Using bioinformatics to reduce the search for genes down to a known 4% of the cow genome

Geoff Pollott

SAC, Sustainable Livestock Systems Research Group,
Edinburgh, U.K.

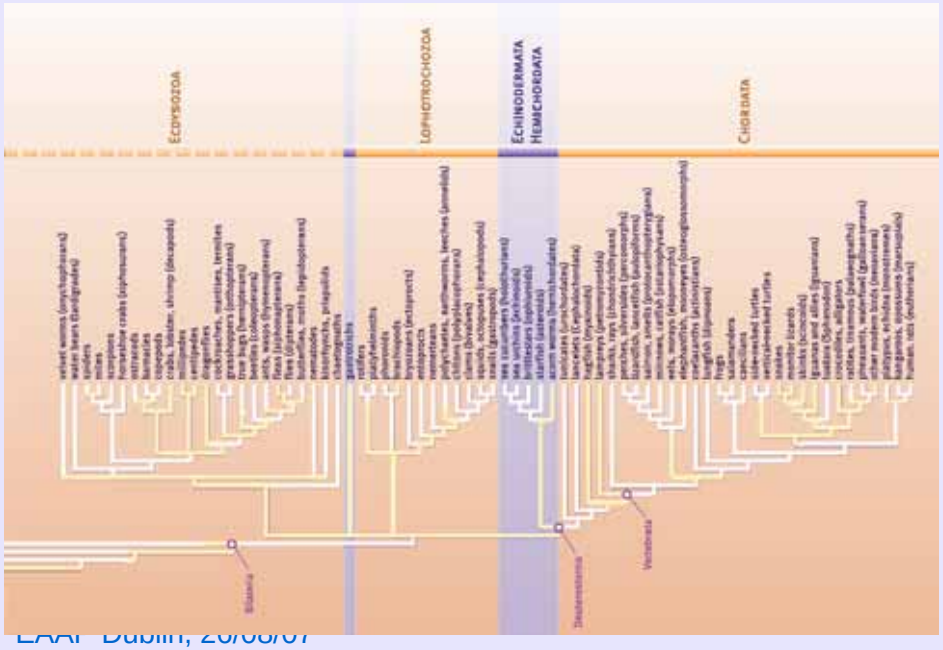
Motivation

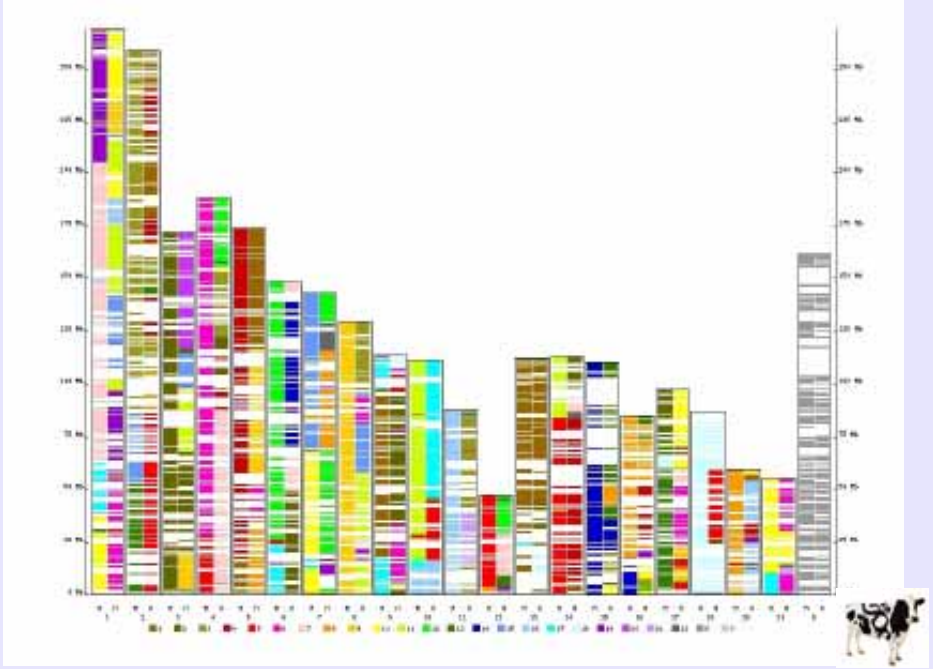
- Moving into the post-'black box' era of livestock genetics
- Genome-wide scans becoming common
- Basic question in genetics – How can we find DNA features of interest in this vast string of bases?
- Can we identify areas of the genome known to contain functional DNA?

Use Neutral Indel Model proposed by Lunter *et al.* (2006)

Basics of the Neutral Indel Model

Uses combination of ideas from molecular evolution and comparative genomics to identify segments of the genome containing functional DNA





Alignments

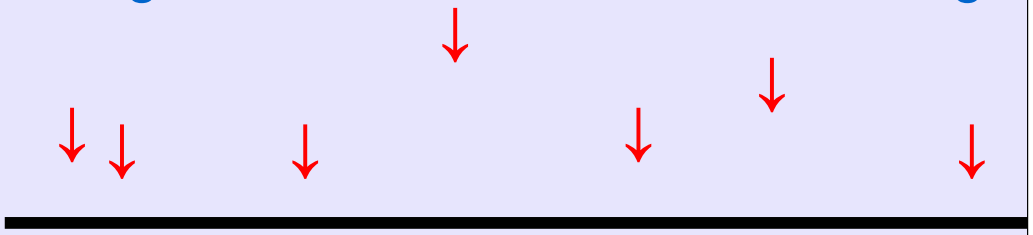
- Publicly available databases contain alignments of complete genomes – segments of genome with large similarity
- Produced by Blastz over whole of both genomes

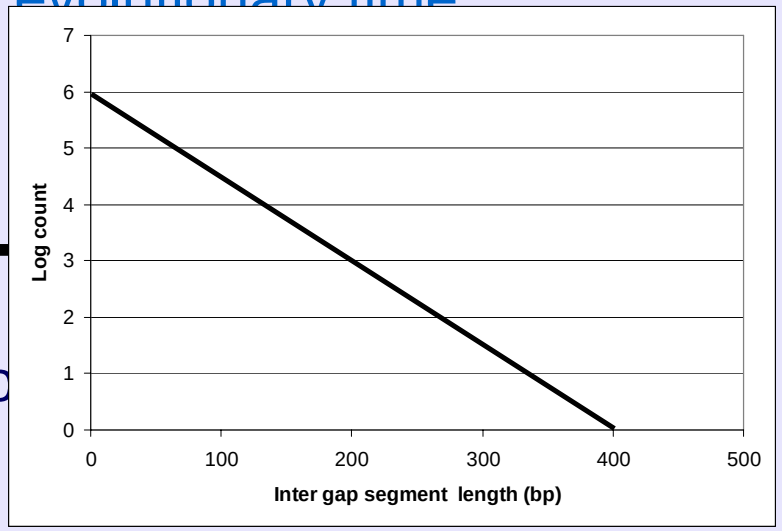
```
2 chr1 3569 5509 chr21 32540445 32542324 + 37620
TCTTTTGTTGGAAAGAA-------TCAGCAGG---CAGC     Cow
TCATTTGCTGGAAAGAGAAGCTGGGCCAGCAGAGCTCAGC     Human
```

Insertion or deletion (indel)

Neutral Indel Model (NIM)

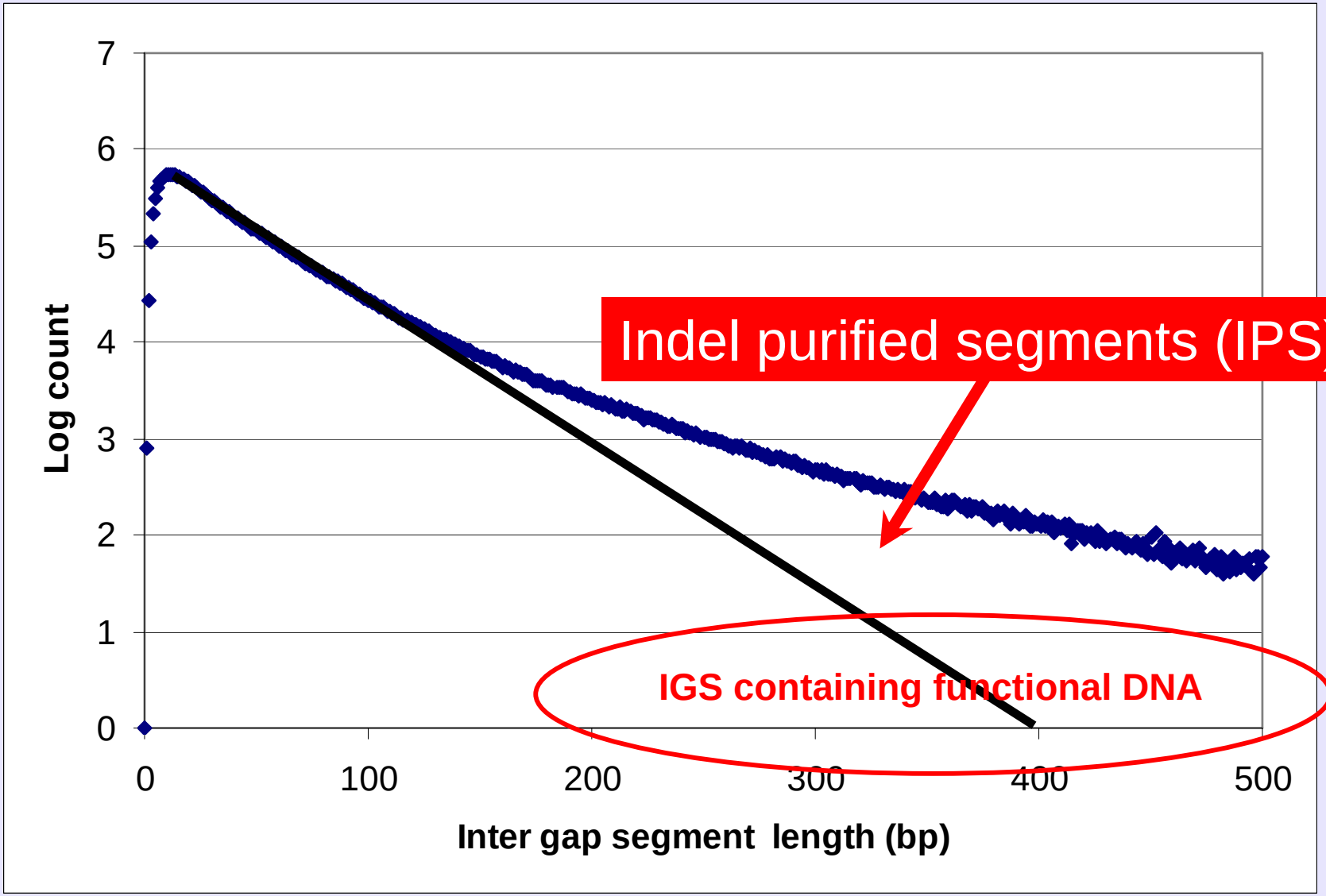
Indels raining down on a chromosome through evolutionary time





- Indels potentially randomly distributed acro
- Disruptive indels in segments of functional purifying selection
- Segments between remaining indels called Inter Gap Segments (IGS)
- Lunter et al (2006) proposed that under neutral selection the distribution of IGS has a geometric distribution
- Distribution of IGS differs in neutral DNA compared to functional DNA

Actual plot of Inter Gap Segment distribution



Neutral indel model and *Bos taurus* genome

Bos taurus genome length (build 2)

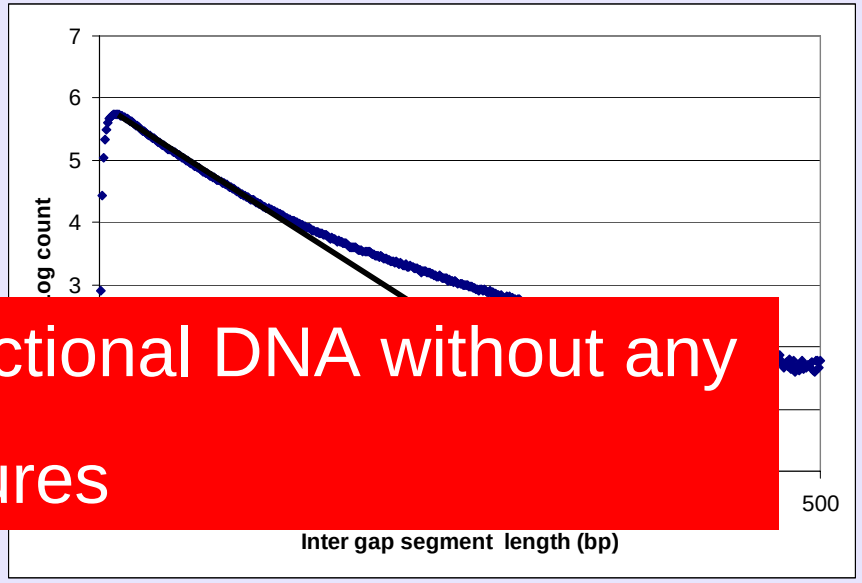
1,741 million bp

% aligned with human genome

48.1%

% ir

43.6%



Neutral Indel Model finds functional DNA without any prior knowledge of DNA features

~20 million IGS (1-2000bp)

Range in functional DNA estimated from IPS (0.1 FDR)

3.9 to 4.8% of genome

~250,000 IPS

How can we test this result?

- Take a class of known functional DNA and see how much of it is found in the identified IPS
- Protein-coding genes are the obvious class of functional DNA to use to test the model
- Use the *Bos taurus* Genscan file from internet
- Set up file of exons
- See what proportion of them overlap with IPS

Finding exonic DNA in Bovine IPS

- Exons in Genscan file

  25,012,251bp – 1.44% of g

- Exons in alignments

  24,010,839bp – 96% of exc

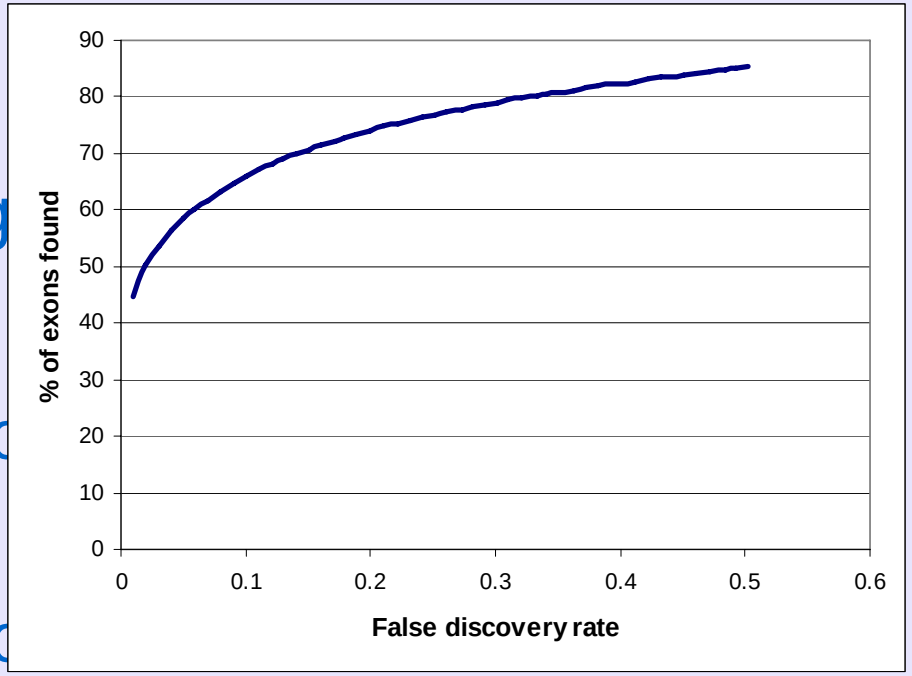
- Exons in inter gap segments

  22,726,172bp – 91% of exc

- Exons in indel-purified segments

  15,000,425bp – 64% of exonic DNA (at FDR of 0.1)

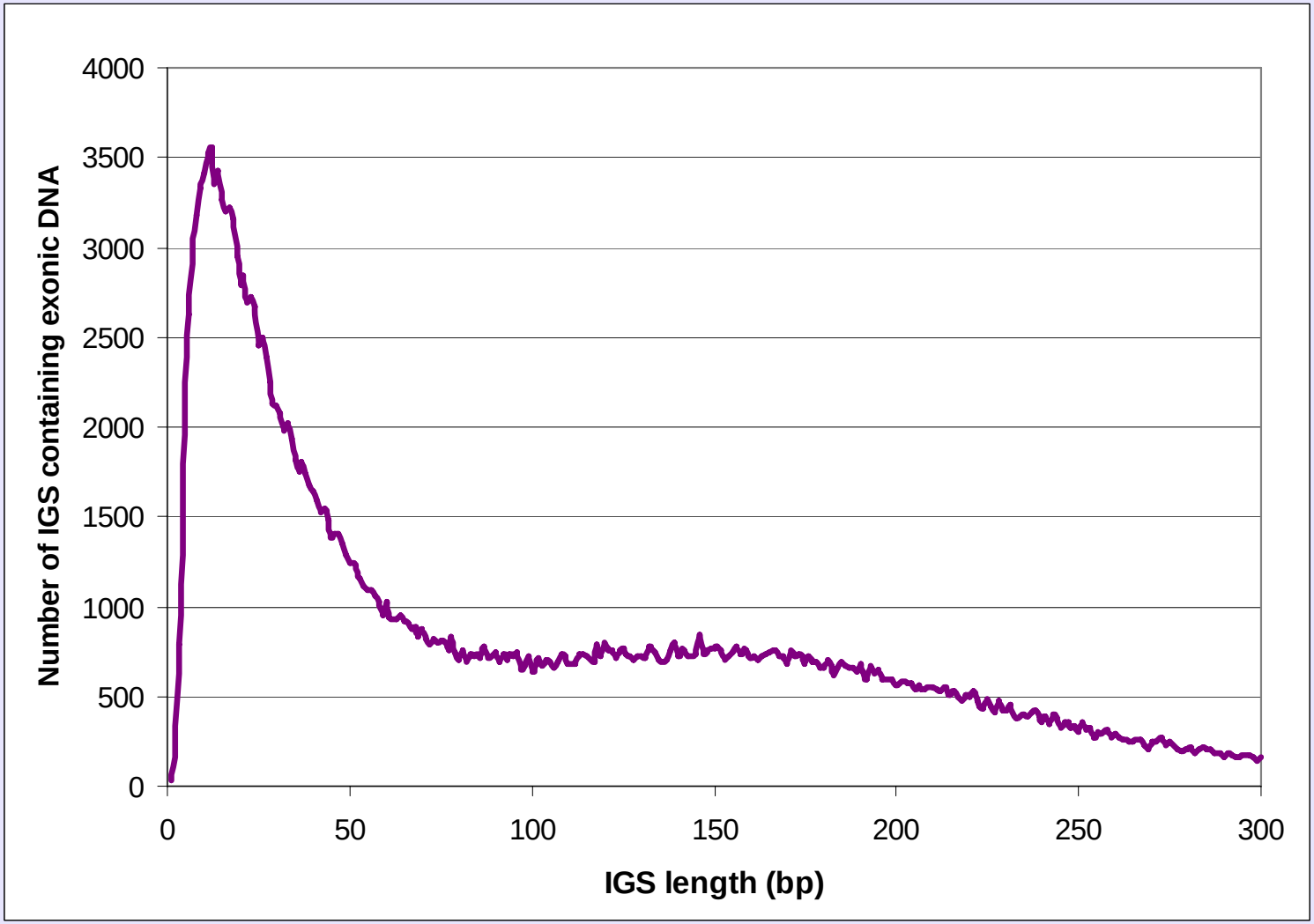
  20,156,372bp – 81% of exonic DNA (at FDR of 0.5)



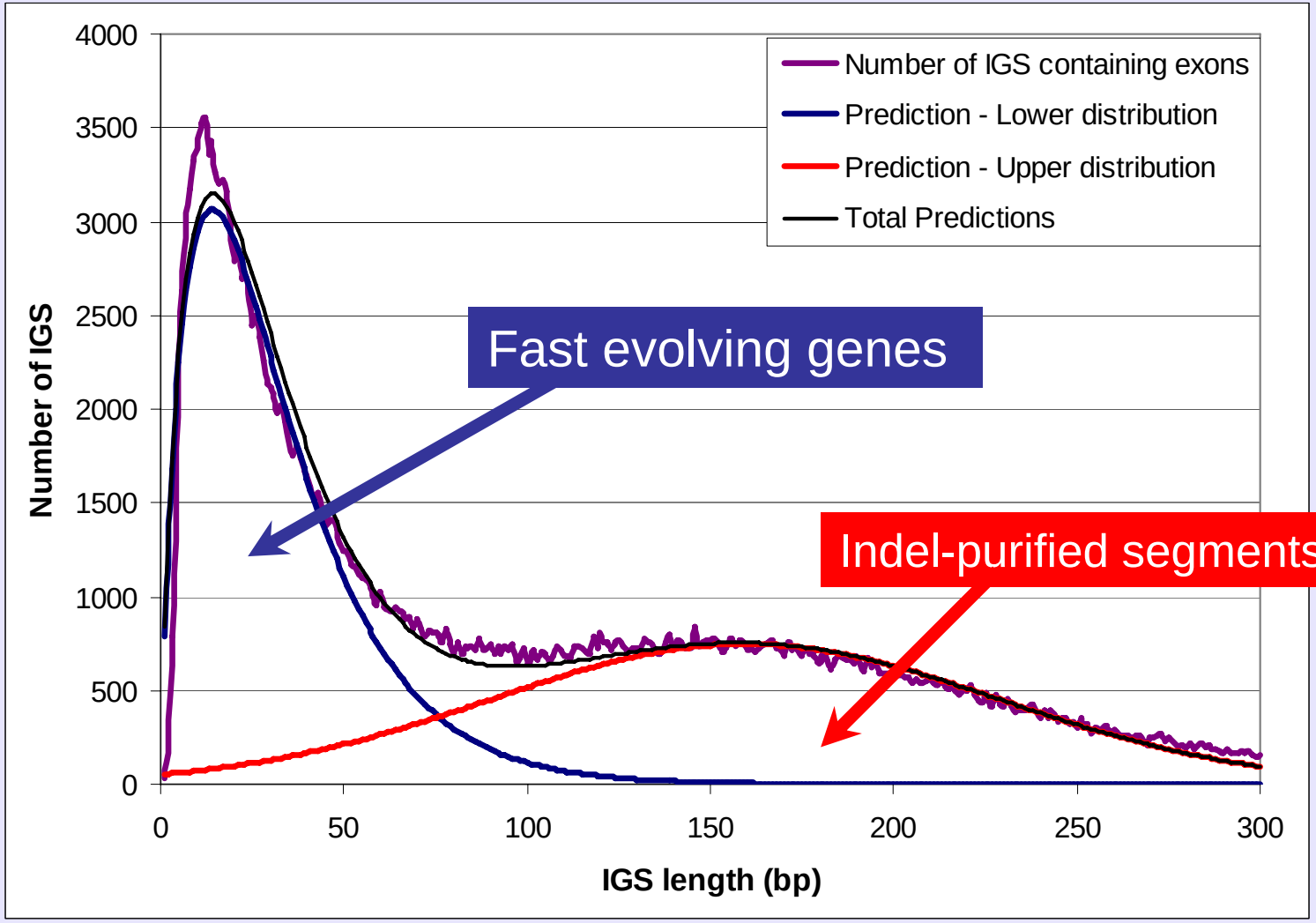
Gene and Exon hit rate

- Number of genes - 22,050
- Number of exons - 150,797
- Genes containing IPS – 19,055 (86.5%)
- Exons containing IPS – 77,317 (51.3%)
- Increase to 94.8% and 79.9% at 0.5 FDR
- IPS from 0.5 FDR includes 10.7% of genome
- Several attempts made to increase hit rate using additional criteria – all failed to improve result

Why does the NIM not find all functional DNA?



Distribution of IGS containing exonic DNA



Conclusions

- NIM method can find most highly conserved functional DNA without knowing its function
- Methods need to be developed to find the remaining exonic DNA
- Other poorly conserved functional DNA features may also require additional methods

Acknowledgements

Gerton Lunter – method and software

Baylor College and University of California at Santa Cruz - data

SAC – time to attend Oxford MSc Bioinformatics courses