


Fall 1-1-2017

Primate proteomic composition of seminal plasma and prostate-specific transglutaminase activity in relation to sexual selection.

Amanda M.C. Zielen
Duquesne University

Follow this and additional works at: <https://dsc.duq.edu/etd>

 Part of the [Biochemistry Commons](#), [Biology Commons](#), [Evolution Commons](#), and the [Molecular Biology Commons](#)

Recommended Citation

Zielen, A. M. (2017). Primate proteomic composition of seminal plasma and prostate-specific transglutaminase activity in relation to sexual selection. (Doctoral dissertation, Duquesne University). Retrieved from <https://dsc.duq.edu/etd/212>

This One-year Embargo is brought to you for free and open access by Duquesne Scholarship Collection. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Duquesne Scholarship Collection. For more information, please contact phillips@duq.edu.

PRIMATE PROTEOMIC COMPOSITION OF SEMINAL PLASMA AND PROSTATE-
SPECIFIC TRANSGLUTAMINASE ACTIVITY IN RELATION TO SEXUAL SELECTION.

A Dissertation

Submitted to the Bayer School of Natural and Environmental Sciences

Duquesne University

In partial fulfillment of the requirements for
the degree of Doctor of Philosophy

By

Amanda M. Colvin Zielen

December 2017

Copyright by

Amanda M. Colvin Zielen

2017

PRIMATE PROTEOMIC COMPOSITION OF SEMINAL PLASMA AND PROSTATE-SPECIFIC TRANSGLUTAMINASE ACTIVITY IN RELATION TO SEXUAL SELECTION.

By

Amanda M. Colvin Zielen

Approved June 15, 2017

Dr. Michael I. Jensen-Seaman
Associate Professor of Biology
(Committee Chair)

Dr. Brady Porter
Associate Professor of Biology
(Committee Member)

Dr. Jan Janecka
Assistant Professor of Biology
(Committee Member)

Dr. Michael Cascio
Associate Professor of Biochemistry and
Chemistry
(Committee Member)

Dr. Philip Reeder
Dean, Bayer School of Natural and
Environmental Sciences

Dr. Joseph McCormick
Chair, Department of Biological Sciences
Associate Professor of Biology

ABSTRACT

PRIMATE PROTEOMIC COMPOSITION OF SEMINAL PLASMA AND PROSTATE-SPECIFIC TRANSGLUTAMINASE ACTIVITY IN RELATION TO SEXUAL SELECTION.

By

Amanda M. Colvin Zielen

December 2017

Dissertation supervised by Dr. Michael I. Jensen-Seaman

Humans (*Homo sapiens*), chimpanzees (*Pan troglodytes*), and gorillas (*Gorilla gorilla*) have diverse mating systems with varying levels of sperm competition. Several seminal plasma genes have been claimed to evolve under positive selection, while others are altered or lost. This study aims to identify biologically relevant differences among seminal plasma proteomes of primates in relation to mating systems and previous genomic studies. Seminal plasma from three individuals of each species were run in triplicate in shotgun liquid chromatography – tandem mass spectrometry (LC-MS/MS) and confirmed with Western blots. Over 7,000 peptides were identified across all individuals; 168 proteins were identified with high confidence, 70 seminal plasma proteins were identified for human, 64 proteins for chimpanzee, and 34 proteins for gorilla. The gorilla seminal plasma proteome has higher variation among individuals and many proteins involved in semen coagulation and liquefaction have been lost. Chimpanzees have

approximately 7-fold higher prostate specific transglutaminase (TGM4) expression than humans. TGM4 was not detected in gorillas, supporting pseudogenization of this gene. The structural semenogelin proteins, SEMG1 and SEMG2, were detected in high abundance in only one of three gorilla individuals, and in all three human and chimpanzee individuals. Chimpanzees have significantly higher expression of SEMG1 (~2.5-fold) compared to human; whereas, they only produce a small amount of SEMG2; ~6.5 –fold less than humans. Chimpanzees have roughly 34-fold higher expression of a serine protease inhibitor, SERPINA3 (Serpin Family A Member 3), than humans. SERPINA3 paralogs, SERPINA1 and SERPINA5, also have increased expression (~2.5 –fold) compared to human, and only SERPINA1 was detected in gorilla. SERPINAs may delay protease dissolution of the copulatory plug in chimpanzees. Recombinant human TGM4 and the reconstructed ancestral TGM4 sequence of our last common ancestor (LCA) with chimpanzees (the human-chimpanzee ancestor) proteins were produced and incubated with casein and monodansylcaverdine to determine enzymatic activity. The human-chimpanzee ancestor TGM4 had higher activity compared to human TGM4. Considering the importance of TGM4 in semen coagulation and copulatory plug formation in chimpanzee, the increased activity of the human-chimpanzee ancestor TGM4 may be indicative of elevated female promiscuity of our LCA, perhaps similar to a chimp-like mating system.

DEDICATION

To my husband, family, and friends,
thank you for providing immense support,
comradery, and mostly for believing in me,
even when I did not believe in myself.

To my research mentor, Dr. Michael I. Jensen-Seaman,
for choosing me for this research project, for challenging me,
and helping me develop as an independent scientist.

To the little girl [me],
who loved science, learning, and the outdoors.
Who was brave and determined to push beyond expectations,
and who was strong enough to overcome obstacles and failures.

To my undergraduate research advisor,
for telling me that earning a Ph.D is beyond intelligence,
that it requires perseverance, humility, and determination.

Lastly, to all the fortune cookies,
with inspiring words and pearls of wisdom,
which provided me hope that my dreams would come true.

ACKNOWLEDGEMENTS

Dr. Michael I. Jensen-Seaman's guidance was pivotal throughout this research project and for the development of my research skills and independent experimental design. I am immensely appreciative of the mentorship he provided me that was both supportive and facilitated independence, which is a hard balance to achieve. During my tenure in his lab, I have become a better writer, researcher, and thinker; I attribute this progression to both his effort for challenging me to improve and my desire to learn.

I would like to recognize my past and present committee members, Dr. Brady Porter, Dr. Jan Janecka, Dr. Michael Cascio, and Dr. Partha Basu for their useful discussions and advice during my committee meetings. Moreover, I am appreciative of their willingness for collaboration and discussion throughout the year. I was unafraid to request to use any of their resources or equipment. I hope to return the support in years to come.

The Bayer School of Natural Environmental Sciences is fortunate to have accommodating and knowledgeable faculty, staff, and graduate students. Throughout my tenure at Duquesne University, I could always find someone to help me with any given issue. I am appreciative of my senior lab members, Dr. Sarah Carnahan-Craig and Dr. Scott Hergenrother, for their exceptional mentorship; especially, when I first joined the lab. Overall the Jensen-Seaman lab had a great group of graduate and undergraduate students, but I would like to specifically thank Jennifer (Vill) Doyle, Lindsay Loughner, Megan Hockman, and Dr. Ranajit Das for their collaboration, support, and comradery over the years. Dr. Tara Allison, Dr. Kate Sadler, Sumedha Sethi, Jackie Shane, and Joe Sallmen were invaluable fellow graduate students

in discussing science and troubleshooting experiments. I sincerely appreciate their knowledge, time, and friendship.

I would like to thank specific people for aiding in experimental development and/or acquisition of data discussed in this dissertation and future publications. Dr. John Thomas processed all the LC/MS-MS injections and initial peptide analysis through spectrum mills software. John and his advisor, Dr. Partha Basu, were immensely helpful in experimental design and data analysis. Dr. Lauren O'Donnell was exceptionally helpful in aiding in our transition of detecting Western blots with Li-COR infrared imaging. Dr. Allyson O'Donnell and Dr. Michael Cascio provided advice and equipment for protein production, purification, and enzymatic assays.

TABLE OF CONTENTS

ABSTRACT	iv
DEDICATION	vi
ACKNOWLEDGEMENTS	vii
LIST OF TABLES	xiv
LIST OF FIGURES	xvi
CHAPTER 1: Introduction	1
1.1 Hominid evolution and physiology	1
1.1.1 Hominid phylogeny	1
1.1.2 Hominid physiology	2
1.1.2.1 Gorilla species, biogeography, and life history	2
1.1.2.2 Chimpanzee species, biogeography, and life history	4
1.1.2.3 Human evolution and life history	5
1.2 Evolution of reproductive proteins	6
1.2.1 Prediction of selection	6
1.2.2 Evolution of proteins with activity	8
1.2.3 Eukaryotic regulation and its role in protein evolution.....	10
1.3 Sexual selection	12
1.3.1 Precopulatory strategies	13
1.3.1.1 Intersexual precopulatory strategies	13
1.3.1.2 Intrasexual precopulatory strategies	14
1.3.2 Postcopulatory strategies	15
1.3.2.1 Intersexual postcopulatory strategies	15
1.3.2.2 Intrasexual postcopulatory strategies	16
1.4 Primate sperm competition, physiology, and seminal plasma proteins	20
1.4.1 Primate mating systems, sperm competition, and reproductive physiology	20
1.4.2 Seminal plasma proteins.....	23
1.4.2.1 Semenogelin proteins	24
1.4.2.1.1 Structure of SEMG1 and SEMG2 genes	24
1.4.2.1.2 Expression and function of Semenogelin proteins	27
1.4.2.2 Prostate specific transglutaminase.....	28
1.4.2.3 Kallikrein related peptidase.....	29
1.4.2.4 Prostatic acid phosphatase ACPP.....	30
1.5 Experimental goals	32
1.6 References	33

CHAPTER 2: Proteomic composition of seminal plasma proteins in gorilla, human, and chimpanzee	43
2.1 Introduction	43
2.1.1 Primate mating systems and sperm competition	43
2.1.2 Shotgun proteomics and seminal plasma proteins	44
2.2 Methods	46
2.2.1 Seminal plasma sample preparation	46
2.2.1.1 Sample preparation	46
2.2.1.2 Protein quantification	47
2.2.2 SDS-PAGE optimization.....	47
2.2.3 Western blot optimization	48
2.2.3.1 Abnova primary antibodies for highly abundant seminal proteins	48
2.2.3.2 Dot blot to confirm secondary antibody effectiveness	50
2.2.3.3 Purified actin dilution series	51
2.2.3.4 Optimization of secondary antibodies	51
2.2.3.5 Membrane transfer optimization	52
2.2.3.6 Anti-pilin detection.....	54
2.2.4 Quality control checks.....	55
2.2.4.1 Loading control Western blot.....	55
2.2.4.2 Protease inhibitor assay	55
2.2.5 Methods for comparative proteomics among human, chimpanzee, and gorilla seminal plasma	56
2.2.5.1 Shotgun Liquid Chromatography Tandem Mass Spectrometry LC-MS/MS.....	56
2.2.5.1.1 LC/MS-MS sample preparation	56
2.2.5.1.2 LC/MS-MS conditions	56
2.2.5.1.3 LC/MS-MS spectra data processing	58
2.2.5.1.4 Manual mapping of SEMG peptides	59
2.2.5.1.5 Protein quantification and statistical analysis.....	60
2.2.5.2 SDS PAGE and Western blot methods	61
2.3 Results.....	63
2.3.1 Protein quantification	63
2.3.2 SDS PAGE optimization	65
2.3.3 Western blot optimization	66
2.3.3.1 Bio-Rad wet transfer system provides band clarity in Western blots compared to iBlot™ transfer system.....	66

2.3.3.2 Non-specific bands are due to cross-linking and degradation of seminal proteins and not due to our Western blot protocol	67
2.3.3.3 Alkaline phosphatase conjugated secondary antibody and FastRed colorimetric detection reduces signal to noise ratio compared to chemiluminescent detection	69
2.3.4 Western blot quality control checks	70
2.3.4.1 Increased abundance of heterodimeric clusterin in gorilla seminal plasma prevents it from being an effective extracellular loading control across hominid seminal plasma	70
2.3.4.2 Majority of proteolytic degradation of proteins, particularly the semenogelins, occurs before we receive semen samples	71
2.3.5 LC/MS-MS and Western blot identification of seminal plasma proteins among human, chimpanzee, and gorilla individuals	74
2.3.5.1 SEMGs, if expressed, are in relatively high abundance across hominid species with differences among SEMG1 and SEMG2 expression	74
2.3.5.2 Nine proteins are shared among human, chimpanzee, and gorilla seminal plasma, with a vast majority of proteins being species specific contributing to distinct proteomes.	76
2.3.5.3 Biodiversity indexes indicate moderate protein richness and evenness in human, chimpanzee, and gorilla seminal plasma without significant differences between species	78
2.3.5.4 Chimpanzee seminal plasma has increased expression of proteins involved in semen coagulation and prevention of liquefaction	79
2.3.5.5 Gorilla seminal plasma has loss of proteins involved in semen coagulation and liquefaction pathways	80
2.4 Discussion	91
2.4.1 Semenogelins	92
2.4.2 Up-regulation and pseudogenization of seminal coagulation pathway	93
2.5 Future Directions	98
2.6 References	100
CHAPTER 3: Prostate specific transglutaminase TGM4 activity	103
3.1 Introduction	103
3.2 Methods	106
3.2.1 Cloning of Human SEMGs and TGM4 into pFastBac1 vector	106
3.2.1.1 Traditional and TOPO® cloning	106
3.2.1.2 Site-directed mutagenesis of human TGM4	108
3.2.2 Gibson assembly of human and chimpanzee KLK3s and chimpanzee and human-chimpanzee ancestor TGM4s into the pFastBac1 vector	110
3.2.2.1 Designing of gBlocks	110
3.2.2.2 Gibson assembly	113

3.2.2.3 Site-directed mutagenesis of chimpanzee and ancestor TGM4s in pFastBac1 vector	115
3.2.3 Cloning of chimpanzee SEMGs into pFastBac1	117
3.2.4 Site directed mutagenesis to produce chimpanzee SEMG2-TYR mutant and human- chimpanzee ancestor KLK3 in pFastBac1 vector	120
3.2.5 Transfer of pFastBac1 clones to pCMV mammalian expression vectors	121
3.2.5.1 Chimpanzee SEMG1	121
3.2.5.2 KLK3s, TGM4s, and SEMGs subcloning procedure into pCMV vector	123
3.2.6 Production/Purification of recombinant proteins	126
3.2.6.1 Gene expression in insect cell culture	126
3.2.6.2 LNCaP cell culture	126
3.2.6.3 293T mammalian cell culture	128
3.2.6.4 Protein quantification	129
3.2.6.5 Acetone precipitation	129
3.2.6.6 His purification and dialysis	130
3.2.6.7 Antibody detection of recombinant proteins	130
3.2.7 TGM4 assays	131
3.2.7.1 1D gel based assay	131
3.2.7.2 96 well plate reader assay	132
3.3 Results	133
3.3.1 Generation of recombinant plasmids with human, chimpanzee, and human-chimpanzee ancestral mRNA sequences	133
3.3.2 Expression of recombinant proteins in mammalian cell culture	134
3.3.3 Transglutaminase assay	135
3.3.3.1 Optimization of the 1D gel assay	135
3.3.3.2 Human SEMG1 Ia peptides are cross-linked by transglutaminases	136
3.3.3.3 Recombinant TGM4 proteins lose function after HIS purification	137
3.3.4 Comparison of TGM4 enzymatic activity of human and the human-chimp ancestor	138
3.4 Discussion	142
3.5 Future Directions	144
3.6 References	146
APPENDIX	148
A.1 Script in R for protein quantification and protease inhibitor analysis	148
A.1.1 Quantification method mini-comparison	148
A.1.2 Seminal plasma concentrations	152
A.1.3 Statistical comparison of seminal concentration across species	154
A.1.4 Protease Inhibitor Bradford Assay	155

A.2 Protein sequence alignments.....	158
A.3 Outline of LC/MS-MS data sorting, labeling, and normalization.....	161
A.3.1 Annotation and Normalization:	161
A.4 Script in R for PLGEM statistics	164
A.4.1 Load libraries and set working directory.....	164
A.4.2 Length adjusted pairwise PLGEM of human and chimp	165
A.4.3 Length adjusted pairwise PLGEM of human and gorilla.....	167
A.4.4 Length adjusted pairwise PLGEM of chimp and gorilla.....	169
A.4.5 Raw normalization pairwise PLGEM of human and chimp	172
A.4.6 Raw normalization pairwise PLGEM of human and gorilla	174
A.4.7 Raw normalization pairwise PLGEM of chimp and gorilla.....	176
A.5 Shannon and Simpson diversity index calculations.....	179
A.5.1 Script in R for comparing biodiversity index statistics between raw and length adjusted data	190
A.6 Complete tables of LC/MS-MS high confidence identified proteins.....	191
A.7 Gene Ontology.....	215
A.8 Sequencing and PCR primers.....	223
A.9 Human-chimpanzee TGM4 ancestor sequence reconstruction.....	225
A.10 Human, chimpanzee, and human-chimpanzee ancestor sequences in pCMV vector	226

LIST OF TABLES

TABLE 1.1: TYPES OF SEXUAL SELECTION AND THEIR ANATOMICAL AND PHYSIOLOGICAL ADAPTATIONS	12
TABLE 1.2: RATES OF SEMG2 EVOLUTION COMPARED TO SPECIES LEVEL OF POLYANDRY*	27
TABLE 2.1: PARAMETERS FOR COLLISION ENERGY CALCULATION.	58
TABLE 2.2: SEMG SEQUENCES USED FOR PEPTIDE MAPPING	59
TABLE 2.3: EQUATIONS USED TO CALCULATE SHANNON AND SIMPSON BIODIVERSITY INDEXES.....	61
TABLE 2.4: PRIMARY ANTIBODIES AND THEIR DILUTIONS	62
TABLE 2.5: KNOWN CONCENTRATIONS ARE LOWER IN BRADFORD AND HIGHER IN QBIT ASSAYS THAN EXPECTED	64
TABLE 2.6: HOMINID SPECIES SEMINAL PLASMA CONCENTRATIONS	64
TABLE 2.7: NORMALIZED RAW SPECTRAL ABUNDANCES OF SEMGS ARE RELATIVELY HIGH BUT VARIABLE ACROSS HOMINIDS.....	75
TABLE 2.8: DIVERSITY INDEXES INDICATE RELATIVELY HIGH PROTEIN DIVERSITY AND MODERATELY EVEN DISTRIBUTION OF SEMINAL PLASMA PROTEOMES FOR HOMINIDS.	79
TABLE 2.9: RELATIVE ABUNDANCE OF PROTEINS DETECTED IN HOMINID SEMINAL PLASMA	82
TABLE 3.1: MODIFIED PRIMERS USED FOR HUMAN GENE CONSTRUCT CLONING.....	107
TABLE 3.2: MUTAGENIC PRIMERS USED FOR HUMAN TGM4 MODIFICATION	110
TABLE 3.3: GBLOCK PRIMERS OF KLK3 AND TGM4 USED FOR GIBSON ASSEMBLY.....	113
TABLE 3.4: MUTAGENIC PRIMERS USED FOR CHIMPANZEE AND ANCESTOR TGM4 MODIFICATION	117
TABLE 3.5: PRIMERS USED FOR CHIMPANZEE SEMG1 AND SEMG2 AMPLIFICATION.....	118
TABLE 3.6: MUTAGENIC PRIMERS USED FOR CHIMPANZEE AND ANCESTOR TGM4 MODIFICATION	121
TABLE 3.7: UNIVERSAL PRIMERS USED FOR pFASTBAC1 TGM4 AMPLIFICATION	124
TABLE 3.8: PRIMERS USED FOR VIRAL PARTICLE AMPLIFICATION	126
TABLE 3.9: SPECTRAMAX® I3X PARAMETERS FOR TGM ASSAYS	132
TABLE 3.10: SUMMARY OF RECOMBINANT PLASMIDS	133
TABLE A.1: SHANNON DIVERSITY CALCULATION WITH RAW NORMALIZATION.....	179
TABLE A.2: SIMPSON DIVERSITY CALCULATION WITH RAW NORMALIZATION	185
TABLE A.3: ALL PROTEINS IDENTIFIED WITH HIGH CONFIDENCE IN EACH HUMAN INDIVIDUAL	191
TABLE A.4: ALL PROTEINS IDENTIFIED WITH HIGH CONFIDENCE IN EACH CHIMPANZEE INDIVIDUAL.....	195

TABLE A.5: ALL PROTEINS IDENTIFIED WITH HIGH CONFIDENCE IN EACH GORILLA INDIVIDUAL	199
TABLE A.6: RAW NORMALIZED ABUNDANCES OF ALL PROTEINS IDENTIFIED IN EACH HUMAN INDIVIDUAL (IN TRIPPLICATE) ARRANGED BY PREVALENCE.....	201
TABLE A.7: RAW NORMALIZED ABUNDANCES OF ALL PROTEINS IDENTIFIED IN EACH CHIMPANZEE INDIVIDUAL (IN TRIPPLICATE) ARRANGED BY PREVALENCE.....	204
TABLE A.8: RAW NORMALIZED ABUNDANCES OF ALL PROTEINS IDENTIFIED IN EACH GORILLA INDIVIDUAL (IN TRIPPLICATE) ARRANGED BY PREVALENCE.....	207
TABLE A.9: SUMMARY OF SIGNIFICANT DIFFERENCES BETWEEN SPECIES	209
TABLE A.10: SIGNIFICANT OVEREXPRESSION OF GO MOLECULAR FUNCTIONS IN HUMAN SEMINAL PLASMA COMPARED TO THE HUMAN REFERENCE GENOME	216
TABLE A.11: SIGNIFICANT OVEREXPRESSION OF GO BIOLOGICAL PROCESSES IN HUMAN SEMINAL PLASMA COMPARED TO THE HUMAN REFERENCE GENOME	216
TABLE A.12: SIGNIFICANT OVEREXPRESSION OF GO MOLECULAR FUNCTIONS IN CHIMPANZEE SEMINAL PLASMA COMPARED TO THE HUMAN AND CHIMPANZEE REFERENCE GENOMES	218
TABLE A.13: SIGNIFICANT OVEREXPRESSION OF GO BIOLOGICAL PROCESSES IN CHIMPANZEE SEMINAL PLASMA COMPARED TO THE HUMAN AND CHIMPANZEE REFERENCE GENOMES	218
TABLE A.13: SIGNIFICANT OVEREXPRESSION OF GO BIOLOGICAL PROCESSES IN GORILLA SEMINAL PLASMA COMPARED TO THE HUMAN AND CHIMPANZEE REFERENCE GENOMES	221
TABLE A.14: VECTOR AND GENE SPECIFIC SEQUENCING PRIMERS	223
TABLE A.15: CONTROL PCR PRIMERS FOR GENOMIC DNA	224

LIST OF FIGURES

FIGURE 1.1: HOMINID PHYLOGENIC TREE.....	1
FIGURE 1.2: PRIMATE SPERM COMPETITION PHENOTYPES	18
FIGURE 1.3: MORPHOLOGICAL PENILE DIFFERENCES AMONG MULTI-PARTNER AND SINGLE-PARTNER MATING SYSTEMS	19
FIGURE 1.4: SUMMARY OF HOMINID MATING SYSTEMS, SPERM COMPETITION, AND PHYSIOLOGY	22
FIGURE 1.5: SEMGS GENOMIC AND PROTEIN STRUCTURE	26
FIGURE 2.1: BRADFORD ASSAY WITH SEMINAL PLASMA SAMPLE ABSORBENCIES	65
FIGURE 2.2: COMMERCIAL NUPAGE GELS YIELD BETTER CLARITY THAN POURED STACKING GELS	66
FIGURE 2.3: BIO-RAD WET TRANSFER SYSTEM HAS A CLEARER SIGNAL TO BACKGROUND WESTERN BLOT RATIO COMPARED TO iBLOT™ DRY TRANSFER.....	67
FIGURE 2.4: NONGLYCOSYLATED PILIN WAS CLEANLY DETECTED FROM HOSPITAL SAMPLES	68
FIGURE 2.5: COLORIMETRIC FASTRED DETECTION HAS REDUCED BACKGROUND NOISE COMPARED TO CHEMILUMINESCENCE.....	69
FIGURE 2.6: CLUSTERIN IS NOT A RELIABLE EXTRACELLULAR LOADING CONTROL.....	71
FIGURE 2.7: NO SIGNIFICANT DIFFERENCES IN PROTEIN CONCENTRATION IN SEMINAL PLASMA WITH AND WITHOUT PROTEASE INHIBITORS.....	72
FIGURE 2.8: SEMINAL PLASMA SAMPLES WITH AND WITHOUT PROTEASE INHIBITORS LOOK SIMILAR.....	73
FIGURE 2.9: SEMG1 AND SEMG2 ARE DEGRADED BEFORE ADDITION OF PROTEASE INHIBITORS	73
FIGURE 2.10: MANUAL PEPTIDE MAPPING OF SEMGS	75
FIGURE 2.11: WESTERN BLOT SUPPORTS LC/MS-MS PEPTIDE MAPPING OF SEMG2 IN SEMINAL PLASMA OF CHIMPANZEE INDIVIDUALS AND ONE GORILLA INDIVIDUAL.....	76
FIGURE 2.12: DISTINCT SPECIES-SPECIFIC BANDING PATTERN OF HOMINID SEMINAL PLASMA.....	77
FIGURE 2.13: PAIRWISE PLGEM GRAPHS WITH RAW NORMALIZATION	89
FIGURE 2.14: CHIMPANZEE SEMINAL PLASMA HAS AN EXCESS OF CRTAC1	90
FIGURE 2.15: WESTERN BLOTS CONFIRM LC/MS-MS RESULTS.....	90
FIGURE 2.16: EXPRESSION DIFFERENCES IN SEMINAL COAGULATION AND LIQUEFACTION PATHWAYS OF HOMINIDS: HUMAN, CHIMPANZEE, AND GORILLA.....	97

FIGURE 3.1: PROTEINS INVOLVED IN SEMEN COAGULATION AND LIQUEFACTION UPON EJACULATION	105
FIGURE 3.2: HUMAN AND CHIMPANZEE KLK3 GBLOCK DESIGN	111
FIGURE 3.3: HUMAN-CHIMPANZEE ANCESTOR AND CHIMPANZEE TGM4 GBLOCK DESIGN	112
FIGURE 3.4: DIFFERENTIAL TIMING OF SECRETION OF RECOMBINANT PROTEINS.....	134
FIGURE 3.5: TRANSGLUTAMINASE ACTIVITY DETECTED AS LOW AS 0.1MG OF GP TGM ENZYME.....	135
FIGURE 3.6: GUINEA PIG TRANSGLUTAMINASE HAS ACTIVITY ON hSEMG1.IA PEPTIDES AND TRANSFECTED hTGM4 MEDIA HAS ACTIVITY WITH CASEIN.....	136
FIGURE 3.7: RECOMBINANT HUMAN AND ANCESTOR TGM4 HAVE ENZYMATIC ACTIVITY BEFORE BUT NOT AFTER HIS PURIFICATION.....	137
FIGURE 3.8: RECOMBINANT HUMAN TGM4 IS PURIFIED FROM MEDIA USING HIS COLUMNS.....	138
FIGURE 3.9: HUMAN AND ANCESTOR TGM4 LYSATES ARE MORE EFFICIENT THAN MEDIA	139
FIGURE 3.10: HUMAN-CHIMPANZEE ANCESTOR TGM4 HAS SIGNIFICANTLY HIGHER ENZYMATIC ACTIVITY COMPARED TO HUMAN TGM4.....	140
FIGURE 3.11: RECOMBINANT HUMAN-CHIMPANZEE ANCESTOR TGM4 HAS MORE ACTIVITY THAN RECOMBINANT HUMAN TGM4.....	141
FIGURE A.1: BIOLOGICAL PROCESS FUNCTIONS OF ALL IDENTIFIED PROTEINS IN HUMAN, CHIMPANZEE, AND GORILLA SEMINAL PLASMA.....	215
FIGURE A.2: BIOLOGICAL PROCESS FUNCTIONS OF ALL IDENTIFIED PROTEINS IN EACH SPECIES: HUMAN, CHIMPANZEE, AND GORILLA SEMINAL PLASMA	222

CHAPTER 1: Introduction

1.1 Hominid evolution and physiology

1.1.1 Hominid phylogeny

Hominoids, or apes, are divided into two families: hylobatidae (gibbons) and hominidae (hominids) and diverged from old world monkeys approximately 20 million years ago (mya). The hominids, or great apes, consist of human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*), western and eastern gorilla (*Gorilla gorilla* and *Gorilla beringei*, respectively), and orangutan (*Pongo pygmaeus*) species (Figure 1.1). The major divergences within the hominid clade arose approximately 15 million years ago (orangutan), 7 Mya (gorilla), and 6 Mya (human-chimpanzee split) (Chen and Li, 2001). The taxonomic ranks assigned to some hominoid groups and their divergence times remain somewhat contentious, while the branching order is nearly universally accepted.

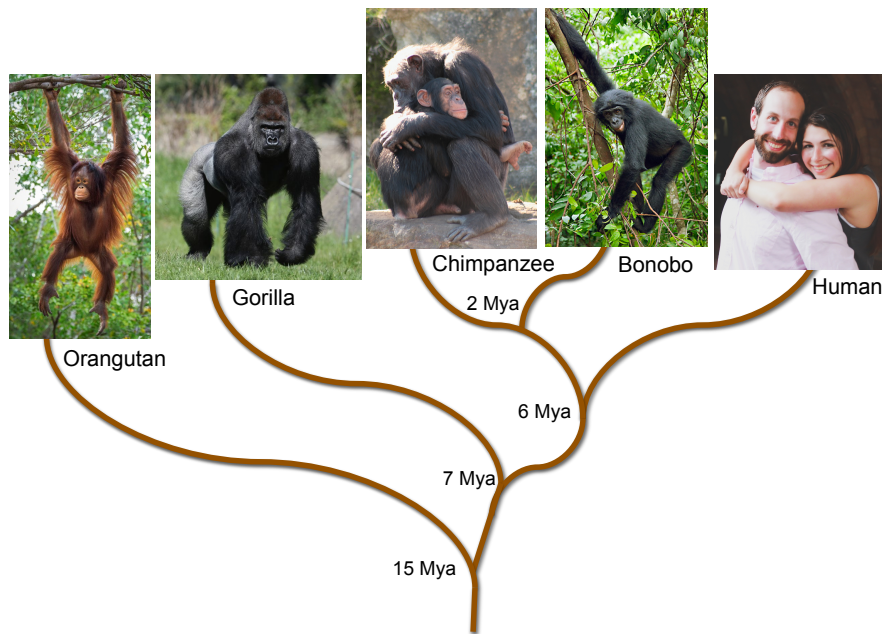


Figure 1.1: Hominid phylogenetic tree

Orangutan diverged approximately 15 million years ago (mya) from the other hominids. Gorilla separated from the human-chimpanzee clade approximately 7 Mya, and the human / chimpanzee split occurred 6 Mya. Chimpanzee and Bonobo were separated geographically by the Congo River approximately 2 Mya, which led to speciation. Note: All photos are personal or available for free usage.

1.1.2 Hominid physiology

The extant primates have multiple universal morphological characters, which distinguish them from other mammals. Perhaps the most known, is that primates have opposable thumbs and big toes, which allow them to grasp and hold objects with ease. Furthermore on all digits, the primates' nails are flattened as opposed to claws seen in other mammals. The olfactory region in primates has been greatly reduced in most species, which allows for the expansion of the cerebrum. The areas of the brain that correlate with hand-eye coordination and stereoscopic vision (forward facing eyes) are enhanced compared to other mammals. Primates also have slower reproduction rates than mammals of similar size, which is a result of having an extended infant/juvenile stage in offspring. Typically primate species only have one to three offspring per pregnancy, and hominids usually only have one (Fleagle, 2013). My research mainly focuses on the hominids: human, chimpanzee, and gorilla species; therefore, they will be discussed in more detail.

1.1.2.1 Gorilla species, biogeography, and life history

There are two species of gorilla, the western gorilla (*Gorilla gorilla*) and eastern gorilla (*Gorilla beringei*), existing in two distinct geographical ranges north and east of the Congo River basin, separated by about 1000 kilometers. Western gorillas are found throughout west Central Africa, as far north as southern Nigeria. Eastern gorillas are found east of the Congo River basin to the extreme north and southwestern edges of Rwanda and Uganda (Mitchell and Gonder, 2013). The western gorilla comprises two subspecies: the western lowland gorilla (*G. g. gorilla*) and the Cross River gorilla (*G. g. dielhi*). The eastern gorilla comprises two subspecies: the eastern lowland gorilla (*G. b. graueri*) and the mountain gorilla (*G. b. beringei*) (Das et al., 2014).

Gorilla is the largest living primate, with females weighing up to 98 kg and males up to 181 kg in the wild (Miller-Schroeder, 1997). In general, gorillas have black skin with thick dark brown or black hair covering their body, except for their face, hands, feet, and the male chest (Rowe, 1996). Dominant males, or silverbacks, have a pronounced sagittal crest and distinct gray, or “silver”, patch of hair on their backs and posterior end (Groves, 2004). Gorillas walk on their finger knuckles of the two digits closest to the thumb. This leaves the hand open, allowing them to carry objects as they walk on all fours (Fleagle, 2013).

Around six years of age, females experience menarche, or their first menstrual cycle. However, female gorillas remain infertile for two years post menarche. Female gorillas first parturition, or their first delivery, is around ten years of age (Czekala and Robbins, 2001). Gorilla infant stage is up to three years, and then they begin their juvenile stage (3-6 years of age), which shows a decrease in maternal grooming, weaning, and no longer sleeping in a nest with the mother (Stewart, 2001). Because of their long period of maternal dependence, gorilla mothers can expect to invest years caring for their altricial offspring, which coincides with their interbirth span of approximately 4 years (Czekala and Robbins, 2001). Although male gorillas do not take an active role in caring for infants, they are essential for their socialization, often associating with older infants and juveniles, and fighting off aggressive adult males (Stewart, 2001). It is difficult to assess male gorilla reproductive maturity because they can be fertile before exhibiting secondary physical sexual characteristics (like gray hair). Males between the ages of 8 and 12 are called blackbacks (Robbins et al., 2004). Usually by age 12 or 13, males can be considered silverbacks, but most will not reach their full adult size until the age of 15 (Czekala and Robbins, 2001).

1.1.2.2 Chimpanzee species, biogeography, and life history

Chimpanzees live north of the Congo River, and their congeneric species, bonobo, lives to the south. There are currently three geographically defined chimpanzee subspecies: *P.t. troglodytes* from central Africa, *P.t. schweinfurthii* from east Africa, and *P. t. verus* from west Africa (Ely et al., 2005). Although a fourth subspecies, *P.t. vellerosus*, located in northern Cameroon and Nigeria was proposed, it has not become widely accepted (Gonder et al., 1997).

Chimpanzees are black or dark brown in skin and hair color, have beards, and can become bald. They are born with a pale face and hands, which darken with age (Rowe, 1996). Chimpanzees have a less pronounced snout and sagittal crest, and are smaller, as a male chimpanzee's average weight is 34-70 kg and females are typically between 26-50 kg in the wild (Nowak, 1999). Chimpanzees are quadrupedal knuckle walkers, and walk with open hands, like gorilla (Fleagle, 2013). Chimpanzees have limited bipedalism, and are also able to be arboreal (Doran, 1996).

Female chimpanzees experience menarche between 8-11 years of age, although they typically have their first offspring around 13 years of age, which coincides with female natal dispersal. The interbirth interval is about four to six years, with a life span of about 40 years in the wild (Atsalis and Videan, 2009) In chimpanzees, hierarchy is mediated by male dominance, where the more dominant males copulate more than the less dominant males (Goodall, 1986; Whiten et al., 1999). Chimpanzee females may covertly, forced or willingly, leave the group with another male and/or mate with males from neighboring groups (Vigilant et al., 2001). Though these behaviors increase the female's chance of reproductive success, they also increase the chance of male mediated infanticide of their offspring (Nishida and Kawanaka, 1985).

1.1.2.3 Human evolution and life history

Although the only extant hominin is modern humans (*Homo sapiens*), there are multiple extinct hominins that are on our lineage after the human-chimpanzee split (~6 Mya). The earliest hominin fossils were found in eastern and southern Africa from the Pliocene, and possibly as early as the late Miocene (Robson and Wood, 2008). *Ardipithecus* fossils have been found in eastern Africa and have been dated 4.2 - 5.8 Mya. *Australopithecus* fossils were found in eastern and southern Africa from the time period of 1.9 - 4.2 Mya. *Paranthropus* most likely lived between 1.2 – 2.7 Mya, and their fossils were mostly found in eastern and southern Africa. The genus *Homo* dates back to 2.3 Mya, and its species have a wide distribution of fossils found in Africa, Europe, Asia, and Indonesia (Potts, 2013). Genetic evidence supports an African origin of modern human (*Homo sapiens*), as well as at least two instances of admixture with other species of archaic humans (*Homo neanderthalensis* and *Homo sapiens ssp. denisova*) once modern humans had radiated out of Africa (Green et al., 2010; Reich et al., 2010).

Humans are unique from other apes with long legs, upright bodies, and large brains. Hair color in human populations ranges between shades of black, red, brown, and white; skin color ranges between multiple shades of brown. Men often have beards, but beard density and coverage varies greatly. Although hair covers the majority of the human body, most of it is shorter, fragile, less pigmented, and less dense than in other primates. This has led to the humans being described as the “naked ape”, or simply hairless (Newman, 1970; Pagel and Bodmer, 2003). The human brain is bigger with an enlarged cranium, which lacks well-defined brow ridges and crests. In addition, human canine tooth size is greatly reduced, our premolars are wider, and often the third molar is absent or reduced in size (Fleagle, 2013). Human weight

varies, but on average humans are 42 to 78 kg. Men, on average, are 1.2 times heavier than women, and 1.06 times taller (Fleagle, 2013; Dixson, 2009)

Human females experience menarche between 9 and 13 years of age, and unlike other primates, will go through reproductive senescence, commonly known as menopause, around 50 years of age. Females will then continue to live several decades as a post-reproductive individual (Kuhle, 2007). Human male fertility develops during the same time frame as females. Though different human societies have describable natal dispersal, there is no real gender-specific trend, with humans of both genders dispersing as well as maintaining familial relationships throughout life (Fleagle, 2013).

1.2 Evolution of reproductive proteins

1.2.1 Prediction of selection

Genes involved in reproduction have shown positive selection and rapid divergence among diverse animal groups as a result of sexual selection (Koisol et al., 2008; Ramm et al., 2008; 2009). As selection has different effects on synonymous (dS) and nonsynonymous (dN) substitution rates, estimation of these rates is used to understand the selective forces (purifying, neutral, or positive) of sequences coding for proteins.

Synonymous and nonsynonymous nucleotide changes occur in a protein sequence. The encoded amino acid remains the same in synonymous base substitutions, while the amino acid is changed in nonsynonymous substitutions. The dN/dS value is calculated as the ratio of the rate of nonsynonymous substitutions divided by the rate of synonymous substitutions (dN/dS). A dN/dS value equal to 1 is consistent with the gene experiencing neutral evolution because the rate of nonsynonymous change is equal to the rate of synonymous change. If the dN/dS value is less than 1, then the gene is likely experiencing purifying selection. The rate of nonsynonymous

change is less than rate of synonymous change because the gene's function is important, and nonsynonymous changes tend to be deleterious and selected against. A dN/dS value greater than 1 is indicative of positive selection, in the form of recurrent selective sweeps. In this case, the rate of nonsynonymous change is greater than the rate of synonymous change because multiple changes in the amino acid sequence are adaptive. Defining selective forces using dN/dS ratios may be misleading as relaxation of selective constraint (i.e., pseudogenization, loss of function, etc.) can appear as positive selection. For example, the mammalian insulin gene is highly conserved across species and dN/dS ratios indicate rapid positive selection in guinea pig and chinchilla insulin, while it is later argued that it is due to the absence of the zinc ion in their insulin and corresponding relaxation of conserved sites involved with zinc (King, 1969; Kimura, 1986; Opazo et al., 2005).

Nonsynonymous substitutions are more likely to change the function of a protein than synonymous substitutions because they ultimately change the amino acid composition. However, some synonymous substitutions may affect protein abundance and interactions through codon usage bias, changes in regulatory binding sites, and changes in translation synthesis and folding (Sauna and Kimchi-Sarfaty, 2011). Some nonsynonymous changes are considered conservative (or neutral) and others are considered radical. The ratio of radical to conservative amino acid changes are used in predicting selection upon protein sequences in a similar fashion to how nonsynonymous/synonymous changes are used with gene sequences. Radical changes can be classified by volume and polarity, charge, or Grantham's distance (1974): a physiochemical measure. Many human disease Amino Acid Substitution (AAS) prediction methods find radical/conservative ratios useful in a context of protein structure and sequence (Ng and Henikoff, 2006). Mutations that affect function are usually at evolutionarily conserved sites or

buried in protein structure (Saunders and Baker, 2002). Disease causing AASs are overabundant at conserved sites (Miller and Kumar, 2001) and 83% of disease causing mutations affect protein stability (Wang and Moulton, 2001).

1.2.2 Evolution of proteins with activity

Functional protein sequences are often located near other functional sequences (Keefe and Szostak, 2001; Taverna and Goldstein, 2002). However, mutations far away from catalytic active sites can influence protein function (Spiller et al., 1999; Shimotohno et al., 2001). A large fraction of random substitutions do not detectably change structure and function of a protein; they are usually modest and additive (Shortle and Lin, 1985; Pakula et al., 1986; Loeb et al., 1989; Serrano et al., 1993; Bloom et al., 2006), yet small changes in structure or chemical properties can have substantial effects on catalytic function (Romero and Arnold, 2009). This may be explained by the reduction of fitness effects (either positive or negative) as an enzyme reaches optimal efficiency (Hartl et al., 1985). Most random mutations are considered destabilizing (Pakula et al., 1986; Matthews, 1995; Kumar et al., 2006). It has been shown that natural proteins are marginally stable with free energies (ΔG) between -5 and -15 kcal/mol (Fersht, 1999). If a mutation does not shift the thermodynamic stability of the overall protein it may persist.

Compensatory mutations are common in adaptive evolution and occur in the same gene as a deleterious mutation (Wang et al., 2002; Poon and Chao, 2005). Usually ten to twelve compensatory mutations are required for each deleterious mutation in a protein. A single amino acid change that is favorable in one protein dimension (function, stability, etc.) is most likely negative in other dimensions. Therefore, adaptive change in function probably demands a series of events to restore biochemical constraints (DePristo et al., 2005). However, globally expressed

protein-folding chaperones and heat shock proteins may offset unfavorable mutations in various adaptive proteins, reducing the number of gene-specific compensatory mutations, which is evident when they are inhibited or overexpressed (Rutherford, 2003; Sangster et al., 2004).

Most contemporary enzymes are thought to have evolved from promiscuous functions of ancestral proteins (Babtie et al., 2010). A promiscuous activity is an enzyme's ability to catalyze an alternative function that is not its primary function. If promiscuous activity of a protein is physiologically insignificant to the overall function and stability of the protein, the promiscuous activity could persist (Jensen, 1976; O'Brien and Herschlag, 1999; Khersonsky et al., 2006). Promiscuous activities are usually less efficient than the primary function of an enzyme (O'Brien and Herschlag, 1999; Wolfenden, 2006; Jonas and Hollfelder, 2008; Tawfik, 2010). Promiscuity could play a biological role in organismal survival to environmental changes, and has been shown to contribute to observed tolerance of gene deletion in metabolic pathways (Kim and Copley, 2007). An experimental study of dehydroxyacetone kinase (DHAK) in *Citrobacter freundii* showed changes in amino acid composition had little effect on native enzymatic activity but increased promiscuous activity nearly 50-fold (Sánchez-Moreno et al., 2009). Promiscuous activity can occur in numerous ways such as substrate similarities in hydrophobic binding, substrates utilizing the same catalytic functional groups, or the influence of different metal cofactors in active sites (Wang et al., 2003; Khersonsky and Tawfik, 2005; Afriat et al., 2006; Villiers and Hollfelder, 2009).

Evolutionary protein biochemists propose a conceptual framework for protein evolution similar to Wright's adaptive landscape metaphor. Neutral Networks conceptualize a protein sequence space where one functional protein is linked to all other functional proteins that have a single nucleotide difference (Smith, 1970; Huynen et al., 1996; Tiana et al., 2000). Therefore,

evolution can move one mutation at a time (an adaptive walk) through a continuous network of functional proteins. Otherwise, mutations would always lead to lower fitness (Smith, 1970). Mutations that spread through neutral genetic drift may impact future evolution through altering the functional neutral network (DePristo et al., 2005). Therefore neutral and adaptive protein evolution may be coupled through protein thermodynamics and the ever changing neutral network (Bloom et al., 2007; Bloom and Arnold, 2009).

Proteins are undeniably complex and have to balance promiscuity, functionality, and stability to survive. Protein evolution depends on two critical factors: 1) induction of new phenotypic functions by a relatively low number of mutations (the simplest evolutionary path) and 2) reduction of lethality from beneficial mutations, in essence compensatory or back mutations (Aharoni et al., 2005). Utilizing neutrality tests, the prevalence of advantageous and compensatory mutations would be identified as positive selection (through increased abundance of nonsynonymous mutations), which would prove beneficial as a first indicator of protein evolution. Though the neutral theory may not be entirely correct, nor the statistical neutrality tests error free, they may provide a basis for identifying selection. However, neutrality tests can also produce false negative and positive results, and it is extremely important for further investigation. Nearly neutral mutations may increase a protein's robustness, or ability to withstand additional (maybe beneficial) mutations (Bloom et al., 2006; Bloom et al., 2007). These nearly neutral mutations may open new paths for adaptation, including but not limited to optimizing functions, increasing promiscuity, or through gene duplication division of function.

1.2.3 Eukaryotic regulation and its role in protein evolution

King and Wilson suggested that evolutionary changes between species could not come from protein sequence changes alone, as sequence similarity is high, particularly between human

and chimpanzee, yet the expression is quite different (1975). They hypothesized that the differences must be due to gene regulation. Regulation of genes occurs at every stage from DNA transcription to the final product, protein. Initiation of transcription is intensely regulated by both DNA sequences (*cis*-regulatory) and DNA-binding proteins (*trans*-regulatory). The basic genomic structure of a gene includes a promoter with a TATA box and transcription start site (TSS) commonly upstream of the first exon. However, additional *cis*-regulatory sequences, like enhancers, silencers, or insulators may be involved, and are usually located away from the gene (Maston, 2006). DNA mutations in these *cis*-regulatory regions can drastically affect the expression of their associated gene; whereas mutations in *trans*-regulatory factors would affect the expression of many genes involved in a pathway, or general transcription. Therefore, *trans*-regulatory factors are most likely conserved and slow evolving, while *cis*-regulatory regions still have constraint, they are able to evolve faster than *trans*-regulatory factors (Arnone and Davidson, 1997).

After a eukaryotic gene is transcribed, pre-mRNA undergoes processing that ultimately affects the expression of the protein. Pre-mRNA transcripts can be alternatively spliced, which results in different isoforms of proteins. Depending on particular conditions one isoform may be expressed over another. The 5' cap and 3' poly-A tail aid in mRNA stability, and the longer the mRNA is stable, the longer the transcript is used to produce protein (Proudfoot et al., 2002). In addition, a protein's function and abundance is regulated through post-translational modifications, such as phosphorylation, glycosylation, lipidation, and ubiquitination. Essentially, a protein's activity and function can evolve through regulation, protein-coding sequences, or a combination of the two. Though this dissertation focuses mainly on species differences in

abundance or function of proteins, these results should be supplemented with research addressing how those changes developed.

1.3 Sexual selection

Darwin originally described sexual selection as the advantage some individuals have over others (of the same sex) in a species with respect to reproduction alone (Darwin, 1859; 1883). Darwin focused on “secondary characteristics” which involve ornamentation to attract mates or physical characteristics to combat other individuals, which today is often associated with precopulatory sexual selection. Though these attributes help acquire mates, there is also a behind the scenes battle of the sexes, referred to as postcopulatory sexual selection. Jones and Rattlerman suggested that “sexual selection arises from differences in reproductive success caused by competition for access to mates or fertilization opportunities” (2009). This modified definition incorporates both precopulatory and postcopulatory sexual selection (Table 1.1), which will be further defined in the following subsections.

Table 1.1: Types of sexual selection and their anatomical and physiological adaptations

	Precopulatory (Mating success)	Postcopulatory (Fertilization success)
Intersexual selection	Female choice	Female cryptic choice
	<ul style="list-style-type: none"> • Nuptial gifts • Courtship (visual or vocal displays) • Ornamentation • ‘Healthier males’ 	<ul style="list-style-type: none"> • Modifying pH levels • Growing eggs to maturity • Preparing uterus for implantation • Allowing completion of copulation • Ejecting sperm
Intrasexual selection	Male to male combat	Sperm competition
	<ul style="list-style-type: none"> • Sexual dimorphism • Competition weaponry • Canines • Antlers • Increased levels of testosterone 	<ul style="list-style-type: none"> • Increasing sperm viability • Faster sperm motility • Stimulation of the reproductive tract • Formation of copulatory plug • Suppression of female chemical and immunological challenges to sperm

In most cases of intersexual selection the female is the ‘choosey’ gender, while males are competitive for mates, and if given the opportunity, are more likely to mate with any female (Table 1.1). This gender bias in reproductive strategies seems likely due to the fundamental asymmetry of sex. In many species, females typically have a relatively large investment in producing offspring, beginning with using a lot of energy to produce a limited number of eggs, while their male counterparts can produce millions of sperm with ease. In addition to differences in gamete investment, females provide extensive care to offspring. However, there are several species where the sex roles are reversed and the males have a larger investment in offspring, like seahorses, where males are choosier and females become competitive (Berglund et al., 2005).

1.3.1 Precopulatory strategies

Precopulatory strategies are mechanisms occurring before insemination, which increase a male’s chance of mating with a female. Both males and females influence the probability of individual females obtaining the sperm from certain male individuals. Mating success is determined by the number of successful matings an individual has over their lifetime.

1.3.1.1 Intersexual precopulatory strategies

Female choice is the main form of precopulatory intersexual selection. A female may choose a mate for direct or indirect benefits. Direct benefits would include, but are not limited to, parental care, nuptial gifts, and territory defense. An indirect benefit is mating with a male that has perceived superior genes to aid in offspring survivability (Jones and Ratterman, 2009). These sexual signals come in many forms: acoustic (mating calls), visual (coloration or appendages), and chemical (i.e. pheromones), which provide information to females about male quality and potential mating benefits.

Fisher proposed that this type of ‘artificial selection’ by the female could lead to runaway selection, where if certain traits have a selective advantage in males, and become more prevalent in the population, then females will also evolve genes to prefer that phenotype (1915). If sexual selection was the only selective pressure, this ‘runaway selection’ could continue indefinitely with the males’ evolving more elaborate ornamentations and the females’ preference evolving to favor the extreme. Sexual selection is not the only prevalent force; thus, there must be an upper limit to ornamentation that intersects with an individual’s ability to survive (Pomiankowski and Iwasa, 1998).

Sometimes elaborate signals provide accurate information about the individual’s health, and other times they are falsely representative. An example of a ‘honest’ signal is that the survival of offspring positively correlates with the number of eye spots in a male peacock’s tail (Møller and Petrie, 2002; Loyau et al., 2005). In fiddler crabs, females prefer males with a larger claw; which also warns other males to choose their battles. If a fiddler loses its claw, it is able to regenerate a new claw to signal its ability to fight; however, the regenerated claw is a false signal, as it is actually weak (Backwell et al., 2000).

1.3.1.2 Intrasexual precopulatory strategies

Precopulatory intrasexual selection includes male to male combat. Species that are adapted through intrasexual selection are typically sexually dimorphic (Gould and Gould, 1989), where males have an enlarged body size and enhanced fighting appendages. Male gorillas are much larger than females with well-developed muscles and strong jaws to aid in combat, in order to protect his territory and sole reproductive access of his harem of females. In addition to large size, animals have competition weaponry, Examples of competition weaponry include: large toothed claws in fiddler crabs, antlers in ungulates, tusks in walruses, and horned mandibles in

many beetle species (Emlen, 2008). While enlarged competition weaponry certainly aids in competition success, in return, they can be costly to the individual by requiring a lot of nutrition or maintenance, or the appendage weight and size could impede common tasks, like locomotion or eating, which are essential to survival.

1.3.2 Postcopulatory strategies

Postcopulatory strategies increase a male's probability of fertilizing the egg. The number of progeny sired per copulation describes fertilization success. Both males and females have adapted strategies suited for their respective reproductive goals in the battlegrounds of the female reproductive tract.

1.3.2.1 Intersexual postcopulatory strategies

Intersexual postcopulatory strategies are those that involve cryptic female choice. These are female controlled processes that modulate a male's chance for insemination. It is considered 'cryptic' because it is hidden in the female reproductive tract (Birkhead and Pizzari, 2001). Some female controlled processes are modified pH levels, growing more eggs to maturity, producing eggs with more nutrients, preparing the uterus for implantation, ejecting sperm, and allowing completion of copulation (Eberhard, 2009). When seminal fluid enters the female reproductive tract, several alterations occur in the female. In some species, there is an increased clearance of microorganisms and extra spermatozoa, the maternal immune response is activated, tissue is remodeled, and cytokines and growth factors, are activated for pre-implantation embryo growth development (Robertson et al., 2002; Robertson, 2005).

Seminal plasma significance in uterine clearance has been studied in livestock species, and has indicated that increased inflammatory response to seminal plasma is positively correlated with implantation success (Robertson et al., 2002). The increased inflammatory response

increases the number of phagocytic cells in the cervix. Phagocytic cells target microorganisms and sperm with abnormal morphologies. The female reproductive tract may inactivate and sequester sperm to favor genetically superior or phenotypically more competent sperm (Robertson, 2005). It is inevitable that some viable spermatozoa will be destroyed as well.

Immunoregulatory molecules present in the seminal plasma, like TGF β , protect the majority of sperm surface antigens from feminine rejection. In mice, TGF β 1 is originally inactive in the seminal vesicles and it becomes activated in the female uterus (Robertson, 2005). TGF β was identified as the principle active immune component in human seminal plasma. This promotes partner-specific embryo implantation and is considered to be “priming.” The female reproductive tract is primed for an individual male until their immunoregulatory molecules are cleared (Robertson et al., 2002).

1.3.2.2 Intrasexual postcopulatory strategies

Sperm competition is the major form of male-male postcopulatory intrasexual selection. Sperm competition was originally defined as “competition between the sperm of two or more males for the fertilization of a given set of ova” (Parker, 1970). Sperm competition strategies include increasing sperm viability, faster sperm mobility and capacitation, stimulation of the female reproductive tract, suppression of female chemical and immunological challenges to sperm, and in some species the formation of a copulatory plug (Ramm, 2008). A copulatory plug is a semi-solid gelatinous ejaculate that produces a barrier in the female reproductive tract, which is thought to aid in fertilization or have a “chastity-belt” function by blocking access to the ova from future mating males (Dixson and Anderson, 2004). In mammalian species with increased female promiscuity, there is heightened sperm competition. Increased sperm competition leads to changes in sperm morphology and abundance, physiological anatomy, and behavior in males.

Sperm morphology and number are influenced by increased sperm competition. Sperm tend to be slightly smaller with a larger midpiece for better movement in species with high sperm competition (Dixon and Anderson, 2004; Ramm et al., 2009). The mitochondria that are required for flagellar movement are located in the midpiece, which would explain increased midpiece volume, for faster motility. In primates, midpiece volume has been correlated with mating systems and relative testis size (Figure 1.2: B; Anderson and Dixson, 2002; Dixson and Anderson, 2004). Under increased sperm competition, males have a higher sperm count, larger volume of ejaculate, and a greater number of ejaculations than species with lower sperm competition. There is a fitness cost to increased sperm production, and male gametic investment should reflect species levels of female promiscuity and degree of sperm competition (Ramm et al., 2008).

To amplify spermatogenesis production, seminiferous tubular tissue is expanded, resulting in larger testes size in relation to body weight. This has been shown in some bats (Hosken, 1997), porpoises (Fontaine and Barrette, 1997), primates (Short, 1979; Harcourt et al., 1981; Harcourt et al., 1995; Kappeler, 1997), metatherian mammals (Rose et al., 1997; Taggart et al., 1998), and eutherian mammals (Kenagy and Trombulak, 1986). Mammalian species with larger relative testes have been shown to produce higher sperm count per ejaculate, and within species those with larger relative testes size have increased reproductive success (Dixson, 1998).

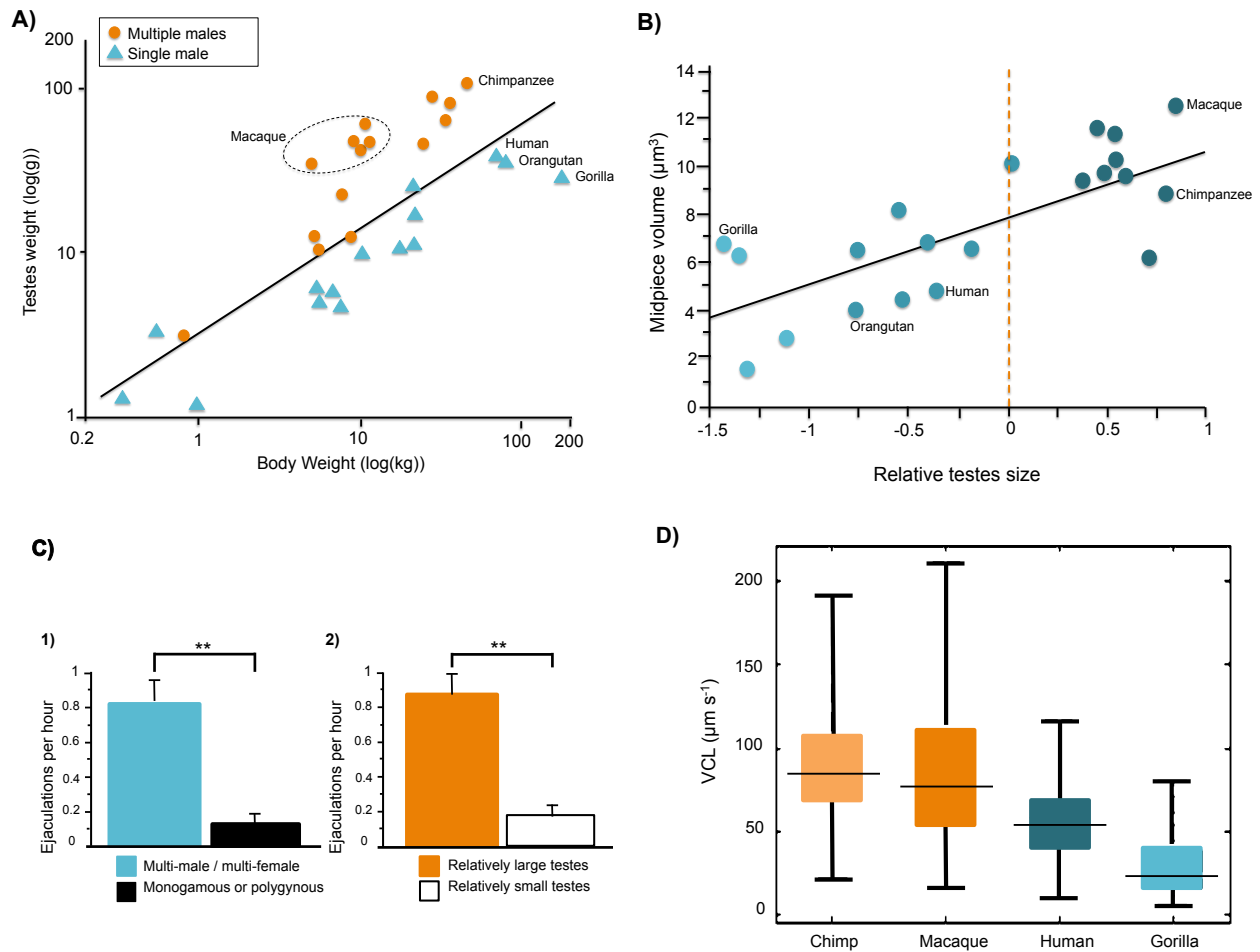


Figure 1.2: Primate sperm competition phenotypes

A) Primates in multi-male mating systems (orange circles) have relatively large testes compared to body size than males with single-male mating systems (blue circles). Data are from Harcourt et al., 1981 and the graph was redrawn from Dixson, 2012. **B)** Sperm midpiece volume is correlated with relative testes to body size ratio. Primates with multi-male mating systems have a larger midpiece volume and a larger testes to body size ratio. The graph was redrawn from Dixson, 2012, which Dixson incorporated data from their previous publications. **C)** Average frequency of ejaculations in primate genera having 1) multiple partner mating systems (blue bar, n=6) compared to single partner mating systems (black bar, n=6), and 2) large relative testes size (orange bar, n=5) compared to small relative testes sizes (white bar, n=7). Primates of multi-male mating systems and those with a relatively large testes to body size ratio ejaculate more frequently (**p < 0.01; (Mann–Whitney U-test)) than primates of single-male mating systems and those with a smaller testes to body size ratio, respectively. This graph was redrawn from Dixson and Anderson, 2004; which incorporated unpublished data and data from Dixson, 1995. **D)** Chimpanzee and macaque, primates with multi-partner mating systems (orange bars), have increased velocity of sperm compared to human and gorilla, with single-male mating systems (blue bars). This graph was redrawn and modified from Nascimento et al., 2008.

Anatomical reproductive features may be reduced or lost in species with relatively low sperm competition. There are significant anatomical differences between single-partner and multi-partner mating systems, which are associated with penile length, baculum length, and distal complexity (Figure 1.3; Dixson, 2009). A human-specific loss of regulatory DNA in front of the androgen receptor locus was hypothesized to have led to the loss of male penile spines in humans. It was concluded that this regulatory DNA deletion was not deleterious to humans because of our mostly monogamous mating system (McLean et al., 2001). However, these anatomical traits are also associated with copulation time, which has not been well correlated with mating systems. In species with shorter copulation times, the penile bone has been reduced or lost. This has been shown in hyenas, humans, and tarsiers. Likewise, in species with longer copulation times, mammals have an elongated penile bone, which has been shown in some pinnepeds, bats, stumptail macaques, and galagos (Dixson and Anderson, 2004).

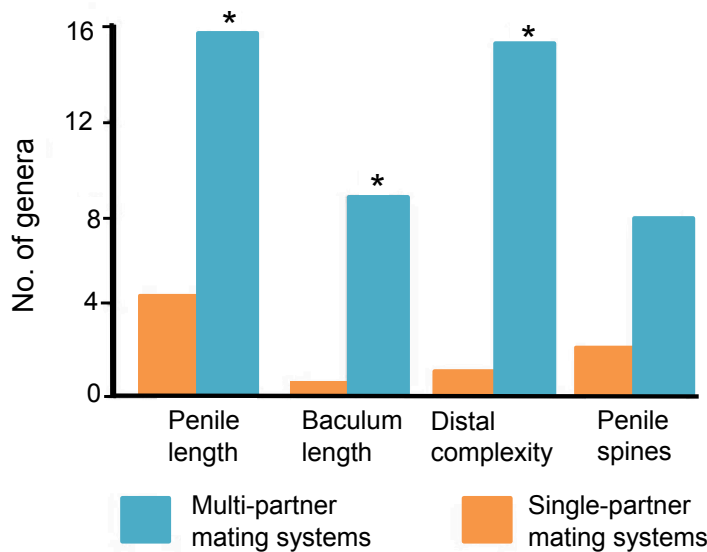


Figure 1.3: Morphological penile differences among multi-partner and single-partner mating systems

Penile morphologies are more complex in primate genera where females mate with multiple males (multi-male / multi-female or dispersed mating systems, defined as multi-partner) than genera where females typically mate with a single male during estrus (monogamous or polygynous mating systems, defined as single partner). Data are from 48 primate genera and show the number of genera that received either a 4 or 5 (on a 5 point scale) for various traits (*p < 0.05). Graph was redrawn from Dixson (2012) and data was from Dixson, 2009.

1.4 Primate sperm competition, physiology, and seminal plasma proteins

1.4.1 Primate mating systems, sperm competition, and reproductive physiology

Gorillas (*Gorilla gorilla*) have a polygynous mating system where multiple females mate with a single dominant male during estrus (Watts, 1990). A gorilla troop usually has a silverback male, several females, and dependent offspring, with an average size of 10 individuals with a maximum of 20. The smallest group size begins with one male and one female (Harcourt, 1978). The troop may also include a young reproductive subordinate male, or 'black back', from within the group for a short period of time (Robbins et al., 2005). The subordinate males in the group form non-threatening relationships with the females consisting of protection, food, and grooming. Conversely, they try to sneak fertilizations with estrous females, which are often interrupted by the silverback who may guard the females from these advances (Watts, 1996). These subordinate males may migrate from the group alone, in search of a mate, wait for the dominant male to die, or leave with a non-related female group member (Bradley et al., 2005). Mountain gorillas are unique in that some groups may have two related, or unrelated reproductively successful silverbacks, with a dominant silverback siring the majority, but not all, of the offspring (Bradley et al., 2005). Gorilla sperm competition is likely to be low and gorillas have relatively small ejaculate volume (Seager et al., 1982; Dixson, 1997).

Chimpanzees (*Pan troglodytes*) have multi-male/multi-female mating systems, where a female will mate with multiple males in rapid succession during estrus (White, 1996). During estrus, female chimpanzees exhibit prominent "sexual swelling" of the urogenital skin, presumably to attract males. Groups have many adult males with high levels of male kinship, and many adult females with little to no kinship. Chimpanzees have a fission-fusion society (Fleagle, 2013; White, 1996; Dixson, 2012) with small foraging groups regularly splitting and rejoining.

Chimpanzees and closely related bonobos (*Pan paniscus*) are the only hominoids to form a copulatory plug and produce a relatively large volume of ejaculate (Dixson, 1997; Dorus et al., 2004; Dixson and Anderson, 2004).

The human mating system is a challenge to classify due to societal differences, but most humans are monogamous or polygynous, where we define monogamy as the tendency to mate with a single male during any given cycle (Marlowe, 2000). Humans, when compared to all other primates, have the most diversity in their social organization (Fleagle, 2013). Notably, humans of both genders participate in fission-fusion, where often familial members will separate themselves from their mates and offspring to participate in activities with other individuals of both genders before returning to their familial unit (Aureli et al., 2008). Humans live in many social settings, with any number of individuals that are related or unrelated, male or female, and of various ages. There are cultural constraints specific to each population that determine how humans interact/live within their environment, and how they think about their interactions (Costa et al., 2001; Özer et al., 2013, Fleagle, 2013). The aforementioned variability makes it difficult to categorize humans into any one social or mating system (Fleagle, 2013). Relative testis weight to body size suggests that humans are polygynous, but with more sperm competition than gorillas (Harcourt et al., 1981; Dixson 2009). Using sperm midpiece volume and mitochondrial density, humans look more monandrous than chimpanzee (Anderson and Dixson, 2002; Anderson et al., 2007). Humans have sexually dimorphic characteristics, where males are larger than females and have facial hair, a more muscular build, an enlarged larynx, and a deepened voice (Dixson, 2012). This dissertation will consider humans as mostly monogamous with relatively moderate levels of sperm competition, compared to low levels of sperm competition in gorilla, and higher levels of sperm competition in chimpanzees.

Although these are generalizations, mating systems with higher female promiscuity are projected to have higher sperm competition (Karr and Picnick, 1999; Dixson and Anderson, 2001; Birkhead and Pizzari, 2002). Primate species that experience high sperm competition have been shown to have larger testes to body size ratio (Figure 1.2: A; Harcourt et al., 1981). Sperm midpiece volume, likely important in sperm motility, is positively correlated with relative testes size (Figure 1.2: B; Anderson and Dixson, 2002; Dixson and Anderson, 2004). Similarly, seminal vesicles are larger in species with higher sperm competition (Dixson, 1997). The number of ejaculations (Figure 1.2: C; Dixson, 1995; Dixson and Anderson, 2004), concentration of motile sperm, volume of sperm midpiece, and movement of sperm increase (Figure 1.2: D; Nascimento et al., 2008) with higher sperm competition.




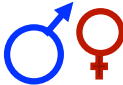


	Human	Chimpanzee	Gorilla
Mating system	<i>Mostly Monogamous</i> with some polyandry and polygyny	Multi-male / Multi-female	Single-male Polygynous
Ejaculate	Moderate volume Viscous	Large volume High sperm count Formation of copulatory plug	Small volume Liquid
Testes weight to body size ratio	 Moderate	 Large	 Small
Male to female body size ratio			
Inferred level of sperm competition	Low to moderate	High	Low

Figure 1.4: Summary of hominid mating systems, sperm competition, and physiology

This summary presentation is original and visually describes mating system, ejaculate volume and consistency, relative testes size, level of sexual size dimorphism, and relative level of sperm competition of hominids (reviewed in section 1.4.1). For testes weight to body size ratio, the size of the circles represent the relative size of testes, with the larger circles indicating a larger ratio. For male to female body size, the male/female symbols represent approximate size of each gender in relation to the other. Section 1.1 provides more detail about body size differences of these primates.

1.4.2 Seminal plasma proteins

Semen is the product of five main organs in primates: the testes, epididymis, prostate, seminal vesicles, and the Cowper gland (Mann and Lutwak-Mann, 2012). Seminal plasma, the liquid portion of semen separated from spermatozoa, affects various aspects of fertilization, including: sperm motility, female immunological suppression, female receptivity, sperm storage, and copulatory plug formation. Seminal plasma provides energy and buffers the acidity of the vagina, containing proteins, lipids, sugars, potassium, sodium, calcium, magnesium, and zinc (Owen and Katz, 2005).

Proteomic studies show that seminal plasma proteins vary between individuals in humans and rodents (Oliva et al., 2008; Ramm et al., 2008); however, there is a need for proteomic studies assessing differences between species. Currently, there has only been one other study conducted on non-human primate seminal plasma, and as of yet, it has only been written in a dissertation (Claw, 2013). They reported that the most abundant proteins across multiple primate species were involved in semen coagulation and liquefaction, and particularly the copulatory plug pathway for some mating systems with high female promiscuity. These proteins were semenogelin 1 and 2, prostate specific transglutaminase, kallikrein related peptidase 3, and prostatic acid phosphatase; at least one of these proteins was within the top five most abundant proteins in each species, despite the associated mating system (Claw, 2013). The aforementioned abundant seminal plasma proteins are important to all subsequent chapters of this dissertation; therefore, they will be discussed extensively in this introduction chapter.

1.4.2.1 Semenogelin proteins

1.4.2.1.1 Structure of SEMG1 and SEMG2 genes

The human semenogelin genes, *SEMG1* and *SEMG2*, are duplicated genes located adjacent to one another on human chromosome 20. The locus in which they are located is called the Wap four-disulfide core (WFDC) domain locus, which contains the semenogelin genes, *PI3*, *EPPIN*, *SLPI*, 3 *SPINT* genes, and 11 *WFDC* genes (Peter et al., 1998; Clauss et al., 2002). Both *SEMG1* and *SEMG2* have a conserved gene structure consisting of three exons (Figure 1.5:A; Ulvsbäck et al., 1992). The first exon has a leader and a signal peptide coding sequence, the second exon encodes for the entire secreted peptide sequence, and the third exon is non-coding. The timing of the gene duplication, which resulted in primate *SEMG1* and *SEMG2*, is estimated to be 61 million years ago and is perhaps thought to be the result of homologous recombination between LINE - L1 elements (Lundwall, 1996).

SEMG1 and *SEMG2* genes vary in length among and within primate species due to expansion of 180 base pair repetitive units in the second exon. These repetitive units account for greater than 80% of the coding sequence, and are split into three groups based on sequence similarity (Lundwall and Lazure, 1995). Exon 2 of human *SEMG1* has two (a and b) of each type of 60 amino acid repeats (I, II, and III) and exon 2 of human *SEMG2* has four (a, b, c, and d) repeats of type I, two (a, b) repeats of type II, two (a, b) repeats of type III (Figure 1.5:B; Lilja and Lundwall, 1992). Type I repetitive units are the putative sites for transglutaminase cross-linking activity (Zalensky et al., 1993; Robert et al., 1997; Ulvsbäck and Lundwall, 1997; Robert and Gagnon, 1999), which is believed to be important in the functionality of the SEMG proteins; it is interesting that the SEMGs differ among species mainly in the number of type I repeats. The

number of repeats in exon 2 of *SEMG1* is more variable than the number of repeats in *SEMG2* across primates (Figure 1.5:C; Jensen-Seaman and Li, 2003).

Since the duplication in primates, the *SEMGs* have undergone numerous rearrangements, pseudogenizations, open reading frame expansions, deletions, and homogenizations. This is often a trademark of duplicated gene fate, where conservation of both genes in the ancestral form is rare. Neo-functionalization may occur when one of the genes gains a new function, but most commonly, sub-functionalization occurs with function preservation in one gene and complementary loss of function in the other gene (Zhang, 2003). Several deletions have occurred in New World monkeys; the cotton-top tamarin (*Saguinus oedipus*) only has *SEMG1* due to the deletion of the *SEMG2* gene by insertion of a *LINE1* element (Ulvsbäck and Lundwall, 1997; Lundwall and Olsson, 2001). The owl monkey (*Atous nancymaae*) has a single chimeric semenogelin due to an *Alu* insertion, which resulted in a *SEMG1*-like exon 1 and intron 1 and a *SEMG2*-like exon 2, intron 2, and exon 3 (Hurle et al., 2007).

There is a wide distribution of premature stop codons within the protein coding sequence (Figure 1.5: D) among the primates. Only human, orangutan, vervet, colobus, and the marmoset seem to have full-length functional copies of both *SEMGs*. Chimpanzees, bonobos, and baboons have a premature stop codon in the coding region of *SEMG2* and gibbons have a premature stop codon in *SEMG1* (Jensen-Seaman and Li, 2003; Dorus et al., 2004; Hurle et al., 2007). Gorillas have premature stop codons in exon 2 of both *SEMGs*; in *SEMG1*, gorillas are polymorphic for three premature stop codons, in *SEMG2* they are polymorphic for two premature stop codons (Jensen-Seaman and Li, 2003).

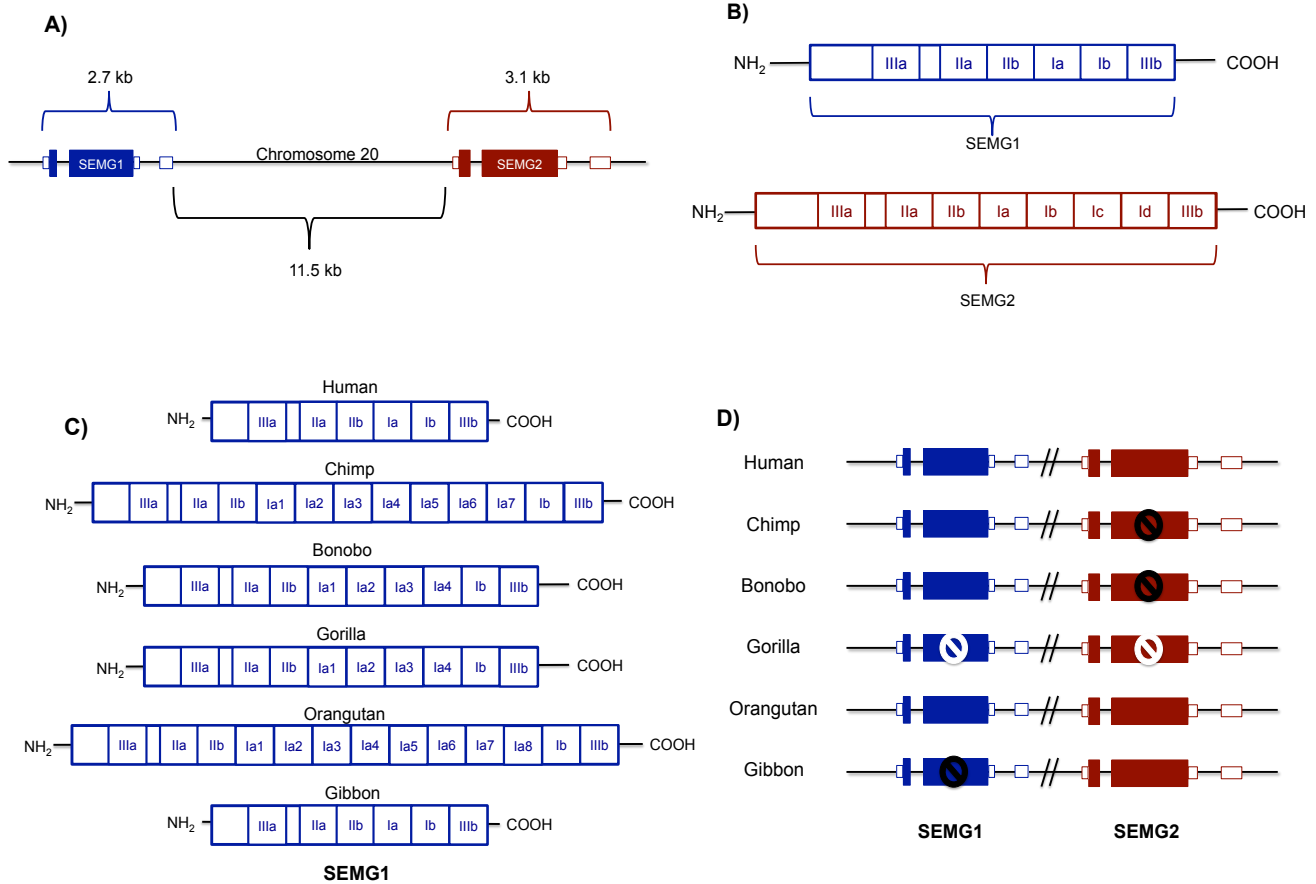


Figure 1.5: SEMGs genomic and protein structure

A) Schematic of *SEMG1* (blue) and *SEMG2* (red) on the q arm of human chromosome 20 (12-13.1). Both paralogs have three exons, with the solid color region representing either the signal peptide or protein coding sequence, and the outlined regions represent regulatory regions. *SEMG1* is 2.7 kb in length, there is 11.5 kilobases between paralogs, and *SEMG2* is 3.1 kb in length (UCSC genome browser). **B)** Schematic of human *SEMG1* (blue) and *SEMG2* (red) proteins and location of 60 amino acid repeats: I, II, and II (Lilja and Lundwall, 1992). **C)** *SEMG1* has variability in amino acid repeats across hominoid species. Expansion of Ia repeats are numbered, with orangutan having the most (Jensen-Seaman and Li, 2003). **D)** *SEMG1* and *SEMG2* genes have premature stop codons located in some of the hominoid species. Black symbols indicate fixed premature stop codons located in either *SEMG1* (blue) or *SEMG2* (red) exons; whereas, white symbols represent polymorphic premature stop codons (Jensen-Seaman and Li, 2003). All figures are original or redrawn from sources indicated above.

The protein-coding region of semenogelins shows evidence of rapid sequence change. A dN/dS ratio above 1 indicates positive selection (or rapid evolution); whereas, a value below 1 indicates purifying selection (or conservation) of the gene. *SEMG2* shows a higher dN/dS value on chimpanzee and bonobo lineages with maximum likelihood comparison to other hominid

lineages (Table 1.2; Dorus et al., 2004). Strikingly, the pattern of selection on both genes seems to correlate with higher levels of polyandry, or female promiscuity, and the degree of semen coagulation (Dorus et al., 2004) with the more rapidly evolving lineages being those that have the highest level of sperm competition.

Table 1.2: Rates of SEMG2 evolution compared to species level of polyandry*

Species	Mating system	Mean number of male partners per estrus	dN/dS
Gorilla	Polygynous	~1	0.61
Colobus monkey	Polygynous	~1	0.70
Gibbon	Monogamous	~1	0.89
Orangutan	Dispersed	1-2	0.88
Human	Various	1-2	0.91
Macaque	Multi-male / Multi-female	~3	1.28
Chimpanzee	Multi-male / Multi-female	~8	2.52

*Reproduced from Dorus et al., 2004

1.4.2.1.2 Expression and function of Semenogelin proteins

SEMG1 and SEMG2 are mainly expressed in the seminal vesicles in vast quantity (Bjartell et al., 1996; Lundwall et al., 2002; Koistinen et al., 2002; Yoshida et al., 2003), but also have expression (although lower) in the vas deferens, prostate, epididymis, and trachea. In addition to the above, SEMG1 has some expression in the gastro-intestinal tract and skeletal muscle, while SEMG2 has some expression in the kidney and testes (Lundwall et al., 2002).

In humans, the semenogelin paralogs share 78% identity in their amino acid sequence (Lilja et al., 1989); however, they differ in molecular weight by approximately 20 kilodaltons (kDa). SEMG2 is between 71 and 76 kDa (depending on glycosylation state) and SEMG1 is approximately 52 kDa (Lilja et al., 1985; Lilja and Lundwall, 1992; Malm et al., 1996). Non-

covalent interactions and the covalent cross-linking of semenogelins generate a coagulum, which is stabilized by zinc and traps spermatozoa (Jonsson et al, 2006; Malm et al., 2007). In addition to semen coagulation, the semenogelins have known functions in inhibiting sperm motility (Robert and Gagnon, 1995), sperm capacitation (de Lamirande et al., 2001), and antimicrobial properties (Bourgeon et al., 2004; Edström et al., 2008; Mitra et al., 2010). Upon ejaculation semenogelin proteins adhere to the surface of the sperm and form an eppin complex through covalent crosslinkage. The eppin complex provides antimicrobial protection for sperm (Mitra et al, 2010; O’Rand et al., 2011).

After seminal coagulation, the SEMGs are cleaved by seminal fluid peptidases, KLK3 (Kallikrein related peptidase 3, also known as PSA; Lilja, 1985) and to a lesser extent by ACP (Prostate acid phosphatase, also known as PAP; Brillard-Bourdet, et al. 2002), facilitating liquefaction of semen. Semenogelins are a specific substrate for KLK3 and are rapidly fragmented when in solution together (Peter et al., 1998; Robert and Gagnon, 1999; Jonsson et al., 2006). The cleavage of semenogelin releases bound and trapped spermatozoa (Robert et al., 1997; Flori et al., 2008).

1.4.2.2 Prostate specific transglutaminase

Prostate specific transglutaminase (TGM4) is a calcium dependent enzyme involved in semen coagulation and copulatory plug formation by crosslinking semenogelin (SEMG1 and SEMG2) proteins at glutamine and lysine residues (Greenberg et al., 1991; Folk, 1983; Esposito and Caputo, 2004). The gene coding for TGM4 is located on human chromosome 3 (Dubink et al., 1996; Dubink et al., 1998). In gibbons, gorillas, and other species with relatively low sperm competition, *TGM4* has become a pseudogene (Clark and Swanson, 2005; Carnahan and Jensen-Seaman, 2008; Tian et al., 2009). *TGM4* knockout mice had reduced fertility and were unable to

form a copulatory plug, a common phenotype in mice (Dean, 2013). The expression of TGM4 can be upregulated by retinoic acid (Pasquali et al., 1999; Rivera-Gonzalez et al., 2012) and is down regulated by the steroid hormone androgen (Dubbink et al., 1998; Rivera-Gonzalez et al., 2012).

1.4.2.3 Kallikrein related peptidase

Kallikrein related peptidase (KLK3), otherwise known as prostate specific antigen (PSA) is a trypsin-like serine protease involved in semen liquefaction by cleaving semenogelin (SEMG1 and SEMG2) proteins at serine residues (Lilja et al., 1987; Jonsson et al., 2005; Jonsson et al., 2006; Lundwall and Brattsand, 2008; Andrade et al., 2010). *KLK3* originated from a gene duplication of *KLK2* approximately 42 million years ago (Olsson et al., 2004). A study has identified chimeric *KLK3-KLK2* fusion genes in Gorilla (*Gorilla gorilla*) and gibbon (*Nomascus leucogenys*), a result of gene duplication and relaxation of selective pressures (Marques et al., 2012). Active and functional *KLK2* and *KLK3* genes are associated with higher female promiscuity (chimpanzees), while the dysfunction of one or both genes is associated with lower female promiscuity or monogamy (Marques et al., 2012).

Kallikreins are known to work in cooperation with one another in proteolytic cascade pathways, with one KLK activating another (Pampalakis and Sotiropoulou, 2007). KLK2 self-activates, and then cleaves and activates proenzyme KLK3 in the liquefaction process of seminal plasma (Pampalakis and Sotiropoulou, 2007; Lundwall and Brattsand, 2008). KLK2 and KLK3 are expressed in the prostate (Lovgren et al., 1997). KLK3 is mostly studied in association with prostate cancer because increasing levels of KLK3 (PSA – prostate specific antigen) is a clinical indication of prostate cancer (Schieferstein, 1999; Loeb and Catalona, 2007). Though KLK3 has been studied in association with fertility, there have not been any significant findings in humans;

moreover, mice knockout models are fertile. However, there is a slight negative correlation between sperm motility and KLK3, and a negative correlation of KLK3 to fructose concentration (Schieferstein, 1999).

1.4.2.4 Prostatic acid phosphatase ACPP

Tissues such as the prostate, brain, kidney, liver, lung, muscle, placenta, salivary gland, spleen, thyroid, and thymus express prostatic acid phosphatase (ACPP) (Goldfarb et al., 1986; Solin et al., 1990; Kong and Byun, 2013). In fact, ACPP is the most abundant phosphatase in human prostatic tissue (Hassan et al., 2010). ACPP is a homodimer of two non-covalently associated subunits, each of which is approximately 50kDa. Homodimerization is necessary for catalytic activity (Kuciel et al., 1990).

ACPP is located on human chromosome 3. There are two protein isoforms, intracellular and extracellular, which are encoded from this single gene. The smaller of the two, isoform 1, has 10 exons and encodes the secreted form of ACPP found in semen. The longer of the two, isoform 2, with 11 exons, contains a transmembrane domain and is intracellular (Winqvist et al., 1989, Li and Sharief, 1993). The intracellular and secreted isoproteins differ in isoelectric point values and biochemical properties, but are similar in that they both hydrolyze organic phosphomonoesters in acidic conditions (pH 4-6) (Vihanko et al., 1978; Lin et al., 1983; Lad et al., 1984). The secreted isoprotein is a 100kDa homodimeric glycoprotein, which is only secreted by columnar epithelial cells in the prostate gland at a concentration around 1mg/mL (Röhnberg et al., 1981; Solin et al., 1990; Lee et al., 1991; Shan et al., 2003). Each ACPP monomer has two domains. The larger domain is a seven-stranded β -sheet with α -helices on each side. The smaller domain is composed of six α -helices (Lee et al., 1991; Kuciel et al., 1996; Ortlund et al., 2003).

As a member of a family of enzymes known as acid phosphatases, the phosphatase activity of ACP is optimal in acidic conditions. Dephosphorylation of orthophosphoric monoesters and phosphorylated proteins by this enzyme are optimal in a pH range of 3-6 (Zelivianski et al., 1998; Brillard-Bourdet et al., 2002). In addition to the phosphatase activity, ACP acts as a protease, hydrolyzing the cleavage of SEMG1 substrates preferentially at Tyr136, Tyr292, and Gln266 amino acid residues (Brillard-Bourdet et al., 2002). The proteolytic role of ACP is associated with its extracellular function related to seminal liquefaction and has an optimal activity around a pH of 9 (Brillard-Bourdet et al. 2002).

An unknown protease within seminal fluid cleaves ACP. Cleaved peptides that remain stable contain amino acids 85-120 or 248-286, which then form amyloid fibrils called Semen-derived Enhancer of Virus Infection (SEVI) (Münch et al., 2007). Amyloid fibrils, which are insoluble aggregates formed from cleaved fragments of ACP or SEMG1, enhance HIV infectivity (Münch et al., 2007; Roan et al., 2011; 2014). SEVI increases infectivity several magnitudes (Münch et al., 2007), with its positively charged residues which interact with negatively charged HIV virions in order to target the cell membrane to promote attachment and uptake of the virus (Roan et al., 2011). Similarly, an amyloid fibril produced by SEMG1 (amino acids between 86-107), enhances HIV infection by 30-fold *in vitro* (Roan et al., 2011; 2014). This peptide possesses an overall positive charge that helps it bind to HIV-1 virions and augments infection (Münch et al., 2007).

1.5 Experimental goals

In this dissertation I aim to fully identify abundant seminal plasma proteins among multiple human, chimpanzee, and gorilla seminal plasma samples (Chapter 2). Proteins that are expressed in high abundance in one species compared to another species are likely important in producing the phenotypic differences in seminal fluid as a result of high or low sperm competition. I will also include a closer look at proteins (like SEMG1 and SEMG2) that have been predicted by others to be under positive selection and/or pseudogenization across the hominids. Finally, I take a comparative approach to understand the functional molecular evolution of an important enzyme in semen coagulation and copulatory plug formation, prostate-specific transglutaminase (TGM4), by examining the activity of recombinant human, chimpanzee, and the hypothetical human-chimpanzee ancestor proteins (Chapter 3).

1.6 References

- Afriat, Livnat, et al. “The Latent Promiscuity of Newly Identified Microbial Lactonases Is Linked to a Recently Diverged Phosphotriesterase.” *Biochemistry*, vol. 45, no. 46, 2006, pp. 13677–13686.
- Aharoni, Amir, et al. “The ‘evolvability’ of Promiscuous Protein Functions.” *Nature Genetics*, vol. 37, no. 1, 2004, pp. 73–76.
- Anderson, Matthew J., et al. “Functional Evidence for Differences in Sperm Competition in Humans and Chimpanzees.” *American Journal of Physical Anthropology*, vol. 134, no. 2, 2007, pp. 274–280, doi:10.1002/ajpa.20674.
- Anderson, Matthew J., and Alan F. Dixson. “Sperm Competition: Motility and the Midpiece in Primates.” *Nature*, vol. 416, no. 6880, 2002, pp. 496–496.
- Andrade, Douglas, et al. “Substrate Specificity and Inhibition of Human Kallikrein-Related Peptidase 3 (KLK3 or PSA) Activated with Sodium Citrate and Glycosaminoglycans.” *Archives of Biochemistry and Biophysics*, vol. 498, 2010, pp. 74–82.
- Arnone, Maria I., and Eric H. Davidson. “The Hardwiring of Development: Organization and Function of Genomic Regulatory Systems.” *Development*, vol. 124, no. 10, 1997, pp. 1851–1864.
- Atsalis, Sylvia, and Elaine Videan. “Reproductive Aging in Captive and Wild Common Chimpanzees: Factors Influencing the Rate of Follicular Depletion.” *American Journal of Primatology*, vol. 71, no. 4, 2009, pp. 271–282.
- Aureli, Filippo, et al. “Fission-Fusion Dynamics: New Research Frameworks.” *Current Anthropology*, vol. 49, no. 4, 2008, pp. 627–654.
- Babtie, Ann, et al. “What Makes an Enzyme Promiscuous?” *Current Opinion in Chemical Biology*, vol. 14, no. 2, 2010, pp. 200–207.
- Backwell, Patricia R. Y., et al. “Dishonest Signalling in a Fiddler Crab.” *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 267, no. 1444, Apr. 2000, p. 719, doi:10.1098/rspb.2000.1062.
- Berglund, Anders, et al. “Sex-Role Reversal Revisited: Choosy Females and Ornamented, Competitive Males in a Pipefish.” *Behavioral Ecology*, vol. 16, no. 3, 2005, pp. 649–655.
- Birkhead, Timothy R., and Tommaso Pizzari. “Post Copulatory Sexual Selection.” *Nature*, vol. 3, 2002, pp. 262–273.
- Bjartell, Anders, et al. “Distribution and Tissue Expression of Semenogelin I and II in Man as Demonstrated by in Situ Hybridization and Immunocytochemistry.” *Journal of Andrology*, vol. 17, no. 1, 1996, pp. 17–26.
- Bloom, Jesse D., Philip A. Romero, et al. “Neutral Genetic Drift Can Alter Promiscuous Protein Functions, Potentially Aiding Functional Evolution.” *Biol Direct*, vol. 2, 2007, p. 17.
- Bloom, Jesse D., Sy T. Labthavikul, et al. “Protein Stability Promotes Evolvability.” *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, 2006, pp. 5869–5874.
- Bloom, Jesse D., Alpan Raval, et al. “Thermodynamics of Neutral Protein Evolution.” *Genetics*, vol. 175, no. 1, 2007, pp. 255–266.
- Bloom, Jesse D., and Frances H. Arnold. “In the Light of Directed Evolution: Pathways of Adaptive Protein Evolution.” *Proceedings of the National Academy of Sciences*, vol. 106, no. Supplement 1, 2009, pp. 9995–10000.
- Bourgeon, Frédéric, et al. “Involvement of Semenogelin-Derived Peptides in the Antibacterial Activity of Human Seminal Plasma.” *Biology of Reproduction*, vol. 70, no. 3, 2004, pp. 768–774.
- Bradley, Brenda J., et al. “Mountain Gorilla Tug-of-War: Silverbacks Have Limited Control over Reproduction in Multimale Groups.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 26, 2005, pp. 9418–9423.
- Brillard-Bourdet, Michèle, et al. “Amidolytic Activity of Prostatic Acid Phosphatase on Human Semenogelins and Semenogelin-derived Synthetic Substrates.” *European Journal of Biochemistry*, vol. 269, no. 1, 2002, pp. 390–395.

- Carnahan, Sarah J., and Michael I. Jensen-Seaman. "Hominoid Seminal Protein Evolution and Ancestral Mating Behavior." *American Journal of Primatology*, vol. 70, 2008, pp. 939–948.
- Chabot, Adrien, et al. "Using Reporter Gene Assays to Identify Cis Regulatory Differences between Humans and Chimpanzees." *Genetics*, vol. 176, 2007, pp. 2069–2076.
- Chen, Feng-Chi, and Wen-Hsiung Li. "Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees." *The American Journal of Human Genetics*, vol. 68, no. 2, 2001, pp. 444–456.
- Clauss, Adam, et al. "A Locus on Human Chromosome 20 Contains Several Genes Expressing Protease Inhibitor Domains with Homology to Whey Acidic Protein." *Biochemical Journal*, vol. 368, no. 1, 2002, pp. 233–242.
- Claw, Katrina G. *Proteomic Identification and Evolutionary Analysis of Primate Reproductive Proteins*. University of Washington, 2013.
- Costa Jr, Paul, et al. *Gender Differences in Personality Traits across Cultures: Robust and Surprising Findings*. 2001.
- Czekala, Nancy, and Martha M. Robbins. "12 Assessment of Reproduction and Stress through Hormone Analysis in Gorillas." *Mountain Gorillas: Three Decades of Research at Karisoke*, vol. 27, 2001, p. 317.
- Darwin, Charles. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.
- Darwin, Charles. *The Descent of Man and Selection in Relation to Sex*. 1883.
- Das, Ranajit, et al. "Complete Mitochondrial Genome Sequence of the Eastern Gorilla (*Gorilla beringei*) and Implications for African Ape Biogeography." *Journal of Heredity*, 2014, p. esu056.
- de Lamirande, Eve, et al. "Semenogelin, the Main Protein of Semen Coagulum, Inhibits Human Sperm Capacitation by Interfering with the Superoxide Anion Generated during This Process." *Journal of Andrology*, vol. 22, no. 4, 2001, pp. 672–679.
- de Lamirande, Eve, et al. "Semenogelin, the Main Protein of the Human Semen Coagulum, Regulates Sperm Function." *Seminars in Thrombosis and Hemostasis*, vol. 33, no. 1, 2007, pp. 60–68.
- Dean, Matthew D. "Genetic Disruption of the Copulatory Plug in Mice Leads to Severely Reduced Fertility." *PLoS Genet*, vol. 9, no. 1, Jan. 2013, p. e1003185, doi:10.1371/journal.pgen.1003185.
- DePristo, Mark A., et al. "Missense Meanderings in Sequence Space: A Biophysical View of Protein Evolution." *Nature Reviews Genetics*, vol. 6, no. 9, 2005, pp. 678–687.
- Dixon, AF. "Sexual Selection and Ejaculatory Frequencies in Primates." *Folia Primatologica*, vol. 64, no. 3, 1995, pp. 146–152.
- Dixon, Alan. *Primate Sexuality*. Wiley Online Library, 2012.
- Dixon, Alan, and Matthew Anderson. "Sexual Selection and the Comparative Anatomy of Reproduction in Monkeys, Apes, and Human Beings." *Annual Review of Sex Research*, vol. 12, no. 1, 2001, pp. 121–144.
- Dixon, Alan F. "Evolutionary Perspectives on Primate Mating Systems and Behavior." *Annals New York Academy of Sciences*, vol. 807, 1997, pp. 21–61.
- Dixon, Alan F. *Sexual Selection and the Origins of Human Mating Systems*. Oxford University Press, 2009.
- Dixon, Alan F., and Matthew J. Anderson. "Sexual Behavior, Reproductive Physiology Sperm Competition in Male Mammals." *Physiology and Behavior*, vol. 15, no. 83, 2004, pp. 361–371.
- Dorus, Steve, et al. "Rate of Molecular Evolution of the Seminal Protein Gene *SEMG2* Correlates with Levels of Female Promiscuity." *Nature Genetics*, vol. 36, no. 12, 2004, pp. 1326–1329.
- Dubbink, Hendrikus J., Leon de Waal, et al. "The Human Prostate-Specific Transglutaminase Gene (*TGM4*): Genomic Organization, Tissue-Specific Expression, and Promoter Characterization." *Genomics*, vol. 51, no. 3, 1998, pp. 434–444.

- Dubbink, Hendrikus J., Nicole S. Verkaik, et al. "Tissue-Specific and Androgen-Regulated Expression of Human Prostate-Specific Transglutaminase." *Biochemical Journal*, vol. 315, no. 3, 1996, pp. 901–908.
- Eberhard, William G. "Postcopulatory Sexual Selection: Darwin's Omission and Its Consequences." *PNAS*, vol. 106, 2009, pp. 100025–10032.
- Edström, Anneli ML, et al. "The Major Bactericidal Activity of Human Seminal Plasma Is Zinc-Dependent and Derived from Fragmentation of the Semenogelins." *The Journal of Immunology*, vol. 181, no. 5, 2008, pp. 3413–3421.
- Ely, John J., et al. "Subspecies Composition and Founder Contribution of the Captive US Chimpanzee (*Pan troglodytes*) Population." *American Journal of Primatology*, vol. 67, no. 2, 2005, pp. 223–241.
- Emlen, Douglas J. "The Evolution of Animal Weapons." *Annual Review of Ecology, Evolution, and Systematics*, vol. 39, 2008, pp. 387–413.
- Esposito, Carla, and Ivana Caputo. "Mammalian Transglutaminases." *FEBS Journal*, vol. 272, no. 3, 2004, pp. 615–631.
- Fersht, Alan. "Structure and Mechanism in Protein Sciences: A Guide to Enzyme Catalysis and Protein Folding." *W. H Freeman and Company*, 1999.
- Fisher, Ronald A. "The Evolution of Sexual Preference." *The Eugenics Review*, vol. 7, no. 3, 1915, p. 184.
- Fleagle, John G. *Primate Adaptation and Evolution*. Academic Press, 2013.
- Flori, Federica, et al. "The GPI-anchored CD52 Antigen of the Sperm Surface Interacts with Semenogelin and Participates in Clot Formation and Liquefaction of Human Semen." *Molecular Reproduction and Development*, vol. 75, no. 2, 2008, pp. 326–335.
- Folk, JE. "Mechanism and Basis for Specificity of Transglutaminase-catalyzed-(G-Glutamyl) Lysine Bond Formation." *Adv. Enzymol. Relat. Areas Mol. Biol.*, vol. 54, 1983, pp. 1–56.
- Fontaine, PM, and C. Barrette. "Megatestes: Anatomical Evidence for Sperm Competition in the Harbor Porpoise." *Mammalia*, vol. 61, no. 1, 1997, pp. 65–72.
- Goldfarb, DA, et al. "Age-Related Changes in Tissue Levels of Prostatic Acid Phosphatase and Prostate Specific Antigen." *The Journal of Urology*, vol. 136, no. 6, 1986, pp. 1266–1269.
- Gonder, M.Katherine, et al. "A New West African Chimpanzee Subspecies?" *Nature*, vol. 388, no. 6640, 1997, p. 337.
- Good, Jeffrey M., et al. "Comparative Population Genomics of the Ejaculate in Humans and the Great Apes." *Molecular Biology and Evolution*, vol. 30, no. 4, 2013, pp. 964–976.
- Goodall, Jane. *The Chimpanzees of Gombe: Patterns of Behavior*. 1986.
- Gould, James L., and Carol Grant Gould. *Sexual Selection*. Scientific American Library New York, 1989.
- Grantham, R. "Amino Acid Difference Formula to Help Explain Protein Evolution." *Science*, vol. 185, no. 4154, 1974, pp. 862–864.
- Green, Richard E., et al. "A Draft Sequence of the Neandertal Genome." *Science*, vol. 328, no. 5979, 2010, pp. 710–722.
- Greenberg, Charles S., et al. "Transglutaminases: Multifunctional Cross-Linking Enzymes That Stabilize Tissues." *The FASEB Journal*, vol. 5, no. 15, 1991, pp. 3071–3077.
- Groves, Colin. "The What, Why and How of Primate Taxonomy." *International Journal of Primatology*, vol. 25, no. 5, 2004, pp. 1105–1126.
- Harcourt, AH, et al. "Sperm Competition: Mating System, Not Breeding Season, Affects Testes Size of Primates." *Functional Ecology*, 1995, pp. 468–476.
- Harcourt, Alexander H., KS Stewart, et al. *Male Emigration and Female Transfer in Wild Mountain Gorilla*. 1976.
- Harcourt, Alexander H. "Strategies of Emigration and Transfer by Primates, with Particular Reference to Gorillas." *Ethology*, vol. 48, no. 4, 1978, pp. 401–420.
- Harcourt, Alexander H., Paul H. Harvey, et al. "Testis Weight, Body Weight and Breeding System in Primates." *Nature*, vol. 293, no. 5827, 1981, pp. 55–57.

- Hartl, Daniel L., et al. "Limits of Adaptation: The Evolution of Selective Neutrality." *Genetics*, vol. 111, no. 3, 1985, pp. 655–674.
- Hassan, Md Imtaiyaz, et al. "Structural and Functional Analysis of Human Prostatic Acid Phosphatase." *Expert Review of Anticancer Therapy*, vol. 10, no. 7, 2010, pp. 1055–1068.
- Hosken, DJ. "Sperm Competition in Bats." *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 264, no. 1380, 1997, pp. 385–392.
- Hurle, Belen, et al. "Comparative Sequence Analyses Reveal Rapid and Divergent Evolutionary Changes of the *WFDC* Locus in the Primate Lineage." *Genome Research*, vol. 17, 2007, pp. 276–286.
- Huynen, Martijn A., et al. "Smoothness within Ruggedness: The Role of Neutrality in Adaptation." *Proceedings of the National Academy of Sciences*, vol. 93, no. 1, 1996, pp. 397–401.
- Jensen, Roy A. "Enzyme Recruitment in Evolution of New Function." *Annual Reviews in Microbiology*, vol. 30, no. 1, 1976, pp. 409–425.
- Jensen-Seaman, Michael I., and Wen-Hsiung Li. "Evolution of the Hominoid Semenogelin Genes, the Major Proteins of Ejaculated Semen." *Journal of Molecular Evolution*, vol. 57, 2003, pp. 261–270.
- Jonas, Stefanie, and Florian Hollfelder. "Mechanism and Catalytic Promiscuity: Emerging Mechanistic Principles for Identification and Manipulation of Catalytically Promiscuous Enzymes." *Protein Engineering Handbook, Volume 1 & Volume 2*, 2008, pp. 47–79.
- Jones, Adam G., and Nicholas L. Ratterman. "Mate Choice and Sexual Selection: What Have We Learned since Darwin?" *PNAS*, vol. 106, 2009, pp. 10001–10008.
- Jonsson, Magnus, Sara Linse, et al. "Semenogelins I and II Bind Zinc and Regulate the Activity of Prostate Specific Antigen." *Biochemical Society*, vol. 387, 2005, pp. 447–453.
- Jonsson, Magnus, Åke Lundwall, et al. "Truncated Semenogelin I Binds Zinc and Is Cleaved by Prostate-Specific Antigen." *Journal of Andrology*, vol. 27, no. 4, 2006, pp. 542–547.
- Kappeler, Peter M. "Intrasexual Selection in *Mirza coquereli*: Evidence for Scramble Competition Polygyny in a Solitary Primate." *Behavioral Ecology and Sociobiology*, vol. 41, no. 2, 1997, pp. 115–127.
- Karr, Timothy L., and Scott Pitnick. "Sperm Competition: Defining the Rules of Engagement." *Current Biology*, vol. 9, no. 20, Oct. 1999, pp. R787–R790, doi:10.1016/S0960-9822(00)80014-7.
- Keefe, Anthony D., and Jack W. Szostak. "Functional Proteins from a Random-Sequence Library." *Nature*, vol. 410, no. 6829, 2001, pp. 715–718.
- Kenagy, GJ, and Stephen C. Trombulak. "Size and Function of Mammalian Testes in Relation to Body Size." *Journal of Mammalogy*, 1986, pp. 1–22.
- Khersonsky, Olga, et al. "Enzyme Promiscuity: Evolutionary and Mechanistic Aspects." *Current Opinion in Chemical Biology*, vol. 10, no. 5, 2006, pp. 498–508.
- Khersonsky, Olga, and Dan S. Tawfik. "Structure-Reactivity Studies of Serum Paraoxonase PON1 Suggest That Its Native Activity Is Lactonase." *Biochemistry*, vol. 44, no. 16, 2005, pp. 6371–6382.
- Kim, Juhan, and Shelley D. Copley. "Why Metabolic Enzymes Are Essential or Nonessential for Growth of *Escherichia coli* K12 on Glucose." *Biochemistry*, vol. 46, no. 44, 2007, pp. 12501–12511.
- Kimura, Motoo. "DNA and the Neutral Theory." *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 312, no. 1154, 1986, pp. 343–354.
- King, Jack L., et al. *Non-Darwinian Evolution*. Bobbs-Merrill, 1969.
- King, Mary-Claire, and A. C. Wilson. "Evolution at Two Levels in Humans and Chimpanzees." *Science*, vol. 188, no. 4184, 1975, pp. 107–116.
- Koistinen, Hannu, et al. "Monoclonal Antibodies, Immunofluorometric Assay, and Detection of Human Semenogelin in Male Reproductive Tract: No Association with in Vitro Fertilizing Capacity of Sperm." *Biology of Reproduction*, vol. 66, no. 3, 2002, pp. 624–628.
- Kong, Hoon Young, and Jonghoe Byun. "Emerging Roles of Human Prostatic Acid Phosphatase." *Biomolecules and Therapeutics*, vol. 21, no. 1, 2013, pp. 10–20.

- Kosiol, Carolin, et al. "Patterns of Positive Selection in Six Mammalian Genes." *PLoS Genetics*, vol. 4, no. 8, 2008.
- Kuciel, R., et al. "Is the Subunit of Prostatic Phosphatase Active? Reversible Denaturation of Prostatic Acid Phosphatase." *Biochemistry International*, vol. 22, no. 2, 1990, pp. 329–334.
- Kuciel, Radoslaw, et al. "The Folding Intermediate of Reversibly Denatured Human Prostatic Acid Phosphatase." *International Journal of Biological Macromolecules*, vol. 18, no. 3, 1996, pp. 167–175.
- Kuhle, Barry X. "An Evolutionary Perspective on the Origin and Ontogeny of Menopause." *Maturitas*, vol. 57, 2007, pp. 329–337.
- Kumar, Anand, et al. "Duration of Hypotension before Initiation of Effective Antimicrobial Therapy Is the Critical Determinant of Survival in Human Septic Shock." *Critical Care Medicine*, vol. 34, no. 6, 2006, pp. 1589–1596.
- Lad, Pramod M., et al. "Distribution of Prostatic Acid Phosphatase Isoenzymes in Normal and Cancerous States." *Clinica Chimica Acta*, vol. 141, no. 1, 1984, pp. 51–65.
- Lee, H., et al. "Homodimer and Heterodimer Subunits of Human Prostate Acid Phosphatase." *Biochemical Journal*, vol. 277, no. 3, 1991, pp. 759–765.
- Li, Steven S. L., and Farida S. Sharief. "The Prostatic Acid Phosphatase (ACPP) Gene Is Localized to Human Chromosome 3q21-q23." *Genomics*, vol. 17, no. 3, 1993, pp. 765–766.
- Lilja, H. "A Kallikrein-like Serine Protease in Prostatic Fluid Cleaves the Predominant Seminal Vesicle Protein." *Journal of Clinical Investigation*, vol. 76, no. 5, 1985, p. 1899.
- Lilja, Hans, Per-Anders Abrahamsson, et al. "Semenogelin, the Predominant Protein in Human Semen. Primary Structure and Identification of Closely Related Proteins in the Male Accessory Sex Glands and on the Spermatozoa." *J Biol Chem*, vol. 264, 1989, pp. 1894–1900.
- Lilja, Hans, Jorgen Oldbring, et al. "Seminal-Vesicle Secreted Proteins and Their Reactions during Gelation and Liquefaction of Human Semen." *J. Clin. Invest.*, vol. 80, 1987, pp. 281–285.
- Lilja, Hans, and Ake Lundwall. "Molecular Cloning of Epididymal and Seminal Vesicular Transcripts Encoding a Semenogelin-Related Protein." *Proceedings of the National Academy of Sciences*, vol. 89, no. 10, 1992, pp. 4559–4563.
- Lin, Ming Fong, et al. "Purification and Characterization of a New Human Prostatic Acid Phosphatase Isoenzyme." *Biochemistry*, vol. 22, no. 5, 1983, pp. 1055–1062.
- Loeb, Daniel D., et al. "Complete Mutagenesis of the HIV-1 Protease." *Nature*, vol. 340, no. 6232, 1989, pp. 397–400.
- Loeb, Stacy, and William J. Catalona. "Prostate-Specific Antigen in Clinical Practice." *Cancer Letters*, vol. 249, no. 1, 2007, pp. 30–39.
- Lövgren, Janita, et al. "Enzymatic Action of Human Glandular Kallikrein 2 (hK2)." *European Journal of Biochemistry*, vol. 262, no. 3, 1999, pp. 781–789.
- Loyau, Adeline, et al. "Multiple Sexual Advertisements Honestly Reflect Health Status in Peacocks (*Pavo Cristatus*)." *Behavioral Ecology and Sociobiology*, vol. 58, no. 6, 2005, pp. 552–557.
- Lundwall, Å., and M. Brattsand. "Kallikrein-Related Peptidases." *Cell Mol Life Sci*, vol. 65, 2008, pp. 2019–2038.
- Lundwall, Åke, et al. "Semenogelin I and II, the Predominant Human Seminal Plasma Proteins, Are Also Expressed in Non-Genital Tissues." *Mol Human Reproduction*, vol. 8, 2002, pp. 805–810.
- Lundwall, Åke, et al. "The Cloning of a Rapidly Evolving Seminal-vesicle-transcribed Gene Encoding the Major Clot-forming Protein of Mouse Semen." *European Journal of Biochemistry*, vol. 235, no. 1-2, 1996, pp. 424–430.
- Lundwall, Åke, and Claude Lazure. "A Novel Gene Family Encoding Proteins with Highly Differing Structure because of a Rapidly Evolving Exon." *FEBS Letters*, vol. 374, no. 1, 1995, pp. 53–56.

- Lundwall, Åke, and A.Yvonne Olsson. "Semenogelin II Gene Is Replaced by a Truncated LINE1 Repeat in the Cotton-Top Tamarin." *Biology of Reproduction*, vol. 65, 2001, pp. 420–425.
- Malm, Johan, Jukka Hellman, et al. "Isolation and Characterization of the Major Gel Proteins in Human Semen, Semenogelin I and Semenogelin II." *Eur. J. Biochem*, vol. 238, 1996, pp. 48–53.
- Malm, Johan, Magnus Jonsson, et al. "Structural Properties of Semenogelin I." *FEBS Journal*, vol. 274, 2007, pp. 4503–4510.
- Mann, Thaddeus, and Cecilia Lutwak-Mann. *Male Reproductive Function and Semen: Themes and Trends in Physiology, Biochemistry and Investigative Andrology*. Springer Science & Business Media, 2012.
- Marlowe, F. "Paternal Investment and the Human Mating System." *Behavioural Processes*, vol. 51, no. 1–3, Oct. 2000, pp. 45–61, doi:10.1016/S0376-6357(00)00118-2.
- Marques, Patrícia Isabel, et al. "Birth and Death of *KLK3* and *KLK2* in Primates: Evolution Driven by Reproductive Biology?" *Genome Biology and Evolution*, vol. Advanced Access, 2012.
- Maston, Glenn A., et al. "Transcriptional Regulatory Elements in the Human Genome." *Annu. Rev. Genomics Hum. Genet.*, vol. 7, 2006, pp. 29–59.
- Matthews, Brian W. "Studies on Protein Stability with T4 Lysozyme." *Advances in Protein Chemistry*, vol. 46, 1995, pp. 249–278.
- McLean, Cory Y., et al. "Human-Specific Loss of Regulatory DNA and the Evolution of Human-Specific Traits." *Nature*, vol. 471, 2011, pp. 216–219.
- Miller, Mark P., and Sudhir Kumar. "Understanding Human Disease Mutations through the Use of Interspecific Genetic Variation." *Human Molecular Genetics*, vol. 10, no. 21, 2001, pp. 2319–2328.
- Miller-Schroeder, Patricia. *Gorillas*. Raintree, 1997.
- Mitchell, MW, and MK Gonder. "Primate Speciation: A Case Study of African Apes." *Nature Education Knowledge*, vol. 4, no. 2, 2013, p. 1.
- Mitra, Anurag, et al. "Analysis of Recombinant Human Semenogelin as an Inhibitor of Human Sperm Motility." *Biology of Reproduction*, vol. 82, 2010, pp. 489–496.
- Møller, Anders Pape. "Ejaculate Quality, Testes Size and Sperm Competition in Primates." *Journal of Human Evolution*, vol. 17, no. 5, Aug. 1988, pp. 479–488, doi:10.1016/0047-2484(88)90037-1.
- Møller, Anders Pape, and Marion Petrie. "Condition Dependence, Multiple Sexual Signals, and Immunocompetence in Peacocks." *Behavioral Ecology*, vol. 13, no. 2, 2002, pp. 248–253.
- Nascimento, Jaclyn M., et al. "The Use of Optical Tweezers to Study Sperm Competition and Motility in Primates." *Journal of The Royal Society Interface*, vol. 5, no. 20, Mar. 2008, pp. 297–302, doi:10.1098/rsif.2007.1118.
- Newman, Russell W. "Why Man Is Such a Sweaty and Thirsty Naked Animal: A Speculative Review." *Human Biology*, 1970, pp. 12–27.
- Ng, Pauline C., and Steven Henikoff. "Predicting the Effects of Amino Acid Substitutions on Protein Function." *Annu. Rev. Genomics Hum. Genet.*, vol. 7, 2006, pp. 61–80.
- Nishida, Toshisada, and Kenji Kawanaka. "Within-Group Cannibalism by Adult Male Chimpanzees." *Primates*, vol. 26, no. 3, 1985, pp. 274–284.
- Nowak, Ronald M. "Order Artiodactyla." *Walker's Mammals of the World*. John Hopkins University Press, London, 1999.
- O'Brien, Patrick J., and Daniel Herschlag. "Catalytic Promiscuity and the Evolution of New Enzymatic Activities." *Chemistry & Biology*, vol. 6, no. 4, 1999, pp. R91–R105.
- Oliva, Rafael, et al. "Proteomics in the Study of the Sperm Cell Composition, Differentiation and Function." *Systems Biology in Reproductive Medicine*, vol. 54, no. 1, 2008, pp. 23–36.
- Olsson, A.Yvonne, et al. "The Evolution of the Glandular Kallikrein Locus: Identification of Orthologs and Pseudogenes in the Cotton-Top Tamarin." *Gene*, vol. 343, no. 2, 2004, pp. 347–355.
- Opazo, Juan C., et al. "Adaptive Evolution of the Insulin Gene in Caviomorph Rodents." *Molecular Biology and Evolution*, vol. 22, no. 5, 2005, pp. 1290–1298.

- O’Rand, Michael G., et al. “Functional Studies of Eppin.” *Biochemical Society Transactions*, vol. 39, 2011, pp. 1447–1449.
- Ortlund, Eric, et al. “Crystal Structure of Human Complement Protein C8 γ at 1.2 Å Resolution Reveals a Lipocalin Fold and a Distinct Ligand Binding Site†.” *Biochemistry*, vol. 41, no. 22, 2002, pp. 7030–7037.
- Owen, Derek H., and David F. Katz. “A Review of the Physical and Chemical Properties of Human Semen and the Formulation of a Semen Simulant.” *Journal of Andrology*, vol. 26, no. 4, 2005, p. 459.
- Özer, Ali, et al. “Sociodemographic Variables and Depression in Turkish Women from Polygamous versus Monogamous Families.” *Health Care for Women International*, vol. 34, no. 11, 2013, pp. 1024–1034.
- Pagel, Mark, and Walter Bodmer. “A Naked Ape Would Have Fewer Parasites.” *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 270, no. Suppl 1, 2003, pp. S117–S119.
- Pakula, Andrew A., et al. “Bacteriophage Lambda Cro Mutations: Effects on Activity and Intracellular Degradation.” *Proceedings of the National Academy of Sciences*, vol. 83, no. 23, 1986, pp. 8829–8833.
- Pampalakis, Georgios, and Georgia Sotiropoulou. “Tissue Kallikrein Proteolytic Cascade Pathways in Normal Physiology and Cancer.” *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1776, no. 1, 2007, pp. 22–31.
- Parker, G. A. “Sperm Competition and Its Evolutionary Consequences in the Insects.” *Biological Reviews*, vol. 45, no. 4, 1970, pp. 525–567, doi:10.1111/j.1469-185X.1970.tb01176.x.
- Peter, Anders, et al. “Semenogelin I and Semenogelin II, the Major Gel-Forming Proteins in Human Semen, Are Substrates for Transglutaminase.” *Eur. J. Biochem*, vol. 252, 1998, pp. 216–221.
- Pomiankowski, Andrew, and Yoh Iwasa. “Runaway Ornament Diversity Caused by Fisherian Sexual Selection.” *Proceedings of the National Academy of Sciences*, vol. 95, no. 9, 1998, pp. 5106–5111.
- Poon, Art, and Lin Chao. “The Rate of Compensatory Mutation in the DNA Bacteriophage ϕ X174.” *Genetics*, vol. 170, no. 3, 2005, pp. 989–999.
- Potts, Richard. “Hominin Evolution in Settings of Strong Environmental Variability.” *Quaternary Science Reviews*, vol. 73, 2013, pp. 1–13.
- Proudfoot, Nick J., et al. “Integrating mRNA Processing with Transcription.” *Cell*, vol. 108, no. 4, 2002, pp. 501–512.
- Ramm, Steven A., Lucy McDonald, et al. “Comparative Proteomics Reveals Evidence for Evolutionary Diversification of Rodent Seminal Fluid and Its Functional Significance in Sperm Competition.” *Mol. Biol. Evol.*, vol. 26, no. 1, 2009, pp. 189–198.
- Ramm, Steven A., Sarah A. Cheetham, et al. “Encoding Choosiness: Female Attraction Requires Prior Physical Contact with Individual Male Scents in Mice.” *Proc. R. Soc. B*, vol. 275, 2008, pp. 1727–1735.
- Ramm, Steven A., Peter I. Oliver, et al. “Sexual Selection and the Adaptive Evolution of Mammalian Ejaculate Proteins.” *Mol. Biol. Evol.*, vol. 25, 2008, pp. 207–219.
- Reich, David, et al. “Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia.” *Nature*, vol. 468, no. 7327, 2010, pp. 1053–1060.
- Rivera-Gonzalez, Guillermo C., et al. “Retinoic Acid and Androgen Receptors Combine to Achieve Tissue Specific Control of Human Prostatic Transglutaminase Expression: A Novel Regulatory Network with Broader Significance.” *Nucleic Acids Research*, 2012, p. gks143.
- Roan, Nadia R., Haichuan Liu, et al. “Liquefaction of Semen Generates and Later Degrades a Conserved Semenogelin Peptide That Enhances HIV Infection.” *Journal of Virology*, vol. 88, no. 13, 2014, pp. 7221–7234.
- Roan, Nadia R., Janis A. Müller, et al. “Peptides Released by Physiological Cleavage of Semen Coagulum Proteins Form Amyloids That Enhance HIV Infection.” *Cell Host & Microbe*, vol. 10, no. 6, 2011, pp. 541–550.

- Robbins, Andrew M., and Martha M. Robbins. "Fitness Consequences of Dispersal Decisions for Male Mountain Gorillas (*Gorilla Beringei Beringei*)." *Behavioral Ecology and Sociobiology*, vol. 58, no. 3, 2005, pp. 295–309.
- Robbins, Martha M., et al. "Social Structure and Life-history Patterns in Western Gorillas (*Gorilla gorilla gorilla*)." *American Journal of Primatology*, vol. 64, no. 2, 2004, pp. 145–159.
- Robert, M., and C. Gagnon. "Sperm Motility Inhibitor from Human Seminal Plasma: Association with Semen Coagulum." *Molecular Human Reproduction*, vol. 1, no. 6, 1995, pp. 292–297.
- Robert, Martin, et al. "Characterization of Prostate-Specific Antigen Proteolytic Activity on Its Major Physiological Substrate, the Sperm Motility Inhibitor/Precursor/Semenogelin I." *Biochemistry*, vol. 36, no. 13, 1997, pp. 3811–3819.
- Robert, Martin, and Claude Gagnon. "Semenogelin I: A Coagulum Forming, multifunctional Seminal Vesicle Protein." *Cell Mol Life Sci*, vol. 55, 1999, pp. 944–960.
- Robertson, Sarah A. "Seminal Plasma and Male Factor Signalling in the Female Reproductive Tract." *Cell Tissue Research*, vol. 322, 2005, pp. 43–52.
- Robertson, Sarah A. "Transforming Growth Factor Beta—a Mediator of Immune Deviation in Seminal Plasma." *Journal of Reproductive Immunology*, vol. 57, no. 1–2, 2002, pp. 109–128.
- Robson, Shannen L., and Bernard Wood. "Hominin Life History: Reconstruction and Evolution." *Journal of Anatomy*, vol. 212, no. 4, 2008, pp. 394–425.
- Romero, Philip A., and Frances H. Arnold. "Exploring Protein Fitness Landscapes by Directed Evolution." *Nature Reviews Molecular Cell Biology*, vol. 10, no. 12, 2009, pp. 866–876.
- Rönneberg, L., et al. "Clomiphene Citrate Administration to Normogonadotropic Subfertile Men: Blood Hormone Changes and Activation of Acid Phosphatase in Seminal Fluid." *International Journal of Andrology*, vol. 4, no. 1-6, 1981, pp. 372–378.
- Rose, RW, et al. "Testes Weight, Body Weight and Mating Systems in Marsupials and Monotremes." *Journal of Zoology*, vol. 243, no. 3, 1997, pp. 523–531.
- Rowe, Noel. *Pictorial Guide to the Living Primates*. Pogonias Press, 1996.
- Rutherford, Suzanne L. "Between Genotype and Phenotype: Protein Chaperones and Evolvability." *Nature Reviews Genetics*, vol. 4, no. 4, 2003, pp. 263–274.
- Sánchez-Moreno, Israel, et al. "From Kinase to Cyclase: An Unusual Example of Catalytic Promiscuity Modulated by Metal Switching." *ChemBioChem*, vol. 10, no. 2, 2009, pp. 225–229.
- Sangster, Todd A., et al. "Under Cover: Causes, Effects and Implications of Hsp90-mediated Genetic Capacitance." *Bioessays*, vol. 26, no. 4, 2004, pp. 348–362.
- Sauna, Zuben E., and Chava Kimchi-Sarfaty. "Understanding the Contribution of Synonymous Mutations to Human Disease." *Nature Reviews Genetics*, vol. 12, no. 10, 2011, pp. 683–691.
- Saunders, Christopher T., and David Baker. "Evaluation of Structural and Evolutionary Contributions to Deleterious Mutation Prediction." *Journal of Molecular Biology*, vol. 322, no. 4, 2002, pp. 891–901.
- Schieferstein, G. "Prostate-Specific Antigen (PSA) in Human Seminal Plasma." *Archives of Andrology*, vol. 42, no. 3, 1999, pp. 193–197.
- Seager, S. W. J., et al. "Semen Collection and Evaluation in *Gorilla gorilla gorilla*." *American Journal of Primatology*, vol. 3, no. S1, 1982, pp. 13–13, doi:10.1002/ajp.1350030506.
- Serrano, Manuel, et al. "A New Regulatory Motif in Cell-Cycle Control Causing Specific Inhibition of Cyclin D/CDK4." *Nature*, vol. 366, no. 6456, 1993, pp. 704–707.
- Shan, Jingdong, et al. "Tissue-Specific Expression of the Prostatic Acid Phosphatase Promoter Constructs." *Biochemical and Biophysical Research Communications*, vol. 311, no. 4, 2003, pp. 864–869.
- Shimotohno, Akie, et al. "Demonstration of the Importance and Usefulness of Manipulating Non-Active-Site Residues in Protein Design." *Journal of Biochemistry*, vol. 129, no. 6, 2001, pp. 943–948.

- Short, Roger V. "Sexual Selection and Its Component Parts, Somatic and Genital Selection, as Illustrated by Man and the Great Apes." *Advances in the Study of Behavior*, vol. 9, 1979, pp. 131–158.
- Shortle, David, and Beth Lin. "Genetic Analysis of Staphylococcal Nuclease: Identification of Three Intragenic 'global' suppressors of Nuclease-Minus Mutations." *Genetics*, vol. 110, no. 4, 1985, pp. 539–555.
- Smith, John Maynard. *Natural Selection and the Concept of a Protein Space*. 1970.
- Solin, Timo, et al. "Gene Expression and Prostate Specificity of Human Prostatic Acid Phosphatase (PAP): Evaluation by RNA Blot Analyses." *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, vol. 1048, no. 1, 1990, pp. 72–77.
- Spiller, Ben, et al. "A Structural View of Evolutionary Divergence." *Proceedings of the National Academy of Sciences*, vol. 96, no. 22, 1999, pp. 12305–12310.
- Taggart, D., et al. *Reproduction, Mating Strategies and Sperm Competition in Marsupials and Monotremes*. Academic Press, Harcourt Brace and Company, Publishers, 1998.
- Taverna, Darin M., and Richard A. Goldstein. "Why Are Proteins Marginally Stable?" *Proteins: Structure, Function, and Bioinformatics*, vol. 46, no. 1, 2002, pp. 105–109.
- Tawfik, Olga Khersonsky and Dan S. "Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective." *Annual Review of Biochemistry*, vol. 79, 2010, pp. 471–505.
- Tiana, G., et al. "Hiking in the Energy Landscape in Sequence Space: A Bumpy Road to Good Folders." *Proteins: Structure, Function, and Bioinformatics*, vol. 39, no. 3, 2000, pp. 244–251.
- Ulvsbäck, M., et al. "Gene Structure of Semenogelin I and II. The Predominant Proteins in Human Semen Are Encoded by Two Homologous Genes on Chromosome 20." *Journal of Biological Chemistry*, vol. 267, no. 25, 1992, pp. 18080–18084.
- Ulvsbäck, Magnus, and Åke Lundwall. "Cloning of the Semenogelin II Gene of the Rhesus Monkey Duplications of 360 Bp Extend the Coding Reaching in Man, Rhesus Monkey and Baboon." *Eur. J. Biochem*, vol. 248, 1997, pp. 25–31.
- Vigilant, Linda, et al. "Paternity and Relatedness in Wild Chimpanzee Communities." *Proceedings of the National Academy of Sciences*, vol. 98, no. 23, 2001, pp. 12890–12895.
- Vihko, Pirkko, et al. "Purification of Human Prostatic Acid Phosphatase by Affinity Chromatography and Isoelectric Focusing. Part I." *Clinical Chemistry*, vol. 24, no. 3, 1978, pp. 466–470.
- Villiers, Benoit RM, and Florian Hollfelder. "Mapping the Limits of Substrate Specificity of the Adenylation Domain of TycA." *ChemBioChem*, vol. 10, no. 4, 2009, pp. 671–682.
- Wang, Susan C., et al. "The 4-Oxalocrotonate Tautomerase-and Ywhb-Catalyzed Hydration of 3 E-Haloacrylates: Implications for the Evolution of New Enzymatic Activities." *Journal of the American Chemical Society*, vol. 125, no. 47, 2003, pp. 14282–14283.
- Wang, Xiaojun, et al. "Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-Offs." *Journal of Molecular Biology*, vol. 320, no. 1, 2002, pp. 85–95.
- Wang, Zhen, and John Mould. "SNPs, Protein Structure, and Disease." *Human Mutation*, vol. 17, no. 4, 2001, pp. 263–270.
- Watts, David P. "Ecology of Gorillas and Its Relation to Female Transfer in Mountain Gorillas." *International Journal of Primatology*, vol. 11, no. 1, Feb. 1990, pp. 21–45, doi:10.1007/BF02193694.
- White, Frances J. "Comparative Socio-Ecology of *Pan paniscus*." *Great Ape Societies*, edited by WC McGrew et al., Cambridge University Press, 1996, pp. 29–41, <http://dx.doi.org/10.1017/CBO9780511752414.005>.
- Whiten, Andrew, et al. "Cultures in Chimpanzees." *Nature*, vol. 399, no. 6737, 1999, pp. 682–685.
- Winqvist, R., et al. "Chromosomal Localization to 3q21→ Qter and Two *TaqI* RFLPs of the Human Prostate-Specific Acid Phosphatase Gene (ACPP)." *Cytogenetic and Genome Research*, vol. 52, no. 1–2, 1989, pp. 68–71.

- Wolfenden, Richard. "Degrees of Difficulty of Water-Consuming Reactions in the Absence of Enzymes." *Chemical Reviews*, vol. 106, no. 8, 2006, p. 3379.
- Yoshida, Karou, et al. "Quantification of Seminal Plasma Motility Inhibitor/Semenogelin in Human Seminal Plasma." *Journal of Andrology*, vol. 24, no. 6, 2003, pp. 878–884.
- Zalensky, Andrei O., et al. "The Abundant 19-kilodalton Protein Associated with Human Sperm Nuclei That Is Related to Seminal Plasma A-inhibins." *Molecular Reproduction and Development*, vol. 36, no. 2, 1993, pp. 164–173.
- Zelivianski, Stanislav, et al. "Cloning and Analysis of the Promoter Activity of the Human Prostatic Acid Phosphatase Gene." *Biochemical and Biophysical Research Communications*, vol. 245, no. 1, 1998, pp. 108–112.
- Zhang, Jianzhi. "Evolution by Gene Duplication: An Update." *Trends in Ecology & Evolution*, vol. 18, no. 6, 2003, pp. 292–298.

CHAPTER 2: Proteomic composition of seminal plasma proteins in gorilla, human, and chimpanzee

2.1 Introduction

Many reproductive, immunological, and sensory genes appear to have evolved under positive selection in mammalian lineages (Kosiol et al., 2008). Specifically, male reproductive proteins are claimed to evolve rapidly in primates and other taxa (Torgerson, 2002; Jensen-Seaman and Li, 2003; Dorus et al., 2004; Clark and Swanson, 2005; Panhuis, 2006, Haerty, 2007; Carnahan and Jensen-Seaman, 2008). This rapid evolution has been attributed to forces including sperm competition, pathogen resistance, sexual conflict, and heterospecific avoidance (Swanson and Vacquier, 2002). Though many seminal proteins have shown this pattern in single gene studies, in large-scale computational studies, sexual selective pressures do not appear to affect the overall evolution of extracellular reproductive proteins (Dean et al., 2009; Findlay and Swanson, 2010; Wong, 2011). Rather sexual selection exerts strong pressure on proteins that are either tissue specific or have functions related to immunity and peptidase activity (Good et al., 2013, Carnahan-Craig and Jensen-Seaman, 2014).

2.1.1 Primate mating systems and sperm competition

Extant hominids are closely related but differ dramatically in social structure and mating systems. Gorilla (*Gorilla gorilla*) has a polygynous mating system where a female usually mates with a single dominant male during estrus (Watts, 1990). Chimpanzee (*Pan troglodytes*) has multi-male/multi-female mating systems, where a female will mate with multiple males in rapid succession during estrus (White, 1996). The human (*Homo sapiens*) mating system is a challenge to classify due to societal differences, but most humans are monogamous or polygynous, where

monogamy is defined as the tendency to mate with a single male during any given cycle (Marlowe, 2000).

Although these are generalizations, mating systems with higher female promiscuity are projected to have higher sperm competition (Karr and Picnick, 1999; Dixson and Anderson, 2001; Birkhead and Pizzari, 2002). Primate species that experience high sperm competition have been shown to have larger testes to body size ratio (Harcourt et al., 1981); similarly, seminal vesicles are larger in species with higher sperm competition (Dixson, 1997). Concentration of motile sperm (Møller, 1988), volume of sperm midpiece (Dixson and Anderson, 2004), and movement of sperm (Nascimento et al., 2008) increases with higher sperm competition. Chimpanzee and closely related bonobo (*Pan paniscus*) are the only hominoids to form a copulatory plug and produce a relatively large volume of ejaculate (Dixson, 1997; Dorus et al., 2004; Dixson and Anderson, 2004). Gorilla sperm competition is likely to be low (Dixson, 1997), and they have relatively small ejaculate volume (Seager et al., 1982; Dixson, 1997). Different levels of sperm competition lead to varying levels of positive, purifying, or neutral evolution on reproductive genes. Genes involved in semen coagulation have been shown to be positively selected for in species with higher sperm competition (Dixson, 1997; Jensen-Seaman and Li, 2003; Dorus et al., 2004; Carnahan and Jensen-Seaman, 2008).

2.1.2 Shotgun proteomics and seminal plasma proteins

As DNA sequencing becomes more efficient and economical, the increased availability of complete genome sequences of species has accelerated research in multiple fields, especially proteomics. Shotgun proteomics, a term created by Yates (1998), is a method of analyzing a complex mixture of proteins through liquid chromatography of proteolytically digested proteins. Shotgun proteomics has been used to identify proteins from reproductive fluids, including but

not limited to, seminal fluid, prostatic secretions, and follicular fluid (Pilch and Mann, 2006; Drake et al., 2010; Ambekar et al., 2013). Within primates, human seminal fluid is the only primate in which the proteome has been comprehensively characterized (Pilch and Mann, 2006; Fung et al., 2004). Proteomic studies show that seminal plasma proteins vary between individuals in humans and rodents (Martinez-Heredia et al., 2008; Ramm et al., 2008); however, there is a need for proteomic studies assessing differences between species.

Currently, there has only been one other study conducted on non-human primate seminal fluid, and as of yet, it has only been written in a dissertation (Claw, 2013). They reported that the most abundant proteins, across multiple primate species, were involved in semen coagulation and liquefaction. These proteins were SEMG1, SEMG2, TGM4, KLK3, and ACPP (all were discussed extensively in section 1.4.2); at least one of these proteins was within the top five most abundant proteins in each species, despite the associated mating system (Claw, 2013). As mentioned before, genes encoding proteins involved in semen coagulation have been shown to be positively selected for in primate species with increased female promiscuity (Dixson, 1997; Jensen-Seaman and Li, 2003; Dorus et al., 2004; Carnahan and Jensen-Seaman, 2008).

2.2 Methods

This methods section is broken into five subcategories: 1) Seminal plasma sample preparation, 2) SDS-PAGE optimization, 3) Western blot optimization, 4) Quality control checks, and 5) Methods for comparative proteomics among human, chimpanzee, and gorilla seminal plasma. Subsections 2.2.1.1. and 2.2.5 provide descriptions of optimized methods utilized to address my experimental goals.

2.2.1 Seminal plasma sample preparation

2.2.1.1 Sample preparation

Semen samples were provided by three healthy non-vasectomized human volunteers (Lee BioSystems, St. Louis, MO) and from three healthy non-vasectomized chimpanzees trained to voluntarily provide samples (ID#95A016, age 16; ID#99A003, age 12; and ID#95A018, age 16), housed at the New Iberia Research Center (New Iberia, LA). Semen was obtained from three gorillas (Studbook IDs 835, 883, and 612), via electrojaculation under sedation. Individuals 835 and 883 are brothers. Human samples were obtained under approval from the Institutional Review Board of Lee BioSystems. Chimpanzee and gorilla samples were obtained under Institutional Animal Care and Use Committee approval of the University of Louisiana at Lafayette and the Henry Doorly Zoo (Omaha, NE), respectively. Samples were stored at -80°C until analysis. In the case of human and gorilla samples, an aliquot of each individual was centrifuged (13,000xg, 5 min, 4° C) to pellet sperm cells and the supernatant was used for further analysis. For chimpanzee, the coagulated semen samples were first allowed to partially liquefy at 37°C for 3 hours. From the resulting liquid fraction of each individual an aliquot was centrifuged (13,000xg, 5 min, 4° C), and the supernatant was used for further analysis.

2.2.1.2 Protein quantification

Protein concentration was determined using either the Qubit® Protein Assay Kit (ThermoFisher Scientific, Waltham, MA) or the Bradford protein assay (Bio-Rad, Hercules, CA). For the Qubit® Protein Assay, seminal samples were diluted 1:20 or 1:200 before assaying. Either 1, 5, or 10µl of diluted sample was added to Qubit® protein assay buffer (1µl reagent in 199µl buffer) with the total volume being 200µl. Samples were vortexed, incubated for 5 minutes, and then the fluorescence was measured with the Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA). This kit used three standards to generate a standard curve and calculated sample concentration in relation to fluorescence. In the Bradford assay, seminal samples were undiluted or diluted 1:20 before assaying. Then 20µl of sample was added to 1mL of Bio-Rad Bradford assay dye, mixed, and incubated for 5 minutes. Absorbance at 595nm was measured using the Genesys 10 UV spectrophotometer (Thermo Spectronic, Rochester, NY). Utilizing 7 BSA standards, a standard curve was manually generated relating absorbance to mg/mL concentration. Sample concentration was calculated by solving for “x” using the standard slope equation generated by the standards. For comparison purposes, 1-10 µl of the 7 BSA standards were cross-assayed using the Qubit® protein assay; likewise 20 µl of the 3 Qubit protein standards were cross-assayed using the Bradford assay.

2.2.2 SDS-PAGE optimization

Chimpanzee ‘Little Joe’ seminal plasma was used initially for SDS-PAGE optimization of seminal fluid. A volume of seminal plasma containing 10µg of protein (1:20 diluted and undiluted) was mixed 1:1 with Laemmli buffer (Bio-Rad) and 5% β-mercaptoethanol; and was incubated at 95°C for 10 minutes. Samples were loaded onto a 10% SDS-PAGE gel (Invitrogen) and run at 200 volts for 20 minutes. The gel was stained with coomassie stain (0.05% coomassie

R-250, 50% methanol, 40% dH₂O, and 10% glacial acetic acid) for thirty minutes and destained overnight. A pooled human semen sample (donor number: T3402; named 'Human #4') was ordered from Lee BioSystems, St. Louis, MO. Varying concentrations (5µg, 10µg, and 15µg) of Human #4 seminal plasma were separated on a 10% SDS-PAGE gel and stained as described above.

On several occasions, commercial gels were compared to poured 10% gels, as well as a poured 6-10% stacking gel. Either 10µg or 20µg of seminal plasma (human or chimpanzee) was mixed 1:1 with Laemmli buffer (Bio-Rad) and 10% β-mercaptoethanol and was incubated at 95°C for 10 minutes. Samples were loaded into either a poured or commercial gel and ran at 110 volts for 15 minutes, and then 220 volts for thirty minutes. Gels were stained and destained as described above. However, after 2014, a commercial coomassie (Li-COR, Lincoln, NE) stain was used instead, followed with an overnight destain with dH₂O, and imaged on the FC odyssey imager (Li-COR).

2.2.3 Western blot optimization

2.2.3.1 Abnova primary antibodies for highly abundant seminal proteins

Initial Westerns were performed with a polyclonal mouse anti-SEMG1 antibody because SEMG1 and SEMG2 are the most abundant proteins in human semen (Lilja et al., 1989). A volume of human seminal plasma containing 5, 10, or 15µg of protein was separated on a 10% SDS-PAGE (Invitrogen), as described above. An Immuno-Blot® PVDF (Bio-Rad) membrane was activated in methanol for one minute, and incubated in chilled Towbin Transfer Buffer (25mM Tris, 192 M glycine, and 20% methanol, with a pH 8.3) with the separated gel and other cassette components for 10 minutes. The Western sandwich was assembled and put into the mini Trans-Blot® (Bio-Rad) chamber with an ice pack. The apparatus was placed on a stir plate with

agitation, and proteins were transferred at 95 volts for 75 minutes. After transfer several control checks were made. The SDS gel was stained and destained, to check if any proteins remained. Before continuing, the PVDF membrane was incubated in a 0.5% Ponceau S (Sigma Aldrich, St. Louis, MO) and 1% glacial acetic acid solution for one minute and rinsed with dH₂O to observe transferred protein. The PVDF membrane was incubated in dH₂O until all of the Ponceau S dye dissipated. After the control checks, the PVDF membrane was incubated in SuperBlock PBS (Thermofisher Scientific) for 30 to 90 minutes. Mouse anti-SEMG1 (Abnova, Jhongli, Taiwan) was diluted (1:5,000) in SuperBlock PBS and placed on the PVDF membrane and incubated overnight at room temperature. The blot was washed three times with PBS for 15 minutes. Then a 1:12,000 dilution of Immunopure goat anti-mouse IgG peroxidase conjugated secondary antibody (Thermofisher Scientific) in SuperBlock PBS was added to the PVDF membrane and incubated for one hour at room temperature. Afterwards the blot was washed three times with PBS for fifteen minutes, and detected with the Pierce ® ECL Western blotting substrate kit (Thermofisher Scientific) using the typhoon 8600 (Molecular Dynamics, Sunnyvale, CA).

Human albumin, an abundant protein in bodily serum (Pilch and Mann, 2006) was Western blotted. A volume of human seminal plasma containing 5, 10, or 15µg of protein was separated on a 10% SDS-PAGE (Invitrogen), as described above. Proteins were transferred to an Immuno-Blot® PVDF (Bio-Rad) and blocked in SuperBlock PBS (Thermofisher Scientific) as described above, except transfer occurred at 100 volts for 75 minutes. Monoclonal mouse anti-Albumin (Abnova) was diluted (1:5,000) in SuperBlock PBS, dispensed over the PVDF membrane and incubated overnight at room temperature. Three different secondary antibodies and detection methods were tested, including Immunopure goat anti-mouse IgG peroxidase conjugated antibody and Piece ® ECL detection as described above. In addition, peroxidase

conjugated antibody incubation was combined with SIGMAFAST™ 3, 3'- Diaminobenzidine (Sigma Aldrich) tablet detection following the product's protocol. A different secondary antibody, goat anti-mouse IgG alkaline phosphatase conjugated antibody (ThermoFisher Scientific), was diluted 1:1,000 with SuperBlock PBS and covered the transferred membrane for one hour at room temperature. Following normal PBS washes, the PVDF membrane was washed with 0.05M Tris-HCL, pH 8, for ten minutes at room temperature. FastRed detection solution was made immediately before use and was made with 0.01g Naphthol AS-MX phosphate and 0.02g FastRed (Sigma Aldrich) in 10mL of 0.05M Tris-HCL, pH 8, solution. FastRed detection solution covered the membrane and once bands developed (within a few minutes), the solution was removed and the membrane was rinsed with dH₂O, and imaged on a scanner. In addition to seminal plasma, purified human albumin (Sigma Aldrich) and bovine serum albumin (ThermoFisher Scientific) were made in a dilution series (62.5ng to 5,000ng and 125ng to 5,000ng, respectively) and separated on a 10% SDS-PAGE (Invitrogen), as described above. Proteins were transferred to a PVDF membrane at 100 volts for 60 minutes, and blocked with SuperBlock PBS for 30 minutes. The membrane was incubated in mouse anti-albumin primary antibody diluted in Superblock PBS (1:5,000 or 1:50,000) overnight, and detected with FastRed.

2.2.3.2 Dot blot to confirm secondary antibody effectiveness

To determine if goat anti-mouse IgG alkaline phosphatase conjugated antibody (ThermoFisher Scientific) was active, a dot blot was performed. Both a dry PVDF membrane (not activated in methanol) and a dry nitrocellulose membrane absorbed 2µl of secondary antibody. The membrane was incubated with SuperBlock PBS for ten minutes, washed with PBS twice for ten minutes, and was detected with FastRed, as described above.

2.2.3.3 Purified actin dilution series

In order to compare several different Western blotting protocols and detection systems, a dilution series of bovine actin was detected with a monoclonal mouse anti-actin antibody. Bovine actin (Sigma Aldrich) was reconstituted in Tris-EDTA, pH 8, to a concentration of 1mg/mL. A dilution series (3ng/μl, 6 ng/μl, 12 ng/μl, 25ng/μl, 50ng/μl, and 100ng/μl) of actin was made, which was chosen based on Urban and Woo (2007). The dilution series (5μl) was prepared with 2:1 Laemmli buffer and 10% β-ME separated on a 10% SDS-PAGE (Invitrogen), as described above. Proteins were transferred at 100 volts for 75 minutes. Post-transfer, the PVDF membrane was blocked in SuperBlock PBS for 30 minutes, and was incubated in monoclonal mouse anti-actin antibody (Sigma Aldrich) overnight, which was either diluted 1:5,000 or 1:50,000 in SuperBlock PBS. Washes were performed in PBS for either three times for 15 minutes or three times for 5 minutes. Both goat anti-mouse alkaline phosphatase conjugated and peroxidase conjugated antibodies were diluted in SuperBlock PBS in the following dilutions: 1,1,000 or 1:50,000. Both FastRed and Pierce® ECL detection systems were used. Several membranes were stripped and re-probed. After detection, the PVDF membrane was incubated in stripping buffer (1.5% glycine, 0.1% SDS, and 1% Tween20) twice for 10 minutes at room temperature, then washed twice with PBS for 10 minutes, and lastly washed twice with TBST for 5 minutes. The stripped membrane was incubated in SuperBlock PBS for 30 minutes and downstream procedures continued.

2.2.3.4 Optimization of secondary antibodies

Varying concentrations of secondary goat anti-mouse IgG alkaline phosphatase conjugated antibody (Themofisher Scientific) were tested using 5μl of Magic Mark™ (Invitrogen) ladder separated on the same 10% SDS-PAGE gel and PVDF membrane. Once

transferred, the PVDF membrane was cut into 5 strips, blocked with SuperBlock PBS, incubated with the following secondary antibody dilutions: 1:1,000; 1:5,000; 1:25,000; 1:50,000; and 1:100,000. After washes, the individual strips were detected with FastRed solution, described previously.

After several attempts with Abnova primary antibodies, rabbit polyclonal anti- SEMG1 and rabbit polyclonal anti- SEMG2 (Abcam, Cambridge, UK) were optimized with human and chimpanzee seminal plasma. First 1, 5, 10, 15, 20, and 50 μ g of human seminal plasma were separated on a 10% SDS-PAGE (Invitrogen), as described above. Proteins were transferred to an Immuno-Blot® PVDF (Bio-Rad) membrane at 100 volts for 60 minutes and blocked in SuperBlock PBS (Thermofisher Scientific) for thirty minutes. Membranes were incubated with either 1:1,000¹ or 1:5,000 primary anti-SEMG1 or anti-SEMG2 antibody diluted in SuperBlock PBS overnight. Membranes were washed three times for five minutes in PBS. Goat anti-rabbit alkaline phosphatase (Abcam) conjugated antibody was diluted (1:25,000²; 1:40,000³; 1:50,000) in Superblock PBS and the membrane was incubated for an hour at room temperature. FastRed detection was used as described above.

2.2.3.5 Membrane transfer optimization

Besides optimizing wet transfer (using the mini Trans-Blot® Bio-Rad apparatus) voltage and times, a semi-dry iBlot® (Thermofisher Scientific) system was tested for comparison purposes. Volumes of human and chimpanzee seminal plasma containing 20 μ g of protein (in duplicate) were separated on a 10% SDS-PAGE (Invitrogen) at 150 volts for 90 minutes. The 10% SDS-PAGE gel was carefully cut between lanes 5 and 6, separating identical halves. One

¹ Optimal primary antibody dilution for both anti-SEMG1 and anti-SEMG2 Abcam antibodies

² Optimal secondary antibody dilution for SEMG2 detection

³ Optimal secondary antibody dilution for SEMG1 detection

half was transferred using the semi-dry iBlot® transfer system in seven minutes to an Immuno-Blot® PVDF (Bio-Rad) membrane. The membrane was kept wet in dH₂O, until the second half of the gel was transferred using the Trans-Blot® system to an Immuno-Blot® PVDF membrane (100 volts for 60 minutes). Both membranes were blocked for 45 minutes in SuperBlock PBS, rinsed three times for five minutes in PBS, and then incubated with rabbit polyclonal anti-SEMG2 antibody (1:10,000; Abcam) overnight. After incubation, the membrane was washed three times with PBS for five minutes, and incubated with goat anti-rabbit alkaline phosphatase conjugated antibody (1:25,000) for one hour at room temperature. FastRed detection was used, as described above.

Previously my lab detected cartilage acidic protein (CRTAC1) in high abundance in chimpanzee seminal plasma using liquid chromatography tandem mass spectrometry (LC/MS-MS), which was 142.2 fold higher relative abundance compared to human seminal plasma (Chovanec et al., unpublished). To further characterize this difference, 10µg of three human and three chimpanzee seminal plasma samples were separated on a 10% SDS-PAGE (Invitrogen), as described above. Proteins were transferred to an Immuno-Blot® PVDF (Bio-Rad) membrane at 100 volts for 75 minutes and blocked in SuperBlock PBS (ThermoFisher Scientific) for thirty minutes. Monoclonal mouse anti-CRTAC1 (Abnova) was diluted (1:1,000⁴; 1:5,000; or 1:50,000) in SuperBlock PBS, dispensed over the PVDF membrane and incubated overnight at room temperature. The membrane was washed three times in PBS for five minutes. Several blots were incubated with goat anti-mouse alkaline phosphatase conjugated antibody (1:1,000; 1:50,000) for one hour at room temperature. FastRed detection was used, as described above. In addition, a single blot was cut in half post transfer and incubated in either PBS or TBS for all the washes, and detected using FastRed and alkaline phosphatase conjugated secondary antibody.

⁴ Optimal primary antibody dilution for anti-CRTAC1 Abnova antibody

Several blots were incubated with goat anti-rabbit peroxidase conjugated antibody (1:1,000; 1:15,000; 1:20,000) for one hour at room temperature. After washes, blots were detected with either Pierce ® ECL or Pierce ® ECL Supersignal Western blotting kit, in a comparison study, using X-ray film.

2.2.3.6 Anti-pilin detection

Because natural seminal plasma samples were used for our Western blots, bands were often not single, and multiple variants of cross-linked, monomeric, and degraded of proteins were detected. Furthermore, it was sometimes unclear if the signal was truly related to our protein of interest, or if it was background signal due to our blotting protocol. In order to address these concerns during our optimizations of Western blotting protocols, hospital samples of *Pseudomonas aeruginosa* isolates were borrowed from Dr. Castric's lab, along with their optimized primary antibody. Samples were separated on a 10% SDS-PAGE (Invitrogen), at 150 volts for 60 minutes. Proteins were transferred to an Immuno-Blot® PVDF (Bio-Rad) membrane at 100 volts for 30 minutes and blocked in SuperBlock PBS (Thermofisher Scientific) for thirty minutes. The PVDF membrane was incubated in primary monoclonal mouse anti-pilin (1:1,000) antibody overnight at room temperature, and then was washed three times in PBS for five minutes. The blot was incubated with either goat anti-mouse alkaline phosphatase or peroxidase conjugated antibody (1:10,000) for sixty minutes. FastRed or Pierce® Supersignal Western blotting substrate kit were used for detection with their respective secondary antibodies.

2.2.4 Quality control checks

2.2.4.1 Loading control Western blot

All of the seminal proteins we study are extracellular proteins; therefore common intracellular controls, like β -actin, are not applicable. In addition, there are few, if any, published papers that include an extracellular protein control for seminal plasma. Using LC/MS-MS data of human and chimpanzee seminal plasma (Chovanec et al., unpublished), a few candidate proteins had similar moderate abundances in each individual and species. Each candidate protein's NCBI sequence for human, chimpanzee, gorilla, orangutan, and macaque were aligned using Clustal Omega to identify differences among species. Clusterin was moderately abundant, conserved, and had several commercial antibodies available. Clusterin was Western blotted as described in section 2.2.5.2.

2.2.4.2 Protease inhibitor assay

A new-pooled individual semen sample (donor number: T2566, named: 'Human 5') was ordered from Lee BioSystems, St. Louis, MO. The semen was frozen at -80°C upon arrival; the sample was thawed on ice and aliquoted ($250\mu\text{l}$) into three tubes. One tube had $2.5\mu\text{l}$ of the following protease inhibitors added: BCI-, apo, benzamide, and DMSF (gifted from the Cascio lab). Another tube had $2.5\mu\text{l}$ mammalian protease inhibitor premade cocktail (Sigma Aldrich) added. Subsequent aliquots ($50\mu\text{l}$) were made and incubated at the following temperatures: 4°C , 23°C , and 37°C for 80 minutes. Samples were quantified using the Bradford assay (Bio-Rad) and $10\mu\text{g}$ was loaded onto a 10% SDS-PAGE gel (Invitrogen) and either coomassie stained or Western blotted (*Li-COR method*) with an anti-SEMG1 or anti-SEMG2 antibody (Abcam).

2.2.5 Methods for comparative proteomics among human, chimpanzee, and gorilla seminal plasma

2.2.5.1 Shotgun Liquid Chromatography Tandem Mass Spectrometry LC-MS/MS

2.2.5.1.1 LC/MS-MS sample preparation

Three human, chimpanzee, and gorilla individual seminal plasma samples (described in section 2.2.1.1) were precipitated as follows. Equal volume of trifluoroethanol, TFE (Sigma Aldrich), was added to 25µl 100mM ammonium bicarbonate, pH 7.5 (Sigma Aldrich), 2.5µl 200mM dithiothreitol, DTT (Bio-Rad), and a volume of seminal plasma containing 30µg of total protein. Samples were heated at 90°C for 20 minutes; 10µl of 100mM iodoacetamide, IAM (Sigma Aldrich), was added and incubated for 60 minutes in the dark. To destroy excess IAM, 2.5µl of 200mM DTT was added and incubated in the dark for 60 minutes. Samples were dried with a speedvac and reconstituted with 100mM ammonium bicarbonate (pH 7.5). Samples were cold acetone (2X) precipitated, dried with a speedvac, and reconstituted in 50µl of 100mM ammonium bicarbonate (pH 7.5). Samples were digested with trypsin (Promega, Madison, WI) overnight at 37°C in a 1:50 enzyme to substrate concentration. Formic acid was added to stop trypsin activity and samples were dried with a speedvac; then reconstituted in 20µl of 0.1% formic acid.

2.2.5.1.2 LC/MS-MS conditions

A collaborator, John Thomas, working in the laboratory of Dr. Partha Basu, completed the analysis of generated peptides from the tryptic digestion. The system used for peptide analysis was a 6530 Q-TOF mass spectrometer (Agilent Technologies, Santa Clara, CA), 1200 series liquid chromatography (Agilent Technologies), HPLC-Chip Cube MS interface (Agilent Technologies), and an HPLC-Chip with a 160 nL enrichment column (Agilent Technologies).

Peptides were separated using a 150 mm × 75 mm analytical column packed with Zorbax 300 SB-C18 (5 mm particles). For each injection, 2 µl of peptides were loaded onto the enrichment column with 95% solvent A (95% water, 5% acetonitrile, 0.1% formic acid) with a sample flush out factor of 10.0 times the injection volume. The system maintained a 95% solvent A concentration for the first 5 minutes of the run. They were then eluted with a gradient from 5% B (95% acetonitrile, 5% water, 0.1% formic acid) to 55% B in 45 minutes, 80% in 55 minutes, at a flow rate of 0.5 µl/min. The total runtime, including column reconditioning, was 60 minutes. The column effluent was directly coupled to the 6530 Q-TOF mass spectrometer via a HPLC-Chip Cube nanospray source operated at capillary voltage 1950 V with a capillary voltage in 2GHz extended dynamic range. The MS data was acquired in the positive ionization mode using MassHunter Workstation (Agilent Technologies, Build 5.01.5125.1). Gas temperature, drying gas volume, fragmentor voltage, skimmer voltage, and octopole RF were set to 365°C, 4 L/min, 175 V, 65 V, and 750 V, respectively. Auto MS/MS was performed with a total cycle time of 1.3 seconds. In each cycle, MS spectra was acquired at 4 Hz (5 spectra per sec; m/z 250-2400) and the three most-abundant ions (with charge states +2, +3, > +3, and unknown) exceeding 1000 counts were selected for MS/MS at 3 Hz (3 spectra per sec; m/z 50-3200). Medium isolation (4 m/z) window was used for precursor isolation. Equation 2.1 was used to determine the collision energy based on charge state and precursor m/z. Table 2.1 presents the parameters for the calculation of the collision energy. Reference mass correction was activated using reference masses of 299.294437 m/z and 1221.990637. Precursors were set in an exclusion list for 0.5 min after two MS/MS spectra.

$$\text{Collision Energy} = \frac{(\text{slope}) \times \left(\frac{m}{z}\right)}{100} + \text{Offset}$$

Equation 2.1: Equation used to calculate Collision energy

Table 2.1: Parameters for collision energy calculation.

Charge	Slope	Offset
2	3	2
3	3.6	-4.8
>3	3.6	-4.8
Unknown	3.6	-4.8

2.2.5.1.3 LC/MS-MS spectra data processing

Collision induced dissociation (CID) data was searched against *Homo sapiens*, *Pan troglodytes*, and *Gorilla gorilla gorilla* databases. The data files were extracted to generate peak lists using Spectrum Mills Proteomic Workbench (Agilent Technologies, Rev B.04.01.141) with the following attributions: precursors were between 250 to 2400 Da over the scan time of 0 to 60 minutes, scans with the same precursor ± 1.4 m/z were merged within a time frame of ± 30 seconds. Precursor ions needed to have a minimum signal to noise value of 25. Charges up to a maximum of 7 were assigned to the precursor ion, and the 12C peak was determined by the Data Extractor. The databases were searched for tryptic peptides with a product mass tolerance of ± 0.7 Da for the precursor ions and a tolerance of ± 1.2 Da for the fragment ions. Two missed cleavages were allowed and the minimum tag length was >3 with at least 4 detected peaks. The search mode was set to “Identity” with carboamidomethylation as fixed modification. A Spectrum Mills autovalidation was performed first in the protein details followed by peptide mode using a.) in protein details mode; protein score ≥ 20 , peptide score (scored percent intensity [SPI]) charge +2 (>6 , $>60\%$), peptide charge +1 (>6 , $>70\%$), peptide charge +3 (>8 , $>70\%$),

peptide charge +4 (>8, >70%), peptide charge +5 (>12, >70%), peptide charge +2 (>6, >90%), and rank 1 minus rank 2 score threshold 2 for the first 5 rules and 1 for the final rule; b.) in peptide mode: SPI charge +2 (>11, >60%), peptide charge +1 (>13, >70%), peptide charge +3 (>13, >70%), peptide charge +4 (>13, >70%), peptide charge +5 (>15, >70%), and rank 1 minus rank 2 score threshold ≥ 2 . All protein hits found in a distinct database search by Spectrum Mill are non-redundant. When an individual peptide matched to more than one species database, it was prioritized as follows: a) if the database match was to a protein from the same species as the sample origin, that annotation was used; b) else, if the peptide matched to the human database, that annotation was used; c) else, if the peptide matched to another species, that annotation was used.

2.2.5.1.4 Manual mapping of SEMG peptides

The automated mapping to NCBI databases described above misidentified SEMG1 and SEMG2 peptides. Therefore, all the SEMG identified peptides were manually mapped; unique peptides were identified to either SEMG1 or SEMG2 sequences. Table 2.2 includes the accession numbers to the SEMG sequences utilized for each species. Downstream quantification analysis used my manual SEMG assignment opposed to the computational database search assignment.

Table 2.2: SEMG sequences used for peptide mapping

Protein name	Accession number
Human SEMG1	AAB59506.1
Human SEMG2	AAA60562.1
Chimpanzee SEMG1	ABO52927.1
Chimpanzee SEMG2	Q5U7N4.1
Gorilla SEMG1	ABO52952
Gorilla SEMG2	Q5U7N3

2.2.5.1.5 Protein quantification and statistical analysis

The resulting peptide and protein matches were filtered to produce a set of “high confidence” proteins by eliminating those found in only one run for any given species, and those found as only one unique peptide. These data were then normalized by dividing the number of peptides matching each protein by the total number of peptides in that run. An alternative method of normalization was also used, in which the number of peptides were standardized by the length of the protein (using the length of the longest isoform in the human UniProt database, except for SEMG1 in which case the length used was the species-specific length). These protein abundances are referred to as the “normalized spectral abundance frequencies”, or NSAF. The NSAFs were averaged across the three technical replicates of each individual. Two methods were used to assess the significance of the differences between pairs of species in protein abundance, for both the length-standardized and the non-standardized (“raw”) data sets. First, a two-tailed t-test was used with a Benjamini-Hochberg false discovery rate of 0.1 (Benjamini and Hochberg, 1995). Second, a power law global error method (PLGEM) commonly used for microarray and proteomic data described by Pavelka et al. (2008) was used. For this, the number of subsets sampled during the model-fitting steps was changed to 5, the number of iterations was increased to 10,000, and the false discovery rate was set to 0.01. Biodiversity indexes, Shannon and Simpson, were utilized on length and raw normalized abundances. Biodiversity index scores were calculated based on species averaged protein abundance, using the equations listed in Table 2.3. Subsequently proteins were analyzed for gene ontology on Panther and DAVID online bioinformatic tools.

Table 2.3: Equations used to calculate Shannon and Simpson biodiversity indexes

Diversity Index	Equations
Shannon Diversity index values	$H' = - \sum_{i=1}^R p_i \ln p_i$
	$E_H = \frac{\ln(n)}{- \sum_{i=1}^R p_i \ln p_i}$
	$e^H = Ln(- \sum_{i=1}^R p_i \ln p_i)$
Simpson Diversity index value	$\frac{1}{D} = \frac{1}{\sum_{i=1}^R p_i^2}$

2.2.5.2 SDS PAGE and Western blot methods

Volumes of human, chimpanzee, and gorilla seminal plasma containing 10µg of protein were mixed with dH₂O (to 10µl). Samples were diluted (2:1) with Laemmli buffer (Bio-Rad) and 10% β-ME and incubated at 95°C for ten minutes. Samples were loaded into 10% SDS-PAGE gel (Invitrogen) in 1X NuPAGE MOPS (Invitrogen) running buffer and run at 110 volts for 15 minutes, and then 220 volts for 30 minutes. For SDS-PAGE imaging, the gel was incubated with coomassie stain (Li-COR) for 30 minutes and destained with dH₂O overnight. After destaining, the gel was imaged on the FC Odyssey imager (Li-COR) under the 700nm channel.

SDS-PAGE gels that underwent protein-specific antibody detection followed the protocol outlined below. An Immobilon®-FL PVDF (Millipore, Merck, Germany) was activated in methanol for one minute, and incubated in 1X NuPAGE (Invitrogen) transfer buffer with the separated gel and other cassette components for 10 minutes. The Western sandwich was assembled and put into the mini Trans-Blot® (Bio-Rad) chamber with an ice pack. The apparatus

was placed on a stir plate with agitation, and proteins were transferred at 100 volts for 45 minutes. The PVDF membrane was rinsed with water, and incubated in 50%odyssey block/50% PBS for an hour and subsequently rinsed in dH₂O, three times for five minutes. Primary antibody was diluted, specific to each antibody (Table 2.4), in 50%odyssey block/50% PBST. The membrane was incubated overnight in primary antibody, washed three times for five minutes in PBS, and then incubated for an hour in secondary IRDye® 680RD or 800CW (LI-COR) antibody diluted either 1:10,00 or 1:15,000. Following three washes in PBST for five minutes, the Western blot was imaged on the Odyssey FC (LI-COR) imager. Western blot samples were quantified using LI-COR Image Studio Lite software. All previous blots were repeated and detected using the Odyssey.

Table 2.4: Primary antibodies and their dilutions

Primary Antibody	Dilution	Secondary Antibody
Mouse anti-ACPP	1:3,000	1:10,000
Mouse anti-KLK3	1:1,000	1:10,000
Rabbit anti-KLK3	1:1,000	1:10,000
Mouse anti-TGM4	1:1,000	1:10,000
Goat anti-Clusterin	1:1,000	1:15,000

* All antibodies were from Abcam (Cambridge, UK).

2.3 Results

2.3.1 Protein quantification

The Bradford and Qbit protein quantification assays were compared to one another using each assay's standards. The difference between means (expected and calculated concentration) and a paired t-test were used to analyze results, and were calculated based on three categories: low standards, high standards, and all standards (Table 2.5). Analysis was completed in R studio and is provided in the Appendix (A.1).

In general, the Bradford assay reported lower concentrations than expected, which were all significant ($p \leq 0.006$) for known standards under $400 \mu\text{g/mL}$. Bradford standards used for generating the curve are between $125 - 2,000 \mu\text{g/mL}$, which does not explain why $200 \mu\text{g/mL}$ and $400 \mu\text{g/mL}$ Qbit standards would be significantly different. The Qbit assay reported higher concentrations than expected, and were significantly different for higher Bradford standards (or all standards). However, lower Bradford standards ($125 \mu\text{g/mL} - 500 \mu\text{g/mL}$) were not significantly different ($p=0.09$). It is important to note, that higher Bradford standards were diluted before assaying in order to fall within the assay's standard range, and their calculated diluted concentrations were multiplied by their dilution factor to obtain their final concentration. It is probable that the extrapolation of the diluted concentration may have contributed to the significant difference between calculated concentration and expected concentration. Overall, each quantification assay yields a different concentration for the same sample, which is likely attributed to the mechanism behind quantification and the samples amino acid composition. This comparison indicates a trend in the Bradford assay reporting lower than known concentration values while the Qbit assay reports higher than known concentration, with the caveat that the commercial known concentrations are correct.

Table 2.5: Known concentrations are lower in Bradford and higher in Qbit assays than expected

Standards	Assay	Sum of differences between expected 'known' concentration and calculated concentration.	Paired t-test of differences between means
Low Qbit Standards	Bradford	-203.24	p = 0.006
High Qbit Standards	Bradford	-95.36	p = 0.003
All Qbit Standards	Bradford	-149.30	p = 0.002
Low Bradford Standards	Qbit	+221.3	p = 0.09
High Bradford Standards	Qbit	+1212.75	p = 0.007
All Bradford Standards	Qbit	+787.85	p= 0.012

Semen samples were measured using the Bradford protein assay in order to be consistent for comparison studies. Concentrations are listed in Table 2.6 and the standard curve and sample absorbencies are graphed in Figure 2.1. The average concentration of human seminal plasma was 9.44 ± 1.21 mg/mL, chimpanzee seminal plasma was 16.91 ± 1.97 mg/mL, and gorilla seminal plasma was 14.20 ± 1.63 mg/mL. The average chimpanzee seminal plasma concentration was significantly higher compared to human ($p= 0.019$), but not gorilla ($p=0.331$). Gorilla and human seminal plasma concentrations were not significantly different ($p=0.059$).

Table 2.6: Hominid species seminal plasma concentrations

Species	Name Assignment	Sample ID	Sample concentration (mg/mL)
Human <i>Species average:</i> 9.44 ± 1.21 mg/mL	"Hu#1"	Donor ID: T3027	11.97
	"Hu#2"	Donor ID: T3206	12.24
	"Hu#3"	Donor ID: T3127	5.77
	"Hu#4"	Donor ID: T3402	8.93
	"Hu#5"	Donor ID: T2566	8.28
Chimpanzee <i>Species Average:</i> 16.91 ± 1.97 mg/mL	"Ch#1"	Kent, ID#95A016	18.05
	"Ch#2"	BJ, ID#99A003	14.51
	"Ch#3"	Jared, ID#95A018	13.13
	"Ch#4"	Little Joe	21.95
Gorilla <i>Species Average:</i> 14.20 ± 1.63 mg/mL	"Go#1"	Mo, Studbook ID#835	11.53
	"Go#2"	Tubby, Studbook ID#883	12.62
	"Go#3"	Bom, Studbook ID#612	18.91
	"Go#4"	Kit	13.71

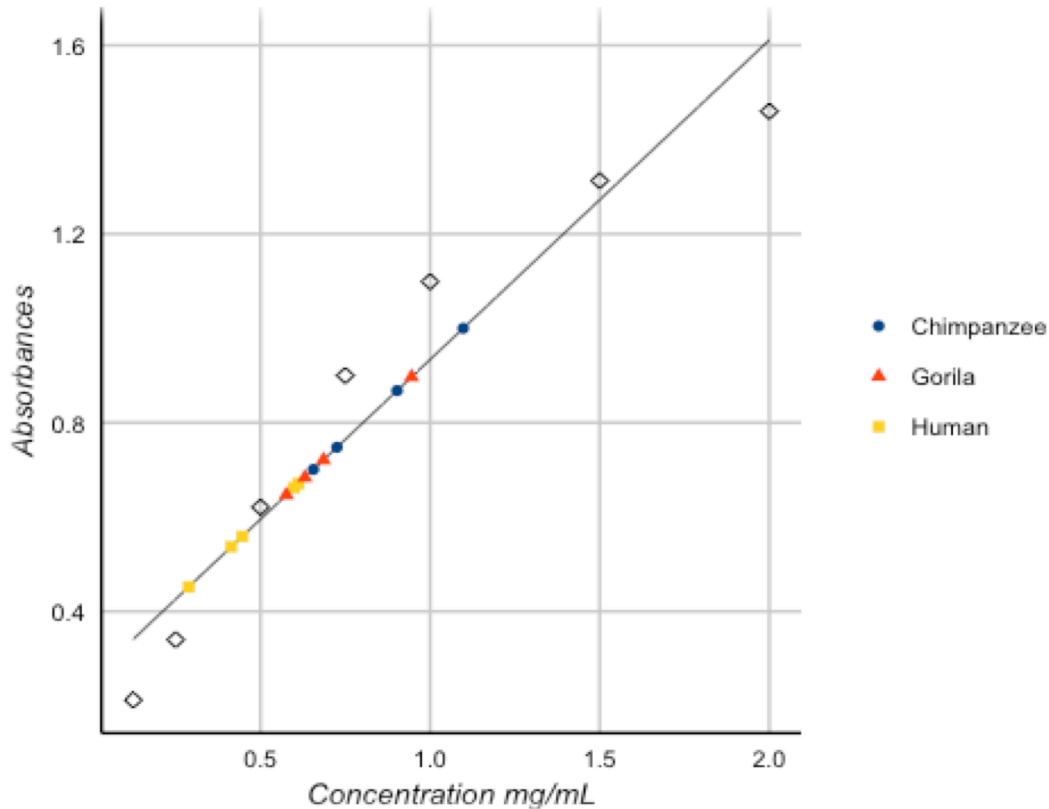


Figure 2.1: Bradford assay with seminal plasma sample absorbencies

Standard BSA samples in the Bradford assay were plotted (black outlined diamonds) in R and a linear regression line was calculated ($y = 0.677x + 0.256$, $R^2 = 0.93$). Chimpanzee, gorilla, and human seminal plasma sample concentrations were calculated by inputting their diluted absorbencies for y and solving for x . Dilution concentrations were multiplied by 20 in order to calculate the total concentration for the sample (Table 2.6). The graph above has the absorbencies (of dilutions) plotted alongside the standard absorbencies. All samples fell within the standard range, which ensures accurate predicted concentrations.

2.3.2 SDS PAGE optimization

Multiple attempts at pouring 6-10% stacking gels or 10% SDS-PAGE gels were attempted. Consistently, commercial 10% NuPAGE (Invitrogen) gels provided better sample separation and clarity. Figure 2.2 shows a representative experiment where human seminal plasma (30 μ g) and mock- transfected 293T media (10 μ g) were separated in a 10% NuPAGE and a 6-10% poured stacking gel in the same electrophoresis chamber. Post electrophoresis, gels were stained using the same coomassie stain and destain. Therefore, the only difference between gels was the preparation (poured or commercial).

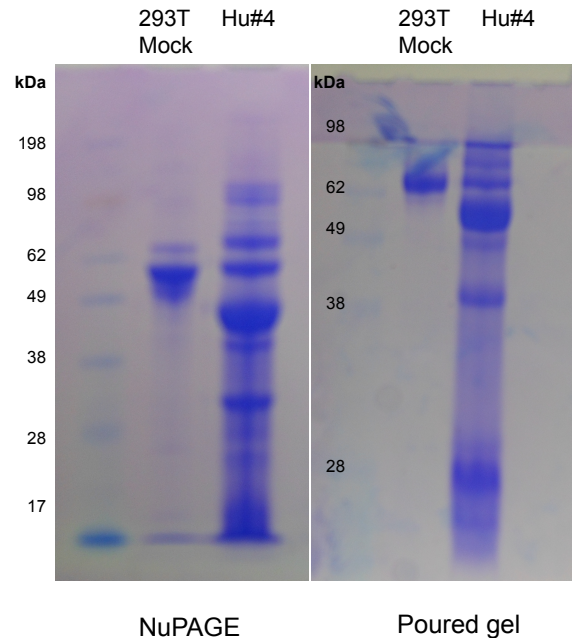


Figure 2.2: Commercial NuPAGE gels yield better clarity than poured stacking gels
 Mock- transfected 293T media (10 μ g) and human seminal plasma (30 μ g) were separated in a 10% NuPAGE gel (left) and a 6-10% poured stacking gel (right) and stained with coomassie. The commercial NuPAGE gel (left) yielded better separation of samples.

2.3.3 Western blot optimization

2.3.3.1 Bio-Rad wet transfer system provides band clarity in Western blots compared to iBlot™ transfer system

Polyclonal rabbit anti-SEMG1 and anti-SEMG2 antibodies were ordered from ABCAM, and SEMG2 primary antibody was used to optimize several detection methods. An iBlot™ (Invitrogen) dry transfer system was temporarily available in our department for trial experiments. The iBlot™ system transfers protein from an SDS-PAGE to a membrane within 7 minutes, whereas our Bio-Rad wet system transferred proteins between 45-75 minutes. Human and chimpanzee seminal plasma (20 μ g) in quadruplicate was separated on the same SDS-PAGE gel. Then the gel was cut into equal halves and proteins were transferred utilizing either the iBlot™ system or the Bio-Rad wet transfer system. Post-transfer the PVDF membranes were incubated in antibodies and washes in the same container. The only difference between

membrane detection was the transfer system. When using the anti-SEMG2 (ABCAM) primary antibody, Bio-Rad wet transfer yielded cleaner results with FastRed detection (Figure 2.3). Distinct bands around 70kDa were detected for human seminal plasma using the wet transfer system, while the iBlot™ transfer system did not detect the full size of SEMG2. It is important to note that SEMG proteins are known substrates for prostate-specific transglutaminase and for Kallikrein 3, and the smearing pattern of protein is probably a result of protein cross-linking and degradation.

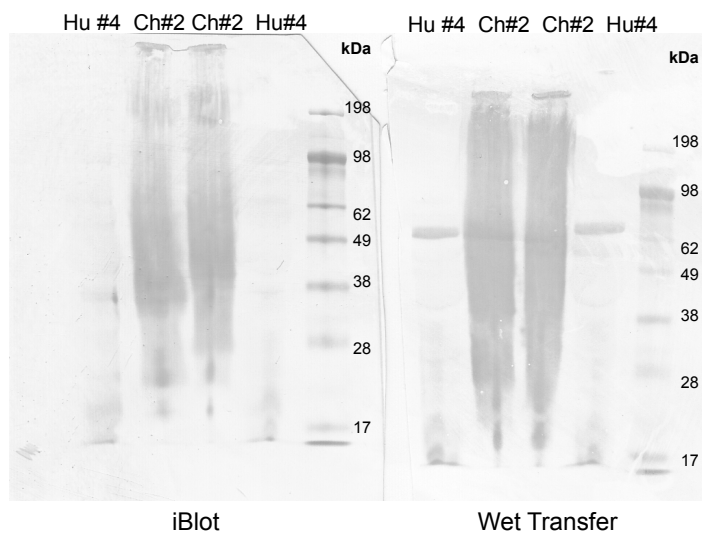


Figure 2.3: Bio-Rad wet transfer system has a clearer signal to background Western blot ratio compared to iBlot™ dry transfer

Human and chimpanzee seminal plasma (20µg) was separated on the same SDS-PAGE gel in quadruplicate, transferred by the iBlot™ dry transfer system or the Bio-Rad wet transfer system, and incubated with the same antibodies and FastRed detection substrates.

2.3.3.2 Non-specific bands are due to cross-linking and degradation of seminal proteins and not due to our Western blot protocol

In order to address the smeary appearance of our seminal plasma Western blots (seen in Figure 2.3), we used our protocols and equipment to blot samples that are not known to be cross-linked or degraded. Hospital samples of nonglycosylated *Pseudomonas aeruginosa* isolates (B017, B018, and B029) were gifted from Dr. Pete Castric's lab, along with their optimized

monoclonal mouse anti-pilin primary antibody. Our lab equipment and FastRed Western blot protocol were utilized with their samples and primary antibody. Nonglycosylated mature pilin protein is around 16 kDa molecular weight while glycosylated pilin protein is around 17kDa (Castric et al., 2001). Nonglycosylated pilin samples were detected at the expected 16kDa with low background signal (Figure 2.4). This Western blot confirmed that our equipment and protocol were sufficient for detection of protein using optimized working antibodies. In addition, if the protein of interest (pilin) was not susceptible to degradation and cross-linking (unlike SEMG1 and SEMG2), our methods would detect clean single bands within the expected molecular weight range.

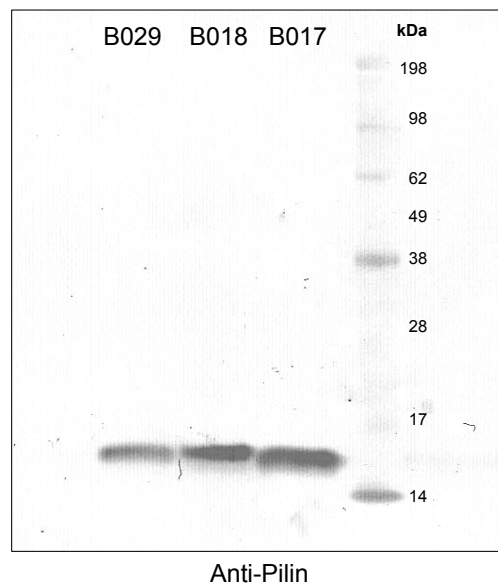


Figure 2.4: Nonglycosylated pilin was cleanly detected from hospital samples
Pseudomonas aeruginosa isolates (B017, B018, and B029) from Children’s Hospital of Pittsburgh were separated on an SDS-PAGE and mature nonglycosylated pilin was detected around 16kDa.

2.3.3.3 Alkaline phosphatase conjugated secondary antibody and FastRed colorimetric detection reduces signal to noise ratio compared to chemiluminescent detection

I wanted to assay the sensitivity of chemiluminescent detection with horseradish peroxidase (HRP) conjugated secondary antibody and the FastRed colorimetric detection with alkaline phosphatase (AP) conjugated secondary antibody. For this experiment, purified human actin (Sigma Aldrich) with different amounts of protein (15ng, 31ng, 62.5ng, 125ng, 250ng, 500ng, and 5,000ng) were separated and transferred to PVDF membranes under the same conditions. The only difference between membranes was the secondary antibody applied (either HRP or AP conjugated) and the detection substrate (ECL chemiluminescent or FastRed colorimetric, respectively). Both detection methods were able to detect 250ng of actin (Figure 2.5), but the FastRed colorimetric Western blot had relatively less background noise compared to signal.

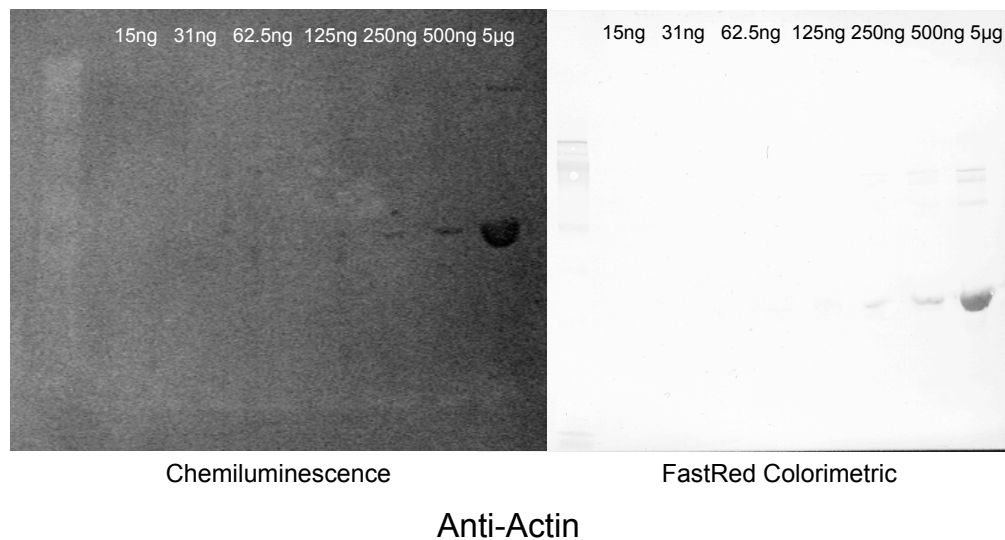


Figure 2.5: Colorimetric FastRed detection has reduced background noise compared to chemiluminescence

Human actin protein was Western blotted using horseradish peroxidase (HRP) conjugated secondary antibody and chemiluminescent ECL substrate and imaged on the typhoon (left) and was also Western blotted using alkaline phosphatase (AP) conjugated secondary antibody and FastRed colorimetric detection (right). Lanes are labeled with total amount of protein that was loaded into the SDS-PAGE gel. Both detection methods were able to visualize actin in greater amounts than 125ng.

2.3.4 Western blot quality control checks

2.3.4.1 Increased abundance of heterodimeric clusterin in gorilla seminal plasma prevents it from being an effective extracellular loading control across hominid seminal plasma

Currently there is no universal loading control for seminal plasma, like β -actin, or other housekeeping proteins, in intracellular experiments. We searched for a protein that was similar in expression across primate seminal plasma, which could act as a loading control between species. After reviewing previous shotgun Liquid Chromatography Tandem Mass Spectrometry (LC/MS-MS) data comparing human and chimpanzee seminal plasma (Chovanec et al., unpublished), two candidate proteins, albumin and clusterin, were both highly abundant and similar in concentration between species. Both albumin and clusterin have seven amino acid differences between human and chimpanzee (Appendix A.2, Alignments 1 and 2). Clusterin was chosen because it was less likely to have false signal compared to albumin. Bovine serum albumin is used as a blank in LC/MS-MS and is in many blocking buffers, which could provide additional signal for human albumin in an experiment. When human and bovine serum albumin were aligned there were four fragments greater than six amino acids that could be cross identified between human and bovine serum albumin (Appendix A.2, Alignment 3).

Initially, human and chimpanzee seminal plasma (10 μ g and 20 μ g) was Western blotted for clusterin. Human clusterin detection was 30% higher than chimpanzee (opposite of the LC/MS-MS data). Both human and chimpanzee detected protein was the appropriate molecular weight for clusterin (Figure 2.6:A) in its monomeric form (~48kDa). When human, chimpanzee, and gorilla seminal plasma were Western blotted the results were less clear (Figure 2.6:B). Gorilla seminal plasma inconsistently blots both the monomeric form and the heterodimeric disulfide cross-linked form of clusterin (70-80kDa), which provides a challenge to quantify

clusterin signal. Human clusterin detection was 30% higher than chimpanzee and gorilla was the lowest signal, which is opposite of my LC/MS-MS discussed later in this chapter.

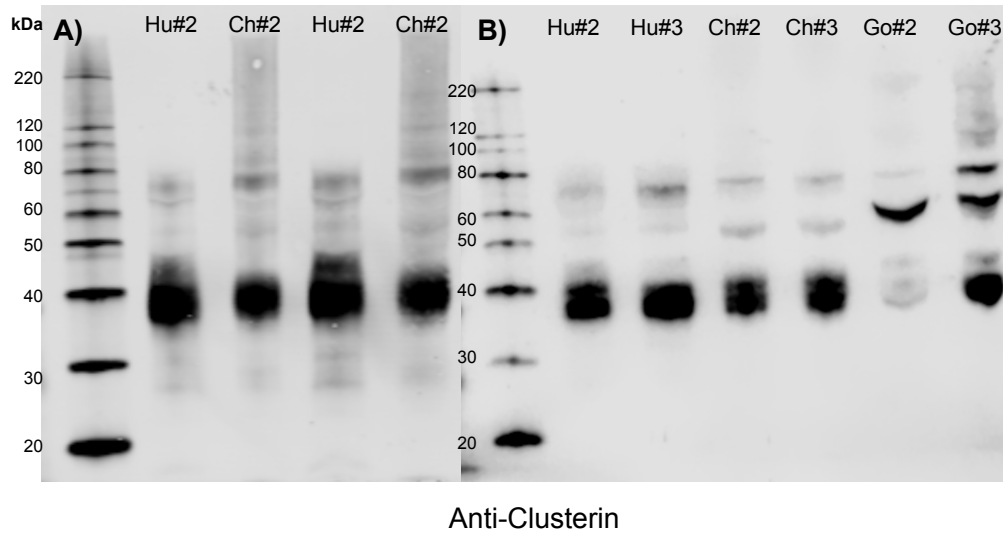


Figure 2.6: Clusterin is not a reliable extracellular loading control

Anti-clusterin antibody from ABCAM was used to detect clusterin in human, chimpanzee, and gorilla seminal plasma. **A)** Lanes 2 and 3 have 10µg of human or chimpanzee seminal plasma loaded and lanes 4 and 5 have 20µg of the same human or chimpanzee seminal plasma loaded. Human signal was roughly 30% higher than chimpanzee in both loading conditions. **B)** Two individuals of each species' seminal plasma (20 µg) was loaded and detected with anti-clusterin antibody. Monomeric clusterin protein should be around 48kDa; however, heterodimeric clusterin which is disulfide cross-linked is expected to be around 70-80kDa. Though gorilla clusterin detection looks different than human and chimpanzee, it most likely has both monomeric and heterodimeric forms of clusterin.

2.3.4.2 Majority of proteolytic degradation of proteins, particularly the semenogelins, occurs before we receive semen samples

The semenogelins are known substrates for cross-linking and degradation, which explains the appearance of multiple bands in my Western blots detecting either protein (i.e., Figure 2.3). Protease inhibitor cocktails were added to semen samples (BCI-, apo, benzamide, and DMSF – “CPI-Hu#5” or Sigma mammalian protease inhibitors- “SPI-Hu#5”) to determine if a majority of protein degradation could be prevented. Semen with and without inhibitors was centrifuged to separate seminal plasma proteins from spermatozoa, and then aliquoted into different tubes and each tube was incubated at 4°C, 23°C, and 37°C for an hour, potentially allowing enzymatic

activity. Samples were quantified (Figure 2.7), and samples with protease inhibitors had a slightly higher concentration than without inhibitors. However, there were minimal differences in protein concentration across the different aliquots incubated at different temperatures.

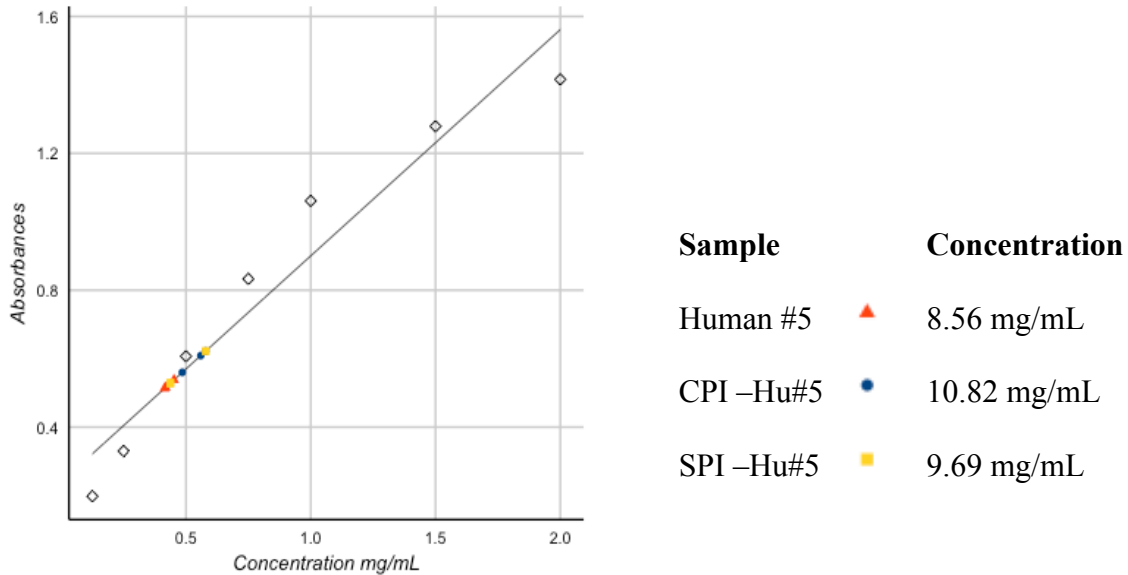


Figure 2.7: No significant differences in protein concentration in seminal plasma with and without protease inhibitors

Human semen (Hu#5), human semen with a mix of protease inhibitors (CPI), and human semen with a commercial sigma protease inhibitor (SPI) cocktail were aliquoted and incubated at different temperatures, and their absorbencies were measured using BioRad protein assay dye. Standard BSA samples in the Bradford assay were plotted (diamonds) in R and a linear regression line was calculated ($y = 0.661x + 0.239$, $R^2 = 0.94$). The average calculated concentration of each sample is listed on the right and their concentrations were not significantly different ($p = 0.124$; see Appendix A.1.4).

There were no noticeable visual differences between samples (inhibitor vs. non-inhibitor or across temperature incubation) when samples were separated by SDS-PAGE (Figure 2.8) or blotted with anti-SEMG antibodies (Figure 2.9). Both anti-SEMG1 and SEMG2 antibodies detected proteins at their respective full-length molecular weights; however, they also had smaller portions of protein detected. Essentially, a large amount of degradation occurs before we receive our semen samples. Although adding protease inhibitors may prevent subsequent degradation, unless inhibitors are immediately added during collection, majority of proteins susceptible to proteolytic cleavage will be degraded.

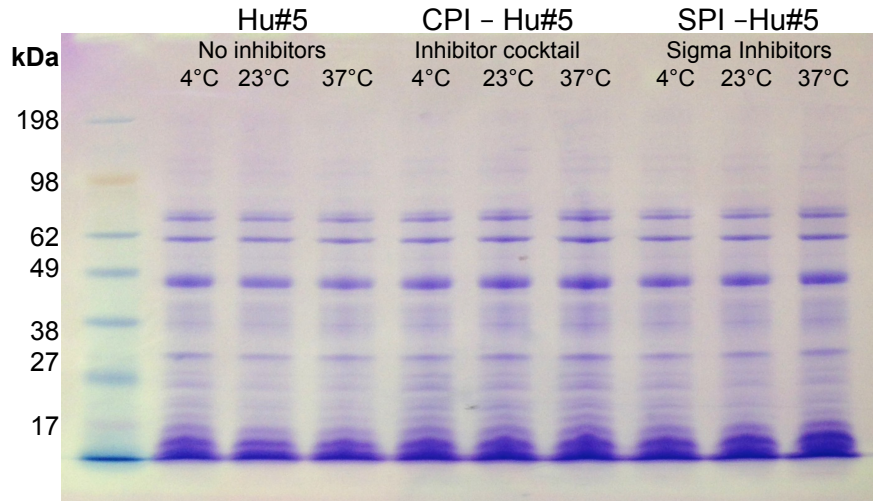


Figure 2.8: Seminal plasma samples with and without protease inhibitors look similar

Human semen was aliquoted and immediately inhibitors (BCL-, apo, benzamide, and DMSF) were added (inhibitor cocktail/CPI) or a premade mammalian protease inhibitor cocktail (Sigma inhibitor mix/SPI) was added. Centrifugation removed sperm cells and seminal plasma was subjected to varying temperature (4°C, 23°C, or 37°C) incubations for an hour. Seminal plasma with and without protease inhibitors visually look the same when stained with coomassie.

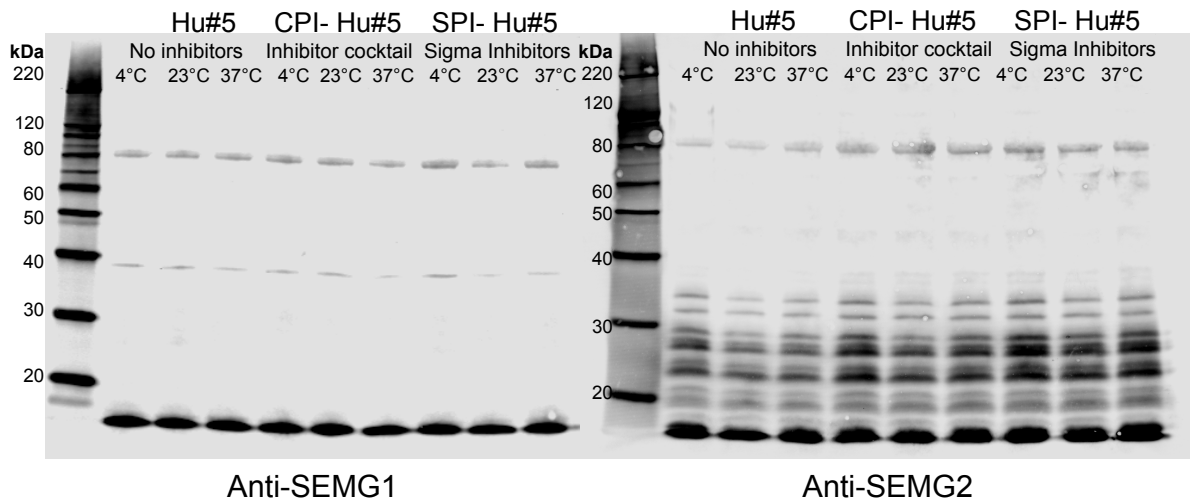


Figure 2.9: SEMG1 and SEMG2 are degraded before addition of protease inhibitors

Human semen was aliquoted and immediately inhibitors (BCL-, apo, benzamide, and DMSF) were added (inhibitor cocktail/CPI) or a premade mammalian protease inhibitor cocktail (Sigma inhibitor mix/SPI) was added. Centrifugation removed sperm cells and seminal plasma was subjected to varying temperature (4°C, 23°C, or 37°C) incubations for an hour. There were minimal or no differences of anti-SEMG binding between 10µg seminal plasma with and without protease inhibitors. Both full length and varying sizes of degraded SEMG peptides were detected across samples.

2.3.5 LC/MS-MS and Western blot identification of seminal plasma proteins among human, chimpanzee, and gorilla individuals

A total of 7,735 peptides were identified across all three species, nine individuals, and twenty-seven runs. Initially, all chimpanzee semenogelin peptides were identified as SEMG2. Considering that genomic chimpanzee *SEMG2* analysis predicts pseudogenization (discussed in section 1.4.2.1), but anti-SEMG2 antibody (Figure 2.3) detected protein from chimpanzee seminal plasma, SEMGs were manually identified by peptide sequence before further analyzing LC/MS-MS data.

2.3.5.1 SEMGs, if expressed, are in relatively high abundance across hominid species with differences among SEMG1 and SEMG2 expression

For each species, semenogelin peptides were manually mapped to reference sequences (Table 2.2) and identified as either SEMG1 or SEMG2 (Table 2.7). Several peptides were ambiguous as they matched well to either paralog (Figure 2.10) in each species, and were removed from the data set for further quantification. There were five unique peptides identified in chimpanzee that appear more likely to be from SEMG2 (Figure 2.10); however, a majority of chimpanzee semenogelin peptides mapped to SEMG1. Only one of the three gorilla individuals had semenogelin peptides identified, although when peptides were identified, they were in relatively high abundance (Table 2.7). Using an anti-SEMG2 antibody, semenogelin expression was confirmed in one of the three gorillas (Figure 2.11). Previous genomic sequence analysis of gorilla SEMGs indicated multiple premature stop codons in the semenogelin genes that are polymorphic (Figure 1.5:D).

Table 2.7: Normalized raw spectral abundances of SEMGs are relatively high but variable across hominids

Species	SEMG1 % Abundance	SEMG2 % Abundance	SEMG total number of peptides identified
Human (Average)	13.9 ± 6.7 %	20.3 ± 2.0 %	1,050
Chimpanzee (Average)	43.6 ± 2.7 %	1.47 ± 0.4 %	1,258
Gorilla (Average)	3.25 ± 3.25%	10.6 ± 10.6 %	81
Gorilla (Individual #3)	9.74 ± 1.74 %	31.7 ± 11.5 %	81

Human SEMG1 (AAB59506.1)
 MKPNIIIFVLSLLILEKQAAMVGGKGGSKGRLPSEFSQFPHGQKGOHYSGQKQKQQTESK
 GSFSIQTYHYVDANDHDQSRKSSQYDLNALHKTTSQRHLGGSQQLLNKQEGRDHDKSK
 GHFHRVVIHHKGGKAHRGTQNPSPQDQGNPSGKGISSQYSNTEERLWVHGLSKEQTSVSG
 AQKGRKQGGSSSYVLQTEELVANKQRETKNSHQKGYQNVVEVREEHSSKVTSLCP
 AHQDKLQHGSKDIFSTQDELLVYNKNQHQTKNLNQDQHGKANKISYQSSSTEERLHY
 GENGVQKDVSSSIYSQTEEKAQKSKQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GENGVQKDVSSSIYSQTEKLVAGKSIQAPNPKQEPWHGENAKGESGQSTNREQDLLSH
 EQKGRHQHSGSHGLDIVIIEQEDSDRHLAQHLNDRNPLFT

Chimpanzee SEMG1 (ABO52927.1)
 MKPNIIIFVLSLLILEKQAAMVGGKGGSKGRLPSESSQFPHGQKGOHYSGQKQKQQTESK
 GSFSIQTYHYVDANDHDQTRKSSQYDLNALHKTTSERHLGGSQQLLNKQEGRDHDKSK
 GHFHRVVIHHKGGKAHRWTQNPSPQDQGNPSGKGISSQYSNTEERLWVHGLSKEQTSVSG
 AQKGRKQGGSSSYVLQTEELVANKQRETKNSHQKGYQNVVEVREEHSSKVTSLCP
 AHQDKLQHGSKDIFSTQDELLVYNKNQHQTKNLNQDQHGKANKISYQSSSTEERLHY
 GENGVQKDVSSSIYSQTEEKAHGSQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GENGVQKDVSSSIYSQTEEKAHGSQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GENGVQKDVSSSIYSQTEEKAHGSQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GENGVQKDVSSSIYSQTEEKAHGSQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GENGVQKDVSSSIYSQTEKLVAGKSIQAPNPKQEPWHGENAKGESGQSTNREQDLLSRE
 EQKGRHQHSGSHGLDIVIIEEEDSDHHLAQHLNDRNPLFT

Human SEMG2 (AAA60562.1)
 MKSIILFVLSLLILEKQAAMVGGKGGKGLPSGSSQFPHGQKGOHYFGQKQDQHTKSK
 GSFSIQTYHYVDINDHDQTRKSSQYDLNALHKTTSKQHLGGSQQLLNKQEGRDHDKSK
 GHFHRVVIHHKGGQAHCCTQNPSPQDQGNPSGKGLSSQYSNTEERLWVHGLSKEQASASG
 AQKGRKQGGSSSYVLQTEELVANKQRETKNSHQKGYQNVVVDREHSSKLTQSLHP
 AHQDRLOHGPKDIFSTQDELLVYNKNQHQTKNLNQDQEHGKANKISYQSSSTEERLHY
 GEKSVQKDVSSSIYSQTEEKAHGSQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GEKGIQKGVSKGSIYSQTEEKAHGSQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GEKDVQKGVSKGSIYSQTEEKAHGSQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GKGSTQKDVSSSIYSQTEEKAHGSQKQITIPSNPNQDQWSGQNAKKGSGQASDSDKQDLLSH
 EQKGRYKQESSSESHNIVITEHEVAQDDHLTQQYNEDRNPIST

Chimpanzee SEMG2 (Q5U7N4.1)
 MKSIILFVLSLLILEKQAAMVGGKGGKGLPSGSSQFPHGQKGOHYFGQKQDQHTKSK
 GSFSIQTYHYVDINDHDQTRKSSQYDLNALHKTTSKQHLGGSQQLLNKQEGRDHDKSE
 GHFHRVVIHHKGGQAHCCTQNPSPQDQGNPSGKGLSSQYSNTEERLWVHGLSKEQASASG
 AQKGRKQGGSSSYVLQTEELVANKQRETKNSHQKGYQNVVVDREHSSKLTQSLHP
 AHQDRLOHGPKDIFSTQDELLVYNKNQHQTKNLNQDQEHGKANKISYQSSSTEERLHY
 GEKSVQKDVSSSIYSQTEEKAHGSQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GEKGIQKGVSKGSIYSQTEEKAHGSQKQITIPSQEQEHSQKANKISYQSSSTEERLHY
 GEKDVQKGVSKGSIYSQTEEKAHGSQKQITIPSNPNQDQWSGQNAKKGSGQASDSDKQDLLSH
 EQKGRYKQESSSESHNIVITEHEVAQDDHLTQQYNEDRNPIST

Figure 2.10: Manual peptide mapping of SEMGs

LC-MS/MS peptide spectra were each matched to either SEMG1 (light grey), SEMG2 (black), both SEMGs (dark grey) protein sequence. The signal peptide sequence is struck through, indicating that this peptide portion should not have been detected, and it was not. A manual coverage map was generated for gorilla semenogelins as well.

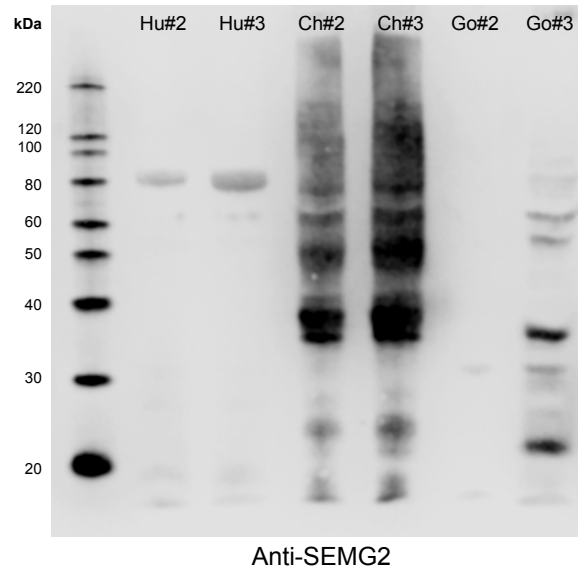


Figure 2.11: Western blot supports LC/MS-MS peptide mapping of SEMG2 in seminal plasma of chimpanzee individuals and one gorilla individual

Human seminal plasma samples bound anti-SEMG2 antibody around 80kDa. Expected weight of human SEMG2 is between 65 -76 kDa, depending on glycosylation. Chimpanzee samples all bound anti-SEMG2 antibody at various molecular weights, as well as one gorilla individual.

2.3.5.2 Nine proteins are shared among human, chimpanzee, and gorilla seminal plasma, with a vast majority of proteins being species specific contributing to distinct proteomes.

A protein was deemed to be present in a species if a peptide was identified in at least two runs and there were at least two unique peptides identified. Utilizing the ‘two-peptide’ rule, 168 proteins were identified with high confidence; 71 seminal plasma proteins were identified for human, 64 proteins for chimpanzee, and 34 proteins for gorilla. Two methods for abundance normalization were used, which are described in section 2.2.5.1.5, the first method does not adjust for length of protein, while the second method is normalized incorporating the length of the protein. To differentiate between each method, “length” will label results normalized with protein length, and “raw” will be relative abundances not normalized by length.

Nine proteins were shared among the three species, which include: albumin, clusterin, keratin 1 and 10, lacto-transferrin, serpin 1, semenogelin I and II, and titin (also known as connectin). Through gene ontology, these core proteins were identified as extracellular structural proteins or enzyme modulators (Panther). Except for the semenogelins, the core protein sequences are conserved (with a $dN/dS < 0.5$), and these proteins are expressed in various tissues throughout the body. There is individual variation within species, evident in SDS-PAGE (Figure 2.12) and more specifically LC/MS-MS (Appendix A.6 tables); however, there are greater differences between species, emphasized by the species-specific banding patterns evident in SDS-PAGE (Figure 2.12). Human seminal plasma has 41 unique proteins compared to chimpanzee and gorilla, while chimpanzee and gorilla have 33 and 24, respectively, unique seminal plasma proteins.

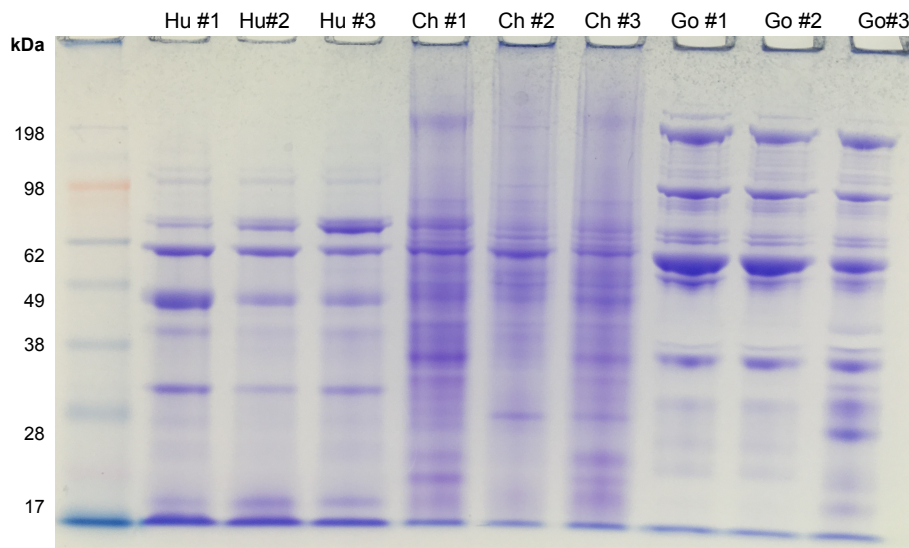


Figure 2.12: Distinct species-specific banding pattern of hominid seminal plasma
Seminal plasma (20 μ g) of three individuals of each species (human, chimpanzee, and gorilla) were separated on a 10% SDS PAGE gel and stained with coomassie.

2.3.5.3 Biodiversity indexes indicate moderate protein richness and evenness in human, chimpanzee, and gorilla seminal plasma without significant differences between species

Both Shannon and Simpson's indexes are typically used to assess biodiversity of species within an ecological community. However, I applied these mathematical calculations to assess the diversity and evenness of proteins within each species seminal plasma proteome (based on three individuals) and for both normalization methods (Table 2.8). Simpson's diversity index is sensitive to abundances of the more plentiful species in a sample, therefore, is associated as a measure of 'dominant' concentration of species (Whittaker, 1965). Simpson index ($1/D$) values are between 0 and 1, where the closer the value is to 1 the greater diversity within the population. Human, chimpanzee, and gorilla seminal plasma proteomes have a relatively high ($1/D > 0.80$ for all samples) Simpson index value, and all three species had similar diversity values. The Simpson and Shannon index values did not change (not significant⁵) much between the length and raw normalization methods.

Normalized abundance quantifications incorporating protein length, appear to fit the data better when calculating Shannon's diversity and evenness statistics, and are more consistent with the Simpson's diversity index. An evenness (E_H) score is between 0 and 1, where the closer the value is to 1 indicates an even distribution of proteins (i.e., there is not a dominant protein contributing to most of the sample). Human, chimpanzee, and gorilla E_H values were all 0.70 ± 0.02 (adjusting for protein length), which indicates a moderately even distribution of proteins. A Shannon index (H) value assesses richness and evenness of diversity within a population, and the higher the value the more species (or proteins) are present, which are also evenly distributed. However, it is argued that Shannon indexes are susceptible to sampling size (or area in

⁵ A simple paired t-test with Welch's correction for both indexes reported an insignificant p-value comparing length and raw normalization statistics (Simpson, $p=0.848$; Shannon, $p=0.819$).

ecological studies); in order to assess diversity indexes across populations, H should be adjusted. A common adjustment is to convert H using the natural log scale, which infers the equivalent or effective numbers of species by calculating e^H . Moreover, e^H values are comparable across populations because they represent the number of equally abundant species required to produce the same value of diversity (Hill, 1972; Peet, 1974; Jost, 2006). H and e^H values are reported for each species with both normalization data sets in Table 2.8. Human and chimpanzee seminal plasma proteomes have similar diversity in the length-adjusted data set ($\sim H=2.92$, $\sim e^H=18.6$), while gorilla has lower diversity in the length-adjusted data set ($H=2.43$, $\sim e^H=11.4$).

Table 2.8: Diversity indexes indicate relatively high protein diversity and moderately even distribution of seminal plasma proteomes for hominids.

Species	Shannon Index		Simpson Index	
	Raw	Length	Raw	Length
Human	$E_H = 0.71$ $H = 3.04$ $e^H = 20.90$	$E_H = 0.69$ $H = 2.92$ $e^H = 18.54$	$1/D = 0.92$	$1/D = 0.90$
Chimpanzee	$E_H = 0.62$ $H = 2.57$ $e^H = 13.07$	$E_H = 0.71$ $H = 2.93$ $e^H = 18.72$	$1/D = 0.80$	$1/D = 0.90$
Gorilla	$E_H = 0.78$ $H = 2.78$ $e^H = 16.12$	$E_H = 0.68$ $H = 2.43$ $e^H = 11.38$	$1/D = 0.92$	$1/D = 0.86$

2.3.5.4 Chimpanzee seminal plasma has increased expression of proteins involved in semen coagulation and prevention of liquefaction

Chimpanzee has multiple serpin (1, 3, and 5) protease inhibitors in a significantly higher concentration than human. Serpin 3 has the largest difference (~ 34 -fold, length) between human and chimpanzee. In addition, only one serpin protease inhibitor, serpin 1, was detected in gorilla, although, it was significantly higher by ~ 3 -fold (length) compared to human (Table 2.9;

Figure 2.13). Prolactin inducible protein (PIP) was not detected in chimpanzee seminal plasma, even though it is high in abundance (~28%, length) in human and has moderate abundance (~4%, length) in gorilla (Table 2.9). In addition, human is significantly 7-fold higher (length) than gorilla seminal plasma (Figure 2.13). Interestingly, Cartilage Acidic Protein 1 (CRTAC1) was found in chimpanzee seminal plasma ~1.3% (length) abundance (Table 2.9), and not detected in either human or gorilla seminal plasma. A follow up Western blot confirmed CRTAC1 abundance in chimpanzee, which was ~100.8-fold higher compared to human (Figure 2.14).

2.3.5.5 Gorilla seminal plasma has loss of proteins involved in semen coagulation and liquefaction pathways

Prostate-specific transglutaminase, TGM4, was not detected in gorilla seminal plasma, which supports pseudogenization shown in genomic studies. Moreover, TGM4 is in a significantly higher concentration (~7-fold, length) in chimpanzee compared to human (Table 2.9). A follow up Western blot confirmed our LC/MS-MS findings, where chimpanzees had ~7-fold higher expression compared to human seminal plasma, and gorilla TGM4 was not detected (Figure 2.15).

Prostatic acid phosphatase (ACPP) was found around 3-4% (length) abundance for both human and chimpanzee seminal plasma, and was not detected in gorilla. There was no significant difference between human and chimpanzee (Table 2.9). A follow up Western confirmed similar abundance between human and chimpanzee ACPP; however, one of the gorilla samples also had ACPP antibody bind around its respective molecular weight, but was much lower in abundance, approximately 17-fold lower than human (Figure 2.15).

Alternatively, gorilla has a significant 4- 6.6 fold increase of albumin compared to chimpanzee and human. Likewise, gorilla has a significant 2-4.5 fold increase of clusterin compared to chimpanzee and human seminal plasma. Additionally, proteins like ARNT (1.6%, length), BCL3 (1.63%, length), CCDC88C (1.96%, length), NDUFS1 (1.41%, length), PAEP (26.74%, length), TSSC1 (2.01%, length), and ZCCHC11 (1.11%, length) are found in relatively high abundance in gorilla but are either not detected or in low abundance in the other species.

Table 2.9: Relative abundance of proteins detected in hominid seminal plasma

Protein ^a	Method ^b	Human Average ^c	Chimp Average ^c	Gorilla Average ^c
ACE	Length	0.00% ± 0.00%	0.16% ± 0.10%	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.43% ± 0.27%*	0.00% ± 0.00%
ACO2	Length	0.00% ± 0.00%	0.04% ± 0.02%	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.07% ± 0.04%	0.00% ± 0.00%
ACOT13	Length	0.00% ± 0.00%	0.53% ± 0.34%*	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.15% ± 0.10%	0.00% ± 0.00%
ACPP	Length	3.86% ± 1.03%*	3.04% ± 0.97%*	0.00% ± 0.00%
	Raw	4.12% ± 0.92%*	2.37% ± 0.78%*	0.00% ± 0.00%
ACR	Length	0.00% ± 0.00%	0.18% ± 0.06%	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.15% ± 0.05%	0.00% ± 0.00%
ACRBP	Length	0.00% ± 0.00%	0.56% ± 0.10%*	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.60% ± 0.12%*	0.00% ± 0.00%
ACTA2	Length	0.54% ± 0.05%*	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.58% ± 0.05%*	0.00% ± 0.00%	0.00% ± 0.00%
ACTB	Length	0.00% ± 0.00%	0.93% ± 0.23%*	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.68% ± 0.14%*	0.00% ± 0.00%
ALB	Length	3.37% ± 0.65%	6.70% ± 1.13%*	24.71% ± 10.51%*
	Raw	5.66% ± 0.81%	8.03% ± 1.29%*	25.91% ± 10.07%*
AMY2B	Length	0.00% ± 0.00%	1.10% ± 0.12%*	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	1.13% ± 0.15%*	0.00% ± 0.00%
ANPEP	Length	0.25% ± 0.11%	0.06% ± 0.06%	0.00% ± 0.00%
	Raw	0.72% ± 0.32%*	0.12% ± 0.12%	0.00% ± 0.00%
APOB	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.34% ± 0.15%*
	Raw	0.07% ± 0.04%	0.00% ± 0.00%	2.71% ± 1.09%*
ARHGAP31	Length	0.02% ± 0.01%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.07% ± 0.04%	0.00% ± 0.00%	0.00% ± 0.00%
ARNT	Length	0.00% ± 0.00%	0.00% ± 0.00%	1.60% ± 0.87%*
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	1.82% ± 0.91%*
ASAP2	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.33% ± 0.24%
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	0.58% ± 0.32%*
AZGP1	Length	2.85% ± 0.57%*	1.53% ± 0.64%*	0.00% ± 0.00%
	Raw	2.43% ± 0.48%*	0.92% ± 0.39%*	0.00% ± 0.00%
B2M	Length	0.82% ± 0.40%*	0.45% ± 0.30%	0.00% ± 0.00%
	Raw	0.24% ± 0.11%	0.11% ± 0.07%	0.00% ± 0.00%
B4GALT1	Length	0.00% ± 0.00%	0.26% ± 0.14%	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.20% ± 0.11%	0.00% ± 0.00%

Protein ^a	Method ^b	Human Average ^c	Chimp Average ^c	Gorilla Average ^c
BCL3	Length	0.00% ± 0.00%	0.00% ± 0.00%	1.63% ± 0.91%*
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	1.09% ± 0.62%*
BIRC6	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.08% ± 0.08%	0.00% ± 0.00%	0.00% ± 0.00%
CAP1	Length	0.15% ± 0.09%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.19% ± 0.10%	0.00% ± 0.00%	0.00% ± 0.00%
CCDC88C	Length	0.00% ± 0.00%	0.00% ± 0.00%	1.96% ± 1.20%*
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	6.14% ± 3.96%*
CEP162	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.53% ± 0.29%*
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	1.38% ± 0.89%*
CFAP69	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.32% ± 0.23%
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	0.68% ± 0.44%*
CKB	Length	0.22% ± 0.15%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.23% ± 0.16%	0.00% ± 0.00%	0.00% ± 0.00%
CLEC11A	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.63% ± 0.63%*
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	0.59% ± 0.59%*
CLIP2	Length	0.03% ± 0.03%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.08% ± 0.08%	0.00% ± 0.00%	0.00% ± 0.00%
CLU	Length	1.85% ± 0.29%	4.54% ± 1.10%	8.27% ± 1.23%*
	Raw	2.36% ± 0.53%	4.57% ± 0.94%	7.09% ± 0.42%*
CPE	Length	0.37% ± 0.19%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.45% ± 0.20%	0.00% ± 0.00%	0.00% ± 0.00%
CPSF1	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.37% ± 0.37%
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	0.74% ± 0.74%*
CRISP1	Length	1.38% ± 0.27%*	0.33% ± 0.16%	0.00% ± 0.00%
	Raw	0.96% ± 0.26%*	0.17% ± 0.08%	0.00% ± 0.00%
CRTAC1	Length	0.00% ± 0.00%	1.35% ± 0.09%*	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	1.76% ± 0.17%*	0.00% ± 0.00%
CST2	Length	0.09% ± 0.09%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.09% ± 0.05%	0.00% ± 0.00%	0.00% ± 0.00%
CST3	Length	0.56% ± 0.32%*	1.79% ± 0.23%*	0.00% ± 0.00%
	Raw	0.19% ± 0.15%	0.52% ± 0.08%*	0.00% ± 0.00%
CST4	Length	1.40% ± 0.53%*	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.57% ± 0.24%*	0.00% ± 0.00%	0.00% ± 0.00%
CTSB	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.40% ± 0.40%
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	0.34% ± 0.34%
CTSD	Length	0.29% ± 0.19%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.32% ± 0.21%	0.00% ± 0.00%	0.00% ± 0.00%

Protein ^a	Method ^b	Human Average ^c		Chimp Average ^c		Gorilla Average ^c	
DEFA1	Length	1.43%	± 0.44%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.38%	± 0.13%*	0.00%	± 0.00%	0.00%	± 0.00%
DHX57	Length	0.01%	± 0.01%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.05%	± 0.05%	0.00%	± 0.00%	0.00%	± 0.00%
DNAH7	Length	0.00%	± 0.00%	0.00%	± 0.00%	0.03%	± 0.03%
	Raw	0.00%	± 0.00%	0.00%	± 0.00%	0.42%	± 0.42%*
DPEP3	Length	0.00%	± 0.00%	0.58%	± 0.05%*	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.56%	± 0.04%*	0.00%	± 0.00%
DUT	Length	0.09%	± 0.05%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.07%	± 0.04%	0.00%	± 0.00%	0.00%	± 0.00%
ECM1	Length	0.46%	± 0.07%*	0.91%	± 0.08%*	0.00%	± 0.00%
	Raw	0.67%	± 0.11%*	0.97%	± 0.07%*	0.00%	± 0.00%
ENO1	Length	0.00%	± 0.00%	0.09%	± 0.09%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.08%	± 0.08%	0.00%	± 0.00%
FAM184B	Length	0.00%	± 0.00%	0.00%	± 0.00%	0.82%	± 0.28%*
	Raw	0.00%	± 0.00%	0.00%	± 0.00%	1.04%	± 0.04%*
FAM3D	Length	0.00%	± 0.00%	0.12%	± 0.12%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.06%	± 0.06%	0.00%	± 0.00%
FN1	Length	0.39%	± 0.16%	1.13%	± 0.22%*	0.00%	± 0.00%
	Raw	2.57%	± 0.94%*	5.25%	± 0.83%*	0.00%	± 0.00%
FOLH1	Length	0.09%	± 0.09%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.15%	± 0.15%	0.00%	± 0.00%	0.00%	± 0.00%
GDF15	Length	0.42%	± 0.42%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.33%	± 0.33%	0.00%	± 0.00%	0.00%	± 0.00%
GOLGB1	Length	0.01%	± 0.01%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.08%	± 0.08%	0.00%	± 0.00%	0.00%	± 0.00%
GPI	Length	0.00%	± 0.00%	0.26%	± 0.09%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.29%	± 0.11%	0.00%	± 0.00%
HEXB	Length	0.23%	± 0.10%	0.12%	± 0.12%	0.00%	± 0.00%
	Raw	0.33%	± 0.12%	0.14%	± 0.14%	0.00%	± 0.00%
HK1	Length	0.00%	± 0.00%	0.19%	± 0.09%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.34%	± 0.17%	0.00%	± 0.00%
HSP90AA1	Length	0.00%	± 0.00%	0.07%	± 0.07%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.11%	± 0.11%	0.00%	± 0.00%
HSPA1A	Length	0.00%	± 0.00%	0.11%	± 0.01%*	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.14%	± 0.02%*	0.00%	± 0.00%
HSPA1L	Length	0.13%	± 0.04%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.23%	± 0.09%	0.00%	± 0.00%	0.00%	± 0.00%

Protein ^a	Method ^b	Human Average ^c		Chimp Average ^c		Gorilla Average ^c	
HSPA5	Length	0.11%	± 0.03%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.20%	± 0.03%*	0.00%	± 0.00%	0.00%	± 0.00%
IDH1	Length	0.17%	± 0.05%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.18%	± 0.05%	0.00%	± 0.00%	0.00%	± 0.00%
IGHG1	Length	0.52%	± 0.26%*	2.21%	± 0.32%*	0.00%	± 0.00%
	Raw	0.47%	± 0.21%	1.44%	± 0.18%*	0.00%	± 0.00%
IGKC	Length	0.00%	± 0.00%	4.03%	± 1.28%*	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.83%	± 0.25%*	0.00%	± 0.00%
KLK3	Length	2.83%	± 0.36%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	2.09%	± 0.29%*	0.00%	± 0.00%	0.00%	± 0.00%
KRT1	Length	0.32%	± 0.08%	0.38%	± 0.20%	2.93%	± 2.23%*
	Raw	0.54%	± 0.09%	0.50%	± 0.27%	3.74%	± 2.75%*
KRT10	Length	0.29%	± 0.23%	0.09%	± 0.09%	0.23%	± 0.23%
	Raw	0.44%	± 0.32%	0.12%	± 0.12%	0.34%	± 0.34%
KRT2	Length	0.47%	± 0.33%	0.08%	± 0.08%	0.00%	± 0.00%
	Raw	0.78%	± 0.53%*	0.11%	± 0.11%	0.00%	± 0.00%
LAMP2	Length	0.31%	± 0.12%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.24%	± 0.13%	0.00%	± 0.00%	0.00%	± 0.00%
LCP1	Length	0.14%	± 0.07%	0.17%	± 0.07%	0.00%	± 0.00%
	Raw	0.40%	± 0.12%	0.22%	± 0.10%	0.00%	± 0.00%
LDHC	Length	0.00%	± 0.00%	0.57%	± 0.57%*	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.39%	± 0.39%*	0.00%	± 0.00%
LGALS3BP	Length	1.19%	± 0.35%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	1.94%	± 0.52%*	0.00%	± 0.00%	0.00%	± 0.00%
LRPPRC	Length	0.01%	± 0.01%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.00%	± 0.00%	0.00%	± 0.00%
LRRK1	Length	0.00%	± 0.00%	0.00%	± 0.00%	0.24%	± 0.24%
	Raw	0.00%	± 0.00%	0.00%	± 0.00%	0.72%	± 0.72%*
LTF	Length	3.53%	± 0.68%	2.96%	± 1.74%	3.48%	± 1.39%
	Raw	7.27%	± 1.80%	3.99%	± 2.21%	4.25%	± 1.38%
MDH2	Length	0.00%	± 0.00%	0.48%	± 0.48%*	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.34%	± 0.34%	0.00%	± 0.00%
MGAM	Length	0.00%	± 0.00%	0.05%	± 0.03%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.18%	± 0.10%	0.00%	± 0.00%
MME	Length	0.00%	± 0.00%	0.04%	± 0.04%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.06%	± 0.06%	0.00%	± 0.00%
MSMB	Length	1.61%	± 1.14%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.46%	± 0.31%*	0.00%	± 0.00%	0.00%	± 0.00%

Protein ^a	Method ^b	Human Average ^c		Chimp Average ^c		Gorilla Average ^c	
MUC6	Length	0.37%	± 0.09%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	2.53%	± 0.62%*	0.00%	± 0.00%	0.00%	± 0.00%
NBPF3	Length	0.00%	± 0.00%	0.00%	± 0.00%	0.94%	± 0.78%*
	Raw	0.00%	± 0.00%	0.00%	± 0.00%	0.79%	± 0.57%*
NDUFS1	Length	0.00%	± 0.00%	0.00%	± 0.00%	1.41%	± 0.97%*
	Raw	0.00%	± 0.00%	0.00%	± 0.00%	1.51%	± 0.87%*
NEMF	Length	0.00%	± 0.00%	0.00%	± 0.00%	0.13%	± 0.13%
	Raw	0.00%	± 0.00%	0.00%	± 0.00%	0.34%	± 0.34%
NPC2	Length	3.17%	± 0.93%*	6.58%	± 0.76%*	0.00%	± 0.00%
	Raw	1.36%	± 0.40%*	1.98%	± 0.28%*	0.00%	± 0.00%
ORM1	Length	0.24%	± 0.24%	0.40%	± 0.40%	0.00%	± 0.00%
	Raw	0.14%	± 0.14%	0.17%	± 0.17%	0.00%	± 0.00%
OXCT2	Length	0.00%	± 0.00%	0.12%	± 0.07%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.13%	± 0.07%	0.00%	± 0.00%
PAEP	Length	0.00%	± 0.00%	1.50%	± 0.79%*	26.74%	± 5.05%*
	Raw	0.00%	± 0.00%	0.55%	± 0.28%*	10.68%	± 3.61%*
PATE1	Length	0.53%	± 0.53%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.21%	± 0.21%*	0.08%	± 0.08%	0.00%	± 0.00%
PGAM2	Length	0.00%	± 0.00%	0.16%	± 0.16%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.00%	± 0.00%	0.00%	± 0.00%
PGC	Length	0.22%	± 0.06%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.26%	± 0.06%*	0.00%	± 0.00%	0.00%	± 0.00%
PGK2	Length	0.00%	± 0.00%	0.19%	± 0.14%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.15%	± 0.11%	0.00%	± 0.00%
PIP	Length	28.40%	± 6.23%*	0.00%	± 0.00%	3.71%	± 3.71%*
	Raw	11.80%	± 3.22%*	0.00%	± 0.00%	1.40%	± 1.40%*
PKM	Length	0.00%	± 0.00%	0.48%	± 0.24%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.49%	± 0.25%*	0.00%	± 0.00%
PLA1A	Length	0.00%	± 0.00%	0.17%	± 0.13%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.15%	± 0.11%	0.00%	± 0.00%
PLS3	Length	0.00%	± 0.00%	0.10%	± 0.06%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.12%	± 0.07%	0.00%	± 0.00%
PPIB	Length	0.00%	± 0.00%	1.07%	± 0.48%	1.57%	± 1.57%
	Raw	0.00%	± 0.00%	0.47%	± 0.21%	0.90%	± 0.90%
PRDX6	Length	0.71%	± 0.32%	0.91%	± 0.28%	0.00%	± 0.00%
	Raw	0.42%	± 0.17%	0.39%	± 0.11%	0.00%	± 0.00%
PSAP	Length	1.07%	± 0.20%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	1.52%	± 0.18%*	0.00%	± 0.00%	0.00%	± 0.00%

Protein ^a	Method ^b	Human Average ^c		Chimp Average ^c		Gorilla Average ^c	
PSCA	Length	0.00%	± 0.00%	2.67%	± 0.56%*	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.66%	± 0.15%*	0.00%	± 0.00%
QSOX1	Length	0.09%	± 0.06%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.18%	± 0.11%	0.00%	± 0.00%	0.00%	± 0.00%
RLTPR	Length	0.00%	± 0.00%	0.03%	± 0.03%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.08%	± 0.08%	0.00%	± 0.00%
RNASET2	Length	0.00%	± 0.00%	1.68%	± 0.06%*	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.84%	± 0.01%*	0.00%	± 0.00%
S100A8	Length	1.13%	± 0.72%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.33%	± 0.21%*	0.00%	± 0.00%	0.00%	± 0.00%
SEMG1	Length	10.96%	± 5.32%*	28.61%	± 2.50%*	1.90%	± 1.90%
	Raw	13.88%	± 6.73%*	43.62%	± 2.70%*	3.25%	± 3.25%
SEMG2	Length	12.46%	± 1.60%*	1.86%	± 0.63%	7.44%	± 7.44%*
	Raw	20.29%	± 2.00%*	1.47%	± 0.44%	10.58%	± 10.58%*
SERPINA1	Length	0.90%	± 0.03%	2.24%	± 0.53%*	3.08%	± 1.99%*
	Raw	1.05%	± 0.09%	1.88%	± 0.47%*	2.50%	± 1.37%*
SERPINA3	Length	0.06%	± 0.06%	2.16%	± 0.54%*	0.00%	± 0.00%
	Raw	0.08%	± 0.08%	1.84%	± 0.50%*	0.00%	± 0.00%
SERPINA5	Length	1.70%	± 0.45%*	4.29%	± 0.81%*	0.00%	± 0.00%
	Raw	1.91%	± 0.39%*	3.47%	± 0.71%*	0.00%	± 0.00%
SERPINF2	Length	0.13%	± 0.10%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.19%	± 0.15%	0.00%	± 0.00%	0.00%	± 0.00%
SLC9B2	Length	0.08%	± 0.04%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.13%	± 0.08%	0.00%	± 0.00%	0.00%	± 0.00%
SMG1	Length	0.01%	± 0.01%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.08%	± 0.08%	0.00%	± 0.00%	0.00%	± 0.00%
SOD2	Length	0.00%	± 0.00%	0.15%	± 0.08%	0.00%	± 0.00%
	Raw	0.00%	± 0.00%	0.07%	± 0.03%	0.00%	± 0.00%
SOD3	Length	0.13%	± 0.13%	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.07%	± 0.07%	0.00%	± 0.00%	0.00%	± 0.00%
SORD	Length	0.21%	± 0.05%*	0.00%	± 0.00%	0.00%	± 0.00%
	Raw	0.19%	± 0.04%*	0.00%	± 0.00%	0.00%	± 0.00%
SRCAP	Length	0.00%	± 0.00%	0.00%	± 0.00%	0.10%	± 0.07%
	Raw	0.00%	± 0.00%	0.00%	± 0.00%	0.52%	± 0.32%*
TF	Length	0.85%	± 0.32%*	0.71%	± 0.38%*	0.00%	± 0.00%
	Raw	1.64%	± 0.55%*	0.95%	± 0.48%*	0.00%	± 0.00%
TGM4	Length	0.19%	± 0.06%	1.35%	± 0.60%*	0.00%	± 0.00%
	Raw	0.36%	± 0.13%	1.88%	± 0.87%*	0.00%	± 0.00%

Protein ^a	Method ^b	Human Average ^c	Chimp Average ^c	Gorilla Average ^c
TIMP1	Length	0.96% ± 0.22%*	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.56% ± 0.12%*	0.00% ± 0.00%	0.00% ± 0.00%
TKFC	Length	0.00% ± 0.00%	0.05% ± 0.05%	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.06% ± 0.06%	0.00% ± 0.00%
TMPRSS2	Length	0.17% ± 0.12%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.25% ± 0.20%	0.00% ± 0.00%	0.00% ± 0.00%
TRRAP	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.09% ± 0.06%
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	0.49% ± 0.28%*
TSSC1	Length	0.00% ± 0.00%	0.00% ± 0.00%	2.01% ± 1.76%*
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	1.28% ± 0.99%*
TTLL5	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.41% ± 0.41%
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	1.11% ± 1.11%*
TTN	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.01% ± 0.01%
	Raw	0.20% ± 0.14%	0.07% ± 0.04%	0.70% ± 0.35%*
TUBA1A	Length	0.00% ± 0.00%	0.24% ± 0.13%	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.22% ± 0.12%	0.00% ± 0.00%
TUBB	Length	0.00% ± 0.00%	0.12% ± 0.12%	0.00% ± 0.00%
	Raw	0.00% ± 0.00%	0.11% ± 0.11%	0.00% ± 0.00%
UBA1	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.47% ± 0.24%
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	0.73% ± 0.36%*
UBA52	Length	1.35% ± 0.68%*	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.52% ± 0.27%*	0.00% ± 0.00%	0.00% ± 0.00%
WFDC2	Length	0.68% ± 0.27%*	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.24% ± 0.09%	0.00% ± 0.00%	0.00% ± 0.00%
WTIP	Length	0.06% ± 0.06%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.08% ± 0.08%	0.00% ± 0.00%	0.00% ± 0.00%
YWHAB	Length	0.33% ± 0.33%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.19% ± 0.19%	0.00% ± 0.00%	0.00% ± 0.00%
ZCCHC11	Length	0.00% ± 0.00%	0.00% ± 0.00%	1.11% ± 0.75%*
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	2.94% ± 1.98%*
ZFH2	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.05% ± 0.05%
	Raw	0.00% ± 0.00%	0.00% ± 0.00%	0.34% ± 0.34%
ZNF148	Length	0.00% ± 0.00%	0.00% ± 0.00%	0.00% ± 0.00%
	Raw	0.07% ± 0.07%	0.00% ± 0.00%	0.00% ± 0.00%

a) Proteins listed in this table were detected with high confidence (utilizing the ‘two-peptide’ rule) and are organized by alphabetical order **b)** Normalization method (length adjusted or raw) is specified in this column for the adjacent relative abundance values. **c)** Values in this column represent species relative protein abundance with standard error of the mean (among three individuals) and are different depending on normalization method. Values with a “*” were significantly higher in abundance compared to one or both species. Refer to Table A.9 for statistical comparison.

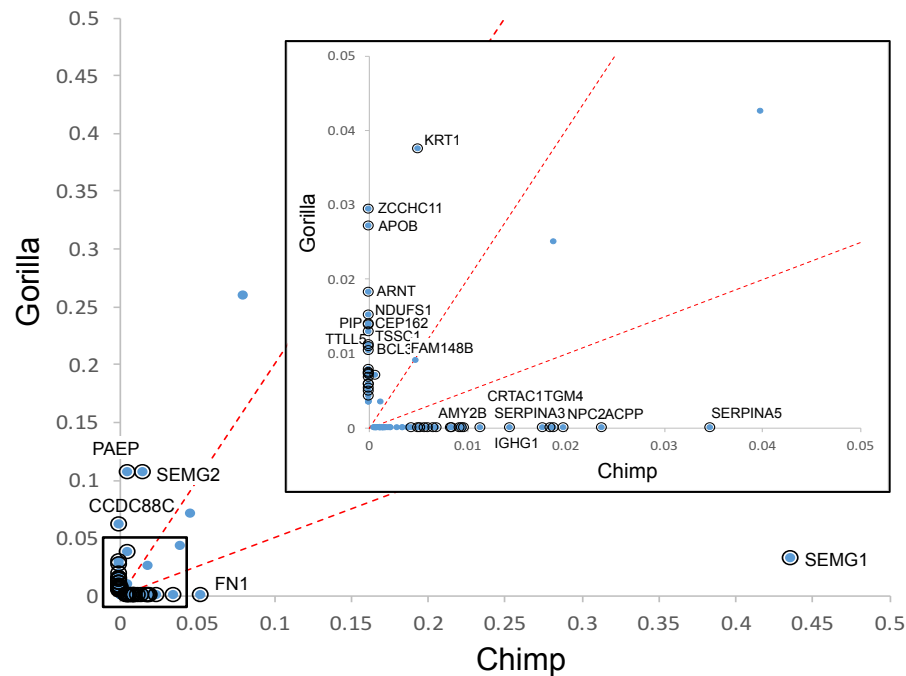
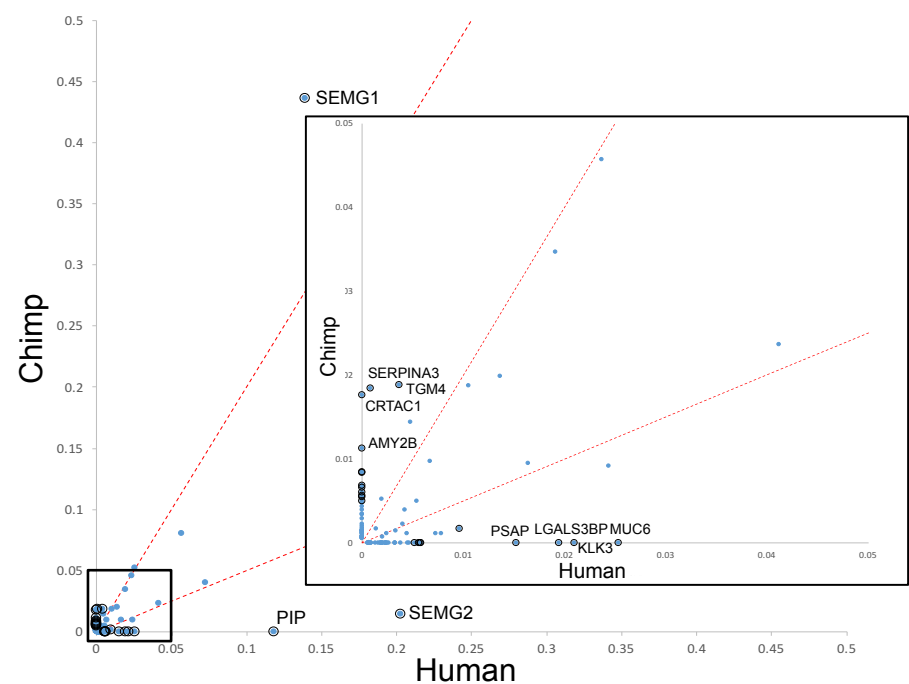
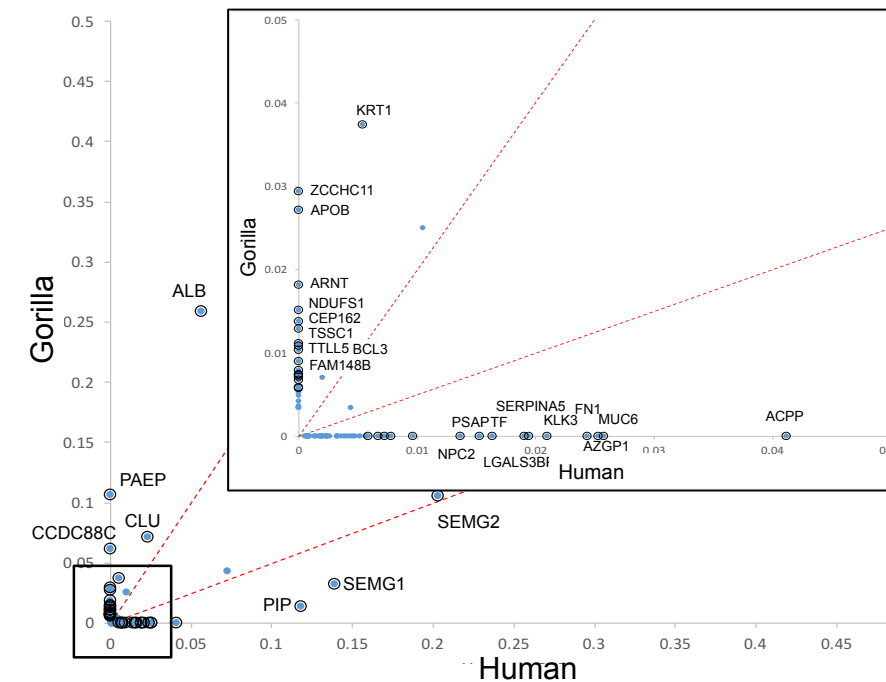


Figure 2.13: Pairwise PLGEM graphs with raw normalization
 Normalized spectral abundance frequencies (NSAF) of high-confidence proteins identified in human, chimpanzee, and gorilla seminal plasma. Dashed lines indicate the margins of two-fold difference in abundance. Circled points indicate those proteins whose abundance differed significantly between these species, as inferred from PLGEM analysis and applying a 1% FDR. For clarity, *inset plots* show proteins whose NSAF values are less than 0.05 for both species, indicated by the black square in the larger plot; this inset plot does not obscure any data points in the larger plot. Names are given for proteins with significant differences and found at NSAF values greater than 0.01.

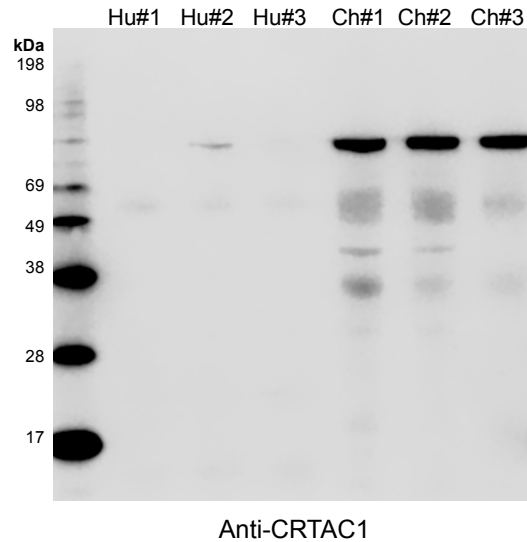


Figure 2.14: Chimpanzee seminal plasma has an excess of CRTAC1

Seminal plasma (20 μ g) of three individuals of human and chimpanzee were separated on a 10% SDS PAGE and proteins were transferred to a PVDF and Western blotted with an anti-CRTAC1 antibody. Only one of three human seminal plasma samples bound to anti-CRTAC1 antibody while all three chimpanzee seminal plasma samples bound to anti-CRTAC1 (100.8-fold higher than human).

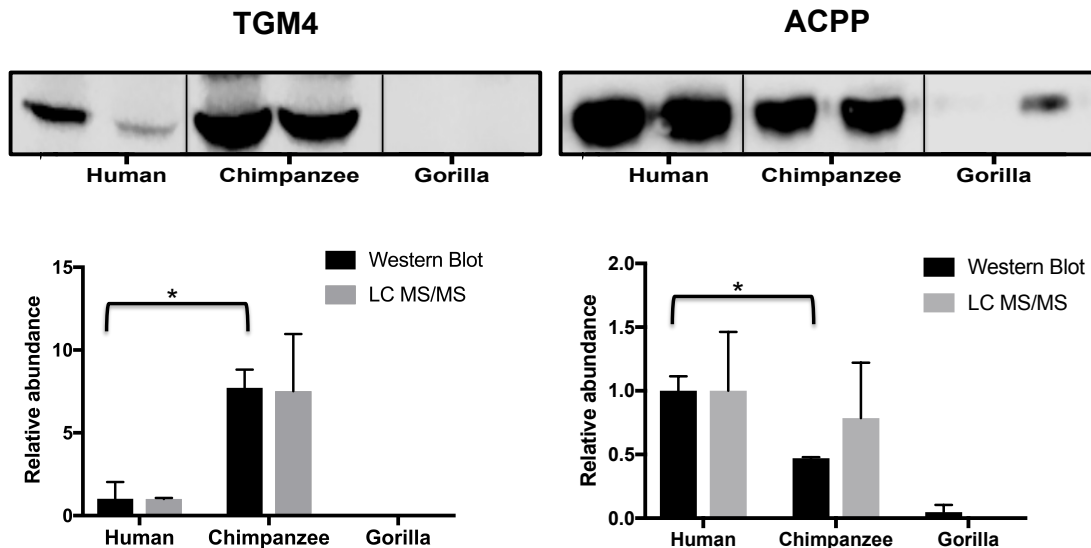


Figure 2.15: Western blots confirm LC/MS-MS results

In both LC-MS/MS and Western blot methods, TGM4 was not detected in gorilla seminal plasma; chimpanzee seminal plasma had 7-fold higher expression than human that was significant via Western blot (and significant using PLGEM for LC/MS-MS). ACPP was only detected in only one gorilla's seminal plasma via Western blot and was not detected in any gorilla individual's seminal plasma via LC-MS/MS. ACPP is expressed 1.5 fold higher in human seminal plasma compared to chimpanzee, which was significant via Western blot (* $p < 0.05$, Holm-Sidak method).

2.4 Discussion

Seminal plasma is a complex mixture of proteins, sugars, lipids, and ions. Our LC-MS/MS data in conjunction with the use of biodiversity indexes support that the seminal plasma proteome is diverse (Table 2.8; 2.9). There were no significant differences in overall diversity of proteins between human, chimpanzee, and gorilla seminal plasma. In addition, gene ontology analysis did not reveal any species-specific differences in proteome molecular function, biological process, or protein classes (further analyzed in Appendix A.7).

We observed a statistically significant difference in average concentration of chimpanzee seminal plasma (14.20 ± 1.63 mg/mL) compared to human seminal plasma (9.44 ± 1.21 mg/mL). This difference is likely attributed to sample collection, such as pooling semen samples from the same individual, and all subsequent reactions (LC/MS-MS and Western blots) had the same total protein concentration loaded for each sample. Therefore detection and relative abundance differences were not accounted for by total protein content. There were over >50% (detection only, not factoring in abundance) of proteins identified being unique to each species (human 41/71; chimpanzee 33/64; gorilla 24/34) and separation of seminal plasma in SDS-PAGE has species-specific banding patterns (Figure 2.12). It is abundantly clear that these species proteomes are quite different. The challenge is interpreting which differences are responding to sexual selective pressures and generating the different semen consistencies observed in these species. The following subsections will focus on seemingly striking differences which functions are unknown or biologically relevant to fertilization or semen coagulation. It is likely that our analysis may have missed proteins that play an important role, which are not highly abundant nor significantly different between species.

2.4.1 Semenogelins

SEMG1 and SEMG2 are among the most abundant proteins in human seminal plasma, together accounting for 23.4% of MS spectra identified across the three human samples (Table 2.7). Similarly, these two proteins account for 30.9% of spectra from chimpanzees. However, whereas SEMG1 and SEMG2 are found in nearly equal abundance in human semen, nearly all of the semenogelin protein in chimpanzees is derived from the longer *SEMG1* gene (Figure 1.5:C). Considering previous studies have shown a fixed premature stop codon in the *SEMG2* gene (Figure 1.5:D; Jensen-Seaman and Li, 2003) for chimpanzees, it was surprising that SEMG2 was even detected. However, another study analyzing a single chimpanzee individual also detected SEMG2 with an approximate abundance of 2.8% (and ~15.9% of SEMG1) in a separate lab run under their conditions (Claw, 2013).

Western blots of seminal plasma revealed high-molecular weight targets of anti-semenogelin antibodies especially in chimpanzees (Figure 2.11). This is consistent with covalent cross-linking of semenogelin monomers to form large insoluble protein complexes. It is therefore possible that the high-molecular weight proteins bound by anti-SEMG2 and seen in our Western blots are similarly produced. Among the three species examined in this study, chimpanzees are the only one to produce a firm copulatory plug (Dixson and Anderson, 2001). In a proteomic study analyzing the copulatory plug, a common phenotype in mice, 63 different proteins were identified (Dean et al., 2011), including the semenogelins paralog, SVS2. It is likely that the detection of high molecular weight semenogelins is a result of their cross-linking interaction with themselves and other unknown proteins.

With both LC-MS/MS and Western blotting, only one of the three gorilla individuals sampled possessed peptides derived from the semenogelin genes, although in this individual the

amount produced was substantial (Table 2.7; Figure 2.11), ~28% relative abundance. This variation of expression among gorillas correlates to the presence of polymorphic stop codons in both their semenogelin genes (Figure 1.5:D).

Western blots of all three hominid species also show smaller, low molecular weight peptides when probed with anti-SEMG1 or anti-SEMG2 (Figure 2.11), consistent with degradation by endogenous seminal proteases, as seen in other studies (Lwaleed et al., 2004; de Lamirande, 2007). Attempts to use protease inhibitors to prevent or stop degradation were largely unsuccessful (Figure 2.8, 2.9), indicating that at least a large amount of this proteolytic degradation is happening almost immediately upon ejaculation.

2.4.2 Up-regulation and pseudogenization of seminal coagulation pathway

Serpina1, Serpina3, Serpina5, SEMG1, Fibronectin (FN1), CRTAC1, and TGM4 were expressed higher in chimpanzees compared to humans and have functions related to semen coagulation and copulatory plug formation (Figure 2.16). Serpina5, or Protein C Inhibitor, is known to inactivate Kallikrein related peptidase 3 (KLK3) (Lwaleed et al., 2004). KLK3 is an important serine protease involved in cleavage of semenogelin peptides to release spermatozoa, which also liquefies semen. The up regulation of serine protease inhibitors may delay the dissolution of the copulatory plug formed in chimpanzees.

SEMG1 is a known substrate of TGM4, and is one of the major components, along with FN1, in forming viscosity, or the copulatory plug in chimpanzees. Cross-linking of the rodent paralogs SVS1, 2 and 3 to form a copulatory plug is mediated by the rodent TGM4 paralog (Tseng et al., 2008, 2011). *Tgm4* or *Svs2* knockout male mice are unable to form a copulatory plug, indicating that both are required, even though additional proteins are present in the copulatory plug (Dean et al., 2011; Dean et al., 2013; Kawano et al., 2014). CRTAC1 is an

extracellular matrix protein in cartilage tissue, and may help form the copulatory plug; however, functional studies would need to support this hypothesis.

Prolactin inducible protein (PIP) was not detected in our chimpanzee samples, though it was in moderate to high abundance in human and gorilla (Table 2.9). However, PIP was detected in one chimpanzee individual in another LC/MS-MS data set around 2.5% abundance (Claw, 2013). Human and mouse tissue expression data suggest that PIP is expressed from the prostate (Unigene). PIP has been shown to cleave FN1 (Caputo, 2000), which may aid in the dissolution of the copulatory plug. Subsequently, PIPs down regulation or absence in chimpanzee semen may be connected to its interaction with FN1.

Multiple proteins in this pathway were not detected in gorilla, such as: TGM4, KLK3, Serpina3, and Serpina5 (Figure 2.16). In addition, the SEMGs were only found in one individual, and ACPP was only detected in one individual via Western blot (Figure 2.15). As discussed in Chapter 1, gorilla ejaculate is “liquidy” and in a smaller volume compared to human and chimpanzee. This semen phenotype may be a result of gene loss involved in the coagulation pathway. Gorilla has increased expression of albumin and clusterin compared to human and chimpanzees, and possibly these two proteins are making up the consistency of gorilla semen.

Intriguingly, Progesterone-associated endometrial protein (PAEP) nearly accounts for ~26% (length-adjusted) of the gorilla seminal plasma proteome. PAEP is significantly higher in abundance in gorilla than chimpanzee (~1.5%, length adjusted), and PAEP was not detected in human at all. However, other studies have detected PAEP in human seminal fluid and there is some gene expression of PAEP in human testes, although expression is mainly associated with human mammary glands, uterus, and cervix (Unigene). PAEP is the official gene symbol, but has numerous identifiers in literature, the most common being Glycodelin, which is indicative of it

being a glycoprotein. PAEP has four forms, with the same protein backbone but different glycosylation profiles, likely influencing its function and are named after their location (Glycodelin-A; amniotic fluid, Glycodelin-C; cumulus cells, Glycodelin-F; follicular fluid, Glycodelin-S; seminal fluid). These forms have essential roles in regulating a uterine environment for embryo implantation and development (Uchida et al., 2013). Glycodelin-S, likely the form we detected from gorilla and chimpanzee seminal plasma, binds spermatozoa immediately after ejaculation, subsequently preventing capacitation, until is disassociated (Uchida et al., 2013). This increased abundance of PAEP in gorilla seminal plasma is interesting, and perhaps it is playing a role in delaying capacitation of spermatozoa to lengthen the viability time of sperm, as capacitation is required for fertilization of the ova, but sperm are not functional two hours post capacitation (Cohen-Dayag et al., 1995).

Only SEMG2 and PIP are expressed in higher abundance in human compared to the other species (Figure 2.16). However, human has maintained expression of all the proteins involved in the semen coagulation and liquefaction pathway. This is not surprising, as human ejaculate is moderate compared to chimpanzee and gorilla, and retaining proteins in both coagulation and liquefaction would lead to this phenotype.

In conclusion, chimpanzee has an up regulation of proteins that are likely involved in semen coagulation and copulatory plug formation. Likewise, serine protease inhibitors have increased expression that may slow the dissolution of the copulatory plug, while proteases (like KLK3 and ACPP) are absent or under-expressed. Copulatory plug formation is thought to be advantageous, by either aiding in fertilization or preventing subsequent males sperm from fertilizing the females egg. Likely this seminal phenotype is a response to increased sperm competition due to their multi-male/multi-female mating system. Contrarily, gorillas with

relatively low sperm competition have many proteins lost in both the coagulation and liquefaction pathways. This may be due to a relaxation of selective pressures on semen, while other sexual selective pressures are at play increasing their ability to combat other males for access to females. The high abundance of PAEP in gorilla seminal plasma, was unexpected, but likely is important in the sperm capacitation process and timing of fertilization, which may be unique to gorillas. Human, with perceived moderate sperm competition, has all proteins involved in the semen coagulation and liquefaction pathway, likely contributing to a comparatively average semen consistency and proteome.

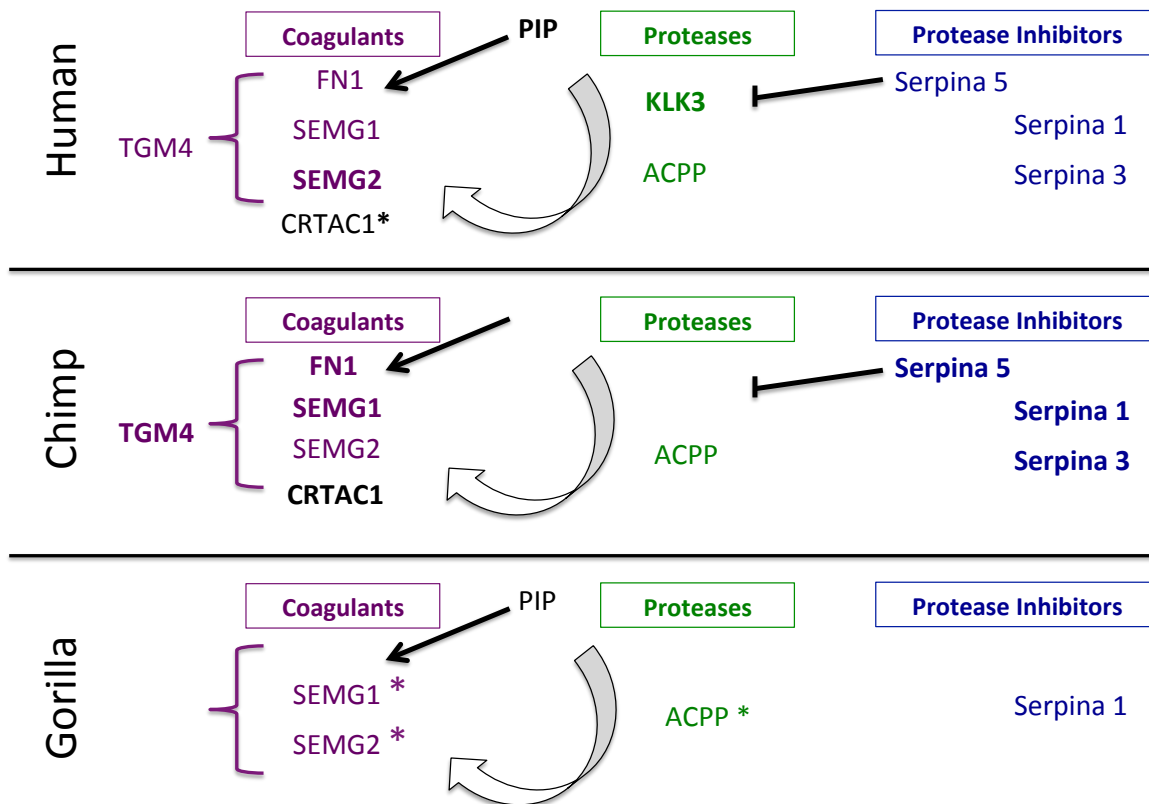


Figure 2.16: Expression differences in seminal coagulation and liquefaction pathways of hominds: human, chimpanzee, and gorilla

Proteins involved in the semen coagulation (TGM4 and “coagulants” proteins) and semen liquefaction (“proteases”; “protease inhibitors”) are represented for each species: human, chimpanzee, and gorilla. Proteins are color-coded based on their function. Proteins that are black, PIP and CRTAC1, have unknown functions directly related to semen coagulation/liquefaction pathways, however related functions and observed differences in abundance suggest involvement in the liquefaction and coagulation pathway, respectively. PIP has been shown to have proteolytic activity on fibronectin *in vitro*, therefore, its unknown function in this pathway is likely associated with cleaving fibronectin and liquefaction of semen. CRTAC1 is an extracellular matrix protein involved in cartilage formation and may be a coagulant in semen. All the proteins were detected in human seminal plasma (either by LC/MS-MS or Western blot). Subsequently, protein names that are listed in human and not listed in another species indicate that they were not detected via LC/MS-MS in that species. Protein names, which are bolded, are in the species that had the highest significant expression of that particular protein compared to other species. Protein names with an *, were not identified in all individuals of the same species. Arrows and blunt end lines have either proteolytic or inhibitory activity, respectively, on the proteins that they are pointed to.

2.5 Future Directions

It is strikingly clear that there are differences among these seminal plasma proteomes, and the presence of all the proteins detected in this study likely have some physiologic function in seminal plasma which is a result of multiple selective forces. In essence, there are 171 different future directions this study could continue – the same number of proteins identified with high confidence. This study and previous studies suggest that the proteins involved in semen coagulation and liquefaction (Figure 2.16) may be of most interest. Specifically, I followed up with TGM4 (Chapter 3), but recently I have been thinking about the importance of proteins unique to one species and not the other. Where did they come from? What is their function?

Anne-Ruxandra Carvunis coined the term ‘proto-gene hypothesis’, which proposes that *de novo* genes arise from ancestral promoters driving constitutive expression of non-genic sequences (2012). The constitutive promoters are signatures left from ancestral genes, which have become pseudogenized or lost. The nongenic short transcripts in front of these promoters are quickly degraded if they do not provide a benefit, but if they are functionally advantageous, can rapidly evolve into a functional gene generating a novel protein in a species. In essence, a gene is preserved until it is no longer needed; then it becomes a pseudogene, in which its sequence can primitively become a new gene with a different function. I hypothesize that the pseudogenization events in both *SEMG1* and *SEMG2* of gorilla have occurred relatively recently, and in accordance with the model proposed by Carvunis, the promoter is still very active—as evidenced by the high expression in one gorilla who lacks a premature stop codon. If possible, I would like to sequence the DNA of the same gorilla individuals, whose semen was assayed, to determine: 1) if they had premature stop codons corresponding to their proteomic expression and

2) if their promoter sequences were conserved. I would then follow up with an in vitro model whereby reporter constructs carrying the promoter sequence found in front of expressed and non-expressed SEMG genes are transfected into mammalian cells in culture to assay their promoter strength. If possible, SEMG mRNA transcripts could be analyzed through RT-PCR of gorilla testes tissue.

CRTAC1 expression in chimpanzee seminal plasma and its relative high abundance is intriguing to me. I have hypothesized that its roles in cartilage formation may have functions in the construction of the copulatory plug, but how, I am unsure. There are a few pilot experiments I would design to assay its importance. First, is CRTAC1, or its paralogs present in other species, known to form copulatory plugs? If in mice, do *Crtac1* knockout mice produce a plug? Or are male mice unable to form plugs like *Svs2* and *Tgm4* knockout mice? Second, is CRTAC1 a substrate for prostate specific transglutaminase? I would use either recombinant chimpanzee CRTAC1 or commercially available CRTAC1 (most likely human) in the TGM4 assays to determine if CRTAC1 is cross-linked.

I recently went to a lecture where Mark Chance, Ph.D., was presenting on his team's innovative project, CrosstalkerTM, which is a bioinformatics tool to search interactive gene-protein networks. He claimed that users could input a list of genes, proteins, or metabolites, and search multiple connected resources, like String, to identify additional genes/proteins/metabolites involved in associated pathways. The generation of these large 'Omics' data sets is a tremendous advancement in research; however, the bioinformatic tools and expertise for analysis is lagging behind. I think there are still biologically interesting proteins which are present in these data but not resolved through our current analysis. The largest future direction of this project should be to utilize these data as a resource for future bioinformatic tools to identify interesting leads.

2.6 References

- Ambekar, Aditi S., et al. "Proteomic Analysis of Human Follicular Fluid: A New Perspective towards Understanding Folliculogenesis." *Journal of Proteomics*, vol. 87, 2013, pp. 68–77.
- Benjamini, Yoav, and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995, pp. 289–300.
- Birkhead, Timothy R., and Tommaso Pizzari. "Post Copulatory Sexual Selection." *Nature*, vol. 3, 2002, pp. 262–73.
- Caputo, Emilia, et al. "A Novel Aspartyl Proteinase from Apocrine Epithelia and Breast Tumors." *Journal of Biological Chemistry*, vol. 275, no. 11, 2000, pp. 7935–41.
- Carnahan, Sarah J., and Michael I. Jensen-Seaman. "Hominoid Seminal Protein Evolution and Ancestral Mating Behavior." *American Journal of Primatology*, vol. 70, 2008, pp. 939–48.
- Carnahan-Craig, S. J., and M. I. Jensen-Seaman. "Rates of Evolution of Hominoid Seminal Proteins Are Correlated with Function and Expression, Rather than Mating System." *Journal of Molecular Evolution*, Nov. 2013, pp. 1–13, doi:10.1007/s00239-013-9602-z.
- Carvunis, Anne-Ruxandra, et al. "Proto-Genes and de Novo Gene Birth." *Nature*, vol. 487, no. 7407, 2012, pp. 370–74.
- Castric, Peter, et al. "Structural Characterization of the *Pseudomonas aeruginosa* 1244 Pilin Glycan." *Journal of Biological Chemistry*, vol. 276, no. 28, 2001, pp. 26479–85.
- Clark, Nathaniel L., and Willie J. Swanson. "Pervasive Adaptive Evolution in Primate Seminal Protein." *PLoS Genetics*, vol. 1, no. 35, 2005, pp. 335–42.
- Claw, Katrina G. *Proteomic Identification and Evolutionary Analysis of Primate Reproductive Proteins*. University of Washington, 2013.
- de Lamirande, Eve. "Semenogelin, the Main Protein of the Human Semen Coagulum, Regulates Sperm Function." *Seminars in Thrombosis and Hemostasis*, vol. 33, no. 1, 2007, pp. 60–68.
- Dean, Matthew D. "Genetic Disruption of the Copulatory Plug in Mice Leads to Severely Reduced Fertility." *PLoS Genet*, vol. 9, no. 1, Jan. 2013, p. e1003185, doi:10.1371/journal.pgen.1003185.
- Dean, Matthew D., et al. "Identification of Ejaculated Proteins in the House Mouse (*Mus Domesticus*) via Isotopic Labeling." *BMC Genomics*, vol. 12, no. 1, 2011, p. 306.
- Dean, Matthew D., et al. "Proteomics and Comparative Genomic Investigations Reveal Heterogeneity in Evolutionary Rate of Male Reproductive Proteins in Mice (*Mus Domesticus*)." *Mol. Biol. Evol.*, vol. 26, no. 8, 2009, pp. 1733–43.
- Dixon, Alan, and Matthew Anderson. "Sexual Selection and the Comparative Anatomy of Reproduction in Monkeys, Apes, and Human Beings." *Annual Review of Sex Research*, vol. 12, no. 1, 2001, pp. 121–44.
- Dixon, Alan F. "Evolutionary Perspectives on Primate Mating Systems and Behavior." *Annals New York Academy of Sciences*, vol. 807, 1997, pp. 21–61.
- Dixon, Alan F., and Matthew J. Anderson. "Sexual Behavior, Reproductive Physiology Sperm Competition in Male Mammals." *Physiology and Behavior*, vol. 15, no. 83, 2004, pp. 361–71.
- Dorus, Steve, et al. "Rate of Molecular Evolution of the Seminal Protein Gene SEMG2 Correlates with Levels of Female Promiscuity." *Nature Genetics*, vol. 36, no. 12, 2004, pp. 1326–29.
- Drake, Richard R., et al. "In-Depth Proteomic Analyses of Direct Expressed Prostatic Secretions." *Journal of Proteome Research*, vol. 9, no. 5, 2010, pp. 2109–16.
- Findlay, Geoffrey D., and Willie J. Swanson. "Proteomics Enhances Evolutionary and Functional Analysis of Reproductive Proteins." *Bioessays*, vol. 32, 2010, pp. 26–36.

- Fung, Kim YC, et al. "A Comprehensive Characterization of the Peptide and Protein Constituents of Human Seminal Fluid." *The Prostate*, vol. 61, no. 2, 2004, pp. 171–81.
- Good, Jeffrey M., et al. "Comparative Population Genomics of the Ejaculate in Humans and the Great Apes." *Molecular Biology and Evolution*, vol. 30, no. 4, 2013, pp. 964–76.
- Haerty, Wilfried, et al. "Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in *Drosophila*." *Genetics*, vol. 177, no. 3, 2007, pp. 1321–35.
- Harcourt, Alexander H., et al. "Testis Weight, Body Weight and Breeding System in Primates." *Nature*, vol. 293, no. 5827, 1981, pp. 55–57.
- Hill, Mark O. "Diversity and Evenness: A Unifying Notation and Its Consequences." *Ecology*, vol. 54, no. 2, 1973, pp. 427–32.
- Jensen-Seaman, Michael I., and Wen-Hsiung Li. "Evolution of the Hominoid Semenogelin Genes, the Major Proteins of Ejaculated Semen." *Journal of Molecular Evolution*, vol. 57, 2003, pp. 261–70.
- Jost, Lou. "Entropy and Diversity." *Oikos*, vol. 113, no. 2, 2006, pp. 363–75.
- Karr, Timothy L., and Scott Pitnick. "Sperm Competition: Defining the Rules of Engagement." *Current Biology*, vol. 9, no. 20, Oct. 1999, pp. R787–90, doi:10.1016/S0960-9822(00)80014-7.
- Kawano, Natsuko, et al. "Seminal Vesicle Protein SVS2 Is Required for Sperm Survival in the Uterus." *Proceedings of the National Academy of Sciences*, vol. 111, no. 11, 2014, pp. 4145–50.
- Kosiol, Carolin, et al. "Patterns of Positive Selection in Six Mammalian Genes." *PLoS Genetics*, vol. 4, no. 8, 2008.
- Lilja, Hans, et al. "Semenogelin, the Predominant Protein in Human Semen. Primary Structure and Identification of Closely Related Proteins in the Male Accessory Sex Glands and on the Spermatozoa." *J Biol Chem*, vol. 264, 1989, pp. 1894–900.
- Lwaleed, Bashir, et al. "Seminal Clotting and Fibrinolytic Balance; a Possible Physiological Role in the Male Reproductive System." *Thromb Haemost*, vol. 92, 2004, pp. 752–66.
- Marlowe, F. "Paternal Investment and the Human Mating System." *Behavioural Processes*, vol. 51, no. 1–3, Oct. 2000, pp. 45–61, doi:10.1016/S0376-6357(00)00118-2.
- Martínez-Heredia, Juan, et al. "Identification of Proteomic Differences in Asthenozoospermic Sperm Samples." *Human Reproduction*, vol. 23, no. 4, 2008, pp. 783–91.
- Møller, Anders Pape. "Ejaculate Quality, Testes Size and Sperm Competition in Primates." *Journal of Human Evolution*, vol. 17, no. 5, Aug. 1988, pp. 479–88, doi:10.1016/0047-2484(88)90037-1.
- Nascimento, Jaclyn M., et al. "The Use of Optical Tweezers to Study Sperm Competition and Motility in Primates." *Journal of The Royal Society Interface*, vol. 5, no. 20, Mar. 2008, pp. 297–302, doi:10.1098/rsif.2007.1118.
- Pavelka, Norman, et al. "Statistical Similarities between Transcriptomics and Quantitative Shotgun Proteomics Data." *Molecular & Cellular Proteomics*, vol. 7.4, 2008, pp. 631–44.
- Peet, Robert K. "The Measurement of Species Diversity." *Annual Review of Ecology and Systematics*, vol. 5, no. 1, 1974, pp. 285–307.
- Pilch, Bartosz, and Matthias Mann. "Large-Scale and High-Confidence Proteomic Analysis of Human Seminal Plasma." *Genome Biology*, vol. 7, no. 5, 2006, p. R40.
- Ramm, Steven A., et al. "Sexual Selection and the Adaptive Evolution of Mammalian Ejaculate Proteins." *Mol. Biol. Evol.*, vol. 25, 2008, pp. 207–19.
- Seager, S. W. J., et al. "Semen Collection and Evaluation in *Gorilla gorilla gorilla*." *American Journal of Primatology*, vol. 3, no. S1, 1982, pp. 13–13, doi:10.1002/ajp.1350030506.
- Swanson, Willie J., and Victor D. Vacquier. "Reproductive Protein Evolution." *Annual Review of Ecology and Systematics*, vol. 33, 2002, pp. 161–79.
- Swanson, Willie J., and Victor D. Vacquier. "The Rapid Evolution of Reproductive Proteins." *Nature*, vol. 3, 2002, pp. 137–44.

- Torgerson, Dara G., et al. "Mammalian Sperm Proteins Are Rapidly Evolving: Evidence of Positive Selection in Functionally Diverse Genes." *Molecular Biology and Evolution*, vol. 19, no. 11, 2002, pp. 1973–80.
- Tseng, Huan-Chin, Han-Jia Lin, Jyh-Bing Tang, et al. "Identification of the Major Transglutaminase 4 Cross-Linking Sites in the Androgen-Dependent SVS I Exclusively Expressed in Mouse Seminal Vesicle." *Journal of Cellular Biochemistry*, vol. 107, 2011, pp. 899–907.
- Tseng, Huan-Chin, Jyh-Bing Tang, et al. "Mutual Adaptation between Mouse Transglutaminase 4 and Its Native Substrates in the Formation of Copulatory Plug." *Springer*, 2011.
- Tseng, Huan-Chin, Han-Jia Lin, P. S. Sudhakar Gandhi, et al. "Purification and Identification of Transglutaminase from Mouse Coagulating Gland and Its Cross-Linking Activity among Seminal Vesicle Secretion Proteins." *Journal of Chromatography B*, vol. 876, 2008, pp. 198–202.
- Uchida, Hiroshi, et al. "Glycodelin in Reproduction." *Reproductive Medicine and Biology*, vol. 12, no. 3, 2013, pp. 79–84.
- Urban, Michael, and Lily Woo. *Molecular Weight Estimation and Quantification of Protein Samples Using Precision Plus Protein WesternC Standards, the Immun-Star WesternC Chemiluminescence Detection Kit, and the ChemiDoc XRS Imaging System*. Bio-Rad bulletin 5676, 2007.
- Watts, David P. "Ecology of Gorillas and Its Relation to Female Transfer in Mountain Gorillas." *International Journal of Primatology*, vol. 11, no. 1, Feb. 1990, pp. 21–45, doi:10.1007/BF02193694.
- White, Frances J. "Comparative Socio-Ecology of Pan Paniscus." *Great Ape Societies*, edited by WC McGrew et al., Cambridge University Press, 1996, pp. 29–41, <http://dx.doi.org/10.1017/CBO9780511752414.005>.
- Whittaker, Robert H. "Dominance and Diversity in Land Plant Communities." *Science*, vol. 147, no. 3655, 1965, pp. 250–60.
- Wong, Alex. "The Molecular Evolution of Animal Reproductive Tract Proteins: What Have We Learned from Mating-System Comparisons?" *International Journal of Evolutionary Biology*, vol. 2011, 2011.
- Yates, John R. "Mass Spectrometry and the Age of the Proteome." *Journal of Mass Spectrometry*, vol. 33, no. 1, 1998, pp. 1–19.

CHAPTER 3: Prostate specific transglutaminase TGM4 activity

3.1 Introduction

Humans and chimpanzees diverged approximately 6 million years ago. Our last common ancestor with chimpanzees is extinct and the fossils they left behind can only allude to behavioral and social organization. There is debate about what our last common ancestor's mating system was, and several models have been proposed using anatomical fossil evidence, proposing either a chimpanzee-like mating system or a gorilla-like mating system. A chimpanzee-like mating system model transitions from polyandry to increased pair-bonded monogamy driven by female choice and male provisioning, concealed ovulation, and greater paternal care in the hominin lineage (Lovejoy et al, 2009; Lovejoy, 2009; Gavrillets, 2012). A monoandrous gorilla-like mating system model requires reduced aggression between males first within social groups and then between groups, while simultaneously shifting toward mate pair-bonding (Nakahashi and Horiuchi, 2012; Chapais, 2013). Either model is plausible, but many researchers tend to favor the chimpanzee-like mating system (Dixson, 2009; Chapais, 2009).

Ancestral sequence reconstruction (ASR) enables restoration of an extinct common ancestor sequence through maximum likelihood comparison of extant species and genetic engineering to produce the ancestral protein (Harms and Thornton, 2010). ASR has not been commonly used in human evolutionary studies, but has been implemented in studies of other species (Chang et al., 2002; Yokoyama et al., 2008). For example, the common ancestor of reptiles and birds, the archosaur, was determined to have nocturnal behavior, after its rhodopsin protein was reconstructed and compared to extant vertebrate rhodopsins (Chang et al., 2002). My study reconstructs the human-chimpanzee ancestral prostate specific transglutaminase (TGM4) and aims to assay its activity compare to human and chimpanzee TGM4s.

TGM4 is a calcium-dependent enzyme involved in semen coagulation and copulatory plug formation in chimpanzee by crosslinking semenogelin (SEMG1 and SEMG2) proteins at glutamine and lysine residues (Greenberg et al., 1991; Folk, 1983; Esposito and Caputo, 2004). In gibbons, gorillas, and other species with relatively low sperm competition, *TGM4* has become a pseudogene (Clark and Swanson, 2005; Carnahan and Jensen-Seaman, 2008; Tian et al., 2009). After seminal coagulation, the semenogelin proteins are cleaved by KLK3 (Kallikrein related peptidase 3, Lilja, 1985) and to a lesser extent by ACPP (Prostate acid phosphatase; Brillard-Bourdet, et al., 2002) facilitating liquefaction of semen. The cleavage of semenogelin releases bound and trapped spermatozoa (Robert et al., 1997; Flori et al., 2008). (Figure 3.1)

TGM4 is approximately 7 fold higher in chimpanzee seminal plasma compared to humans (discussed in Chapter 2), and it is important in copulatory plug formation (Dean, 2013). Because of these attributes, TGM4's activity is an excellent choice to study in order to assess mating system behavior in our last common ancestor with chimpanzee. If the ancestral TGM4 functions more like chimpanzee TGM4 compared to human TGM4, then a chimpanzee-like mating system with polyandry would be supported. However, if ancestral TGM4 activity is more like human TGM4 activity, then the gorilla-like mating system model would be supported.

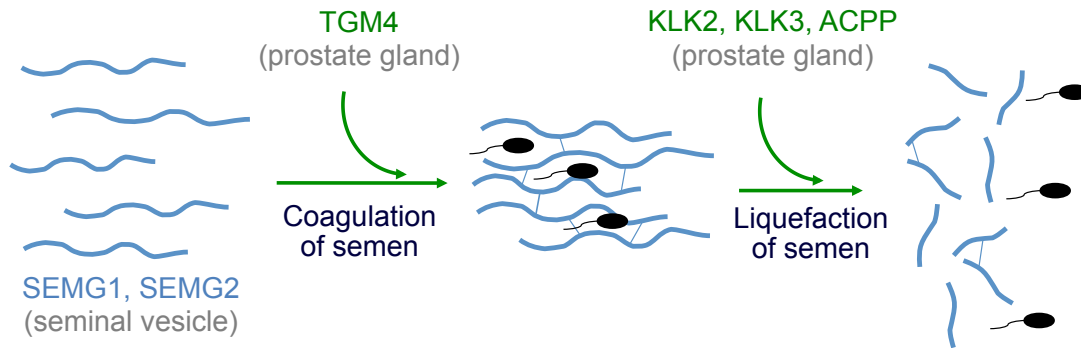


Figure 3.1: Proteins involved in semen coagulation and liquefaction upon ejaculation

SEMG1 and SEMG2 proteins are transcribed and translated in the seminal vesicle. Subsequently they are cross-linked together by TGM4 (expressed from the prostate gland) at glutamine and lysine residues forming an extracellular matrix that traps spermatozoa. Proteases, like ACPP and the KLKs, cleave SEMGs, releasing spermatozoa and liquefying semen within 5-20 minutes upon ejaculation.

3.2 Methods

This section incorporates methods for cloning *TGM4*, *SEMG1*, *SEMG2*, and *KLK3* cDNA into expression vectors with the objective to produce recombinant proteins with the human, chimpanzee, or the human-chimpanzee ancestor amino acid sequence. *TGM4* constructs were subsequently analyzed in this dissertation; however, the other constructs will be part of a larger project.

3.2.1 Cloning of Human SEMGs and TGM4 into pFastBac1 vector

3.2.1.1 Traditional and TOPO® cloning

Dr. Hiroshi Miyamoto from the University of Rochester Medical Center, generously donated pSG5 plasmids containing full-length human *SEMG1* and *SEMG2* cDNA. A clone containing the full coding sequence of human *TGM4* (IMAGE clone ID:3950865; GenBank ID:BC007003.1) was purchased from ThermoFisher Scientific (Waltham, MA). I designed primers that matched the beginning and end of *SEMG1*, *SEMG2*, and *TGM4* sequence with the addition of restriction enzyme sites (*SEMG1*: *SalI/PstI*; *XhoI/SphI*, *SEMG2*: *SalI/XbaI*; *XhoI/SphI*, *TGM4*: *SalI/XbaI*; *XhoI/SphI*) matching pFastBac1 (Invitrogen, Carlsbad, CA) vector, a C-terminal HIS tag sequence, and random nucleotides on the ends to allow restriction enzymes to work properly (Table 3.1). The semenogelins and *TGM4* were amplified with iProof™ DNA polymerase (BioRad, Hercules, CA) under the following conditions: (98°C, 30'')/(98°C, 10''/55°C, 15''/72°C, 1')35 / (72°C, 10'')/(4°C, ∞). Initially, traditional cloning (digestion, ligation, and transformation) of the PCR products was attempted, but was unsuccessful. Subsequently, the gene amplified products were purified with the Wizard® SV gel and PCR clean-up system (Promega, Madison, WI), and incubated at 72°C for 15 minutes with Taq polymerase, thus incorporating additional A's to the 3' ends before TOPO® TA cloning (Invitrogen) and plating

onto LB kanamycin/IPTG/X-gal plates. TOPO® transformants were grown overnight at 37°C in LB ampicillin and DNA was extracted using the IBI high speed plasmid mini kit (Midsci, St. Louis, MO). Candidates were sequenced with BigDye® Terminator v3.1 (Thermofisher Scientific) with the following conditions: (96.0°C, 10"/50°C 5"/60°C, 4')35 /(4°C, ∞), and analyzed on the Avant 3100 sequencer (Applied Biosystems, Foster City, CA). Vector and gene specific sequencing primers were used and are included in Appendix A.7 (Table A.14).

Table 3.1: Modified primers used for human gene construct cloning

Gene	Primer Sequence
Human <i>SEMG1</i> Forward: Reverse:	5' - CAGAATGCCGTCGACCTGCAGATGAAGCCCAACATCATCTTTGTACTTTCCCTG-3' 5' - ACGGTGATCGCATGCCTCGAGTTAATGGTGATGGTGATGGTGTGTAAATAATGGGTTT CGGTCGTTGTTAAG-3'
Human <i>SEMG2</i> Forward: Reverse:	5' - GGCTATCGATCGTCGACTCTAGAATGAAGTCCATCATCCTCTTTGTCC-3' 5' - GCCTATGCGCATGCCTCGAGTTAATGGTGATGGTGATGGTGTGTAGATATTGGATTTC TGTCTTCATTATA-3'
Human <i>TGM4</i> Forward: Reverse:	5' - CGAGTCAGGTCGACTCTAGAATGATGGATGCATCAAAAGAGCTGC-3' 5' - TAGACTCATCGGCATGCCTCGAGTTAATGGTGATGGTGATGGTGCCTTGGTGATGAGAA CAATCTTCTGAGCATTAAATC-3'

SEMGs and *TGM4* gene orientation in the TOPO® vector was determined; and 20µl of the TOPO® - *SEMG1*, *SEMG2*, or *TGM4* clone (approximately 1.5- 7µg DNA) was digested with *NotI* / *SalI* or *NotI*/ *Acc65I* in multicore buffer (Promega) as well as 20µl of pFastBac1 (Invitrogen) vector (about 3.5 µg of vector DNA), for four hours at 37°C and then terminated at 65°C for ten minutes. The pFastBac1 vector was then incubated with alkaline phosphatase (Promega) at 37°C for one hour. Alkaline phosphatase removes 5' phosphates in order to prevent

re-circularization of the vector. The digested and/or dephosphorylated products were extracted with the Wizard® SV gel and PCR clean-up system (Promega) and quantified with the Qubit® dsDNA BR assay kit on the Qubit® 2.0 fluorometer (Invitrogen). Both 1:1 and 3:1 (insert:vector) DNA ratios (~100-200ng DNA total) were added to separate ligation reactions (total volume 10µl) and incubated with T4 ligase (Promega) in the thermocycler at the following increasing temperatures for sixty minute increments (4°C, 6°C, 8-16°C). Entire ligation reactions (10µl) were transferred to 200µl of thawed TG1 chemically competent *E. coli* cells and put on ice for 20 minutes. Cells were heatshocked at 42°C for 45 seconds and immediately transferred to ice for two minutes. Cells were added to 500µl of LB broth in a culture tube and incubated at 37°C and 250 rpm for one hour. After incubation, 50µl and 250µl of each culture were spread onto LB carbenicillin plates and incubated 16-18 hours at 37°C. Single colonies were transferred to a 3mL LB ampicillin tube, and to an LB carbenicillin replicate plate, and incubated overnight at 37°C. Freezer stocks were made from 750µl of culture with 250µl 60% glycerol and stored at -80°C. The remaining 2.25mL culture's DNA was extracted with the IBI high-speed plasmid mini kit (Midsci). Candidates were sequenced with BigDye® Terminator (Thermofisher Scientific) at the following conditions: (96.0°C, 10"/50°C 5"/60°C, 4')₃₅/(4°C, ∞), and analyzed on the Avant 3100 sequencer (Applied Biosystems). Verified constructs were grown in 50mL LB ampicillin cultures and DNA was extracted using the Pureyield™ plasmid midiprep system (Promega).

3.2.1.2 Site-directed mutagenesis of human TGM4

The pcDNA-*TGM4* plasmid used to subclone human *TGM4* into pFastBac1 had one nonsynonymous mutation at 1125 bases in the sequence. This missense variant converts a proline (CCG) codon to a serine (TCG) codon. After searching for common single nucleotide polymorphisms (SNPs) in humans for *TGM4*, this variant did not appear to be present. In

addition, PolyPhen and SIFT missense effect indexes indicated this mutation would most likely be deleterious. Therefore, QuickChange II Site-directed mutagenesis (Agilent Technologies, Santa Clara, CA) was utilized to modify the mutated human *TGM4* to the common proline (CCG) variant in the population.

Human *TGM4* pFastBac1 template (10ng and 50ng) was amplified with 5.7 μ l of 20nM mutagenetic primers (Table 3.2) at the following conditions: (95°C, 1')/(95°C, 50'')/60°C, 50'')/68°C, 16'')₁₈/(68°C, 8'')/(4°C, ∞). Then 1 μ l of *DpnI* was added to each reaction and incubated at 37°C for 2 hours and 80°C for 20 minutes. XL1-Blue chemically competent *E. coli* cells (Agilent Technologies) were aliquoted (50 μ l) into pre-chilled tubes and incubated on ice for several minutes. Then 1 μ l (5ng reaction) or 4 μ l (10ng reaction) of *DpnI* digested DNA was added to the cells and incubated on ice for thirty minutes. Cells were heat shocked at 42°C for 45 seconds, and immediately incubated on ice for two minutes. Transformed cells were transferred to a 0.5mL NZY+ broth (preheated to 42°C) and incubated at 37°C and 250 rpm for one hour. After incubation, 250 μ l of cells were spread onto LB carbenicillin/ 2%XGAL/ 10mM IPTG plates and grown overnight at 37°C. Freezer stocks were made from 750 μ l of culture with 250 μ l 60% glycerol and stored at -80°C. The remaining 2.25mL culture's DNA was extracted with the IBI high speed plasmid mini kit (Midsci). Candidates were sequenced with BigDye® Terminator (Thermofisher Scientific) at the following conditions: (96.0°C, 10'')/50°C 5'')/60°C, 4'')₃₅/(4°C, ∞), and analyzed on the Avant 3100 sequencer (Applied Biosystems). Verified constructs were grown in 50mL LB ampicillin cultures and DNA was extracted using the Pureyield™ plasmid midiprep system (Promega).

Table 3.2: Mutagenic primers used for human TGM4 modification

Gene	Mutagenic Primer Sequence
Human TGM4	1: 5' - GCTGTGGACGCAACGCCGCAGGAGCGAAGCC-3'
	2: 5' - GGCTTCGCTCCTGCGGCGTTGCGTCCACAGC-3'

3.2.2 Gibson assembly of human and chimpanzee KLK3s and chimpanzee and human-chimpanzee ancestor TGM4s into the pFastBac1 vector

3.2.2.1 Designing of gBlocks

Human and chimpanzee *KLK3*s were cloned using Gibson assembly of DNA fragments. Double stranded DNA fragments, called gblocks, were ordered through Integrated DNA Technologies (IDT). Currently, 125-3,000 base pairs of DNA can be synthesized; however, during the production of these clones, only up to 700 base pairs were commercially available. Because *KLK3* mRNA with a HIS-tag would be more than 800 base pairs, three gblocks were designed by Jennifer (Vill) Doyle and Michael Seaman (Figure 3.2) to clone human and chimpanzee *KLK3*. Because species-specific nucleotide differences (that are nonsynonymous) occur in the first half of *KLK3* mRNA, a human specific gblock “A” and a chimpanzee specific gblock “A” were designed (Figure 3.2). Since no nonsynonymous mutations occur in the second half of *KLK3* mRNA, a combined human and chimpanzee gblock “B” was designed with a HIS-tag sequence inserted before the stop codon (Figure 3.2). Both gblocks include restriction enzyme sites and the end of block A and beginning of block B overlapped (underlined in Figure 3.2) approximately 60 nucleotides (for Gibson assembly to work).

```

HumanKLK3_blockA
AGGTGACAGAGCTAGCGTCGAATTCATGTGGGTCCCGGTTGTCTTCCTCACC
CTGTCCGTGACGTGGATTGGTGTGCACCCCTCATCTGTCTCGGATTGTGG
GAGGCTGGGAGTGCGAGAAGCATTCCCAACCTGGCAGGTGCTTGTGGCCTC
TCGTGGCAGGGCAGTCTGCGGCGGTGTTCTGGTGCACCCCAAGTGGGTCCCTC
ACAGCTGCCACTGCATCAGGAACAAAAGCGTGATCTTGTCTGGGTGGGCACA
GCCTGTTTCATCTGAAGACACAGGCCAGGTATTTTCAGGTCAGCCACAGCTT
CCCACACCGCTCTACGATATGAGCCTCCTGAAGAATCGATTCCCTCAGGCCA
GGTGATGACTCCAGCCACGACCTCATGTGCTCCGCTGTCTCAGAGCCTGCCG
AGCTCACGGATGCTGTGAAGGTCATGGACCTGCCCAACCCAGGAGCCAGCACT
GGGGACCACCTGCTAC

ChimpKLK3_blockA
AGGTGACAGAGCTAGCGTCGAATTCATGTGGGTCCCGGTTGTCTTCCTCACC
CTGTCCGTGACGTGGATTGGTGTGCACCCCTCATCTGTCTCGGATTGTGG
GAGGCTGGGAGTGCAAGAAGCATTCCCAACCTGGCAGGTGCTTGTGGCCTC
TCGTGGCAGGGCAGTCTGCGGCGGTGTTCTGGTGCACCCCAAGTGGGTCCCTC
ACAGCTGCCACTGCATCAGGAACAAAAGCGTGATCTTGTCTGGGTGGGCACA
GCCTGTTTCATCTGAAGACACAGGCCAGGTATTTTCAGGTCAGCCACAGCTT
CCCACACCGCTCTACGATATGAGCCTCCTGAAGAATCGATTCCCTCAGGCCA
GGTGATGACTCCAGCCACGACCTCATGTGCTCCGCTGTCTCAGAGCCTGCCG
AGATACGGATGCTGTGAAGGTCATGGACCTGCCCAACCCAGGAGCCAGCACT
GGGGACCACCTGCTAC

HumanChimpKLK3_blockB
GATGCTGTGAAGGTCATGGACCTGCCCAACCCAGGAGCCAGCACTGGGGACCA
CCTGCTACGCCTCAGGCTGGGGCAGCATTGAACCAGAGGAGTTCCTTGACCCC
AAAGAACTTCAGTGTGTGGACCTCCATGTTATTTCCAAATGACGTGTGTGGC
CAAGTTCACCTTCAGAAGGTGACCAAGTTCATGCTGTGTGTGGACGCTGGA
CAGGGGGCAAAGCACCTGCTCGGGTGATTCGGGGGCCCACTTGTCTGTAA
TGGTGTGCTTCAAGGTATCACGTATGGGGCAGTGAACCATGTGCCCTGCC
GAAAGGCCTTCCTGTACACCAAGTGGTGCATTACCGAAGTGGATCAAGG
ACACCATCGTGCCCAACCCCACCATCACCATCACCATTGAGGTACCGAGAT
CAAGGCTCCATCTCC

```

Figure 3.2: Human and chimpanzee *KLK3* gblock design

Human *KLK3* gblock A and chimpanzee gblock A have two nonsynonymous differences highlighted in yellow. Both gblock As overlap with HumanChimp*KLK3* gblock B where the sequence is underlined. The combined gblock B has a HIS-tag sequence (in blue) inserted before the stop codon (red). Restriction enzyme sites (*Eco*RI and *Acc*65I) were included before the start codon (green) and after the stop codon (red) and are bolded.

Chimpanzee and human-chimpanzee ancestor *TGM4s* were cloned using Gibson assembly of DNA fragments. At the time, Integrated DNA Technologies (IDT) could synthesize up to 2,000 base pairs of DNA, which was just shy of the 2,100 base pair minimum requirement to design HIS-tag *TGM4s*. Therefore each *TGM4* would have to be assembled using two gblocks with a forty base pair overlap. Nonsynonymous mutations between chimpanzee and human-chimpanzee ancestor are throughout the mRNA sequence, therefore, a total of four gblocks were required (Ancestor Block A, Ancestor Block B, Chimpanzee Block A, and Chimpanzee Block B) and were otherwise designed similar to the *KLK3* gblocks (Figure 3.3). Human-chimpanzee ancestor sequence was determined by comparing the amino acid sequence between human, chimpanzee, and orangutan as an outgroup (Appendix A.8).

ChimpTGM4_BlockA
GGATCCCGGTCCGAGCGCGCGGAATTCAAAGGCC**TACGTCGAC**
ACATGATGGATGCATCAAAGAGCTGCAAGTTCTCCACGTTGACTT
CTTGAAGCAGGACAACGCCGTTTCTCACCACACATGGGAGTTCC
AAACGAGCA**CT**CTGTGTCCGGCGAGGACAGGTGTTTCCACTG
CGGCTGGTGTGAACCAGCCCTACAATCTACCACCAACTGAA
ACTGGAATTCAGCACAGGGCCGAATCCTAGCATCGCCAAACACA
CC**G**TGGTGGTCTCGACCT**G**GAGGACGCCCTCAGACCACTAAC
TGGCAGGCAACCC**TT**CAAAATGAGTCTGGCAAAGAGGTCACAGT
GGCTGTCAACAGTTCCCCCAATGCCATCTGGGCAAGTACCAAC
TAAATGTGAAAATCGAAACCACATCCTTAAGTCTGAAGAAAAC
ATCTATACCTTCTTCAACCCATGGTGTAAAGAGGACATGGT
TTTCACTGCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATG
ACACGGGCTGCCATTACGTGGGGCTGCCAGAAGTATCA**G**ATAC
AAACCC**T**GGAACTTTGGT**C**AGTTTGAGAAAAATGTCTGGACTG
GTGCATTTCCCTGTGACTGAGAGCTCCCTCAAGCCACAGATA
GGAGGGACCCCGTGTGGTGTGTCAGGGCCATGTGTCTATGATG
AGCTTTGAGAAAGGCCAAGGCCGTGCTCATTTGGGAATGGACTGG
GGACTACGAAGTGGCACAGCCCATACAAGTGGACAGGCACTG
CCCGATCCTGCAGCACTACAACACGAAGCAGGCTGTGTGC
TTTGGCCAGTCTGGGTGTTGTCTGGGATCCTGACTACAGTCT
GAGAGCTTTGGGCTCCAGCACGCACTGTGACAGGCTTCGATT
CAGCTCACGACAGAAAGGAACCTCACGGTGGACACCTATGTG
AATGAGAATGGCGAGAAAATCACCAATATGA

ChimpTGM4_BlockB
ACACCTATGTGAATGAGAATGGCGAGAAAATCACCAATATGACCC
CGACTCTGTCTGGAATTTCCATGTGTGGACGGATCGCTGGATGA
CGACCGGATCTGCCAAGGGCTACGACGGCTGGCAGGCTGTGGACG
CAACGCCGAGGAGCGAAGCCAGGGTGTCTTCTGTCTGTGGGCCAT
ACCCTGACCCATCCGCAAAGGTGACATCTTTATTGTCTATGAC
ACCAGATTCGTCTTCTCAGAAGTGAATGGTGACAGGCTCATCTGGT
TGGTGAAGATGGTGAATGGCGAGGAGGTTACACGTAAATTTCAAT
GGAGACCACAAGCATCGGAAAAACATCAGCACAAGGCAAGTGGGC
CAAGACAGGCGGAGAGATATCACCTATGAGTACAAAGTATCCAGAAG
GCTCCTCTGAGGAGAGGCAGGTCATGGATCATGCCTTCTCCTTCT
CAGTCTCTGAGAGGGACAGACAGCACTGTAAAAGAGAATTTCTT
CACATGTCTGTAACAATCAGATGATGTGTCTGGGAAACCC**T**GTTA
ATTTACCCGTGATTTCTAAAAGGAAGACCGTGCCTACAGAATGT
CAACATCTTGGGCTCCTTTGAAC**T**ACAGTTGTACTGGCAAGAAG
GTGGCAAACCTGTGTGACCTCAATAAGACCTCGCAGATCAAAGGTC
AAGTATCAGAAGTGAATCTGACCTGGACTCCAAGACCTACATCAA
CAGCTTGGCTATATTAGATGATGAGCAGTATCAGAGGTTTCTATC
ATTGGCGAAATTTGTGGAGTCTAAGGAAATCATGGCCTCTGAAGTAT
TCACGTCTTTCCAGTACCCTGAG**A**TCCTATAGAGTTGCC**T**AAAC
AGGCAGAATTTGGCCAGCTACTTGTCTGCAATTTGATCTTCAAGAT
ACCTGGCCATCCCTTTGACTGACCTCAAGTCTCTTTGGAAAGCC
TGGGCATCTCTCACTACAGACCTCTGACCATGGGACCGTGCAGCC
TGGTGAGACCATCCAATCCCAAATAAAATGCACCCCAATAAAAACT
GGACCCAAAGAAATTTATCGTCAAGTTAAGTTCCAACCAAGTGAAG
AGATTAATGCTCAGAAGATTGTTCTCATCACCAAG**CACCATCACCA**
TCACCATTAACGGCGCCGCTTTTGGAACTTAGAGCTGCAGTCTCGA
GGCATGGGTACCAAGC

Human/ChimpAncestorTGM4_BlockA
GGATCCCGGTCCGAGCGCGCGGAATTCAAAGGCC**TACGTCG**
ACATGATGGATGCATCAAAGAGCTGCAAGTTCTCCACGTTGACTT
ACTTCTTGAAGCAGGACAACGCCGTTTCTCACCACACATGGG
AGTTCAAACGAGCA**CT**CTGTGTCCGGCGAGGACAGGTGTT
TTCACCTGCGGCTGGTGTGAACCAGCCCTACAATCTTACC
ACCAACTGAAACTGGAATTCAGCACAGGCGCGAATCTTAGCA
TCGCCAAACACAC**CT**GGTGGTGTCTGACCC**G**GAGGACGCCCT
CAGACCACTACAAC**T**GGCAGGCAACCC**TT**CAAAATGAGTCTG
GCAAAGAGGTCACAGTGGCTGTACCAGTTC**CC**CAATG**CC**A
TCCTTGGCAAGTACCAACTAAATGTGAAA**ACT**GGAAACCACA
TCCTTAAGTCTGAAGAAACATCTTATACCTTCTTCAACC
CATGGTGAAGAGGACATGGT**TT**TCATGCCTGATGAGGAC
AGCGCAAAGAGTACATCTCAATGACACGGGCTGCCATTAGC
TGGGGCTGCCAGAAGTATCA**AA**ATACAACCC**T**GGAACTTTG
GTCAGTTTGAGAAAAATGTCTGGACTGCTGCATTTCCCTGC
TGACTGAGAGCTCCCTCAAGCCACAGATAGGAGGGACCCCG
TGCTGTGTGTCAGGGCCATGTGTCTATGATGAGCTTTGAGA
AAGGCCAAGCGTCTCATTTGGGAATTTGGACTGGGGACTAGC
AAGGTGGCACAGCCCATACAAGTGGACAGGCACTGCCCCGA
TCCTGCAGCAGTACTACAACAGCAAGCAGGCTGTGTCTTTG
CCGACTGGGTGCTTCTGCTGGGATCCTGACTGACTGCTGTA
GAGCGTTGGGCTCCAGCACGCACTGTGACAGGCTTCGATT
CAGCTCACGACAGAAAGGAACCTCACGGTGGACACCTATG
TGAATGAGAATGGCGAGAAAATCACCAATATGA

Human/ChimpAncestorTGM4_BlockB
ACACCTATGTGAATGAGAATGGCGAGAAAATCACCAATATGACCC
ACGACTCTGTCTGGAATTTCCATGTGTGGACGGATCGCTGGATGA
AGGCACCGGATCTGCCAAGGGCTACGACGGCTGGCAGGCTGTGG
ACGCAACGCCGAGGAGCGAAGCCAGGGTGTCTTCTGTCTGTGGC
CATCACCACTGACCCCATCCGCAAGGTGACATCTTATTGTCT
ATGACACCAGATTCGTCTTCTCAGAAGTGAATGGTGACAGGCTCA
TCGGTGGTGAAGATGGTGAATGGGCAGGAGGATACACGTAA
TTTCAATGGAGACCACAAGCATCGGAAAAACATCAGCACAAGG
CAGTGGCCAAAGCAGGCGGAGAGATATCACCTATGAGTACAAGT
ATCCAGAAGGCTCCTCTGAGGAGAGGCAGGTCATGGATCATGCCT
TCCTCCTCTCAGTCTCTGAGAGGGACACAGCAGCTGTAAAAG
AGAACTTTCTTACATGTCTGATACATCAGATGATGTCTGTGG
GAAACCC**T**GTAAATTTACCCGTGATTTCTTAAAAGGAAGACCCGTG
CCCTACAGAATGTCAACATCTTGGGCTCCTTTGAAC**T**ACAGTTGT
ACACTGGCAAGAAAGTGGCAAACCTGTGTGACCTCAATAAGACCT
CGCAGATCCAAGGTCAAGTATCAGAAGTGA**CT**CTGACCTTGGACT
CCAAGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAG
TTATCAGAGGTTTCTATCATTTGGGAAATTTGGAGTCTAAGGAAA
TCATGGCCTCTGAAGTATTCAGTCTTTCCAGTACCCTGAG**CT**CT
CTATAGAGTTGCC**T**AAACAGGCAGAAATGGCCAGCTACTTGTCT
GCAATTTGATCTTCAAGAAATACCCTGGCCATCCCTTTGACTGAGC
TCAAGTTCTCTTTGGAAAGCCTGGGCATCTCTCACTACAGACCT
CTGACCATGGGACGGTGCAGCCTGGTGGAGACCATCCAATCCAAA
TAAATGCACCCCAATAAA**ACT**GGACCAAGAAATTTATCGTCA
AGTTAAGTTCCAAACAAGTGAAGAGATTAATGCTCAGAAGATG
TTCTCATCACCAAG**CACCATCACCATCACCATTAACGGCGCCGCT**
TTGAACTTAGAGCTGCAGTCTCGAGGCATGGGTACCAAGC

Figure 3.3: Human-chimpanzee ancestor and chimpanzee TGM4 gblock design

Human-chimpanzee ancestor TGM4 gblock A and B and chimpanzee gblock A and B have different sequences, highlighted in yellow. Both gblock As overlap with their respective gblock Bs where the sequence is underlined. Both gblock Bs have a HIS-tag sequence (in blue) inserted before the stop codon (red). Restriction enzyme sites (*Sall* and *NotI*) were included before the start codon (green) and after the stop codon (red) and are bolded. In addition, the surrounding sequence matched the pFastBac1 vector sequence.

The ordered gblocks (Figures 3.2 and 3.3) from Integrated DNA Technology (IDT) came lyophilized and were reconstituted with 20µl TrisEDTA (10ng/µl). The gblocks were diluted (1:100) and amplified with 1µl of 20nmol gblock primers (Table 3.3) using 0.5µl TaKaRa DNA polymerase (Clontech, Mountain View, CA) under the following conditions: (94°C, 2')/(98°C, 15'')/ 55°C, 15'')/ 72°C, 1')₃₅/(4°C, ∞) in a 50µl reaction. The amplified gblocks were separated on a 1% agarose gel with 0.1% crystal violet buffer, and the amplified products were extracted with the Wizard® SV gel and PCR clean-up system (Promega) and quantified with the Qubit® dsDNA BR assay kit on the Qubit® 2.0 fluorometer (Invitrogen).

Table 3.3: gBlock primers of *KLK3* and *TGM4* used for gibson assembly

Gene	gBlock Primer Sequence
KLK3	
Block A_Fwd:	5' - AGGTGACAGAGCTAGCGTC-3'
Block A_Rev:	5' - GTAGCAGGTGGTCCCCAG-3'
Block B_Fwd:	5' - GATGCTGTGAAGGTCATGGA-3'
Block B_Rev:	5' - GGAGATGGAGCCTTGATCTC-3'
TGM4	
Block A_Fwd:	5' - GGATCCCGGTCCGAAGCGCGCGG-3'
Block A_Rev:	5' - CATATTGGTGATTTTCTCGCCATTC-3'
Block B_Fwd:	5' - CACCTATGTGAATGAGAATGGCGAG-3'
Block B_Rev:	5' - GCTTGGTACCGCATGCCTCG-3'

3.2.2.2 Gibson assembly

Chimpanzee *KLK3* gblocks were amplified separately by Jennifer (Vill) Doyle and were incubated with Gibson assembly mix (New England Biolabs®, Ipswich, MA) for an hour at 50°C and then 80°C for 20 minutes. Gibson assembled chimpanzee *KLK3* and pFastBac1 (Invitrogen) vector were digested with *EcoRI/Acc65I* in multicore buffer (Promega), as well as 20µl of pFastBac1 (Invitrogen) vector for four hours at 37°C and then terminated at 65°C for ten minutes. The pFastBac1 vector was then incubated with alkaline phosphatase (Promega) at 37°C

for one hour. Ligation, transformation, and verification of chimpanzee *KLK3* pFastBac1 was similar to human *SEMGs* and human *TGM4* described previously.

Human *KLK3* gblocks A and B were amplified together with Block A forward primer and Block B reverse primer using iProof™ DNA polymerase (BioRad) under the following conditions: (98°C, 2')/(98°C, 10')/ 55°C, 15')/ 72°C, 1')₃₅/(72°C, 10')/(4°C, ∞). Amplified human *KLK3* gblocks AB DNA and pFastBac1 (Invitrogen) were digested with fast *EcoRI* and *Acc65I* (Thermofisher Scientific) at 37°C for 45 minutes and terminated at 80°C for 20 minutes. The pFastBac1 vector was simultaneously dephosphorylated with Fast Alkaline Phosphatase (Thermofisher Scientific). The digested and/or dephosphorylated products were extracted with the Wizard® SV gel and PCR clean-up system (Promega) and quantified with the Qubit® dsDNA BR assay kit on the Qubit® 2.0 fluorometer (Invitrogen). Different insert:vector DNA ratios (1:1, 1:3 and 3:1) were added to separate ligation reactions (total volume 10µl) and incubated with Fast T4 ligase (Thermofisher Scientific) in the thermocycler at the following temperatures: (22°C, 10')/(21°C, 5')/ (20°C, 5')/ (18°C, 5')/ (16°C, 5')/ (14°C, 5')/(65°C, 10')/(4°C, ∞). Half of each ligation reaction (5µl) was transferred to 200µl of thawed TG1 chemically competent *E. coli* cells and put on ice for 20 minutes. Cells were heatshocked at 42°C for 45 seconds and immediately transferred to ice for two minutes. Cells were added to 500µl of LB broth in a culture tube and incubated at 37°C and 250 rpm for one hour. After incubation, 50µl and 250µl of each culture were spread onto LB carbenicillin plates and incubated 16-18 hours at 37°C. Single colonies were transferred to a 3mL LB ampicillin tube, and to an LB carbenicillin replicate plate, and incubated overnight at 37°C. Freezer stocks were made from 750µl of culture with 250µl 60% glycerol and stored at -80°C. The remaining 2.25mL culture's DNA was extracted with the IBI high speed plasmid mini kit (Midsci). Candidates were

sequenced with BigDye® Terminator (Thermofisher Scientific) at the following conditions: (96.0°C, 10'')/50°C 5'')/60°C, 4')₃₅/(4°C, ∞), and analyzed on the Avant 3100 sequencer (Applied Biosystems). Verified constructs were grown in 50mL LB ampicillin cultures and DNA was extracted using the Pureyield™ plasmid midiprep system (Promega).

Two to three molar excess of ancestor and chimpanzee *TGM4* gblocks were incubated with *NotI/SalI* digested pFastBac1, Gibson assembly mix (New England Biolabs), and water for an hour at 50°C and terminated at 80°C for 10 minutes. NEB® 5-alpha competent *E. coli* (high efficiency) cells were thawed on ice for approximately thirty minutes. Then 10µl of chimpanzee or ancestor *TGM4* gblock mixed DNA was added to the cells and incubated for another thirty minutes on ice. Cells were heatshocked at 42°C for 45 seconds, and 0.5mL of LB broth was added to the cells and incubated at 37°C and 250 rpm for one hour. After incubation, 50µl and 200µl were plated on LB carbenicillin plates and incubated overnight at 37°C. Single colonies were transferred to a 3mL LB ampicillin tube, and to an LB carbenicillin replicate plate, and incubated overnight at 37°C. Freezer stocks were made from 750µl of culture with 250µl 60% glycerol and stored at -80°C. The remaining 2.25mL culture's DNA was extracted with the IBI high speed plasmid mini kit (Midsci). Candidates were sequenced with BigDye® Terminator (Thermofisher Scientific) at the following conditions: (96.0°C, 10'')/50°C 5'')/60°C, 4')₃₅/(4°C, ∞), and analyzed on the Avant 3100 sequencer (Applied Biosystems). Verified constructs were grown in 50mL LB ampicillin cultures and DNA was extracted using the Pureyield™ plasmid midiprep system (Promega).

3.2.2.3 Site-directed mutagenesis of chimpanzee and ancestor TGM4s in pFastBac1 vector

Gibson assembly of chimpanzee and ancestor *TGM4* into the pFastBac1 plasmid produced a nonsynonymous mutation in each construct. Therefore, Quick-change lightening site-

directed mutagenesis (Agilent Technologies) was utilized to modify the mutated TGM4s to normal. Table 3.4 provides the mutagenic primer sequences. Ancestor and chimpanzee TGM4 template (5ng, 25ng, and 50ng) was amplified with 125ng of respective mutagenetic primers at the following conditions: $(95^{\circ}\text{C}, 2')$ / $(95^{\circ}\text{C}, 20' / 60^{\circ}\text{C}, 10' / 68^{\circ}\text{C}, 5')$ ₁₈/ $(68^{\circ}\text{C}, 5')$ / $(4^{\circ}\text{C}, \infty)$. Then 2 μl of *DpnI* was added to each reaction and incubated at 37°C for 5 minutes. XL10-gold chemically competent *E. coli* cells (Agilent Technologies) were aliquoted (45 μl) into pre-chilled tubes and 2 μl of β -mercaptoethanol was added to the cells and incubated on ice for two minutes. Then 2 μl of *DpnI* digested DNA was added to the cells and incubated on ice for thirty minutes. Cells were heat shocked at 42°C for 30 seconds, and immediately incubated on ice for two minutes. Transformed cells were transferred to 0.5mL NZY+ broth and incubated at 37°C and 250 rpm for one hour. After incubation, 250 μl of cells were spread onto LB carbenicillin plates and grown overnight at 37°C. Freezer stocks were made from 750 μl of culture with 250 μl 60% glycerol and stored at -80°C. The remaining 2.25mL culture's DNA was extracted with the IBI high speed plasmid mini kit (Midsci). Candidates were sequenced with BigDye® Terminator (Thermofisher Scientific) at the following conditions: $(96.0^{\circ}\text{C}, 10' / 50^{\circ}\text{C} 5' / 60^{\circ}\text{C}, 4')$ ₃₅/ $(4^{\circ}\text{C}, \infty)$, and analyzed on the Avant 3100 sequencer (Applied Biosystems). Verified constructs were grown in 50mL LB ampicillin cultures and DNA was extracted using the Pureyield™ plasmid midiprep system (Promega).

Table 3.4: Mutagenic primers used for chimpanzee and ancestor TGM4 modification

Gene	Mutagenic Primer Sequence
Chimpanzee TGM4	1: 5' - CAAGCCCACAGATAGGAGGGACCCCGTGCTGGTGTGCAGGGCC-3'
	2: 5' - GGCCCTGCACACCAGCACGGGGTCCCTCCTATCTGTGGGCTTG-3'
Human-chimpanzee ancestor TGM4	1: 5' - CAGATGATGTGCTGCTGGGAAACCCTGTTAATTTACCG-3'
	2: 5' - CGGTGAAATTAACAGGGTTTCCCAGCAGCACATCATCTG-3'

3.2.3 Cloning of chimpanzee SEMGs into pFastBac1

The coding portion of *SEMG1* in exon 2 from chimpanzee genomic DNA was amplified using TaKaRa DNA polymerase (Clonetechn) under the following conditions: (98°C, 30'')/(98°C, 10'')/ 57°C, 15'')/ 72°C, 1')₁₀/(98°C, 10'')/ 55°C, 15'')/ 72°C, 1')₂₀/(72°C, 10')/(4°C, ∞) in a 50µl reaction. This PCR is a modified nested PCR, where four primers were added, two external primers to amplify a larger region of *SEMG1*, and two internal primers to amplify a smaller section of *SEMG1* with added restriction sites and addition of a HIS tag (Table 3.5). In addition, TaKaRa taq (0.5µl) and dH₂O (5µl) were added to the reaction immediately before the first cycle's 57°C annealing stage, which is often referred to a "Hot start" PCR.

The coding portion of *SEMG2* in exon 2 from chimpanzee genomic DNA was amplified using Phusion High Fidelity polymerase (Thermofisher Scientific) under the following conditions: (98°C, 30'')/(98°C, 10'')/ 57°C, 15'')/ 72°C, 1')₁₀/(98°C, 10'')/ 55°C, 15'')/ 72°C, 1')₂₀/(72°C, 10')/(4°C, ∞) in a 50µl reaction with provided GC buffer. This PCR is a modified nested PCR, where four primers were added, two external primers to amplify a larger region of *SEMG2*,

and two internal primers to amplify a smaller section of *SEMG2* with added restriction sites and addition of a HIS tag (Table 3.5). Like *SEMG1* amplification, 0.5µl of Phusion polymerase and 5µl of dH₂O was added to the reaction during the first primer annealing cycle.

Table 3.5: Primers used for chimpanzee *SEMG1* and *SEMG2* amplification

Gene	Primer Sequence
Chimpanzee <i>SEMG1</i> (External)	
MS_11F:	5' - GAAAGCTGCTCAGACAGCTA-3'
MS_12R:	5' - AGATATCTCCCCTATTTGCTA-3'
(Internal)	
2CS1_PvuIIEx1/2 Fwd:	5' - CGATGACTAGCAGCTGTGATGGGACAAAAAGGTGGATCAAAAG GCCGATTACCAAGTGAAT-3'
2CS1_HISstopNotIRev:	5' - GATCGATGACGCGGCCGCTTAATGGTGATGGTGATGGTGTGTA AATAATGGGTTTCGGTCGTTG-3'
Chimpanzee <i>SEMG2</i> (External)	
MS_21F:	5' - TCTCCACCCAACGCTGTAGGC-3'
MS_22R:	5' - CCTGACATTGCAAGTGTCC-3'
(Internal)	
2CS2_PvuIIEx1/2 Fwd:	5' - CGTTGCAAGGCAGCTGTGATGGGACAAAAAGATGGATCAAAAG GCCAATT- 3'
2CS2_HISstopKpnIRev:	5' - GATGGCATGGGTACCTTAATGGTGATGGTGATGGTGTGTAGAT ATTGGATTTCTGTCTTCATTATATTGTTGTGTC- 3'

Multiple attempts to digest *SEMGs* PCR products (*SEMG1*: *PvuII/NotI*, *SEMG2*: *PvuII/Acc65I*) and extract with the Wizard® SV gel and PCR clean-up system (Promega) were attempted. However, after quantification with the Qubit® dsDNA BR assay kit on the Qubit® 2.0 fluorometer (Invitrogen), concentrations were too low for ligation. We decided instead of amplifying from genomic DNA, specificity and yield might be higher if amplification was from BAC (Bacterial Artificial Chromosome) DNA known to contain the chimpanzee *SEMGs* (Hurle

et al., 2007; PubMed ID 17267810). An *E. coli* culture harboring the CH251-562O11 BAC (BACPAC Resources/CHORI, Oakland, CA) was streaked onto a 2XYT (1.6% Bactotryptone, 1% yeast, and 0.5% NaCl) chloramphenicol plate and grown overnight at 37°C for 20 hours and at 22°C for 48 hours. Single CH251-562O11 colonies were transferred to 3mL of 2XYT chloramphenicol broth and incubated at 37°C overnight with 250 rpm agitation. Freezer stocks were made from 750µl of culture with 250µl 60% glycerol and stored at -80°C. The remaining 2.25mL culture's BAC DNA was extracted with the ZR BAC DNA miniprep kit (Zymo Research, Irvine, CA). Then 1µl of CH251-562O11 BAC DNA was amplified under the same conditions described above for both *SEMG1* and *SEMG2*.

Chimpanzee *SEMG1* PCR and human *SEMG1* pFastBac1 were digested with *PvuII* and *NotI* for 45 minutes at 37°C and terminated at 72°C for 10 minutes. Chimpanzee *SEMG2* PCR and human *SEMG2* pFastBac1 were digested with *PvuII* and *Acc65I* for 45 minutes at 37°C and terminated at 72°C for 10 minutes. Both digested human *SEMG1* and *SEMG2* pFastBac1 were dephosphorylated with 2µl of FastAP (ThermoFisher Scientific) for 15 minutes at 37°C and terminated at 75°C for 5 minutes. Digested products were separated on a 1% agarose gel with 0.1% crystal violet buffer and were purified with the Zymoclean gel DNA recovery kit (Zymo Research). Samples were quantified with the Qubit® dsDNA BR assay kit on the Qubit® 2.0 fluorometer (Invitrogen). Different insert:vector DNA ratios (1:1, 1:3 and 3:1) were added to separate ligation reactions (total volume 10µl) and incubated with Fast T4 ligase (ThermoFisher Scientific) in the thermocycler at the following temperatures: (22°C, 10')/(21°C, 5')/(20°C, 5')/(18°C, 5')/(16°C, 5')/(14°C, 5')/(65°C, 10')/(4°C, ∞). Half of each ligation reaction (5µl) was transferred to 200µl of thawed TG1 chemically competent *E. coli* cells and put on ice for 20 minutes. Cells were heatshocked at 42°C for 45 seconds and immediately transferred to ice for

two minutes. Cells were added to 500µl of LB broth in a culture tube and incubated at 37°C and 250 rpm for one hour. After incubation, 50µl and 200µl of each culture were spread onto LB carbenicillin plates and incubated 16-18 hours at 37°C. Single colonies were transferred to a 3mL LB ampicillin tube, and to an LB carbenicillin replicate plate, and incubated overnight at 37°C. Freezer stocks were made from 750µl of culture with 250µl 60% glycerol and stored at -80°C. The remaining 2.25mL culture's DNA was extracted with the IBI high speed plasmid mini kit (Midsci). Candidates were sequenced with BigDye® Terminator (Thermofisher Scientific) at the following conditions: (96.0°C, 10"/50°C 5"/60°C, 4')₃₅/(4°C, ∞), and analyzed on the Avant 3100 sequencer (Applied Biosystems). Only chimpanzee *SEMG2* was verified via sequencing and size analysis through gel electrophoresis. TA TOPO cloning of chimpanzee *SEMG1* products was attempted, but also yielded shorter constructs.

3.2.4 Site directed mutagenesis to produce chimpanzee SEMG2-TYR mutant and human-chimpanzee ancestor KLK3 in pFastBac1 vector

Quick-change lightning site-directed mutagenesis (Agilent Technologies) was utilized to make both human-chimpanzee ancestor *KLK3* and the chimpanzee *SEMG2* Tyrosine mutant in pFastBac1 (Invitrogen). Human *KLK3* and chimpanzee *SEMG2* constructs in pFastBac1 (5ng, 25ng, and 50ng) were amplified with 125ng mutagenic primers (Table 3.6) at the following conditions: (95°C, 2')/(95°C, 20"/60°C, 10"/68°C, 5')₁₈/(68°C, 5')/(4°C, ∞). *DpnI* digestion, transformation, and verification were performed similar to chimpanzee and ancestor *TGM4* site-directed mutagenesis described previously.

Table 3.6: Mutagenic primers used for chimpanzee and ancestor TGM4 modification

Gene	Mutagenic Primer Sequence
Chimpanzee SEMG2-TYR	1: 5' - GGAAAATAAAAATATCATACCAATCTTCGAGTAC-3'
	2: 5' - GTACTCGAAGATTGGTATGATATTTTATTTTCC-3'
Human-chimpanzee ancestor <i>KLK3</i>	1: 5' - CCTGTCAGAGCCTGCCGAGATCACGGATGCTGTGAAGGTC-3'
	2: 5' - GACCTTCACAGCATCCGTGATCTCGGCAGGCTCTGACAGG-3'

3.2.5 Transfer of pFastBac1 clones to pCMV mammalian expression vectors

3.2.5.1 Chimpanzee *SEMG1*

After multiple attempts to clone full-length chimpanzee *SEMG1*, we ordered a construct cloned into pCMV-3Tag-3a (Agilent Technologies) from GenScript (Piscataway, NJ) containing chimpanzee *SEMG1*-HIS sequence (Hurle et al., 2007; Accession number: DP000037). First we received 4µg of DNA, which 20, 40, or 60ng of DNA was transformed into One Shot® TOP10 chemically competent cells (Thermofisher Scientific) and 75µl or 250µl of culture was spread onto LB kanamycin plates. Plates were incubated overnight at 37°C; however, there was no growth. We requested a glycerol stock of chimpanzee *SEMG1*-HIS pCMV-3Tag-3a from GenScript. The stock was labeled as TOP10 *E. coli* cells, but I confirmed with the company that the cells were *Stb13* cells, which aid in reduced recombination between repetitive regions. The glycerol stock was streaked onto LB kanamycin plates and grown at 37°C for approximately 26 hours. A single colony was transferred to 3mL of LB kanamycin broth and grown overnight at 37°C with 250rpm agitation. Freezer stocks were made from 750µl of culture with 250µl 60% glycerol and stored at -80°C. In addition, 30µl was transferred from the overnight culture to another 3mL LB kanamycin broth and grown for 6 hours. Then 50µl of the starter culture was transferred to a 50mL LB kanamycin broth and grown at 37°C (250rpm) for 18 hours. DNA was extracted using the Pureyield™ plasmid midiprep system (Promega). Midipreps were digested

with *HindIII* and *XhoI* to determine size compared to other digested semenogelins. However, midipreps grown at 37°C were the incorrect size (smaller), I decided to restart from the GenScript glycerol stock and grow all cultures at 30°C. Midiprep digestion with *HindIII* and *XhoI* showed the 30°C were also the incorrect size; however, the original 4µg stock DNA from GenScript when digested, was the correct size.

The Puthenveedu lab from Carnegie Mellon Research Institute provided chemically competent *Stbl3* cells, and 1µl (200ng) of chimpanzee *SEMG1*-HIS pCMV-3tag-3a DNA was added to thawed *Stbl3* competent cells. Cells were incubated on ice for 30 minutes and then heatshocked at 42°C for 45 seconds and immediately placed on ice for 2 minutes. Transformed cells were added to 500µl of LB broth and incubated at 30°C for one hour at 250 rpm. Then 200µl of culture was transferred to three LB kanamycin plates and incubated at three different temperatures: 22°C, 30°C, and 37°C. In addition, 100µl of cells were grown on LB plates at 30°C to determine if the *Stbl3* cells were viable. Single colonies were picked from 22°C and 30°C plates and grown in 3mL LB kanamycin at either 22°C or 30°C, respectively, overnight with agitation. DNA was extracted using the IBI high speed plasmid mini kit (Midsci), and 250µl of culture was mixed with 750µl of 60% glycerol and stored at -80°C. Both 22°C and 30°C grown cultures were digested with *HindIII* and *XhoI* to determine size. Candidates were sequenced with BigDye® Terminator (ThermoFisher Scientific) at the following conditions: (96.0°C, 10' / 50°C 5' / 60°C, 4')₃₅ / (4°C, ∞), and analyzed on the Avant 3100 sequencer (Applied Biosystems). Verified constructs were grown in 50mL LB kanamycin cultures at 22°C for 48-72 hours and DNA was extracted using the Pureyield™ plasmid midiprep system (Promega).

3.2.5.2 *KLK3s, TGM4s, and SEMGs subcloning procedure into pCMV vector*

Human, chimpanzee, and human-chimpanzee ancestor *KLK3* in pFastBac1 constructs and pCMV-flag vector (with *GFP*) were digested with Fast *EcoRI* and *Acc65I*. Human, chimpanzee, and human-chimpanzee ancestor *TGM4* in pFastBac1 were amplified using Phusion High Fidelity polymerase (Thermofisher Scientific) under the following conditions: (98°C, 1')/(98°C, 10'"/ 60°C, 30'"/ 72°C, 1')₃₀/ (72°C, 8')/(4°C, ∞) in a 50µl reaction with provided HF buffer and with universal primers (Table 3.7). Amplified *TGM4* products and pCMV-flag vector (with *GFP*) were digested with Fast *BgIII*⁶ and *Acc65I* (Thermofisher Scientific). Human *SEMG1* in pFastBac1 was digested with *SalI*, and then incubated with 0.2µl of C-dNTP and T-dNTPs and 2µl Klenow polymerase (Promega) at room temperature for 30 minutes and terminated at 75°C for 10 minutes. Likewise, pCMV-flag vector (with *GFP*) was digested with *BgIII*, and then incubated with 0.2µl of G-dNTP and A-dNTPs and 2µl of Klenow polymerase at room temperature for 30 minutes and terminated at 75°C for 10 minutes. Both Klenow base-filled in digests were redigested with *Acc65I*, and the pCMV-flag vector was dephosphorylated with Fast Alkaline Phosphatase (Thermofisher Scientific). Human *SEMG2*, chimpanzee *SEMG2*, and chimpanzee *SEMG2*-Tyr constructs in pFastBac1 and pCMV-flag (with *GFP*) were digested with *NotI* and *Acc65I*. All digestions were carried out at 37°C for 90 minutes with termination at 80°C for 20 minutes. All pCMV-flag vectors were dephosphorylated with Fast Alkaline Phosphatase (Thermofisher Scientific), simultaneously. Products were separated on a 1% agarose gel with 0.1% crystal violet buffer and were purified with the IBI high speed plasmid mini kit (Midsci). Samples were quantified with the Qubit® dsDNA BR assay kit on the Qubit® 2.0 fluorometer (Invitrogen). A 1:1 (insert:vector) DNA molar ratio was added to separate ligation reactions

⁶ Chimpanzee *TGM4* mRNA has a *BgIII* cut site. These constructs were not correct. Subsequently, pCMVscript-c*TGM4* was digested and transferred to pCMV-flag vector and used for reported *TGM4* assays

(total volume 10µl) and incubated with Fast T4 ligase (Thermofisher Scientific) in the thermocycler at the following temperatures: (22°C, 20')/(21°C, 5')/ (20°C, 5')/ (18°C, 5')/ (16°C, 5')/ (14°C, 5') / (65°C, 10')/(4°C, ∞). All of each ligation reaction (10µl) was transferred to 200µl of thawed TG1 chemically competent *E. coli* cells and put on ice for 20 minutes. Cells were heatshocked at 42°C for 45 seconds and immediately transferred to ice for two minutes. Cells were added to 500µl of LB broth in a culture tube and incubated at 37°C and 250 rpm for one hour. After incubation, 50µl and 200µl of each culture were spread onto LB carbenicillin plates and incubated 16-18 hours at 37°C. Single colonies were transferred to 3mL LB ampicillin broth and to an LB carbenicillin replicate plate, and incubated overnight at 37°C. Freezer stocks were made from 750µl of culture with 250µl 60% glycerol and stored at -80°C. The remaining 2.25mL culture's DNA was extracted with the IBI high speed plasmid mini kit (Midsci). Candidates were sequenced with BigDye® Terminator (Thermofisher Scientific) at the following conditions: (96.0°C, 10'')/50°C 5'')/60°C, 4')₃₅/(4°C, ∞), and analyzed on the Avant 3100 sequencer (Applied Biosystems).

Table 3.7: Universal primers used for pFastBac1 TGM4 amplification

Gene	Primer Sequence
TGM4 in pFastBac1	
Forward:	5' - GTACGATGACAGATCTAAAGGCCCTACGTCGACATGATGGATGCATCAAAGAGCTG-3'
Reverse:	5' - TCCTCTAGTACTTCTCGACAAGCTTGGTACCGCATGCCTCGA-3'

Our pCMV-flag (with *GFP*) vector worked with all our constructs; however, it was challenging to clone into because of the position of *GFP*. We ordered pCMV-script (Agilent Technologies) vector for easier cloning. TG1 chemically competent cells were thawed on ice, and 1µl of pCMV-script vector was transferred to the cells using a filtered tip. Cells were incubated on ice for 20 minutes, heat shocked at 42°C for 45 seconds, and then immediately

placed on ice for 2 minutes. The transformed mixture was transferred to 500µl of LB media in a culture tube and incubated for an hour at 37°C and 250 rpm. Both 50µl and 250µl were plated on LB kanamycin plates and incubated at 37°C overnight. pCMV-script colonies were selected and streaked onto a LB kanamycin replicate plate and grown overnight at 37°C. Two replicates were selected and grown overnight in 3mL of LB kanamycin broth, 250µl of each culture was mixed with 750µl of 60% glycerol and stored at -80°C. A 2mL LB kanamycin broth was inoculated and grown at 37°C for 6 hours before transferring 50µl into a 50mL LB kanamycin broth. 50mL cultures were grown at 37°C for 18 hours and DNA was extracted with the Pureyield™ plasmid midiprep system (Promega).

Chimpanzee *TGM4*, human-chimpanzee ancestor *TGM4*, and chimpanzee *SEMG2-Tyr* in pFastBac1 were cloned into the pCMV-script vector. First pCMV-script was digested with *SalI/Acc65I* and *NotI/Acc65I*, with and without fast Alkaline Phosphatase (ThermoFisher Scientific). However, none of the single *Acc65I* digests worked. Then pCMV-script, chimpanzee *TGM4*, human-chimpanzee ancestor *TGM4*, and chimpanzee *SEMG2-Tyr* were digested with *SalI/KpnI* (*TGM4s*) and *NotI/KpnI* (*SEMG2-Tyr*). *KpnI* single digests worked for the pCMV-script vector. Gel-purification, ligation, transformation, and verification processes were exactly like pCMV-Flag (with *GFP*) cloning.

Subsequently, chimpanzee *TGM4* in pCMV-script was subcloned into pCMV-flag vector because the original pCMV-flag chimpanzee *TGM4* was incorrect (see footnote 6, page 123). Both pCMV-script chimpanzee *TGM4* and pCMV-flag human *TGM4* were digested with *SalI* and *KpnI*. Both the chimpanzee *TGM4* and pCMV-flag backbone bands were gel-purified, ligated, and transformed as described previously.

3.2.6 Production/Purification of recombinant proteins

3.2.6.1 Gene expression in insect cell culture

Our pFastBac1 constructs were sent to our collaborators to transfect in *Sf9* insect cells. Media containing viral particles for human *TGM4*, chimpanzee *KLK3*, and human *SEMG1* were received. Samples were heated at 95°C for 5 minutes, and amplified with GoTaq® (Promega) at the following conditions: (95°C, 30'')/(95°C, 10''/ 55°C, 15''/ 72°C, 1')₃₅/ (72°C, 10') / 4°C, ∞) using the primers listed in Table 3.8. PCR products were separated on a 0.8% agarose gel with 0.008% ethidium bromide and imaged under UV light.

Table 3.8: Primers used for viral particle amplification

Gene	Primer Sequence
TGM4	
F420:	5' - CCCCCAATGCCATCCTGGGCAAGTACCAAC-3'
R720:	5' - CACACATGGCCCTGCACACCAGCACGGGG-3'
R1020:	5' - TGGTGATTTTCTCGCCATTCTCATTAC-3'
SEMG1	
F450:	5' - CATCTGGAAAGGGAATATCCAG-3'
R950:	5' - GGGATACATCTTTCTGCACACC-3'
R1200:	5' - GCCATGGCTCTTGCTTAGGA-3'
KLK3	
SeqInternal_Fwd:	5' - CTCATCCTGTCTCGGATTG-3'
Block_B_F:	5' - GATGCTGTGAAGGTCATGGA-3'
His_Rev:	5' - AATGGTGATGGTGATGGTG-3'

3.2.6.2 LNCaP cell culture

Androgen sensitive human metastatic prostate LNCaP cells were ordered from ATCC (Manassas, Virginia, lot number: 59410738). Cells were grown at 37°C with 5% CO₂ in Dulbeccos Modified Eagle's Medium, DMEM (GE Healthcare, Little Chalfont, United Kingdom) with 10% FBS (Atlanta Biologicals, Flowery Branch, GA) and 1% penicillin streptomycin (Cellgro, Corning, New York). Cells were dispersed by Trypsin EDTA (Life

Technologies) every 3-4 days and transferred to a new 25cm² flask (Sarstedt, Nümbrecht, Germany) in a 1:4 dilution.

Endogenous Kallikrein 3 should be expressed from LNCaP cells (Väisänen et al, 1999). Approximately 3µg of protein from LNCaP media was separated in a 10% SDS-PAGE (Invitrogen). An Immobilon®-FL PVDF (Millipore, Merck, Germany) was activated in methanol for one minute, and incubated in 1X NuPAGE (Invitrogen) transfer buffer with the separated gel and other cassette components for 10 minutes. The Western sandwich was assembled and put into the mini Trans-Blot® (Bio-Rad) chamber with an ice pack. The apparatus was placed on a stir plate with agitation, and proteins were transferred at 100 volts for 45 minutes. The PVDF membrane was rinsed with water, and incubated in 50%odyssey block/50% PBS for an hour and subsequently rinsed in dH₂O, three times for five minutes. Polyclonal rabbit anti-KLK3 and mouse anti-TGM4 primary antibodies (ABCAM) were diluted (1:1,000) in 50%odyssey block/50% PBST. The membrane was incubated overnight in primary antibody, washed three times for five minutes in PBS, and then incubated for an hour in secondary donkey anti-mouse IRDye® 680RD and anti-rabbit IRDye® 800CW (Licor, Lincoln, NE) antibodies (1:15,000). Following three washes in PBST for five minutes, the Western blot was imaged on the Odyssey FC (Licor) imager.

Approximately 80,000 LNCaP cells were plated in each well of a 12 well plate (Corning, Corning, NY) 24 hours before transfection. *SEMG1* and *SEMG2* (1µg) in pSG5 (provided by Miyamoto) were mixed with 7.5µl of FuGENE (Promega), and DMEM media up to 250µl, and were incubated at room temperature for 5 minutes. Then 50µl of each mixture was slowly dropped in a circular fashion to a designated well. The plate was returned to the incubator (37°C and 5% CO₂). 24 hours after transfection, 1mL of media was removed from each well and

replaced with 0.5mL of DMEM with 1% penicillin streptomycin. Media was collected and saved after 48 hours post transfection. Media was quantified, acetone precipitated, and Western blotted described later.

3.2.6.3 293T mammalian cell culture

Human embryonic kidney cells with T antigen (293T) were ordered from ATCC (lot number: 59587035). Cells were grown at 37°C with 5% CO₂ in DMEM (GE Healthcare), with 10% FBS (Atlanta Biologicals) and 1% penicillin streptomycin (Cellgro). Cells were dispersed by Trypsin EDTA (Life Technologies) every 2-5 days and transferred to a new 25cm² flask (Sarstedt) in a 1:4 or 1:10 dilution. Initially, *SEMG1* and *SEMG2* in pSG5 vectors were transfected as described for LNCaP cells. However, Jennifer (Vill) Doyle, optimized transfections with pCMV-*GFP* with polyethylinamine (PEI) transfection reagent.

Between 200,000-250,000 293T cells were plated with 1mL of DMEM, 10% FBS, and 1% penicillin streptomycin in a BioLite 12 well plate (Thermofisher Scientific) 24 hours before transfection. Then 1-2µg of pCMV- DNA, 1.63µl of PEI, and DMEM media up to 160µl was mixed together and incubated at room temperature for 20 minutes and then dropped into a designated well. Between 16-20 hours post transfection, total media was removed from the cells and replaced with 0.6 mL DMEM and 1% penicillin streptomycin. Media was collected at various time points (48, 72, 96, and 120 hours) post transfection. During optimization, cell lysate was also collected at various time points using Promega passive lysis buffer.

Transfections were up-scaled by plating 1,500,000 cells in 5mL DMEM, 10% FBS, and 1% penicillin streptomycin in a 25cm² flask. 5µg of pCMV-DNA, 1µg of pCMV-*GFP* DNA, 8.2µl of PEI, and DMEM media up to 800µl was mixed together and incubated for 20 minutes in a 15mL conical tube. To ensure a uniform transfection, media was removed from the 25cm² flask

and then 4.2mL of total media was added to the mixture and added to the 25cm² flask. GFP expression was imaged 48-72 hours post transfection using the confocal microscope. Media and lysate were quantified, purified, and assayed described later.

3.2.6.4 Protein quantification

Protein concentration was determined using either the Qubit® Protein Assay Kit (ThermoFisher Scientific) or the Bradford protein assay (Bio-Rad). For the Qubit® Protein Assay, samples were undiluted and diluted 1:20 before assaying. 2µl of diluted sample was added to Qubit® protein assay buffer (1µl reagent in 199µl buffer) with the total volume being 200µl. Samples were vortexed, incubated for 5 minutes, and then protein concentration was measured with the Qubit 2.0 (Life Technologies). In the Bradford assay, samples were undiluted before assaying. Then 20µl of sample was added to 1mL of Bio-Rad Bradford assay dye, mixed, and incubated for 5 minutes. Absorbance at 595nm was measured using the Genesys 10 UV spectrophotometer (Thermo Spectronic, Rochester, NY). Utilizing 7 BSA standards, a standard curve was generated relating absorbance to mg/mL concentration. Sample concentration was calculated by solving for “x” using the standard slope equation generated by the standards.

3.2.6.5 Acetone precipitation

Some transfected media from both LNCaP and 293T cell lines were acetone precipitated. Four times the volume of -20°C acetone was added to the sample in a 15mL conical tube and incubated on ice for an hour. Samples were centrifuged at 4°C for 10 minutes at 14,000 rpm. The supernatant was decanted and the pellet was dried. Once dried, PBS was used to re-suspend the sample to its original volume.

3.2.6.6 His purification and dialysis

Transfected media from 293T cells was purified with His-Pur™ cobalt columns (ThermoFisher Scientific). After purification, columns were washed with regeneration MES buffer (20mM 2-(N-morpholine)-ethanesulfonic acid, 0.1M sodium chloride: pH 5.0) and stored with 20% ethanol. Columns were labeled with the species and protein name and were only re-used with the same species/protein sample. Samples were dialyzed with slide-a-lyzer™ dialysis cassettes (ThermoFisher Scientific). TGM4 proteins were dialyzed in 7.5mM CaCl₂ and 5mM Tris (pH 7.5) buffer because it was the reaction buffer used in downstream applications.

3.2.6.7 Antibody detection of recombinant proteins

Media and cell lysate from transfected 293T cells were analyzed with traditional Western and dot blotting methods, with primary antibodies specific to the protein, or most commonly, an anti-HIS primary antibody (Gen Script). 10µg of media or lysate was separated in a 10% SDS-PAGE (Invitrogen). An Immobilon®-FL PVDF (Millipore, Merck, Germany) was activated in methanol for one minute, and incubated in 1X NuPAGE (Invitrogen) transfer buffer with the separated gel and other cassette components for 10 minutes. The Western sandwich was assembled and put into the mini Trans-Blot® (Bio-Rad) chamber with an ice pack. The apparatus was placed on a stir plate with agitation, and proteins were transferred at 100 volts for 45 minutes. The PVDF membrane was rinsed with water, and incubated in 50%odyssey block/50% PBS for an hour and subsequently rinsed in dH₂O, three times for five minutes. Primary antibodies specific to TGM4, KLK3, SEMG1, SEMG2, or HIS were diluted (1:1,000) in 50%odyssey block/50% PBST. The membrane was incubated overnight in primary antibody, washed three times for five minutes in PBST, and then incubated for an hour in secondary donkey anti-mouse IRDye® 680RD or anti-rabbit IRDye® 800CW (Licor) antibodies, diluted

1:15,000. Following three washes in PBST for five minutes, the Western blot was imaged on the Odyssey FC (Licor) imager.

Media and cell lysate samples (10 μ l, 100 μ l, and 200 μ l) were added to a dot blot apparatus (BioRad) over a dry nitrocellulose-FL (Licor) membrane. Samples were absorbed using a vacuum pump attached to a faucet. The membrane was blocked with 50% odyssey block/50% PBS for one hour, and rinsed three times with dH₂O for five minutes. Anti-HIS (or anti-TGM4) antibody, diluted 1:1,000 in 50%PBST/50% odyssey block, covered the membrane and was incubated overnight. The membrane was washed three times for five minutes in PBST, and then incubated for an hour in secondary goat anti-mouse IRDye® 800CW (Licor) antibody (diluted 1:15,000). After three washes in PBST for five minutes, the dot blot was imaged on the Odyssey FC (Licor) imager.

3.2.7 TGM4 assays

3.2.7.1 1D gel based assay

Peter et al. utilized monodansylcaverdine (MDC), calcium chloride reaction buffer, and dimethylcasein to assay prostate specific transglutaminase (TGM4) activity (1998). This assay was initially optimized using guinea pig transglutaminase (gpTGM). gpTGM (0.001 μ g, 0.001 μ g, 0.01 μ g, 0.5 μ g⁷, or 1 μ g) was incubated with either 2.5 or 5 μ g of casein (Sigma-Aldrich) in 10 μ l of reaction buffer (0.3mM MDC⁸, 7.5mM CaCl₂, 5mM Tris: pH 7.5) for one, two, or three⁹ hours at 37°C. 10 μ l of human seminal plasma, TGM4 transfected media, or TGM4 transfected cell lysate were incubated with 2.5 μ g of casein (Sigma-Aldrich) in 6 μ l of reaction buffer (0.3mM MDC, 7.5mM CaCl₂, 5mM Tris: pH 7.5) for three hours at 37°C. Each reaction was prepared in duplicate, and one had 2 μ l of 500mM EDTA (pH 8), which prevented enzymatic activity.

⁷ 0.5 μ g gpTGM is optimal when comparing activity to transfected media or cell lysate

⁸ Monodansylcaverdine (MDC) solubility is 10 mg/mL

⁹ Incubation for three hours was preferred

After enzyme/substrate incubation, 2 μ l of 500mM EDTA (pH 8) was added to all reactions to ensure enzyme termination. Then 5 μ l of loading buffer (Lammeli buffer (BioRad) with 10% β - Mercaptoethanol (Sigma Aldrich)) was added to each reaction and incubated at 95°C for 10 minutes. Samples were loaded into a 10% NuPAGE gel (Invitrogen) and electrophoresed at 150 volts for approximately 45 minutes. The gel was imaged under UV and then stained with coomassie (Licor) for thirty minutes. The gel was destained with dH₂O overnight and then imaged on the Odyssey FC (Licor) imager.

3.2.7.2 96 well plate reader assay

Zedira transglutaminase fluorescent assay was modified for a 96 well plate format (Darmstadt, Germany). Total volume of the reaction was adjusted to 250 μ l for each well (opposed to 1mL). First, 125 μ l of reaction buffer was warmed at 37°C (for a minimum of 10 minutes) in the Spectramax® i3x platform (Molecular Devices, Sunnyvale, CA) in a Corning® 96 well flat clear bottom black polystyrene plate (Thermofisher Scientific). Transfected media/lysate, HIS purified media/lysate, or guinea pig transglutaminase (Sigma Aldrich) was aliquoted into 0.2mL tubes with a 125 μ l total volume. Using a multichannel pipette, transglutaminase enzymes were transferred to the reaction buffer immediately before measurements were taken on the Spectramax® i3x platform under the parameters in Table 3.9.

Table 3.9: Spectramax® i3x parameters for TGM assays

Parameter	Specification
Excitation wavelength	332nm
Emission wavelength	500nm
Emission slit	15nm
Excitation slit	15nm
Read height (from top)	4.35- 4.53 mm
Time intervals between reads	1 minute
Total time	3.5 – 7 hours

3.3 Results

3.3.1 Generation of recombinant plasmids with human, chimpanzee, and human-chimpanzee ancestral mRNA sequences

TGM4 and *KLK3* cDNA (or gBlocks) were cloned into pFastBac1 for human, chimpanzee, and the hypothetical human-chimpanzee ancestor and then transferred into the pCMV vector with methods described in section 3.2.1 (Table 3.10). Alignments for human, chimpanzee, and human-chimpanzee ancestor *TGM4* in pCMV are included in the appendix (A.14), since these constructs' enzymatic activity (post-expression) were assayed. Human and chimpanzee *SEMG1* and *SEMG2* cDNAs (with the exception of chimpanzee *SEMG1*) were cloned into pFastBac1, sequence confirmed, and transferred into the pCMV vector with methods described in section 3.2.1 (Table 3.10). All pCMV constructs were transfected into mammalian 293T cells and some of them had confirmed expression, in the media or lysate (summarized in Table 3.10; expanded on in section 3.3.2).

Table 3.10: Summary of recombinant plasmids

Clone	pFastBac1	pCMV vector
Human <i>TGM4</i> -HIS	Sequence confirmed	Sequence confirmed; protein expressed in the lysate
Human <i>SEMG1</i> -HIS	Sequence confirmed	Sequence confirmed; protein expressed in the media
Human <i>SEMG2</i> -HIS	Sequence confirmed	Sequence confirmed; protein expressed in the media
Human <i>KLK3</i> -HIS	Sequence confirmed	Sequence confirmed
Chimp <i>TGM4</i> -HIS	Sequence confirmed	Sequence confirmed; protein <i>probably</i> expressed in the lysate
Chimp <i>SEMG1</i> -HIS	-----	Sequence confirmed; protein expressed in the media
Chimp <i>SEMG2</i> -HIS	Sequence confirmed	Sequence confirmed; protein expressed in the media
Chimp <i>SEMG2</i> Tyr-HIS	Sequence confirmed	Sequence confirmed
Chimp <i>KLK3</i> -HIS	Sequence confirmed	Sequence confirmed
H/C Anc <i>TGM4</i> -HIS	Sequence confirmed	Sequence confirmed; protein expressed in the lysate
H/C Anc <i>KLK3</i> -HIS	Sequence confirmed	Sequence confirmed

3.3.2 Expression of recombinant proteins in mammalian cell culture

Seminal proteins by nature are extracellular; therefore, their coding sequence should contain a secretion signal peptide. All of our recombinant plasmids (Table 3.10) were cloned with a HIS-tag before the native stop codon but were otherwise natural in sequence. In early optimization experiments, recombinant plasmids containing GFP had intercellular expression peaking between 48 and 72 hours post-transfection (*data not shown*). However, expression of our recombinant proteins in media occurred at different time points. Recombinant TGM4 proteins are first detected in media 72 hours post-transfection but have increased expression at 96 hours post-transfection (Figure 3.4). Recombinant SEMG proteins were detected in the media 48 hours post-transfection (Figure 3.4).

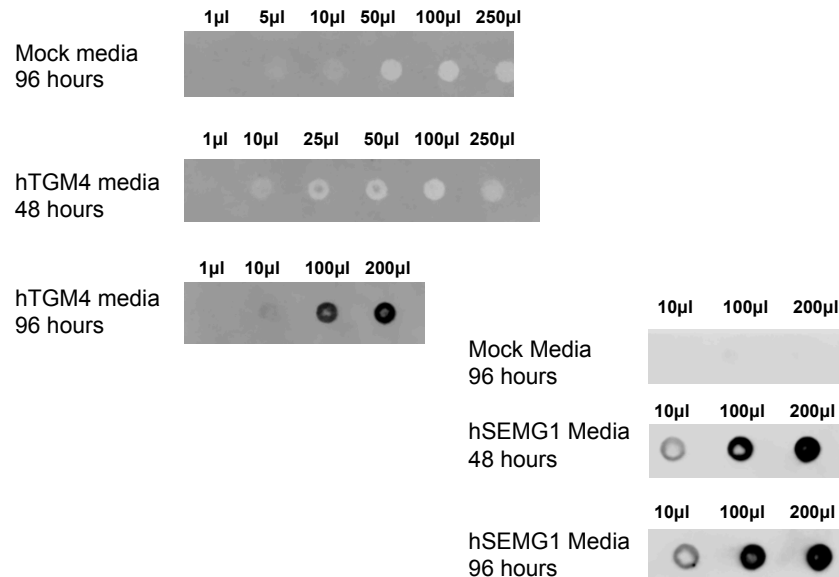


Figure 3.4: Differential timing of secretion of recombinant proteins

An anti-HIS antibody was used to detect recombinant proteins (designed with a HIS tag) from transfected cell media aliquoted onto a nitrocellulose membrane. Mock media, where no DNA was transfected in 293T cells was used as a negative control in each experiment. Human TGM4 (hTGM4) and human SEMG1 (hSEMG1) are representative orthologs of other recombinant proteins (H/C ancestor TGM4, human SEMG2, chimpanzee SEMG1, and chimpanzee SEMG2). TGM4s are first detected in media collected 96 hours after transfection; whereas, SEMGs are detected in the media collected from 48 hours after transfection.

3.3.3 Transglutaminase assay

3.3.3.1 Optimization of the 1D gel assay

This assay was initially optimized using 1 μg of guinea pig transglutaminase (gpTGM) in reaction buffer containing either 2.5 μg or 5 μg of casein. Either amount of casein yielded a positive reaction, and visually there was not a large difference between reactions (*data not shown*). For subsequent reactions, 2.5 μg of casein was used. gpTGM (0.001 μg , 0.01 μg , 0.1 μg , 1 μg) was incubated with reaction buffer/casein for one hour at 37°C, and activity was only visible for 0.1 μg and 1 μg of gpTGM (Figure 3.5).

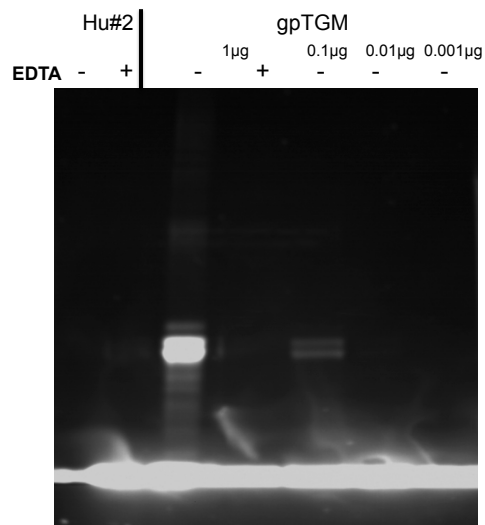


Figure 3.5: Transglutaminase activity detected as low as 0.1 μg of gpTGM enzyme

Human seminal plasma (Hu#2) does not appear to have functional transglutaminase activity. Both 0.1 μg and 1 μg of gpTGM is able to cross-link MDC to casein.

Incubation at 37°C for variable lengths of time (one hour, two hours, or three hours) was tested with 0.01 μg , 0.1 μg , and 1 μg of gpTGM and activity was visualized on a 1D gel.

Transglutaminase activity was detected after one hour of incubation for 0.1 (faintly) and 1 μg of enzyme; however, activity was easier to detect after a two hour incubation. Transglutaminase activity was not detected after 3 hours of incubation (at 37°C) for 0.01 μg gpTGM (*data not shown*).

3.3.3.2 Human SEMG1 Ia peptides are cross-linked by transglutaminases

Megan Hockman, a previous undergraduate in our lab produced human SEMG1.Ia fragments for her honors thesis. The SEMG Ia fragments are enriched with glutamine and lysine residues and are the substrates for TGM4. Likely, SEMG Ia fragments would be a substrate for gpTM. In fact after a 3 hour incubation at 37°C, gpTGM was able to crosslink hSEMG1.Ia proteins together (Figure 3.6). However, enzyme efficiency was noticeably lower than crosslinking casein. Transfected human TGM4 media was able to crosslink casein, but not hSEMG1.Ia peptides.

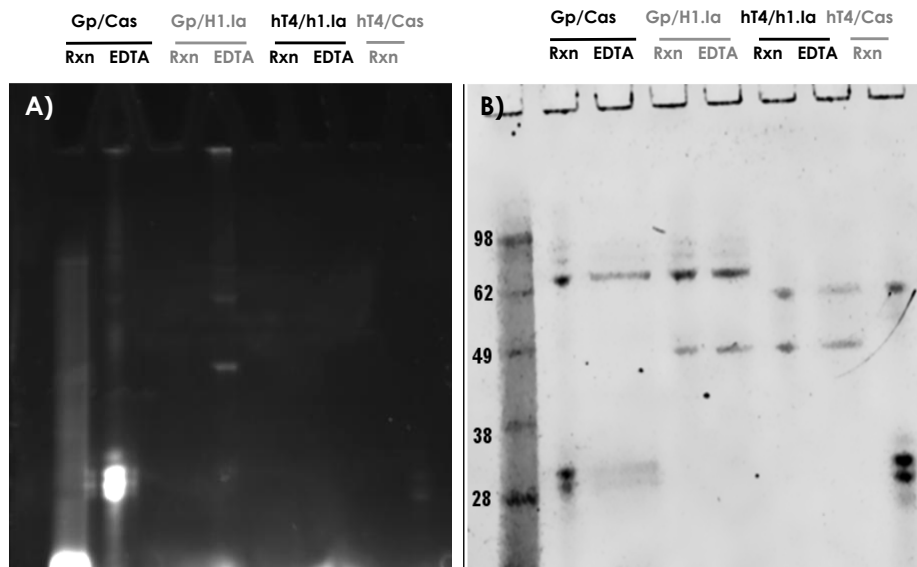


Figure 3.6: Guinea pig transglutaminase has activity on hSEMG1.Ia peptides and transfected hTGM4 media has activity with casein

A) GpTGM is able to crosslink casein (Gp/Cas); gpTGM is able to weakly crosslink hSEMG1.Ia (Gp/H1.Ia), and as expected, the reaction is terminated in the presence of EDTA. Human TGM4 media is able to crosslink casein (hT4/Cas), but not hSEMG1.Ia (ht4/h1.Ia). **B)** The SDS-PAGE gel was stained with coomassie and imaged on the Licor odyssey imager after UV detection.

3.3.3.3 Recombinant TGM4 proteins lose function after HIS purification

TGM4 loses activity after HIS-tag purification (Figure 3.7), even though multiple Western and dot blots confirmed retention of the protein (Figure 3.8). Dialysis was attempted to regain activity of TGM4, but was not successful (*data not shown*).

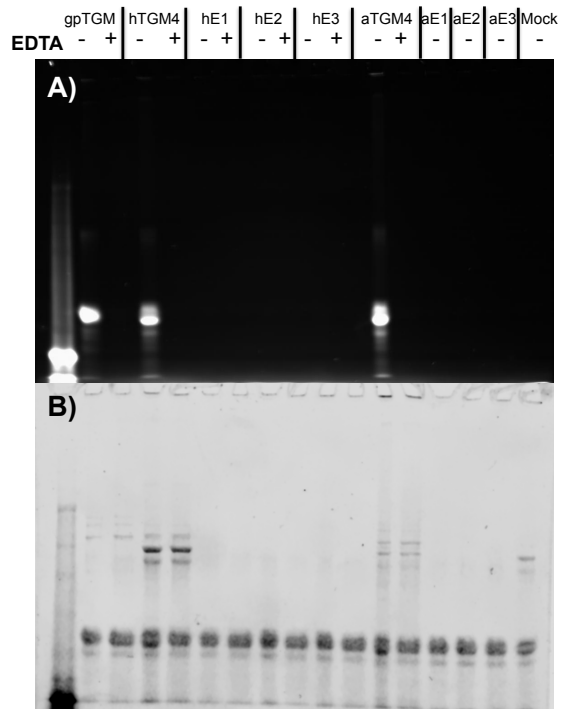


Figure 3.7: Recombinant human and ancestor TGM4 have enzymatic activity before but not after HIS purification

Guinea pig transglutaminase (gpTGM) is able to crosslink casein (gpTGM/-) and as expected, the reaction is terminated in the presence of EDTA (gpTGM/+). Human TGM4 media is able to crosslink casein (hT4/-); however, post-HIS purification activity is lost (hE1, hE2, and hE2). Ancestor TGM4 media actively crosslinks casein (aTGM4/-), but activity is lost after HIS purification (aE1, aE2, and aE3). Mock transfected media has no detectable cross-linking activity. **A)** SDS-PAGE gel detected under UV light. **B)** The same SDS-PAGE gel stained with coomassie and imaged on the Licor Odyssey FC.

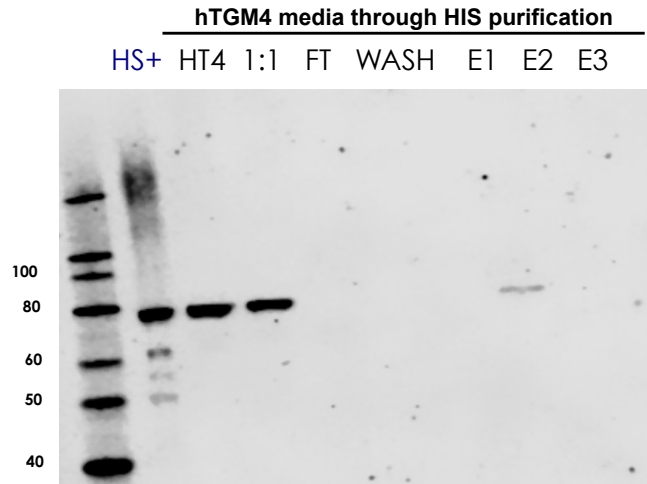


Figure 3.8: Recombinant human TGM4 is purified from media using HIS columns

Anti-TGM4 Western blot detects TGM4 in human seminal plasma (HS+), human transfected media (HT4), HT4 diluted 1:1 with equilibrium buffer (1:1), as well as in elution 2 (E2) post HIS purification. Nothing is detected in the flow through (FT- proteins unbound to cobalt resin) or washes.

3.3.4 Comparison of TGM4 enzymatic activity of human and the human-chimp ancestor

Consistently, transfected TGM4 lysates have more MDC cross-linked to casein (brighter banding) compared to transfected media (Figure 3.9). In addition, mock-transfected lysates do not have any enzymatic activity (no banding), which supports that all the activity shown in TGM4 transfected lysate is from TGM4. Chimpanzee TGM4 transfected media or lysates yielded no or minimal activity (Figure 3.9) despite multiple attempts. Human-chimpanzee ancestor TGM4 has significantly ($p < 0.0007632$, Welch's two sample t-test) higher activity compared to human TGM4 (Figure 3.10; 3.11).

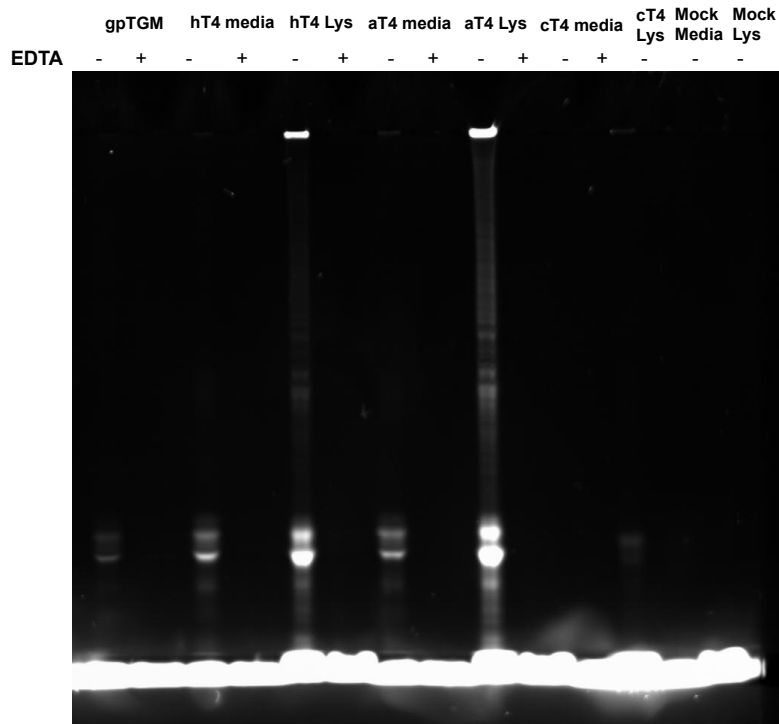


Figure 3.9: Human and ancestor TGM4 lysates are more efficient than media

Guinea pig transglutaminase (gpTGM) was used as a positive control. Any reaction in the presence of EDTA (+) should not have activity. Equal volumes of transfected TGM4 media or lysate were assayed with MDC and casein. Mock media and lysate have no activity. Chimpanzee TGM4 lysate has very little activity. Both human and ancestor TGM4 lysates have a lot of activity, and have some activity in their transfected media.

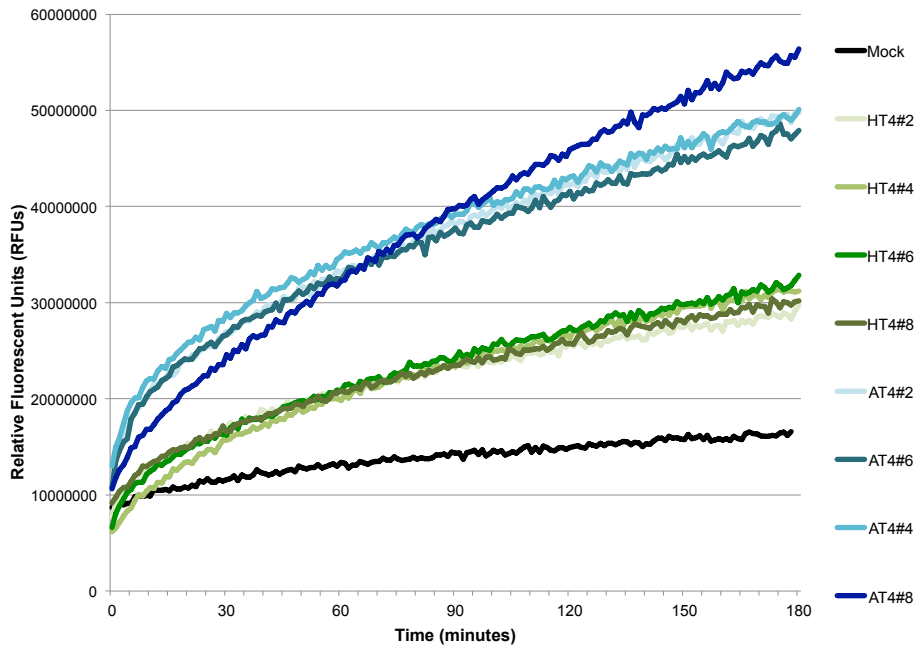
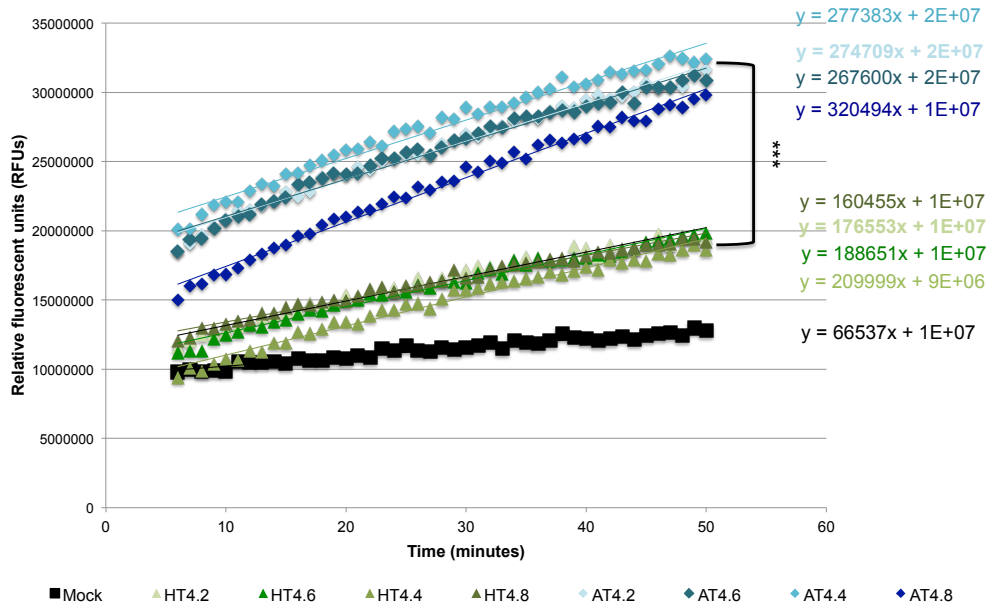


Figure 3.10: Human-chimpanzee ancestor TGM4 has significantly higher enzymatic activity compared to human TGM4.

Equal concentrations of TGM4 enzymes were assayed with MDC and casein and fluorescence was measured overtime. All samples were from transfected cell lysates. Graphs A and B are from the same data set, but differ in the length of the x-axis. All human-chimpanzee ancestor TGM4 transfected recombinant protein replicates (2, 4, 6, 8) are blue lines, and represent the average fluorescence across four technical replicates. Likewise, human TGM4 transfected recombinant protein replicates (2, 4, 6, 8) are green lines, and data points represent the average fluorescence across four technical replicates. Mock lysate, the black line, is from one mock- transfection, but represents the average of four technical replicates. Slopes derived from samples in the linear phase of the assay (Graph A) are proportional to the enzyme’s ability to catalyze MDC-casein cross-linkage. A steeper slope indicates greater velocity. The average slope of human TGM4 was 183914.5 while the average slope of ancestor TGM4 was 212935.24. Ancestor TGM4 has ~1.2 fold increase in velocity, which is significant ($p < 0.0007632$, Welch’s two sample t-test), compared to human TGM4s activity.

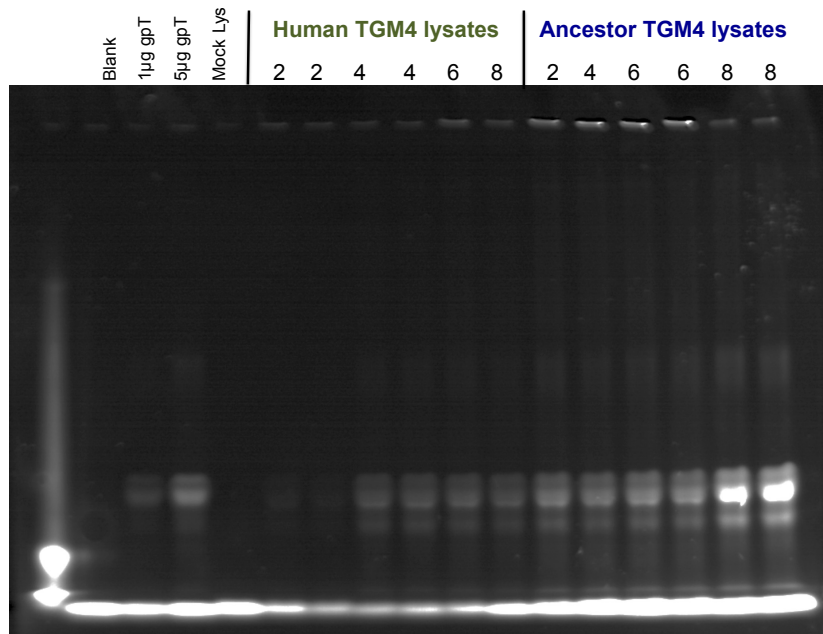


Figure 3.11: Recombinant human-chimpanzee ancestor TGM4 has more activity than recombinant human TGM4

Samples assayed in Figure 3.10, were separated on an SDS PAGE gel after their fluorescence was measured over time. This gel confirms that mock lysate does not have any activity. Human-chimpanzee ancestor TGM4 reactions have cross-linked more MDC to casein compared to human TGM4. Numbers under human or ancestor TGM4 titles identify the transfected recombinant protein replicate. Each lane has two of the four assay reaction technical replicates loaded.

3.4 Discussion

Both human and human-chimpanzee ancestor TGM4s express more activity in lysates from transfected 293T cells than the media (Figure 3.9). My initial experiments concluding that TGM4 activity was present 96 hours after transfection in the media, may be a result of dead cells lysing and releasing the proteins, opposed to the TGM4s being slowly secreted into the media (Figure 3.4). Signal peptide sequence analysis of TGM4s did not recognize a secretion signal peptide. TGM4s are secreted from the prostate gland, and it is possible that they have an unknown tissue-specific signal peptide for secretion. Therefore, LnCAP (derived from prostate) cells may be able to natively secrete transfected TGM4s better than 293T cells. However, we initially chose to use 293T cells because of their transfection efficiency. Alternatively, a secretion signal peptide sequence could be genetically engineered into the current pCMV constructs to increase secretion in our optimized 293T cells.

Production of recombinant chimpanzee TGM4 has consistently been less than both human and ancestor TGM4 (Figure 3.9). I have thoroughly analyzed the DNA sequence, in both the TGM4 and vector promoter sequences for any mutations that could cause this effect (Appendix A.10). I have cloned chimpanzee TGM4 several times, have transformed it into two different types of competent *E. coli* strains, and I have attempted recombinant plasmid purification with multiple methods. There are a few synonymous differences between my clones, which could account for variation if there was codon usage bias. Upon review, some codons had decreased usage in human cell lines, although all codons were utilized. My hypothesis is that chimpanzee TGM4 is incorrectly folding in vitro; this could either be because the native protein needs additional chaperones to fold correctly, or potentially the HIS tag is interfering with folding.

Recombinant human-chimpanzee ancestor TGM4 significantly has ~1.2 fold greater velocity compared to recombinant human TGM4. This indicates that our last common ancestor with chimpanzee had a functional TGM4 with faster cross-linking activity than extant humans. It would have been ideal if we were able to produce and assay chimpanzee TGM4's activity, which would have provided a more complete picture.

However, we do know that gorillas have presumably a functionless TGM4, as it was not identified in seminal plasma (Chapter 2) and an 11bp deletion in its gene sequence has predicted pseudogenization (Clark and Swanson, 2005; Carnahan and Jensen-Seaman, 2008). The relative abundance of TGM4 in our ancestor's seminal plasma will always be unknown; however, we know that chimpanzees have 7-fold higher expression of TGM4 compared to humans (Chapter 2). It is likely that over the past 6Mya, humans have decreased expression of TGM4 and have also reduced transglutaminase activity. Considering that TGM4 is important in semen coagulation and the formation of the copulatory plug in chimpanzees, increased efficiency and/or abundance may be associated with high levels of sperm competition. The increased activity of human-chimpanzee ancestor TGM4 compared to human TGM4 indicates that the ancestor may have had higher sperm competition. This study supports a more chimpanzee-like mating system, with polyandrous females, in our last common ancestor with chimpanzee.

3.5 Future Directions

First and foremost, I would make a few experimental changes to have all the TGM4s secreted into the media and elevate chimpanzee TGM4 protein expression to relatively similar levels as human TGM4. I would clone all the TGM4s into a new vector that included a secretion signal peptide, or I would insert a secretion signal peptide in front of the TGM4 sequence in the current vectors. The addition of the signal peptide should allow expression of TGM4s to be secreted from mammalian cells and into the media. Concurrently, these changes may increase the expression of chimpanzee TGM4, but if not, I would first remove the HIS-tag sequence from the recombinant DNA plasmid. The positively charged HIS-tag may be interacting strangely with the chimpanzee TGM4 protein causing it to misfold and be degraded.

In addition to optimizing production of TGM4 proteins, I would make a few modifications to the plate-reader assay. Initially, I attempted to make my own reaction mixture for the plate reader, but was unsuccessful. Once I began using the Zedira transglutaminase assay kit, I was able to see enzymatic activity and presented those results in this dissertation. However, I would want to experimentally test different amounts of substrate, different substrates, change the pH conditions, or change the calcium concentrations of the reaction buffer. Therefore, I would revisit developing my own reaction mixture that worked as well as the Zedira transglutaminase assay kit. I would start by reducing the concentration of MDC or increasing the concentration of calcium in my original reaction base.

Currently this study only looked at the enzymatic activity of two recombinant TGM4s on a common substrate, casein. However, I would want to assay the activity of recombinant TGM4s on physiological substrates like SEMG1, SEMG2, and possibly CRTAC1. In addition, I would design an experiment to increase the amount of substrate to determine the K_m as well as the

V_{max} of TGM4. I would want to look at the specificity of enzyme/substrate pairings, by testing same species TGM4/substrate activity in comparison to one species TGM4 with another species substrate activity.

This study focused on human and human/chimpanzee ancestor TGM4 activity. Besides assaying chimpanzee TGM4 activity, it would be interesting to generate the chimpanzee/gorilla ancestor TGM4, the orangutan TGM4, and the macaque TGM4. These species TGM4s would provide a more complete picture of the evolution of TGM4 across species with differing sexual selection. Depending on these findings, I would suggest looking at “transitional” TGM4 proteins, by using site-directed mutagenesis to modify TGM4 one evolutionary step (amino acid change) at a time to either its derived or ancestral protein.

3.6 References

- Brillard-Bourdet, Michèle, et al. "Amidolytic Activity of Prostatic Acid Phosphatase on Human Semenogelins and Semenogelin-derived Synthetic Substrates." *European Journal of Biochemistry*, vol. 269, no. 1, 2002, pp. 390–395.
- Carnahan, Sarah J., and Michael I. Jensen-Seaman. "Hominoid Seminal Protein Evolution and Ancestral Mating Behavior." *American Journal of Primatology*, vol. 70, 2008, pp. 939–948.
- Carnahan-Craig, S. J., and M. I. Jensen-Seaman. "Rates of Evolution of Hominoid Seminal Proteins Are Correlated with Function and Expression, Rather than Mating System." *Journal of Molecular Evolution*, Nov. 2013, pp. 1–13, doi:10.1007/s00239-013-9602-z.
- Chang, Belinda SW, et al. "Recreating a Functional Ancestral Archosaur Visual Pigment." *Molecular Biology and Evolution*, vol. 19, no. 9, 2002, pp. 1483–1489.
- Chapais, Bernard. "Monogamy, Strongly Bonded Groups, and the Evolution of Human Social Structure." *Evolutionary Anthropology: Issues, News, and Reviews*, vol. 22, no. 2, 2013, pp. 52–65.
- Chapais, Bernard. *Primeval Kinship: How Pair-Bonding Gave Birth to Human Society*. Harvard University Press, 2009.
- Clark, Nathaniel L., and Willie J. Swanson. "Pervasive Adaptive Evolution in Primate Seminal Protein." *PLoS Genetics*, vol. 1, no. 35, 2005, pp. 335–342.
- Dean, Matthew D. "Genetic Disruption of the Copulatory Plug in Mice Leads to Severely Reduced Fertility." *PLoS Genet*, vol. 9, no. 1, Jan. 2013, p. e1003185, doi:10.1371/journal.pgen.1003185.
- Dixon, Alan F. *Sexual Selection and the Origins of Human Mating Systems*. Oxford University Press, 2009.
- Esposito, Carla, and Ivana Caputo. "Mammalian Transglutaminases." *FEBS Journal*, vol. 272, no. 3, 2004, pp. 615–631.
- Flori, Federica, et al. "The GPI-anchored CD52 Antigen of the Sperm Surface Interacts with Semenogelin and Participates in Clot Formation and Liquefaction of Human Semen." *Molecular Reproduction and Development*, vol. 75, no. 2, 2008, pp. 326–335.
- Folk, JE. "Mechanism and Basis for Specificity of Transglutaminase-catalyzed (G-Glutamyl) Lysine Bond Formation." *Adv. Enzymol. Relat. Areas Mol. Biol.*, vol. 54, 1983, pp. 1–56.
- Gavrilets, Sergey. "Human Origins and the Transition from Promiscuity to Pair-Bonding." *PNAS*, vol. 109, no. 25, 2012, pp. 9923–9928.
- Greenberg, Charles S., et al. "Transglutaminases: Multifunctional Cross-Linking Enzymes That Stabilize Tissues." *The FASEB Journal*, vol. 5, no. 15, 1991, pp. 3071–3077.
- Harms, Michael J., and Joseph W. Thornton. "Analyzing Protein Structure and Function Using Ancestral Gene Reconstruction." *Current Opinion in Structural Biology*, vol. 20, 2010, pp. 360–366.
- Hurle, Belen, et al. "Comparative Sequence Analyses Reveal Rapid and Divergent Evolutionary Changes of the WFDC Locus in the Primate Lineage." *Genome Research*, vol. 17, 2007, pp. 276–286.
- Lilja, H. "A Kallikrein-like Serine Protease in Prostatic Fluid Cleaves the Predominant Seminal Vesicle Protein." *Journal of Clinical Investigation*, vol. 76, no. 5, 1985, p. 1899.
- Lovejoy, C. Owen. "Reexamining Human Origins in Light of *Ardipithecus Ramidus*." *Science*, vol. 326, no. 5949, 2009, p. 74–74e8.
- Lovejoy, C. Owen. "The Great Divides: *Ardipithecus ramidus* Reveals the Postcrania of Our Last Common Ancestors with African Apes." *Science*, vol. 326, no. 5949, 2009, pp. 73–106.
- Nakahashi, Wataru, and Shiro Horiuchi. "Evolution of Ape and Human Mating Systems." *Journal of Theoretical Biology*, vol. 296, 2012, pp. 56–64.
- Peter, Anders, et al. "Semenogelin I and Semenogelin II, the Major Gel-Forming Proteins in Human Semen, Are Substrates for Transglutaminase." *Eur. J. Biochem*, vol. 252, 1998, pp. 216–221.

- Robert, Martin, et al. "Characterization of Prostate-Specific Antigen Proteolytic Activity on Its Major Physiological Substrate, the Sperm Motility Inhibitor/Precursor/Semenogelin I." *Biochemistry*, vol. 36, no. 13, 1997, pp. 3811–3819.
- Tian, Xin, et al. "Gene Birth, Death, and Divergence: The Different Scenarios of Reproduction-Related Gene Evolution." *Biology of Reproduction*, vol. 80, no. 4, 2009, pp. 616–621.
- Yokoyama, Shozo, et al. "Elucidation of Phenotypic Adaptations: Molecular Analyses of Dim-Light Vision Proteins in Vertebrates." *Proceedings of the National Academy of Sciences*, vol. 105, no. 36, 2008, pp. 13480–13485.

APPENDIX

A.1 Script in R for protein quantification and protease inhibitor analysis

A.1.1 Quantification method mini-comparison

```
#BRADFORD STANDARDS to generate standard curve
BR_known_Std_conc <- c(0.125, 0.25, 0.5, 0.75, 1, 1.5, 2)
BR_Std_Abs <- c(0.176, 0.397, 0.635, 0.846, 1.021, 1.261, 1.481)
BRStdDF <- data.frame(BR_known_Std_conc, BR_Std_Abs)
BRStd_reg <- lm(BR_Std_Abs ~ BR_known_Std_conc)
summary(BRStd_reg)

##
## Call:
## lm(formula = BR_Std_Abs ~ BR_known_Std_conc)
##
## Residuals:
##      1      2      3      4      5      6      7
## -0.1540 -0.0165  0.0545  0.0985  0.1065  0.0125 -0.1015
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.24650    0.06992   3.525 0.016822 *
## BR_known_Std_conc 0.66800    0.06484  10.303 0.000148 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1081 on 5 degrees of freedom
## Multiple R-squared:  0.955, Adjusted R-squared:  0.946
## F-statistic: 106.1 on 1 and 5 DF,  p-value: 0.0001482

conc.calc <- function(Slope, Intercept, Abs, DF)
{((Abs-Intercept)/(Slope))*DF}

#TEST SAMPLES qbit standards in Bradford
qbSamplenames <- c("200ug/mL", "200ug/mL", "200ug/mL", "400ug/mL", "400ug/mL",
", "400ug/mL")
qbAbsorbances <- c(0.265, 0.234, 0.234, 0.445, 0.448, 0.457)
qbSampleconcs <- conc.calc(BRStd_reg$coefficients[2], BRStd_reg$coefficients[
1], qbAbsorbances, 1)
qbSampleconcs_in_ug_mL <- qbSampleconcs *1000
Qb_concs_DF <- data.frame(qbSamplenames, qbAbsorbances, qbSampleconcs, qbSam
pleconcs_in_ug_mL)

#Stats
two_qbitstds <- c(200, 200, 200)
two_qbit_measured_concs <- c(27.69, -18.71, -18.71) #in ug/mL
```

```

Twoqt <- t.test(two_qbitstds, two_qbit_measured_concs, paired=TRUE)
Twoqt

## Paired t-test
##
## data: two_qbitstds and two_qbit_measured_concs
## t = 13.141, df = 2, p-value = 0.005741
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 136.6956 269.7910
## sample estimates:
## mean of the differences
## 203.2433

four_qbitstds <-c(400, 400, 400)
four_qbit_measured_concs <-c(297.15, 301.65, 315.12)
fourqt <- t.test(four_qbitstds, four_qbit_measured_concs, paired=TRUE)
fourqt

##
## Paired t-test
##
## data: four_qbitstds and four_qbit_measured_concs
## t = 17.664, df = 2, p-value = 0.00319
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 72.13161 118.58839
## sample estimates:
## mean of the differences
## 95.36

all_qbitstds <-c(200, 200, 200, 400, 400, 400)
all_qbit_measured_concs <-c(27.69, -18.71, -18.71, 297.15, 301.65, 315.12)
allqt <- t.test(all_qbitstds, all_qbit_measured_concs, paired=TRUE)
allqt

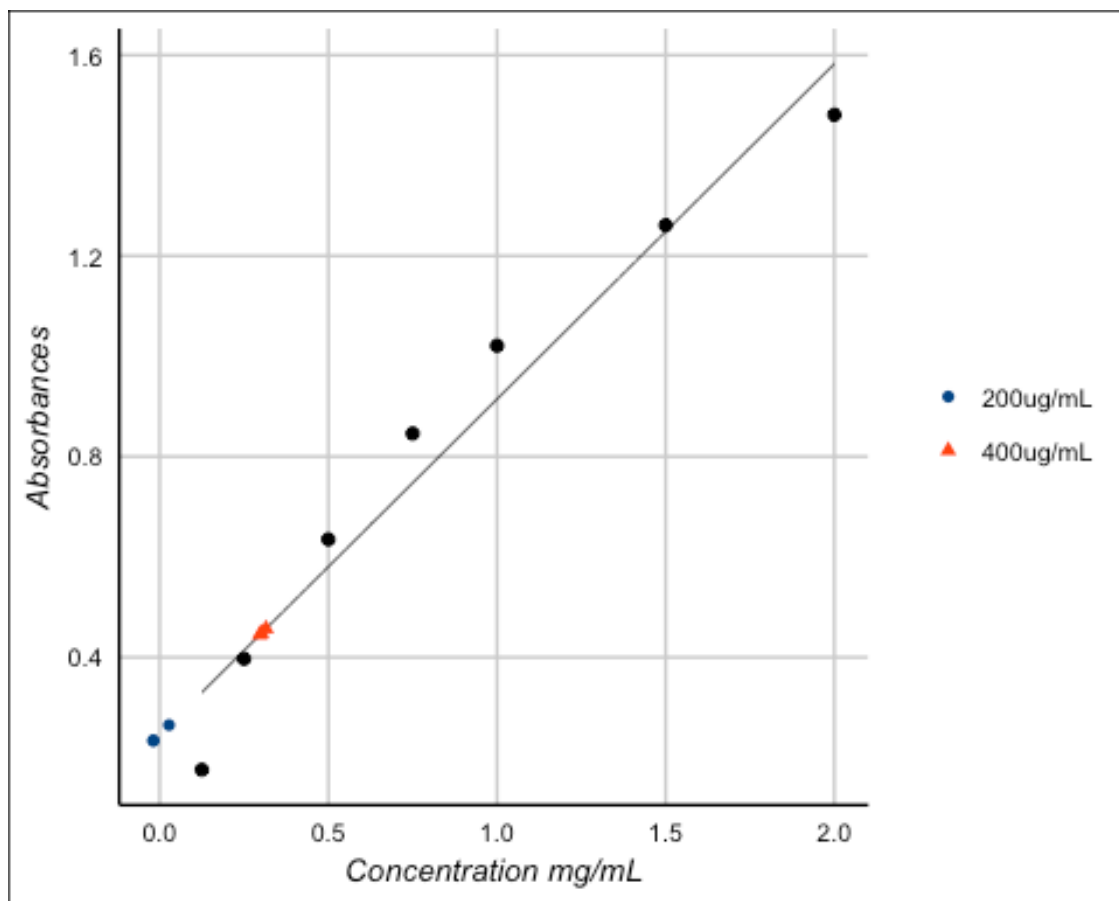
##
## Paired t-test
##
## data: all_qbitstds and all_qbit_measured_concs
## t = 5.922, df = 5, p-value = 0.001957
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 84.49377 214.10956
## sample estimates:
## mean of the differences
## 149.3017

library(ggplot2)
library(ggthemes)

```

```
#Qbit in Bradford Graph
```

```
ggplot(data= BRStdDF,  
       aes(x= BR_known_Std_conc,  
           y= BR_Std_Abs)) +  
geom_point() +  
geom_smooth(method = lm,  
            se = FALSE, size = 0.25, color = "black") +  
geom_point(data = Qb_concs_DF,  
           aes(x = qbSampleconcs,  
               y = qbAbsorbances,  
               color = qbSamplenames,  
               shape= qbSamplenames)) +  
theme_gdocs() + scale_color_calc() +  
theme(legend.title=element_blank()) +  
labs(x="Concentration mg/mL") + labs(y= "Absorbances") +  
theme(text=element_text(size=10, family="Arial"))
```



```
#Bradford standards in qbit
```

```
Brstandards_in_ug_ml <- c(125, 250, 500, 750, 1000, 1500, 2000)  
Brstandards_measured_concs <- c(328, 605, 606, 1451, 2200, 3000, 3450)  
BRinqb <- data.frame(Brstandards_in_ug_ml, Brstandards_measured_concs)  
BRinqbt <- t.test(Brstandards_in_ug_ml, Brstandards_measured_concs, paired=TR)
```

```

UE)
BRinQBt

##
## Paired t-test
##
## data: Brstandards_in_ug_ml and Brstandards_measured_concs
## t = -3.5089, df = 6, p-value = 0.01269
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1337.2574 -238.4569
## sample estimates:
## mean of the differences
## -787.8571

LowBRinQB <- data.frame("lstd"= c(125, 250, 500), "lmeasured"= c(328, 605, 606))
lBRinQBt <- t.test(LowBRinQB$lstd, LowBRinQB$lmeasured, paired=TRUE)
lBRinQBt

##
## Paired t-test
##
## data: LowBRinQB$lstd and LowBRinQB$lmeasured
## t = -3.0545, df = 2, p-value = 0.09255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -533.11323 90.44657
## sample estimates:
## mean of the differences
## -221.3333

HighBRinQB <- data.frame("hstd"= c(750, 1000, 1500, 2000), "hmeasured"= c(1451, 2200, 3000, 3450))
HBRinQBt <- t.test(HighBRinQB$hstd, HighBRinQB$hmeasured, paired=TRUE)
HBRinQBt

##
## Paired t-test
##
## data: HighBRinQB$hstd and HighBRinQB$hmeasured
## t = -6.6355, df = 3, p-value = 0.006973
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1794.4001 -631.0999
## sample estimates:
## mean of the differences
## -1212.75

```

A.1.2 Seminal plasma concentrations

```
# STANDARDS
seStd_conc <- c(0.125, 0.25, 0.5, 0.75, 1, 1.5, 2)
seStd_Abs <- c(0.212, 0.34, 0.621, 0.9, 1.099, 1.313, 1.46)
seStdDF <- data.frame(seStd_conc, seStd_Abs)
seStd_reg <- lm(seStd_Abs ~ seStd_conc)
summary(seStd_reg)

##
## Call:
## lm(formula = seStd_Abs ~ seStd_conc)
##
## Residuals:
##      1      2      3      4      5      6      7
## -0.12934 -0.08600  0.02569  0.13537  0.16506  0.04043 -0.15120
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.25668    0.08906   2.882 0.034504 *
## seStd_conc   0.67726    0.08259   8.200 0.000439 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1377 on 5 degrees of freedom
## Multiple R-squared:  0.9308, Adjusted R-squared:  0.917
## F-statistic: 67.25 on 1 and 5 DF,  p-value: 0.0004388

#SEMEN SAMPLES
seSamplenames <- c("H1", "H2", "H3", "H4", "H5", "C1", "C2", "C3", "CLJ", "G1",
", "G2", "G3", "G4")
seAbsorbances <- c(0.662, 0.671, 0.452, 0.559, 0.537, 0.868, 0.748, 0.701, 1.
00, 0.647, 0.684, 0.897, 0.721)
seSpecies <- c("Human", "Human", "Human", "Human", "Human", "Chimpanzee", "Ch
impanzee", "Chimpanzee", "Chimpanzee", "Gorila", "Gorila", "Gorila", "Gorila"
)

conc.calc <- function(Slope, Intercept, Abs, DF)
{((Abs-Intercept)/(Slope))*DF}

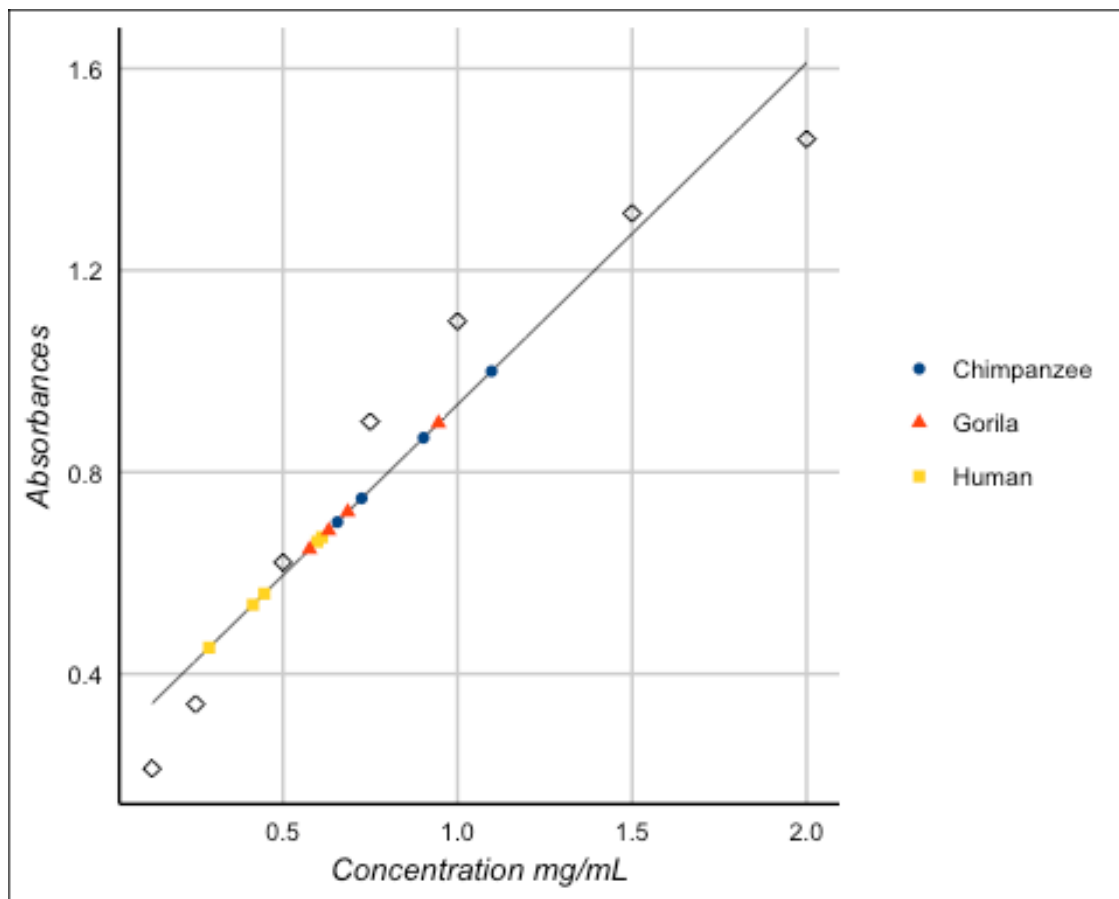
seSampleconcs <- conc.calc(seStd_reg$coefficients[2], seStd_reg$coefficients[
1], seAbsorbances, 1)
seSampleconcs20 <- conc.calc(seStd_reg$coefficients[2], seStd_reg$coefficien
s[1], seAbsorbances, 20)
Semen_concs_DF <- data.frame(seSpecies, seSamplenames, seAbsorbances, seSampl
econcs, seSampleconcs20)

library(ggplot2)
library(ggthemes)
```

```

ggplot(data= seStdDF,
      aes(x= seStd_conc,
          y= seStd_Abs)) +
geom_point(shape=5) +
geom_smooth(method = lm,
            se = FALSE,
            size = 0.25,
            color = "black") +
geom_point(data = Semen_concs_DF,
          aes(x = seSampleconcs,
              y = seAbsorbances,
              color = seSpecies,
              shape = seSpecies)) +
theme_gdocs() + scale_color_calc() +
theme(legend.title=element_blank()) +
labs(x="Concentration mg/mL") + labs(y= "Absorbances") +
theme(text=element_text(size=10, family="Arial"))

```



A.1.3 Statistical comparison of seminal concentration across species

```
#HumanSEM
Human_concentrations <- c(11.97,12.24,5.77,8.93,8.28)
Mean_Human <- mean(Human_concentrations)
Human_SD <- (sd(Human_concentrations))
Human_SE <- Human_SD/sqrt(length(Human_concentrations))

#ChimpSEM
Chimp_concentrations <- c(18.05,14.51,13.13,21.95)
Mean_Chimp <- mean(Chimp_concentrations)
Chimp_SD <- (sd(Chimp_concentrations))
Chimp_SE <- Chimp_SD/sqrt(length(Chimp_concentrations))

#GorillaSEM
Gorilla_concentrations <-c(11.53,12.62,18.91,13.71)
Mean_Gorilla <- mean(Gorilla_concentrations)
Gorilla_SD <- (sd(Gorilla_concentrations))
Gorilla_SE <- Gorilla_SD/sqrt(length(Gorilla_concentrations))

SemenMeans_DF <- data.frame("Species"= c("Human", "Chimp", "Gorilla"), "Means
" =c(Mean_Human, Mean_Chimp, Mean_Gorilla), "SEs"= c(Human_SE, Chimp_SE, Gori
lla_SE))
SemenMeans_DF

##   Species  Means      SEs
## 1   Human  9.4380  1.210692
## 2   Chimp 16.9100  1.973778
## 3 Gorilla 14.1925  1.634250

#ANOVA
Semenanova <- aov(seSampleconcs20 ~seSpecies, data= Semen_concs_DF)
summary(Semenanova)

##           Df Sum Sq Mean Sq F value Pr(>F)
## seSpecies    2  129.8   64.90   5.998 0.0194 *
## Residuals   10  108.2   10.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

t.test(Chimp_concentrations, Gorilla_concentrations)

##
## Welch Two Sample t-test
##
## data:  Chimp_concentrations and Gorilla_concentrations
## t = 1.0605, df = 5.7982, p-value = 0.3311
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.606071  9.041071
## sample estimates:
```



```

## mean of x mean of y
## 16.9100 14.1925

t.test(Human_concentrations, Gorilla_concentrations)

##
## Welch Two Sample t-test
##
## data: Human_concentrations and Gorilla_concentrations
## t = -2.3377, df = 5.8704, p-value = 0.05896
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -9.7579078 0.2489078
## sample estimates:
## mean of x mean of y
## 9.4380 14.1925

t.test(Chimp_concentrations, Human_concentrations)

##
## Welch Two Sample t-test
##
## data: Chimp_concentrations and Human_concentrations
## t = 3.2269, df = 5.1368, p-value = 0.0224
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 1.567154 13.376846
## sample estimates:
## mean of x mean of y
## 16.910 9.438

```

A.1.4 Protease Inhibitor Bradford Assay

```

#STANDARDS
piStd_conc <- c(0.125, 0.25, 0.5, 0.75, 1, 1.5, 2)
piStd_Abs <- c(0.198, 0.33, 0.607, 0.833, 1.061, 1.279, 1.416)
piStdDF <- data.frame(piStd_conc, piStd_Abs)
piStd_reg <- lm(piStd_Abs ~ piStd_conc)
summary(piStd_reg)

##
## Call:
## lm(formula = piStd_Abs ~ piStd_conc)
##
## Residuals:
##      1      2      3      4      5      6      7
## -0.12394 -0.07457 0.03717 0.09791 0.16066 0.04814 -0.14538
##
## Coefficients:

```

```

##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.23931    0.08238   2.905 0.033598 *
## piStd_conc  0.66103    0.07639   8.653 0.000341 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1274 on 5 degrees of freedom
## Multiple R-squared:  0.9374, Adjusted R-squared:  0.9249
## F-statistic: 74.88 on 1 and 5 DF,  p-value: 0.0003406

#SAMPLES
piTemperatures <- c("4C", "23C", "37C", "4C", "23C", "37C", "4C", "23C", "37C")
piSamplenames <- c("H5", "H5", "H5", "CPI", "CPI", "CPI", "SPI", "SPI", "SPI")
piAbsorbances <- c(0.516, 0.513, 0.538, 0.56, 0.609, 0.622, 0.528, 0.529, 0.622)
conc.calc <- function(Slope, Intercept, Abs, DF)
  {((Abs-Intercept)/(Slope))*DF}
piSampleconcs <- conc.calc(0.66103, 0.23931, piAbsorbances, 1)
piSampleconcs20 <- conc.calc(piStd_reg$coefficients[2], piStd_reg$coefficients[1], piAbsorbances, 20)
ProteaseDF <- data.frame(piTemperatures, piSamplenames, piAbsorbances, piSampleconcs, piSampleconcs20)

```

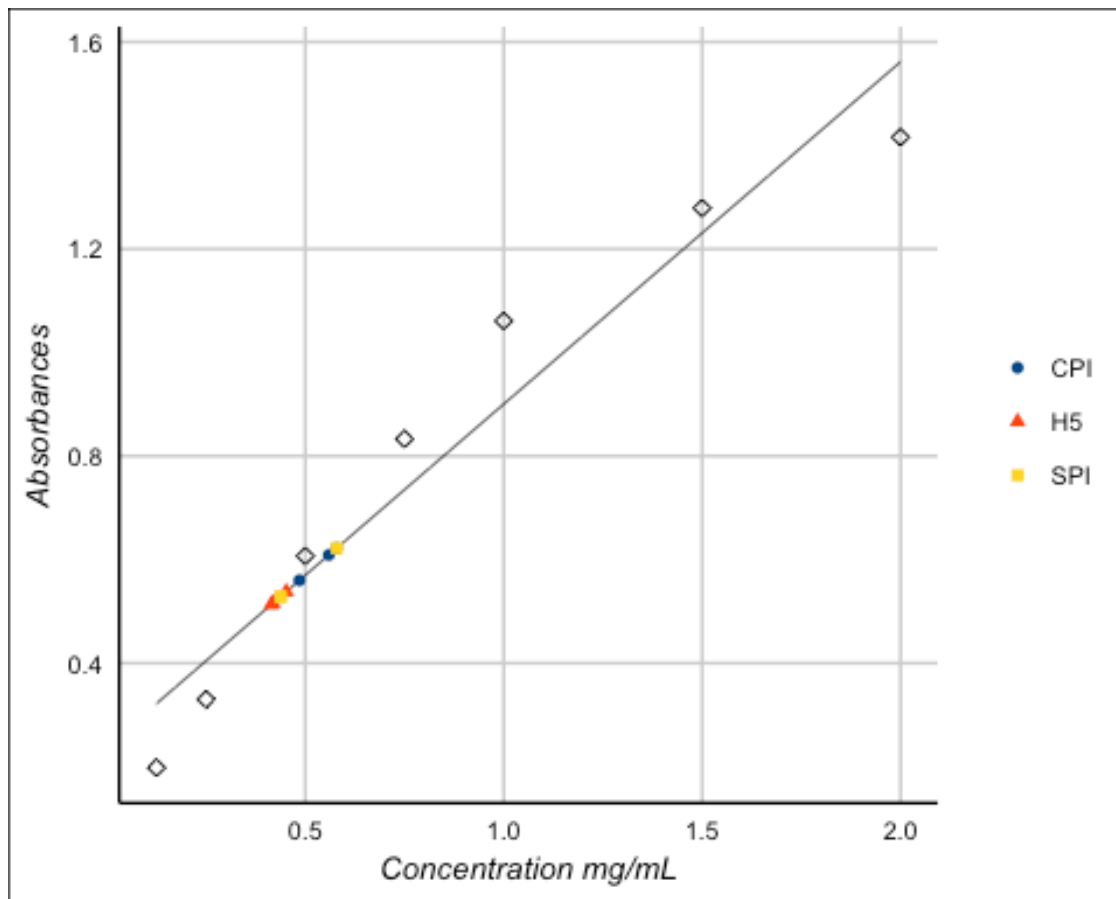
Graph in ggplots

```

library(ggplot2)
library(ggthemes)

ggplot(data= piStdDF,
       aes(x= piStd_conc,
          y= piStd_Abs)) +
  geom_point(shape=5) +
  geom_smooth(method = lm,
             se = FALSE, size = 0.25, color = "black") +
  geom_point(data = ProteaseDF,
            aes(x = piSampleconcs,
               y = piAbsorbances,
               color = piSamplenames,
               shape = piSamplenames)) +
  theme_gdocs() + scale_color_calc() +
  theme(legend.title=element_blank()) +
  labs(x="Concentration mg/mL") + labs(y= "Absorbances") +
  theme(text=element_text(size=10, family="Arial"))

```



```
Semenanova <- aov(piSampleconcs20 ~piSamplenames, data= ProteaseDF)
summary(Semenanova)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## piSamplenames  2  7.655   3.828   3.008  0.124
## Residuals     6  7.634   1.272
```

A.2 Protein sequence alignments

Alignment 1: Human and chimpanzee clusterin sequences

```
>HumanClus gi|355594753|ref|NP_001822.3| clusterin preproprotein [Homo sapiens]
MMKTLFFFVGLLLTWESGQVLGDQTVSDNELQEMSNQGSKYVNKEIQNAVNGVKQIKTLIEKTNEERKTLLSNLEEAKKKKEDALN
ETRESETKLKEKELPGVCNETMMALWEECKPCKLQTCMKFYARVCRSGSGLVGRQLEEFNLQSSPFYFWMNGDRIDSLLENDRQQTHM
LDVMQDHFSSRASSIIDELFQDRFFFTREPQDTYHYLPFSLPHRRPHFFFPKSRIVRSLMPFSPYEPLNFHAMFQPFLEMIHEAQQAM
DIHFHSPAFQHPPTFEFIREGDDDRTVCREIRHNSTGCLRMKDQCDKREILSVDCSTNNPSQAKLRRELDLQVAERLTRKYNEL
LKSQWKMLNTSSLEQLNEQFNWVSRLANLTQGEDQYYLRVTTVASHTSDSDVPSGVTEVVVKLFDSDPITVTVPVEVSRKNPKF
METVAEKALQEYRKKHREE
```

```
>ChimpClus gi|410340483|gb|JAA39188.1| clusterin [Pan troglodytes]
MMKTLFFFVGLLLTWESGQVLGDQTVSDNELQEMSDQGSKYVNKEIQNAVNGVKQIKTLIEKTNEERKTLLSNLEEAKKKKEDALN
ETRESETKLKEKELPGVCNETMMALWEECKPCKLQTCMKFYARVCRSGSGLVGRQLEEFNLQSSPFYFWMNGDRIDSLLENDRQQTHM
LDVMQDHFSSRASSIMDELQDRFFAREPQDTYHYLPFSLPHRRPHFFFPKSRIVRSLMPFSPYEPLNFHAMFQPFLEMIHEAQQAM
DIHFHSPAFQHPPTFEFIREGDDDRTVCREIRHNSTGCLRMKDQCDKREILSVDCSASNPSQAQLRRELDLQVAEKLTRKYNEL
LKSQWKMLNTSSLEQLNEQFNWVSRLANLTQGEDQYYLRVTTVASHTSDSDIPSGVTEVVVKLFDSDPITVTVPVEVSRKNPKF
METVAEKALQEYRKKHREE
```

```
HumanClus   MMKTLFFFVGLLLTWESGQVLGDQTVSDNELQEMSNQGSKYVNKEIQNAVNGVKQIKTLI
ChimpClus   MMKTLFFFVGLLLTWESGQVLGDQTVSDNELQEMSDQGSKYVNKEIQNAVNGVKQIKTLI
*****:*****

HumanClus   EKTNEERKTLLSNLEEAKKKKEDALNETRESETKLKEKELPGVCNETMMALWEECKPCKLQTC
ChimpClus   EKTNEERKTLLSNLEEAKKKKEDALNETRESETKLKEKELPGVCNETMMALWEECKPCKLQTC
*****:*****

HumanClus   CMKFYARVCRSGSGLVGRQLEEFNLQSSPFYFWMNGDRIDSLLENDRQQTHMLDVMQDHF
ChimpClus   CMKFYARVCRSGSGLVGRQLEEFNLQSSPFYFWMNGDRIDSLLENDRQQTHMLDVMQDHF
*****:*****

HumanClus   SRASSIIDELFQDRFFFTREPQDTYHYLPFSLPHRRPHFFFPKSRIVRSLMPFSPYEPLNF
ChimpClus   SRASSIMDELQDRFFAREPQDTYHYLPFSLPHRRPHFFFPKSRIVRSLMPFSPYEPLNF
*****:*****

HumanClus   HAMFQPFLEMIHEAQQAMDIHFHSPAFQHPPTFEFIREGDDDRTVCREIRHNSTGCLRMKD
ChimpClus   HAMFQPFLEMIHEAQQAMDIHFHSPAFQHPPTFEFIREGDDDRTVCREIRHNSTGCLRMKD
*****:*****

HumanClus   QCDKREILSVDCSTNNPSQAKLRRELDLQVAERLTRKYNELLSYQWKMLNTSSLE
ChimpClus   QCDKREILSVDCSASNPSQAQLRRELDLQVAEKLTRKYNELLSYQWKMLNTSSLE
*****:*****

HumanClus   QLNEQFNWVSRLANLTQGEDQYYLRVTTVASHTSDSDVPSGVTEVVVKLFDSDPITVTVP
ChimpClus   QLNEQFNWVSRLANLTQGEDQYYLRVTTVASHTSDSDIPSGVTEVVVKLFDSDPITVTVP
*****:*****

HumanClus   VEVSRKNPKFMETVAEKALQEYRKKHREE
ChimpClus   VEVSRKNPKFMETVAEKALQEYRKKHREE
*****
```

::: indicates amino acid differences between human and chimpanzee clusterin which may affect antibody detection

*** indicates where the ab104652 antibody binds

Alignment 2: Human and chimpanzee albumin sequences

```

>HumanALB gi|4502027|ref|NP_000468.1| serum albumin preproprotein [Homo sapiens]
MKWVTFISLLFLFSSAYSRGVFRDAHKSEVAHRFKDLGEENFKALVLI AFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENC
DKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNP NLPRLV RPEVDVMCTAFHDNEETFLKKYLYEIARRHPY
FYAPELLFFAKRYKAAFTTECCQAADKAAACLLPKLDEL RDEGKASSAKQRLK CASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKL
VTDLTKVHTECCHGD LLECADDRADLAKYICENQDSISSK LKECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYA
EAKDVFLGMFLY EYARRHPDYSV LLLRLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNC ELFQGEYKFQNA
LLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCTESLVNRRPCFSALE
VDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCKADDKETCFAEEGKKLVAA
SQAALGL

>ChimpALB gi|332819547|ref|XP_517233.3| PREDICTED: serum albumin [Pan troglodytes]
MKWVTFISLLFLFSSAYSRGVFRDAHKSEVAHRFKDLGEENFKALVLI AFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENC
DKSLHTLFGDKLCTVATLREKYGEMADCCAKQEPERNECFLQHKDDNP NLPRLV RPEVDVMCTAFHDNEGTFLKKYLYEVARRHPY
FYAPELLFFAERYKAAFTTECCQAADKAAACLLPKLDEL RDEGKASSAKQRLK CASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKL
VTDLTKVHTECCHGD LLECADDRADLAKYICENQDSISSK LKECCEKPLLEKSHCLAEVENDEMPADLPSLAADFVESKEVCKNYA
EAKDVFLGMFLY EYARRHPDYSV LLLRLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNC ELFQGEYKFQNA
LLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCTESLVNRRPCFSALE
VDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCKADDKETCFAEEGKKLVAA
SQAALGL

HumanALB      MKWVTFISLLFLFSSAYSRGVFRDAHKSEVAHRFKDLGEENFKALVLI AFAQYLQQCPF
ChimpALB      MKWVTFISLLFLFSSAYSRGVFRDAHKSEVAHRFKDLGEENFKALVLI AFAQYLQQCPF
*****:*****

HumanALB      EDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEP
ChimpALB      EDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLREKYGEMADCCAKQEP
*****:*****

HumanALB      ERNECFLQHKDDNP NLPRLV RPEVDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLF
ChimpALB      ERNECFLQHKDDNP NLPRLV RPEVDVMCTAFHDNEGTFLKKYLYEVARRHPYFYAPELLF
*****:*****

HumanALB      FAKRYKAAFTTECCQAADKAAACLLPKLDEL RDEGKASSAKQRLK CASLQKFGERAFKAWAV
ChimpALB      FAERYKAAFTTECCQAADKAAACLLPKLDEL RDEGKASSAKQRLK CASLQKFGERAFKAWAV
*:*****

HumanALB      ARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGD LLECADDRADLAKYICENQDSISSK LK
ChimpALB      ARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGD LLECADDRADLAKYICENQDSISSK LK
*****

HumanALB      ECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLY EYAR
ChimpALB      ECCEKPLLEKSHCLAEVENDEMPADLPSLAADFVESKEVCKNYAEAKDVFLGMFLY EYAR
*****:*****

HumanALB      RHPDYSV LLLRLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNC ELF
ChimpALB      RHPDYSV LLLRLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNC ELF
*****

HumanALB      QLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCKHPEAKRMPCAEDYLSV
ChimpALB      QLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCKHPEAKRMPCAEDYLSV
*****

HumanALB      LNQLCVLHEKTPVSDRVTKCTESLVNRRPCFSALEVDETYVPKEFNAETFTFHADICTL
ChimpALB      LNQLCVLHEKTPVSDRVTKCTESLVNRRPCFSALEVDETYVPKEFNAETFTFHADICTL
*****

HumanALB      SEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCKADDKETCFAEEGKKLV
ChimpALB      SEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCKADDKETCFAEEGKKLV
*****

HumanALB      AASQAALGL
ChimpALB      AASQAALGL
*****

```

::: indicates amino acid differences between human and chimpanzee albumin which may affect antibody detection

Alignment 3: Human and bovine albumin sequences

```
>BSA sp|P02769|ALBU_BOVIN Serum albumin OS=Bos taurus
MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQCPFDEHVKLVNELTEFAKTCVADESHAGC
EKSLHTLFGDELCKVASLRETYGDMADCCEKQEPERNECFLSHKDDSPDLPKLKPDPNTLCDEFKADEKKFWGKLYE IARRHPYF
YAPELLYANKYNGVFQEQCAEDKGACLLPKIETMREKVLASSARQLRCASIQKFGERALKAWSVARLSQKFPKAEFVEVTKLV
TDLTKVHKECCHGDLLECADDRADLAKYICDNQDTISSKLKECCDKPILLEKSHCIAEVEKDAIPENLPPLTADFAEDKDVCKNYQE
AKDAFLGSFLEYSRRHPEYAVSVLLRLAKEYEATLECCAKDDPHACYSTVFDKLLKHLVDEPQNLIKQNCQDFEKLGEYGFQNALI
VRYTRKVPQVSTPTLVEVSRSLGKVGTRCCTKPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTCCTESLVNRRPCFSALTPD
ETYVPKAFDEKLFTHADICTLPDTEKQIKKQATALVELLKHKPKATEEQKLTVMENFVAFVDKCCAADDKEACFAVEGPKLVVSTQ
TALA
>HSA sp|P02768|ALBU_HUMAN Serum albumin OS=Homo sapiens
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEEHFKALVLI AFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENC
DKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNL PRLVLRPEVDVMCTAFHDNEETFLKLYE IARRHPY
FYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELDRDEGKASSAKQRLKCASIQKFGERAFKAWAVARLSQRFKAEFAEVSKL
VTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYA
EAKDVFLGMFLYFYARRHPDYSVLLLRLLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNCLEFQELGEYKFNQA
LLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNRRPCFSALE
VDETYVPKEFNAETFTFHADICTLSEKERQIKKQATALVELLVKHKPKATKEQLKAVMDDFAAFVEKCKADDKETCFAEEGKLVAA
SQAALGL
```

```
CLUSTAL O(1.2.1) multiple sequence alignment
BSA      MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQCPF
HSA      MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEEHFKALVLI AFAQYLQQCPF
*****:*****:****:*****.*.******:*****
BSA      DEHVKLVNELTEFAKTCVADESHAGCEKSLHTLFGDELCKVASLRETYGDMADCCEKQEP
HSA      EDHVKLVNEVTEFAKTCVADESAENC DKSLHTLFGDKLCTVATLRETYGEMADCCAKQEP
.:*****:*****.*.******:*.**.******:***** **
BSA      ERNECFLSHKDDSPDLPKL-KPDPNTLCDEFKADEKKFWGKLYE IARRHPYFYAPELLY
HSA      ERNECFLQHKDDNPNL PRLVLRPEVDVMCTAFHDNEETFLKLYE IARRHPYFYAPELLF
*****.***.*:***:* :*: :.* * : :*. * *****:
BSA      YANKYNGVFQEQCAEDKGACLLPKIETMREKVLASSARQLRCASIQKFGERALKAWSV
HSA      FAKRYKAAFTECCQAADKAACLLPKLDELDRDEGKASSAKQRLKCASIQKFGERAFKAWAV
:*.:.*. * ** **.******: :*: : *****:***:***:***:***:***:
BSA      ARLSQKFPKAEFVEVTKLVTDLTKVHKECCHGDLLECADDRADLAKYICDNQDTISSKLK
HSA      ARLSQRFKAEFAEVSKLVTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLK
*****:*****.*.*:*****.*.******:*****:***:*****
BSA      ECCDKPILLEKSHCIAEVEKDAIPENLPPLTADFAEDKDVCKNYQEAKDAFLGSFL-EYSR
HSA      ECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYFYAR
***:*****:***:* :* :** *:* **.****** *****.*** ** **:*
BSA      RHPEYAVSVLLRLAKEYEATLECCAKDDPHACYSTVFDKLLKHLVDEPQNLIKQNCQDFE
HSA      RHPDYSVLLLRLLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNCLEF
***.*:* :*** **:***.*** ** **.***:* :* **:******: **
BSA      KLGEYGFQNALIVRYTRKVPQVSTPTLVEVSRSLGKVGTRCCTKPESERMPCTEDYLSLI
HSA      QLGEYKFNQALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCKHPEAKRMPCAEDYLSVV
:*** *****:***:*****.*.*****:*.**.***:***:***:***:*:
BSA      LNRLCVLHEKTPVSEKVTCCTESLVNRRPCFSALTPDETYVPKAFDEKLFTHADICTL
HSA      LNQLCVLHEKTPVSDRVTKCCTESLVNRRPCFSALEVDETYVPKEFNAETFTFHADICTL
**:******:*****.****** ***** * : *****
BSA      PDTEKQIKKQATALVELLKHKPKATEEQKLTVMENFVAFVDKCCAADDKEACFAVEGPKLV
HSA      SEKERQIKKQATALVELLVKHKPKATKEQLKAVMDDFAAFVEKCKADDKETCFAEEGKLV
.*:***.******:*****.***:***:***:***.***.***.***.***.***.***
BSA      VSTQTALA-
HSA      AASQAALGL
.:*:**.
```

*** indicates trypsin digested fragments that are the same between human and bovine serum albumin

A.3 Outline of LC/MS-MS data sorting, labeling, and normalization

Please note that the initial data processing and normalization was done by Michael I. Jensen-Seaman. Below is his outline on the files and Pearl scripts he used.

Data analysis of human, chimpanzee, and gorilla shotgun proteomics.

Experiments performed in February 2015. All analyses done on data set

“All_Data_JThomas_02112016”, located in

“Projects/Proteomics/HCG_Shotgun_2015/All_Data_JThomas_02110216/”.

This “re-run” is using cutoff score of 11, and “parent charge” up to 7.

A.3.1 Annotation and Normalization:

Initial annotation is by matching peptides from each species to all species with SpectrumMill software.

Results from each run (three runs per individual) and each individual (three individuals per species) and each “cross-matching annotation” are a separate file. Data files are in folder “Projects/Proteomics/HCG_Shotgun_2015/All_Data_JThomas_02110216/”, with the names “Chimp_SMReRun”, “Gorilla_SMReRun”, and “Human_SMReRun”. Within each folder there are three files (e.g., Ch_Ch, Ch_Go, Ch_Hu); these refer to the species that was run, and the species against which it was annotated. Within each of these folders are four folders, with “Pep” and “Prot” for peptide and protein data, and “T” and “NT” for truncated and not truncated by score cutoff of 11.

We are using “T” data, and “Pep” data.

Step 1: Concatenate all peptide files for each species and cross-match (e.g., combining nine runs of self-match into one file). Now have three files for each species (e.g., C_H_T_all.ssv is all nine runs of chimp annotated against human).

Step 2: Concatenate three files into one:

(e.g., `cat C_C_T_all.ssv C_G_T_all.ssv C_H_T_all.ssv > C_T_all.ssv`)

Step 3: Add gene name and length from UniProt (accessed February 2016), using perl script

“addgene_peptide.pl”: e.g., `perl addgene_peptide.pl C_T_all.ssv`

`Prot_Pan_UP000002277.tab`)

Resulting files are *.outnew (e.g., C_T_all.ssv.outnew).

Step 4: Prioritize annotation as follows:

1. If match to self-species, take it.
2. Else, if match to human, take it.
3. Else, if match to third species, take it.

This is done using prioritize.pl, where command line defines priority order

e.g., `perl prioritize.pl C_T_all.ssv.outnew PANTR HUMAN GORGO`

Resulting files are *.prior (e.g., C_T_all.ssv.outnew.prior)

Results from annotation:

Species	Total Pep	Total named	Human annot	Chimp annot	Gorilla annot
Human	3469	3455	3229	155	71
Chimp	3275	3256	208	2998	50
Gorilla	991	967	200	67	724

Step 5: Filter by confidence in hits:

To be considered “high-confidence ID”, protein must:

- Be found in two or more runs for that species
- Be found as two or more unique peptide

Results from filtering:

Species	Peptides/Protein	Protein Found in >1 run?	Protein ID'd from >1 peptide?
Human, peptide	3469	2861	
Human, protein	475	140	71
Chimp, peptide	3275	2849	
Chimp, protein	332	111	64
Gorilla, peptide	991	359	
Gorilla, protein	503	81	36

Note: This reduced the number of proteins a lot.

These proteins are now called “high confidence” proteins. They are found in *.uniq file (e.g., “C_T_HiConf.uniq”, in folder

Projects/Proteomics/HCG_Shotgun_2015/All_Data_JThomas_02110216/Annotation/Final_Annotations/)

After adding Amanda’s manual curation of the semenogelin annotation, I have the three files:

C_T_all_anno_MJS_02222016.xlsx (“keep” tab of spreadsheet)

G_T_all_anno_MJS_02222016.xlsx (“keep” tab of spreadsheet)

H_T_all_anno_MJS_02222016.xlsx (“keep” tab of spreadsheet)

Step 6: Normalize data (without accounting for length):

Re-splitting into separate files for each run (e.g., C12.T.spec.anno), where the first letter is the species, the first number is the individual, and the second number is the run). Using “normalize.pl” as follows:

```
perl normalize.pl *spec.anno
```

output is *spec.anno.norm

This divides the spectral count for each protein by the total spectral count of each run.

These are now placed into a spreadsheet “HCG_compiled.normal_new.xlsx”.

There are 131 unique proteins among human, chimp, and gorilla.

Step 7: Repeat above but accounting for length.

For this, combine number of spectra across the three technical replicates, to get a sum of spectra for each biological replicate. To produce a normalized length-controlled NSAF for each protein k for each biological replicate:

$$NSAF_k = \frac{(\text{number of spectra/length})_k}{\sum_{i=1}^N (\text{number of spectra/length})_i}$$

note: lengths for all species were taken from the human protein database (because annotation for other species is incomplete and unreviewed). The exception is SEMG1 and SEMG2.

Results are now given as “NSAF_length” in three spreadsheets:
C123.spec.anno.length.xlsx, G123.spec.anno.length.xlsx, H123.spec.anno.length.xlsx

And combined into HCG.nsaf.length

Step 8: Statistics.

1. Simple t-test where each individual is a value.
2. Count two-fold or greater as ‘interesting’.
3. PLGEM analysis, in R.

A.4 Script in R for PLGEM statistics

A.4.1 Load libraries and set working directory

```
library(plgem)

##
## Welcome to plgem version 1.46.0

library(Biobase)

## Loading required package: BiocGenerics
## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##   clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##   clusterExport, clusterMap, parApply, parCapply, parLapply,
##   parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##   IQR, mad, xtabs

## The following objects are masked from 'package:base':
##
##   anyDuplicated, append, as.data.frame, cbind, colnames,
##   do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##   grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
##   match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##   Position, rank, rbind, Reduce, rownames, sapply, setdiff,
##   sort, table, tapply, union, unique, unsplit, which, which.max,
##   which.min

## Welcome to Bioconductor
##
##   Vignettes contain introductory material; view with
##   'browseVignettes()'. To cite Bioconductor, see
##   'citation("Biobase)", and for packages 'citation("pkgname)".

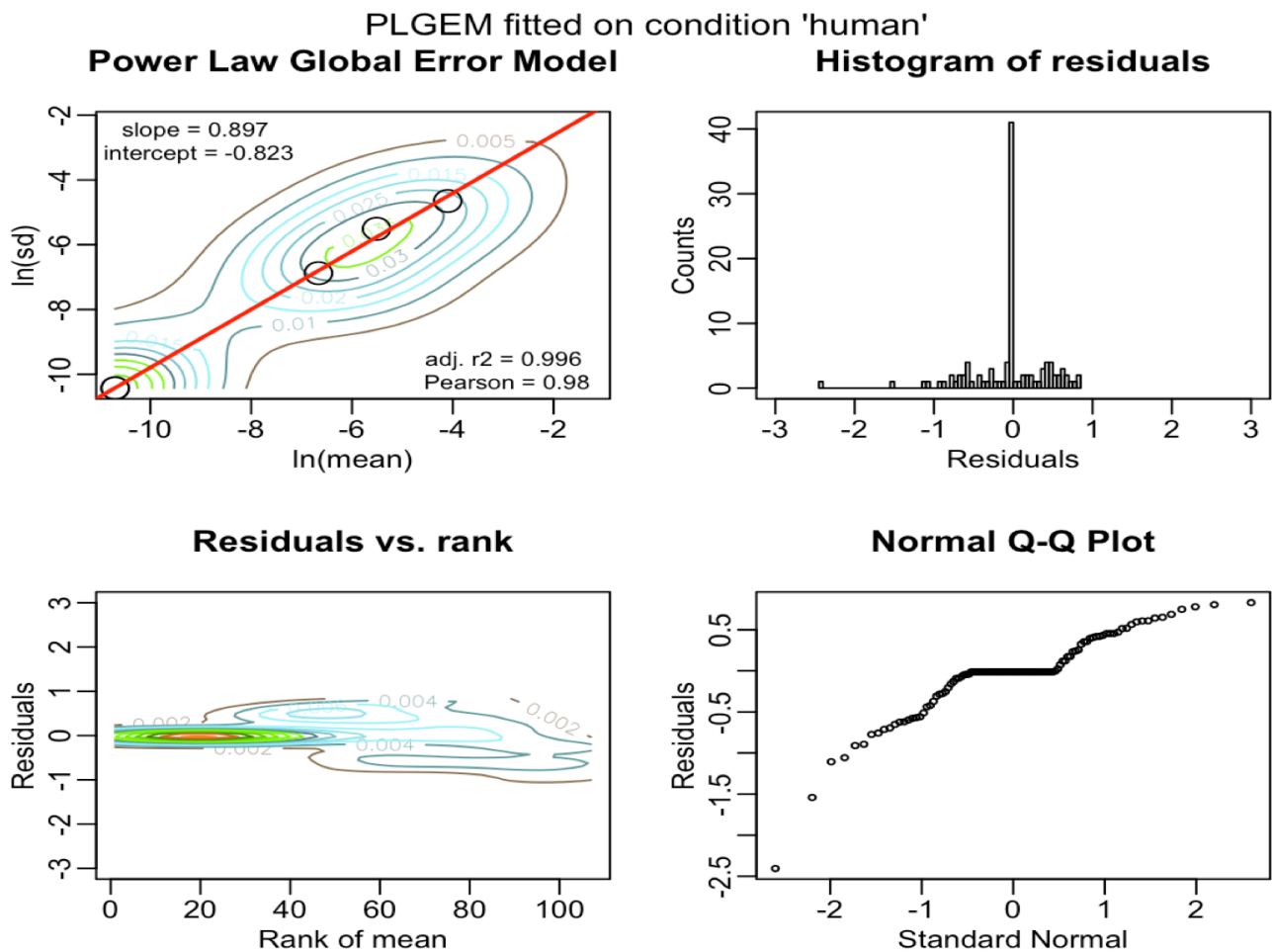
setwd("~/Desktop/PLGEM") #all files listed in this script must be in the same
folder without any of the output 'result' files.
```

A.4.2 Length adjusted pairwise PLGEM of human and chimp

```
esetLCH <- readExpressionSet("length.HC.txt", "phenoDataFile_CH_L.txt")

NSAfitLCH <- plgem.fit(data=esetLCH, covariate = 1, fitCondition = 'human',
p=5, q=0.5, plot.file = FALSE, fittingEval = TRUE, verbose = TRUE)

## Fitting PLGEM...
## samples extracted for fitting:
##   species
## H1_NSAF  human
## H2_NSAF  human
## H3_NSAF  human
## replacing 36 non-positive means with smallest positive mean...
## replacing 36 zero standard deviations with smallest non-zero standard de
viation...
## determining modelling points...
## fitting data and modelling points...
```



```

## done with fitting PLGEM.

NSAFobsSTNLCH <- plgem.obsStn(data=esetLCH, covariate=1, baselineCondition =
1, plgemFit = NSAFfitLCH, verbose = TRUE)

## calculating observed PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## working on baseline human ...
## H1_NSAF H2_NSAF H3_NSAF
## working on condition chimp ...
## C1_NSAF C2_NSAF C3_NSAF
## done with calculating PLGEM-STN statistics.

set.seed(123)
NSAFresampledStnLCH <- plgem.resampledStn(data=esetLCH, plgemFit = NSAFfitLCH
, iterations = 10000, verbose = TRUE)

## calculating resampled PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## baseline samples:
## H1_NSAF H2_NSAF H3_NSAF
## resampling on samples:
## H1_NSAF H2_NSAF H3_NSAF
## Using 10000 iterations...
## working on cases with 3 replicates...
##      Iterations: 100 200 300 400 500 600 700 800 900 1000
## 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000
## 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000
## 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000
## 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
## 5100 5200 5300 5400 5500 5600 5700 5800 5900 6000
## 6100 6200 6300 6400 6500 6600 6700 6800 6900 7000
## 7100 7200 7300 7400 7500 7600 7700 7800 7900 8000
## 8100 8200 8300 8400 8500 8600 8700 8800 8900 9000
## 9100 9200 9300 9400 9500 9600 9700 9800 9900 10000
##
## done with calculating resampled PLGEM-STN statistics.

NSAFpValuesLCH <- plgem.pValue(observedStn = NSAFobsSTNLCH, plgemResampledStn
= NSAFresampledStnLCH, verbose = TRUE)

## calculating PLGEM p-values... done.

NSAFdegListLCH <- plgem.deg(observedStn = NSAFobsSTNLCH, plgemPval = NSAFpVal
uesLCH, delta= 0.01, verbose= TRUE)

## selecting significant DEG:found 1 condition(s) compared to the baseline.
## Delta = 0.01
## Condition = chimp_vs_human
## delta: 0.01 condition: chimp_vs_human found 28 DEG
## done with selecting significant DEG.

```



```

## done with fitting PLGEM.

NSAFobsSTNLHG <- plgem.obsStn(data=esetLHG, covariate=1, baselineCondition =
1, plgemFit = NSAFfitLHG, verbose = TRUE)

## calculating observed PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## working on baseline human ...
## H1_NSAF H2_NSAF H3_NSAF
## working on condition gorilla ...
## G1_NSAF G2_NSAF G3_NSAF
## done with calculating PLGEM-STN statistics.

set.seed(123)
NSAFresampledStnLHG <- plgem.resampledStn(data=esetLHG, plgemFit = NSAFfitLHG
, iterations = 10000, verbose = TRUE)

## calculating resampled PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## baseline samples:
## H1_NSAF H2_NSAF H3_NSAF
## resampling on samples:
## H1_NSAF H2_NSAF H3_NSAF
## Using 10000 iterations...
## working on cases with 3 replicates...
##      Iterations: 100 200 300 400 500 600 700 800 900 1000
## 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000
## 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000
## 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000
## 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
## 5100 5200 5300 5400 5500 5600 5700 5800 5900 6000
## 6100 6200 6300 6400 6500 6600 6700 6800 6900 7000
## 7100 7200 7300 7400 7500 7600 7700 7800 7900 8000
## 8100 8200 8300 8400 8500 8600 8700 8800 8900 9000
## 9100 9200 9300 9400 9500 9600 9700 9800 9900 10000
##
## done with calculating resampled PLGEM-STN statistics.

NSAFpValuesLHG <- plgem.pValue(observedStn = NSAFobsSTNLHG, plgemResampledStn
= NSAFresampledStnLHG, verbose = TRUE)

## calculating PLGEM p-values... done.

NSAFdegListLHG <- plgem.deg(observedStn = NSAFobsSTNLHG, plgemPval = NSAFpVal
uesLHG, delta= 0.01, verbose= TRUE)

## selecting significant DEG:found 1 condition(s) compared to the baseline.
## Delta = 0.01
## Condition = gorilla_vs_human
## delta: 0.01 condition: gorilla_vs_human found 35 DEG
## done with selecting significant DEG.

```

```

plgem.write.summary(NSAFdegListLHG, prefix = "HG_length_PLGEM", verbose = TRUE)
## Writing files...
##   HG_length_PLGEM_fit.csv
##   HG_length_PLGEM_stn_p-value.csv
##   HG_length_PLGEM_gorilla_vs_human_0.01.txt
## ...to folder /Users/Amanda/Desktop/PLGEM

```

A.4.4 Length adjusted pairwise PLGEM of chimp and gorilla

```

esetLCG <- readExpressionSet("length.CG.txt", "phenoDataFile_CG_L.txt")

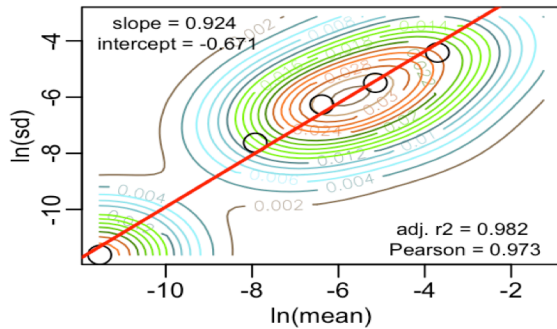
NSAFfitLCG <- plgem.fit(data=esetLCG, covariate = 1, fitCondition = 'chimp',
p=5, q=0.5, plot.file = FALSE, fittingEval = TRUE, verbose = TRUE)

## Fitting PLGEM...
## samples extracted for fitting:
##   species
## C1_NSAF  chimp
## C2_NSAF  chimp
## C3_NSAF  chimp
## replacing 24 non-positive means with smallest positive mean...
## replacing 24 zero standard deviations with smallest non-zero standard deviation...
## determining modelling points...
## fitting data and modelling points...

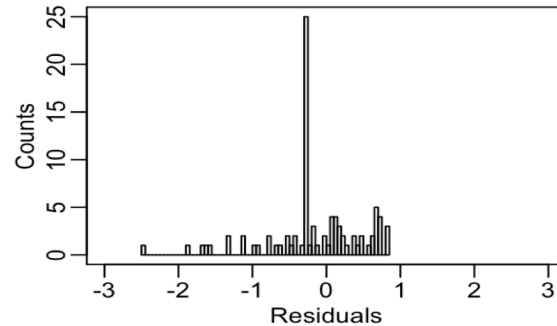
```

PLGEM fitted on condition 'chimp'

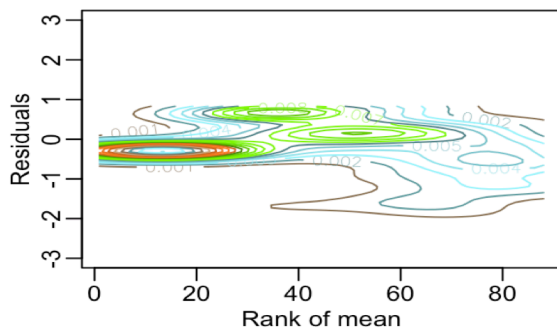
Power Law Global Error Model



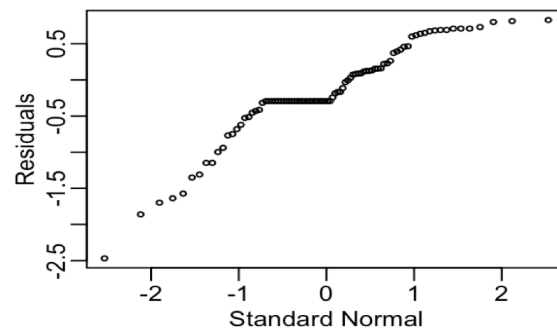
Histogram of residuals



Residuals vs. rank



Normal Q-Q Plot



```
## done with fitting PLGEM.
```

```
NSAFobsStnLCG <- plgem.obsStn(data=esetLCG, covariate=1, baselineCondition =  
1, plgemFit = NSAFfitLCG, verbose = TRUE)
```

```
## calculating observed PLGEM-STN statistics:found 1 condition(s) to compare  
to the baseline.
```

```
## working on baseline chimp ...
```

```
## C1_NSAF C2_NSAF C3_NSAF
```

```
## working on condition gorilla ...
```

```
## G1_NSAF G2_NSAF G3_NSAF
```

```
## done with calculating PLGEM-STN statistics.
```

```
set.seed(123)
```

```
NSAFresampledStnLCG <- plgem.resampledStn(data=esetLCG, plgemFit = NSAFfitLCG  
, iterations = 10000, verbose = TRUE)
```

```
## calculating resampled PLGEM-STN statistics:found 1 condition(s) to compare  
to the baseline.
```

```
## baseline samples:
```

```
## C1_NSAF C2_NSAF C3_NSAF
```

```
## resampling on samples:
```

```
## C1_NSAF C2_NSAF C3_NSAF
```

```
## Using 10000 iterations...
```

```
## working on cases with 3 replicates...
```



```

##      Iterations: 100  200  300  400  500  600  700  800  900  1000
## 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000
## 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000
## 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000
## 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
## 5100 5200 5300 5400 5500 5600 5700 5800 5900 6000
## 6100 6200 6300 6400 6500 6600 6700 6800 6900 7000
## 7100 7200 7300 7400 7500 7600 7700 7800 7900 8000
## 8100 8200 8300 8400 8500 8600 8700 8800 8900 9000
## 9100 9200 9300 9400 9500 9600 9700 9800 9900 10000
##
## done with calculating resampled PLGEM-STN statistics.

NSAFpValuesLCG <- plgem.pValue(observedStn = NSAFobsSTNLCG, plgemResampledStn
 = NSAFresampledStnLCG, verbose = TRUE)

## calculating PLGEM p-values... done.

NSAFdegListLCG <- plgem.deg(observedStn = NSAFobsSTNLCG, plgemPval = NSAFpVal
uesLCG, delta= 0.01, verbose= TRUE)

## selecting significant DEG:found 1 condition(s) compared to the baseline.
## Delta = 0.01
## Condition = gorilla_vs_chimp
## delta: 0.01 condition: gorilla_vs_chimp found 33 DEG
## done with selecting significant DEG.

plgem.write.summary(NSAFdegListLCG, prefix = "CG_length_PLGEM", verbose = TRU
E)

## Writing files...
##      CG_length_PLGEM_fit.csv
##      CG_length_PLGEM_stn_p-value.csv
##      CG_length_PLGEM_gorilla_vs_chimp_0.01.txt
## ...to folder /Users/Amanda/Desktop/PLGEM

```

A.4.5 Raw normalization pairwise PLGEM of human and chimp

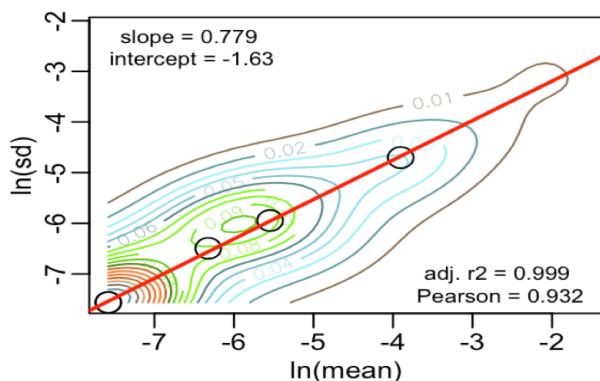
```
esetRCH <- readExpressionSet("NSAF_CH_10042016.txt", "phenoDataFile_CH_R.txt")
)

NSAfitRCH <- plgem.fit(data=esetRCH, covariate = 1, fitCondition = 'human',
p=5, q=0.5, plot.file = FALSE, fittingEval = TRUE, verbose = TRUE)

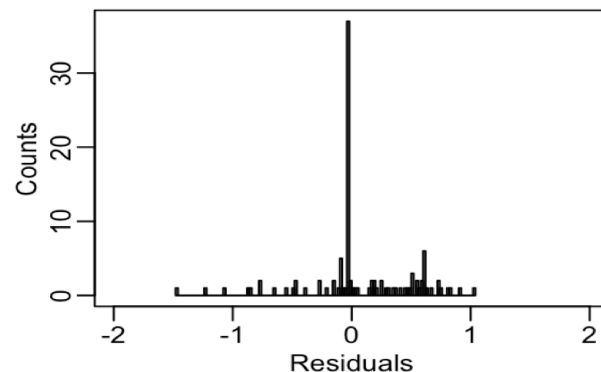
## Fitting PLGEM...
## samples extracted for fitting:
##     species
## H1_ave   human
## H2_ave   human
## H3_ave   human
## replacing 36 non-positive means with smallest positive mean...
## replacing 36 zero standard deviations with smallest non-zero standard deviation...
## determining modelling points...
## fitting data and modelling points...
```

PLGEM fitted on condition 'human'

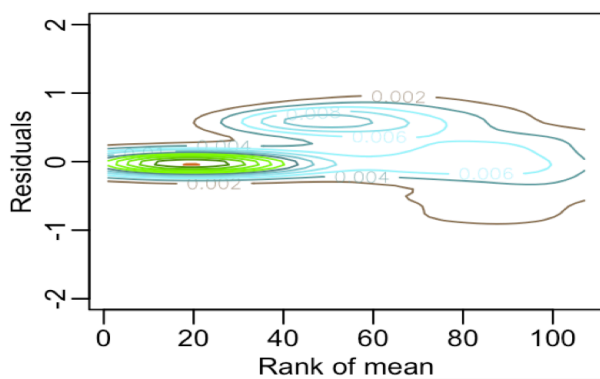
Power Law Global Error Model



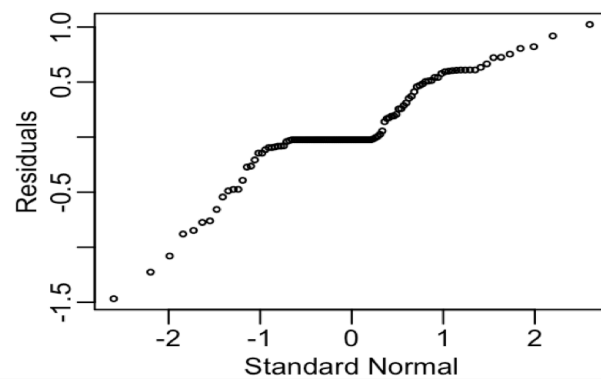
Histogram of residuals



Residuals vs. rank



Normal Q-Q Plot



```

## done with fitting PLGEM.

NSAFobsSTNRCH <- plgem.obsStn(data=esetRCH, covariate=1, baselineCondition =
1, plgemFit = NSAFfitRCH, verbose = TRUE)

## calculating observed PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## working on baseline chimp ...
## C1_ave C2_ave C3_ave
## working on condition human ...
## H1_ave H2_ave H3_ave
## done with calculating PLGEM-STN statistics.

set.seed(123)
NSAFresampledStnRCH <- plgem.resampledStn(data=esetRCH, plgemFit = NSAFfitRCH
, iterations = 10000, verbose = TRUE)

## calculating resampled PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## baseline samples:
## C1_ave C2_ave C3_ave
## resampling on samples:
## H1_ave H2_ave H3_ave
## Using 10000 iterations...
## working on cases with 3 replicates...
##      Iterations: 100 200 300 400 500 600 700 800 900 1000
## 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000
## 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000
## 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000
## 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
## 5100 5200 5300 5400 5500 5600 5700 5800 5900 6000
## 6100 6200 6300 6400 6500 6600 6700 6800 6900 7000
## 7100 7200 7300 7400 7500 7600 7700 7800 7900 8000
## 8100 8200 8300 8400 8500 8600 8700 8800 8900 9000
## 9100 9200 9300 9400 9500 9600 9700 9800 9900 10000
##
## done with calculating resampled PLGEM-STN statistics.

NSAFpValuesRCH <- plgem.pValue(observedStn = NSAFobsSTNRCH, plgemResampledStn
= NSAFresampledStnRCH, verbose = TRUE)

## calculating PLGEM p-values... done.

NSAFdegListRCH <- plgem.deg(observedStn = NSAFobsSTNRCH, plgemPval = NSAFpVal
uesRCH, delta= 0.01, verbose= TRUE)

## selecting significant DEG:found 1 condition(s) compared to the baseline.
## Delta = 0.01
## Condition = human_vs_chimp
## delta: 0.01 condition: human_vs_chimp found 24 DEG
## done with selecting significant DEG.

```

```

plgem.write.summary(NSAFdegListRCH, prefix = "CH_raw_PLGEM", verbose = TRUE)

## Writing files...
##   CH_raw_PLGEM_fit.csv
##   CH_raw_PLGEM_stn_p-value.csv
##   CH_raw_PLGEM_human_vs_chimp_0.01.txt
## ...to folder /Users/Amanda/Desktop/PLGEM

```

A.4.6 Raw normalization pairwise PLGEM of human and gorilla

```

esetRHG <- readExpressionSet("NSAF_GH_10042016.txt", "phenoDataFile_GH_R.txt"
)

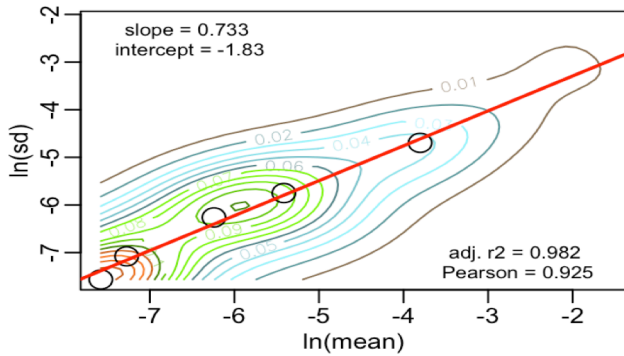
NSAFfitRHG <- plgem.fit(data=esetRHG, covariate = 1, fitCondition = 'human',
p=5, q=0.5, plot.file = FALSE, fittingEval = TRUE, verbose = TRUE)

## Fitting PLGEM...
## samples extracted for fitting:
##   species
## H1_ave   human
## H2_ave   human
## H3_ave   human
## replacing 26 non-positive means with smallest positive mean...
## replacing 26 zero standard deviations with smallest non-zero standard deviation...
## determining modelling points...
## fitting data and modelling points...

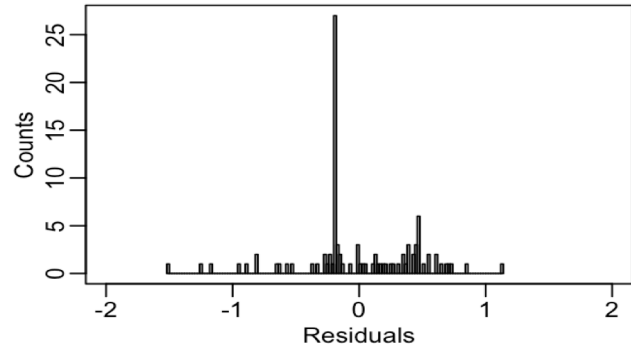
```

PLGEM fitted on condition 'human'

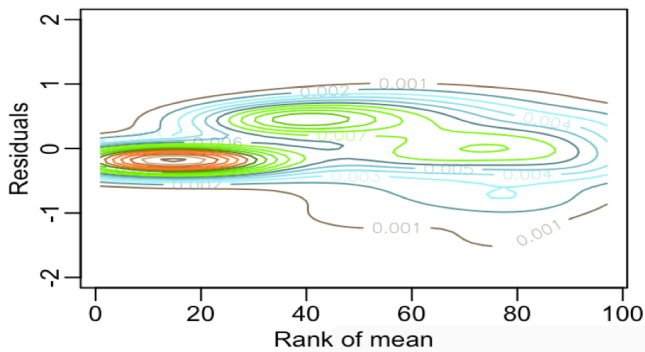
Power Law Global Error Model



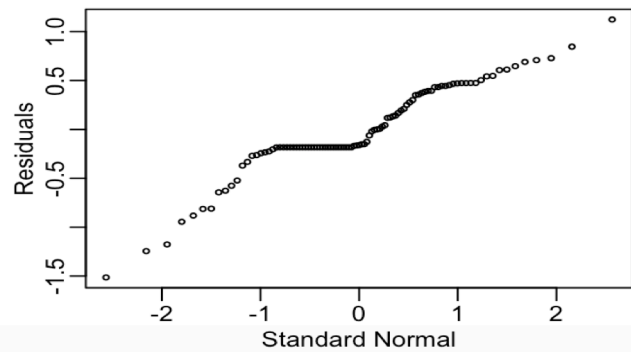
Histogram of residuals



Residuals vs. rank



Normal Q-Q Plot



```
## done with fitting PLGEM.

NSAFobsSTNRHG <- plgem.obsStn(data=esetRHG, covariate=1, baselineCondition =
1, plgemFit = NSAFfitRHG, verbose = TRUE)

## calculating observed PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## working on baseline gorilla ...
## G1_ave G2_ave G3_ave
## working on condition human ...
## H1_ave H2_ave H3_ave
## done with calculating PLGEM-STN statistics.

set.seed(123)
NSAFresampledStnRHG <- plgem.resampledStn(data=esetRHG, plgemFit = NSAFfitRHG
, iterations = 10000, verbose = TRUE)

## calculating resampled PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## baseline samples:
## G1_ave G2_ave G3_ave
## resampling on samples:
## H1_ave H2_ave H3_ave
```

```

## Using 10000 iterations...
## working on cases with 3 replicates...
##      Iterations: 100 200 300 400 500 600 700 800 900 1000
## 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000
## 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000
## 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000
## 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
## 5100 5200 5300 5400 5500 5600 5700 5800 5900 6000
## 6100 6200 6300 6400 6500 6600 6700 6800 6900 7000
## 7100 7200 7300 7400 7500 7600 7700 7800 7900 8000
## 8100 8200 8300 8400 8500 8600 8700 8800 8900 9000
## 9100 9200 9300 9400 9500 9600 9700 9800 9900 10000
##
## done with calculating resampled PLGEM-STN statistics.

NSAFpValuesRHG <- plgem.pValue(observedStn = NSAFobsSTNRHG, plgemResampledStn
 = NSAFresampledStnRHG, verbose = TRUE)

## calculating PLGEM p-values... done.

NSAFdegListRHG <- plgem.deg(observedStn = NSAFobsSTNRHG, plgemPval = NSAFpVal
uesRHG, delta= 0.01, verbose= TRUE)

## selecting significant DEG:found 1 condition(s) compared to the baseline.
## Delta = 0.01
## Condition = human_vs_gorilla
## delta: 0.01 condition: human_vs_gorilla found 40 DEG
## done with selecting significant DEG.

plgem.write.summary(NSAFdegListRHG, prefix = "HG_raw_PLGEM", verbose = TRUE)

## Writing files...
##      HG_raw_PLGEM_fit.csv
##      HG_raw_PLGEM_stn_p-value.csv
##      HG_raw_PLGEM_human_vs_gorilla_0.01.txt
## ...to folder /Users/Amanda/Desktop/PLGEM

```

A.4.7 Raw normalization pairwise PLGEM of chimp and gorilla

```

esetRCG <- readExpressionSet("NSAF_CG_10042016.txt", "phenoDataFile_CG_R.txt"
)

NSAFfitRCG <- plgem.fit(data=esetRCG, covariate = 1, fitCondition = 'chimp',
p=5, q=0.5, plot.file = FALSE, fittingEval = TRUE, verbose = TRUE)

## Fitting PLGEM...
## samples extracted for fitting:
##      species
## C1_ave  chimp
## C2_ave  chimp
## C3_ave  chimp

```

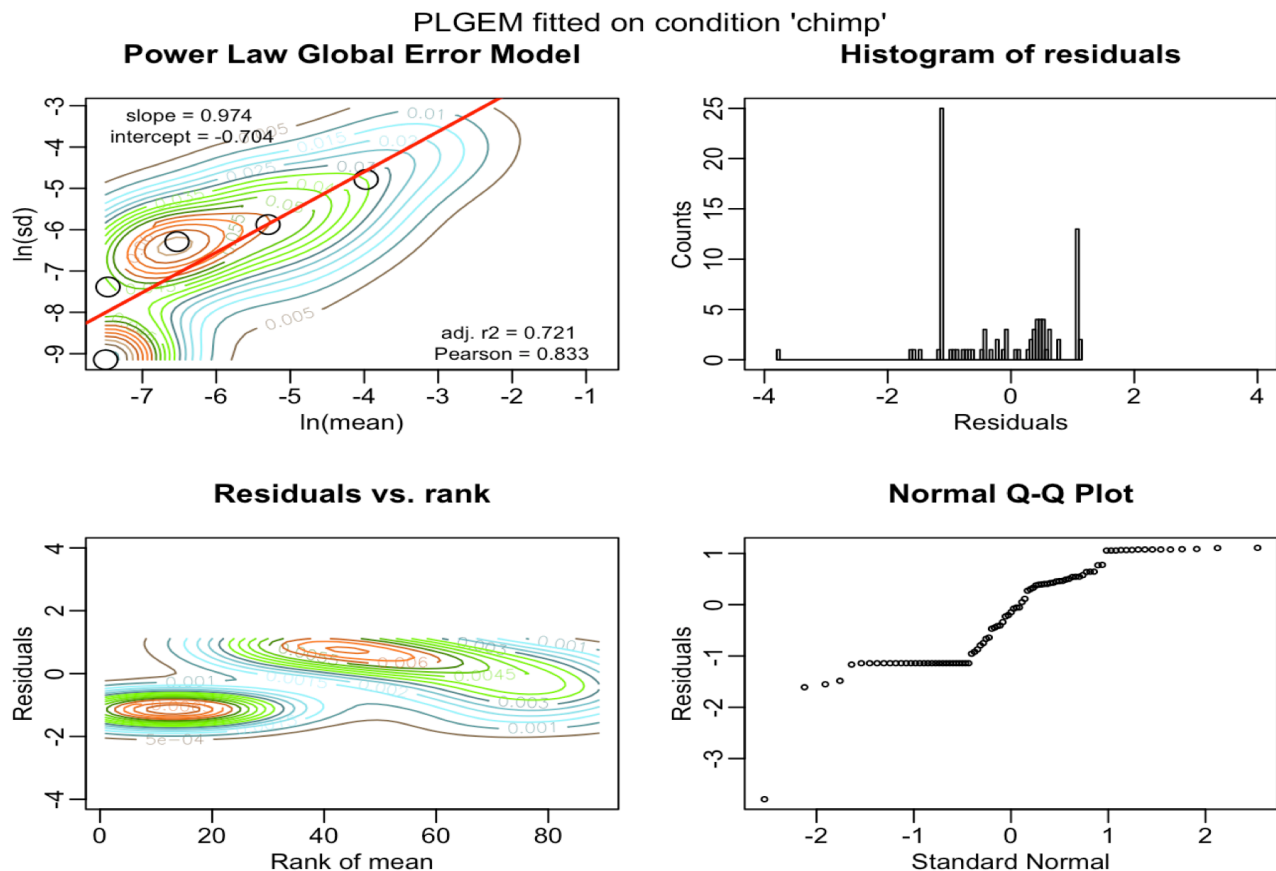
```

## replacing 25 non-positive means with smallest positive mean...
## replacing 25 zero standard deviations with smallest non-zero standard deviation...
## determining modelling points...
## fitting data and modelling points...

## Warning in plgem.fit(data = esetRCG, covariate = 1, fitCondition =
## "chimp", : Adjusted r^2 is lower than 0.95

## Warning in plgem.fit(data = esetRCG, covariate = 1, fitCondition =
## "chimp", : Pearson correlation coefficient is lower than 0.85

```



```

## done with fitting PLGEM.

NSAFobsSTNRCG <- plgem.obsStn(data=esetRCG, covariate=1, baselineCondition =
1, plgemFit = NSAFfitRCG, verbose = TRUE)

## calculating observed PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## working on baseline chimp ...
## C1_ave C2_ave C3_ave
## working on condition gorilla ...
## G1_ave G2_ave G3_ave
## done with calculating PLGEM-STN statistics.

```

```

set.seed(123)
NSAFresampledStnRCG <- plgem.resampledStn(data=esetRCG, plgemFit = NSAFfitRCG
, iterations = 10000, verbose = TRUE)

## calculating resampled PLGEM-STN statistics:found 1 condition(s) to compare
to the baseline.
## baseline samples:
## C1_ave C2_ave C3_ave
## resampling on samples:
## C1_ave C2_ave C3_ave
## Using 10000 iterations...
## working on cases with 3 replicates...
##      Iterations: 100 200 300 400 500 600 700 800 900 1000
## 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000
## 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000
## 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000
## 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
## 5100 5200 5300 5400 5500 5600 5700 5800 5900 6000
## 6100 6200 6300 6400 6500 6600 6700 6800 6900 7000
## 7100 7200 7300 7400 7500 7600 7700 7800 7900 8000
## 8100 8200 8300 8400 8500 8600 8700 8800 8900 9000
## 9100 9200 9300 9400 9500 9600 9700 9800 9900 10000
##
## done with calculating resampled PLGEM-STN statistics.

NSAFpValuesRCG <- plgem.pValue(observedStn = NSAFobsSTNRCG, plgemResampledStn
= NSAFresampledStnRCG, verbose = TRUE)

## calculating PLGEM p-values... done.

NSAFdegListRCG <- plgem.deg(observedStn = NSAFobsSTNRCG, plgemPval = NSAFpVal
uesRCG, delta= 0.01, verbose= TRUE)

## selecting significant DEG:found 1 condition(s) compared to the baseline.
## Delta = 0.01
## Condition = gorilla_vs_chimp
## delta: 0.01 condition: gorilla_vs_chimp found 47 DEG
## done with selecting significant DEG.

plgem.write.summary(NSAFdegListRCG, prefix = "CG_raw_PLGEM", verbose = TRUE)

## Writing files...
##      CG_raw_PLGEM_fit.csv
##      CG_raw_PLGEM_stn_p-value.csv
##      CG_raw_PLGEM_gorilla_vs_chimp_0.01.txt
## ...to folder /Users/Amanda/Desktop/PLGE

```


A.5 Shannon and Simpson diversity index calculations

Shannon and Simpson diversity indexes were hand calculated in excel. Calculation tables are shown only for raw normalization, however the same process was utilized for length adjusted normalized data. A t-test in R determined that the biodiversity statistics were not significantly different between raw and length adjust normalizations. The script and analysis follows the tables.

Table A.1: Shannon diversity calculation with raw normalization

Protein	Human pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Chimp pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Gorilla pi=sample/ sum	Ln(pi)	pi*Ln(pi)
ACE	0.0000			0.0043	-5.4400	-0.0236	0.0000		
ACO2	0.0000			0.0007	-7.2813	-0.0050	0.0000		
ACOT13	0.0000			0.0015	-6.4696	-0.0100	0.0000		
ACPP	0.0412	-3.1895	-0.1314	0.0237	-3.7439	-0.0886	0.0000		
ACR	0.0000			0.0015	-6.4930	-0.0098	0.0000		
ACRBP	0.0000			0.0060	-5.1154	-0.0307	0.0000		
ACTA2	0.0058	-5.1489	-0.0299	0.0000			0.0000		
ACTB	0.0000			0.0068	-4.9839	-0.0341	0.0000		
ALB	0.0566	-2.8714	-0.1626	0.0803	-2.5222	-0.2025	0.2591	-1.3504	-0.3499
AMY2B	0.0000			0.0113	-4.4844	-0.0506	0.0000		
ANPEP	0.0072	-4.9315	-0.0356	0.0012	-6.7649	-0.0078	0.0000		
APOB	0.0000			0.0000			0.0271	-3.6069	-0.0979
ARHGAP31	0.0007	-7.2592	-0.0051	0.0000			0.0000		
ARNT	0.0000			0.0000			0.0182	-4.0086	-0.0728
ASAP2	0.0000			0.0000			0.0058	-5.1499	-0.0299
AZGP1	0.0243	-3.7154	-0.0905	0.0092	-4.6876	-0.0432	0.0000		
B2M	0.0024	-6.0318	-0.0145	0.0011	-6.7957	-0.0076	0.0000		
B4GALT1	0.0000			0.0020	-6.2141	-0.0124	0.0000		
BCL3	0.0000			0.0000			0.0109	-4.5215	-0.0492
BIRC6	0.0008	-7.1208	-0.0058	0.0000			0.0000		

Protein	Human pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Chimp pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Gorilla pi=sample/ sum	Ln(pi)	pi*Ln(pi)
CAP1	0.0019	-6.2644	-0.0119	0.0000			0.0000		
CCDC88C	0.0000			0.0000			0.0614	-2.7902	-0.1713
CEP162	0.0000			0.0000			0.0138	-4.2827	-0.0591
CFAP69	0.0000			0.0000			0.0068	-4.9967	-0.0338
CKB	0.0023	-6.0635	-0.0141	0.0000			0.0000		
CLEC11A	0.0000			0.0000			0.0059	-5.1324	-0.0303
CLIP2	0.0008	-7.1495	-0.0056	0.0000			0.0000		
CLU	0.0236	-3.7450	-0.0885	0.0457	-3.0859	-0.1410	0.0709	-2.6462	-0.1877
CPE	0.0045	-5.4066	-0.0243	0.0000			0.0000		
CPSF1	0.0000			0.0000			0.0074	-4.9053	-0.0363
CRISP1	0.0096	-4.6425	-0.0447	0.0017	-6.3955	-0.0107	0.0000		
CRTAC1	0.0000			0.0176	-4.0375	-0.0712	0.0000		
CST2	0.0009	-7.0091	-0.0063	0.0000			0.0000		
CST3	0.0019	-6.2688	-0.0119	0.0052	-5.2584	-0.0274	0.0000		
CST4	0.0057	-5.1665	-0.0295	0.0000			0.0000		
CTSB	0.0000			0.0000			0.0034	-5.6785	-0.0194
CTSD	0.0032	-5.7346	-0.0185	0.0000			0.0000		
DEFA1	0.0038	-5.5677	-0.0213	0.0000			0.0000		
DHX57	0.0005	-7.5609	-0.0039	0.0000			0.0000		
DNAH7	0.0000			0.0000			0.0042	-5.4744	-0.0230
DPEP3	0.0000			0.0056	-5.1798	-0.0292	0.0000		
DUT	0.0007	-7.2734	-0.0050	0.0000			0.0000		
ECM1	0.0067	-5.0042	-0.0336	0.0097	-4.6349	-0.0450	0.0000		
ENO1	0.0000			0.0008	-7.0866	-0.0059	0.0000		
FAM183B	0.0000			0.0000			0.0036	-5.6312	-0.0202
FAM184B	0.0000			0.0000			0.0104	-4.5661	-0.0475
FAM3D	0.0000			0.0006	-7.4713	-0.0043	0.0000		

Protein	Human pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Chimp pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Gorilla pi=sample/ sum	Ln(pi)	pi*Ln(pi)
FN1	0.0257	-3.6602	-0.0942	0.0525	-2.9474	-0.1547	0.0000		
FOLH1	0.0015	-6.4737	-0.0100	0.0000			0.0000		
GDF15	0.0033	-5.7272	-0.0186	0.0000			0.0000		
GOLGB1	0.0008	-7.1245	-0.0057	0.0000			0.0000		
GPI	0.0000			0.0029	-5.8387	-0.0170	0.0000		
HEXB	0.0033	-5.7142	-0.0189	0.0014	-6.5573	-0.0093	0.0000		
HK1	0.0000			0.0034	-5.6719	-0.0195	0.0000		
HSP90AA1	0.0000			0.0011	-6.7907	-0.0076	0.0000		
HSPA1A	0.0000			0.0014	-6.5987	-0.0090	0.0000		
HSPA1L	0.0023	-6.0936	-0.0138	0.0000			0.0000		
HSPA5	0.0020	-6.2242	-0.0123	0.0000			0.0000		
IDH1	0.0018	-6.3232	-0.0113	0.0000			0.0000		
IGHG1	0.0047	-5.3524	-0.0254	0.0144	-4.2413	-0.0610	0.0000		
IGKC	0.0000			0.0083	-4.7873	-0.0399	0.0000		
KLK3	0.0209	-3.8660	-0.0810	0.0000			0.0000		
KRT1	0.0054	-5.2261	-0.0281	0.0050	-5.2935	-0.0266	0.0374	-3.2848	-0.1230
KRT10	0.0044	-5.4236	-0.0239	0.0012	-6.7592	-0.0078	0.0034	-5.6785	-0.0194
KRT2	0.0078	-4.8523	-0.0379	0.0011	-6.7907	-0.0076	0.0000		
LAMP2	0.0024	-6.0485	-0.0143	0.0000			0.0000		
LCP1	0.0040	-5.5241	-0.0220	0.0022	-6.1143	-0.0135	0.0000		
LDHC	0.0000			0.0039	-5.5342	-0.0219	0.0000		
LGALS3BP	0.0194	-3.9417	-0.0765	0.0000			0.0000		
LRPPRC	0.0005	-7.5791	-0.0039	0.0000			0.0000		
LRRK1	0.0000			0.0000			0.0072	-4.9381	-0.0354
LTF	0.0727	-2.6208	-0.1907	0.0399	-3.2222	-0.1285	0.0425	-3.1587	-0.1342
MDH2	0.0000			0.0034	-5.6918	-0.0192	0.0000		
MGAM	0.0000			0.0018	-6.3335	-0.0112	0.0000		

Protein	Human pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Chimp pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Gorilla pi=sample/ sum	Ln(pi)	pi*Ln(pi)
MME	0.0000			0.0006	-7.4966	-0.0042	0.0000		
MSMB	0.0046	-5.3809	-0.0248	0.0000			0.0000		
MUC6	0.0253	-3.6772	-0.0930	0.0000			0.0000		
NBPF3	0.0000			0.0000			0.0079	-4.8419	-0.0382
NDUFS1	0.0000			0.0000			0.0151	-4.1928	-0.0633
NEMF	0.0000			0.0000		0.0000	0.0034	-5.6785	-0.0194
NPC2	0.0136	-4.2951	-0.0586	0.0198	-3.9200	-0.0778	0.0000		
ORM1	0.0014	-6.6030	-0.0090	0.0017	-6.3850	-0.0108	0.0000		
OXCT2	0.0000			0.0013	-6.6615	-0.0085	0.0000		
PAEP	0.0000			0.0055	-5.2115	-0.0284	0.1068	-2.2368	-0.2389
PATE1	0.0021	-6.1658	-0.0129	0.0000		0.0000	0.0000		
PGAM2	0.0000			0.0008	-7.1566	-0.0056	0.0000		
PGC	0.0026	-5.9390	-0.0156	0.0000		0.0000	0.0000		
PGK2	0.0000			0.0015	-6.5251	-0.0096	0.0000		
PIP	0.1180	-2.1369	-0.2522	0.0000		0.0000	0.0140	-4.2722	-0.0596
PKM	0.0000			0.0049	-5.3089	-0.0263	0.0000		
PLA1A	0.0000			0.0015	-6.5313	-0.0095	0.0000		
PLS3	0.0000			0.0012	-6.7067	-0.0082	0.0000		
PPIB	0.0000			0.0047	-5.3597	-0.0252	0.0090	-4.7118	-0.0424
PRDX6	0.0042	-5.4655	-0.0231	0.0039	-5.5408	-0.0217	0.0000		
PSAP	0.0152	-4.1844	-0.0637	0.0000			0.0000		
PSCA	0.0000			0.0066	-5.0223	-0.0331	0.0000		
QSOX1	0.0018	-6.3376	-0.0112	0.0000			0.0000		
RLTPR	0.0000			0.0008	-7.1566	-0.0056	0.0000		
RNASET2	0.0000			0.0084	-4.7741	-0.0403	0.0000		
S100A8	0.0033	-5.7281	-0.0186	0.0000			0.0000		
SEMG1	0.1388	-1.9750	-0.2741	0.4362	-0.8297	-0.3619	0.0325	-3.4271	-0.1113

Protein	Human pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Chimp pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Gorilla pi=sample/ sum	Ln(pi)	pi*Ln(pi)
SEMG2	0.2029	-1.5953	-0.3236	0.0147	-4.2226	-0.0619	0.1058	-2.2462	-0.2376
SERPINA1	0.0105	-4.5594	-0.0477	0.0188	-3.9753	-0.0746	0.0250	-3.6906	-0.0921
SERPINA3	0.0008	-7.1208	-0.0058	0.0184	-3.9952	-0.0735	0.0000		
SERPINA5	0.0191	-3.9594	-0.0755	0.0347	-3.3617	-0.1166	0.0000		
SERPINF2	0.0019	-6.2688	-0.0119	0.0000			0.0000		
SLC9B2	0.0013	-6.6701	-0.0085	0.0000			0.0000		
SMG1	0.0008	-7.1245	-0.0057	0.0000			0.0000		
SOD2	0.0000			0.0007	-7.3065	-0.0049	0.0000		
SOD3	0.0007	-7.1971	-0.0054	0.0000			0.0000		
SORD	0.0019	-6.2593	-0.0120	0.0000			0.0000		
SRCAP	0.0000			0.0000			0.0052	-5.2501	-0.0275
TF	0.0164	-4.1131	-0.0673	0.0095	-4.6590	-0.0441	0.0000		
TGM4	0.0036	-5.6166	-0.0204	0.0188	-3.9724	-0.0748	0.0000		
TIMP1	0.0056	-5.1880	-0.0290	0.0000			0.0000		
TKFC	0.0000			0.0006	-7.4580	-0.0043	0.0000		
TMPRSS2	0.0025	-6.0040	-0.0148	0.0000			0.0000		
TRRAP	0.0000			0.0000			0.0049	-5.3218	-0.0260
TSSC1	0.0000			0.0000			0.0128	-4.3545	-0.0560
TTL5	0.0000			0.0000			0.0111	-4.4998	-0.0500
TTN	0.0020	-6.2359	-0.0122	0.0007	-7.2855	-0.0050	0.0070	-4.9614	-0.0347
TUBA1A	0.0000			0.0022	-6.1051	-0.0136	0.0000		
TUBB	0.0000			0.0011	-6.7715	-0.0078	0.0000		
UBA1	0.0000			0.0000			0.0073	-4.9215	-0.0359
UBA52	0.0052	-5.2623	-0.0273	0.0000			0.0000		
WFDC2	0.0024	-6.0501	-0.0143	0.0000			0.0000		
WTIP	0.0008	-7.1351	-0.0057	0.0000			0.0000		
YWHAB	0.0019	-6.2420	-0.0121	0.0000			0.0000		

Protein	Human pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Chimp pi=sample/ sum	Ln(pi)	pi*Ln(pi)	Gorilla pi=sample/ sum	Ln(pi)	pi*Ln(pi)
ZCCHC11	0.0000			0.0000			0.0294	-3.5268	-0.1037
ZFHX2	0.0000			0.0000			0.0034	-5.6785	
ZNF148	0.0007	-7.2755	-0.0050	0.0000			0.0000		
	Sums	71.0000	-3.0438		64.0000	-2.5728		36.0000	-2.7769

Human

Hmax = Ln(71)

H (diversity)

Eveness = Hmax/E(pi*Ln(pi))

4.26

3.04

0.71

Chimp

Hmax = Ln(64)

H (diversity)

Eveness = Hmax/E(pi*Ln(pi))

4.16

2.57

0.62

Gorilla

Hmax = Ln(35)

H (diversity)

Eveness = Hmax/E(pi*Ln(pi))

3.56

2.78

0.78

Table A.2: Simpson diversity calculation with raw normalization

Protein	Human Avg	n	n(n-1)	Chimp Avg	n	n(n-1)	Gorilla Avg	n	n(n-1)
ACE	0.0000	0.0000	0.0000	0.0043	0.2777	-0.2006	0.0000	0.0000	0.0000
ACO2	0.0000	0.0000	0.0000	0.0007	0.0440	-0.0421	0.0000	0.0000	0.0000
ACOT13	0.0000	0.0000	0.0000	0.0015	0.0992	-0.0894	0.0000	0.0000	0.0000
ACPP	0.0412	2.9247	5.6293	0.0237	1.5144	0.7790	0.0000	0.0000	0.0000
ACR	0.0000	0.0000	0.0000	0.0015	0.0969	-0.0875	0.0000	0.0000	0.0000
ACRBP	0.0000	0.0000	0.0000	0.0060	0.3842	-0.2366	0.0000	0.0000	0.0000
ACTA2	0.0058	0.4122	-0.2423	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ACTB	0.0000	0.0000	0.0000	0.0068	0.4382	-0.2462	0.0000	0.0000	0.0000
ALB	0.0566	4.0200	12.1401	0.0803	5.1380	21.2607	0.2591	9.0694	73.1841
AMY2B	0.0000	0.0000	0.0000	0.0113	0.7222	-0.2006	0.0000	0.0000	0.0000
ANPEP	0.0072	0.5123	-0.2498	0.0012	0.0738	-0.0684	0.0000	0.0000	0.0000
APOB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0271	0.9497	-0.0478
ARHGAP31	0.0007	0.0500	-0.0475	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ARNT	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0182	0.6355	-0.2316
ASAP2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0058	0.2030	-0.1618
AZGP1	0.0243	1.7285	1.2592	0.0092	0.5894	-0.2420	0.0000	0.0000	0.0000
B2M	0.0024	0.1705	-0.1414	0.0011	0.0716	-0.0665	0.0000	0.0000	0.0000
B4GALT1	0.0000	0.0000	0.0000	0.0020	0.1281	-0.1117	0.0000	0.0000	0.0000
BCL3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0109	0.3805	-0.2357
BIRC6	0.0008	0.0574	-0.0541	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CAP1	0.0019	0.1351	-0.1169	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CCDC88C	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0614	2.1493	2.4703
CEP162	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0138	0.4832	-0.2497
CFAP69	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0068	0.2366	-0.1806
CKB	0.0023	0.1652	-0.1379	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CLEC11A	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0059	0.2066	-0.1639
CLIP2	0.0008	0.0558	-0.0526	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CLU	0.0236	1.6781	1.1379	0.0457	2.9241	5.6263	0.0709	2.4823	3.6796

Protein	Human Avg	n	n(n-1)	Chimp Avg	n	n(n-1)	Gorilla Avg	n	n(n-1)
HSPA5	0.0020	0.1407	-0.1209	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IDH1	0.0018	0.1274	-0.1112	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IGHG1	0.0047	0.3363	-0.2232	0.0144	0.9209	-0.0728	0.0000	0.0000	0.0000
IGKC	0.0000	0.0000	0.0000	0.0083	0.5335	-0.2489	0.0000	0.0000	0.0000
KLK3	0.0209	1.4869	0.7240	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
KRT1	0.0054	0.3816	-0.2360	0.0050	0.3216	-0.2182	0.0374	1.3107	0.4072
KRT10	0.0044	0.3132	-0.2151	0.0012	0.0743	-0.0687	0.0034	0.1197	-0.1053
KRT2	0.0078	0.5545	-0.2470	0.0011	0.0719	-0.0668	0.0000	0.0000	0.0000
LAMP2	0.0024	0.1677	-0.1396	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LCP1	0.0040	0.2833	-0.2030	0.0022	0.1415	-0.1215	0.0000	0.0000	0.0000
LDHC	0.0000	0.0000	0.0000	0.0039	0.2528	-0.1889	0.0000	0.0000	0.0000
LGALS3BP	0.0194	1.3785	0.5217	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LRPPRC	0.0005	0.0363	-0.0350	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
LRRK1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0072	0.2509	-0.1879
LTF	0.0727	5.1650	21.5124	0.0399	2.5516	3.9591	0.0425	1.4869	0.7239
MDH2	0.0000	0.0000	0.0000	0.0034	0.2159	-0.1693	0.0000	0.0000	0.0000
MGAM	0.0000	0.0000	0.0000	0.0018	0.1137	-0.1007	0.0000	0.0000	0.0000
MME	0.0000	0.0000	0.0000	0.0006	0.0355	-0.0343	0.0000	0.0000	0.0000
MSMB	0.0046	0.3269	-0.2200	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MUC6	0.0253	1.7958	1.4291	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
NBPF3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0079	0.2762	-0.1999
NDUFS1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0151	0.5286	-0.2492
NEMF	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0034	0.1197	-0.1053
NPC2	0.0136	0.9681	-0.0309	0.0198	1.2698	0.3426	0.0000	0.0000	0.0000
ORM1	0.0014	0.0963	-0.0870	0.0017	0.1080	-0.0963	0.0000	0.0000	0.0000
OXCT2	0.0000	0.0000	0.0000	0.0013	0.0819	-0.0752	0.0000	0.0000	0.0000
PAEP	0.0000	0.0000	0.0000	0.0055	0.3490	-0.2272	0.1068	3.7381	10.2350
PATE1	0.0021	0.1491	-0.1269	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PGAM2	0.0000	0.0000	0.0000	0.0008	0.0499	-0.0474	0.0000	0.0000	0.0000

Protein	Human Avg	n	n(n-1)	Chimp Avg	n	n(n-1)	Gorilla Avg	n	n(n-1)
TKFC	0.0000	0.0000	0.0000	0.0006	0.0369	-0.0356	0.0000	0.0000	0.0000
TMPRSS2	0.0025	0.1753	-0.1446	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
TRRAP	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0049	0.1709	-0.1417
TSSC1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0128	0.4497	-0.2475
TTLL5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0111	0.3889	-0.2377
TTN	0.0020	0.1390	-0.1197	0.0007	0.0439	-0.0419	0.0070	0.2451	-0.1850
TUBA1A	0.0000	0.0000	0.0000	0.0022	0.1428	-0.1224	0.0000	0.0000	0.0000
TUBB	0.0000	0.0000	0.0000	0.0011	0.0733	-0.0680	0.0000	0.0000	0.0000
UBA1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0073	0.2551	-0.1900
UBA52	0.0052	0.3680	-0.2326	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
WFDC2	0.0024	0.1674	-0.1394	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
WTIP	0.0008	0.0566	-0.0534	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
YWHAB	0.0019	0.1382	-0.1191	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ZCCHC11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0294	1.0290	0.0298
ZFHX2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0034	0.1197	-0.1053
ZNF148	0.0007	0.0492	-0.0467	0.0000	0.0000	0.0000	0.00000	0.0000	0.0000

	n	n(n-1)		n	n(n-1)		n	n(n-1)
Sum of columns	71	380.85		64	788.12		35	96.17

	Human		Chimp		Gorilla	
Statistics	D=	0.08	D=	0.20	D=	0.08
	1/d=	0.92	1/d=	0.80	1/d=	0.92

A.5.1 Script in R for comparing biodiversity index statistics between raw and length adjusted data

```
setwd("~/Desktop")
ShannonsDI <- read.csv("Shannon_csv.csv")
SimpsonDI <- read.csv("Simpson_csv.csv")

ShannonsDI

##      Species Index  Raw Length
## 1      Human      H 3.04   2.92
## 2      Human      EH 0.71   0.69
## 3 Chimpanzee      H 2.57   2.93
## 4 Chimpanzee      EH 0.62   0.71
## 5      Gorilla      H 2.78   2.43
## 6      Gorilla      EH 0.78   0.68

SimpsonDI

##      Species Index  Raw Length
## 1      Human      1/D 0.92   0.90
## 2 Chimpanzee      1/D 0.80   0.90
## 3      Gorilla      1/D 0.92   0.87

Shannon_tTEST <- t.test(ShannonsDI$Raw, ShannonsDI$Length, paired=TRUE)
Simpson_tTEST <- t.test(SimpsonDI$Raw, SimpsonDI$Length, paired=TRUE)
Shannon_tTEST

##
## Paired t-test
##
## data:  ShannonsDI$Raw and ShannonsDI$Length
## t = 0.24078, df = 5, p-value = 0.8193
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.2257760  0.2724426
## sample estimates:
## mean of the differences
##              0.02333333

Simpson_tTEST
## Paired t-test
##
## data:  SimpsonDI$Raw and SimpsonDI$Length
## t = -0.21822, df = 2, p-value = 0.8475
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.2071723  0.1871723
## sample estimates:
## mean of the differences
##              -0.01
```

A.6 Complete tables of LC/MS-MS high confidence identified proteins

Table A.3: All proteins identified with high confidence in each human individual

Protein	Method	Hu#1	Hu#2	Hu#3	Human Avg \pm SEM	
ACPP	Length	5.87%	2.43%	3.30%	3.86%	\pm 1.03%
	Raw	5.77%	2.57%	4.01%	4.12%	\pm 0.92%
ACTA2	Length	0.64%	0.47%	0.53%	0.54%	\pm 0.05%
	Raw	0.62%	0.49%	0.64%	0.58%	\pm 0.05%
ALB	Length	4.67%	2.69%	2.75%	3.37%	\pm 0.65%
	Raw	7.23%	4.50%	5.26%	5.66%	\pm 0.81%
ANPEP	Length	0.40%	0.03%	0.33%	0.25%	\pm 0.11%
	Raw	1.06%	0.08%	1.03%	0.72%	\pm 0.32%
APOB	Length	0.00%	0.00%	0.00%	0.00%	\pm 0.00%
	Raw	0.12%	0.09%	0.00%	0.07%	\pm 0.04%
ARHGAP31	Length	0.03%	0.02%	0.00%	0.02%	\pm 0.01%
	Raw	0.12%	0.09%	0.00%	0.07%	\pm 0.04%
AZGP1	Length	3.72%	1.77%	3.07%	2.85%	\pm 0.57%
	Raw	2.89%	1.48%	2.94%	2.43%	\pm 0.48%
B2M	Length	1.62%	0.49%	0.33%	0.82%	\pm 0.40%
	Raw	0.45%	0.15%	0.12%	0.24%	\pm 0.11%
BIRC6	Length	0.00%	0.00%	0.00%	0.00%	\pm 0.00%
	Raw	0.00%	0.00%	0.24%	0.08%	\pm 0.08%
CAP1	Length	0.30%	0.00%	0.14%	0.15%	\pm 0.09%
	Raw	0.33%	0.00%	0.24%	0.19%	\pm 0.10%
CKB	Length	0.51%	0.15%	0.00%	0.22%	\pm 0.15%
	Raw	0.54%	0.15%	0.00%	0.23%	\pm 0.16%
CLIP2	Length	0.00%	0.00%	0.08%	0.03%	\pm 0.03%
	Raw	0.00%	0.00%	0.24%	0.08%	\pm 0.08%
CLU	Length	1.72%	1.43%	2.39%	1.85%	\pm 0.29%
	Raw	1.91%	1.77%	3.41%	2.36%	\pm 0.53%
CPE	Length	0.71%	0.06%	0.33%	0.37%	\pm 0.19%
	Raw	0.78%	0.09%	0.48%	0.45%	\pm 0.20%
CRISP1	Length	1.16%	1.06%	1.92%	1.38%	\pm 0.27%
	Raw	0.67%	0.73%	1.49%	0.96%	\pm 0.26%
CST2	Length	0.00%	0.00%	0.28%	0.09%	\pm 0.09%
	Raw	0.00%	0.15%	0.12%	0.09%	\pm 0.05%
CST3	Length	0.00%	0.60%	1.09%	0.56%	\pm 0.32%
	Raw	0.00%	0.09%	0.48%	0.19%	\pm 0.15%

Protein	Method	Hu#1	Hu#2	Hu#3	Human Avg ± SEM	
CST4	Length	0.34%	1.87%	1.98%	1.40%	± 0.53%
	Raw	0.10%	0.72%	0.89%	0.57%	± 0.24%
CTSD	Length	0.23%	0.64%	0.00%	0.29%	± 0.19%
	Raw	0.24%	0.73%	0.00%	0.32%	± 0.21%
DEFA1	Length	1.54%	0.62%	2.12%	1.43%	± 0.44%
	Raw	0.38%	0.15%	0.62%	0.38%	± 0.13%
DHX57	Length	0.00%	0.04%	0.00%	0.01%	± 0.01%
	Raw	0.00%	0.16%	0.00%	0.05%	± 0.05%
DUT	Length	0.00%	0.12%	0.16%	0.09%	± 0.05%
	Raw	0.00%	0.09%	0.12%	0.07%	± 0.04%
ECM1	Length	0.54%	0.33%	0.52%	0.46%	± 0.07%
	Raw	0.67%	0.49%	0.85%	0.67%	± 0.11%
FN1	Length	0.69%	0.13%	0.35%	0.39%	± 0.16%
	Raw	4.18%	0.91%	2.62%	2.57%	± 0.94%
FOLH1	Length	0.26%	0.00%	0.00%	0.09%	± 0.09%
	Raw	0.46%	0.00%	0.00%	0.15%	± 0.15%
GDF15	Length	1.25%	0.00%	0.00%	0.42%	± 0.42%
	Raw	0.98%	0.00%	0.00%	0.33%	± 0.33%
GOLGB1	Length	0.03%	0.00%	0.00%	0.01%	± 0.01%
	Raw	0.24%	0.00%	0.00%	0.08%	± 0.08%
HEXB	Length	0.43%	0.11%	0.14%	0.23%	± 0.10%
	Raw	0.57%	0.18%	0.24%	0.33%	± 0.12%
HSPA1L	Length	0.15%	0.05%	0.19%	0.13%	± 0.04%
	Raw	0.22%	0.08%	0.38%	0.23%	± 0.09%
HSPA5	Length	0.15%	0.13%	0.06%	0.11%	± 0.03%
	Raw	0.22%	0.23%	0.14%	0.20%	± 0.03%
IDH1	Length	0.23%	0.07%	0.19%	0.17%	± 0.05%
	Raw	0.22%	0.08%	0.24%	0.18%	± 0.05%
IGHG1	Length	1.02%	0.18%	0.36%	0.52%	± 0.26%
	Raw	0.87%	0.15%	0.40%	0.47%	± 0.21%
KLK3	Length	3.32%	2.13%	3.05%	2.83%	± 0.36%
	Raw	2.24%	1.54%	2.50%	2.09%	± 0.29%
KRT1	Length	0.45%	0.32%	0.19%	0.32%	± 0.08%
	Raw	0.69%	0.55%	0.38%	0.54%	± 0.09%
KRT10	Length	0.74%	0.00%	0.14%	0.29%	± 0.23%
	Raw	1.06%	0.00%	0.26%	0.44%	± 0.32%

Protein	Method	Hu#1	Hu#2	Hu#3	Human Avg ± SEM	
KRT2	Length	1.13%	0.09%	0.19%	0.47%	± 0.33%
	Raw	1.83%	0.15%	0.36%	0.78%	± 0.53%
LAMP2	Length	0.47%	0.07%	0.39%	0.31%	± 0.12%
	Raw	0.50%	0.09%	0.12%	0.24%	± 0.13%
LCP1	Length	0.08%	0.28%	0.06%	0.14%	± 0.07%
	Raw	0.15%	0.53%	0.52%	0.40%	± 0.12%
LGALS3BP	Length	1.81%	0.60%	1.16%	1.19%	± 0.35%
	Raw	2.74%	0.96%	2.13%	1.94%	± 0.52%
LRPPRC	Length	0.00%	0.04%	0.00%	0.01%	± 0.01%
	Raw	0.00%	0.00%	0.00%	0.00%	± 0.00%
LTF	Length	3.26%	2.51%	4.82%	3.53%	± 0.68%
	Raw	6.04%	4.97%	10.82%	7.27%	± 1.80%
MSMB	Length	3.80%	1.03%	0.00%	1.61%	± 1.14%
	Raw	1.05%	0.33%	0.00%	0.46%	± 0.31%
MUC6	Length	0.51%	0.19%	0.39%	0.37%	± 0.09%
	Raw	3.28%	1.31%	3.00%	2.53%	± 0.62%
NPC2	Length	4.47%	1.36%	3.69%	3.17%	± 0.93%
	Raw	1.78%	0.57%	1.75%	1.36%	± 0.40%
ORM1	Length	0.72%	0.00%	0.00%	0.24%	± 0.24%
	Raw	0.41%	0.00%	0.00%	0.14%	± 0.14%
PATE1	Length	0.00%	0.00%	1.58%	0.53%	± 0.53%
	Raw	0.00%	0.00%	0.63%	0.21%	± 0.21%
PGC	Length	0.25%	0.30%	0.10%	0.22%	± 0.06%
	Raw	0.30%	0.35%	0.14%	0.26%	± 0.06%
PIP	Length	16.17%	32.46%	36.55%	28.40%	± 6.23%
	Raw	5.72%	13.04%	16.65%	11.80%	± 3.22%
PRDX6	Length	1.29%	0.65%	0.18%	0.71%	± 0.32%
	Raw	0.72%	0.41%	0.14%	0.42%	± 0.17%
PSAP	Length	1.47%	0.89%	0.84%	1.07%	± 0.20%
	Raw	1.88%	1.30%	1.39%	1.52%	± 0.18%
QSOX1	Length	0.06%	0.20%	0.00%	0.09%	± 0.06%
	Raw	0.15%	0.38%	0.00%	0.18%	± 0.11%
S100A8	Length	0.52%	0.31%	2.57%	1.13%	± 0.72%
	Raw	0.15%	0.09%	0.73%	0.33%	± 0.21%
SEMG1	Length	7.09%	21.47%	4.31%	10.96%	± 5.32%
	Raw	8.11%	27.29%	6.23%	13.88%	± 6.73%

Protein	Method	Hu#1	Hu#2	Hu#3	Human Avg ± SEM	
SEMG2	Length	12.83%	15.03%	9.51%	12.46%	± 1.60%
	Raw	19.33%	24.12%	17.41%	20.29%	± 2.00%
SERPINA1	Length	0.92%	0.84%	0.95%	0.90%	± 0.03%
	Raw	0.96%	0.94%	1.24%	1.05%	± 0.09%
SERPINA3	Length	0.00%	0.00%	0.19%	0.06%	± 0.06%
	Raw	0.00%	0.00%	0.24%	0.08%	± 0.08%
SERPINA5	Length	2.61%	1.23%	1.28%	1.70%	± 0.45%
	Raw	2.67%	1.37%	1.68%	1.91%	± 0.39%
SERPINF2	Length	0.00%	0.06%	0.32%	0.13%	± 0.10%
	Raw	0.00%	0.09%	0.48%	0.19%	± 0.15%
SLC9B2	Length	0.09%	0.00%	0.15%	0.08%	± 0.04%
	Raw	0.10%	0.00%	0.28%	0.13%	± 0.08%
SMG1	Length	0.03%	0.00%	0.00%	0.01%	± 0.01%
	Raw	0.24%	0.00%	0.00%	0.08%	± 0.08%
SOD3	Length	0.40%	0.00%	0.00%	0.13%	± 0.13%
	Raw	0.22%	0.00%	0.00%	0.07%	± 0.07%
SORD	Length	0.27%	0.25%	0.11%	0.21%	± 0.05%
	Raw	0.22%	0.23%	0.12%	0.19%	± 0.04%
TF	Length	1.31%	1.01%	0.23%	0.85%	± 0.32%
	Raw	2.38%	1.98%	0.55%	1.64%	± 0.55%
TGM4	Length	0.07%	0.26%	0.23%	0.19%	± 0.06%
	Raw	0.10%	0.48%	0.51%	0.36%	± 0.13%
TIMP1	Length	1.40%	0.71%	0.77%	0.96%	± 0.22%
	Raw	0.80%	0.40%	0.48%	0.56%	± 0.12%
TMPRSS2	Length	0.10%	0.00%	0.40%	0.17%	± 0.12%
	Raw	0.10%	0.00%	0.64%	0.25%	± 0.20%
TTN	Length	0.00%	0.00%	0.00%	0.00%	± 0.00%
	Raw	0.47%	0.00%	0.12%	0.20%	± 0.14%
UBA52	Length	1.88%	0.00%	2.18%	1.35%	± 0.68%
	Raw	0.68%	0.00%	0.88%	0.52%	± 0.27%
WFDC2	Length	1.17%	0.24%	0.64%	0.68%	± 0.27%
	Raw	0.39%	0.08%	0.24%	0.24%	± 0.09%
WTIP	Length	0.00%	0.00%	0.19%	0.06%	± 0.06%
	Raw	0.00%	0.00%	0.24%	0.08%	± 0.08%
YWHAB	Length	0.98%	0.00%	0.00%	0.33%	± 0.33%
	Raw	0.58%	0.00%	0.00%	0.19%	± 0.19%
ZNF148	Length	0.00%	0.00%	0.00%	0.00%	± 0.00%
	Raw	0.21%	0.00%	0.00%	0.07%	± 0.07%

Table A.4: All proteins identified with high confidence in each chimpanzee individual

Protein	Method	Ch#1	Ch#2	Ch#3	Chimp Average \pm SEM	
ACE	Length	0.15%	0.34%	0.00%	0.16%	\pm 0.10%
	Raw	0.37%	0.93%	0.00%	0.43%	\pm 0.27%
ACO2	Length	0.00%	0.05%	0.08%	0.04%	\pm 0.02%
	Raw	0.00%	0.08%	0.12%	0.07%	\pm 0.04%
ACOT13	Length	0.00%	1.16%	0.43%	0.53%	\pm 0.34%
	Raw	0.00%	0.34%	0.12%	0.15%	\pm 0.10%
ACPP	Length	1.35%	3.05%	4.71%	3.04%	\pm 0.97%
	Raw	0.96%	2.47%	3.67%	2.37%	\pm 0.78%
ACR	Length	0.15%	0.10%	0.29%	0.18%	\pm 0.06%
	Raw	0.12%	0.09%	0.25%	0.15%	\pm 0.05%
ACRBP	Length	0.48%	0.75%	0.45%	0.56%	\pm 0.10%
	Raw	0.47%	0.85%	0.48%	0.60%	\pm 0.12%
ACTB	Length	1.39%	0.76%	0.65%	0.93%	\pm 0.23%
	Raw	0.96%	0.59%	0.50%	0.68%	\pm 0.14%
ALB	Length	8.14%	7.48%	4.48%	6.70%	\pm 1.13%
	Raw	9.10%	9.53%	5.46%	8.03%	\pm 1.29%
AMY2B	Length	0.89%	1.11%	1.31%	1.10%	\pm 0.12%
	Raw	0.84%	1.19%	1.36%	1.13%	\pm 0.15%
ANPEP	Length	0.00%	0.17%	0.00%	0.06%	\pm 0.06%
	Raw	0.00%	0.35%	0.00%	0.12%	\pm 0.12%
AZGP1	Length	0.44%	1.50%	2.65%	1.53%	\pm 0.64%
	Raw	0.24%	0.93%	1.59%	0.92%	\pm 0.39%
B2M	Length	0.00%	0.34%	1.02%	0.45%	\pm 0.30%
	Raw	0.00%	0.09%	0.25%	0.11%	\pm 0.07%
B4GALT1	Length	0.33%	0.00%	0.46%	0.26%	\pm 0.14%
	Raw	0.24%	0.00%	0.36%	0.20%	\pm 0.11%
CLU	Length	6.55%	2.75%	4.33%	4.54%	\pm 1.10%
	Raw	6.23%	2.96%	4.51%	4.57%	\pm 0.94%
CRISP1	Length	0.00%	0.49%	0.49%	0.33%	\pm 0.16%
	Raw	0.00%	0.25%	0.25%	0.17%	\pm 0.08%
CRTAC1	Length	1.18%	1.48%	1.38%	1.35%	\pm 0.09%
	Raw	1.44%	2.03%	1.82%	1.76%	\pm 0.17%
CST3	Length	1.34%	1.95%	2.08%	1.79%	\pm 0.23%
	Raw	0.36%	0.60%	0.61%	0.52%	\pm 0.08%
DPEP3	Length	0.67%	0.58%	0.50%	0.58%	\pm 0.05%
	Raw	0.60%	0.60%	0.49%	0.56%	\pm 0.04%

Protein	Method	Ch#1	Ch#2	Ch#3	Chimp Average ± SEM	
ECM1	Length	0.97%	0.75%	1.01%	0.91%	± 0.08%
	Raw	0.96%	0.85%	1.10%	0.97%	± 0.07%
ENO1	Length	0.00%	0.28%	0.00%	0.09%	± 0.09%
	Raw	0.00%	0.25%	0.00%	0.08%	± 0.08%
FAM3D	Length	0.00%	0.36%	0.00%	0.12%	± 0.12%
	Raw	0.00%	0.17%	0.00%	0.06%	± 0.06%
FN1	Length	1.53%	0.77%	1.09%	1.13%	± 0.22%
	Raw	6.70%	3.82%	5.22%	5.25%	± 0.83%
GPI	Length	0.12%	0.22%	0.43%	0.26%	± 0.09%
	Raw	0.12%	0.26%	0.49%	0.29%	± 0.11%
HEXB	Length	0.00%	0.37%	0.00%	0.12%	± 0.12%
	Raw	0.00%	0.43%	0.00%	0.14%	± 0.14%
HK1	Length	0.14%	0.35%	0.07%	0.19%	± 0.09%
	Raw	0.24%	0.68%	0.12%	0.34%	± 0.17%
HSP90AA1	Length	0.00%	0.22%	0.00%	0.07%	± 0.07%
	Raw	0.00%	0.34%	0.00%	0.11%	± 0.11%
HSPA1A	Length	0.10%	0.13%	0.09%	0.11%	± 0.01%
	Raw	0.12%	0.17%	0.12%	0.14%	± 0.02%
IGHG1	Length	2.77%	2.22%	1.65%	2.21%	± 0.32%
	Raw	1.68%	1.54%	1.10%	1.44%	± 0.18%
IGKC	Length	6.15%	4.22%	1.72%	4.03%	± 1.28%
	Raw	1.21%	0.94%	0.36%	0.83%	± 0.25%
KRT1	Length	0.10%	0.76%	0.28%	0.38%	± 0.20%
	Raw	0.12%	1.03%	0.36%	0.50%	± 0.27%
KRT10	Length	0.00%	0.28%	0.00%	0.09%	± 0.09%
	Raw	0.00%	0.35%	0.00%	0.12%	± 0.12%
KRT2	Length	0.00%	0.25%	0.00%	0.08%	± 0.08%
	Raw	0.00%	0.34%	0.00%	0.11%	± 0.11%
LCP1	Length	0.10%	0.32%	0.10%	0.17%	± 0.07%
	Raw	0.12%	0.42%	0.12%	0.22%	± 0.10%
LDHC	Length	0.00%	1.71%	0.00%	0.57%	± 0.57%
	Raw	0.00%	1.18%	0.00%	0.39%	± 0.39%
LTF	Length	6.43%	1.09%	1.37%	2.96%	± 1.74%
	Raw	8.40%	1.62%	1.95%	3.99%	± 2.21%
MDH2	Length	0.00%	1.44%	0.00%	0.48%	± 0.48%
	Raw	0.00%	1.01%	0.00%	0.34%	± 0.34%
MGAM	Length	0.11%	0.04%	0.00%	0.05%	± 0.03%
	Raw	0.36%	0.18%	0.00%	0.18%	± 0.10%

Protein	Method	Ch#1	Ch#2	Ch#3	Chimp Average ± SEM	
MME	Length	0.00%	0.11%	0.00%	0.04%	± 0.04%
	Raw	0.00%	0.17%	0.00%	0.06%	± 0.06%
NPC2	Length	6.05%	8.08%	5.62%	6.58%	± 0.76%
	Raw	1.69%	2.55%	1.71%	1.98%	± 0.28%
ORM1	Length	0.00%	1.21%	0.00%	0.40%	± 0.40%
	Raw	0.00%	0.51%	0.00%	0.17%	± 0.17%
OXCT2	Length	0.00%	0.24%	0.12%	0.12%	± 0.07%
	Raw	0.00%	0.26%	0.12%	0.13%	± 0.07%
PAEP	Length	0.00%	1.81%	2.70%	1.50%	± 0.79%
	Raw	0.00%	0.68%	0.96%	0.55%	± 0.28%
PATE1	Length	0.00%	0.00%	0.00%	0.00%	± 0.00%
	Raw	0.00%	0.00%	0.23%	0.08%	± 0.08%
PGAM2	Length	0.00%	0.00%	0.48%	0.16%	± 0.16%
	Raw	0.00%	0.00%	0.00%	0.00%	± 0.00%
PGK2	Length	0.47%	0.10%	0.00%	0.19%	± 0.14%
	Raw	0.36%	0.08%	0.00%	0.15%	± 0.11%
PIP	Length	0.00%	0.00%	0.00%	0.00%	± 0.00%
	Raw	0.00%	0.00%	0.00%	0.00%	± 0.00%
PKM	Length	0.74%	0.69%	0.00%	0.48%	± 0.24%
	Raw	0.72%	0.77%	0.00%	0.49%	± 0.25%
PLA1A	Length	0.43%	0.09%	0.00%	0.17%	± 0.13%
	Raw	0.36%	0.08%	0.00%	0.15%	± 0.11%
PLS3	Length	0.10%	0.19%	0.00%	0.10%	± 0.06%
	Raw	0.12%	0.25%	0.00%	0.12%	± 0.07%
PPIB	Length	0.30%	0.94%	1.97%	1.07%	± 0.48%
	Raw	0.12%	0.43%	0.86%	0.47%	± 0.21%
PRDX6	Length	1.46%	0.73%	0.54%	0.91%	± 0.28%
	Raw	0.60%	0.34%	0.24%	0.39%	± 0.11%
PSCA	Length	1.59%	2.97%	3.45%	2.67%	± 0.56%
	Raw	0.36%	0.76%	0.86%	0.66%	± 0.15%
RLTPR	Length	0.00%	0.00%	0.08%	0.03%	± 0.03%
	Raw	0.00%	0.00%	0.23%	0.08%	± 0.08%
RNASET2	Length	1.78%	1.59%	1.66%	1.68%	± 0.06%
	Raw	0.83%	0.85%	0.85%	0.84%	± 0.01%
SEMG1	Length	31.21%	23.60%	31.01%	28.61%	± 2.50%
	Raw	44.29%	38.65%	47.92%	43.62%	± 2.70%
SEMG2	Length	3.05%	0.90%	1.64%	1.86%	± 0.63%
	Raw	2.28%	0.77%	1.35%	1.47%	± 0.44%

Protein	Method	Ch#1	Ch#2	Ch#3	Chimp Average ± SEM	
SERPINA1	Length	1.25%	2.43%	3.05%	2.24%	± 0.53%
	Raw	0.96%	2.12%	2.55%	1.88%	± 0.47%
SERPINA3	Length	1.08%	2.69%	2.72%	2.16%	± 0.54%
	Raw	0.84%	2.38%	2.31%	1.84%	± 0.50%
SERPINA5	Length	3.05%	4.01%	5.83%	4.29%	± 0.81%
	Raw	2.27%	3.39%	4.74%	3.47%	± 0.71%
SOD2	Length	0.00%	0.18%	0.27%	0.15%	± 0.08%
	Raw	0.00%	0.08%	0.12%	0.07%	± 0.03%
TF	Length	1.40%	0.64%	0.09%	0.71%	± 0.38%
	Raw	1.79%	0.93%	0.12%	0.95%	± 0.48%
TGM4	Length	0.38%	2.44%	1.24%	1.35%	± 0.60%
	Raw	0.47%	3.47%	1.70%	1.88%	± 0.87%
TKFC	Length	0.00%	0.14%	0.00%	0.05%	± 0.05%
	Raw	0.00%	0.17%	0.00%	0.06%	± 0.06%
TTN	Length	0.00%	0.00%	0.00%	0.00%	± 0.00%
	Raw	0.00%	0.09%	0.12%	0.07%	± 0.04%
TUBA1A	Length	0.00%	0.45%	0.27%	0.24%	± 0.13%
	Raw	0.00%	0.42%	0.25%	0.22%	± 0.12%
TUBB	Length	0.00%	0.37%	0.00%	0.12%	± 0.12%
	Raw	0.00%	0.34%	0.00%	0.11%	± 0.11%

Table A.5: All proteins identified with high confidence in each gorilla individual

Protein	Method	Go#1	Go#2	Go#3	Gorilla Average \pm SEM		
ALB	Length	40.24%	29.22%	4.67%	24.71%	\pm	10.51%
	Raw	41.27%	29.52%	6.95%	25.91%	\pm	10.07%
APOB	Length	0.52%	0.47%	0.04%	0.34%	\pm	0.15%
	Raw	4.29%	3.23%	0.63%	2.71%	\pm	1.09%
ARHGAP31	Length	0.00%	0.00%	0.00%	0.00%	\pm	0.00%
	Raw	0.00%	0.00%	0.00%	0.00%	\pm	0.00%
ARNT	Length	3.01%	1.80%	0.00%	1.60%	\pm	0.87%
	Raw	2.86%	2.59%	0.00%	1.82%	\pm	0.91%
ASAP2	Length	0.79%	0.00%	0.20%	0.33%	\pm	0.24%
	Raw	1.11%	0.00%	0.63%	0.58%	\pm	0.32%
BCL3	Length	1.74%	3.14%	0.00%	1.63%	\pm	0.91%
	Raw	1.11%	2.15%	0.00%	1.09%	\pm	0.62%
CCDC88C	Length	1.56%	4.21%	0.10%	1.96%	\pm	1.20%
	Raw	4.13%	13.78%	0.51%	6.14%	\pm	3.96%
CEP162	Length	0.56%	1.01%	0.00%	0.53%	\pm	0.29%
	Raw	1.11%	3.03%	0.00%	1.38%	\pm	0.89%
CFAP69	Length	0.00%	0.76%	0.22%	0.32%	\pm	0.23%
	Raw	0.00%	1.52%	0.51%	0.68%	\pm	0.44%
CLEC11A	Length	0.00%	0.00%	1.89%	0.63%	\pm	0.63%
	Raw	0.00%	0.00%	1.77%	0.59%	\pm	0.59%
CLU	Length	10.56%	7.93%	6.33%	8.27%	\pm	1.23%
	Raw	7.46%	6.26%	7.56%	7.09%	\pm	0.42%
CPSF1	Length	1.10%	0.00%	0.00%	0.37%	\pm	0.37%
	Raw	2.22%	0.00%	0.00%	0.74%	\pm	0.74%
CTSB	Length	0.00%	0.00%	1.20%	0.40%	\pm	0.40%
	Raw	0.00%	0.00%	1.03%	0.34%	\pm	0.34%
DNAH7	Length	0.00%	0.00%	0.10%	0.03%	\pm	0.03%
	Raw	0.00%	0.00%	1.26%	0.42%	\pm	0.42%
FAM184B	Length	0.75%	1.34%	0.38%	0.82%	\pm	0.28%
	Raw	0.95%	1.08%	1.09%	1.04%	\pm	0.04%
KRT1	Length	7.37%	1.11%	0.32%	2.93%	\pm	2.23%
	Raw	9.21%	1.52%	0.51%	3.74%	\pm	2.75%
KRT10	Length	0.00%	0.00%	0.70%	0.23%	\pm	0.23%
	Raw	0.00%	0.00%	1.03%	0.34%	\pm	0.34%
LRRK1	Length	0.00%	0.71%	0.00%	0.24%	\pm	0.24%
	Raw	0.00%	2.15%	0.00%	0.72%	\pm	0.72%

Protein	Method	Go#1	Go#2	Go#3	Gorilla Average \pm SEM		
LTF	Length	5.57%	4.01%	0.86%	3.48%	\pm	1.39%
	Raw	6.35%	4.74%	1.65%	4.25%	\pm	1.38%
NBPF3	Length	2.50%	0.00%	0.32%	0.94%	\pm	0.78%
	Raw	1.90%	0.00%	0.46%	0.79%	\pm	0.57%
NDUFS1	Length	3.26%	0.98%	0.00%	1.41%	\pm	0.97%
	Raw	3.02%	1.52%	0.00%	1.51%	\pm	0.87%
NEMF	Length	0.00%	0.00%	0.38%	0.13%	\pm	0.13%
	Raw	0.00%	0.00%	1.03%	0.34%	\pm	0.34%
PAEP	Length	17.57%	27.69%	34.97%	26.74%	\pm	5.05%
	Raw	5.56%	8.85%	17.64%	10.68%	\pm	3.61%
PIP	Length	0.00%	0.00%	11.13%	3.71%	\pm	3.71%
	Raw	0.00%	0.00%	4.19%	1.40%	\pm	1.40%
PPIB	Length	0.00%	0.00%	4.70%	1.57%	\pm	1.57%
	Raw	0.00%	0.00%	2.70%	0.90%	\pm	0.90%
SEMG1	Length	0.00%	0.00%	5.69%	1.90%	\pm	1.90%
	Raw	0.00%	0.00%	9.74%	3.25%	\pm	3.25%
SEMG2	Length	0.00%	0.00%	22.33%	7.44%	\pm	7.44%
	Raw	0.00%	0.00%	31.74%	10.58%	\pm	10.58%
SERPINA1	Length	0.00%	6.81%	2.43%	3.08%	\pm	1.99%
	Raw	0.00%	4.74%	2.75%	2.50%	\pm	1.37%
SRCAP	Length	0.24%	0.00%	0.06%	0.10%	\pm	0.07%
	Raw	1.11%	0.00%	0.46%	0.52%	\pm	0.32%
TRRAP	Length	0.20%	0.00%	0.05%	0.09%	\pm	0.06%
	Raw	0.95%	0.00%	0.51%	0.49%	\pm	0.28%
TSSC1	Length	0.00%	5.52%	0.52%	2.01%	\pm	1.76%
	Raw	0.00%	3.23%	0.63%	1.28%	\pm	0.99%
TTL5	Length	1.23%	0.00%	0.00%	0.41%	\pm	0.41%
	Raw	3.33%	0.00%	0.00%	1.11%	\pm	1.11%
TTN	Length	0.00%	0.02%	0.01%	0.01%	\pm	0.01%
	Raw	0.00%	1.08%	1.03%	0.70%	\pm	0.35%
UBA1	Length	0.75%	0.67%	0.00%	0.47%	\pm	0.24%
	Raw	1.11%	1.08%	0.00%	0.73%	\pm	0.36%
ZCCHC11	Length	0.48%	2.60%	0.25%	1.11%	\pm	0.75%
	Raw	0.95%	6.89%	0.98%	2.94%	\pm	1.98%
ZFHX2	Length	0.00%	0.00%	0.16%	0.05%	\pm	0.05%
	Raw	0.00%	0.00%	1.03%	0.34%	\pm	0.34%

Table A.6: Raw normalized abundances of all proteins identified in each human individual (in triplicate) arranged by prevalence

Protein	H1.1 %	H1.2 %	H1.3 %	Hu#1 % Avg ± SEM	H2.1 %	H2.2 %	H2.3 %	Hu#2 % Avg ± SEM	H3.1 %	H3.2 %	H3.3 %	Hu#3 % Avg ± SEM	Human % Avg ± SEM
SEMG2	22.7	18.1	17.1	19.3 ± 1.72	24.12	25.1	23.2	24.1 ± 0.54	17.4	16.3	18.6	17.4 ± 0.66	20.29 ± 2.00
SEMG1	6.82	7.25	10.3	8.11 ± 1.09	25.75	27.5	28.6	27.3 ± 0.83	5.39	5.65	7.64	6.23 ± 0.71	13.88 ± 6.73
PIP	4.55	3.26	9.35	5.72 ± 1.85	12.47	14.0	12.7	13.0 ± 0.48	14.1	17.7	18.2	16.7 ± 1.28	11.80 ± 3.22
LTF	7.27	6.16	4.67	6.04 ± 0.75	5.69	3.84	5.39	4.97 ± 0.57	12.0	10.6	9.82	10.8 ± 0.65	7.27 ± 1.80
ALB	7.27	7.25	7.17	7.23 ± 0.03	4.07	4.51	4.92	4.50 ± 0.25	5.39	5.65	4.73	5.26 ± 0.28	5.66 ± 0.81
ACPP	5.91	5.80	5.61	5.77 ± 0.09	2.44	2.71	2.58	2.57 ± 0.08	4.15	3.89	4.00	4.01 ± 0.08	4.12 ± 0.92
FN1	4.09	4.71	3.74	4.18 ± 0.28	1.36	0.45	0.94	0.91 ± 0.26	2.49	2.47	2.91	2.62 ± 0.14	2.57 ± 0.94
MUC6	4.09	3.26	2.49	3.28 ± 0.46	1.63	1.13	1.17	1.31 ± 0.16	2.90	3.18	2.91	3.00 ± 0.09	2.53 ± 0.62
AZGP1	3.64	2.54	2.49	2.89 ± 0.37	1.90	1.13	1.41	1.48 ± 0.22	4.15	2.12	2.55	2.94 ± 0.62	2.43 ± 0.48
CLU	0.91	3.26	1.56	1.91 ± 0.70	1.63	1.81	1.87	1.77 ± 0.07	4.15	3.18	2.91	3.41 ± 0.38	2.36 ± 0.53
KLK3	2.73	1.81	2.18	2.24 ± 0.27	1.63	1.35	1.64	1.54 ± 0.09	2.49	2.47	2.55	2.50 ± 0.02	2.09 ± 0.29
LGALS3BP	3.18	2.54	2.49	2.74 ± 0.22	0.81	0.90	1.17	0.96 ± 0.11	2.07	2.12	2.18	2.13 ± 0.03	1.94 ± 0.52
SERPINA5	2.73	2.17	3.12	2.67 ± 0.27	1.36	1.58	1.17	1.37 ± 0.12	2.90	1.41	0.73	1.68 ± 0.64	1.91 ± 0.39
TF	2.73	2.54	1.87	2.38 ± 0.26	2.71	1.58	1.64	1.98 ± 0.37	1.66	0.00	0.00	0.55 ± 0.55	1.64 ± 0.55
PSAP	0.45	3.62	1.56	1.88 ± 0.93	1.63	1.58	0.70	1.30 ± 0.30	1.66	1.41	1.09	1.39 ± 0.16	1.52 ± 0.18
NPC2	2.27	1.81	1.25	1.78 ± 0.30	0.54	0.45	0.70	0.57 ± 0.07	1.66	1.77	1.82	1.75 ± 0.05	1.36 ± 0.40
SERPINA1	0.91	0.72	1.25	0.96 ± 0.15	0.54	1.35	0.94	0.94 ± 0.23	0.83	1.06	1.82	1.24 ± 0.30	1.05 ± 0.09
CRISP1	0.00	1.09	0.93	0.67 ± 0.34	0.81	0.68	0.70	0.73 ± 0.04	1.24	2.12	1.09	1.49 ± 0.32	0.96 ± 0.26
KRT2	1.82	1.81	1.87	1.83 ± 0.02	0.00	0.23	0.23	0.15 ± 0.08	0.00	0.71	0.36	0.36 ± 0.20	0.78 ± 0.53
ANPEP	1.82	0.72	0.62	1.06 ± 0.38	0.00	0.00	0.23	0.08 ± 0.08	1.66	0.71	0.73	1.03 ± 0.31	0.72 ± 0.32
ECM1	0.00	1.09	0.93	0.67 ± 0.34	0.54	0.45	0.47	0.49 ± 0.03	0.41	1.41	0.73	0.85 ± 0.29	0.67 ± 0.11
ACTA2	0.45	1.09	0.31	0.62 ± 0.24	0.54	0.45	0.47	0.49 ± 0.03	0.83	0.35	0.73	0.64 ± 0.14	0.58 ± 0.05
CST4	0.00	0.00	0.31	0.10 ± 0.10	0.54	0.45	1.17	0.72 ± 0.23	1.24	1.41	0.00	0.89 ± 0.45	0.57 ± 0.24
TIMP1	1.36	0.72	0.31	0.80 ± 0.31	0.27	0.45	0.47	0.40 ± 0.06	0.00	0.71	0.73	0.48 ± 0.24	0.56 ± 0.12
KRT1	0.45	0.36	1.25	0.69 ± 0.28	0.27	0.90	0.47	0.55 ± 0.19	0.41	0.35	0.36	0.38 ± 0.02	0.54 ± 0.09
UBA52	1.36	0.36	0.31	0.68 ± 0.34	0.00	0.00	0.00	0.00 ± 0.00	0.83	0.71	1.09	0.88 ± 0.11	0.52 ± 0.27

Protein	H1.1 %	H1.2 %	H1.3 %	Hu#1 % Avg ± SEM	H2.1 %	H2.2 %	H2.3 %	Hu#2 % Avg ± SEM	H3.1 %	H3.2 %	H3.3 %	Hu#3 % Avg ± SEM	Human % Avg ± SEM
IGHG1	1.36	0.00	1.25	0.87 ± 0.44	0.00	0.23	0.23	0.15 ± 0.08	0.83	0.00	0.36	0.40 ± 0.24	0.47 ± 0.21
MSMB	0.45	1.45	1.25	1.05 ± 0.30	0.54	0.45	0.00	0.33 ± 0.17	0.00	0.00	0.00	0.00 ± 0.00	0.46 ± 0.31
CPE	0.00	1.09	1.25	0.78 ± 0.39	0.27	0.00	0.00	0.09 ± 0.09	0.00	0.71	0.73	0.48 ± 0.24	0.45 ± 0.20
KRT10	0.91	0.72	1.56	1.06 ± 0.25	0.00	0.00	0.00	0.00 ± 0.00	0.41	0.00	0.36	0.26 ± 0.13	0.44 ± 0.32
PRDX6	0.45	1.09	0.62	0.72 ± 0.19	0.54	0.45	0.23	0.41 ± 0.09	0.41	0.00	0.00	0.14 ± 0.14	0.42 ± 0.17
LCP1	0.45	0.00	0.00	0.15 ± 0.15	1.36	0.00	0.23	0.53 ± 0.42	0.83	0.35	0.36	0.52 ± 0.16	0.40 ± 0.12
DEFA1	0.45	0.36	0.31	0.38 ± 0.04	0.00	0.23	0.23	0.15 ± 0.08	0.41	0.71	0.73	0.62 ± 0.10	0.38 ± 0.13
TGM4	0.00	0.00	0.31	0.10 ± 0.10	0.27	0.45	0.70	0.48 ± 0.13	0.83	0.71	0.00	0.51 ± 0.26	0.36 ± 0.13
HEXB	0.00	1.09	0.62	0.57 ± 0.31	0.54	0.00	0.00	0.18 ± 0.18	0.00	0.35	0.36	0.24 ± 0.12	0.33 ± 0.12
S100A8	0.45	0.00	0.00	0.15 ± 0.15	0.27	0.00	0.00	0.09 ± 0.09	0.41	1.06	0.73	0.73 ± 0.19	0.33 ± 0.21
GDF15	0.91	1.09	0.93	0.98 ± 0.06	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.33 ± 0.33
CTSD	0.00	0.72	0.00	0.24 ± 0.24	0.81	0.90	0.47	0.73 ± 0.13	0.00	0.00	0.00	0.00 ± 0.00	0.32 ± 0.21
PGC	0.91	0.00	0.00	0.30 ± 0.30	0.81	0.00	0.23	0.35 ± 0.24	0.41	0.00	0.00	0.14 ± 0.14	0.26 ± 0.06
TMPRSS2	0.00	0.00	0.31	0.10 ± 0.10	0.00	0.00	0.00	0.00 ± 0.00	0.83	0.35	0.73	0.64 ± 0.14	0.25 ± 0.20
WFDC2	0.45	0.72	0.00	0.39 ± 0.21	0.00	0.23	0.00	0.08 ± 0.08	0.00	0.35	0.36	0.24 ± 0.12	0.24 ± 0.09
B2M	0.00	0.72	0.62	0.45 ± 0.23	0.00	0.23	0.23	0.15 ± 0.08	0.00	0.35	0.00	0.12 ± 0.12	0.24 ± 0.11
LAMP2	0.45	0.72	0.31	0.50 ± 0.12	0.27	0.00	0.00	0.09 ± 0.09	0.00	0.00	0.36	0.12 ± 0.12	0.24 ± 0.13
HSPA1L	0.00	0.36	0.31	0.22 ± 0.11	0.00	0.23	0.00	0.08 ± 0.08	0.41	0.35	0.36	0.38 ± 0.02	0.23 ± 0.09
CKB	0.91	0.72	0.00	0.54 ± 0.28	0.00	0.23	0.23	0.15 ± 0.08	0.00	0.00	0.00	0.00 ± 0.00	0.23 ± 0.16
PATE1	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.83	1.06	0.00	0.63 ± 0.32	0.21 ± 0.21
HSPA5	0.00	0.36	0.31	0.22 ± 0.11	0.00	0.23	0.47	0.23 ± 0.14	0.41	0.00	0.00	0.14 ± 0.14	0.20 ± 0.03
TTN	0.00	1.09	0.31	0.47 ± 0.32	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.36	0.12 ± 0.12	0.20 ± 0.14
SORD	0.00	0.36	0.31	0.22 ± 0.11	0.00	0.23	0.47	0.23 ± 0.14	0.00	0.35	0.00	0.12 ± 0.12	0.19 ± 0.04
CAP1	0.00	0.36	0.62	0.33 ± 0.18	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.73	0.24 ± 0.24	0.19 ± 0.10
CST3	0.00	0.00	0.00	0.00 ± 0.00	0.27	0.00	0.00	0.09 ± 0.09	0.00	0.71	0.73	0.48 ± 0.24	0.19 ± 0.15
SERPINF2	0.00	0.00	0.00	0.00 ± 0.00	0.27	0.00	0.00	0.09 ± 0.09	0.00	0.71	0.73	0.48 ± 0.24	0.19 ± 0.15

Protein	H1.1 %	H1.2 %	H1.3 %	Hu#1 % Avg ± SEM	H2.1 %	H2.2 %	H2.3 %	Hu#2 % Avg ± SEM	H3.1 %	H3.2 %	H3.3 %	Hu#3 % Avg ± SEM	Human % Avg ± SEM
YWHAB	0.45	0.36	0.93	0.58 ± 0.18	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.19 ± 0.19
IDH1	0.00	0.36	0.31	0.22 ± 0.11	0.00	0.00	0.23	0.08 ± 0.08	0.00	0.71	0.00	0.24 ± 0.24	0.18 ± 0.05
QSOX1	0.45	0.00	0.00	0.15 ± 0.15	0.00	0.90	0.23	0.38 ± 0.27	0.00	0.00	0.00	0.00 ± 0.00	0.18 ± 0.11
FOLH1	0.45	0.00	0.93	0.46 ± 0.27	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.15 ± 0.15
ORM1	0.91	0.00	0.31	0.41 ± 0.27	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.14 ± 0.14
SLC9B2	0.00	0.00	0.31	0.10 ± 0.10	0.00	0.00	0.00	0.00 ± 0.00	0.83	0.00	0.00	0.28 ± 0.28	0.13 ± 0.08
CST2	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.23	0.23	0.15 ± 0.08	0.00	0.35	0.00	0.12 ± 0.12	0.09 ± 0.05
BIRC6	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.73	0.24 ± 0.24	0.08 ± 0.08
CLIP2	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.71	0.00	0.24 ± 0.24	0.08 ± 0.08
GOLGB1	0.00	0.72	0.00	0.24 ± 0.24	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.08 ± 0.08
SERPINA3	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.73	0.24 ± 0.24	0.08 ± 0.08
SMG1	0.00	0.72	0.00	0.24 ± 0.24	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.08 ± 0.08
WTIP	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.35	0.36	0.24 ± 0.12	0.08 ± 0.08
ARHGAP31	0.00	0.36	0.00	0.12 ± 0.12	0.27	0.00	0.00	0.09 ± 0.09	0.00	0.00	0.00	0.00 ± 0.00	0.07 ± 0.04
DUT	0.00	0.00	0.00	0.00 ± 0.00	0.27	0.00	0.00	0.09 ± 0.09	0.00	0.35	0.00	0.12 ± 0.12	0.07 ± 0.04
SOD3	0.00	0.36	0.31	0.22 ± 0.11	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.07 ± 0.07
ZNF148	0.00	0.00	0.62	0.21 ± 0.21	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.07 ± 0.07
DHX57	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.47	0.16 ± 0.16	0.00	0.00	0.00	0.00 ± 0.00	0.05 ± 0.05
LRPPRC	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.23	0.23	0.15 ± 0.08	0.00	0.00	0.00	0.00 ± 0.00	0.05 ± 0.05

Table A.7: Raw normalized abundances of all proteins identified in each chimpanzee individual (in triplicate) arranged by prevalence

Protein	C1.1 %	C1.2 %	C1.3 %	Ch#1 % Avg \pm SEM	C2.1 %	C2.2 %	C2.3 %	Ch#2 % Avg \pm SEM	C3.1 %	C3.2 %	C3.3 %	Ch#3 % Avg \pm SEM	Chimp % Avg \pm SEM
SEMG1	46.81	43.4	42.7	44.3 \pm 1.28	37.2	38.8	39.9	38.7 \pm 0.78	51.6	44.8	47.4	47.9 \pm 1.98	43.62 \pm 2.70
ALB	7.45	11.0	8.82	9.10 \pm 1.04	8.35	10.4	9.85	9.53 \pm 0.60	5.96	4.85	5.56	5.46 \pm 0.33	8.03 \pm 1.29
FN1	5.32	8.54	6.25	6.70 \pm 0.96	4.81	3.46	3.20	3.82 \pm 0.50	5.61	4.48	5.56	5.22 \pm 0.37	5.25 \pm 0.83
CLU	6.03	6.05	6.62	6.23 \pm 0.19	3.54	2.39	2.96	2.96 \pm 0.33	3.86	5.60	4.07	4.51 \pm 0.55	4.57 \pm 0.94
LTF	8.87	6.76	9.56	8.40 \pm 0.84	1.77	1.60	1.48	1.62 \pm 0.09	1.75	2.99	1.11	1.95 \pm 0.55	3.99 \pm 2.21
SERPINA5	2.13	2.85	1.84	2.27 \pm 0.30	3.54	3.19	3.45	3.39 \pm 0.11	4.91	5.22	4.07	4.74 \pm 0.34	3.47 \pm 0.71
ACPP	0.71	1.07	1.10	0.96 \pm 0.13	3.29	2.39	1.72	2.47 \pm 0.45	2.46	4.48	4.07	3.67 \pm 0.62	2.37 \pm 0.78
NPC2	1.06	1.42	2.57	1.69 \pm 0.46	3.29	2.39	1.97	2.55 \pm 0.39	1.05	1.87	2.22	1.71 \pm 0.35	1.98 \pm 0.28
SERPINA1	0.71	0.71	1.47	0.96 \pm 0.25	2.03	2.13	2.22	2.12 \pm 0.06	2.81	1.87	2.96	2.55 \pm 0.34	1.88 \pm 0.47
TGM4	0.35	1.07	0.00	0.47 \pm 0.31	4.05	2.93	3.45	3.47 \pm 0.33	1.75	1.49	1.85	1.70 \pm 0.11	1.88 \pm 0.87
SERPINA3	1.77	0.00	0.74	0.84 \pm 0.51	2.28	2.39	2.46	2.38 \pm 0.05	2.46	2.61	1.85	2.31 \pm 0.23	1.84 \pm 0.50
CRTAC1	1.77	1.07	1.47	1.44 \pm 0.20	1.77	1.86	2.46	2.03 \pm 0.22	1.75	1.49	2.22	1.82 \pm 0.21	1.76 \pm 0.17
SEMG2	1.06	2.85	2.94	2.28 \pm 0.61	0.76	0.80	0.74	0.77 \pm 0.02	0.70	1.49	1.85	1.35 \pm 0.34	1.47 \pm 0.44
IGHG1	1.77	1.07	2.21	1.68 \pm 0.33	1.01	2.13	1.48	1.54 \pm 0.32	1.05	1.49	0.74	1.10 \pm 0.22	1.44 \pm 0.18
AMY2B	0.71	0.71	1.10	0.84 \pm 0.13	0.51	1.33	1.72	1.19 \pm 0.36	0.35	2.24	1.48	1.36 \pm 0.55	1.13 \pm 0.15
ECM1	1.42	0.36	1.10	0.96 \pm 0.32	0.76	1.06	0.74	0.85 \pm 0.11	0.70	0.37	2.22	1.10 \pm 0.57	0.97 \pm 0.07
TF	1.77	2.14	1.47	1.79 \pm 0.19	1.01	0.53	1.23	0.93 \pm 0.21	0.00	0.37	0.00	0.12 \pm 0.12	0.95 \pm 0.48
AZGP1	0.35	0.36	0.00	0.24 \pm 0.12	0.51	1.06	1.23	0.93 \pm 0.22	1.05	2.61	1.11	1.59 \pm 0.51	0.92 \pm 0.39
RNASET2	1.06	1.07	0.37	0.83 \pm 0.23	0.76	1.06	0.74	0.85 \pm 0.11	1.05	0.75	0.74	0.85 \pm 0.10	0.84 \pm 0.01
IGKC	0.71	1.07	1.84	1.21 \pm 0.33	0.25	1.33	1.23	0.94 \pm 0.34	0.70	0.00	0.37	0.36 \pm 0.20	0.83 \pm 0.25
ACTB	0.35	1.07	1.47	0.96 \pm 0.33	0.51	0.53	0.74	0.59 \pm 0.07	0.00	1.49	0.00	0.50 \pm 0.50	0.68 \pm 0.14
PSCA	0.35	0.36	0.37	0.36 \pm 0.00	1.27	0.27	0.74	0.76 \pm 0.29	0.35	1.12	1.11	0.86 \pm 0.25	0.66 \pm 0.15
ACRBP	1.06	0.36	0.00	0.47 \pm 0.31	0.51	0.80	1.23	0.85 \pm 0.21	0.70	0.75	0.00	0.48 \pm 0.24	0.60 \pm 0.12
DPEP3	0.35	0.71	0.74	0.60 \pm 0.12	0.51	0.80	0.49	0.60 \pm 0.10	0.35	0.75	0.37	0.49 \pm 0.13	0.56 \pm 0.04
PAEP	0.00	0.00	0.00	0.00 \pm 0.00	1.01	0.53	0.49	0.68 \pm 0.17	1.75	0.75	0.37	0.96 \pm 0.41	0.55 \pm 0.28
CST3	0.71	0.00	0.37	0.36 \pm 0.20	1.01	0.53	0.25	0.60 \pm 0.22	0.70	0.37	0.74	0.61 \pm 0.12	0.52 \pm 0.08

Protein	C1.1 %	C1.2 %	C1.3 %	Ch#1 % Avg ± SEM	C2.1 %	C2.2 %	C2.3 %	Ch#2 % Avg ± SEM	C3.1 %	C3.2 %	C3.3 %	Ch#3 % Avg ± SEM	Chimp % Avg ± SEM
KRT1	0.35	0.00	0.00	0.12 ± 0.12	1.52	1.33	0.25	1.03 ± 0.40	0.70	0.00	0.37	0.36 ± 0.20	0.50 ± 0.27
PKM	0.71	0.71	0.74	0.72 ± 0.01	0.76	0.80	0.74	0.77 ± 0.02	0.00	0.00	0.00	0.00 ± 0.00	0.49 ± 0.25
PPIB	0.00	0.00	0.37	0.12 ± 0.12	0.25	0.53	0.49	0.43 ± 0.09	0.35	1.87	0.37	0.86 ± 0.50	0.47 ± 0.21
ACE	0.00	0.00	1.10	0.37 ± 0.37	1.27	0.80	0.74	0.93 ± 0.17	0.00	0.00	0.00	0.00 ± 0.00	0.43 ± 0.27
PRDX6	0.71	0.71	0.37	0.60 ± 0.11	0.51	0.27	0.25	0.34 ± 0.08	0.35	0.37	0.00	0.24 ± 0.12	0.39 ± 0.11
LDHC	0.00	0.00	0.00	0.00 ± 0.00	1.01	1.06	1.48	1.18 ± 0.15	0.00	0.00	0.00	0.00 ± 0.00	0.39 ± 0.39
HK1	0.71	0.00	0.00	0.24 ± 0.24	1.01	0.53	0.49	0.68 ± 0.17	0.35	0.00	0.00	0.12 ± 0.12	0.34 ± 0.17
MDH2	0.00	0.00	0.00	0.00 ± 0.00	1.52	0.53	0.99	1.01 ± 0.29	0.00	0.00	0.00	0.00 ± 0.00	0.34 ± 0.34
GPI	0.00	0.00	0.37	0.12 ± 0.12	0.51	0.27	0.00	0.26 ± 0.15	0.00	0.00	1.48	0.49 ± 0.49	0.29 ± 0.11
LCP1	0.35	0.00	0.00	0.12 ± 0.12	0.51	0.27	0.49	0.42 ± 0.08	0.00	0.00	0.37	0.12 ± 0.12	0.22 ± 0.10
TUBA1A	0.00	0.00	0.00	0.00 ± 0.00	0.51	0.27	0.49	0.42 ± 0.08	0.00	0.37	0.37	0.25 ± 0.12	0.22 ± 0.12
B4GALT1	0.71	0.00	0.00	0.24 ± 0.24	0.00	0.00	0.00	0.00 ± 0.00	0.35	0.00	0.74	0.36 ± 0.21	0.20 ± 0.11
MGAM	0.35	0.71	0.00	0.36 ± 0.21	0.00	0.53	0.00	0.18 ± 0.18	0.00	0.00	0.00	0.00 ± 0.00	0.18 ± 0.10
CRISP1	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.27	0.49	0.25 ± 0.14	0.00	0.37	0.37	0.25 ± 0.12	0.17 ± 0.08
ORM1	0.00	0.00	0.00	0.00 ± 0.00	0.76	0.27	0.49	0.51 ± 0.14	0.00	0.00	0.00	0.00 ± 0.00	0.17 ± 0.17
ACR	0.00	0.36	0.00	0.12 ± 0.12	0.00	0.27	0.00	0.09 ± 0.09	0.00	0.00	0.74	0.25 ± 0.25	0.15 ± 0.05
ACOT13	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.53	0.49	0.34 ± 0.17	0.00	0.00	0.37	0.12 ± 0.12	0.15 ± 0.10
PGK2	0.35	0.71	0.00	0.36 ± 0.21	0.25	0.00	0.00	0.08 ± 0.08	0.00	0.00	0.00	0.00 ± 0.00	0.15 ± 0.11
PLA1A	0.71	0.36	0.00	0.36 ± 0.20	0.00	0.00	0.25	0.08 ± 0.08	0.00	0.00	0.00	0.00 ± 0.00	0.15 ± 0.11
HSPA1A	0.00	0.36	0.00	0.12 ± 0.12	0.25	0.27	0.00	0.17 ± 0.09	0.35	0.00	0.00	0.12 ± 0.12	0.14 ± 0.02
HEXB	0.00	0.00	0.00	0.00 ± 0.00	0.25	0.53	0.49	0.43 ± 0.09	0.00	0.00	0.00	0.00 ± 0.00	0.14 ± 0.14
OXCT2	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.53	0.25	0.26 ± 0.15	0.00	0.37	0.00	0.12 ± 0.12	0.13 ± 0.07
PLS3	0.35	0.00	0.00	0.12 ± 0.12	0.25	0.00	0.49	0.25 ± 0.14	0.00	0.00	0.00	0.00 ± 0.00	0.12 ± 0.07
ANPEP	0.00	0.00	0.00	0.00 ± 0.00	0.51	0.53	0.00	0.35 ± 0.17	0.00	0.00	0.00	0.00 ± 0.00	0.12 ± 0.12
KRT10	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.80	0.25	0.35 ± 0.24	0.00	0.00	0.00	0.00 ± 0.00	0.12 ± 0.12
B2M	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.27	0.00	0.09 ± 0.09	0.00	0.00	0.74	0.25 ± 0.25	0.11 ± 0.07

Protein	C1.1 %	C1.2 %	C1.3 %	Ch#1 % Avg ± SEM	C2.1 %	C2.2 %	C2.3 %	Ch#2 % Avg ± SEM	C3.1 %	C3.2 %	C3.3 %	Ch#3 % Avg ± SEM	Chimp % Avg ± SEM
HSP90AA1	0.00	0.00	0.00	0.00 ± 0.00	0.25	0.27	0.49	0.34 ± 0.08	0.00	0.00	0.00	0.00 ± 0.00	0.11 ± 0.11
KRT2	0.00	0.00	0.00	0.00 ± 0.00	0.25	0.27	0.49	0.34 ± 0.08	0.00	0.00	0.00	0.00 ± 0.00	0.11 ± 0.11
TUBB	0.00	0.00	0.00	0.00 ± 0.00	0.25	0.53	0.25	0.34 ± 0.09	0.00	0.00	0.00	0.00 ± 0.00	0.11 ± 0.11
PGAM2	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.70	0.00	0.00	0.23 ± 0.23	0.08 ± 0.08
RLTPR	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.70	0.00	0.00	0.23 ± 0.23	0.08 ± 0.08
ENO1	0.00	0.00	0.00	0.00 ± 0.00	0.51	0.00	0.25	0.25 ± 0.15	0.00	0.00	0.00	0.00 ± 0.00	0.08 ± 0.08
TTN	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.27	0.00	0.09 ± 0.09	0.35	0.00	0.00	0.12 ± 0.12	0.07 ± 0.04
ACO2	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.25	0.08 ± 0.08	0.00	0.37	0.00	0.12 ± 0.12	0.07 ± 0.04
SOD2	0.00	0.00	0.00	0.00 ± 0.00	0.25	0.00	0.00	0.08 ± 0.08	0.35	0.00	0.00	0.12 ± 0.12	0.07 ± 0.03
FAM3D	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.27	0.25	0.17 ± 0.09	0.00	0.00	0.00	0.00 ± 0.00	0.06 ± 0.06
MME	0.00	0.00	0.00	0.00 ± 0.00	0.25	0.00	0.25	0.17 ± 0.08	0.00	0.00	0.00	0.00 ± 0.00	0.06 ± 0.06
TKFC	0.00	0.00	0.00	0.00 ± 0.00	0.25	0.27	0.00	0.17 ± 0.09	0.00	0.00	0.00	0.00 ± 0.00	0.06 ± 0.06

Table A.8: Raw normalized abundances of all proteins identified in each gorilla individual (in triplicate) arranged by prevalence

Protein	G1.1 %	G1.2 %	G1.3 %	Go#1 % Avg ± SEM	G2.1 %	G2.2 %	G2.3 %	Go#2 % Avg ± SEM	G3.1 %	G3.2 %	G3.3 %	Go#3 % Avg ± SEM	Gorilla % Avg ± SEM
ALB	50.00	37.1	36.7	41.4 ± 4.37	27.3	38.7	22.6	29.5 ± 4.79	9.72	1.89	9.23	6.95 ± 2.53	25.9 ± 10.1
PAEP	7.14	2.86	6.67	5.56 ± 1.36	13.6	6.45	6.45	8.85 ± 2.39	9.72	34.0	9.23	17.6 ± 8.16	10.7 ± 3.61
SEMG2	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	52.8	13.2	29.2	31.7 ± 11.5	10.6 ± 10.6
CLU	7.14	8.57	6.67	7.46 ± 0.57	9.09	3.23	6.45	6.26 ± 1.70	5.56	9.43	7.69	7.56 ± 1.12	7.09 ± 0.42
CCDC88C	0.00	5.71	6.67	4.13 ± 2.08	9.09	16.1	16.1	13.8 ± 2.35	0.00	0.00	1.54	0.51 ± 0.51	6.14 ± 3.96
LTF	7.14	8.57	3.33	6.35 ± 1.56	4.55	3.23	6.45	4.74 ± 0.94	0.00	1.89	3.08	1.65 ± 0.90	4.25 ± 1.38
KRT1	14.29	0.00	13.3	9.21 ± 4.61	4.55	0.00	0.00	1.52 ± 1.52	0.00	0.00	1.54	0.51 ± 0.51	3.74 ± 2.75
SEMG1	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	8.33	13.2	7.69	9.74 ± 1.74	3.25 ± 3.25
ZCCHC11	0.00	2.86	0.00	0.95 ± 0.95	4.55	6.45	9.68	6.89 ± 1.50	1.39	0.00	1.54	0.98 ± 0.49	2.94 ± 1.98
APOB	7.14	5.71	0.00	4.29 ± 2.18	0.00	3.23	6.45	3.23 ± 1.86	0.00	1.89	0.00	0.63 ± 0.63	2.71 ± 1.09
SERPINA1	0.00	0.00	0.00	0.00 ± 0.00	4.55	6.45	3.23	4.74 ± 0.94	1.39	3.77	3.08	2.75 ± 0.71	2.50 ± 1.37
ARNT	0.00	8.57	0.00	2.86 ± 2.86	4.55	0.00	3.23	2.59 ± 1.35	0.00	0.00	0.00	0.00 ± 0.00	1.82 ± 0.91
NDUFS1	0.00	5.71	3.33	3.02 ± 1.66	4.55	0.00	0.00	1.52 ± 1.52	0.00	0.00	0.00	0.00 ± 0.00	1.51 ± 0.87
PIP	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	4.17	3.77	4.62	4.19 ± 0.24	1.40 ± 1.40
CEP162	0.00	0.00	3.33	1.11 ± 1.11	9.09	0.00	0.00	3.03 ± 3.03	0.00	0.00	0.00	0.00 ± 0.00	1.38 ± 0.89
TSSC1	0.00	0.00	0.00	0.00 ± 0.00	0.00	3.23	6.45	3.23 ± 1.86	0.00	1.89	0.00	0.63 ± 0.63	1.28 ± 0.99
TTLL5	7.14	2.86	0.00	3.33 ± 2.08	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	1.11 ± 1.11
BCL3	0.00	0.00	3.33	1.11 ± 1.11	0.00	6.45	0.00	2.15 ± 2.15	0.00	0.00	0.00	0.00 ± 0.00	1.09 ± 0.62
FAM184B	0.00	2.86	0.00	0.95 ± 0.95	0.00	3.23	0.00	1.08 ± 1.08	1.39	1.89	0.00	1.09 ± 0.56	1.04 ± 0.04
PPIB	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	2.78	3.77	1.54	2.70 ± 0.65	0.90 ± 0.90
NBPF3	0.00	5.71	0.00	1.90 ± 1.90	0.00	0.00	0.00	0.00 ± 0.00	1.39	0.00	0.00	0.46 ± 0.46	0.79 ± 0.57
CPSF1	0.00	0.00	6.67	2.22 ± 2.22	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.74 ± 0.74
UBA1	0.00	0.00	3.33	1.11 ± 1.11	0.00	3.23	0.00	1.08 ± 1.08	0.00	0.00	0.00	0.00 ± 0.00	0.73 ± 0.36
LRRK1	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	6.45	2.15 ± 2.15	0.00	0.00	0.00	0.00 ± 0.00	0.72 ± 0.72
TTN	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	3.23	1.08 ± 1.08	0.00	0.00	3.08	1.03 ± 1.03	0.70 ± 0.35
CFAP69	0.00	0.00	0.00	0.00 ± 0.00	4.55	0.00	0.00	1.52 ± 1.52	0.00	0.00	1.54	0.51 ± 0.51	0.68 ± 0.44

Protein	G1.1 %	G1.2 %	G1.3 %	Go#1 % Avg ± SEM	G2.1 %	G2.2 %	G2.3 %	Go#2 % Avg ± SEM	G3.1 %	G3.2 %	G3.3 %	Go#3 % Avg ± SEM	Gorilla % Avg ± SEM
CLEC11A	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.00	3.77	1.54	1.77 ± 1.10	0.59 ± 0.59
ASAP2	0.00	0.00	3.33	1.11 ± 1.11	0.00	0.00	0.00	0.00 ± 0.00	0.00	1.89	0.00	0.63 ± 0.63	0.58 ± 0.32
SRCAP	0.00	0.00	3.33	1.11 ± 1.11	0.00	0.00	0.00	0.00 ± 0.00	1.39	0.00	0.00	0.46 ± 0.46	0.52 ± 0.32
TRRAP	0.00	2.86	0.00	0.95 ± 0.95	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	1.54	0.51 ± 0.51	0.49 ± 0.28
DNAH7	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.00	3.77	0.00	1.26 ± 1.26	0.42 ± 0.42
FAM183B	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	3.23	1.08 ± 1.08	0.00	0.00	0.00	0.00 ± 0.00	0.36 ± 0.36
CTSB	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	3.08	1.03 ± 1.03	0.34 ± 0.34
KRT10	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	3.08	1.03 ± 1.03	0.34 ± 0.34
ZFHX2	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	0.00	0.00 ± 0.00	0.00	0.00	3.08	1.03 ± 1.03	0.34 ± 0.34

Table A.9: Summary of significant differences between species

Protein^a	Statistical Test^b	Human – Chimp^c	Human – Gorilla^d	Chimp – Gorilla^e
ACE	PLGEM			C > G (R)
	t test			
	2x > 0.01			
ACOT13	PLGEM	C > H (L)		
	t test			
	2x > 0.01			
ACPP	PLGEM		H > G	C > G
	t test			
	2x > 0.01		H > G	C > G
ACRBP	PLGEM	C > H		C > G
	t test	C > H		C > G
	2x > 0.01			
ACTA2	PLGEM	H > C	H > G	
	t test	H > C	H > G	
	2x > 0.01			
ACTB	PLGEM	C > H		C > G
	t test	C > H		C > G
	2x > 0.01			
ALB	PLGEM		G > H	
	t test			
	2x > 0.01	C > H	G > H	G > C
ANPEP	PLGEM		H > G (R)	
	t test			
	2x > 0.01			
AMY2B	PLGEM	C > H		C > G
	t test	C > H		C > G
	2x > 0.01	C > H		C > G
APOB	PLGEM		G > H (R)	G > C (R)
	t test			
	2x > 0.01			G > C (R)
ARNT	PLGEM		G > H	G > C
	t test			
	2x > 0.01		G > H	G > C
ASAP2	PLGEM		G > H (R)	G > C (R)
	t test			
	2x > 0.01			
AZGP1	PLGEM		H > G	C > G
	t test		H > G	
	2x > 0.01	H > C (R)	H > G	C > G (L)

Protein ^a	Statistical Test ^b	Human – Chimp ^c	Human – Gorilla ^d	Chimp – Gorilla ^e
B2M	PLGEM		H > G (L)	
	t test			
	2x > 0.01			
BCL3	PLGEM		G > H	G > C
	t test			
	2x > 0.01		G > H	G > C
CCDC88C	PLGEM		G > H	G > C
	t test			
	2x > 0.01		G > H	G > C
CEP162	PLGEM		G > H	G > C
	t test			
	2x > 0.01			G > C (R)
CFAP69	PLGEM		G > H (R)	G > C (R)
	t test			
	2x > 0.01			
CLEC11A	PLGEM		G > H	G > C
	t test			
	2x > 0.01			
CLU	PLGEM		G > H (R)	
	t test		G > H	
	2x > 0.01		G > H	
CPSF1	PLGEM		G > H (R)	G > C (R)
	t test			
	2x > 0.01			
CRISP1	PLGEM	H > C (R)	H > G	
	t test		H > G	
	2x > 0.01		H > G	
CRTAC1	PLGEM	C > H		C > G
	t test	C > H		C > G
	2x > 0.01	C > H		C > G (R)
CST3	PLGEM		H > G (L)	C > G
	t test			C > G
	2x > 0.01	C > H (L)		C > G (L)
CST4	PLGEM	H > C	H > G (L)	
	t test			
	2x > 0.01	H > C	H > G	
DEFA1	PLGEM	H > C (L)	H > G (L)	
	t test			
	2x > 0.01	H > C	H > G	

Protein ^a	Statistical Test ^b	Human – Chimp ^c	Human – Gorilla ^d	Chimp – Gorilla ^e
DNAH7	PLGEM			G > C (R)
	t test			
	2x > 0.01			
DPEP3	PLGEM	C > H		C > G
	t test	C > H		C > G
	2x > 0.01			
ECM1	PLGEM		H > G (R)	C > G
	t test	H > C (L)	H > G	C > G
	2x > 0.01			
FAM184B	PLGEM		G > H	G > C
	t test			G > C (R)
	2x > 0.01			G > C (R)
FN1	PLGEM		H > G (R)	C > G
	t test			C > G
	2x > 0.01	C > H		C > G
HSPA1A	PLGEM			
	t test	C > H		C > G
	2x > 0.01			
HSPA5	PLGEM			
	t test	H > C		
	2x > 0.01			
IGHG1	PLGEM		H > G (L)	C > G
	t test	C > H (L)		C > G
	2x > 0.01	C > H		C > G
IGKC	PLGEM	C > H		C > G
	t test			
	2x > 0.01	C > H (L)		C > G
KLK3	PLGEM	H > C	H > G	
	t test	H > C	H > G	
	2x > 0.01	H > C	H > G	
KRT1	PLGEM		G > H (R)	G > C (R)
	t test			
	2x > 0.01		G > H	G > C
KRT2	PLGEM		H > G (R)	
	t test			
	2x > 0.01			
LDHC	PLGEM	C > H		C > G (L)
	t test			
	2x > 0.01			

Protein ^a	Statistical Test ^b	Human – Chimp ^c	Human – Gorilla ^d	Chimp – Gorilla ^e
LGALS3BP	PLGEM	H > C	H > G	
	t test			
	2x > 0.01	H > C	H > G	
LRRK1	PLGEM	G > H (R)		G > C (R)
	t test			
	2x > 0.01			
MDH2	PLGEM	C > H (L)		
	t test			
	2x > 0.01			
MSMB	PLGEM	H > C (L)	H > G	
	t test			
	2x > 0.01	H > C	H > G	
MUC6	PLGEM	H > C (R)	H > G (R)	
	t test	H > C		
	2x > 0.01	H > C (R)		
NBPF3	PLGEM		G > H	G > C
	t test			
	2x > 0.01			
NDUF51	PLGEM		G > H	G > C
	t test			
	2x > 0.01		G > H	G > C
NPC2	PLGEM		H > G	C > G
	t test			C > G
	2x > 0.01	C > H (L)	H > G	C > G
PAEP	PLGEM	C > H	G > H	G > C
	t test		G > H	G > C (L)
	2x > 0.01	C > H (L)	G > H	G > C
PATE1	PLGEM	H > C	H > G (L)	
	t test			
	2x > 0.01			
PGC	PLGEM			
	t test	H > C (R)		
	2x > 0.01			
PIP	PLGEM	H > C	H > G	G > C
	t test			
	2x > 0.01	H > C	H > G	G > C
PKM	PLGEM	C > H (R)		C > G (R)
	t test			
	2x > 0.01			

Protein^a	Statistical Test^b	Human – Chimp^c	Human – Gorilla^d	Chimp – Gorilla^e
PPIB	PLGEM	C > H (L)	G > H	
	t test			
	2x > 0.01	C > H (L)	G > H	
PRDX6	PLGEM		H > G (L)	C > G (L)
	t test			
	2x > 0.01			
PSAP	PLGEM	H > C	H > G	
	t test	H > C	H > G	
	2x > 0.01	H > C	H > G	
PSCA	PLGEM	C > H		C > G
	t test	C > H		C > G
	2x > 0.01	C > H (L)		C > G (L)
RNASET2	PLGEM	C > H		C > G
	t test	C > H		C > G
	2x > 0.01	C > H		C > G (L)
S100A8	PLGEM	H > C (L)	H > G (L)	
	t test			
	2x > 0.01	H > C	H > G	
SEMG1	PLGEM	C > H (R)	H > G (R)	C > G
	t test	C > H (R)		C > G
	2x > 0.01	C > H	H > G	C > G
SEMG2	PLGEM	H > C	H > G (R)	G > C (R)
	t test	H > C		
	2x > 0.01	H > C		G > C
SERPINA1	PLGEM			
	t test			
	2x > 0.01	C > H (L)	G > H	
SERPINA3	PLGEM	C > H		C > G
	t test	C > H (L)		C > G (L)
	2x > 0.01	C > H		C > G
SERPINA5	PLGEM		H > G	C > G
	t test			C > G
	2x > 0.01	C > H (L)	H > G	C > G
SORD	PLGEM			
	t test	H > C		
	2x > 0.01			
SRCAP	PLGEM			G > C (R)
	t test			
	2x > 0.01			

Protein ^a	Statistical Test ^b	Human – Chimp ^c	Human – Gorilla ^d	Chimp – Gorilla ^e
TF	PLGEM		H > G	C > G
	t test			
	2x > 0.01			
TGM4	PLGEM	C > H (R)		C > G
	t test			
	2x > 0.01	C > H		C > G
TIMP1	PLGEM	H > C	H > G (L)	
	t test	H > C		
	2x > 0.01	H > C (L)		
TRRAP	PLGEM			G > C (R)
	t test			
	2x > 0.01			
TSSC1	PLGEM		G > H	G > C
	t test			
	2x > 0.01		G > H	G > C
TTLL5	PLGEM		G > H (R)	G > C (R)
	t test			
	2x > 0.01			G > C (R)
TTN	PLGEM			G > C (R)
	t test			
	2x > 0.01			
UBA1	PLGEM		G > H (R)	G > C (R)
	t test			
	2x > 0.01			
UBA52	PLGEM	H > C	H > G (L)	
	t test			
	2x > 0.01	H > C (L)	H > G	
WFDC2	PLGEM	H > C (L)	H > G (L)	
	t test			
	2x > 0.01			
ZCCHC11	PLGEM		G > H	G > C
	t test			
	2x > 0.01		G > H	G > C

a) Each protein listed was significantly higher in one species compared to another using at least one of three statistical tests. **b)** Statistical tests included Power Law Global Error Model with a 1% false discovery rate (PLGEM; Appendix A.4), traditional t-test with Benjamini-Hochberg correction with a 10% false discovery rate (t-test) and proteins that were all proteins absent in one species, plus those present at greater than two-fold abundance, but only including those found with NSAF greater than 0.01 (2x > 0.01) **c-d)** Capital letters represent species (i.e. “H” is human) and the “>” indicates that the species on the right had significantly less (or none) of that protein compared to the species on the left. Protein abundance is summarized for each species in Table 2.7. (L) indicates statistical significance was only using length-adjusted normalization; likewise (R), indicates statistical significance was only using raw-normalization. Without either (L) or (R) the significant difference was found in both normalization data sets.

A.7 Gene Ontology

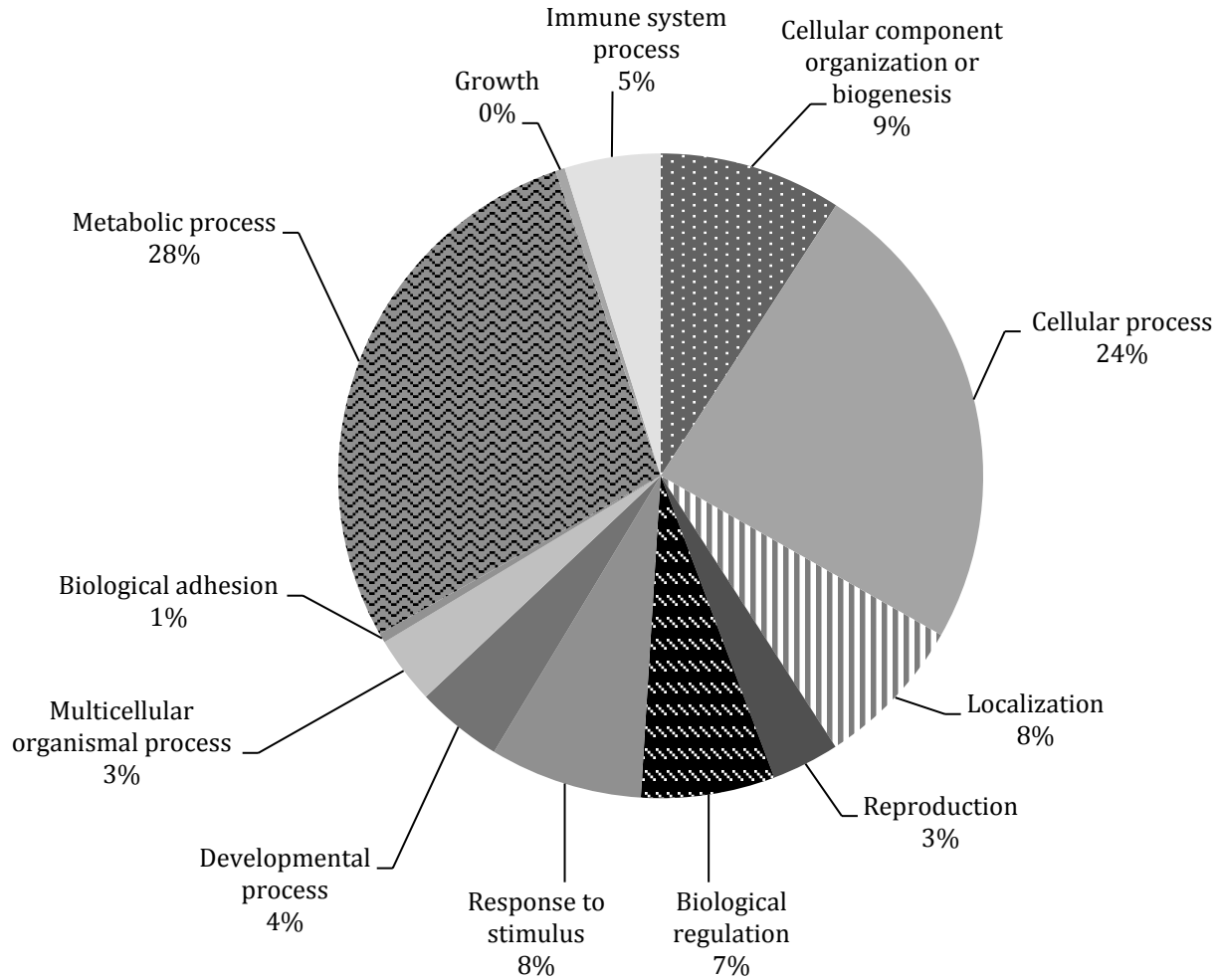


Figure A.1: Biological process functions of all identified proteins in human, chimpanzee, and gorilla seminal plasma

Seminal plasma proteins (identified across all three species) are associated with 12 biological processes identified through gene ontology (Panther). Metabolic processes (28%) and cellular processes (17%) are the largest categories.

Table A.10: Significant overexpression of GO molecular functions in human seminal plasma compared to the human reference genome

GO molecular function	#	Expected	Fold Enrichment	+/-	P value
<u>Cysteine-type endopeptidase inhibitor activity</u>	<u>6</u>	.22	27.50	+	3.27E-04
<u>Endopeptidase inhibitor activity</u>	<u>11</u>	.66	16.62	+	2.42E-07
<u>Endopeptidase regulator activity</u>	<u>11</u>	.68	16.07	+	3.43E-07
<u>Peptidase regulator activity</u>	<u>12</u>	.83	14.44	+	1.55E-07
<u>Enzyme regulator activity</u>	<u>15</u>	3.91	3.84	+	2.16E-02
<u>Peptidase inhibitor activity</u>	<u>11</u>	.70	15.81	+	4.07E-07
<u>Enzyme inhibitor activity</u>	<u>12</u>	1.53	7.84	+	1.36E-04
<u>Protease binding</u>	<u>12</u>	.50	23.99	+	4.54E-10
<u>Enzyme binding</u>	<u>25</u>	8.19	3.05	+	6.93E-04
<u>Protein binding</u>	<u>65</u>	43.13	1.51	+	6.79E-04
<u>Serine-type endopeptidase activity</u>	<u>8</u>	.96	8.34	+	1.80E-02
<u>Peptidase activity, acting on L-amino acid peptides</u>	<u>13</u>	2.66	4.88	+	7.50E-03
<u>Peptidase activity</u>	<u>13</u>	2.75	4.73	+	1.06E-02
<u>Serine-type peptidase activity</u>	<u>9</u>	1.06	8.45	+	4.06E-03
<u>Serine hydrolase activity</u>	<u>9</u>	1.08	8.31	+	4.68E-03
<u>Unclassified</u>	<u>1</u>	13.42	< 0.2	-	0.00E00

GO molecular functions are hyperlinked to their functional description on Panther's website. The “#” column identified how many human proteins (out of 71) were associated with that molecular function. In addition, these underlined numbers are hyperlinked to tables identifying which proteins matched that function.

Table A.11: Significant overexpression of GO biological processes in human seminal plasma compared to the human reference genome

GO biological process complete	#	Expected	Fold enrichment	+/-	P value
<u>Antibacterial humoral response</u>	<u>5</u>	.14	35.93	+	2.96E-03
<u>Antimicrobial humoral response</u>	<u>7</u>	.36	19.18	+	8.39E-04
<u>Response to stimulus</u>	<u>56</u>	30.13	1.86	+	3.20E-05
<u>Humoral immune response</u>	<u>9</u>	1.16	7.77	+	2.21E-02
<u>Immune response</u>	<u>28</u>	6.29	4.45	+	4.79E-08
<u>Immune system process</u>	<u>30</u>	9.60	3.13	+	3.69E-05
<u>Defense response</u>	<u>23</u>	4.62	4.98	+	7.21E-07
<u>Response to stress</u>	<u>35</u>	12.54	2.79	+	1.92E-05
<u>Acute-phase response</u>	<u>5</u>	.15	32.42	+	4.89E-03
<u>Acute inflammatory response</u>	<u>8</u>	.26	30.38	+	2.78E-06
<u>Platelet degranulation</u>	<u>15</u>	.48	31.40	+	1.71E-14
<u>Regulated exocytosis</u>	<u>29</u>	2.58	11.25	+	4.50E-19
<u>Exocytosis</u>	<u>29</u>	2.89	10.03	+	1.05E-17
<u>Secretion by cell</u>	<u>29</u>	3.61	8.04	+	4.04E-15
<u>Secretion</u>	<u>29</u>	4.01	7.24	+	6.55E-14
<u>Transport</u>	<u>41</u>	16.58	2.47	+	1.18E-05
<u>Establishment of localization</u>	<u>41</u>	16.99	2.41	+	2.53E-05
<u>Localization</u>	<u>47</u>	20.68	2.27	+	3.91E-06
<u>Vesicle-mediated transport</u>	<u>33</u>	6.71	4.92	+	1.40E-11
<u>Retina homeostasis</u>	<u>8</u>	.27	29.95	+	3.11E-06
<u>Tissue homeostasis</u>	<u>10</u>	.66	15.11	+	1.26E-05
<u>Multicellular organismal homeostasis</u>	<u>10</u>	1.08	9.23	+	1.24E-03
<u>Multicellular organismal process</u>	<u>45</u>	25.19	1.79	+	3.25E-02
<u>Regulation of biological quality</u>	<u>32</u>	13.37	2.39	+	5.33E-03
<u>Anatomical structure homeostasis</u>	<u>11</u>	1.10	10.01	+	1.23E-04
<u>Peptide cross-linking</u>	<u>5</u>	.22	22.53	+	2.86E-02
<u>Protein metabolic process</u>	<u>39</u>	16.76	2.33	+	2.50E-04

GO biological process complete	#	Expected	Fold enrichment	+/-	P value
<u>Organonitrogen compound metabolic process</u>	<u>46</u>	20.78	2.21	+	1.88E-05
<u>Negative regulation of endopeptidase activity</u>	<u>11</u>	.93	11.84	+	2.23E-05
<u>Negative regulation of peptidase activity</u>	<u>12</u>	.97	12.37	+	2.44E-06
<u>Negative regulation of hydrolase activity</u>	<u>12</u>	1.58	7.61	+	5.09E-04
<u>Negative regulation of catalytic activity</u>	<u>14</u>	3.36	4.17	+	4.90E-02
<u>Negative regulation of molecular function</u>	<u>17</u>	4.41	3.86	+	1.21E-02
<u>Regulation of molecular function</u>	<u>32</u>	12.71	2.52	+	1.67E-03
<u>Regulation of catalytic activity</u>	<u>26</u>	9.43	2.76	+	7.24E-03
<u>Regulation of hydrolase activity</u>	<u>19</u>	5.31	3.58	+	7.80E-03
<u>Regulation of peptidase activity</u>	<u>17</u>	1.58	10.79	+	2.67E-09
<u>Regulation of proteolysis</u>	<u>22</u>	3.16	6.96	+	3.24E-09
<u>Regulation of protein metabolic process</u>	<u>29</u>	10.40	2.79	+	9.57E-04
<u>Regulation of biological process</u>	<u>63</u>	42.80	1.47	+	1.67E-02
<u>Negative regulation of proteolysis</u>	<u>12</u>	1.31	9.14	+	6.98E-05
<u>Regulation of cellular protein metabolic process</u>	<u>27</u>	9.48	2.85	+	2.08E-03
<u>Negative regulation of protein metabolic process</u>	<u>16</u>	4.21	3.80	+	3.17E-02
<u>Regulation of endopeptidase activity</u>	<u>15</u>	1.47	10.17	+	1.75E-07
<u>Neutrophil degranulation</u>	<u>20</u>	1.81	11.03	+	1.02E-11
<u>Neutrophil mediated immunity</u>	<u>20</u>	1.87	10.72	+	1.75E-11
<u>Myeloid leukocyte mediated immunity</u>	<u>20</u>	1.94	10.30	+	3.65E-11
<u>Leukocyte mediated immunity</u>	<u>22</u>	2.77	7.95	+	2.34E-10
<u>Immune effector process</u>	<u>23</u>	3.86	5.96	+	2.02E-08
<u>Leukocyte degranulation</u>	<u>20</u>	1.89	10.57	+	2.27E-11
<u>Neutrophil activation involved in immune response</u>	<u>20</u>	1.82	10.99	+	1.11E-11
<u>Myeloid cell activation involved in immune response</u>	<u>20</u>	1.93	10.36	+	3.27E-11
<u>Leukocyte activation involved in immune response</u>	<u>21</u>	2.28	9.20	+	6.09E-11
<u>Leukocyte activation</u>	<u>22</u>	3.31	6.65	+	8.10E-09
<u>Cell activation</u>	<u>24</u>	3.86	6.22	+	2.40E-09
<u>Cell activation involved in immune response</u>	<u>21</u>	2.30	9.14	+	6.91E-11
<u>Myeloid leukocyte activation</u>	<u>21</u>	2.13	9.86	+	1.57E-11
<u>Neutrophil activation</u>	<u>20</u>	1.85	10.81	+	1.50E-11
<u>Granulocyte activation</u>	<u>20</u>	1.86	10.74	+	1.68E-11
<u>Positive regulation of biological process</u>	<u>40</u>	21.06	1.90	+	4.27E-02
<u>Unclassified</u>	<u>1</u>	13.46	< 0.2	-	0.00E00

GO biological processes are hyperlinked to their functional description on Panther's website. The “#” column identified how many human proteins (out of 71) were associated with that biological process. In addition, these underlined numbers are hyperlinked to tables identifying which proteins matched that process.

Table A.12: Significant overexpression of GO molecular functions in chimpanzee seminal plasma compared to the human and chimpanzee reference genomes

GO molecular function complete	#	Expected	Fold enrichment	+/-	P value
<u>Structural constituent of epidermis</u>	<u>3</u>	.05	66.18	+	4.53E-02
<u>Protease binding</u>	<u>9</u>	.43	20.90	+	1.82E-06
<u>Enzyme binding</u>	<u>22</u>	7.05	3.12	+	2.37E-03
<u>Protein binding</u>	<u>51</u>	19.61	2.60	+	1.28E-11
<u>Binding</u>	<u>61</u>	36.99	1.65	+	9.45E-07
<u>Serine-type endopeptidase activity</u>	<u>6</u>	.50	12.05	+	3.19E-02
<u>Peptidase activity, acting on L-amino acid peptides</u>	<u>10</u>	1.88	5.31	+	4.93E-02
<u>Catalytic activity</u>	<u>39</u>	18.29	2.13	+	3.35E-04
<u>Hydrolase activity</u>	<u>24</u>	7.71	3.11	+	5.53E-04
<u>Serine-type peptidase activity</u>	<u>7</u>	.57	12.19	+	5.27E-03
<u>Serine hydrolase activity</u>	<u>7</u>	.59	11.82	+	6.46E-03
<u>Identical protein binding</u>	<u>16</u>	4.63	3.45	+	2.89E-02
<u>Metal ion binding</u>	<u>24</u>	9.30	2.58	+	1.64E-02
<u>Cation binding</u>	<u>25</u>	9.58	2.61	+	7.73E-03
<u>Ion binding</u>	<u>34</u>	15.73	2.16	+	3.48E-03

GO molecular functions are hyperlinked to their functional description on Panther's website. The “#” column identified how many chimpanzee proteins (out of 64) were associated with that molecular function. In addition, these underlined numbers are hyperlinked to tables identifying which proteins matched that function.

Table A.13: Significant overexpression of GO biological processes in chimpanzee seminal plasma compared to the human and chimpanzee reference genomes

GO biological process complete	#	Expected	Fold enrichment	+/-	P value
<u>Canonical glycolysis</u>	<u>5</u>	.08	61.77	+	2.02E-04
<u>Glucose catabolic process to pyruvate</u>	<u>5</u>	.08	61.77	+	2.02E-04
<u>Pyruvate metabolic process</u>	<u>7</u>	.22	32.27	+	2.37E-05
<u>Organic substance metabolic process</u>	<u>51</u>	30.72	1.66	+	4.85E-03
<u>Metabolic process</u>	<u>51</u>	32.13	1.59	+	2.55E-02
<u>Cellular process</u>	<u>66</u>	48.61	1.36	+	4.11E-04
<u>Glucose catabolic process</u>	<u>5</u>	.09	53.25	+	4.20E-04
<u>Glucose metabolic process</u>	<u>6</u>	.37	16.40	+	1.72E-02
<u>Hexose metabolic process</u>	<u>8</u>	.51	15.74	+	4.18E-04
<u>Monosaccharide metabolic process</u>	<u>8</u>	.64	12.48	+	2.41E-03
<u>Carbohydrate metabolic process</u>	<u>13</u>	1.52	8.52	+	3.10E-05
<u>Primary metabolic process</u>	<u>48</u>	29.56	1.62	+	4.86E-02
<u>Hexose catabolic process</u>	<u>6</u>	.16	37.82	+	1.35E-04
<u>Monosaccharide catabolic process</u>	<u>6</u>	.19	31.41	+	4.00E-04
<u>Catabolic process</u>	<u>20</u>	6.28	3.18	+	1.79E-02
<u>Carbohydrate catabolic process</u>	<u>9</u>	.36	24.82	+	1.12E-06
<u>Organic substance catabolic process</u>	<u>19</u>	5.33	3.56	+	6.57E-03
<u>NADH regeneration</u>	<u>5</u>	.08	61.77	+	2.02E-04
<u>NADH metabolic process</u>	<u>6</u>	.12	51.48	+	2.19E-05
<u>NAD metabolic process</u>	<u>6</u>	.22	26.86	+	9.98E-04
<u>Nicotinamide nucleotide metabolic process</u>	<u>7</u>	.36	19.48	+	7.27E-04

GO biological process complete	#	Expected	Fold enrichment	+/-	P value
<u>Pyridine nucleotide metabolic process</u>	<u>7</u>	.36	19.48	+	7.27E-04
<u>Pyridine-containing compound metabolic process</u>	<u>7</u>	.38	18.48	+	1.04E-03
Organonitrogen compound metabolic process	41	17.88	2.29	+	3.30E-05
<u>Oxidoreduction coenzyme metabolic process</u>	<u>7</u>	.41	17.02	+	1.79E-03
<u>Coenzyme metabolic process</u>	<u>8</u>	.94	8.55	+	3.94E-02
<u>Organophosphate metabolic process</u>	<u>14</u>	3.20	4.38	+	2.46E-02
<u>Glycolytic process through glucose-6-phosphate</u>	<u>5</u>	.08	59.39	+	2.45E-04
<u>Glycolytic process through fructose-6-phosphate</u>	<u>5</u>	.08	59.39	+	2.45E-04
<u>Glycolytic process</u>	<u>6</u>	.12	48.77	+	3.01E-05
<u>ATP generation from ADP</u>	<u>6</u>	.13	47.52	+	3.51E-05
<u>ADP metabolic process</u>	<u>6</u>	.15	41.18	+	8.16E-05
<u>Purine ribonucleoside monophosphate metabolic process</u>	<u>8</u>	.80	9.96	+	1.28E-02
<u>Purine nucleoside monophosphate metabolic process</u>	<u>8</u>	.81	9.92	+	1.32E-02
<u>Nucleoside monophosphate metabolic process</u>	<u>8</u>	.89	9.02	+	2.67E-02
<u>Ribonucleoside monophosphate metabolic process</u>	<u>8</u>	.84	9.50	+	1.82E-02
<u>Purine ribonucleoside diphosphate metabolic process</u>	<u>6</u>	.20	30.38	+	4.86E-04
<u>Purine nucleoside diphosphate metabolic process</u>	<u>6</u>	.20	30.38	+	4.86E-04
<u>Nucleoside diphosphate metabolic process</u>	<u>6</u>	.27	22.60	+	2.72E-03
<u>Ribonucleoside diphosphate metabolic process</u>	<u>6</u>	.20	29.41	+	5.87E-04
<u>Nucleoside diphosphate phosphorylation</u>	<u>6</u>	.18	33.09	+	2.95E-04
<u>Nucleotide phosphorylation</u>	<u>6</u>	.20	29.89	+	5.35E-04
<u>Generation of precursor metabolites and energy</u>	<u>9</u>	1.01	8.88	+	6.98E-03
<u>ATP metabolic process</u>	<u>8</u>	.66	12.05	+	3.12E-03
<u>Purine ribonucleoside triphosphate metabolic process</u>	<u>8</u>	.76	10.51	+	8.63E-03
<u>Ribonucleoside triphosphate metabolic process</u>	<u>8</u>	.78	10.30	+	1.01E-02
Retina homeostasis	8	.23	34.80	+	9.09E-07
Tissue homeostasis	9	.57	15.79	+	5.56E-05
<u>Multicellular organismal homeostasis</u>	<u>9</u>	.93	9.65	+	3.52E-03
<u>Multicellular organismal process</u>	<u>43</u>	21.68	1.98	+	9.47E-04
<u>Homeostatic process</u>	<u>17</u>	4.76	3.57	+	2.82E-02
<u>Regulation of biological quality</u>	<u>28</u>	11.51	2.43	+	1.78E-02
<u>Anatomical structure homeostasis</u>	<u>10</u>	.95	10.58	+	3.25E-04
Acute inflammatory response	6	.23	26.47	+	1.09E-03
Defense response	18	3.97	4.53	+	4.06E-04

GO biological process complete	#	Expected	Fold enrichment	+/-	P value
<u>Response to stress</u>	<u>29</u>	10.79	2.69	+	1.22E-03
Peptide cross-linking	5	.19	26.17	+	1.36E-02
Platelet degranulation	9	.41	21.89	+	3.34E-06
Regulated exocytosis	27	2.22	12.17	+	1.16E-18
<u>Exocytosis</u>	<u>27</u>	2.49	10.84	+	2.23E-17
<u>Secretion by cell</u>	<u>27</u>	3.11	8.70	+	5.98E-15
<u>Secretion</u>	<u>28</u>	3.45	8.12	+	6.41E-15
<u>Transport</u>	<u>36</u>	14.27	2.52	+	5.84E-05
<u>Establishment of localization</u>	<u>37</u>	14.62	2.53	+	2.68E-05
<u>Localization</u>	<u>43</u>	17.80	2.42	+	1.34E-06
<u>Vesicle-mediated transport</u>	<u>30</u>	5.77	5.20	+	4.40E-11
Neutrophil degranulation	21	1.56	13.46	+	2.16E-14
<u>Neutrophil mediated immunity</u>	<u>22</u>	1.61	13.70	+	1.97E-15
<u>Myeloid leukocyte mediated immunity</u>	<u>22</u>	1.67	13.17	+	4.50E-15
<u>Leukocyte mediated immunity</u>	<u>25</u>	2.38	10.49	+	2.16E-15
<u>Immune effector process</u>	<u>27</u>	3.32	8.13	+	3.25E-14
<u>Immune system process</u>	<u>32</u>	8.26	3.87	+	1.24E-08
<u>Leukocyte degranulation</u>	<u>21</u>	1.63	12.89	+	5.06E-14
<u>Neutrophil activation involved in immune response</u>	<u>21</u>	1.57	13.40	+	2.35E-14
<u>Myeloid cell activation involved in immune response</u>	<u>21</u>	1.66	12.64	+	7.48E-14
<u>Leukocyte activation involved in immune response</u>	<u>22</u>	1.97	11.19	+	1.32E-13
<u>Leukocyte activation</u>	<u>23</u>	2.85	8.07	+	2.25E-11
<u>Cell activation</u>	<u>25</u>	3.32	7.53	+	4.83E-12
<u>Cell activation involved in immune response</u>	<u>22</u>	1.98	11.12	+	1.51E-13
<u>Immune response</u>	<u>28</u>	5.42	5.17	+	6.06E-10
<u>Myeloid leukocyte activation</u>	<u>22</u>	1.83	12.00	+	3.09E-14
<u>Neutrophil activation</u>	<u>21</u>	1.59	13.18	+	3.25E-14
<u>Granulocyte activation</u>	<u>21</u>	1.60	13.10	+	3.67E-14
Regulation of endopeptidase activity	9	1.27	7.09	+	4.33E-02
<u>Regulation of peptidase activity</u>	<u>11</u>	1.36	8.11	+	9.25E-04
<u>Regulation of proteolysis</u>	<u>16</u>	2.72	5.88	+	7.46E-05
Multi-organism process	27	7.56	3.57	+	9.91E-06

GO biological processes are hyperlinked to their functional description on Panther's website. The “#” column identified how many chimpanzee proteins (out of 64) were associated with that biological process. In addition, these underlined numbers are hyperlinked to tables identifying which proteins matched that process.

Table A.13: Significant overexpression of GO biological processes in gorilla seminal plasma compared to the human and chimpanzee reference genomes

GO biological process complete	#	Expected	Fold enrichment	+/-	P value
<u>Antimicrobial humoral response</u>	<u>5</u>	.18	28.49	+	8.02E-03
<u>Positive regulation of serine-type endopeptidase activity</u>	<u>2</u>	.00	> 100	+	1.52E-02
<u>Positive regulation of serine-type peptidase activity</u>	<u>2</u>	.00	> 100	+	1.52E-02
<u>Neutrophil degranulation</u>	<u>4</u>	.01	> 100	+	3.68E-07
<u>Neutrophil mediated immunity</u>	<u>4</u>	.03	> 100	+	2.92E-04
<u>Myeloid leukocyte mediated immunity</u>	<u>4</u>	.06	70.33	+	2.69E-03
<u>Leukocyte mediated immunity</u>	<u>6</u>	.24	25.25	+	1.05E-03
<u>Leukocyte degranulation</u>	<u>4</u>	.03	> 100	+	2.26E-04
<u>Regulated exocytosis</u>	<u>7</u>	.18	39.61	+	3.96E-06
<u>Exocytosis</u>	<u>7</u>	.29	24.44	+	1.07E-04
<u>Secretion by cell</u>	<u>7</u>	.57	12.31	+	1.07E-02
<u>Secretion</u>	<u>7</u>	.71	9.85	+	4.60E-02
<u>Neutrophil activation involved in immune response</u>	<u>4</u>	.01	> 100	+	1.08E-05
<u>Myeloid cell activation involved in immune response</u>	<u>4</u>	.05	72.94	+	2.33E-03
<u>Leukocyte activation involved in immune response</u>	<u>5</u>	.22	22.58	+	2.17E-02
<u>Cell activation involved in immune response</u>	<u>5</u>	.23	21.79	+	2.58E-02
<u>Myeloid leukocyte activation</u>	<u>5</u>	.15	32.82	+	3.52E-03
<u>Neutrophil activation</u>	<u>4</u>	.04	> 100	+	4.66E-04
<u>Granulocyte activation</u>	<u>4</u>	.04	98.47	+	7.08E-04
<u>Platelet degranulation</u>	<u>4</u>	.01	> 100	+	1.08E-05
<u>Antimicrobial humoral response</u>	<u>5</u>	.10	52.38	+	3.55E-04
<u>Humoral immune response</u>	<u>6</u>	.22	27.61	+	6.25E-04
<u>Retina homeostasis</u>	<u>4</u>	.11	37.16	+	3.33E-02
<u>Interspecies interaction between organisms</u>	<u>6</u>	.40	15.15	+	2.01E-02
<u>Protein complex assembly</u>	<u>9</u>	1.33	6.78	+	3.82E-02
<u>Protein complex biogenesis</u>	<u>9</u>	1.33	6.77	+	3.87E-02
<u>Unclassified</u>	<u>4</u>	9.00	.44	-	0.00E00

GO biological processes are hyperlinked to their functional description on Panther's website. The “#” column identified how many gorilla proteins (out of 34) were associated with that biological process. In addition, these underlined numbers are hyperlinked to tables identifying which proteins matched that process.

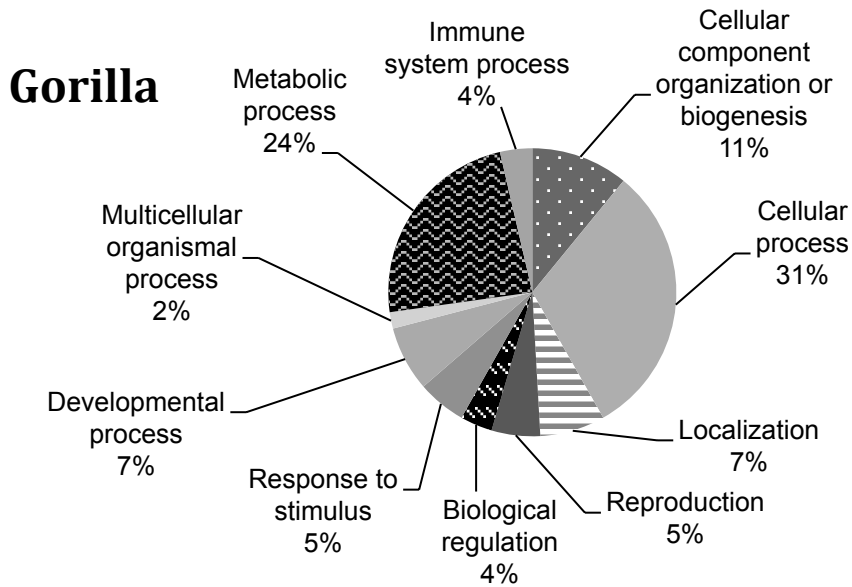
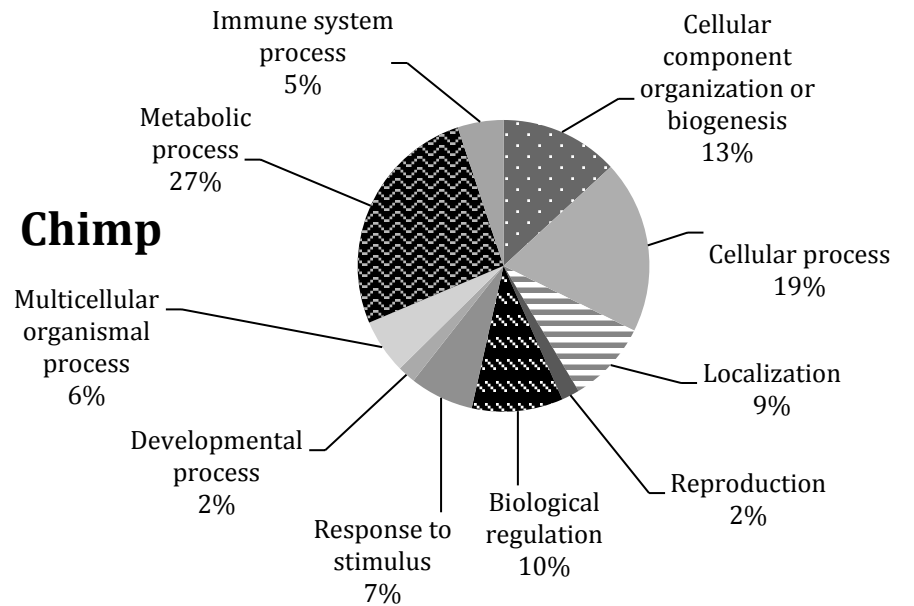
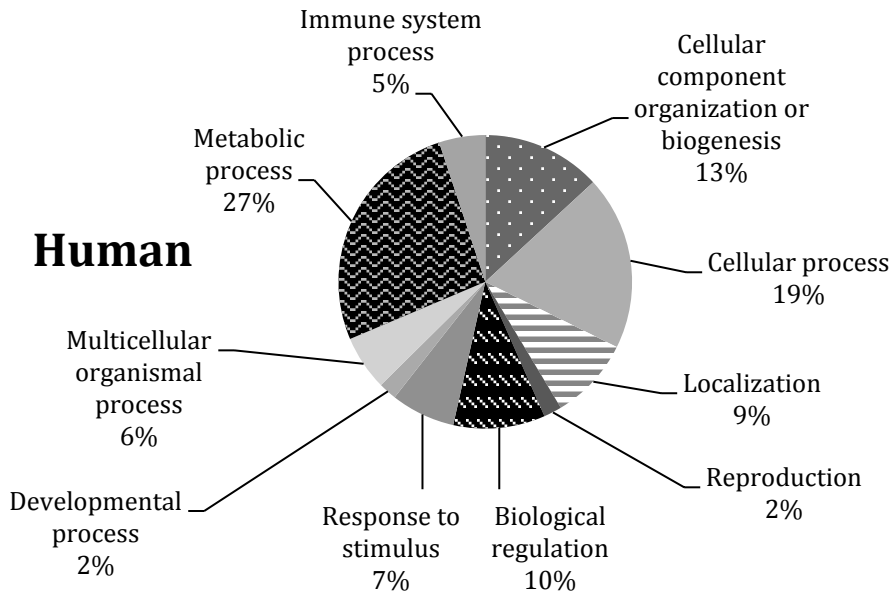


Figure A.2: Biological process functions of all identified proteins in each species: human, chimpanzee, and gorilla seminal plasma

Seminal plasma proteins are associated with 12 biological processes identified through gene ontology (Panther). These three pie charts indicate the species specific percentage distribution of the biological processes that are summarized in Figure A.1.

A.8 Sequencing and PCR primers

Table A.14: Vector and gene specific sequencing primers

Vector	Primer
Ebv_Rev_pFB	5' - GATGAGTTTGGACAAACCAC-3'
pFLAG_N-CMV-30_For	5' - AATGTCGTAATAACCCCGCCCCGTTGACGC-3'
pFLAG_C-CMV-24_Rev	5' - TATTAGGACAAGGCTGGTGGGCAC-3'
HIS_Reverse	5' - AATGGTGATGGTGATGGTG-3'
AC_M13_Rev	5' - AACAGCTATGACCATG-3'
AC_M13-20_Fwd	5' - GTAAAACGACGGCCAGT-3'
Poly_Fwd_pFB	5' - AAATGATAACCATCTCGC-3'
AC_pSG5_Rev	5' - AGTTTGGACAAACCACAACACTAG-3'
SG5_RevMCS	5' - CTGCAATAAACAAGTTCTGC-3'
T3_Fwd	5' - AATTAACCCTCACTAAAGGG-3'
T7_Rev	5' - CATTATGCTGAGTGATATCCCG-3'
Gene	Primer
R_KLK3_80	5' - CACTCCCAGCCTCCCACAATC-3'
KLK3_SeqInternal_Fwd	5' - CTCATCCTGTCTCGGATTG-3'
KLK3_SeqInternal_Rev	5' - GAGGGTGAACCTTGCGCAC-3'
R_SEMG_100	5' - GCTGCTTGCTTCTCCAAG-3'
AC_SgI_450_Reverse	5' - CTTTCTTCTGTGTTTGAATATTG-3'
AC_SgI_mRNA_Forward_450	5' - CATCTGGAAAGGGAATATCCAG-3'
AC_SgI_mRNA_Forward_650	5' - AGCTAACAAACAACAACGTGAGA-3'
AC_SgI_mRNA_Reverse_950	5' - GGGATACATCTTTCTGCACACC-3'
AC_SgI_mRNA_Reverse_1200	5' - GCCATGGCTCTTGCTTAGGA-3'
AC_SgI_Fwd_1200	5' - CAAATCCTAAGCAAGAGCCATG-3'
AC_SgII_Rev_480	5' - GTTTGAACATTGACTGGATAATCC-3'
AC SgII mRNA Forward 480	5' - GGATTATCCAGTCAATGTTCAAAC-3'
AC SgII mRNA Forward 673	5' - GGGCATTACCAAAATGTGGTTGACGTG-3'

Gene	Primer
AC SgII mRNA Reverse 976	5' - CAGTTTGGATAGAAATGCTGCCTTTGG-3'
AC SgII mRNA Reverse 1209	5' - GCTTGACTAGGAATTCTTACCTGG-3'
AC_SGII_Fwd_1290	5' - GCTTGACTAGGAATTCTTACCTGG-3'
TGM4_R80	5' - CTCCCATGTGTGGTGAG-3'
AC_TGM4_Rev420	5' - GTTGGTACTTGCCCAGGATGGCATTGGGGG-3'
AC_TGM4_Fwd 420	5' - CCCCCAATGCCATCCTGGGCAAGTACCAAC-3'
AC_TGM4_Fwd720	5' - CCCCCTGCTGGTGTGCAGGGCCATGTGTG-3'
AC_TGM4_Rev720	5' - CACACATGGCCCTGCACACCAGCACGGGG-3'
AC_TGM4_Fwd1020	5' - GTGAATGAGAATGGCGAGAAAATCACCA-3'
AC_TGM4_Rev1020	5' - TGGTGATTTTCTCGCCATTCTCATTAC-3'
AC_TGM4_Fwd1620	5' - CCTTTGAACCTACAGTTGTACTGCGCAAG-3'
AC_TGM4_Rev1620	5' - CTTGCCAGTGTACAACCTGTAGTTCAAAGG-3'
AC_TGM4_Fwd1920	5' - CAAGAATACCCTGGCCATCCCTTTGA-3'
AC_TGM4_Rev1920	5' - TCAAAGGGATGGCCAGGGTATTCTTG-3'
AC_TGM4_Fwd2220	5' - CAGATTATGGATGATTAATTTGATGAC-3'
AC_TGM4_Fwd2520	5' - CCCTTGAGAAGCTGCCATATCTTCAGGCC-3'

Table A.15: Control PCR primers for genomic DNA

Gene	Primer
CFTR_ex10_F	5' - CATGTCCTCTAGAAACCGTATGC-3'
CFTR_ex10_R	5' - CCCATACATTCTCCTAATGCTCA-3'
Mt_5269_F	5' - CCYCTGTCTTTAGATTTACAGTCC-3'
Mt_6266_R	5' - GCGAGYCAGCTRAATACTTTGACG-3'

A.9 Human-chimpanzee TGM4 ancestor sequence reconstruction

```

Human Change
Chimp Change
HumanTGM4      MMDASKELQVLHIDFLNQDNAVSHHTWEFQTS SPVFRRGQVFHLRLVLNQPLQSYHQLKL 60
ChimpanzeeTGM4 MMDASKELQVLHIDFLNQDNAVSHHTWEFQTS TPVFRRGQVFHLRLVLNQPLQSYHQLKL 60
OrangutanTGM4  MMDTSKELQVLHIDFLKQENAVSHHTWEFQTS SPVFRRGQVFHLRLVLNQPLQSYHQLKL 60
GibbonTGM4     MMDTSKELQVLHIDFLKQENAVSHHTWEFQTS SPVFRRGQVFHLRLVLNQPLQSYHQLKL 60
*****:*****:*****:*****:*****:*****:*****:*****

HumanTGM4      EFSTGPNPSIAKHTLVVLDERTPSDHYNWQATLQNESGKEVTVAVTSSPNAILGKYQLNV 120
ChimpanzeeTGM4 EFSTGPNPSIAKHTLVVLDERTPSDHYNWQATLQNESGKEVTVAVTSSPNAILGKYQLNV 120
OrangutanTGM4  EFSTGPNPSIAKHTLVVLDERTPSDHYNWQATLQNESGKEVTVAVTSSPNAILGKYQLNV 120
GibbonTGM4     EFSTGPNPSIAKHTLVVLDERTPSAHYNWQAALQNESGKEVTVAVTSSPNAILGKYQLNV 120
*****:*****:*****:*****:*****:*****:*****:*****

HumanTGM4      KTGNIHLKSEENILYLLFNPWCKEDMVFMPDEDERKEYILNDTGCHYVGAARSIKKPKWN 180
ChimpanzeeTGM4 KTGNIHLKSEENILYLLFNPWCKEDMVFMPDEDERKEYILNDTGCHYVGAARSIRYKPKWN 180
OrangutanTGM4  KTGNIHLKSEENILYLLFNPWCKEDMVFMPDEDERKEYILNDTGCHYVGAARSIKKPKWN 180
GibbonTGM4     KTGNIHLKSEENILYLLFNPWCKEDMVFMPDEDERKEYILNDTGCHYVGAARSIRYKPKWN 180
*****:*****:*****:*****:*****:*****:*****

HumanTGM4      FGQFEKNVLDCCISLLETSSLKPTDRRDPVLVCRAMCAMMSFEKGGVLIQNWTDGYEGG 240
ChimpanzeeTGM4 FGQFEKNVLDCCISLLETSSLKPTDRRDPVLVCRAMCAMMSFEKGGVLIQNWTDGYEGG 240
OrangutanTGM4  FGQFEKNVLDCCISLLETSSLKPTDRRDPVLVCRAMCAMMSFEKGGVLIQNWTDGYEGG 240
GibbonTGM4     FGQFEENVLDCCISLLETSSLKPTDRRDPVLVCRAMCAMMSFEKGGVLIQNWTDGYQGG 240
*****:*****:*****:*****:*****:*****:*****

HumanTGM4      TAPYKWTGSAPILQQYYNTKQAVCFGQCWVFAGILTTVLRALGIPARSVTGFDSAHDTER 300
ChimpanzeeTGM4 TAPYKWTGSAPILQQYYNTKQAVCFGQCWVFAGILTTVLRALGIPARSVTGFDSAHDTER 300
OrangutanTGM4  TAPYKWTGSAPILQQYYNTKQAVCFGQCWVFAGILTTVLRALGIPARSVTAFDSAHDTER 300
GibbonTGM4     TAPYKWTGSAPILQQYYNTKQAVCFGQCWVFAGILTTVLRALGIPARSVTAFDSAHDTER 300
*****:*****:*****:*****:*****:*****

HumanTGM4      NLTVDTYVNGEKITMTHDSVWNFHVWTDAMMKRPDLPGKYDGWQAVDATPQERSQGV 360
ChimpanzeeTGM4 NLTVDTYVNGEKITMTHDSVWNFHVWTDAMMKRPDLPGKYDGWQAVDATPQERSQGV 360
OrangutanTGM4  NLTVDTYVNGEKITMTHDSVWNFHVWTDAMMKRPDLPGKYDGWQAVDATPQERSQGV 360
GibbonTGM4     NLTVDTYVNGEKITMTHDSVWNFHVWTDAMMKRPDLPGKYDGWQAVDATPQERSQGV 360
*****:*****:*****:*****:*****:*****

HumanTGM4      FCCGSPSLTAIRKGDIFIVYDTRFVSEVNGDRLIWLKVMVNGQEELHVISMETTSIGKN 420
ChimpanzeeTGM4 FCCGSPSLTAIRKGDIFIVYDTRFVSEVNGDRLIWLKVMVNGQEELHVISMETTSIGKN 420
OrangutanTGM4  FCCGSPSLTAIRKGDIFIVYDTRFVSEVNGDRLIWLKVMVNGQEELHVISMETTSIGKN 420
GibbonTGM4     FCCGSPSLTAIRKGDIFIVYDTRFVSEVNGDRLIWLKVMVNGQEELHVISMETTSIGKN 420
*****:*****:*****:*****:*****:*****

HumanTGM4      ISTKAVGQDRRDITYEYKYPEGSSEERQVMDHAFLLSSEREHRRPVKENFLHMSVQSD 480
ChimpanzeeTGM4 ISTKAVGQDRRDITYEYKYPEGSSEERQVMDHAFLLSSEREHRRPVKENFLHMSVQSD 480
OrangutanTGM4  ISTKAVGQDRRDITYEYKYPEGSSEERQVMDHAFLLSSEREHRRPVKENFLHMSVQSD 480
GibbonTGM4     ISTKAVGQDRRDITYEYKYPEGSSEERQVMDHAFLLSSEREHRRPVKENFLHMSVQSD 480
*****:*****:*****:*****:*****:*****

HumanTGM4      DVLLGNFVNFTVILKRKTAALQNVNLSGFEQLYTGKKWAKLCDLNKTSQIQGVSEVT 540
ChimpanzeeTGM4 DVLLGNFVNFTVILKRKTAALQNVNLSGFEQLYTGKKWAKLCDLNKTSQIQGVSEVT 540
OrangutanTGM4  DVLLGNFVNFTVILKRKTAALQNVNLSGFEQLYTGKKWAKLCDLNKTSQIQGVSEVT 540
GibbonTGM4     DVLLGNFVNFTVILKRKTAALQNVNLSGFEQLYTGKKWAKLCDLNKTSQIQGVSEVT 540
*****:*****:*****:*****:*****:*****

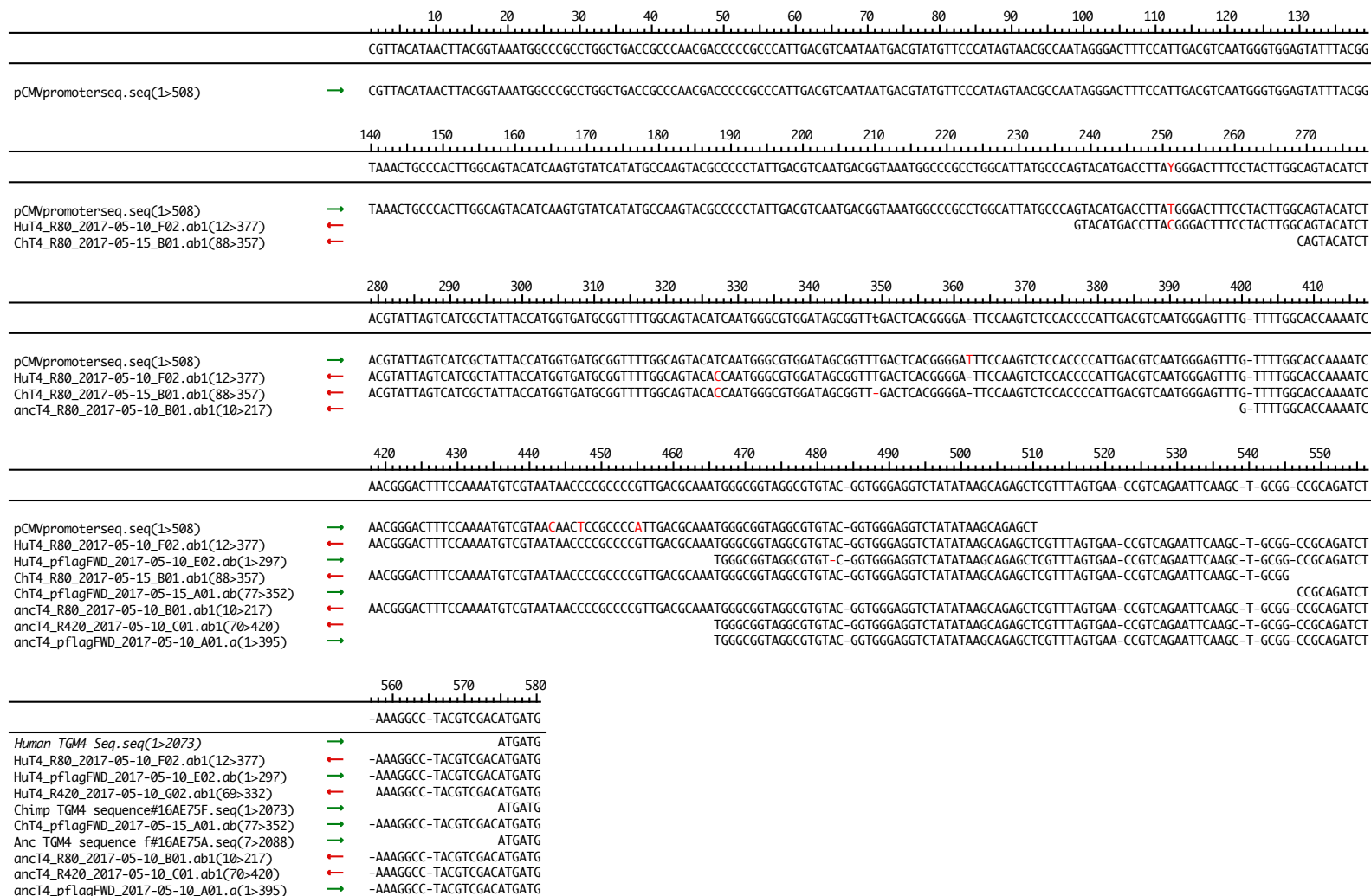
HumanTGM4      LTLDSKTYINSLAILDDEPVIRGFIIAEIVESKEIMASEVFTSFQYPEFSIELPNTGRIG 600
ChimpanzeeTGM4 LTLDSKTYINSLAILDDEPVIRGFIIAEIVESKEIMASEVFTSFQYPEFSIELPNTGRIG 600
OrangutanTGM4  LTLDSKTYINSLAILDDEPVIRGFIIAEIVESNEIMASEVFXSFQYPEFSIELPNTGRIG 600
GibbonTGM4     LTLDSKTYINSLAILDDEPVIRGFIIAEIVESNEIMASEVFTSFQYPEFSIELPNTGRIG 600
*****:*****:*****:*****:*****:*****

HumanTGM4      QLLVCNCFKNTLAIPLTDVKFSLLESLSLQTS DHGTVQPGETIQSQIKCTPIKGTGPK 660
ChimpanzeeTGM4 QLLVCNCFKNTLAIPLTDVKFSLLESLSLQTS DHGTVQPGETIQSQIKCTPIKGTGPK 660
OrangutanTGM4  QLLVCNCFKNTLAIPLTDVKFSLLESLSLQTS DHGTVQPGETIQSQIKCTPIKGTGPK 660
GibbonTGM4     QLLVCNCFKNTLAIPLTDVKFSLLESLSLQTS DHGTVQPGETIQSQIKCAPIKGTGPK 660
*****:*****:*****:*****:*****:*****

HumanTGM4      KFIVKLSSKQVKEINAQKIVLITK 684
ChimpanzeeTGM4 KFIVKLSSKQVKEINAQKIVLITK 684
OrangutanTGM4  KFIVKLSSKQVKEINAQKIVLITK 684
GibbonTGM4     KFIKLSSKQVKEINAQKIVLITK 684
*****:*****:*****:*****

```

A.10 Human, chimpanzee, and human-chimpanzee ancestor sequences in pCMV vector



10 20 30 40 50 60 70

ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACATTGACTTCTT-G-AATCAGGACAACGCCG-TTTCTCAC

Top
 M M D A S K E L Q V L H I D F L N Q D N A V S H
 • W M H Q K S C K F S T L T S • I R T T P F L T
 D G C I K R A A S S P H • L L E S G Q R R F S P

Human TGM4 Seq.seq(1>2073) → ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACATTGACTTCTT-G-AATCAGGACAACGCCG-TTTCTCAC
 HuT4_R80_2017-05-10_F02.ab1(12>49) ← ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACAT
 HuT4_pflagFWD_2017-05-10_E02.ab(103>297) → ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACATTGACTTCTT-G-AATCAGGACAACGCCG-TTTCTCAC
 HuT4_R420_2017-05-10_G02.ab1(69>316) ← ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACATTGACTTCTT-G-AATCAGGACAACGCCG-TTTCTCAC
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACGTTGACTTCTT-G-AAGCAGGACAACGCCG-TTTCTCAC
 ChT4_pflagFWD_2017-05-15_A01.ab(103>352) → ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCAGTTGACTTCTT-G-AAGCAGGACAACGCCG-TTTCTCAC
 ChT4_R420_2017-05-15_C01.ab1(87>301) ← AAGTTCTCCACGTTGACTTCT--G-AAGCAGGACAACGCCG-TTTCTCAC
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACGTTGACTTCTT-G-AAGCAGGACAACGCCG-TTTCTCAC
 ancT4_R80_2017-05-10_B01.ab1(10>50) ← ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACGT-GAC
 ancT4_R420_2017-05-10_C01.ab1(70>318) ← ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACGTTGACTTCTT-G-AAGCAGGACAACGCCG--TTCTCAC
 ancT4_pflagFWD_2017-05-10_A01.a(103>395) → ATGATGGATGCATCAAAA-GAGCTG-CAAGTTCTCCACGTTGACTTCTT-G-AAGCAGGACAACGCCG-TTTCTCAC

80 90 100 110 120 130 140 150

CACACATGGGAGTTCCAAACGAGCAGTCTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA

Top
 H T W E F Q T S S P V F R R G Q V F H L R L V L N Q
 T H G S S K R A V L C S G E D R C F T C G W C • T S
 H M G V P N E Q S C V P A R T G V S P A A G A E P

Human TGM4 Seq.seq(1>2073) → CACACATGGGAGTTCCAAACGAGCAGTCTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA
 HuT4_pflagFWD_2017-05-10_E02.ab(103>297) → CACACATGGGAGTTCCAAACGAGCAGTCTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA
 HuT4_R420_2017-05-10_G02.ab1(69>316) ← CACACATGGGAGTTCCAAACGAGCAGTCTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → CACACATGGGAGTTCCAAACGAGCACTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA
 ChT4_pflagFWD_2017-05-15_A01.ab(103>352) → CACACATGGGAGTTCCAAACGAGCACTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA
 ChT4_R420_2017-05-15_C01.ab1(87>301) ← CACACATGGGAGTTCCAAACGAGCACTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → CACACATGGGAGTTCCAAACGAGCAGTCTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA
 ancT4_R420_2017-05-10_C01.ab1(70>318) ← CACACATGGGAGTTCCAAACGAGCAGTCTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA
 ancT4_pflagFWD_2017-05-10_A01.a(103>395) → CACACATGGGAGTTCCAAACGAGCAGTCTCTGTGTTCCGGCGAGGACAGGTGTTTACCTGCGGCTGGTGCTGAACCA

160 170 180 190 200 210 220 230

GCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-

Top
 P L Q S Y H Q L K L E F S T G P N P S I A K H T L
 P Y N P T T N • N W N S A Q G R I L A S P N T P
 A P T I L P P T E T G I O H R A E S • H R Q T H P

Human TGM4 Seq.seq(1>2073) → GCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-
 HuT4_pflagFWD_2017-05-10_E02.ab(103>297) → GCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG
 HuT4_R420_2017-05-10_G02.ab1(69>316) ← GCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → GCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-
 ChT4_pflagFWD_2017-05-15_A01.ab(103>352) → GCCCTAC-ATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-
 ChT4_R420_2017-05-15_C01.ab1(87>301) ← GCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-

		160	170	180	190	200	210	220	230	
		GCCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-								
Top		P L Q S Y H Q L K L E F S T G P N P S I A K H T L P Y N P T T N • N W N S A Q G R I L A S P N T P A P T I L P P T E T G I Q H R A E S • H R Q T H P								
Human TGM4 Seq.seq(1>2073)	→	GCCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-								
Anc TGM4 sequence f#16AE75A.seq(7>2088)	→	GCCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-								
ancT4_R420_2017-05-10_C01.ab1(70>318)	←	GCCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-								
ancT4_R720_2017-05-10_F01.ab1(100>450)	←	CAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-								
ancT4_pflagFWD_2017-05-10_A01.a(103>395)	→	GCCCCTACAATCCTACCACCAACTGAAACTGGAATTCAGCACAGGG-CCGAAT-CCTAGCATCGCCAAACACACCC-								
		240	250	260	270	280	290	300		
		TGGT-GGTGCTCGACCCGA-GGACGCC-C-TCAGACC-ACTACAACCTGGC-AGGCAACCCCTTCAAATGAGTCTGGC								
Top		V V L D P R T P S D H Y N W Q A T L Q N E S G W W C S T R G R P Q T T T T G R Q P F K M S L A G G A R P E D A L R P L Q L A G N P S K • V W Q								
Human TGM4 Seq.seq(1>2073)	→	TGGT-GGTGCTCGACCCGA-GGACGCC-C-TCAGACC-ACTACAACCTGGC-AGGCAACCCCTTCAAATGAGTCTGGC								
HuT4_R420_2017-05-10_G02.ab1(69>316)	←	TGGT-GGTGCTCGACCCGA-GGACGCC								
HuT4_R720_2017-05-10_B03.ab1(68>386)	←	GACGCC-C-TCAGACC-A-TNCAACTGGC-AGGCAACCCCTTCAAATGAGTCTNGC								
Chimp TGM4 sequence#16AE75F.seq(1>2073)	→	TGGT-GGTGCTCGACCTGA-GGACGCC-C-TCAGACC-ACTACAACCTGGC-AGGCAACCCCTTCAAATGAGTCTGGC								
ChT4_pflagFWD_2017-05-15_A01.ab(103>352)	→	TGGT-GGTGCTCGACCTGA-GGACGCC-C-T								
ChT4_R420_2017-05-15_C01.ab1(87>301)	←	TGGT-GGTGCTCGACCTGA								
Anc TGM4 sequence f#16AE75A.seq(7>2088)	→	TGGT-GGTGCTCGACCCGA-GGACGCC-C-TCAGACC-ACTACAACCTGGC-AGGCAACCCCTTCAAATGAGTCTGGC								
ancT4_R420_2017-05-10_C01.ab1(70>318)	←	TGGT-GGTGCTCGACCCGA-GGACGCC-C-T								
ancT4_R720_2017-05-10_F01.ab1(100>450)	←	TGGT-GGTGCTCGACCCGA-GGACGCC-C-TCAGACC-A-TACAACCTGGC-AGGCAACCCCTTCAAATGAGTCTGGC								
ancT4_pflagFWD_2017-05-10_A01.a(103>395)	→	TGGT-GGTGCTCGACCCGA-GGACGCC-C-TCAGACC-ACTACAACCTGGC-AGGCAACCCCTTCAAATGAGTCTGG								
		310	320	330	340	350	360	370	380	
		AAAG-AGGTCAC-AGTGGCT-GTCACCAGTT-CCCCC-AATG-CCATCCT-GGGCAAGTACCAA-CTAAAC-GTGAA								
Top		K E V T V A V T S S P N A I L G K Y Q L N V K K R S Q W L S P V P P M P S W A S T N • T • K R G H S G C H Q F P Q C H P G Q V P T K R E								
Human TGM4 Seq.seq(1>2073)	→	AAAG-AGGTCAC-AGTGGCT-GTCACCAGTT-CCCCC-AATG-CCATCCT-GGGCAAGTACCAA-CTAAAC-GTGAA								
HuT4_R720_2017-05-10_B03.ab1(68>386)	←	AAAG-AGGTCAC-AGTNGCT-GTCACCAGTT-CCCCC-AATG-CCATCCT-GGGCAAGTACCAA-CTAAAC-GTGAA								
Chimp TGM4 sequence#16AE75F.seq(1>2073)	→	AAAG-AGGTCAC-AGTGGCT-GTCACCAGTT-CCCCC-AATG-CCATCCT-GGGCAAGTACCAA-CTAAAT-GTGAA								
ChT4_R720_2017-05-15_F01.ab1(82>319)	←	AGTGGCT-GTCACCAGTT-CCCCC-AATG-CCATCCT-GGGCAAGTACCAA-CTAAAT-GTGAA								
Anc TGM4 sequence f#16AE75A.seq(7>2088)	→	AAAG-AGGTCAC-AGTGGCT-GTCACCAGTT-CCCCC-AATG-CCATCCT-GGGCAAGTACCAA-CTAAAT-GTGAA								
ancT4_R720_2017-05-10_F01.ab1(100>450)	←	AAAG-AGGTCAC-AGTGGCT-GTCACCAGTT-CCCCC-AATG-CCATCCT-GGGCAAGTACCAA-CTAAAT-GTGAA								

390 400 410 420 430 440 450 460

AACTGGAAACACATCCTTAAGTCTGAAGAAAACATCCTATACCTTCTCTTCAACCCATGGTGTAAAGAGGACATGG

Top

T G N H I L K S E E N I L Y L L F N P W C K E D M V
 L E T T S L S L K K T S Y T F S S T H G V K R T W
 N W K P H P • V • R K H P I P S L Q P M V • R G H G

Human TGM4 Seq.seq(1>2073) → AACTGGAAACACATCCTTAAGTCTGAAGAAAACATCCTATACCTTCTCTTCAACCCATGGTGTAAAGAGGACATGG
 HuT4_R720_2017-05-10_B03.ab1(68>386) ← AACTGGAAACACATCCTTAAGTCTGAAGAAAACATCCTATACCTTCTCTTCAACCCATGGTGTAAAGAGGACATGG
 HuT4_F420_2017-05-10_H02.ab1(62>351) → CCCATGGTGTAAA-AGGACATGG
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → AACTGGAAACACATCCTTAAGTCTGAAGAAAACATCCTATACCTTCTCTTCAACCCATGGTGTAAAGAGGACATGG
 ChT4_R720_2017-05-15_F01.ab1(82>319) ← AACTNGAAACACATCCTTAAGTCTGAAGAAAACATCCTATACCTTCTCTTCAACCCATGGTGTAAAGAGGACATGG
 ChT4_F420_2017-05-15_D01.ab1(19>243) → ATGG
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → AACTGGAAACACATCCTTAAGTCTGAAGAAAACATCCTATACCTTCTCTTCAACCCATGGTGTAAAGAGGACATGG
 ancT4_R720_2017-05-10_F01.ab1(100>450) ← AACTGGAAACACATCCTTAAGTCTGAAGAAAACATCCTATACCTTCTCTTCAACCCATGGTGTAAAGAGGACATGG
 ancT4_F420_2017-05-10_D01.ab1(22>362) → ATGG

470 480 490 500 510 520 530

TTTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA

Top

F M P D E D E R K E Y I L N D T G C H Y V G A A R
 F S C L M R T S A K S T S S M T R A A I T W G L P E
 F H A • • G R A Q R V H P Q • H G L P L R G G C Q K

Human TGM4 Seq.seq(1>2073) → TTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA
 HuT4_R720_2017-05-10_B03.ab1(68>386) ← TTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA
 HuT4_F420_2017-05-10_H02.ab1(62>351) → TTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → TTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA
 ChT4_R720_2017-05-15_F01.ab1(82>319) ← TTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA
 ChT4_F420_2017-05-15_D01.ab1(19>243) → TTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → TTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA
 ancT4_R720_2017-05-10_F01.ab1(100>450) ← TTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA
 ancT4_F420_2017-05-10_D01.ab1(22>362) → TTTTCATGCCTGATGAGGACGAGCGCAAAGAGTACATCCTCAATGACACGGGCTGCCATTACGTGGGGGCTGCCAGA

540 550 560 570 580 590 600 610

AGTATCAAATGCAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTCCCTGCTGACT

Top

S I K C K P W N F G Q F E K N V L D C C I S L L T
 V S N A N P L G T L V S L R K M S W T A A F P C • L
 Y Q M O T L E L W S V • E K C P G L L H F P A D •

Human TGM4 Seq.seq(1>2073) → AGTATCAAATGCAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTCCCTGCTGACT
 HuT4_R720_2017-05-10_B03.ab1(68>386) ← AGTATCAAATGCAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTCCCTGCTGACT
 HuT4_F420_2017-05-10_H02.ab1(62>351) → AGTATCAAATGCAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTCCCTGCTGACT
 HuT4_R1020_2017-05-10_D03.ab1(67>346) ← TTCCCTGCTGACT
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → AGTATCAGATACAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTCCCTGCTGACT
 ChT4_R720_2017-05-15_F01.ab1(82>319) ← AGTATCAGATACAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTCCCTGCTGACT
 ChT4_F420_2017-05-15_D01.ab1(19>243) → CAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTCCCTGCTGACT
 ChT4_R1020_2017-05-15_H01.ab1(81>389) ← CAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATNTCC-T-GACTGCTGCA-TTCCCTG-TGACT

		540	550	560	570	580	590	600	610	
		AGTATCAAATGCAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTTCCTGCTGACT								
Top		S I K C K P W N F G Q F E K N V L D C C I S L L T V S N A N P G T L V S L R K M S W T A A F P C • L Y Q M Q T L E L W S V • E K C P G L L H F P A D •								
Human TGM4 Seq.seq(1>2073)	→	AGTATCAAATGCAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTTCCTGCTGACT								
Anc TGM4 sequence f#16AE75A.seq(7>2088)	→	AGTATCAAATACAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTTCCTGCTGACT								
ancT4_R720_2017-05-10_F01.ab1(100>450)	←	AGTATCAAATACAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTTCCTGCTGACT								
ancT4_F420_2017-05-10_D01.ab1(22>362)	→	AGTATCAAATACAAACCTGGAACCTTTGGTCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTTCCTGCTGACT								
ancT4_R1020_2017-05-10_H01.ab1(69>374)	←	TCAGTTTGAGAAAAATGTCC-TGGACTGCTGCA-TTTCCTGCTGACT								
		620	630	640	650	660	670	680	690	
		GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
Top		E S S L K P T D R R D P V L V C R A M C A M M S R A P S S P Q I G G T P C W C A G P C V L • E E L P Q A H R • E G P R A G V Q G H V C Y D E								
Human TGM4 Seq.seq(1>2073)	→	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
HuT4_F420_2017-05-10_H02.ab1(62>351)	→	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
HuT4_R1020_2017-05-10_D03.ab1(67>346)	←	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
HuT4_F720_2017-05-10_A03.ab1(1>326)	→	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
Chimp TGM4 sequence#16AE75F.seq(1>2073)	→	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
ChT4_F420_2017-05-15_D01.ab1(19>243)	→	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
ChT4_R1020_2017-05-15_H01.ab1(81>389)	←	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
Anc TGM4 sequence f#16AE75A.seq(7>2088)	→	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
ancT4_F420_2017-05-10_D01.ab1(22>362)	→	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
ancT4_R1020_2017-05-10_H01.ab1(69>374)	←	GAGAGCT-CCCTCAAG-CCCACAGATAGGA-GGGA-CCCCGTGCTGGTGTGCA-GGGCCATGTGTGCTATGA-TGAG								
		700	710	720	730	740	750	760	770	
		CTTTGAGAAA-GGCCAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
Top		F E K G Q G V L I G N W T G D Y E G G T A P Y K L R K A R A C S L G I G L G T T K V A Q S P H T L • E R P G R A H W E L D W G L R R W H S P I Q								
Human TGM4 Seq.seq(1>2073)	→	CTTTGAGAAA-GGCCAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
HuT4_F420_2017-05-10_H02.ab1(62>351)	→	CTTTGAGAAA-GGC-AGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-AC								
HuT4_R1020_2017-05-10_D03.ab1(67>346)	←	CTTTGAGAAA-GGCCAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
HuT4_F720_2017-05-10_A03.ab1(1>326)	→	CTTTGAGAAA-GGCCAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
Chimp TGM4 sequence#16AE75F.seq(1>2073)	→	CTTTGAGAAA-GGCCAAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
ChT4_R1020_2017-05-15_H01.ab1(81>389)	←	CTT-GAGAAA-GGCCAAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
ChT4_F720_2017-05-15_E01.ab1(75>316)	→	CTTTGAGAAA-GGCCAAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
Anc TGM4 sequence f#16AE75A.seq(7>2088)	→	CTTTGAGAAA-GGCCAAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
ancT4_F420_2017-05-10_D01.ab1(22>362)	→	CTTTGAGAAA-GGCCAAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
ancT4_R1020_2017-05-10_H01.ab1(69>374)	←	CTTTGAGAAA-GGCCAAGGGCGTGCTCA-TTGGG-AA-TTGGACT-GGGG-ACTACGAAGGTGGCACAGCCCCATACA								
ancT4_F720_2017-05-10_E01.ab1(55>381)	→	AAGGTGGCACAGCCCCATACA								

780 790 800 810 820 830 840

AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG

Top

W T G S A P I L Q Q Y Y N T K Q A V C F G Q C W V
 S G Q A V P R S C S S T T T R S R L C A L A S A G C
 V D R Q C C P D P A A V L Q H E A G C V L W P V L G V

Human TGM4 Seq.seq(1>2073) → AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG
 HuT4_R1020_2017-05-10_D03.ab1(67>346) ← AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG
 HuT4_F720_2017-05-10_A03.ab1(1>326) → AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG
 ChT4_R1020_2017-05-15_H01.ab1(81>389) ← AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG
 ChT4_F720_2017-05-15_E01.ab1(75>316) → AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG
 ancT4_F420_2017-05-10_D01.ab1(22>362) → AGTGGACAGGCAGTGCCC-GATCCTGCAGCAGTACTACAACAC
 ancT4_R1020_2017-05-10_H01.ab1(69>374) ← AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG
 ancT4_F720_2017-05-10_E01.ab1(55>381) → AGTGGACAGGCAGTGCCCCGATCCTGCAGCAGTACTACAACACGAAGCAGGCTGTGTGCTTTGGCCAGTCTGGGTG

850 860 870 880 890 900 910 920

TTTGCTGGGATCCTGACTACAGTGCTGAGAGCGTTGGGCAT-CCCAGCAC-GCAGTGTGACAGGCTTCGATTCAGCT

Top

F A G I L T T V L R A L G I P A R S V T G F D S A
 L L G S • L Q C • E R W A S Q H A V • Q A S I Q L
 C W D P D Y S A E S V G H P S T Q C D R L R F S S

Human TGM4 Seq.seq(1>2073) → TTTGCTGGGATCCTGACTACAGTGCTGAGAGCGTTGGGCAT-CCCAGCAC-GCAGTGTGACAGGCTTCGATTCAGCT
 HuT4_R1020_2017-05-10_D03.ab1(67>346) ← TTTGCTGGGATCCTGACTACAGTGCTGAGAGCGTTGGGCAT-CCCAGCAC
 HuT4_F720_2017-05-10_A03.ab1(1>326) → TTTGCTGGGATCCTGACTACAGTGCTGAGAGCGTTGGGCAT-CCCAGCAC-GCAGTGTGACAGGCTTCGATTCAGCT
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → TTTGCTGGGATCCTGACTACAGTGCTGAGAGCGTTGGGCAT-CCCAGCAC-GCAGTGTGACAGGCTTCGATTCAGCT
 ChT4_R1020_2017-05-15_H01.ab1(81>389) ← TTTGCTGGGATCCTGACTACAGTGCTGAG
 ChT4_F720_2017-05-15_E01.ab1(75>316) → TTTGCTGGGATCCTGACTACAGTGCTGAGAGCGTTGGGCAT-CCCAGCAC-GCAGTGTGACAGGCTTCGATTCAGCT
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → TTTGCTGGGATCCTGACTACAGTGCTGAGAGCGTTGGGCAT-CCCAGCAC-GCAGTGTGACAGGCTTCGATTCAGCT
 ancT4_R1020_2017-05-10_H01.ab1(69>374) ← TTTGCTGGGATCCTGACTACAGTGCTGAGAGCGTTGGGCAT
 ancT4_F720_2017-05-10_E01.ab1(55>381) → TTTGCTGGGATCCTGACTACAGTGCTGAGAGCGTTGGGCAT-CCCAGCAC-GCAGTGTGACAGGCTTCGATTCAGCT

930 940 950 960 970 980 990 1000

CACGACACAGAAAGGAA-CCTCACGGTGGACACCTATGTGAATGAGAATGGCGAGAAAATCACCAGTATGACCCACG

Top

H D T E R N L T V D T Y V N E N G E K I T S M T H D
 T T Q K G T S R W T P M • M R M A R K S P V • P T
 R H R K E P H G G H L C E • E W R E N H Q Y D P R

Human TGM4 Seq.seq(1>2073) → CACGACACAGAAAGGAA-CCTCACGGTGGACACCTATGTGAATGAGAATGGCGAGAAAATCACCAGTATGACCCACG
 HuT4_F720_2017-05-10_A03.ab1(1>326) → CACGACACAGAAAGGAA-CCTCACGGTGGACACCTATGTGAATGAGAATGGCGAGAAAATCACCAGTATGACCCACG
 HuT4_F1020_2017-05-10_C03.ab1(4>284) → CACGACACAGAAAGGAA-CCTCACGGTGGACACCTATGTGAATGAGAATGGCGAGAAAATCACCAGTATGACCCACG
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → CACGACACAGAAAGGAA-CCTCACGGTGGACACCTATGTGAATGAGAATGGCGAGAAAATCACCAGTATGACCCACG
 ChT4_F720_2017-05-15_E01.ab1(75>316) → CACGACACAGAAAGGAA-CCTCACGGTGGACACCTATGTGAATGAGANTGGCGAGAAAATCACCAGTATGACCCACG
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → CACGACACAGAAAGGAA-CCTCACGGTGGACACCTATGTGAATGAGAATGGCGAGAAAATCACCAGTATGACCCACG
 ancT4_F720_2017-05-10_E01.ab1(55>381) → CACGACACAGAAAGGAA-CCTCACGGTGGACACCTATGTGAATGAGAATGGCGAGAAAATCACCAGTATGACCCACG

1010 1020 1030 1040 1050 1060 1070

ACTCTGTCTGGAATTTCCATGTGTGGA-CGGATGCCTGG-ATGAAGC-GACCGG-ATCTGCCCAAGGGCTACGACGG

Top

S V W N F H V W T D A W M K R P D L P K G Y D G
 T L S G I S M C G R M P G • S D R I C P R A T T A
 L C L E F P C V D G C L D E A T G S A Q G L R R

Human TGM4 Seq.seq(1>2073) → ACTCTGTCTGGAATTTCCATGTGTGGA-CGGATGCCTGG-ATGAAGC-GACCGG-ATCTGCCCAAGGGCTACGACGG
 HuT4_F1020_2017-05-10_A03.ab1(1>326) → ACTCTGTCTGGAATTTCC
 HuT4_F1020_2017-05-10_C03.ab1(4>284) → ACTCTGTCTGGAATTTCCATGTGTGGA-CGGATGCCTGG-ATGAAGC-GACCGG-ATCTGCCCAAGGGCTACGACGG
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → ACTCTGTCTGGAATTTCCATGTGTGGA-CGGATGCCTGG-ATGAAGC-GACCGG-ATCTGCCCAAGGGCTACGACGG
 ChT4_F720_2017-05-15_E01.ab1(75>316) → ACTCTGTCTGGAA
 ChT4_F1020_2017-05-15_G01.ab1(23>298) → TGTGTGGA-CGGATGCCTGG-ATGAAGC-GACCNCN-CTGCCA-GGGCTACAACGG
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → ACTCTGTCTGGAATTTCCATGTGTGGA-CGGATGCCTGG-ATGAAGC-GACCGG-ATCTGCCCAAGGGCTACGACGG
 ancT4_F720_2017-05-10_E01.ab1(55>381) → ACTCTGTCTGGAATTT-CATGTGTGGA-CGGATGCCTGG-ATGAAGC-GACCGG-ATCTGCCCAAGGGCTACNACGG
 ancT4_F1020_2017-05-10_G01.ab1(12>331) → CTGTCTGG-ATTTCCATGTGTGGA-CGGATGCCTGG-ATGAAGC-GNCCCCATCTGCC-AGGGCTACGACGG

1080 1090 1100 1110 1120 1130 1140 1150

CT-GGCAGGCTGTGGACGCAACGCCGCGAGGAGCGAAGCCAGGGTGCTTCTGTTGTGGCCATCACCACCTGACCGCC

Top

W Q A V D A T P Q E R S Q G V F C C G P S P L T A
 G R L W T Q R R R S E A R V S S V V G H H • P P
 L A G C G R N A A G A K P G C L L L W A I T T D R H

Human TGM4 Seq.seq(1>2073) → CT-GGCAGGCTGTGGACGCAACGCCGCGAGGAGCGAAGCCAGGGTGCTTCTGTTGTGGCCATCACCACCTGACCGCC
 HuT4_F1020_2017-05-10_C03.ab1(4>284) → CT-GGCAGGCTGTGGACGC-ACGCCGCGAGGAGCGAAGCCAGGGTGCTTCTGTTGTGGCCATCACCACCTGACCGCC
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → CT-GGCAGGCTGTGGACGCAACGCCGCGAGGAGCGAAGCCAGGGTGCTTCTGCTGTGGCCATCACCACCTGACCGCC
 ChT4_F1020_2017-05-15_G01.ab1(23>298) → CT-GGCAGGCTGTGGACGC-ACGCCGCGAGGAGCGAAGCCAGGGTGCTTCTGCTGTGGCCATCACCACCTGACCGCC
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → CT-GGCAGGCTGTGGACGCAACGCCGCGAGGAGCGAAGCCAGGGTGCTTCTGCTGTGGCCATCACCACCTGACCGCC
 ancT4_F720_2017-05-10_E01.ab1(55>381) → CT-GGCA
 ancT4_F1020_2017-05-10_G01.ab1(12>331) → CT-GGCAGGCTGTGGACGCAACGCCGCGAGGAGCGAAGCCAGGGTGCTTCTGCTGTGGCCATCACCACCTGACCGCC

1160 1170 1180 1190 1200 1210 1220 1230

ATCCGCAAAGGTGACATCTTTATTGTCTATGACACC-AGATTCGTCTTCTCAGAAGTGAATGGTGACAGGCTCATCT

Top

I R K G D I F I V Y D T R F V F S E V N G D R L I W
 S A K V T S L L S M T P D S S S Q K • M V T G S S
 P Q R • H L Y C L • H Q I R L L R S E W • Q A H L

Human TGM4 Seq.seq(1>2073) → ATCCGCAAAGGTGACATCTTTATTGTCTATGACACC-AGATTCGTCTTCTCAGAAGTGAATGGTGACAGGCTCATCT
 HuT4_F1020_2017-05-10_C03.ab1(4>284) → ATCCGCAAAGGTGACATCTTTATTGTCTATGACACC-AGATTCGTCTTCTCAGAAGTGAATGGTGACAGGCTCATCT
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → ATCCGCAAAGGTGACATCTTTATTGTCTATGACACC-AGATTCGTCTTCTCAGAAGTGAATGGTGACAGGCTCATCT
 ChT4_F1020_2017-05-15_G01.ab1(23>298) → ATCCGCAAAGGTGACATCTTTATTGTCTATGATACC-AGATTCGTCTTCTCAGAAGTGAATGGTGACAGGCTCATCT
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → ATCCGCAAAGGTGACATCTTTATTGTCTATGACACC-AGATTCGTCTTCTCAGAAGTGAATGGTGACAGGCTCATCT
 ancT4_F1020_2017-05-10_G01.ab1(12>331) → ATCCGCAAAGGTGACATCTTTATTGTCTATGACACC-AGATTCGTCTTCTCAGAAGTGAATGGTGACAGGCTCATCT

1240 1250 1260 1270 1280 1290 1300

GGTTGGTGAAGATGGTGAATGGG-CAGGAGGAGTTACACGTAATTTCAATGGA-GACCACAAGCATCGGGAAAAA-C

Top

L V K M V N G Q E E L H V I S M E T T S I G K N
 G W • R W • M G R R S Y T • F Q W R P Q A S G K T
 V G E D G E W A G G V T R N F N G D H K H R E K H

Human TGM4 Seq.seq(1>2073) → GGTTGGTGAAGATGGTGAATGGG-CAGGAGGAGTTACACGTAATTTCAATGGA-GACCACAAGCATCGGGAAAAA-C
 HuT4_F1020_2017-05-10_C03.ab1(4>284) → GGTTGGTGAAGATGGTGAATGGG-CAGGAGGAGTTACACGTAATTTCAATGGA
 HuT4_R1620_2017-05-10_E03.ab1(58>281) ← A-GACCACAAGCATCGGGAAAAA-C
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → GGTTGGTGAAGATGGTGAATGGG-CAGGAGGAGTTACACGTAATTTCAATGGA-GACCACAAGCATCGGGAAAAA-C
 ChT4_F1020_2017-05-15_G01.ab1(23>298) → GGTTGGTGAAGATGGTGAATGGG-CAGGAGGAGTTACACGTAATTTCAATGGA-TGGA-GACCACAAGCATCGGGAAAAA
 ChT4_R1620_2017-05-15_A02.ab1(72>289) ← ATTTCAATGGA-GACCACAAGCATCGGGAAAAA-C
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → GGTTGGTGAAGATGGTGAATGGG-CAGGAGGAGTTACACGTAATTTCAATGGA-GACCACAAGCATCGGGAAAAA-C
 ancT4_F1020_2017-05-10_G01.ab1(12>331) → GGTTGGTGAAGATGGTGAATGGG-CAGGAGGAGTTACACGTAATTTCAATGGA-GACCACAAGCATCGGGAAAAA-C
 ancT4_R1620_2017-05-10_A02.ab1(55>314) ← GAATGGG-CAGGAGGAGTTACACGTAATTTCAATGGA-GACCACAAGCATCGGGAAAAA-C

1310 1320 1330 1340 1350 1360 1370 1380

ATCAGC-ACCAAGG-CAGTGGGCCAAGA-CAGG-CGGAGAGA-TATCACC-TATGA-GTACAAGTATCCAGAAGGCT

Top

I S T K A V G Q D R R R D I T Y E Y K Y P E G S
 S A P R Q W A K T G A G E I S P M S T S I Q K A
 Q H Q G S G P R Q A E R Y H L • V Q V S R L

Human TGM4 Seq.seq(1>2073) → ATCAGC-ACCAAGG-CAGTGGGCCAAGA-CAGG-CGGAGAGA-TATCACC-TATGA-GTACAAGTATCCAGAAGGCT
 HuT4_R1620_2017-05-10_E03.ab1(58>281) ← ATCAGC-ACCAAGG-CAGTGGGCCAAGA-CAGG-CGGAGAGA-TATCACC-TATGA-GTACAAGTATCCAGAAGGCT
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → ATCAGC-ACCAAGG-CAGTGGGCCAAGA-CAGG-CGGAGAGA-TATCACC-TATGA-GTACAAGTATCCAGAAGGCT
 ChT4_R1620_2017-05-15_A02.ab1(72>289) ← ATCAGC-ACCAAGG-CAGTGGGCCAAGA-CAGG-CGGAGAGA-TATCACC-TATGA-GTACAAGTATCCAGAAGGCT
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → ATCAGC-ACCAAGG-CAGTGGGCCAAGA-CAGG-CGGAGAGA-TATCACC-TATGA-GTACAAGTATCCAGAAGGCT
 ancT4_F1020_2017-05-10_G01.ab1(12>331) → ATCAGC-AC-AAGG-CTGTGGGCCAAGA
 ancT4_R1620_2017-05-10_A02.ab1(55>314) ← ATCAGC-ACCAAGG-CTGTGGGCCAAGA-CAGG-CGGAGAGA-TATCACC-TATGA-GTACAAGTATCCAGAAGGCT

1390 1400 1410 1420 1430 1440 1450 1460

CCTCTGAGGAGAGGCAGGTCATGGATCATGCCCTTCTCCTTCTCAGTTCTGAGAGGGAGCACAGACGACCTGTAAAA

Top

S E E R Q V M D H A F L L L S S E R E H R R P V K
 P L R R G R S W I M P S S F S V L R G S T D D L • K
 L • G E A G H G S C L P P S O F • E G A O T T C K R

Human TGM4 Seq.seq(1>2073) → CCTCTGAGGAGAGGCAGGTCATGGATCATGCCCTTCTCCTTCTCAGTTCTGAGAGGGAGCACAGACGACCTGTAAAA
 HuT4_R1620_2017-05-10_E03.ab1(58>281) ← CCTCTGAGGAGAGGCAGGTCATGGATCATGCCCTTCTCCTTCTCAGTTCTGAGAGGGAGCACAGACGACCTGTAAAA
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → CCTCTGAGGAGAGGCAGGTCATGGATCATGCCCTTCTCCTTCTCAGTTCTGAGAGGGAGCACAGACGACCTGTAAAA
 ChT4_R1620_2017-05-15_A02.ab1(72>289) ← CCTCTGAGGAGAGGCAGGTCATGGATCATGCCCTTCTCCTTCTCAGTTCTGAGAGGGAGCACAGACGACCTGTAAAA
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → CCTCTGAGGAGAGGCAGGTCATGGATCATGCCCTTCTCCTTCTCAGTTCTGAGAGGGAGCACAGACGACCTGTAAAA
 ancT4_R1620_2017-05-10_A02.ab1(55>314) ← CCTCTGAGGAGAGGCAGGTCATGGATCATGCCCTTCTCCTTCTCAGTTCTGAGAGGGAGCACAGACGACCTGTAAAA

		1470	1480	1490	1500	1510	1520	1530	1540	
		GAGAACTTTCTTCACATGTCGGTACAATCAGATGATGTGCTGCTGGGAAACTCTGTTAATTTACACGTGATTCTTAA								
Top		E N F L H M S V Q S D D V L L G N S V N F T V I L K R T F F T C R Y N Q M M C C W E T L L I S P • F L K E L S S H V G T I R • C A A G K L C • F H R D S •								
Human TGM4 Seq.seq(1>2073)	→	GAGAACTTTCTTCACATGTCGGTACAATCAGATGATGTGCTGCTGGGAAACTCTGTTAATTTACACGTGATTCTTAA								
HuT4_R1620_2017-05-10_E03.ab1(58>281)	←	GAGAACTTTCTTCACATGTCGGTACAATCAGATGATGTGCTGCTGGGAAACTCT								
Chimp TGM4 sequence#16AE75F.seq(1>2073)	→	GAGAACTTTCTTCACATGTCGGTACAATCAGATGATGTGCTGCTGGGAAACCTGTTAATTTACACGTGATTCTTAA								
ChT4_R1620_2017-05-15_A02.ab1(72>289)	←	GAGAACTTTCTTCACATGTCGGTACAATCAGATGATGT								
Anc TGM4 sequence f#16AE75A.seq(7>2088)	→	GAGAACTTTCTTCACATGTCGGTACAATCAGATGATGTGCTGCTGGGAAACCTGTTAATTTACACGTGATTCTTAA								
ancT4_R1620_2017-05-10_A02.ab1(55>314)	←	GAGAACTTTCTTCACATGTCGGTACAATCAGATGATGTGCTGCTGGGAAACCTG								
		1550	1560	1570	1580	1590	1600	1610		
		AAGGAAGACCGCTGCCCTACAGAATGTCAACATCTTTGGGTTCCCTTTGAACTACAGTTGTACACTGGCAAGAAGATGG								
Top		R K T A A L Q N V N I L G S F E L Q L Y T G K K M A G R P L P Y R M S T S W V P L N Y S C T L A R R W K E D R C P T E C Q H L G F L • T T V V H W Q E D G								
Human TGM4 Seq.seq(1>2073)	→	AAGGAAGACCGCTGCCCTACAGAATGTCAACATCTTTGGGTTCCCTTTGAACTACAGTTGTACACTGGCAAGAAGATGG								
Chimp TGM4 sequence#16AE75F.seq(1>2073)	→	AAGGAAGACCGCTGCCCTACAGAATGTCAACATCTTTGGGCTCCTTTGAACTACAGTTGTACACTGGCAAGAAGGTTGG								
Anc TGM4 sequence f#16AE75A.seq(7>2088)	→	AAGGAAGACCGCTGCCCTACAGAATGTCAACATCTTTGGGCTCCTTTGAACTACAGTTGTACACTGGCAAGAAGGTTGG								
		1620	1630	1640	1650	1660	1670	1680	1690	
		CAAAACTGTGTGACCTCAATAAGACCTCGCAGATCCAAGG-TCAAGTATCAGAAGTG-ACTCTGACCTTGGACTCCA								
Top		K L C D L N K T S Q I Q G Q V S E V T L T L D S K Q N C V T S I R P R R S K V K Y Q K • L • P W T P K T V • P Q • D L A D P R S S I R S D S D L G L Q								
Human TGM4 Seq.seq(1>2073)	→	CAAAACTGTGTGACCTCAATAAGACCTCGCAGATCCAAGG-TCAAGTATCAGAAGTG-ACTCTGACCTTGGACTCCA								
HuT4_F1620_2017-05-10_F03.ab1(65>286)	→	ACTCTGACCTTGGACTCCA								
Chimp TGM4 sequence#16AE75F.seq(1>2073)	→	CAAAACTGTGTGACCTCAATAAGACCTCGCAGATCCAAGG-TCAAGTATCAGAAGTG-ACTCTGACCTTGGACTCCA								
ChT4_F1620_2017-05-15_B02.ab1(3>273)	→	AAACTGTGTGACCTCAATAAGACCTCGCAGATCCA-GG-TC-AGTATCAGAAGTG-NCNCTGACCTTGGACTCCA								
Anc TGM4 sequence f#16AE75A.seq(7>2088)	→	CAAAACTGTGTGACCTCAATAAGACCTCGCAGATCCAAGG-TCAAGTATCAGAAGTG-ACTCTGACCTTGGACTCCA								
ancT4_F1620_2017-05-10_B02.ab1(62>362)	→	ACTCTGACCTTGGACTCCA								
		1700	1710	1720	1730	1740	1750	1760	1770	
		AGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAGTTATCAGAGGTTTCATCATTGCGGAAATTGTGGAG								
Top		T Y I N S L A I L D D E P V I R G F I I A E I V E R P T S T A W L Y • M M S Q L S E V S S L R K L W S D L H Q Q P G Y I R • • A S Y Q R F H H C G N C G V								
Human TGM4 Seq.seq(1>2073)	→	AGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAGTTATCAGAGGTTTCATCATTGCGGAAATTGTGGAG								
HuT4_F1620_2017-05-10_F03.ab1(65>286)	→	AGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAGTTATCAGAGGTTTCATCATTGCGGAAATTGTGGAG								

1700 1710 1720 1730 1740 1750 1760 1770
 AGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAGTTATCAGAGGTTTCATCATTGCGGAAATTGTGGAG

Top
 T Y I N S L A I L D D E P V I R G F I I A E I V E
 R P T S T A W L Y • M M S Q L S E V S S L R K L W S
 D L H Q Q P G Y I R • A S Y Q R F H H C G N C G V

Human TGM4 Seq.seq(1>2073) → AGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAGTTATCAGAGGTTTCATCATTGCGGAAATTGTGGAG
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → AGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAGTTATCAGAGGTTTCATCATTGCGGAAATTGTGGAG
 ChT4_F1620_2017-05-15_B02.ab1(3>273) → AGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAGTTATCAGAGGTTTCATCATTGCGGAAATTGTGGAG
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → AGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAGTTATCAGAGGTTTCATCATTGCGGAAATTGTGGAG
 ancT4_F1620_2017-05-10_B02.ab1(62>362) → AGACCTACATCAACAGCCTGGCTATATTAGATGATGAGCCAGTTATCAGAGGTTTCATCATTGCGGAAATTGTGGAG

1780 1790 1800 1810 1820 1830 1840
 TCT-AA-GGAAATCATGGCCTCT-GAAGTATTCACGTCTTTCCAGTACCCTG-AGTTCTCTATAGAGTTGCCTAACA

Top
 S K E I M A S E V F T S F Q Y P E F S I E L P N T
 L R K S W P L K Y S R L S S T L S S L • S C L T
 • G N H G L • S I H V F P V P • V L Y R V A • H

Human TGM4 Seq.seq(1>2073) → TCT-AA-GGAAATCATGGCCTCT-GAAGTATTCACGTCTTTCCAGTACCCTG-AGTTCTCTATAGAGTTGCCTAACA
 HuT4_F1620_2017-05-10_F03.ab1(65>286) → TCT-AA-GGAAATCATGGCCTCT-GAAGTATTCACGTCTTTCCAGTACCCTG-AGTTCTCTATAGAGTTGCCTAACA
 HuT4_HISR_2017-05-10_G03.ab1(61>268) ← GTTGCTAACA
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → TCT-AA-GGAAATCATGGCCTCT-GAAGTATTCACGTCTTTCCAGTACCCTG-AGTTCTCTATAGAGTTGCCTAACA
 ChT4_F1620_2017-05-15_B02.ab1(3>273) → TCT-AA-GGAAATCATGGCCTCT-GAAGTATTCACGTCTTTCCAGTACCCTG-AGTTCTCTATAGAGTTGCCTAACA
 ChT4_HIISR_2017-05-15_C02.ab1(64>299) ← TCCAGTCTTTCCAGTACCCTG-AGTTCTCTATAGAGTTGCCTAACA
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → TCT-AA-GGAAATCATGGCCTCT-GAAGTATTCACGTCTTTCCAGTACCCTG-AGTTCTCTATAGAGTTGCCTAACA
 ancT4_F1620_2017-05-10_B02.ab1(62>362) → TCT-AA-GGAAATCATGGCCTCT-GAAGTATTCACGTCTTTCCAGTACCCTG-AGTTCTCTATAGAGTTGCCTAACA
 ancT4_HISR_2017-05-10_C02.ab1(66>312) ← -GAAGTATTCACGTCTTTCCAGTACCCTG-AGTTCTCTATAGAGTTGCCTAACA

1850 1860 1870 1880 1890 1900 1910 1920
 CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT

Top
 G R I G Q L L V C N C I F K N T L A I P L T D V
 Q A E L A S Y L S A I V S S R I P W P S L • L T S
 R Q N W P A T C L Q L Y L Q E Y P G H F D • L R

Human TGM4 Seq.seq(1>2073) → CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT
 HuT4_F1620_2017-05-10_F03.ab1(65>286) → CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT
 HuT4_HISR_2017-05-10_G03.ab1(61>268) ← CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT
 HuT4_pFlagR_2017-05-10_H03.ab1(66>322) ← TTTGACTGACGT
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT
 ChT4_F1620_2017-05-15_B02.ab1(3>273) → CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT
 ChT4_HIISR_2017-05-15_C02.ab1(64>299) ← CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT
 ancT4_F1620_2017-05-10_B02.ab1(62>362) → CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT
 ancT4_HISR_2017-05-10_C02.ab1(66>312) ← CAGGCAG-AATTGGCC-AGCTACTTGTCTGCAATTGTATCTTCAAGAATACCCTGG-CCATCCC-TTTGACTGACGT

1930 1940 1950 1960 1970 1980 1990 2000

CAAG-TTCTCTTTGGAAAGCCTGGGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-GCAGCCTGGTGAGA

Top

K F S L E S L G I S S L Q T S D H G T V Q P G E T
 S S L W K A W A S P H Y R P L T M G R C S L V R
 Q V L F G K P G H L L T T D L • P W D G A A W • D

Human TGM4 Seq.seq(1>2073) → CAAG-TTCTCTTTGGAAAGCCTGGGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-GCAGCCTGGTGAGA
 HuT4_HISR_2017-05-10_G03.ab1(61>268) ← CAAG-TTCTCTTTGGAAAGCCTGGGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-GCAGCCTGGTGAGA
 HuT4_pFlagR_2017-05-10_H03.ab1(66>322) ← CAAG-TTCTCTTTGGAAAGCCTGGGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-GCAGCCTGGTGAGA
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → CAAG-TTCTCTTTGGAAAGCCTGGGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-GCAGCCTGGTGAGA
 ChT4_HIISR_2017-05-15_C02.ab1(64>299) ← CAAG-TTCTCTTTGGAAAGCCTGGGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-GCAGCCTGGTGAGA
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → CAAG-TTCTCTTTGGAAAGCCTGGGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-GCAGCCTGGTGAGA
 ancT4_F1620_2017-05-10_B02.ab1(62>362) → CAAG-TTCTCTTTGGAA-GCCTGGGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-
 ancT4_HISR_2017-05-10_C02.ab1(66>312) ← CAAG-TTCTCTTTGGAAAGCCTGGGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-GCAGCCTGGTGAGA
 ancT4_pFlagRev_2017-05-10_D02.a(74>329) ← GGCATCTCCTCACT-ACAGACCTCTGACCATGGGACGGT-GCAGCCTGGTGAGA

2010 2020 2030 2040 2050 2060 2070

CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCCAAGAAATTTATCGTCAAGTTAAGTTCCAAAC

Top

I Q S Q I K C T P I K T G P K K F I V K L S S K
 P S N P K • N A P Q • K L D P R N L S S S • V P N
 H P I P N K M H P N K N W T Q E I Y R Q V K F Q T

Human TGM4 Seq.seq(1>2073) → CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCCAAGAAATTTATCGTCAAGTTAAGTTCCAAAC
 HuT4_HISR_2017-05-10_G03.ab1(61>268) ← CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCCAAGAAA
 HuT4_pFlagR_2017-05-10_H03.ab1(66>322) ← CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCCAAGAAATTTATCGTCAAGTTAAGTTCCAAAC
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCCAAGAAATTTATCGTCAAGTTAAGTTCCAAAC
 ChT4_HIISR_2017-05-15_C02.ab1(64>299) ← CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCC
 ChT4_pFlagR_2017-05-15_D02.ab1(145>274) ← CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCC TAAGTTCCAAAC
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCCAAGAAATTTATCGTCAAGTTAAGTTCCAAAC
 ancT4_HISR_2017-05-10_C02.ab1(66>312) ← CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCCAAGA
 ancT4_pFlagRev_2017-05-10_D02.a(74>329) ← CCATCCAATCCC-AAATAAAATGCA-CCCAATAAAAACTGGACCCAAGAAATTTATCGTCAAGTTAAGTTCCAAAC

2080 2090 2100 2110 2120 2130 2140 2150

-AAGTAAAAGAGATTAATGCTCAGAAGATTGTTCTCATCACCAAGCACCATCACCATCACCATTAAcggcgctt

Top

V K E I N A Q K I V L I T K H H H H H • R G R F
 K • K R L M L R R L F S S P S T I T I T I N A A A F
 S E R D • C S E D C S H H O A P S P S P L T R P L

Human TGM4 Seq.seq(1>2073) → -AAGTAAAAGAGATTAATGCTCAGAAGATTGTTCTCATCACCAAGCACCATCACCATCACCATTAA
 HuT4_pFlagR_2017-05-10_H03.ab1(66>322) ← -AAGTAAAAGAGATTAATGCTCAGAAGATTGTTCTCATCACCAAGCACCATCACCATCACCATTAA-----
 Chimp TGM4 sequence#16AE75F.seq(1>2073) → -AAGTAAAAGAGATTAATGCTCAGAAGATTGTTCTCATCACCAAGCACCATCACCATCACCATTAA
 ChT4_pFlagR_2017-05-15_D02.ab1(145>274) ← -AAGTAAAAGAGATTAATGCTCAGAAGATTGTTCTCATCACCAAGCACCATCACCATCACCATTAAACGGCGCCGCTT
 Anc TGM4 sequence f#16AE75A.seq(7>2088) → -AAGTAAAAGAGATTAATGCTCAGAAGATTGTTCTCATCACCAAGCACCATCACCATCACCATTAAACGGCGCCGCT
 ancT4_pFlagRev_2017-05-10_D02.a(74>329) ← -AAGTAAAAGAGATTAATGCTCAGAAGATTGTTCTCATCACCAAGCACCATCACCATCACCATTAAACGGCGCCGCTT

