OMICS A Journal of Integrative Biology
Volume 15, Number 3, 2011

DOI: 10.1089/omi.2010.0107

Huge Splicing Frequency in Human Y Chromosomal *UTY* Gene

Ingeborg Laaser,[1] Fabian J. Theis,[2] Martin Hrabé de Angelis,[1,4] Hans-Jochem Kolb,[3] and Jerzy Adamski[1,4]

Abstract

Over 90% of human genes produce more than one mRNA by alternative splicing (AS). Human *UTY* (ubiquitously transcribed tetratricopeptide repeat protein on the chromosome Y) has six mRNA-transcripts. *UTY* is subject to interdisciplinary approaches such as Y chromosomal genetics or development of leukemia immunotherapy based on UTY-specific peptides. Investigating *UTY* expression in a normal and leukemic setting we discovered an exceptional splicing phenomenon fostering huge transcript diversity. Transcript sequencing identified 90 novel AS-events being almost randomly combined in 284 new transcripts. We uncovered a novel system of transcript architecture and genomic organization in *UTY*. On a basis of a new *UTY*-splicing multigraph including a mathematical model we calculated the theoretical yield to exceed 1.3 billion distinct transcripts. To our knowledge, this is the greatest estimated transcript diversity by AS. On protein level we demonstrated interaction of AS-derived proteins with new interactors by yeast-two-hybrid assay. For translational research we predicted new UTY-peptide candidates for leukemia therapy development. Our study provides new insights into the complexity of human alternative splicing and its potential contribution to the transcript diversity of the transcriptome.

Introduction

Alternative splicing (AS) is a fundamental process that increases transcriptome and proteome variety by producing different mRNA isoforms of a single gene. Recent studies on the contribution of AS to human transcript diversity have received high attention. Several global transcriptome analyses identified significant numbers of novel splice junctions (Sultan et al., 2008); furthermore, they revealed that over 90% of human genes produce more than one mRNA transcript (Castle et al., 2008; Pan et al., 2008; Wang et al., 2008). This demonstrates clearly that the extent of AS in human genes is not well established as yet. Further exploration that focuses on transcript diversity as produced by AS of a single gene is therefore necessary. The human Y-chromosomal gene *UTY* (ubiquitously transcribed tetratricopeptide repeat protein on the Y chromosome) is subject to interdisciplinary research. This covers major aspects such as Y-chromosomal molecular genetics or translational research on the development of immunotherapy of leukemia relapse, via UTY-specific peptides (Ivanov et al., 2005; Lahn & Page, 1997; Rozen et al., 2009; Skaletsky et al., 2003; Warren et al., 2000). Despite this, little is known still about the expression of *UTY*. One intriguing question is whether AS produces yet undetected transcripts in *UTY*. They may contribute new views to important effects of the aspects depicted above, which are still difficult to explain.

The world of RNA is complex and involves not only coding and noncoding categories (Dinger et al., 2008). Multiple mechanisms are instrumental in transcript diversification or regulation, like the coupling of premature termination codon (PTC) containing transcripts with nonsense-mediated decay (NMD). It is often involved in expression regulation (Chang et al., 2007; Lewis et al., 2003). The occurrence of mRNA transcript diversity, as produced by AS, is not fully understood as yet. Its multiple roles include involvement in generation of protein variety and regulation of gene expression. Moreover, the splicing extend can differ between genes (Wang et al., 2008). Whereas the basic mechanisms of alternative splice forms are known (Breitbart et al., 1987), a common splicing code in pre-mRNAs is just becoming deciphered (Barash et al., 2010). Thus also, on a more general level, no universal model for transcript diversity generating modes exists as yet. As these phenomena are just becoming explored,

[1]Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Experimental Genetics, Genome Analysis Center, and [2]Institute of Bioinformatics and Systems Biology, Neuherberg, Germany.
[3]Clinical Cooperative Group, Haematopoietic Cell Transplantation, José-Carreras Transplantation Unit, Ludwig Maximilians-Universität, Munich, Germany.
[4]Lehrstuhl für Experimentelle Genetik, Technische Universität München, Freising-Weihenstephan, Germany.

further study of *UTY* transcript diversity, its potential extend and transcript generating modes may contribute new perspectives to the understanding of the expression in *UTY*. This basic research may also contribute valuable new aspects for translational research on UTY-based immunotherapy targeting leukemia relapse.

Human *UTY* belongs to the single-copy X-degenerate genes on the Y chromosome. As the recombination between the homologous genes *UTY* and X-chromosomal *UTX* had become suppressed, the corresponding mRNA transcripts of *UTY* and *UTX* differ, mainly by point mutations, whereas the reading frames are conserved (Lahn and Page, 1997; Skaletsky et al., 2003). According to present records the human gene *UTY* codes for six mRNA transcripts. They represent six different polyadenylation variants, and comprise of three major reference sequences (RefSeqs; Fig. 1), and three short transcripts (Lahn and Page, 1997; Strausberg et al., 2002). Further, three internal AS-events are known (independent partial sequence and EST; GenBank CR936684, BX090888). This indicates that the present annotation of *UTY* might be incomplete.

Proteins of UTY and UTX exhibit conserved domain structure (see Fig. 1B for explanation of UTY). Their C-terminal jmjC domain is suggested to be essential for protein function (Hong et al., 2007). Although both proteins share interactors of their N-terminus (Grbavec et al., 1999), two recent studies using purified proteins identified yet only human UTX, and not UTY, as demethylating epigenetic mark H3K27 (Hong et al., 2007; Lan et al., 2007). This might point at unknown cofactors being necessary for the function of UTY. Therefore, further characterization of UTY by screening for new protein interactors is required, in particular, in the context of possibly enhanced UTY diversity.

UTY is ubiquitously expressed (Lahn and Page, 1997). In humans and the mouse, UTY-derived peptides elicit immunorecognition (Greenfield et al., 1996, 1998; Riddell et al., 2002). UTY has received attention in translational research on immunotherapy of leukemia, with regard to the very important prevention of tumor relapse in posttransplant patients (Kolb, 2008). In a male-recipient/female-donor setting UTY-specific peptides exhibit a gender specific antitumor effect, observed mainly *in vitro* (Ivanov et al., 2005; Riddell et al., 2002; Warren et al., 2000). These UTY-derived peptides act as minor histocompatibility antigens and elicit tissue restricted recognition by cytotoxic donor T cells. It is unclear, however, how tissue-restricted UTY-antigen recognition can function in the presence of ubiquitous *UTY* expression, as the encoding mRNA transcripts mainly represent distinct polyadenylation variants only. As the observed graft-versus-leukemia effect suggests medical application potential, further investigation into *UTY* expression and its potential transcript diversity is crucial.

Here we report the discovery of highly frequent *UTY* splicing in normal and leukemic cells. We provide the first evidence that far more *UTY* transcripts exist than previously

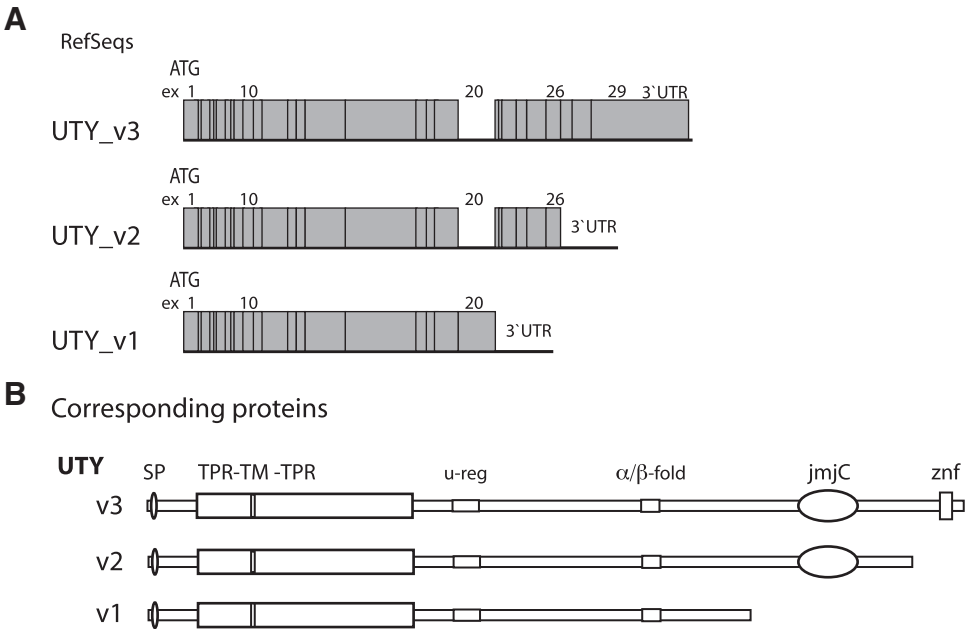

FIG. 1. Reference sequences of human *UTY*. (**A**) Organization of the open reading frames of the three polyadenylation variants, *UTY*_v1, v2, v3, illustrated on transcript level (for the transcript graph see also Methods). In *UTY*_v2 and v3 alternative splicing skips exon 20. The flanking 5′UTRs are not shown, they are part of exon 1. Concerning 3′UTRs, the final exons of v1 and v2 are flanked by characteristic 3′elongations in conjunction with poly(A) tails, whereas the 3′UTR of *UTY*_v3 is part of exon 29. Gray upright bars, exons; large dark bars, 3′ extensions; ex, exon; ex20, exon number 20. Data from GenBank NM_182660.1, NM_182659.1, NM_007125.3 and NM_021140. (**B**) Predicted UTY protein domain structure (Ginalski et al., 2004). Location of putative domains, motives, and structural elements. UTY/UTX are predicted to share two domains: the tetratricopeptide repeat domain region (TPR) with nin single TPRs and the jumonjiC (jmjC) domain, and an N-terminal signal peptide. For the UTY transmembrane domain and treble-clef-zinkfinger motive further structural elements such as an unstructured region and α/β-fold are predicted. SP, signal peptide; TM, transmembrane region; u-reg, unstructured region; znf, treble-clef-zinkfinger motif.

annotated. The huge diversity severely restricted the spectrum of methods. We therefore developed in this study an alternative strategy to single transcript approach, which allowed a new in-depth analysis on *UTY*. Our observations introduce a new dimension into the potential contribution of AS to the transcript diversity of the transcriptome. For human *UTY*, we provide new molecular basis and new views for research on UTY-based immunotherapy.

Materials and Methods

Accession numbers

Sequences of all variants have been deposited in GenBank (www.ncbi.nlm.nih.gov). Their accession numbers (acc.nos) are given in the Supplementary Table 2.

RNA preparation

Total RNA was isolated from primary cells by RNeasy Kit (DNase treated; Qiagen, Hilden, Germany). Subsequently, the RNA quality was controlled by an Agilent Bioanalyzer, with an RNA Nano LabChip and expert 2100 software. Both procedures were conducted according to the manufacturers' instructions.

Selective UTY *RT-PCR analyses*

RNA was reverse transcribed according to the manufacturer's instructions (Roche, Germany), using $Oligo_{dt\ 16}$ primers to target mRNAs. Selective PCR for each of the RefSeqs, *UTY*_v1, v2, and v3 was then established and subsequently conducted in all probes. PCR primer design was based on GenBank NM_182660.1, NM_182659.1, NM_007125.3 RefSeq sequences. Primers flanked the full-length open reading frame (ORF) and the adjoining regions of the UTRs.

PCR primers: *UTY*_v1, v2, v3 F GTCGCCCGGGTGTTTCCATGAAATCCTGC, *UTY*_v1 R ATCGGTCGACGTAGGAACTTTATTTCTCCATTAG, *UTY*_v2 R ATCGGTCGACTAAGGGAAGGCAATATTTTAATC, and *UTY*_v3 R ATCGGTCGACTATATCAAGATGAGGATGA. *UTY* primers contained XmaI and reverse SalI restriction sites for other projects outside the focus of this study. PCR was conducted using pfu/Taq mixture (Fermentas, Hanover, MD, USA), Robo Cycler 96 (Strategene, La Jolle, CA, USA), and PCR conditions: 94°C (2 min); 4 cycles 94°C (20 s), 49°C (1 min), 68°C (3 min 30 s), 32 cycles 94°C (20 s), 67°C (1 min), 68° C (3 min 30 s), and elongation 68° C (15 min).

Cloning and sequencing

PCR products were separated by agarose (1%) gel electrophoresis and purified (Qiagen). Subsequently, the purified products were cloned into pCR-XL-TOPO-vector and transfected into *Escherichia coli* (Top 10) (Invitrogen, Carlsbad, CA, USA). Gel extraction, cloning, and transfection procedures were conducted according to the manufacturers' instructions. Transfection was then validated by PCR, with primers and PCR conditions as above, and 95°C (5 min) denaturation and 72°C elongation steps, using Taq (MBI). Plasmids were then purified in 96-well format (Macherey-Nagel). Sequencing primer design was based on *UTY* RefSeqs. Sequencing-PCR was performed using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Darmstadt, Germany), Robo Cycler 96 and PCR conditions: 95°C (1 min), 37 cycles 95°C (30 s), primer specific annealing (49–57°C, 45 s), and a final 60°C (4 min). PCR products were purified in 96-well format (Qiagen or Millipore, Schwalbach, Germany) and sequenced with ABI 3730 Analyzer (Applied Biosystems). Sequences were evaluated by Sequencing Analysis Program V 5.1 (Applied Biosystems) and carefully manually inspected.

Transcript analysis and identification of novel AS-events and new isoforms in UTY

UTY sequences retrieved from sequencing were compared to human mRNA RefSeqs (*UTY*_v1, v2, v3) and to the respective genomic sequence (ENSEMBL: *UTY* ENSG00000183878) by a single-transcript approach and consequent manual segmenting. For this analysis, EMBOSS pairwise alignment algorithm with Needle (global alignment) (www.ebi.ac.uk/emboss/align) was applied and sequences were subsequently manually evaluated. In this way also the selectivity against almost identical UTX had become validated. For splice-site analysis the ENSEMBL *UTY* gene sequence (see above) was used. This analysis focused on the terminal dinucleotides of both introns, which flank AS-events or which become newly generated by AS-events. A subsequent homology search by Blastn (www.ncbi.nlm.nih.gov/blast/Blast.cgi) was conducted for specificity. Such identified novel transcripts were submitted to GenBank. To test the predictability of AS-events GENSCAN (http://genes.mit.edu/GENSCAN.html) was applied. Further, ENSEMBL Vega/Havana Project Databases (www.ensembl.org/index.html) were searched for predicted *UTY* isoforms.

Prediction and classification of in-frame and PTC-events

AS-events' effects on the reading frame were first predicted upon conceptual translation using EMBOSS transeq (www.ebi.ac.uk/emboss/align). PTC introducing events were then defined according to the rule for NMD (Lejeune and Maquat, 2005). Therefore, in this study all AS-events that inserted a premature stop codon >50 bases upstream of the final exon/exon border were categorized as a PTC-event.

Nomenclature for UTY *splicing and graphical illustration*

In this work, all modifications of the constitutive exons and their consecutive organization in transcripts were defined as AS-events. To structure the diversity of *UTY* splicing effectively, all internal AS-events observed in this study were then named and abbreviated by an individual description. It is comprised of the respective splice form(s) and their regions of occurrence. The latter were denoted by the respective constitutive introns or exons affected in pre-mRNAs. Thus, each modification of constitutive splicing in *UTY* by AS can be easily assigned to splice forms and region.

Concerning the classification of splice forms, we faced the challenge, that besides the observed five internal basic splice forms, as alternative cassette exon, intron retention, alternative 5′ and 3′ splice sites and exon skipping (Blencowe, 2006), no universal terminology or graphs for complex splicing

exists (Sammeth et al., 2008). Thus, in this work, complex AS events were classified and termed according to their combined splice forms (Table 1, and Supplementary Table 1).

In this work the following abbreviations were used: constitutive exons were abbreviated "ex" and denoted 5′ to 3′ by serial numbers (ex1–ex29). Constitutive introns were named to their flanking exons (e.g., 12/13). AS-events were abbreviated as follows: for all AS-events alphabetical naming like "a," "b," "c," represents 5′ to 3′ order, alternative cassette exons (Ex) were denoted according to their location in introns (Ex12/13a, Ex12/13b). Retained introns (rI) were denoted by adjacent exons (rI21–22). To describe alternative 5′ and 3′ splice sites (ss) precisely, they were abbreviated 5′ and 3′ of the respective exons (3′a20) denotes 3′a ss of ex20. This description was preferred over an intron-wise, as the multitude of AS-events' combinations created a huge number of distinct novel observed and potential intron isoforms, which did or may contain all together the same alternative ss. Further observed substantial reductions of >90% of the exon length by alternative 5′ or 3′ss were subgrouped and denoted as large, for example, 5′L-15 for a 5′ss modification of ex15.

In view of the numerous distinct single and multiexon skipping AS-events and their combination with other splice forms, only one abbreviation "D" was used for all these AS-events, to facilitate understanding of our focus on transcript-diversity. "D" refers to the common observation that on transcript level the diversity of AS-events results only in two categories either "insertions" or "deletions." "D" AS-events were further defined by the excised respective constitutive exons, for example, (D5–15); combined splice forms were indicated, for example, (D5′2–3′b11). Finally, the observed internal segment exclusions of ex16 were termed central deletions. They were comprised of a short (pDcs16) and a long variant (pDcl16). The third AS-event whose classification depends on the central deletions, was termed pDc16 + 3′16.

For a graphical representation on alternative splicing, pre-mRNA splice graphs were used (Blencowe, 2006). However, as our study focused on the great diversity of *UTY* in both AS-events and transcripts, a graphical presentation that illustrates the key features of both levels, modified complex transcript-architecture and underlying AS-events, was created to enhance comprehension. This graph displays the number of modifications per transcript, the regions affected by AS, underlying AS-events, and their putative effects on the reading frame (in-frame or PTC). Transcripts are illustrated by constitutive exons, modifications are denoted by symbols that indicate the group of underlying splice forms and the putative effects in-frame—blue, or PTC introducing—orange. Underlying AS-events are denoted by abbreviations.

Quantitative PCR

qPCR was performed on Taqman ABI-Prism 7900HT. iTaq SYBR Green Supermix with ROX (Bio-Rad, Munich, Germany) was used with conditions: 95°C (10 min), 38×95°C (15 s) and 60°C (60 s), and final dissociation step 95°C (15 s), 60°C (15 s) 95°C (15 s). Primers for both analyses were preselected by sequencing the respective qPCR-products at conditions as above, with 59°C annealing temperature. General *UTY* expression levels were determined by qPCR, at which one primer covered the conserved exon border. For the verification of selected AS-events, primers used covered the respective exon border(s). Subsequent sample analysis for the detection of AS-events was then performed with such determined dissociation curves.

Estimation of the UTY *splicing extent*

Basic assumptions. To determine the potential transcript diversity we built a new *UTY* splicing graph and integrated our observations in the calculation. We included the existence

Table 1. Alternative Splicing Classes and Number of Internal Alternative Splice Events in Human *UTY*

	Total AS-events			*In-frame AS-events*			*PTC AS-events*		
Alternative splicing classes	*Sum*	*New*	*Known*	*Sum*	*New*	*Known*	*Sum*	*New*	*Known*
Single splice forms									
Cassette exon	21	19	2	4	2	2	17	17	—
Alternative 3′ ss	13	13	—	7	7	—	6	6	—
Alternative 5′ ss	5	5	—	2	2	—	3	3	—
Retained intron	2	2	—	—	—	—	2	2	—
Exon skipping	14	12	2	6	5	1	8	7	1
Complex AS-events									
Cassette exons + alt 5′ or 3′ss	7	7	—	1	1	—	6	6	—
Retained intron + cassette exon	2	2	—	—	—	—	2	2	—
Multiexon skipping	22	22	—	12	12	—	10	10	—
Multiexon skipping + alt 5′, 3′ss	4	4	—	1	1	—	3	3	—
Combined alt 5′ + 3′ss	1	1	—	—	—	—	1	1	—
Extraordinary AS-events									
ex16 central deletion	3	3	—	1	1	—	2	2	—
Total	94	90	4	34	31	3	60	59	1

Observed in this study for the region between ex1 and ex29. For each alternative splicing class the total number of AS—events and the portion of in—frame and PTC events are given, "new" denotes AS—events discovered in this work, "known" AS—events annotated in databases before this study. Lack of any known variants is denoted by "—." Identities of AS—events are listed in Supplementary Table 1; for GenBank accession numbers of the previously annotated sequences, see Figure 3A. After our annotation was completed, one AS—event D16 was annotated by another group (partial sequence *UTY*, GenBank AK308952). Further abbreviations: ss, splice site; alt 5′ 3′ss, alternative 5′ or 3′ splice sites; ex16, constitutive exon 16; PTC AS—event, premature termination codon introducing AS—event. For AS—events classification see also sections Results and Methods.

of three polyadenylation variants (ex1–29, ex1–26, ex1–29), the occurrence of 94 internal AS-events, one to eight internal splice events are considered per transcript, a limited number of PTC and in-frame events per transcript (one to four PTC and one to four in-frame-events, also one to three PTC and one to five in-frame events). We considered as well the existence of many mutually exclusive AS-events.

We represented all possible transcripts in a gene translation using a splicing graph G, which somewhat differs from the ones previously defined (Chacko and Ranganathan, 2009; Heber et al., 2002; Leipzig et al., 2004). In this work we ordered AS-events by increasing position in the gene, and gave them according numbers $1,\ldots,N$. A transcript is then uniquely defined by an N-tuple×with entries from {0,1}, that is, a number from 2^N, where x_i is 1 if splice event i has been used in the transcript.

Clearly, not all numbers from 2^N are valid due to mutual exclusivity of many events. Hence, we defined a splicing graph as a directed acyclic graph (DAG) on the nodes $1,\ldots,N$ and an additional terminal node. We added an edge from $i \to j$ if event j can follow directly after event i, and label it with 2 if the splicing event is taken and with 1 if not. In particular, this implies that links are only allowed from lower to higher numbers, and that the graph is binary in the sense that each parent has precisely two children or none if it is a leaf. Note that the resulting DAG is a multigraph, because we need to include cases where independent on whether event i has been used, it is followed by event $i+1$.

Enumeration of transcripts using the splicing graph

We enumerated all possible transcripts (i.e., subset of 2^N) as all maximal paths through our graph G. Here, a maximal path was defined as a path of G that is not a proper subset of another graph of G. Due to construction of G all such paths necessarily start at node 1, and end at the end node $N+1$. Without constraints, we can determine the number of paths from 1 to $N+1$ as follows. Let A denote the adjacency matrix of G, which consists of entries A(i,j) counting the number of edges from node i to node j. By construction A is an upper triangular graph with zeros on its diagonal and $A(i,i+1) > 0$. Because A is a multigraph it may have entry 2 instead of 1 if two edges exist between i and j.

It is well known that in general the (i,j)-th entry of A^n contains the number of directed paths of length n from node i to node j, and it can be easily seen that this also holds in the case of a multigraph as constructed above. Because we are dealing with a finite DAG, all paths are of finite length, and therefore A nilpotent, in particular, $A^{N+1} = 0$. So we can determine *all* paths from 1 to $N+1$ simply by $\sum_{n=0}^{N} A^n(1, N+1)$. In order to reduce computational effort, we can use the geometrical series for matrices to see that this equals $(1-A)^{-1}$ $(1,N+1)$. This matrix inversion however needs to be calculated at a high precision.

To count paths from 1 to $N+1$ with additional constraints such as a limit on the number of events to be taken, we recursively transversed the graph splitting up at each parent into both directions up to a maximal number. This 100-dimensional graph transversal may be costly but can be calculated very efficiently by memorization, that is, by storing previously calculated results. Then, for example, AS-event 1 takes place, then the transcription stops, if not also event 2 might take place. So we have transcripts 00,10,01 but not 11, so 3 out of the $2^2 = 4$ possible ones.

Determination of AS-event frequencies by a probabilistic splicing graph and a mixture model approach

In the second step of the analysis we added probabilities to the graph, estimated from observed transcripts. These probabilities reflect the dependency structure of the AS-events. Here we are in the special case of binary graphs, so we can simplify the probability definitions at each edge and only specify a splice event probability at each node. Because the graph is a DAG, we represented the underlying joint probability using a Bayesian network.

A probabilistic splicing graph therefore was defined as a splicing graph G together with N probabilities $\mathbf{p} = (p_1,\ldots,p_N)$ from $[0,1]^N$. The random variable S_i captures the case (2) if an event is taken and (1) if not. The key idea of the whole approach now is that we did not assume independence, that is, $P(S_i, S_j) = P(S_i)P(S_j)$, but instead used conditional probabilities, derived from the probabilistic splice graph by $P(S_j \mid S_i = 2) = p_i$ $P(S_k \mid S_i = 1) = 1 - p_i$.

Assuming a memory-free process for transcript generation we determined the probability of a transcript by the product of these conditional probabilities along its defining path in G.

We estimated the probabilities from the data as follows. Instead of counting the relative probability of all AS-events S_i (as done with the assumption of full independence) we only used the subset of samples that have a path containing node i at all. There we determined relative probability of S_i as relative frequency. Both events S1 and S2 have high probabilities of 0.8. However, if we only count total relative probabilities, we would correctly estimate S1 to have 0.8, but S2 to only have $0.2 \times 0.8 = 0.16$, and hence, incorrectly disregard it as infrequent.

Exponential mixture model

The distribution of conditional probabilities of AS-events is to be analyzed with respect to contributions from possibly multiple sources. For this an exponential mixture model was hypothesized:

$$f(x; w_1, \mu_1, \ldots, w_r, \mu_r) = \sum_{i=1}^{r} w_i \frac{1}{\mu_i} e^{-\frac{x}{\mu_i}} \text{ subject to } \sum_{i=1}^{r} w_i = 1.$$

This mixture model was fitted for varying number r of exponential distributions to the data using maximum likelihood estimation. Confidence intervals for each parameter have been determined to the 95% confidence level.

Analyses on the effect of splicing on the UTY *protein domain structure*

In all protein analyses a direct translation of the experimentally identified transcripts sequences by EMBOS transeq (see above) was used to reveal amino acid sequence. Analyses of encoding exons, all protein domains, and structure assignments were performed as described (Ginalski et al., 2004).

Yeast-two-hybrid assay

Yeast-two-hybrid screening was performed according to the manufacturer's protocol (matchmaker Gal4-Two-Hybrid

System3, Clontec, Heidelberg, Germany). The central region of *UTY*_v26 (aa 278–672, bp 834–2016) served as bait. It was amplified by PCR, cloned into the pGBKT7 vector via NDE I and Bam Hi restriction sites, and was then transformed into the yeast strain AH109. As prey, a pretransformed human bone marrow cDNA library was used (pGADT7-Rec, strain Y187, Clontec). Yeast strains were mated, and diploids grown at 30°C on SD plates under highest stringency conditions lacking the nutritional components *Ade*, *His*, *Leu*, and *Trp*. Colony growth was recorded over a period of 12 days. Colonies larger than 3 mm and of pink color were further tested for MEL1 expression with X-α-Gal (BD Bioscience, San Jose, CA, USA), according to the manufacturer's instructions. Positive clones were then analyzed for their bait and prey by PCR/sequencing. Prey sequences were then identified by Blastn and further selected for in-frame sequences by EMBOSS Transeq. Retransfection of identified prey and bait, according to the manufacturer's, confirmed interaction.

T-cell epitope prediction

MHC binding probabilities were defined by "epitope prediction" of SYFPEITHY data (Rammensee et al., 1999). UTY-specificity was determined by comparison to UTX with EMBOSS align and by protein Blast using human protein and Swissprot databases.

Tissue preparation

Human hematopoietic cells, mainly enriched fractions of $CD34^+$ progenitor cells and leukemic blasts, and fibroblasts were studied. Samples were taken from diagnostic material or from healthy volunteers after informed consent.

$CD34^+$ progenitor cells were collected after granulocyte colony-stimulating factor mobilization, enriched fractions comprised (70–97%). For qPCR, the corresponding $CD34^-$ fractions and peripheral blood lymphocytes of nonmobilized donors were additionally tested. Leukemic samples had been diagnosed as acute myeloid leukemia (AML), blast content (40–70%), and one related sample as Hodgkin. Probes originated from bone marrow or periphery.

CD34-positive selection was conducted according to the manufacturer's (Miltenyi, Bergisch Gladbach, Germany) with CD34-Clinimacs-kit and LS-Separation columns. Selection-quality was controlled by flow-cytometry, with CD34 (PE)-antibody (clone 8G12, IgG1, BD-Bioscience), isotype controls (BD), and propidium iodide staining.

Fresh leukemic samples were obtained after Ficoll separation. However, because the majority of leukemic samples was cryo-preserved, no further enrichment of blasts was performed to avoid false positives. Cryo-preserved samples were incubated for 30 min, bone marrow in DMEM, with 10% serum, PBLs in VLE-RPMI 1640 with 7% HSA. Subsequently, blast content was selectively reestimated by multiparameter flow-cytometry (fourfold analysis of CD3, CD14, CD19, CD33), monitoring the CD33 content against CD14. For all FACS analyses, FACSCalibur flow cytometer (Becton Dickinson, Heidelberg, Germany) and Cell-Quest-Pro-program were used. Cryo-preserved primary fibroblasts were recultured in DMEM, supplemented with 10% fetal bovine serum (FBS) (Gibco, Grand Island, NY, USA), and 1% L-glutamine. For RNA isolation, cells were harvested by trypsination. In all cells and tissue, each sample became retested for vitality by Trypan blue staining (Sigma, St. Louis, MO, USA), and was washed twice in PBS prior to RNA isolation.

Results

Evidence for a multitude of novel UTY *splice variants*

In our research on the role of *UTY* in leukemia, we analyzed the intriguing effect of tissue-specific UTY peptide recognition despite ubiquitous *UTY* expression. To test if yet unknown modulation of *UTY* on mRNA level exists, we conducted expression analyses in primary cells, enriched leukemic blasts, hematopoietic progenitors, and primary fibroblasts. To examine first whether all three RefSeqs were generated in our tissues, we amplified them by specific RT-PCR (see Methods). Surprisingly, all subsequent agarose gel electrophoreses showed multiple bands. These bands ranged in molecular size from 0.5 to 4.5 kb, occurred in all samples, and displayed no tissue specificity. Control PCRs of cloned amplicons confirmed numerous distinct molecular mass sizes (Fig. 2A–B). Subsequent validation by DNA sequencing identified 622 clones (over 90% of the amplicons cloned), as *UTY* transcripts. We had focused this analysis primarily on high molecular mass bands, 1.8–4.5 kb, to avoid possible degradation effects. First approaches to fully align sequences to *UTY* displayed very complex splicing patterns, which precluded the use of standard multiple alignment algorithms.

We then developed a strategy to effectively target this diversity. We applied a single-transcript approach using long range DNA-sequencing segments (>700–900 bp) as "anchor" to identify *UTY* and to locate AS-events in the transcript. This was followed by stepwise manual segmenting and pairwise alignment to both RefSeqs' exons and the *UTY* gene sequence.

This first analysis revealed that indeed the alternative splicing was abundant in all probes and in the all three polyadenylation variants. Of the 622 *UTY* transcripts isolated, 85% were spliced differently from the RefSeqs. These results indicated that we discovered evidence for a novel *UTY* splicing phenomenon. Our observations prompted us to focus our study on the new transcript diversity in *UTY*.

New genomic exon organization in UTY

To target *UTY* splicing complexity we systematically analyzed transcript architecture and genomic organization. In the first step, we examined whether *UTY* transcripts indicate the existence of new exons. At present, the genomic organization in *UTY* has 31 exons (Fig. 2F, below). Of these, 29 are included in the RefSeqs and two further exons occur in a partial sequence (GenBank CR936684). In our *UTY* clone set, we identified both the known 31 exons and further novel 19 internal exons (see below). These observations indicated a novel genomic exon/intron organization in *UTY*. It differs considerably by exon number, that raises from 31 to 50 (Fig. 2F, above).

As the high percentage of transcripts was spliced differentially from the RefSeqs, we examined in the second step which exons represent the constitutive organization. As constitutive exons are often defined by prevalent exon inclusion in transcripts, we determined the predominant exons in our

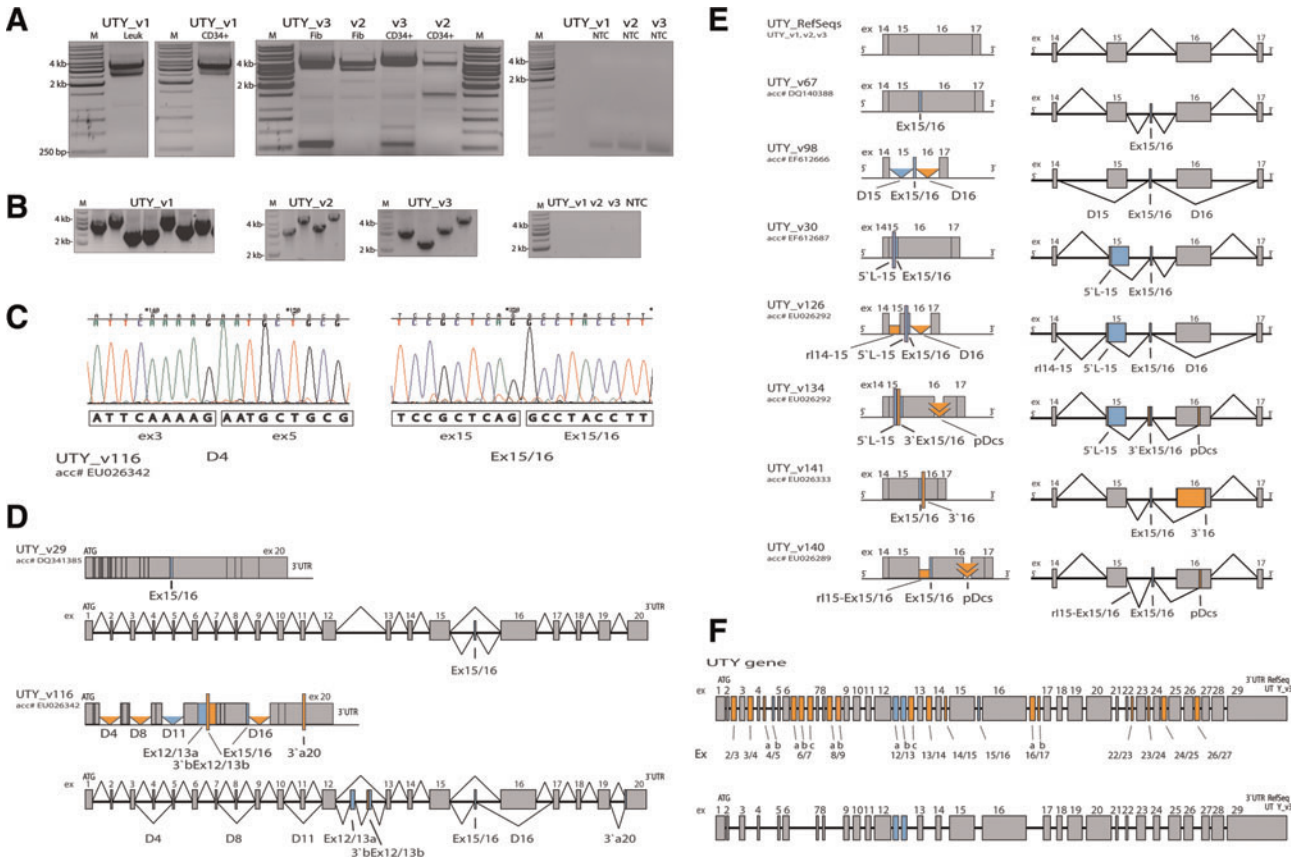

FIG. 2. Novel *UTY* transcript isoforms are abundant, display very complex splicing patterns, and indicate a novel genomic exon/intron organization. (**A**) Gel electrophoresis of *UTY* RT-PCR products. $CD34^{+}$, enriched hematopoietic progenitors; Leuk, enriched leukemic blasts; Fib, fibroblasts. (**B**) Gel electrophoresis of cloned *UTY* amplicons display large differences in molecular masses. (**A, B**) M, molecular mass standard; NTC, nontemplate control; *UTY*_v1 v2, v-3, *UTY* mRNA RefSeqs, whose full ORF had been covered by selective RT-PCR (see Methods). (**C**) Examples of novel alternative splice events identified by DNA-sequencing. (**D**) New transcript modification patterns in human *UTY:* illustration of minimum (1) and maximum (8) number of modifications per transcript as observed in this study. Graphical illustration of the transcript level (above) (see also Methods) and corresponding pre-mRNA splicing scheme (below), pre-mRNA splicing graph—constitutive splicing above, alternative splicing events, as observed in the respective transcript, below. (**E**) Very complex *UTY* transcript compositions by almost randomly combined underlying AS events, as newly observed in this study. Graphs exemplify transcript modifications in the region between constitutive exon 14–17 as created by the inclusion of cassette exon 15/16 and its distinct combinations with other AS-events. Seven out of 11 observed distinct combinations with cassette exon 15/16 are shown. Graphical illustration of the transcript level (left), corresponding pre-mRNA splicing scheme (right). Pre-mRNA splicing graph; only the observed splicing of the respective transcript is shown; above, constitutive splicing; below, alternative splicing events. (**D–E**) Transcript level graph: symbols indicate grouped splice forms, Pre-mRNA splicing scheme: central black line: introns (not to scale). The underlying internal AS-events are denoted by abbreviations, numbers indicate the constitutive region affected (see also Methods). (**F**) Novel genomic exon/intron organization in human *UTY*. Above—as indicated by this study, below—as previously known. Exons are to scale, whereas introns are not, due to very different sizes (about 0.4–37 kb). (**C–F**) ex—constitutive exons (1–29), 12/13—constitutive intron, Ex—cassette exon, rI—retained intron, 5′ 3′-alternative spice sites, 5′L substantial reduction of exon length, D—summarizes AS-events involving exon skipping, pDcs—exceptional exon 16 splicing, short variant. Symbols: small gray upright bars—constitutive exons, small colored upright bars—alternative exons, large upright bars 5′or 3′ ss, in the pre-mRNA splicing scheme 5′ and 3′ ss are illustrated by inner exon bars, triangle—deletion (exon skipping), double triangle—central deletion small (ex16), rectangle—retained intron, large black central line—introns, blue—in-frame-events, orange—PTC-events (for abbreviations, see also Methods, Databases for C–F Supplementary Tables 1 and 2).

above clone set (data not shown). We found all 29 exons of the RefSeqs (Fig. 1A) to prevail. We used this exon organization and its respective introns as basic constitutive setting and defined each modification as AS-event. In this work, constitutive exons (ex) were denoted 5′ to 3′ by serial numbers (ex1–ex29) and constitutive introns were named to their flanking exons (e.g., 12/13) (see also Methods).

Detection of novel AS-events in human UTY

Systematic transcript analysis revealed 90 novel internal AS-events in *UTY*. In addition, we confirmed four known internal AS-events, and also partial 3′ UTR elongations of ex20, 26, and 29 (Table 1, new AS-events exemplified see Fig. 2C, 3′UTRs see Supplementary Table 2). In our set of 94

internal AS-events, we observed five single basic forms of alternative splicing (Blencowe, 2006; Breitbart et al., 1987): alternative 5′ and 3′ ss, intron retention, exon skipping, and alternative cassette exon inclusion. In addition, we found 36 complex AS-events, in which these single forms were combined in many different ways. Finally we observed two different AS-events of inner exon splicing in exon 16 (Table 1 and Supplementary Table 1). For complex AS-events, we observed 11 subgroups of multi exon skipping, that ranged in numbers of consecutive exons skipped from 2 to 24. Further, four multiexon skipping events used alternative ss, either in both or in the 3′-position. Alternative 5′ or 3′ss occurred also combined with five cassette exons, of which two 3′ss existed for Ex12/13b and Ex15/16, respectively. Moreover, we identified combined alternative 5′ss and 3′ss between the two constitutive exons 10 and 11. Intron retention occurred between a constitutive and a cassette exon and also between consecutive cassette exons. Besides complex events, we also found an exceptional size reduction (>90%) of three exons by alternative 5′ or 3′ ss. The largest reduction, from 396 nucleotides (nt) to 24 nt, occurred in exon 15. Finally, we observed two cases of inner exon splicing in exon 16. A large segment (435 nt) and a small part of it (91 nt) became excised. This can be interpreted in two ways: either as an exceptional form of splicing, as inner exon splicing, or as a presumed reactivation of an ancient retained intron with a variable 5′ ss. The latter is corroborated by the dinucleotides GT/AG at the 5′–3′ terminal sites of both excluded segments. As the focus of this study was on the systematic exploration of the *UTY* splicing extent, we did not address this observation further. We termed it central deletion, long and short. A third AS-event was included in this group for classification reasons (Supplementary Table 1; see also Table 1).

In our set of 94 internal AS-events, the largest group was the multiple exon skipping group, followed by the cassette exon group (Table 1). Further analysis on the effects of AS on the transcript length revealed very great diversity: indels by AS-events ranged in molecular size from 3 nt (3′21) to 3,503 nt (D3–26) in relation to the corresponding RefSeqs (Supplementary Table 1).

Verification and further examination of AS-events

We first validated the alternative splicing of *UTY* transcripts by analysis of position and of splice sites: we found that each AS-event occurred only once per transcript, and that its position corresponded to the new genomic organization. We then examined the *UTY* gene sequence for bases, which correspond to pre-mRNA splice sites. To cover also potential, yet unknown intron isoforms, we focused on the terminating dinucleotides of introns, which flank AS-events. Eighty-nine percent of AS-events were flanked by the standard forms GT (5′ ss), and/or AG, (3′ ss), or were GT/AG introns (Supplementary Table 1). As constitutive introns were also GT/AG introns, the common GT/AG sites clearly prevail in the here observed *UTY* splicing. Further verification of AS-events by common hybridization techniques was severely constrained by the complexity of splicing. Therefore, we validated selected AS-events by sequencing, qPCR and consequent dissociation curve analysis. Additionally determined general *UTY* expression in tissues showed low levels (Ct ~25–28). Finally, AS-events were categorized as putative in-frame or PTC introducing event (PTC-event), by prediction (see Methods). We found 34 in-frame and 60 PTC-events (Table 1 and Supplementary Table 1). Concerning the tissues tested, about 30% of AS-events occurred in all and about 73% in normal tissue. The AS-events newly identified here, significantly increased the number of known internal (ex1–29) AS-events for *UTY* from 4 to 94.

Identification and analysis of 284 distinct novel transcripts of UTY

We then examined systematically the splicing patterns in our clone set. We identified 284 novel distinct patterns in *UTY*, which were spliced differently from the RefSeqs. Of these novel isoforms, 69 were in-frame and 215 were PTC transcripts. Their transcript architecture displayed a wide diversity (described in more detail below, and Supplementary Table 2). Also, consecutive AS-events occurred, that ranged in number from two (e.g., *UTY*_v111) to four (e.g., *UTY*_v261) (Supplementary Table 2; illustration of consecutive AS-events, see also Fig. 2E). In our set of 287 different *UTY* transcripts (284 new and 3 RefSeqs) 72% corresponded in their in-frame/PTC splicing patterns in a way, that in PTC transcripts PTC-events occurred additionally to the common in-frame pattern. These corresponding patterns occurred in all tissues. The novel *UTY* splice variants were annotated at GenBank and named *UTY*_v4 to v287, as recommended by HUGO (Supplementary Table 2). Our observations increased the number of annotated distinct isoforms of human *UTY* from 6 to 290, and that of encoded *UTY* proteins from 6 to 73. The new *UTY* isoforms were not predicted by GENSCAN or ENSEMBL Vega/Havana. As all samples of cells and tissues displayed very great transcript diversity (about 50–70% of the respective clones isolated), we assumed the observed *UTY* splicing diversity as a general phenomenon. We therefore focused our study on the further systematic bioinformatic analysis on its inherent key features and its potential splicing extend.

UTY *splicing system architecture provides the basis for extraordinary diversity*

We examined whether inherent key components exist that structure *UTY* transcript diversity. We first analyzed the distribution of internal AS-events in our transcript set (287 distinct isoforms, 284 new, and 3 RefSeqs). We identified four key components of *UTY* transcript architecture that allow to promote diversity: (1) transcripts contain 1 up to 8 AS-events (exemplified in Fig. 2D), (2) AS-events were combined in a multitude of ways (Fig. 2E), (3) frequent occurrence of mutually excluding AS-events. In addition, as splice sites occur 5′ to 3′ in pre-mRNAs, theoretically 91 AS-events have mutually excluding events, which number from 1 up to 15. Only three AS-events, D5′2–3′b11, Ex26/27, and D27-28 have no excluding counterpart; (4) identical splice patterns of the three polyadenylation variants tested occur (for data to all points see Supplementary Table 2).

We then mapped AS-events to their corresponding splice sites in the *UTY* gene, and revealed a novel very complex genomic organization (Fig. 3A–B). AS effects all but one constitutive exon–exon border (ex1–ex2); further, alternative splice sites were almost evenly spread in *UTY* (Fig. 3B). By these features the new genomic organization provides a

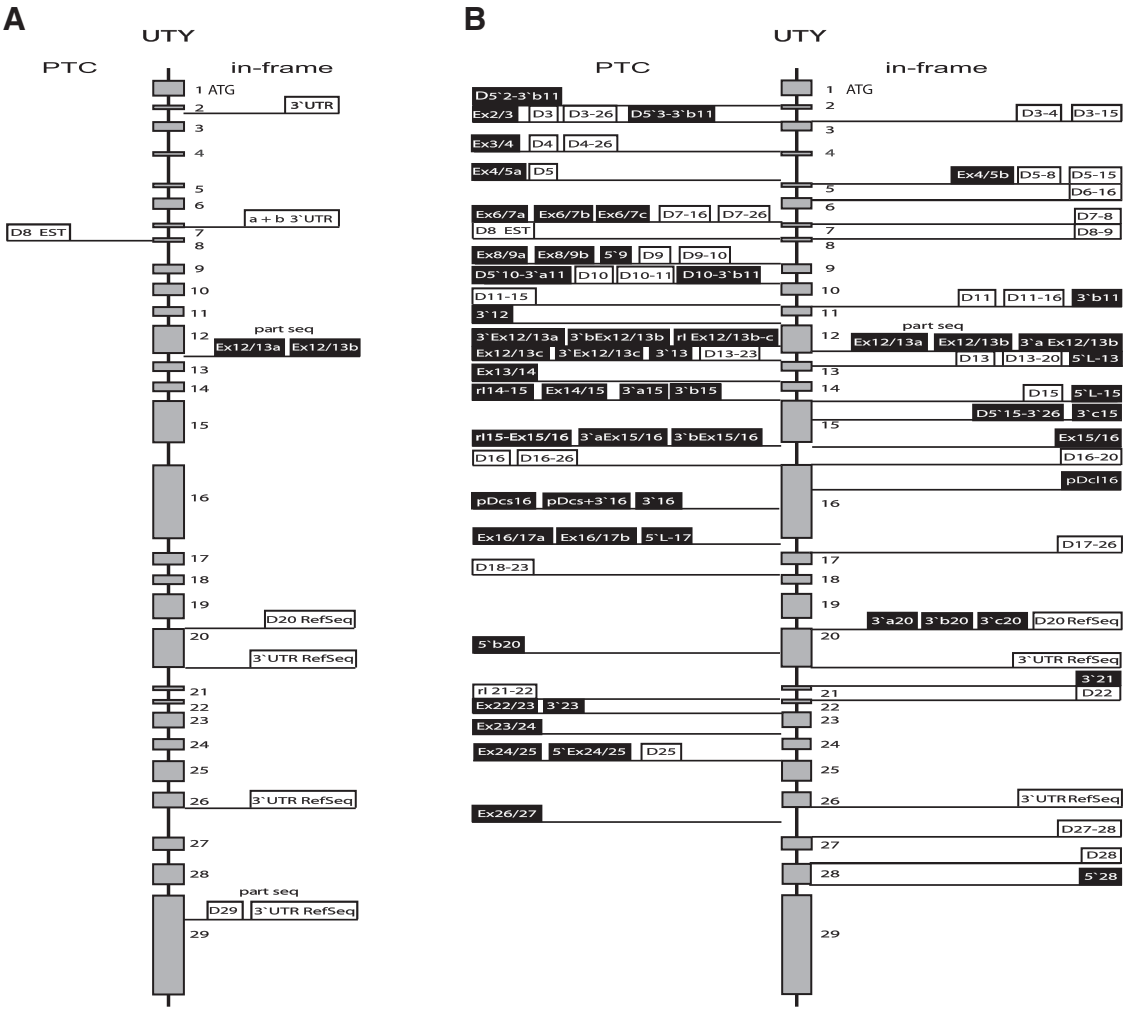

FIG. 3. Alternative splicing affects almost all regions of the *UTY* gene. Overview on genomic sites in human *UTY*, which correspond to regions affected during pre-mRNA AS. AS-events are assigned 5′ to 3′ to the regions containing the respective splice site(s), the precise regions affected are denoted by AS-events abbreviations in the boxes right and left of the central bars. 3′UTRs are additionally indicated. (**A**) Previously known AS-events. RefSeqs and independent short transcripts, partial sequences and EST (GenBank RefSeqs: NM_007125.3, NM_182659.1, NM_182660.1; further polyadenylation variants BC012581, BC029389, BC071744; Cassette exons: CR936684, exon skipping: BX090888). (**B**) Internal (ex1–29) AS-events added by this study, including four confirmed events. Confirmed previously known AS-events are noted by EST and part seq. Concerning encoded proteins, note that constitutive exons 3–12 code for the tetratricopeptide repeat domain (TPR) and exons 23–26 code for the jmjC domain. (**A** + **B**) Central bars, constitutive genomic exon organization: gray boxes denoted by numbers—constitutive exons, black vertical line—introns. Boxes right and left of the central bars—AS-events, black boxes indicate use of new splice sites. Confirmed AS-events: Part seq.—partial sequence, EST—expressed sequence taq, RefSeq refers to their 3′UTRs.

substantial basis for multiple ways of AS-event combinations. Therefore, both levels, genomic and transcript, contribute structural key components, which endorse observed transcript diversity. In summary, our analysis shows for the first time, that huge transcript diversity is an inherent key feature of *UTY* alternative splicing.

Estimation of splicing extent by a new model multigraph

Our data on *UTY* splicing complexity strongly suggested that further distinct isoforms exist. We therefore estimated the *UTY* splicing extent. To cover the diversity, we build a model multigraph and included our observations in the calculation (see Methods). In this new *UTY* splicing-graph (Supplementary Fig. 1), each possible transcript uniquely corresponds to a path, with AS-events taken or not, according to the choice of edges at each AS-event node. We determined the total number of distinct transcripts by counting the number of distinct paths. Without restrictions this amounts to 1.57×10^{17} transcripts. We have observed, however, that *UTY* transcripts contain only up to eight different internal AS-events; further, out of these at most four are in-frame or PTC events. This constraint reduces the number of transcripts to $1{,}077{,}609{,}760 \approx 1.08\times10^{9}$. In some additional cases, inclusion of at most three PTC-events and precisely five in-frame events occurred. This amounts to $240{,}549{,}403 \approx 2.41\times10^{8}$ transcripts. Therefore, without further restrictions, AS of *UTY* has still the potential to produce $39{,}238 \approx 3.92\times10^{4}$ in-frame

isoforms and in total 1,318,159,163 $\approx 1.32\times10^9$ distinct transcripts. To our knowledge, this is the greatest estimated transcript diversity, coded by a single gene via alternative splicing.

Modes of transcript generation

This outstanding potential raised the question on modes of its generation. We analyzed whether *UTY* splicing is composed of one component, as, for example, supposed solely random choice, or of more components, as, for example, additional selective choice of AS-events. To test whether AS-events differ significantly in their frequencies, we calculated the conditional AS-event probabilities of *UTY*. Using a Bayesian network, we described conditional splicing probabilities and estimated them in the data set (Supplementary Table 3). Clearly, a few prominent high-probability AS-events are followed by many low probability events. We therefore estimated a mixture model for AS-event probabilities, based on the assumption that random choice of AS-events lead to a purely exponential distribution of AS-events (see Methods).

Three models with $r = 1, 2, 3$ exponential distributions were fitted (Fig. 4A). The resulting mixture parameters (contribution of first/second component) were $w_1 = 0.34 \pm 0.13$ for $r = 2$ and $w_1 = 0.32 \pm 0.20$ and $w_2 = 0.045 \pm 0.16$ for $r = 3$. Therefore, we cannot assume significant deviation of 0 of the mixture parameters for the three-component model, in contrast to the two-component model. This implies that we can interpret the AS-event probabilities within a two-component mixture model (Fig. 4A).

Their mean values are $\mu_1 = 0.11 \pm 0.044$ and $\mu_2 = 0.0072 \pm 0.0026$. So the second stronger exponential, which accounts for roughly 70% of the data, is quickly decaying. It explains the low probabilities, which we interpret as result of random splicing. This is superimposed at threshold 0.0258 (Fig. 4B) by a slower decaying smaller power exponential, which explains the low-frequency high probabilities.

In our transcript set 28 internal AS-events displayed high frequent occurrence (Supplementary Table 3). They occurred in all tissues and accounted for about 88% of all modifications observed, which suggests preferred selection of AS-events. The large low frequent group showed dispersed inclusions, that is, random splicing. The additionally observed occurrence of many highly related in-frame and PTC transcripts in all tissue samples points at subsequent PTC/NMD driven regulation.

jmjC domain is conserved in proteins encoded by novel splice variants

We then assessed the implications of AS on the protein level upon direct translation. The new *UTY* isoforms of our here isolated set (72 isoforms, 69 novel, and 3 RefSeqs) encoded a huge protein heterogeneity. Among novel isoforms the predicted protein size ranged from 376–1,444 amino acids (aa) and surpassed the hitherto longest isoform RefSeq v3 (1,347aa) by an additional 97aa (*UTY*_v59; Supplementary Table 2). Further, in our set, AS substantially modifies proteins in the TPR domain, central region and C-terminus, whereas protein topologies of putative signal peptide, jmjC domain and α/β-fold were mainly conserved. Very rarely jmjC domain became deleted in total. No in-frame alterations of jmjC domain occurred, which may indicate functional selection (for regions affected by AS, see Fig. 3B; database Supplementary Table 2).

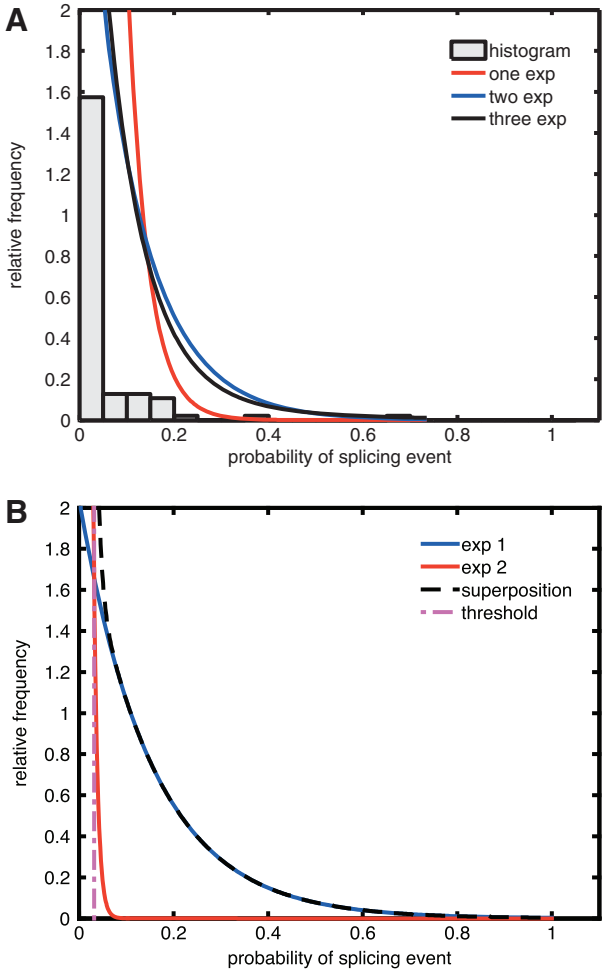

FIG. 4. Modes of *UTY*-transcript generation. (**A**) Fitting a linear, weighted mixture model of exponential densities to the probability distribution of AS-events (data visualized by histogram with bin size 0.05). The three curves show theoretical distributions of iteratively more complex mixture models. Confidence interval analysis shows that $r = 2$ exponentials are sufficient. (**B**) Determination of a threshold in the splicing probabilities. The mixture model for $r = 2$ is visualized. The two already weighted exponentials *exp1* and *exp2* are shown.

Analyses of protein functionality by interaction study

To gain new data on UTY functionality we then checked if besides the N-terminus a further region may be involved in protein interaction. Our analyses had shown already, that AS substantially modifies central UTY that also contains a large unstructured region (Fig. 5A). Therefore, we screened central UTY segments for protein interaction by yeast-two-hybrid assay, using directly translated baits. For a bait containing the two high frequent cassette exons, Ex12/13b and Ex15/16, we identified two novel UTY protein interactors: aryl hydrocarbon receptor interacting protein (AIP) and guanine nucleotide binding protein (GNB2L1) (Fig. 5B), and further verified this by coimmunoprecipitation *in vitro* (data not shown).

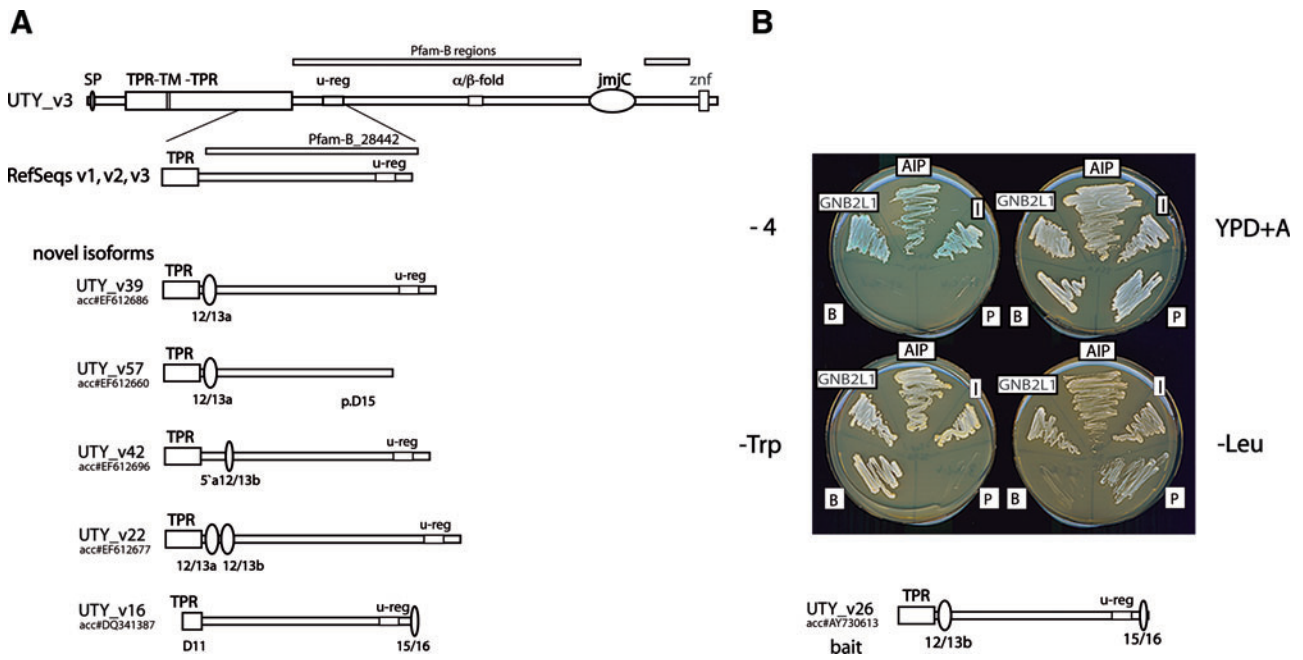

FIG. 5. Identification of novel protein-interactors of central UTY isoform by yeast-two-hybrid screen. (**A**) The encoded central regions of novel UTY isoforms display intense alterations (aa insertions/deletions) due to alternative splicing. (**B**) Yeast-two-hybrid assay reveals interaction between alternatively spliced UTY-bait and aryl hydrocarbon receptor interacting protein (AIP), and also with guanine nucleotide binding protein (GNB2L1). −4, positive clones revealing protein–protein interaction for AIP and GNB2L1, growth under high stringency conditions and blue stain indicate the expression of all three reporter genes. I, third interactor. P, prey; B, bait, YPD + A, full medium with additional adenine; − trp, drop-out medium selective for transformation efficiency of baits; − leu, drop-out medium selective for the transformation efficiency of preys; (**A–B**) ovals denote aa insertions due to alternative splicing, compared to the RefSeqs encoded proteins; AS-events are indicated in the graph.

Although we could not link protein interaction precisely to the cassette exons, we have shown the central region is involved in protein interaction.

New aspects of UTY *alternative splicing to translational research*

As our observations and analyses indicated new encoded sequences, we finally checked by *in silico* analysis the medical application potential of *UTY* alternative splicing. As the UTY-peptide elicited T-cell recognition has been considered as a mechanism for immunotherapy of leukemia (Ivanov et al., 2005; Riddell et al., 2002; Warren et al., 2000), we tested, whether AS could be responsible for the generation of newly encoded UTY-specific peptide candidates. We detected new peptides with predicted high binding probabilities for MHCs. This pool included several MHCs, as exemplified by at least six different MHCI for peptides of Ex12/13b in Table 2.

Discussion

In this study, we report the discovery of highly frequent *UTY* splicing in normal and leukemic cells. Thereby, we provide the first evidence, that far more transcripts exist than previously annotated and that the diversity and complexity of *UTY* transcripts had been largely underestimated. Only rarely does AS produce very large numbers of distinct transcripts. The common relevance of these observations is not well understood. Present maximum estimations of transcript diversity yields nearly 90,000 isoforms for human basonuclin 2 gene, including 2,000 coding transcripts (Vanhoutteghem and Djian, 2007), and for *dscam* over 38,000 coding isoforms in *Drosophila melanogaster* (Schmucker et al., 2000). In our study we revealed that *UTY* has to our knowledge the greatest estimated transcript diversity coded by alternative splicing of a single gene. Its base content exceeds several hundred-fold the human genome, that contains ∼3.24 billion base pairs (Genome_Database_36v2). In terms of coding transcripts the potential diversity in *UTY* exceeds slightly the above maximum estimation of *dscam*. Our observations introduce a new dimension into the potential contribution of AS to the transcriptome.

The estimation of the splicing extend required the detailed representation of *UTY* alternative splicing. By our new model *UTY*-splicing multigraph we described and calculated in detail *UTY* splicing, and created a valuable tool that enabled us to explore for the first time the huge potential complexity of *UTY*. We are aware that yet unidentified AS-events might exist. These would even enlarge *UTY* transcript diversity.

With regard to the question of the underlying transcript generating modes, we revealed that despite its outstanding potential, the generation of *UTY* transcript diversity can be explained by a two-component model. As no universal model of transcript diversity generating modes nor a model for *UTY* exists as yet, we propose a simple model for *UTY* transcript generation. It includes our observations of AS-events' features, and interprets *UTY* transcript diversity as a result of the combination of two distinct hypothetical mechanisms. Primarily, a basic novel mechanism of random alternative

Table 2. *In Silico* Analysis Detects Novel UTY-Specific Peptides with High MHC-Binding Probabilities

Peptide	*HLA*	*Score*
Octamer	MHCI	
SP**A**K**K**KRT	HLA-B*08	18
AK**K**K**R**TSS	HLA-B*08	18
TSSQVEGL	HLA-B*37	18
D**P**NTEHVL	HLA-B*5101	23
Nonamer	MHCI	
I**T**SSQVEG**L**	HLA-A*0201	21
ITSSQVEGL	HLA-A*26	21
V**L**NHSQTP**I**	HLA-A*0201	20
P**I**LQQSLS**L**	HLA-A*0201	20
PILQQSLSL	HLA-B*1402	20
15-mer	MHCII	
EHV**L**N**H**SQ**T**PILQQS	HLA-DRB1*0701	30
TPI**L**QQSL**S**LHMITS	HLA-DRB1*0701	30

Data are exemplified for Ex12/13b. UTY-specific peptide sequences differed in one up to six amino acids from homologous UTX and other peptides of GenBank and Swiss protein human data bases. Only selected UTY-peptides, which displayed higher probabilities than their X-homologs, are shown. Epitopes were predicted by ligation strength via SYFPEITHY. Only peptides above thresholds 20 (MHC I) and 28 (MHC II) are shown. For octamers a threshold of 17 was chosen, because an experimentally verified known peptide has a score of 16 for HLA * B08 (Warren et al., 2000). Octamer, nonamer, 15-mer correspond to peptide sizes, anchor amino acids are given in bold face, auxiliary amino acids are underlined.

splicing generates a huge diversity of low-frequent transcripts. This reflects the quickly decaying exponential in our two-component mixture model. It may also explain the high number of low-frequency in-frame/PTC-events observed by us, which originate from widely dispersed locations and belong to many distinct splice classes. The second mechanism promotes inclusion of selected AS-events as implied by the second exponential of the two-component model. It may also explain our identification of high frequent AS-events. Concerning *UTY* gene expression regulation, we have observed prevalent existence of corresponding in-frame and PTC-transcripts in all three tissues studied. This coincides with the widespread occurrence of regulated unproductive splicing and translation mechanism (Lewis et al., 2003; Ni et al., 2007). In our model we therefore assume a coupling between PTC-transcripts and NMD. By proposing two combined transcript generating mechanisms and assumed coupling to NMD, our model meets the requirements of *UTY* splicing plasticity. It allows for the production of high/low degrees of *UTY* transcript diversity as well as for the coverage of its outstanding theoretical yield.

Very likely our discovery had been hitherto precluded by the complexity of the *UTY* splicing system. Its key properties such as large-sized transcripts of very diverse architecture, very complex splicing patterns containing AS-events of many different sizes, and further, high *UTY*/*UTX* sequence identity, cause severe technical constraints. This applies to common methods like EST screens, Northern blot, hybridization, splicing arrays, and multiple alignment tools. It also refers to the new high throughput techniques that focus on AS-events in short sequences. Hence, also latest global analyses on alternative splicing complexity in the human transcriptome did not uncover *UTY* diversity (Pan et al., 2008; Sultan et al., 2008; Wang et al., 2008). The complex situation also imposes severe constraints on the spectrum of methods suited for this study. Further, the primary hematopoietic cells studied here have severe limitations in cell culture. We therefore developed in this work an alternative strategy to a single-transcript approach and combined experimental and bioinformatic approaches to reveal the inherent structure of the *UTY* splicing complexity.

Our observations on *UTY* splicing have several important consequences for further research on *UTY*. First, recent research on *UTY* focuses mainly on the available RefSeqs. Therefore, our data provide new essential basis for further interdisciplinary research on *UTY*. In addition, our observations require new standards for the research on *UTY*. Our data on the new transcript diversity in *UTY* clearly demonstrate that research on gene expression now requires both careful discrimination between *UTY* and *UTX* isoforms and to take transcript diversity into account. Finally, on the protein level, we contribute new aspects for further research on the presumed functionality of *UTY*. We have identified new protein interaction of the central region in *UTY*. Further, in the context of abundant alternative splicing, we have now revealed that functionally important jmjC domain is mainly conserved in encoded proteins. This may indicate functional selection.

With regard to research on adoptive immunotherapy of leukemic relapses via UTY-specific peptides, we have provided the first evidence of abundant *UTY* splicing in leukemic and healthy tissue. Via AS, gender-specific new UTY-derived peptides could be generated. Besides this potential for new therapeutic peptides, our observations on transcript diversity provide a new starting point for research on the intriguing question of tissue restricted recognition of UTY peptides despite ubiquitous *UTY* expression. Because peptide presentation by MHCI also depends on high intracellular peptide concentrations (Yewdell, 2003), it should be now investigated if AS is involved in this process.

Conclusions

In this work we have provided the first evidence for a novel highly frequent and very complex *UTY* splicing system. In normal and leukemic cells far more *UTY* transcripts exist than previously annotated. As the great diversity imposes severe constraints on the spectrum of analytical methods, we developed in this study a new approach to *UTY* splicing diversity. A single-transcript approach allowed a new in-depth analysis that uncovered a novel *UTY* transcript architecture and new genomic organization. Further, we determined an outstanding potential for transcript diversity in *UTY* and propose a simple model mechanism for its generation. Our study provides new insights into the complexity of human alternative splicing and its potential contribution to the transcript diversity of the transcriptome. For human *UTY* our contribution substantially enlarges its molecular bases. Furthermore, we provide new views for research on UTY-based immunotherapy.

Acknowledgments

We express our gratitude to Dr. H. Adler for his continuing support and fruitful discussions. We thank I. Bigalke, M.

Leeping, and Drs. H. Schmetzer, S. Kaiser, and A. Moosmann for kindly providing samples. Further, we thank K. Schwerdtner for excellent technical assistance in DNA sequencing.

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

Supplementary Data

Supplementary data can be found at http://www.gacmunich.de/supplementary/omics.

References

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., et al. (2010). Deciphering the splicing code. Nature 465, 53–59.

Blencowe, B.J. (2006). Alternative splicing: new insights from global analyses. Cell 126, 37–47.

Breitbart, R.E., Andreadis, A., and Nadal-Ginard, B. (1987). Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annu Rev Biochem 56, 467–495.

Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., et al. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet 40, 1416–1425.

Chacko, E., and Ranganathan, S. (2009). Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse. BMC Genomics 10(Suppl 1), S5.

Chang, Y.F., Imam, J.S., and Wilkinson, M.F. (2007). The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem 76, 51–74.

Dinger, M.E., Pang, K.C., Mercer, T.R., and Mattick, J.S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol 4, e1000176.

Genome_Database_36v2 Human genome statistics: http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=36&ver=2. Homo sapiens Genome Statistics 36.

Ginalski, K., Rychlewski, L., Baker, D., and Grishin, N.V. (2004). Protein structure prediction for the male-specific region of the human Y chromosome. Proc Natl Acad Sci USA 101, 2305–2310.

Grbavec, D., Lo, R., Liu, Y., Greenfield, A., and Stifani, S. (1999). Groucho/transducin-like enhancer of split (TLE) family members interact with the yeast transcriptional co-repressor SSN6 and mammalian SSN6-related proteins: implications for evolutionary conservation of transcription repression mechanisms. Biochem J 337(Pt 1), 13–17.

Greenfield, A., Scott, D., Pennisi, D., Ehrmann, I., Ellis, P., Cooper, L., et al. (1996). An H-YDb epitope is encoded by a novel mouse Y chromosome gene. Nat Genet 14, 474–478.

Greenfield, A., Carrel, L., Pennisi, D., Philippe, C., Quaderi, N., Siggers, P., et al. (1998). The UTX gene escapes X inactivation in mice and humans. Hum Mol Genet 7, 737–742.

Heber, S., Alekseyev, M., Sze, S.H., Tang, H., and Pevzner, P.A. (2002). Splicing graphs and EST assembly problem. Bioinformatics 18(Suppl 1), S181–S188.

Hong, S., Cho, Y.W., Yu, L.R., Yu, H., Veenstra, T.D., and Ge, K. (2007). Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. Proc Natl Acad Sci USA 104, 18439–18444.

Ivanov, R., Hol, S., Aarts, T., Hagenbeek, A., Slager, E.H., and Ebeling, S. (2005). UTY-specific TCR-transfer generates potential graft-versus-leukaemia effector T cells. Br J Haematol 129, 392–402.

Kolb, H.J. (2008). Graft-versus-leukemia effects of transplantation and donor lymphocytes. Blood 112, 4371–4383.

Lahn, B.T., and Page, D.C. (1997). Functional coherence of the human Y chromosome. Science 278, 675–680.

Lan, F., Bayliss, P.E., Rinn, J.L., Whetstine, J.R., Wang, J.K., Chen, S., et al. (2007). A histone H3 lysine 27 demethylase regulates animal posterior development. Nature 449, 689–694.

Leipzig, J., Pevzner, P., and Heber, S. (2004). The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. Nucleic Acids Res 32, 3977–3983.

Lejeune, F., and Maquat, L.E. (2005). Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. Curr Opin Cell Biol 17, 309–315.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci USA 100, 189–192.

Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., et al. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev 21, 708–718.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40, 1413–1415.

Rammensee, H., Bachmann, J., Emmerich, N.P., Bachor, O.A., and Stevanovic, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50, 213–219.

Riddell, S.R., Murata, M., Bryant, S., and Warren, E.H. (2002). Minor histocompatibility antigens-targets of graft versus leukemia responses. Int J Hematol 76(Suppl 2), 155–161.

Rozen, S., Marszalek, J.D., Alagappan, R.K., Skaletsky, H., and Page, D.C. (2009). Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection. Am J Hum Genet 85, 923–928.

Sammeth, M., Foissac, S., and Guigo, R. (2008). A general definition and nomenclature for alternative splicing events. PLoS Comput Biol 4, e1000147.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., et al. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 101, 671–684.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423, 825–837.

Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., et al. (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci USA 99, 16899–16903.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321, 956–960.

Vanhoutteghem, A., and Djian, P. (2007). The human basonuclin 2 gene has the potential to generate nearly 90,000 mRNA isoforms encoding over 2000 different proteins. Genomics 89, 44–58.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–476.
Warren, E.H., Gavin, M.A., Simpson, E., Chandler, P., Page, D.C., Disteche, C., et al. (2000). The human UTY gene encodes a novel HLA-B8-restricted H-Y antigen. J Immunol 164, 2807–2814.
Yewdell, J.W. (2003). Immunology. Hide and seek in the peptidome. Science 301, 1334–1335.

Address correspondence to:
Jerzy Adamski
Helmholtz Zentrum München
German Research Center for Environmental Health
Institute of Experimental Genetics
Genome Analysis Center
Ingolstädter Landstraße 1
85764 Neuherberg, Germany

E-mail: adamski@helmholtz-muenchen.de