Evolutionary Reconstruction and Population Genetics Analysis of Aurora Kinases

Balu Kamaraj[☯], **Ambuj Kumar**[☯], **Rituraj Purohit**[*]

Bioinformatics Division, School of Bio Sciences and Technology, Vellore Institute of Technology University, Vellore, Tamil Nadu, India

Abstract

Background: Aurora kinases belong to the highly conserved kinase family and play a vital role in cell cycle regulation. The structure and function of these kinases are inter-related and sometimes they also act as substitutes in case of knockdown of other aurora kinases.

Method: In this work we carried out the evolutionary reconstruction and population genetic studies of aurora kinase proteins. Substitution saturation test, CAI (Codon adaptation index), gene expression and RSCU (Relative synonymous codon usage) values were computed for all the three aurora kinases. Linear regression method was used to check the dependency of gene expression on their CAI values.

Results: The results suggested that aurora-B and aurora-C has shown convergence in their evolutionary pathway. Moreover, the aurora-A I57V mutation showed high penetrance in human population and exist at very high frequency (84.4%) when compared to the native residue (15.6%). The mutation showed notable range of functional gain and seemed to be promising for the evolution of aurora-A function. Mutant allele might also become a challenging prospect for understanding the pattern of evolution followed by cell cycle kinases.

Conclusion: The overall result suggested that the aurora-A is currently under the evolutionary transition and to determine the functional significance of the mutation further investigation are required.

Citation: Kamaraj B, Kumar A, Purohit R (2013) Evolutionary Reconstruction and Population Genetics Analysis of Aurora Kinases. PLoS ONE 8(9): e75763. doi:10.1371/journal.pone.0075763

Editor: Ali Torkamani, The Scripps Research Institute, United States of America

Received April 3, 2013; **Accepted** August 16, 2013; **Published** September 24, 2013

Funding: The authors have no support or funding to report.

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

* E-mail: riturajpurohit@gmail.com

☯ These authors contributed equally to this work.

Introduction

Aurora kinases belong to the highly conserved protein kinase family and play a vital role in regulating cell cycle progression mechanism. They help the dividing cell to dispense its genetic materials to its daughter cells. More specifically, Aurora kinases play a crucial role in cellular division by controlling chromatid segregation [1]. The activity of Aurora kinases are strictly regulated by a group of events like as phosphorylation and degradation. Special interest in Aurora kinases has arisen since people discovered their defects in association with severe mitotic abnormalities and cancers. Deregulation of Aurora kinase activity can result in mitotic abnormality and genetic instability, leading to defects in centrosome function, spindle assembly, chromosome alignment, and cytokinesis [2]. The human aurora family of serine-threonine kinases comprises three members, which act in concert with many other proteins to control chromosome assembly and segregation during mitosis [3]. Aurora A expression in tumours is often associated with gene amplification, genetic instability, poor histologic differentiation, and poor prognosis [1]. Aurora B is frequently expressed at high levels in a variety of tumours, often coincidently with aurora A, and expression level has also been associated with increased genetic instability and clinical outcome [1]. Further, aurora kinase gene polymorphisms are associated with increased risk or early onset of cancer. The expression of aurora C in cancer is less studied. In recent years, several small aurora kinase inhibitor molecules have been developed that exhibit preclinical activity against a wide range of solid tumours [3].

Aurora kinases comprise mainly two domains: a regulatory domain in the NH2 terminus and a catalytic domain in the COOH terminus. The regulatory domain is largely diverse, whereas the catalytic domain with a short segment of diverse COOH terminus shares >70% homology among Aurora-A, Aurora-B, and Aurora-C. There is a D-Box in the COOH terminus and an A-Box in the NH2 terminus of Aurora kinases, which are responsible for degradation [2,4-7]. These features have been confirmed by experiments in both Aurora-A and Aurora-B; yet, the structure of Aurora-C is deduced by sequence alignment and seems to lack the A-Box. Aurora-A and Aurora-B are distinct in subcellular distribution during mitosis [2]. Despite great similarities in sequences and structures, Aurora kinases are completely different in their subcellular distribution. Aurora-A localizes to the pericentriolar material from the end of S phase to the beginning of the next G1 and spreads to the pole proximal ends of spindle microtubules during mitosis. In contrast, Aurora-B remains in the nucleus and moves to centromeres from prometaphase to metaphase [2]. After anaphase begins, Aurora-B relocates gradually to the midzone and persists at the midbody until cytokinesis is completed. Moreover, Aurora-B undergoes different localizations along with at least three other partners: INCENP, Survivin, and Borealin [2]. They form a tight complex within the cell during mitosis and are named “chromosomal passengers” as they move precisely from site to site at specific times. By G2 phase, Aurora-A localizes to pericentriolar material and persists throughout cell cycle. Additionally, it spreads to the minus ends of the mitotic spindle microtubules and midzone microtubules during mitosis [2]. Diversely, Aurora-B and its “passenger” partners remain a part of centromeres from prometaphase to metaphase. After chromatids begin to separate, they relocate to midzone and persist at the midbody until cytokinesis is completed. Moreover, Aurora-A and Aurora-B exhibit divergent functions in mitotic control, corresponding to different subcellular distribution of the two kinases. Aurora-A is mainly involved in centrosome function, mitotic entry, and spindle assembly, whereas Aurora-B participates in chromatin modification, microtubule-kinetochore attachment, spindle checkpoint, and cytokinesis. Along with these two kinases, different partners and substrates take part in these processes [2]. But in contrast to the wide range of distinct variation in their functional regulation, when N158 is mutated to G, the characteristic amino acid at that position in Aurora A, Aurora B became 350-fold more active than wild-type enzyme, similar to the activity level of wild type Aurora A [8]. Also it is remarkable that the large difference in intrinsic activity of the two kinases in the absence of regulatory subunits can be account for a single amino acid difference [8]. It seems as if by compromising at some specific amino acid position cause greater convergence in the activity regulation of Aurora-A and Aurora-B which explains the interlinking between their evolutionary pathways.

Although some vital experiments have handsomely explained the function of Aurora-C, the explicit insights have not been obtained. It has been suggested that Aurora-C is a Serine/threonine-protein kinase component of the chromosomal passenger complex (CPC), a complex that acts as a key regulator of mitosis. The CPC complex has essential functions at the centromere in ensuring correct chromosome alignment and segregation and is required for chromatin-induced microtubule stabilization and spindle assembly. It also plays a vital role in meiosis and more particularly in spermatogenesis. It has redundant cellular functions with Aurora B and can rescue an Aurora B knockdown. Like Aurora B, Aurora C phosphorylates histone H3 at 'Ser-10' and 'Ser-28'. It also phosphorylates TACC1, another protein involved in cell division, at 'Ser-228' [9-11]. It is likely to be understood that the Aurora-B and Aurora-C shares significant functional coordination and thus it is likely to be evolved in convergent manner. Also, by observing the traces of evolutionary linkage between the Aurora-A and Aurora-B proteins, it is evident that all three Aurora kinases might have shared some common evolutionary road map.

In this study we have investigated the evolutionary relationship between the three aurora kinases and compiled the crucial factors that might have governed their evolutionary pattern. Phylogenetic reconstructions were conducted to study their evolutionary divergence. Furthermore, the sequence based analyses were carried to investigate the functional regions evolving under different rates and also crucial factors governing their evolution. Moreover the population genetic studies were conducted to determine the substitution based evolution occurring in the aurora kinases in present human population. The overall results obtained from this study would aid in determining the key amino acid residues and the regions involved in functional regulation of aurora kinases.

Materials and Methods

Datasets

The coding sequence data for Aurora kinases were retrieved from the ENSEMBLE/GenBank gene databases [12]. We considered the protein coding mRNA sequence (CDS) by removing the partial and internal stop codon sequences. Partial and the incomplete sequences were omitted from the dataset.

Phylogenetic Reconstruction

We constructed phylogenetic trees using distance maximum parsimony (MP), maximum likelihood quartet puzzling (QP) and Bayesian posterior probabilities (BP). The programs SEQBOOT and CONSENSE were used to estimate the confidence limits of branching points from 1000 bootstrap replications. ML tree topologies were constructed using the software PUZZLE 4.0 [13], employing 1000 puzzling steps, the JTT substitution matrix, estimation of rate heterogeneity using the gamma distribution model with eight rate categories, and the gamma-parameter estimation from the dataset. MP analysis was performed using PAUP4.0b5 software [14] where the number and lengths of minimal trees were estimated from 100 random sequence additions, while confidence limits of branch points were estimated by 1000 bootstrap replications. BP trees were constructed using the software MrBayes v3.0B4 [15,16]. Bayesian analysis used the mixed model of sequence evolution with random starting trees. Markov chains were run for 106 generations, burn-in values were set for 104 generations, and trees sampled every 100 generations. All trees were visualized using the program Dendroscope [17]. The multiple sequence alignment for the Aurora kinase proteins were obtained using CLUSTALW [18].

Substitution saturation and relative synonymous codon usage analysis

Substitution saturation in all three aurora kinase protein coding region was determined by using DAMBE (data analysis in molecular biology and evolution) software V.5.2.9 [19,20]. The relative synonymous codon usage (RSCU) [21] method was applied to calculate the codon usage for the dataset. Relative synonymous codon usage (RSCU) is defined as the ratio of the observed frequency of codons to the expected frequency given that all the synonymous codons for the same amino acids are used equally. RSCU values have no relation to the amino acids usage and the abundance ratio of synonymous codons, which can directly reflect the bias of synonymous codon usage [22].

CAI and Expression analysis

The codon adaptation index (CAI) was used to estimate the extent of bias towards codons that were known to be preferred in highly expressed genes [21]. It is now proved that CAI values mostly approach the theoretical values to reflect the expression level of a gene. Thus it has been widely utilized to measure the gene expression level [23,24]. A CAI value ranges from 0 to 1.0, and a higher value indicates a stronger codon usage bias. CAI provides an indication of gene expression level under the assumption that there is translational selection to optimize gene sequences according to their expression levels. Relationship between the evolution of codon usage pattern and the corresponding gene expression values were computed using DAMBE package. Linear regression analyses were conducted to examine the dependence of expression on the CAI values of aurora kinases. The dependencies were computed using the standard linear regression model given below.

$$\alpha = \beta x + \varepsilon$$

Here α is the dependent variable (expression value) and the β is the independent variable (CAI). The x is the slope and ε is the intercept of the line on Y axis.

Positive selection detection

PAML [25] and HyPhy [26] packages were implemented to compute the positively selected residues in Aurora kinases. The type of selection constraint for the identified sites can be tested based on the posterior probability values. The inferred codon sites with significant posterior probability (>50%) are likely to be under variable selection pressure as estimated by the Bayes empirical Bayes (BEB) and naive empirical Bayes (NEB) approach. Codeml models of PAML (phylogenetic analysis maximum likelihood) package were applied to predict the codon sites and the type of selective pressure in a protein-coding sequence, as different codon would be under different selective pressures [27]. The models compute the rate of nucleotide substitution per codon site using non-synonymous (amino acid changing, dN) to synonymous (no change in amino acid, dS) ratio (ω) which indicates the selective pressure that is inferred through maximum likelihood method. Furthermore, the Random Effect Likelihood (REL) program from HyPhy package was implemented to examine the positively selected residues in Aurora kinases. It is a reliable model for low divergence or fewer numbers (5 to 15) of sequences for conducting a selection test. The program automatically screens the submitted set of aurora kinase sequences to identify the best nucleotide substitution models. These models were computed using hierarchical testing procedure together with LRT and AIC. The REL codon site model was fitted to the aurora kinase subclass genes by incorporating the nucleotide substitution model with the generated NJ trees. For every codon site, REL computes two Bayes factor values; if the computed value for a site was >50, then the amino acid sites are said to be positively selected otherwise negatively selected. Finally both PAML and HyPhy packages were used altogether to compute the dN/dS ratios to predict the most likely position of substitution selection.

Population genetics studies

SPSmart [28] was used to examine the penetrance level of Aurora kinases large scale genomic mutations in different human populations. The SPSmart engine encompasses a range of data sets: Hap-Map release, the Stanford University and University of Michigan CEPH-HGDP, i.e., Centre d'Etude du Polymorphisme Humain (CEPH) Human Genome Diversity Panel (HGDP) and the Perlegen SNP Data Set [28]. The mutations with penetrance frequency > 50% was considered as significant. Further, the mutations were subjected to SNAP (screening for non-acceptable polymorphisms) package [29] which is a neural network-based method for the prediction of the functional effects of non-synonymous SNPs and used to examine its linkage disequilibrium values. MutPred [30], SNPeffect 4.0 [31], SIFT [32], Polyphen 2.0 [33] and I-mutant2.0 [34] tools were implemented to study the molecular changes associated with the above compiled mutations. MutPred is a web based tool, used to predict the molecular cause of disease related amino acid substitution [30]. It utilizes several attributes related to protein structure, function, and evolution. Associated phenotype changes for the predicted mutation were examined using SNPeffect4.0 tool [31]. SIFT uses sequence homology-based approach to classify amino acid substitutions [32]. The prediction score < 0.05 is considered to be deleterious. PolyPhen 2.0 is based on combination of sequence and structure based attributes and uses naive Bayesian classifier for the identification of an amino acid substitution and the impact of mutation. The output levels of probably damaging and possibly damaging were classified as functionally significant (≤0.5) and the benign level being classified as tolerated (≥0.51) [33]. I-Mutant 2.0 is a support vector machine (SVM) based tool for the automatic prediction of protein stability changes upon single point mutations. I-Mutant 2.0 can be used both as a classifier for predicting the sign of the protein stability change upon mutation and as a regression estimator for predicting the related DDG values [34]. From the above tools, it was clearly showed the molecular changes upon mutations.

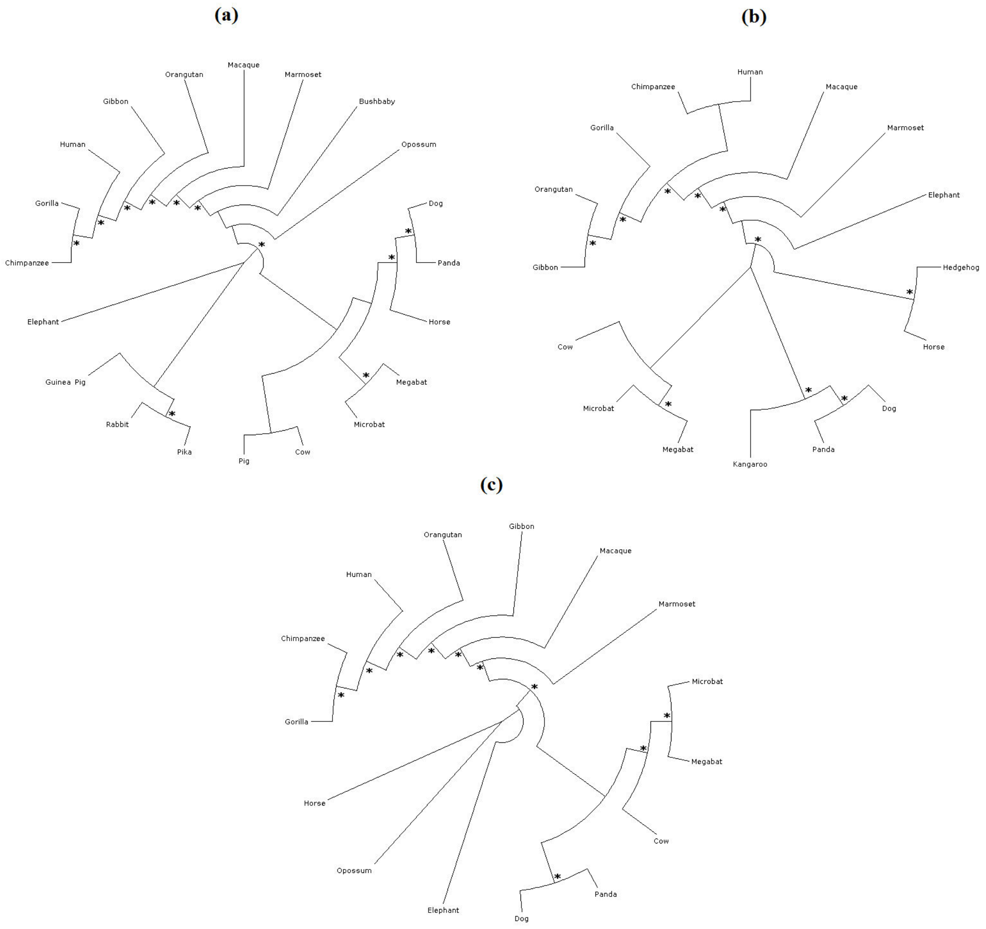


Figure 1. Reconstructed phylogenetic tree of Aurora kinase protein. Asterisks ("*") indicate those nodes supported 70% or greater by distance maximum parsimony (MP) and maximum likelihood quartet puzzling (QP) tree-building methods and 0.90 Bayesian posterior probability. (a) Aurora-A, (b) Aurora-B, (c) Aurora-C.

doi: 10.1371/journal.pone.0075763.g001

Results and Discussion

Nucleotide sequence for the coding region of Aurora kinases were collected from the ENSEMBLE database specifically for the primate species. Since some of the primate species Aurora kinases are not completely sequenced, we omitted them from our study. Moreover, the coding sequence of rat and mouse were not included due to the presence of additional functional motiff in their coding region which might cause errors while calculating the substitution saturation rates. Furthermore, few species such as Bushbaby, Pika, Rabbit and Guinea pig sequences were not included while constructing the evolutionary tree for Aurora kinase B and C due to the unavailability of nucleotide sequence for their complete coding region in these species. Multiple sequence alignments showed that the Aurora kinase family is highly conserved. Pairwise sequence comparisons had suggested that the mean proportion of similar amino acids is much higher among all the different families of Aurora-A, Aurora-B and Aurora-C of vertebrates than within the same family (Aurora-A or Aurora-B) between vertebrates and invertebrates species. This suggested a recent evolutionary radiation of Aurora families within vertebrates [35]. Moreover, the phylogenetic tree shown in Figure 1 suggested that the Aurora kinases had followed a similar evolutionary road map. We further conducted substitution saturation test, to investigate if the proteins had reached the level of substitution saturation threshold level. The transition transversion plots against the K80 distances clearly showed that the evolution driven substitution in aurora kinase proteins were well under the saturation threshold level (Figure 2). Moreover, a greater divergence between the transition-transversion rates were observed in case of Aurora-A, depicting that the Aurora-A is more likely to evolve when compared to Aurora-B and Aurora-C (Figure 2).

The evolution driven changes in the codon usage pattern significantly accounted for the distinct functional behaviour existing within the proteins. The results obtained in our RSCU analysis further suggested a supportive evidence of multiple codon usage patterns among the aurora kinases. It was interesting to see that the codons such as GCU, UGC, GAU,

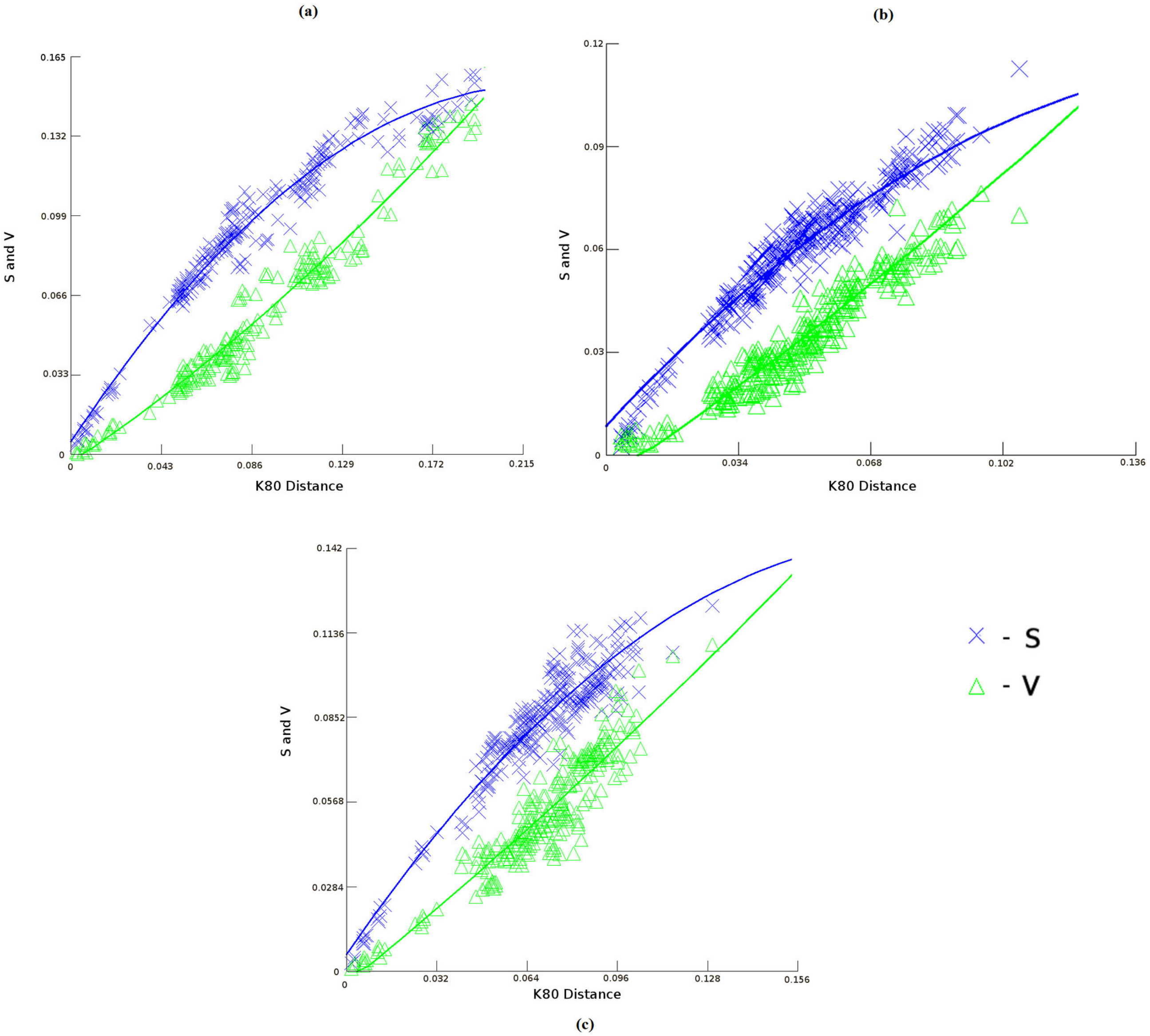


Figure 2. Transition (S) and transversion (V) values plotted against the K80 distance for (a) Aurora-A, (b) Aurora-B, (c) Aurora-C.

doi: 10.1371/journal.pone.0075763.g002

GGA, CUG, UUG, AUG, CCU, CCA, AGA, UCU, ACC and UGG were shown to exist at very high usage frequency (Table 1). It made sense to say that these codons might construct the signature patterns of Aurora kinase family. Moreover, there were several other codons which were showing higher usage frequency specific to the individual protein. Furthermore, it was interesting to see that the codon triplets used for coding serine amino acids did not show high usage frequency in Aurora-B and Aurora-C, whereas in Aurora-A, the frequency of serine coding triplets showed relatively higher usage pattern. This might account for its specific scaffolding behaviour, allowing their amino acid residues to be phosphorylated in a different manner.

Evolution can affect the phenotype either by modifying the sequences of proteins or by changing their pattern of expression. Although there were evidences of evolution showing co-regulatory behaviour on the gene expression level and the corresponding codon sequences. Thus, we further extended our study to examine if these variations in the codon usage pattern caused any changes in its expression patterns. CAI is a widely used index for characterizing gene expression in general and translation efficiency in particular. The CAI values for each of these genes were compiled and plotted against their expression values obtained from in-silico experiment (Figure 3). The expression values obtained in this study were calculated using the isoelectric gel profiling method

Table 1. RCSU value of codons for AURKA, AURKB and AURKC genes.

Amino Acid	CODON		RCSU	
		AURKA	AURKB	AURKC
ALA	**GCU**	**1.202**	**1.380**	**1.121**
	GCC	0.875	1.023	1.067
	GCA	1.503	0.865	1.058
	GCG	0.421	0.732	0.754
CYS	UGU	0.906	0.678	0.750
	UGC	**1.094**	**1.322**	**1.250**
ASP	**GAU**	**1.294**	**1.629**	**1.112**
	GAC	0.706	0.371	0.888
GLU	GAA	1.322	1.403	0.871
	GAG	0.678	0.597	1.129
PHE	UUU	0.993	1.245	1.024
	UUC	1.007	0.755	0.976
GLY	GGU	0.968	0.650	0.577
	GGC	0.758	0.782	0.694
	GGA	**1.675**	**1.606**	**1.215**
	GGG	0.599	0.962	1.514
HIS	CAU	0.934	1.050	0.964
	CAC	1.066	0.950	1.065
ILE	AUU	1.482	0.846	1.065
	AUC	1.042	1.254	0.935
	AUA	0.476	0.900	0.619
LYS	AAA	1.298	1.164	0.671
	AAG	0.702	0.836	1.329
LEU	CUU	0.995	1.489	0.878
	CUC	0.885	0.370	0.780
	CUA	1.000	1.072	0.618
	CUG	**1.121**	**1.069**	**1.724**
	UUA	0.917	0.961	0.528
	UUG	**1.083**	**1.039**	**1.472**
MET	**AUG**	**1.000**	**1.000**	**1.381**
ASN	AAU	0.860	1.076	1.179
	AAC	1.140	0.924	0.821
PRO	**CCU**	**1.440**	**1.326**	**1.197**
	CCC	0.849	1.068	1.201
	CCA	**1.372**	**1.097**	**1.161**
	CCG	0.340	0.510	0.440
GLN	CAA	1.372	1.346	0.846
	CAG	0.931	0.654	1.154
ARG	CGU	0.814	0.766	0.709
	CGC	0.945	0.574	0.880
	CGG	0.989	1.313	1.547
	CGA	1.252	1.347	0.863
	AGA	**1.398**	**1.094**	**1.384**
	AGG	0.602	0.906	1.757
SER	**UCU**	**1.027**	**1.510**	**1.394**
	UCC	1.323	0.861	1.429
	UCA	1.313	0.970	0.806
	UCG	0.338	0.660	0.371
	AGU	1.195	1.121	0.736
	AGC	0.805	0.879	1.264
THR	ACU	1.296	0.936	0.675
	ACC	**1.046**	**1.406**	**1.293**
	ACA	1.215	0.872	1.434
	ACG	0.443	0.786	0.598
VAL	GUU	0.998	0.927	0.736
	GUC	1.449	1.201	0.643
	GUA	0.850	0.952	0.834
	GUG	0.702	0.920	1.787
TRP	**UGG**	**1.000**	**1.000**	**1.193**
TYR	UAU	1.132	0.957	0.945
	UAC	0.868	1.043	1.055

Codons highlighted in bold were having RCSU values greater than 1 in all three genes.

doi: 10.1371/journal.pone.0075763.t001

included in DAMBE package, which has been trained on the empirical data's of acid-resistant gastric pathogen *Helicobacter pylori*. The pI values from DAMBE using the same method had also been used to study the adaptive evolution of the matrix extracellular phosphoglycoprotein in mammals and the implication of its change on protein folding [36]. A notable pattern of gene expression changes and their corresponding CAI values were observed for Aurora-B and Aurora-C, whereas for Aurora-A there was no strict fashion of CAI dependent expression changes. The plot clearly depicted that the CAI had greater role in optimizing the expression values of Aurora-B and Aurora-C genes (Figure 3). With the rise in CAI values, the gene expression reduced consistently and extended up to ~ 0.06 in both the genes. Following dependencies were observed for Aurora-B.

$$\mathit{Expressionvalue} = -0.874 \times \mathit{CAI} + 0.78$$

whereas for Aurora-C, the dependency was

$$\mathit{Expressionvalue} = -0.873 \times \mathit{CAI} + 0.69$$

In aurora-A, no such uniform and strict regression patterns were observed, suggesting that the Aurora-B and Aurora-C showed a similar fashion of CAI based gene expression changes.

The extent to which selection affected genes and genomes is a key question in genetics and molecular evolution. Selection might modulate gene sequence evolution in different ways, for example, by constraining potential changes of amino acid sequences (purifying or negative selection) or by favouring new and adaptive genetic variants (positive selection) [37]. To quantify selection in the simplest case, the number of non-synonymous differences in a pair of protein-coding sequences could be estimated. However, substitution rates vary across the genome and between species that made direct comparisons solely based on non-synonymous substitutions difficult. To control variation in the underlying mutation rate, a standard way had to take the ratio of the number of non-synonymous differences per total number of possible non-synonymous changes (dN) to the number of synonymous differences per

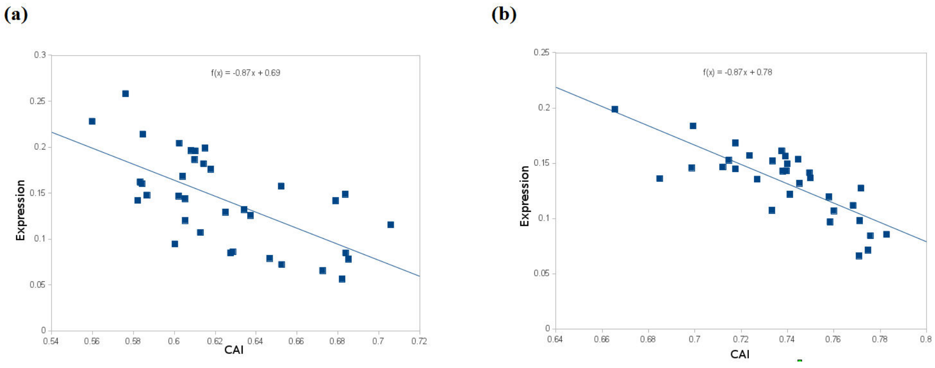


Figure 3. Gene expression values plotted against the codon adaptation indices. (a) Aurora-B, (b) Aurora-C.
doi: 10.1371/journal.pone.0075763.g003

total number of synonymous changes (dS) [38]. This ratio (dN/dS) was then used as a measure of ‘‘functional divergence’’ that accounted for the underlying local or regional variation in the substitution rate for which dS is taken as a proxy. The ratio dN/dS were calculated using PAML and HyPhy packages. It was interesting to observe that all three aurora kinases mainly evolved under the influence of negative selection showing 58 negatively selected sites in Aurora-A, 64 in Aurora-B and 93 negatively selected sites in Aurora-C protein coding region. At any given moment of time, positive selection, which favoured currently uncommon derived alleles, affected only a small fraction of sites in the genome and, thus, was much rarer than negative selection, which favoured common ancestral alleles [39]. Since the higher number of neucleotide substitutions were very harmful for the conserved proteins, specially the aurora kinases, and the negative selection played an important role in maintaining the long-term stability of their structures and function by removing deleterious mutations. When the required optimizations were achieved, there was a danger of losing that improvement by a deleterious mutation. The negative selection made sure that deleterious mutations could not take over a population and that any improved structures, once fixed in a population, were maintained as long as they were needed [40]. The specificity of scaffolding in aurora kinases requires evolution driven optimization of their structure which further made the negative selection as the key criteria for its evolution. The substitution saturation results suggested that more or less all the three aurora kinase structures were tending towards their required optimization levels and hence any further substitution was most likely to cause major deleterious effect in their structure as well as function.

Positive selection had undoubtedly played a critical role in the evolution of *Homo sapiens* and promoted the emergence of new phenotypes. Comparative genetics/genomics studies in recent years have uncovered a growing list of genes that might have experienced positive selection during the evolution of human and/or primates. Moreover we investigated the positively selected residues that could have played significant role in evolution of aurora kinases function. PAML predicted a total of two residue positions including 31F and 66S in Aurora-A, 17A, 69T, 68L, 294A and 300A in Auora-B and 5R, 12K, 13A, 26T and 33P in Aurora-C. REL model predicted 23K, 31F and 66S in Aurora-A as positively selected, 16T, 17A, 69T, 294A and 300A in Aurora-B as positively selected, whereas 5R, 7V, 12K, 13A, 26T and 245S as positively selected in Aurora-C. Significance of these selections could be directly seen, where the Aurora A (F31I) polymorphism showed preferential amplification associated with increased aneuploidy in colon cancers and was a low-penetrance cancer susceptibility allele affecting multiple cancer types. Residues position 31F was shown to be essential for the binding of E2 ubiquitin-conjugating enzyme (UBE2N) protein, resulted in co-localization of UBE2N with Aurora-A at the centrosomes during mitosis. The 31I variant showed notable loss in the binding potential, ultimately leading to the degree of aneuploidy in human colon tumours [41]. This directly indicates the role of positive selection in optimizing the function of Aurora-A protein, which in turn aids in its essential interaction with UBE2N protein. Furthermore, the protein kinases exert control over their protein targets by covalent modification of a Ser, Thr, or Tyr residue with the γ-phosphate group cleaved from ATP. The selection of 66S in Aurora-A, 16T and 69T in Aurora-B and 26T and 245S in Aurora-C might correspond to an additional phosphorylation sites required for their recruitment and scaffolding activity. Moreover, the distribution of positively selected residues suggested that the majority of selections were concentrated in the N terminal flanking region. The N-terminal region determines selectivity during protein-protein interactions [42]. The N lobe was responsible for positioning the ATP phosphate group through a αC helix, whereas the activation loop within the C lobe was able to harbour substrates. Only when the two lobes arrive at a certain conformation could the Aurora kinases fulfil their kinase functions. Low conservation and high evolvability in this region might be a crucial factor in the evolution of complex network of its scaffolding activity. Although the clear picture of their functional role had not been established, which was likely be essential for maintaining the activity of the aurora kinases.

We further extended our observation to the evolutionary selection of Aurora kinases occurring in the current human population. In the process of substitution, a previously non-

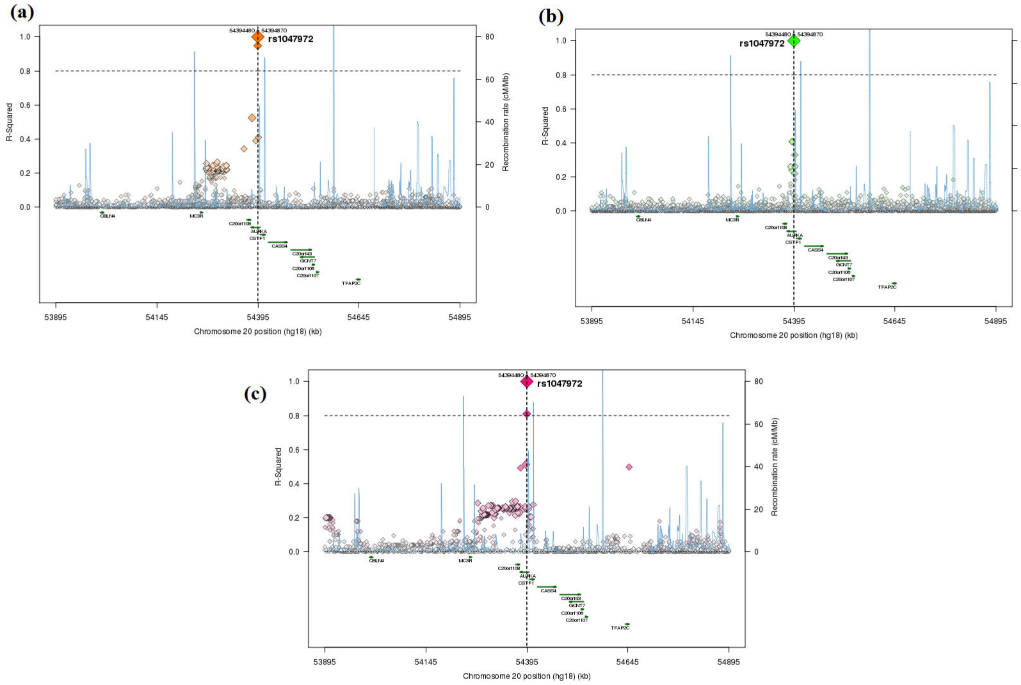
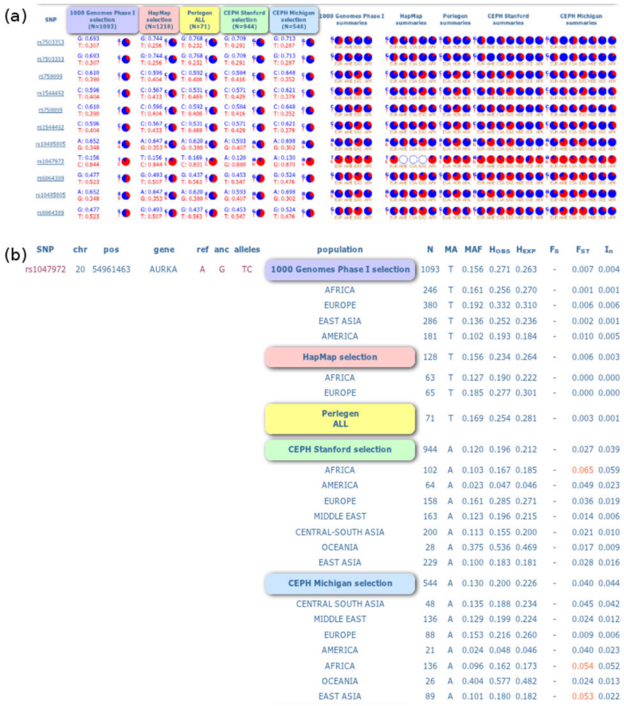


Figure 5. The R-Squared values and recombination rates of rs1047972 in (a) CEU; (b) YRI and (c) CHB/JPT populations.
doi: 10.1371/journal.pone.0075763.g005

existent allele arises by mutation and undergoes fixation by spreading through the population by random genetic drift and/or positive selection. Once the frequency of the allele is at 100%, it is said to be fixed in the population. Detecting the possibility of mutation fixation rates might provide a clear picture of current evolutionary trend followed by aurora kinases in human. I57V mutation in Aurora-A was shown to be highly penetrant, prevailing in 84.4% of human populations (Figure 4a-b). This directly showed that the mutant allele had been strongly under the influence of evolutionary selection and had invaded up to a very large mass of human population. To examine the molecular effect of the mutation, we further subjected it to SNPeffect4.0, I-mutant2.0, SIFT, Polyphen and MutPred tools. No damaging or deleterious consequences were observed in SIFT and polyphen scores. Moreover, SNPeffect4.0 depicted no change in LIMBO, TANGO and WALTZ score, suggesting no impact on the aggregation tendency, chaperone binding tendency and the amyloid propensity of protein respectively. The results clearly suggested that the mutation had been completely accepted in a large mass of human population genome, in terms of its functional and structural effects and did not cause any negative consequences on the protein activity. Furthermore, the I-mutant2.0 scores suggested that the free energy change induced by the mutations were also tolerable, indicating negligible tendency of mutation to cause structural damages. Further, we implemented MutPred tool to examine the molecular changes invoked by the I57V mutation. It was interesting to observe that the mutation has tendency to invoke three significant changes in the protein, including gain of Molecular Recognition Features (MoRFs) binding, gain of phosphorylation at S53 amino acid position and gain of glycosylation at P58. MoRFs were short, interaction-prone segments of protein disorder that undergo disorder-to-order transitions upon specific binding, representing a specific class of intrinsically disordered regions that exhibited molecular recognition and binding functions. MoRFs were common in various proteomes and occupy a unique structural and functional niche in which function was a direct consequence of intrinsic disorder. Gain of MoRF binding would significantly aid in scaffold regulation of aurora kinase A. Furthermore, gain of phosphorylation and glycosylation at the corresponding residues also presented an insight of the gain of functional properties induced by the mutation. Moreover, this mutation seemed to cross the fixation threshold and could be an essential part of the Aurora-A evolutionary road map. The observations indicated strong fixation tendency of mutant allele and was further confirmed by the linkage disequilibrium scores obtained from the SNAP tool (Figure 5). Figure 5 showed that the mutation had very high linkage disequilibrium scores and recombination rates in Utah residents with ancestry from northern and western Europe (CEU), Yoruba (YRI), Han Chinese in Beijing (CHB) and Japanese in Tokyo (JPT) populations. Higher recombination rates of alleles were generally useful to introduce genetic novelty into the existing genotypes, which further explained the chances for the selection of I57V mutation in Aurora-A kinase. The overall results strongly suggested that the Aurora A I57V mutation had deeply invaded the human population and was likely to induce significant gain in the functional activity of protein. It would be interesting to examine the molecular and phenotypic changes associated with I57V mutation and its benefits.



(a)

SNP	1000 Genomes Phase I selection (N=1093)	HapMap selection (N=1218)	Perlegen ALL (N=71)	CEPH Stanford selection (N=944)	CEPH Michigan selection (N=546)
rs7503353	G: 0.693 T: 0.307	G: 0.744 T: 0.256	G: 0.768 T: 0.232	G: 0.709 T: 0.291	G: 0.713 T: 0.287
rs7503353	G: 0.693 T: 0.307	G: 0.744 T: 0.256	G: 0.768 T: 0.232	G: 0.709 T: 0.291	G: 0.713 T: 0.287
rs758099	C: 0.610 T: 0.390	C: 0.596 T: 0.404	C: 0.592 T: 0.408	C: 0.584 T: 0.416	C: 0.648 T: 0.352
rs1544492	C: 0.596 T: 0.404	C: 0.567 T: 0.433	C: 0.531 T: 0.469	C: 0.571 T: 0.429	C: 0.621 T: 0.379
rs758099	C: 0.610 T: 0.390	C: 0.596 T: 0.404	C: 0.592 T: 0.408	C: 0.584 T: 0.416	C: 0.648 T: 0.352
rs1544492	C: 0.596 T: 0.404	C: 0.567 T: 0.433	C: 0.531 T: 0.469	C: 0.571 T: 0.429	C: 0.621 T: 0.379
rs10485805	A: 0.652 G: 0.348	A: 0.647 G: 0.353	A: 0.620 G: 0.380	A: 0.593 G: 0.407	A: 0.698 G: 0.302
rs1047972	T: 0.156 C: 0.844	T: 0.156 C: 0.844	T: 0.169 C: 0.831	A: 0.120 G: 0.880	A: 0.130 G: 0.870
rs6064389	G: 0.477 T: 0.523	G: 0.493 T: 0.507	G: 0.437 T: 0.563	G: 0.453 T: 0.547	G: 0.524 T: 0.476
rs10485805	A: 0.652 G: 0.348	A: 0.647 G: 0.353	A: 0.620 G: 0.380	A: 0.593 G: 0.407	A: 0.698 G: 0.302
rs6064389	G: 0.477 T: 0.523	G: 0.493 T: 0.507	G: 0.437 T: 0.563	G: 0.453 T: 0.547	G: 0.524 T: 0.476

(b)

SNP	chr	pos	gene	ref	anc	alleles	population	N	MA	MAF	H_{OBS}	H_{EXP}	F_S	F_{ST}	I_n
rs1047972	20	54961463	AURKA	A	G	TC	1000 Genomes Phase I selection	1093	T	0.156	0.271	0.263	-	0.007	0.004
							AFRICA	246	T	0.161	0.256	0.270	-	0.001	0.001
							EUROPE	380	T	0.192	0.332	0.310	-	0.006	0.006
							EAST ASIA	286	T	0.136	0.252	0.236	-	0.002	0.001
							AMERICA	181	T	0.102	0.193	0.184	-	0.010	0.005
							HapMap selection	128	T	0.156	0.234	0.264	-	0.006	0.003
							AFRICA	63	T	0.127	0.190	0.222	-	0.000	0.000
							EUROPE	65	T	0.185	0.277	0.301	-	0.000	0.000
							Perlegen ALL	71	T	0.169	0.254	0.281	-	0.003	0.001
							CEPH Stanford selection	944	A	0.120	0.196	0.212	-	0.027	0.039
							AFRICA	102	A	0.103	0.167	0.185	-	0.065	0.059
							AMERICA	64	A	0.023	0.047	0.046	-	0.049	0.023
							EUROPE	158	A	0.161	0.285	0.271	-	0.036	0.019
							MIDDLE EAST	163	A	0.123	0.196	0.215	-	0.014	0.006
							CENTRAL-SOUTH ASIA	200	A	0.113	0.155	0.200	-	0.021	0.010
							OCEANIA	28	A	0.375	0.536	0.469	-	0.017	0.009
							EAST ASIA	229	A	0.100	0.183	0.181	-	0.028	0.016
							CEPH Michigan selection	544	A	0.130	0.200	0.226	-	0.040	0.044
							CENTRAL SOUTH ASIA	48	A	0.135	0.188	0.234	-	0.045	0.042
							MIDDLE EAST	136	A	0.129	0.199	0.224	-	0.024	0.012
							EUROPE	88	A	0.153	0.216	0.260	-	0.009	0.006
							AMERICA	21	A	0.024	0.048	0.046	-	0.040	0.023
							AFRICA	136	A	0.096	0.162	0.173	-	0.054	0.052
							OCEANIA	26	A	0.404	0.577	0.482	-	0.024	0.013
							EAST ASIA	89	A	0.101	0.180	0.182	-	0.053	0.022

Figure 4. Results from population genetics analysis of Aurora kinases. (a) Frequency distribution of Aurora-A single neucleotide polymorphism's (SNPs) in different human population. (b) Heterozygosity values of the minor and major alleles in different human population

doi: 10.1371/journal.pone.0075763.g004

Acknowledgements

We gratefully acknowledge the management of Vellore Institute of Technology University for providing the facilities to carry out this work. The authors would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the manuscript.

Author Contributions

Conceived and designed the experiments: AK RP. Performed the experiments: BK. Analyzed the data: AK BK. Contributed reagents/materials/analysis tools: AK BK. Wrote the manuscript: BK AK.

References

1. Macarulla T, Ramos FJ, Tabernero J (2008) Aurora kinase family: a new target for anticancer drug. Recent Patents Anticancer Drugs Discov 3(2): 114–122. doi:10.2174/157489208784638785. PubMed: 18537754.
2. Fu J, Bian M, Jiang Q, Zhang C (2007) Roles of aurora kinases in mitosis and tumorigenesis. Mol Cancer Res 5: 1–10. doi: 10.1158/1541-7786.MCR-06-0208. PubMed: 17259342.
3. Gautschi O, Heighway J, Mack PC, Purnell PR Jr., Gandara DR et al. (2008) Aurora kinases as anticancer drug targets. Clin Cancer Res 14: 1624–1633. PubMed: 18347165.
4. Nguyen HG, Chinnappan D, Urano T, Ravid K (2005) Mechanism of Aurora-B degradation and its dependency on intact KEN and A-boxes: identification of an aneuploidy-promoting property. Mol Cell Biol 25: 4977-4992. doi:10.1128/MCB.25.12.4977-4992.2005. PubMed: 15923616.
5. Castro A, Vigneron S, Bernis C, Labbé JC, Prigent C et al. (2002) The D-Boxactivating domain (DAD) is a new proteolysis signal that stimulates the silent DBox sequence of Aurora-A. EMBO Rep 3: 1209-1214. doi:10.1093/embo-reports/kvf241. PubMed: 12446569.
6. Castro A, Arlot-Bonnemains Y, Vigneron S, Labbé JC, Prigent C et al. (2002) APC/Fizzy-Related targets Aurora-A kinase for proteolysis. EMBO Rep 3: 457-462. doi:10.1093/embo-reports/kvf095. PubMed: 11964384.
7. Crane R, Kloepfer A, Ruderman JV (2004) Requirements for the destruction of human Aurora-A.J. J Cell Sci 117: 5975-5983. doi: 10.1242/jcs.01418. PubMed: 15536123.
8. Eyers PA, Churchill ME, Maller JL (2005) The Aurora A and Aurora B protein kinases: a single amino acid difference controls intrinsic activity and activation by TPX2. Cell Cycle. 4: 784–789. doi:10.4161/cc. 4.6.1693. PubMed: 15908779.
9. Yan X, Cao L, Li Q, Wu Y, Zhang H et al. (2005) Aurora C is directly associated with survivin and required for cytokinesis. Genes Cells 10: 617–626. doi:10.1111/j.1365-2443.2005.00863.x. PubMed: 15938719.
10. Gabillard JC, Ulisse S, Baldini E, Sorrenti S, Cremet JY et al. (2011) Aurora-C interacts with and phosphorylates the transforming acidic coiled-coil 1 protein 408. Biochem Biophys Res Co. pp. 647-653.
11. Avo Santos M, van de Werken C, de Vries M, Jahr H, Vromans MJ, et al. (2011) A role for Aurora C in the chromosomal passenger complex during human preimplantation embryo development. Hum Reprod 26:1868–1881. PubMed: 21531210.
12. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K et al. (2013) Ensembl 2013. Nucleic Acids Res 41: D48-D55. doi:10.1093/nar/gks1236. PubMed: 23203987.
13. Strimmer K, von Haesler A (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol 13: 964-969. doi:10.1093/oxfordjournals.molbev.a025664.
14. Swpfford DL (1999). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sunderland, MA: Sinauer Associates; 1999

15. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755. doi:10.1093/bioinformatics/17.8.754. PubMed: 11524383.
16. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574. doi:10.1093/bioinformatics/btg180. PubMed: 12912839.
17. Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357-358. PubMed: 8902363.
18. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680. doi:10.1093/nar/22.22.4673. PubMed: 7984417.
19. Xia X, Lemey P (2009) Assessing substitution saturation with DAMBE. The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny. New York: Kluwer Publishing House Academic Publishers. pp. 615–630.
20. Xia X (2013) DAMBE5: A Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution. Mol Biol Evol 30(7): 1720–1728. doi:10.1093/molbev/mst064. PubMed: 23564938.
21. Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15(3): 1281–1295. doi:10.1093/nar/15.3.1281. PubMed: 3547335.
22. Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24(1-2): 28–38. doi:10.1007/BF02099948. PubMed: 3104616.
23. Naya H, Romero H, Carels N, Zavala A, Musto H (2001) Translational selection shapes codon usage in the GC-rich genomes of *Chlamydomonas reinhardtii*. FEBS Lett. 501(2-3): 127–130. doi:10.1016/S0014-5793(01)02644-8. PubMed: 11470270.
24. Gupta SK, Bhattacharyya TK, Ghosh TC (2004) Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. J Biomol Struct Dyn 21(4): 527–536. doi:10.1080/07391102.2004.10506946. PubMed: 14692797.
25. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591. doi:10.1093/molbev/msm088. PubMed: 17483113.
26. Kosakovsky PSL, Frost SDW (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21: 2531–2533. doi:10.1093/bioinformatics/bti320. PubMed: 15713735.
27. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogenous selection pressure at amino acid sites. Genetics 155: 431–449. PubMed: 10790415.
28. Amigo J, Salas A, Phillips C, Carracedo A (2008) SPSmart: Adapting population based SNP genotype databases for fast and comprehensive web access. BMC Bioinformatics, 9: 428–433. doi:10.1186/1471-2105-9-428. PubMed: 18847484.
29. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res, 35(11): 3823-3835. doi:10.1093/nar/gkm238. PubMed: 17526529.
30. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK et al. (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics, 25: 2744-2750. doi:10.1093/bioinformatics/btp528. PubMed: 19734154.
31. De Baets G, Van Durme J, Reumers J, Maurer-Stroh S, Vanhee P et al. (2012) SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res 40: D935–9(Database issue): D935-9. PubMed: 22075996
32. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc, 4: 1073-1081. doi:10.1038/nprot.2009.86. PubMed: 19561590.
33. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249. doi:10.1038/nmeth0410-248. PubMed: 20354512.
34. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant 2.0: Predicting stability changes upon Mutation from the protein sequence or structure. Nucleic Acids Res 33: 306–310. doi:10.1093/nar/gki375.
35. Brown JR, Koretke KK, Birkeland ML, Sanseau P, Patrick DR (2004) Evolutionary relationships of Aurora kinases: implications for model organism studies and the development of anti-cancer drugs. BMC Evol Biol, 4: 39. doi:10.1186/1471-2148-4-39. PubMed: 15476560.
36. Machado JP, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A (2011) Adaptive evolution of the matrix extracellular phosphoglycoprotein in mammals. BMC Evol Biol 11: 342. doi:10.1186/1471-2148-11-342. PubMed: 22103247.
37. Chen J, Sun Y (2011) Variation in the analysis of positively selected sites using nonsynonymous/synonymous rate ratios: an example using influenza virus. PLOS ONE. 6(5): e19996. doi:10.1371/journal.pone.0019996. PubMed: 21629696.
38. Wolf JB, Künstner A, Nam K, Jakobsson M, Ellegren H (2009) Nonlinear Dynamics of Nonsynonymous (dN) and Synonymous (dS) Substitution Rates Affects Inference of Selection. Genome Biol Evol 1: 308-319. PubMed: 20333200.
39. Bazykin GA, Kondrashov AS (2011) Detecting past positive selection through ongoing negative selection. Genome Biol Evol 3: 1006-1013. doi:10.1093/gbe/evr086. PubMed: 21859804.
40. Loewe L (2008) Negative selection. Nature. Education: 1.
41. Ewart-Toland A, Briassouli P, de Koning JP, Mao JH, Yuan J et al. (2003) Identification of Stk6/STK15 as a candidate low-penetrance tumor-susceptibility gene in mouse and human. Nat Genet, 34: 403–412. doi:10.1038/ng1220. PubMed: 12881723.
42. Bolanos-Garcia VM (2005) Aurora kinases. Int J Biochem Cell Biol, 37: 1572–1577. doi:10.1016/j.biocel.2005.02.021. PubMed: 15896667.