Sequence searching and sequence alignments – MBV-INFX410

In this exercise we will start with a bacterial DNA repair protein called Nth and identify its homologs in different species, including humans, using BLAST and PSI-BLAST, and then identify conserved sequence motifs using multiple alignments. It is a good idea to create a report document in Word (or a similar editor) were you describe briefly what you do, save the sequences that you work with and answer the questions that are asked. You must also save screen shots of what you do in your report. *Make sure you know how to do this!*

1. Find the RefSeq protein sequence of the Endonuclease III (Nth) protein from the bacterium *Escherichia coli*, strain K-12, substrain MG1655, using Entrez, in the NCBI protein database ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)). *First try yourself, without checking below!*

**There are many ways to do this, but what we are looking for is the RefSeq sequence NP\_416150. One possibility is to search for “Escherichia coli” AND “Endonuclease III” AND MG1655 in the Protein database and then filter for RefSeq in “Source databases”. You will then have something like 13 candidates and among these the only one that is both MG1655 and “Endonuclease III” is NP\_416150. *Make sure you understand how you find a sequence by searching like this!***

1. Get the FASTA sequence for the protein and paste it into your report document.

**>gi|16129591|ref|NP\_416150.1| DNA glycosylase and apyrimidinic (AP) lyase (endonuclease III) [Escherichia coli str. K-12 substr. MG1655]**

**MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIAVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE**

**GVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH**

**IFRVCNRTQFAPGKNVEQVEEKLLKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVD**

**I**

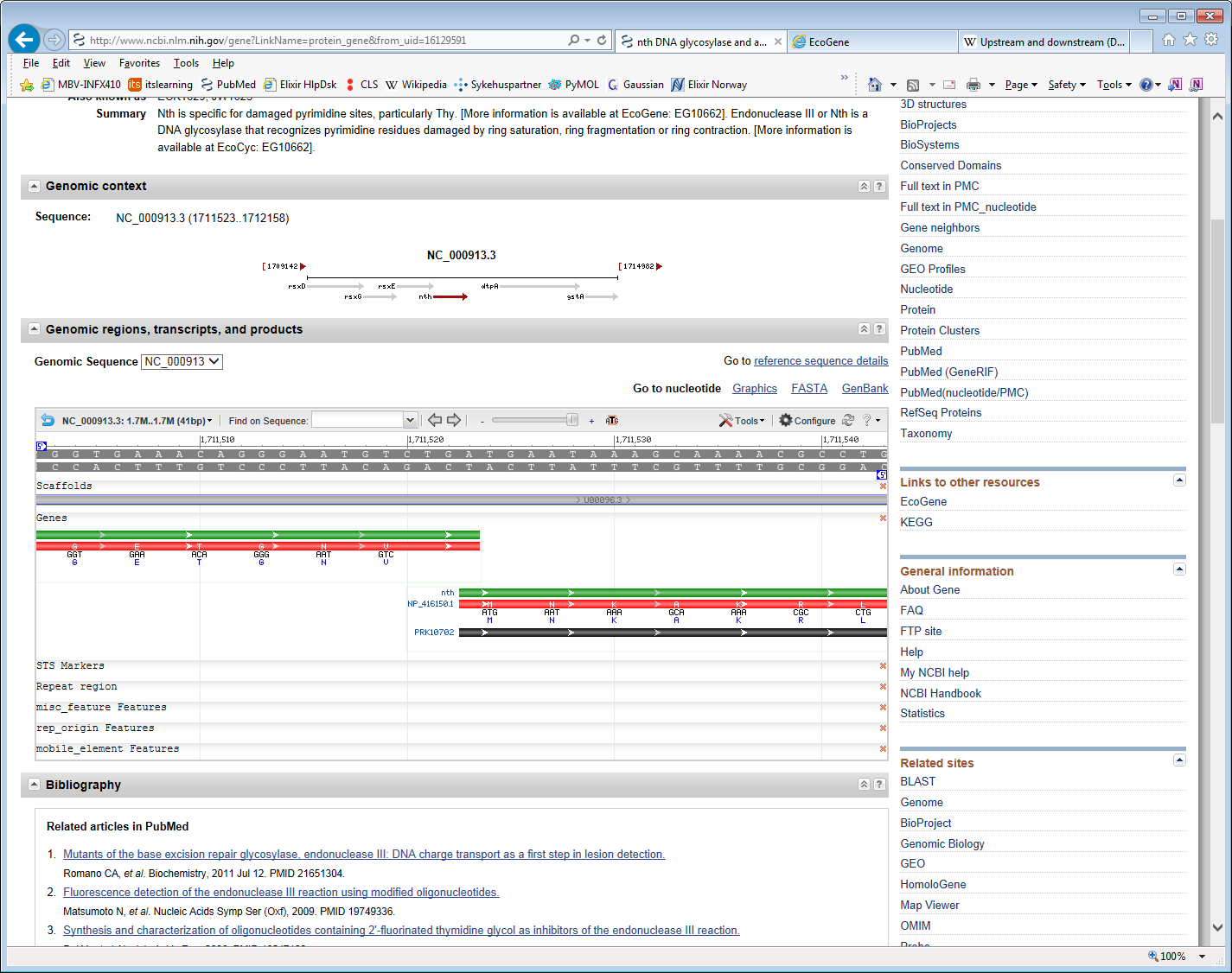
1. Follow the link to the corresponding Gene (in the list of “Related information” at the right hand side). What is the NCBI gene identifier (Gene ID) for the gene? What is the Swiss-Prot identifier for this protein? Which genes are found directly upstream and downstream of *nth*? Are the three genes transcribed in the same direction?

**The Gene ID is 947122 and the Swiss-Prot ID is P0AB83. The two neighbouring genes are *rsxE* and *dtpA*. All three genes are transcribed in same direction.**

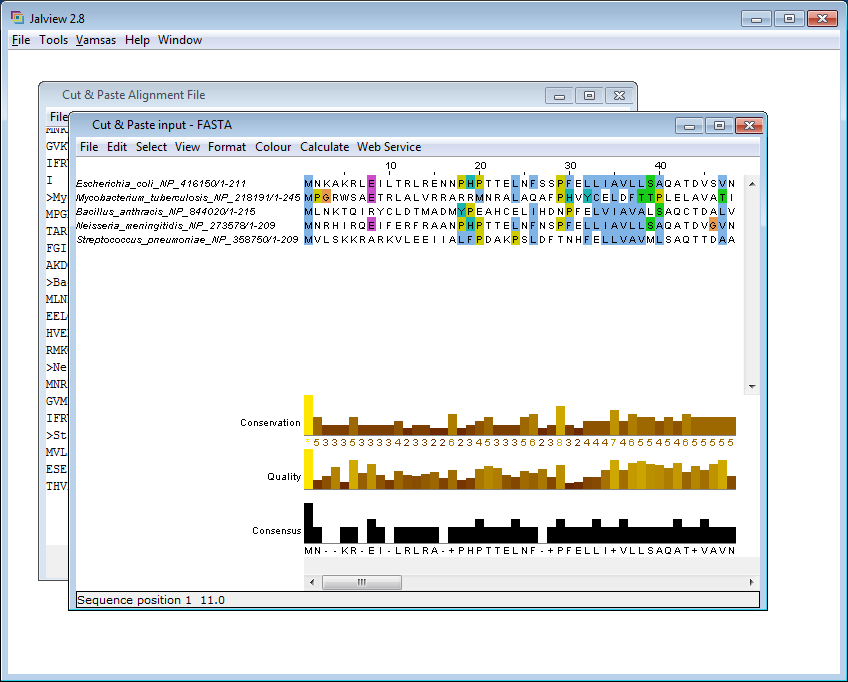
1. In the simple genome browser on the NCBI Gene page (“Genomic regions, transcripts, and products” section) click and drag the genome to centre the region where you have the start of *nth* and the stop of *rsxE*. Then zoom in all the way to the highest possible magnification by using the slider and/or the “+” and “ATG” buttons. Make sure the start of *nth* stays in the middle of your browser by click-dragging, if necessary. What are the three nucleotides of the codon encoding the last (C-terminal) amino acid in *rsxE*? What are the nucleotides of the stop codon of *rsxE* and the start codon of *nth*? How many nucleotides are there between the start of *nth* and the stop of *rsxE*?

**The last amino acid of *rsxE* is Val, encoded by GTC, and the stop codon is TGA. The start codon of *nth* is encoded by ATG. In this case the A of TGA (stop) is the same as the A of ATG (start). The two genes overlap by a single nucleotide, and there are, obviously, no nucleotides between them.**

**Notice how densely packed the genes are in bacteria compared to, for example, the vertebrates.**



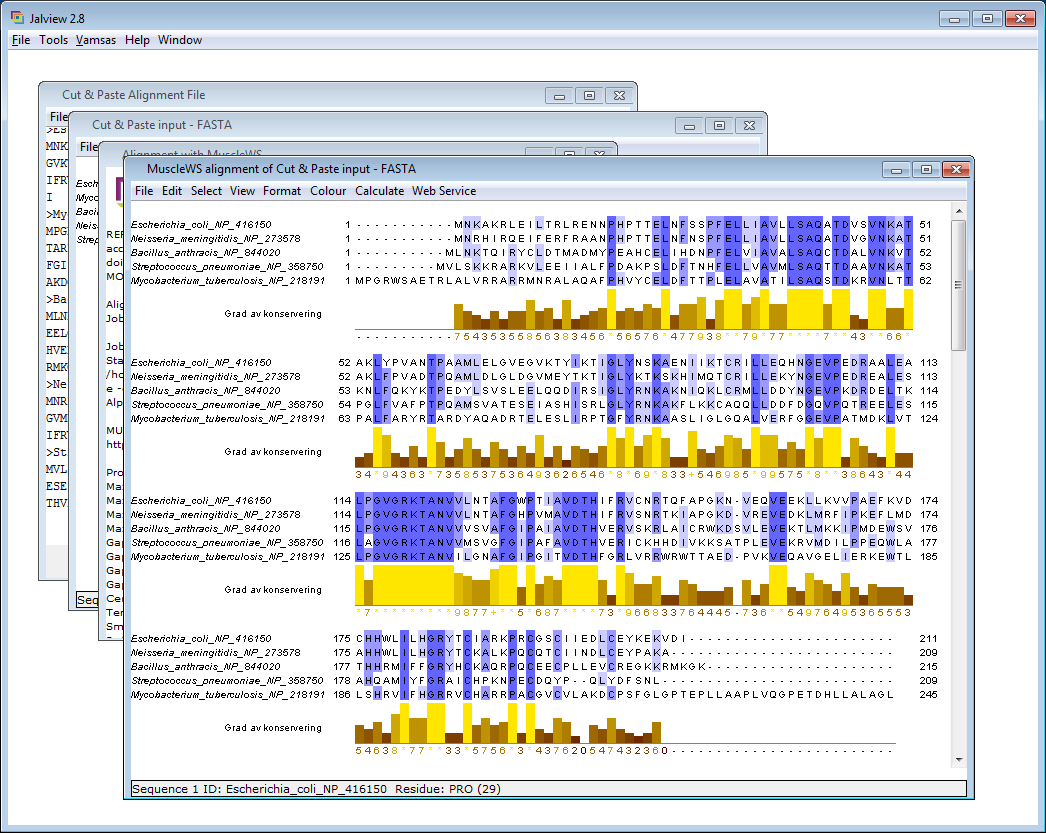
1. Get the homologous sequences of the Nth protein from *Mycobacterium tuberculosis* strain H37Rv (GI number 57117142), *Bacillus antracis* strain Ames (GI number 30261643), *Neisseria meningitidis* strain MC58 (GI number 15676439), and *Streptococcus pneumoniae* strain R6 (GI number 15903200) in FASTA format, and copy them into your report.
2. Edit the sequence titles to contain only the name of the bacteria and the RefSeq identifiers. Replace the spaces with the underscore character (“\_”), but keep the initial larger-than character (“>”). For your first sequence, the header will then be “>Escherichia\_coli\_NP\_416150”.
3. Start Desktop JalView. Use “File” → “Input Alignment” → “from Textbox” to enter the five bacterial Nth sequences by copying and pasting. Click on “New Window”. Take a screenshot of Jalview with the input sequences and paste the image into your report.



1. Use the MUSCLE-algorithm web service (found under “Web Service” → “Alignment”) (with “Muscle with Defaults”) to generate a multiple sequence alignment (MSA). Can you say anything about on which computer the MUSCLE-algorithm is running? Where on the planet? Hint: Look in “Tools” → “Preferences” → “Web Services”.

**The job is running as a web service on a server that at least contains the name “dundee”. The service is, very likely, running in Dundee, Scotland, on a server belonging to the group of Professor Geoff Barton. This is the group that is developing Jalview.**

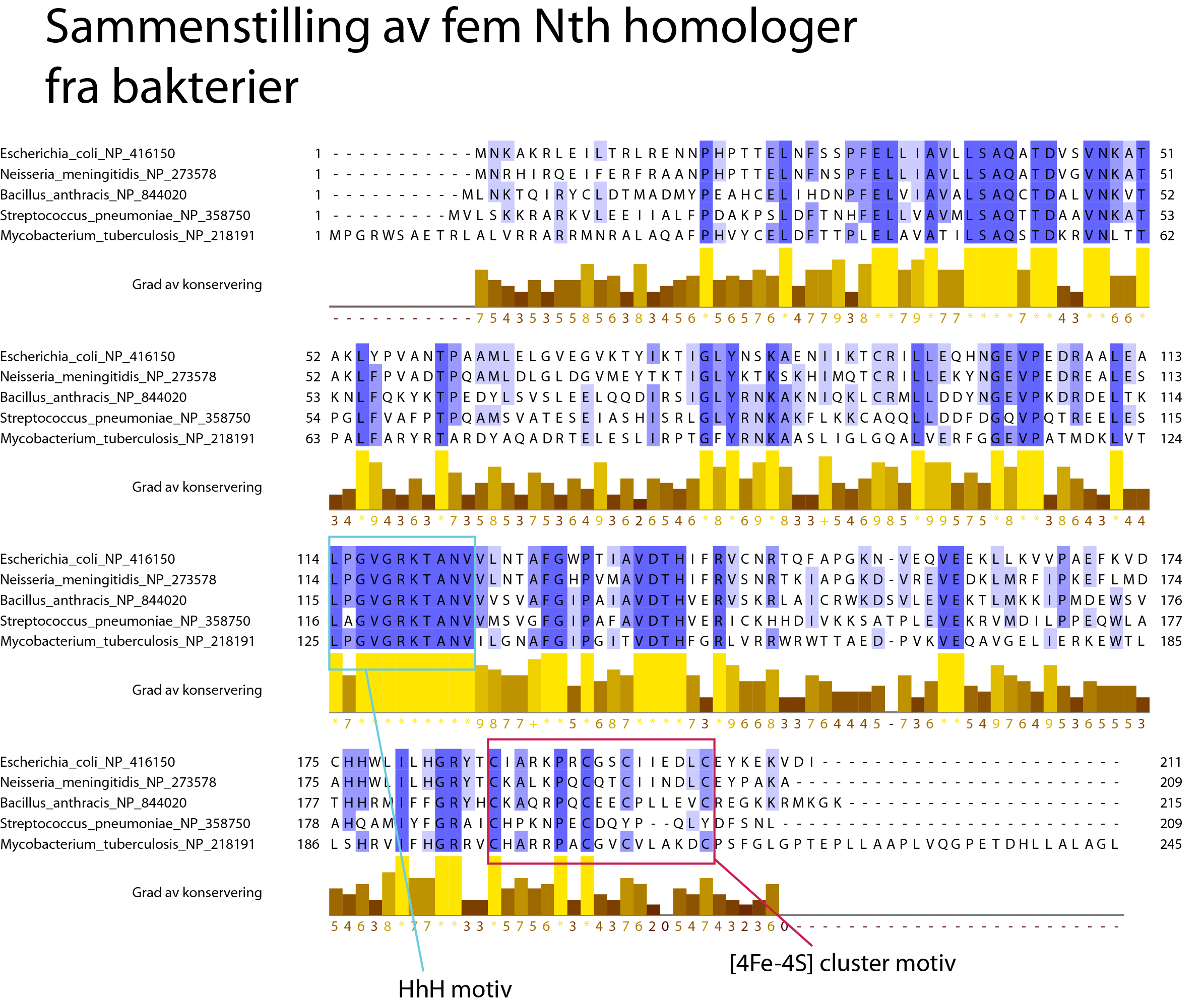
1. Colour the amino acids according to “Percentage identity”. Remove the “Quality” annotation information in the lower part of the window by right-clicking on the word “Quality” and choose “Hide this row”. Do the same for the “Consensus” annotation. Right-click on “Conservation” and choose “Edit Label/Description”. Change the “Annotation name” to your native language. For example, in Norwegian use the text “Grad av konservering”, then click “OK”. Sort the sequences by pairwise similarity (“Calculate” → “Sort” → “By Pairwise Identity”). Reformat the alignment to make it more compact (“Format” → “Wrap”). Adjust the width of the window so that you get the MSA split into 4 lines/blocks. Also remove the tick mark at “Show Sequence Limits” under “Format”.



Export the alignment in PNG format, and import it into PowerPoint, Adobe Illustrator, or a similar program in order to add some extra information in the MSA. Indicate the residues involved in the helix-hairpin-helix (HhH) motif (LxGVGxK) and the [4Fe–4S] (iron sulphur) cluster motif (CxxxxxxCxxCxxxxxC). See Fig. 3 in the article below for more information about these motifs. Copy the resulting figure into your report. Are both motifs fully conserved in all sequences?

N. Goosen & G.F. Moolenaar, “Repair of UV damage in bacteria”, DNA Repair **7**, 353 (2008) [**http://dx.doi.org/10.1016/j.dnarep.2007.09.002**](http://dx.doi.org/10.1016/j.dnarep.2007.09.002)

**The HhH motif is conserved in all species, with GVGRKTANV being fully conserved. The [4Fe-4S] cluster is conserved in all species except Streptococcus, which lacks the two last cysteines.**



1. Find the percentage sequence identity between the five sequences. First select all the sequences in the Jalview MSA window, for example by typing <ctrl>-a. Btw, if you want to select nothing, press the <Esc> key. Try this. Select all sequences again and do “Calculate” → “Pairwise Alignments…”. Look at the pairwise alignments and find which two sequences are the most similar. Which are they? What is the sequence identity between those two sequences?

***E. coli* and *N. meningitidis* are 72% identical while no other pairs are above 47%**

1. Using the sequence from *E. coli* Nth as query, perform a protein BLAST (blastp) search at the NCBI website (<http://blast.ncbi.nlm.nih.gov>). Use “Basic BLAST” and “protein blast” and search in the RefSeq protein database. Limit the search to protein sequences from vertebrates. Set the max target sequences options to 500 under algorithm parameters. Why do we choose blastp in this case and not tblastn?

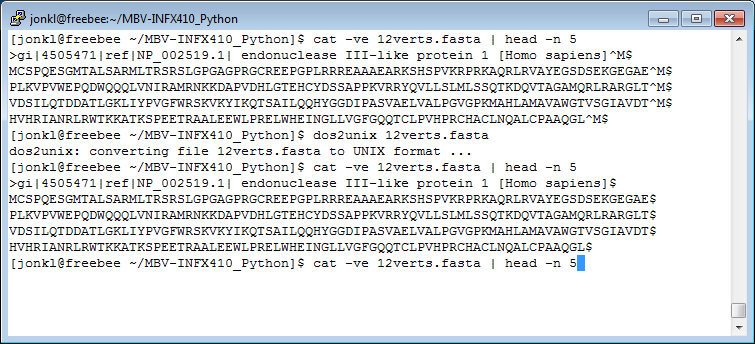
**We are searching with a protein query sequence in a protein sequence database, hence blastp. tblastn is used for searching with a protein sequence in a translated nucleotide database.**

1. From the resulting BLAST hits, select the following sequences: endonuclease III-like (Nth) (approx. 280-360 amino acids) and A/G-specific adenine glycosylase (also known as MutY) (approx. 510-720 amino acids) from man (*Homo sapiens*), mouse (*Mus musculus*), cow (*Bos taurus*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis*), and the fugu pufferfish (*Takifugu rubripes*). If there are several isoforms of the proteins, choose the one with the lowest isoform number. Also, if there are several entries for the same protein, select the one who has an accession starting with “NP\_”, or alternatively with “XP\_”. We do not have time to look very closely at all the sequences and their splice variants, but if we wanted to do serious work with these sequences, we would have to do that. Retrieve the sequences in FASTA format, and paste them into the report. Make sure you are able to do this properly, at least for human, mouse, and cow, before you continue below.
2. We want, as for the bacterial protein sequences, to shorten each sequence title to contain only the species name and RefSeq identifier. We could do this manually, in a text editor, as we did above for the bacterial homologs. However, the task here will be to use a little program or script to do this. If we had hundreds of sequences, making a script would certainly be quicker and less error prone. If we had even more sequences, a script would be the only option.

Log onto freebee.abel.uio.no, and create a new directory that you will work in. Download the zipped tarball MSA\_exercise.tar.gz and put it in the new directory. Uncompress the file and extract all the files from the tarball.



1. You find the 12 vertebrate Nth homologs in the file 12verts.fasta. Make sure the file has correct Unix format with Unix end-of-lines by using *cat -ve*. Use *man cat* to find out what the “*-ve*” is doing. If the end-of-lines are not correct, fix the problem with the command *dos2unix.*



1. Your task is to open the file 12verts.fasta and change all headers to the correct format (*e.g.* “>Homo\_sapiens\_NP\_002519” for the human Nth homolog). Change all spaces to “\_” and, of course, keep the initial “>”. Then write out a new Fasta file, identical to the original one, but with modified and simplified Fasta headers. Call the new file 12verts\_new.fasta.

>gi|4505471|ref|NP\_002519.1| endonuclease III-like protein 1 [Homo sapiens]

MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE

PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT

...

should become

>Homo\_sapiens\_NP\_002519

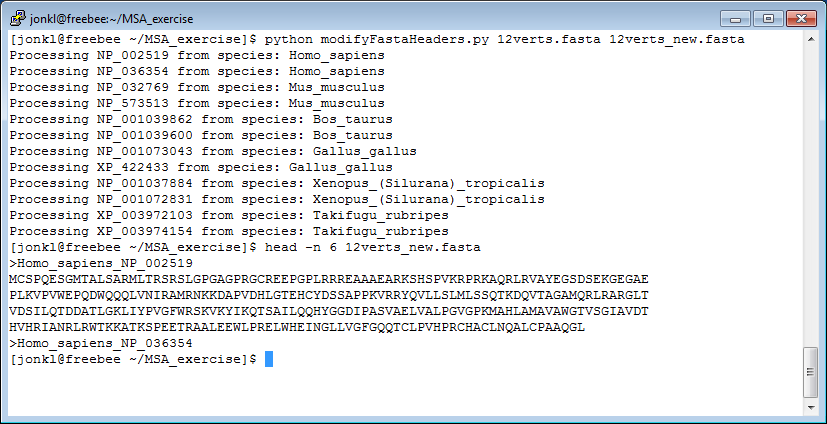
MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE

PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT

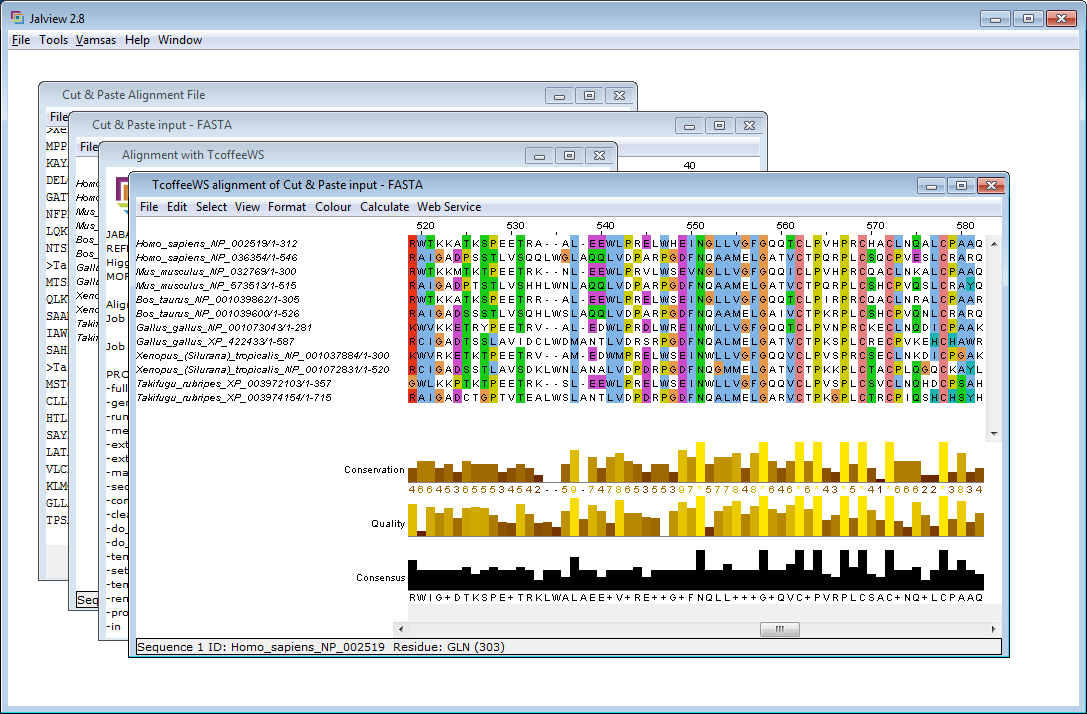
...

and so on.

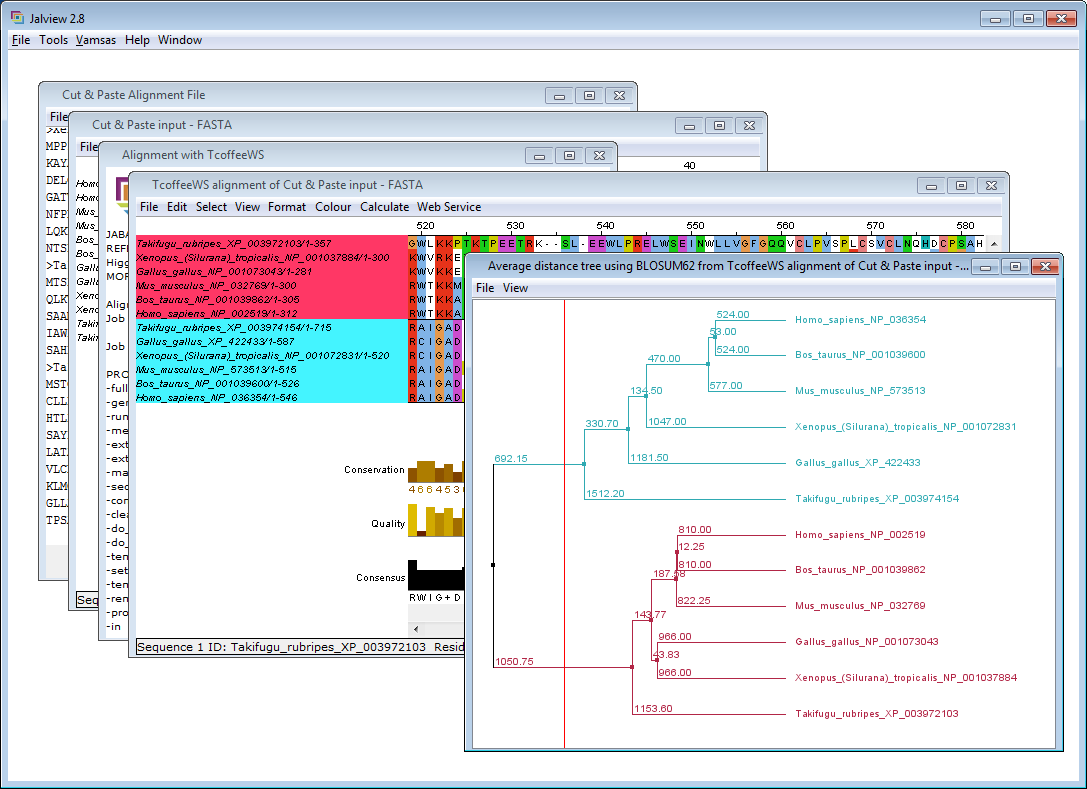
1. Write down the steps, or a little flowchart, that shows what the script has to do in order to solve the task.
2. If you have done any programming before *or* you want a challenge, choose (a) below, otherwise do (b).
   1. Make a script in a programming language of your own choice that does the file conversion described above
   2. Take a look at the python script modifyFastaHeaders.py you found in the tarball MSA\_exercise.tar.gz. Go through it step by step and make sure you understand what it will do. Use this script to do the file conversion



1. As you did for the bacterial sequences, use Jalview to generate an MSA for the twelve vertebrate sequences, but this time use the T-Coffee algorithm (with default settings).



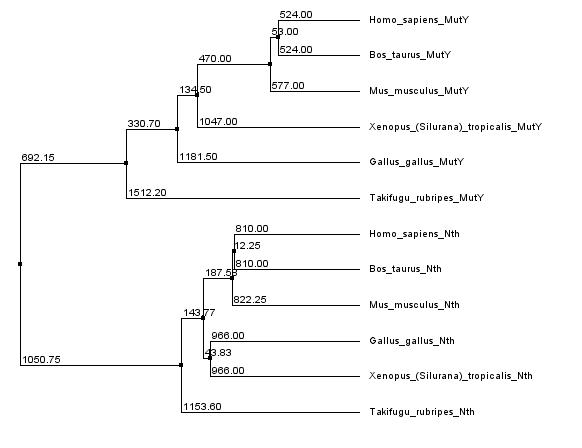
1. In Jalview, generate a phylogenetic tree from the alignment of the twelve proteins (Choose “Calculate” → “Calculate Tree” → ”Average Distance Using BLOSUM62”). Click in the tree window to get different colours on the two clades in the tree (See below). Then, in the MSA windows, do “Calculate” → “Sort” → “By Tree Order” and choose sorting according to the tree you just generated.



1. Use the terms homologs, paralogs, and orthologs to describe the relationships between these proteins/genes.

**NP\_036354 is human MutY (with official gene name and symbol “mutY homolog” and MUTYH) while NP\_002519 is human Nth (with official gene name and symbol “nth endonuclease III-like 1 (E. coli)” and NTHL1). All the other “blue” sequences/nodes in the figure above are orthologs of human MUTYH. They are unique genes/proteins due to a speciation event. Similarly, all the nodes in the “red” clade are orthologs of NTHL1. NTHL1 and MUTYH are paralogs, due to a gene duplication. All the sequences are homologs.**

1. We now change the names of the sequences a final time and put Nth in all the headers of the NTHL1 orthologs and MutY in all the headers of the MUTYH orthologs. Open the file 12verts\_new.fasta in a text editor and change the headers so that all the Nth orthologs are named by their species and Nth (*e.g.* Homo\_sapiens\_Nth), while all MutY homologs are named with MutY (*e.g.* Homo\_sapiens\_MytY). Do this manually, or with a script. Save the new Fasta file as 12verts\_final.fasta.
2. Get the sequences from 12verts\_final.fasta into Jalview and generate an MSA with the T-Coffee algorithm, as above. Also generate a phylogenetic tree and sort as above. Save the tree in PNG format, and import it into your report. Are all the clades as you would expect?

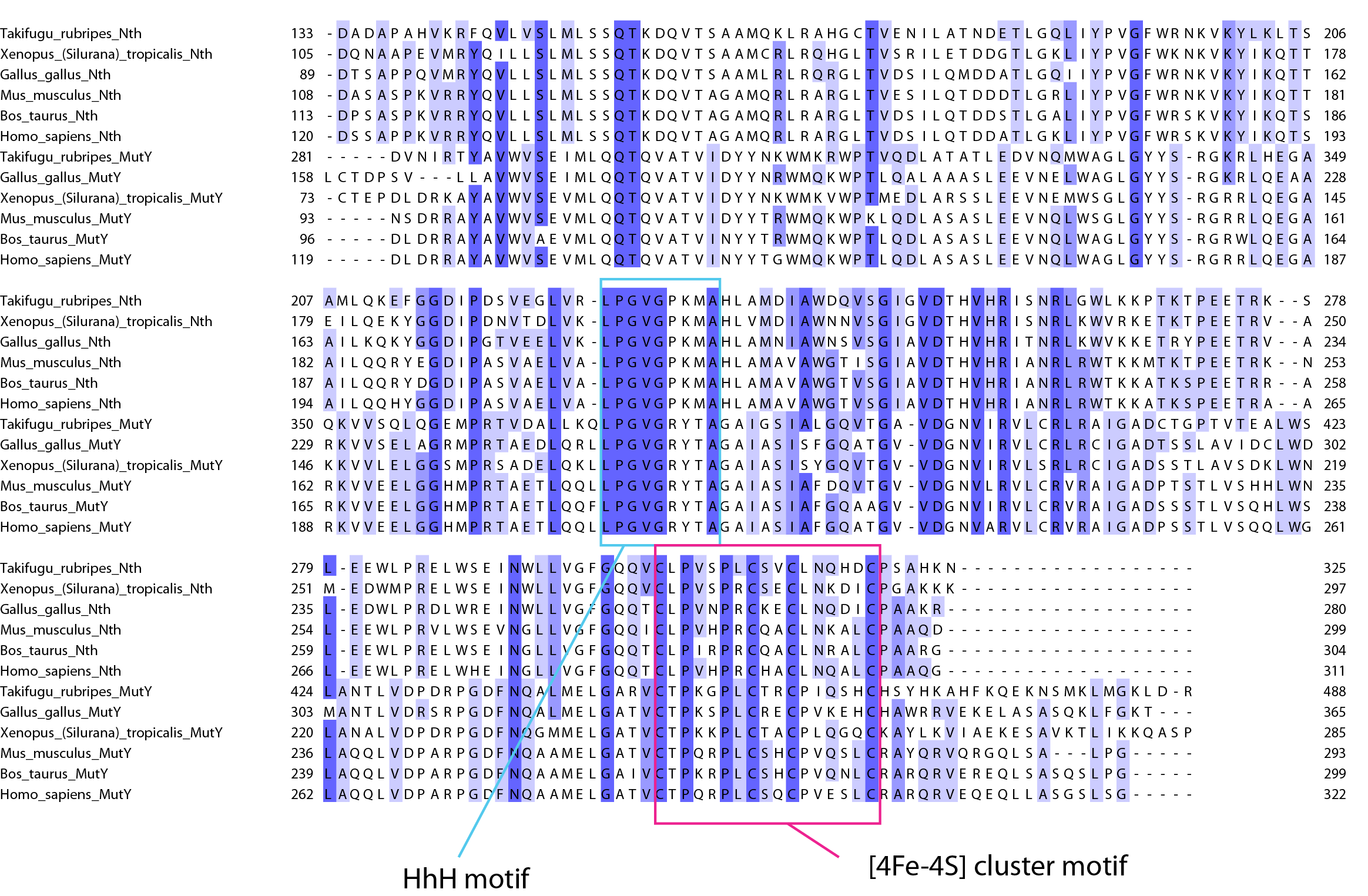


**One would expect, in each of the two main clades, that the orthologs in human, mouse and cow were most similar, as they also are. Further out one would expect to encounter first chicken, then frog, and finally the fish orthologs. The fish clades are where they should be, but the frog/chicken is swapped (MutY) or in their own clade (Nth).**

**It is certainly not possible to make reliable phylogenetic species trees from just a single gene, but another problem here is that the phylogenetic tree building algorithms in Jalview are *very* simple. They can certainly not be used in a publication, for example!**

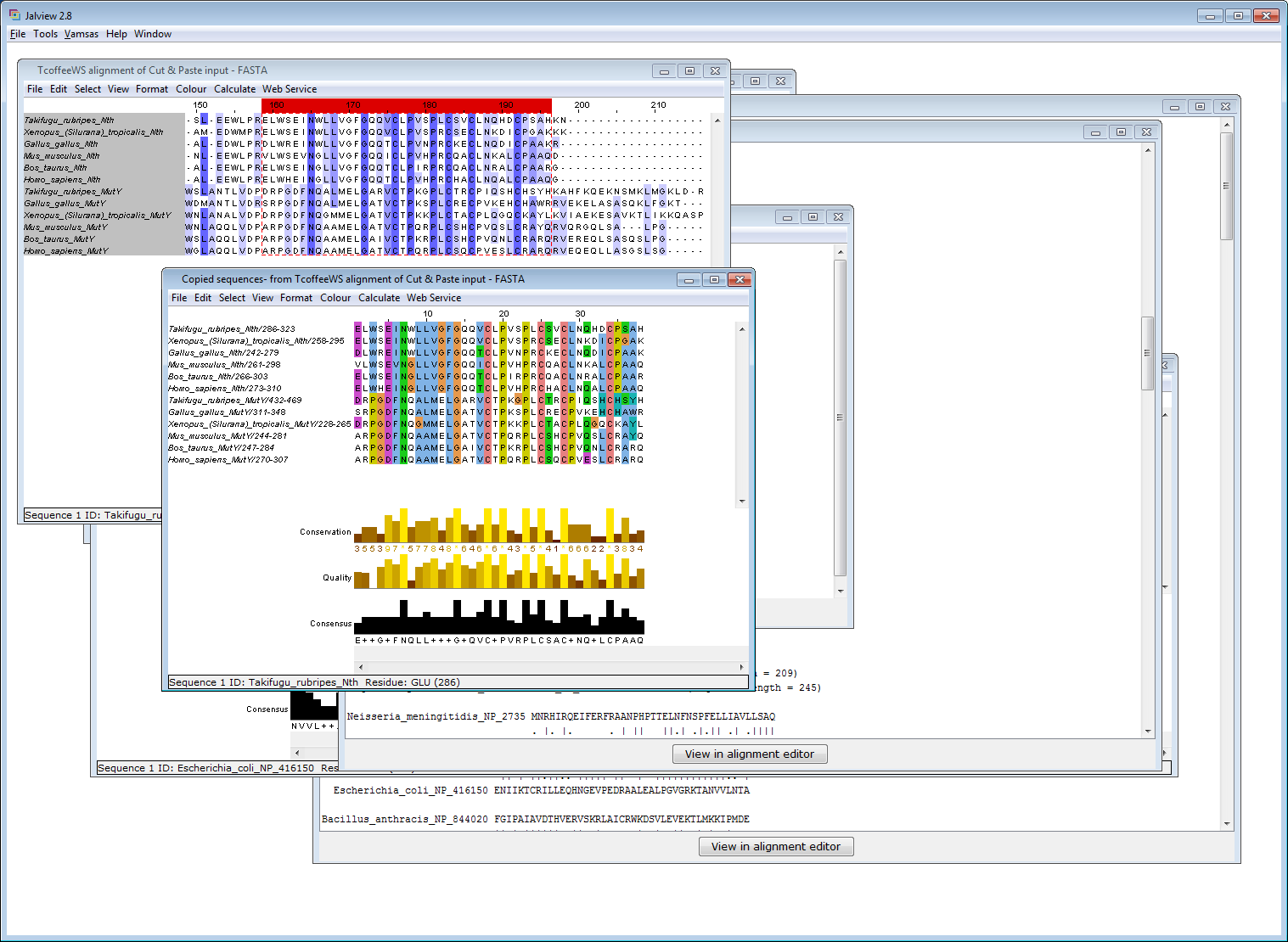
**This is a good review on how to make reliable trees: Z. Yang & B. Rannala, "Molecular phylogenetics: principles and practice", Nat. Rev. Genet. 13, 303 (2012).**

1. For the MSA, colour by percentage identity, turn off all annotations (remove the tick mark at “View” → “Show Annotations”), and use “Edit” → “Remove Left/Right” to trim the MSA and only keep the core part that is relatively conserved in all the sequences (roughly corresponding to human MutY residues 120 – 320). Turn on wrapping, and export the MSA as a PNG file. Import this alignment into PowerPoint or a similar program, and indicate the two sequence motifs. Copy the resulting figure into your report. Are both motifs fully conserved in all sequences?

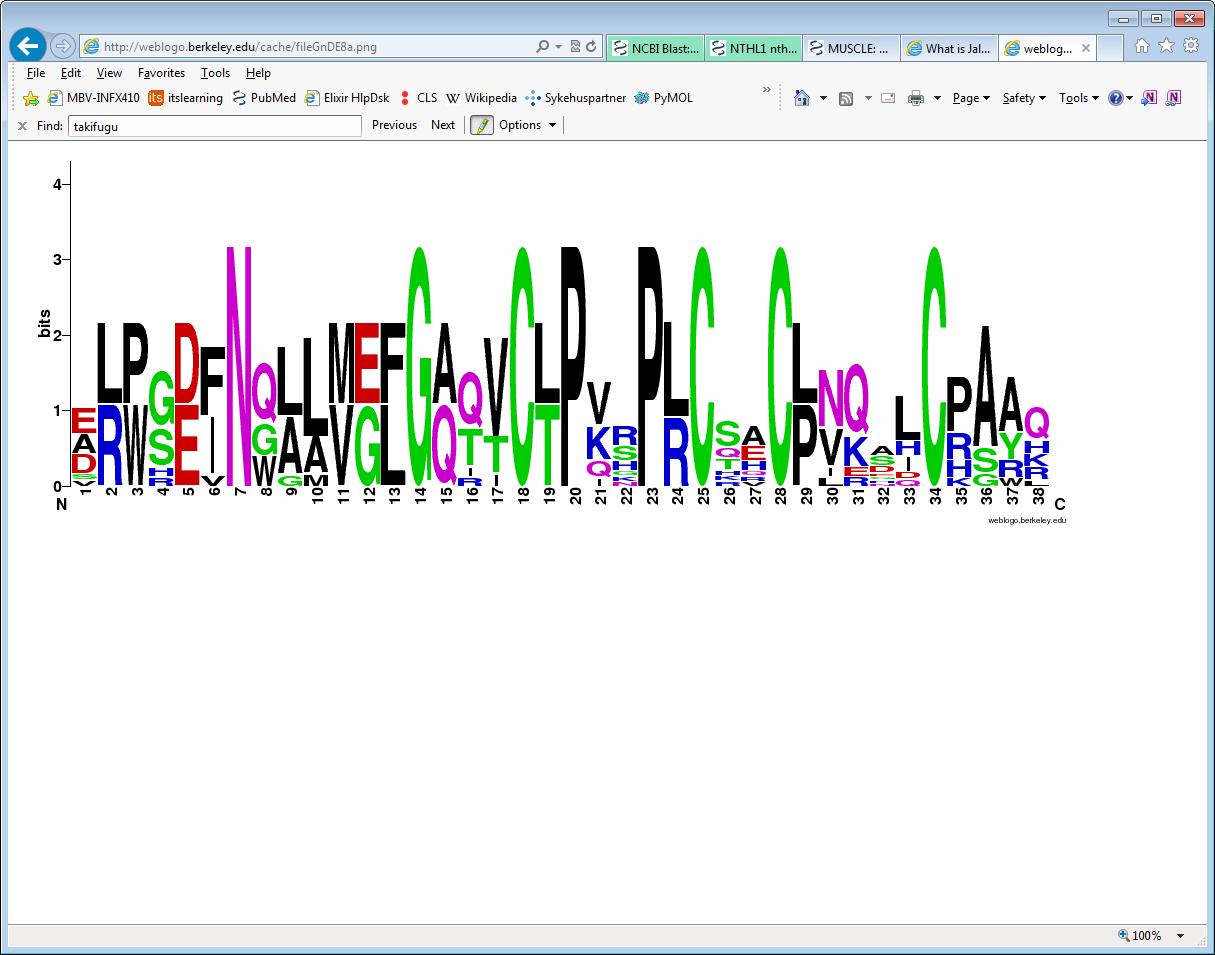


**The 4 Cys residues of the [4Fe-4S] cluster motif are 100% conserved in all homologs. The motif LPGVGxxxA is conserved in all sequences where xxx is PKM or RYT in Nth and MutY, respectively.**

1. Select a chunk of the MSA between human MutY residues 270 and 307 containing the [4Fe-4S] cluster motif. Do this by left-clicking just above the MSA, next to “160” (See below), and pull to the right while holding down the mouse button. Select the “red region” below. Copy this segment by pressing <ctrl>-c, and paste this into a new window by pressing <ctrl>-<shift>-v. See below. It should look something like this!:



1. Now let us make a sequence logo for this segment. Go to the following website, <http://weblogo.berkeley.edu>, and follow the link “create”. In Jalview, get the MSA for our [4Fe-4S] cluster motif segment in Fasta format by doing “File” → “Output to Textbox” → “FASTA”. Copy the Fasta format text into the window on the WebLogo website. Then press “Create Logo”. Put the logo in your report. Take a screen-shot, for example.



**The logo gives a good illustration of which residues are conserved in this protein family and which are not.**

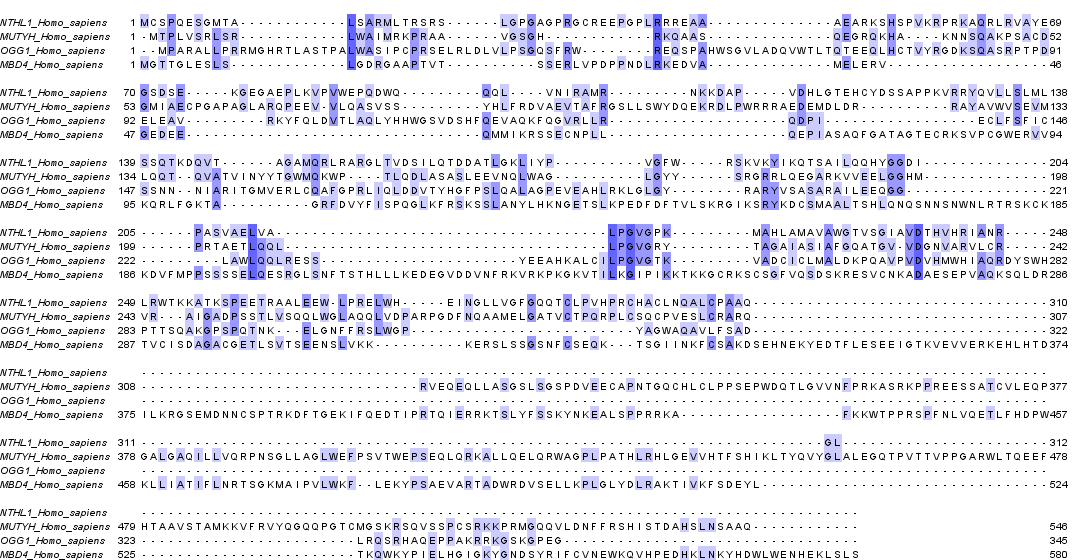
1. Using the sequence of *E. coli* Nth as query, perform an iterative protein PSI-BLAST search against the NCBI Reference protein sequence database (Refseq protein). Before doing the search, limit the search to vertebrate sequences, and change the “PSI-BLAST threshold” from the default value of 0.005 to 0.00005 (=5e-5). After convergence (or at least three iterations), reformat the results to include only human (*Homo sapiens*) sequences. From the results, select sequences corresponding to the four human homologs denoted Endonuclease III-like protein 1 (NTHL1) (312 aa), A/G-specific adenine DNA glycosylase isoform 1 (MUTYH) (546 aa), N-glycosylase/DNA lyase isoform 1a (OGG1) (345 aa) and methyl-CpG-binding domain protein 4 (MBD4) (580 aa). Give the sequences short names.

Make a multiple sequence alignment of the four sequences, using the MUSCLE program from JalView. Format the alignment as earlier. Then try the MAFFT and ClustalW programs. Import the three sequence alignments into your report.

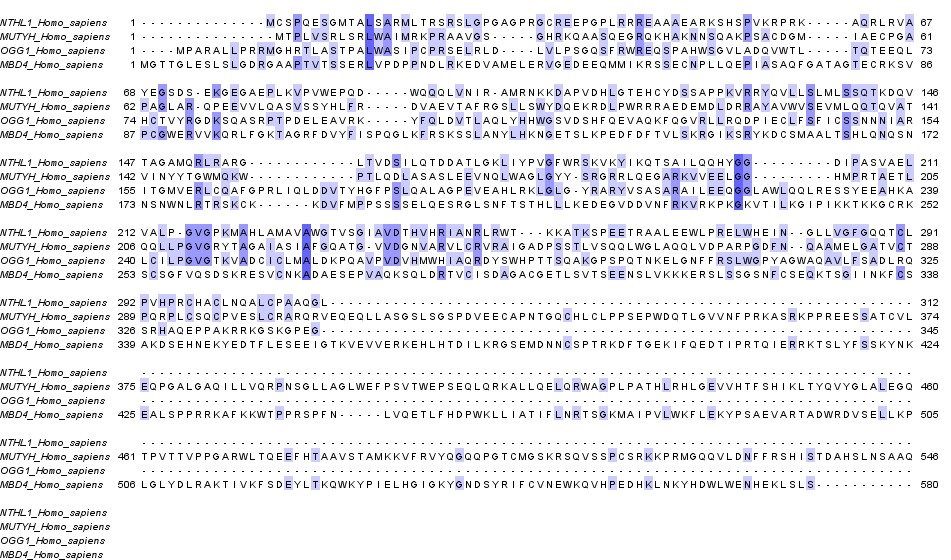
**MUSCLE:**



**MAFFT:**

****

**CLUSTAL:**

****

1. Are the HhH motif and the [4Fe–4S] cluster motif present in all four sequences? Note that the first 400 residues in the N-terminal of MBD4 are unrelated to the other proteins, and any similarity to that N-terminal part of the MBD4 protein is completely random.

**The HhH motif is well conserved in NTHL1, MUTYH and OGG1. However, ClustalW does not align the initial L and P correctly for NTHL1.**

**The [4Fe-4S] cluster motif is fully conserved in NTHL1 and MUTYH, but not nicely aligned above due to the other two sequences that are lacking the motif. Actually, there is no [4Fe-4S] cluster in OGG1 or MBD4. Hence, there is no need to conserve, during evolution, the Cys residues that are complexing the iron-sulphur cluster in the other homologs.**

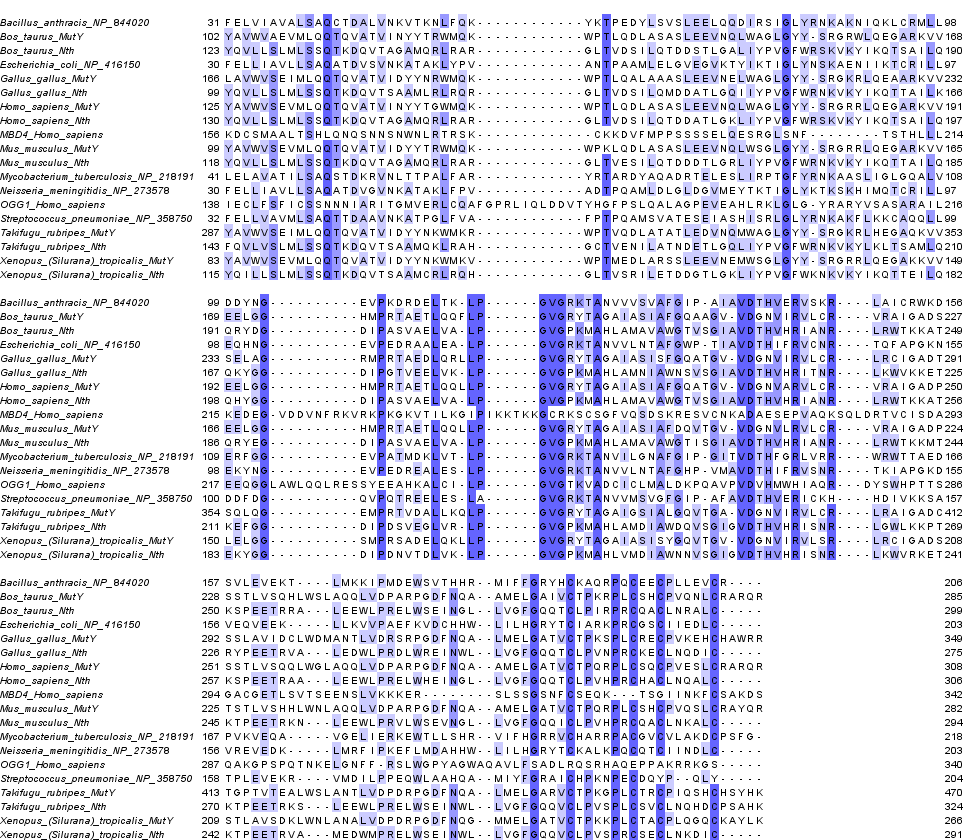
**MBD4 is also aligned to both these motifs with all three programs, but to the wrong part of MBD4.**

1. Judging from the proper alignment of residues in the two motifs, which of the programs has produced the worst alignment?

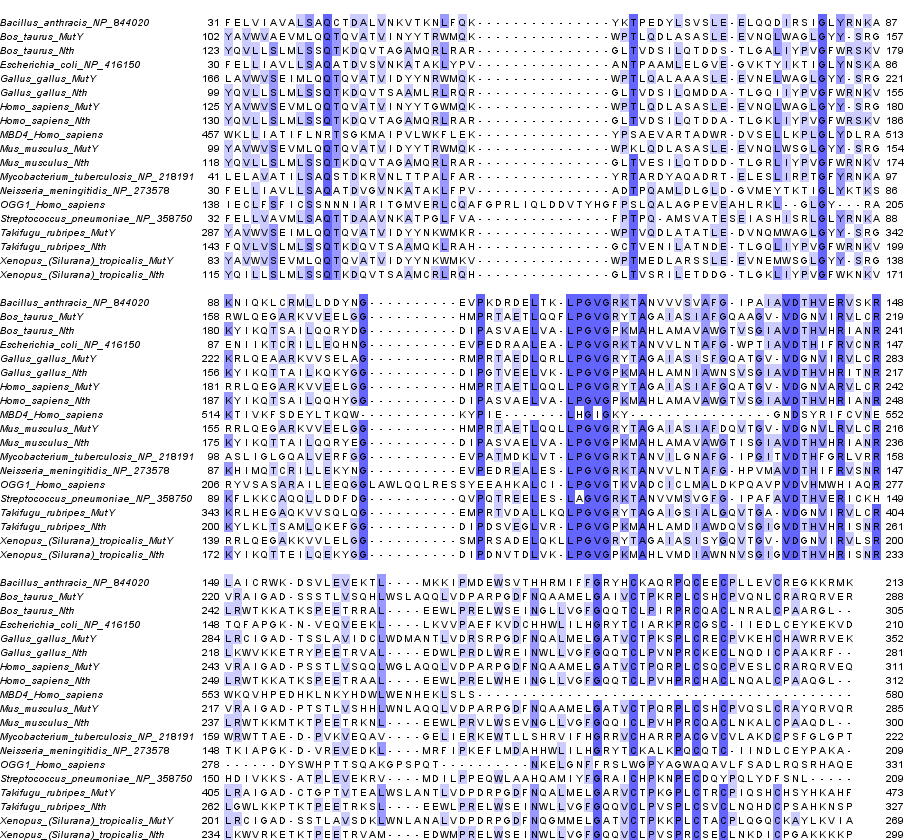
**The Clustal W program seemes to produce the worst alignment, as the HhH motif was not well aligned in NTHL1.**

1. Finally, make MUSCLE and MAFFT alignments where you also include the bacterial and vertebrate Nth and MutY sequences that we worked with earlier. Do not duplicate the human Nth and MutY. Format the alignment as earlier, but sort “by ID”. Include the alignments in your report, but crop the images so that only the region “core region” with the HhH motif is shown. Are any of the programs able to correctly align the HhH motif when all sequences are included? Which important lesson can we learn from this? Which program performed best?

**MUSCLE:**

****

**MAFFT:**

****

**MAFFT was able to correctly align the HhH motif of MBD4 with all the others when all sequences are included. MUSCLE did not. MAFFT performed best, but this is no general rule. MUSCLE and T-Coffee are also excellent MSA programs.**

***Important:* Very often, you get a better alignment of two, or a few sequences, if you align these sequences together with many homologs!**

1. If you have more time, experiment and modify the script for example to
   1. Use H\_sapiens, M\_musculus, and so on in the headers
   2. Automatically generate 12verts\_final.fasta from 12verts.fasta
   3. Or download a few hundred Nth vertebrate homologs from the BLAST results and test the script on this bigger data set. If necessary, modify the script to be more robust

**APPENDIX 1:**

**Bacterial Nth homologs, original sequences**

>gi|16129591|ref|NP\_416150.1| DNA glycosylase and apyrimidinic (AP) lyase (endonuclease III) [Escherichia coli str. K-12 substr. MG1655]

MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIAVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE

GVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH

IFRVCNRTQFAPGKNVEQVEEKLLKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVD

I

>gi|57117142|ref|NP\_218191.2| Probable endonuclease III Nth (DNA-(apurinic or apyrimidinic site)lyase) (AP lyase) (AP endonuclease class I) (endodeoxyribonuclease (apurinic or apyrimidinic)) (deoxyribonuclease (apurinic or apyrimidinic)) [Mycobacterium tuberculosis H37Rv]

MPGRWSAETRLALVRRARRMNRALAQAFPHVYCELDFTTPLELAVATILSAQSTDKRVNLTTPALFARYR

TARDYAQADRTELESLIRPTGFYRNKAASLIGLGQALVERFGGEVPATMDKLVTLPGVGRKTANVILGNA

FGIPGITVDTHFGRLVRRWRWTTAEDPVKVEQAVGELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVL

AKDCPSFGLGPTEPLLAAPLVQGPETDHLLALAGL

>gi|30261643|ref|NP\_844020.1| endonuclease III [Bacillus anthracis str. Ames]

MLNKTQIRYCLDTMADMYPEAHCELIHDNPFELVIAVALSAQCTDALVNKVTKNLFQKYKTPEDYLSVSL

EELQQDIRSIGLYRNKAKNIQKLCRMLLDDYNGEVPKDRDELTKLPGVGRKTANVVVSVAFGIPAIAVDT

HVERVSKRLAICRWKDSVLEVEKTLMKKIPMDEWSVTHHRMIFFGRYHCKAQRPQCEECPLLEVCREGKK

RMKGK

>gi|15676439|ref|NP\_273578.1| endonuclease III [Neisseria meningitidis MC58]

MNRHIRQEIFERFRAANPHPTTELNFNSPFELLIAVLLSAQATDVGVNKATAKLFPVADTPQAMLDLGLD

GVMEYTKTIGLYKTKSKHIMQTCRILLEKYNGEVPEDREALESLPGVGRKTANVVLNTAFGHPVMAVDTH

IFRVSNRTKIAPGKDVREVEDKLMRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTCIINDLCEYPAKA

>gi|15903200|ref|NP\_358750.1| endonuclease III [Streptococcus pneumoniae R6]

MVLSKKRARKVLEEIIALFPDAKPSLDFTNHFELLVAVMLSAQTTDAAVNKATPGLFVAFPTPQAMSVAT

ESEIASHISRLGLYRNKAKFLKKCAQQLLDDFDGQVPQTREELESLAGVGRKTANVVMSVGFGIPAFAVD

THVERICKHHDIVKKSATPLEVEKRVMDILPPEQWLAAHQAMIYFGRAICHPKNPECDQYPQLYDFSNL

**APPENDIX 2:**

**Bacterial Nth homologs, modified headers**

>Escherichia\_coli\_NP\_416150

MNKAKRLEILTRLRENNPHPTTELNFSSPFELLIAVLLSAQATDVSVNKATAKLYPVANTPAAMLELGVE

GVKTYIKTIGLYNSKAENIIKTCRILLEQHNGEVPEDRAALEALPGVGRKTANVVLNTAFGWPTIAVDTH

IFRVCNRTQFAPGKNVEQVEEKLLKVVPAEFKVDCHHWLILHGRYTCIARKPRCGSCIIEDLCEYKEKVD

I

>Mycobacterium\_tuberculosis\_NP\_218191

MPGRWSAETRLALVRRARRMNRALAQAFPHVYCELDFTTPLELAVATILSAQSTDKRVNLTTPALFARYR

TARDYAQADRTELESLIRPTGFYRNKAASLIGLGQALVERFGGEVPATMDKLVTLPGVGRKTANVILGNA

FGIPGITVDTHFGRLVRRWRWTTAEDPVKVEQAVGELIERKEWTLLSHRVIFHGRRVCHARRPACGVCVL

AKDCPSFGLGPTEPLLAAPLVQGPETDHLLALAGL

>Bacillus\_anthracis\_NP\_844020

MLNKTQIRYCLDTMADMYPEAHCELIHDNPFELVIAVALSAQCTDALVNKVTKNLFQKYKTPEDYLSVSL

EELQQDIRSIGLYRNKAKNIQKLCRMLLDDYNGEVPKDRDELTKLPGVGRKTANVVVSVAFGIPAIAVDT

HVERVSKRLAICRWKDSVLEVEKTLMKKIPMDEWSVTHHRMIFFGRYHCKAQRPQCEECPLLEVCREGKK

RMKGK

>Neisseria\_meningitidis\_NP\_273578

MNRHIRQEIFERFRAANPHPTTELNFNSPFELLIAVLLSAQATDVGVNKATAKLFPVADTPQAMLDLGLD

GVMEYTKTIGLYKTKSKHIMQTCRILLEKYNGEVPEDREALESLPGVGRKTANVVLNTAFGHPVMAVDTH

IFRVSNRTKIAPGKDVREVEDKLMRFIPKEFLMDAHHWLILHGRYTCKALKPQCQTCIINDLCEYPAKA

>Streptococcus\_pneumoniae\_NP\_358750

MVLSKKRARKVLEEIIALFPDAKPSLDFTNHFELLVAVMLSAQTTDAAVNKATPGLFVAFPTPQAMSVAT

ESEIASHISRLGLYRNKAKFLKKCAQQLLDDFDGQVPQTREELESLAGVGRKTANVVMSVGFGIPAFAVD

THVERICKHHDIVKKSATPLEVEKRVMDILPPEQWLAAHQAMIYFGRAICHPKNPECDQYPQLYDFSNL

**APPENDIX 3:**

**12 vertebrate homologs, original sequences**

>gi|4505471|ref|NP\_002519.1| endonuclease III-like protein 1 [Homo sapiens]

MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE

PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT

VDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDT

HVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFGQQTCLPVHPRCHACLNQALCPAAQGL

>gi|6912520|ref|NP\_036354.1| A/G-specific adenine DNA glycosylase isoform 1 [Homo sapiens]

MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPEEVVLQASVSS

YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVATVINYYTGWMQKWPTLQDLA

SASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFGQATGVVDGNVARVL

CRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSL

SGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWE

FPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHT

AAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ

>gi|227908769|ref|NP\_032769.2| endonuclease III-like protein 1 [Mus musculus]

MNSGVRMVTRSRSRATRIASEGCREELAPREAAAEGRKSHRPVRHPRRTQKTHVAYEAANGEEGEDAEPLKVPVWEPQNW

QQQLANIRIMRSKKDAPVDQLGAEHCYDASASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVESILQTDDDTL

GRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIANRLRWT

KKMTKTPEETRKNLEEWLPRVLWSEVNGLLVGFGQQICLPVHPRCQACLNKALCPAAQDL

>gi|227330621|ref|NP\_573513.2| A/G-specific adenine DNA glycosylase [Mus musculus] MKKLQASVRSHKKQPANHKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSNLLSWYDQEKR

DLPWRNLAKEEANSDRRAYAVWVSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEEVNQLWSGLGYYSRGRRLQEG

ARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGVVDGNVLRVLCRVRAIGADPTSTLVSHHLWNLAQQL

VDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAYQRVQRGQLSALPGRPDIEECALNTRQCQLCLTSSSPWDPS

MGVANFPRKASRRPPREEYSATCVVEQPGAIGGPLVLLVQRPDSGLLAGLWEFPSVTLEPSEQHQHKALLQELQRWCGPL

PAIRLQHLGEVIHIFSHIKLTYQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRQGTRKGSKRSQ

VCPPSSRKKPSLGQQVLDTFFQRHIPTDKPNSTTQ

>gi|114051958|ref|NP\_001039862.1| endonuclease III-like protein 1 [Bos taurus]

MNAAGVRMVVTRARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSPVKRRRKAQRLSVAYEASEGEGGEGAEHLQAPSW

QPQDWRQQLDNIRTMRSGKDAPVDQLGAEHCFDPSASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVDSILQT

DDSTLGALIYPVGFWRSKVKYIKQTSAILQQRYDGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIAN

RLRWTKKATKSPEETRRALEEWLPRELWSEINGLLVGFGQQTCLPIRPRCQACLNRALCPAARGL

>gi|281485563|ref|NP\_001039600.2| A/G-specific adenine DNA glycosylase [Bos taurus]

MKKSRAAVGNRSGRRKQASSQEGKEKCAFGSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTALQESLLDWYDR

KKRDLPWRRLVEDEVDLDRRAYAVWVAEVMLQQTQVATVINYYTRWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRWL

QEGARKVVEELGGHMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGVVDGNVIRVLCRVRAIGADSSSTLVSQHLWSLA

QQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLCRARQRVEREQLSASQSLPGNCDVEECAPNTGQCPLCAPPT

EPWDQTLGVTNFPRKASRKPPREECSAICVLEQPKALGGAHILLVQRPNSGLLAGLWEFPSVSVNAEASGQHQRAALLQE

LQSWVGPLPDTRLQHLGQVVHTFSHIKMTYQVYSLALEEHTPVTIVPPGARWLTREDFHTAAVSTAMKKVFRMYEGQQPG

TCKGSKRSQVATLSKRKKPSPGQQVLESFFWPHVPTDAPSLNTAAQ

>gi|118601744|ref|NP\_001073043.1| endonuclease III-like protein 1 [Gallus gallus]

MCAAAPRGGGRAARRLGAATAGSRVPSAAPRYSRRTRRVPIAYEAEPKPESPGPKWEPENWQQQLERIREMRRHRDAPVD

EMGVDKCYDTSAPPQVMRYQVLLSLMLSSQTKDQVTSAAMLRLRQRGLTVDSILQMDDATLGQIIYPVGFWRNKVKYIKQ

TTAILKQKYGGDIPGTVEELVKLPGVGPKMAHLAMNIAWNSVSGIAVDTHVHRITNRLKWVKKETRYPEETRVALEDWLP

RDLWREINWLLVGFGQQTCLPVNPRCKECLNQDICPAAKRF

>gi|513197809|ref|XP\_422433.3| PREDICTED: A/G-specific adenine DNA glycosylase isoform X5 [Gallus gallus]

MGGAAVRARRSVKVRAGGEHVGPGLGSPAIALRTHRRCCDPTPVPVSRQGLPLDHMHCISSVTPSRSMPYVAACSPGMTK

AGGTFPGGRWLQLSWMLTGGRMQLLGLLVASERQVWARERSCGDEDGEEGCWVGFCCSSWNQQHGDRGACCEKWHWHLCT

DPSVLLAVWVSEIMLQQTQVATVIDYYNRWMQKWPTLQALAAASLEEVNELWAGLGYYSRGKRLQEAARKVVSELAGRMP

RTAEDLQRLLPGVGRYTAGAIASISFGQATGVVDGNVIRVLCRLRCIGADTSSLAVIDCLWDMANTLVDRSRPGDFNQAL

MELGATVCTPKSPLCRECPVKEHCHAWRRVEKELASASQKLFGKTTLVPDVEDCGPGGCPLCLPAAEPWDSSLGVTNFPR

KAAKKQPRVEWTATCVLERRGRLGAPEYLIVQRPSSGLLAGLWEFPSLPLAPGLQEEQQKEVLADHLRAWTRQPVQTQSL

CFIGEVVHIFSHIHQTYVVYSLCLDGDVALDAASSPSRWVTEEEFRASAVSTAMKKVLKARETQRGVQSGRAKGSKRKRE

SKLGAAGSTPTGMQLSLRAFLRAQPPP

>gi|113205550|ref|NP\_001037884.1| nth endonuclease III-like 1 [Xenopus (Silurana) tropicalis]

MSGSLRPLGRRGRRGVLKAVGGKDQQDGTSKGQVIDDSEDEKPSSPKERSKRRVSVEYEQAASETVAKRPKWQPKNWAQH

LENIRQMRSRRDAPVDQMGAEKCYDQNAAPEVMRYQILLSLMLSSQTKDQVTSAAMCRLRQHGLTVSRILETDDGTLGKL

IYPVGFWKNKVKYIKQTTEILQEKYGGDIPDNVTDLVKLPGVGPKMAHLVMDIAWNNVSGIGVDTHVHRISNRLKWVRKE

TKTPEETRVAMEDWMPRELWSEINWLLVGFGQQVCLPVSPRCSECLNKDICPGAKKKKPR

>gi|118403607|ref|NP\_001072831.1| mutY homolog [Xenopus (Silurana) tropicalis]

MPPPRTKTSLGRSAAASGKRKSPKQAFPKREEHVLQSSIYHSFTSQETEIIRDKLLAWYDKSKRDLPWRTMACTEPDLDR

KAYAVWVSEVMLQQTQVATVIDYYNKWMKVWPTMEDLARSSLEEVNEMWSGLGYYSRGRRLQEGAKKVVLELGGSMPRSA

DELQKLLPGVGRYTAGAIASISYGQVTGVVDGNVIRVLSRLRCIGADSSTLAVSDKLWNLANALVDPDRPGDFNQGMMEL

GATVCTPKKPLCTACPLQGQCKAYLKVIAEKESAVKTLIKKQASPIAKDVGDIEDCDLGPGLCALCVPTSDPWDSSLGVA

NFPRKSAKKPSRMEQTAICVWEKCGDHGELEYLIVQRPSSGLLAGLWEFPSILLDEKFTEQNRQHSLLGLLQDLSGHAVP

LQKLQYKGEVVHIFSHIHQTYVVYFLSLNTTENCSVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAV

NTSKKRKGDLAKVQLPSGRIKTEKGKQQSIQSFFKLATEK

>gi|410917257|ref|XP\_003972103.1| PREDICTED: endonuclease III-like protein 1-like [Takifugu rubripes]

MTSHYFAQSRSVVTRRGAQNAAHKPATSLKSKLTIQPEKDDLVSSSAGVKLEEEEAKISGNALKPETDAPTLSSHSRRRR

QLKVEYDKDGSMPQLKTEPWEPPRWKTQLENIRAMRSGRDAPVDNMGADKCHDADAPAHVKRFQVLVSLMLSSQTKDQVT

SAAMQKLRAHGCTVENILATNDETLGQLIYPVGFWRNKVKYLKLTSAMLQKEFGGDIPDSVEGLVRLPGVGPKMAHLAMD

IAWDQVSGIGVDTHVHRISNRLGWLKKPTKTPEETRKSLEEWLPRELWSEINWLLVGFGQQVCLPVSPLCSVCLNQHDCP

SAHKNSPVRRPKFLERSPWIKSPRLFITPGANTFLIR

>gi|410921366|ref|XP\_003974154.1| PREDICTED: A/G-specific adenine DNA glycosylase-like [Takifugu rubripes]

MSTQGEPVQVGKVLSFLREWDRGDRSARGRMLSSFLGRSAGRTQGELEYLGEFVGHGGVVTLLEVLTQPQSNEETKAEAL

CLLLAISDAGRKYKELICQSCGAMAAAECLTHSGTGETQESAWMLLESLSHGNPKYEGEIYKGLIGHLTCTSAKAQQFVL

HTLHTLQSKMEIAHHSIVEPLLGVLTSLHPDVQSEVARLIFELRRYDVRPMLLRALCGLGLNVARAPTYPEEESSHASSP

SAYHFFHDAADVALLRSRLLAWYDQEKRELPWRTLALTEPDVNIRTYAVWVSEIMLQQTQVATVIDYYNKWMKRWPTVQD

LATATLEDVNQMWAGLGYYSRGKRLHEGAQKVVSQLQGEMPRTVDALLKQLPGVGRYTAGAIGSIALGQVTGAVDGNVIR

VLCRLRAIGADCTGPTVTEALWSLANTLVDPDRPGDFNQALMELGARVCTPKGPLCTRCPIQSHCHSYHKAHFKQEKNSM

KLMGKLDRKSSALPDIEDCLSSGTCTLCLSEPWDDELGVQNFPRKPAKKPPRAERCLTCVVIRQGEGGEHEFLLTQRPSK

GLLAGLWEFPCINHEEKNAVEEKKVLCAEINRILGTSLTHGLLQYVGEVVHIFSHIHQTYVVHTLRLKDAVSQSENMQWL

TPSALQEAAVSTGVKKIMKLCNSALGQQGAPDGEEKRPKKDRKGQITKRPRLSGANSRSRQLSLSSFFQTVKQDC

**APPENDIX 4:**

**12 vertebrate homologs, modified headers**

>Homo\_sapiens\_NP\_002519

MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE

PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT

VDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDT

HVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFGQQTCLPVHPRCHACLNQALCPAAQGL

>Homo\_sapiens\_NP\_036354

MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPEEVVLQASVSS

YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVATVINYYTGWMQKWPTLQDLA

SASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFGQATGVVDGNVARVL

CRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSL

SGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWE

FPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHT

AAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ

>Mus\_musculus\_NP\_032769

MNSGVRMVTRSRSRATRIASEGCREELAPREAAAEGRKSHRPVRHPRRTQKTHVAYEAANGEEGEDAEPLKVPVWEPQNW

QQQLANIRIMRSKKDAPVDQLGAEHCYDASASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVESILQTDDDTL

GRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIANRLRWT

KKMTKTPEETRKNLEEWLPRVLWSEVNGLLVGFGQQICLPVHPRCQACLNKALCPAAQDL

>Mus\_musculus\_NP\_573513

MKKLQASVRSHKKQPANHKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSNLLSWYDQEKR

DLPWRNLAKEEANSDRRAYAVWVSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEEVNQLWSGLGYYSRGRRLQEG

ARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGVVDGNVLRVLCRVRAIGADPTSTLVSHHLWNLAQQL

VDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAYQRVQRGQLSALPGRPDIEECALNTRQCQLCLTSSSPWDPS

MGVANFPRKASRRPPREEYSATCVVEQPGAIGGPLVLLVQRPDSGLLAGLWEFPSVTLEPSEQHQHKALLQELQRWCGPL

PAIRLQHLGEVIHIFSHIKLTYQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRQGTRKGSKRSQ

VCPPSSRKKPSLGQQVLDTFFQRHIPTDKPNSTTQ

>Bos\_taurus\_NP\_001039862

MNAAGVRMVVTRARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSPVKRRRKAQRLSVAYEASEGEGGEGAEHLQAPSW

QPQDWRQQLDNIRTMRSGKDAPVDQLGAEHCFDPSASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVDSILQT

DDSTLGALIYPVGFWRSKVKYIKQTSAILQQRYDGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIAN

RLRWTKKATKSPEETRRALEEWLPRELWSEINGLLVGFGQQTCLPIRPRCQACLNRALCPAARGL

>Bos\_taurus\_NP\_001039600

MKKSRAAVGNRSGRRKQASSQEGKEKCAFGSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTALQESLLDWYDR

KKRDLPWRRLVEDEVDLDRRAYAVWVAEVMLQQTQVATVINYYTRWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRWL

QEGARKVVEELGGHMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGVVDGNVIRVLCRVRAIGADSSSTLVSQHLWSLA

QQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLCRARQRVEREQLSASQSLPGNCDVEECAPNTGQCPLCAPPT

EPWDQTLGVTNFPRKASRKPPREECSAICVLEQPKALGGAHILLVQRPNSGLLAGLWEFPSVSVNAEASGQHQRAALLQE

LQSWVGPLPDTRLQHLGQVVHTFSHIKMTYQVYSLALEEHTPVTIVPPGARWLTREDFHTAAVSTAMKKVFRMYEGQQPG

TCKGSKRSQVATLSKRKKPSPGQQVLESFFWPHVPTDAPSLNTAAQ

>Gallus\_gallus\_NP\_001073043

MCAAAPRGGGRAARRLGAATAGSRVPSAAPRYSRRTRRVPIAYEAEPKPESPGPKWEPENWQQQLERIREMRRHRDAPVD

EMGVDKCYDTSAPPQVMRYQVLLSLMLSSQTKDQVTSAAMLRLRQRGLTVDSILQMDDATLGQIIYPVGFWRNKVKYIKQ

TTAILKQKYGGDIPGTVEELVKLPGVGPKMAHLAMNIAWNSVSGIAVDTHVHRITNRLKWVKKETRYPEETRVALEDWLP

RDLWREINWLLVGFGQQTCLPVNPRCKECLNQDICPAAKRF

>Gallus\_gallus\_XP\_422433

MGGAAVRARRSVKVRAGGEHVGPGLGSPAIALRTHRRCCDPTPVPVSRQGLPLDHMHCISSVTPSRSMPYVAACSPGMTK

AGGTFPGGRWLQLSWMLTGGRMQLLGLLVASERQVWARERSCGDEDGEEGCWVGFCCSSWNQQHGDRGACCEKWHWHLCT

DPSVLLAVWVSEIMLQQTQVATVIDYYNRWMQKWPTLQALAAASLEEVNELWAGLGYYSRGKRLQEAARKVVSELAGRMP

RTAEDLQRLLPGVGRYTAGAIASISFGQATGVVDGNVIRVLCRLRCIGADTSSLAVIDCLWDMANTLVDRSRPGDFNQAL

MELGATVCTPKSPLCRECPVKEHCHAWRRVEKELASASQKLFGKTTLVPDVEDCGPGGCPLCLPAAEPWDSSLGVTNFPR

KAAKKQPRVEWTATCVLERRGRLGAPEYLIVQRPSSGLLAGLWEFPSLPLAPGLQEEQQKEVLADHLRAWTRQPVQTQSL

CFIGEVVHIFSHIHQTYVVYSLCLDGDVALDAASSPSRWVTEEEFRASAVSTAMKKVLKARETQRGVQSGRAKGSKRKRE

SKLGAAGSTPTGMQLSLRAFLRAQPPP

>Xenopus\_(Silurana)\_tropicalis\_NP\_001037884

MSGSLRPLGRRGRRGVLKAVGGKDQQDGTSKGQVIDDSEDEKPSSPKERSKRRVSVEYEQAASETVAKRPKWQPKNWAQH

LENIRQMRSRRDAPVDQMGAEKCYDQNAAPEVMRYQILLSLMLSSQTKDQVTSAAMCRLRQHGLTVSRILETDDGTLGKL

IYPVGFWKNKVKYIKQTTEILQEKYGGDIPDNVTDLVKLPGVGPKMAHLVMDIAWNNVSGIGVDTHVHRISNRLKWVRKE

TKTPEETRVAMEDWMPRELWSEINWLLVGFGQQVCLPVSPRCSECLNKDICPGAKKKKPR

>Xenopus\_(Silurana)\_tropicalis\_NP\_001072831

MPPPRTKTSLGRSAAASGKRKSPKQAFPKREEHVLQSSIYHSFTSQETEIIRDKLLAWYDKSKRDLPWRTMACTEPDLDR

KAYAVWVSEVMLQQTQVATVIDYYNKWMKVWPTMEDLARSSLEEVNEMWSGLGYYSRGRRLQEGAKKVVLELGGSMPRSA

DELQKLLPGVGRYTAGAIASISYGQVTGVVDGNVIRVLSRLRCIGADSSTLAVSDKLWNLANALVDPDRPGDFNQGMMEL

GATVCTPKKPLCTACPLQGQCKAYLKVIAEKESAVKTLIKKQASPIAKDVGDIEDCDLGPGLCALCVPTSDPWDSSLGVA

NFPRKSAKKPSRMEQTAICVWEKCGDHGELEYLIVQRPSSGLLAGLWEFPSILLDEKFTEQNRQHSLLGLLQDLSGHAVP

LQKLQYKGEVVHIFSHIHQTYVVYFLSLNTTENCSVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAV

NTSKKRKGDLAKVQLPSGRIKTEKGKQQSIQSFFKLATEK

>Takifugu\_rubripes\_XP\_003972103

MTSHYFAQSRSVVTRRGAQNAAHKPATSLKSKLTIQPEKDDLVSSSAGVKLEEEEAKISGNALKPETDAPTLSSHSRRRR

QLKVEYDKDGSMPQLKTEPWEPPRWKTQLENIRAMRSGRDAPVDNMGADKCHDADAPAHVKRFQVLVSLMLSSQTKDQVT

SAAMQKLRAHGCTVENILATNDETLGQLIYPVGFWRNKVKYLKLTSAMLQKEFGGDIPDSVEGLVRLPGVGPKMAHLAMD

IAWDQVSGIGVDTHVHRISNRLGWLKKPTKTPEETRKSLEEWLPRELWSEINWLLVGFGQQVCLPVSPLCSVCLNQHDCP

SAHKNSPVRRPKFLERSPWIKSPRLFITPGANTFLIR

>Takifugu\_rubripes\_XP\_003974154

MSTQGEPVQVGKVLSFLREWDRGDRSARGRMLSSFLGRSAGRTQGELEYLGEFVGHGGVVTLLEVLTQPQSNEETKAEAL

CLLLAISDAGRKYKELICQSCGAMAAAECLTHSGTGETQESAWMLLESLSHGNPKYEGEIYKGLIGHLTCTSAKAQQFVL

HTLHTLQSKMEIAHHSIVEPLLGVLTSLHPDVQSEVARLIFELRRYDVRPMLLRALCGLGLNVARAPTYPEEESSHASSP

SAYHFFHDAADVALLRSRLLAWYDQEKRELPWRTLALTEPDVNIRTYAVWVSEIMLQQTQVATVIDYYNKWMKRWPTVQD

LATATLEDVNQMWAGLGYYSRGKRLHEGAQKVVSQLQGEMPRTVDALLKQLPGVGRYTAGAIGSIALGQVTGAVDGNVIR

VLCRLRAIGADCTGPTVTEALWSLANTLVDPDRPGDFNQALMELGARVCTPKGPLCTRCPIQSHCHSYHKAHFKQEKNSM

KLMGKLDRKSSALPDIEDCLSSGTCTLCLSEPWDDELGVQNFPRKPAKKPPRAERCLTCVVIRQGEGGEHEFLLTQRPSK

GLLAGLWEFPCINHEEKNAVEEKKVLCAEINRILGTSLTHGLLQYVGEVVHIFSHIHQTYVVHTLRLKDAVSQSENMQWL

TPSALQEAAVSTGVKKIMKLCNSALGQQGAPDGEEKRPKKDRKGQITKRPRLSGANSRSRQLSLSSFFQTVKQDC

**APPENDIX 5:**

**12 vertebrate homologs, final headers**

>Homo\_sapiens\_Nth

MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEGSDSEKGEGAE

PLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLT

VDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDT

HVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGFGQQTCLPVHPRCHACLNQALCPAAQGL

>Homo\_sapiens\_MutY

MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPEEVVLQASVSS

YHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVATVINYYTGWMQKWPTLQDLA

SASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFGQATGVVDGNVARVL

CRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAMELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSL

SGSPDVEECAPNTGQCHLCLPPSEPWDQTLGVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWE

FPSVTWEPSEQLQRKALLQELQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHT

AAVSTAMKKVFRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ

>Mus\_musculus\_Nth

MNSGVRMVTRSRSRATRIASEGCREELAPREAAAEGRKSHRPVRHPRRTQKTHVAYEAANGEEGEDAEPLKVPVWEPQNW

QQQLANIRIMRSKKDAPVDQLGAEHCYDASASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVESILQTDDDTL

GRLIYPVGFWRNKVKYIKQTTAILQQRYEGDIPASVAELVALPGVGPKMAHLAMAVAWGTISGIAVDTHVHRIANRLRWT

KKMTKTPEETRKNLEEWLPRVLWSEVNGLLVGFGQQICLPVHPRCQACLNKALCPAAQDL

>Mus\_musculus\_MutY

MKKLQASVRSHKKQPANHKRRRTRALSSSQAKPSSLDGLAKQKREELLQASVSPYHLFSDVADVTAFRSNLLSWYDQEKR

DLPWRNLAKEEANSDRRAYAVWVSEVMLQQTQVATVIDYYTRWMQKWPKLQDLASASLEEVNQLWSGLGYYSRGRRLQEG

ARKVVEELGGHMPRTAETLQQLLPGVGRYTAGAIASIAFDQVTGVVDGNVLRVLCRVRAIGADPTSTLVSHHLWNLAQQL

VDPARPGDFNQAAMELGATVCTPQRPLCSHCPVQSLCRAYQRVQRGQLSALPGRPDIEECALNTRQCQLCLTSSSPWDPS

MGVANFPRKASRRPPREEYSATCVVEQPGAIGGPLVLLVQRPDSGLLAGLWEFPSVTLEPSEQHQHKALLQELQRWCGPL

PAIRLQHLGEVIHIFSHIKLTYQVYSLALDQAPASTAPPGARWLTWEEFCNAAVSTAMKKVFRMYEDHRQGTRKGSKRSQ

VCPPSSRKKPSLGQQVLDTFFQRHIPTDKPNSTTQ

>Bos\_taurus\_Nth

MNAAGVRMVVTRARSRGTGASLRRRGEKAAPLRSGEAAAEERKSYSPVKRRRKAQRLSVAYEASEGEGGEGAEHLQAPSW

QPQDWRQQLDNIRTMRSGKDAPVDQLGAEHCFDPSASPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVDSILQT

DDSTLGALIYPVGFWRSKVKYIKQTSAILQQRYDGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIAN

RLRWTKKATKSPEETRRALEEWLPRELWSEINGLLVGFGQQTCLPIRPRCQACLNRALCPAARGL

>Bos\_taurus\_MutY

MKKSRAAVGNRSGRRKQASSQEGKEKCAFGSSQAKPSAPSAGPARQQKALLQASVSPYHLFRDVAEVTALQESLLDWYDR

KKRDLPWRRLVEDEVDLDRRAYAVWVAEVMLQQTQVATVINYYTRWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRWL

QEGARKVVEELGGHMPRTAETLQQFLPGVGRYTAGAIASIAFGQAAGVVDGNVIRVLCRVRAIGADSSSTLVSQHLWSLA

QQLVDPARPGDFNQAAMELGAIVCTPKRPLCSHCPVQNLCRARQRVEREQLSASQSLPGNCDVEECAPNTGQCPLCAPPT

EPWDQTLGVTNFPRKASRKPPREECSAICVLEQPKALGGAHILLVQRPNSGLLAGLWEFPSVSVNAEASGQHQRAALLQE

LQSWVGPLPDTRLQHLGQVVHTFSHIKMTYQVYSLALEEHTPVTIVPPGARWLTREDFHTAAVSTAMKKVFRMYEGQQPG

TCKGSKRSQVATLSKRKKPSPGQQVLESFFWPHVPTDAPSLNTAAQ

>Gallus\_gallus\_Nth

MCAAAPRGGGRAARRLGAATAGSRVPSAAPRYSRRTRRVPIAYEAEPKPESPGPKWEPENWQQQLERIREMRRHRDAPVD

EMGVDKCYDTSAPPQVMRYQVLLSLMLSSQTKDQVTSAAMLRLRQRGLTVDSILQMDDATLGQIIYPVGFWRNKVKYIKQ

TTAILKQKYGGDIPGTVEELVKLPGVGPKMAHLAMNIAWNSVSGIAVDTHVHRITNRLKWVKKETRYPEETRVALEDWLP

RDLWREINWLLVGFGQQTCLPVNPRCKECLNQDICPAAKRF

>Gallus\_gallus\_MutY

MGGAAVRARRSVKVRAGGEHVGPGLGSPAIALRTHRRCCDPTPVPVSRQGLPLDHMHCISSVTPSRSMPYVAACSPGMTK

AGGTFPGGRWLQLSWMLTGGRMQLLGLLVASERQVWARERSCGDEDGEEGCWVGFCCSSWNQQHGDRGACCEKWHWHLCT

DPSVLLAVWVSEIMLQQTQVATVIDYYNRWMQKWPTLQALAAASLEEVNELWAGLGYYSRGKRLQEAARKVVSELAGRMP

RTAEDLQRLLPGVGRYTAGAIASISFGQATGVVDGNVIRVLCRLRCIGADTSSLAVIDCLWDMANTLVDRSRPGDFNQAL

MELGATVCTPKSPLCRECPVKEHCHAWRRVEKELASASQKLFGKTTLVPDVEDCGPGGCPLCLPAAEPWDSSLGVTNFPR

KAAKKQPRVEWTATCVLERRGRLGAPEYLIVQRPSSGLLAGLWEFPSLPLAPGLQEEQQKEVLADHLRAWTRQPVQTQSL

CFIGEVVHIFSHIHQTYVVYSLCLDGDVALDAASSPSRWVTEEEFRASAVSTAMKKVLKARETQRGVQSGRAKGSKRKRE

SKLGAAGSTPTGMQLSLRAFLRAQPPP

>Xenopus\_(Silurana)\_tropicalis\_Nth

MSGSLRPLGRRGRRGVLKAVGGKDQQDGTSKGQVIDDSEDEKPSSPKERSKRRVSVEYEQAASETVAKRPKWQPKNWAQH

LENIRQMRSRRDAPVDQMGAEKCYDQNAAPEVMRYQILLSLMLSSQTKDQVTSAAMCRLRQHGLTVSRILETDDGTLGKL

IYPVGFWKNKVKYIKQTTEILQEKYGGDIPDNVTDLVKLPGVGPKMAHLVMDIAWNNVSGIGVDTHVHRISNRLKWVRKE

TKTPEETRVAMEDWMPRELWSEINWLLVGFGQQVCLPVSPRCSECLNKDICPGAKKKKPR

>Xenopus\_(Silurana)\_tropicalis\_MutY

MPPPRTKTSLGRSAAASGKRKSPKQAFPKREEHVLQSSIYHSFTSQETEIIRDKLLAWYDKSKRDLPWRTMACTEPDLDR

KAYAVWVSEVMLQQTQVATVIDYYNKWMKVWPTMEDLARSSLEEVNEMWSGLGYYSRGRRLQEGAKKVVLELGGSMPRSA

DELQKLLPGVGRYTAGAIASISYGQVTGVVDGNVIRVLSRLRCIGADSSTLAVSDKLWNLANALVDPDRPGDFNQGMMEL

GATVCTPKKPLCTACPLQGQCKAYLKVIAEKESAVKTLIKKQASPIAKDVGDIEDCDLGPGLCALCVPTSDPWDSSLGVA

NFPRKSAKKPSRMEQTAICVWEKCGDHGELEYLIVQRPSSGLLAGLWEFPSILLDEKFTEQNRQHSLLGLLQDLSGHAVP

LQKLQYKGEVVHIFSHIHQTYVVYFLSLNTTENCSVKTEETERPLTRWVTKKEFLNSAVPTAMKKIMKLCESHGSSCTAV

NTSKKRKGDLAKVQLPSGRIKTEKGKQQSIQSFFKLATEK

>Takifugu\_rubripes\_Nth

MTSHYFAQSRSVVTRRGAQNAAHKPATSLKSKLTIQPEKDDLVSSSAGVKLEEEEAKISGNALKPETDAPTLSSHSRRRR

QLKVEYDKDGSMPQLKTEPWEPPRWKTQLENIRAMRSGRDAPVDNMGADKCHDADAPAHVKRFQVLVSLMLSSQTKDQVT

SAAMQKLRAHGCTVENILATNDETLGQLIYPVGFWRNKVKYLKLTSAMLQKEFGGDIPDSVEGLVRLPGVGPKMAHLAMD

IAWDQVSGIGVDTHVHRISNRLGWLKKPTKTPEETRKSLEEWLPRELWSEINWLLVGFGQQVCLPVSPLCSVCLNQHDCP

SAHKNSPVRRPKFLERSPWIKSPRLFITPGANTFLIR

>Takifugu\_rubripes\_MutY

MSTQGEPVQVGKVLSFLREWDRGDRSARGRMLSSFLGRSAGRTQGELEYLGEFVGHGGVVTLLEVLTQPQSNEETKAEAL

CLLLAISDAGRKYKELICQSCGAMAAAECLTHSGTGETQESAWMLLESLSHGNPKYEGEIYKGLIGHLTCTSAKAQQFVL

HTLHTLQSKMEIAHHSIVEPLLGVLTSLHPDVQSEVARLIFELRRYDVRPMLLRALCGLGLNVARAPTYPEEESSHASSP

SAYHFFHDAADVALLRSRLLAWYDQEKRELPWRTLALTEPDVNIRTYAVWVSEIMLQQTQVATVIDYYNKWMKRWPTVQD

LATATLEDVNQMWAGLGYYSRGKRLHEGAQKVVSQLQGEMPRTVDALLKQLPGVGRYTAGAIGSIALGQVTGAVDGNVIR

VLCRLRAIGADCTGPTVTEALWSLANTLVDPDRPGDFNQALMELGARVCTPKGPLCTRCPIQSHCHSYHKAHFKQEKNSM

KLMGKLDRKSSALPDIEDCLSSGTCTLCLSEPWDDELGVQNFPRKPAKKPPRAERCLTCVVIRQGEGGEHEFLLTQRPSK

GLLAGLWEFPCINHEEKNAVEEKKVLCAEINRILGTSLTHGLLQYVGEVVHIFSHIHQTYVVHTLRLKDAVSQSENMQWL

TPSALQEAAVSTGVKKIMKLCNSALGQQGAPDGEEKRPKKDRKGQITKRPRLSGANSRSRQLSLSSFFQTVKQDC

**APPENDIX 6:**

**4 human homologs, original headers**

>gi|4505471|ref|NP\_002519.1| endonuclease III-like protein 1 [Homo sapiens]

MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEG

SDSEKGEGAEPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSS

QTKDQVTAGAMQRLRARGLTVDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAE

LVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEING

LLVGFGQQTCLPVHPRCHACLNQALCPAAQGL

>gi|6912520|ref|NP\_036354.1| A/G-specific adenine DNA glycosylase isoform 1 [Homo sapiens]

MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPE

EVVLQASVSSYHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVA

TVINYYTGWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLP

GVGRYTAGAIASIAFGQATGVVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAM

ELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTL

GVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQE

LQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHTAAVSTAMKKV

FRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ

>gi|4505495|ref|NP\_002533.1| N-glycosylase/DNA lyase isoform 1a [Homo sapiens]

MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLTQTE

EQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQDPIEC

LFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGYRARYVSA

SARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYS

WHPTTSQAKGPSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG

>gi|4505121|ref|NP\_003916.1| methyl-CpG-binding domain protein 4 [Homo sapiens]

MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPIA

SAQFGATAGTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPED

FDFTVLSKRGIKSRYKDCSMAALTSHLQNQSNNSNWNLRTRSKCKKDVFMPPSSSSELQESRGLSNFTST

HLLLKEDEGVDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRKSCSGFVQSDSKRESVCNKADAESEPVAQ

KSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYED

TFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRTQIERRKTSLYFSS

KYNKEALSPPRRKAFKKWTPPRSPFNLVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKFLEKYPSAEV

ARTADWRDVSELLKPLGLYDLRAKTIVKFSDEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPED

HKLNKYHDWLWENHEKLSLS

**APPENDIX 7:**

**4 human homologs, modified headers**

>NTHL1\_Homo\_sapiens

MCSPQESGMTALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSHSPVKRPRKAQRLRVAYEG

SDSEKGEGAEPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSS

QTKDQVTAGAMQRLRARGLTVDSILQTDDATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAE

LVALPGVGPKMAHLAMAVAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEING

LLVGFGQQTCLPVHPRCHACLNQALCPAAQGL

>MUTYH\_Homo\_sapiens

MTPLVSRLSRLWAIMRKPRAAVGSGHRKQAASQEGRQKHAKNNSQAKPSACDGMIAECPGAPAGLARQPE

EVVLQASVSSYHLFRDVAEVTAFRGSLLSWYDQEKRDLPWRRRAEDEMDLDRRAYAVWVSEVMLQQTQVA

TVINYYTGWMQKWPTLQDLASASLEEVNQLWAGLGYYSRGRRLQEGARKVVEELGGHMPRTAETLQQLLP

GVGRYTAGAIASIAFGQATGVVDGNVARVLCRVRAIGADPSSTLVSQQLWGLAQQLVDPARPGDFNQAAM

ELGATVCTPQRPLCSQCPVESLCRARQRVEQEQLLASGSLSGSPDVEECAPNTGQCHLCLPPSEPWDQTL

GVVNFPRKASRKPPREESSATCVLEQPGALGAQILLVQRPNSGLLAGLWEFPSVTWEPSEQLQRKALLQE

LQRWAGPLPATHLRHLGEVVHTFSHIKLTYQVYGLALEGQTPVTTVPPGARWLTQEEFHTAAVSTAMKKV

FRVYQGQQPGTCMGSKRSQVSSPCSRKKPRMGQQVLDNFFRSHISTDAHSLNSAAQ

>OGG1\_Homo\_sapiens

MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLTQTE

EQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQDPIEC

LFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGYRARYVSA

SARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYS

WHPTTSQAKGPSPQTNKELGNFFRSLWGPYAGWAQAVLFSADLRQSRHAQEPPAKRRKGSKGPEG

>MBD4\_Homo\_sapiens

MGTTGLESLSLGDRGAAPTVTSSERLVPDPPNDLRKEDVAMELERVGEDEEQMMIKRSSECNPLLQEPIA

SAQFGATAGTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPED

FDFTVLSKRGIKSRYKDCSMAALTSHLQNQSNNSNWNLRTRSKCKKDVFMPPSSSSELQESRGLSNFTST

HLLLKEDEGVDDVNFRKVRKPKGKVTILKGIPIKKTKKGCRKSCSGFVQSDSKRESVCNKADAESEPVAQ

KSQLDRTVCISDAGACGETLSVTSEENSLVKKKERSLSSGSNFCSEQKTSGIINKFCSAKDSEHNEKYED

TFLESEEIGTKVEVVERKEHLHTDILKRGSEMDNNCSPTRKDFTGEKIFQEDTIPRTQIERRKTSLYFSS

KYNKEALSPPRRKAFKKWTPPRSPFNLVQETLFHDPWKLLIATIFLNRTSGKMAIPVLWKFLEKYPSAEV

ARTADWRDVSELLKPLGLYDLRAKTIVKFSDEYLTKQWKYPIELHGIGKYGNDSYRIFCVNEWKQVHPED

HKLNKYHDWLWENHEKLSLS