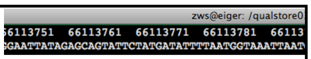
The effect of read depth in whole genome sequencing data

Christine Baes, Peter von Rohr, Marlies Dolezal, Eric Fritz-Waters,
James Koltes, Beat Bapst, Christine Flury, Heidi Signer-Hasler,
Dorian Garrick, Christian Stricker, Birgit Gredler

Outline

- What are SNPs?
- Read depth
- Gaps in the reference genome
- Summary

What are SNPs?

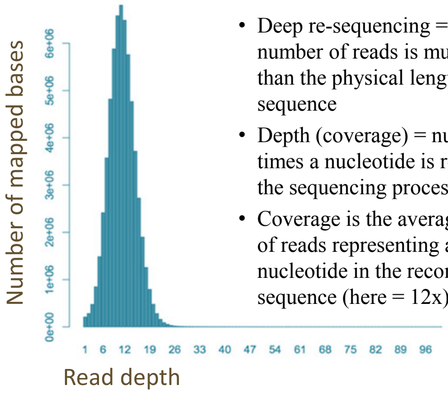


```
zws@eiger: /qualstore0
56113751   66113761   66113771   66113781   66113
GGAATTATAGAGCAGTATTCTATGATATTTTAATGGTAAATTAAT
```

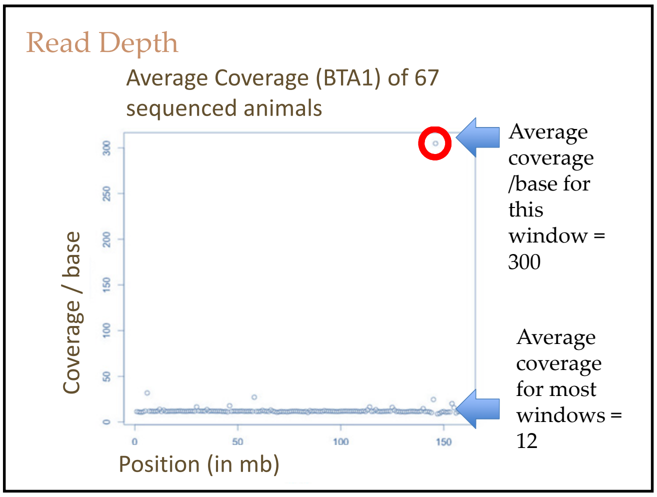
- Reference genome (Dominettes' mosaic)
- Genome of sample split into billions of small pieces of DNA (reads)
- Each read is 101 bases long
- Reads are aligned to Dominettes genome
- Variants are identified if single bases in the reads differ from the reference

BAM = Binary Alignment Mapping File

Read Depth



- Deep re-sequencing = total number of reads is much higher than the physical length of sequence
- Depth (coverage) = number of times a nucleotide is read during the sequencing process
- Coverage is the average number of reads representing a given nucleotide in the reconstructed sequence (here = 12x)

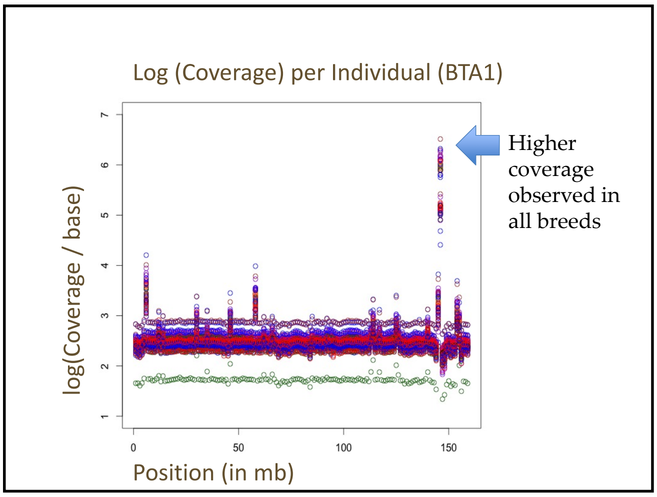


Read Depth

Average Coverage (BTA1) of 67 sequenced animals

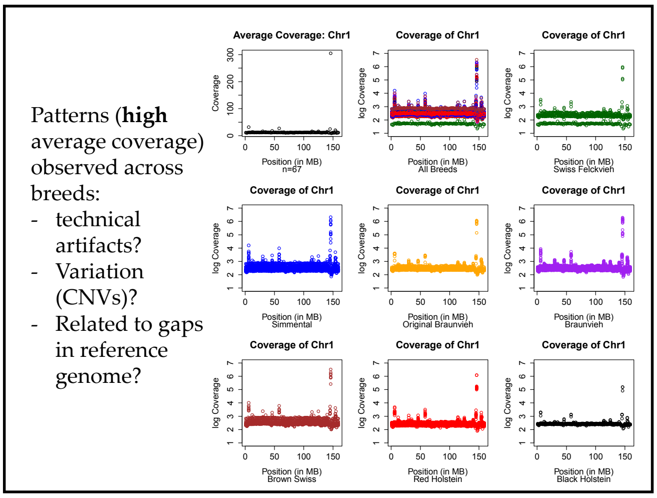
Average coverage /base for this window = 300

Average coverage for most windows = 12



Log (Coverage) per Individual (BTA1)

Higher coverage observed in all breeds

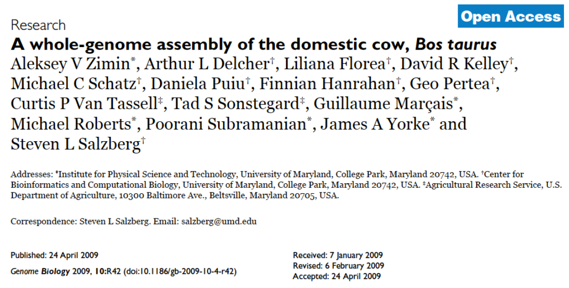


Patterns (**high** average coverage) observed across breeds:

- technical artifacts?
- Variation (CNVs)?
- Related to gaps in reference genome?

...there are gaps in the reference genome?!?

Since 2009, we have considered the genome of the domestic cow “finished“…



Research

Open Access

A whole-genome assembly of the domestic cow, *Bos taurus*

Aleksey V Zimin*, Arthur L Delcher†, Liliana Florea†, David R Kelley†, Michael C Schatz†, Daniela Puiu†, Finnian Hanrahan†, Geo Pertea†, Curtis P Van Tassell‡, Tad S Sonstegard‡, Guillaume Marçais*, Michael Roberts*, Poorani Subramanian*, James A Yorke* and Steven L Salzberg†

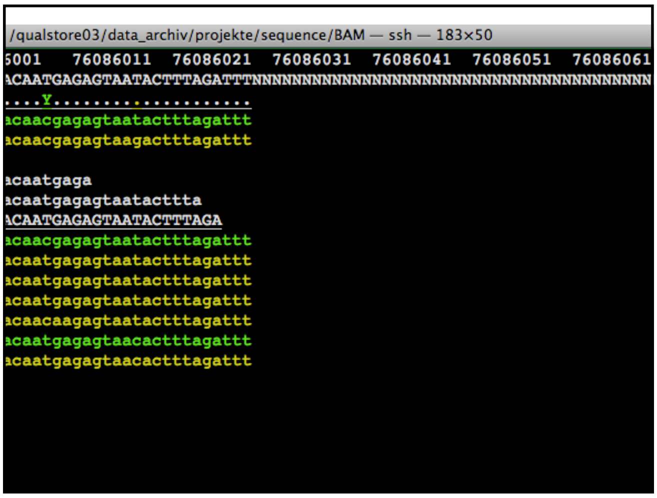
Addresses: *Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA. †Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA. ‡Agricultural Research Service, U.S. Department of Agriculture, 10300 Baltimore Ave., Beltsville, Maryland 20705, USA.

Correspondence: Steven L Salzberg. Email: salzberg@umd.edu

Published: 24 April 2009

Genome ***Biology*** 2009, **10:**R42 (doi:10.1186/gb-2009-10-4-r42)

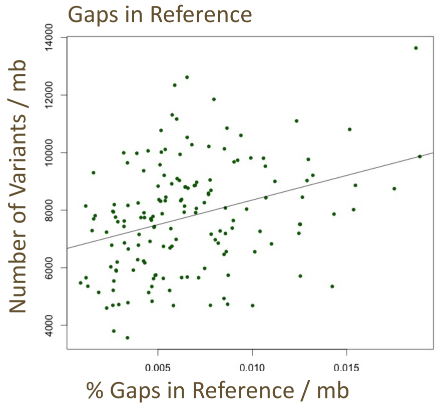
Received: 7 January 2009
Revised: 6 February 2009
Accepted: 24 April 2009



```
/qualstore03/data_archiv/projekte/sequence/BAM — ssh — 183×50
5001      76086011  76086021  76086031  76086041  76086051  76086061
ACAATGAGAGTAATACTTTAGATTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
....Y....................
acaacgagagtaatactttagattt
acaacgagagtaagactttagattt

acaatgaga
acaatgagagtaatacttta
ACAATGAGAGTAATACTTTAGA
acaacgagagtaatactttagattt
acaatgagagtaatactttagattt
acaatgagagtaatactttagattt
acaatgagagtaatactttagattt
acaacaagagtaatactttagattt
acaatgagagtaacactttagattt
acaatgagagtaacactttagattt
```

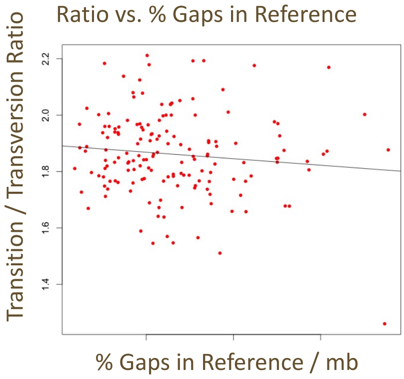
Number of Variants Identified vs. Gaps in Reference



The number of variants found is positively correlated with the number of gaps in the reference.

(r = 0.33)

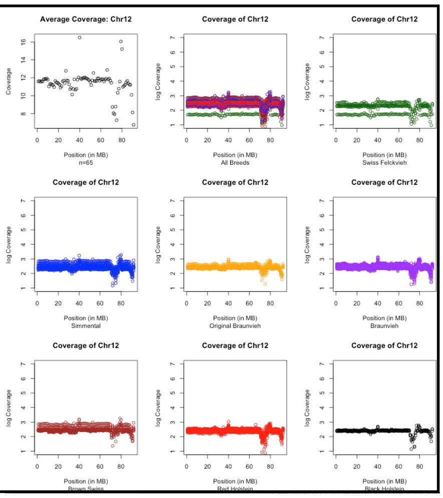
Average Transition / Transversion Ratio vs. % Gaps in Reference



The quality of variants is negatively correlated with the number of gaps in the reference.

(r = -0.11)

Low average coverage is more clearly related to gaps in the reference genome (e.g. BTA12).



Summary

1. Average coverage varies across chromosomes
2. Patterns observed across breeds:
   - technical artifacts
   - Variation (CNVs)
   - Related to gaps in reference genome
3. Number of variants found is positively correlated with the number of gaps in the reference
4. The quality of variants is negatively correlated with the number of gaps in the reference
5. We need a new reference (preferably one for each breed)

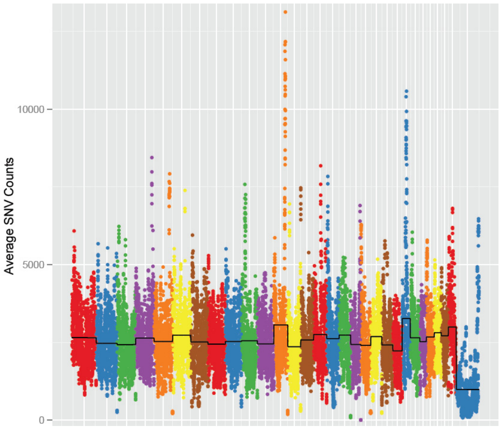
KTI/CTI

ASR
Arbeitsgemeinschaft Schweizerischer Rinderzüchter
Communauté de travail des éleveurs bovins suisses

Financial assistance from the Swiss Commission for Technology and Innovation and the ASR is gratefully acknowledged.

What to expect:

1. Not all SNPs are created equally:
   - SNP arrays are great (call rate > 98%)
   - Sequencing data is also great, but varies in read depth & quality for many, many, many positions
2. Rapidly evolving technology (Batch effects, software, etc.)
3. Expect many ugly variants
   - Array variants are pre-selected (minimum MAF, found in many populations, etc)
   - Sequencing reveals variants across the entire allele frequency spectrum
4. Variant QC is an art
   - The genotype table from a GWAS with array data is a beautiful thing…



Large variation in the average number of SNP found across the genome