Molecular Pharmacology Supplemental Materials

Activation of Cryptic Donor Splice Sites Within the UGT1A First-Exon Region Generates Variant Transcripts That Encode UGT1A Proteins With Truncated Aglycone-binding Domains Dong Gui Hu, Shashikanth Marri, Julie-Ann Hulin, Radwan Ansaar, Peter I. Mackenzie, Ross A. McKinnon, and Robyn Meech

Legends of Supplemental Figures

Supplemental Fig. 1. Sequences of variant UGT1A first exons. Shown are the sequences for eight variant UGT1A first exons that are generated using cryptic donor sites within the UGT1A first-exon region, including *1A1E1v* (A), *1A3E1v1* (B), *1A3E1v2* (C), *1A4E1v* (D), *1A5E1v* (E), *1A8E1v* (F), *1A9E1v* (G), *1A10E1v* (H). The nucleotide sequences of variant first exons (BLUE) are positioned at the right according to the human GRCh38/hg38 genome assembly. The start ATG codons and the dinucleotide GT splice signals of the novel cryptic donor splice sites are also indicated.

Supplemental Fig. 2. Predicted sequences for variant UGT1A proteins. Shown are the sequences for eight predicted UGT1A variant proteins, including 1A1_in1 (A), 1A3_in3 (B), 1A3_in4 (C), 1A4_in4 (D), 1A5_in1 (E), 1A8_in2 (F), 1A9_in2 (G), and 1A10_in7 (H). The sequences encoded by the first exons and exons 2-5 are indicated in BLUE and RED, respectively. 1A3_in3 has a novel 77-aa C-terminal peptide (GREEN) but lacks the sequence encoded by exons 2-5.

Supplemental Figure 3. Sequence reads for variant UGT1A transcripts. Shown are the sequence reads (100 nucleotides) for variant transcripts 1A4_n4 (A), 1A8_n2 (B) and 1A9_n2 (C) identified from the UGT-enriched CaptureSeq datasets (GSE80463) using transcript-specific probes and the Sequence Read Archive (SRA) platform. The transcript-specific splice

junctions are indicated by a vertical RED line. The nucleotide positions of the 3' ends of the three variant first exons (1A4E1v, 1A8E1v, 1A9E1v) are also indicated.

Supplemental Fig. 4. Expression of variant transcripts 1A8_n2 and 1A9_n2 in normal and drug-metabolizing tissues. Using transcript-specific probes and the Sequence Read Archive (SRA) platform, the sequence reads of canonical (1A8_v1, 1A9_v1) and variant (1A8_n2, 1A9_n2) transcripts were identified in fifteen CaptureSeq samples (GSE80463) generated from normal and cancerous drug-metabolizing tissues as indicated. The number of sequence reads for each transcript was normalized using the number of total sequence reads in the same sample and then presented as the relative reads of this transcript per 10⁹ reads of the total sequence reads. Shown are the expression level of 1A8_n2 (A), 1A9_n2 (C), and the expression ratio for 1A8_n2/1A8_v1 (B), or 1A9_n2/1A9_v1 (D) in fifteen CaptureSeq samples.

Supplemental Figure 5. Variant UGT1A transcripts and proteins. (A) RT-PCR was conducted using cDNA samples of colorectal cancer HT-29 cells and primers to clone the full coding sequence of canonical UGT1A8 mRNA (1A8_v1) or UGT1A10 mRNA (1A10_v1). The resultant amplicons were run on an ethidium-bromide-stained agarose gel and imaged using UV-illumination. (B) HEK293T cells were transfected with constructs expressing no UGT protein (control), wildtype (1A8_i1) and variant (1A8_i3) UGT1A8 proteins alone and in combination (1A8_i1 + 1A8_i3) as indicated. Lysates of transfected cells were subjected to standard Western blotting assays using a pan-UGT1A antibody and imaged using chemiluminescent agents as described in *Materials and Methods*. (C) Glucuronidation of HEK293T lysates transfected with vectors expression 1A8_i1 or 1A8_i3 alone and in combination were conducted using HPLC assays. The activity of 4MU glucuronidation was

normalized to the band intensities of western blots obtained using equal amounts of HEK293T lysates of the same samples used for 4MU activity assays as described in *Materials and Methods*. Shown is the mean plus SD of 4MU-glucuronidation activity of $1A8_i1/_i3$ -transfected cells normalized to that of $1A8_i1$ -transfected cells (set as a value of 100%) from three independent experiments. Student's t-test, p < 0.05 is considered statistically significant.

Supplemental Fig. 6. **Sequencing results identified the 1A8_n2 transcript in HT-29 cells**. The 1A8_n2 transcript has a variant first exon (*1A8E1v*) and common UGT1A exons 2-5. Shown are the sequencing chromatograms of a cloned pEF_IRESpuro6 construct from the HT-29 cell line that contains the 1A8_n2 transcript-specific splice junction (1A8E1v/1AE2) and all other four UGT1A common splice junctions [1AE2/1AE3, 1AE3/1AE4, 1AE4/1AE5]. All splice junctions are indicated by a vertical line. nt: nucleotide; AA: amino acid.

Supplemental Fig. 7. Sequencing results identified the 1A10_n7 transcript in HT-29 cells.

The 1A10_n7 transcript has a variant first exon (*1A10E1v*) and common UGT1A exons 2-5. Shown are the sequencing chromatograms of a cloned pEF_IRESpuro6 construct from the HT-29 cell line that contain the 1A10_n7 transcript-specific splice junction (1A10E1v/1AE2) and all UGT1A common splice junctions [1AE2/1AE3, 1AE3/1AE4, 1AE4/1AE5]. All splice junctions are indicated by a vertical line. nt: nucleotide; AA: amino acid

Supplemental Fig. 8. Variant transcripts UGT1A8_n2 and UGT1A10_n7 generated using novel cryptic donor splice sites within UGT1A first exons in colorectal cancer HT-29 cells. (A) Shown are the exon structures of the UGT1A8 pre-mRNA (Aa), mRNA (1A8_v1) (Ab), and variant transcript (1A8_n2). Pre-mRNA splicing using the 1A8 canonical and cryptic donor splice sites generates 1A8 mRNA (1A8_v1) (Ab) and variant 1A8_n2 (Ac), respectively. (B) Shown are the exon structures of the UGT1A10 pre-mRNA (Ba), mRNA (1A10_v1) (Bb), and variant transcript (1A10_n7). Pre-mRNA splicing using the 1A10 canonical and cryptic donor splice sites generates 1A10 mRNA (1A10_v1) (Bb) and variant 1A10_n7 (Bc), respectively. The sequencing results covering the 1A8_n2 (Ad) or 1A10_n7 (Bd) novel splice junction are also shown. The donor and acceptor splice signal dinucleotides are indicated GT and AG, respectively.













| | | star | t ATG codon | | | | | |
|---|--|--|--|--|---|--|--|--|
| F | catgtattec tttttttta <u>GIGCCIGIAG</u> TGCTGGCTCG | tqttcttatq tqacaqgata TPCTICCGCC GGCTGCAGTT | agtaaatcat aatacacgcc FACYGTAYCA CTCTCA2GGC | tqqcaqtqaq ctctattqqq TAGCAGCYTA TCGCACAGGG | lqlqatttt gtcagqtttt GAATCCCAGC TGGACCAGCC | 233617582 233617632 233617682 233617782 | | |
| | GGGAAGCTGC GTCGGTGGTG TGCCAGAGGT | AUGIGUTUU IGGTAGIGCC GAGAAACTIA GAGTIGGCAA | CATGGATGGG TCCTCAGGGG CTGGGAAAAT | AGICACIGGI GCATGAGGIG CACIGAATIG | TCACCATGCA GTIGTAGTCA CACAGTGAAG | 23361782 233617832 233617882 233617932 | | |
| | ACTIACTCAA TITCGCCGAT LLCLGAGLLC | GCTCAATGGA alccaalggl | AAGCACAAgt LUULULaacu | acgaagtttg Lattitttc | ttttctctat gcallgcagg | 233617982 233618032 233618082 | | |
| | GT: donor splice site | | | | | | | |
| | | end of | LASE1v | | | | | |

start of UGT1A9 exon 1 (NM_021027.3) start ATG codon

| | | end of IA9E1v | | | | | | | |
|-----|------------|---------------|------------|-----------------------|------------|-----------|--|--|--|
| | | | | GT: donor splice site | | | | | |
| | | | | | | | | | |
| | tatattetet | allaalgggl | teatacaate | acalttilga | citatitit | 233672297 | | | |
| | GGAGTTCAAG | GCTITTGCCC | ATGCTCAATG | GAAAGCACAA | gtacgaagta | 233672247 | | | |
| | TGCACAGTGA | AGACTTATTC | AACTTCATAT | ACCCTGGAGG | ATCTGGACCG | 233672197 | | | |
| | TGGTTCTACT | CATOCCAGAG | GTCAGTTOCC | AACTGOCAAC | ATCACTGAAT | 233672147 | | | |
| U U | GTTCACCATG | AGGICGGIGG | TGGAGAAACT | CATTCTCAGG | GGGCATGAGG | 233672097 | | | |
| G | TTTGCCGAGG | CAGGGAAGCT | ACTEGTAGTE | CCCATGGATG | GGAGCCACTG | 233672047 | | | |
| | CGTCGACCAC | CCCCCTTCCT | CTATGTGTGT | CICICCICT | GACCTGTCGC | 233671997 | | | |
| | AGATTOCCAG | CIECTICIC | TCAGCTGCAG | TTCTCTGATG | GCTTGCACAG | 233671947 | | | |
| | ggleagglil | lgigelggia | llicicecae | clacigiate | alaggagett | 233671897 | | | |
| | cagtgactga | ttttttttt | atgaaaggat | aaaaacaege | cctctattgg | 233671847 | | | |
| | | | | | | | | | |

| | | start ATG codon | | | | | | | | |
|----|------------|-----------------|------------|------------|------------|-----------|--|--|--|--|
| | tgttatcgtt | cttatgagta | aatcattggc | agtgagtgtg | atttttttt | 233636402 | | | | |
| | ttttatgaaa | ggataaatac | acgeceteta | ttggggtcag | gttttgtgcc | 233636452 | | | | |
| | TGTACITCIT | CCGCCTACTG | TATCATAGCA | GCTTAGAATC | CCAGCIGCIG | 233636502 | | | | |
| Н | GCTCGGGCTG | CAGTTCTCTC | ATGGCTCGCG | CAGGGTGGAC | CAGCCCCGTT | 233636552 | | | | |
| •• | CCTTTATGTG | TGTGTCTACT | GCTGACCTGT | GGCTTTGCCG | AGGCAGGGAA | 233636602 | | | | |
| | GCTGCTGGTA | GTGCCCATGG | ATGGGAGTCA | CTGGTTCACC | ATGCAGTOGG | 233636652 | | | | |
| | TGGTGGAGAA | ACTTATCCTC | AGGGGGCATG | AGGTGGTTGT | AGTCATGCCA | 233636702 | | | | |
| | GAGgtgagtt | ggcaactgga | aagatcactg | aattgcacag | tgaagactta | 233636752 | | | | |
| | ctcaacctcg | tacactetqg | aagatcagaa | cogggaatte | atggttttcg | 233636802 | | | | |
| | | | | | | | | | | |
| | GT: dono | or splice site | | | | | | | | |

end of 1A10E1v

Protein predicted from transcript 1A1_n1 (487 aa)

MAVESQGGRPLVLGLLLCVLGPVVSHAGKILLIPVDGSHWLSMLGAIQQLQQRGHEIVVLAPDASLYIRDGAFYTLKTYPVPFQREDVKE SFVSLGHNVFENDSFLQRVIKTYKKIKKDSAMLLSGCSHLLHNKELMASLAESSFDVMLTDPFLPCSPIVAQYLSLPTVFFLHALPCSLE FEATQCPNPFSYVPRPLSSHSDHMTFLQRVKNMLIAFSQNFLCDVVYSPYATLASEFLQREEFEAYINASGEHGIVVFSLGSMVSEIPEK KAMAIADALGKIPOTVLWRYTGTRPSNLANNTILVKWLPONDLLGHPMTRAFITHAGSHGVYESICNGVPMVMMPLFCDOMDNAKRMETK

A FEATQCENEFSYVERPLSSHSDHMTFLQRVKNMLIAFSQMFLCDVVYSPYATLASEFLQREEFEAYINASGEHGIVVFSLGSMVSEIPEK KAMAIADALGKIPQTVLWRYTGTRPSNLANNTILVKWLPQNDLLGHPMTRAFITHAGSHGVYESICNGVPMVMMPLFGDQMDNAKRMETK GAGVTLNVLEMTSEDLENALKAVINDKSYKENIMRLSSLHKDRPVEPLDLAVFWVEFVMRHKGAPHLRPAAHDLTWYQYHSLDVIGFLLA VVLTVAFITFKCCAYGYRKCLGKKGRVKKAHKSKTH

Protein predicted from transcript 1A3_n3 (239 aa)

B MATGLQVPLPWLATGLLLLLSVQPWAESGKVLVVPIDGSHWLSMREVLRELHARGHQAVVLTPEVNMHIKEENFFTLTTYAISWTQDEFD RHVLGHTQLYFETEHFLKKFFRSMAMLNNMSLVYHRSCVELLHNEALIRHLNATSFDVVLTDPVNLCAAVLAKNLKPTLMLLENMELWFS LWDQWSQKFQRRKLWQLLMLWAKSLRQSCGGTLEPDHRILRTTRYLLSGYPKTICLVTR

Predicted protein from transcript 1A3_n4 (309 aa)

MATGLQVPLPWLATGLLLLLSVQPWAESGKVLVVPIDGSHWLSMREVLRELHARGHQAVVLTPEEFEAYINASGEHGIVVFSLGSMVSEI PEKKAMAIADALGKIPQTVLWRYTGTRPSNLANNTILVKWLPQNDLLGHPMTRAFITHAGSHGVYESICNGVPMVMMPLFGDQMDNAKRM ETKGAGVTLNVLEMTSEDLENALKAVINDKSYKENIMRLSSLHKDRPVEPLDLAVFWVEFVMRHKGAPHLRPAAHDLTWYQYHSLDVIGF

LLAVVLTVAFITFKCCAYGYRKCLGKKGRVKKAHKSKTH

С

Predicted protein from transcript 1A4_n4 (309 aa)

D MARGLQVPLPRLATGLLLLLSVQPWAESGKVLVVPTDGSPWLSMREALRELHARGHQAVVLTPÆFEAYINASGEHGIVVFSLGSMVSEI PEKKAMAIADALGKIPQTVLWRYTGTRPSNLANNTILVKWLPQNDLLGHPMTRAFITHAGSHGVYESICNGVPMVMMPLFGDQMDNAKRM ETKGAGVTLNVLEMTSEDLENALKAVINDKSYKENIMRLSSLHKDRPVEPLDLAVFWVEFVMRHKGAPHLRPAAHDLTWYQYHSLDVIGF LLAVVLTVAFITFKCCAYGYRKCLGKKGRVKKAHKSKTH

Predicted protein from transcript 1A5_n1 (309 aa)

E MATGLQVPLPQLATGLLLLLSVQPWAESGKVLVVPTDGSHWLSMREALRDLHARGHQVVVLTLEEFEAYINASGEHGIVVFSLGSMVSEI PEKKAMAIADALGKIPQTVLWRYTGTRPSNLANNTILVKWLPQNDLLGHPMTRAFITHAGSHGVYESICNGVPMVMMPLFGDQMDNAKRM ETKGAGVTLNVLEMTSEDLENALKAVINDKSYKENIMRLSSLHKDRPVEPLDLAVFWVEFVMRHKGAPHLRPAAHDLTWYQYHSLDVIGF LLAVVLTVAFITFKCCAYGYRKCLGKKGRVKKAHKSKTH

Predicted proetin (termed 1A8_i3) from transcript 1A8_n2 (346 aa)

 $\label{eq:martgwtspiplcvsllltcgfaeagkllvvpmdgshwftmqsvveklilrghevvvvmpevswqlgkslnctvktystsytledldref for an and a state of the state$

Predicted protein from transacript 1A9 n2 (346 aa)

G MACTGWTSPLPLCVCLLLTCGFAEAGKLLVVPMDGSHWFTMRSVVEKLILRGHEVVVVMPEVSWQLGRSLNCTVKTYSTSYTLEDLDREF KAFAHAQWKAQEFEAYINASGEHGIVVFSLGSMVSEIPEKKAMAIADALGKIPQTVLWRYTGTRPSNLANNTILVKWLPQNDLLGHPMTR AFITHAGSHGVYESICNGVPMVMMPLFGDQMDNAKRMETKGAGVTLNVLEMTSEDLENALKAVINDKSYKENIMRLSSLHKDRPVEPLDL AVFWVEFVMRHKGAPHLRPAAHDLTWYQYHSLDVIGFLLAVVLTVAFITFKCCAYGYRKCLGKKGRVKKAHKSKTH

Predicted protein(termed 1A10_i3)from transcript 1A10_n7(306 aa)

MARAGWTSPVPLCVCLLLTCGFAEAGKLLVVPMDGSHWFTMQSVVEKLILRGHEVVVVMPEEFEAYINASGEHGIVVFSLGSMVSEIPEK KAMAIADALGKIPQTVLWRYTGTRPSNLANNTILVKWLPQNDLLGHPMTRAFITHAGSHGVYESICNGVPMVMMPLFGDQMDNAKRMETK GAGVTLNVLEMTSEDLENALKAVINDKSYKENIMRLSSLHKDRPVEPLDLAVFWVEFVMRHKGAPHLRPAAHDLTWYQYHSLDVIGFLLA VVLTVAFITFKCCAYGYRKCLGKKGRVKKAHKSKTH



Supplemental Fig. 3.



Supplemental Fig. 4.





В

С



Supplemental Fig. 5

A







| | NCBI RefSeq | NCBI RefSeq | Experimental validation of transcript expression | | | | | |
|---|-------------|--------------------|--|----------------------|---|---|-------------------------------------|---|
| | (mRNAs) | (proteins) | Predicted proteins (aa) | RT-PCR | Cloning | CaptureSeq with UGT- enrichment | RNAseq without UGT enrichment | Western blotting |
| UGT1A1 UGT1A1_v1 UGT1A1_v2 UGT1A1_v3 UGT1A1_n1* UGT1A1_n2 UGT1A1_n3 | NM_000463.3 | NP_000454 (533 aa) | 487 | イイ | イイン | | イ イ イ イ | イイン |
| UGT1A2P UGT1A2P_n1 UGT1A2P_n2 UGT1A2P_n3 UGT1A2P_n4 UGT1A2P_n5 UGT1A2P_n6 UGT1A2P_n7 UGT1A2P_n8 UGT1A2P_n9 UGT1A2P_n10 UGT1A2P_n11 | | | | | | インマン | | |
| UGT1A3_V1 UGT1A3_V2 UGT1A3_V3 UGT1A3_n1 UGT1A3_n2 UGT1A3_n3 UGT1A3_n4 | NM_019093.4 | NP_061966 (534 aa) | 239 309 | イイ | イイイ | イイン | $\frac{1}{\sqrt{2}}$ | イイ |
| UGT1A4_v1 UGT1A4_v2 UGT1A4_v2 UGT1A4_v3 UGT1A4_n1 UGT1A4_n2 UGT1A4_n3 UGT1A4_n4 | NM_007120.3 | NP_009051 (534 aa) | 309 | イイ | $\begin{array}{c} \checkmark \\ \checkmark \\ \checkmark \\ \checkmark \end{array}$ | インシン | \checkmark | $\begin{array}{c} \checkmark \\ \checkmark \\ \checkmark \end{array}$ |
| UGT1A5_v1 UGT1A5_v1 UGT1A5_v2 UGT1A5_v3 UGT1A5_n1 | NM_019078.2 | NP_061951 (534 aa) | 309 | イイ | $\sqrt[]{}$ | イイ | \checkmark | イン |
| UGT1A6_v1 UGT1A6_v2 UGT1A6_v2 UGT1A6_n3 UGT1A6_n3 UGT1A6_n4 | NM_001072.4 | NP_001063 (532 aa) | | イイ | イイ | イ イ イ イ イ イ イ イ イ イ イ イ イ イ イ イ イ イ イ | | イイ |
| UGT1A7_v1 UGT1A7_v1 UGT1A7_v2 UGT1A7_v3 UGT1A7_n1 | NM_019077.3 | NP_061950 (530 aa) | | \sim \sim \sim | $\frac{1}{\sqrt{2}}$ | インシン | | $\sqrt[n]{\sqrt{1}}$ |

Supplemental Table 7: Known UGT1A transcripts and novel UGT1A variant transcripts identified in this study that are named using the current UGT1A nomenclature (Tourancheau A et al 2016)

| UGT1A9 UGT1A9_v1 UGT1A9_v2 UGT1A9_v3 UGT1A9_n1 UGT1A9_n2 | NM_021027.3 | NP_066307 (530 aa) | 346 | イン | イ イ イ | インシン | V | $\sqrt[]{}$ |
|---|-------------|--------------------|-----|--------------|----------------------|------------------|--------------|------------------------------|
| UGT1A10 UGT1A10_v1 UGT1A10_v2 UGT1A10_v3 UGT1A10_n4 UGT1A10_n5 | NM_019075.4 | NP_061948 (530 aa) | | イン | $\frac{1}{\sqrt{2}}$ | インシン | | \checkmark \checkmark |
| UGT1A10_n6 UGT1A10_n7 | | | 306 | \checkmark | \checkmark | | | |
| UGT1A8 UGT1A8_v1 UGT1A8_v2 UGT1A8_v3 UGT1A8_n1 UGT1A8_n2 | NM_019076.5 | NP_061949 (530 aa) | 346 | イント | イ イ イ | イ イ イ イ | \checkmark | |
| Other UGT1A UGT1A_n1 UGT1A_n2 UGT1A_n3 UGT1A_n4 UGT1A_n5 UGT1A_n6 UGT1A_n7 UGT1A_n7 UGT1A_n8 UGT1A_n9 UGT1A_n10 UGT1A_n11 UGT1A_n12 UGT1A_n13 UGT1A_n14 UGT1A_n15 UGT1A_n16 UGT1A_n17 UGT1A_n18 UGT1A_n20 UGT1A_n21 UGT1A_n22 | | | | | | | ~~~~~~~~~~~ | |

UGT1A transcripts highlighted in bold are reported in the present study and all others are reported by Tourancheau et al 2016 and several other studies (e.g. Levesque E et al 2007 and Giard H et al 2007). Also listed are the evidence for the synthesis of these transcripts in human tissues and cell lines from one or multiple experimental approaches, such as RT-PCR, Cloning, CaptureSeq, RNA-seq and Western Blotting assays. * This variant was also previously described in Tourancheau et al 2016.