Genomics101

BIOL647
Digital Biology

Rodolfo Aramayo

Genomics101

Genes, Genome Files, and Genome Browsers

Open Reading Frames 101

Genomics101

Genes, Genome Files, and Genome Browsers

Gene Models 101
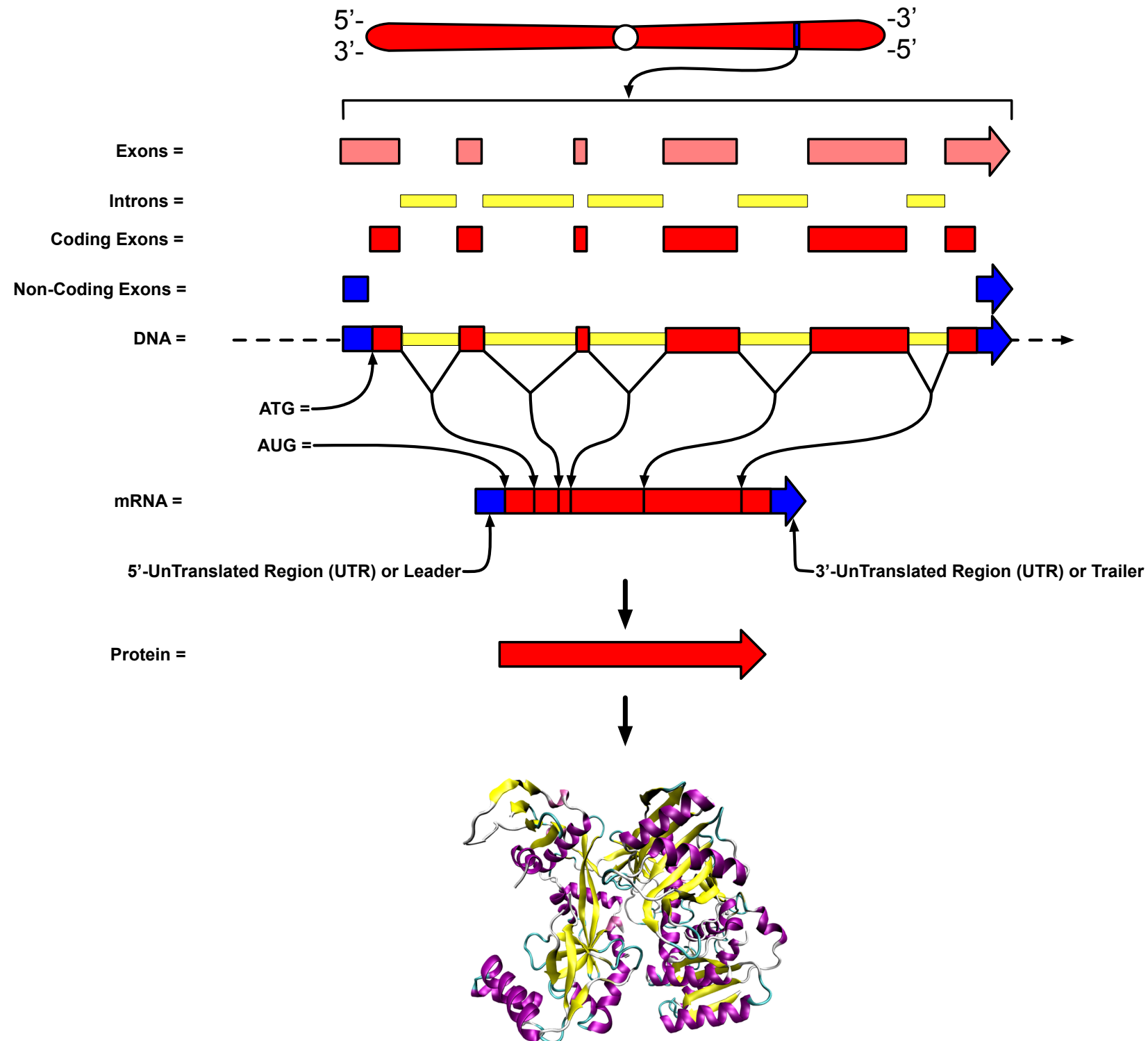
Genomics101

Genes, Genome Files, and Genome Browsers

BRCA2-ENSEMBL

Gene: BRCA2 ENSG00000139618

Description	BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101 ⧉]
Gene Synonyms	BRCC2, FACD, FAD, FAD1, FANCD, FANCD1, XRCC11
Location	Chromosome 13: 32,315,086-32,400,268 forward strand.
	GRCh38:CM000675.2
About this gene	This gene has 10 transcripts (splice variants), 175 orthologues and is associated with 171 phenotypes.
Transcripts	Hide transcript table

Show/hide columns (1 hidden)   Filter

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000380152.8	BRCA2-201	11954	3418aa	Protein coding	CCDS9344 ⧉	P51587 ⧉	NM_000059.4 ⧉	MANE Select v0.95 Ensembl Canonical GENCODE basic APPRIS P1 TSL:5
ENST00000680887.1	BRCA2-210	11880	3418aa	Protein coding	CCDS9344 ⧉	-	-	APPRIS P1
ENST00000544455.6	BRCA2-206	11854	3418aa	Protein coding	CCDS9344 ⧉	P51587 ⧉	-	GENCODE basic APPRIS P1 TSL:1
ENST00000530893.6	BRCA2-204	2011	481aa	Protein coding	-	A0A590UJI7 ⧉	-	TSL:1 CDS 3' incomplete
ENST00000614259.2	BRCA2-207	11763	2649aa	Nonsense mediated decay	-	-	-	TSL:2
ENST00000665585.1	BRCA2-208	2598	438aa	Nonsense mediated decay	-	A0A590UJU6 ⧉	-	CDS 5' incomplete
ENST00000470094.1	BRCA2-202	842	186aa	Nonsense mediated decay	-	H0YE37 ⧉	-	TSL:5 CDS 5' incomplete
ENST00000666593.1	BRCA2-209	523	58aa	Nonsense mediated decay	-	A0A590UJ24 ⧉	-	CDS 5' incomplete
ENST00000528762.1	BRCA2-203	495	64aa	Nonsense mediated decay	-	H0YD86 ⧉	-	TSL:4 CDS 5' incomplete
ENST00000533776.1	BRCA2-205	523	No protein	Retained intron	-	-	-	TSL:3

Genomics101

Genes, Genome Files, and Genome Browsers

BRCA2-Exons-ENSEMBL

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence						..........gtcacgtggccagcgcgggcttgtggcgcgagcttctgaaactaggcggc
1	ENSE00001184784	32,315,508	32,315,667	-	-	160	AGAGGCGGAGCCGCTGTGGCACTGCTGCGCCTCTGCTGCGCCTCGGGTGTCTTTTGCGGC GGTGGGTCGCCGCCGGGAGAAGCGTGAGGGGACAGATTTGTGACCGGCGCGGTTTTTGTC AGCTTACTCCGGCCAAAAAAGAACTGCACCTCTGGAGCGG
	Intron 1-2	32,315,668	32,316,421			754	gttagtggtggtggtagtgggttgg..........tgcatttttggtcttctgtttttgcag
2	ENSE00001484009	32,316,422	32,316,527	-	1	106	ACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAAAAATGCCTATTGGATCCAAAGAG AGGCCAACATTTTTTGAAATTTTTAAGACACGCTGCAACAAAGCAG
	Intron 2-3	32,316,528	32,319,076			2,549	gtattgacaaatttttatataaacttttt..........aaggtgggatttttttttttaaatag
3	ENSE00003666217	32,319,077	32,319,325	1	1	249	ATTTAGGACCAATAAGTCTTAATTGGTTTGAAGAACTTTCTTCAGAAGCTCCACCCTATA ATTCTGAACCTGCAGAAGAATCTGAACATAAAAACAACAATTACGAACCAAACCTATTTA AAACTCCACAAAGGAAACCATCTTATAATCAGCTGGCTTCAACTCCAATAATATTCAAAG AGCAAGGGCTGACTCTGCCGCTGTACCAATCTCCTGTAAAAGAATTAGATAAAATTCAAAT TAGACTTAG
	Intron 3-4	32,319,326	32,325,075			5,750	gtaagtaatgcaatatggtagactg..........tcactgaattattgtactgtttcag
4	ENSE00003659301	32,325,076	32,325,184	1	2	109	GAAGGAATGTTCCCAATAGTAGACATAAAAGTCTTCGCACAGTGAAAACTAAAATGGATC AAGCAGATGATGTTTCCTGTCCACTTCTAAATTCTTGTCTTAGTGAAAG
	Intron 4-5	32,325,185	32,326,100			916	gtatgatgaagctattatattaaaa..........agggatttgctttgtttttatttttag
5	ENSE00003739878	32,326,101	32,326,150	2	1	50	TCCTGTTGTTCTACAATGTACACATGTAACACCACAAAGAGATAAGTCAG
	Intron 5-6	32,326,151	32,326,241			91	gtatgattaaaaacaatgctttttta..........aattttcccctttttttttacccccag
6	ENSE00003747332	32,326,242	32,326,282	1	0	41	TGGTATGTGGGAGTTTGTTTCATACACCAAAGTTTGTGAAG
	Intron 6-7	32,326,283	32,326,498			216	gtaaatattctacctggtttattttt..........aataaactattttctttcctccag
7	ENSE00003749714	32,326,499	32,326,613	0	1	115	GGTCGTCAGACACCAAAACATATTTCTGAAAGTCTAGGAGCTGAGGTGGATCCTGATATG TCTTGGTCAAGTTCTTTAGCTACACCACCCACCCTTAGTTCTACTGTGCTCATAG
	Intron 7-8	32,326,614	32,329,442			2,829	gtaataatagcaaatgtgtatttac..........atacacataaattttttatcttcag
8	ENSE00003714754	32,329,443	32,329,492	1	0	50	TCAGAAATGAAGAAGCATCTGAAACTGTATTTCCTCATGATACTACTGCT
	Intron 8-9	32,329,493	32,330,918			1,426	gtaagtaaaatatgacattgattaga..........tgatttttaaactataattttttgcag
9	ENSE00003731761	32,330,919	32,331,030	0	1	112	AATGTGAAAAGCTATTTTTCCAATCATGATGAAAGTCTGAAGAAAAATGATAGATTTATC GCTTCTGTGACAGACAGTGAAAACACAAATCAAAGAGAAGCTGCAAGTCATG
	Intron 9-10	32,331,031	32,332,271			1,241	gtaagtcctctgtttagttgaacta..........gtgcttctgtttttatactttaacag
10	ENSE00000939167	32,332,272	32,333,387	1	1	1,116	GATTTGGAAAAACATCAGGGAATTCATTTAAAGTAAATAGCTGCAAAGACCACATTGGAA AGTCAATGCCAAATGTCCTAGAAGATGAAGTATATGAAACAGTTGTAGATACCTCTGAAG AAGATAGTTTTTCATTATGTTTTTCTAAATGTAGAACAAAAAATCTACAAAAAGTAAGAA CTAGCAAGACTAGGAAAAAAAATTTTCCATGAAGCAAACGCTGATGAATGTGAAAAATCTA AAAACCAAGTGAAAGAAAAATACTCATTTGTATCTGAAGTGGAACCAAATGATACTGATC CATTAGATTCAAATGTAGCAAATCAGAAGCCCTTTGAGAGTGGAAGTGACAAAATCTCA AGGAAGTTGTACCGTCTTTGGCCTGTGAATGGTCTCAACTAACCCTTTCAGGTCTAAATA GAGCCCAGATGGAAAATACCCCTATTGCATATTTCTCATGATGTGACCAAAATATTTCAG AAAAAGACCTATTAGACACAGAGAACAAAAGAAAGAAGATTTTCTTACTTCAGAGAATT CTTTGCCACGTATTTCTAGCCTACCAAAATCAGAGAAGCCATTAAATGAGGAAAACAGTG TAAATAAGAGAGATGAAGAGCAGCATCTTGAATCTCATCACAGACTGCATTCTTGCAGTA AGCAGGCAATATCTGGAACTTCTCCAGTGGCTTCTTCATTTCAGGGTATCAAAAAGTCTA TATTCAGAATAAGCAATTGAAGAACTTTCAATGCCAAGTTTCAGGTTCAGGTCATATGA

Genomics101

Genes, Genome Files, and Genome Browsers

BRCA2-cDNA-ENSEMBL

```
  1 AGAGGCGGAGCCGCTGTGGCACTGCTGCGCCTCTGCTGCGCCTCGGGTGTCTTTTGCGGC    60
    ............................................................
    ............................................................

 61 GGTGGGTCGCCGCCGGGAGAAGCGTGAGGGGACAGATTTGTGACCGGCGCGGTTTTTGTC   120
    ............................................................
    ............................................................

121 AGCTTACTCCGGCCAAAAAAGAACTGCACCTCTGGAGCGGACTTATTTACCAAGCATTGG   180
    ............................................................
    ............................................................

181 AGGAATATCGTAGGTAAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAAT   240
    ...................ATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAAT    41
    ...................-M--P--I--G--S--K--E--R--P--T--F--F--E--I    14

241 TTTTAAGACACGCTGCAACAAAGCAGATTTAGGACCAATAAGTCTTAATTGGTTTGAAGA   300
 42 TTTTAAGACACGCTGCAACAAAGCAGATTTAGGACCAATAAGTCTTAATTGGTTTGAAGA   101
 14 --F--K--T--R--C--N--K--A--D--L--G--P--I--S--L--N--W--F--E--E    34

301 ACTTTCTTCAGAAGCTCCACCCTATAATTCTGAACCTGCAGAAGAATCTGAACATAAAAA   360
102 ACTTTCTTCAGAAGCTCCACCCTATAATTCTGAACCTGCAGAAGAATCTGAACATAAAAA   161
 34 --L--S--S--E--A--P--P--Y--N--S--E--P--A--E--E--S--E--H--K--N    54

361 CAACAATTACGAACCAAACCTATTTAAAACTCCACAAAGGAAACCATCTTATAATCAGCT   420
162 CAACAATTACGAACCAAACCTATTTAAAACTCCACAAAGGAAACCATCTTATAATCAGCT   221
 54 --N--N--Y--E--P--N--L--F--K--T--P--Q--R--K--P--S--Y--N--Q--L    74

421 GGCTTCAACTCCAATAATATTCAAAGAGCAAGGGCTGACTCTGCCGCTGTACCAATCTCC   480
222 GGCTTCAACTCCAATAATATTCAAAGAGCAAGGGCTGACTCTGCCGCTGTACCAATCTCC   281
 74 --A--S--T--P--I--I--F--K--E--Q--G--L--T--L--P--L--Y--Q--S--P    94

481 TGTAAAAGAATTAGATAAATTCAAAATTAGACTTAGGAAGGAATGTTCCCAATAGTAGACA   540
282 TGTAAAAGAATTAGATAAATTCAAAATTAGACTTAGGAAGGAATGTTCCCAATAGTAGACA   341
 94 --V--K--E--L--D--K--F--K--L--D--L--G--R--N--V--P--N--S--R--H   114

541 TAAAAGTCTTCGCACAGTGAAAACTAAAATGGATCAAGCAGATGATGTTTCCTGTCCACT   600
342 TAAAAGTCTTCGCACAGTGAAAACTAAAATGGATCAAGCAGATGATGTTTCCTGTCCACT   401
114 --K--S--L--R--T--V--K--T--K--M--D--Q--A--D--D--V--S--C--P--L   134

601 TCTAAATTCTTGTCTTAGTGAAAGTCCTGTTGTTCTACAATGTACACATGTAACACCACA   660
402 TCTAAATTCTTGTCTTAGTGAAAGTCCTGTTGTTCTACAATGTACACATGTAACACCACA   461
134 --L--N--S--C--L--S--E--S--P--V--V--L--Q--C--T--H--V--T--P--Q   154

661 AAGAGATAAGTCAGTGGTATGTGGGAGTTTGTTTCATACACCAAAGTTTGTGAAGGGTCG   720
462 AAGAGATAAGTCAGTGGTATGTGGGAGTTTGTTTCATACACCAAAGTTTGTGAAGGGTCG   521
154 --R--D--K--S--V--V--C--G--S--L--F--H--T--P--K--F--V--K--G--R   174

721 TCAGACACCAAAACATATTTCTGAAAGTCTAGGAGCTGAGGTGGATCCTGATATGTCTTG   780
522 TCAGACACCAAAACATATTTCTGAAAGTCTAGGAGCTGAGGTGGATCCTGATATGTCTTG   581
174 --Q--T--P--K--H--I--S--E--S--L--G--A--E--V--D--P--D--M--S--W   194

781 GTCAAGTTCTTTAGCTACAGCACCCACCCTTAGTTCTACTGTGCTCATAGTCAGAAATGA   840
```

Genomics101

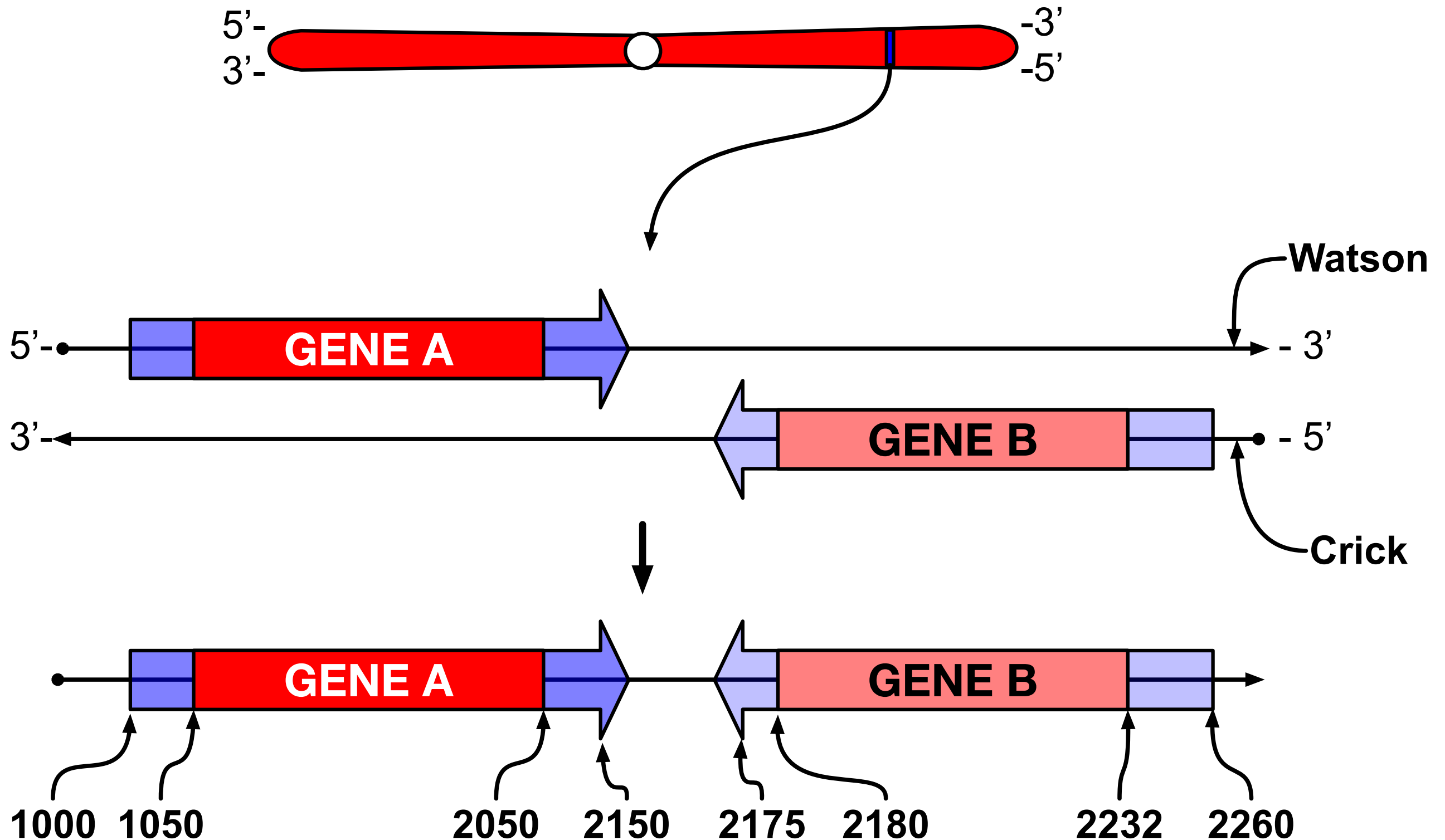## Genes, Genome Files, and Genome Browsers

BRCA2-Protein-ENSEMBL

```
MPIGSKERPTFFEIFKTRCNKADLGPISLNWFEELSSEAPPYNSEPAEESEHKNNNYEPN
LFKTPQRKPSYNQLASTPIIFKEQGLTLPLYQSPVKELDKFKLDLGRNVPNSRHKSLRTV
KTKMDQADDVSCPLLNSCLSESPVVLQCTHVTPQRDKSVVCGSLFHTPKFVKGRQTPKHI
SESLGAEVDPDMSWSSSLATPPTLSSTVLIVRNEEASETVFPHDTTANVKSYFSNHDESL
KKNDRFIASVTDSENTNQREAASHGFGKTSGNSFKVNSCKDHIGKSMPNVLEDEVYETVV
DTSEEDSFSLCFSKCRTKNLQKVRTSKTRKKIFHEANADECEKSKNQVKEKYSFVSEVEP
NDTDPLDSNVANQKPFESGSDKISKEVVPSLACEWSQLTLSGLNGAQMEKIPLLHISSCD
QNISEKDLLDTENKRKKDFLTSENSLPRISSLPKSEKPLNEETVVNKRDEEQHLESHTDC
ILAVKQAISGTSPVASSFQGIKKSIFRIRESPKETFNASFSGHMTDPNFKKETEASESGL
EIHTVCSQKEDSLCPNLIDNGSWPATTTQNSVALKNAGLISTLKKKTNKFIYAIHDETSY
KGKKIPKDQKSELINCSAQFEANAFEAPLTFANADSGLLHSSVKRSCSQNDSEEPTLSLT
SSFGTILRKCSRNETCSNNTVISQDLDYKEAKCNKEKLQLFITPEADSLSCLQEGQCEND
PKSKKVSDIKEEVLAAACHPVQHSKVEYSDTDFQSQKSLLYDHENASTLILTPTSKDVLS
NLVMISRGKESYKMSDKLKGNNYESDVELTKNIPMEKNQDVCALNENYKNVELLPPEKYM
RVASPSRKVQFNQNTNLRVIQKNQEETTSISKITVNPDSEELFSDNENNFVFQVANERNN
LALGNTKELHETDLTCVNEPIFKNSTMVLYGDTGDKQATQVSIKKDLVYVLAEENKNSVK
QHIKMTLGQDLKSDISLNIDKIPEKNNDYMNKWAGLLGPISNHSFGGSFRTASNKEIKLS
EHNIKKSKMFFKDIEEQYPTSLACVEIVNTLALDNQKKLSKPQSINTVSAHLQSSVVVSD
CKNSHITPQMLFSKQDFNSNHNLTPSQKAEITELSTILEESGSQFEFTQFRKPSYILQKS
TFEVPENQMTILKTTSEECRDADLHVIMNAPSIGQVDSSKQFEGTVEIKRKFAGLLKNDC
NKSASGYLTDENEVGFRGFYSAHGTKLNVSTEALQKAVKLFSDIENISEETSAEVHPISL
SSSKCHDSVVSMFKIENHNDKTVSEKNNKCQLILQNNIEMTTGTFVEEITENYKRNTENE
DNKYTAASRNSHNLEFDGSDSSKNDTVCIHKDETDLLFTDQHNICLKLSGQFMKEGNTQI
KEDLSDLTFLEVAKAQEACHGNTSNKEQLTATKTEQNIKDFETSDTFFQTASGKNISVAK
ESFNKIVNFFDQKPEELHNFSLNSELHSDIRKNKMDILSYEETDIVKHKILKESVPVGTG
NQLVTFQGQPERDEKIKEPTLLGFHTASGKKVKIAKESLDKVKNLFDEKEQGTSEITSFS
HQWAKTLKYREACKDLELACETIEITAAPKCKEMQNSLNNDKNLVSIETVVPPKLLSDNL
CRQTENLKTSKSIFLKVKVHENVEKETAKSPATCYTNQSPYSVIENSALAFYTSCSRKTS
VSQTSLLEAKKWLREGIFDGQPERINTADYVGNYLYENNSNSTIAENDKNHLSEKQDTYL
SNSSMSNSYSYHSDEVYNDSGYLSKNKLDSGIEPVLKNVEDQKNTSFSKVISNVKDANAY
PQTVNEDICVEELVTSSSPCKNKNAAIKLSISNSNNFEVGPPAFRIASGKIVCVSHETIK
KVKDIFTDSFSKVIKENNENKSKICQTKIMAGCYEALDDSEDILHNSLDNDECSTHSHKV
FADIQSEEILQHNQNMSGLEKVSKISPCDVSLETSDICKCSIGKLHKSVSSANTCGIFST
ASGKSVQVSDASLQNARQVFSEIEDSTKQVFSKVLFKSNEHSDQLTREENTAIRTPEHLI
SQKGFSYNVVNSSAFSGFSTASGKQVSILESSLHKVKGVLEEFDLIRTEHSLHYSPTSRQ
NVSKILPRVDKRNPEHCVNSEMEKTCSKEFKLSNNLNVEGGSSENNHSIKVSPYLSQFQQ
DKQQLVLGTKVSLVENIHVLGKEQASPKNVKMEIGKTETFSDVPVKTNIEVCSTYSKDSE
NYFETEAVEIAKAFMEDDELTDSKLPSHATHSLFTCPENEEMVLSNSRIGKRRGEPLILV
GEPSIKRNLLNEFDRIIENQEKSLKASKSTPDGTIKDRRLFMHHVSLEPITCVPFRTTKE
RQEIQNPNFTAPGQEFLSKSHLYEHLTLEKSSSNLAVSGHPFYQVSATRNEKMRHLITTG
RPTKVFVPPFKTKSHFHRVEQCVRNINLEENRQKQNIDGHGSDDSKNKINDNEIHQFNKN
NSNQAVAVTFTKCEEEPLDLITSLQNARDIQDMRIKKKQRQRVFPQPGSLYLAKTSTLPR
ISLKAAVGGQVPSACSHKQLYTYGVSKHCIKINSKNAESFQFHTEDYFGKESLWTGKGIQ
LADGGWLIPSNDGKAGKEEFYRALCDTPGVDPKLISRIWVYNHYRWIIWKLAAMECAFPK
EFANRCLSPERVLLQLKYRYDTEIDRSRRSAIKKIMERDDTAAKTLVLCVSDIISLSANI
SETSSNKTSSADTQKVAIIELTDGWYAVKAQLDPPLLAVLKNGRLTVGQKIILHGAELVG
SPDACTPLEAPESLMLKISANSTRPARWYTKLGFFPDPRPFPLPSSLFSDGGNVGCVDV
IIQRAYPIQWMEKTSSGLYIFRNEREEEKEAAKYVEAQQKRLEALFTKIQEEFEEHEENT
TKPYLPSRALTRQQVRALQDGAELYEAVKNAADPAYLEGYFSEEQLRALNNHRQMLNDKK
QAQIQLEIRKAMESAEQKEQGLSRDVTTVWKLRIVSYSKKEKDSVILSIWRPSSDLYSLL
```

Genomics101

Genes, Genome Files, and Genome Browsers

Fasta File Format

- *FASTA originally stood for a fast algorithm that could align any letters (not just nucleotides or protein symbols), but came to represent the file format inputted to that algorithm*

- *The Fasta file format is the standard format for interchanging simple sequence data*

Genomics101

Genes, Genome Files, and Genome Browsers

Fasta File Format

1. **FASTA**

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is:

>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)   ← **Header**

QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP   ← **Sequence**
FLFLIKHNPTNTIVYFGRYWSP

Blank lines are not allowed in the middle of FASTA input.

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below). Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue). The nucleic acid codes supported are:

A	adenosine	C	cytidine	G	guanine
T	thymidine	N	A/G/C/T (any)	U	uridine
K	G/T (keto)	S	G/C (strong)	Y	T/C (pyrimidine)
M	A/C (amino)	W	A/T (weak)	R	G/A (purine)
B	G/T/C	D	G/A/T	H	A/C/T
V	G/C/A	–	gap of indeterminate length		

Genomics101

Genes, Genome Files, and Genome Browsers

Fasta File Format

- *Fasta files can contain DNA, RNA or Protein sequences*

- *A Fasta file can contain one record (i.e., one DNA, RNA or Protein sequence), or many (i.e., more than one DNA, RNA or Protein sequence) records*

  - *Records are separated by the symbol ">"*

  - *A Fasta library is a Fasta file containing multiple Fasta records (i.e., one DNA, RNA or Protein records)*

  - *Different Sequence types (protein, DNA, or RNA) should NEVER be mixed in Fasta libraries*

- *It is recommended, not required, to use the following suffixes:*

  - *Use '.fna' or 'fasta-nucleotides-a' for fasta-nucleotides*

  - *Use '.faa' or 'fasta-aminoacides-a' for fasta-polypeptides*

Genomics101

Genes, Genome Files, and Genome Browsers

Comparing BRCA2 unmasked with BRCA2 soft_masked Regions

```
>13 dna:chromosome chromosome:GRCh38:13:32311612:32403740:1
CACTGCACAGCCTTGTCCACCCGCGGGCTCACCTGCACGCCCACCTCCTTGGTGTTGGGC
TTGCACAGCCGCGGGCTCAGGCTGGGGTTCATCTGGGAGAGAATGGCCTTAAGCTGCGCC
CTCTTGTAAGGGTCAATGCAGTAGTCAGGGGCGTTCGCGGGCACCAGCAGCCCTGGCCTG
GCCAGAAAAGTGGGAGGACCCATATTTTGCCTCCAGTCGGGCTGTTTGTGCCCTGAGAGT
CCAGGCTGGCCCAAAGGCACTGTGCTCCCATAACCCTGGTACAAGCCATAGGGAACACGG
ACAAAGCGCTCCATCCGCTCTCAGGTGCTCAGGCGCAGTCTAATATCCTAGCCAGTTTGT
TTGGGTCCTTATTTTGCCTTCCTCCTTCATCCGCCCCTTCTTTCTCTTCATCAGGTTGGT
TCAGATTCATTCCTGGCTTCTCCATTTATTTCAGATTCTAATGGGAGCTGCTGCTTATTA
CCCTGATTGAGGGAGAGAAGCTCTATCTACAGATGTCTAATTGCCACCTGATTCCTACCT
AATTGCTTTTCCCTACCTCTTCACTGTTCGGGAATCTTTTTTTCAAAGCTGTTTTTCATT
TCCAGCTTATTCACACATATGCTTTAATTAGTGATTAATTCATATATTCCTGTAGTGACC
ATTCTAAAGATCCTTTGTAAGGTAATGTGGTAATGTTAGATAATGTTTTAAGGTACAGCC
ATTATGGACAACAGTATGAAGGTTCCTTAAAAAAGTTAATAGAAGTACCAGATAATCCAG
TAATCCCACCACTGCGTATTTACCCAAAGGAAATGAAATCAATATCCTGAAGAGATACCT
TCACTTCCTTGTTCGTTGTAGCATTATGCACAATAGCCAAGTTATGGAAACAATCTAAGT
GTCTGTTAACAGAAAAATGGATAGAGAAATTGTGGCATGTATTTATACAATGGAATATTA
TTCAGCCTTAAAACAGGAGATCCTGGCCAGGCGCGGTGGCTCACACCTGTAATCCCAGCA
GTTTGGGAGGCCGAGGCAGGCAGATGACCTAAGGTCAGGAGTTCGAGACCAACCTGGCCA
ACATGATGAAACCCTGCCTCTAATAAAAATACAAAAATCAGCCGGGTATGGTGGCACGCT
CCTGTAATCCCAGCTACTCGGGAGGCTGAGGCTGGAGAATCACCTGAACCCGGGAGGCAA
AGGTTGCAGTGAGCTGAGATCACACTACTGCACTCCAGTCTGGGCAACAGAGAGAGAAAG
ACTCCGTCTCAAACAAACAAACAAACAAAAAACAGGAGATCCTGCCATTTGCAACAACAT
GGGTGAAACTGGGGAACATTATGCTAAGTGAAATAAGCCAGATAGAAAAATACTGCATAA
ACTCACTGATATGTGGAATCTGAAAAAGTCAAATACATAGAAGCAGAAAGTAGAATGGTG
```

```
>13 dna:chromosome chromosome:GRCh38:13:32311612:32403740:1
CACTGCACAGCCTTGTCCACCCGCGGGCTCACCTGCACGCCCACCTCCTTGGTGTTGGGC
TTGCACAGCCGCGGGCTCAGGCTGGGGTTCATCTGGGAGAGAATGGCCTTAAGCTGCGCC
CTCTTGTAAGGGTCAATGCAGTAGTCAGGGGCGTTCGCGGGCACCAGCAGCCCTGGCCTG
GCCAGAAAAGTGGGAGGACCCATATTTTGCCTCCAGTCGGGCTGTTTGTGCCCTGAGAGT
CCAGGCTGGCCCAAAGGCACTGTGCTCCCATAACCCTGGTACAAGCCATAGGGAACACGG
ACAAAGCGCTCCATCCGCTCTCAGGTGCTCAGGCGCAGTCTAATATCCTAGCCAGTTTGT
TTGGGTCCTTATTTTGCCTTCCTCCTTCATCCGCCCCTTCTTTCTCTTCATCAGGTTGGT
TCAGATTCATTCCTGGCTTCTCCATTTATTTCAGATTCTAATGGGAGCTGCTGCTTATTA
CCCTGATTGAGGGAGAGAAGCTCTATCTACAGATGTCTAATTGCCACCTGATTCCTACCT
AATTGCTTTTCCCTACCTCTTCACTGTTCGGGAATCTTTTTTTCAAAGCTGTTTTTCATT
TCCAGCTTATTCACACATATGCTTTAATTAGTGATTAATTCATATATTCCTGTAGTGACC
ATTCTAAAGATCCTTTGTAAGGTAATGTGGTAATGTTAGATAATGTTTTAAggtacagcc
attatggacaacagtatgaaggttccttaaaaaagttaatagaagtaccagataatccag
taatcccaccactgcgtatttacccaaaggaaatgaaatcaatatcctgaagagatacct
tcacttccttgttcgttgtagcattatgcacaatagccaagttatggaaacaatctaagt
gtctgttaacagaaaaatggatagagaaattgtggcatgtatttatacaatggaatatta
ttcagccttaaaacaggagatcctggccaggcgcggtggctcacacctgtaatcccagca
gtttgggaggccgaggcaggcagatgacctaaggtcaggagttcgagaccaacctggcca
acatgatgaaaccctgcctctaataaaaatacaaaaatcagccgggtatggtggcacgct
cctgtaatcccagctactcgggaggctgaggctggagaatcacctgaacccgggaggcaa
aggttgcagtgagctgagatcacactactgcactccagtctgggcaacagagagagaaag
actccgtctcaaacaaacaaacaaacaaaaaacaggagatcctgccatttgcaacaacat
gggtgaaactggggaacattatgctaagtgaaataagccagatagaaaaatactgcataa
actcactgatatgtggaatctgaaaaagtcaaatacatagaagcagaaagtagaatggtg
```

Genomics101

Genes, Genome Files, and Genome Browsers

Comparing BRCA2 unmasked with BRCA2 hard_masked Regions

```
>13 dna:chromosome chromosome:GRCh38:13:32311612:32403740:1
CACTGCACAGCCTTGTCCACCCGCGGGCTCACCTGCACGCCCACCTCCTTGGTGTTGGGC
TTGCACAGCCGCGGGCTCAGGCTGGGGTTCATCTGGGAGAGAATGGCCTTAAGCTGCGCC
CTCTTGTAAGGGTCAATGCAGTAGTCAGGGGCGTTCGCGGGCACCAGCAGCCCTGGCCTG
GCCAGAAAAGTGGGAGGACCCATATTTTGCCTCCAGTCGGGCTGTTTGTGCCCTGAGAGT
CCAGGCTGGCCCAAAGGCACTGTGCTCCCATAACCCTGGTACAAGCCATAGGGAACACGG
ACAAAGCGCTCCATCCGCTCTCAGGTGCTCAGGCGCAGTCTAATATCCTAGCCAGTTTGT
TTGGGTCCTTATTTTGCCTTCCTCCTTCATCCGCCCCTTCTTTCTCTTCATCAGGTTGGT
TCAGATTCATTCCTGGCTTCTCCATTTATTTCAGATTCTAATGGGAGCTGCTGCTTATTA
CCCTGATTGAGGGAGAGAAGCTCTATCTACAGATGTCTAATTGCCACCTGATTCCTACCT
AATTGCTTTTCCCTACCTCTTCACTGTTCGGGAATCTTTTTTTCAAAGCTGTTTTTCATT
TCCAGCTTATTCACACATATGCTTTAATTAGTGATTAATTCATATATTCCTGTAGTGACC
ATTCTAAAGATCCTTTGTAAGGTAATGTGGTAATGTTAGATAATGTTTTAAGGTACAGCC
ATTATGGACAACAGTATGAAGGTTCCTTAAAAAAGTTAATAGAAGTACCAGATAATCCAG |
TAATCCCACCACTGCGTATTTACCCAAAGGAAATGAAATCAATATCCTGAAGAGATACCT |
TCACTTCCTTGTTCGTTGTAGCATTATGCACAATAGCCAAGTTATGGAAACAATCTAAGT |
GTCTGTTAACAGAAAAATGGATAGAGAAATTGTGGCATGTATTTATACAATGGAATATTA |
TTCAGCCTTAAAACAGGAGATCCTGGCCAGGCGCGGTGGCTCACACCTGTAATCCCAGCA |
GTTTGGGAGGCCGAGGCAGGCAGATGACCTAAGGTCAGGAGTTCGAGACCAACCTGGCCA |
ACATGATGAAACCCTGCCTCTAATAAAAATACAAAAATCAGCCGGGTATGGTGGCACGCT |
CCTGTAATCCCAGCTACTCGGGAGGCTGAGGCTGGAGAATCACCTGAACCCGGGAGGCAA |
AGGTTGCAGTGAGCTGAGATCACACTACTGCACTCCAGTCTGGGCAACAGAGAGAGAAAG |
ACTCCGTCTCAAACAAACAAACAAACAAAAAACAGGAGATCCTGCCATTTGCAACAACAT |
GGGTGAAACTGGGGAACATTATGCTAAGTGAAATAAGCCAGATAGAAAAATACTGCATAA |
ACTCACTGATATGTGGAATCTGAAAAAGTCAAATACATAGAAGCAGAAAGTAGAATGGTG |
```

```
>13 dna:chromosome chromosome:GRCh38:13:32311612:32403740:1
CACTGCACAGCCTTGTCCACCCGCGGGCTCACCTGCACGCCCACCTCCTTGGTGTTGGGC
TTGCACAGCCGCGGGCTCAGGCTGGGGTTCATCTGGGAGAGAATGGCCTTAAGCTGCGCC
CTCTTGTAAGGGTCAATGCAGTAGTCAGGGGCGTTCGCGGGCACCAGCAGCCCTGGCCTG
GCCAGAAAAGTGGGAGGACCCATATTTTGCCTCCAGTCGGGCTGTTTGTGCCCTGAGAGT
CCAGGCTGGCCCAAAGGCACTGTGCTCCCATAACCCTGGTACAAGCCATAGGGAACACGG
ACAAAGCGCTCCATCCGCTCTCAGGTGCTCAGGCGCAGTCTAATATCCTAGCCAGTTTGT
TTGGGTCCTTATTTTGCCTTCCTCCTTCATCCGCCCCTTCTTTCTCTTCATCAGGTTGGT
TCAGATTCATTCCTGGCTTCTCCATTTATTTCAGATTCTAATGGGAGCTGCTGCTTATTA
CCCTGATTGAGGGAGAGAAGCTCTATCTACAGATGTCTAATTGCCACCTGATTCCTACCT
AATTGCTTTTCCCTACCTCTTCACTGTTCGGGAATCTTTTTTTCAAAGCTGTTTTTCATT
TCCAGCTTATTCACACATATGCTTTAATTAGTGATTAATTCATATATTCCTGTAGTGACC
ATTCTAAAGATCCTTTGTAAGGTAATGTGGTAATGTTAGATAATGTTTTAANNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Genomics101

Genes, Genome Files, and Genome Browsers

BED File Format

ENST00000380152.7 BRCA2-201 Region

13	32315314	32315363	ENSE00002167541	0	−
13	32314329	32314550	ENSE00002196734	0	−
13	32311271	32312093	ENSE00002140998	0	−
13	32310638	32310731	ENSE00001385658	0	−
13	32308685	32308760	ENSE00001365416	0	−
13	32303698	32304022	ENSE00001380497	0	−
13	32311271	32311954	ENSE00001375742	0	−
13	32310638	32310731	ENSE00001385658	0	−
13	32308685	32308760	ENSE00001365416	0	−
13	32303770	32304022	ENSE00003301848	0	−
13	32315085	32315145	ENSE00003856928	0	+
13	32316421	32316527	ENSE00001484009	0	+
13	32315473	32315667	ENSE00001184784	0	+
13	32316421	32316527	ENSE00001484009	0	+
13	32319076	32319325	ENSE00003666217	0	+
13	32325075	32325184	ENSE00003659301	0	+
13	32326100	32326150	ENSE00003739878	0	+
13	32326241	32326282	ENSE00003747332	0	+
13	32326498	32326613	ENSE00003749714	0	+
13	32329442	32329492	ENSE00003714754	0	+
13	32330918	32331030	ENSE00003731761	0	+
13	32332271	32333387	ENSE00000939167	0	+
13	32336264	32341196	ENSE00000939168	0	+
13	32344557	32344653	ENSE00000939169	0	+

Genomics101

Genes, Genome Files, and Genome Browsers

GFF3 File Format

ENST00000380152.7 BRCA2-201 Region

```
##gff-version 3
# ##sequence-region 13 1 114364328
13 Ensembl transcript 32303699 32315363 . - . ID=ENST00000533490.7;Name=ENST00000533490.7;Parent=ENSG00000189167.12;biotype=protein_coding
13 Ensembl transcript 32303771 32311954 . - . ID=ENST00000345108.6;Name=ENST00000345108.6;Parent=ENSG00000189167.12;biotype=protein_coding
13 Ensembl transcript 32315086 32316527 . + . ID=ENST00000671466.1;Name=ENST00000671466.1;Parent=ENSG00000139618.15;biotype=protein_coding
13 Ensembl transcript 32315474 32400266 . + . ID=ENST00000380152.7;Name=ENST00000380152.7;Parent=ENSG00000139618.15;biotype=protein_coding
13 Ensembl transcript 32315505 32333291 . + . ID=ENST00000530893.6;Name=ENST00000530893.6;Parent=ENSG00000139618.15;biotype=protein_coding
13 Ensembl transcript 32316072 32316527 . + . ID=ENST00000670614.1;Name=ENST00000670614.1;Parent=ENSG00000139618.15;biotype=protein_coding
13 Ensembl transcript 32316461 32362659 . + . ID=ENST00000614259.1;Name=ENST00000614259.1;Parent=ENSG00000139618.15;biotype=processed_transcript
13 Ensembl transcript 32356428 32398233 . + . ID=ENST00000665585.1;Name=ENST00000665585.1;Parent=ENSG00000139618.15;biotype=nonsense_mediated_decay
13 Ensembl transcript 32370971 32379495 . + . ID=ENST00000528762.1;Name=ENST00000528762.1;Parent=ENSG00000139618.15;biotype=nonsense_mediated_decay
13 Ensembl transcript 32379840 32398272 . + . ID=ENST00000470094.1;Name=ENST00000470094.1;Parent=ENSG00000139618.15;biotype=nonsense_mediated_decay
```
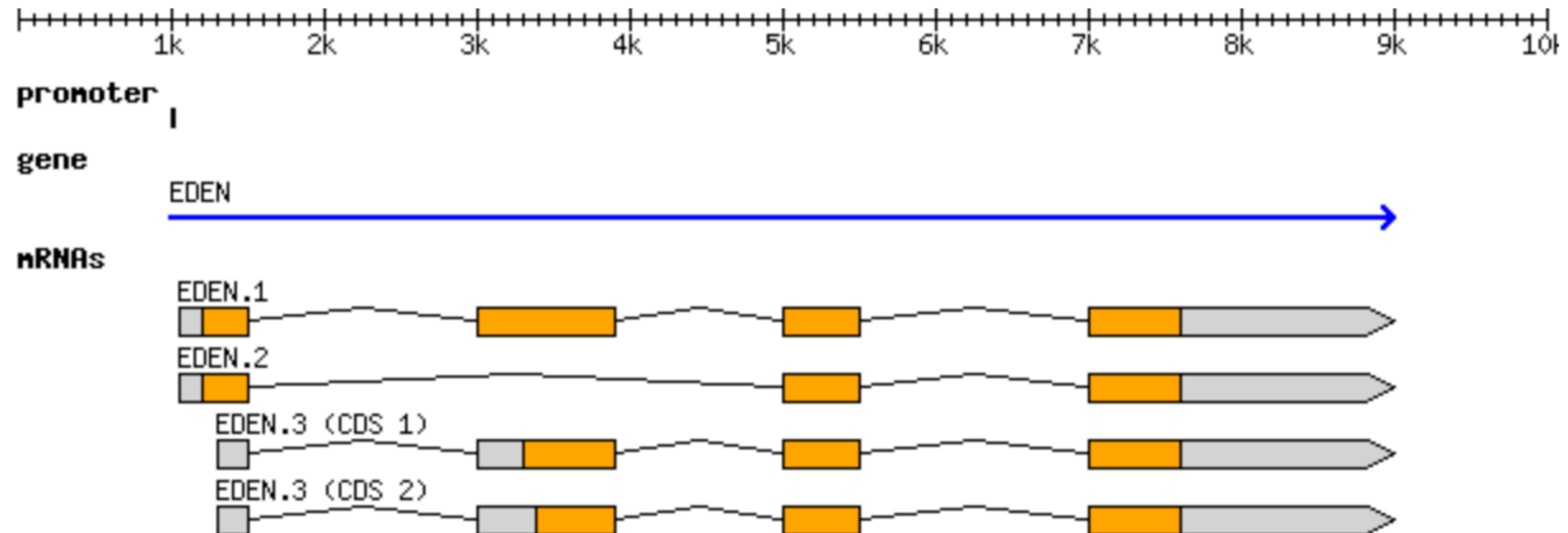
Genomics101

Genes, Genome Files, and Genome Browsers

GFF3 File Format

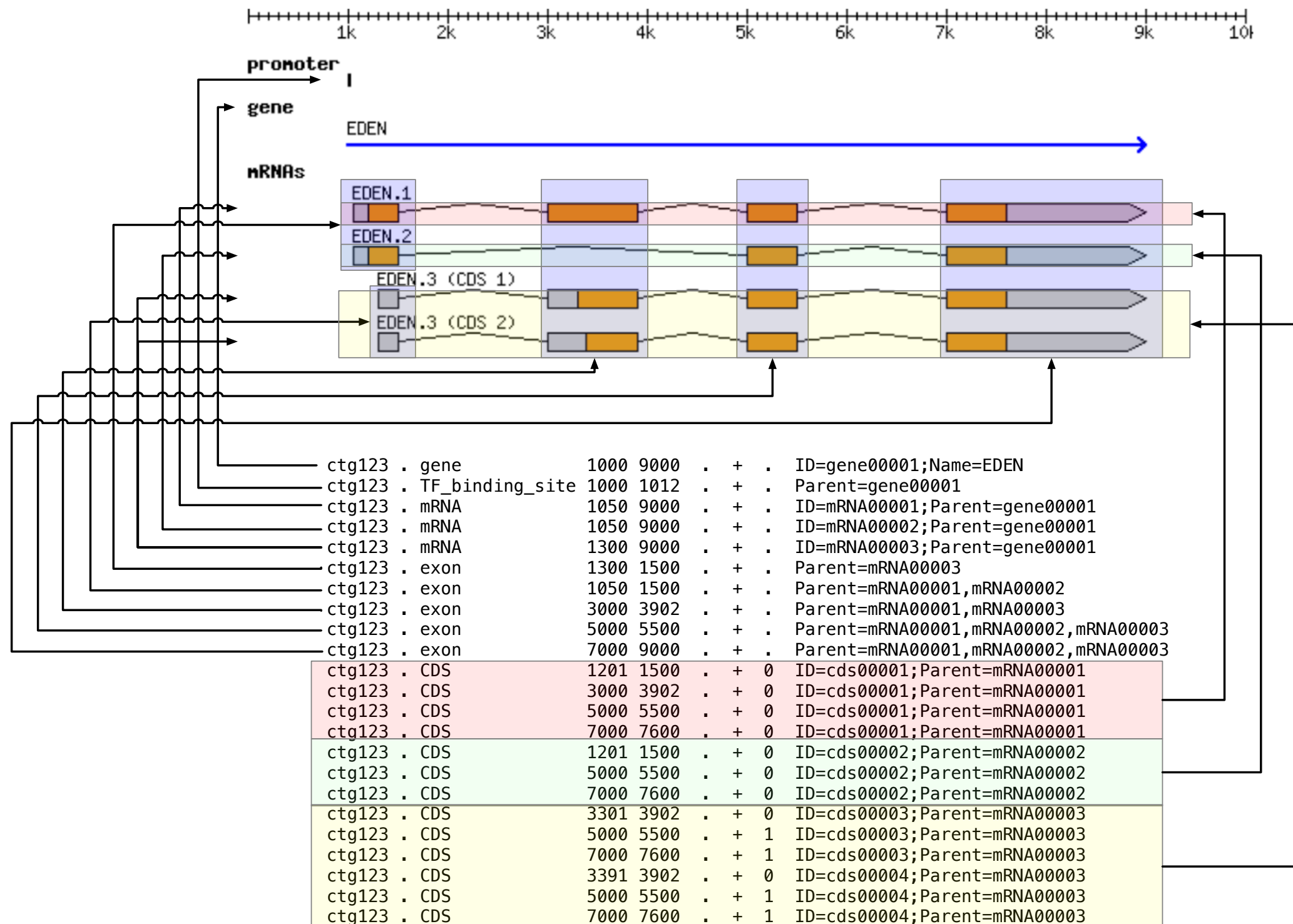### *Generic Feature Format Version 3 (GFF3)*

The Canonical Gene

Genomics101

Genes, Genome Files, and Genome Browsers

GFF3 File Format

Genomics101

Genes, Genome Files, and Genome Browsers

GFF/GTF File Format

ENST00000380152.7 BRCA2-201 Region

```
13 Ensembl gene 32315315 32315363 . − 0 ENSG00000189167.12 ENST00000533490.7
13 Ensembl gene 32314330 32314550 . − 0 ENSG00000189167.12 ENST00000533490.7
13 Ensembl gene 32311272 32312093 . − 0 ENSG00000189167.12 ENST00000533490.7
13 Ensembl gene 32310639 32310731 . − 0 ENSG00000189167.12 ENST00000533490.7
13 Ensembl gene 32308686 32308760 . − 0 ENSG00000189167.12 ENST00000533490.7
13 Ensembl gene 32303699 32304022 . − 0 ENSG00000189167.12 ENST00000533490.7
13 Ensembl gene 32311272 32311954 . − 0 ENSG00000189167.12 ENST00000345108.6
13 Ensembl gene 32310639 32310731 . − 0 ENSG00000189167.12 ENST00000345108.6
13 Ensembl gene 32308686 32308760 . − 0 ENSG00000189167.12 ENST00000345108.6
13 Ensembl gene 32303771 32304022 . − 0 ENSG00000189167.12 ENST00000345108.6
13 Ensembl gene 32315086 32315145 . + −1 ENSG00000139618.15 ENST00000671466.1
13 Ensembl gene 32316422 32316527 . + −1 ENSG00000139618.15 ENST00000671466.1
13 Ensembl gene 32315474 32315667 . + −1 ENSG00000139618.15 ENST00000380152.7
```

Genomics101

Genes, Genome Files, and Genome Browsers

GenBank and GenPept Records

Human breast cancer susceptibility (BRCA2) mRNA, complete cds

```
LOCUS       HSU43746              10987 bp    mRNA    linear   PRI 03-SEP-1996
DEFINITION  Human breast cancer susceptibility (BRCA2) mRNA, complete cds.
ACCESSION   U43746
VERSION     U43746.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (sites)
  AUTHORS   Teng,D.H., Bogden,R., Mitchell,J., Baumgard,M., Bell,R., Berry,S.,
            Davis,T., Ha,P.C., Kehrer,R., Jammulapati,S., Chen,Q., Offit,K.,
            Skolnick,M.H., Tavtigian,S.V., Jhanwar,S., Swedlund,B., Wong,A.K.
            and Kamb,A.
  TITLE     Low incidence of BRCA2 mutations in breast carcinoma and other
            cancers
  JOURNAL   Nat. Genet. 13 (2), 241-244 (1996)
   PUBMED   8640236
REFERENCE   2  (bases 1 to 10987)
  AUTHORS   Tavtigian,S.V., Rommens,J.M., Couch,F.J., Neuhausen,S., Bell,R.,
            Berry,S., Bogden,R., Chen,Q., Davis,T., Frye,C., Hattier,T.,
            Jammulapati,S., Janecki,T., Jiang,P., Kehrer,R., Schroeder,M.,
            Snyder,S., Stringfellow,M., Stroup,C., Swedlund,B., Teng,D.,
            Thomas,A., Tran,T., Weaver-Feldhaus,J., Wong,A., Leblanc,J.-F.,
            Belanger,C., Tranchant,M., Samson,C., Dumont,M.,
            McArthur-Morrison,J., McSweeney,D., Peng,Y., Shizuya,H., Slepak,T.,
            Simon,M.I., Labrie,F., Shattuck-Eidens,D., Skolnick,M., Goldgar,D.,
            Weber,B.L., Simard,J. and Kamb,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-DEC-1995) Sean Tavtigian, Myriad Genetics Inc.,
            Department of Research, 390 Wakara Way, Salt Lake City, 84106, USA
COMMENT     On Jan 19, 1996 this sequence version replaced gi:1147801.
            Tavtigian,S.V., Bell,R., Berry,S., Bogden,R., Chen,Q., Davis,T.,
            Frye,C., Hattier,T., Jammulapati,S., Janecki,T., Jiang,P.,
            Kehrer,R., Schroeder,M., Snyder,S., Stringfellow,M., Stroup,C.,
            Swedlund,B., Shattuck-Eidens,D., Skolnick,M., and  Kamb,A.
              Myriad Genetics Inc.
              390 Wakara Way
              Salt lake City, Utah 84106
```

```
                             University of Utah Medical Center.
FEATURES             Location/Qualifiers
     source          1..10987
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="13"
                     /map="13q12-q13"
     gene            1..10987
                     /gene="BRCA2"
     CDS             229..10485
                     /gene="BRCA2"
                     /codon_start=1
                     /product="BRCA2"
                     /protein_id="AAB07223.1"
                     /translation="MPIGSKERPTFFEIFKTRCNKADLGPISLNWFEELSSEAPPYNS
                     EPAEESEHKNNNYEPNLFKTPQRKPSYNQLASTPIIFKEQGLTLPLYQSPVKELDKFK
                     LDLGRNVPNSRHKSLRTVKTKMDQADDVSCPLLNSCLSESPVVLQCTHVTPQRDKSVV
                     CGSLFHTPKFVKGRQTPKHISESLGAEVDPDMSWSSSLATPPTLSSTVLIVRNEEASE
                     TVFPHDTTANVKSYFSNHDESLKKNDRFIASVTDSENTNQREAASHGFGKTSGNSFKV
                     NSCKDHIGKSMPNVLEDEVYETVVDTSEEDSFSLCFSKCRTKNLQKVRTSKTRKKIFH
                     EANADECEKSKNQVKEKYSFVSEVEPNDTDPLDSNVAHQKPFESGSDKISKEVVPSLA
                     CEWSQLTLSGLNGAQMEKIPLLHISSCDQNISEKDLLDTENKRKKDFLTSENSLPRIS
                     SLPKSEKPLNEETVVNKRDEEQHLESHTDCILAVKQAISGTSPVASSFQGIKKSIFRI
                     RESPKETFNASFSGHMTDPNFKKETEASESGLEIHTVCSQKEDSLCPNLIDNGSWPAT
                     TTQNSVALKNAGLISTLKKKTNKFIYAIHDETFYKGKKIPKDQKSELINCSAQFEANA
                     FEAPLTFANADSGLLHSSVKRSCSQNDSEEPTLSLTSSFGTILRKCSRNETCSNNTVI
                     SQDLDYKEAKCNKEKLQLFITPEADSLSCLQEGQCENDPKSKKVSDIKEEVLAAACHP
                     VQHSKVEYSDTDFQSQKSLLYDHENASTLILTPTSKDVLSNLVMISRGKESYKMSDKL
                     KGNNYESDVELTKNIPMEKNQDVCALNENYKNVELLPPEKYMRVASPSRKVQFNQNTN
                     LRVIQKNQEETTSISKITVNPDSEELFSDNENNFVFQVANERNNLALGNTKELHETDL
                     TCVNEPIFKNSTMVLYGDTGDKQATQVSIKKDLVYVLAEENKNSVKQHIKMTLGQDLK
                     SDISLNIDKIPEKNNDYMNKWAGLLGPISNHSFGGSFRTASNKEIKLSEHNIKKSKMF
                     FKDIEEQYPTSLACVEIVNTLALDNQKKLSKPQSINTVSAHLQSSVVVSDCKNSHITP
                     QMLFSKQDFNSNHNLTPSQKAEITELSTILEESGSQFEFTQFRKPSYILQKSTFEVPE
                     NQMTILKTTSEECRDADLHVIMNAPSIGQVDSSKQFEGTVEIKRKFAGLLKNDCNKSA
                     SGYLTDENEVGFRGFYSAHGTKLNVSTEALQKAVKLFSDIENISEETSAEVHPISLSS
                     SKCHDSVVSMFKIENHNDKTVSEKNNKCQLILQNNIEMTTGTFVEEITENYKRNTENE
                     DNKYTAASRNSHNLEFDGSDSSKNDTVCIHKDETDLLFTDQHNICLKLSGQFMKEGNT
                     QIKEDLSDLTFLEVAKAQEACHGNTSNKEQLTATKTEQNIKDFETSDTFFQTASGKNI
```

Genomics101

Genes, Genome Files, and Genome Browsers
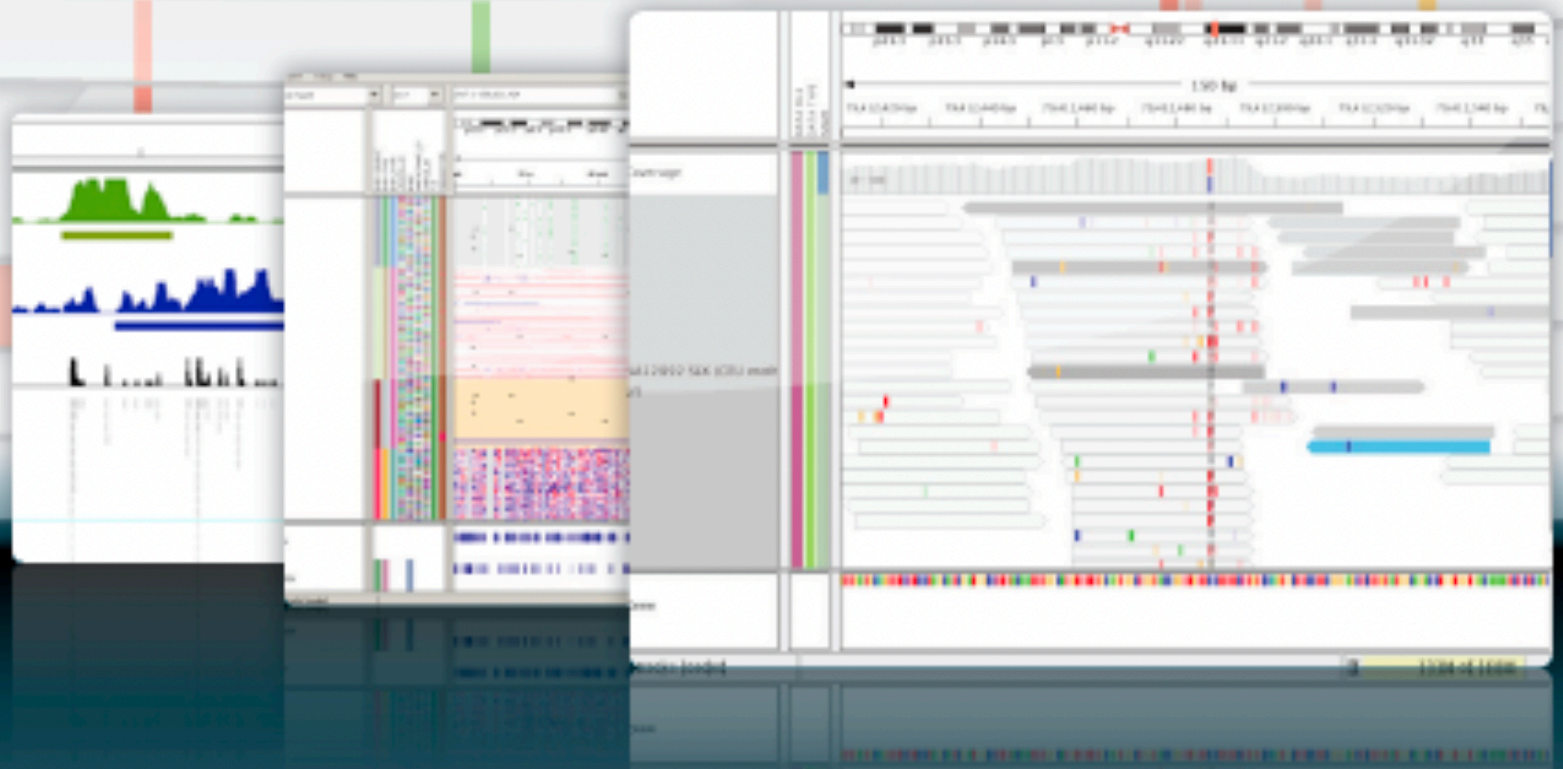
NCBI Genome Browser

Genome Data Viewer
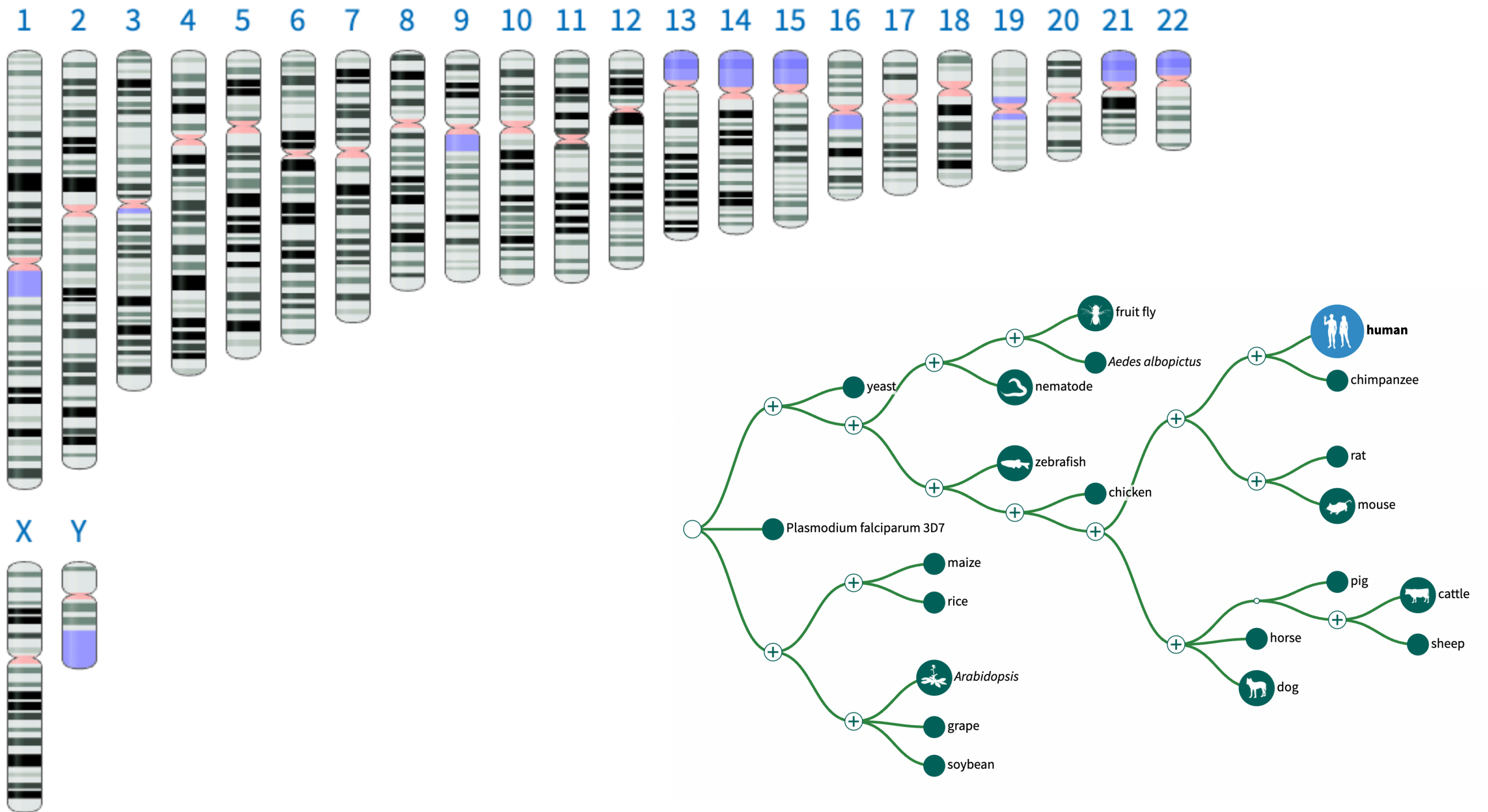
Genomics101

Genes, Genome Files, and Genome Browsers

UCSC Genome Browser

ENSEMBL Genome Browser

IGV Genome Browser

Genomics101

BIOL647
Digital Biology

Rodolfo Aramayo