Journal of Biomedical Informatics 58 (2015) 104–113

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Whole-exome sequencing enhances prognostic classification of myeloid malignancies

Matthew Ruffalo [a], Holleh Husseinzadeh [b], Hideki Makishima [b], Bartlomiej Przychodzen [b], Mohamed Ashkar [b], Mehmet Koyutürk [a], Jaroslaw P. Maciejewski [b], Thomas LaFramboise [a,c,*]

[a] Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA
[b] Department of Translational Hematology and Oncology Research, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH, USA
[c] Department of Genetics and Genome Science, Case Western Reserve University, Cleveland, OH, USA

ARTICLE INFO

Article history:
Received 5 May 2015
Revised 14 August 2015
Accepted 2 October 2015
Available online 8 October 2015

Keywords:
Myeloid malignancy
Patient prognosis
Mutation
Exome sequencing
Random forests

ABSTRACT

Purpose: To date the standard nosology and prognostic schemes for myeloid neoplasms have been based on morphologic and cytogenetic criteria. We sought to test the hypothesis that a comprehensive, unbiased analysis of somatic mutations may allow for an improved classification of these diseases to predict outcome (overall survival).
Experimental design: We performed whole-exome sequencing (WES) of 274 myeloid neoplasms, including myelodysplastic syndrome (MDS, $N = 75$), myelodysplastic/myeloproliferative neoplasia (MDS/MPN, $N = 33$), and acute myeloid leukemia (AML, $N = 22$), augmenting the resulting mutational data with public WES results from AML ($N = 144$). We fit random survival forests (RSFs) to the patient survival and clinical/cytogenetic data, with and without gene mutation information, to build prognostic classifiers. A targeted sequencing assay was used to sequence predictor genes in an independent cohort of 507 patients, whose accompanying data were used to evaluate performance of the risk classifiers.
Results: We show that gene mutations modify the impact of standard clinical variables on patient outcome, and therefore their incorporation hones the accuracy of prediction. The mutation-based classification scheme robustly predicted patient outcome in the validation set (log rank $P = 6.77 \times 10^{-21}$; poor prognosis vs. good prognosis categories HR 10.4, 95% CI 3.21–33.6). The RSF-based approach also compares favorably with recently-published efforts to incorporate mutational information for MDS prognosis.
Conclusion: The results presented here support the inclusion of mutational information in prognostic classification of myeloid malignancies. Our classification scheme is implemented in a publicly available web-based tool (http://myeloid-risk.case.edu/).

1. Introduction

Traditional classification of myeloid neoplasms relies heavily on morphologic and cytogenetic features to define major sub-entities and risk categories. The goals of nosologic schemes are to define disease by the most likely outcomes, biologic behavior, and therapy responses. Despite constant improvement, classical and current categorization schemes [10,11,21,3] suffer from heterogeneity within classes and poor distinction between classes. Moreover, morphology-defined sub-entities may not reflect underlying disease severity due to redundancies in the function of various molecular defects as well as phenocopies.

Recent technological advances facilitate new levels of understanding of the molecular pathogenesis of cancer in general, and specifically in myeloid neoplasms where progress has been particularly rapid. Whole-exome sequencing (WES) allows detection of nearly all amino acid coding changes in an individual genome [20,17,25], revealing the diversity of mutations and the complexity of mutational patterns, which likely explains a portion of clinical heterogeneity. The technology therefore opens the door to incorporating somatic mutations into diagnostic and prognostic applications, with the potential to augment current schemes.

In general, the most common statistical approach for identifying patient features that have an impact on outcome is Cox regression applied to right-censored data. However, challenges arise when using gene mutations as classifying features in the Cox

[*] Corresponding author at: Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine, 10900 Euclid Avenue, Cleveland, OH 44106, USA. Tel.: +1 216 368 0150; fax: +1 216 368 3432.
E-mail address: Thomas.LaFramboise@case.edu (T. LaFramboise).

http://dx.doi.org/10.1016/j.jbi.2015.10.003
1532-0464/© 2015 Elsevier Inc. All rights reserved.

model since a cohort of hundreds of myeloid malignancy patients will collectively carry mutations in thousands of distinct genes. The Cox proportional hazards assumption may not hold, and non-linear effects and interactions are difficult to capture without severe risk of overfitting, especially with very large numbers of classifying features. The problems are particularly acute when relationships among mutations, patient survival, and other variables are complex. For example, *NPM1* mutations have been associated with improved AML patient outcome in the absence of *FLT3* internal tandem duplication (*FLT3*-ITD) [5], and also in the presence of *IDH1* or *IDH2* [30]. On the other hand, adverse prognosis has been reported [29] for AML patients with *IDH1* or *IDH2* mutations and absence of *FLT3*-ITD. These types of complex dependencies cannot be captured by linear regression models such as the Cox model unless interaction terms are included. Power to detect interactions in this way would require much larger sample sizes than are typically available.

As an alternative approach we have adopted random survival forests (RSFs) to integrate gene mutations into a risk classifier of myeloid malignancy patients. An extension of random forests [4], RSFs are specifically designed for survival data. Briefly, the procedure constructs hundreds to thousands of decision trees, each of which uses a random subset of predictors (clinical features and/or mutation status in our case) to iteratively split the patient set into subsets with similar survival. The result of the procedure is an ensemble forest that collectively assigns each patient a predicted mortality. RSFs have the advantage of being able to handle non-linear effects and interactions automatically, owing to the complex multiple-decision-tree structure of the forest. Overfitting is minimized by randomizing the samples and variables used for each tree and split, respectively. In addition, the importance of the classifying variables may be assessed using the “minimal depth of maximal subtree” metric [14].

The RSF methodology is particularly well-suited to studies involving large numbers of classifying variables, as is the case with data produced by genomic biotechnologies. Since their introduction, RSFs have been applied to various human diseases. In cancer, researchers have used the approach to assess risk from genomic and non-genomic data alike. Outcomes for colon [22] and thyroid [1] cancer patients have been predicted by applying RSFs to data from large non-genomic databases, while genomic/proteomic data of various types were recently used to develop RSF-based predictors of outcome in 12 different tumor types [39]. In leukemias, RSFs were used to predict survival from DNA methylation data in AML [37] and from gene expression data in pediatric T-cell acute lymphoblastic leukemia [35].

Here, we performed WES on myeloid neoplasms of various types to test the hypothesis that a comprehensive analysis of somatic mutations may allow for an enhanced classification of these diseases, facilitating separation of the entities based on outcome (overall survival). We postulate that mutational features reflect the distinct pathogenesis of individual subtypes because of their direct foundation in molecular defects. We incorporate both gene mutation information from WES and more standard clinical variables such as cytogenetics, diagnoses, and morphological features. Our patient cohort comprises hundreds of individuals diagnosed with myelodysplastic/myeloproliferative neoplasms (MDS/MPN), myelodysplastic syndrome (MDS), or acute myeloid leukemia (AML). Our goals were to use RSFs (i) construct risk classifiers based on mutations and clinical features; (ii) assess the performance of the classifiers on an independent validation set; (iii) determine the improvement in prognostic accuracy gained by incorporating the mutation data in addition to the standard variables; and (iv) compare the performance of the RSF-based approach with extant MDS-specific prognostic schemes. Our mutation-based risk classifier is distributed as a publicly available web-based tool, allowing the user to input clinical variable values and mutational status of relevant genes, producing the corresponding risk category.

2. Materials and methods

2.1. Patient cohorts

All samples were collected after written informed consent was obtained. For the training (WES) cohort, paired tumor (bone marrow aspirate) and normal (CD3+ T-cells) DNA was obtained from Cleveland Clinic patients diagnosed with MDS, MDS/MPN, and AML. Data from these patients were augmented with clinical and WES data for AML patients from the Cancer Genome Atlas (TCGA) [18]. For the validation (targeted sequencing) set, tumor DNA (from bone marrow aspirate or peripheral blood) was subjected to targeted gene enrichment and deep-sequenced. Cytogenetic anomalies for all samples were called using FISH or single nucleotide polymorphism (SNP) arrays. SNP array analysis was performed using Affymetrix 250K and 6.0 platforms (Affymetrix, Santa Clara, CA) according to the standard protocols, followed by copy number analysis using CNAG (v3.0) [26] or Genotyping Console (Affymetrix). Patients positive for the t(15;17) translocation or for chromosomal lesions affecting the core binding factor transcription complex were omitted from downstream analysis. Malignancies with these characteristics are considered to be separate entities, are already known to be driven by very specific rearrangements, and are associated with much more favorable outcomes than other forms of myeloid malignancies [41]. Patient characteristics for both the training and validation cohorts are shown in Table 1.

2.2. Whole-exome sequencing and mutation calling

For WES, total genomic DNA was enriched for approximately 50 Mb of protein coding sequences by liquid phase hybridization using SureSelect version 4 (Agilent), followed by massively parallel sequencing with the HiSeq 2000 (Illumina), according to the manufacturers’ protocols. Sequence data was aligned using BWA [19] and candidate variants were detected with the GATK pipeline [24] (Supplementary Fig. S1). Somatic protein-altering and splice site mutations were identified using Genomon (http://genomon.hgc.jp/exome/en/index.html). TCGA patient mutation calling was performed as previously described [18].

2.3. Target assay and sequencing for classifier validation

A TruSeq (Illumina) custom enrichment kit was designed to include the coding regions of genes with mutations that were deemed likely to influence patient outcome. The kit was used to extract these genomic regions from the validation cohort DNA for sequencing on the MiSeq (Illumina), and mutations were called with the GATK pipeline [24] (Supplementary Fig. S1).

2.4. Statistical analysis

The probability of m patients in our WES cohort harboring a mutation in a gene with patient population mutation frequency f was calculated as $P(X = m)$, where X is distributed as a *Binomial* $(274,f)$ random variable (since our cohort size is 274). Survival analysis was conducted using R [31] version 3.0.2, with the survival [33] and randomForestSRC [13] packages. Mosaic plots were created in R using the mosaicplot function, and Kendall’s τ computed using the cor.test function with method = “Kendall”. Association between prognostic group and binary variables (mutation and cytogenetic status) was assessed using Fisher’s exact test on

Table 1
Patient characteristics (training and validation cohorts). IQR denotes interquartile range.

	Training cohort	Validation cohort
Patient count	274	507
MDS	75 (27.4%)	236 (46.6%)
Histologically high risk	26 (34.7%)	95 (40.3%)
Histologically low risk	49 (65.3%)	141 (59.7%)
IPSS high risk	9 (11.7%)	12 (5.1%)
IPSS intermediate-2	20 (26.0%)	39 (16.7%)
IPSS intermediate-1	30 (39.0%)	93 (39.7%)
IPSS low risk	18 (23.3%)	90 (38.5%)
MDS/MPN	33 (12.0%)	97 (19.1%)
AML	166 (60.6%)	174 (34.3%)
Clinical measures (median ± IQR)		
Hemoglobin (g/dL)	9.8 ± 1.7	9.8 ± 2.4
Bone marrow blasts	8.5 ± 43.3	4.0 ± 18.0
Platelets (10^9/L)	64.0 ± 90.8	74.0 ± 122.0
Absolute neutrophil count (10^9/L)	2.0 ± 3.9	2.15 ± 4.0
Cytogenetics		
Normal	105 (38.3%)	262 (51.7%)
3q-	6 (2.2%)	12 (2.4%)
inv(3)	1 (0.36%)	3 (0.59%)
t(3q)	0 (0.0%)	1 (0.20%)
5q-	43 (15.7%)	56 (11.0%)
7-	16 (5.8%)	49 (9.7%)
7q-	27 (9.9%)	41 (8.1%)
Trisomy 8	27 (9.9%)	56 (11.0%)
t(9;11)	2 (0.73%)	1 (0.20%)
11q-	2 (0.72%)	14 (2.8%)
12p-	10 (3.6%)	14 (2.8%)
inv(17q)	1 (0.36%)	4 (0.79%)
Trisomy 19	2 (0.73%)	10 (2.0%)
20q-	12 (4.4%)	41 (8.1%)
Y-	4 (1.5%)	14 (2.8%)
Complex	20 (7.3%)	27 (5.3%)
Demographics		
Age (mean ± sd)	61.9 ± 15.3	66.1 ± 12.9
Gender	61% male	63% male
Outcome		
Alive at last followup	28%	31%
Time to death (months, median ± IQR)	10.1 ± 14.4	7.6 ± 12.5
$RSF_{clin+mut}$ prognostic categories		
Poor	28%	9%
Intermediate Poor	29%	46%
Intermediate Good	25%	42%
Good	18%	3%

patient counts. Continuous measures (clinical variables) were compared using the non-parametric Wilcoxon test.

To test interactions between gene mutations and clinical features, we fit a Cox proportional hazards model, including gene mutation, the clinical feature, and interaction terms. The *P*-values reported correspond to the interaction terms. In the case of bone marrow blasts, hazard ratios and their confidence intervals are shown separately for patients with and without the relevant gene mutation.

2.5. Constructing the risk score from RSF

The R package randomForestSRC was used to construct a RSF from the data. *FLT3* mutations were separately classified as internal tandem duplications (*FLT3*-ITD) or tyrosine kinase domain mutations (*FLT3*-TKD). Cytogenetic lesion status was encoded as a binary variable (presence/absence) for del(11q), del(12p), del(20q), del(3q), del(5q), del(7), del(7q), del(Y), inv(17q), inv(3), t(3q), t (9;11), trisomy 19, trisomy 8, and complex. Numerical values for absolute neutrophil count, bone marrow blast percentage, hemoglobin, and platelets were stratified into categories corresponding to those in [11]. Each stratum was coded as a binary indicator variable. Indicator values for missing numeric variables were set to zero. RSF_{clin} was constructed using only clinical variables, while $RSF_{clin+mut}$ additionally incorporated gene mutation status. The ensemble mortality output by the RSF was used as the risk score for each patient.

2.6. Stratifying risk score to obtain prognostic categories

Given a risk score for each individual, categories were assigned by constructing a survival tree [9] using the R package rpart with default parameters. Briefly, at each split, the tree assigns the score threshold that optimally separates the patients by overall survival. Splitting is recursively performed until there are too few patients to meaningfully split based on risk score. The leaf nodes in the final tree determine the class assignments, which are therefore determined by risk score thresholds.

3. Results

3.1. Training and validation cohorts

Our training cohort comprised 274 myeloid malignancy patients, including patients diagnosed with MDS (*N* = 75), MDS/MPN (*N* = 33), and AML (*N* = 22), augmented with 144 AML patients from TCGA [18]. The validation set comprised 507 myeloid malignancy patients (*N* = 236, 97, and 174 for MDS, MDS/MPN, and AML, respectively). Patient characteristics are summarized in Table 1. We reasoned that a pan-diagnosis classifier would perform well provided it was trained on a cohort with substantial representation from all three diagnoses. Furthermore, using the diagnoses themselves as input variables will account for any diagnosis-specific effects. Our strategy was to build two different RSFs on the training set data, the first using clinical data only and the second also incorporating mutation data. Both RSF classifiers were then tested on the validation set (Fig. 1; see Section 2). This strategy allows us to assess the additional prognostic accuracy conferred by incorporating gene mutations.

3.2. Random forest classification using standard clinical variables

We first built a RSF using the clinical variables – cytogenetics, diagnoses, absolute neutrophil count, bone marrow blasts, hemoglobin, and platelets – shown in Table 1, to obtain a risk classification termed RSF_{clin} (Supplementary Fig. S2). Classifying features comprise both continuous (*e.g.* hemoglobin) and binary (cytogenetic features) variables. The variable importance measure in the randomForestSRC package is known to unfairly favor continuous variables over categorical ones [14]. We therefore deliberately treated all classifying features as binary variables as this allows all variables to be considered equally. The continuous variables were stratified by well-established clinically relevant thresholds [11].

As expected, the RSF_{clin} risk categories corresponded closely with actual overall survival in the training set ($P = 5.34 \times 10^{-47}$; Fig. 2A). The classifier remained robust when applied to the validation set as a whole ($P = 8.32 \times 10^{-12}$; Poor vs. Good prognosis categories hazard ratio 3.56; Fig. 2B), supporting the use of the RSF as an alternative approach to assessing patient risk from standard myeloid malignancy clinical features. The classifier was somewhat predictive within specific diagnoses ($P = 6.6 \times 10^{-3}$, 0.0369, and 5.01×10^{-3} for MDS, MDS/MPN, and AML, respectively; Supplementary Fig. S3), but left room for improvement. We next sought

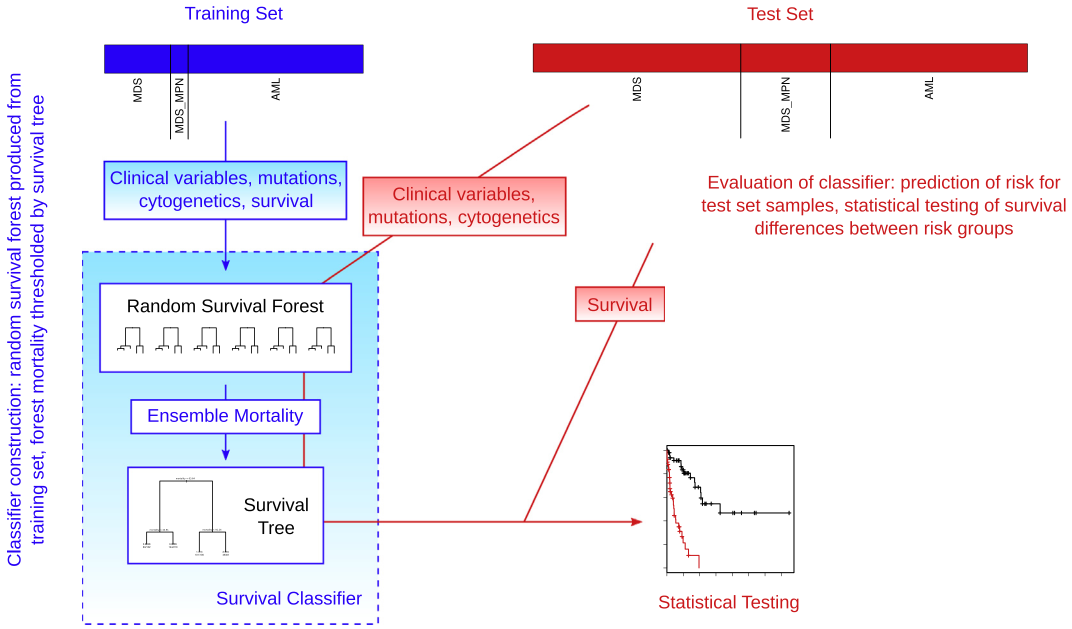

Training Set
Test Set
MDS
MDS_MPN
AML
MDS
MDS_MPN
AML
Classifier construction: random survival forest produced from training set, forest mortality thresholded by survival tree
Clinical variables, mutations, cytogenetics, survival
Clinical variables, mutations, cytogenetics
Evaluation of classifier: prediction of risk for test set samples, statistical testing of survival differences between risk groups
Random Survival Forest
Survival
Ensemble Mortality
Survival Tree
Survival Classifier
Statistical Testing

Fig. 1. Conceptual overview of prognostic classifier construction and validation. The random survival forest is built on the training cohort, incorporating clinical variables (for RSF_{clin}) or clinical variables and WES data (for $\text{RSF}_{\text{clin+mut}}$). The resulting ensemble mortality values are stratified by a survival tree to determine categories. The same forest and stratifying thresholds are then applied to the validation cohort and performance is assessed.

to determine whether somatic mutation status could be used to improve accuracy in patient risk classification.

3.3. Whole-exome sequencing shows a heterogeneous mutational spectrum across patients

WES of our training cohort of 274 myeloid malignancy patients was followed by a bioanalytic and validation pipeline (Supplementary Fig. S1, Supplementary Table S1; see Section 2). Overall, 5548 genes had a validated non-synonymous or splice site mutation in at least one patient. The median number of mutations per patient was 14, and $\sim$97% carried at least one mutation. Four genes – *DNMT3A*, *FLT3*, *NPM1*, and *TET2* – harbored mutations in more than 10% of patients. Given the cohort size of 274, theoretical expectation (Supplementary Fig. S4) is that a gene with a general patient population mutation frequency of 10% would be mutated in approximately 27 patients (95% CI 18–37) in our cohort, while those with frequency 3% would be expected to be mutated in 8 patients (95% CI 3–14). It is unlikely that genes mutated at frequencies lower than these would have detectable impact on patient outcome. Therefore, we restricted the feature space to the mutational status of the 71 genes mutated in at least five patient samples (training cohort shown in Supplementary Fig. S5, validation cohort shown in Supplementary Fig. S6).

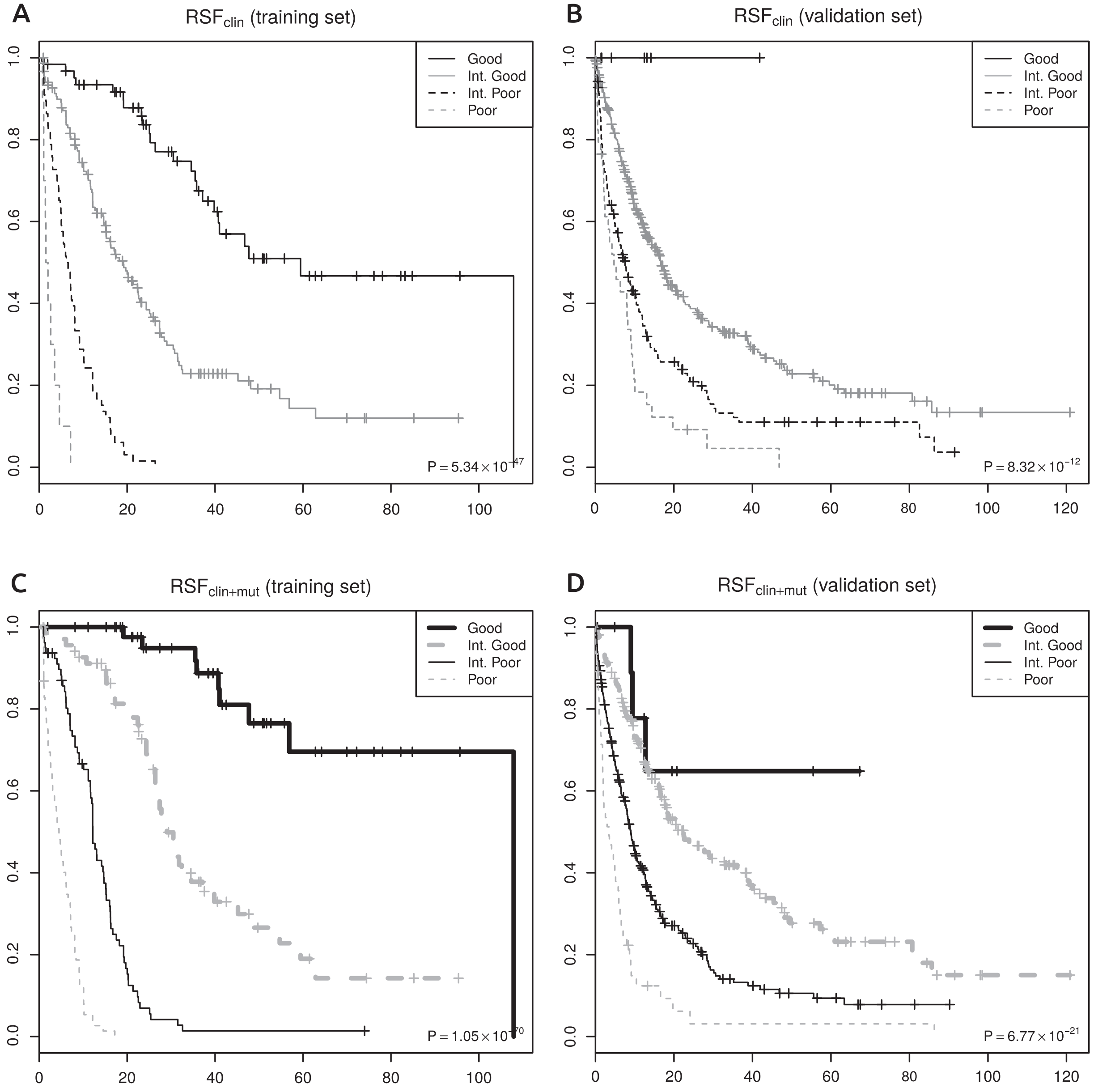
3.4. Incorporating mutational information into the risk classifier

To select the gene mutations most associated with patient outcome, we first applied the RSF variable selection procedure to the clinical variables and 71 gene mutations, ranking these classifying features in terms of importance (Supplementary Fig. S7). As downstream classifying mutational features, we only included genes ranked among the most important, as well as those previously reported as significant in myeloid malignancies [18,28,12] (Supplementary Table S2). These 30 classifying genes, along with the clinical variables, served as input to construct the second RSF classifier, termed $\text{RSF}_{\text{clin+mut}}$ (Supplementary Fig. S8). $\text{RSF}_{\text{clin+mut}}$ has improved performance over RSF_{clin} on the training cohort ($P = 1.05 \times 10^{-70}$; Fig. 2C).

To apply $\text{RSF}_{\text{clin+mut}}$ to the validation cohort, we designed a custom enrichment panel that included the 30 classifying genes. The enrichment products for each gene were deep-sequenced, and each patient was annotated with presence or absence of mutation in each gene. These values, as well as the clinical variable values, were run through the $\text{RSF}_{\text{clin+mut}}$ classifier for each patient. The risk classifier was very predictive of actual patient outcome in the validation cohort ($P = 6.77 \times 10^{-21}$; Fig. 2D and E). The two-year survival rates for the patients assigned to the good, intermediate good, intermediate poor, and poor prognostic categories were 64%, 47%, 23%, and 6%, respectively. The classifier's performance remained robust even within individual diagnoses ($P = 3.41 \times 10^{-7}$, 3.4×10^{-5}, and 4.28×10^{-5}, respectively within MDS, MDS/MPN, and AML; Supplementary Fig. S9).

3.5. Gene mutations modify the effects of clinical variables

Closer examination of the relationships between the risk groups and classifying variables gives a complex picture of the variables' impact on patient outcome (Supplementary Fig. S10; Supplementary Fig. S11). The improved performance of $\text{RSF}_{\text{clin+mut}}$ over RSF_{clin} implies that the mutational information provides additional prognostic information not available from clinical variables, and therefore is able to modify risk categories assigned by clinical variables to make predictions more accurate. On a univariate level, mutations in six genes – *DNMT3A*, *EZH2*, *FLT3*-TKD, *RUNX1*, *SF3B1*, and *TP53* – are strongly associated with RSF-determined mortality (Supplementary Fig. S11; Supplementary Table S2). Mutations in all of these genes, save *SF3B1*, are associated with poorer outcomes. Individuals with wild-type SF3B1 collectively have higher mortality, suggesting prognostic benefit from *SF3B1* mutation. Individual patient risk categories assigned by $\text{RSF}_{\text{clin+mut}}$ and RSF_{clin} show some concordance (Kendall $\tau = 0.502$; Fig. 3A), but inclusion of the mutational information tunes risk assignment in a way that yields clear improvement in prognostic accuracy (Fig. 3B). We next investigated the reasons for this improvement by examining the



E

	RSF_{clin}		$RSF_{clin+mut}$	
	HR vs. Good	Two-year survival	HR vs. Good	Two-year survival
Good	-	44.4%	-	64.9%
Int. 1	1.46	28.2%	2.24	47.3%
Int. 2	2.22	20.0%	4.73	23.3%
Poor	3.56	8.9%	10.39	6.2%

Fig. 2. Kaplan–Meier curves for random survival survival forest classifiers. (A) RSF_{clin} applied to the training set. (B) RSF_{clin} applied to the validation set. (C) $RSF_{clin+mut}$ applied to the training set. (D) $RSF_{clin+mut}$ applied to the validation set. (E) Hazard ratio estimates and median two-year survival for RSF_{clin} and $RSF_{clin+mut}$ categories in validation set.

modifying effect of the mutations on the clinical features' impact on patient outcome.

Analysis of statistical interactions between mutations and clinical variables show that specific gene mutations appear to exacerbate the prognostic impact of certain clinical variables. These risk modifications likely partially explain the enhanced prognostic accuracy conferred by mutational information. In particular, the presence of *NRAS* or *PHF6* mutation in combination with trisomy (8) is associated with a particularly poor prognosis (interaction $P = 0.002$ and 0.02, respectively; Fig. 4A). Mutations in *CBL* or *TP53* appear to worsen the impact of elevated bone marrow blasts (interaction $P = 0.09$ and 0.002, respectively; Fig. 4B).

3.6. Comparison with existing prognostic classifiers

A number of recent studies have incorporated mutational information for MDS patient risk classification. The studies use either both clinical and mutational data [28,38] or only mutational data

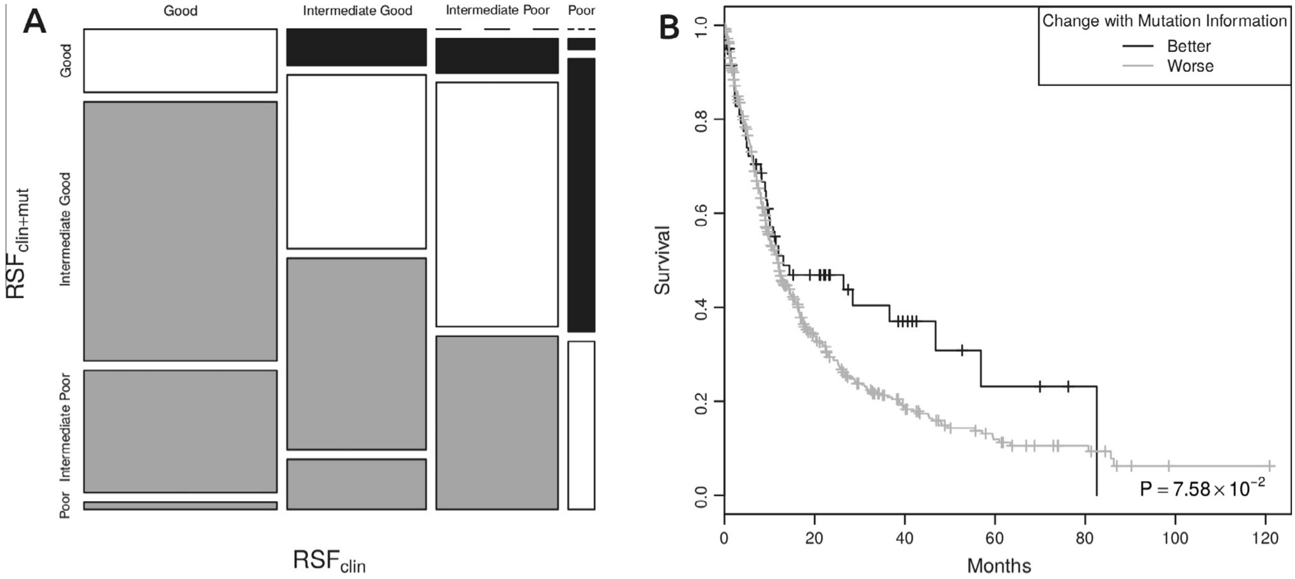


Fig. 3. Gene mutations modify RSF_{clin}-derived patient prognosis. (A) Mosaic plot comparing RSF_{clin} and $RSF_{clin+mut}$ in total cohort (training and validation). The rectangle areas indicate relative numbers of patients in each prognostic category intersection. Shading indicates patients whose prognostic category improved (black) or worsened (gray) from RSF_{clin} to $RSF_{clin+mut}$ (B) Kaplan–Meier curves for patients whose risk category changed between RSF_{clin} and $RSF_{clin+mut}$ classifiers. Here the columns correspond to the RSF classification, and are each divided into four blocks, ordered according to $RSF_{clin+mut}$ classification (from Poor at bottom to Good at top). The dashed lines indicate that there are no patients in the corresponding prognostic category intersection.

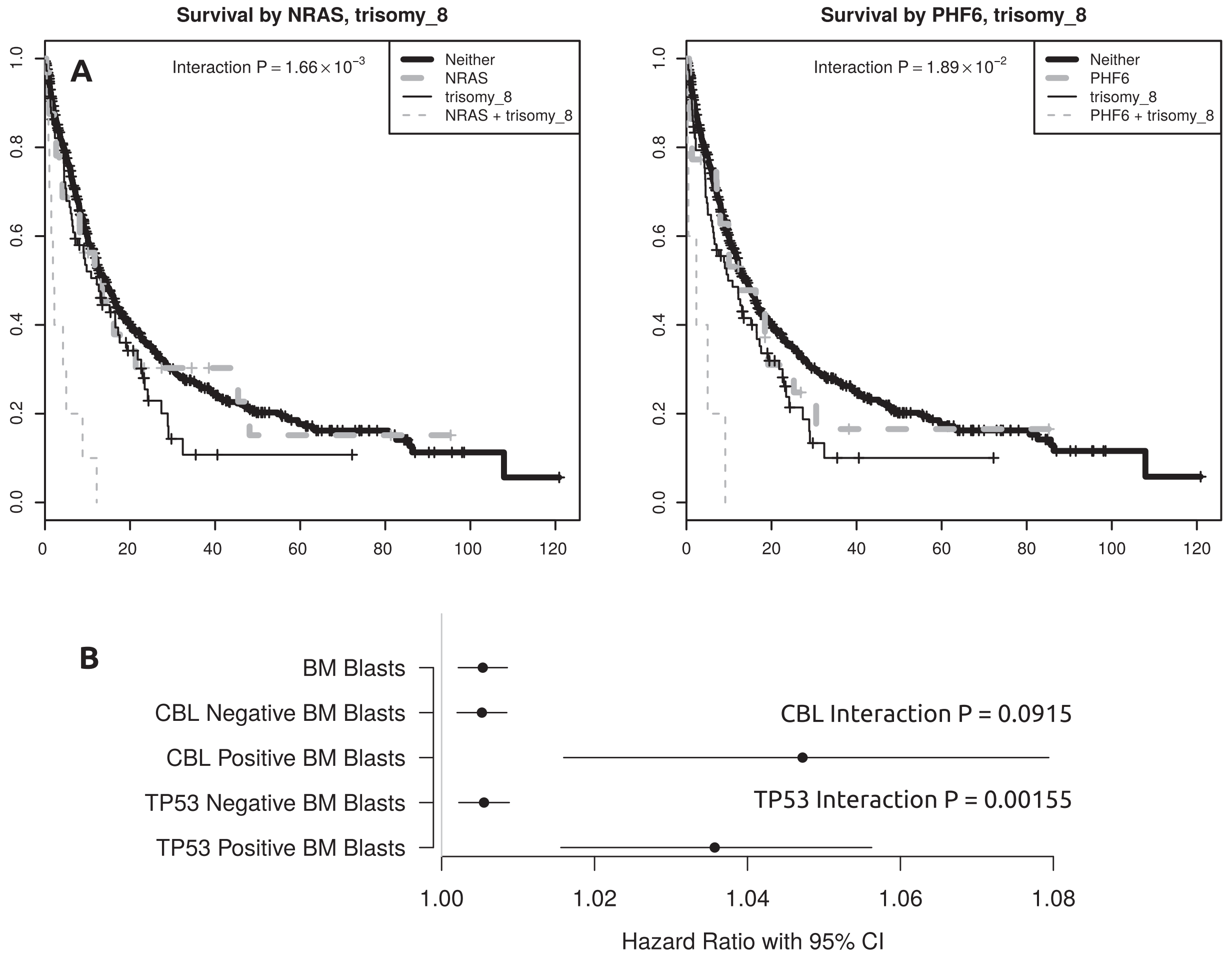


Fig. 4. Statistical interactions between gene mutations and clinical variables. (A) Kaplan–Meier curves showing interactions between gene mutations and cytogenetic features. In both cases, mutations in the gene (*NRAS* and *PHF6*) combined with trisomy(8) are associated with significantly worse prognosis. (B) Hazard ratios and confidence intervals showing interactions between gene mutations and bone marrow blasts. Hazard ratios are shown overall (top of the panel), as well as for patients stratified by gene mutation status (positive and negative for mutation).

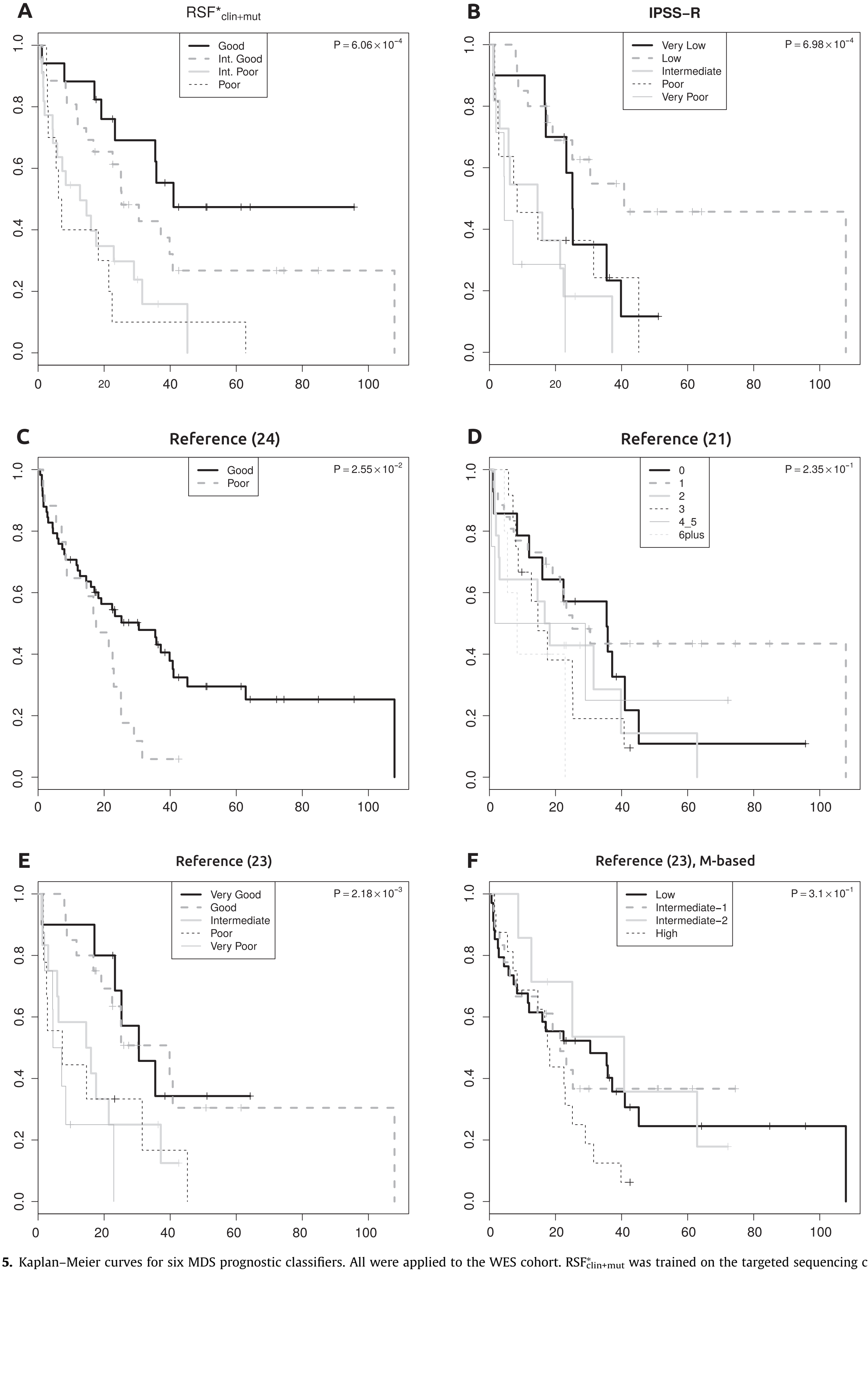


Fig. 5. Kaplan–Meier curves for six MDS prognostic classifiers. All were applied to the WES cohort. $RSF^{*}_{clin+mut}$ was trained on the targeted sequencing cohort.

[2,38]. Two of the classifiers [28,38] incorporate mutational status of genes that were not included as part of our validation cohort targeted sequencing assay. Therefore, we assess prognostic accuracy for all classification schemes using our WES (training) cohort, since all schemes' classifying genes are assayed. To ensure a fair comparison with our RSF approach, we exchange our training and validation cohorts from above, training a RSF on the targeted sequencing cohort to assess its performance on the WES cohort. The resulting RSF classifier, termed $\text{RSF}^*_{\text{clin+mut}}$ is restricted to MDS patients. The performance of $\text{RSF}^*_{\text{clin+mut}}$ surpasses that of the other methods, and is comparable to that of IPSS-R. [11] (Fig. 5; Supplementary Fig. S12).

3.7. Implementation and public availability

One disadvantage of a RSF classifier is the lack of transparency. Methods that use smaller numbers of classifying variables may be represented as scoring systems that can be communicated as a look-up table [38] or as lists of genes [28], facilitating biological interpretation. The RSFs presented here involved thousands of decision trees that collectively yield the mortality score. Straightforward presentation of the classification criteria is therefore challenging. Nonetheless, the risk assignments of $\text{RSF}_{\text{clin+mut}}$ are deterministic in the sense that a patient's combination of clinical parameters and mutation status in the 30 genes uniquely determines risk category. We have implemented the risk assignment as a web-based tool (http://myeloid-risk.case.edu/) that allows the user to input the value of each classifying variable, and returns risk category based on these values. The RSF object itself is also freely downloadable as an R [31] object from the same site.

4. Discussion

In this study, we have developed and implemented an RSF-based procedure to classify patients into prognostic categories based on histomorphological, cytogenetic, and molecular features. Ours is the first study, to our knowledge, to build and independently validate a classifier that initially considers all coding genes affected by somatic mutations. This is only possible with WES data. With comprehensive gene mutation data, finding clinically relevant combinations of mutations is highly nontrivial. The RSF structure enables simultaneous consideration of many genes, automatically accommodating complex conditional dependencies between mutation status and patient outcome.

We have deliberately constructed a pan-classifier that is designed to work across three patient diagnoses. The rationale for this is that diagnostic categories are somewhat arbitrary (*e.g.* thresholds on blast counts) and subjective. Here our risk categories are predictive across the patient diagnostic subsets, and indeed compare favorably against MDS-specific schemes. Testing on a larger independent patient set would be necessary for a more conclusive comparison, however.

Our results shed additional light on the impact of specific genes' mutations on patient outcome. Integrating mutational information into the clinical variable-based classifier has the effect of shifting a subset of patients into different risk categories (Fig. 3). Testing associations between the RSF-generated mortality measure and mutational status underscored the impact of gene mutations previously reported as being myeloid malignancy-related (Supplementary Fig. S11; Supplementary Table S2). Among these, mutation of *TP53* is a well-established indicator of poor outcome [16,15,2]. The same holds true for *FLT3* mutations, particularly for patients with a normal karyotype [7]. *DNMT3A* mutations have been linked to poor survival in AML [34,32] and MDS [36], as have *RUNX1* mutations [2]. Mutations in *EZH2* were first reported in MDS/MPN as being associated with poor outcome [6]. On the other hand, we found that *SF3B1* mutation was associated with more positive patient outcome, which has previously been reported [27,12] in myeloid malignancy.

The study presented here does have some limitations. First, we only consider the presence or absence of a mutation, and do not consider clonal abundance. A recent study reported that the impact of a mutation on patient outcome does not depend upon its allelic abundance [28]. Therefore, the dichotomous approach may be optimal. Second, we are treating all non-synonymous mutations equally and do not consider their differential impacts on protein as measured by various bioinformatics tools [42,43]. We are also disregarding synonymous and non-coding region mutations, which can have function impact. Third, our sample size is smaller relative to some recent myeloid malignancy classification studies. The results presented here should be validated and improved using larger cohorts. Finally, as mentioned above, the nature of RSFs limits the ability to gain biological insight from their output. Nonetheless, we have shown that the RSF results may be used to identify genes whose mutations specifically modify the effects of clinical variables.

We view the approach presented here as a general framework to incorporate WES data into prognostic systems for cancer patients. Applied to gene-specific mutation data, the RSF can be used to agnostically identify relevant patient subgroups from cancer of any type. Although routine whole-exome sequencing on patients is currently prohibitively expensive for most centers, precipitous cost decreases will make such sequencing feasible in the near future [23]. Furthermore, mutation-based classification has the advantage of being less subjective than morphology-based criteria [8]. The framework proposed here could also be adapted to other complex tumor genome data such as non-coding RNA, DNA methylation, and histone modification. Ongoing research efforts will make large genomic data sets increasingly publicly available, improving the community's ability to accurately determine the impact of genomic features on clinical outcomes. As genomic data becomes ubiquitous and more complex, correspondingly complex prognostic schemes like the one presented here may be the wave of the future. One may envision a direct link between a patient's electronic medical record and software to compute a more accurate predicted outcome. Already, collaborations between IBM and U.S. cancer centers are underway to train a highly sophisticated computer system, using vast amounts of patient data, to inform clinical decision making (IBM Watson Oncology [40]). In that spirit, accompanying this study is a web tool implementing our classification software. Achieving the highest degree of prognostic precision will ultimately rely on optimal use of multiple sources of patient-specific data.

Translational relevance

Currently, prognostic schemes that are in clinical use for myeloid malignancies suffer from heterogeneity within classes and poor distinction between classes. Since these schemes are based on morphological and cytogenetic features, we sought to determine whether improvement could be gained by including gene-specific mutation information. Large-scale incorporation of mutational data has only recently become feasible because of advances in DNA sequencing technology. We proceeded by training a mutation-based risk classifier on a large set of myeloid malignancy patients, and validating its accuracy on an independent cohort. Improved prognostic accuracy would serve to better guide treatment choices. Indeed, patients that are placed into a favorable risk class with a high degree of confidence could avoid unnecessary hazard associated with certain treatments. Additionally, individuals solidly in a high-risk category may benefit from more experimental treatments.

Financial support

This work was supported by National Institutes of Health grants R01CA-131341 (T.L.), R01HL-082983 (J.P.M.), U54 RR019391 (J.P. M.), and K24 HL-077522 (J.P.M.), as well as American Cancer Society grant 123436-RSG-12-159-01-DMC (T.L.), and a Scott Hamilton CARES grant (H.M.).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

The authors would like to thank Professors S. Miyano and S. Ogawa for access to and help with the Genomon software.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.10.003.

References

[1] M. Banerjee, D.G. Muenz, J.T. Chang, M. Papaleontiou, M.R. Haymart, Tree-based model for thyroid cancer prognostication, J. Clin. Endocrinol. Metab. 99 (10) (2014) 3737–3745.

[2] R. Bejar, K. Stevenson, O. Abdel-Wahab, N. Galili, B. Nilsson, G. Garcia-Manero, H. Kantarjian, A. Raza, R.L. Levine, D. Neuberg, B.L. Ebert, Clinical effect of point mutations in myelodysplastic syndromes, N. Engl. J. Med. 364 (26) (2011) 2496–2506 (6).

[3] J.M. Bennett, D. Catovsky, M.T. Daniel, G. Flandrin, D.A. Galton, H.R. Gralnick, C. Sultan, Proposals for the classification of the myelodysplastic syndromes, Br. J. Haematol. 51 (2) (1982) 189–199.

[4] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[5] H. Dohner, E.H. Estey, S. Amadori, F.R. Appelbaum, T. Buchner, A.K. Burnett, H. Dombret, P. Fenaux, D. Grimwade, R.A. Larson, F. Lo-Coco, T. Naoe, D. Niederwieser, G.J. Ossenkoppele, M.A. Sanz, J. Sierra, M.S. Tallman, B. Lowenberg, C.D. Bloomfield, Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet, Blood 115 (3) (2010) 453–474.

[6] T. Ernst, A.J. Chase, J. Score, C.E. Hidalgo-Curtis, C. Bryant, A.V. Jones, K. Waghorn, K. Zoi, F.M. Ross, A. Reiter, A. Hochhaus, H.G. Drexler, A. Duncombe, F. Cervantes, D. Oscier, J. Boultwood, F.H. Grand, N.C. Cross, Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders, Nat. Genet. 42 (8) (2010) 722–726.

[7] E.H. Estey, Acute myeloid leukemia: 2013 update on risk-stratification and management, Am. J. Hematol. 88 (4) (2013) 318–327.

[8] P. Font, J. Loscertales, C. Benavente, A. Bermejo, M. Callejas, L. Garcia-Alonso, A. Garcia-Marcilla, S. Gil, M. Lopez-Rubio, E. Martin, C. Munoz, P. Ricard, C. Soto, P. Balsalobre, A. Villegas, Inter-observer variance with the diagnosis of myelodysplastic syndromes (MDS) following the 2008 WHO classification, Ann. Hematol. 92 (1) (2013) 19–24.

[9] L. Gordon, R.A. Olshen, Tree-structured survival analysis, Cancer Treat. Rep. 69 (10) (1985) 1065–1069.

[10] P. Greenberg, C. Cox, M.M. LeBeau, P. Fenaux, P. Morel, G. Sanz, M. Sanz, T. Vallespi, T. Hamblin, D. Oscier, K. Ohyashiki, K. Toyama, C. Aul, G. Mufti, J. Bennett, International scoring system for evaluating prognosis in myelodysplastic syndromes, Blood 89 (6) (1997) 2079–2088.

[11] P.L. Greenberg, H. Tuechler, J. Schanz, G. Sanz, G. Garcia-Manero, F. Sole, J.M. Bennett, D. Bowen, P. Fenaux, F. Dreyfus, H. Kantarjian, A. Kuendgen, A. Levis, L. Malcovati, M. Cazzola, J. Cermak, C. Fonatsch, M.M. Le Beau, M.L. Slovak, O. Krieger, M. Luebbert, J. Maciejewski, S.M. Magalhaes, Y. Miyazaki, M. Pfeilstocker, M. Sekeres, W.R. Sperr, R. Stauder, S. Tauro, P. Valent, T. Vallespi, A.A. van de Loosdrecht, U. Germing, D. Haase, Revised international prognostic scoring system for myelodysplastic syndromes, Blood 120 (12) (2012) 2454–2465.

[12] T. Haferlach, Y. Nagata, V. Grossmann, Y. Okuno, U. Bacher, G. Nagae, S. Schnittger, M. Sanada, A. Kon, T. Alpermann, K. Yoshida, A. Roller, N. Nadarajah, Y. Shiraishi, Y. Shiozawa, K. Chiba, H. Tanaka, H.P. Koeffler, H.U. Klein, M. Dugas, H. Aburatani, A. Kohlmann, S. Miyano, C. Haferlach, W. Kern, S. Ogawa, Landscape of genetic lesions in 944 patients with myelodysplastic syndromes, Leukemia 28 (2) (2014) 241–247.

[13] H. Ishwaran, U.B. Kogalur, Random Forests for Survival, Regression and Classification (RF-SRC), R Package Version 1.4, 2013.

[14] Hemant Ishwaran, Udaya B. Kogalur, Eiran Z. Gorodeski, Andy J. Minn, Michael S. Lauer, High-dimensional variable selection for survival data, J. Am. Stat. Assoc. 105 (489) (2010) 205–217.

[15] M. Jadersten, L. Saft, A. Smith, A. Kulasekararaj, S. Pomplun, G. Gohring, A. Hedlund, R. Hast, B. Schlegelberger, A. Porwit, E. Hellstrom-Lindberg, G.J. Mufti, TP53 mutations in low-risk myelodysplastic syndromes with del(5q) predict disease progression, J. Clin. Oncol. 29 (15) (2011) 1971–1979.

[16] A.G. Kulasekararaj, A.E. Smith, S.A. Mian, A.M. Mohamedali, P. Krishnamurthy, N.C. Lea, J. Gaken, C. Pennaneach, R. Ireland, B. Czepulkowski, S. Pomplun, J.C. Marsh, G.J. Mufti, TP53 mutations in myelodysplastic syndrome are strongly correlated with aberrations of chromosome 5, and correlate with adverse prognosis, Br. J. Haematol. 160 (5) (2013) 660–672.

[17] T. LaFramboise, Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances, Nucl. Acids Res. 37 (13) (2009) 4181–4193.

[18] T.J. Ley, C. Miller, L. Ding, B.J. Raphael, A.J. Mungall, A. Robertson, K. Hoadley, T. J. Triche, P.W. Laird, J.D. Baty, L.L. Fulton, R. Fulton, S.E. Heath, J. Kalicki-Veizer, C. Kandoth, J.M. Klco, D.C. Koboldt, K.L. Kanchi, S. Kulkarni, T.L. Lamprecht, D.E. Larson, L. Lin, C. Lu, M.D. McLellan, J.F. McMichael, J. Payton, H. Schmidt, D.H. Spencer, M.H. Tomasson, J.W. Wallis, L.D. Wartman, M.A. Watson, J. Welch, M. C. Wendl, A. Ally, M. Balasundaram, I. Birol, Y. Butterfield, R. Chiu, A. Chu, E. Chuah, H.J. Chun, R. Corbett, N. Dhalla, R. Guin, A. He, C. Hirst, M. Hirst, R.A. Holt, S. Jones, A. Karsan, D. Lee, H.I. Li, M.A. Marra, M. Mayo, R.A. Moore, K. Mungall, J. Parker, E. Pleasance, P. Plettner, J. Schein, D. Stoll, L. Swanson, A. Tam, N. Thiessen, R. Varhol, N. Wye, Y. Zhao, S. Gabriel, G. Getz, C. Sougnez, L. Zou, M.D. Leiserson, F. Vandin, H.T. Wu, F. Applebaum, S.B. Baylin, R. Akbani, B. M. Broom, K. Chen, T.C. Motter, K. Nguyen, J.N. Weinstein, N. Zhang, M.L. Ferguson, C. Adams, A. Black, J. Bowen, J. Gastier-Foster, T. Grossman, T. Lichtenberg, L. Wise, T. Davidsen, J.A. Demchok, K.R. Shaw, M. Sheth, H.J. Sofia, L. Yang, J.R. Downing, G. Eley, S. Alonso, B. Ayala, J. Baboud, M. Backus, S.P. Barletta, D.L. Berton, A.L. Chu, S. Girshik, M.A. Jensen, A. Kahn, P. Kothiyal, M.C. Nicholls, T.D. Pihl, D.A. Pot, R. Raman, R.N. Sanbhadti, E.E. Snyder, D. Srinivasan, J. Walton, Y. Wan, Z. Wang, J.P. Issa, M. Le Beau, M. Carroll, H. Kantarjian, S. Kornblau, M.S. Bootwalla, P.H. Lai, H. Shen, D.J. Van Den Berg, D.J. Weisenberger, D.C. Link, M.J. Walter, B.A. Ozenberger, E.R. Mardis, P. Westervelt, T.A. Graubert, J.F. DiPersio, R.K. Wilson, Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia, N. Engl. J. Med. 368 (22) (2013) 2059–2074.

[19] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.

[20] J.P. Maciejewski, G.J. Mufti, Whole genome scanning as a cytogenetic tool in hematologic malignancies, Blood 112 (4) (2008) 965–974.

[21] L. Malcovati, M.G. Porta, C. Pascutto, R. Invernizzi, M. Boni, E. Travaglino, F. Passamonti, L. Arcaini, M. Maffioli, P. Bernasconi, M. Lazzarino, M. Cazzola, Prognostic factors and life expectancy in myelodysplastic syndromes classified according to WHO criteria: a basis for clinical decision making, J. Clin. Oncol. 23 (30) (2005) 7594–7603 (10).

[22] E.A. Manilich, R.P. Kiran, T. Radivoyevitch, I. Lavery, V.W. Fazio, F.H. Remzi, A novel data-driven prognostic model for staging of colorectal cancer, J. Am. Coll. Surg. 213 (5) (2011) 579–588.

[23] E.R. Mardis, A decade's perspective on DNA sequencing technology, Nature 470 (7333) (2011) 198–203 (2).

[24] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (9) (2010) 1297–1303.

[25] M. Meyerson, S. Gabriel, G. Getz, Advances in understanding cancer genomes through second-generation sequencing, Nat. Rev. Genet. 11 (10) (2010) 685–696.

[26] Y. Nannya, M. Sanada, K. Nakazaki, N. Hosoya, L. Wang, A. Hangaishi, M. Kurokawa, S. Chiba, D.K. Bailey, G.C. Kennedy, S. Ogawa, A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays, Cancer Res. 65 (14) (2005) 6071–6079.

[27] E. Papaemmanuil, M. Cazzola, J. Boultwood, L. Malcovati, P. Vyas, D. Bowen, A. Pellagatti, J.S. Wainscoat, E. Hellstrom-Lindberg, C. Gambacorti-Passerini, A.L. Godfrey, I. Rapado, A. Cvejic, R. Rance, C. McGee, P. Ellis, L.J. Mudie, P.J. Stephens, S. McLaren, C.E. Massie, P.S. Tarpey, I. Varela, S. Nik-Zainal, H.R. Davies, A. Shlien, D. Jones, K. Raine, J. Hinton, A.P. Butler, J.W. Teague, E.J. Baxter, J. Score, A. Galli, M.G. Della Porta, E. Travaglino, M. Groves, S. Tauro, N. C. Munshi, K.C. Anderson, A. El-Naggar, A. Fischer, V. Mustonen, A.J. Warren, N. C. Cross, A.R. Green, P.A. Futreal, M.R. Stratton, P.J. Campbell, Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts, N. Engl. J. Med. 365 (15) (2011) 1384–1395.

[28] E. Papaemmanuil, M. Gerstung, L. Malcovati, S. Tauro, G. Gundem, P. Van Loo, C.J. Yoon, P. Ellis, D.C. Wedge, A. Pellagatti, A. Shlien, M.J. Groves, S.A. Forbes, K. Raine, J. Hinton, L.J. Mudie, S. McLaren, C. Hardy, C. Latimer, M.G. Della Porta, S. O'Meara, I. Ambaglio, A. Galli, A.P. Butler, G. Walldin, J.W. Teague, L. Quek, A. Sternberg, C. Gambacorti-Passerini, N.C. Cross, A.R. Green, J. Boultwood, P. Vyas, E. Hellstrom-Lindberg, D. Bowen, M. Cazzola, M.R. Stratton, P.J. Campbell, Clinical and biological implications of driver mutations in myelodysplastic syndromes, Blood 122 (22) (2013) 3616–3627.

[29] P. Paschka, R.F. Schlenk, V.I. Gaidzik, M. Habdank, J. Kronke, L. Bullinger, D. Spath, S. Kayser, M. Zucknick, K. Gotze, H.A. Horst, U. Germing, H. Dohner, K. Dohner, IDH1 and IDH2 mutations are frequent genetic alterations in acute myeloid leukemia and confer adverse prognosis in cytogenetically normal acute myeloid leukemia with NPM1 mutation without FLT3 internal tandem duplication, J. Clin. Oncol. 28 (22) (2010) 3636–3643.

[30] J.P. Patel, M. Gonen, M.E. Figueroa, H. Fernandez, Z. Sun, J. Racevskis, P. Van Vlierberghe, I. Dolgalev, S. Thomas, O. Aminova, K. Huberman, J. Cheng, A.

Viale, N.D. Socci, A. Heguy, A. Cherry, G. Vance, R.R. Higgins, R.P. Ketterling, R.E. Gallagher, M. Litzow, M.R. van den Brink, H.M. Lazarus, J.M. Rowe, S. Luger, A. Ferrando, E. Paietta, M.S. Tallman, A. Melnick, O. Abdel-Wahab, R.L. Levine, Prognostic relevance of integrated genetic profiling in acute myeloid leukemia, N. Engl. J. Med. 366 (12) (2012) 1079–1089.

[31] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2012, ISBN 3-900051-07-0.

[32] A.F. Ribeiro, M. Pratcorona, C. Erpelinck-Verschueren, V. Rockova, M. Sanders, S. Abbas, M.E. Figueroa, A. Zeilemaker, A. Melnick, B. Lowenberg, P.J. Valk, R. Delwel, Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia, Blood 119 (24) (2012) 5824–5831.

[33] Terry Therneau, A Package for Survival Analysis in S, R Package Version 2.36-14, 2012.

[34] F. Thol, F. Damm, A. Ludeking, C. Winschel, K. Wagner, M. Morgan, H. Yun, G. Gohring, B. Schlegelberger, D. Hoelzer, M. Lubbert, L. Kanz, W. Fiedler, H. Kirchner, G. Heil, J. Krauter, A. Ganser, M. Heuser, Incidence and prognostic influence of DNMT3A mutations in acute myeloid leukemia, J. Clin. Oncol. 29 (21) (2011) 2889–2896.

[35] C.S. Tremblay, T. Hoang, T. Hoang, Early T cell differentiation lessons from T-cell acute lymphoblastic leukemia, Prog. Mol. Biol. Trans. Sci. 92 (2010) 121–156.

[36] M.J. Walter, L. Ding, D. Shen, J. Shao, M. Grillot, M. McLellan, R. Fulton, H. Schmidt, J. Kalicki-Veizer, M. O'Laughlin, C. Kandoth, J. Baty, P. Westervelt, J.F. DiPersio, E.R. Mardis, R.K. Wilson, T.J. Ley, T.A. Graubert, Recurrent DNMT3A mutations in patients with myelodysplastic syndromes, Leukemia 25 (7) (2011) 1153–1158.

[37] G.B. Wertheim, C. Smith, M. Luskin, A. Rager, M.E. Figueroa, M. Carroll, S.R. Master, Validation of DNA methylation to predict outcome in acute myeloid leukemia by use of xMELP, Clin. Chem. (2014).

[38] L. Xu, Z.H. Gu, Y. Li, J.L. Zhang, C.K. Chang, C.M. Pan, J.Y. Shi, Y. Shen, B. Chen, Y. Y. Wang, L. Jiang, J. Lu, X. Xu, J.L. Tan, Y. Chen, S.Y. Wang, X. Li, Z. Chen, S.J. Chen, Genomic landscape of CD34+ hematopoietic cells in myelodysplastic syndrome and gene mutation profiles as prognostic markers, Proc. Natl. Acad. Sci. USA 111 (23) (2014) 8589–8594.

[39] Y. Yuan, E.M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K.R. Hess, L. Diao, L. Han, X. Huang, M.S. Lawrence, J.N. Weinstein, J.M. Stuart, G.B. Mills, L.A. Garraway, A.A. Margolin, G. Getz, H. Liang, Assessing the clinical utility of cancer genomic and proteomic data across tumor types, Nat. Biotechnol. 32 (7) (2014) 644–652.

[40] Marjorie Glass Zauderer, Ayca Gucalp, Andrew S. Epstein, Andrew David Seidman, Aryeh Caroline, Svetlana Granovsky, Julia Fu, Jeffrey Keesing, Scott Lewis, Heather Co, et al. Piloting IBM Watson oncology within memorial Sloan Kettering's regional network, in: ASCO Annual Meeting Proceedings, vol. 32, p. e17653, 2014.

[41] G. Marcucci, T. Haferlach, H. Döhner, Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications, J. Clin. Oncol. 29 (5) (2011) 475–486.

[42] P.C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucl. Acids Res. 31 (13) (2003) 3812–3814.

[43] I. Adzhubei, D.M. Jordan, S.R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2, Curr. Protoc. Hum. Genet., Unit7.20, 2013 (Chapter 7).