Representative cancer-associated U2AF2 mutations alter RNA interactions and splicing

Received for publication, July 22, 2020, and in revised form, September 21, 2020 Published, Papers in Press, October 5, 2020, DOI 10.1074/jbc.RA120.015339

Debanjana Maji, Eli Ezra Glasser, Steven Henderson, Justin Galardi, Mary J. Pulvino, Jermaine L. Jenkins®, and Clara L. Kielkopf®

From the Center for RNA Biology, Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, New York, USA

Edited by Alex Toker

High-throughput sequencing of hematologic malignancies and other cancers has revealed recurrent mis-sense mutations of genes encoding pre-mRNA splicing factors. The essential splicing factor U2AF2 recognizes a polypyrimidine-tract splicing site signal and initiates spliceosome assembly. Here, we investigate representative, acquired U2AF2 mutations, namely N196K or G301D amino acid substitutions associated with leukemia or solid tumors, respectively. We determined crystal structures of the wild-type (WT) compared with N196K- or G301D-substituted U2AF2 proteins, each bound to a prototypical AdML poly(pyrimidine tract, at 1.5, 1.4, or 1.7 Å resolutions. The N196K residue appears to stabilize the open conformation of U2AF2 with an inter-RNA recognition motif hydrogen bond, in agreement with an increased apparent RNA-binding affinity of the N196K-substituted protein. The G301D residue remains in a similar position as the WT residue, where unfavorable proximity to the RNA phosphodiester could explain the decreased RNA-binding affinity of the G301D-substituted protein. We found that expression of the G301D-substituted U2AF2 protein reduces splicing of a minigene transcript carrying prototypical splice sites. We further show that expression of either N196K- or G301D-substituted U2AF2 can subtly alter splicing of representative endogenous transcripts, despite the presence of endogenous, WT U2AF2 such as would be present in cancer cells. Altogether, our results demonstrate that acquired U2AF2 mutations such as N196K and G301D are capable of dysregulating gene expression for neoplastic transformation.

Large-scale sequencing projects, together with an emerging plethora of protein structures, have revealed statistically significant clustering of disease-associated mutations at protein–ligand interfaces (1–4). This revelation explains how mis-sense mutations of the same gene can cause different diseases by affecting distinct functional interfaces of an encoded protein product. Conversely, trans-acting mutations that modify mutually interacting surfaces of a multisubunit complex often produce similar disease symptoms. For example, mis-sense mutations in different domains of the WAS (Wiskott–Aldrich syndrome) protein cause clinically distinct disorders such as Wiskott–Aldrich syndrome or X-linked neutropenia (1). On the other hand, mutations clustered at the mutual interfaces of either complement factor H and component C3 proteins result in hemolytic uremic syndrome (1). Indeed, understanding the 3D aspects of the gene-to-disease process is of great interest to pharmaceutical and medical industries seeking to identify druggable targets for precision medicine. For example, an Phe508 deletion remodels the interactome of the cystic fibrosis transmembrane conductance regulator (ΔF508). Reduced levels of specific ΔF508 cystic fibrosis transmembrane conductance regulator interactors can partially restore channel function (5). As a second example, oncogenic mutations of the p53 tumor suppressor protein can stimulate interactions with the transcription factor Nrf2, which in turn increases expression of proteasome genes and confers proteasome inhibitor resistance to cancer cells (6). The premise that disease-relevant mutations tend to cluster at 3D protein interfaces has been incorporated in several recent computational approaches for distinguishing neutral passenger mutations from candidate drivers of human disease (7–9).

Among hematologic malignancies and certain types of cancers, acquired mis-sense mutations frequently affect pre-mRNA splicing factors involved in the early stages of 3′ splice-site selection (10). Most often, mutational hot spots affect clusters of residues at the protein–RNA interfaces of SF3B1, SRSF2, and U2AF1 (11). The cancer-associated mutations of SF3B1 further are believed to modify the toroidal structure of the protein and alter its recruitment of RNA helicases (12, 13). The recurrent mutations of SF3B1, SRSF2, and U2AF1 in turn dysregulate the functions of the encoded proteins for gene expression and are thought to be drivers of cancer progression. Lower-frequency mutations of other splicing factors also may represent potential drug targets that alter proto-oncogenic functional events. Precedents for clinical consequences from such “long-tail” mutations in the frequency distributions of somatically mutated genes already have been established outside the field of pre-mRNA splicing, as exemplified for the paralogous mutations of Ras superfamily members (14).

We previously documented cancer-associated mutations for U2AF2, the heterodimeric partner of the U2AF1 pre-mRNA splicing factor (15). In most cases, U2AF2 mutations affect residues that are located in discrete domains of the protein, most prominently the two central RNA recognition motifs (RRM1 and RRM2), as well as an N-terminal region for heterodimerization with U2AF1 and a C-terminal protein-interaction motif (Fig. 1). The U2AF2 RRM1 and RRM2 are responsible for
recognizing a polypyrimidine (Py) tract signal preceding the major class of 3′ splice sites (16, 17). Structure determinations by NMR and X-ray crystallography demonstrate that the two RRM2s recognize a continuous nine-uridine Py tract in an open, side-by-side configuration (18–21). In the absence of RNA, the inter-RRM conformation is dynamic and can adopt a range of RRM1/RRM2 proximities (20, 22). A closed U2AF2 conformation, in which the RNA-binding surface of RRM1 is masked by RRM2 (18), is stabilized in the heterodimer with the U2AF1 subunit (23). Notably, many of the cancer-associated U2AF2 mutations are predicted to cluster near the RNA or RRM1/RRM2 interface of the respective open and closed conformations (15). This observation suggests that cancer-associated mutations of the U2AF2 RRM2 could modulate binding to the Py tract and dysregulate pre-mRNA splicing or other U2AF2-RNA-dependent processes. However, the structural and functional consequences of such long-tail U2AF2 mutations have yet to be investigated empirically.

Here, we use X-ray crystallography, RNA-binding assays, and pre-mRNA splicing assays of minigene reporters and endogenous transcripts to investigate the consequences of two representative cancer-associated mutations of U2AF2. We focused on an N196K substitution of the N-terminal RRM1 that recurs among patients with acute myeloid leukemia (AML) and a distinct, G301D substitution of the C-terminal U2AF2 RRM2 observed in cases of colon adenocarcinoma and castration-resistant prostate carcinoma. We compared the respective 1.4 and 1.7 Å resolution structures of the N196K- and G301D-substituted proteins bound to a prototypical Py tract with a baseline, 1.5 Å resolution structure of the wild-type (WT) complex. The U2AF2 loop containing the N196K mutation shifts to form an inter-RRM hydrogen bond that appears to stabilize the open conformation. The G301D conformation remains unchanged but may disfavor binding to the nearby phosphodiester backbone of the RNA. Accordingly, the N196K mutation increases the apparent RNA-binding affinity of U2AF2, whereas the G301D mutation has a converse effect. The mutations differently affect splicing of representative transcripts, with the G301D mutation showing effects similar to siRNA-mediated reductions in U2AF2 levels. These different structural and functional consequences affirm that the N196K and G301D mutations of U2AF2 have the potential to drive dysregulated gene expression in leukemias and cancers and, in addition, resolve how mutations in the same U2AF2 gene can result in clinically distinct disorders.

Results
Structure of human U2AF2<sup>12L</sup> bound to the AdML Py tract
As a baseline to distinguish the structural influence of the N196K and G301D mutations on a bona fide splice-site complex, we first determined the crystal structure of the U2AF2 RRM-containing region (U2AF2<sup>12L</sup>; Fig. 1B) bound to a modified Py tract oligonucleotide (5′-UUUU(dU)(5BrdU)CC-3′) (Table 1 and Fig. 2A). The sequence of the co-crystallized oligonucleotide matched the prototypical adenovirus major late promoter transcript (AdML), which also is identical to the preferred U2AF2 binding site as determined by in vitro selection (16). The two terminal cytidines differed from the primarily uridine sequences of our prior U2AF2<sup>12L</sup> structures (20). As for prior structures, including a deoxyuridine (dU) and 5-bromo-2′-deoxyuridine in the oligonucleotide marked the sequence register and facilitated high-quality crystals. The U2AF2<sup>12L</sup>–AdML Py tract structure was determined at 1.5 Å resolution by molecular replacement. Based on similar crystallization conditions (succinate, pH 7.0), crystal packing, and resolution limits, the prior structure of U2AF2<sup>12L</sup> bound to an all-uridine oligonucleotide (PDB code 5EV3) was an appropriate starting model for refinement and comparison of structural changes to bind the terminal cytidines. Other high-resolution U2AF2<sup>12L</sup>-structures bound to uridines (PDB code 5EV1 and 5EV2) were available; however, the nonphysiological low pH of the crystallization conditions (pH 4.0) altered relevant terminal nucleotide interactions.

The overall structure of the U2AF2<sup>12L</sup>–AdML complex is nearly identical to the uridine counterpart (RMSD 0.1 Å for 185 matching Ca atoms of PDB code 5EV3), apart from a large, ~3 Å shift in the positions of the terminal cytidines (Fig. 2B). These tandem cytidines are well-ordered in all unbiased, feature-enhanced electron density map (24) (Fig. 2C), where they are engaged by a combination of direct and water-mediated hydrogen bonds with U2AF2 Arg<sup>146</sup>, Arg<sup>148</sup>, and Asp<sup>231</sup> (Fig. 2E). The well-defined contacts with the Asp<sup>231</sup> side chain are consistent with the ability of a D231V-variant U2AF2 to alter specificity for the terminal nucleotides of the Py tract (25). Despite comparable resolutions and similar crystallization conditions, the terminal uridines of the prior structure are poorly defined in the electron density (Fig. 2D). A Gln<sup>147</sup> side chain that mediates hydrogen bonds with the uracil base edges (Fig. 2F) has been displaced by arginine side chains in the current cytidine-bound structure. Although a prior deoxyribose substitution may contribute structural differences at the terminal nucleotides, the base-specific interactions and absence of 2′ hydroxyl contacts supports the conclusion that U2AF2 preferentially secures the tandem cytidines of the AdML Py tract compared with the uridine counterpart.

**Figure 1. U2AF2 RNA recognition motifs recognize the polypyrimidine splice-site signal and can be affected by cancer-associated mutations.** A, schematic diagram of the U2AF2 complex with the 3′ splice site. The C-terminal domain (UHM) is thought to sequentially bind SF1 and then SF3B1 during spliceosome assembly with the splice sites. B, location of the N196K and G301D mutations on the U2AF2 domains. The domain boundaries of the construct used for RNA-binding and crystallization experiments (U2AF2<sup>12L</sup>) are shown at the bottom.
Cancer-associated mutations alter U2AF2 structure–function

Table 1
Crystallographic data collection and refinement statistics

| U2AF2<sup>32L</sup> variant + UUUU(dU)U(5BrdU)CC | Wildtype | N196K | G301D |
|---|---|---|---|
| PDB accession code | 6XLW | 6XLV | 6XLX |
| Data collection<sup>a</sup> | | | |
| Wavelength (Å) | 1.033 | 0.979 | 0.979 |
| Resolution range (Å) | 38.71–1.50 (1.52–1.50) | 38.35–1.40 (1.42–1.40) | 38.31–1.70 (1.73–1.70) |
| Space group | P 2<sub>1</sub> 2<sub>1</sub> 2<sub>1</sub> | P 2<sub>1</sub> 2<sub>1</sub> 2<sub>1</sub> | P 2<sub>1</sub> 2<sub>1</sub> 2<sub>1</sub> |
| Unit cell (Å) | 43.3, 62.7, 77.4 | 43.3, 63.1, 76.9 | 43.5, 61.9, 77.4 |
| Total no. reflections | 230,397 | 251,647 | 136,172 |
| Multiplicity | 6.9 (6.0) | 6.2 (3.8) | 5.8 (5.2) |
| Completeness (%) | 97.0 (85.0) | 96.0 (70.7) | 99.7 (97.1) |
| Mean | I/σ(I) | 22.3 (9.9) | 13.1 (1.3) | 12.3 (2.4) |
| R<sub>merge</sub> (%)<sup>b</sup> | 5.3 (9.8) | 5.5 (67.7) | 6.3 (57.0) |
| R<sub>reim</sub> (%)<sup>c</sup> | 3.2 (6.2) | 2.4 (38.1) | 2.8 (26.5) |
| CC<sub>1/2</sub> (%)<sup>d</sup> | 99.8 (99.1) | 99.8 (70.3) | 99.8 (89.5) |
| Refinement | | | |
| No. reflections (work/test) | 33,522/2,223 | 76,426/5,061 | 44,172/2,922 |
| R<sub>work</sub>/R<sub>free</sub> (%)<sup>e</sup> | 14.1/15.3 | 14.1/16.9 | 16.3/19.4 |
| No. of atoms | 1,560 | 1,566 | 1,524 |
| Macromolecules | 160 | 160 | 160 |
| DNA/RNA | 266 | 282 | 178 |
| Solvent | 0.008 | 0.010 | 0.004 |
| RMSD | 1.041 | 1.209 | 0.835 |
| Bonds (Å) | 99.5 | 99.5 | 98.5 |
| Angles (°) | 0.5 | 0.5 | 1.5 |
| Ramachandran (%) | 0.0 | 0.0 | 0.0 |
| Outliers | 1.12 | 0.98 | 0.94 |
| MolProbity score<sup>f</sup> | 22.0 | 27.8 | 37.4 |
| B-factor (Å<sup>2</sup>) | 20.6 | 27.3 | 37.5 |
| Protein | 20.9 | 28.8 | 35.7 |
| DNA/RNA | 30.7 | 34.6 | 37.9 |
| Solvent | | | |
| Solvent | | | |

<sup>a</sup> Statistics for the highest-resolution shell are shown in parentheses.

<sup>b</sup> R<sub>merge</sub> = Σ(|I<sub>i</sub> – nI<sub>i</sub>)/nI<sub>i</sub> I<sub>i</sub>, where I<sub>i</sub> is an intensity for the ith measurement of a reflection with indices hkl and is the weighted mean of all measurements of I<sub>i</sub>.

<sup>c</sup> R<sub>reim</sub> = Σ(|I<sub>i</sub> – nI<sub>i</sub>)/nI<sub>i</sub>, where n is the number of observations of the intensity I<sub>i</sub>.

<sup>d</sup> CC<sub>1/2</sub>: correlation coefficient between intensities of random-half data set (42).

<sup>e</sup> R<sub>work</sub> = Σ(|F<sub>calc</sub>| – |F<sub>obs</sub>|)/Σ|F<sub>calc</sub>| for the working set of reflections. R<sub>free</sub> is R<sub>work</sub> for ~7% of the reflections excluded from the refinement. All data were used in the refinement.

<sup>f</sup> Calculated using the program MolProbity (43).

Structure of N196K U2AF2<sup>32L</sup> bound to the AdML Py tract

To view the structural changes caused by a representative cancer-relevant substitution of the U2AF2 RRM1, we determined the crystal structure of an N196K-substituted U2AF2<sup>32L</sup> bound to the AdML Py tract at 1.7 Å resolution (Table 1). This N196K substitution is among the most common U2AF2 mutations, resulting from A → T transversions in four AML patients (COSMIC code COSU544). The precipitant (malonate, pH 7.0) and crystal packing environment is similar to the WT AdML counterpart described above, such that structural changes can be attributed to the amino acid substitution. The Asn<sup>196</sup> residue is located in a loop of the N-terminal RRM1 near the bound oligonucleotide and at the RRM1/RRM2 interface (Fig. 2A). Apart from local movement of this loop region, the overall structure of the N196K-substituted protein remains similar to the WT complex (RMSD 0.6 between 197 matching Cα atoms). In the WT complex, the Asn<sup>196</sup> side chain is poorly ordered and modeled as two alternative conformations, one of which mediates a hydrogen bond with the uracil-O2 atom (Fig. 3, A and C). By contrast, the mutant lysine side chain is well-defined in the electron density (Fig. 3, B and D). While remaining well-defined, the Ser<sup>294</sup> residue shows evidence of two alternative conformations in the N196K-substituted structure. Rather than directly interacting with the nucleotide, the position of the Lys<sup>196</sup>-containing loop has shifted to achieve a hydrogen bond with one conformation of the Ser<sup>294</sup> backbone carbonyl in the opposite RRM. This interaction could potentially stabilize the open U2AF2 conformation for association with uridine-rich Py tracts.

Structure of G301D U2AF2<sup>32L</sup> bound to the AdML Py tract

We next investigated the structural changes caused by a representative cancer-relevant substitution of the U2AF2 RRM2 by determining the crystal structure of a G301D-substituted U2AF2<sup>32L</sup> bound to the AdML Py tract at 1.4 Å resolution (Table 1). This G301D substitution results from a A → G transition identified in colon adenocarcinoma and castration-resistant prostate carcinoma patients (26, 27). A related G301S substitution also occurs in papillary renal cell carcinoma (International Cancer Genome Consortium code DO48476). The crystallization conditions and packing environment of the G301D structure remained similar to the N196K and unmodified counterparts. The Gly<sup>301</sup> residue is located preceding the third β-strand of RRM2 and adjacent the first uridine of the bound oligonucleotide (Fig. 2A). There, the Gly<sup>301</sup> residue appears to stabilize the open U2AF2 conformation for association with uridine-rich Py tracts.
between 188 matching Ca atoms). The mutant Asp\textsuperscript{301} side chain is anchored by hydrogen bonds to the Lys\textsuperscript{328} and Asn\textsuperscript{268} side chains of neighboring RRM2 loops, from which it displaces an ordered water molecule of the Gly\textsuperscript{301} structure (Fig. 4, B and D). The partial negative charge of the acidic Asp\textsuperscript{301} side chain is located within van der Waals packing distance of the electronegative terminal phosphate of the oligonucleotide (3.7 Å oxygen–oxygen). This close proximity is expected to disfavor U2AF\textsubscript{2}–RNA association.

N196K and G301D substitutions alter U2AF\textsubscript{2} affinity for the AdML Py tract

Based on the structures of N196K and G301D U2AF\textsubscript{2}, we predicted that these amino acid substitutions would influence the RNA-binding affinities of the mutated proteins. To test this prediction, we measured the fluorescence anisotropy changes during titration of the WT and mutant proteins into a consensus 3’ splice site labeled with 5’ fluorescein and fit the apparent binding affinities (Fig. 5). The N196K substitution increased the apparent binding affinity of U2AF\textsubscript{2} for this splice-site RNA by approximately 4-fold. Conversely, the G301D substitution decreased the RNA-binding affinity of U2AF\textsubscript{2} by nearly 12-fold. The magnitudes of the mutation-induced effects on U2AF\textsubscript{2}–RNA association (~1–1.5 kcal mol\(^{-1}\)) agree with the positive charge and inter-RRM contacts of the Lys\textsuperscript{196} residue, as well as with the Asp\textsuperscript{301} negative charge introduced near the phosphodiester backbone.
N196K and G301D substitutions alter splicing of a U2AF2-responsive minigene model

We hypothesized that the altered RNA binding caused by the N196K and G301D substitutions would in turn affect splice-site selection. We first tested this hypothesis by RT-PCR and quantitative real-time RT-PCR of a well-characterized pyPY minigene (Fig. 6 and Fig. S1A), which comprises IgM M1 and partial M2 exons fused to an intron and exon from AdML (28). These alternative 3' splice sites are marked by uridine-poor (py) or uridine-rich (PY) Py tracts. Because the PY sequence of the minigene corresponds to the AdML prototype used for our U2AF212L co-crystal structures, this reporter is well-suited for evaluating U2AF2 structure–function relationships. In our cell line (HEK 293T) and culture conditions ("Experimental procedures"), the pyPY transcript was primarily unspliced (Fig. 6B). A small amount of splicing was detected at the cryptic splice site resulting in an 330-bp band representative of the "AG" consensus closest to the 5' splice-site donor. This site lacks a detectable Py tract and remains unchanged by U2AF2 expression, unlike the U2AF2-sensitive py and PY 3' splice sites. STD, molecular size standards.
the strong, consensus PY site. A minor product of a cryptic splice site corresponds in size to a proximal AG, which lacks a distinguishable Py tract and is not expected to respond to U2AF2 levels. Transfection of a plasmid expressing WT U2AF2 increased splicing, particularly at the weak py site as noted previously (20). Expression of the N196K-substituted U2AF2 also increased use of the py splice site, consistent with a gain in RNA-binding affinity. Conversely, expression of the G301D-substituted U2AF2 reduced py splicing nearly to background levels. These results agree with the differences in structure and RNA-binding affinities of the two mutant proteins, although downstream consequences for other interactions (e.g. regulation of U2AF1) may play roles in the altered splicing.

N196K and G301D substitutions of U2AF2 alter splicing of representative endogenous pre-mRNAs

We next tested the influence of expressing N196K- and G301D-substituted U2AF2 on splicing of representative endogenous pre-mRNAs in a human cell line (HEK 293T) (Fig. 7). We focused on three transcripts known to exhibit U2AF2-responsive exon-skipping (4): GSK3B, THYN1, and SAT1. To mimic the heterozygous context of acquired mutations in cancers, we overexpressed either WT, N196K, or G301D variants in the presence of endogenous U2AF2. We compared the effects of siRNA-mediated reductions in U2AF2 levels (Fig. S2). As expected (4), loss of U2AF2 increased inclusion of the GSK3B cassette exon and decreased inclusion of THYN1 and SAT1 cassette exons. The N196K and G301D substitutions lead to subtle but reproducible changes in splicing (Fig. 7). Overexpression of the G301D-mutant U2AF2 had a lesser but similar effect as U2AF2 knockdown, supporting that this mutant can stall the splicing process. The N196K variant slightly enhanced or had similar effects as WT U2AF2 on splicing of the GSK3B, THYN1, and SAT1 sites, consistent with its RNA-binding properties. Together with the alterations in pyPY splicing, these differences demonstrate that the N196K- and G301D-mutant U2AF2 can influence splicing of endogenous gene transcripts even in the presence of the normal U2AF2 counterpart.

Discussion

A number of the cancer-associated mutations of U2AF2 affect residues at the RNA interface of the U2AF2 RRMs (15). In the present study, we demonstrate structural and functional consequences for two different representatives of this mutational class, including a leukemia-associated N196K mutation and solid tumor-associated G301D mutation. The positively charged N196K mutation of RRM1 favors U2AF2 binding to the negatively charged RNA. The mutant lysine also promotes a local conformational change and mediates a hydrogen bond bridge to the neighboring RRM2, which appears to promote the open U2AF2 conformation for RNA binding. Accordingly, the N196K mutation increases the RNA-binding affinity of U2AF2. The G301D mutation, on the other hand, introduces a negative charge abutting the 5´ terminal phosphate of the bound oligonucleotide in the crystal structure. Although the G301D protein and RNA conformations remain similar to the WT structure, the aspartate side chain is expected to repulse the neighboring phosphate. In support of this conclusion, the G301D mutation penalizes the RNA-binding affinity of U2AF2.

For a prototypical pyPY minigene, expressing the N196K variant invokes a similar effect as WT U2AF2, most likely by...
binding to and activating the weak py tract. Conversely, the G301D-substituted U2AF2 is unable to activate splicing of the pyPY minigene, in agreement with its reduced RNA-binding affinity. Because U2AF2 in turn regulates expression of its heterodimeric U2AF1 subunit and likely other splicing factors (4, 28), it is possible that the U2AF2-associated effects on pyPY splicing are indirect. However, the clear response of pyPY splicing to U2AF2 levels and substitutions is consistent with a primarily direct effect under the conditions of this experiment.

The splicing of endogenous transcripts in human cells is even more complex because of competing factors and coupled processes such as transcription and polyadenylation. Moreover, we expressed the cancer-associated U2AF2 variants in the presence of normal U2AF2 to mimic the expected heterozygous state of cancer cells that have acquired the U2AF2 mutations. Nevertheless, expression of the N196K or G301D U2AF2 variants subtly but detectably alters splicing of endogenous transcripts. Although small, such changes could destabilize gene expression sufficiently to promote a cancerous state. Moreover, differences in the RNA-binding properties of the mutant proteins appear to affect splicing in different ways that could contribute to separate disease outcomes (e.g. AML for N196K versus colorectal/prostate carcinomas for G301D variants).

An ongoing challenge in the field of personalized medicine is to distinguish cancer “driver” mutations from the millions of neutral variants that have been documented in nearly every human gene (29). For example, inherited single-nucleotide variants can in some cases predispose carriers to cancers yet more often are simply neutral passengers. Although we did not detect significant growth differences for HEK 293T cells expressing the mutant U2AF2 proteins in the time frame of our experiments (data not shown), the WT U2AF2 allele is essential for cell viability, and accordingly, its acquired mutations typically are heterozygous. Our finding that the N196K or G301D variants of U2AF2 change splicing of representative gene transcripts, coupled with the apparent absence of N196K- or G301D-encoding mutations among inherited U2AF2 single-nucleotide variants (30), suggests that critical functional consequences could prevent passage of these mutations through the germline. Taken together, these findings support the notion that N196K or G301D mutations of U2AF2 are capable of contributing to the oncogenic dysregulation of gene expression.

Conversely, mutations that affect mutual interfaces of distinct subunits can have analogous functional consequences and cause the same disease, i.e. “guilt by association” (1, 31, 32). Because U2AF2 contacts the majority of 3’ splice sites (4), a mutant Py tract signal would have little impact on the transcriptome compared with a mis-sense mutation of the U2AF2 protein itself. However, acquired mutations affecting protein partners of U2AF2 (including U2AF1, SF3B1, SF1, or RNA unwindases) could trigger similar downstream effects. Accordingly, our crystal structure suggests that the AML-associated N196K substitution stabilizes the open U2AF2 conformation, as observed for the AML/myelodysplasia–associated S34F mutation of U2AF1 in complex with a subset of splice sites (23).

As a second example, the cancer-associated K700E mutation of SF3B1 is expected to weaken SF3B1–RNA contacts and modulate RNA unwindases (11–13), which could mimic the G301D-dependent destabilization of the U2AF2–Py tract complex. Third, a search using cBioPortal (33, 34) indicates that SF1 alterations recur among castration-resistant prostate cancers and colorectal adenocarcinomas, which are the same cancer types associated with the U2AF2 G301D mutation. Indeed, the most common mis-sense mutations of SF1 (R255Q/W and R135C/H) are located at its RNA interface, which is expected to reduce RNA binding to the ternary SF1–U2AF2–U2AF1 complex, as would the G301D substitution of U2AF2. These observations suggest that in certain contexts, the long-tail N196K and G301D U2AF2 mutations may evoke similar consequences as more common mutations in other splicing factors, thereby dysregulating pre-mRNA splicing and contributing to neoplastic transformation.

In conclusion, the results presented here offer a concrete molecular mechanism for N196K and G301D mis-sense mutations of U2AF2 to contribute to the progression of malignancies by altering splice-site signal recognition. By analogy, other known 3D interfaces of the U2AF2 protein can explain the enrichment of cancer-associated mutations in its splicing factor partners. Taken collectively, these examples of long-tail splicing factor mutations represent a source of pre-mRNA splicing aberrations among cancers that may be more widespread than apparent based on inspection of the relatively rare, individual occurrences. Beyond the N196K/G301D-containing cluster of mutations at the U2AF2–RNA interface, other subsets of cancer-associated U2AF2 mutations are located at the inter-RRM interface of its closed conformation or in the C-terminal SF1–SF3B1-interaction motif. More studies are needed to distinguish the relevance of other long-tail mutations for U2AF2 function, and associated cancers. Meanwhile, our confirmation that the cancer-associated N196K and G301D mutations modify the structural and functional properties of U2AF2 raises the possibility of targeting U2AF2 and its partners as a potential means to investigate and treat leukemias and cancers.

### Experimental procedures

#### Expression and purification

For crystallization and RNA-binding experiments, the WT, N196K, or G301D variants of the U2AF2 RNA-binding domain included the N/C-terminal extensions of RRMI/RRM, RRMI, RRM2, and the inter-RRM linker (residues 141–342 of NCBI RefSeq NP_009210). These U2AF212L proteins were expressed and purified as described (20). Following a final step of size-exclusion chromatography on a Superdex-75 prep-grade column (Cytiva Inc.) equilibrated with 100 mM NaCl, 15 mM HEPES, pH 6.8, 0.2 mM TCEP, the purified U2AF212L proteins were concentrated using a Vivaspin 15 (Sartorius Corp.) centrifugal concentrator with a 10-kDa molecular mass cutoff. The protein concentration was estimated using the calculated extinction coefficient of 8,940 M⁻¹ cm⁻¹ and absorbance at 280 nm. Purified, deprotected oligonucleotides for co-crystallization were purchased from Integrated DNA Technologies Inc. Purified, fluorescein-labeled RNA oligonucleotides were purchased from Horizon Discovery Ltd. and deprotected according to the manufacturer’s instructions.
**Cancer-associated mutations alter U2AF2 structure–function**

_Crystallization and structure determination_

Prior to crystallization, WT, N196K, or G301D U2AF2^{12L} variants were mixed in a 1:2 molar ratio with purified oligonucleotide (5′-UUUU(dU)(U5BrdU)CC-3′) and incubated on ice for 20 min. The final protein concentration was ~20 mg ml\(^{-1}\). Diffraction quality crystals were obtained within approximately 1 week from a hanging drop of 1 μl of macromolecule layered with 1 μl of precipitant equilibrated over a 0.7-ml reservoir at 4 °C. The precipitant of the WT complex was 1 M succinic acid, 0.1 M HEPES, pH 7.0, 3% (w/v) PEG mono methyl ether 2000. The precipitant of the N196K or G301D variants was 0.24 M sodium malonate, pH 7.0, 5% sucrose, and 20–25% w/v PEG 3350. Additionally, 0.1 μl of 5% w/v LDAO detergent (Hampton Research) was added to the N196K or G301D protein-oligonucleotide mixtures immediately prior to the precipitant solution. Each crystal was coated with a mixture of 1:1 (v/v) paratone-N and silicone oil and then flash-cooled in liquid nitrogen before data collection at 100 K. Crystallographic data sets were collected remotely at the Stanford Synchrotron Radiation Light Source Beamline 12-2 (35). The data were processed using the Stanford Synchrotron Radiation Light Source AUTOXDS script (A. Gonzalez and Y. Tsai) implementation of XDS (36) and CCP4 packages (37). The structures were determined using the Fourier synthesis method starting from PDB code 5EV3. The models were adjusted using COOT (38) and refined using PHENIX (39). The crystallographic data and refinement statistics are given in Table 1.

**Fluorescence anisotropy RNA-binding assays**

Protocols for the RNA-binding experiments were essentially as described (40). A 5′-fluorescein-labeled, 32-mer RNA oligonucleotide contained a near-consensus 3′ splice site (5′-CCU-GUCCCUUUUUUUUUUUAGGUCUGGGCA, with the AG consensus underlined). The purified proteins and RNA were diluted separately by >100-fold into a binding buffer comprising 100 mM NaCl, 15 mM HEPES at pH 6.8, 0.2 mM TCEP, 0.1 unit ml\(^{-1}\) Superase-In\textsuperscript{TM} (Invitrogen). The final RNA concentration in the cuvette was 30 nM. The volume changes during addition of the protein were <10% to minimize dilution effects. The fluorescence anisotropy changes during titration were measured using a FluoroMax-3 spectrophotometer and temperature-controlled at 23 °C by a circulating water bath. The samples were excited at 490 nm, and the emission intensities were recorded at 520 nm with a slit width of 5 nm. The fluorescence emission spectra also were monitored for similarity throughout the experiment. Each titration was fit as described (40) to obtain the apparent equilibrium dissociation constant (\(K_D\)). These fits and the \(p\) values of a two-tailed unpaired \(t\) test with Welch’s correction were calculated using Prism version 6.0 (GraphPad Software Inc.). The apparent equilibrium affinities (\(K_A\)) are the reciprocals of each \(K_D\). The average \(K_D\) or \(K_A\) values and standard deviation among three replicates sequences are listed in Table S1.

**Data availability**

The atomic coordinates and structure factors of the WT, N196K, and D301G variants of U2AF\(^{12L}\) bound to AdML oligonucleotide (accession codes 6XLV, 6XLW, and 6XLX) have been deposited at the Protein Data Bank.

**Acknowledgments**—We are grateful to Maria Carmo-Fonseca (University of Lisbon, Portugal) for providing the pyPPY and wild-type HA-U2AF2 plasmids, Hartmut Land (University of Rochester) for pBABE-puro, Paul Bourtz (University of Rochester) for advice with PCR optimization, and Irimpan Mathews (Stanford Synchrotron Radiation Light Source) for assistance with remote data collection.

**Author contributions**—D. M., E. G., S. H., J. G., M. J. P., J. L. J., and C. L. K. investigation; D. M. and C. L. K. writing–original draft; D. M., M. J. P., J. L. J., and C. L. K. writing–review and editing; M. J. P. and C. L. K. conceptualization; M. J. P., J. L. J., and C. L. K. supervision; M. J. P. and J. L. J. validation; M. J. P., J. L. J., and C. L. K. methodology; C. L. K. resources; C. L. K. funding acquisition.

**Funding and additional information**—This study was supported by National Institutes of Health Grant R01 GM070503 (to C. L. K.). The crystallographic data were collected at the Stanford Synchrotron Radiation Light Source, which is supported by U.S. Department of Energy Contract DE-AC02-76SF00515 and National Institutes of Health Grant P41 GM103393. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Conflict of interest**—The authors declare that they have no conflicts of interest with the contents of this article.

**Abbreviations**—The abbreviations used are: AdML, adenovirus major late promoter transcript; AML, acute myeloid leukemia; BrdU, 5-bromo-2′-deoxyuridine; dU, deoxyuridine; HEK, human embryonic kidney; PDB, Protein Data Bank; Py, polypyrimidine; RMSD, root mean square deviation; RRMA, RNA recognition motif; TCEP, tris(2-carboxyethyl)phosphine; GAPDH, glyceraldehyde-3-phosphate dehydrogenase.

**References**

1. Wang, X., Wei, X., Thijsen, B., Das, J., Lipkin, S. M., and Yu, H. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat. Biotechnol. 30, 159–164 CrossRef Medline
2. Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., Lander, E. S., and Getz, G. (2015) Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc. Natl. Acad. Sci. U.S.A. 112, E5486–E5495 CrossRef Medline
3. Gao, M., Zhou, H., and Skolnick, J. (2015) Insights into disease-associated mutations in the human proteome through protein structural analysis. Structure 23, 1362–1369 CrossRef Medline
4. Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., Zhou, J., Qiu, J., Jiang, L., Li, H., Chen, G., Sun, H., Zhang, Y., Denise, A., Zhang, D. E., et al. (2014) Mechanisms for U2AF to define 3′ splice sites and regulate alternative splicing in the human genome. Nat. Struct. Mol. Biol. 21, 997–1005 CrossRef Medline
5. Pankow, S., Bamberger, C., Calzolari, D., Martinez-Bartolomé, S., Lavallée-Adam, M., Balch, W. E., and Yates, J. R., 3rd (2015) ΔF508 CFTR
interactome remodelling promotes rescue of cystic fibrosis. *Nature* **528**, 510–516 CrossRef Medline

6. Walerych, D., Lisek, K., Sommaggio, R., Piazza, S., Clani, Y., Dalla, E., Rajkowska, K., Gaweda-Walerych, K., Ingallina, E., Tonelli, C., Morelli, M. J., Amato, A., Eterno, V., Zambelli, A., Rosato, A., et al. (2016) Proteasome system functionality is instrumental in a common gain-of-function program of the p53 missense mutants in cancer. *Nat. Cell Biol.* **18**, 897–909 CrossRef Medline

7. Shim, J. E., Kim, J. H., Shin, I., Lee, J. E., and Lee, I. (2019) Pathway-specific protein domains are predictive for human diseases. *PLoS Comput. Biol.* **15**, e1007052 CrossRef Medline

8. Ashford, P., Pang, C. S. M., Moya-García, A. A., Adeyelu, T., and Orenko, C. A. (2019) A CATH domain functional family based approach to identify putative cancer driver genes and driver mutations. *Sci. Rep.* **9**, 263 CrossRef Medline

9. Chen, S., Fragoza, R., Klei, L., Liu, Y., Wang, J., Roeder, K., Devlin, B., and Yu, H. (2018) An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. *Nat. Genet.* **50**, 1032–1040 CrossRef Medline

10. Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R. K. (2016) RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16**, 413–430 CrossRef Medline

11. Jenkins, J. L., and Kielkopf, C. L. (2017) Splicing factor mutations in myelodysplasias: Insights from spliceosome structures. *Trends Genet.* **33**, 336–348 CrossRef Medline

12. Tang, Q., Rodríguez-Santiago, S., Wang, J., Pu, J., Yuste, A., Gupta, V., Moldón, A., Xu, Y. Z., and Query, C. C. (2016) SF3B1/Hsh155 HEAT motif mutations affect interaction with the spliceosomal ATPase Prp5, resulting in altered branch site selectivity in pre-mRNA splicing. *Genes Dev.* **30**, 2710–2723 CrossRef Medline

13. Carrocci, T. I., Zoerner, D. M., Paulson, J. C., and Hoskins, A. A. (2017) SF3B1 mutations associated with myelodysplastic syndromes alter the fidelity of branchsite selection in yeast. *Nucleic Acids Res.* **45**, 4837–4852 CrossRef Medline

14. Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., Gao, J., Socci, N. D., Solit, D. B., Oshen, A. B., Schultz, N., and Taylor, B. S. (2016) Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 CrossRef Medline

15. Glasser, E., Agrawal, A. A., Jenkins, J. L., and Kielkopf, C. L. (2017) Cancer-associated mutations mapped on high-resolution structures of the U2Aδ2 RNA recognition motifs. *Biochemistry* **56**, 4757–4761 CrossRef Medline

16. Singh, R., Valcárcel, J., and Green, M. R. (1995) Distinct binding specificities of the main splicing factor U2AF65: a combined NMR and SAXS study. *J. Am. Chem. Soc.* **117**, 7068–7076 CrossRef Medline

17. Agrawal, A. A., Salsi, E., Chatriki, R., Henderson, S., Jenkins, J. L., Green, M. R., Ermolchenko, D. N., and Kielkopf, C. L. (2016) An extended U2AF65-RNA-binding domain recognizes the 3’ splice site signal. *Nat. Commun.* **7**, 10950 CrossRef Medline

18. Voith von Voithenberg, L., Sánchez-Rico, C., Kang, H. S., Madl, T., Zanier, K., Barth, A., Warner, L. R., Sattler, M., and Lamb, D. C. (2016) Recognition of the 3’ splice site RNA by the U2AF heterodimer involves a dynamic population shift. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E7169–E7175 CrossRef Medline

19. Huang, J. R., Warner, L. R., Sanchez, C., Abel, E., Hultman, T., Gollub, R. D., Meyerson, M., Spit, R. S., Lander, E. S., and Getz, G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 CrossRef Medline

20. Wang, L., Lawrence, M. S., Yan, Y., Stojanov, P., Sougnez, C., Sivachenko, A., DeLuca, D. S., Zhang, L., Zheng, W., Vartanov, A. R., Fernandes, S. M., Goldstein, N. R., Folco, E. G., et al. (2011) SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *Nat. Engl. J. Med.* **365**, 2497–2506 CrossRef Medline

21. Olivier, S. (2000) Giuliet-by-association goes global. *Nature* **403**, 601–603 CrossRef Medline

22. Jenkins, J. L., Laird, K. M., and Kielkopf, C. L. (2012) A broad range of conformational contributions to the solution ensemble of the essential splicing factor U2AF65. *Biochemistry* **51**, 5223–5225 CrossRef Medline

23. Afonine, P. V., Moriarty, N. W., Mustyakimov, M., Sobolev, O. V., Terwilliger, T. C., Turk, D., Urzhumtsev, A., and Adams, P. D. (2015) FEM: feature-enhanced map. *Acta Crystallogr. B* **71**, 646–666 CrossRef Medline

24. Agrawal, A. A., McLaughlin, K. J., Jenkins, J. L., and Kielkopf, C. L. (2014) Structure-guided U2AF65 variant improves recognition and splicing of a defective pre-mRNA. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17420–17425 CrossRef Medline

25. Agrawal, A. A., McLaughlin, K. J., Jenkins, J. L., and Kielkopf, C. L. (2014) Structure-guided U2AF65 variant improves recognition and splicing of a defective pre-mRNA. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17420–17425 CrossRef Medline

26. Cancer Genome Atlas, N (2012) Comprehensive molecular characterization of solid tumours reveals tissue- androgen receptor.
38. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501. [CrossRef Medline](https://doi.org/10.1107/S0907444909025529)

39. Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol Crystallogr.* **66**, 213–221. [CrossRef Medline](https://doi.org/10.1107/S0907444909054630)

40. Jenkins, J. L., Shen, H., Green, M. R., and Kielkopf, C. L. (2008) Solution conformation and thermodynamic characteristics of RNA binding by the splicing factor U2AF65. *J. Biol. Chem.* **283**, 33641–33649. [CrossRef Medline](https://doi.org/10.1074/jbc.M805573200)

41. Morgenstern, J. P., and Land, H. (1990) Advanced mammalian gene transfer: high titre retroviral vectors with multiple drug selection markers and a complementary helper-free packaging cell line. *Nucleic Acids Res.* **18**, 3587–3596. [CrossRef Medline](https://doi.org/10.1093/nar/18.18.3587)

42. Karplus, P. A., and Diederichs, K. (2012) Linking crystallographic model and data quality. *Science* **336**, 1030–1033. [CrossRef Medline](https://doi.org/10.1126/science.1219096)

43. Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12–21. [CrossRef Medline](https://doi.org/10.1107/S0907444909054630)