Supplementary Material

Evolution and diversity of copy number variation in the great ape lineage

Peter H. Sudmant, John Huddleston, Claudia R. Catacchio, Maika Malig, LaDeana W. Hillier, Carl Baker, Kiana Mohajeri, Ivanela Kondova, Ronald E. Bontrop, Stephan Persengiev, Francesca Antonacci, Mario Ventura, Javier Prado-Martinez, Tomas Marques-Bonet, and Evan E. Eichler

Additional resources available online: **http://eichlerlab.gs.washington.edu/greatape-cnv**

Section 1: Genome mapping and quality control

Genome mapping for read-depth-based copy number analysis: Genomes were mapped to the human reference assembly Build 36 (UCSC HG18) using the mrsfastc mapping software(Sudmant et al. 2010; Hach et al. 2010). Reads were first divided into their 36 bp constituents and mapped with a maximum edit distance of 2 to a repeat-masked reference. Subdividing reads into their 36 bp constituents served firstly to allow us to align reads with an ungapped alignment algorithm (mrsfastc) quickly. Additionally, subdividing reads served to normalize the varied read lengths between genomes. Masking was performed with RepeatMasker (default UCSC masking version 3 Repeat Masking) and Tandem Repeats Finder(Marques-Bonet et al. 2009; Benson 1999; Cheng et al. 2005) using the following parameters: *match 2, mismatch 3, delta 5, PM 80, PI 10, minscore 30, maxperiod 1000*. Masked

regions were extended out by 36 bp on either side after mapping to eliminate mapping edge effects adjacent to masked regions.

Read-depth-based GC correction and copy number prediction: Copy numbers were estimated in tiled 500 bp, 1 kbp, and 3 kbp windows of unmasked sequence across the genome of each individual as described in Sudmant *et al* 2010(1000 Genomes Project Consortium et al. 2012; Sudmant et al. 2010). First, a GC correction step was performed to eliminate biases in sequence coverage introduced during library construction and associated with the GC content of loci. For each individual a multiplicative GC correction factor was calculated and applied to the genome. This correction factor was calculated by determining the average read-depth across putative invariant diploid regions of the human genome (potentially copy number variable loci subtracted from HG18, namely, the Database of Genomic Variants, gaps, segmental duplications (SDs), and copy number variants identified by Conrad *et al*(Meyer et al. 2012; Conrad et al. 2010) binned by GC content as computed in 401 base-pair windows across the genome. A correction factor k_{GC} was then calculated as $k_{GC}=\mu_{total}/\mu_{GC}$ where μ_{total} is the total average read-depth and μ_{GC} is the read-depth within a particular GC content bin. The corrected read-depth at a base *x* where *d(x)* is the read-depth at base *x* is then calculated as $d'(x) = d(x)*k_{GC}$ where k_{GC} is chosen to match the particular GC content at base *x.* Copy numbers were then estimated by using a linear regression model fit to regions of known copy number 2 in all primates.

Genome-wide copy number 2 assessment: Read-depth-based copy number estimates are based on the assumption that the number of reads shotgun sequenced over a particular locus will be directly proportional to the copy number of that locus. As such, extreme biases introduced during sequencing and library construction or resulting from DNA degradation may reduce the power of this technique even after corrections are applied. To test the quality of each of our genomes, we focused on 4836 regions encompassing ~1.1 Gbp of sequence, which we predict to be largely devoid of any structural variation and thus fixed at copy number 2. These loci were selected by removing from the human reference genome regions of known or likely structural variation, including the Database of Genomic Variation, genomic gaps, SDs, and variants detected by Conrad *et al*(Miller et al. 1993; Conrad et al. 2010). We then further eliminated any regions <100 kbp in length to avoid over-fragmenting. The resulting set of loci should largely be fixed at copy number 2 and invariant among all human genomes. Among different species of nonhuman primates, some these loci may indeed be copy number variable.

However, we can assume that not only will the number of such variable sites will be low but that among individuals of the same species the same number of regions will be variable and therefore the total number of invariant loci within a species should be consistent. We thus estimated the copy number of these loci in tiled 3 kbp windows of unmasked sequence. 131242 such windows were considered within each individual genome (**Figure 1.1**). Additionally, the proportion of loci correctly predicted as copy number 2 as a function of the GC content of a region was also quantified to assess the effect of library construction induced GC biases (**Figure 1.2**).

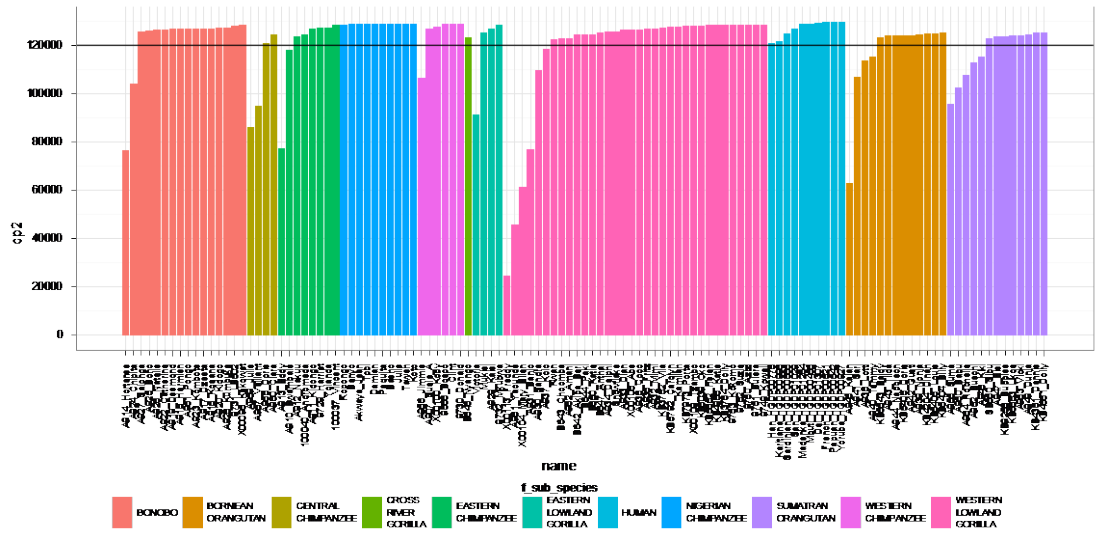


Figure 1.1: 131,242 3 kbp windows encompassing of 1.1 Gbp of putative diploid invariant sequence were analyzed in each sequenced genome. The number of 3 kbp windows correctly determined as copy number 2 is displayed for each genome sorted and colored by species and subspecies. The black line indicates a cutoff of 120,000, below which corresponds to genomes with poor copy number estimation across the entire spectrum of GC content. Genomes failing this cutoff are plotted with dotted lines in Figure 1.2 below.

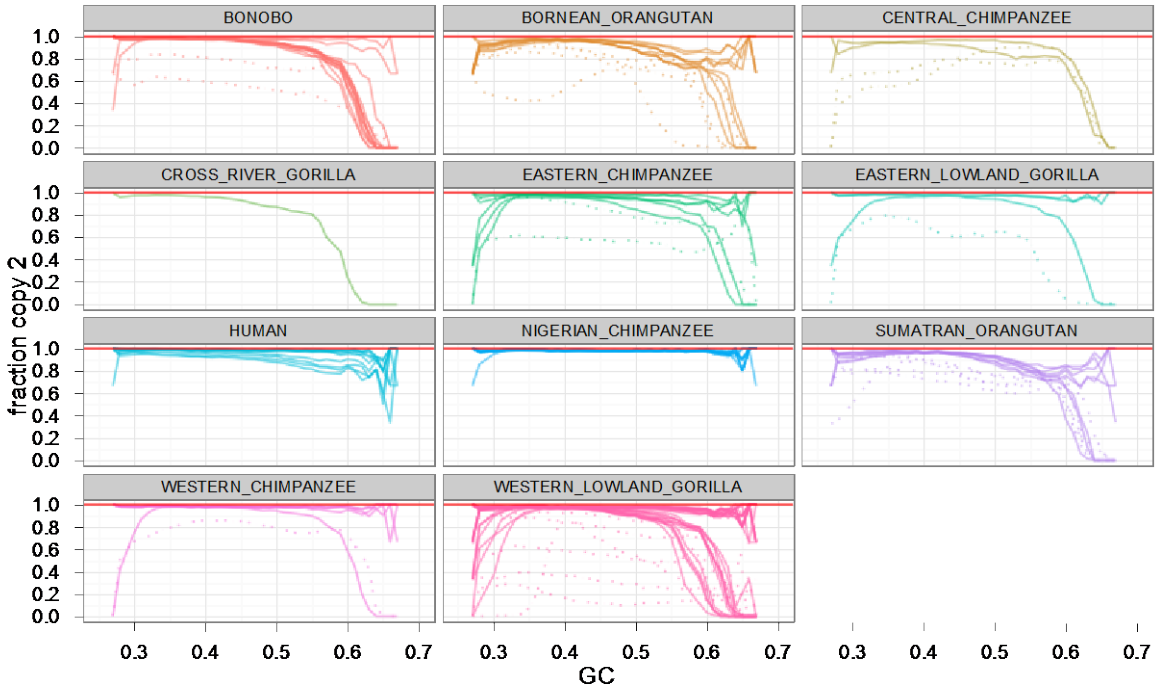


Figure 1.2: The fraction of 3 kbp windows correctly determined to be copy number 2 is plotted as a function of GC content grouped and colored by subspecies. Regions of increased GC content are expected to have reduced coverage and increased variance and, thus, are more difficult to accurately assay. Reduced power to accurately estimate copy

number across the entire spectrum of GC content is, however, an indication of a genome that has been subject to multiple confounding factors during sequencing or poor starting DNA quality. Genomes exhibiting a trend of reduced copy number prediction power across the entire spectrum of GC content were discarded and are plotted with dotted lines. These genomes correspond to those falling below the cutoff threshold indicated in Figure 1.1 and are listed in Table 1.1.

Table 1.1: Individuals filtered out by copy 2 region analyses (23).

name	sub_species	fraction of diploid regions correctly estimated
A914_Hortense	BONOBO	0.59
A924_Chipita	BONOBO	0.80
A959_Julie	CENTRAL_CHIMPANZEE	0.66
A957_Vaillant	CENTRAL_CHIMPANZEE	0.73
A913_Eddy	EASTERN_CHIMPANZEE	0.59
A910_Bwambale	EASTERN_CHIMPANZEE	0.90
A956_Jimmie	WESTERN_CHIMPANZEE	0.81
Victoria	EASTERN_LOWLAND_GORILLA	0.70
X00113_Freddy	WESTERN_LOWLAND_GORILLA	0.19
A937_Kolo	WESTERN_LOWLAND_GORILLA	0.90
X00104_Mary_Ellen	WESTERN_LOWLAND_GORILLA	0.47
A931_Banjo	WESTERN_LOWLAND_GORILLA	0.59
A961_Yaounde	WESTERN_LOWLAND_GORILLA	0.35
A930_Sandra	WESTERN_LOWLAND_GORILLA	0.84
A941_Sari	BORNEAN_ORANGUTAN	0.82
A938_Lotti	BORNEAN_ORANGUTAN	0.87
A940_Temmy	BORNEAN_ORANGUTAN	0.88
A946_Kajan	BORNEAN_ORANGUTAN	0.48
A952_Buschi	SUMATRAN_ORANGUTAN	0.86
A948_Kiki	SUMATRAN_ORANGUTAN	0.88
A955_Suma	SUMATRAN_ORANGUTAN	0.73
A964_Rochelle	SUMATRAN_ORANGUTAN	0.82
A950_Babu	SUMATRAN_ORANGUTAN	0.78

We previously reported that at regions of increased GC content, Illumina-sequenced genomes often demonstrate reduced and more variable depth-of-coverage(Olson 1999; Sudmant et al. 2010; Alkan et al. 2009). To a certain degree this can be corrected for and additionally only a small fraction of the human genome lies within these high GC regions; >90% of the human genome is at <50% GC **(Figure 1.3)**. 4.2%, 1.7% and 0.056% of the genome are at >55%, >60% and >65% GC content, respectively. Our analysis of those 1.1 Gbp of putative copy number 2 loci among all of the genomes indicated, however, that a small fraction of these genomes displayed reduced power to accurately estimate copy number across the entire spectrum of GC content **(Figure 1.2** dashed lines, **Table 1.1)**. The genomes demonstrating this

trend corresponded to those for which the fewest number of copy number 2 loci were accurately assessed **(Figure 1.1)** and thus a cutoff of correctly estimating the copy of 120,000/131,242 windows (91.4%) was selected to flag any genomes of reduced quality due to processes in the library construction and sequencing or perhaps in the quality of the DNA sequenced. These genomes (n=23) were discarded from the read-depth-based analyses of copy number and SDs. Additionally, regions of extreme GC content (>0.57% GC corresponding to 69,119,920 bp, 2.23%) were masked from the analysis to avoid regions of potential biased sequencing.

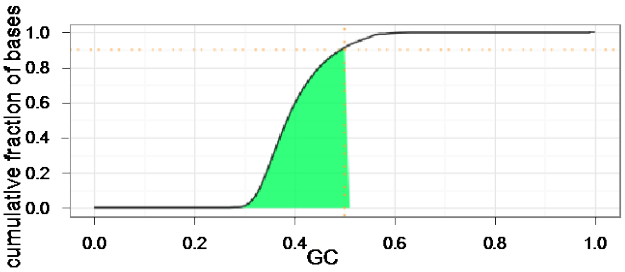


Figure 1.3: A curve of the fraction of base pairs represented in the human genome cumulatively as a function of GC content. >90% of the human genome is represented at 50% GC or less as exhibited in the shaded green portion below the curve.

Genomes passing quality-control dataset: In total, 97 genomes were assessed in this study, including 10 human genomes and the genome of an archaic hominid Denisovan individual (referred to as Homo Denisova here). 75 of the individuals assessed here were sequenced as part of the Great Ape Genome Project—of which 13 showed low levels of contamination (Prado/Sudmant 2012, under review, **Table 1.2**). Ten of the individuals demonstrated <2% contamination while the remaining three showed <4% contamination. Our read-depth-based copy number prediction methodology is highly robust to such low levels of contamination as read-depth typically fluctuates with a mean and variance proportional to the underlying copy number and sequencing coverage of the sample(Brawand et al. 2011; Scally et al. 2012; Sudmant et al. 2010; Ventura et al. 2011; Meyer et al. 2012; Locke et al. 2011). To ensure no biases were introduced, all analyses which may be sensitive to cross-species contamination were performed/confirmed in the subset of non-contaminated individuals.

Table 1.2: Individuals assessed in this study. In addition to 75 genomes sequenced as part of the Great Ape Diversity project, 22 previously published samples were analyzed(Siepel et al. 2005; Scally et al. 2012; Ventura et al. 2011; Meyer et al. 2012; Locke et al. 2011)**.**

Species/sub-species	individual	Source	estimated % contamination
Gorilla beringei graueri	9732_Mkubwa	GAGP	
Gorilla beringei graueri	A929_Kaisi	GAGP	0.2
Gorilla beringei graueri	Mukisi	GAGP	
Gorilla gorilla diehli	B646_Nyango	GAGP	
Gorilla gorilla gorilla	9749_Kowali	GAGP	
Gorilla gorilla gorilla	9750_Azizi	GAGP	
Gorilla gorilla gorilla	9751_Bulera	GAGP	
Gorilla gorilla gorilla	9752_Suzie	GAGP	
Gorilla gorilla gorilla	9753_Kokomo	GAGP	
Gorilla gorilla gorilla	A932_Mimi	GAGP	0.093
Gorilla gorilla gorilla	A933_Dian	GAGP	
Gorilla gorilla gorilla	A934_Delphi	GAGP	0.06
Gorilla gorilla gorilla	A935_Fritz	GAGP	3.8
Gorilla gorilla gorilla	A936_Coco	GAGP	
Gorilla gorilla gorilla	A962_Amani	GAGP	
Gorilla gorilla gorilla	B642_Akiba_Beri	GAGP	
Gorilla gorilla gorilla	B643_Choomba	GAGP	
Gorilla gorilla gorilla	B644_Paki	GAGP	
Gorilla gorilla gorilla	B647_Anthal	GAGP	
Gorilla gorilla gorilla	B650_Katie	GAGP	
Gorilla gorilla gorilla	Kamilah	Scally *et al, 2012*	
Gorilla gorilla gorilla	KB3782_Vila	GAGP	
Gorilla gorilla gorilla	KB3784_Dolly	GAGP	
Gorilla gorilla gorilla	KB4986_Katie	GAGP	
Gorilla gorilla gorilla	KB5792_Carolyn	GAGP	
Gorilla gorilla gorilla	KB5852_Helen	GAGP	
Gorilla gorilla gorilla	KB6039_Oko	GAGP	
Gorilla gorilla gorilla	KB7973_Porta	GAGP	
Gorilla gorilla gorilla	Kwan	Ventura *et al, 2011*	
Gorilla gorilla gorilla	Snowflake	GAGP	
Gorilla gorilla gorilla	X00108_Abe	GAGP	
Gorilla gorilla gorilla	X00109_Tzambo	GAGP	
Homo denisova	Denisova_30x	Meyer *et al,* 2012	
Homo sapiens	Dai_HGDP01307	Meyer *et al,* 2012	
Homo sapiens	French_HGDP00521	Meyer *et al,* 2012	
Homo sapiens	Han_HGDP00778	Meyer *et al,* 2012	
Homo sapiens	Karitiana_HGDP00998	Meyer *et al,* 2012	
Homo sapiens	Madenka_HGDP01284	Meyer *et al,* 2012	
Homo sapiens	Mbuti_HGDP00456	Meyer *et al,* 2012	
Homo sapiens	Papuan_HGDP00542	Meyer *et al,* 2012	
Homo sapiens	San_HGDP01029	Meyer *et al,* 2012	
Homo sapiens	Sardinian_HGDP00665	Meyer *et al,* 2012	
Homo sapiens	Yoruba_HGDP00927	Meyer *et al,* 2012	
Pan paniscus	9731_LB502	GAGP	
Pan paniscus	A915_Kosana	GAGP	
Pan paniscus	A917_Dzeeta	GAGP	
Pan paniscus	A918_Hermien	GAGP	
Pan paniscus	A919_Desmond	GAGP	
Pan paniscus	A920_Kidogo	GAGP	
Pan paniscus	A922_Catherine	GAGP	
Pan paniscus	A923_Kombote	GAGP	
Pan paniscus	A925_Bono	GAGP	
Pan paniscus	A926_Natalie	GAGP	
Pan paniscus	A927_Salonga	GAGP	
Pan paniscus	A928_Kumbuka	GAGP	0.6
Pan paniscus	A951_Pongo	GAGP	
Pan paniscus	X00095_Kakowet	GAGP	0.018
Pan troglodytes ellioti	Akwaya_Jean	GAGP	
Pan troglodytes ellioti	Banyo	GAGP	
Pan troglodytes ellioti	Basho	GAGP	
Pan troglodytes ellioti	Damian	GAGP	
Pan troglodytes ellioti	Julie	GAGP	
Pan troglodytes ellioti	Kopongo	GAGP	
Pan troglodytes ellioti	Koto	GAGP	

Pan troglodytes ellioti	Paquita	GAGP	
Pan troglodytes ellioti	Taweh	GAGP	
Pan troglodytes ellioti	Tobi	GAGP	
Pan troglodytes schweinfurthii	100037_Vincent	GAGP	
Pan troglodytes schweinfurthii	100040_Andromeda	GAGP	0.8
Pan troglodytes schweinfurthii	9729_Harriet	GAGP	
Pan troglodytes schweinfurthii	A911_Kidongo	GAGP	
Pan troglodytes schweinfurthii	A912_Nakuu	GAGP	0.8
Pan troglodytes schweinfurthii	Yolanda	GAGP	2
Pan troglodytes troglodytes	A958_Doris	GAGP	
Pan troglodytes troglodytes	A960_Clara	GAGP	
Pan troglodytes verus	9668_Bosco	GAGP	
Pan troglodytes verus	9730_Donald	GAGP	
Pan troglodytes verus	A907_Susie_A	GAGP	3.2
Pan troglodytes verus	Clint	GAGP	
Pan troglodytes verus	X00100_Koby	GAGP	
Pongo abelii	A947_Elsi	GAGP	0.9
Pongo abelii	A949_Dunja	GAGP	
Pongo abelii	A953_Vicki	GAGP	3.8
Pongo abelii	KB4361_Dennis	Locke *et al*, 2011	
Pongo abelii	KB4661_Dolly	Locke *et al*, 2011	
Pongo abelii	KB5883_Likoe	Locke *et al*, 2011	
Pongo abelii	KB9258_Bubbles	Locke *et al*, 2011	
Pongo abelii	SB550_Sibu	GAGP	
Pongo pygmaeus	A939_Nonja	GAGP	
Pongo pygmaeus	A942_Gusti	GAGP	1.2
Pongo pygmaeus	A943_Tilda	GAGP	
Pongo pygmaeus	A944_Napoleon	GAGP	
Pongo pygmaeus	KB4204_Dinah	Locke *et al*, 2011	
Pongo pygmaeus	KB5404_Billy	Locke *et al*, 2011	
Pongo pygmaeus	KB5405_Louis	Locke *et al*, 2011	
Pongo pygmaeus	KB5406_Doris	Locke *et al*, 2011	
Pongo pygmaeus	KB5543_Baldy	Locke *et al*, 2011	

Table 1.3: Coverage of individuals assessed in this study. Coverage calculations are shown for bwa and mrsfast kmer mappings.

indiv	bwa coverage	kmer-coverage
Denisova_30x	-	23.464668
Gorilla_beringei_graueri-9732_Mkubwa	15.22	7.5223
Gorilla_beringei_graueri-A929_Kaisi	27.095	15.252012
Gorilla_gorilla_dielhi-B646_Nyango	19.3176	14.538121
Gorilla_gorilla_gorilla-B642_Akiba_Beri	17.2405	12.863963
Gorilla_gorilla_gorilla-B643_Choomba	18.8362	13.976078
Gorilla_gorilla_gorilla-B644_Paki	18.7551	13.912575
Gorilla_gorilla_gorilla-B647_Anthal	17.8673	13.23363
Gorilla_gorilla_gorilla-B650_Katie	16.4672	12.118194
Gorilla_gorilla_gorilla-9749_Kowali	16.1971	11.872634
Gorilla_gorilla_gorilla-9750_Azizi	16.0982	9.904
Gorilla_gorilla_gorilla-9751_Bulera	16.1769	11.951256
Gorilla_gorilla_gorilla-9752_Suzie	16.4371	12.025414
Gorilla_gorilla_gorilla-9753_Kokomo	13.4676	9.667402
Gorilla_gorilla_gorilla-A932_Mimi	24.7175	1.833891
Gorilla_gorilla_gorilla-A933_Dian	26.2103	22.277507
Gorilla_gorilla_gorilla-A934_Delphi	30.2397	23.113322
Gorilla_gorilla_gorilla-A935_Fritz	26.1051	7.291716
Gorilla_gorilla_gorilla-A936_Coco	22.2133	16.636503
Gorilla_gorilla_gorilla-A962_Amani	33.9349	23.197861
Gorilla_gorilla_gorilla-Snowflake	18.9214	12.456781
Gorilla_gorilla_gorilla-X00108_Abe	15.6194	8.953561
Gorilla_gorilla_gorilla-X00109_Tzambo	18.5012	11.487812
Homo_sapiens-Dai_HGDP01307	19.2118	9.009638
Homo_sapiens-French_HGDP00521	23.5054	12.040849
Homo_sapiens-Han_HGDP00778	-	10.267159
Homo_sapiens-Karitiana_HGDP00998	17.6082	7.576881
Homo_sapiens-Madenka_HGDP01284	22.5465	11.475714
Homo_sapiens-Mbuti_HGDP00456	15.8098	8.238722
Homo_sapiens-Papuan_HGDP00542	16.2915	8.487436
Homo_sapiens-San_HGDP01029	32.8205	14.390751

Homo_sapiens-Sardinian_HGDP00665	21.9033	9.473815
Homo_sapiens-Yoruba_HGDP00927	28.5036	14.729003
Pan_paniscus-9731_LB502	9.66735	7.051574
Pan_paniscus-A915_Kosana	37.5078	28.006446
Pan_paniscus-A917_Dzeeta	39.4451	30.510164
Pan_paniscus-A918_Hermien	38.5	28.862005
Pan_paniscus-A919_Desmond	39.6459	24.840106
Pan_paniscus-A920_Kidogo	26.732	16.105185
Pan_paniscus-A922_Catherine	23.424	17.85244
Pan_paniscus-A923_Kombote	27.2995	26.391115
Pan_paniscus-A925_Bono	27.0355	16.973132
Pan_paniscus-A926_Natalie	28.0149	26.687373
Pan_paniscus-A927_Salonga	22.9217	14.590245
Pan_paniscus-A928_Kumbuka	30.7078	26.832489
Pan_paniscus-A951_Pongo	32.7053	24.546877
Pan_paniscus-X00095_Kakowet	18.2247	11.423614
Pan_troglodytes_ellioti-Akwaya_Jean	22.4677	5.745486
Pan_troglodytes_ellioti-Banyo	9.91445	6.682296
Pan_troglodytes_ellioti-Basho	14.8194	6.699436
Pan_troglodytes_ellioti-Damian	20.5135	4.298133
Pan_troglodytes_ellioti-Julie	22.3339	8.427398
Pan_troglodytes_ellioti-Koto	23.8682	5.709182
Pan_troglodytes_ellioti-Paquita	13.532	9.440522
Pan_troglodytes_ellioti-Taweh	20.6315	4.507583
Pan_troglodytes_ellioti-Tobi	-	9.009664
Pan_troglodytes_ellioti-Kopongo	-	7.946675
Pan_troglodytes_schweinfurthii-100037_Vincent	21.339	13.587553
Pan_troglodytes_schweinfurthii-100040_Andromeda	20.0961	15.853668
Pan_troglodytes_schweinfurthii-A911_Kidongo	47.1899	34.386432
Pan_troglodytes_schweinfurthii-A912_Nakuu	39.6525	26.502671
Pan_troglodytes_schweinfurthii-Yolanda	25.8635	28.46547
Pan_troglodytes_schweinfurthii-9729_Harriet	11.0701	8.636543
Pan_troglodytes_troglodytes-A958_Doris	35.7141	21.921671
Pan_troglodytes_troglodytes-A960_Clara	23.9787	16.498607
Pan_troglodytes_verus-9730_Donald	19.2662	9.404114
Pan_troglodytes_verus-A907_Susie_A	28.076	21.334752
Pan_troglodytes_verus-9668_Bosco	15.6902	7.808031
Pan_troglodytes_verus-Clint	33.7102	16.566894
Pan_troglodytes_verus-X00100_Koby	17.8774	10.910666
Pongo_abelii-A947_Elsi	31.4027	22.223994
Pongo_abelii-A949_Dunja	34.0396	23.289702
Pongo_abelii-A953_Vicki	29.8808	20.514495
Pongo_pygmaeus-A939_Nonja	25.6488	18.264823
Pongo_pygmaeus-A942_Gusti	29.6212	12.912687
Pongo_pygmaeus-A943_Tilda	30.2099	12.133029
Pongo_pygmaeus-A944_Napoleon	29.1711	7.543612
Gorilla_gorilla_gorilla-KB3782_Vila	10.8599	7.763543
Gorilla_gorilla_gorilla-KB3784_Dolly	14.7481	10.054133
Gorilla_gorilla_gorilla-KB4986_Katie	12.6238	8.955044
Gorilla_gorilla_gorilla-KB5792_Carolyn	10.4592	7.50207
Gorilla_gorilla_gorilla-KB5852_Helen	13.1221	9.352174
Gorilla_gorilla_gorilla-KB6039_Oko	13.8467	9.738127
Gorilla_gorilla_gorilla-KB7973_Porta	10.4381	7.553997
Gorilla_gorilla_gorilla-Kwan	-	8.935158
Gorilla_beringei_graueri-Mukisi	-	2.313648
Gorilla_gorilla_gorilla-Kamilah	-	22.561756
Pongo_pygmaeus-KB5404_Billy	-	12.425859
Pongo_pygmaeus-KB4204_Dinah	-	2.84881
Pongo_abelii-KB4361_Dennis	-	4.627225
Pongo_abelii-KB4661_Dolly	-	2.420915
Pongo_pygmaeus-KB5405_Louis	-	3.508526
Pongo_pygmaeus-KB5406_Doris	-	4.041577
Pongo_pygmaeus-KB5543_Baldy	-	3.784019
Pongo_abelii-KB5883_Likoe	-	2.955035
Pongo_abelii-KB9258_Bubbles	-	5.161534
Pongo_abelii-SB550_Sibu	-	3.676026

Section 2: Nonhuman primate reference genome comparisons

Mapping to nonhuman primate reference assemblies: In order to assess human-specific deletions and the copy number of genomic sequence not represented in the human reference, we identified nonhuman primate-specific sequence. The gorGor3(McLean et al. 2011; Chimpanzee Sequencing and Analysis Consortium 2005; Scally et al. 2012), panTro3(Varki et al. 2008; Locke et al. 2011; Chimpanzee Sequencing and Analysis Consortium 2005) and ponAbe2(Elsea and Girirajan 2008; Kent 2002; Locke et al. 2011) references were each mapped in 1 kbp chunks with BLAT(Varki et al. 2008; Altschul et al. 1990; Olson 1999; Kent 2002) against the human reference genome (default parameters and *–minIdentity=90*) to identify regions ≥ 1 kbp that could not be aligned to the human reference (**Table 2.1**). Sex chromosomes were excluded and only regions with ≥ 500 bp of unmasked sequence were considered in the analysis.

Table 2.1: The total sum of unmasked base pairs considered from each nonhuman primate assembly that did not place in the human reference genome.

Reference	Species	Unmasked base pairs	Homologous control sequence base pairs included in mapping
panTro3	Chimpanzee	6015364	2848731
gorGor3	Gorilla	6124188	2827964
ponAbe2	Orangutan	21110529	2797810

All genomes were mapped to this nonhuman primate-specific sequence in addition to four orthologous collections of ~2.85 Mbp of control diploid sequence present in all primate references in order to calibrate copy number counts. Copy number calling was performed on nonhuman primate sequence not present in the human reference, as explained above, adjusted to use the appropriate set of control sequence.

In total, 30,033 loci containing ≥ 500 bp of unmasked sequence were identified in the nonhuman primate genomes assayed; however, this set is redundant as sequence not present in the human reference genome may be present in multiple nonhuman primate reference genomes. We, thus, searched all sequences against each other using Mega BLAST(Cooper et al. 2011; Sudmant et al. 2010; Emerson et al. 2008; Altschul et al. 1990) (Mega BLAST version

2.2.11, parameters *-D 2 -p 95 -F m -U T -s 220 -R 1*) and clustered together all sequences that could align over 20% of their length with ≥97% identity. A total of 22,114 clusters were then identified and from each cluster the sequence containing the maximum number of unmasked base pairs was selected for consideration. Finally, masked sequence in each of the regions considered was padded (extended out on both sides) by 36 bp to eliminate mapping edge effects(Witkin 1984; Sudmant et al. 2010) as was done in the full genome mappings described above. After repeat masking, 9855 regions with ≥500 bp remained. These loci, encompassing 25,374,943 total base pairs, were analyzed for their duplicated and deleted content (**Figure 2.1, see Section 6**).

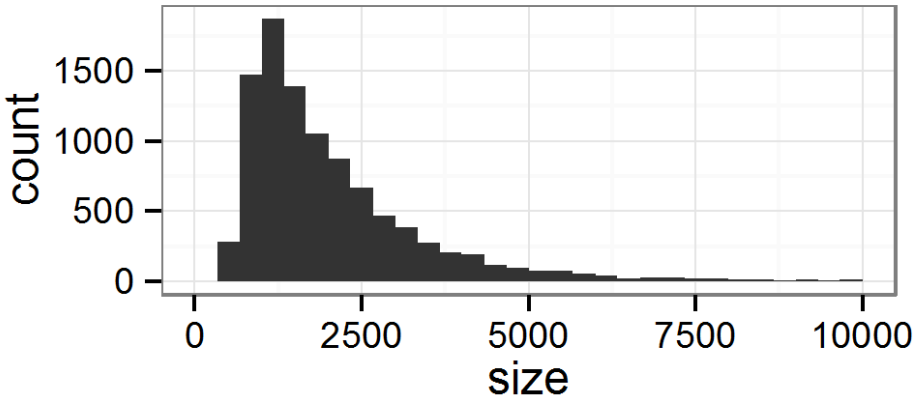


Figure 2.1: Size distribution of 9855 nonredundant loci not present in the human reference genome analyzed.

Section 3: Copy number variant calling methods

Segmentation of duplication and deletion loci: To detect duplications and deletions within genomes, we developed a new algorithm based on the concept of scale-space filtering(Ventura et al. 2011; Witkin 1984). Genome copy numbers were predicted in tiled contiguous windows at a resolution of 500 bp of unmasked sequence in each genome analyzed. We strived to segment these windowed estimates into diploid, duplicated, or deleted regions encompassing contiguous sets of windows. Additionally, we aimed to segment complex duplication architectures to inform the underlying SD structure (**Figure 3.1**).

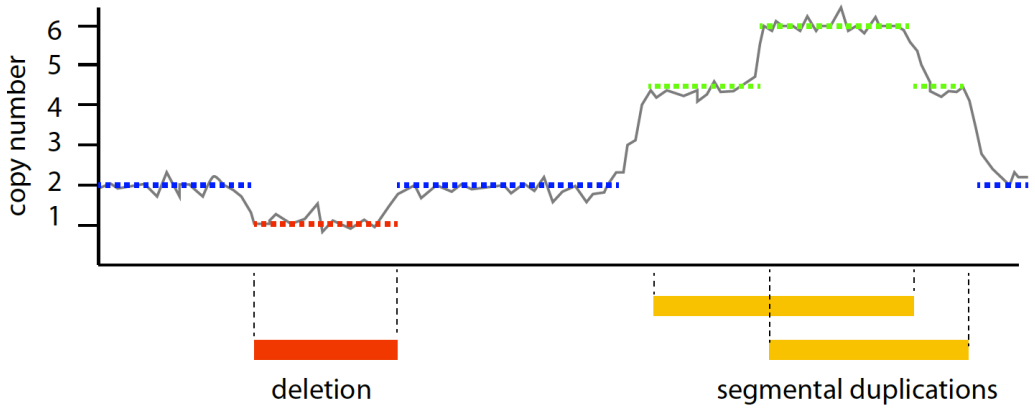


Figure 3.1: A schematic of the desired segmentation output. The gray line represents the predicted copy number of contiguous tiled windows. An ideal segmentation should distinguish deletion (red), diploid (blue), and duplication (green) segments. Also, complex duplication segments should be sub-segmented into constitutive duplications of differing copy number.

To construct this segmentation we first applied scale-space filtering over the windowed copy number estimations. Scale-space filtering is a multi-resolution technique that transforms a waveform signal $f(x)$ into a set of waveforms $f(x,\sigma)$, parameterized by the '*scale*' variable σ. The transformation is achieved by applying a Gaussian convolution over the waveform for multiple values of σ, the standard deviation of the Gaussian. The zero crossings of the second derivative of the transformed waveform, corresponding to peaks of the first derivative of the convolution, are the loci of inflection of the original curve, smoothed over a particular scale. Zero contours of the second derivative within scale-space are assumed to correspond to the same underlying event and events are assumed to originate from the x location where $\sigma \rightarrow 0$. Thus, for each scale-space σ_i, we traverse the scale-space zero contours from σ_i to σ_{min} and construct a set of contour intercepts corresponding to inflections on our original signal. Finally, we select a scale-space cutoff, below which all contour intercepts are discarded, and use the remaining intercepts to construct an initial set of segments. Finally, we hierarchically cluster these segments assigning each segment a value equivalent to the median of the copy number of the windows it encompasses and greedily merging the pair of adjacent segments that shows the least difference in estimated copy number. This merging procedure is repeated until the minimum difference between adjacent segments reaches a designated threshold, in our case 0.5, corresponding to a copy number difference of 0.5 between adjacent segments.

Identifying deletions - methods and sensitivity: We constructed a set of deletions for each species by running our segmentation algorithm on each of the genomes analyzed individually, then merging the total set of events identified in all individuals into a species deletion set. In this way, the population of individuals is used to discover the total possible set of deletion events and these events can then be genotyped in all individuals. To test the accuracy of this method, we then compared the set of deletions called in gorillas to those recently identified in the Western lowland gorilla Kwan(Hormozdiari et al. 2009; Ventura et al. 2011) using an orthologous paired-end-read approach (Variation Hunter(Hormozdiari et al. 2009)) and validated by array comparative genomic hybridization (arrayCGH) **(Figure 3.2)**. Using segmentations based on 500 bp windows in the single individual Kwan, we were able to rediscover 97.6% (81/83) of deletions encompassing 3 kbp or more of unmasked sequence that were previously identified by paired-end analysis in this same individual and 89.6% of deletions encompassing 1 kbp or more of unmasked sequence (457/510). As smaller deletions were more challenging to discover, we leveraged the fact that 32 gorilla genomes were sequenced and compared the total set of gorilla deletions discovered by our algorithm to those identified by paired-end sequencing in Kwan. 95.3% (486/510) and 98.8% (159/161) of events encompassing greater than 1 kbp or greater than 2 kbp unmasked base pairs, respectively, were rediscovered in the total gorilla deletion set, demonstrating the power of leveraging a population of sequenced individuals.

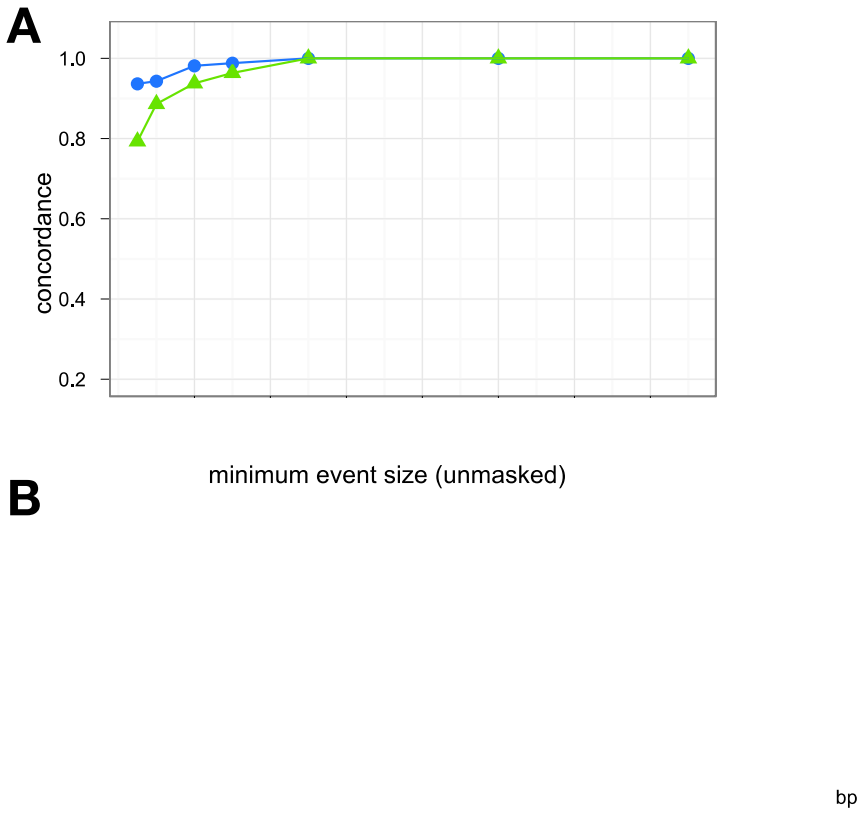


Figure 3.2: The concordance rate of deletion calls detected by the read-depth-based segmentation algorithm compared to those made by paired-end structural variation detection using Variation Hunter(Bailey et al. 2002; Hormozdiari et al. 2009)**. Concordance rates are shown for different size thresholds using the number of unmasked base pairs (A) underlying an event (i.e., the number of base pairs used in the segmentation) and the total event size (B). Concordance rates are plotted for 500 bp segmentations and for events discovered in just Kwan and events discovered across the set of all gorillas (n=32).**

Identifying duplication segmentations: For each species we constructed a consensus set of duplication segments by merging the individual segmentations of all individuals in a species with the purpose of leveraging information across multiple individuals to inform the true SD boundaries. Segmentations across multiple individuals are largely similar. Additionally, as demonstrated from the detection of deletion loci across multiple individuals above, duplicated loci discovered in one individual can then be genotyped across the population of individuals. Merged duplication segmentations were constructed by first considering all duplicated regions (duplicated regions may or may not encompass multiple segments; **Figure 3.1**) detected across

any individual. As demonstrated for the deletions above, we have significant power to detect events at a threshold of 3 kbp with 96.4% of deletions identified from discovery in a single individual. For each duplicated region, we assessed all individuals where the duplication was present and determined the median number of segments (*k)* identified in individuals in this duplicated region. Each edge between adjacent segments was also assigned a count equal to the total number of individuals in which this edge was identified. A consensus segmentation *k* segments long was then constructed by greedily selecting the *k+1* edges with the highest counts **(Figure 3.3)**.

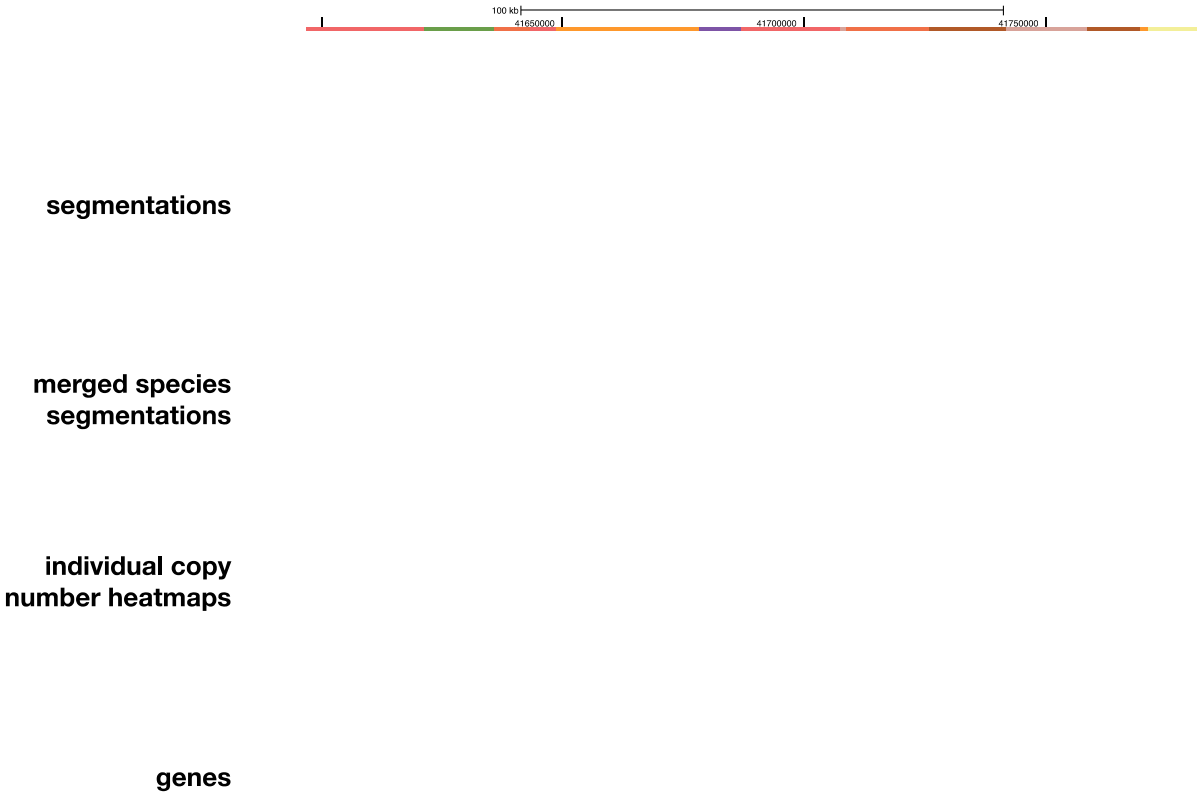


Figure 3.3: An example of a merged species segmentation overlaid over copy number heatmaps for 32 Western gorillas for a duplicated region on 17q21. The individual duplication segmentations (top, false colored) are merged (middle bars and black lines) to construct a more robust species segmentation from which copy numbers can be assessed and compared between individuals. Each window represents 500 bp of unmasked sequence scaled to genomic coordinates.

To test the effectiveness of our method, we compared the consensus duplication segmentation made across the 10 humans we analyzed to annotated SDs in the reference genome(Bailey et al. 2002) called from genome self-alignments (whole-genome alignment comparison, WGAC) of

a single individual. We identified 172,817,129 (**non-copy number corrected**) duplicated bp among the 10 humans analyzed (excluding sex chromosomes), 139,560,079 bp of which were present in at least 5/10 individuals, compared to 130,447,077 duplicated bp identified in the human reference genome by WGAC. 95.4% of all SDs containing >3 kbp of unmasked sequence in the human reference genome were identified in our analysis of these 10 human individuals and 92.7% of all annotated duplications were identified in at least 5/10 of these individuals (**Figure 3.4**).

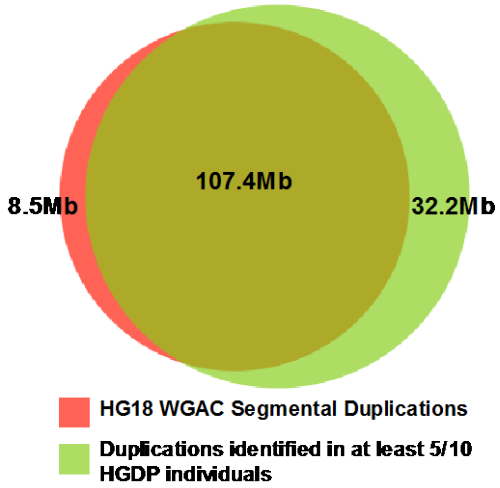


Figure 3.4: The overlap between duplicated base pairs identified from WGAC and identified by read-depth followed by segmentation in 10 diverse human individuals and present in at least 5/10 individuals containing at least 3 kbp of unmasked sequence.

We next assessed our segmentation by comparing the boundaries of known SDs, identified from self-alignments of the human genome(Bailey et al. 2002), to those identified by our segmentation algorithm. We, thus, took all segments overlapping known SDs and compared the edges of each segmentation to the set of known human SD edge boundaries. Our duplication segmentation edges, which were based on copy number estimates across 500 unmasked bp tiled windows, had a median distance of 367 unmasked bp to the nearest known SD edge **(Figures 3.5 and 3.6)** and a median total distance of 1373 total base pairs to the nearest known SD boundary. Since the distance between known duplication edges and our segmentation edges is at the resolution of our windowed copy number estimates, we conclude that our segmentation is accurately reflecting the underlying duplication structure of the genomes analyzed.

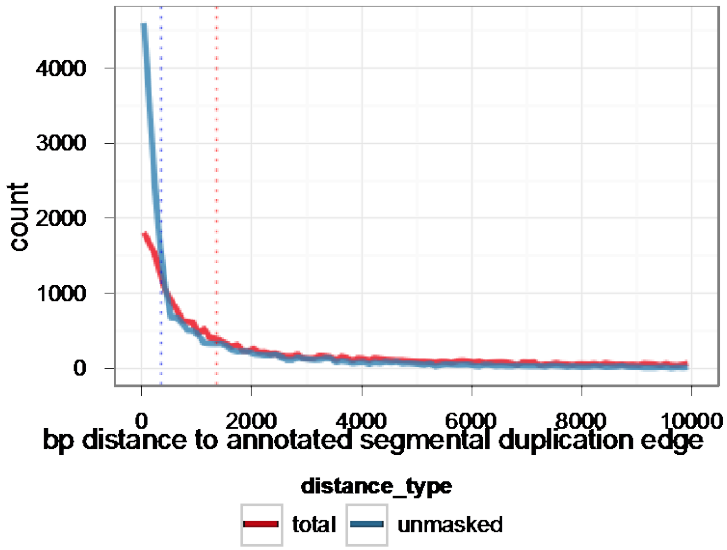


Figure 3.5: Histograms of the distance in total and unmasked base pairs from predicted duplication segmentation edges to known SD edges in the human genome. The median distance from a predicted duplication boundary to a known duplication boundary (dotted lines) is 367 bp of unmasked sequence or 1373 total bp. Human duplication segmentations were identified from duplications identified in 10 human genomes.

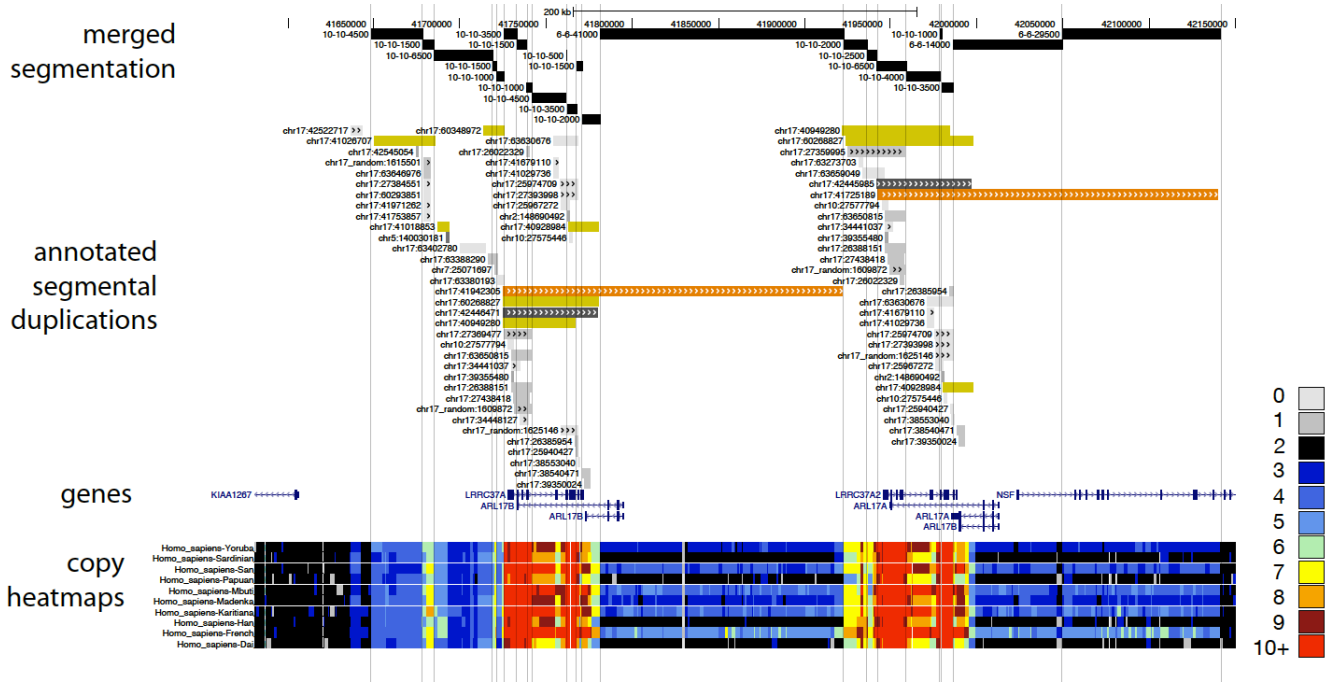


Figure 3.6: An example of a complex duplicated locus on 17q21 in human. Human-annotated SDs are shown overlaid with predicted merged human species segmentations and copy number heatmaps of the 10 individuals analyzed. Boundaries of predicted human SD edges correspond well with annotated SD boundaries.

Section 4: Validations

ArrayCGH validations: We designed three custom arrays to validate our call-sets: a duplication array, a deletion array, and an array targeted to sequence present in nonhuman primate reference genomes but not in the human reference. A total of 58 total hybridization experiments (29 test/reference, reference/test dye-swap pairs) were performed on these platforms using the human HapMap individual NA12878 as a reference against 18 different nonhuman primate individuals (4 gorillas, 2 bonobos, 4 orangutans, and 8 chimpanzees). Primate and human DNA were labeled with Cy3 or Cy5 dye (NimbleGen labeling kit). Labeled DNA were mixed in equal amounts and hybridized to either a NimbleGen 3 X 720K or Agilent 2 X 400K array for 40 hours at 42 degrees. Slides were then washed, scanned, and analyzed. Individual noisy and poorly performing probes were identified as those giving opposite signals in dye swap experiments and were discarded. Additionally, only loci with three or more probes were considered. Lineage-specific deletions were considered confirmed if a mean log2ratio of ≤-0.5 across a locus was seen in all individuals in that lineage. Duplications were considered identically using a log2ratio cutoff of 0.25. Lineage-specific duplications along the human lineage were additionally considered confirmed if they intersected with known SDs identified by WGAC(Lichter et al. 1990; Bailey et al. 2002). The validation rate among all experiments ranged from 96.9% to 98.1% for fixed sites and was 85% for copy number polymorphisms (CNPs) demonstrating our call-set to be highly robust (**Table 4.1**).

Table 4.1: Validation results summary. In total 9034 loci were experimentally tested, of which 8641 were validated, resulting in an overall validation rate of 95.65%.

Method	Targeted variation	Experiments performed	Loci experimentally tested	Loci confirmed	Validation rate
arrayCGH	**Duplications**	14	3776	3660	96.93%
FISH	**Duplications**	104	104	102	98.08%
arrayCGH	**Deletions**	14	2503	2476	98.92%
arrayCGH	**Nonhuman reference sequence deletions (absent from human reference)**	30	1518	1490	98.16%
arrayCGH	**CNPs**	15	1520	1294	85.13%
	Total	**177**	**9421**	**9022**	95.76%

Fluorescent *in situ* hybridization (FISH) validations: To further assess the structure of duplications and deletions identified, we performed 109 FISH experiments on human and nonhuman great ape cell lines across 53 different loci. Metaphase and interphase nuclei from nonhuman primates (common chimpanzee, gorilla, bonobo, and orangutan) were obtained from lymphoblastoid or fibroblast cell lines; human metaphase spreads were derived from PHA-stimulated peripheral lymphocytes of normal donors by standard procedures. Briefly, cells were treated with colcemid (D1925; Sigma) and collected by centrifugation, incubated in a hypotonic solution (KCl 0.56%), and then washed three times and stored in a fixative solution (one part acetic acid, three parts methanol). DNA from human fosmid clones was extracted using a Plasmid Miniprep kit (Bio-Rad, Cat# 732-6100). FISH assays were performed as previously described by Lichter et al(Cheng et al. 2005; Lichter et al. 1990). Fosmids were directly labeled either with Cy3-dUTP or fluorescein-dCTP by nick-translation reactions. Three hundred nanograms of labeled probe were used for the FISH experiments. Hybridization was performed at 37°C in 2XSSC, 50% (v/v) formamide, 10% (w/v) dextran sulfate, and 3 μg of sonicated salmon sperm DNA in a volume of 10 μL. High stringency post-hybridization washings (three) were at 60°C in 0.1X SSC. Digital images were obtained using a Leica DMRXA epifluorescence microscope equipped with a CCD camera. Fluorescence signals detected with Cy3 and fluorescein filters and chromosomes and nuclei images detected with DAPI filter were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

Of the 109 *in situ* experiments performed, we confirmed in 102 cases the predicted duplication and were able to resolve the duplication locus (**Table 4.2**). Five experiments were inconclusive showing only background signal resulting in an overall validation rate of 98% (102/104). Each

of the two remaining duplications not resolved by FISH showed strong signals by arrayCGH, suggesting a FISH false-negative (**Table 4.2**). Overall, the FISH and arrayCGH validation rates results are highly concordant, resulting in an overall validation rate of ~97% (**Table 4.1**).

<TABLE 4.2>

Section 5: Lineage-specific great ape segmental duplications and deletions

Lineage-specific SDs: Lineage-specific duplication loci were assigned by assessing all duplicated segments and labeling them as lineage-specific SDs if >50% of individuals in a particular lineage were at least copy number 4 while <10% of individuals in all other lineages showed any signature of duplication (<10% ≥copy 3). All segments were assigned a duplication state. The same approach was taken to identify expansions, requiring all individuals in a particular lineage to have a higher copy number than all other individuals. The total number of duplicated base pairs in a specific lineage was thus calculated by summing the duplication content of all duplicated segments corrected for the copy number of SDs present in the human genome (this process has previously been termed *copy number correction*(Cheng et al. 2005; Marques-Bonet et al. 2009; Ventura et al. 2011)) and excluding the ancestral locus of the duplication (**Table 5.1**). We note that in previous analyses of lineage-specific SDs we have not excluded the ancestral locus from total counts(Meyer et al. 2012; Cheng et al. 2005; Marques-Bonet et al. 2009; Ventura et al. 2011). However, to directly compare different rates in the accumulation of genetic variation, excluding the ancestral locus is necessary. Duplicated base pair counts were also computed over the nonhuman primate reference assembly sequence that was not present in the human reference genome. The overall contribution of this sequence to the total number of duplicated base pairs was minimal however (~2%).

Table 5.1: Counts of the number of events and total base pairs specifically duplicated in different lineages. Base-pair counts are corrected for the SD content of the human reference genome and to exclude the ancestral duplication locus.

lineage	copy corrected duplicated base pairs	number of sites
Gbeg-Ggod-Ggog-Hde-Hsa-Pab-Ppa-Ppy-Ptre-Ptrs-Ptrt-Ptrv	184296495	7118
Gbeg-Ggod-Ggog	29413779	278
Pab-Ppy	27854532	885
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	24491334	1172
Ppa-Ptre-Ptrs-Ptrt-Ptrv	6050953	121
Hde-Hsa	5715010	282
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	4227312	269
Ppa	860900	67
Ptre-Ptrs-Ptrt-Ptrv	802219	30
Ppy	487342	48
Gbeg	446285	89
Pab	361732	38

Ggod-Ggog	339371	35
Hsa	280445	19
Hde	142093	6
Ggod	111722	31
Ptrv	31250	15
Ptrt	20823	10
Ptrs	13994	2
Ptre-Ptrv	13830	2
Ptrs-Ptrt	1790	1
Ptre	1302	1
Ggog	5600	2
BONOBO-CHIMP-EASTERNGORILLA-WESTERNGORILLA	25555717	103
EASTERNGORILLA-HUMAN-WESTERNGORILLA	2646209	204
EASTERNGORILLA-ORANGUTAN-WESTERNGORILLA	2002046	78
EASTERNGORILLA-HUMAN-ORANGUTAN-WESTERNGORILLA	1776400	211
BONOBO-CHIMP-HUMAN-ORANGUTAN	1492574	146
HUMAN-ORANGUTAN	1015592	86
CHIMP-EASTERNGORILLA-HUMAN-WESTERNGORILLA	991242	64
CHIMP-EASTERNGORILLA-HUMAN-ORANGUTAN-WESTERNGORILLA	705223	55
BONOBO-CHIMP-EASTERNGORILLA-ORANGUTAN-WESTERNGORILLA	642072	27
BONOBO-CHIMP-ORANGUTAN	435420	17
Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	426930	6
Hde-Hsa-Pab-Ppy	420272	9
HUMAN-WESTERNGORILLA	243475	65
Gbeg-Ggod-Ggog-Pab-Ppy	172364	17
BONOBO-EASTERNGORILLA-HUMAN-WESTERNGORILLA	148340	16
CHIMP-HUMAN	140254	30
BONOBO-EASTERNGORILLA-ORANGUTAN-WESTERNGORILLA	122483	8
HUMAN-ORANGUTAN-WESTERNGORILLA	118243	53
ORANGUTAN-WESTERNGORILLA	111772	19
CHIMP-EASTERNGORILLA-WESTERNGORILLA	97792	3
BONOBO-CHIMP-WESTERNGORILLA	93547	23
BONOBO-HUMAN	46010	12
CHIMP-EASTERNGORILLA	34776	1
CHIMP-HUMAN-ORANGUTAN	27975	10
BONOBO-ORANGUTAN	23010	5
BONOBO-EASTERNGORILLA-WESTERNGORILLA	18598	1
Pab-Ppa-Ppy-Ptre-Ptrs-Ptrt-Ptrv	13736	5
BONOBO-ORANGUTAN-WESTERNGORILLA	13600	2
BONOBO-EASTERNGORILLA-HUMAN-ORANGUTAN-WESTERNGORILLA	10933	7
CHIMP-HUMAN-WESTERNGORILLA	8363	4
CHIMP-EASTERNGORILLA-ORANGUTAN-WESTERNGORILLA	7058	2
EASTERNGORILLA-ORANGUTAN	6920	2
BONOBO-CHIMP-ORANGUTAN-WESTERNGORILLA	5015	2
BONOBO-CHIMP-EASTERNGORILLA	4623	2
Gbeg-Ggod-Ggog-Hde-Hsa	4616	2
CHIMP-ORANGUTAN	2466	1
BONOBO-CHIMP-HUMAN-ORANGUTAN-WESTERNGORILLA	2054	3
BONOBO-CHIMP-EASTERNGORILLA-HUMAN-ORANGUTAN	1924	3
BONOBO-HUMAN-ORANGUTAN	1778	1

BONOBO-CHIMP-HUMAN-WESTERNGORILLA	1303	2
EASTERNGORILLA-HUMAN-ORANGUTAN	931	5
BONOBO-HUMAN-WESTERNGORILLA	519	1
BONOBO-HUMAN-ORANGUTAN-WESTERNGORILLA	507	2
EASTERNGORILLA-HUMAN	301	1
BONOBO-CHIMP-EASTERNGORILLA-HUMAN	145	1

Resolving duplications inconsistent with the phylogeny: Some of the duplication events identified were inconsistent with the species' phylogeny and, thus, could not simply be assigned to any branch. For example, ~25.6 Mbp of sequence were identified as duplicated in chimpanzees, bonobos, and gorillas, but not in the human lineage. The most parsimonious explanation for this result in many cases is that this duplicated sequence was present in the ancestor of African great apes and was deleted in the human lineage. This sequence was subsequently added to correct lineage using the most parsimonious explanation.

Duplication rates: The rate of duplication was directly compared to the rate of single nucleotide substitution from single nucleotide polymorphism (SNP) data calculated for 86 of the same individuals assessed for duplication content (see companion Prado/Sudmant 2012; **Table 5.2**). Briefly, all individuals were aligned to the HG18 reference genome using BWA with relaxed edit distance parameters to account for the increased divergence of these species. SNPs were called using the GATK pipeline on a population level and were filtered for depth-of-coverage, proximity to an indel, and allele balance. Genetic distances between all species were then calculated by computing the genetic distances between all individuals and taking the mean pairwise divergence between individuals of different populations. A tree was constructed from this distance matrix by neighbor joining. The Denisova individual was placed as a sister group to the human genome splitting at 12.35% the branch length of the human-chimpanzee common ancestor with a branch length 1.2% shorter than human as reported in Meyer *et al* 2012(Sun et al. 2012; Meyer et al. 2012; Langergraber et al. 2012). Divergence times were calculated using a mutation rate scaled to a human-chimpanzee divergence time of 6.5 million years ago, (8.79e-10 substitutions/year per base **Table 5.2**). These speciation times match well with the literature(Sun et al. 2012; Langergraber et al. 2012). The rate of duplication along each branch was calculated by dividing the per-base-pair duplication rate (total copy number corrected duplicated bp/2.867e9) by the per-base-pair substitution rate and additionally by the estimated divergence times in millions of years (**Table 5.3**).

As we previously reported(1000 Genomes Project Consortium et al. 2012; Hach et al. 2010; Marques-Bonet et al. 2009; Ventura et al. 2011), we find a burst of duplication activity in the ancestor of African great apes (**Figure 5.1**) corresponding to a duplication rate along the African great ape lineage approximately 2.6-fold the rate of substitution and ~7-fold higher than the rate along the branch leading to the human lineage. This increased rate of duplication continued along the gorilla lineage and the chimpanzee-human ancestral lineage.

<Table 5.3>

Table 5.2: Divergence estimates from Prado and Sudmant *et al* (under review). Divergence times in years are estimated using a mutation rate of $8.79X10^{-10}$ bp/year, calibrated to a 6.5 million human-chimpanzee divergence time.

lineage	branch distance to leaf	branch length	time to leaf (MYA)	branch length (million years)
Gbeg-Ggod-Ggog-Hde-Hsa-Pab-Ppa-Ppy-Ptre-Ptrs-Ptrt-Ptrv	0.014354	-	16.3	-
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	0.006928	0.007422	7.9	8.4
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	0.005713	0.00126	6.5	1.4
Ppa-Ptre-Ptrs-Ptrt-Ptrv	0.001752	0.00391	2	4.4
Pab-Ppy	0.001585	0.012788	1.8	14.5
Gbeg-Ggod-Ggog	0.000997	0.00584	1.1	6.6
Ptre-Ptrs-Ptrt-Ptrv	0.000937	0.00081	1.1	0.9
Ptrs-Ptrt	0.00085	0.00006	1	0.1
Ggod-Ggog	0.00081	0.00017	0.9	0.2
Ptre-Ptrv	0.000835	0.00013	0.9	0.1
Hde-Hsa	0.000737	0.005233	0.8	6
Ppy	0	0.00162	0	1.8
Pab	0	0.00155	0	1.8
Gbeg	0	0.00103	0	1.2
Hde	0	0.000666	0	0.8
Hsa	0	0.000737	0	0.8
Ppa	0	0.00177	0	2
Ggod	0	0.00079	0	0.9
Ggog	0	0.00083	0	0.9
Ptrv	0	0.00086	0	1
Ptre	0	0.00081	0	0.9
Ptrs	0	0.00087	0	1
Ptrt	0	0.00083	0	0.9

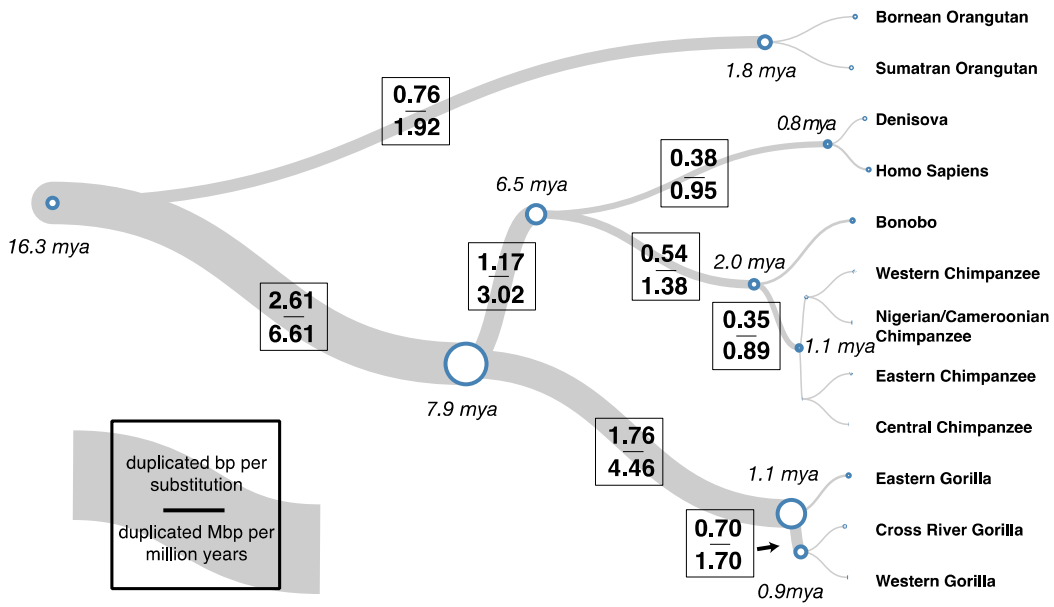


Figure 5.1: A species/subspecies tree with branch lengths proportional to the number of substitutions and branch widths scaled to the base-pair duplication rate per-nucleotide substitution and labeled for nonterminal branches. A burst of duplications in the ancestor of African great apes is apparent with duplication rates along the African great ape lineage of ~7-fold the rate along the human lineage per million years.

Lineage-specific deletions: Lineage-specific deletions were assigned as above for duplications, including sequence not present in the human reference genome (see **Section 2**). We required at least 80% of individuals to have a complete deletion signature in a particular lineage and all other lineages to show evidence of at least a single copy of the segment in 100% of all individuals assayed to ensure we were capturing fixed lineage-specific deletion events (**Figure 5.2**, **Table 5.4**). In contrast to the rate of duplication, we find the rate of deletion is largely clocklike with the major exception of the chimpanzee ancestral branch (**Table 5.3, Figure 5.3**), on which we observe a ~2-fold acceleration in the rate of chimpanzee deletions. This excess in deletion events correlates with a predicted severe reduction in effective population size, possibly the result of an extreme bottleneck in the chimpanzee-bonobo ancestor (**Figure 5.4,** Prado/Sudmant under review).

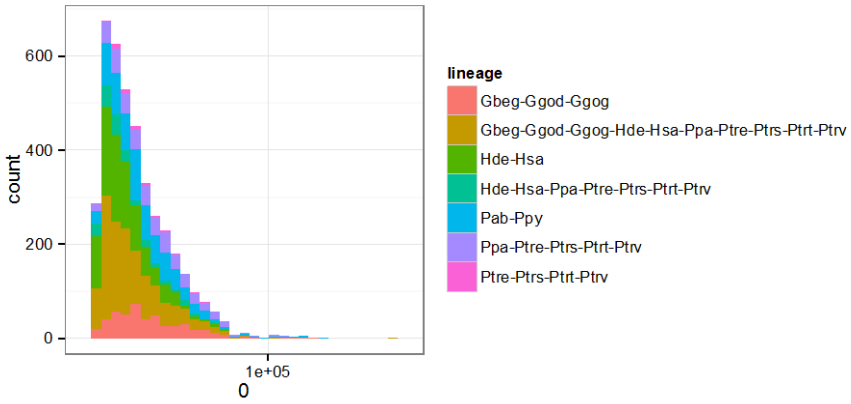


Figure 5.2: Size distribution (in log space) of lineage-specific duplications and deletions identified in great apes.

Table 5.4: Counts of all lineage-specific deletions identified in great apes and the number of sites identified.

species	total_bp	total_events	total_bp_gt_5kb	total_events_gt_5kb
Ptrs	7887	2	0	0
Ptrt	10655	4	0	0
Ptrv	276145	32	228196	17
Ggod-Ggog	174258	8	171072	7
Ppy	411190	36	352421	19
Ppa-Ptre-Ptrs-Ptrt-Ptrv	6495864	575	5727585	339
Ptre	35715	7	21178	2
Hde-Hsa	32339	2	32339	2
Hde	931589	113	755770	58
Ppa-Ptre-Ptrs-Ptrt-Ptrv-Pab-Ppy	373174	13	354786	8
Gbeg-Ggod-Ggog-Pab-Ppy	356638	31	330272	23
Ptrs-Ptrt	8730	1	8730	1

Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	1072867	76	1008570	57
Ppa	1644331	149	1456279	86
Ptre-Ptrv	21208	4	12213	1
Pab-Ppy	10562674	1067	8739431	518
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv-Pab-Ppy	165022599	264	164936373	237
Gbeg-Ggod-Ggog	6271686	663	5236407	353
Gbeg	527951	70	422322	37
Ptre-Ptrs-Ptrt-Ptrv	441826	62	326028	29
Ggod	1194717	146	956540	74
Hsa	20353	2	17909	1
Pab	210551	10	198160	6

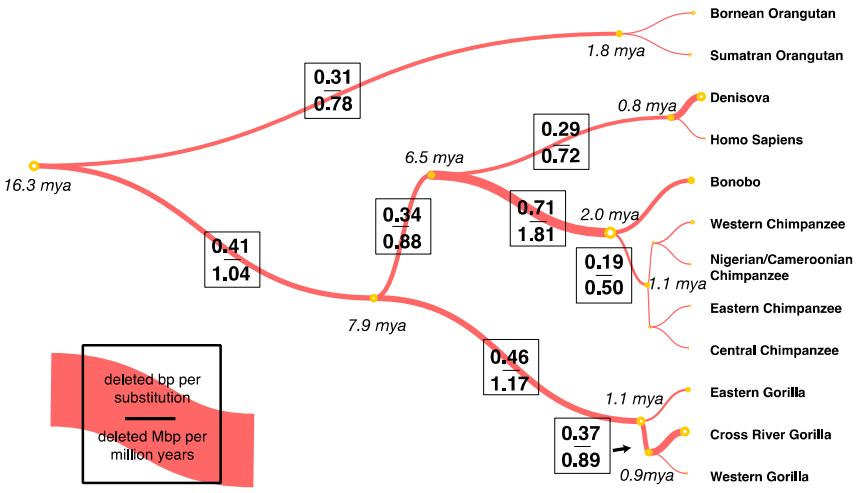


Figure 5.3: A species/subspecies tree with branch lengths proportional to the number of substitutions and branch widths scaled to the base-pair deletion rate per nucleotide substitution and labeled for nonterminal branches. The rate of deletion is largely clocklike with respect to the substitution rate; however, we find a ~2-fold increase in the rate of deletion in the chimpanzee ancestor.

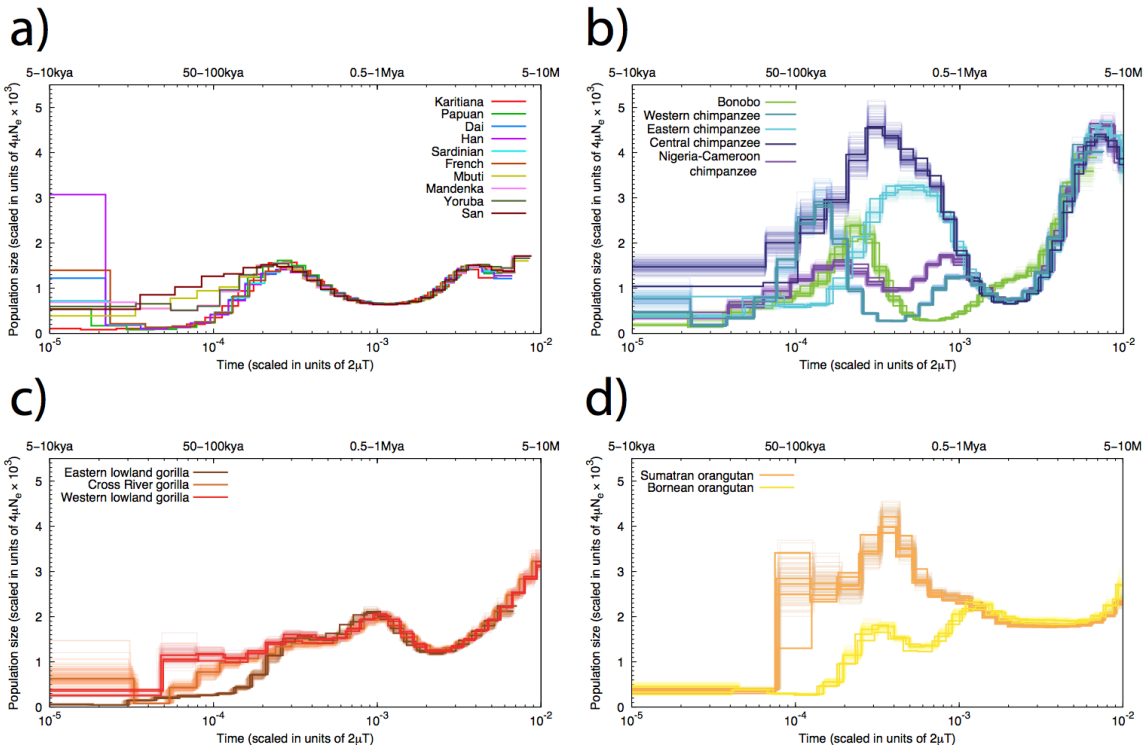


Figure 5.4: Pairwise Sequentially Markovian Coalescent (PSMC) profiles of effective population size (by scaled mutation rate) as a function of time (measured by pairwise sequence divergence, lower x-axis, and in years assuming a mutation rate ranging 10^{-9} – $5x10^{-10}$, upper x-axis) from Prado and Sudmant (under review). In panels b,c,d thin lines

indicate 100 bootstrapped replicates to indicate confidence. A severe crash in the effective population size of the chimpanzee-bonobo ancestor is observed just prior to the separation of the chimpanzee and bonobo species. Steep declines are also observed in Sumatran orangutans, bonobos, and Western chimpanzees.

In concert, these analyses demonstrate how two diametric forms of structural variation have differently shaped the genomes of great apes (**Figure 5.5**). While a burst of duplications characterized the African great ape ancestral genome, this rate subsequently rapidly decreased in the chimpanzee-human ancestral lineage and decreased, at a slower rate, in the gorilla lineage. The rate of deletion, in contrast, has remained largely consistent with the exception of an increase in the chimpanzee ancestor. This striking excess of deletions coincides with a drop in the predicted effective population size, suggesting a severe bottleneck may have resulted in the collapse of this population and, thus, a surfeit of deletions accumulating. While the overall number of base pairs affected by duplication far exceeds that affected by deletion, the relative contribution of duplicated base pairs to deleted base pairs is markedly diverse among different lineages. Duplicated base pairs outnumber deleted base pairs by a factor 6.3 in the African great ape ancestor. In the human ancestor and the chimpanzee ancestor, these ratios are 1.3 and 0.76, respectively, exhibiting a markedly higher relative impact of deletions in these lineages.

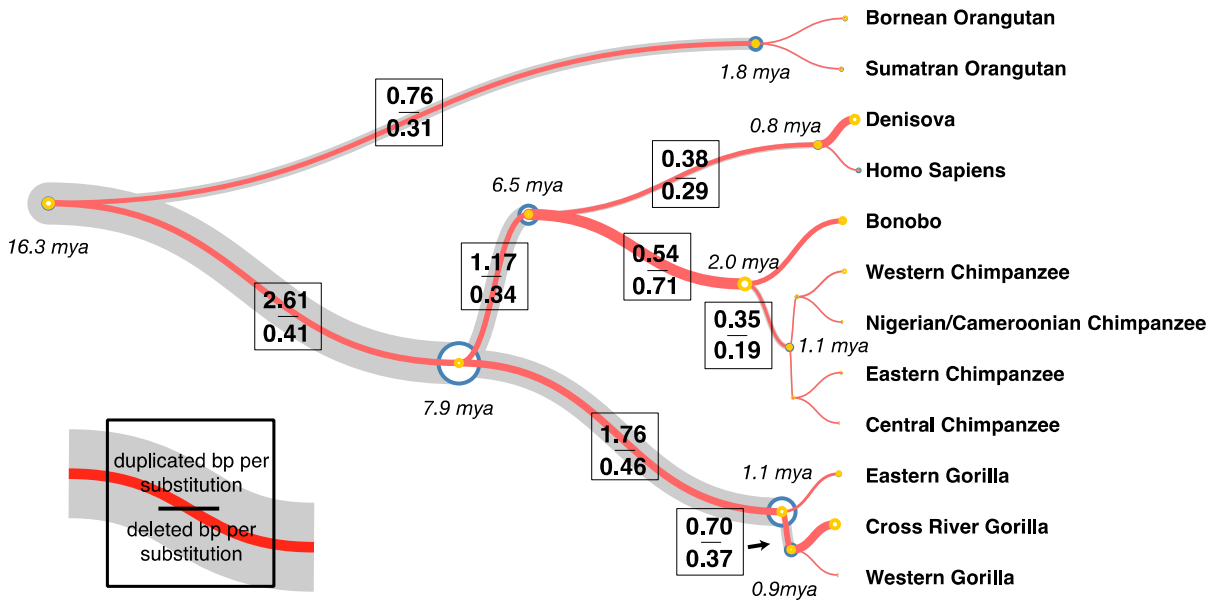


Figure 5.5: The superimposed duplication and deletion rate trees with all nonterminal branches labeled with the duplication and deletion rates per substitution. In the *Pan*

lineage the rate of deletion has dramatically increased with respect to the rate of duplication.

Section 6: Lineage-specific deletions and diversity of sequence not represented in the human reference genome

Evolution and diversity of human reference deletions: Of the 9855 loci present in nonhuman primate reference genomes but not represented in the human reference genome, 5777 loci (13.54 Mbp) corresponded to lineage-specific fixed deletions and were assigned to their respective branches (see **Section 5**, **Table 6.1**, **Figure 6.1**). In addition to these loci identified as fixed deletions, 52 loci encompassing 179 kbp of sequence were not present in any of the individuals assessed, suggesting that these loci largely represent mis-assemblies in addition to rapidly diverging sequence and copy number polymorphic loci.

Table 6.1: Summary of sequence absent from the human reference genome.

***Sequence identified to be segregating at >5% frequency in 620 diverse humans.**

Lineage	Fixed deletions bp (loci)	Duplicated copy corrected bp (loci)	Copy number polymorphic bp (loci)
Human	3588396 (917)	168165 (2)	695007 (324)*
Human-Pan	1232172 (213)	2529 (1)	-
Human-Pan-Gorilla	8729420 (1185)	11850 (3)	-
Human-Pan-Gorilla-Orangutan	179252 (52)	307249 (46)	-

Figure 6.1: A heatmap of sequence absent from the reference genome. Each column represents a locus; blocks are colored to represent per-population frequency estimates of the presence or absence of each particular sequence. The cladogram to the left represents hierarchical clustering of the populations surveyed and recapitulates all population relationships.

We next strived to identify which of these loci not present in the human reference genome could be found segregating in the human population. Accordingly, we mapped short-read shotgun sequence from 620 diverse humans from 13 different populations sequenced as part of the 1000 Genomes Project(Siepel et al. 2005; 1000 Genomes Project Consortium et al. 2012) to loci not present in the human reference genome and determined the approximate frequency of each of these regions in individual populations based on the *presence/absence* of signal. In total, we identified 316 loci corresponding 683 kbp of sequence segregating in at >5% frequency in at least one population (**Table 6.1**). We note that these are not strictly allele frequencies but frequencies of the presence/absence of any particular sequence, as accurate copy number genotyping of these sequences is problematic due to their divergence, small size, and the low coverage of these samples. Hierarchical clustering of the 13 populations based on these frequency estimates completely recapitulates all of the inter-population relationships (**Figure 6.2**), including those between admixed populations, and places the African populations

as an outgroup to the European, Asian and North American populations as is expected for an out-of-Africa origin of Homo sapiens.

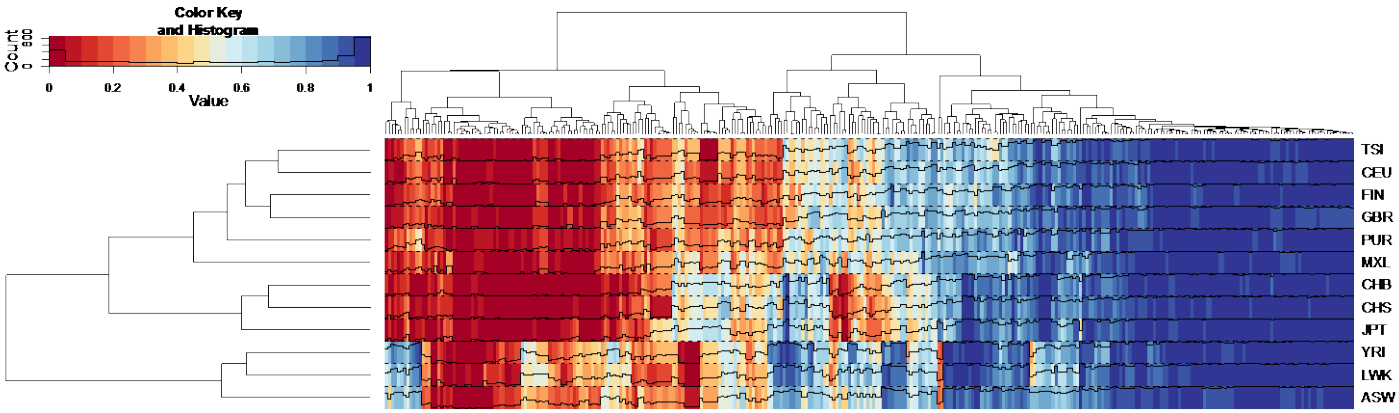


Figure 6.2: A heatmap of sequences genotyped for presence/absence in 620 diverse humans from the 1000 Genomes Project. Columns consist of 324 loci with a frequency >5% in at least one of the populations assessed. Blocks are colored by the per-population frequency and both rows and columns are clustered by hierarchical clustering recapitulating the inter-population relationships.

We calculated a modified Fst (indicated here as Fst*) statistic for each of these loci using the frequency of the presence of a sequence as a surrogate for allele frequency (**Figure 6.3**). Ranking events by these Fst* values, we identified a number of sites of extreme population differentiation, including 53.8 kbp of continental-population-specific sequence absent from all other continental populations groups (52.5 kbp African- and 1.4 kbp European-specific sequence segregating at least 1% frequency) in addition to 77.4 kbp of sequence segregating exclusively in a single continental population and 152.9 kbp of sequence segregating in a single subpopulation at a frequency of at least 1% while fixed in all other populations.

Table 6.1: Population-specific segregating copy number variant loci not found in the human reference genome. (CNVs=copy number variations)

Population	YRI	CHB	ASW	TSI	MXL	LWK	CEU	CHS	GBR	PUR	FIN	JPT	CLM
# of CNVs	5	2	1	4	3	2	2	2	14	11	8	1	5
Total bp sum	44978	4900	871	4319	4895	2002	5494	10092	31903	18243	16614	2744	5798

Functional impact of deleted sequence absent from the human reference: Compared to the human reference genome, sequences absent from the human reference yet present in other nonhuman primate reference genomes are less likely to be annotated for functional elements. We, thus, performed a number of analyses to assess the functional relevance of sequence deleted throughout the great ape lineage and absent in the human reference. We first assessed the conservation of these elements using Phastcons highly conserved elements (HCEs(Siepel et al. 2005), UCSC genome browser). Among the 13.54 Mbp of human-deleted sequence, we identify 180,522 bp of conserved sequence (2219 total elements) with 9910/2385 human deleted loci (38.1%) encompassing an element. As HCEs encompass 3-8% of the human reference genome(Brawand et al. 2011; Siepel et al. 2005), this represents a depletion of conserved loci in human-deleted loci. We next assessed RefSeq gene and mRNA annotations to identify deleted coding elements. We found three complete or partially deleted genes (**Table 6.2**).

Table 6.2: Annotated genic elements absent from the reference genome and deleted in 620 diverse humans.

Locus	Gene (approved name)	Lineage lost	Description
chr19:56319337-56321545	*SIGLEC13*	Human	Complete gene deletion – previously identified in Want *et al,* 2012. This gene is expressed in chimpanzee monocytes which are responsible for innate immunity to bacteria.
chr19:7839944-7842326	*CD209L2 (CLECM4)*	Human	3 exon deletion of the gene encoding CD209 antigen like protein E. CD209 genes encode for C-type lectins which recognize bacteria, mycobacteria, mycobacteria and, viruses and protozoa. CD209L2 has been previously identified as a human-specific deletion. Oritz *et al,* 2008.
chr1_random:2270147-2276474	*LOC100171780*	Gorilla-Chimp-Human	Uncharacterized protein

As nonhuman primate gene annotations are potentially incomplete, we next assessed RNAseq data from six tissues each in gorillas, humans, chimps, bonobos, and orangutans (55 experiments assessed total) generated in Brawand *et al*(Brawand et al. 2011) (**Table 6.3)**.

Table 6.3: Summary table of 55 RNAseq experiments from Brawand *et al* (Langmead and Salzberg 2012; Brawand et al. 2011; Trapnell et al. 2012) **across all great ape lineages mapped with Bowtie 2.0 to each respective primate reference genome. cb=cerebellum, ts=testis, kd=kidney, ht=heart, br=brain, lv=liver**

ID	Species	Tissue	Sex	Reads Placed (Primate Reference)
GSM752656	ggo	cb	M	11108268
GSM752663	ggo	ts	M	12161822
GSM752659	ggo	kd	F	12879458
GSM752658	ggo	ht	M	15376965
GSM752655	ggo	cb	F	15966074
GSM752657	ggo	ht	F	16347127
GSM752660	ggo	kd	M	19509289
GSM752653	ggo	br	F	20146569
GSM752661	ggo	lv	F	22299848
GSM752662	ggo	lv	M	25664046
GSM752696	hsa	br	M	2145713
GSM752707	hsa	ts	M	3933518
GSM752692	hsa	br	M	9305888
GSM752703	hsa	kd	M	12157025
GSM752701	hsa	ht	M	12557133
GSM752694	hsa	br	M	13052578
GSM752699	hsa	ht	F	13115883
GSM752700	hsa	ht	M	14363375
GSM752702	hsa	kd	F	14780007
GSM752704	hsa	kd	M	15587456
GSM752691	hsa	br	F	15713369
GSM752706	hsa	lv	M	15724554
GSM752697	hsa	cb	F	18974942
GSM752708	hsa	ts	M	20019814
GSM752698	hsa	cb	M	23395954
GSM752705	hsa	lv	M	24466576
GSM752690	ppa	ts	M	7470392
GSM752689	ppa	lv	M	10507559
GSM752685	ppa	ht	M	13320503
GSM752680	ppa	br	F	14135815
GSM752684	ppa	ht	F	14282705

GSM752687	ppa	kd	M	15046031
GSM752686	ppa	kd	F	16010719
GSM752682	ppa	cb	F	16159020
GSM752683	ppa	cb	M	16411415
GSM752688	ppa	lv	F	18219997
GSM752681	ppa	br	M	22253609
GSM752646	ppy	cb	F	11143394
GSM752648	ppy	ht	M	13409196
GSM752649	ppy	kd	F	14743211
GSM752651	ppy	lv	F	15581764
GSM752650	ppy	kd	M	16265868
GSM752644	ppy	br	F	18775829
GSM752647	ppy	ht	F	19782162
GSM752652	ppy	lv	M	22657746
GSM752677	ptr	lv	M	8012561
GSM752664	ptr	br	F	8738268
GSM752671	ptr	cb	M	9975348
GSM752678	ptr	ts	M	11938608
GSM752674	ptr	kd	F	17641193
GSM752670	ptr	cb	F	17948479
GSM752672	ptr	ht	F	18475125
GSM752676	ptr	lv	F	19954292
GSM752673	ptr	ht	M	20050737
GSM752675	ptr	kd	M	24779565

Reads were mapped to each of their respective species' reference genomes separately with the Bowtie 2 aligner using TopHat(Trapnell et al. 2012; Langmead and Salzberg 2012). Transcripts were then assembled using the Cufflinks package(Trapnell et al. 2012) (**Table 6.4**). Both TopHat and Cufflinks were run with default parameters, thus, only transcripts supported by at least 10 RNAseq fragments were reported. Additionally we required transcripts to show intron-exon structure to minimize false positives. Among the identified transcripts, we rediscovered 2/3 annotated loci described above. No evidence of the CD209L2 annotated transcript was found. In macaque CD209L2 is expressed in the liver, spleen, lymph nodes, heart, and skin; however, its paralogs are specifically expressed in dendritic cells and lymph nodes, suggesting that the transcript may simply be absent from the tissues assessed here. We identified 115 deletions along the human lineage encompassing expressed transcripts. Of the 115 regions containing expressed transcripts, 57 (49.6%) additionally contained conserved elements, representing a ~6-fold enrichment for conserved elements in regions containing an expressed transcript compared to genomic background.

Table 6.4: Fixed deletions absent from the human reference genome containing *de novo* assembled transcripts.

Species Reference	Regions containing expressed transcripts (previously annotated transcripts identified)	Regions containing an expressed transcript and conserved elements
ponAbe2	83 (2)	38
gorGor3	1	0
panTro3	57 (1)	19

Great ape gene deletions: We assessed lineage-specific gene deletions throughout the great ape lineage for underlying gene content. For deletions identified from mappings to the human reference genome, RefSeq gene models were used. As nonhuman primate reference genomes are not as well annotated, for assessing the genic content of deletions identified among these references, we use *de novo* assembled transcripts that contain an open reading frame (ORF). ORFs were identified within assembled transcripts using the TransDecoder package (http://transdecoder.sourceforge.net/). In total, we identified 340 genes in which at least one exon was specifically deleted in the great ape lineage (**Table 6.5**). 29 gene deletions were present in multiple lineages and inconsistent with the overall species tree hence representing cases of either homoplasy or incomplete lineage sorting.

Table 6.5: Lineage-specific genic deletion counts in the great ape lineage. Gene counts represent the total number of fixed lineage-specific genes containing at least a single exon deletion. Rates are calculated per million years (MY) along each branch. A surfeit of gene deletions is observed along the chimpanzee-bonobo lineage.

lineage	genes	exons	genes >50% deleted	gene deletions/MY
Gbeg	1	1	0	0.833333
Gbeg-Ggod-Ggog	35	128	21	5.30303
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	40	67	32	4.761905
Ggod	41	120	11	45.555556
Ggod-Ggog	2	3	0	10
Hde	11	24	6	13.75
Hde-Hsa	37	61	20	6.166667
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	6	6	4	4.285714
Pab	7	7	5	3.888889
Pab-Ppy	90	149	64	6.206897
Ppa	7	18	5	3.5
Ppa-Ptre-Ptrs-Ptrt-Ptrv	57	144	34	12.954545
Ppy	4	5	4	2.222222
Ptre	1	1	1	1.111111
Ptre-Ptrs-Ptrt-Ptrv	1	1	0	1.111111
Ptre-Ptrv	0	0	0	0
Ptrs	1	1	0	1
Ptrs-Ptrt	0	0	0	0
ils_Gbeg-Ggod-Ggog-Hde-Hsa	3	3	2	-
ils_Gbeg-Ggod-Ggog-Pab-Ppy	7	11	6	-
ils_Gbeg-Ggod-Ggog-Ppa-Ptre-Ptrs-Ptrt-Ptrv	20	55	13	-

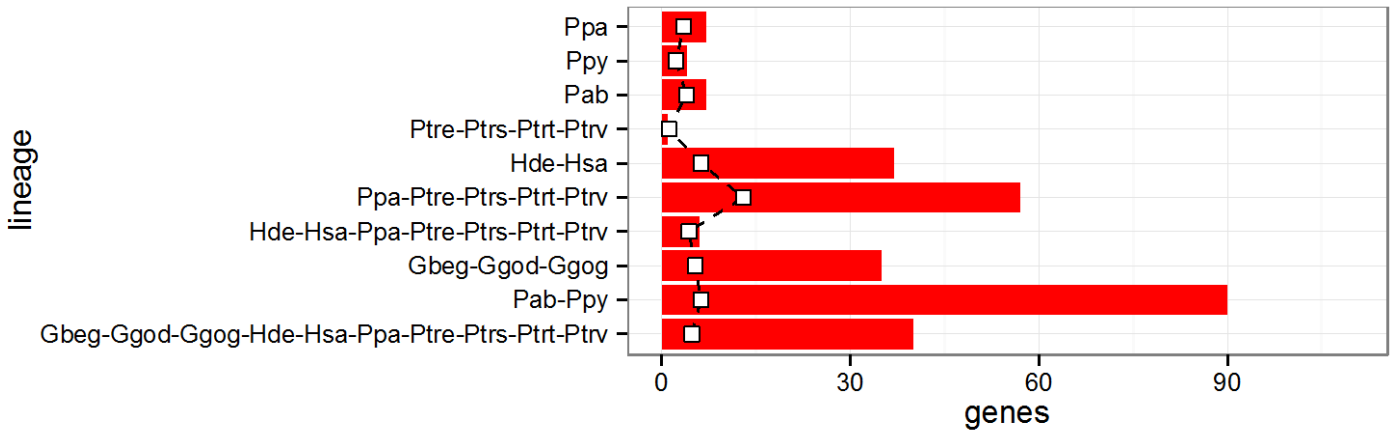


Figure 6.3: A histogram of the number of genes with exons deleted per lineage is plotted in red. White boxes connected by black dotted lines represent the number of gene deletions per million years in each lineage. A 3-fold acceleration in the number of gene deletions is present in the chimpanzee-bonobo lineage.

A list of deleted genes is provided in supplementary **Table 6.6.**

<TABLE 6.6>

Homology of human lineage gene losses: To determine if any of the ORFs deleted along the human lineage had homology to genes identified in other organisms or had been previously predicted as genes, we searched them against the NCBI RefSeq protein database. Of the 86 ORFs lost along the human lineage, 60 contained homology to 42 different genes (**Table 6.7**).

Table 6.7: BLAST results of searching ORFs lost along the human lineage against the RefSeq protein database. We identified 42 significant hits to previously predicted genes/homologs.

lineage	reference	position	exons lost	BLAST result	e-value	HCEs
Hde-Hsa	panTro3	chr10:133269290-133306499	1	gi|397467791|ref|XP_003805587.1| PREDICTED: putative scavenger receptor cysteine-rich domain-containing protein LOC619207-like [Pan paniscus]	0.00E+00	30
Hde-Hsa	panTro3	chr16:21145204-21181787	3	gi|402907936|ref|XP_003916716.1| PREDICTED: ATP-binding cassette sub-family A member 3-like [Papio anubis]	8.65E-51	4
Hde-Hsa	panTro3	chr11:58411158-58441246	2	gi|402893240|ref|XP_003909808.1| PREDICTED: membrane-spanning 4-domains subfamily A member 5-like [Papio anubis]	3.52E-19	2
Hde-Hsa	ponAbe2	chr2a:16396704-16432475	1	gi|297667005|ref|XP_002811788.1| PREDICTED: sulfotransferase 1C3-like [Pongo abelii]	1.96E-119	6
Hde-Hsa	panTro3	chr19:43024841-43025176	1	gi|403293043|ref|XP_003937533.1| PREDICTED: WD repeat-containing protein 87-like [Saimiri boliviensis boliviensis]	6.79E-44	0
Hde-Hsa	panTro3	chr2B:171197850-171216575	1	gi|397507773|ref|XP_003824361.1| PREDICTED: xin actin-binding repeat-containing protein 2 [Pan paniscus]	8.54E-77	14
Hde-Hsa	ponAbe2	chr1:282764-292085	9	gi|297281586|ref|XP_002802121.1| PREDICTED: hypothetical protein LOC100427314 [Macaca mulatta]	1.51E-84	0
Hde-Hsa	ponAbe2	chr19:49725476-49755401	1	gi|426389531|ref|XP_004061173.1| PREDICTED: caspase recruitment domain-containing protein 8 [Gorilla gorilla gorilla]	0.00E+00	0
Hde-Hsa	ponAbe2	chr19:53072028-53077831	3	gi|110835743|ref|NP_001036087.1| sialic acid-binding Ig-like lectin 13 precursor [Pan troglodytes]	0.00E+00	4
Hde-Hsa	panTro3	chr6:28457142-28462847	1	gi|55626286|ref|XP_527293.1| PREDICTED: 60S ribosomal protein L30-like [Pan troglodytes]	6.23E-70	1
Hde-Hsa	panTro3	chr2A:4996171-5029725	1	gi|114575954|ref|XP_001152938.1| PREDICTED: trafficking protein particle complex subunit 12 isoform 6 [Pan troglodytes] >gi|114575956|ref|XP_001153001.1| PREDICTED: trafficking protein particle complex subunit 12 isoform 7 [Pan troglodytes]	5.05E-131	15
Hde-Hsa	panTro3	chr19:43029614-43030908	1	gi|332206687|ref|XP_003252428.1| PREDICTED: hypothetical protein LOC100595862 [Nomascus leucogenys]	0.00E+00	1
Hde-Hsa	panTro3	chr7:130910015-130968051	1	gi|426357869|ref|XP_004046252.1| PREDICTED: protein FAM40B [Gorilla gorilla gorilla]	0.00E+00	62
Hde-Hsa	panTro3	chr14:21311933-21312805	2	gi|403264885|ref|XP_003924697.1| PREDICTED: uncharacterized protein LOC101031436 [Saimiri boliviensis boliviensis]	4.98E-12	1
Hde-Hsa	panTro3	chr19:56318609-56324555	2	gi|332241162|ref|XP_003269753.1| PREDICTED: LOW QUALITY PROTEIN: sialic acid-binding Ig-like lectin 13-like [Nomascus leucogenys]	8.88E-131	0
Hde-Hsa	panTro3	chr8:912532-913202	1	gi|156369634|ref|XP_001628080.1| predicted protein [Nematostella vectensis]	3.86E-11	0
Hde-Hsa	panTro3	chr5:138992707-139049493	1	gi|297676086|ref|XP_002815977.1| PREDICTED: M-phase inducer phosphatase 3 isoform 3 [Pongo abelii]	0.00E+00	40
Hde-Hsa	ponAbe2	chr8:2039648-2040136	1	gi|156389414|ref|XP_001634986.1| predicted protein [Nematostella vectensis]	4.87E-12	0

Hde-Hsa	panTro3	chr16:21194196-21194660	2	gi|403277334|ref|XP_003930322.1| PREDICTED: ATP-binding cassette sub-family A member 3-like [Saimiri boliviensis boliviensis]	1.76E-48	3
Hde-Hsa	ponAbe2	chr22:42137786-42176624	1	gi|426394877|ref|XP_004063711.1| PREDICTED: ceramide kinase [Gorilla gorilla gorilla]	0.00E+00	16
Hde-Hsa	panTro3	chr6:52707714-52798280	1	gi|297678352|ref|XP_002817041.1| PREDICTED: fibrocystin, partial [Pongo abelii]	0.00E+00	102
Hde-Hsa	panTro3	chr20:23615403-23618960	3	gi|332263715|ref|XP_003280898.1| PREDICTED: cystatin-12-like [Nomascus leucogenys]	3.58E-97	1
Hde-Hsa	panTro3	chr6:41677792-41687234	2	gi|332824003|ref|XP_003311331.1| PREDICTED: adenylate cyclase type 10-like [Pan troglodytes]	2.33E-153	16
ils_Gbeg-Ggod-Ggog-Hde-Hsa	panTro3	chr1:190489209-190523588	1	gi|397486258|ref|XP_003814247.1| PREDICTED: putative uncharacterized protein encoded by LINC00467-like [Pan paniscus] >gi|410034417|ref|XP_001170038.2| PREDICTED: putative uncharacterized protein encoded by LINC00467-like [Pan troglodytes]	2.93E-44	0
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr12:96567618-96581976	1	gi|359065249|ref|XP_002687265.2| PREDICTED: uncharacterized protein LOC100336892 [Bos taurus]	1.99E-24	4
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr22:17209875-17273295	1	gi|395753054|ref|XP_003779528.1| PREDICTED: immunoglobulin omega chain-like [Pongo abelii]	8.11E-64	30
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr10:132702048-132702523	1	gi|410044564|ref|XP_003951836.1| PREDICTED: uncharacterized protein LOC101059682 [Pan troglodytes]	4.63E-16	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr11:101294540-101308678	2	gi|395743436|ref|XP_003777927.1| PREDICTED: caspase-12-like [Pongo abelii]	0.00E+00	9
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr2b:134898141-134898734	2	gi|291233023|ref|XP_002736453.1| PREDICTED: hypothetical protein [Saccoglossus kowalevskii]	2.14E-10	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr2a:112233060-112237384	2	gi|259500926|ref|ZP_05743828.1| conserved hypothetical protein [Lactobacillus iners DSM 13335]	7.09E-10	1
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr18:76813671-76836046	3	gi|395749962|ref|XP_002828334.2| PREDICTED: uncharacterized protein LOC100442861 [Pongo abelii]	0.00E+00	9
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr18:17861680-17863215	1	gi|338728618|ref|XP_003365712.1| PREDICTED: zinc finger protein 791-like [Equus caballus]	9.54E-25	2
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr11:1058088-1058715	1	gi|50290505|ref|XP_447684.1| hypothetical protein [Candida glabrata CBS 138]	3.61E-09	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr16:61628302-61629469	1	gi|395748070|ref|XP_003778708.1| PREDICTED: exosome complex component MTR3 [Pongo abelii]	7.26E-80	2
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr8:153430204-153431184	1	gi|46094068|ref|NP_689818.2| zinc finger protein 781 [Homo sapiens]	1.26E-10	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr7:149328103-149493757	1	gi|33186925|ref|NP_057287.2| 5'-AMP-activated protein kinase subunit gamma-2 isoform a [Homo sapiens] >gi|332870104|ref|XP_003318972.1| PREDICTED: 5'-AMP-activated protein kinase subunit gamma-2 isoform 1 [Pan troglodytes]	7.45E-113	28
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr19:40740646-40741970	1	gi|402905533|ref|XP_003915572.1| PREDICTED: eosinophil lysophospholipase-like [Papio anubis]	1.75E-04	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr6:27923306-27923676	1	gi|350586447|ref|XP_001928408.4| PREDICTED: hypothetical protein LOC100155756 [Sus scrofa]	1.69E-07	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr18:17828711-17861419	4	gi|426387972|ref|XP_004060436.1| PREDICTED: zinc finger protein 14 [Gorilla gorilla gorilla]	2.92E-36	2
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr1:67094355-67101144	1	gi|114568366|ref|XP_001163353.1| PREDICTED: neutrophil cytosol factor 2 isoform 3 [Pan troglodytes] >gi|114568368|ref|XP_001163464.1| PREDICTED: neutrophil cytosol factor 2 isoform 5 [Pan troglodytes]	5.01E-49	3
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr19:40600114-40600861	1	gi|297277044|ref|XP_001088117.2| PREDICTED: galactoside-binding soluble lectin 13-like [Macaca mulatta]	3.79E-09	0
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr19:53342284-53344161	1	gi|426389932|ref|XP_004061370.1| PREDICTED: zinc finger protein 649 [Gorilla gorilla gorilla]	0.00E+00	3
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chr2b:90877161-90887811	1	gi|109100506|ref|XP_001089951.1| PREDICTED: formimidoyltransferase-cyclodeaminase-like [Macaca mulatta]	6.42E-112	16
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	ponAbe2	chrUn:28367950-28368823	1	gi|297265250|ref|XP_001107481.2| PREDICTED: hypothetical protein LOC716637 [Macaca mulatta]	4.68E-36	0

Section 7: Gene duplication analysis

Gene duplications: Lineage-specific duplication segments were each assessed for their underlying gene content. RefSeq gene models, collapsed by 80% reciprocal overlap, were used for this analysis. Since many duplicated genes are represented multiple times within the human reference, we avoid over counting lineage-specific gene duplicates for each set of genes identified in a lineage by performing pairwise *bl2seq* (NCBI) alignments between all genic, duplicated segments. If two duplicated genes were found to align to each other with >90% identity reciprocally over at least 10% of their length, they were clustered together as paralogous. Finally, paralogous gene clusters were manually curated.

In total, 405 genes were identified as lineage-specific duplications with the criteria that at least 50% of the gene model be duplicated (**Figure 7.1, Figure 7.2, Table 7.1**)**.** An increased rate of gene duplication was observed in the African great ape, gorilla, and the human-chimpanzee ancestors with the rate in the human-chimpanzee ancestor showing an ~1.5-fold higher rate than the next highest rate of duplication—the African great ape ancestor (19.3 genes/MY versus 13 genes/MY; **Table 7.2**). **Table 7.1** contains a list of all lineage-specific duplicated genes in great apes. We additionally identified 103 genes duplicated in multiple lineages yet expanded to markedly higher copy in one particular lineage (**Table 7.3**). A heatmap of gene duplications, deletions, and expansions is shown in **Figure 7.3.**

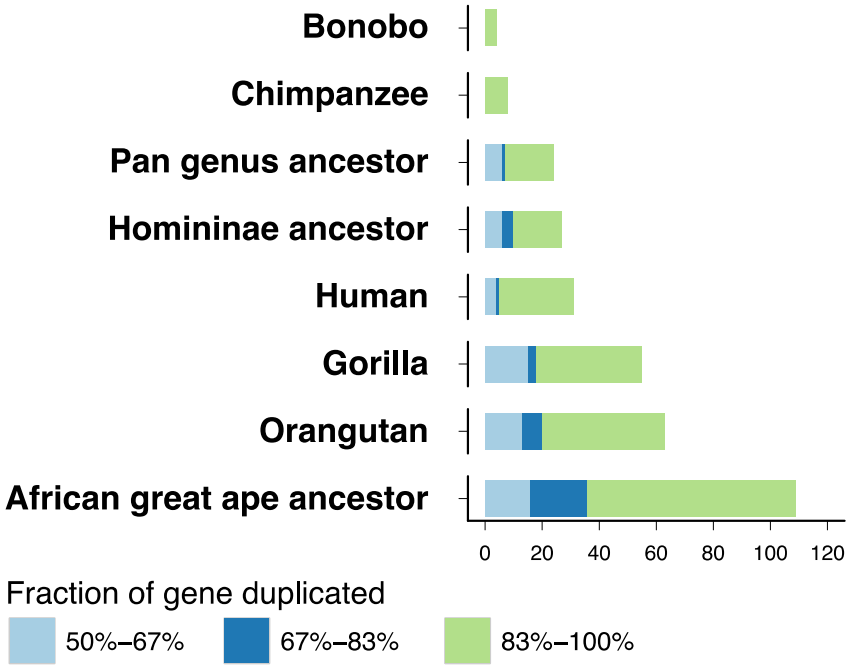


Figure 7.1: Bar chart of the number of lineage-specific gene duplications stratified by the fraction of the gene duplicated.

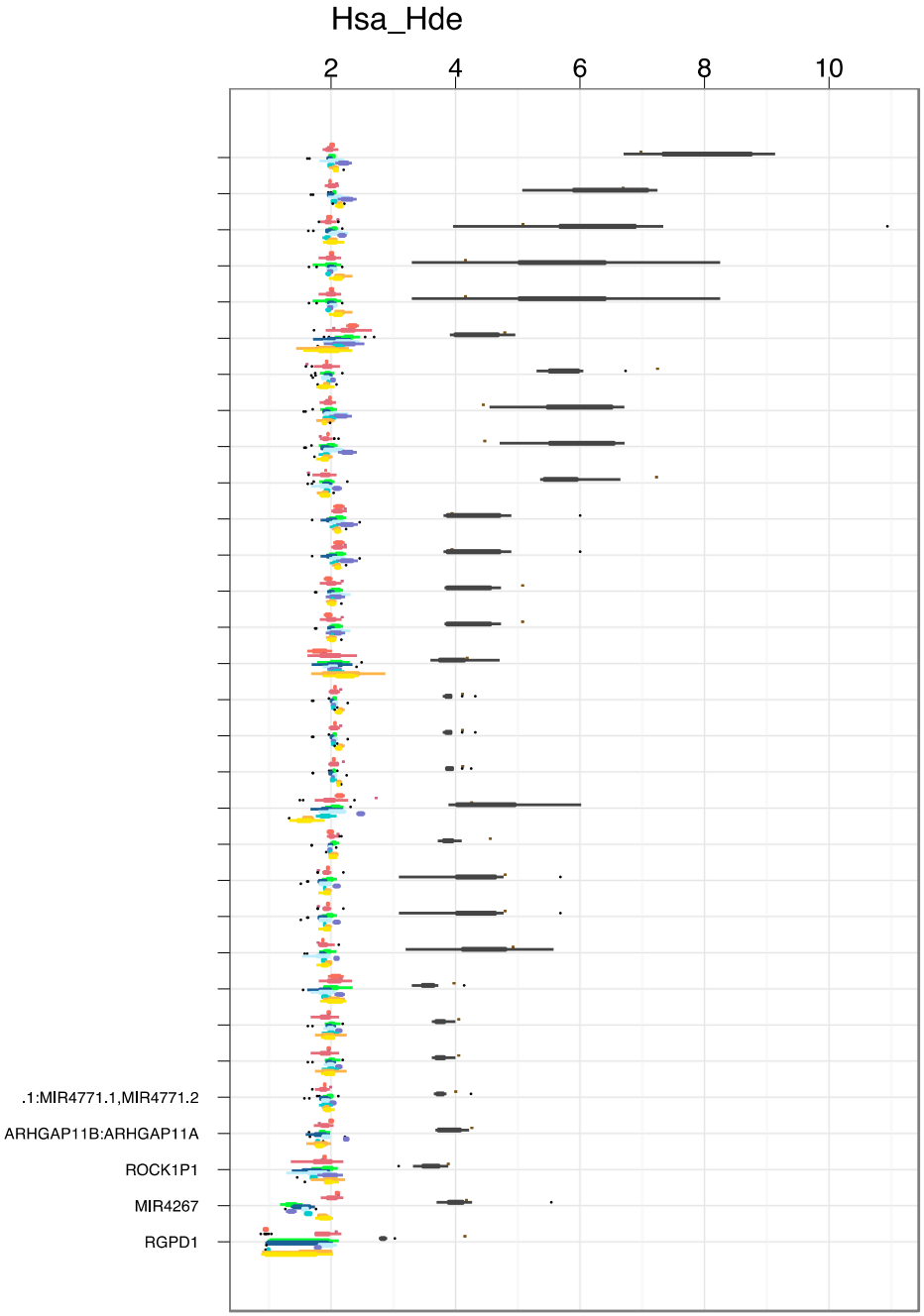
<TABLE 7.1>

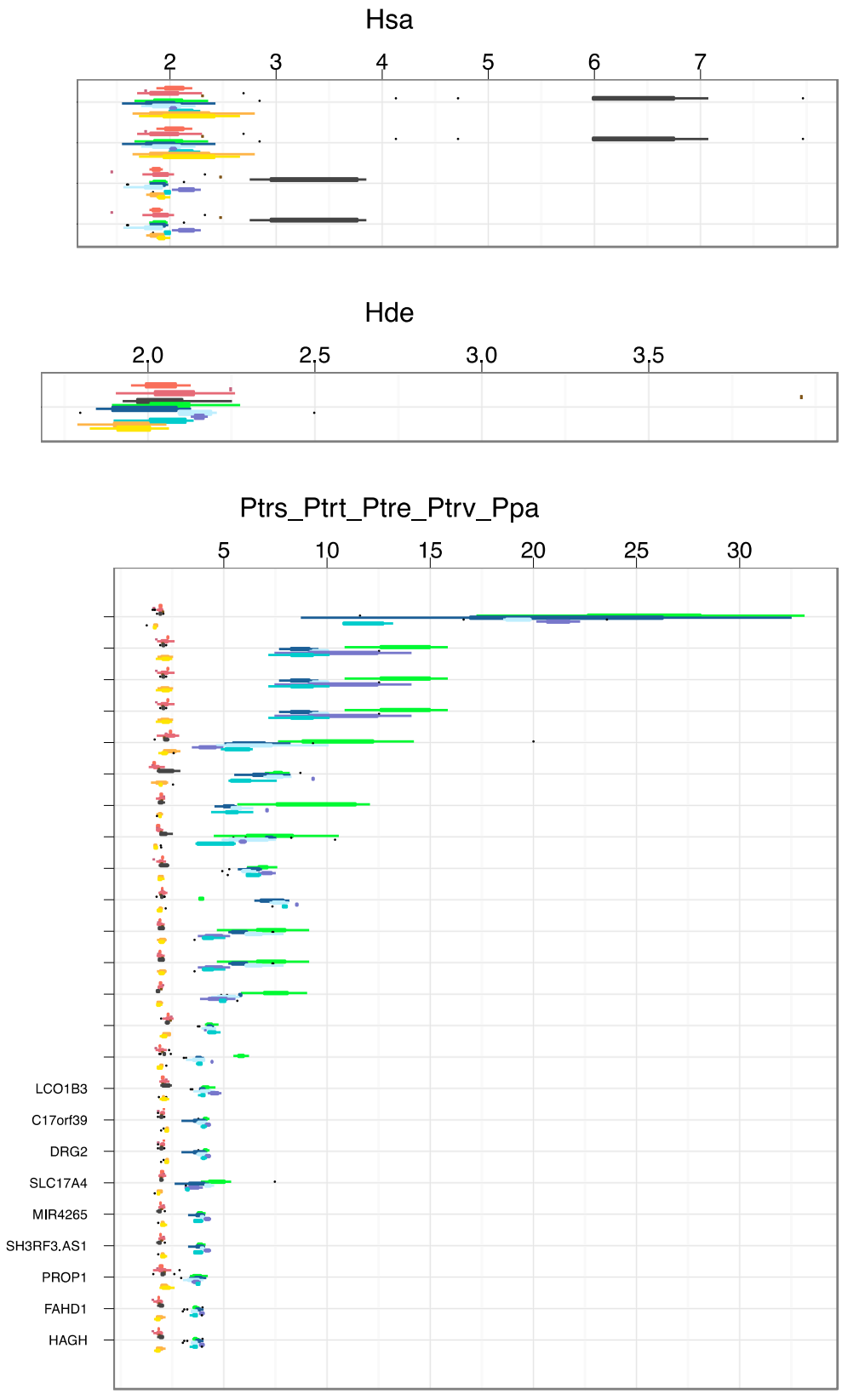
Table 7.2: Counts of lineage-specific gene duplications and rates per million years (MY) of gene duplication in each lineage.

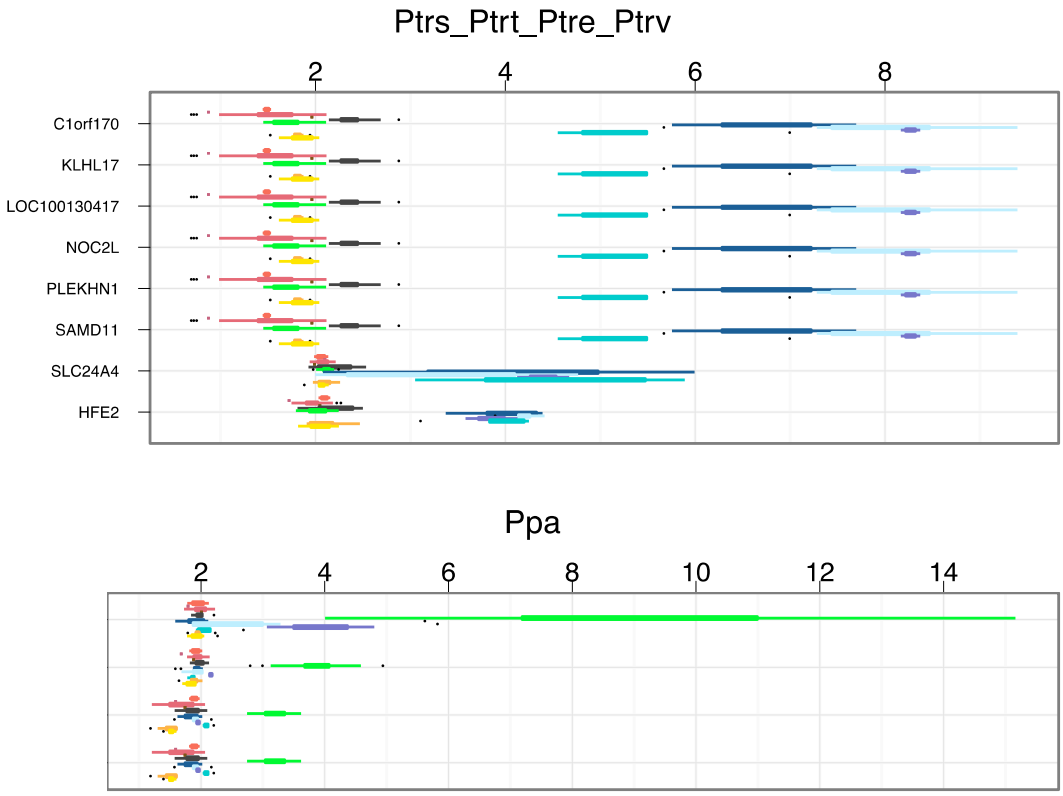
lineage	duplicated_genes	duplications/MY
Ggod-Ggog	2	10
Gbeg-Ggod-Ggog	55	8.333333333
Gbeg-Ggod-Ggog-Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	109	12.97619048
Hde	1	1.25
Hde-Hsa	31	5.166666667
Hde-Hsa-Ppa-Ptre-Ptrs-Ptrt-Ptrv	27	19.28571429
Ppa-Ptre-Ptrs-Ptrt-Ptrv	24	5.454545455
Ppa	4	2
Ppy	1	0.555555556
Pab-Ppy	63	4.344827586
Gbeg	1	0.833333333
Ptre-Ptrs-Ptrt-Ptrv	8	8.888888889
Hsa	4	5
Pab	2	1.111111111

<Table 7.3>

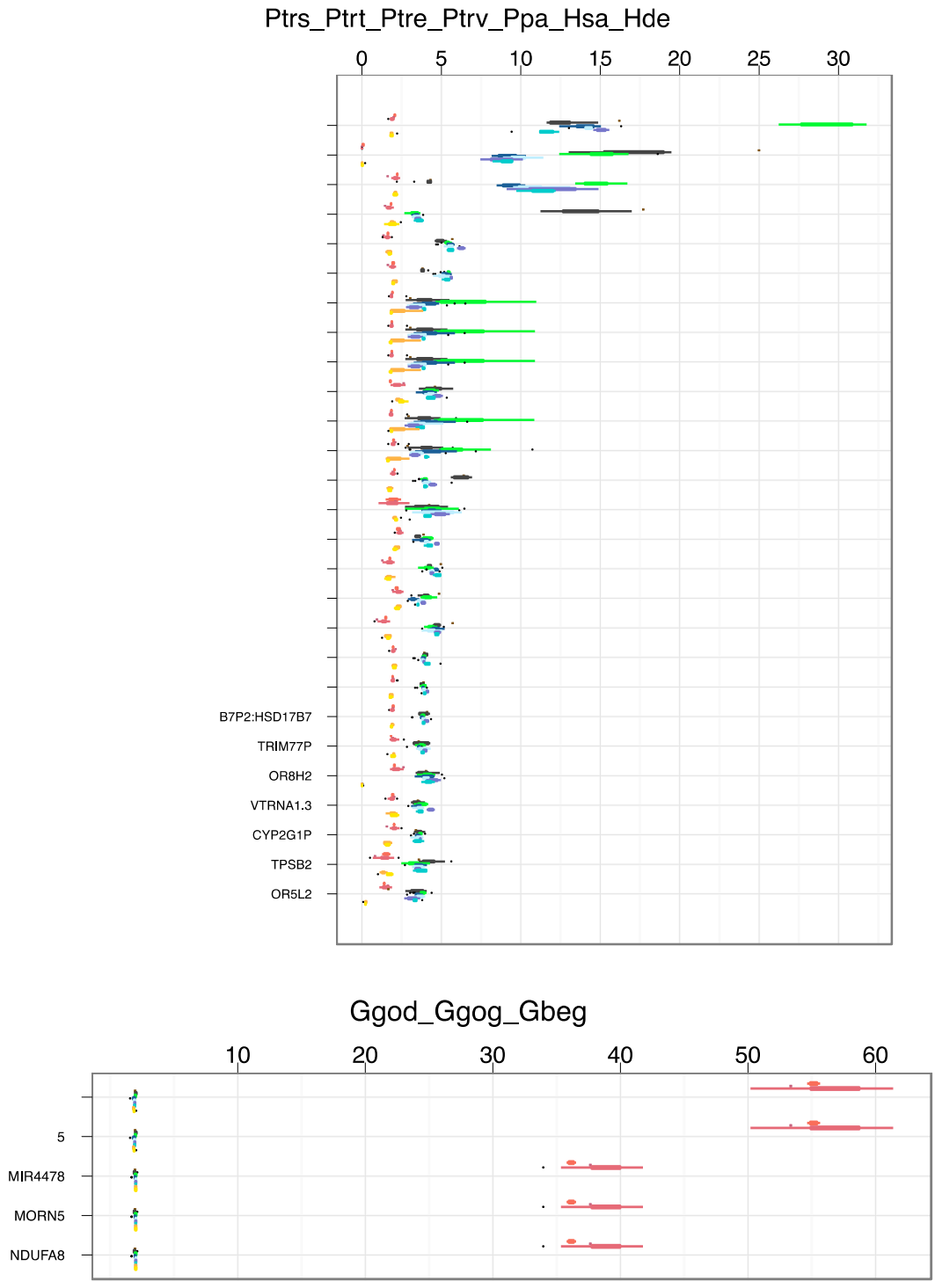
Figure 7.2 (following pages): Lineage-specific gene duplication boxplots for the copy number of each gene in all species assessed in this study.

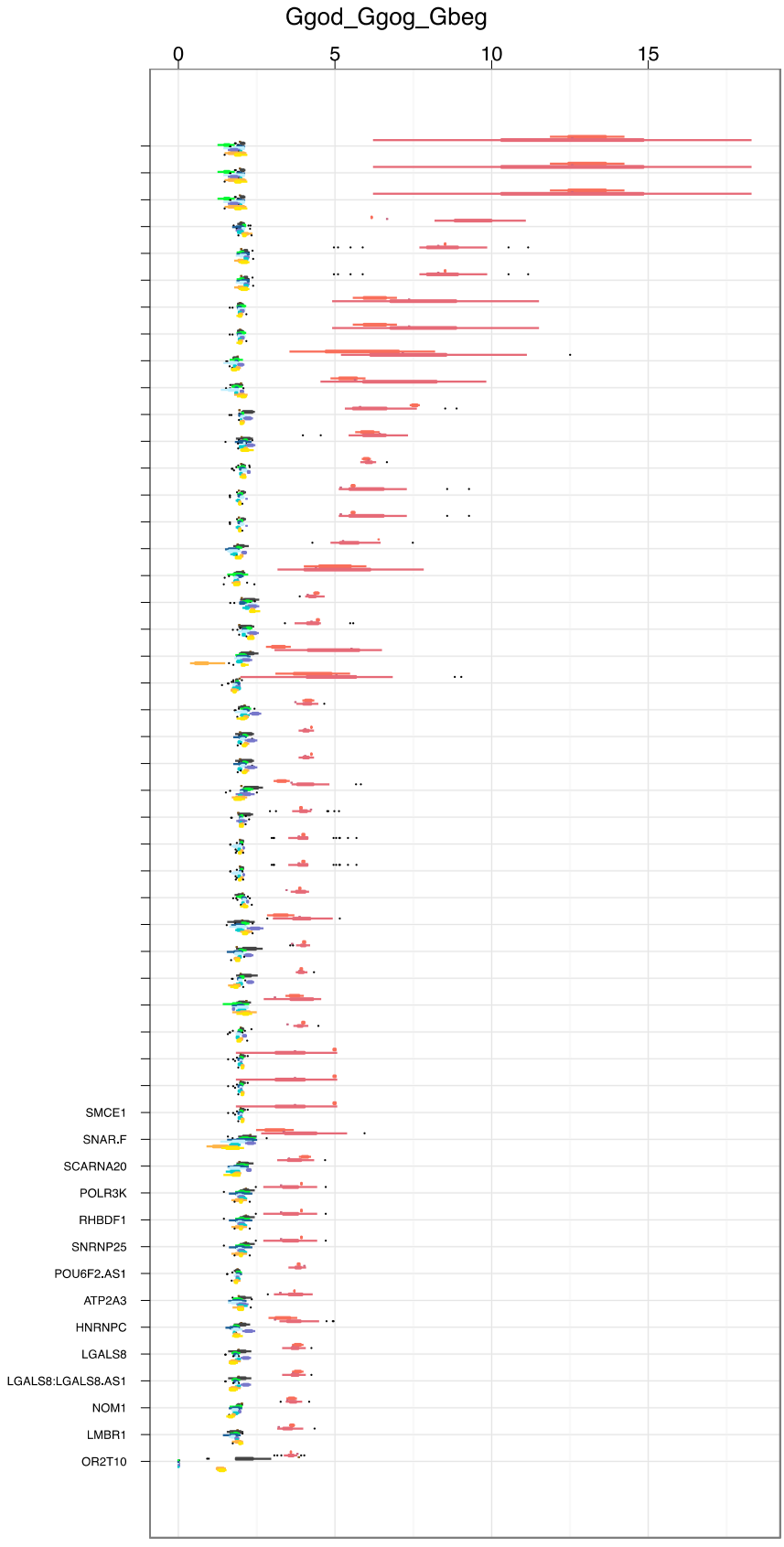


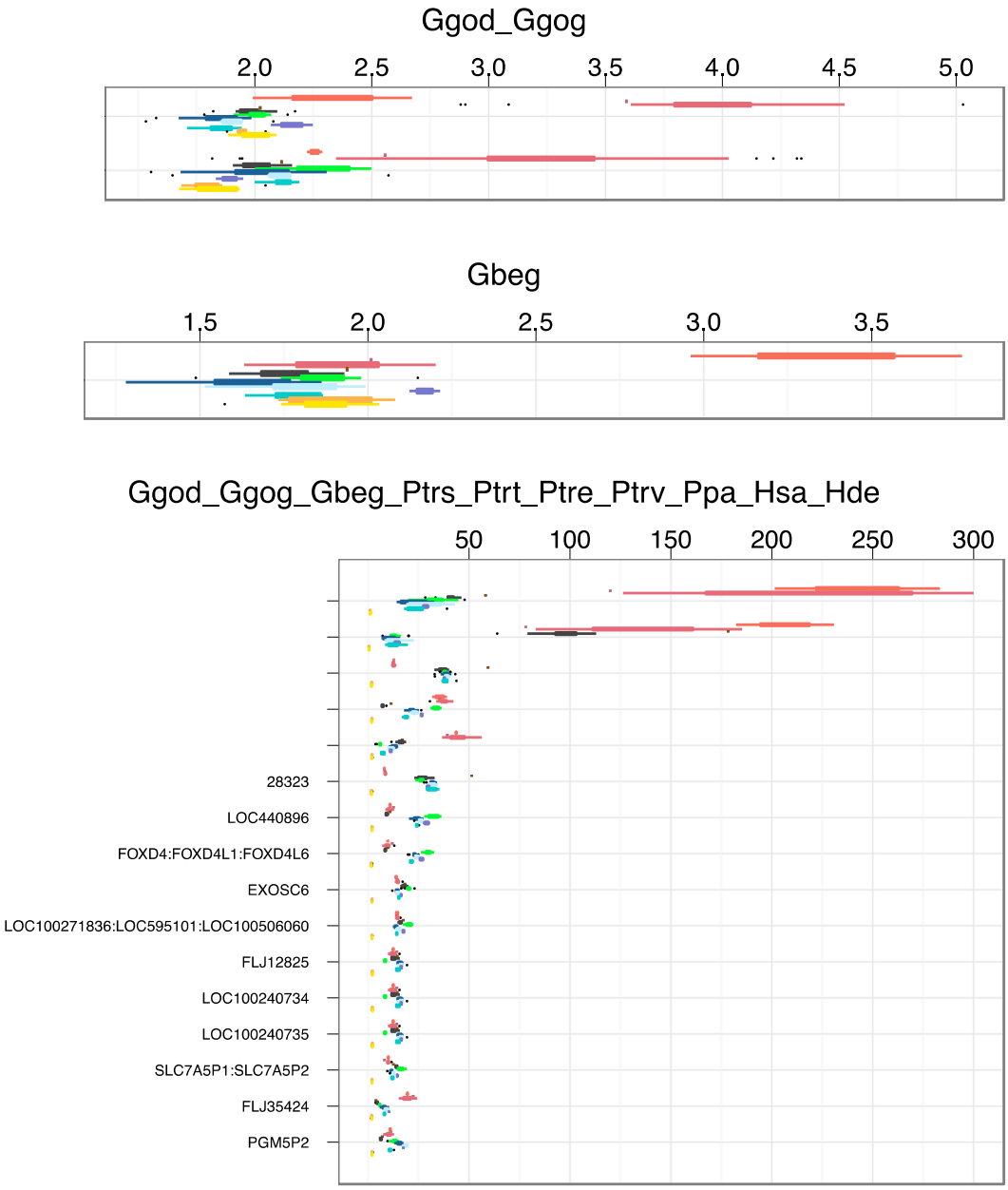

Hsa
2
3
4
5
6
7
Hde
2.0
2.5
3.0
3.5
Ptrs_Ptrt_Ptre_Ptrv_Ppa
5
10
15
20
25
30
LCO1B3
C17orf39
DRG2
SLC17A4
MIR4265
SH3RF3.AS1
PROP1
FAHD1
HAGH

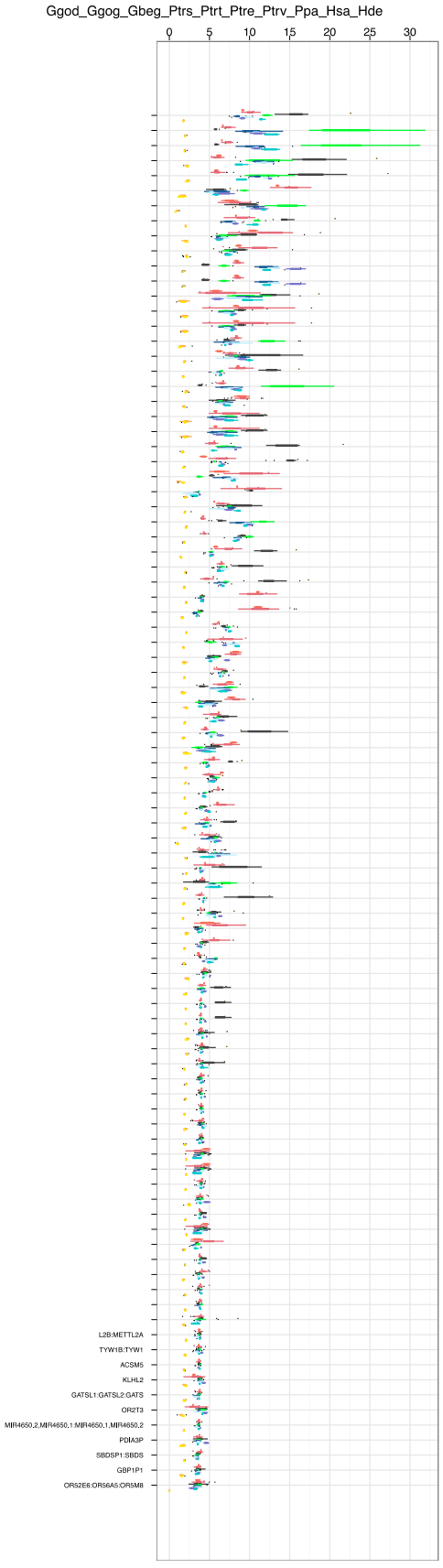

Ptrs_Ptrt_Ptre_Ptrv
2
4
6
8
C1orf170
KLHL17
LOC100130417
NOC2L
PLEKHN1
SAMD11
SLC24A4
HFE2

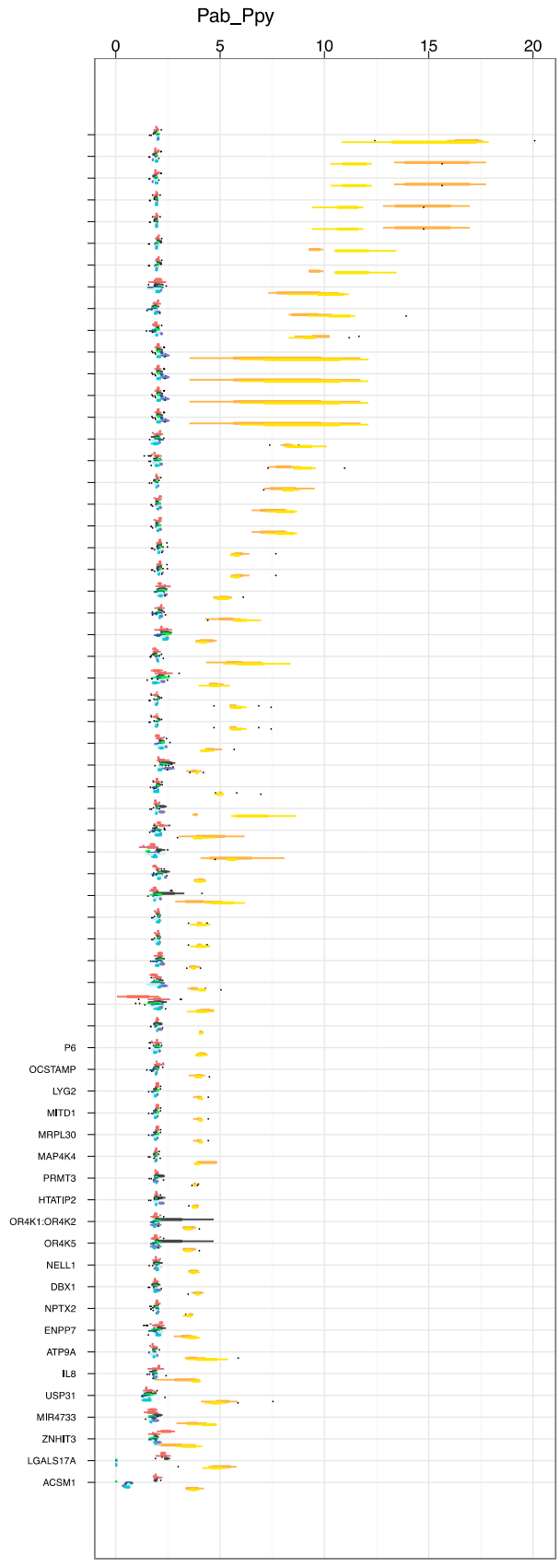
Ppa
2
4
6
8
10
12
14

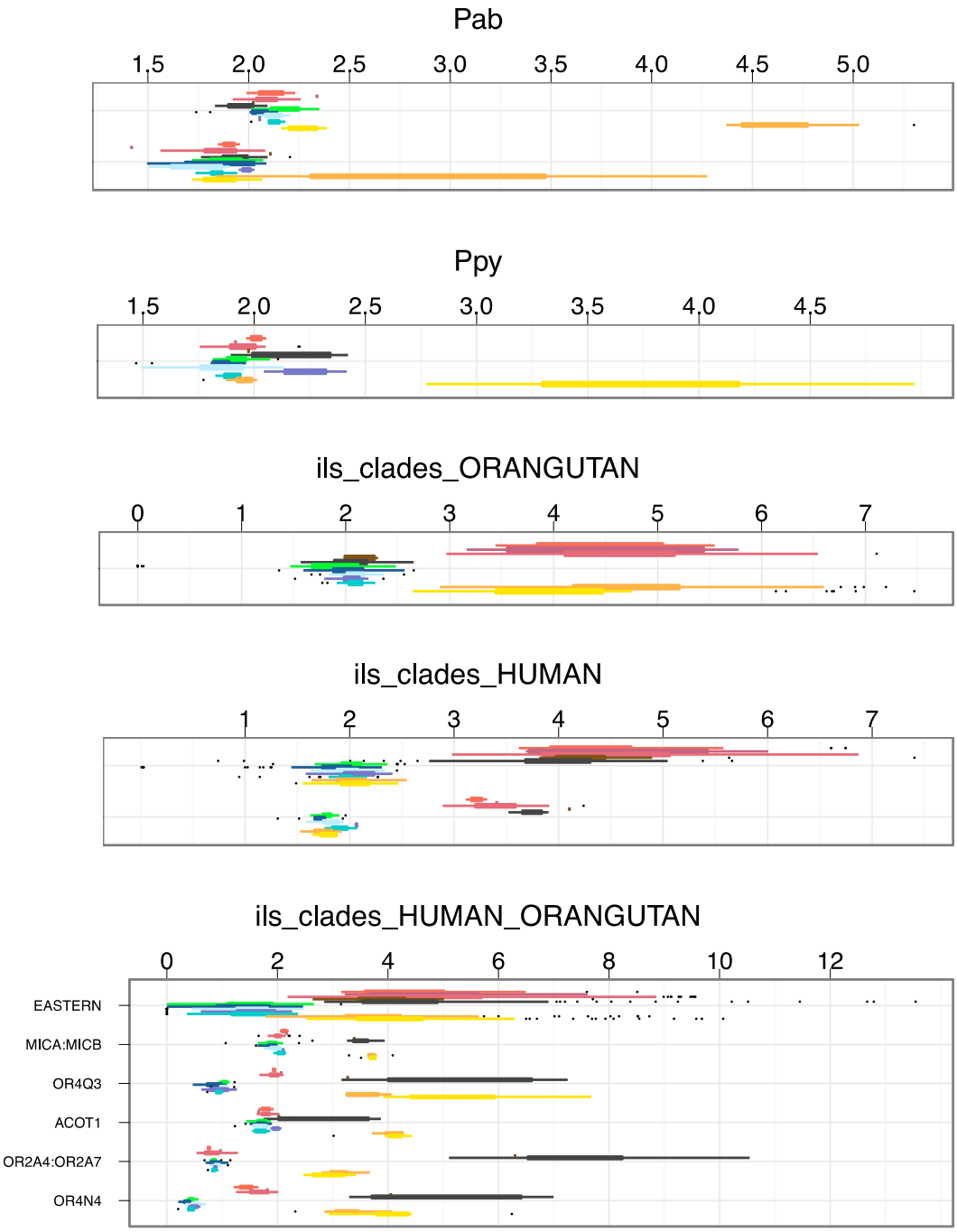

Ptrs_Ptrt_Ptre_Ptrv_Ppa_Hsa_Hde
0
5
10
15
20
25
30
B7P2:HSD17B7
TRIM77P
OR8H2
VTRNA1.3
CYP2G1P
TPSB2
OR5L2
Ggod_Ggog_Gbeg
10
20
30
40
50
60
5
MIR4478
MORN5
NDUFA8


Ggod_Ggog_Gbeg
0
5
10
15
SMCE1
SNAR.F
SCARNA20
POLR3K
RHBDF1
SNRNP25
POU6F2.AS1
ATP2A3
HNRNPC
LGALS8
LGALS8:LGALS8.AS1
NOM1
LMBR1
OR2T10

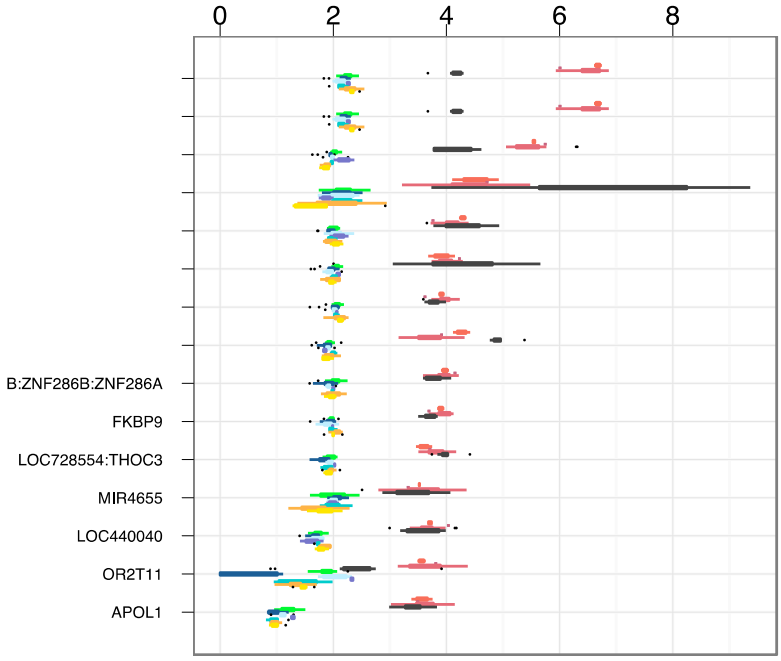

Ggod_Ggog
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Gbeg
1.5
2.0
2.5
3.0
3.5
Ggod_Ggog_Gbeg_Ptrs_Ptrt_Ptre_Ptrv_Ppa_Hsa_Hde
50
100
150
200
250
300
28323
LOC440896
FOXD4:FOXD4L1:FOXD4L6
EXOSC6
LOC100271836:LOC595101:LOC100506060
FLJ12825
LOC100240734
LOC100240735
SLC7A5P1:SLC7A5P2
FLJ35424
PGM5P2


Ggod_Ggog_Gbeg_Ptrs_Ptrt_Ptre_Ptrv_Ppa_Hsa_Hde
0
5
10
15
20
25
30
L2B:METTL2A
TYW1B:TYW1
ACSM5
KLHL2
GATSL1:GATSL2:GATS
OR2T3
MIR4650.2,MIR4650.1:MIR4650.1,MIR4650.2
PDIA3P
SBDSP1:SBDS
GBP1P1
OR52E6:OR56A5:OR5M8

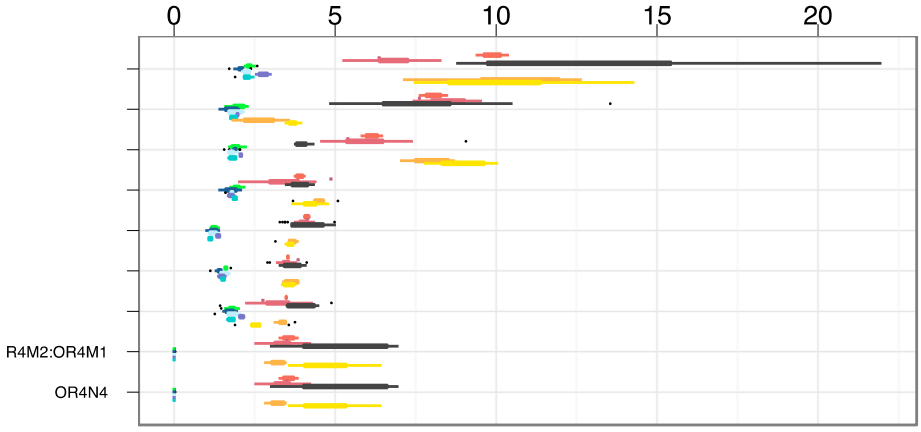

Pab_Ppy
0
5
10
15
20
P6
OCSTAMP
LYG2
MITD1
MRPL30
MAP4K4
PRMT3
HTATIP2
OR4K1:OR4K2
OR4K5
NELL1
DBX1
NPTX2
ENPP7
ATP9A
IL8
USP31
MIR4733
ZNHIT3
LGALS17A
ACSM1


Pab
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Ppy
1.5
2.0
2.5
3.0
3.5
4.0
4.5
ils_clades_ORANGUTAN
0
1
2
3
4
5
6
7
ils_clades_HUMAN
1
2
3
4
5
6
7
ils_clades_HUMAN_ORANGUTAN
0
2
4
6
8
10
12
EASTERN
MICA:MICB
OR4Q3
ACOT1
OR2A4:OR2A7
OR4N4

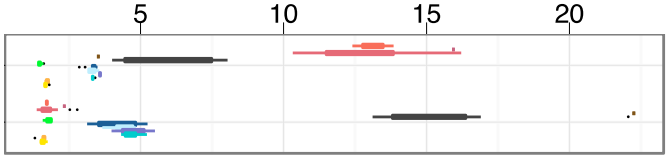
ils_clades_HUMAN_GORILLA



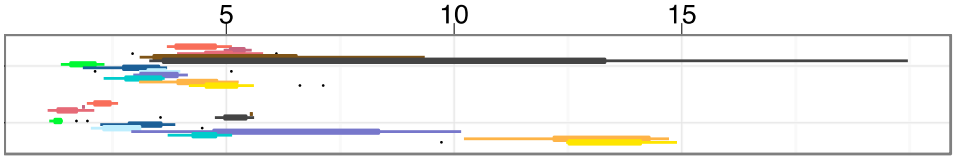
ils_clades_HUMAN_GORILLA_ORANGUTAN



ils_clades_HUMAN_CHIMP

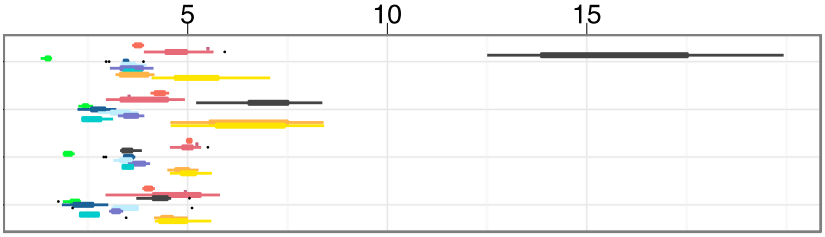


ils_clades_HUMAN_CHIMP_ORANGUTAN

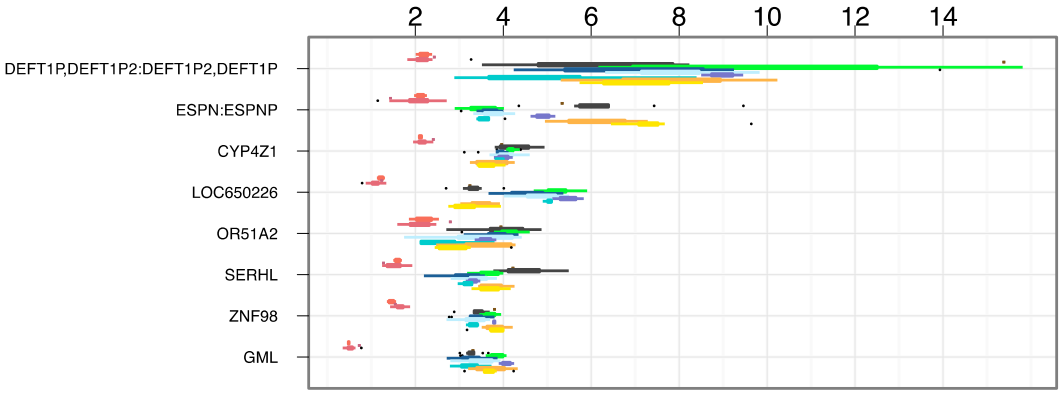


ils_clades_HUMAN_CHIMP_GORILLA

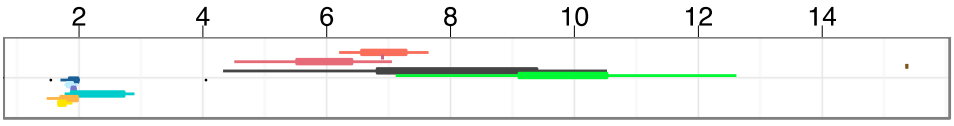
ils_clades_HUMAN_CHIMP_GORILLA_ORANGUTAN



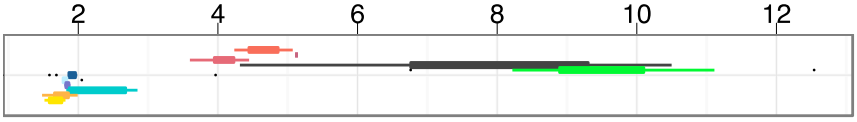
ils_clades_HUMAN_CHIMP_BONOBO_ORANGUTAN

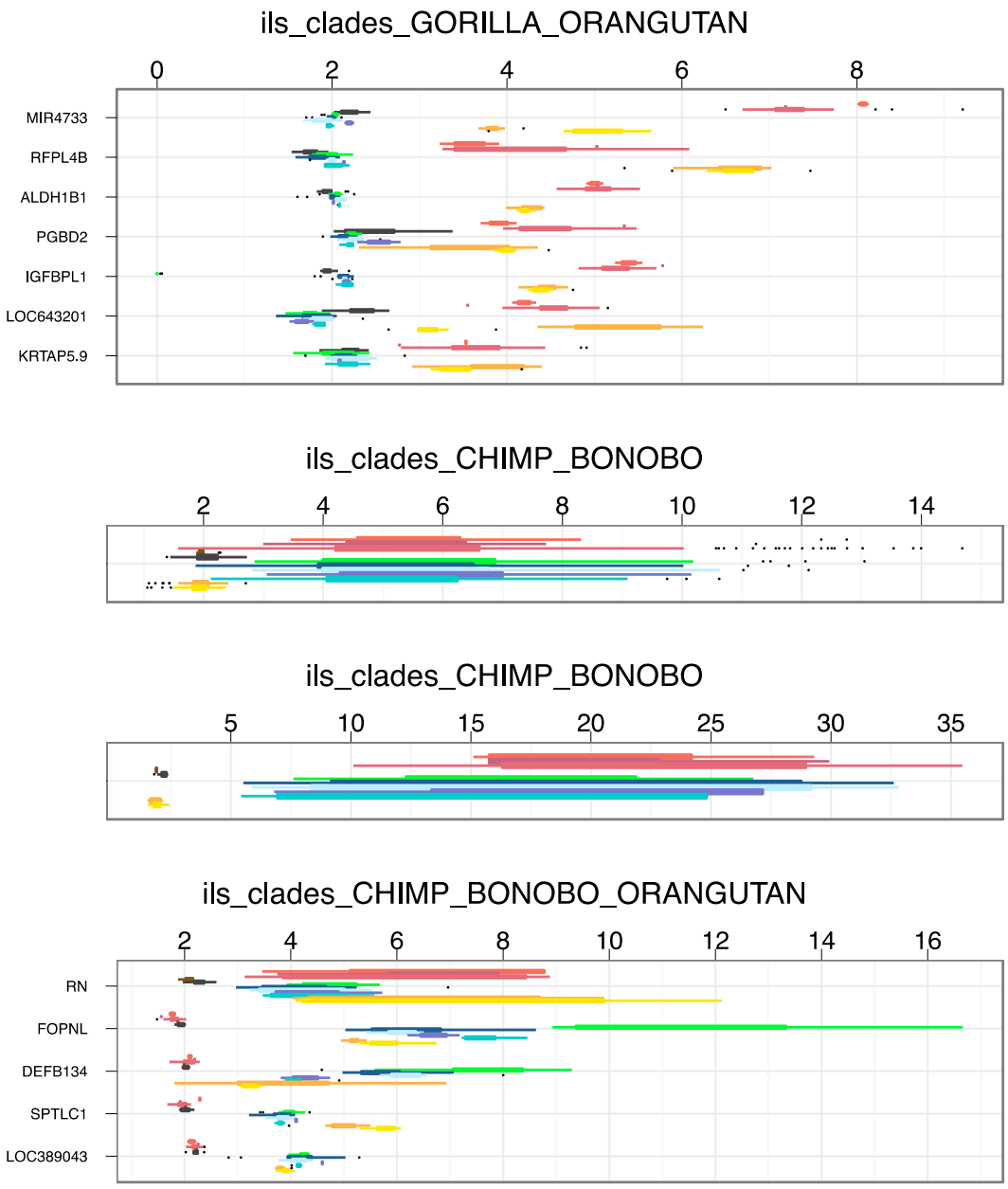


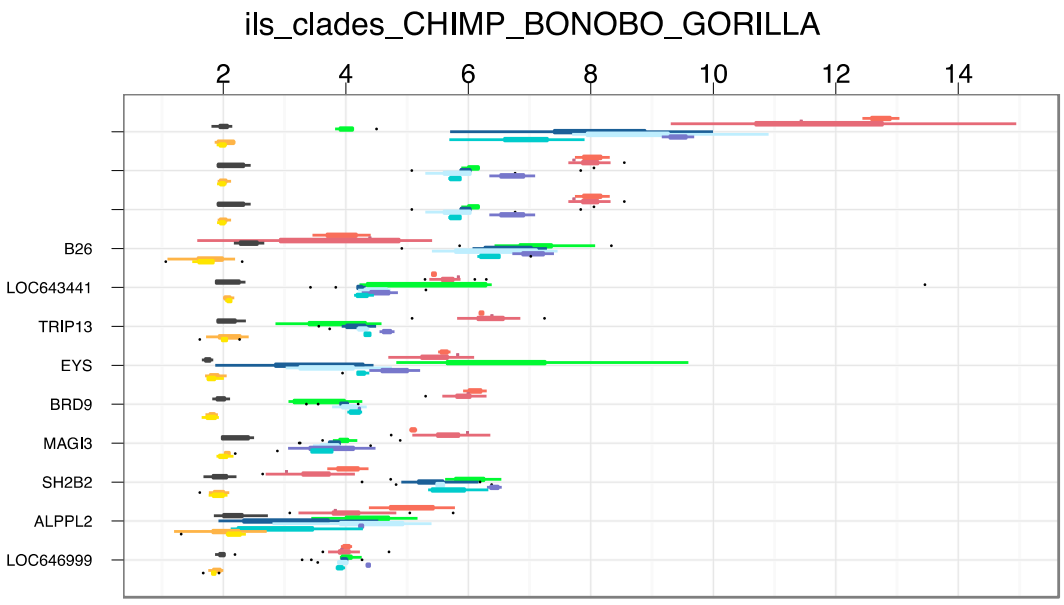
ils_clades_HUMAN_BONOBO

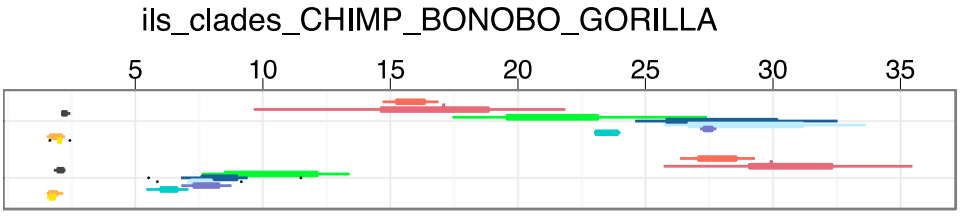


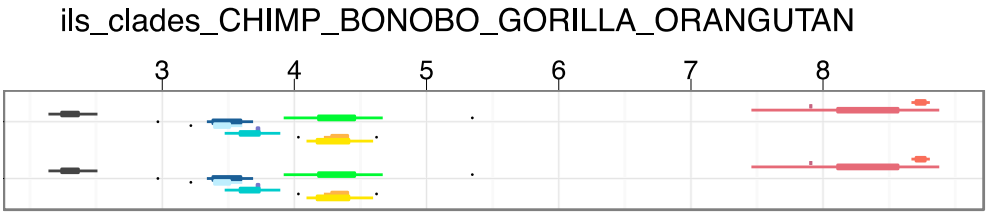
ils_clades_HUMAN_BONOBO_GORILLA

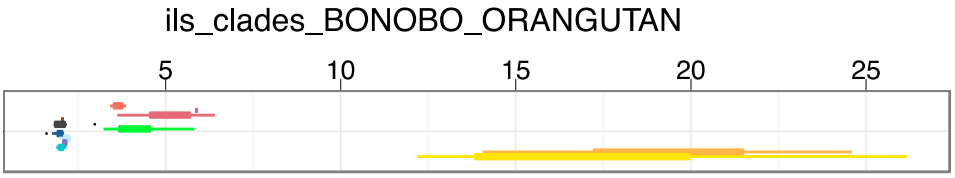


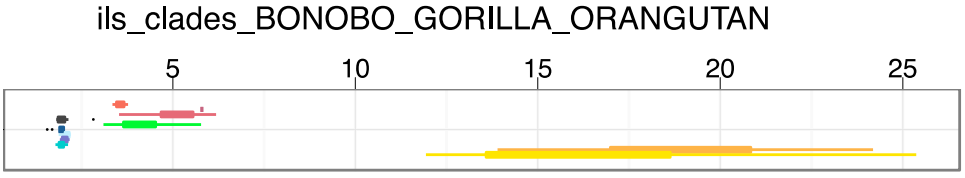

ils_clades_GORILLA_ORANGUTAN
0
2
4
6
8
MIR4733
RFPL4B
ALDH1B1
PGBD2
IGFBPL1
LOC643201
KRTAP5.9
ils_clades_CHIMP_BONOBO
2
4
6
8
10
12
14
ils_clades_CHIMP_BONOBO
5
10
15
20
25
30
35
ils_clades_CHIMP_BONOBO_ORANGUTAN
2
4
6
8
10
12
14
16
RN
FOPNL
DEFB134
SPTLC1
LOC389043


ils_clades_CHIMP_BONOBO_GORILLA
2
4
6
8
10
12
14
B26
LOC643441
TRIP13
EYS
BRD9
MAGI3
SH2B2
ALPPL2
LOC646999


ils_clades_CHIMP_BONOBO_GORILLA
5
10
15
20
25
30
35


ils_clades_CHIMP_BONOBO_GORILLA_ORANGUTAN
3
4
5
6
7
8


ils_clades_BONOBO_ORANGUTAN
5
10
15
20
25


ils_clades_BONOBO_GORILLA_ORANGUTAN
5
10
15
20
25

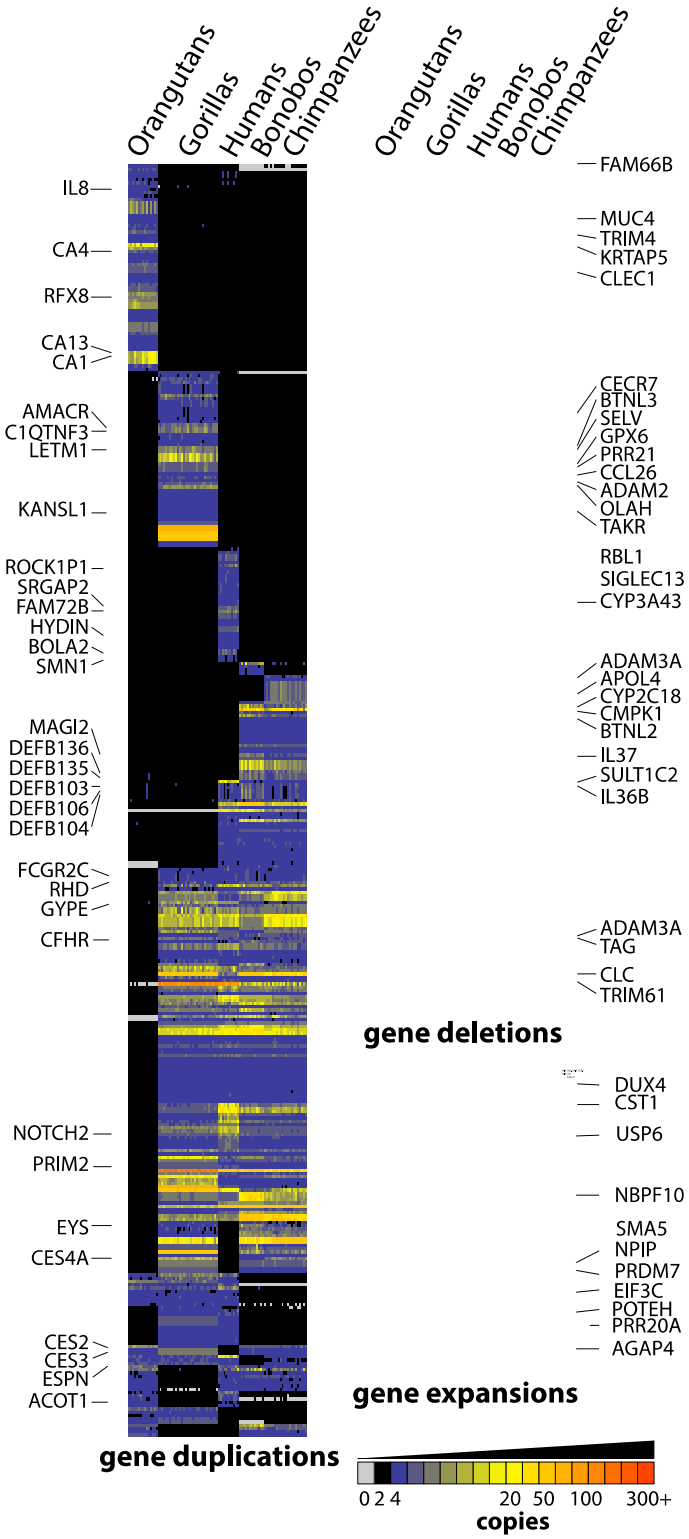


Figure 7.3: Heatmap representation of all lineage-specific gene duplications, deletions, and expansions identified throughout the great ape lineage.

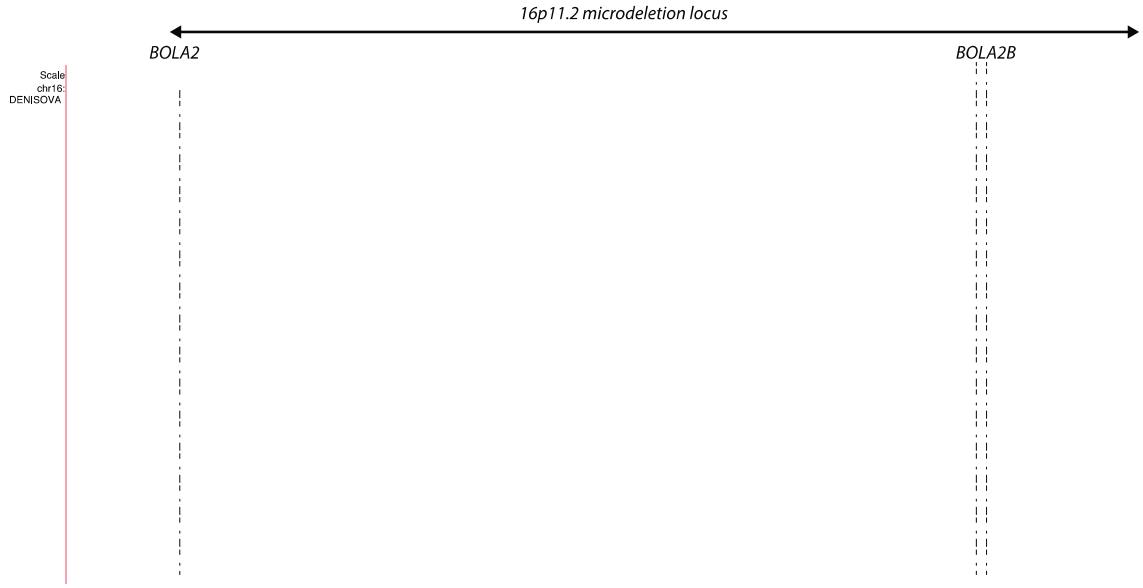


Figure 7.4: Heatmap of the 16p11.2 locus. The *BOLA2* and *BOLA2B* gene are highlighted in addition to the approximate breakpoints of the 16p11.2 micro-duplication/micro-deletion. *BOLA2* exhibits the ancestral copy number state.

Section 8: Distribution of duplication and deletion events

We assessed the position of lineage-specific gene duplications in relation to each other along the genome and found them to be significantly non-uniformly distributed (**Figures 8.1, 8.2 a,b,c,d,e**). Duplicated genes instead clustered into distinct cores of activity.

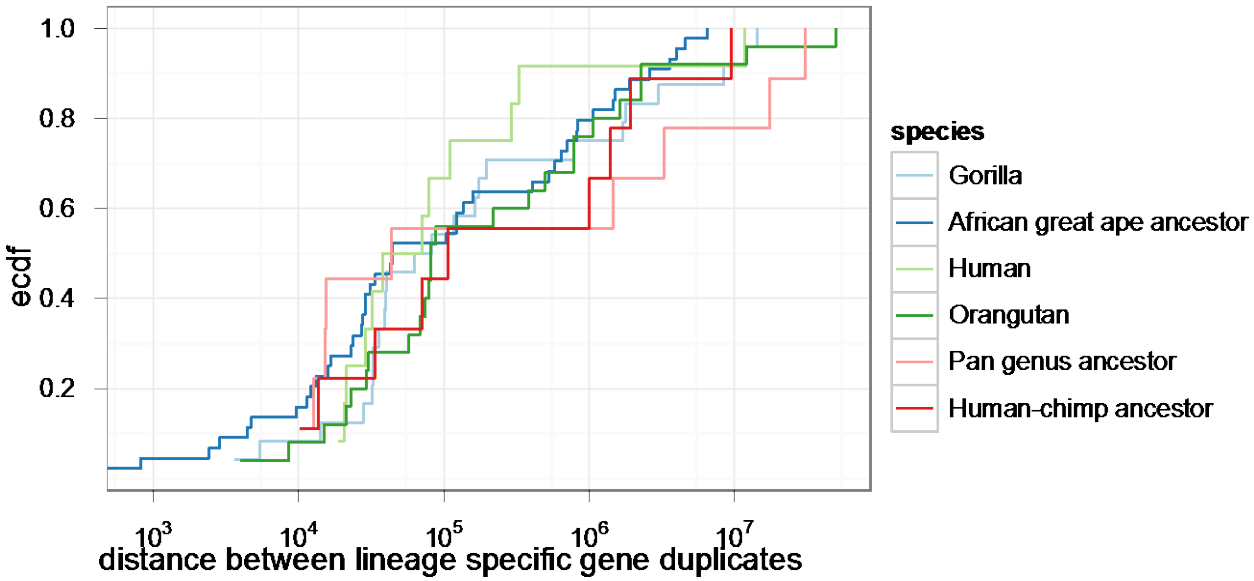


Figure 8.1: Cumulative histogram of the distance between lineage-specific gene duplicates. More than 50% of all gene duplicates lie within 100 kbp of each other.

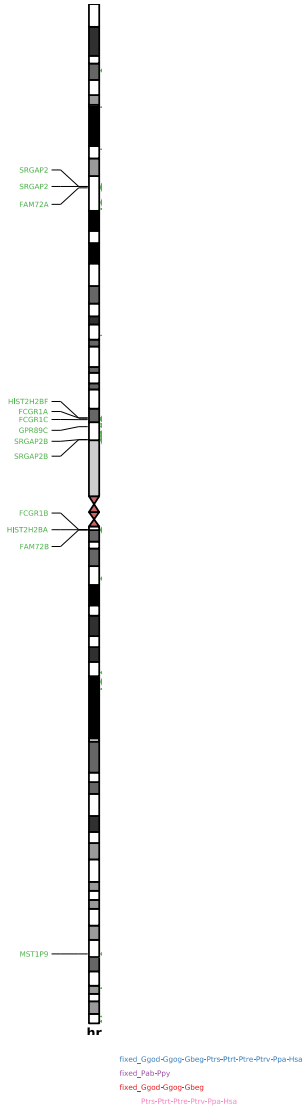

SRGAP2
SRGAP2
FAM72A
HIST2H2BF
FCGR1A
FCGR1C
GPR89C
SRGAP2B
SRGAP2B
FCGR1B
HIST2H2BA
FAM72B
MST1P9
hr
fixed_Ggod-Ggog-Gbeg-Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa
fixed_Pab-Ppy
fixed_Ggod-Ggog-Gbeg
Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa

Figure 8.2a: Human-specific gene duplications plotted on a chromosome ideogram show gene duplications clustered together. Lineage-specific duplicated segments are plotted to the right of each chromosome ideogram as arrows with their bases scaled by 10 for visibility.

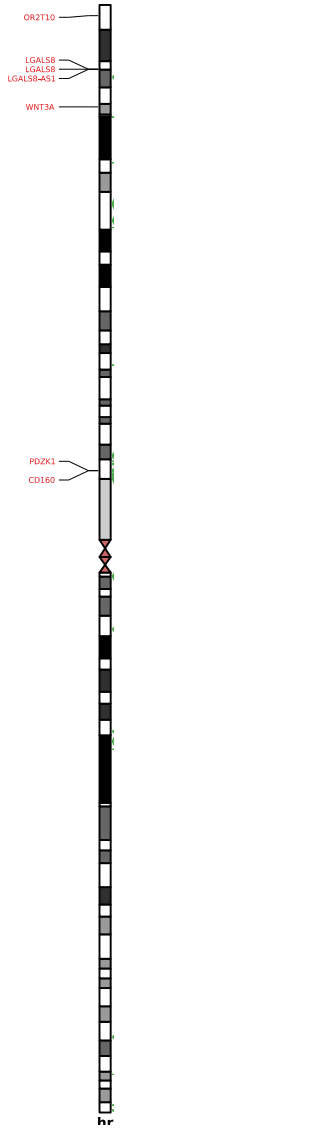

OR2T10
LGALS8
LGALS8
LGALS8-AS1
WNT3A
PDZK1
CD160
hr
Ppa-Hsa

Figure 8.2b: Gorilla-specific gene duplications plotted on a chromosome ideogram are also clustered together.

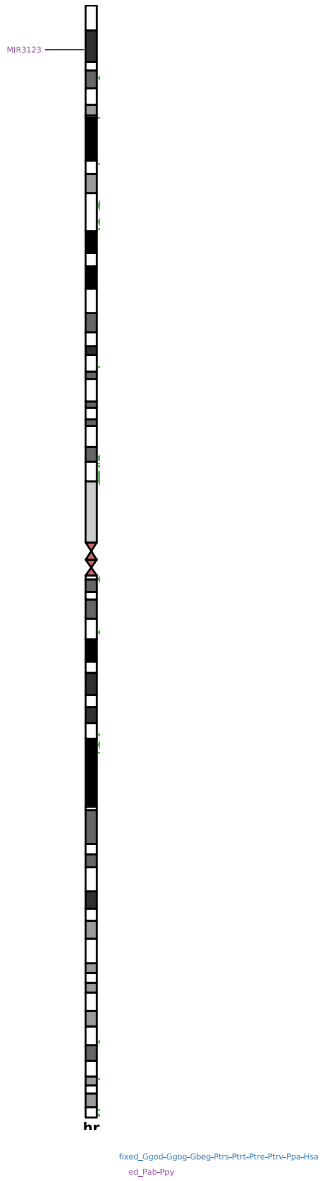


Figure 8.2c: Orangutan-specific gene duplications plotted on a chromosome ideogram.

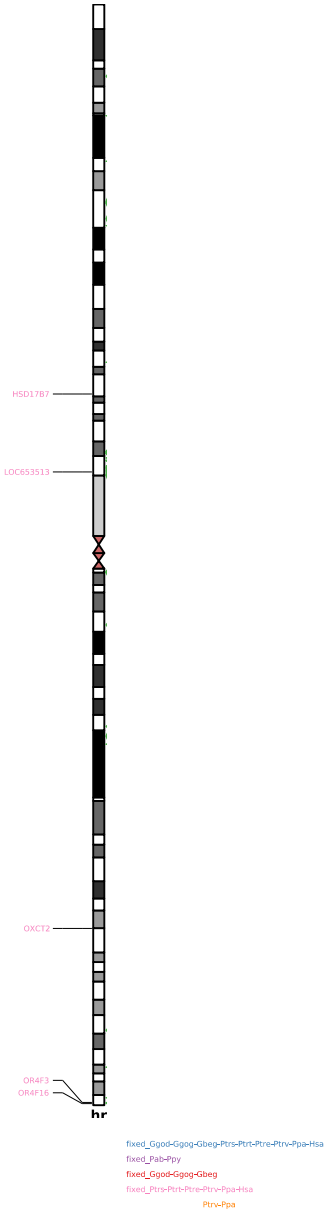
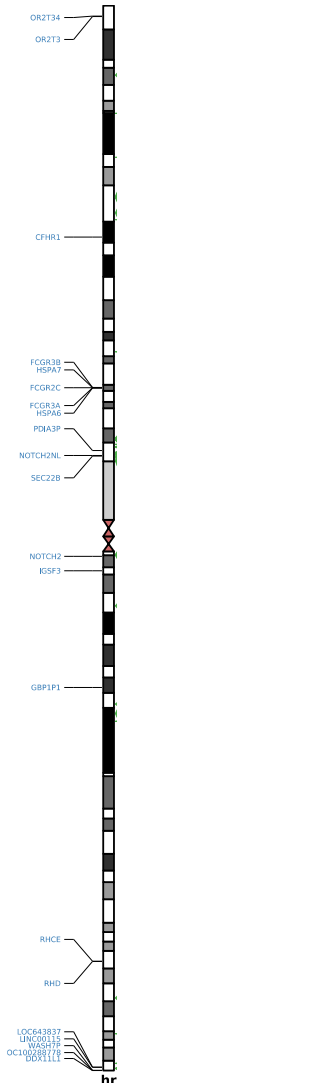


Figure 8.2d: Human, chimpanzee, and bonobo shared gene duplications plotted on a chromosome ideogram.



Gbeg-Ptrs-Ptrt-Ptre-Ptrv-Ppa-Hsa

Figure 8.2e: African great ape-specific gene duplications plotted on a chromosome ideogram.

Distribution of SDs and deletions: We have previously observed that lineage-specific duplications tend to arise adjacent to ancestral duplications, a phenomenon we have termed *duplication shadowing*(Cheng et al. 2005; Marques-Bonet et al. 2009). We tested that association again here (using a permutation test method)(Marques-Bonet et al. 2009) and found that indeed lineage-specific duplications were highly enriched near to ancestral duplications (P<0.0002, **Table 8.1**). We were curious whether this phenomenon held true for deletions as well. Lineage-specific deletions did not tend to arise adjacent to ancestral deletions (P>0.2, **Table 8.1**), and additionally, no association with SDs was observed either (P>0.2, **Table 8.1**).

Table 8.1: Association of lineage-specific duplication and deletion events with each other. As we previously reported, we find that lineage-specific SDs tend to emerge adjacent to ancestral SDs. We also tested if deletions emerged adjacent to ancestral deletions or SDs and found no association. For loci lost along the human lineage, only deletions flanked by regions that could unambiguously be placed in HG18 were considered.

Comparison	test events	reference events	test within 5 kbp of reference	P-value
human duplications with human/chimp duplications	279	268	30	P<0.0002
chimp duplications with human/chimp duplications	115	268	12	P<0.0002
Bonobo duplications with human/chimp duplications	51	268	2	P=0.0214
human/chimp duplications with human/chimp/gorilla duplications	268	1172	125	P<0.0002
gorilla duplications with human/chimp/gorilla duplications	238	1172	28	P<0.0002
orangutan duplications with human/chimp/gorilla/orangutan ancestral duplications	549	7077	116	P<0.0002
gorilla deletions with gorilla duplications	469	238	1	P=0.8200
gorilla deletions with human/chimp/gorilla duplications	469	1172	5	P=0.8900
gorilla deletions with human/chimp duplications	459	268	2	P=0.5500
human deletions with human/chimp/ duplications	331	268	1	P=0.7600
human/chimp deletions with human/chimp/gorilla duplications	65	1172	0	P=1.0000
human/chimp/gorilla deletions with human/chimp/gorilla/orangutan duplications	372	7077	47	P=1.0000
human deletions with human/chimp ancestral deletions	331	74	7	P=0.2300
gorilla deletions with human/chimp/gorilla ancestral deletions	469	422	39	P=1.0000
human/chimp deletions with human/chimp/gorilla ancestral deletions	65	422	10	P=0.6500

Section 9: Rates of segmental duplication and deletion and underlying genes

As we have identified SNPs and substitutions in the same individuals assessed for structural variants, we can compute precise estimates of the rates of duplication and deletion as a function of the number of substitutions along a branch. Additionally, the divergence times of

each of the species assessed can be computed precisely from genetic divergence estimates. We have previously described a likelihood framework for testing the significance of changes in duplication rate observed throughout a phylogeny(Marques-Bonet et al. 2009). In this likelihood framework, the null model assumes a constant rate among all branches while in the test model the rates vary among different branches. All tests were performed in units of duplications per million years, with divergence times estimated from exact substitution rates, and in units of duplications per substitution.

Rates of SD: We have previously shown an acceleration in the rate of duplication in the African great ape ancestor(Marques-Bonet et al. 2009) and additionally an acceleration of the duplication rate in the gorilla lineage(Ventura et al. 2011). We confirmed these two findings by assessing the rate of ancient duplication among all branches in the tree. The four subspecies of chimpanzee were considered as a single population. Three models were tested against a null model of a constant duplication rate of duplication along all branches:

- **An accelerated rate in the African great ape ancestor,**
- **An accelerated rate in the African great ape ancestor, the gorilla ancestor, and the human-chimpanzee ancestor, and**
- **Differing rates of duplication among all branches of the tree.**

Each of these three models is highly significant compared the null model (**Table 9.1**); however, the tree is best explained by a model in which there is an accelerated rate of duplication in the African great ape and gorilla ancestors in addition to the ancestor of human and chimpanzee. This supports a model in which a burst of duplication activity occurred in the African great ape ancestor which subsequently rapidly declined in the ancestor of human and chimpanzee and declined more slowly in the gorilla lineage.

Table 9.1: Duplication rates along the great ape lineage.

description	units	Model 1	Model 2	degrees of freedom	p-value
acceleration in African great ape against all other branches	Mb/My	λ=3.030	λ1 = 2.127 λ2 = 7.501	1	6.73E-13
	bp/sub	λ=1.197	λ1 = 0.841 λ2 = 2.961	1	7.00E-13
	sites/MY	λ=30.006	λ1 = 17.659 λ2 = 91.155	1	<1.138e-202
acceleration in African great ape and gorilla and human-	Mb/My	λ=3.030	λ1 = 1.496	1	1.21E-17

chimp linages against all branches			λ2 = 6.173		
	bp/sub	λ=1.197	λ1 = 0.592	1	1.53E-17
			λ2 = 2.432		
	sites/MY	λ=30.006	λ1 = 14.943	1	<7.529e-157
			λ2 = 60.866		
different rates among all branches	Mb/My	λ=3.030	λ1 = 0.266	12	3.00E-14
			λ2 = 7.501		
			λ3 = 4.984		
			λ4 = 3.809		
			λ5 = 1.649		
			λ6 = 1.119		
			λ7 = 0.547		
			λ8 = 1.210		
			λ9 = 0.470		
			λ10 = 0.528		
			λ11 = 2.147		
			λ12 = 0.341		
			λ13 = 2.104		
	bp/sub	λ=1.197	λ1 = 0.108	12	3.48E-14
			λ2 = 2.961		
			λ3 = 1.965		
			λ4 = 1.476		
			λ5 = 0.647		
			λ6 = 0.434		
			λ7 = 0.216		
			λ8 = 0.484		
			λ9 = 0.178		
			λ10 = 0.215		
			λ11 = 0.849		
			λ12 = 0.132		
			λ13 = 0.863		

Gene duplication rates: We next tested if the increase in duplication rate in African great ape ancestor, gorilla and human-chimpanzee ancestral lineages had led to a significant increase in the rate of gene duplication (**Table 9.2**). Indeed, the rate of gene duplication in these lineages was ~2.8-fold higher than throughout the rest of the tree (P=1.66e-20). We also observed a ~1.5-fold excess in the rate of gene duplication in the human-chimpanzee ancestral branch compared to the African great ape ancestral branch and the gorilla branch, which was weakly significant (P=0.0107).

Table 9.2: Rates of lineage-specific gene duplication.

description	units	Model 1	Model 2	degrees of freedom	p-value

acceleration in the rate of gene duplication in the African great ape, gorilla and human-chimpanzee lineages against all branches	genes/MY	λ=6.620	λ1 = 4.167 λ2 = 11.646	1	1.66E-20
	genes/sub	λ=2.616	λ1 = 1.649 λ2 = 4.588	1	2.42E-20
acceleration in the rate of gene duplication in human-chimp lineage compared to the gorilla and African great ape lineages	genes/MY	λ=11.646	λ1 = 10.933 λ2 = 19.286	1	0.0107
	genes/sub	λ=4.588	λ1 = 4.313 λ2 = 7.474	1	0.01323

Rates of lineage-specific deletion and gene loss: We observed a ~2-fold increase in the rate of deletion along the chimpanzee-bonobo lineage. We found this increase to be highly statistically significant (**Table 9.3,** $P=4.79X10^{-9}$). In addition, the associated rate of gene loss along this branch is also highly statistically significant ($P=4.4X10^{-8}$). Amongst the other branches, the rate of deletion and the associated rate of gene loss appear to be more clocklike. We cannot reject the null model (of clocklike) in either of these cases (P=0.03 and P=0.6). Analyses were also performed just for deletions >5kbp, (**Table 9.4**), showing a strong trend

Table 9.3: Rates of lineage-specific deletion and gene loss.

description	units	Model 1	Model 2	degrees of freedom	p-value
acceleration in the rate of deletion along the chimpanzee-bonobo ancestral branch compared to all other branches	ckbp/MY	λ=9.072	λ1 = 8.189	1	4.79E-09
			λ2 = 18.054		
	bp/kb-sub	λ=3.587	λ1 = 3.240	1	6.63E-09
			λ2 = 7.086		
	sites/MY	λ=114.842	λ1 = 110.829	1	7.13E-13
			λ2 = 151.136		
	genes/MY	λ=6.034	λ1 = 5.318	1	4.40E-08
			λ2 = 12.955		
	genes/subs	λ=2.385	λ1 = 2.103	1	5.57E-08
			λ2 = 5.085		
differing rate of deletion in all non-chimpanzee-bonobo ancestor branches (reject clocklike)	kbp/MY	λ=8.918	λ1 = 7.626	5	0.03123
			λ2 = 8.919		
			λ3 = 8.801		
			λ4 = 10.392		
			λ5 = 7.172		
			λ6 = 11.668		
	bp/kb-sub	λ=3.526	λ1 = 3.012	5	0.03557
			λ2 = 3.660		
			λ3 = 3.411		
			λ4 = 4.102		
			λ5 = 2.868		
			λ6 = 4.600		
	genes/MY	λ=5.660	λ1 = 6.207	5	0.6271
			λ2 = 10.000		
			λ3 = 4.286		
			λ4 = 4.762		
			λ5 = 6.167		
			λ6 = 5.303		
	sites/MY	λ=113.767	λ1 = 76.621	4	1.82E-72
			λ2 = 152.143		
			λ3 = 141.071		
			λ4 = 153.167		
			λ5 = 116.667		
excluding orangutan	sites/MY	λ=137.812	λ1 = 116.667	3	1.12E-07
			λ2 = 152.143		
			λ3 = 141.071		
			λ4 = 153.167		

Table 9.4: Rates of lineage-specific deletion for sites >5kb.

description	units	Model 1	Model 2	degrees of freedom	p-value
acceleration in the rate of deletion along the chimpanzee-bonobo ancestral branch compared to all other branches	ckbp/MY	λ=7.778	λ1 = 6.836	1	2.08E-09
			λ2 = 16.304		
	bp/10-sub	λ=3.072	λ1 = 2.702	1	2.72E-09
			λ2 = 6.399		
	sites/MY	λ=50.837	λ1 = 45.704	1	1.85E-38
			λ2 = 97.273		
	genes/MY	λ=6.282	λ1 = 5.527	1	1.60E-07
			λ2 = 12.955		
	genes/subs	λ=2.482	λ1 = 2.186	1	2.00E-07
			λ2 = 5.085		
differing rate of deletion in all non-chimpanzee-bonobo ancestor branches (reject clock like)	kbp/MY	λ=6.812	λ1 = 6.365	4	0.0003297
			λ2 = 5.761		
			λ3 = 7.478		
			λ4 = 3.655		
			λ5 = 10.100		
	bp/10-sub	λ=2.693	λ1 = 2.514	4	0.000414
			λ2 = 2.233		
			λ3 = 2.952		

			λ4 = 1.462		
			λ5 = 3.981		
	genes/MY	λ=5.637	λ1 = 6.207	4	0.5693
			λ2 = 4.286		
			λ3 = 4.762		
			λ4 = 6.167		
			λ5 = 5.303		
	sites/MY	λ=46.152	λ1 = 38.759	4	3.35E-20
			λ2 = 47.143		
			λ3 = 46.548		
			λ4 = 37.333		
			λ5 = 69.697		
excluding orangutan	sites/MY	λ=50.938	λ1 = 69.697	3	8.61E-15
			λ2 = 47.143		
			λ3 = 46.548		
			λ4 = 37.333		

Section 10: Copy number variation and duplication diversity

Copy number variants and copy number polymorphic loci: Copy number polymorphic loci and copy number variants were identified using a digital array comparative genomic hybridization (dCGH) strategy(Sudmant et al. 2010), in combination with the above scale-space filtering-based segmentation approach. Briefly, for a particular species, one individual is assigned to be a *reference* sample. A comparative copy number signal between *test* individuals from each species and their respective references is then generated from the log2 ratio of the 500 bp tiled window copy number estimates between the two samples. These log2 ratios are then segmented into blocks of respective 'gain' and 'loss' using the above segmentation procedure. CNPs identified in high GC loci were then filtered out if the GC content of the CNP was >0.55 or if the mean copy number of individuals with known GC biases differed by >0.5 from unbiased individuals. 18 individuals were excluded from this analysis as GC bias prevented them from being compared to nonbiased samples. CNPs encompassing fewer than four 500 bp windows of unmasked sequence were discarded to enrich for true positives effectively imposing a 2 kbp size constraint on discovered events. To maximize sensitivity, for each species every individual in turn was selected as the *reference* genome and compared against all other genomes. A total of 2062 individual dCGH comparisons were performed and a total of 6406 events discovered (**Table 10.1**). Though the minimum size of CNPs discovered was 2 kbp,

as a minimum of four 500 bp windows of unmasked sequence was required to discover a CNP, the median size was much larger (8061 bp; **Figure 10.1**).

Table 10.1: CNPs discovered among 80 great apes.

	private		segregating			total		θ	
species	**deletion**	**duplication**	**duplication**	**deletion**	**duplication /deletion**	**private**	**segregating**	**CNVs (per genome)**	**bp (per base)**
Sumatran_orangutan	430	71	226	114	65	501	405	122.05	0.0029
Bornean_orangutan	296	61	149	108	53	357	310	97.48	0.002
Human	203	182	113	38	29	385	180	49.38	0.0013
Bonobo	340	68	147	89	29	408	265	77.04	0.0012
Eastern_chimpanzee	466	135	107	78	21	601	206	68.21	0.002
Western_chimpanzee	242	61	49	37	20	303	106	37.47	0.001
Nigerian-Cameroon_chimpanzee	198	43	132	35	30	241	197	55.53	0.0018
Central_chimpanzee	581	232	-	-	-	813	-	-	0.0024
Western_gorilla	402	130	287	92	39	532	418	98.27	0.002
Eastern_gorilla	114	60	-	-	-	174	0	-	0.0009

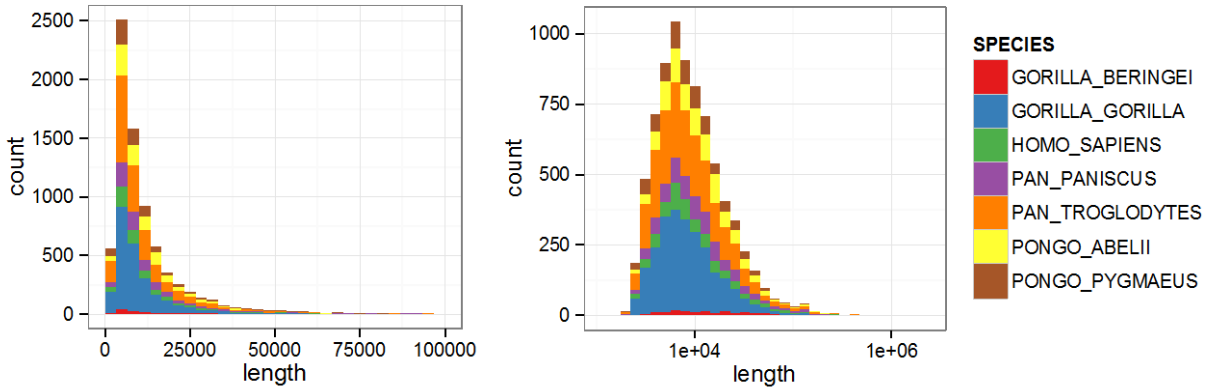


Figure 10.1: Distribution of CNP sizes shown in standard and log-scale. The median CNP size was 8061 bp.

We next assessed the number, and size distribution of discovered CNVs on a per-individual basis, to determine if any individuals in particular showed aberrantly more or larger CNVs (**Table 10.2**). The results are summarized in **Figure 10.2** with all individuals of a particular species demonstrating a similar number of CNVs on average, with four potential outliers showing excess numbers of deletions. The first outlier, Nyango, is the only Cross River Gorilla sample assessed in our study. We note that this individual shows both signals of recent inbreeding and exceedingly low diversity, (Prado and Sudmant *et al* – *in* press) in addition to a unique demographic history and reduced N_e. Though interesting, it is hard to draw conclusions from this single individuals excess of deletions. The other three individuals are a central chimpanzee, of which only two individuals were assessed, a Bornean

orangutan and a Sumatran orangutan. The two orangutans are not outliers among the orangutan samples with respect to sequence coverage, nor is the central chimpanzee. We thus find no apparent reason to purge these individuals from our analyses, however, we note that all reported results on population diversity are robust to the removal of these individuals, and indeed, none of the significant signals reported is related to any of these individuals. Finally, we assessed the distribution of CNV lengths among the 80 great apes analyzed for CNV diversity in this study (**Figure 10.3**). We noted that for almost all individuals the length distributions of deletions and duplications were very similar as one would expect for a random sampling of CNVs. Additionally, size distributions were consistent among individuals of the same species and largely between species as well. Interestingly, the human individuals consistently exhibited a slight excess of longer deletions, perhaps as a result of the human reference genome being biased towards a particular set of CNV alleles.

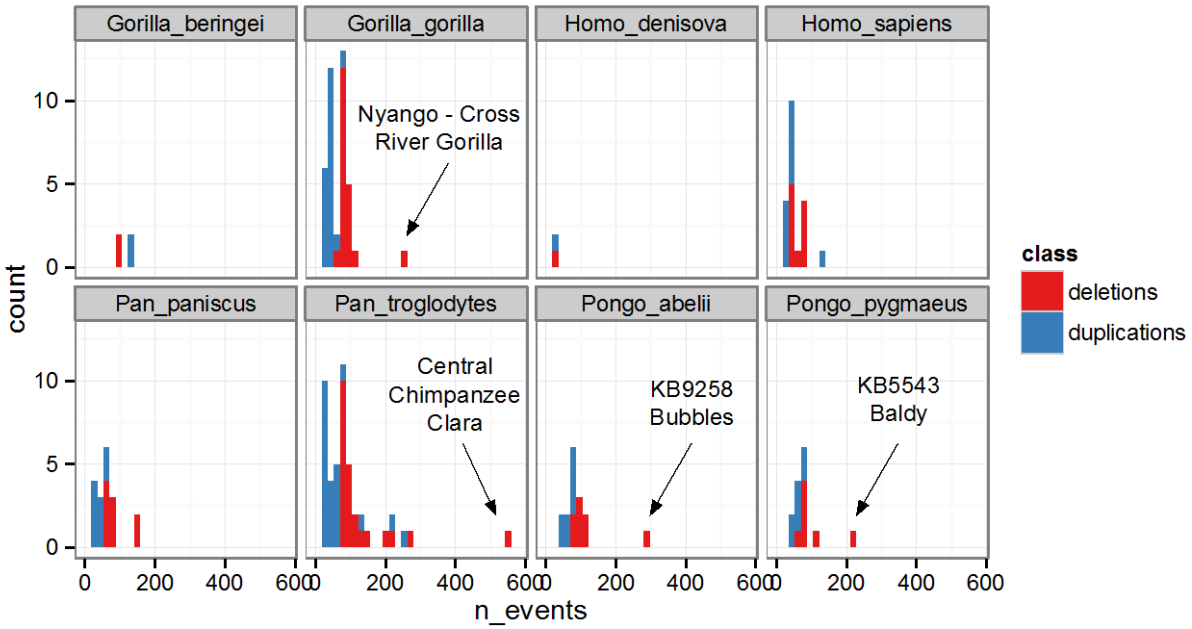


Figure 10.2: Summary of the number of duplications and deletions discovered per individual and grouped by species.

Table 10.2: The number of CNV duplications and deletions discovered per individual.

indiv	species	duplications	deletions
Pongo_abelii-KB9258_Bubbles	Pongo_abelii	77	293
Pongo_abelii-A953_Vicki	Pongo_abelii	87	90

Pongo_abelii-A949_Dunja	Pongo_abelii	70	70
Pongo_abelii-KB4661_Dolly	Pongo_abelii	66	107
Pongo_abelii-KB5883_Likoe	Pongo_abelii	70	97
Pongo_abelii-A947_Elsi	Pongo_abelii	67	70
Pongo_abelii-SB550_Sibu	Pongo_abelii	50	109
Pongo_abelii-KB4361_Dennis	Pongo_abelii	45	97
Pongo_pygmaeus-KB5543_Baldy	Pongo_pygmaeus	83	212
Pongo_pygmaeus-KB4204_Dinah	Pongo_pygmaeus	44	73
Pongo_pygmaeus-KB5406_Doris	Pongo_pygmaeus	55	107
Pongo_pygmaeus-A939_Nonja	Pongo_pygmaeus	81	80
Pongo_pygmaeus-KB5404_Billy	Pongo_pygmaeus	48	69
Pongo_pygmaeus-KB5405_Louis	Pongo_pygmaeus	64	87
Pongo_pygmaeus-A944_Napoleon	Pongo_pygmaeus	61	84
Homo_denisova-Denisova_30x	Homo_denisova	32	28
Homo_sapiens-Han_HGDP00778	Homo_sapiens	36	45
Homo_sapiens-Dai_HGDP01307	Homo_sapiens	18	50
Homo_sapiens-Yoruba_HGDP00927	Homo_sapiens	35	73
Homo_sapiens-Mbuti_HGDP00456	Homo_sapiens	35	76
Homo_sapiens-San_HGDP01029	Homo_sapiens	36	74
Homo_sapiens-Karitiana_HGDP00998	Homo_sapiens	129	42
Homo_sapiens-Sardinian_HGDP00665	Homo_sapiens	37	48
Homo_sapiens-French_HGDP00521	Homo_sapiens	23	55
Homo_sapiens-Papuan_HGDP00542	Homo_sapiens	20	47
Homo_sapiens-Madenka_HGDP01284	Homo_sapiens	29	86
Pan_paniscus-A925_Bono	Pan_paniscus	43	78
Pan_paniscus-A923_Kombote	Pan_paniscus	31	73
Pan_paniscus-9731_LB502	Pan_paniscus	59	140
Pan_paniscus-A918_Hermien	Pan_paniscus	38	66
Pan_paniscus-A919_Desmond	Pan_paniscus	30	68
Pan_paniscus-A927_Salonga	Pan_paniscus	40	83
Pan_paniscus-X00095_Kakowet	Pan_paniscus	62	153
Pan_paniscus-A926_Natalie	Pan_paniscus	28	58
Pan_paniscus-A915_Kosana	Pan_paniscus	26	67
Pan_troglodytes_ellioti-Tobi	Pan_troglodytes	21	96
Pan_troglodytes_schweinfurthii-A911_Kidongo	Pan_troglodytes	53	147
Pan_troglodytes_verus-9668_Bosco	Pan_troglodytes	46	98
Pan_troglodytes_verus-9730_Donald	Pan_troglodytes	53	103
Pan_troglodytes_schweinfurthii-Yolanda	Pan_troglodytes	48	223
Pan_troglodytes_ellioti-Julie	Pan_troglodytes	26	85
Pan_troglodytes_ellioti-Taweh	Pan_troglodytes	23	81
Pan_troglodytes_ellioti-Basho	Pan_troglodytes	28	84
Pan_troglodytes_schweinfurthii-A912_Nakuu	Pan_troglodytes	53	105
Pan_troglodytes_ellioti-Banyo	Pan_troglodytes	30	76
Pan_troglodytes_schweinfurthii-9729_Harriet	Pan_troglodytes	77	116
Pan_troglodytes_ellioti-Paquita	Pan_troglodytes	23	84
Pan_troglodytes_schweinfurthii-100040_Andromeda	Pan_troglodytes	127	262
Pan_troglodytes_schweinfurthii-100037_Vincent	Pan_troglodytes	55	98
Pan_troglodytes_troglodytes-A960_Clara	Pan_troglodytes	214	542
Pan_troglodytes_ellioti-Koto	Pan_troglodytes	25	77
Pan_troglodytes_ellioti-Kopongo	Pan_troglodytes	21	90

Pan_troglodytes_verus-A907_Susie_A	Pan_troglodytes	53	84
Pan_troglodytes_verus-X00100_Koby	Pan_troglodytes	50	198
Pan_troglodytes_ellioti-Akwaya_Jean	Pan_troglodytes	29	79
Pan_troglodytes_ellioti-Damian	Pan_troglodytes	26	86
Pan_troglodytes_troglodytes-A958_Doris	Pan_troglodytes	247	126
Pan_troglodytes_verus-Clint	Pan_troglodytes	43	86
Gorilla_gorilla_gorilla-9753_Kokomo	Gorilla_gorilla	26	92
Gorilla_gorilla_gorilla-A936_Coco	Gorilla_gorilla	35	68
Gorilla_gorilla_gorilla-A935_Fritz	Gorilla_gorilla	38	84
Gorilla_gorilla_gorilla-KB5852_Helen	Gorilla_gorilla	39	87
Gorilla_gorilla_gorilla-9751_Bulera	Gorilla_gorilla	33	93
Gorilla_gorilla_gorilla-A932_Mimi	Gorilla_gorilla	38	91
Gorilla_gorilla_gorilla-Snowflake	Gorilla_gorilla	57	86
Gorilla_gorilla_gorilla-KB7973_Porta	Gorilla_gorilla	33	86
Gorilla_gorilla_gorilla-KB5792_Carolyn	Gorilla_gorilla	39	87
Gorilla_gorilla_gorilla-KB4986_Katie	Gorilla_gorilla	36	75
Gorilla_beringei_graueri-A929_Kaisi	Gorilla_beringei	130	101
Gorilla_gorilla_gorilla-9752_Suzie	Gorilla_gorilla	35	94
Gorilla_gorilla_gorilla-9749_Kowali	Gorilla_gorilla	20	87
Gorilla_gorilla_gorilla-X00109_Tzambo	Gorilla_gorilla	41	108
Gorilla_beringei_graueri-9732_Mkubwa	Gorilla_beringei	127	90
Gorilla_gorilla_dielhi-B646_Nyango	Gorilla_gorilla	77	252
Gorilla_gorilla_gorilla-KB3782_Vila	Gorilla_gorilla	35	79
Gorilla_gorilla_gorilla-9750_Azizi	Gorilla_gorilla	37	101
Gorilla_gorilla_gorilla-A934_Delphi	Gorilla_gorilla	27	80
Gorilla_gorilla_gorilla-KB6039_Oko	Gorilla_gorilla	40	84
Gorilla_gorilla_gorilla-KB3784_Dolly	Gorilla_gorilla	25	87
Gorilla_gorilla_gorilla-A933_Dian	Gorilla_gorilla	36	81

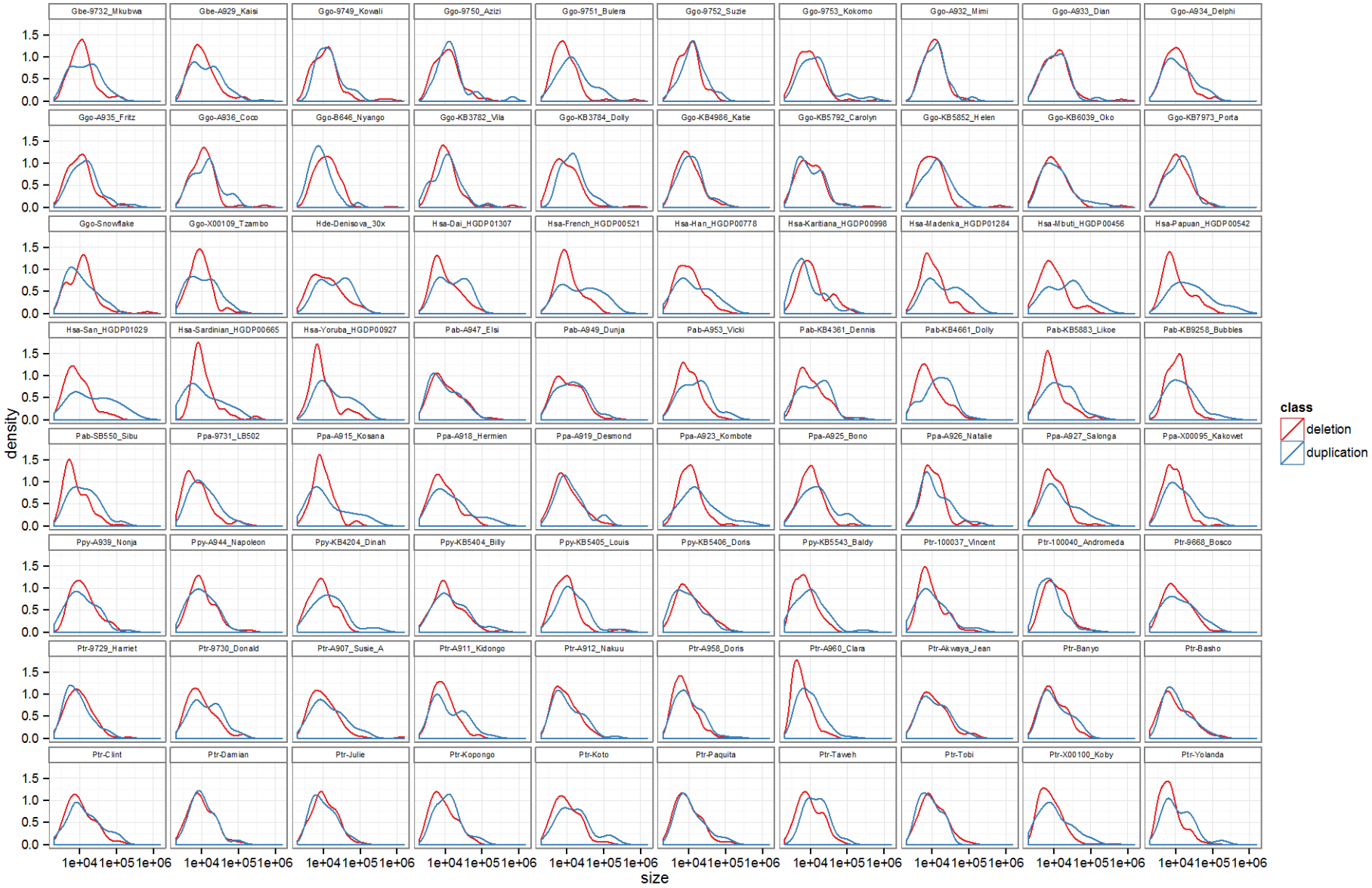


Figure 10.3 – Size distributions plotted in log-space of CNVs identified among all the great apes assessed for CNV diversity in this study.

Diversity of CNPs in great apes: In order to explore the diversity of great ape copy number variants, we first compared the copy number diversity of each of the populations analyzed to the SNP diversity (Prado and Sudmant, companion under review). Using segregating CNVs, we computed an estimate of Watterson's θ for segregating CNVs/genome and compared this to Watterson's theta for SNPs/bp. As expected, the two were highly correlated (r^2=0.52, p=0.026; **Figure 10.4**).

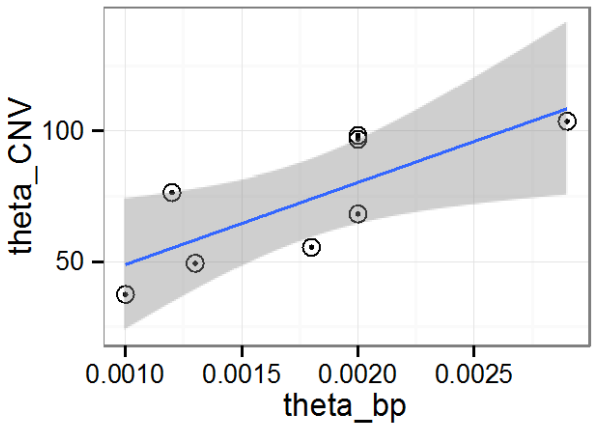


Figure 10.4: Estimates Watterson's theta (θ) for segregating CNVs/genome plotted versus segregating SNPs/bp shows a strong correlation between the diversity of SNP and CNP diversity (r^2 =0.52, p=0.026).

We next performed a Principle Components Analysis (PCA(Patterson et al. 2006)) using only deletion CNVs that could be assigned a state of 0, 1, or 2 (**Figure 10.5**). All known species' and subspecies' relationships and classifications are observed as clusters in the PCAs. Additionally, the relative levels of diversity are captured in the clusterings with more diverse populations, such as the Western gorilla, showing a more dispersed pattern. Sumatran orangutan show more diversity than Bornean orangutan, as suggested by the estimates of θ in both SNPs and CNPs. The PCAs additionally show two clusters in Sumatran orangutans, possibly demonstrating Sumatran orangutan subpopulations. Among human, the first PC discriminates African from non-African populations with the Denisova individual falling intermediate. The Denisova is closest to the Papuan and Asian individuals along PC2, which is interesting as gene flow from Denisovans to Papuan individuals has been reported(Meyer et al. 2012).

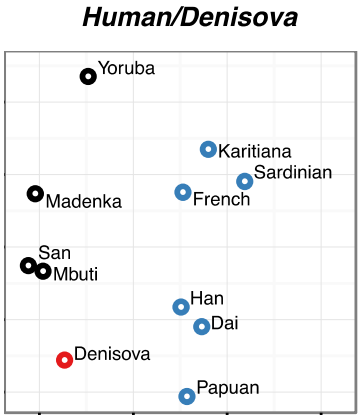


Figure 10.5: PCA plots based on CNV genotypes for chimpanzee, gorilla, and orangutan, respectively. All individual species' and subspecies' relationships are confirmed in addition to relative diversity between populations. Western lowland gorillas, for example, show far more diversity than Eastern lowland gorillas. Between orangutan species, the Sumatran orangutans show more diversity and appear to cluster into two groups, potentially a signature of Sumatran orangutan subpopulations.

We next constructed neighbor-joining trees using the genotypes of fixed and segregating variant calls to compute the hamming distance between all genomes. Species' trees were first considered using all variants >2 kbp, 3.5 kbp, and 5 kbp, respectively (**Figure 10.6**). As a function of millions of years since divergence, all trees show a significant increase in the number fixed deletions in the chimpanzee-bonobo ancestor (**Section 9**); however, we find there is a marked enrichment in the chimpanzee-bonobo branch for larger events.

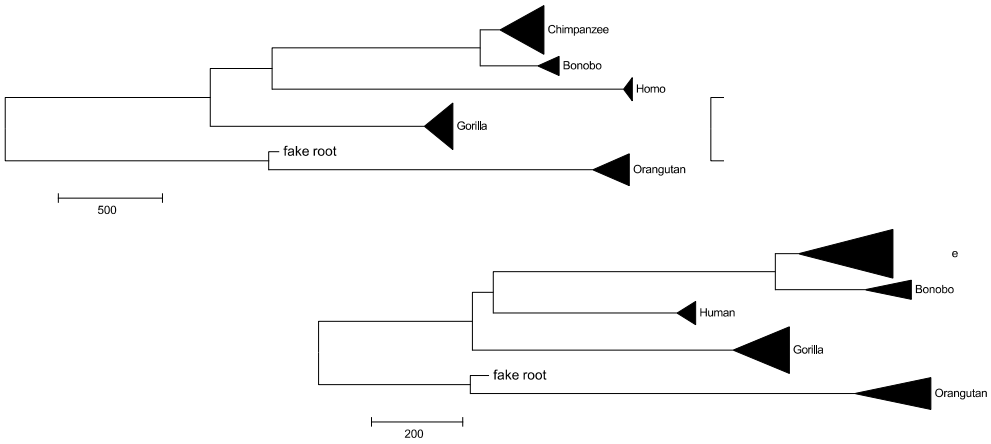


Figure 10.6: Neighbor-joining species trees constructed from segregating variants and fixed variants ≥2 kbp, ≥3.5 kbp and ≥5 kbp (a,b,c), respectively. Though the chimpanzee-bonobo branch shows an increased rate of deletion for each of these size thresholds, the effect is more pronounced for larger deletions.

We next constructed a neighbor-joining tree for all great ape individuals (**Figure 10.7**). Confidence values on branches are generated from repeatedly randomly subsampling 50% of the variants and re-computing the tree. All known species' relationships were captured with 100% confidence along with the correct species-tree topology. All the individual subspecies' topologies were additionally reconstructed with high confidence and were identical to those identified by SNPs with the exception of Eastern chimpanzee, which are placed as an outgroup to all chimpanzee subspecies. Eastern lowland gorillas place as an outgroup to all Western gorillas and the single Cross River gorilla individual as an outgroup to all Western lowland gorillas. Within the chimpanzee phylogeny, the two Nigerian-Cameroon individuals Tobi and Julie cluster together, as observed in the SNP-based tree, and the individuals Yolanda, Andromeda and Vincent, all Eastern chimpanzees from Gombe reserve in Tanzania, cluster together with 100% support. Within the human phylogeny, Africans and non-Africans form two high-confidence clades and the Denisova archaic hominid places as an outgroup to all humans robustly with 92% confidence.

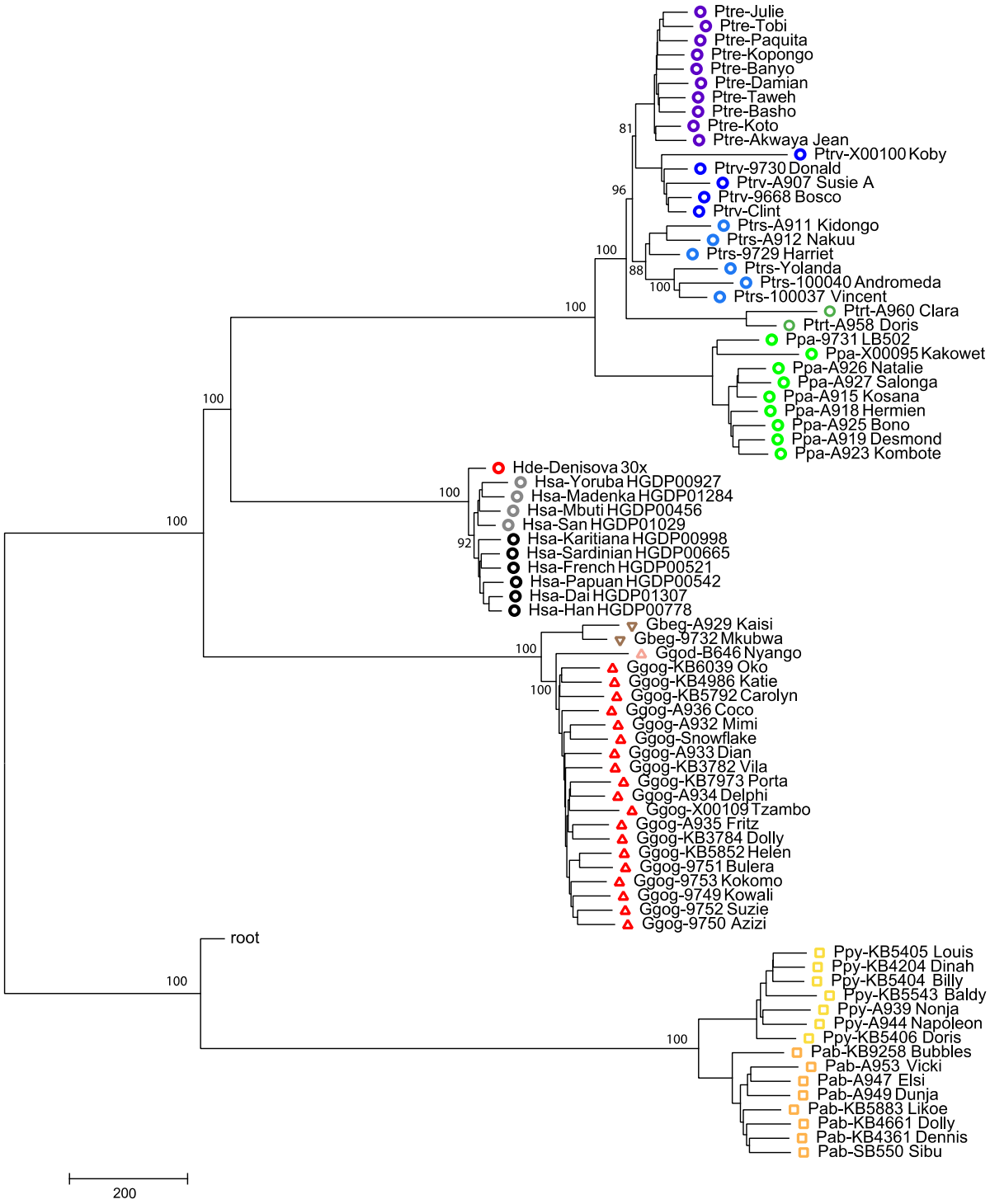


Figure 10.7: A neighbor-joining tree of great ape individuals constructed from segregating structural variation and fixed deletions >5 kbp. Confidence values on branches are generated from repeatedly randomly subsampling 50% of the variants and re-computing the tree.

CNV burden among great ape populations: Though our per-population sample sizes are small, we attempted to assess the CNV burden among different primate populations by comparing the total load of deletion and duplication events between different populations. Analysis was limited

to non-segmentally duplicated regions of the genome and events >30 kbp to limit false positives. We first assessed the relative duplication load of different populations (**Figure 10.8**).

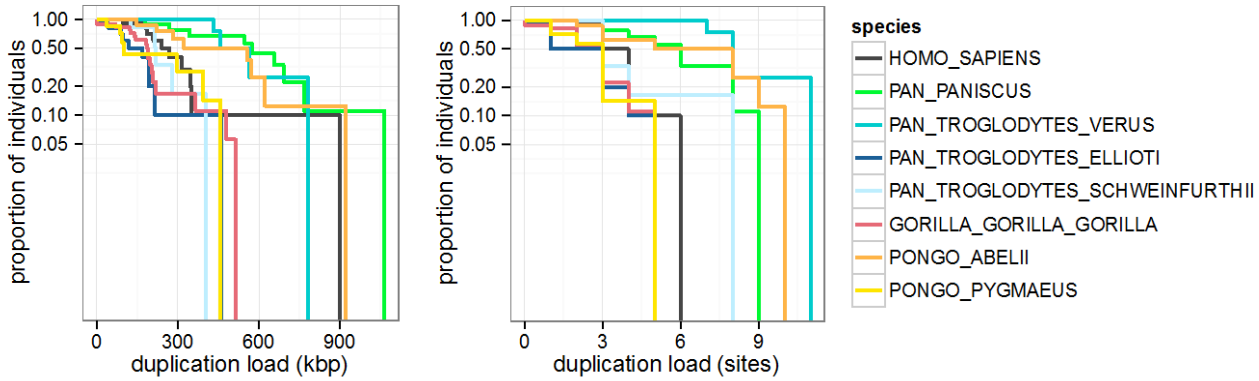


Figure 10.8: Survival functions of the proportion of individuals with at least a given total base-pair load of duplications plotted by total kbp and by number of sites. Bonobo, Western chimpanzee, and Sumatran orangutan have an increased load of duplications. The human Papuan individual harbors two large private duplications increasing the total kbp load in this individual.

Bonobo, Western chimpanzee, and Sumatran orangutan all showed a significantly higher burden of duplications compared to human, Eastern chimpanzee, Nigerian-Cameroon chimpanzee, gorilla, and Bornean orangutan (**Table 10.3**). A single human Papuan individual showed an increased duplication load as measured in base pairs as a result of two large private duplications.

Table 10.3: Statistical significance of the relative base-pair load and the relative site load of duplications between different primate populations.

Species 1	Species 2	p-value – bp load	p-value # of sites load
Bonobos	*Humans, Eastern chimpanzees, Nigerian-Cameroon chimpanzees, gorillas and Bornean orangutans*	0.0014	0.005
Sumatran orangutans	*Humans, Eastern chimpanzees, Nigerian-Cameroon chimpanzees, gorillas and Bornean orangutans*	0.0088	0.014
Western chimpanzees	*Humans, Eastern chimpanzees, Nigerian-Cameroon chimpanzees, gorillas and Bornean orangutans*	0.02	0.004
Bonobos, Sumatran orangutans and Western chimpanzees	*Humans, Eastern chimpanzees, Nigerian-Cameroon chimpanzees, gorillas and Bornean orangutans*	8.66e-6	3.09e-5

We next assessed the relative load of deletions between different populations identifying a significantly increased deletion burden in Western chimpanzees compared to all other primate populations (**Figure 10.9**; p=0.00289, for base-pair load, p=0.04 for site load). One particular Western chimpanzee, Susie-A, harbored a single large ~1.6 Mbp deletion event (**Section 11**). To ensure that this single individual was not driving the significance of the load comparison, we removed her from the analysis and tested the relative load again. Though this left only three Western chimpanzees in the analysis, the result was still significant by base-pair load (p=0.02) though not by number of sites (p=0.055).

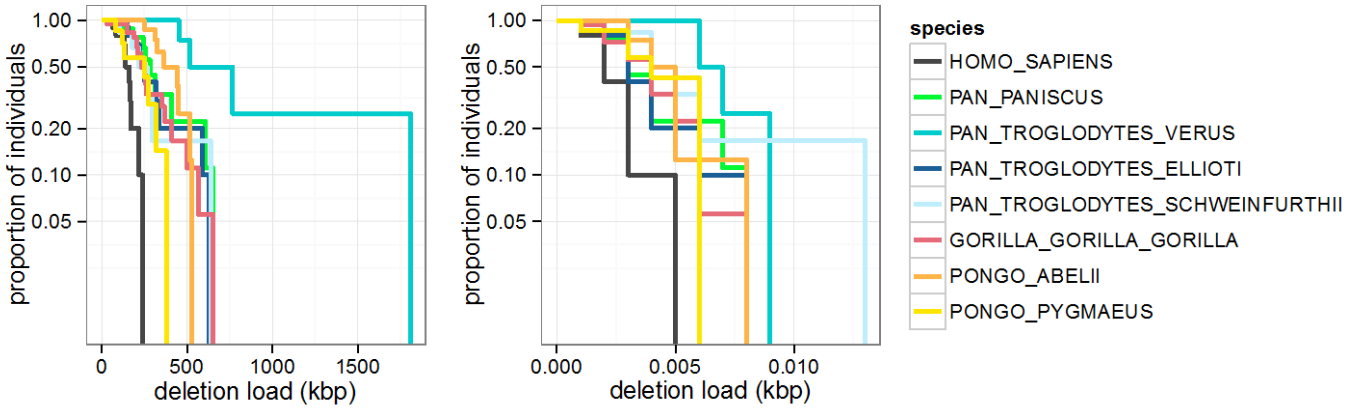


Figure 10.9 Survival functions of the proportion of individuals with at least a given total base-pair load of deletions plotted by total kbp and by number of sites. Western chimpanzees have an increased load of deletions.

Section 11: A human genomic disorder identified in a nonhuman primate

One particularly striking structural variant we identified was a ~1.6 Mbp microdeletion on 17p11.2 in the Western chimpanzee Susie-A (chr17:16,658,625-18,271,593). This deletion event encompasses 29 genes, notably including the gene retinoic acid-induced 1 (*RAI1*), deletions of which cause Smith-Magenis syndrome (SMS) in humans, a rare syndrome with an incidence of 1 in 15,000-25,000 human births. SMS is a complex neurobehavioral disorder resulting mental retardation and developmental delay, behavioral abnormalities, and facial and skeletal dismorphologies(Elsea and Girirajan 2008). We validated this event by arrayCGH against the Western chimpanzee Rappa (**Figure 11.1**).

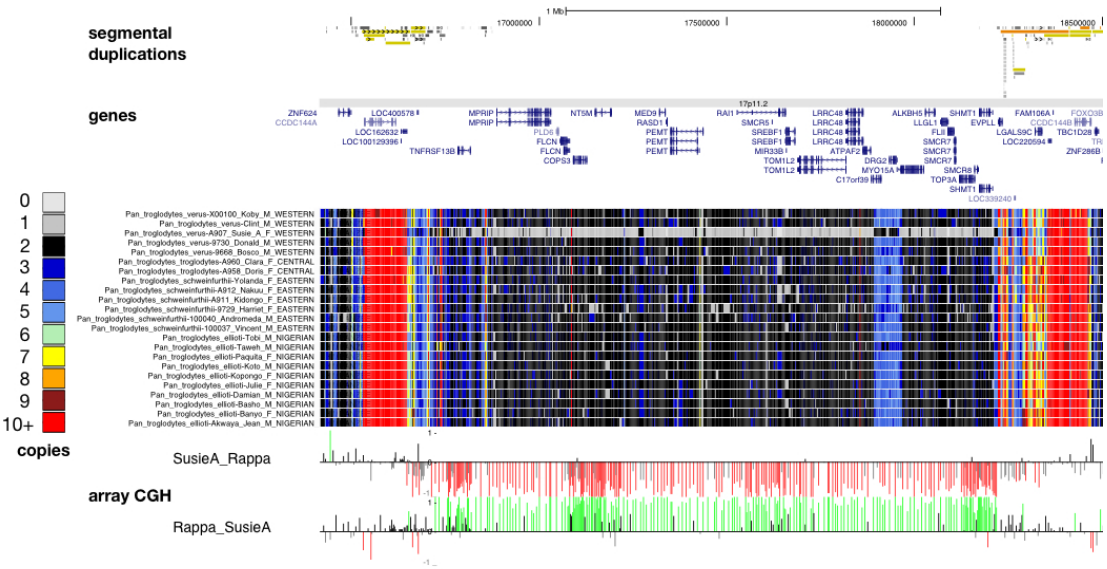


Figure 11.1: A 1.74 Mbp microdeletion on 17p11.2 in the chimpanzee Susie-A. The deletion overlaps 29 genes including *RAI1*; haploinsufficiency in *RAI1* results in SMS in humans, a severe neurobehavioral.

Though Susie-A, who was captured in 1975, died in 1997, we were able to obtain a report of her clinical features from the records of her handlers and veterinarians. These descriptions bare striking similarity to many of the phenotypes observed in SMS patients **(Table 11.1**); notably, maladaptive behaviors such as aggression and disobedience were identified in Susie-A, who was described a "mean and more aggressive than usual" chimpanzee. Susie-A also presented a hump on her back similar to many SMS patients who exhibit abnormal spinal curvatures, was obese—a common characteristic of SMS patients, and died of interstitial nephritis. Renal abnormalities are common in SMS patients. Finally, Susie-A exhibited tracheitis and had grossly overlapping tracheal cartilage ends, which contributed to noted breathing noises she would make; 50-75% of SMS patients exhibit tracheobronchial problems and velopharyngeal insufficiency.

Table 11.1: Common clinical features of SMS and related features of Susie-A.

Clinical features of Smith-Magenis syndrome	**Related clinical features of Susie-A**
Maladaptive behavioral issues including: • Frequent outbursts and tantrums • Aggression • Disobedience • Emotional volatility • Tendency toward attention-seeking behaviors • Lack of respect for personal space during conversation.	Susie-A is described as having exhibiting • 'marked impairment in her behavioral skills,' • 'mean and more aggressive than usual' behavior • When in the close proximity to people Susie-A would palpate her vagina, which is known to be a challenging "culturally abnormal" sexualized behavior in chimpanzees.

Abnormal curvature of the spine and scoliosis are present in 50-75% of individuals with SMS(Greenberg et al. 1996; Pagon et al. 1993).	Susie-A had kyphoscoliosis as a result of abnormal spine curvature.
Edelman *et al*(Edelman et al. 2007) found that individuals with *RAI1* mutations and deletions were likely to be obese. Mouse models of *RAI1* deletions additionally demonstrated an obese phenotype for deletions but not duplications of the critical locus(Walz et al. 2003; Yan et al. 2004); duplications conferred an underweight phenotype. A null *RAI1* allele in mice generated by Bi *et al*(Bi et al. 2005) also exhibited obesity.	Susie-A was an obese chimp with a body weight of ~90 kg. The normal body weight of a mature chimpanzee is between 50 kbp and 65 kg and less for female Western chimpanzees.
>75% of SMS patients exhibit otolaryngologic abnormalities including a hoarse, deep voice and 50-75% exhibit tracheobronchial problems and velopharyngeal insufficiency(Pagon et al. 1993).	Susie-A exhibited tracheitis and had grossly overlapping tracheal cartilage ends, which are suspected to contribute to documented breathing noises she would make.
Renal abnormalities have been shown to occur in 20-35% of SMS patients(Greenberg et al. 1996; Pagon et al. 1993). Additionally, 25%-50% of SMS patients have been found to have cardiac abnormalities.	Susie-A died of chronic interstitial nephritis and was found to have increased creatinine and leukocytosis consistent with renal failure. Numerous eosinophils were found in her gastrointestinal tract with no detection of parasites. Obesity and tracheitis presumably played a role in her final cardiorespiratory failure as well.

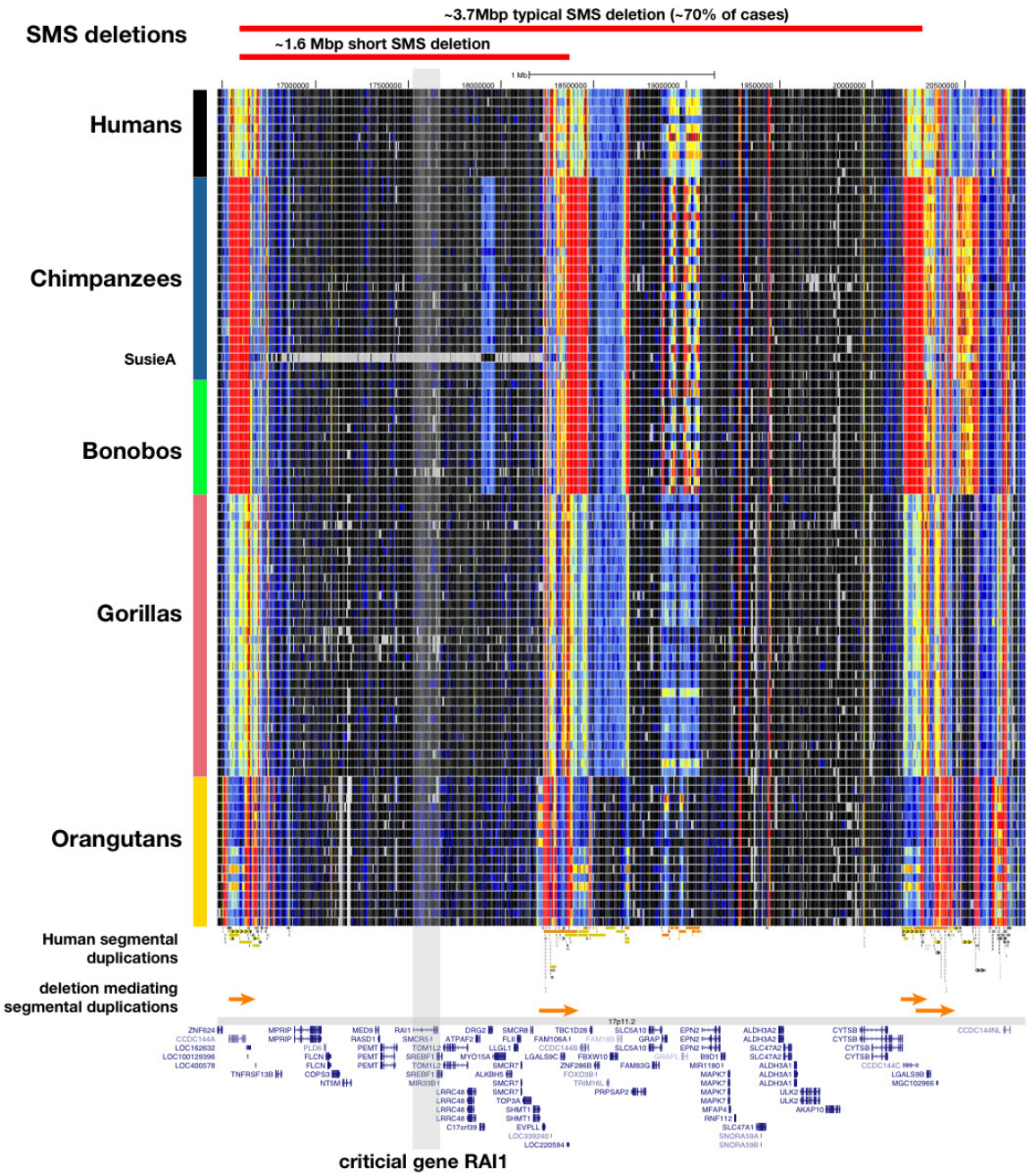


Figure 11.2: Architecture of the SMS/PTLS locus 17q11.2 in great apes. In humans, SDs in direct orientation mediate two different deletions sharing a distal breakpoint, both of which delete the critical gene *RAI1*. Susie-A exhibits the rarer of the two deletions.

SMS is caused by haploinsufficiency of *RAI1*, which is most often the result of a ~3.7 Mbp deletion mediated by two highly homologous SDs (~70% of cases(Elsea and Girirajan 2008)). Atypical deletions encompassing *RAI1* and *RAI1* mutations are responsible for an additional fraction of cases. The event we identify in Susie-A is much smaller than the typical human deletion (**Figure 11.2**). Notably, the SDs of the 17p11.2 locus have increased in copy number

specifically in the *Pan* genus (**Figure 11.3, Figure 11.4**), perhaps modifying the architecture of bonobo and chimpanzee SDs to predispose to this smaller deletion event.

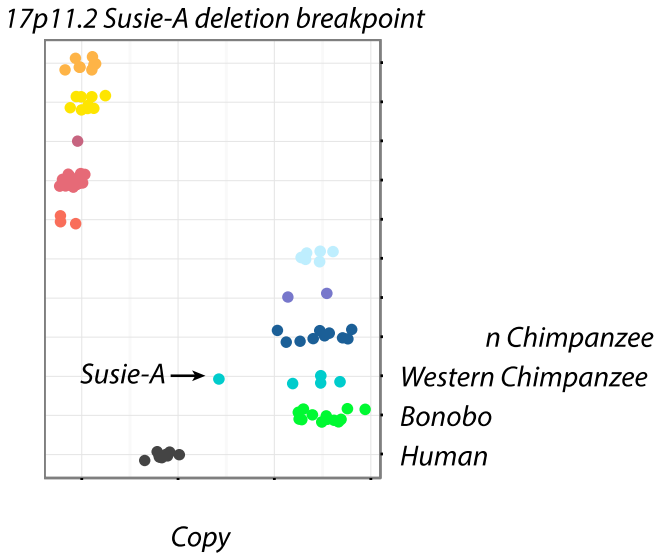


Figure 11.4: Copy number of the H duplicon at 17p11.2 is shown for different individuals clustered by great ape species. Chimpanzees and bonobos have increased copy number of the duplication ranging from 6-8 copies, with the exception of Suzie-A who has lost a copy. Humans exhibit 4 copies of the region.

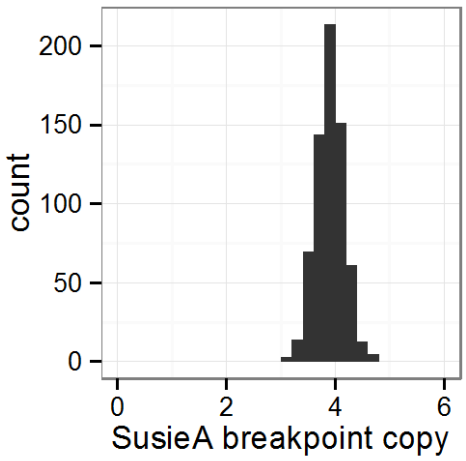


Figure 11.4: Copy number of the H duplicon, which is lost in Susie-A, genotyped across 675 individuals. 100% of humans robustly show 4 copies of this duplication.

To resolve the architecture of the chimpanzee 17p11.2 locus, we turned to large-insert clone sequencing. We began by searching for chimpanzee BACs that mapped to the SDs of the SMS region in 17p11.2 of HG18 to determine the structure of SDs present in that region. We identified 37 such BACs that mapped to this locus and were annotated as mapping to chromosome 17 or

unknown. Two of these BACs had previously been Sanger sequenced and we sequenced the remaining clones using a PacBio RS system, successfully assembling 20 more nonredundant clones (**Table 11.2**). To determine true anchor positions for each BAC, we ran whole-genome shotgun sequence detection (WSSD(Bailey et al. 2002)) of duplication content in BACs using human, chimpanzee, gorilla, and orangutan reads and determined regions in BACs with no coverage by WSSD in any species. We searched HG18 with these unique regions to find the highest identity location and considered these positions the "anchor" for each BAC. We were able to anchor 13 BACs in this manner. Aligning BACs to each other, we were then able to construct three super-contigs from seven BACs. Finally, we used DupMasker(Jiang et al. 2008) to determine the content of the duplication blocks in humans and chimpanzees and aligned our contigs to both the human and chimpanzee reference genomes to assess the differences in content and structure (**Figure 11.5**). BACs that we could not anchor with 100% certainty were placed to their most likely location given their structure and content in the context of Susie-A's deletion.

Table 11.2 : Sequenced BACs assessed and sequencing technology used.

Clone name	Source	Clone name	Source
CH251-545A24 (AC183294.3)	Sanger	**CH251-35C10**	PacBio
CH251-59G18	PacBio	**CH251-48B19**	PacBio
CH251-60C22 (AC183837.3)	Sanger	**CH251-354I21**	PacBio
CH251-21N5	PacBio	**CH251-118C16**	PacBio
CH251-182P19	PacBio	**CH251-13D21**	PacBio
CH251-425F18	PacBio	**CH251-173I3**	PacBio
CH251-33G11	PacBio	**CH251-184C19**	PacBio
CH251-179D15	PacBio	**CH251-18I21**	PacBio
CH251-133B4	PacBio	**CH251-251H7**	PacBio
CH251-200P13	PacBio	**CH251-30P23**	PacBio
CH251-433E19	PacBio	**CH251-75B17**	PacBio

Table 11.3: Summary of BACs used to resolve the architecture of the SMS region in chimpanzees. Clone information is grouped by sequencing technology (PacBio and Sanger). PacBio clones were sequenced and assembled in-house; Sanger clones were sequenced at The Genome Institute at Washington University.

Clone assembly	Number	Size (bp)	Unique by WSSD (bp)	Discontiguous wrt HG18 (bp)	Contiguous and inverted wrt HG18 (bp)	Collapsed duplications in assembly (bp)
PacBio	20	2,935,821	2,181,309	1,194,224	71,582	268,000
Sanger	2	413,710	95,750	219,862	0	N/A
Total	22	3,349,531	2,277,059	1,414,086	71,582	268,000

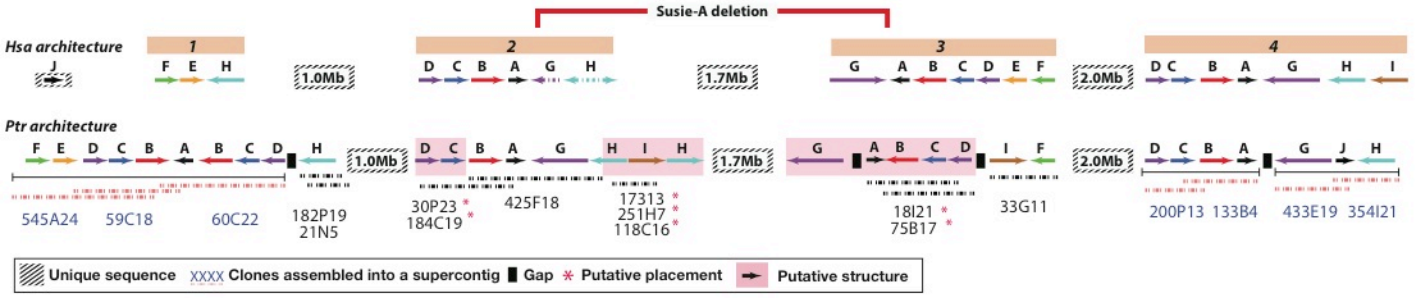


Figure 11.5: Duplicon structure of the human and chimpanzee 17p11.2 locus. Duplication blocks are labeled 1-4 and annotated with the underlying structure of SDs. We hypothesize that the directly oriented G-duplicon in chimpanzees is responsible for mediating the Susie-A deletion.

Section 12: References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**: 1061–1067.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW,

Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Bi W, Ohyama T, Nakamura H, Yan J, Visvanathan J, Justice MJ, Lupski JR. 2005. Inactivation of Rai1 in mice recapitulates phenotypes observed in chromosome engineered mouse models for Smith-Magenis syndrome. *Hum Mol Genet* **14**: 983–995.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**: 88–93.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704–712.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**: 838–846.

Edelman EA, Girirajan S, Finucane B, Patel PI, Lupski JR, Smith ACM, Elsea SH. 2007. Gender, genotype, and phenotype differences in Smith-Magenis syndrome: a meta-analysis of 105 cases. *Clin Genet* **71**: 540–550.

Elsea SH, Girirajan S. 2008. Smith–Magenis syndrome. *Eur J Hum Genet* **16**: 412–421.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* **320**: 1629–1631.

Greenberg F, Lewis RA, Potocki L, Glaze D, Parke J, Killian J, Murphy MA, Williamson D, Brown F, Dutton R, et al. 1996. Multi-disciplinary clinical study of Smith-Magenis syndrome (deletion 17p11.2). *Am J Med Genet* **62**: 247–254.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC. 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**: 576–577.

Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2009. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19**: 1270–1278.

Jiang Z, Hubley R, Smit A, Eichler EE. 2008. DupMasker: a tool for annotating primate segmental duplications. *Genome Res* **18**: 1362–1368.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Langergraber KE, Prüfer K, Rowney C, Boesch C, Crockford C, Fawcett K, Inoue E, Inoue-Muruyama M, Mitani JC, Muller MN, et al. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci USA* **109**: 15716–15721.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lichter P, Tang CJ, Call K, Hermanson G, Evans GA, Housman D, Ward DC. 1990. High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247**: 64–69.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**: 529–533.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**: 877–881.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**: 216–219.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**: 222–226. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22936568&retmode=ref&cmd=prlinks.

Miller DT, Nasir R, Sobeih MM, Shen Y, Wu B-L, Hanson E. 1993. *16p11.2 Microdeletion - GeneReviews*. University of Washington, Seattle, Seattle (WA) http://www.ncbi.nlm.nih.gov/books/NBK11167/.

Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet* **64**: 18–23.

Pagon RA, Bird TD, Dolan CR, Stephens K, Adam MP, Smith AC, Boyd KE, Elsea SH, Finucane BM, Haas-Givler B, et al. 1993. *Smith-Magenis Syndrome*. University of Washington, Seattle, Seattle (WA).

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2**: e190.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169–175.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, et al. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330**: 641–646.

Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, et al. 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet* **44**: 1161–1165.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with

TopHat and Cufflinks. *Nat Protoc* **7**: 562–578.

Varki A, Geschwind DH, Eichler EE. 2008. Explaining human uniqueness: genome interactions with environment, behaviour and culture. *Nat Rev Genet* **9**: 749–763.

Ventura M, Catacchio CR, Alkan C, Marques-Bonet T, Sajjadian S, Graves TA, Hormozdiari F, Navarro A, Malig M, Baker C, et al. 2011. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res* **21**: 1640–1649.

Walz K, Caratini-Rivera S, Bi W, Fonseca P, Mansouri DL, Lynch J, Vogel H, Noebels JL, Bradley A, Lupski JR. 2003. Modeling del(17)(p11.2p11.2) and dup(17)(p11.2p11.2) contiguous gene syndromes by chromosome engineering in mice: phenotypic consequences of gene dosage imbalance. *Mol Cell Biol* **23**: 3646–3655.

Witkin A. 1984. Scale-space filtering: A new approach to multi-scale description. **9**: 150–153.

Yan J, Keener VW, Bi W, Walz K, Bradley A, Justice MJ, Lupski JR. 2004. Reduced penetrance of craniofacial anomalies as a function of deletion size and genetic background in a chromosome engineered partial mouse model for Smith-Magenis syndrome. *Hum Mol Genet* **13**: 2613–2624.