SUPPLEMENTAL NOTE 1

Files, Applications and Results / Data

The raw and aligned sequence files, the applications used to generate results, and resulting data sets of this research are available at http://cetaceanresearch.org/research/gyrencephaly_2013.

The specific parameters used in each program to generate results are detailed below:

RAxML

To generate maximum likelihood analyses with thorough bootstraps, the most recent version of raxmlGUI v1.3 was downloaded from http://sourceforge.net/projects/raxmlgui/. The following options were specified in raxmlGUI, which is equivalent to running the most recent version of RAxML with the following parameters:

ramxmlGUI options:
Add alignment file = <*gene_name*>_aligned.phy
Outgroup = unspecified; let raxmlGUI make its best assessments
Run mode = ML + thorough bootstrap
Run = 1
Replicates = 5000
BS brL = selected
Model = GTRGAMMA

RAxML command-line equivalence for these options is:

$raxmlHPC-PTHREADS-SSE3-Mac –T 2 –f c –m GTRGAMMA –s /Users/dave/Dropbox/__MarGen/<*gene_name*>/<*gene_name*>_aligned.phy –n <gene_name>_aligned.phy_red –w /Users/dave/Dropbox/__MarGen/<gene_name>/" -0

Once raxmlGUI maximum likelihood analysis and bootstraps were completed, a further step was taken in raxmlGUI to make the consensus bootstrap tree compatible with FigTree for Mac OS X, which appends node and branch bootstrap values to the tree file(s). This was done from raxmlGUI using the following user-interface seqeuence:

Utilites → Convert tree file to FigTree format

MrBayes

MrBayes version 3.2.1 for 64-bit Mac OS X systems was downloaded from http://mrbayes.sourceforge.net. MrBayes analysis was run from the command line using iTerm (http://iterm.sourceforge.net), using the

following command sequences for all Bayesian analyses and posterior probability generations:

For each respective gene, where the gene's name = <*gene_name*>:

```
$ mb

> execute <gene_name>_aligned.nexus
> lset nst=6 rates=invgamma
> outgroup=X
> mcmc ngen=200000 samplefreq=100 printfreq=100 diagnfreq=1000
```

If analysis split frequency was not less than 0.01, an additional 200,000 replicates would have been performed. However, in all cases, 200,000 replicates was sufficient.

Once analysis was completed for *PAFAH1B1*, *NDEL1* and *NEUROG1*, results were saved to files using the following two commands:

```
> sump
> sumt
```

It is worth noting that many of the extended NEXUS file formats are not easily interpreted by MrBayes. As such this specific format of the NEXUS file should be used:

--

```
#NEXUS
begin data;
        dimensions ntax=XX nchar=YYY;
        format datatype=dna missing=? gap=-;
matrix
Orcinus_orca_XM_004282101.1           ATGCCAGCC…
< … all remaining taxa sequences… >
;
end;
```

--

PAML and CODEML

PAML, which includes the CODEML module, was downloaded from http://abacus.gene.ucl.ac.uk/software/paml.html. In order to make PAML's CODEML module work with the .phy files used in raxmlGUI, a copy of the file used in raxmlGUI was made, then edited to ensure that (1) each taxa had two spaces after its name, and (2) that the letter "i" for interleaves or the letter "s" for sequential was appended to the end

```
of the first line, with two spaces between the number of taxa, the
number of DNA sequences in the alignment, and the added letter. This
is a peculiarity of paml's expectations for .phy files, but these
edits made paml work perfectly with these slightly modified .phy
files.

--------------------------------

H(0)
m0 (one-ratio) -- assumes one w (=dN/dS) for all codons in the
sequence
to compare m0 vs. m3
where P << 0.001 rejects m0
with df = 4

    seqfile = <genename>_aligned.phy  * NOTE: edit to include i on
first line
    treefile = <genename>_alltaxa.tree * NOTE: generate from HyPhy
    outfile = <genename>_m0_results.txt
    noisy = 3
    verbose = 1
    runmode = 0
    seqtype = 1
    CodonFreq = 2
    model = 0
    NSSites = 0
    icode = 0
    fix_kappa = 0
    kappa = 2
    fix_omega = 0
    omega = 5


--

H(A)
m3 (discrete) -- uses an unconstrained discrete distribution with all
three site
    classes estimated from the data, with w(0) < 1 and w(1) = 1
to compare m3 vs. m0

    seqfile = <genename>_aligned.phy  * NOTE: edit to include i on
first line
    treefile = <genename>_alltaxa.tree * NOTE: generate from HyPhy
    outfile = <genename>_m3_results.txt
    noisy = 3
    verbose = 1
    runmode = 0
    seqtype = 1
```

```
    CodonFreq = 2
    model = 0
    NSSites = 3
    icode = 0
    fix_kappa = 0
    kappa = 2
    fix_omega = 0
    omega = 5


--------------------------------

H(0)
m1a (nearly neutral) -- assumes two site classes estimated with data,
with
    w(0) < 1 and w(1) = 1
to compare m1a vs. m2a --  -- tests whether or not the analyzed region
evolves under
    positive selection, using comparisons to their nested neutral
models
where P < 0.001 rejects m1a
with df =  2

    seqfile = <genename>_aligned.phy  * NOTE: edit to include i on
first line
    treefile = <genename>_alltaxa.tree * NOTE: generate from HyPhy
    outfile = <genename>_m1a_results.txt
    noisy = 3
    verbose = 1
    runmode = 0
    seqtype = 1
    CodonFreq = 2
    model = 0
    NSSites = 1
    icode = 0
    fix_kappa = 0
    kappa = 2
    fix_omega = 0
    omega = 5


--

H(A)
m2a (positive selection - alternative hypothesis model) -- adds a
third class of sites
    to m1a, with w(2) > 1
to compare m1a vs. m2a
```

```
    seqfile = <genename>_aligned.phy  * NOTE: edit to include i on
first line
    treefile = <genename>_alltaxa.tree * NOTE: generate from HyPhy
    outfile = <genename>_m2a_results.txt
    noisy = 3
    verbose = 1
    runmode = 0
    seqtype = 1
    CodonFreq = 2
    model = 0
    NSSites = 2
    icode = 0
    fix_kappa = 0
    kappa = 2
    fix_omega = 0
    omega = 5


--------------------------------

H(0)
m7 (beta) -- a flexible null model, in which the w ratio for a codon
is a random draw
    with a beta distribution with 0 < w < 1
to compare m7 vs. m8 -- tests whether or not the analyzed region
evolves under
    positive selection, using comparisons to their nested neutral
models
with df = 2

    seqfile = <genename>_aligned.phy  * NOTE: edit to include i on
first line
    treefile = <genename>_alltaxa.tree * NOTE: generate from HyPhy
    outfile = <genename>_m7_results.txt
    noisy = 3
    verbose = 1
    runmode = 0
    seqtype = 1
    CodonFreq = 2
    model = 0
    NSSites = 7
    icode = 0
    fix_kappa = 0
    kappa = 2
    fix_omega = 0
    omega = 5


--
```

```
H(A)
m8 (beta and w) -- adds an extra class site to model m7, with a
proportion of
    w(s) > 1 estimated from the data
to compare m7 and m8a vs. m8

    seqfile = <genename>_aligned.phy  * NOTE: edit to include i on
first line
    treefile = <genename>_alltaxa.tree * NOTE: generate from HyPhy
    outfile = <genename>_m8_results.txt
    noisy = 3
    verbose = 1
    runmode = 0
    seqtype = 1
    CodonFreq = 2
    model = 0
    NSSites = 8
    icode = 0
    fix_kappa = 0
    kappa = 2
    fix_omega = 0
    omega = 5


--

H(0)
m8a (beta and w(s)=1) -- introduced by Swanson et al.; similar to m8
except that the
    category w(s) is fixed at w(s) = 1, specified in CODEML using
NSSite = 8
to compare m8 vs. m8a -- tests for evidence of positive selection
while eliminating
    the potential identification of relaxed purifying selection
with df = 1

    seqfile = <genename>_aligned.phy  * NOTE: edit to include i on
first line
    treefile = <genename>_alltaxa.tree * NOTE: generate from HyPhy
    outfile = <genename>_m8a_results.txt
    noisy = 3
    verbose = 1
    runmode = 0
    seqtype = 1
    CodonFreq = 2
    model = 0
    NSSites = 8
    icode = 0
```

```
    fix_kappa = 0
    kappa = 2
    fix_omega = 1
    omega = 1

--------------------------------
```

After installing paml to /usr/local/paml, codeml is run for each model using codeml.ctl files above, as follows:

```
$ /usr/local/paml/bin/codeml ./<model_number>.ctl
```

Assign significance of detection of positive selection on the selected branch, as follows:

Retrieve likelihood values lnL(H(A)) and lnL(H(0)) from alternative and null hypothesis
results files generated above.

Then reconstruct the Likelihood Ratio Test (LRT), as follows:

deltaLRT = 2 · (lnL(H(A)) - lnL(H(0)) (e.g.: 2 * ((-5710) - (-5712)) = 4)

In the above line, if deltaLRT = 4, and if chi^2 curve has one degree of freedom
(check the results of "$grep lnL *.results" for np: XX values of respective tests),
so p-value for chi^2 test = some small value under chi^2, so result is significant.

In cases where the result is significant, it is possible to retrieve the sites under positive selection using Bayes Empirical Bayes (BEB) method, which is described here: http://dx.doi.org/10.1093/molbev/msi097

```
    e.g.:
    Positive sites for foreground lineages Prob(w>1):
        36 K 0.971*
        159 C 0.993**
    Amino acids K and C refer to the first sequence in the alignment.
    Position 36 has  a high probability (97.1%) of being under
    positive selection.
    Position 159 has a very high probability (99.3%) of being under
    positive selection.

    See Table 4 for results that show this happening.
```