Structure-based and evolutionary analyses of alternative splicing patterns in carbonic anhydrases

Ramesh Karki
Master's Thesis
MDP Programme in Digital Health and Life Sciences
Department of Future Technologies
University of Turku
October 2019

UNIVERSITY OF TURKU
Department of Future Technologies/Faculty of Science and Engineering
Ramesh Karki: Structure-based and evolutionary analyses of
alternative splicing patterns in carbonic anhydrases
Pro gradu –thesis, 73 p., 23 p. appendices
Bioinformatics
October 2019

Abstract

Alternative splicing is a biological process which generates multiple distinct mature mRNAs from a single primary transcript. Several isoforms of alpha carbonic anhydrase exist due to this phenomenon which are categorized into several subgroups (cytoplasmic, mitochondrial, secreted, membrane-associated/extracellular, CA-related proteins). The research investigates and analyzes the missing/extra exon pattern in these isoforms. The transcript information of each CA gene of human and mouse was extracted using ensemble database. Protein-coding isoforms were then classified based on the presence of signal peptide, complete catalytic domain, active site, metal-ion binding site and TM helix. Ensembl database flags (such as APPRIS, GENCODE basic, TSL) and catalytic domain structure was observed to find out principle isoforms for each CAs. EST and cDNA evidence for missing/extra exon was examined for each principle isoforms using ensemble and Genomic browser. Structural feasibility of the isoforms with missing/extra exons was studied. The result shows that human and mouse extracellular and secretory CAs have extra/missing exons in various exon positions. Some of these exons are within catalytic domain and some in the linker region after the catalytic domain and before the transmembrane helix. Transcripts like human CAXII and mouse Car XIV have missing 9^{th} exon between catalytic and transmembrane domains which implies the possibility of function-altering variant. The CA XII transcript with missing 9th exon (11aa) seems to be common in astrocytomas (type of cancer that forms in brain or spinal cord). There are no evidence of missing/extra exons in cytoplasmic and mitochondrial CAs. Ensembl shows no evidence of missing/extra exons in extracellular and secreted CAs of zebrafish and cow (not enough data for these species). Homologous missing/extra exons in human and mouse in specific groups of CAs (i.e. extracellular and secreted) confirm the biological relevance of the pattern to some extent. Further studies are needed to fully understand and confirm evolutionary significance of this pattern.

Keywords
Alternative Splicing, Principle Isoform, Carbonic Anhydrase, Exon, Catalytic Domain, Membrane-associated CAs, Secreted CA, Protein-coding

Acknowledgements

This research has been possible because of continuous support and guidance from my thesis supervisor Dr. Martti Tolvanen. I am very thankful to Martti for his critical insights and effective supervision during the project.

I would like to thank my fellow students for their feedback and co-operation. I would also like to expess my gratitude to Dr. Juho Heimonen for reviewing my thesis. Last but not the least, I am deeply indebted to my parents for providing me good atmosphere for my growth and progress. Thank you for supporting and believing in me at all times.

Abbreviations

CA	Carbonic anhydrase
PDB	Protein Data Bank
Zn	Zinc
His	Histidine
OH	Hydroxide
ORF	Open Reading Frame
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
cDNA	Complimentary DNA
mRNA	messenger RNA
aa	Amino Acid
bp	Base Pair
hCA	Human Carbonic Anhydrase
TSL	Transcript Support Level
ncRNA	Non-Coding RNA
CCDS	Consensus Coding Sequence Project
EST	Expressed Sequence Tag

Table of Contents

1 Introduction

Carbonic Anhydrase

Carbonic anhydrase enzymes are metalloproteins that contain zinc and plays a major role in reversible conversion of carbon dioxide to bicarbonate and release proton. They are essential for several physiological and pathophysiological function (Imtaiyaz Hassan et al., 2013).

$$CO2 + H2O \rightleftharpoons HCO3- + H+$$

16 different CA isozymes exists in mammals based on their sequence, bio- chemical properties and subcellular location (A. J. J. Esbaugh and Tufts, 2006).

The α-CA gene family contains 3 subfamilies. The cytoplasmic CAs are found in the cytoplasm of various tissues which includes CA I, II, III, VII and XIII. Another group of isozymes consists of CA IV, IX, XII, XIV and XV which are membrane-associated CAs. The secreted CAs contains CA VI and mitochondrial CAs contain CAVA and CA VB (A. J. Esbaugh and Tufts, 2006). Large cone-shaped cavity forms the active site of CA where Zn2+ ion resides at the bottom.

Structural analysis of CA isozymes I, II, III, IV,V, XII, XIII, and XIV shows a high degree of structural similarity (Imtaiyaz Hassan et al., 2013).

Table 1 CO2 hydration activity and organ/tissue distribution of the 12 human catalytically active alpha CA isozymes (Supuran and Simone, 2015)

Enzyme	Catalytic activity	Organ/tissue distribution
CA I	Low	Erythrocytes, gastrointestinal tract, eye
CA II	High	Erythrocytes, eye, gastrointestinal tract, bone osteoclasts, kidney, lung, testis, brain
CA III	Very low	Skeletal muscle, adipocytes
CA IV	Medium	Kidney, lung, pancreas, brain capillaries, colon, heart muscle, eye
CA VA	Low	Liver
CA VB	High	Heart and skeletal muscle, pancreas, kidney, spinal cord, gastrointestinal tract
CA VI	Low	Salivary and mammary glands
CA VII	High	Central nervous system
CA IX	High	Tumors, gastrointestinal mucosa
CA XII	Low	Renal, intestinal, reproductive epithelia, eye, tumors
CA XIII	Low	Kidney, brain, lung, gut, reproductive tract
CA XIV	Low	Brain, liver, eye, skeletal muscle

Human CAs' Catalytic Features

Carbonic anhydrase catalyzes reaction in two-steps. Bicarbonate ion coordinated to zinc is formed in first step due to nucleophilic attack of hydroxide ion on carbon dioxide molecule. Catalytically active form of enzyme is formed through a proton transfer reaction from zinc-coordinated water molecule to the external buffer in second step (Supuran and Simone, 2015).

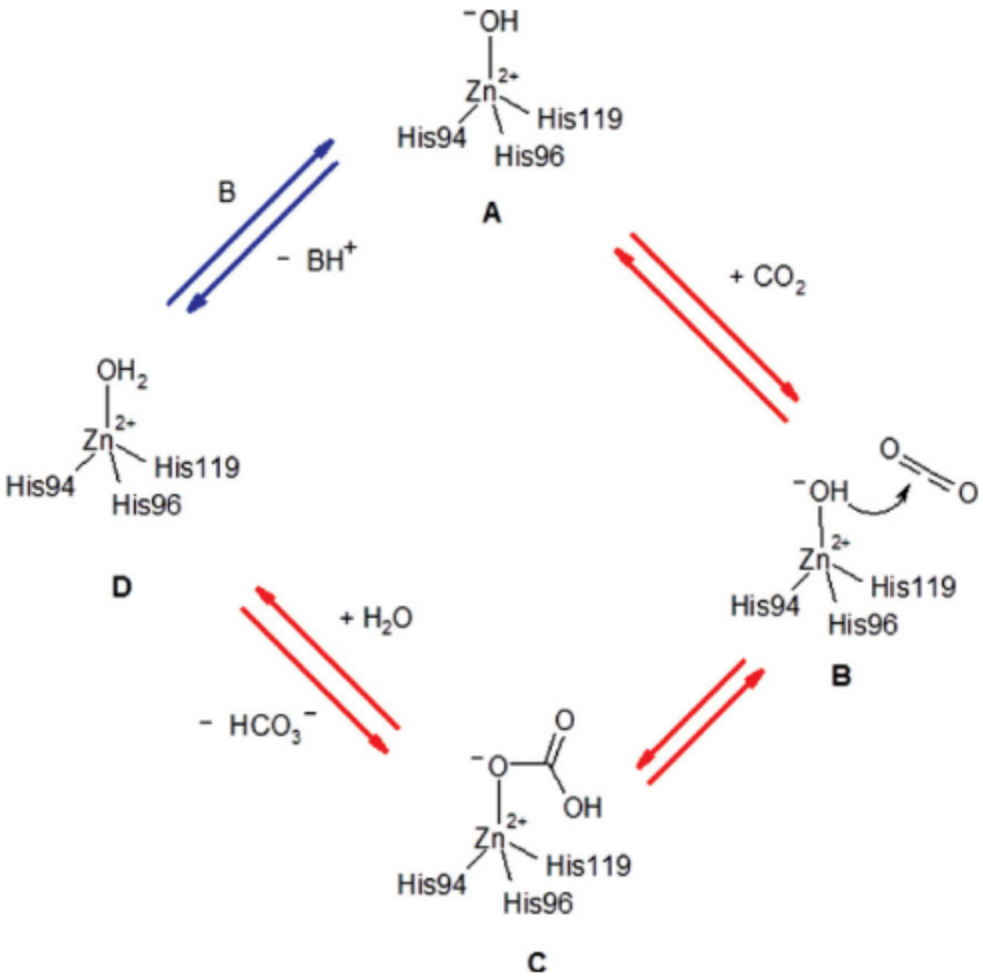


Figure 1 Catalytic mechanism of a-CAs. Light gray arrows indicate reactions defining the first step of the catalytic mechanism, while the black arrows specify the reaction of the second step (Supuran and Simone, 2015).

2 Literature Review

2.1 Types of human carbonic anhydrases

2.1.1 CA I

CA I has 260 residues where its principal transcript contains 261 residues. To date, there are 31 crystallographic structures available of CA I. It is mostly located in erythrocytes including kidneys, gastrointestinal tract, lungs, brain and eyes. It is less active than CA II isoform (KANNAN et al., 1984; Kumar and Kannan, 1994).

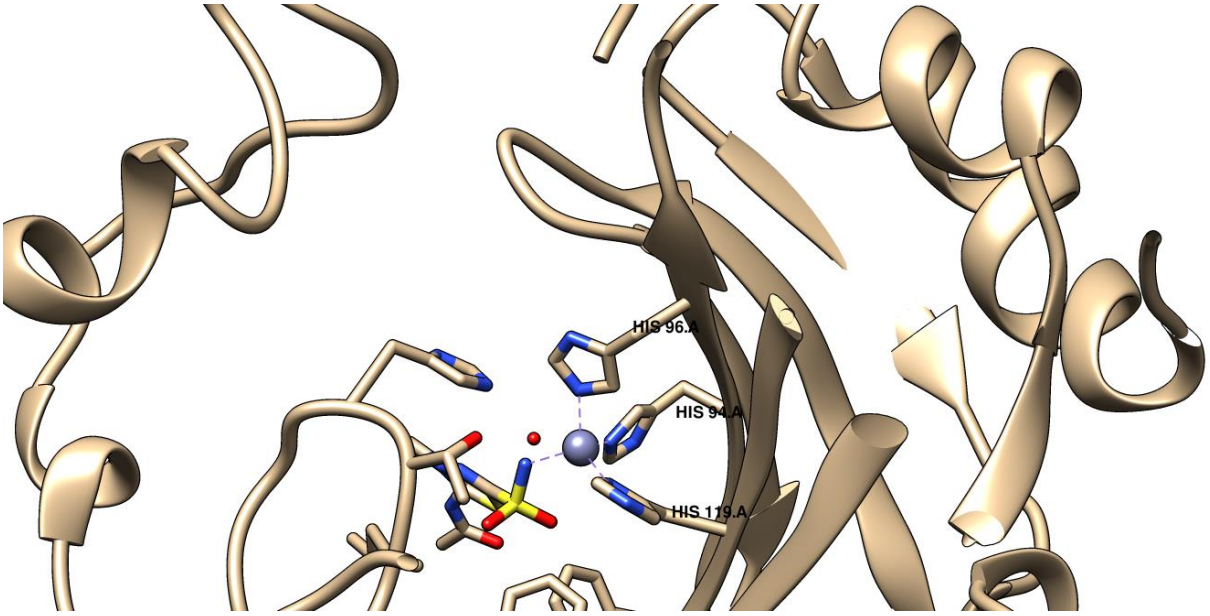


Figure 2 CA I catalytic active site (PDB ID: 1BZM) depiction using chimera. Histidine residues (94, 96 and 119) and water molecule (small red sphere) coordinates with zinc cofactor (big sphere)

Histine64 is conserved in highly active isoforms of human carbonic anhydrases which acts as proton shuttle thereby increasing deprotonation rate of zinc-bound water molecule. CA I is closely identical with other cytosolic isoforms II, III, VII, and XIII due to its high number of conserved residues and 3D structure (Chakravarty and Kannan, 1994; Ferraroni et al., 2002).

2.1.2 CA II

Human CA II contains 259 amino acid residues where its principal transcript contains 260aa. This enzyme helps in bone resorption and osteoclast differentiation (Kim et al., 2002). It regulates fluid secretion into the anterior chamber of the eye and also balances pH in duodenal upper villous epithelium (Briganti et al., 1999; Eriksson et al., 1988).

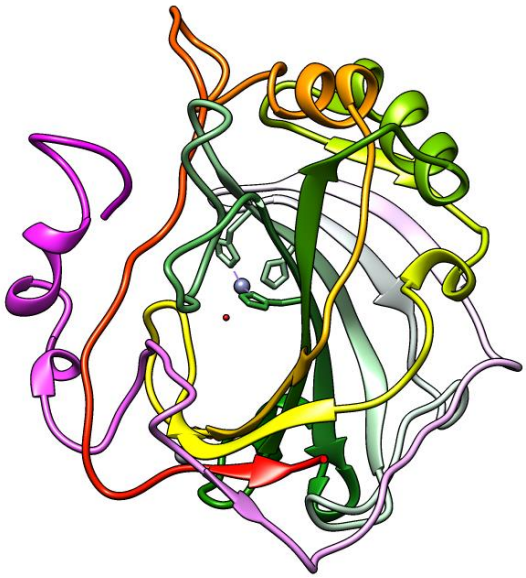


Figure 3 Structure of CA II (PDB ID: 12CA) as viewed in chimera

This isoform of carbonic anhydrase is most studied and forms a basis for better understanding of other CAs. CA II has high catalytic activity and has highest expression level in colonic mucosa (Alvarez et al., 2005; Kim et al., 2002).

2.1.3 CA III

CAIII is a monomeric cytosolic protein that is most abundant in red fibers of skeletal muscle, liver, and adipose tissue (Di Fiore et al., 2009). It is comparatively very weak in catalysis nearly 300 fold less than that of CA II.

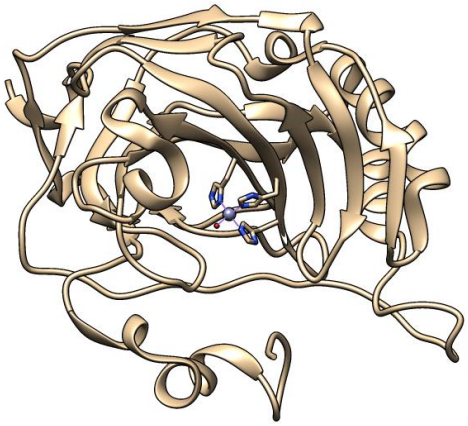


Figure 4 Structure of carbonic anhydrase III (PDB ID: 1Z93) as viewed in Chimera

Principal isoform of this enzyme in human has catalytic domain residues starting from position 3-259 with residues 127 acting as active site. Mutagenesis of residue F to L enhances activity by at least 10 fold (Crocetti et al., 2009a).

Peptide sequence of principal isoform of CAIII (black and blue color represents alternating exons with red representing splice site):

MAKEWGYASHNGPDHWHELFPNAKGENQSPVELHTKDIRHDPSLQPWSVSYDGGSAKTIL

NNGKTCRVVFDDTYDRSMLRGGPLPGPYRLRQFHLHWGSSDDHGSEHTVDGVKYAAELHL

VHWNPKYNTFKEALKQRDGIAVIGIFLKIGHENGEFQIFLDALDKIKTKGKEAPFTKFDP

SCLFPACRDYWTYQGSFTTPPCEECIVWLLLKEPMTVSSDQMAKLRSLLSSAENEPPVPL

VSNWRPPQPINNRVVRASFK

2.1.4 CA IV

CA IV has 10 PDB entries derived from x-ray crystallography. The gene for this enzyme is located in chromosome 17q23 (Temperini et al., 2006, 2007c). The mutagenesis of residue R to S at position 219 leads to impaired SLC4A4 cotransporter activity stimulation with no catalytic activity (Köhler et al., 2007b; Stams et al., 1996). It is mostly expressed in colon, fat cells, lung, thyroid and other tissues (Yang et al., 2005).

Peptide structure of principal isoform of CA IV (ensembl ID: ENST00000300900.9) where black and blue color represents alternating exons with red representing splice site):

MRMLLALLALSAARPSASAESHWCYEVQAESSNYPCLVPVKWGGNCQKDRQSPINIVTTK

AKVDKKLGRFFFSGYDKKQTWTVQNNGHSVMMLLENKASISGGGLPAPYQAKQLHLHWSD

LPYKGSEHSLDGEHFAMEMHIVHEKEKGTSRNVKEAQDPEDEIAVLAFLVEAGTQVNEGF

QPLVEALSNIPKPEMSTTMAESSLLDLLPKEEKLRHYFRYLGSLTTPTCDEKVVWTVFRE

PIQLHREQILAFSQKLYYDKEQTVSMKDNVRPLQQLGQRTVIKSGAPGRPLPWALPALLG

PMLACLLAGFLR

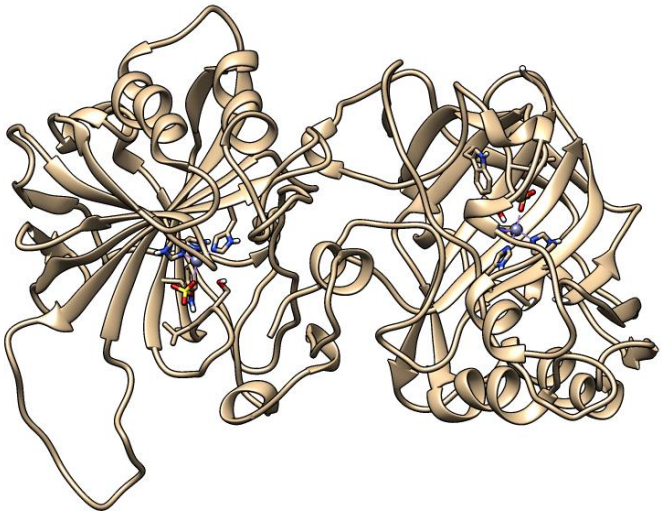


Figure 5 Structure of human CA IV (PDB ID: 1ZNC) with chain A and B as viewed in chimera

2.1.5 CA V

Human carbonic anhydrase consist of two isozymes, CA VA and CA VB. They are both located in mitochondria (Nagao et al., 1993). CA VA is located in the liver, kidney and skeletal muscle whereas CA VB is located in pancreas, kidneys, salivary glands, skeletal muscle, heart and spinal cord (Maresca et al., 2009; Van Karnebeek et al., 2014).

Peptide structure of principal isoform of CA VA (ensembl ID: ENST00000649794.2) where black and blue color represents alternating exons with red representing splice site):

MLGRNTWKTSAFSFLVEQMWAPLWSRSMRPGRWCSQRSCAWQTSNNTLHPLWTVPVSVPG

GTRQSPINIQWRDSVYDPQLKPLRVSYEAASCLYIWNTGYLFQVEFDDATEASGISGGPL

ENHYRLKQFHFHWGAVNEGGSEHTVDGHAYPAELHLVHWNSVKYQNYKEAVVGENGLAVI

GVFLKLGAHHQTLQRLVDILPEIKHKDARAAMRPFDPSTLLPTCWDYWTYAGSLTTPPLT

ESVTWIIQKEPVEVAPSQLSAFRTLLFSALGEEEKMMVNNYRPLQPLMNRKVWASFQATN

EGTRS

Yellow highlighted region is a transit peptide (residue 1-38). Mutagenesis of residue ‘S’ to ‘P’ at position 233 results in reduced enzymatic activity (Yang et al., 2005).

Peptide structure of principal isoform of CA VB (ensembl ID: ENST00000318636.8) where black and blue color represents alternating exons with red representing splice site):

MVVMNSLRVILQASPGKLLWRKFQIPRFMPARPCSLYTCTYKTRNRALHPLWESVDLVPG

GDRQSPINIRWRDSVYDPGLKPLTISYDPATCLHVWNNGYSFLVEFEDSTDKSVIKGGPL

EHNYRLKQFHFHWGAIDAWGSEHTVDSKCFPAELHLVHWNAVRFENFEDAALEENGLAVI

GVFLKLGKHHKELQKLVDTLPSIKHKDALVEFGSFDPSCLMPTCPDYWTYSGSLTTPPLS

ESVTWIIKKQPVEVDHDQLEQFRTLLFTSEGEKEKRMVDNFRPLQPLMNRTVRSSFRHDY

VLNVQAKPKPATSQATP

Yellow highlighted region is a transit peptide (1-33).

2.1.6 CA VI

CA VI is the only secretory carbonic anhydrase in mammals which is localized in serous acinar cells, ductal cells of excretory glands and various non-glandular cells (Crocetti et al., 2009b; Maresca et al., 2009). It has only one PDB structure modelled by x-ray crystallography.

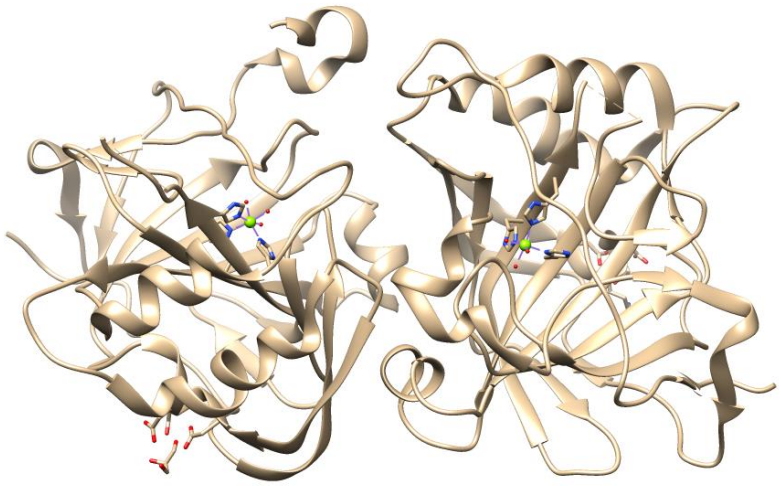


Figure 6 Structure of human carbonic anhydrase 6 (PDB ID: 3FE4) as viewed in chimera

The principal isoform (ensemble ID: ENST00000377443.7) has a signal peptide in the region 1-17. The length of CA VI (3FE4) is 278 amino acids whereas length of principal isoform (ENST00000377443.7) is 308 amino acids (Kannan et al., 1972; Temperini et al., 2007a).

2.1.7 CA VII

CA VII is located in chromosome 16q22 and encodes a protein of 263 residues (Yang et al., 2005). It is located in some brain tissues in mammals and in the stomach, colon, duodenum, liver, and skeletal muscle of mice (Montgomery et al., 1991).

Peptide structure of principal isoform of CA VII (ensembl ID: ENST00000338437.7) where black and blue color represents alternating exons with red representing splice site):

MTGHHGWGYGQDDGPSHWHKLYPIAQGDRQSPINIISSQAVYSPSLQPLELSYEACMSLS

ITNNGHSVQVDFNDSDDRTVVTGGPLEGPYRLKQFHFHWGKKHDVGSEHTVDGKSFPSEL

HLVHWNAKKYSTFGEAASAPDGLAVVGVFLETGDEHPSMNRLTDALYMVRFKGTKAQFSC

FNPKCLLPASRHYWTYPGSLTTPPLSESVTWIVLREPICISERQMGKFRSLLFTSEDDER

IHMVNNFRPPQPLKGRVVKASFRA

The principal isoform (ensemble ID: ENST00000338437.7) has 264 amino acid residues whereas PDB structure(6H38) has 270 amino acids. There are 6 PDB entries for this isozyme.

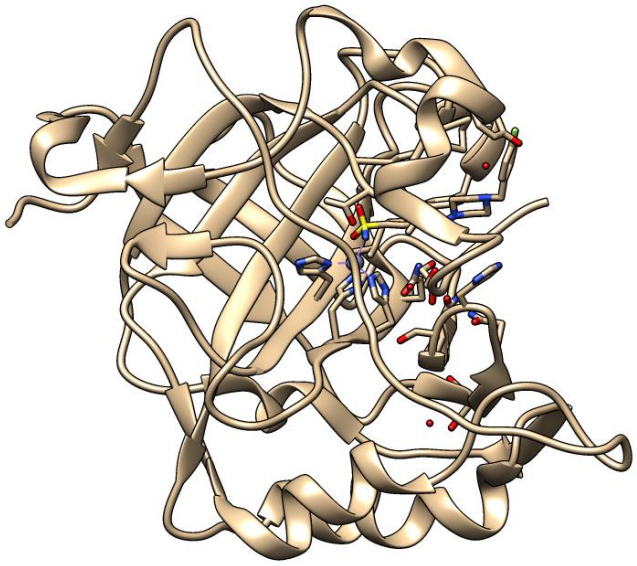


Figure 7 Ribbon diagram of CA VII structure (PDB ID: 6H38) as viewed in chimera

2.1.8 CA VIII

CA VIII has a protein coding transcript (ensemble ID: ENST00000317995.5) with 290 amino acids. It has one x-ray structure PDB entry(2W2J). It has Arg-116 instead of the conserved histidine due to which it lacks carbonic anhydrase activity (Yang et al., 2005).

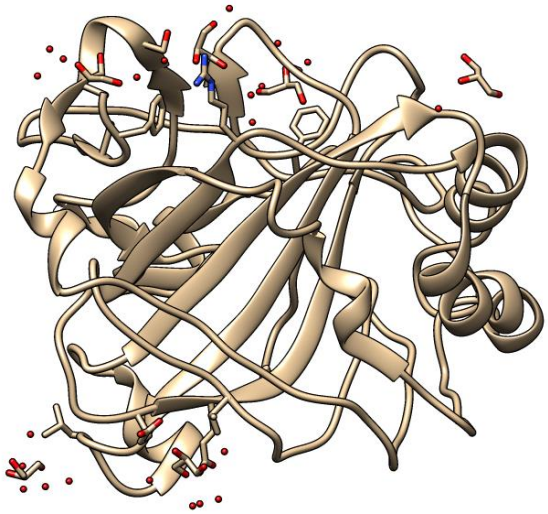


Figure 8 Ribbon diagram of CA VIII structure (PDB ID: 2W2J) as viewed in chimera.

2.1.9 CA IX

CA IX consists of extracellular domain, transmembrane region, and an intracellular tail which can exist is monomer and dimer. It is expressed in basolateral membranes of alimentary tract, testis, ovary, skeletal system, lining cells of body cavity, etc. (Supuran and Simone, 2015). It is inhibited by sulfonamide derivatives such as acetazolamide (AZA), saccharin, coumarins and Foscarnet (phosphonoformate trisodium salt) (Temperini et al., 2007b). Its optimum pH is 6.5 (Alterio et al., 2009). The signal peptide position is 1-37 (Alterio et al., 2009)

It has 11 PDB structure entries. The PDB structure 2HKF has 3 chains (L, H, and P).

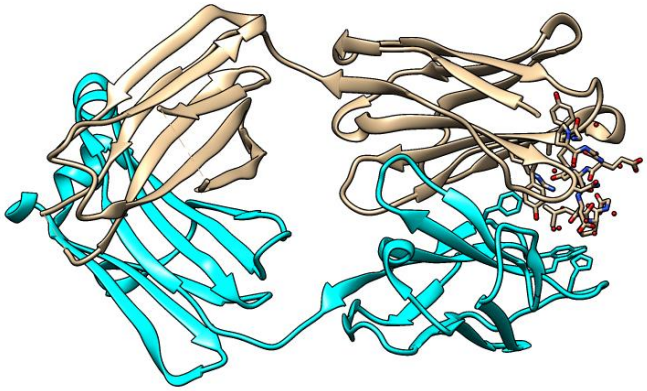


Figure 9 Ribbon diagram of CA IX structure (2HKF) with chain L highlighted in cyan and chain H in grey as viewed in chimera.

2.1.10 CA X

CA X is located in chromosome 17 and does not have catalytic activity. It is expressed in 111 organs with highest expression in brain and central nervous system (Alterio et al., 2009). It is not expressed in fetal brain. It has no PDB structure as of now.

Peptide structure of principal isoform of CA X (ensembl ID: ENST00000442502.6) where black and blue color represents alternating exons with red representing splice site):

MEIVWEVLFLLQANFIVCISAQQNSPKIHEGWWAYKEVVQGSFVPVPSFWGLVNSAWNLC

SVGKRQSPVNIETSHMIFDPFLTPLRINTGGRKVSGTMYNTGRHVSLRLDKEHLVNISGG

PMTYSHRLEEIRLHFGSEDSQGSEHLLNGQAFSGEVQLIHYNHELYTNVTEAAKSPNGLV

VVSIFIKVSDSSNPFLNRMLNRDTITRITYKNDAYLLQGLNIEELYPETSSFITYDGSMT

IPPCYETASWIIMNKPVYITRMQMHSLRLLSQNQPSQIFLSMSDNFRPVQPLNNRCIRTN

INFSLQGKDCPNNRAQKLQYRVNEWLLK

2.1.11 CA XI

This isoform is located in chromosome19 and does not have a catalytic activity. Its signal peptide is located in position 1-23. It has N-linked glycosylation site at positions 118, 170 and 260 (Alterio et al., 2009). It is expressed mainly in the brain with weak expression in spinal cord and thyroid.

Peptide structure of principal isoform of CA XI (ensembl ID: ENST00000084798.9) where black and blue color represents alternating exons with red representing splice site):

MGAAARLSAPRALVLWAALGAAAHIGPAPDPEDWWSYKDNLQGNFVPGPPFWGLVNAAWS

LCAVGKRQSPVDVELKRVLYDPFLPPLRLSTGGEKLRGTLYNTGRHVSFLPAPRPVVNVS

GGPLLYSHRLSELRLLFGARDGAGSEHQINHQGFSAEVQLIHFNQELYGNFSAASRGPNG

LAILSLFVNVASTSNPFLSRLLNRDTITRISYKNDAYFLQDLSLELLFPESFGFITYQGS

LSTPPCSETVTWILIDRALNITSLQMHSLRLLSQNPPSQIFQSLSGNSRPLQPLAHRALR

GNRDPRHPERRCRGPNYRLHVDGVPHGR

Yellow highlighted region is a signal peptide.

2.1.12 CA XII

It is located in chromosome 15 and its activity is inhibited by saccharin, sulfonamide derivatives such as acetazolamide (AZA), benzenesulfonamide and derivatives, coumarins, etc. (Alterio et al., 2009) (Crocetti et al., 2009a) (Whittington et al., 2001) (Köhler et al., 2007a). Its zinc metal binding site is located in residue positions 119, 121, 145 (Di Fiore et al., 2009). It is highly expressed in kidney, prostate, colon, activated lymphocytes and intestine with moderate expression in pancreas, testis and ovary (Di Fiore et al., 2009).

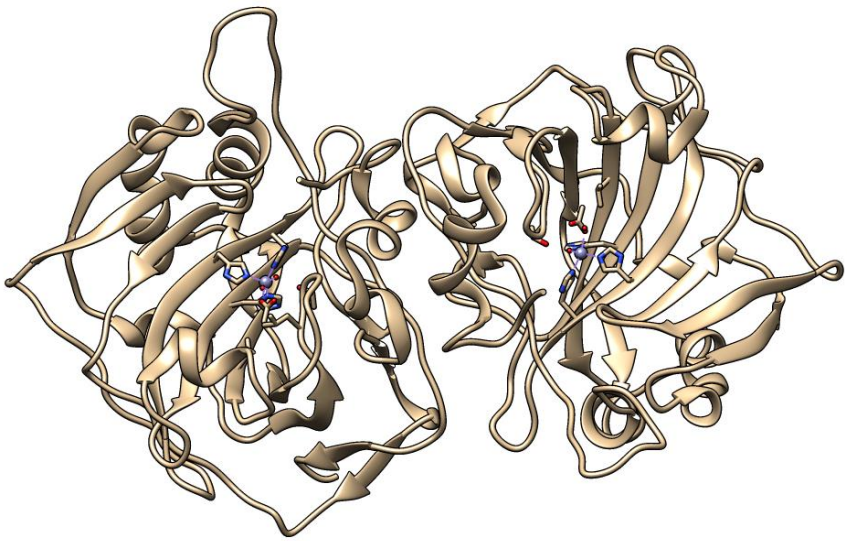


Figure 10 Ribbon diagram of CA XII structure (PDB ID: 1JCZ) as viewed in chimera.

2.1.13 CA XIII

CA XIII consists of 262 amino acids and closely related to CAs I, II, and III. It is generally expressed in the small intestine, thymus, testis, prostate, spleen, ovary, and colon with the exception in leucocytes (Di Fiore et al., 2009; Lehtonen et al., 2004). It is located in chromosome 8.

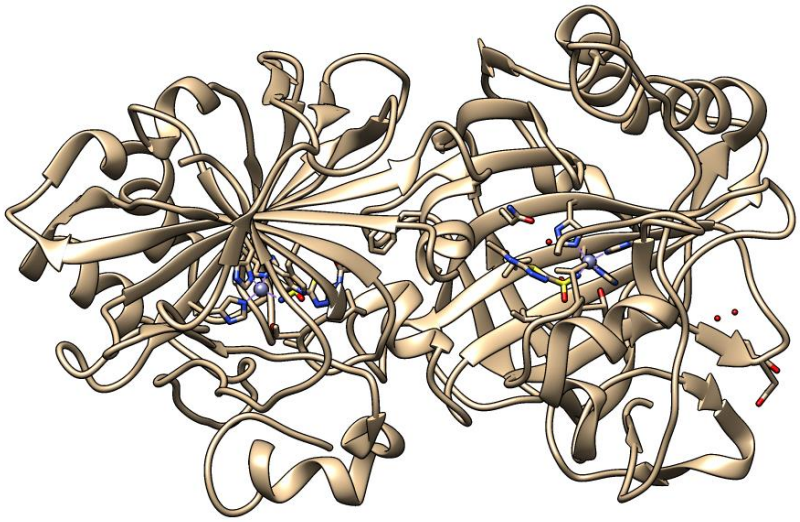


Figure 11 Ribbon diagram of CA XIII structure (PDB ID: 3CZV) as viewed in chimera.

Peptide structure of principal isoform of CA XIII (ensembl ID: ENST00000321764.4) where black and blue color represents alternating exons with red representing splice site):

MSRLSWGYREHNGPIHWKEFFPIADGDQQSPIEIKTKEVKYDSSLRPLSIKYDPSSAKII

SNSGHSFNVDFDDTENKSVLRGGPLTGSYRLRQVHLHWGSADDHGSEHIVDGVSYAAELH

VVHWNSDKYPSFVEAAHEPDGLAVLGVFLQIGEPNSQLQKITDTLDSIKEKGKQTRFTNF

DLLSLLPPSWDYWTYPGSLTVPPLLESVTWIVLKQPINISSQQLAKFRSLLCTAEGEAAA

FLVSNHRPPQPLKGRKVRASFH

2.1.14 CA XIV

This isoform along with CA XIII is most recently discovered. It is located in chromosome 1 with zinc metal binding residue at positions 109, 111 and 135 (Alterio et al., 2014). It is highly expressed in central nervous system and low expression in heart, liver, colon, small intestine, kidney, urinary bladder and skeletal muscle (Alterio et al., 2014; Temperini et al., 2006). It has two PDB structures derived from x-ray crystallography.

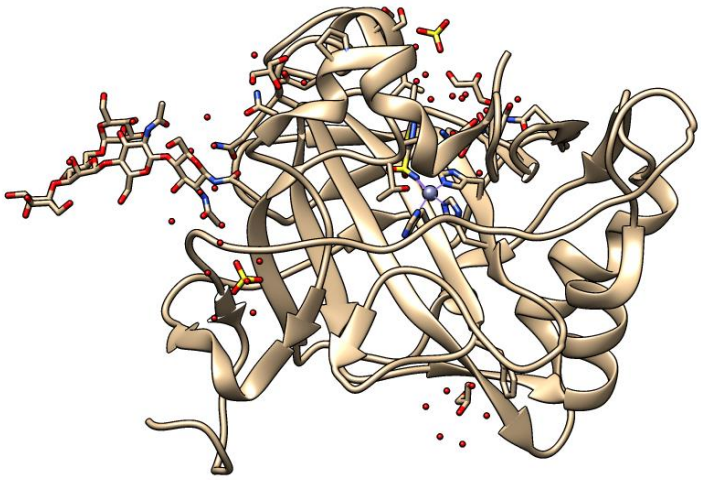


Figure 12 Ribbon diagram of CA XIV structure (PDB ID: 4LU3) as viewed in chimera.

2.2 Alternative Splicing

Alternative splicing is a process in which single primary transcript generates multiple distinct mature mRNAs with different structural and functional properties (Ghigna et al., 2008). Alternative splicing involves interaction of various cis-acting elements and trans-acting factors guided by coupling between transcription and splicing (WANG et al., 2015). There are about 25,000 protein coding genes in humans which codes for more than 90,000 different proteins. This is the result of alternative splicing process (WANG et al., 2015). Most prevalent alternative splicing pattern in vertebrates and invertebrates is the cassette-type alternative exon(exon skipping) whereas intron retention is prevalent in lower metazoans (WANG et al., 2015). Neurofibromatosis type 1(NF1) is caused by one frameshift, two nonsense and two missense mutations in RNA splicing of NF1 gene(Ars et al., 2000). Mutually exclusive splicing of 95

variable exons results in the encode of 38,016 distinct axon guidance receptors in insect Dscam (Graveley, 2005).

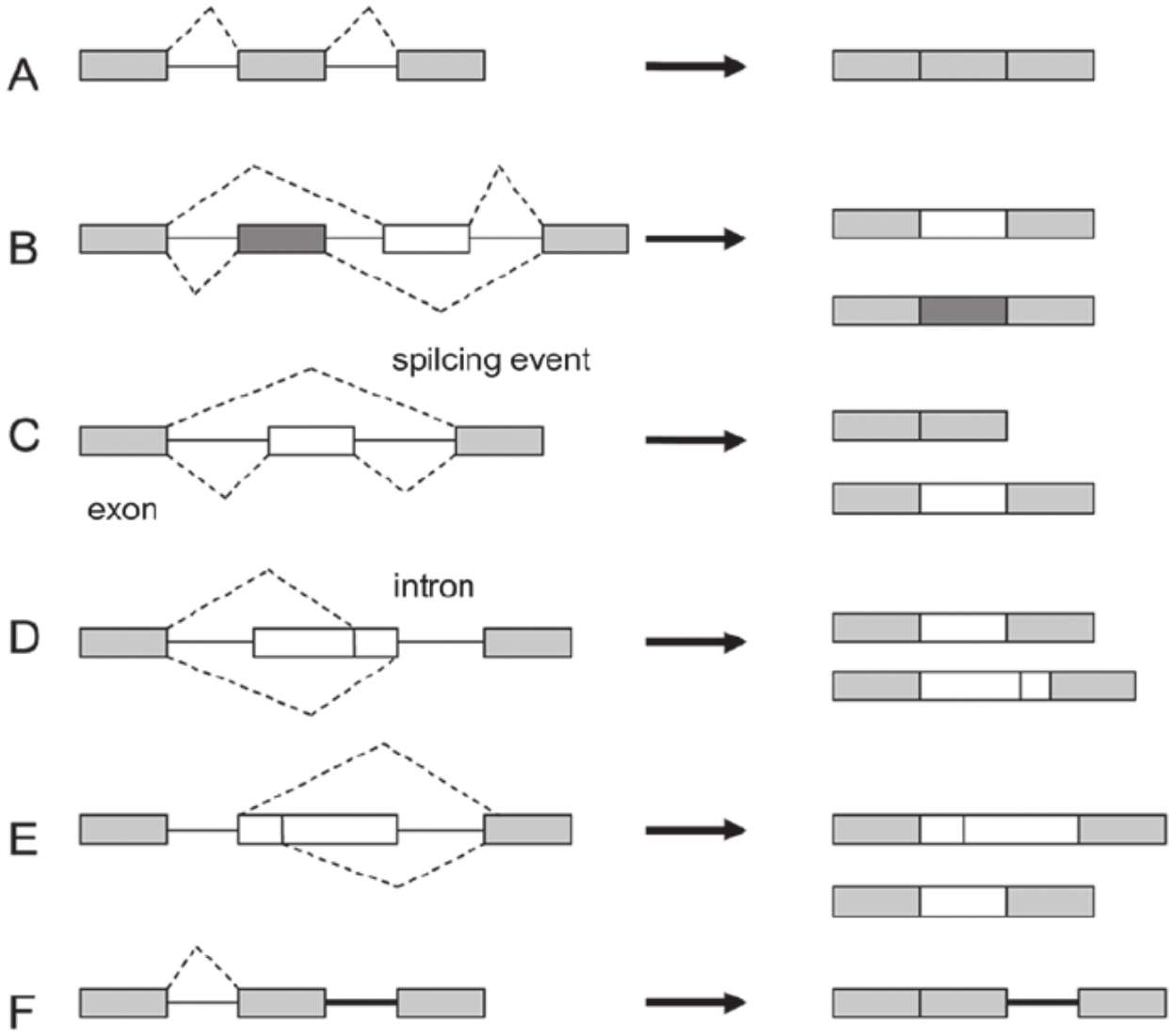


Figure 13 Five main types of alternative splicing events are depicted. (A) Constitutive splicing; (B) mutually exclusive exons; (C) cassette alternative exon; (D) alternative 3' splice site; (E) alternative 5' splice site; and (F) intron retention (WANG et al., 2015)

It has been observed that rate of alternative splicing in mammals is greater than in invertebrates. Ars et al. have predicted this observation using two data sets , first being mRNA sequences and second EST contig sequences(Kim et al., 2004).

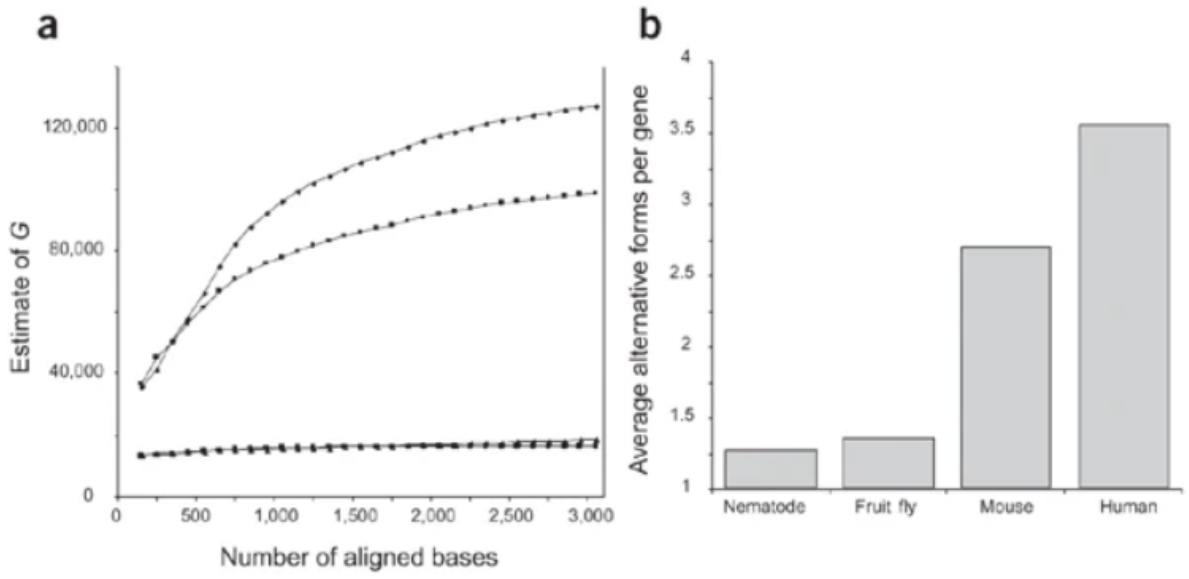


Figure 14 (a) Estimate of the gene count G with varying minimum number of aligned bases for H. sapiens (diamonds), M. musculus (squares), D. melanogaster (triangles) and C. elegans (circles). (b) Average number of alternative splice forms per gene for each organism (Kim et al., 2004).

Alternative splicing of CAIX isoform generates a transcript lacking exons 8-9 which is detected in cancer cells. The resulting protein lacks transmembrane region, the intracellular tail and the C-terminal of the catalytic domain (Malentacchi et al., 2009).

The CAXII transcript with missing 9th exon (11aa) seems to be common in astrocytomas (type of cancer that forms in brain or spinal cord) (Malentacchi et al., 2009).

3 Materials and Methods

3.1 Human Carbonic Anhydrase Isoforms

Carbonic Anhydrase 1

This gene has 25 transcripts as shown in the transcript table below:

CA1-222	ENST00000523953.5	2785	261aa	Protein coding
CA1-224	ENST00000542576.5	1236	261aa	Protein coding
CA1-201	ENST00000431316.3	1232	261aa	Protein coding
CA1-219	ENST00000523022.5	1208	261aa	Protein coding
CA1-225	ENST00000626824.1	2405	127aa	Protein coding
CA1-204	ENST00000517618.5	832	251aa	Protein coding
CA1-203	ENST00000517590.5	696	175aa	Protein coding
CA1-223	ENST00000524324.5	675	194aa	Protein coding
CA1-214	ENST00000521846.5	666	149aa	Protein coding
CA1-218	ENST00000522814.5	618	148aa	Protein coding
CA1-216	ENST00000522579.5	614	149aa	Protein coding
CA1-213	ENST00000521679.5	606	178aa	Protein coding
CA1-215	ENST00000522389.5	551	127aa	Protein coding
CA1-208	ENST00000519991.5	527	137aa	Protein coding
CA1-221	ENST00000523858.5	518	99aa	Protein coding
CA1-217	ENST00000522662.5	501	118aa	Protein coding
CA1-207	ENST00000519129.5	491	22aa	Protein coding
CA1-210	ENST00000520663.5	454	87aa	Protein coding
CA1-202	ENST00000517429.5	652	81aa	Nonsense mediated decay
CA1-206	ENST00000518341.5	655	No protein	Processed transcript
CA1-209	ENST00000520093.5	594	No protein	Retained intron
CA1-220	ENST00000523712.5	574	No protein	Retained intron
CA1-212	ENST00000520990.5	525	No protein	Retained intron
CA1-205	ENST00000518233.5	355	No protein	Retained intron
CA1-211	ENST00000520692.1	297	No protein	Retained intron

Among them 18 transcripts are protein coding.

Carbonic Anhydrase 2

This gene has 5 transcripts out of which only one is protein-coding as shown in the transcript table below:

Name	Transcript ID	bp	Protein	Biotype
CA2-201	ENST00000285379.10	1562	260aa	Protein coding
CA2-203	ENST00000520127.5	875	97aa	Nonsense mediated decay
CA2-205	ENST00000522742.1	691	101aa	Nonsense mediated decay
CA2-204	ENST00000520996.5	684	No protein	Retained intron
CA2-202	ENST00000518231.1	587	No protein	Retained intron

Carbonic Anhydrase 3

This gene has 3 transcripts:

Name	Transcript ID	bp	Protein	Biotype
CA3-201	ENST00000285381.3	1721	260aa	Protein coding
CA3-202	ENST00000520921.1	571	19aa	Protein coding
CA3-203	ENST00000522207.1	645	No protein	Retained intron

Two of them are protein coding.

Carbonic Anhydrase 4

This gene has 6 transcripts out of which 4 are protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA4-201	ENST00000300900.9	1123	312aa	Protein coding
CA4-204	ENST00000587265.1	702	99aa	Protein coding
CA4-205	ENST00000590203.1	677	184aa	Protein coding
CA4-206	ENST00000591725.1	565	38aa	Protein coding
CA4-203	ENST00000586876.1	941	106aa	Nonsense mediated decay
CA4-202	ENST00000585705.5	598	No protein	Retained intron

Carbonic Anhydrase 5A

This gene has 3 transcripts out of which 1 is protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA5A-201	ENST00000309893.3	7567	305aa	Protein coding
CA5A-207	ENST00000649794.2	1113	305aa	Protein coding
CA5A-206	ENST00000649158.1	1298	300aa	Protein coding
CA5A-205	ENST00000648177.1	1113	191aa	Protein coding
CA5A-204	ENST00000648022.1	1248	222aa	Nonsense mediated decay
CA5A-203	ENST00000568801.1	506	No protein	Processed transcript
CA5A-202	ENST00000566402.2	725	No protein	Retained intron

Carbonic Anhydrase 5B

This gene has 10 transcripts out of which 4 are protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA5B-201	ENST00000318636.8	6837	317aa	Protein coding
CA5B-204	ENST00000454127.2	2205	317aa	Protein coding
CA5B-208	ENST00000479740.5	529	136aa	Protein coding
CA5B-210	ENST00000498004.5	488	76aa	Protein coding
CA5B-206	ENST00000478341.1	517	70aa	Nonsense mediated decay
CA5B-202	ENST00000380313.1	510	No protein	Processed transcript
CA5B-203	ENST00000380319.2	652	No protein	Retained intron
CA5B-207	ENST00000478923.1	578	No protein	Retained intron
CA5B-205	ENST00000474624.5	565	No protein	Retained intron
CA5B-209	ENST00000496188.1	515	No protein	Retained intron

Carbonic Anhydrase 6

This gene has 6 transcripts out of which 5 are protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA6-203	ENST00000377443.7	1334	308aa	Protein coding
CA6-201	ENST00000377436.6	942	313aa	Protein coding
CA6-202	ENST00000377442.3	747	248aa	Protein coding
CA6-205	ENST00000480186.7	1493	179aa	Protein coding
CA6-206	ENST00000549778.5	582	187aa	Protein coding
CA6-204	ENST00000476083.1	787	No protein	Processed transcript

Carbonic Anhydrase 7

This gene has 3 transcripts. Two of them are protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA7-201	ENST00000338437.7	1518	264aa	Protein coding
CA7-202	ENST00000394069.3	1713	208aa	Protein coding
CA7-203	ENST00000548332.6	1161	44aa	Nonsense mediated decay

Carbonic Anhydrase 8

This gene has 4 transcripts. Only 1 of them is protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA8-201	ENST00000317995.5	5735	290aa	Protein coding
CA8-203	ENST00000528666.1	510	No protein	Processed transcript
CA8-202	ENST00000524872.5	1825	No protein	Retained intron
CA8-204	ENST00000529918.1	1697	No protein	Retained intron

Carbonic Anhydrase 9

This gene has 4 transcripts. Two of them are protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA9-201	ENST00000378357.9	1546	459aa	Protein coding
CA9-204	ENST00000617161.1	1374	356aa	Protein coding
CA9-203	ENST00000493245.1	697	No protein	Processed transcript
CA9-202	ENST00000485665.1	329	No protein	Processed transcript

Carbonic Anhydrase 10

This gene has 9 transcripts. Six of them are protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA10-203	ENST00000451037.7	3179	328aa	Protein coding
CA10-202	ENST00000442502.6	2935	328aa	Protein coding
CA10-201	ENST00000285273.8	2914	328aa	Protein coding
CA10-204	ENST00000570565.5	2276	253aa	Protein coding
CA10-209	ENST00000575181.1	1474	276aa	Protein coding
CA10-208	ENST00000575097.1	410	56aa	Protein coding
CA10-205	ENST00000571371.5	2699	42aa	Nonsense mediated decay
CA10-206	ENST00000571918.1	628	No protein	Processed transcript
CA10-207	ENST00000573294.1	567	No protein	Processed transcript

Carbonic Anhydrase 11

This gene has 4 transcripts. Two of them are protein coding.

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
CA11-201	ENST00000084798.8	1844	328aa	Protein coding	CCDS12729	O75493	NM_001217 NP_001208 NR_136241	TSL:1 GENCODE basic APPRIS P1
CA11-203	ENST00000596080.1	498	113aa	Protein coding	-	M0QXK8	-	CDS 5' incomplete TSL:3
CA11-204	ENST00000599267.1	381	No protein	Retained intron	-	-	-	TSL:2
CA11-202	ENST00000594088.1	275	No protein	Retained intron	-	-	-	TSL:2

Carbonic Anhydrase 12

This gene has 6 transcripts. Three of them are protein coding.

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
CA12-203	ENST00000422263.2	2556	283aa	Protein coding	CCDS76767	B3KUB4	NM_001293642 NP_001280571	TSL:2 GENCODE basic
CA12-202	ENST00000344366.7	2744	343aa	Protein coding	CCDS10186	O43570	NM_206925 NP_996808	TSL:1 GENCODE basic APPRIS ALT1
CA12-201	ENST00000178638.7	6413	354aa	Protein coding	CCDS10185	O43570	NM_001218 NP_001209 NR_135511	TSL:1 GENCODE basic APPRIS P4
CA12-206	ENST00000560666.1	580	No protein	Processed transcript	-	-	-	TSL:3
CA12-204	ENST00000558287.1	561	No protein	Retained intron	-	-	-	TSL:3
CA12-205	ENST00000560293.1	538	No protein	Retained intron	-	-	-	TSL:4

Carbonic Anhydrase 13

This gene has 5 transcripts. Only 1 of them is protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA13-201	ENST00000321764.4	3884	262aa	Protein coding
CA13-202	ENST00000517298.5	1400	No protein	Processed transcript
CA13-203	ENST00000517831.5	865	No protein	Processed transcript
CA13-205	ENST00000522631.1	700	No protein	Processed transcript
CA13-204	ENST00000518392.1	461	No protein	Retained intron

Carbonic Anhydrase 14

This gene has 5 transcripts. Three of them are protein coding.

Name	Transcript ID	bp	Protein	Biotype
CA14-201	ENST00000369111.9	1788	337aa	Protein coding
CA14-206	ENST00000647854.1	1531	337aa	Protein coding
CA14-204	ENST00000607082.1	594	130aa	Protein coding
CA14-202	ENST00000483993.3	938	42aa	Nonsense mediated decay
CA14-203	ENST00000582010.3	1514	No protein	Retained intron
CA14-205	ENST00000607652.5	1382	No protein	Retained intron

3.2 Mouse Carbonic Anhydrase Isoforms

Carbonic Anhydrase 1

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car1-203	ENSMUST00000181860.7	1194	261aa	Protein coding	CCDS17248	P13634	NM_009799 NP_033929	TSL:1 GENCODE basic APPRIS P1
Car1-201	ENSMUST00000094365.10	1180	261aa	Protein coding	CCDS17248	P13634	NM_001083957 NP_001077426	TSL:1 GENCODE basic APPRIS P1
Car1-202	ENSMUST00000144327.2	367	92aa	Protein coding	-	D3YYQ4	-	CDS 3' incomplete TSL:3

Number of transcrips:3

Protein coding:3

Carbonic Anhydrase 2

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car2-201	ENSMUST00000029078.8	1788	260aa	Protein coding	CCDS17251	P00920	NM_001357334 NM_009801 NP_001344263 NP_033931	TSL:1 GENCODE basic APPRIS P1
Car2-202	ENSMUST00000192609.5	587	115aa	Protein coding	-	A0A0A6YX78	-	CDS 3' incomplete TSL:3
Car2-203	ENSMUST00000195520.1	881	No protein	Retained intron	-	-	-	TSL:2

Number of transcripts: 3

Protein coding: 2

Carbonic Anhydrase 3

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car3-201	ENSMUST00000029076.5	1679	260aa	Protein coding	CCDS17250	P16015	NM_007606 NP_031632	TSL:1 GENCODE basic APPRIS P1
Car3-202	ENSMUST00000195575.1	1395	No protein	Retained intron	-	-	-	TSL:1
Car3-203	ENSMUST00000195834.1	804	No protein	Retained intron	-	-	-	TSL:2

Number of transcripts: 3

Protein Coding: 1

Carbonic Anhydrase 4

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car4-201	ENSMUST00000103194.9	1256	305aa	Protein coding	CCDS25190	Q64444	NM_007607 NP_031633	TSL:1 GENCODE basic APPRIS P1
Car4-202	ENSMUST00000108076.2	732	164aa	Protein coding	-	F6ST32	-	CDS 5' incomplete TSL:3
Car4-206	ENSMUST00000150596.7	692	38aa	Nonsense mediated decay	-	D6RCZ3	-	TSL:5
Car4-203	ENSMUST00000127827.1	516	38aa	Nonsense mediated decay	-	D6RCZ3	-	TSL:2
Car4-204	ENSMUST00000138331.1	443	No protein	Processed transcript	-	-	-	TSL:3
Car4-205	ENSMUST00000139416.1	435	No protein	Retained intron	-	-	-	TSL:3

Number of transcripts: 6

Protein coding: 2

Carbonic anhydrase 5a

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car5a-201	ENSMUST00000057653.7	1249	299aa	Protein coding	CCDS22731	P23589	NM_007608 NP_031634	TSL:1 GENCODE basic APPRIS P1
Car5a-202	ENSMUST00000151462.1	458	No protein	Processed transcript	-	-	-	TSL:3

Number of transcripts: 2

Protein coding: 1

Carbonic anhydrase 5b

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car5b-201	ENSMUST00000033739.4	3436	317aa	Protein coding	CCDS30515	Q9QZA0	NM_181315 NP_851832	TSL:1 GENCODE basic APPRIS P1
Car5b-202	ENSMUST00000126650.1	3157	No protein	Retained intron	-	-	-	TSL:2

Number of transcripts: 2

Protein coding: 1

Carbonic anhydrase 6

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car6-201	ENSMUST00000030817.4	1567	317aa	Protein coding	CCDS51386	P18761	NM_009802 NP_033932	TSL:1 GENCODE basic APPRIS P1
Car6-202	ENSMUST00000105683.8	1460	261aa	Protein coding	-	B1ARR4	-	TSL:1 GENCODE basic
Car6-204	ENSMUST00000134648.7	749	No protein	Processed transcript	-	-	-	TSL:3
Car6-203	ENSMUST00000126449.1	727	No protein	Processed transcript	-	-	-	TSL:5

Number of transcripts: 4

Protein coding: 2

Carbonic anhydrase 7

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car7-202	ENSMUST00000159416.7	1943	208aa	Protein coding	CCDS80920	G3XA26	NM_001301164 NP_001288093	TSL:1 GENCODE basic
Car7-201	ENSMUST00000056051.10	1538	264aa	Protein coding	CCDS22581	Q9ERQ8	NM_053070 NP_444300	TSL:1 GENCODE basic APPRIS P1
Car7-204	ENSMUST00000162761.1	1505	208aa	Protein coding	CCDS80920	G3XA26	NM_001301165 NP_001288094	TSL:1
Car7-205	ENSMUST00000212942.1	1702	No protein	Retained intron	-	-	-	TSL:NA
Car7-203	ENSMUST00000162399.1	592	No protein	Retained intron	-	-	-	TSL:3

Number of transcripts: 5

Protein coding: 3

Carbonic anhydrase 8

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car8-201	ENSMUST00000066674.7	3853	291aa	Protein coding	CCDS17954	P28651	NM_007592 NP_031618	TSL:1 GENCODE basic APPRIS P1

Number of transcripts: 1

Protein coding: 1

Carbonic anhydrase 9

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car9-201	ENSMUST00000030183.9	2023	437aa	Protein coding	CCDS18099	Q3UUZ9 Q8VHB5	NM_139305 NP_647466	TSL:1 GENCODE basic APPRIS P1
Car9-206	ENSMUST00000138073.1	712	237aa	Protein coding	-	F6XXU0	-	CDS 5' and 3' incomplete TSL:1
Car9-202	ENSMUST00000124114.7	1667	No protein	Processed transcript	-	-	-	TSL:5
Car9-205	ENSMUST00000129996.1	480	No protein	Processed transcript	-	-	-	TSL:5
Car9-207	ENSMUST00000154251.1	398	No protein	Processed transcript	-	-	-	TSL:5
Car9-203	ENSMUST00000126750.1	374	No protein	Processed transcript	-	-	-	TSL:3
Car9-204	ENSMUST00000128232.1	334	No protein	Processed transcript	-	-	-	TSL:3

Number of transcripts: 7

Protein coding:2

Carbonic anhydrase 10

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car10-206	ENSMUST00000107863.3	3334	328aa	Protein coding	CCDS25244	P61215 Q3V1V7	NM_028296 NP_082572	TSL:1 GENCODE basic APPRIS P1
Car10-201	ENSMUST00000042943.12	3293	328aa	Protein coding	CCDS25244	P61215 Q3V1V7	NM_001361707 NM_001361708 NP_001348636 NP_001348637	TSL:1 GENCODE basic APPRIS P1
Car10-203	ENSMUST00000107858.8	2742	304aa	Protein coding	-	E9Q2V1	-	TSL:1 GENCODE basic
Car10-205	ENSMUST00000107861.7	2613	169aa	Protein coding	-	Q3TRQ4	-	TSL:1 GENCODE basic
Car10-204	ENSMUST00000107859.7	1738	103aa	Protein coding	-	Q9CZQ3	-	TSL:1
Car10-207	ENSMUST00000149611.1	303	No protein	Processed transcript	-	-	-	TSL:2
Car10-202	ENSMUST00000092780.8	1885	No protein	Retained intron	-	-	-	TSL:1

Number of transcripts: 7

Protein coding: 5

Carbonic anhydrase 11

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car11-201	ENSMUST00000003360.9	1585	328aa	Protein coding	CCDS21259	O70354 Q541E9	NM_009800 NP_033930	TSL:1 GENCODE basic APPRIS P1
Car11-202	ENSMUST00000209796.1	599	82aa	Nonsense mediated decay	-	A0A1B0GRJ7	-	CDS 5' incomplete TSL:5
Car11-203	ENSMUST00000210027.1	252	26aa	Nonsense mediated decay	-	A0A1B0GT65	-	CDS 5' incomplete TSL:5
Car11-204	ENSMUST00000210872.1	858	No protein	Retained intron	-	-	-	TSL:3
Car11-205	ENSMUST00000211259.1	706	No protein	Retained intron	-	-	-	TSL:1

Number of transcripts: 5

Protein coding: 1

Carbonic anhydrase 12

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car12-201	ENSMUST00000071889.12	3716	354aa	Protein coding	CCDS23306	A0A0R4J0W4	NM_178396 NP_848483	TSL:1 GENCODE basic APPRIS P3
Car12-202	ENSMUST00000085420.11	3639	344aa	Protein coding	CCDS81026	Q8K2J1	NM_001306148 NP_001293077	TSL:1 GENCODE basic APPRIS ALT2
Car12-204	ENSMUST00000134829.1	614	204aa	Protein coding	-	F6W0I8	-	CDS 5' and 3' incomplete TSL:5
Car12-206	ENSMUST00000217394.1	3263	No protein	Retained intron	-	-	-	TSL:NA
Car12-205	ENSMUST00000152011.7	1652	No protein	Retained intron	-	-	-	TSL:2
Car12-203	ENSMUST00000123195.1	710	No protein	Retained intron	-	-	-	TSL:2

Number of transcripts: 6

Protein coding:3

Carbonic anhydrase 13

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car13-201	ENSMUST00000029071.8	2299	262aa	Protein coding	CCDS17247	Q9D6N1	NM_024495 NP_078771	TSL:1 GENCODE basic APPRIS P1

Number of transcripts: 1

Protein coding: 1

Carbonic anhydrase 14

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car14-201	ENSMUST00000036181.14	1709	337aa	Protein coding	CCDS17625	Q9WVT6	NM_001355750 NM_001355751 NM_011797 NP_001342679 NP_001342680 NP_035927	TSL:1 GENCODE basic APPRIS P1
Car14-203	ENSMUST00000147962.2	895	171aa	Protein coding	-	D3Z4J8	-	CDS 3' incomplete TSL:3
Car14-204	ENSMUST00000149202.1	984	No protein	Retained intron	-	-	-	TSL:1
Car14-202	ENSMUST00000126722.1	452	No protein	Retained intron	-	-	-	TSL:2

Number of transcripts:4

Protein coding: 2

Carbonic anhydrase 15

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
Car15-201	ENSMUST00000118960.1	1283	324aa	Protein coding	CCDS49784	Q99N23	NM_030558 NP_085035	TSL:1 GENCODE basic APPRIS P1
Car15-204	ENSMUST00000232529.1	789	No protein	Retained intron	-	-	-	
Car15-202	ENSMUST00000231865.1	595	No protein	Retained intron	-	-	-	
Car15-203	ENSMUST00000232516.1	520	No protein	Retained intron	-	-	-	

Number of transcripts: 4

Protein coding: 1

3.3 Classification of human protein-coding CA transcript variants

Table 2 Classification of human protein-coding CAs transcript variants based on various properties. Transcripts highlighted in yellow represent principal isoform based on ensemble Flags (APPRIS, GENCODE basic and TSL) and complete catalytic domain.

Name	Transcript ID	Protein	Biotype	Signal Peptide	Catalytic Domain	Active Site	Metal-ion Binding Site	TM Helix
CA I-222	ENST00000523953.5	261aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA I-219	ENST00000523022.5	261aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA I-215	ENST00000522389.5	127aa	Protein coding	Missing	Incomplete	No	No	Missing
CA I-225	ENST00000626824.1	127aa	Protein coding	Missing	Incomplete	No	No	Missing
CA I-224	ENST00000431316.3	261aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA I-201	ENST00000431316.3	261aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA I-204	ENST00000517618.5	251aa	Protein coding	Missing	Incomplete	Yes	Yes	Missing
CA I-223	ENST00000524324.5	194aa	Protein coding	Missing	Incomplete	No	No	Missing
CA I-208	ENST00000519991.5	137aa	Protein coding	Missing	Incomplete	No	No	Missing
CA I-217	ENST00000522662.5	118aa	Protein coding	Missing	Incomplete	No	No	Missing
CA I-214	ENST00000521846.5	149aa	Protein coding	Missing	Incomplete	Yes	No	Missing
CA I-218	ENST00000522814.5	148aa	Protein coding	Missing	Incomplete	Yes	No	Missing
CA I-216	ENST00000522579.5	149aa	Protein coding	Missing	Incomplete	Yes	No	Missing
CA I-221	ENST00000523858.5	99aa	Protein coding	Missing	Incomplete	No	No	Missing
CA I-210	ENST00000520663.5	87aa	Protein coding	Missing	Incomplete	No	No	Missing
CA I-203	ENST00000517590.5	175aa	Protein coding	Missing	Incomplete	Yes	No	Missing
CA I-207	ENST00000519129.5	22aa	Protein coding	Missing	Incomplete	No	No	Missing
CA I-213	ENST00000521679.5	178aa	Protein coding	Missing	Incomplete	No	No	Missing
CA II-201	ENST00000285379.10	260aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA III-201	ENST00000285381.3	260aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA III-202	ENST00000520921.1	19aa	Protein coding	Missing	Incomplete	No	No	Missing
CA IV-201	ENST00000300900.9	312aa	Protein coding	Present	Complete	Yes	Yes	Present
CA IV-206	ENST00000591725.1	38aa	Protein coding	Missing	Incomplete	No	No	Missing
CA IV-204	ENST00000587265.1	99aa	Protein coding	Missing	Incomplete	No	No	Missing
CA IV-205	ENST00000590203.1	184aa	Protein coding	Missing	Incomplete	Yes	No	Present

CA VA-206	ENST00000649794.2	305aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA VA-205	ENST00000649158.1	300aa	Protein coding	Missing	Incomplete	Yes	Yes	Missing
CA VA-204	ENST00000648177.1	191aa	Protein coding	Missing	Incomplete	No	No	Missing
CA VB-201	ENST00000318636.8	317	Protein coding	Missing	Complete	Yes	Yes	Missing
CA VB-204	ENST00000454127.2	317aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA VB-208	ENST00000479740.5	136aa	Protein coding	Missing	Incomplete	No	No	Missing
CA VB-210	ENST00000498004.5	76aa	Protein coding	Missing	Incomplete	No	No	Missing
CA VI-202	ENST00000377442.3	248aa	Protein coding	Present	Incomplete	No	No	Missing
CA VI-203	ENST00000377436.6	313aa	Protein coding	Present	complete	Yes	Yes	Missing
CA VI-201	ENST00000377443.7	308aa	Protein coding	Present	Complete	Yes	Yes	Missing
CA VI-205	ENST00000480186.7	179aa	Protein coding	Present	Incomplete	No	No	Missing
CA VII-201	ENST00000338437.7	264aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA VII-202	ENST00000394069.3	208aa	Protein coding	Missing	Incomplete	Yes	Yes	Missing
CA VIII-201	ENST00000317995.5	290aa	Protein coding	Missing	Complete	Yes	Yes	Missing
CA IX-201	ENST00000378357.9	459aa	Protein coding	Present	Complete	Yes	Yes	Present
CA IX-204	ENST00000617161.1	356aa	Protein coding	Present	Incomplete	Yes	Yes	Missing
CA X-202	ENST00000571371.5	328aa	Protein coding	Present	Complete	-	-	Missing
CA X-201	ENST00000285273.8	328aa	Protein coding	Present	Complete	-	-	Missing
CA X-203	ENST00000451037.7	328aa	Protein coding	Present	Complete	-	-	Missing
CA X-209	ENST00000575181.1	276aa	Protein coding	Present	Incomplete	-	-	Missing
CA X-204	ENST00000570565.5	253aa	Protein coding	Present	Incomplete	-	-	Missing
CA X-208	ENST00000575097.1	56aa	Protein coding	Absent	Incomplete	-	-	Missing
CA X-205	ENST00000571371.5	42aa	Protein coding	Absent	Incomplete	-	-	Missing
CA XI-201	ENST00000084798.9	328aa	Protein coding	Present	Complete	-	-	Missing
CA XI-201	ENST00000084798.9	328aa	Protein coding	Present	Complete	-	-	Missing
CA XI-203	ENST00000596080.1	113aa	Protein coding	Present	Incomplete	-	-	Missing
CA XII-201	ENST00000178638.8	354aa	Protein coding	Present	Complete	Yes	Yes	Yes
CA XII-202	ENST00000344366.7	343aa	Protein coding	Present	Complete	Yes	Yes	Yes

CA XII-203	ENST00000422263.2	283aa	Protein coding	Present	Incomplete	No	Yes	Yes
CA XIII-201	ENST00000321764.4	262aa	Protein coding	No	Complete	Yes	Yes	No
CA XIV-201	ENST00000369111.9	337aa	Protein coding	Yes	Complete	Yes	Yes	Yes
CA XIV-206	ENST00000647854.1	337aa	Protein coding	Yes	Complete	Yes	Yes	Yes
CA XIV-204	ENST00000607082.1	130aa	Protein coding	No	Incomplete	No	No	No

CA VA-206 and CA VA-205 and CA VA-204 transcripts have transit peptide (in position 1-38).

CA VB-201, CA VB-208, CA VB-208 and CA VB-210 transcripts have transit peptide (in position 1-33)

4 Results

4.1 Exon-oriented protein visualization of CA II (human)

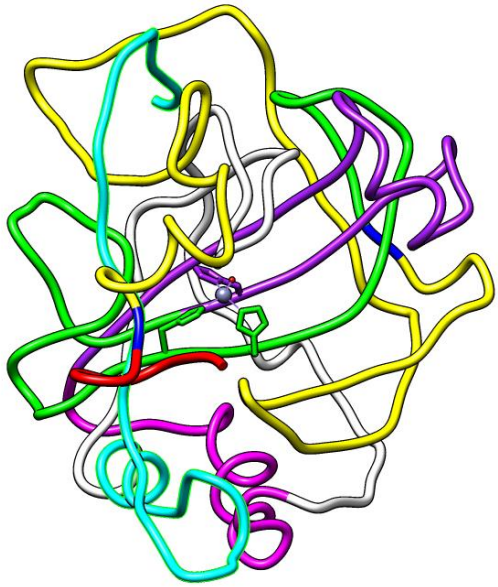
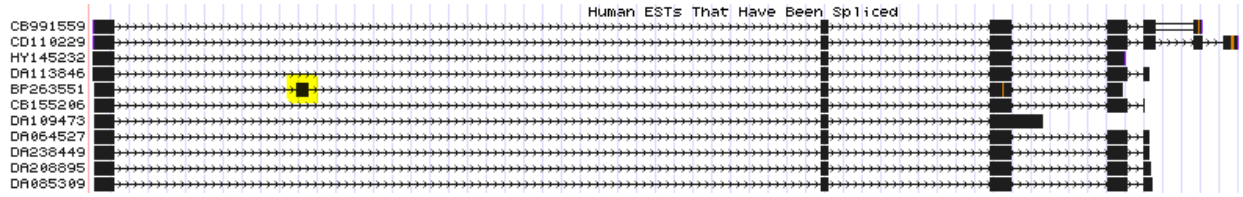


Figure 15 Exon-oriented protein visualization of human CA II using chimera. Exon 1 is color-coded with red, exon 2- yellow, exon 3- green, exon 4- purple, exon-5- magenta, exon 6- white, exon 7- cyan. Spliced codon is color coded blue. 1CA II is used as the PDB identifier.

4.2 Missing/extra exons in Human Extracellular Carbonic anhydrases

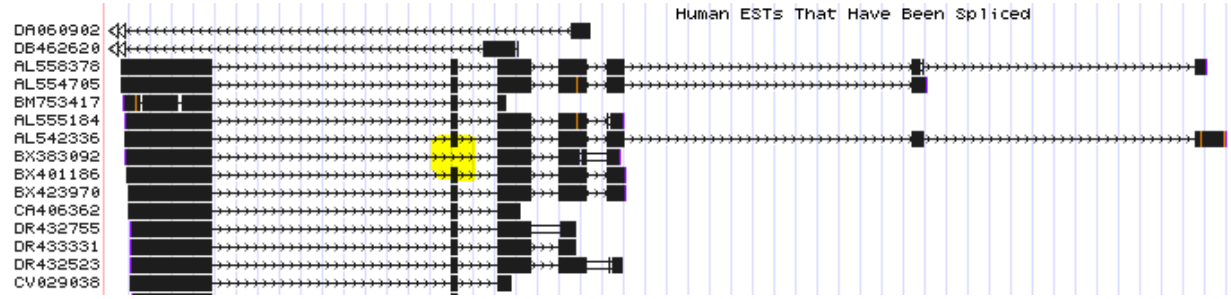
4.2.1 CA IV

Extra second-exon in spliced EST view in UCSC. Ensembl has no transcripts with extra-exon.

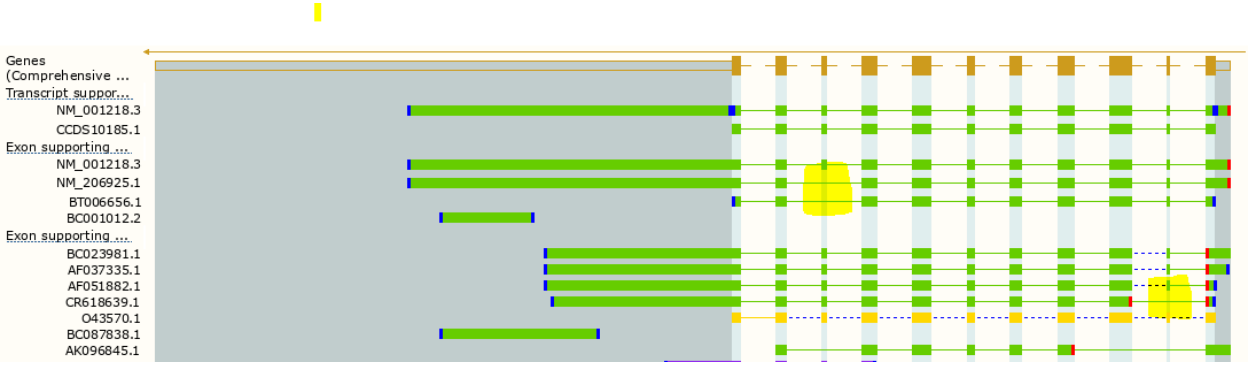


4.2.2 CA IX

Second short exon in human CA IX is missing as shown in the UCSC spliced-Est track. Ensembl does not show any transcripts with second exon skipped.

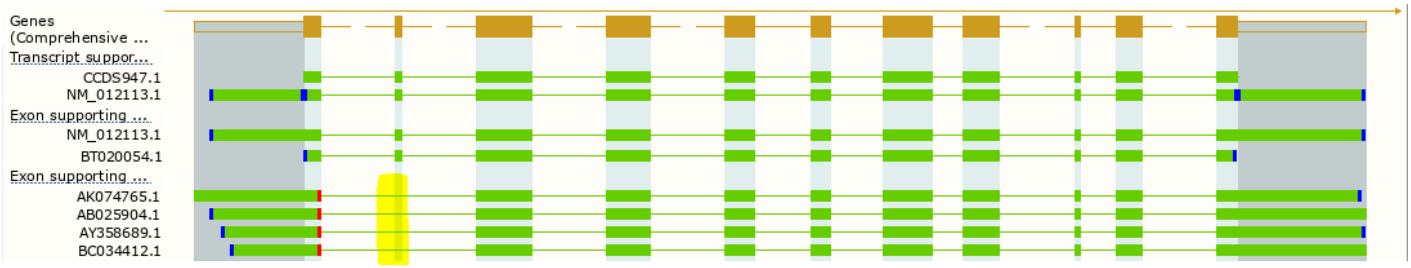


4.2.3 CA XII



Ensembl transcript/supporting evidence shows second and ninth exon missing as highlighted in the image. There is even Refseq transcript with missing ninth exon.

4.2.4 CA XIV



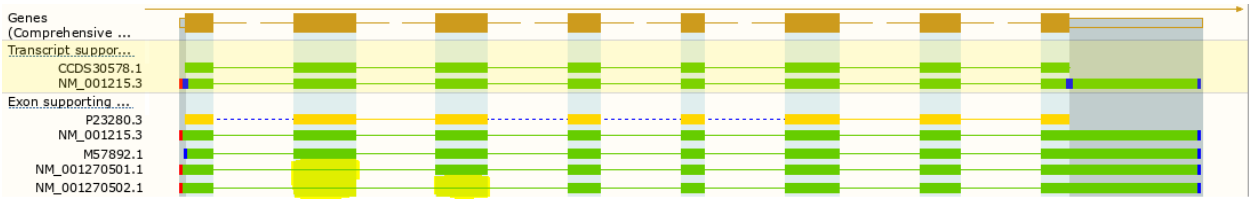
Ensembl supporting evidence shows second exon missing in some cDNAs as highlighted in the image.

CA I5 has no protein-coding transcripts in ensemble.

4.3 Missing/extra exons in Human Secretory Carbonic Anhydrase

4.3.1 CA VI

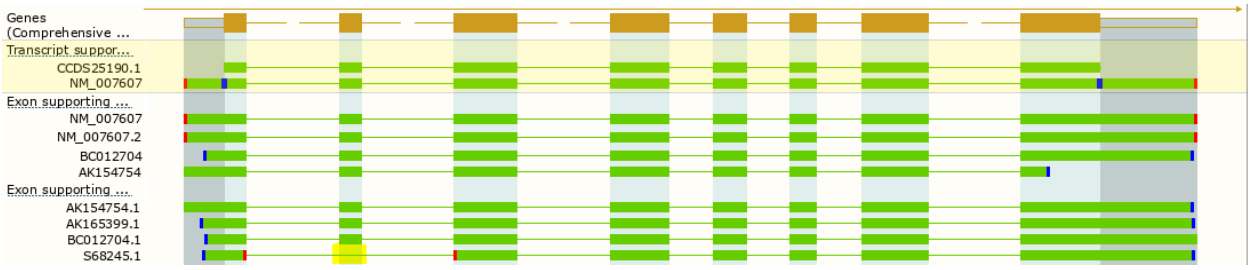
For CA VI, there is even a RefSeq entries NM_001270501.1 and NM_001270502.1 in which the second and third exons are skipped.



4.4 Missing/extra exons in Mouse Extracellular Carbonic Anhydrase

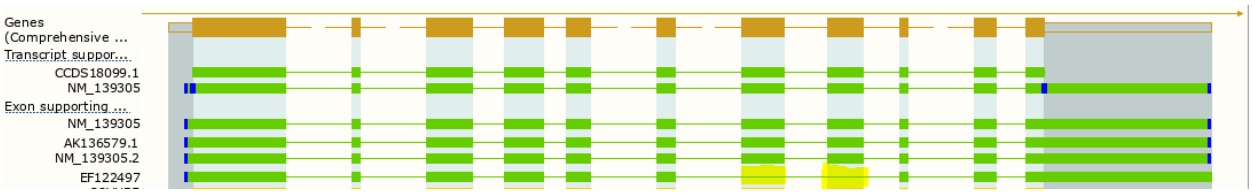
4.4.1 Car IV

Ensembl shows one EST evidence with missing second exon.



4.4.2 Car IX

Ensembl shows missing 7th and 8th exons in EST evidence.

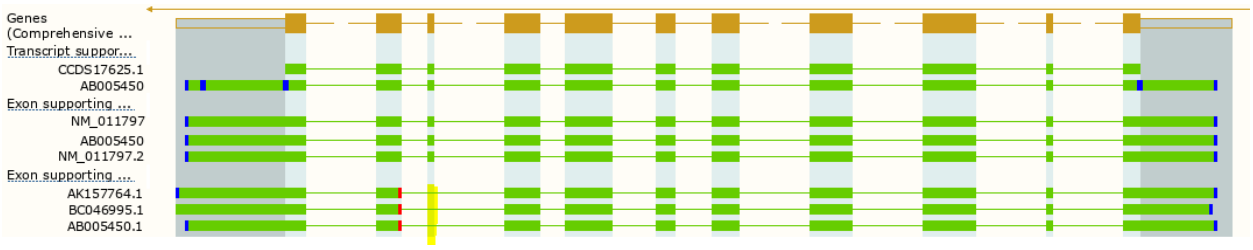


4.4.3 Car XII

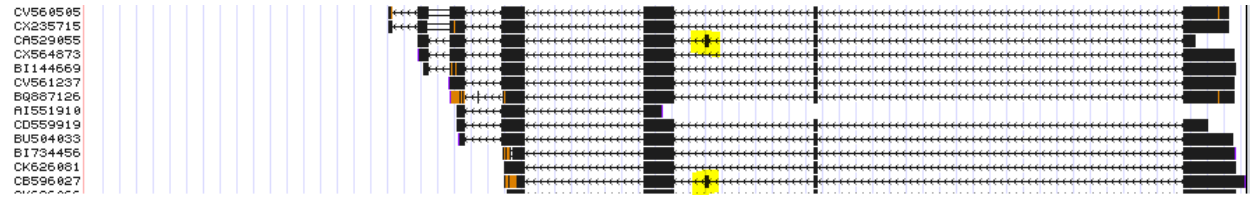
Ensembl and genome browser shows no missing exons.

4.4.4 Car XIV

Ensembl shows missing 9th exon in EST evidence track.



UCSC genome browser shows extra 3rd exon in spliced EST.



4.5 Missing/extra exons in Mouse Secretory Carbonic Anhydrase

4.5.1 Car VI

Ensembl as well as UCSC genome browser have no transcripts/ESTs with missing/extra exon.

4.6 Missing/extra exons in Extracellular/Secreted Carbonic anhydrase in Zebrafish and Cow

Ensembl shows no evidence of missing or extra exons for these species.

4.7 Structural feasibility of Human carbonic anhydrase isoforms

4.7.1 CA IV

>Peptide sequence of Human_CaIV principal isoform with catalytic domain highlighted in gray

MRMLLALLALSAARPSASAE

SHWCYEVQAESSNYPCLV -2nd exon

PVKWGGNCQKDRQSPINIVTTKAKVDKKLGRFFFSGYDKKQTWTVQNNGHSV

MMLLENKASISGGGLPAPYQAKQLHLHWSDLPYKGSEHSLDGEHFAME

MHIVHEKEKGTSRNVKEAQDPEDEIAVLAFLVE

AGTQVNEGFQPLVEALSNIPKPE

MSTTMAESSLLDLLPKEEKLRHYFRYLGSLTTPTCDEKVVWTVFREPIQLHREQ

ILAFSQKLYYDKEQTVSMKDNVRPLQQLGQRTVIKSGAPGRPLPWALPALLGPMLACLL AGFLR

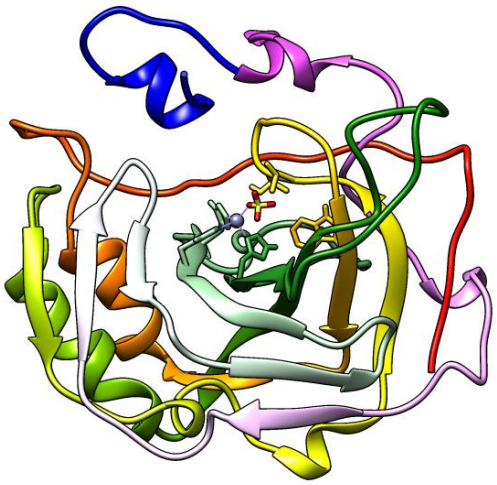


Figure 16 PDB structure of Human_CaIV (1ZNC) with 2nd exon highlighted in blue as viewed in Chimera. Residue 'S' and 'H' is missing in the highlighted area.

Second exon is the part of catalytic domain of the principal isoform so the addition of extra 2nd exon affects the structure of protein.

4.7.2 CA VI

Comparing the Carbonic anhydrase 6 principle isoform with its EST (supporting evidence with missing second exon/ NM_001270501.1)

>peptide sequence of Human_CAVI with catalytic domain highlighted in gray

MRALVLLLSLFLLGGQAQHVSDWTYSE

GALDEAHWPQHYPACGGQRQSPINLQRTKVRYNPSLKGLNMTGYETQAGEFPMVNNG
HTV

QISLPSTMRMTVADGTVYIAQQMHFHWGGASSEISGSEHTVDGIRHVIE

IHIVHYNSKYKSYDIAQDAPDGLAVLAAFVE

VKNYPENTYYSNFISHLANIKYPG

QRTTLTGLDVQDMLPRNLQHYYTYHGSLTTPPCTENVHWFVLADFVKLSRTQ

VWKLENSLLDHRNKTIHNDYRRTQPLNHRVVESNFPNQE

YTLGSEFQFYLHKIEEILDYLRRALN

Aligning the protein sequences of two isoforms shows that EST form is devoid of some catalytic residues (EGALDEAHWPQHYPACGGQRQSPINLQRTKVRYNPSLKGLNMTGYETQAGEFPMVNNGHT/27-86) which is important catalytic domain in principle isoform. It also has missing catalytic site residues QSP (47,48,49), H(85))

```
P23280 MRALVLLLSLFLLGGQAQHVSDWTYSEGALDEAHWPQHYPACGGQRQSPI
P23280 NLQRTKVRYNPSLKGLNMTGYETQAGEFPMVNNGHTVQISLPSTMRMTVA
P23280 DGTVYIAQQMHFHWGGASSEISGSEHTVDGIRHVIEIHIVHYNSKYKSYD
P23280 IAQDAPDGLAVLAAFVEVKNYPENTYYSNFISHLANIKYPGQRTTLTGLD
P23280 VQDMLPRNLQHYYTYHGSLTTPPCTENVHWFVLADFVKLSRTQVWKLENS
P23280 LLDHRNKTIHNDYRRTQPLNHRVVESNFPNQEYTLGSEFQFYLHKIEEIL
P23280 DYLRRALN
```

Figure 17 Highlighted area showing splice variant/ missing second exon in EST/isoform NM_001270501.1. (source: Chimera)

It lacks a sequence in position 26/C i.e. Cysteine which plays role in disulfide bond formation.

Also, it lacks Glycine in sequence position 67 which acts as a glycosylation site.

CA VI missing 2nd exon visualization:

Figure 18 PDB structure of CA VI (PDB id: 3FE4/chain A) with missing second exon highlighted in purple. QSP catalytic residue is the part of 2nd exon highlighted in blue.

The missing 2nd exon forms the major part of catalytic domain as seen clearly in fig. Hence, absence of 2nd exon in alternative transcript affects the structure of protein.

4.7.3 CA IX

Human CAIX exon structure with gray highlighted area representing catalytic domain:

>peptide sequence of human CA_IX principle isoform(ENST00000378357.9) with catalytic domain highlighted in gray

MAPLCPSPWLPLLIPAPAPGLTVQLLLSLLLLVPVHPQRLPRMQEDSPLGGGSSGEDDPL
GEEDLPSEEDSPREEDPPGEEDLPGEEDLPGEEDLPEVKPKSEEEGSLKLEDLPTVEAPG
DPQEPQNNAHRDKEG
DDQSHWRYGG
DPPWPRVSPACAGRFQSPVDIRPQLAAFCPALRPLELLGFQLPPLPELRLRNNGHSV
QLTLPPGLEMALGPGREYRALQLHLHWGAAGRPGSEHTVEGHRFPAE
IHVVHLSTAFARVDEALGRPGGLAVLAAFLEE
GPEENSAYEQLLSRLEEIAEEG
SETQVPGLDISALLPSDFSRYFQYEGSLTTPPCAQGVIWTVFNQTVMLSAKQ
LHTLSDTLWGPGDSRLQLNFRATQPLNGRVIEASFPAGVDSSPRAAEPV
QLNSCLAAG
DILALVFGLLFAVTSVAFLVQMRRQHR
RGTKGGVSYRPAEVAETGA

Second exon is missing in alternative transcript(**BX383092**). Second exon is the part of catalytic domain in principle isoform as shown in the image below.

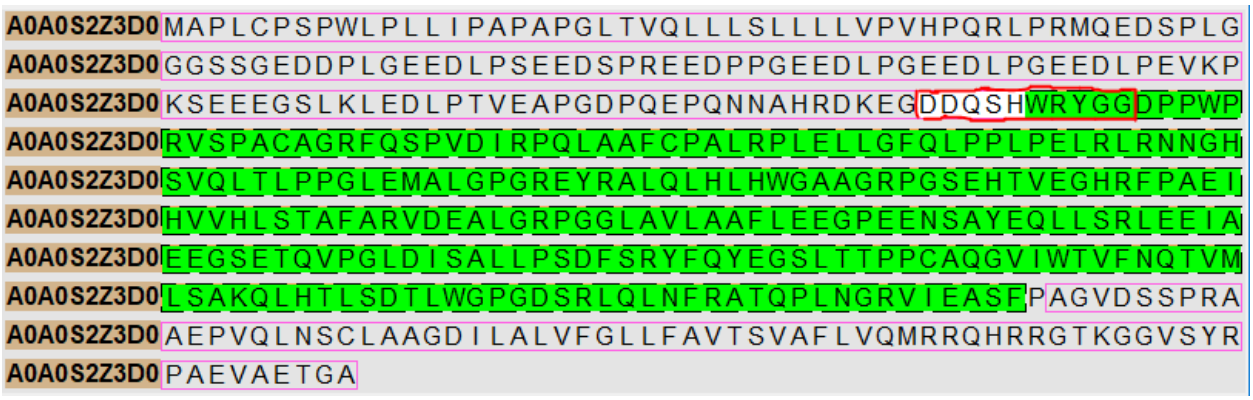


Figure 19 Protein sequence of principle isoform of CA IX. Red-box outline is second exon, green highlighted area is the catalytic domain. (source: chimera)

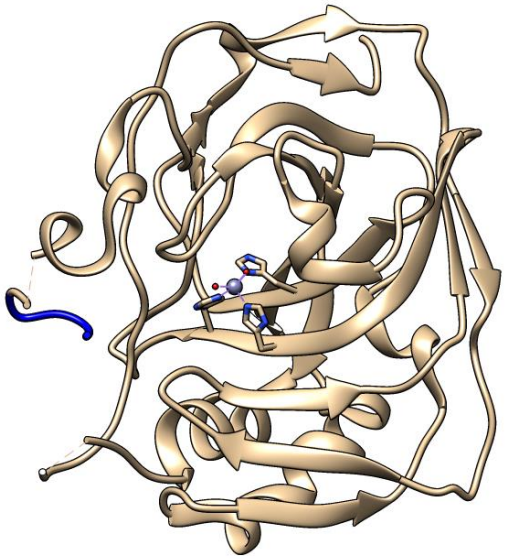


Figure 20 PDB structure of CA IX/chain A(6FE2) with second exon highlighted in blue as viewed in chimera.

PDB structure(6FE2) has no 5aa sequence (DDQSH) in second exon.

The second exon with 10aa is missing in alternative transcript. Major portion of catalytic domain is still intact along with QSP (catalytic site residue).

4.7.4 CA XII

>Peptide sequence of Human_CAXII principal isoform(ENST00000178638.8) with catalytic domain highlighted in gray

MPRRSLHAAAVLLLVILKEQPSSPAPVNG
SKWTYFG
PDGENSWSKKYPSCGGLLQSPIDLHSDILQYDASLTPLEFQGYNLSANKQFLLTNNGHSV
KLNLPSDMHIQGLQSRYSATQLHLHWGNPNDPHGSEHTVSGQHFAAE
LHIVHYNSDLYPDASTASNKSEGLAVLAVLIE
MGSFNPSYDKIFSHLQHVKYKG
QEAFVPGFNIEELLPERTAEYYRYRGSLTTPPCNPTVLWTVFRNPVQISQEQ
LLALETALYCTHMDDPSPREMINNFRQVQKFDERLVYTSFSQV
QVCTAAGLSLG- 9th exon
IILSLALAGILGICIVVVVSIWLFRRKS
IKKGDNKGVIYKPATKMETEAHA

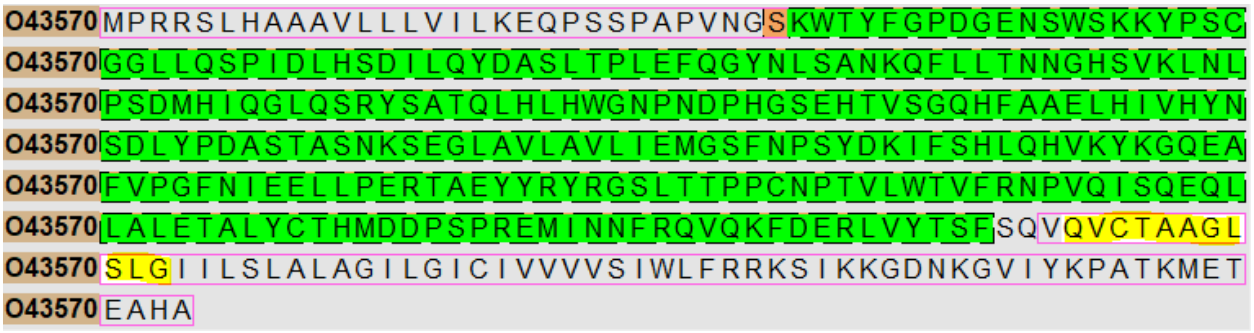


Figure 21 Peptide sequence of principle isoform viewed in chimera (ENST00000178638.8) with catalytic domain highlighted in green and ninth exon highlighted in yellow.

CA XII has Refseq cDNA transcript NM_206925.1 with missing ninth exon. The missing ninth exon in Refseq cDNA transcript NM_206925.1 is outside of catalytic domain. The peptide sequence is exactly same as ensemble transcript CA XII-202 with 343aa. Therefore, the functional protein structure is feasible for this alternate transcript.

Supporting evidence also shows cDNA record(CR618639) with missing 2nd exon for carbonic anhydrase 12.

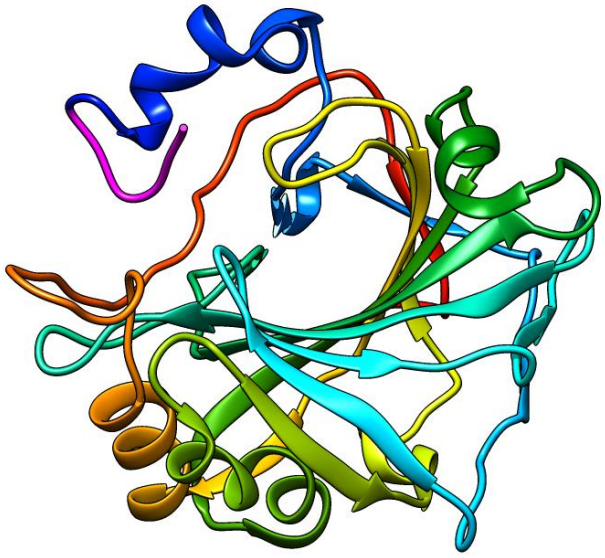


Figure 22 PDB structure of CA XII/chain A(4HT2) with second exon highlighted in magenta as viewed in chimera. Residue 'S' in 2nd exon is missing in the highlighted region.

The missing 2nd exon is the part of catalytic domain of principal isoform. The 2nd exon contains 'WTY' motif so absence of this exon affects the structural feasibility.

4.7.5 CA XIV

>Peptide sequence of Human_CAXIV principal isoform (ENST00000369111.9) with catalytic domain highlighted in gray

MLFSALLLEVIWILAADGG

QHWTYEG – **2nd exon**

PHGQDHWPASYPECGNNAQSPIDIQTDSVTFDPDLPALQPHGYDQPGTEPLDLHNNGHTV

QLSLPSTLYLGGLPRKYVAAQLHLHWGQKGSPGGSEHQINSEATFAE

LHIVHYDSDSYDSLSEAAERPQGLAVLGILIE

VGETKNIAYEHILSHLHEVRHKD

QKTSVPPFNLRELLPKQLGQYFRYNGSLTTPPCYQSVLWTVFYRRSQISMEQ

LEKLQGTLFSTEEEPSKLLVQNYRALQPLNQRMVFASFIQA

GSSYTTG

EMLSLGVGILVGCLCLLLAVYFIARKIR

KKRLENRKSVVFTSAQATTEA

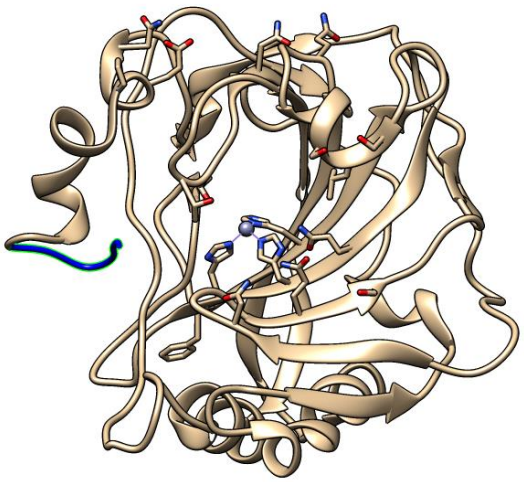


Figure 23 PDB structure of CA XIV(4LU3) with second exon highlighted in blue as viewed in chimera

Second exon with 'WTY' motif belongs to catalytic domain in principal isoform so absence of this exon leads to instability in protein structure.

4.8 Structural feasibility of Mouse carbonic anhydrase isoforms

4.8.1 Car IV

>Peptide sequence of Mouse_CarIV principal isoform (ENSMUST00000103194.9)

MQLLLALLALAYVAPSTED

SGWCYEIQTKDPRSSCLG -2nd exon

PEKWPGACKENQQSPINIVTART

KVNPRLTPFILVGYDQKQQWPIKNNQHTV

EMTLGGGACIIGGDLPARYEAVQLHLHWSNGNDNGSEHSIDGRHFAME

MHIVHKKLTSSKEDSKDKFAVLAFMIE

VGDKVNKGFQPLVEALPSISKPH

STSTVRESSLQDMLPPSTKMYTYFRYNGSLTTPNCDETVIWTVYKQPIKIHKNQ

FLEFSKNLYYDEDQKLNMKDNVRPLQPLGKRQVFKSHAPGQLLSLPLPTLLVPTLTCLV
ANFLQ

Supporting evidence in ensemble shows cDNA(S68245.1) with missing 2nd exon.

```
Q64444 MQLLLALLALAYVAPSTEDSGWCYEIQTKDPRSSCLGPEKWPGACKENQQ
Q64444 SPINIVTARTKVNPRLTPFILVGYDQKQQWPIKNNQHTVEMTLGGGACIIJ
Q64444 GGDLPARYEAVQLHLHWSNGNDNGSEHSIDGRHFAMEMHIVHKKLTSSKEJ
Q64444 DSKDKFAVLAFMIEVGDKVNKGFQPLVEALPSISKPHSTSTVRESSLQDMJ
Q64444 LPPSTKMYTYFRYNGSLTTPNCDETVIWTVYKQPIKIHKNQFLEFSKNLYJ
Q64444 YDEDQKLNMKDNVRPLQPLGKRQVFKSHAPGQLLSLPLPTLLVPTLTCLV
Q64444 ANFLQ
```

Figure 24 Catalytic domain of Mouse_CarIV principle isoform highlighted in orange with 2nd exon inside red-box viewed in chimera.

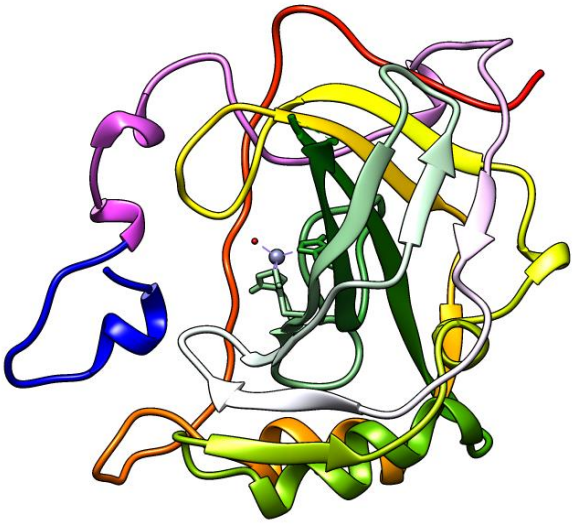


Figure 25 PDB structure of Mouse_CarIV (2ZNC) with 2nd exon highlighted in blue as viewed in chimera. Residue 'S' and 'G' is missing in the highlighted region.

The 2nd exon is the part of catalytic domain in principle isoform and contains residue that forms disulfide bond (residue 'C' in position 23 and residue 'C' in position 35). Thus missing of this exon affects the structure of protein.

4.8.2 Car IX (PDB structure missing)

>Peptide sequence of Mouse_CarIX principal isoform with catalytic domain highlighted in gray

MASLGPSPWAPLSTPAPTAQLLLFLLLQVSAQPQGLSGMQGEPSLGDSSSGEDELGVDV
L

PSEEDAPEEADPPDGEDPPEVNSEDRMEESLGLEDLSTPEAPEHSQGSHGDEKG

GGHSHWSYGG

TLLWPQVSPACAGRFQSPVDIRLERTAFCRTLQPLELLGYELQPLPELSLSNNGHTV

QLTLPPGLKMALGPGQEYRALQLHLHWGTSDHPGSEHTVNGHRFPAE

IHVVHLSTAFSELHEALGRPGGLAVLAAFLQ

ESPEENSAYEQLLSHLEEISEEG

SKIEIPGLDVSALLPSDLSRYYRYEGSLTTPPCSQGVIWTVFNETVKLSAKQ-**7th exon**

LHTLSVSLWGPRDSRLQLNFRATQPLNGRTIEASFPAAEDSSPEPV-**8th exon**

HVNSCFTAG

DILALVFGLLFAVTSIAFLLQLRRQHR

HRSGTKDRVSYSPAEMTETGA

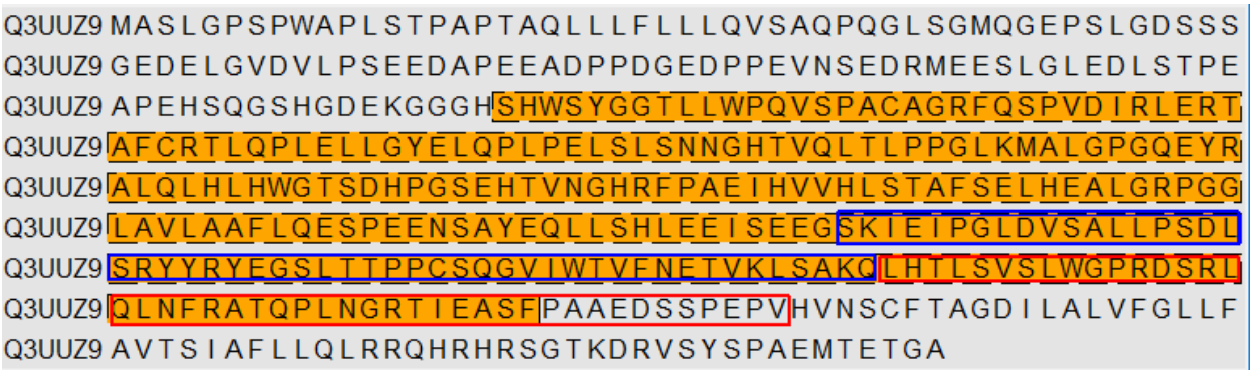


Figure 26 Peptide sequence of principal isoform of Mouse_CarIX with catalytic domain highlighted in orange as viewed in Chimera. Blue rectangular box represents 7th exon and red-box represents 8th exon.

The missing 7th and 8th exon in cDNA (EF122497) constitutes the major part of catalytic domain of principal isoform. The absence of these exons affect the structure of protein.

.

4.8.3 Car XIV (9th exon outside of PDB structure)

>Peptide sequence of Mouse_CarXIV principal isoform with catalytic domain highlighted in gray

MLFFALLLKVTWILAADGG

HHWTYEG

PHGQDHWPTSYPECGGDAQSPINIQTDSVIFDPDLPAVQPHGYDQLGTEPLDLHNNGHT
V

QLSLPPTLHLGGLPRKYTAAQLHLHWGQRGSLEGSEHQINSEATAAE

LHVVHYDSQSYSSLSEAAQKPQGLAVLGILIE

VGETENPAYDHILSRLHEIRYKD

QKTSVPPFSVRELFPQQLEQFFRYNGSLTTPPCYQSVLWTVFNRRAQISMGQ

LEKLQETLSSTEEDPSEPLVQNYRVPQPLNQRTIFASFIQA

GPLYTTG-9th exon

EMLGLGVGILAGCLCLLLAVYFIAQKIR

KKRLGNRKSVVFTSARATTEA

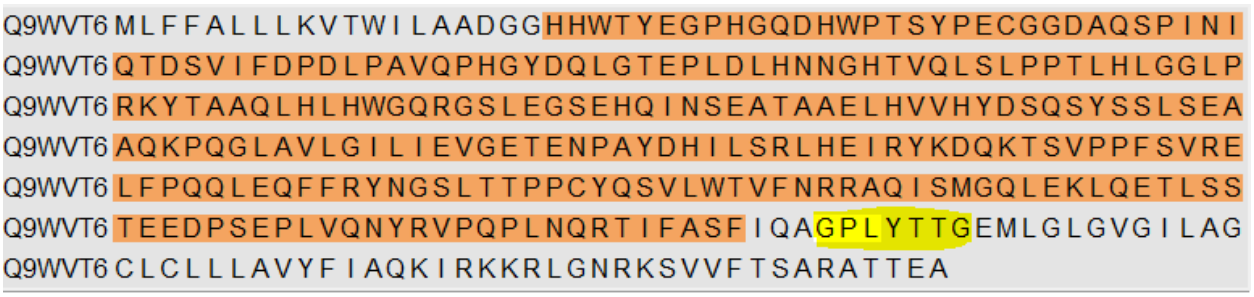


Figure 27 Peptide sequence of Mouse_CarXIV with orange highlighted area showing catalytic domain as viewed in Chimera. The yellow highlighted area represents 9th exon of principal isoform.

AK157764.1, BC046995.1 and AB005450.1 cDNAs has 9th exon missing. The 9th exon is outside main catalytic domain which signifies the feasibility of functional protein structure.

5 Discussion

EST/cDNA evidence in ensembl and genomic browser shows missing/extra exons in extracellular and secretory carbonic anhydrases. Some of these exons are within catalytic domain and some in the linker region after the catalytic domain and before the transmembrane helix. The homologous missing/extra exons is observed in human and mouse CAs to confirm its biological relevance. The lack of enough information relating this matter on other animals like Zebra-fish and cow hinders the efficacy of the study.

CA IV

CA IV EST shows extra second-exon in human. This exon is the part of catalytic domain of principal isoform so addition of extra exon in this region affects the structure and function of the protein.

Car IV

EST evidence shows missing second exon which is the part of catalytic domain in principal isoform and contains residue that forms disulfide bond (residue ‘C’ in position 23 and residue ‘C’ in position 35). Therefore, absence of this exon affects the protein structure.

CA VI

RefSeq entries NM_001270501.1 and NM_001270502.1 shows second and third exons missing respectively. Second and third exon forms the major portion of catalytic domain so absence of these exons affect the structure of protein product.

Car VI

This isoform has no missing exons.

CA IX

UCSC spliced-EST track shows missing second exon(10aa) in this transcript. The PDB structure (6FE2) shows 5aa (DDQSH) missing as compared with the second exon of principal isoform. Since the major portion of catalytic domain is intact along with QSP catalytic site residue, protein is structurally feasible.

Car IX

7th and 8th exons are missing according to EST evidence. These exons constitute the major part of catalytic domain which in turn affects the structure of protein. In human, CAIX alternative splicing isoform with missing exon 8-9 was detected in cancer cells (Malentacchi et al., 2009).

CA XII

Ensemble transcript evidence shows second and ninth exon missing. This isoform has alternative Refseq transcript NM_206925.1 with missing ninth exon. This exon is outside the catalytic domain and its peptide sequence is identical to transcript CA XII-202.Thus functional protein structure is possible.

The CA XII transcript with missing 9th exon (11aa) seems to be common in astrocytomas (type of cancer that forms in brain or spinal cord) (Malentacchi et al., 2009).

Another transcript evidence shows missing second exon which is the part of catalytic domain in principal isoform and contains 'WTY' motif. Absence of 'WTY' motif affects the structure of the final protein product.

Car XII

Ensembl and Genome Brower shows no missing exons.

CA XIV

Ensembl EST shows second exon missing in some cDNAs. This second exon contains 'WTY' motif and forms a part of catalytic domain therefore absence of this exon destabilizes the functionality of protein.

Car XIV

EST evidence shows 9th exon missing which is outside of the main catalytic domain therefore the structural feasibility is possible without this exon.

6 Conclusion

Many membrane-associated isoforms in human and mouse have EST/cDNA evidence that shows missing/extra exon in various exon positions unlike cytoplasmic and mitochondrial CAs. Membrane-associated transcripts like human CAXII and mouse Car XIV have missing 9^{th} exon between catalytic and transmembrane domains which implies the possibility of function-altering variant. Homologous missing/extra exons in human and mouse confirm the biological relevance of the findings to some extent. Further study is needed to fully understand and confirm the evolutionary significance.

References

Alterio, V., Hilvo, M., Di Fiore, A., Supuran, C.T., Pan, P., Parkkila, S., Scaloni, A., Pastorek, J., Pastorekova, S., Pedone, C., Scozzafava, A., Monti, S.M., De Simone, G., 2009. Crystal structure of the catalytic domain of the tumor-associated human carbonic anhydrase IX. Proc. Natl. Acad. Sci. 106, 16233–16238. https://doi.org/10.1073/pnas.0908301106

Alterio, V., Pan, P., Parkkila, S., Buonanno, M., Supuran, C.T., Monti, S.M., De Simone, G., 2014. The structural comparison between membrane-associated human carbonic anhydrases provides insights into drug design of selective inhibitors. Biopolymers 101, 769–778. https://doi.org/10.1002/bip.22456

Alvarez, B. V., Vilas, G.L., Casey, J.R., 2005. Metabolon disruption: A mechanism that regulates bicarbonate transport. EMBO J. 24, 2499–2511. https://doi.org/10.1038/sj.emboj.7600736

Ars, E., Serra, E., García, J., Kruyer, H., Gaona, A., Lázaro, C., Estivill, X., 2000. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum. Mol. Genet. 9, 237–247. https://doi.org/10.1093/hmg/9.2.237

Briganti, F., Mangani, S., Scozzafava, A., Vernaglione, G., Supuran, C.T., 1999. Carbonic anhydrase catalyzes cyanamide hydration to urea: Is it mimicking the physiological reaction? J. Biol. Inorg. Chem. 4, 528–536. https://doi.org/10.1007/s007750050375

Chakravarty, S., Kannan, K.K., 1994. Drug-protein interactions. Refined structures of three sulfonamide drug complexes of human carbonic anhydrase I enzyme. J. Mol. Biol. 243, 298–309. https://doi.org/10.1006/jmbi.1994.1655

Crocetti, L., Maresca, A., Temperini, C., Hall, R.A., Scozzafava, A., Mühlschlegel, F.A., Supuran, C.T., 2009a. A thiabendazole sulfonamide shows potent inhibitory activity against mammalian and nematode α-carbonic anhydrases. Bioorg. Med. Chem. Lett. 19, 1371–1375. https://doi.org/10.1016/j.bmcl.2009.01.038

Crocetti, L., Maresca, A., Temperini, C., Hall, R.A., Scozzafava, A., Mühlschlegel, F.A., Supuran, C.T., 2009b. A thiabendazole sulfonamide shows potent inhibitory activity against mammalian and nematode α-carbonic anhydrases. Bioorganic Med. Chem. Lett. 19, 1371–1375. https://doi.org/10.1016/j.bmcl.2009.01.038

Di Fiore, A., Monti, S.M., Hilvo, M., Parkkila, S., Romano, V., Scaloni, A., Pedone, C., Scozzafava, A., Supuran, C.T., De Simone, G., 2009. Crystal structure of human carbonic anhydrase XIII and its complex with the inhibitor acetazolamide 74, 164–175. https://doi.org/10.1002/prot.22144

Eriksson, A.E., Jones, T.A., Liljas, A., 1988. Refined structure of human carbonic anhydrase II at 2.0 Å resolution. Proteins Struct. Funct. Bioinforma. 4, 274–282. https://doi.org/10.1002/prot.340040406

Esbaugh, A.J., Tufts, B.L., 2006. The structure and function of carbonic anhydrase isozymes in

the respiratory system of vertebrates. Respir. Physiol. Neurobiol. 154, 185–198. https://doi.org/10.1016/J.RESP.2006.03.007

Esbaugh, A.J.J., Tufts, B.L.L., 2006. The structure and function of carbonic anhydrase isozymes in the respiratory system of vertebrates. Respir. Physiol. Neurobiol. 154, 185–198. https://doi.org/10.1016/J.RESP.2006.03.007

Ferraroni, M., Tilli, S., Briganti, F., Chegwidden, W.R., Supuran, C.T., Wiebauer, K.E., Tashian, R.E., Scozzafava, A., 2002. Crystal structure of a zinc-activated variant of human carbonic anhydrase I, CA I Michigan 1: Evidence for a second zinc binding site involving arginine coordination. Biochemistry 41, 6237–6244. https://doi.org/10.1021/bi0120446

Ghigna, C., Valacca, C., Biamonti, G., 2008. Alternative splicing and tumor progression. Curr. Genomics 9, 556–70. https://doi.org/10.2174/138920208786847971

Graveley, B.R., 2005. Mutually Exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures. Cell 123, 65–73. https://doi.org/10.1016/j.cell.2005.07.028

Imtaiyaz Hassan, M., Shajee, B., Waheed, A., Ahmad, F., Sly, W.S., 2013. Structure, function and applications of carbonic anhydrase isozymes. Bioorganic Med. Chem. 21, 1570–1582. https://doi.org/10.1016/j.bmc.2012.04.044

Kannan, K.K., Liljas, A., Waara, I., Bergstén, P.C., Lövgren, S., Strandberg, B., Bengtsson, U., Carlbom, U., Fridborg, K., Järup, L., Petef, M., 1972. Crystal structure of human erythrocyte carbonic anhydrase C. VI. The three-dimensional structure at high resolution in relation to other mammalian carbonic anhydrases. Cold Spring Harb. Symp. Quant. Biol. 36, 221–231. https://doi.org/10.1101/SQB.1972.036.01.030

KANNAN, K.K., RAMANADHAM, M., JONES, T.A., 1984. Structure, Refinement, and Function of Carbonic Anhydrase Isozymes: Refinement of Human Carbonic Anhydrase I. Ann. N. Y. Acad. Sci. 429, 49–60. https://doi.org/10.1111/j.1749-6632.1984.tb12314.x

Kim, C.Y., Whittington, D.A., Chang, J.S., Liao, J., May, J.A., Christianson, D.W., 2002. Structural aspects of isozyme selectivity in the binding of inhibitors to carbonic anhydrases II and IV. J. Med. Chem. 45, 888–893. https://doi.org/10.1021/jm010163d

Kim, H., Klein, R., Majewski, J., Ott, J., 2004. Estimating rates of alternative splicing in mammals and invertebrates. Nat. Genet. 36, 915–916. https://doi.org/10.1038/ng0904-915

Köhler, K., Hillebrecht, A., Schulze Wischeler, J., Innocenti, A., Heine, A., Supuran, C.T., Klebe, G., 2007a. Saccharin Inhibits Carbonic Anhydrases: Possible Explanation for its Unpleasant Metallic Aftertaste. Angew. Chemie Int. Ed. 46, 7697–7699. https://doi.org/10.1002/anie.200701189

Köhler, K., Hillebrecht, A., Schulze Wischeler, J., Innocenti, A., Heine, A., Supuran, C.T., Klebe, G., 2007b. Saccharin inhibits carbonic anhydrases: Possible explanation for its unpleasant metallic aftertaste. Angew. Chemie - Int. Ed. 46, 7697–7699. https://doi.org/10.1002/anie.200701189

Kumar, V., Kannan, K.K., 1994. Enzyme-substrate interactions structure of human carbonic anhydrase I complexed with bicarbonate. J. Mol. Biol. 241, 226–232.

https://doi.org/10.1006/jmbi.1994.1491

Lehtonen, J., Shen, B., Vihinen, M., Casini, A., Scozzafava, A., Supuran, C.T., Parkkila, A.K., Saarnio, J., Kivelä, A.J., Waheed, A., Sly, W.S., Parkkila, S., 2004. Characterization of CA XIII, a Novel Member of the Carbonic Anhydrase Isozyme Family. J. Biol. Chem. 279, 2719–2727. https://doi.org/10.1074/jbc.M308984200

Malentacchi, F., Simi, L., Nannelli, C., Andreani, M., Janni, A., Pastorekova, S., Orlando, C., 2009. Alternative splicing variants of carbonic anhydrase IX in human non-small cell lung cancer. Lung Cancer 64, 271–276. https://doi.org/10.1016/j.lungcan.2008.10.001

Maresca, A., Temperini, C., Vu, H., Pham, N.B., Poulsen, S.A., Scozzafava, A., Quinn, R.J., Supuran, C.T., 2009. Non-zinc mediated inhibition of carbonic anhydrases: Coumarins are a new class of suicide inhibitors. J. Am. Chem. Soc. 131, 3057–3062. https://doi.org/10.1021/ja809683v

Montgomery, J.C., Venta, P.J., Eddy, R.L., Fukushima, Y.S., Shows, T.B., Tashian, R.E., 1991. Characterization of the human gene for a newly discovered carbonic anhydrase, CA VII, and its localization to chromosome 16. Genomics 11, 835–848. https://doi.org/10.1016/0888-7543(91)90006-Z

Nagao, Y., Platero, J.S., Waheed, A., Sly, W.S., 1993. Human mitochondrial carbonic anhydrase: cDNA cloning, expression, subcellular localization, and mapping to chromosome 16. Proc. Natl. Acad. Sci. U. S. A. 90, 7623–7627. https://doi.org/10.1073/pnas.90.16.7623

Stams, T., Nair, S.K., Okuyama, T., Waheed, A., Sly, W.S., Christianson, D.W., 1996. Crystal structure of the secretory form of membrane-associated human carbonic anhydrase IV at 2.8-Å resolution. Proc. Natl. Acad. Sci. U. S. A. 93, 13589–13594. https://doi.org/10.1073/pnas.93.24.13589

Supuran, C.T., Simone, G. De, 2015. Carbonic Anhydrases as Biocatalysts From Theory to Medical and Industrial Applications. https://doi.org/10.1016/B978-0-444-63258-6/00022-6

Temperini, C., Innocenti, A., Guerri, A., Scozzafava, A., Rusconi, S., Supuran, C.T., 2007a. Phosph(on)ate as a zinc-binding group in metalloenzyme inhibitors: X-ray crystal structure of the antiviral drug foscarnet complexed to human carbonic anhydrase I. Bioorganic Med. Chem. Lett. 17, 2210–2215. https://doi.org/10.1016/j.bmcl.2007.01.113

Temperini, C., Innocenti, A., Guerri, A., Scozzafava, A., Rusconi, S., Supuran, C.T., 2007b. Phosph(on)ate as a zinc-binding group in metalloenzyme inhibitors: X-ray crystal structure of the antiviral drug foscarnet complexed to human carbonic anhydrase I. Bioorg. Med. Chem. Lett. 17, 2210–2215. https://doi.org/10.1016/j.bmcl.2007.01.113

Temperini, C., Innocenti, A., Scozzafava, A., Mastrolorenzo, A., Supuran, C.T., 2007c. Carbonic anhydrase activators: l-Adrenaline plugs the active site entrance of isozyme II, activating better isoforms I, IV, VA, VII, and XIV. Bioorganic Med. Chem. Lett. 17, 628–635. https://doi.org/10.1016/j.bmcl.2006.11.027

Temperini, C., Scozzafava, A., Vullo, D., Supuran, C.T., 2006. Carbonic anhydrase activators. Activation of isozymes I, II, IV, VA, VII, and XIV with L- and D-histidine and

crystallographic analysis of their adducts with isoform II: Engineering proton-transfer processes within the active site of an enzyme. Chem. - A Eur. J. 12, 7057–7066. https://doi.org/10.1002/chem.200600159

Van Karnebeek, C.D., Sly, W.S., Ross, C.J., Salvarinova, R., Yaplito-Lee, J., Santra, S., Shyr, C., Horvath, G.A., Eydoux, P., Lehman, A.M., Bernard, V., Newlove, T., Ukpeh, H., Chakrapani, A., Preece, M.A., Ball, S., Pitt, J., Vallance, H.D., Coulter-Mackie, M., Nguyen, H., Zhang, L.H., Bhavsar, A.P., Sinclair, G., Waheed, A., Wasserman, W.W., Stockler-Ipsiroglu, S., 2014. Mitochondrial carbonic anhydrase VA deficiency resulting from CA5A alterations presents with hyperammonemia in early childhood. Am. J. Hum. Genet. 94, 453–461. https://doi.org/10.1016/j.ajhg.2014.01.006

WANG, Y., LIU, J., HUANG, B., XU, Y.-M., LI, J., HUANG, L.-F., LIN, J., ZHANG, J., MIN, Q.-H., YANG, W.-M., WANG, X.-Z., 2015. Mechanism of alternative splicing and its regulation. Biomed. Reports 3, 152–158. https://doi.org/10.3892/br.2014.407

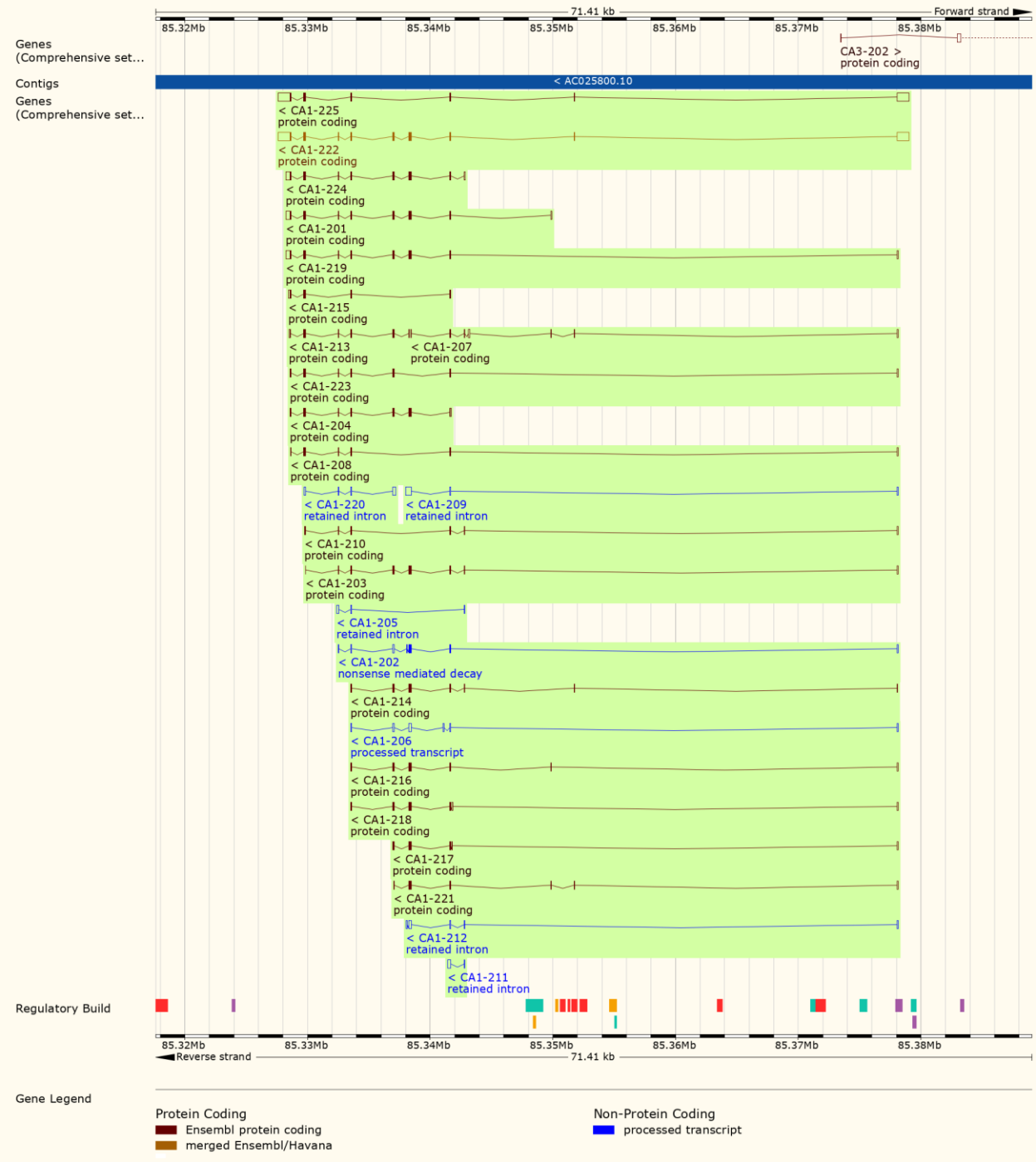
Whittington, D.A., Waheed, A., Ulmasov, B., Shah, G.N., Grubb, J.H., Sly, W.S., Christianson, D.W., 2001. Crystal structure of the dimeric extracellular domain of human carbonic anhydrase XII, a bitopic membrane protein overexpressed in certain cancer tumor cells. Proc. Natl. Acad. Sci. 98, 9545–9550. https://doi.org/10.1073/pnas.161301298

Yang, Z., Alvarez, B. V., Chakarova, C., Jiang, L., Karan, G., Frederick, J.M., Zhao, Y., Sauvé, Y., Li, X., Zrenner, E., Wissinger, B., Den Hollander, A.I., Katz, B., Baehr, W., Cremers, F.P., Casey, J.R., Bhattacharya, S.S., Zhang, K., 2005. Mutant carbonic anhydrase 4 impairs pH regulation and causes retinal photoreceptor degeneration. Hum. Mol. Genet. 14, 255–265. https://doi.org/10.1093/hmg/ddi023

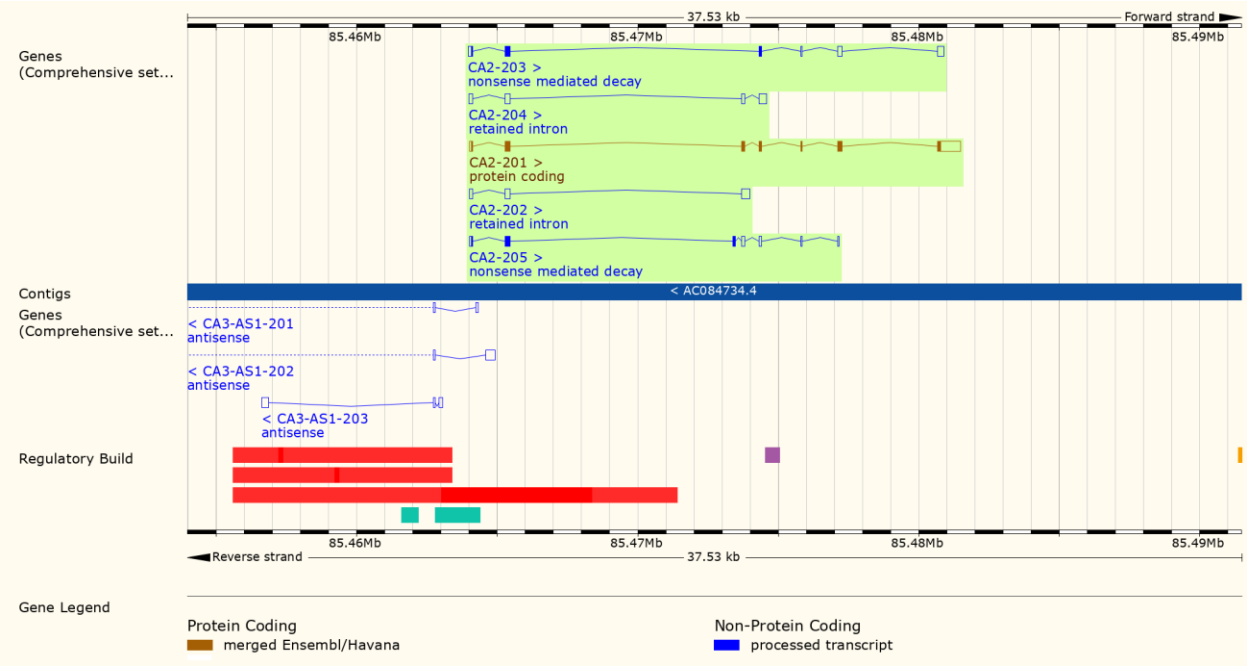
Appendix

Transcript Graphics of Human Carbonic Anhydrase Isoforms

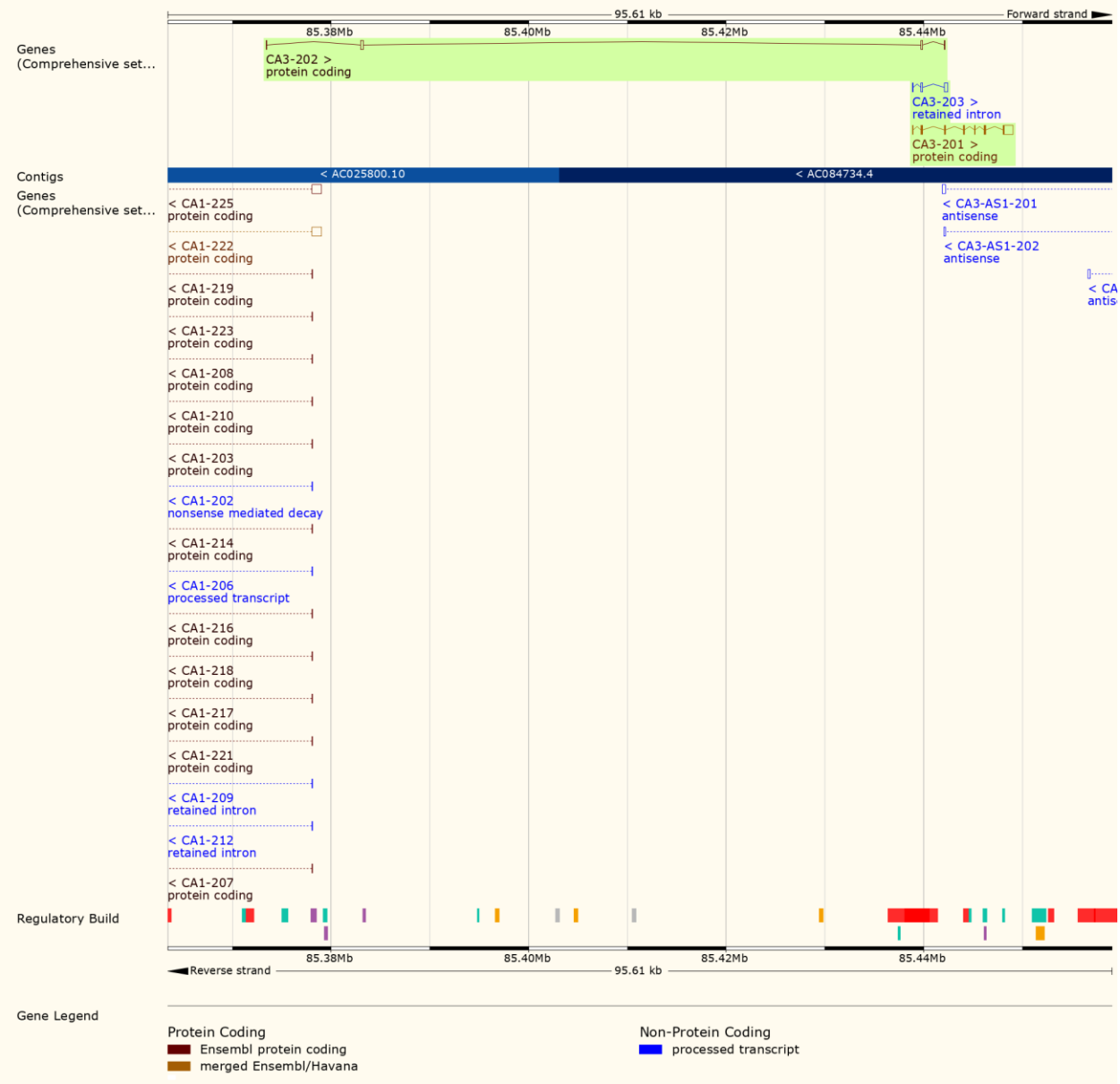
Carbonic anhydrase I



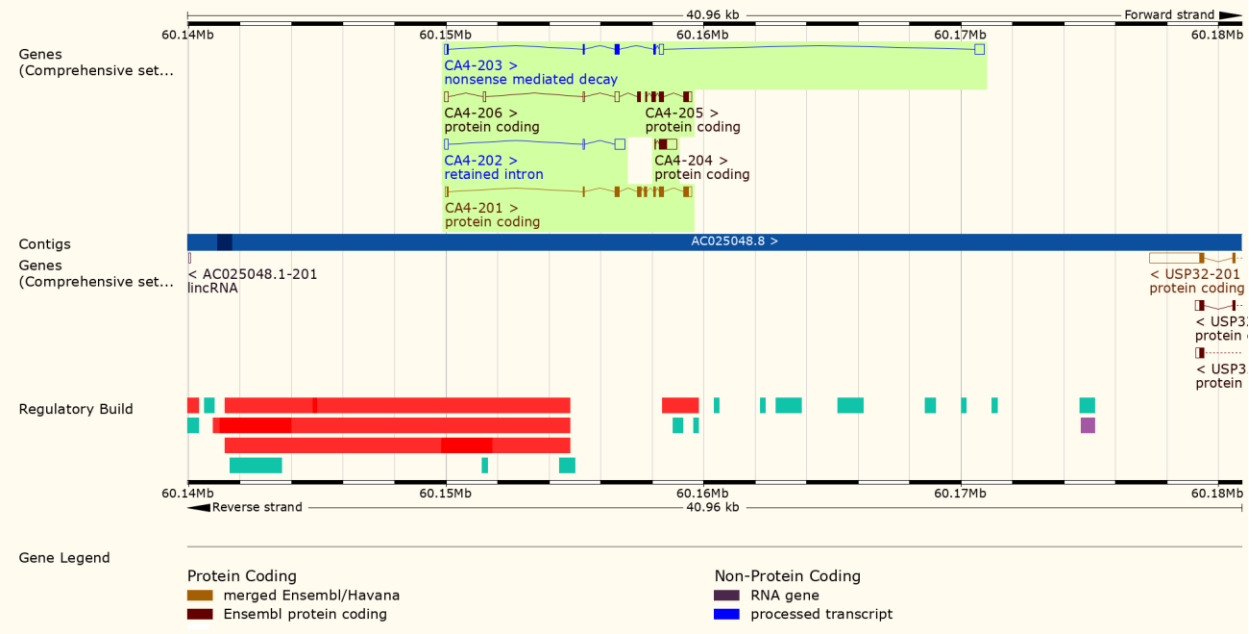
Carbonic Anhydrase II



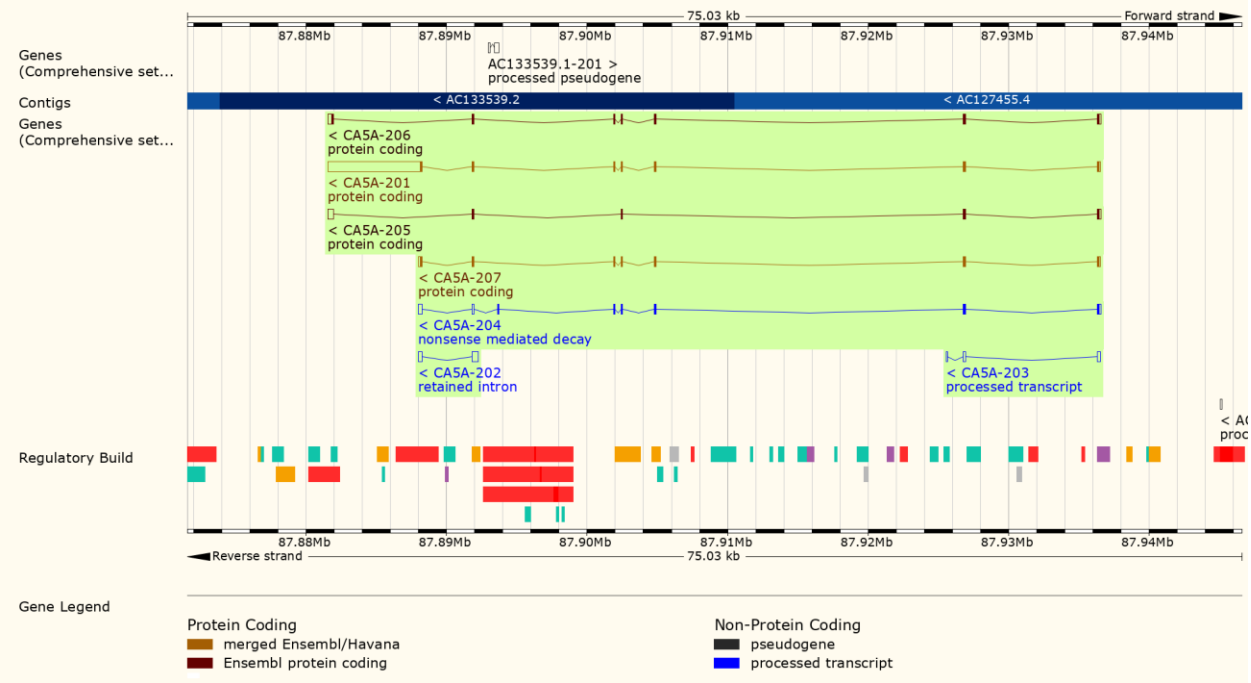
Carbonic Anhydrase III



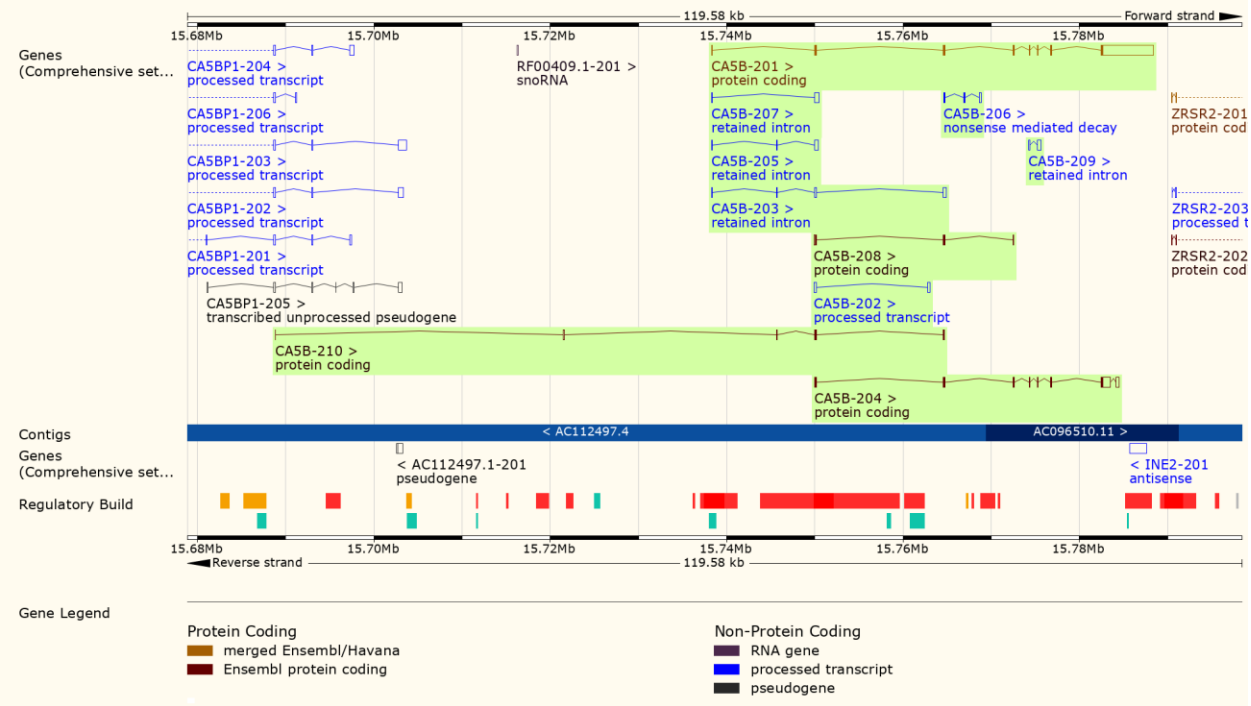
Carbonic anhydrase IV



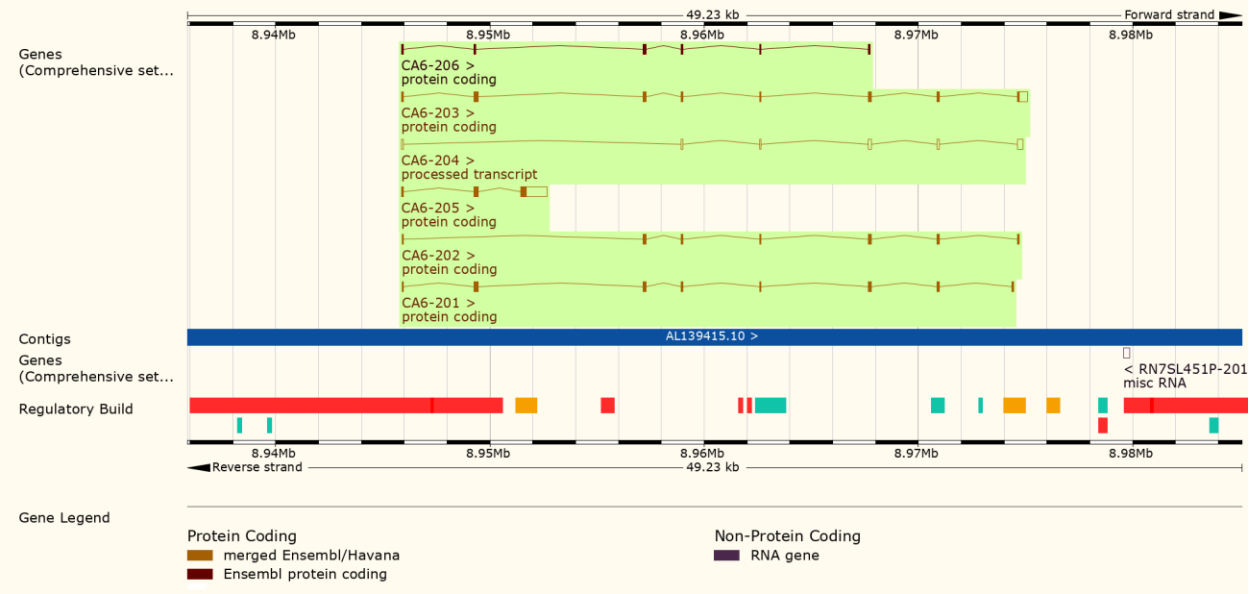
Carbonic Anhydrase VA



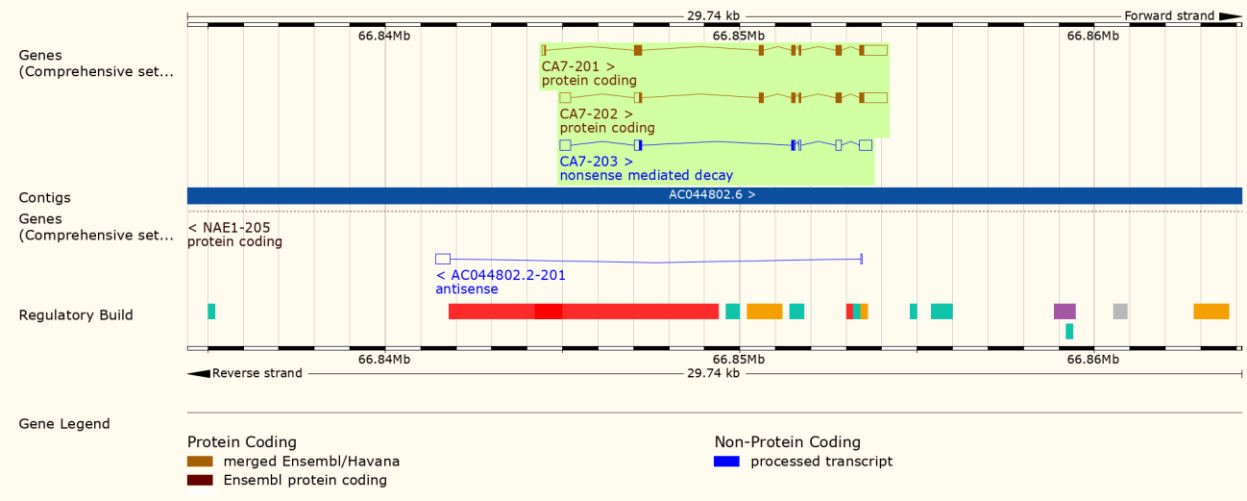
Carbonic Anhydrase VB



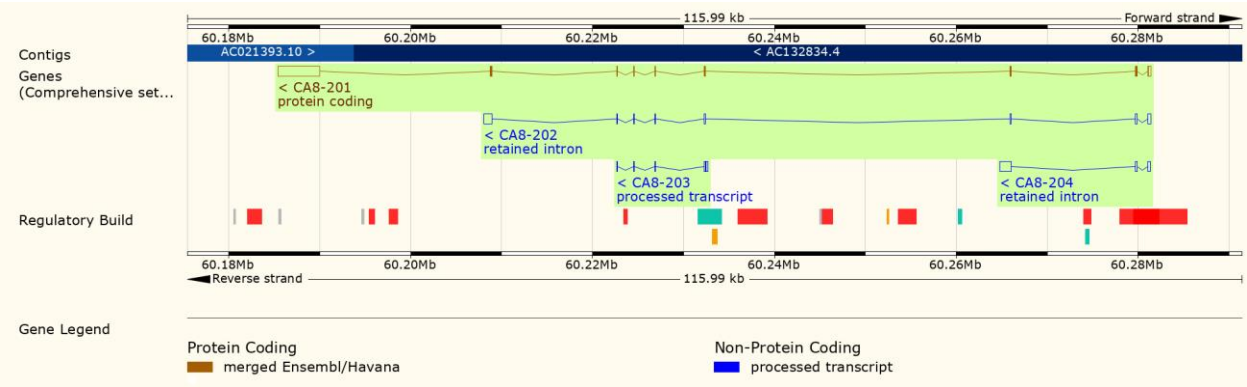
Carbonic Anhydrase VI



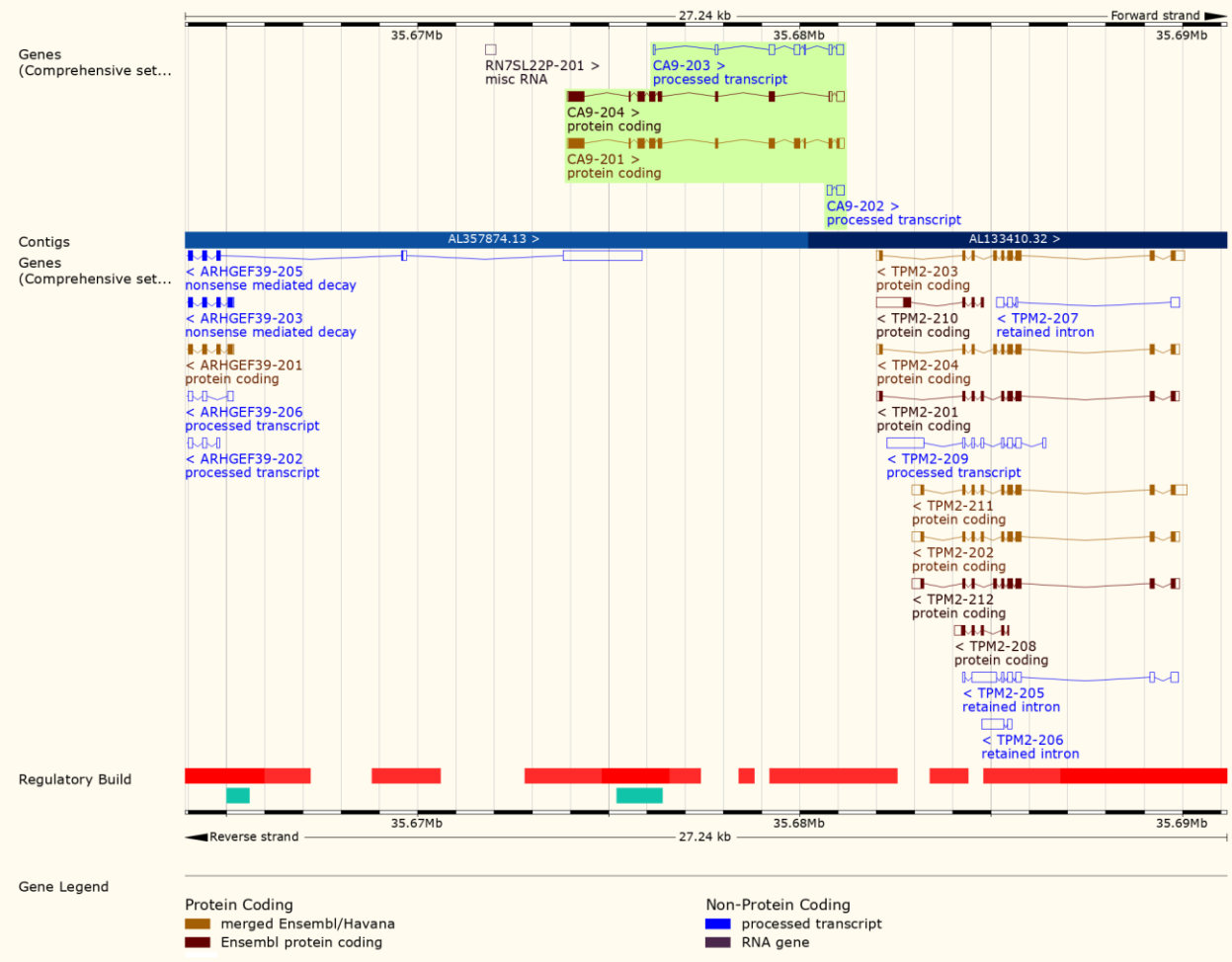
Carbonic Anhydrase VII



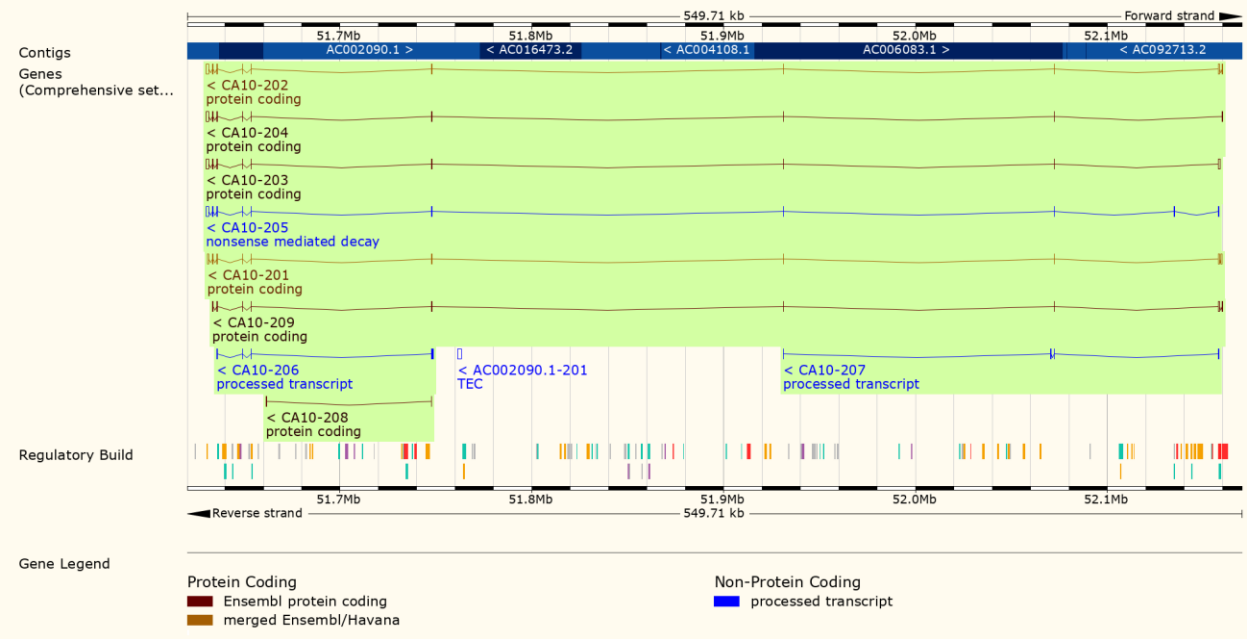
Carbonic Anhydrase VIII



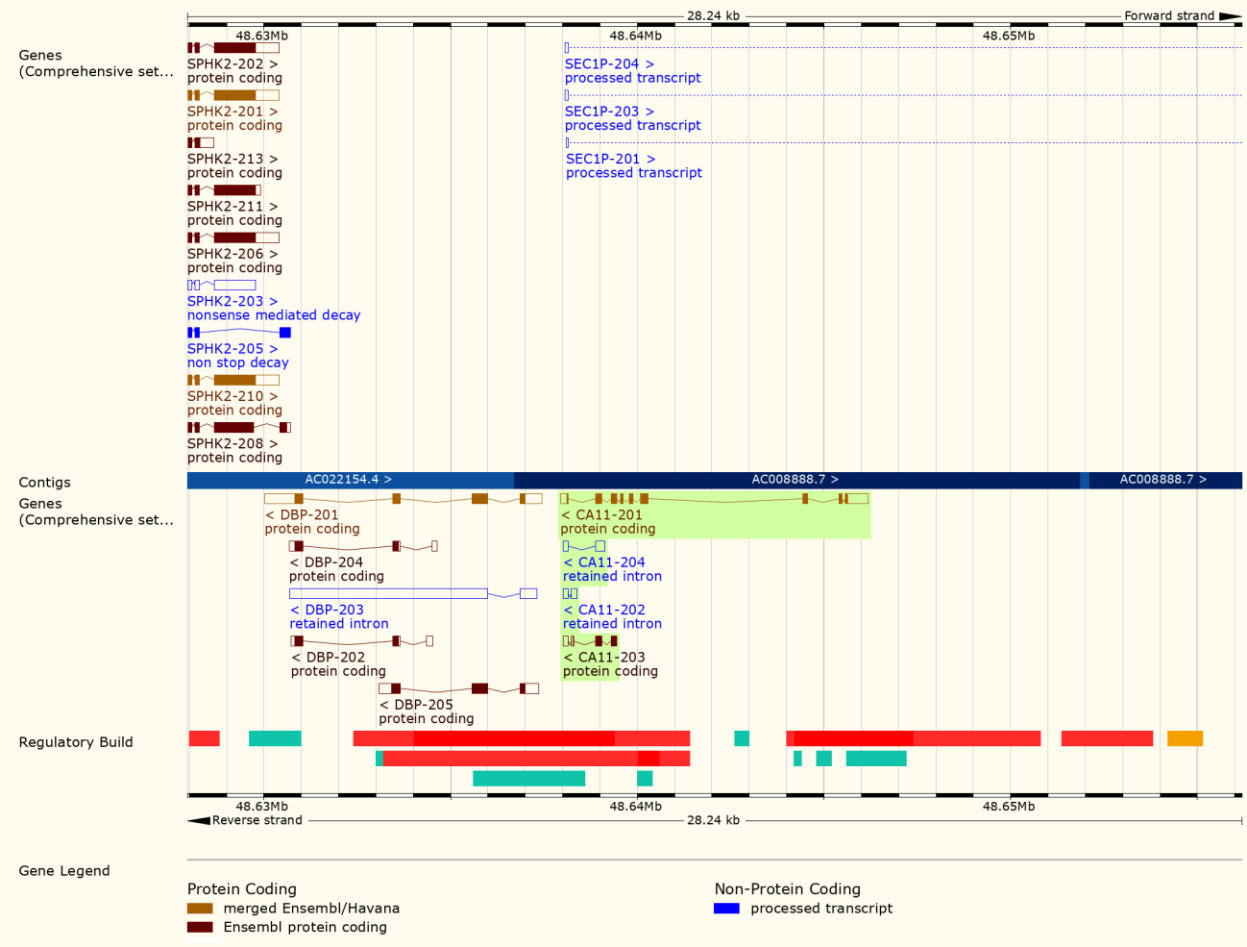
Carbonic Anhydrase IX



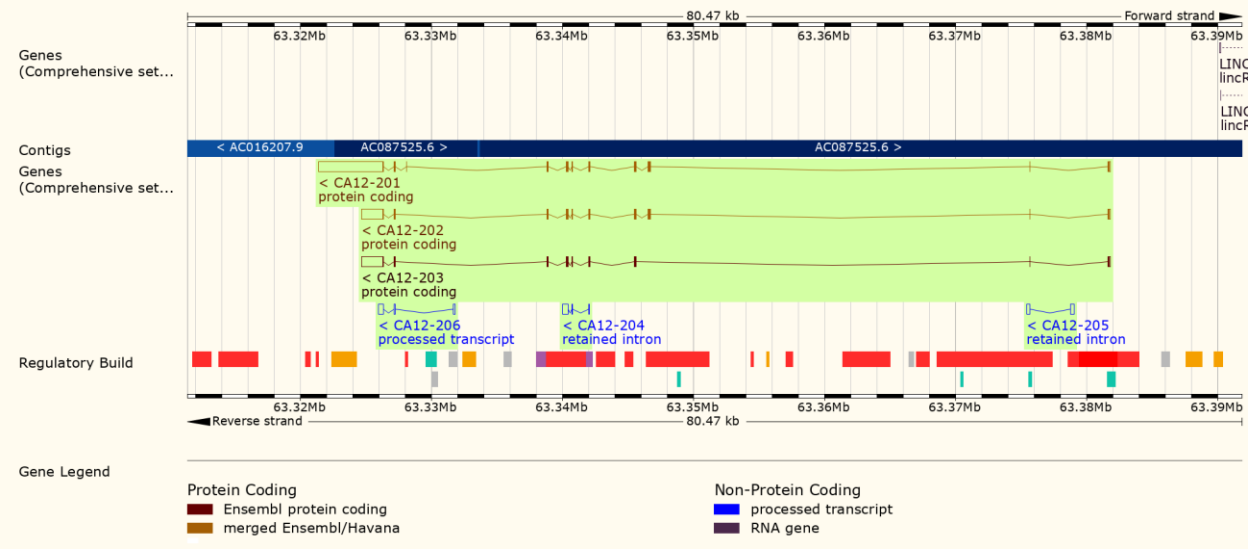
Carbonic Anhydrase X



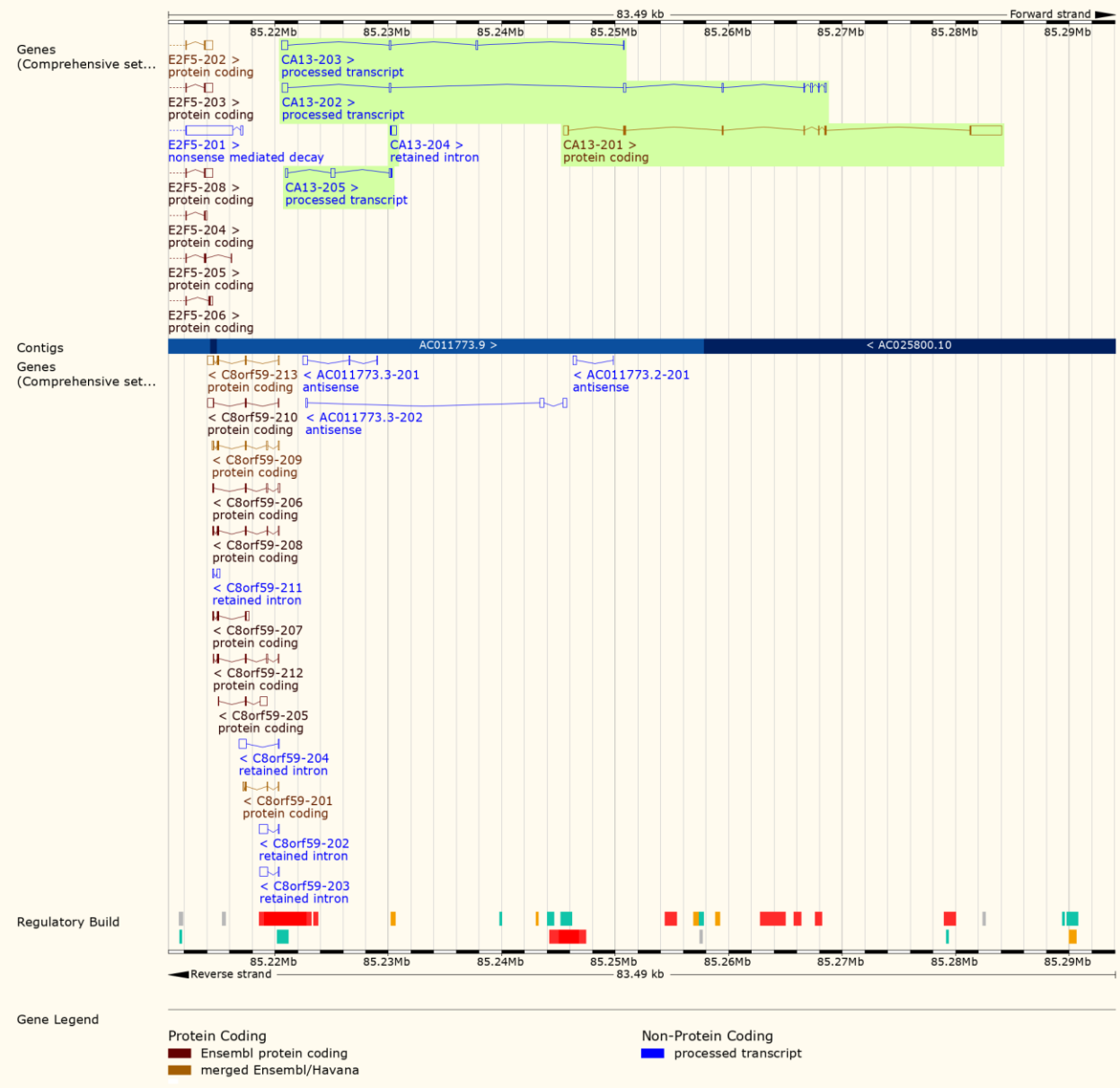
Carbonic Anhydrase XI



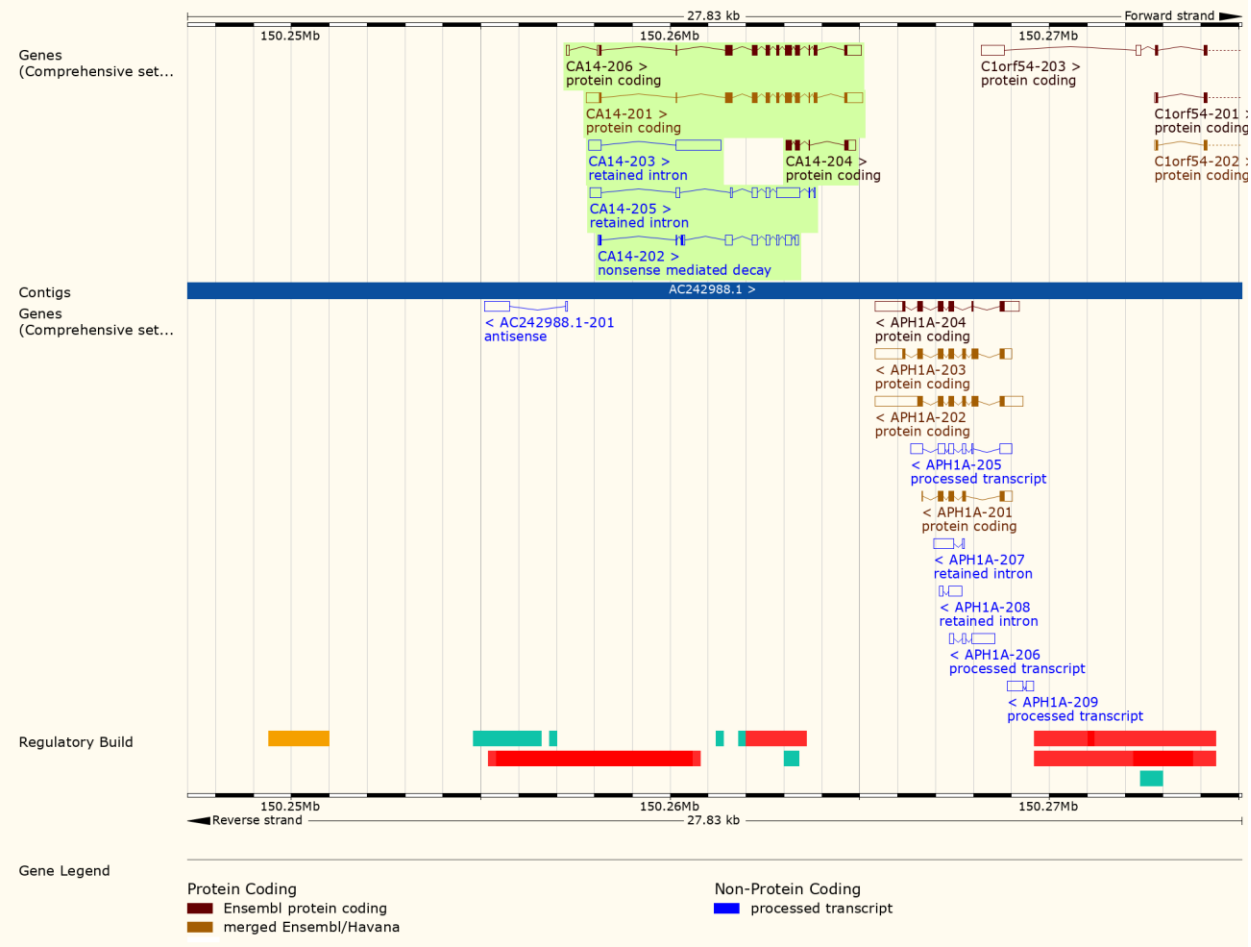
Carbonic Anhydrase XII



Carbonic Anhydrase XIII

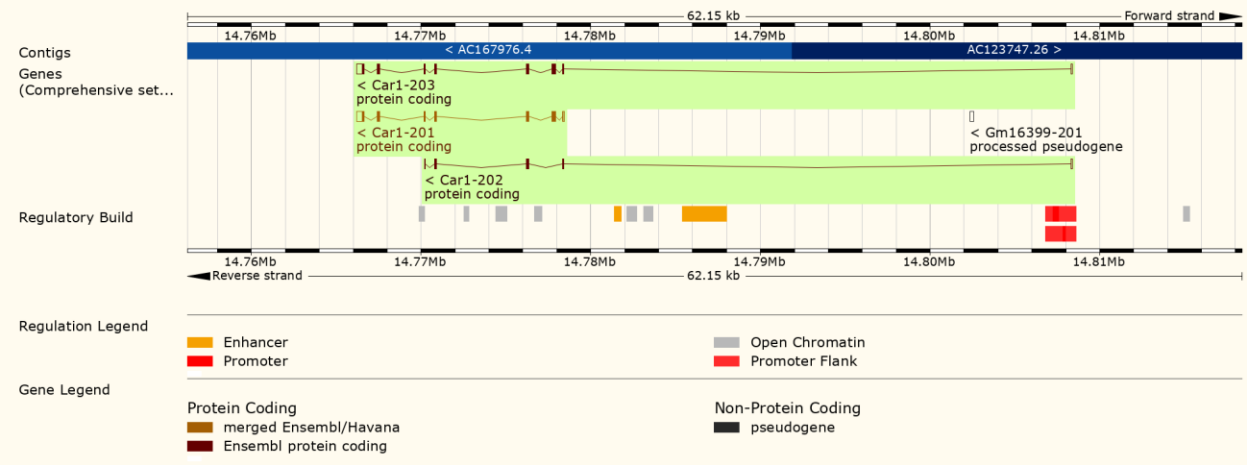


Carbonic Anhydrase XIV

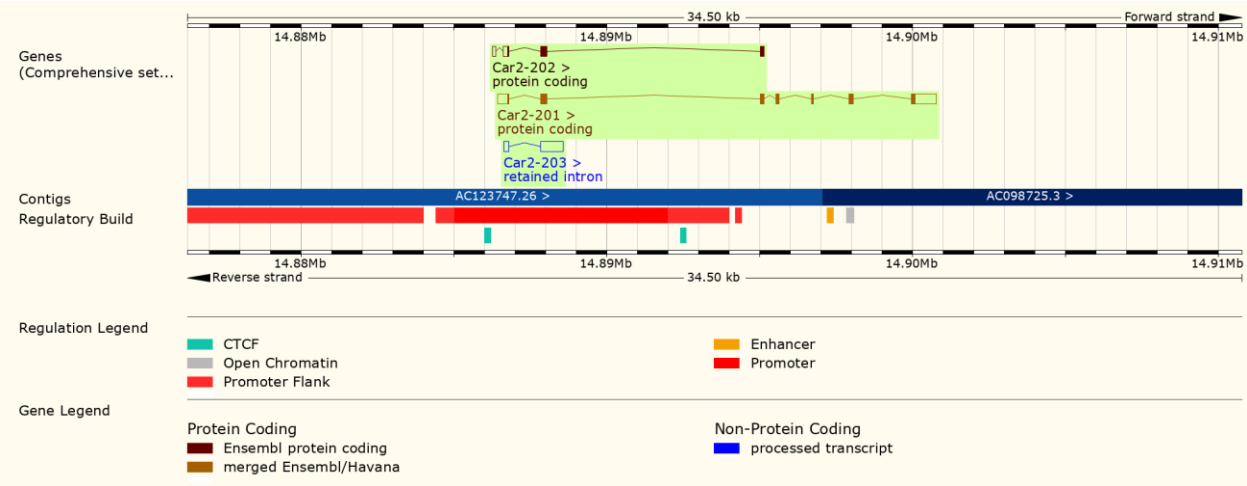


Transcript Graphics of Mouse Carbonic Anhydrase Isoforms

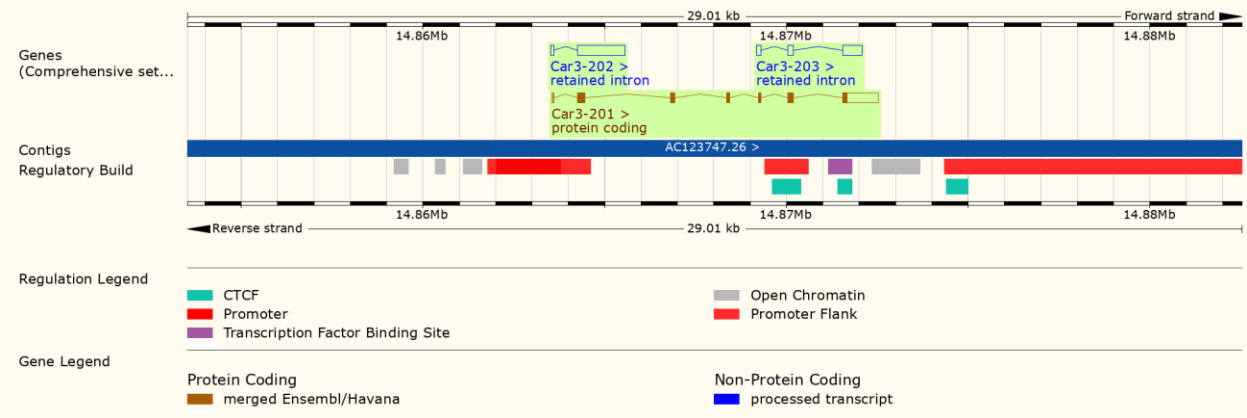
Car I



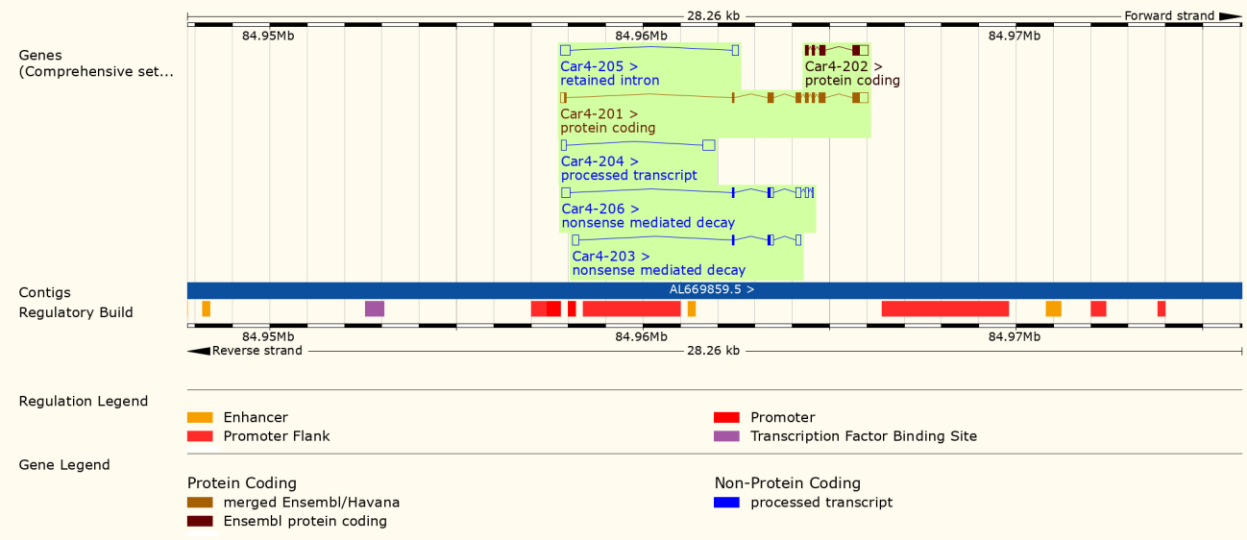
Car II



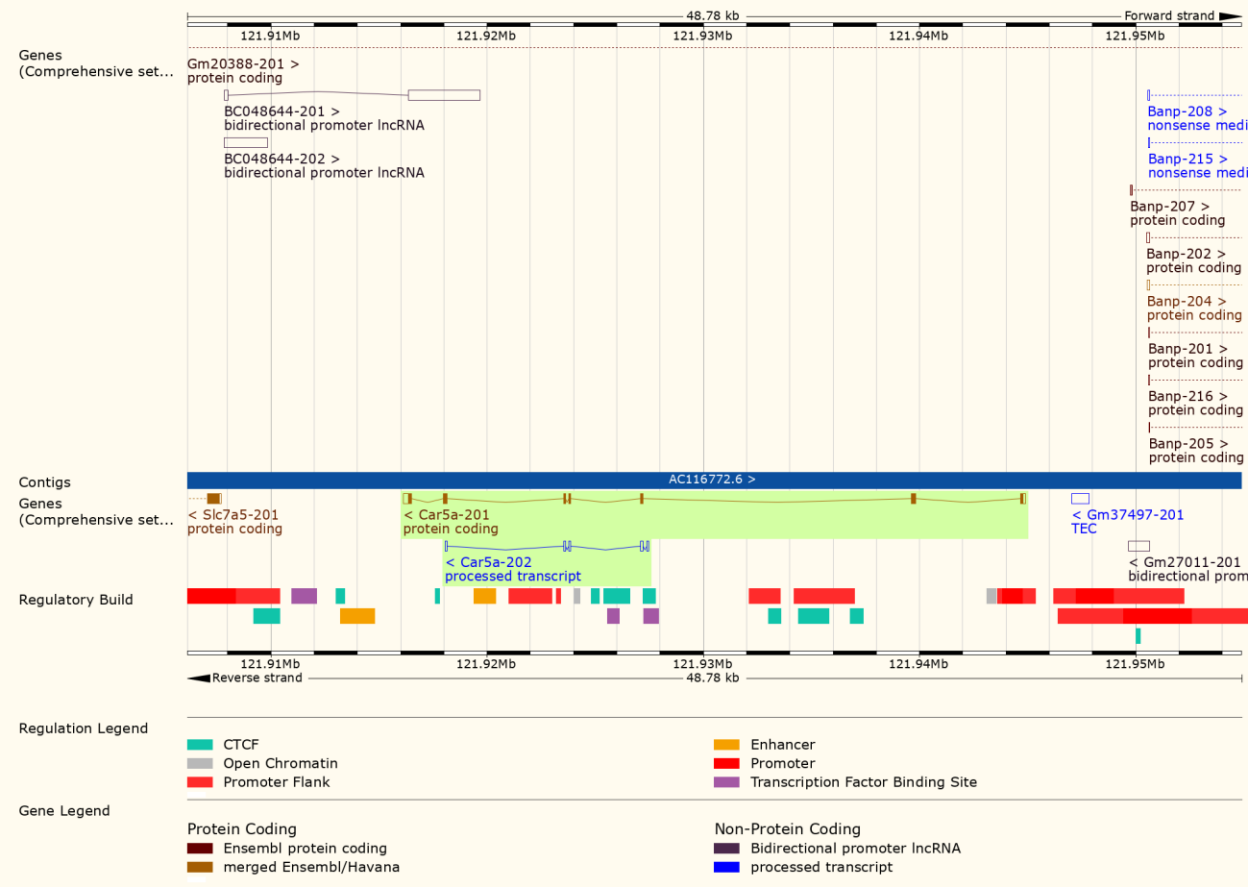
Car III



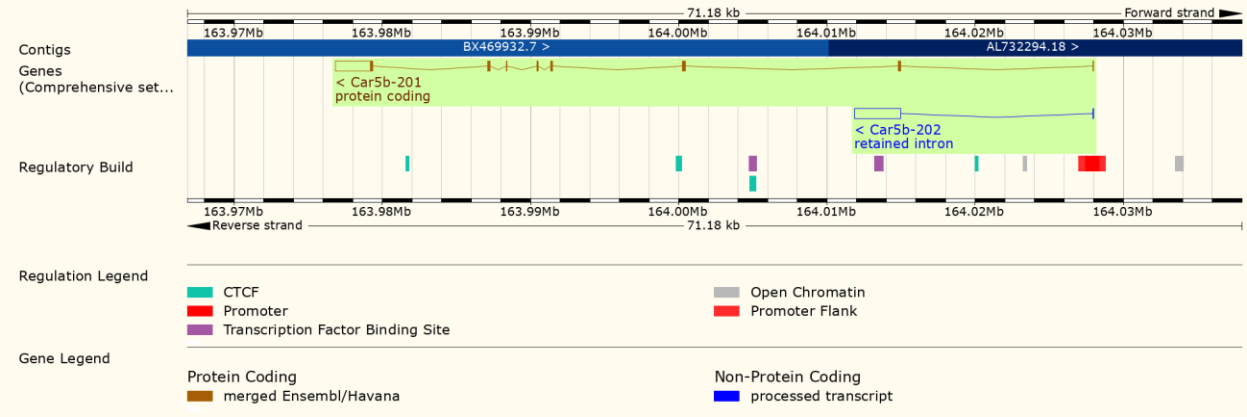
Car IV



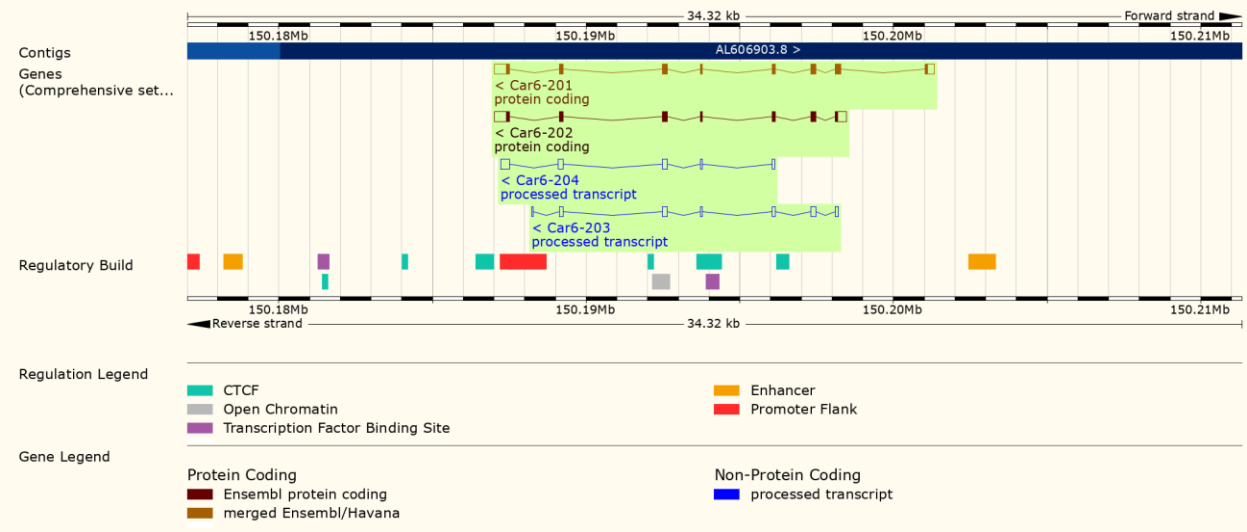
Car VA


48.78 kb
Forward strand
121.91Mb
121.92Mb
121.93Mb
121.94Mb
121.95Mb
Genes
(Comprehensive set...
Gm20388-201 >
protein coding
BC048644-201 >
bidirectional promoter lncRNA
BC048644-202 >
bidirectional promoter lncRNA
Banp-208 >
nonsense medi
Banp-215 >
nonsense medi
Banp-207 >
protein coding
Banp-202 >
protein coding
Banp-204 >
protein coding
Banp-201 >
protein coding
Banp-216 >
protein coding
Banp-205 >
protein coding
Contigs
AC116772.6 >
Genes
(Comprehensive set...
< Slc7a5-201
protein coding
< Car5a-201
protein coding
< Car5a-202
processed transcript
< Gm37497-201
TEC
< Gm27011-201
bidirectional prom
Regulatory Build
121.91Mb
121.92Mb
121.93Mb
121.94Mb
121.95Mb
Reverse strand
48.78 kb
Regulation Legend
CTCF
Open Chromatin
Promoter Flank
Enhancer
Promoter
Transcription Factor Binding Site
Gene Legend
Protein Coding
Ensembl protein coding
merged Ensembl/Havana
Non-Protein Coding
Bidirectional promoter lncRNA
processed transcript

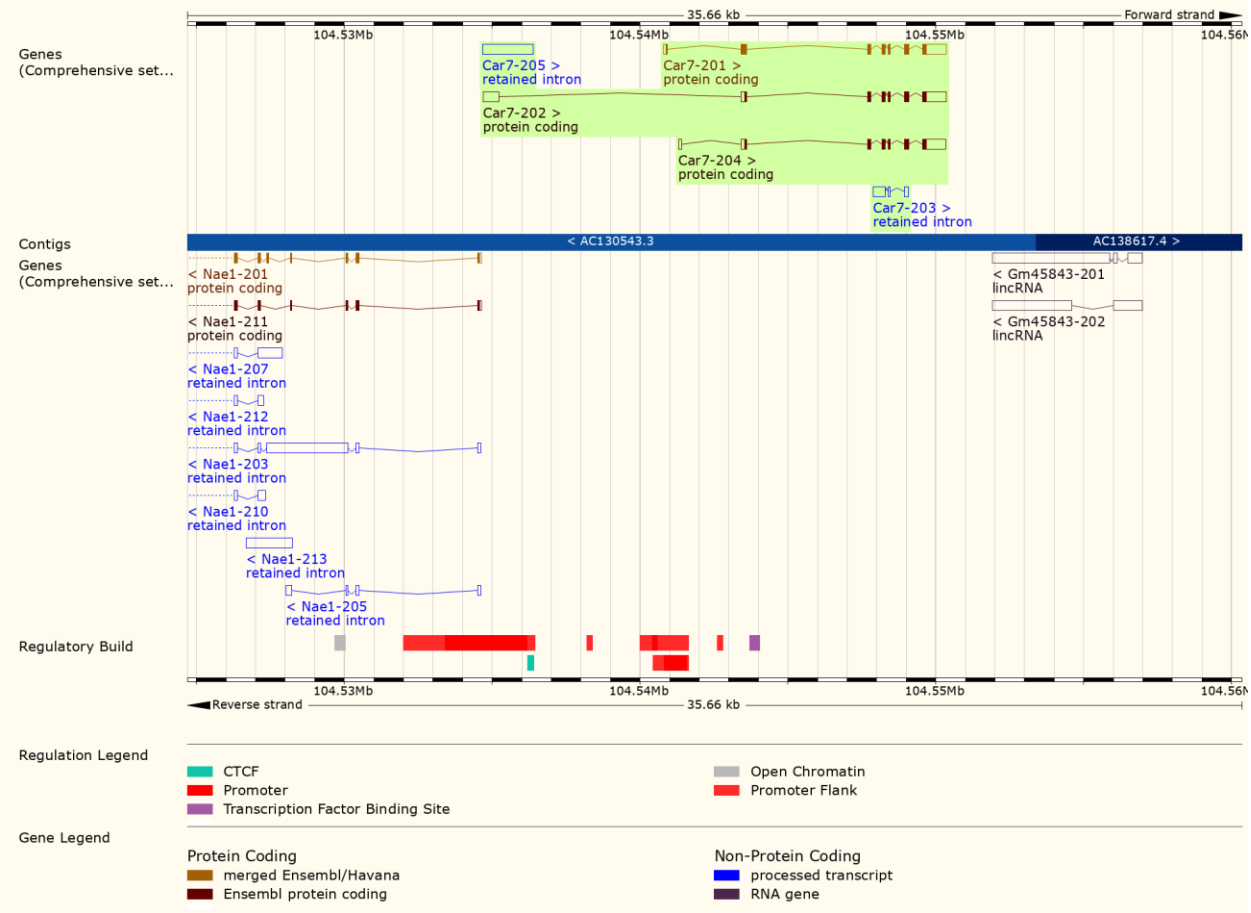
Car VB


71.18 kb
Forward strand
163.97Mb
163.98Mb
163.99Mb
164.00Mb
164.01Mb
164.02Mb
164.03Mb
Contigs
BX469932.7 >
AL732294.18 >
Genes
(Comprehensive set...
< Car5b-201
protein coding
< Car5b-202
retained intron
Regulatory Build
163.97Mb
163.98Mb
163.99Mb
164.00Mb
164.01Mb
164.02Mb
164.03Mb
Reverse strand
71.18 kb
Regulation Legend
CTCF
Promoter
Transcription Factor Binding Site
Open Chromatin
Promoter Flank
Gene Legend
Protein Coding
merged Ensembl/Havana
Non-Protein Coding
processed transcript

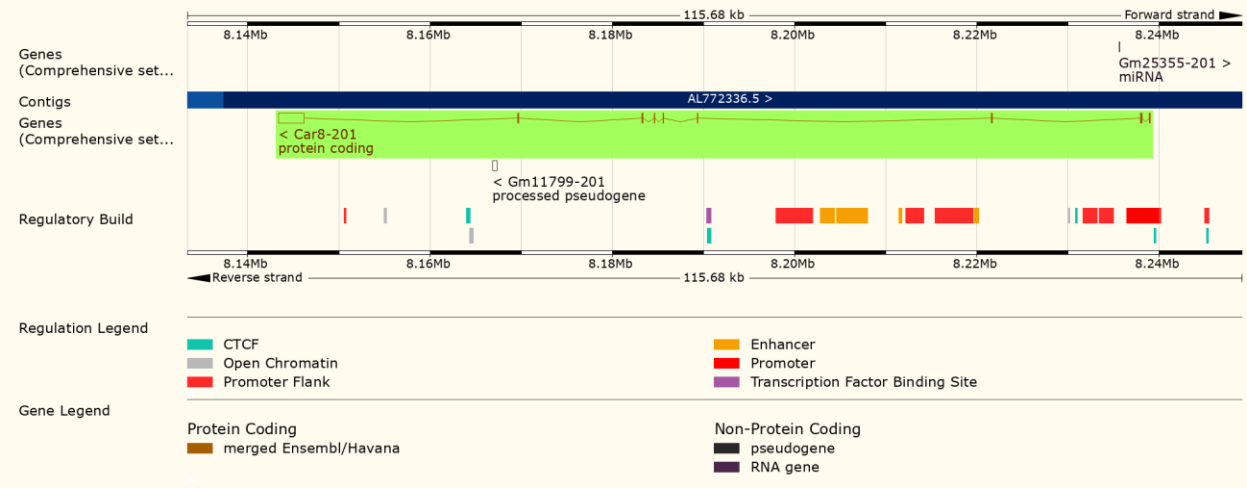
Car VI


34.32 kb
Forward strand
150.18Mb
150.19Mb
150.20Mb
150.21Mb
Contigs
AL606903.8 >
Genes
(Comprehensive set...
< Car6-201
protein coding
< Car6-202
protein coding
< Car6-204
processed transcript
< Car6-203
processed transcript
Regulatory Build
150.18Mb
150.19Mb
150.20Mb
150.21Mb
Reverse strand
34.32 kb
Regulation Legend
CTCF
Open Chromatin
Transcription Factor Binding Site
Enhancer
Promoter Flank
Gene Legend
Protein Coding
Ensembl protein coding
merged Ensembl/Havana
Non-Protein Coding
processed transcript

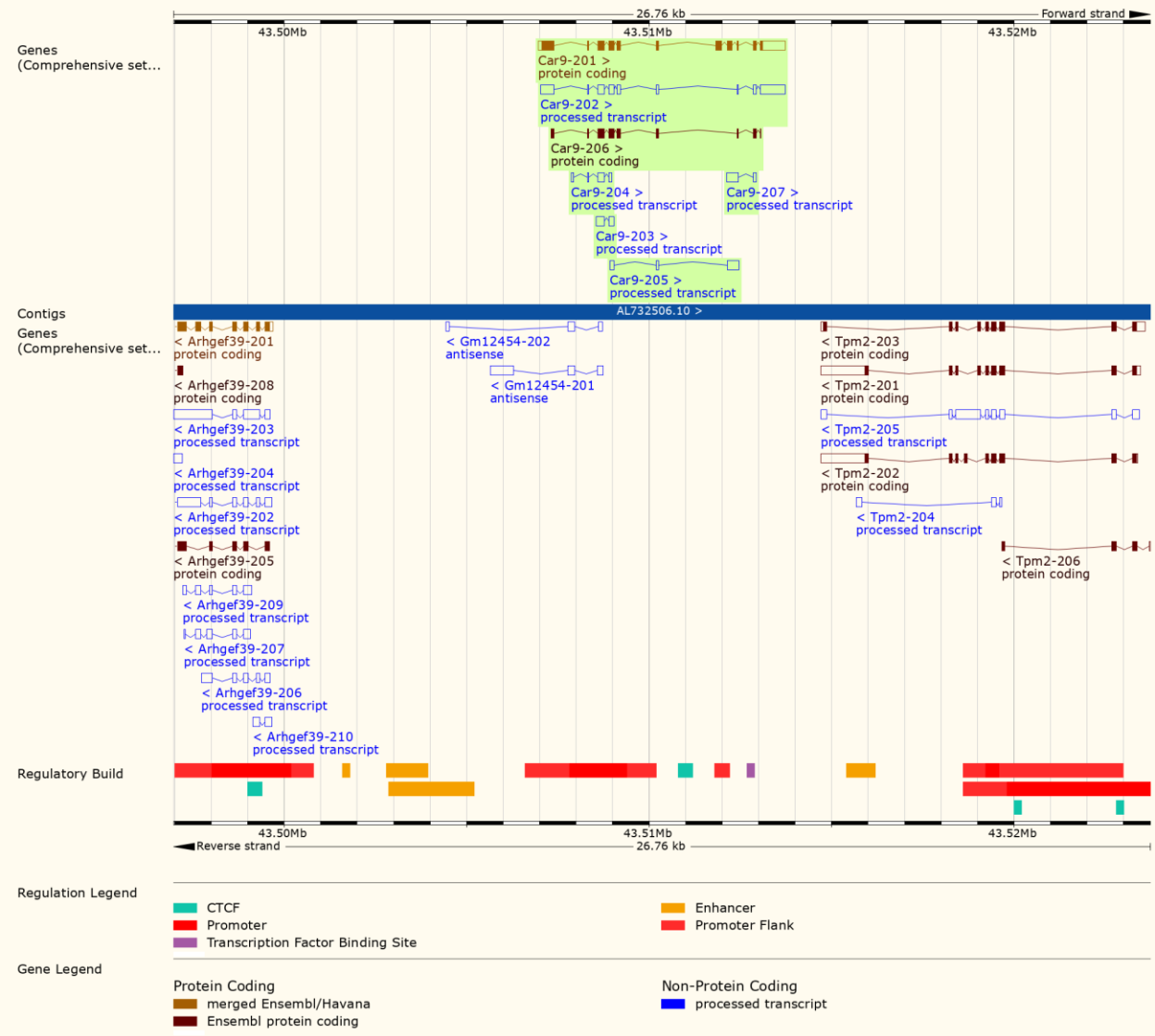
Car VII


35.66 kb
Forward strand
104.53Mb
104.54Mb
104.55Mb
104.56Mb
Genes
(Comprehensive set...
Car7-205 >
retained intron
Car7-201 >
protein coding
Car7-202 >
protein coding
Car7-204 >
protein coding
Car7-203 >
retained intron
Contigs
< AC130543.3
AC138617.4 >
Genes
(Comprehensive set...
< Nae1-201
protein coding
< Nae1-211
protein coding
< Nae1-207
retained intron
< Nae1-212
retained intron
< Nae1-203
retained intron
< Nae1-210
retained intron
< Nae1-213
retained intron
< Nae1-205
retained intron
< Gm45843-201
lincRNA
< Gm45843-202
lincRNA
Regulatory Build
104.53Mb
104.54Mb
104.55Mb
104.56Mb
Reverse strand
35.66 kb
Regulation Legend
CTCF
Promoter
Transcription Factor Binding Site
Open Chromatin
Promoter Flank
Gene Legend
Protein Coding
merged Ensembl/Havana
Ensembl protein coding
Non-Protein Coding
processed transcript
RNA gene

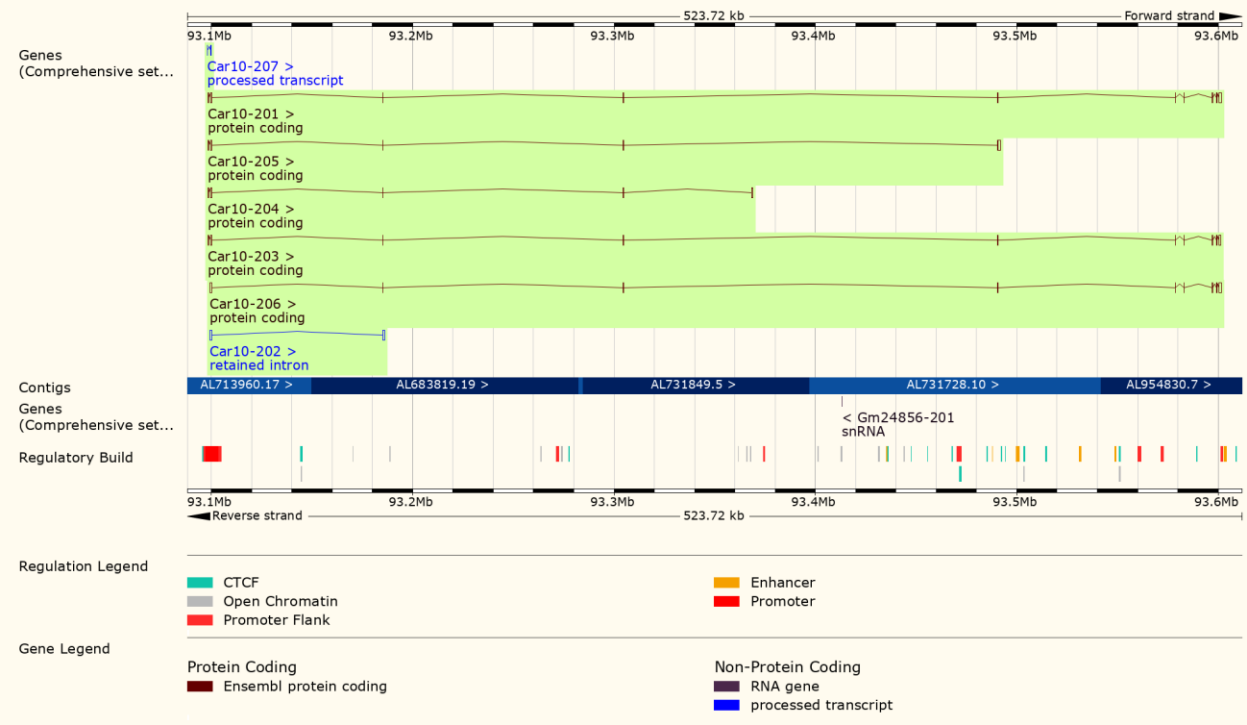
Car VIII


115.68 kb
Forward strand
8.14Mb
8.16Mb
8.18Mb
8.20Mb
8.22Mb
8.24Mb
Genes (Comprehensive set...
Gm25355-201 > miRNA
Contigs
AL772336.5 >
Genes (Comprehensive set...
< Car8-201 protein coding
< Gm11799-201 processed pseudogene
Regulatory Build
Reverse strand
Regulation Legend
CTCF
Open Chromatin
Promoter Flank
Enhancer
Promoter
Transcription Factor Binding Site
Gene Legend
Protein Coding
merged Ensembl/Havana
Non-Protein Coding
pseudogene
RNA gene

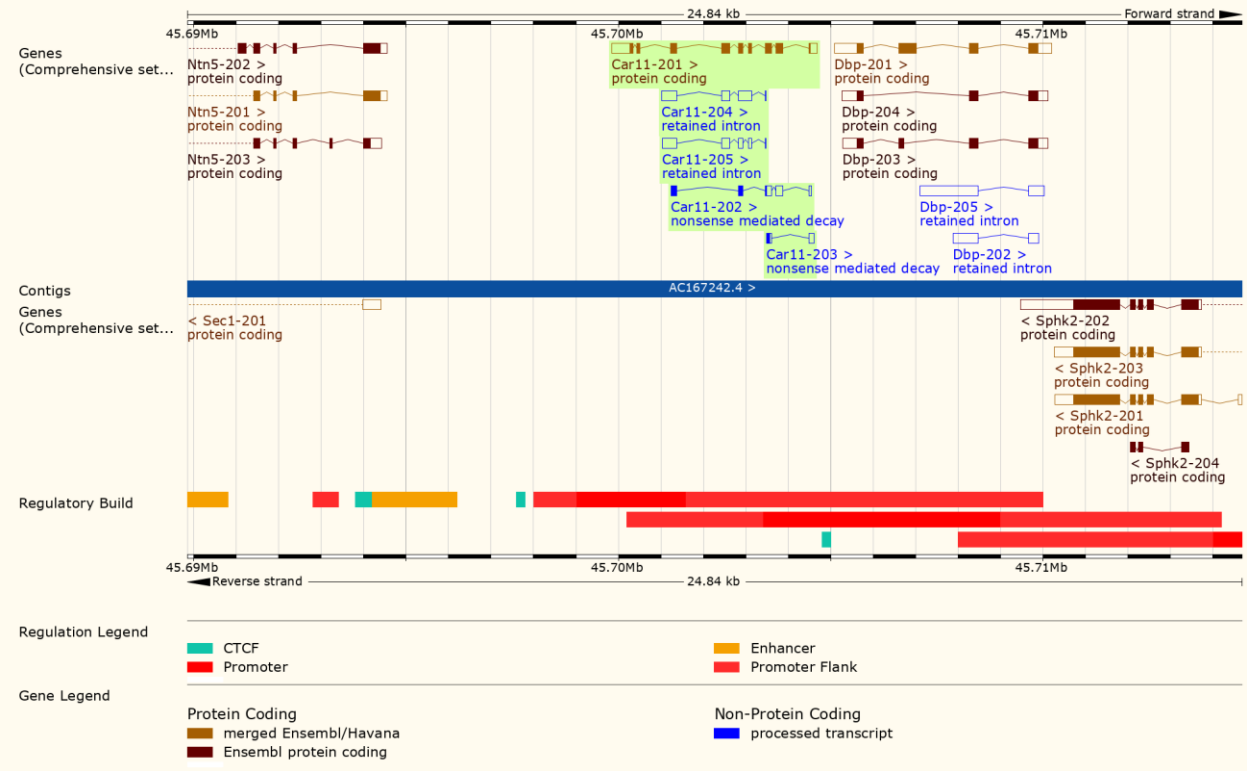
Car IX


26.76 kb
Forward strand
43.50Mb
43.51Mb
43.52Mb
Genes
(Comprehensive set...
Car9-201 >
protein coding
Car9-202 >
processed transcript
Car9-206 >
protein coding
Car9-204 >
processed transcript
Car9-207 >
processed transcript
Car9-203 >
processed transcript
Car9-205 >
processed transcript
Contigs
AL732506.10 >
Genes
(Comprehensive set...
< Arhgef39-201
protein coding
< Arhgef39-208
protein coding
< Arhgef39-203
processed transcript
< Arhgef39-204
processed transcript
< Arhgef39-202
processed transcript
< Arhgef39-205
protein coding
< Arhgef39-209
processed transcript
< Arhgef39-207
processed transcript
< Arhgef39-206
processed transcript
< Arhgef39-210
processed transcript
< Gm12454-202
antisense
< Gm12454-201
antisense
< Tpm2-203
protein coding
< Tpm2-201
protein coding
< Tpm2-205
processed transcript
< Tpm2-202
protein coding
< Tpm2-204
processed transcript
< Tpm2-206
protein coding
Regulatory Build
43.50Mb
43.51Mb
43.52Mb
Reverse strand
26.76 kb
Regulation Legend
CTCF
Promoter
Transcription Factor Binding Site
Enhancer
Promoter Flank
Gene Legend
Protein Coding
merged Ensembl/Havana
Ensembl protein coding
Non-Protein Coding
processed transcript

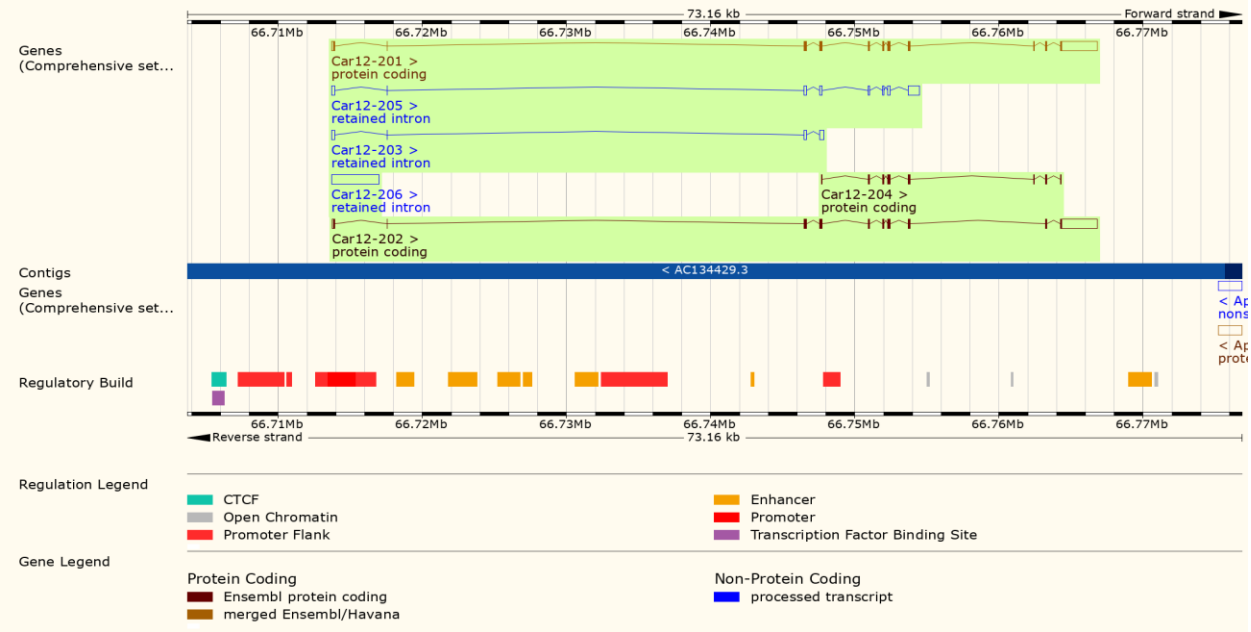
Car X


523.72 kb
Forward strand
93.1Mb
93.2Mb
93.3Mb
93.4Mb
93.5Mb
93.6Mb
Genes (Comprehensive set...
Car10-207 > processed transcript
Car10-201 > protein coding
Car10-205 > protein coding
Car10-204 > protein coding
Car10-203 > protein coding
Car10-206 > protein coding
Car10-202 > retained intron
Contigs
AL713960.17 >
AL683819.19 >
AL731849.5 >
AL731728.10 >
AL954830.7 >
Genes (Comprehensive set...
< Gm24856-201 snRNA
Regulatory Build
93.1Mb
93.2Mb
93.3Mb
93.4Mb
93.5Mb
93.6Mb
Reverse strand
523.72 kb
Regulation Legend
CTCF
Open Chromatin
Promoter Flank
Enhancer
Promoter
Gene Legend
Protein Coding
Ensembl protein coding
Non-Protein Coding
RNA gene
processed transcript

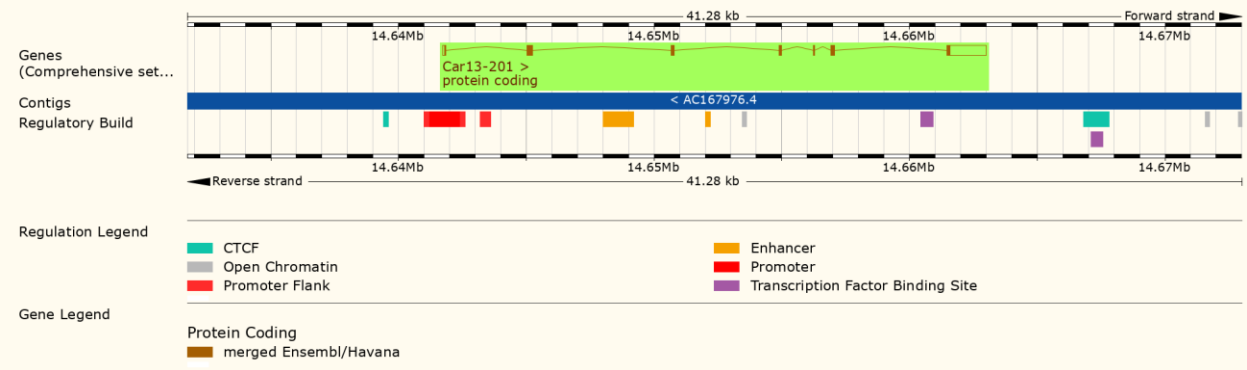
Car XI



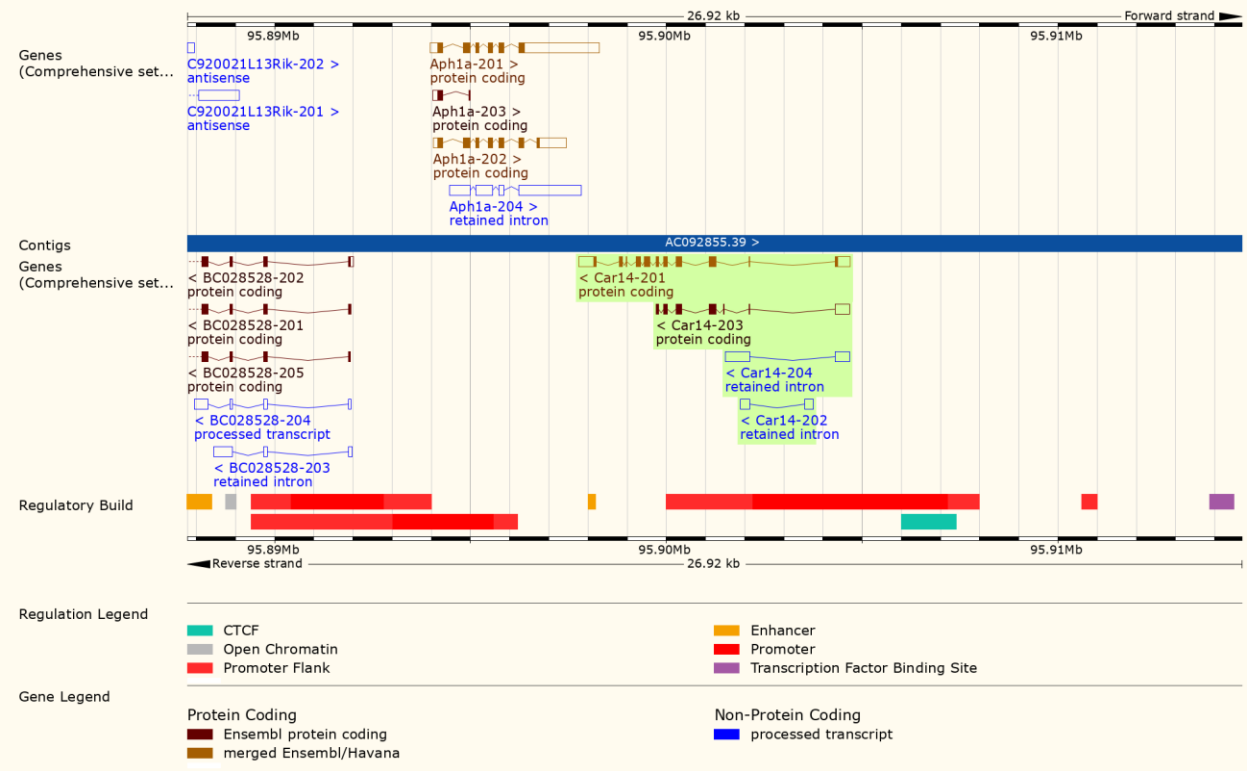
Car XII



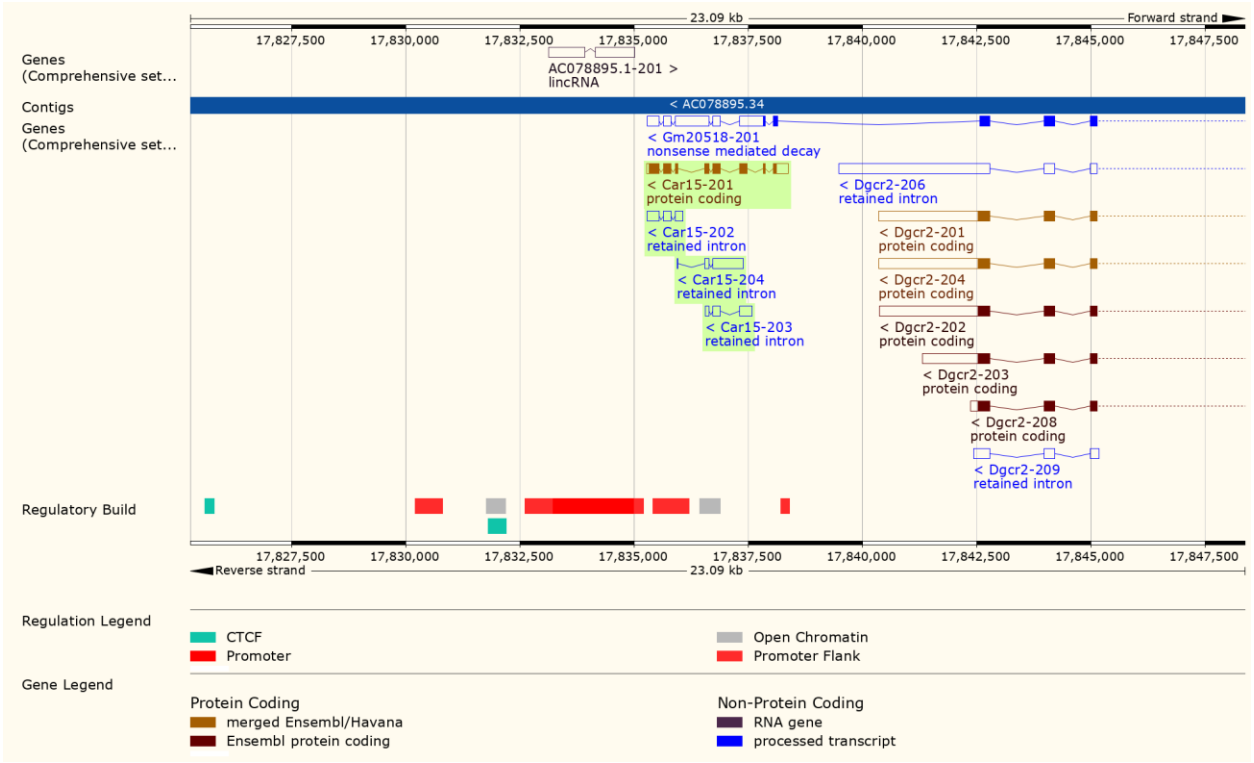
Car XIII



Car XIV


26.92 kb
Forward strand
95.89Mb
95.90Mb
95.91Mb
Genes
(Comprehensive set...
C920021L13Rik-202 >
antisense
C920021L13Rik-201 >
antisense
Aph1a-201 >
protein coding
Aph1a-203 >
protein coding
Aph1a-202 >
protein coding
Aph1a-204 >
retained intron
Contigs
AC092855.39 >
Genes
(Comprehensive set...
< BC028528-202
protein coding
< BC028528-201
protein coding
< BC028528-205
protein coding
< BC028528-204
processed transcript
< BC028528-203
retained intron
< Car14-201
protein coding
< Car14-203
protein coding
< Car14-204
retained intron
< Car14-202
retained intron
Regulatory Build
95.89Mb
95.90Mb
95.91Mb
Reverse strand
26.92 kb
Regulation Legend
CTCF
Open Chromatin
Promoter Flank
Enhancer
Promoter
Transcription Factor Binding Site
Gene Legend
Protein Coding
Ensembl protein coding
merged Ensembl/Havana
Non-Protein Coding
processed transcript

Car XV


23.09 kb
Forward strand
17,827,500
17,830,000
17,832,500
17,835,000
17,837,500
17,840,000
17,842,500
17,845,000
17,847,500
Genes (Comprehensive set...
AC078895.1-201 > lincRNA
Contigs
< AC078895.34
Genes (Comprehensive set...
< Gm20518-201 nonsense mediated decay
< Car15-201 protein coding
< Dgcr2-206 retained intron
< Car15-202 retained intron
< Dgcr2-201 protein coding
< Car15-204 retained intron
< Dgcr2-204 protein coding
< Car15-203 retained intron
< Dgcr2-202 protein coding
< Dgcr2-203 protein coding
< Dgcr2-208 protein coding
< Dgcr2-209 retained intron
Regulatory Build
Reverse strand
Regulation Legend
CTCF
Promoter
Open Chromatin
Promoter Flank
Gene Legend
Protein Coding
merged Ensembl/Havana
Ensembl protein coding
Non-Protein Coding
RNA gene
processed transcript