Tandem repeats structure of gel-forming mucin domains could be revealed by SMRT sequencing data

Tiange Lang (✉ langtiange@jnu.edu.cn)
Jinan University

Research Article

Keywords: mucin, domains, forming, sequence, muc2

Posted Date: October 13th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-112828/v2>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Tandem repeats structure of gel-forming mucin domains could be revealed by SMRT sequencing data

Author number: 1.

Name of the author: Tiange Lang

Institution of the author: Big Data Decision Institution, Jinan University, China

Country of the author: China

Email address of the author: langtiange@jnu.edu.cn

Full post address of the submitting author: No 601 Huangpu Avenue West, Tianhe District, Guangzhou, China. Post code: 510632.

Abstract

Mucins are large glycoproteins that cover and protect epithelial surface of the body. Gel-forming mucin domains of mucin genes are rich in proline, threonine, and serine that are heavily glycosylate. These domains show great complexity with tandem repeats (TRs), thus make it difficult to study the sequences. With the coming of single molecule real-time (SMRT) sequencing technologies, we manage to present sequence structure of mucin domains via SMRT long reads for gel-forming mucins MUC2, MUC5AC, MUC5B and MUC6. Our study shows that for different individuals, single nucleotide polymorphisms (SNPs) could be found in mucin domains of MUC2, MUC5AC, MUC5B and MUC6, while different number of tandem repeats could be found in mucin domains of MUC2 and MUC6. Furthermore, we get the sequence of MUC2, MUC5AC, and MUC5B mucin domain in a Chinese individual at accuracy of possibly maximum 99.98%, 99.93%, and 99.76%, respectively. We report a new method to obtain DNA sequence of gel-forming mucin domains. This method will provided new insights on getting the sequence for Tandem Repeat parts which locate in coding region. With the sequences we obtained with this method, we can give more information for people to study the sequences of gel-forming mucin domains.

Background

The gel-forming mucins are large glycosylated proteins essential for mucus layer to cover epithelial cells(1). The gel-like structure of mucins could protect against harmful microorganisms(2). There are five gel-forming mucins in mammals(1). They are MUC2, MUC5AC, MUC5B, MUC6 and MUC19. Each protein contains a mucin domain which is rich in amino acids proline (P), threonine (T) and serine (S). The mucin domain, also called PTS domain, is heavily glycosylated and therefore has a stiff conformation which could give the mucin protein the function of protection(3).

For each gel-forming mucin, the sequence before the mucin domain is called N-terminal sequence and the sequence afterwards is called C-terminal sequence. All the N-terminal sequences of gel-forming mucins share a similar structure which contains three VWD domains, and all the C-terminal sequences have Cysteine knot (CK) domains(4-7). Both the cysteine number and their positions are extremely conserved in those domains, which play an essential role in forming dimers and trimers.

In human, the genes of MUC2, MUC5AC, MUC5B and MUC6 are clustered in a complex of 400 kb very rich in CpG islands on chromosome 11 in region p15.5(8, 9). Computational and phylogenetic analyses suggested an evolutionary history of the four human mucin genes from an ancestor gene common to the human von Willebrand factor gene (7, 10).

In human, DNA sequence of mucin domain in MUC2/MUC5AC/MUC5B/MUC6 is only one exon. This exon is organized in tandem repeats. These tandem repeats are very large, and the length of each repeat could vary (9, 11). Therefore, it is very difficult to get the exact sequence of this part.

SMRT DNA sequencing technology could generate reads up to 10k-20k bases (12). This length could span the whole mucin domain. Therefore, the sequence of mucin domain could be assembled with SMRT reads (13). However, this technology has high error rate (14). Thus a high coverage of the reads is needed to obtain the sequence of mucin domain (15).

Several human reference genomes are available, and these could be used to locate mucin domains (13, 14, 16). Moreover, a whole human genome SMRT reads of a Chinese individual (HX1) are available on the website <http://hx1.wglab.org/>, and this could be used to make the assembly of mucin domains (13).

Methods

Downloading of Chr11p15.5 of human genome of reference sequence (Refseq), a Korean individual, and an American individual

From NCBI, we downloaded whole chromosome 11 of human genome of Refseq , a Korean individual (BioProject PRJNA294231), and an American individual (BioProject PRJNA294231). Then we extracted whole chromosome 11 from 1M to 1.4M.

In Refseq, this region includes six genes. They are APA2 (forward strand, position 925,809 to 1,012,245), MUC6 (reverse strand, position 1,012,823 to 1,036,706), MUC2 (forward strand, position 1,074,875 to 1,110,508), MUC5AC (forward strand, position 1,157,953-1,201,138), MUC5B (forward strand, position 1,239,777 to 1,249,564), and TOLLIP (reverse strand, position 1,274,368 to 1,309,662).

Downloading of whole human genome sequence read archive (SRA) data of a Chinese individual (HX1) and converting of SRA data to fasta sequence data

We downloaded whole human genome SRA data of a Chinese individual (HX1) from NCBI. There are two types of SRA data. One is PACBIO_SMRT data from PacBio RS II, and the other is ILLUMINA data from Illumina Hiseq 2000. We only downloaded PACBIO_SMRT data and the size is 2,773.6Gbp.

For the genome data of HX1 downloaded, we used perl and python scripts to convert SRA data to fasta sequence data (supplementary S1 and S2). In the converting process, we only kept the reads which were longer than 5k bases. The number of reads kept was 24,290,526. The number of bases kept was 284,581,254,229, and that was about 94.86X coverage.

Obtaining of mucin domain of MUC2/MUC5AC/MUC5B/ MUC6 in the assembly of Refseq, Korean individual and American individual

In human genome, the mucin domains of MUC2/MUC5AC/MUC5B/ MUC6 have only one exon. Thus first we found the DNA sequence of mucin domains in NCBI MUC2 (Accession number NM_002457.4), MUC5AC (Accession number NM_001304359.1), MUC5B (Accession number NM_002458.2), and MUC6 (Accession number NM_005961.2). Next we used the sequence found as query sequences doing similarity search into genome assembly of Refseq in Chr11p15.5, Korean individual in Chr11 and American individual.

Obtaining of mucin domain of MUC2/MUC5AC/MUC5B/ MUC6 in the SMRT reads of HX1

First we found the DNA sequence of the exons of mucin domains in NCBI MUC2 (Accession number NM_002457.4), MUC5AC (Accession number NM_001304359.1), MUC5B (Accession number NM_002458.2) and MUC6 (Accession number NM_005961.2). Next for each mucin, we took introns

up-stream and down-stream of the mucin domain exon. Then we used the combination of the three fragments as query sequences doing similarity search into SMRT reads of HX1.

Bioinformatic methods

For similarity search we used BLAST (17). For multiple sequence alignment we used CLUSTALW (18). Genetic codon table for human was used for translation. HMMER was used for identifying protein domains (19). Perl and python scripts were used for other tasks (*supplementary materials*).

Results

Mucin domain of MUC2/MUC5AC/MUC5B/ MUC6 in Refseq, Korean, and American assembly

The length of mucin domains of MUC2 of Refseq, Korean and American assembly are 5884 bases, 4471 bases and 8773 bases, respectively. In NCBI there is a nucleotide entry of human MUC2 DNA sequence with Accession number NM_002457.4. The mucin domain of this entry is 9270 bases and only a part of it could be found in Refseq. However, we still regard the mucin domain of NM_002457.4 as the most complete assembly among all available MUC2 mucin domain assemblies by similarity search.

The length of mucin domains of MUC5AC of Refseq, Korean and American assembly are 10371 bases, 10576 bases and 11196 bases, respectively. We regard the mucin domain of MUC5AC in Refseq as the most accurate assembly among all the three assemblies by similarity search.

The length of mucin domains of MUC5B of Refseq, Korean and American assembly are 10893 bases, 11589 bases and 10772 bases, respectively. We regard the mucin domain of MUC5B in Refseq as the most accurate assembly among all the three assemblies by similarity search.

The lengths of mucin domains in MUC6 of Refseq, Korean and American assembly are 3009 bases, 13299 bases and 8727 bases, respectively. For the American mucin domain of MUC6, the heads goes into gap region. We regard the mucin domain of MUC6 in Korean individual as the most complete assembly among the three assemblies by similarity search.

We could see that even by length, none of these domains are identical.

Programing pipelines to get consensus sequence with SMRT reads

We used quality control and length control on reads downloaded with programming scripts (supplementary S3 and S4). By similarity search with BLAST, we managed to find reads which could cover each mucin domain (17). Then we used multiple alignment method CLUSTALW to get an alignment (18). In the alignment, for each column we took the nucleotide (including insertion, i.e.

A/T/C/G/-) which appears maximum number of times. Then we removed all the insertions and got the consensus sequence (supplementary S5). Next we aligned back the consensus to all reads, and corrected the errors which were caused by several insertions together with one nucleotide or several same nucleotides together with one insertion (Fig. 1). For example in the part of three '-'s and one 'A', in the alignment some reads give '- - - A' and some give 'A - - -'. Thus a nucleotide could be replaced by an insertion and later be removed in the consensus, and this will cause a missing of a nucleotide. Same principle, several same nucleotides together with an insertion might cause a redundant nucleotide in the consensus. This problem could cause a frame shift, and we managed to correct it according to the translation result. If we regard the error rate of SMRT as 15% (20), for each position, the error rate is 0.15 to the power of the number of reads which could be aligned in this position.

Assembly of mucin domain of MUC2 in HX1 with SMRT reads

For MUC2, in all the SMRT reads downloaded, 4 reads could be found to cover both the intron up-stream of mucin domain exon and the intron down-stream of mucin domain exon (Fig. 2). In human MUC2 mucin domain there is a CysD domain in the middle of a domain which is full of proline, threonine and serine (PTS). Although TR structure of PTS domain makes it impossible to be identified precisely with similarity search, CysD domain could be precisely found. Therefore, we took use of the CysD domain in the middle of PTS exon. We regarded the last "cysteine" in CysD domain as delimiter. Thus we searched for two types of read: I. intron up-stream + up-stream part of exon; II. down-stream part of exon + intron down-stream. We found 18 type I and 9 type II reads. Combining with the 4 reads which could be found to cover introns both up-stream and down-stream, 22 reads could be used to build up-stream part and 13 reads could be used to build down-stream part (Fig. 2). For up-stream part, in one position at least 11 reads could be aligned, thus the maximum error rate is 0.15 to the power 11 and the accuracy is 99.9999999%. For down-stream part, in one position at least 7 reads could be aligned, thus the maximum error rate is 0.15 to the power of 7 and the accuracy is 99.9998%. The whole mucin domain exon of MUC2 in HX1 has 8994 bases.

MUC2 mucin domain TR structure

The protein sequence of up-stream part of MUC2 mucin domain in HX1 has 2 CysD domains at the beginning and the end (Fig. 2). They have 95 and 97 amino acids, respectively. Between the two CysD domains is the PTS TR up-stream part (Fig. 3A). The TR lengths vary a lot. We define each TR with a symbol "PS" at the start of each TR. Therefore, the PTS TR up-stream part has 28 TRs. The shortest TR has 7 amino acids and the longest TR has 26 amino acids. The protein sequences of PTS TR up-stream part as well as two CysD domains of MUC2 mucin domain in NCBI (Nucleotide accession number NM_002457.4; Protein accession number NP_002448.4) are exactly the same as those in HX1.

The PTS sequence after 2nd CysD domain of MUC2 mucin domain in HX1 is PTS TR down-stream part and it has 101 TRs (Fig. 3B). 98 TRs have 23 amino acids, respectively. 3rd TR has 24 amino acids. 4th TR has 22 amino acids. 57th TR has 21 amino acids.

The protein sequence of PTS TR down-stream part of MUC2 mucin domain in NCBI (Nucleotide accession number NM_002457.4; Protein accession number NP_002448.4) has 105 TRs. Comparing with that of MUC2 mucin domain in HX1, it has 4 more TR CNVs directly after 89th repeat and 18 SNPs at repeat 15, 18, 25, 25, 38, 38, 39, 39, 39, 40, 40, 40, 41, 41, 42, 42, 60, and 104 (100 for HX1), respectively (Fig. 3C and 3D).

Assembly of mucin domain of MUC5AC in HX1 with SMRT reads

For MUC5AC, in all the SMRT reads downloaded, 10 reads could be found to cover both the intron up-stream of mucin domain exon and the intron down-stream of mucin domain exon. In one position at least 6 reads could be aligned, thus the maximum error rate is 0.15 to the power 6 and the accuracy is minimum 99.9988%. The whole mucin domain exon of MUC5AC in HX1 has 10371 bases.

MUC5AC mucin domain TR structure

The protein sequence of MUC5AC mucin domain in HX1 has one main head, one main tail, 6 CysD domains, 2 Long Tandem Repeat (LTR) groups, 4 Short Tandem Repeat (STR) groups and 1 unique short piece (Fig. 4A). The main head is composed of a PTS domain of 45 amino acids long and a CysD like domain of 99 amino acids long (Fig. 4B). The main tail is composed of a PTS domain of 50 amino acids long and a short piece of 12 amino acids long (Fig. 4C). For all 6 CysD domains, each has 105 amino acids and locates after each LTR/STR group (Fig. 4L). For 2 LTR groups, each is composed of one PTS domain of 95 amino acids long, one CysD like domain of 101 amino acids long, and one PTS domain of 65 amino acids long (Fig. 4E). Other than one CysD domain, there is one small PTS piece of 7 amino acids long between the 2 LTR groups (Fig. 4D). For 4 STR groups, each has one PTS head of 36 amino acids long and one PTS tail of 13 amino acids long (Fig. 4J and 4K). 1st, 2nd, 3rd and 4th STR group have 119, 18, 35, and 65 STRs, respectively (Fig. 4F, 4G, 4H and 4J). Each repeat has 8 amino acids except that 14th repeat of 3rd STR group has only 7 amino acids (Fig. 4F, 4G, 4H and 4I). As the delimiters of LTR/STR groups, 6 CysD domains are quite similar (Fig. 4L).

The protein sequence of MUC5AC mucin domain in NCBI (Nucleotide accession number NM_001304359.1; Protein accession number NP_001291288.1) has same length and TR structure as the protein sequence of MUC5AC mucin domain HX1. There are only 3 SNPs. One is in 99th repeat in 1st STR group; another two are in 3rd and 4th CysD, respectively (Fig. 4M).

Assembly of mucin domain of MUC5B in HX1 with SMRT reads

For MUC5B, in all the SMRT reads downloaded, 9 reads could be found to cover introns both up-stream and down-stream of mucin domain exon. In one position at least 5 reads could be aligned, thus the maximum error rate is 0.15 to the power 5 and the accuracy is 99.99%. The whole mucin domain exon of MUC5B in HX1 has 10893 bases.

MUC5B mucin domain TR structure

The protein sequence of MUC5B mucin domain in HX1 has one main head, one main tail, 1 Cys-similar domain, 6 CysD domains, and 7 PTS domains (Fig. 5A). The main head is composed of a small piece of 8 amino acids long (Fig. 5B). The main tail is composed of a small piece of 12 amino acids long (Fig. 5C). The CysD-similar domain has 100 amino acids (Fig. 5M). 2nd CysD domain has 102 amino acids (Fig. 5L). For other 5 CysD domains, each has 101 amino acids (Fig. 5L). For all 7 CysD and CysD-similar domains, each locates before each PTS domain (Fig. 5A). The first 2 PTS domains have no repeats, but a long piece of 70 and 180 amino acids, respectively (Fig. 5D and 5E). Each of the last 5 PTS domains has some STRs and one PTS tail (Fig. 5F, 5G, 5H, 5I, 5J, 5K, and 5N). The number of STRs of the bodies of 3rd, 4th, 5th, 6th, and 7th PTS domain are 10, 11, 16, 11, and 22, respectively (Fig. 5F, 5G, 5H, 5I and 5J). For all the STRs in the bodies of last 5 PTS domains, 5 have 24 amino acids, respectively; 8 have 26 amino acids, respectively; 8 have 28 amino acids, respectively; 48 have 29 amino acids, respectively; one has 34 amino acids (Fig. 5F, 5G, 5H, 5I and 5J). The PTS tails of 3rd, 4th, 5th and 6th PTS domains are homologous and they all have 147 amino acids, respectively (Fig. 5K). The PTS tail of 7th PTS domain has 87 amino acids (Fig. 5N). As the delimiters of 7 PTS domains, 2nd, 3rd, 4th, 5th, and 6th CysD domains are quite similar (Fig. 5L).

The protein sequence of MUC5B mucin domain in NCBI (Nucleotide accession number NM_002458.2; Protein accession number NP_002449.2) has same length and TR structure as that in HX1. There are only 7 SNPs. Three are in PTS tail of 3rd PTS domain, one is in 11th repeat of 5th PTS domain, one is in 4th repeat of 6th PTS domain, and three are in 7th, 8th and 9th repeats of 7th PTS domain, respectively (Fig. 5O).

Assembly of mucin domain of MUC6 in HX1 with SMRT reads

For MUC6, in all the SMRT reads downloaded, only 3 reads could be found to cover introns both up-stream and down-stream of mucin domain exon. Therefore, it is impossible to get the exactly correct nucleotide in each position. However, due to the TR structure, the number of TRs and the lengths of each TR could be obtained. The MUC6 refseq of NCBI has all the non-TR part of mucin domain, thus we can use this as the template to get the whole mucin domain exon of MUC6 in HX1 which has 13470 bases.

MUC6 mucin domain TR structure

The protein sequence of MUC6 mucin domain in HX1 has one head, one tail and 27 TRs (Fig. 6A). The head has 60 amino acids and the tail has 265 amino acids (Fig. 6B and 6C). 27 TRs could be found between the head and the tail. 1st, 2nd, 3rd, 4th, 5th, 7th, 8th, 9th, 12th, 13th, 14th, 18th, 22nd, and 26th TRs have 169 amino acids, respectively. This number is most of the case among all TRs, thus we call this type of TRs “typical TR”. 6th TR has 171 amino acids, and there is a “TG” insertion comparing with the typical TR. 10th, 11th, 15th, 19th, and 23rd TRs have 168 amino acids, respectively, and there is a deletion comparing with the typical TR. 16th, 20th, and 24th TRs have 150 amino acids, respectively, and they are first 150 amino acids of the typical TR. 17th, 21st, and 25th TRs have 74 amino acids, respectively, and they are last 74 amino acids of the typical TR. 27th TR has 115 amino acids, and it is the first 115 amino acids of the typical TR (Fig. 6D).

The protein sequence of MUC6 mucin domain in NCBI (Nucleotide accession number NM_005961.2; Protein accession number NP_005952.2) only has head, 1st TR, first 33 amino acids of 2nd TR, last 117 amino acids of 24th TR, 25th TR, 26th TR, 27th TR, and tail (Fig. 6E). Since we cannot be sure the exact nucleotide in each position due to only 3 reads available, we cannot say SNP information.

Estimation of number of TRs in down-stream part of mucin domain of MUC2 for another individual

In the result from the pipeline, all frameshifts are caused by several same nucleotides together. In HX1, in the DNA sequence of TRs in down-stream part of mucin domain of MUC2 mucin domain, no continuous multiple “T”s could be found other than two SNPs at 46th and 96th TR, respectively, which cause two “T”s together (Fig. 7). Therefore, the number of “T”s in the TR part from pipeline consensus could be used to estimate the number of repeats without arranging each frameshift (Table 1).

In down-stream part of mucin domain of MUC2 for HX1, the number of “T”s keeps same after adjustment. For each repeat, in most cases “T” comes 2 or 3 or 4 or 5 times, in some less number of cases “T” comes 6 times, and only once “T” comes 7 times. Therefore, for another individual, after checking “T” number we could get repeat number roughly by comparing with HX1. In HX1, the down-stream part has 101 TRs and 364 “T”s. 46th and 96th TR have one “TT”, respectively. Since in common repeats we cannot find “TT”, we regard such cases as SNPs and count two “T”s as one. Therefore “T” number of 362 shall correspond to repeat number 101, and on average each repeat has 3.58 “T”s (Fig. 7). If we divide “T” number difference by 3.58, we can roughly get repeat number difference. For instance, DNA sequence of most complete MUC2 protein in NCBI (accession number NP_002448.4) has 13 more “T”s and 4 more repeats. However, some SNPs (T/A, T/C, or T/G) might affect T number difference. Anyway a roughly estimation of repeat number could be obtained in this way (Fig. 8).

Discussion

The PacBio SMRT data of HX1 human genome was used to get the mucin domain. Since the whole data set has 44.2 million reads, the data with a size of several tens of Terabytes was dealt with.

Therefore, it is essential to make an efficient pipeline to obtain the reads which could be used to assemble the mucin domain.

HX1 human genome assembly was made with Illumina Hiseq data which has 2.8 billion reads (N50 length 151) as well as PacBio SMRT data which has 44.2 million reads (N50 length 12.1 k). MUC2 mucin domain could not be found in the HX1 human genome assembly. MUC5AC mucin domain and MUC5B mucin domain could be found in one contig with 10368 bases and 10789 bases, respectively. MUC6 mucin domain could be found in another contig with 11992 bases. With PacBio SMRT reads and the pipeline, we got the DNA lengths of mucin domains of MUC2, MUC5AC, MUC5B and MUC6 at 8994 bases, 10371 bases (same as NCBI Refseq), 10893 bases (same as NCBI Refseq), and 13470 bases, respectively.

Due to the TR character, it is hard for the alignment softwares like clustalw to make a good alignment if only PTS exon was used. Therefore, it is essential to include the introns up-stream and down-stream. For MUC2, since a CysD island could be found in the middle of PTS region, the alignment could be divided into two parts. The first part could contain the intron up-stream, PTS TRs, and the CysD island, and the second part could contain the CysD island, PTS TRs, and the intron down-stream.

For clustalw, we set gap_opening_penalty and gap_extension_penalty to 0. The reason is that if you consider to align two sequences with different TRs numbers, the default numbers of gap_opening_penalty and gap_extension_penalty in clustalw will cause the interruption of one complete repeat unit. However, the alignment of two common sequences (none TR) has no such problem.

Stop codons or frameshift could appear in the regions where same nucleotide appears several times continuously. To make a better prediction of frameshift, a useful way is to make an algorithm to calculate the probability of having a frameshift in these regions.

CysD domain is a non-PTS motif in the protein sequences encoded by PTS exon. For MUC5AC and MUC5B, CysD domains are more conserved than the repeated regions between individuals. In MUC6, we did not find any CysD domain.

Conclusion

We report a new method to obtain DNA sequence of gel-forming mucin domains. In this method we used SMRT long reads which could cover the whole mucin domains of human MUC2, MUC5AC,

MUC5B and MUC6. We made a programming pipeline to get consensus sequence of the long reads we picked, and then adjusted the consensus sequence by examining the reads. We also found a quick way to estimate the number of repeats in MUC2 by checking the number of nucleotide “T”. With the sequences we obtained, we found that for different human individuals, SNPs could be found in mucin domains of MUC2, MUC5AC, MUC5B and MUC6, while different number of TRs could be found in mucin domains of MUC2 and MUC6. This result could give more information for people to study the sequences of gel-forming mucin domains. Moreover, the method we used could reveal the number of tandem repeats in gel-forming mucin domains in a rather simple way. Therefore, by further knowing the repeat character of these mucin domains, one can understand better about the relationship between sequence features and biological functions of this region.

Abbreviations

TR(s) : tandem repeat(s).
LTR : long tandem repeat.
STR : short tandem repeat.
SMRT : single molecule real-time.
SNP(s) : single nucleotide polymorphism(s).
CNV(s) : copy number variation(s).
PTS : proline, threonine and serine.
VWD : von Willebrand D.
CK : cysteine knot.
HX1 : a Chinese individual.
Refseq : reference sequence.
Chr : chromosome.
NCBI : National Center for Biotechnology Information.
SRA : sequence read archive.
BAC : bacterial artificial chromosome.
Gbp : gigabases.

Declarations

Ethics approval and consent to participate

No need.

Consent for publication

No need.

Availability of data and material

All data are downloaded from NCBI website.

Whole chromosome 11 of human genome Refseq (Accession number NC_000011):

https://www.ncbi.nlm.nih.gov/nuccore/NC_000011.10

Chromosome 11 of human genome assembly of a Korean individual (Accession number GCA_001712695.1): https://www.ncbi.nlm.nih.gov/assembly/GCA_001712695.1

Whole human genome assembly of an American individual (Accession number GCA_001013985.1): https://www.ncbi.nlm.nih.gov/assembly/GCA_001013985.1

Whole human genome SRA data of a Chinese individual (HX1, Accession number GCA_001708065.2): <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA301527/>

Competing interests

No.

Funding

No funding was obtained in this study.

Authors' contributions

TL is the sole author and did everything.

Acknowledgements

The author thanks Gunnar C. Hansson, Malin Johansson, and Frida Svensson of Mucin Biology Group, Department of Medical Biochemistry and Cell Biology, University of Gothenburg, Sweden, for comments and discussion.

References

1. J. Perez-Vilar, R. L. Hill, The structure and assembly of secreted mucins. *The Journal of biological chemistry* **274**, 31751–31754 (1999).
2. G. C. Hansson, M. E. Johansson, The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Gut microbes* **1**, 51–54 (2010).
3. C. G. Davis *et al.*, Deletion of clustered O-linked carbohydrates does not impair function of low density lipoprotein receptor in transfected fibroblasts. *The Journal of biological chemistry* **261**, 2828–2838 (1986).
4. J. R. Gum, J. W. Hicks, Y. S. Kim, Identification and characterization of the MUC2 (human intestinal mucin) gene 5'–flanking region: promoter activity in cultured cells. *The Biochemical journal* **325** (Pt 1), 259–267 (1997).

5. D. Baeckstrom, G. C. Hansson, The transcripts of the apomucin genes MUC2, MUC4, and MUC5AC are large and appear as distinct bands. *Glycoconjugate journal* **13**, 833–837 (1996).
6. K. Rousseau *et al.*, The complete genomic organization of the human MUC6 and MUC2 mucin genes. *Genomics* **83**, 936–939 (2004).
7. G. Javitt *et al.*, Assembly Mechanism of Mucin and von Willebrand Factor Polymers. *Cell* **183**, 717–+ (2020).
8. J. L. Desseyn *et al.*, Evolutionary history of the 11p15 human mucin gene family. *Journal of molecular evolution* **46**, 102–106 (1998).
9. L. E. Vinall *et al.*, Variable number tandem repeat polymorphism of the mucin genes located in the complex on 11p15.5. *Human genetics* **102**, 357–366 (1998).
10. N. Moniaux, F. Escande, N. Porchet, J. P. Aubert, S. K. Batra, Structural organization and classification of the human mucin genes. *Frontiers in Bioscience-Landmark* **6**, D1192–D1206 (2001).
11. J. L. Desseyn, V. Guyonnet-Dupérat, N. Porchet, J. P. Aubert, A. Laine, Human mucin gene MUC5B, the 10.7-kb large central exon encodes various alternate subdomains resulting in a super-repeat. Structural evidence for a 11p15.5 gene family. *J Biol Chem* **272**, 3168–3178 (1997).
12. C.-S. Chin *et al.*, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**, 563–+ (2013).
13. L. Shi *et al.*, Long-read sequencing and de novo assembly of a Chinese genome. *Nature Communications* **7**, (2016).
14. M. Pendleton *et al.*, Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods* **12**, 780–786 (2015).
15. M. Wang *et al.*, PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *Bmc Genomics* **16**, (2015).
16. Y. S. Cho *et al.*, An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nature Communications* **7**, (2016).
17. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389–3402 (1997).
18. J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673–4680 (1994).
19. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29–37 (2011).
20. E. Bao, L. Lan, HALC: High throughput algorithm for long read error correction. *Bmc Bioinformatics* **18**, (2017).

Figure Legends

1. Programing pipeline to get consensus with SMRT reads.
2. MUC2 mucin domain of HX1 and available reads for building MUC2 mucin domain.
3. MUC2 mucin domain of HX1. A) Up-stream part. B) Down-stream part. C) SNP comparison of up-stream part with most complete NCBI entry. D) CNV comparison of down-stream part with most complete NCBI entry.
4. MUC5AC mucin domain of HX1. A) All domains. B) Main head. C) Main tail. D) Small PTS piece. E) Alignment of LTR1 and LTR2. F) STR1 body. G) STR2 body. H) STR3 body. I) STR4 body. J) Alignment of STR heads. K) Alignment of STR tails. L) Alignment of 6 CysD domains. M) SNP comparison with most complete NCBI entry.
5. MUC5B mucin domain of HX1. A) All domains. B) Main head. C) Main tail. D) PTS1. E) PTS2. F) PTS3 body. G) PTS4 body. H) PTS5 body. I) PTS6 body. J) PTS7 body. K) Alignment of tails of PTS3, PTS4, PTS5 and PTS6. L) Alignment of CysD2, CysD3, CysD4, CysD5 and CysD6. M) CysD1. N) PTS7 tail. O) SNP comparison with most complete NCBI entry.
6. MUC6 mucin domain of HX1. A) All domains. B) Head. C) Tail. D) TRs. E) CNV comparison with most complete NCBI entry.
7. DNA sequence of down-stream part TRs of MUC2 mucin domain in HX1. Other than 2 “TT” SNPs at 17th and 96th TR, no continuous “T”s could be found in this part.
8. Steps to estimate number of repeats in down-stream part of MUC2 mucin domain.

Table Legends

1. Nucleotide statistics of MUC2 mucin domain down-stream TR part. We can see that the number of only letter “T” keeps same after adjustment. Therefore that number could be used to estimate the number of tandem repeats before adjustment.

Figures

Figure 1

Programing pipeline to get consensus with SMRT reads.

Figure 2

MUC2 mucin domain of HX1 and available reads for building MUC2 mucin domain.

VTQPTMTTITEN		ITT	
PTTPTTTTTVTPPTPSTQS	1	PSPTTTT	1
TTPTPTTTTTVTPPTPSTQS	2	PSPTPTTTTLPPTTT	2
PTTPTTTTTVTPPTPSTQS	3	PSPTPTTTTTPPTTT	3
PTTPTTTTTVTPPTPSTQS	4	PSPTPTTTTTPPTTT	4
TTPTPTTTTTVTPPTPSTQS	5	PSPTPTTTTTPPTTT	5
PTTPTTTTTVTPPTPSTQS	6	PSPTPTTTTTPPTTT	6
PTTPTTTTTVTPPTPSTQS	7	PSPTPTTTTTPPTTT	7
TTPTPTTTTTVTPPTPSTQS	8	PSPTPTTTTTPPTTT	8
PTTPTTTTTVTPPTPSTQS	9	PSPTPTTTTTPPTTT	9
PTTPTTTTTVTPPTPSTQS	10	PSPTPTTTTTPPTTT	10
TTPTPTTTTTVTPPTPSTQS	11	PSPTPTTTTTPPTTT	11
PTTPTTTTTVTPPTPSTQS	12	PSPTPTTTTTPPTTT	12
PTTPTTTTTVTPPTPSTQS	13	PSPTPTTTTTPPTTT	13
TTPTPTTTTTVTPPTPSTQS	14	PSPTPTTTTTPPTTT	14
PTTPTTTTTVTPPTPSTQS	15	PSPTPTTTTTPPTTT	15
TTPTPTTTTTVTPPTPSTQS	16	PSPTPTTTTTPPTTT	16
PTTPTTTTTVTPPTPSTQS	17	PSPTPTTTTTPPTTT	17
TTPTPTTTTTVTPPTPSTQS	18	PSPTPTTTTTPPTTT	18
PTTPTTTTTVTPPTPSTQS	19	PSPTPTTTTTPPTTT	19
TTPTPTTTTTVTPPTPSTQS	20	PSPTPTTTTTPPTTT	20
PTTPTTTTTVTPPTPSTQS	21	PSPTPTTTTTPPTTT	21
TTPTPTTTTTVTPPTPSTQS	22	PSPTPTTTTTPPTTT	22
PTTPTTTTTVTPPTPSTQS	23	PSPTPTTTTTPPTTT	23
TTPTPTTTTTVTPPTPSTQS	24	PSPTPTTTTTPPTTT	24
PTTPTTTTTVTPPTPSTQS	25	PSPTPTTTTTPPTTT	25
TTPTPTTTTTVTPPTPSTQS	26	PSPTPTTTTTPPTTT	26
PTTPTTTTTVTPPTPSTQS	27	PSPTPTTTTTPPTTT	27
TTPTPTTTTTVTPPTPSTQS	28	PSPTPTTTTTPPTTT	28
PTTPTTTTTVTPPTPSTQS	29	PSPTPTTTTTPPTTT	29
TTPTPTTTTTVTPPTPSTQS	30	PSPTPTTTTTPPTTT	30
PTTPTTTTTVTPPTPSTQS	31	PSPTPTTTTTPPTTT	31
TTPTPTTTTTVTPPTPSTQS	32	PSPTPTTTTTPPTTT	32
PTTPTTTTTVTPPTPSTQS	33	PSPTPTTTTTPPTTT	33
TTPTPTTTTTVTPPTPSTQS	34	PSPTPTTTTTPPTTT	34
PTTPTTTTTVTPPTPSTQS	35	PSPTPTTTTTPPTTT	35
TTPTPTTTTTVTPPTPSTQS	36	PSPTPTTTTTPPTTT	36
PTTPTTTTTVTPPTPSTQS	37	PSPTPTTTTTPPTTT	37
TTPTPTTTTTVTPPTPSTQS	38	PSPTPTTTTTPPTTT	38
PTTPTTTTTVTPPTPSTQS	39	PSPTPTTTTTPPTTT	39
TTPTPTTTTTVTPPTPSTQS	40	PSPTPTTTTTPPTTT	40
PTTPTTTTTVTPPTPSTQS	41	PSPTPTTTTTPPTTT	41
TTPTPTTTTTVTPPTPSTQS	42	PSPTPTTTTTPPTTT	42
PTTPTTTTTVTPPTPSTQS	43	PSPTPTTTTTPPTTT	43
TTPTPTTTTTVTPPTPSTQS	44	PSPTPTTTTTPPTTT	44
PTTPTTTTTVTPPTPSTQS	45	PSPTPTTTTTPPTTT	45
TTPTPTTTTTVTPPTPSTQS	46	PSPTPTTTTTPPTTT	46
PTTPTTTTTVTPPTPSTQS	47	PSPTPTTTTTPPTTT	47
TTPTPTTTTTVTPPTPSTQS	48	PSPTPTTTTTPPTTT	48
PTTPTTTTTVTPPTPSTQS	49	PSPTPTTTTTPPTTT	49
TTPTPTTTTTVTPPTPSTQS	50	PSPTPTTTTTPPTTT	50
PTTPTTTTTVTPPTPSTQS	51	PSPTPTTTTTPPTTT	51
TTPTPTTTTTVTPPTPSTQS	52	PSPTPTTTTTPPTTT	52
PTTPTTTTTVTPPTPSTQS	53	PSPTPTTTTTPPTTT	53
TTPTPTTTTTVTPPTPSTQS	54	PSPTPTTTTTPPTTT	54
PTTPTTTTTVTPPTPSTQS	55	PSPTPTTTTTPPTTT	55
TTPTPTTTTTVTPPTPSTQS	56	PSPTPTTTTTPPTTT	56
PTTPTTTTTVTPPTPSTQS	57	PSPTPTTTTTPPTTT	57
TTPTPTTTTTVTPPTPSTQS	58	PSPTPTTTTTPPTTT	58
PTTPTTTTTVTPPTPSTQS	59	PSPTPTTTTTPPTTT	59
TTPTPTTTTTVTPPTPSTQS	60	PSPTPTTTTTPPTTT	60
PTTPTTTTTVTPPTPSTQS	61	PSPTPTTTTTPPTTT	61
TTPTPTTTTTVTPPTPSTQS	62	PSPTPTTTTTPPTTT	62
PTTPTTTTTVTPPTPSTQS	63	PSPTPTTTTTPPTTT	63
TTPTPTTTTTVTPPTPSTQS	64	PSPTPTTTTTPPTTT	64
PTTPTTTTTVTPPTPSTQS	65	PSPTPTTTTTPPTTT	65
TTPTPTTTTTVTPPTPSTQS	66	PSPTPTTTTTPPTTT	66
PTTPTTTTTVTPPTPSTQS	67	PSPTPTTTTTPPTTT	67
TTPTPTTTTTVTPPTPSTQS	68	PSPTPTTTTTPPTTT	68
PTTPTTTTTVTPPTPSTQS	69	PSPTPTTTTTPPTTT	69
TTPTPTTTTTVTPPTPSTQS	70	PSPTPTTTTTPPTTT	70
PTTPTTTTTVTPPTPSTQS	71	PSPTPTTTTTPPTTT	71
TTPTPTTTTTVTPPTPSTQS	72	PSPTPTTTTTPPTTT	72
PTTPTTTTTVTPPTPSTQS	73	PSPTPTTTTTPPTTT	73
TTPTPTTTTTVTPPTPSTQS	74	PSPTPTTTTTPPTTT	74
PTTPTTTTTVTPPTPSTQS	75	PSPTPTTTTTPPTTT	75
TTPTPTTTTTVTPPTPSTQS	76	PSPTPTTTTTPPTTT	76
PTTPTTTTTVTPPTPSTQS	77	PSPTPTTTTTPPTTT	77
TTPTPTTTTTVTPPTPSTQS	78	PSPTPTTTTTPPTTT	78
PTTPTTTTTVTPPTPSTQS	79	PSPTPTTTTTPPTTT	79
TTPTPTTTTTVTPPTPSTQS	80	PSPTPTTTTTPPTTT	80
PTTPTTTTTVTPPTPSTQS	81	PSPTPTTTTTPPTTT	81
TTPTPTTTTTVTPPTPSTQS	82	PSPTPTTTTTPPTTT	82
PTTPTTTTTVTPPTPSTQS	83	PSPTPTTTTTPPTTT	83
TTPTPTTTTTVTPPTPSTQS	84	PSPTPTTTTTPPTTT	84
PTTPTTTTTVTPPTPSTQS	85	PSPTPTTTTTPPTTT	85
TTPTPTTTTTVTPPTPSTQS	86	PSPTPTTTTTPPTTT	86
PTTPTTTTTVTPPTPSTQS	87	PSPTPTTTTTPPTTT	87
TTPTPTTTTTVTPPTPSTQS	88	PSPTPTTTTTPPTTT	88
PTTPTTTTTVTPPTPSTQS	89	PSPTPTTTTTPPTTT	89
TTPTPTTTTTVTPPTPSTQS	90	PSPTPTTTTTPPTTT	90
PTTPTTTTTVTPPTPSTQS	91	PSPTPTTTTTPPTTT	91
TTPTPTTTTTVTPPTPSTQS	92	PSPTPTTTTTPPTTT	92
PTTPTTTTTVTPPTPSTQS	93	PSPTPTTTTTPPTTT	93
TTPTPTTTTTVTPPTPSTQS	94	PSPTPTTTTTPPTTT	94
PTTPTTTTTVTPPTPSTQS	95	PSPTPTTTTTPPTTT	95
TTPTPTTTTTVTPPTPSTQS	96	PSPTPTTTTTPPTTT	96
PTTPTTTTTVTPPTPSTQS	97	PSPTPTTTTTPPTTT	97
TTPTPTTTTTVTPPTPSTQS	98	PSPTPTTTTTPPTTT	98
PTTPTTTTTVTPPTPSTQS	99	PSPTPTTTTTPPTTT	99
TTPTPTTTTTVTPPTPSTQS	100	PSPTPTTTTTPPTTT	100
PTTPTTTTTVTPPTPSTQS	101	PSPTPTTTTTPPTTT	101
PTTPTTTTTVTPPTPSTQS		PSPTPTTTTTPPTTT	

Figure 3

MUC2 mucin domain of HX1. A) Up-stream part. B) Down-stream part. C) SNP comparison of up-stream part with most complete NCBI entry. D) CNV comparison of down-stream part with most complete NCBI entry.

Figure 4

MUC5AC mucin domain of HX1. A) All domains. B) Main head. C) Main tail. D) Small PTS piece. E) Alignment of LTR1 and LTR2. F) STR1 body. G) STR2 body. H) STR3 body. I) STR4 body. J) Alignment of STR heads. K) Alignment of STR tails. L) Alignment of 6 CysD domains. M) SNP comparison with most complete NCBI entry.

Figure 5

MUC5B mucin domain of HX1. A) All domains. B) Main head. C) Main tail. D) PTS1. E) PTS2. F) PTS3 body. G) PTS4 body. H) PTS5 body. I) PTS6 body. J) PTS7 body. K) Alignment of tails of PTS3, PTS4, PTS5 and PTS6. L) Alignment of CysD2, CysD3, CysD4, CysD5 and CysD6. M) CysD1. N) PTS7 tail. O) SNP comparison with most complete NCBI entry.

Figure 6

MUC6 mucin domain of HX1. A) All domains. B) Head. C) Tail. D) TRs. E) CNV comparison with most complete NCBI entry.

Number of "T"s

Figure 7

DNA sequence of down-stream part TRs of MUC2 mucin domain in HX1. Other than 2 “TT” SNPs at 17th and 96th TR, no continuous “T”s could be found in this part.

Figure 8

Steps to estimate number of repeats in down-stream part of MUC2 mucin domain.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table1.pdf](#)
- [Supplementaryinformation.pdf](#)