ISSN 2472-1972

Single-Nucleotide Variations of the Human Nuclear Hormone Receptor Genes in 60,000 Individuals

Rafah Mackeh,[1] Alexandra K. Marr,[1] Soha R. Dargham,[2] Najeeb Syed,[3] Khalid A. Fakhro,[1,4] and Tomoshige Kino[1]

[1]Department of Human Genetics, Division of Translational Medicine, Sidra Medical and Research Center, Doha 26999, Qatar; [2]Biostatistics, Epidemiology and Biomathematics Research Core, Weill Cornell Medicine in Qatar, Doha 24811, Qatar; [3]Division of Biomedical Informatics, Sidra Medical and Research Center, Doha 26999, Qatar; and [4]Department of Genetic Medicine, Weill Cornell Medicine in Qatar, Doha 24144, Qatar

Nuclear hormone receptors (NRs) mediate biologic actions of lipophilic molecules to gene transcription and are phylogenetically and functionally categorized into seven subfamilies and three groups, respectively. Single-nucleotide variations (SNVs) or polymorphisms are genetic changes influencing individual response to environmental factors and susceptibility to various disorders, and are part of the genetic diversification and basis for evolution. We sorted out SNVs of the human *NR* genes from 60,706 individuals, calculated three parameters (percentage of all variants, percentage of loss-of-function variants, and ratio of nonsynonymous/synonymous variants in their full protein-coding or major domain–coding sequences), and compared them with several valuables. Comparison of these parameters between *NRs* and control groups identified that *NRs* form a highly conserved gene family. The three parameters for the full coding sequence are positively correlated with each other, whereas four *NR* genes are distinct from the others with much higher tolerance to protein sequence-changing variants. DNA-binding domain and *N*-terminal domain are respectively those bearing the least and the most variation. NR subfamilies based on their phylogenetic proximity or functionality as well as diversity of tissue distribution and numbers of partner molecules are all not correlated with the variation parameters, whereas their gene age demonstrates an association. Our results suggest that the natural selection driving the *NR* family evolution still operates in humans. Gene age and probably the potential to adapt to various new ligands, but not current functional diversity, are major determinants for SNVs of the human *NR* genes.

Copyright © 2018 Endocrine Society

This article has been published under the terms of the Creative Commons Attribution Non-Commercial, No-Derivatives License (CC BY-NC-ND; https://creativecommons.org/licenses/by-nc-nd/4.0/).

Freeform/Key Words: evolution, functional diversity, gene age, ligand-binding pocket (LBP), major domains

The nuclear hormone receptors (NRs) are ligand-dependent transcription factors, forming a large family consisting of >200 members throughout the metazoans [1, 2]. NRs mediate biologic actions of small lipophilic molecules, including metabolites of cholesterol and fatty acids, steroid hormones, retinoids, and vitamin D [1, 2]. Some NRs do not have ligands, and thus they are called orphan receptors [3]. NRs have diverse regulatory actions in a wide range of biological processes, such as cell proliferation and differentiation, circadian rhythms, growth and aging, immunity, reproduction, and intermediary metabolism [4]. They exert

Abbreviations: AF-2, activation function 2; DBD, DNA-binding domain; ExAC, Exome Aggregation Consortium; HoR, hormonal receptor; HR, hinge region; LBD, ligand-binding domain; LBP, ligand-binding pocket; LoF, loss of function; MeR, metabolic receptor; NR, nuclear hormone receptor; NS, nonsynonymous; NTD, *N*-terminal domain; OR, orphan receptor; PCA, principal component analysis; S, synonymous; SNV, single-nucleotide variation.

Received 18 October 2017
Accepted 28 November 2017
First Published Online 4 December 2017

their effects primarily by regulating the transcriptional activity of their responsive genes through binding to specific DNA recognition motifs called response elements [5, 6].

NRs consists of four major functional domains/region: the *N*-terminal domain (NTD, also called A/B region), the DNA-binding domain (DBD, also called C region), the hinge region (HR, also called D region), and the ligand-binding domain (LBD, also called E/F region) [1, 7]. NTD has ligand-independent transactivation domain(s) and contains several amino acids subjected to posttranslational modification [1, 5]. DBD has two C4-type zinc finger motifs through which it binds a DNA response element [8]. LBD has a ligand-binding pocket (LBP) and a ligand-dependent transactivation domain called activation function 2 (AF-2) [1, 9]. Finally, HR acts as a linker for DBD and LBD and determines in part the number of spacing nucleotides in tandem response elements [1, 5]. NTD is the most variable region throughout NRs for both amino acid sequence and peptide size, whereas DBD is the most conserved domain [10–13].

NRs can be observed from simple metazoans to humans [14]. NRs are highly conserved molecules and originated from a common ancestral protein that harbors DBD and LBD [15–19]. These pieces of evidence indicate that the current human *NR* genes have been created through a long history of natural selection, coping with random occurrence of various genetic changes [20, 21].

Single-nucleotide variations (SNVs) or polymorphisms are genetic alterations accounting for >80% of all variations with a frequency of >1% [22]. A single human genome contains millions of SNVs, indicating that they can be identified virtually in every 300 base pairs [23]. Within the protein-coding genes, SNVs can be of two types: (1) synonymous (S) variation (the nucleotide replacements that do not change coding amino acids) and (2) nonsynonymous (NS) variation (the nucleotide replacements that do change coding amino acids) [24]. SNVs also occur within the non-protein-coding area where they sometimes interfere with messenger RNA expression of the nearby genes, such as by affecting RNA splicing and by changing access of the transcription factors to gene regulatory elements [25–27]. Although >90% of SNVs are not associated with obvious biological consequences in humans [28, 29], some SNVs may cause beneficial or adverse effects [30]. Furthermore, they are among the strong driving forces for promoting organism evolution and diversification. Indeed, the NS variants previously inserted into the *NR* genes and surviving against genetic selection have contributed to the creation of the current NRs that demonstrate broad specificities to ligands and downstream biological activities [14]. Thus, the gene variations found in the current human *NR* genes may underlie the future sequence alteration to be observed in forthcoming NRs. Therefore, we quantified SNVs in 47 out of all 48 human *NR* genes by browsing a collection of sequencing data from ~60,000 unrelated individuals and revealed a "snapshot" of the genetic variation found in the *NR* genes of the current human population. Our results suggest that the rule that previously drove NR evolution/diversification in their major domains still operates in the human *NR* genes. Furthermore, gene age and a high adaptation potential to new ligands, but not functional diversity, are major determinants for the accumulation of SNVs in the human *NR* genes.

1. Materials and Methods

A. Determination of Nucleotide Sequences for Major Domains/Regions of NRs

We used 47 human *NR* genes in this study (Supplemental Table 1). We omitted *NR2E3* (*PNR*) from our analysis, as variation data for this gene were not available in the Exome Aggregation Consortium (ExAC) Browser (Beta) (http://exac.broadinstitute.org/). Various reports use slightly different amino acid sequences of the four major domains/region (NTD, DBD, HR, and LBD) of each NR, and thus we used a single data resource relevant to all NRs to determine their domain/region sequences. We retrieved the information from the Pfam source in the Ensembl database (www.ensembl.org), accessing to its "domains & features display" for the canonical transcripts only. NTD was considered the domain spanning from the translation

start site of the respective NR to the nucleotide prior to the start codon of its DBD. HR was similarly determined based on the end and the start nucleotides of DBD and LBD, respectively. The end nucleotide of LBD was determined by the stop codon of the NR protein.

B. Counting of SNVs

SNVs in each NR protein-coding sequence were retrieved from the ExAC browser by entering the gene name and by selecting the "canonical transcript." For members of the human leukocyte antigen (*HLA*) or histone deacetylase (*HDAC*) families, their SNVs were also retrieved from their canonical transcripts found in the same database (Supplemental Tables 2 and 3). The following variations were counted: NS (same as missense), S, "in-frame insertion/deletion," "start lost," "stop gained," "stop lost," "frameshift," "splice donor," and "splice acceptor." Among these, "start lost," "stop gained," "stop lost," "frameshift," "splice donor," and "splice acceptor" variations were considered loss-of-function (LoF) variation. For data quality control, any variants with an allele frequency <0.001 and observed $<20{,}000$ alleles were excluded from analysis. Three parameters—percentage of all variants, percentage of LoF variants, and ratio of NS/S—variants were then calculated for the whole protein and four major domains/region of each NR (Supplemental Table 4.1, 4.2, 4.3, 4.4, and 4.5). Percentage of all variants and percentage of LoF variants were calculated by dividing the number of corresponding variants with the length of the coding nucleotides.

C. Grouping of NRs Into Phylogenetic or Functional Subfamilies/Groups

NRs were classified phylogenetically into seven subfamilies (NR0 to NR6: groups of the *NR* genes with high sequence similarity and often paralogous relationship in vertebrates) [31] or functionally into three groups: metabolic receptors (MeRs; 21 NRs: RARα, RARβ, RARγ, PPARα, PPARβ/δ, PPARγ, Rev-erbα, Rev-erbβ, RORα, RORβ, RORγ, LXRα, LXRβ, FXRα, PXR, CAR, HNF4α, HNF4γ, RXRα, RXRβ, and RXRγ), orphan receptors [ORs; 14 NRs: TR2, TR4, TLX (TLL), COUP-TFI, COUP-TFII, EAR2, NUR77, NURR, NOR1, SF1, LRH-1, GCNF, DAX-1, and SHP], and hormonal receptors (HoRs; 12 NRs: TRα, TRβ, VDR, ERα, ERβ, ERRα, ERRβ, ERRγ, GR, MR, PR, and AR) [1].

D. Data Extraction for NR Gene Age, Organ/Tissue Expression, and Interacting Proteins

Gene age of the 46 human *NR* genes was sorted out from the GenTree browser (http://gentree.ioz.ac.cn) in which gene age is determined with the 13 timeframes based on branching of the vertebrate phylogenetic tree [32]. Thirty-eight *NR* genes were found in branch 0 (~454.6 million years ago), whereas seven and one *NR* genes were in branch 1 (~361.2 million years ago) and in branch 3 (~220.2 million years ago), respectively (Supplemental Table 5). Gene age of the *NR1D1* (*REV-ERB*α) was not available in this site. We extracted the number of NR-expressing tissues in 27 different organs/tissues from the Human Protein Atlas (http://www.proteinatlas.org/) that integrates RNA and protein expression data for ~80% of the human protein-coding genes [33] (Supplemental Table 5). We considered that NRs are expressed if fragments per kilobase of exon per million mapped of a *NR* messenger RNA are >1, as previously suggested [34]. The number of proteins interacting with NRs was retrieved from the Human Integrated Protein-Protein Interaction Reference browser (http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie), where the interaction is determined based on the amount and quality of the evidence for a given protein-protein interaction [35] (Supplemental Table 5).

E. Statistical Analysis

Kruskal-Wallis, two-way analysis of variance, or Mann-Whitney tests were used, depending on the data distribution and the sample size. The significance level was set to 0.05. Although we performed statistical analyses for all possible comparisons, only the results of comparisons

with statistical significance were demonstrated in the figures to avoid complexity. Most of the data analyses were performed with the GraphPad Prism software version 7.0b (GraphPad Software, La Jolla, CA). A bioconductor package Consensusclusterplus (https://academic.oup.com/bioinformatics/article/26/12/1572/281699/ConsensusClusterPlus-a-class-discovery-tool-with) with an agglomerative hierarchical clustering option was used for unsupervised clustering of the data. Principal component analysis (PCA) was also performed, and PC1 and PC2 coordinates were plotted to show the variance.

2. Results

A. NR Genes Form a Highly Conserved Gene Family

We first counted SNVs in the 47 human *NR* genes and calculated three parameters: (1) percentage of all variants, (2) percentage of LoF variants, and (3) ratio of NS/S variants (Supplemental Tables 1 and 4.1). Percentages of all variants ranged between 22.03% (*NR2F6*: *EAR2*) and 68.87% (*NR3A2*: *ERβ*) and averaged 43.61%. To assess overall tolerance of the human *NR* genes against SNVs, we compared the three variation parameters of all *NR* genes to those of the families known to be highly (*HLAs*) or little (*HDACs*) tolerant to gene variation [36] (Fig. 1). Such characteristics of these control families are evident in the ExAC browser, where *HLAs* and *HDACs* respectively demonstrated statistically significant low and high scores of pLI (probability of being LoF intolerant) (Supplemental Tables 2 and 3). *NRs*, *HDACs*, and *HLAs* demonstrated similar levels of the percentage of all variants (Fig. 1A). *NRs* showed a significantly lower percentage of LoF variants and lower ratio of NS/S variants to those of *HLAs* but equivalent to *HDACs* (Fig. 1B and 1C). LoF variation potentially and largely affects functions of the encoded proteins or knock outs their expression, and a low ratio of NS/S variants indicates a tendency of preserving protein sequences against SNVs [37]. Thus, these results indicate that *NR*s form a conserved gene family like *HDACs* that are highly intolerant to (or easily eliminated through natural selection) the SNVs potentially affecting protein sequences and/or functions of the encoded proteins.

B. Three Variation Parameters of the Human NR Genes are Correlated with Each Other, Whereas Four NR Genes Demonstrate High Variation Profiles Distinct From the Others

We next evaluated the relationship between the three variation parameters of the human *NR* genes. We found that they were strongly and positively correlated with each other in linear regression analyses ($P < 0.0001$) (Fig. 2A). In these analyses, four *NR* genes—*NR0B2* (*SHP*),

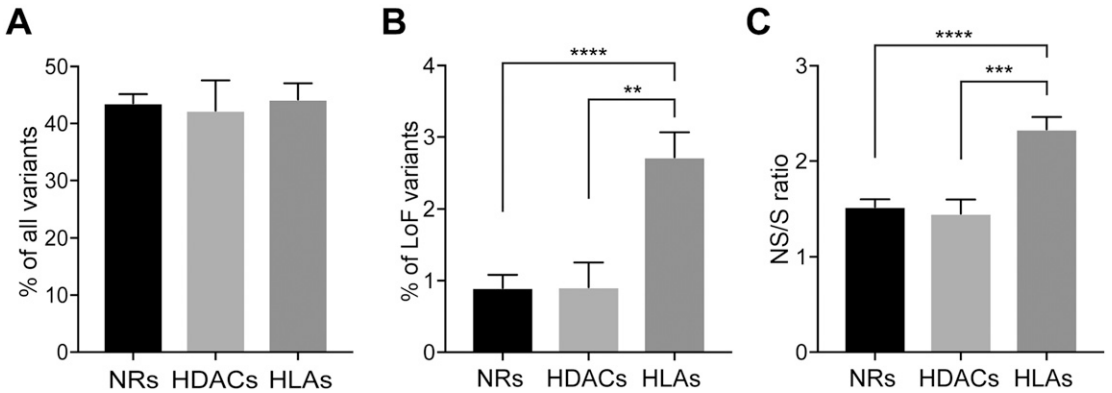

Figure 1. The *NR* genes form a highly conserved family. Percentages of (A) all or (B) LoF variants or (C) ratios of NS/S variants between *NR*, *HDAC*, and *HLA* families are shown. Bars represent mean ± standard error of the mean values of the indicated parameters. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$, compared between the two gene families indicated. Only the results of statistical analyses having significant difference are shown.

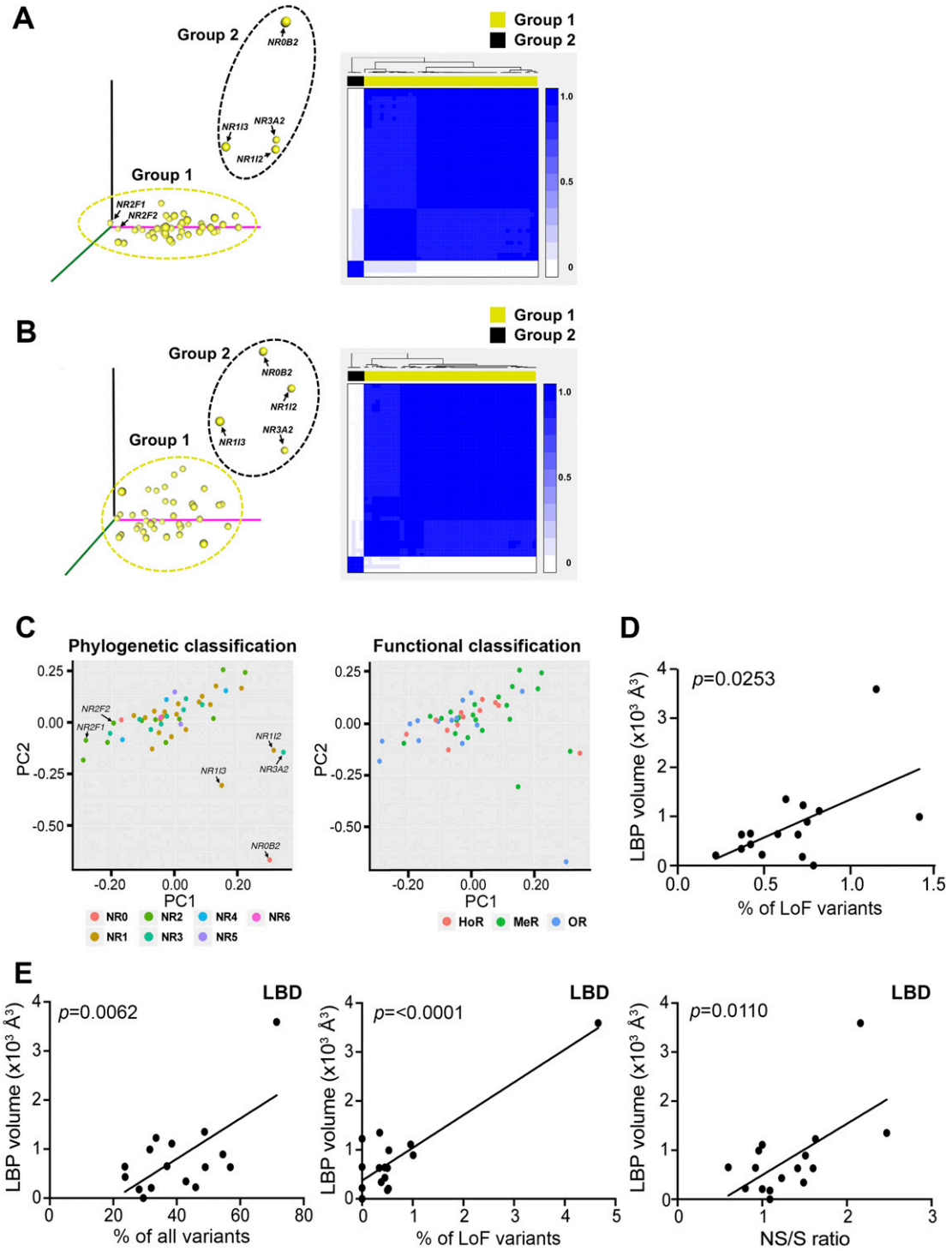

Figure 2. Four *NR* genes demonstrate high variation profiles distinct from the other *NR* genes. (A, B) Relationship between the three variation parameters of the *NR* genes. The three variation parameter values of the (A) full protein-coding (left panel) or (B) LBD-coding (left panel) sequence of the human *NR* genes are shown in three-dimensional plots. Pink, gray, and green axes indicate percentage of all variants, percentage of LoF variants, and ratio of NS/S variants, respectively. Results of the hierarchical cluster analysis for the full protein- or LBD-coding sequence of the human *NR* genes (right panels of A and B) demonstrate that the four *NR* genes [*NR0B2* (*SHP*), *NR1I2* (*PXR*), *NR1I3* (*CAR*), and *NR3A2* (*ERβ*): group 2] are distinct from the other *NR* genes (group 1), due to their high percentage of LoF variants and high ratio of NS/S variants. Scale legends are shown in the right side of the

heatmaps. *NR0B2* (*SHP*), *NR1I2* (*PXR*), *NR1I3* (*CAR*), and *NR3A2* (*ERβ*), as well as *NR2F1* (*COUP-TFI*) and *NR2F2* (*COUP-TFII*), are indicated in panel A, whereas the former four genes are pointed in panel B. (C) PCA for the three variation parameters of the full protein-coding sequence of the *NR* genes. Results of the PCA labeled with phylogenetic (left panel) or functional (right panel) classification of the 47 *NR* genes are shown. The four distinct *NR* genes—*NR0B2* (*SHP*), *NR1I2* (*PXR*), *NR1I3* (*CAR*), and *NR3A2* (*ERβ*)—are indicated in the left panel. (D) LBP size is moderately and positively correlated with the percentage of LoF variants of the full protein-coding sequence in 16 *NR* genes. LBP size of the 16 NRs (RARα, PPARγ, RORα, RORβ, LXRα, FXRα, VDR, PXR, HNF4α, HNF4γ, RXRα, ERα, ERRγ, GR, PR, and LRH-1) previously evaluated in the structural analysis [40] is moderately correlated with the percentage of LoF variants of their full protein-coding sequence in the linear regression analysis. (E) LBP size is strongly and positively correlated with all three variation parameters of the LBD-coding sequence in 16 *NR* genes. LBP size of the 16 NRs is strongly correlated with the percentage of all variants (left panel), percentage of LoF variants (middle panel), and ratio of NS/S variants (right panel) of their LBD-coding sequence in the linear regression analysis.

NR1I2 (*PXR*), *NR1I3* (*CAR*), and *NR3A2* (*ERβ*) (group 2)—were distinct from the remaining *NR* genes (group 1), as evident in the cluster analysis (Fig. 2A, right panel) as well as in the PCA (Fig. 2C). These genes demonstrated a much higher percentage of LoF variants and a higher ratio of NS/S variants against the percentage of all variants compared with the group 1 *NR* genes, indicating that they are highly tolerant to (thus allow) the SNVs potentially affecting their amino acid sequences and/or functions. On the other hand, *NR2F1* (*COUP-TFI*) and *NR2F2* (*COUP-TFII*) showed the least variation profiles compared with the other *NR* genes, although they did not make a distinct group (Fig. 2A).

We investigated whether the excess number of variants observed in the four *NR* genes is randomly distributed in their full-coding sequence or accumulated in specific domains/regions. Thus, we examined the association of the three variation parameters in the coding sequence of the four domains/region of all *NR* genes and found that the LBD-coding sequence, but not the other domains/region-coding sequences, of the four distinct *NR* genes [*NR0B2* (*SHP*), *NR1I2* (*PXR*), *NR1I3* (*CAR*), and *NR3A2* (*ERβ*)] demonstrated similar profiles of a high percentage of LoF variants and a high ratio of NS/S variants against the percentage of all variants (Fig. 2B and Supplemental Table 4.2, 4.3, 4.4, and 4.5). This result indicates that LBD is responsible for the high variation profiles found in the full protein-coding sequence of the four distinct *NR* genes.

PXR and CAR are known as xenobiotic receptors with their ability to bind a variety of endogenous and exogenous compounds (39), whereas ERβ has a higher affinity to various estrogenic endocrine disruptors (xenoestrogens) compared with ERα [39]. Thus, we hypothesized that the high variation property of these *NR* genes may support their flexibility in interacting with broad ranges of ligands, including those newly appearing in the environment. As LBP size may be associated with ligand plasticity [40], we compared LBP volume of the 16 NRs already under structural analysis with the three variation parameters of their full protein-coding sequence and found that their LBP volume was moderately and positively correlated with the percentage of LoF variants among the three variation parameters (Fig. 2D). Because LBD of the four *NR* genes is responsible for their high variation profiles observed in their full protein-coding sequence, we performed the comparison using the LBD data of the 16 *NR* genes and found a much stronger correlation of their LBP volume to all three variation parameters (Fig. 2E). These results indicate that the high variation character of these *NR* genes may underlie their high flexibility in interacting with various compounds appearing in the environment based on their high capacity to maintain sequence diversity.

C. NTD and DBD are Respectively the Domains Bearing the Most and Least Variation

We next examined distribution of SNVs in four major domains/region of the NRs: NTD, DBD, HR, and LBD (Fig. 3). DBD demonstrated significantly lower percentage of all variants than NTD and HR but had no statistical difference against LBD. HR showed a statistically significant higher percentage of all variants than LBD but not NTD (Fig. 3A). There was no

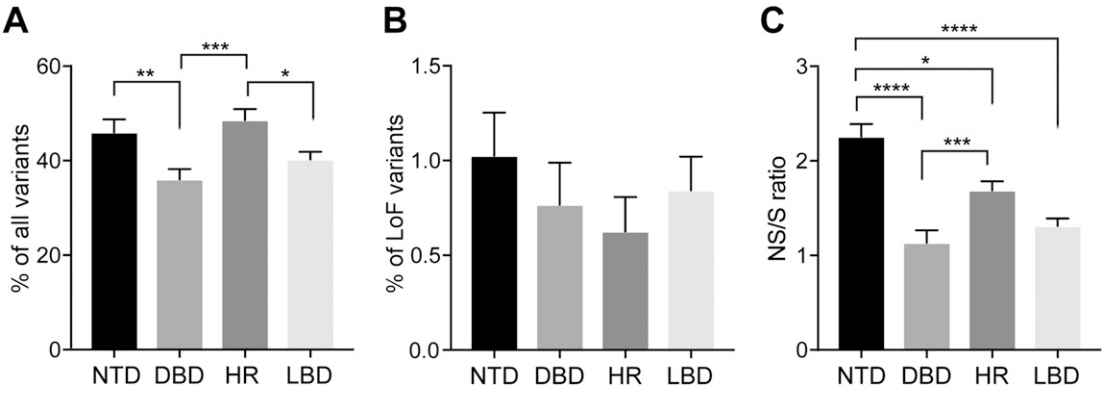

Figure 3. DBD and NTD are respectively the domains with the least and the most gene variation. Percentages of (A) all or (B) LoF variants or (C) ratios of NS/S variants in the coding sequence of the NTD, DBD, HR, and LBD of all *NR* genes are shown. Bars represent mean ± standard error of the mean values of the indicated parameters. $*P < 0.05$, $**P < 0.01$, *** < 0.001, $****P < 0.0001$, compared between the two domains indicated. Only the results of statistical analyses having significant difference are shown.

statistical difference between the four domains/region for percentage of LoF variants (Fig. 3B). DBD again demonstrated a statistically significant lower ratio of NS/S variants than NTD and HR but not LBD. On the other hand, NTD showed a significantly higher ratio of NS/S variants than HR and LBD (Fig. 3C). Taken together, these results suggest that DBD is a domain of the least genetic variation, whereas NTD is of the most variation. However, all domains/regions showed no significant difference in the percentage of LoF variants; thus, they appear to be equally sensitive to the variations that significantly alter or knock out functions/expression of the encoding proteins.

D. Phylogenetic or Functional NR Subfamilies/Groups Do Not Strongly Influence Their Gene Variation

We then examined SNVs within the NR subfamilies/groups due to their phylogenetic proximity (NR0 to NR6) or functionality (MeRs, ORs, and HoRs). For the former subfamilies, NR0, NR5, and NR6 were excluded from our analyses because they have insufficient numbers of subfamily members for statistical evaluation.

We found that NR1 to NR4 subfamilies showed no difference in their percentage of all variants and percentage of LoF variants (Fig. 4A–4C). We further examined these parameters in four major domains/region and found that percentage of all variants and percentage of LoF variants were similar between the same domains of all indicated subfamilies (Fig. 4D–4F). Similarly, functional NR groups (MeRs, ORs, and HoRs) showed no difference in their percentage of all variants, percentage of LoF variants, and ratio of NS/S variants (Fig. 5A–5C). Furthermore, they did not demonstrate any difference in all of these variation parameters between the groups inside the same domains, except for percentage of all variants for NTD between MeRs and ORs (Fig. 5D–5F). These results were also confirmed in the PCA employing all *NR* genes (Fig. 2C) where members of the same subfamilies/groups did not form distinct clusters. Using the same data sets, we compared the difference between four major domains/region inside the same phylogenetic or functional subfamilies/groups and found that, although weak, the tendency identified in all *NR* genes for DBD and NTD to harbor the least and the most variation, respectively, was consistent across different NR subfamilies/groups (Supplemental Fig. 1). Taken together, our results indicate that phylogenetic proximity or functional difference has a limited influence on the diversity of SNVs in the human *NR* genes.

E. NR Gene Age is Associated With SNV Accumulation, But the Number of Their Expressing Organs/Tissues and of Interacting Proteins is Not

We explored the factors that potentially influence SNVs of the human *NR* genes. We first focused on gene age, which was recently determined for over half of the human protein-coding

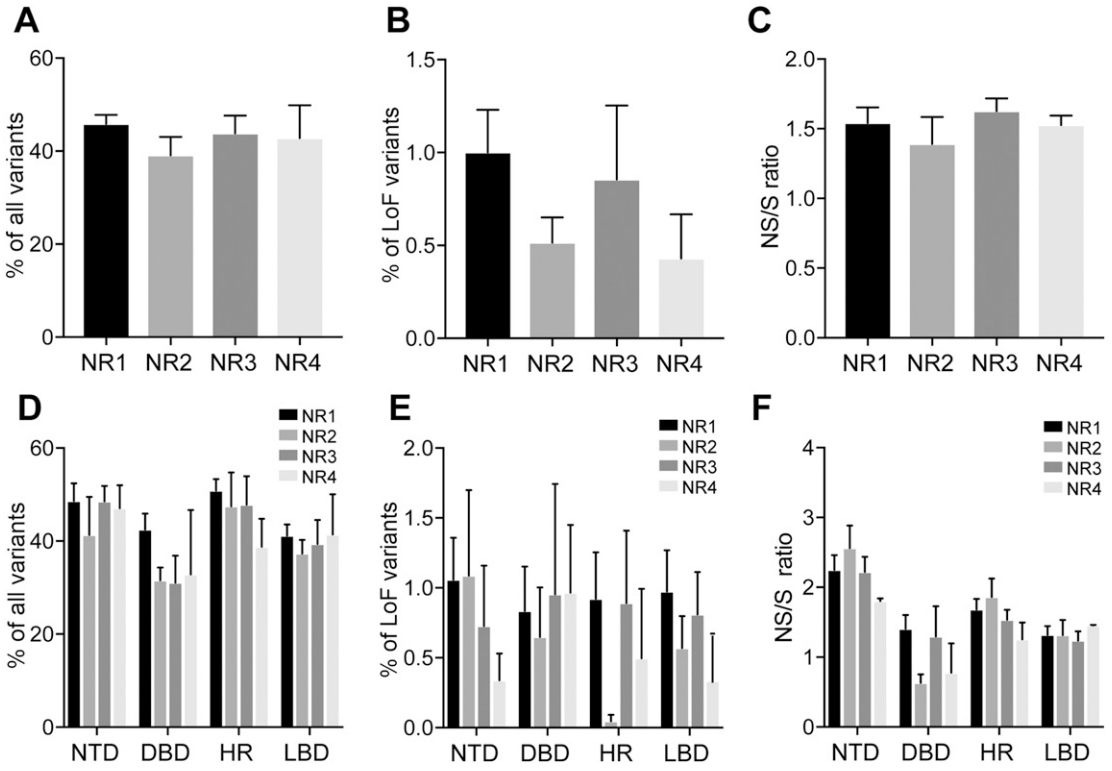

Figure 4. Phylogenetic proximity of the *NR* genes little influences *NR* gene variation. Percentages of (A, D) all or (B, E) LoF variants or (C, F) ratios of NS/S variants in (A–C) different NR subfamilies or in (D–F) their major domains/region are shown. NR0, NR5, and NR6 were excluded from the comparisons due to insufficient numbers of family members for statistical evaluation. Bars represent mean ± standard error of the mean values of the indicated parameters. Comparisons in panels D, E, and F were made inside the same domains/regions. Kruskal-Wallis test was used for statistical analyses. All possible comparisons do not reach statistical significance.

genes [32]. We successfully retrieved gene age for 46 human *NRs*: 38 *NRs* are in branch 0 (~454.6 million years ago), whereas 7 *NRs* are in branch 1 (~361.2 million years ago) and 1 *NR* is in branch 3 (~220.2 million years ago) (Supplemental Table 5). Because most of the *NR* genes were found in branch 0 but only one was in branch 3, we categorized them into two groups: (1) old *NR* genes found in branch 0 and (2) young *NR* genes in branch 1 or 3. We found that young *NR* genes demonstrated a significantly higher percentage of all variants and percentage of LoF variants than old *NR* genes, whereas these two groups showed no differences in the ratio of NS/S variants (Fig. 6A–6C). We thus concluded that gene age of the human *NR* genes is a factor potentially influencing the abundance of their SNVs, likely as a reflection of genetic selection pressure over organismal evolution. We also examined the correlation of the number of organs/tissues expressing respective NR or the number of partner proteins that interact with each NR to our three variation parameters by retrieving the data from publicly available data resources. We found that neither of these correlated with any of our three parameters (Fig. 6D–6F and Fig. 6G–6I, respectively). These results indicate that SNVs of the human *NR* genes are independent of their functional diversity over human organs/tissues as well as in communicating with other biological pathways, although the latter data set for NR-interacting proteins might contain some biases caused by difference in our current understanding of each NR.

3. Discussion

We present a thorough investigation on SNVs in most of the human *NR* genes by focusing on the three variation parameters: percentage of all variants, percentage of LoF variants, and

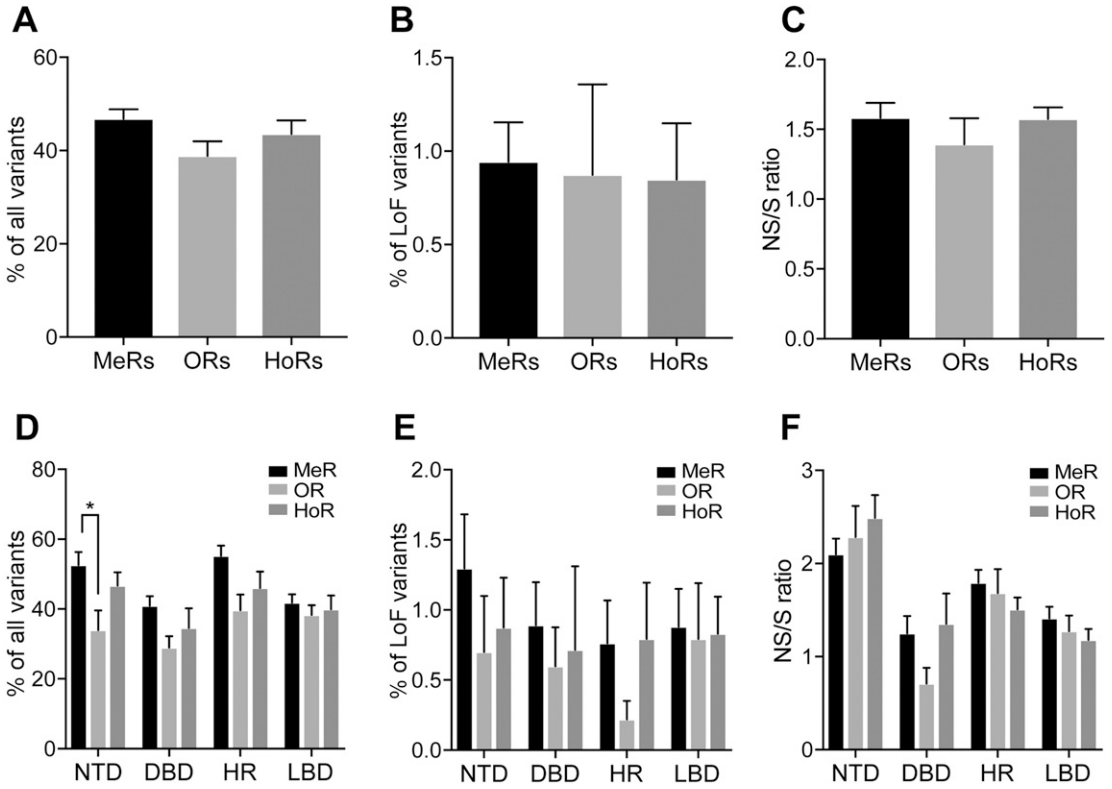

Figure 5. Functionality of NRs is not associated with *NR* gene variation. Percentages of (A, D) all or (B, E) LoF variants or (C, F) ratios of NS/S variants in (A–C) different functional NR groups or in (D–F) their major domains/region are shown. Bars represent mean ± standard error of the mean values of the indicated parameters. Comparisons in panels D, E, and F were made inside the same domains/region. Kruskal-Wallis test was used for statistical analyses. *$P < 0.05$, compared between the two functional groups indicated. Only the result of statistical analyses having a significant difference is shown.

ratio of NS/S variants. We found that, similar to the *HDAC* genes, *NRs* form a gene family with few gene variations that potentially affect coding peptide sequences and functions of the expressed proteins. Their percentage of all variants is, however, similar to that of the *HLA* and *HDAC* genes respectively known to be highly and little tolerant to (or allowing) genetic changes, indicating that gene variation may happen evenly in these three gene families. Regarding the fact that most of the *NR* genes appeared at a very early evolutionary time point in vertebrates (38 are in branch 0 in the evolutional tree), our results indicate that NRs have fundamental functions that are conserved across most of the vertebrates expressing NRs [2].

Four human *NR* genes—*NR0B2* (*SHP*), *NR1I2* (*PXR*), *NR1I3* (*CAR*), and *NR3A2* (*ERβ*)—demonstrated surprisingly high variation profiles compared with the other *NR* genes. These four *NR* genes are also exceptional among the members of their phylogenetic subfamilies or functional groups. This character of *PXR*, *CAR*, and *ERβ* may support a high potential to interact with new endogenous or exogenous ligands appearing in the environment, similar to the *HLA* genes, whose encoding proteins also have a remarkable ability to interact with new antigens [41, 42]. Indeed, ERβ has a higher affinity to estrogenic endocrine disruptors (xenoestrogens) compared with ERα, suggesting that ERβ might act as a receptor for integrating environmental changes into estrogen-organizing biological activities through its flexible LBP for recognizing new ligands, whereas ERα tends to mediate classic actions of endogenous estrogens [39]. We also found that the *NR0B2* (*SHP*) gene is highly valuable. In fact, it demonstrated the most variation compared with all other *NRs*, including the three genes with high variation profiles. Its encoding protein, SHP, is an atypical NR missing DBD [43]. SHP functions as a fine modulator for many other NRs, influencing a variety of biological

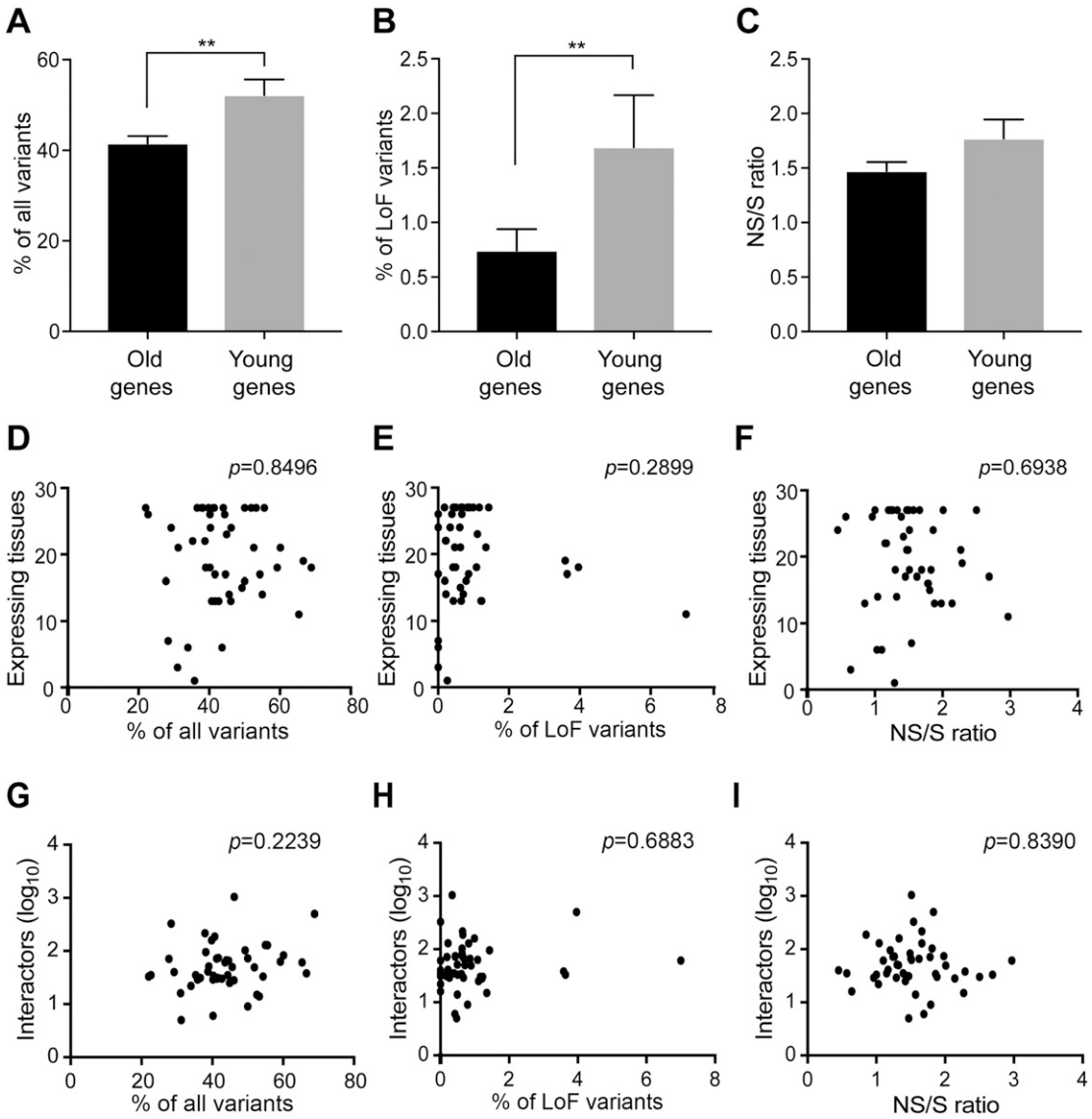

Figure 6. *NR* gene age is associated with gene variation, but diversity in tissue distribution of NRs and NR-interacting partner proteins is not. (A–C) *NR* gene age is correlated with two gene variation parameters. Percentages of (A) all or (B) LoF variants or (C) ratios of NS/S variants of old (branch 0: 38 NRs) and young (branches 1 and 3: 7 and 1 members, respectively) *NR* genes are shown. Bars represent mean ± standard error of the mean values of the indicated parameters. $^{**}P < 0.01$, compared between the two groups indicated. Only the results of statistical analyses having significant difference are shown. (D–F) Diversity of NR organ/tissue expression is not associated with *NR* gene variation. Plots between the number of organs/tissues expressing each NR and percent values of (A) all or (B) LoF variants or (C) ratios of NS/S variants are shown. Linear regression analyses show no statistical significance. (G–I) The number of NR-interacting proteins is not associated with *NR* gene variation. Plots between the number of proteins (in $\log_{10}$ scale) interacting with each NR and percent values of (A) all or (B) LoF variants or (C) ratios of NS/S variants are shown. Linear regression analyses show no statistical significance.

processes, such as lipid, glucose, bile acid and xenobiotic metabolism, and steroidogenesis, through direct physical interaction with the AF-2 surface of its partner NRs via a LxxLL motif [43–45]. The characteristic high tolerance of *PXR*, *CAR*, and *ERβ* to gene variation may be associated with their high plasticity in interacting with various ligands, and thus SHP might have similar flexibility in its modulatory activity to partner NRs, possibly through high structural variation in its peptide area harboring the LxxLL motif, which would increase its

potential to change the strength of the interaction with current partner NRs or to interact with new ones. Furthermore, SHP might have high flexibility in interacting with yet unknown ligands similar to the other three NRs, as this receptor might have them based on the evidence that synthetic retinoid-like compounds can bind SHP and regulate its activity [46]. The high variation character of SHP is in contrast to the same phylogenetic subfamily member DAX-1, which has a similar protein conformation lacking DBD [47] but is among the *NRs* with the lowest tolerance to genetic variation. This evidence suggests that the domain feature shared between these receptors does not underlie their tolerance to SNVs.

We found that DBD and NTD are respectively domains bearing the least and the most gene variation. This is consistent with the reported evidence that these domains are the most and the least conserved domains among NRs [5, 48]. Thus, our results suggest that the evolutionary process generating the current human NRs is continuously active in humans and potentially influences creation of future human NR proteins. The results also suggest that recognition of the DNA response elements by DBD is the most important determinant for NR functions among various NR activities mediated by the four major domains/regions in terms of evolutionary selection on the *NR* genes, whereas NTD may contribute to diversity of the NR activities in the human population with its higher levels of genetic variation. LBD demonstrated the characteristics similar to those of DBD, indicating substantial evolutionary constraint on its functions, such as ligand-binding activity and AF-2–mediated transcriptional regulation through interaction with ligands/transcriptional cofactors [2]. However, statistical significance of LBD against other domains/region is less than that of DBD, and thus this domain may have more flexibility than DBD for harboring genetic variations that potentially affect such molecular interactions and functions. All domains/region of the NRs demonstrated no significant difference in the percentage of LoF variants, indicating that they are equally sensitive to the variations that highly damage or knock out their functions/expression. This result may explain the clinical findings that the pathologic *NR* point mutations creating truncated proteins or abolishing entire protein expression seem to distribute evenly in their domains/regions [49, 50].

In our analyses, phylogenetic or functional grouping of the *NR* genes does not correlate with any of the three variation parameters. These unexpected findings indicate that *NRs* in different phylogenetic branches or with various functions are equally sensitive to SNVs at least for those except the four *NRs* with high variation profiles. In addition, SNVs potentially affecting NR protein functions were found with similar frequencies over all *NR* genes, further strengthening the hypothesis that most *NRs* are uniformly important for humans. Given the presence of significant difference among *NR* genes in the number of identified pathologic mutations and disease-associated SNVs [7, 49, 51, 52], it is likely that clinical biases, such as difficulty in recognizing/identifying the variation-associated phenotypes, may exist among NRs.

We found that gene age of the *NR* genes influences their accumulation/preservation of SNVs. This is expected based on the premise that younger genes may have more flexibility for accepting sequence alterations than older genes [53]. On the other hand, the number of organs/tissues expressing NRs and the number of partner protein molecules interacting with NRs do not show correlation with any of our variation parameters. These findings suggest that diversity of the current NR functions in terms of their expression sites in human organs/tissues and breadth of their interaction with other intracellular biologic pathways are maintained despite the accumulation of SNVs in the human *NR* genes.

4. Conclusions

Our examination on SNVs in the human *NR* genes revealed that they occur mainly under the driving force promoting evolution of the *NR* genes but are not related to their current functional diversity of most of the *NR* genes in humans. Based on the results for *NR1I2* (*PXR*), *NR1I3* (*CAR*), and *NR3A2* (*ERβ*), functional plasticity of NRs for interacting with new ligands may be one of the determinants for the high SNV character (thus genetic diversity) of some *NR* genes.

Acknowledgments

We thank Dr. M. Long (University of Chicago) for retrieving the data from the GenTree browser.

Financial Support: This study was supported by an internal fund of the Sidra Medical and Research Center to T.K.

Correspondence: Tomoshige Kino, MD, PhD, Division of Translational Medicine, Sidra Medical and Research Center, Out Patient Clinic, 6th Floor, Room C6-332, PO Box 26999, Al Luqta Street, Education City North Campus, Doha 26999, Qatar. E-mail: tkino@sidra.org.

Disclosure Summary: The authors have nothing to disclose.

References and Notes

1. Mackeh R, Marr AK, Fadda A, Kino T. C2H2-type zinc finger proteins: evolutionally old and new interaction partners of the nuclear hormone receptors. *Nucl Recept Signal*. 2017. In press.
2. Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schütz G, Umesono K, Blumberg B, Kastner P, Mark M, Chambon P, Evans RM. The nuclear receptor superfamily: the second decade. *Cell*. 1995;**83**(6): 835–839.
3. Burris TP, Busby SA, Griffin PR. Targeting orphan nuclear receptors for treatment of metabolic diseases and autoimmunity. *Chem Biol*. 2012;**19**(1):51–59.
4. Aranda A, Pascual A. Nuclear hormone receptors and gene expression. *Physiol Rev*. 2001;**81**(3): 1269–1304.
5. Kino T. Glucocorticoid receptor. In: De Groot LJ, Beck-Peccoz P, Chrousos G, Dungan K, Grossman A, Hershman JM, Koch C, McLachlan R, New M, Rebar R, Singer F, Vinik A, Weickert MO, eds. *Endotext*. South Dartmouth, MA: MDText.com, Inc.; 2000.
6. Chrousos GP, Kino T. Intracellular glucocorticoid signaling: a formerly simple system turns stochastic. *Sci STKE*. 2005;**2005**:pe48.
7. Kino T, Chrousos GP. Glucocorticoid and mineralocorticoid receptors and associated diseases. *Essays Biochem*. 2004;**40**:137–155.
8. Härd T, Kellenbach E, Boelens R, Maler BA, Dahlman K, Freedman LP, Carlstedt-Duke J, Yamamoto KR, Gustafsson JA, Kaptein R. Solution structure of the glucocorticoid receptor DNA-binding domain. *Science*. 1990;**249**(4965):157–160.
9. Hurt DE, Suzuki S, Mayama T, Charmandari E, Kino T. Structural analysis on the pathologic mutant glucocorticoid receptor ligand-binding domains. *Mol Endocrinol*. 2016;**30**(2):173–188.
10. Shao D, Lazar MA. Modulating nuclear receptor function: may the phos be with you. *J Clin Invest*. 1999; **103**(12):1617–1618.
11. Luisi BF, Xu WX, Otwinowski Z, Freedman LP, Yamamoto KR, Sigler PB. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature*. 1991;**352**(6335):497–505.
12. Schwabe JW, Chapman L, Finch JT, Rhodes D. The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. *Cell*. 1993; **75**(3):567–578.
13. Pratt WB, Toft DO. Steroid receptor interactions with heat shock protein and immunophilin chaperones. *Endocr Rev*. 1997;**18**(3):306–360.
14. Escriva H, Bertrand S, Laudet V. The evolution of the nuclear receptor superfamily. *Essays Biochem*. 2004;**40**:11–26.
15. Amero SA, Kretsinger RH, Moncrief ND, Yamamoto KR, Pearson WR. The origin of nuclear receptor proteins: a single precursor distinct from other transcription factors. *Mol Endocrinol*. 1992;**6**(1):3–7.
16. Laudet V, Hänni C, Coll J, Catzeflis F, Stéhelin D. Evolution of the nuclear receptor gene superfamily. *EMBO J*. 1992;**11**(3):1003–1013.
17. Escrivá García H, Laudet V, Robinson-Rechavi M. Nuclear receptors are markers of animal genome evolution. *J Struct Funct Genomics*. 2003;**3**(1–4):177–184.
18. Thornton JW, DeSalle R. A new method to localize and test the significance of incongruence: detecting domain shuffling in the nuclear receptor superfamily. *Syst Biol*. 2000;**49**(2):183–201.
19. Bridgham JT, Eick GN, Larroux C, Deshpande K, Harms MJ, Gauthier ME, Ortlund EA, Degnan BM, Thornton JW. Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol*. 2010;**8**(10):e1000497.
20. Zhang Z, Burch PE, Cooney AJ, Lanz RB, Pereira FA, Wu J, Gibbs RA, Weinstock G, Wheeler DA. Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome. *Genome Res*. 2004;**14**(4):580–590.

21. Krasowski MD, Yasuda K, Hagey LR, Schuetz EG. Evolutionary selection across the nuclear hormone receptor superfamily with a focus on the NR1I subfamily (vitamin D, pregnane X, and constitutive androstane receptors). *Nucl Recept*. 2005;**3**(1):2.
22. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*. 1999;**22**(3):231–238.
23. Miller RD, Phillips MS, Jo I, Donaldson MA, Studebaker JF, Addleman N, Alfisi SV, Ankener WM, Bhatti HA, Callahan CE, Carey BJ, Conley CL, Cyr JM, Derohannessian V, Donaldson RA, Elosua C, Ford SE, Forman AM, Gelfand CA, Grecco NM, Gutendorf SM, Hock CR, Hozza MJ, Hur S, In SM, Jackson DL, Jo SA, Jung SC, Kim S, Kimm K, Kloss EF, Koboldt DC, Kuebler JM, Kuo FS, Lathrop JA, Lee JK, Leis KL, Livingston SA, Lovins EG, Lundy ML, Maggan S, Minton M, Mockler MA, Morris DW, Nachtman EP, Oh B, Park C, Park CW, Pavelka N, Perkins AB, Restine SL, Sachidanandam R, Reinhart AJ, Scott KE, Shah GJ, Tate JM, Varde SA, Walters A, White JR, Yoo YK, Lee JE, Boyce-Jacino MT, Kwok PY; SNP Consortium Allele Frequency Project. High-density single-nucleotide polymorphism maps of the human genome. *Genomics*. 2005;**86**(2):117–126.
24. Savas S, Tuzmen S, Ozcelik H. Human SNPs resulting in premature stop codons and protein truncation. *Hum Genomics*. 2006;**2**(5):274–286.
25. Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin*. 2015;**8**(1):57.
26. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet*. 1992;**90**(1–2):41–54.
27. Li G, Pan T, Guo D, Li LC. Regulatory variants and disease: the E-cadherin–160C/A SNP as an example. *Mol Biol Int*. 2014;**2014**:967565.
28. Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM. The sounds of silence: synonymous mutations affect function. *Pharmacogenomics*. 2007;**8**(6):527–532.
29. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MMA. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*. 2007;**315**(5811): 525–528.
30. van Rossum EF, Russcher H, Lamberts SW. Genetic polymorphisms and multifactorial diseases: facts and fallacies revealed by the glucocorticoid receptor gene. *Trends Endocrinol Metab*. 2005;**16**(10): 445–450.
31. Nuclear Receptors Nomenclature Committee. A unified nomenclature system for the nuclear receptor superfamily. *Cell*. 1999;**97**(2):161–163.
32. Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol*. 2010;**8**(10):e1000494.
33. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F. Tissue-based map of the human proteome. *Science*. 2015;**347**(6220):1260419.
34. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, Asplund A, Sjöstedt E, Lundberg E, Szigyarto CA, Skogs M, Takanen JO, Berling H, Tegel H, Mulder J, Nilsson P, Schwenk JM, Lindskog C, Danielsson F, Mardinoglu A, Sivertsson A, von Feilitzen K, Forsberg M, Zwahlen M, Olsson I, Navani S, Huss M, Nielsen J, Ponten F, Uhlén M. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics*. 2013;**13**(2):397–406.
35. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One*. 2012;**7**(2): e31826.
36. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*. 2013;**9**(8):e1003709.
37. Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, Mahowald M, Wyckoff GJ, Malcom CM, Lahn BT. Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell*. 2004; **119**(7):1027–1040.
38. Benoit G, Malewicz M, Perlmann T. Digging deep into the pockets of orphan nuclear receptors: insights from structural studies. *Trends Cell Biol*. 2004;**14**(7):369–376.

39. Timsit YE, Negishi M. CAR and PXR: the xenobiotic-sensing receptors. *Steroids*. 2007;**72**(3):231–246.
40. Gustafsson JA. Estrogen receptor β—a new dimension in estrogen mechanism of action. *J Endocrinol*. 1999;**163**(3):379–383.
41. Meyer D, Aguiar C, Bitarello BD, C Brandt DY, Nunes K. A genomic perspective on HLA evolution. *Immunogenetics*. 2017, https://doi.org/10.1007/s00251-017-1017-3.
42. Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, Noé F, Freund C. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Front Immunol*. 2017;**8**:292.
43. Chanda D, Park JH, Choi HS. Molecular basis of endocrine regulation by orphan nuclear receptor small heterodimer partner. *Endocr J*. 2008;**55**(2):253–268.
44. Amoutzias GD, Pichler EE, Mian N, De Graaf D, Imsiridou A, Robinson-Rechavi M, Bornberg-Bauer E, Robertson DL, Oliver SG. A protein interaction atlas for the nuclear receptors: properties and quality of a hub-based dimerisation network. *BMC Syst Biol*. 2007;**1**(1):34.
45. Zhi X, Zhou XE, He Y, Zechner C, Suino-Powell KM, Kliewer SA, Melcher K, Mangelsdorf DJ, Xu HE. Structural insights into gene repression by the orphan nuclear receptor SHP. *Proc Natl Acad Sci USA*. 2013;**111**(2):839–844.
46. Miao J, Choi SE, Seok SM, Yang L, Zuercher WJ, Xu Y, Willson TM, Xu HE, Kemper JK. Ligand-dependent regulation of the activity of the orphan nuclear receptor, small heterodimer partner (SHP), in the repression of bile acid biosynthetic CYP7A1 and CYP8B1 genes. *Mol Endocrinol*. 2011;**25**(7):1159–1169.
47. McCabe ER. DAX1: increasing complexity in the roles of this novel nuclear receptor. *Mol Cell Endocrinol*. 2007;**265-266**:179–182.
48. Germain P, Staels B, Dacquet C, Spedding M, Laudet V. Overview of nomenclature of nuclear receptors. *Pharmacol Rev*. 2006;**58**(4):685–704.
49. Charmandari E, Kino T, Ichijo T, Chrousos GP. Generalized glucocorticoid resistance: clinical aspects, molecular mechanisms, and implications of a rare genetic disorder. *J Clin Endocrinol Metab*. 2008;**93**(5):1563–1572.
50. Onigata K, Szinnai G. Resistance to thyroid hormone. *Endocr Dev*. 2014;**26**:118–129.
51. Herynk MH, Fuqua SA. Estrogen receptor mutations in human disease. *Endocr Rev*. 2004;**25**(6):869–898.
52. Hay C, Wu F. Genetics and hypogonadotrophic hypogonadism. *Curr Opin Obstet Gynecol*. 2002;**14**(3):303–308.
53. Popadin KY, Gutierrez-Arcelus M, Lappalainen T, Buil A, Steinberg J, Nikolaev SI, Lukowski SW, Bazykin GA, Seplyarskiy VB, Ioannidis P, Zdobnov EM, Dermitzakis ET, Antonarakis SE. Gene age predicts the strength of purifying selection acting on gene expression variation in humans. *Am J Hum Genet*. 2014;**95**(6):660–674.