# Applied Bioinformatics Exercise – Learning to know a new protein and working with sequences

In this Exercise we will explore some databases and tools that can be used to get more insight about a new protein coding gene and the corresponding protein.

1. Open your internet browser and go to the website http://www.genenames.org. Here you find the database of the HUGO Gene Nomenclature Committee (HGNC) approved gene names. Search for OGG1 and go to the OGG1 page. What is the approved symbol and name for this gene and what is the chromosomal location?

   **Approved symbol: OGG1**
   **Approved name: 8-oxoguanine DNA glycosylase**
   **Location: 3p26**

2. There are many links to other databases from the OGG1 page. Follow the "UniProt" link under "Protein resources". What is the UniProtKB identifier for human OGG1? Is it in the Swiss-Prot or TrEMBL part of UniProtKB? (Check under "Status" close to the top of the page) What does that tell you about the quality of this entry? Can you find any information about the function of this protein? Under the "Sequences" heading you will find the sequence of OGG1 (Isoform 1A, the "canonical" sequence). How many residues/amino acids are there in this variant of OGG1? How many other isoforms are there?

   **The identifier is O15527 and this is an entry in UniProtKB/Swiss-Prot. This means that a human curator/expert actually has checked this entry. Function: "DNA repair enzyme that incises DNA at 8-oxoG residues. Excises 7,8-dihydro-8-oxoguanine and 2,6-diamino-4-hydroxy-5-N-methylformamidopyrimidine (FAPY) from damaged DNA. Has a beta-lyase activity that nicks DNA 3' to the lesion." There are 345 residues in the UniProtKB "canonical" sequence. There are 7 additional isoforms.**

3. Go back to the genenames.org webpage for OGG1 and follow the "GenBank" link under the "Nucleotide sequences" category. What are the accession identifier and GI number for this GenBank entry? What is the identifier/accession for the corresponding protein? What is the length of this protein (click on the "/protein_id" link to find out)?

   **Accession: U96710**
   **GI: 2078293**
   **Protein Accession: AAB81132**
   **Length of protein is 351 residues**

4. Go back to the genenames.org webpage for OGG1 and follow the "RefSeq" link under the "Nucleotide sequences" category. What are the accession identifier and GI number for this RefSeq entry? What is the identifier/accession for the corresponding RefSeq protein? What is the length of this protein (follow the "/protein_id" link 2/3 down the page)? Is this the same length you found when following the GenBank link? Is it the same as the UniProtKB "canonical" sequence? Do you have any idea why?

   **Accession: NM_016821**
   **GI number: 197276617**

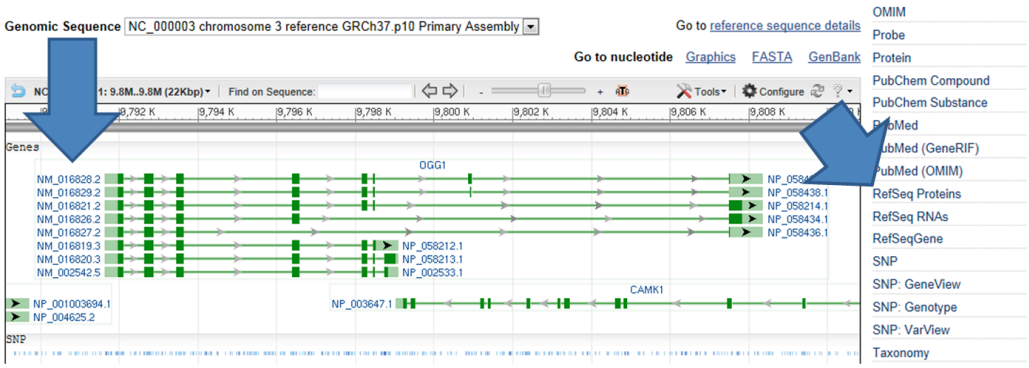
**Protein Accession: NP_058214**
**Length of protein is 424 residues**
**The lengths are not the same, most likely because they represent three different, predicted splice variants**

5. Go back to the genenames.org webpage for OGG1 and follow the "Ensembl" link under the "Gene resources" category. Use the first link. What is the identifier for the human OGG1 gene? How many transcripts have Ensembl listed for OGG1? How many of these are protein coding? Do any of the Ensembl proteins have the same length as any of the three splice variants investigated above? If so, what are the transcript IDs for these?

   **Human OGG1 gene Ensembl identifier is ENSG00000114026**
   **It is 18 transcripts, but only 14 (or 12) are predicted to encode a protein**
   **Protein with transcript ID ENST00000302036 has 424 residues**
   **None of the 14 has 351**
   **Protein with transcript ID ENST00000344629 has 345 residues**

6. Go back to the genenames.org webpage for OGG1 and follow the "Entrez Gene" link under the "Gene resources" category. What is the identifier (ID) for the human OGG1 gene in this database? How many transcripts are there listed for OGG1 here? In the "Related information" link on the right, click on "RefSeq Proteins". How many proteins do you find here? Do any of the RefSeq proteins have the same lengths as the splice variants investigated above (except Ensembl)? If so, what are the protein accession identifiers for these? Click on the links to these proteins to find the accessions for the corresponding transcripts. What are they?

   **The identifier is 4968**
   **There are 10 transcripts and 10 RefSeq proteins**
   **None are 351 residues**
   **NP_058214 has 424 residues, and the transcript is NM_016821**
   **NP_002533 has 345 residues, and the transcript is NM_002542**




7. Based on all the information above, what do you believe are the main functional isoforms of human OGG1 *in vivo*?

   **In brief, it is one big mess with 8 proteins in UniProtKB, 10 in RefSeq, and 14 proteins in Ensembl. None of these are the 351 residues protein which is the translation of the GenBank entry we were pointed to. UniProt/Swiss-Prot has a "canonical" sequence, but we find 7 other isoforms even in manually curated Swiss-Prot. It is far from obvious which variants are most important, if any. In this case, as often, the solution is *to read the literature!***

**According to Klungland *et al.* "OGG1: From structural analysis to the knockout mouse", in *Oxidative Damage to Nucleic Acids* (Landes Bioscience, Springer, 2007):**

*70* *Oxidative Damage to Nucleic Acids*

The cloning of the human *OGG1* gene also uncovered the existence of two splice variants, *α-hOGG1* and *β-hOGG1* with an open reading frame (ORF) coding for peptides of 345 and 424 amino acids, respectively.[12,27] The *β-OGG1* gene structure appears to be absent in rodents. Later studies have identified several alternatively spliced forms of human OGG1 mRNAs, with the two variants previously mentioned being predominant.[30] These two alternative spliced forms are localized to the nucleus (*α-OGG1*) and mitochondria (*β-OGG1*).[30]

**The 345 and 424 residues variants are α-OGG1 and β-OGG1, respectively, the nuclear and mitochondrial targeted isoforms. Quite possibly, many of the other splice variants we found above (in RefSeq, UniProt, and GenBank) represent "transcriptional junk". In addition, many of the additional Ensembl variants are only computationally predicted with some bioinformatics algorithm. These variants most likely exist only on the computer and not in human cells at all!**

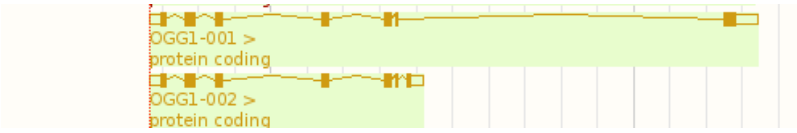
**This is important:**

1. **Some protein sequences in databases correspond to proteins that actually are present and have an important biological function in the organism – these are of course those that we are interested in studying**
2. **Some protein sequences in databases are translations of mRNAs that in reality never are properly translated in the organism – they do not represent biologically useful information**
3. **Many protein sequences in databases are translations of mRNAs that have been computationally predicted from the genome sequence. In many (most?) cases, not the mRNA and certainly not the protein, exists in the organism**

8. Go to the Ensembl page for the OGG1 gene again:

   http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000114026

   Note the names for α-OGG1 and β-OGG1. Halfway down the page you find a graphical illustration of the two isoforms under "Genes (Comprehensive Gene Annotations from GENCODE 23)". What are the differences between the two isoforms (in terms of splicing)?




**OGG1-001 is β-OGG1 while OGG1-002 is α-OGG1. The six 5' exons appear to be the same, but the 7th exon is different in the two variants.**

9. At the top of the page, click on the ENST00000344629 transcript (*i.e.* α-OGG1 or the 345 residues variant). At the left hand side click on "Protein" under "Sequence" to see the protein sequence. It is displayed with alternating black and blue colouring, corresponding to sequence encoded by exon 1, 2, 3 and so on. Residues spanning an intron are show in red. Take a screen shot of the window with the displayed sequence and paste it into PowerPoint, Word or some other suitable program. Crop the picture to show only the protein sequence and the explaining text.

**After cropping, the screenshot looks like this (actually slightly different in 2015):**

**Protein sequence**

Key

Exons | Alternating exons | Alternating exons | Residue overlap splice site

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
NFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG
```

10. Do the same for β-OGG1. What are the five last residues of the OGG1 protein that are identical in the two isoforms?

**Protein sequence**

Key

Exons | Alternating exons | Alternating exons | Residue overlap splice site

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
NFFRSLWGPYAGWAQAGLLGNAFDGHQLLRPLIFCQDHLREGPPIGRGDSQGEELEPQLF
SSLSSIPYGFCDHCWTKDVDDPPLVTHPSPGSRDGHMTQAWPVKVVSPLATVIGHVMQAS
LLAL
```

**Everything is identical until the sequence GWAQA, *i.e.* until the last intron. The 3' exon is different in the two variants.**

11. In order to see the exonic sequences click on "Exons" at the left hand side under "Sequence". Again take a screenshot and crop to get a nice picture. Which two nucleotides are at the start of all introns? And at the ends? What is the start codon of OGG1 and what is the stop codon? What are the lengths of the 5' and 3' introns in β-OGG1?

**Exons**

Show All entries | Show/hide columns | Filter

| No. | Exon / Intron | Start | End | Start Phase | End Phase | Length | Sequence |
|---|---|---|---|---|---|---|---|
| | 5' upstream sequence | | | | | | .....cctcctcctcggattggctacctctaggtgaaatgagcggtggttgagcc |
| 1 | ENSE00002259329 | 9,791,628 | 9,792,107 | - | 2 | 480 | CTACTTCCGGTGGTGCTGTGGTCTGCCCCTGGAGAACCCAGAAGAACACAGCTGTGCGCG<br>CCCACAGGCTCTGGGGGCGGGAGAAGATAAGTCGCAAGGAGGGGGCGGGACCTACACCTC<br>AGGAAAGCCGGAGAATTGGGGCACGAAGCGGGGCTTTGATGACCCGCAAAGGGCGAGGCA<br>TGCAGGAGGTGGAGGAATTAAGTGAAACAGGGAAGGTTGTTAAACAGCACCGTGTGGGCG<br>AGGCCTTAAGGGTCGTGGTCCTTGTCTGGGCGGGGTCTTTGGGCGTCGACGAGGCCTGGT<br>TCTGGGTAGGCGGGGCTACTACGGGGCGGTGCCTGCTGTGGAAATGCCTGCCCGCGCGCT<br>TCTGCCCAGGCGCATGGGGCATCGTACTCTAGCCTCCACTCCTGCCCTGTGGGCCTCCAT<br>CCCGTGCCCTCGCTCTGAGCTGCGCCTGGACCTGGTTCTGCCTTCTGGACAATCTTTCCG |
| | Intron 1-2 | 9,792,108 | 9,792,628 | | | 521 | gtgagtgactgagcctgagaagcct.....gggttgtcatgtgccttgggctcag |
| 2 | ENSE00001669842 | 9,792,629 | 9,792,876 | 2 | 1 | 248 | GTGGAGGGAGCAAAGTCCTGCACACTGGAGTGGTGTACTAGCGGATCAAGTATGGACACT<br>GACTCAGACTGAGGAGCAGCTCCACTGCACTGTGTACCGAGGAGACAAGAGCCAGGCTAG<br>CAGGCCCACACCAGACGAGCTGGAGGCCGTGCGCAAGTACTTCCAGCTAGATGTTACCCT<br>GGCTCAACTGTATCACCACTGGGGTTCCGTGGACTCCCACTTCCAAGAGGTGGCTCAGAA<br>ATTCCAAG |
| | Intron 2-3 | 9,792,877 | 9,793,453 | | | 577 | gtgagtacaggacctgggctggggt.....taccccctgcatttctggtctccag |
| 3 | ENSE00001644854 | 9,793,454 | 9,793,633 | 1 | 1 | 180 | GTGTGCGACTGCTGCGACAAGACCCCATCGAATGCCTTTTCTCTTTTATCTGTTCCTCCA<br>ACAACAACATCGCCCGCATCACTGGCATGGTGGAGCGGCTGTGCCAGGCTTTTGGACCTC<br>GGCTCATCCAGCTTGATGATGTCACCTACCATGGCTTCCCCAGCCTGCAGGCCCTGGCTG |
| | Intron 3-4 | 9,793,634 | 9,796,387 | | | 2,754 | gtgagtaggtgggtcccctgccccc.....ttgccctcctcctcactccccgcag |
| 4 | ENSE00001165146 | 9,796,388 | 9,796,569 | 1 | 0 | 182 | GGCCAGAGGTGGAGGCTCATCTCAGGAAGCTGGGCCTGGGCTATCGTGCCCGTTACGTGA<br>GTGCCAGTGCCCGAGCCATCCTGGAAGAACAGGGCGGGCTAGCCTGGCTGCAGCAGCTAC<br>GAGAGTCCTCATATGAGGAGGCCCACAAGGCCCTCTGCATCCTGCCTGGAGTGGGCACCA<br>AG |
| | Intron 4-5 | 9,796,570 | 9,798,154 | | | 1,585 | gtgaggccccagggggtaggagctg.....agggctcattctgtgtctgtcaaag |
| 5 | ENSE00002823518 | 9,798,155 | 9,798,305 | 0 | 1 | 151 | GTGGCTGACTGCATCTGCCTGATGGCCCTAGACAAGCCCCAGGCTGTGCCCGTGGATGTC<br>CATATGTGGCACATTGCCCAACGTGACTACAGCTGGCACCCTACCACGTCCCAGGCGAAG<br>GGACCGAGCCCCCAGACCAACAAGGAACTGG |
| | Intron 5-6 | 9,798,306 | 9,798,450 | | | 145 | gtgaggaaagtgggctgcaggggct.....cttagtctcatcacttctgatttag |
| 6 | ENSE00001764469 | 9,798,451 | 9,798,500 | 1 | 0 | 50 | GAAACTTTTTCCGGAGCCTGTGGGGACCTTATGCTGGCTGGGCCCAAGCG |
| | Intron 6-7 | 9,798,501 | 9,807,492 | | | 8,992 | gtgagtgtacctaggtgtcctccct.....ccatgctgccttccttgccctccag |
| 7 | ENSE00002896204 | 9,807,493 | 9,808,421 | 0 | - | 929 | GGCCTCCTTGGCAATGCATTTGATGGCCACCAGCTTCTGCGTCCTCTTATCTTCTGCCAG<br>GATCACCTCCGAGAAGGCCCCCCTATCGGGAGAGGGGATTCACAAGGTGAAGAACTGGAA<br>CCCCAGCTTCCCTCCAGCCTCTCCTCCATTCCCTATGGGTTCTGTGACCACTGCTGGACC<br>AAGGACGTGGATGACCCTCCCCTAGTCACTCATCCATCCCCTGGCTCCAGAGATGGTCAC<br>ATGACCCAGGCCTGGCCAGTCAAAGTAGTCTCTCCCCTGGCCACAGTAATTGGTCATGTG<br>ATGCAAGCCAGCTTACTAGCACTTTGAGAATGAGTCTCCTGTTGAGCTGGTAGGATGTAA<br>GCCTGGAGCTAATGGCGATCATCTTTGCCACCACCTGGGGAGAGCCTGCTTGGGAATGAA<br>ATTAACACAAAGGAAGTCCAACCTGAGAAATGGCCAAATATATTTCCTGATAACATTATG<br>TGGCCCTCTGGATCCAGCCATGCCTGAGGTCTACCCCTGGGCTTTTGGATTATGTGTACA<br>GTTGGTTCATCCCTTTTTCTGCTAATTCGAGTCATGGCTAATTTAACACCCTTTAGAACC<br>TTAAAGAACCATCAGCATCACCCGGGAACTTTTTTAGAAATGCAAAATCTCTACTGCTTT<br>GGATCCTGGGTCAAAAAAAAAGAAAAAAAAAAGAAATGCAAAACTTTAGGCCCTGCCCCAG |

**All introns start with GT and end with AG. This is very common. According to a large study (http://nar.oxfordjournals.org/content/34/14/3955.full.pdf) >98% of all introns are of the GT-AG type. Less than 1% is GC-AG type, and these often have less efficient splicing. This may lead to alternative splicing (See below). Start and stop codons are ATG and TGA, respectively. The 5' intron has 521 nucleotides, while the 3' intron has 8992 nucleotides.**

**Pagani and Baralle (*Nat. Rev. Genet.* 5, 389 (2004)) explain splicing as follows,**

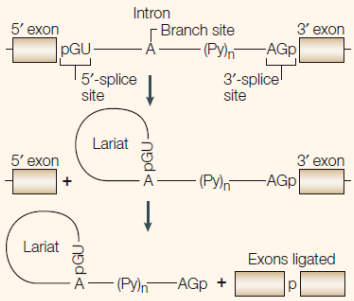
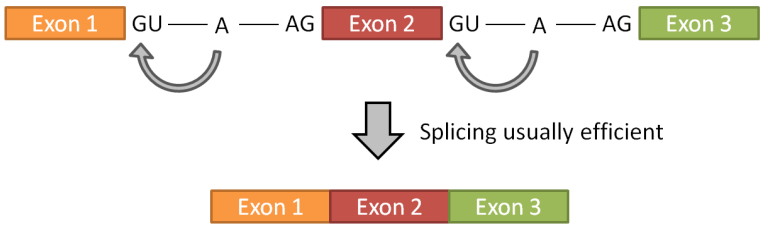
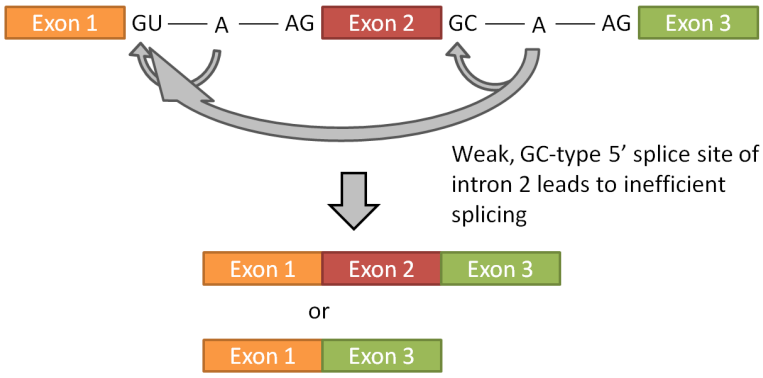



Figure 1 | **Splicing reactions and essential splicing signals.** There are several conserved motifs in the nucleotide sequences near the intron–exon boundaries that act as essential splicing signals: GU and AG dinucleotides at the exon–intron and intron–exon junctions, respectively (5′- and 3′-splice sites), a polypyrimidine tract $(Py)_n$ and an A nucleotide at the branch site. Splicing takes places in two TRANSESTERIFICATION steps. In the first step, the 2′-hydroxyl group of the A residue at the branch site attacks the phosphate at the GU 5′-splice site. This leads to cleavage of the 5′ exon from the intron and the formation of a lariat intermediate. In the following step, a second transesterification reaction, which involves the phosphate (p) at the 3′ end of the intron and the 3′-hydroxyl group of the detached exon, ligates the two exons. This reaction releases the intron, still in the form of a lariat.

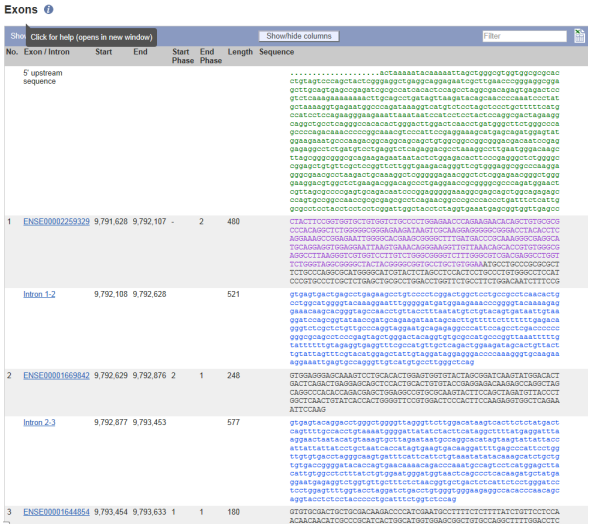


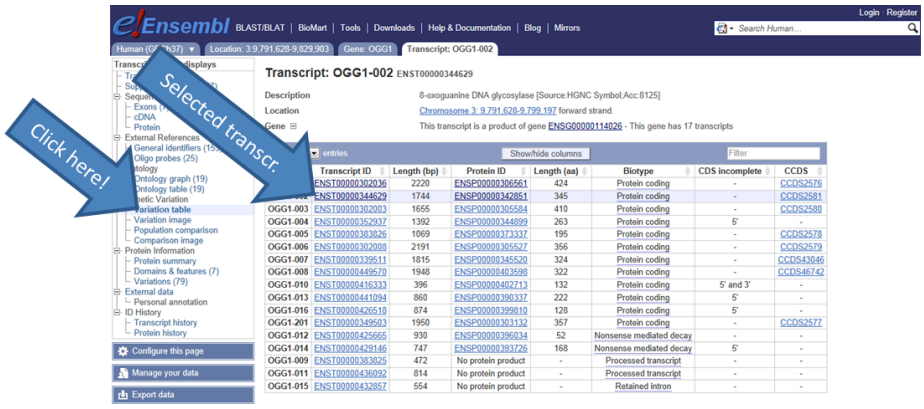
12. Check the "Exons" link for the other main transcript. What are the lengths of the 5' and 3' introns in α-OGG1?

    **The 5' intron has 521 nucleotides, obviously. It is exactly the same intron as in β-OGG1. The 3' intron has 244 nucleotides.**

13. Staying on the α-OGG1 "Exons" page, click on the "Configure this page" button on the left. Change "Flanking sequence at either..." to 1000 and tick "Show full intronic sequence". Then click on the little "tick mark" in the upper right-hand corner to save the changes. Again take a screenshot and crop to get a nice picture of the first exons and introns.




14. Make sure that the "OGG1-002"-line in the table is selected, *i.e.* it is light blue (as seen below) and click on "Variation table" under "Genetic variation" on the left.




You get a table with many SNPs and other variants found in human individuals/populations. Use the "Consequence Type" filter and show only "Missense variants". There are many columns. Try to find these columns: the ID for the SNP, position in chromosome, the two alleles, the allele frequency if it is known, the type of variation, the source database, the amino acid change (AA) and position of this residue in the protein (AA cord). The two columns at the right shows the output from two bioinformatics algorithms, SIFT and PolyPhen, that try to estimate if the amino acid change will affect the function of the protein.

Variation table

Variant table

This table shows known variants for this gene. Use the 'Consequence Type' filter to view a subset of these.

Show/hide columns

Filter × Consequence Type: Missense variant

| Variant ID | Chr: bp | Alleles | Global MAF | Class | Source | Evidence | Clin. Sig. | Type | AA | AA coord | SIFT | PolyPhen |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs1052133 | 3:9757089 | C/G | 0.302 (G) | SNP | dbSNP | | - | Missense variant | S/C | 326 | 0.18 | 0.23 |
| rs1805373 | 3:9754824 | G/A | 0.028 (A) | SNP | dbSNP | | - | Missense variant | R/Q | 229 | 0.09 | 0.168 |
| rs113561019 | 3:9756791 | G/A/T | 0.002 (A) | SNP | dbSNP | | - | Missense variant | G/E | 308 | 0 | 1 |
| rs113561019 | 3:9756791 | G/A/T | 0.002 (A) | SNP | dbSNP | | - | Missense variant | G/V | 308 | 0 | 1 |
| rs104893751 | 3:9750423 | G/A | 0.001 (A) | SNP | dbSNP | | | Missense variant, Splice region variant | R/Q | 46 | 0 | 0.999 |
| rs201580660 | 3:9751042 | C/G/T | 0.001 (G) | SNP | dbSNP | | - | Missense variant | R/G | 79 | 0.21 | 0.018 |

Are there any missense variants in human α-OGG1 that have a relatively high frequency? Which residue does it affect and what is this residue mutated to? What do SIFT and PolyPhen think about this mutation?

**The SNP rs1052133 has a minor allele frequency of 30.2%. This is a Ser326 to Cys mutation, often written as Ser326Cys or S326C. No other SNPs have a frequency of more than 2.8%. For S326C SIFT predicts "tolerated" and PolyPhen predicts "benign". According to these predictors, S326C will most likely not change the function of α-OGG1 significantly.**

15. Click on the rs1052133 link to go to the Ensembl page for this SNP/variation. Then click on the "Population genetics" link at the left. In the 1000 Genomes project data (Phase 3), which population has the highest ratio of G|G genotype (which means homozygote for Cys326)? Which population has the lowest? In each case, what are the G|G genotype ratios?

    **The highest is in the CDX population, Chinese Dai in Xishuangbanna, China. 42% have a G|G genotype. The lowest is in the TSI population (Toscana, Italy) with 1.0%. Also YRI (Yoruba, Nigaria), and other African and European populations are low (less than 2.5%).**

16. Can you find anything in the scientific literature about this mutation? Are their any studies on the effect of the OGG1 Cys/Cys genotype? Look for example here: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0035970

    **According to several studies, individuals with the OGG1 Cys/Cys variant are more prone to get lung cancer.**

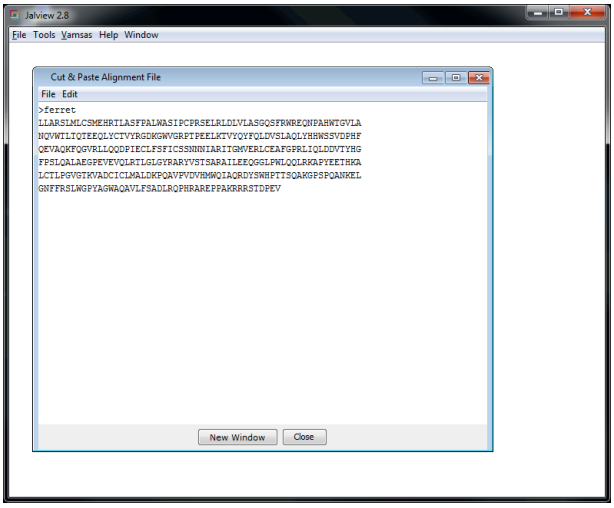
17. Back in Ensembl, lick on the "Gene:OGG1" tab at the top of the page to go back to the information about the gene. Click on the "Orthologues" link at the left. What are orthologous genes? Is there an OGG1 ortholog in the ferret? Click on the ferret OGG1 gene identifier link (marked ENSMPUG00000017122). Click on the "Transcript ID" and then on "Protein" under "Sequence" at the left.

    **"Homologous sequences are orthologous if they were separated by a speciation event: when a species diverges into two separate species, the copies of a single gene in the two resulting species are said to be orthologous." Yes, there is an OGG1 gene in the ferret genome.**

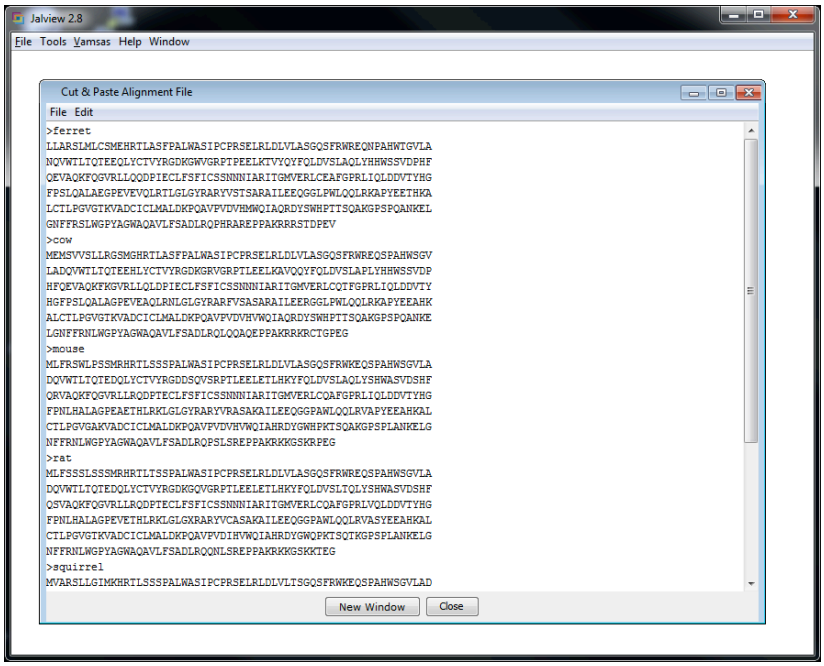
    Now go to the Jalview website (http://www.jalview.org) and "Launch Jalview Desktop" in the upper right-hand corner. Do "File" - "Input Alignment" - "from Textbox". Copy the protein

sequence from Ensembl and paste it into the textbox window with the header "ferret". Use FASTA format, as you see below. If you do not know what that is, read about it here:

http://en.wikipedia.org/wiki/FASTA_format




18. Go back to the human OGG1 "Orthologues" page. Do exactly as you did for the ferret OGG1 for the following species: cow, mouse (longest sequence), rat, squirrel, and human (α-OGG1). That is, go to the gene, find the protein sequence, copy the sequence and paste it into the Jalview textbox under the ferret sequence (See below). When you believe you know how to do this and you have had enough practising, you can copy the remaining sequences from below.

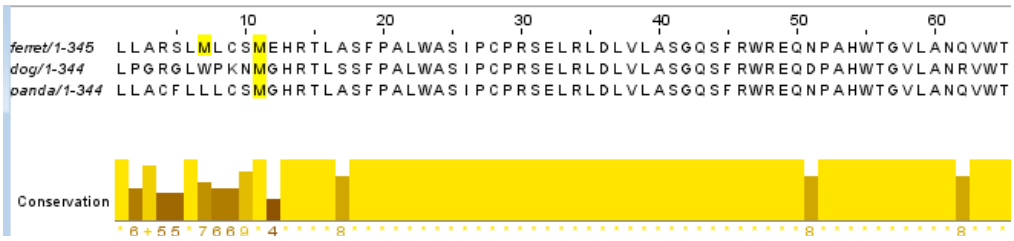



```
>ferret
LLARSLMLCSMEHRTLASFPALWASIPCPRSELRLDLVLASGQSFRWREQNPAHWTGVLA
NQVWTLTQTEEQLYCTVYRGDKGWVGRPTPEELKTVYQYFQLDVSLAQLYHHWSSVDPHF
QEVAQKFQGVRLLQQDPIECLFSFICSSNNNIARITGMVERLCEAFGPRLIQLDDVTYHG
FPSLQALAEGPEVEVQLRTLGLGYRARYVSTSARAILEEQGGLPWLQQLRKAPYEETHKA
LCTLPGVGTKVADCICLMALDKPQAVPVDVHMWQIAQRDYSWHPTTSQAKGPSPQANKEL
GNFFRSLWGPYAGWAQAVLFSADLRQPHRAREPPAKRRRSTDPEV
>cow
MEMSVVSLLRGSMGHRTLASFPALWASIPCPRSELRLDLVLASGQSFRWREQSPAHWSGV
LADQVWTLTQTEEHLYCTVYRGDKGRVGRPTLEELKAVQQYFQLDVSLAPLYHHWSSVDP
```

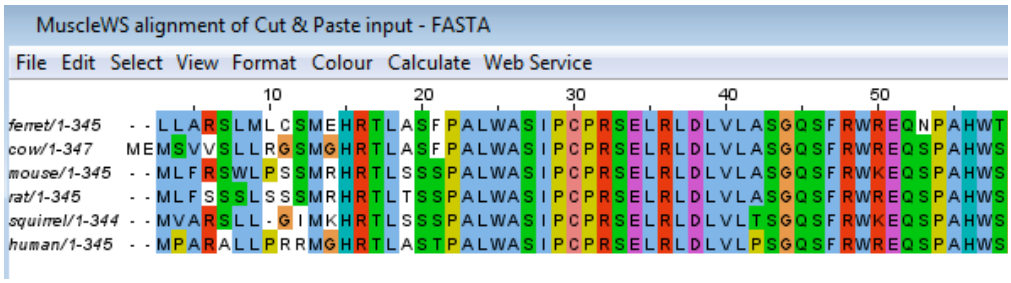
```
HFQEVAQKFKGVRLLQLDPIECLFSFICSSNNNIARITGMVERLCQTFGPRLIQLDDVTY
HGFPSLQALAGPEVEAQLRNLGLGYRARFVSASARAILEERGGLPWLQQLRKAPYEEAHK
ALCTLPGVGTKVADCICLMALDKPQAVPVDVHVWQIAQRDYSWHPTTSQAKGPSPQANKE
LGNFFRNLWGPYAGWAQAVLFSADLRQLQQAQEPPAKRRKRCTGPEG
>mouse
MLFRSWLPSSMRHRTLSSSPALWASIPCPRSELRLDLVLASGQSFRWKEQSPAHWSGVLA
DQVWTLTQTEDQLYCTVYRGDDSQVSRPTLEELETLHKYFQLDVSLAQLYSHWASVDSHF
QRVAQKFQGVRLLRQDPTECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPNLHALAGPEAETHLRKLGLGYRARYVRASAKAILEEQGGPAWLQQLRVAPYEEAHKAL
CTLPGVGAKVADCICLMALDKPQAVPVDVHVWQIAHRDYGWHPKTSQAKGPSPLANKELG
NFFRNLWGPYAGWAQAVLFSADLRQPSLSREPPAKRKKGSKRPEG
>rat
MLFSSSLSSSMRHRTLTSSPALWASIPCPRSELRLDLVLASGQSFRWREQSPAHWSGVLA
DQVWTLTQTEDQLYCTVYRGDKGQVGRPTLEELETLHKYFQLDVSLTQLYSHWASVDSHF
QSVAQKFQGVRLLRQDPTECLFSFICSSNNNIARITGMVERLCQAFGPRLVQLDDVTYHG
FPNLHALAGPEVETHLRKLGLGXRARYVCASAKAILEEQGGPAWLQQLRVASYEEAHKAL
CTLPGVGTKVADCICLMALDKPQAVPVDIHVWQIAHRDYGWQPKTSQTKGPSPLANKELG
NFFRNLWGPYAGWAQAVLFSADLRQQNLSREPPAKRKKGSKKTEG
>squirrel
MVARSLLGIMKHRTLSSSPALWASIPCPRSELRLDLVLTSGQSFRWKEQSPAHWSGVLAD
QVWTLTQTEELLYCTVYRGDKGWVGKPTPEELETVHKYFQLDVSLAQLYSHWSSVDSHFQ
KMAQKFQGVRLLRLDPIECLFSFICSSNNNITRITGMVERLCQAFGPRLIQLDDVTYHGF
PTLQALAGSEVEACLRKLGLGYRAQYVSASARAILEEQGGLAWLQQLREAPYEEAHKALC
TLPGVGTKVADCICLMALEKPQAVPVDVHVWQIAQRDYSWHPTTSQAKGPSPQANKELGN
FFRRLWGPYAGWAQAVLFSADLHQPHRSQEPPAKRKKRSKGPEV
>human
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
NFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG
```

19. In the Jalview textbox window, press "New Window". You get a window with 6 sequences, but they are not optimally aligned. In this window, do "Web service" - "Alignment" - "Muscle with defaults". This will run the Muscle multiple sequence alignment program on a machine somewhere else on the internet (most likely in Dundee, Scotland) and send the result back to your Jalview session. Do all the sequences have a Met residue at the N-terminus? If not, which one is missing it? Why do you think it is missing?

**The N-terminal residue of the ferret sequence is Leu, all the other have Met. It is unlikely that the Leu is the actual start codon of this sequence. More likely, the start codon is Met7 (which in that case is Met1), or possibly Met11. Actually, there are some indications that it is Met11 in the ferret sequence that is the actual start codon. If you get the other available carnivore sequences from Ensembl (the cat sequence is missing the N-terminus), you see (below) that Met11 is conserved in all 3 while no other Met (ATG) codons are near the N-terminus. Also the 3 sequences are nearly 100% conserved downstream of Met11, but not upstream, indicating that this actually might be part of the 5' UTR. If you take a look at the 5' UTRs for these 3 carnivore sequences in Ensembl, you will find that there are no in-frame ATG start codons nearby and upstream in the sequence.**

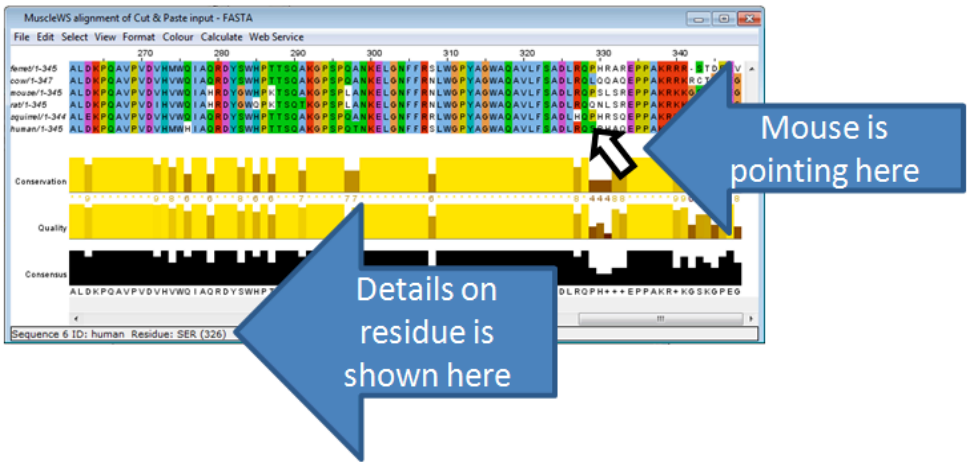



20. Are all the other sequences the same length at the N-terminus? If not, which is longer? Is it certain that it actually should be longer than the others?




**The cow sequence appears to have an extra 2 N-terminal residues. However, it also has a Met3. The nucleotide sequence is therefore ATGxxxATG. It is difficult from the sequence alone to determine if it is the first or second ATG that is the actual start codon. Quite likely no-one has checked experimentally if bovine protein OGG1 has MEMSVVS or MSVVS at the N-terminus, so we cannot know for sure. It is also possible that both forms exist, and that the function of the protein is identical for the two variants.**

21. Show the multiple sequence alignment with Clustalx colouring. In the Jalview multiple sequence alignment window, you can move the mouse over the residues. You will then see the identity of the residue you are pointing to in the lower left corner of the window (See below).




Use this technique to find human α-OGG1 Ser326. Is this residue conserved in other mammals? Is Ser320 conserved?

**Ser326, the residue affected by SNP rs1052133 investigated above, is not conserved in the other mammals. Ser320 is 100% conserved in the 6 mammals.**

22. Find ferret OGG1 Glu189. Is there anything odd here? Can you find out what?

**There is a single residue insertion in ferret OGG1 here, in a quite conserved segment of the protein. We better check it in Ensembl... The protein is found here:**

**http://www.ensembl.org/Mustela_putorius_furo/Transcript/Sequence_Protein?db=core;g=ENSMPUG00000017122;r=GL896899.1:32715972-32721358;t=ENSMPUT00000017267**

**Glu189 is encoded partly by exon 3 and exon 4. We click on the cDNA link at the left and locate the relevant part (below).**

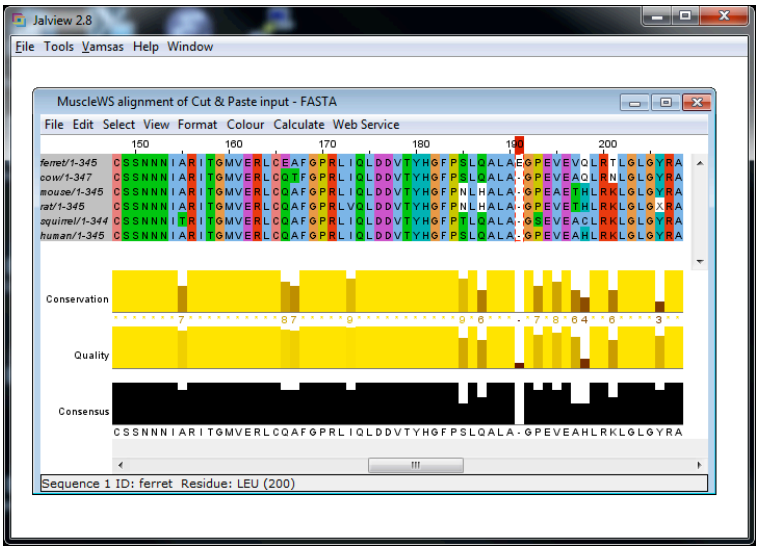
```
481 CGACTCTGTGAGGCCTTTGGACCTCGGCTCATCCAGCTTGATGATGTCACCTACCATGGC
481 CGACTCTGTGAGGCCTTTGGACCTCGGCTCATCCAGCTTGATGATGTCACCTACCATGGC
161 -R--L--C--E--A--F--G--P--R--L--I--Q--L--D--D--V--T--Y--H--G-

541 TTCCCTAGCCTGCAGGCCCTGGCTGAAGGGCCAGAGGTAGAGGTGCAGCTCAGGACGCTG
541 TTCCCTAGCCTGCAGGCCCTGGCTGAAGGGCCAGAGGTAGAGGTGCAGCTCAGGACGCTG
181 -F--P--S--L--Q--A--L--A--E--G--P--E--V--E--V--Q--L--R--T--L-

601 GGCCTGGGGTACCGTGCCCGCTATGTGAGTACCAGTGCCCGAGCCATCCTAGAAGAACAG
601 GGCCTGGGGTACCGTGCCCGCTATGTGAGTACCAGTGCCCGAGCCATCCTAGAAGAACAG
201 -G--L--G--Y--R--A--R--Y--V--S--T--S--A--R--A--I--L--E--E--Q-
```

**We have the codons CTG-GCT-GAA-GGG-CCA, that encodes L-A-E-G-P. However, if we guess that AAG with the yellow background is actually a part of the intron we get the protein sequence L-A-G-P. This splicing is also perfectly ok, since the intron ends with AG. We also get rid of the insertion, the extra Glu residue. If we click on the "Exons" link, we see that introns 1, 2 and 4-6 are normal GT-AG introns. However, intron 3 that we are fiddling with now is GT-AA! Impossible!!! For some unknown and stupid reason, the gene searching algorithm made a GT-AA intron, instead of a perfectly ok GT-AG intron. This created a most likely completely artificial insertion in the protein. Extremely likely, Glu189 does not exist *in vivo*, just in the computer...**

23. Let us remove ferret Glu189 in JalView. Select the column containing Glu189 by clicking just above the ferret sequence. You get a red square as seen below. Then click "Edit" - "Delete" to get rid of this quite likely erroneous insertion.




24. In the middle of ferret OGG1 intron 3 (back in the Ensembl "Exons" page), there are a lot of nnnnnnnn. Why is that?

**This is because the genomic sequence here is unknown. This part of the genome has not yet been sequenced.**

25. Go back to UniProt (http://www.uniprot.org) and click on the UniRef link. Search (at the top) in UniRef for human OGG1, *i.e.* O15527. Ideally, O15527 should be in a single UniRef50 cluster, *i.e.* a group of homologs with more than 50% sequence identity. However, the various splice variants are currently (Nov 2015) found in 3 UniRef50 clusters. We will also see below that sequences that are clearly more than 50% identical to human OGG1 is found in none of the 3 clusters. Clustering in UniRef does not appear to function very well… Focus on the UniRef50 cluster with the canonical

O15527 variant. Click on the identifier for this cluster. How many proteins are there in this OGG1 50% cluster?

**There are currently 46 proteins, but this might change with the next UniProt update.**

26. Select the following sequences (by setting a tick mark at the left): O15527 (human α-OGG1), K7AAZ7 (Chimpanzee), F7B0F5 (Rhesus macaque), F7IC54 (marmoset), and G3RZB6 (gorilla). Choose "Add to basket" and select "Selected cluster members". Open your shopping basket and click "Full View". Choose "Download" – "FASTA (canonical) – Preview and copy the sequences into Jalview (using the input from textbox option). Also find, i UniProt, the sequence H0XK07 from the galago and add it to the text box. Show the alignment by clicking "New window". We do not need to run Muscle. The sequences are already aligned and they all have the same length. Find the human α-OGG1 Ser326 residue. Does it appear to be conserved in primates?

**Ser326 is conserved in all great apes, in the New World monkey (marmoset), and Old World monkey (macaque), that is all the "higher primates". It is not conserved in galago, a prosimian, related to lemurs.**

**This could mean that Ser326 has some function in higher apes, but it could also be that, by chance, no mutation has yet happened at this position.**

27. Go back to http://www.uniprot.org/uniprot/O15527 again. Can you find any information about S326C on this page?

**It says "Common polymorphism in the Japanese population" (with some literature references) and "Corresponds to variant rs1052133".**

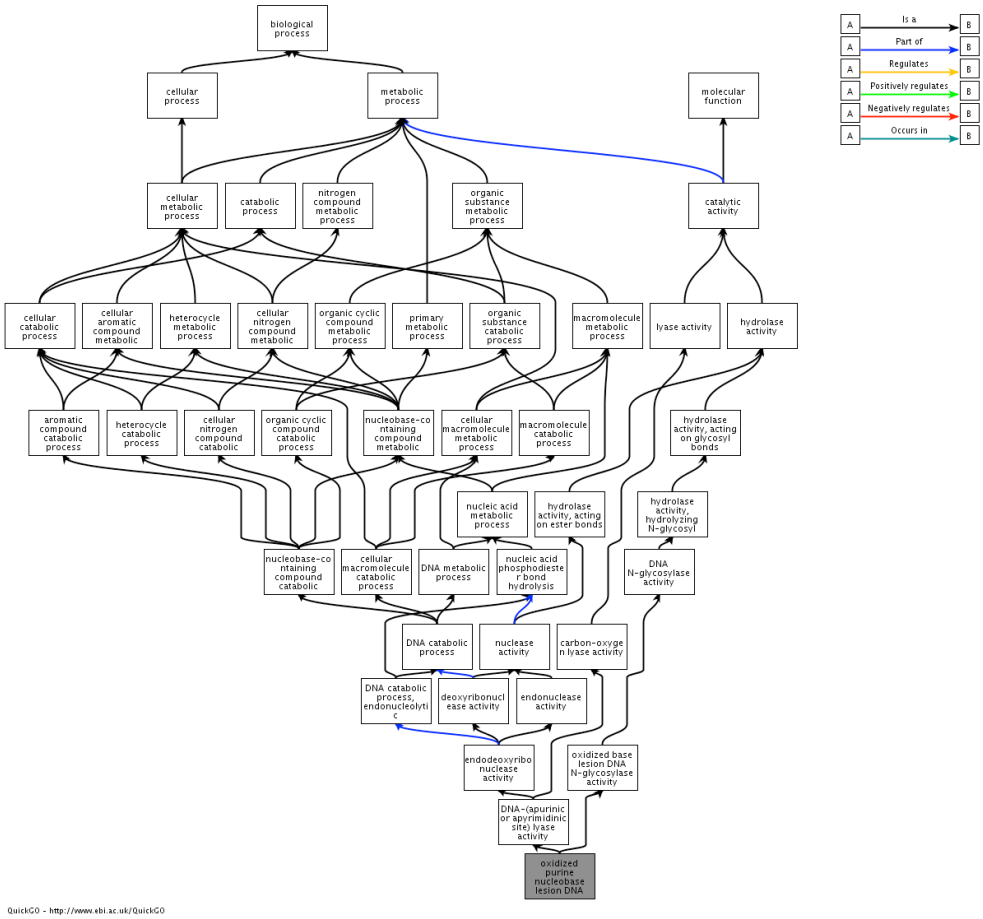
28. Can you find any Molecular function GO terms for OGG1 on this page? Give the ID for one of them.

**Yes, "damaged DNA binding", "endonuclease activity", and "oxidized purine nucleobase lesion DNA N-glycosylase activity" (GO:0008534, had to click on the link to find this).**

29. Click on the "oxidized purine nucleobase lesion DNA N-glycosylase activity" GO term link. Click on the "Protein Annotation" tab to see other proteins given this annotation. Are their many proteins here? More than 50?

**Yes, there are 8,989 proteins at present, and this number will most likely increase fast.**

30. Click on the "Ancestor Chart" to see the ancestor GO terms. Take a screenshot and paste the chart below.




31. Let us go back to the human OGG1 Ensembl entry to get an impression of how much data you can find for OGG1 homologs there (and how messy it is):

http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000114026

Click on the "Ensembl protein families" under "Comparative Genomics" at the left. Ensembl currently has 97 proteins in this family (Nov 2015). Including other proteins from UniProtKB there is also a cluster of 533 OGG1 homologs. Download the Ensembl OGG1 homologs and open the multiple sequence alignment in Desktop Jalview. **Try this first:** Click on "Jalview" in "97 Ensembl members of this family JalView". Then click on the "Start Jalview" button. You have hopefully opened the JalviewLite applet with less functionality that Desktop Jalview. ***If not, see below!*** Choose "File" - "Output to text box", choose "FASTA". Then copy all the text in the window and paste it into a textbox in Desktop Jalview. If you are unable to open JalviewLite, **try this instead:** On the "Ensembl protein families" page, click on the "Family ID" link ENSFM00250000003270 to get to the page for this protein family. Click on "Download family alignment", choose FASTA file format, Uncompressed format and "Preview". We only want the 97 Ensembl sequences. Either copy only the 97 first sequences (the ones that has identifiers starting with ENS, but not the UniProtKB sequences such as G3GX91) into Desktop Jalview or download all 533 sequences as a file, open the file in Desktop Jalview and remove all the sequences at the bottom that are not from Ensembl.

You have now opened the multiple sequence alignment of 97 Ensembl OGG1 homologs in Desktop Jalview. Try to tidy up the data by removing sequences that "obviously are wrong", for example sequences that are missing one or more exons, have long segments of missing sequence (that is XXXXXXXXXX), or have insertions that are found in one species only. "Edit" - "Remove Empty Columns" is useful for tidying up. See my result below…

In order to get a good set of sequences to work with one can, as here, start with a large amount and clean up by removing sequences that are very likely to be incorrect. The other possibility is to start with one or a few sequences and add more sequences that appear fine, one by one, similar to what we did earlier in this exercise. What is obvious, however, is that "all homologs" contains such a mess that is cannot be used for much!


*Make sure you are understand everything in this exercise and are able to do all the manipulations. If you are stuck, please ask for help!*