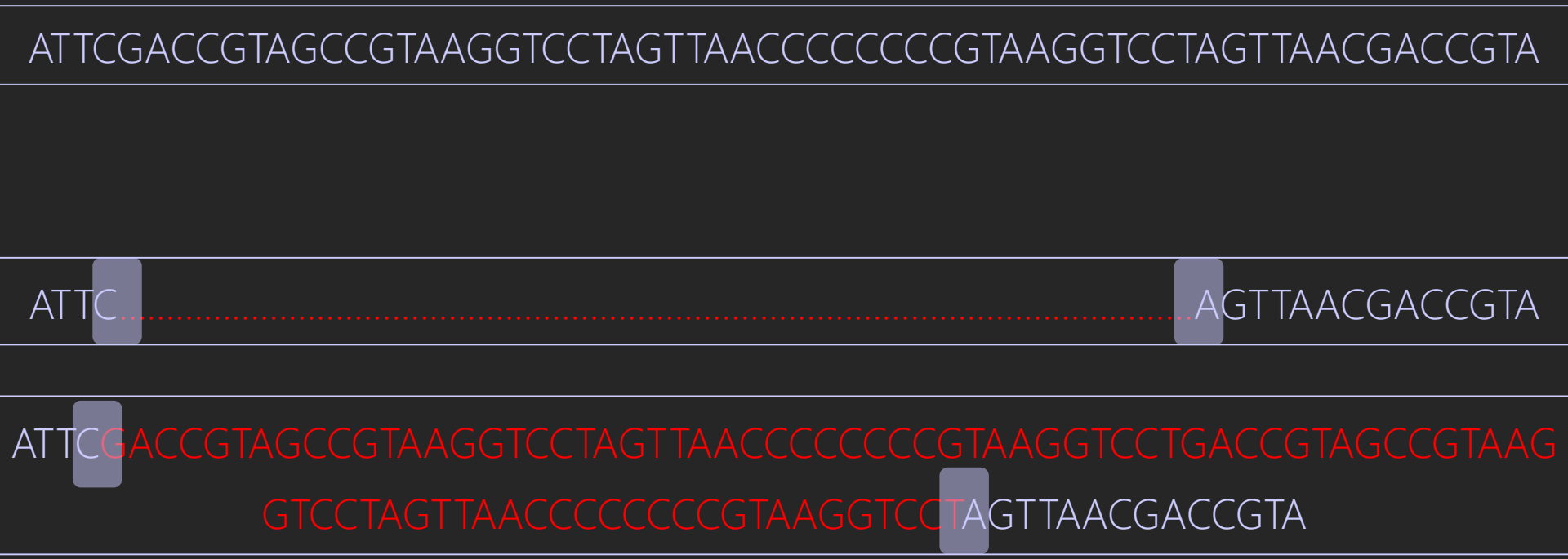
Functional annotation of CNV breakpoints in Holstein-Friesian cows

Joanna Szyda, Magda Mielczarek, Magdalena Frąszczak

Riccardo Giannico, Giulietta Minozzi, Ezequiel Nicolazzi

Katarzyna Wojdak-Maksymiec

CNVs – Copy Number Variants



Are CNV breakpoints formed at random in the bovine genome ?

Why breakpoints ?

Outline

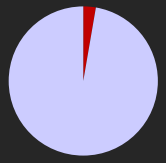
1. Material
2. CNV detection
3. Functional annotation of CNV breakpoints
   - SO terms
   - Gene functional category
   - All genes
4. Conclusions

Whole genome sequence data

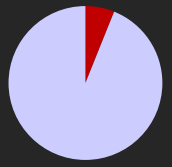
- 32 HF cows → 29 HF cows
- 16 paternal half-sibs
- WGS – Illumina HiSeq 2000
  - Paired end 100 bp
  - Raw reads 164,984,147 – 472,265,620
  - Coverage 5x – 17x

Copy Number Variation

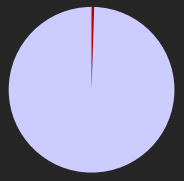
1. Alignment to UMD 3.1 – BWA.MEM
2. Read editing – SAMtools & Picard
3. CNV detection – CNVnator
   - Deletion 200 bp – 724,000 bp



2.71 % – 6.09 %

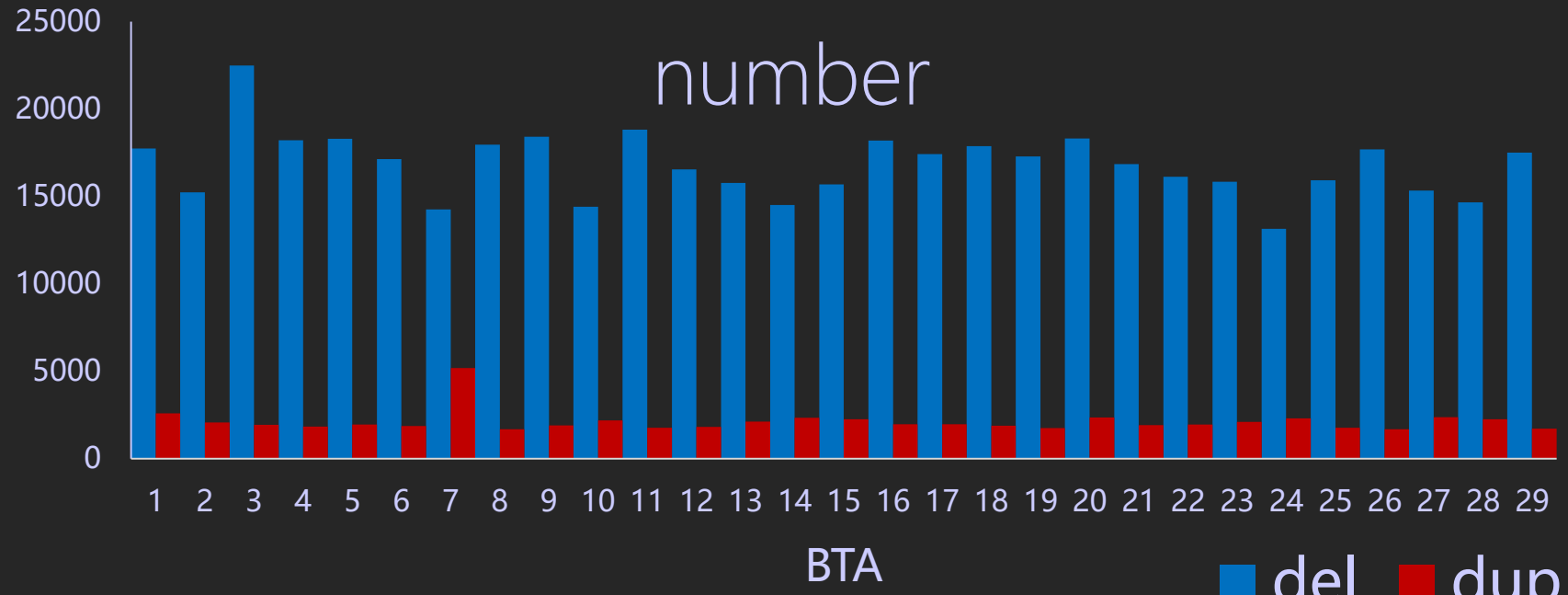


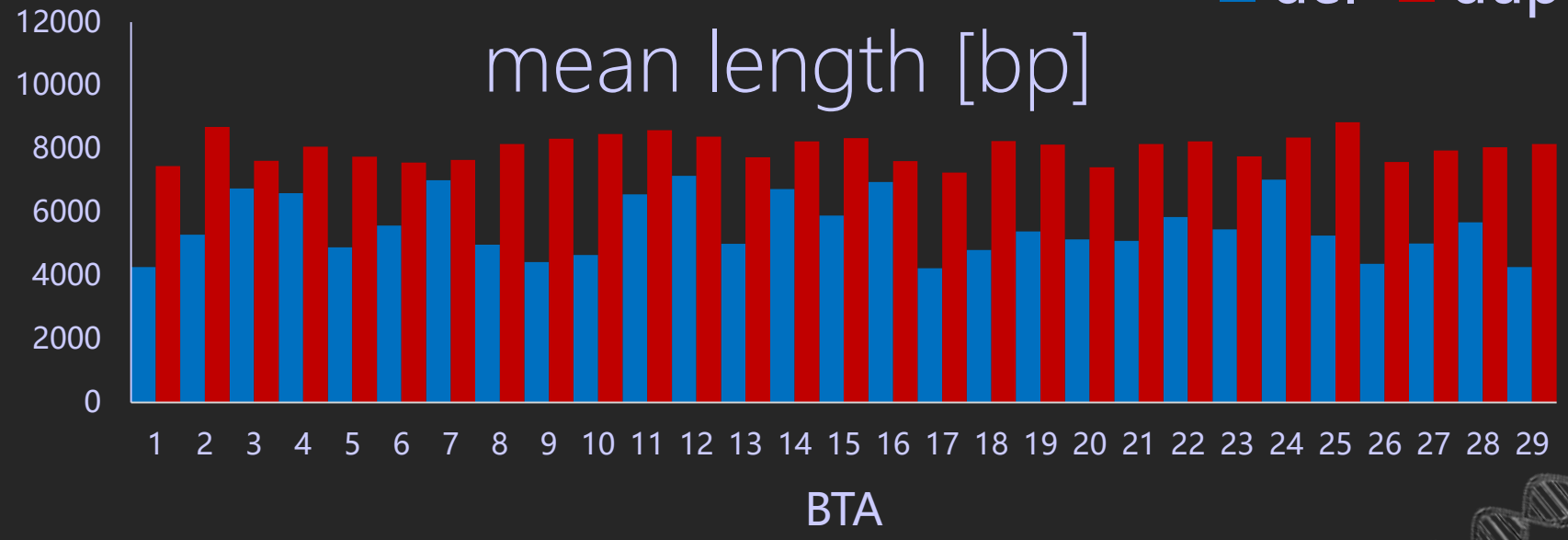
   - Duplication 200 bp – 439,300 bp



0.51 % – 1.58 %

Copy Number Variation

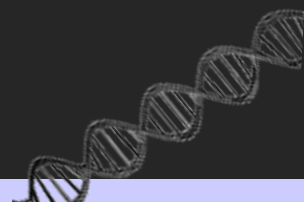




Functional annotation

Genome assembly UMD 3.1 (release 2011)

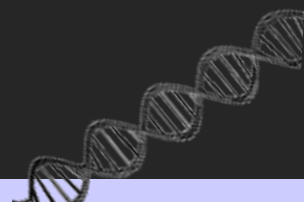
1. Sequence Ontology Terms
   - define features of DNA sequence e.g. transcript ablation, stop gained, frameshift variant, …
   - cumulated to 8 categories: coding, intron, splice region, non-coding transcript, UTR, up/down stream gene, intergenic
2. Gene functional categories
3. All genes



Functional annotation

Genome assembly UMD 3.1 (release 2011)

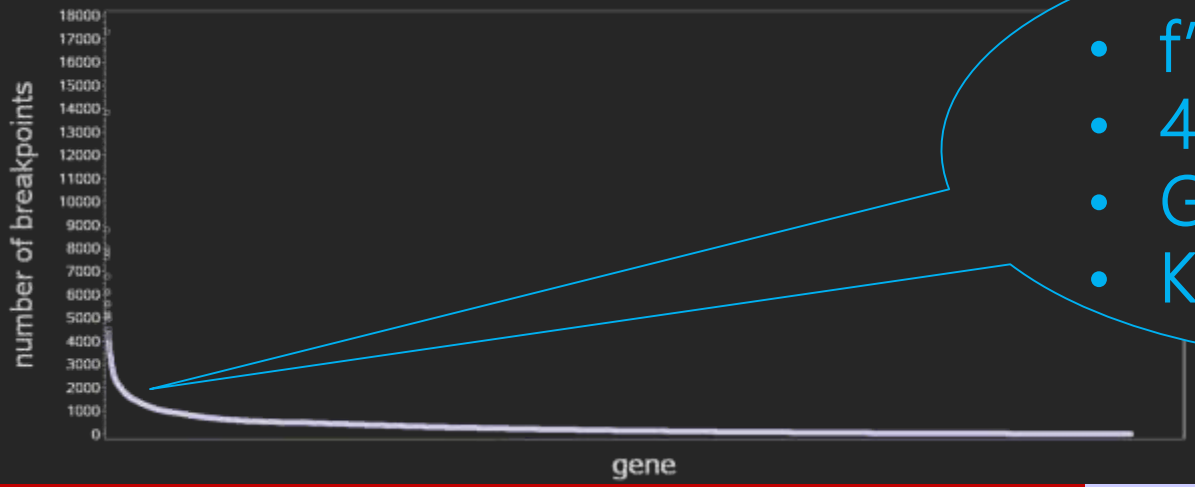
1. Sequence Ontology Terms
2. Gene functional categories
   - housekeeping, under strong selection, neutral
3. All genes



Functional annotation

Genome assembly UMD 3.1 (release 2011)

1. Sequence Ontology Terms
2. Gene functional categories
3. All genes
   - genes with a large number of breakpoints

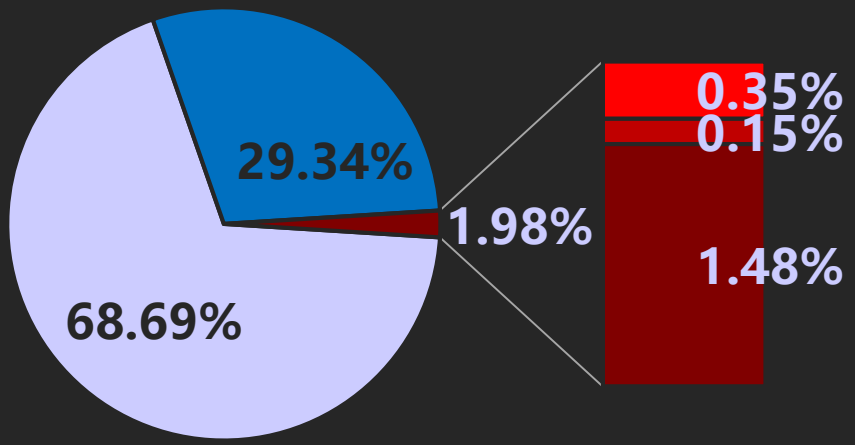


- f'(y=log*(x)) ≥ 1
- 456 genes
- GO
- KEGG

Functional annotation – results

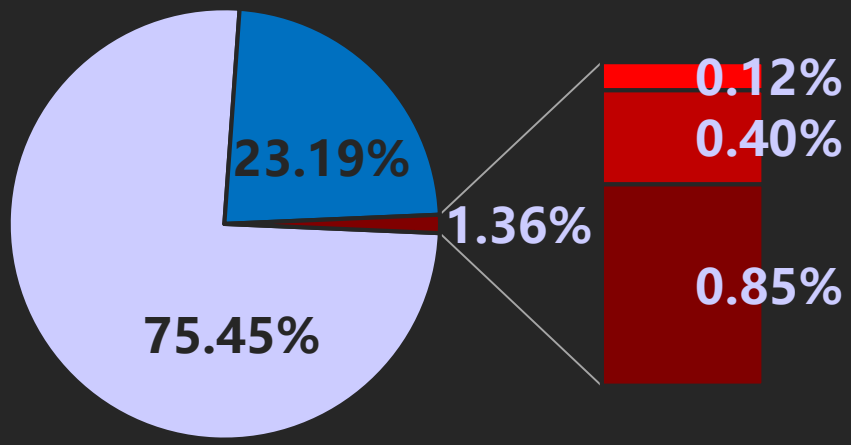
Annotation to SO terms

Deletions (487 746)



- intergenic
- intron
- non-coding regions of transcripts
- splice regions
- coding regions

Duplications (61 789)

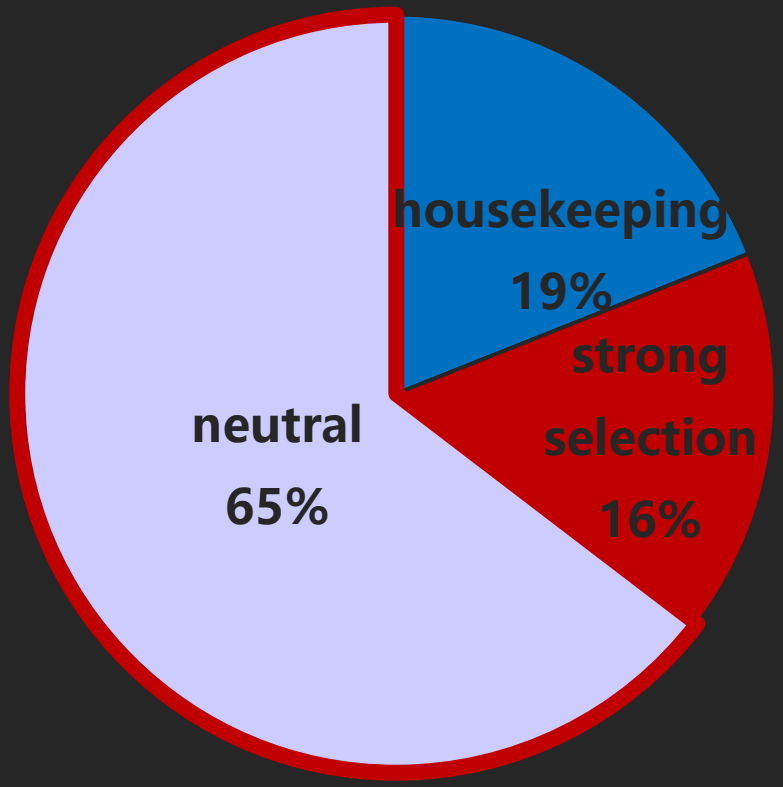


- intergenic
- intron
- non-coding regions of transcripts
- splice regions
- coding regions

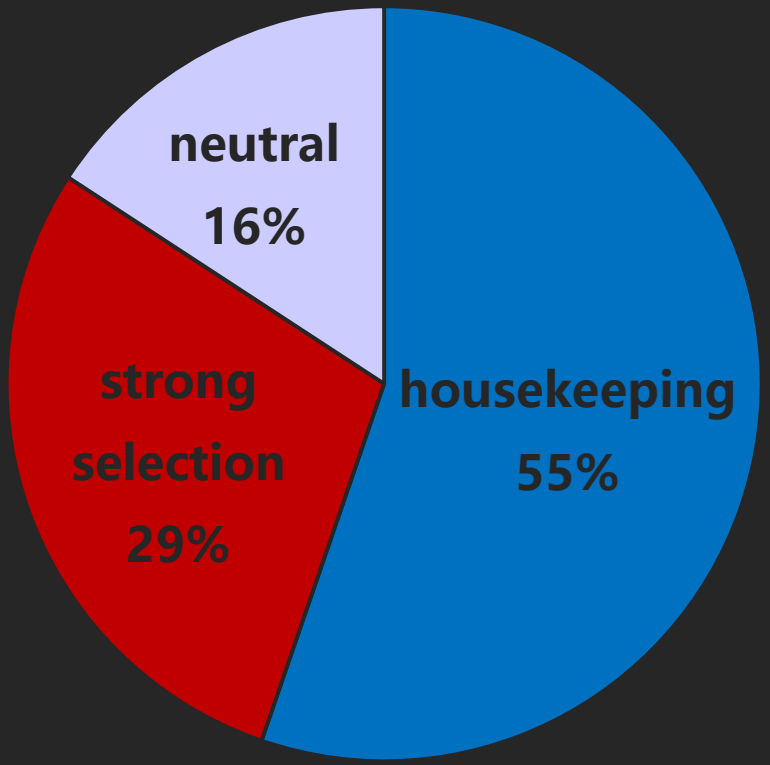
Functional annotation – results

Annotation to gene categories

Deletions (199)



Duplications (38)



Functional annotation – results

Annotation to genes with many breakpoints - KEGG

- "Olfactory transduction"
- significantly underrepresented ($P < 2.83 \cdot 10^{-50}$)

2 vs. 607

genes with many breakpoints

all genes

Functional annotation – results

Annotation to genes with many breakpoints - GO

- 45 GO terms significant with $P<0.05$
- 38 clusters
- Highest enrichment score – regulation of cell adhesion

Conclusions

- Dynamic landscape of CNVs
- Variability across:
  - Individuals
  - Chromosomes
  - Functional genome elements
- CNV formation
  - Supressed in coding regions, but …
  - Deletions → supressed in housekeeping genes and genes under strong selection
  - Duplications → most common in housekeeping genes

Thank you for attention