Journal of Human Genetics (2009) 54, 488–492
© 2009 The Japan Society of Human Genetics All rights reserved 1434-5161/09 $32.00
www.nature.com/jhg

ORIGINAL ARTICLE

A survey of the population genetic variation in the human kinome

Wei Zhang[1], Daniel VT Catenacci[1], Shiwei Duan[1] and Mark J Ratain[1,2,3]

Protein kinases are key regulators of various biological processes, such as control of cell growth, metabolism, differentiation and apoptosis. Therefore, protein kinases have been an important class of targets for anticancer drugs. Health-related disparities such as differential drug response have been observed between human populations. A survey of the human kinases and their ligand genes for those containing population-specific genetic variants could provide new insights into the mechanisms of these health disparities and suggest novel targets for ethnicity-specific personalized medicine. Using the International HapMap Project genotypic data on single-nucleotide polymorphisms (SNPs), the protein kinase complement of the human genome (kinome) and some experimentally verified ligand genes were scanned for the existence of population-specific SNPs (eSNPs). In general, protein kinases were found to contain a much higher proportion of eSNPs than the whole genome background, indicating a stronger pressure for adaptation in individual populations. In contrast, the proportion of ligand genes containing eSNPs was not different from that of the whole genome background. Although with some important limitations, our results suggest that human kinases are more likely to be under recent positive selection than ligands. Our findings suggest that the health-related disparities associated with kinase signaling pathways are more likely to be driven by the genetic variation in the kinase genes than their cognate ligands. Illustrating the role of molecular evolution in the genetic variation of the human kinome could provide a promising route to understand the ethnic differences in cancer and facilitate the realization of ethnicity-based individualized medicine.

Journal of Human Genetics (2009) **54,** 488–492; doi:10.1038/jhg.2009.72; published online 31 July 2009

Keywords: HapMap; kinase; kinome; ligand; single-nucleotide polymorphism

INTRODUCTION

Protein kinases are key regulators of cell function by adding phosphate groups to substrate proteins. Phosphorylation by protein kinases is the most widespread and well-studied signaling mechanism in eukaryotic cells. Phosphorylation can regulate almost every property of a protein and is involved in the activity, localization and overall function of many proteins. It serves to orchestrate the activity of almost all cellular processes.[1] The protein kinase complement ($\sim$500 genes) of the human genome (kinome) constitutes one of the largest and the most functionally diverse gene families and has been comprehensively cataloged by the Human Kinome Project.[2]

As protein phosphorylation has a central role in diverse biological processes, such as control of cell growth, metabolism, differentiation and apoptosis, abnormal phosphorylation has been implicated in the cause of human cancer. The development of selective protein kinase inhibitors that can block or modulate diseases caused by abnormalities in these signaling pathways is widely considered a promising approach for drug development.[3] Several new cancer treatments are designed to inhibit aberrantly activated kinases within cancer cells in an effort to prevent cell division. FDA (Food and Drug Administration)-approved kinase inhibitors that are used to treat various cancers include, for example, erlotinib and gefitinib, which target the epidermal growth factor receptor (EGFR),[4,5] and sorafenib, which was designed as an inhibitor of Raf kinase, but also targets the vascular endothelial growth factor receptors.[6,7] Protein kinases have now become the second most important group of drug targets, after G-protein-coupled receptors.[8]

Although socioeconomic status could affect health-related disparities, for some diseases, there are well-established relationships between ancestry and disease risk/pharmacological response. For example, African-American, Hispanic, Asian and Native American women have a lower incidence of breast cancer but higher mortality compared with non-Hispanic white women.[9] A significant difference in response and pulmonary toxicity to gefitinib, an inhibitor of the EGFR kinase, has been observed between patients with advanced non-small-cell lung cancer from Asia and Europe/North America.[10] In addition, there are clear population differences in *EGFR*, which may explain some of the clinical population differences.[11] Thus, we hypothesized that there may be clinically important population differences in other kinase genes, and sought to comprehensively assess the entire kinome and the relevant cognate ligands.[12]

[1]Section of Hematology/Oncology, Department of Medicine, The University of Chicago, Chicago, IL, USA; [2]Committee on Clinical Pharmacology and Pharmacogenomics, The University of Chicago, Chicago, IL, USA and [3]Cancer Research Center, The University of Chicago, Chicago, IL, USA
Correspondence: Dr MJ Ratain, Section of Hematology/Oncology, Department of Medicine, The University of Chicago, 5841 S Maryland Ave. MC 2115, Chicago, IL 60637, USA.
E-mail: mratain@medicine.bsd.uchicago.edu

Received 27 March 2009; accepted 3 July 2009; published online 31 July 2009

To catalog the genetic variation in protein kinase genes, we used a resource of single-nucleotide polymorphisms (SNPs) from the International HapMap Project (http://www.hapmap.org/).[13,14] The Phase 1/2 HapMap genotypic database, which comprises >3 million SNPs,[15] has proven to be a key resource for researchers investigating the genetic contribution to human diseases, variation in gene expression and drug response.[16] A comprehensive survey was carried out to identify protein kinase genes as well as ligand genes that contained SNPs with differential frequencies (eSNPs) among a panel of human lymphoblastoid cell lines derived from apparently healthy individuals of northern and western European ancestry (CEU: 60 unrelated Caucasian individuals from Utah, USA), YRI (60 unrelated Yoruba people from Ibadan, Nigeria) of African ancestry and ASN (CHB: 45 unrelated Han Chinese from Beijing, China; JPT: 45 unrelated Japanese from Tokyo, Japan) of Asian ancestry. As the three major continental populations (Asians, Europeans and Africans) have been separated geographically during the past 50 000–100 000 years, recent positive selection has been shown to contribute to the genetic[17] and phenotypic (for example, gene expression)[18] differences in the current populations. Therefore, the evidence for recent positive selection[17] among the kinase and ligand genes was also searched using the HapMap SNP genotypic data.

MATERIALS AND METHODS

Human protein kinase genes

The protein kinase complement of the human genome was previously cataloged using public and proprietary genomic, cDNA and expressed sequence tag sequences.[2] The list of human protein kinase genes was downloaded from the Human Kinome Project database (http://kinase.com/mammalian/).[2] This updated list (December 2007) is comprised of 514 putative human kinase genes belonging to 10 groups and 133 families.[2] The 102 protein kinase pseudogenes in the database were excluded from this study.

Human ligand genes

The Database of Ligand-Receptor Partners (DLRP)[19] is a subset of the Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/),[20] which lists protein pairs that are known to interact with each other. In particular, the DLRP is a database of protein ligand and protein receptor pairs that are experimentally known to interact with each other.[19] In total, 181 unique ligands and 133 unique receptors (473 ligand–receptor relationships) are included in the current DLRP database (November 2001).[19] Among them, 35 unique kinase receptors (cross-checked with the Human Kinome Project database[2]) and 58 unique ligands representing 183 ligand–kinase receptor relationships were included in our analysis.

Identifying kinase and ligand genes containing eSNPs

SNP@Ethnos (http://variome.kobic.re.kr/SNPatETHNIC/),[21] a catalog of SNPs and genes that contains human population variation, was queried for variant kinases and ligands containing eSNPs across human populations. The database contains results for detecting natural selection and population differences using the ~3.6 million Phase 1 (release 16c) HapMap Project[13,14] SNPs. In particular, the nearest shrunken centroid method (NSCM) score[22] was calculated by SNP@Ethnos[21] to detect population differences in the allele frequencies of ~1 million common SNPs in the genic regions across the following three HapMap populations:[13,14] CEU (60 unrelated Caucasian individuals from Utah, USA) of northern and western European ancestry, YRI (60 unrelated Yoruba people from Ibadan, Nigeria) of African ancestry and ASN (CHB: 45 unrelated Han Chinese from Beijing, China; JPT: 45 unrelated Japanese from Tokyo, Japan) of Asian ancestry. A detailed mathematical explanation of the NSCM is described in the study by Tibshirani *et al.*[22] For example, three similar scores obtained for CHB+JPT, CEU and YRI indicate that the SNP is not critical, whereas one score differing from the other two indicates that the SNP is specific to that population. An SNP is called specific in population A (that is, eSNP for population A) if $(|s(A)-s(B)|+|s(A)-s(C)|)/2>0.3$,[21] where, for example, s(A) is the score of population A. In addition to the NSCM score, various other related information such as minor allele frequency can be obtained by searching SNP@Ethnos.[21]

Enriched kinase groups

The enrichment of a particular kinase group was detected by a binomial test using the entire human kinome as reference. The annotations for kinase groups were retrieved from the Human Kinome Project database.[2] The entire human kinome comprises 10 major groups.[2] A false discovery rate of 5% after the Benjamini–Hochberg correction[23] was used for significance in this enrichment analysis. In addition, only groups with a minimum of three genes were considered to minimize the small sample size effect.

Genes under recent positive selection

Happlotter (http://hg-wen.uchicago.edu/selection/)[17] was used to evaluate whether a particular gene had been a target of recent positive selection. Haplotter is a web application that has been developed to display the results of a scan for positive selection in the human genome using the HapMap data. In particular, we used the Happlotter-calculated iHS (integrated haplotype score) (HapMap Phase 1 data) to measure the possibility of a gene undergone recent positive selection. The empirical *P*-values, quantified by the proportion of SNPs with $|\text{iHS}|>2$ for each bin of 50 neighboring SNPs, were generated by Happlotter.[17] Simulations indicate that this criterion provides a powerful signal of selection.[17] The empirical *P*-value of 0.05 was used as the cutoff for significance.

RESULTS

Variant kinases containing eSNPs

By searching the SNP@Ethnos database, 268 unique kinase genes (Supplementary Table 1) were found to contain eSNPs across the three HapMap populations. The proportion of kinase genes (~52%) containing eSNPs across populations was much higher than that of the whole genome, whose ~38% genes (10 138 out of 26 280 genes in the Phase 1 HapMap data[21]) contain eSNPs (binomial test P=8.4E-11). Table 1 lists some examples of these kinase genes. In total, 77 genes had eSNPs in the CEU samples, 240 genes had eSNPs in the YRI samples and 53 genes had eSNPs in the ASN samples. Among them, 39 genes had eSNPs in both the CEU and YRI samples, 8 genes had eSNPs in both the CEU and ASN samples and 15 genes had eSNPs in both the YRI and ASN samples. Furthermore, 20 kinase genes had eSNPs in all of the three HapMap populations (Figure 1).

Enriched kinase groups among genes containing eSNPs

After the Benjamini–Hochberg correction ($P_{\text{adjusted}}<0.05$), no kinase groups were found to be enriched among the 268 kinase genes containing eSNPs relative to the distribution of the entire human kinome (514 kinase genes). The top-ranking kinase group among the 268 genes was classified as ‘Other’ (that is, kinases not belonging to other major groups) (25 genes, nominal P=0.00502, P_{adjusted}=0.0502). In addition, at $P_{\text{adjusted}}<0.05$, none of the kinase groups were enriched among the 77, 240 and 53 genes that contained eSNPs in the CEU, YRI and ASN samples, respectively.

Variant ligands containing eSNPs

Among the 58 ligands of protein kinases, 23 ligand genes were found to contain eSNPs. In contrast to the protein kinase genes, this proportion (~39%) was not different from that of the whole genome background (binomial test P=0.79). Table 1 lists some examples of these variant ligand genes. In total, 7 ligand genes had population-specific SNPs in the CEU samples, 19 genes had population-specific SNPs in the YRI samples and 3 genes had population-specific SNPs in the ASN samples. Among them, two ligand genes had population-specific SNPs in both the CEU and YRI samples. Furthermore, two ligand genes had population-specific SNPs in all of the three HapMap

Table 1 Some examples of kinase and ligand genes containing eSNPs

Symbol	*Kinase group*[a]	*Ligand/ Kinase*[b]	*Population under selection*	*P (liHSI)*[c]	*eSNP population*[d]
EFNB2		*L*			*YRI*
EPHB1	TK	R	ASN	0.037	CEU/YRI
EPHB2	TK	R			YRI/ASN
EPHB3	TK	R			CEU/YRI
EPHB4	TK	R			ASN
EPHB6	TK	R	ASN	0.044	YRI
EPHA4	TK	R			CEU/YRI
BTC		*L*			*YRI*
EGFR	TK	R			YRI
ERBB4	TK	R	ASN	0.044	CEU/YRI/ASN
NRG1		*L*			*CEU*[e]*/YRI*[e]*/ASN*
ERBB4	TK	R	ASN	0.044	CEU/YRI/ASN
ERBB2	TK	R			YRI[e]
TGFB1		*L*			*CEU*
TGFBR1	TKL	R	YRI	0.006	ASN
TGFB2		*L*			*YRI*
TGFBR1	TKL	R	YRI	0.006	ASN
TGFB3		*L*			*YRI*
TGFBR1	TKL	R	YRI	0.006	ASN
BMP3		*L*	*CEU*	*0.0033*	*YRI*
BMPR1A	TKL	R			YRI[f]/ASN[f]
BMPR1B	TKL	R			CEU[f]/YRI[f]
BMPR2	TKL	R	CEU	0.042	YRI
BMP5		*L*	*ASN*	*0.044*	*YRI*
BMPR1A	TKL	R			YRI/ASN
BMPR1B	TKL	R			CEU/YRI
BMPR2	TKL	R	CEU	0.042	YRI
BMP7		*L*			*YRI*
BMPR1A	TKL	R			YRI/ASN
BMPR1B	TKL	R			CEU/YRI
BMPR2	TKL	R	CEU	0.042	YRI
BMP15		*L*			*YRI*[g]
BMPR1A	TKL	R			YRI/ASN
BMPR1B	TKL	R			CEU/YRI
BMPR2	TKL	R	CEU	0.042	YRI

Abbreviations: L, ligand gene; R, kinase receptor gene; SNP, single-nucleotide polymorphism; TK, tyrosine kinase; TLK, tyrosine kinase like.

[a] Represents TK and TLK.

[b] Represents *L* and R. Kinase genes follow the genes encoding their cognate ligands, which are italicized.

[c] Empirical *P*-values obtained from the Happlotter.[15]

[d] CEU, Caucasian individuals from Utah, USA; YRI, Yoruba people from Ibadan, Nigeria; ASN, Asian individuals from Beijing, China and Tokyo, Japan; all eSNPs are located in introns if not otherwise indicated.

[e] Indicating the existence of non-synonymous eSNPs in coding regions.

[f] Indicating the existence of eSNPs in untranslated regions.

[g] Indicating the existence of synonymous eSNPs in coding regions.

populations. These 23 ligand genes represent 74 ligand–kinase receptor relationships with 29 unique kinase genes belonging to two groups (TK: tyrosine kinase and TLK: tyrosine kinase like), among which 22 kinase genes contained eSNPs. Supplementary Table 2 shows the complete list of these 74 ligand–kinase receptor relationships.

Kinase–ligand pairs under recent positive selection

Among the 74 ligand–kinase receptor relationships (Supplementary Table 2), two ligand genes had evidence for recent positive selection: *BPM3* in the CEU samples and *BMP5* in the ASN samples. In addition, 11 kinase genes had evidence for recent positive selection (CEU: 4, YRI: 3 and ASN: 4 genes). Furthermore, two ligand–kinase pairs involving *BMPR2* (bone morphogenetic protein receptor, type II), *BMP3-BMPR2* and *BMP5-BMPR2*, showed evidence for recent positive selection in both the ligand and kinase genes. In particular, *BMP3* and its target kinase gene *BMPR2* have been under recent positive selection in the CEU samples, whereas *BMP5* showed evidence for recent positive selection in the ASN samples. Some examples are illustrated in Table 1 and in Figure 2.

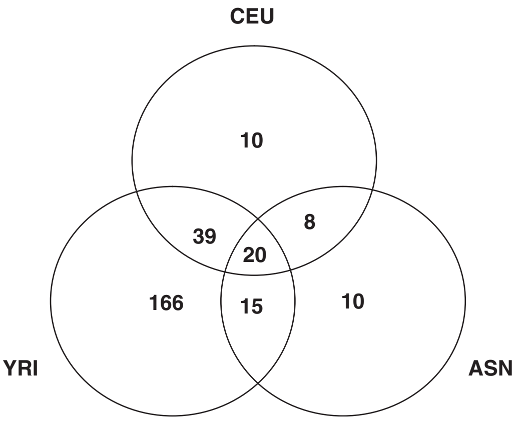

Figure 1 A Venn diagram of the kinases that contained population-specific SNPs (eSNPs) in different populations. CEU, Caucasian individuals from Utah, USA; YRI, Yoruba people from Ibadan, Nigeria; ASN, Asian individuals from Beijing, China and Tokyo, Japan.

DISCUSSION

The results of our study show that the human kinome, important in many different diseases, manifests significant genetic variation among major continental populations, in excess of that expected across the general genome background. This increased population diversity suggests a stronger adaptation of these genes within each population. Given that the kinases are hubs of various cellular functions, the stronger adaptation of these genes could have been critical for different populations to adapt to their new environments as *Homo sapiens* migrated from Africa to other continents. Another observation was that the African individuals represented by the YRI samples from Nigeria had a much larger number of kinases (240 genes) that contained eSNPs than did the Asian (53 genes) and Caucasian samples (77 genes), consistent with the observation that the YRI samples (∼74%) are more diverse than the CEU (∼15%) and ASN (∼7%) samples in terms of the proportion of total eSNPs.[21] Again, this difference may reflect the evolutionary history of human populations migrating from Africa to other continents, as the Africans are older populations containing more genetic variation. On the other hand, no particular kinase group was exceptionally enriched among the kinase genes containing eSNPs, suggesting that no particular kinase group(s) underwent faster evolution relative to the other groups.

As many kinases perform their cellular function by the activation of ligands, surveying the relationship of population genetic variation between the kinase and ligand pairs could shed some light on the evolution of these dynamic cellular components. Using a list of experimentally verified ligands that were obtained from the DLRP,[19] we found that the proportion of ligand genes that contained eSNPs was not different from that of the whole genome background. This suggests that the ligand genes may not be the major targets of adaptation for the signaling pathways involving kinases. On the other hand, some eSNP-containing ligand genes were found to share common kinase targets, but the ligand genes and their kinase targets may not necessarily contain

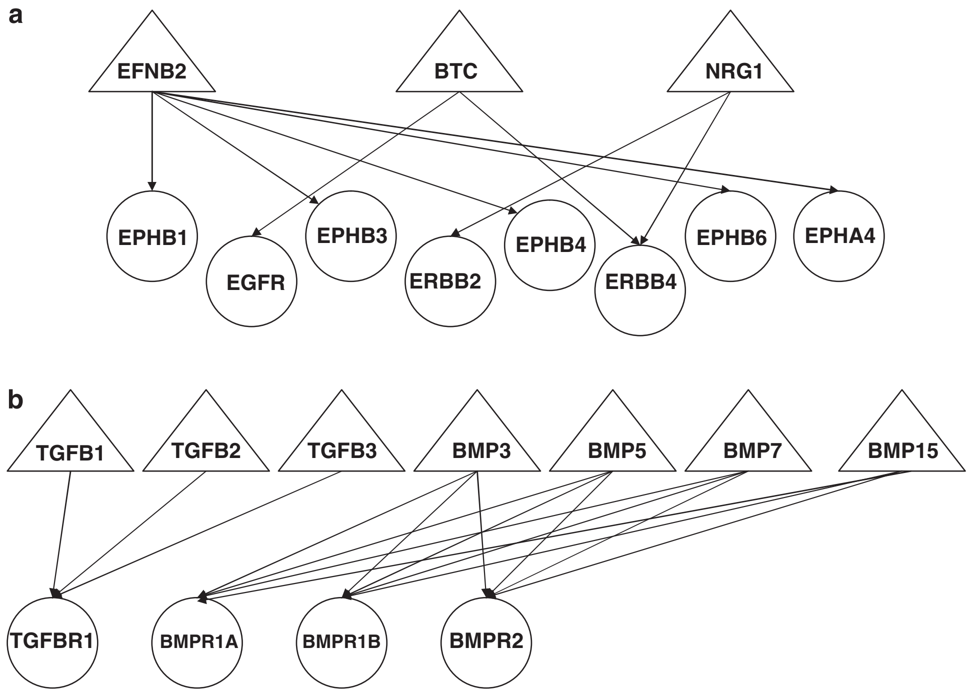

Figure 2 Population-specific SNPs (eSNPs)-containing ligand genes and their kinase targets. Triangles indicate ligand genes. Circles indicate kinases. Arrows link ligands to their targets. (**a**) The kinases belonging to the TK (tyrosine kinase) group; (**b**) The kinases belonging to the TLK (tyrosine kinase-like) group.

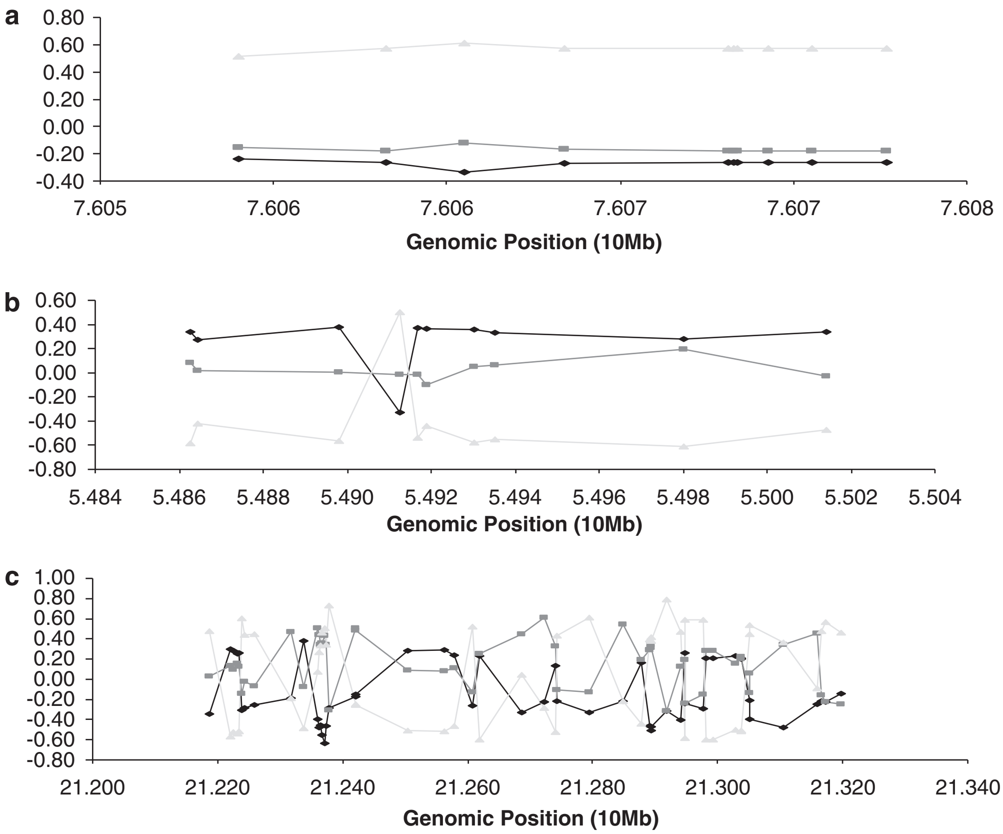

Figure 3 The nearest shrunken centroid method (NSCM) scores of the population-specific SNPs (eSNPs) of a ligand gene *BTC* and its kinase targets. X axis is the genomic position based on NCBI build 36. Y axis is the NSCM score, which was used to identify eSNPs. Gray: the CEU samples; Black: the ASN samples; and Light gray: the YRI samples. (**a**) *BTC* (Chr4). Ten intronic eSNPs are shown. (**b**) *EGFR* (Chr7), encoding a kinase receptor for BTC. Ten intronic eSNPs are shown. (**c**) *ERBB4* (Chr2), encoding a kinase receptor for BTC. Fifty-two intronic eSNPs are shown.

eSNPs in the same population(s) (Table 1). For example, BTC (betacellulin) and NRG1 (neuregulin 1) are common ligands of ERBB4 (v-erb-a erythroblastic leukemia viral oncogene homolog 4) (Figure 2a); bone morphogenetic proteins BMP3, BMP5, BMP7 and BMP15 are common ligands of BMPR1A (BMP receptor, type IA), BMPR1B (BMP receptor, type IB) and BMPR2 (BMP receptor, type II) (Figure 2b). However, these ligand genes and their kinase targets often contained eSNPs in different populations, as illustrated in Figure 3 for *BTC* and the genes encoding its two kinase targets (*EGFR* and *ERBB4*). Although *BTC* and *EGFR* contained eSNPs in the YRI samples, *ERBB4* contained eSNPs in all the three populations, suggesting that ligands and their kinase targets could be under different adaptation in each population. In addition, a search for signatures of recent positive selection in the kinase and ligand pairs showed that more kinases (11 genes) had significant |iHS|[17] scores than did ligands (2 genes) (Table 1). Therefore, it seems that the health-related disparities associated with kinase signaling pathways are more likely to be driven by the genetic variation in the kinase genes than by the genes encoding their cognate ligands. However, one limitation is that the current DLRP comprises only a subset of the ligands of kinase genes. In fact, only the TK and TLK groups of kinases are represented in the database. A more comprehensive list of kinase and ligand pairs may be necessary to evaluate the relationship of the evolutionary history of these genes.

Technically, the NSCM score[22] was used to identify eSNPs in the HapMap samples. It is a discriminating value, which is small if there is little difference between the classes or if the variation of the SNP distribution is large.[21] The NSCM has been proposed as a suitable approach to solving the classification problem when there are a large number of features (for example, ~1 million HapMap SNPs) from which to predict a relatively small number of classes (for example, three HapMap populations).[22] A limitation of this score is that the cutoff is empirical and a comprehensive evaluation relative to other metrics is lacking. There are also some important limitations of using the HapMap genotypic data. As the CEU samples were collected decades earlier[24] than the YRI and ASN samples,[13,14] certain biases may occur because of the differences in cell line culture and transformation techniques.[25] In addition, the current HapMap samples were obtained from individuals of the three major human populations. More samples from other populations (for example, the Phase 3 HapMap samples such as the Mexican Americans) will greatly benefit the investigation of population genetic variation in these genes. Furthermore, although the HapMap genotypic data are extensive (>3 million SNPs), the project was designed to cover only common genetic variants (minor allele frequency >5%).[13,14] As the human genome may contain ~10 million SNPs, untyped or unknown genetic variants may also contribute significantly to the population differences in the human kinome and their ligands. Deep resequencing projects (for example, the 1000 Genomes Project[26]) using next-generation sequencing technologies[27,28] may allow researchers to more comprehensively catalog the human genetic variants,[29] thus improving our understanding of the genetic variation in these important genes in the future.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This Pharmacogenetics of Anticancer Agents Research (PAAR) Group (http://www.pharmacogenetics.org/) study was supported by the NIH/NIGMS grant U01GM61393. Data will be deposited into PharmGKB (supported by the NIH/NIGMS Pharmacogenetics Research Network and Database grant U01GM61374, http://www.pharmgkb.org/). We are grateful to Drs Anna Di Rienzo, Chang Sun and Wanqing Liu for their helpful discussions.

1 Hunter, T. Signaling—2000 and beyond. *Cell* **100,** 113–127 (2000).
2 Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298,** 1912–1934 (2002).
3 Fabbro, D., Ruetz, S., Buchdunger, E., Cowan-Jacob, S. W., Fendrich, G., Liebetanz, J. *et al.* Protein kinases as targets for anticancer agents: from inhibitors to useful drugs. *Pharmacol. Ther.* **93,** 79–98 (2002).
4 Dassonville, O., Bozec, A., Fischel, J. L. & Milano, G. EGFR targeting therapies: monoclonal antibodies versus tyrosine kinase inhibitors. Similarities and differences. *Crit. Rev. Oncol. Hematol.* **62,** 53–61 (2007).
5 Harari, P. M., Allen, G. W. & Bonner, J. A. Biology of interactions: antiepidermal growth factor receptor agents. *J. Clin. Oncol.* **25,** 4057–4065 (2007).
6 Ratain, M. J., Eisen, T., Stadler, W. M., Flaherty, K. T., Kaye, S. B., Rosner, G. L. *et al.* Phase II placebo-controlled randomized discontinuation trial of sorafenib in patients with metastatic renal cell carcinoma. *J. Clin. Oncol.* **24,** 2505–2512 (2006).
7 Escudier, B., Eisen, T., Stadler, W. M., Szczylik, C., Oudard, S., Siebels, M. *et al.* Sorafenib in advanced clear-cell renal-cell carcinoma. *N. Engl. J. Med.* **356,** 125–134 (2007).
8 Cohen, P. Protein kinases—the major drug targets of the twenty-first century? *Nat. Rev. Drug. Discov.* **1,** 309–315 (2002).
9 Fejerman, L. & Ziv, E. Population differences in breast cancer severity. *Pharmacogenomics* **9,** 323–333 (2008).
10 Taron, M., Ichinose, Y., Rosell, R., Mok, T., Massuti, B., Zamora, L. *et al.* Activating mutations in the tyrosine kinase domain of the epidermal growth factor receptor are associated with improved survival in gefitinib-treated chemorefractory lung adenocarcinomas. *Clin. Cancer. Res.* **11,** 5878–5885 (2005).
11 Liu, W., Wu, X., Zhang, W., Montenegro, R. C., Fackenthal, D. L., Spitz, J. A. *et al.* Relationship of EGFR mutations, expression, amplification, and polymorphisms to epidermal growth factor receptor inhibitors in the NCI60 cell lines. *Clin. Cancer. Res.* **13,** 6788–6795 (2007).
12 Gomase, V. S. & Tagore, S. Kinomics. *Curr. Drug. Metab.* **9,** 255–258 (2008).
13 The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
14 The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
15 Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
16 Zhang, W., Ratain, M. J. & Dolan, M. E. The HapMap resource is providing new insights into ourselves and its application to pharmacogenomics. *Bioinform. Biol. Insights* **2,** 15–23 (2008).
17 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4,** e72 (2006).
18 Zhang, W. & Dolan, M. E. Exploring the evolutionary history of the differentially expressed genes between human populations: action of recent positive selection. *Evol. Bioinform.* **4,** 171–179 (2008).
19 Graeber, T. G. & Eisenberg, D. Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat. Genet.* **29,** 295–300 (2001).
20 Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32,** D449–D451 (2004).
21 Park, J., Hwang, S., Lee, Y. S., Kim, S. C. & Lee, D. SNP@Ethnos: a database of ethnically variant single-nucleotide polymorphisms. *Nucleic Acids Res.* **35,** D711–D715 (2007).
22 Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA* **99,** 6567–6572 (2002).
23 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **B** (57), 289–300 (1995).
24 Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J. M. & White, R. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6,** 575–577 (1990).
25 Zhang, W. & Dolan, M. E. On the challenges of the HapMap resource. *Bioinformation* **2,** 238–239 (2008).
26 The 1000 Genomes Project Meeting Report: a workshop to plan a deep catalog of human genetic variation (http://www.1000genomes.org) (2007).
27 Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–141 (2008).
28 Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387–402 (2008).
29 Zhang, W. & Dolan, M. E. Beyond the HapMap genotypic data: prospects of deep resequencing projects. *Curr. Bioinformatics* **3,** 178–182 (2008).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)