Genome Biology

Genome Biology
Biology for the post-genomic era

Fast-evolving non-coding sequences in the human genome

Genome Biology 2007, **8**:R118 doi:10.1186/gb-2007-8-6-r118

Christine P Bird (cpb@sanger.ac.uk)
Barbara E Stranger (bes@sanger.ac.uk)
Maureen Liu (ml5@sanger.ac.uk)
Daryl J Thomas (daryl@soe.ucsc.edu)
Catherine E Ingle (ci2@sanger.ac.uk)
Claude Beazley (cb9@sanger.ac.uk)
Webb Miller (webb@cse.psu.edu)
Matthew E Hurles (meh@sanger.ac.uk)
Emmanouil T Dermitzakis (md4@sanger.ac.uk)

ISSN	1465-6906
Article type	Research
Submission date	19 December 2006
Acceptance date	19 June 2007
Publication date	19 June 2007
Article URL	http://genomebiology.com/2007/8/6/R118

Fast-evolving non-coding sequences in the human genome

Christine P Bird[1], Barbara E Stranger[1], Maureen Liu[1], Daryl J Thomas[2], Catherine E Ingle[1], Claude Beazley[1], Webb Miller[3], Matthew E Hurles[1], Emmanouil T Dermitzakis[1]

[1] The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK
[2] Center for Biomolecular Science and Engineering, University of California Santa Cruz, CA, USA
[3] Department of Computer Science & Engineering, Penn State University, PA, USA

Corresponding author
Emmanouil T. Dermitzakis
The Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus,
Hinxton, CB10 1SA
UK
e-mail: md4@sanger.ac.uk
Tel : +44-1223-494866
Fax : +44-1223-494919

Abstract

Background

Gene regulation is considered one of the driving forces of evolution. Although protein-coding DNA sequences and RNA genes have been subject to recent evolutionary events in the human lineage, it has been hypothesized that the large phenotypic divergence between humans and chimpanzees has been driven mainly by changes in gene regulation rather than altered protein-coding gene sequences. Comparative analysis of vertebrate genomes has revealed an abundance of evolutionarily conserved but non-coding sequences. These conserved non-coding (CNC) sequences may well harbour critical regulatory variants that have driven recent human evolution.

Results

Here we identify 1356 conserved non-coding sequences that appear to have undergone dramatic human-specific changes in selective pressures, at least 15% of which have substitution rates significantly above that expected under neutrality. The 1356 ‘accelerated CNCs’ (or ANCs) are enriched in recent segmental duplications suggesting a recent change in selective constraint following duplication. In addition, SNPs within ANCs have a significant excess of high frequency derived alleles and high F_{ST} values relative to controls indicating that acceleration and positive selection are recent in human populations. Finally, a significant number of SNPs within ANCs are associated with changes in gene expression. The probability of variation in an ANC being associated with a gene expression phenotype is 5-fold higher than variation in a control CNC.

Conclusion

Our analysis suggests that ANC sequences have until very recently played a role in human evolution, potentially through lineage-specific changes in gene regulation.

Background

The manner in which the expression of genes is regulated defines and determines many of the cellular and developmental processes in an organism. It has been hypothesized that variation in gene regulation is responsible for much of the phenotypic diversity within and between species [1]. In particular, it was proposed a few decades ago that the phenotypic divergence between human and chimpanzees is largely due to changes in gene regulation rather than changes in the protein-coding sequences of genes [2]. Although it has been long recognized that regulatory sequences play an important role in genome function, the fine structure and evolutionary patterns of such sequences are not well understood [3], mainly due to the fact that such sequences have a much more complex functional code and appear not to be restricted to particular sequence motifs. One of the most powerful approaches to identify regulatory sequences has been the use of multiple-species comparative sequence analysis to look for conserved non-coding sequences [4], but these sequences only represent a subset of regulatory elements in the genome and only a subset of them are regulatory elements [5].

Conserved non-coding (CNC) sequences are distributed throughout the genome in a manner independent of gene density [6, 7]. Studies of nucleotide variation have revealed strong selective constraints on CNCs in human populations [8], so there is little doubt that a large number of them have a functional role. The abundance and genomic distribution of CNCs has raised intriguing questions about the functions of such sequences in the genome. Although a small fraction of the CNCs can be associated with transcriptional regulation (the majority of the most highly conserved examples of CNCs appear to be enhancers of early development genes [5, 9]), there remains a large number of CNCs with unexplained function.

Although the identification of CNCs relies on sequence conservation, it is conceivable that some of the most interesting functional non-coding elements are also evolving under positive (directional) selection in particular lineages. Studies in Drosophila have suggested such a pattern observed in untranslated regions (UTRs) and some introns and intergenic DNA [10]. Moreover, loss-of-function mutations as well as mutations that lead to gain of novel functions are also likely to contribute to evolutionary change [11, 12]. A relatively recent model for the evolution of novel gene function following gene duplication proposed that the reciprocal degeneration of regulatory elements after duplication (Duplication-Degeneration-Complementation or DDC) [13] could drive gene subfunctionalization, and an older model of gene duplication proposed an important role for positive selection after duplication [14-16]. All of the above evolutionary processes could contribute to phenotypic evolution in the human lineage, and would result in a lineage-specific acceleration of the substitution rate of associated functional non-coding DNA.

In the present study, we have performed an analysis of lineage-specific acceleration of previously identified CNCs in vertebrates. By comparing the CNC sequences of three genomes, the human, chimpanzee and macaque, we identify 1356 CNCs that have an excess of human-specific substitutions relative to the chimpanzee lineage. By analyzing the genomic distribution and nucleotide variation of these fast-evolving (accelerated) CNCs we find that significant numbers of them are found in the most recent (mostly human-specific) segmental duplications and single nucleotide polymorphisms (SNPs) within them are associated with changes in gene expression. We also find a strong signal of recent directional selection in the human lineage.

Results

Searching for fast-evolving (accelerated) CNCs

We have selected 304,291 of the most conserved non-coding sequences of at least 100 bps in length to look for evidence of accelerated substitution rate in the human lineage (see methods), by comparing the orthologous sequences of CNCs between human and chimpanzee. We used a chi-square based test to detect regions of CNC sequence that are diverging at an accelerated rate in either the human or chimpanzee lineage [17]. The test requires at least 4 substitutions between human and chimpanzee. Of the 304,291 CNCs, only 26,475 have at least 4 human-chimpanzee substitutions. For those 26,475 CNCs, we generated human-chimpanzee-macaque three-way alignments to infer the direction of substitutions and performed Tajima's one-tailed chi-square test to detect human-or chimpanzee-specific substitution rate acceleration, applying the Yate's correction for continuity to correct for small substitution counts [17]. The chosen P-value threshold was P = 0.08 because it was the P-value with the minimum False Discovery Rate (see methods) in the range of P-values between 0.05 and 0.15, (FDR = 75%). At this threshold we detected a total of 2794 (10.6%) accelerated CNCs (hereafter referred to as ANC – Accelerated Non-Coding) in either the human (1356 ANCs or 5.1%) or chimpanzee (1438 ANCs or 5.3%) lineages (Figure 1A) with P less than or equal to 0.08, while we expect only 2118 in total by chance. The FDR of 75% is likely to be an overestimate since the Yate's correction is generally considered conservative.

Comparison of the human and chimpanzee chromosomes in the alignments reveals that only 20 out of 1356 are not on the expected syntenic chromosome (see Additional Data File 1). We have also performed visual and manual examination of a random sample of 5% of the ANCs across the whole spectrum of significance (see Additional Data File 1) to confirm that the signals we detect are not a result of misalignments and we have concluded that this is very rare (only 2 out of 72 cases are potentially problematic). Some of the ANCs overlap with features that could potentially create such patterns (segmental duplications, retroposed genes and pseudogenes), but in all the cases that we tested, the result cannot be explained by misalignment. In fact, if we exclude sequences that could generate potential alignment artefacts (segmental duplications, retroposed genes and pseudogenes – see below) we then detect 1145 human ANCs (Figure 1B) relative to 18,289 *power CNCs*. The false discovery rate is estimated at 40%, which suggests that 688 (60%) of ANCs are true positives, a larger proportion than estimated above. We discuss below the relevance of such overlaps to real biological signals and hence their inclusion. However we also perform all the analysis below excluding the ANCs in the above features to confirm the validity of the obtained results.

Two recent studies have also described accelerated non-coding sequences in the human genome [18, 19]. A total of 37 of the 202 HARs (Human Accelerated Regions) (18%) of the Pollard study and 159 of the 992 accelerated CNSs (Conserved Non-coding Sequences) of the Prabhakar study (16%) overlap our set of ANCs. The overlap between these sets is also low, 51 of the 202 HARs (25%) overlap the Prabhakar study. The overlap between studies (Figure 2) is highly significant and all three studies are capturing similar signals but obviously the overlap is not complete. One explanation for the limited overlap between the three studies is that there are many accelerated non-coding sequences, most of which can't be detected because of a lack of power. However, it is difficult to distinguish this explanation from the differences expected from three methods relying on different assumptions. In particular, our study uses a methodology that specifically detects human lineage-specific acceleration relative to the chimpanzee and the identification of ANCs is mutually exclusive in the two species, which is not the case in the two other studies.

Throughout this analysis we use the following sets of DNA sequences as genomic controls against which we compare the human ANCs: 1) the 23,681 non-accelerated CNCs with at least 4 substitutions sufficient to detect significant acceleration (excluding human and chimpanzee ANCs, and hereafter referred to as *power CNCs*), and 2) all remaining 277,814 non-accelerated CNCs (excluding *power CNCs*).

Positive selection vs. loss of constraint

The analysis above allows us to identify CNCs that have accelerated rates of substitutions in humans relative to chimpanzees. This acceleration can be due either to loss of selective constraint or positive selection and the biological interpretation of the two is different. Loss of selective constraint should result in sequences adopting the neutral rate of evolution, whereas sequences under positive selection might be expected to be evolving more rapidly than under neutral evolution. In order to obtain a minimum estimate of the fraction of the 1356 ANCs that are undergoing positive selection we compared the human lineage-specific substitution rate of ANCs to that of 50,846 and 50,627 regions of the same size distribution as the CNCs that are 10 Kb away and 500 Kb from a CNC, respectively and with at least 4 substitutions between human and chimpanzee. As a threshold to determine whether an ANC has a substitution rate higher than neutral we defined the 5% tail of the distributions of human lineage-specific divergence of the two sets. These thresholds are $d_{0.05at10Kb}$ = 0.0267 and $d_{0.05at500Kb}$ = 0.0268. A total of 260 (19%) and 259 (19%) ANCs have rates higher than these thresholds respectively, while only 5% (68 ANCs) are expected by chance. This suggests that at least 191 ANCs have undergone sequence divergence consistent with positive selection. If we exclude potentially confounding ANCs we observe that 200 of the 1145 ANCs (17.5%) have a human lineage-specific rate above the neutral threshold and that this accounts for at least 143 ANCs presumably under positive selection.

In an alternative approach we compared the human lineage-specific rate to the synonymous substitution rate estimated from human and chimpanzee [20], which in some cases may serve as a neutral proxy. The average synonymous substitution rate was computed as Ks = 0.0141 +/-0.0132 (mean +/- stdev), an estimate of the expected human Ks rate is taken as half that. We consider two upper bounds of neutral rate as Ks_{2stdev}= mean + 2Stdev = 0.0203 and Ks_{3stdev} = mean + 3Stdev = 0.0270. With

Ks_{2stdev} and Ks_{3stdev}, 515 ANCs (38%) and 253 ANCs (18%) respectively are estimated to have undergone positive selection. Similar results are obtained if we consider the observed distribution of Ks values to determine the 95% ($p < 0.05$) and 99% ($p < 0.01$) upper confidence limits. We conclude that at least 15% and potential more than a third of the ANCs are evolving faster than the neutral substitution rate. Synonymous sites can be constrained but the fact that all three methods give similar results suggests that 15-19% of ANCs have substitutions rates above what is expected by neutral evolution.

Genomic location of ANCs

We investigated the possibility that ANCs are degenerate regulatory elements associated with subfunctionalized genes or elements that have decayed in function following duplication in a manner similar to pseudogenes. We explored the distribution of ANCs, *power CNCs* and non-accelerated CNCs in recent Segmental Duplications (SDs) of the human genome as defined in recent studies [21, 22]. Approximately 5-6% of the genome is included in SDs but we find 8% of the ANCs, 10% of the *power CNCs* and only 5% of non-accelerated CNCs (Table1) within SDs. This suggests an enrichment of ANCs and *power CNCs* in SDs and this is significantly different from the density of non-accelerated CNCs in SDs (chi-square test, $P < 10^{-4}$).

We subsequently considered the age of the SDs containing ANCs, *power CNCs*, and non-accelerated CNCs, by comparing the distribution of percent identity between paralogs of SDs overlapping each of the 3 sets above. The distribution for SDs containing ANCs reveals that ANCs are highly enriched within recent SDs of low divergence (less than 2%; Figure 3). The distributions of the two controls are both significantly skewed toward an excess of old and highly diverged SDs (Mann-Whitney-U-test; $P < 10^{-4}$). This strongly suggests that some ANCs have undergone modification of their selective pressures (either loss of selective constraint or positive selection) after very recent duplication.

To test for enrichment of ANCs in variable genomic duplications segregating in human populations, we intersected ANCs, *power CNCs*, and non-accelerated CNCs with human copy number variants (CNVs) from a public database (Database of Genomic Variants in Toronto [23]). The enrichment we observed was entirely due to high overlap between CNVs and SDs suggesting no enrichment of ANCs in CNVs per se.

We further explored the overlap of ANCs, *power CNCs* and the non-accelerated CNCs with retroposed genes and pseudogenes. Only 8% of ANCs overlap these elements compared to an overlap of 15% for the *power CNCs* (chi-square test, $P < 10^{-4}$) (see Table 1). This supports the idea that the detection of acceleration in ANCs is not due to misalignments since one of our control sets, the *power CNCs*, are more enriched for retroposed genes and pseudogenes. Normally, most studies exclude such sequences from the analysis because they are considered noise, but in light of recent studies that associated function with repetitive elements [24, 25], we retained all ANCs and CNCs overlapping such elements for subsequent analysis, but in most cases we also perform the analysis without them to control for any biases they might introduce.

Historical and recent patterns of nucleotide variation

We further explored the patterns and levels of nucleotide variation in ANCs in human populations to determine whether the processes that shape the evolution of ANCs are historical (predating human coalescent time) and/or recent in human populations. We used the derived allele frequency (DAF) spectrum of SNPs from the phase II HapMap [26, 27]. The state of the allele (either derived -new- or ancestral) was inferred by aligning the SNP position to the chimpanzee genome and using parsimonious assumptions (see methods). Regions with an excess of SNPs with high DAF relative to the expectations of a neutral equilibrium model are likely to be evolving under positive selection [28].

We defined 5 sets of SNPs from the Yoruba (YRI) population of the HapMap [26] project: SNPs within ANCs (n = 682), *power CNCs* (n = 28,722), non-accelerated CNCs (n = 48,811), and two new control sets of SNPs (n = 28,408 and 28,722) from 1356 20Kb windows located 500Kb 5' and 3' of the ANCs. The DAF spectrum of the ANCs has a significant excess of high-frequency derived alleles relative to the DAF spectrum of all control sets (Mann-Whitney-U test; $P < 10^{-4}$) (Figure 4A). The DAF spectrum of the *power CNCs* is more similar to the neutral controls than to that of the non-accelerated CNCs, possibly suggesting that *power CNCs* are a mix of ANCs and non-accelerated CNCs. The other HapMap populations exhibit very similar patterns (data not shown).

As SNPs in SDs and CNVs can exhibit odd patterns of variation such as those caused by genotyping errors, we have also performed the analysis excluding any SNPs in ANCs that map to SDs, CNVs, pseudogenes of retroposed genes (n = 610) and observed that the pattern of excess of high frequency derived alleles remains strong and significant (Figure 4A). This overall analysis suggests that recent, possibly positive selection in ANCs has shaped the pattern of nucleotide variation in similar ways as the pattern of fixed nucleotide changes between species.

We then compared the DAF spectrum of SNPs in ANCs with those of SNPs within HARs [18] (n = 84) and accelerated CNSs [19] (n = 328). We observe that SNPs in HARs show an excess of high derived allele frequency, similar to SNPs in ANCs, consistent with recent positive selection, while SNPs in accelerated CNSs of Prabhakar et al. show a pattern more similar to those neutrally evolving (See additional data file 2), indicating once again the heterogeneity of these three sets of accelerated sequences.

Population differentiation of SNPs within ANCs

In order to further characterise the recent evolutionary pressures on ANCs and detect recent population-specific patterns of selection, we calculated F_{ST}, a common measure of population differentiation [29] for SNPs in ANCs and non-accelerated CNCs and compared these two distributions of F_{ST} values. We excluded all SNPs on the X-chromosome, which tend to have higher F_{ST} values due to its lower effective population size [26]. We find that F_{ST} values in ANCs are higher than those for non-accelerated CNCs but at marginal statistical significance (Mann-Whitney-U-test; $P = 0.0504$) (Figure 4B). The signal of higher F_{ST} values in ANC SNPs becomes significant if we then exclude the SNPs in retroposed genes, pseudogenes, SDs or CNVs (Mann-Whitney-U-test; $P = 0.0363$). SNPs from the Pollard and Prabhakar studies do not any statistically significant skew in F_{ST} values (See additional data file 2).

Analysis of ANCs associated with differential gene expression

To assess the functional impact of nucleotide variation in ANCs on phenotypic variation we looked for associations between SNPs from the phase II HapMap [26, 27] within ANCs or *power CNCs* and gene expression levels from the 210 unrelated HapMap individuals using recently generated gene expression data from [30, 31] see Methods). We performed a linear regression between quantitative gene expression values for 14,925 probes and numerically coded genotypes of each SNP within a 10 Mb window centred on the midpoint of each transcript probe. The statistical significance was evaluated through the use of 10,000 permutations performed separately for each gene to give adjusted significance thresholds of 0.0001, 0.001, and 0.01 (Table 2). At these thresholds we find 3, 58 and 458 SNP to gene expression associations for ANCs and 43, 135 and 960 SNP to gene expression associations for *power CNCs*, respectively across all populations. At the 0.01 threshold 16% of the tested ANCs (59 out of 366) contain SNPs that are significantly associated with the expression of a gene, contrasting with only 3% of the tested *power CNCs* (165 out of 5968) (Table 2). This means that a SNP within an ANC is 7 times more likely to be associated with variation in gene expression levels than is a SNP within a *power CNC*, and that nucleotide variation within ANCs is 5 times more likely to be associated with gene expression levels than variation in a *power CNC*. At the most stringent threshold there are 3 genes associated with ANCs: *C13orf7* of unknown function, *SLC35B3* a probable sugar transporter and *RBPSUH* (Recombining Binding Protein SUppressor of Hairless), a J kappa-recombination signal-binding protein.

We further explored the biological properties of the associated genes at the significance threshold of 0.01 by counting the occurrences of each of the Gene Ontology (GO) slim terms associated with these genes. We compared the proportions of genes with and without a GO slim term for ANC associated genes versus those tested with the same counts for *power CNCs* (Fisher’s exact test). Genes associated with ANC variation are deficient for the GO slim term “binding” and enriched for the GO slim term “physiological process” relative to *power CNCs*. Overall, this suggests that ANC nucleotide variation affects expression of different types of genes than does nucleotide variation within *power CNCs* (after controlling for the types of genes that went into the analysis), but the counts are too small to draw specific conclusions about the nature of the effect.

Discussion

We have detected 1356 CNCs that have an accelerated substitution rate in the human relative to the chimpanzee lineage (human ANCs). Misalignment of paralogous sequences is unlikely to explain the overall signal, and manual curation confirms that this only potentially occurs in less than 3% of cases. The lower quality of the other two genomes has minimal effect on the human ANC analysis, since for a substitution to be classified as human-specific, both the chimpanzee and the macaque sequences must have the same nucleotide, and differ from the human nucleotide. We therefore expect this test to be conservative since many chimpanzee-specific substitutions could be sequencing errors, leading to an overestimate of these. The comparison of the human substitution rate in control regions 10Kb or 500Kb from *power CNC*s or the expected human synonymous substitution rate (Ks) to that of the ANCs suggests that 15-19% of the ANCs have not just simply diverged from the sequence of the common ancestor due to loss of constraint but that the rate of divergence has increased 2 to 4-fold above that expected under neutrality, suggesting that they have undergone positive selection.

An interesting possibility is that some ANCs are degenerate regulatory elements associated with subfunctionalized duplicate genes as described in the DDC model [13], or elements that have decayed in function in a similar way to pseudogenes. We found an enrichment of the ANCs within the most recent SDs (less than 2% divergence) relative to both *power CNCs* and non-accelerated CNCs. The general enrichment in SDs is not surprising, as it has been observed that sequence divergence is elevated in duplicated sequences [32, 33]. The most recent SDs in the human genome have occurred after the human-chimpanzee split, and differential evolution between these copies would explain the human-specific acceleration caused by loss of selective constraint due to redundancy or positive selection due to gain of a new function. The DAF analysis suggests that many newly derived alleles within ANCs are undergoing positive selection, but unfortunately there is a paucity of SNPs genotyped within SDs, therefore, insufficient to test this for ANCs in SDs alone. If the signal of ANCs were due to misalignments, we would have observed an excess of ANCs in older and more divergent SDs. We therefore conclude that the recent change in selective forces of some ANCs may be a result of duplication.

The overlap of ANCs with elements such as retroposed genes and pseudogenes is not surprising as these elements are thought to undergo degradation or change when released from the selective constraint placed on active genes. They are, however, more enriched in the *power CNCs* than ANCs. By parallel analysis we demonstrate that our observations are generally robust to inclusion of ANCs in the above elements.

Regions with an excess of SNPs with high DAF relative to the expectations of a neutral equilibrium model are likely to be evolving under positive selection [28]. The DAF spectrum of the ANCs shows an excess of high-frequency derived alleles relative to the DAF spectrum of all control sets. In addition, the observation of higher population differentiation (higher F_{ST} values) in ANC SNPs suggests that ANCs have not only contributed to evolutionary change along the human lineage since the time of the human-chimpanzee common ancestor, but also that some have contributed to recent differentiation between human populations. The *power CNC* set is expected to contain regions that have high substitution rates and also regions with human lineage-specific acceleration that failed to meet the significance threshold for inclusion into the ANC category, or previously fast-evolving regions that have switched selective pressures before the human-chimpanzee split that therefore have similar rates in both human and chimpanzee. This hypothesis is strengthened by the recent Pollard study [18] as 112 out of the 202 HARs overlap the *power CNCs* of the present study. The overlap of 112 HARs with *power CNCs* is not due to low power in our study but mainly due to the fact that our analysis makes the explicit assumption that the human lineage is significantly faster than the chimpanzee, which is not the case for the Pollard study. Interestingly the most significant ANC in our analysis completely overlaps with the most significant element in the Pollard study (HAR1)[34].

We have observed that SNPs within ANCs are significantly associated with gene expression phenotypes and the probability that SNP variation within an ANC being associated is 5-fold higher than for a *power CNC*. The pattern of enrichment in gene expression associations provides our strongest evidence that ANCs contain functionally evolving sequence that is associated with changes in gene expression. There is a tendency for the derived alleles within ANCs to be associated with low gene expression levels though this is not statistically significant. As the derived allele is high in frequency in SNPs within ANCs this could indicate that low expression could be potentially advantageous for some genes but this cannot be tested formally with this dataset due to the small sample size.

The presence of ANCs in the human genome suggests that the evolution of noncoding DNA contributes substantially to species differentiation. Our analysis relies on the identification of these ANCs by initially requiring conservation across multiple vertebrate species, so it is conservative with respect to the contribution of functional non-coding elements to species differentiation. Previous studies have shown that the proportion of functional non-coding sequences can be large and not necessarily conserved above neutral expectation [3]. When additional genomes become available, increasingly rigorous analyses and detection methodologies can be developed to elucidate the degree of non-coding and regulatory evolution and the birth-and-death process of regulatory elements. Nevertheless, the ANCs identified in this study can serve as a baseline for the elucidation of biological processes in non-coding DNA that contribute to species differentiation.

Materials and Methods

Detection of ANCs: alignments and calling of ANCs

CNCs were detected with a phylogenetic hidden Markov model (phyloHMM) [35] and the top 5% of the conserved genome (PhastCons conserved elements, 17-way vertebrate MULTIZ alignment) as available at the UC Santa Cruz browser [36]. The top 5% represents the minimal selectively constrained genome as inferred from the Mouse genome analysis [37].We selected elements of at least 100 bases to increase our power to detect acceleration and intersected those elements with Ensembl gene predictions (v40 - Aug2006) [38] to obtain the set of elements that did not overlap any part of the processed transcript. CNCs with more than 4 substitutions between human and chimpanzee were aligned among human, chimpanzee, and macaque, and lineage-specific substitutions were inferred assuming parsimony. Alignments of these elements were obtained from a 3-way MULTIZ alignment [39] of human finished sequence (hg18), chimpanzee assembly (panTro2), and macaque (draft assembly). The human and chimp genome sequences were aligned with the blastz program [40] with the following substitution scores and penalizing a gap of length k by 600 + 150k.

A	C	G	T
90	-330	-236	-356
-330	100	-318	-236
-236	-318	100	-330
-356	-236	-330	90

For human-rhesus alignments, we used the following and 600 + 130k.

A	C	G	T
87	-226	-129	-255
-226	100	-212	-129
-129	-212	100	-226
-255	-129	-226	87

A three-way alignment of human, chimp and rhesus was computed with the multiz program [39], and searched for intervals of interest (e.g., at least 4 mismatches) using software written for just that purpose.

For the following analysis the human coordinates were mapped from NCBI 36 (hg18) to NCBI 35 (hg17) using the *liftOver* program [41].

Since we are testing for differences in the relative rates of substitution along the lineages, paralogous alignments of duplicates after the (macaque, (chimpanzee, human)) split will not generate a signal since the length of the branches are the same. The only scenario that can generate a false signal is if the duplication occurred before the (macaque, (chimpanzee, human)) split, giving rise to copies X and Y, and the alignment is between the chimpanzee and macaque copy X and the human copy Y. This scenario requires that the human copy X has been lost and that the macaque and chimpanzee copies of Y are either not included in the assembly or have also both been lost. The fact that this requires 3 losses/misses makes the scenario unlikely, and inspection of the data does not suggest that it is occurring.

Fast-evolving non-coding human sequences

We applied the chi-square based relative rate test [17] to detect sequences that are accelerated in either the human or chimpanzee lineage. As this method could potentially be affected by small counts of substitutions we applied the Yates' correction for continuity which is conservative in estimating the p-value of the test. We then selected the threshold that had the lowest FDR in the range of p-values between 0.05 and 0.15. This threshold was P = 0.08 with estimated FDR of 75%, so we subsequently analyzed all human ANCs that has a p-value equal to or less than 0.08. Note that the Yate's correction is generally overcorrecting so our FDR is likely to be an overestimate.

As a control of our ability to detect human accelerated region we compared the relative enrichment of our ANCs and *power CNCs* in those detected as accelerated in humans by alternative methods[18, 19]. Although the tests differ in their approaches (ours for example conditions on human lineage acceleration versus the chimpanzee lineage only) we find a 6-fold enrichment of previously detected accelerated regions (HARs and accelerated CNSs) in our ANC set relative to the *power CNCs* control set.

Due to the lower quality of the chimpanzee and macaque genome sequences relative to the human genome sequence, we only considered sequences accelerated in the human lineage. As a control we also performed alignments of human-chimpanzee-macaque at coordinates 10 and 500 Kb away from the initial CNC coordinates to use as controls for the neutral substitution rate.

Segmental duplications

A set of genomic coordinates corresponding to segmental duplications (SDs) defined by [21, 22] were used as points of reference in the genome. Accelerated, non-accelerated, and *power CNCs* were then mapped to those SDs, and the abundance of ANCs was compared to the observed abundance of non-accelerated or *power CNCs* in SDs as well as the estimated coverage of the genome by SDs (5-6%). CNV genomic coordinates were obtained from the Database of Genomic Variants in Toronto [23].

Pseudogenes and Retroposed genes

Genomic coordinates for retroposed genes and two set of pseudogenes (Yale and Vega annotations available at the UC Santa Cruz browser [36] were used. Accelerated, non-accelerated and *power CNCs* were then mapped to those coordinates and an overlap was defined whenever at least a single base was common between the two sets of features under comparison.

SNPs and F_{ST} values

SNPs from phase I and phase II from release 19 of the HapMap project [26, 27] were mapped from NCBI 34 (hg16) to NCBI 35 (hg17) using the *liftOver* program [41]. SNPs that did not map to hg17 were ignored and derived alleles were inferred based on the chimpanzee alignment to the hg17 version of the human genome. For those SNPs that did not have a reliable chimpanzee alignment, the alignment to the Rhesus macaque was used. Inference of the derived allele was based on parsimony, and the common allelic state between the human and the chimpanzee (or macaque in few cases) was considered the ancestral allele. The derived allele frequency (DAF) was estimated and DAF spectra were compared with the non-parametric Mann-Whitney-U-test. One potential caveat of this analysis is that, because we required the reference human sequence to be quite divergent from the chimpanzee, we have selected a large number of CNCs with an excess of derived alleles by chance, which specifically enriches for SNPs with high DAFs. We find this unlikely since only 4.2% of the fixed differences (281 of the 6660) that produced the signal of acceleration can be explained by the derived alleles of HapMap SNPs in the reference sequence, and this can only increase to approximately 8% if ungenotyped SNPs are accounted for. Therefore, the bulk of the signal for acceleration was independent of the DAFs of the SNPs within the ANCs. The SNP ascertainment does not affect the analysis since we are using both phase I and II SNPs of the HapMap, which together provide a relatively unbiased view of SNP density and allele frequencies. In addition, any potential bias towards genic regions would not create a bias in our analysis since all of the frequency spectra we compare are independent of genes. The phase II HapMap is estimated to contain more than half of the common SNPs in the tested Yoruban (YRI) Hap Map population as has been estimated by the resequenced ENCODE regions [26], so the contribution of SNPs to divergence is not expected to be more than 8%. This together with the comparison to the accelerated sequences at 10 kb and 500kb suggests that small confounding effects of divergence and DAF spectrum are not the reason for our signal. F_{ST} values for each SNP in ANCs and non-accelerated CNCs were calculated according to the Weir and Cockerham [29] method. Distributions of F_{ST} values were compared using the Mann-Whitney-U-test excluding the X chromosome SNPs. This comparison of distributions was repeated with *power CNCs*, and ANCs, excluding any SNPs in SDs, CNV, retroposed genes, or pseudogenes.

Gene Expression Associations

We used gene expression data of 47,294 transcripts in lymphoblastoid cell lines of all 210 HapMap [26] unrelated individuals from the 4 populations, in 4 technical replicates. The gene expression values of 47,294 transcripts interrogated by the array were then normalized and averages taken for each probe across replicates. We downloaded the HapMap [26, 27] genotypes (release 21) for each population of all the phase II SNPs (with a minor allele frequency >5%) within ANCs and *power CNCs*. A linear regression was then performed (separately within each population) between quantitative gene expression values for 14,925 probes (a subset chosen on the basis of sufficient measurable expression levels and variability) and numerically coded genotypes (0, 1, 2) of each SNP within a 10 Mb window centred on the midpoint of each transcript probe. The statistical significance was evaluated through the use of 10,000 permutations performed separately for each gene. In each permutation of a single gene, the most significant p-value was retained, so that there were 10,000 p-values for each gene. From these distributions, for each gene, we determined significance thresholds of 0.0001, 0.001, and 0.01. For each gene tested for association with SNPs in ANCs or *power CNCs* the Gene Ontology (GO) slim terms were tabulated in a non-redundant list (multiple transcripts were removed). For each GO slim term the counts of genes with and without the GO slim term in significantly associated genes (at threshold 0.01) and the total genes tested were compared using 2x2 contingency tables tested by the Fisher's exact test for genes associated with SNPs in accelerated and the *power CNCs*.

Figure Legends

Figure 1 - Substitution rates of 1356 human-specific ANCs.

The relative rates (p-distance) of substitutions of (A) the 1356 ANCs in the human (y-axis) and chimpanzee (x-axis) lineages and (B) the 1145 ANCs excluding those within potential confounding features (SDs, CNVs, pseudogenes, retroposons).

Figure 2 - Venn diagram of overlap between accelerated sequences in the three studies.

The figure shows the overlap between the present study, the Pollard study and the Prabhakar study.

Figure 3 - Segmental duplication divergence in ANCs and CNCs.

The figure shows that the divergence of paralogs in segmental duplications (SDs) where CNCs (red) and *power CNCs* (purple) are found is skewed to high divergence values, while the ANCs (yellow) have a strong enrichment in recent SDs as expected if the acceleration is due to a recent change in selective forces (positive selection or loss of selective constraint).

Figure 4 - Patterns and levels of nucleotide variation in ANCs.

(A) The comparative DAF spectrums for phase II HapMap SNPs in non-accelerated CNCs (n = 48,811), ANCs (n = 682), ANCs outside of SDs, CNVs, retroposed genes or pseudogenes (n = 610), in the two controls (n = 28,408 and n = 28,722) in the *power CNCs* (n = 10,882), in the 60 individuals of the Yoruban (YRI) population. **(B)** The comparative distributions of F_{ST} values for all phase II HapMap SNPs in ANCs (n = 688,) ANCs outside of SDs, CNVs, retroposed genes or pseudogenes (n = 620), *power CNCs* (n = 11,267) and non-accelerated CNCs (n = 52,210).

Tables

Table 1 - Percentage overlap between sets of genomic features with ANCs, power CNCs and non-accelerated CNCs.

	All	SD		CNV		SD or CNV		Pseudogene		Retroposed gene		Pseudogene or Retroposed gene		SD, CNV, Pseudogene or Retroposed gene	
ANC	1356	108	8%	62	5%	138	10%	72	5%	102	8%	111	8%	211	16%
Power CNC	23681	2346	10%	1240	5%	3087	13%	2207	9%	3489	15%	3576	15%	5392	23%
NonAcc CNC	277814	13889	5%	10514	4%	21874	8%	9094	3%	15988	6%	16836	6%	32405	12%

Table 2 - Summary of SNPs within ANCs and power CNCs associated to differential gene expression.

Results for four populations the Yoruba people from Ibadan Nigeria (YRI), US residents with Northern and Western European ancestry (CEU), Han Chinese from Beijing (CHB) and Japanese from Tokyo (JPT).

						No. of significant ANC/CNC to gene associations			No. of significant ANCs/CNCs of those tested					
Population		No. of tested ANCs/CNCs	No. of SNPs	No. of probes tested	No. of associations	0.01	0.001	0.0001	0.01		0.001		0.0001	
CEU	ANC	387	555	8673	23330	77	9	0	59	15%	9	2%	0	0
	Power	6232	8388	14906	350309	181	36	18	149	2%	33	1%	17	0
CHB	ANC	356	499	8092	21291	83	13	0	56	16%	11	3%	0	0
	Power	5737	7579	14893	317518	202	41	15	159	3%	39	1%	15	0
CHB&JPT	ANC	342	466	7919	20163	109	11	1	59	17%	9	3%	1	0
	Power	5474	7162	14852	301636	203	12	1	149	3%	12	0	1	0
JPT	ANC	355	490	8197	21166	88	12	0	59	17%	11	3%	0	0
	Power	5674	7531	14852	315476	241	48	20	194	3%	42	1%	19	0
YRI	ANC	391	583	9118	24310	113	15	2	64	16%	15	4%	2	1%
	Power	6724	9218	14908	381407	196	32	15	173	3%	30	0	14	0

List of Abbreviations

CNC	Conserved Non-coding
ANC	Accelerated Conserved Non-coding
HAR	Human Accelerated Region
CNS	Conserved Non-coding Sequence
DDC	Duplication-Degeneration-Complementation
SNP	Single Nucleotide Polymorphism
UTR	Untranslated Region
SD	Segmental Duplications

CNV	Copy Number Variants
DAF	Derived Allele Frequency
YRI	Yoruba people from Ibadan Nigeria
CEU	US residents with Northern and Western European ancestry
CHB	Han Chinese from Beijing
JPT	Japanese from Tokyo
GO	Gene Ontology

Additional data files

The following additional data files are available with the online version of this paper. Additional data file 1 is a table listing the co-ordinates for ANCs, highlighting those manually checked and overlapping other elements. Additional data file 2 is a figure of the patterns and levels of nucleotide variation in ANCs compared to the alternatively defined fast evolving CNCs.

Acknowledgements

We would like to thank the Rhesus Macaque Genome Sequencing Consortium (RMGSC) and the HapMap consortium for making data available, and the Wellcome Trust for funding ETD, MEH and CPB. CPB is also funded by the Medical Research Council (MRC). WM, ETD, and DJT were supported by the National Human Genome Research Institute (NHGRI). We would also like to thank the Center for Comparative Genomics and Bioinformatics at Penn State University that funded ETD's visit, during which the project was conceived.

References

1. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes**. *Mol Biol Evol* 2003, **20**(9):1377-1419.
2. King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees**. *Science* 1975, **188**(4184):107-116.
3. Dermitzakis ET, Clark AG: **Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover**. *Mol Biol Evol* 2002, **19**(7):1114-1121.
4. Hardison RC: **Conserved noncoding sequences are reliable guides to regulatory elements**. *Trends Genet* 2000, **16**(9):369-372.
5. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM: **Scanning human gene deserts for long-range enhancers**. *Science* 2003, **302**(5644):413.
6. Dermitzakis ET, Kirkness E, Schwarz S, Birney E, Reymond A, Antonarakis SE: **Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment**. *Genome Res* 2004, **14**(5):852-859.

7. Dermitzakis ET, Reymond A, Antonarakis SE: **Conserved non-genic sequences - an unexpected feature of mammalian genomes**. *Nat Rev Genet* 2005, **6**(2):151-157.
8. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET *et al*: **Conserved noncoding sequences are selectively constrained and not mutation cold spots**. *Nat Genet* 2006, **38**(2):223-227.
9. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D: **Ultraconserved elements in the human genome**. *Science* 2004, **304**(5675):1321-1325.
10. Andolfatto P: **Adaptive evolution of non-coding DNA in Drosophila**. *Nature* 2005, **437**(7062):1149-1152.
11. Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, Sabeti P, Chen Y, Stalker J, Huckle E *et al*: **Spread of an inactive form of caspase-12 in humans is due to recent positive selection**. *American Journal of Human Genetics* 2006, **78**(4):659-670.
12. Wang X, Grus WE, Zhang J: **Gene losses during human origins**. *PLoS Biol* 2006, **4**(3):e52.
13. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 1999, **151**(4):1531-1545.
14. Ohta T: **Simulating evolution by gene duplication**. *Genetics* 1987, **115**(1):207-213.
15. Ohta T: **Role of gene duplication in evolution**. *Genome* 1989, **31**(1):304-310.
16. Ohno S: **Evolution by Gene Duplication**. New York: Springer-Verlag 1970.
17. Tajima F: **Simple methods for testing the molecular evolutionary clock hypothesis**. *Genetics* 1993, **135**(2):599-607.
18. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R *et al*: **Forces shaping the fastest evolving regions in the human genome**. *PLoS Genetics* 2006, **2**(10):1599-1611.
19. Prabhakar S, Noonan JP, Paabo S, Rubin EM: **Accelerated evolution of conserved noncoding sequences in humans**. *Science* 2006, **314**(5800):786.
20. Consortium CS: **Initial sequence of the chimpanzee genome and comparison with the human genome**. *Nature* 2005, **437**(7055):69-87.
21. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome**. *Science* 2002, **297**(5583):1003-1007.
22. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S *et al*: **A genome-wide comparison of recent chimpanzee and human segmental duplications**. *Nature* 2005, **437**(7055):88-93.
23. **Database of Genomic Variants [http://projects.tcag.ca/variation/]**.
24. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D: **A distal enhancer and an ultraconserved exon are derived from a novel retroposon**. *Nature* 2006, **441**(7089):87-90.
25. Nishihara H, Smit AF, Okada N: **Functional noncoding sequences derived from SINEs in the mammalian genome**. *Genome Res* 2006, **16**(7):864-874.
26. Consortium IH: **A haplotype map of the human genome**. *Nature* 2005, **437**(7063):1299-1320.

27. **The International HapMap Project [http://www.hapmap.org]**.
28. Fay JC, Wu CI: **Hitchhiking under positive Darwinian selection**. *Genetics* 2000, **155**(3):1405-1413.
29. Weir BS, Cockerham CC: **Estimating F-statistics for the analysis of population structure.** . *Evolution* 1984, **38**(6):1358-1370.
30. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S *et al*: **Genome-wide associations of gene expression variation in humans**. *PLoS Genet* 2005, **1**(6):e78.
31. **GENe Expression VARiation [http://www.sanger.ac.uk/genevar]**.
32. She X, Liu G, Ventura M, Zhao S, Misceo D, Roberto R, Cardone MF, Rocchi M, Green ED, Archidiacano N *et al*: **A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications**. *Genome Res* 2006, **16**(5):576-583.
33. Hurles ME, Willey D, Matthews L, Hussain SS: **Origins of chromosomal rearrangement hotspots in the human genome: evidence from the AZFa deletion hotspots**. *Genome Biol* 2004, **5**(8):R55.
34. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A *et al*: **An RNA gene expressed during cortical development evolved rapidly in humans**. *Nature* 2006, **443**(7108):167-172.
35. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**(8):1034-1050.
36. **UC Santa Cruz Browser [http://genome.ucsc.edu].**
37. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**(6915):520-562.
38. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T *et al*: **Ensembl 2006**. *Nucleic Acids Res* 2006, **34**(Database issue):D556-561.
39. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED *et al*: **Aligning multiple genomic sequences with the threaded blockset aligner**. *Genome Res* 2004, **14**(4):708-715.
40. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ**. *Genome Res* 2003, **13**(1):103-107.
41. **UCSC Genome Bioinformatics [http://genome.ucsc.edu/cgi-bin/hgLiftOver]**.

Figure 1

Figure 1

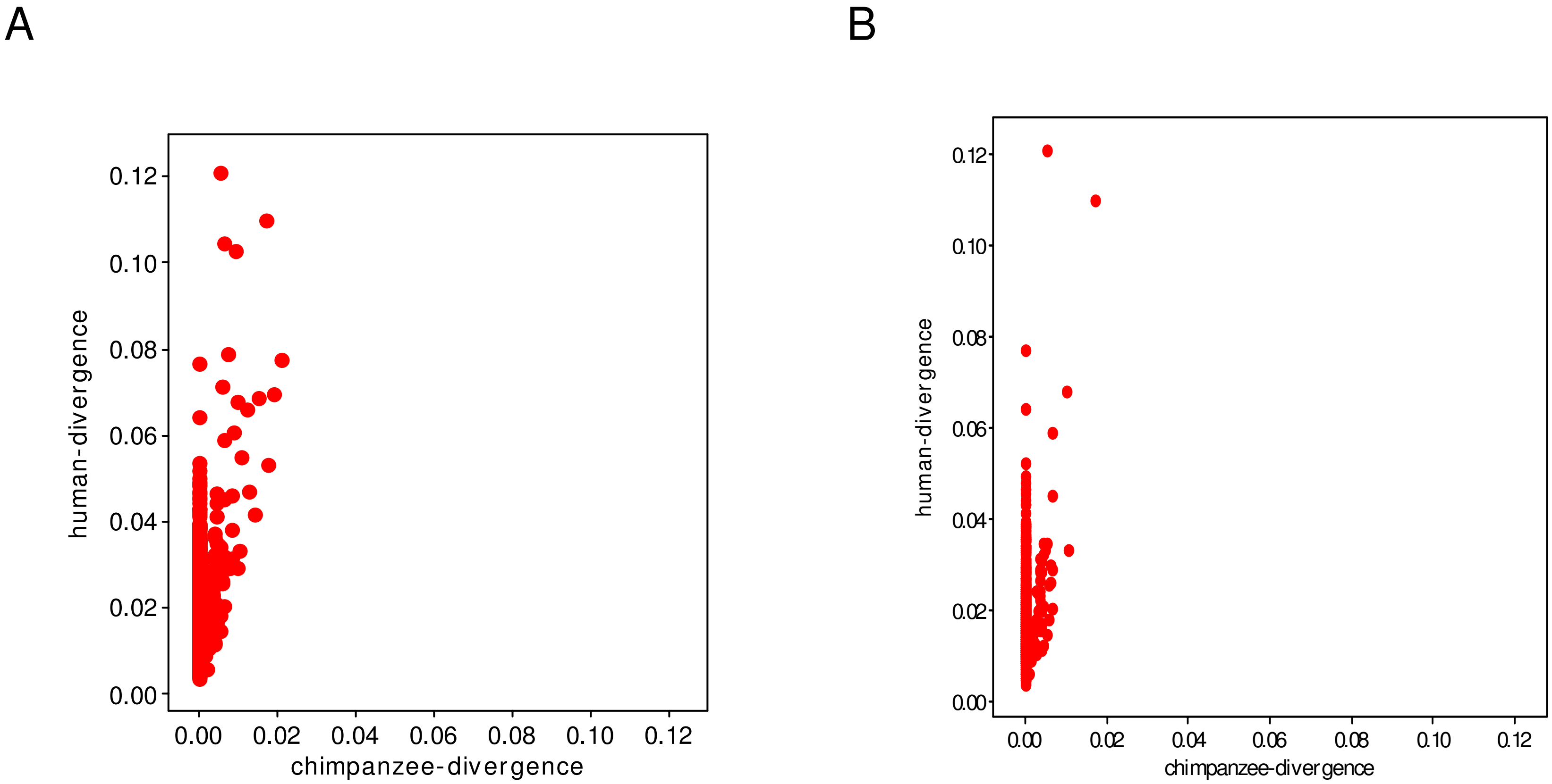


Figure 2

Figure 2

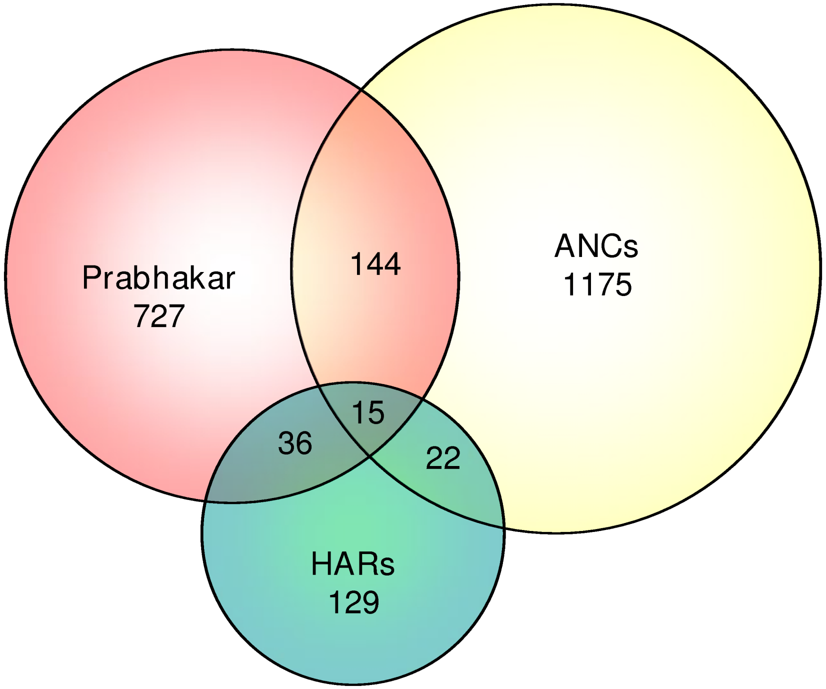


Figure 3

Figure 3

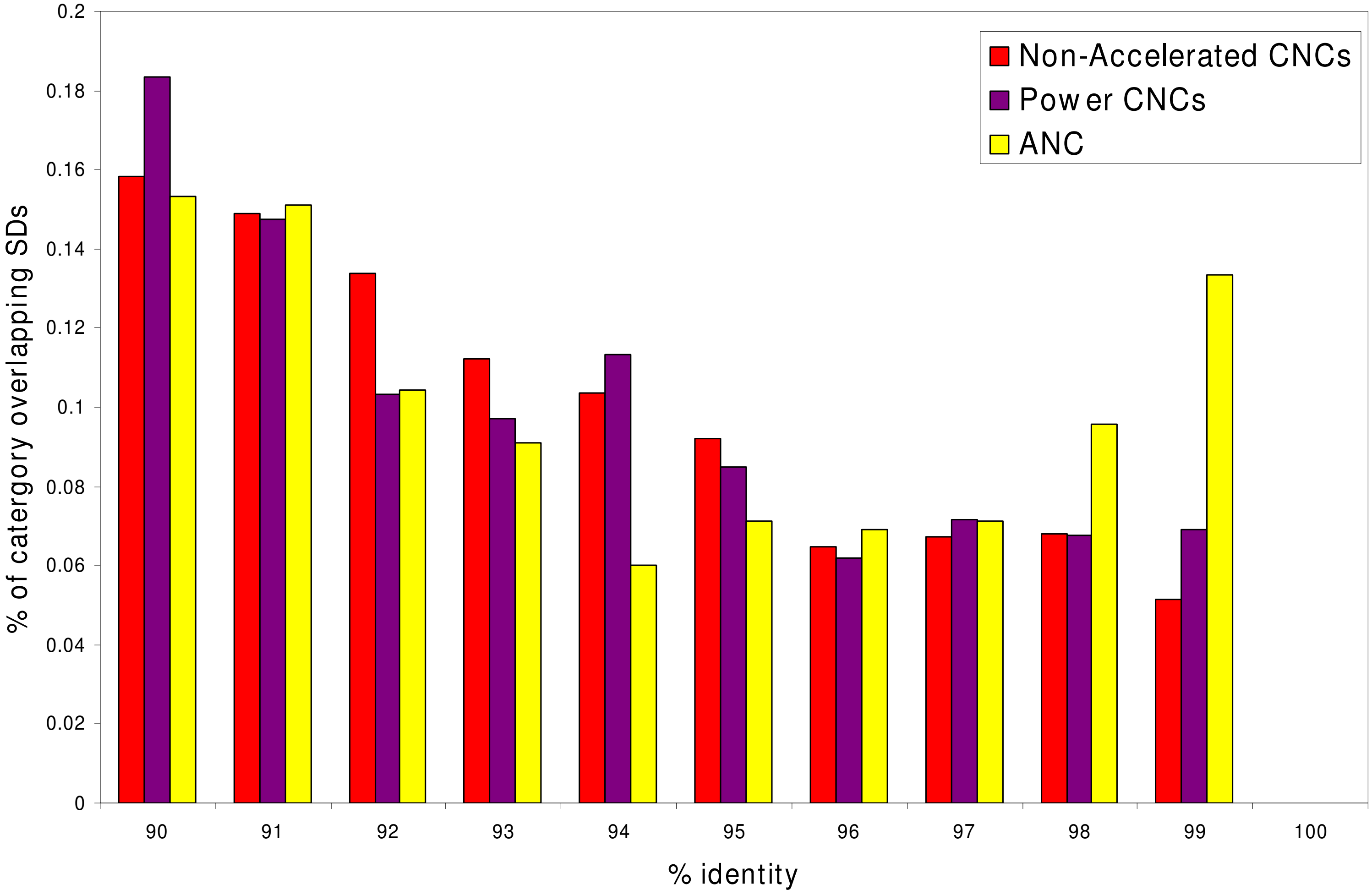
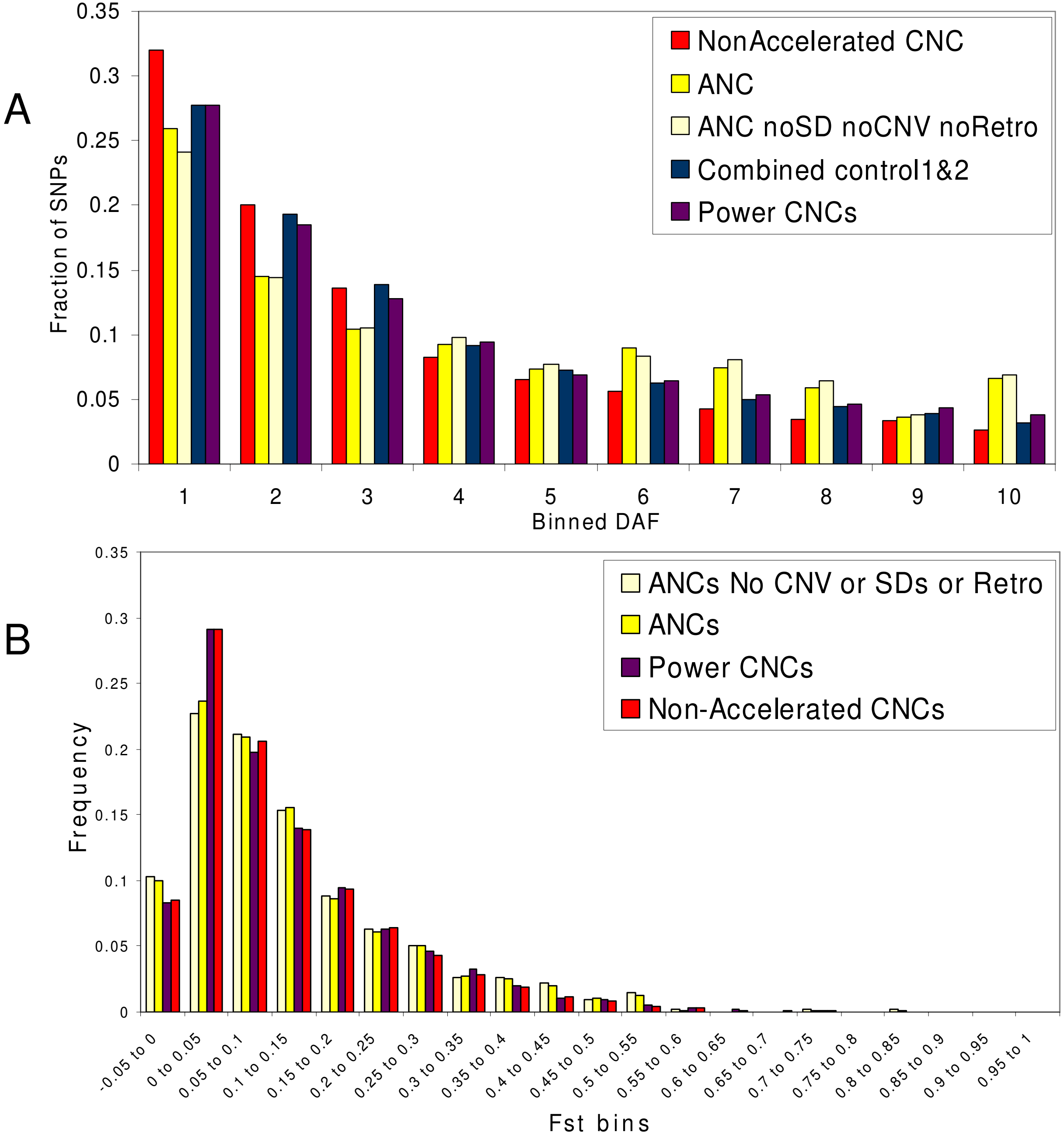


Figure 4

Figure 4



Additional files provided with this submission:

Additional file 1: bird-additional-data-file1.xls, 270K
http://genomebiology.com/imedia/1268645716148614/supp1.xls
Additional file 2: bird-supplementary-figures1.pdf, 22K
http://genomebiology.com/imedia/2886607851486152/supp2.pdf