Positive Selection Drives the Rapid Fixation of Dephosphorylation in IRF9

Jianhai Chen (jianhaichen@scu.edu.cn)
West China Hospital, Sichuan university
Xuefei He
West China Hospital, Sichuan university
Ivan Jakovlić
State Key Laboratory of Grassland Agro-Ecosystems, and College of Ecology, Lanzhou University, 730000, Lanzhou, China

Research Article

Keywords: Protein phosphorylation, positive selection, purifying selection, branch-site model, branch model, IRF9

Posted Date: April 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1556042/v1

Abstract

The arms race between humans and pathogens drives the evolution of the human genome. Here, based on the HPO (Human Phenotype Ontology) annotation, we focused on human genes related to recurrent infections of virus, (RVI), bacterial (RBI) and recurrent fungal infections (RFI) to understand positive selection on pathogen-responsive genes. Interestingly, cross-species positive selection analyses revealed that the proportion of unique genes under positive selection was higher for RVI (78.57%) than for RFI (57.14%) and RBI (58.68%). Based on results of the branch-site test, we further focused on the amino acid site Val129 of *IRF9*, which has a significant signal of positive selection based on multiple evidence. Interestingly, this novel and derived amino acid (V) has been rapidly fixed before the "out-of-Africa" event ~ 500,000 years ago from the ancestral state S, which is conserved among 88.5% of mammalian lineages. Phosphorylation analysis revealed that the conserved ancestral S may serve as the phosphorylation site of *IRF9*. Further analyses suggested that the rapid dephosphorylation of *IRF9* via the change of S to V may have conferred potential molecular adaptations by boosting and extending the immune activity of *IRF9*. This study provides an interesting mode in which strong positive Darwinian selection drives the rapid fixation of a hominin specific amino acid leading to molecular adaptation for immune response.

Introduction

How positive Darwinian selection has shaped the evolution of immune-related genes has long been spotlighted. The never-ending arms race between pathogens, such as viruses, bacteria, and fungi, and the immune-related genes is one of the most long-lasting forces to shape the infection susceptibility of humans [1, 2]. A genomic scanning of primate species revealed that the most enriched pathways with positively selected genes in the human genome are related to the immune system [3]. It was estimated that over 30% of adaptive protein changes in the human genome were driven by virus-human interactions over millions of years [4]. A study incorporating data from birds and mammals suggested a general pattern of positive selection on immune-related genes [5]. Despite growing evidence of immune-related positive selection, it is still unclear which pathogen, viruses, bacteria, or fungi, has higher fraction of respective human genes under positive selection.

Susceptibility of pathogens infection and related genes have been summarized in the HPO (Human Phenotype Ontology, https://hpo.jax.org), a project of the Monarch Initiative in the Global Alliance for Genomics and Health (GA4GH) strategic roadmap [6, 7]. Specifically, in HPO, the most relevant phenotype for infection susceptibility is "recurrent infections (RI)" (HP:0002719). This phenotype comprises three child phenotypes: "recurrent bacterial infections" (RBI, HP:0002718), "recurrent fungal infections" (RFI, HP:0002841) and "recurrent viral infections (RVI, HP:0004429)". As the names suggest, these phenotypes are defined as increased susceptibility to bacterial, fungal, and viral infections, respectively, manifested by recurrent episodes of infection. For the objectives of this study, RVI is particularly interesting because it suggests a possibility of long-term host-virus interactions.

To understand how frequently the HPO registered RVI genes may evolve under positive selection in the evolutionary past, we conducted selection analyses across mammalian species for RVI genes and compared them to other pathogens-related genes and background levels (autoimmunity genes, and genome-wide background genes). We revealed that RVI has higher fraction of unique genes under positive selection (f_S) than RBI, RFI, and background groups. We uncovered evidence of molecular adaptation based on signals of positive selection on a specific site Val129 of *IRF9* (Interferon Regulatory Factor 9). Interesting, Val129 is a novel substitution from ancestral amino acid serine (“S”), appeared since the common ancestor of humans, Neanderthals, and Denisovans at ~ 500,000 years ago. Further analysis revealed that this change may result in the loss of phosphorylation. We proposed that advantages underlying the molecular adaptation of this rapid dephosphorylation may correlate with the elevation and extension of the immune activity of *IRF9*.

Materials And Methods

Datasets: recurrent viral infection genes and orthologs

Before filtering, we obtained 62 genes related to "recurrent viral infections" (HP:0004429), 178 related to "recurrent bacterial infections" (RBI, HP:0002718), and 62 related to "recurrent fungal infections" (RFI, HP:0002841) from the Human Phenotype Ontology database (HPO, https://hpo.jax.org/, NOV-2021) [6, 7]. By comparing the percentages of positively selected genes in RVI, RFI, and RBI datasets, we aimed to understand to what extend viruses may differ from bacteria and fungi in triggering the natural selection on pathogen defense-related genes.

To better understand the trajectories of these genes during mammalian evolution, we further compared them to an orthologous dataset from multiple mammalian species. While increasing the sample size can increase the power of detecting positive selection [31], the larger number of taxa would also increase the computational burden, and even more importantly reduce the effective length of protein alignments, thus decreasing the numbers of sites that can be analyzed. To balance the number of species and sites, we firstly conducted selection analysis using a dataset comprising 20 mammalian species, covering phylogenetic clades of Primates, Euarchontoglires, Boreoeutheria, Eutheria, and Theria. The following species were included in the dataset: human, chimpanzee, bonobo, gorilla, gibbon, macaque, golden snub-nose monkey, Ma's night monkey, marmoset, tarsier, mouse lemur, mouse, rat, rabbit, dog, horse, cow, megabat, elephant, and opossum. According to the above-described human genome datasets (RVI, RBI, RFI, autoimmunity, and genome-wide; Supplementary table 1 and 2), we further extracted orthologous genes of the remaining 19 mammalian species from the Ensembl database Release 105 [32]. We conducted analyses using only on "1 to 1" orthologs identified with high confidence based on Ensembl and protein-level identity over 60%. We further limited these orthologous genes to genes with conserved gene synteny where at least one neighboring gene also exhibited a conserved gene order in other mammals. The genes with fewer than 7 orthologs (out of 20) were removed.

After pruning, we obtained 58 genes in the RVI dataset, 167 in the RBI dataset, and 60 in the RFI dataset (Supplementary table 1). As further background data to conduct comparisons (controls), we retrieved two more categories from the same database: 145 genes associated with autoimmune diseases ("autoimmunity", HP:0002960) and all 16,625 genome-wide protein-coding genes (Supplementary table 2). Subsequently, for our focus phenotype, RVI, we expanded the taxonomic dataset to up to 46 taxa to examine whether the observations inferred using the more limited dataset are stable (Supplementary table 3).

Selective pressure analyses based on the branch model and branch-site model

We used the above datasets to conduct two types of likelihood-ratio tests: branch model and branch-site model. The CODEML package of PAML 4.9 [33] was utilized for the maximum likelihood analysis of dN/dS ratios (ω). For the branch model of each gene, we estimated dN/dS ratios for different lineages using the free-ratio model. The significance of the free-ratio model was estimated by comparing the statistical parameters to those of the one-ratio model. The free-ratio model allows evolutionary rate variations across species, while the one-ratio model assumes no difference in selective pressure among species. We compared different phenotypes based on the fraction of positively selected genes for a given phenotype (f_s), following the formula of $f_s = \frac{s \cap p}{T}$, where s represents the number of genes with statistical significance (χ^2 test, $p < 0.05$), p indicates the number of genes with at least one dN/dS ratio > 1, and T is the total number of genes related to a phenotype.

The branch-site model in PAML was used for detecting positive selection on both branches and sites [34]. The null hypothesis was set by "fix_omega = 1" and "omega = 1". The statistical significance was computed by a chi-square distribution, with two times the difference in log-likelihood values and degree of freedom as the difference in the number of parameters for the two models. Moreover, for the significantly selected genes and sites of interest, we further confirmed their selection using different tools in the HyPhy package [35, 36]: BUSTED (branch-site unrestricted statistical test for episodic diversification) method allows ω to vary from branch to branch and provides descriptive information about the potential selection status for a given site [37]; the MEME (mixed effects model of evolution) method aims to detect individual sites under episodic positive selection or diversifying selection [38]; the FEL (fixed effects likelihood) method can be used to test which sites in a gene may be associated with adaptation to a different environment [39]; the aBSREL (adaptive branch-site random effects likelihood) method is an improved version of "branch-site" models, which models both site- and branch-heterogeneity, though it does not test for selection at specific sites [40].

Protein Structure and Motifs Prediction

Positively selected genes from the previous step were used to conduct further structural analyses. Differences between the secondary and 3-dimensional structures of human and rat proteins were predicted using PSIPRED 4.0 [41] and AlphaFold2 protein structure database [42, 43], respectively.

SCANSITE 4.0 was used to predict the specific sites of kinase phosphorylation and binding domains using 81 mammalian kinases/domains as the background database. The result was further filtered with stringency set as "high". The linker region and domains were then visualized manually.

Population allele frequency, gene expression, and ancestral state reconstruction of positively selected sites

The population allele frequency can be used to understand whether a specific amino acid mutation has been fixed in human populations, which is important for evolutionary and medical studies [44]. We conducted this analysis using Genome Aggregation Database (gnomAD) 3.0 (https://gnomad.broadinstitute.org), which has curated allele frequencies for human variants [45]. We further checked whether the positively selected genes are expressed in the immune-related tissues or cells using GTEx [46]. To trace the origin, or ancestral state, of positively selected sites, we first examined whether the sites are divergently fixed between humans and outgroups. If substitutions diverged between humans and outgroup species, we distinguished the ancestral and derived states based on phylogenetic distribution: the ancestral sites are expected to have a wider distribution across ancestral phylogenetic nodes. This intuitive method was further reexamined with the Maximum likelihood (ML) approaches in both PAML [12] and MEGA [13]. Finally, we checked the origin branch of derived novel substitutions using the UCSC genome browser by focusing on the comparative mapped reads between the Denisovan, Neanderthal and human genomes.

Results

Genes under positive selection estimated using the branch model

The branch model was designed to detect branches or lineages with faster evolution [8]. Under a likelihood-ratio test framework, this methodology has been extensively used in screening positively selected genes in a specific lineage [9]. After filtering out genes without Ensembl-annotated orthologs, a total of 58 RVI genes were used for selection analyses. We compared the HPO-annotated responsive genes of recurrent viral, fungal, and bacterial infections (RVI, RFI and RBI respectively), which are phenotypes involving different classes of pathogens. The free-ratio model revealed that 77.59% of genes (45/58) showed signals of positive selection in at least one lineage (Supplementary table 1a and 1b). When we compared the free-ratio model with the one-ratio model, which assumes the absence of inter-lineage differences in the evolutionary rate, we found evidence that 70.69% of genes (41/58, χ^2 test, $p < 0.05$) evolved at a rate that differed from the assumptions of neutral evolution. Among these, 55.17% of genes (32/58) were identified as positively selected genes. Notably, genes were identified as positively selected only if they conformed to two parameters simultaneously: positive selection identified with statistical significance (χ^2 test, $p < 0.05$) and positive selection in at least one evolutionary lineage ($\omega > 1$) (Supplementary Table 3 and Fig. 1a).

To make the fractions of positively selected genes (f_S) of RVI comparable among the HPO-defined phenotypes, we firstly compared RVI with RBI and RFI, which are also phenotypes involved in pathogen-responses. Following the filtering of shared genes among the three phenotypic categories (RVI, RFI and RBI), we then focused on unique genes for each category. We found higher f_S for unique genes of RVI (78.57%, 11/14) than for unique genes of RFI (57.14%, 12/21) and RBI (58.68%, 71/121) (Supplementary Table 1b, 1c and 1d). We further used the genes related to the "autoimmunity" category as a (proxy of) control, because autoimmunity is also within the scope of the immune system but not necessarily related to pathogens. When we focused on unique genes between the autoimmunity and RVI, the f_S of the autoimmunity category (55.00%, 66/120) was also lower than that of the RVI (66.67%, 22/33) (Supplementary table 1e). Interestingly, on average, the background f_S level of genome-wide protein-coding genes was 49.43% (8218/16625, Supplementary Table 2), which is also lower than that of RVI. These comparisons tentatively suggest that the probability of historical recurrent positive selection on genes responsive for antiviral responses is higher than the average genome-level positive selection, the non-pathogen immune phenotype (autoimmunity), and other pathogen-related immune phenotypes (bacteria and fungi).

To understand the long-term evolution of these genes, we mapped the genes with signals of significant positive selection (dN/dS > 1) onto the mammalian phylogenetic tree. All branches were decomposed into 12 phylostra or phylostratigraphic stages (PS), also the ancestral or outgroup branches. The genes with dN/dS ratios significantly larger than 1 in at least one branch were assigned into these stages. Interestingly, after counting the number of positively selected genes distributed across species, we found that the Old-World monkey species have the highest number of genes (19) with a positive selection signal (Fig. 1 and Supplementary Table 4, χ^2 test, $p < 0.05$). Among all primates, 26 genes had significant positive selection signals (Supplementary table 5, χ^2 test, $p < 0.05$). The Laurasiatherian species (cattle, cat, camel, dog, horse, pig, and yak) also demonstrated a higher number of genes (14) under positive selection than other remaining mammalian clades including bats and rodents (Fig. 1 and Supplementary Table 6).

The branch-site model revealed two sites of IRF9 are under a significant positive selection

To increase the resolution of positive selection analyses, we tried to understand whether the above 32 genes under positive selection have some specific sites under significant positive selection. To identify positively selected sites in the human lineage, we predefined the human branch as the foreground in the branch-site model analysis. We identified a significant signal of one gene, *IRF9*, with two sites under positive selection (χ^2 test, $p = 0.01036$). The probabilities of positive selection on the two sites Val129 and Val333 were 0.992 and 0.652, respectively, based on the Naive Empirical Bayes (NEB) analysis (Table 1).

To assess whether the selection signals identified in *IRF9* are robust, we attempted to corroborate these results using different tools in the HyPhy package. The aBSREL method found evidence of episodic diversifying selection on the human branch (p = 0.0114). For the likelihood test on sites, due to a coordinate change caused by an alignment gap, the site coordinate of Val333 in the branch-site model of

PAML was changed into site 340. The MEME method confirmed that Val129 and Val333 (i.e., 340) are positively selected sites (LRT, $p < 0.05$). The BUSTED method with synonymous rate variation found evidence (LRT, $p = 0.027$) of episodic diversifying selection acting upon the *IRF9* in the human lineage (selected as the foreground branch). Therefore, there is evidence that at least one site on the human branch has experienced diversifying selection (Fig. 2a). The FEL method also supported the positive selection on two sites Val129 and Val333 (i.e., 340) with statistical significance (Fig. 2b and Table 2, LRT $p < 0.01$). All these analyses based on different algorithms strongly suggest that *IRF9* and its two sites Val129 and Val333 are under positive selection in the human lineage, though Val129 has stronger signal than Val 333.

We further assessed whether these two sites were conserved in non-human species (Fig. 2c and 2d). Among 26 species, site 129 had a conserved "S" amino acid in 23 species. This suggests that the ancestral state of this site is "S", and that it is highly conserved during the long-term evolution history in mammals (23/26 = 88.5%). In this light, only three species exhibited derived states (amino acid mutations) in this site: humans, megabat, and orangutan. The derived state in the human genome is "V", whereas in megabat is L, and orangutan P. As for site 333, 23/26 species had the "A" amino acid. Only two Old-world monkeys and humans had the amino acid "V". This suggests that "V" is the derived state, whereas "A" is the ancestral state. The ancestral states of these two sites were also confirmed with the Maximum likelihood (ML) inference based on packages of PAML [12] and MEGA [13] (Supplementary Fig. 1).

Table 1
The two sites of the human *IRF9* detected to be under positive selection by the Naive Empirical Bayes (NEB) analysis.

Codon	Amino acid	Location in protein	Location in genome
GTA	V	129	chr14:24163398–24163400
GTC	V	333	chr14:24165852–24165854

Table 2
The sites of *IRF9* detected to be under positive selection by the FEL algorithm. Note: Site 340 corresponds to the site 333 in the branch-site model of PAML due to coordinate shift caused by alignment gap. α and β denote synonymous and non-synonymous rates

Codon	α	β	α = β	LRT	p value	class
129	1.148	383.28	2.201	16.854	0	Diversifying
340	0	116.021	0.649	10.083	0.0015	Diversifying
119	1.422	86.249	2.345	5.483	0.0192	Diversifying
377	10000	0	53.61	4.645	0.0312	Purifying

The structure of IRF9 **in humans and rats**

To examine whether the amino acid substitutions may have caused a change in the structure, we compared both secondary and tertiary structures between human and rat orthologues. Based on the prediction of an online tool PSIPRED and the machine learning software Alphafold2, we revealed that, despite the above-mentioned changes in the conserved sites, the secondary and tertiary structures did not change between the human and rat orthologues (Fig. 3). The site Val129 in humans and its orthologous site Ser129 in rats were both located in the coil region. In addition, the site Val333 in humans and its orthologous site in rats also shared the structure of the helix.

The secondary structural similarities of the IRFs gene family

The gene family of *IRFs* has been investigated extensively. For example, *IRF3* is one of the most well-characterized transcription factors involving innate immune responses [14]. The IAD domain was reported to be conserved in all *IRFs*, except *IRF1* and *IRF2* [15]. Here, we tried to understand the structural differences between the homologous proteins within the *IRFs* family with a focus on positively selected sites in *IRF9.* The homologous alignment revealed that common features of *IRF* structures comprised DBD (DNA Binding Region), IAD (IRF Associated Domain), and the linker region (Fig. 4a). Alignment of homologous genes revealed that Val129 can be aligned to the amino acid "S" within *IRF3* (Fig. 4b). DBD can bind to its interferon-stimulated response element (ISRE), while IAD is responsible for binding with the signal transducer and activator of transcription 2 (STAT2) [16, 17]. The human Val129 lies within the linker region between DBD and IAD, while the human Val333 is located within the IAD region (Fig. 5a). The alignment similarities for these homologous proteins were higher in DBD and IAD regions than in the linker region (Fig. 5b), suggesting that the linker region, comprised of the random coil, might be the hotspot for positive selection due to its comparative abundance in amino acid differences among homologous genes.

The ancestral state "S" of Val129 might be phosphorylation site

To understand the functional effects of evolutionary changes from "S" to "V" in the alignment site of 129 in human gene *IRF9*, we conducted the motif prediction analysis using SCANSITE 4.0. The Val129 from the human reference sequence ("PGTQK[V]PSKRQ", Val129 is emphasized with brackets) was not predicted to be a phosphorylation site. Interestingly, however, after we manually changed the human Val129 into its ancestral state Ser129 (hereafter referred to as V129S), it was identified as a phosphorylation site (Table 3). The potential kinase was identified as the Cyclin-dependent kinase 1 (CDK1). The ancestral core sequence around the human 129 site ("PGTQK[S]PSKRQ", S129 is emphasized with brackets) conforms to the requirement of the CDK1 consensus phosphorylation site (Ser/Thr-Pro-X-Lys/Arg) [18]. Since the human V129 site is very conserved across mammals, and only three outgroup species (rat, megabat, and orangutan) showed different amino acids (Fig. 2c), we selected these three species to predict potential kinases (Table 3). Interestingly, in rats, where the amino acid in this site corresponds to the ancestral state "S" of human *IRF9* Val129, S129 was also predicted to be a phosphorylation site. This further indirectly supported the pseudo-mutant prediction in humans.

Table 3
The prediction of phosphorylation site and potential kinases using SCANSITE 4.0 across three outgroup species with different amino acids around the orthologous region of human Val129 (see Fig. 2). The amino acids in different species that are highlighted with brackets correspond to the positively selected site of human Val129. The human pseudo-mutant to ancestral state is named "Human_V129S". The other identified phosphorylation sites are underlined.

Species	*IRF9* Site	Sequence	Potential kinases
Human_V129 (reference)	V129	PGTQK[V]PSKRQ	No
	T348	PKFQVTLNFWE	NEK1/NEK3
Human_V129S (ancestral)	S129	PGTQK[S]PSKRQ	CDK1/CDK5
	T348	PKFQVTLNFWE	NEK1/NEK3
Rat	S129	PRNQK[S]PEKRS	CDK1/CDK5
	T215	YSLLLTFIYGG	NEK5
	T354	PSSSHTQENLI	PRKDC
Megabat	P307	ISWSAPQAPPG	ABL1
	T352	PKFQVTLNFWE	NEK1/NEK3
Orangutan	T347	PKFQVTLNFWE	NEK1/NEK3

The evolutionary origin and population allele frequency of Val129

Based on the human population RNAseq data (GTEX), we found that the top two highest expression organs/tissues/cells of *IRF9* are the spleen and EBV-transformed lymphocytes (Supplementary Fig. 2). Interestingly, based on the GTEx database, *IRF9* is expressed higher than other *IRF* genes in the spleen. To further understand whether there are polymorphisms at the site of Val129 in human populations, we examined the gnomAD Database, which has curated allele frequency data for 76,156 individuals. We found that there is no "S" type in any human population. This result suggests that the Val129 should have been fixed before the "out-of-Africa" event. To identify the evolutionary stage when the change from "S" to "V" occurred, we manually assessed the mapping results of Denisovan and Neanderthal genomic reads publicly available in the UCSC browser against the human genome. We found that Val129 was shared among all three *Homo* species (Supplementary Fig. 3), which suggests that the fixation of this mutation dates at least 500,000 years ago.

Discussion

To understand how natural selection has shaped the evolution of genes involved in pathogen responses, we focused on genes related to "recurrent viral/bacterial/fungal infects" from the HPO database. Intriguingly, RVI had a higher proportion of unique genes with signals of positive selection than the

presumably similar RBI and RFI categories, as well as the background categories: autoimmunity and genome-wide level. This suggests that RVI had a disproportionately strong influence on shaping the mammalian immune genes.

In humans, we found that at least two sites in *IRF9* are evolving under a positive Darwinian selection. The signals were confirmed with several methods, supporting the reliability of our results. We found a particularly interesting pattern for the site Val129. Evolutionary alignment across species and ancestral sequence reconstruction support "S" as the ancestral state of Val129, suggesting the evolutionary adaptation for the establishment of "V" and the loss of "S". The natural selection was further supported by the rapid fixation of "V" within the last 500,000 years since the common ancestor of humans, Neanderthals and Denisovans. Though evolutionarily conserved, the "S" site is located within the random coil region of *IRF9* (supported by both secondary and tertiary structure predictions). A study on *Drosophila* has revealed that random coil structure is the preferred hotspot for positive selection [19]. Our analyses support their findings and indicate that mutations in the random coil region of IRF proteins may contribute to the rapid evolution of human immune-responsive genes.

IRF9, the interferon regulatory factor 9, is a transcription factor critical for mediating the type I interferon antiviral immunity. It is associated with the human disease phenotype that was summarized as the Immunodeficiency 65 Viral Infections. Specifically, the inherited *IRF9* deficiency is related to life-threatening influenza pneumonitis in early infancy [20] and impaired control of multiple viral infections [21]. We summarized the *IRF9*-related process inferred via a literature review in Fig. 6. Briefly, the canonical type I IFN signaling initiates from the binding of type I IFNs (IFN-alpha and IFN-beta) to two types of cell surface transmembrane receptors: IFNAR1 and IFNAR2. This binding results in the activation of two cytoplasmic Jak kinases, TYK2 and JAK1, to further phosphorylate two transcription factors, STAT1 and STAT2. The ISGF3, a transcriptional activator with a multi-subunit structure, is then formed via an interaction between the phosphorylated STAT1:STAT2 dimer and IRF9. IRF9 can facilitate the DNA binding activity of the ISGF3 complex to stabilize the complex with the aid of STAT1. The transcriptionally active ISGF3 then enters the nucleus and directly binds to the promoter regions of IFN-stimulated response elements (ISRE) [22]. This binding process can then activate the transcription of interferon-stimulated genes (ISGs) [23], thereby triggering the full-blown antiviral response in the cell [24, 25] (Fig. 6). The type I IFN signaling can effectively stimulate the transcription of IRF genes themselves (*IRF1* to *IRF9*), thereby serving as a positive feedback-amplifier circuit [23].

Herein, based on the motif prediction, we found that the ancestral state "S" is within a linker region and that it is a conserved putative phosphorylation site across multiple species. Interestingly, the alignment between *IRF3* and *IRF9* revealed that the corresponding amino acid in *IRF3* is also "S" (S123). Indeed, in Eukaryotes, serine (S), threonine (T), and tyrosine (Y) residues are the most used phosphorylation sites [26–28]. Through the linker region phosphorylation, the EBV-encoded kinase, *BGLF4*, can down-regulate the *IRF3* transactivation [26]. In contrast, a phosphorylation-defect mutant, *IRF3*_S123A, showed higher activity [26]. In addition, Glycogen synthase kinase 3 (*GSK3*) inhibitor can enhance the activity of *IRF3* [26]. Recent studies revealed that unphosphorylated STAT2 (U-STAT2), IRF9, and U-STAT1 can form

unphosphorylated ISGF3, which can prolong and sustain the resistance to virus infection and DNA damage [23]. Thus, we propose that the evolutionary change from "S" to "V", which corresponds to the loss of the phosphorylation site, may have resulted in the molecular adaptation of transactivation activity in *IRF9*.

Protein phosphorylation is an evolutionarily ancient process in all living organisms ranging from prokaryotes to eukaryotes [29]. Despite the deep conservation of phosphoproteins, the phosphorylation site can be under rapid evolution and divergence to enable phenotypic plasticity and diversity [30]. Protein phosphorylation is indeed a rapid and versatile mechanism to drive signal tuning and protein regulation. The novel amino acid "V" from the ancestral "S" in *IRF9* was rapidly fixed during the hominin evolution due to positive selection. The potential advantage of the novel form could be linked to the loss of phosphorylation, which may enhance and prolong the transactivation activity during the anti-viral immune response. Therefore, this study provides an elegant case of positive Darwinian selection underlying the rapid fixation of a specific amino acid leading to molecular adaptation for immune response.

Conclusions

We found that the RVI category had a higher fraction (f_S) of unique genes with significant signals of positive selection in at least one mammalian lineage (78.57%) than RBI (58.68%) and RFI (57.14%). This fraction was also higher than in autoimmunity genes (55.00%) and genome-wide genes (49.43%). Interestingly, among positively selected RVI genes, we identified *IRF9* with a significant signal of positive selection on the amino acid change from S to V, which may involve the process of dephosphorylation. The novel *IRF9* amino acid (Val129) may have been adaptive in humans due to its enhanced and prolonged transactivation activity.

Declarations

Acknowledgements

This study was supported by the fifth batch of technological innovation research projects in Chengdu (2021-YF05 -01331-SN) and Postdoctoral Research and Development Fund of West China Hospital of Sichuan University (2020HXBH087).

Author contributions

J.H.C. designed the research. J.H.C. and X.F.H. analyzed data. J.H.C. and I.J. wrote the manuscript.

Competing interests

The authors declare no competing financial interests.

Availability of data and materials

PAML computing files for branch and branch-site models:

The datasets generated and/or analysed during the current study are available in the datadryad repository, https://datadryad.org/stash/share/E37GVHBsNIMl_ksTIOTk-Sf5DVgVWqZvhHJ7sD83oBE

References

1. Savoy SKA, Boudreau JE: **The Evolutionary Arms Race between Virus and NK Cells: Diversity Enables Population-Level Virus Control**. Viruses 2019, **11**(10):959.
2. Barreiro LB, Quintana-Murci L: **From evolutionary genetics to human immunology: how selection shapes host defence genes**. Nature Reviews Genetics 2010, **11**(1):17–30.
3. Zhao S, Zhang T, Liu Q, Wu H, Su B, Shi P, Chen H: **Identifying Lineage-Specific Targets of Natural Selection by a Bayesian Analysis of Genomic Polymorphisms and Divergence from Multiple Species**. Molecular biology and evolution 2019, **36**(6):1302–1315.
4. Enard D, Cai L, Gwennap C, Petrov DA: **Viruses are a dominant driver of protein adaptation in mammals**. Elife 2016, **5**:e12469.
5. Shultz AJ, Sackton TB: **Immune genes are hotspots of shared positive selection across birds and mammals**. eLife 2019, **8**:e41815.
6. Kohler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, Gargano M, Harris NL, Matentzoglu N, McMurry JA *et al*: **Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources**. Nucleic Acids Res 2019, **47**(D1):D1018-D1027.
7. Kohler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, Danis D, Balagura G, Baynam G, Brower AM *et al*: **The Human Phenotype Ontology in 2021**. Nucleic Acids Res 2021, **49**(D1):D1207-D1217.
8. Yang Z: **Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution**. Molecular biology and evolution 1998, **15**(5):568–573.
9. Chen J, Ni P, Li X, Han J, Jakovlić I, Zhang C, Zhao S: **Population size may shape the accumulation of functional mutations following domestication**. BMC Evolutionary Biology 2018, **18**(1):4.
10. Jin G, Ma P-F, Wu X, Gu L, Long M, Zhang C, Li D-Z: **New Genes Interacted With Recent Whole-Genome Duplicates in the Fast Stem Growth of Bamboos**. Molecular Biology and Evolution 2021, **38**(12):5752–5768.
11. Domazet-Lošo T, Tautz D: **An Ancient Evolutionary Origin of Genes Associated with Human Genetic Diseases**. Molecular biology and evolution 2008, **25**:2699–2707.
12. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. Molecular biology and evolution 2007, **24**(8):1586–1591.

13. Kumar S, Tamura K, Nei M: **MEGA: molecular evolutionary genetics analysis software for microcomputers**. Bioinformatics 1994, **10**(2):189–191.
14. Yanai H, Chiba S, Hangai S, Kometani K, Inoue A, Kimura Y, Abe T, Kiyonari H, Nishio J, Taguchi-Atarashi N: **Revisiting the role of IRF3 in inflammation and immunity by conditional and specifically targeted gene ablation in mice**. *Proceedings of the National Academy of Sciences* 2018, **115**(20):5253–5258.
15. Behr M, Schieferdecker K, Bühr P, Büter M, Petsophonsakul W, Sirirungsi W, Redmann-Müller I, Müller U, Prempracha N, Jungwirth C: **Interferon-stimulated response element (ISRE)-binding protein complex DRAF1 is activated in Sindbis virus (HR)-infected cells**. Journal of Interferon & Cytokine Research 2001, **21**(11):981–990.
16. Rengachari S, Groiss S, Devos JM, Caron E, Grandvaux N, Panne D: **Structural basis of STAT2 recognition by IRF9 reveals molecular insights into ISGF3 function**. *Proceedings of the National Academy of Sciences* 2018, **115**(4):E601-E609.
17. Veals SA, Santa Maria T, Levy DE: **Two domains of ISGF3 gamma that mediate protein-DNA and protein-protein interactions during transcription factor assembly contribute to DNA-binding specificity**. Molecular and cellular biology 1993, **13**(1):196–206.
18. Au - Cui H, Au - Loftus KM, Au - Noell CR, Au - Solmaz SR: **Identification of Cyclin-dependent Kinase 1 Specific Phosphorylation Sites by an In Vitro Kinase Assay**. JoVE 2018(135):e57674.
19. Ridout KE, Dixon CJ, Filatov DA: **Positive selection differs between protein secondary structure elements in Drosophila**. Genome biology and evolution 2010, **2**:166–179.
20. Hernandez N, Melki I, Jing H, Habib T, Huang SS, Danielson J, Kula T, Drutman S, Belkaya S, Rattina V: **Life-threatening influenza pneumonitis in a child with inherited IRF9 deficiencyIRF9 deficiency**. The Journal of experimental medicine 2018, **215**(10):2567–2585.
21. García-Morato MB, Apalategi AC, Bravo-Gallego LY, Moreno AB, Simón-Fuentes M, Garmendia JV, Echevarría AM, del Rosal Rabes T, Domínguez-Soto Á, López-Granados E: **Impaired control of multiple viral infections in a family with complete IRF9 deficiency**. Journal of Allergy and Clinical Immunology 2019, **144**(1):309–312. e310.
22. Jefferies CA: **Regulating IRFs in IFN Driven Disease**. Frontiers in Immunology 2019, **10**.
23. Michalska A, Blaszczyk K, Wesoly J, Bluyssen HAR: **A Positive Feedback Amplifier Circuit That Regulates Interferon (IFN)-Stimulated Gene Expression and Controls Type I and Type II IFN Responses**. Frontiers in Immunology 2018, **9**.
24. Fu XY, Kessler DS, Veals SA, Levy DE, Darnell JE, Jr.: **ISGF3, the transcriptional activator induced by interferon alpha, consists of multiple interacting polypeptide chains**. Proc Natl Acad Sci U S A 1990, **87**(21):8555–8559.
25. Veals SA, Schindler C, Leonard D, Fu XY, Aebersold R, Darnell JE, Jr., Levy DE: **Subunit of an alpha-interferon-responsive transcription factor is related to interferon regulatory factor and Myb families of DNA-binding proteins**. Mol Cell Biol 1992, **12**(8):3315–3324.

26. Wang JT, Doong SL, Teng SC, Lee CP, Tsai CH, Chen MR: **Epstein-Barr virus BGLF4 kinase suppresses the interferon regulatory factor 3 signaling pathway**. J Virol 2009, **83**(4):1856–1869.
27. Wei Y, Zhou J, Yu H, Jin X: **AKT phosphorylation sites of Ser473 and Thr308 regulate AKT degradation**. Bioscience, biotechnology, and biochemistry 2019, **83**(3):429–435.
28. Gong J, Holewinski RJ, Van Eyk JE, Steinberg SF: **A novel phosphorylation site at Ser130 adjacent to the pseudosubstrate domain contributes to the activation of protein kinase C-δ**. Biochemical Journal 2016, **473**(3):311–320.
29. Studer RA, Rodriguez-Mias RA, Haas KM, Hsu JI, Viéitez C, Solé C, Swaney DL, Stanford LB, Liachko I, Böttcher R *et al*: **Evolution of protein phosphorylation across 18 fungal species**. Science 2016, **354**(6309):229–232.
30. Anisimova M, Bielawski JP, Yang Z: **Accuracy and Power of Bayes Prediction of Amino Acid Sites Under Positive Selection**. Molecular Biology and Evolution 2002, **19**(6):950–958.
31. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T *et al*: **An overview of Ensembl**. Genome Res 2004, **14**(5):925–928.
32. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. Mol Biol Evol 2007, **24**(8):1586–1591.
33. Yang Z, dos Reis M: **Statistical Properties of the Branch-Site Test of Positive Selection**. Molecular Biology and Evolution 2010, **28**(3):1217–1228.
34. Pond SLK, Muse SV: **HyPhy: hypothesis testing using phylogenies**. In: *Statistical methods in molecular evolution*. Springer; 2005: 125–181.
35. Kosakovsky Pond SL, Poon AF, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A: **HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies**. Molecular biology and evolution 2020, **37**(1):295–299.
36. Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM *et al*: **Gene-Wide Identification of Episodic Selection**. Molecular Biology and Evolution 2015, **32**(5):1365–1371.
37. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL: **Detecting Individual Sites Subject to Episodic Diversifying Selection**. PLOS Genetics 2012, **8**(7):e1002764.
38. Kosakovsky Pond SL, Frost SDW: **Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection**. Molecular Biology and Evolution 2005, **22**(5):1208–1222.
39. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL: **Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection**. Molecular Biology and Evolution 2015, **32**(5):1342–1353.
40. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server**. Bioinformatics 2000, **16**(4):404–405.
41. Cramer P: **AlphaFold2 and the future of structural biology**. Nature Structural & Molecular Biology 2021, **28**(9):704–705.

42. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A: **Highly accurate protein structure prediction with AlphaFold**. Nature 2021, **596**(7873):583–589.
43. Chen J, Zhang P, Chen H, Wang X, He X, Zhong J, Zheng H, Li X, Jakovlić I, Zhang Y *et al*: **Whole-genome sequencing identifies rare missense variants of WNT16 and ERVW-1 causing the systemic lupus erythematosus**. Genomics 2022, **114**(3):110332.
44. Karczewski K, Francioli L: **The genome aggregation database** (**gnomAD**). *MacArthur Lab* 2017.
45. Consortium G, Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T: **The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans**. Science 2015, **348**(6235):648–660.
46. Consortium G, Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T: **The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans**. *Science* 2015, **348**(6235):648–660.

Figures

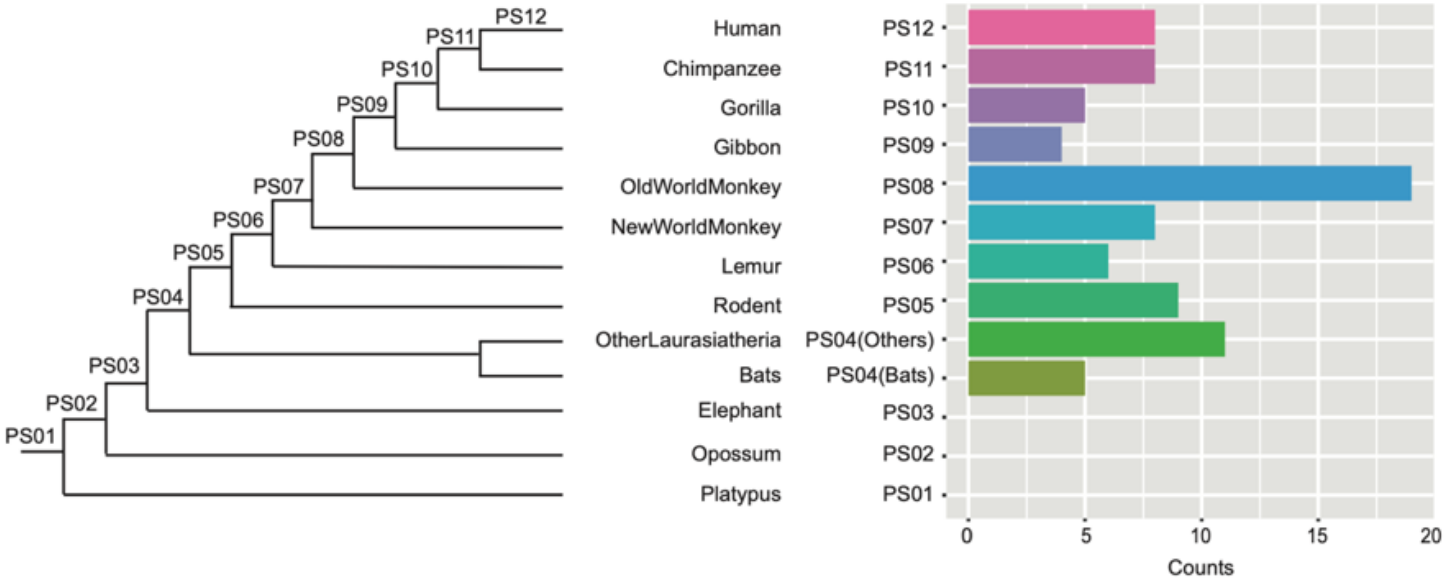

Figure 1

The number of genes under significant positive selection in studied mammalian lineages. The “PS” indicates “phylostratum or phylostratigraphic stage” [10, 11]. PS01 to PS12 are ancestral nodes leading to human.

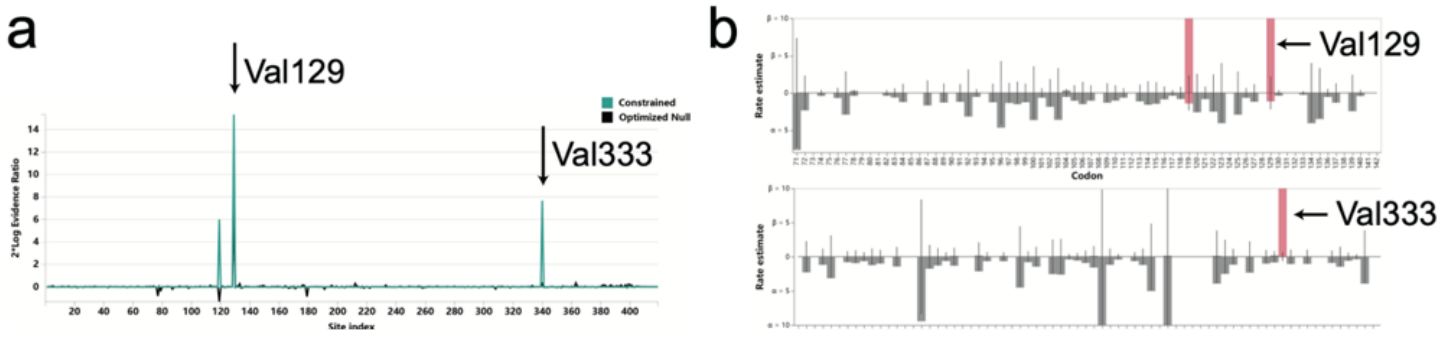

Figure 2

a) The site model test statistics based on the BUSTED analysis. Val129 could be under positive selection,0 as supported by the highest 2*Log evidence ratio. b) The FEL method identified positive selection on Val129 and Val333 (i.e., 340). The red bars indicate the significantly positively selected sites. Maximum likelihood estimates of synonymous (α) and non-synonymous rates (β) at each site are shown as bars. The line shows the estimates under the null model (α=β). c) and d) The regional alignments for the two positively selected sites of *IRF9* respectively. Black arrows show the sites and "**" shows the NEB significance level.

Figure 3

The secondary and tertiary structures of *IRF9* with the focus on the two sites in the human orthologue (p.219V and p.333V) detected by NEB as the positively selected sites. a) The secondary structure of *IRF9* and regions around the two sites based on the prediction of PSIPRED in human and rat orthologues. b) The tertiary structure of *IRF9* in human and rat orthologues based on inference by Alphafold2. The black arrows show the two sites under positive selection based on the NEB inference.

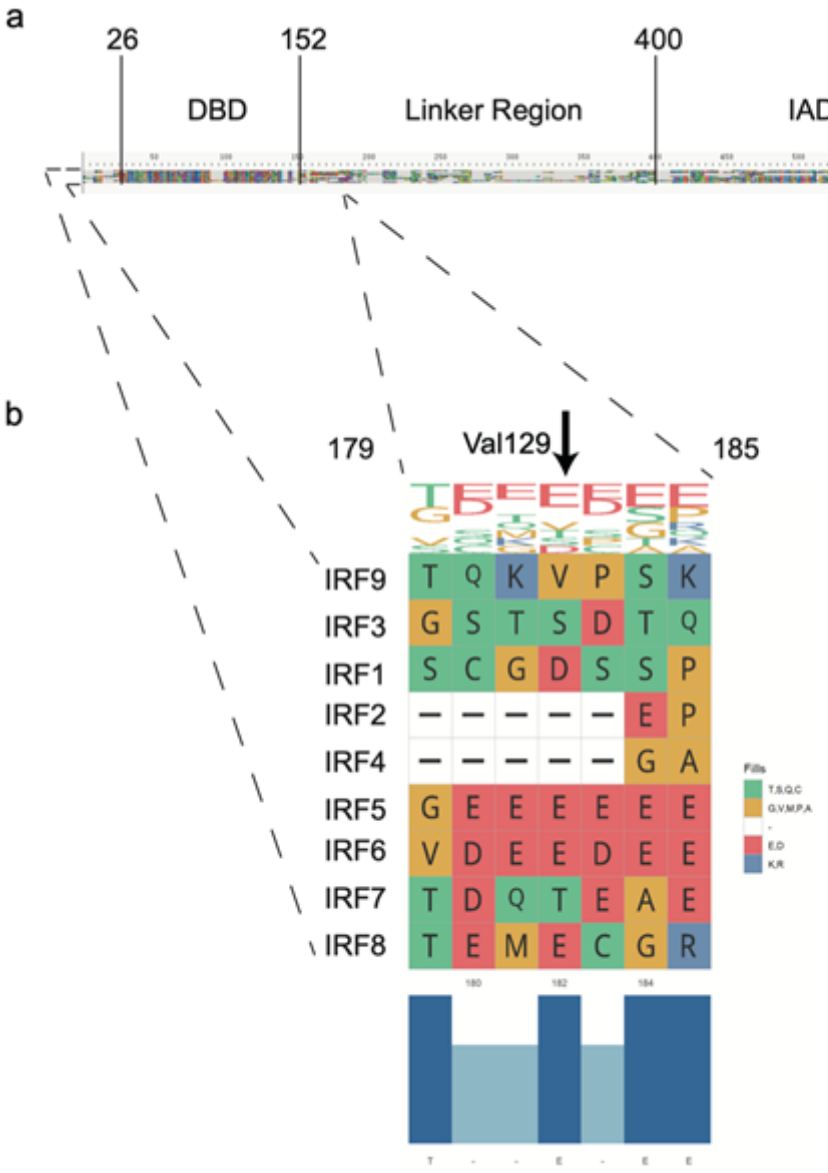

Figure 4

The alignment of homologous *IRF* genes. a) a miniature of multiple sequence alignment of *IRFs*, showing DBD, a linker region, and IAD; b). The enlarged region around the positively selected site Val129. Note: except the Val129, coordinates in (a) and (b) are based on *IRF* protein alignment including gaps.

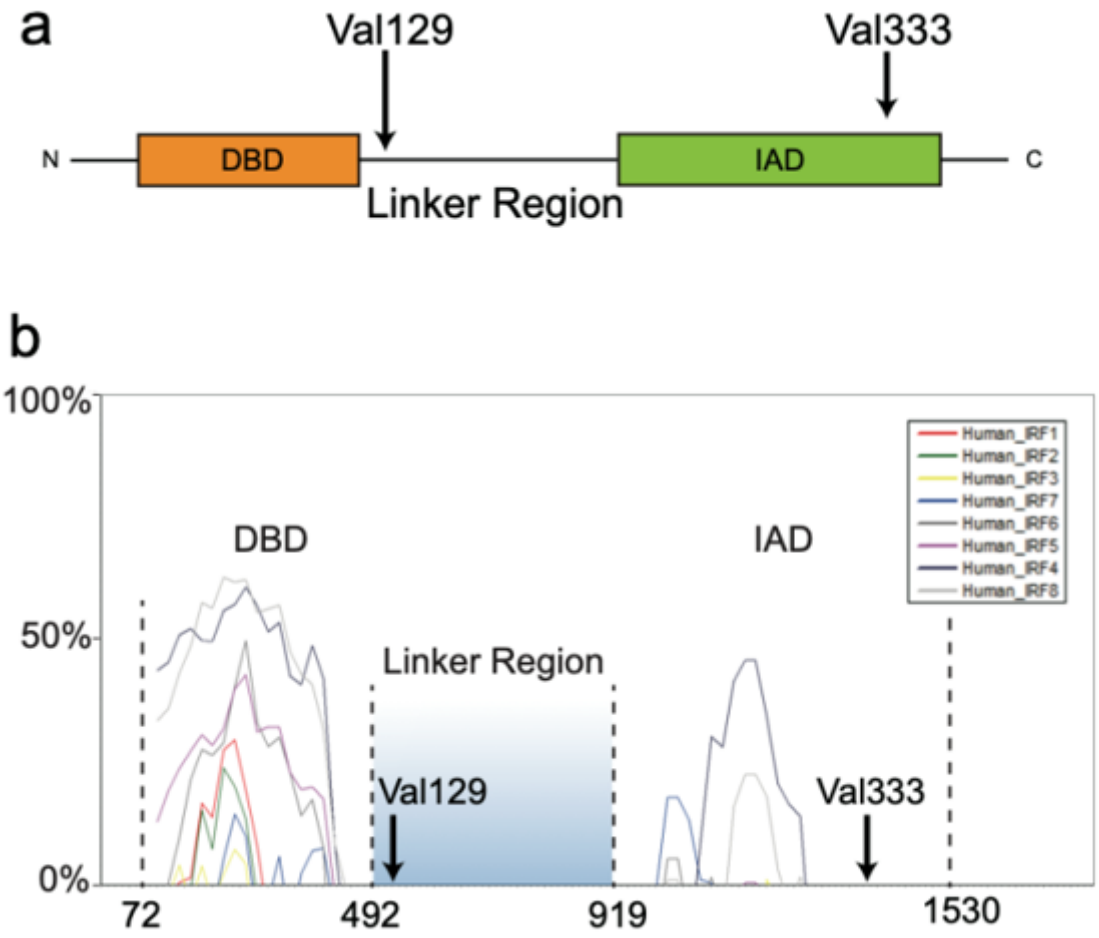

Figure 5

Structural similarities among homologous *IRF* genes. a) a simplified scheme of the *IRF9 domain*, showing DBD, a linker region, and IAD; b) The nucleotide similarities between *IRF9* and other homologous genes. The positively selected sites are shown by black arrows. The horizontal axis in (b) indicates CDS coordinates. The vertical axis shows the similarity between *IRF9* and its homologs.

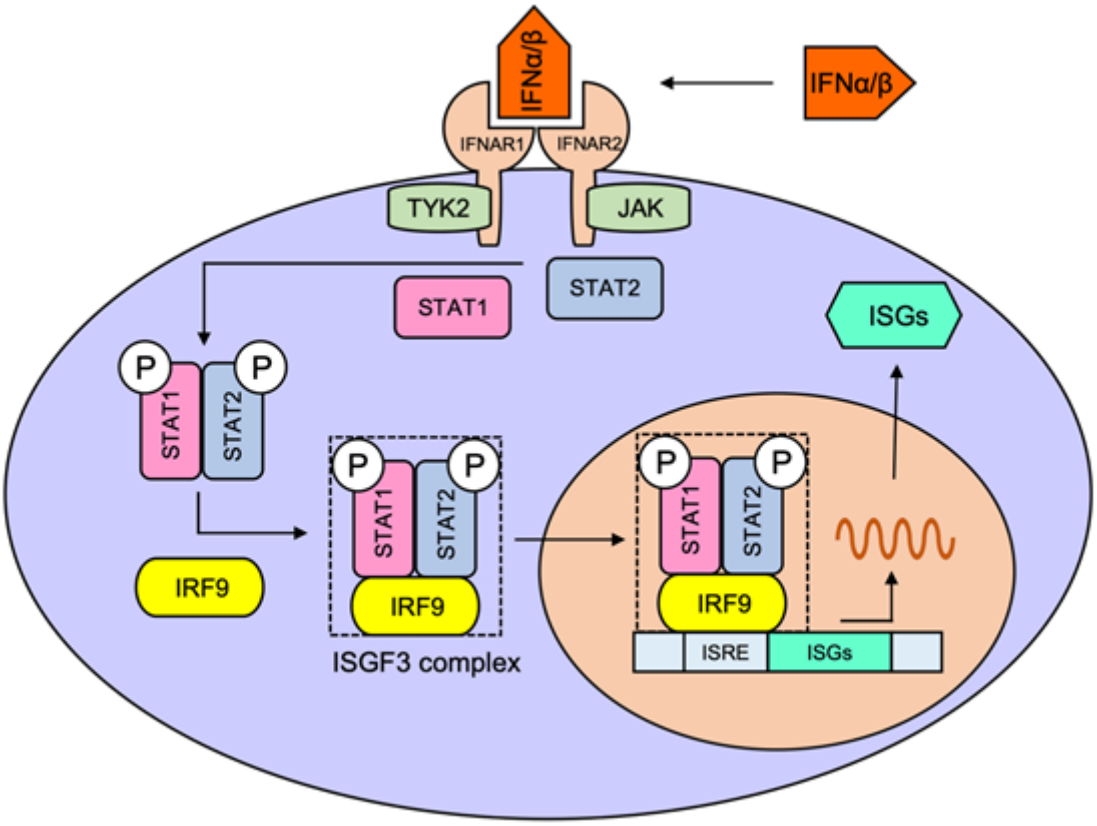

Figure 6

Schematic molecular processes of the type I IFN signaling based on literature review. The abbreviations and full names are IFNAR1 (Interferon Alpha and Beta Receptor Subunit 1), IFNAR2 (Interferon Alpha and Beta Receptor Subunit 2), TYK2 (Tyrosine Kinase 2), JAK (Janus kinase), STAT1 (Signal Transducer And Activator Of Transcription 1), STAT2 (Signal Transducer And Activator Of Transcription 2), ISRE (IFN stimulated response elements), ISGs (interferon-stimulated genes).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.docx
- Supplementarytable.xlsx