene-expression based breast cancer risk prediction is a notable example of a genomics-based assay that is in routine clinical use[14,25]. In another example, an ‘omic profiling of ependymoma, a common childhood brain tumour, found no evidence of recurrent somatic mutations; instead, two disease subtypes with different clinical characteristics were distinguished by distinct mRNA expression levels and DNA methylation signatures in CpG islands[26]. More recently, data from The Cancer Genome Atlas project has been used to investigate the relative value of different genomic datatypes in predicting cancer survival for individual tumour types, with good success[27].

With the utility of genomics clearly proven, data collection continues to increase (Figure 2B). Several consortia now exist for genomic profiling at population-scale. Public-sector initiatives in the US include the veteran-focused Million Veterans Program[28] and the US National Institute of Health's "All of Us" targeted to the general US population; both initiatives have the specific goal of accelerating precision medicine. The UK BioBank is a similar initiative supported by multiple government and not-for-profit agencies in the UK[29]. Private-sector projects include the AstraZeneca bid to sequence two million genomes to identify rare genetic variants that affect drug response[30], and the repository of the direct-to-consumer genetic testing service, 23andMe Inc., which are now being used for focused GWAS[31,32]. Crowd-sourcing initiatives such as PatientsLikeMe (http://patientslikeme.come) and patient-led foundations, such as Dragon Masters (https://www.dragonmasterfoundation.org/) are another growing source of genetic and patient data. Initiatives are also going deeper, generating multiple genomics data types for particular diseases. The Cancer Genome Atlas (TCGA) profiled 33 tumour types on 11,000 patients with up to six different genomic data types profiled for a given tumour type (http://cancergenome.nih.gov/abouttcga), the Alzheimer's Disease Neuroimaging Initiative (ADNI)[33] collects genetics and brain imaging data, STARNET[34] collects genetic and gene expression data for cardiovascular risk, and the Lundbeck study collects genetic data to study treatment response in major depression; see Box 1 for more information.

Continued investment in data collection is driving the technology, infrastructure, reduced costs and logistics necessary to achieve a future in which clinical decision-making routinely incorporates genomic data. Historically, improvements in the depth of patient data have led to a substantial improvement in patient care and new large-scale phenotype data are already leading to more precise care. For example, breast cancer is now treated by subtype (e.g. ER+, HER2+, triple negative) using targeted drugs that outperform the older one-size-fits-all therapies[35].

Most genomic risk models are currently genetic, with limited use of other 'omics data types

With its established model of heritability and increasing cost-effectiveness, genetic data is understandably the single most common genomic data source used in contemporary clinical prediction models (Figure 2B). Associated mutations can have a wide range of effect sizes and this influences the number of genetic variables included in a risk model. For instance, BRCA1/2 mutations predict an up to 80% lifetime risk of breast cancer, and clinical models use this and a variety of other information types to predict risk[13]. At the other end of the spectrum are highly polygenic disorders, where hundreds of common variants contribute small effects to disease risk; these include cardiovascular disease, schizophrenia and bipolar disorder[22,23,34,36,37]. A polygenic risk score is typically used to model these weak contributions, where an individual's genetic risk is the sum of all the risk alleles carried by that individual, weighted by the significance of the corresponding allele in a GWAS study[23,38-40]. Another strategy is to extend this idea to include all genomic SNPs regardless of their marginal disease association, as an overall estimate of genetic contribution to the phenotype[41-43]. Unsurprisingly, the predictive value of genome-wide SNP data is greater in diseases with more polygenic architecture. Celiac disease, characterized by autoimmune reactivity to ingested gluten, has multiple interacting genetic contributors both within and outside the HLA region of the genome[44]. In this scenario, models that included half a million SNPs and considered inter-SNP correlations outperformed those that relied on a genetic risk score limited to significant SNPs from GWAS studies[43,45,46]. Increasing the number of SNPs has increased predictive power in celiac disease (e.g. from AUC of 0.82 with HLA-based GWAS hits to 0.90 when considering all SNPs in the genome[43,45,46]). Additional types of genomics data have also been successfully used. For example, multiple commercial diagnostic tests of breast cancer are available that use gene expression (Mammaprint, Oncotype DX, ProSigna) or protein expression by immunohistochemistry (MammoStrat, IHC4)[14,25,47-53]. The Oncotype DX test uses the expression of 21 genes to calculate the risk of recurrence and chemotherapy response in ER+, node-negative breast cancers[48] and can identify a greater than fourfold difference in recurrence risk between low and high risk score groups[9,14].

Clinical models must be interpretable

While excellent predictive power is ultimately the goal, model interpretability is valuable for multiple reasons. First, understanding the mechanism for how specific variables relate to outcome is useful to gain confidence in the generalizability of the method, especially with smaller data sets that cannot support high confidence statistical predictions using "black box" methods. Second, a transparent model can help us understand the causal molecules or processes underlying a clinical condition that can then be targeted for rational treatment design. Popular machine learning methods provide different levels of out-of-the-box interpretability. Support vector machines provide individual feature weights as output; interpretation of these weights, especially when decision boundaries are nonlinear, requires additional domain-specific method development (e.g. in neuroimaging[54,55]). Random forests are considered more interpretable than SVMs because they provide an explicit decision tree of successive choices used in classification. More recently, popular software libraries of machine-learning methods have been extended to expose the rules used by the algorithms to learn decision boundaries (sklearn in Python and inTrees in R). But in general, machine learning methods are typically non-trivial to interpret mechanistically.

Patient similarity networks as a framework for clinical prediction

The patient similarity network paradigm is a recently developed analytical framework that addresses a number of challenges in data analytics and is naturally interpretable.

The patient similarity network paradigm

In a patient similarity network, each node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given feature. In this paradigm, each input patient data feature (e.g. age, sex, mutation status) is represented as a network of pairwise patient similarities (Figure 3). Each feature is represented as a different "view" of patient similarity that can be integrated with all the other views to identify patient subgroups or predict outcome. As a simple example of the concept, we can represent smoking frequency as a patient similarity network (Figure 3). Patients (nodes) who are frequent smokers would be tightly connected to each other and 'never smokers' would separately be tightly connected. This network is highly predictive of lung cancer status, as lung cancer cases would be enriched in the 'smoker' network region, while healthy controls would be enriched in the 'never smoker' region. If a new patient is a 'never smoker', they would be more similar to healthy controls and a classifier would predict them as such.

Advantages of Patient Similarity Networks

The patient similarity network (PSN) framework enables classifiers that are accurate, generalizable, able to integrate heterogeneous data, and naturally handle missing information. This paradigm also provides excellent model interpretability and additionally, may be better suited to protect patient privacy than most established machine learning methods.

The PSN paradigm is novel; at the time of this writing, only two methods have used this framework for patient clustering[56,57] and only one, netDx, for supervised classification[58]. When compared to other clustering and classification approaches, these methods can demonstrate superior performance (see details in the following sections). As a clustering algorithm, Similarity Network Fusion (SNF) has been used to identify clinically homogeneous patient subgroups in multiple cancers by integrating gene expression, miRNA and DNA methylation data. Using a different approach, PSNs were used to discover new subtypes of type 2 diabetes, each with distinctive genetic enrichments and clinical features[56,57]. Supervised applications have also been able to accurately predict patient survival given genomic data from various cancers.

PSNs naturally handle heterogeneous data, as any data type can be converted into a similarity network by defining a similarity measure. Once converted, all data is represented in the same manner, as a network that can be directly input into analysis methods. Missing data is also naturally handled, as a patient missing in one network may be in another and could still be used. Further, patient similarity measures, like Pearson correlation, are robust even if part of the input data vectors are missing.

Representing patients by similarity is conceptually intuitive because it can convert the data into network views where the decision boundary can be visually evident. As a clinical research tool, it is conceptually analogous to clinical diagnosis, which often involves a physician relating a patient to a mental database of similar patients they have seen. Feature engineering can help further improve interpretability. For instance, creating features at the level of biological pathways helps identify cellular processes that may be causal mechanisms for a given patient subgroup phenotype.

Algorithms that take patient similarity networks as input have the added advantage that the data have been transformed from the raw values and thus sensitive raw data need not be directly used. As the research community increasingly pools its patient cohorts to increase sample sizes for clinical discovery, protocols and technologies for maintaining patient privacy have been in parallel development (https://beacon-network.org, encryption[59]). Sharing PSNs enables clustering and classification applications without the need to share sensitive raw patient data.

Patient similarity networks for clustering

Clustering or class discovery is an important precursor to supervised learning algorithms. Class discovery can help identify patient labels or subtypes based on homogenous molecular signatures. A classifier could then be built for each of the subtypes, or multi-task learning could be used to build a single multi-way classifier[60]. Two PSN-based clustering methods have been reported to date. The first identified subgroups of type 2 diabetes patients using 73 clinical variables obtained from electronic medical records of ~11,000 patients[56]. Networks were generated using singular value decomposition and cosine similarity, the latter being a popular similarity metric in text mining applications. Using medical records and genotype data on the same individuals, the authors demonstrated that identified patient clusters were enriched for different comorbidities and biological pathways. In Similarity Network Fusion (SNF), a patient similarity network is generated from each input data type; for continuous-valued measures, similarity is based on Euclidean distance followed by exponential scaling[57]. The set of networks is then fused by iteratively boosting – or increasing – the weights of edges that are concordant among different layers, and dampening – or decreasing – the weights of those that are only present in some but not all layers. Spectral clustering is then applied to “cut” the final network into highly-interconnected clusters.

SNF performance was benchmarked against naïve integrative clustering – namely, data concatenation – and a method based on joint latent variable models. Patient subgroups were identified in five tumours by integrating mRNA expression, DNA methylation and miRNA expression[57,61]. SNF significantly outperformed the other approaches in identifying clinically-distinct clusters in all cases, and demonstrated consistent fast algorithm run times regardless of the number of genes included in the input data[57]. Since its development, SNF has been used in various applications, including subtyping medulloblastoma patients from DNA methylation and gene expression, and clustering pancreatic ductal adenocarcinoma tumours from RNA, DNA methylation and miRNA expression[62,63].

netDx: Patient classification by similarity networks

We recently developed netDx, a supervised machine learning method for patient classification, based on the patient similarity network paradigm (submitted)[58]. The workflow starts with the definition of a classification problem, such as "predict patients that respond to a drug". A cohort of patients containing positive and negative examples for the classification problem (i.e. cases and controls), and associated patient-level data is required as input. Each available data feature (e.g. patient age, gene expression profile) is converted into a patient similarity network. The resulting patient similarity networks are directly used as input for netDx. In the feature selection and training phase, netDx uses a machine learning algorithm to identify which input networks best characterize each patient category (e.g. cases vs. controls), and builds an optimal predictor from these features. Samples are partitioned into a training set, which is used to score input networks based on their discriminative power, and a test set, which is used to validate the predictor created on the training set. This step identifies a set of selected features (networks) per patient class, that best capture the similarity for that class. Standard cross-validation methods (e.g. 10-fold) are used to estimate generalization performance across different subsets of training and test samples. The final phase is predictor validation. A 'blind test' set, held out before the feature selection and training phase (i.e. not used in predictor training), is used to test the predictor. This process is repeated many times with different cohort splits to increase the strength of the generalization estimate and to optimize the feature selection. The optimized predictor (i.e. using selected networks) is used to score and rank each patient for each of the classes. The patient is then assigned to the class with the highest similarity score. netDx output includes a list of all networks and their prediction value, various predictor performance measures, and an overall patient similarity network integrating all feature-selected networks, which can be visualized and interpreted.

netDx relies on the GeneMANIA classifier originally developed for gene function prediction, which demonstrated excellent accuracy, generalization and ability to integrate heterogeneous data in this task. For example, GeneMANIA outperformed previous models in predicting mouse gene function by integrating gene expression, protein sequence data, protein interactions, phenotypes, conservation, and disease annotation[64,65]. netDx adapts the GeneMANIA algorithm to classify patients instead of genes. GeneMANIA scores each input network based on how well it can classify an input set of patients known to be in the same class (the query; e.g. all patients non-responsive to a medication). The ideal network for classification would perfectly connect all input patients to each other in a clique and would not connect to any other patients outside of the input list. This network would support perfect classification, since any artificially held out patient would perfectly connect only to other patients in the same class. GeneMANIA will weight such a network highly (i.e.

1.0). On the other hand, a network that does not connect any of the input patients to each other is not useful for classification. GeneMANIA will assign a low weight to such a network (i.e. 0.0). Almost all real networks are expected to be between these two extremes and get scored accordingly. Weighting is accomplished by representing the input networks as a single matrix of patient edges, and by applying ridge regression to this single matrix[65]. Once all input networks are weighted based on their informativeness for classification of input patients, a linear combination of networks is used to create a composite network. Label propagation is used to score all non-input patients based on similarity to the input patients. Label propagation uses the edge weights in the composite network, and assigns node discriminant values by solving a sparse linear system with a global minimum. In netDx, the process is repeated for each known patient class (e.g. cases and controls), and patients are assigned to the class that they are closest to.

Compared to other machine-learning methods used for classification, netDx demonstrates consistently excellent performance. Using a benchmark data set for predicting binary cancer survival in four tumours, netDx was compared to a panel of eight popular machine-learning methods, such as support vector machines and random forests. This prediction task required the integration of up to six data types including clinical data, mRNA expression, miRNA expression, DNA methylation, somatic copy number aberrations and proteomic profiles[58]. On average, netDx significantly outperforms other approaches for three of four tumours, and is at par for the fourth tumour; moreover, its top model outperforms all other models for two of the tumours. Therefore, as a machine-learning algorithm, the PSN-based netDx can perform as well as or better than standard machine-learning approaches.

In a feature not readily available in other machine-learning methods, netDx can also be used to provide mechanistic insight by grouping gene-level features into pathway-level features. When predicting breast cancer subtype from gene expression, netDx correctly feature selects pathways related to DNA damage repair and cell cycle progression. In contrast, when predicting case/control status in asthma, feature-selected themes reflect cellular processes involved in inflammation[58]. These different themes highlight netDx's ability to identify cellular processes that reflect the particular biology of the condition under study. Grouping variables at the pathway level provides two major advantages. Feature selected pathways provide mechanistic insight into differences between classified patient groups. Second, pathways help address sparse data. For instance, somatic mutations may not provide enough information to compute patient similarity (e.g. patients may not have mutations in common). Merging these into pathways increases the chances that patients will have mutated pathways in common and thus can be related in terms of similarity.

Case study: Predicting tumour subtype in ependymoma with netDx

To illustrate the use of patient similarity network based classification, we use netDx to classify patients as belonging to one of two ependymoma subtypes. Ependymoma is the third most common type of pediatric brain tumour, with nearly half the cases being incurable. Witt et al. identified two types of tumours originating in the posterior fossa of the brain (Group A and Group B), each subtype showing different demographic, clinical, and molecular profiles. We obtained normalized microarray gene expression data from Witt et al.[66] (total of 96 samples; 53 of Group A and 43 of Group B), and used regression to correct for batch effects. We first ran netDx with a single input network based on all genes, using pairwise Pearson correlation as a similarity metric. Lasso regression was used within the cross-validation loop to prefilter genes. Cross validation (10 train/test splits x 10-fold CV) was used to calculate predictor performance. This predictor achieved an AUROC of 0.90 (SEM=0.02), AUPR of 0.82 (SEM=0.02) and accuracy of 81% (SEM=0.02).

While the single-network design is simple and a good first-pass to estimate the signal in a given datatype, it does not provide mechanistic insight into what the predictor has learnt. We therefore implemented a pathway-based design, where all genes were grouped into 2,118 networks, one per pathway. Cross-validation and similarity were computed as before. Pathway definitions were aggregated from HumanCyc[67] IOB's NetPath[68], Reactome[69,70], NCI Curated Pathways[71], mSigDB[72], and Panther[73] (http://download.baderlab.org/EM_Genesets/February_01_2018/Human/symbol/Human_AllPathways_February_01_2018_symbol.gmt)[74]. The overall score for a feature was defined as the highest score it consistently obtained in >=70% of the trials. This resulted in comparable class separation than that with the single network, with an average AUROC of 0.92 (SEM=0.02), AUPR of 0.84 (SEM=0.02) and accuracy of 80% (SEM=3%). Top-scoring pathways predictive of Group A tumours were related to processes involved in interactions of the cell membrane with the extra-cellular matrix (Figure 4). These include terms related to the basement membrane, integrins, laminins and chondroitin sulfate proteoglycans (Figure 4B). These themes are consistent with those identified in the original paper describing the two tumour subtypes[66]. Figure 4C shows the integrated patient similarity network that results from combining the top-scoring networks and shows the clear separation between the two clusters. This example illustrates the utility of netDx as a classification tool and as a tool for generating mechanistic hypotheses for precision medicine.

The road ahead: Challenges and outlook for patient similarity network analytics in precision medicine

Network-based approaches have only recently started being applied for precision medicine and many challenges must be solved for them to reach their full potential. First, analytical methods must be improved to: 1) handle large data sizes (e.g. thousands of genomes); 2) identify the most relevant features for prediction, including non-linear interactions between features; 3) automate ways to generally improve signal-to-noise ratio; 4) automate ways to characterize patient heterogeneity, like disease subtypes[26,51,66,75]; 5) make the best use of complementary genomics layers which may have complex relationships (e.g. gene expression is modulated by genetic variants in a tissue-specific manner[76]); 6) improve performance by tuning parameters and hyperparameters (similar to the way in which Google's AutoML aims to solve this for particular problem domains - https://cloud.google.com/automl/). Scalability for patient networks can be improved by keeping only the strongest similarities by sparsification or by applying dimensionality reduction, as performed by the Mashup algorithm[77]. Deep learning is also promising[78], as recently demonstrated by the deep network fusion method[79]; this method performs classifications with neural networks, by using similarity networks as input. Both of these methods could be used in netDx.

Another major challenge is to improve the use of prior knowledge. For example, for sparse genetic data, such as somatic mutations of CNVs, "smoothing" mutations over a network of known gene-gene interactions has improved patient clustering[80]. Given that 43% of disease-associated genetic variants are located in intergenic regions (88% lie in noncoding regions, which includes introns)[81], incorporating non-coding and epigenetic information about gene regulation is important for a model seeking to explain clinical outcome. Pan-tissue atlases, such as the Roadmap Epigenomic Consortium and GTEx, as well as tissue-specific atlases such as the PsychENCODE project, are increasingly available to support this extension. New information about chromatin structure is also being mapped, such as topological associated domains (or TADs), characterized by high within-region chromatin looping, relative to interactions outside the region[82,83], and enhancer-promoter loops that activate transcription. Given this information, genetic variants could be limited to those known or predicted to affect gene function[84,85], such as via modulating gene expression. Use of this knowledge will increase the number of patients to whom a predictor is applicable, because some patients only have mutations in non-coding regions.

Perspective: Towards the future clinical visit

Based on these ideas, it is exciting to envision a doctor's clinic of the future, similar to the one described by Friend and Ideker[86], that uses network-based approaches for clinical decision-making (Figure 5). Such a system would initially be used by researchers to identify and validate successful predictors. A clinical researcher would identify patients to include in model training, and select which types of clinical and genomic data to include. Model training would run on centralized high performance computing systems, and results could then be interactively visualized in a web-based interface using Cytoscape.js[87], similar to the gene function prediction tool at http://genemania.org. Following completion of a research study, similarity networks could be uploaded to a repository such as NDEx[88] for sharing with the research community. Eventually, as the technology matures and as classifiers are validated, it would evolve to be useful to practicing physicians for use with their patients. This would require the development of additional reporting tools tailored for use in clinical decision-making. These would include a summary report card of overall confidence in the predictor as well as classification accuracy for a given patient, graphical summaries of relevant features used, and alerts about specific patient details that would affect result interpretation (e.g. ethnicity, lifestyle, genetic variants). It would also include links to relevant medical literature associating specific features with the disorder, to provide the clinician with information on prior knowledge to aid in decision-making. It would also provide the history of success rate for specific treatment choices for this condition in the health system, which would improve with data collection over time.

Algorithms like SNF and netDx advance several ideas to achieve this goal. They permit the integration of several genomic layers of patient data for patient subtyping or classification to directly answer specific clinical questions. From genomic data, netDx also can identify biological pathways whose alteration is predictive of patient outcome. This variation provides insight into mechanistic differences in patient subgroups that could be useful for rational treatment design. The integrated patient similarity network enables individual patients to be examined in the context of patients with clinically similar profiles (“neighbours”). This context enables the clinical researcher to identify the features the selected patient either conforms with, or deviates from, relative to the typical group profile. For instance, a patient classified as a treatment responder, but whose metabolic similarity is an outlier relative to other responders, may need to be more closely monitored for non-response, as compared to another patient whose profile is typical for a responder. Such network exploration would identify the pathophysiology unique to the patient, thereby enabling tailoring of their personal treatment plan.

Conclusion

Network-based approaches have the conceptual and technical features necessary to enable precision medicine that is grounded in biologically-informative, interpretable models. We predict that this paradigm will become increasingly useful in the next five years as it is used for subtype identification, prediction of clinical outcome, and the identification of biomarkers and targetable therapies in disease-related multi-omic studies.

Box items

Box 1. Online risk calculators and popular software for clinical prediction in the context of precision medicine.

Disease risk calculators in use
Cardiovascular disease risk (ASCVD):
http://tools.acc.org/ASCVD-Risk-Estimator-Plus/
Cardiovascular disease risk (Framingham Risk Score):
https://www.cvdriskchecksecure.com/framinghamriskscore.aspx
Neonatal sepsis risk: https://neonatalsepsiscalculator.kaiserpermanente.org/al
Melanoma risk: http://www.cancer.gov/melanomarisktool/
Diabetes risk: http://www.diabetes.org/are-you-at-risk/diabetes-risk-test/?loc=atrisk-slabnav

Clinical risk predictors that use genetic or genomic data
Non-Invasive Prenatal Testing: cell-free circulating DNA testing for trisomies:
http://www.perinatalservicesbc.ca/health-professionals/professional-resources/screening/prenatal-genetic/non-invasive-prenatal-testing-nipt
BOADICEA:breast and ovarian cancer risk: http://ccge.medschl.cam.ac.uk/boadicea/
oncotypeIQ: gene expression-based tests for breast, lung, prostate, colon cancer: http://www.oncotypeiq.com

Disease-specific genomic profiling initiatives
Alzheimer's Disease: Alzheimer's Disease Neuroimaging Initiative[33]
Autism: Autism Genetic Resource Exchange (https://research.agre.org), Simon's Simplex Collection[89], National Database for Autism Research (https://ndar.nih.gov)
Depression Treatment: Lundbeck (http://www.lundbeck.com/global/about-us/features/2017/flying-start-to-huge-depression-genetics-study); Canadian Biomarker Integration Network in Depression[90]
Cardiovascular risk: STARNET[34]
Cancers: The Cancer Genome Project (http://cancergenome.nih.gov)
Schizophrenia: PsychENCODE[91]

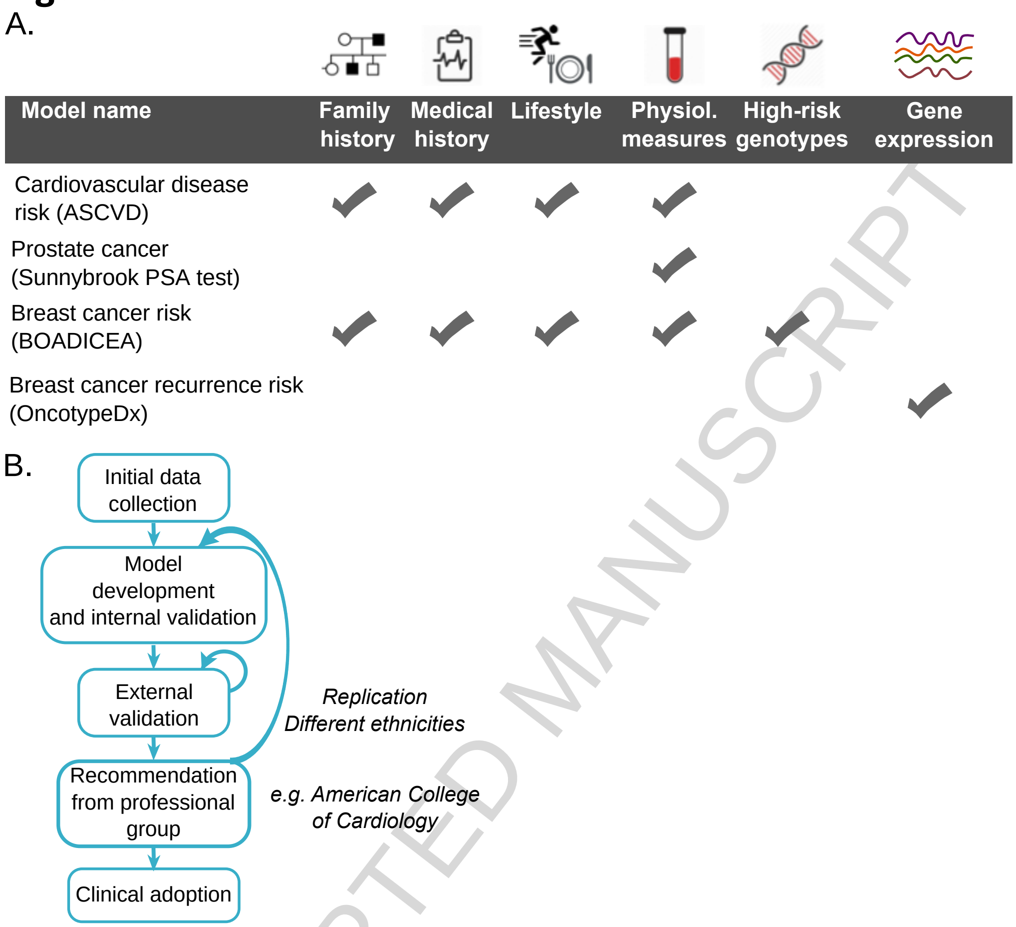
Software to compute clinical risk from genomic data:
Polygenic risk score: PRSice http://prcise.info[92]
Patient similarity networks for multi-omic integration: Similarity Network Fusion[57]
http://netdx.org[58]
General machine-learning software libraries: scikit-learn (python), caret (R), Weka, keras, tensorflow

Box 2: Concepts in patient risk modeling using machine learning

Key predictive model concepts

- **Machine-learning:** Algorithms that identify patterns in data by iterative exposure to data samples to either find subgroups within them (unsupervised learning, clustering, class discovery) or classify them (supervised learning). Examples of machine-learning algorithms include K-means clustering, regression, support vector machines, random forests and deep learning.
- **Supervised learning:** A class of machine-learning algorithm that learns to classify samples or predict output from provided examples in a training set. Examples should represent all outcomes in roughly equal proportions.
- **Features:** Inputs provided to the model for training. These could include individual measures (gene-level features in a gene expression matrix) or grouped measures (e.g. patient similarity network where genes were grouped by pathway).
- **Feature selection:** Step of model development that assigns weights to features such that more predictive features have higher weights.
- **Train and test set:** Model building has two phases: training the model, which identifies feature weights, and testing the model, where the model generalization is tested on independent data. Input samples are randomly partitioned into two groups; the set used to train the model is called the *training set* and that used to validate the model following the training is called the *test set*.
- **Overfitting:** Overfitting is the phenomenon of a model learning patterns that are biased to the data it is trained on and that limits its generalizability to new data. A symptom of overfitting is a model that performs excellently on training data but poorly on new data. Overfitting is a common occurrence in machine learning, especially with high dimensional data, and model training must incorporate strategies such as cross-validation to reduce the overfitting risk.
- **Cross-validation:** Repeatedly holding out a portion of the training data, training on the remaining data, and computing prediction error on the held-out set. Repeating 10 times is called “10-fold cross validation”. Performance over all runs is used to estimate the generalization of the model.
- **Performance measures:** Metrics that evaluate how well a predictor works, such as the balance between true and false positive rate. Common measures are specificity, sensitivity, ROC curves, precision-recall curves, PPV, F1 and MCC.
- **Regularization:** A constraint applied in model fitting problems with large number of features (e.g. 20,000 gene-level measures) to limit the number of features with non-zero weights. Reduces redundancy and improves interpretability.

Figures



A.

Model name	Family history	Medical history	Lifestyle	Physiol. measures	High-risk genotypes	Gene expression
Cardiovascular disease risk (ASCVD)	✓	✓	✓	✓		
Prostate cancer (Sunnybrook PSA test)				✓		
Breast cancer risk (BOADICEA)	✓	✓	✓	✓	✓	
Breast cancer recurrence risk (OncotypeDx)						✓

B.
Initial data collection
Model development and internal validation
External validation
Replication
Different ethnicities
Recommendation from professional group
e.g. American College of Cardiology
Clinical adoption

Figure 1. **Contemporary risk calculators and their development process.**

A. Examples of risk models in current clinical use (rows) and the patient data required for each (columns). See Box 1 for details.

B. Process for risk model development. The first model is developed by testing performance of a variety of models on subsets of the training data (internal validation). Following successful internal validation, model generalizability is then assessed by external validation on similar populations. Generalizability is also tested on similar populations with specific differences (e.g. geographic origin). This step would identify whether it is possible to develop a general model for multiple populations or whether subpopulation-specific models are needed. A well-validated model is recommended in professional clinical practice guidelines, but a clinician may choose to adopt a sufficiently validated model earlier in this process. This process is iterative and refinements continue to be made on decades-long models in clinical use.

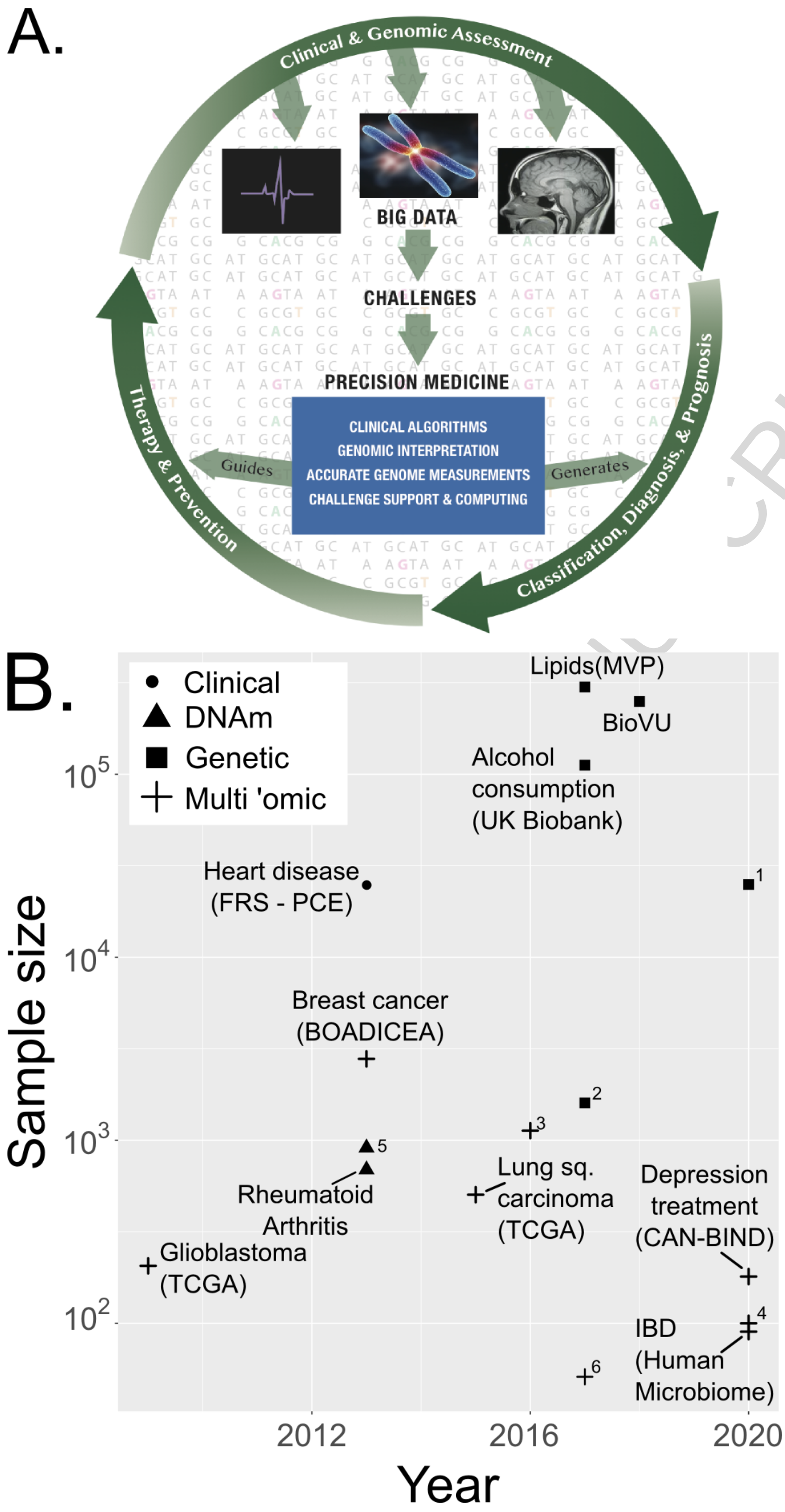


Figure 2. **Genomics in clinical risk models.**
A. Vision of genomic analyses as part of a process for clinical decision-making.

The outer ring tracks patient interactions with the healthcare system in a future genomic era of medicine. Clinical and genomic assessment generates patient data, whereupon physicians diagnose patients, prescribe therapy and counsel about prevention based on disease risk. Patients iterate this process with follow-up visits. The field of computational biology will catalyze precision medicine by developing tools that help generate patient classification, diagnosis and prognosis, and guide therapy and prevention.

B. Current and projected ‘omic cohorts for precision medicine. The x-axis shows the year of the publication or update; values at 2020 are projected by the authors based on public information. Y-axis shows the sample size on which the project was or is projected to run (powers of 10). [90,93-101] (IBD: https://ibdmdb.org/; https://victr.vanderbilt.edu/pub/biovu/). Unlabelled points are for: 1. http://www.lundbeck.com/global/about-us/features/2017/flying-start-to-huge-depression-genetics-study, 2. Blood lipids GWAS[98]; 3. Glioma[95]; 4. Type 2 diabetes microbiome: http://med.stanford.edu/ipop.html; 5. Breast cancer[51] 6 - Cholangiocarcoma[102]. MVP: Million Veterans Program.

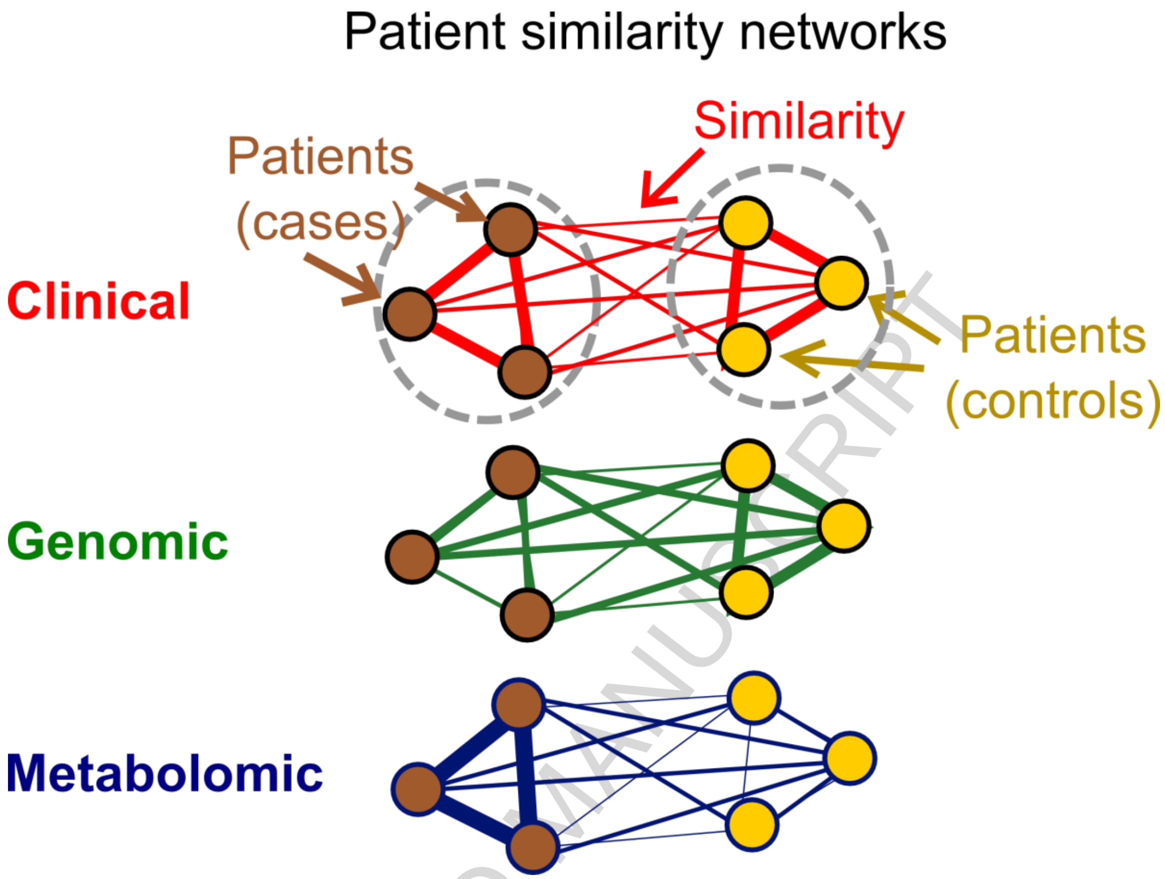


Figure 3. **Patient similarity networks for hypothetical example of predicting lung cancer risk**

Nodes are patients and edge weights reflect datatype similarity. This example shows similarity from clinical (red), gene expression (green) and metabolomics (blue) data. Here, cases and controls form separate densely connected parts of the network based on clinical data (red; e.g. smoking frequency), and a similar clique in metabolomics data (blue). The predictor would therefore select clinical data and metabolomic data as predictive of case status.

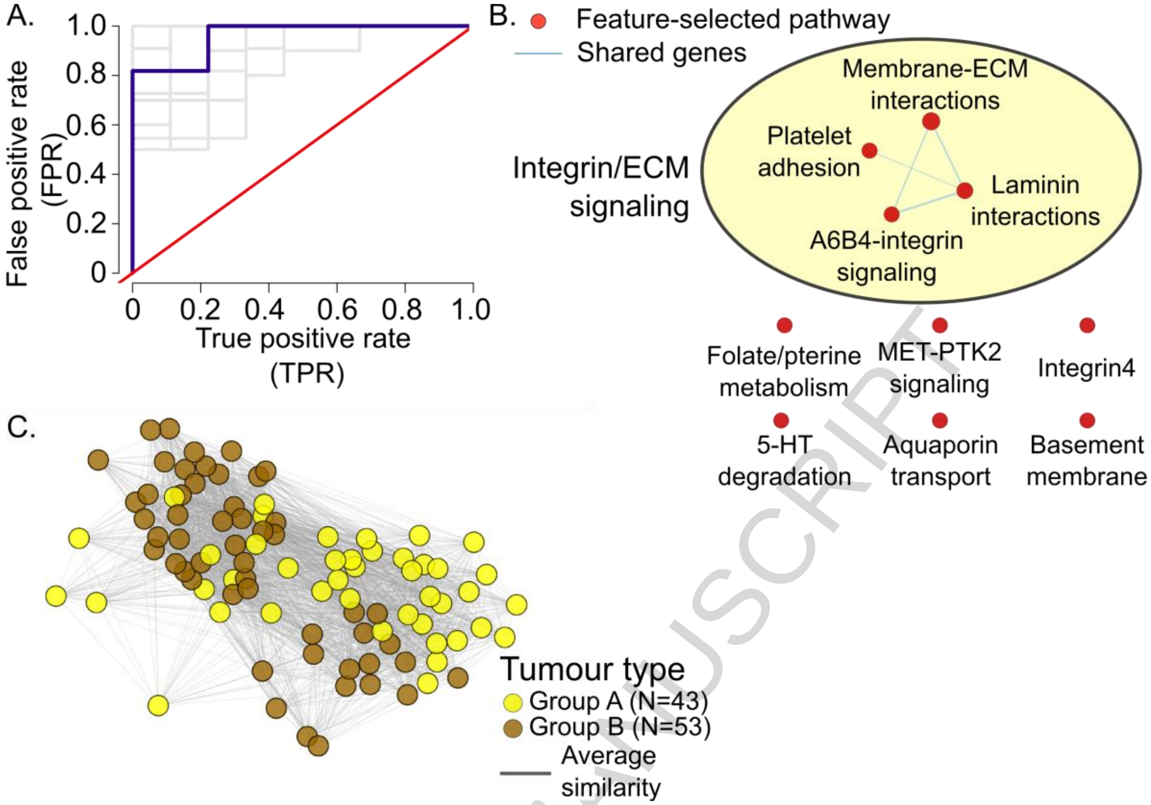
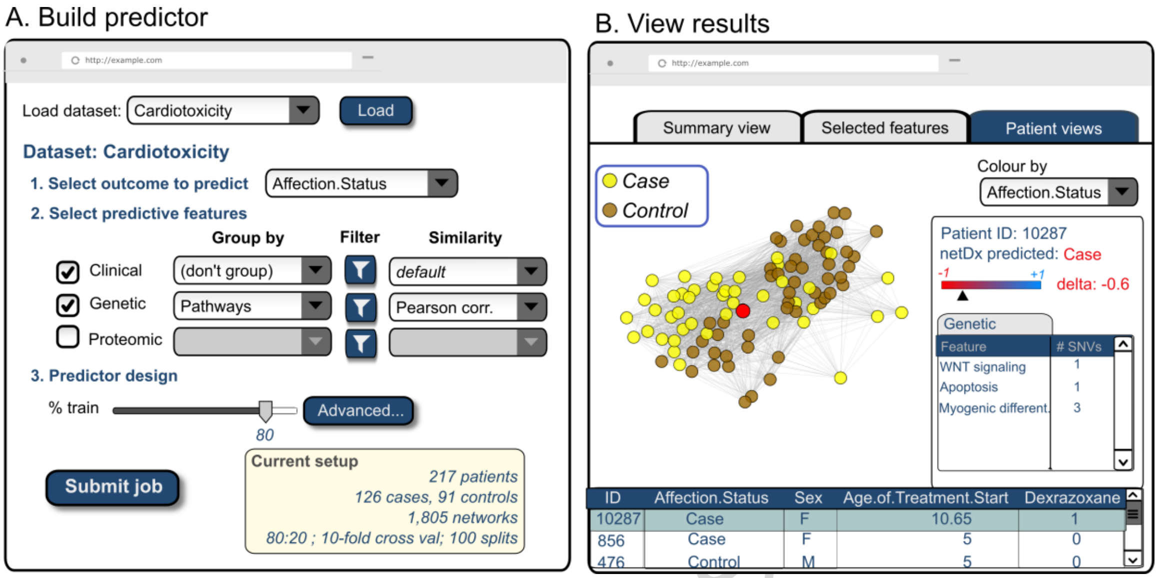


Figure 4. **Predicting ependymoma subtype with netDx**

A. ROC curve showing performance over 10 train/test splits (grey) and the average (blue).
B. Pathway-level scores for Group A tumours. Nodes show pathway-level features that scored 10/10 in >= 7 out of 10 trials; edges connect pathways with shared genes. AutoAnnotate was used to cluster pathways.[74,103]
C. Integrated patient similarity network following feature selection. Nodes show the two types of tumours. Edges show patient similarity for pathways scoring 10/10 in all splits for either class. For visualization, the top 90% edges were included; edge-weighted spring-embedded layout was used to lay out the network in Cytoscape.



Load dataset: Cardiotoxicity ▾ Load

Dataset: Cardiotoxicity

1. Select outcome to predict: Affection.Status ▾
2. Select predictive features

	Group by	Filter	Similarity
☑ Clinical	(don't group)		default
☑ Genetic	Pathways		Pearson corr.
□ Proteomic			

3. Predictor design

% train 80 Advanced...

Submit job

Current setup
217 patients
126 cases, 91 controls
1,805 networks
80:20 ; 10-fold cross val; 100 splits

Summary view | Selected features | Patient views

● Case
● Control

Colour by: Affection.Status ▾

Patient ID: 10287
netDx predicted: Case
-1 +1 delta: -0.6

Genetic

Feature	# SNVs
WNT signaling	1
Apoptosis	1
Myogenic different.	3

ID	Affection.Status	Sex	Age.of.Treatment.Start	Dexrazoxane
10287	Case	F	10.65	1
856	Case	F	5	0
476	Control	M	5	0

Figure 5. **Vision for a network-based classification tool for precision medicine.**

A. User interface for a network-based patient classifier software tool, such as netDx, in the near future. Such a system could be integrated with a research hospital Electronic Medical Record system and in-house genomics database. A clinical researcher could use this to build a predictor by selecting data of interest and predictor options.

B. User interface for visualizing predictor results, represented as multiple tabs. Here, the active tab shows a hypothetical integrated patient similarity network. The user has interactively highlighted a single patient for detailed study (red node) as shown in the right panel.

Tables

Table 1. **Methods used in clinical risk models**

Method	Advantages	Disadvantages	Applications
Similarity networks	* Interpretable * Handles missing data * History of success in gene/protein function prediction	* New paradigm, yet to be extensively applied * Scalability needs to be improved * Currently supports only categorical outcomes	Gene function prediction[34,65,77], protein function prediction[79], cancer survival prediction, asthma case/control prediction, breast cancer subtype prediction[58]
Linear/ logistic regression	* Simple to use	* Requires imputation of missing data * Effective only for linearly separable data * Requires coding of categorical variables	Diabetes, Prostate cancer[11,12]
Cox proportional hazards models	* Well-suited for modeling survival (time-dependence of risk) in censored data * Moderate assumptions about underlying model	* Requires imputation of missing data * Assumes risk is always proportional to base * Limited usefulness for problems not related to survival	Risk of developing cardiovascular disease or stroke; breast cancer risk[10,13,14]
Polygenic risk score	* Easy to calculate	Limited to genetic data, limited by GWAS results	Coronary heart disease, schizophrenia.[23,38,39]
MultiBLUP (or realized relationship matrix)	* Captures full contribution of genetic component	* Current framework specific to quantifying effect of large number of genetic contributors with varied effect sizes and local correlation structure, on a complex trait. Suitability for other problem domains to be yet demonstrated.	Celiac disease, yeast QTL analysis [41,42]
Support vector machines	Consistently well-performing	* Requires imputation of missing data * Compute-intensive * Requires tuning * Interpretation requires work	Cancer survival prediction; Celiac disease genetic risk[21,27,45]

Random forest	* Handles continuous and categorical data * Improved interpretability , compared to support vector machines	• Potentially slow for real-time predictions, especially for models with many decision trees • May perform poorly with rare outcomes	Survival prediction in various cancer types; autism case/control prediction from CNVs[27,38,39,104,105]
Deep learning	* Can model complex nonlinearities based on structure of neural net * Consistently excellent performance	* Requires additional work to interpret * Computationally intensive to run * Requires computational expertise to tune	Deep survival models (cancer survival)[106]

Acknowledgements

This work was supported by NRNB (U.S. National Institutes of Health, National Center for Research Resources grant number P41 GM103504). We thank Ruth Isserlin for providing prepared data for the ependymoma classification example.

References

1. Katsnelson, A. Momentum grows to make 'personalized' medicine more 'precise'. *Nat Med* **19**, 249 (2013).
2. Meyer, U., Feldon, J. & Dammann, O. Schizophrenia and autism: both shared and disorder-specific pathogenesis via perinatal inflammation? *Pediatr Res* **69**, 26r-33r (2011).
3. Truett, J., Cornfield, J. & Kannel, W. A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chronic Dis* **20**, 511-24 (1967).
4. Dawber, T.R., Meadors, G.F. & Moore, F.E., Jr. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health* **41**, 279-81 (1951).
5. Rana, J.S. et al. Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Population. *J Am Coll Cardiol* **67**, 2118-2130 (2016).
6. Hippisley-Cox, J. et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *Bmj* **336**, 1475-82 (2008).
7. Hippisley-Cox, J. et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *Bmj* **335**, 136 (2007).
8. Brindle, P. et al. Primary prevention of cardiovascular disease: a web-based risk score for seven British black and minority ethnic groups. *Heart* **92**, 1595-602 (2006).

9. Gradishar, W.J. et al. NCCN Guidelines Insights: Breast Cancer, Version 1.2017. *J Natl Compr Canc Netw* **15**, 433-451 (2017).
10. Zhang, Z., Gillespie, C., Bowman, B. & Yang, Q. Prediction of atherosclerotic cardiovascular disease mortality in a nationally representative cohort using a set of risk factors from pooled cohort risk equations. *PLoS One* **12**, e0175822 (2017).
11. Zhang, L., Loblaw, A. & Klotz, L. Modeling prostate specific antigen kinetics in patients on active surveillance. *J Urol* **176**, 1392-7; discussion 1397-8 (2006).
12. Schmidt, M.I. et al. Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study. *Diabetes Care* **28**, 2013-8 (2005).
13. Lee, A.J. et al. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. *Br J Cancer* **110**, 535-45 (2014).
14. Tang, G. et al. Comparison of the prognostic and predictive utilities of the 21-gene Recurrence Score assay and Adjuvant! for women with node-negative, ER-positive breast cancer: results from NSABP B-14 and NSABP B-20. *Breast Cancer Res Treat* **127**, 133-42 (2011).
15. Kuzniewicz, M.W. et al. A Quantitative, Risk-Based Approach to the Management of Neonatal Early-Onset Sepsis. *JAMA Pediatr* **171**, 365-371 (2017).
16. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896-d901 (2017).
17. Corder, E.H. et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921-3 (1993).
18. Verghese, P.B., Castellano, J.M. & Holtzman, D.M. Apolipoprotein E in Alzheimer's disease and other neurological disorders. *Lancet Neurol* **10**, 241-52 (2011).
19. Noble, J.A. & Valdes, A.M. Genetics of the HLA region in the prediction of type 1 diabetes. *Curr Diab Rep* **11**, 533-42 (2011).
20. Harrison, P.J. Recent genetic findings in schizophrenia and their therapeutic relevance. *J Psychopharmacol* **29**, 85-96 (2015).
21. Deloukas, P. et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* **45**, 25-33 (2013).
22. Ikeda, M. et al. A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. *Mol Psychiatry* **23**, 639-647 (2018).
23. Purcell, S.M. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-52 (2009).
24. Boycott, K.M., Vanstone, M.R., Bulman, D.E. & MacKenzie, A.E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14**, 681-91 (2013).
25. Wallden, B. et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med Genomics* **8**, 54 (2015).
26. Mack, S.C. et al. Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature* **506**, 445-50 (2014).

27. Yuan, Y. et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* **32**, 644-52 (2014).
28. Gaziano, J.M. et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214-23 (2016).
29. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
30. Ledford, H. AstraZeneca launches project to sequence 2 million genomes. *Nature* **532**, 427 (2016).
31. Hu, Y. et al. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. *Nat Commun* **7**, 10448 (2016).
32. Sanchez-Roige, S. et al. Genome-wide association study of delay discounting in 23,217 adult research participants of European ancestry. *Nat Neurosci* **21**, 16-18 (2018).
33. Mueller, S.G. et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement* **1**, 55-66 (2005).
34. Franzen, O. et al. Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science* **353**, 827-30 (2016).
35. den Hollander, P., Savage, M.I. & Brown, P.H. Targeted therapy for breast cancer prevention. *Front Oncol* **3**, 250 (2013).
36. Ripatti, S. et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* **376**, 1393-400 (2010).
37. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).
38. Andersen, A.M. et al. Polygenic Scores for Major Depressive Disorder and Risk of Alcohol Dependence. *JAMA Psychiatry* **74**, 1153-1160 (2017).
39. Escott-Price, V., Shoai, M., Pither, R., Williams, J. & Hardy, J. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiol Aging* **49**, 214.e7-214.e11 (2017).
40. Power, R.A. et al. Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat Neurosci* **18**, 953-5 (2015).
41. de Los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C. & Sorensen, D. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet* **9**, e1003608 (2013).
42. Martens, K., Hallin, J., Warringer, J., Liti, G. & Parts, L. Predicting quantitative traits from genome and phenome with near perfect accuracy. *Nat Commun* **7**, 11512 (2016).
43. Speed, D. & Balding, D.J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res* **24**, 1550-7 (2014).
44. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43**, 1193-201 (2011).
45. Abraham, G. et al. Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genet* **10**, e1004137 (2014).

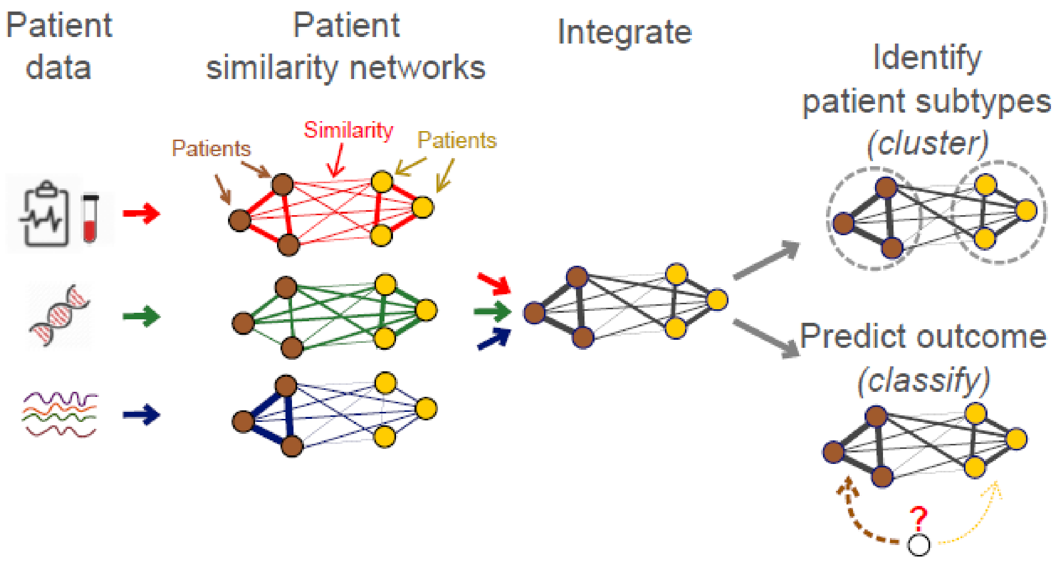
46. Romanos, J. et al. Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. *Gastroenterology* **137**, 834-40, 840.e1-3 (2009).
47. Perou, C.M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747-52 (2000).
48. Paik, S. et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817-26 (2004).
49. Prat, A. et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, R68 (2010).
50. Prat, A. & Perou, C.M. Deconstructing the molecular portraits of breast cancer. *Mol Oncol* **5**, 5-23 (2011).
51. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
52. Stephen, J. et al. Time dependence of biomarkers: non-proportional effects of immunohistochemical panels predicting relapse risk in early breast cancer. *Br J Cancer* **111**, 2242-7 (2014).
53. Cardoso, F. et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N Engl J Med* **375**, 717-29 (2016).
54. Gaonkar, B., R, T.S. & Davatzikos, C. Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. *Med Image Anal* **24**, 190-204 (2015).
55. Rasmussen, P.M., Madsen, K.H., Lund, T.E. & Hansen, L.K. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *Neuroimage* **55**, 1120-31 (2011).
56. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med* **7**, 311ra174 (2015).
57. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **11**, 333-7 (2014).
58. Pai, S. et al. netDx: Interpretable patient classification using integrated patient similarity networks. *bioRXiv preprint.* (2016).
59. Chen, F. et al. PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensionS. *Bioinformatics* **33**, 871-878 (2017).
60. Ruffalo, M., Stojanov, P., Pillutla, V.K., Varma, R. & Bar-Joseph, Z. Reconstructing cancer drug response networks using multitask learning. *BMC Syst Biol* **11**, 96 (2017).
61. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).
62. Cavalli, F.M.G. et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell* **31**, 737-754.e6 (2017).
63. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32**, 185-203.e13 (2017).
64. Pena-Castillo, L. et al. A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biol* **9 Suppl 1**, S2 (2008).

65. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* **9 Suppl 1**, S4 (2008).
66. Witt, H. et al. Delineation of two clinically and molecularly distinct subgroups of posterior fossa ependymoma. *Cancer Cell* **20**, 143-57 (2011).
67. Romero, P. et al. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6**, R2 (2005).
68. Kandasamy, K. et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol* **11**, R3 (2010).
69. Croft, D. et al. The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-7 (2014).
70. Fabregat, A. et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-7 (2016).
71. Schaefer, C.F. et al. PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674-9 (2009).
72. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
73. Mi, H. et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* **33**, D284-8 (2005).
74. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G.D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).
75. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350-6 (2015).
76. Battle, A., Brown, C.D., Engelhardt, B.E. & Montgomery, S.B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
77. Cho, H., Berger, B. & Peng, J. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Syst* **3**, 540-548.e5 (2016).
78. Webb, S. Deep learning for biology. *Nature* **554**, 555-557 (2018).
79. Gligorijević, V., Barot, M. & Bonneau, R. deepNF: Deep network fusion for protein function prediction. *bioRXiv preprint.* (2017).
80. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* **10**, 1108-15 (2013).
81. Hindorff, L.A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).
82. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).
83. Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-5 (2012).
84. Bendl, J. et al. PredictSNP2: A Unified Platform for Accurately Evaluating SNP Effects by Exploiting the Different Characteristics of Variants in Distinct Genomic Regions. *PLoS Comput Biol* **12**, e1004962 (2016).
85. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).

86. Friend, S.H. & Ideker, T. Point: Are we prepared for the future doctor visit? *Nat Biotechnol* **29**, 215-8 (2011).
87. Franz, M. et al. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* **32**, 309-11 (2016).
88. Pratt, D. et al. NDEx, the Network Data Exchange. *Cell Syst* **1**, 302-305 (2015).
89. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-5 (2010).
90. Kennedy, S.H. et al. The Canadian Biomarker Integration Network in Depression (CAN-BIND): advances in response prediction. *Curr Pharm Des* **18**, 5976-89 (2012).
91. Akbarian, S. et al. The PsychENCODE project. *Nat Neurosci* **18**, 1707-12 (2015).
92. Euesden, J., Lewis, C.M. & O'Reilly, P.F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466-8 (2015).
93. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-25 (2012).
94. Braun, K.V.E. et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clin Epigenetics* **9**, 15 (2017).
95. Ceccarelli, M. et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550-63 (2016).
96. Clarke, T.K. et al. Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). *Mol Psychiatry* **22**, 1376-1384 (2017).
97. Farshidfar, F. et al. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. *Cell Rep* **18**, 2780-2794 (2017).
98. Liu, D.J. et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat Genet* **49**, 1758-1766 (2017).
99. Liu, Y. et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* **31**, 142-7 (2013).
100. Verhaak, R.G. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98-110 (2010).
101. Xu, Z. et al. Epigenome-wide association study of breast cancer using prospectively collected sister study samples. *J Natl Cancer Inst* **105**, 694-700 (2013).
102. Farshidfar, F. et al. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. *Cell Rep* **19**, 2878-2880 (2017).
103. Kucera, M., Isserlin, R., Arkhangorodsky, A. & Bader, G.D. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. *F1000Res* **5**, 1717 (2016).
104. Engchuan, W. et al. Performance of case-control rare copy number variation annotation in classification of autism. *BMC Med Genomics* **8 Suppl 1**, S7 (2015).

105. Gerstung, M. et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun* **6**, 5901 (2015).

106. Yousefi, S. et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci Rep* **7**, 11707 (2017).



Graphical abstract

Highlights

* Future clinics will combine clinical and genomic data with cellular models for precision medicine.

* Statistical risk calculators using genomics need to be interpretable due to small sample sizes.

* Patient similarity networks (PSN) are a new model to integrate data to cluster/classify patients.

* PSN are accurate, intuitive, preserve patient privacy and supply mechanistic insight.