Table 3. Univariate Analysis of Risks for Kidney Tubular Dysfunction in Patients With HIV Infection Treated With Tenofovir

Characteristic	OR	95% CI	*P* Value
Female sex	1.844	.204–16.67	.586
Age per 1 year	1.165	1.100–1.233	<.001
Weight per 1 kg decrement	1.076	1.021–1.135	.007
CD4 count per 1/μL decrement	1.002	.999–1.004	.261
Baseline eGFR per 1 mL/minutes/1.73 m^2 decrement	1.052	1.016–1.090	.004
Concurrent use of nephrotoxic drugs	1.559	.322–7.555	.581
Hepatitis B	0.721	.156–3.319	.674
C-reactive protein per 1 mg/dL	1.551	.689–3.494	.289
Hypertension	2.234	.843–5.922	.106
Dyslipidemia	0.578	.183–1.823	.349
Duration of treatment with tenofovir disoproxil fumarate (weeks)	0.999	.992–1.007	.888
ABCC2			
–24 CC	10.50	1.369–80.55	.024
1249 AA	7.828	1.609–38.10	.011
–24 CC plus 1249 AA	31.88	3.131–324.5	.003
2934 GG	1.358	.167–11.07	.775
ABCC4			
559 TT	4.912	.837–28.81	.078
912 TT	1.466	.531–4.042	.460
2269 AA	2.756	.530–14.34	.228
3348 GG	1.950	.510–7.463	.329
4135 GG	1.254	.450–3.494	.665
4976 CC	2.462	.925–6.547	.071
ABCC10			
526 GG	1.158	.360–3.725	.805
2759 TT	0.619	.220–1.738	.363
ABCB1			
2677 AA	7.828	1.609–38.10	.011

Abbreviations: CI, confidence interval; eGFR: estimated glomerular filtration rate; HIV, human immunodeficiency virus; OR, odds ratio.

[a] Due to low prevalence of minor alleles, rs56220353, rs11568630, and rs2274407 were not included in this analysis.

associated with tenofovir-induced KTD (OR, 2.497; 95% CI, .902–6.949; $P = .077$).

DISCUSSION

The present study demonstrated that genotype CC at position −24 and genotype AA at position 1249 of *ABCC2* gene are associated with tenofovir-induced KTD in Japanese patients with HIV-1 infection. The effect of SNPs was more evident in patients with both −24 CC and 1249 AA homozygotes than in those with either homozygote only. The findings of this study resolve long-term controversy over the role of genetic

Table 4. Multivariate Analysis for the Risk of Tenofovir-Induced Kidney Tubular Dysfunction With Homozygotes at −24 and 1249 of *ABCC2* in Patients With HIV Infection

ABCC2	Adjusted OR	95% CI	*P* Value
Homozygote at −24 CC	20.08	1.711–235.7	.017
Homozygote at 1249 AA	16.21	1.630–161.1	.017
Homozygotes at −24 CC plus 1249 AA	38.44	2.051–720.4	.015

Each variable was adjusted for sex, age, weight, estimated glomerular filtration rate, and hypertension.

Abbreviations: CI, confidence interval; OR, odds ratio.

polymorphisms in tenofovir-induced KTD and confirm the effect of the SNPs in *ABCC2* gene in tenofovir-induced KTD.

CA haplotype (−24C, 1249A) of *ABCC2* was associated with tenofovir-induced KTD, whereas TG was a protective haplotype (Table 5). Izzedine et al [13] reported the role of CATC haplotype (−24C, 1249A, 3563T, 3972C) of *ABCC2* in KTD. However, 3563T did not play such role in this haplotype analysis, because the prevalence of 3563T is 0% in the Japanese, according to the HapMap data, and haplotype with only −24C plus 1249A still exhibited its effect on tenofovir-induced KTD (Table 5; www.hapmap.org). The reported association between tenofovir-induced KTD and 526G and 2759C of *ABCC10* described by Pushpakom et al [21] was also not reproduced in this study. Furthermore, SNPs in *ABCC4*, *SLC22A6*, and *ABCB1* investigated in the present study did not show a significant association with tenofovir-induced KTD (Table 3).

Three main aspects of our study are important. First, this is the first study to our knowledge that elucidated the effect of SNPs on tenofovir-induced KTD conducted in a country other than European countries or the United States. Our study examined Japanese patients of genetic background different from patients of previous studies, which consisted mostly of whites. While SNPs −24C and 1249A of *ABCC2* have been speculated to correlate with tenofovir-induced KTD in previous studies, the present study confirmed that these SNPs are risk factors for tenofovir-induced KTD in nonwhites.

The result that the SNPs in *ABCC2* are a risk for tenofovir-induced KTD can also be applied to patients with other genetic backgrounds who host SNPs −24C and 1249A. Notably, the impact of SNPs on tenofovir-induced KTD might be more significant in Africans and Indians than in Japanese or whites, considering that the allele frequencies of −24C and 1249A are higher in these population according to the HapMap data (−24C; Africans 96.9%, Indians 92.6%, Japanese 80.8%, whites 81.9%, 1249A; Africans 21.7%, Indians 30.7%, Japanese 8.9%, whites 23.7%; www.hapmap.org).

Second, the study was designed to evaluate the exclusive effect of SNPs on tenofovir-induced KTD by excluding

Table 5. Association Between Haplotype in *ABCC2* and *ABCC4* and Kidney Tubular Dysfunction

SNP Marker/Haplotype	Allele	Allele/Haplotype Frequency, %		OR (95% CI)[a]	*P* Value
		KTD Group (n = 19)	Control Group (n = 171)		
ABCC2					
−24 C → T	C	97.4	78.4	10.22 (1.658–419.8)	.003
1249 G → A	A	28.9	12.3	2.91 (1.345–6.296)	.011
ABCC2 haplotype	CA	28.9	12.3	2.91 (1.295–6.221)	.011
	TG	2.6	21.6	0.098 (.002–.603)	.003
ABCC4					
559 G → T	T	21.1	12.3	1.905 (.705–4.614)	.213
4976 T → C	T	48	55.3	0.746 (.375–1.470)	.399
ABCC4 haplotype					
TT	TT	17.6	7.9	2.497 (.902–6.949)	.077

Abbreviations: CI, confidence interval; KTD, kidney tubular dysfunction; OR, odds ratio; SNP, single-nucleotide polymorphism.

[a] ORs and *P* values are for comparisons of allele/haplotype frequencies between the kidney tubular dysfunction and control groups.

possible predisposing factors for KTD, for example, active infection, malignancies, diabetes mellitus, and preexisting renal impairment, which are known risks for KTD [35]. Patients who showed no HIV-1 viral suppression were also excluded. Furthermore, the enrolled patients were Japanese only, and this helped to examine a study population with comparatively similar genetic background. The study population was also on the same antiretroviral regimen (ritonavir-boosted darunavir plus tenofovir/emtricitabine), and this also helped to evaluate more precisely the effect of SNPs, because plasma concentration of tenofovir is affected by concomitant antiretrovirals and the delta change in plasma tenofovir concentration likely differs in the presence of each concomitant drug [26].

Third, SNPs were examined in 190 patients in this study. To our knowledge, the number of enrolled patients is the largest among the studies that have so far examined the effect of SNPs on tenofovir-induced KTD. Thus, this feature provided the study a higher statistical power than previous studies.

Why are polymorphisms in *ABCC2* a risk for tenofovir-induced KTD, even though it is controversial whether MRP2 plays a role in the excretion of tenofovir via the luminal membrane? [18, 20] The exact mechanism has not been determined yet, but we speculate 2 hypotheses. First, there might be unknown endogenous substances that influence tenofovir nephrotoxicity in renal tubular cells, and SNPs in *ABCC2* modulate the function or transportation of such substances [15]. Second, MRP2 may indeed take part in transporting tenofovir, because various substances including methotrexate are reported to be a substrate of MRP2, and *ABCC2* mutation alters excretion of those substances [36, 37]. Further studies are warranted to elucidate the exact mechanism of these SNPs on tenofovir-induced KTD. Furthermore, the impact of these SNPs on KTD with long-term TDF use needs to be evaluated in prospective studies.

Several limitations need to be acknowledged. First, not all polymorphisms in genes of the targeted transporter proteins were examined. Thus, we might have missed other important SNPs on the function of tenofovir transportation. There might be other unknown transporter proteins for tenofovir excretion in the kidney that contribute to susceptibility to tenofovir-induced KTD as well. Second, the diagnostic criteria for TDF-induced KTD are not uniformly established in the field and are different in the published studies. The criteria applied in this study are not entirely similar to the ones used in previous studies that examined the role of SNPs in tenofovir-induced KTD. However, by excluding other predisposing factors for KTD and enrolling a large number of patients, this study succeeded in providing a clear-cut association between SNPs and tenofovir-induced KTD.

In conclusion, the present study demonstrated that SNPs in *ABCC2* associate with tenofovir-induced KTD in Japanese patients, in a setting that excluded other predisposing factors. Assessment of renal tubular function is more cumbersome and costly to monitor than serum creatinine. However, monitoring tubular function is clinically important, because undetected long-term tubular dysfunction might lead to premature osteopenia due to phosphate wasting and accelerated progression of renal dysfunction. Close monitoring of tubular function is warranted in patients with *ABCC2* −24C and 1249A under TDF treatment.

Notes

Acknowledgments. The authors thank Ryo Yamada, Takuro Shimbo, Fumihiko Hinoshita, Yoshimi Kikuchi, Katsuji Teruya, Kunihisa Tsukada, Junko Tanuma, Hirohisa Yazaki, Haruhito Honda, Ei Kinai, Koji

Watanabe, Takahiro Aoki, Daisuke Mizushima, Yohei Hamada, Michiyo Ishisaka, Mikiko Ogata, Mai Nakamura, Akiko Nakano, Fumihide Kanaya, and all other staff at the AIDS Clinical Center for their help in completion of this study.

Financial support. This work was supported by a Grant-in-Aid for AIDS research from the Japanese Ministry of Health, Labour, and Welfare (H23-AIDS-001), and the Global Center of Excellence Program (Global Education and Research Center Aiming at the Control of AIDS) from the Japanese Ministry of Education, Science, Sports, and Culture.

Potential conflicts of interest. S. O. has received honorariums and research grants from MSD K.K., Abbott Japan, Janssen Pharmaceutical K.K., Pfizer, and Roche Diagnostics K.K.; has received honorariums from Astellas Pharmaceutical K.K., Bristol-Myers K.K., Daiichisankyo, Dainippon Sumitomo Pharma, GlaxoSmithKline, K.K., Taisho Toyama Pharmaceutical, Torii Pharmaceutical, and ViiV Healthcare. H. G. has received honorariums from MSD K.K., Abbott Japan, Janssen Pharmaceutical K.K., Torii Pharmaceutical, Roche Diagnostics K.K., and ViiV Healthcare. The remaining authors declare no conflict of interest.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Sax PE, Tierney C, Collier AC, et al. Abacavir/lamivudine versus tenofovir DF/emtricitabine as part of combination regimens for initial treatment of HIV: final results. J Infect Dis **2011**; 204:1191–201.
2. Post FA, Moyle GJ, Stellbrink HJ, et al. Randomized comparison of renal effects, efficacy, and safety with once-daily abacavir/lamivudine versus tenofovir/emtricitabine, administered with efavirenz, in antiretroviral-naive, HIV-1-infected adults: 48-week results from the ASSERT study. J Acquir Immune Defic Syndr **2010**; 55:49–57.
3. Arribas JR, Pozniak AL, Gallant JE, et al. Tenofovir disoproxil fumarate, emtricitabine, and efavirenz compared with zidovudine/lamivudine and efavirenz in treatment-naive patients: 144-week analysis. J Acquir Immune Defic Syndr **2008**; 47:74–8.
4. de Vries-Sluijs TE, Reijnders JG, Hansen BE, et al. Long-term therapy with tenofovir is effective for patients co-infected with human immunodeficiency virus and hepatitis B virus. Gastroenterology **2010**; 139:1934–41.
5. Izzedine H, Hulot JS, Vittecoq D, et al. Long-term renal safety of tenofovir disoproxil fumarate in antiretroviral-naive HIV-1-infected patients. Data from a double-blind randomized active-controlled multicentre study. Nephrol Dial Transplant **2005**; 20:743–6.
6. Nelson MR, Katlama C, Montaner JS, et al. The safety of tenofovir disoproxil fumarate for the treatment of HIV infection in adults: the first 4 years. AIDS **2007**; 21:1273–81.
7. Cooper RD, Wiebe N, Smith N, Keiser P, Naicker S, Tonelli M. Systematic review and meta-analysis: renal safety of tenofovir disoproxil fumarate in HIV-infected patients. Clin Infect Dis **2010**; 51:496–505.
8. Verhelst D, Monge M, Meynard JL, et al. Fanconi syndrome and renal failure induced by tenofovir: a first case report. Am J Kidney Dis **2002**; 40:1331–3.
9. Schaaf B, Aries SP, Kramme E, Steinhoff J, Dalhoff K. Acute renal failure associated with tenofovir treatment in a patient with acquired immunodeficiency syndrome. Clin Infect Dis **2003**; 37:e41–3.
10. Peyriere H, Reynes J, Rouanet I, et al. Renal tubular dysfunction associated with tenofovir therapy: report of 7 cases. J Acquir Immune Defic Syndr **2004**; 35:269–73.
11. Kohler JJ, Hosseini SH, Hoying-Brandt A, et al. Tenofovir renal toxicity targets mitochondria of renal proximal tubules. Lab Invest **2009**; 89:513–9.
12. Herlitz LC, Mohan S, Stokes MB, Radhakrishnan J, D'Agati VD, Markowitz GS. Tenofovir nephrotoxicity: acute tubular necrosis with distinctive clinical, pathological, and mitochondrial abnormalities. Kidney Int **2010**; 78:1171–7.
13. Izzedine H, Hulot JS, Villard E, et al. Association between ABCC2 gene haplotypes and tenofovir-induced proximal tubulopathy. J Infect Dis **2006**; 194:1481–91.
14. Kiser JJ, Aquilante CL, Anderson PL, King TM, Carten ML, Fletcher CV. Clinical and genetic determinants of intracellular tenofovir diphosphate concentrations in HIV-infected patients. J Acquir Immune Defic Syndr **2008**; 47:298–303.
15. Rodriguez-Novoa S, Labarga P, Soriano V, et al. Predictors of kidney tubular dysfunction in HIV-infected patients treated with tenofovir: a pharmacogenetic study. Clin Infect Dis **2009**; 48:e108–16.
16. Uwai Y, Ida H, Tsuji Y, Katsura T, Inui K. Renal transport of adefovir, cidofovir, and tenofovir by SLC22A family members (hOAT1, hOAT3, and hOCT2). Pharm Res **2007**; 24:811–5.
17. Mallants R, Van Oosterwyck K, Van Vaeck L, Mols R, De Clercq E, Augustijns P. Multidrug resistance-associated protein 2 (MRP2) affects hepatobiliary elimination but not the intestinal disposition of tenofovir disoproxil fumarate and its metabolites. Xenobiotica **2005**; 35:1055–66.
18. Imaoka T, Kusuhara H, Adachi M, Schuetz JD, Takeuchi K, Sugiyama Y. Functional involvement of multidrug resistance-associated protein 4 (MRP4/ABCC4) in the renal elimination of the antiviral drugs adefovir and tenofovir. Mol Pharmacol **2007**; 71:619–27.
19. Kohler JJ, Hosseini SH, Green E, et al. Tenofovir renal proximal tubular toxicity is regulated by OAT1 and MRP4 transporters. Lab Invest **2011**; 91:852–8.
20. Ray AS, Cihlar T, Robinson KL, et al. Mechanism of active renal tubular efflux of tenofovir. Antimicrob Agents Chemother **2006**; 50:3297–304.
21. Pushpakom SP, Liptrott NJ, Rodriguez-Novoa S, et al. Genetic variants of ABCC10, a novel tenofovir transporter, are associated with kidney tubular dysfunction. J Infect Dis **2011**; 204:145–53.
22. Hoffmeyer S, Burk O, von Richter O, et al. Functional polymorphisms of the human multidrug-resistance gene: multiple sequence variations and correlation of one allele with P-glycoprotein expression and activity in vivo. Proc Natl Acad Sci USA **2000**; 97:3473–8.
23. Horinouchi M, Sakaeda T, Nakamura T, et al. Significant genetic linkage of MDR1 polymorphisms at positions 3435 and 2677: functional relevance to pharmacokinetics of digoxin. Pharm Res **2002**; 19:1581–5.
24. Kurata Y, Ieiri I, Kimura M, et al. Role of human MDR1 gene polymorphism in bioavailability and interaction of digoxin, a substrate of P-glycoprotein. Clin Pharmacol Ther **2002**; 72:209–19.
25. Han WK, Waikar SS, Johnson A, et al. Urinary biomarkers in the early diagnosis of acute kidney injury. Kidney Int **2008**; 73:863–9.
26. Kiser JJ, Carten ML, Aquilante CL, et al. The effect of lopinavir/ritonavir on the renal clearance of tenofovir in HIV-infected patients. Clin Pharmacol Ther **2008**; 83:265–72.
27. Goicoechea M, Liu S, Best B, et al. Greater tenofovir-associated renal function decline with protease inhibitor-based versus nonnucleoside reverse-transcriptase inhibitor-based therapy. J Infect Dis **2008**; 197:102–8.
28. Rodriguez-Novoa S, Labarga P, Soriano V. Pharmacogenetics of tenofovir treatment. Pharmacogenomics **2009**; 10:1675–85.
29. Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. Nephron **1976**; 16:31–41.
30. Salem MA, el-Habashy SA, Saeid OM, el-Tawil MM, Tawfik PH. Urinary excretion of n-acetyl-beta-D-glucosaminidase and retinol binding protein as alternative indicators of nephropathy in patients with type 1 diabetes mellitus. Pediatr Diabetes **2002**; 3:37–41.
31. Ezinga M, Wetzels J, van der Ven A, Burger D. Kidney tubular dysfunction is related to tenofovir plasma concentration, abstract 603. In: Program and abstracts of the 19th Conference on Retroviruses and Opportunistic Infections. 5–8 March 2012, Seattle, Washington.
32. Gupta SK, Eustace JA, Winston JA, et al. Guidelines for the management of chronic kidney disease in HIV-infected patients: recommendations of the HIV medicine association of the infectious diseases society of America. Clin Infect Dis **2005**; 40:1559–85.

33. Gatanaga H, Tachikawa N, Kikuchi Y, et al. Urinary beta2-microglobulin as a possible sensitive marker for renal injury caused by tenofovir disoproxil fumarate. AIDS Res Hum Retroviruses **2006**; 22:744–8.
34. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res **2002**; 30:158–62.
35. Ando M, Yanagisawa N, Ajisawa A, Tsuchiya K, Nitta K. Kidney tubular damage in the absence of glomerular defects in HIV-infected patients on highly active antiretroviral therapy. Nephrol Dial Transplant **2011**; 26:3224–9.
36. Hulot JS, Villard E, Maguy A, et al. A mutation in the drug transporter gene ABCC2 associated with impaired methotrexate elimination. Pharmacogenet Genomics **2005**; 15:277–85.
37. Suzuki H, Sugiyama Y. Single nucleotide polymorphisms in multidrug resistance associated protein 2 (MRP2/ABCC2): its impact on drug disposition. Adv Drug Deliv Rev **2002**; 54:1311–31.

Gene 505 (2012) 324–332

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Extremely slow rate of evolution in the HOX cluster revealed by comparison between Tanzanian and Indonesian coelacanths

Koichiro Higasa [a,1], Masato Nikaido [b,1], Taro L. Saito [a,1], Jun Yoshimura [a,1], Yutaka Suzuki [c], Hikoyu Suzuki [b], Hidenori Nishihara [b], Mitsuto Aibara [b], Benjamin P. Ngatunga [d], Hassan W.J. Kalombo [e], Sumio Sugano [c], Shinichi Morishita [a,*], Norihiro Okada [b,**]

[a] Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-15 Kashiwanoha, Kashiwa City, Chiba 277-0882, Japan
[b] Department of Biological Sciences, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259-B21 Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan
[c] Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-15 Kashiwanoha, Kashiwa City, Chiba 277-0882, Japan
[d] Tanzania Fisheries Research Institute, P.O. Box 9750, Dar es Salaam, Tanzania
[e] Regional Commissioner's Office Tanga, Box 5095, Tanga, Tanzania

ARTICLE INFO

Article history:
Accepted 21 May 2012
Available online 12 June 2012

Keywords:
Living fossil
Synonymous substitution rate
Next generation sequencer

ABSTRACT

Coelacanths are known as "living fossils" because their morphology has changed very little from that in the fossil record. To elucidate why coelacanths have evolved so slowly is thus of primary importance in evolutionary biology. In the present study, we determined the entire sequence of the HOX cluster of the Tanzanian coelacanth (*Latimeria chalumnae*) and compared it with that of the Indonesian coelacanth (*L. menadoensis*), which was available in the literature. The most intriguing result was the extremely small genetic divergence between the two coelacanths. The synonymous divergence of the HOX coding region between the two coelacanths was estimated to be 0.07%, which is ~11-fold smaller than that of human–chimp. When we applied the estimated divergence time of the two coelacanths of 6 million years ago (MYA) and 30 MYA, which were proposed in independent mitochondrial DNA analyses, the synonymous substitution rate of the coelacanth HOX cluster was estimated to be ~11-fold and 56-fold smaller than that of human–chimp, respectively. Thus, the present study implies that the reduction of the nucleotide substitution rate in coelacanth *HOX* genes may account for the conservation of coelacanth morphology during evolution.

1. Introduction

Coelacanths were initially recognized as a distinct taxonomic group by fossil records, which range in age from the Early Devonian to the Late Cretaceous periods (Maisey, 1996). Because no fossil records of coelacanths have been found since 65 million years ago (MYA), coelacanths were believed to have been extinct (Maisey, 1996). Therefore, the discovery in 1938 of the first living coelacanth, *Latimeria chalumnae*, off the coast of South Africa created a sensation in the field of evolutionary biology (Smith, 1939). One of the most interesting observations is that coelacanth morphology has changed very little (Holder et al., 1999; Smith, 1939); most of the characteristics unique to coelacanths (fleshy-lobed fins, hollow nerve cord, poor ossification of the skeleton, lack of defined ribs, and bilobed caudal region) have been maintained from the Devonian era (Carroll, 1988). Accordingly, coelacanths are called "evolutionary relics" or "living fossils". The elucidation of the molecular mechanism of such slow morphological change in coelacanths is of primary importance to understand the morphological evolution of animals from genotype to phenotype.

After the discovery of a second living coelacanth in the Comoros archipelagos (Smith, 1953), the existence of a viable coelacanth population in this region was confirmed. In addition to the Comoros archipelagos, several coelacanths have been captured off the coasts of Mozambique (Schliewen et al., 1993), Madagascar (Heemstra et al., 1996), and Kenya (De Vos and Oyugi, 2002). Nikaido et al. (2011) recently found a genetically distinct coelacanth population off the northern coastal region of Tanzania, indicating that coelacanths are widely distributed throughout the western Indian Ocean. Apart from the western Indian Ocean, two coelacanth individuals were captured off the coast of Manado, Sulawesi, Indonesia (Erdman et al., 1998). These coelacanths are the first individuals recorded from a location outside the western Indian Ocean and were described as a new species, *Latimeria menadoensis*. The divergence time between the two coelacanth species was estimated in three independent studies. Holder et al. (1999) used partial mtDNA sequences and estimated the divergence time at 6 MYA. On the other hand, using the entire mtDNA sequences

Abbreviations: MYA, million years ago; PCR, polymerase chain reaction; SNV, single nucleotide variation; SINE, short interspersed repetitive element.

* Corresponding author. Tel.: +81 47 136 3985; fax: +81 47 136 3977.
** Corresponding author. Tel.: +81 45 924 5742; fax: +81 45 924 5835.
E-mail addresses: moris.utokyo@gmail.com (S. Morishita), nokada@bio.titech.ac.jp (N. Okada).

[1] These authors equally contributed to this work.

0378-1119/$ – see front matter.
doi:10.1016/j.gene.2012.05.047

(except for the d-loop) and Bayesian methods, Inoue et al. (2005) estimated the divergence time at about 30 MYA. Sudarto et al. (2010) also used Bayesian analysis and proposed the divergence time to be 28 MYA.

HOX genes encode a highly conserved family of transcription factors possessing a 60-amino acid residue motif called a homeodomain, and they are involved in morphogenesis during embryonic development (Krumlauf, 1994). Most of the jawed vertebrates have four separate HOX clusters–HOXA, HOXB, HOXC and HOXD–in which about 40 *HOX* genes are arranged. However, the number and composition of HOX clusters of teleost fish are distinct from those of the other vertebrates owing to a whole-genome duplication event specific to teleost fish (Meyer and Málaga-Trillo, 1999). In particular, duplication of the HOX clusters led to eight HOX clusters in an ancestor of teleost fish. Subsequently, some *HOX* genes were lost independently in the lineage of each teleost fish species during evolution. As a result, euteleosts have seven HOX clusters, in which 46 (medaka) to 49 (zebrafish) *HOX* genes were identified (Kurosawa et al., 2006). Amemiya et al. (2010) characterized the complete HOX clusters in the coelacanth genome. Although the HOX clusters of coelacanth were not remarkable relative to those from other species with four clusters, characterization of the complete *HOX* genes of coelacanth allowed us to reconstruct the evolutionary history of HOX clusters among vertebrates.

Consistent with the slow rate of phenotypic changes in coelacanth, several genetic studies showed a slow rate of evolution at the molecular level. Noonan et al. (2004) indicated that the content and organization of the procadherin gene cluster were more conserved in coelacanths than in teleost fish. Furthermore, they indicated fewer amino acid substitutions in coelacanths than in zebrafish and humans. Amemiya et al. (2010) also showed a significantly slower rate of amino acid substitution in the coelacanth HOX cluster than in teleost fish and tetrapods. These observations suggested that the coelacanth genes have evolved under strong purifying selection or that mutation rate has been slowed down in the coelacanth genome. To examine these possibilities, we directly estimated the absolute value of synonymous divergence of the two coelacanths. Namely, we determined the entire HOX cluster sequence for Tanzanian coelacanth *L. chalumnae* and compared it to that of *L. menadoensis*, which was already available in the literature (Amemiya et al., 2010).

2. Materials and methods

2.1. Genomic DNA of Tanzanian coelacanth

Frozen or ethanol-preserved coelacanth materials were transferred from the Tanzania Fisheries Research Institute to the Tokyo Institute of Technology in accordance with international regulations under the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). In the present study, we used a juvenile coelacanth individual TCC041-1 to determine the HOX cluster sequence. The juvenile was found in the body of female coelacanth individual (TCC041) captured off the Tanga region, Tanzania, in 2007. The capture date, locality, and identification number of the specimen were reported in Nikaido et al. (2011). To examine the insertion of a short interspersed nuclear element (SINE) at particular loci, we used the genomic DNA of six Tanzanian and two Comoran adult coelacanths (TCC035, TCC040, TCC041, TCC042, TCC043, TCC044, Comoro_1 and Comoro_2). The two Comoran coelacanth materials were provided by Kyoto University and Aquamarine Fukushima. Total genomic DNA was extracted from frozen or ethanol-preserved tissues using the DNeasy Tissue kit (Qiagen) and stored at 4 °C in TE buffer until use.

2.2. HOX cluster sequence of Indonesian coelacanth

We downloaded the currently available genomic sequence (~1.6 Mb) for Indonesian coelacanth (*L. menadoensis*) from GenBank (ID: 220898172, 220898186, 220898198, 220898210) at NCBI (Amemiya et al., 2010).

2.3. Genome sequencing

We prepared 150-bp insert paired-end libraries and single-end libraries and sequenced the samples using the Illumina genome analyzer II (GAII) according to the manufacturer's instructions. Two runs were performed. In total, 553.1 million 76-bp paired-end reads and 24.6 million 76-bp single-end reads of *L. chalumnae* genome fragments were collected. The accession number for the sequence data of this study is SRP011573.

2.4. Mapping and identification of repetitive sequences

All reads were aligned to the HOX cluster sequences of *L. menadoensis* using Burrows–Wheeler Alignment software (Li and Durbin, 2009). Each alignment was assigned a mapping quality score by Burrows–Wheeler Alignment, which is the Phred score that the alignment is incorrect. The PCR amplification step will lead to the sequencing of identical DNA fragments. Not removing these PCR duplicates can lead to the miscalling of single nucleotide variants (SNVs) by overrepresentation of one allele. This was corrected by a quality control step to remove these potential PCR duplicates with SAMtools software (Li et al., 2009). Known repetitive elements were defined by RepeatMasker (http://www.repeatmasker.org/), and novel repetitive elements were identified by pairwise alignment with the lastz program (Schwartz et al., 2003).

2.5. Calling of SNVs and short insertions/deletions (indels)

After all the reads were aligned to the reference genome using Burrows–Wheeler Alignment (Li and Durbin, 2009), we used the SAMtools to produce a consensus genotype for each genomic position. SNVs and indels were then identified based on the differences between the consensus genotype of *L. chalumnae* and the *L. menadoensis* allele at that position. These SNVs and indels were then filtered by the SAMtools variation filter, changing the minimum and maximum read depth parameter to call variants from its default values (3 and 100) to 6 and 39, to prevent the inclusion of SNVs and indels at repetitive regions. The lists of SNVs/indels were then annotated by custom Perl scripts with the SQLite database, which was specifically designed to annotate the identified SNVs/indels using information from GenBank.

2.6. Validation of SNVs

To validate the ambiguous SNVs with low conQ values, we performed PCR and direct sequencing. The PCR protocol consisted of 30 cycles with denaturation at 94 °C for 30 s, annealing at 55 °C for 45 s, and extension at 72 °C for 1 min. The PCR mixture contained 2.5 U Ex Taq polymerase™ (Takara), 1× Ex Taq buffer, 0.4 mM dNTPs, 0.1 μM of each primer, and 10 ng of template genomic DNAs in a final volume 20 μl. PCR products were confirmed by electrophoresis in a 3.0% agarose gel (Takara) and stained with ethidium bromide. The PCR products were then purified through precipitation with isopropanol. Purified PCR products were used for direct sequencing, with 25 cycles of denaturation at 96 °C for 30 s, annealing at 50 °C for 15 s, and extension at 60 °C for 1 min. Reactions contained 1 μl BigDye® ver. 3.1 terminator premix (Applied Biosystems), 1× sequencing buffer (Applied Biosystems), 1 μM sequence primer, and 2 μl of purified PCR product in a final volume of 5 μl. Sequences were determined using an automated sequencer (Applied Biosystems, model 3130).

2.7. Evolutionary analyses

The sequences were edited by GENETYX-Windows version 10 (GENETYX). Clustal W (Thompson et al., 1994) was used to align deduced amino acid sequences of HOX coding sequences of the two coelacanths, human, chimpanzee, mouse, and rat. For nucleotide sequence comparison, CodonAlign 2.0 (http://www.sinauer.com/hall/2e/) was used to introduce gaps into HOX coding sequences at the positions corresponding to the gaps in the aligned protein sequences. MEGA 5.0 software (Tamura et al., 2011) was used for genetic distance calculation for aligned HOX coding nucleotide sequences. The values of the synonymous divergence between mouse–rat, human–chimp and two coelacanth pairs were estimated and plotted for each HOX gene separately. Because the synonymous divergences were 0 in some comparisons due to the absence of no synonymous substitution, the ratios of dN/dS were estimated by concatenating the sequences of all HOX genes used for the analysis.

2.8. SINE insertion analysis

To confirm the insertion of SINEs in the coelacanths in the western Indian ocean (Tanzania and Comoros), we examined 16 loci, into which LF SINEs (Bejerano et al., 2006) were apparently inserted in the HOX cluster sequence of Tanzanian coelacanth. PCR primers were designed to amplify the SINE unit and its flanking genomic region using Primer3 (Rozen and Skaletsky, 2000). The primer sequences are summarized in Table S1. The PCR protocol was mostly the same as that shown in "Validation of SNVs" (2.6).

3. Results and discussion

3.1. Whole-genome sequencing and mapping

We performed two sequencing runs on the Illumina Genome Analyzer II platform and produced 85.9 Gbp of sequence data (Table 1). The short reads were aligned with the Burrows–Wheeler Alignment tool (Li and Durbin, 2009); we mapped 2.8% of the total reads to the *L. menadoensis* HOX cluster sequences, and 1.6% of the total reads were uniquely mapped. Insert size distribution showed not only major broad peaks around 150 bp but also a minor peak around 320 bp. This minor peak was created by the reads derived from repetitive sequences (Fig. 1; discussed in Section 3.3).

3.2. Repetitive elements and coverage

We also examined the distribution of repetitive elements defined by RepeatMasker, estimating a rate of repetitive element occupancy of ~13.1% (211,850 bp) in the HOX cluster sequences. Coverage of HOX clusters differed widely from region to region (Fig. 2). Especially, repetitive regions showed high coverage because all reads that were derived from repetitive sequences of the coelacanth genome were mapped to repetitive elements within that HOX cluster (Fig. 3). On the other hand, non-repetitive exonic regions showed a mean coverage of 22.2 (±5.9 s.d.) (Fig. 3A). These results indicate that the sequences of the exons of HOX cluster genes are probably unique in the coelacanth genome because the genome size estimated from the coverage of those exons was consistent with the size estimated from the weight of nuclear DNA content (~3 Gbp) (Makapedua et al., 2011).

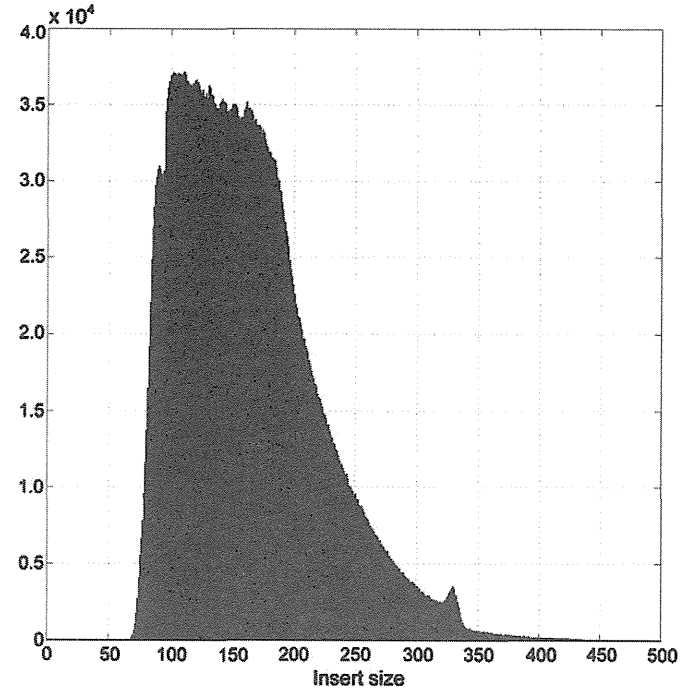

Fig. 1. Insert size distribution of paired-end reads. Insert size distribution showed two peaks, of which the minor peak at ~320 bp was derived from reads from repetitive sequences.

Table 1
Data production and mapping results for the coelacanth genome.

Data type	Number of reads	Number of mapped reads	Total bases (Mb)	Mapped bases (Mb)	Percentage with unique placement
SE	24,649,274	471,070	1873	36	84.18
PE	553,079,477	15,690,405	84,068	2385	55.93
Total	577,728,751	16,161,475	85,941	2421	56.75

Single-end (SE) and paired-end (PE) sequencing reads were aligned onto the HOX cluster sequence of *L. menadoensis* (~1.6 Mbp). 'Unique placement' means a read had only one best placement.

3.3. Identification of novel coelacanth-specific repetitive sequences

Although we excluded known repetitive elements using RepeatMasker, ~20% of intergenic sequences still had high coverage (blue line in Fig. 3C). We extracted those sequences and compared them with themselves. A dot plot provided evidence for the existence of other repetitive elements in the coelacanth genomes (Fig. 4). These elements were 2336 bp (c1) and 7736 bp (c2) in length. Although we did not characterize c1, we identified c2 to be Harbinger transposons, which were reported recently by Smith et al. (2012). As shown in Fig. 2, paired-end reads with an insert size of ~320 bp colocalized with previously identified coelacanth-specific repeats (red boxes) as well as with novel repeats identified in this study (green boxes). Their abundance suggests that these sequences are active mobile elements in the coelacanth genome. This notion is consistent with the study by Smith et al. (2012), in which they showed that the Harbinger transposons are still active *in vivo*.

3.4. Identifying SNVs and small indels

Assuming that sequences of the exons of *HOX* genes are unique (non-repetitive) in the coelacanth genome, we can define the unique region using the coverage of these exons as a reference. According to the read depth of the exons, we defined the regions with a mean exon depth ±3 s.d. as unique. After excluding extremely low and high read coverage regions, genetic differences between our sequenced *L. chalumnae* genome and the *L. menadoensis* genome were identified using modified settings in SAMtools (Li et al., 2009). We identified

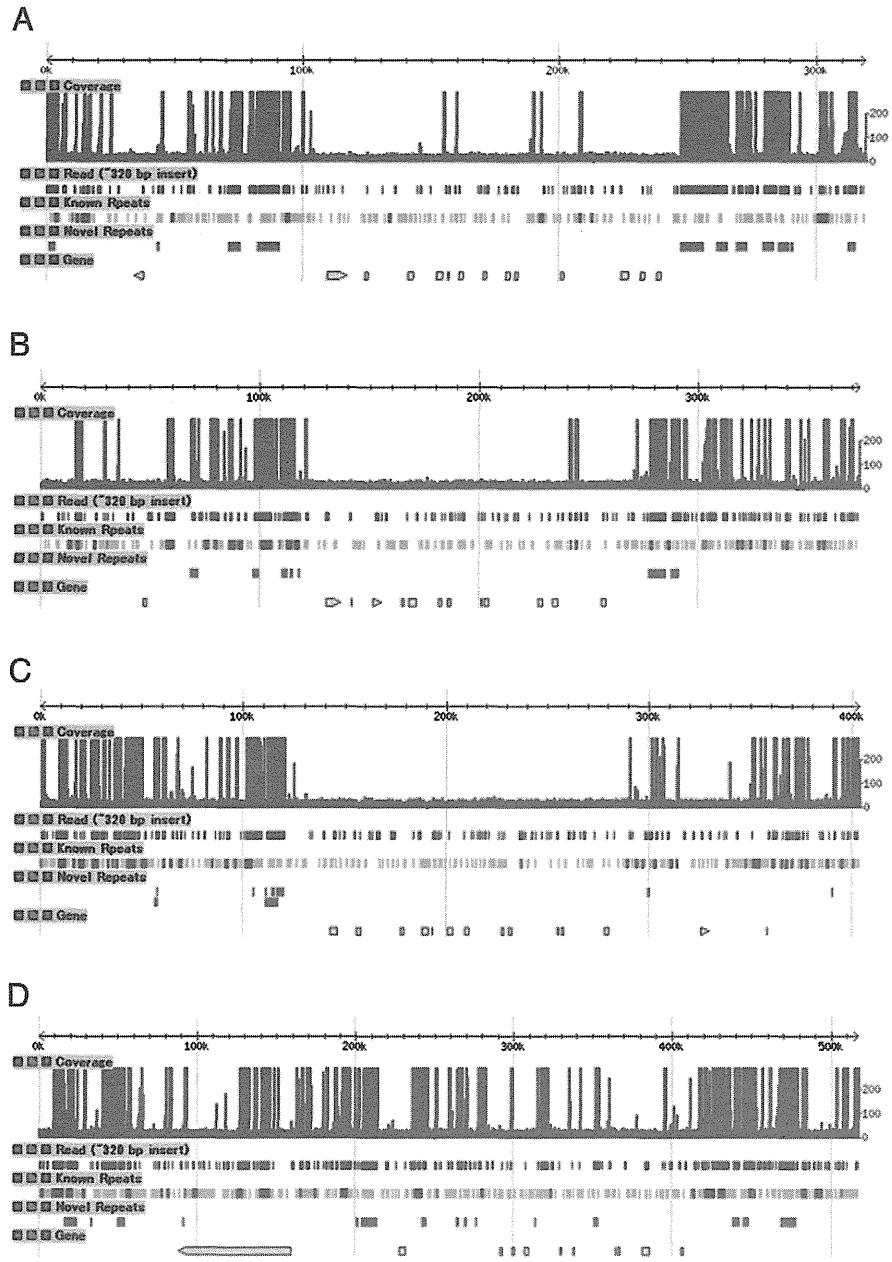

Fig. 2. Overview of HOX cluster regions. From top to bottom track, coverage of short reads, paired-end reads with ~320-bp insert size, known repetitive elements, novel repetitive elements identified in this study, and genes are shown for the HoxA (A), HoxB (B), HoxC (C), and HoxD (D) clusters. Red boxes in the known repetitive element track show previously identified coelacanth-specific repetitive sequences.

2280 SNVs and 837 indels, in which 18, 273, and 2826 variants were located in exons, introns, and intergenic regions, respectively (Table 2).

3.5. Validation of SNVs in exons

Although we filtered the possible errors during the process of variant calling, some of the SNVs shown in Table 2 were still ambiguous in terms of the low-quality scores (about 25 in conQ). To more accurately estimate the subsequent genetic divergence of HOX coding exons between the two coelacanth species, we needed to validate these SNVs to eliminate possible misidentified SNVs. Therefore, we amplified such genomic regions by PCR using the genomic DNA of the juvenile coelacanth, TCC041-1, as the template. Then, we directly determined the sequences of these PCR products by Sanger sequencing. We found that one heterozygous SNV and one homozygous indel shown in Table 2 were absent in the sequence, indicating that they were misidentified variations not eliminated during the filtering process. Accordingly, we used the modified sequences for further analyses. The modified number of SNVs and indels in coding exons is shown in parentheses (Table 2).

3.6. Estimation and comparison of synonymous divergence

Next, we estimated the genetic divergence at synonymous sites in coelacanths by focusing on whether the substitution rate is slow even at these sites. To examine this possibility, we compared the synonymous divergence of the two coelacanth species with that of the human–chimpanzee pair and that of the mouse–rat pair. Forty orthologous *HOX* genes (including two *EVX* genes) were common in all six species, whereas HOXA14, HOXB10, HOXC1, and HOXC3 were absent in mammals, and HOXD9 was absent in coelacanths.

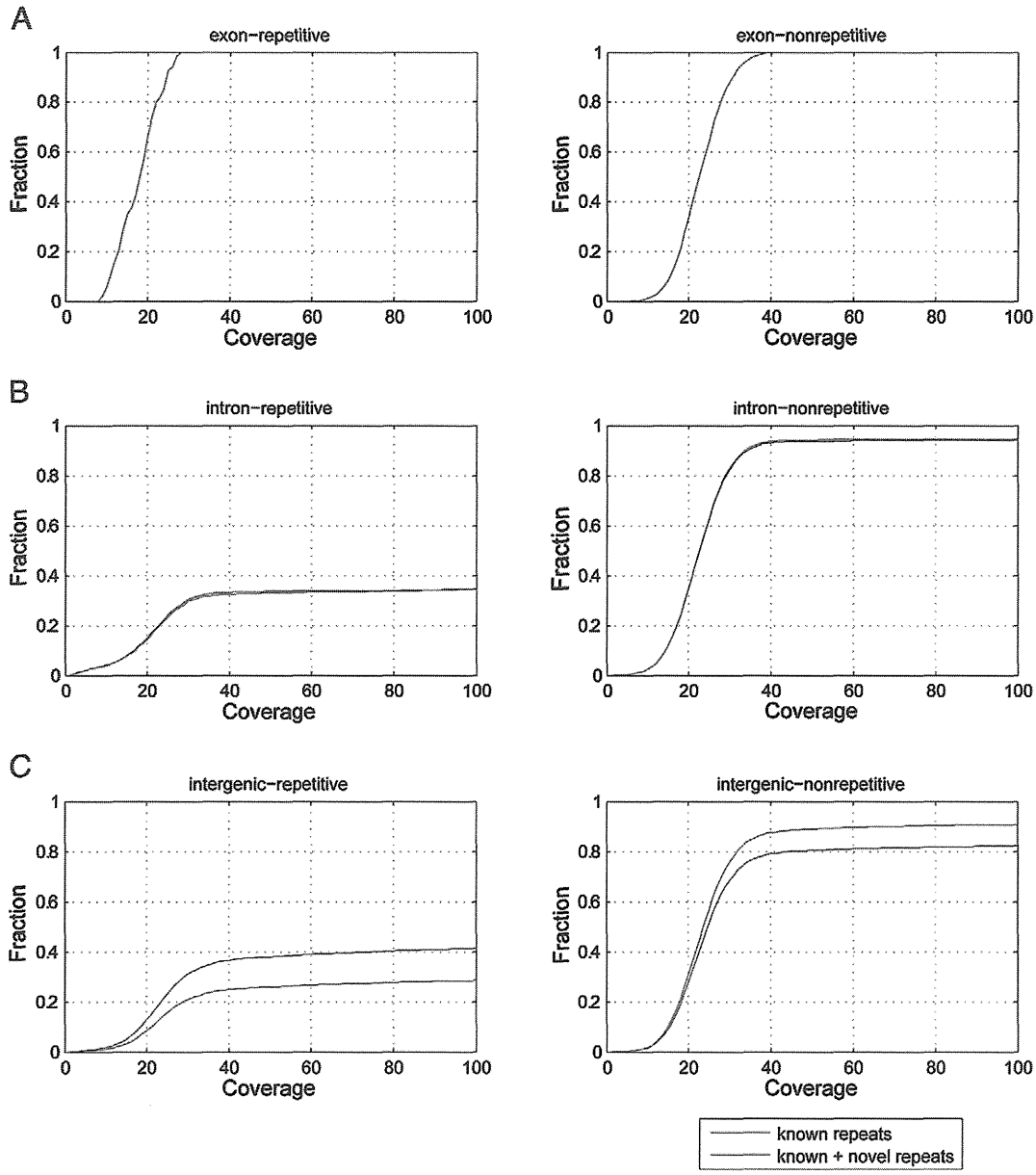

Fig. 3. Cumulative distribution of HOX cluster coverage. Cumulative distribution of read coverage in exons (A), introns (B), and intergenic (C) regions was plotted separately for repetitive (left) and non-repetitive (right) regions. Repetitive sequences were defined by RepeatMasker. Twenty percent of non-repetitive intergenic regions have extremely high coverage (blue line in C), of which 50% were derived from novel repetitive elements identified in this study (red line in C), but the remaining was unknown.

Furthermore, the putative cDNA sequences of rat HOXD8 and HOXD11 were incomplete because of the low quality of the genome sequence data for the corresponding region. Accordingly, we removed them from the analysis. In total, we used the 38 *HOX* gene orthologs for the subsequent analyses to compare the same genomic region among six species. The protein sequences were first aligned using Clustal W implemented in MEGA 5 with the default settings. Then, gaps were introduced into the cDNA sequences based on the protein sequence alignment. The aligned nucleotide sequences were used to estimate genetic divergence using MEGA 5. Fig. 5 shows the plots of the uncorrected synonymous divergence (y axis) against the corrected nucleotide divergence (x axis) of 38 *HOX* genes for each of the three species pairs (human–chimp, mouse–rat, and Tanzania–Indonesia coelacanths), indicating no signs of saturation of synonymous substitutions. Accordingly, we could compare the synonymous divergence among the three species pairs without regard to the possibility of saturation.

Table 3 shows the resultant averaged estimated synonymous divergence of *HOX* genes for each of the three species pairs. In the human–chimp and mouse–rat pairs, the synonymous divergence was 0.78% and 7.5%, respectively. Although these values are smaller than those estimated at the whole-genome level (human–chimp 1.1%, Chen et al., 2001; mouse–rat 19%, Rat Genome Sequencing Project Consortium, 2004), they are consistent with previous studies in which the authors indicated that *HOX* genes were conserved even at synonymous sites (Lin et al., 2008; Woltering and Duboule, 2009). The most intriguing result is that the synonymous divergence was much smaller in the coelacanth species pair; the value was 0.07%, which is about 11-fold smaller than that of the human–chimp pair.

When we consider that the divergence time of human–chimp and the two coelacanth species is almost the same (see Section 3.7), the difference in the synonymous divergence between these groups was statistically significant ($p<10^{-14}$ by two-sample proportion z-test, Snedecor and Cochran, 1989).

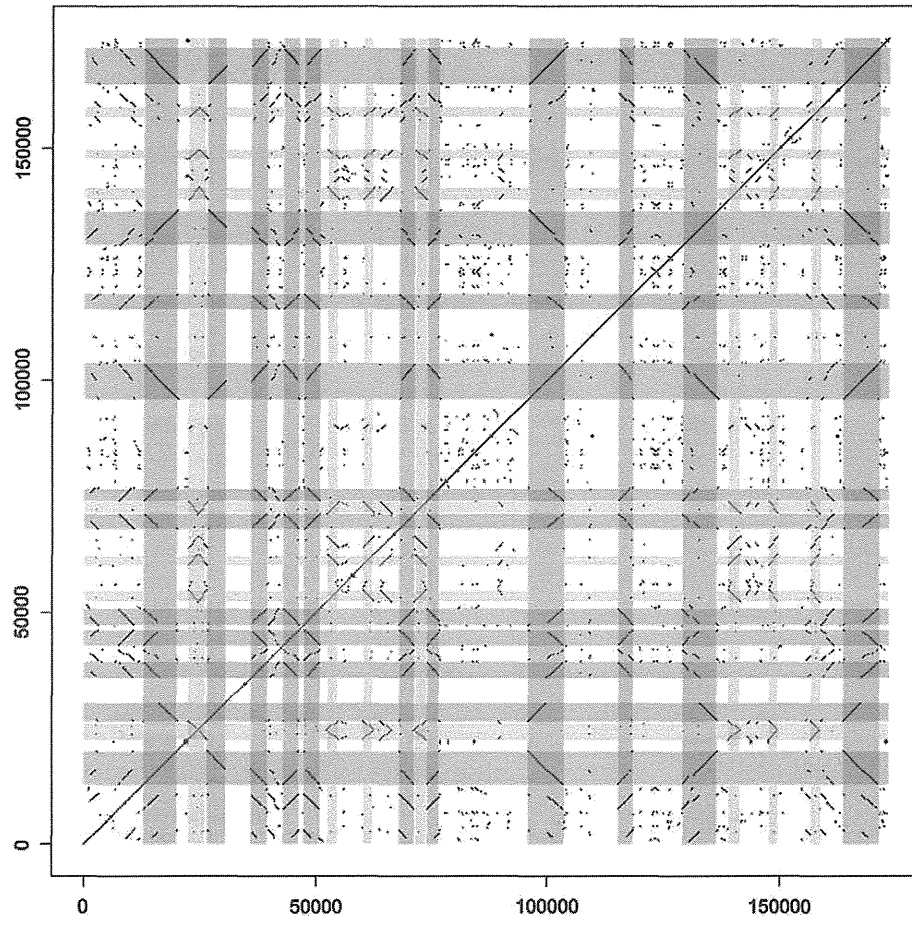

Fig. 4. Dot plot of HOX cluster region. After excluding known repetitive sequences identified by RepeatMasker, the concatenated remaining sequence was compared with itself by lastz (Schwartz et al., 2003). Green and red shadows indicate two newly identified repetitive sequences specific to coelacanth.

3.7. Estimation and comparison of synonymous substitution rates

To directly compare the absolute values of the synonymous substitution rate among the three species pairs, we interpolated the divergence times estimated by the independent studies. We used 6 MYA and 16 MYA for the divergence of human–chimp (Glazko and Nei 2003) and mouse–rat (Hasegawa et al., 2003) pairs, respectively. For coelacanths, there are two main alternative hypotheses for their divergence time (6 MYA and 30 MYA), which have not been validated (Holder et al., 1999; Inoue et al., 2005; Sudarto et al., 2010). The estimated value of 6 MYA by Holder et al. (1999) appears to be less reliable than that of 30 MYA by Inoue et al. (2005) because the former study analyzed just partial mitochondrial sequences and applied the substitution rate of tetrapods. However, we here used both 6 MYA and 30 MYA to see the difference between the resultant values. The synonymous substitution rates of human–chimp and mouse–rat were calculated to be 0.65×10^{-9} and 2.33×10^{-9} (per site per year), respectively (Table 3). The higher substitution rate in rodents relative to that of primates is consistent with previous studies (Kitano and Saitou, 2005; Li et al., 1987). As for the coelacanths, when we applied the divergence time of 6 MYA, the substitution rate was calibrated to be 0.059×10^{-9} (per site per year, Table 3), which is 11-fold slower than that of human–chimp. Furthermore, the divergence time of 30 MYA, which appears to be more reliable dating, led to the significantly slow substitution rate of 0.0012×10^{-9} (per site per year, Table 3). This value is 56-fold slower than that of human–chimp.

Table 2
SNVs and indels in unique (non-repetitive) regions.

Locus	SNV		Indel		Total	Percentage	Length of region (bp)
	Homo	Hetero	Homo	Hetero			
Exon	15	1 (0)	2 (1)	0	18 (16)	0.0445% (0.0395%)	40,471
Intron	147	15	104	7	273	0.2373%	115,045
Intergenic	1444	658	695	29	2826	0.2514%	1,124,218
Total	1606	674	801	36	3117	0.2515%	1,239,263

The modified number of SNVs and indels in coding exons after validation is shown in parentheses (see Section 3.5).

3.8. SINE insertion polymorphisms

To test whether the insertion of transposable elements occurred after the split of *L. chalumnae* and *L. menadoensis*, we searched the alignments of the four HOX clusters for indels between these two species. Six indels of >50 nt were identified, in which the sequences were found in *L. menadoensis* but were absent in *L. chalumnae*. However, these differences did not result from insertions of transposable elements (data not shown). Specifically, three of the six differences were copy number variations of tandem duplications of a ~60-nt unit, and two of them were due to partial deletions of Harbinger-like transposons in *L. chalumnae*. The remaining sequence showed no homology with known transposable elements. This result suggests that there is no difference in the presence/absence pattern of transposable elements in the HOX clusters. However, owing to the mapping strategy of short read fragments, it is still possible that the polymorphisms of the short-length retroposons such as SINEs were actually absent. To examine such a possibility, we examined 16 loci, into which LF SINEs (Bejerano et al., 2006) were apparently inserted in *L. chalumnae*.

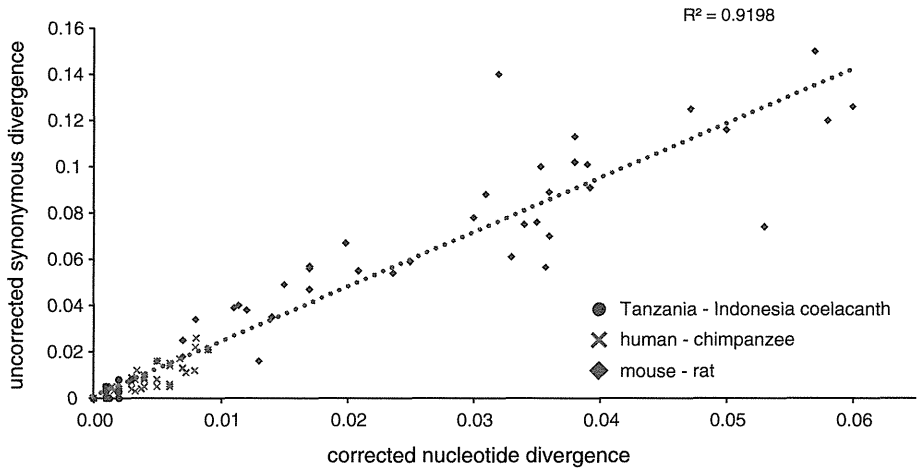

Fig. 5. Saturation plot for the synonymous divergence of *HOX* genes of three species pairs. The uncorrected synonymous divergence from the pairwise comparison plotted against the pairwise corrected nucleotide divergence. The synonymous divergence increases almost linearly, indicating that the saturation is apparently absent in this comparison. The uncorrected synonymous divergence was estimated using the modified Nei–Gojobori method, and the corrected nucleotide divergence was estimated with the maximum composite likelihood method.

The 16 SINEs were chosen to represent recently inserted copies based on the divergence from the consensus sequence. Then, we performed SINE-flanking PCR to investigate the presence or absence of a SINE at each locus. If a SINE is absent at a particular locus, the PCR band becomes shorter than expected (Okada et al., 2004). We detected SINE insertions in all of the loci examined (Fig. 6), indicating very few, if any, SINE insertion polymorphisms between *L. chalumnae* and *L. menadoensis*. Accordingly, low genetic diversity of the HOX cluster region between the two coelacanth species was further indicated in terms of SINE insertions.

3.9. Reduction of substitution rate, purifying selection, or recent divergence?

An important question is why the genetic divergence of the coelacanth HOX cluster is so small, even at the synonymous sites. There are three possible explanations: (1) strong purifying selection in the open reading frames of *HOX* genes, including the synonymous sites, (2) reduction of the nucleotide substitution rate in the coelacanth lineage, and (3) exceptionally more recent divergence between *L. chalumnae* and *L. menadoensis* than expected from the mtDNA analyses. Regarding the first explanation, Lin et al. (2008) indicated that the mammalian *HOX* genes have ultraconserved coding regions, in which the sequences are highly conserved at the nucleotide level. Their study implied that the ultraconserved coding regions have some gene regulatory functions, which led to purifying selection without regard to synonymous and non-synonymous sites. Accordingly, it is possible that the smaller genetic divergence in coelacanth *HOX* genes may reflect substantially stronger purifying selection compared with that in rodents and primates. However, the dN/dS ratios of *HOX* genes among human–chimp, rodents and two coelacanth pairs were 0.23, 0.42 and 0.39, respectively, indicating that the purifying selection in coelacanth was not strong in comparison with the other groups. Accordingly, the smaller synonymous divergence in the coelacanth *HOX* genes may not be caused by the operation of stronger purifying selection. Regarding the second possibility, an exceptionally slow substitution rate has been reported in the mtDNA of shark, which is 7- to 8-fold slower than that of primates or ungulates (Martin et al., 1992). Given the similarity of coelacanths to sharks in their ecology, large body size, and long generation time, it is plausible that coelacanths also exhibit a slow nucleotide substitution rate in their genomes similar to that observed in sharks. The small genetic divergence even at introns and intergenic regions (0.25%, Table 2) of coelacanth *Hox* genes and the apparent absence of SINE insertion polymorphisms, are consistent with this possibility. However, there are also arguments for the third possibility, namely, the more recent divergence time of *L. chalumnae* and *L. menadoensis*. Because coelacanths have been a single lineage since the Devonian era, they do not have closely related taxa, making it difficult to fix the calibration point for divergence time estimation (Sudarto et al., 2010). Although there is a large difference in the estimates between the two major hypotheses, even the later estimate of 6 MYA leads to an 11-fold slower substitution rate in coelacanths. However, if the two coelacanth species diverged much more recently than what has been estimated, the conclusion would be drastically changed. Indeed, Schartl et al. (2005) raised the possibility that the western Indian Ocean could have been colonized recently by coelacanth drifters from the Pacific Province by Indonesian throughflow, which has probably existed for only 3 to 4 MY. Gordon (1998) also pointed out the presence of an oceanographic connection between Sulawesi and the Comoran Islands region that has enabled very recent gene flow from Indonesia to Comoros. Furthermore, if such dispersal occurred only in male individuals, the genetic divergence should be smaller in the nuclear genome than in the mitochondrial genome. This may provide us with a compromise resolution for the exceptionally smaller genetic divergence in the HOX cluster of coelacanths compared with the divergence time based on mtDNA.

Table 3
Estimated genetic divergences and substitution rates at synonymous sites.

	Divergence time from literature	Estimated synonymous divergence	Estimated synonymous substitution rate (per site per year)
Human–chimp	6 MYA[a]	0.0078 (0.0012)[b]	$0.65\ (0.10)\times10^{-9}$
Tanzanian–Indonesian coelacanths	6 MYA[c]	0.00070 (0.00011)	$0.059\ (0.0090)\times10^{-9}$
	30 MYA[d]		$0.012\ (0.0018)\times10^{-9}$
Mouse–rat	16 MYA[e]	0.075 (0.012)	$2.33\ (0.37)\times10^{-9}$

[a] Inoue et al. (2005).
[b] The values in parentheses indicate the standard error.
[c] Holder et al. (1999).
[d] Glazko and Nei (2003) and Sudarto et al. (2010).
[e] Hasegawa et al. (2003).

3.10. Conclusion

In the present study, we performed a large-scale comparison of the HOX cluster sequence between two coelacanth species, showing a significantly small genetic divergence, even at silent sites, relative

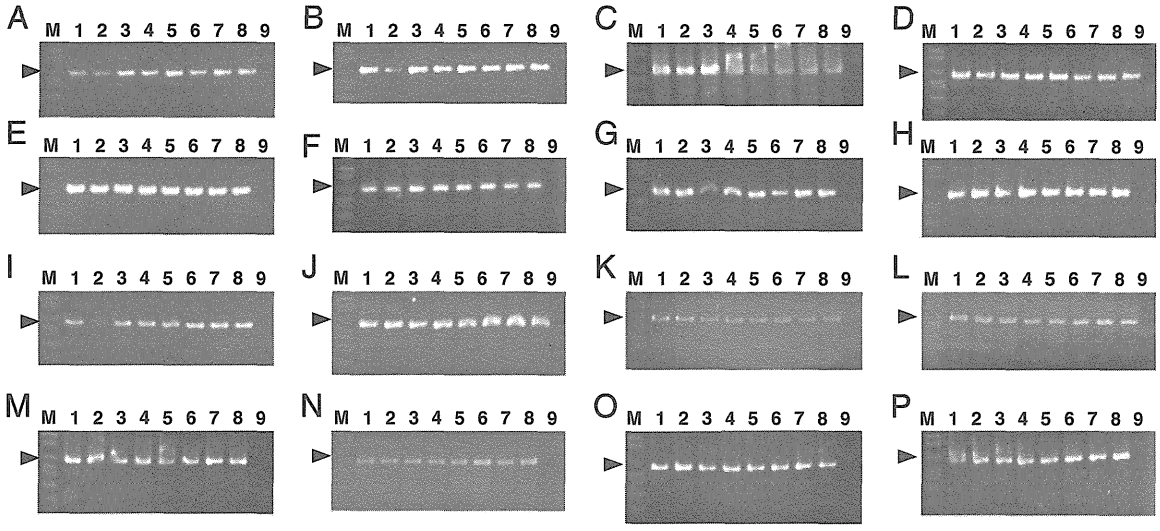

Fig. 6. SINE insertion analysis for the 16 loci of coelacanths in the western Indian Ocean. (A–P) Agarose gel electrophoresis of SINE-flanking PCR products from locus 1 to 16. Lane M indicates the size marker, and lanes 1–9 indicate the *L. chalumnae* samples (ID: Comoro_1, Comoro_2, TCC039, TCC040, TCC041, TCC042, TCC043, TCC044, and negative control). The arrowheads indicate the expected size of the PCR products, which were deduced from the sequence of *L. chalumnae*. In all loci, PCR bands were detected at ~600 bp, which is expected for the presence of a SINE.

to human–chimp and mouse–rat pairs. When we applied the divergence times deduced by the mtDNA sequence data, the substitution rate was calibrated to be significantly slower than that of the other vertebrates reported. To explain the phenomenon, we propose two alternative possibilities, which are not necessarily mutually exclusive: reduction of the nucleotide substitution rate in coelacanth lineages or an unexpectedly recent divergence of the two coelacanth species. Although it is difficult to settle this issue at present, whole genome-wide analysis and large-scale population genetic analyses, which will become available soon, may shed light on the subject. Such genome-wide analyses may eventually elucidate why coelacanths could be evolutionary relics from the Devonian time.

Acknowledgments

This work was supported by the JSPS AA Science Platform Program, a Grant-in-Aid for Scientific Research (S) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (to N.O.), and the Global COE Program "Deciphering Biosphere from Genome Big Bang" (to S.M.).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2012.05.047.

References

Amemiya, C.T., Powers, T.P., Prohaska, S.J., Grimwood, J., Schmutz, J., Dickson, M., Miyake, T., Schoenborn, M.A., Myers, R.M., Ruddle, F.H., Stadler, P.F., 2010. Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. Proc. Natl. Acad. Sci. U. S. A. 107, 3622–3627.
Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J., Haussler, D., 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441, 87–90.
Carroll, R.L., 1988. Vertebrate Paleontology and Evolution. H. Freeman and Co., New York.
Chen, F.C., Vallender, E.J., Wang, H., Tzeng, C.S., Li, W.H., 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. J. Hered. 92, 481–489.
De Vos, L., Oyugi, D., 2002. First capture of a coelacanth, *Latimeria chalumnae* Smith, 1939 (Pisces: Latimeriidae), off Kenya. S. Afr. J. Sci. 98, 345–347.
Erdman, M.V., Caldwell, R.L., Moosa, M.K., 1998. Indonesian 'king of the sea' discovered. Nature 395, 335.
Glazko, G.V., Nei, M., 2003. Estimation of divergence times for major lineages of primate species. Mol. Biol. Evol. 20, 424–434.
Gordon, A.L., 1998. Coelacanth populations may go with the flow. Nature 395, 634.
Hasegawa, M., Thorne, J.L., Kishino, H., 2003. Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. Genes Genet. Syst. 78, 267–283.
Heemstra, P.C., Freeman, A.L., Wong, H.Y., Hensley, D.A., Rabesandratana, H.D., 1996. First authentic capture of a coelacanth, *Latimeria chalumnae* (Pisces: Latimeriidae), off Madagascar. S. Afr. J. Sci. 92, 150–151.
Holder, M.T., Erdmann, M.V., Wilcox, T.P., Caldwell, R.L., Hillis, D.M., 1999. Two living species of coelacanths? Proc. Natl. Acad. Sci. U. S. A. 96, 12616–12620.
Inoue, J.G., Miya, M., Venkatesh, B., Nishida, M., 2005. The mitochondrial genome of Indonesian coelacanth *Latimeria menadoensis* (Sarcopterygii: Coelacanthiformes) and divergence time estimation between the two coelacanths. Gene 349, 227–235.
Kitano, T., Saitou, N., 2005. Evolutionary conservation of 5′ upstream sequence of nine genes between human and great apes. Genes Genet. Syst. 80, 225–232.
Krumlauf, R., 1994. *Hox* genes in vertebrate development. Cell 78, 191–201.
Kurosawa, G., Takamatsu, N., Takahashi, M., Sumitomo, M., Sanaka, E., Yamada, K., Nishii, K., Matsuda, M., Asakawa, S., Ishiguro, H., Miura, K., Kurosawa, Y., Shimizu, N., Kohara, Y., Hori, H., 2006. Organization and structure of hox gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. Gene 370, 75–82.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.
Li, W.H., Tanimura, M., Sharp, P.M., 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J. Mol. Evol. 25, 330–342.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.
Lin, Z., Ma, H., Nei, M., 2008. Ultraconserved coding regions outside the homeobox of mammalian Hox genes. BMC Evol. Biol. 8, 260.
Maisey, J.G., 1996. Discovering Fossil Fishes. Henry Holt, New York.
Makapedua, D.M., Barucca, M., Forconi, M., Antonucci, N., Bizzaro, D., Amici, A., Carradori, M.R., Olmo, E., Canapa, A., 2011. Genome size, GC percentage and 5mC level in the Indonesian coelacanth *Latimeria menadoensis*. Mar. Genomics 4, 167–172.
Martin, A.P., Naylor, G.J., Palumbi, S.R., 1992. Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. Nature 357, 153–155.
Meyer, A., Málaga-Trillo, E., 1999. Vertebrate genomics: more fishy tales about *Hox* genes. Curr. Biol. 9, R210–R213.
Nikaido, M., Sasaki, T., Emerson, J.J., Aibara, M., Mzighani, S.I., Budeba, Y.L., Ngatunga, B.P., Iwata, M., Abe, Y., Li, W.H., Okada, N., 2011. Genetically distinct coelacanth population off the northern Tanzanian coast. Proc. Natl. Acad. Sci. U. S. A. 108, 18009–18013.
Noonan, J.P., Grimwood, J., Danke, J., Schmutz, J., Dickson, M., Amemiya, C.T., Myers, R.M., 2004. Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. Genome Res. 14, 2397–2405.
Okada, N., Shedlock, A.M., Nikaido, M., 2004. Retroposon mapping in molecular systematics. Methods Mol. Biol. 260, 189–226.
Rat Genome Sequencing Project Consortium, 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493–521.
Rozen, S., Skaletsky, H.J., 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz, S., Misener, S. (Eds.), Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 365–386.
Schartl, M., Hornung, U., Hissmann, K., Schauer, J., Fricke, H., 2005. Genetics: relatedness among east African coelacanths. Nature 435, 901.
Schliewen, U., Fricke, H., Schartl, M., Epplen, J.T., Pääbo, S., 1993. Which home for coelacanth? Nature 363, 405.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., Miller, W., 2003. Human–mouse alignments with BLASTZ. Genome Res. 13, 103–107.
Smith, J.L.B., 1939. A living fish of Mesozonic type. Nature 143, 455–456.
Smith, J.L.B., 1953. The second Coelacanth. Nature 171, 99–101.
Smith, J.J., Sumiyama, K., Amemiya, C.T., 2012. A living fossil in the genome of a living fossil: harbinger transposons in the coelacanth genome. Mol. Biol. Evol. 29, 985–993.
Snedecor, G.W., Cochran, W.G., 1989. Statistical Methods, 8 edn. Iowa State University Press.
Sudarto, Lalu, X.C., Kosen, J.D., Tjakrawidjaja, A.H., Kusumah, R.V., Sadhotomo, B., Kadarusman, Pouyaud, L., Slembrouck, J., Paradis, E., 2010. Mitochondrial genomic divergence in coelacanths (*Latimeria*): slow rate of evolution or recent speciation? Mar. Biol. 157, 2253–2262.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Thompson, J.D., Higgins, D.G., Gibson, T.J., 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.
Woltering, J.M., Duboule, D., 2009. Conserved elements within open reading frames of mammalian Hox genes. J. Biol. 8, 17.

OPEN ACCESS Freely available online

PLOS | GENETICS

Functional Variants in *NFKBIE* and *RTKN2* Involved in Activation of the NF-κB Pathway Are Associated with Rheumatoid Arthritis in Japanese

Keiko Myouzen[1], Yuta Kochi[1,2]*, Yukinori Okada[1,2,3], Chikashi Terao[4,5], Akari Suzuki[1], Katsunori Ikari[6], Tatsuhiko Tsunoda[7], Atsushi Takahashi[3], Michiaki Kubo[8], Atsuo Taniguchi[6], Fumihiko Matsuda[4,9,10], Koichiro Ohmura[5], Shigeki Momohara[6], Tsuneyo Mimori[5], Hisashi Yamanaka[6], Naoyuki Kamatani[11], Ryo Yamada[12], Yusuke Nakamura[13], Kazuhiko Yamamoto[1,2]

1 Laboratory for Autoimmune Diseases, Center for Genomic Medicine (CGM), RIKEN, Yokohama, Japan, **2** Department of Allergy and Rheumatology, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan, **3** Laboratory for Statistical Analysis, CGM, RIKEN, Yokohama, Japan, **4** Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan, **5** Department of Rheumatology and Clinical Immunology, Graduate School of Medicine, Kyoto University, Kyoto, Japan, **6** Institute of Rheumatology, Tokyo Women's Medical University, Tokyo, Japan, **7** Laboratory for Medical Informatics, CGM, RIKEN, Yokohama, Japan, **8** Laboratory for Genotyping Development, CGM, RIKEN, Yokohama, Japan, **9** CREST Program, Japan Science and Technology Agency, Kawaguchi, Saitama, Japan, **10** Institut National de la Santé et de la Recherche Médicale (INSERM), Unité U852, Kyoto University Graduate School of Medicine, Kyoto, Japan, **11** Laboratory for International Alliance, CGM, RIKEN, Yokohama, Japan, **12** Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan, **13** Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan

Abstract

Rheumatoid arthritis is an autoimmune disease with a complex etiology, leading to inflammation of synovial tissue and joint destruction. Through a genome-wide association study (GWAS) and two replication studies in the Japanese population (7,907 cases and 35,362 controls), we identified two gene loci associated with rheumatoid arthritis susceptibility (*NFKBIE* at 6p21.1, rs2233434, odds ratio (OR) = 1.20, $P = 1.3 \times 10^{-15}$; *RTKN2* at 10q21.2, rs3125734, OR = 1.20, $P = 4.6 \times 10^{-9}$). In addition to two functional non-synonymous SNPs in *NFKBIE*, we identified candidate causal SNPs with regulatory potential in *NFKBIE* and *RTKN2* gene regions by integrating *in silico* analysis using public genome databases and subsequent *in vitro* analysis. Both of these genes are known to regulate the NF-κB pathway, and the risk alleles of the genes were implicated in the enhancement of NF-κB activity in our analyses. These results suggest that the NF-κB pathway plays a role in pathogenesis and would be a rational target for treatment of rheumatoid arthritis.

Citation: Myouzen K, Kochi Y, Okada Y, Terao C, Suzuki A, et al. (2012) Functional Variants in *NFKBIE* and *RTKN2* Involved in Activation of the NF-κB Pathway Are Associated with Rheumatoid Arthritis in Japanese. PLoS Genet 8(9): e1002949. doi:10.1371/journal.pgen.1002949

Editor: Panos Deloukas, The Wellcome Trust Sanger Institute, United Kingdom

Received March 31, 2012; **Accepted** July 12, 2012; **Published** September 13, 2012

Funding: This work was conducted as a part of the BioBank Japan Project that was supported by the Ministry of Education, Culture, Sports, Sciences, and Technology of the Japanese government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: ykochi@src.riken.jp

Introduction

Rheumatoid arthritis (RA [MIM 180300]) is an autoimmune disease [1] with a complex etiology involving several genetic factors as well as environmental factors. Previous genome-wide association studies (GWAS) for RA have discovered many genetic loci [2–6], although the causal mechanisms linking the variants in these loci and disease etiology are largely unknown, except for in a few cases [6–8]. In contrast to mutations in Mendelian, monogenic diseases, most disease-associated variants in complex diseases, including autoimmune diseases, have moderate effects on disease susceptibility. This is because the disease causal variants in complex diseases are thought to have moderate effects on gene function, while amino acid changes introduced by the mutations of monogenic diseases have critical impacts on protein function [9]. Moreover, it has been demonstrated that the majority of autoimmune disease loci are expression quantitative trait loci (eQTLs) [10,11], indicating that accumulation of quantitative, but not qualitative, changes in gene function likely predisposes individuals to the disease. This renders it difficult to pinpoint the causal variants in the GWAS loci, especially in eQTLs, because all the variations in strong linkage disequilibrium (LD) with the marker SNP in a GWAS, the majority of which are not covered by the SNP array, are possible candidates for causal variants.

In recent years, with the emergence of next-generation sequencing technologies, the way we approach disease-causing variants has dramatically changed. First, a comprehensive map of human genetic variations is now available owing to the 1000 Genome Project [12], which allows us to grasp most of the potential common variants. This also enables us to perform genotype imputation of SNPs that are not directly genotyped in the GWAS, and consequently, to test them for association. Second, genomic studies using new technologies, such as chromatin immunoprecipitation-sequencing (ChIP-seq) and DNase I hypersensitive sites sequencing (DNase-seq), have advanced our understanding of how each genomic cluster regulates gene

Author Summary

Rheumatoid arthritis (RA) is a chronic autoimmune disease affecting approximately 1% of the general adult population. More than 30 susceptibility loci for RA have been identified through genome-wide association studies (GWAS), but the disease-causal variants at most loci remain unknown. Here, we performed replication studies of the candidate loci of our previous GWAS using Japanese cohorts and identified variants in *NFKBIE* and *RTKN2* gene loci that were associated with RA. To search for causal variants in both gene regions, we first examined non-synonymous (ns)SNPs that alter amino-acid sequences. As *NFKBIE* and *RTKN2* are known to be involved in the NF-κB pathway, we evaluated the effects of nsSNPs on NF-κB activity. Next, we screened *in silico* variants that may regulate gene transcription using publicly available epigenetic databases and subsequently evaluated their regulatory potential using *in vitro* assays. As a result, we identified multiple candidate causal variants in *NFKBIE* (2 nsSNPs and 1 regulatory SNP) and *RTKN2* (2 regulatory SNPs), indicating that our integrated *in silico* and *in vitro* approach is useful for the identification of causal variants in the post–GWAS era.

transcription. If disease-associated variants are present in a critical site for gene regulation suggested by the ChIP-seq and DNase-seq studies, the disease-associated variants might possibly influence gene transcription levels such as through altered transcription factor-DNA binding avidity.

In the present study, we first performed replication studies of candidate loci in our previous GWAS and identified two association signals with genome-wide significance ($P<5\times10^{-8}$) in nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon (*NFKBIE* [MIM 604548]) and rhotekin 2 (*RTKN2*) loci. By utilizing publicly available datasets yielded by the above-mentioned genomic studies, we then performed integrated *in silico* and *in vitro* analysis to identify plausible causal variants in *NFKBIE* and *RTKN2* loci.

Results

Identification of rheumatoid arthritis susceptibility genes

We previously performed a GWAS of RA using a Japanese case-control cohort (2,303 cases and 3,380 controls) and identified significant associations in major histocompatibility complex, class II, DR beta 1 (*HLA-DRB1* [MIM 142857]), and chemokine (C-C motif) receptor 6 (*CCR6* [MIM 601835]) loci ($P_{GWAS}<5\times10^{-8}$) [6]. To reveal additional risk loci from those showing moderate associations in the GWAS (31 loci, $5\times10^{-8}<P_{GWAS}<5\times10^{-5}$), we selected a landmark SNP from each locus and genotyped it for an additional cohort (replication-1: 2,187 cases and 28,219 controls) (Table S1, S2). Among the 31 SNPs genotyped, seven SNPs were nominally associated with RA ($P<0.05$), which included SNPs in the tumor necrosis factor, alpha-induced protein 3 (*TNFAIP3* [MIM 191163]), and signal transducer and the activator of transcription 4 (*STAT4* [MIM 600558]) gene loci that were previously reported to be associated with RA [13,14] (Table S2). In a combined analysis of the GWAS and the 1st replication study, we identified two associations with genome-wide significance ($P<5\times10^{-8}$) in *NFKBIE* (6p21.1, rs2233434, $P=4.1\times10^{-11}$, odds ratio (OR) = 1.21, 95% confidence interval (CI) = 1.14–1.28) and in *RTKN2* (10q21.2, rs3125734, $P=3.7\times10^{-8}$, OR = 1.23, 95% CI = 1.14–1.32) (Table 1 and Figure 1). *NFKBIE* was previously reported as a novel RA susceptibility gene locus in a meta-analysis of three GWASs for RA in the Japanese population, which included the GWAS set that the present study used [15]. *RTKN2* is located in the same region (10q21) as *ARID5B*, in which a significant association signal was also reported in the meta-analysis [15]. In our GWAS set, however, two significant signals were observed at rs3125734 (*RTKN2*: $P=4.8\times10^{-5}$) and rs10821944 (*ARID5B*: $P=7.4\times10^{-4}$), the former of which was tested as a landmark in the replication study. These two SNPs were in weak LD ($r^2=0.11$) and the independent effect of each SNP was observed after conditioning on each SNP (*RTKN2*: $P=1.5\times10^{-3}$, *ARID5B*: $P=0.024$, respectively). This indicated that two independent associations existed in this region, and the association of *RTKN2* is novel. We also confirmed the association in the *STAT4* locus [14] with genome-wide significance (2q32.2, rs10168266, $P=3.2\times10^{-8}$, OR = 1.16, 95% CI = 1.10–1.22) (Table S2). The associations in *NFKBIE* and *RTKN2* were further replicated in the 2nd replication cohort (3,417 cases and 3,763 controls; rs2233434, $P=1.1\times10^{-5}$, OR = 1.19, 95% CI = 1.10–1.30 and rs3125734, $P=0.016$, OR = 1.14, 95% CI = 1.02–1.26, respectively), confirming the associations in these loci (a combined analysis of three sets; rs2233434, $P=1.3\times10^{-15}$, OR = 1.20, 95% CI = 1.15–1.26 and rs3125734, $P=4.6\times10^{-9}$, OR = 1.20, 95% CI = 1.13–1.27, respectively) (Table 1 and Figure 1). We also genotyped these SNPs for individuals with systemic lupus erythematosus (SLE [MIM 152700]) ($n=657$) and Graves' disease (GD [MIM 275000]) ($n=1,783$). We identified a significant association of *RTKN2* (rs3125734) with GD ($P=3.4\times10^{-5}$, OR = 1.24, 95% CI = 1.12–1.37), whereas no significant associations were detected in *NFKBIE* (rs2233434) with either disease or in *RTKN2* (rs3125734) with SLE (Table S3).

Functional analysis of non-synonymous SNPs

NFKBIE and *RTKN2* genes are both involved in the NF-κB pathway: *NFKBIE* encodes IκB epsilon (IκBε), a member of the IκB family [16], and its binding to NF-κB inhibits the nuclear translocation of NF-κB [17]; *RTKN2* encodes a member of Rho-GTPase effector proteins highly expressed in $CD4^+$ T cells [18] and is implicated in the activation of the NF-κB pathway [19]. Considering that the NF-κB pathway is critical for the pathogenesis of RA [20], these two genes could be strong candidates in these regions. To identify disease-causing variants, we first sequenced the coding regions of the genes using DNA from patients ($n=48$) to find variants that alter amino acid sequences. We identified four non-synonymous (ns)SNPs, which were all registered in the dbSNP database: two nsSNPs in *NFKBIE* (rs2233434 (Val194Ala) and rs2233433 (Pro175Leu)) and two in *RTKN2* (rs3125734 (Arg462His) and rs61850830 (Ala288Thr)), where rs2233434 and rs3125734 were the same as the landmark SNPs in the GWAS (Figure 1 and Figure 2A). The two nsSNPs of each locus were in strong LD (Figure 2B) and were both associated with disease (Table S4). In the haplotype analysis, a single common risk haplotype with a frequency >0.05 was observed in each locus, and significant associations with disease risk were detected (*NFKBIE*, $P=5.3\times10^{-8}$, Table S5; *RTKN2*, $P=5.7\times10^{-5}$, Table S6).

To investigate the effect of these nsSNPs on protein function, we evaluated them by *in silico* analysis using PolyPhen and SIFT software, which predicts possible impacts of amino acid substitutions on the structure and function of proteins, but all four nsSNPs were predicted to have little effect (Table S7), contrasting with the effect of Mendelian disease mutations [9]. We next examined their influence on the NF-κB activity in cells by performing NF-κB

Table 1. Association analysis of *NFKBIE* and *RTKN2* with rheumatoid arthritis.

		Allele		Number of subjects		Frequency of allele 1			
Gene	dbSNP ID	(1/2)	Study set	Case	Control	Case	Control	Odds ratio (95% CI)	*P*-value[a]
NFKBIE	rs2233434	G/A	GWAS	2,303	3,380	0.254	0.216	1.24 (1.13–1.35)	2.2×10^{-6}
			Replication study-1	2,186	28,204	0.245	0.215	1.19 (1.10–1.27)	4.2×10^{-6}
			Replication study-2	3,396	3,756	0.239	0.209	1.19 (1.10–1.30)	1.1×10^{-5}
			Combined analysis	7,885	35,340	0.245	0.215	1.20 (1.15–1.26)	1.3×10^{-15}
RTKN2	rs3125734	T/C	GWAS	2,303	3,380	0.125	0.101	1.27 (1.13–1.43)	4.8×10^{-5}
			Replication study-1	2,185	28,218	0.129	0.110	1.20 (1.09–1.31)	1.4×10^{-4}
			Replication study-2	3,402	3,751	0.115	0.103	1.14 (1.02–1.26)	0.016
			Combined analysis	7,890	35,349	0.122	0.108	1.20 (1.13–1.27)	4.6×10^{-9}

[a]: Cochran-Armitage trend test was used for the GWAS and replication studies. Mantel-Haenszel method was used for the combined analysis.
doi:10.1371/journal.pgen.1002949.t001

reporter assays with haplotype-specific expression vectors. In *NFKBIE*, the non-risk haplotype (A-C: rs2233434 (non-risk allele (NR))-rs2233433 (NR)) displayed an inhibitory effect on NF-κB activity compared with the mock construct, which reflected compulsorily binding of exogenous IκBε to the endogenous NF-κB, as shown in a previous study [16]. Of note, the risk haplotype (G-T: risk allele (R)-R) showed higher NF-κB activity than A-C (NR-NR) (Figure 3A), suggesting impaired inhibitory potential of G-T (R-R) products. No haplotypic difference was detected in the protein expression levels of these constructs (Figure 3C). We also examined two additional constructs of G-C (R-NR) and A-T (NR-R) haplotypes to evaluate the effect of each nsSNP (Figure S1A, S1B). Because NF-κB activity increased in the order of A-C<G-C<A-T<G-T (rs2233434-rs2233433: NR-NR<R-NR<NR-R<R-R) when cells were stimulated with TNF-α, the C>T substitution (Pro175Leu) in rs2233433 may have more impact on the protein function of IκBε compared with the A>G substitution (Val194Ala) in rs2233434. In contrast to the observations in *NFKBIE*, no clear difference was detected between the two common haplotype products of *RTKN2* in either their effect on NF-κB activity or protein expression levels, although both products enhanced NF-κB activity as reported previously (Figure 3B, 3D) [19]. These functional analyses of nsSNPs suggest that two nsSNPs (rs2233434 and rs2233433) in the *NFKBIE* region are candidates for causal SNPs.

ASTQ analysis suggested the existence of regulatory variants

As the majority of autoimmune disease loci have been implicated as eQTL [11], we speculated that variants in the *NFKBIE* and *RTKN2* loci would influence gene function by regulating gene expression, in addition to changing the amino acid sequences. To address this possibility, we performed allele-specific transcript quantification (ASTQ) analysis by using allele-specific probes targeting the nsSNPs in exons (rs2233434 for *NFKBIE* and rs3125734 for *RTKN2*, both of which were the GWAS landmarks). The genomic DNAs and cDNAs were extracted from peripheral blood mononuclear cells (PBMCs) in individuals with heterozygous genotype ($n=14$ for *NFKBIE* and $n=6$ for *RTKN2*) and from lymphoblastoid B-cell lines ($n=9$) for *NFKBIE*. As the expression levels of *RTKN2* were low in lymphoblastoid B cells, only PBMCs were used. When quantified by allele-specific probes, transcripts from the risk allele of *NFKBIE* showed 1.1-fold and 1.2-fold lower amounts (in PBMCs and lymphoblastoid B cells, respectively) than those from non-risk alleles ($P=0.012$ and 5.3×10^{-4}, respectively; Figure 3E and Figure S2). In contrast, 1.5-fold higher amounts of transcripts were observed in the risk allele of *RTKN2* ($P=0.016$; Figure 3F). These allelic imbalances suggested that both gene loci were eQTL and that there existed variants with *cis*-regulatory effects. Moreover, considering the inhibitory effects of *NFKBIE* and the activating potential of *RTKN2* on NF-κB activity, which might both be dose dependent (Figure 3G, 3H), these regulatory variants in the risk alleles should enhance NF-κB activity *in vivo*.

Integrated *in silico* and *in vitro* analysis to search for regulatory variants

To comprehensively search the two genomic regions for causal regulatory variants, we performed an integrated *in silico* and *in vitro* analysis with multiple steps (Figure 4 and Figures S3, S4). We first determined the target genomic region by selecting LD blocks containing disease-associated SNPs ($P_{GWAS}<1.0\times10^{-3}$) (Step 1). We then extracted SNPs with frequencies of >0.05 from HapMap and 1000 Genome Project databases in the region (Step 2). We excluded uncommon variants (MAF<0.05) from the analysis because of their low imputation accuracy in the GWAS (93% of uncommon variants in *NFKBIE* and 76% in *RTKN2* exhibited *Rsq* <0.6). There is neither structural variation (>1 kbps) nor indels (100 bps to 1 kbs) that are common in the population (frequency >0.01) in these loci. To evaluate the *cis*-regulatory potential of sequences around the SNPs *in silico*, we used the regulatory potential (RP) score [21,22]. This score was calculated based on the extent of sequence conservation among species or similarity with known regulatory motifs. We selected SNPs from the genomic elements with an RP score >0.1 (Step 3a). Subsequently, we selected SNPs from sites of transcriptional regulation as demonstrated by previous ChIP-seq studies (transcription factor binding sites [23,24] and histone modification sites [25,26]) or a DNase-seq study (DNase I hypersensitivity sites) [27] (Step 3b). Finally, these SNPs with regulatory potential were further screened out by the disease-association status ($P<0.05$) using an imputed GWAS dataset (Step 4). As a consequence, we selected 14 SNPs in *NFKBIE* and 10 SNPs in *RTKN2* that had regulatory potential predicted *in silico*.

To further investigate the regulatory potential of the SNPs, we evaluated 31-bp sequences around the SNPs by *in vitro* assays. First, we examined their ability to bind nuclear proteins by EMSAs (Step 5a) using nuclear extracts from lymphoblastoid B cells (PSC cells) and Jurkat cells. Of the 24 SNPs examined, nine

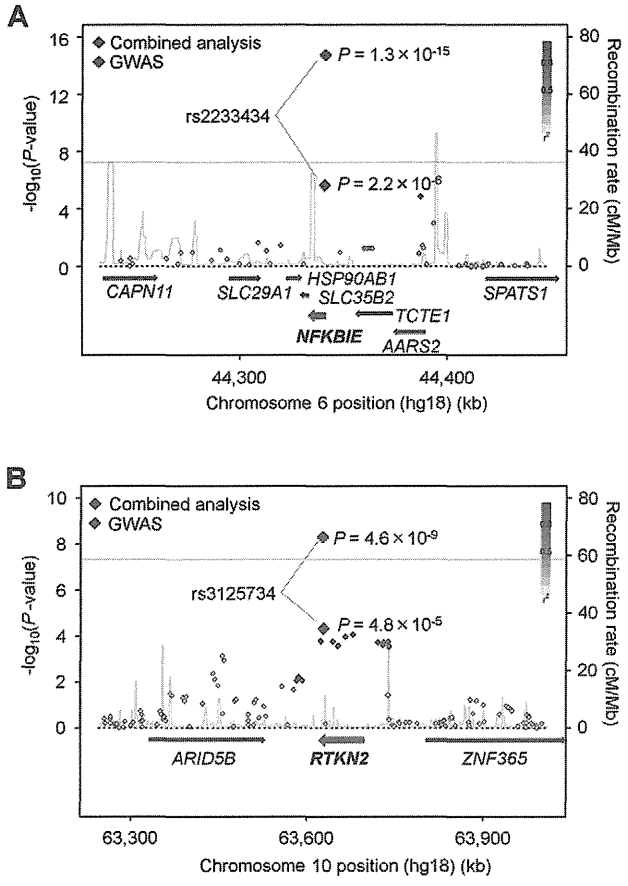

Figure 1. Association plots of *NFKBIE* and *RTKN2* regions. The diamonds represent the -$\log_{10}$ of the Cochran-Armitage trend *P*-values. Large diamonds show landmark SNPs in *NFKBIE* (rs2233434: A) and *RTKN2* (rs3125734: B). Red: GWAS, Blue: combined analysis. Red colors of each SNP indicate its r^2 with landmark SNP. Gray lines indicate the genome-wide significance threshold ($P<5\times10^{-8}$). For each plot, the -$\log_{10}$ *P*-values (y-axis) of the SNPs are presented according to their chromosomal positions (x-axis). Physical positions are based on NCBI build 36.3 of the human genome. Genetic recombination rates, estimated using the 1000 Genome Project (JPT and CHB), are represented by the blue line.
doi:10.1371/journal.pgen.1002949.g001

SNPs displayed allelic differences, implying differential potential of transcriptional activity between these alleles (Figure 5A and Figure S5). We then evaluated the enhancing or repressing activity of the sequences by luciferase reporter assays (Step 5b). We cloned them into the pGL4.24 vector, which has minimal promoter activity, and transfected these constructs into HEK293A cells (for *NFKBIE* and *RTKN2*), lymphoblastoid B cells (for *NFKBIE*), and Jurkat cells (for *RTKN2*). Among the three SNPs examined in *NFKBIE*, the risk allele of rs2233424 (located −396 bps from the 5′ end) displayed stronger repression activity (Figure 2A and Figure 5B) than that of the non-risk allele. Among the six SNPs in *RTKN2*, the risk alleles of rs12248974 (approximately 10 kb from the 3′ end) and rs61852964 (−215 bps from the 5′ end) showed higher enhancing activity compared with the non-risk alleles (Figure 2A and Figure 5B). These results corresponded to the results of ASTQ analyses (Figure 3E, 3F). Other SNPs showed no allelic differences or had the opposite trend of transcriptional activity in the risk allele compared to the results of ASTQ analysis (Figure S6).

To confirm the regulatory potential of these SNPs, we investigated the correlation between genotypes and gene expression levels in lymphocytes utilizing the data from the previous eQTL studies. We evaluated the expression of *RTKN2* in primary T cells from Western European individuals by using Genevar software [28,29]. Though *NFKBIE* is also expressed in primary T cells, the genotypes of rs2233424 are not available. We thus evaluated gene expression data of lymphoblastoid B-cell lines obtained from HapMap individuals (Japanese (JPT) + Han Chinese in Beijing (CHB), European (CEU), and African (YRI)) [30,31] instead. The *NFKBIE* expression level decreased with the number of risk alleles of rs2233424 ($R=-0.18$, $P=0.020$), and the *RTKN2* expression levels increased with that of rs1432411 (a proxy for rs12248974, $r^2=0.97$) ($R=0.27$, $P=0.018$) (Figure 5C), corresponding to the results of the *in vitro* assays. The data for rs61852964 in *RTKN2* was not available. Among the SNPs that displayed opposite transcriptional activities in the reporter assays compared to the results of ASTQ, the data for rs2233434, rs77986492, and rs3852694 (a proxy for rs1864836, $r^2=1.0$) were available (Figure S7 and S8). These SNPs displayed the opposite direction of the correlation trend as compared to the results of reporter assays, but parallel to ASTQ, implying that the regulatory effects observed in the *in vitro* assays were cancelled out by the effects of other regulatory variants on the same haplotype *in vivo*.

Finally, we validated the associations of these regulatory (r)SNPs observed in the imputed GWAS dataset. We directly genotyped them by TaqMan assay and confirmed significant associations (Table S8). As the candidate causal variants (nsSNPs and rSNPs) and the landmark SNPs of GWAS were in strong LD at each locus (Figure 2A, 2B), we evaluated the independent effect of each SNP by haplotype analysis in both loci (Table S9 and S10) and the conditional logistic regression analysis in *RTKN2* (Table S11). The conditional analysis was not performed in *NFKBIE* because three candidate causal variants were in strong LD ($r^2>0.9$). However, the analyses for these two loci did not demonstrate any evidence of primary or independent effects across the candidate causal variants, and it remains a possibility that all of the functional variants were involved in the pathogenesis. In addition, although the landmark nsSNP (rs3125734) in *RTKN2* did not display any influence on NF-κB activity in our *in vitro* assays, rs3125734 might influence functions of *RTKN2* other than those in the NF-κB pathway; alternatively, it is still possible that rs3125734 tags the effects of other unknown variants, such as rare variants, in addition to the other two rSNPs (rs12248974 and rs61852964).

Discussion

In the present study, we performed a replication study of our previously reported GWAS and identified variants in *NFKBIE* and *RTKN2* loci that were associated with RA susceptibility. The associations of *NFKBIE* and *RTKN2* loci have not been reported in other populations with genome-wide significance. However, rs2233434 in *NFKBIE* showed a suggestive association (589 cases vs. 1,472 controls, $P=0.0099$, OR = 1.57, 95% CI = 1.11–2.21) in a previous meta-analysis in European populations [32]. The weak association signal in Europeans may be partially due to the lower frequency of the risk allele (0.04 in Europeans compared to 0.22 in Japanese). On the other hand, the association of rs3125734 in *RTKN2* was not observed in a GWAS meta-analysis of European populations (cases 5,539 vs. controls 20,169, $P=0.11$, OR = 1.04, 95% CI = 0.99–1.09). As the association of *RTKN2* locus was also implicated in Graves' disease in a Han Chinese population [33], the association in *RTKN2* locus may be unique to Asian populations.

To find the disease causal variants in disease-associated loci, target re-sequencing and variant genotyping with a large sample set followed by conditional association analysis examining the

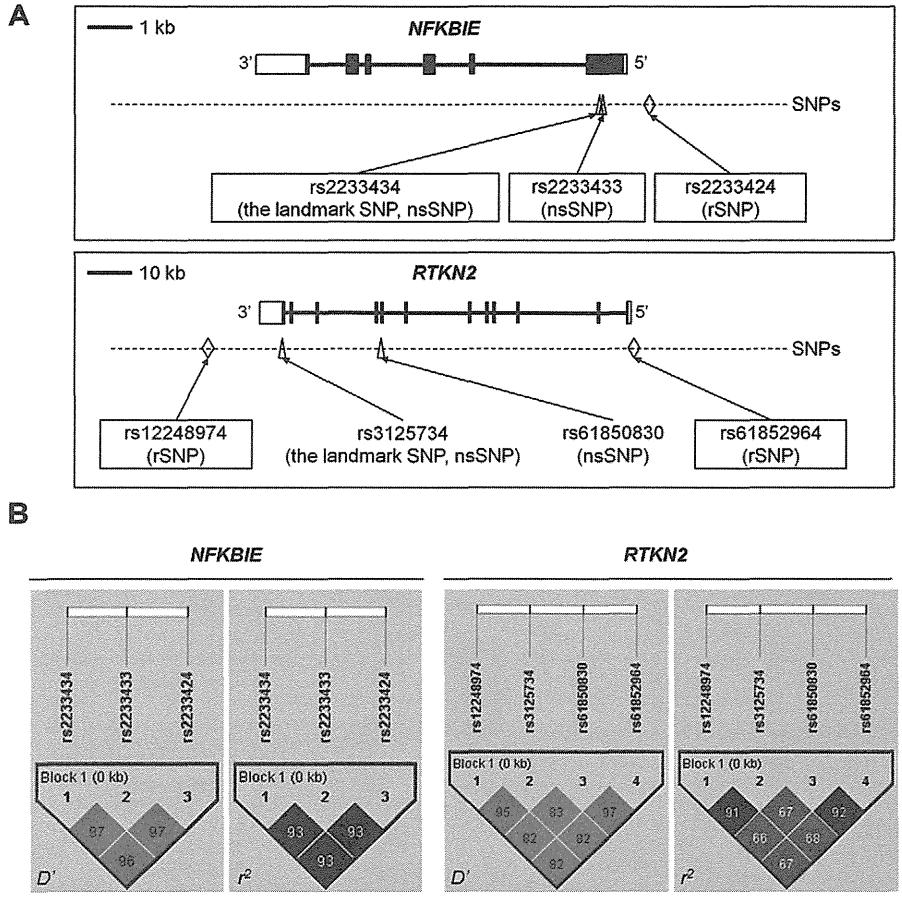

Figure 2. Genomic position and LD blocks. (A) Genomic position of non-synonymous (ns)SNPs and regulatory (r)SNPs in *NFKBIE* and *RTKN2*. *NFKBIE* (top) and *RTKN2* (bottom) correspond to transcripts NM_004556.2 and NM_145307.2, respectively. Exons are shown as boxes, where black boxes represent coding regions and open boxes represent untranslated regions. Intron sequences are drawn as lines. Open triangles represents nsSNPs and open diamond shapes indicate candidate rSNPs. dbSNP IDs of candidate causal variants were boxed in a solid line. (B) LD patterns for nsSNPs and candidate rSNPs in *NFKBIE* (left) and *RTKN2* (right) gene regions. LD blocks were constructed from genotype data of 3,290 control individuals of the GWAS. The diagrams show pairwise LD values as quantified using the D' and r^2 values.
doi:10.1371/journal.pgen.1002949.g002

independent effects of each variant would be the first step. For this purpose, a recent attempt to fine-map the known autoimmunity risk loci in Celiac disease (MIM 212750) using an "Immunochip" brought us several insights [34]. First, no stronger signals compared to the GWAS signals were detected in most of the known loci, while additional independent signals were found in several loci. Second, none of the genome-wide significant common SNP signals could be explained by any rare highly penetrant variants. Third, although the fine-mapping strategy could localize the association signals into finer scale regions, it could not identify the actual causal variants due to strong LD among the variants, indicating that an additional approach, such as functional evaluation of candidate variants, is needed.

In the present study, we focused on common variants to find causal variants. Instead of re-sequencing additional samples, we utilized the 1000 Genome Project dataset, where the theoretically estimated cover rate for common variants (frequency of >0.05) in our population is >0.99 [12,35]. To fine-map the association signals, we performed imputation-based association analysis, where we could not find any association signals that statistically exceeded the effect of landmark SNPs (rs2233434 for *NFKBIE* and rs3125734 for *RTKN2*) in both gene regions (Figures S3 and S4). We also performed a conditional logistic regression analysis, and found no additional independent signals of association when conditioned on each landmark SNP (data not shown). Although the imputation-based association tests may yield some bias compared to direct genotyping of the variants, these results suggested that variants in strong LD with the landmark SNPs were strong candidates for causal variants.

Following the analysis of nsSNPs, we evaluated *cis*-regulatory effects of variants in the two regions by ASTQ analysis using both B-cell lines and primary cells (PBMC), the majority of which consisted of T and B lymphocytes. As the mechanism of gene-regulation is substantially different between cell types [26], ASTQ analysis in more specific cell types that are relevant to the disease etiology, such as Th1 and Th17 cells, would be ideal to evaluate the *cis*-regulatory effects of variants. In this context, a more comprehensive catalog of the eQTL database of multiple cell types should be established for genetic study of diseases. As our ASTQ analysis demonstrated *cis*-regulatory effects of variants in both regions, we then performed an integrated *in silico* and *in vitro* analysis to identify candidate regulatory variants. Accumulating evidence by recent ChIP-seq and DNase-seq studies suggested that *cis*-regulatory variants are

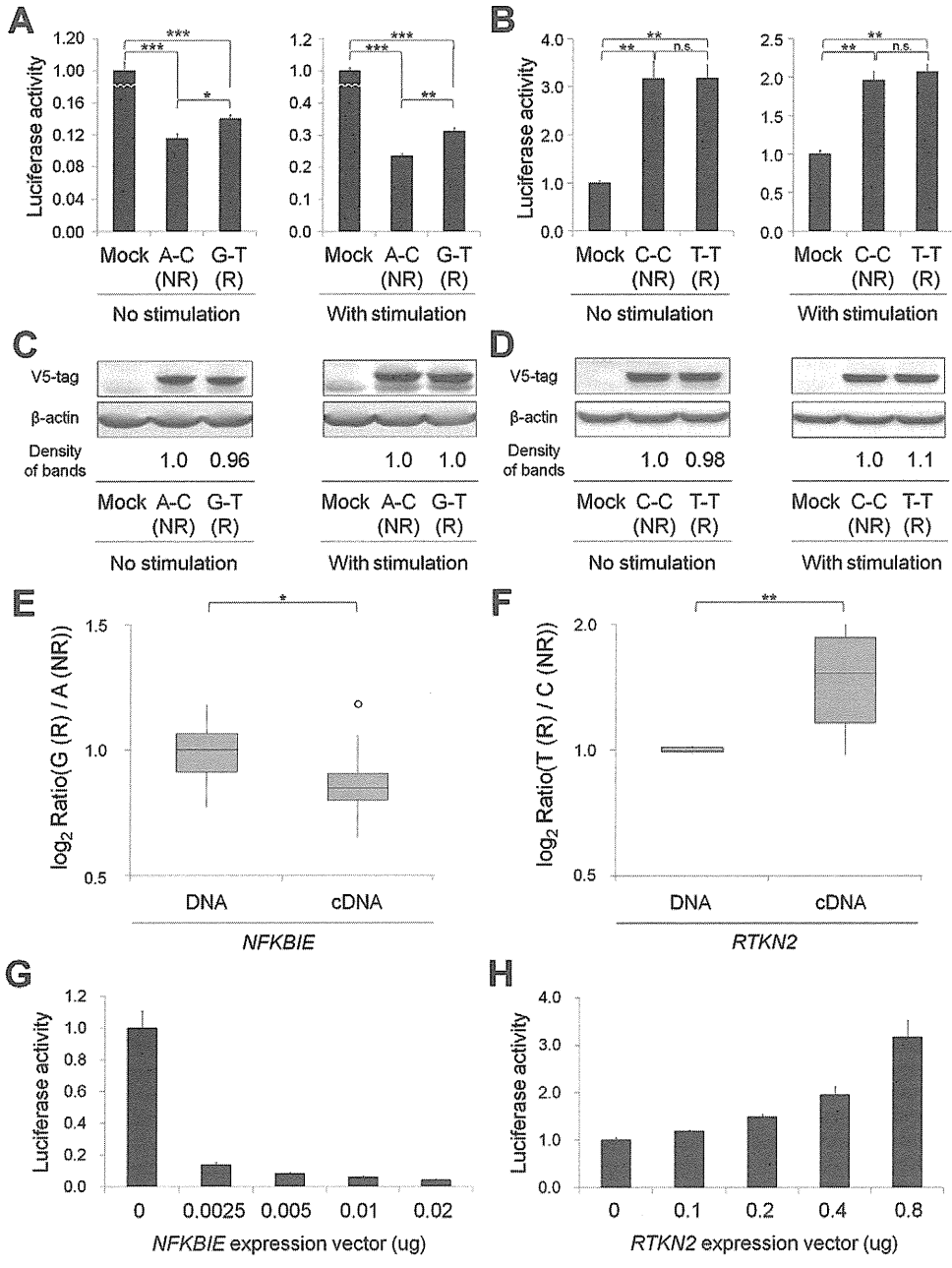

Figure 3. Functional evaluation of nsSNPs and allelic imbalance of expression in *NFKBIE* and *RTKN2*. (A, B) Effects of nsSNPs in *NFKBIE* (A) and *RTKN2* (B) on NF-κB activity by luciferase assays. Two haplotype constructs (A-C (rs2233434-rs2233433; non-risk (NR)) and G-T (risk (R)) for *NFKBIE* and C-C (rs3125734-rs61850830; NR) and T-T (R) for *RTKN2*) were used. The expression vector of each construct, pGL4.32[*luc2P*/NF-κB-RE] vector and pRL-TK vector were transfected into HEK293A cells. Data represent the mean ± s.d. Each experiment was performed in sextuplicate, and experiments were independently repeated three times. $^{*}P<0.05$, $^{**}P<1.0\times10^{-5}$, and $^{***}P<1.0\times10^{-10}$ by Student's *t*-test. n.s.: not significant. (C, D) Protein expression levels of each haplotype construct. Anti-V5 tag antibody was used in the Western blotting analysis to detect the expression of exogenous IκBε (C) and RTKN2 (D). Beta-actin expression was used as an internal control. The densities of the bands were quantified and normalized to that of the risk allele. (E, F) Allelic imbalance of expression in *NFKBIE* (E) and *RTKN2* (F). ASTQ was performed using samples from individuals heterozygous for rs2233434 (G/A) in *NFKBIE* and rs3125734 (T/C) in *RTKN2*. Genomic DNAs and cDNAs were extracted from PBMCs ($n=14$ for *NFKBIE* and $n=6$ for *RTKN2*). The y-axis shows the $\log_2$ ratio of the transcript amounts in target SNPs (risk allele/non-risk allele). The top bar of the box-plot represents the maximum value and the lower bar represents the minimum value. The top of box is the third quartile, the bottom of box is the first quartile, and the middle bar is the median value. The circle is an outlier. $^{*}P=0.012$, $^{**}P=0.016$, by Student's *t*-test. (G, H) Dose-dependent inhibition of *NFKBIE* (G) and activation of *RTKN2* (H) on NF-κB activity. Various doses of expression vectors carrying the non-risk allele of each gene were transfected into HEK293A cells with pGL4.32 and pRL-TK vectors.
doi:10.1371/journal.pgen.1002949.g003

located in the key regions of transcriptional regulation [26,36], warranting the prioritization of variants before evaluation by *in vitro* assays. This could also minimize false-positive results of the *in vitro* assays. However, there may be additional causal variants, including rare variants, unsuccessfully selected at each step of our integrated screening. Therefore, the screening strategy

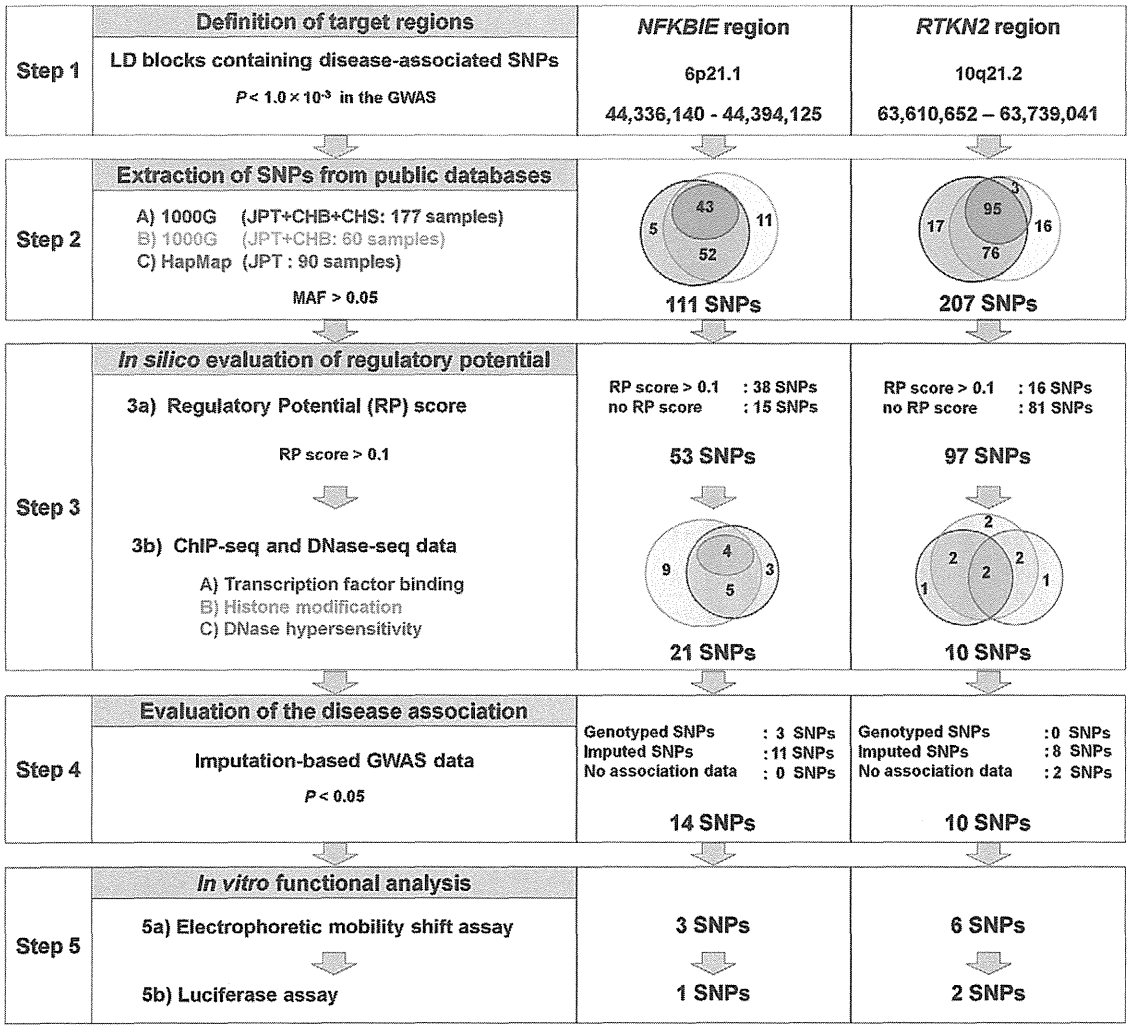

Figure 4. Overview of SNP selection using integrated *in silico* and *in vitro* approaches. The figure shows the SNP selection process (left) and the results of *NFKBIE* (middle) and *RTKN2* (right). (Step 1) LD blocks that contain disease-associated SNPs ($P_{GWAS}<1.0\times10^{-3}$) were selected. (Step 2) SNPs were extracted from three databases (A–C). 1000G, 1000 Genome Project; HapMap, International HapMap Project. A) JPT, CHB, and CHS samples ($n=177$) from the 1000G (the August 2010 release). B) JPT and CHB samples ($n=60$) from the pilot 1 low coverage study data of 1000G (the March 2010 release). C) JPT samples ($n=90$) from HapMap phase II+III (release #27). SNPs with minor allele frequency >0.05 were selected. (Step 3) Prediction of regulatory potential *in silico*. 3a) Regulatory potential (RP) scores were used for SNP selection, where an RP score >0.1 indicated the presence of regulatory elements. SNPs without RP scores were also selected. 3b) Prediction of regulatory elements by ChIP-seq data and DNase-seq data. (A) Transcription factor binding sites, (B) histone modification sites (CTCF binding, H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac), and (C) DNase I hypersensitivity sites were evaluated. ChIP-seq and DNase-seq data derived from GM12878 EBV-transformed B cells were used for *NFKBIE* and *RTKN2*. DNase-seq data of Th1, Th2, and Jurkat cells were also used for *RTKN2*. (Step 4) Association data of the imputation-based GWAS using 1000G reference genotypes were used. SNPs with a significance level of $P<0.05$ were selected. SNPs without association data were also selected. (Step 5) EMSAs and luciferase assays were performed for evaluation of regulatory potentials *in vitro*.
doi:10.1371/journal.pgen.1002949.g004

should be refined as the quality and quantity of genomic databases improves in the future.

We identified multiple candidate causal variants in *NFKBIE* (two nsSNPs and one rSNP) and *RTKN2* (two rSNPs). We could not statistically distinguish the primary effect of each candidate causal variant, because these variants are in strong LD and on the same common haplotype. However, multiple causal variants could be involved in a single locus, which is also seen in another well-known autoimmune locus in 6q23 (*TNFAIP3* gene locus), where both an nsSNP and a regulatory variant have been shown to be functionally related to the disease [8,37]. The risk haplotype of nsSNPs in *NFKBIE* (rs2233433 and rs2233434) showed an enhancement of NF-κB activity, which might reflect an impaired inhibitory effect of IκB-ε on nuclear translocation of NF-κB. On the other hand, down-regulated *NFKBIE* expression and up-regulated *RTKN2* expression were observed at the risk haplotypes, which may be regulated in *cis* by the rSNPs (rs2233424 in *NFKBIE*, rs12248974 and rs61852964 in *RTKN2*). As overexpression studies have also demonstrated dose-dependent attenuation of NF-κB activity by *NFKBIE*, and dose-dependent enhancement by *RTKN2*, the *cis*-regulatory effects of these rSNPs should enhance the NF-κB activity in the risk allele. Taken together with the effect of nsSNPs in *NFKBIE*, the enhancement of NF-κB activity may play a role in the pathogenesis of the disease. This is further supported by evidence that previous GWAS for RA have also identified genes related to the NF-κB pathway, such as *TNFAIP3* [13], v-rel