Sequence diversity of *Pan troglodytes* subspecies and the impact of *WFDC6* selective constraints in reproductive immunity.

Zélia Ferreira, Belen Hurle, Aida M. Andrés, Warren Kretzschmar, Jim Mullikin, Praveen Cherukuri, Pedro Cruz, Mary Katherine Gonder, Anne Stone, Sarah Tishkoff, Willie Swanson, NISC Comparative Sequencing Program, Eric Green, Andrew G. Clark, and Susana Seixas

Supplementary Tables

Supplementary Table S1: Samples sequenced.

Population	Sample numbers						
Pan troglodytes troglodytes (15)	Etrange	Ewake		Kita	Mac	Maya	Mokolo	Nanga-Eboko	Nikita	Papaie	Pauldina	Pecos	Silva	Suzanne	Lada	Jules
Pan troglodytes ellioti (31)	Achidi	Akwaya-Bernadette	Jean	Alex	Bally	Bankim	Basho	Carlos	Eve	Gah	George	Jack	Jacob	Julie	Kopongo	Louisa	Margaret	Mesange	Nemo	Nicolene	Papa	Paquita	TKC	Taweh	Tinto-Mbu	Tobi	Bana	Bergkamp	Koto	Damian	Banyo
Pan troglodytes verus (22)	Pt100	Pt101	Pt102	Pt103	Pt104	Pt105	Pt106	Pt107	Pt112	Pt115	Pt117	Pt120	Pt121	Pt122	Pt124	Pt125	Pt126	Pt81	Pt87	Pt88	Pt97	Pt98							

Supplementary Table S2: Regions of the genome sequenced.

Chr 20 coordinates*	Gene
	WFDC-CEN
43,171,507-43,177,217	*WFDC5* (WAP four-disulfide core domain 5)
43,185,481-43,186,520	*WFDC12* (WAP four-disulfide core domain 12)
43,236,912-43,238,599	*PI3* (elafin)
43,269,088-43,271,823	*SEMG1* (semenogelin 1)
43,269,088-43,286,513	*SEMG2* (semenogelin 2)
43,314,293-43,316,620	*SLPI* (secretory leukocyte peptidase inhibitor)
	WFDC-TEL
43,541,899-43,543,586	*WFDC2* (WAP four-disulfide core domain 2)
43,574,515-43,577,678	*SPINT3*
43,596,250-43,601,548	*WFDC6* (WAP four-disulfide core domain 6)
43,602,679-43,609,442	*SPINLW1* (WAP four-disulfide core domain 7)
43,613,815-43,641,379	*WFDC8* (WAP four-disulfide core domain 8)
43,669,992-43,693,321	*WFDC9* (WAP four-disulfide core domain 9)
43,691,799-43,693,245	*WFDC10a* (WAP four-disulfide core domain 10a)
43,710,616-43,732,292	*WFDC11* (WAP four-disulfide core domain 11)
43,746,704-43,767,072	*WFDC10b* (WAP four-disulfide core domain 10b)
43,764,069-43,770,870	*WFDC13* (WAP four-disulfide core domain 13)
43,784,402-43,787,749	*SPINT4*
43,836,254-43,853,954	*WFDC3* (WAP four-disulfide core domain 3)

[#]Control Regions (Andrés, et al. 2010)

Pseudogene ID	Chrromosome coordinate	Gene and chr of origin	Processed Pseudogenes per genome			
			Human	Chimpanzee	Orangutan	Rhesus
ENCODE Pseudogene:79794	chr11:5505093-5505962	Unknown	1	1	1	1
NON-ENCODE						
Pseudogene:89	chr1:38020719-38021144	C14ORF138 @ Chr14	1	1	1	0
Pseudogene:127271	chr1:224666087-224666538	Unknown	1	3	2	4
Pseudogene:511	chr1:181191136-181191557	NCR1 @ Chr19	1	1	1	1
Pseudogene:716	chr1:245414326-245414818	PEX5 @ Chr12	1	1	1	1
Pseudogene:189	chr1:61892410-61893166	SPR @ Chr2	1	1	1	1
Pseudogene:414	chr1:156412265-156414162	ELL2 @ Chr5	3	3	3	3

Pseudogene:725	chr2:7794021-7794677	PSMB1	@	Chr6	1	1	ORF	2?		
Pseudogene:881	chr2:76338812-76339387	NP	@	Chr14	1	1	1	1 + ORF		
Pseudogene:919	chr2:101503725-101504888	PRCP	@	Chr11	1	1	1	1		
Pseudogene:10377	chr2:146951393-146952148	Unknown			1	1	1	1		
Pseudogene:10800	chr2:158526147-158527646	MTA3	@	Chr2	1	1	1	ND		
Pseudogene:1317	chr3:34889168-34890350	FECH	@	Chr18	1	1	1	1		
Pseudogene:1390	chr3:59496630-59497181	Unknown			2	2	2	2		
Pseudogene:131502	chr3:81498830-81499811	Unknown			1	1	1	ND		
Pseudogene:1588	chr3:146885050-146885571	GM2A	@	Chr5	2	2	2	2		
Pseudogene:75498	chr4:36184061-36186883	FBXO38	@	Chr5	1	1	1	1		
Pseudogene:1886	chr4:89667737-89668341	CD53	@	Chr1	1	1	1	1		
Pseudogene:18262	chr4:176732748-176734338	ADAM29	@	Chr4	2	2	2	2		
Pseudogene:1835	chr4:68730627-68731081	Unknown			1	1	1	1		
Pseudogene:2117	chr5:29141432-29142101	C14ORF45	@	Chr14	1	1	1	1		
Pseudogene:19453	chr5:133084092-133084524	DPH4	@	Chr11	1	1	1	1		
Pseudogene:2934	chr6:139701013-139702410	DNAJC7	@	Chr17	2	2	2	3		
Pseudogene:3153	chr7:55279435-55280868	SLC19A3	@	Chr2	2	2	2	2		
Pseudogene:20740	chr7:128120741-128121217	IMP3	@	Chr15	2	2	2	2		
Pseudogene:3607	chr8:60632574-60632998	NUDT15	@	Chr13	2	2 + 1 ORF	2	2		
Pseudogene:3693	chr8:97208019-97209206	Unknown			1	1	1	1		
Pseudogene:69522	chr9:6185935-6186853	GTF3A	@	Chr13	5	5	5	5		
Pseudogene:21854	chr9:21685170-21686842	KHSRP	@	Chr19	1	1	1	1		
Pseudogene:4067	chr9:106130426-106131316	WDR45L	@	Chr17	3	2 + 1 ORF	2 + 1 ORF	2		
Pseudogene:4320	chr10:79498396-79499304	GNAI2	@	Chr3	3	3	2	1		
Pseudogene:4547	chr11:40061840-40062643	ZCCHC9	@	Chr5	3	2	3	3		
Pseudogene:4837	chr11:113829007-113830057	CCRN4L	@	Chr4	1	1	1	2		
Pseudogene:4765	chr11:91708853-91709310	NDUFB11	@	ChrX	1	1	1	1 + 1 ORF		
Pseudogene:4913	chr12:7650698-7651196	CLTA	@	Chr9	2	2	2	2		
Pseudogene:5244	chr12:107199014-107200142	APOBEC3F	@	Chr22	1	1	1	1		
Pseudogene:5055	chr12:44124847-44125485	MESDC2	@	Chr15	1	1	1	1		
Pseudogene:73106	chr13:24217272-24218374	IRX1	@	Chr5	1	1	1	1		
Pseudogene:5504	chr13:89442087-89443147	PEX12	@	Chr17	1	1	1	0		
Pseudogene:5477	chr13:74300713-74302220	RIOK3	@	Chr18	1	1	1	1		
Pseudogene:5700	chr14:61368336-61368840	COX4I1	@	Chr16	2	1	2	2 + 1 ORF		
Pseudogene:62081	chr14:88646206-88647320	MPPE1	@	Chr18	2	1	1	1		
Pseudogene:6310	chr16:82710324-82711542	PLK-1	@	Chr16	2	2	2	2		
Pseudogene:6549	chr17:66136577-66137001	Unknown			1	1	1	1		
Pseudogene:26649	chr19:18012148-18012866	APOA1BP	@	Chr1	2	2	2	2		
Pseudogene:6997	chr20:21429964-21430620	GSTM3	@	Chr1	1	1	1	2		
Pseudogene:7045	chr20:37391075-37391973	ATG3	@	Chr3	1	2	1	1		
Pseudogene:7112	chr21:15051781-15052541	POLR2C	@	Chr16	1	1	1	1		
Pseudogene:7176	chr21:45316169-45316649	Unknown			1	1	1	1		

* Chromosome coordinates are based in the human march 2006 assembly (hg18; genome.ucsc.edu)

[#] These loci were selected using the following filters: Processed pseudogenes (according to the annotation in pseudogene.org); Minimum length of 400 bp; Unlinked to each other; ot OR, not

ribosomal RNA (according to the annotation in pseudogene.org); With human, orangutan and rhesus orthologues; No overlap with UCSC genes; No overlap with highly conserved regions (most conserved, UCSC); Average genome recombination rate; Processed pseudogenes in single copy in human and chimpanzee genomes were preferred. (When this was not possible, one pseudogene per family was accepted provided that the members of the family had 90% identity or less among them.)

Supplementary Table S3: Coding substitutions in *WFDC* genes.

a) Nonsynonymous substitutions

SNP location in chr20 (Pantro2)	Protein	Frequency	Expected Residue	Found Residue	Expected Codon	Found Codon	SIFT (Kumar, et al. 2009)	PolyPhen (Adzhubei, et al. 2010)
42453065	WFDC5	0.06	R	H	CGC	CAC	Tolerated	Benign
42453208	WFDC5	0.04	V	I	GTC	ATC	Tolerated	Possibly Damaging
42453271	WFDC5	0.02	V	M	GTG	ATG	Tolerated	Possibly Damaging
42466176	WFDC12	0.04	W	R	TGG	CGG	Damaging	Benign
42466200	WFDC12	0.04	S	P	TCA	CCA	Damaging	Benign
42466206	WFDC12	0.04	D	H	GAT	CAT	Damaging	Probably Damaging
42466400	WFDC12	0.01	K	N	AAG	AAT	Tolerated	Probably Damaging
42466463	WFDC12	0.02	D	E	GAC	GAA	Tolerated	Damaging
42466492	WFDC12	0.02	D	N	GAT	AAT	Tolerated	Benign
42466510	WFDC12	0.04	V	I	GTA	ATA	Tolerated	Benign
42519166	PI3	0.02	T	M	ACG	ATG	Tolerated	Benign
42520083	PI3	0.12	V	I	GTT	ATT	Tolerated	Possibly Damaging
42568115	SEMG2	0.02	G	D	GGT	GAT	Affect Protein Funcion	Probably Damaging
42568172	SEMG2	0.05	G	V	GGC	GTC	Tolerated	Possibly Damaging
42568204	SEMG2	0.02	H	D	CAT	GAT	Tolerated	Benign
42568441	SEMG2	0.02	A	T	GCT	ACT	Tolerated	Possibly Damaging
42568613	SEMG2	0.01	S	N	AGC	AAC	Tolerated	Possibly Damaging
42568780	SEMG2	0.01	H	Y	CAT	TAT	Tolerated	Probably Damaging
42568831	SEMG2	0.02	K	E	AAG	GAG	Tolerated	Probably Damaging
42568883	SEMG2	0.02	K	M	AAG	ATG	Tolerated	Probably Damagin
42568888	SEMG2	0.02	H	Y	CAT	TAT	Tolerated	Probably Damaging
42569052	SEMG2	0.01	E	D	GAG	GAT	Tolerated	Probably Damaging
42569213	SEMG2	0.07	I	T	ATT	ACT	Tolerated	Benign
42569468	SEMG2	0.02	R	P	CGA	CCA	Tolerated	Probably Damaging
42569698	SEMG2	0.02	H	Y	CAT	TAT	Tolerated	Benign
42826868	WFDC2	0.02	S	L	TCG	TTG	Tolerated	Possibly Damaging
42826905	WFDC2	0.04	S	R	AGC	AGG	Tolerated	Benign
42897365	WFDC6	0.01	C	R	TGT	CGT	Tolerated	Probably Damaging
42897376	WFDC6	0.02	R	H	CGT	CAT	Tolerated	Benign
42897403	WFDC6	0.03	E	G	GAA	GGA	Tolerated	Benign
42897449	WFDC6	0.01	E	K	GAA	AAA	Tolerated	Benign
42897457	WFDC6	0.07	V	G	GTG	GGG	Tolerated	Probably Damaging
42898710	WFDC6	0.01	I	V	ATC	GTC	Tolerated	Benign
42972808	WFDC9	0.03	I	T	ATT	ACT	Torerated	Possibly Damaging
42992587	WFDC10A	0.09	T	P	ACT	CCT	Affect Protein Funcion	Benign

42992632	WFDC10A	0.01	Q	K	CAG	AAG	Torelated Affect Protein Funcion	Possibly Damaging
42993698	WFDC10A	0.01	C	R	TGT	CGT		Probably Damaging
43044472	WFDC10B	0.01	C	*	TGT	TGA		
43044552	WFDC10B	0.02	L	V	CTA	GTA	Torelated Affect Protein Funcion	Benign
43044585	WFDC10B	0.01	I	V	ATC	GTC		Benign
43045539	WFDC10B	0.07	R	C	CGT	TGT	Tolerated	Possibly Damaging
43045602	WFDC10B	0.08	P	T	CCC	ACC	Tolerated	Benign
43071581	WFDC10B	0.02	R	T	AGG	ACG	Affect Protein Function	Probably Damaging
43142946	WFDC3	0.01	E	K	GAA	AAA	Tolerated	Benign
43142988	WFDC3	0.02	K	E	AAA	GAA	Tolerated	Benign
43143039	WFDC3	0.02	C	S	TGT	AGT	Affect Protein Function	Probably Damaging
43155352	WFDC3	0.44	S	F	TCT	TTT	Tolerated	Benign
43155425	WFDC3	0.45	T	P	ACT	CCT	Tolerated	Benign
43155466	WFDC3	0.20	P	L	CCT	CTT	Tolerated	Benign
43155467	WFDC3	0.03	P	S	CCT	TCT	Affect Protein Function	Benign

b) Non-synonymous chimpanzee specific fixed differences

Protein	Position	Human	Chimp	Orang	Macaque	SIFT (Kumar, et al. 2009)	PolyPhen (Adzhubei, et al. 2010)
WFDC12	G27D	C	T	C	C	Tolerated	Benign
SEMG2	A93V	C	T	C	C	Tolerated	Benign
SEMG2	G101D	G	A	G	G	Tolerated	Benign
SEMG2	K120E	A	G	A	A	Tolerated	Benign
SEMG2	H136Y	C	T	C	C	Tolerated	Benign
SEMG2	S232G	A	G	A	A	Tolerated	Benign
SEMG2	H401R	A	G	A	A	Tolerated	Benign
SEMG2	H461R	A	G	A	A	Tolerated	Benign
SEMG2	T485S	A	T	A	A	Affect Protein Function	Benign
SEMG2	G504D	G	A	G	G	Tolerated	Benign
SLPI	L13F	G	A	G	G	Tolerated	Benign
WFDC6	P34S	G	A	G	G	Tolerated	Benign
EPPIN	K79E					Affect Protein Function	
WFDC8	L207P	A	G	?	A	Tolerated	Benign
WFDC8	D64E	G	T	G	A	Tolerated	Benign
WFDC8	S30C	G	C	G	G	Tolerated	Benign
WFDC9	F31L	A	G	A	T	Tolerated	Benign
WFDC10B/13	L16S	A	G	A	A	Tolerated	Benign

c) Synonymous substitutions for the *WFDC* genes.

SNP location	Protein	Frequency	Residue	Expected	Found

in chr20 (Pantro2)				Codon	Codon
42452973	WFDC5	0.44	R	CGG	AGG
42452992	WFDC5	0.45	S	AGC	AGT
42453007	WFDC5	0.2	K	AAG	AAA
42453097	WFDC5	0.02	V	GTG	GTA
42453292	WFDC5	0.21	L	CTA	TTA
42466156	WFDC12	0.45	G	GGC	GGT
42466499	WFDC12	0.14	F	TTC	TTT
42520145	PI3	0.02	G	GGT	GGA
42569232	SEMG2	0.13	E	GAG	GAA
42569682	SEMG2	0.02	Q	CAG	CAA
42826908	WFDC2	0.05	A	GCG	GCC
42836863	WFDC2	0.03	L	CTC	CTT
42921553	WFDC8	0.12	S	AGC	AGT
42993772	WFDC10A	0.01	I	ATC	ATT
43010530	WFDC11	0.02	T	ACC	ACT
43010584	WFDC11	0.01	R	AGG	AGA
43071581	WFDC13	0.02	S	TCC	TCG
43072457	WFDC10B	0.05	S	AGC	AGT
43156519	WFDC3	0.01	L	CTG	CTA
43156543	WFDC3	0.09	P	CCC	CCT

Supplementary Table S.4: 2.5 percentile resulting from 100000 coalescent simulations using ms, under three demographic models (Hudson 2002).

Gene	Sub-species	S	π (10^{-4})	Length	Θ_w	Tajima's *D*	Fu & Li *D**	Fay and Wu's *H*	P(HKA)
WFDC5	*Pan troglodytes*	45	7.7796	5536	8.25	-0.1739	-0.1817	1.4109	0.6746
WFDC12	*Pan troglodytes*	24	2.4808	1323	4.39	-1.2459	-1.0808	0.7227	0.6637
PI3	*Pan troglodytes*	32	4.3176	3377	5.85	-0.7764	-0.3782	6.5226	0.4454
SEMG1	*Pan troglodytes*	45	3.9938	3305	8.2	-1.5677	0.3357	-1.0394	0.1462
SEMG2	*Pan troglodytes*	47	3.5766	4324	8.57	-1.7859	-0.352	5.0682	0.8163
SLPI	*Pan troglodytes*	46	6.6832	4709	8.38	-0.6211	0.6394	7.6619	0.247
WFDC2	*Pan troglodytes*	34	5.3535	3984	6.2	-0.4058	-0.2482	0.656	0.1038
SPINT3	*Pan troglodytes*	44	5.2971	3727	8.04	-1.0425	-1.5881	36.879	0.0084
WFDC6	*Pan troglodytes*	31	1.6333	2807	5.67	-2.1039	-2.8978	-1.5222	0.0131
WFDC7	*Pan troglodytes*	38	3.165	3233	6.96	-1.648	-2.0926	16.8892	0.0175
WFDC8	*Pan troglodytes*	56	4.8159	7179	10.2	-1.6397	-2.1769	6.131	0.1153
WFDC9/10A	*Pan troglodytes*	67	7.2409	6863	12.2	-1.2844	-1.8648	19.3656	0.3861
WFDC11	*Pan troglodytes*	71	10.094	5037	13	-0.7082	-0.5289	25.782	0.0145
WFDC10B/13	*Pan troglodytes*	59	6.8495	7365	10.8	-1.1309	-1.54	-4.1285	0.1052
SPINT4	*Pan troglodytes*	30	2.2099	3527	5.5	-1.7644	-1.9467	16.0833	0.1734
WFDC3	*Pan troglodytes*	93	10.64	7572	17	-1.1874	-1.0211	17.8464	0.0719

Supplementary Table S.5: Parameter Estimates and Likelihood Scores under Different Branch models (Yang 1997; Zhang, et al. 2005).

	Parameters for Branches	**Likelihoods**
One ratio	$\omega_{EPPIN\text{-}WFDC6} = 0.4739$	-1388.2014
Two ratios	$\omega_{EPPIN} = 0.4738$	-1387.6820
	$\omega_{WFDC6} = 0.7782$	
Three ratios	$\omega_{EPPIN} = 0.4739$	-1387.5959
	$\omega_{WFDC6others} = 0.7091$	
	$\omega_{WFDC6ancHomoPan} = 1.1656$	
Three ratios	$\omega_{EPPIN} = 0.4738$	-1387.6534
	$\omega_{WFDC6\ others} = 0.8119$	
	$\omega_{WFDC6Pan} = 0.5891$	

Supplementary Tables References

Adzhubei IA, et al. 2010. A method and server for predicting damaging missense mutations. Nature methods 7: 248-249. doi: 10.1038/nmeth0410-248

Andrés AM, et al. 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. PLoS Genetics 6: e1001157. doi: 10.1371/journal.pgen.1001157

Fay JC, Wyckoff GJ, Wu C-I 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415: 3.

Fu Y-X, Li W-H 1993. Statistical Tests of Neutrality of Mutations. Genetics 133: 15.

Hudson RR 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Hudson RR, Kreitman M, Aguadé M 1987. A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics 116: 6.

Kumar P, Henikoff S, Ng PC 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4: 1073-1082. doi: DOI 10.1038/nprot.2009.86

Tajima F 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Watterson GA 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256-276.

Wegmann D, Excoffier L 2010. Bayesian inference of the demographic history of chimpanzees. Molecular Biology and Evolution 27: 1425-1435. doi: 10.1093/molbev/msq028

Yang Z 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.

Zeng K, Fu YX, Shi S, Wu CI 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431-1439. doi: genetics.106.061432 [pii]

10.1534/genetics.106.061432

Zhang J, Nielsen R, Yang Z 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular Biology and Evolution 22: 2472-2479. doi: 10.1093/molbev/msi237