The twenty badger samples donated by Secret World Wildlife Rescue (SWWR) were all cubs orphaned following recent culls and flooding, but were in good condition except where noted in Table 4. Locations were mapped against the Woodchester Park study group, giving seven distinct new locations for this trial, seen in Figure 6. Seven badger samples were donated by the University College Dublin, from badgers raised on site as part of their academic observations, which have a shared familial history within the same location. The three sibling cohorts are listed in Table 5. The roadkill Badger ‘G’ was found in a mangled, decomposed, state which made sexing the animal impossible. Samples of its hide were taken for DNA extraction. Badger G were found close to Leamington Spa, with latitude and longitude 52°11'25.8"N 1°32'39.0"W. This location has been added to the map in Figure 6.

Table 4. Recording details for Secret World Badger Samples

<table>
<thead>
<tr>
<th>ID</th>
<th>Location</th>
<th>Distance from WP</th>
<th>On map</th>
<th>Age</th>
<th>Gender</th>
<th>Weight</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Shepton Mallet</td>
<td>43.7 miles S</td>
<td>A</td>
<td>16 weeks</td>
<td>Male</td>
<td>4.82kg</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>BA5 1QQ</td>
<td>46.4 miles S</td>
<td>C</td>
<td>14 weeks</td>
<td>Male</td>
<td>3.94kg</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>BA5 1QS</td>
<td>46.7 miles S</td>
<td></td>
<td>7 weeks</td>
<td>Male</td>
<td>Not given</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Locking BS24 8PL</td>
<td>44.3 miles SW</td>
<td>B</td>
<td>12 weeks</td>
<td>Both M</td>
<td>Both 2.7kg</td>
<td>Siblings *</td>
</tr>
<tr>
<td>1</td>
<td>Weston Super Mare</td>
<td>46.7 miles SW</td>
<td>D</td>
<td>16 weeks</td>
<td>Both M</td>
<td>4.28kg</td>
<td>Siblings</td>
</tr>
<tr>
<td>10</td>
<td>BA21 4PW</td>
<td>62.8 miles S</td>
<td>E</td>
<td>12 weeks</td>
<td>Female</td>
<td>3.5kg</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>CV11 6LA</td>
<td>76.3 miles NE</td>
<td>F</td>
<td>16 weeks</td>
<td>Male</td>
<td>Not given</td>
<td>Siblings</td>
</tr>
<tr>
<td>15</td>
<td>CV11 6LA</td>
<td>76.3 miles NE</td>
<td></td>
<td></td>
<td>Male</td>
<td>Not given</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CV11 6LA</td>
<td>76.3 miles NE</td>
<td></td>
<td></td>
<td>Female</td>
<td>Not given</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Shropshire</td>
<td>76.8 miles N</td>
<td>G</td>
<td>11 weeks</td>
<td>Male</td>
<td>Not given</td>
<td>*</td>
</tr>
<tr>
<td>16</td>
<td>Shropshire</td>
<td>Unknown</td>
<td></td>
<td>8 weeks</td>
<td>Male</td>
<td>2.99kg</td>
<td>Siblings</td>
</tr>
<tr>
<td>12</td>
<td>Shropshire</td>
<td>Unknown</td>
<td></td>
<td>10 weeks</td>
<td>Male</td>
<td>2.85kg</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Venn Ottery,</td>
<td>95.0 miles SW</td>
<td>H</td>
<td>8 weeks</td>
<td>Male</td>
<td>1.75kg</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>EX10 9EX.</td>
<td>202.0 miles SE</td>
<td>I</td>
<td>12 weeks</td>
<td>Male</td>
<td>4.64kg</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>St. Ives TR26 2JB</td>
<td>202.0 miles SE</td>
<td>I</td>
<td>8 weeks</td>
<td>Male</td>
<td>2.46kg</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Not Given</td>
<td>Unknown</td>
<td></td>
<td>14 weeks</td>
<td>Male</td>
<td>Not given</td>
<td>Siblings</td>
</tr>
<tr>
<td>3</td>
<td>Not Given</td>
<td>Unknown</td>
<td></td>
<td>12 weeks</td>
<td>Male</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Not Given</td>
<td>Unknown</td>
<td></td>
<td>12 weeks</td>
<td>Male</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend: SWWR badgers are grouped into 8 groups according to geographical distribution, marked by alternation of the colours yellow and green. "*" indicates found emaciated next to deceased mother (culled).
Figure 6. Map showing relative geographical location of outgroup badgers.

Legend: Red labels correspond to locations in figure X by alphabetical reference, Blue=WP, Green= Badger G. The map gives only a sense of the breadth of geographical distribution and is cropped for clarity, there were also samples from St. Ives (off map.)
Table 5. UCD Badger sample details

<table>
<thead>
<tr>
<th>ID</th>
<th>Age (yrs)</th>
<th>Gender</th>
<th>Weight</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>2</td>
<td>F</td>
<td>7.7kg</td>
<td>Siblings</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>F</td>
<td>8.3kg</td>
<td>B. 2012</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>F</td>
<td>8.5kg</td>
<td>Siblings</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>M</td>
<td>8.8kg</td>
<td>B. 2013</td>
</tr>
</tbody>
</table>

DNA extraction from was performed after practice on a non-project related primate blood source, Sample J, which yielded 290.8 ng/μl DNA with a purity of 1.81/2 and was deemed satisfactory. Samples 1 through 20 comprised 400μl to 800μl red blood cells per tube. Additional extraction attempts were made on 2, 3, 6, 19 and 20 after first attempts yielded low DNA levels. No sample remained for re-extraction after the first attempt on 10. Badger 20 was excluded from the trial as DNA remained impure following further extraction and clean up. Dilutions were made to reach a concentration of 100ng/μl. Extraction protocols for badger samples 21 to 27 were optimized to extract from the 40μl white blood cells per sample, using 100μl lysis buffer, with succeeding stages performed with 90μl phenol-chloform per step. While this gave good purity ratings, DNA yields were low (Table 6, left).

Prior to extraction from Badger G two trials were performed using store-bought beef. This resulted in strong streaking when run on agarose gel and spectrophotometry testing gave purity readings of 1.93/2 and 1.94/2 and concentrations of 1469.1 µg/ml and 1687.0 µg/ml.

Tubulin PCR testing of extracted DNA was performed. Tubulin is a highly repeated gene found in multiple copies throughout the genome. A test for tubulin indicates that DNA is capable of amplification, lacking inhibitors that may prevent this (Terry et al 2001). The primers and PCR settings used were as follows: Primer MtubF (5’-CGTGAGTGCATCTCCATCCAT-3’) and MtubR (5’-GCCCTCACCATACACCAGTG-3’) underwent PCR at 94°C for 5 minutes, 40 cycles of 94°C for 50 seconds, 55°C for 1 minute and 72°C for 1 minute 30 seconds, then 72°C for 10 minutes. This was performed in order to ensure that the DNA is viable for amplification with untested primers. However, all extractions from badger G failed initial PCR testing using tubulin primers. Spectrophotometery (Table 6, right) showed badger G DNA was highly contaminated, but had good concentrations of DNA. Clean up of all five extractions combined achieved higher purity (1.55) and a concentration of 513.1 µg/ml, and were retained for PCR.
Table 6. Results of Spectrophotometer purity and quantity test for outgroup badgers

<table>
<thead>
<tr>
<th>Badger G only</th>
<th>ID</th>
<th>Purity n/2</th>
<th>Conc. µg/ml</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.89</td>
<td>361.3</td>
</tr>
<tr>
<td>2</td>
<td>2*</td>
<td>1.65</td>
<td>133.6</td>
</tr>
<tr>
<td>3</td>
<td>3*</td>
<td>1.77</td>
<td>327.3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1.81</td>
<td>258.1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1.81</td>
<td>224.6</td>
</tr>
<tr>
<td>6*</td>
<td>6*</td>
<td>1.85</td>
<td>334.3</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>1.82</td>
<td>229.8</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>1.79</td>
<td>167.9</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1.85</td>
<td>298.9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>1.37</td>
<td>128.8</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>1.82</td>
<td>203.4</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>1.75</td>
<td>100.4</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>1.79</td>
<td>127.6</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>1.78</td>
<td>170.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ex.</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>15</td>
<td>1.80</td>
<td>155.0</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>1.78</td>
<td>134.0</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>1.78</td>
<td>149.1</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>1.88</td>
<td>476.8</td>
</tr>
<tr>
<td>19*</td>
<td>19</td>
<td>1.87</td>
<td>124.4</td>
</tr>
<tr>
<td>20**</td>
<td>20</td>
<td>1.55</td>
<td>10.3</td>
</tr>
<tr>
<td>21*</td>
<td>21</td>
<td>1.74</td>
<td>40.7</td>
</tr>
<tr>
<td>22*</td>
<td>22</td>
<td>2.0</td>
<td>2.3</td>
</tr>
<tr>
<td>23*</td>
<td>23</td>
<td>1.7</td>
<td>0.6</td>
</tr>
<tr>
<td>24*</td>
<td>24</td>
<td>1.71</td>
<td>18.4</td>
</tr>
<tr>
<td>25*</td>
<td>25</td>
<td>1.74</td>
<td>5.6</td>
</tr>
<tr>
<td>26*</td>
<td>26</td>
<td>1.79</td>
<td>52.1</td>
</tr>
<tr>
<td>27*</td>
<td>27</td>
<td>1.38</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Legend: Samples highlighted in red were excluded from the PCR stage due to extreme low purity or concentration. * indicates additional attempts at extraction. Ex= Extraction number for Badger G. All= Result of clean up of all 5 badger G extractions combined.

To find the optimal DNA concentration in preparation for the PCR stage, two fold serial dilutions of the second beef steak extraction were made and tested using a Tubulin PCR. The best result, seen in Figure 7, was produced at around 100 µg/ml, because this dilution produced the strongest band with the greatest similarity to the strength of the positive. This was taken as the optimal concentration and all extracted DNA was diluted to this level.

Figure 7. Beef DNA dilution experiment for tubulin PCR.

Legend: Lanes 1-9 are tubulin primer PCR of the beef steak extraction in decreasing concentration; Lane 1=210.8 µg/ml, 2=105.4 µg/ml, 3= 52.7 µg/ml, 4= 26.36 µg/ml, 5= 13.18 µg/ml, 6= 6.59 µg/ml, 7= 3.29 µg/ml, 8= 1.65 µg/ml, 9= 0.82 µg/ml, 10= positive tubulin control, 0= negative control.
In order to design primers to select targets for amplification in the badger, known TLR sequences from other species were collated. The following databases were investigated for published badger TLR sequences, or a whole badger genome, and all TLRs in all species: ARNIE, ARSA, BLAST, Bioinformatic Harvester, Cosmic, DECIPHER, DGVa, DNA Databank of Japan, DRALError, EBI, EBI-metagenomics, ENCODE, Ensembl, euGenes, European Genome Archive, EXOMISER, Gene Cards, GeneDB, GLIDES, Google, GWAS, MEROPS, PubMed-NCBI, PATRIC, Pfam, Phenogrid, Rfam, Sanger Centre, Source (Stanford,) Quantomics, Tiffin, Treefam, TXSearch, Vega, Wormbase, and ZF-models. Of these, PubMed-NCBI and ENSEMBL held the greatest abundance of TLR sequences for different species. Cosmic, EBI, and VEGA held some additional, limited, TLR data. None of the databases bore a TLR sequence for the badger, or the full badger genome at the time of searching (September, 2013). Relevant TLRs for all available mammals were searched, and over forty were found with sequences available for TLR2 and TLR4; *Mustela putorius furo, Odobenus rosmarus divergens, Pteropus alecto, Myotis brandtii, Equus asinus, Equus caballus, Orcinus orca, Lageno rhynchos obliquidens, Tursiops truncatus, Bos indicus, Bos taurus, Bison bison, Bubalus bubalis, Cerco cebusatys, Macaca fascicularis, Otolemur garnetii, Loxodonta africana, Pantherolophus verus, Hylobateslar, Callithrix jacchus, Sorex araneus, Homo sapiens, Pan paniscus, Papio anubus, Cricetulus griseus, Felis catus, Gorilla gorilla, Pongo pygmaeus, Ailuropoda melanoleuca, Sus scrofa, Boselaphustrago camelus, Myodes, Eothenomys, Arvicola, Microtus, Chionomys, Micromys, Apodemus, Canis familiaris, Mus musculus, Rattus norvegicus, Condylura cristata and Capra hircus*. The most relevant were the ferret (*Mustela putorius furo*), cat (*Felis catus*), dog (*Canis familiaris*), panda (*Ailuropoda melanoleuca*) and walrus (*Odobenus rosmarus divergens*) which are all found in the order Carnivora along with the badger. Having gathered sequences of interest, there was scope to produce Multiple Sequence Alignments.

First, it was important to identify the targetted proteins so as to differentiate exons and introns. Exons form the DNA code which results in the final protein configuration, while introns do not code for proteins. Since the targets were TLR proteins, the exons were of greatest interest and would form the target of any primers designed. A Matlab code was written in order to convert strings of base pairs of any length into their final protein sequence
This was then used to compare nucleotide sequences to protein sequences found in online databanks and thus identify exons and introns.

**Figure 8. ‘DNAToProtein’ Matlab function (written by Andrew Whiteoak.)**

```matlab
function Amino Acid Sequence = DNAToProtein(bpSequence)
N = floor(length(bpSequence)/3); M = N*3;
Threesomes = reshape(bpSequence(1:M),3,N)';
Result = zeros(1,N);
Result(strmatch('AAA',Threesomes)) = 'K';
Result(strmatch('AAG',Threesomes)) = 'K';
Result(strmatch('AAC',Threesomes)) = 'N';
Result(strmatch('AAT',Threesomes)) = 'N';
Result(strmatch('ACA',Threesomes)) = 'T';
Result(strmatch('ACC',Threesomes)) = 'T';
Result(strmatch('ACG',Threesomes)) = 'T';
Result(strmatch('ACT',Threesomes)) = 'T';
Result(strmatch('ACN',Threesomes)) = 'T';
Result(strmatch('AGA',Threesomes)) = 'R';
Result(strmatch('AGG',Threesomes)) = 'R';
Result(strmatch('CGA',Threesomes)) = 'R';
Result(strmatch('CGC',Threesomes)) = 'R';
Result(strmatch('CGG',Threesomes)) = 'R';
Result(strmatch('CGT',Threesomes)) = 'R';
Result(strmatch('CGN',Threesomes)) = 'R';
Result(strmatch('AGC',Threesomes)) = 'S';
Result(strmatch('AGT',Threesomes)) = 'S';
Result(strmatch('TCA',Threesomes)) = 'S';
Result(strmatch('TCC',Threesomes)) = 'S';
Result(strmatch('TCG',Threesomes)) = 'S';
Result(strmatch('TCT',Threesomes)) = 'S';
Result(strmatch('TCN',Threesomes)) = 'S';
Result(strmatch('ATA',Threesomes)) = 'I';
Result(strmatch('ATC',Threesomes)) = 'I';
Result(strmatch('ATT',Threesomes)) = 'I';
Result(strmatch('ATG',Threesomes)) = 'M';
Result(strmatch('CAA',Threesomes)) = 'Q';
Result(strmatch('CAG',Threesomes)) = 'Q';
Result(strmatch('CAT',Threesomes)) = 'H';
Result(strmatch('CAC',Threesomes)) = 'H';
Result(strmatch('CCA',Threesomes)) = 'P';
Result(strmatch('CCC',Threesomes)) = 'P';
Result(strmatch('CCT',Threesomes)) = 'P';
Result(strmatch('CCN',Threesomes)) = 'P';
Amino Acid Sequence = char(Result);
End
```

In order to infer the position of exons and introns, the TLR2 and TLR4 protein sequences for the closest relatives of the badger; the ferret, cat, dog, panda and walrus (most closely related due to being of the order Carnivora) were compared to their nucleotide sequences.

To this end it was found that TLR2 has a single, unbroken, 2355 bp exon (Figure 9), and TLR4 has three exons and two introns (Figure 10). This was an incredibly useful venture because it allows the researcher to make useful decisions as to where to design primers. Clearly, for TLR2, it is simply a matter of crawling along the single exon searching for areas of high
sequence conservation between the collection of sequences. While for TLR 4, avoidance of intron spaces between the three exons would be crucial.

Figure 9. The overall nucleotide structure of TLR2

<table>
<thead>
<tr>
<th>Intron</th>
<th>Exon</th>
<th>Intron</th>
</tr>
</thead>
<tbody>
<tr>
<td>600bp</td>
<td>2355bp</td>
<td>603bp</td>
</tr>
</tbody>
</table>

Figure 10. The overall nucleotide structure of TLR4

<table>
<thead>
<tr>
<th>Exon</th>
<th>First Intron</th>
<th>Exon</th>
<th>Second Intron</th>
<th>Exon</th>
</tr>
</thead>
<tbody>
<tr>
<td>96 bp</td>
<td>6059 bp</td>
<td>165 bp</td>
<td>3243 bp</td>
<td>2238 bp</td>
</tr>
</tbody>
</table>

Multiple Sequence Alignments (MSA) allow researchers to see where nucleotides or proteins are conserved across species. These regions indicate possibilities for primer design for unknown species sequences due to their higher probability of their also being conserved in the same manner. Current bioinformatics researchers recommend Clustal software as the fastest (Sievers et al 2011) and most sensitive method (Soding, 2005) of performing MSAs, outperforming other packages (Stevens et al 1998). It is recommended due to its capacity to align any number of protein sequences using modest amounts of memory (Blackshields et al 2010). To sequence large regions of DNA, it is necessary to subdivide a large fragment into smaller ones to be individually sequenced (Slatko et al 1999). A full protein alignment for TLR4 for 46 species was made, which showed that the minke whale Balenoptera acutostrata, Omura's whale Balenoptera omuraim, the Chinese white dolphin Sousa chinensis, the killer whale Orcinus orca, the common dolphin Delphinus capensis and the bottle nosed dolphin Tursiops truncates do not carry exon 1 or exon 2, but begin their protein sequence at exon 3. Alignments amongst the order Carnivora were used to identify conserved sequences, so that primers could be drawn from them. Conserved areas were identified from which primers were designed to cover all of TLR2 and exons 2 and 3 of TLR4. A Computer Code was created to give reverse primer sequences which was highly useful during the primer design stage. Forward primers are read straight from a given sequence, in a 5' to 3' direction. However, reverse primers are the reverse complement of a given sequence. A Matlab function was created to perform this task (McMahon, 2007). The code is shown in Figure 11. Letters other
than GATC indicate base pair redundancy (degenerate bases) and their reverse complement.

Using this function researchers can copy-paste in the sequence where a primer possibility lies, and have Matlab output the sequence as a reverse primer.

Figure 11. Function for creating Reverse Primers in Matlab (written by Whiteoak, 2013).

```matlab
function Comp = ReversePrimer(Primer)
%
%This function gives the reverse primer including degenerate bases
b=fliplr(Primer);
%%fliplr(Primer) reverses the direction of the forward primer
% but does not complement it.
Comp(b=='A')='T';
Comp(b=='T')='A';
Comp(b=='G')='C';
Comp(b=='C')='G';
Comp(b=='B')='V';
Comp(b=='V')='B';
Comp(b=='D')='H';
Comp(b=='H')='D';
Comp(b=='K')='M';
Comp(b=='M')='K';
Comp(b=='R')='Y';
Comp(b=='Y')='R';
Comp(b=='S')='W';
Comp(b=='W')='S';
Comp(b=='N')='N';
End
```

Lvovsky et al (1998) discussed the effect of secondary structure on very small primer lengths with only eight bases (8-mers) and two degenerate positions. Degenerate bases are manufactured replacements for single nucleotides which can code for more than one nucleotide and therefore attach to more than one possible complementary. These short 8-mer primers were significantly affected by local structure, and hence the folding energy of the primers, leaving them more highly prone to dimerisation and requiring strong PCR optimization with respect to temperature and cycling to improve their efficiency.

Dimerisation causes the amplification of non-target lengths of amplified DNA which form when primers complement one another and join together, amplifying themselves rather than the target band. Even when the target band is amplified, the dimer creates contamination which will hamper sequencing results, and thus must be avoided. This can occur when forward and reverse primers complement, or when either the forward or the reverse complements itself. There is method by which primers can be checked manually where the reverse complement of each primer is written by first writing the primer letters out backwards, then writing out their complement. If this is a close match to either the other primer or itself then a primer dimer is likely. The programme created in MatLab ‘ReversePrimer’ can also perform the former function while the latter is manual. An example of this is given in Figure 12, showing the various dimers given with two hypothetical primers.

During primer design, conserved sequences around target areas were investigated until
primer pairs with no dimerization were found. Lvovsky et al (1998) pointed out that this became less of an issue for conventional longer primers that have more selectivity.

Figure 12. Two hypothetical primers and their dimers

<table>
<thead>
<tr>
<th>ExampleFwd</th>
<th>AGTCGTTAGCTCGAAT</th>
<th>ExampleRev</th>
<th>CGCTAGGCTTACCTAG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-Dimers</strong></td>
<td></td>
<td><strong>Cross Primer Dimers</strong></td>
<td></td>
</tr>
<tr>
<td>5-cgctaggctacctag-&gt;</td>
<td>5-agtctaggctcgaat-&gt;</td>
<td>&lt;-gatccattcggatgcgc-5</td>
<td>&lt;-gatccattcggatgcgc-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-cgctaggctacctag-&gt;</td>
<td>5-agtctaggctcgaat-&gt;</td>
<td>&lt;-gatccattcggatgcgc-5</td>
<td>&lt;-gatccattcggatgcgc-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-cgctaggctacctag-&gt;</td>
<td>5-agtctaggctcgaat-&gt;</td>
<td>&lt;-gatccattcggatgcgc-5</td>
<td>&lt;-gatccattcggatgcgc-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Jaric et al (2013) looked at the design of universal primers, in which they combined the standard design rules with an automated database search and optimization to design primers that could be used over a range of species more effectively to identify their presence as well as their prevalence. Here, the bases were optimized to achieve a limited specificity that was able to work efficiently across a large range species of species. By contrast, our research requires a very high degree of specificity. We were concerned about the **specificity** of primers with a large number of full or partial degeneracies and could find no literature regarding this effect. Hence we developed the following simple combinatorial model to evaluate the effect of degeneracies on the primer’s **specificity**.

Table 7. Standard code for degenerate base specification, paired by complement.

<table>
<thead>
<tr>
<th>Degenerate base</th>
<th>B</th>
<th>V</th>
<th>D</th>
<th>H</th>
<th>K</th>
<th>M</th>
<th>R</th>
<th>Y</th>
<th>S</th>
<th>W</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bases covered</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

All degenerate bases available are show in Table 7. Six of them allow selection to waver between two possible target bases ("x"). Four allow selection to waver between three possible target bases ("y"). One allows selection to waver between all four possible target bases ("z").
An equation was derived to measure relative specificity if the primer length ("L") the the amount and type of degenerate bases are known using the following deduction. Where no degenerate bases are involved, specificity must be a function of length. Let us take, for simplicity, a primer with a length of three nucleotides. There are 64 possible combinations using four base pairs over those three nucleotides; this is a function of $4^3$ or $4^L$ where 4 equals the amount of bases in DNA (G, A, T and C,) and L=primer length. With no degenerate bases, the primer would have a specificity of 1 in 64, as it would target a specific 3 bp sequence from 64 possibilities; and would rewrite as $\frac{1}{4L}$. If one of the three bases in this primer is a two-way degenerate base, then while there are only 64 permutations of GATC in three positions, there are now two chances of matching one of them. Therefore specificity drops to 2 in 64. Likewise, if two bases are two-way degenerate bases, specificity will drop to 4 in 64. From the deductions above we can see that specificity becomes a function of the power of the amount of variation in the degenerate bases. The final equation was written as:

$$Specificity = \frac{2^x \times 3^y \times 4^z}{4^L}$$

This equation was extended to show that length is more critical for specificity than the ratio of specific nucleotides to degenerate bases. The presence of degenerate bases reduces specificity but not as much as reducing primer length does. To express the reduction in effective sequence length in base pairs as a product of any combination of degenerate bases on an (x,y,z) plot it was useful to know that a four way degenerate base has an equivalent effect on specificity as two two-way degenerate bases, both reduce specificity by $\frac{1}{4}$. The equation was shown as follows:

$$Specificity = \frac{2^x \times 3^y \times 2^{2z}}{4^L} = \frac{2^{(x+2z)} \times 3^y}{4^L}$$

This gave a function in two variables. Given that $2 = \sqrt{4} = 4^{1/2}$, then,

$$Specificity = \frac{4^{(\frac{x}{2}+z)} \times 3^y}{4^L} = 4^{-L-(\frac{x}{2}+z)-log4(3^y)}$$

Reduction of specificity in effective length in bp $= \frac{x}{2} + log_4(3^y) + z$

Figure 13 was plotted using two and three way degenerate bases to show their combined effect on specificity, as expressed by how many base pairs their presence would effectively remove. The reductive slope caused by three-way degenerate bases is much steeper than that caused by two-way degenerate bases, as would be expected. Length has the largest
impact on a primer’s specificity. Any increase in sensitivity caused by a degenerate base is far outweighed by simply increasing the nucleotide length of the designed primer. For each additional nucleotide, primers become four times more specific. However, researchers have found that the longer the primer, the smaller the fraction of primed templates will occur in each annealing step, leading to a significant decrease in the volume of final product. Dieffenbach et al (1993) recommend between 18-24 nucleotide length primers.

Figure 13. Graph showing reduction in effective primer length due to degenerate bases.

Legend: The X and Y axes show possible numbers of two-way and three-way degenerate bases, with each cross-hair above either axis being a possible combination of marked amounts of both types. The Z axis takes into account the relative specificity levels of each combination with respect to reduced length; if a primer of ‘N’ length in base pairs possessed the combination of amounts of degenerate bases shown by the X and Y axis, its specificity would be reduced by ‘N−r’ where r is the number on the Z axis. Specificity refers to how specific a primer is, it is the opposite of sensitivity. Highly specific primers will only capture the intended target and thus may miss the target if there are any minor variables; highly sensitive primers will capture the intended target but may also capture a number of unwanted non-targets. The colours go from dark red (most specific) to pale yellow (least specific).

The position of the degenerate base effects efficiency not specificity (Jaric et al 2013) and this equation looks to specificity alone. It is considered useful to have at least three G or C bases at the 3’ end, which helps primers bind to their targets and therefore increases yield; in the same respect, degenerate bases grouped at the 3’ end should be avoided as this will lower
primer efficiency (http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html). Clearly this is only part of the design. Primers also must be designed with regard to primer-dimers, GC content, and melting and annealing temperatures as described in the following analysis. An investigation was made into the best methods for Tm calculation. The most commonly given Tm equation was: \( Tm = 4(G+C)+2(A+T) \). The primer manufacturer, Eurofins MWG, provided another equation:

\[
TM \ [^\circ C] = 69.3 + \frac{41(nG+nC)}{L} - \frac{650}{L}
\]

Where \( n \) = number of nucleosides of type G or C, and, \( L \) = number of all nucleotides per sequence (personal communication Wamack, 2013). To compare equations, both were reformulated as functions of the percentage GC content.

The crude equation became: \( TM \ [^\circ C] = 4(pL) + 2(L - pL) \)

The MWG equation became: \( TM \ [^\circ C] = 69.3 + \frac{41(pL)}{L} - \frac{650}{L} \)

Where, \( p \) = the percentage GC content expressed as a decimal e.g. 0.10 = 10% GC content, and, \( L \) = Nucleotide length of primer. A comparative graph was plotted (Figure 14). It shows that for the crude equation, the longer the primer becomes, the hotter Tm is predicted to be. This prediction is done in a linear fashion (given it is a linear equation) so that with a length off 23bp and 50% GC content, the predicted Tm is already equal to the normal Extension Temperature, which would make PCR impossible if it were correct. The graph also shows, by the nature of spaces between lines representing the MWG equation in its application to different primer length, that it allows for a logarithmic reduction in Tm with length.

The MWG equation was then plotted alone (Figure 15) to examine the relationship between GC content and length with respect to their effect on Tm according to this equation. We can see the equation has been formulated so that Tm never becomes as high as the extension temperature. In this graph it is not length that makes the biggest difference to Tm but GC content. For example at 50% GC content, the 16bp primer (red line) with 8 total G+C, and the 24bp primer (violet line) with 12 total GC, both have a Tm of 60\(^\circ\)c even though they are 33% to 50% different in length. Meanwhile a 10% difference in GC content, for example for a 20bp primer with 10 total GC vs a 20bp primer with 9 total GC the Tm changes from 57\(^\circ\)c to 55\(^\circ\)c. Neither of the above equations take into account the temporal aspect of primer design for PCR, in which one base pair must always first 'unzip' from its complement before the next in
a step wise fashion. Dieffenbach et al (1993) indicated that Tm prediction based on nearest neighbour thermodynamic parameters are slightly more accurate. 'Nearest Neighbour' is an expression of adjacent nucleotide interaction. This uses a boundary element method to calculate melting temperature taking into account that a G next to a T will be different from a G next to an A and so forth. A website was found that performed these calculations (http://www.thermoscientificbio.com/webtools/multipleprimer/).

Figure 14. Graph showing T<sub>m</sub> values given by the crude vs. MWG equation.

Legend: Full lines= MWG equation, dotted lines= Crude equation.
Gradient testing in lab found optimal Tm settings for successful PCR. A table of Tm results for all successful primers across all three equations compared to actual Tm results as found by gradient PCR tests is shown in Table 8. The mean of each primer pair was taken as the predicted balanced Tm for each equation and was shown as a graph (Figure 16.) The graph shows that the Nearest Neighbour and Crude equations come out with surprisingly similar estimates for each primer. Points were joined by lines on the graph, in spite of primers being unrelated, in order to capture the overall pattern of predictions. The MWG equations’ pattern was consistently in line with the Nearest Neighbour predicts but on average 5°C lower with little variance. Meanwhile, the actual successful Tm of each primer found by gradient PCR did not correspond to the pattern of any equation, veering up and down in Tm result; matching the MWG equation prediction only with primer set 4p2, and the crude equation with primer set 2F, and otherwise adhering to know particular set of predictions by any equation. The actual Tm must therefore be reliant on far wider parameters than any equation can predict, which starkly shows the importance of the Gradient PCR step.
Figure 16. Graph showing comparison of all three equations to actual lab-tested Tm.

**Legend:** Each reaction is independent of one another, thus dotted lines do not indicate a connection but simply allow the reader to more easily visualise the dispersal of each separate equation or actual Tm. Some data points overlay one another, the lines show what is otherwise obscured by data sharing the same point.

Table 8. Comparison of equations versus final Tm used in PCR

<table>
<thead>
<tr>
<th>Equations for $TM$ [°C] and balances between two primer results</th>
<th>Actual Tm</th>
</tr>
</thead>
<tbody>
<tr>
<td>4(G+C)+2(A+T) Mean $69.3 + \frac{41.0G + 6C}{L} - \frac{650}{L}$</td>
<td></td>
</tr>
<tr>
<td>2s Fwd 60</td>
<td>61</td>
</tr>
<tr>
<td>2s Rev 62</td>
<td>55.3</td>
</tr>
<tr>
<td>2x Fwd 68</td>
<td>65</td>
</tr>
<tr>
<td>2x Rev 62</td>
<td>57.9</td>
</tr>
<tr>
<td>2fin Fwd 68</td>
<td>64</td>
</tr>
<tr>
<td>2fin Rev 60</td>
<td>62.4</td>
</tr>
<tr>
<td>43p1 Fwd 76</td>
<td>69</td>
</tr>
<tr>
<td>43p1 Rev 62</td>
<td>57.9</td>
</tr>
<tr>
<td>43p2 Fwd 52</td>
<td>52</td>
</tr>
<tr>
<td>43p2 Rev 52</td>
<td>51.4</td>
</tr>
</tbody>
</table>
Once primer design was complete, primers were manufactured by MWG Eurofins. All badger blood samples were tested for capacity to amplify by using a mammalian tubulin PCR to confirm good quality DNA and to check for the presence of PCR inhibitors (Terry et al 2001). Figure 17 shows an example of a successful tubulin run on Woodchester samples 34K, 8i, 15K and 8L, as well as a negative (H2O) and positive control DNA. The positive control in column 6 is taken from a sample of cattle blood. Expected band size varies from mammalian species to species and its size in badgers has not been published. This band size is not easily predictable by bioinformatics because the tubulin genes are a family of genes. Band sizes vary across different mammalian species, however analysis of the gel in Figure 17 shows it is 1301bp. Table 9 shows band sizes vs distance travelled along the gel, this data was used to compare distance travelled along the gel with length in base pairs, and plotted in Figure 18. This gave a closest fit equation of $y = e^{(bx+c)}$, where $b = -0.00873$ and $c = 8.44589$. The mean error for hyperladder bands was 4.28%.

Figure 17. Example of a successful tubulin PCR across four badgers

![Figure 17](image)

Legend: Lane 1= 1K, 2= 2P, 3= 4K, 4= 5K, 5= Negative Control, 6= Positive Control

Table 9. Band sizes vs distance travelled along the gel for tubulin PCR.

<table>
<thead>
<tr>
<th>Hyperladder (bp)</th>
<th>200</th>
<th>400</th>
<th>600</th>
<th>800</th>
<th>1000</th>
<th>BT 1500</th>
<th>2000</th>
<th>2500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance (px)</td>
<td>360</td>
<td>287</td>
<td>236</td>
<td>197</td>
<td>170</td>
<td>146</td>
<td>126</td>
<td>98</td>
</tr>
<tr>
<td>Predicted length by closest fit (bp)</td>
<td>201</td>
<td>380</td>
<td>593</td>
<td>834</td>
<td>1055</td>
<td>1301</td>
<td>1550</td>
<td>1979</td>
</tr>
<tr>
<td>% Error</td>
<td>0.50</td>
<td>5.00</td>
<td>1.17</td>
<td>4.25</td>
<td>5.50</td>
<td>3.33</td>
<td>1.05</td>
<td>5.76</td>
</tr>
</tbody>
</table>

Legend: BT= Badger tubulin band
Column 6 has more than one band, which occurs in some but not all species; within the tandemly repeated genes of some species there are some repeats that are different lengths. All Woodchester samples tested positive for tubulin. Of the samples donated by Secret World 1-8, 11-14, and 16, 18 and 19 tested positive for tubulin. As was expected, samples 10 and 20 failed. Samples 9, 15 and 17 were trialled additional times and later run through all TLR PCRs but failed to produce any results. All samples donated by UCD tested positive.

Figure 18. Standard curve of band size vs distance travelled for tubulin PCR.

PCR optimisation was performed for all primer sets. TLR 2 was described by three of four redundant primer sets, 2s, 2x and 2 fin. 2p2 was later abandoned in favour of the other primers for reasons described in the following sections. Primer set 2s covered the first 1110bp of the single exon gene from the 5’ end. The reverse primer of 2s overlaps the forward primer of 2x; this gene set covers the middle 854bp of TLR2. Primer set 2fin overlaps the region covered by primer set 2x, and covers the final 822bp as well as 79bp of the intron. Primer set 2p2 overlapped the regions covered by the 5’ end of primer sets 2s and the 3’ end of primer set 2x. Final optimization for all TLR2 primers is discussed here.

According to nearest neighbour calculations primer 2s Fwd had a Tm of 64.5°C while 2s Rev had a Tm of 59.2°C. In order to find which temperature was best for amplification, gradient
PCRs were run between 52°C and 64°C. The strongest bands appeared at the 60°C to 62°C level. A problematic, faded non-target band (slightly larger than the desired band size of 1110bp) was removed by increasing the annealing time to 1:10 (m:s). Presence of this non-target band was further reduced by lowering the cycle time to 35 cycles.

Table 10. PCR amplification using primer 2s and predicted length using the Hyperladder

<table>
<thead>
<tr>
<th>Hyperladder</th>
<th>Distance</th>
<th>Pred.</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500 bp</td>
<td>791px</td>
<td>1501bp</td>
<td>0.67%</td>
</tr>
<tr>
<td>Band 2s Exp.</td>
<td>1070px</td>
<td>1066bp</td>
<td>3.60%</td>
</tr>
<tr>
<td>1110bp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000 bp</td>
<td>1113px</td>
<td>1011bp</td>
<td>1.10%</td>
</tr>
<tr>
<td>800 bp</td>
<td>1297px</td>
<td>807bp</td>
<td>0.88%</td>
</tr>
<tr>
<td>600 bp</td>
<td>1547px</td>
<td>594bp</td>
<td>1.00%</td>
</tr>
<tr>
<td>400 bp</td>
<td>1891px</td>
<td>389bp</td>
<td>2.75%</td>
</tr>
</tbody>
</table>

Legend: Lane 1= 3SP, 2= 12L, 3= 27K, 4= 1L, 0= Negative Control, Pred= Predicted Length by Closest Fit

The sample gel photo was measured by counting pixels. This data (Table 10) was used to compare distance travelled along the gel with length in base pairs, and plotted in Figure 19. This gave a closest fit equation of \( y = e^{(bx+c)} \), where \( b = -0.0013 \) and \( c = 8.3093 \). The amplified bands were estimated to be 1066 base pairs, 40 bp more than the desired product. The mean error for hyperladder bands was 1.28%.

Figure 19. Graph of distanced travelled vs. hyperladder size for 2s
According to nearest neighbour calculations primer 2p2 Fwd had a Tm of 61.4°C while 2p2 Rev had a Tm of 61.6°C. In order to find which Tm was best for amplification, gradient PCRs were run between 54°C and 64°C. The strongest bands appeared at 60°C. Using this primer, a non-target band appeared no matter what was done to optimise the reaction. To aid in analysis of this contamination, discussed later, the non-target band was isolated by increasing the MgCl₂ concentration during PCR. The sample gel photo shown in Table 11 was measured by counting pixels on a high resolution screen. It is clear, given the nature of bands in this gel that accuracy would be questionable, and errors were calculated. This data was used to compare distance travelled along the gel with length in base pairs, and plotted in Figure 20. This gave a closest fit equation of \( y = e^{(bx+c)} \), where \( b = -0.0012 \) and \( c = 8.3856 \). The amplified bands were estimated to be 915 base pairs, 21 bp less than the desired product. The mean error for hyperladder bands was 0.91%.

Table 11. PCR amplification using primer 2p2 and predicted length using the Hyperladder

<table>
<thead>
<tr>
<th>Hyperladder</th>
<th>Distance</th>
<th>Pred.</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 bp</td>
<td>1241px</td>
<td>1010bp</td>
<td>1.00%</td>
</tr>
<tr>
<td><strong>2p2 (Exp 936bp)</strong></td>
<td>1325px</td>
<td>915bp</td>
<td>2.24%</td>
</tr>
<tr>
<td>800 bp</td>
<td>1442px</td>
<td>797bp</td>
<td>0.38%</td>
</tr>
<tr>
<td>600 bp</td>
<td>1680px</td>
<td>601bp</td>
<td>0.17%</td>
</tr>
<tr>
<td>400 bp</td>
<td>2041px</td>
<td>392bp</td>
<td>2.00%</td>
</tr>
<tr>
<td><strong>Non-target</strong></td>
<td>2071px</td>
<td>379bp</td>
<td></td>
</tr>
<tr>
<td>200 bp</td>
<td>2601px</td>
<td>202bp</td>
<td>1.00%</td>
</tr>
</tbody>
</table>

Legend: Lane 1= 3K, 2= 13F, 3= 13N, 4= 15R, 0= Negative control, Pred= Predicted Length by Closest Fit.

The digital method for measuring distance travelled along the gel allowed for greater accuracy, however a clearer gel image would have been more desirable. At the time of experimentation equipment failure in the lab meant that electrophoresis equipment was running above set voltage, which heated gels and caused DNA to smear. Primer set 2p2 was later abandoned for reasons discussed in depth in the following pages, and returning to this primer set simply to produce a better image was not considered critical to completion of the necessary research given more successful redundant primers had been found.
Figure 20. Graph of distance travelled vs. hyperladder size for 2p2

The following describes the same working, done for the isolated non-target band in primer set 2p2. The sample gel photo was measured by counting pixels. This data (Table 12) was used to compare distance travelled along the gel with length in base pairs, and plotted in Figure 21. This gave a closest fit equation of $y = e^{(bx+c)}$, where $b = -0.0011$ and $c = 8.0944$. The amplified bands were estimated to be 354 base pairs. The mean error for hyperladder bands was 1.68%. The isolated band, when sequenced, came to 374bp; the area contaminated in target sequence 2p2 is 406 bp from the 255th nucleotide to the end. Thus, the estimate given by the first hyperladder is more accurate both because of its lower mean error and its closer match to the actual strand length measured.

Table 12. Example gel 2p2^MgCl and predicted length using Hyperladder

<table>
<thead>
<tr>
<th>Hyperladder</th>
<th>Distance</th>
<th>Predicted Length</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000bp</td>
<td>1085px</td>
<td>989bp</td>
<td>1.10%</td>
</tr>
<tr>
<td>800 bp</td>
<td>1254px</td>
<td>821bp</td>
<td>2.63%</td>
</tr>
<tr>
<td>600 bp</td>
<td>1537px</td>
<td>601bp</td>
<td>0.16%</td>
</tr>
<tr>
<td>400 bp</td>
<td>1933px</td>
<td>388bp</td>
<td>3.00%</td>
</tr>
<tr>
<td><strong>Non-target</strong></td>
<td>2015px</td>
<td>354bp</td>
<td>N/A</td>
</tr>
<tr>
<td>200 bp</td>
<td>2520px</td>
<td>203bp</td>
<td>1.50%</td>
</tr>
</tbody>
</table>

**Legend:** 1= 46P, 0= -ve Control
Figure 21. Graph of distanced travelled vs. hyperladder size for 2p2 non target

According to nearest neighbour calculations primer 2X Fwd had a Tm of 64.9°C while 2x Rev had a Tm of 65.2°C. In order to find which temperature was best for amplification, gradient PCRs were run between 54°C and 64°C. The strongest bands appeared at 56°C level. The sample gel photo was measured by counting pixels. This data (Table 13) was used to compare distance travelled along the gel with length in base pairs, and plotted in Figure 22.

Table 13. PCR amplification using primer 2X and predicted length using Hyperladder

<table>
<thead>
<tr>
<th>Hyperladder</th>
<th>Distance</th>
<th>Pred</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500 bp</td>
<td>728px</td>
<td>1377</td>
<td>8.13%</td>
</tr>
<tr>
<td>1000 bp</td>
<td>958px</td>
<td>1041</td>
<td>4.10%</td>
</tr>
<tr>
<td>800 bp</td>
<td>1136px</td>
<td>838</td>
<td>4.75%</td>
</tr>
<tr>
<td><strong>Band 2x Exp. 845bp</strong></td>
<td><strong>1216px</strong></td>
<td><strong>777</strong></td>
<td><strong>8.05%</strong></td>
</tr>
<tr>
<td>600 bp</td>
<td>1384px</td>
<td>619</td>
<td>3.17%</td>
</tr>
<tr>
<td>400 bp</td>
<td>1750px</td>
<td>397</td>
<td>0.75%</td>
</tr>
</tbody>
</table>

**Legend:** Lane 0= Negative Control, 1= 82W, 2= 26K, 3= 31L, 4= 2F, Pred= Predicted length by closest fit.

This gave a closest fit equation of $y = e^{(bx+c)}$, where $b = -0.0012$ and $c = 8.1143$. The amplified bands were estimated to be 777 base pairs, 68 less than the predicted 845bp. The mean
error for hyperladder bands was 4.18%; this is possibly due to the smeared nature of the hyperladder.

Figure 22. Graph of distanced travelled vs. hyperladder size for 2x

According to nearest neighbour calculations primer 2F Fwd had a Tm of 65.4°C while 2 fin Rev had a Tm of 66.0°C. In order to find which temperature was best for amplification, gradient PCRs were run from 56°C to 66°C. The strongest bands appeared at 64°C. A larger non-target was removed by increasing Ta to 1:10 (m:s) and reducing cycles to x35.

Table 14. PCR amplification using primer 2F and predicted lengths using the hyperladder

<table>
<thead>
<tr>
<th>Hyperladder</th>
<th>Distance in px</th>
<th>Pred.</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500 bp</td>
<td>620</td>
<td>1466</td>
<td>2.27%</td>
</tr>
<tr>
<td>1000 bp</td>
<td>953</td>
<td>998</td>
<td>0.20%</td>
</tr>
<tr>
<td>800 bp</td>
<td>1115</td>
<td>827</td>
<td>3.38%</td>
</tr>
<tr>
<td>Band 2f Exp. 743bp</td>
<td>1205</td>
<td>746</td>
<td>0.40%</td>
</tr>
<tr>
<td>600 bp</td>
<td>1389</td>
<td>603</td>
<td>0.50%</td>
</tr>
<tr>
<td>400 bp</td>
<td>1749</td>
<td>399</td>
<td>0.25%</td>
</tr>
</tbody>
</table>

Legend: Lane 0= Negative Control, 1= 58R, 2= 41L, 3 =40L, 4= 9P, Pred= Predicted length by closest fit.
The sample gel photo was measured by counting pixels. This data (Table 14) was used to compare distance travelled along the gel with length in base pairs, and plotted in Figure 23. This gave a closest fit equation of \( y = e^{(bx+c)} \), where \( b = -0.0012 \) and \( c = 8.0064 \). The amplified bands were estimated to be 746 base pairs, 3 more than the predicted 743bp. The mean error for hyperladder bands was 1.32%.

Figure 23. Graph of distance travelled vs. hyperladder size for 2F

![Graph of distance travelled vs. hyperladder size for 2F](image)

Most of Exon 3 of TLR4 was described by two primer sets (Figure 24). This section will go on to show what was done to successfully work with these two primer sets. Exon 2 of TLR4 was also sequenced in its entirety. According to nearest neighbour calculations primer 4p1 Fwd had a Tm of 65°C while 4p1 Rev had a Tm of 64.9°C. In order to find which temperature was best for amplification, gradient PCRs were run for all temperatures between 54°C and 66°C. The strongest bands appeared at 56°C. A larger non-target band was removed by increasing annealing time to 1:10 and reducing cycles to x35.
Figure 24. TLR4 Exon 3, as described by two successful primer sets.

The sample gel photo was measured by counting pixels. This data (Table 15) was used to compare distance travelled along the gel with length in base pairs, and plotted in Figure 25. This gave a closest fit equation of $y = e^{(bx+c)}$, where $b = -0.0012$ and $c = 8.3516$. The amplified bands were estimated to be 1146 base pairs, 36 more than the predicted 1110bp. The mean error for hyperladder bands was 1.79%.

Table 15. PCR amplification using primer 4p1 and predicted length using Hyperladder

<table>
<thead>
<tr>
<th>Hyperladder</th>
<th>Distance</th>
<th>Pred.</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500 bp</td>
<td>845</td>
<td>1508</td>
<td>5.33%</td>
</tr>
<tr>
<td>Band 4p1</td>
<td>1090</td>
<td>1146bp</td>
<td>3.24%</td>
</tr>
<tr>
<td>Exp. 1110bp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000 bp</td>
<td>1189</td>
<td>990</td>
<td>1.00%</td>
</tr>
<tr>
<td>800 bp</td>
<td>1352</td>
<td>812</td>
<td>1.50%</td>
</tr>
<tr>
<td>600 bp</td>
<td>1604</td>
<td>596</td>
<td>0.67%</td>
</tr>
<tr>
<td>400 bp</td>
<td>1942</td>
<td>395</td>
<td>1.25%</td>
</tr>
<tr>
<td>200 bp</td>
<td>2490</td>
<td>202</td>
<td>1.00%</td>
</tr>
</tbody>
</table>

Legend: Lane 1= 26K, Lane 2= 26N, Lane 3= 27K, Lane 4= 30L, Lane 0= Negative, Pred= Predicted Length

According to nearest neighbour calculations primer 4p2 Fwd had a Tm of 48.1°C while Primer 4p2 Rev had a Tm of 51.4°C. In order to find which temperature was best for amplification, gradient PCRs were run between 44°C and 54°C. The strongest bands appeared at 50 °C. The sample gel photo was measured by counting pixels. This data (Table 16) was used to compare distance travelled along the gel with length in base pairs, and plotted in Figure 26. This gave a closest fit equation of $y = e^{(bx+c)}$, where $b = -0.0127$ and $c = 4.9067$. The amplified bands were estimated to be 847 base pairs, just 9bp more than the desired product. The mean error for hyperladder bands was 1.37%.
Figure 25. Graph of distanced travelled vs. hyperladder size for 4p1

Table 16. PCR amplification using primer 4p2 and predicted length using Hyperladder

<table>
<thead>
<tr>
<th>Hyperladder</th>
<th>Distance</th>
<th>Pred.</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000bp</td>
<td>158 px</td>
<td>999bp</td>
<td>0.10%</td>
</tr>
<tr>
<td>Band 4p2</td>
<td>145px</td>
<td>847bp</td>
<td>1.07%</td>
</tr>
<tr>
<td>Exp 838bp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>800bp</td>
<td>142px</td>
<td>816bp</td>
<td>2.00%</td>
</tr>
<tr>
<td>600bp</td>
<td>117px</td>
<td>594bp</td>
<td>1.00%</td>
</tr>
<tr>
<td>400bp</td>
<td>84px</td>
<td>391bp</td>
<td>2.25%</td>
</tr>
<tr>
<td>200bp</td>
<td>32px</td>
<td>203bp</td>
<td>1.50%</td>
</tr>
</tbody>
</table>

Legend: Lane 1= 12L, 2= 13L, 3= 14L, 4= 14N, 0= Negative Control, Pred= Predicted length by closest fit.

Table 17. PCR amplification using primer 4e2 and predicted length using Hyperladder

<table>
<thead>
<tr>
<th>Hyperladder</th>
<th>Distance</th>
<th>Pred.</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500 bp</td>
<td>819px</td>
<td>1484bp</td>
<td>1.07%</td>
</tr>
<tr>
<td>1000 bp</td>
<td>1103px</td>
<td>1016bp</td>
<td>1.60%</td>
</tr>
<tr>
<td>800 bp</td>
<td>1267px</td>
<td>816bp</td>
<td>2.00%</td>
</tr>
<tr>
<td>600 bp</td>
<td>1504px</td>
<td>594bp</td>
<td>1.00%</td>
</tr>
<tr>
<td>400 bp</td>
<td>1825px</td>
<td>387bp</td>
<td>3.25%</td>
</tr>
<tr>
<td>200 bp</td>
<td>2306px</td>
<td>204bp</td>
<td>2.00%</td>
</tr>
<tr>
<td>4e2 (Exp 178bp)</td>
<td>2349px</td>
<td>192bp</td>
<td>7.87%</td>
</tr>
</tbody>
</table>

Legend: Lane 1= 9P, 2= 50N, 0= Negative Control, Pred= Predicted length by closest fit.

According to nearest neighbour calculations primer 4e2 Fwd had a Tm of 64.6°C while 4e2 Rev had a Tm of 60.9°C. In order to find which temperature held the best compromise for amplification gradient PCRs were run for all temperatures between 46°C and 64°C, before
some success was had at 48°C. Best results occurred when cycle timings were extended as much as possible, here Td:Ta:Te was 1:00;1:00;2:15 for 40 cycles. Optimal settings were not found and redesign of primers in this area proved even less successful. However three samples yielded good bands, giving 178bp amplicons, and were sequenced. The sample gel photo was measured by counting pixels. This data (Table 17) was used to compare distance travelled along the gel with length in base pairs, and plotted in Figure 27. This gave a closest fit equation of \( y = e^{(bx+c)} \), where \( b = -0.0013 \) and \( c = 8.3969 \). The amplified bands were estimated to be 192 base pairs, 14 more than the predicted 178bp. The mean error for hyperladder bands was 1.82%.

Figure 26. Graph of distanced travelled vs. hyperladder size for 4p2

![Graph of distanced travelled vs. hyperladder size for 4p2](image)

Figure 27. Graph of distanced travelled vs. hyperladder size for 4e2

![Graph of distanced travelled vs. hyperladder size for 4e2](image)
3.2 Sequencing and Analysis of TLRs

PCR products were frozen at -20°C before being sent for forward and reverse sequencing. Read outs were returned in the form of .abi graphics which were checked by eye for variants. Fifty-two samples from Woodchester Park, as well as nine outgroup samples provided by Secret World Wildlife Rescue, were sequenced across TLR2 in its entirety (Table 18).

Table 18. Sixty-one Badgers sequenced across ALL primers in TLR2.

| Outgroups | 1, 2, 3, 4, 6, 13, 14, 16, 19 |

A further three Woodchester badger samples 9N, 37K, and 38Y were sequenced across primers 2S and 2X only. Four additional SWWR badgers were partially sequenced: badger 12 across 2S and 2X, and badgers 5 and 18 were across 2X only. Badger 24 provided by University College Dublin was sequenced across 2S and 2X. Forty-nine samples from Woodchester Park, as well as ten outgroup samples provided by Secret World Wildlife Rescue were sequenced across all primers for TLR4 exon 3 (Table 19). A further eleven Woodchester samples were sequenced across primer 4p2 only: 3K, 13F, 13N, 15R, 17F, 22K, 28F, 28K, 31K, 32K, and 46P, as well as a further three SWWR samples: 1, 4 and 18.

Table 19. Fifty-nine Badgers sequenced for across ALL primers in TLR4 exon 3.

| Outgroups | 2, 3, 5, 6, 11, 12, 13, 14, 16, 19 |

Sequences were run through multiple sequence alignment to gain an overall consensus sequence for each primer set, and the .abi graphic of every sample was scanned by eye, for every nucleotide, in a search for variant or heterozygous bases. Lyons (2014) advised that heterozygotes are marked by the presence of both peaks with matching curves sitting at
roughly half the height of homozygous peaks, or where one peak is directly under the other but lower in height. There is a precedent for sequencing both strands, as heterozygous sequences can sometimes be missed on one strand and seen on the other. Taylor et al (1999) found a T>A heterozygosity which was called correctly in one direction, but missed on reverse strand, and an A>G heterozygosity missed on the forward stand, but called on the reverse strand. They found that due to signal weakening as the sequence progresses, heterozygotes were lost unless read from the complimentary strand. Error rates increase rapidly after 690 bases, due to weakened peak resolution. No variations were found using primer 2S, though noise in the data sometimes produced false positives that were proven not to be a variation through further trials. One of these was a C/G Heterozygous nucleotide in sample 3K which would have equated to a stop codon had its veracity been proven. This was the 146th base along TLR2. Several trials were performed to check its veracity (Figure 28.) A repeated trial always refers to where both repeated PCR and repeated sequencing in both directions has occurred. In this case a large amount of repeated trials occurred in order to ensure due diligence. Sequences were often returned with ‘noise’ so that any possible polymorphism was visible but questionable. A single retrial returning a polymorphism with low noise levels was not considered definitive, so trials were done until at least three produced sequences with noise low enough to determine that the polymorphism could not be dismissed as noise within noisy data.

Further sample testing revealed no variations across all samples for primer set 2s. Noise in other sample data was resolved by further PCR. For example, the first run of 4K was noisy and was resampled to show peaks identical in sequence to all other badgers tested (Figure 29.)

Figure 29. Two PCR samplings and sequencings of sample 4K
The first sampling of 24L produced a messy forward sequence. T appeared where a G was expected. The resampled data made it clear that the nucleotide was in fact a G (Figure 30). Early sampling of 66B gave a G peak underneath the expected A peak. Resampling found no G peak, showing that even in cleaner samples, noise can be an issue (figure 31). Other samples required multiple trials. The first and second trials of 82W were contaminated from the 909th nucleotide. The sequence identified as consensus across all other badgers can be seen clearly underneath the taller waves which graph the contamination (figure 32). A third and fourth sampling revealed the data points to be identical to the majority consensus data.
TLR2 primer set 2p2 was sequenced across only 13 Woodchester badgers; 1K, 7K, 8L, 9N, 12K, 14L, 24L, 25K, 26N, 32N, 38K, 63Y and 66B before sequencing issues led to a redesign of primers for this area. The forward strand in all 13 samples showed a perfectly 'clean' sequence with a single, evenly spaced peak at every nucleotide. However the reverse strands in all 13 began to be noisy at the exact same point, with increasing noise until the 3' end was reached. This suggested that there was a contaminating sequence, and this underlying sequence looked to be the same each time. It was hypothesised that the reverse primer may have captured a smaller region, perhaps a pseudogene. In order to solve this mystery, first, an attempt was made to decipher the contaminating sequence by eye. Noting underlying peaks the sequence in Figure 33 was reached. In an attempt to isolate this sequence, further PCR was performed using these primers. MgCl₂ loading was doubled and a single short band appeared. This band was the same size as the length of strand contaminating the target, and was sequenced (Figure 34).
Figure 32. The first two PCR samplings and sequencings of sample 82W

Figure 33. Primer 2p2 Underlying sequence called by eye
AAGTNCCGNNNGTGATNGAGCWCWNTNYNGRNTAATNTNGKNGTTTTANRRAWAAATTTTRMARGAGAGGNKKTANG
ATANAGGYATNTCTNNCTTNNNGCDANANNTTNCACNTNCRTKMTATNTNAGNAANCARNCKRAANNNNNAT
NAGANAAGGGAGRGGNTGCTGTAANANNAGAGNTGTGTYANGACAGCCCTTGGAGAARTGNCNCNTAAGRCAC
NNNMTATGCGTNNNNNCTTGANGNNNNGCAGGACANNNNCATGAAANTGCNTWGANAAGRGCCNGTG

Legend: To obtain this sequence the abi graphic was viewed by eye. All peaks had at least two bases and the smaller, or ‘underlying’, sequence peaks were noted. Where peaks could be one of a number of base calls a degenerate base value was used to indicate this.

Figure 34. Isolated band sequence (double MgCl₂ with 2p2 primers)
AAGTCCCGCTTGTGCCCGNCGTTCNCCNNNGGATAAATNNNCCTCAAGGAAAGAATACACCGGCCTCAACGTGAG
AAGGCTTGNCAATTCAGCTNCGACNNNCAACACNNNTGTAGNNATGGTTGCAAGCAAGCAGCATNGGGGGGAG
AAAGGAGGACGAATTCCCTGCTCTCAAAATTAGGTGTNTTTTATGCAGCAGACACGAAGTCTTGAGCAG
GCAGATCGCCACGAGANTCCAGACAGAGAGNTATGTGNNCAAGNGATATTCCGGAGAAAGGGCCNMACGGGA
GGCAANGAAGCCTTGAGCAAGAGCNGCTNNNGTCGNNNTNNCINNNNNNNNNNNNNNNNNNNNNNNNG

The MgCl₂ band and underlying sequence were matched closely via multiple sequence alignment (Figure 35). Efforts were made to understand where the contamination had appeared. It was found that the reverse complement of sequence 2p2 also matched the underlying sequence closely. It was postulated that the contaminating sequence was in fact a section of reverse strand read forwards. The sequence at the start of contamination was also closely complementary to the forward primer. This created a unpredicted pseudo dimer. This differs from a primer dimer which is simply the interation of either two primers attaching to one another, or a primer self-attaching. Instead the reverse strand of the amplified sequence from primer set 2p2, contained a central sequence which complemented the forward primer of 2p2, so that during PCR the forward primer was able to palindromically attach to both strands of the target sequence, creating an additional, shorter amplicon of one section of the target sequence.
Figure 35. First 120 bases of Multiple sequence alignment

Fwd Primer 5' CTGAGGAMAATTTAGGGTT

2p2RevStrand 3' CAGTCTCTTTTGAGGCA

MgCl2 band 3' AAGTCCCGCTTGTGAAGGCAC

underlying 3' AAGTCCCGNNNGTGATNGAGCWCTNTNYNGRNTAAATNGKNGNNTTTANRRNAAATTTRMA

U-match-M ******* *** *** *** *** *** *** *** *** ***

U-match-Rev ******* *** *** *** *** *** *** *** *** *** ***

Legend: U-match-M indicates, with asterixes, where the underlying contaminating sequence matches the isolated MgCl2 band. U-match-Rev indicates, with asterixes, where the underlying contaminating sequence matches the reverse strand of amplicon 2p2. "!!" indicate where the Fwd primer complements these sequences.

Figure 36. Resolution of primer set 2p2 self-contamination.

Alignment was strong for the first few bases; 42/60 bp of the underlying sequence match the MgCl2 isolated band. Twenty-nine of these match the reverse complement of the 2p2 amplicon. The source of the problem was the palindromic nature of the forward target strand, which in collaboration with the nearly complimentary sequence midway through the reverse strand, caused self contamination to occur (figure 36). The solution was to create primer sets 2X and 2F.

Two confirmed haplotype variations at one nucleotide position were found by primer set 2X. Outgroup badger 1 showed a homozygous T, and outgroup badger 14 showed a heterozygous C/T (Figure 37), for a base homozygous for C in all other badgers.
Just as with primer 2S, other samples suffered noise and were checked again by resampling. An example is sample 58R. On first sampling an A peak was seen in place of an expected G. Further sampling still showed the A close underneath the normal G peak on an otherwise clear graphic. However, this was still questionable; unlike the heterozygous nucleotide given by matching half-size peaks in badger 14, the 58R peaks were of similar height to peaks surrounding them. A third sampling showed that there was actually only a G at this point. These three trials are shown in Figure 38, clockwise from left to right.

In sample 3K there there seemed a very real possibility that a heterozygous variation had been found. Though the data was noisy, the C peak noted superimposed over the expected G peak, and was half the height of surrounding peaks. However a second trial found only the expected G peak (Figure 39). Negotiable amounts of noise in other samples meant that even where it seemed unlikely that a variation would be found, samples were repeated. In 37K there is clearly a large amount of noise. However, there appears to be a T and G curve simultaneously overlapping at a lower height than surrounding peaks where normally the sequence is G homozygous. A clean resampling shows that it is indeed homozygous for G as expected (Figure 40).

Primer set 2F bore one confirmed variation. Multiple samplings of badger 30L showed C/G heterozygous peak. Early trials were noisy but further PCR came up with clean data where the double peak remained (Figure 41).
Figure 38. Resolution of noise in sample 58R for primer 2X

Figure 39. Two PCR samplings and sequencings of sample 3K

Figure 40. Two PCR samplings and sequencings of sample 37K
Figure 41. Three sequencings of PCR samples from 30L showing C/G heterozygous base.

Figure 42. Two PCR samplings and sequencings of sample 22N

Figure 43. Two PCR samplings and sequencings of sample 1L
Noise in other samples was investigated and bases were proven to be none variant. Noise created the question of an A nucleotide insertion in the first sampling of 22N, and was shown to be noise which created a software based sequence misreading after resampling (Figure 42). Noise in sample 1L was resolved with further sampling (Figure 43).

By aligning sequenced sections end to end, a full sequence for TLR2 in the badger was obtained (Figure 45) which was then converted to an amino acid sequence using the function designed in Matlab (Figure 44). Variations found were examined to find their affect on the final protein. All variations were SNPs which are either transitions or transversions (Seabury et al. 2007). Transitions are like replaced by like; a purine replaced by a purine (A or G) or a pyrimidine by a pyrimidine (C or T). Transversions are like replaced by non-like; e.g. a purine replaced by a pyrimidine (Friedberg and Lawrence, 2001).

Figure 44. Amino Acid Sequence of Badger TLR2.

```
MSRVLWTVVLAVKLSKEEPDPQASSLSCDLTGVCDGRFRSLKSIPLSGLTAAVRSLDLSNNEITYIRNRLRG
CVNLKALKLASRINAIIEEDSFISLRSLEHLDSLNSLNSWFRFLSSLKFLNLLGNYKSLGEMPLFSPLT
NLQILKVGIDSFTELQKDFAGLSFLEELEIDASNLQRYPESLKSIGNISILMKRMQIPFLLIEIFGDLRSRL
KHELRDTHLNTFQSKEASIRENTNTIKWTRFNKVIDGSFSELVKNLNCVSGVEVFEGCTLDDGHNFDISD
MDKIKNIGGETLIVVRAIHFSYDMSSSYYSLAVNVRVTVESSKVFLVPCILLSQHLKSLLEYLDSDLNLVE
ESLRNSACQANPLLQTLILRHNRKSELEGLTTLLSLKNLTKLDISKNNYVSPMCTCQPDKLKYLNLNSNRIY
SVTRCIPIWMEILDISNNLDSDKLPIKVELIISGNKLTLPADSLPFTLRILRISRNISTFTKEQLDSTST
LEALEAGNNFCCSCFELSFTREQQSLAQILTDWPNDLCSDSFSSVRGQRVKDTVRLPAECHRVALVSVCVLF
LLILLGLVCLHHFGLWYLMWMWQLAQKRPKAPRPRDCYDAFVSYEHDSYVWNMMVQLEHFDPPFKLCL
HKDFIPIGRKWWIINDIIEIKSHKIFVLSFVENFVSENCKEYELDFSSHFLRDFENDNAILVLLFPIEKKAIQRF
CKLRKINMTKTYLENPTDETQEGFWLNLRAIKS
```

**Legend:** Pink highlights amino acids in consensus sequence which are changed by non-synonymous variation.

Sixty-one badgers were sequenced using all TLR2 primers covering all 2355bp of this single exon gene. Four haplotypes were found, all of which led to nonsynonymous mutations. Two percent of badgers from the Woodchester Park set showed variation from the majority consensus sequence, while in outgroup sets 22% of badgers showed TLR2 sequence variation. This refers to 1/52 badgers from the core sample set and 2/9 badgers from outgroups. A Fisher’s Exact Test was performed to resolve whether there was a significant difference between core and out group variation (Table 20).
Table 20. Fisher’s Exact Test on TLR2 variation levels in core and out group badgers.

<table>
<thead>
<tr>
<th></th>
<th>Woodchester</th>
<th>Outgroup</th>
<th>column sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variant</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Non-Variant</td>
<td>51</td>
<td>7</td>
<td>58</td>
</tr>
<tr>
<td>row sum</td>
<td>52</td>
<td>9</td>
<td>61</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td>0.054</td>
</tr>
</tbody>
</table>

The test found a marginally significant difference between the two groups; significant at the 90% level and just outside of the 95% significance level with a P value of 0.054. This may suggest variation due to Geographical Isolation by Distance (GID) which implies that a broader geographical study across badger TLR2 is in order.

Figure 45. Complete single exon TLR2 for the badger

ATGTCACGTGTTTTTGAGCGACGCTTGCTGAGCTACCTACGGGTCCCTCTCCCTGTTGGCTGCTGGAAATTTTAGATATTAGCAATAACAACCTCGATTTCCTTTTCCCTGATTTTGCCACGACTCAAAGAACTTTATATTTCCGGAAATAAGTTGAAGACCCTACCAGATGCCTCCTTCTTACCCAGTTACGCATTTGGAATACCGAGTTAGAATATCATTGACTCCATCGAGAAGAGCCACAAGACC

Legend: Red letters: Intron, Pink highlights: loci at which variation was found, Underlined: Base-pairs coding for amino acid codons affected by nucleotide variation.
A summary of badgers sampled is shown in Table 21. A haplotype network was created by aligning the 57 H1 sequences against the single H4 and linked H2 and H3 sequences, and is displayed in Figure 46.

Table 21. Summary of sixty-one badgers sequenced and their haplotypes for TLR2

<table>
<thead>
<tr>
<th>WP</th>
<th>H</th>
<th>WP</th>
<th>H</th>
<th>WP</th>
<th>H</th>
<th>WP</th>
<th>H</th>
<th>WP</th>
<th>H</th>
<th>SWWR</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1K</td>
<td>H1</td>
<td>8L</td>
<td>H1</td>
<td>17K</td>
<td>H1</td>
<td>26N</td>
<td>H1</td>
<td>36K</td>
<td>H1</td>
<td>50N</td>
<td>H1</td>
</tr>
<tr>
<td>1L</td>
<td>H1</td>
<td>9P</td>
<td>H1</td>
<td>17L</td>
<td>H1</td>
<td>27K</td>
<td>H1</td>
<td>38K</td>
<td>H1</td>
<td>58R</td>
<td>H1</td>
</tr>
<tr>
<td>2F</td>
<td>H1</td>
<td>12K</td>
<td>H1</td>
<td>21K</td>
<td>H1</td>
<td>28F</td>
<td>H1</td>
<td>38L</td>
<td>H1</td>
<td>59B</td>
<td>H1</td>
</tr>
<tr>
<td>2P</td>
<td>H1</td>
<td>12L</td>
<td>H1</td>
<td>22K</td>
<td>H1</td>
<td>30L</td>
<td>H1</td>
<td>39K</td>
<td>H1</td>
<td>63Y</td>
<td>H1</td>
</tr>
<tr>
<td>3K</td>
<td>H1</td>
<td>13L</td>
<td>H1</td>
<td>22N</td>
<td>H1</td>
<td>31L</td>
<td>H1</td>
<td>40K</td>
<td>H1</td>
<td>66B</td>
<td>H1</td>
</tr>
<tr>
<td>4K</td>
<td>H1</td>
<td>13N</td>
<td>H1</td>
<td>24L</td>
<td>H1</td>
<td>32N</td>
<td>H1</td>
<td>40L</td>
<td>H1</td>
<td>67I</td>
<td>H1</td>
</tr>
<tr>
<td>5K</td>
<td>H1</td>
<td>14L</td>
<td>H1</td>
<td>24P</td>
<td>H1</td>
<td>33Y</td>
<td>H1</td>
<td>41L</td>
<td>H1</td>
<td>82W</td>
<td>H1</td>
</tr>
<tr>
<td>7K</td>
<td>H1</td>
<td>14N</td>
<td>H1</td>
<td>25K</td>
<td>H1</td>
<td>34K</td>
<td>H1</td>
<td>45P</td>
<td>H1</td>
<td></td>
<td>H1</td>
</tr>
<tr>
<td>8I</td>
<td>H1</td>
<td>15K</td>
<td>H1</td>
<td>26K</td>
<td>H1</td>
<td>35P</td>
<td>H1</td>
<td>45R</td>
<td>H1</td>
<td></td>
<td>H1</td>
</tr>
</tbody>
</table>

Legend: H= Haplotype, WP= Woodchester Park, SWWR= Secret World Wildlife Rescue

Figure 46. Haplotype network showing low variation from majority consensus sequence

Legend: Circle size proportional to number of badgers per haplotype (H1=57, H4=1, linked haplotypes H2 and H3=2). Dark grey= Woodchester Park sample set, Pale grey= Outgroup sample set.

Haplotypes 2 and 3 (H2 and H3) are both transitions occurring at nucleotide 1007 which form a homozygote, and heterozygote, respectively, changing cytosine to thymine. This causes a missense mutation translating a threonine to an isoleucine at amino acid 336. Haplotype 4 (H4) is a transversion of guanine to cytosine at nucleotide 1722, leading to a missense mutation of glutamine to histidine at amino acid 574 (Table 22).
Table 22. Three variations found in TLR2

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Nt. Variation</th>
<th>SNP type</th>
<th>A.A. Variation</th>
<th>Mutation</th>
<th>Badger</th>
<th>% Pop.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WILD TYPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>95%</td>
</tr>
<tr>
<td>2</td>
<td>T HoZ 1007</td>
<td>Transition</td>
<td>Threonine (T)</td>
<td>Missense</td>
<td>1</td>
<td>1.6%</td>
</tr>
<tr>
<td></td>
<td>A &gt; A</td>
<td></td>
<td>&gt;Isoleucine (I)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>C/T HeZ 1007</td>
<td>Transition</td>
<td>Threonine (T)</td>
<td>Missense</td>
<td>14</td>
<td>1.6%</td>
</tr>
<tr>
<td></td>
<td>A &gt; A</td>
<td></td>
<td>&gt;Isoleucine (I)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>C/G HeZ 1722</td>
<td>Transversion</td>
<td>Glutamine (Q)</td>
<td>Missense</td>
<td>30L</td>
<td>1.6%</td>
</tr>
<tr>
<td></td>
<td>CAG &gt; CAC</td>
<td></td>
<td>&gt;Histidine (H)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Legend:** HoZ= Homozygote, HeZ= Heterozygote, number denotes position along nt or AA sequence.

H2 and H3 were seen in outgroup badgers located 73 miles apart. It is interesting to note that H2, the homozygous mutation of nt1007 was found in a badger with a sibling of identical age and gender (both 2.7kg males when found) and yet this sibling was homozygous for the wild type nucleotide at this point. These were both found in Locking, 44.3 miles southwest of Woodchester Park. Linked haplotype H3 was found in a badger found in Venn Ottery, 95 miles from Woodchester Park. This badger was sampled along with a cub from the same sett, but which bore the wildtype nucleotide at this point. H4, the unlinked haplotype, was found in a Woodchester Park badger, but was not shared by any member of its sett, nor of the setts in the area. Figure 47 shows locations where badgers with variable genes were found, but does not include the entire range of badgers sampled and found to bear the wild type from locations near Birmingham to St. Ives.

No variation was found across all samples for TLR4. In primer set 4p1, a sampling of 36K showed what looked to be an A/C heterozygote polymorphism. Peaks matched and were smaller than surrounding peaks. However, further PCR and sequencing showed that 36K was homozygous for C as with all other badgers (Figure 48). Sequences for almost all 4p1 primed PCR samples were very clear and showed no variation. It was only in some of the more degraded out group samples that noise occurred enough to create need for a second round of PCR. A low resolution sample of Badger 12 from Secret World Wildlife Rescue, had a questionable A peak in an area normally homozygous for C. A second trial had clear, sharp peaks, and proved there to be no variation (Figure 49).
Figure 47. Map showing locations where variable haplotypes were found.

Legend: WP=Woodchester Park (location of badger with H4), Lo= Locking (location of badger with H2), SM= Shepton Mallet (sample area, no variable haplotypes found), VO=Venn Ottery (location of badger with H3).

Figure 48. Two PCR samplings and sequencings of sample 36K

Figure 49. Anomalous peak in low resolution data resolved after a second PCR trial

Neither was there any variation found in primer 4p2. A possible heterozygous G/T nucleotide was seen in the first sampling of 24P. This sample underwent fresh PCR and was shown to in fact bare no such variation (Figure 50).
Figure 50. Questionable peak for 24P primer 4p2 resolved within two PCR trials.

Two possible variations were disproved in sample 36K; A/G peaks were seen where wild types are normally G homozygous, are seen at points 358 and 387 in the first trial, and at points 349 and 379 in second trial. Finally, a third trial found these points to be G homozygous once all noise was cleared up (points 357 and 386) (Figure 51).

Figure 51. Three trials to find no variation for 36K with primer 4p2

There were no variations found across all nucleotides sequenced, for all fifty-nine badgers tested over 1909 base pairs of TLR4 exon 3 which covers all of the LRR region. The primer set for TLR4 exon 2 was sequenced across only three Woodchester badgers; 9P, 13L and 50N. All three produced an identical sequence but no more were pursued due to extreme difficulty in
amplifying this area. By aligning sequence sections so that the end of one primer set was aligned with the beginning of the next, the known TLR4 sequence was arrived at (Table 23).

This was translated to a partial amino acid sequence using the function in MatLab (Table 24).

Table 23. Badger sequence for TLR4

<table>
<thead>
<tr>
<th>Exon</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 1 (96bp)</td>
<td></td>
</tr>
<tr>
<td>Exon 2 (165bp)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exon</th>
<th>Complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 3 (1909/2238 bp)</td>
<td></td>
</tr>
</tbody>
</table>

Table 24. Partial TLR4 Amino acid sequence for the badger

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPNITYKCMELNKLKNP1PTSKKLDSFLSNPLRLHNLGNSHFNSFPFLQVLDDLTS****TR1AYQGLHLSLLITLNGP1K</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSTEAFSGLSSLQLTLAVETNLRSLRLDLPIGHKSLKELVNHALLCSFSKLPYFPCNLTNLLELYDLDSNNN1KDYHSDLQVLH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>QMPMLNLSLNSNLPSNFQPGAPKEIKHELTLRSLONSTVDNKTCIQGLLKL1HHLVLGFCFKNERLSEFDKYLLEGLC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTEIEKFLAFNEFSEIDTDLPNCALNCTSTILMLFHLPKLNRLQWELVMECCEEFEPKWELELDKLFEVFTANKYS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFEKMLKESLEDFLDLSRNHLSFKCCSYESLGDGATRLKLHLDLSFMDIITMS5NFLGELQELYLDFFSNLQGFSDFVLSLRRN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRLYLDITHSHTVQVRPGFIFGDLVSSLQVLKMGANPGQFDFNLPHF1FKLDTNTILDLASKQLCLGVSQFFGSLFQKLQIMSSHN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSLDL1YPEFLSLSQ1LDSNFRIASTVEQVRHFFSNLVSMLTQDFCSFMCVEQHNLQWHKDRLELUVFMEKCTKFLD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MQMPLLSLRFINATCQRSKLITTSVSTFLVMVSLVAVYVKYFHLMLACGKYSRGKTSTYDFIVYIYQSMQEDWNRNELYKNL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEVVFQFQLCHYHRDFPGVIAAN11QEGF</td>
</tr>
</tbody>
</table>

**Legend:** The first 32 amino acids of exon 1 are unknown, the 55 amino acids of exon 2 are shown. The first 5 amino acids of exon 3 are unknown as are the final 105; the central 636 amino acids are shown.
3.3 Variation in relation to trypanosome Infection

Badger infection status was determined by Eze Justin Ideozu, using DNA extracted from the Woodchester Park sample group by Hadil Alkathiry by the phenol-chloroform method described earlier. PCR methods used are described in section 2.0. Twenty-nine out of eighty-two badger DNA samples amplified using ITS-Nested PCR producing band sizes of 1271 bp which indicated the badgers were positive for trypanosomes. This gave a prevalence of 35.4% (25.9% - 46.2%; 95% CI) (Ideozu et al unpublished). Raw data giving infected and non-infected individuals is shown in Table 25.

Table 25. Raw data showing badger infection status.

<table>
<thead>
<tr>
<th>Badger ID</th>
<th>Status</th>
<th>Badger ID</th>
<th>Status</th>
<th>Badger ID</th>
<th>Status</th>
<th>Badger ID</th>
<th>Status</th>
<th>Badger ID</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>24P</td>
<td>1</td>
<td>28K</td>
<td>1</td>
<td>12K</td>
<td>1</td>
<td>24L</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59B</td>
<td>0</td>
<td>2P</td>
<td>0</td>
<td>58R</td>
<td>1</td>
<td>67Y</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10N</td>
<td>0</td>
<td>15K</td>
<td>0</td>
<td>7K</td>
<td>0</td>
<td>21K</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13L</td>
<td>1</td>
<td>38K</td>
<td>0</td>
<td>38L</td>
<td>1</td>
<td>40K</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26K</td>
<td>1</td>
<td>50N</td>
<td>1</td>
<td>13N</td>
<td>0</td>
<td>5L</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3K</td>
<td>0</td>
<td>31K</td>
<td>1</td>
<td>66B</td>
<td>0</td>
<td>33K</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1K</td>
<td>0</td>
<td>25N</td>
<td>1</td>
<td>9N</td>
<td>1</td>
<td>23N</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17K</td>
<td>0</td>
<td>1L</td>
<td>0</td>
<td>9L</td>
<td>0</td>
<td>2K</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39K</td>
<td>0</td>
<td>40L</td>
<td>0</td>
<td>33Y</td>
<td>0</td>
<td>24K</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22K</td>
<td>0</td>
<td>31L</td>
<td>0</td>
<td>30L</td>
<td>0</td>
<td>29K</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17F</td>
<td>1</td>
<td>4K</td>
<td>1</td>
<td>37K</td>
<td>0</td>
<td>29N</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2F</td>
<td>0</td>
<td>14N</td>
<td>1</td>
<td>32K</td>
<td>0</td>
<td>30K</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36K</td>
<td>0</td>
<td>34K</td>
<td>0</td>
<td>23K</td>
<td>0</td>
<td>6L</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38Y</td>
<td>1</td>
<td>27K</td>
<td>0</td>
<td>41K</td>
<td>0</td>
<td>10K</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5K</td>
<td>0</td>
<td>35K</td>
<td>1</td>
<td>8L</td>
<td>0</td>
<td>11K</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12L</td>
<td>1</td>
<td>17L</td>
<td>1</td>
<td>41L</td>
<td>1</td>
<td>13K</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35P</td>
<td>0</td>
<td>28F</td>
<td>0</td>
<td>15L</td>
<td>1</td>
<td>9K</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>82W</td>
<td>0</td>
<td>8I</td>
<td>0</td>
<td>63Y</td>
<td>1</td>
<td>3N</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45P</td>
<td>1</td>
<td>25K</td>
<td>0</td>
<td>32N</td>
<td>0</td>
<td>22N</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26N</td>
<td>0</td>
<td>13F</td>
<td>1</td>
<td>45R</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9P</td>
<td>1</td>
<td>34P</td>
<td>0</td>
<td>14L</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total 8 9 7 5 29

Legend: 1=infected (shaded green for visual reference), 0=Not infected.

The Infection status and its relation to the badger’s movements were compared. Sixty-four badgers never left their home sett, while fifteen were found in a new location. A null hypothesis was formed, stating that infection status is the same for both static and mobile badgers. For this data it was not possible to use Chi-square test because the number of
mobile infected badgers is only four, which well is below the statistical threshold for Chi-square validity for one degree of freedom, which suggests that all elements in the table should be above ten. This is because Chi Square uses the normal distribution as an approximation for the binomial distribution which is valid when ‘n’ is large.

Because badgers are either infected or not, and mobile or not, we are dealing with dichotomous categorical variables that can have a value of ‘1’ or ‘0’ and thus Fisher’s Exact Test may be used. Unlike Chi-square it directly uses a combinatorial calculation that relates the probability to the number of successes in a set of ‘n’ events, and so is valid when dealing with very small numbers. Fisher’s Exact Test also takes account of the fact that for every ‘success’ the probability of further success is altered as proportions are changed. For example, when testing for infected badgers in a fixed population, the discovery of an infected badger is a ‘success’. This alters the probability of finding another as proportions of ‘successful’ badgers in the remaining, untested, population are lowered. This behaviour means that the hyper geometric distribution should be used as opposed to the binomial distribution. By using the hyper geometric distribution Fisher’s Exact Test allows us to examine whether our small data set has significant differences between mobile and static badgers regarding trypanosome infection. The number of infected vs. noninfected and mobile vs. static badgers gives a two by two contingency table (Table 26).

<table>
<thead>
<tr>
<th></th>
<th>Infected</th>
<th>Not infected</th>
<th>Column sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile Badgers</td>
<td>4</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>Static Badgers</td>
<td>25</td>
<td>39</td>
<td>64</td>
</tr>
<tr>
<td>Row sum</td>
<td>29</td>
<td>50</td>
<td>79</td>
</tr>
</tbody>
</table>

This can then be used to work out the exact probability that the observed data could happen by chance. The contingency table shows what was observed, and these observations indicate that infection rates are not the same; so the test asks the data whether the observed difference is significant.
Fisher’s Exact Test calculates the P value using the following equation:

\[ p = \frac{(a + b)! \ (c + d)! \ (a + c)! \ (b + d)!}{a! \ b! \ c! \ d! \ n!} \]

Where \( a \) = mobile infected badgers, \( b \) = mobile non-infected badgers, \( c \) = static infected badgers, \( d \) = static non-infected badgers and \( n \) = all badgers. When this is calculated in Excel the P value is 0.16. Thus there is 16% chance the data could have occurred randomly. Therefore, we must fail to reject the null hypothesis. Either, there is no significant relationship between infection status and badger movement, or, we do not have enough diversity of observations to find significance.

Of the fifty-two badgers which were fully sequenced across TLR2, only one (30L) held a variant haplotype (Table 27) Fifteen were infected by trypanosomes, while thirty-seven were not. 30L was among those not infected. Due to the low diversity of variant gene data it is not possible to perform a correlation which would show any significant relation between variation and infection.

Table 27. Fifty-two Woodchester Badgers sequenced across ALL primers in TLR2.

| Red highlight= Infected with Trypanosomes 15/52 |
| Black underlined (30L) = Single Woodchester badger with variant haplotype G1722C/Q574H |
3.4 Comparing Badger TLR Sequences with Those of Other Species

Variant haplotypes exist at two locations on TLR2. The first part of analysis involved plotting where in the protein these haplotypes are located, and what changes they make to the protein. Basic Local Alignment Search Tool (BLAST) software was used to compare the amino-acid sequences found to homologous proteins across species. A Position-Specific Iterated BLAST (PSI-BLAST) was chosen because it outputs all known proteins within the query sequence according to their position along the protein, so that their precise location along the TLR can be identified. Table 28 was created by zooming in on a BLAST output of TLR2 by eye to examine each amino acid one by one and identify where each known protein begins, ends or overlaps another. Haplotypes two and three move the position of a Leucine Rich Repeat named COG4886 (unknown function) from amino acids 330-494 to 359-535, effectively enlarging this domain by 12 amino acids. These same haplotypes also alter the position of an E3 Ubiquitin Protein Ligase named PRK15370 SirP from amino acids 334-559 to 392-538, shortening this domain by 79 proteins. Haplotype 4 adds a three additional amino acids to a Leucine Rich Repeat called SMART0082 in the C-terminal domain, which normally sits between amino acids 534-584, but in haplotype 4 sits between 534-587.

Table 28. Badger TLR2 with locations of related proteins highlighted

MSRVLWTVVWLGVATKLKEEPAEDQASSLSCDLTGVCDGRFRSLSIFGSGLTAAVRSLDLSNNEITYI RNRDLRCGVLKALKLASNRINAIEEDSFISLRSLEHLDLSYNLSSNLSSWFRPLSSLKFLNLLGNYKSLGEMPLFSPLTNQLKVGISDSTELQKDFAGLSLEELIDASNLQYEEPESLSQINISY ALRMKQPIFLEIFGDLRSLEHLDLELRTDHTLNTFQFSKASIRENTNTLIKWTFRNVKITDSGSELVK LLNCVSGLVEVEFCTLDGLNGFSMKPNGGNETLIVRRALIHPFYSYDDMSIIYSLFANVK RVTVESKVLCPFLSQQHLKSLEYLDLSNMLVEESLRNSACDAWPLLQTLIRHRNLKSLKTGE TLSSKLNTKLDISRSNNYVSMFTCQWPDKLKYLNLSNTRYSVTRCIPWMLELDISSLNDDSFLITLPRLKELYISGNKKTLPDasFPLPTLRILRISNIISTFTKELDSFSTLEALEAGGNFFCSCEFLLSFTREQSLSQILTDWPNDYNLCSDFSFSVRCVRKTDRLPASECHRVALVSCAVSLFLILLGTGCHFHGLWYLMMMWALQLAKRKRAPPDRDCYDAFVSYEHDSEYVENMVQEUHEDFDDDFKLPKLCDDHRDFIPGKIIDNIIIDSIEKSHKTIFVLSENFGKSEWCKYLEDFSHFRLFDFDENDAAILVLEPIEKKAPAQ RFCKLKRIMMTKTYLWEPTDDETQEGFVNLRLMAIKS

Legend: Four Leucine Rich Repeats (LRR) were identified; three LRR8 (shown by text coloured magenta, purple and red) and one LRRC (text colour lime). Two of the LRR8 proteins were found to overlap and this area is highlighted by green text. There is also an LRR multi domain, COG4886, highlighted in grey. The function of these Leucine Rich Repeats are as yet unknown. There is also one Toll-interleukin 1 receptor domain (TIR) 2 superfamily (blue text) known to function in signal transduction and is found at the cell surface interface.
A new search of TLR2 in the order Carnivora was made (September, 2014) and two new TLR2 sequences were found, *Ursus maritimus* and *Panthera tigris*. These, in addition to the TLR2 sequences collected at the beginning of the project, as well as badger haplotype 1 were collated for phylogenetic analysis. A clustal alignment of these sequences was performed to check whether the changes were in regions of mutability for other species. Haplotypes 2 and 3 refer to T>I variation at aa position 336, which at the nucleotide level are homozygous and heterozygous respectively. This was found to be a T across all Carnivores with no variation, so the local variation here is unusual (Figure 52.)

![Figure 52. Clustal Alignment showing amino acid position 336 in TLR2.](image)

**Legend:** Green highlight T amino acid for all Carnivores at posn 336.

Haplotype 4 is a Q>H heterozygous variation at position 574. It too shows no variation from Q across all Carnivores, and thus the local H variation is unusual (Figure 53.)

![Figure 53. Clustal Alignment showing amino acid position 574 in TLR2.](image)

**Legend:** Green highlight T amino acid for all Carnivores at posn 336.

In order to understand how TLR2 compares to a published phylogenetic tree of the order Carnivora, phylogenetic tree maker software was used. Prefered software changes quickly, and while PHYLIP was the product of choice until recently, today available programmes are no longer maintained. A current programme of choice is FigTree (Dudas and Rambaut, 2014; Bouckaert et al 2014), both as a graphical viewer of phylogenetic trees and a program for
producing publication-ready figures. This method estimates phylogenies from distance matrix data under the "additive tree model" where distances are equal to the sums of branch lengths between species. Clustal alignments for TLR2 were made by ClustalW Phylogeny, formatted with FigTree, and placed next to a published Carnivora phylogenetic tree for comparison (Figure 54). While our presentation of the tree used for comparison is a simplification of Nyakatura and Bininda-Emonds’ (2012) tree containing sixty animals (distances were divided down as distance increases with the amount of species included) it is noted that the tree using TLR2 sequences involves an originally limited array of carnivores as only this limited set of TLR2 sequences are currently available. Due to this the relative lengths of branches are artificial, and the following interpretations, though interesting, are merely putative and will benefit from further work into TLR sequence studies across other carnivore populations. Our comparison postulates that, for TLR2, the polar bear may have evolved far more rapidly than the panda; and, to a more subtle degree the tiger TLR2 shows a more rapid arms race than the cat; while the badger seems to have far more evolved TLR2 than the ferret; and the panda appears to have comparatively slow development in TLR2. PSI-BLAST was performed for all available carnivores in order to compare TLR2 configurations with *Meles meles*. Percentage similarities between TLRs across species were calculated (Cargill and Womack, 2007) (Figures 55 and 56).

Figure 54. Phylogenetic tree of Carnivora adapted from Nyakatura and Bininda-Emonds, (2012) vs. TLR2 Cladogram using distance matrix method.

Legend: Published tree left, TLR2 based tree right. Scale shows relative evolutionary distances from a common ancestor.
Figure 55. Badger TLR2, then ferret, panda, walrus and polar bear TLR2 which have 95%, 89%, 88% and 87% similarity to the badger.

(Two orange vertical lines roughly indicate the the position of variations found in this research.)

Legend: COG4886, CD00116, SMART00082, pfam12799, pfam13855 and PLN00113 are all Leucine Rich Repeat domains. TIGR00084 is a Polycystin Cation Channel protein. PRK15370 and PRK15387 are E3 Ubiquitin protein ligases. smart00255 and pfam01582 are in the TIR homology domain. Functions unknown.
Figure 56. Cat, tiger and dog TLR2 which have 86%, 85% and 83% similarity to the badger.

Legend: COG4886, CD00116, SMART00082, pfam12799, pfam13855 and PLN00113 are all Leucine Rich Repeat domains. TIGR00864 is a Polycystin Cation Channel protein. PRK15370 and PRK15387 are E3 Ubiquitin protein ligases. smart00255 and pfam01582 are in the TIR homology domain. Functions unknown.

The protein domains of each animal were compared in detail and it was found that while protein types were homologous across species, the numbers of homologous proteins varied between species (Table 29). These comprise Leucine Rich Repeat domains; COG4886, CD00116, SMART00082, pfam12799, pfam13855 and PLN00113: The Polycystin Cation Channel protein TIGR00864: The E3 Ubiquitin protein ligases PRK15370 and PRK15387 and smart00255 and pfam01582 of unknown function, which are found in the TIR homology domain. The biggest difference between TLR2 genes of different species is the total number of Leucine Rich repeat domains COG4886 and pfam13855, while the number of proteins located in the TIR homology domain were most similar between species (Figure 57). The discovery of polymorphisms within the LRR region is in line with other TLR studies. Cuscó et al 2014 listed locations of 31 non-synonymous SNPs found across all ten canine TLRs across 435 canids. Six were in or close to Leucine Rich Repeat regions, three were in or close to TIR domains, and the remaining twenty two were located in the sensor domains of the TLRs.
Table 29. Domain Hits for all Carnivores listed for TLR2

<table>
<thead>
<tr>
<th>Domain</th>
<th>Badger</th>
<th>Ferret</th>
<th>Panda</th>
<th>Walrus</th>
<th>Polar Bear</th>
<th>Cat</th>
<th>Tiger</th>
<th>Dog</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD00116</td>
<td>358-491</td>
<td>58-197</td>
<td>359-514</td>
<td>37-135</td>
<td>58-135</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend: One | Two | Three | Four | Five
In the domain column, proteins highlighted in red are Leucine Rich Repeats, all are of unknown function, however it is known that COG00116 belongs to a Ribonuclease Inhibitor-like subfamily, SMART00082 is found in the C-terminal domain, pfam12799 and pfam13855 are structurally different and these structures are identified as LRR4 and LRR8 respectively, and PLN00113 is an LRR-like protein kinase. The protein highlighted in green is a Polycystin Cation Channel protein, proteins highlighted in purple are E3 Ubiquitin protein ligases and proteins highlighted in orange belong to the TIR homology domain.

Figure 57. Leucine Rich Repeat density across species in TLR2

**Legend:** COG4886 is a Leucine Rich Repeat superfamily and pfam13855 is a Leucine Rich Repeat, which are most variable in their numbers found in TLR2 are most variable across species.
As with TLR2, phylogenetic analysis was also performed for TLR4 and compared to the phylogenetic tree for Carnivora (Nyakatura and Bininda-Emonds, 2012) (Figure 58). The walrus was excluded as its relative position in the TLR4 tree was extremely advanced in comparison to all other carnivores. Again, due to the low number of carnivore TLR4 sequences available, the relative lengths of branches in the TLR4 tree are artificial and all interpretations are putative and would benefit from further investigation into TLR4 sequences in other carnivores. The putative tree suggests that TLR4 phylogeny the cat is possibly more positively selected than the tiger, which contrasts with published phylogenetic relationships. It also postulates that phylogeny whereas the polar bear appeared to show positive in the TLR2 phylogeny, in TLR4 its position does not vary from the published tree. And, that the badger and ferret show no putative increase in positive selection on TLR4, which may explain why no variation was found in the badger population for TLR4 while variation was found for TLR2.

Figure 58. Phylogenetic tree of Carnivora adapted from Nyakatura and Bininda-Emonds, (2012) vs. TLR4 Cladogram using distance matrix method.

Legend: Published tree left, TLR4 based tree right. Scale shows relative evolutionary distances from a common ancestor

PSI-BLAST was performed for all available carnivores in order to compare TLR4 configurations with *Meles meles*. Percentage similarities between TLRs across species were calculated for the portion sequenced in the badger (Cargill and Womack, 2007) (Figures 59 and 60).
Figure 59. Badger TLR4, then ferret, polar bear, panda and dog TLR4 which have 95%, 91%, 90% and 84% similarity to the badger.

(Areas NOT yet sequenced in the badger are highlighted in orange)

Legend: COG4886, CD00116, SMART00082, pfam12799, pfam13855 and PLN00113 are all Leucine Rich Repeat domains. TIGR00864 is a Polycystin Cation Channel protein. PRK15370 and PRK15387 are E3 Ubiquitin protein ligases. smart00255 and pfam01582 are in the TIR homology domain. Functions unknown.
Figure 60. Tiger, cat and walrus TLR4 which have 84%, 83% and no similarity to the badger.

Legend for X through Y: COG4886, CD00116, SMART00082, pfam12799, pfam13855 and PLN00113 are all Leucine Rich Repeat domains. TIGR00864 is a Polycystin Cation Channel protein. PRK15370 and PRK15387 are E3 Ubiquitin protein ligases. smart00255 and pfam01582 are in the TIR homology domain. Functions unknown.

The protein domains of each animal were compared in detail and it was found that while protein types were homologous across species, the numbers of homologous proteins varied somewhat less between species than they had done in TLR2 (Table 30). The proteins found were the same as those found in TLR2. The TIR domain is the largest area not sequenced in this project and also appears to be the area where variation is least likely as it shows no significant variation across species. TLR4 shows greater homology than TLR2 across species. There is no variation in the amounts of protein COG4886 which was a highly variant factor across TLR2 in different species. In protein Pfam13855 only the felids differ from other carnivores in levels of this protein.
Table 30. Domain Hits for all Carnivores listed for TLR4

<table>
<thead>
<tr>
<th>Domain</th>
<th>Badger</th>
<th>Ferret</th>
<th>Polar Bear</th>
<th>Panda</th>
<th>Dog</th>
<th>Tiger</th>
<th>Cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>COG4886</td>
<td>44-190; 400-607</td>
<td>49-199; 400-607</td>
<td>44-190; 392-607</td>
<td>44-190; 391-605</td>
<td>43-341; 400-605</td>
<td>27-190; 373-579</td>
<td>27-190; 350-605</td>
</tr>
<tr>
<td>SMART00082</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pfam12799</td>
<td>150-190; 496-538</td>
<td>150-188; 496-538</td>
<td>150-190; 495-537</td>
<td>496-539</td>
<td>96-538; 152-190</td>
<td>150-190; 496-538</td>
<td></td>
</tr>
<tr>
<td>pfam13855</td>
<td>57-114; 126-187; 373-434; 422-483; 472-532;</td>
<td>57-114; 126-187; 373-434; 422-483; 472-532;</td>
<td>57-114; 103-197; 373-434; 422-483; 471-531; 421-482; 496-556;</td>
<td>57-114; 103-162; 373-434; 422-483; 471-531; 421-482; 496-556;</td>
<td>78-138; 99-458; 496-556;</td>
<td>79-138; 192-190; 399-458; 496-556;</td>
<td></td>
</tr>
<tr>
<td>PLN00113</td>
<td>47-532; 73-581</td>
<td>47-532; 73-581</td>
<td>48-532; 49-187</td>
<td>59-187; 44-532; 352-580</td>
<td>47-532; 47-532</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIGR00864</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRK15370</td>
<td>42-190</td>
<td>42-187</td>
<td>42-190; 42-187</td>
<td>42-190; 31-190</td>
<td>42-187</td>
<td>31-190</td>
<td></td>
</tr>
<tr>
<td>smart00255</td>
<td>674-816</td>
<td>674-816</td>
<td>674-816</td>
<td>674-816</td>
<td>674-816</td>
<td>674-816</td>
<td></td>
</tr>
<tr>
<td>pfam01582</td>
<td>677-815</td>
<td>677-815</td>
<td>677-815</td>
<td>677-815</td>
<td>677-815</td>
<td>677-815</td>
<td></td>
</tr>
</tbody>
</table>

Legend: Internal structure count
- One
- Two
- Four
- Five

In the domain column, proteins highlighted in red are Leucine Rich Repeats, all are of unknown function, however it is known that CD00116 belongs to a Ribonuclease Inhibitor-like subfamily, SMART00082 is found in the C-terminal domain, pfam12799 and pfam13855 are structurally different and these structures are identified as LRR4 and LRR8 respectively, and PLN00113 is an LRR-like protein kinase. The protein highlighted in green is a Polycystin Cation Channel protein, proteins highlighted in purple are E3 Ubiquitin protein ligases and proteins highlighted in orange belong to the TIR homology domain.
4.0 Discussion

DNA was extracted from the tissues of one hundred and six badgers, which were represented by eighty-two from Woodchester Park and twenty-eight from outgroup sources. Bioinformatic searches at the beginning and end of the study found no TLRs have yet been sequenced for the badger. Primers designed in this study successfully amplified TLR2 and TLR4 (exon 3) after PCR optimization. The single exon gene of TLR2 was sequenced in its entirety across sixty-one badgers, fifty-two from Woodchester Park and nine from outgroup sources. TLR4 exon 3 was sequenced across fifty-nine badgers; forty-nine from Woodchester Park and ten from outgroup sources. TLR2 sequencing showed three variant haplotypes of non-synonymous SNPs which produced missense mutations at the amino acid level. No DNA sequence variations were found in TLR4. No correlation was found between TLR variation and trypanosome infection status. Badger TLR2 was found to bear positive selection due to its higher ratio of nonsynonymous to synonymous mutations. When Badger TLR consensus sequences were compared to those of other species, phylogenetic comparisons showed that badger TLR2 was strongly selected for, while badger TLR4 showed little positive selection.

Early stages of the investigation revealed an absence of TLR4 (exon 3) DNA sequence variation across the first fifty-nine Woodchester Park badgers sampled, and no variation was found in early stages (first 32 badgers) sequenced across TLR2. Due to this it became clear that the results would preclude the analysis of any form of correlation with the 40% trypanosome infection found in this collection of badgers. In order to mitigate against the possibility of not being able to achieve the objective of investigating TLR variation with respect to trypanosome infection, outgroup badgers were sought to further explore TLR polymorphism. A variety of sources (veterinarians, rescue workers and academics) were contacted to investigate possible sources of specimens to act as geographical outgroups. Two groups responded, Secret World Wildlife Rescue (SWWR) and University College Dublin were able to donate samples. This could only occur from April onwards, in relation to the badger season and when cubs were two months old and able to given health checks. Furthermore, the beginning of the delayed 2013 badger cull determined when badgers were being rescued in southern England and the midlands. Thus parameters beyond the control of the study restricted the diversity of samples. Samples from SWWR ranged in all directions from
Woodchester Park, for up to 200 miles, which made it possible to rule out any inbreeding should low variation be found due to the geographical isolation by distance. Samples from Woodchester Park were given with a map detailing their sett locations and a spreadsheet details place of birth and the sett each badger was inhabiting at the time of sampling. It was interesting to note that most badgers never left their home sett. Extraction was highly successful when applied to the 400ul filled tubes of red blood cells donated by SWWR. However, the standard phenol-chloroform extraction method, used in this study may not be the best method for production of DNA from small quantities of white blood cells. Unfortunately, the small quantities of material and time constraints precluded further investigation of extraction methods on this occasion.

The bioinformatic search for TLR genes proved very fruitful. Due to the novel nature of TLR investigation, the number of available TLR gene sequences on web based databases nearly doubled during the course of the project, providing a substantial bank for comparison to the sequenced badger TLRs. However, even on the basis of the sparser amount of TLR information at the beginning of the study, primer design was very successful. Once exon regions were identified, Multiple Sequence Alignments (MSAs) allowed the conserved regions across species to be noted. From these regions the first primers were selected. It was only after some sequencing had been undertaken that it was realised how very close badger and ferret sequences were. After this it was possible to use the ferret TLR sequence as a reference when making MSAs to specifically select conserved regions were the least variance was found between the ferret and all other creatures. Investigations into primer design proved very useful in selecting the optimal parameters to target specific sequences. Primer length was found to improve specificity more than any other variable, and detection of possible primer dimers was also a key parameter. Balanced temperature settings were found to be of relatively low import, it was also found that no known equation, including nearest neighbour evaluations, can accurately predict optimal temperature settings better than performing a gradient PCR in lab. Adding degenerate bases to code for any of two or three nucleotides at one base position was found to have little relative impact on specificity, and so could be embarked upon with the confidence of knowing that countering such choices with increased primer length would prevent the possibility of amplifying non-target regions.
Defining the most important parameters of primer design in this way lead to cost- and time-efficient creation of successful primers.

During PCR optimisation MgCl₂ gradients were performed as suggested in the Stratagene Robocycler manual. Following gradient PCR a handful of samples still did not amplify at optimum temperatures which successfully amplified TLRs in all other samples, or would produce additional bands not seen in the majority of samples. Where non-target bands were smaller than the target band size MgCl₂ gradients were run where MgCl₂ levels were incrementally decreased in proportion to the master mix until the larger band appeared alone. Where additional bands appeared that were larger than the target amplicon, MgCl₂ levels were incrementally increased in proportion to the master mix until the smaller target band appeared alone. This was especially useful when attempting to isolate and identify the contaminating band from primer set 2p2. In this case the band was isolated when MgCl₂ loading was doubled. It was also found that increasing cycles allowed more amplicon to be produced, this helped maintain strong bands were annealing time was reduced.

As well as being automatically analysed by FinchTV software, the data were also examined manually. All variations and noise were confirmed or retested via additional rounds of PCR and sequencing. TLR4 showed no sequence variation across both Woodchester and outgroup badgers. However, non-synonymous variation which caused missense mutations at the amino acid level was found in TLR2. Two percent of badgers from the Woodchester Park set showed variation from the majority consensus sequence, while in outgroup sets twenty-two percent of badgers showed TLR2 sequence variation, this may point to Geographical Isolation by Distance (GID). Analyses of parasite burden in Woodchester park found no significant correlation between place of birth and infection, nor any significant variation across setts; disease was fairly evenly distributed across the whole of Woodchester Park with a mean 37.2% of badgers infected. There were too few roaming badgers to make any significant correlation between roaming and disease susceptibility. As well as finding non-synonymous mutations that indicated positive selection in TLR2 but not TLR4 in the population sampled in this study; phylogenetic tree comparisons of these genes to those of other species showed the consensus badger TLR2 gene positioned much further along the distance matrix in comparison to other species than the badger TLR4 gene, inferring stronger evolutionary
selection on badger TLR2. This suggests that badger TLR2 is being driven to such selection by an evolutionary arms race against an as yet unidentified pathogen or parasite, and that further investigation of badger TLR2 in relation to key badger diseases would be useful. TLR population studies are a recent phenomena and data is sparse. In human studies the number of possible haplotypes and total gene variance is often discarded in favour of investigating less rare, known, polymorphisms. In a study of 640 volunteers from Romania and the Netherlands Ioana et al (2012) focused on three TLR2 polymorphisms and found consistent levels of their appearance. For all three a variant homozygous nucleotide was rare, comprising from none to 0.16% of the population. The heterozygous variations found were at levels of 0.78%, 4.53% and 6.09% in populations. These results are very similar to our own findings for TLR2, with respect to low amounts of variation. Across all badgers in this study, heterozygous variation occurred in 1.6% of the total population and homozygous variation in 3.2% of the population. Similarly, in volunteer populations of 580 and 905, respectively, Mockenhaupt et al (2006) and Reismann (2009) found 1.21%, 3.45% and 0.11% had homozygous variation, while between 4.83% 22.24% and 11.49% had heterozygous variation, across three pre-selected nucleotides on TLR4. In contrast to this, badgers in this study were found to have absolutely no variation across TLR4 gene sequences. Given the populations studied above are over ten times larger than our study population, and yet yield some, low, variation in TLR4, this may point to a need to continue the TLR4 across a wider sample range.

Hashemi-Shahri et al (2014) examined known TLR8 variations in humans and found that 27% of a population of 175 had a homozygous variation, while 29% had a heterozygous variation. Liu et al (2012) examined a known TLR9 variation and found that 19% of a population of 432 had a homozygous variation, while 33% had a heterozygous variation. Zhang et al (2014) examined four known TLR9 variations and found that on average across all four variations examined 13% of a population of 854 had a homozygous variation, while 26% had a heterozygous variation. Carvalho et al (2007) assayed known polymorphisms across 388 Portuguese blood donors. 11.1% and 10.8% were heterozygous for TLR4 Asp299Gly and Thr399Ile (TLR4) while 0% were homozygous for either. A TLR9 variant allele (T-1237C) was present in 17.3% as heterozygous mutation and 2.1% as homozygous. Martínez-Ríos et al (2013) studied two known polymorphisms within 740 Mexican volunteers. They found that for TLR-4 Asp299Gly, the AA homozygote occurred in 93% of the population, the AG
heterozygote in 7% and a GG homozygote never occurred; and that for TLR-4 Thr399Ile, the CC homozygote occurred in 97% of the population, the CT heterozygote in 3% and a TT homozygote never occurred.

These studies, taken alone, possibly indicates how unusual our findings were in terms of low levels of TLR sequence variation across a population. However, other studies of known polymorphisms found equal, or even less, variant polymorphisms. Kim et al (2012) investigated six known TLR polymorphisms; TLR1 (Arg80Thr), TLR2 (Arg753Gln and Arg677Trp), TLR4 (Asp299Gly and Thr399Ile), TLR6 (Ser249Pro) across 322 Korean volunteers and found no polymorphisms between them. Rodriguez-Osorio et al (2013) found unusually low variation in two known SNPs, which was ascribed to the ethnicity of the population tested. Of 170 Mexicans, only 5 were heterozygous for both TLR4 D299G and TLR4 T399I polymorphisms. Emingil et al (2007) found an unusually low amount of known polymorphisms in 245 Turkish subjects. For SNPs TLR2 Arg753Gln, TLR4 Asp299Gly and TLR4 Thr399Ile they found 10.1%:0%, 4.5%:0.3% and 3.4%:0% homozygous: heterozygous variation respectively. Weng et al (2014) looked at population variation in ten known TLR4 polymorphisms and found that a maximum of >2% of the 11,319 people studied bore variant for any SNP studied.

Some studies went further, looking for all polymorphisms in a single gene fragment, but often failed to report the distribution of these genes within the population studied. For TLR2 and TLR4 Mukherjee et al (2014) found eight and nine non-synonymous (NS) mutations in a population of 266 people, while Grueber et al (2012) found one and four NS mutations in a smaller study of 23 Petroica australis takiura. Charbonnel et al (2014) found ten NS mutations for TLR4 across 300 Myodes glareolus. Abrantes et al (2013) reported on the genetic diversity of TLR3 across 80 European rabbits from Portugal, France and Spain, where they found 14 non-synonymous variations mostly in the LRR region except for only one in the TIR domain. Morger et al (2014) performed a population study comparable to our own study of sets across Woodchester Park, examining woodmice collected from sites 500m apart around a quadrant roughly 6750m² near Malham Tarn. This looked at variation in TLR11 and TLR12, and found six and thirteen haplotypes, bearing respectively four and nine nonsynonymous mutations, respectively, across 120 Apodemus Sylvaticus. Half that number of badgers, from
Woodchester Park, were examined in this study, in an area roughly equal to the Malham study, but only two TLR2 haplotypes were found in this area; the wild type consensus sequence borne by fifty-eight badgers sequence, and haplotype 2, an SNP leading to one nonsynonymous mutation. No variation was found for TLR4 across these badgers. Additional variation found due to outgroup sampling. Thus our findings show unusually low variation within a core sample population of animals in a geographical range where contact between them is not heeded by landscape barriers.

Figure 61. SNPs in bovine TLRs from Seabury et al (2007) and Cargill and Womack (2007).

Seabury et al (2007) found 92 SNPs and 6 indels, across three TLRs in 10 cow breeds, 45 of which were located in intron areas. Cargill and Womack (2007) found 130 SNPs (97 of which were in introns) and 9 indels across four TLRs in the same range of cows. This study focused only on the exons of TLR5, however looking at the exon data alone there was still a greater range of variations found in the cow studies; a mean of 15.6 per TLR in the Seabury study and 8.25 per TLR in the Cargill and Womack study. The data is still not entirely comparable as in each of the above studies, two cow species were sampled: Bos taurus, Bos indicus as well as eight breeds derived from them; Angus, Charolais, Holstein, Limousin, Brahman, Nelore, Braford, Piedmontese and Romagnola, whereas a single species with no hybrids were
considered in this study. Geographical distribution analysis was not possible as these breeds cover an international range. Given this, the higher level of variation found may well be a factor of the amount of breeds and the massive geographical distribution. Further analysis of the data (Figure 61) presented in both papers showed that the level of variations found did not correlate to the amount of length of fragments analysed, thus variation is not evenly spread across the TLRs and some (TLR10 in cattle especially) are more positively selected than others. Unfortunately, neither paper looked at TLR2 or TLR4, as comparative levels of variation here would have been fascinating in relation to this study.

To date, only Tschirren et al (2014) have produced a study which investigates both the range of haplotypes and their distribution within a population: Ten out of fifteen haplotypes were found to be nonsynonymous across a fragment of 49.87% of the TLR2 exon (1173/2352bp) were shared across 726 Myodes glareolus, in a 0.25km² area of Kalvs Moss, distributed across 292, 203, 50, 25, 24, 21, 20, 14, 2 and 1 M. glareolus respectively. Compared to our own, over ten times as many animals were sampled in this study, but just over three times as many nonsynonymous haplotypes were found. Thus, just as with our study variation was relatively low, and the comparison suggests that further investigation of badgers could well expand our current repertoire of TLR sequence variation.

TLR variation is important in disease. Studies reveal that TLR polymorphisms lead to a change in the degree of innate responsiveness, either increasing or decreasing the inflammatory response can be both destructive or protective depending on the context of specific diseases tested. TLR polymorphisms have been found to be significantly related to both disease onset risk and prevention. In patients with pre-existing disease, where polymorphism prevalences were not significantly different to healthy controls, instead significant links were often found between either the diseases’ development or protection against its development.

The following studies all found polymorphisms which bestowed a significantly increased risk of disease onset. A great deal of work focused on TLR2 and TLR4, just as in our study, but was limited to testing previously known polymorphisms. ALL disease onset associations listed here are ‘significant’. TLR4 Asp299Gly, TLR4 Thr399Ile, and TLR2 Arg753Gln were associated with early-onset and severe preeclampsia in a study of 270 pregnant women (Xie et al 2010).
TLR4 Asp299Gly was also significantly higher among 137 scrub typhus patients versus 134 controls from South India (Janardhanan et al 2013). TLR2 indel (196 to 174) was associated with risk of gall bladder cancer, and, non-synonymous TLR4 Exon 4 1936C>T was associated with cancer risk in females with gallstones in a study of 490 people from India (Srivastava et al 2010). TLR2 rs893629 was associated with arterial thrombosis in a study of 3587 North American systemic lupus erythrmus patients (Kaiser et al 2014).

The prevalence of this, like a great number of disease-significant SNPs, proved to be dependent on ethnicity. TLR2 rs893629 was more common in African American patients (9% prevalence) than in white or hispanic patients (less than 2%). TLR2 (P631H) and TLR6 (P249S) were associated with increased susceptibility to skin infection across a study of 646 caucasians. TLR1 R80T resulted in lower IL-6 cytokine responses to Staph aureus infection (Stappers et al 2014). TLR9 1237-CC was significantly linked with cerebral malaria in 117 children studied in Uganda (Sam-Agudu et al 2010). TLR5 C1174T encodes a variant that destroys flagellin-induced signaling, and is significantly associated with increased risk recurrent cystitis (Hawn et al 2009). In a study of 358 Holstein cows four TLR9 polymorphisms (A945G, G1187A, G1401A, and C2788T), increased risks of metritis occurrence, while TLR2 polymorphism C9564T increased risk of cytologic endometritis (Pinedo et al 2013).

Further studies showed that relationships between TLRs and disease onset are not always 1:1; multiple polymorphisms can combine to promote disease where one polymorphism may have no significant affect. Sampath et al (2013) found that when two TLR4 SNPS coexisted (rs4986790 and rs4986791) there was significantly increased risk of Gram-negative infections in the 408 low birth-weight babies tested. Liao et al (2010) found that if two TLR9 gene polymorphisms (rs287084 and rs352140) occurred in tandem, they significantly increased risk of Graves’ ophthalmopathy in males only, in a study of 471 Taiwanese volunteers. This study points to another issue when analysing TLR polymorphisms and their contribution to disease; many studies have found gender dimorphic affect. The reasons for this are not elucidated for TLR9, but are more clear cut for TLR7 and TLR8, which, in humans, are located on the X chromosome. Alagarasu et al (2015) found a significantly higher frequency of TLR8 rs3764879–rs3764880 haplotype C-A was in males with dengue hemorrhagic fever compared to healthy controls, but no such relationship occurred in females. Thus TLR analysis is a
multifaceted field which must take into account race, gender, and interacting effects of multiple polymorphisms. For example, eleven TLR polymorphisms, across 276 volunteers, were both significantly associated with HIV status and were also race specific: Nine SNPs in TLR1, TLR4, TLR6 and TLR8 in Caucasians, and two other SNPs, in TLR4 and TLR8, in African Americans (Willie et al 2014). Seventeen SNPs across TLRs 1 and 2, TLR4, and TLRs 6 through 10, were linked with allergic rhinitis and asthma. Prevalence of each SNP varied significantly between ethnicities (Gao et al 2010).

Conversely, the following studies all reveal polymorphisms which bestow a significantly increased protection against disease onset. Again, ALL associations listed here are significant. TLR2 Arg677Trp was associated with a lower risk of cytomegalovirus infection in adults but not infants in a study of 229 volunteers (Jablonska et al 2014). TLR4 (Asp299Gly) significantly protected those with chronic gastritis from reaching the active ulcerative stage among 195 patients infected with Helicobacter pylori (Bagheri et al 2014). TLR4 T610C and T610C and TLR6 G14578A, produce lower risk of clinical endometritis (Pinedo et al 2013). TLR4 A896G was associated with protection from recurrent cystitis while TLR1 G1805T was associated with protection from pyelonephritis in a study of 987 menstruating caucasian women (Hawn et al 2009). TLR1 rs5743551-CC was associated with reduced risk of disease in 300 healthy controls compared to 702 Russian patients with coronary artery disease (Golovkin et al 2014). TLR3 rs3775296 formed significant resistance to photosensitivity and anemia compared to wild type carriers (Wang et al 2014). TLR6 Pro249Ser showed a significantly reduced risk for atherosclerosis in a study of 503 heart disease patients and 605 healthy controls (Hamann et al 2013). TLR6 rs3775073-CC was significantly associated with decreased risk of infective endocarditis in a study of 410 Russian caucasians (Golovkin et al 2015).

Where TLR polymorphisms were not found to be significantly related to either disease onset or its prevention, they were often found to play a significant role in its development in patients with disease. The following TLR polymorphisms are significantly related to disease development from early to later stages once a disease has already taken hold. ALL associations listed here are significant. Spelatas et al (2009) associated TLR4 T399I with a 2.4-fold increased risk of COPD development, in a study of 240 smokers. Nachtigall et al (2014) found that TLR2 Arg753Gln and TLR4 Asp299Gly were related to progression from sepsis to
septic shock in 145 critically ill patients. Oliveira et al (2013) found that TLR4 rs4986790 and TLR4 rs4986791 were associated with presence of autoimmune thyroiditis in 572 bipolar disorder patients vs. 202 healthy controls. Pine et al (2009) studied 201 acute-stage caucasian HIV patients from Seattle and found that TLR4 D299G and TLR4 T399I were significantly more frequent, but TLR9 1635G was significantly less frequent, among individuals with high viral loading. Zidi et al (2014) studied 130 cervical cancer patients and 200 healthy controls in Tunisia, and found that TLR3 rs3775290-CC and TLR4 rs4986790-CC are significantly associated with higher risk of developing cervical cancer following herpes infection. Matas-Cobos et al (2014) found that TLR3 rs3775291 and TLR6 rs5743795 play a significant role in disease development to severe and acute stages (respectively) across a study of 260 pancreatitis patients. Apinjoh et al (2013) found that TLR9 rs187084 was significantly associated with susceptibility to malaria, while a TLR1 rs4833095 significantly weakened bodily defences against hyperparasitaemia; both contributed to the onset of severe malaria, in a study of 971 children with malaria and 891 healthy volunteers from the Cameroon. Yang et al (2012) found a significant association between TLR9 1237T/C and increased risk of chronic kidney disease, and higher mean plasma IL-6 levels in 630 end stage renal disease patients from the Han Chinese population.

Combinations of TLR polymorphisms also acted to speed disease development where single polymorphisms had not weakened the immune system. The following combinations made a significant difference in hastening disease development. Peric et al (2014) found that while single polymorphisms taken alone showed no significant correlation to chronic Hepatitis C, if taken in combination, then, a polymorphic heterozygous TLR4 Asp299Gly together with a homozygous TLR4 Thr399Ile correlated with significantly higher viral loading. Wang et al (2014) found that intron based TLR7 SNP rs3853839-G>C played a significant role in systemic lupus erythematosus development in a study of 1957 females. Another non-synonymous polymorphism TLR8 rs3764880-G>C, based in an exon, was associated with oral ulcers. These two SNPs acting in combination were associated with pericardial effusion.

Just as some polymorphisms hasten disease progression, others protect against it. ALL of the following TLR polymorphisms confer significant protection against disease. Alagarasu et al (2015) sampled 120 dengue cases and 109 healthy controls from Maharashtra, Western
India, and found significantly lower frequencies of the TLR3 rs3775291 T allele in those with dengue hemorrhagic fever. This was associated with damage to the TLR resulting in a decreased inflammatory response which had a protective effect against disease progression. Devaraju et al (2014) studied a TLR9 (1237C/T) polymorphism and found that a wild type C allele conferred significant risk of disease development across 300 systemic lupus erythematous patients, while a polymorphic T allele was found in a significant proportion of the 460 healthy controls. Also, just as while some polymorphisms had limited affect alone, polymorphisms also work in tandem to prevent disease progression. ALL of the following TLR polymorphic combinations had significant affect in the prevention of disease development.

A combined heterozygous TLR4 D299G/T399I was protective against heart disease among 125 Chagas Disease patients in northern Chile (Weitzel et al 2012). Holla et al (2010) found that 222 patients with chronic periodontitis bore different TLR9 haplotypes with a complex range of SNPs to the 259 unrelated controls studied in Czechoslovakia. No significant relationship was found between individual SNPs and this disease but analysis showed that the SNPs worked in combination to alter disease outcome. Castaño-Rodríguez et al (2014) found that a combination of TLR4 polymorphisms (rs10759931, rs1927911 and rs10116253) were protected against progression to gastric cancer in 310 Chinese patients infected with Helicobacter pylori.

As we have seen above some of the known TLR polymorphisms tested were found to be either detrimental or protective against disease and its progression. A study by Liadaki et al (2011) makes it clear that even for specific polymorphisms, results are not clear cut. In a study of two nonsynonymous TLR4 SNPs across 327 tonsillectomy patients with a history of tonsillitis, Liadaki et al found that TLR4 polymorphisms D299G and T399I predispose individuals to streptococcus infection on a three-fold scale compared to 245 healthy bone marrow donors, while the former polymorphism was protective against Haemophilus influenzae by two-fold.

As has already been discussed in detail in section “1.1.1 Badgers and their role as Reservoirs for Infection,” badgers are implicated in bovine TB. They, alongside the brushtail possum in New Zealand, the African buffalo in South Africa, the white-tailed deer in the USA, and the European wild boar in Spain are all considered maintenance hosts (Le Roex et al 2013).
Current cull-based strategies to intervene with this are complex, controversial and generate polarised public feeling. Badger vaccination trials have shown success, but receive little financial support and no government backing. Current tuberculin skin tests on cattle have proven costly and unreliable, and depend on antibody detection which cannot differentiate between active infection and a host’s immune defenses after overcoming disease.

Currently TB infects one-third of the world’s population and causes 1.6 million deaths per year (Kulchavenya et al 2012). Only 5% to 15% of cases develop into active TB suggesting a key mediating role for genes of the innate immune system (Khan et al 2013). Studies suggesting that TLR polymorphism correlates with TB susceptibility are particularly pertinent to this study, and are detailed below. The absence of TLR2 in mice has been associated with TB susceptibility (Stenger et al 2002). However, more specifically, in a study of 474 African-Americans (Velez et al 2009) and a linked study of 381 Caucasians and 667 Africans (Velez et al 2010) an indel polymorphism (−196 to −174) in TLR2 was significantly associated with increased risk of TB. Khan et al (2013) added that this same polymorphism was significantly associated with TB susceptibility across 187 volunteers from Pakistan.

Not only TLR2, but others such as TLR1 have been implicated in bTB susceptibility (Le Roex et al 2013). Sun et al (2012) found a TLR1-G1596A gene variant was significantly associated with bTB susceptibility due to a reduction in PAMP recognition in a study of 586 Chinese Holstein cows. Wujcicka et al (2014) pointed out that polymorphisms, especially of TLR2 and TLR4, have been found to enhance congenital transmission of TB. It is thus very possible that TLRs and the innate immune system might be involved in badger susceptibility to this disease. However, the research required to resolve this may require looking at multiplicit factors, such as polymorphisms acting in combination rather than alone, as well as variation between genders and between distinct population heritages, such as that between Irish and British badgers as discussed in section “1.1.2 Badger Evolutionary History”. Simply looking at a small range of known polymorphisms may not be enough, and instead looking at full genes we did in this study, may be called for. For example, Selvaraj et al (2010) found no association between five known polymorphisms and TB prevalence across 206 TB patients and 212 healthy controls, having tested for TLR-1 1805T/G (Ile602Ser), TLR-2 2258G/A (Arg753Gln),
TLR-4 896A/G (Asp299Gly), TLR-4 1196C/T (Thr399Ile), and TLR-6 745C/T (Ser249Pro).

However a study of all TLR variations may well have had more positive results.

As far as we are aware there are currently no publications on badger TLRs. No badger TLR has yet been published. There are no badger TLRs nor any badger full genome sequences on any DNA database. This research has begun to fill that gap.

Further development to the groundwork laid out here might include sequencing TLR2 and TLR4 from a wider sample set of badgers including outside UK. Or, sequencing all other TLRs from badgers. Cuscó et al (2014) studied the exon regions of all ten canine TLRs across 335 dogs from seven reeds and 100 wolves from two populations, and found TLR5, then TLR4, to be the most polymorphic TLR among canines. Indeed, while polymorphisms in both TLR4 and TLR5 are associated with increased resistance to inflammatory bowel disease in German Shepards, only the TLR5 polymorphism confers IBD resistance in 38 other breeds (Kathrani et al 2011). It would be interesting, therefore, to sequence all badger TLRs and deduce from their phylogenies which are most positively selected. We might also develop studies that investigate links between diseases and TLR polymorphisms, as described above. Or, investigate intron and flanking sequences for possible epigenetic marks that link with disease (e.g. CpG methylation); investigate other molecules of the innate immune system (e.g, MyD88, inflamasome components); or develop whole genome sequencing of badgers to reveal all possible SNPs involved with disease.

The TLR data generated by this study suggests a badger population bottleneck. However, the possibility exists that this could be an artefact of considering only this study, thus a literature search was performed which revealed a further eleven papers which considered variation in genetic markers across populations. The following studies also suggested low genetic variability in the European badger. Domingo-Roura et al (2003) examined twelve microsatellites across a badger population in Wytham Woods, Oxfordshire, and found that only five showed variability. However targets regions were short (94bp to 359bp) and the sample group small; one primer was tested across only 13 badgers, three across 16 badgers, and the remaining eight across 40 badgers. Badgers were potentially related individuals from a 6 km² area. Annavi et al (2014) performed a temporal study of badgers sampled from 1987
to 2002 across the same 6 km² area in Wytham Woods, genotyping 1170 individuals across 35 microsatellite loci, and found an average of only 4.46 alleles per locus. Huck et al (2008) studied twenty microsatellite loci across seventy-four badgers found at a sample site of less than 2 km² in Brighton, UK, and found a mean allelic richness value of 3·4 indicating low genetic variability. Dawnay et al (2008) sequenced ten Short Tandem Repeats from 1083 badgers from twenty populations across England and Wales and found very low allelic diversity even in this large sample range.

It is possible that postulations on genetic variability in badgers may in fact be dependent on the choice of sampled genes. The same seven microsatellites were used across three studies, all of which distinguished siblings with >99% certainty, both in British and continental populations. Scheppers et al (2007) used the seven loci to compare genetic methods of estimating population size via hair-trap sampling to observation. The sample population lived across 13 km² of adjoining sites between Eppeldorf and Medernach in Luxembourg. While sample population was small and geographically limited, the particular loci tested proved variant enough to identify more individuals (n=55) than was estimated by eye (n=49). In a similar study, Frantz et al (2004) expanded on an observed sample of 13 badgers to show that 15 individuals lived in an area of 5·4 km² between Ermsdorf and Eppeldorf in Luxembourg. The same was true across a British badger population: Frantz et al 2003 investigated badger faecal samples from Woodchester Park, the core sample area used in this study, using the same seven microsatellite loci and was able to distinguish between 33 individual badgers using this method.

Higher genetic variability appears to exist within badgers on the continent vs. Britain (Frantz et al 2004; Scheppers et al 2007), and also in Asia vs. Europe (Tashima et al 2011). Indeed, Pope et al (2002) sampled allelic variation across twenty-two microsatellites in 354 badgers ranging across England, Wales and Scotland were compared to 108 badgers across nine countries in continental Europe. Mean number of alleles per locus per individual was found to be much lower in the UK (0.14) compared to the continent (0.30). O’Meara et al (2012) studied six microsatellite loci across badgers in Holland (N = 29), Ireland (N = 28), and Spain (N = 16), and found that genetic variation on a loci by loci basis was not correlated with geographical but dependent on the loci being tested. Using the microsatellite Mel116 a high
diversity of 7.6 alleles per locus and per sample was found in Northern England, whereas other loci suggested low diversity across the entire range. However, overall loci the Irish samples showed lower diversity while continental samples showed greater differentiation. Tashima et al (2011) studied mitochondrial DNA across 112 badgers from a wide geographical range spanning the European continent to the Japanese islands. They traced a common ancestor of Eastern and western lineages to the Volga River in Russia, and showed that Japanese Meles meles have greater diversity, which was attributed to Japan’s island structure. Badgers based on the European continent were found to have twelve mtDNA haplotypes, while Japanese badgers had twenty-seven. A possible answer to low variability in British badgers may be to import their more genetically diverse counterparts on the continent to promote extra-group breeding which may in turn have positive effect on genetic diversity and thus immuno-adaptive capability in British badgers.

On the other hand a more recent study by Frantz et al (2010) compared sixteen microsatellite loci across 180 badgers over 500km² in Glouestershire and 149 badgers over 600km² in the Broye region of Switzerland, and found the genetic diversity across both populations not significantly different: Mean number of alleles per locus were 5.6 and 6.2 respectively. These results taken alone are contradictory and inconclusive. Genetic variability on the continent may also have regional variation. Further work should look at wider sample ranges across Britain and continental Europe, a wider variety genes, and larger sample groups.
5.0 References


Barreiro L. B. Marioni J. C., Blekhman R., Stephens M. and Gilad Y. (2010): Universal TLR response is more conserved than the immune responses to specific bacterial or viral infections. PLOS Genetics. 10.1371/journal.pgen.1001249.g002.


population decline of the critically endangered Australian marsupial, the brush tailed bettong or woylie (*Bettongia penicillata*). International Journal for Parasitology 2 77-89.


**Carrera-Silva** E. A., Cano R. C., Guinazu N., Aoki M. P., Pellegrini A. and Gea S. (2008) TLR2, TLR4 and TLR9 are differentially modulated in liver lethally injured from BALB/c and C57BL/6 mice during *Trypanosoma cruzi* acute infection. Molecular Immunology 45 3580-3588.


Devarajua P., Gulatib R., Antonya P. T., Mithuna C. B. and Negia V. S. (2014) Susceptibility to SLE in South Indian Tamils may be influenced by genetic selection pressure on TLR2 and TLR9 genes. Molecular Immunology 4524 Author’s copy.


Donnelly C. A and Hone J. (2010) Is there an association between levels of bovine tuberculosis in cattle herds and badgers? Statistical Communications in Infectious Diseases 2:1 Accession Number 20133149921.


**Finelle** P. (1972) African animal trypanosomiasis. Disease and Chemotherapy 1-26


Liadaki K., Petinaki E., Skoulakis C., Tsirevelou P., Klapsa D., Gemenis A. E. and Speletas M.
(2011) Toll-Like Receptor 4 Gene (TLR4), but Not TLR2, Polymorphisms Modify the Risk of Tonsillar Disease Due to Streptococcus pyogenes and Haemophilus influenzae. Clinical And Vaccine Immunology, Feb. 2011 217–222


