Title:

Phylogenetic analysis supports a link between DUF1220 domain number and primate brain expansion

Authors:
Fabian Zimmer
Stephen H. Montgomery*

Affiliations:
Dept. Genetics, Evolution & Environment, University College London, Gower Street, London, UK, WC1E 6BT

*Corresponding author
Email: Stephen.Montgomery@cantab.net

Key words: Autistic Spectrum Disorders, DUF1220 domains, NBPF, primate brain evolution.
Abstract

The expansion of DUF1220 domain copy number during human evolution is a dramatic example of rapid and repeated domain duplication. However, the phenotypic relevance of DUF1220 dosage is unknown. Although patterns of expression, homology and disease associations suggest a role in cortical development, this hypothesis has not been robustly tested using phylogenetic methods. Here, we estimate DUF1220 domain counts across 12 primate genomes using a nucleotide Hidden Markov Model. We then test a series of hypotheses designed to examine the potential evolutionary significance of DUF1220 copy number expansion. Our results suggest a robust association with brain size, and more specifically neocortex volume. In contradiction to previous hypotheses we find a strong association with postnatal brain development, but not with prenatal brain development. Our results provide further evidence of a conserved association between specific loci and brain size across primates, suggesting human brain evolution occurred through a continuation of existing processes.
Introduction

The molecular targets of selection favoring brain expansion during human evolution have been sought by identifying dramatic, lineage-specific shifts in evolutionary rate. The increase in DUF1220 domains during human evolution provides one of the most dramatic increases in copy number (Popesco et al., 2006; Dumas et al., 2012). A single copy of this protein domain is found in PDE4DIP in most mammalian genomes. In primates, this ancestral domain has been duplicated many times over, reaching its peak abundance in humans where several hundred DUF1220 domains exist across 20-30 genes in the Nuclear Blastoma Breakpoint Family (NBPF) (Vandepoele et al., 2005; Dumas et al., 2012). The majority of these map to 1q21.1, a chromosomal region with complex, and unstable genomic architecture (O’Bleness et al., 2012, 2014).

Interspecific DUF1220 counts show a pattern of phylogenetic decay with increasing distance from humans (Popesco et al., 2006; Dumas and Sikela, 2009; Dumas et al., 2012). In humans, DUF1220 dosage has also been linked to head circumference (Dumas et al., 2012), and severe neurodevelopmental disorders, including autism spectrum disorders (ASD) and microcephaly (Dumas et al., 2012; Davis et al., 2014). The severity of ASD impairments is also correlated with 1q21.1 DUF1220 copy number suggesting a dosage effect (Davis et al., 2014). Taken together, these observations led to the suggestion that the expansion of DUF1220 copy number played a primary role in human brain evolution (Dumas and Sikela, 2009; Keeney et al., 2014a).

The strength of this hypothesis is difficult to assess given the paucity of information on the developmental function of NBPF genes and the DUF1220 domain. DUF1220 domains are highly expressed during periods of cortical neurogenesis, suggesting a potential role in prolonging the proliferation of neural progenitors by regulating centriole and microtubule dynamics to control key cell fate switches critical for neurogenesis (Keeney et al., 2014b). PDE4DIP, which contains the ancestral DUF1220 domain, does indeed associate with the spindle poles (Popesco et al., 2006) and is homologous to CDK5RAP2, a centrosomal protein essential for neural proliferation (Bond et al., 2005; Buchman et al., 2010), which co-evolved with brain mass across primates (Montgomery et al., 2011).

Two previous analyses reported a significant association between DUF1220 copy number and brain mass, cortical neuron number (Dumas et al., 2012), cortical...
gray and white matter, surface area and gyrification (Keeney et al., 2014b). However, several limitations in these analyses restrict confidence in the results. First, DUF1220 copy number was assessed across species using a BLAT/BLAST analysis with a query sequence from humans, which introduces a bias that may partly explain the observed phylogenetic decay. Second, counts were not restricted to those domains occurring in functional exonic sequence. Finally, the analyses were limited to a small number of species (4-8 primates), and did not correct for phylogenetic non-independence (Felsenstein, 1985) or autocorrelation between traits.

Here, we use nucleotide Hidden Markov Models (HMMER3; Eddy, 2011) to more accurately query the DUF1220 domain number of distantly related genomes. After filtering these counts to limit the analysis to exonic sequence, we use phylogenetic comparative methods that correct for non-independence to test whether DUF1220 copy number is robustly associated with brain size, whether this is due to an association with pre- or postnatal brain development, and whether the association is specific to the neocortex.

Results

We find evidence that CM-associated exonic DUF1220 counts (Table 1) are associated with brain mass across primates (n = 12, posterior mean =1.927, 95% CI = 0.800-3.040, pMCMC = 0.001). This association is robust to the exclusion of Homo (posterior mean =1.271, 95% CI = 0.490-2.019, pMCMC = 0.003), and found when hominoids (n = 5, posterior mean = 3.679, 95% CI = 0.966-6.258, pMCMC = 0.018) or anthropoids (n = 9, posterior mean = 2.019, 95% CI = 0.352-3.684, pMCMC = 0.010) are analyzed alone, suggesting a consistent phylogenetic association. When body mass is included as a co-factor in the model, the positive association is restricted to brain mass (Table 2a).

Separation of pre- and postnatal development specifically links DUF12220 number to postnatal brain growth. Analysed separately, the association with prenatal brain growth is weaker (n = 11, posterior mean =1.758, 95% CI = -0.039-3.543, pMCMC = 0.023) than with postnatal brain growth (posterior mean =1.839, 95% CI = 0.895-2.808, pMCMC = 0.001). If both traits are included in the same model, only the positive association with postnatal brain growth remains (Table 2b). Multiple
regression analysis also confirms the association is specific to postnatal brain growth, rather than postnatal body growth (Table 2b).

Finally, we examined the hypothesized relationship with neocortex volume (e.g. Keeny et al., 2014a,b), but also consider cerebellum volume, as this region co-evolves with the neocortex (Barton and Harvey, 2000), has expanded in apes (Barton and Venditti, 2014), and shows high levels of NBPF expression (Popesco et al., 2006). When the rest-of-the-brain (RoB) is included as a co-factor, to account for variation in overall brain size, a positive association is found for neocortex volume but not cerebellum volume (Table 2c).

Discussion

Our phylogenetic analyses support the hypothesis that the increase in DUF1220 number co-evolves with brain mass, and contributes to the proximate basis of primate brain evolution. We also find evidence of specific associations with neocortex volume and postnatal brain growth. Previous hypotheses concerning the phenotypic relevance of DUF1220 domain number have focused on their possible contribution to neurogenesis (Dumas and Sikela, 2009; Keeney et al., 2014a; b). This is supported by homology to genes with known functions in cell cycle dynamics (Popesco et al., 2006; Thornton and Woods, 2009), relevant spatial and temporal expression patterns (Keeney et al., 2014b), and an effect on the proliferation of neuroblastoma cell cultures (Vandepoele et al., 2008). However, a direct effect of variation in DUF1220 domain number on neural proliferation has not been demonstrated (Keeney et al., 2015).

If DUF1220 domains do regulate neurogenesis, we would expect them to co-evolve with prenatal brain growth, as cortical neurogenesis is restricted to prenatal development (Bhardwaj et al., 2006). Our results instead suggest a robust and specific relationship with postnatal brain development. Existing data on DUF1220 domain function suggest two potential roles that may explain this association: i) a contribution to axonogenesis via initiating and stabilizing microtubule growth in dendrites; and ii) a potential role in apoptosis during brain maturation. Both hypotheses are consistent with the reported association between variation in DUF1220 dosage and ASD (Davis et al., 2014). Indeed, an emphasis on postnatal brain growth is potentially more relevant for ASD, which develops postnatal, accompanied by a period of accelerated brain growth (Courchesne et al., 2001).
Microtubule assembly is essential for dendritic growth and axonogenesis (Conde and Cáceres, 2009). PDEADIP, which contains the ancestral DUF1220 domain, has known functions in microtubule nucleation, growth, and cell migration (Roubin et al., 2013). There is also evidence NBPF1 interacts with a key regulator of Wnt signaling (Vandepoele et al., 2010), which has important roles in neuronal differentiation, dendritic growth and plasticity (Inestrosa and Varela-Nallar, 2014). Consistent with this function, DUF1220 domains are highly expressed in the cell bodies and dendrites of adult neurons (Popesco et al., 2006). A role for DUF1220 domains in synaptogenesis could potentially explain the association with ASD severity (Davis et al., 2014). ASDs are associated with abnormalities in cortical minicolumns (Casanova et al., 2002) and cortical white matter (Hazlett et al., 2005; Courchesne et al., 2011), both of which suggest a disruption of normal neuronal maturation (Courchesne and Pierce, 2005; Minshew and Williams, 2007).

Alternatively, NBPF genes are also known to interact with NF-κB (Zhou et al., 2013), a transcription factor implicated in tumor progression, with a range of roles including apoptosis and inflammation (Karin and Lin, 2002; Perkins, 2012). Postnatal apoptosis has a significant influence on brain growth (Kuan et al., 2000; Polster et al., 2003; Madden et al., 2007), including regulating neuronal density (Sanno et al., 2010), and apoptotic genes may have been targeted by selection in relation to primate brain expansion (Vallender and Lahn, 2006). Disruption of apoptosis causes microcephaly (Poulton et al., 2011), potentially explaining the association between DUF1220 dosage and head circumference (Dumas et al., 2012). The association of NF-κB with inflammatory diseases (Tak et al., 2001) is also intriguing, given the growing evidence that the inflammatory response is linked to the risk and severity of ASD (Meyer et al., 2011; Depino, 2012).

If DUF1220 domain number does contribute to the evolution of postnatal brain growth, this contrasts with results of previously studied candidate genes with known roles in neurogenesis that co-evolve with prenatal brain growth (Montgomery et al., 2011). This suggests a two-component model of brain evolution where selection targets one set of genes to bring about an increase in neuron number (e.g. Montgomery et al., 2011; Montgomery and Mundy, 2012a;b), and an independent set of genes to optimize neurite growth and connectivity (e.g. Charrier et al., 2012). NBPF genes may fall into the latter category. This two-component model is consistent with comparative analyses that indicate pre- and postnatal brain development evolve...
independently, and must therefore be relatively free of reciprocal pleiotropic effects (Barton and Capellini, 2011).

Finally, these results add further evidence that many of the genetic changes that contribute to human evolution will be based on the continuation or exaggeration of conserved gene-phenotype associations that contribute to primate brain evolution (Montgomery et al., 2011; Scally et al., 2012). Understanding the commonalities between human and non-human primate brain evolution is therefore essential to understand the genetic differences that contribute the derived aspects of human evolution.

Materials and methods

Counting DUF1220 domains

HMMER3.1b (Eddy, 2011) was used to build a Hidden Markov Model (HMM) from the DUF1220 (PF06758) seed alignment stored in the PFAM database (Finn et al., 2014). The longest isoforms for all proteomes of 12 primate genomes from Ensembl v.78 (Cunningham et al., 2014) (Figure 1A), were searched using the protein DUF1220 HMM (hmmsearch, E-value < 1e-10) (Table S1). We extracted the corresponding cDNA regions to build a DUF1220 nucleotide profile HMM (nHMM), allowing for more sensitive analysis across a broad phylogenetic range. The DUF1220 nHMM was used to search the complete genomic DNA for all 12 species. These counts were filtered to remove any DUF1220 domains not located in annotated exonic sequence, or located in known pseudogenes.

We next filtered our counts to limit them to exonic sequence in close proximity to the NBPF-specific Conserved-Mammal (CM) promoter (O’Bleness et al. 2012). To do so, we built a nucleotide HMM for the CM promoter based on a MAFFT (Katoh et al., 2002) alignment of the 900bp CM region upstream of human genes NBPF4, NBPF6 and NBPF7. Using this CM promoter nHMM, we searched 1000bp up- and downstream of genes containing DUF1220 domains for significant CM promoter hits (nhmmer, E-value < 1e-10). This provided final counts for DUF1220 domains within exonic regions and associated with the CM promoter (Table 1). These counts were used in subsequent phylogenetic analyses. In the Supplementary Information we compare our counts with previous estimates and discuss possible sources of error. All scripts and data used in the analysis are freely available from:

https://github.com/qfma/duf1220
**Phylogenetic gene-phenotype analysis**

Phylogenetic multivariate generalized mixed models were implemented using a Bayesian approach in MCMCglmm (Hadfield, 2010), to test for phylogenetically-corrected associations between DUF1220 counts and log-transformed phenotypic data (Table S2). All analyses were performed using a Poisson distribution, as recommended for count data (O’Hara and Kotze, 2010), with uninformative, parameter expanded priors for the random effect (G: V = 1, n ν = 1, alpha.ν = 0, alpha.V = 1000; R: V = 1, ν = 0.002) and default priors for the fixed effects. Phylogenetic relationships were taken from the 10k Trees project (Arnold et al., 2010). We report the posterior mean of the co-factor included in each model and its 95% confidence intervals (CI), and the probability that the parameter value is >0 (pMCMC) as we specifically hypothesize a positive association (Dumas et al., 2012). Alternative data treatments lead to similar conclusions (Supplementary Information).

**Acknowledgments and funding information**

We thank James Sikela, Majesta O’Bleness, Chris Venditti, Charlotte Montgomery, Andrew Moore and Jarrod Hadfield for advice and comments, and Judith Mank’s lab at UCL for support. SHM thanks the Leverhulme Trust for funding.

**References**


Keeney J, Dumas L, Sikela J. 2014a. The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion. Front Hum Neurosci 8:1–11.


### Table 1: DUF1220 count data

<table>
<thead>
<tr>
<th>Species</th>
<th>O'Bleness et al. (2012)</th>
<th>nHMM whole genome</th>
<th>Functional exonic with CM promoter</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>272</td>
<td>302</td>
<td>262</td>
</tr>
<tr>
<td><em>Pan troglodytes</em></td>
<td>125</td>
<td>138</td>
<td>34</td>
</tr>
<tr>
<td><em>Gorilla gorilla</em></td>
<td>99</td>
<td>97</td>
<td>32</td>
</tr>
<tr>
<td><em>Pongo abelii</em></td>
<td>92</td>
<td>101</td>
<td>27</td>
</tr>
<tr>
<td><em>Nomascus leucogenys</em></td>
<td>53</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td><em>Papio anubis</em></td>
<td>-</td>
<td>75</td>
<td>15</td>
</tr>
<tr>
<td><em>Chlorocebus sabaeus</em></td>
<td>-</td>
<td>48</td>
<td>16</td>
</tr>
<tr>
<td><em>Macaca mulatta</em></td>
<td>35</td>
<td>74</td>
<td>10</td>
</tr>
<tr>
<td><em>Callithrix jacchus</em></td>
<td>31</td>
<td>75</td>
<td>9</td>
</tr>
<tr>
<td><em>Tarsius syrichta</em></td>
<td>-</td>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td><em>Microcebus murinus</em></td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td><em>Otolemur garnettii</em></td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>
### a) Brain mass and body mass

<table>
<thead>
<tr>
<th>Model</th>
<th>Posterior mean</th>
<th>95% CI</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. log(brain mass)</td>
<td>4.105</td>
<td>2.163 - 6.000</td>
<td>0.001</td>
</tr>
<tr>
<td>+ log(body mass)</td>
<td>-1.986</td>
<td>-3.544 - -3.900</td>
<td>0.988</td>
</tr>
</tbody>
</table>

### b) Prenatal and postnatal growth

<table>
<thead>
<tr>
<th>Model</th>
<th>Posterior mean</th>
<th>95% CI</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. log(prenatal brain growth)</td>
<td>-2.158</td>
<td>-4.471 - 0.106</td>
<td>0.967</td>
</tr>
<tr>
<td>+ log(postnatal brain growth)</td>
<td>3.319</td>
<td>1.470 - 4.982</td>
<td>0.002</td>
</tr>
<tr>
<td>2. log(postnatal brain growth)</td>
<td>2.910</td>
<td>1.641 - 4.151</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>+ log(postnatal body growth)</td>
<td>-1.241</td>
<td>-2.442 - -0.052</td>
<td>0.977</td>
</tr>
</tbody>
</table>

### c) Brain regions

<table>
<thead>
<tr>
<th>Model</th>
<th>Posterior mean</th>
<th>95% CI</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. log(neocortex volume)</td>
<td>5.961</td>
<td>0.720 - 11.173</td>
<td>0.014</td>
</tr>
<tr>
<td>+ log(RoB volume)</td>
<td>-5.817</td>
<td>-13.322 - 1.120</td>
<td>0.953</td>
</tr>
<tr>
<td>2. log(cerebellum volume)</td>
<td>3.699</td>
<td>-5.857 - 12.611</td>
<td>0.186</td>
</tr>
<tr>
<td>+ log(RoB volume)</td>
<td>-2.435</td>
<td>-13.869 - 10.132</td>
<td>0.681</td>
</tr>
<tr>
<td>3. log(neocortex volume)</td>
<td>6.076</td>
<td>-0.139 - 12.5712</td>
<td>0.025</td>
</tr>
<tr>
<td>+ log(cerebellum volume)</td>
<td>-0.369</td>
<td>-9.5128 - 8.961</td>
<td>0.526</td>
</tr>
<tr>
<td>+ log(RoB volume)</td>
<td>-5.494</td>
<td>-15.814 - 5.288</td>
<td>0.872</td>
</tr>
</tbody>
</table>
Figures

Figure 1: A) Phylogeny of Ensembl primates. B) Posterior means of the association between DUF1220 count and brain mass (red) and body mass (black). C) Posterior means of the association between DUF1220 count and postnatal brain growth (red) and prenatal brain growth (black). D) Posterior means of the association between DUF1220 count and neocortex volume (red), cerebellum volume (solid black) and rest-of-brain volume (dashed black).