Session 4 gpollott@rvc.ac.uk

Refining bioinformatic methods to locate functional DNA in the bovine genome

Geoff Pollott Royal Veterinary College, Royal College Street, London, NW1 0TU, UK.

1 Motivation

- Functional DNA (fDNA) comprises all segments of the genome which contribute to make an organism work
- fDNA includes protein coding genes, various RNAs, transcription factors, promoter regions, etc.
- A method to locate fDNA in advance of any genomic analysis could reduce the task dramatically

2 Main findings

- 64% of exonic bp located in a known 3% of the bovine genome using the Neutral Indel Model (NIM; Lunter et al., 2006)
- This can be improved to 78% of exonic bp in a known 11% of the genome using G+C content of genome windows
- A further 11% of exonic bp can be located using percentage identity scores (PID) of DNA segments with a closely related species

3 Molecular evolution and fDNA

- Three types of selection operate on any given base pair (bp) site
- 1) Neutral selection – site subject to random mutations
- 2) Purifying (negative selection) – site shows no evidence of mutation because any mutations are removed from population
- 3) Positive selection – mutation at a site can have a beneficial effect on the fitness of the organisms – mutation retained
- Neutral Indel Model finds segments of DNA subjected to purifying selection using indel (insertions/deletions) gap distribution between two moderately related species (e.g. cow and human)
- Sites of positive selection identified by low percentage identity score (PID) between two closely related species (e.g. human and chimpanzee)
- Exons are associated with DNA having high guanine and cytosine (G+C) content

4 Materials and methods

- BlastZ 16-multiway alignments downloaded, including bovine (bosTau4), human (hg18), mouse (mm8) and chimpanzee (panTro2)
- NIM used to locate inter-gap segments (IGS) under purifying selection in bovine genome using the human genome as the secondary species
- G+C content of segments calculated and arranged in 20 ‘bins’
- Short inter-gap segments not containing purified DNA had their PID calculated using human and chimpanzee alignments
- Genscan gene prediction file used to test results for presence of fDNA
- Exonic bp in Genscan file ‘found’ in various segments used to test for fDNA

Figure 1. Distribution of exonic base pairs (bp) by inter-gap segment (IGS) length and G+C proportion

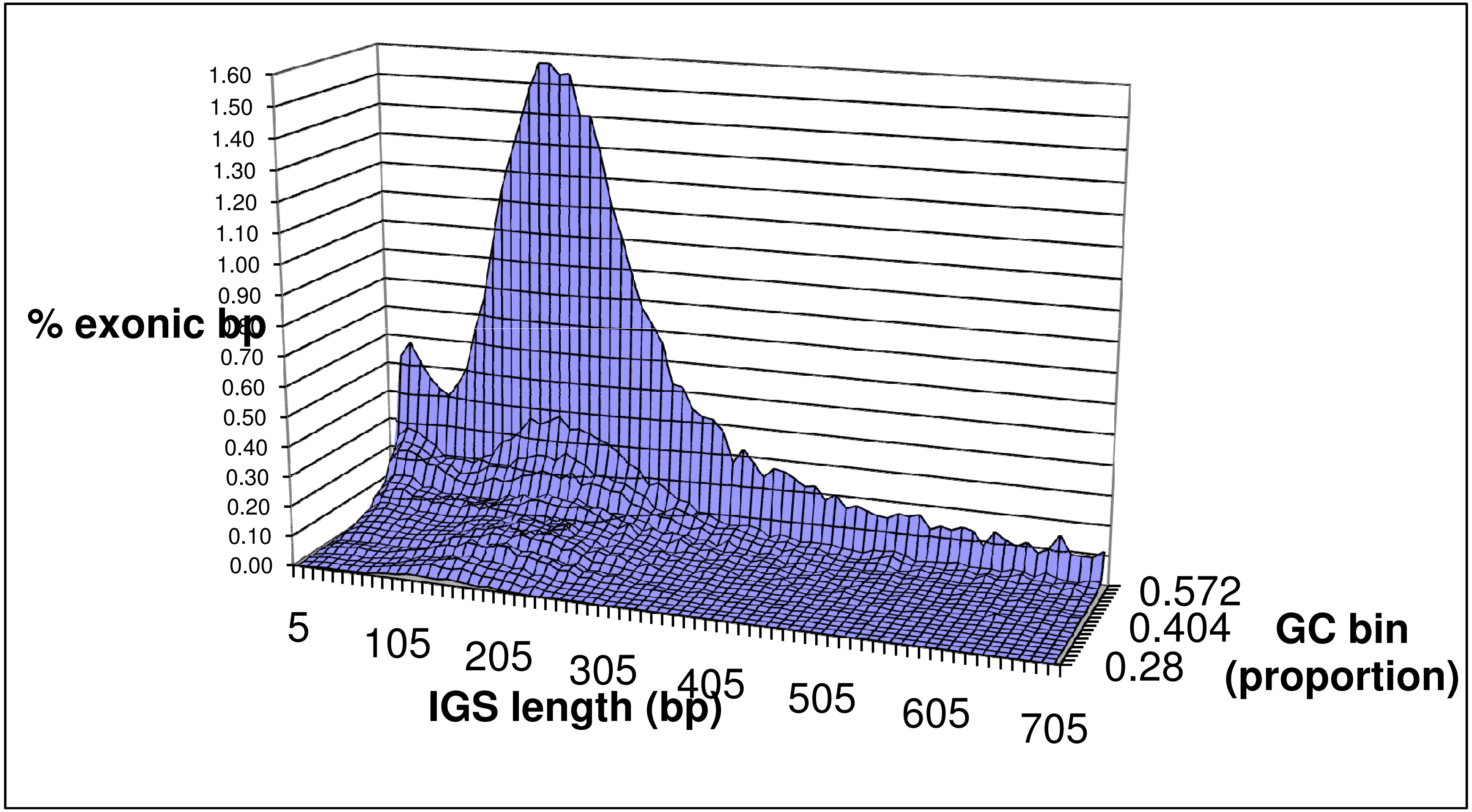
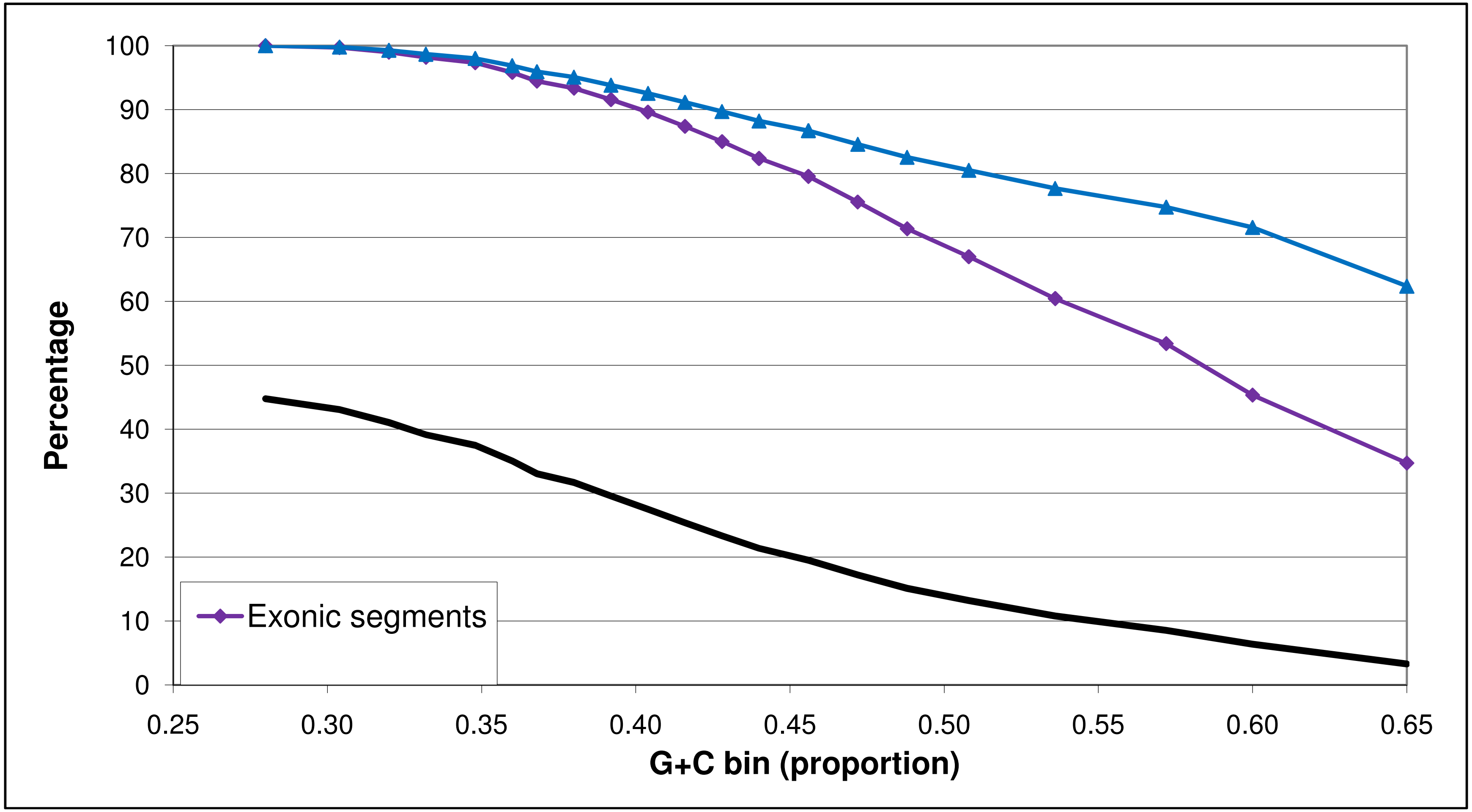


Figure 2. Reverse cumulative proportion (%) of exonic DNA found in combinations of different G+C bins (0.65 bin represents no G+C bin data)



5 Results 1 *Sites under purifying selection*

- Majority of exonic bp found in longer segments and in segments with higher G+C content (Figure 1)
- Segments in the highest four G+C content bins add 19% of the exonic bp for an additional 11% of the genome (Figure 2)

6 Results 2 *Sites under positive selection*

- Sites of positive selection analysed in all segments not subject to purifying selection
- Segments with a PID < 0.99 contained 11% of exonic bp in ~8% of the genome
- About 20% of exonic bp still not located in segments analysed for both purifying and positive selection

7 Discussion points

- Using both high G+C content segments and those under positive
- selection its is possible to locate 70% of exonic bp in 20% of the genome
- Promising first attempt at finding fDNA using NIM is difficult to improve on
- Many short segments contain exonic bp – must be many protein-coding genes which readily accept indels
- Need to look for additional method which can characterise these genes – could look for alignments which contain several short exonic segments

Reference

- Lunter, G., Ponting, C. P. and Hein, J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *Computational Biology*, **2**:e5.