Supplemental Text S1

Morphological features of monocytes and monoblasts, and erythroid dysplastic changes

To address phenotypic features of *SRSF2* mutation we investigated CMML patients with a *SRSF2* mutation and patients without mutation. We could not detect statistically significant discrepancies in dysgranulopoiesis, dysmegakaryopoiesis or dyserythropoiesis between the two groups. However, according to the WHO criteria for dysplasia, 76.2% of patients were dysplastic in at least one lineage, 32.0% in two lineages and only 8.2% of patients in all 3 lineages. These aspects were not different in patients with *SRSF2* mutation or patients with *SRSF2* wild-type.

In addition, we investigated the number of peripheral monoblasts as well as monocytes in cases being mutated or non-mutated, and also did not see statistically significant differences. This was also true for any other morphological aspect such as vacuolisation etc.

In silico analyses

Effects of molecular variants

In order to estimate the damaging character of the missense mutations at Pro95 as well as the novel mutations p.[Pro96_Arg103del;Pro107His], p.Arg86_Gly93dup, and p.Arg94_Pro95insArg we used three different algorithms: SIFT (http://sift-jcvi.org),[1] PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/index.shtml),[2] and MutationTaster (www.mutationtaster.org).[3] The table below summarizes the predicted effects for all mutations.

Mutation	**PolyPhen-2**	**SIFT**	**MuatationTaster**
p.Pro95His	probably damaging	damaging	disease causing
p.Pro95Leu	possibly damaging	tolerated	disease causing
p.Pro95Arg	probably damaging	tolerated	disease causing
p.Pro95Ala	probably damaging	tolerated	disease causing
p.Pro95Thr	probably damaging	tolerated	disease causing
p.Pro96_Arg103del;Pro107His	-	-	disease causing
p.Arg86_Gly93dup	-	-	disease causing
p.Arg94_Pro95insArg	-	-	polymorphism

Protein structure prediction

So far, a complete structure of SRSF2 is not available. Therefore, we applied *ab inito* prediction methods to obtain a full protein model for SRSF2. Based on this task, we choose the Robetta Server (http://robetta.bakerlab.org),[4] which is among the best *ab inito* prediction servers. The robustness and accuracy of Robetta was validated in various "Critical Assessment of Techniques for Protein Structure Prediction" (CASP). The main goal of CASP is an objective assessment of current methods for protein structure prediction. Within these tasks, *ab inito* servers like Robetta were validated by comparing their predicted models with known but unpublished protein structures. Based on the last CASP5 to CASP8 tasks, Robetta was among the top performing *ab inito* prediction servers.[4-8] Robetta is a full-chain protein structure prediction server. In a first step, Robetta separates the protein sequence into putative domains using the Ginzu protocol. Ginzu attempts to determine regions of protein sequences that will fold into globular units (domains). For each of these domains, Robetta predicts the folding using either homology modelling or *ab initio* modelling.

In a first iteration, we applied Robetta to predict models for the known RRM-domain (2KN4.pdb) of the SRSF2 wild-type protein. Based on the resulting models, the 3D-full model option was applied to obtain a complete model of SRSF2. Next, we evaluated the accuracy of Robetta. We compared the result of Robetta with the known structure of the RRM domain. Here, we used the common Cα-RMSD value to display the differences between the obtained crystal structure and the predicted model. The root-mean-square deviation (RMSD) reflects the average distance between atoms of two protein structures. In case of two globular protein conformations, one commonly measures the similarity in three-dimensional structure by the RMSD of the Cα atomic coordinates.[9] Robetta predicts five models. To select the best fitting model, we compared each model with the known structure of the RRM-domain (2KN4.pdb), and the best fitting one was selected as reference model for the following calculations. The differences of the five generated reference models to the known RRM-domain was determined by calculating the Cα-RMSD value between the two corresponding AA covering the RRM-domain (AA 14-92). The mean distance of all 79 analyzed AA for the best SRSF2 wild-type model is quite marginal (Cα-RMSD: 2.2 Å), showing that our calculated reference model reflects well the crystallized RRM-domain structure.

To gain insights into the extent by which *SRSF2* mutations might alter the protein folding and therefore the protein function, we submitted the altered protein sequences to Robetta and compared the resulting full models with the selected SRSF2 wild-type model. For each submitted sequence, we selected the best model based on the Cα-Cα distance for the RRM domain of the SRSF2 wild-type model. The differences between these selected models to the wild-type model were also determined by calculating Cα-RMSD value between the two corresponding AA of the SRSF2 wild-type model and mutant SRSF2 for AA 88 to 99.

This area covers the mutation hotspot Pro95 and represents the linker sequence (AA 92-117) and therefore reflects the proper folding of the two functional domains (RRM and SR) relative to each other. The analyzed novel mutations (p.[Pro96_Arg103del;Pro107His], p.Arg86_Gly93dup, and p.Arg94_Pro95insArg) all demonstrated distinct distances relative to SRSF2 wild-type model, summarized in the main text in Figure 2 and the table below. The 3 bp duplication showed the smallest divergence to the reference model with a distance range of 0.4 – 6.3 Å. The 24 bp deletion and the 24 bp insertion models show greater differences with distances ranging from 0.2 – 20.1 Å and 0.5 – 22.7 Å, respectively. Since only one AA is changed by the missense mutations of Pro95, the models for the missense mutations show only slight divergences, being very congruent with the SRSF2 wild-type model (see table below). These data show that all calculated models fit very well with the known crystal structure of the RRM-domain up to AA 92 and larger changes appear within the mutated linker sequences.

	Cα-Cα distances of reference to mutation model for AA 88 to 99 in Å											
	88	89	90	91	92	93	94	95	96	97	98	99
Calculated mutation model	RRM					linker						
p.Pro95His	0.3	0.4	0.6	1.4	0.5	0.8	0.4	0.5	0.7	0.4	0.8	0.2
p.Pro95Leu	0.4	0.1	0.6	1.0	0.5	0.3	0.6	0.8	1.0	0.4	0.2	0.2
p.Pro95Arg	0.2	0.4	0.7	1.4	1.2	0.7	0.1	0.1	0.5	0.5	1.0	0.2
p.Pro95Ala	0.3	0.4	0.5	1.3	1.1	0.8	0.4	0.9	1.1	0.3	1.2	0.3
p.Pro95Thr	0.3	0.5	0.6	1.2	1.1	0.9	0.5	0.6	0.8	0.4	1.0	0.5
p.Pro96_Arg103del;Pro107His	0.2	0.3	2.3	4.3	7.0	13.8	17.1	20.5	16.7	19.7	20.0	16.3
p.Arg86_Gly93dup	0.7	0.5	0.6	1.2	4.0	7.9	19.6	22.9	19.1	19.6	22.7	22.3
p.Arg94_Pro95insArg	0.4	0.6	2.4	3.6	5.9	6.3	6.3	3.8	3.8	6.6	7.8	3.5

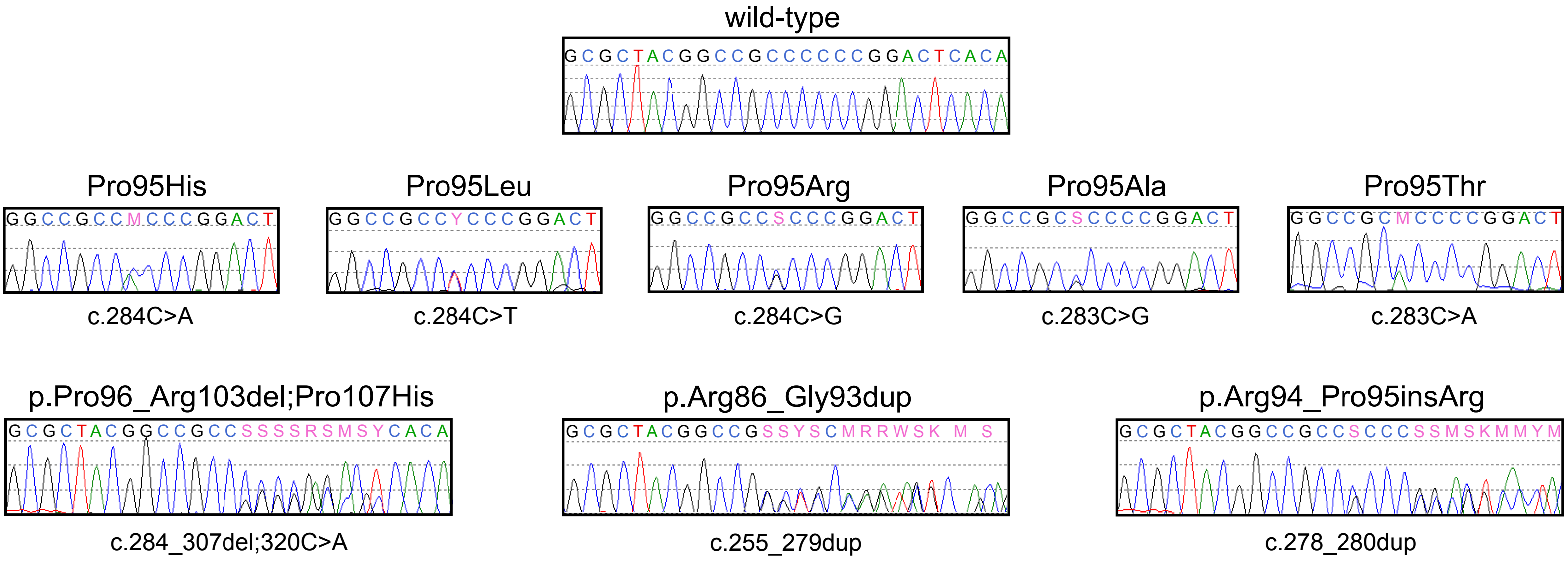
Reference list

1. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat.Protoc.* 2009;4(7):1073-1081.
2. Adzhubei IA, Schmidt S, Peshkin L et al. A method and server for predicting damaging missense mutations. *Nat.Methods* 2010;7(4):248-249.
3. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat.Methods* 2010;7(8):575-576.
4. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* 2004;32(Web Server issue):W526-W531.
5. Chivian D, Kim DE, Malmstrom L et al. Automated prediction of CASP-5 structures using the Robetta server. *Proteins* 2003;53 Suppl 6(524-533.
6. Chivian D, Kim DE, Malmstrom L et al. Prediction of CASP6 structures using automated Robetta protocols. *Proteins* 2005;61 Suppl 7(157-166.
7. Das R, Qian B, Raman S et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* 2007;69 Suppl 8(118-128.
8. Raman S, Vernon R, Thompson J et al. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 2009;77 Suppl 9(89-99.
9. Coutsias EA, Seok C, Dill KA. Using quaternions to calculate RMSD. *J Comput.Chem.* 2004;25(15):1849-1857.

Supplemental Figure S2

A

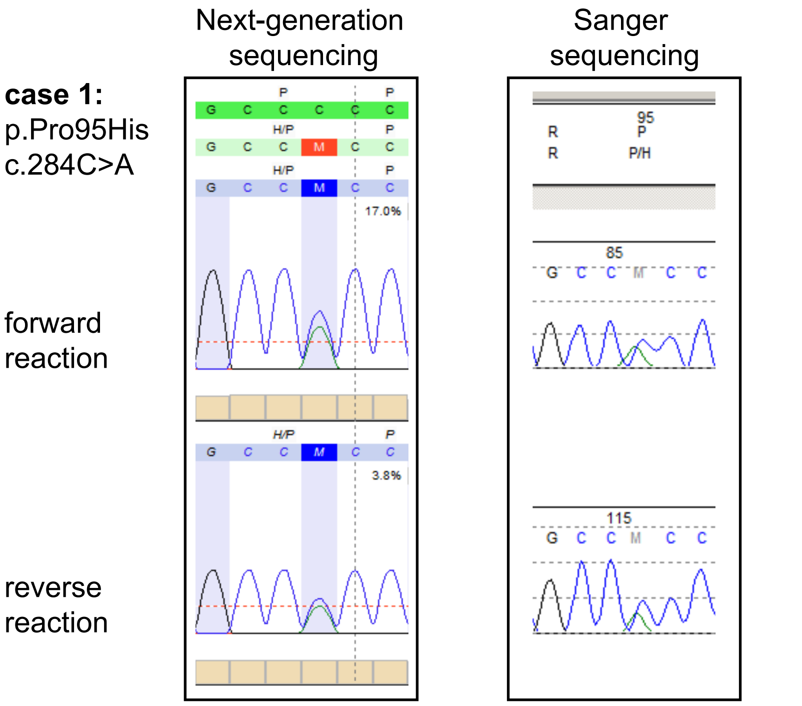
Overview of *SRSF2* sequences



Sequence representations around the codon for Pro95 of each mutation type are given as electropherograms of the Sanger sequencing reaction.

B

Correlation of mutational loads received by Next-generation and Sanger sequencing



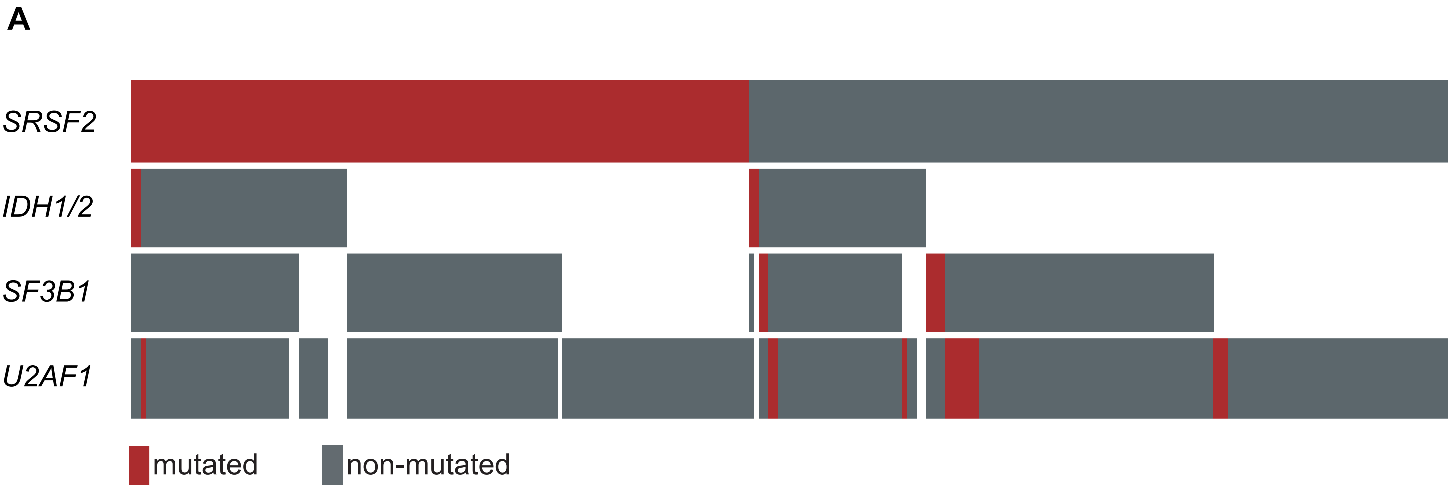
mutational load:
42,5%

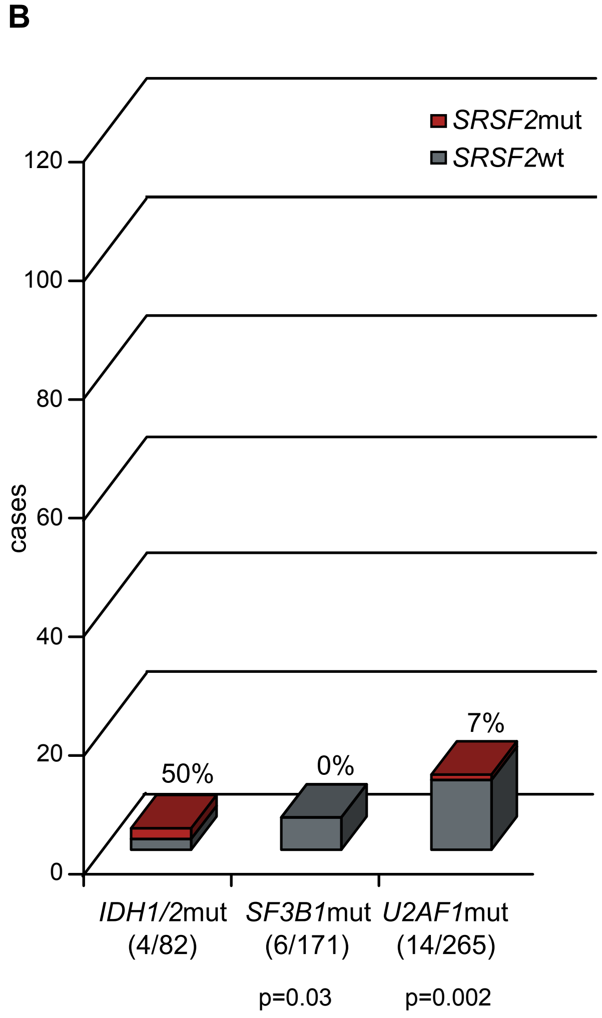
estimated mutational load:
40%

case	Sanger sequencing	NGS
1	40%	43%
2	37%	32%
3	35%	42%
4	25%	28%
5	40%	35%
6	45%	56%
7	50%	31%
8	40%	26%
9	35%	33%
10	50%	51%

The estimation of the mutational load was based on the electropherograms of the forward and reverse Sanger sequencing reactions. For comparison 10 cases were additionally analyzed by Next-generation sequencing. The electropherograms of both sequencing methods are shown for case 1, showing the comparability of both mutational loads received by Next-generation sequencing and estimation by Sanger sequencing. The table gives the results of all 10 cases.

Supplemental Figure S3



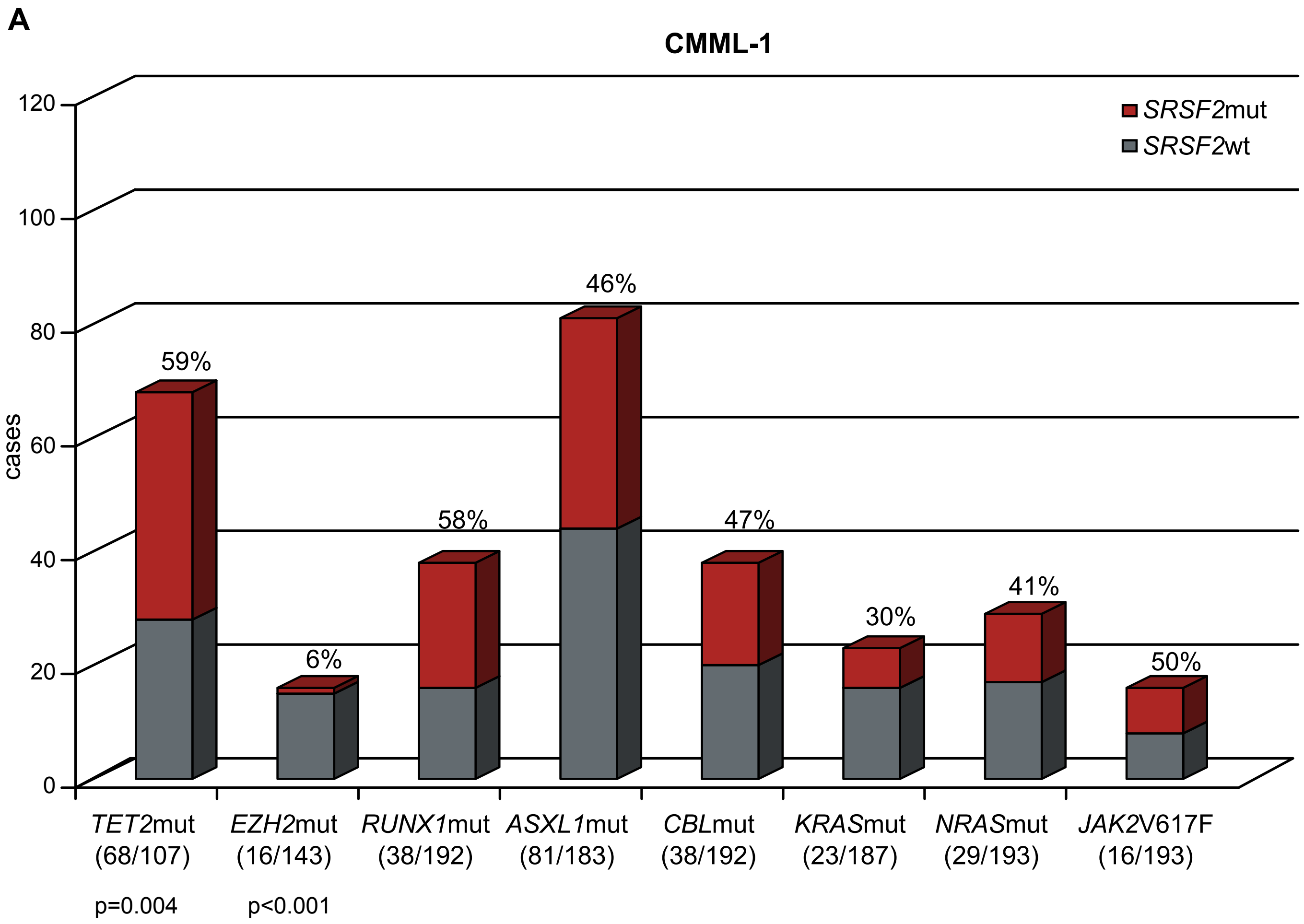


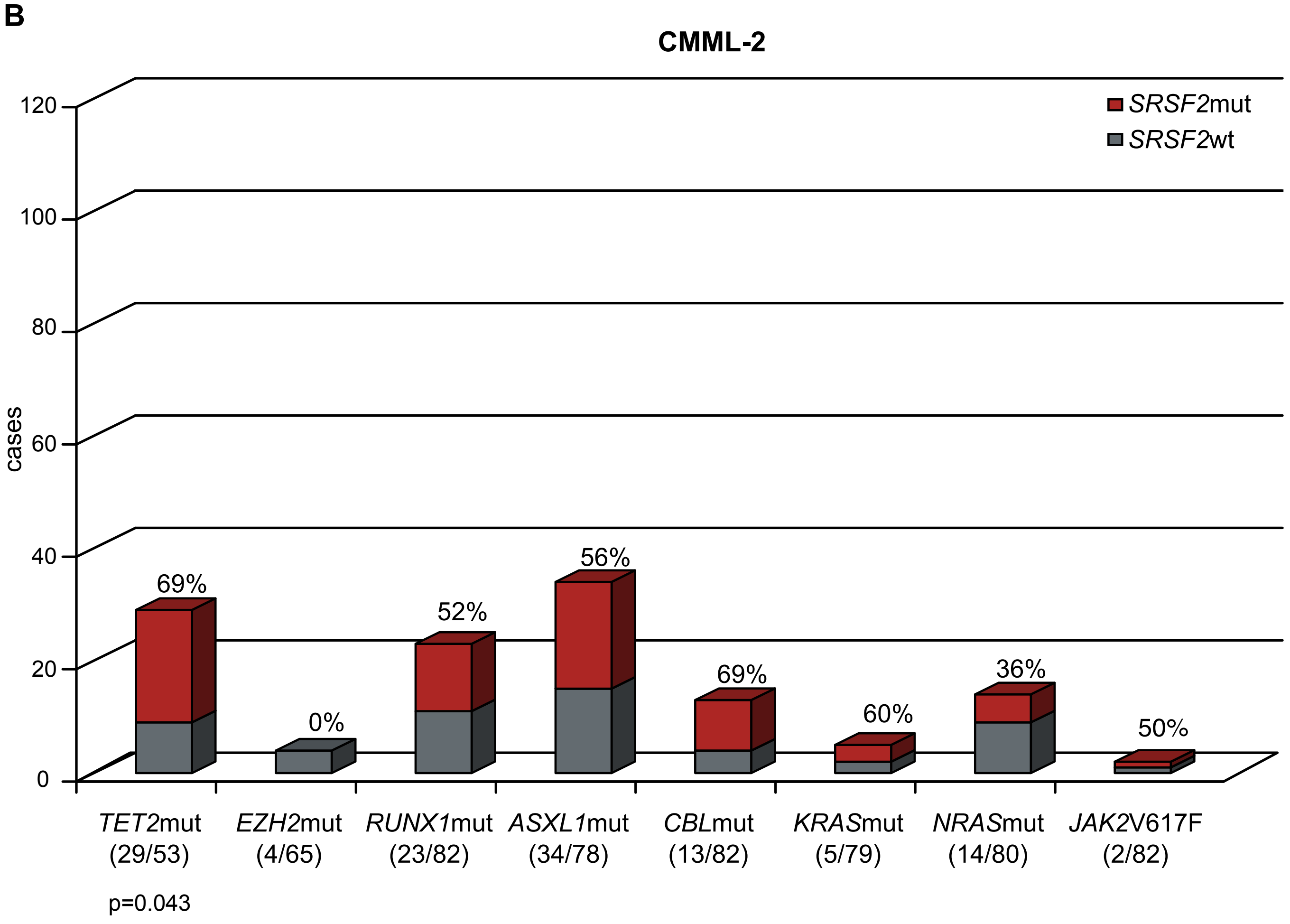
IDH1/2 is mutated in 4 of 82 cases (5%), *SF3B1* in 6/171 (4%), and *U2AF1* in 14/265 cases (5%). All 6 *SF3B1*mut cases are *SRSF2*wt, and only one case of the 14 *U2AF1*mut cases carries an additional *SRSF2* mutation, indicating that *SRSF2*mut is mutually exclusive of *SF3B1*mut and *U2AF1*mut.

(A) shows the alignment of the gene mutations for *SRSF2*, *IDH1/2*, *SF3B1*, and *U2AF1*. Each column represents one of the analyzed samples. Red bar: mutated gene; grey bar: non-mutated gene; white bar: no data available.

(B) Concomitant events of *SRSF2* with mutations in *IDH1/2*, *SF3B1*, and *U2AF1* are shown additionally as a bar chart. The grey part represents *SRSF2*wt, the red one *SRSF2*mut within the analyzed subcohorts. *SRSF2*mut frequencies and significances (p-values) are denoted; numbers of mutated/analyzed cases of the subcohorts are given in parenthesis below the bars.

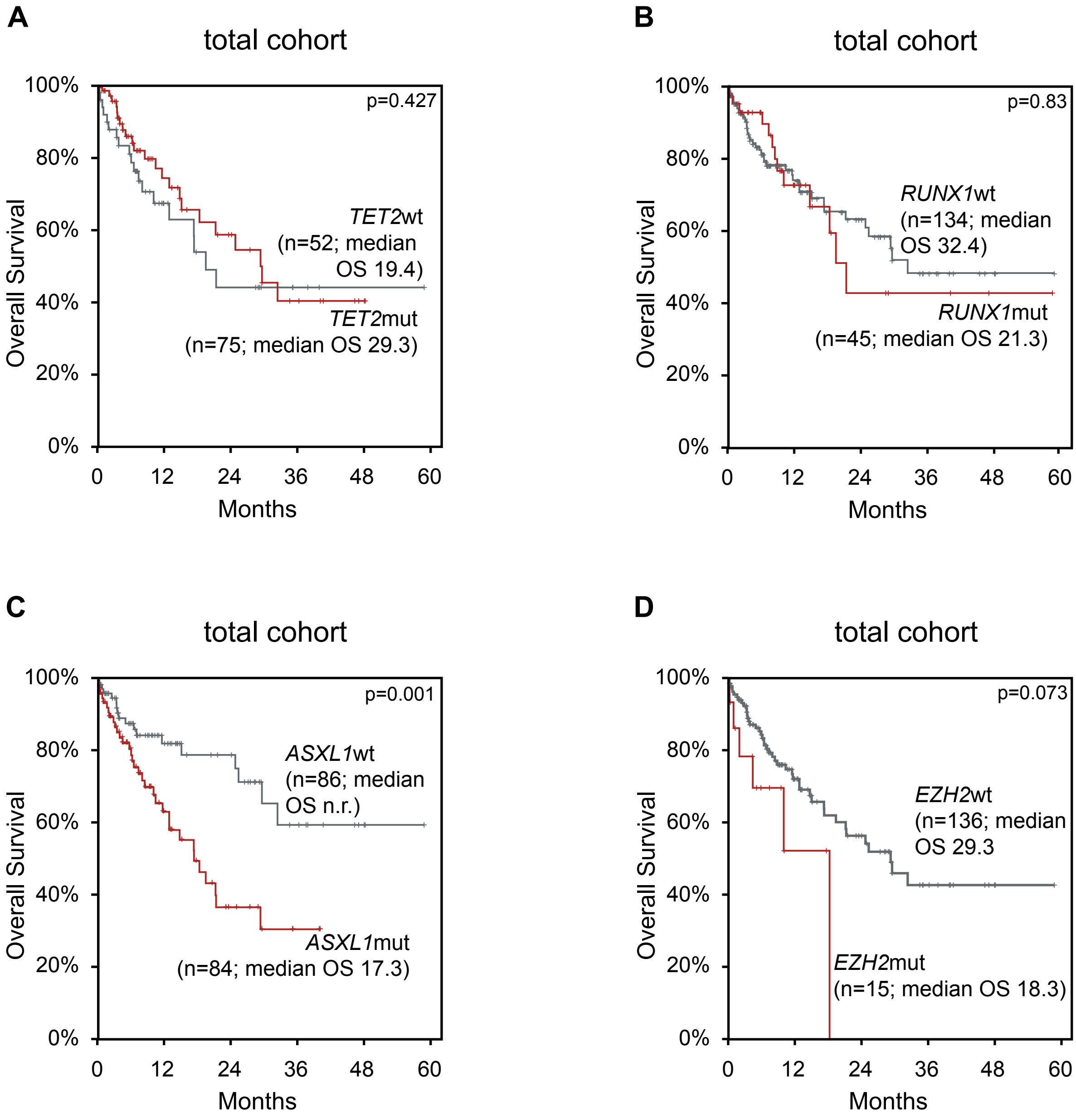
Supplemental Figure S4





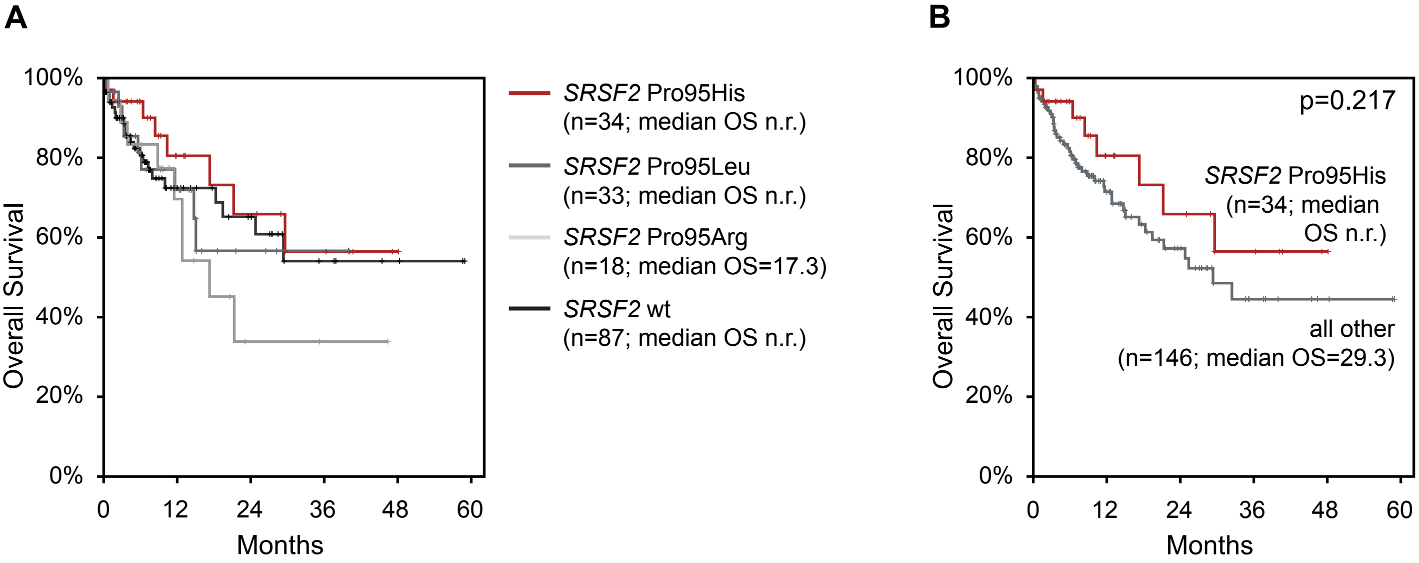
Concomitant events of *SRSF2* with other mutations are shown as bar charts separately for CMML-1 (**A**) and CMML-2 (**B**) cases. The grey part represents *SRSF2*wt, the red one *SRSF2*mut within the analyzed subcohorts. *SRSF2*mut frequencies and significances (p-values) are denoted; numbers of mutated/analyzed cases of the subcohorts are given in parenthesis below the bars.

Supplemental Figure S5



Overall survival (Kaplan-Meier) of CMML patients with mutations in *TET2*, *RUNX1*, *ASXL1,* and *EZH2*. The overall survival of patients with *TET2*mut (**A**) or *RUNX1*mut (**B**) did not differ compared to patients with the corresponding wild-type. In contrast, mutations in *ASXL1* (**C**) or *EZH2* (**D**) resulted in shorter overall survival. Overall survival is indicated in months and was compared using the two-sided log rank test. p-values are indicated in each graph, respectively.

Supplemental Figure S6



Overall survival (Kaplan-Meier) of patients with *SRSF2* missense mutations. (**A**) The overall survival curves of the separate *SRSF2* cohorts Pro95His, Pro95Leu, Pro95Arg, and wild-type indicated that patients with a *SRSF2* Pro95His mutation show a better overall survival compared to patients with either a Pro95Leu, Pro95Arg or no *SRSF2* (*SRSF2*wt) mutation. (**B**) Based on the finding shown in (A) all patients with other *SRSF2* mutations than Pro95His and patients with wild-type *SRSF2* were grouped (="all other" group). The overall survival of the *SRSF2* Pro95His group indicated that the Pro95His mutation has a favorable impact compared to the "all other" group. Overall survival is indicated in months and was compared using the two-sided log rank test. p-value is indicated.