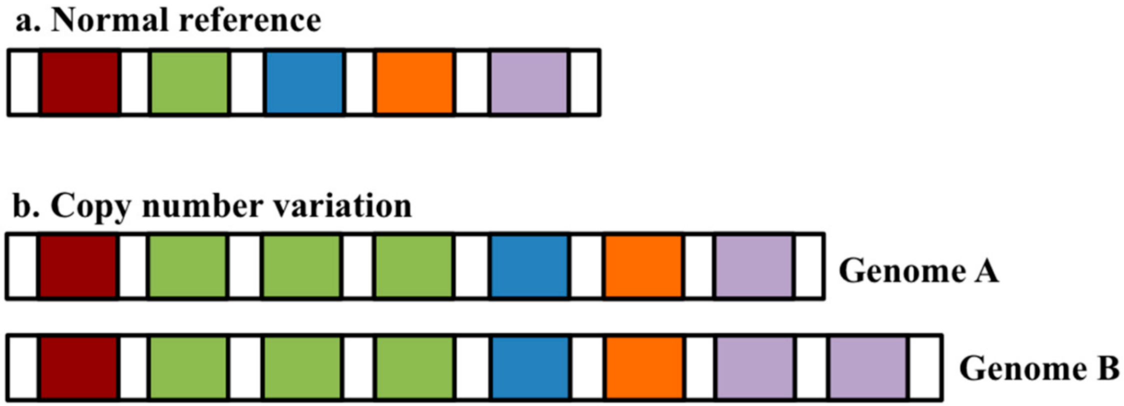
Identification of bovine copy number variants from array and sequence data

A. Butty[1], F. Miglior[1,2], K. Krivushin[3], J. Grant[3], A. Kommadath[3], F. S. Schenkel[1], P. Stothard[3], and C. Baes[1]*

[1]CGIL – University of Guelph, ON; [2]Canadian Dairy Network, Guelph, ON;
[3]AFNS – University of Alberta, Edmonton, AB

Copy Number Variants (CNV)

- Could be important for economically relevant traits
- No consensus on definition or quality criteria in cattle data analysis
- Validation in silico possible: use of multiple detection methods



Adapted from Hurgobin et al., 2017

Objective

- Identify CNVs using array data
  - Can we find the same CNVs in array and sequence data for the same animals?
  - Do they overlap with previously identified CNVs?

CNV Identification - WGS

- Whole-genome re-sequences of 38 Holstein bulls
  - Aligned to UMD3.1 following protocol of the 1,000 Bull Genomes Project
  - Average coverage: 14X
- cn.MOPS, version 3.5 (Klambauer et al., 2012)
  - All samples analyzed simultaneously
  - Read depth approach
  - Identifies only copy number variants

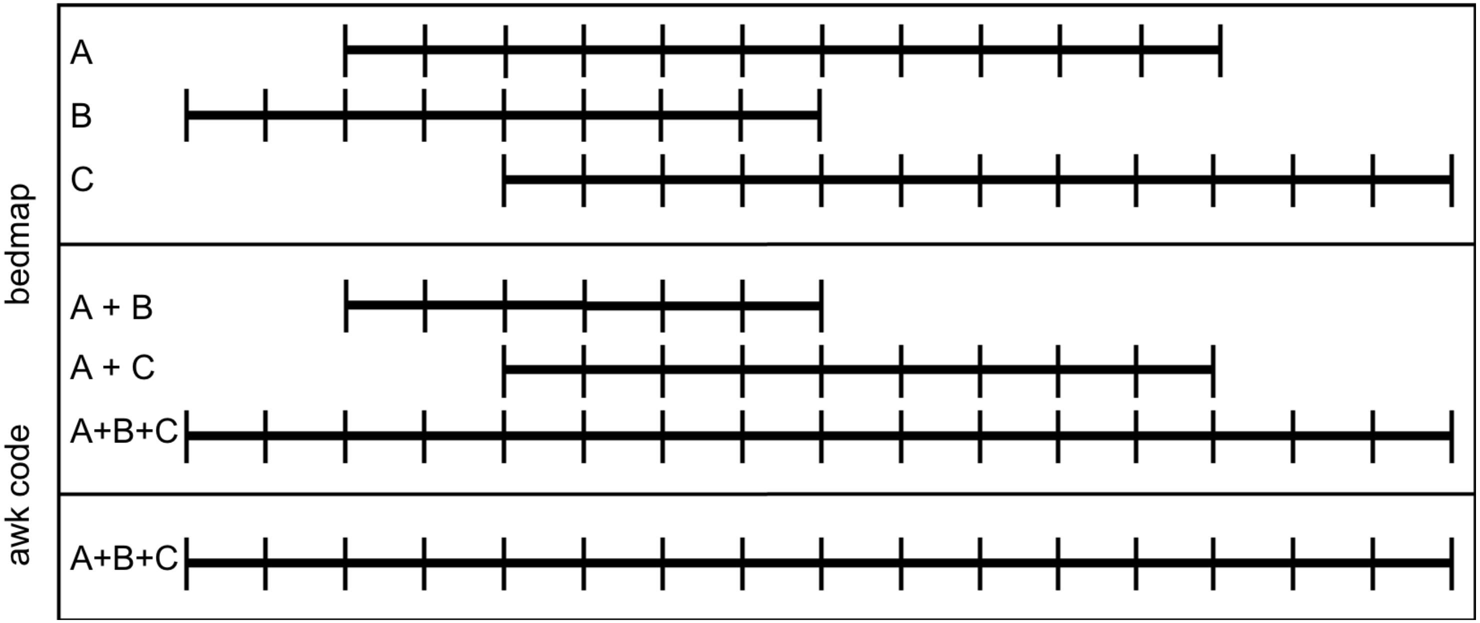
CNV Identification – SNP Arrays

- High-density array genotypes of the same 38 Holstein bulls
  - Information for 777,962 markers per animal

- PennCNV, version 1.0.4 (Wang et al., 2007)
  - One sample at a time
  - Relies on:
    - Genotyping signal intensities
    - B-allele frequency
  - Identifies only copy number variants

CNV Regions

- **CNV regions (CNVR):** Two CNV were considered one region if they had a reciprocal overlap of at least 50% of their lengths
- Redundant information was removed

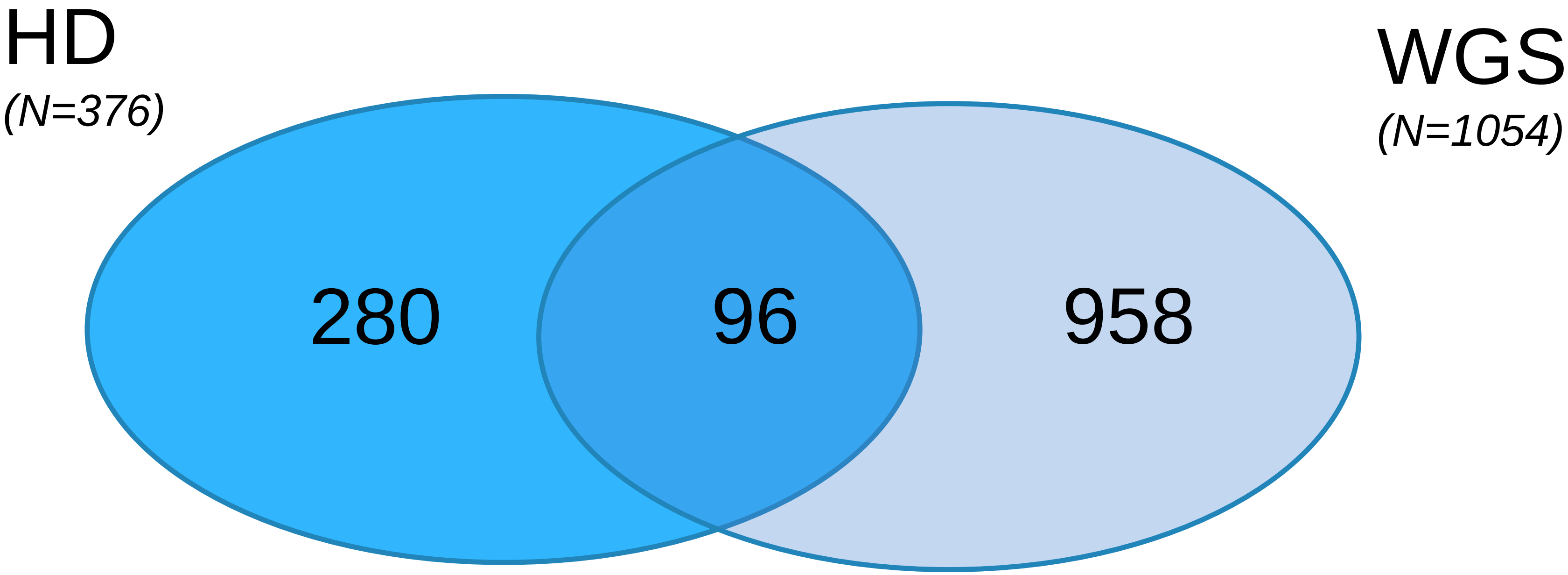


Comparisons

Set	WGS	HD
Number of CNVR	1,054	376
Average length (bp)	204,834	134,378
Genome coverage	7.2 %	1.7 %

- More WGS compared to HD CNVR were private (i.e. only found in one sample)
- Difference in average length was due to map density
- Proportion of genome coverage is low for HD CNVR
  - Potentially impacted by small sample size

CNVR Overlaps



The Genomic Variant Archive (GVa)

- EMBL-EBI database (February 2018)
- CNV datasets from 4 WGS and 4 array studies available

  (Liu et al., 2010; Hou et al., 2011; Bickhart et al., 2012; Hou et al., 2012; Boussaha et al., 2015; Keel et al., 2016; Menzi et al., 2016; Karimi et al., 2017)
- Arrays studies relied on PennCNV
- Only one WGS study used a multi-sample CNV discovery approach (Keel et al,. 2016)

GVa Studies

Study by	Data type	# samples	# breeds
Liu et al, 2010	Array	90	17*
Hou et al, 2011	Array	539	21*
Bickhart et al, 2012	WGS	6	4*
Hou et al, 2012	Array	472	1
Boussaha et al, 2015	WGS	62	3
Keel et al, 2016	WGS	175	20
Menzi et al, 2016	Array / WGS	4	1
Karimi et al, 2017	Array	50	8*

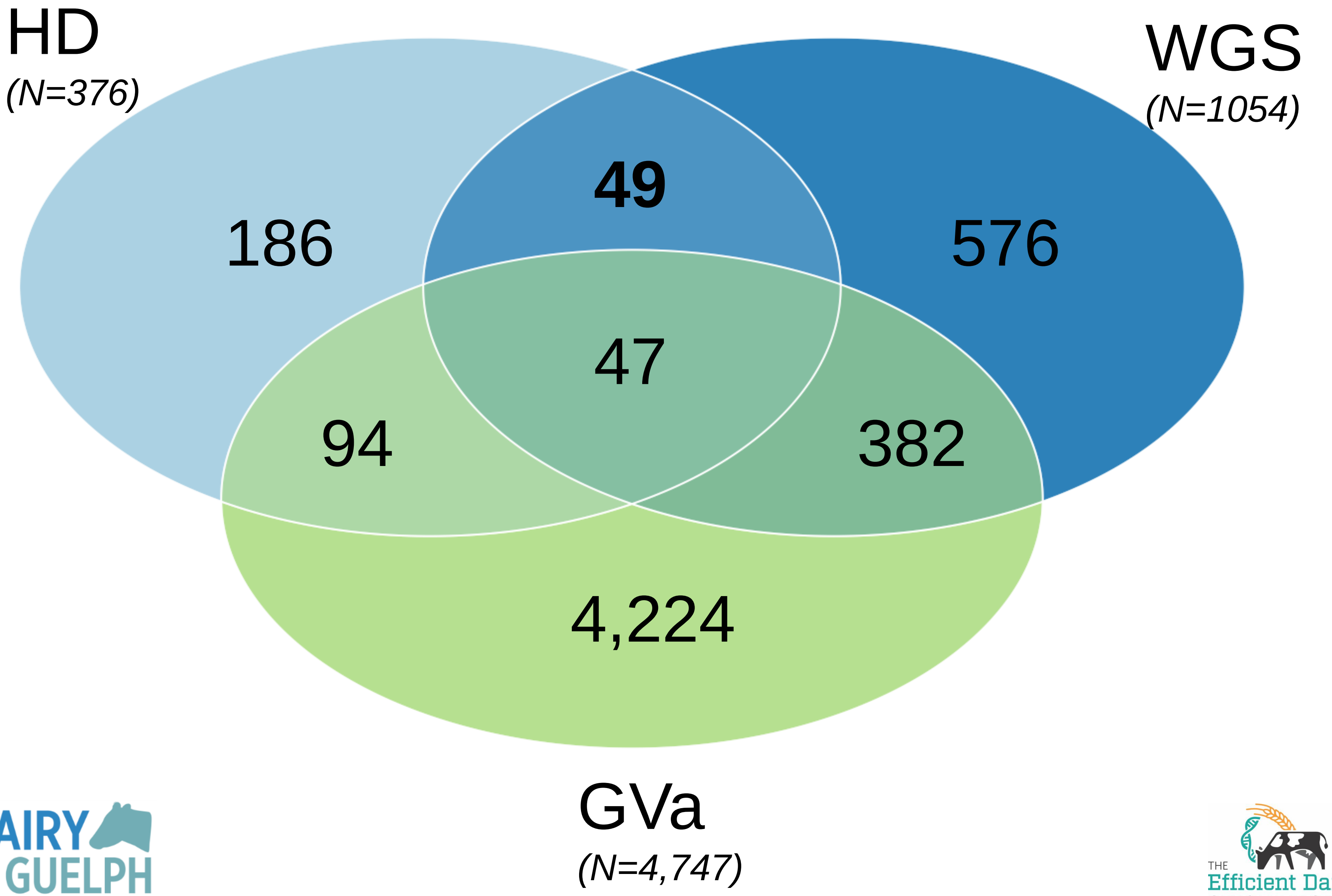
* Bos Indicus animals were also included in the study

Comparisons

Set	WGS	HD	GVa
Number of CNVR	1,054	376	4,747
Average length (bp)	204,834	134,378	181,955
Genome coverage (~)	7.2 %	1.7 %	32.3 %

- High number of CNVR ➔ high genome coverage in GVa
  - probably due to heterogeneous study parameters, mostly to inclusion of indicine breeds

CNVR Overlaps



“New” CNVR

- Average length of “new” CNVR is shorter than the HD, the WGS, or the GVa CNVR (114,037 bp)

- 20% of newly discovered CNVR are on chromosome 12, positions 72.4 Mb - 76.6 Mb
  - Previous studies found high proportion of CNV on this chromosome (e.g. Upadhyay et al., 2017)
  - This BTA is syntenic with human chromosome 13, which is recognized for CNV hotspots (Letaief et al., 2017)

Conclusions

- CNVR identified from SNP arrays and from WGS do not overlap much, even when analyzing the same individuals

- Lack of consistency in CNVR detection is found between analyses

- With less private CNVR and more CNVR validated through other studies, the HD CNVR set appears more accurate than the WGS CNVR set
  - HD is limited compared to WGS, as marker density is low
  - Studies need to be validated with many more samples

Acknowledgements

THE Efficient Dairy GENOME PROJECT

GenomeCanada

GenomeAlberta

Ontario Genomics

Alberta Agriculture and Forestry

Ontario MINISTRY OF RESEARCH & INNOVATION

Ontario Ministry of Agriculture, Food and Rural Affairs

Réseau laitier canadien CDN Canadian Dairy Network

UNIVERSITY of GUELPH

UNIVERSITY OF ALBERTA QUAECUMQUE VERA

GrowSafe SYSTEMS®

ALBERTA milk

USDA

SRUC

QUALITAS

SEMEX® Genetics for Life®

valacta

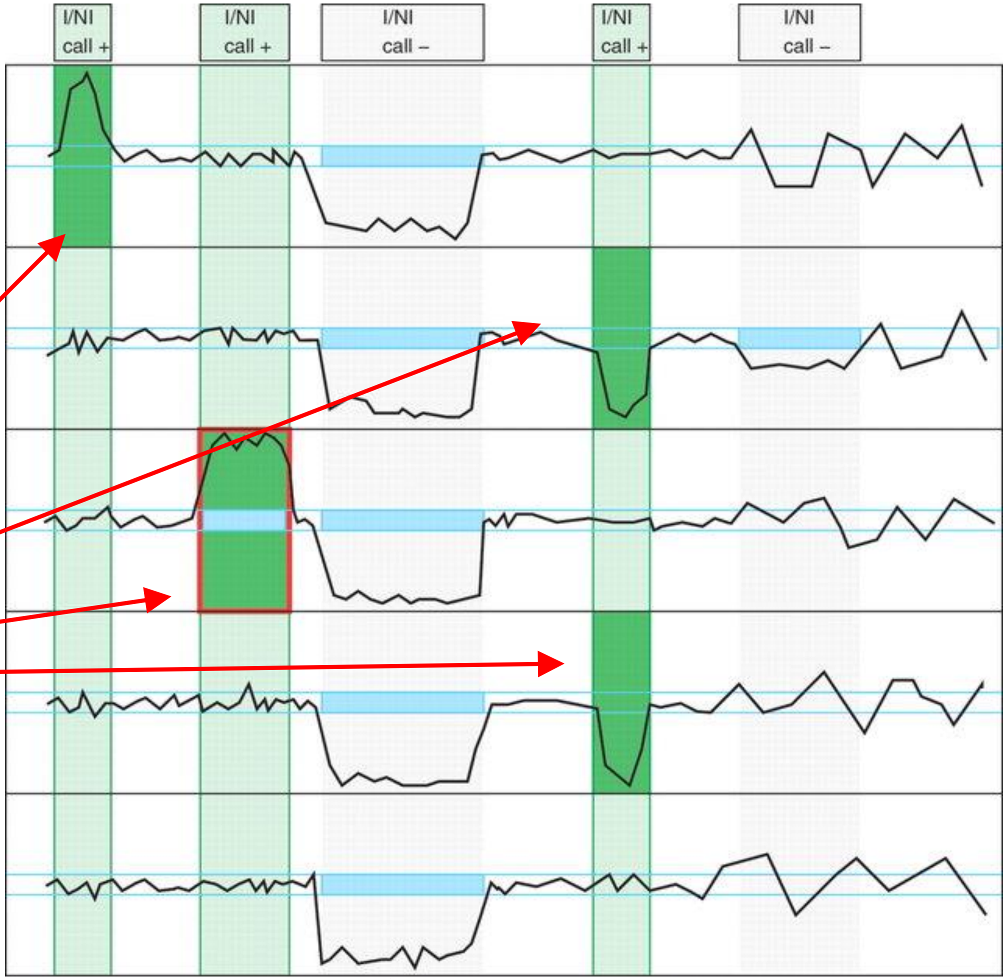
AGRICULTURE VICTORIA

AARHUS UNIVERSITY

CANWEST DHI

cn.MOPS

Cn.MOPS identifies CNV based on read depth variation in segments along each chromosome.
Poisson distributions are applied to disentangle variation of both technical and biological origin.

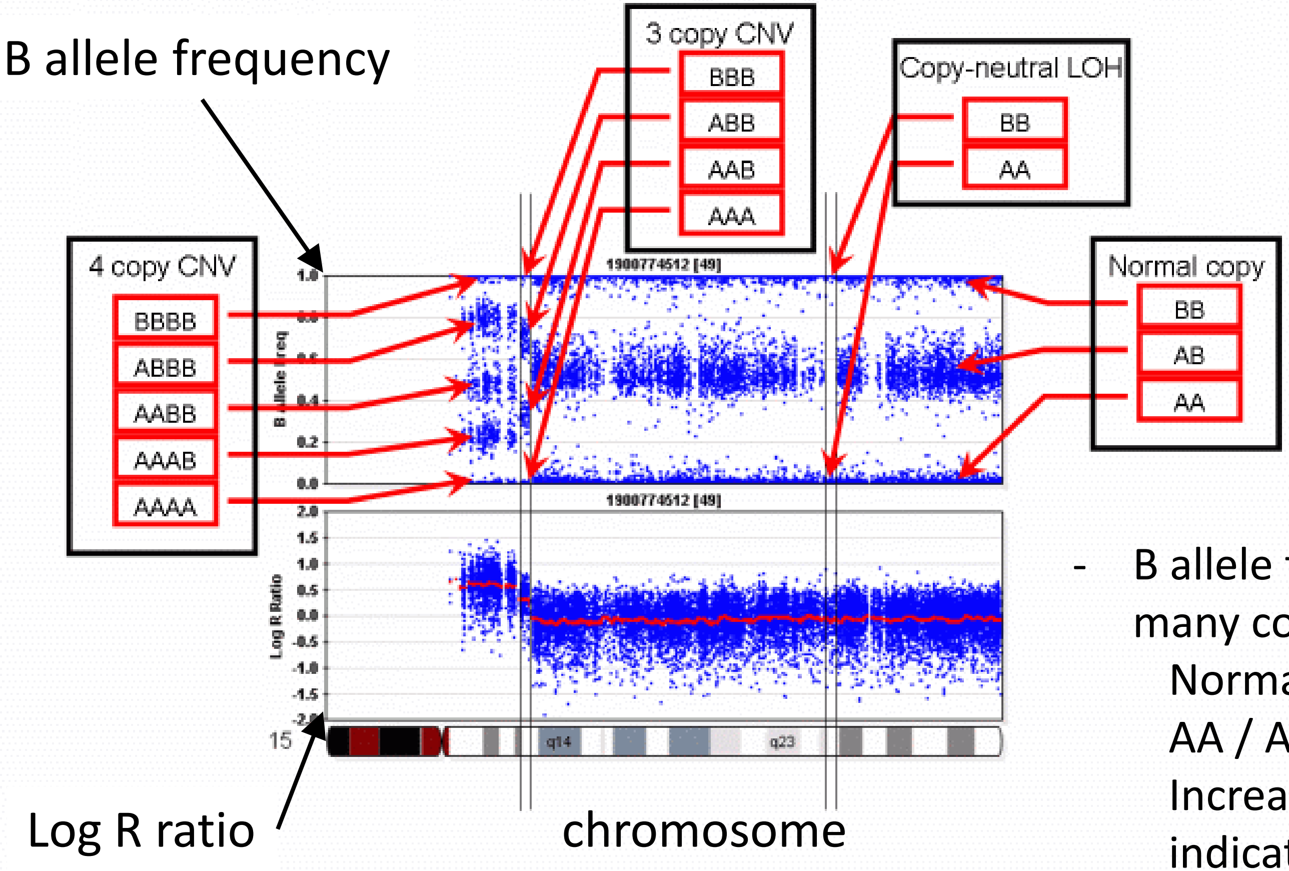


Only some samples show variation in read depth ➔ CNV identified

All samples show the same pattern ➔ no CNV identified

(from Klambauer et al, 2012)

PennCNV



B allele frequency

Log R ratio

chromosome

(from Wang et al, 2007)

PennCNV is based on signal intensities measured at genotyping and observed B allele frequencies:

- Higher Log R ration indicate the presence of genomic material in a sample ➔ possible CNVR
- B allele frequencies indicate how many copies are found for a CNV;
  Normal regions have 3 clusters AA / AB / BB.
  Increased number of clusters indicate CNV regions.
  (4 clusters = CN3, 5 clusters = CN4, etc...)

GVa Studies

Study by	Data type	# samples	# breeds
Liu et al, 2010	Array	90	17*
Hou et al, 2011	Array	539	21*
Bickhart et al, 2012	WGS	6	4*
Hou et al, 2012	Array	472	1
Boussaha et al, 2015	WGS	62	3
Keel et al, 2016	WGS	175	20
Menzi et al, 2016	Array / WGS	4	1
Karimi et al, 2017	Array	50	8*

* Bos Indicus animals were also included in the study