Evolutionarily conserved non-protein-coding regions in the chicken genome harbor functionally important variation

Christian Groß[1,2,*], Chiara Bortoluzzi[3*], Dick de Ridder[1], Hendrik-Jan Megens[3], Martien AM Groenen[3], Marcel Reinders[2], Mirte Bosse[3]

[1] Bioinformatics Group, Wageningen University & Research, 6708 PB, Wageningen, The Netherlands

[2] Delft Bioinformatics Lab, University of Technology Delft, 2600GA, Delft, The Netherlands

[3] Animal Breeding and Genomics Group, Wageningen University & Research, 6708 PB, Wageningen, The Netherlands

* These authors equally contributed to this work

Abstract

The availability of genomes for many species has advanced our understanding of the non-protein-coding fraction of the genome. Comparative genomics has proven to be an invaluable approach for the systematic, genome-wide identification of conserved non-protein-coding elements (CNEs). However, for many non-mammalian model species, including chicken, our capability to interpret the functional importance of variants overlapping CNEs has been limited by current genomic annotations, which rely on a single information type (e.g. conservation). We here studied CNEs in chicken using a combination of population genomics and comparative genomics. To investigate the functional importance of variants found in CNEs we develop a ch(icken) Combined Annotation-Dependent Depletion (chCADD), a variant effect prediction tool first introduced for humans and later on for mouse and pig. We show that 73 Mb of the chicken genome has been conserved across more than 280 million years of vertebrate evolution. The vast majority of the conserved elements are in non-protein-coding regions, which display SNP densities and allele frequency distributions characteristic of genomic regions constrained by purifying selection. By annotating SNPs with the chCADD score we are able to pinpoint specific subregions of the CNEs to be of higher functional importance, as supported by SNPs found in these subregions are associated with known disease genes in humans, mice, and rats. Taken together, our findings indicate that CNEs harbor variants of functional significance that should be object of further investigation along with protein-coding mutations. We therefore anticipate chCADD to be of great use to the scientific community and breeding companies in future functional studies in chicken.

Introduction

The rapidly increasing availability of genomes has considerably advanced our understanding of the non-protein-coding fraction of the genome. With the sequencing of the human genome (1) and the first ENCODE project (2,3) it was soon realized that protein-coding genes constitute a small fraction of a species functional genome and that the remaining non-protein-coding DNA is not simply ´junk´ DNA as initially thought. Nevertheless, the functional importance of these non-protein-coding regions remained for long time unknown, as determining (molecular) function was far more difficult than for protein-coding genes (4). A better understanding of the functional importance of these non-protein-coding regions comes from comparative genomics, which has allowed the systematic, genome-wide identification of conserved non-protein-coding elements (CNEs) (5,6).

Comparative genomics relies on the genome comparison of a group of species related by a narrow or wide time-scale (i.e. phylogenetic scope). Regions in the genome that share some minimum sequence similarity across two or more species are an indication of a selection constraint. Moreover, conservation often implies a biological function (7). Based on this principle, CNEs can be identified in any species included in the alignment, as reported in recent studies in the collared flycatcher (8), fruit flies (9), and plants (6). However, the phylogenetic scope (10) and species included in the alignment (11) can have important implications for the identification of CNEs. For instance, by including the spotted gar genome in their alignment, (11) recently identified numerous CNEs previously undetectable in direct human-teleost comparisons, supporting the importance of a bridging species in the alignment.

CNEs have been the subject of intense recent interest. The identification of CNEs has had important implications in enhancing genome annotation (12), investigating signatures of adaptive evolution (13–15), and identifying putative trait loci (16). CNEs and sequence conservation have also proven crucial in studying the genetic basis of phenotypic diversity. In fact, non-protein-coding SNPs have been linked to traits and diseases in genome-wide association studies (17,18).

Although the methodological advantages of a comparative genomic approach are well recognized, the functional interpretation of CNEs is incomplete if based on conservation alone, as conservation provides information on restrictions, but not on functionality. A possible solution is combining conservation with other complementary types of data that characterize the biological role of genetic sequences at a genome-wide scale (7). Such data include, for instance, RNA sequencing (RNA-seq) for the identification of transcriptionally active regions(19) and chromatin immunoprecipitation followed by sequencing (ChIP-seq) for regulatory-factor-binding regions (RFBRs) (20). In human genetics, integrative annotations such as Combined Annotation-Dependent Depletion (CADD) (21) have been developed. The main advantage of such frameworks is the combination, into a unique score, of diverse genomic features derived from,

among others, gene model annotations, evolutionary constraints, epigenetic measurements, and functional predictions (21,22).

Compared to humans, for many non-mammalian model species, including chicken (*Gallus gallus*), the situation is quite different. First, comparative genomic studies that made use of the very first genome assemblies (23–25) may have provided an incomplete and biased picture of avian CNEs and avian genome evolution, as recently pointed out by (26). Second, the lack of species-specific methods that can identify and score functional non-protein-coding mutations throughout the genome has restricted most of the research interest to protein-coding genes. In fact, in the context of protein-coding genes generic predictors such as SIFT (27), PolyPhen2 (28), and Provean (29) can be used.

We here addressed these limitations using a combination of comparative genomic and population genomic approaches to accurately predict CNEs in the chicken genome. Furthermore, we used machine learning to develop a ch(icken) Combined Annotation-Dependent Depletion (chCADD), in the tradition of previous CADD models for non-human species, including mouse (mCADD) (30) and pig (pCADD) (31). As we show, chCADD has the potential of providing new insights into the functional role of non-protein-coding regions of the chicken genome at a single base pair resolution.

Even though deciphering the function of the non-protein-coding portion of a species genome has been a challenging task, we expect our study to provide a new framework for decoding the still largely unknown function of CNEs and their relative variants in chicken, an ideal non-mammalian model and anchor species in evolutionary studies

Results

Conserved non-protein-coding elements cover a large fraction of the chicken genome

To define CNEs, we first identified conserved elements (CEs) using the UCSC PhastCons most conserved track approach (32). PhastCons predicted in the 23 sauropsids multiple sequence alignment (MSA) 1.14 million CEs encompassing ~8% of the chicken genome for a total of 73 Mb. In line with the density of genes and regulatory features characteristic of the chicken genome (33), we found that most of the predicted CEs are on micro-chromosomes (GGA11-GGA33), followed by intermediate (GGA6-GGA10) and macro-chromosomes (GGA1-GGA5) (**Figure S1**). Even though the length of predicted CEs ranged from 4 bp to a maximum of ~ 2,000 bp, the vast majority was short (< 100 bp) (**Figure S2**). Therefore, we do not expect any length bias in our final set of CEs.

We annotated CEs by genomic features, considering only genes for which the transcript had a proper annotated start and stop codon, as defined by the Ensembl´s annotation files (n = 14,828 genes). Overall, we found that 23% of the predicted CEs were associated with exonic sequences (i.e. CDS, 5’ UTR, 3’ UTR, promoter, and RNA genes) spanning 17.14 Mb of the chicken genome (**Table 1**). The majority of the exon-

associated CEs overlapped known coding regions (85% of total exon-associated CEs), followed by 3' UTRs (8% of total), and promoter regions (4% of total). Although we observed conservation in exon sequences, most CEs overlapped non-protein-coding sequences, including lncRNA (15% of total non-exon associated CEs), intronic (36% of total), and intergenic regions (49% of total). We further examined the biological processes and molecular functions of known genes overlapped by CEs in coding regions, 5' UTRs, 3' UTRs, and introns. These genes are associated with basic functions, including cell differentiation and development, anatomical structure development, morphogenesis, and growth (**Table 2**). Most of these GO categories have also been previously associated with mammalian and vertebrate ultraconserved elements (UCEs) (33,34).

In total we identified 259,688 CEs in protein-coding regions, leaving 850,920 CNEs spanning over 51 Mb of the chicken genome (**Table 1**), with a genome-wide distribution of 92.10 CNEs/100-kb. We further observed noticeable differences in the length distribution of CEs associated with different types of annotations. Among the conserved exon-associated CEs, those found in CDSs are, on average, the longest (~68 bp), followed by 3' UTRs (61 bp), RNA genes (52 bp), promoters (47 bp), and 5' UTRs (38 bp) (**Figure S3**). On the contrary, CEs found in non-protein-coding regions show a homogenous length distribution, ranging from 56 bp in introns to 63 bp in lncRNAs (**Figure S4**).

CNEs populate regions not occupied by genes

We further investigated the genomic location of CNEs as this might provide important clues to their functional role. We found that the distribution of CNEs in windows of 100 kb is significantly negatively correlated (r = -0.20; *p*-value: <2.2x10-16) with the distribution of exons (**Figure 1**). We subsequently analyzed chicken polymorphism data to address the mutational or evolutionary forces shaping CNEs, following previous studies in humans (35) and *Drosophila (9,36).* We used polymorphism densities to investigate whether these forces could still be acting on the chicken genome or they could have acted in other species and may no longer be relevant for chicken. SNP density, which reflects events within the chicken lineage, was calculated in the genomes of 169 chickens from different traditional breeds of divergent demographic and selection history. Specifically, we compared the SNP density found in CNEs with that in non-protein-coding elements that were identified not to be conserved (non-CNEs; i.e. not conserved intronic, lncRNA and intergenic regions), following (9,35,36). Overall, we found that CNEs are less enriched in SNPs (SNP density = 0.0092) than non-CNEs (SNP density = 0.02).

CNEs are selectively constrained in chicken

To test whether low local mutation rates in CNEs or purifying selection is responsible for the observed low SNP density, we looked at the derived allele frequency (DAF) distribution in CNEs and non-CNEs. This is

because mutation rate differences are not expected to affect the allele frequency spectra. On the contrary, selective constraint is responsible for the shift in allele frequency distribution of constrained alleles towards lower values. Allele frequencies for derived (new) alleles were compiled using the sequence of the inferred ancestor between chicken and turkey. The ancestral allele was determined for a total of ~9 million SNPs that passed several filtering criteria (see Methods). We observed an excess of rare (≤ 10%) derived alleles of SNPs within CNEs in all chicken populations (**Table 3**). Overall, 57% of SNPs within CNEs had a DAF ≤ 10%, compared to only ~48% in non-CNEs (the same pattern was observed for each SNP functional class; see also **Table 3**). Non-CNEs displayed on the contrary a higher proportion of common SNPs (DAF>10%) (~52% versus 43% within CNEs) independent of their functional class (**Table 3**). Therefore, the low proportion of derived alleles in CNEs indicates that evolutionary pressure has suppressed CNE-derived allele frequencies.

chCADD scores for the investigation of CNE and SNP evaluation

To investigate CNEs further, we developed a model that can evaluate individual SNPs or entire sequences based on a per-base score, with respect to its putative deleteriousness. This model is based on the CADD approach, hence it is labeled ch(icken) CADD. chCADD is a linear logistic model that is trained to differentiate between two classes of variants, one being relatively more enriched in potentially deleterious variants than the other. To obtain these two classes, one class is generated from derived variants, alleles that have accumulated since the last ancestor with turkey and became fixed or almost fixed (>90% AF) in our chicken populations. These are depleted in deleterious variants and can be assumed to be benign or at least neutral in their nature. The set of putative deleterious variants contains simulated *de novo* variants that are not depleted of deleterious variants. The feature weights obtained during training are shown in Supplementary file 2. Performance on a held out test set to determine an optimal penalization term are shown in **Figure S5**.

chCADD scores potentially causal variants higher

We evaluated the performance and applicability of chCADD on two different sets of variants before we annotated non-coding SNPs.

First, we assigned a chCADD score to all SNPs found in the genomes of the 169 chickens previously used in the SNP density and DAF analysis and compared these to functional predictions as annotated by the Ensembl VEP (**Figure S6**). To this end, we categorized VEP predictions into 14 categories (**Table S1**). The purpose of this was to test whether chCADD correctly scores SNPs with respect to their potential to cause a deleterious or phenotype-changing effect, as indicated (mostly for protein-coding mutations) by the VEP functional predictions. We observed that mutations with a relatively large deleterious potential, such as

stop-gained mutations and splice-site altering mutations, were scored higher than regular missense and synonymous mutations (**Figure S6**). SNPs in potentially regulatory active regions were also evaluated to be potentially more deleterious than synonymous SNPs (**Figure S6**). We performed a similar analysis considering only protein-coding and regulatory mutations found in the Online Mendelian Inheritance in Animals (OMIA) database (**Table 4**). We annotated only SNPs whose genomic positions were uniquely mapped to the chicken GRCg6a reference genome and the reference/alternative allele matched that in the genome assembly. Of the 15 annotated SNPs associated with a change of phenotype, 5 were reported to cause a deleterious phenotype change in the affected individual, and an average chCADD score of 27.1. These 5 variants (3 stop-gained, 2 missense) have a chCADD score above 20 and are putatively responsible for dwarfism, scaleless, analphalipoproteinaemia, muscular dystrophy, and wingless phenotypes (**Table 4**). All these phenotypes display a strong severity and may lead to an early death in uncontrolled environments.

chCADD detects evolutionary constraints within CNEs

As we showed, chCADD can score functionally important protein-coding variants. We therefore decided to take a step further by annotating SNPs found in CNEs with chCADD to predict their deleteriousness and function (**Table 3**). We assume that highly scored SNPs can help us to identify truly functionally active regions among CNEs. We observed that rare non-protein-coding variants located within CNEs (DAF ≤ 10%) have an overall higher chCADD score compared to rare variants found in non-CNEs (**Table 3**). This result supports our previous conclusion based on the derived allele frequency spectrum that evolutionarily conserved non-protein-coding variants are likely functional. As expected, this trend was most pronounced in lncRNAs, followed by introns and intergenic regions.

We further used the chCADD score to identify specific subregions of potentially higher functional importance within each CNE, assuming that the high scoring SNPs would indicate that. We applied a change point analysis to search for a center region that has high chCADD scores as opposed to the two outer regions (see Methods). We ranked CNEs based on positive chCADD score differences between the center region and the outer regions and filtered for significant difference (p-value of ≤ 0.05, t-test).

The top 3 ranked CNEs that overlap with lncRNAs, intronic and intergenic regions, respectively, are shown in Figure 3**A.1, B.1** and **C.1**.

Analogous to this subregion analysis based on chCADD score, we performed a subregion analysis based on the 23 sauropsids PhastCons scores. **A.2-C.2** show the identified regions for the PhastCons score for the same CNEs as **Figure 3A.1, 4C.1**, respectively. These figures indicate that chCADD generates more discriminative subregions than PhastCons. Particularly interesting are the chCADD scores for the top intergenic regions (**C.1**). The chCADD score increased from ~5 to ~15 at the subregion change point. This

is equal to an increase of predicted deleteriousness by one magnitude, from the top 33% highest scored sites in the entire genome to the top 3%.

To further investigate the subregion partitioning of the CNEs, we computed the SNP density in each region, for both the chCADD induced regions (Figure 4**,** blue bars) as well as the 23 sauropsids PhastCons induced regions (Figure 4**,** orange bars). In both bases, the SNP densities of the center region are lower than those of the outer regions. Moreover, all CNE subregions display a lower density than regions up- and downstream the CNE, supporting the functional importance of the CNEs in general. Interestingly, the center regions, as identified by the chCADD score, have in general a ~0.07% lower SNP density than the center regions detected using the PhastCons scores. Therefore, our findings suggest that chCADD is more effective in pinpointing potentially regions of interest.

Conserved non-protein-coding subregions are detected on the basis of a limited number of genomic annotations

As part of the investigation into subregions we identified two change points, splitting each CE into three subregions, starting from 5' to 3', 1st-, 2nd- and 3rd subregion (Figure 5**)**. Next we were interested how genomic annotations that were used in the creation of chCADD, differ between the three subregions. The model coefficients with the largest weights (**Table S2**) point to the importance of the PhastCons conservation scores calculated on the 4 sauropsids alignment. Other important model features are secondary structure predictions and combinations with the intronic identifier from VEP. Over all CNEs, we compared the chCADD model features, especially the conservation scores that are based on different phylogenies, excluding the chicken reference sequence in their computation. For all genomic annotations, we computed absolute Cohen's D values (standardized mean difference) (64,65). We observed that the conservation scores based on the largest 77 vertebrate alignments cannot properly distinguish between the 1st-,2nd- and 3rd subregions. Conservation scores based on smaller phylogenies (4 sauropsids and 37 amniote/mammalia) are more discriminative between these (**Table 5**; see columns **1st-2nd**, **2nd-3rd**).

Considering the three PhastCons scores, based on differently large phylogenies, the average absolute Cohen's D between the 1st- and 2nd- and the 2nd- to the 3rd- subregions differ less between different genomic features (intergenic, lncRNA and introns) than between genomic annotations (**Table 5**; see columns **1st-2nd**, **2nd-3rd**). The average absolute Cohen's D between the three subregions of a CNE ranges from 0.259 to 0.276. In comparison, the average absolute Cohen's D between the same subregions, taking the three conservation scores individually, range from 0.137 to 0.338. The effect sizes between the different multiple sequence alignment PhastCons score (i.e. 4 sauropsids, 37 amniote/mammalia, 77 vertebrates) differ by more than 2-fold.

Intronic CNE, differentially scored between the 1st , 2nd and 3rd subregions overlap functionally important genes

Intronic CNEs were associated with genes for which we obtained phenotype annotations of their orthologs in human, mouse, and rat. We investigated the top 10 CNEs that are located in introns, with the largest p-value differences between the 1st and 3rd to the 2nd section. 6 CNEs were associated with homologous genes that have annotated phenotypes in other species. Among the phenotypes found for human genes are mental retardation and non-syndromic male infertility. For mouse, these included neuronal issues and abnormal shape of heart and limbs (**Table S3**). The link to highly severe phenotypes in other species highlights the potential importance of regulatory features for orthologous genes in chicken.

Discussion

The prediction of CNEs depend on the phylogenetic scope

Non-protein-coding elements are typically identified by sequence-level similarity across species, which is a generally applicable criterion of conservation and biological function (10). However, when predicting CEs, and subsequently CNEs, the evolutionary distance among species included in the alignment (or phylogenetic scope) is an important parameter that can considerably affect the prediction and resolution of CEs. If the evolutionary distance among species is too narrow, the specificity of constraint is reduced, but if it is too broad, the number of CEs rapidly declines and lineage-specific conservation is lost (10,37).

One of the first studies to address the impact of the phylogenetic scope on CEs prediction was that of (12). In their study on the 29 mammalian multiple sequence alignment the authors identified 3.6 million conserved elements spanning 4.2% of the genome at a resolution of 12 bp (12). When comparing these results to a 5-vertebrate alignment, Lindblad-Toh and colleagues observed that only 45% of the 5-taxa CEs were covered by the 29-taxa alignment. This partial overlap indicates that most of the CEs derived from the 29-taxa alignment were mammalian-specific (12). The issue resulting from a broad phylogenetic scope on CNEs has also recently been reported by (38) where authors identified CNEs between chicken and four mammalian species, including human, mouse, dog, and cattle (38). By applying a minimum length of 100 bp, Babarinde and Saitou (2016) identified 21,584 CNEs in chicken, a small number as expected from the divergence time between human and chicken ~310 million years ago (33). Therefore, CNEs detected among distant species are better predictions of ultraconserved CNEs than CNEs between closely related species (i.e. human-mouse) (39), as they were already present in the ancient common ancestor of the considered species.

In this study we chose the 23 sauropsids multiple sequence alignment for two reasons. First, the phylogenetic distance between crocodilian and bird species (240 million years ago) (40) is large enough to detect likely functional CNEs. Second, the alignment is reference free allowing the identification of

lineage-specific CEs. Reference-free alignments should always be preferred over reference-based ones (41). In fact, genomic regions shared within a certain clade, which would be missed in a reference-based alignment (e.g. MULTIZ), can also be detected. As a result, reference free alignments better enable the study of genome evolution along all phylogenetic branches equally.

Avian genomes have similar genomic characteristics

According to our study, 8% of the chicken genome is covered by CEs for a total of 1.14 million CEs. These results are comparable to those on the collared flycatcher genome (*Ficedula albicollis*) (8). By means of the same alignment, (8) identified 1.28 million CEs covering 7% of the flycatcher genome. Compared to the flycatcher, the slightly lower number of CEs we report in chicken could be explained by its smaller genome size, as small genomes require fewer regulatory sequences involved in the organization of chromatin structure (8). For instance, the chicken genome is nearly 4 times smaller (i.e. GRCg6a: 1.13 Gb) than that of human (i.e. GRCh38.p13: 4.53 Gb), but of nearly equal size to that of the collared flycatcher (i.e. FicAlb1.5: 1.11 Gb). The similarity in genome size between chicken and flycatcher reflects the little cross-species variation characteristic of birds (42).

The limited number of CEs often identified in birds relative to mammals has repeatedly been linked to gene loss (23,25,43). However, the role of gene loss in avian evolution, genome size, and prediction of CEs has recently been questioned. According to (26), gene loss was incorrectly hypothesized from the absence of genes clustering in GC-rich regions in the earlier chicken genome assemblies (26). In fact, these regions are often difficult to sequence and assemble. This issue is particularly prominent in the GC-rich micro-chromosomes, which, as we show, contribute disproportionately to the total density of functional sequence (**Figure S1**). We therefore recommend future comparative genomics studies in chicken to make use of the most recent and complete genome assembly to avoid any erroneous link of CEs to gene loss in chicken

Conserved non-protein-coding elements are maintained by purifying selection

A fundamental question in the study of CNEs is the role of purifying selection. Purifying selection can be discriminated from a low mutation rate by comparing the derived allele frequency (DAF) spectra in constrained regions (i.e. CNEs) with that of neutral regions (i.e. non-CNEs) (9,35). This is because new mutations are unlikely to increase in frequency in constrained regions. Although CNEs are identified using an interspecific comparative genomic approach, the evolution and dynamics of these regions are generally analyzed at an intraspecific scale by looking at polymorphism data (9,44). In this study, we showed that the evolutionary constraint acting on the 23 sauropsids is correlated with constraint within the chicken populations, as assessed from chicken polymorphism data. Consistent with studies in humans

(12,35), plants (6), and *Drosophila (9,36)*, the derived allele frequency spectra of our chicken populations is shifted towards an excess of rare variants in CNEs. These results indicate that the conservation of CNEs in the chicken genome is mainly driven by selective constraints, and not by local variation in mutation rate. The role of purifying selection was also confirmed by the reduced SNP density in CNEs compared to non-CNEs and by the reduced SNP density in specific conserved non-protein-coding subregions. The concordance in SNP density is a clear indication of reduced levels of population diversity and functional roles of CNEs as confirmed by the association of subregions within CNEs to highly severe phenotypes in humans, mouse, and rat. However, future population diversity comparisons in terms of nucleotide diversity (π) (45) or Watterson's estimator (θw) (46) between outbred and inbred populations would further elucidate our understanding of purifying selection in CNEs.

Integrating comparative and functional genomics into a single score

We developed a ch(icken) Combined Annotation-Dependent Depletion (chCADD) approach that provides scores for all SNPs throughout the chicken genome. These scores are indicative of putative SNP deleteriousness and can be used to prioritize variants.

The annotation of chCADD relies on the combination of a diverse set of genomic features, including evolutionary constraints and functional data (21,22). Multiple sequence alignments of distantly related species are better suited to differentiate conserved sites that can reliably be used to identify functionally important regions. However, these regions are often large enough to question the functional role of the entire region. Our findings show that chCADD outperforms any conservation-based method alone (e.g. PhastCons) in the identification of functionally important subregions within CNEs. Therefore, methods, such as chCADD, are required to fine-tune in one step CNEs to identify subregions directly linked to - in some cases deleterious – phenotypes.

According to the authors of the original human CADD (21), SNPs with a score above 20 (i.e. the SNP is among the top 1% highest scored potential SNPs in the genome) could be considered deleterious. This means that the higher the score, the higher the chance the variant has a functional effect or may even be deleterious. When annotating protein-coding and regulatory mutations found in OMIA, we observed that SNPs with a chCADD score of 15 can already be considered functional. Therefore, our findings indicate that by setting an arbitrary threshold of 20 may underestimate the fraction of the genome that is actually functional. This is particularly pronounced when the variants in question are located outside protein-coding regions. Therefore we recommend future chCADD users to evaluate the variants identified in their populations to see if they are particularly highly scored compared to other variants in the same genomic region.

Future uses of chCADD

The high scoring of non-protein-coding variants in subregions of CNEs has important implications for future functional and genome-wide association studies (GWAS) in chicken. A very large fraction of trait- or disease-associated loci identified in GWAS are intronic or intergenic. This is expected considering the preponderance of non-protein-coding SNPs on genotyping arrays (5) or along the genome. However, because of a lack of understanding of the function of non-protein-coding mutations, most of the causal mutations reported in the OMIA database are coding. Moreover, in the presence of non-protein-coding mutations, many studies stop at the general locus or - understandably - assume that the closest neighboring gene is affected. However, these assumptions on genomic distance are simplistic. Our findings in chicken demonstrate that chCADD can accurately pinpoint non-protein and protein-coding variants associated with important phenotypes in chicken. Therefore we expect future genome-wide association studies combined with chCADD to identify novel causal mutations or substantially narrow down the list of potential causal variants in large quantitative trait loci (QTLs). We also expect chCADD to accelerate the discovery and understanding of the biology and genetic basis of phenotypes.

Conclusions

Deciphering the function of the non-coding portion of a species genome has been a challenging task. However, the availability of genomes from a great variety of species, along with the development of new computational approaches at the interface of machine learning and bioinformatics, has made this task possible in model and non-model organisms. Our findings indicate an accurate assessment of selective pressure at individual sites becomes an achievable goal. We have also shown that chCADD is a reliable score for the analysis of non-protein-coding SNPs, which should be targeted along with protein-coding mutations in future genome-wide association studies. We therefore anticipate chCADD to be of great use to the scientific community and breeding companies in future functional studies in chicken.

Materials and methods

Chicken genomic data

We used a dataset by Bortoluzzi and colleagues available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession number PRJEB34245 (47) and PRJEB36674 (18). The 169 chicken samples included in the dataset were sequenced at the French Institute of Agricultural Research (INRA), France, on an Illumina HiSeq 3000. Reads were processed following standard bioinformatics pipelines. Reads were aligned to the chicken GRCg6a reference genome (GenBank Accession:

GCA_000002315.5) with the Burrows-Wheeler alignment (BWA-mem) algorithm v0.7.17 (48). After removal of duplicate reads with the *markdup* option in sambamba v0.6.3 (49), we performed population-based variant calling in Freebayes (50), retaining only sites with a mapping and base quality >20. We reduced the false discovery rate by additional filtering using BCFtools v1.4.1 (48).

Multiple whole-genome sequence alignment

Conserved elements (CE) were identified using the 23 sauropsids multiple whole-genome sequence alignment (MSA) generated using Progressive Cactus (https://github.com/glennhickey/progressiveCactus) (51) by (40). The MSA downloaded in the hierarchical alignment format (HAL) was converted into multiple alignment format (MAF) using the HAL tools command *hal2maf (52)* with the following parameters: -refGenome galGal4 (GenBank Accession: GCA_000002315.2) to extract alignments referenced to the chicken genome assembly, -noAncestors to exclude any ancestral sequence reconstruction, -onlyOrthologs to include only sequences orthologous to chicken, and -noDupes to ignore paralogy edges. During reformatting, only blocks of sequences where chicken aligned to at least two other species were considered for a total chicken genome alignability of 90.88%. Genomic coordinates were converted to the GRCg6a genome assembly using the pyliftover library in python v3.6.3.

Prediction of evolutionarily conserved elements

Conserved elements were predicted from the whole-genome alignment using PhastCons (53). We chose PhastCons because this approach does not use a fixed-size window approach, but can take advantage of the fact that most functional regions involve several consecutive sites (54). We first generated a neutral evolutionary model from the 114,709 four-fold degenerate (4D) sites previously extracted from the alignment by (40). The topology of the phylogeny was also identical to that derived by (40). PhastCons was run using the set of parameters used by the UCSC genome browser to produce the 'most conserved' tracks (top 5% of the conserved genome): expected length = 45, target coverage = 0.3, and rho = 0.31 (32). Conserved elements were subsequently excluded if falling or overlapping assembly gaps and/or if their size was < 4 bp.

Annotation of conserved elements by genomic feature

We use the Ensembl (release 95) chicken genome annotation files to extract sequence coordinates of CDS, exons, 5' and 3' UTRs, pseudogenes, and lncRNAs. Sequence information was extracted from 14,828 genes (out of the 15,636 genes found in the Ensembl annotation), as transcripts of these genes had a properly annotated start and stop codon. For protein-coding genes with an annotated 5' UTR of at least 15 bp, the promoter was defined as the 2-kb region upstream of the transcription start site (TSS) (8).

Sequence coordinates of miRNAs, rRNAs, snoRNAs, snRNAs, ncRNAs, tRNAs, and scRNAs were also extracted from the annotation file. For the identification of intergenic regions we considered all annotated protein-coding genes and defined intergenic regions as DNA regions located between genes that did not overlap any protein-coding genes in either of the DNA strands. The intersection between CEs and the various annotated genomic features was found following the approach of (12) of assigning a CE overlapping two or more genomic features to a single one in a hierarchical format: CDS, 5' UTR, 3' UTR, promoter, RNA genes, lncRNA, intronic, and intergenic region. Conserved non-protein-coding elements (CNEs) were defined as CEs without any overlap with exon-associated features (CDS, 5' UTR, 3' UTR, promoter, and RNA genes) and include lncRNAs, introns, and intergenic regions.

Gene ontology analysis

Genes in conserved regions overlapping CDS, 5' UTR, 3' UTR, and introns were separately used to perform a Gene Ontology analysis in g:Profiler (55) using *Gallus gallus* as organism. We only considered annotated genes that passed Bonferroni correction for multiple testing with a threshold < 0.05.

Genome-wide distribution and density of conserved non-protein-coding regions

CNE density and the density of exon-associated features were calculated in non-overlapping 100 kb windows along the genome. Windows that included assembly gaps between scaffolds were discarded, resulting in a total of 9,196 windows. Correlation between density of exons and CNEs was calculated in R v3.2.0 using the Pearson's correlation test.

Annotation of variants by functional class

Polymorphic, bi-allelic SNPs belonging to all functional classes predicted by the Variant Effect Predictor (VEP) (56) were considered. However, to improve the reliability of the set of annotated variants, we applied additional filtering steps. SNPs were discarded if they overlapped repetitive elements or if their call rate was <70%. The rationale for excluding variants found in repetitive elements was to reduce erroneous functional prediction as a result of mapping issues, as regions enriched for repetitive elements are usually difficult to assemble. Intronic and intergenic SNPs were further discarded if they overlapped spliced intronic ESTs (35). Protein-coding variants were also discarded if they were found outside coding sequences, whose genomic coordinates were obtained from the Ensembl chicken GTF file (release 95).

Ancestral allele and derived allele frequency

The sequence of the inferred ancestor between chicken and turkey (*Meleagris gallopavo*; Turkey_2.01) (57) reconstructed from the Ensembl EPO 4 sauropsids alignment (release 95) was used to determine the

ancestral and derived state of an allele, along with its derived allele frequency. We considered only SNPs for which either the reference or alternative allele matched the ancestral allele. Ancestral alleles that did not match either chicken allele were discarded. We generated derived allele frequency (DAF) distributions for sets of SNPs based on functional class and whether they were within or outside of CNEs. A derived allele frequency cutoff of 10% was used to distinguish rare from common SNPs.

Chicken Combined Annotation Dependent Depletion (chCADD)

The chicken CADD scores are the -10 log relative ranks of all possible alternative alleles of all autosomes and Z chromosome of the chicken GRCg6a reference genome, according to the following formula:

$$chCADD_i = -10\log_{10}\left(\frac{n_i}{N}\right)$$

where N represents the number of all possible alternative alleles (3,073,805,640) on the investigated chromosomes and n is the rank of the i^{th} SNP. The ranks are based on the model posteriors of a ridge penalized logistic regression model trained to classify simulated and derived SNPs.

Chicken derived SNPs were defined as those sites where the chicken reference genome differs from the chicken-turkey ancestral genome inferred from the Ensembl EPO 4 sauropsids alignment. Sites for which the ancestral allele occurs at a minor allele frequency greater than 5% were excluded. In addition, derived SNPs that are observed with frequency above 90% in our population of 169 individuals were included. In total we identified 17,237,778 SNPs.

The dataset of simulated variants was simulated based on derived nucleotide substitution rates between the inferred ancestor of chicken, turkey, zebra finch (*Taeniopygia guttata*; taeGut3.2.4) (58) and green anole lizard (*Anolis carolinensis*; AnoCar2.0) (59). These derived nucleotide substitution rates were obtained for windows of 100 kb and used to simulate *de novo* variants which have a larger probability to have a deleterious effect than the set of derived variants. All SNPs which have a known ancestral site are retained in the dataset. In total 17,233,727 SNPs were simulated in this way. 17,233,722 SNPs of each dataset were joined and randomly assigned to train and test sets of sizes 15,667,020 and 1,566,702, respectively.

The datasets were annotated with various genomic annotations: among others, PhyloP and PhastCons (**Table S4**) conservation scores based on three differently deep phylogenies (i.e. 4 sauropsids, 37 amniote/mammalia, 77 vertebrate, all excluding the chicken genome), secondary DNA structure predictions (**Table S4**), Ensembl Consequence predictions, amino acid substitution scores such as Grantham (**Table S4**) and amino acid substitution deleterious scores such as SIFT (**Table S4**).

Annotations for which values were missing were imputed, categorical values were one hot-encoded (60). In the one hot-encoding process, an annotation is a series of binary annotations, each indicating the

presence of a specific category for a given variant. For scores that are by definition not available for certain parts of the genome, such as SIFT which is found only for missense mutations, columns indicating their availability were introduced.

Combinations of annotations were created of Ensembl Variant Effect Predictor consequences and other annotations, such as distance to transcription start site and conservation scores. The total number of all features used in training was 874. An extensive list of all annotations, combinations of annotations and their learned model weights is shown in Supplementary File 2. Finally, each feature column is scaled by its standard deviation. The logistic regression is trained via the Python Graphlab module. We selected a penalization term of 1, based on results on the test set (**Figure S5**).

Investigation of likely causal SNPs from the OMIA database

We downloaded the likely causal variants of phenotype changes from the Online Mendelian Inheritance in Animals (OMIA) (61) database (last accessed 25.11.2019). SNPs whose location was reported for older genome assemblies such as Galgal4 and Galgal5 were mapped to the chicken GRCg6a reference genome via CrossMap (62). We only consider bi-allelic SNPs whose genomic position was successfully mapped to GRCg6a and whose substitution remained the same. In total, 15 SNPs were left and annotated with chCADD.

Change point analysis

To identify sub-regions of particular importance within each CE, we annotated all with the maximum chCADD score found at each site or the 23-sauropsids PhastCons scores that were used to identify conserved elements in the first place. Our basic assumption was that highly important subregions within a CE are preceded and succeeded by less important sites which would result in a relatively higher score region surrounded by two lower scored regions. Each CE was treated similarly to time series data by conducting an offline change point analysis, once based on maximum chCADD scores and once based on 23-sauropsids PhastCons scores. To this end, we used the Python ruptures module (63) and applied a binary segmentation algorithm with radial basis function (RBF). It first identifies a single change point, if one is detected, the the algorithm investigates each sub-sequence independently to identify the next change point We were looking particularly for 2 change points, which would divide the CE into three subregions, numbered from 1 to 3, starting at the 5’ end of the sequence. We added 5 bp upstream and downstream of each CE to allow that the borders of the 2nd region coincide with the borders of the CE (Figure 5). After computing the change points, we conducted t-tests between the scores of the 1st and 2nd, as well as 3rd and 2nd subregions, to identify CEs that have a significantly different score in the 2nd section than in the other two. We applied a *p*-value cutoff of 0.05. We sorted CNEs with respect to the largest

difference between the mean chCADD score of the inner and the two outer subregions and selected those with a higher scored 2nd section than either of the other two outer ones.

SNP density distribution within conserved non-protein-coding regions

SNP density was calculated as the number of SNPs identified in the 169 chicken individuals divided by the number of bases found in the sequence. SNP density was computed for conserved coding (CC) and conserved non-protein-coding (CNE) regions, as well as for the subregions identified in the change point analysis of CNEs overlapping lncRNAs, introns, and intergenic regions. We repeated this analysis once for the change points identified using chCADD scores and once for the 23-sauropsids PhastCons based change points.

Homologous phenotypes

We obtained phenotypes from the Ensembl database (release 95) for genes associated with the lncRNA and intronic CNEs. Beside chicken, these phenotypes encompass the observed phenotypes for orthologous genes associated with disease studies in humans (*Homo sapiens*) and gene-knockout studies in mouse (*Mus musculus*) and rat (*Rattus norvegicus*).

Data access

Raw sequences the 169 individuals used in this study are available at the European Nucleotide Archive under accession number PRJEB34245 and PRJEB36674. chCADD scores partitioned per chromosomes can be downloaded from the Open Science Framework project page(https://osf.io/d6wxp/).

Acknowledgement

C.G. was funded by the TTW-Breed4Food Partnership, project number 14283: From sequence to phenotype: detecting deleterious variation by prediction of functionality. C.B was funded by the European Union´s Horizon 2020 Research and Innovation Programme under the Grant Agreement No. 677353 (Innovative Management of Animal Genetic Resources – IMAGE). M.B. is financially supported by the NWO-VENI grant no.016.Veni.181.050.

Disclosure declaration

The authors declare that they do not have any conflict of interest.

List of figures

Figure 1. Correlation between exons and conserved non-protein-coding elements (CNEs) along the chicken genome. CNEs and exons count per 100 kb windows are shown with the Pearson correlation coefficient *r* and corresponding *p-value* in the top left corner.

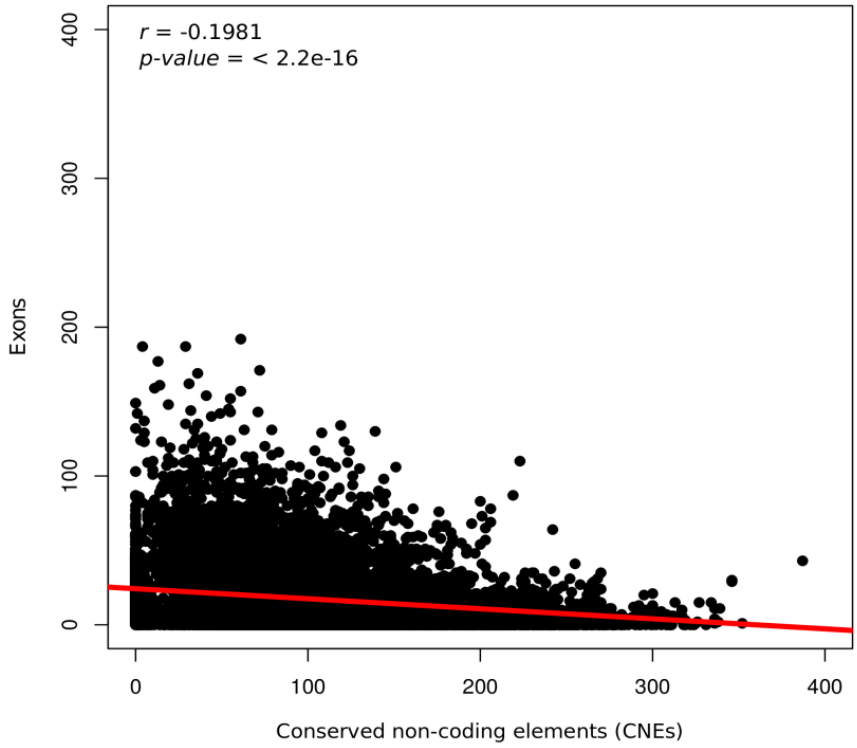


Figure 2. Derived allele frequency (DAF) distribution of SNPs in CNEs and non-CNEs.

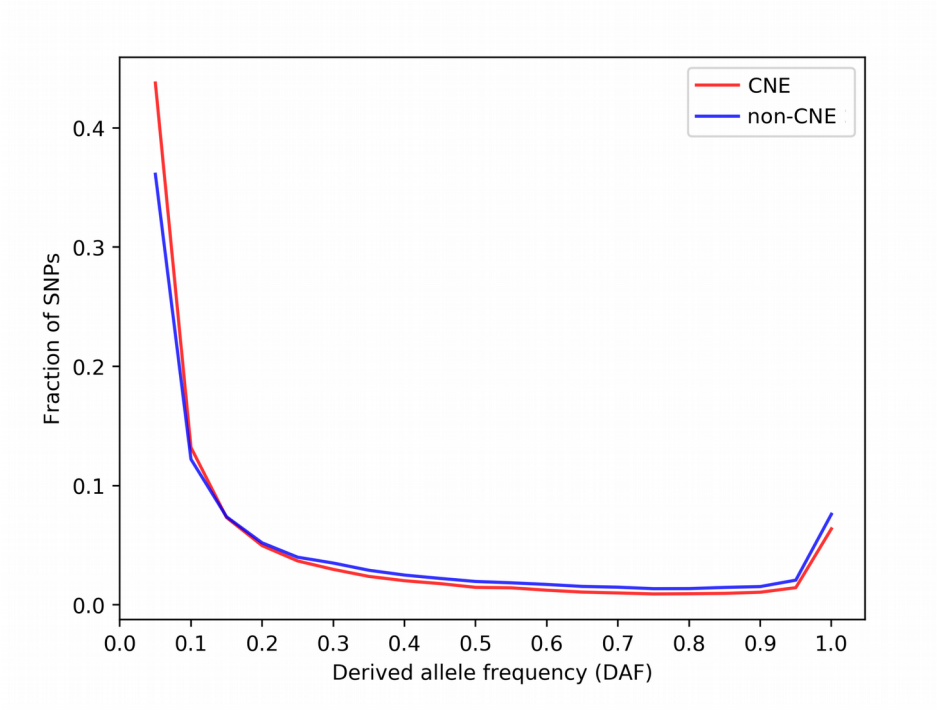


Figure 3. Change point analysis of the top 3 CNEs for each genomic feature, respectively (lncRNA, intronic, intergenic). CNEs are sorted based on the largest difference between the 2nd section and 1st or 3rd section for each of the three CNE classes respectively (lncRNA, intronic, intergenic). Change points were once computed based on maximum chADD score per site (A.1,B.1,C.1) and once on 23 sauropsids PhastCons scores (A.2,B.2,C.2). The dots in each plot display the scores for the 5 bp up- and downstream regions. The transition from blue to red background indicates the identified change points. A.1) lncRNA - maximum chCADD A.2) lncRNA - PhastCons scores. B.1) intronic - maximum chCADD. B.2) intronic - PhastCons. C.1) intergenic - maximum chCADD. C2) intergenic - PhastCons.

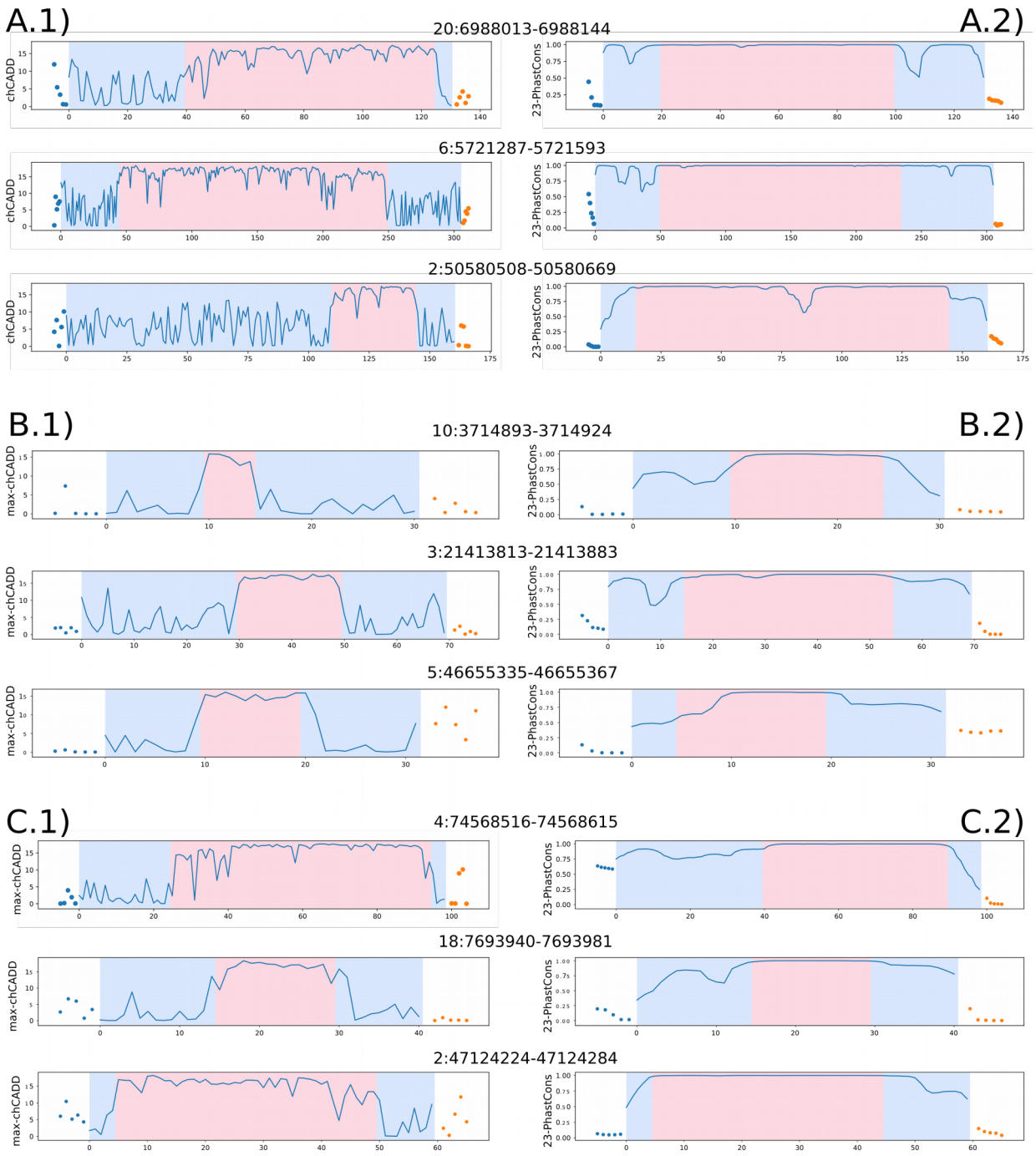


Figure 4. SNP densities computed for each section of the three different CNEs (lncRNA, Intronic, Intergenic). The orange bars represent the SNP densities for that section based on change points derived from 23 sauropsids alignment PhastCons scores, the blue bars represent the SNP densities based on change points identified via chCADD.

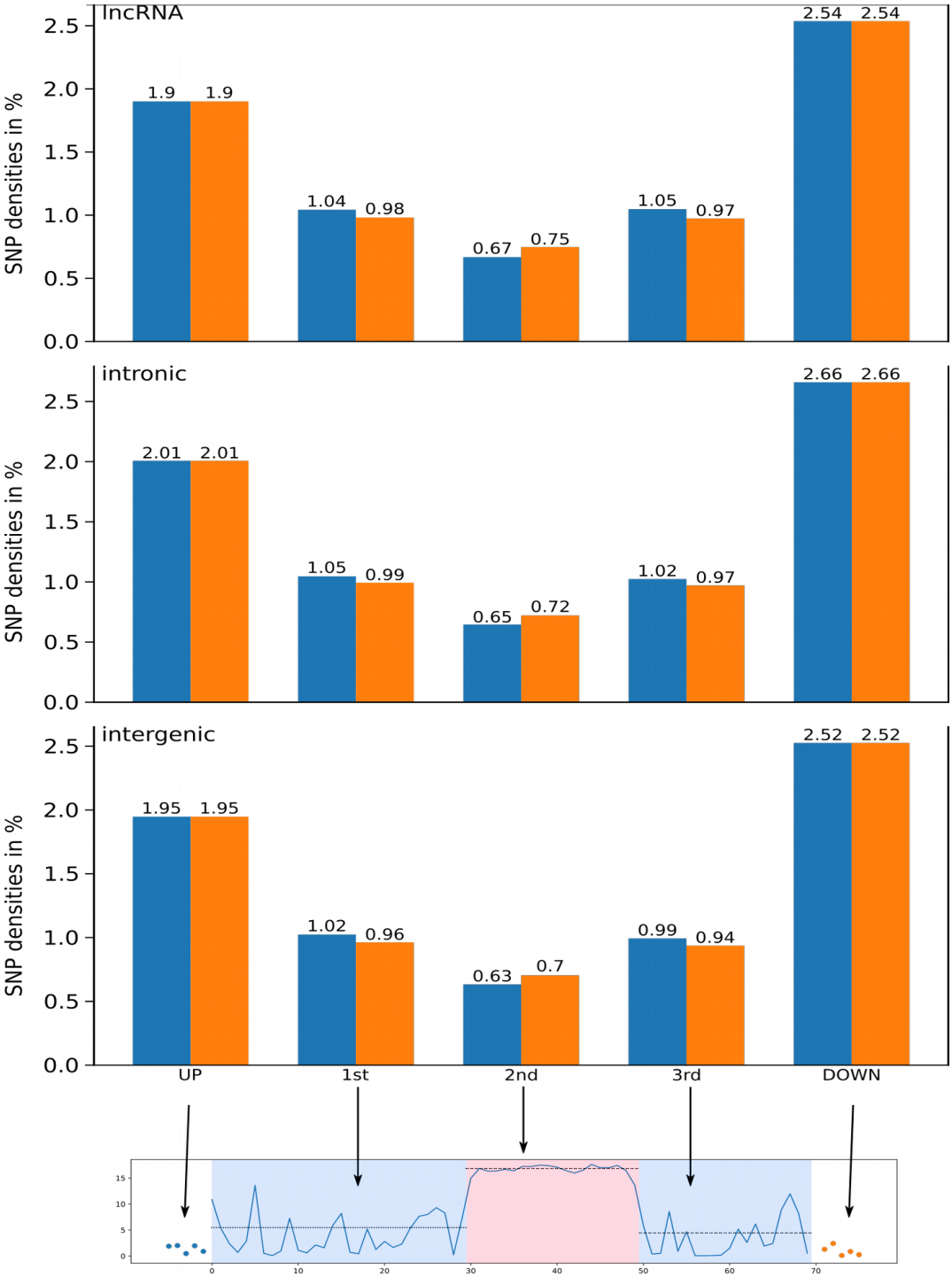
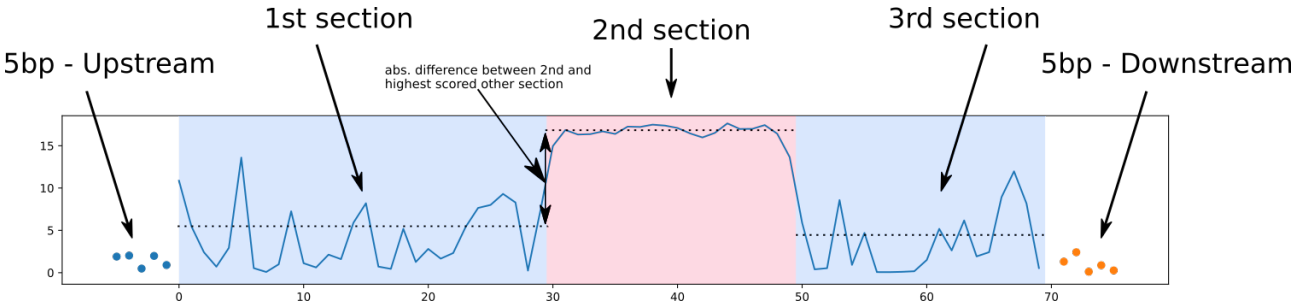


Figure 5. Approach used to identify subregions within CNEs via change point analysis. The scores used to annotate the CE region are displayed on the y-axis. The position in the investigated CE region is shown on the x-axis. In total there are five sections, 5 bp up and downstream, 1st, 2nd and 3rd subregions. The transitions from blue to red background indicate the position of the two identified change points. The up and downstream scores are shown as dots while the scores in the CE region are a continuous blue line.



List of tables

Table 1. Statistics of predicted conserved elements (CEs) based by gene annotations. The fraction of CEs per sites class is presented, for protein-coding gene annotations, in percentages of the exonic CEs (17,148,879 bp). For non-protein-coding gene annotations, the fraction is relative to the non-exonic CEs (51,224,645 bp). Abbreviations: CC, conserved coding; CNE, conserved non-protein-coding elements

Genomic feature	No. overlapping CEs	Total overlap (bp)	Genome coverage (%)	Fraction of site class conserved (%)
CDS	213,787	14,683,183	1.38	85.62
5’ UTRs	5,457	207,320	0.02	1.21
3’ UTRs	23,721	1,460,144	0.15	8.51
Promoters	16,022	761,504	0.08	4.44
RNA genes	701	36,728	0.00	0.21
LncRNAs	121,840	7,696,557	0.80	15.03
Introns	328,579	18,520,675	1.93	36.16
Intergenic	400,501	25,007,413	2.60	48.82
Total CC	**259,688**	**17,148,879**	**1.78**	**100.00**
Total CNE	**850,920**	**51,224,645**	**5.33**	**100.00**

Table 2. GO term enrichment analysis of exonic-associated CE and intronic CEs

Term ID	Term description	Target size	3 UTR				Intron			
			Term size	Query size	Overlap size	p-value	Term size	Query size	Overlap size	p-value
GO:0048856	Anatomical structure development	12,514	3,293	4,736	1,475	$1.24e^{-17}$	3,293	6,971	2,128	$1.09\ e^{-29}$
GO:0010646	Regulation of cell communication	12,514	2,038	4,736	917	$3.67\ e^{-09}$	2,038	6,971	1,329	$1.33\ e^{-17}$
GO:0010604	Positive regulation of macromolecule metabolic process	12,514	2,118	4,736	952	$1.49\ e^{-09}$	2,118	6,971	1,331	$2.21\ e^{-09}$
GO:0023051	Regulating of signaling	12,514	2,056	4,736	926	$2\ e^{-09}$	2,056	6,971	1,339	$1.88\ e^{-17}$
GO:0048583	Regulation of response to stimulus	12,514	2,332	4,736	1,032	$1.44\ e^{-08}$	2,332	6,971	1,477	$9.79\ e^{-13}$
GO:0048468	Cell development	12,514	1,364	4,736	625	$1.27\ e^{-06}$	1,364	6,971	927	$1.12\ e^{-18}$
GO:0031325	Positive regulation of cellular metabolic process	12,514	2,091	4,736	936	$9.01\ e^{-09}$	2,091	6,971	1,304	$1.09\ e^{-07}$
Term ID	**Term description**	**Target size**	**CDS**				**5 UTR**			
			Term size	**Query size**	**Overlap size**	**p-value**	**Term size**	**Query size**	**Overlap size**	**p-value**
GO:0048856	Anatomical structure development	12,514	3,293	9,703	2,713	$2.06\ e^{-11}$	3,293	1,896	654	$5.13\ e^{-14}$
GO:0010646	Regulation of cell communication	12,514	2,038	9,703	1,686	$2.64\ e^{-06}$	2,038	1,896	381	$9.33\ e^{-03}$
GO:0010604	Positive regulation of macromolecule metabolic process	12,514	2,118	9,703	1,749	$3.53\ e^{-06}$	2,118	1,896	403	$5.06\ e^{-04}$
GO:0023051	Regulating of signaling	12,514	2,056	9,703	1,699	$4.46\ e^{-06}$	2,056	1,896	384	$9.24\ e^{-03}$
GO:0048583	Regulation of response to stimulus	12,514	2,332	9,703	1918	$5.55\ e^{-06}$	2,332	1,896	424	$4.39\ e^{-02}$
GO:0048468	Cell development	12,514	1,364	9,703	1,142	$1.78\ e^{-05}$	1,364	1,896	282	$3.38\ e^{-05}$

GO:0031325	Positive regulation of cellular metabolic process	12,514	2,091	9,703	1,723	1.91 e^{-05}	2,091	1,896	388	1.60e^{-02}

Table 3. Derived allele frequency distribution for SNPs in CNEs and non-CNEs by SNP functional class.

Genomic feature	DAF	Within CNEs	Outside CNEs	chCADD within CNEs	chCADD outside CNEs
		Number of SNPs (%)	Number of SNPs (%)	Average (± sd)	Average (± sd)
All	≤0.10	137,871 (57%)	482,685 (48.4%)	9.78 (4.18)	3.21 (3.18)
	> 0.10	103,726 (43%)	513,935 (51.5%)	8.81 (4.25)	2.74 (2.83)
LncRNA	≤0.10	24,364 (57.4%)	26,429 (47.6%)	10.02 (4.00)	3.49 (3.33)
	> 0.10	18,081 (42.5%)	29,014 (52.4%)	9.10 (4.13)	3.03 (2.99)
Intron	≤0.10	43,790 (56.8%)	159,203 (47.4%)	9.81 (4.46)	3.00 (3.11)
	> 0.10	33,171 (43.2%)	176,650 (52.6%)	8.71 (4.53)	2.46 (2.74)
Intergenic	≤0.10	69,717 (57%)	297,053 (44.6%)	9.68 (4.05)	3.31 (3.20)
	> 0.10	52,474 (43%)	308,271 (55,4%)	8.78 (4.11)	2.87 (2.86)

Table 4. OMIA chicken SNPs with chCADD annotations, locations are reported for Gal6.

OMIA ID(s)	Variant Phenotype	Gene	Type of Variant	Deleterious?	g. or m.	chCADD
OMIA 001622-9031	Resistance to avian sarcoma and leukosis viruses, subgroup C	BTN1A1	stop-gain	no	28:g.903289G>T	17.83409
OMIA 000889-9031	Scaleless	FGF20	stop-gain	yes	4:g.63270401A>T	33.02083
OMIA 001534-9031	Resistance to myxovirus	MX1	missense	no	1:g.110260061G>A	14.26893
OMIA 000915-9031	Feather colour, silver	SLC45A2	missense	no	Z:g.10336596G>T	21.72641
OMIA 000915-	Feather colour, silver	SLC45A2	missense	no	Z:g.10340909T>C	15.69336

9031						
OMIA 000679-9031	Muscular dystrophy	WWP1	missense	yes	2:g.123014353G>A	26.29866
OMIA 000303-9031	Dwarfism, autosomal	C1H12ORF23	stop-gain	yes	1:g.53638233C>T	35.29646
OMIA 001302-9031	Resistance to avian sarcoma and leukosis viruses, subgroup B	TNFRSF10B	stop-gain	no	22:g.1418711C>T	17.63145
OMIA 000810-9031	Polydactyly	LMBR1	regulatory	yes	2:g.8553470G>T	17.41378
OMIA 000913-9031	Silky/Silkie feathering	PDSS2	regulatory	unknown	3:g.67850419C>G	3.8812
OMIA 001547-9031	Wingless-2	RAF1	stop-gain	yes	12:g.5374854G>A	23.44641
OMIA 000374-9031	Feather colour, extended black	MC1R	missense	no	11:g.18840857T>C	18.05882
OMIA 000374-9031	Feather colour, extended black	MC1R	missense	no	11:g.18840919G>A	18.88983
OMIA 000374-9031	Feather colour, buttercup	MC1R	missense	no	11:g.18841289A>C	17.41773
OMIA 000374-9031	Feather colour, extended black	MC1R	regulatory; 5 'UTR	no	11:g.18840609C>T	6.74322

Table 5. Differences between genomic annotations utilized for the chCADD model, between CNE subregions defined by chCADD located in intronic, lncRNA and intergenic regions, measured in absolute Cohen's D.

INTRONIC	**UP-1st**	**1st-2nd**	**2nd-3rd**	**3rd-Down**
4PhastCons	0.594	0.307	0.361	0.609
37PhastCons	0.446	0.328	0.369	0.448
77PhastCons	1.25	0.096	0.195	1.32
4PhyloP	0.43	0.09	0.126	0.428
37PhyloP	0.351	0.187	0.214	0.35
77PhyloP	0.776	0.186	0.237	0.778
GerpS	0.272	0.182	0.196	0.257
GerpN	0.212	0.112	0.11	0.214
dnaMGW	0.103	0.009	0.007	0.104
dnaProT	0.08	0.013	0.012	0.08
dnaHelT	0.082	0.002	0.002	0.083
GC	0.121	0.045	0.047	0.12
CpG	0.034	0.034	0.034	0.034
OChrom-Peaknb	0.058	0.001	0.091	0.015
OChrom-logFC	0.062	0.087	0.138	0.017
OChrom-pval	0.006	0.013	0.070	0.055
lncRNA	**UP-1st**	**1st-2nd**	**2nd-3rd**	**3rd-Down**
4PhastCons	0.608	0.289	0.338	0.623
37PhastCons	0.469	0.31	0.342	0.482
77PhastCons	1.29	0.086	0.184	1.37
4PhyloP	0.428	0.083	0.117	0.43
37PhyloP	0.343	0.161	0.18	0.348
77PhyloP	0.788	0.17	0.22	0.792
GerpS	0.267	0.17	0.181	0.259
GerpN	0.212	0.086	0.098	0.201
dnaMGW	0.097	0.006	0.008	0.095
dnaProT	0.096	0.009	0.009	0.093
dnaHelT	0.089	0.003	0.0	0.086
GC	0.114	0.037	0.041	0.109
CpG	0.024	0.033	0.029	0.028
OChrom-Peaknb	0.059	-0.02	0.064	0.023
OChrom-logFC	0.102	0.093	0.137	0.055
OChrom-pval	0.012	0.096	0.103	0.005
INTERGENIC	**UP-1st**	**1st-2nd**	**2nd-3rd**	**3rd-Down**
4PhastCons	0.61	0.281	0.341	0.619
37PhastCons	0.474	0.319	0.359	0.481
77PhastCons	1.29	0.084	0.179	1.37
4PhyloP	0.431	0.084	0.119	0.432
37PhyloP	0.351	0.162	0.185	0.351
77PhyloP	0.79	0.167	0.215	0.795
GerpS	0.29	0.169	0.183	0.274
GerpN	0.209	0.091	0.088	0.215
dnaMGW	0.096	0.008	0.008	0.096
dnaProT	0.097	0.014	0.012	0.096
dnaHelT	0.086	0.003	0.002	0.084
GC	0.136	0.062	0.062	0.136
CpG	0.039	0.037	0.036	0.041
OChrom-Peaknb	0.017	0.004	0.02	0.005
OChrom-logFC	0.089	0.005	0.012	0.077
OChrom-pval	0.00	0.005	0.052	0.023

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;412(6846):565–6.
2. Feingold EA, Good PJ, Guyer MS, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia of DNA Elements) Project. Science (80-). 2004;306(5696):636–40.
3. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799–816.
4. Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. 2007;17(6):760–74.
5. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. Nat Rev Genet. 2010;11(8):559–71.
6. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. 2013;45(8):891–8.
7. Alföldi J, Lindblad-Toh K. Comparative genomics as a tool to understand evolution and disease. Genome Res. 2013;23(7):1063–8.
8. Craig RJ, Suh A, Wang M, Ellegren H. Natural selection beyond genes: Identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (Ficedula albicollis). Mol Ecol. 2018;27(2):476–92.
9. Berr T, Peticca A, Haudry A. Evidence for purifying selection on conserved noncoding elements in the genome of Drosophila melanogaster. bioRxiv. 2019;623744.
10. Harmston N, Barešić A, Lenhard B. The mystery of extreme non-coding conservation. Philos Trans R Soc B Biol Sci. 2013;368(1632).
11. Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat Genet. 2016;48(4):427–37.
12. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478(7370):476–82.
13. Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, et al. Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. PLoS Genet. 2013;9(12).
14. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A. in Recent Human Evolution. Science

(80-). 2011;257(February):920–4.

15. Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, et al. Evidence for Widespread Positive and Negative Selection in Coding and Conserved Noncoding Regions of Capsella grandiflora. PLoS Genet. 2014;10(9).
16. Marcovitz A, Jia R, Bejerano G. “reverse Genomics” Predicts Function of Human Conserved Noncoding Elements. Mol Biol Evol. 2016;33(5):1358–69.
17. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science (80-). 2012;337(6099):1190–5.
18. Bortoluzzi, Chiara; Megens, Hendrik-Jan; Bosse, Mirte; Derks, Martijn, Dibbits, Bert; Lamport, Kimberley; Weigend, Steffe; Groenen, Martien; Crooijmans R. Parallel genetic origin of foot feathering in birds. Mol Biol Evol. 2020;
19. Wang Z, Gerstein M, Snyder M. Nihms229948. 2010;10(1):57–63.
20. Park P. Applications of next-generation sequencing: ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669.
21. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet [Internet]. 2014;46(3):310–5.
22. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–94.
23. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. and adaptation. :1311–21.
24. Meredith RW, Zhang G, Gilbert MTP, Jarvis ED, Springer MS. Evidence for a single loss of mineralized teeth in the common avian ancestor. Science (80-). 2014;346(6215).
25. Lovell P V., Wirthlin M, Wilhelm L, Minx P, Lazar NH, Carbone L, et al. Conserved syntenic clusters of protein coding genes are missing in birds. Genome Biol. 2014;15(12):565.
26. Bornelöv S, Seroussi E, Yosefi S, Pendavis K, Burgess SC, Grabherr M, et al. Correspondence on Lovell et al.: Identification of chicken genes previously assumed to be evolutionarily lost. Genome Biol. 2017;18(1):1–4.
27. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
28. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Vol. 2, Current Protocols in Human Genetics. 2013.
29. Choi Y, Chan AP. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(16):2745–7.
30. Groß C, de Ridder D, Reinders M. Predicting variant deleteriousness in non-human species: Applying the CADD approach in mouse. BMC Bioinformatics. 2018;19(1):1–10.

31. Groß C, Derks M, Megens HJ, Bosse M, Groenen MAM, Reinders M, et al. PCADD: SNV prioritisation in Sus scrofa. Genet Sel Evol. 2020;52(1):1–15.

32. Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, et al. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. 2007;17(12):1797–808.

33. Chicken I, Sequencing G. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432(7018):695–716.

34. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved elements in the human genome. Science (80-). 2004;304(5675):1321–5.

35. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, et al. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet. 2006;38(2):223–7.

36. Casillas S, Barbadilla A, Bergman CM. Purifying selection maintains highly conserved noncoding sequences in Drosophila. Mol Biol Evol. 2007;24(10):2222–34.

37. Cooper GM, Shendure J. Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011;12(9):628–40.

38. Babarinde IA, Saitou N. Genomic Locations of Conserved Noncoding Sequences and Their Proximal Protein-Coding Genes in Mammalian Expression Dynamics. Mol Biol Evol. 2016;33(7):1807–17.

39. Polychronopoulos D, King JWD, Nash AJ, Tan G, Lenhard B. Conserved non-coding elements: Developmental gene regulation meets genome organization. Nucleic Acids Res. 2017;45(22):12611–24.

40. Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science (80-). 2014;346(6215).

41. Armstrong J, Hickey G, Diekhans M, Deran A, Fang Q, Xie D, et al. Progressive alignment with Cactus: a multiple-genome aligner for the thousand-genome era. BioRxiv. 2019

42. Zhang G. The bird's-eye view on chromosome evolution. Genome Biol. 2018;19(1):18–20.

43. Jarvis ED, Ye C, Liang S, Yan Z, Zepeda ML, Campos PF, et al. A Phylogeny of Modern Birds. Science (80-). 2014;346(6215):1126–38.

44. Steige KA, Laenen B, Reimegård J, Scofield DG, Slotte T. Genomic analysis reveals major determinants of cis-regulatory variation in Capsella grandiflora. Proc Natl Acad Sci U S A. 2017;114(5):1087–92.

45. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979;76(10):5269–73.

46. Watterson GA. 7to recombination between sites. Of course, this assumption is meaningful only if the gametes making up our population go through a diploid phase, as they do in Burrows and Cockerham's model to be described below. For 256. Theor Popul Biol. 1975;276(7):256–76.

47. Bortoluzzi C, Bosse M, Derks MFL, Crooijmans RPMA, Groenen MAM, Megens HJ. The type of

bottleneck matters: Insights into the deleterious variation landscape of small managed populations. Evol Appl. 2019;(September 2019):330–41.

48. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

49. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. Bioinformatics. 2015;31(12):2032–4.

50. Garrison E, Marth M. Haplotype-based variant detection from short-read sequencing Erik Garrison and Gabor Marth January 12, 2016 Abstract. 2016;1–20.

51. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. Cactus: Algorithms for genome multiple sequence alignment. Genome Res. 2011;21(9):1512–28.

52. Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: A hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics. 2013;29(10):1341–2.

53. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res [Internet]. 2005;15(8):1034–50.

54. Sadri J, Diallo AB, Blanchette M. Predicting site-specific human selective pressure using evolutionary signatures. Bioinformatics. 2011;27(13):266–74.

55. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191–8.

56. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. 2016;17(1):122.

57. Dalloul RA, Long JA, Zimin A V., Aslam L, Beal K, Blomberg LA, et al. Multi-platform next-generation sequencing of the domestic Turkey (Meleagris gallopavo): Genome assembly and analysis. PLoS Biol. 2010;8(9).

58. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, et al. The genome of a songbird. Nature. 2010;464(7289):757–62.

59. Alföldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature. 2011;477(7366):587–91.

60. Draper, N.R; Smith H. Applied regression analysis. John Wiley & Sons; 1998. Vol. 326.

61. Lenffer J. OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. Nucleic Acids Res. 2006;34(90001):D599–601.

62. Zhao H, Sun Z, Wang J, Huang H, Kocher J, Wang L. CrossMap : a versatile tool for coordinate conversion between genome assemblies. 2014;30(7):1006–7.

63. Truong C, Oudre L, Vayatis N. ruptures: change point detection in Python. 2018;1–5.

64. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 1998(67-).

65. Sawilowsky SS.New effect size rules of thumb. *Journal of Modern Applied Statistical Methods.* 2009; (26)2..