In-silico identification and characterisation of 17 polymorphic anonymous non-coding sequence markers (ANMs) for red grouse (*Lagopus lagopus scotica*)

Marius A. Wenzel* and Stuart B. Piertney

December 16, 2014

Institute of Biological and Environmental Sciences, University of Aberdeen, Zoology Building, Tillydrone Avenue, Aberdeen AB24 2TZ, UK

* corresponding author. email address: marius.a.wenzel.08@aberdeen.ac.uk. Phone number: +44 1224 272395

Word counts: 133 (abstract), 779 (main text), 86 (legends)
Keywords: anonymous nuclear markers; non-coding DNA; neutral evolution; red grouse; Lagopus

Abstract

Anonymous non-coding sequence markers (ANMs) are powerful neutral genetic markers with great utility in phylogeography, population genetics and population genomics. Developing ANMs has previously relied on sequencing random fragments of genomic DNA in the target species and then querying bioinformatics databases to identify unannotated, putatively neutral fragments. Here, we describe an alternative *in silico* approach that is based on identifying large unannotated genomic regions in model species to provide *a priori* neutral targets for candidate ANMs that are remote from exonic regions. We illustrate this approach by developing a set of 17 polymorphic ANMs for red grouse (*Lagopus lagopus scotica*) from c. 1 Mbp non-coding chromosome regions of chicken, turkey and zebrafinch genomes. This pipeline represents a powerful and efficient approach when appropriate model genomes are available for the target species of interest.

The ability to isolate and characterise nuclear DNA sequence polymorphisms remains a major priority for studies resolving population history, estimating demographic parameters and examining the genetic basis of divergence, adaptation and speciation (Thomson et al, 2010). In non-model species, one classic approach is to use exon-primed intron-crossing markers (EPICs) or comparative anchor-tagged sequences (CATS), which target nuclear intronic sequences by anchoring primers in conserved flanking exonic regions (Backström et al, 2008). These markers are considered useful for phylogenetics, gene mapping and population genetics because of high variability, cross-species utility and presumed neutrality (Brito and Edwards, 2009; Slate et al, 2009). However, they are unlikely to be truly neutral because purifying selection on flanking exons may affect intronic polymorphism through hitchhiking (Thomson et al, 2010). In contrast, nuclear anonymous non-coding markers (ANMs) that are located in regions remote from exonic domains are unlikely to be under selection and are substantially more polymorphic than EPICs or CATS (Thomson et al, 2010). Additionally, ANMs are more abundant and easier to type than microsatellites, making them ideal tools for population genetics and phylogeography (Rosenblum et al, 2007; Lee and Edwards, 2008; Thomson et al, 2010).

Isolating ANMs is usually based on sequencing random fragments of genomic DNA following shearing (Rosenblum et al, 2007; Lee and Edwards, 2008) or enzymatic digestion (Barlow et al, 2012; Ren et al, 2013), or via whole-genome massive parallel sequencing (Bertozzi et al, 2012; Lewis et al, 2014). Non-coding sequences can then be identified from absence of annotations following BLAST (Altschul et al, 1997) queries against bioinformatics databases, and primers are designed accordingly (Bertozzi et al, 2012; Lewis et al, 2014). One issue with this strategy is that primer design on library clone sequences may be compromised because unidentified polymorphism in binding sites may cause null-alleles, PCR failure and poor cross-species utility (Thomson et al, 2010). Most crucially, however, neutrality cannot be established from mere absence of BLAST results. Confirming remoteness from exonic domains as a criterion for neutrality requires examining the genomic context of the sequences in model genomes, but direct sequence mapping may be difficult if no taxonomically close model genome is available.

Here, we describe an alternative strategy to identifying ANMs that is purely based on available bioinformatics resources and provides *a priori* candidate targets for designing primers in non-coding regions that are remote from exonic regions and hence likely to be truly neutral. We illustrate this strategy by developing ANMs from avian model genomes for red grouse (*Lagopus lagopus scotica*), an economically important game bird endemic to upland heather moors in Scotland and northern England (Martínez-Padilla et al, 2014).

The UCSC Table Browser (Karolchik et al, 2004) provides tabulated annotations from published genomes. RefSeq annotations were downloaded for the chicken genome (*Gallus gallus* galGal4 assembly) and analysed using custom scripts in R 3.0.3 (R Core Team, 2014). The table fields *txStart* and *txEnd* were used to calculate genomic distances (bp) between consecutive transcription blocks across each autosome. The maximum region size per autosome ranged from 0.1 Mbp to 5.1 Mbp (median 1 Mbp) and a total of 113, 19 and 7 regions of at least 1 Mbp, 2 Mbp and 3 Mbp, respectively, were available across all autosomes (Figure 1). Nine c. 1 Mbp regions in nine autosomes were arbitrarily selected as candidate target regions (Figure 1). The central 10 kbp portion of these regions was extracted from GENBANK chromosome sequences, and homologous sequences in turkey (*Meleagris gallopavo* melGal1 assembly) and zebrafinch (*Taeniopygia guttata* taeGut1 assembly) genomes were identified using the BLAST-like alignment tool BLAT (Kent, 2002). Alignments of all three species and also chicken and turkey alone were generated in GENEIOUS v5.6.3 (Drummond et al, 2012). Non-degenerate primers (200–800 bp amplicon size, 18–27 bp primer length, 20–80 % GC content, 50–64 ºC melting temperature) were then designed opportunistically on small conserved regions using PRIMER3 (Rozen and Skaletsky, 2000) as implemented in GENEIOUS. Primer specificity was tested using UCSC IN-SILICO PCR amplicon prediction (Hinrichs et al, 2006) on the chicken, turkey and zebrafinch genomes.

Sequence polymorphism was ascertained in three red grouse individuals from locations that maximise geographic variation across a network of grouse moors in north-east Scotland (Glenlivet 57.29 °N 3.18 °W, Mar Lodge 56.95 °N 3.66 °W and Invermark 56.89 °N 2.88 °W). PCR conditions followed Wenzel et al (2014), with annealing temperatures as detailed in Table 1. Amplicons were Sanger sequenced in both directions, sequences were aligned in GENEIOUS and heterozygote sites were coded as IUPAC degenerate bases. Absence of exonic annotations was re-confirmed using BLASTN against the GENBANK NT database (Altschul et al, 1997). Polymorphic sites, numbers of haplotypes, nucleotide diversity, haplotype diversity and Tajima's D were then computed on reconstructed haplotypes derived from the PHASE algorithm in DNASP v5 (Librado and Rozas, 2009).

Twenty-two out of thirty primer pairs (73 %) amplified in red grouse, demonstrating a high success rate of

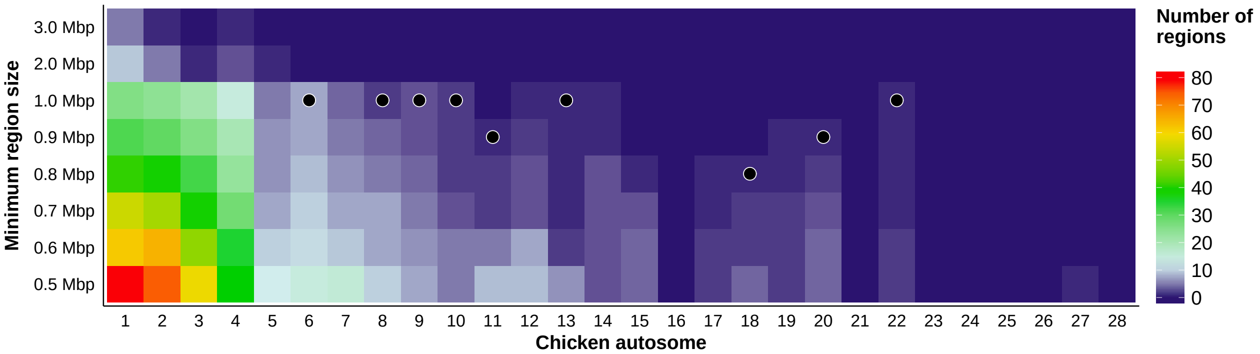

Figure 1: Numbers of unannotated genomic regions of particular minimum sizes in chicken autosomes, based on distances between consecutive transcription blocks. Black dots represent candidate regions selected for ANM design (Table 1).

our development strategy. Polymorphic sequence alignments were obtained for seventeen loci (57 %), containing 1–18 SNPs that define 2–6 haplotypes with no evidence of deviation from neutral sequence evolution (Table 1). Insertions/deletions of 1–10 bp were present in five loci. These polymorphic ANMs provide a valuable resource for a range of population genetics or genomics applications in red grouse. The zebrafinch genome impeded primer design in many cases due to its taxonomic distance (Table 1), but considering the taxonomic distance between red grouse, chicken and turkey, these markers should be conserved and hence useful across a range of closely related galliform species.

Acknowledgements

This study was funded by a BBSRC studentship (MA Wenzel) and NERC grants NE/H00775X/1 and NE/D000602/1 (SB Piertney). We are grateful to Mario Röder, Keliya Bai and Marianne James for help with fieldwork, and all grouse estate factors, owners and keepers, most particularly Alistair Mitchell, Shaila Rao, Christopher Murphy, Richard Cooke and Fred Taylor, for providing access to estate game larders.

References

Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Backström N, Fagerberg S, Ellegren H (2008) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. Mol Ecol 17(4):964–980

Barlow A, Grail W, de Bruyn M, Wüster W (2012) Anonymous nuclear markers for the African adders (Serpentes: Viperidae: *Bitis*). Conserv Gen Res 4(4):967–969

Bertozzi T, Sanders KL, Sistrom MJ, Gardner MG (2012) Anonymous nuclear loci in non-model organisms: making the most of high-throughput genome surveys. Method Biochem Anal 28(14):1807–1810

Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica 135(3):439–455

Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, S S, Thierer T, Wilson A (2012) Geneious v5.6.3. Available from http://www.geneious.com

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al (2006) The UCSC genome browser database: update 2006. Nucleic Acids Res 34(suppl 1):D590–D598

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32(suppl 1):D493–D496

Kent WJ (2002) BLAT–The BLAST-Like Alignment Tool. Genome Res 12:656–664

Lee JY, Edwards S (2008) Divergence Across Australia's Carpentarian Barrier: Statistical Phylogeography of the Red-Backed Fairy Wren (*Malurus melanocephalus*). Mem New York Botan G 62(12):3117–3134

Lewis CJ, Maddock ST, Day JJ, Nussbaum RA, Morel C, Wilkinson M, Foster PG, Gower DJ (2014) Development of anonymous nuclear markers from Illumina paired-end data for Seychelles caecilian amphibians (Gymnophiona: Indotyphlidae). Conserv Gen Res 6(2):289–291

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Method Biochem Anal 25(11):1451–1452

Table 1: Characterisation of 17 anonymous non-coding sequence markers (ANMs) for red grouse. Primer GC content, melting temperature T_m and annealing temperature T_a (^{TD}=TouchDown) are presented alongside genomic locations in three bird genomes and sequence diversity statistics derived from three red grouse individuals (segregating sites S, nucleotide diversity π, haplotypes H, haplotype diversity H_d, Tajima's D).

					in silico amplification (genomic location and predicted amplicon size)			*in vitro* amplification						
Primer name	Primer sequence (5' → 3')	GC (%)	T_m (ºC)	T_a (ºC)	Chicken	Turkey	Zebrafinch	Size	S	π	H	H_d	D	GENBANK accession
Lls_ANM_6_1F	ACCCTTGTGAGCTGAGAGCTT	52	59.7	60–50^{TD}	chr6:5360100+5360504	chr8:2879375+2879779	–	406	9	0.011	4	0.800	0.793	KM379116
Lls_ANM_6_1R	TCACACCTATGGCAAACAAAACAC	41	56.7		405 bp	405 bp	–							
Lls_ANM_8_1F	TGGCCAGGGTTATCTGGAGTGC	59	63.3	68	chr8:9422697+9423060	chr10:1058050-1058413	–	287^{ab}	3	0.005	4	0.867	0.600	KM379117
Lls_ANM_8_1R	TGCCCTCTGAAGAAGCCATTTGA	47	59.8		364 bp	364 bp	–							
Lls_ANM_8_2F	TCTGTCACTGTTCTCACATTTT	36	52.1	60–50^{TD}	chr8:9424802+9425186	chr10:1055959-1056338	–	388^{b}	2	0.002	2	0.333	-1.132	KM379118
Lls_ANM_8_2R	CACTCAATTTGATTTTCTCAGTAACC	34	52.5		385 bp	380 bp	–							
Lls_ANM_9_1F	AGTCTGAGACATTTTCCCCATCC	47	57.6	65	chr9:20994525+20994916	chr11:21699903+21700290	–	390^{b}	5	0.007	4	0.867	0.708	KM379119
Lls_ANM_9_1R	AGAACTCATTCTGCTTTGCAGC	45	56.8		392 bp	388 bp	–							
Lls_ANM_9_2F	TGAAATGTACTTCCTAACACATGC	37	53.4	60–50^{TD}	chr9:20992310+20992689	chr11:21697593+21697976	–	385	3	0.004	4	0.867	1.386	KM379120
Lls_ANM_9_2R	TGTTTTTCTTTTCTGATTTATGTGGA	26	50.8		380 bp	384 bp	–							
Lls_ANM_9_3F	CTCCAGGATACTCAAGCCACA	52	57.5	65	chr9:20999542+20999958	chr11:21704861+21705255	–	407	2	0.002	3	0.733	-0.050	KM379121
Lls_ANM_9_3R	TCCTTGCAGTTTTTAGACTTGGA	39	54.6		417 bp	395 bp	–							
Lls_ANM_10_1F	CACTCAGCCCTCTGTGTAAGTGC	56	61.4	65–55^{TD}	chr10:15610965-15611270	chr12:16100168-16100473	chr10:16580442-16580734	305	1	0.001	2	0.333	-0.933	KM379122
Lls_ANM_10_1R	TGAGTTGTTAGACCACACGCGA	50	59.6		306 bp	306 bp	292 bp							
Lls_ANM_10_2F	ACTCGCGTGTGGTCTAACAACTC	52	60.4	65–55^{TD}	chr10:15610580-15610988	chr12:16099786-16100191	chr10:16580040-16580466	393	2	0.002	3	0.733	0.311	KM379123
Lls_ANM_10_2R	ACTGCATGGTGGGGAATGCCA	57	63.7		409 bp	406 bp	426 bp							
Lls_ANM_10_3F	TTGCTGCCTGCCACTGCTTA	55	61.3	65–55^{TD}	chr10:15611228+15611646	chr12:16100431+16100850	–	278^{ab}	18	0.034	6	1.000	0.723	KM379124
Lls_ANM_10_3R	AGCCACACCTCCCCCATTCA	60	63.1		419 bp	420 bp	–							
Lls_ANM_11_1F	AGTTGACATCAAAGTGGAGACA	40	54.3	65–55^{TD}	chr11:5087733+5088037	chr13:5398213+5398522	chr11:9467370+9467669	304	4	0.004	2	0.333	-1.295	KM379125
Lls_ANM_11_1R	GTGTCTGGTTTCACATCTGGC	52	57.5		305 bp	310 bp	299 bp							
Lls_ANM_13_1F	GGACATTTAGCAACAAGGTCACA	43	56.1	65–55^{TD}	chr13:5734379+5734751	chr15:5954643+5955002	–	373	3	0.004	3	0.800	1.124	KM379126
Lls_ANM_13_1R	GGATGATTAGGCTGTGTAAAACCC	45	56.6		373 bp	360 bp	–							
Lls_ANM_13_3F	TGTGGATGTATACTACCTGGCA	45	55.8	60–50^{TD}	chr13:5741550+5741896	chr15:5961829+5962181	–	244^{ab}	3	0.006	3	0.733	0.338	KM379127
Lls_ANM_13_3R	GCTGATACCTTTATAAACTTTGGTGT	34	53.3		347 bp	353 bp	–							
Lls_ANM_18_1F	TGGAAGCCATGAGGAAGGGGA	57	62.2	67	chr18:8776322+8776692	chr20:6558242-6558622	–	379	7	0.009	4	0.867	0.508	KM379128
Lls_ANM_18_1R	AGGAAGGAAGAATGCAAGGCA	47	57.8		371 bp	381 bp	–							
Lls_ANM_18_2F	TCAGCCAATATTGCTTCAAAGG	40	54	60–50^{TD}	chr18:8779401+8779815	chr20:6555203-6555600	–	322^{a}	7	0.008	4	0.867	-0.631	KM379129
Lls_ANM_18_2R	TCCAATGAAATGAAGCTGTATGC	39	53.9		415 bp	398 bp	–							
Lls_ANM_20_2F	ATTCCTCGCTGGTTGCTGGC	60	62.6	68	chr20:4665027+4665427	chr22:4275127+4275522	–	403	9	0.008	3	0.600	-0.818	KM379130
Lls_ANM_20_2R	CTGCACTTGTGGGCAGACCC	65	63.5		401 bp	396 bp	–							
Lls_ANM_22_2F	GCGATGCTACACCCTCCAAG	60	59.9	60–50^{TD}	chr22:3864116+3864470	chr24:3873443+3873798	–	357	2	0.002	2	0.333	-1.132	KM379131
Lls_ANM_22_2R	ACAAAATGCTACTGACAAATCTGA	33	52.6		355 bp	356 bp	–							
Lls_ANM_22_3F	GCTTTCCCTCCTCTATTTCCTTC	47	56	66	chr22:3862528+3862926	chr24:3871853+3872244	–	397	6	0.008	3	0.733	1.392	KM379132
Lls_ANM_22_3R	AGAATCCCAAAGCCTTTCCCT	47	57.4		399 bp	392 bp	–							

[a]: partial alignment due to unresolvable electropherogram peaks (multiple heterozygote INDEL mutations)

[b]: alignment contains INDEL mutations

Martínez-Padilla J, Redpath SM, Zeineddine M, Mougeot F (2014) Insights into population ecology from long-term studies of red grouse *Lagopus lagopus scoticus*. J Anim Ecol 83(1):85–98

R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org

Ren QP, Fan Z, Zhou XM, Jiang GF, Wang YT, Liu YX (2013) Identification and characterization of anonymous nuclear markers for the double-striped cockroach, *Blattella bisignata*. B Entomol Res 103(01):29–35

Rosenblum E, Belfiore N, Moritz C (2007) Anonymous nuclear markers for the eastern fence lizard, *Sceloporus undulatus*. Mol Ecol Notes 7(1):113–116

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. Meth Mol Biol 132:365–386

Slate J, Gratten J, Beraldi D, Stapley J, Hale M, Pemberton JM (2009) Gene mapping in the wild with SNPs: guidelines and future directions. Genetica 136(1):97–107

Thomson RC, Wang IJ, Johnson JR (2010) Genome-enabled development of DNA markers for ecology, evolution and conservation. Mol Ecol 19(11):2184–2195

Wenzel MA, Webster LMI, Paterson S, Piertney SB (2014) Identification and characterisation of 17 polymorphic candidate genes for response to parasitic nematode (*Trichostrongylus tenuis*) infection in red grouse (*Lagopus lagopus scotica*). Conserv Gen Res