Distributed bioinformatics analyses on an SGE cluster, for variant calling on bovine whole-genome samples

Alexandru E. Mizeranschi

Research and Development Station for Bovine, Arad

Short intro to genomics

- Human Genome Project completed in 2003
- The **base letter (A, G, C, T) sequences** of (most of) the chromosomes of an organism (reference genome)
- Reference genomes are becoming available for more and more species
- The 23 human chromosomes consist of more than 3 billion base pairs
- Each human cell (around 10 µm in diameter) contains approx. 2 meters of DNA
- The “central dogma of biology”:
  genes -> mRNA -> proteins

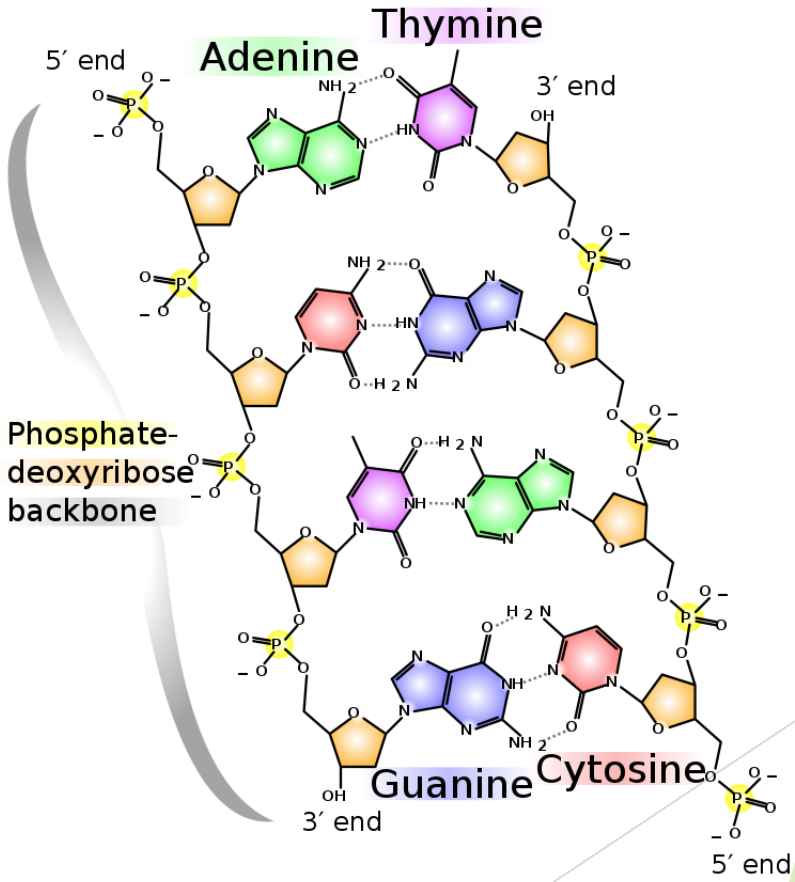


Image source: https://en.wikipedia.org/wiki/File:DNA_chemical_structure.svg

Gene transcription into mRNA

- A gene consists of:
  - A promoter, where the transcription machinery assembles
  - **Coding regions** (exons), which are later used as a blueprint for proteins
  - **Noncoding regions** (introns)
- Introns are removed from the mRNA sequence
- Alternative splicing

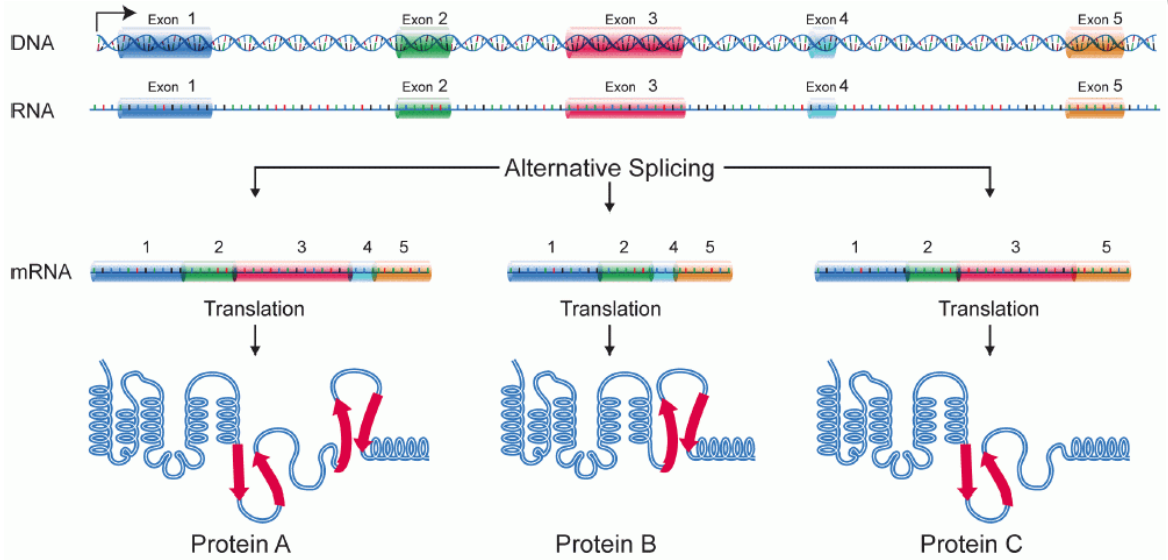
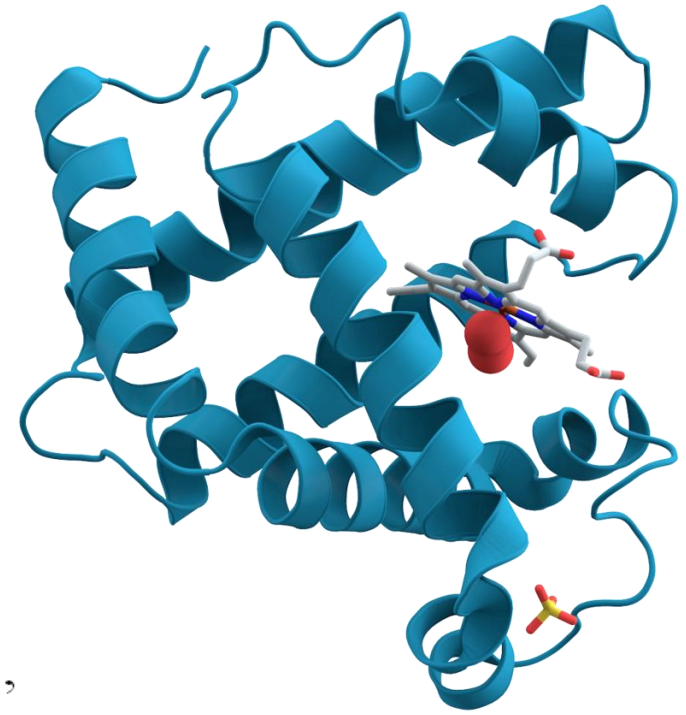


Image source: https://en.wikipedia.org/wiki/File:DNA_alternative_splicing.gif

mRNA translation into protein

- **Triplets of base pairs** are called **codons**
- Codons encode one of **20 amino acids** (AA)
- [AGCT]* -> [20 AA]*
- Translated AA chains then fold into 3D-shaped proteins, as **determined by the physical properties** of the AA
- mRNA sequences can be “read” in three ways (called **reading frames**) in a single direction



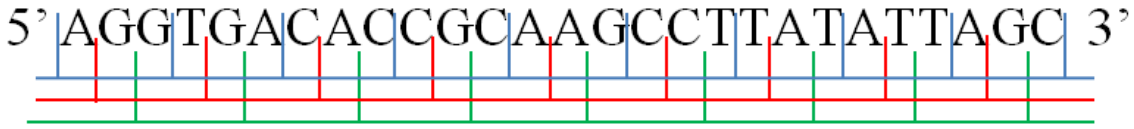


Image source (top right): https://en.wikipedia.org/wiki/File:Myoglobin.png
Image source (bottom): https://en.wikipedia.org/wiki/File:Reading_Frame.png

Codon table

Amino acids biochemical properties	nonpolar	polar	basic	acidic

Termination: stop codon

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)[B]	TGA	Stop (*Opal*)[B]	A
	TTG[A]		TCG		TAG	Stop (*Amber*)[B]	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG[A]		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

Source: https://en.wikipedia.org/wiki/DNA_codon_table

Genetic mutations

- Mutations are defined **relative to a reference genome sequence**
  - SNPs (single nucleotide polymorphisms)
  - Indels (insertions or deletions)
- Mutations are usually important when they fall within the coding region of a gene and **change a codon** into one representing a different AA, which results in a **change to the 3D protein structure**
- Genome-wide association studies (GWAS): statistically associate SNPs with measurable traits (phenotypes), such as disease resistance or risk

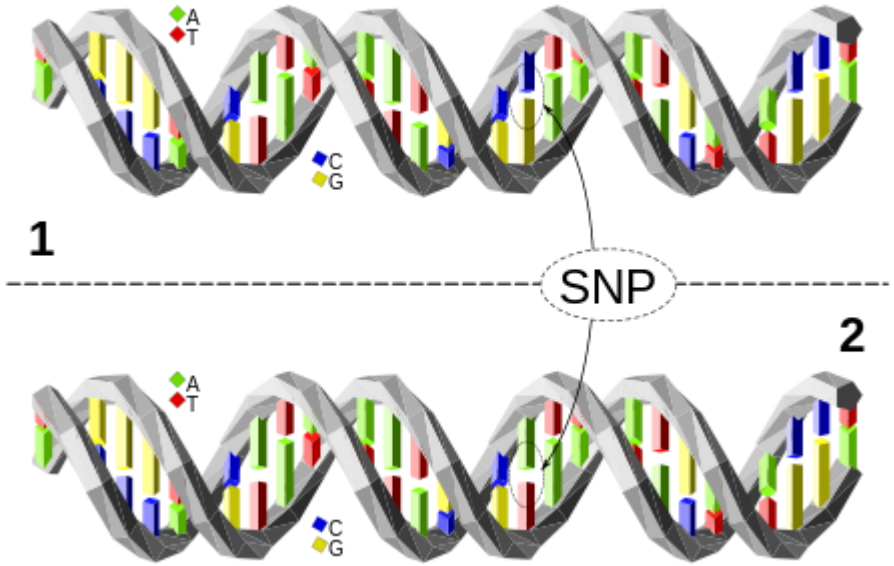


Image source: https://en.wikipedia.org/wiki/File:Dna-SNP.svg

Genetic mutations

- GWAS require **large sample sizes** (from hundreds or thousands to millions)
- GWAS results are visualized using Manhattan plots

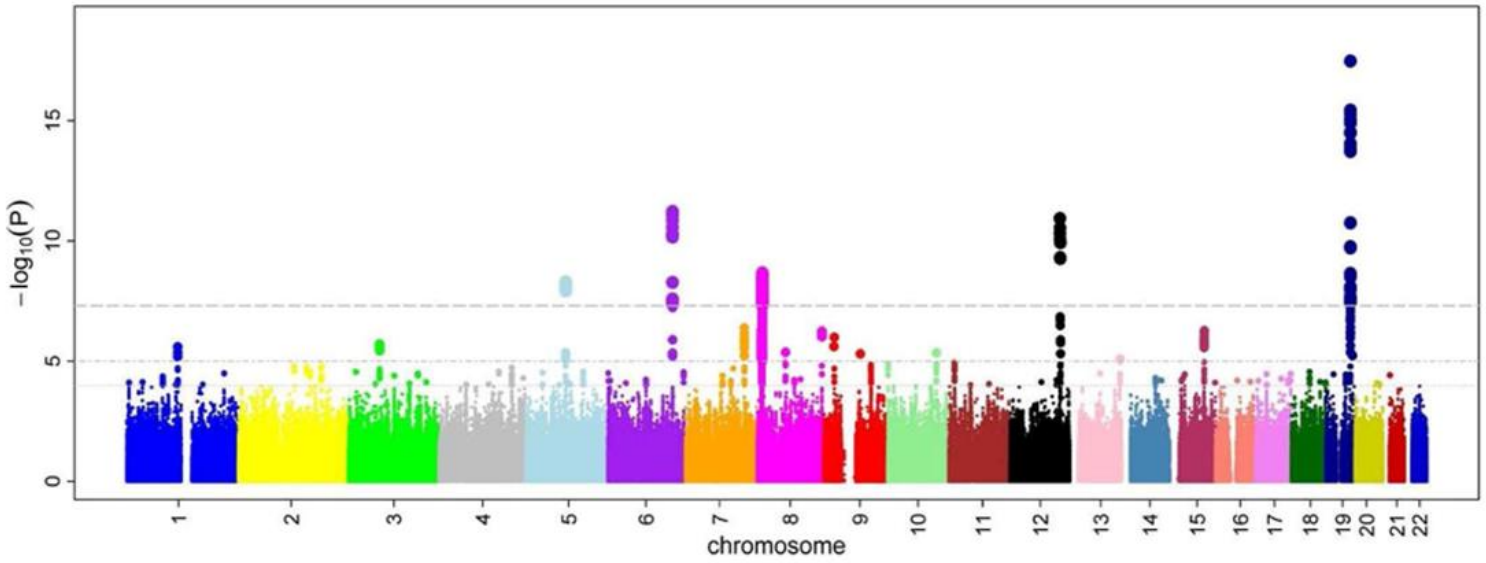


Image source: https://en.wikipedia.org/wiki/File:Manhattan_Plot.png

Next-generation sequencing (NGS)

- DNA is extracted from cells, broken into pieces of approx. equal size and sequenced, one base at a time
- **High-throughput**, **massively-parallel** DNA sequencing
- The result is a **text file** in the FASTQ format, containing millions of sequence reads. For each read, there are four lines in the FASTQ file:
  - `@SEQ_ID`
  - `GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT`
  - `+`
  - `!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65`
- The final row contains ASCII-coded quality values

NGS data analysis

- Data analysis **pipelines** – use Unix pipes to avoid writing data to disk as much as possible
  - Instead of: dataFile2 = tool1 dataFile1; dataFile3 = tool2 dataFile2
  - Write: dataFile3 = tool1 dataFile1 | tool2
- Data analysis steps:
  - Quality control (QC)
  - **Read mapping** to reference genome
  - **Variant calling**
  - Functional variant annotation (influence on translated AA sequence)
- Challenges:
  - Storage: **data volumes** are increasing year by year;
  - Computational complexity: **not all analysis steps can be parallelized**; HPC is required
  - Interpretation of results: experts are required

Computational experiments

- Variants were called for **40 Bos Taurus samples** obtained from the 1000 Bull Genomes project, relative to the UMD3.1 (bosTau6) reference genome
- The **Bcbio-nextgen pipeline** was run on an **SGE cluster** maintained by the Politehnica University of Bucharest
- Analyses were set up using 2, 4, 6, 8 or 10 cluster nodes (32 CPU cores each)
- The **wall-time** was recorded for read mapping, variant calling, as well as the total runtime
- The input data set was also reduced to 20 and 10 samples, while maintaining **the same average read depth** (coverage) as for the original set of 40 samples
- Read coverage is: (ReadLength · NumReads) / GenomeLength

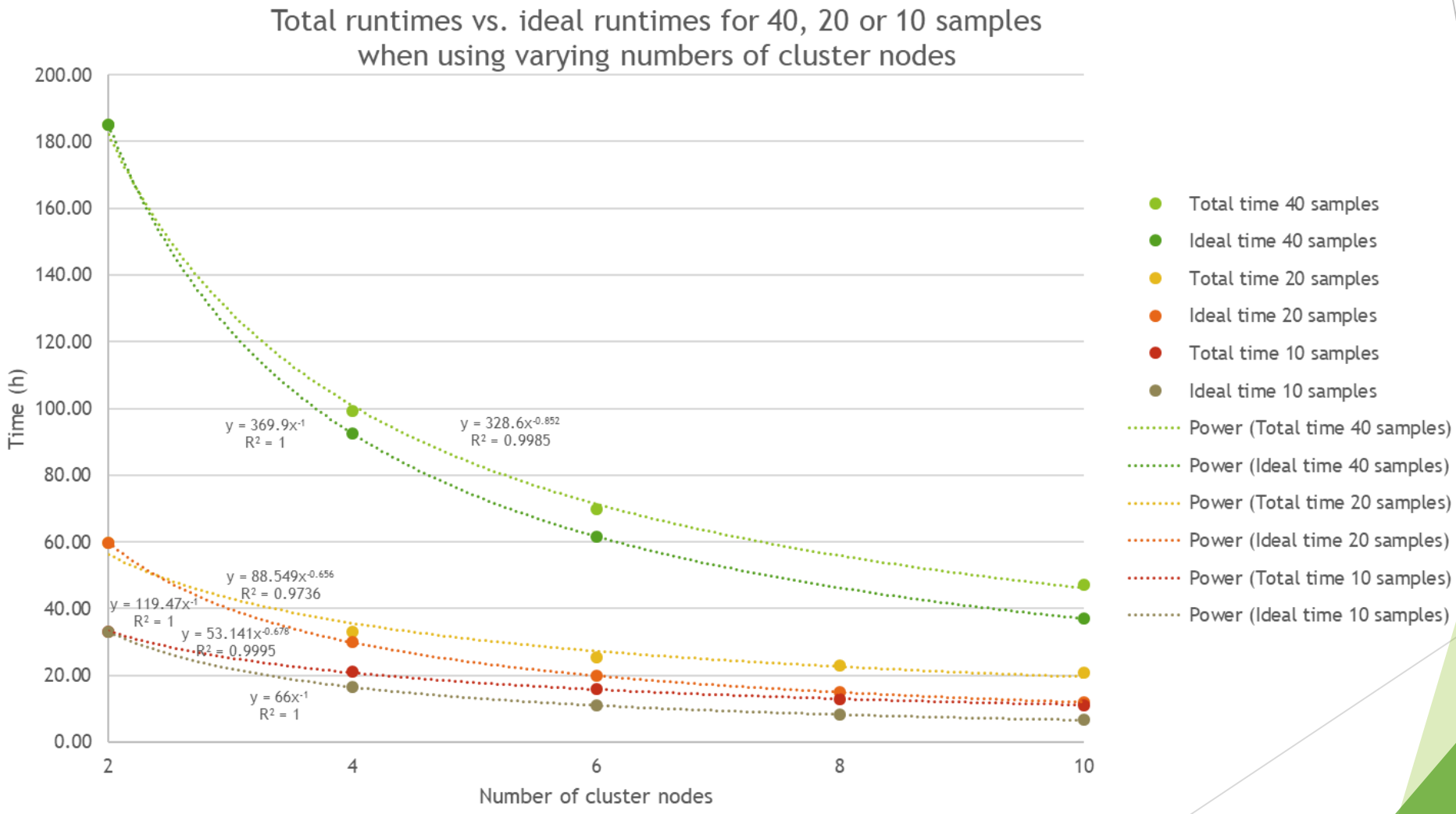
Bcbio-nextgen home page: https://bcbio-nextgen.readthedocs.io/en/latest/

Computational experiments

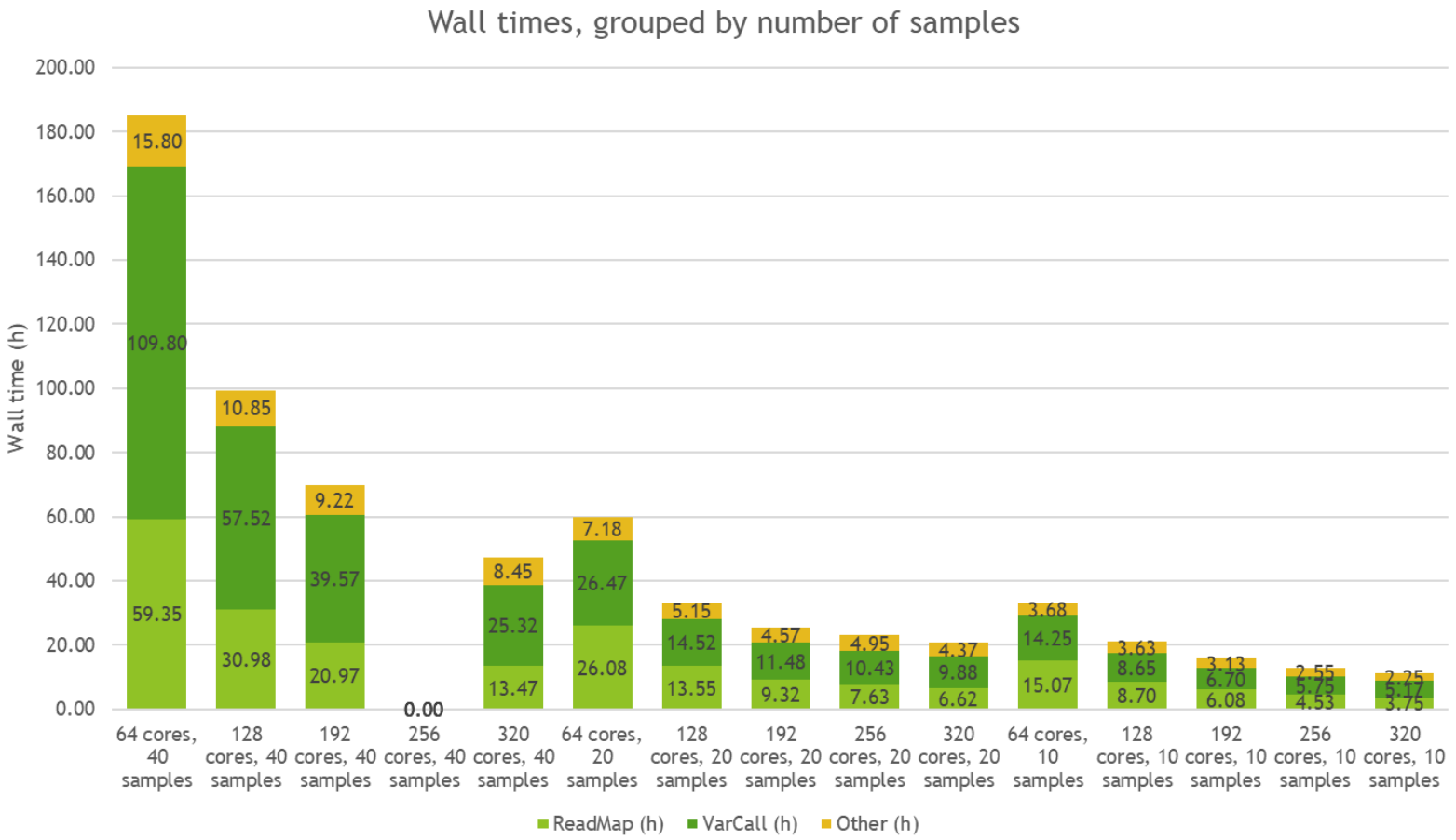
- The complexity of the cow genome is comparable to that of humans in terms of total genome length and estimated number of protein-coding genes
- The 40-sample analyses occupied approx. **3.6 TB of storage space** in total (not including the input data), which is almost 20% of the total of 19 TB available on the cluster
- The 20- and 10-sample analyses occupied approx. 1.7 TB and 900 GB, respectively (not including the input data)
- The compressed raw sequence data (FASTQ format) for the 40 samples occupied **814 GB of storage**

Results

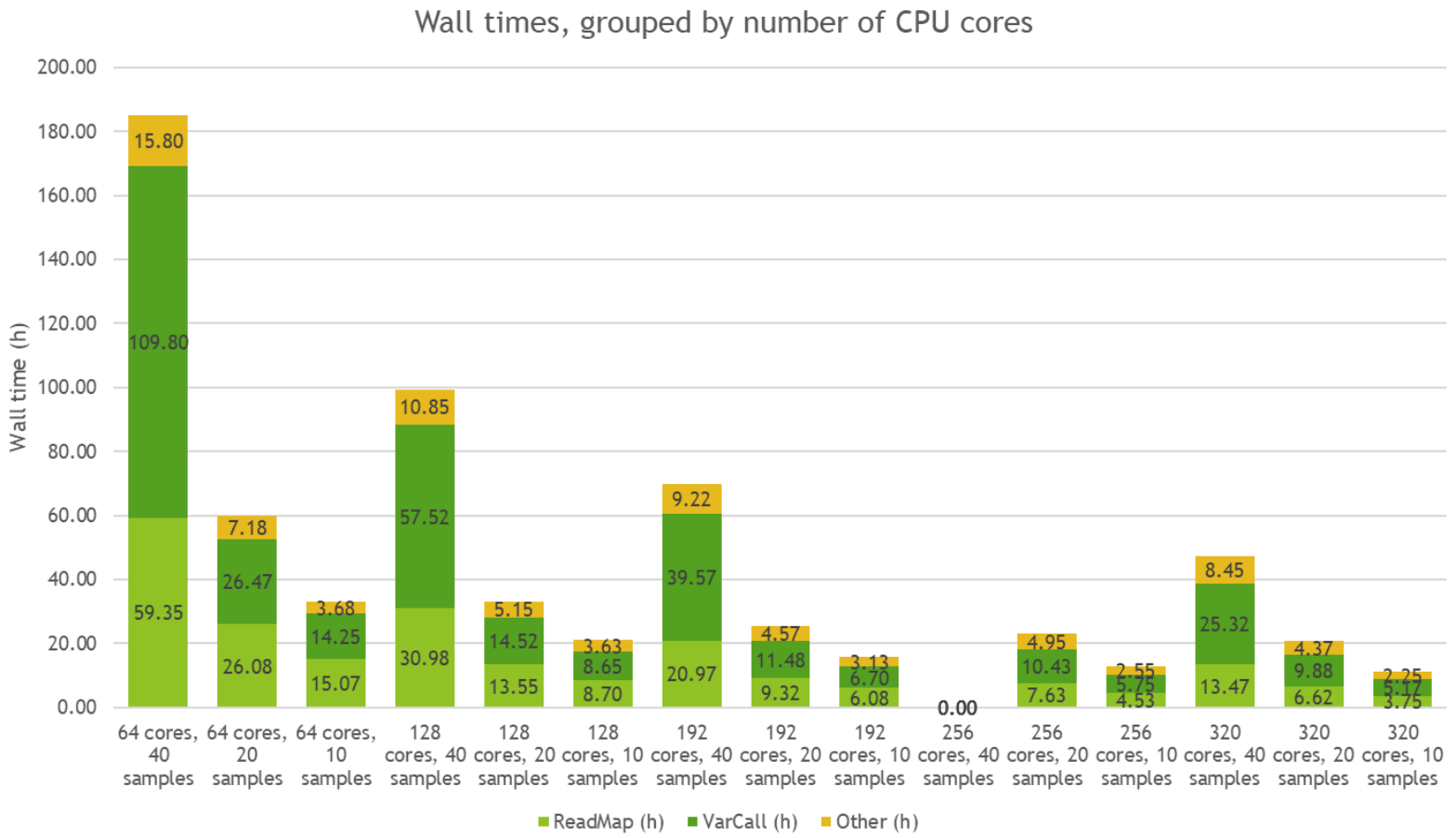
Samples	Cluster nodes	CPU cores	Total time (h)	Ideal time (h)	ReadMap (h)	VarCall (h)	Other (h)
40	2	64	184.95	184.95	59.35	109.80	15.80
40	4	128	99.35	92.48	30.98	57.52	10.85
40	6	192	69.75	61.65	20.97	39.57	9.22
40	8	256	0.00	46.24	0.00	0.00	0.00
40	10	320	47.23	36.99	13.47	25.32	8.45
20	2	64	59.73	59.73	26.08	26.47	7.18
20	4	128	33.22	29.87	13.55	14.52	5.15
20	6	192	25.37	19.91	9.32	11.48	4.57
20	8	256	23.02	14.93	7.63	10.43	4.95
20	10	320	20.87	11.95	6.62	9.88	4.37
10	2	64	33.00	33.00	15.07	14.25	3.68
10	4	128	20.98	16.50	8.70	8.65	3.63
10	6	192	15.92	11.00	6.08	6.70	3.13
10	8	256	12.83	8.25	4.53	5.75	2.55
10	10	320	11.17	6.60	3.75	5.17	2.25


Total runtimes vs. ideal runtimes for 40, 20 or 10 samples when using varying numbers of cluster nodes
Time (h)
200.00
180.00
160.00
140.00
120.00
100.00
80.00
60.00
40.00
20.00
0.00
2
4
6
8
10
Number of cluster nodes
y = 369.9x-1
R² = 1
y = 328.6x-0.852
R² = 0.9985
y = 88.549x-0.656
R² = 0.9736
y = 119.47x-1
R² = 1
y = 53.141x-0.678
R² = 0.9995
y = 66x-1
R² = 1
Total time 40 samples
Ideal time 40 samples
Total time 20 samples
Ideal time 20 samples
Total time 10 samples
Ideal time 10 samples
Power (Total time 40 samples)
Power (Ideal time 40 samples)
Power (Total time 20 samples)
Power (Ideal time 20 samples)
Power (Total time 10 samples)
Power (Ideal time 10 samples)

Results


Wall times, grouped by number of samples
Wall time (h)
200.00
180.00
160.00
140.00
120.00
100.00
80.00
60.00
40.00
20.00
0.00
15.80
109.80
59.35
10.85
57.52
30.98
9.22
39.57
20.97
0.00
8.45
25.32
13.47
7.18
26.47
26.08
5.15
14.52
13.55
4.57
11.48
9.32
4.95
10.43
7.63
4.37
9.88
6.62
3.68
14.25
15.07
3.63
8.65
8.70
3.13
6.70
6.08
2.55
5.75
4.53
2.25
5.47
3.75
64 cores, 40 samples
128 cores, 40 samples
192 cores, 40 samples
256 cores, 40 samples
320 cores, 40 samples
64 cores, 20 samples
128 cores, 20 samples
192 cores, 20 samples
256 cores, 20 samples
320 cores, 20 samples
64 cores, 10 samples
128 cores, 10 samples
192 cores, 10 samples
256 cores, 10 samples
320 cores, 10 samples
ReadMap (h)
VarCall (h)
Other (h)

Results


Wall times, grouped by number of CPU cores
Wall time (h)
200.00
180.00
160.00
140.00
120.00
100.00
80.00
60.00
40.00
20.00
0.00
15.80
109.80
59.35
7.18
26.47
26.08
3.68
14.25
15.07
10.85
57.52
30.98
5.15
14.52
13.55
3.63
8.65
8.70
9.22
39.57
20.97
4.57
11.48
9.32
3.13
6.70
6.08
0.00
4.95
10.43
7.63
2.55
5.75
4.53
8.45
25.32
13.47
4.37
9.88
6.62
2.25
5.47
3.75
64 cores, 40 samples
64 cores, 20 samples
64 cores, 10 samples
128 cores, 40 samples
128 cores, 20 samples
128 cores, 10 samples
192 cores, 40 samples
192 cores, 20 samples
192 cores, 10 samples
256 cores, 40 samples
256 cores, 20 samples
256 cores, 10 samples
320 cores, 40 samples
320 cores, 20 samples
320 cores, 10 samples
ReadMap (h)
VarCall (h)
Other (h)

MC1R gene: controls the production of black and red pigments

- In Bos Taurus, the MC1R gene encodes the protein melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor), involved in the production of melatonin, which **influences the skin and hair color**
- Two versions (alleles) of the MC1R gene are known:
  - E^D, contains a **codon-changing SNP**: g.14757910T>C and is associated with **black hair** in different breeds such as the Holstein breed
  - e, contains a **frameshift single-base deletion**: g.14757924delG and is associated with **red hair** in breeds such as the Fleckvieh/Simmental breed

Source: https://omia.org/OMIA001199/9913/
Image source 1: https://en.wikipedia.org/wiki/Holstein_Friesian_cattle#/media/File:Cow_female_black_white.jpg
Image source 2: https://upload.wikimedia.org/wikipedia/commons/c/cb/Schecken_in_Rettenbach.JPG

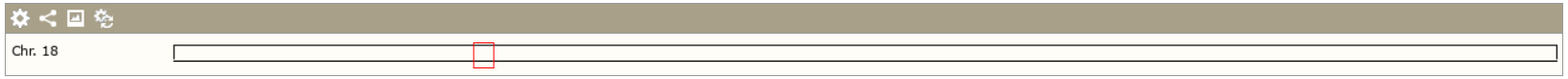
Data visualized at a specific region inside the MC1R gene

Holstein

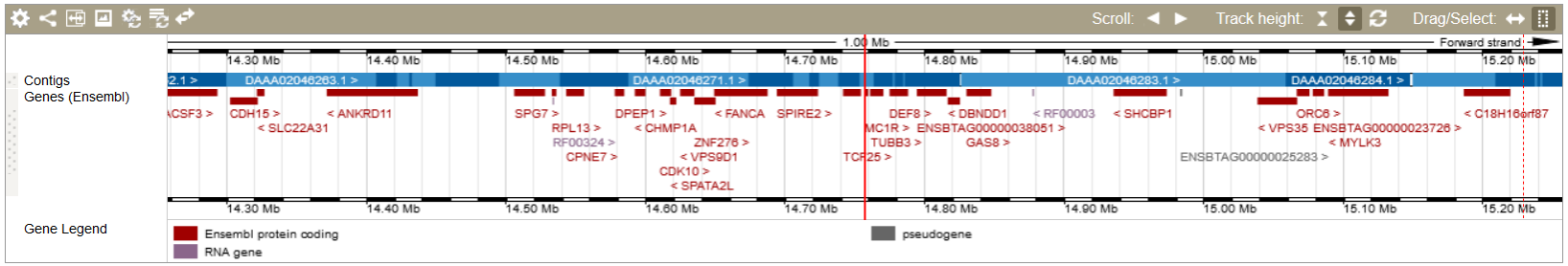
Simmental

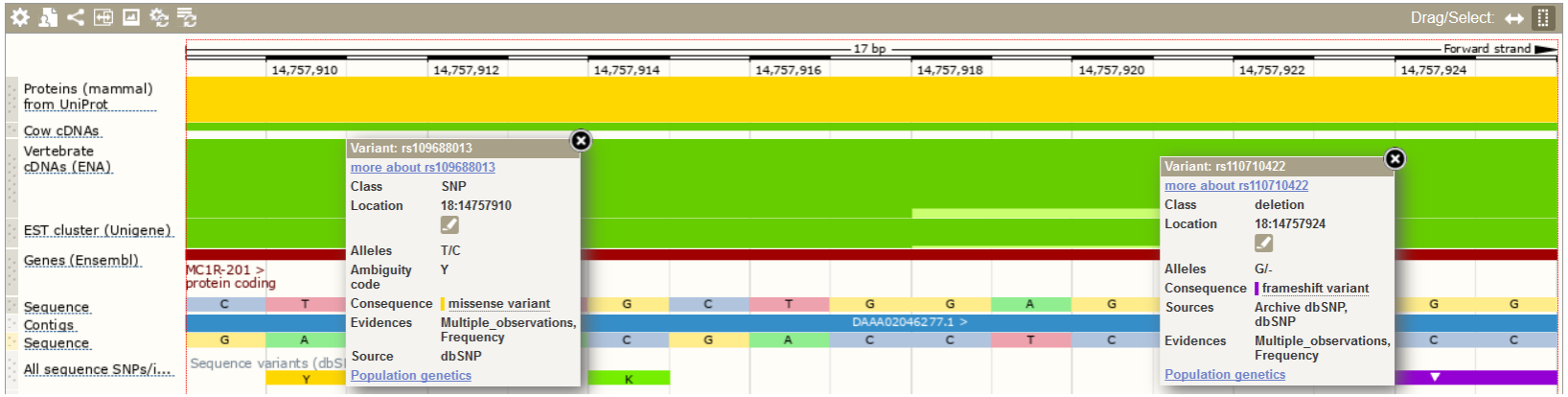
The two variants viewed on Ensembl.org

Chromosome 18: 14,757,909-14,757,925



Region in detail





Source: https://oct2018.archive.ensembl.org/Bos_taurus/Location/View?r=18:14757909-14757925;db=core

Conclusion

- The Bcbio-nextgen NGS data analysis pipeline shows **good scalability** when using from 64 up to 320 CPU cores
- More computing resources (**CPU cores** and, especially, **storage**) will be required for NGS data analysis of **larger populations**, in order to support more complex scenarios such as GWAS
- 100 samples with read coverage similar to the data used in this study will require approx. 2 TB for the raw sequence data and approx. 10 TB for the data analysis

Acknowledgments

- **Mihai Carabaș** from the **Politehnica University of Bucharest**, for providing access to one of their computing clusters
- **Amanda Chamberlain** and **Hans Daetwyler** from the **1000 Bovine Genomes project**, for providing a list of 1784 Bos Taurus samples that have been sequenced and whose data is publicly available