Science Advances
AAAS

Supplementary Materials for

A mechanism of gene evolution generating mucin function

Petar Pajic *et al.*

Corresponding author: Omer Gokcumen, omergokc@buffalo.edu; Stefan Ruhl, shruhl@buffalo.edu

Sci. Adv. **8**, eabm8757 (2022)
DOI: 10.1126/sciadv.abm8757

The PDF file includes:

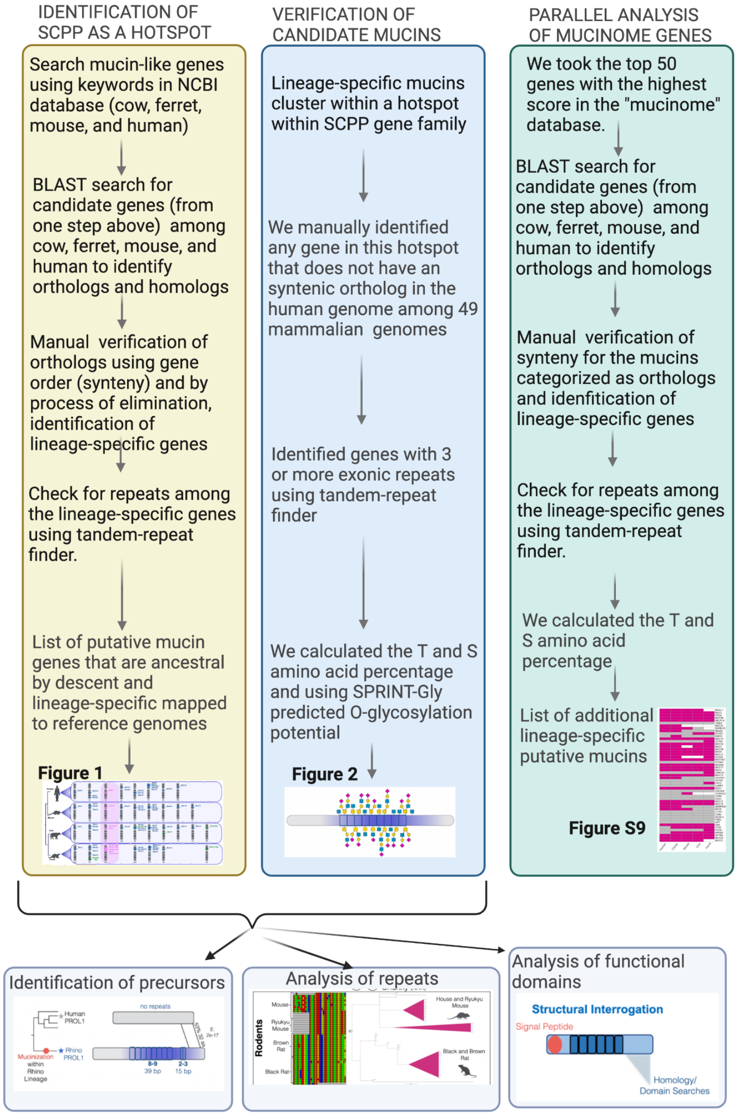
Figs. S1 to S9
Legends for tables S1 and S2
Legend for sequence file S1
References

Other Supplementary Material for this manuscript includes the following:

Tables S1 and S2
Sequence file S1

Supplementary Figures

Figure S1. Bioinformatic analysis pipeline for discovery and analysis of novel mucins. Schematic display of the bioinformatic pipeline we used for mucin identification.



IDENTIFICATION OF SCPP AS A HOTSPOT

Search mucin-like genes using keywords in NCBI database (cow, ferret, mouse, and human)

BLAST search for candidate genes (from one step above) among cow, ferret, mouse, and human to identify orthologs and homologs

Manual verification of orthologs using gene order (synteny) and by process of elimination, identification of lineage-specific genes

Check for repeats among the lineage-specific genes using tandem-repeat finder.

List of putative mucin genes that are ancestral by descent and lineage-specific mapped to reference genomes

Figure 1

VERIFICATION OF CANDIDATE MUCINS

Lineage-specific mucins cluster within a hotspot within SCPP gene family

We manually identified any gene in this hotspot that does not have an syntenic ortholog in the human genome among 49 mammalian genomes

Identified genes with 3 or more exonic repeats using tandem-repeat finder

We calculated the T and S amino acid percentage and using SPRINT-Gly predicted O-glycosylation potential

Figure 2

PARALLEL ANALYSIS OF MUCINOME GENES

We took the top 50 genes with the highest score in the "mucinome" database.

BLAST search for candidate genes (from one step above) among cow, ferret, mouse, and human to identify orthologs and homologs

Manual verification of synteny for the mucins categorized as orthologs and idenfitication of lineage-specific genes

Check for repeats among the lineage-specific genes using tandem-repeat finder.

We calculated the T and S amino acid percentage

List of additional lineage-specific putative mucins

Figure S9

Identification of precursors

Analysis of repeats

Analysis of functional domains

Figure S2: Lineage-specific mucinization events (a) Table showing all discovered lineage-specific mucins arranged based on homology indicating ancestral mucinization events. Colored boxes with numbers denote the number of independent mucinization events found through identity-by-descent during BLAST searches. **(b)** Comparison of gene locations (synteny) in the SCPP locus. Pseudogenes are indicated by dotted grey lines. Lineage-specific mucins are indicated in color. The lineage-specific mucins that show cross-species homology are indicated by connecting lines. The synthetic location of *Muc7* across species is indicated for reference. **(c)** Phylogeny of placental mammals showing potential events of mucinization. The dots in the phylogeny matching in color to **panels a** and **b** indicate estimated emergence of lineage-specific mucins.

a.

1	Sheep MUC2-like	Cow MUC2-like	Camel MUC2-like	Alpaca MUC2-like	Antelope MUC2-like	yak, water buffalo, bison, deer
2	Camel MUC19-like	Ferret MUC5ac-like	Pig MUC5ac-like	Seal MUC19-like	Grizzly and Polar Bear	
3	Seal MUC5ac-like	Ferret MUC5ac-like	Grizzly and Polar Bear			
4	Seal MUC2-like	Walrus MUC17-like	Bears, Dingo, Dog			
5	Ferret MUC5ac-like					
6	Ferret MUC C.1-like	Dog				
7	Cat MUC2	Ferret SMR3A	Dog SMR3A	Puma and Tiger		
8	Mole MUC3a-like					
9-15	ElephantShrew MUC2-like	Elephant MUC5ac-like	Opossum MUC5ac-like	Bat MUC C.1-like		
	Green Monkey C.1-like	Rhino PROL1	Mouse/Rat Muc 10			

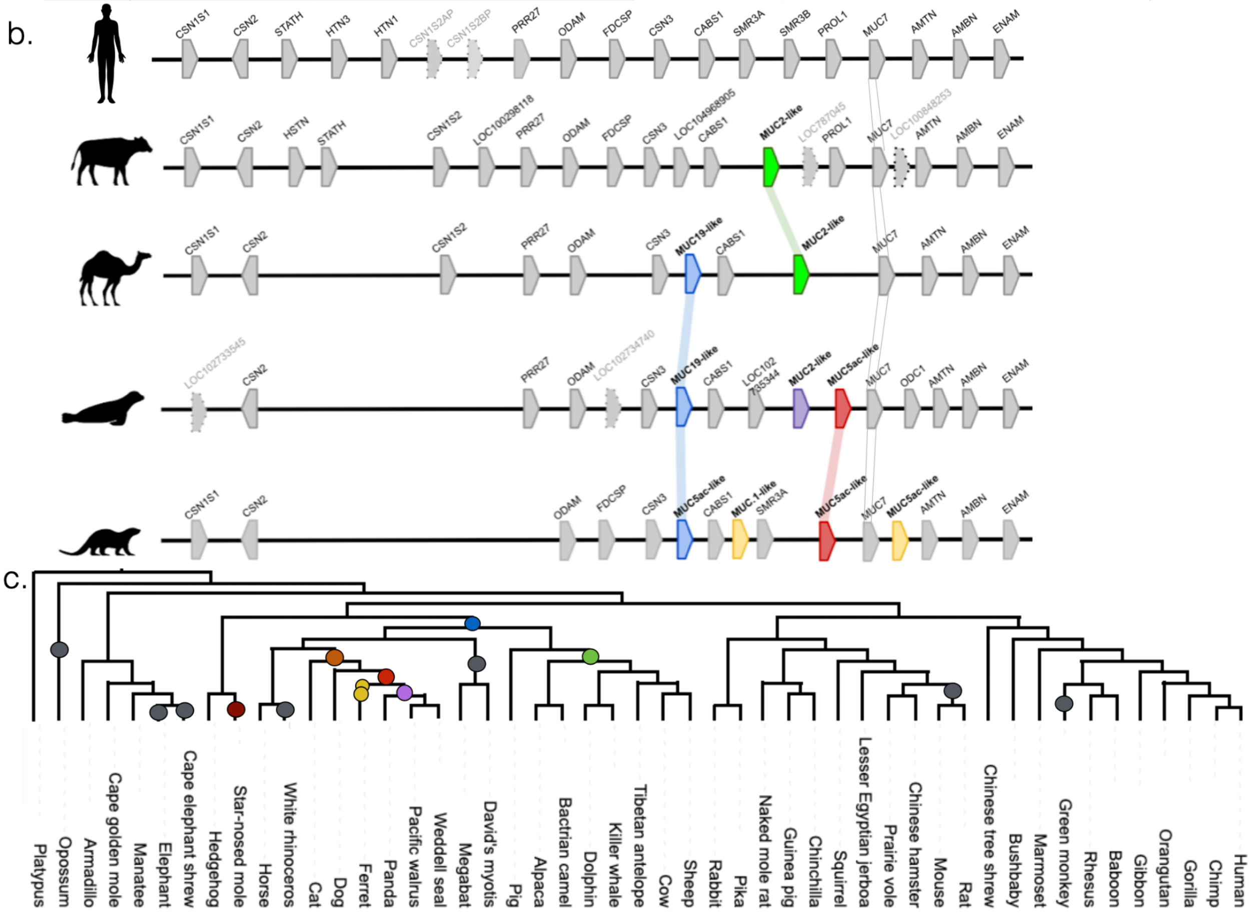


Figure S3: Lineage-specific and orphan mucin structure. Examples of O-glycosylation and repeat content of lineage-specific or orphan mucin genes. The red dots correspond to predicted O-glycosylation sites, with darker shades indicating a higher predicted chance of glycosylation based on the amino acid sequence and surrounding motifs. PTS-rich domains are highlighted with a pink border. The dark grey squares represent repeats. The copy number of the repeats and the length of each individual repeat (period size) are noted above each domain.

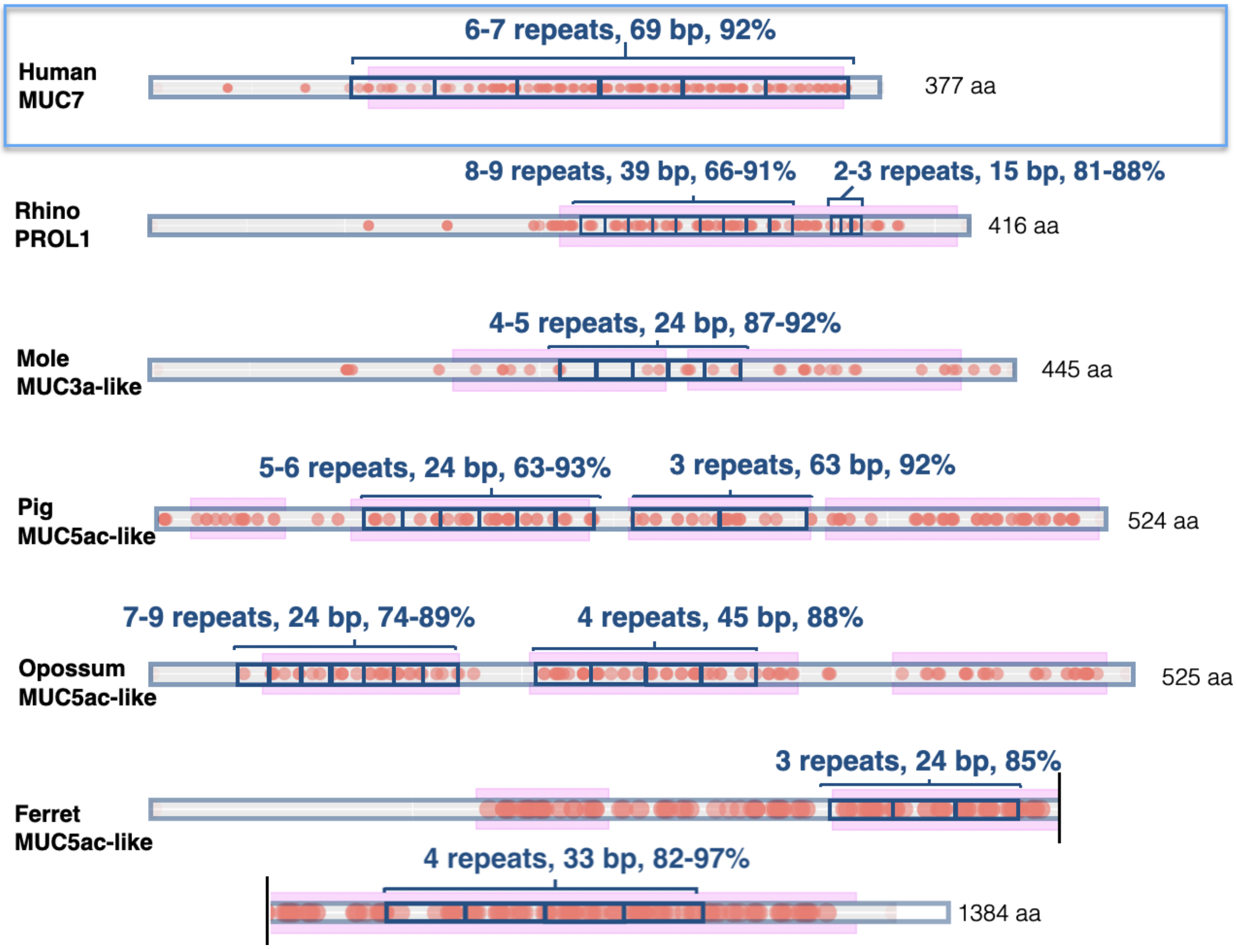


Figure S4: Chinese Pangolin genome harbors two lineage-specific mucin genes that likely originated from precursor genes coding for proteins rich in proline. UCSC Genome Browser screenshot of the Chinese pangolin reference genome zoomed into the SCPP locus with the RefSeq gene annotation using data from other mammalian species. The consolidated versions of gene annotations for the two putative mucin genes (*SMR3A/B* and *OPRPN* based on their homology to human genes) are magnified and shown below the predicted gene annotations. Parts of these two novel mucin genes that do not show homology to sequences in other species (pink ellipses) and harbor PTS-repeats similar to other mucins. We labeled these sequences as “Mucin-like” in the diagram. A study found transcripts matching to *OPRPN* based on RNAseq reads from the Sunda pangolin salivary glands (*75*) (transcripts: *gene22848/new gene_9802)*. These transcripts map back to the Chinese Pangolin reference genome OPRPN mucin-like domain (red circle). The histogram at the bottom right shows an analysis of the abundances of transcripts expressed in Sunda pangolin salivary glands. We found that the expression level of the mucin-like sequences of this gene (corresponding to the transcript *new gene_9802*) is higher than 94% of all transcripts expressed in Sunda pangolin salivary glands.

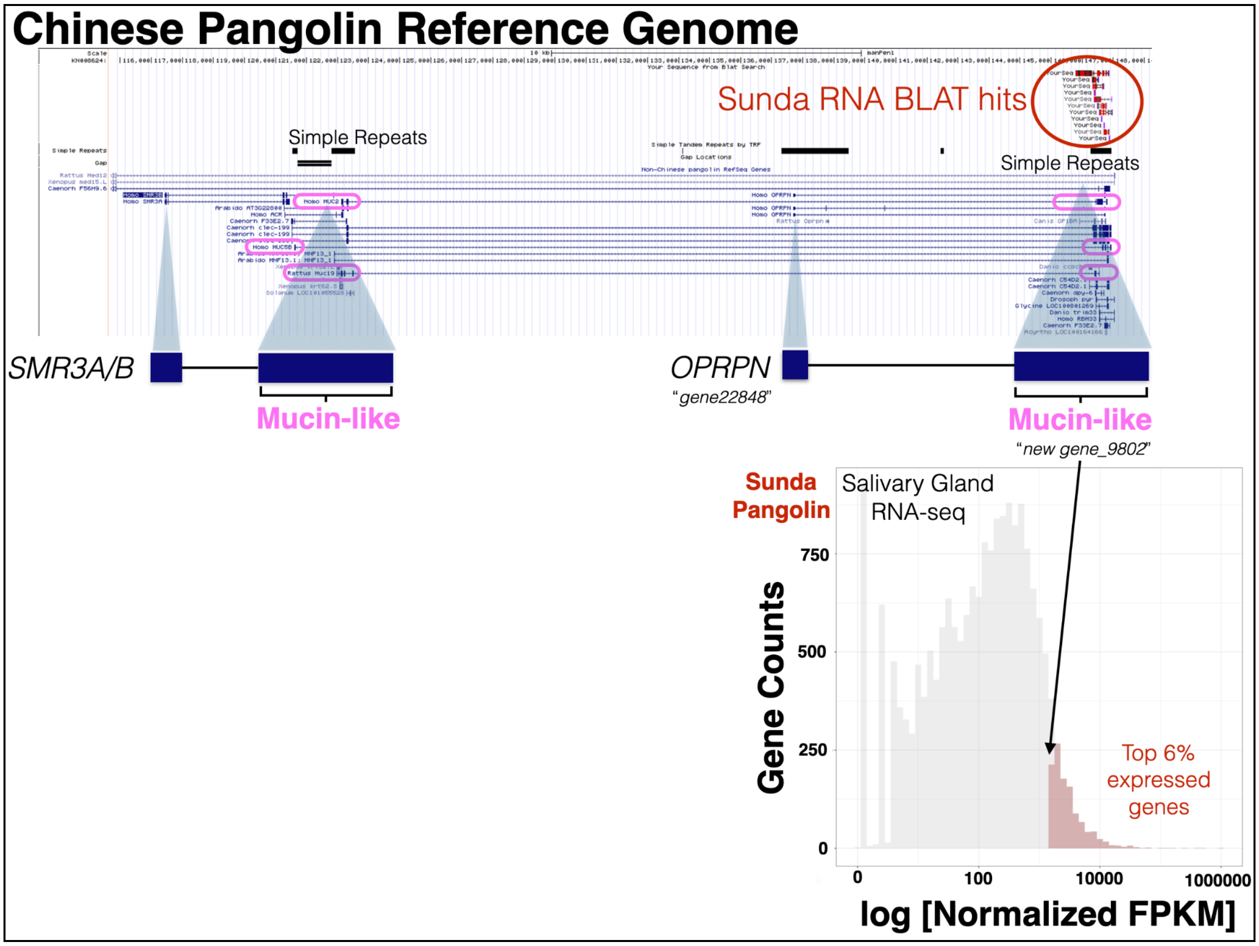


Figure S5: Comparison of lineage-specific mucin exonic repeat period size with other exonic repeats in their respective genomes. The x-axis of the density plots shows the log sequence length of individual repeat units (period size) as documented by Tandem Repeat Finder. The top (pink), middle (tan), and bottom (green) plots show period size distribution of lineage-specific mucins, known mucins in humans, and all exonic repeats in the human genome, respectively. The data show that the period size of mucin exonic repeats falls within the average range of other known exonic repeat units.

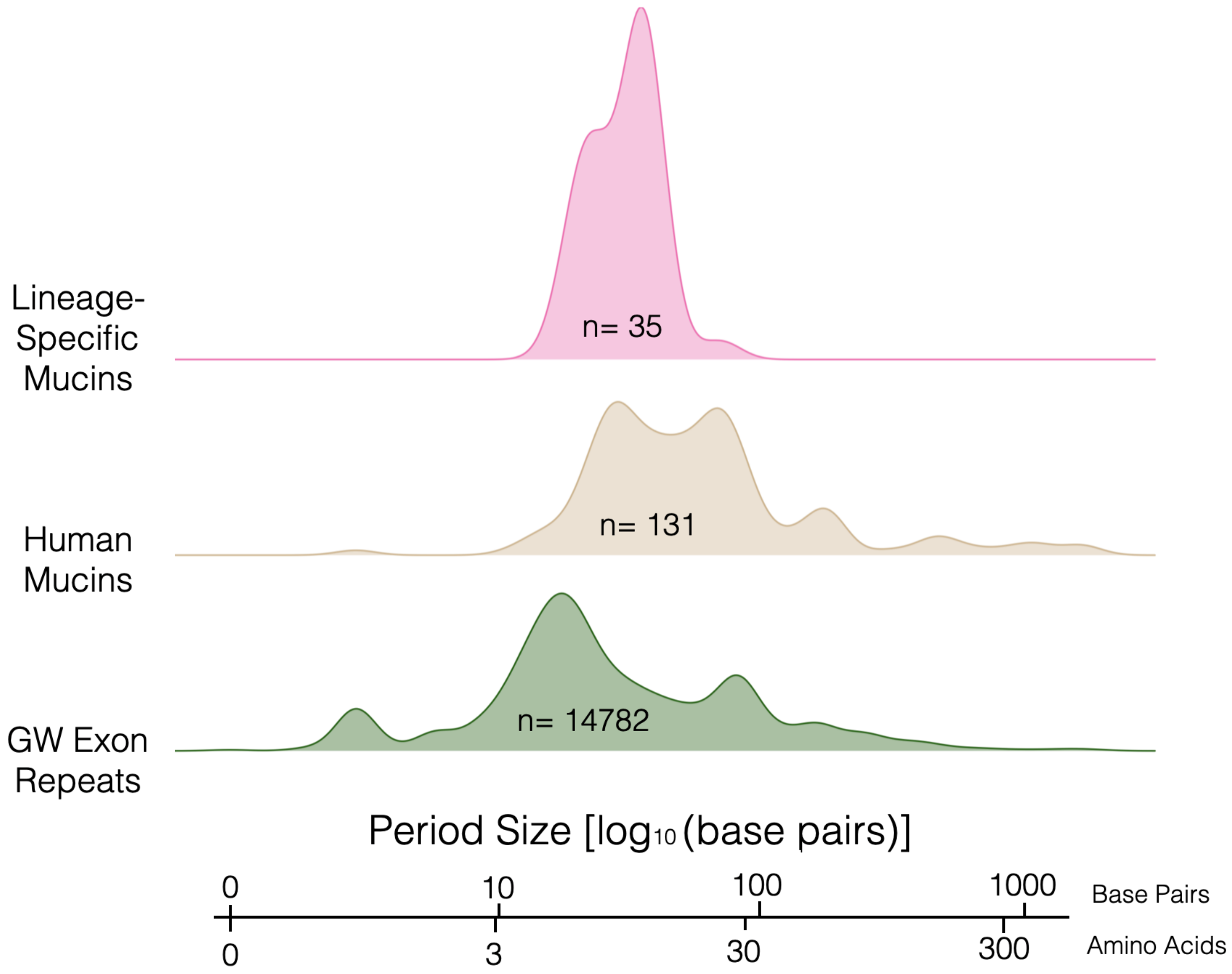


Figure S6: Mechanisms of repeat expansion. As depicted in the schematic model at the top, we considered two potential mechanisms through which exonic repeats could have expanded. One mechanism involves "step-wise" copy number gains and losses of exonic repeats, where only one repeat unit is gained or lost, likely through replication slippage. The other mechanism involves multiple repeat units to be gained or lost in larger "chunks", likely through non-allelic homologous recombination. The bottom panel shows our analysis of sequence diversity of *Muc10* in rodents (upper) and *Muc2-like* in felines (lower). We found evidence for the occurrence of both "step-wise" and "chunky" evolution of exonic repeats.

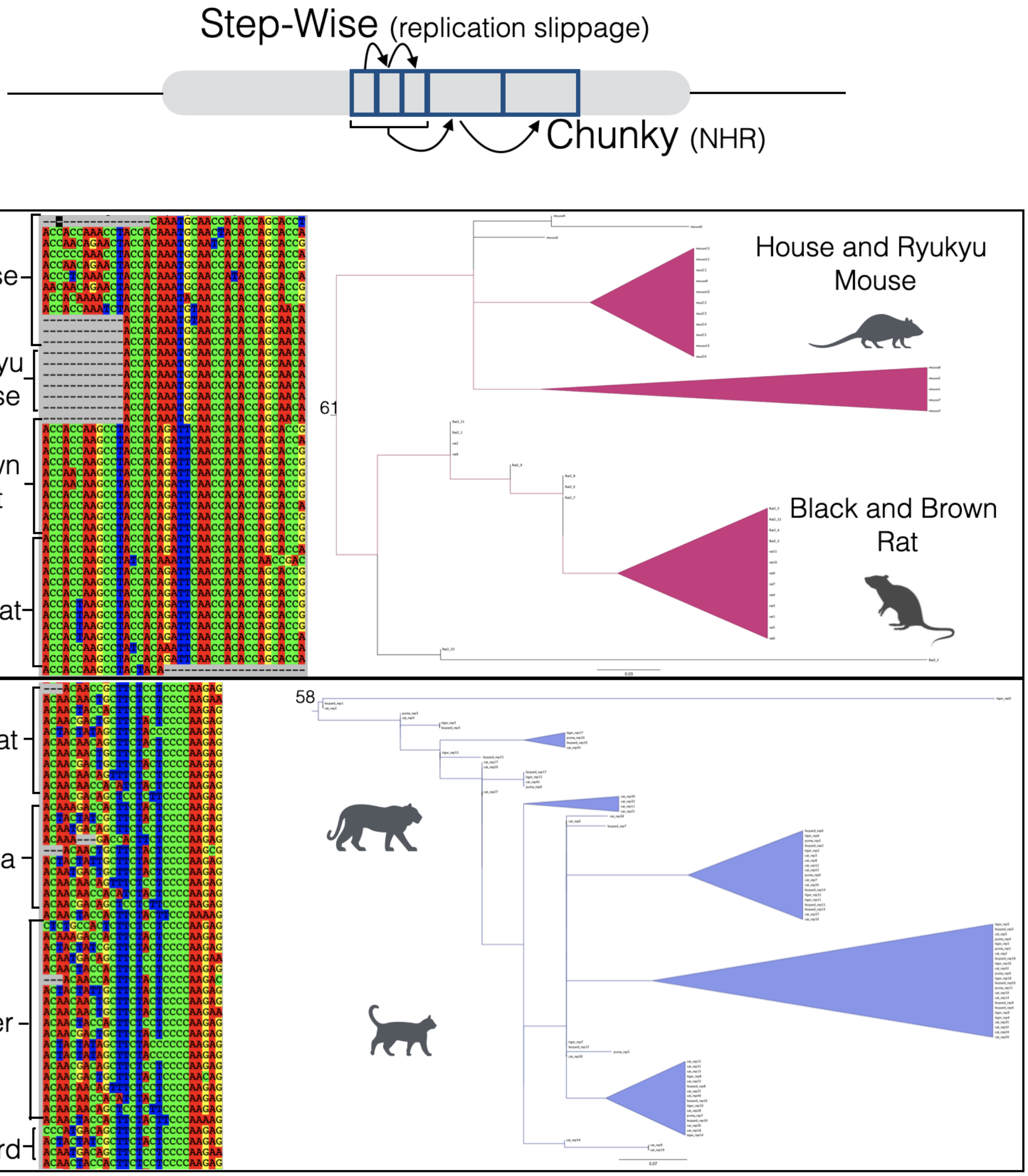


Figure S7: Pairwise comparison of repeats within lineage-specific mucins of *Felinae* and *Rodentia*. We analyzed the sequence differences among individual repeat units both within and across species for *Muc2-like* in *Felinae* (house cat, puma, tiger, and leopard) (red, top panels) and *Muc10 in rodentia* (house mouse, brown rat, black rat, and ryukyu mouse) (blue, bottom panels). Each data point represents a comparison of two different repeat unit sequences. Left panels show the number of nucleotide changes (x-axis) versus the number of amino acid changes (y-axis) as a result of nonsynonymous differences. The right panels show the same x-axis data plotted against the number of threonine and serine (TS) amino acid changes. The regression line and R^2 for each plot is shown.

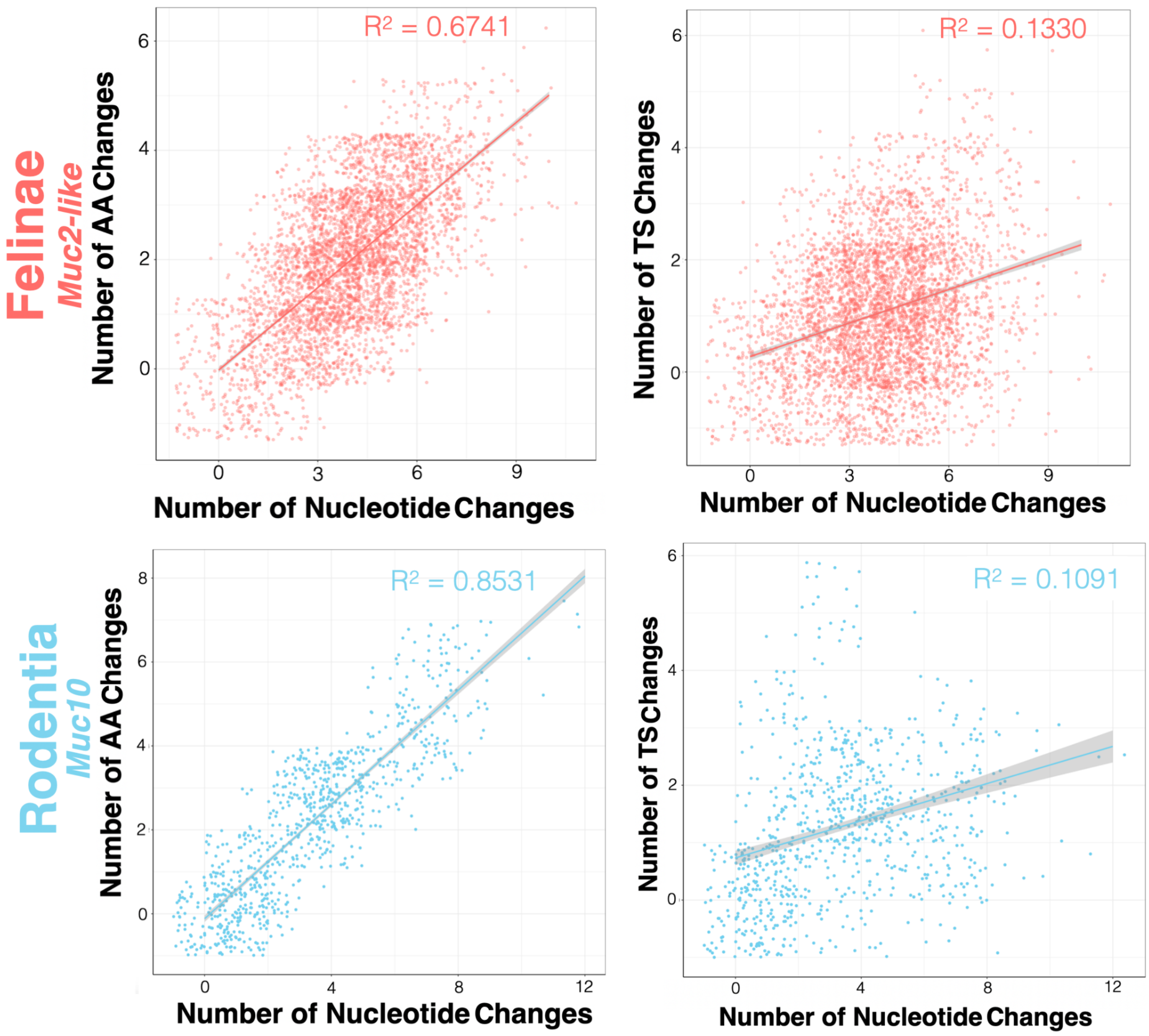


Figure S8: The expression of lineage-specific mucins in species with available transcriptome data. The plot indicates whether individual lineage-specific mucin genes are expressed (red) or not (cream) based on available RNAseq data for multiple tissues (columns) across species (rows). We highlighted three genes (*Muc2-like*, *Muc7*, and *Smr3a*) to exemplify the RNA-seq data mapping to mucin genes. Lineage-specific mucins with a * indicate a shortened version of the gene annotation.

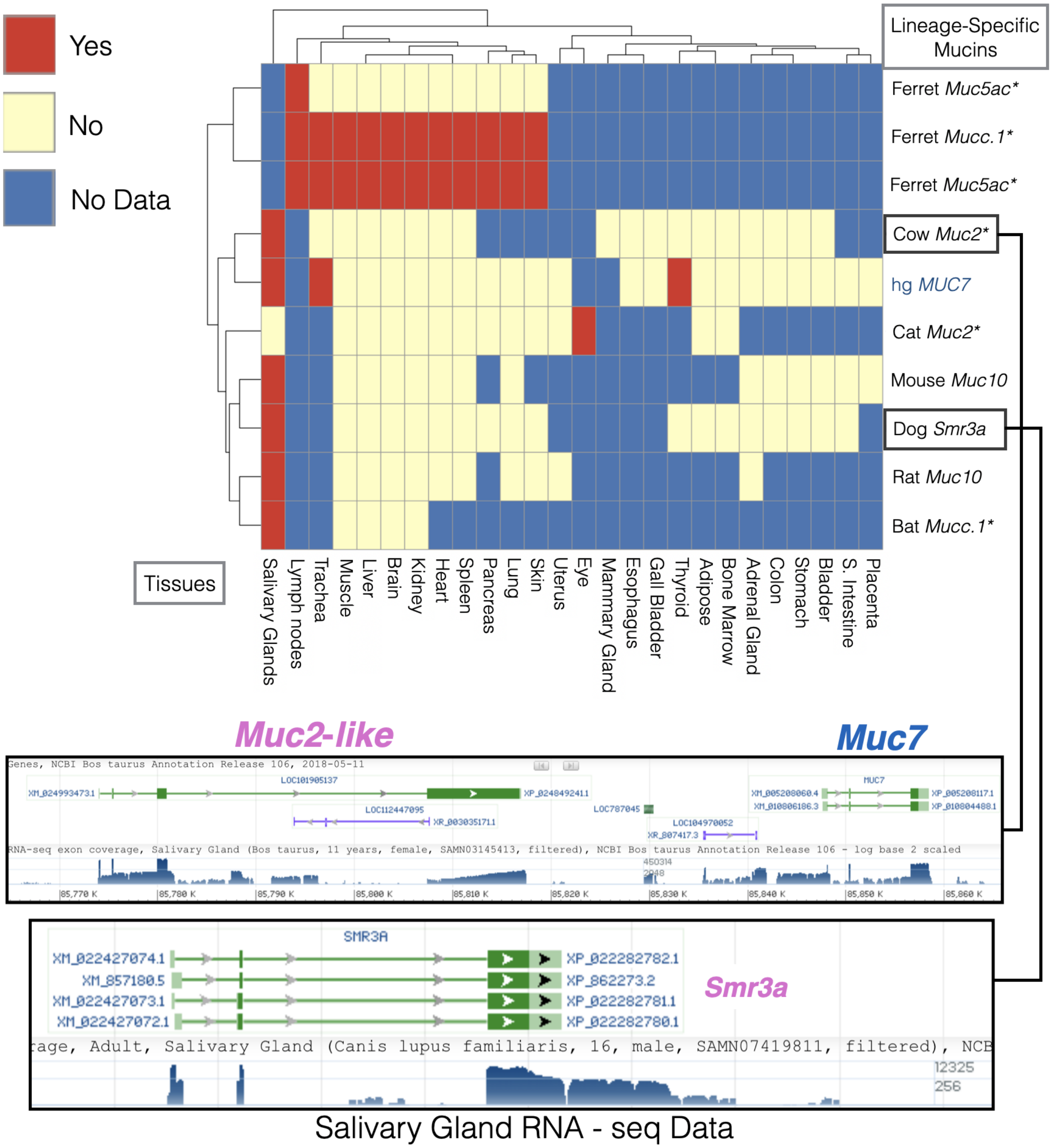
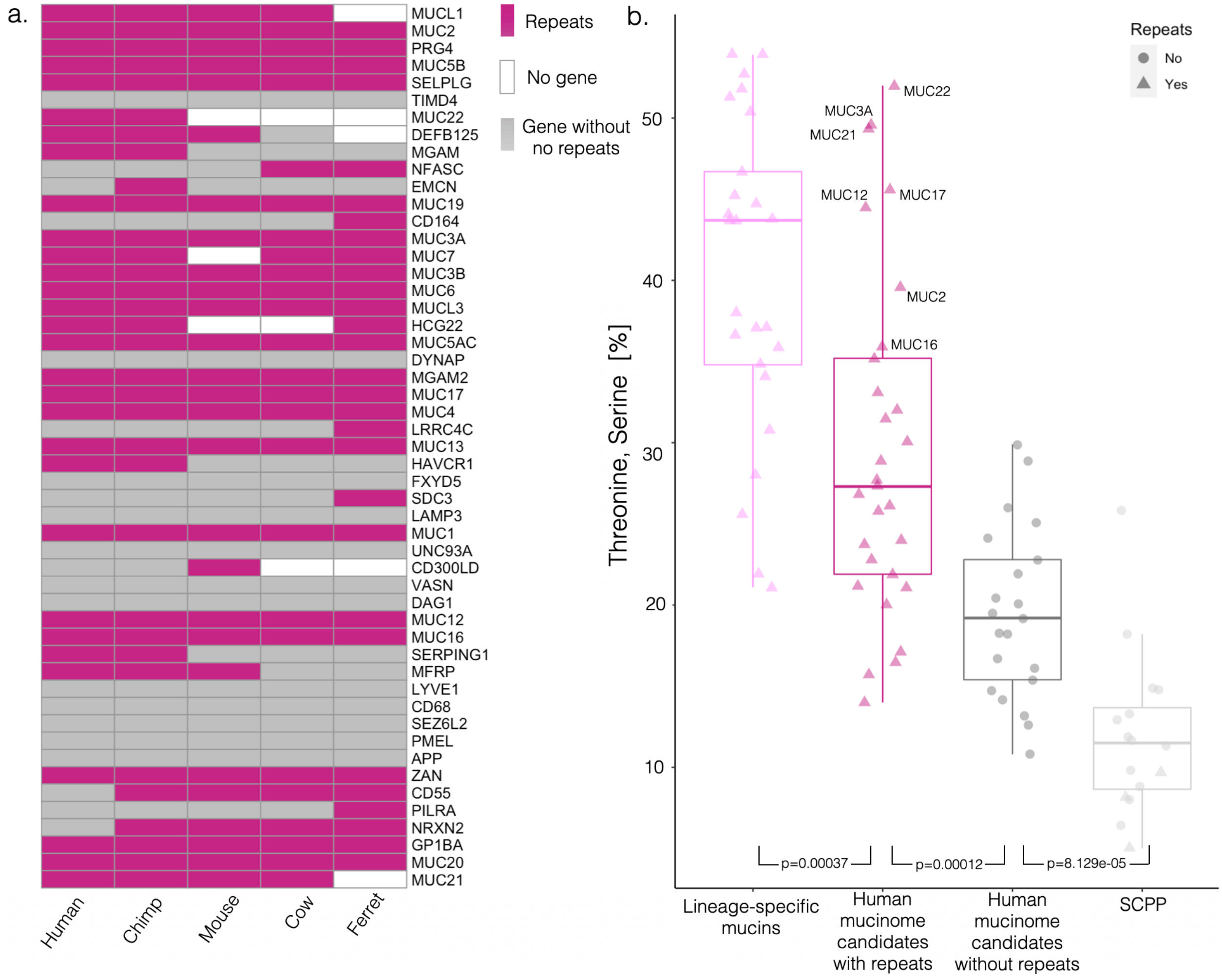


Figure S9. Analysis of other candidate human mucins as defined by the "mucinome". (a) Heatmap of the top 50 mucin candatiate human proteins (plus MUC20 & MUC21) as ranked by the mucin score (defined by (*38*)), across the five mammalian species: human, chimpanzee, mouse, cow, and ferret. Purple squares indicate the presence of the gene in the given genome and that the gene contains a repeat domain (repeat tandems ≥ 3). Grey squares indicate presence of the gene without a repeat domain, and white squares represent absence of the gene in the species' genome. **(b)** Box plot representing the percentage of T and S amino acids within the proteins. Proteins are categorized into lineage-specific mucins (pink; as defined in our cross-species mucin pipeline), human mucinome candidates with repeats (magenta), human mucinome candidates without repeats (dark grey), and SCPP proteins (light grey). Individual proteins of each species are indicated by dots representing a protein without repeats and triangles representing proteins with repeats.



Supplementary Tables

Table S1. Data for Figures. 1,2,3,4 and Supplementary Figures

Table S2. Data for Fig. 5: LC-MS WS and Gel, Saliva Samples

Sequence File S1. Sequences and alignments used in the study.

1. D. L. Stern, The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).

2. T. B. Sackton, N. Clark, Convergent evolution in the genomics era: New insights and directions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190102 (2019).

3. P. Pajic, P. Pavlidis, K. Dean, L. Neznanova, R.-A. Romano, D. Garneau, E. Daugherity, A. Globig, S. Ruhl, O. Gokcumen, Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife* **8**, e44628.

4. M. R. Patel, Y.-M. Loo, S. M. Horner, M. Gale Jr., H. S. Malik, Convergent evolution of escape from hepaciviral antagonism in primates. *PLOS Biol.***10**, e1001282 (2012).

5. F. Denoeud, L. Carretero-Paulet, A. Dereeper, G. Droc, R. Guyot, M. Pietrella, C. Zheng, A. Alberti, F. Anthony, G. Aprea, J.-M. Aury, P. Bento, M. Bernard, S. Bocs, C. Campa, A. Cenci, M.-C. Combes, D. Crouzillat, C. D. Silva, L. Daddiego, F. De Bellis, S. Dussert, O. Garsmeur, T. Gayraud, V. Guignon, K. Jahn, V. Jamilloux, T. Joët, K. Labadie, T. Lan, J. Leclercq, M. Lepelley, T. Leroy, L.-T. Li, P. Librado, L. Lopez, A. Muñoz, B. Noel, A. Pallavicini, G. Perrotta, V. Poncet, D. Pot, Priyono, M. Rigoreau, M. Rouard, J. Rozas, C. Tranchant-Dubreuil, R. Van Buren, Q. Zhang, A. C. Andrade, X. Argout, B. Bertrand, A. de Kochko, G. Graziosi, R. J. Henry, Jayarama, R. Ming, C. Nagai, S. Rounsley, D. Sankoff, G. Giuliano, V. A. Albert, P. Wincker, P. Lashermes, The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**, 1181–1184 (2014).

6. T. D. Kazandjian, D. Petras, S. D. Robinson, J. van Thiel, H. W. Greene, K. Arbuckle, A. Barlow, D. A. Carter, R. M. Wouters, G. Whiteley, S. C. Wagstaff, A. S. Arias, L.-O. Albulescu, A. P. Laing, C. Hall, A. Heap, S. Penrhyn-Lowe, C. V. McCabe, S. Ainsworth, R. R. da Silva, P. C. Dorrestein, M. K. Richardson, J. M. Gutiérrez, J. J. Calvete, R. A. Harrison, I. Vetter, E. A. B. Undheim, W. Wüster, N. R. Casewell, Convergent evolution of pain-inducing defensive venom components in spitting cobras. *Science* **371**, 386–390 (2021).

7. N. R. Casewell, D. Petras, D. C. Card, V. Suranse, A. M. Mychajliw, D. Richards, I. Koludarov, L.-O. Albulescu, J. Slagboom, B.-F. Hempel, N. M. Ngum, R. J. Kennerley, J. L. Brocca, G. Whiteley, R. A. Harrison, F. M. S. Bolton, J. Debono, F. J. Vonk, J. Alföldi, J. Johnson, E. K. Karlsson, K. Lindblad-Toh, I. R. Mellor, R. D. Süssmuth, B. G. Fry, S. Kuruppu, W. C. Hodgson, J. Kool, T. A. Castoe, I. Barnes, K. Sunagar, E. A. B. Undheim, S. T. Turvey, Solenodon genome reveals convergent evolution of venom in eulipotyphlan mammals. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 25745–25755 (2019).

8. J. Dekker, J. W. A. Rossen, H. A. Büller, A. W. C. Einerhand, The MUC family: An obituary. *Trends Biochem. Sci.* **27**, 126–131 (2002).

9. A. R Vahdati, A. Wagner, Parallel or convergent evolution in human population genomic data revealed by genotype networks. *BMC Evol. Biol.***16**, 154 (2016).

10. A. McShane, J. Bath, A. M. Jaramillo, C. Ridley, A. A. Walsh, C. M. Evans, D. J. Thornton, K. Ribbeck, Mucus. *Curr. Biol.* **31**, R938–R945 (2021).

11. S. K. Linden, P. Sutton, N. G. Karlsson, V. Korolik, M. A. McGuckin, Mucins in the mucosal barrier to infection. *Mucosal Immunol.* **1**, 183–197 (2008).

12. J. L. McAuley, S. K. Linden, C. W. Png, R. M. King, H. L. Pennington, S. J. Gendler, T. H. Florin, G. R. Hill, V. Korolik, M. A. McGuckin, MUC1 cell surface mucin is a critical element of the mucosal barrier to infection. *J. Clin. Invest.* **117**, 2313–2324 (2007).

13. J. D. Li, A. F. Dohrman, M. Gallup, S. Miyata, J. R. Gum, Y. S. Kim, J. A. Nadel, A. Prince, C. B. Basbaum, Transcriptional activation of mucin by Pseudomonas aeruginosa lipopolysaccharide in the pathogenesis of cystic fibrosis lung disease. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 967–972 (1997).

14. Y. Ning, H. Zheng, Y. Zhan, S. Liu, Y. Yang, H. Zang, J. Luo, Q. Wen, S. Fan, Comprehensive analysis of the mechanism and treatment significance of Mucins in lung cancer. *J. Exp. Clin. Cancer Res.* **39**, 162 (2020).

15. D. W. Kufe, Mucins in cancer: Function, prognosis and therapy. *Nat. Rev. Cancer* **9**, 874–885 (2009).

16. S. A. Malaker, K. Pedram, M. J. Ferracane, B. A. Bensing, V. Krishnan, C. Pett, J. Yu, E. C. Woods, J. R. Kramer, U. Westerlind, O. Dorigo, C. R. Bertozzi, The mucin-selective protease StcE enables molecular and functional analysis of human cancer-associated mucins. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 7278–7287 (2019).

17. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, 1970).

18. D. Tautz, T. Domazet-Lošo, The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).

19. K. Kawasaki, K. M. Weiss, SCPP gene evolution and the dental mineralization continuum. *J. Dent. Res.* **87**, 520–531 (2008).

20. C. P. Ponting, Biological function in the twilight zone of sequence conservation. *BMC Biol.* **15**, 71 (2017).

21. A. Varki, R. Cummings, J. Esko, H. Freeze, G. Hart, J. Marth, *O-Glycans* (Cold Spring Harbor Laboratory Press, 1999).

22. L. A. Bobek, H. Tsai, A. R. Biesbrock, M. J. Levine, Molecular cloning, sequence, and specificity of expression of the gene encoding the low molecular weight human salivary mucin (MUC7). *J. Biol. Chem.* **268**, 20563–20569 (1993).

23. S. B. Van Oss, A.-R. Carvunis, De novo gene birth. *PLOS Genet.* **15**, e1008160 (2019).

24. A. McLysaght, D. Guerzoni, New genes from non-coding sequence: The role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140332 (2015).

25. C. M. Weisman, The origins and functions of de novo genes: Against all odds? *J. Mol. Evol.* **90**, 244–257 (2022).

26. M. Saitou, E. A. Gaylord, E. Xu, A. J. May, L. Neznanova, S. Nathan, A. Grawe, J. Chang, W. Ryan, S. Ruhl, S. M. Knox, O. Gokcumen, Functional specialization of human salivary glands and origins of proteins intrinsic to human saliva. *Cell Rep.* **33**, 108402 (2020).

27. S. Thamadilok, K.-S. Choi, L. Ruhl, F. Schulte, A. L. Kazim, M. Hardt, O. Gokcumen, S. Ruhl, Human and nonhuman primate lineage-specific footprints in the salivary proteome. *Mol. Biol. Evol.* **37**, 395–405 (2020).

28. D. Xu, P. Pavlidis, S. Thamadilok, E. Redwood, S. Fox, R. Blekhman, S. Ruhl, O. Gokcumen, Recent evolution of the salivary mucin MUC7. *Sci. Rep.* **6**, 31791 (2016).

29. P. C. Denny, L. Mirels, P. A. Denny, Mouse submandibular gland salivary apomucin contains repeated N-glycosylation sites. *Glycobiology* **6**, 43–50 (1996).

30. D. P. Dickinson, M. Thiesse, cDNA cloning of an abundant human lacrimal gland mRNA encoding a novel tear protein. *Curr. Eye Res.* **15**, 377–386 (1996).

31. X. Gao, M. S. Oei, C. E. Ovitt, M. Sincan, J. E. Melvin, Transcriptional profiling reveals gland-specific differential expression in the three major salivary glands of the adult mouse. *Physiol. Genomics* **50**, 263–271 (2018).

32. A. M. Ozyildirim, G. J. Wistow, J. Gao, J. Wang, D. P. Dickinson, H. F. Frierson Jr., G. W. Laurie, The lacrimal gland transcriptome is an unusually rich source of rare and poorly characterized gene transcripts. *Invest. Ophthalmol. Vis. Sci.* **46**, 1572–1580 (2005).

33. *Encyclopedia of Life Sciences* (John Wiley & Sons Ltd., 2001), vol. 26, p. 323.

34. R. Gemayel, M. D. Vinces, M. Legendre, K. J. Verstrepen, Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–477 (2010).

35. E. Schaper, O. Gascuel, M. Anisimova, Deep conservation of human protein tandem repeats within the eukaryotes. *Mol. Biol. Evol.* **31**, 1132–1148 (2014).

36. N. Jonckheere, N. Skrypek, F. Frénois, I. Van Seuningen, Membrane-bound mucin modular domains: From structure to function. *Biochimie* **95**, 1077–1086 (2013).

37. J. Perez-Vilar, R. L. Hill, The structure and assembly of secreted mucins. *J. Biol. Chem.* **274**, 31751–31754 (1999).

38. S. A. Malaker, N. M. Riley, D. J. Shon, K. Pedram, V. Krishnan, O. Dorigo, C. R. Bertozzi, Revealing the human mucinome. *Nat. Commun.* **13**, 3542 (2022).

39. B. W. Cross, S. Ruhl, Glycan recognition at the saliva–oral microbiome interface. *Cell. Immunol.* **333**, 19–33 (2018).

40. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

41. A. J. Hannan, Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018).

42. P. Gagneux, M. Aebi, A. Varki, in *Essentials of Glycobiology*, A. Varki, R. D. Cummings, J. D. Esko, P. Stanley, G. W. Hart, M. Aebi, A. G. Darvill, T. Kinoshita, N. H. Packer, J. H. Prestegard, R. L. Schnaar, P. H. Seeberger, Eds. (Cold Spring Harbor Laboratory Press, 2017).

43. S. H. Church, C. G. Extavour, Null hypotheses for developmental evolution. *Development* **147**, dev178004 (2020).

44. K. M. Wheeler, G. Cárcamo-Oyarce, B. S. Turner, S. Dellos-Nolan, J. Y. Co, S. Lehoux, R. D. Cummings, D. J. Wozniak, K. Ribbeck, Mucin glycans attenuate the virulence of Pseudomonas aeruginosa in infection. *Nat. Microbiol.* **4**, 2146–2154 (2019).

45. S. Pinzón Martín, P. H. Seeberger, D. V. Silva, Mucins and pathogenic mucin-like molecules are immunomodulators during infection and targets for diagnostics and vaccines. *Front. Chem.* **7**, 710 (2019).

46. M. Cohen, A. Varki, Modulation of glycan recognition by clustered saccharide patches. *Int. Rev. Cell Mol. Biol.* **308**, 75–125 (2014).

47. S. Thamadilok, H. Roche-Håkansson, A. P. Håkansson, S. Ruhl, Absence of capsule reveals glycan-mediated binding and recognition of salivary mucin MUC7 by Streptococcus pneumoniae. *Mol. Oral Microbiol.* **31**, 175–188 (2016).

48. K. N. Barnard, B. K. Alford-Lawrence, D. W. Buchholz, B. R. Wasik, J. R. LaClair, H. Yu, R. Honce, S. Ruhl, P. Pajic, E. K. Daugherity, X. Chen, S. L. Schultz-Cherry, H. C. Aguilar, A. Varki, C. R. Parrish, Modified sialic acids on mucus and erythrocytes inhibit influenza a virus hemagglutinin and neuraminidase functions. *J. Virol.* **94**, e01567-19 (2020).

49. L. Deng, B. A. Bensing, S. Thamadilok, H. Yu, K. Lau, X. Chen, S. Ruhl, P. M. Sullam, A. Varki, Oral streptococci utilize a Siglec-like domain of serine-rich repeat adhesins to preferentially target platelet sialoglycans in human blood. *PLOS Pathog.* **10**, e1004540 (2014).

50. P. Gagneux, A. Varki, Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology* **9**, 747–755 (1999).

51. R. Matsuo, Role of saliva in the maintenance of taste sensitivity. *Crit. Rev. Oral Biol. Med.* **11**, 216–229 (2000).

52. H. Y. Çelebioğlu, S. Lee, I. S. Chronakis, Interactions of salivary mucins and saliva with food proteins: A review. *Crit. Rev. Food Sci. Nutr.* **60**, 64–83 (2020).

53. L. E. Tailford, E. H. Crost, D. Kavanaugh, N. Juge, Mucin glycan foraging in the human gut microbiome. *Front. Genet.* **6**, 81 (2015).

54. K. Oliphant, E. Allen-Vercoe, Macronutrient metabolism by the human gut microbiome: Major fermentation by-products and their impact on host health. *Microbiome* **7**, 91 (2019).

55. M. Derrien, M. W. van Passel, J. H. van de Bovenkamp, R. G. Schipper, W. M. de Vos, J. Dekker, Mucin-bacterial interactions in the human oral cavity and digestive tract. *Gut. Microbes* **1**, 254–268 (2010).

56. H. Inoue, K. Ono, W. Masuda, T. Inagaki, M. Yokota, K. Inenaga, Rheological properties of human saliva and salivary mucins. *J. Oral Biosci.* **50**, 134–141 (2008).

57. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

58. S. F. Altschul, J. C. Wootton, E. M. Gertz, R. Agarwala, A. Morgulis, A. A. Schäffer, Y.-K. Yu, Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* **272**, 5101–5109 (2005).

59. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N.-C. Chen, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, A. M. Phillippy, The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

60. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

61. T. Lang, M. Alexandersson, G. C. Hansson, T. Samuelsson, Bioinformatic identification of polymerizing and transmembrane mucins in the puffer fish Fugu rubripes. *Glycobiology* **14**, 521–527 (2004).

62. J. J. A. Armenteros, K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, H. Nielsen, SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).

63. S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto, R. D. Finn, The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

64. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

65. A. Garg, G. P. S. Raghava, A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biol.* **8**, 129–140 (2008).

66. L. Zhao, G. Poschmann, D. Waldera-Lupa, N. Rafiee, M. Kollmann, K. Stühler, OutCyte: A novel tool for predicting unconventional protein secretion. *Sci. Rep.* **9**, 19448 (2019).

67. G. Taherzadeh, A. Dehzangi, M. Golchin, Y. Zhou, M. P. Campbell, SPRINT-Gly: Predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. *Bioinformatics* **35**, 4140–4146 (2019).

68. C. Steentoft, S. Y. Vakhrushev, H. J. Joshi, Y. Kong, M. B. Vester-Christensen, K. T.-B. G Schjoldager, K. Lavrsen, S. Dabelsteen, N. B. Pedersen, L. Marcos-Silva, R. Gupta, E. P. Bennett, U. Mandel, S. Brunak, H. H. Wandall, S. B. Levery, H. Clausen, Precision

mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488 (2013).

69. 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 Res.* **22**, 4673–4680 (1994).

70. G. Stecher, K. Tamura, S. Kumar, Molecular evolutionary genetics analysis (MEGA) for macOS. *Mol. Biol. Evol.* **37**, 1237–1239 (2020).

71. S. Shen, B. An, X. Wang, S. P. Hilchey, J. Li, J. Cao, Y. Tian, C. Hu, L. Jin, A. Ng, C. Tu, M. Qu, M. S. Zand, J. Qu, Surfactant cocktail-aided extraction/precipitation/on-pellet digestion strategy enables efficient and reproducible sample preparation for large-scale quantitative proteomics. *Anal. Chem.* **90**, 10350–10359 (2018).

72. X. Shen, S. Shen, J. Li, Q. Hu, L. Nie, C. Tu, X. Wang, B. Orsburn, J. Wang, J. Qu, An ionstar experimental strategy for MS1 ion current-based quantification using ultrahigh-field orbitrap: Reproducible, in-depth, and accurate protein measurement in large cohorts. *J. Proteome Res.* **16**, 2445–2456 (2017).

73. G. C. McAlister, D. P. Nusinow, M. P. Jedrychowski, M. Wühr, E. L. Huttlin, B. K. Erickson, R. Rad, W. Haas, S. P. Gygi, MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86**, 7150–7158 (2014).

74. Y. Perez-Riverol, J. Bai, C. Bandla, D. García-Seisdedos, S. Hewapathirana, S. Kamatchinathan, D. J. Kundu, A. Prakash, A. Frericks-Zipper, M. Eisenacher, M. Walzer, S. Wang, A. Brazma, J. A. Vizcaíno, The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).

75. J.-E. Ma, L.-M. Li, H.-Y. Jiang, X.-J. Zhang, J. Li, G.-Y. Li, L.-H. Yuan, J. Wu, J.-P. Chen, Transcriptomic analysis identifies genes and pathways related to myrmecophagy in the Malayan pangolin (Manis javanica). *PeerJ* **5**, e4140 (2017).