Copy Number Variation Analysis In The Great Apes Reveals Species-Specific Patterns Of Structural Variation

Elodie Gazave*, Fleur Darré*, Carlos Morcillo-Suarez, Natalia Petit-Marty, Angel Carreño, Urko M. Marigorta, Oliver A. Ryder, Antoine Blancher, Mariano Rocchi, Elena Bosch, Carl Baker, Tomàs Marquès-Bonet, Evan E. Eichler and Arcadi Navarro

*These authors contributed equally to this work

Supplementary Material and Methods

Sample collection

A list of individuals, origins and hybridization arrays can be found in Supplementary Table S11. The geographical origins of the ancestors of our samples are sometimes unavailable. Gathering all possible information, we could determine that, among the individuals used on the oligonucleotide array, 5 chimpanzees are Central chimpanzees (*Pan troglodytes troglodytes*), 1 is a Western chimpanzee (*Pan troglodytes verus*), and 1 is of at least of 3/4 Pan troglodytes troglodytes ancestry. All the gorillas were Western lowland gorillas (Gorilla gorilla gorilla). Five orangutans were of Sumatran origin (*Pongo abelii*) and three were from Borneo (*Pongo pygmaeus*).

BAC array hybridization

Hybridizations were performed as described by Wang et al. (2004) (2004). Arrays were scanned using an Agilent G2565BA MicroArray Scanner System (Agilent Inc., Palo Alto, Ca) and the acquired images were analyzed using GenePix Pro 6.0 software (Axon, Molecular Devices) using the irregular feature finding option. Raw data were filtered and normalized (loess method) using the Limma R package (Smyth and Speed 2003).

CNV-calling from BAC tiling-path array

The aCGH data obtained from the hybridizations with the BAC array were analyzed with the R statistical software package (Ihaka and Gentleman 1996). Data from each species (chimpanzee, gorilla, orangutan) were analyzed separately with identical procedures. Only autosomes were considered in the analysis. For each individual, a combination of two approaches was applied to detect clones with a significant signal. In the first approach, the values obtained with each dye combination were not averaged. The algorithm we used had two steps: First, it looked for two consecutive clones or 3 out of 4 consecutive clones that would have log2 ratios above 1.5 times the standard deviation of the whole individual. Then, it checked for consistency between direct and dye-swap hybridizations, and only kept the regions that were called in both dye combinations. This method is very sensitive to the quality of the data and is not adapted to cases where data are dispersed or if there are local trends in the data. In the second approach, we applied algorithms designed to deal with these problems. The normalized log ratios from each dye-swap were averaged before the analysis. Then, a denoising step was applied using the method described in Hsu et al. (2005), using the Haar wavelet family and the sure estimator for thresholding, with J_0 (the level up to which the wavelet coefficients are subject to thresholding) equal to 4. The wavelet decomposition and reconstruction functions were from the WAVESLIM package. Finally, the

Circular Binary Segmentation algorithm described in Olshen et al. (2004) was used for clone calling. For this purpose, the functions *CNA*, *smoothCNA* and *segment*, available in the DNAcopy package, were used with default parameter values. Data from the two approaches were combined. Overall, the second analysis method is more conservative, but it has the ability to rescue some regions that would not pass the very strict threshold based on standard deviations alone. Consequently, the overlap between the regions called with the two methods is large but not complete (212 calls common in both methods over 531 calls in total, corresponding to 56.3% in length). This is why the clones that were called by only one method were manually checked in the data file for validity of signal. At the end of the BAC-array calling process, regions larger than 1 Mb were also manually checked. All clone coordinates are given for Human Genome Build 35, hg17.

Choice of parameter values for the oligonucleotide array CNV detection

For each array, an optimal set of *HMMseg* parameters was manually determined. To do so, we started by performing several (up to 20) runs of the algorithm under a wide range of emission probabilities. Increasing or decreasing the values of parameters in opposite directions for both, call extension and signal detection. After checking the results of every run, we selected the set of parameter values that would simultaneously optimize number and extension of calls. Regarding call extension, we selected parameter values that avoided over-extension or over-fragmentation. For signal detection, the chosen values were the ones better favoring the detection of consistent signals over long stretches of probes while avoiding the detection of many short (probably noisy) calls. The final set of parameters used for each array is available in Supplementary Table S12. Examples of signal intensity (log2 ratio) are shown for a selection of genomic regions in Supplementary Figure S10a, b, c, d and e. On these figures, data are shown for all the individuals and for each dye combination. Figure 11a, 11b, 11c and 11d presents a CNVR in bonobo, chimpanzee, gorilla and orangutan, respectively. On Figure 11e, we can see that this CNVR shared by the four species is formed by CNV largely or fully overlapping among individuals or among species.

Statistical tests for CNV length

The differences in mean CNV length were tested by permutation tests in which individuals were permuted among species. P-values correspond to the number of times the difference between the averages of the two groups was larger than the observed difference. Given our sample size, the exact number of possible different pairwise individual randomizations ranges from 352 (5 vs. 6 individuals) to 32,032 (5 vs. 9 individuals). For each pairwise comparison, we performed 100,000 permutations to guarantee balanced and complete exploration of the space of possible randomizations. We did not perform permutation tests on CNVR length since re-computing CNVRs from the CNV calls of individuals belonging different species would generate artifactually large and short CNVRs.

Supplementary Results

Basic description of the CNV detected in the discovery phase

The BAC array hybridizations allow us to call 251 CNV in 9 chimpanzees, 129 CNVs in 8 gorillas, and 151 CNVs in 7 orangutans (see Supplementary Table S14). The full list of CNVs is available on Supplementary Table S15. The CNV regions plotted on a karyotype of human chromosomes show that they are distributed everywhere in the genome (Supplementary Figure S5) with considerable overlap with known SD regions.

The largest dataset of CNVs in great apes available so far comes from a study on chimpanzees by Perry et al. (2008). In this paper, the authors use a method based on BAC-arrays, which makes their results roughly comparable to the chimpanzee CNVs made in the discovery phase of the present study. As to the differences in number of calls, call length and region length between our data (see Supplementary Table S14) and those from Perry et al. (2008), they are mainly explained by differences in the calling methods. We were more conservative since we did not consider CNVs detected by a single positive BAC (see Supplementary Material and Methods) and thus called larger and less numerous CNVs than Perry et al. (2008).

Supplementary Table S14 also shows that the proportion of individuals per region in the present study is higher (29% here, vs. 15% in Perry et al. 2008). The fact that the data set of Perry et al. (2008) presents more low frequency CNVs is probably explained by the differences in sample size between both studies. In Perry et al. (2008), 30 individuals were examined as opposed to the 9 individuals in our study. To test for this difference in sample size effect, we performed 30,000,000 resamplings of 9 out of 30 individuals from the Perry et al. (2008) dataset. To these resampled individuals, we applied our calling filters so that they could be compared with our own set. In addition we recomputed the average proportion of individuals per region from the CNV list available as supplementary information in Perry et al. (2008). The mean proportion of individuals per regions remains significantly lower (P= 0.0442) in Perry et al. (2008), but the average over all the resamplings of the proportion of individuals per region is 27%, which is much closer to the proportion observed in the present study.

In addition, we plotted the frequency of CNVs among individuals. Supplementary Figure S6 shows that, although Perry et al.'s (2008) study presented on average more low-frequency CNVs than ours, their distribution of CNVR frequency is very similar to ours. When we further split CNVs into two categories, overlapping SD and not overlapping SD, we see that the patterns of frequency are again similar in both studies (Supplementary Figure S7). This confirmed that, although our sample size was not as large as in other studies, it was large enough to detect with accuracy species-specific patterns of structural variability.

Suitability of the targeted oligonucleotide array for genome-wide analyses

For the individuals that were common in both phases (discovery and refinement), we validated 150 CNV BAC-array calls made in discovery phase with 307 CNV calls in the oligonucleotide array (Supplementary Table S13c), the higher number of oligonucleotide calls being due to the better resolution of the oligonucleotide array, that tends to split larger calls made with the BAC array. The oligonucleotide validation represents between 54 to

90% of the BAC calls, depending on the individual. In addition, 305 CNVs were detected with the oligonucleotide array in places where the BAC array had not produced any signal for the analyzed individual, but had been called with the BAC array in other individuals of the same species. These later cases may have been real CNV that we did not detect with the BAC array as a consequence of either the lower resolution of the BAC array or the conservative criteria we applied for its analysis. Finally, 427 oligonucleotide calls were made in regions that were never called in our discovery phase. This is mostly due to the fact that in the design of the targeted array we covered regions of SD and CNV described in the literature (see Material and Methods). Altogether, these results show first, that the detection of CNV by BAC and by the oligonucleotide array are consistent, and second, that the validation phase efficiently allow a better definition of CNVs.

Further confirmation of the consistency of our detection methods comes from the similarity of results obtained for CNVRs detected with the BAC array and the oligonucleotide array. For example, a plot of the proportion of individuals supporting CNVR detected in the discovery phase (Supplementary Figure S8) shows that, on average, CNV regions in chimpanzee are supported by more individuals, and conversely CNV regions supported by few individuals are more frequent in orangutan and gorilla. As to the relationship between SDs and CNVs in our focus species, we see that there is an enrichment of CNVs in places of SDs, with 57% of the chimpanzee CNVs overlapping SDs (35.6% in gorilla and 59% in orangutan). This enrichment is roughly the same as the one observed when CNVs were detected with the oligonucleotide array (chimpanzee 72%, gorilla 58.8%, orangutan 74%), although it is a bit higher in the validation phase because the array was targeting SDs (see Material and Methods). Altogether, the similarity of patterns observed in the genome-wide approach and in the targeted approach demonstrates that although the targeted array is not covering the entire genome, it covers the most representative CNV regions and therefore is a good proxy for a genome-wide study with the added value of a higher resolution.

Finally, we analyzed the possibility that the use of arrays based on human sequences would decrease our power to detect CNVs in distant species because a larger number of primate DNA sequences would not show a perfect match with the human probe. Since all our hybridizations are performed using a sample and reference of the same species, if a possible bias exists it should not affect intra-specific CNV frequency, but it may account for the variations in total number of CNVs discovered among species. To test for this effect, we counted the total number of CNVs called in one dye combination and in the other. In our calling strategy, we kept only overlapping calls so, for each individual, the total inconsistency between one dye combination and the other reflects the amount of noise on each hybridization (note that it does not constitute anyhow an estimate of false positive or false negative calling, since we never adapted our methods to single-dye detection of CNVs). If sequence divergence introduced any bias, we would expect that, as phylogenetic distance from humans increases, the primate DNA would hybridize less consistently. Consequently, we should observe increased noise in the hybridization signal, detectable as higher inconsistency between the two dye combinations for the same individual. Supplementary Table S13b shows the percentages of inconsistency for each individuals and species. We can see that this value does not increase with phylogenetic distance to humans, but rather varies with the individual sample, reflecting variations in DNA quality. Thus, at least for the species under study, sequence divergence cannot explain our observations. For instance, sequence divergence is not an adequate explanation for the lower number of CNVs in the orangutan. The lower number of CNVs in the orangutan is very likely to be a property of this species compared to the African Apes, as described elsewhere (Locke et al. 2011).

Differences in CNV length

Using the oligonucleotide array we observed important differences in mean CNV length among species (see Supplementary Table S1). Permutations test show that the comparisons are significant, except for chimpanzee versus orangutan and chimpanzee versus bonobo (Supplementary Table S3a). However, we noticed the presence of a few outlier CNVs (2 in bonobo, and 5 in chimpanzee, see Supplementary Figures S9a, S9b, S9c, and S9d) with extreme length (>1.3 Mb, which is almost twice as large as the next larger call and represents a clear discontinuity in the distribution of length). To test whether some individuals with extreme CNV lengths are inducing the differences observed among species, we performed the permutation test without these outliers. Results show that the difference in call length between orangutan and chimpanzee is no longer significant (Supplementary Table S3b), meaning that the difference in average call length between these two species was due to the presence of a few extremely long calls in chimpanzees. On the contrary, the difference between bonobo and chimpanzee which was not significant with the full list of calls becomes significant when tested without outliers. This indicates that a few extremely long calls in both species rendered them artifactually similar, when chimpanzee has actually smaller calls on average if we do not consider these outliers.

However, it is important to note that measures of CNV (and CNVR) length are not fully reliable for two reasons. First, 48% of our CNV calls have at least one of their edges located exactly at the limit of the regions covered in the array and, thus, length measures tend to be underestimates. Second, some of these differences might be attributed to the design of our array. Full information on bonobo and gorilla SDs is not publicly available and bonobos were not used in the CNV discovery phase, so the array was designed without that information. Therefore, CNVR length estimates in these species are probably less reliable than in chimpanzee and orangutan, whose SDs were fully considered and that present no significant differences in call length. To assess the relevance of the latter issue, we focused on CNVRs overlapping SDs that are shared between humans, chimpanzees and orangutans (that is SDs that were targeted by our oligonucleotide array and are likely to be shared among all the great apes and therefore present less ascertainment bias in bonobo and gorilla). We observed that average CNVR lengths were increased in gorilla and dramatically reduced in bonobo. It is clear that only full consideration of segmental duplication maps from all the great apes, probably by full-genome sequencing, will allow reliable assessment of species-specific differences in CNV length.

CNV frequencies

The average bonobo CNVR is shared by 2.10 out of 5 individuals, while these values are 3.10 out of 9 individuals in chimpanzee, 1.93 out of 6 in gorilla and 2.62 out of 9 in orangutan. Given the difference in sample size and CNVR number described in the four species, comparisons among species were performed by looking at proportions of the total sample (see main text) instead of absolute numbers of individuals. These proportions are not driven by the amount of singletons in a given species. For example, even if CNVRs tend to be low frequency events in the orangutan (29%), this species shows the lowest proportion of singletons (39.8%, see Table 1 and Figure 4). This cannot be explained only by a larger sample size in orangutan, because chimpanzee has the same number of individuals and the same amount of singletons (40.3%) but a higher average CNV frequency (32%). Gorillas have the highest proportion of singletons (55%), which is reflected in Figure 3. There, we can see that the first gorilla cluster (edge number 17), that is, the two closest gorillas, is built late relatively to clusters formed within the three other species, showing that gorilla individuals are less similar to each other because they tend to have singletons.

The relationship between CNVs and SDs

We observed that around two thirds of the CNVRs of a given species overlap at least partially with known SDs (ranging from 59% in gorilla to 72% in chimpanzee, Table 2a). This overlap is higher than expected by chance in all the species (permutation test considering only the regions tiled by the array, $P<10^{-5}$). However, the association between CNVRs and SD could be due to some CNVRs being longer and, thus, having higher probabilities of overlapping with SDs. Independently of the biological relevance of the CNV lengths observed in this study (see above), we performed a simple test of the lengths of observed CNVRs according to their relation to known SDs. This test allows us to ascertain the possibility of an artifactual association. We found that in bonobo and gorilla, CNVRs overlapping known SDs are actually shorter than those away from SDs (Mann-Whitney tests, $P=7.48*10^{-12}$ and P=0.00031 respectively). In orangutan there is no difference in median length (P= 0.13) and in chimpanzee CNVRs that overlap SD are marginally larger (P=0.038). Thus, there is not a common trend and therefore CNVRs do tend to co-occur with SDs independently of their length.

References

Hsu L, Self SG, Grove D, Randolph T, Wang K, Delrow JJ, Loo L, Porter P. 2005. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6:** 211-226.

Ihaka R and Gentleman R. 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5:** 299-314.

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

Olshen AB, Venkatraman ES, Lucito R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5:** 557-572.

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C et al. 2008. Copy number variation and evolution in humans and chimpanzees. *Genome Res* **18:** 1698-1710.

Smyth GK and Speed T. 2003. Normalization of cDNA microarray data. Methods 31: 265-273.

Wang NJ, Liu D, Parokonny AS, Schanen NC. 2004. High-resolution molecular characterization of 15q11-q13 rearrangements by array comparative genomic hybridization (array CGH) with detection of gene dosage. *Am J Hum Genet* **75:** 267-281.