SUPPLEMENTAL INFORMATION

Datasets

Dataset 1 (Separate Excel sheet) - Overview of sample properties and sequencing statistics

Dataset 2 (Separate Excel sheet) Summary of PCR verification experiments for deletions, duplications, non-reference MEIs (mobile element insertions), SNPs and Indels

Dataset 3 (Separate Excel sheet) List of all SVs (deletions, duplications, MEIs) discovered in this study and list of indels and SNPs affecting protein-coding sequences that our study discovered in chimpanzee, orang-utan and rhesus macaque.

Dataset 4 (Separate Excel sheet) List of inter-species gene duplications

Dataset 5 (Separate Excel sheet) Table of genes inferred to acquire expression in a new tissue

Supplemental Tables

Table S1. Sequencing statistics and non-redundant discovered genetic variants across five individuals in chimpanzee, orang-utan and rhesus macaque.

Statistic	Chimpanzee	Orang-utan	Rhesus macaque
Total raw bases (Gb)	358.94	332.43	299.37
Total mapped bases (Gb) / (%)	296.45 (82.59%)	264.1 (79.44%)	240.5 (80.34%)
Mean coverage per species	19X	17X	17X
Total coverage per species	96X	86X	85X
Polymorphic deletions (not including MEIs)	2680	4983	3905
Polymorphic duplications (not including MEIs)	1499	1095	807
Fixed unannotated duplications (not including MEIs)	1910	540	625
Novel polymorphic MEI insertions (“non-reference MEI”)	764	2548	15566
Polymorphic MEIs (“reference MEI“)	94	315	1124
SNPs	6643502	12806033	13762492
SNPs, predicted functional effect	24313	29379	39209
Indels	541080	952602	1205735
Indels, predicted functional effect	3894	3402	5508

TABLE S2. Estimated rates (μ, per site per generation) for the formation of different sequence variant classes in primates. We estimated genome-wide SV formation rates for different SV formation mechanisms based on estimated SNP mutation rates (the latter of which are comparable to published results [27,28,29]. We surmised that SNP formation rates are proxies of demographic trends and overall diversity among the individuals that we studied in each species, and inferred formation rates of NAHR, NHR, Alu and L1 by relating the number of observations for each mechanism with the observed number of SNPs. Note that these inferred rates are likely under-estimating the rate of SV formation, owing to our low sensitivity for characterizing SVs in repeat-rich areas of the genome prone to form SVs, as well as to the increased relative impact of purifying selection on SVs. These results further support our observation that NAHR is more abundant, whereas Alu insertions are less prevalent, in great apes as compared to macaques. Furthermore, the differences in rates reported here are concordant with previous reports that Alu polymorphisms are highly active in rhesus macaques [4] and L1s are relatively more active in orang-utans [5].

Statistic	Chimpanzee	Orang-utan	Rhesus macaque
Estimated θ	6.8×10^{-4}	1.15×10^{-3}	1.29×10^{-3}
N_e	11,413	37,590	80,000
Calculated μ (SNP)	1.5×10^{-8}	7.6×10^{-9}	4.0×10^{-9}
Calculated μ (NAHR)	2.0×10^{-12}	1.4×10^{-12}	1.2×10^{-13}
Calculated μ (NHR)	1.6×10^{-12}	9.0×10^{-13}	5.0×10^{-13}
Calculated μ (Alu)	5.1×10^{-13}	7.4×10^{-14}	3.3×10^{-12}
Calculated μ (L1)	6.1×10^{-13}	1.0×10^{-12}	3.9×10^{-13}

Supplemental Figures

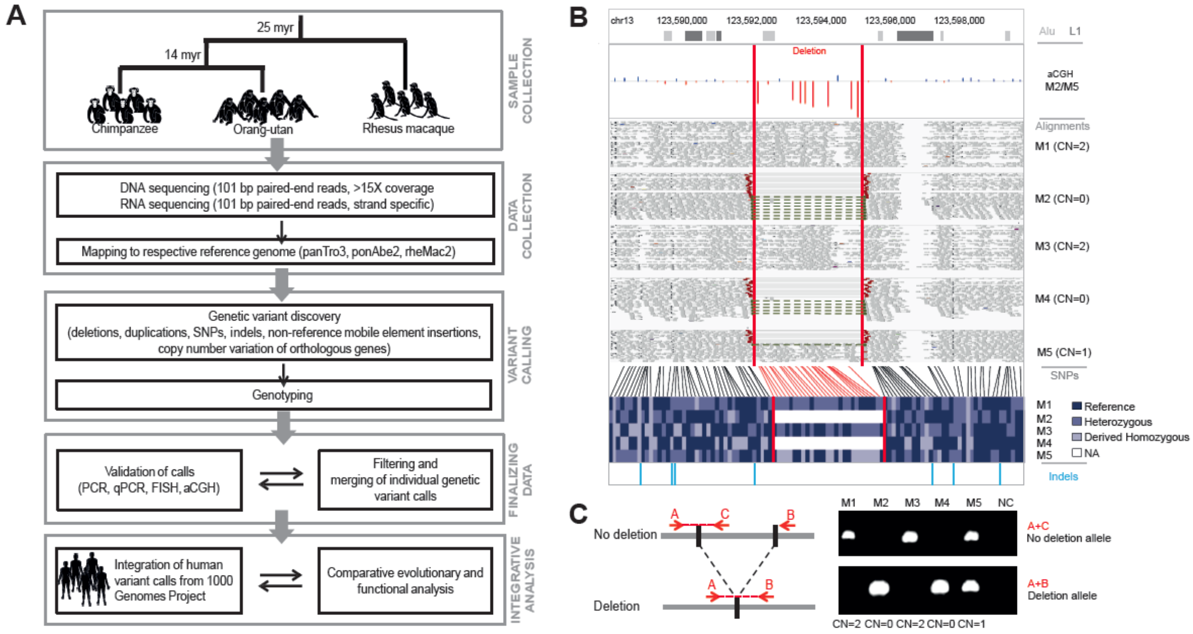

A
25 myr
14 myr
Chimpanzee
Orang-utan
Rhesus macaque
SAMPLE COLLECTION
DNA sequencing (101 bp paired-end reads, >15X coverage
RNA sequencing (101 bp paired-end reads, strand specific)
Mapping to respective reference genome (panTro3, ponAbe2, rheMac2)
DATA COLLECTION
Genetic variant discovery
(deletions, duplications, SNPs, indels, non-reference mobile element insertions, copy number variation of orthologous genes)
Genotyping
VARIANT CALLING
Validation of calls
(PCR, qPCR, FISH, aCGH)
Filtering and merging of individual genetic variant calls
FINALIZING DATA
Integration of human variant calls from 1000 Genomes Project
Comparative evolutionary and functional analysis
INTEGRATIVE ANALYSIS
B
chr13
123,590,000
123,592,000
123,594,000
123,596,000
123,598,000
Alu
L1
Deletion
aCGH
M2/M5
Alignments
M1 (CN=2)
M2 (CN=0)
M3 (CN=2)
M4 (CN=0)
M5 (CN=1)
SNPs
M1
M2
M3
M4
M5
Reference
Heterozygous
Derived Homozygous
NA
Indels
C
No deletion
Deletion
A
C
B
A
B
M1
M2
M3
M4
M5
NC
A+C
No deletion allele
A+B
Deletion allele
CN=2 CN=0 CN=2 CN=0 CN=1

Figure S1. Analysis workflow with an example of a deletion event. A. Variant discovery, genotyping, and validation approaches for characterizing SVs in non-human primates and performing comparative analyses with humans. **B.** Illustration of a region with a polymorphic deletion in macaques. The uppermost section marks the genomic coordinates (based on the rheMac2 reference assembly). Mobile elements already annotated in the reference genome are depicted by gray bars below. Blue and red bars below depict positive and negative aCGH signal intensity $\log_2$ ratios (sample M2 was hybridized relative to sample M5). Homozygous deletions in M2 and M4 and a heterozygous deletion (i.e. copy-number (CN) equals 1) in M5 were independently confirmed by read depth (gray dashes), abnormally mapped paired-end reads (red dashes on each side of the deletion), and split reads (green dashes). The locations of SNPs are indicated with vertical lines, and SNP genotypes in all five macaque samples using a blue color scheme. Locations of small indels (light blue) are in the lowermost panel. **C.** Polymerase chain reaction **(**PCR) based verification of the rhesus macaque deletion (amplicons span the deletion, or one junction of the deletion). NC, negative control.

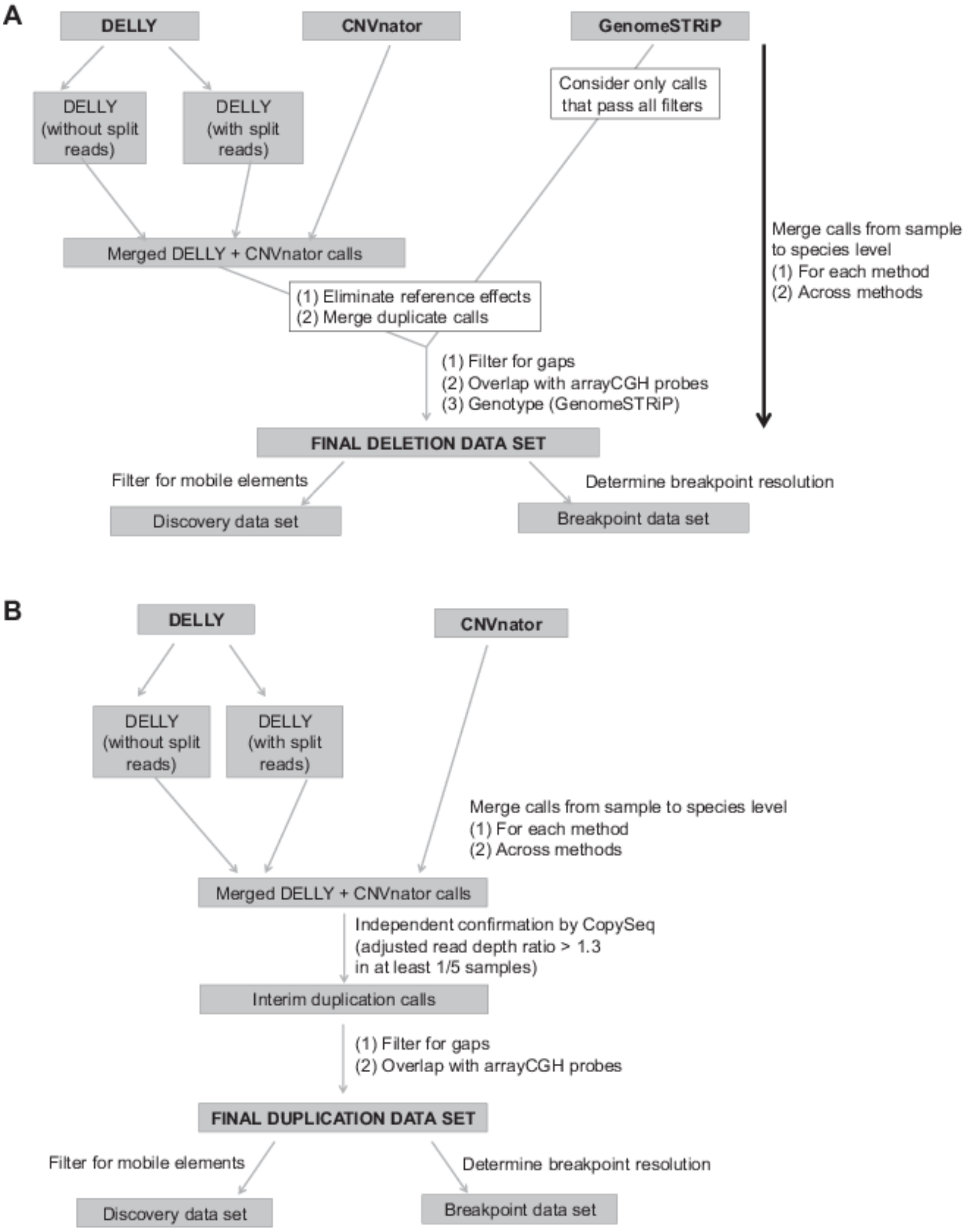

A
DELLY
CNVnator
GenomeSTRiP
DELLY (without split reads)
DELLY (with split reads)
Consider only calls that pass all filters
Merge calls from sample to species level
(1) For each method
(2) Across methods
Merged DELLY + CNVnator calls
(1) Eliminate reference effects
(2) Merge duplicate calls
(1) Filter for gaps
(2) Overlap with arrayCGH probes
(3) Genotype (GenomeSTRiP)
FINAL DELETION DATA SET
Filter for mobile elements
Determine breakpoint resolution
Discovery data set
Breakpoint data set
B
DELLY
CNVnator
DELLY (without split reads)
DELLY (with split reads)
Merge calls from sample to species level
(1) For each method
(2) Across methods
Merged DELLY + CNVnator calls
Independent confirmation by CopySeq
(adjusted read depth ratio > 1.3
in at least 1/5 samples)
Interim duplication calls
(1) Filter for gaps
(2) Overlap with arrayCGH probes
FINAL DUPLICATION DATA SET
Filter for mobile elements
Determine breakpoint resolution
Discovery data set
Breakpoint data set

Figure S2. Structural variant (SV) discovery. A. Flowchart of comprehensive deletion discovery pipeline. **B.** Flowchart of comprehensive duplication discovery pipeline.

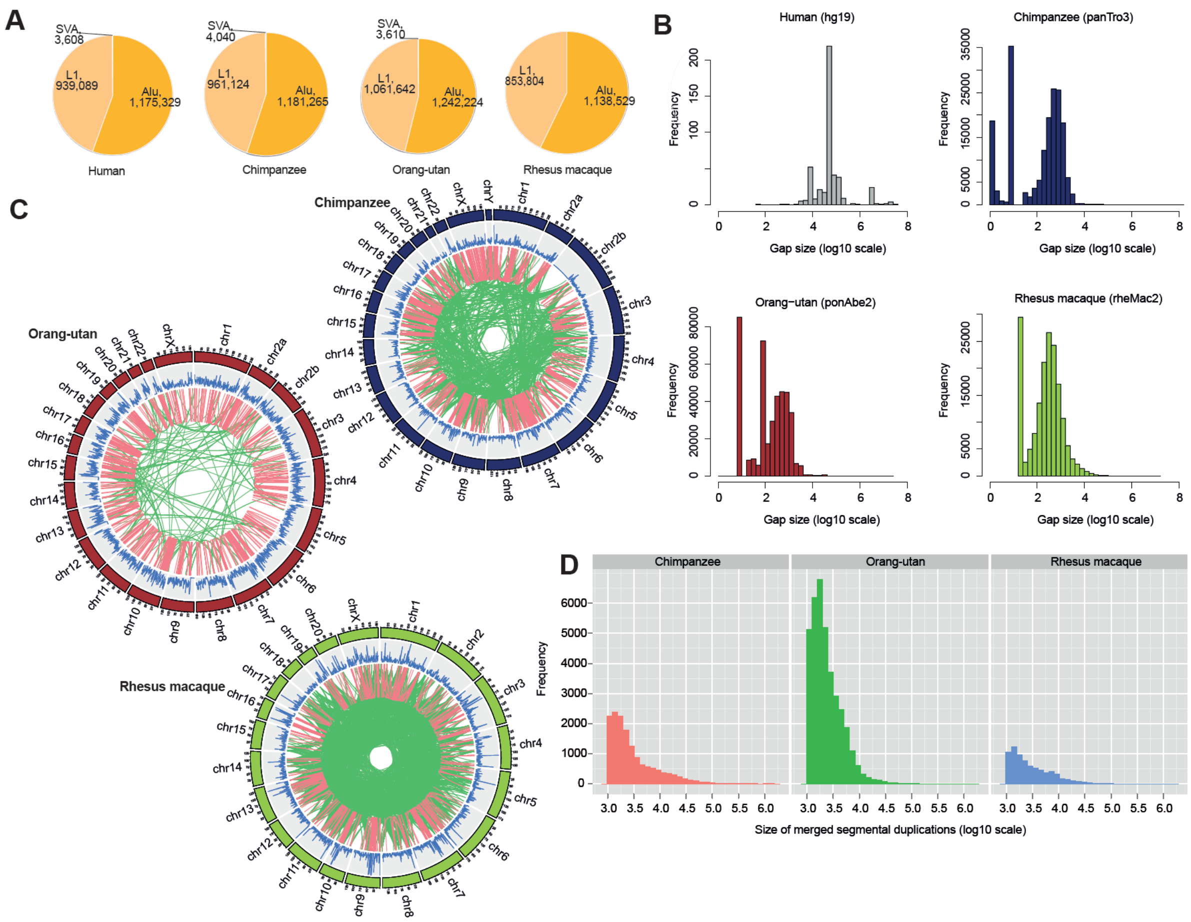


Figure S3. Reference genome properties. A. Annotation of major mobile element classes in the reference genomes of human (hg19), chimpanzee (panTro3), orang-utan (ponAbe2) and rhesus macaque (rheMac2), based on the RepeatMasker [1] track provided through the UCSC genome browser [2]. **B.** Distribution of gap sizes in different primate species as described in their respective reference assemblies [3,4,5]. X-axes show the size of gaps in $\log_{10}$ scale, y-axes the frequency of gaps of a given size; y-axes are scaled differently due to high variance in the numbers of gaps between the respective reference genomes. **C.** Genome-wide distribution of pairwise segmental duplications (SDs). Intrachromosomal SD calls are shown in red and interchromosomal SD calls in green. SD density along the genome is depicted in blue. To enhance visibility, only those SDs between 7.5 and 50kb in size are depicted (similar results are observed for different size windows). **D.** Size distribution of all SDs in three primate genomes (with the x-axis using a $\log_{10}$ scale).

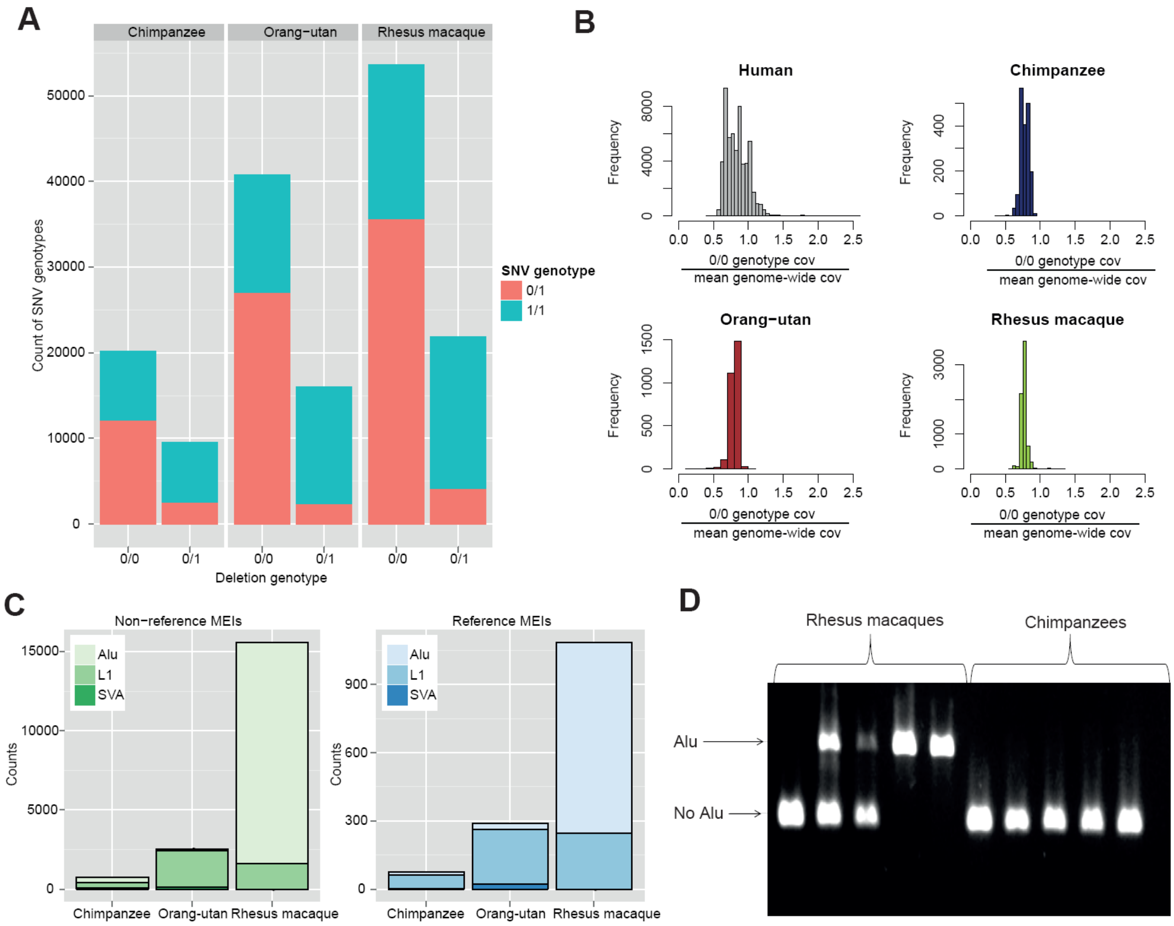
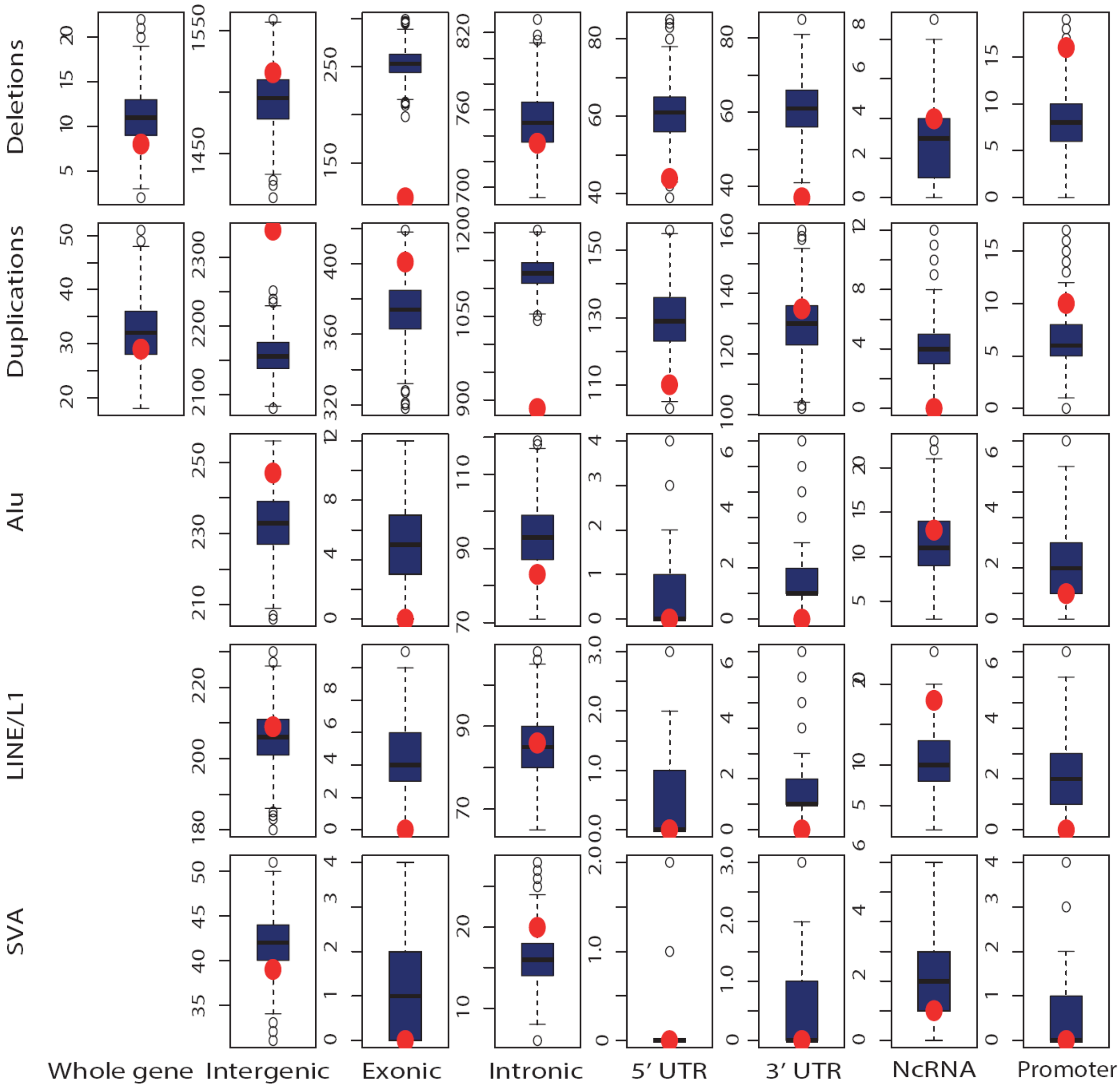
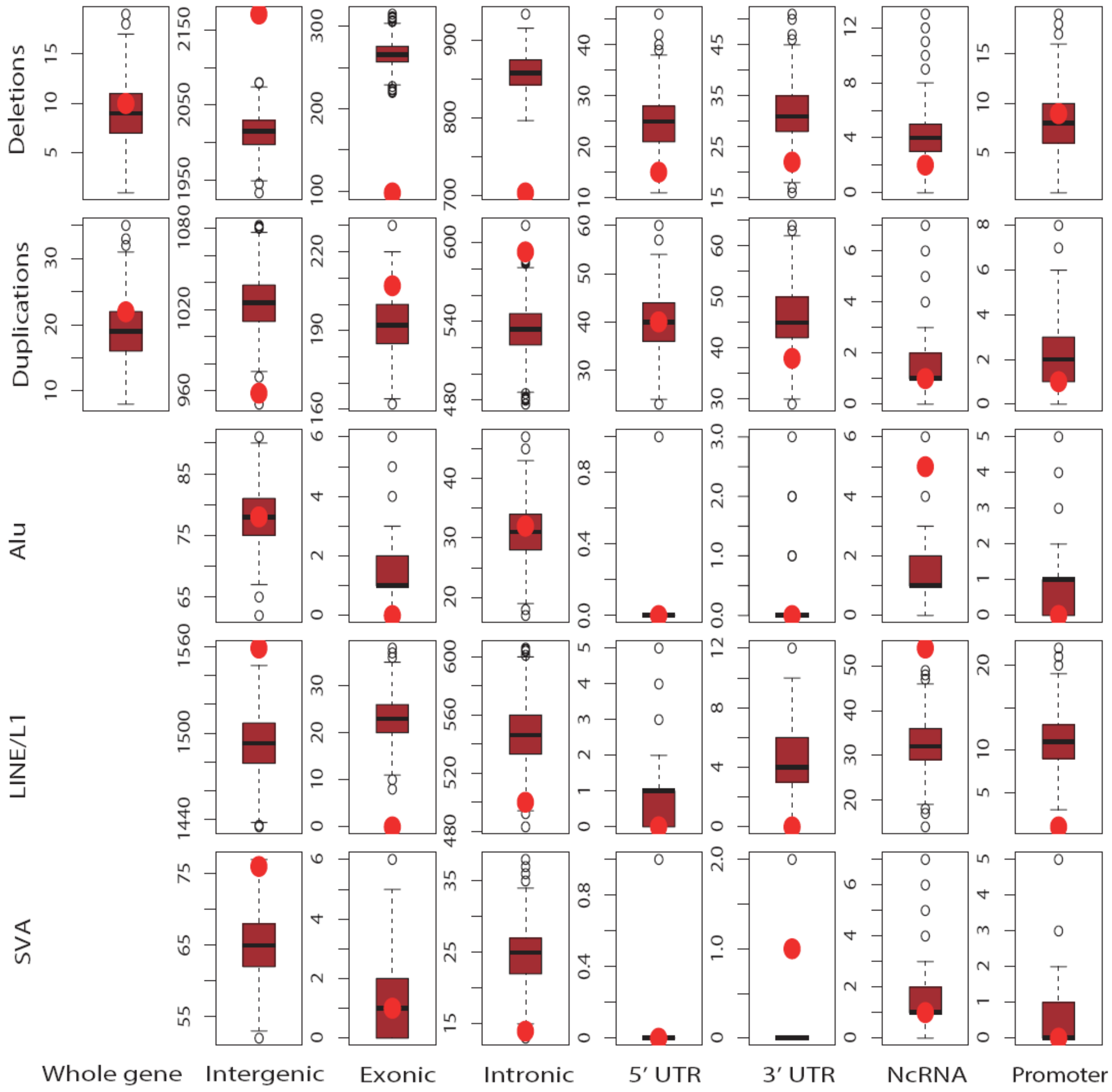


Figure S4. Assessment of the quality of SV calls in non-human primates. A. Single nucleotide polymorphism (SNP) heterozygosity within the inferred boundaries (*i.e.*, breakpoints) of genotyped heterozygous deletions (regions with copy-number=1, *i.e.* deletion genotype 0/1). SNP heterozygosity for individuals with copy-number=2 (deletion genotype 0/0), is shown for comparison. SNP genotype 0/1 are heterozygous SNPs, SNP genotype 1/1 are homozygous SNPs. **B.** Read coverage at deletion loci for samples with homozygous reference allele (deletion genotype 0/0) relative to genome-wide read coverage. Human data are from 1000 Genomes Project (deletion calls reported in [6]). "cov" is read coverage. **C.** Comparison of the distributions of different MEI classes discovered by two different calling algorithms. On the left: ‘Novel’ MEIs (non-reference MEIs) identified by the TEA algorithm On the right: MEIs that are annotated in the reference genome, but discovered as polymorphic deletions and classified as reference MEIs by BreakSeq. While overall numbers of discovered MEIs differed depending on the approach, relative portions of identified MEI classes were highly robust. **D.** Quality assessment of non-reference MEIs. We designed PCR assays to verify that non-reference MEIs discovered by our pipeline are polymorphic within the given species (rhesus macaques, in this example), but absent in the other species (chimpanzee, in this example).



Deletions
Duplications
Alu
LINE/L1
SVA
Whole gene
Intergenic
Exonic
Intronic
5′ UTR
3′ UTR
NcRNA
Promoter
5
10
15
20
1450
1550
150
250
700
760
820
40
60
80
20
30
40
50
2100
2200
2300
320
360
400
900
1050
1200
110
130
150
100
120
140
160
0
4
8
12
210
230
250
70
90
110
0
1
2
3
4
0
2
4
6
5
10
15
20
180
200
220
0.0
1.0
2.0
3.0
35
40
45
50
10
20


Deletions
Duplications
Alu
LINE/L1
SVA
Whole gene
Intergenic
Exonic
Intronic
5′ UTR
3′ UTR
NcRNA
Promoter
5
10
15
1950
2050
2150
100
200
300
700
800
900
10
20
30
40
15
25
35
45
0
4
8
12
5
10
15
10
20
30
960
1020
1080
160
190
220
480
540
600
30
40
50
60
30
40
50
60
0
2
4
6
0
2
4
6
8
65
75
85
0
2
4
6
20
30
40
0.0
0.4
0.8
0.0
1.0
2.0
3.0
0
2
4
6
0
1
2
3
4
5
1440
1500
1560
0
10
20
30
480
520
560
600
0
1
2
3
4
5
0
4
8
12
20
30
40
50
5
10
20
55
65
75
0
2
4
6
15
25
35
0
0.4
0.8
0
1.0
2.0
0
2
4
6
0
1
2
3
4
5

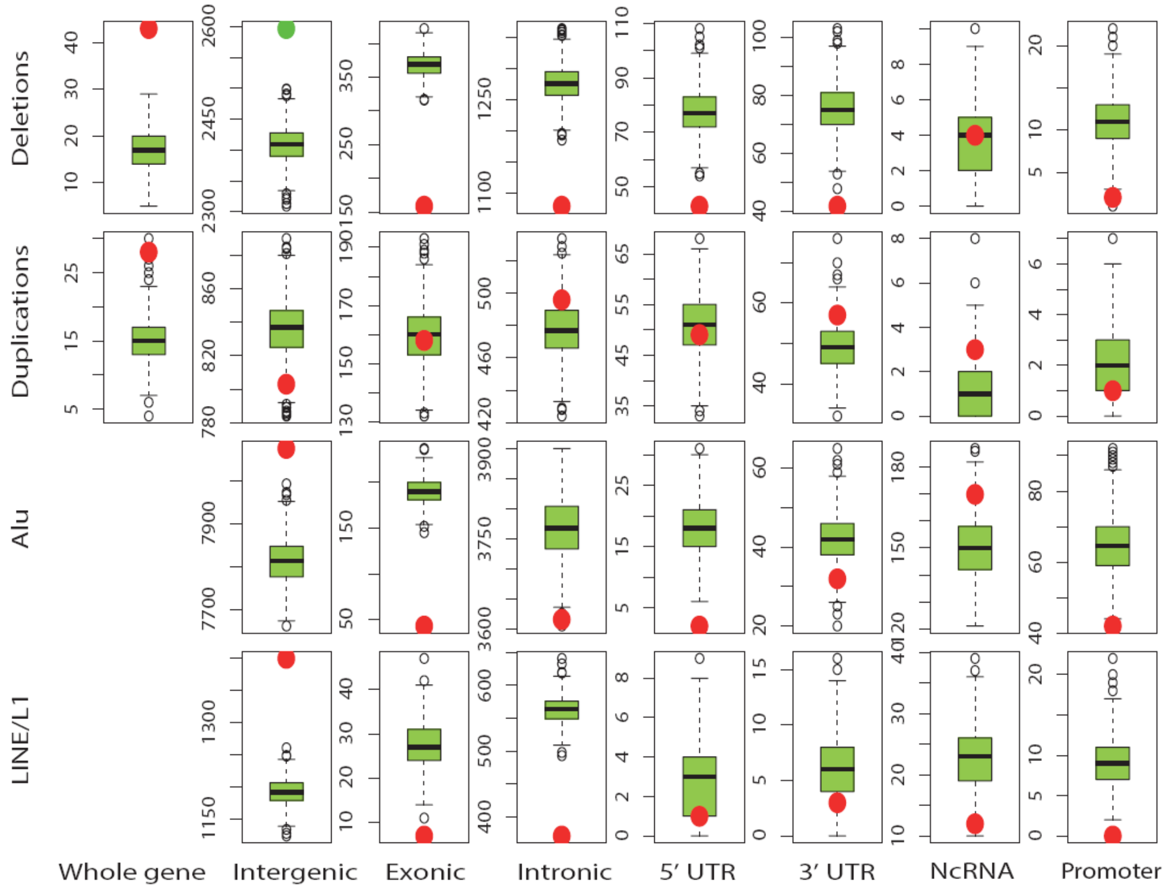


Figure S5. Estimated functional impact of structural variation in primate genomes. Functional impact of copy number variants (deletions and duplications) and mobile elements (Alu, LINE/L1, SVA) in chimpanzee, orang-utan and rhesus macaque. Different species are displayed in different colors (chimpanzee in blue, orang-utan in red, and rhesus macaque in green). Categories of whole gene, intergenic, exonic, intronic, 5’ UTR and 3’ UTR annotations were defined based on available gene annotations in each species (obtained through Ensembl BioMart, version 62 (April 2011)). Promoter predictions were obtained from MPromDB (http://mpromdb.wistar.upenn.edu/; May 2013), a curated mammalian promoter database that includes promoters inferred by ENCODE (http://genome.ucsc.edu/ENCODE/). Promoter coordinates were lifted over from hg19 to panTro3, ponAbe2 and rheMac2 respectively (using LiftOver). NcRNA = union of lincRNAs and microRNAs, whereby human lincRNAs were obtained from Ensembl BioMart (version 71, April 2013) and Cabili et al. [7] and lifted to panTro3, ponAbe2 and rheMac2 coordinates using LiftOver, and microRNAs were obtained from Ensembl BioMart (version 71, April 2013). To assess whether there is an enrichment or depletion of functional variation types, we performed 1,000 permutations in which we shuffled observed variant calls along chromosomes and determined their overlap with the respective genomic elements. The number of SVs overlapping with a certain type of

genome annotation is shown on the x-axes. Red dots represent the number of SVs overlapping with genome annotation, boxplots represent distributions of randomized SVs (1000 randomizations) intersecting with different genome annotation classes in the genome.

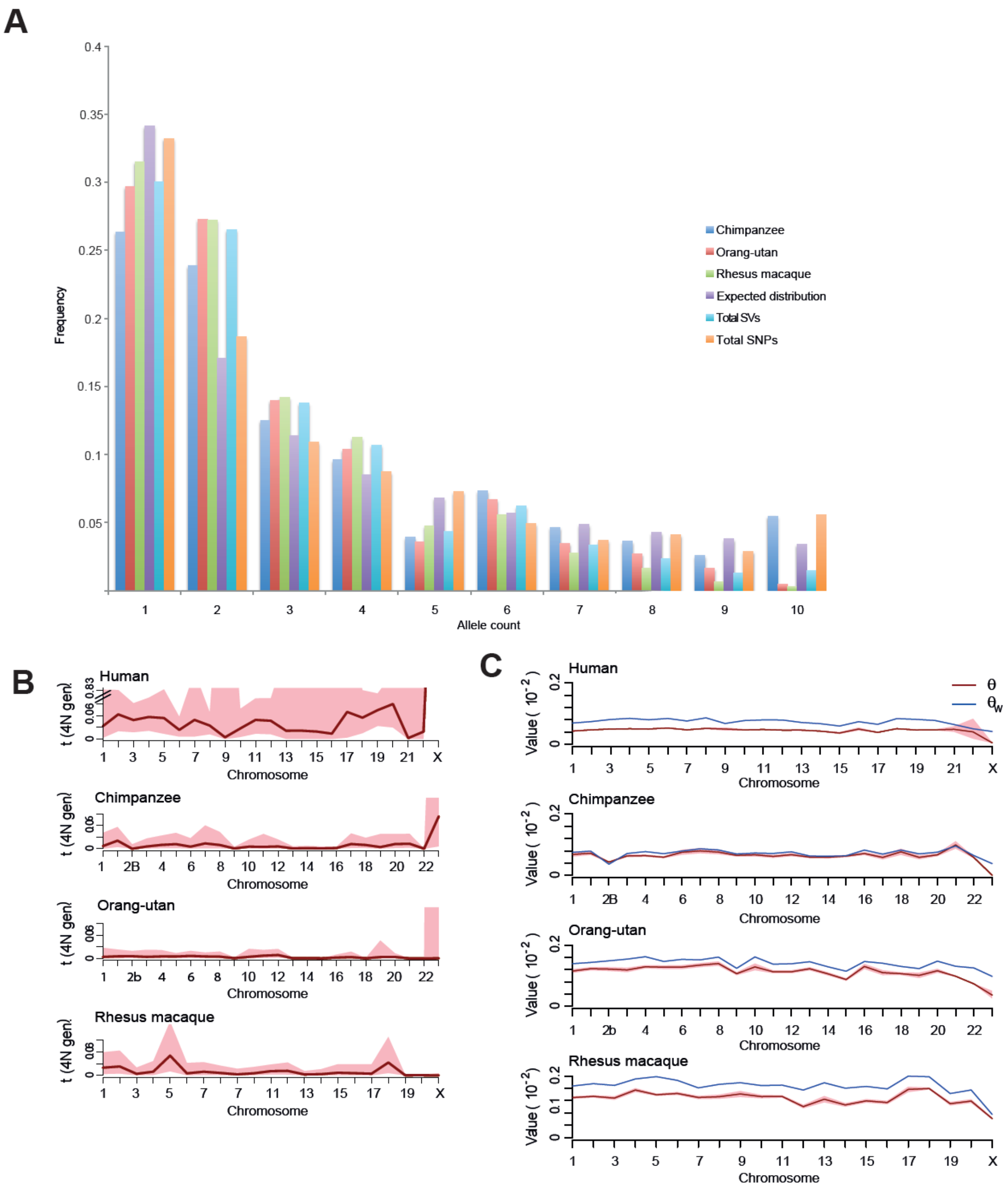


Figure S6. Selection analysis in primates A. Site Frequency spectrum for SVs per non-human primate species, for all SVs identified in the non-human primate genomes studied, and for all SNPs identified in these non-human primate genomes. Allele counts of deletions are on the x-axis, and observed frequencies in the respective non-human primate population are on the y-axis. The purple bars represent the expected distribution of allele counts under the neutral model 1/*i*. **B.** Estimation of parameter *t* for all chromosomes of human,

chimpanzee, orang-utan and rhesus macaque. The parameter *t* quantifies the time period for which chromosomes belonging to different individuals have been isolated from each other (thus, it reflects population substructure). The shaded (pink) area denotes the credible interval (0.05 - 0.95) for the parameter *t*. Observed credible intervals for most chromosomes indicate that chromosomes from different human, chimpanzee, orang-utan and rhesus individuals have been isolated from each other (*i.e.*, the individuals are 'non-relatives'). **C**. Estimation of the parameter theta (*θ*) and comparison with Watterson's estimate. Since the demographic model deviates from the standard neutral model and involves population substructure the Watterson's estimate is different than the value of the parameter theta. Specifically for the population substructure model Watterson's estimate is higher than the value of the parameter theta, which agrees with our testing hypothesis that while individuals are unrelated there is population substructure in the populations of human, chimpanzee, orang-utan and rhesus macaque assessed by this study.

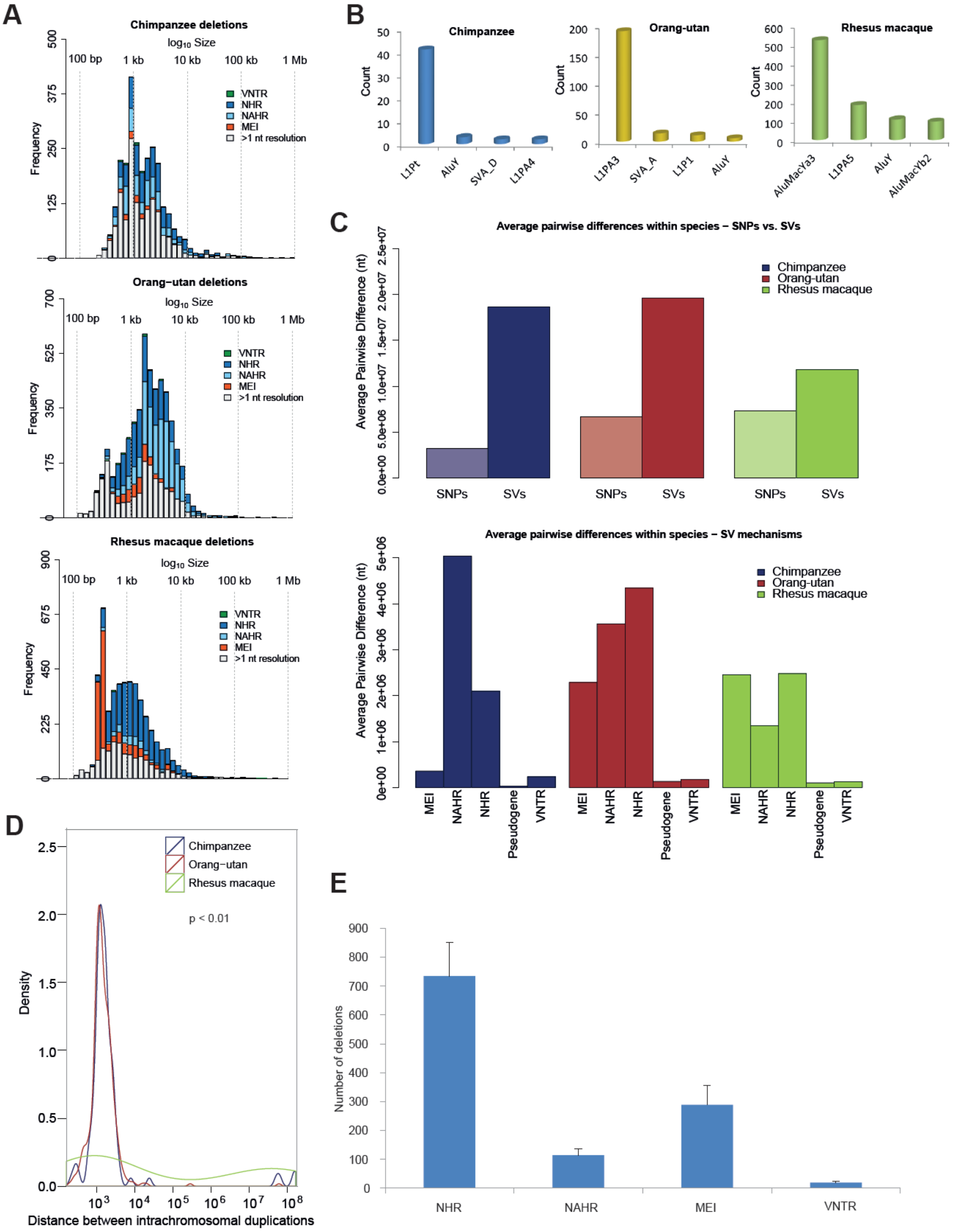
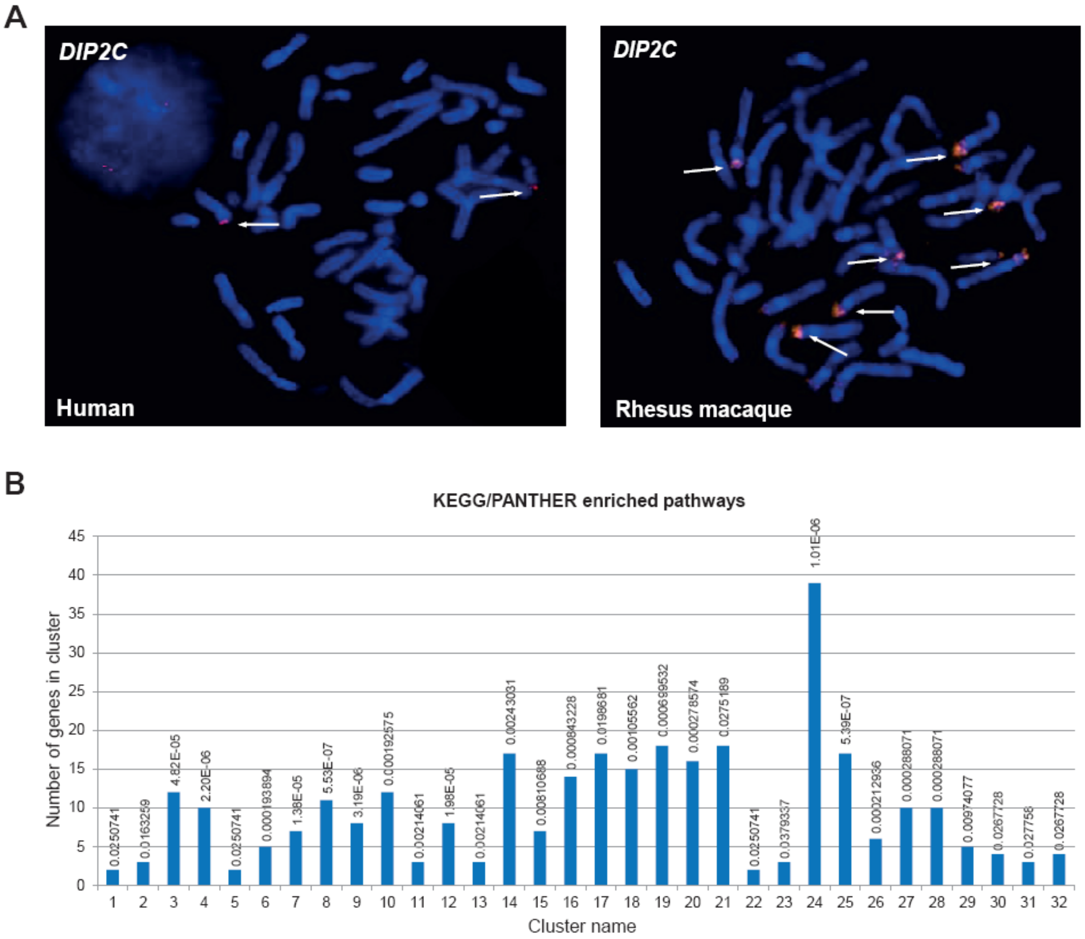


Figure S7. Differences in SV *de novo* formation mechanisms in non-human primates.

A. Size distributions of deletions in chimpanzee, orang-utan and rhesus macaque.

B. Abundance of the four most abundant MEI subfamilies annotated in the respective reference genomes. **C.** Average pairwise within-species differences of genetic variation in chimpanzee, orang-utan and rhesus macaque. Top panel: Average pairwise differences

between SNPs and SVs, the latter of which include deletions, duplications, as well as reference and non-reference MEIs. Lower panel: SVs separated by mechanism (note that SVs lacking nucleotide resolution breakpoints were not categorized into a mechanism).
D. Density plots depicting distributions of distances between intrachromosomal duplicative insertions and the respective genomic loci these insertions have arisen from. Great ape duplicative insertions typically arise from nearby genomic loci, different from macaque duplicative insertions. **E.** Estimation of the margin of error introduced by sampling five individuals from a species. We performed sampling experiments 10 times, whereby we randomly drew five human samples sequenced by the 1000 Genomes Project [6], obtained GenomeSTRiP, CNVnator, and DELLY deletion calls for all sets of 5 samples, and subsequently applied the merging steps from our deletion discovery pipeline depicted in Fig. S1. Bar graphs depict the overall numbers of deletions inferred in each sampling experiment, broken down by those inferred to be formed by NHR, NAHR, MEI, and VNTR expansion/contraction (with the portions of NHR, NAHR, MEI, and VNTR being at a similar level to those reported by Mills et al. [6] for the given SV size range). On average, we detected 803 deletion polymorphisms in each sample. The inferred standard deviation (represented by error bars) was 21.3% for deletions inferred to be formed by NHR, 21.6% for NAHR, 17.3% for MEI and 17.5% for VNTR, which suggests that reproducible SV callset numbers and formation mechanism breakdowns can be retrieved by sampling 5 individuals from a species.



Cluster name	Cluster details
1	Cytokine-cytokine receptor interaction,Inflammation mediated by chemokine and cytokine signaling pathway,Endocytosis,Chemokine signaling pathway,Interleukin signaling pathway,Epithelial cell signaling in Helicobacter pylori infection
2	Cytokine-cytokine receptor interaction,Chemokine signaling pathway,Epithelial cell signaling in Helicobacter pylori infection
3	Retinol metabolism
4	Retinol metabolism,Metabolism of xenobiotics by cytochrome P450,Drug metabolism - cytochrome P450
5	Retinol metabolism,Metabolism of xenobiotics by cytochrome P450,Drug metabolism - cytochrome P450,Bile secretion,Steroid hormone biosynthesis,Drug metabolism - other enzymes
6	Retinol metabolism,Metabolism of xenobiotics by cytochrome P450,Drug metabolism - cytochrome P450,Pentose and glucuronate interconversions,Ascorbate and aldarate metabolism,Steroid hormone biosynthesis,Starch and sucrose metabolism,Other types of O-glycan biosynthesis,Porphyrin and chlorophyll metabolism,Drug metabolism - other enzymes
7	Retinol metabolism,Metabolism of xenobiotics by cytochrome P450,Drug metabolism - cytochrome P450,Steroid hormone biosynthesis,Drug metabolism - other enzymes
8	Retinol metabolism,Drug metabolism - cytochrome P450
9	Retinol metabolism,Drug metabolism - cytochrome P450,Drug metabolism - other enzymes
10	Metabolism of xenobiotics by cytochrome P450
11	Metabolism of xenobiotics by cytochrome P450,Bile secretion
12	Metabolism of xenobiotics by cytochrome P450,Steroid hormone biosynthesis
13	Inflammation mediated by chemokine and cytokine signaling pathway,Endocytosis,Chemokine signaling pathway,Epithelial cell signaling in Helicobacter pylori infection
14	Oxidative phosphorylation
15	Oxidative phosphorylation,Cardiac muscle contraction,Alzheimer's disease,Parkinson's disease,Huntington's disease
16	Oxidative phosphorylation,Alzheimer's disease,Parkinson's disease,Huntington's disease
17	Alzheimer's disease
18	Alzheimer's disease,Huntington's disease
19	Parkinson's disease
20	Parkinson's disease,Huntington's disease
21	Huntington's disease
22	Fc gamma R-mediated phagocytosis,Systemic lupus erythematosus,Phagosome,Tuberculosis,Leishmaniasis,Osteoclast differentiation,Staphylococcus aureus infection
23	Regulation of actin cytoskeleton,Focal adhesion,Leukocyte transendothelial migration,Adherens junction,Tight junction
24	Olfactory transduction
25	Ribosome
26	Pentose and glucuronate interconversions,Ascorbate and aldarate metabolism
27	Steroid hormone biosynthesis
28	Drug metabolism - other enzymes
29	Chemokine signaling pathway,Epithelial cell signaling in Helicobacter pylori infection
30	Leukocyte transendothelial migration,Adherens junction,Tight junction
31	One carbon pool by folate,Formyltetrahydroformate biosynthesis
32	Folate biosynthesis

Figure S8. FISH verification and pathway enrichment analysis for gene duplications A. Verification of duplication copy number of *DIP2C* in macaques (right panel), a gene for which we inferred a diploid copy-number of 20 using fluorescence *in situ* hybridization. *DIP2C* was previously reported to be duplicated in macaques, with an inferred diploid copy-number of 12 based on arrays [4]. Humans (left panel) have a diploid *DIP2C* copy number of two. **B.** KEGG/PANTHER pathway enrichment analysis of gene duplications; values on top of each bar represent FDR-adjusted hypergeometric p-values.

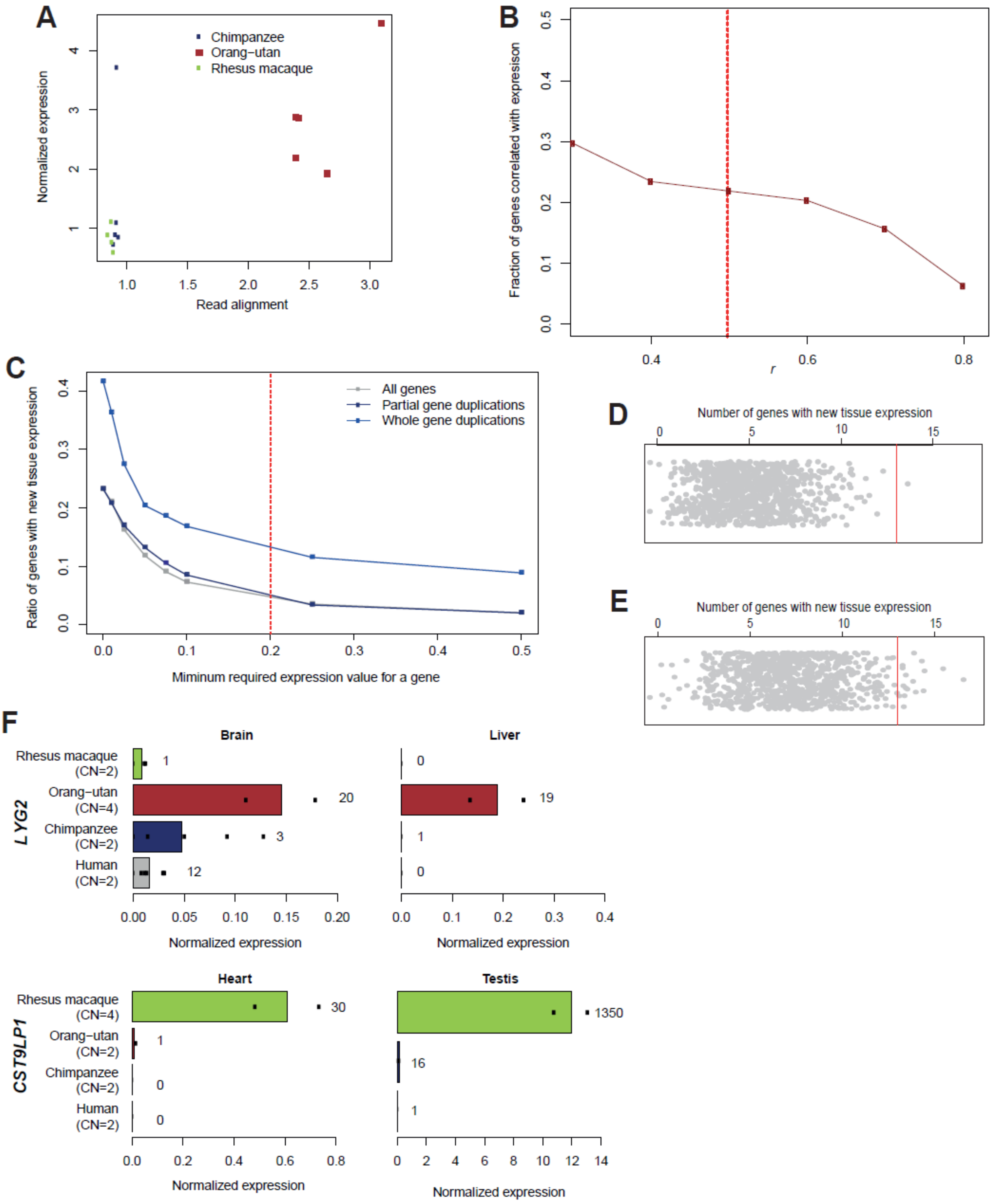


Figure S9 Impact of gene duplications on expression A. Effect of *USP31* gene duplication on expression. *USP31* shows increased expression read depth in orang-utan, where it is also gained to copy number (CN 4-5, corresponding to a mrsFAST read depth measurement of 2 – 2.5; with a mrsFAST read depth ratio RD_{MF} of 1 representing normal

diploid copy number). **B.** Correlation between DNA-seq and RNA-seq based read coverages. The x-axis depicts different thresholds for adjusted correlation (*r*) values, whereas the y-axis depicts the fraction of whole gene duplications correlating with expression at a given *r* threshold. **C.** Exploration of the parameters space to define a threshold for identification of whole gene duplications with newly acquired tissue expression. The red dashed line indicates the minimum normalized expression value for defining a gene as expressed (median expression value = 1). The expression value is defined as the number of exonic reads per gene, normalized by GC, overall read count and median expression in a sample. **D.** The observed number of whole-gene duplications coinciding with observed gene expression in a new tissue (red line) is markedly higher than expected by chance, based on 1000 Monte Carlo simulations (p=0.003; permutation-based empirical p-value). The red line depicts the original value of 13 genes with inferred gain of expression in a new tissue. Gray dots represent simulation outcomes. **E.** Whole-gene duplications show a significant enrichment for 'gains' in expression in new tissues (compared to numbers expected by chance, based on 1000 Monte Carlo simulations) when expression levels proportionally increased to gene copy number (dosage-effect) are modeled (p = 0.02; permutation-based empirical p-value). The red line depicts the original value of 13 genes with inferred gain of expression in a new tissue. Gray dots represent simulation outcomes. **F.** Barplots depicting normalized expression values of *LYG2* and *CST9LP1* for different tissues. *LYG2* is commonly expressed in brain and shows new tissue expression in orang-utan liver, whereas *CST9LP1* shows new tissue expression in rhesus macaque heart and highly increased expression in testis. The length of each bar corresponds to the mean normalized expression per tissue and species (based on multiple matching reads). Black dots represent individual normalized read counts of samples. Numbers at each bar show the mean number of reads generating the respective expression signal.

Supplemental Material and Methods

Non-human primate samples

Following the acquisition of federal (Federal Fish and Wildlife Permit, USA – Permit: MA232608-0) and institutional permissions, five samples each of fibroblast-derived cell lines from unrelated chimpanzee, orang-utan and macaque individuals were obtained from the Coriell Cell repository (**Dataset 1**).

Library preparation and sequencing

Genomic DNA library preparation was carried out using paired-end protocols according to the vendor's guideline (Illumina, Inc.). In brief, 5μg of high molecular weight genomic DNA were fragmented to 250-350bp insert size with a Covaris S2 device (Covaris, Inc.), followed by size selection through agarose gel excision and sequencing on an Illumina Hiseq2000 instrument. Sequenced reads were aligned to the respective reference genomes of each species in paired-end mode using the proprietary Illumina alignment software *ELAND*, version 2 (Illumina, Inc.). Aligned reads were merged from lane-level data to sample-level data. We converted aligned reads to the SAM/BAM format using SAMtools [8].

Human samples

To facilitate a comparison of non-human primates to human, we used sequencing data from the 1000 Genomes Project (http://www.1000genomes.org/). We aimed for the most comparable sequenced human samples to our non-human primate data and required (1) that the samples were sequenced using the Illumina sequencing platform, (2) using 101bp paired-end reads, and (3) with high sequencing coverage of ~20X. The most comparable dataset at the conception of this project were the genetically unrelated trio parents NA12891, NA12892, NA19238, NA19239, sequenced to 80X each, which we downsampled to 20X using Downsample.jar from PicardTools version 1.52 (http://picard.sourceforge.net/command-line-overview.shtml), and one additional low coverage (~5X) sample NA06894. Where applicable we integrated results from these samples with our non-human primate based data analyses.

Discovery of deletions and duplications

We discovered SVs by evaluating complementary signatures allowing SV computational inference in deep sequencing data (recently reviewed in [9]): (1) discordantly mapped paired-end reads, (2) split reads, and (3) abnormal read depth signatures. Presently available algorithms utilize one or a combination of these approaches for SV discovery. To combine different approaches and aim for the most comprehensive dataset, we used a combination of different algorithms to infer deletions and duplications in non-human primate genomes (see also **Figure S1 and Figure S2**): (1) DELLY version 0.0.4 [10], which uses paired-end mapping and split-reads to define breakpoint-resolution SV calls, (2) CNVnator version 0.2.2 [11], which is based on a read depth approach and (3) GenomeSTRiP version 1.03 [12], a population based deletion caller, which integrates read-depth and paired-read based discovery approaches. DELLY was used to detect tandem duplications and deletions. CNVnator was used for tandem and dispersed duplication discovery, as well as for deletion discovery. GenomeSTRiP and DELLY were applied with default parameters. For DELLY we required at least 2 supporting read pairs to trigger a split-read analysis in search for deletion and duplication breakpoints. CNVnator window sizes were chosen according to recommendations of the CNVnator developers, *i.e.,* by applying window sizes between 100bp and 300bp depending on the genomic read coverage of a samples, whereby the recommended window size scales inversely with genomic read coverage [11].

Filtering and merging of deletion calls from sample level to species level

Calls generated by each of the three methods were filtered and merged differently, to account for conceptual differences in the respective discovery approaches. Deletion calls made by either DELLY or CNVnator were merged from sample-level to species-level using custom scripts. In brief, we first merged DELLY and CNVnator calls separately based on our confidence in the inferred breakpoint coordinates: DELLY deletion calls with split read support have breakpoint (*i.e.,* nucleotide) resolution and thus were merged only if two calls have exactly the same start and end coordinates. By comparison, DELLY deletion calls supported by discordantly mapped paired reads were merged if they displayed intersecting confidence intervals, assuming intervals of +/- 100 bp at the breakpoints. CNVnator deletion calls based on read depth were merged assuming a confidence interval of +/- 300bp at the breakpoints, since read-depth based calls have lower boundary resolution than paired-end calls. Since GenomeSTRiP is a population based caller (*i.e.,* considers the input of multiple samples to call SVs at the population level), no merging of discovered SVs across individuals of a species was required. We considered GenomeSTRiP deletion calls passing all internal filters in our discovery set (flag "PASS"). Additionally, we included calls not identified by GenomeSTRiP, if (and ony if) they were supported by both CNVnator (read depth based analysis) and DELLY (paired-end based, or paired-end *and* split-read based analysis). We merged GenomeSTRiP calls displaying >50% reciprocal overlap with a DELLY/CNVnator call. For these merged calls, we used the start and end coordinates of the DELLY call if nucleotide resolution breakpoint information (*i.e.*, split-read support) was available. Otherwise, we used the GenomeSTRiP coordinates. To eliminate reference effects, we filtered out deletion calls observed in all 5 samples of a species. SV calls showing a >50% reciprocal overlap with reference assembly gaps were removed to ensure high quality of the deletion set. Note that this filter shows different impact on distinct species, depending on the gap content in the respective reference genome (**Figure S3B**). GenomeSTRiP was used for genotyping deletions (default parameters). Our final dataset was categorized into the discovery dataset and the breakpoint dataset (*i.e.*, SV calls with DELLY-based split read support; **Figure S2**, **Dataset 3**). We separated reference MEIs that were detected as deletions relative to the reference genome from our deletion set; these were analysed separately along with the non-reference MEIs in our MEI set. The discovered SVs with breakpoint resolution were used for assessment of SV formation mechanisms and for ancestral state determination.

Filtering and merging of duplication calls from sample level to species level

Similar to the approach applied for deletions, we initially merged DELLY and CNVnator duplication calls based on our confidence in their breakpoint coordinates using a custom approach. In brief, DELLY duplication calls with split read support (nucleotide resolution) were merged if two calls displayed the same start and end coordinates. DELLY duplication calls supported by discordantly mapped paired reads were merged if they displayed intersecting confidence intervals at the breakpoints (+/- 100 bp). CNVnator duplication calls were merged assuming confidence intervals of +/- 300bp at the breakpoints. We further verified our initial duplication calls independently using the read depth based copy-number genotype assessment algorithm CopySeq version 1.7.1 [13], using default parameters. For this purpose, GC content and mappability maps, as well as the genomic intervals for variance model parameter estimation were generated for each reference genome assembly (panTro3, ponAbe2 and rheMac2) as described in [13]. Using these parameters, we applied CopySeq to our merged duplication discovery set (based on DELLY and CNVnator), and further filtered the duplication calls based on the adjusted read-depth ratio reported by CopySeq (in the following referred to as 'CopySeq read depth ratio RD_{CS}'; see [13] for details). Specifically, we required at least one sample with a CopySeq read depth ratio RD_{CS} > 1.3, which is indicative of a markedly elevated read depth in the interrogated region (note: for a normal diploid copy, the expected CopySeq read depth ratio RD_{CS} is ~1). These interim duplication calls were further filtered for gaps (50% reciprocal overlap) to ensure high callset quality. Using LOWESS we normalized CopySeq read depth ratios RD_{CS} across samples. Following normalization, suspected duplication regions no longer having at least one sample with CopySeq read depth ratio RD_{CS} >=1.3 were removed.

Validation of structural variants

Deletion and MEI calls were validated by PCR, using individuals that we inferred to lack the DNA variant in question as controls. As a sanity check, we further verified the absence of predicted novel MEIs in orthologous regions of other primate species (**Figure S4C**). Primers can be found in **Dataset 2**.

In addition to PCR we conducted custom high-density (>9 Million probe-based) tiling aCGH experiments for each non-human primate species. The aCGH probes, designed based on the reference genomes of each species, comprised sequences uniquely mapping to each reference genomes (with the probes being as uniformly distributed as feasible, given repeat and gap content). For each species, aCGH reference and sample were chosen randomly (**Dataset 1**). Genomic DNA for both reference and sample was sheared by sonication (10-second bursts at 100 amps for 1 min) and labeled with Cy3 (reference) or Cy5 (sample). Labeling reactions were cleaned by column filtration and then pre-hybridized to cot-1 DNA (Life Technologies). Afterwards, both sample and reference labeled DNAs were co-hybridized to the nine 1-million probe arrays for autosomes and to the 400K probe array for the sex chromosomes. Hybridizations occurred at 60°C for 40 hours with constant rotation of arrays. Arrays were then washed and scanned at 3 μm resolution. Fluorescent intensities at each DNA spot on the arrays were determined using Agilent Feature Extraction software (version 10.5), which also performs background and dye-bias corrections to normalize the intensity values. Only arrays having a derivative of $\log_2$ ratios (DLR) of 0.3 or lower, indicating a high quality array run with tolerable background noise, were used for further analysis.

When assessing deletions and duplications inferred by our deep sequencing based pipeline (**Figure S1** and **Figure S2**) by aCGH, we required a minimum overlap of 3 probes per call, for which the median $\log_2$ ratio of intensities from sample to reference (*i.e.,* $\log_2$(sample intensity/reference intensity)) was calculated. FDR was computed by considering assessable calls based on these requirements and applying a $\log_2$ ratio cutoff of ≤ -0.1 or ≥ 0.1. We estimated FDRs of 13.48% (60/445 calls) in chimpanzee, 9.70% (40/412 calls) in orang-utan and 7.38% (79/1070 calls) for genotyped deletions in rhesus macaque.

We further made use of aCGH to identify suitable cut-offs for inferring duplications as fixed or variable. Specifically, we assumed that a variable duplication in a species would yield aCGH $\log_2$ ratio deviating from 0 when different individuals from that species are compared by aCGH. For instance, a gain in the sample relative to the reference would be expected to result in aCGH $\log_2 \geq 0.1$, whereas a gain in the reference relative to the sample would be expected to result in ≤ -0.1. Likewise, sequencing based read depth $\log_2$ ratios can be used to infer variable duplications. We calculated the ratio of sequencing read depth between sample and reference as follows:

$$RD_{seq} = log_2 \ \frac{adjRD_{sample}}{adjRD_{reference}}$$

adjRD = CopySeq read depth ratio RD_{CS}

To determine which duplications were variable or fixed, we examined different RD_{seq} cutoffs of sample versus reference for concordance with aCGH. For each species, we determined RD_{seq} cutoffs yielding >80% concordance with aCGH (*i.e.* the RD_{seq} cutoff where 80% of all assessable duplications were confirmed by aCGH) – with ≥0.1 for gains in the sample versus the reference and ≤ -0.1 for gains in the reference versus the sample. The RD_{seq} cutoff yielding 80% concordance of variable duplications with aCGH was 0.7 for chimpanzee, 0.6 for orang-utan and 0.8 for rhesus macaque.

Next, we calculated the RD_{seq} for all samples of a given species (not just the array sample and reference). If each pairwise comparison yielded a RD_{seq} below the species-specific cutoff, the duplication was considered as “fixed”. If at least one of the pairwise comparisons of RD_{seq} of all 5 samples was above the species-specific cutoff, the duplication was considered as “variable”.

Validation of fixed duplications

To verify inferred fixed duplications, we assessed their concordance with previously published cross-species aCGH experiments (here denoted “common arrays” when referring to common aCGH probes aligning perfectly to the human, chimp and macaque reference genomes) [21]. For this purpose, fixed duplication calls were lifted over to hg18 using the LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). We required at least 3 array probes to overlap a duplication call (intersection by at least 1 bp). The duplication was considered to be validated in a species if all samples of a species had an average $\log_2$ ratio of intensities ≥ 0.1. By this means, 81.5% (53/65 calls), 84.3% (43/51 calls) and 82.3% (14/17 calls) of fixed duplications were verified in chimpanzee, orang-utan, and rhesus macaque, respectively.

Additionally we applied the mapping algorithm mrsFAST (version 2.3.0.2, using default parameters and requiring >94% identity with the reference genome for a read to map) on all inferred fixed duplication calls, to verify their fixation with an independent approach. mrsFAST aligns reads to all possible mapping locations; hence fixed duplications would show an elevated read depth in all samples of a species. Using a cutoff of RD_{MS} > 1.3, we could confirm ~80 % of our predicted fixed duplications by this independent approach.

SV formation mechanism assignment and in depth analysis of mechanisms

We used BreakSeq [6,14] (version 1.3 with default parameters) to infer formation mechanisms for deletions and duplications mapped with nucleotide resolution breakpoints, and used the same approach to infer mechanisms of duplications formation. We were able to map on average ~ 51% of all deletions and ~ 18% of all duplications at breakpoint resolution (**Dataset 3**). We observed an excess of NAHR-mediated SVs in the great apes, with the size of NAHR-mediated SVs being larger than the size of NHR-mediated SVs on average. The mean sizes of NAHR events were 16.5kb in chimpanzee, 7.4kb in orang-utan, and 11.3kb in rhesus macaque. For NHR we observed mean sizes of 7.8kb in chimpanzee, 5.7kb in orang-utan, and 3.1kb in rhesus macaque. We used a Monte Carlo simulation based approach to assess whether NAHR events contained significantly more basepairs, *i.e.* were significantly larger than expected based on random mechanism assignments. The total amount of genomic sequence occupied by NAHR and NHR events was 11.6Mb and 5.8Mb in chimpanzee, 12.7Mb and 8.9Mb in orang-utan and 4.4Mb and 6.6Mb in rhesus macaque, respectively, reflecting the marked excess of NAHR events in the great apes. We performed

1000 permutations in each primate species, by keeping SV size assignments constant and permuting the mechanism assignments. In each iteration, we calculated the total amount of genomic sequence occupied by randomly-assigned NAHR-labeled and NHR-labeled SVs. NAHR mediated SVs were larger than SVs formed by other mechanisms in all three primate species ($p<0.001$, $p=0.037$ and $p<0.001$ in chimpanzee, orang-utan, and rhesus macaque, respectively; empirically calculated p-values based on permutation). By comparison, NHR-mediated SVs did not display a trend towards larger SVs ($p=0.41$, $p=0.64$ and $p=0.99$ in chimpanzee, orang-utan, and rhesus macaque, respectively; empirically calculated p-values).

We used a simulation based approach to assess whether the 5-fold increase of NAHR-mediated SVs in the great apes can explain the increased impact of SVs at the genomic sequence level in this lineage compared to the rhesus macaque. Performing 1000 Monte Carlo simulations in chimpanzee and orang-utan, we randomly picked 20% out of all NAHR-mediated SVs (to mimic the 5-fold difference between great apes and rhesus macaque) and calculated the total size of sequence occupied. Indeed, when considering only 20% of all NAHR-mediated SVs, both chimpanzee and orang-utan displayed a smaller genomic impact of NAHR-mediated SVs than rhesus macaque ($p<0.005$; permutation-based empirical p-value), with an average of 2.3Mb and 2.5Mb of sequence occupied in chimpanzee and orang-utan, respectively, compared to the 4.4 Mb that are occupied by NAHR-mediated SVs in the macaque. We hence concluded that the 5-fold increase in the activity of NAHR formation in the great apes compared to rhesus macaque significantly contributed to the markedly increased nucleotide-level impact of SVs in the great apes.

Ancestral state inference

We determined the ancestral state of SVs using the BreakSeq package [14]. For deletions or duplications as identified relative to the respective reference genome, two different alleles were taken into account for ancestral state determination: (1) the reference allele, for which +/- 500bp flanking sequences were extracted at each breakpoint representing both left and right reference junction sequences; (2) the alternative allele, for which also +/- 500bp breakpoint flanking sequences were extracted (see [14]). The respective junction sequences were extracted from each species and were aligned to the genomes of the other species (*e.g.*, rhesus macaque junction sequences (query species) were aligned on the marmoset (calJac3), orang-utan (ponAbe2), chimpanzee (panTro3) and human (hg19) reference genomes, and so forth). The alignment was performed using BLAT [15] on the syntenic regions of the corresponding SV (top levels of the Net alignments downloaded from UCSC (http://genome.ucsc.edu/cgi-bin/hgGateway) for each species). For example, when assessing SVs that were inferred as deletions by our SV discovery pipeline, if the alternative junction sequence from one species mapped with better sequence identity and length (compared with the reference junction sequence) onto one of the four corresponding syntenic regions, the event was rectified as "insertion"; if the reference junction sequences mapped better than alternative junction sequences, the event was rectified as "deletion" (see [14] for details). Events were "unrectifiable", if we failed to identify an alignment between the junction sequences obtained from the query species and the corresponding syntenic regions from the other species. To address the origin of duplicative insertions, we focused on those deletion and duplication calls that were rectified as insertions, indicating that an insertion into ancestral genomic sequence, rather than a sequence deletion, has occurred. The respective sequences were subjected to BLAT analysis to determine the donor locus (**Figure S3**).

Fluorescent *in situ* hybridization (FISH)

To perform FISH, metaphases were prepared using primate fibroblast and human β-lymphocyte cell lines (Coriell) by incubation with colcemid (0.08µg/mL) for three hours. Cells were placed in hypotonic solution (0.075M KCl) at 37 ºC for 20 minutes, then fixed three times in 3:1 methanol:acetic acid. The fixed cells were dropped onto slides, baked for one hour at 55 ºC, and aged 1 day at room temperature. The slides were stored at -20 ºC until hybridization. BAC clones were extracted according the Large-Construct Kit protocol (Qiagen). The DNA was then labeled with SpectrumRed-dUTP (Abbott Molecular) using the nick translation kit (Roche Pharmaceutical), with the following modifications: the labeled probe was precipitated using human Cot-1 DNA only, and it was resuspended in hybrisol (50% formamide, 2X SSC, 10% dextran sulfate) to 0.05µg/µL. Fibroblast slides were pre-treated with Digest-All (Invitrogen) under Parafilm for 5 minutes at 37 ºC on a Hybrite (Thermo-Fisher) and washed twice in PBS. All slides were dehydrated for two minutes in 70%, 90%, and 100% ethanol. Probes were added to selected regions of the slide, coverslipped, and sealed with rubber cement. The slides were denatured at 73 ºC for 2 minutes and hybridized at 37 ºC overnight in a humidified chamber. Coverslips were removed, and the slides were washed twice in 50% formamide/2X SSC at 42 ºC for 7 minutes, in 2X SSC at 42 ºC for 5 minutes, and for 3 minutes at 25 ºC 1X PBS with 0.1% IGEPAL. Slides were mounted in DAPI II (Abbott Molecular) and sealed with nail polish. The slides were then viewed and imaged using Olympus BX51 fluorescent microscope with appropriate filters and Cytovision 3.6 software (Applied Imaging).

Estimation of SV calling accuracy and error rates

As a further means of assessing variant calling accuracy and the quality of our SV callsets, we performed additional quality control analyses. First, we investigated the read depth-of-coverage in deletion regions by extracting reads from samples inferred to have a normal (disomic) copy number (based on genotyping) and calculated the coverage for those regions using "samtools view" and "samtools depth". In all species investigated the read depth-of-coverage for deletion regions was similar to the genome-wide depth-of-coverage (**Figure S4B**). The slight reduction in depth by ~20% we observed is mirrored in 1000 Genomes Project pilot project samples [6] and likely reflects the known association of SVs with repeat-rich regions leading to reduced read mappability [13].

Second, we inferred the number of heterozygous and homozygous SNPs within the boundaries of deletions separately for each sample. We observed that the number of heterozygous SNPs (SNP genotype 0/1) in samples genotyped as 'homozygous reference allele' (SV genotype 0/0; *i.e.* 'no deletion') is approximately twice as high as homozygous SNP calls (SNP genotype 1/1), whereas for heterozygous deletions (SV genotype 0/1) most SNPs were homozygous (**Figure S4A**) – as would be expected for true deletion sites. Altogether 82%, 90% and 88% of the genotyped heterozygous deletions in chimpanzee, orang-utan, and macaque, respectively, showed an excess of homozygous over heterozygous SNPs. (Note that owing to remaining uncertainties in SV boundary coordinates, few remaining heterozygous SNPs (*i.e.*, SNP genotype '0/1') were expected to occur even within accurately genotyped heterozygous deletions).

Third, using published SVs in humans [6], we assessed the errors that may have been introduced (*i*) by sampling 5 individuals from a given primate species, and (*ii*) by our SV calling and filtering approach. We merged 1000 Genomes Project pilot phase deletions (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/paper_data_sets/companion_papers/mapping_structural_variation/MasterValidation.Pilot1.deletion.leftmost.061510a_mergedValPlus.txt) and SV formation mechanism assignments (see Supplementary Table 11 of [6]) using a 50% reciprocal overlap criterion. 5 human samples were randomly selected 10 times, and published CNVnator calls [6] were extracted for each sample. Since, GenomeSTRiP, unlike CNVnator, is sensitive to number of samples used as input, we generated GenomeSTRiP calls for each of the 10 sets of 5 randomly selected individuals. As DELLY was not used in the pilot phase of the 1000 Genomes Project, we additionally generated DELLY calls for each randomly sampled sets, and merged these with the CNVnator and GenomeSTRiP calls using the same criteria and cutoffs that we used for generating non-human primate SV sets. We estimated the standard deviation for sampling 5 individuals of a population as ~19% (based on the number of SVs called in each resampling experiment). Importantly, when proportions of inferred SV formation mechanisms were considered, the observed standard deviations were low in relation to the primate inter-species differences in SV formation mechanism activity that we inferred for several key formation mechanisms: *i.e.*, for NAHR and MEI displaying pronounced inter-species differences (see main text), the proportion of reference MEI-associated SVs varied (due to sampling) between 0.085 – 0.103 (mean = 0.0966; standard deviation = 0.0072), and the proportion of inferred NAHR-associated deletions varied between 0.138-0.162 (mean = 0.150; standard deviation= 0.0084). These variations due to sampling were substantially lower than observed inter-species differences in SV formation mechanisms, *i.e.*, 0.35-0.44 in great apes vs. 0.86 in rhesus macaques for MEIs, and 0.28 in great apes vs. 0.02 in rhesus macaques for NAHR – which suggests that our findings of inter-species differences in these mechanisms are highly robust.

We note that the proportions of different mechanisms are slightly different to what is observed based on all genotyped deletions in the 1000 Genomes Project pilot phase [6]. Differences in proportions were, however, generally <5%, and are explainable by the subset of SV calling algorithms that we applied (and their association with SVs of a particular size range [6]) out of the whole panel of SV calling algorithms used in the 1000 Genomes Project pilot phase. It is also of note that since human individuals in the 1000 Genomes Project pilot phase were sequenced with a comparably low coverage (*i.e.*, ~5-fold, vs. 15-20-fold in our primate study), we expect a relative decrease in the numbers of false negative SVs missed in each sample in our non-human primate based study. Hence, the re-sampling based assessments we performed should provide a conservative estimate for the margin of error (*i.e.*, the margin of error would be expected to be comparably reduced, rather than comparably increased, in sequencing data generated at higher genomic coverage).

To further assess the accuracy for our callset, we also realigned reads from all 5 chimpanzee individuals to the human reference genome assembly (hg19) using BWA with default parameters and pursued SV (deletion and duplication) calling on these alignments. We applied the same SV calling algorithms as applied to the non-human primate reference genome assemblies on chimpanzee reads aligned to hg19 and merged calls from sample to species level as for non-human primates (**Figure S2**). As sequencing gaps and sequence divergence between species have an influence on the SV calling, we generated a stringent high-quality lifted-over SV dataset by filtering out SVs intersecting with sequencing gaps in panTro3, and requiring 95% of bases to remap in the lift-over process. Applying these criteria, 671 and 1776 deletions and duplications were lifted over successfully to hg19. (Note that proportionally more duplications than deletions were successfully lifted over, since deletions are on average comparably larger, and hence more often interspersed by small sequencing gaps.)

When using a 50% overlap criterion, 60% of all SVs (514/671 deletions and 937/1776 duplications) that were lifted over from panTro3 to hg19 could be re-identified in humans based on chimp reads aligned to hg19. Widely used Illumina read mappers (BWA or ELAND) were not designed to map DNA reads from a given species (chimpanzee) onto a distinct reference (human) – with sequence divergence impeding with read mapping. As such, we also used custom read-depth and paired-end analysis approaches to partially recover SVs that failed to be recalled due to DNA read alignment issues. First, we extracted DNA reads within each SV region and within ranges of 2000bp on either side of the SV region, ranges of sufficient length for read depth analysis [13]. If the read depth ratio within the SV region, compared to its vicinity, was reduced by at least 0.3 reads/bp (or increased by at least 0.3 reads/bp for duplications, respectively), we considered calls as verified – a cutoff consistent with the cutoff we applied for CopySeq predicted duplications. We additionally considered single instances of discordant paired-end reads (as defined through DELLY) among the hg19-aligned reads to verify deletion and duplication calls made in the chimpanzee (thereby accounting for potentially missed calls due to the more sparse chimp reads in the human-based alignments). Using these custom approaches we were able to recall in total 82% of the original chimpanzee calls (605/671 (90%) deletions and 1410/1776 (79%) duplications) based on alignments of chimp DNA reads to the human reference genome assembly.

SNP analysis

SNPs were identified using the Genome Analysis Toolkit (GATK, [34]) and Samtools [8]. We subsequently applied GATK base quality score recalibration and realignment, and performed SNP discovery and genotyping across all samples simultaneously using standard hard filtering parameters. The consensus of multiple primary callsets from GATK and Samtools was used for further analysis. In human validation tests we obtained >90% specificity and sensitivity with this approach (see below). For each sample, a series of filters were applied to remove potential false positives. First, we removed candidate SNPs mapping to gaps in the reference. Second, SNPs with a Phred quality score 10 or below were discarded. Third, SNPs intersecting with segmental duplications (see “Assembly-based segmental duplications”) were excluded. Finally, we did not allow for SNPs within 10 bp of each other, in order to minimize the rate of false positives caused by recent segmental duplications. For orang-utan and rhesus macaque, respectively, we also removed those SNPs falling into regions in the reference genome with low consensus quality score <90 (on a scale of 1-97, based on the Phred scores of underlying whole-genome shotgun reads) and <60 (on a scale of 1-60), respectively.

We tested the sensitivity and specificity of our SNP calling pipeline by comparison to recently released SNP calls by the 1000 Genomes Project, which were identified by the Illumina Omni platform, using the CEU (European ancestry) trio samples NA12891 and NA12892. These samples were sequenced to high coverage (~80X) by the 1000 Genomes Project. The Omni genotypes and the CEU trio Bam files were downloaded from the 1000 Genomes Project webpage[1]. Omni genotypes from samples NA12891 and NA12892 were extracted using ‘SelectVariant’ from GATK. The original Bam files of NA12891 and NA12892 were downsampled to 20X and 5X coverage by DownsampleSam.jar from the PicardTools package version 1.57 (http://picard.sourceforge.net/command-line-overview.shtml). We then called SNPs with the same pipeline as applied to non-human primate SNP calling, using 80X, 20X, and 5X coverage data, respectively. To establish an evaluation on sensitivity and specificity, we calculated false positive (FP), false negative (FN), true positive (TP) and true negative (TN) rates for each coverage depth using Omni genotypes as a golden standard and calculated sensitivity and specificity at the level of Omni genotypes, using sensitivity = TP / (TP + FN), and specificity = TN / (FP + TN). Using Omni genotypes as a gold standard, we obtained >90% sensitivity and specificity for 20X and 80X sequencing coverage data. By comparison, using 5x coverage data we obtained similarly high specificity whereas the sensitivity was ~80%. We hence concluded that a sequencing coverage ~20X allows for high SNP calling sensitivity and specificity, using our dedicated SNP calling pipeline.

In addition, we amplified132 and 60 regions around randomly selected SNPs and indel polymorphisms, respectively (**Dataset 2**). By Sanger sequencing of these amplified fragments, we validated 238 out of 241 SNPs, as well as 42 out of 43 indels we were able to interrogate. These results gave us false discovery rates (FDRs) of 1.2% and 2.3% for SNPs and indels, respectively. We further calculated an false negative rate of 7% for SNPs, by determining the number of heterozygous SNPs that we detected by Sanger sequencing, but

[1] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110921_phase2_omni_genotypes/ ,ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110915_CEUtrio_b37_decoy_alignment/

unable to detect by our next generation sequencing based pipelines. Please note that since we required very stringent criteria for detection of indels, we expect our indel dataset to be highly accurate as supported by low FDR, but not sensitive.

Calculating pairwise differences in nucleotides for SNPs and SVs

Using SNP genotypes we generated from the non-human primate samples, we calculated the average number of SNPs differing between any two samples of a species, using pair-wise comparisons. If an SV was found in only one of the two samples, then the size of the SV was added to the estimate of nucleotide differences between the two samples. The average of all pairwise comparisons was computed for each species (**Figure S7C).** MEIs inferred using the Tea pipeline were aligned to a library of subfamily consensus sequences for *Alu*, L1, and SVA. The best scoring alignment was chosen and used to estimate the size of the MEI. Since not all of the inferred non-reference MEIs have reconstructed sequences comprising the entire mobile element, we used the average sequence length for a given subfamily within each species (for example, we averaged the lengths for all novel *Alu* insertions within chimpanzee). We then used these averaged lengths for our estimate on the number of nucleotides differing between pairs of samples within each species.

Novelty of variant calls

We extracted primate SNPs from dbSNP (version 129) and compared these known SNPs to our dataset. Merged SV calls (high confidence discovery dataset, deletions and duplications) were compared to published SV calls from aCGH experiments. Calls were compiled from dbVar [16,17,18] and signature papers [19,20,21] and converted to the respective reference genomes panTro3, ponAbe2 and rheMac2 using the liftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The overlap cutoff was set to a minimum of 1bp between known SV and novel SV from our datasets.

Functional annotation of SNPs

SNP annotation was performed using ANNOVAR [22] . To prepare the annotation database, we downloaded the non-human primate Ensembl genes from the UCSC genome browser (ftp://hgdownload.cse.ucsc.edu/). SNPs were categorized into exonic, splicing, intronic, ncRNA (exonic, intronic, and splicing), 5′-untranslated regions, 3′-untranslated regions, upstream, downstream and intergenic regions (**Dataset 3**).

Population genetics analysis

We applied the Approximate Bayesian Computation (ABC, [23]) approach to select between (i) a model in which all individuals are potential partners and can mate randomly (panmictic population), and (ii) a model where individuals belong to separate isolated sub-populations (there would be no random mating between individuals from different sub-populations). The evolutionary scenarios (i) and (ii) were implemented with msABC [24]. Each chromosome was split in segments of 10kB each. For model (ii), we used a uniform prior for *t* on the logarithmic scale. Arbitrary lower and upper boundaries were set as 0.001 and 2.0 coalescent units (one coalescent unit corresponds to 4 N generations), respectively. For model (i) (panmictic model), time *t*=0, that is coalescence of chromosomes from different individuals was allowed at any time point.

We performed 350,000 simulations for each model. Model choice was accomplished by the 'abc' package in R [25]. Regression was accomplished with the "mnlogistic" method and tolerance 0.01, i.e., 3,500 simulation instances were kept from each model for inference. For all four species, the average posterior probability of the non-random mating model was higher than the average posterior probability of the random mating model. The posterior probability values for the non-random mating model were 1.0, 0.85, 0.96, 0.86 for human, chimpanzee, orang-utan and rhesus macaque, respectively. Furthermore, the average estimated median value of *t* for humans was 0.073, for chimpanzee 0.0064, for orang-utan 0.0042 and for rhesus macaque 0.0096. Notably, the *t* value for the X chromosome of human (0.83) and chimpanzee (0.08) were considerably higher than the average *t* value over the whole genome. The higher average posterior probability of the non-random mating model and the positive time t in all four primates imply that the non-random mating model fits the data better than the random mating model, suggesting that each primate population we study is sub-structured (**Figure S6BC**).

We additionally used population genetics based analyses to test whether the primate individuals used in this study are relatives. We tested a scenario in which individuals of each species are relatives between themselves (*i.e.,* non-random sampling scenario). Under this scenario, a hypothesis of a recent founder event or a population bottleneck will fit the coancestry of individuals, since individuals will originate from a single common ancestor in the recent past. Using the ABC methodology we excluded the population bottleneck scenario. Thus, there is no evidence that individuals are relatives, leading us to be able to exclude kinship.

Inference of SNP and SV mutation rates

We inferred the population mutation rate ($\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per site per generation) and the population recombination rate ($\rho = 4N_e r$, *r* is the recombination rate per base per generation) using the ABC methodology described above. We tested whether the recombination rate and the mutation rate are constant or variable along the chromosomes. Since data are unphased, we adapted msABC to simulate unphased data.

We used msABC to generate two simulated datasets: (1) mutation and recombination rates follow uniform distributions $U(\mu_{min}, \mu_{max})$ and $U(r_{min}, r_{max})$, respectively, and (2) mutation and recombination rates are constant on each chromosome. Performing 230,000 simulations, we found that the posterior probability of the variable mutation and recombination rate model is higher than the constant rate model. For all species, the posterior probability for the model with variable mutation and recombination rate is 1.0.

The average population mutation and recombination rates of autosomes for human are 4.9 x 10^{-4} per bp and 4.2 x 10^{-4} per bp, for chimpanzee 6.8 x 10^{-4} per bp and 4.0 x 10^{-4} per bp, for orang-utan 1.15 x 10^{-3} per bp and 2.1 x 10^{-4} per bp and for rhesus macaque 1.29 x 10^{-3} per bp and 3.7 x 10^{-4} per bp, respectively. The population mutation and recombination rates are considerably lower for the X chromosome for all species. For humans the population mutation rate for the X chromosome is estimated as 9 x 10^{-5} per bp (CI: 8.1x10^{-5} per bp, 1 x 10^{-4} per bp) and the population recombination rate is 3 x 10^{-8} per bp (CI: 1.5 x 10^{-12} per bp, 1.8x10^{-5} per bp). For chimpanzee mutation and recombination rates are 4.6 x 10^{-5} per bp (CI: 4.4 x 10^{-5} per bp, 5.2 x 10^{-5} per bp) and 1 x 10^{-6} per bp (CI: 4.5 x 10^{-8} per bp, 9.6 x 10^{-6} per bp), respectively. For orang-utans mutation and recombination rates are 3.9 x 10^{-4} per bp (CI: 2.9 x 10^{-4} per bp, 5.2 x 10^{-4} per bp) and 2.51 x 10^{-7} per bp (CI: 7.3 x 10^{-8} per bp, 5.2 x 10^{-7} per bp), respectively. Finally, for rhesus macaque mutation and recombination rates are 6.3 x 10^{-4} per bp (CI: 6 x 10^{-4} per bp, 6.6 x 10^{-4} per bp) and 3.8 x 10^{-6} per bp (CI: 6.6 x 10^{-7} per bp, 1.1 x 10^{-5} per bp), respectively.

Based on estimates of average values of θ computed for autosomes and reported effective population sizes N_e for chimpanzee [26], orang-utan [5], and rhesus macaque [27], we calculated μ as follows:

$$\mu_{SNP} = \frac{\theta}{4N_e}$$

We obtained average species mutation rates (per base pair per generation) of $1.5x10^{-8}$ for chimpanzee, $7.6x10^{-9}$ for orang-utan and $4x10^{-9}$ for rhesus macaque. These rates are in the range of previously published results [27,28,29]. Based on average numbers of SNPs and SVs inferred to be formed by different formation mechanisms (NAHR, NHR, *Alu* and LINE/L1), we estimated the average SV formation rate as:

$$\mu_{SV} = \frac{n_{SV}}{n_{SNP}} \cdot \mu_{SNP}$$

n_{SV} = number of SVs per species

n_{SNP} = number of SNPs per species

Reference assembly-based segmental duplication maps

To construct comparable segmental duplication (SD) sets for chimpanzee (panTro3), orang-utan (ponAbe2) and rhesus macaque (rheMac2) we applied the following approach to each non-human primate reference genome: High-copy repeats annotated in the UCSC RepeatMasker table (http://hgdownload.soe.ucsc.edu/) were initially removed from each reference assembly (to avoid that they affect the seeding of SD regions). Subsequently, SDs were identified by aligning each chromosome with itself (to detect intrachromosomal duplications) and to all other chromosomes (interchromosomal analysis). Specifically, maximal exact matches (MEMs) of minimal length 17 bp were computed using the vmatch software (http://www.vmatch.de). MEMs were then connected using CHAINER [30] with the following parameters: -length 34 -gc 100 -lw 8. To obtain extended chains, the resulting segments were used as an input for CHAINER for a second time, this time using the parameters: -length 100 -gc 1000 -lw 14 (this recursive chaining strategy is outlined in [31]). Afterwards, high-copy repeat sequences were re-inserted into the resultant chains. Chains smaller than 1000 bp were discarded. The remaining chains were globally aligned using either stretcher or needle from the EMBOSS package [32], depending on the chain size (*i.e.*, when the product of the sequence lengths was greater than 100 Mb stretcher was used, otherwise needle was used). Alignments of smaller than 90% identity, or a gap percentage larger than 30% were discarded. Application of this approach resulted in 20,680, 14,520, 20,379 and 4,742 intrachromosomal SDs and 17,423, 11,406, 5,021 and 23,216 interchromosomal SDs in human, chimpanzee, orang-utan, and macaque respectively – with both the overall numbers and occupied nucleotide basepairs being comparable to previously published results in humans [33]. SD calls were elevated in the great apes, with 36,943 (5.4%) and 14,831 (4.7%) in orang-utan and chimpanzee compared to macaques (7,295; 1.6%). Global distribution and size spectra of SDs are depicted in **Figure S3C** and **Figure S3D**.

Indel discovery

Indels were discovered using three different algorithms: GATK [34], SAMtools [8] and Pindel [35]. We applied several filters to remove potential false positives: We filtered out raw Pindel predicted indels supported by only one read. For GATK and SAMtools indel calls, we required the Phred scaled quality to be ≥ 10. Furthermore, we removed indels in poorly mapped regions (defined as regions with the GATK filtering flag "HARD_TO_VALIDATE"; those regions are characterized by 4 or more aligned reads having a mapping quality of 0 and the number of aligned reads with mapping quality 0 are more than a tenth of all alignments), in segmental duplications, and in indel clusters (where indels were identified within 10 bp of each other). After initial filtering, we extracted the consensus indel calls from these 3 algorithms, only considering these for further analyses. We further removed indels in regions where the reference genome quality was lower than 90 in orang-utan or lower than 60 in rhesus macaque.

Non-reference mobile element insertion (MEI) discovery

We used the Tea pipeline [36] to perform non-reference MEI discovery. The repeat sequence library required by Tea was constructed by concatenating multiple consensus subfamily sequences separated by multiple 'N' nucleotide spacers. To represent L1/LINE elements, consensus subfamily sequences for L1HS, L1PA3, L1PA5, L1Pt were used; for *Alu* elements, consensus subfamily sequences for AluJb, AluSx, AluY, AluMacYa3, AluYe5a2_Pongo, AluYc1a5_Pongo, and AluYe5b5_Pongo [37] were used; for SVA elements, the sequences of six SVA subfamilies (SVA_A/B/C/D/E/F) and of the general SVA consensus sequence were used.

Candidate insertion sites were considered as high-confidence if they satisfied the following criteria: (1) more than three supporting repeat-anchored mate (RAM) reads were observed, and at least one RAM on each side of the insertion was observed; (2) at least one positive and negative strand soft-clipped read was observed within the RAM cluster boundary; (3) the gap between two insertion breakpoints defined by negative and positive strand clipped reads was within [-20, 50]; (4) the ratio of well-aligned clipped reads over all clipped reads was at least 0.5. Insertion loci within 500 bp margin from the instances of the same mobile element family annotated in the reference genome were removed. Mobile element insertions located in gapped regions of the reference genome were annotated as such and removed from the final data set. Following their discovery in individual samples, we merged non-reference *Alu*, L1 and SVA insertions across samples. Our list of non-reference MEIs was merged with our list of reference MEIs (mobile element insertions identified as a deletion relative to the respective reference genome) for pursuing SV formation mechanism analyses.

Functional impact of structural variation at different scales

To study the functional impact of deletions, duplications, mobile element insertions (MEIs), SNPs, and indels we investigated the intersection of these different variation classes with protein-coding sequences. Using a 1bp overlap criterion, the number of genes with protein-coding sequences affected by SVs was 443 (2.5%), 368 (1.8%) and 318 (1.5%) in chimpanzee, orang-utan and rhesus macaque, respectively (p=0.012; Fisher's exact test). Out of these, 138 and 113 SVs in chimpanzee and orang-utan versus only 59 SVs in rhesus macaque were inferred to be formed by NAHR (see **Dataset 3**), which suggests a link between the proportionally low rate of NAHR in rhesus macaque versus great apes and the functional impact of SVs.

To assess whether there is an enrichment or depletion of functional variation types, we performed 1,000 iterations in which we shuffled the observed genetic variant calls along chromosomes and determined their overlap with genomic elements (**Figure S5**). For this purpose, we extracted non-human primate whole gene, exon, intron, 5'UTR and 3'UTR information based on annotations from Ensembl BioMart (version 62; April 2011). Non-coding RNAs were defined as lincRNAs and microRNAs. Human lincRNAs were collected from [7] and lifted to hg19 using LiftOver. Lifted human lincRNAs were combined with human lincRNAs from Ensembl BioMart (version 71, April 2013). LincRNAs intersecting with a 50% reciprocal overlap were merged. Combined datasets of lincRNAs were lifted over to panTro3, ponAbe2 and rheMac2 (http://genome.ucsc.edu/cgi-bin/hgLiftOver). MicroRNAs for panTro3, ponAbe2 and rheMac2 were obtained from Ensembl BioMart (version 71, April 2013). Human promoter coordinates were obtained from MPromDB (http://mpromdb.wistar.upenn.edu/; obtained May 2013) and lifted over to panTro3, ponAbe2 and rheMac2 using the UCSC liftOver tool.

Inference of inter-species gene duplications

To characterize gene duplications, we assigned read depth values to all genes in the primate species studied, using two independent orthogonal approaches. We applied the mrsFAST read mapping approach [38] to compute aggregate read depth ratios for each repeat-masked set of genes in each species (including in non-unique, recently duplicated genomic regions). The mrsFAST read depth ratio 'RD_{MF}' for each gene is defined as:

$$RD_{MF} = \frac{basepairs_{mapped}}{unmasked\_basepairs_{gene} \cdot genome\_wide\_coverage}$$

genome_wide_coverage = sequencing coverage based on ELAND aligned reads

We additionally applied the CopySeq algorithm in conjunction with ELAND based read alignments to unique genomic regions to compute paralog-specific read depth ratios [13]. The CopySeq based read depth ratio 'RD_{CS} measures the normalized depth of uniquely mapping sequence reads in a specific locus (see [13] for details).

We computed mrsFAST and CopySeq based read depth ratios for each gene represented in eggNOG, version 3, occurring in primates (prNOG) [39]. Inter-species gene duplications were inferred if, and only if, (1) a minimum mrsFAST read depth ratio RD_{MF} of > 1.3 was observed in this species, *i.e.,* the individual read depth ratio of each sample in this species is > 1.3, with this cutoff being selected based on concordance with recently developed cross-species arrays (≥ 75% concordance for whole gene duplications) [21]; (2) in at least one other species no duplication of the orthologous gene locus was observed; (3) mrsFAST based read depth ratios RD_{MF} identified in the species harboring the duplication were significantly different from at least one other species that was lacking duplications at the orthologous gene locus (with $p < 0.01$; based on the Mann-Whitney-U-test);

Since CopySeq evaluates duplication copy-number by considering reads mapping to the unique proportion of a gene (*i.e.*, singly unique nucleotide stretches, or SUNs), we were able to distinguish between duplicated genes that were annotated in the respective genome assembly from those that were not annotated in the reference genome – *i.e.* by evaluating CopySeq based read depth ratios RD_{CS}. Namely, for gene duplications annotated in the reference genome, we observed abnormal mrsFAST based read depth ratios RD_{MF} (indicative for a duplication), but normal CopySeq based read depth ratios RD_{CS} (indicative for no duplication affecting the corresponding singly unique nucleotide (SUN) positions of the respective paralogs). By comparison, gene duplications that are not annotated in the reference genome displayed abnormal read depth ratios both by mrsFAST ('RD_{MF}') and CopySeq ('RD_{CS}').

We distinguished between whole and partial inter-species gene duplications by evaluating the mrsFAST based read-depth. Specifically, each gene was partitioned into five equally sized bins (including exonic and intronic sequence). For each bin, the mrsFAST read depth ratio RD_{MF} and the length of unmasked sequence (regions not intersecting with repeat-masked sequence) were determined. The mrsFAST read depth ratios RD_{MF} were determined for unmasked sequence, requiring a minimum of 150 unmasked basepairs for each assessable bin. If each of the assessable bins showed $RD_{MF} >1.3$, the gene was classified as a whole gene duplication. Conversely, if not all assessable bins showed $RD_{MF} > 1.3$, the inter-species gene duplication was classified as "partial". Gene duplications were classified as ambiguous ("potential whole/partial"), if the number of assessable bins was less than 3 or if paralogous genes showed a mixture of whole and partial gene duplications (**Dataset 4**).

Evolutionary timing of inter-species gene duplications

To investigate the timing of gene duplications, inferred inter-species gene duplications were further filtered for: (1) consistent information in all studied species, *i.e.* genes with missing orthology annotation in at least one species, based on eggNOG, were not considered; (2) a maximum mrsFAST read depth ratio $RD_{MF} < 1.3$ in all non-duplicated species; (3) mrsFAST read depth ratios RD_{MF} in species harboring the duplicate were required to be significantly different from all other species at that locus ($p < 0.01$; Mann-Whitney-U test). By applying these filters we obtained a subset of our initial gene duplication data set containing the necessary information to time gene duplications in the primate tree. Overall we timed the duplication of 316 genes using this approach (see **Dataset 4**). For each gene duplication, we assessed whether the respective gene locus intersected a segmental duplication by at least 80% – for genes fulfilling this condition, we computed the pairwise sequence identity between ancestral and derived paralogs (**Figure 3A**). Ancient gene duplications displayed more sequence divergence between the original gene and its duplicate than more recent duplications (p<0.01; two-sided KS-test), in keeping with an accumulation of sequence changes over time.

qPCR validations of gene duplications

We performed qPCR across 8 primate species (human, chimpanzee, bonobo, orang-utan, olive baboon, savanna baboon, guinea baboon and rhesus macaque) to verify fixed gene duplications. We designed cross-species qPCR primers from conserved sequences (i.e., perfect alignments to the reference genomes for human (hg19), chimpanzee (panTro3), orang-utan (ponAbe2), rhesus macaque (rheMac2) and baboon (papHam1), **Dataset 2**). Our control primers amplified an ultraconserved element that is known to be diploid in all primates as described in [21]. The estimated haploid copy number for each sample was calculated using the standard curve method. In brief, a human sample was chosen to be the qPCR reference (NA10851). Serial dilutions of this reference were used to create a standard curve. The copy number for each sample was calculated by inputting the cycle threshold (Ct) into the standard curve to obtain an estimated amount of input DNA. This was then divided by the sample's actual input DNA (using DNA concentration and volume added) to determine its copy number relative to the reference. All reactions were run using SYBR Green 2X Master Mix (Life Technologies) on a 7900H real-time PCR machine (ABI) at default settings with an increased annealing temperature (65C).

mRNA-Seq data preparation and analysis

Strand specific RNA libraries were prepared with the ScriptSeq mRNA-Seq Library generation kit (Epicentre Biotechnologies) using 3-10μg of total fibroblast cell line RNA, followed by mRNA selection with Sera-Mag oligo (dT) beads (Distrilab). All samples were sequenced on an Illumina HiSeq2000 instrument in paired-end (101bp reads) mode using manufacturer's protocols. The fibroblast cell line derived mRNAseq reads, and mRNAseq reads derived from a compendium of six primate tissues [40], were aligned onto each reference genome using GSNAP [41] based on two complementary approaches: (1) accounting for reads mapping to multiple genomic positions by allowing a read to map up to 100 times (in analogy to using mrsFAST for duplication copy-number analysis);

(2) mapping reads to unique genomic positions (in analogy to using CopySeq for duplication copy-number analysis). We used non-unique alignments to measure the expression of orthologous genes across different primate species in the eggNOG database. The unique alignments were used to distinguish between the parent and duplicated copies. To assess the impact of gene duplications on tissue-specific gene expression, we searched for genes expressed in a tissue- and species-specific manner. We normalized mRNA-Seq reads as described previously [42], by (1) mapping reads onto exons, (2) performing GC-content correction, (3) merging reads falling into exons of the same gene, and (4) normalizing aligned reads by total sample read depth. We merged all samples from each species into one matrix per species to enable applying steps (1) to (3) for all samples from the same species simultaneously. Subsequently, raw expression values of genes (*i.e.* the outcome of step (3)) from different primate species were merged into a single matrix according to eggNOG orthologous group relationships [39]. In step (4), normalization by read depth was first performed on all samples from each primate species individually. To enable the comparison of expression values across samples from different species, we further normalized each sample by the median gene expression value, which was inferred after excluding non-expressed genes. In that way we obtained a median gene expression value of 1 in each sample, with harmonized gene expression values across species.

Impact of gene duplications on expression

To understand the impact of gene duplications on expression we considered all gene duplications for which at least one species had detectable expression levels (defined here as a normalized gene expression value of ≥0.2). We related the expression of fixed whole gene duplications at orthologous gene loci to the expression of these loci in species not harboring the duplication (**Figure 4A**) using a two-sided KS-test.

Impact of gene duplications on the acquisition of gene expression in new tissues

To assess the impact of gene duplications on gains of expression in new tissues, we analyzed a compendium of transcriptome data generated in different primate species and tissues [40]. We defined gains of expression in new tissues as follows: a gene that is expressed in one tissue in a species with a gene duplicate (normalized gene expression value of ≥0.2; defined as the number of exonic reads per gene, normalized by GC, overall read count and median expression in a sample), and is not expressed in another species lacking the gene duplicate (normalized gene expression value = 0).

We performed Monte Carlo simulations to assess whether the number of fixed whole gene duplications we observed to coincide with the emergence of new tissue expression (13 duplications) may have occurred by chance. Performing 1000 permutations, we randomly picked, without replacement, 113 genes (*i.e.* the total number of whole gene duplications encompassing expressed genes) from the whole matrix of expression values and assessed how many of these randomly drawn genes show a pattern of gained tissue expression. We verified that fixed whole gene duplications coincide with the emergence of expression in a new tissue significantly more often than expected (p=0.003, based on permutations, **Figure S9D**).

We further considered the possibility that whole-gene duplications gaining expression in new tissues can be explained simply by a ‘dosage-effect’, *i.e.* an increase in the gene's expression across *all* tissues of a species (rather than a *tissue-specific* gain in expression) proportional to the copy-number change, which may drive the expression level above our threshold for identifying genes as “expressed” in certain tissues. We hence also performed Monte Carlo simulations to assess whether our findings of whole gene duplications coinciding with acquired tissue expression can be explained solely by proportional increases in gene expression (‘dosage effect’). Performing 1000 simulations, we randomly picked, without replacement, 113 genes (*i.e.* the total number of whole gene duplications) and randomly assigned a copy-number status (picking, without replacement, from the set of actually observed gene copy-numbers) to these genes. We then increased the expression of each randomly picked gene proportionally to the duplication-specific copy-number change (using mrsFAST read depth ratios RD_{MF} as a basis; see **Figure S9E**). The new expression value is defined as follows:

$$expression_{new} = expression_{old} \cdot \frac{1}{5}\sum_{i=1}^{5}(RD_{MF})_i$$

expression = expression value

$(RD_{MF})_i$ *= mrsFAST read depth ratio of an individual sample i*

Notably, the modeled dosage effect failed to explain the gains of gene expression in new tissues that we identified in this study (p=0.02; based on permutation; **Figure S9E**) – suggesting that gains of expression in new tissues are not due to dosage, but involve other scenarios (such as an altered *cis* regulatory context of the newly emerged gene duplicate leading to acquisition of gene expression in new tissues).

To investigate *CST9LP1* as a candidate for newly acquired tissue expression in macaque, we applied BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) in two modes, requiring >94% sequence identity to identify BLAT hits: (1) we used the human DNA sequence of *CST9LP1* as a query in each species, (2) we used the specific DNA sequence of the annotated ortholog (based on eggNOG annotation) of *CST9LP1* in each species. Both analysis modes confirmed that *CST9LP1* is duplicated in rhesus macaque, in a species-specific manner. As *CST9LP1* is annotated as a tentative gene (potential pseudogene) in Ensembl, we investigated whether *CST9LP1* and its orthologs harbor premature stop codons indicative for a pseudogene. We extracted the sequence for each annotated ortholog (based on eggNOG annotation) and analysed predicted open reading frames (ORFs) using SMS2 [43]. We could not find any premature stop codons in any *CST9LP1* ortholog, including the new duplicate in rhesus macaque with the exon structure, transcript length and protein length of *CST9LP1* being very well conserved across all species, a finding that we considered as strong evidence that *CST9LP1* represents a functional gene. To understand the contribution of the new gene duplicate of *CST9LP1* in rhesus macaque to the overall expression signal, we assessed alignments of unique and perfectly matching reads in IGV (IGV version 2.0.34) and considered exonic reads only (**Figure 4D**).

SUPPLEMENTAL REFERENCES

1. Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16, 418-420.
2. Fujita, PA et al. (2010). The UCSC Genome Browser database: update 2011. Nucleic Acids Res., 39(Database issue):D876-82.
3. Chimpanzee Genome Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055):69-87.
4. Gibbs RA, et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316(5822):222-234.
5. Locke DP, et al. (2011) Comparative and demographic analysis of orang-utan genomes. Nature 469(7331):529-533.
6. Mills RE, et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470(7332):59-65.
7. Cabili MN, Trapnell C, Goff L, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & Development:1915-1927.
8. Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment / Map format and SAMtools. Bioinformatics;25(16):2078-2079.
9. Alkan C, Coe BP, Eichler EE. (2011) Genome structural variation discovery and genotyping. Nature reviews. Genetics;(March):15-18.
10. Rausch T, Zichner T, Schlattl A, et al. (2012) DELLY : structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics;28:333-339.
11. Abyzov A, Urban A E, Snyder M, Gerstein M. (2011) CNVnator: An approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research.
12. Handsaker RE, Korn JM, Nemesh J, & McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43(3):269-276.
13. Waszak SM, Hasin Y, Zichner T, et al. (2010) Systematic Inference of Copy-Number Genotypes from Personal Genome Sequencing Data Reveals Extensive Olfactory Receptor Gene Content Diversity. PLoS Computational Biology;6(11).
14. Lam HY, et al. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol 28(1):47-55.

15. Kent WJ, Kent WJ. (2002) BLAT -- The BLAST-Like Alignment Tool BLAT — The BLAST-Like Alignment Tool. Genome Research:656-664.
16. Perry GH, Tchinda J, McGrath SD, et al. (2006) Hotspots for copy number variation in chimpanzees and humans. Proceedings of the National Academy of Sciences of the United States of America;103(21):8006-11.
17. Perry GH, et al. (2008) Copy number variation and evolution in humans and chimpanzees. Genome Research 18(11):1698-1710.
18. Lee AS, et al. (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. Hum Mol Genet 17(8):1127-1136.
19. Gokcumen O, et al. (2011) Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. Genome Biol 12(5):R52.
20. Gazave E, Darre F, Morcillo-Suarez C, et al. (2011) Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome Research
21. Iskow RC, et al. (2012) Regulatory element copy number differences shape primate expression profiles. Proc Natl Acad Sci U S A 109(31):12656-12661.
22. Wang K, Li M, Hakonarson H. (2010) ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research: 38(16):1-7.
23. Beaumont, MA, Zhang, W & Balding, DJ.(2002) Approximate Bayesian computation in population genetics. Genetics, 162(4), pp.2025–35.
24. Pavlidis, P, Laurent, S & Stephan, W. (2010) msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. Molecular ecology resources, 10(4), pp.723–7
25. Csilléry, K, François, O & Blum, M.G. (2011). abc: an R package for Approximate Bayesian Computation (ABC). Methods in Ecology and Evolution, pp.1 –19.
26. Auton A, et al. (2012) A fine-scale chimpanzee genetic map from population sequencing. Science 336(6078):193-198.
27. Yuan Q, Zhou Z, Lindell SG, Higley JD, Ferguson B et al. (2012): The rhesus macaque is three times as diverse but more closely equivalent in damaging coding variation as compared to the human. BMC Genetics, 13:52.
28. Conrad DF, Keebler JEM, Depristo MA, et al. (2011) Variation in genome-wide mutation rates within and between human families. Nature Genetics, 43(7):712-715.
29. Scally A, Durbin R. (2012) Revising the human mutation rate: implications for understanding human evolution. Nature Reviews Genetics,13:745-753.

30. Abouelhoda M, Ohlebusch E.(2004) CHAINER: Software for Comparing Genomes. Proc. 12th International Conference on Intelligent Systems for Molecular Biology/3rd European Conference on Computational Biology

31. Abouelhoda, M.I., Kurtz, S., and Ohlebusch, E. (2008). CoCoNUT: an efficient system for the comparison and analysis of genomes. BMC Bioinformatics 9, 476.

32. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16, 276-277.

33. Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. (2005). Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77, 78-88.

34. McKenna A, Hanna M, Banks E, et al. (2010) The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research:1297-1303.

35. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics;25(21):2865-71.

36. Lee E*, et al.* (2012) Landscape of somatic retrotransposition in human cancers. Science 337(6097):967-971.

37. Walker JA, Konkel MK, Ullmer B, et al. (2012) Orangutan Alu quiescence reveals possible source element: support for ancient backseat drivers. Mobile DNA.;3(1):8

38. Hach F et al. (2011) mrsFast: a cache-oblivious algorithm for short-read mapping. Nature methods;7(8):576-577.

39. Powell S, et al. (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res 40(Database issue):D284-289.

40. Brawand D, et al. (2011) The evolution of gene expression levels in mammalian organs. Nature 478(7369):343-348.

41. Wu TD, Nacu S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics;26(7):873-881

42. Schlattl A, Anders S, Waszak SM, et al. (2011) Relating CNVs to transcriptome data at fine resolution : Assessment of the effect of variant size , type , and overlap with functional regions Relating CNVs to transcriptome data at fine resolution. Genome Research;(21):2004-2013.

43. Stothard P (2000) The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102-1104.