nature COMMUNICATIONS

ARTICLE

Received 15 Mar 2016 | Accepted 2 Sep 2016 | Published 6 Oct 2016

DOI: 10.1038/ncomms13107 OPEN

Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper

Wei Yin[1,*], Zong-ji Wang[2,3,4,5,*], Qi-ye Li[3,4,6], Jin-ming Lian[3], Yang Zhou[3], Bing-zheng Lu[7], Li-jun Jin[3], Peng-xin Qiu[7], Pei Zhang[3], Wen-bo Zhu[7], Bo Wen[8], Yi-jun Huang[7], Zhi-long Lin[8], Bi-tao Qiu[3,9], Xing-wen Su[7], Huan-ming Yang[8,10], Guo-jie Zhang[3,4,9], Guang-mei Yan[7] & Qi Zhou[2,11]

Snakes have numerous features distinctive from other tetrapods and a rich history of genome evolution that is still obscure. Here, we report the high-quality genome of the five-pacer viper, *Deinagkistrodon acutus*, and comparative analyses with other representative snake and lizard genomes. We map the evolutionary trajectories of transposable elements (TEs), developmental genes and sex chromosomes onto the snake phylogeny. TEs exhibit dynamic lineage-specific expansion, and many viper TEs show brain-specific gene expression along with their nearby genes. We detect signatures of adaptive evolution in olfactory, venom and thermal-sensing genes and also functional degeneration of genes associated with vision and hearing. Lineage-specific relaxation of functional constraints on respective *Hox* and *Tbx* limb-patterning genes supports fossil evidence for a successive loss of forelimbs then hindlimbs during snake evolution. Finally, we infer that the ZW sex chromosome pair had undergone at least three recombination suppression events in the ancestor of advanced snakes. These results altogether forge a framework for our deep understanding into snakes' history of molecular evolution.

[1] Department of Biochemistry, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510089, China. [2] Life Sciences Institute, The Key Laboratory of Conservation Biology for Endangered Wildlife of the Ministry of Education, Zhejiang University, Hangzhou 310058, China. [3] China National Genebank, BGI-Shenzhen, Shenzhen 518083, China. [4] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China. [5] School of Bioscience & Bioengineering, South China University of Technology, Guangzhou 510006, China. [6] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, Copenhagen K 1350, Denmark. [7] Department of Pharmacology, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510089, China. [8] BGI-Shenzhen, Shenzhen 518083, China. [9] Centre for Social Evolution, Department of Biology, University of Copenhagen, Universitetsparken 15, Copenhagen DK-2100, Denmark. [10] James D. Watson Institute of Genome Sciences, Hangzhou 310058, China. [11] Department of Integrative Biology, University of California, Berkeley, Berkeley, California 94720, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to G.-m.Y. (email: ygm@mail.sysu.edu.cn) or to Q.Z. (email: zhouqi1982@zju.edu.cn).

Snakes have undergone a massive adaptive radiation with ~3,400 extant species successfully inhabiting almost all continents except for the polar regions[1]. This process has culminated in 'advanced snakes' (Caenophidia, ~3,000 species), involved numerous evolutionary changes in body form, chemo and thermo-perception, venom and sexual reproductive systems, which altogether distinguish snakes from the majority of other squamates (lizards and worm lizards). Some of these dramatic changes can be tracked from fossils, which have established that the ancestor of snakes had already evolved an elongated body plan, probably as an adaptation to a burrowing and crawling lifestyle, but had lost only the forelimbs[2–4]. Extant boa and python species retain rudimentary hindlimbs, whereas advanced snakes have completely lost them. Limblessness, accompanied by degeneration in visual and auditory perception, has not compromised snakes' dominant role as top predators, largely due to the evolution of infrared sensing and/or venom, and the development of corresponding facial pit and fangs (specialized teeth for venom injection) independently in different lineages[5,6].

These extreme adaptations have sparked strong and standing interest into their genetic basis. Snakes are used as a model for studying various basic questions about the mechanisms of axial patterning and limb development[3,7,8], 'birth-and-death' of venom proteins[9–11] and sex chromosome evolution[12]. Cytogenetic findings in snakes first drove Ohno[13] to propose that sex chromosomes in vertebrates evolved from ancestral autosomes, such as those of insects[14] and plants[15]. Insights into these questions have been advanced recently by the application of next-generation sequencing. Analyses of python and king cobra genomes and transcriptomes have uncovered the metabolic gene repertoire involved in feeding, and inferred massive expansion and adaptive evolution of toxin families in elapids (an 'advanced' group)[10,16]. However, comparative studies of multiple snake genomes unraveling their evolutionary trajectories since the divergence from lizards are lacking, and so far only a few specific developmental 'toolbox' (for example, *Hox*[7,17,18] and Fgf signalling pathway[19]) genes have been studied between snakes and lizards. This deficiency hampers our comprehensive understanding into the molecular basis of stepwise or independent acquisition of snake-specific traits. We bridge this gap here by deep-sequencing the genomes and transcriptomes of the five-pacer viper, *Deinagkistrodon acutus* (Fig. 1a), a member of the Viperidae family. This pit viper is a paragon of infrared sensing, heteromorphic ZW sex chromosomes, and distinctive types of fangs and toxins (its common name exaggerates that victims can walk no more than five paces) from other venomous snake families[6,20].

Here we conduct comparative genomic analyses of the five-pacer viper with the available genomes of three species from other major snake families, that is, *Boa constrictor* (Boidae)[21], *Python bivittatus* (Pythonidae)[16], *Ophiophagus hannah* (Elapidae)[10] and several reptile outgroups. We show that all analysed snake genomes have a distinctive distribution of GC content compared with that of the green anole lizard, and that different snake species' genome architectures are shaped by lineage-specific expansion of respective TE families. We show evidence for adaptive evolution of olfactory receptor and venom genes, as well as relaxation of functional constraints on limb-patterning, visual and auditory genes among snake lineages. Altogether with the inferred recombination suppression events between the ZW sex chromosomes, we have reconstructed the major genomic changes during the snake evolution in this work.

Results

Evolution of snake genome architecture. We sequenced a male and a female five-pacer viper (*Deinagkistrodon acutus*) to high-coverage (♀ 238 fold, ♂114 fold, Supplementary Table 1), and estimated the genome size to be 1.43 Gb based on k-mer frequency distribution[22] (Supplementary Table 2 and Supplementary Fig. 1). Fewer than 10% of the reads, which have a low quality or are probably derived from repetitive regions, were excluded from the genome assembly (Supplementary Table 3). We generated a draft genome using only male reads for constructing the contigs, and female long-insert (2–40 kb) library reads for joining the contigs into scaffolds. The draft genome has an assembled size of 1.47 Gb, with a slightly better quality than the genome assembled using only female reads. The draft genome has high continuity (contig N50: 22.42 kb, scaffold N50: 2.12 Mb) and integrity (gap content 5.6%, Supplementary Table 4), and thus was chosen as the reference genome for further analyses. It includes a total of 21,194 predicted protein-coding genes, as estimated using known vertebrate protein sequences and transcriptome data generated in this study from eight tissues (Fig. 1a, Methods). For comparative analyses, we also annotated 17,392 protein-coding genes in the boa genome (the SGA version from[21]). 80.84% (17,134) of the viper genes show robust expression (normalized expression level RPKM > 1) in at least one tissue, comparable to 70.77% in king cobra (Supplementary Table 5). On the basis of 5,353 one-to-one orthologous gene groups of four snake species (five-pacer viper, boa[21], python[16] and king cobra[10]), the green anole lizard[23] and several other sequenced vertebrate genomes (Methods), we constructed a phylogenomic tree with high bootstrapping values at all nodes (Fig. 1b). We estimated that advanced snakes diverged from boa and python about 66.9 (47.2–84.4) million years ago (MYA), and five-pacer viper and king cobra diverged 44.9 (27.5–65.0) MYA assuming a molecular clock. These results are consistent with the oldest snake and viper fossils from 140.8 and 84.7 MYA, respectively[24].

The local GC content of snakes (boa and five-pacer viper) shows variation (GC isochores) similar to the genomes of turtles and crocodiles, and intermediate between mammals/birds and lizard (Fig. 1c, Supplementary Fig. 2), confirming the loss of such a genomic feature in the green anole lizard[23]. Cytogenetic studies showed that, like most other snakes, the five-pacer viper karyotype has $2n = 36$ chromosomes (16 macro and 20 micro-chromosomes)[25]. Previous work showed that there is extensive inter-chromosomal conservation between the rat snake and the butterfly lizard[26]. This information enables us to organize 56.50% of the viper scaffold sequences into linkage groups, based on their homology with sequences of known green anole lizard macro-chromosomes (Supplementary Table 6). As expected, autosomal sequences have the same read coverage in both sexes, whereas scaffolds inferred to be located on the viper Z chromosome (homologous to green anole lizard chr6) have coverage in the female that is half that in the male (Fig. 1c). Additionally, the frequency of heterozygous variants on the Z chromosome is much lower in the female than in the male (0.005 versus 0.08%, Wilcoxon signed rank test, P value < 2.2e-16,) due to the nearly hemizygous state of Z chromosome in female, while those of autosomes (~0.1%) are very similar between sexes. These results indicate that our assembly mostly assigns genes to the correct chromosome, which is further supported by comparison of 172 genes' locations with previous fluorescence *in situ* hybridization results (Supplementary Data 1)[26]. The pattern of heterozygous variants also suggests that the viper sex chromosomes are highly differentiated from each other (see below).

47.47% of the viper genome consists of transposable element sequences (TEs), a higher percentage than in any other snake so far analysed (33.95–39.59%), which cannot be explained solely by the higher assembly quality of the viper[10,16,21] (Supplementary Tables 7 and 8). The TEs in the viper genome

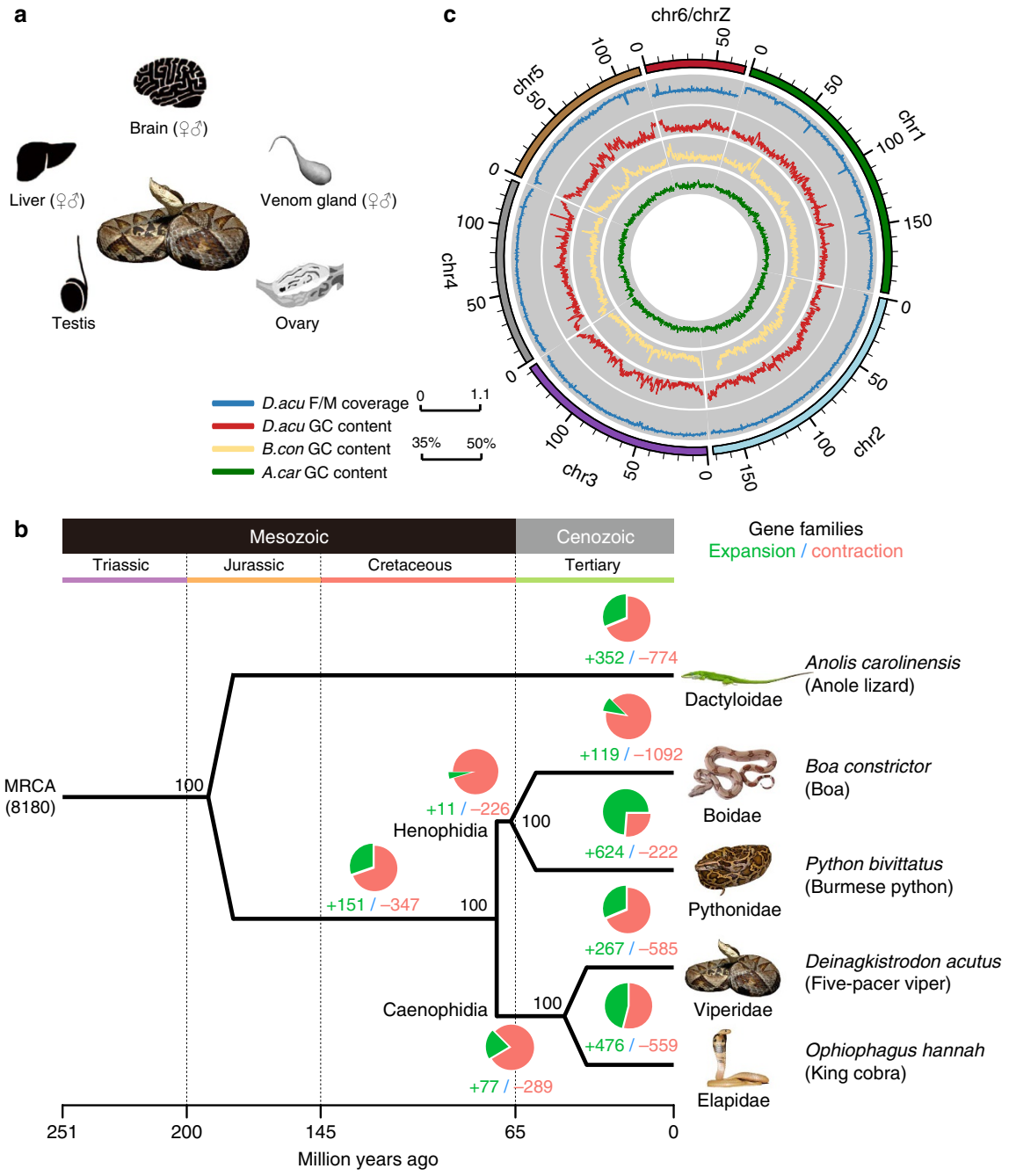

Figure 1 | The comparative genomic landscape of five-pacer viper. (**a**) *Deinagkistrodon acutus* (five-pacer viper) and eight adult tissues used in this study. The viper photo is contributed by Ren-jie Wang. (**b**) Circos plot showing the linkage group assignment using lizard chromosomes as reference (outmost circle), normalized female versus male mapped read coverage ratio (blue line) and GC-isochore structures of five-pacer viper (red), boa (yellow) and green anole lizard (green). Both snake genomes have a much higher variation of local GC content than that of green anole lizard. (**c**) Phylogenomic tree constructed using fourfold degenerate sites from 5,353 single-copy orthologous genes. We also showed bootstrapping percentages, the numbers of inferred gene family expansion (in green) and contraction (red), and corresponding phylogenetic terms at each node. MRCA: most recent common ancestor. Animal photos are contributed by Mike Graziano, Sid Ewing, Camilla Bjerke, Ren-jie Wang and Zill Niazi.

are mostly long interspersed elements (LINE, 13.84% of the genome) and DNA transposons (7.96%, Supplementary Table 7). Sequence divergence of individual families from inferred consensus sequences uncovered recent rampant activities in the viper lineage of LINEs (CR1), DNA transposons (hAT and TcMar) and retrotransposons (Gypsy and DIRS). In particular, there is an excess of low-divergence (<10% divergence level) CR1 and hAT elements in the viper genome only (Fig. 2a). We also inferred earlier propagation of TEs shared by viper and king cobra, which thus probably occurred in the ancestor of advanced snakes. Altogether, these derived insertions resulted in an at least three-fold difference in the CR1 and hAT content between viper and more basal-branching snakes such as the boa and python (Fig. 2b). Meanwhile, the boa and python have undergone independent expansion of L2 and CR1 repeats, so that their overall LINE content is at a similar level to that of the viper and cobra (Fig. 2a, Supplementary Table 7).

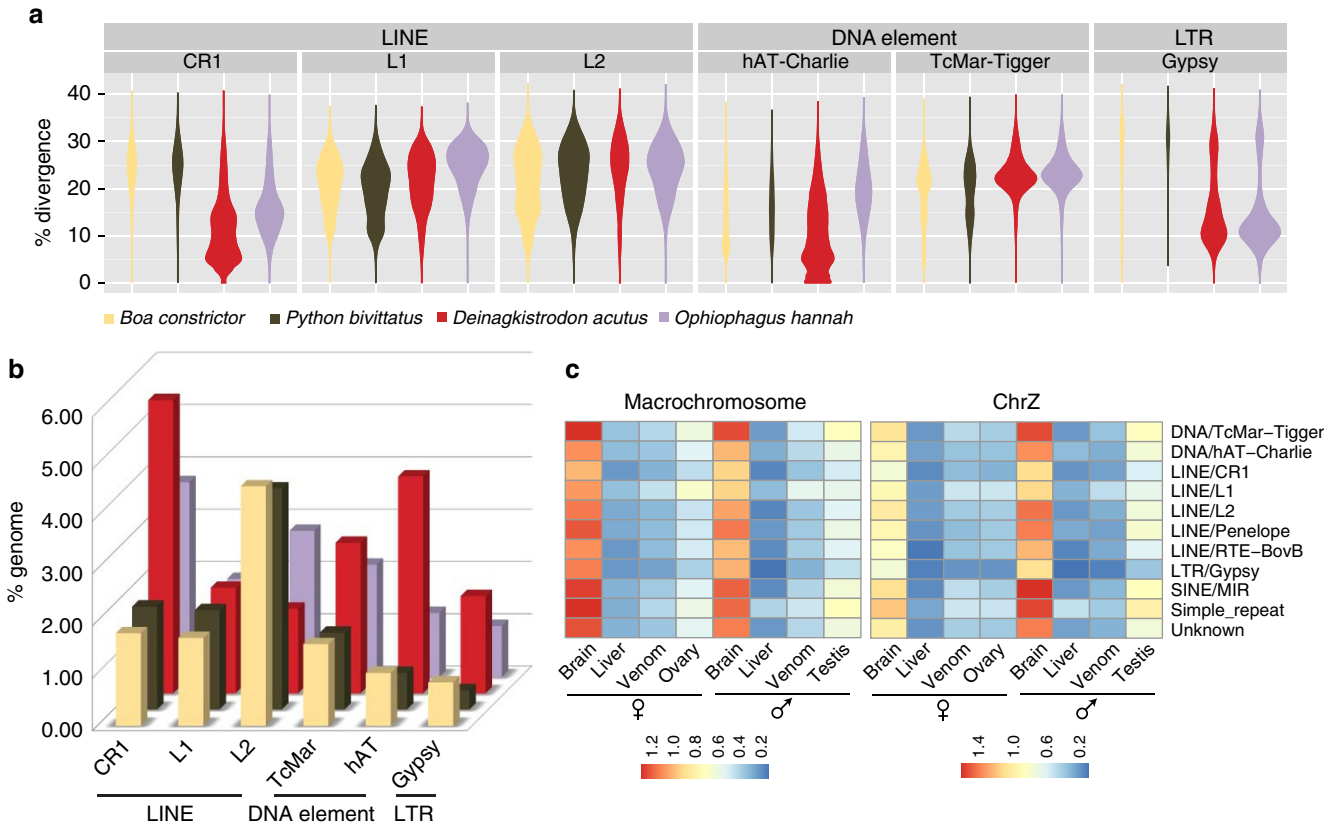

Figure 2 | Genomic and transcriptomic variation of snake transposable elements. (**a**) Violin plots showing each type of TE's frequency distribution of sequence divergence level from the inferred ancestral consensus sequences. Clustering of TEs with similar divergence levels, manifested as the 'bout' of the violin, corresponds to the burst of TE amplification. (**b**) Bar plots comparing the genome-wide TE content between four snake species. TE families were annotated combining information of sequence homology and *de novo* prediction. (**c**) TE's average normalized expression level (measured by RPKM) across different tissues in five-pace viper.

Most TEs are presumably silenced through epigenetic mechanisms to prevent their deleterious effects of transposition and mediation of genomic rearrangements. Indeed, very few TEs are transcribed in all of the tissues examined, except, unexpectedly, in the brain (Fig. 2c). This brain-specific expression prompted us to test whether some snake TE families might have been co-opted into brain gene regulatory networks. Focusing on highly expressed (RPKM > 5) TEs that are located within 5 kb flanking regions of genes, we found that these nearby genes also show a significantly higher expression in brain than in any other tissues (Wilcoxon test, P value < 1.1e-40, Supplementary Fig. 3). The expression levels of individual genes are strongly correlated (Spearman's test, P value < 1.35e-08) with those of nearby TEs. These genes are predominantly enriched (Fisher's Chi-square test, Q-value < 0.05, Supplementary Figs 4–9) in functional domains of 'biological process' compared with 'cellular component' and 'molecular function', and particularly enriched categories include environmental response ('response to organic substance', 'regulation of response to stimulus' and 'sensory perception of light stimulus') and brain signalling pathways ('neuropeptide signalling pathway', 'opioid receptor signalling pathway' and 'regulation of cell communication' and so on). Further experimental studies are required to elucidate how some of these TEs evolved to regulate gene expression in the brain; these results nevertheless highlight the evolutionary dynamics and potential functional contribution of TEs in shaping snake genome evolution.

Evolution of snake genes and gene families. To pinpoint the critical genetic changes underlying the phenotypic innovations of snakes, we next mapped protein coding genes' gain and loss (Fig. 1b) onto the phylogenetic tree. We also characterized signatures of lineage-specific adaptive or degenerative evolution (Fig. 3a) measured by their ratios (ω) of nonsynonymous versus synonymous substitution rates (Supplementary Data 2). We inferred a total of 1,725 gene family expansion and 3,320 contraction events, and identified 610 genes that appear to have undergone positive selection and 6,149 with relaxed selective constraints at different branches, using a likelihood model and conserved lineage-specific test[27]. Genes of either scenario were separated for analysis of their enriched gene ontology (GO) and mouse orthologs' mutant phenotype terms, assuming most of them have a similar function in snakes.

Significantly (Fisher's exact test, P value < 0.05) enriched mutant phenotype terms integrated with their branch information illuminated the molecular evolution history of snake-specific traits (Fig. 3a). For example, as adaptations to a fossorial lifestyle, the four-legged snake ancestor[28] had evolved an extreme elongated body plan without limbs, and also fused eyelids ('spectacles', presumably for protecting eyes against soil[29]). The latter is supported by the results for the positively selected gene *Ereg*[30] and the genes under relaxed selection *Cecr2* (ref. 31) and *Ext1* (ref. 32) at the snake ancestor branch (Supplementary Data 3), whose mouse mutant phenotype is shown as prematurely opened or absent eyelids. The limbless body plan has already driven many comparisons of expression domains and coding-sequences of the responsible *Hox* genes between snakes and other vertebrates[7,17]. We here refined the analyses to within snake lineages, focusing on sequence evolution of *Hox* and other genes

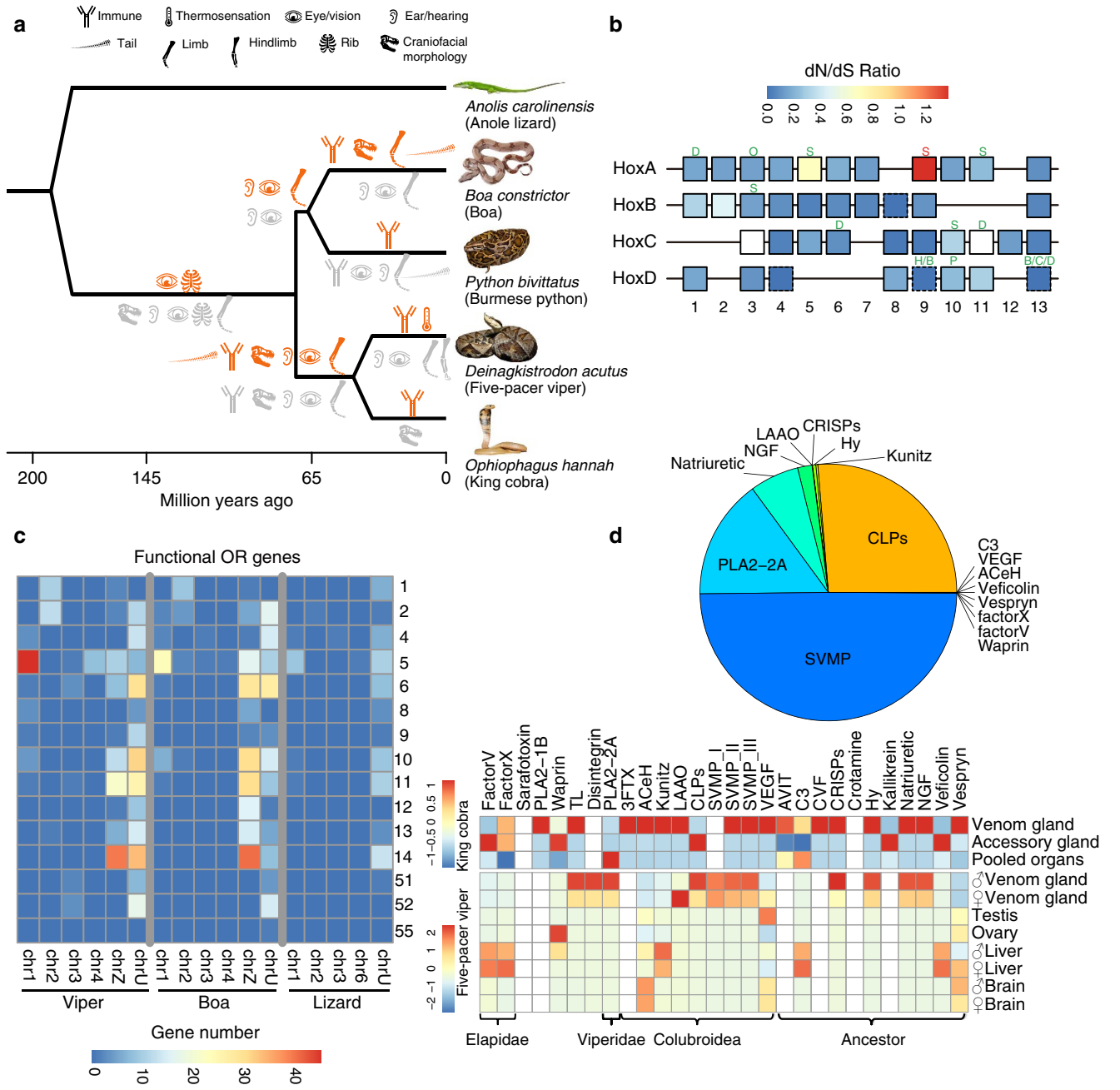

Figure 3 | Evolution of snake genes and gene families. (**a**) Phylogenetic distribution of mutant phenotypes (MP) of mouse orthologs of snakes. Each MP term is shown by an organ icon, and significantly enriched for snake genes undergoing positive selection (red) or relaxed selective constraints (grey) inferred by lineage-specific PAML analyses. (**b**) We show the four *Hox* gene clusters of snakes, with each box showing the ratio of nonsynonymous (dN) over synonymous substitution (dS) rate at the snake ancestor lineage. White boxes represent genes that haven't been calculated for their ratios due to the genome assembly issue in species other than five-pacer viper. Boxes with dotted line refer to genes with dS approaching 0, therefore the dN/dS ratio cannot be directly shown. Each cluster contains up to 13 *Hox* genes with some of them lost during evolution. We also marked certain *Hox* genes undergoing positive selection (in red) or relaxed selective constraints (in green) at a specific lineage above the box. Each lineage was denoted as: S: *Serpentes* (ancestor of all snakes), H: *Henophidia* (ancestor of boa and python), B: *Boa constrictor*, P: *Python bivittatus*, C: *Colubroidea* (ancestor of five-pacer viper and king cobra), D: *Deinagkistrodon acutus*, O: *Ophiophagus hannah*. (**c**) Comparing olfactory receptor (OR) gene repertoire between boa, viper and lizard. Each cell corresponds to a certain OR family (shown at the *y*-axis) gene number on a certain chromosome (*x*-axis). (**d**) Pie chart shows the composition of normalized venom gland transcripts of male five-pacer viper. The heatmap shows the normalized expression level (in RPKM) across different tissues of viper and king cobra. We grouped the venom genes by their time of origination, shown at the bottom *x*-axis.

involved in limb development and somitogenesis. We annotated the nearly complete sequences of 39 *Hox* genes organized in four clusters (*HoxA*–*HoxD*) of the five-pacer viper. Compared with the green anole lizard, the four studied snake species have *Hox* genes whose sizes are generally reduced, due to the specific accumulation of DNA transposons in the lizard's introns and intergenic regions (Supplementary Fig. 10). However, snakes have accumulated particularly higher proportions of simple tandem repeat and short interspersed element sequences within *Hox* clusters (Supplementary Fig. 11), either as a result of relaxed

selective constraints and/or evolution of novel regulatory elements. We identified 11 *Hox* genes as under relaxed selective constraint and one (*Hoxa9*) as under positive selection (Fig. 3b). Their combined information of gene function and affected snake lineage informed the stepwise evolution of snake body plan. In particular, *Hoxa5* (ref. 33), *Hoxa11* (ref. 34) and *Tbx5* (ref. 35), which specifically pattern the forelimbs in mouse, have been identified as genes under relaxed selective constraint in the common ancestor of all four snakes. Meanwhile, *Hoxc11* and *Tbx4* (ref. 36), which pattern the hindlimbs in the mouse, and many other limb-patterning genes (for example, *Gli3*, *Tbx18*, *Alx4*) were identified as genes under relaxed selective constraint that evolved independently on external snake branches (Fig. 3b, Supplementary Data 3). These results provide robust molecular evidence supporting the independent loss of hindlimbs after the complete loss of forelimbs in snake ancestors. In the snake ancestor branch, we also identified the genes under relaxed selective constraint *Hoxa11, Hoxc10* and *Lfng*, which are respectively associated with sacral formation[37], rib formation[8] and somitogenesis speed[38] in vertebrates. Their changed amino acids and the expression domains that have expanded in snakes relative to lizards[17,19] might have altogether contributed to the 'de-regionalization'[17] and elongation of the snake body plan. In several external branches, we identified *Hoxd13* independently as under relaxed selective constraint. Besides its critical roles in limb/digit patterning[39], *Hoxd13* is also associated with termination of the somitogenesis signal and is specifically silenced at the snake tail relative to the lizard tail[7]. This finding suggests that body elongation may have evolved more than once among snake lineages. Overall, limb/digit/tail development mutant phenotype terms are significantly enriched in genes under relaxed selective constraint at both ancestral and external snake branches (Fig. 3a), and we identified many such genes in different snake lineages for future targeted experimental studies (Supplementary Data 3 and 4).

Another important adaption to the snakes' ancestrally fossorial and later ground surface lifestyle is the shift of their dominant source of environmental sensing from visual/auditory to thermal/chemical cues. Unlike most other amniotes, extant snake species do not have external ears, and some basal species (for example, blindsnake) have completely lost their eyes. Consistently, we found mutant phenotype terms associated with hearing/ear and vision/eye phenotypes (for example, abnormal ear morphology, abnormal vision and abnormal cone electrophysiology) are enriched among genes under relaxed selection along all major branches of snakes starting from their common ancestor (Fig. 3a, Supplementary Fig. 12, Supplementary Data 3). Gene families that have contracted in the ancestor of the four studied snake species, and specifically in the viper, are also significantly enriched in GOs' of 'sensory perception of light stimulus (GO:0050953)' or 'phototransduction (GO:0007602)' (Fisher's Exact Test, Q-value $<$ 9.08e-4; Fig. 1c, Supplementary Data 5). In particular, only three (*RH1, LWS* and *SWS1*) out of 13 opsin genes' complete sequences can be identified in the viper genome, consistent with the results found in python and cobra[16]. By contrast, infrared receptor gene *TRPA1* (ref. 5) and ubiquitous taste-signalling gene *TRPM5* (ref. 40) have respectively undergone adaptive evolution in five-pacer viper and the ancestor of boa and python. Gene families annotated with the GO term 'olfactory receptor (OR) activity' have a significant (Fisher's Exact Test, Q-value $<$ 1.63e-4) expansion in all snake species studied and at some of their ancestral nodes, except for the king cobra (Supplementary Figs 13–26). In the boa and viper, whose genome sequences have much better quality than the other two snake genomes, we respectively annotated 369 and 412 putatively functional OR genes, based on homology search and the characteristic 7-TM (transmembrane) structure (Methods). Both terrestrial species have an OR repertoire predominantly comprised of class II OR families (OR1–14, presumably for binding airborne molecules, Fig. 3c), and their numbers are much higher than the reported numbers in other squamate genomes[41]. Some (ranging from 18 to 24) class I (OR51–56, for water-borne molecules) genes have also been found in the two species, indicating this OR class is not unique to python as previously suggested[41]. Compared with the green anole lizard, the boa and viper exhibit a significant size expansion of OR family 5, 11 and 14 (Fisher's exact test $P < 0.05$), and also a bias towards being located on the Z chromosome (Fig. 3c), leading to higher expression of many OR genes in males than in females (see below). In particular, OR5 in the viper probably has experienced additional expansion events and become the most abundant (with 71 members) family in the genome. Intriguingly, this family is specifically enriched in birds of prey[42] relative to other birds, and in non-frugivorous bats versus frugivorous bats[43]. Therefore, its expansion in the five-pacer viper could have been positively selected for a more efficient detection of prey.

Besides acute environmental sensing, specialized fangs[6] and venoms[11] (for example, hemotoxins of viper or neurotoxins of elapid) arm the venomous snakes ($\sim$650 species) to immediately immobilize much larger prey for prolonged ingestion, which probably comprised one of the most critical factors that led to the advanced snakes' species radiation. It has been proposed that the tremendous venom diversity probably reflects snakes' local adaption to prey[44] and was generated by changes in the expression of pre-existing or duplicated genes[11,45]. Indeed, we found that the five-pacer viper's venom gland gene repertoire has a very different composition compared with other viper[46] or elapid species[10] (Fig. 3d). We have annotated a total of 35 venom genes or gene families using all the known snake venom proteins as the query. Certain gene families, including snake venom metalloproteinases (SVMP), C-type lectin-like proteins (CLPs), thrombin-like snake venom serine proteinases (TL), Kunitz and disintegrins, have more genomic copies in the five-pacer viper than other studied snakes or the green anole lizard (Supplementary Table 9), whereas characteristic elapid venom genes such as three-finger toxins (3FTx) are absent from the viper genome. Most venom proteins of both the viper and king cobra have expression restricted to venom or accessary glands, and for both species this is particularly seen for those genes that originated in the ancestor of snakes or of advanced snakes (Fig. 3d). However, elapid- and viper-specific venom genes, that is, those that originated more recently, are usually expressed in the liver of the other species. Such cases include FactorV, FactorX of king cobra, which are expressed in the liver of five-pacer viper, and PLA2-2A of viper (Fig. 3d), which is expressed in the pooled organ of king cobra. This expression pattern suggests that these venom genes may have originated from metabolic proteins and undergone neo-/sub-functionalization, with altered expression.

Evolution of snake sex chromosomes. Different snake species exhibit a continuum of sex chromosome differentiation. Pythons and boas possess homomorphic sex chromosomes, which is assumed to be the ancestral state; the lack of differentiation between the W and Z chromosomes in these species suggests that most regions of this chromosome pair recombine like the autosomes[47]. Advanced snakes usually have heteromorphic sex chromosomes that have undergone additional recombination suppression[47,48]. We found that the five-pacer viper probably has suppressed recombination throughout almost the entire sex chromosome pair, as the read coverage in the female that we sequenced is half that in the male (figs 1c and 4). By contrast,

the boa's homologous chromosomal regions show a read coverage pattern that does not differ from that of autosomes and between sexes (Fig. 4). Assuming that these two species share the same ancestral snake sex-determining region, this lack of sex-differentiated region shown in boa suggests that that region is not included in our current chromosomal assembly.

In plants, birds and mammals, it has been found that recombination suppression probably occurred by a succession of events. This stepwise recombination loss has led to the punctuated accumulation of excessive neutral or deleterious mutations on the Y or W chromosome by genetic drift, and produced a gradient of sequence divergence levels over time, which are termed 'evolutionary strata'[49–51]. Advanced snakes have been suggested to have at least two strata[12]. One goal of our much more continuous genome assembly of the five-pacer viper compared with those of any other studied advanced snakes[10,12]

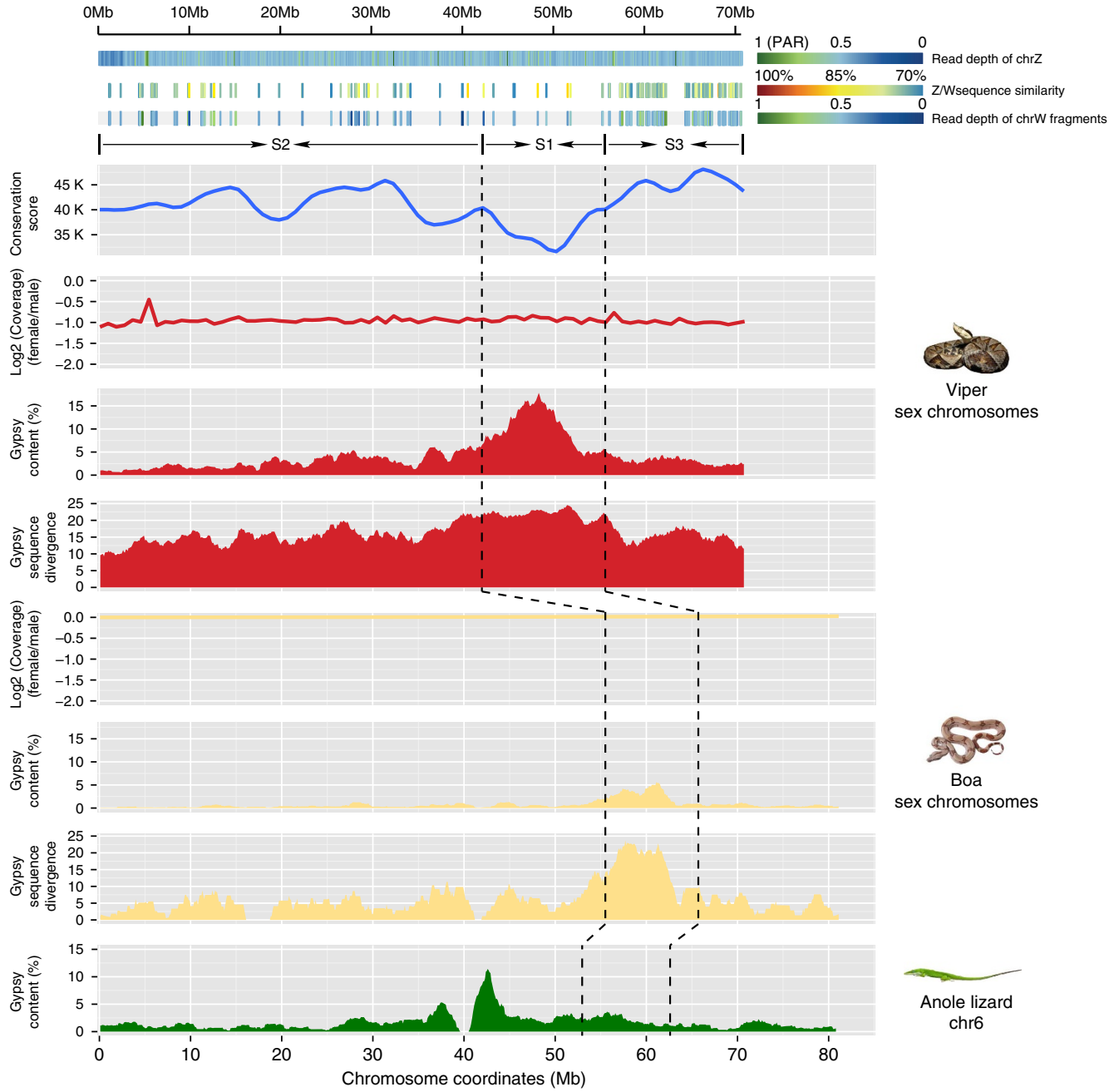

Figure 4 | Snake sex chromosomes have at least three evolution strata. The three tracks in the top panel shows female read depths along the Z chromosome relative to the median depth value of autosomes, Z/W pairwise sequence divergence within intergenic regions, and female read depths of W-linked sequence fragments relative to the median depth value of autosomes. Depths close to 1 suggest that the region is a recombining pseudoautosomal region (PAR), whereas depths of 0.5 are expected in a highly differentiated fully sex-linked region where females are hemizygous. The identifiable W-linked fragments are much denser at the region 56–70 Mb, probably because this region (denoted as stratum 3, S3) has suppressed recombination most recently. S2 and S1 were identified and demarcated by characterizing the sequence conservation level (measured by LASTZ alignment score, blue line) between the chrZs of boa and viper. At the oldest stratum S1 where recombination has been suppressed for the longest time, there is an enrichment of repetitive elements on the affected Z-linked region (Gypsy track in red, 100 kb non-overlapping sliding window). And these Z-linked TEs A similar pattern was found in homologous recombining region of boa, but not in lizard.

(Supplementary Table 4) was to reconstruct a fine history of snake sex chromosome evolution. We assembled 77 Mb Z-linked and 33 Mb W-linked scaffolds (Methods). The reduction of female read coverage along the Z chromosome suggests that there is substantial divergence between Z- and W- linked sequences; this divergence would enable the separate assembly of two chromosomes' scaffolds. Mapping the male reads confirmed that the inferred W-linked scaffold sequences are only present in the female (Supplementary Fig. 27). The W-linked scaffolds' density and pairwise sequence divergence values within putative neutral regions along the Z chromosome indicate at least two 'evolutionary strata', with the older stratum extending 0–56 Mb, and the younger one extending 56–70 Mb. The boundary at 56 Mb region can also be confirmed by analyses of repetitive elements on the Z chromosome (see below). Consistently, identifiable W-linked fragments are found at the highest density per megabase in the 56–70 Mb region (Fig. 4), suggesting that recombination in this region have been suppressed more recently. The older stratum includes much fewer identifiable fragments that can resolve the actual times of recombination suppression events. To study this region further, we inspected the homologous Z-linked region, whose recombination has also been reduced, albeit to a much smaller degree than that of the W chromosome, after the complete suppression of recombination between Z and W in females. In addition, Z chromosome transmission is biased in males. As males usually have a higher mutation rate than females, due to many more rounds of DNA replication during spermatogenesis than during oogenesis ('male-driven evolution')[52], Z-linked regions are expected to have a generally higher mutation rate than any other regions in the genome. This male-driven evolution effect has been demonstrated in other snake species[12] and also been validated for the snakes inspected in this study (Supplementary Fig. 28). As a result, we expected that regions in older strata should be more diverged from their boa autosome-like homologues than those in the younger strata. This expectation enabled us to identify another stratum (0–42 Mb, stratum 2, S2 in Fig. 4) and demarcate the oldest one (42–56 Mb, S1), by estimating the sequence conservation level (measured by LASTZ alignment score, blue line) between the Z chromosomes of boa and viper. The Z-linked region in the inferred oldest stratum S1 exhibits the highest sequence divergence with the homologous W-linked region and also the highest proportion of repetitive elements (CR1, Gypsy and L1 elements; Fig. 4 shows the example of Gypsy; other repeats are shown in Supplementary Fig. 29). This enrichment of repeats can be explained by the effect of genetic drift[53], which has been acting on the Z-linked S1 longer than any other Z-linked regions since the S1 reduced recombination rate in females. As a result, the accumulated repeats of S1 also tend to have a higher divergence level from the inferred ancestral consensus sequences compared with nearby strata (Fig. 4). Unexpectedly, a similar enrichment was found in the homologous region of S1 in boa, despite it being a recombining region and exhibiting the same coverage depth between sexes (Fig. 4, Supplementary Fig. 29). This finding indicates that the pattern is partially contributed by the ancestral repeats that had already accumulated on the proto-sex chromosomes of snake species. Since our current viper sex chromosomal sequences used the green anole lizard chromosome 6 as a reference, rearrangements within this chromosome make it impossible to test whether S2 encompasses more than one stratum.

We dated the three resolved strata by constructing phylogenetic trees with homologous Z- and W-linked gene sequences of multiple snake species. Combining the published CDS sequences of pygmy rattlesnake (Viperidae family species) and garter snake (Colubridae family species)[12], we found 31 homologous Z-W gene pairs, representing the three strata. All of their sequences clustered by chromosome (that is, the Z-linked sequences from all the species cluster altogether, separately from the W-linked ones) rather than by species (Supplementary Figs 30–32). This clustering pattern indicates that all three strata formed before the divergence of the advanced snakes and after their divergence from boa and python, that is, about 66.9 million years ago (Fig. 1c).

We found robust evidence of functional degeneration on the W chromosome. It is more susceptible to the invasion of TEs; the assembled sequences' overall repeat content is at least 1.5 fold higher than that of the Z chromosome, especially in the LINE L1 (2.9 fold) and LTR Gypsy families (4.3 fold) (Supplementary Table 10 and Supplementary Fig. 33). Of 1,135 Z-linked genes, we were only able to identify 137 W-linked homologues. Among these, 62 (45.26%) have probably become pseudogenes due to nonsense mutations (Supplementary Table 11). W-linked loci generally are transcribed at a significantly lower level (Wilcoxon test, $P < 0.0005$), with pseudogenes transcribed at an even lower level relative to their autosomal or Z-linked homologous loci regardless of the tissue type (Supplementary Figs 34 and 35). Given such a chromosome-wide gene loss, as in other snakes[12] and the majority of species with ZW sex chromosomes[54], the five-pacer viper shows a generally male-biased gene expression throughout the Z-chromosome and probably has not evolved global dosage compensation (Supplementary Fig. 36).

Discussion

The elongated body plan has evolved repeatedly in not only snakes but also other tetrapods (for example, worm lizard and caecilians), in which limb reduction/loss seems to have always been accompanied by body elongation. For example, several limb-patterning *Hox* genes (*Hoxc10*, *Hoxd13*) identified as under relaxed selective constraints also have been characterized by previous work with a changed expression domain along the snake body axis[7,17]. Another gene under relaxed selective constraint, *Hoxa5*, which is involved in the forelimb patterning[33], also participates in lung morphogenesis[55]. *Hoxa5* might have been involved in the elimination of one of the snake lungs during evolution. Therefore, the newly identified genes under positive selection or under relaxed selective constraint throughout the snake phylogeny in this work (Supplementary Data 2–4) can provide informative clues for future experimental work to use the snake as an emerging '*evo-devo*' model[56] to understand the genomic architecture of the developmental regulatory networks of organogenesis, or the crosstalk between these networks.

Like many of its reptile relatives, the snake ancestor is very likely to have determined sex by temperature and to have lacked sex chromosomes. Extant species boa can still undergo occasional parthenogenesis and is able to produce viable WW offspring[57], consistent with it having one of the most primitive vertebrate sex chromosome pairs reported to date. In the ancestor of advanced snakes, we inferred that there at least three recombination suppression events occurred between Z and W, leading to the generally degenerated W chromosome that we have observed in the five-pacer viper. How snakes determine sex genetically is an intriguing question to study in the future.

Methods

Genome sequencing and assembly. All animal procedures were carried out with the approval of China National Genebank animal ethics committee. We extracted genomic DNAs from blood of a male and a female five-pacer viper separately. A total of 13 libraries with insert sizes ranging from 250 bp to 40 kb were constructed using female DNA, and three libraries with insert sizes from 250 to 800 bp were constructed using male DNA. We performed paired-end sequencing (HiSeq 2000 platform) following the manufacturer's protocol, and produced 528 Gb raw data (357 Gb for female and 171 Gb for male). We estimated the

genome size based on the K-mer distribution. A K-mer refers to an artificial sequence division of K nucleotides iteratively from sequencing reads. The genome size can then be estimated through the equation G = K_num/Peak_depth, where the K_num is the total number of K-mer, and Peak_depth is the expected value of K-mer depth[58]. We found a single main peak in the male K-mer (K = 17) frequency distribution and an additional minor peak in the female data, the latter of which probably results from the divergence between W and Z chromosomes (Supplementary Fig. 1). Based on the distribution, we estimated that the genome size of this species is about 1.43 Gb (Supplementary Table 2), comparable to that of other snakes (1.44 and 1.66 Gb for Burmese python and King cobra[10,16], respectively).

After filtering out low-quality and duplicated reads, we performed additional filtering using the following criteria: we excluded the reads from short-insert libraries (250, 500 and 800 bp) with 'N's over 10% of the length or having more than 40 bases with the quality lower than 7, and the reads from large-insert libraries (2 to 40 kb) with 'N's over 20% of the length or having more than 30 bases with the quality lower than 7. Finally, 109.20 Gb (73X coverage) male reads and 148.49 Gb (99X coverage) female reads were retained for genome assembly (Supplementary Table 1) using SOAPdenovo[59] (http://soap.genomics.org.cn). To assemble the female and male genomes, reads from small-insert libraries of the female and male individual were used for contig construction separately. Then read-pairs from small- and large-insert libraries were utilized to join the contigs into scaffolds. We also used female long-insert libraries to join the male contigs into the longer scaffolds. At last, small-insert libraries of female and male individuals were used for gap closure for their respective genomes. The final assemblies of female and male have a scaffold N50 length of 2.0 and 2.1 Mb respectively, and the gap content of the two genomes are both less than 6% (♀ 5.29%, ♂ 5.61%) (Supplementary Table 4).

To access the assembly quality, reads from small-insert libraries that passed our filtering criteria were aligned onto the two assemblies using BWA[60] (Version: 0.5.9-r16) allowing 8 mismatches and 1 indel per read. A total of ~97% reads can be mapped back to the draft genome (Supplementary Table 3), spanning 98% of the assembled regions excluding gaps (Supplementary Table 12), and most genomic bases were covered by about 80X reads (Supplementary Fig. 37). Thus, we conclude that we have assembled most part of the five-pacer viper genome. To further test for potential mis-joining of the contigs into scaffolds, we analysed the paired-end information and found that 57% of the paired-end reads can be aligned uniquely with the expected orientation and distance. This proportion of the long insert library is significantly lower than that from the short insert libraries due to a circulization step during the library construction. When such paired-ends were excluded, the proportion increased to 94.98% (Supplementary Table 3). Overall, these tests suggested that the contigs and scaffolds are consistent with the extremely high density of paired-end reads, which in turn indicated the high-quality of the assembly.

Previous cytogenetic studies showed that snake genomes show extensive inter-chromosomal conservation with lizard[26,47]. Thus, we used the chromosomal information from green anole lizard[23] as a proxy to assign the snake scaffolds. We first constructed their orthologous relationship combining information of synteny and reciprocal best BLAST hits. Then gene coordinates and strandedness from the consensus chromosome were used to place and orient the snake scaffolds. Furthermore, we linked scaffolds into chromosomes with 600 'N's separating the adjacent scaffolds. In total, 625 five-pacer viper scaffolds comprising 832 Mb (56.50% scaffolds in length) were anchored to 5 autosomes and Z chromosome (Supplementary Table 6).

Repeat and gene annotation. We identified the repetitive elements in the genome combining both homology-based and *de novo* predictions. We utilized the 'Tetrapoda' repeat consensus library in Repbase[61] for RepeatMasker (http://www.repeatmasker.org) to annotate all the known repetitive elements in the five-pacer viper genome. To maximize the identification and classification of repeat elements, we further used RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) to construct the consensus repeat sequence libraries of the green anole lizard, boa and five-pacer viper, then used them as a query to identify repetitive elements using RepeatMasker. Finally, we retrieved a non-redundant annotation for each species after combining all the annotation results using libraries of 'Tetrapoda', 'green anole lizard', 'boa' and 'five-pacer viper'. For the purpose of comparison, we ran the same pipeline and parameters in all the snake and lizard genomes as shown in Supplementary Table 7. To provide a baseline estimate for the sequence divergence of TEs from the snake ancestral status, we first merged the genomes from boa and five-pacer viper, and constructed the putative ancestral consensus sequences using RepeatModeler. Then TE sequences of each snake species were aligned to the consensus sequence to estimate their divergence level using RepeatMasker.

For gene annotation, we combined resources of sequence homology, *de novo* prediction and transcriptome to build consensus gene models of the reference genome. Protein sequences of green lizard, chicken and human were aligned to the reference assembly using TBLASTN (E-value < = 1E-5)[62]. Then the candidate gene regions were refined by GeneWise[63] for more accurate splicing sites and gene models. We randomly selected 1,000 homology-based genes to train Augustus[64] for *de novo* prediction on the pre-masked genome sequences. We mapped RNA-seq reads of 13 samples to the genome using TopHat (v1.3.1)[65] and then assembled the transcripts by Cufflinks (v1.3.0) (http://cufflinks.cbcb.umd.edu/). Transcripts from different samples were merged by Cuffmerge. Finally, gene models from these three methods were combined into a non-redundant gene set.

We finally obtained 21,194 protein-coding genes with intact open reading frames (ORFs) (Supplementary Table 13). The gene models (measured by gene length, mRNA length, exon number and exon length) are comparable to those of other vertebrates and are well supported by the RNA-Seq data (Supplementary Fig. 38 and Supplementary Table 5). To annotate the gene names for each predicted protein-coding locus, we first mapped all the 21,194 genes to a manually collected Ensembl gene library, which consists of all proteins from *Anolis carolinensis, Gallus gallus, Homo sapiens, Xenopus tropicalis* and *Danio rerio*. Then the best hit of each snake gene was retained based on its BLAST alignment score, and the gene name of this best hit gene was assigned to the query snake gene. Most of the predicted genes can be found for their orthologous genes in the library at a threshold of 80% alignment rate (the aligned length divided by the original protein length), suggesting our annotation has a high-quality (Supplementary Table 14).

RNA-seq and gene expression analyses. Total RNAs were isolated from four types of tissues collected from both sexes, including brain, liver, venom gland and gonad (Supplementary Table 15). RNA sequencing libraries were constructed using the Illumina mRNA-Seq Prep Kit. Briefly, oligo (dT) magnetic beads were used to purify poly-A containing mRNA molecules. The mRNAs were further fragmented and randomly primed during the first strand synthesis by reverse transcription. This procedure was followed by a second-strand synthesis with DNA polymerase I to create double-stranded cDNA fragments. The cDNAs were subjected to end-repairing by Klenow and T4 DNA polymerases and A-tailed by Klenow lacking exonuclease activity. The fragments were ligated to Illumina Paired-End Sequencing adaptors, size selected by gel electrophoresis and then PCR amplified to complete the library preparation. The paired-end libraries were sequenced using Illumina HiSeq 2000 (90/100 bp at each end).

We used TopHat (v1.3.1) for aligning the RNA-seq reads and predicting the splicing junctions with the following parameters: -I/--max-intron-length: 10000, --segment-length: 25, --library-type: fr-firststrand, --mate-std-dev 10, -r/--mate-inner-dist: 20. Gene expression was measured by reads per kilobase of gene per million mapped reads (RPKM). To minimize the influence of different samples, RPKMs were adjusted by a scaling method based on TMM (trimmed mean of M values; M values mean the log expression ratios)[66] which assumes that the majority of genes are common to all samples and should not be differentially expressed.

Evolution analyses. A phylogenetic tree of the five-pacer viper and the other sequenced genomes (*Xenopus tropicalis, Homo sapiens, Mus musculus, Gallus gallus, Chelonia mydas, Alligator mississippiensis, Anolis carolinensis, Boa constrictor, Python bivittatus* and *Ophiophagus hannah*) was constructed using the 5,353 orthologous single-copy genes. Treebest (http://treesoft.sourceforge.net/treebest.shtml) was used to construct the phylogenetic tree. To estimate the divergence times between species, for each species, 4-fold degenerate sites were extracted from each orthologous family and concatenated to one sequence for each species. The MCMCtree program implemented in the Phylogenetic Analysis by Maximum Likelihood (PAML)[67] package was used to estimate the species divergence time. Calibration time was obtained from the TimeTree database (http://www.timetree.org/). Three calibration points were applied in this study as normal priors to constrain the age of the nodes described below. 61.5–100.5 MA for the most recent common ancestor (TMRCA) of human-mouse; 259.7–299.8 MA for TMRCA of Crocodylidae and Lepidosauria; 235–250.4 MA for TMRCA of Aves and Crocodylidae[68].

To examine the evolution of gene families in Squamate reptiles, genes from four snakes (*Boa constrictor, Python bivittatus, Deinagkistrodon acutus, Ophiophagus hannah*) and green anole lizard were clusterred into gene families by Treefam (min_weight = 10, min_density = 0.34 and max_size = 500)[69]. The family expansion or contraction analysis was performed by CAFE[70]. In CAFE, a random birth-and-death model was proposed to study gene gain and loss in gene families across a user-specified phylogenetic tree. A global parameter λ (lambda), which described both gene birth (λ) and death ($\mu = -\lambda$) rate across all branches in the tree for all gene families was estimated using maximum likelihood method. A conditional *P* value was calculated for each gene family, and the families with conditional *P* values lower than 0.05 were considered to have a significantly accelerated rate of expansion and contraction.

For the PAML analyses, we first assigned orthologous relationships for 12,657 gene groups among all Squamata and outgroup (turtle) using the reciprocal best blast hit algorithm and syntenic information. We used PRANK[71] to align the orthologous gene sequences, which takes phylogenetic information into account when placing a gap into the alignment. We filtered the PRANK alignments by gblocks[72] and excluded genes with high proportion of low complexity or repetitive sequences to avoid alignment errors. To identify the genes that evolve under positive selection (PSGs), we performed likelihood ratio test (LRT) using the branch model by PAML[67]. We first performed a LRT of the two-ratio model, which calculates the dN/dS ratio for the lineage of interest and the background lineage, against the one-ratio model assuming a uniform dN/dS ratio across all branches, so that to determine whether the focal lineage is evolving significantly faster (*P* value < 0.05). To differentiate between episodes of positive selection and

relaxation of purifying selection (RSGs), we performed a LRT of two-ratios model against the model that fixed the focal lineage's dN/dS ratio to be 1 (P value <0.05) and also required PSGs with the free-ratio model dN/dS >1 at the focal lineage. For the identified RSGs and PSGs, we used their mouse orthologs' mutant phenotype information[73] and performed enrichment analyses using MamPhEA[74]. Then we grouped the enriched MP terms by different tissue types.

Olfactory receptor (OR), *Hox* and venom gene annotation. To identify the nearly complete functional gene repertoire of OR, *Hox* and venom toxin genes in the investigated species, we first collected known amino acid sequences of 458 intact OR genes from three species (green anole lizard, chicken and zebra finch)[75], all annotated *Hox* genes from *Mus musculus* and *HoxC3* from *Xenopus tropicalis*, and obtained the query sequences of a total of 35 venom gene families[76] from UniProt (http://www.uniprot.org/) and NCBI (http://www.ncbi.nlm.nih.gov/). These 35 venom gene families represent the vast majority of known snake venoms. Then we performed a TBlastN[62] search with the cutoff E-value of 1E-5 against the genomic data using these query sequences. Aligned sequence fragments were combined into one predicted gene using perl scripts if they belonged to the same query protein. Then each candidate gene region was extended for 2 kb from both ends to predict its open reading frame by GeneWise[63]. Obtained sequences were verified as corresponding genes by BlastP searches against NCBI nonredundant (nr) database. Redundant annotations within overlapped genomic regions were removed.

For the OR gene prediction, these candidates were classified into functional genes and nonfunctional pseudogenes. If a sequence contained any disruptive frame-shift mutations and/or premature stop codons, it was annotated as a pseudogene. The remaining genes were examined using TMHMM2.0 (ref. 77). Those OR genes containing more at least 6 transmembrane (TM) structures were considered as intact candidates and the rest were also considered as pseudogenes. Finally, each OR sequence identified was searched against the Human Olfactory Data Explorer (the HORDE) database (http://genome.weizmann.ac.il/horde/) using the FASTA (ftp://ftp.virginia.edu/pub/fasta), and classified into the different families according to their best-aligned human OR sequence. For the venom toxin genes, we only kept these genes with RPKM higher than one in the five-pacer viper and king cobra venom gland tissue as final toxin gene set.

Identification and analyses of sex-linked genes. To identify the Z-linked scaffolds in the male assembly, we aligned the female and male reads to the male genome separately with BWA[60] allowing two mismatches and one indel. Scaffolds with less than 80% alignment coverage (excluding gaps) or shorter than 500 bp in length were excluded. Then single-base depths were calculated using SAMtools[78], with which we calculated the coverage and mean depth for each scaffold. The expected male versus female (M:F) scaled ratio of a Z-linked scaffold is equal to 2, and we defined a Z-linked scaffold with the variation of an observed scaled ratio to be less than 20% (that is, 1.6–2.4). With this criteria, we identified 139 Z-lined scaffolds, representing 76.93 Mb with a scaffold N50 of 962 kb (Supplementary Table 16). These Z-linked scaffolds were organized into pseudo-chromosome sequence based on their homology with green anole lizard. Another characteristic pattern of the Z-linked scaffolds is that there should be more heterozygous SNPs in the male individual than in the female individual resulted from their hemizygous state in female. We used SAMtools[78] for SNP/indel calling. SNPs and indels whose read depths were too low (<10) or too high (>120), or qualities lower than 100 were excluded. As expected, the frequency of heterozygous sites of Z chromosome of the female individual is much lower than that of the male individual (0.005 versus 0.08%), while the heterozygous rate of autosomes are similar in both sex ($\sim$0.1%) (Supplementary Table 17). To identify the W-linked scaffolds, we used the similar strategy as the Z-linked scaffold detection to obtain the coverage and mean depth of each scaffold. Then we identified those scaffolds covered by female reads over 80% of the length, and by male reads with less than 20% of the length. With this method, we identified 33 Mb W-linked scaffolds with a scaffold N50 of 48 kb (Supplementary Table 18).

We used the protein sequences of Z/W gametologs from garter snake and pygmy rattle snake[12] as queries and aligned them to the genomes of boa (the SGA assembly, http://gigadb.org/dataset/100060), five-pacer viper and king cobra with BLAST[62]. The best aligned (cutoff: identity $>=70\%$, coverage $>=50\%$) region with extended flanking sequences of 5 kb at both ends was then used to determine whether it contains an intact ORF by GeneWise[63] ($-$tfor $-$genesf $-$gff $-$sum). We annotated the ORF as disrupted when GeneWise reported at least one premature stop codon or frame-shift mutation. CDS sequences of single-copy genes' Z/W gametologs were aligned by MUSCLE[79] and the resulting alignments were cleaned by gblocks[72] ($-$b4$=$5, $-$t$=$c, $-$e$=-$gb). Only alignments longer than 300 bp were used for constructing maximum likelihood trees by RAxML[80] to infer whether their residing evolutionary stratum is shared among species or specific to lineages.

Data availability. All the genomic reads generated in this study have been deposited on NCBI Short Reads Archive under the BioProject Accession Number PRJNA314443, and all the RNA-seq reads have been deposited under the BioProject Accession Number PRJNA314559. The genome assembly and annotation produced in this study have been deposited in the GigaScience Database http://dx.doi.org/10.5524/100196.

References

1. Greene, H. W. *Snakes: the evolution of mystery in nature* (University of California Press, 1997).
2. Tchernov, E., Rieppel, O., Zaher, H., Polcyn, M. J. & Jacobs, L. L. A fossil snake with limbs. *Science* **287,** 2010–2012 (2000).
3. Cohn, M. J. & Tickle, C. Developmental basis of limblessness and axial patterning in snakes. *Nature* **399,** 474–479 (1999).
4. Apesteguia, S. & Zaher, H. A Cretaceous terrestrial snake with robust hindlimbs and a sacrum. *Nature* **440,** 1037–1040 (2006).
5. Gracheva, E. O. *et al.* Molecular basis of infrared detection by snakes. *Nature* **464,** 1006–1011 (2010).
6. Vonk, F. J. *et al.* Evolutionary origin and development of snake fangs. *Nature* **454,** 630–633 (2008).
7. Di-Poi, N. *et al.* Changes in *Hox* genes' structure and function during the evolution of the squamate body plan. *Nature* **464,** 99–103 (2010).
8. Guerreiro, I. *et al.* Role of a polymorphism in a Hox/Pax-responsive enhancer in the evolution of the vertebrate spine. *Proc. Natl Acad. Sci. USA* **110,** 10682–10686 (2013).
9. Fry, B. G. From genome to 'venome': molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. *Genome Res.* **15,** 403–420 (2005).
10. Vonk, F. J. *et al.* The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc. Natl Acad. Sci. USA* **110,** 20651–20656 (2013).
11. Fry, B. G., Vidal, N., van der Weerd, L., Kochva, E. & Renjifo, C. Evolution and diversification of the Toxicofera reptile venom system. *J. Proteomics* **72,** 127–136 (2009).
12. Vicoso, B., Emerson, J. J., Zektser, Y., Mahajan, S. & Bachtrog, D. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biol.* **11,** e1001643 (2013).
13. Ohno, S. *Sex Chromosomes and Sex-linked Genes* Vol. 1967 (Springer Berlin Heidelberg, 1967).
14. Kaiser, V. B. & Bachtrog, D. Evolution of sex chromosomes in insects. *Annu. Rev. Genet.* **44,** 91–112 (2010).
15. Westergaard, M. The mechanism of sex determination in dioecious flowering plants. *Adv. Genet.* **9,** 217–281 (1958).
16. Castoe, T. A. *et al.* The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proc. Natl Acad. Sci. USA* **110,** 20645–20650 (2013).
17. Woltering, J. M. *et al.* Axial patterning in snakes and caecilians: evidence for an alternative interpretation of the *Hox* code. *Dev. Biol.* **332,** 82–89 (2009).
18. Head, J. J. & Polly, P. D. Evolution of the snake body form reveals homoplasy in amniote *Hox* gene function. *Nature* **520,** 86–89 (2015).
19. Gomez, C. *et al.* Control of segment number in vertebrate embryos. *Nature* **454,** 335–339 (2008).
20. Zhang, B. *et al.* Transcriptome analysis of *Deinagkistrodon acutus* venomous gland focusing on cellular structure and functional aspects using expressed sequence tags. *BMC Genomics* **7,** 152 (2006).
21. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2,** 10 (2013).
22. Li, Z. *et al.* Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct. Genom.* **11,** 25–37 (2012).
23. Alfoldi, J. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477,** 587–591 (2011).
24. Pyron, R. A. & Burbrink, F. T. Extinction, ecological opportunity, and the origins of global snake diversity. *Evolution* **66,** 163–178 (2012).
25. Yunfang, Q., Xingfu, X., Youjin, Y., Fuming, D. & Meihua, H. Chromosomal studies on six species of venomous snakes in Zhejiang. *Acta Zool. Sin.* **273,** 218–227 (1981).
26. Srikulnath, K. *et al.* Karyotypic evolution in squamate reptiles: comparative gene mapping revealed highly conserved linkage homology between the butterfly lizard (*Leiolepis reevesii rubritaeniata*, Agamidae, Lacertilia) and the Japanese four-striped rat snake (*Elaphe quadrivirgata*, Colubridae, Serpentes). *Chromosome Res.* **17,** 975–986 (2009).
27. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15,** 568–573 (1998).
28. Martill, D. M., Tischlinger, H. & Longrich, N. R. A four-legged snake from the Early Cretaceous of Gondwana. *Science* **349,** 416–419 (2015).
29. Held, J. *How the Snake Lost Its Legs: Curious Tales from the Frontier of Evo-Devo* (Cambridge University Press, 2014).
30. Lee, D. *et al.* Epiregulin is not essential for development of intestinal tumors but is required for protection from intestinal damage. *Mol. Cell Biol.* **24,** 8907–8916 (2004).
31. Banting, G. S. *et al.* CECR2, a protein involved in neurulation, forms a novel chromatin remodeling complex with SNF2L. *Hum. Mol. Genet.* **14,** 513–524 (2005).

32. Iwao, K. *et al.* Heparan sulfate deficiency leads to Peters anomaly in mice by disturbing neural crest TGF-β2 signaling. *J. Clin. Invest.* **119,** 1997–2008 (2009).
33. Xu, B. *et al. Hox5* interacts with *Plzf* to restrict *Shh* expression in the developing forelimb. *Proc. Natl Acad. Sci. USA* **110,** 19438–19443 (2013).
34. Boulet, A. M. & Capecchi, M. R. Multiple roles of *Hoxa11* and *Hoxd11* in the formation of the mammalian forelimb zeugopod. *Development* **131,** 299–309 (2004).
35. King, M., Arnold, J. S., Shanske, A. & Morrow, B. E. T-genes and limb bud development. *Am. J. Med. Genet. A* **140,** 1407–1413 (2006).
36. Logan, M. & Tabin, C. J. Role of Pitx1 upstream of Tbx4 in specification of hindlimb identity. *Science* **283,** 1736–1739 (1999).
37. Wellik, D. M. & Capecchi, M. R. *Hox10* and *Hox11* genes are required to globally pattern the mammalian skeleton. *Science* **301,** 363–367 (2003).
38. McGrew, M. J., Dale, J. K., Fraboulet, S. & Pourquie, O. The *lunatic Fringe* gene is a target of the molecular clock linked to somite segmentation in avian embryos. *Curr. Biol.* **8,** 979–982 (1998).
39. Muragaki, Y., Mundlos, S., Upton, J. & Olsen, B. R. Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. *Science* **272,** 548–551 (1996).
40. Kaske, S. *et al.* TRPM5, a taste-signaling transient receptor potential ion-channel, is a ubiquitous signaling component in chemosensory cells. *BMC Neurosci.* **8,** 49 (2007).
41. Dehara, Y. *et al.* Characterization of squamate olfactory receptor genes and their transcripts by the high-throughput sequencing approach. *Genome Biol. Evol.* **4,** 602–616 (2012).
42. Khan, I. *et al.* Olfactory receptor subgenomes linked with broad ecological adaptations in Sauropsida. *Mol. Biol. Evol.* **32,** 2832–2843 (2015).
43. Hayden, S. *et al.* A cluster of olfactory receptor genes linked to frugivory in bats. *Mol. Biol. Evol.* **31,** 917–927 (2014).
44. Daltry, J. C., Wuster, W. & Thorpe, R. S. Diet and snake venom evolution. *Nature* **379,** 537–540 (1996).
45. Hargreaves, A. D., Swain, M. T., Hegarty, M. J., Logan, D. W. & Mulley, J. F. Restriction and recruitment-gene duplication and the origin and evolution of snake venom toxins. *Genome Biol. Evol.* **6,** 2088–2095 (2014).
46. Casewell, N. R., Harrison, R. A., Wuster, W. & Wagstaff, S. C. Comparative venom gland transcriptome surveys of the saw-scaled vipers (Viperidae: *Echis*) reveal substantial intra-family gene diversity and novel venom transcripts. *BMC Genom.* **10,** 564 (2009).
47. Matsubara, K. *et al.* Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. *Proc. Natl Acad. Sci. USA* **103,** 18190–18195 (2006).
48. Ohno, S. *Sex chromosomes and sex-linked genes* Vol. 1 (Springer, 1967).
49. Zhou, Q. *et al.* Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346,** 1246338 (2014).
50. Cortez, D. *et al.* Origins and functional evolution of Y chromosomes across mammals. *Nature* **508,** 488–493 (2014).
51. Nicolas, M. *et al.* A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. *PLoS Biol.* **3,** e4 (2005).
52. Li, W. H., Yi, S. & Makova, K. Male-driven evolution. *Curr. Opin. Genet. Dev.* **12,** 650–656 (2002).
53. Vicoso, B. & Charlesworth, B. Evolution on the X chromosome: unusual patterns and processes. *Nat. Rev. Genet.* **7,** 645–653 (2006).
54. Mank, J. E. The W, X, Y and Z of sex-chromosome dosage compensation. *Trends Genet.* **25,** 226–233 (2009).
55. Boucherat, O. *et al.* Partial functional redundancy between *Hoxa5* and *Hoxb5* paralog genes during lung morphogenesis. *Am. J. Physiol. Lung Cell Mol. Physiol.* **304,** L817–L830 (2013).
56. Guerreiro, I. & Duboule, D. Snakes: hatching of a model system for Evo-Devo? *Int. J. Dev. Biol.* **58,** 727–732 (2014).
57. Booth, W., Johnson, D. H., Moore, S., Schal, C. & Vargo, E. L. Evidence for viable, non-clonal but fatherless Boa constrictors. *Biol. Lett.* **7,** 253–256 (2011).
58. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463,** 311–317 (2010).
59. Luo, R. *et al.* SOAP *de novo* 2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1,** 18 (2012).
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
61. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16,** 418–420 (2000).
62. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
63. Birney, E., Clamp, M. & Durbin, R. Genewise and genomewise. *Genome Res.* **14,** 988–995 (2004).
64. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32,** W309–W312 (2004).
65. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
66. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25 (2010).
67. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
68. Benton, M. J. & Donoghue, P. C. Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* **24,** 26–53 (2007).
69. Li, H. *et al.* TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34,** D572–D580 (2006).
70. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22,** 1269–1271 (2006).
71. Loytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA* **102,** 10557–10562 (2005).
72. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56,** 564–577 (2007).
73. Bult, C. J. *et al.* Mouse genome database 2016. *Nucleic Acids Res.* **44,** D840–D847 (2016).
74. Weng, M. P. & Liao, B. Y. MamPhEA: a web tool for mammalian phenotype enrichment analysis. *Bioinformatics* **26,** 2212–2213 (2010).
75. Steiger, S. S., Kuryshev, V. Y., Stensmyr, M. C., Kempenaers, B. & Mueller, J. C. A comparison of reptilian and avian olfactory receptor gene repertoires: species-specific expansion of group γ genes in birds. *BMC Genom.* **10,** 446 (2009).
76. Mackessy, S. P. *Handbook of venoms and toxins of reptiles* (CRC Press, 2016).
77. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).
78. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
79. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
80. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

Acknowledgements

This project was supported by the Thousand Young Talents Program funding, and a startup funding of Life Science Institute of Zhejiang University to Q.Z., and National Major Scientific and Technological Special Project for 'Significant New Drugs Development' during the Twelfth Five-year Plan Period (no. 2009ZX09102-217, 2009–2010) to W.Y.

Author contributions

Q.Z., W.Y. and G.Y. conceived and supervised the project. B.L., P.Q., W.Z., Y.H. and X.S. collected the samples. Z.L. and B.W. performed the proteomic mass spectrometry-based experiment and data analysis. J.L., Z.W. and Y.Z. performed the genome assembly and annotation. J.L. and Q.L. designed and performed the identification of sex-linked scaffolds. P.Z. performed RNA-seq data analysis. Z.W., L.J. and Y.Z. performed the genome evolution analyses. Z.W., Y.Z., L.J. and B.Q. performed genes and gene family evolution analyses. Z.W. and L.J. performed the sex chromosome evolution analyses. Q.Z., Z.W., G.Z., G.Y. interpreted the results and wrote the manuscript. All of the authors have read and approved the final manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Yin, W. *et al.* Evolutionary trajectories of snake genes and genomes revealed by comparative analyses of five-pacer viper. *Nat. Commun.* **7,** 13107 doi: 10.1038/ncomms13107 (2016).