OPEN ACCESS Freely available online

PLOS ONE

Both Positive and Negative Selection Pressures Contribute to the Polymorphism Pattern of the Duplicated Human *CYP21A2* Gene

Julianna Anna Szabó[1], Ágnes Szilágyi[1], Zoltán Doleschall[2], Attila Patócs[3,4], Henriette Farkas[1], Zoltán Prohászka[1], Károly Rácz[3,5], George Füst[1†], Márton Doleschall[1,3*]

1 3rd Department of Internal Medicine, Semmelweis University, Budapest, Hungary, **2** Department of Pathogenetics, National Institute of Oncology, Budapest, Hungary, **3** Molecular Medicine Research Group, Hungarian Academy of Sciences and Semmelweis University, Budapest, Hungary, **4** "Lendület" Hereditary Endocrine Tumours Research Group, Hungarian Academy of Sciences and Semmelweis University, Budapest, Hungary, **5** 2nd Department of Internal Medicine, Semmelweis University, Budapest, Hungary

Abstract

The human steroid 21-hydroxylase gene (*CYP21A2*) participates in cortisol and aldosterone biosynthesis, and resides together with its paralogous (duplicated) pseudogene in a multiallelic copy number variation (CNV), called RCCX CNV. Concerted evolution caused by non-allelic gene conversion has been described in great ape *CYP21* genes, and the same conversion activity is responsible for a serious genetic disorder of *CYP21A2*, congenital adrenal hyperplasia (CAH). In the current study, 33 *CYP21A2* haplotype variants encoding 6 protein variants were determined from a European population. *CYP21A2* was shown to be one of the most diverse human genes (HHe=0.949), but the diversity of intron 2 was greater still. Contrary to previous findings, the evolution of intron 2 did not follow concerted evolution, although the remaining part of the gene did. Fixed sites (different fixed alleles of sites in human *CYP21* paralogues) significantly accumulated in intron 2, indicating that the excess of fixed sites was connected to the lack of effective non-allelic conversion and concerted evolution. Furthermore, positive selection was presumably focused on intron 2, and possibly associated with the previous genetic features. However, the positive selection detected by several neutrality tests was discerned along the whole gene. In addition, the clear signature of negative selection was observed in the coding sequence. The maintenance of the CYP21 enzyme function is critical, and could lead to negative selection, whereas the presumed gene regulation altering steroid hormone levels via intron 2 might help fast adaptation, which broadly characterizes the genes of human CNVs responding to the environment.

Citation: Szabó JA, Szilágyi Á, Doleschall Z, Patócs A, Farkas H, et al. (2013) Both Positive and Negative Selection Pressures Contribute to the Polymorphism Pattern of the Duplicated Human *CYP21A2* Gene. PLoS ONE 8(11): e81977. doi:10.1371/journal.pone.0081977

Editor: Marc Robinson-Rechavi, University of Lausanne, Switzerland

Received June 23, 2013; **Accepted** October 20, 2013; **Published** November 29, 2013

Funding: The authors acknowledge the financial support from Hungarian Scientific Research Fund (OTKA, PD100648 (AP)), Foundation for the Prevention and Treatment of Fatal Angioedematous Disease and Research, and Technology Innovation Fund, National Developmental Agency (KTIA-AIK-2012-12-1-0010). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: doleschall@med.semmelweis-univ.hu

† Deceased.

Introduction

Duplications of individual genes together with their chromosomal regions have long been considered as the primary source to yield novel gene functions [1]. The models concerning the emergence, maintenance and evolution of duplicated genes rely on positive selection to impart new functions [2] as well as several combinations of relaxed negative selection, neutral processes and functional properties throughout the different phases of fixation [3,4]. However, the vast majority of duplicated genes are disabled shortly after the initial duplication events [5], and consequently there is an enrichment of pseudogenes in duplicated regions [6]. Besides the models of evolutionary fate, two special mechanisms, birth-and-death and concerted evolution, also shape the features and organization of duplicated genes [7]. The birth-and-death process caused by recurrent gene duplications and the subsequent disablement of duplicated genes [8] characterizes multigene families and longer time spans, whereas concerted evolution, where the duplicated genes evolve as a unit and in a non-independent way [9], features single pairs of duplicated genes in the same way as multigene families and shorter time spans. In the model of concerted evolution, newly arisen mutations spread through duplicated regions by recurrent non-

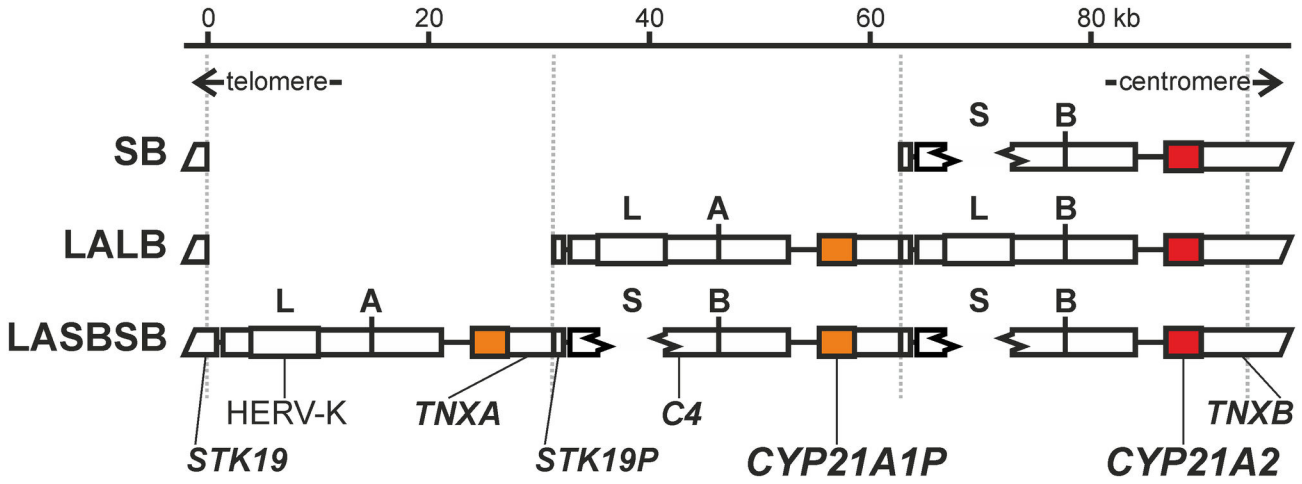

Figure 1. Scaled representation of the organization of human RCCX copy number variation (CNV) depicted by mono-, bi- and trimodular RCCX structure variants. The schematic abbreviations of RCCX structures are indicated on the left side; a module (a repeat) is abbreviated with two letters, the first represents the alleles of HERV-K CNV (L – the long allele or S – short allele [The abbreviation of these alleles comes from the traditional usage of long and short *C4* genes.]), and the second symbolizes the types of *C4* gene (A or B). The duplication of these two letters indicates bimodular RCCX structure, while the triplicate of the two letters means trimodular RCCX structures. Dotted lines indicate the module boundaries, and the directions of the ends of chromosome 6 are indicated by arrows under the scale bar. The variable region of bimodular RCCX contains two pairs of full-length genes, complement component 4 (*C4A* and *C4B*), steroid 21-hydroxylase (*CYP21A1P* and *CYP21A2*), and two pairs of a functional gene and a truncated pseudogene, serine/threonine kinase 19 (*STK19* and *STK19P*) and tenascin-X (*TNXA* and *TNXB*). The illustrated region spans from the telomeric end of exon 4 of *STK19* to the centromeric end of exon 28 of *TNXB*.
doi: 10.1371/journal.pone.0081977.g001

allelic (ectopic, interlocus, interparalog, interparalogue) gene conversion, which maintains sequence homogeneity among duplicated genes [10].

The human steroid 21-hydroxylase gene (*CYP21A2*) is responsible for cortisol and aldosterone biosynthesis in the adrenal glands. *CYP21A2* belongs to the *CYP* multigene family, but shows relatively low homology with other members of the family [11]. *CYP21A2* is not located in the large *CYP* gene clusters, but resides a multiallelic, complex and tandem copy number variation (CNV) of the major histocompatibility complex region [12], called RCCX CNV [13,14]. The multiallelic CNVs consist of at least one CNV allele harboring duplicated regions [15], and are regarded as relatively new duplications going through the polymorphic phase of fixation in populations [16]. The most frequent CNV allele of RCCX CNV is bimodular (duplicated), but monomodular and trimodular CNV alleles are also prevalent in humans (Figure 1) [17]. A haplotypic bimodular RCCX structure (CNV allele) encompasses two duplicated pairs of complete genes, the *CYP21* genes, and the complement component 4 (*C4*) genes. In addition to these, the RCCX CNV has a quite complicated organization, and is discussed elsewhere in detail [13].

A haplotypic RCCX structure usually contains one functional gene in the centromeric, 3'-module and zero, one or two disabled pseudogenes (*CYP21A1P*) in the modules towards the telomeric, 5'-direction (Figure 1). Human *CYP21* paralogues (in general, two homologous genes in a duplicated region, in the case of *CYP21* gene, functional gene and pseudogene(s)) show 97-98% nucleotide identity [18,19], and the partial sequences of the *CYP21* paralogues in great apes are more similar to each other than to the orthologues (homologous genes in the same module of a duplicated region, but in different species), indicating sequence homogenization and concerted evolution [20]. The non-allelic gene conversion in the background of the concerted evolution of *CYP21* paralogues is well studied in humans, because congenital adrenal hyperplasia (CAH), a frequent Mendelian disorder mostly caused by gene conversion from *CYP21A1P* to *CYP21A2*, is a focus of human geneticists [21,22]. However, intense conversion activity has been observed at meiotic (equal) crossover hotspots [23], but RCCX CNV is far away from these hotspots, and a low meiotic (equal) crossover rate characterizes its neighboring genomic region [24,25].

Despite the multitude of population genetics literature concerned with the evolution and selection of non-duplicated genes, there are relatively a few genetic studies based on experimental data dealing with the selection and unique evolution processes in duplicated genes (for example: [26-33]). This paucity applies to the duplicated genes of CNVs to a greater extent (for example: [34-36]). Furthermore, the complex organization of RCCX CNV renders the experimental methods more difficult [13], and therefore a population genetics analysis has never been conducted to reveal the polymorphism pattern of *CYP21A2* and the selection forces acting on it. We assumed beforehand that concerted evolution applied to the entire *CYP21A2*, and the polymorphism pattern in the context of the

evolution of *CYP21A2* haplotypes was examined. In addition to the human population dataset of *CYP21A2* haplotypes, a human *CYP21A1P* polymorphism dataset and a great ape *CYP21A2* and *CYP21A1P* sequences dataset were also compiled and analyzed.

Materials and Methods

Subjects

Genomic DNA samples were collected from 36 healthy, unrelated Hungarian subjects with European ancestry. The study protocol was approved by the Institutional Ethics Committee of Semmelweis University (local ethics committee) and was executed according to the Declaration of Helsinki principles. The subjects gave written informed consent.

Determination of RCCX CNV structures and *CYP21A2* haplotypes

The two haplotypic RCCX CNV structures of a DNA sample were determined principally in the same way as described recently [13]. Briefly, a set of allele-specific long-range (ASLR) polymerase chain reactions (PCRs) [13], *C4*-type-specific quantitative PCRs (qPCR) [37], and HERV-K CNV allele-specific qPCR [38] were applied to the dissection of RCCX CNV. In addition to these published methods, a *CYP21*-type-specific qPCR was developed for the direct gene copy number determination of *CYP21A1P* and *CYP21A2*. The forward primers were *CYP21*-specific and the same as the primers applied to the nested-PCR (see below). The reverse primer was not allele-specific, and the Taqman minor groove binder probe (Applied Biosystems) was labeled with fluorescent dye 6-FAM (see primer and probe sequences in Table S1). *RPPH1* labeled with dye VIC was used as an endogenous reference (RNase P reference assay, Applied Biosystems). Multiplex reactions were carried out in an Applied Biosystems 7500 Fast Real-Time PCR machine by TaqMan Fast Universal PCR master mix, and the reaction conditions were almost the same as that in the manufacturer's protocol, except the annealing temperature was 64 °C. To check the reliability of the Taqman based *CYP21* gene copy number assay, the *CYP21* qPCR primers were tested on some samples in a LightCycler 1.0 qPCR machine (Roche) by LightCycler FastStart DNA Master SYBR Green I mix using *B2M* and *HMBS* genes as references (Table S1).

To determine the *CYP21A2* haplotypes, *CYP21*-specific nested PCRs were performed from particular ASLR-PCR products as described [13]. Using these nested PCR products, whole-gene haplotypes and genotypic alleles of full-length *CYP21A2* on the *CYP21A2* locus were sequenced mostly as described [13], but with some modified sequencing primers. The 5'-forward and the 3'-reverse sequencing primers (SEQ_11F and SEQ_21R) were moved to the ends of nested PCR products to extend the double-sequenced region to 3357 bp (GenBank NT_007592.15: 31946070-31949426). Two primers (SEQ_15F and SEQ_23F) which lie on some polymorphic sites of *CYP21A1P* were changed to ones avoiding the polymorphisms in *CYP21A1P*, and, finally, the SEQ_12F primer was changed to one which could cover the 5'-end of intron 2 in *CYP21* genes (Table S1). The sequence calls of capillary sequencing were assembled with CLC DNA Workbench v6.5 (CLC bio) and were inspected manually by two different operators.

Because the haplotypic RCCX structures and *CYP21A2* haplotypes could not be experimentally determined from all diploid combination of RCCX structures, bioinformatic haplotype reconstruction by PHASE software v2.1.1 [39,40] was applied to resolve the experimentally indeterminable haplotypic RCCX structures and *CYP21A2* haplotypes from genotypic data [13]. *CYP21A2* haplotype sequences encompassed the whole gene, 122 bp of 5'-flanking region (FR) and 7 bp of 3'-FR. To summarize, a human population dataset of RCCX structure-*CYP21A2* haplotypes was generated from 36 unrelated European subjects by a combined molecular and inferred haplotyping approach.

Sequence and polymorphism data of *CYP21* genes from great apes

The sequences of human *CYP21A2* haplotype variants were used from a recent study [13], and full-length *CYP21* sequences and intergenic sequences between *C4* and *CYP21* genes (100 bp from the 3'-end of *C4* genes and 400 bp from the 5'-end of *CYP21* gene to avoid the described and potential gene regulatory sites) from human leukocyte antigen (HLA) homozygous cell lines [41] (Table S2). Primate *CYP21* sequences were collected from GenBank, and the status of gene (functional gene or pseudogene) was assessed (File S1). Human *CYP21A1P* polymorphism data was obtained from a previous study (The studied population was German [42], but Hungarians barely deviate from the German population, the European reference population (CEU) or the majority of European populations based on genome-wide polymorphisms [43,44].), but the segregating sites below 0.01 of minor allele frequency were excluded to preclude the possibility of bias, as suggested recently [45]. The polymorphism dataset from the most 3'-end of *CYP21A1P* (NT_007592.15: 31916635-31916691) is lacking, hence the *CYP21A1P* sequences from HLA-homozygous cell lines and a partial *CYP21A1P* sequence (GenBank: KC493621) were checked, but segregating sites were not found in this short segment. To summarize, a great ape *CYP21* sequence dataset and a human *CYP21A1P* polymorphism dataset were compiled.

Sequence and population data analyses

All alignments were assembled by ClustalX2 v2.0.5 [46], and were edited by CLC DNA Workbench v6.5 and MEGA v5.05 [47]. The Hardy-Weinberg equilibrium and Slatkin's linearized fixation index were calculated by Arlequin v3.5 [48]. Pearson's chi-square (χ^2) and continuous Kolmogorov-Smirnov (KS) tests were calculated under the assumption of independent sites and events by STATISTICA 8 (Statsoft). Power for the χ^2 test was performed with G*Power v3.1.3 [49]. The MEGA v5.05 program [47] was used to construct neighbor-joining (NJ) trees with complete gap deletion and maximum likelihood (ML) trees with gap sites under the assumption that the substitutions followed the Jukes-Cantor (JC) model and uniform rates among sites. The bootstrap tests of NJ and ML trees were performed with

1000 bootstrap replications. Spatial heterogeneity in the phylogenetic signal was detected with the BootScan algorithm built in SimPlot v3.5.1 [50], using a 300 bp window size, 30 bp step size, NJ tree based on the JC model and 1000 bootstrap replications. The Robinson–Foulds tree distance metric [51] was calculated by TOPD/FMTS v3.3 software [52] in order to characterize the similarity among the tree topologies of great ape paralogues and orthologues (The metric skips branch length and bootstrap value, and it evaluates only topology.). Only one human paralogous sequence pair (DBB sequences) was taken into account, because the ambiguous relationships among human *CYP21* sequences were confounding factors for this analysis (It should be noted that the genealogical network is suitable for the analysis of intraspecific sequences with high nucleotide identity [53].). 3Seq [54], BootScan [55], GENECONV [56], MaxChi [57] and RDP [58] algorithms implemented in RDP v3.44 software [59] were chosen to detect potential gene conversion events. The sequences of human *CYP21* genes having high identities were removed following the instructions of RDP v3.44, and a human CYP21 sequence dataset (h08, h21, h30, h35, h41, h42, h44 (QBL), h47, h57, h58, *CYP21A1P* DBB and *CYP21A1P* PGF above a nucleotide difference threshold of 7) and a great ape *CYP21* sequence dataset (h13 (MCF), h28 (COX), h37 (DBB), h44 (QBL), *CYP21A1P* DBB, *CYP21A1P* PGF and all of the chimpanzee, gorilla, orangutan and macaque *CYP21* sequences above a nucleotide difference threshold of 8) were compiled. DnaSP v5.10.01 [60] was applied to perform sliding window analyses using a 300 bp window size and a 30 bp step size. This program was also utilized for assessing the expected haplotype heterozygosity (HHe), the ratio of the nonsynonymous and synonymous average number of pairwise nucleotide differences (π_A/π_S), linkage disequilibrium (LD) and neutrality tests; Tajima's D test [61], Fu's Fs test [62] and Fay and Wu's H test [63] were used with the macaque *CYP21A2* sequence as an outgroup if it was needed (H test relies on an outgroup.). Other neutrality tests, normalized Fay and Wu's H (nH) test [64], Ewens-Watterson (EW) test [65] and the rejection probability of EW test were all carried out using the macaque outgroup by DH software (http://zeng-lab.group.shef.ac.uk). The rejection probabilities of D, Fs, H and nH tests under neutral model and European demography for a 5% significance level were calculated by mlcoalsim v1.42 software [66] based on a previous version of ms software [67]. The rejection probability criteria in DH and mscoalsim softwares were different [65], and the criterion of DH was used. The demography model of European population contained a bottleneck with an effective population size of 1861 from 51 thousand years to 21 thousand years before the present and an expansion from 21 thousand years (about from the retreat of the last glacial age) to the present [68-70]. The gene conversion events were also taken into account omitting the sites of minimum tract length [13] of statistically evident gene conversion events in a similar way to that was described previously [36]. The ratio of the number of nonsynonymous changes per site to the number of synonymous changes (Ka/Ks) and the McDonald-Kreitman test [71] were performed using the gorilla *CYP21A2* sequences (Table S2) by DnaSP v5.10.01 (Hence there were more functional *CYP21A2* coding sequences in gorillas [72] (File S1).).

Results

Reconciling the concerted evolution of *CYP21* genes in great apes

The concerted evolution of orangutan *CYP21* paralogues has been verified by a phylogenetic tree (paralogues resemble each other better than their orthologues) based on partial sequence data [20], hence this phenomenon was tested on the full-length sequences of the great ape *CYP21* sequence dataset. Agreeing with this, the concerted evolution of orangutan paralogues was unequivocal in both the NJ tree and the ML tree, but the *CYP21A2* sequences of human and chimpanzee were separated from *CYP21A1P* sequences with high bootstrap values (Figure 2A, Figure S1 and S2). Gorilla paralogues resided in their own clade with a moderate bootstrap value, also indicating concerted evolution, and the intercalating position of the gorilla clade could be explained by the incomplete lineage sorting of human, chimpanzee and gorilla [73]. The average nucleotide identities of *CYP21* orthologues and paralogues reflected the relationships of phylogenetic trees (Table S3). For instance, the average nucleotide identity of human and chimpanzee *CYP21A2* sequences was slightly higher than that in human paralogous *CYP21A2* and *CYP21A1P* sequences (98.42% against 97.71%), and, vice versa, *CYP21A1P* orthologues in humans and chimpanzee were slightly closer to each other than chimpanzee *CYP21* paralogues (97.85% against 97.09%).

Because concerted evolution is based on homogenization by gene conversion, the unbalanced sequence transfers between the sections of *CYP21* paralogues could lead to subregions with different evolutionary histories. The great ape alignment was scanned by a BootScan algorithm to dissect whether the spatial heterogeneity of the phylogenetic signal (phylogenetic signal is the likelihood of closely related sequences to resemble each other more than random sequences of the same phylogenetic tree) stayed in the background of straggly phylogenetic relationships between *CYP21* sequences. Surprisingly, the most equivocal phylogenetic signal of human *CYP21A2* was located in a narrow region around intron 2 and was related to the human and chimpanzee *CYP21A2* orthologues indicating that there was no homogenization in this region, whereas the expected signs of homogenization between human paralogues were weaker, and the peaks of the signal were dispersed (Figure 3). The similarity of this phylogenetic signal pattern was demonstrated between the *CYP21A1P* orthologues of human and chimpanzee, and between chimpanzee paralogues; the *CYP21A1P* orthologous signal was very high around intron 2, but the chimpanzee paralogous signal ceased at this section. The signals of the paralogues showed fragmented lines with high peaks in each great ape species (Figure 3, Figure S3). These paralogous signals had different spatial patterns as compared to each other implying the different histories of transitions, but the loss of signal in paralogues around intron 2 proved to be a common

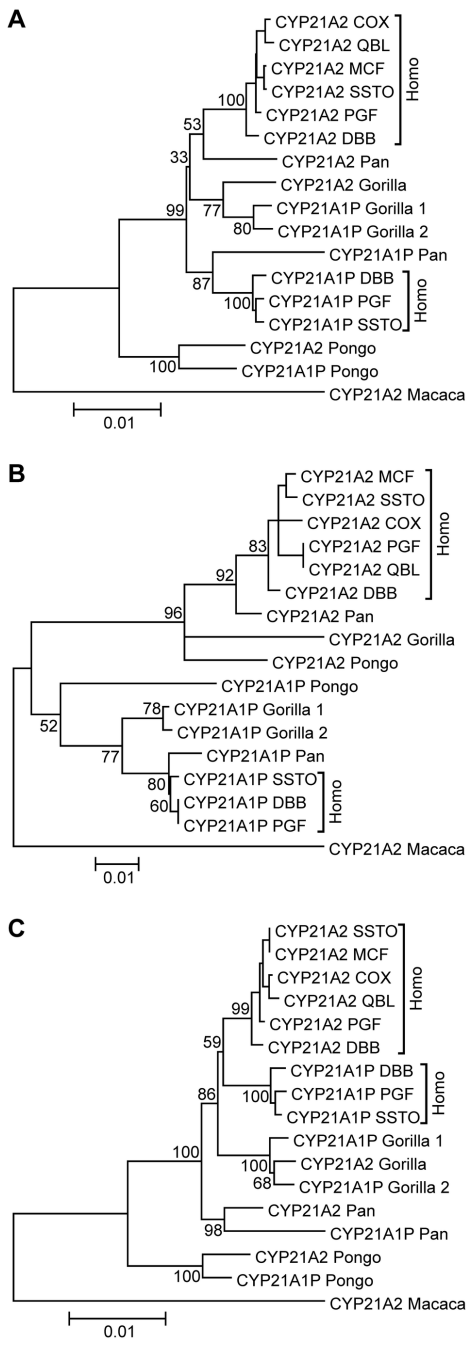

Figure 2. Rooted maximum likelihood phylogenetic trees constructed from selected full-length sequences, intron 2 sequences and *CYP21* gene sequences without intron 2 of the great ape *CYP21* sequence dataset. The names of *CYP21* sequences of HLA-homozygous cell lines available from public databases are represented at the human sequences. Bootstrap values are shown next to the corresponding nodes, but the values within human clades are not presented for clarity. The scale bar indicates genetic distance. (A) Rooted ML tree of great ape *CYP21* full-length gene. (B) Rooted ML tree of great ape *CYP21* intron 2. (C) Rooted ML tree of great ape *CYP21* genes without intron 2.

doi: 10.1371/journal.pone.0081977.g002

feature, and therefore the homogenization was absent from all intron 2 of the great ape *CYP21* genes.

The full-length *CYP21* sequences were divided into two regions based on the spatial distribution of phylogenetic signals; one covered intron 2 and the other encompassed the remaining gene parts. The CYP21A2 and *CYP21A1P* clades were completely separated in the ML tree of intron 2 sequences, showing no signature of concerted evolution or homogenization between paralogous sequences, and the tree followed the model of divergent evolution (Figure 2B). The species in the *CYP21A1P* clade evolved according to the accepted lineage sorting of great apes [73], and only the positions of gorilla and orangutan sequences were vaguely resolved in the *CYP21A2* clade. In the ML tree of the sequences without intron 2, the paralogues from each species were located in separated clades, confirming concerted evolution in each of them (Figure 2C). In addition, the tree topologies of *CYP21* paralogues and orthologues were compared by the Robinson-Foulds metric in ML trees of the full-length gene, the intron 2 subregion and the gene without intron 2. The scale of the metric ranges from 0 to 1 (from complete concordance to complete dissimilarity). The calculated Robinson-Foulds metric was 0.57 for both whole genes vs intron 2 and the whole gene vs gene without intron 2 trees, indicating the medium similarity, whereas the metric was 1 between the trees of the intron 2 and the gene without intron 2, demonstrating the complete lack of similarity (Robinson–Foulds metric range was 0.94-0.95 with 95% confidence interval (CI) of 0.08-0.09 for 100 randomized trees).

To verify the validity of concerted evolution for the coding (cds) region and the non-coding region (non-cds) without intron 2, we further divided the sequences without intron 2 into cds and non-cds subregions, and phylogenetic analyses were performed on them. The clades of different great ape paralogues were separated from each other in both ML trees, however, some discrepancy appeared in the lineage sorting of the gorilla clade (Figure S4). The Robinson-Foulds metric was 0.86 between the cds subregion and the intron 2 subregion, demonstrating very low similarity, while the metric was 1 between the non-cds without intron 2 subregion and intron 2 subregion suggesting complete dissimilartity (for 100 random trees: 0.96 CI: 0.07-0.09). Furthermore, the ML tree of the intergenic region between *C4* and *CYP21* genes also showed the perfect sign of concerted evolution (Figure S4) supporting the idea that the homogenization occurred outside the genes of RCCX CNV as well as inside the *CYP21* genes. The Robinson-Foulds metric was 0.29 for the cds vs. non-cds, the cds vs intergenic and non-cds vs ingergenic trees (for 100 random trees: 0.93-0.96 CI: 0.08-0.09) indicating high concordance. The lack of complete concordance presumably came from the different positions of the gorilla clade and the positions of gorilla paralogues relative to each other.

Search for statistical evidence of the gene conversion in the *CYP21* sequence

We attempted to reveal the potential gene conversion events using several algorithms implemented in the RDP v3.44 software. Two non-allelic conversions confirmed by more

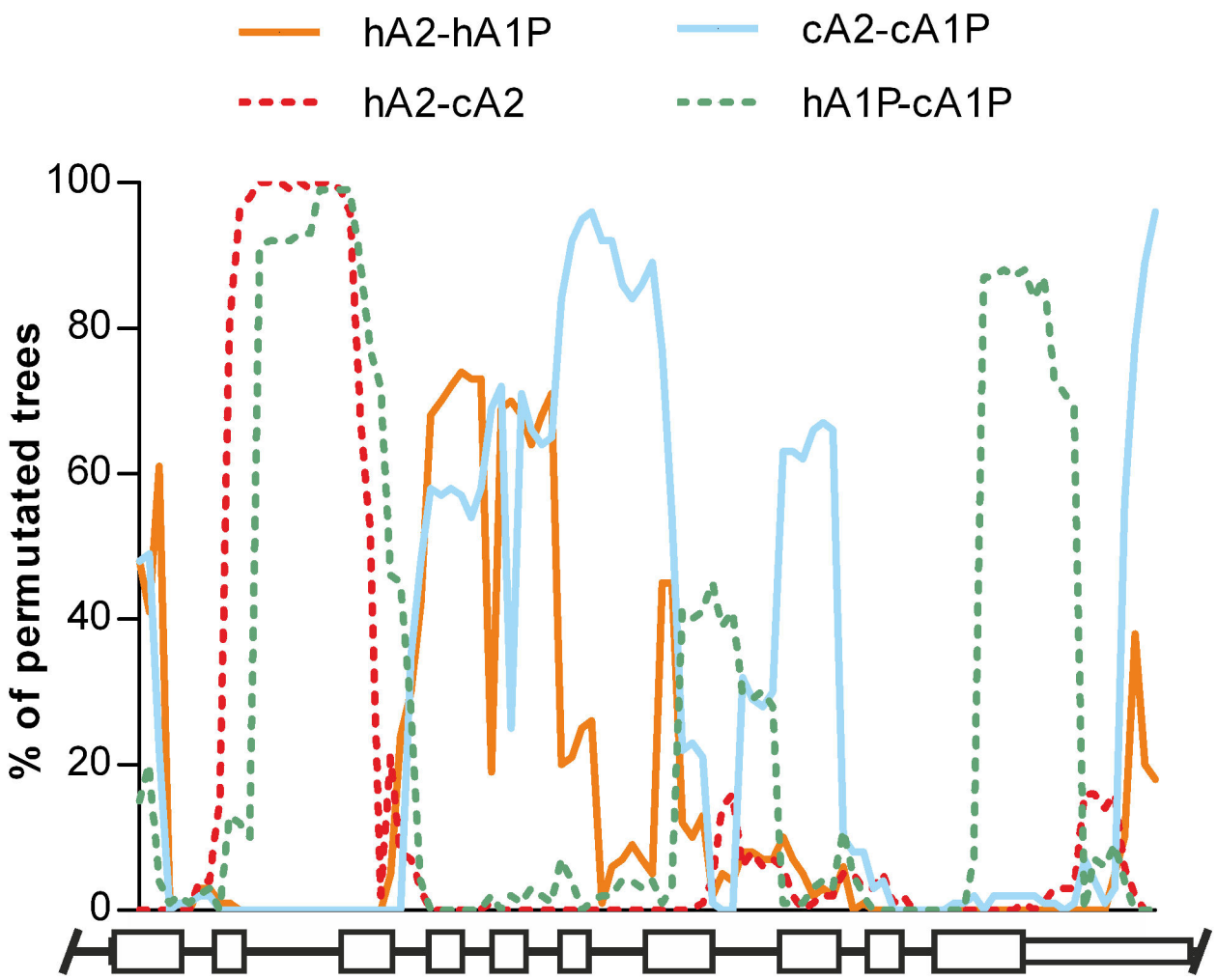

Figure 3. Spatial distributions of phylogenetic signals derived from the different orthologous and paralogous pairs of the human and chimpanzee full-length *CYP21* sequences. Blue line indicates the phlyogenetic signal of human paralogues (hA2-hA1P), red dashed line indicates *CYP21A2* orthologues (hA2-cA2), orange line indicates chimpanzee paralogues (cA2-cA1P) and green dashed line indicates *CYP21A1P* orthologues (hA1P-cA1P). The likelihood of closely related sequences to resemble each other more than random sequences of the same phylogenetic tree is expressed by '% of permutated trees' in y axis. Schematic *CYP21* genes are indicated below the plot, high white boxes symbolize the exons, low white boxes represent the untranslated regions, and black lines indicate the introns and flanking regions.

doi: 10.1371/journal.pone.0081977.g003

algorithms were detected below the 0.05 significance level in the human *CYP21* sequence dataset for RDP v3.44, but their exact breakpoints were undetermined. One (from *CYP21A1P* DBB to h42, GENECONV p=0.0453, confirmed by 3Seq p=0.0271) was nearly identical with an event of the 3'-untranslated region (UTR) covering the alleles of six adjacent segregating sites on the converted sequence (site 3080, 3102-3186) described recently [13]. The other (from *CYP21A1P* to h41, BootScan, p=0.0182, confirmed by GENECONV p=0.0022, 3Seq p=0.0047) was also very similar to a described event of intron 2 spanning four adjacent sites (site 624-634). In the great ape *CYP21* sequence dataset for RDP v3.44, there was no detected gene conversion from chimpanzee, gorilla, orangutan and macaque *CYP21* sequences to human *CYP21* sequences, and there was an nearly identical non-allelic conversion event of 3'-UTR (site 3080, 3102-3186) without clear breakpoints (from *CYP21A1P* PGF to h37, BootScan p=0.0206, confirmed by 3Seq p=0.0037). The parallel observations of these highly similar gene conversion events in the 3'-UTR originated in one gene conversion event, according to the *CYP21A2* genealogical network [13].

Spatial distribution of *CYP21* polymorphisms and related indexes

A total of 33 different *CYP21A2* haplotype variants were observed in the human population dataset, including the determination of a new *CYP21A2* haplotype variant (h62, GenBank: KC493622; Table S4). These 33 haplotype variants encoded 6 different protein variants, which consisted of amino acid changes caused by the alleles of 4 segregating sites. If the deletion mutation (rs61338903) affecting three adjacent nucleotides and causing the loss of one amino acid in exon 1 was considered (The deletion mutations in the human population dataset of *CYP21A2* haplotypes were not examined, because they would violate the majority of population genetic models that form the background of the population genetic analyses.), the number of protein variants would not change (Table S5). The validity of experimental data on haplotypic RCCX CNV structures and *CYP21A2* haplotypes was checked, and only the segregating sites of the human population dataset containing a total of 64 *CYP21A2* haplotypes could be consider as segregating sites, because other sites and site frequencies, which have been disclosed in a recent study [13], did not derive from a homogeneous population (File S1, Table S6).

The association of phylogeny with the different regions might reflect the uneven distribution of polymorphic sites observed along the gene (Figure 4). Hence we further investigated the spatial distribution of the polymorphic sites and related indexes using sliding window analysis. The distribution curves of the number of polymorphic sites (S) and the average number of pairwise nucleotide differences (π) calculated from the human population dataset of *CYP21A2* haplotypes ran parallel to each other (Figure 5A). The curves fluctuated, had a jutting peak around intron 2, and the polymorphic sites were not distributed uniformly along the *CYP21A2* gene (KS: $p<0.01$, χ^2: $p<0.0001$). For the assessment of duplicated genes, the classification of polymorphic sites based on their coexistence in paralogues is a routine procedure [74]. Therefore, the sites from the datasets of human *CYP21A2* haplotypes and human *CYP21A1P* polymorphisms were classified into four types: *CYP21A2*-specific and *CYP21A1P*-specific polymorphic sites, at which polymorphisms were observed in either of the two genes; shared sites, at which polymorphisms were shared by the two paralogues; and fixed sites, at which each paralogue had different fixed alleles (Table S7) [75]. The curves of *CYP21A2*-specific and fixed sites had a peak around intron 2 that was much larger than the fluctuating baseline, while the curves of *CYP21A1P*-specific and shared sites did not have any sharp peaks (Figure 5B). Agreeing with this, the distribution of *CYP21A2*-specific and fixed sites deviated significantly from uniform distribution (both sites: KS: $p<0.01$, χ^2: $p<0.0001$), whereas the distribution of *CYP21A1P*-specific and shared sites did not deviate significantly (KS: non-significant (ns), χ^2: $p=0.2730$ and KS: ns, χ^2: $p=0.1009$). Some neutrality tests detecting skewness from the neutral distribution of allele frequencies in segregating sites [76] were also calculated from the population dataset of human *CYP21A2* haplotypes, and plotted (Figure 5C). The curve of Tajima's D fluctuated around the null point along the full gene, but the curve of Fay and Wu's H dropped sharply around intron 2, indicating that the excess of high-frequency polymorphisms, and not just the excess of *CYP21A2*-specific and fixed sites, characterized this subregion.

Genetic features of the whole *CYP21A2*, intron 2, coding and non-coding without intron 2 subregions

Taking into account the results of the phylogenetic analyses and the spatial distributions of various genetic characteristics in *CYP21A2*, we further used the three subregions; an intron 2 subregion, a cds subregion and a non-cds without intron 2 subregion, and assessed the population genetic features of both the whole gene and the subregions using the population dataset of human *CYP21A2* haplotypes. From all 44 polymorphic sites, 12 sites were observed in intron 2, 21 in the non-cds without intron 2 subregion and 11 in the cds subregion (Table 1). Haplotype heterozygosity was close to its maximum value in the whole gene (HHe=0.949, standard deviation (sd)=0.014), and its values were also high in the different subregions, but the value in intron 2 was only slightly larger than those in the cds and non-cds subregions. As expected from the spatial distribution of the average number of pairwise nucleotide differences, intron 2 and two other subregions showed different nucleotide diversity levels (π), and intron 2 was about 4 times more diverse than the remaining part of the gene. The occurrences of the fixed, shared and *CYP21A2*-specific sites in the subregions highly deviated from each other. This stunning difference in the numbers of sites of the three site classes were highly significant between intron 2 and the remaining part of the gene (χ^2: $p=0.0055$, power=1.0000), and the differences between the subregions were also statistically significant (intron 2 vs. non-cds without intron 2, χ^2: $p=0.0106$, power=1.0000, intron 2 vs. cds, χ^2: $p=0.0091$, power=1.0000, cds and non-cds without intron 2, χ^2: $p=0.0423$, power=0.9989).

In neutrality analyses, the selective neutralities of the whole gene and different subregions were first examined using Tajima's D test. The D test gave no significant values, but there was deviation between the values of the different subregions (Table 1). Rejection probabilities under both neutral and European population demography models were investigated (Table S8). To consider the gene conversion, an additional dataset of human *CYP21A2* haplotypes was generated by omitting the segregating sites affected by statistically evident gene conversion events, and rejection probabilities were also examined. In contrast to D test, Fu's Fs test and the normalized H test, based on the rejection probabilities corrected with demography, produced significant evidence for the non-neutral distribution of the observed allele frequencies in the whole gene (Table 1). The EW test was also significant under neutral model of the DH software (the EW test was also significant under Slatkin's implementation of neutrality [77]). For the subregions, the P-values of the Fs test under demography model could be accepted as significant in all subregions, while P-values of the H and nH tests were highly significant only in the intron 2 subregion. Rejection probability values of the EW test were significant only in the non-cds without intron 2 subregion. The significances of rejection probabilities in the dataset of omitted sites affected by gene conversion were identical to those of the human population dataset of *CYP21A2* haplotypes with the demography model except for the H and

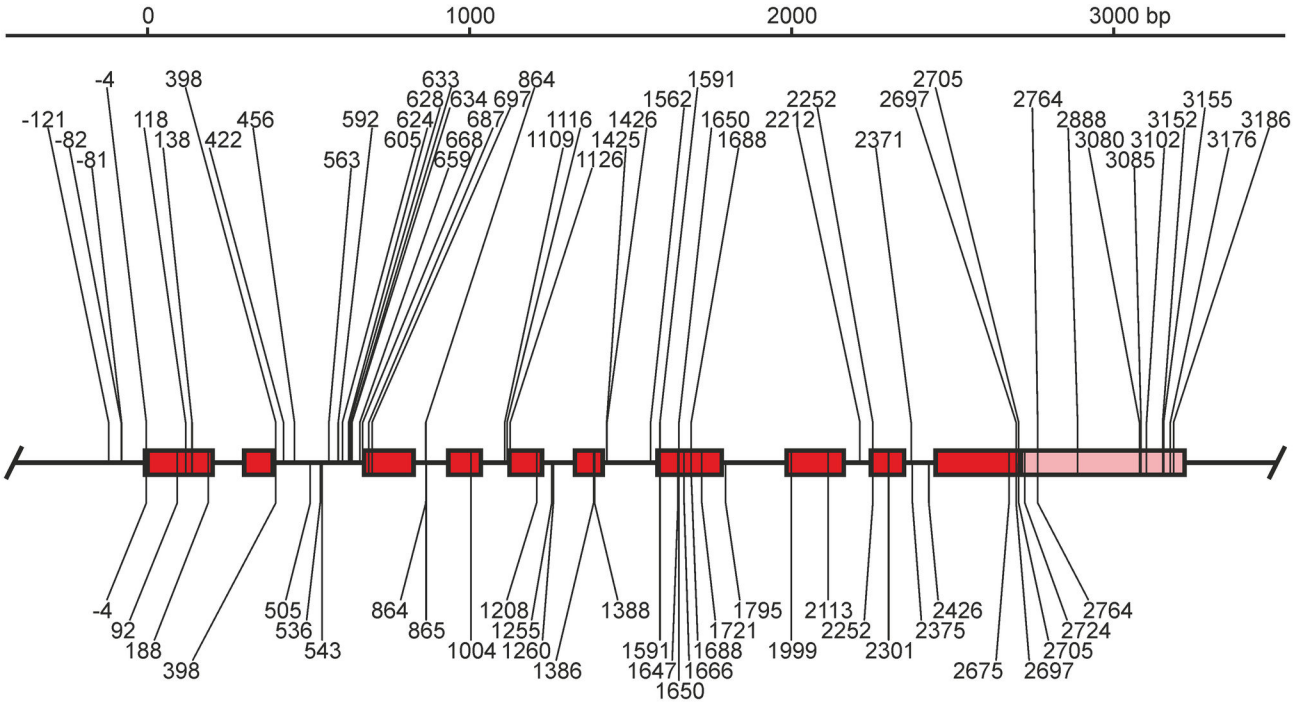

Figure 4. Scale representation of the segregating sites of human *CYP21* genes. Boxes symbolize the exons, red indicates the coding region, pink shows the untranslated regions. Segregating sites are denoted by their position, numbered from the start of the *CYP21A2* coding region in the sequence NT_007592.15: 31945792-31949720. The segregating site of the *CYP21A2* gene can be seen above the depicted gene. The segregating site of the *CYP21A1P* gene derived from an external dataset can be found below the depicted gene.

doi: 10.1371/journal.pone.0081977.g004

nH tests of the intron 2 subregion and the H test of whole gene (Table S8). In addition, the synonymous and nonsynonymous polymorphic sites were also assessed in the cds subregion; π_A/π_S was 0.288, Ka/Ks was 0.389, and the McDonald-Kreitman test did not show any deviation.

Discussion

There are a relatively few theoretical simulation studies on duplicated genes [16,74,75,78], and their study and analysis methods are still in their infancy. In addition, some analysis methods for non-duplicated genes cannot be used for duplicated genes (for instance, Hudson, Kreitman, and Aguade's test [75]). Despite the youth of the research field, some studies provide good examples for the utilization of methods for duplicated genes: Phylogenetic trees were commonly utilized for the detection of concerted evolution of paralogues [26,31-33]. Spatial heterogeneity in the phylogenetic signal were used for the identification of regions under different forces in a previous article [28], and it should be noted that the algorithms for phylogenetic signal analysis are suited to the detection of recombination [79]. The analyses of fixed, specific and shared site classes can be achieved in the datasets of more studies [26,27,29,30,75], and these results can therefore be compared with the result of the current study.

A total of 33 *CYP21A2* haplotype variants encoding 6 protein variants were determined from a European population by a combined molecular and inferred haplotyping approach. As the organization of RCCX CNV is highly complex, the complicated methods ensured the collection of reliable genetic data as normal as the data of non-duplicated genes if we set the advantage of molecular haplotyping aside. The *CYP21A2* gene was proven to be highly diverse. The haplotype diversity of *CYP21A2* (HHe=0.949) was higher than those observed in the comprehensive gene resequencing studies in humans (0.012-0.929 in 313 genes [80] and 0.360-0.910 in genes with <10 kb length from 100 genes [81]), and the nucleotide diversity of *CYP21A2* (π=2.55x10^{-3}) was also much higher than average in humans (0.58x10^{-3} in 313 genes [80] and 0.67 x10^{-3} in 212 genes [82]). The occurrence of fixed, shared and specific sites resembled that in a previous study on duplicated human rhesus genes [29], and this study as well as a theoretical study [78] has predicted mild non-allelic gene conversion activity to be assigned to this kinds of occurrence. The low level of LD between the segregating sites (File S1) also reflected this recombination activity.

Neutrality tests based on different genetic features presented the significant signature of positive selection of the whole *CYP21A2* gene. The population bottleneck and exponential growth characterizing the demographic history of Europeans [68,69] were built into the rejection probabilities of neutrality

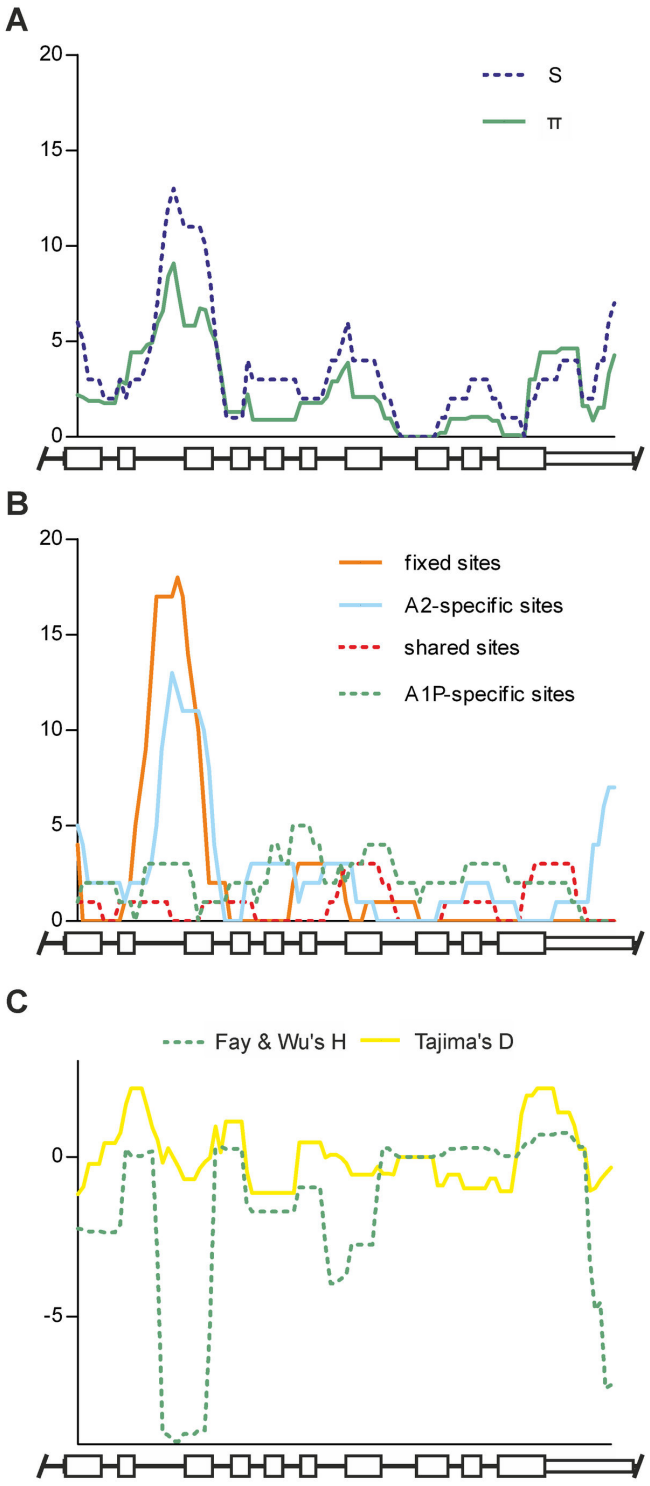

Figure 5. Sliding window plots of different genetic features. Schematic *CYP21* genes are indicated below the plots, high white boxes symbolize the exons, low white boxes represent the untranslated regions, and black lines indicate the introns and flanking regions. (A) Spatial distributions of the number of polymorphic sites (S) and the average number of pairwise nucleotide differences (π) calculated from the population dataset of human *CYP21A2* haplotypes. (B) Spatial distributions of polymorphic site classes (fixed, shared, *CYP21A2*-specific, *CYP21A1P*-specific) based on their coexistence in the human paralogues and calculated from the datasets of human *CYP21A2* haplotypes and human *CYP21A1P* polymorphisms. (C) Spatial distribution of Tajima's D and Fay and Wu's H values calculated from the human population dataset of *CYP21A2* haplotypes.

doi: 10.1371/journal.pone.0081977.g005

Table 1. Genetic features of the full-length *CYP21A2* gene and its subregions.

	length	GC content	S	HHe	π ($x10^{-3}$)	fixed sites	shared sites	CYP21A2-specific sites	D	Fs	H	nH	EW
full-length gene	3357	0.62	44	0.949	2.55	27	10	34	-0.271 (0.200)	**-9.20 (0.024)**	-22.18 (0.063)	**-3.83 (0.039)**	**0.103 (0.008)**
intron 2 subregion	282	0.54	12	0.885	8.97	**17**	**1**	**11**[a,b,c]	-0.019 (0.197)	**-6.43 (<0.001)**	**-8.49 (0.026)**	**-4.73 (0.013)**	0.293 (>0.10)
non-cds with-out intron 2 subregion	1587	0.64	21	0.763	2.06	**5**	**3**	**18**[d]	-0.832 (0.127)	-4.40 (0.028)	-8.87 (0.091)	-2.94 (0.072)	**0.249 (0.012)**
cds subregion	1488	0.62	11	0.860	1.87	**5**	**6**	**5**	0.542 (0.407)	**-6.68 (0.009)**	-4.82 (0.086)	-2.72 (0.074)	0.153 (>0.10)

S – the number of segregating sites, HHe – haplotype diversity, π – nucleotide diversity, D – Tajima's D test, Fs – Fu's Fs test, H – Fay and Wu's H test, nH – normalized Fay and Wu's H test and EW – Ewens-Watterson test. Bold characters indicate significant values, and the rejection probability values are shown in parentheses. The rejection probabilities of D, Fs, H and nH were corrected by the demography model of the European population, and rejection probabilities of the Ewens-Watterson test were under a neutral model. [a]Pearson's chi-square (χ^2) test of fixed, shared and *CYP21A2*-specific sites between intron 2 and the remaining part of the gene, p=0.0055, power (χ^2)=1.0000. [b]χ^2 test between intron 2 and the non-cds subregion without intron 2, p=0.0106, power=1.0000. [c]χ^2 test between intron 2 and the cds subregion, p=0.0091, power=1.0000. [d]χ^2 test between the non-cds subregion without intron 2 and the cds subregion, p=0.0423, power=0.9989.

doi: 10.1371/journal.pone.0081977.t001

tests because neutrality tests are sensitive to these population changes (Tajima's D and Fu's Fs test to population growth [62], Ewens-Watterson test to bottleneck [65] and Fay and Wu's H to recent bottleneck [83]) aside from positive selection. Although Tajima's D test did not show significance, Fu's Fs test, which detects the frequency spectrum of sites with rare alleles (as well as Tajima's D test), was significant agreeing with the result that Fs test is more powerful for positive selection than Tajima's D [62]. The Ewens-Watterson haplotype test, which is conditional on the number of haplotypes, also rejected the hypothesis of neutrality, although the EW test was not tested against the population changes. The normalized Fay and Wu's test under the demography model presented the significant rejection of neutrality in whole gene, but the H test failed to do this. Both H values were around the threshold of rejection probability, and the deviation in the extent of significances between the Fs and nH tests might be that because the sensitivity of the H test to positive selection (hitchhiking) does not persist long after the fixation of an advantageous allele [84]. Besides the population changes, gene conversion may also confound the results of the neutrality tests, and two independent non-allelic gene conversion events between *CYP21* sequences were evident by detection algorithms. We attempted to build these into the rejection probabilities of neutrality test, but models with gene conversion are not well developed [36], and some available programs did not fit with the observed features of the *CYP21A2* gene (for example, only interallelic (allelic, intralocus) gene conversion (conversion between orthologous sequences) is incorporated in the ms coalescent simulation tool [67], but it does not handle the non-allelic gene conversion.) Therefore, our (rough-and-ready) approach was similar to another CNV study on positive selection [36], and the sites affected by the two non-allelic gene conversion events exempted. The omitted sites were only influenced the significance of the H test in the whole gene, where its value became significant. The rejection of neutrality by the H but not by the D test is the unique signature of recent positive selection [63]. The significant result of the three neutrality tests is confirmed by the fact that a recent positive selection on human *CYP21A2* has also been observed based on fixation index in a genome-wide CNV study [85].

The effect of recent positive selection (adaptive protein evolutions) on the cds subregion was not detected by the McDonald-Kreitman test, but the values for synonymous and non-synonymous nucleotide diversity and divergence (π_A/π_S=0.288 and Ka/Ks=0.389) fell below the average values of genomes (π_A/π_S=0.34 [86], Ka/Ks (between human and gorilla)=0.42 [87]). These values reflect the predominant negative selection throughout the human genome [88-90]), and indicated a weak purifying (negative) selection on *CYP21A2*, and this finding was in accord with the significant negative selection in the protein coding sequence of *CYP21A2* described by a previous genome-wide study [88]. The presence of purifying selection is not surprising in the case of *CYP21A2* and CAH, hence the purifying selection called in the formal literature is often genetic disease when the mutation affects humans [91]. The recently published three-dimensional crystal structure of CYP21A2 protein [92] has indicated that

amino acid residues maintaining the enzyme structure are distributed throughout the entire structure, and that negative selection may affect the majority of the coding region. Furthermore, the cds subregion and the non-cds without intron 2 subregion also differed in the occurrences of fixed, shared and *CYP21A2*-specific sites, implying that the occurrences of site classes was actuated by the weak negative selection.

In addition to the negative selection on the cds subregion of *CYP21A2*, a line of evidence from independent datasets verified the separation of genetic features of intron 2 from those of the remaining part of the gene. First, the clades in the phylogenetic tree of intron 2, which could be demonstrated only with primate sequences available from public databases, followed divergent evolution, whereas the remaining part of the gene showed the signs of concerted evolution. The latter applied separately to the cds and the non-cds without intron 2 subregions, and moreover, to the intergenic region between *C4* and *CYP21* genes, and the values of the Robinson–Foulds tree distance metric supported the observed similarities and dissimilarities between the tree topologies. It should be noted that there is little functional or population data from the RCCX CNV of great ape monkeys. For example, the functional and disabled states of different gorilla and orangutan paralogues has not been functionally confirmed, but the sequences applied in this study agreed well with the independent sequences from a previous study [20]. Second, the occurrence of fixed, shared and *CYP21A2*-specific sites, which was classified based on only the genotypic data of polymorphic sites of human paralogues, and was free from the bioinformatic haplotype reconstruction, showed significant difference in intron 2 compared to those in the remaining part of the gene and the other two subregions. A clear excess of the fixed sites, which feature the reduction of effective non-allelic conversion rate and homogenization due to selection [74], were observed in intron 2. The presence of the fixed sites and the potential reduction of effective non-allelic conversion in intron 2 were in concordance with the lacks of homogenization and concerted evolution, however, one of the statistically evident gene conversion event occurred in intron 2 subregion. Third, the Fs, the H and normalized H tests, which were calculated from the human population dataset of *CYP21A2* haplotypes, deviated from neutrality in intron 2 under the European demography model. At first glance, the EW test value of intron 2 contradicted the presence of positive evolution, but observed homozygosity was slightly higher in intron 2 despite the fact that the intron 2 haplotype adequately characterized the full-length haplotype, and the expected homozygosity is conditional on the number of sites, which can greatly change the critical value of shorter sequences with fewer sites. Along the same lines, omitting one-third of segregating sites because of the observed gene conversion in the intron 2 subregion could also affect the rejection probabilities of the H tests, and may not necessarily influence the rejection probabilities in a direct way. We conclude that positive selection presumably focused on intron 2, however, the selective sweep of neutral sites partly slurred and spread the signature of positive selection across the whole *CYP21A2* gene. Therefore, the positive selection, which shapes the diversity and divergence of intronic DNA in eukaryotes [93,94], was potentially associated with the excess of fixed and *CYP21A2*-specific sites in intron 2. In accord with this, the accumulation of fixed sites has been connected to the positive selection site of coding sequences in tandem duplicated genes, although both duplicated genes are functional in these cases [28,29].

Besides the detected positive selection in *CYP21A2*, RCCX structures harboring only the *C4A* gene (one of the two types of *C4*) are associated with different hormone levels [95], implying that the *CYP21A2* haplotype variants can function differently, and some variants may be advantageous. The search for the causative site or sites under positive selection should be conducted mostly in intron 2, which raises the question as to what the functional role of the potential site or sites may be. Gene regulation could be a plausible answer to this question, since an alternative transcript retaining intron 1 and 2 is expressed from *CYP21A2* with a relative abundance of 10-20% compared to the correct transcript [96]. Other alternative transcripts have also been recognized [97], and their abundance and frequency suggest that these alternative transcripts are not aimlessly generated and may contribute to alternative splicing. Furthermore, the *CYP21A2*-specific site 659 (rs6467) was located at the same site as one of the most frequent CAH mutations (an third allele of this site) affecting RNA splicing [22]. Therefore, positive selection may drive the recent adaptive change of cortisol and aldosterone responses through the gene regulation of *CYP21A2*. This matches well with the fact that genes having environmentally responsive functions are amassed in CNVs with duplicated genes, and these genes have long been considered to be subject to rapid adaptive evolution [98].

Supporting Information

Table S1. Primers used in the study. Allele-specific sites are indicated on the sequences by underscore. Non-allele-specific primer sequences avoid the polymorphic sites based on the ENSEMBL database and the sequences of six HLA-homozygous cell lines.
(DOC)

Table S2. Sequences from GenBank used in this study.
(DOC)

Table S3. Average nucleotide identities of higher primate (*Catarrhini*) *CYP21* orthologues and paralogues. Minimum and maximum identity values are shown in parentheses. hA2 indicates human *CYP21A2* sequences, hA1P indicates *CYP21A1P* sequences, c, g, o and m before A2 or A1P indicates chimpanzee, gorilla, orangutan and macaque sequences, respectively.
(DOC)

Table S4. *CYP21A2* haplotype variants and their segregating sites. RCCX structures assigned to a particular haplotype are indicated in column 2. Column 3-56 indicate the segregating sites, bold numbers indicate the segregating sites of the current study, normal numbers indicate the sites being

absent in the current study, but described in a recent study [13]. The concordance with ancestral MHC haplotype from external database (GenBank) is indicated in column RCCX. Four variants of RCCX structure-*CYP21A2* haplotype variants agreed with the sequences of COX, DBB, MCF and QBL cell lines. The *CYP21A2* sequence in the PGF cell line deviated from the h58 haplotype by one nucleotide (PGF -1).
(XLS)

Table S5. Protein variants were encoded by *CYP21A2* haplotype variants in the current study. Protein variants were derived from six segregating sites causing amino acid changes; site 28-30 (rs61338903, amino acid (aa) 12) --- – -, CTG – leucine (L), site 687 – (rs6474, aa 102) A – lysine (K), G – arginine (R), site 1650 (rs6472, aa 286) C – threonine (T), G – serine (S), site 1688 (rs6471, aa 281) G – valine (V), T – leucine (L) and site 2705 (rs6473, aa 493) A – asparagine (N), G – serine (S). The site 28-30 was not included in the genetic analyses of the current study.
(DOC)

Table S6. Frequencies of haplotypic RCCX structure variants in the current study and in a recent family-based study [17]. A module (a repeat) is abbreviated with two letters, the first represents the alleles of HERV-K CNV (L – the long allele or S – short allele), and the second symbolizes the types of *C4* gene (A or B). The multiplication of these two letters indicates bi- and trimodular structures. The number in parentheses indicates the number of *CYP21A2* on the particular haplotypic RCCX structure, but one *CYP21A2* gene is not shown.
(DOC)

Table S7. Segregating sites of human *CYP21* genes and differences between the human paralogues. The sites of the datasets of human *CYP21A2* haplotypes and human *CYP21A1P* polymorphisms were classified based on their coexistence into four types: *CYP21A2*-specific and *CYP21A1P*-specific polymorphic sites, at which polymorphisms were observed in either of the two genes; shared sites, at which polymorphisms were shared by the two paralogues; and fixed sites, at which each paralogue had a different fixed allele.
(DOC)

Table S8. Rejection probabilities of neutrality tests. Rejection probabilities of Tajima's D test, Fu's Fs test, Fay and Wu's H test, normalized Fay and Wu's H (nH) test and Ewens-Watterson (EW) test were under a neutral model, a demography model of the European population or a European demography model on the dataset without sites affected by gene conversion. Rejection probabilities of Ewens-Watterson test were not calculated under the European demography model, and the values of neutrality tests are not shown for dataset without sites affected by gene conversion. [a]The cds subregion was not affected by statistically evident gene conversion events.
(DOC)

Figure S1. Rooted neighbor-joining phylogenetic tree constructed from the full-length sequences of the great ape *CYP21* sequence dataset using complete gap deletion. The names of *CYP21* sequences of HLA-homozygous cell lines available from public databases are represented at the human sequences. Bootstrap values are shown next to the corresponding nodes, but the values within human clades are not presented for clarity. The scale bar indicates genetic distance.
(TIF)

Figure S2. Rooted maximum likelihood phylogenetic tree constructed from the full-length sequences of the great ape *CYP21* sequence dataset using gap sites. The names of *CYP21* sequences of HLA-homozygous cell lines available from public databases are represented at the human sequences. Bootstrap values are shown next to the corresponding nodes, but the values within human clades are not presented for clarity. The scale bar indicates genetic distance.
(TIF)

Figure S3. Spatial distributions of phylogenetic signals derived from the gorilla and orangutan paralogous pairs. Gorilla pair is indicated by gA2-gA1P, and orangutan pair is indicated byoA2-oA1P. The likelihood of closely related sequences to resemble each other more than random sequences of the same phylogenetic tree is expressed by '% of permutated trees' in y axis. Schematic full-length *CYP21* genes are indicated below the plots, high white boxes symbolize the exons, low white boxes represent the untranslated regions, and black lines indicate the introns and flanking regions.
(TIF)

Figure S4. Rooted maximum likelihood (ML) phylogenetic trees constructed from selected sequences of the cds subregion, the non-cds without intron 2 subregion and the intergenic region between *C4* and *CYP21* genes. The names of *CYP21* sequences of HLA-homozygous cell lines available from public databases are represented at the human sequences. Bootstrap values are shown next to the corresponding nodes, but the values within human clades are not presented for clarity. The scale bar indicates genetic distance. (A) Rooted ML tree of the great ape *CYP21* cds subregion. (B) Rooted ML tree of the great ape *CYP21* non-cds without intron 2 subregion. (C) Rooted ML tree of great ape intergenic region between *C4* and *CYP21* genes.
(TIF)

File S1. Assessment of the status of great ape *CYP21* sequences and the validity of experimental data.
(DOC)

Acknowledgements

One of our authors, Prof. George Fust, departed this life in 2012. Rest in peace. We are indebted to Mark Eyre for English proofreading, and we also would like to thank the staff of Biomi

Kft, Dr. Adrienn Micsinai, Dr. Anita Mohr, Dr. Rita Sipos and Réka Szántó-Egész for help with the running of the sequencing reactions. We thank Prof. Mária Sasvári-Székely and Dr. Zsolt Rónai (Department of Medical Chemistry, Molecular Biology and Pathobiochemistry, Semmelweis University, Budapest, Hungary) for their patience when we used their qPCR machine.

Author Contributions

Conceived and designed the experiments: MD. Performed the experiments: JAS ZD MD. Analyzed the data: JAS MD. Wrote the manuscript: AS AP HF ZP KR GF MD.

References

1. Ohno S (1970) Evolution by gene duplication. Berlin, Heidelberg, NY: Springer-Verlag.
2. Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9: 938-950. doi: 10.1038/nrg2482. PubMed: 19015656.
3. Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11: 97-108. doi:10.1038/ni0110-97d. PubMed: 20051986.
4. Proulx SR (2012) Multiple routes to subfunctionalization and gene duplicate specialization. Genetics 190: 737-751. doi:10.1534/genetics.111.135590. PubMed: 22143920.
5. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155. doi:10.1126/science.290.5494.1151. PubMed: 11073452.
6. Khurana E, Lam HY, Cheng C, Carriero N, Cayting P et al. (2010) Segmental duplications in the human genome reveal details of pseudogene formation. Nucleic Acids Res 38: 6997-7007. doi: 10.1093/nar/gkq587. PubMed: 20615899.
7. Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39: 121-152. doi:10.1146/annurev.genet.39.073003.112240. PubMed: 16285855.
8. Ota T, Nei M (1994) Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol Biol Evol 11: 469-482. PubMed: 8015440.
9. Ohta T (1980) Evolution and variation of multigene families; S Levin. Berlin, Heidelberg, NY Springer-Verlag.
10. Nagylaki T, Petes TD (1982) Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. Genetics 100: 315-337. PubMed: 7106560.
11. Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM et al. (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14: 1-18. doi:10.1097/00008571-200401000-00001. PubMed: 15128046.
12. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA et al. (2004) Gene map of the extended human MHC. Nat Rev Genet 5: 889-899. doi:10.1038/nrg1489. PubMed: 15573121.
13. Bánlaki Z, Szabó JA, Szilágyi A, Patócs A, Prohászka Z et al. (2013) Intraspecific evolution of human RCCX copy number variation traced by haplotypes of the *CYP21A2* gene. Genome Biol Evol 5: 98-112. doi: 10.1093/gbe/evs121. PubMed: 23241443.
14. Blanchong CA, Zhou B, Rupert KL, Chung EK, Jones KN et al. (2000) Deficiencies of human complement component C4A and C4B and heterozygosity in length variants of RP-C4-CYP21-TNX (RCCX) modules in caucasians. The load of RCCX genetic diversity on major histocompatibility complex-associated disease. J Exp Med 191: 2183-2196. doi:10.1084/jem.191.12.2183. PubMed: 10859342.
15. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature 464: 704-712. doi:10.1038/nature08516. PubMed: 19812545.
16. Teshima KM, Innan H (2012) The coalescent with selection on copy number variants. Genetics 190: 1077-1086. doi:10.1534/genetics.111.135343. PubMed: 22174068.
17. Bánlaki Z, Doleschall M, Rajczy K, Fust G, Szilágyi A (2012) Fine-tuned characterization of RCCX copy number variants and their relationship with extended MHC haplotypes. Genes Immun 13: 530-535. doi: 10.1038/gene.2012.29. PubMed: 22785613.
18. Higashi Y, Yoshioka H, Yamane M, Gotoh O, Fujii-Kuriyama Y (1986) Complete nucleotide sequence of two steroid 21-hydroxylase genes tandemly arranged in human chromosome: a pseudogene and a genuine gene. Proc Natl Acad Sci U S A 83: 2841-2845. doi:10.1073/pnas.83.9.2841. PubMed: 3486422.
19. White PC, New MI, Dupont B (1986) Structure of human steroid 21-hydroxylase genes. Proc Natl Acad Sci U S A 83: 5111-5115. doi: 10.1073/pnas.83.14.5111. PubMed: 3487786.
20. Kawaguchi H, O'HUigin C, Klein J (1992) Evolutionary origin of mutations in the primate cytochrome P450c21 gene. Am J Hum Genet 50: 766-780. PubMed: 1550121.
21. New MI, Abraham M, Gonzalez B, Dumic M, Razzaghy-Azar M et al. (2013) Genotype-phenotype correlation in 1,507 families with congenital adrenal hyperplasia owing to 21-hydroxylase deficiency. Proc Natl Acad Sci U S A 110: 2611-2616. doi:10.1073/pnas.1303471110. PubMed: 23359698.
22. Speiser PW, White PC (2003) Congenital adrenal hyperplasia. N Engl J Med 349: 776-788. doi:10.1056/NEJMra021561. PubMed: 12930931.
23. Jeffreys AJ, May CA (2004) Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat Genet 36: 151-156. doi:10.1038/ng1287. PubMed: 14704667.
24. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M (2002) High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet 71: 759-776. doi: 10.1086/342973. PubMed: 12297984.
25. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321-324. doi:10.1126/science.1117196. PubMed: 16224025.
26. Araki H, Inomata N, Yamazaki T (2001) Molecular evolution of duplicated amylase gene regions in Drosophila melanogaster: evidence of positive selection in the coding regions and selective constraints in the cis-regulatory regions. Genetics 157: 667-677. PubMed: 11156987.
27. Bettencourt BR, Feder ME (2002) Rapid concerted evolution via gene conversion at the Drosophila hsp70 genes. J Mol Evol 54: 569-586. doi: 10.1007/s00239-001-0044-7. PubMed: 11965431.
28. Goldstone HM, Stegeman JJ (2006) A revised evolutionary history of the CYP1A subfamily: gene duplication, gene conversion, and positive selection. J Mol Evol 62: 708-717. doi:10.1007/s00239-005-0134-z. PubMed: 16752211.
29. Innan H (2003) A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. Proc Natl Acad Sci U S A 100: 8793-8798. doi:10.1073/pnas.1031592100. PubMed: 12857961.
30. Lazzaro BP, Clark AG (2001) Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the Attacin genes of Drosophila melanogaster. Genetics 159: 659-671. PubMed: 11606542.
31. Nikolaidis N, Nei M (2004) Concerted and nonconcerted evolution of the Hsp70 gene superfamily in two sibling species of nematodes. Mol Biol Evol 21: 498-505. PubMed: 14694072.
32. Storz JF, Baze M, Waite JL, Hoffmann FG, Opazo JC et al. (2007) Complex signatures of selection and gene conversion in the duplicated globin genes of house mice. Genetics 177: 481-500. doi:10.1534/genetics.107.078550. PubMed: 17660536.
33. Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM (2004) Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res 14: 354-366. doi:10.1101/gr.2133704. PubMed: 14993203.
34. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256-1260. doi:10.1038/ng2123. PubMed: 17828263.
35. Verrelli BC, Tishkoff SA (2004) Signatures of selection and gene conversion associated with human color vision variation. Am J Hum Genet 75: 363-375. doi:10.1086/423287. PubMed: 15252758.
36. Xue Y, Sun D, Daly A, Yang F, Zhou X et al. (2008) Adaptive evolution of UGT2B17 copy-number variation. Am J Hum Genet 83: 337-346. doi:10.1016/j.ajhg.2008.08.004. PubMed: 18760392.
37. Szilagyi A, Blasko B, Szilassy D, Fust G, Sasvari-Szekely M et al. (2006) Real-time PCR quantification of human complement C4A and C4B genes. BMC Genet 7: 1. doi:10.1186/1471-2350-7-1. PubMed: 16403222.
38. Wu YL, Savelli SL, Yang Y, Zhou B, Rovin BH et al. (2007) Sensitive and specific real-time polymerase chain reaction assays to accurately determine copy number variations (CNVs) of human complement C4A, C4B, C4-long, C4-short, and RCCX modules: elucidation of C4 CNVs

in 50 consanguineous subjects with defined HLA genotypes. J Immunol 179: 3012-3025. PubMed: 17709516.
39. Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73: 1162-1169. doi:10.1086/379378. PubMed: 14574645.
40. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978-989. doi:10.1086/319501. PubMed: 11254454.
41. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ et al. (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics 60: 1-18. doi:10.1007/s00251-007-0262-2. PubMed: 18193213.
42. Cantürk C, Baade U, Salazar R, Storm N, Pörtner R et al. (2011) Sequence Analysis of CYP21A1P in a German Population to Aid in the Molecular Biological Diagnosis of Congenital Adrenal Hyperplasia. Clin Chem 57: 511-517. doi:10.1373/clinchem.2010.156893. PubMed: 21148302.
43. Ralph P, Coop G (2013) The geography of recent genetic ancestry across Europe. PLoS Biol 11: e1001555. PubMed: 23667324.
44. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR et al. (2008) Genes mirror geography within Europe. Nature 456: 98-101. doi:10.1038/nature07331. PubMed: 18758442.
45. Tsai LP, Lee HH (2012) Analysis of CYP21A1P and the duplicated CYP21A2 genes. Gene 506: 261-262. doi:10.1016/j.gene.2012.06.045. PubMed: 22771554.
46. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948. doi:10.1093/bioinformatics/btm404. PubMed: 17846036.
47. Tamura K, Peterson D, Peterson N, Stecher G, Nei M et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739. doi:10.1093/molbev/msr121. PubMed: 21546353.
48. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10: 564-567. doi:10.1111/j.1755-0998.2010.02847.x. PubMed: 21565059.
49. Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39: 175-191
50. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS et al. (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 73: 152-160. PubMed: 9847317.
51. Robinson DR, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53: 131-147. doi:10.1016/0025-5564(81)90043-2.
52. Puigbò P, Garcia-Vallvé S, McInerney JO (2007) TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics 23: 1556-1558. doi:10.1093/bioinformatics/btm135. PubMed: 17459965.
53. Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol 16: 37-45. doi:10.1016/S0169-5347(00)02026-7. PubMed: 11146143.
54. Boni MF, Posada D, Feldman MW (2007) An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176: 1035-1047. PubMed: 17409078.
55. Martin DP, Posada D, Crandall KA, Williamson C (2005) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses 21: 98-102. doi:10.1089/aid.2005.21.98.
56. Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6: 526-538.
57. Smith JM (1992) Analyzing the mosaic structure of genes. J Mol Evol 34: 126-129.
58. Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics 16: 562-563. doi:10.1093/bioinformatics/16.6.562. PubMed: 10980155.
59. Martin DP, Lemey P, Lott M, Moulton V, Posada D et al. (2010) RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26: 2462-2463. doi:10.1093/bioinformatics/btq467. PubMed: 20798170.
60. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452. doi:10.1093/bioinformatics/btp187. PubMed: 19346325.
61. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595. PubMed: 2513255.
62. Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915-925. PubMed: 9335623.
63. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413. PubMed: 10880498.
64. Zeng K, Fu YX, Shi S, Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431-1439. doi:10.1534/genetics.106.061432. PubMed: 16951063.
65. Zeng K, Mano S, Shi S, Wu CI (2007) Comparisons of site- and haplotype-frequency methods for detecting positive selection. Mol Biol Evol 24: 1562-1574. doi:10.1093/molbev/msm078.
66. Ramos-Onsins SE, Mitchell-Olds T (2007) Mlcoalsim: multilocus coalescent simulations. Evol Bioinform Online 3: 41-44.
67. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338. doi:10.1093/bioinformatics/18.suppl_1.S337.
68. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT et al. (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108: 11983-11988. doi:10.1073/pnas.1019276108. PubMed: 21730125.
69. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLOS Genet 5: e1000695.
70. Henn BM, Cavalli-Sforza LL, Feldman MW (2012) The great human expansion. Proc Natl Acad Sci U S A 109: 17758-17764. doi:10.1073/pnas.1212380109. PubMed: 23077256.
71. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654. doi:10.1038/351652a0.
72. Wilming LG, Hart EA, Coggill PC, Horton R, Gilbert JG et al. (2013) Sequencing and comparative analysis of the gorilla MHC genomic sequence. Database (Oxford). p. bat011.
73. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I et al. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169-175. PubMed: 22398555.
74. Innan H (2004) Theories for analyzing polymorphism data in duplicated genes. Genes Genet Syst 79: 65-75. PubMed: 15215672.
75. Innan H (2003) The coalescent and infinite-site model of a small multigene family. Genetics 163: 803-810. PubMed: 12618415.
76. Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39: 197-218. PubMed: 16285858.
77. Slatkin M (1996) A correction to the exact test based on the Ewens sampling distribution. Genet Res 68: 259-260. PubMed: 9062082.
78. Thornton KR (2007) The neutral coalescent process for recent gene duplications and copy-number variants. Genetics 177: 987-1000. PubMed: 17720930.
79. Martin DP, Lemey P, Posada D (2011) Analysing recombination in nucleotide sequences. Mol Ecol Resour 11: 943-955. PubMed: 21592314.
80. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T et al. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293: 489-493. PubMed: 11452081.
81. Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q et al. (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74: 610-622. PubMed: 15015130.
82. Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS et al. (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14: 1821-1831. PubMed: 15364900.
83. Li H (2011) A new test for detecting recent positive selection that is free from the confounding impacts of demography. Mol Biol Evol 28: 365-375.
84. Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160: 1179-1189. PubMed: 11901132.
85. Kato M, Kawaguchi T, Ishikawa S, Umeda T, Nakamichi R et al. (2010) Population-genetic nature of copy number variations in the human genome. Hum Mol Genet 19: 761-773. PubMed: 19966329.
86. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227-1234.
87. Wildman DE, Uddin M, Liu G, Grossman LI, Goodman M (2003) Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo. Proc Natl Acad Sci U S A 100: 7181-7188.
88. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.
89. Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T et al. (2011) Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLOS Genet 7: e1002326.

90. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65. PubMed: 23128226.
91. Hurst LD (2009) Fundamental concepts in genetics: genetics and the understanding of selection. Nat Rev Genet 10: 83-93. PubMed: 19119264.
92. Zhao B, Lei L, Kagawa N, Sundaramoorthy M, Banerjee S et al. (2012) Three-dimensional structure of steroid 21-hydroxylase (cytochrome P450 21A2) with two substrates reveals locations of disease-associated variants. J Biol Chem 287: 10613-10622.
93. Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149-1152. PubMed: 16237443.
94. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X et al. (2009) Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet 5: e1000592. PubMed: 19662163.
95. Bánlaki Z, Raizer G, Acs B, Majnik J, Doleschall M et al. (2012) ACTH-induced cortisol release is related to the copy number of the C4B gene encoding the fourth component of complement in patients with non-functional adrenal incidentaloma. Clin Endocrinol (Oxf) 76: 478-484. PubMed: 21967755.
96. Bristow J, Gitelman SE, Tee MK, Staels B, Miller WL (1993) Abundant adrenal-specific transcription of the human P450c21A "pseudogene". J Biol Chem 268: 12919-12924. PubMed: 7685353.
97. Jin P, Fu GK, Wilson AD, Yang J, Chien D et al. (2004) PCR isolation and cloning of novel splice variant mRNAs from known drug target genes. Genomics 83: 566-571. PubMed: 15028279.
98. Cooper GM, Nickerson DA, Eichler EE (2007) Mutational and selective effects on copy-number variants in the human genome. Nat Genet 39: S22-S29. PubMed: 17597777.