Owner	:	
Recipients	:	
Report description	:	Internship Report
Project name	:	DNA Variant Mining and In silico Analysis of Milk Proteins (BIO-IT Signature Project)
Author(s)	:	Muhamad Rifki Ramadhan
Date	:	2022-07-28

Management summary

1. Objective

Determine new milk protein variants from the known proteins using bull's genomic data.

2. Conclusion

Milk protein variant mining using genomic data is successfully performed on approximately 400 proteins. Variants found in 209 proteins including the known variants of major milk proteins. All variants data are uploaded in the FrieslandCampina's RedShift database. Ten variants are listed as the most interesting variants, including milk protein 4. LC-MS analysis confirms the presence of a milk protein 4 variant.

3. Recommendations and next steps

- Perform activity and digestion analysis on the confirmed milk protein 4 variant and compare it with the original protein.
- Confirm the presence of other protein variants (milk protein 1, milk protein 2, milk protein 5, etc.) by performing LC-MS.
- Perform the variant mining analysis on other bull population to see which variants are unique in Dutch bulls.
- Select bulls with favourable protein variants for breeding

4. Abstract

Milk is a resource of lipids, proteins, amino acids, vitamins, and minerals. Proteins in milk vary in biological activities which include among others antimicrobial, nutrients absorption facilitator, growth factors, hormones, enzymes, and antibodies. Milk proteins also exist in various isoforms which may differ in their activity. However, we still do not know much about the variation of milk proteins other than the major ones (the caseins, β-lactoglobulin and α-lactalbumin). The traditional methods of protein analysis (e.g. 2D PAGE) do not suffice to study minor milk proteins which comprise a very little fraction of milk protein. This project aims to find new milk protein variants in Dutch milk using genomic approach. DNA from 54 Dutch Holstein bulls are sequenced and called for variants. Variants effects on protein are predicted and protein variants from approximately 400 milk proteins are determined. Variants found in 209 proteins including the known variants of major milk proteins. Based on *in silico* protein analysis, ten variants are listed as the most interesting variants, including milk protein 4. LC-MS analysis confirms the presence of a milk protein 4 variant.

5. Layman's Summary

Milk is a resource of lipids, proteins, amino acids, vitamins, and minerals. Proteins in milk have various biological activities such as antimicrobial, hormones, enzymes, and antibodies. Milk proteins also exist in several forms which may differ in their activity. These different forms of protein are called protein variants. There are hundreds of proteins in milk but we only know protein variants of a few proteins, which are the ones with highest concentration (the caseins, β-lactoglobulin and α-lactalbumin). The traditional methods of protein analysis do not suffice to study other proteins which comprise a very little fraction of milk protein. This project aims to find new milk protein variants in Dutch milk using genomic approach. In this approach, instead of looking at proteins directly, we are looking at DNA. Protein is synthesized based on the information on DNA, so if we know the variation in DNA, we can infer that information to determine the variation in protein. DNA from 54 Dutch Holstein bulls are sequenced and called for variants. Variants effects on protein are predicted and protein variants from approximately 400 milk proteins are determined. Variants found in 209 proteins including the known variants of major milk proteins. Based on predictive protein analysis, ten variants are listed as the most interesting variants, including milk protein 4. To confirm the presence of the protein variants, a protein detection method called LC-MS is performed and presence of a milk protein 4 variant is confirmed.

6. Introduction & Background information

Milk Components and Nutritional Value

Milk has a great balance of its different components and is considered a very nutritionally complete food. This is not a surprise since milk is a biologically designed sustenance for fulfilling neonates nutritional requirements. Milk is a resource of lipids, proteins, amino acids, vitamins, and minerals. Different kind of milk varies in their components proportion, depending on several factors such as species, breed, age, nutrition, and period of lactation (Haug et al., 2007).

Vitamins in milk include vitamin E, vitamin A, folate (vitamin B9), riboflavin (vitamin B2), and vitamin B12. These vitamins mainly act as coenzyme and antioxidant. Calcium is a major mineral in milk. Daily intake of milk and milk products has a major role in fulfilling calcium requirement in our body. Calcium has a role in the development and repair of bones and teeth, maintaining neuron function, and help prevent hypertension (Insel et al., 2004). Milk is also an important source of selenium which support immune system, act as antioxidant, and helps the process of DNA synthesis and repair. Iodine, magnesium, and zinc are other essential minerals for our body that are contained in milk (Insel et al., 2004).

Triacylglycerols is the major lipid fraction in milk, accounting for about 95% of all milk lipids. Other milk lipids include diacylglycerol, phospholipid, cholesterol, and free fatty acids. Milk fatty acids are claimed to have negative effects because more than half of the milk fatty acids are saturated (Marckmann et al., 1994; Seidel et al., 2005). Some saturated fatty acid in milk such as myristic and palmitic acid are believed to raise blood cholesterol level (Mensink et al., 2003), whereas high cholesterol level increase the risk of coronary heart disease (CHD) (Mensink et al., 1992; Hegsted et al., 1993). This negative notion on milk consumption has been opposed by some studies that show milk fat consumption is having less pronounced effects on serum lipids than might be expected (Bosaeus, 1991; Eichholzer & Stahelin, 1993) and the association between milk consumption and CHD is non-existent (Stahelin, 1992; Willet et al., 1993; Fehily et al., 1993; Ness et al., 2000). Two studies even shown that cardiovascular risk factors were negatively associated with intake of milk fat (Smedman et al., 1999; Warensjo et al., 2004).

Proteins in milk are composed of two main fractions, caseins and whey proteins. Caseins represent up to 80% of cow's milk protein. Caseins can be obtained by acid precipitation as it is insoluble in milk at pH 4.6. The remaining proteins soluble are whey proteins, or also called serum proteins. Five types of caseins can be distinguished, α_{s1}-, α_{s2}-, β-, γ-, and κ-casein. Whey proteins consists of various minor proteins, with the major ones being β-lactoglobulin which represents about 50% of whey proteins and α-lactalbumin which accounts for about 13%. Another examples of whey proteins are immunoglobulins, bovine serum albumin, lactoferrin, and lactoperoxidase.

Proteins in milk vary in biological activities which include among others antimicrobial, nutrients absorption facilitator, growth factors, hormones,

enzymes, and antibodies. Caseins, the largest proportion of milk proteins, have a role in the binding of calcium and phosphate. It also help digestion in stomach by forming clots. Whey proteins increase the plasma amino acids after meal as it is considered as rapid digested protein. Some milk proteins or peptides derived from it may have functional role inside the digestive tract before being fully digested. Lactoferrin, lactalbumin, and secretory immunoglobulin A are some examples of these proteins.

Several milk bioactive proteins are shown to have positive effect when added as food supplement or as ingredients of food products. Lactoferrin, for example, is believed to reduce the risk of respiratory and gastrointestinal infections in infants (King et al., 2007; Chen et al., 2016; Li et al., 2019). Several companies, including FrieslandCampina, sell lactoferrin as ingredient for infant or medical nutrition. With the completion of lactoferrin factory in Veghel, FrieslandCampina will become the largest producer of lactoferrin in the world.

Variations in Milk Proteins

Major milk proteins are known to exists in various isoforms. β-casein, for example, has 12 variants identified in various breeds of cow. These variants are different in just one or few amino acids. For instance, β-casein A1 variant only differ with A2 variant in amino acid position 67 – histidine in A1 and proline in A2 (Farrell et al., 2004).

These β-casein A1 and A2 variants are also the ones that sparked a lot of discussion in early 2000s. Some researchers from New Zealand and Iceland claimed that A2 variant is more favourable for human consumption because it does not facilitate immunological process that could lead to type I diabetes and coronary heart disease as A1 variant does (Elliot, 1992; Birgisdottir et al., 2002). A company called A2 Corporation was established in New Zealand to select cows and produce milk with only the A2 variant. This product has been marketed as far as USA and Canada. However, the claim received a lot of criticism because it is mostly based on between-countries association studies. Between-countries association study is not a preferred method to discover the health effect of food because there are too many factors at hand and it can not be reproduced. Moreover, the initial experiment in mice was reproduced by other researchers and shows opposing result (Truswell, 2005).

Across all breeds, α_{s1}-casein and α_{s2}-casein has 8 and 4 variants respectively, but Dutch Holstein cattles generally only has the A variant of both protein. K-casein has 11 variants in total, with A, B, and E variants the most common in Dutch cattles. Not only the caseins, whey protein β-lactoglobulin and a-lactalbumin also has protein variations, with 11 variants identified for β-lactoglobulin and 3 for α-lactalbumin (Farrell et al., 2004).

Unfortunately, we still do not know much about the variation of milk proteins other than the major ones (the caseins, β-lactoglobulin and a-lactalbumin). These minor proteins are not thoroughly studied because they account for very little fraction of milk protein. This means that it is very hard to isolate them and they are not likely to affect the overall quality and

properties of the milk. Consequently, they are often overlooked by companies and researchers.

There is an emerging interest in research for several bioactive proteins such as lactoferrin, lactoperoxidase, and the MFGM proteins. These proteins are being studied for their functional properties and are usually extracted to be sold as ingredients for specific-segment nutrition. However, there is still a lot to discover about these proteins, including their variations and the different properties, functionalities, and distribution of their various forms.

Nevertheless, there are a lot more minor proteins, which each accounts for less than 0.1% of milk proteins, that are still not studied at all. We do not know about the role of these proteins in milk, let alone its variations. This is basically a new territory and a good opportunity for FrieslandCampina to be the first one to step its foot in it and understand milk proteins which are always overlooked before.

Our Unique Approach: Looking for Protein Variants through DNA

The major difficulties in studying milk proteins – and proteins in general –, is separation. The traditional method of proteomics to separate protein variants, 2D-PAGE, is a type of electrophoresis method that combines isoelectric and molecular mass separation. Proteins with unique isoelectric point (pI) and mass occupy specific position in the gel and are visualized and quantified by different procedures, such as Coomassie, fluorescence, or silver staining (May et al., 2012). This method is not good enough to detect proteins that have low abundance, extremely acidic or basic, and have masses outside the limit range (Gygi et al., 2000; O'Donnell et al., 2004). Another technique, liquid chromatography (LC), which can be coupled with mass spectrometry (MS), offers better sensibility and dynamic range. However, this method is more of a confirmation method because we have to know beforehand what variation we are expecting. In essence, separating milk minor proteins, whose concentrations are mostly less than 0.1%, is still a huge technical challenge for milk proteomics analysis today (Lorenzo et al., 2018; Agregan et al., 2021).

Another way of looking at protein variants, which are not usually implemented in milk proteins, is by looking at genomics data. This method does not look at the protein directly, but indirectly through its DNA sequence.

DNA is often regarded as the blueprint of life because it contains information and instruction for cell growth, development, survival, and reproduction. These cell functionalities are actually done mostly by proteins, but protein synthesis is based on the information contained in the DNA. Information in DNA is stored in the form of nitrogen base sequence of adenine, cytosine, guanine, and thymine. This information are translated to protein sequence which composed of strings of amino acids. There are 20 kinds of amino acids that make up proteins. Three sets of nitrogen base of DNA, which is called codon, translated to one amino acid. This is why we can know the sequence of a protein by looking at its DNA sequence. If there is a base substitution in the DNA sequence, the resulting effect on protein sequence could also be determined.

DNA sequencing is not an expensive method as it was 20 years ago. Based on NHGRI data (2022), the sequencing cost per mega-base (Mb) dropped from $5292.3 in 2001 to $0.006 in 2021. The invention of Next Generation Sequencing (NGS) makes it much cheaper and easier to sequence a genome – total DNA – of an organism. More and more organisms are being sequenced and a lot of research are now using genomic approach, including cattle breeding.

Using genomic data to look at milk protein variant is a very sensible approach. This method offers great sensibility and resolution as the separation process is not needed at all. There is no limitation on the protein abundance, pH, or mass. This method is also easier, faster, and effective since there is no lengthy, laborious work that consumes a lot of times and resources and prone to human error. In addition, once we build the pipeline, we can apply this method to a lot of proteins at once, and also to other data with different sample or population.

Bull's genomic data is preferred in this analysis compared to cow's. This is because the eventual goal of this project is not only to find protein variants, but also to produce milk with the more favourable protein variant for human consumption. By using bull's data, we will know which bull has a certain protein variant and can encourage farmers to use that bull to sire their cows.

This project aims to find new milk protein variants in Dutch milk using genomic approach. Finding the new milk protein variants will help FrieslandCampina become the frontier in milk protein function discovery and the leading producer of the best milk, milk products, and milk ingredients.

7. Materials & Methods

Workflow overview

This project can be divided into two main works: protein variant identification and in silico protein analysis. Protein variant identification includes DNA variant calling, variant effect prediction, and protein variant determination. Variant calling determines the location of DNA variants in the genome, while variant effect prediction predicts the effect of the DNA variants to its protein. Protein variant determination is a process to determine protein variants of protein of interest, which is inferred from variant effect data.

In silico protein analysis includes variant feature analysis, protein properties analysis, protein digestion analysis, and additional structural damage analysis, all of which will be described further in the following sections. Finally, a liquid chromatography – mass spectrometry (LC-MS) experiment is conducted to confirm the presence of selected protein variants. The overall schematic workflow of this project is shown in Figure 1.

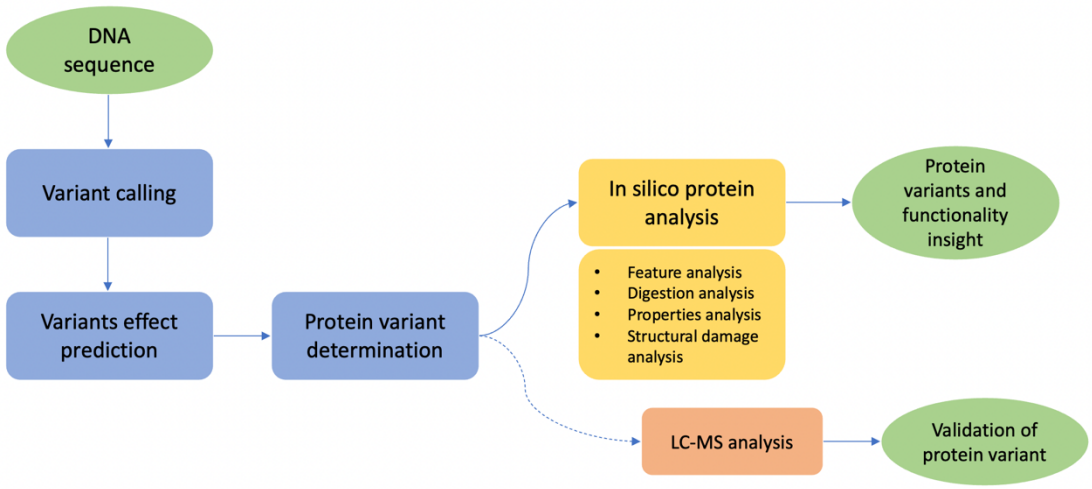


Figure 1. Schematic workflow diagram

Data Collection

Data is obtained from 54 Dutch Holstein bulls which is part of 1000 Bull Genomes Project. DNA sequence data of these bulls are properties of Wageningen University and Research (WUR) and the variant calling process is done by them. The DNA variant data which we obtained from WUR includes 29 chromosomes of bull autosome. The sex chromosome X and Y are not included in the data.

The data is received in vcf format, which is a text file containing lines of information about DNA variant positions in the genome and sample's genotype information of each position. A small section of the vcf files is depicted in Table 1. In this case, each line contains variant position and information (chromosome, position coordinate, variant id, reference base, alternate base, quality of reads, and filtering status) and the variant genotype of each 54 bulls (SIRE01, SIRE02, SIRE03, and so on).

Table 1. Snapshot of vcf file

CHROM	POS	ID	REF	ALT	QUAL	FILTER	SIRE01	SIRE02	SIRE03	SIRE04	SIRE05
6	85111289	.	G	T	.	PASS	0|0:0:1,0,0	0|1:1:0,1,0	0|0:0:1,0,0	0|0:0:1,0,0	0|1:0.932:0.1,0.868,0.032
6	85111561	.	A	C	.	PASS	0|0:0:1,0,0	0|1:1:0,1,0	0|0:0:1,0,0	0|0:0:1,0,0	0|0:0:1,0,0
6	85111621	.	G	C	.	PASS	0|0:0:1,0,0	0|1:1:0,1,0	0|0:0:1,0,0	0|0:0.254:0.751,0.244,0.005	0|0:0:1,0,0
6	85111664	.	A	G	.	PASS	0|0:0:1,0,0	0|1:1:0,1,0	0|0:0:1,0,0	0|0:0:1,0,0	0|1:1:0,1,0
6	85111711	.	G	A	.	PASS	0|0:0:1,0,0	0|1:1:0,1,0	0|0:0:1,0,0	1|0:1:0,1,0	0|1:1:0,1,0
6	85111759	.	G	T	.	PASS	0|0:0:1,0,0	0|1:1:0,1,0	0|0:0:1,0,0	0|0:0:1,0,0	0|1:1:0,1,0
6	85111888	.	G	T	.	PASS	0|0:0:1,0,0	0|1:1:0,1,0	0|0:0:1,0,0	0|0:0:1,0,0	0|1:1:0,1,0
6	85112165	.	C	T	.	PASS	0|0:0:1,0,0	0|0:0:1,0,0	0|0:0:1,0,0	0|0:0:1,0,0	0|0:0:1,0,0
6	85112249	.	T	G	.	PASS	0|0:0:1,0,0	0|0:0:1,0,0	0|0:0:1,0,0	0|0:0.008:0.992,0.008,0	0|0:0.078:0.922,0.078,0

Variant Effect Prediction

Variant effect prediction is performed using variant effect predictor (VEP) (McLaren et al., 2016). VEP predicts the consequences of variants in gene, transcript, and protein sequence. VEP produces insightful information such as gene, transcript, or protein that are affected by the variants, genetic location of the variants (upstream or downstream of a gene, coding region, intergenic region, etc.), and consequences of the variants to protein (missense mutation, synonymous mutation, frameshift mutation, etc.). The snapshot of output file generated from VEP is depicted in Table 2.

For this analysis we included Blosum62 plugin which adds BLOSUM score of the amino acid mutation in the output. BLOSUM score indicates the likelihood of substitution of amino acid based on its side chain characteristics. BLOSUM score is the logarithm of the ratio of two amino acids appearing with a biological sense and the likelihood of those amino acids appearing by chance. The higher the BLOSUM score, the more likely the amino acid mutation occurs in nature.

We also included Phenotypes plugin which adds the known association between a variant and a phenotype. In milk protein, the associated phenotypes mainly involving milk production, such as milk yield, protein percentage, and fat percentage. The phenotypes could also regarding the cow's health, such as udder infection.

Table 2. Snapshot of VEP output file

Uploaded variation	Location	Allele	Gene	Feature	Feature type	Consequence	cDNA pos	CDS pos	Protein pos	Amino acids	Codons
22_113261_T/A	22:113261	A	112443458	XM_024983201.1	Transcript	intron_variant	-	-	-	-	-
22_113261_T/A	22:113261	A	112443460	XR_003031831.1	Transcript	downstream_gene_variant	-	-	-	-	-
22_113273_C/T	22:113273	T	112443458	XM_024983201.1	Transcript	missense_variant,splice_region_variant	92	75	25	M/I	atG/atA
22_113273_C/T	22:113273	T	112443460	XR_003031831.1	Transcript	downstream_gene_variant	-	-	-	-	-
22_113290_T/G	22:113290	G	112443458	XM_024983201.1	Transcript	missense_variant	75	58	20	I/L	Ata/Cta
22_113290_T/G	22:113290	G	112443460	XR_003031831.1	Transcript	downstream_gene_variant	-	-	-	-	-
22_113331_T/C	22:113331	C	112443458	XM_024983201.1	Transcript	missense_variant	34	17	6	N/S	aAc/aGc
22_113331_T/C	22:113331	C	112443460	XR_003031831.1	Transcript	downstream_gene_variant	-	-	-	-	-
22_113336_C/T	22:113336	T	112443458	XM_024983201.1	Transcript	synonymous_variant	29	12	4	S	tcG/tcA
22_113336_C/T	22:113336	T	112443460	XR_003031831.1	Transcript	downstream_gene_variant	-	-	-	-	-
22_113356_G/T	22:113356	T	112443458	XM_024983201.1	Transcript	5_prime_UTR_variant	9	-	-	-	-
22_113356_G/T	22:113356	T	112443460	XR_003031831.1	Transcript	downstream_gene_variant	-	-	-	-	-

Protein Variant Determination

VEP can describe which DNA variant lies within the coding region of which gene, but it does not describe the protein variant resulted from the combination of those DNA variants. VEP also does not calculate the frequency of the DNA variants and the resulting protein variants. Therefore, a function is created to perform this tasks automatically using R (Appendix A).

This function receive input of gene symbol or uniprot entry of a protein and do the following: 1) find DNA variants that lies within the coding region and affects the change in protein sequence of a particular gene, 2) calculate the frequency of the DNA variants, 3) determine protein variants resulted from the combination of the DNA variants, 4) calculate the frequency of the protein variants, 5) determine protein genotypes found in the population, 6) calculate the frequency of the protein genotypes, 7) determine the individual bull haplotypes and genotype of each protein variant.

For each protein, the function will generate 7 main output files: DNA coding variants, protein variants (haplotype), protein genotypes, protein variants-DNA coding variants junction, protein genotypes-protein variants junction, bull haplotypes file, and bull genotypes file. The junction files are the file that links the association of two files, which are the association of protein variants and DNA coding variants (which DNA mutation occurs in a particular protein variant) and the association of protein genotypes and protein variants (which two variants constitute a particular genotype).

FrieslandCampina previously performed a proteomic analysis and listed approximately 400 proteins in bovine milk. The protein variant determination function is performed on all of these milk proteins.

In Silico Feature Analysis

Feature analysis is the first in silico protein analysis performed. Feature analysis is a process to determine whether a protein variant has amino acid substitution in a featured amino acid. Featured amino acid is an amino acid which has been annotated to has particular function.

Table 3. Uniprot feature data

Entry	Source	Type	Start	End	Score	Strand	Frame	Description
Q06002	UniProtKB	Region	184	200	.	.	.	Note=Linker 12
Q06002	UniProtKB	Region	201	319	.	.	.	Note=Coil 2
Q06002	UniProtKB	Region	320	756	.	.	.	Note=Tail
Q06002	UniProtKB	Region	406	436	.	.	.	Note=Disordered;Ontology_term=ECO:0000256;evidence=ECO:0000256|SAM:MobiDB-lite
Q06002	UniProtKB	Region	493	705	.	.	.	Note=Disordered;Ontology_term=ECO:0000256;evidence=ECO:0000256|SAM:MobiDB-lite
Q06002	UniProtKB	Region	532	622	.	.	.	Note=7 X 14 AA tandem repeats
Q06002	UniProtKB	Compositional bias	509	523	.	.	.	Note=Pro residues;Ontology_term=ECO:0000256;evidence=ECO:0000256|SAM:MobiDB-lite
Q06002	UniProtKB	Compositional bias	524	658	.	.	.	Note=Basic and acidic residues;Ontology_term=ECO:0000256;evidence=ECO:0000256|SAM:MobiDB-lite
Q06002	UniProtKB	Site	40	41	.	.	.	Note=Cleavage;Ontology_term=ECO:0000269;evidence=ECO:0000269|PubMed:19875662;Dbxref=PMID:19875662
Q06002	UniProtKB	Site	432	433	.	.	.	Note=Cleavage (by CASP2%2C CASP3%2C and CASP7);Ontology_term=ECO:0000269;evidence=ECO:0000269|PubMed:19875662;Dbxref=PMID:19875662
Q06002	UniProtKB	Site	456	456	.	.	.	Note=Interaction with MIP;Ontology_term=ECO:0000269;evidence=ECO:0000269|PubMed:28259670;Dbxref=PMID:28259670
Q06002	UniProtKB	Modified residue	5	5	.	.	.	Note=Phosphoserine;Ontology_term=ECO:0000269;evidence=ECO:0000269|PubMed:19875662;Dbxref=PMID:19875662
Q06002	UniProtKB	Modified residue	41	41	.	.	.	Note=N-acetylalanine;Ontology_term=ECO:0000269;evidence=ECO:0000269|PubMed:19875662;Dbxref=PMID:19875662

The featured amino acid data is obtained from Uniprot. Uniprot's feature data includes various information of amino acid positions in a protein that are categorized to molecule processing, regions, sites, amino acid modifications, natural variations, experimental information, and secondary structure. For this analysis, we focus on the sites, amino acid modification, and several annotations in regions category. The sites category consists of active site, metal binding, binding site (any other binding such as co-enzyme, prosthetic group, etc.), and site (any other single amino acid site).

The amino acid modification categories includes non-standard residue, modified residue, lipidation, glycosylation, disulfide bond, and cross-link. The regions annotation included in this analysis are calcium binding, nucleotide binding, DNA binding, and motif. Table 3 shows a snippet of Uniprot feature data.

To determine the presence of featured amino acid, amino acid position in variant file is cross-checked with the amino acid position in Uniprot feature data. A function in R is created to do this automatically on all proteins. The script for this function is attached in Appendix B. The resulting DNA variant file will have additional columns regarding the features as shown in Table 4.

Table 4. DNA variant data with feature

symbol	entry	chrom	pos	ref	alt	allele	protein position	ref amino acids	protein pos adjusted	alt amino acids	blosum62	count	freq	feature	start	end	attribute	start adjusted	end adjusted
XXX	XXXXXX	NXX	NXXXXXX	A	G	G	14	F	-3	L	0	14	13	Signal peptide	1	17	Ontology_term=ECO:0000269;evidence=ECO:0000269|PubMed:3458202;Dbxref=PMID:3458202	-16	0
XXX	XXXXXX	NXX	NXXXXXY	G	A	A	210	H	193	Y	2	22	20.4	Sequence conflict	210	210	Note=H->Y;Ontology_term=ECO:0000305;evidence=ECO:0000305	193	193
YYY	YYYYYY	NXY	NXXXXXZ	G	A	A	28	A	4	V	0	49	45.4	Domain	25	142	Note=RNAse_Pc;Ontology_term=ECO:0000259;evidence=ECO:0000259|SMART:SM00092	1	118
ZZZ	ZZZZZZ	NXZ	NXXXXXA	C	G	G	528	R	NA	G	-2	5	4.6	NA	NA	NA	NA	NA	NA
ZZZ	ZZZZZZ	NXZ	NXXXXXB	C	T	T	598	P	NA	L	-3	4	3.7	Region	557	613	Note=Disordered;Ontology_term=ECO:0000256;evidence=ECO:0000256|SAM:MobiDB-lite	NA	NA
ZZZ	ZZZZZZ	NXZ	NXXXXXC	G	A	A	704	G	NA	S	0	5	4.6	NA	NA	NA	NA	NA	NA
ZZZ	ZZZZZZ	NXZ	NXXXXXD	A	G	G	776	E	NA	G	-2	5	4.6	Region	755	783	Note=Disordered;Ontology_term=ECO:0000256;evidence=ECO:0000256|SAM:MobiDB-lite	NA	NA
ZZZ	ZZZZZZ	NXZ	NXXXXXE	A	G	G	776	E	NA	G	-2	5	4.6	Compositional bias	755	782	Note=Polar residues;Ontology_term=ECO:0000256;evidence=ECO:0000256|SAM:MobiDB-lite	NA	NA
ZZZ	ZZZZZZ	NXZ	NXXXXXF	C	T	T	889	A	NA	V	0	4	3.7	Region	879	907	Note=Disordered;Ontology_term=ECO:0000256;evidence=ECO:0000256|SAM:MobiDB-lite	NA	NA

Protein Variant Sequence Generation

Based on the substitution data in variant file, protein variant sequence is created for further analysis. Reference sequence for every protein is downloaded from Uniprot. Amino acid substitution data for each protein variant is extracted from the variant file and the substitution is applied to the reference fasta. A function in R is created to perform these tasks by utilizing seqinR package. The script for this function is attached in the Appendix C.

The output of the protein sequence is stored in fasta format. Fasta is the most common text file format to store protein or DNA sequence and is the standard format used by most application. Fasta file starts with description line that are indicated by ">" and followed by the sequence itself. The protein fasta file example is shown in Figure 2.

```
>A1BG|Q2KJF1|H01
MSAWAALLLLWGLSLSPVTEQATFFDPRPSLWAEAGSPLAPWADVTLTCQSPLPTQEFQL
LKDGVGQEPVHLESPAHEHRFPLGPVTSTTRGLYRCSYKGNNDWISPSNLVEVTGAEPLP
APSISTSPVSWITPGLNTTLLCLSGLRGVTFLLRLEGEDQFLEVAEAPEATQATFPVHRA
GNYSCSYRTHAAGTPSEPSATVTIEELDPPPAPTLTVDRESAKVLRPGSSASLTCVAPLS
GVDFQLRRGAEEQLVPRASTSPDRVFFRLSALAAGDGSGYTCRYRLRSELAAWSRDSAPA
ELVLSDGTLPAPELSAEPAILSPTPGALVQLRCRAPRAGVRFALVRKDAGGRQVQRVLSP
AGPEAQFELRGVSAVDSGNYSCVYVDTSPPFAGSKPSATLELRVDGPLPRPQLRALWTGA
LTPGRDAVLRCEAEVPDVSFLLLRAGEEEPLAVAWSTHGPADLVLTSVGPQHAGTYSCRY
RTGGPRSLLSELSDPVELRVAGS
```

Figure 2. Example of protein fasta file

The sequence of protein variant with frameshift mutation cannot be generated because the amino acids in the new reading frame cannot be converted. Therefore, further analysis of these variants cannot be performed. Nevertheless, the fact that frameshift mutation occurs on this protein variant is already an important enough information since a frameshift mutation will likely damage the structure and function of a protein.

In Silico Protein Properties Analysis

The value of seven protein properties are calculated to assess the characteristics of the protein variants. The properties calculated are molecular weight, charge, hydrophobicity, isoelectric point (pI), peptide length, aliphatic index, and stability index. The calculation of these properties is performed using Peptides package in R. Protein variant sequence is used as input and the output values are incorporated to the protein variant table.

The absolute difference of these properties value between the protein variants and their reference protein is then calculated to give insight on how different these protein variants will behave compared to the reference protein. Table 5 depicts the snippet of protein variants table with properties difference value.

Table 5. Protein variants table with protein properties data

Symbol	Entry	Protein names	Haplotype id	Haplo type	Count	Freq	Ref ere nce	Fra mes hift	aIndex diff abs	Charge diff abs	Hydrop hobicity diff abs	instaInd ex diff abs	Length diff abs	Mw diff abs	pI diff abs
XXX	XXXXXX	Protein X	XXX_XXXXXX_01	0	107	991	1	0	0	0	0	0	0	0	0
XXX	XXXXXX	Protein X	XXX_XXXXXX_02	1	1	9	0	0	1.230	0	0.017	0.030	0	16.043	0
YYY	YYYYYY	Protein Y	YYY_YYYYYY_01	0	107	991	1	0	0	0	0	0	0	0	0
YYY	YYYYYY	Protein Y	YYY_YYYYYY_02	1	1	9	0	0	0	2.917e-4	0.002	2.220	0	28.013	4.475e-5
ZZZ	ZZZZZZ	Protein Z	ZZZ_ZZZZZZ_01	0	106	981	1	0	0	0	0	0	0	0	0
ZZZ	ZZZZZZ	Protein Z	ZZZ_ZZZZZZ_02	1	2	19	0	0	1.421e-14	0	0.002	0.438	0	0	0
AAA	AAAAAA	Protein A	AAA_AAAAAA_01	0	65	602	1	0	0	0	0	0	0	0	0
AAA	AAAAAA	Protein A	AAA_AAAAAA_02	1	43	398	0	0	0.275	0	0.009	1.207	0	26.038	0
BBB	BBBBBB	Protein B	BBB_BBBBBB_01	00	105	972	1	0	0	0	0	0	0	0	0
BBB	BBBBBB	Protein B	BBB_BBBBBB_02	10	2	19	0	1	NA	NA	NA	NA	NA	NA	NA
BBB	BBBBBB	Protein B	BBB_BBBBBB_03	01	1	9	0	0	0	0	0.004	0.754	0	27.026	0

In Silico Protein Digestion Analysis

The digestion of protein variant is performed *in silico* to predict whether there is a difference in peptides produced from a protein variant compared to its reference protein. This is done by comparing the amount of peptides generated in the digestion output file of a protein variant with the amount of peptides in reference protein output file.

The digestion analysis is performed using Rapid Peptide Generator (Maillet, 2019). The enzyme used are pepsin (at pH $\geq$ 2), trypsin, and chymotrypsin to mimic natural protein digestion in human body. The analysis is performed for both concurrent and sequential mode. The concurrent mode means the digestion of protein by the enzymes occurred simultaneously. This mode may allow one enzyme to access cleavage sites that are normally not

available if the enzyme only acts by itself. In sequential mode, the enzymes digest the protein independently and produce distinct result for each enzyme. The script for performing Rapid Peptide Generator is attached in Appendix D.

In Silico Structural Damage Analysis

Structural damage analysis is performed *in silico* using webtool Missense3D (Ittisoponpisan et al., 2019). Missense3D receive PDB or homology predicted structure, the position of the substitution, and the reference and substituted amino acid as input to predict structural changes introduced by that substitution. Missense3D analyze the structural changes based on 16 defined parameters: disulphide breakage, buried proline introduction, clash, buried hydrophilic introduction, buried charge introduction, secondary structure alteration, buried charge switch, disallowed phi/psi, buried charge replacement, buried glycine replacement, buried H-bond breakage, buried salt bridge breakage, cavity alteration, buried/exposed switch, cis proline replacement, and glycine in a bend.

The analysis is performed only to protein variants that have amino acid substitution with very low BLOSUM score (-3 and -2). This is because the limitation of the tool that cannot process the analysis in batch and the fact that substitution with high BLOSUM score will most likely not affect the structure of a protein.

Data Storing

The result of all this variant determination and in silico analysis is stored in RedShift cloud database of FrieslandCampina. They will be uploaded in the giga database. There are a total of 11 tables uploaded from this research. Three tables are the result of variant calling and variant effect prediction: variants, bull variants, and consequences table. Seven tables are a result of protein variant determination and in silico protein analysis, namely DNA coding variants table (including the feature data), protein variants (including protein properties data), protein genotypes, protein variants-DNA coding variants junction, protein genotypes-protein variants junction, bull haplotypes file, and bull genotypes file. The last table is proteins table that are obtained from previous study.

Finding the Most Interesting Variants

A scoring system is created to assess and weigh the importance of the aforementioned analysis. The output of this scoring is a list of protein variants with the most distinct characteristics, thus interesting to study further.

A total score is calculated as the sum of feature, properties, digestion, and substitution score. The feature score is the amount of amino acid substitution that happened at the annotated positions. The properties score is the amount of distinct protein properties values. There are seven properties calculated, so the properties score ranges from 0 to 7. A properties value is considered distinct if it is above the median. Digestion

score is an absolute difference of the amount of peptides resulted from digestion analysis.

The substitution score is the sum of adjusted BLOSUM score at each substituted amino acid position. The adjusted BLOSUM score is calculated by bringing the BLOSUM score to the positive side and inverting the score order. This means that the adjusted BLOSUM score ranges from 1 to 7 with 1 is the most likely amino acid substitution and 7 is the most unlikely.

The weighs for each score is considered to make the most sensible calculation. For example, the feature score is clearly more important than properties score so it should be weighed higher and properties score is closely related to substitution score so they should weighed lower. The final formula to calculate total score is as follows:

$$Total\ score = (5 \times feat.score) + dig.score + prop.score + (0.25 \times subs.score)$$

Protein variants are ranked based on the total score and variants with higher total score are deemed more interesting. In addition, protein variants with frameshift mutation and nonsense mutation (stop codon gained) are automatically top the list regardless of their total score.

LC-MS Analysis as a Proof of Principle

Finally, liquid chromatography-mass spectrometry (LC-MS) analysis is conducted to validate the presence of protein variants found from this genomic analysis. LC-MS combines the physical separation method of liquid chromatography (LC) with the mass measurement method of mass spectromectry (MS). The machine used in this analysis is Waters™ SELECT SERIES Cyclic IMS. This is a type of ion mobility mass spectrometer that separates gas phase ions based on their interaction with a collision gas and their masses (Kanu et al., 2008).

Prior to measurements, protein is digested with trypsin, resulting a mixture of peptides with different sizes. The mixture is then inserted to the chromatography, separating the peptides by the order of its mass. The mass spectrometer ionized and sprayed the peptides into gaseous ions and measure the mass of each peptide while also cleave the peptide mechanically. On the other hand, the protein are digested by trypsin *in silico* to predict the resulting peptides. The expected masses of each peptide is calculated based on the sequence of the peptides.

The LC-MS software matches the measured mass from MS to the expected mass to validate the presence of each peptide. The masses of cleaved peptides are used to further confirm the presence of that particular peptide, as opposed to other peptide with the same total mass. Different protein variants is detected by confirming the presence of peptides in which the mutation occurs.

8. Results & Discussion

Result

Protein variant determination is performed on approximately 400 proteins. Variants found in 48% of proteins while the other 52% of proteins have no variant identified. The identification is failed on 4 proteins because the protein name and uniprot entry are not identified. Overall, 93.5% of protein has no more than 3 variants, with 9 proteins having 4 variants, 4 proteins having 5 variants, and 15 proteins having more than 5 variants. Figure 3 depicted the number of variant distribution of the proteins. Protein with the most number of variant found is an uncharacterized protein with Uniprot entry of G5E5W7 which has 107 variants. The second protein with the highest number of variant is complement factor H (CFH) with 103 variants. Table containing the number of variants found in each protein is attached in Appendix E.

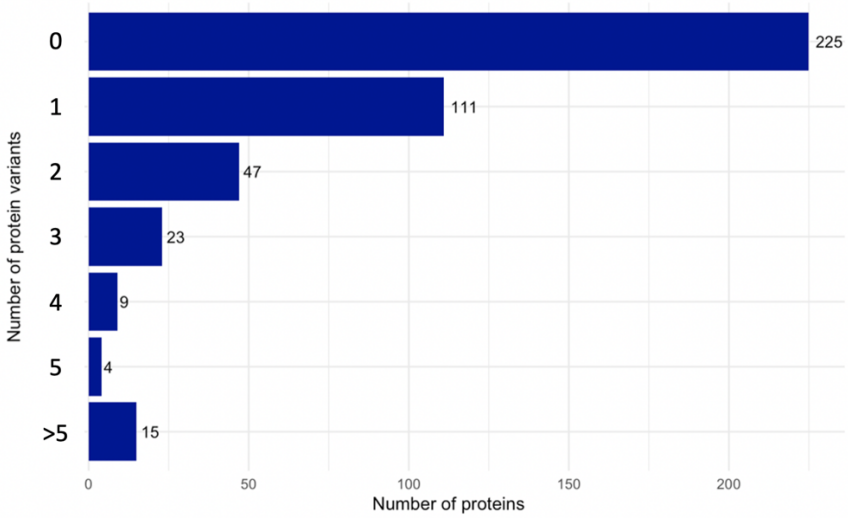


Figure 3. Number of variants distribution

It is important to note that the protein variants are filtered using 5% frequency threshold. This is based on the sequence reads investigation of several proteins that showed variants with very low frequency tend to be a result of false genotyping due to low coverage reads.

The protein variants found in 6 major milk proteins are depicted in Table 6. There is no variant found for α_{s1}-casein, α_{s2}-casein, and α-lactalbumin. Two variants found for β-casein and one variant found for κ-casein. B-lactoglobulin is found to have two variants, with one variant, X, is a new variant. This variant has the A134V mutation like the B variant, but does not have the G80D mutation. Further investigation suggests that this variant is most certainly a result of false genotyping due to low quality of DNA sequence reads.

Table 6. Protein variants found in major milk proteins

Symbol	Protein names	Protein variant	Length
CSN1S1	Alpha-S1-casein	0	214
CSN1S2	Alpha-S2-casein	0	222
CSN2	Beta-casein	2	259
CSN3	Kappa-casein	1	190
LALBA	Alpha-lactalbumin	0	185
PAEP	Beta-lactoglobulin (Beta-LG)	2	178

Beta-casein (CSN2)	
Variant	Freq (%)
A^2	63
A^1	17.6
I	13.9

Kappa-casein (CSN3)	
Variant	Freq (%)
A	54.7
B	37

Beta-lactoglobulin (BLG/PAEP)	
Variant	Freq (%)
A	53.7
B	38
X	7.4

The scoring to find the most interesting variant is successfully performed and the top 15 variants are depicted in Table 7. The result is filtered to protein variant with higher than 5% frequency and of top 40 milk protein with highest concentration. Furthermore, the protein variants with unreviewed protein sequence in Uniprot are reviewed manually to determine whether the substitutions could occur in the reviewed protein sequence or not. If an amino acid substitution occurs in a region that is not present in the reviewed sequence, that substitution is omitted and the new list of variants resulted from the combination of the remaining substitution is determined.

Table 7. The most interesting protein variant

No.	Symbol	Protein names	Variant	Freq (%)	Frameshift	Stop gained	Digestion score	Properties score	Substitution score	Feature score	Total score
1.	MP1	Milk protein 1	C	15.7	1	0	0	0	2	0	0.5
2.	MP2	Milk protein 2	C	12.9	0	0	4	3	11	0	9.75
3.	MP1	Milk protein 1	D	7.4	0	0	4	4	6	0	9.5
4.	MP3	Milk protein 3	B	17.6	0	0	5	4	2	0	9.5
5.	MP4	Milk protein 4	B	18.5	0	0	6	2	3	0	8.75
6.	MP5	Milk protein 5	D	7.4	0	0	2	3	11	0	7.75
7.	MP4	Milk protein 4	C	5.6	0	0	6	1	2	0	7.5
8.	MP6	Milk protein 6	B	16.7	0	0	2	4	4	0	7
9.	MP7	Milk protein 7	B	41.7	0	0	2	4	3	0	6.75
10.	MP8	Milk protein 8	C	9.3	0	0	3	3	3	0	6.75

Only one variant experienced frameshift mutation and there is no variant that has a nonsense mutation (stop codon gained). There is also no variant that has an amino acid substitution in the featured position as depicted in the all zero feature score in the table. The list contains 10 protein variants from 8 different proteins. The details of amino acid substitutions of the protein variants are depicted in Table 8.

Milk protein 1 C variant has a frameshift mutation and therefore tops the list. An insertion of two bases occurs at position NXXXXXX (chromosome NX). This insertion causes a frameshift mutation at position NXZ. The amino acid at position NXZ itself does not change (Ala), but the reading frame is shifted and change the following amino acids. The variant concentration is

adequate (15.7%), and further investigation on the DNA sequence reads reveals that this variant is indeed not a case of false genotyping.

Milk protein 1 D variants also make the list and are interesting to look at as well. It has the most distinct characteristics and digestion results compared to the reference (A) variant. This variant has histidine to arginine substitution at position NXX and lysine to arginine at position NXY.

Variant C of milk protein 2 is the second highest scoring variant. The most interesting feature of this variant is its high substitution score – indicates the unlikeliness of the substitution – courtesy of proline substitution to leucine at position NXX. This variant also has different peptides resulted from enzymes digestion and 3 protein parameters that differ highly from the reference milk protein 2 protein. Because of it has a substitution with low BLOSUM score, a structural damage analysis is performed on this variant. A structural damage is detected as the ProNXXLeu substitution leads to the expansion of cavity volume in the repeat region by 71.496 $Å^3$.

Table 8. Protein variants of selected proteins

Milk Protein 1					
Variant	NXX	NXY	NXZ	NXA	Freq (%)
A	His	Lys	Ala	Ile	44.4
B	Arg	Lys	Ala	Ile	26.9
C	His	Lys	Ala-X	Leu	15.7
D	Arg	Arg	Ala	Ile	7.4

Milk Protein 2			
Variant	NXX	NXY	Freq (%)
A	Pro	Thr	45.4
B	Pro	Ala	40.8
C	Leu	Ala	12.9

Milk Protein 3			
Variant	NXX	NXY	Freq (%)
A	Phe	His	69.4
B	Phe	Tyr	17.6
C	Leu	His	10.2

Milk Protein 4			
Variant	NXX	NXY	Freq (%)
A	Ile	His	71.3
B	Val	Tyr	18.5
C	Ile	Tyr	5.6

Milk Protein 5										
Variant	NXX	NXY	NXZ	NXA	NXB	NXC	NXD	NXE	NXF	Freq (%)
A	Ala	Phe	Gln	Arg	Arg	Ser	Asp	Val	Lys	31.5
B	Ala	Phe	Pro	Arg	Arg	Ser	Glu	Val	Lys	20.4
C	Ala	Phe	Pro	Arg	Lys	Thr	Glu	Ile	Arg	14.8
D	Met	Phe	Pro	His	Arg	Ser	Glu	Val	Lys	7.4
E	Ala	Tyr	Gln	Arg	Arg	Ser	Asp	Val	Lys	7.4

Milk Protein 6					
Variant	NXX	NXY	NXZ	NXA	Freq (%)
A	Ala	Asn	Glu	His	53.7
B	Ala	Ser	Lys	Gln	16.7
C	Ala	Asn	Glu	Gln	13.9
D	Val	Asn	Glu	His	6.5

Milk Protein 7		
Variant	NXX	Freq (%)
A	Arg	55.6
B	Gln	41.7

Milk Protein 8			
Variant	NXX	NXY	Freq (%)
A	Ala	Lys	79.6
B	Ser	Lys	11.1
C	Ala	Glu	9.3

Milk protein 3 B variant has histidine to tyrosine substitution at position NXY and is the fourth highest scoring protein variant. Contrary to milk protein 2 C, this variant ranks high because of its digestion and properties score. This variant is predicted to generate 5 different peptides after digestion and is considered differ in 4 properties compared to the milk protein 3 reference protein.

Milk protein 4 has two protein variants and both of them ranked high in our ranking – fifth and seventh. They both have similarly high digestion score (6), but the B variant has better properties and substitution score. The B variant has substitutions at position NXX and NXY (isoleucine to valine and histidine to tyrosine), while the C variant only has the first substitution. The frequency of the C variant is relatively low, 5.6%. Further investigation on the DNA sequence reads reveals that this variant might be a result of false genotyping.

Milk protein 5 D variant ranked sixth in the list. It has high substitution score because there are 4 substitutions occur in this variant. However, these substitutions are not the most unusual ones. This variant also has decent digestion (2) and properties value (3).

Milk protein 6 B variant has AsnNXYSer, GluNXZLys, and HisNXAGln substitutions. This variant has relatively high properties score, 4, accompanied by fair digestion and substitution score.

Milk protein 7 only has one variant and ranked tenth in our list. A substitution occurred at position NXX from arginine to glutamine. The frequency of this variant is relatively high, 41.7%, making it almost as abundant as the reference variant A.

Milk protein 8 has two substitutions at position NXX and NXY, which comprise of alanine to serine and lysine to glutamic acid substitution respectively. The B variant contains the first substitution while the C variant contains the second. Both variants have the same substitution score, but only the C variant featured in the list because it is predicted to produce different peptides after digestion.

LC-MS analysis is performed on milk protein 4. FrieslandCampina product of pure milk protein 4 (95%) is used. Milk protein 4 variant sequences and the *in silico* digestion result is provided to match the peptides measured by the machine. Peptide sequence position NNX-NNY, which contains isoleucine to valine substitution was detected. This confirms the presence of at least the B variant. On the other hand, peptide sequence position NNI-NNJ which contains histidine to tyrosine substitution was also found but does not pass the filtering threshold. A peptide with similar mass is detected but there are only two peptide fractions that support the presence of the peptide (as opposed to 5 as the filtering threshold). The mass spectra of both peptides are depicted in Figure 4.

(picture removed)

Figure 4. Mass spectra of NNX-NNY peptide (top) and NNI-NNJ peptide (bottom)

The ratio of intensity of substituted NNX-NNY peptide and the original peptide is about 0.25, which is also the ratio of milk protein 4 B variant and A variant frequency derived from the protein variant determination analysis. On the other hand, the ratio of intensity between substituted and original NNI-NNJ peptide is 0.15.

Discussion

It is important to note that the variants found in this research is only based on 54 bulls. There is a real possibility that more variant will emerge and the proportion of variants will be altered if the data are obtained from a larger number and more diverse bulls.

The variant determination is successfully performed on the major milk proteins. α_{s1}-casein and α_{s2}-casein have no identified variant. This result is

fitting since Dutch cattles are heavily selected on α_{s1}-casein A and α_{s2}-casein A. The same case also applies for α-lactalbumin where only B variant is present in Dutch cattles.

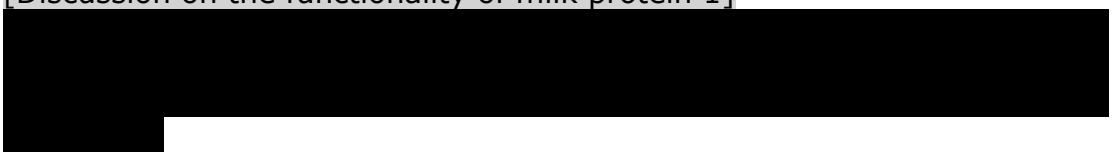
Two variants of β-casein found, making the protein exists in three forms: A2, A1, and I. These are the common variants found in Dutch cattles. The A2 variant accounts for 63% of β-casein, while A1 and I variants account for 17.6% and 13.9% respectively. This composition is slightly different with Demeter et al. (2010) study where the proportion of A2, A1, I, and B β-casein variants are 50.7%, 28.2%, 19%, and 2% respectively.

κ-casein found in two forms, A and B, which are the common κ-casein variants present in Dutch milk. The proportion of A and B variants are 54.7% and 37%, a bit different but still in line with Demeter et al. (2010) study which are 62.6% and 27.9% respectively. Interestingly, we did not find E variant which makes up 9.5% allele population in said study.

Three forms of β-lactoglobulin found – A, B, and one new protein variant X. The X variant is proved to be a false positive because of low quality reads. The proportion of A and B variants are quite similar with Demeter et al. (2010) study (54:38 compared to 58:42).

Our small sample size (54 bulls) and the fact that we use bulls instead of cows might be the cause of the absent of κ-casein E variant and the slight difference in variant proportions of major milk proteins. Cows generally are more uniformed genetically because they are sired from only a handful of bulls. One or a few bulls can influence the majority of cow's genetic make up. In addition, low quality sequence reads which resulted in false genotyping also alter the proportion slightly.

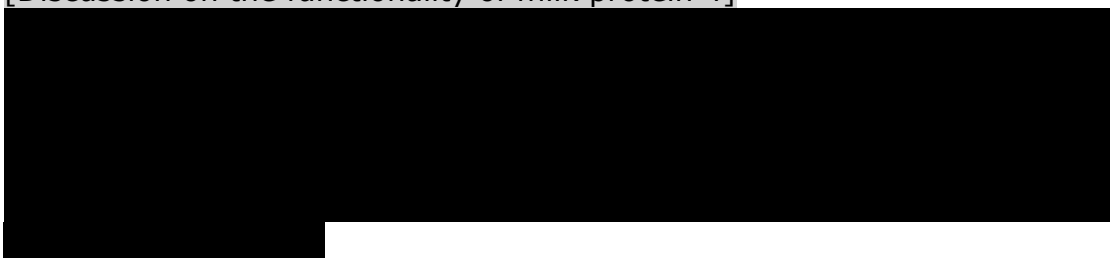
[Discussion on the functionality of milk protein 1]



The frameshift mutation of milk protein 1 C variant occurs at position NXZ, meaning that the rest of the protein will completely altered. This is interesting because the transmembrane region are located in the C-terminus of the protein and will likely be altered. In addition, the frameshift mutation will also affect 4 domains, suggesting that the mutated milk protein 1 protein is not functional at all. Interestingly, three bulls (SIRE18, SIRE24, and SIRE34) has homozygous milk protein 1 C allele. It will be

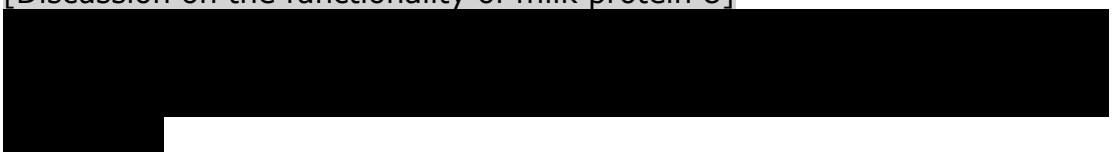
interesting to confirm this finding by proteomic analysis and analyze the immune properties of the homozygous bulls.

[Discussion on the functionality of milk protein 4]



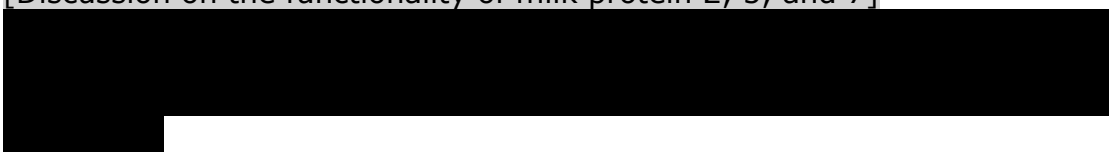
Milk protein 4 B variant differs in position NXX and NXY. The first mutation is located in a binding domain, but not in the binding residue. The mutation of isoleucine to valine in this position may not affect the binding activity because the similar properties of isoleucine and valine side chains. The second mutation is located in a turn and will not likely to affect the activity either. However, this mutation produces 6 different peptides than the reference. Even though it is not altering the bioactive peptide, it will be interesting to see the properties of these peptides.

[Discussion on the functionality of milk protein 8]



Milk protein 8 C variant has different charge and pI compared to the reference protein because of positively charged lysine substitution to negatively charged glutamine at position NXY. This different may change its affinity to substrate and shift its enzymatic activity.

[Discussion on the functionality of milk protein 2, 5, and 7]

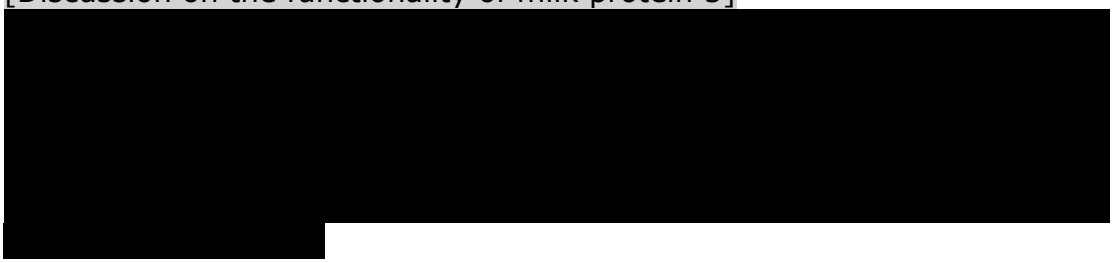


Milk protein 2 C variant has a proline to leucine mutation in its repeated region. This is an interesting mutation because proline is not a common amino acid to be substituted, as indicated by the negative value of the BLOSUM score. Structural damage analysis indicates that this mutation would lead to expansion of cavity volume and damage the structure of the protein.

[Discussion on the functionality of milk protein 2]

The repeated region offers a lot of glycosylation sites, so it will be interesting to see if a structural damage in this region would affect the presence of milk protein 2 in MFGM complex and its binding activity to bacteria.

[Discussion on the functionality of milk protein 5]



In this analysis, milk protein 5 has four variants with no one variant in dominant proportion. Milk protein 5 D is the most interesting one since it has 4 substitution and is predicted to behave differently than the others.

[Discussion on the functionality of milk protein 7]

Milk protein 7 B variant has an arginine to glutamine substitution at position NXX. The substitution does not lead to structural damage but there is a change in charge from positive to uncharged amino acid.

Even though it is much convenient method and the more sensible approach for minor milk proteins, protein variant mining from genomic data has its own challenges and limitations, which all can be traced back to its indirect way of looking at proteins. First, the process of DNA sequence translation into protein is not as straightforward. There is a lot of inbetween process such as alternative splicing and post-translational modification, the processes which are based on genomic region such as intron, exon, termination region, and untranslated region. The problem is there are a lot of genomic annotations for each protein. One protein can have more than one defined gene region, transcript entry, and protein sequence.

Fortunately, a lot of bovine proteins has been manually curated and have one sequence entry as reference. However, VEP frequently cannot annotate the DNA variant to these reviewed entries and instead annotate it to the unreviewed ones. In the end, we still need to manually review the protein annotation ourselves.

There is also a challenge of protein and gene naming convention. Most proteins have more than one gene names, and even though there is an effort to standardize gene names, a lot of bovine protein does not have an official gene name yet.

This research is also limited by the quality of DNA sequence reads. Automatic genotyping of individual bulls based on the sequence read is a tricky process and prone to error. A sequence misread in one read can be interpreted as a heterozygous allele in the region where the read coverage is low. Conversely, a heterozygous allele can be considered as homozygous because there is no alternative allele detected in the region with very low read coverage.

Genotyping is of course in the base of our pipeline and is heavily important in determining the outcome of protein variants finding. There are several

instances in this research that a protein variant turns out to be a false positive because of low quality reads in the particular protein's gene. Again, we need to manually confirm the presence of a variant by looking at the DNA sequence reads data before expecting them to be present in our sample.

As mentioned in the previous section, number of sample is one of the limitation of this research. Fifty four bulls may not be enough to find certain protein variants. A protein variant which has lower than 5% frequency in overall Dutch Holstein population may not be present in our bulls. Another limitation is that the bull population is already selected for certain characteristics – mainly related to milk production. This makes the bulls genetic make up are already quite similar and it is harder to find distinct protein variants.

Nevertheless, this research has successfully discovered new milk protein variants. We discover variants of approximately 400 proteins in very little time and resource compared to traditional proteomic analysis. We found protein variants of major milk proteins that have been described before, suggesting the right principle of our method. We made a scoring model to assess protein variants relevancy and picked 10 variants that we think is the most interesting to study further.

Most importantly, we found proof of our variant existence by LC-MS analysis. Milk protein 4 mutation in position NXX is validated, whereas the second mutation in position NXY is also detected although not as convincing as the first one. All in all, this result suggests the presence of milk protein 4 B variant.

References

Bosaeus, I. (1991). Milk and cholesterol. *Vår Föda*, 43:98-101.

Birgisdottir, B. E., Hill, J. P., Harris, D. P., Thorsdottir, I. (2002). Variation in consumption of cow milk proteins and lower incidence of Type 1 diabetes in Iceland vs the other 4 Nordic countries. *Diabetes Nutr Metab*, 15:240-5.

[redacted]

Demeter, R. M., Markiewicz, K., van Arendonk, J. A. M., & Bovenhuis, H. (2010). Relationships between milk protein composition, milk protein variants, and cow fertility traits in Dutch holstein-friesian cattle. *Journal of Dairy Science*, *93*(11), 5495–5502. https://doi.org/10.3168/jds.2010-3525

[redacted]

Eichholzer, M., Stahelin, H. (1993). Is there a hypocholesterolemic factor in milk and milk products?. *Int J Vitam Nutr Res*, 63(3):158-167.

Elliott, R. B. (1992). Epidemiology of diabetes in Polynesia & New Zealand: In Epidemiology and Etiology of Insulin-Dependent Diabetes in the Young. *eds Levy-Marchal, C. & Czernichow, P.*, Vol 21, pp 66–71. Basel: Karger.

Farrell, H. M., Jimenez-Flores, R., Bleck, G. T., Brown, E. M., Butler, J. E., Creamer, L. K., Hicks, C. L., Hollar, C. M., Ng-Kwai-Hang, K. F., & Swaisgood, H. E. (2004). Nomenclature of the proteins of cows' milk—sixth revision. *Journal of Dairy Science*, *87*(6), 1641–1674. https://doi.org/10.3168/jds.s0022-0302(04)73319-6

Fehily, A. M., Yarnell, J. W., Sweetnam, P. M., & Elwood, P. C. (1993). Diet and incident ischaemic heart disease: The Caerphilly Study. *British Journal of Nutrition*, *69*(2), 303–314. https://doi.org/10.1079/bjn19930035

Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., & Aebersold, R. (2000). Evaluation of two-dimensional gel electrophoresis-based Proteome Analysis Technology. *Proceedings of the National Academy of Sciences*, *97*(17), 9390–9395. https://doi.org/10.1073/pnas.160270797

Insel, P. M., Turner, R. E., & Ross, D. (2004). *Student note-taking guide to accompany nutrition, Second edition*. Jones and Bartlett Publishers.

Ittisoponpisan, S., Islam, S. A., Khanna, T., Alhuzimi, E., David, A., & Sternberg, M. J. E. (2019). Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? *Journal of Molecular Biology*, *431*(11), 2197–2212. https://doi.org/10.1016/j.jmb.2019.04.009

Jensen, R. G. (1995). *Handbook of Milk Composition*. Academic Press.

[redacted]

Kanu, A. B., Dwivedi, P., Tam, M., Matz, L., & Hill, H. H. (2008). Ion mobility-mass spectrometry. *Journal of Mass Spectrometry*, *43*(1), 1–22. https://doi.org/10.1002/jms.1383

[redacted]

[redacted]

[redacted]

[redacted]

Lorenzo, J. M., Munekata, P. E. S., Gómez, B., Barba, F. J., Mora, L., Pérez-Santaescolástica, C., & Toldrá, F. (2018). Bioactive peptides as natural antioxidants in food products – A Review. *Trends in Food Science & Technology*, *79*, 136–147. https://doi.org/10.1016/j.tifs.2018.07.003

Maillet, N. (2019). Rapid peptides generator: FAST and efficient in silico protein digestion. *NAR Genomics and Bioinformatics*, *2*(1). https://doi.org/10.1093/nargab/lqz004

Marckmann, P., Sandström, B., & Jespersen, J. (1994). Low-fat, high-fiber diet favorably affects several independent risk markers of ischemic heart disease: Observations on blood lipids, coagulation, and fibrinolysis from a trial of middle-aged danes. *The American Journal of Clinical Nutrition*, *59*(4), 935–939. https://doi.org/10.1093/ajcn/59.4.935

[redacted]

May, C., Brosseron, F., Pfeiffer, K., Meyer, H. E., & Marcus, K. (2012). Proteome Analysis with classical 2D-page. *Methods in Molecular Biology*, 37–46. https://doi.org/10.1007/978-1-61779-885-6_3

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl variant effect predictor. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0974-4

Mensink, R. P., & Katan, M. B. (1992). Effect of dietary fatty acids on serum lipids and lipoproteins. A meta-analysis of 27 trials. *Arteriosclerosis and Thrombosis: A Journal of Vascular Biology*, *12*(8), 911–919. https://doi.org/10.1161/01.atv.12.8.911

Mensink, R. P., Zock, P. L., Kester, A. D. M., & Katan, M. B. (2003). Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: A meta-analysis of 60 controlled trials. *The American Journal of Clinical Nutrition*, *77*(5), 1146–1155. https://doi.org/10.1093/ajcn/77.5.1146

Ness, A. R. (2001). Milk, coronary heart disease and mortality. *Journal of Epidemiology & Community Health*, *55*(6), 379–382. https://doi.org/10.1136/jech.55.6.379

O'Riordan, N., Kane, M., Joshi, L., & Hickey, R. M. (2014). Structural and functional characteristics of bovine milk protein glycosylation. *Glycobiology*, *24*(3), 220–236. https://doi.org/10.1093/glycob/cwt162

Osorio, D., Rondón-Villarreal, P., & Torres, R. (2015). Peptides: A package for data mining of antimicrobial peptides. *The R Journal*, *7*(1), 4. https://doi.org/10.32614/rj-2015-001

O'Donnell, R., Holland, J. W., Deeth, H. C., & Alewood, P. (2004). Milk proteomics. *International Dairy Journal*, *14*(12), 1013–1023. https://doi.org/10.1016/j.idairyj.2004.04.004

[redacted]

Seidel, C., Deufel, T., & Jahreis, G. (2005). Effects of fat-modified dairy products on blood lipids in humans in comparison with other fats. *Annals of Nutrition and Metabolism*, *49*(1), 42–48. https://doi.org/10.1159/000084176

Smedman, A. E. M., Gustafsson, I.-B., Berglund, L. G. T., & Vessby, B. O. H. (1999). Pentadecanoic acid in serum as a marker for intake of milk fat: Relations between intake of milk fat and metabolic risk factors. *The American Journal of Clinical Nutrition*, *69*(1), 22–29. https://doi.org/10.1093/ajcn/69.1.22

Stähelin, H. B., Eichholzer, M., & Gey, K. F. (1992). Nutritional factors correlating with cardiovascular disease: Results of the basel study. *Nutrition and Cardiovascular Risks*, 24–35. https://doi.org/10.1159/000421431

Warensjö, E., Jansson, J.-H., Berglund, L., Boman, K., Ahrén, B., Weinehall, L., Lindahl, B., Hallmans, G., & Vessby, B. (2004). Estimated intake of milk fat is negatively associated with cardiovascular risk factors and does not increase the risk of a first acute myocardial infarction. A prospective case–control study. *British Journal of Nutrition*, *91*(4), 635–642. https://doi.org/10.1079/bjn20041080

Willett, W. (1993). Intake of trans fatty acids and risk of coronary heart disease among women. *The Lancet*, *341*(8845), 581–585. https://doi.org/10.1016/0140-6736(93)90350-p

Appendices

Appendix A. Protein variant determination script

Appendix B. Feature analysis script

Appendix C. Protein variant sequence generation script

Appendix D. Rapid peptide generator script

Appendix E. Number of variants found in each protein