・遺伝様式を確認し遺伝子の伝わり方や発現について説明できる

7） 遺伝生化学・遺伝薬理学

・生体内分子の機能と代謝について基本的事項を理解し，遺伝医学的に説明できる

8） 生殖・発生遺伝学

・生殖の機構を理解し，その異常を説明できる

・発生の分子機構について基本的事項を理解し，説明できる

9） 体細胞遺伝学

・体細胞遺伝学について基本的事項を理解し，説明できる

10） 腫瘍遺伝学

・癌関連遺伝子を説明できる

・腫瘍の発生機序を遺伝学的に説明できる

・遺伝性腫瘍について説明できる

11） 免疫遺伝学

・免疫応答の遺伝について基本的事項を理解し，説明できる

・血液型の種類と遺伝について理解し，説明できる

・組織適合性とその遺伝について説明できる

12） 遺伝医学・遺伝医療

・遺伝医療を実践するにあたり，臨床遺伝専門医と遺伝カウンセラーの専門的な役割を理解し，医療・保健・福祉システムとの効果的連携について説明できる．

・遺伝カウンセリングが対象とする主な疾患について，臨床像，疫学，診断法，治療，再発予防，ケアの基本事項について理解し，説明できる（メンデル遺伝病，多因子遺伝病，染色異常，ガン，生活習慣病）

・臨床遺伝学における遺伝学的異質性の診断の重要性を理解し，遺伝子診断の概略を説明できる

・遺伝マーカースクリーニングの概略を理解し，説明できる

・出生前医療（受精卵・出生前診断を含む）の基本的事項を理解し，説明できる

・遺伝子治療の現状について理解し，説明できる

・わが国の遺伝医療システムについて理解し，説明できる

・ゲノム機能科学について現状と将来の展望について理解し，説明できる

2．カウンセリング理論と技法

・カウンセリングの主要理論と技法を理解している

・人間発達理論やパーソナリティー理論の基本を理解し，主要な心理検査法を理解している

・アセスメント面接法と行動観察法の基本を理解している

・主要な精神科的疾患の臨床的特徴を理解し，精神科領域の専門職との連携について理解している

・危機介入理論を理解し，危機的状況のアセスメント，危機介入技術について理解している

3．遺伝医療と倫理

・生命倫理学の歴史，インフォームドコンセント，先端医療・生殖医療の現場における生命倫理学的諸問題を理解している

・遺伝医療に関する国内外の規制等を理解している

・遺伝医療特有の倫理問題を理解している

4．遺伝医療と社会

・社会福祉の歴史，社会保障，公的扶助，児童・母子福祉，障害者福祉，老人福祉，地域福祉，医療福祉など社会福祉の基礎を理解している

・社会福祉援助技術（ソーシャルワーク）の基礎を理解している

・保健医療福祉関係法規を理解している

II 遺伝カウンセリングの実践技術の目標

1．クライエントとの人間関係を築くことができる

2．クライエントの問題事・心配事を明確化できる

3．クライエントの持つ遺伝学的背景をアセスメントできる

・家系資料を適切な方法で収集し，家系図を書ける

・必要な遺伝学的情報を得ることができる

・クライエントが受けている医療について必要な情報を得ることができる

・遺伝問題の有無を判断することができる

・再発危険率の推定ができる

・アセスメントの結果を科学的に記録できる

4．遺伝問題から生じる心理・社会的問題を支援できる．

・心理・社会的問題を明確化できる

・クライエントの問題認知状況をアセスメントできる

・クライエントのコーピングをアセスメントし，適切に介入できる

・グリーフカウンセリング，危機介入ができる

・カウンセリングの限界を理解し，他の専門職と連携する時期について判断できる

5．クライアントの課題・問題の明確化・意思決定に必要な情報を提供できる

・人類の遺伝学的荷重とクライエント自身が抱える遺伝学的リスクをわかりやすく説明できる

・検査・診断・治療・生活に関連した情報を提供できる

・クライアントが活用できる専門職・機関に関する情報を提供できる

・クライアントが活用できる社会資源に関する情報を提供できる

・クライアントの理解力に応じた方法で必要な情報に関して説明できる

6．クライアントの意思決定を支持し，支援する

・専門職・機関と連携をとることができる
・家族ダイナミックスを支援できる
・サポートグループへの紹介ができる
・個々の事例について適切にフォローアップを行うことができる

III　カウンセラーの態度目標

1．医療従事者の一員としての自覚をもって行動できる
・遺伝カウンセラーは医療技術を提供する立場ではないが，医療チームの一員であるとの自覚をもって行動できる
・遺伝カウンセラーが担当すべき業務範囲を理解し，クライエントから求められても診断類似行為や治療に関わる判断・指示を行わない．
・クライエントが受けている医療を理解し，主治医との人間関係を損なわないよう配慮できる
・臨床遺伝専門医やその他の専門職の役割を理解し，連携を重視して行動することができる
・最新の医療・遺伝医学に関する情報収集を行い，常に自己研鑽を怠らない
・遺伝カウンセリングの科学的な側面を理解し，科学的な思考ができるよう自己研鑽を怠らない
・カウンセリングについて科学的な記録を残し，適切な方法で管理できる
・守秘義務の原則を理解し，医療人として行動できる

2．カウンセラーとしてクライエントを支援する立場で行動できる
・カウンセラーの立場を理解し，常にクライエントの利益を考えて行動できる
・クライエントの人権を尊重し，家族や人間関係を配慮した態度で接することができる
・クライエントの不安に対しては常に共感的態度で接することができる
・クライエントの自律的決定を尊重し，非指示的態度で接することができる
・コミュニケーション技術や心理学的介入技術について，常に自己研鑽を怠らない

3．ELSI（倫理・法律・社会的事項）の基本的事項を理解し，社会人として公正な立場で行動できる
・生命の尊厳を重視する基本的態度でクライエントに接することができる
・法律，倫理規範，社会通念を配慮する基本的態度と倫理的に公正な態度でクライエントに接することができる
・現代医療や社会的対応の限界を理解し，クライエントにとって最良の選択を可能にするよう調整や支援をすることができる

おわりに

2005年7月，お茶の水女子大学大学院で「遺伝カウンセラーになりたい人のために」という講演会を開催したところ，予想を上回る多くの方々が集まり，とても熱心に私たちの話に耳を傾け，また質問を投げかけてくれました．私たちは遺伝カウンセラーという職業への関心が急速に高まっているのを実感し，驚き，またとても心強く思いました．そして，さまざまな情報や研究成果があふれる今，遺伝カウンセラーを解説する書籍が必要とされていると感じています．

認定遺伝カウンセリング制度委員会は，日本遺伝カウンセリング学会と日本人類遺伝学会の協力のもとに組織されています．本来はこの書籍を「認定遺伝カウンセラー制度委員会編」として出版したかったのですが，専門課程の教育が準備中の大学もあり，編集委員会を組織することが困難でした．そこで，監修者の1人である千代がこれまでの遺伝カウンセリング制度に関する研究班を担当してきた経験から，代表して編集を監督させていただきました．各養成専門課程の紹介に関する原稿を，こころよく執筆してくださった教員の先生方に深謝いたします．また，2006年4月の出版にこぎつけようという無理なお願いに理解を示してくださった，真興交易(株)医書出版部の編集部の皆さんにも心からお礼を申し上げます．

日本の専門職遺伝カウンセラー養成にかける認定養成専門課程教員一同の情熱が，読者の皆さんに届くことを夢見て．

遺伝カウンセラー　その役割と資格取得に向けて

2006年5月20日　第1版第1刷発行

監　修　千代豪昭
　　　　滝澤公子

発行者　梼内千一

発行所　真興交易(株)医書出版部

〒106-0047
東京都港区南麻布2-8-18
電話　(03)3798-3315(代)
振替　00170-0-147227

〈検印省略〉　印刷・製本　東京書籍印刷(株)

※定価はカバーに表示してあります

ISBN 4-88003-771-0 C3047
Printed in Japan

200707021A (2/3)

厚生労働科学研究費補助金

ヒトゲノム・再生医療等研究事業

ゲノムリテラシー向上のための人材育成と教育ツール開発に関する研究

平成19年度　総括研究報告書

研究成果の刊行物・別刷　分冊２/３（雑誌英文）

主任研究者　福嶋　義光

平成20(2008)年３月

研究成果の別刷り（雑誌英文）一覧

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
The International HapMap Consortium (composed of 72 institutes, including Fukushima Y)	A second generation human haplotype map of over 3.1 million SNPs.	Nature	449	851-861	2007
The International HapMap Consortium(composed of 72 institutes, including Fukushima Y)	Genome-wide detection and characterization of positive selection in human populations.	Nature	449	913-918	2007
Yoshida K, Wada T, Sakurai A, Wakui K, Ikeda S, Fukushima Y	Nation wide survey on predictive genetic testing for late-onset, incurable neurological diseases in Japan.	J Hum Genet	52	675-679	2007
Sakurai A, Katai M, Yamashita K, Mori J, Fukushima Y, Hashizume K	Long-term follow-up of patients with multiple endocrine neoplasia type 1.	Endocrine Journal	54	295-302	2007
Rotimi C, Leppert M, Matsuda I, Zeng C, Zhang H, Adebamowo C, Ajayi I, Aniagwu T, Dixon M, Fukushima Y, Macer D, Marshall P, Nkwodimmah C, Peiffer A, Royal C, Suda E, Zhao H, Wang VO, McEwen J, International HapMap Consortium	Community engagement and informed consent in the International HapMap project.	Community Genet	10	186-198	2007
Kosho T, Takahashi J, Momose T, Nakamura A, Sakurai A, Wada T, Yoshida K, Wakui K, Suzuki T, Kasuga K, Nishimura G, Kato H, Fukushima Y	Mandibuloacral dysplasia and a novel LMNA mutation in a woman with severe progressive skeletal changes.	Am J Med Genet	43A	2598-2603	2007

A second generation human haplotype map of over 3.1 million SNPs

The International HapMap Consortium*

We describe the Phase II HapMap, which characterizes over 3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from four geographically diverse populations and includes 25–35% of common SNP variation in the populations surveyed. The map is estimated to capture untyped common variation with an average maximum r^2 of between 0.9 and 0.96 depending on population. We demonstrate that the current generation of commercial genome-wide genotyping products captures common Phase II SNPs with an average maximum r^2 of up to 0.8 in African and up to 0.95 in non-African populations, and that potential gains in power in association studies can be obtained through imputation. These data also reveal novel aspects of the structure of linkage disequilibrium. We show that 10–30% of pairs of individuals within a population share at least one region of extended genetic identity arising from recent ancestry and that up to 1% of all common variants are untaggable, primarily because they lie within recombination hotspots. We show that recombination rates vary systematically around genes and between genes of different function. Finally, we demonstrate increased differentiation at non-synonymous, compared to synonymous, SNPs, resulting from systematic differences in the strength or efficacy of natural selection between populations.

Advances made possible by the Phase I haplotype map

The International HapMap Project was launched in 2002 with the aim of providing a public resource to accelerate medical genetic research. The objective was to genotype at least one common SNP every 5 kilobases (kb) across the euchromatic portion of the genome in 270 individuals from four geographically diverse populations[1,2]: 30 mother–father–adult child trios from the Yoruba in Ibadan, Nigeria (abbreviated YRI); 30 trios of northern and western European ancestry living in Utah from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (CEU); 45 unrelated Han Chinese individuals in Beijing, China (CHB); and 45 unrelated Japanese individuals in Tokyo, Japan (JPT). The YRI samples and the CEU samples each form an analysis panel; the CHB and JPT samples together form an analysis panel. Approximately 1.3 million SNPs were genotyped in Phase I of the project, and a description of this resource was published in 2005 (ref. 3).

The initial HapMap Project data had a central role in the development of methods for the design and analysis of genome-wide association studies. These advances, alongside the release of commercial platforms for performing economically viable genome-wide genotyping, have led to a new phase in human medical genetics. Already, large-scale studies have identified novel loci involved in multiple complex diseases[4,5]. In addition, the HapMap data have led to novel insights into the distribution and causes of recombination hotspots[3,6], the prevalence of structural variation[7,8] and the identity of genes that have experienced recent adaptive evolution[3,9]. Because the HapMap cell lines are publicly available, many groups have been able to integrate their own experimental data with the genome-wide SNP data to gain new insight into copy-number variation[10], the relationship between classical human leukocyte antigen (HLA) types and SNP variation[11], and heritable influences on gene expression[12–14]. The ability to combine genome-wide data on such diverse aspects of genetic variation with molecular phenotypes collected in the same samples provides a powerful framework to study the connection of DNA sequence to function.

*Lists of participants and affiliations appear at the end of the paper.

In Phase II of the HapMap Project, a further 2.1 million SNPs were successfully genotyped on the same individuals. The resulting HapMap has an SNP density of approximately one per kilobase and is estimated to contain approximately 25–35% of all the 9–10 million common SNPs (minor allele frequency (MAF) ≥ 0.05) in the assembled human genome (that is, excluding gaps in the reference sequence alignment; see Supplementary Text 1), although this number shows extensive local variation. This paper describes the Phase II resource, its implications for genome-wide association studies and additional insights into the fine-scale structure of linkage disequilibrium, recombination and natural selection.

Construction of the Phase II HapMap

Most of the additional genotype data for the Phase II HapMap were obtained using the Perlegen amplicon-based platform[15]. Briefly, this platform uses custom oligonucleotide arrays to type SNPs in DNA segmentally amplified via long-range polymerase chain reaction (PCR). Genotyping was attempted at 4,373,926 distinct SNPs, which corresponds, with exceptions (see Methods), to nearly all SNPs in dbSNP release 122 for which an assay could be designed. Additional submissions were included from the Affymetrix GeneChip Mapping Array 500K set, the Illumina HumanHap100 and HumanHap300 SNP assays, a set of ~11,000 non-synonymous SNPs genotyped by Affymetrix (ParAllele) and a set of ~4,500 SNPs within the extended major histocompatibility complex (MHC)[11]. Genotype submissions were subjected to the same quality control (QC) filters as described previously (see Methods) and mapped to NCBI build 35 (University of California at Santa Cruz (UCSC) hg17) of the human genome. The re-mapping of SNPs from Phase I of the project identified 21,177 SNPs that had an ambiguous position or some other feature indicative of low reliability; these are not included in the filtered Phase II data release. All genotype data are available from the HapMap Data Coordination Center (http://www.hapmap.org) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP); analyses described in this paper refer to release 21a. Three data sets are available: 'redundant unfiltered'

©2007 Nature Publishing Group

contains all genotype submissions, 'redundant filtered' contains all submissions that pass QC, and 'non-redundant filtered' contains a single QC+ submission for each SNP in each analysis panel.

The QC filters remove SNPs showing gross errors. However, it is also important to understand the magnitude and structure of more subtle genotyping errors among SNPs that pass QC. We therefore carried out a series of analyses to assess the influence of the long-range PCR amplicon structure on genotyping error, the concordance rates between genotype calls from different genotyping platforms and between those platforms and re-sequencing assays, as well as the rates of false monomorphism and mis-mapping of SNPs (see Supplementary Text 2, Supplementary Figs 1–3 and Supplementary Tables 1–4). We estimate that the average per genotype accuracy is at least 99.5%. However, there are higher rates of missing data and genotype discrepancies at non-reference alleles, with some clustering of errors resulting from the amplicon design and a few incorrectly mapped SNPs.

Table 1 shows the numbers of SNPs attempted and converted to QC+ SNPs in each analysis panel (Supplementary Table 5 shows a breakdown by each major submission). Haplotypes and missing data were estimated for each analysis panel separately using both trio information and statistical methods based on the coalescent model (see Methods). To enable cross-population comparisons, a consensus data set was created consisting of 3,107,620 SNPs that were QC+ in all analysis panels and polymorphic in at least one analysis panel. The equivalent figure from Phase I was 931,340 SNPs. Unless stated otherwise, all analyses have been carried out on the consensus data set. An additional set of haplotypes was created for those SNPs in the consensus where a putative ancestral state could be assigned by comparison of the human alleles to the orthologous position in the chimpanzee and rhesus macaque genomes.

The variation in SNP density within the Phase II HapMap is shown in Fig. 1. On average there are 1.14 genotyped polymorphic SNPs per kilobase (average spacing is 875 base pairs (bp)) and 98.6% of the assembled genome is within 5 kb of the nearest polymorphic SNP. Still, there is heterogeneity in genotyped SNP density at both broad (Fig. 1a) and fine (Fig. 1b) scales. Furthermore, there are systematic changes in genotyped SNP density around genomic features including genes (Fig. 1c).

The Phase II HapMap differs from the Phase I HapMap not only in SNP spacing, but also in minor allele frequency distribution and patterns of linkage disequilibrium (Supplementary Fig. 4). Because the criteria for choosing additional SNPs did not include consideration of SNP spacing or preferential selection for high MAF, the SNPs added in Phase II are, on average, more clustered and have lower MAF than the Phase I SNPs. Because MAF predictably influences the distribution of linkage disequilibrium statistics, the average r^2 at a given physical distance is typically lower in Phase II than in Phase I; conversely, the $|D'|$ statistic is typically higher (data not shown). One notable consequence is that the Phase II HapMap includes a better representation of rare variation than the Phase I HapMap.

The increased resolution provided by Phase II of the project is illustrated in Fig. 2. Broadly, an additional SNP added to a region shows one of three patterns. First, it may be very similar in distribution to SNPs present in Phase I. Second, it may provide detailed resolution of haplotype structure (for example, a group of chromosomes with identical local haplotypes in Phase I can be shown in Phase II to carry

Table 1 | Summary of Phase II HapMap data (release 21)

Phase	SNP categories	Analysis panel		
		YRI	CEU	CHB+JPT
I	Assays submitted	1,304,199	1,344,616	1,306,125
	Passed QC	1,177,312 (90%)	1,217,902 (91%)	1,187,800 (91%)
	Did not pass QC	126,887 (10%)	126,714 (9%)	118,325 (9%)
	>20% missing	82,463 (65%)	95,684 (76%)	78,323 (66%)
	>1 duplicate inconsistent	6,049 (5%)	5,126 (4%)	9,242 (8%)
	>1 mendelian error	18,916 (15%)	11,310 (9%)	N/A
	<0.001 Hardy-Weinberg *P*-value	10,265 (8%)	8,922 (7%)	13,722 (12%)
	Other failures	19,345 (15%)	13,858 (11%)	20,674 (17%)
II	Assays submitted	5,044,989	5,044,996	5,043,775
	Passed QC	3,150,433 (62%)	3,204,709 (64%)	3,244,897 (64%)
	Did not pass QC	1,894,556 (38%)	1,840,287 (36%)	1,798,878 (36%)
	>20% missing	1,419,000 (75%)	1,398,166 (76%)	1,403,543 (78%)
	>1 duplicate inconsistent	0 (0%)	0 (0%)	6,617 (0%)
	>1 mendelian error	172,339 (9%)	127,923 (7%)	N/A
	<0.001 Hardy-Weinberg *P*-value	96,231 (5%)	82,268 (4%)	108,880 (6%)
	Other failures	334,511 (18%)	337,906 (18%)	340,370 (19%)
Overall	Assays submitted	6,349,188	6,389,612	6,349,900
	Passed QC	4,327,745 (68%)	4,422,611 (69%)	4,432,697 (70%)
	Did not pass QC	2,021,443 (32%)	1,967,001 (31%)	1,917,203 (30%)
	>20% missing	1,501,463 (74%)	1,493,850 (76%)	1,481,866 (77%)
	>1 duplicate inconsistent	6,049 (0%)	5,126 (0%)	15,859 (1%)
	>1 mendelian error	191,255 (9%)	139,233 (7%)	N/A
	<0.001 Hardy-Weinberg *P*-value	106,496 (5%)	91,190 (5%)	122,602 (6%)
	Other failures	353,856 (18%)	351,764 (18%)	361,044 (19%)
	Non-redundant (unique) SNPs	3,796,934	3,868,157	3,890,416
	Monomorphic	861,299 (23%)	1,246,183 (32%)	1,410,152 (36%)
	Polymorphic	2,935,635 (77%)	2,621,974 (68%)	2,480,264 (64%)

SNP categories	All analysis panels
Unique QC-passed SNPs	4,000,107
Passed in one analysis panel	88,140 (2%)
Passed in two analysis panels	268,534 (7%)
Passed in three analysis panels (QC+3)	3,643,433 (91%)
QC+3 and monomorphic across three analysis panels	535,813
QC+3 and polymorphic in at least one analysis panel	3,107,620
QC+3 and polymorphic in all three analysis panels	2,006,352
QC+3 and MAF ≥ 0.05 in at least one of three analysis panels	2,819,322

©2007 Nature Publishing Group

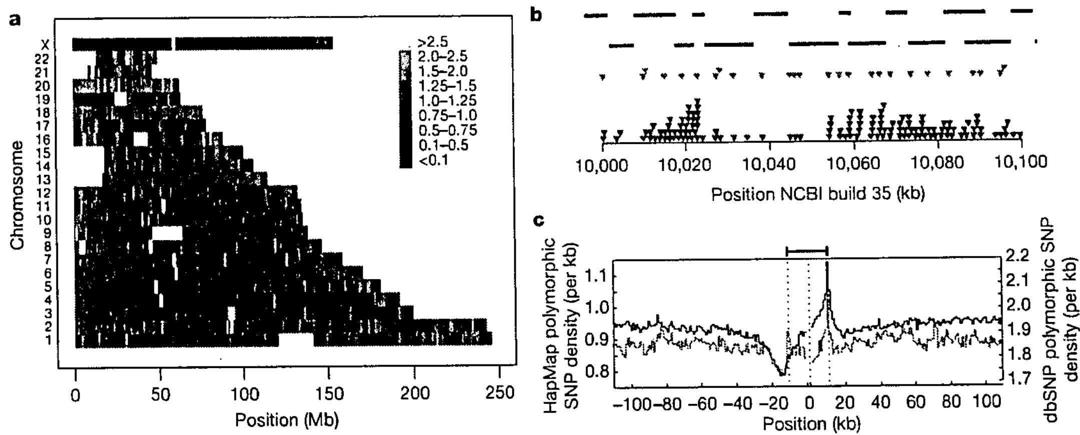

Figure 1 | SNP density in the Phase II HapMap. a, SNP density across the genome. Colours indicate the number of polymorphic SNPs per kb in the consensus data set. Gaps in the assembly are shown as white. **b**, Example of the fine-scale structure of SNP density for a 100-kb region on chromosome 17 showing Perlegen amplicons (black bars), polymorphic Phase I SNPs in the consensus data set (red triangles) and polymorphic Phase II SNPs in the consensus data set (blue triangles). Note the relatively even spacing of Phase I SNPs. **c**, The distribution of polymorphic SNPs in the consensus Phase II HapMap data (blue line and left-hand axis) around coding regions. Also shown is the density of SNPs in dbSNP release 125 around genes (red line and right-hand axis). Values were calculated separately 5′ from the coding start site (the left dotted line) and 3′ from the coding end site (right dotted line) and were joined at the median midpoint position of the coding unit (central dotted line).

multiple related haplotypes). Third, the novel SNP (or group of added SNPs) may reveal previously missed recombinant haplotypes. The extent to which each type of event occurs varies among populations and chromosomal regions. The greatest gains in resolution, in terms of identifying new recombinant haplotypes and haplotype groupings, occur in YRI. Consequently, the Phase II HapMap provides increased resolution in the estimated fine-scale genetic map and improved power to detect and localize recombination hotspots (Fig. 2b).

The use of the Phase II HapMap in association studies

The increased SNP density of the Phase II HapMap has already been extensively exploited in genome-wide studies of disease association. In this section, we quantify the gain in resolution and outline how the HapMap data can be used to improve the power of association studies.

Improved coverage of common variation. We previously predicted that the vast majority of common SNPs would be correlated to Phase II HapMap SNPs by extrapolation from the ten HapMap ENCODE regions[3]. Using the actual Phase II marker spacing and frequency distributions (Table 2), we repeated the simulations and estimate that Phase II HapMap marker sets capture the overwhelming majority of all common variants at high r^2. For common variants (MAF $\geq$ 0.05) the mean maximum r^2 of any SNP to a typed one is 0.90 in YRI, 0.96 in CEU and 0.95 in CHB+JPT. The impact of the

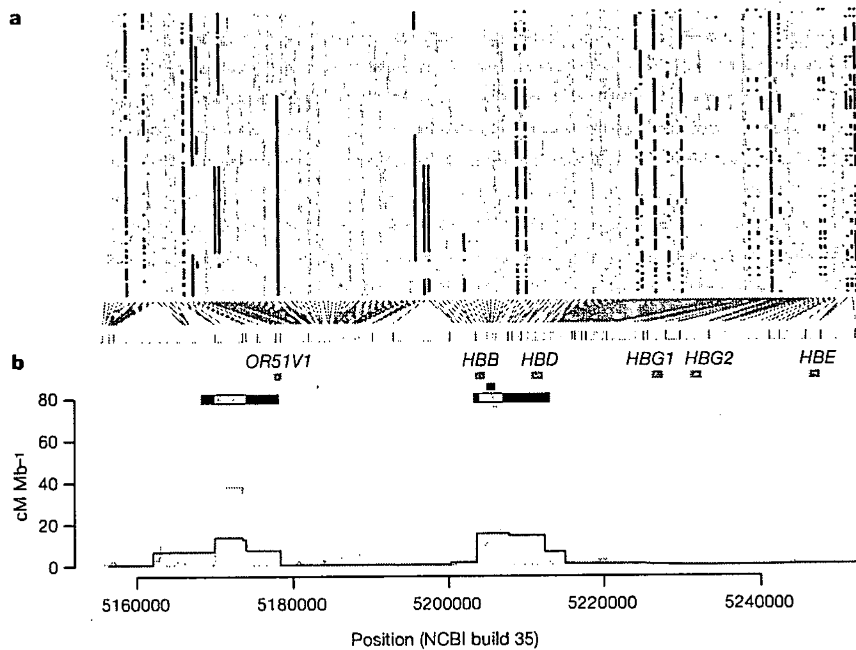

Figure 2 | Haplotype structure and recombination rate estimates from the Phase II HapMap. a, Haplotypes from YRI in a 100 kb region around the β-globin (*HBB*) gene. SNPs typed in Phase I are shown in dark blue. Additional SNPs in the Phase II HapMap are shown in light blue. Only SNPs for which the derived allele can be unambiguously identified by parsimony (by comparison with an outgroup sequence) are shown (89% of SNPs in the region); the derived allele is shown in colour. **b**, Recombination rates (lines) and the location of hotspots (horizontal blue bars) estimated for the same region from the Phase I (dark blue) and Phase II HapMap (light blue) data. Also shown are the location of genes within the region (grey bars) and the location of the experimentally verified recombination hotspot[57,58] at the 5′ end of the *HBB* gene (black bar).

©2007 Nature Publishing Group

Table 2 | Estimated coverage of the Phase II HapMap in the ten HapMap ENCODE regions

Panel	MAF bin	Phase I HapMap[3]		Phase II HapMap			
				Pairwise linkage disequilibrium		Additional 2-SNP tests	
		$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2
YRI	≥0.05	45	0.67	82	0.90	87	0.93
	<0.05			61	0.76	62	0.78
	0.05–0.10			81	0.89	81	0.89
	0.10–0.25			90	0.94	90	0.95
	0.25–0.50			87	0.93	92	0.96
CEU	≥0.05	74	0.85	93	0.96	95	0.97
	<0.05			70	0.79	72	0.81
	0.05–0.10			87	0.92	88	0.93
	0.10–0.25			94	0.96	95	0.97
	0.25–0.50			95	0.97	97	0.98
CHB+JPT	≥0.05	72	0.83	92	0.95	95	0.97
	<0.05			65	0.74	65	0.74
	0.05–0.10			81	0.89	82	0.89
	0.10–0.25			90	0.94	90	0.95
	0.25–0.50			94	0.96	97	0.98

2-SNP tests, linkage disequilibrium to haplotypes formed from two nearby SNPs.

Table 3 | Number of tag SNPs required to capture common (MAF ≥ 0.05) Phase II SNPs

Threshold	YRI	CEU	CHB+JPT
$r^2 \geq 0.5$	627,458	290,969	277,831
$r^2 \geq 0.8$	1,093,422	552,853	520,111
$r^2 = 1.0$	1,616,739	1,024,665	1,078,959

increased density of the Phase II HapMap is most notable in YRI (in the Phase I HapMap the mean maximum r^2 was 0.67). Similar results are found if a threshold of $r^2 \geq 0.8$ is used to determine whether an SNP is captured (Table 2). As expected, very common SNPs with MAF > 0.25 are captured extremely well (mean maximum r^2 of 0.93 in YRI to 0.97 in CEU), whereas rarer SNPs with MAF < 0.05 are less well covered (mean maximum r^2 of 0.74 in CHB+JPT to 0.76 in YRI). The latter figure is probably an overestimate because it is based on lower frequency SNPs discovered via re-sequencing 48 HapMap individuals, and does not include a much larger number of very rare SNPs. We also assessed the increase in coverage provided by using two-SNP haplotypes as proxies for SNPs that are poorly captured by single SNPs[16] (Table 2). These two-SNP haplotypes lead to a modest increase in mean maximum r^2 of 0.01 to 0.03 across all allele frequencies. However, in some regions, particularly where marker density is low, gains from multi-marker and imputation approaches in practical situations can be substantial (see below).

Currently, the Phase II HapMap provides the most complete available resource for selecting tag SNPs genome-wide. Using a simple pairwise tagging approach, we find that 1.09 million SNPs are required to capture all common Phase II SNPs with $r^2 \geq 0.8$ in YRI, with slightly more than 500,000 required in CEU and CHB+JPT (Table 3). These numbers are approximately twice those required to capture SNPs in the Phase I HapMap (which has one-third as many SNPs). The number of SNPs required to achieve perfect tagging ($r^2 = 1.0$) in each analysis panel is almost double that required to achieve the $r^2 \geq 0.8$ threshold. It becomes increasingly expensive to improve the coverage afforded by tags from the Phase I and, now, the Phase II HapMap, because additional tag SNPs are unlikely to capture large groups of additional SNPs.

Phase II HapMap and genome-wide association studies. Although the efficient choice of tag SNPs is one use of the Phase II HapMap, for most disease studies the tag SNPs genotyped will be primarily determined by the choice of a commercial platform for the experiment[17,18]. Using Phase II data, we estimated the coverage of several available products on which genome-wide association studies are already underway (Table 4). Similar to earlier estimates[17,18], these products typically perform well in CEU and CHB+JPT, and some also perform well in YRI. For example, arrays of approximately 500,000 SNPs capture 68–88% (depending on selection method) of all HapMap Phase II variation with $r^2 \geq 0.8$ in CEU. SNPs that are not included in the Phase II HapMap will be covered more poorly because most genotyping products were designed using HapMap data.

HapMap data have several additional roles in the analysis of disease-association studies using fixed marker sets. For example, the high-quality haplotype information within the Phase II HapMap can be used to aid the phasing of genotype data from new samples because additional haplotypes are likely to be locally very similar to at least one haplotype in the Phase II data. By a similar argument, missing genotypes can potentially be inferred through comparison to the Phase II haplotypes. Genotypes may be missing either because of genotyping failure or because the SNP was not assayed within the experiment. Therefore, the HapMap haplotypes provide a way of *in silico* genotyping Phase II SNPs that were not included in the experiment.

Although there is no clear consensus yet about the role of SNP imputation in the analysis of genome-wide association studies, high imputation accuracy can be achieved using model-based methods[19–23] and can lead to an increase in power[23,24]. To illustrate the possibilities, in the 500-kb HapMap ENCODE region on 8q24.11 (Supplementary Fig. 5) we evaluated imputation of Phase II SNPs from the Affymetrix GeneChip 500K array. To do this, we used a

Table 4 | Estimated coverage of commercially available fixed marker arrays

Platform*	YRI		CEU		CHB+JPT	
	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2
Affymetrix GeneChip 500K	46	0.66	68	0.81	67	0.80
Affymetrix SNP Array 6.0	66	0.80	82	0.90	81	0.89
Illumina HumanHap300	33	0.56	77	0.86	63	0.78
Illumina HumanHap550	55	0.73	88	0.92	83	0.89
Illumina HumanHap650Y	66	0.80	89	0.93	84	0.90
Perlegen 600K	47	0.68	92	0.94	84	0.90

* Assuming all SNPs on the product are informative and pass QC; in practice these numbers are overestimates.

©2007 Nature Publishing Group

leave-one-out procedure to assess the accuracy of genotype prediction in the YRI. For SNPs with MAF ≥ 0.2, the average maximum r^2 to a typed SNP in the region is 0.59 compared to an average genotype prediction r^2 of 0.86. Furthermore, whereas 44% of such SNPs in the region have no single-marker proxy with $r^2 \geq 0.5$, fewer than 6% of the SNPs have a genotype imputation accuracy of $r^2 < 0.5$, establishing that accurate imputation can be achieved even in the population where linkage disequilibrium is the weakest.

New insights into linkage disequilibrium structure

The paradigm underlying association studies is that linkage disequilibrium can be used to capture associations between markers and nearby untyped SNPs. However, the Phase II HapMap has revealed several properties of linkage disequilibrium that illustrate the full complexity of empirical patterns of genetic variation. Two striking features are the long-range similarity among haplotypes, and SNPs that show almost no linkage disequilibrium with any other SNP.

The extent of recent common ancestry and segmental sharing. A simplified view of linkage disequilibrium is that genetic variation is organized in relatively short stretches of strong linkage disequilibrium (haplotype blocks), each containing only a few common haplotypes and separated by recombination hotspots across which little association remains[25]. Although this view has heuristic value, if chromosomes share a recent common ancestor then similarity between chromosomes can extend over considerable genetic distance and span multiple recombination hotspots[26]. The extent of such recent ancestry in the four populations surveyed here has not been characterized previously. Therefore we identified stretches of identity between pairs of chromosomes, both within and across individuals, reflecting autozygosity and identity-by-descent (IBD) (Fig. 3a). After first checking for stratification within each analysis panel (see Supplementary Text 3; none was found for YRI, CEU and JPT, and only small stratification was found for CHB), we calculated genome-wide probabilities of sharing 0, 1 or 2 chromosomes identical by descent for each pair of individuals (see Supplementary Text 4). In addition to identifying a few close relationships (as reported in HapMap Phase I[3]), we estimate that, on average, any two individuals from the same population share approximately 0.5% of their genome through recent IBD (Table 5). Using a hidden Markov model approach[27] (see Supplementary Text 5), we searched for such shared segments over 1-megabase (Mb) long and containing at least 50 SNPs, after first pruning the list of SNPs to remove local linkage disequilibrium. We find that 10–30% of pairs in each analysis panel share regions of extended identity resulting from sharing a common ancestor within 10–100 generations. These regions typically span hundreds of SNPs and can extend over tens of megabases (Table 5).

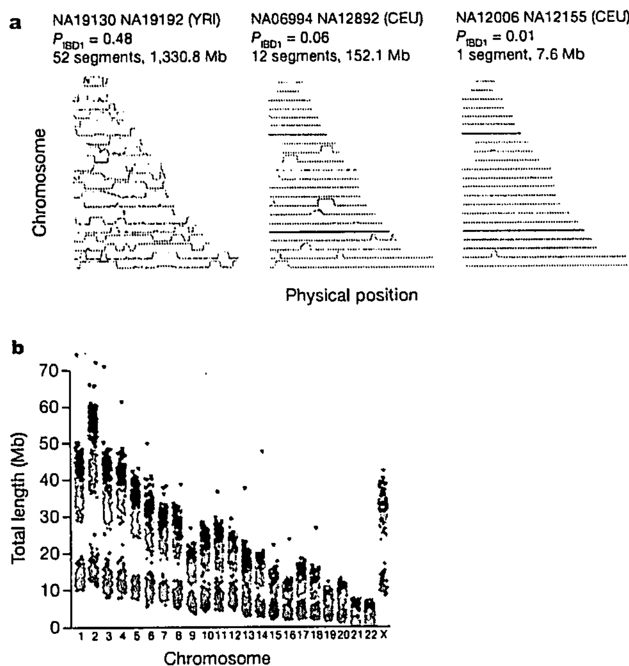

Figure 3 | The extent of recent co-ancestry among HapMap individuals. **a**, Three pairs of individuals with varying levels of identity-by-descent (IBD) sharing illustrate the continuum between very close and very distant relatedness and its relation to segmental sharing. The three pairs are: high sharing (NA19130 and NA19192 from YRI; previously identified as second-degree relatives[3]), moderate sharing (NA06994 and NA12892 from CEU) and low sharing (NA12006 and NA12155 from CEU). Along each chromosome, the probability of sharing at least one chromosome IBD is plotted, based on the HMM method described in Supplementary Text 5. Red sections indicate regions called as segments: in general, the proportion of the genome in segments is similar to each pair's estimated global relatedness. **b**, The extent of homozygosity on each chromosome for each individual in each analysis panel. Excludes segments <106 kb and chromosome X in males. Asterisk, NA12874, length = 107 Mb. YRI, green; CEU, orange; CHB, blue; JPT, magenta.

Similarly, extended stretches of homozygosity are indicative of recent inbreeding within populations[28,29]. Although short runs of homozygosity are commonplace, covering up to one-third of the genome and showing population differences reflective of ancient linkage disequilibrium patterns (Table 5 and Fig. 3b), very long homozygous runs exist that are clearly distinct from this process. Including two JPT individuals who have unusually high levels of homozygosity (NA18987 and NA18992) and one CEU individual (NA12874), we identified 79 homozygous regions over 3 Mb in 51 individuals, with many segments extending over 10 Mb (Supplementary Tables 7 and 8). Segments intersecting with suspected deletions were first removed from the analysis (Supplementary Text 6).

In studies of rare mendelian diseases, the extended haplotype sharing surrounding recent mutations, usually with a frequency of much less than 1%, has been exploited to great advantage through homozygosity mapping[30,31] and haplotype sharing[32] methods. In studies of common disease, extended haplotype sharing among patients potentially offers a route for identifying rare variants (MAF in the range of 1–5%) of high penetrance[33,34], which tend to be poorly captured through single-marker association with genome-wide arrays. To illustrate the idea, we identified SNPs where only two copies of the minor allele are present (referred to as '2-SNPs'), which have minor allele frequencies of 1–2%. We find that these are enriched approximately sevenfold (Table 5) among regions of IBD identified by the hidden Markov model approach. Notably, identification of IBD regions can be performed with the same genome-wide SNP data being

Table 5 | Relatedness, extended segmental sharing and homozygosity

Property	YRI	CEU	CHB	JPT
Number of pairs included	1,767	1,708	990	861
Mean identity by state (IBS) (%)	81.9	83.7	85.0	85.1
Mean identity by descent (IBD) (%)	0.04	0.34	0.36	0.42
Number of pairs with >1% IBD (%)	8.8	20.4	21.1	29.7
Number of pairs with one or more segment (%)	195 (11.0)	350 (20.5)	135 (13.6)	216 (25.1)
Total number of segments	250	427	146	273
Total distance spanned (Mb)	1,416	2,336	704	1,301
Mean segment length (Mb)	5.7	5.5	4.8	4.8
Maximum segment length (Mb)	51.7	56.2	15.0	25.3
Maximum segment length (Mb) (including close relatives)	141.4	128.5	N/A	N/A
Total number of 2-SNPs	6,219	9,220	8,174	8,750
Number of 2-SNPs in segments	109	162	116	132
2-SNP fold increase	6.7	7.3	7.6	7.0
Number of homozygous segments ($\times 10^3$)*	0.9	2.2	2.6	2.6
SNPs in homozygous segments ($\times 10^5$)	1.6	4.2	5.3	5.4
Total length of homozygous segments (Mb)	160	410	510	520

2-SNP, SNPs where only two copies of the minor allele are present.
* Homozygous segments >106 kb.

©2007 Nature Publishing Group

collected in large-scale association studies, making haplotype-sharing approaches an attractive and complementary analysis to standard SNP association tests, with the potential to identify rare variants associated with complex disease.

The distribution and causes of untaggable SNPs. Despite the SNP density of the Phase II HapMap, there are high-frequency SNPs for which no tag can be identified. Among high-frequency SNPs (MAF ≥ 0.2), we marked as untaggable SNPs to which no other SNP within 100 kb has an r^2 value of at least 0.2. In Phase II, approximately 0.5–1.0% of all high-frequency SNPs are untaggable and the proportion in YRI is approximately twice as high as in the other panels. Similar proportions are observed across the ten HapMap ENCODE regions.

To identify factors influencing the location of untaggable SNPs we considered their distribution relative to segmental duplications, repeat sequence, CpG dinucleotide density, regions of low SNP density, unusual allele frequency distribution, linkage disequilibrium patterns and recombination hotspots. We find no evidence for an enrichment of untaggable SNPs in segmental duplications or repeat sequence, as would be expected from mis-mapping of SNPs (2% and 35% of common SNPs lie in segmental duplications and repeat sequence, respectively, compared to 1.8% and 29%, respectively, of untaggable SNPs). Untaggable SNPs are slightly enriched in CpG islands (0.37% of common SNPs are in CpG islands compared to 1.4% of untaggable SNPs) and have slightly reduced MAF (Fig. 4). Most notably, untaggable SNPs are strongly enriched in regions of low linkage disequilibrium, particularly in recombination hotspots. To test whether these untaggable SNPs are themselves responsible for the identification of recombination hotspots, we eliminated them from 100 randomly chosen recombination hotspots and reassessed the evidence for a local peak in recombination. In all cases we still find evidence for a considerable increase in local recombination rate.

Over 50% of all untaggable SNPs lie within 1 kb of the centre of a detected recombination hotspot and over 90% are within 5 kb. Because only 3–4% of all SNPs lie within 1 kb from the centre of a detected recombination hotspot (16% are within 5 kb), this constitutes a marked enrichment and implies that at least 10% of all SNPs within 1 kb of hotspots are untaggable. The implication for association mapping is that when a region of interest contains a known hotspot it may be prudent to perform additional sequencing within the hotspot. Many of the variants identified in this manner will be untaggable SNPs that should be genotyped directly in association studies. From a biological perspective, the proximity of untaggable SNPs to the centre of hotspots suggests that they may lie within gene conversion tracts associated with the repair of double-strand breaks. Double-strand breaks are thought to resolve as crossover events only 5–25% of the time[35]. Consequently, SNPs lying near the centre of a hotspot are liable to be included within gene conversion tracts and will experience much higher effective recombination rates than predicted from crossover rates alone.

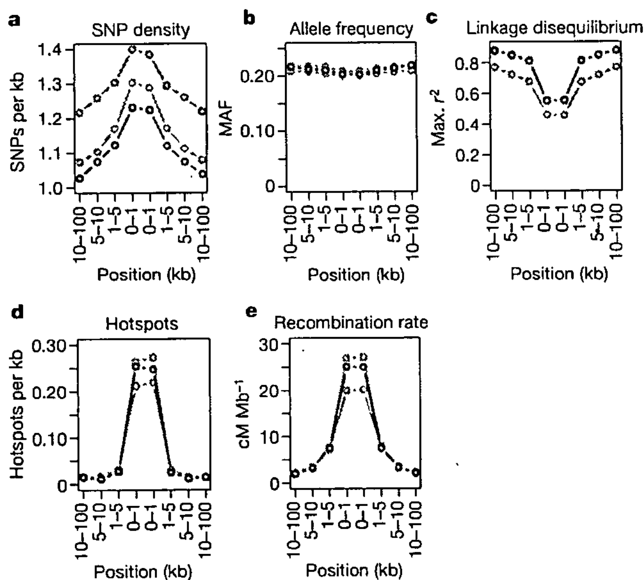

Figure 4 | Properties of untaggable SNPs. **a–e**, Properties of the genomic regions surrounding untaggable SNPs in terms of: **a**, the density of polymorphic SNPs within the consensus data set; **b**, mean minor allele frequency of polymorphic SNPs; **c**, maximum r^2 of SNPs to any others in the Phase II data; **d**, the density of estimated recombination hotspots (defined from hotspot centres); and **e**, the estimated mean recombination rate. YRI, green; CEU, orange; CHB+JPT, purple.

The distribution of recombination

In the Phase II HapMap we identified 32,996 recombination hotspots[3,6,36] (an increase of over 50% from Phase I) of which 68% localized to a region of ≤5 kb. The median map distance induced by a hotspot is 0.043 cM (or one crossover per 2,300 meioses) and the hottest identified, on chromosome 20, is 1.2 cM (one crossover per 80 meioses). Hotspots account for approximately 60% of recombination in the human genome and about 6% of sequence (Supplementary Fig. 6). We do not find marked differences among chromosomes in the concentration of recombination in hotspots, which implies that obligate differences in recombination among chromosomes of different size result from differences in hotspot density and intensity[6].

The increased number of well-defined hotspots allows us to understand better the influence of genomic features on the distribution of recombination. Previous work identified specific DNA motifs that influence hotspot location[6,37] as well as additional influences of local sequence context including the location of genes[6] and base composition[38]. The Phase II HapMap provides the resolution to separate these influences. Figure 5a shows the distribution of recombination, hotspot motifs and base composition around genes. Within the transcribed region of genes there is a marked decrease in the estimated recombination rate. However, 5′ of the transcription start site is a peak in recombination rate with a corresponding local increase in the density of hotspot motifs. This region also shows a marked increase in G+C content, reflecting the presence of CpG islands in promoter regions. There is also an asymmetry in recombination rate across genes, with recombination rates 3′ of transcribed regions being elevated (as are motif density and G+C content) compared to regions 5′ of genes. Studies in yeast have previously suggested an association between promoter regions and recombination hotspots[39]. Our results suggest a significant, although weak, relationship between promoters and recombination in humans. Nevertheless, the vast majority of hotspots in the human genome are not in gene promoters. The association may reflect a general association between regions of accessible chromatin and crossover activity.

Systematic differences in recombination rate by gene class. Previous work has demonstrated differences in the magnitude of linkage disequilibrium, as measured at a megabase scale, among genes associated with different functions[3,40]. Using the fine-scale genetic map estimated from the Phase II HapMap data we can quantify local increases in recombination rate associated with genes of different function using the Panther gene ontology annotation[41]. Average recombination rates vary more than sixfold among such gene classes (Fig. 5b), with defence and immunity genes showing the highest rates (1.9 cM Mb^{-1}) and chaperones showing the lowest rates (0.3 cM Mb^{-1}). Gene functions associated with cell surfaces and external functions tend to show higher recombination rates (immunity, cell adhesion, extracellular matrix, ion channels, signalling) whereas those with lower recombination rates are typically internal to cells (chaperones, ligase, isomerase, synthase). Controlling for systematic differences between gene classes in base composition and gene clustering, the differences between groups remain significant.

©2007 Nature Publishing Group

We also find that the density of hotspot-associated DNA motifs varies systematically among gene classes and that variation in motif density explains over 50% of the variance in recombination rate among gene functions (Supplementary Fig. 7).

These results pose interesting evolutionary questions. Because recombination involves DNA damage through double-strand breaks, hotspots may be selected against in some highly conserved parts of the genome. In regions exposed to recurrent selection (for example, from changes in environment or pathogen pressure) it is plausible that recombination may be selected for. However, because the fine-scale structure of recombination seems to evolve rapidly[42,43] it will be important to learn whether patterns of recombination rate heterogeneity among molecular functions are conserved between species.

Natural selection

The Phase I HapMap data have been used to identify genomic regions that show evidence for the influence of adaptive evolution[3,9], primarily through extended haplotype structure indicative of recent positive selection. Using two established approaches[9,44], we identified approximately 200 regions with evidence of recent positive selection from the Phase II HapMap (Supplementary Table 9). These regions include many established cases of selection, such as the genes *HBB* and *LCT*, the HLA region, and an inversion on chromosome 17. Many other regions have been previously identified in HapMap Phase I including *LARGE*, *SYT1* and *SULT1C2* (previously called *SULT1C1*). A detailed description of the findings from the Phase II HapMap is published elsewhere[45].

The Phase II HapMap also provides new insights into the forces acting on SNPs in coding regions. Effort was made to genotype as many known or putative non-synonymous SNPs as possible. Of the 56,789 non-synonymous SNPs identified in dbSNP release 125, attempts were made to genotype 36,777, which resulted in 17,427 that are QC+ in all three analysis panels and polymorphic. We selected only those SNPs for which ancestral allele information was available (approximately 90%). For comparison, we used patterns of variation at synonymous SNPs. As previously reported[46,47], non-synonymous SNPs show an increase in frequency of rare variants and a slight decrease of common variants compared to synonymous SNPs, compatible with widespread purifying selection against non-synonymous mutations (Fig. 6a). In contrast, we find no excess of high-frequency derived non-synonymous mutations, as might be expected if positive selection were widespread.

Natural selection also influences the extent to which allele frequencies differ between populations, not only through local selective pressures that drive alleles to different frequencies[48,49], but also through local variation in the strength of purifying selection. We compared the distribution of population differentiation (as measured by F_{ST}, the proportion of total variation in allele frequency that is due to differences between populations) at non-synonymous SNPs and synonymous SNPs matched for allele frequency (Fig. 6b). We find a systematic bias for non-synonymous SNPs to show stronger differentiation than synonymous SNPs. Among SNPs showing high levels of differentiation there is a strong tendency for the derived allele to be at higher frequency in non-YRI populations. Among SNPs with $F_{ST} > 0.5$ between CEU and YRI, in 79% and 75% of non-synonymous and synonymous variants, respectively, the derived allele is more common in CEU. Although this difference between non-synonymous and synonymous SNPs is not significant, among the eight exonic SNPs with $F_{ST} > 0.95$, all are non-synonymous. We see no such bias towards increased MAF in CEU at high-differentiation SNPs, indicating that SNP ascertainment is unlikely to explain the difference. Rather, this effect can largely be explained by more genetic drift in the non-African populations, as confirmed by simulations (data not shown). In addition, reduced selection against deleterious mutations and local adaptation within non-African populations will both act to increase the frequency of derived variants in non-African populations.

To assess the evidence for widespread local adaptation influencing non-synonymous mutations we considered the distribution of integrated extended haplotype homozygosity (iEHH) statistics[9,44] (Fig. 6c). We find no evidence for systematic differences between non-synonymous and synonymous SNPs, suggesting that local adaptation does not explain their higher differentiation. Although hitch-hiking effects will tend to obscure differences between selected

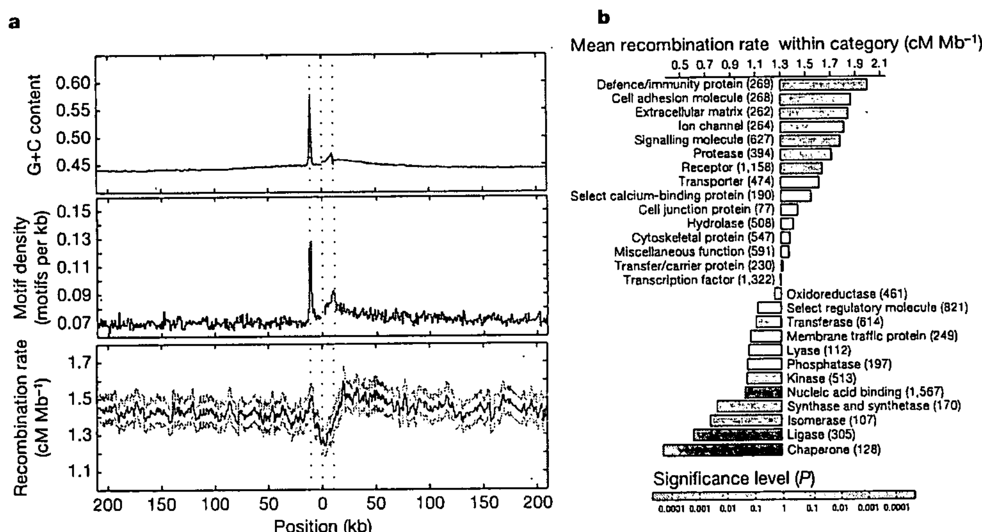

Figure 5 | Recombination rates around genes. **a**, The recombination rate, density of recombination-hotspot-associated motifs (all motifs with up to 1 bp different from the consensus CCTCCCTNNCCAC) and G+C content around genes. The blue line indicates the mean. For the recombination rate, grey lines indicate the quartiles of the distribution. Values were calculated separately 5′ from the transcription start site (the first dotted line) and 3′ from the transcription end site (third dotted line) and were joined at the median midpoint position of the transcription unit (central dotted line). Note the sharp drop in recombination rate within the transcription unit, the local increase around the transcription start site and the broad decrease away from the 3′ end of genes. These patterns only partly reflect the distribution of G+C content and the hotspot-associated motif, suggesting that additional factors influence recombination rates around genes. **b**, Recombination rates within genes of different molecular function[41]. The chart shows the increase or decrease for each category compared to the genome average. *P* values were estimated by permutation of category; numbers of genes are shown in parentheses.

©2007 Nature Publishing Group

and neutral SNPs, these results are consistent with a scenario in which the higher differentiation of non-synonymous SNPs is primarily driven by a reduction in the strength or efficacy of purifying selection in non-African populations.

Discussion and prospects

The International HapMap Project has been instrumental in making well-powered, large-scale, genome-wide association studies a reality. It is now clear that the HapMap can be a useful resource for the design and analysis of disease association studies in populations across the world[50–53]. Furthermore, the decreasing costs and increasing SNP density of standard genotyping panels mean that the focus of attention in disease association studies is shifting from candidate gene approaches towards genome-wide analyses. Alongside developments in technology, new statistical methodologies aimed at improving aspects of analysis, such as genotype calling[21,54], the identification of and correction for population stratification and relatedness[55,56], and imputation of untyped variants[21–23], are increasing the accuracy and reliability of genome-wide association studies.

Within this context, it is important to consider the future of the HapMap Project. Currently, additional samples from the populations used to develop the initial HapMap, as well as samples from seven additional populations (Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; Tuscans in Italy; Gujarati Indian in Houston, Texas, USA; Denver (Colorado) metropolitan Chinese community; people of Mexican origin in Los Angeles, California, USA; and people with African ancestry in the southwestern United States; http://ccr.coriell.org/Sections/Collections/NHGRI/?SsId=11) will be sequenced and genotyped extensively to extend the HapMap, providing information on rarer variants and helping to enable genome-wide association studies in additional populations. There are also ongoing efforts by many groups to characterize additional forms of genetic variation, such as structural variation, and molecular phenotypes in the HapMap samples. Finally, in the future, whole-genome sequencing will provide a natural convergence of technologies to type both SNP and structural variation. Nevertheless, until that point, and even after, the HapMap Project data will provide an invaluable resource for understanding the structure of human genetic variation and its link to phenotype.

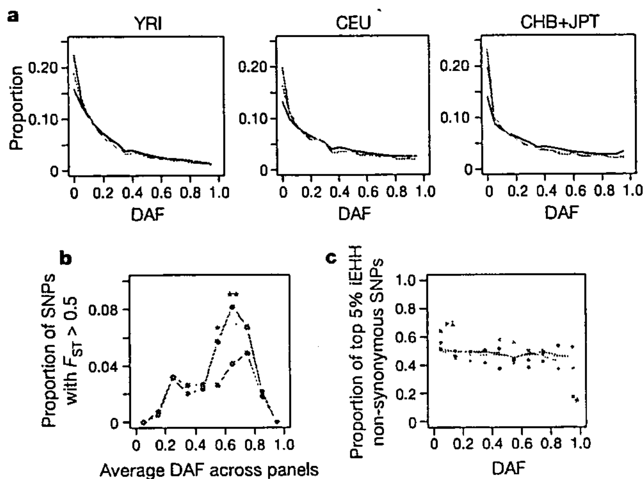

Figure 6 | Properties of non-synonymous and synonymous SNPs. a, The derived allele frequency (DAF) spectrum in each analysis panel for all SNPs (black), synonymous SNPs (green) and non-synonymous SNPs (red). Note the excess of rare variants for coding sequence SNPs but no excess of high-frequency derived variants. **b**, Enrichment of non-synonymous SNPs among genic SNPs showing high differentiation. For each of ten classes of derived allele frequency (averaged across analysis panels) the fraction of non-synonymous (red) and synonymous (green) variants in that class that show $F_{ST} > 0.5$ is shown. Note the strong enrichment of non-synonymous SNPs among SNPs of moderate to high derived-allele frequency (asterisk, $P < 0.05$; double asterisk, $P < 0.01$). **c**, Lack of enrichment of non-synonymous SNPs among those showing long-range haplotype structure. The integrated extended haplotype homozygosity (iEHH) statistic[9] was calculated for non-synonymous and synonymous SNPs in each analysis panel (YRI, green; CEU, orange; CHB+JPT, purple). For each of ten derived allele frequency classes, the proportion of non-synonymous SNPs among those showing the 5% most extreme statistics (within the allele frequency class) is shown (points). Also shown is the proportion of non-synonymous SNPs among SNPs in the coding sequence for each frequency class (dotted lines). Differences between synonymous and non-synonymous SNPs are tested for using a contingency table test.

METHODS SUMMARY

Of approximately 6.9 million SNPs in dbSNP release 122 approximately 4.7 million were selected for genotyping by Perlegen. 2.5 million SNPs were excluded because no assay could be designed and a further 350,000 were excluded for other reasons (see Methods). Perlegen performed genotyping using custom high-density oligonucleotide arrays as previously described[15]. Additional genotype submissions are described in the text. QC filters were applied as previously described[3]. Where multiple submissions met the QC criteria the submission with the lowest missing data rate was chosen for inclusion in the non-redundant filtered data set. Haplotypes were estimated from genotype data as described previously[3]. Ancestral states at SNPs were inferred by parsimony by comparison to orthologous bases in the chimpanzee (panTro2) and rhesus macaque (rheMac2) assemblies. Recombination rates and the location of recombination hotspots were estimated as described previously[3]. Additional details can be found in the Methods section and the Supplementary Information. The data described in this paper are in release 21 of the International HapMap Project.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 April; accepted 18 September 2007.

1. The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nature Rev. Genet.* 5, 467–475 (2004).
2. The International HapMap Consortium. The International HapMap Project. *Nature* 426, 789–796 (2003).
3. The International HapMap Consortium. A haplotype map of the human genome. *Nature* 437, 1299–1320 (2005).
4. Bowcock, A. M. Genomics: guilt by association. *Nature* 447, 645–646 (2007).
5. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nature Genet.* 39, 813–815 (2007).
6. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310, 321–324 (2005).
7. McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nature Genet.* 38, 86–92 (2006).
8. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genet.* 38, 75–81 (2006).
9. 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).
10. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* 444, 444–454 (2006).
11. de Bakker, P. I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genet.* 38, 1166–1172 (2006).
12. Pastinen, T. *et al.* Mapping common regulatory variants to human haplotypes. *Hum. Mol. Genet.* 14, 3963–3971 (2005).
13. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* 1, e78 (2005).
14. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437, 1365–1369 (2005).
15. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072–1079 (2005).
16. de Bakker, P. I. *et al.* Efficiency and power in genetic association studies. *Nature Genet.* 37, 1217–1223 (2005).
17. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genet.* 38, 663–667 (2006).
18. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nature Genet.* 38, 659–662 (2006).
19. Burdick, J. T., Chen, W. M., Abecasis, G. R. & Cheung, V. G. *In silico* method for inferring genotypes in pedigrees. *Nature Genet.* 38, 1002–1004 (2006).
20. Servin, B. R. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* 3, e114 (2007).

©2007 Nature Publishing Group

21. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–668 (2007).
22. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316, 1341–1345 (2007).
23. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genet.* 39, 906–913 (2007).
24. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* 56, 18–31 (2003).
25. Paabo, S. The mosaic that is our genome. *Nature* 421, 409–412 (2003).
26. McVean, G., Spencer, C. C. & Chaix, R. Perspectives on human genetic variation from the HapMap Project. *PLoS Genet.* 1, e54 (2005).
27. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575 (2007).
28. Broman, K. W. & Weber, J. L. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am. J. Hum. Genet.* 65, 1493–1500 (1999).
29. Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* 15, 789–795 (2006).
30. Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 236, 1567–1570 (1987).
31. Leutenegger, A. L. *et al.* Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome. *Am. J. Hum. Genet.* 79, 62–66 (2006).
32. Te Meerman, G. J., Van der Meulen, M. A. & Sandkuijl, L. A. Perspectives of identity by descent (IBD) mapping in founder populations. *Clin. Exp. Allergy* 25 (Suppl 2), 97–102 (1995).
33. Houwen, R. H. *et al.* Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* 8, 380–386 (1994).
34. Durham, L. K. & Feingold, E. Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *Am. J. Hum. Genet.* 61, 830–842 (1997).
35. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genet.* 36, 151–156 (2004).
36. McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* 304, 581–584 (2004).
37. Myers, S. *et al.* The distribution and causes of meiotic recombination in the human genome. *Biochem. Soc. Trans.* 34, 526–530 (2006).
38. Spencer, C. C. *et al.* The influence of recombination on human genetic diversity. *PLoS Genet.* 2, e148 (2006).
39. Petes, T. D. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* 2, 360–369 (2001).
40. Smith, A. V., Thomas, D. J., Munro, H. M. & Abecasis, G. R. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.* 15, 1519–1534 (2005).
41. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141 (2003).
42. Winckler, W. *et al.* Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 308, 107–111 (2005).
43. Ptak, S. E. *et al.* Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genet.* 37, 429–434 (2005).
44. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837 (2002).
45. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* doi:10.1038/nature06250 (this issue).
46. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* 437, 1153–1157 (2005).
47. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* 22, 231–238 (1999).
48. Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12, 1805–1814 (2002).
49. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* 312, 1614–1620 (2006).
50. de Bakker, P. I. *et al.* Transferability of tag SNPs in genetic association studies in multiple populations. *Nature Genet.* 38, 1298–1303 (2006).
51. Conrad, D. F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genet.* 38, 1251–1260 (2006).
52. Service, S., Sabatti, C. & Freimer, N. Tag SNPs chosen from HapMap perform well in several population isolates. *Genet. Epidemiol.* 31, 189–194 (2007).
53. Lim, J. *et al.* Comparative study of the linkage disequilibrium of an ENCODE region, chromosome 7p15, in Korean, Japanese, and Han Chinese samples. *Genomics* 87, 392–398 (2006).
54. Rabbee, N. & Speed, T. P. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* 22, 7–12 (2006).
55. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575 (2007).
56. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* 38, 904–909 (2006).
57. Smith, R. A., Ho, P. J., Clegg, J. B., Kidd, J. R. & Thein, S. L. Recombination breakpoints in the human β-globin gene cluster. *Blood* 92, 4415–4421 (1998).
58. Holloway, K., Lawson, V. E. & Jeffreys, A. J. Allelic recombination and *de novo* deletions in sperm in the human β-globin gene region. *Hum. Mol. Genet.* 15, 1099–1111 (2006).
59. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370 (1984).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank many people who contributed to this project: all members of the genotyping laboratory and the sample, primer, bioinformatics, data quality and IT groups at Perlegen Sciences for technical and infrastructural support; J. Beck, C. Beiswanger, D. Coppock, A. Leach, J. Mintzer and L. Toji for transforming the Yoruba, Japanese and Han Chinese samples, distributing the DNA and cell lines, storing the samples for use in future research, and producing the community newsletters and reports; J. Greenberg and R. Anderson for providing funding and support for cell line transformation and storage in the NIGMS Human Genetic Cell Repository at the Coriell Institute; T. Dibling, T. Ishikura, S. Kanazawa, S. Mizusawa and S. Saito for help with genotyping; C. Hind and A. Moghadam for technical support in genotyping and all members of the subcloning and sequencing teams at the Wellcome Trust Sanger Institute; X. Ke for help with data analysis; Oxford E-Science Centre for provision of high-performance computing resources; H. Chen, W. Chen, L. Deng, Y. Dong, C. Fu, L. Gao, H. Geng, J. Geng, M. He, H. Li, H. Li, S. Li, X. Li, B. Liu, Z. Liu, F. Lu, F. Lu, G. Lu, C. Luo, X. Wang, Z. Wang, C. Ye and X. Yu for help with genotyping and sample collection; X. Feng, Y. Li, J. Ren and X. Zhou for help with sample collection; J. Fan, W. Gu, W. Guan, S. Hu, H. Jiang, R. Lei, Y. Lin, Z. Niu, B. Wang, L. Yang, W. Yang, Y. Wang, Z. Wang, S. Xu, W. Yan, H. Yang, W. Yuan, C. Zhang, J. Zhang, K. Zhang and G. Zhao for help with genotyping; P. Fong, C. Lai, C. Lau, T. Leung, L. Luk and W. Tong for help with genotyping; C. Pang for help with genotyping; K. Ding, B. Qiang, J. Zhang, X. Zhang and K. Zhou for help with genotyping; Q. Fu, S. Ghose, X. Lu, D. Nelson, A. Perez, S. Poole, R. Vega and H. Yonath for help with genotyping; C. Bruckner, T. Brundage, S. Chow, O. Iartchouk, M. Jain, M. Moorhead and K. Tran for help with genotyping; N. Addleman, J. Atilano, T. Chan, C. Chu, C. Ha, T. Nguyen, M. Minton and A. Phong for help with genotyping, and D. Lind for help with quality control and experimental design; R. Donaldson and S. Duan for help with genotyping, and J. Rice and N. Saccone for help with experimental design; J. Wigginton for help with implementing and testing QA/QC software; A. Clark, B. Keats, R. Myers, D. Nickerson and A. Williamson for providing advice to NIH; C. Juenger, C. Bennet, C. Bird, J. Melone, P. Nailer, M. Weiss, J. Witonsky and E. DeHaut-Combs for help with project management; M. Gray for organizing phone calls and meetings; D. Leja for help with figures; the Yoruba people of Ibadan, Nigeria, the people of Tokyo, Japan, and the community at Beijing Normal University, who participated in public consultations and community engagements; the people in these communities who donated their blood samples; and the people in the Utah CEPH community who allowed the samples they donated earlier to be used for the Project. This work was supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology, the Wellcome Trust, Nuffield Trust, Wolfson Foundation, UK EPSRC, Genome Canada, Génome Québec, the Chinese Academy of Sciences, the Ministry of Science and Technology of the People's Republic of China, the National Natural Science Foundation of China, the Hong Kong Innovation and Technology Commission, the University Grants Committee of Hong Kong, the SNP Consortium, the US National Institutes of Health (FIC, NCI, NCRR, NEI, NHGRI, NIA, NIAAA, NIAID, NIAMS, NIBIB, NIDA, NIDCD, NIDCR, NIDDK, NIEHS, NIGMS, NIMH, NINDS, NLM, OD), the W.M. Keck Foundation, and the Delores Dore Eccles Foundation. All SNPs genotyped within the HapMap Project are available from dbSNP (http://www.ncbi.nlm.nih.gov/SNP); all genotype information is available from dbSNP and the HapMap website (http://www.hapmap.org).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.M. (mcvean@stats.ox.ac.uk) or M.D. (mjdaly@chgr.mgh.harvard.edu).

The International HapMap Consortium (Participants are arranged by institution and then alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.)

Genotyping centres: Perlegen Sciences Kelly A. Frazer (Principal Investigator)[1], Dennis G. Ballinger[2], David R. Cox[2], David A. Hinds[2], Laura L. Stuve[2]; **Baylor College of Medicine and ParAllele BioScience** Richard A. Gibbs (Principal Investigator)[3], John W. Belmont[3], Andrew Boudreau[4], Paul Hardenbol[5], Suzanne M. Leal[3], Shiran Pasternak[6], David A. Wheeler[3], Thomas D. Willis[4], Fuli Yu[7]; **Beijing Genomics Institute** Huanming Yang (Principal Investigator)[8], Changqing Zeng (Principal Investigator)[8], Yang Gao[8], Haoran Hu[8], Weitao Hu[8], Chaohua Li[8], Wei Lin[8], Siqi Liu[8], Hao Pan[8], Xiaoli Tang[8], Jian Wang[8], Wei Wang[8], Jun Yu[8], Bo Zhang[8], Qingrun Zhang[8], Hongbin Zhao[8], Hui Zhao[8], Jun Zhou[8]; **Broad Institute of Harvard and Massachusetts Institute of Technology**

©2007 Nature Publishing Group

Stacey B. Gabriel (Project Leader)[7], Rachel Barry[7], Brendan Blumenstiel[7], Amy Camargo[7], Matthew Defelice[7], Maura Faggart[7], Mary Goyette[7], Supriya Gupta[7], Jamie Moore[7], Huy Nguyen[7], Robert C. Onofrio[7], Melissa Parkin[7], Jessica Roy[7], Erich Stahl[7], Ellen Winchester[7], Liuda Ziaugra[7], David Altshuler (Principal Investigator)[7,9]; **Chinese National Human Genome Center at Beijing** Yan Shen (Principal Investigator)[10], Zhijian Yao[10]; **Chinese National Human Genome Center at Shanghai** Wei Huang (Principal Investigator)[11], Xun Chu[11], Yungang He[11], Li Jin[12], Yangfan Liu[11], Yayun Shen[11], Weiwei Sun[11], Haifeng Wang[11], Yi Wang[11], Ying Wang[11], Xiaoyan Xiong[11], Liang Xu[11]; **Chinese University of Hong Kong** Mary M. Y. Waye (Principal Investigator)[13], Stephen K. W. Tsui[13]; **Hong Kong University of Science and Technology** Hong Xue (Principal Investigator)[14], J. Tze-Fei Wong[14]; **Illumina** Luana M. Galver (Project Leader)[15], Jian-Bing Fan[15], Kevin Gunderson[15], Sarah S. Murray[1], Arnold R. Oliphant[16], Mark S. Chee (Principal Investigator)[17]; **McGill University and Génome Québec Innovation Centre** Alexandre Montpetit (Project Leader)[18], Fanny Chagnon[18], Vincent Ferretti[18], Martin Leboeuf[18], Jean-François Olivier[4], Michael S. Phillips[18], Stéphanie Roumy[15], Clémentine Sallée[19], Andrei Verner[18], Thomas J. Hudson (Principal Investigator)[20]; **University of California at San Francisco and Washington University** Pui-Yan Kwok (Principal Investigator)[21], Dongmei Cai[21], Daniel C. Koboldt[22], Raymond D. Miller[22], Ludmila Pawlikowska[21], Patricia Taillon-Miller[22], Ming Xiao[21]; **University of Hong Kong** Lap-Chee Tsui (Principal Investigator)[23], William Mak[23], You Qiang Song[23], Paul K. H. Tam[23]; **University of Tokyo and RIKEN** Yusuke Nakamura (Principal Investigator)[24,25], Takahisa Kawaguchi[25], Takuya Kitamoto[25], Takashi Morizono[25], Atsushi Nagashima[25], Yozo Ohnishi[25], Akihiro Sekine[25], Toshihiro Tanaka[25], Tatsuhiko Tsunoda[25]; **Wellcome Trust Sanger Institute** Panos Deloukas (Project Leader)[26], Christine P. Bird[26], Marcos Delgado[26], Emmanouil T. Dermitzakis[26], Rhian Gwilliam[26], Sarah Hunt[26], Jonathan Morrison[27], Don Powell[26], Barbara E. Stranger[26], Pamela Whittaker[26], David R. Bentley (Principal Investigator)[28]

Analysis groups: Broad Institute Mark J. Daly (Project Leader)[7,9], Paul I. W. de Bakker[7,9], Jeff Barrett[7,9], Yves R. Chretien[7], Julian Maller[7,9], Steve McCarroll[7,9], Nick Patterson[7], Itsik Pe'er[29], Alkes Price[7], Shaun Purcell[9], Daniel J. Richter[7], Pardis Sabeti[7], Richa Saxena[7,9], Stephen F. Schaffner[7], Pak C. Sham[23], Patrick Varilly[7], David Altshuler (Principal Investigator)[7,9]; **Cold Spring Harbor Laboratory** Lincoln D. Stein (Principal Investigator)[6], Lalitha Krishnan[6], Albert Vernon Smith[6], Marcela K. Tello-Ruiz[6], Gudmundur A. Thorisson[30]; **Johns Hopkins University School of Medicine** Aravinda Chakravarti (Principal Investigator)[31], Peter E. Chen[31], David J. Cutler[31], Carl S. Kashuk[31], Shin Lin[31]; **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[32], Weihua Guan[32], Yun Li[32], Heather M. Munro[33], Zhaohui Steve Qin[32], Daryl J. Thomas[34]; **University of Oxford** Gilean McVean (Project Leader)[35], Adam Auton[35], Leonardo Bottolo[35], Niall Cardin[35], Susana Eyheramendy[35], Colin Freeman[35], Jonathan Marchini[35], Simon Myers[35], Chris Spencer[7], Matthew Stephens[36], Peter Donnelly (Principal Investigator)[35]; **University of Oxford, Wellcome Trust Centre for Human Genetics** Lon R. Cardon (Principal Investigator)[37], Geraldine Clarke[38], David M. Evans[38], Andrew P. Morris[38], Bruce S. Weir[39]; **RIKEN** Tatsuhiko Tsunoda (Principal Investigator)[25], Todd A. Johnson[25]; **US National Institutes of Health** James C. Mullikin[40]; **US National Institutes of Health National Center for Biotechnology Information** Stephen T. Sherry[41], Michael Feolo[41], Andrew Skol[42]

Community engagement/public consultation and sample collection groups: Beijing Normal University and Beijing Genomics Institute Houcan Zhang[43], Changqing Zeng[8], Hui Zhao[8]; **Health Sciences University of Hokkaido, Eubios Ethics Institute, and Shinshu University** Ichiro Matsuda (Principal Investigator)[44], Yoshimitsu Fukushima[45], Darryl R. Macer[46], Eiko Suda[47]; **Howard University and University of Ibadan** Charles N. Rotimi (Principal Investigator)[48], Clement A. Adebamowo[49], Ike Ajayi[49], Toyin Aniagwu[49], Patricia A. Marshall[50], Chibuzor Nkwodimmah[49], Charmaine D. M. Royal[48]; **University of Utah** Mark F. Leppert (Principal Investigator)[51], Missy Dixon[51], Andy Peiffer[51]

Ethical, legal and social issues: Chinese Academy of Social Sciences Renzong Qiu[52]; **Genetic Interest Group** Alastair Kent[53]; **Kyoto University** Kazuto Kato[54]; **Nagasaki University** Norio Niikawa[55]; **University of Ibadan School of Medicine** Isaac F. Adewole[49]; **University of Montréal** Bartha M. Knoppers[19]; **University of Oklahoma** Morris W. Foster[56]; **Vanderbilt University** Ellen Wright Clayton[57]; **Wellcome Trust** Jessica Watkin[58]

SNP discovery: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)[3], John W. Belmont[3], Donna Muzny[3], Lynne Nazareth[3], Erica Sodergren[3], George M. Weinstock[3], David A. Wheeler[3], Imtaz Yakub[3]; **Broad Institute of Harvard and Massachusetts Institute of Technology** Stacey B. Gabriel (Project Leader)[7], Robert C. Onofrio[7], Daniel J. Richter[7], Liuda Ziaugra[7], Bruce W. Birren[7], Mark J. Daly[7,9], David Altshuler (Principal Investigator)[7,9]; **Washington University** Richard K. Wilson (Principal Investigator)[59], Lucinda L. Fulton[59]; **Wellcome Trust Sanger Institute** Jane Rogers (Principal Investigator)[26], John Burton[26], Nigel P. Carter[26], Christopher M. Clee[26], Mark Griffiths[26], Matthew C. Jones[26], Kirsten McLay[26], Robert W. Plumb[26], Mark T. Ross[26], Sarah K. Sims[26], David L. Willey[26]

Scientific management: Chinese Academy of Sciences Zhu Chen[60], Hua Han[60], Le Kang[60]; **Genome Canada** Martin Godbout[61], John C. Wallenburg[62]; **Génome Québec** Paul L'Archevêque[63], Guy Bellemare[63]; **Japanese Ministry of Education, Culture, Sports, Science and Technology** Koji Saeki[64]; **Ministry of Science and Technology of the People's Republic of China** Hongguang Wang[65], Daochang An[65], Hongbo Fu[65], Qing Li[65], Zhen Wang[65]; **The Human Genetic Resource Administration of China** Renwu Wang[66]; **The SNP Consortium** Arthur L. Holden[15]; **US National Institutes of Health** Lisa D. Brooks[67], Jean E. McEwen[67], Mark S. Guyer[67], Vivian Ota Wang[67,68], Jane L. Peterson[67], Michael Shi[69], Jack Spiegel[70], Lawrence M. Sung[71], Lynn F. Zacharia[67], Francis S. Collins[72]; **Wellcome Trust** Karen Kennedy[61], Ruth Jamieson[58], John Stewart[58]

[1]The Scripps Research Institute, 10550 North Torrey Pines Road MEM275, La Jolla, California 92037, USA. [2]Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, California 94043, USA. [3]Baylor College of Medicine, Human Genome Sequencing Center, Department of Molecular and Human Genetics, 1 Baylor Plaza, Houston, Texas 77030, USA. [4]Affymetrix, Inc., 3420 Central Expressway, Santa Clara, California 95051, USA. [5]Pacific Biosciences, 1505 Adams Drive, Menlo Park, California 94025, USA. [6]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [7]The Broad Institute of Harvard and Massachusetts Institute of Technology, 1 Kendall Square, Cambridge, Massachusetts 02139, USA. [8]Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 100300, China. [9]Massachusetts General Hospital and Harvard Medical School, Simches Research Center, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [10]Chinese National Human Genome Center at Beijing, 3-707 N. Yongchang Road, Beijing Economic-Technological Development Area, Beijing 100176, China. [11]Chinese National Human Genome Center at Shanghai, 250 Bi Bo Road, Shanghai 201203, China. [12]Fudan University and CAS-MPG Partner Institute for Computational Biology, School of Life Sciences, SIBS, CAS, Shanghai 201203, China. [13]The Chinese University of Hong Kong, Department of Biochemistry, The Croucher Laboratory for Human Genetics, 6/F Mong Man Wai Building, Shatin, Hong Kong. [14]Hong Kong University of Science and Technology, Department of Biochemistry and Applied Genomics Center, Clear Water Bay, Knowloon, Hong Kong. [15]Illumina, 9885 Towne Centre Drive, San Diego, California 92121, USA. [16]Complete Genomics, Inc., 658 North Pastoria Avenue, Sunnyvale, California 94085, USA. [17]Prognosys Biosciences, Inc., 4215 Sorrento Valley Boulevard, Suite 105, San Diego, California 92121, USA. [18]McGill University and Génome Québec Innovation Centre, 740 Dr. Penfield Avenue, Montréal, Québec H3A 1A4, Canada. [19]University of Montréal, The Public Law Research Centre (CRDP), PO Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada. [20]Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 500, Toronto, Ontario M5G 1L7, Canada. [21]University of California, San Francisco, Cardiovascular Research Institute, 513 Parnassus Avenue, Box 0793, San Francisco, California 94143, USA. [22]Washington University School of Medicine, Department of Genetics, 660 South Euclid Avenue, Box 8232, St Louis, Missouri 63110, USA. [23]University of Hong Kong, Genome Research Centre, 6/F, Laboratory Block, 21 Sassoon Road, Pokfulam, Hong Kong. [24]University of Tokyo, Institute of Medical Science, 4-6-1 Sirokanedai, Minato-ku, Tokyo 108-8639, Japan. [25]RIKEN SNP Research Center, 1-7-22 Suehiro-cho, Tsurumi-ku Yokohama, Kanagawa 230-0045, Japan. [26]Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [27]University of Cambridge, Department of Oncology, Cambridge CB1 8RN, UK. [28]Solexa Ltd, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK. [29]Columbia University, 500 West 120th Street, New York, New York 10027, USA. [30]University of Leicester, Department of Genetics, Leicester LE1 7RH, UK. [31]Johns Hopkins University School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Broadway Research Building, Suite 579, 733 North Broadway, Baltimore, Maryland 21205, USA. [32]University of Michigan, Center for Statistical Genetics, Department of Biostatistics, 1420 Washington Heights, Ann Arbor, Michigan 48109, USA. [33]International Epidemiology Institute, 1455 Research Boulevard, Suite 550, Rockville, Maryland 20850, USA. [34]Center for Biomolecular Science and Engineering, Engineering 2, Suite 501, Mail Stop CBSE/ITI, UC Santa Cruz, Santa Cruz, California 95064, USA. [35]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK. [36]University of Chicago, Department of Statistics, 5734 South University Avenue, Eckhart Hall, Room 126, Chicago, Illinois 60637, USA. [37]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA. [38]University of Oxford/Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [39]University of Washington Department of Biostatistics, Box 357232, Seattle, Washington 98195, USA. [40]US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA. [41]US National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, Maryland 20894, USA. [42]University of Chicago, Department of Medicine, Section of Genetic Medicine, 5801 South Ellis, Chicago, Illinois 60637, USA. [43]Beijing Normal University, 19 Xinjiekouwai Street, Beijing 100875, China. [44]Health Sciences University of Hokkaido, Ishikari Tobetsu Machi 1757, Hokkaido 061-0293, Japan. [45]Shinshu University School of Medicine, Department of Medical Genetics, Matsumoto 390-8621, Japan. [46]United Nations Educational, Scientific and Cultural Organization (UNESCO Bangkok), 920 Sukhumwit Road, Prakanong, Bangkok 10110, Thailand. [47]University of Tsukuba, Eubios Ethics Institute, PO Box 125, Tsukuba Science City 305-8691, Japan. [48]Howard University, National Human Genome Center, 2216 6th Street, NW, Washington DC 20059, USA. [49]University of Ibadan College of Medicine, Ibadan, Oyo State, Nigeria. [50]Case Western Reserve University School of Medicine, Department of Bioethics, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA. [51]University of Utah, Eccles Institute of Human Genetics, Department of Human Genetics, 15 North 2030 East, Salt Lake City, Utah 84112, USA. [52]Chinese Academy of Social Sciences, Institute of Philosophy/Center for Applied Ethics, 2121, Building 9, Caoqiao Xinyuan 3 Qu, Beijing 100067, China. [53]Genetic Interest Group, 4D Leroy House, 436 Essex Road, London N130P, UK. [54]Kyoto University, Institute for Research in Humanities and Graduate School of Biostudies, Ushinomiya-cho, Sakyo-ku, Kyoto 606-8501, Japan. [55]Nagasaki University Graduate

©2007 Nature Publishing Group

School of Biomedical Sciences, Department of Human Genetics, Sakamoto 1-12-4, Nagasaki 852-8523, Japan. [56]University of Oklahoma, Department of Anthropology, 455 West Lindsey Street, Norman, Oklahoma 73019, USA. [57]Vanderbilt University, Center for Genetics and Health Policy, 507 Light Hall, Nashville, Tennessee 37232, USA. [58]Wellcome Trust, 215 Euston Road, London NW1 2BE, UK. [59]Washington University School of Medicine, Genome Sequencing Center, Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. [60]Chinese Academy of Sciences, 52 Sanlihe Road, Beijing 100864, China. [61]Genome Canada, 150 Metcalfe Street, Suite 2100, Ottawa, Ontario K2P 1P1, Canada. [62]McGill University, Office of Technology Transfer, 3550 University Street, Montréal, Québec H3A 2A7, Canada. [63]Génome Québec, 630, boulevard René-Lévesque Ouest, Montréal, Québec H3B 1S6, Canada. [64]Ministry of Education, Culture, Sports, Science, and Technology, 3-2-2 Kasumigaseki, Chiyodaku, Tokyo 100-8959, Japan. [65]Ministry of Science and Technology of the People's Republic of China, 15 B. Fuxing Road, Beijing 100862, China. [66]The Human Genetic Resource Administration of China, b7, Zaojunmiao, Haidian District, Beijing 100081, China. [67]US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA. [68]US National Institutes of Health, Office of Behavioral and Social Science Research, 31 Center Drive, Bethesda, Maryland 20892, USA. [69]Novartis Pharmaceuticals Corporation, Biomarker Development, One Health Plaza, East Hanover, New Jersey 07936, USA. [70]US National Institutes of Health, Office of Technology Transfer, 6011 Executive Boulevard, Rockville, Maryland 20852, USA. [71]University of Maryland School of Law, 500 West Baltimore Street, Baltimore, Maryland 21201, USA. [72]US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA.

©2007 Nature Publishing Group

Vol 449|18 October 2007|doi:10.1038/nature06250

LETTERS

Genome-wide detection and characterization of positive selection in human populations

Pardis C. Sabeti[1*], Patrick Varilly[1*], Ben Fry[1], Jason Lohmueller[1], Elizabeth Hostetter[1], Chris Cotsapas[1,2], Xiaohui Xie[1], Elizabeth H. Byrne[1], Steven A. McCarroll[1,2], Rachelle Gaudet[3], Stephen F. Schaffner[1], Eric S. Lander[1,4,5,6] & The International HapMap Consortium†

With the advent of dense maps of human genetic variation, it is now possible to detect positive natural selection across the human genome. Here we report an analysis of over 3 million polymorphisms from the International HapMap Project Phase 2 (HapMap2)[1]. We used 'long-range haplotype' methods, which were developed to identify alleles segregating in a population that have undergone recent selection[2], and we also developed new methods that are based on cross-population comparisons to discover alleles that have swept to near-fixation within a population. The analysis reveals more than 300 strong candidate regions. Focusing on the strongest 22 regions, we develop a heuristic for scrutinizing these regions to identify candidate targets of selection. In a complementary analysis, we identify 26 non-synonymous, coding, single nucleotide polymorphisms showing regional evidence of positive selection. Examination of these candidates highlights three cases in which two genes in a common biological process have apparently undergone positive selection in the same population: *LARGE* and *DMD*, both related to infection by the Lassa virus[3], in West Africa; *SLC24A5* and *SLC45A2*, both involved in skin pigmentation[4,5], in Europe; and *EDAR* and *EDA2R*, both involved in development of hair follicles[6], in Asia.

An increasing amount of information about genetic variation, together with new analytical methods, is making it possible to explore the recent evolutionary history of the human population. The first phase of the International Haplotype Map, including ~1 million single nucleotide polymorphisms (SNPs)[7], allowed preliminary examination of natural selection in humans. Now, with the publication of the Phase 2 map (HapMap2)[1] in a companion paper, over 3 million SNPs have been genotyped in 420 chromosomes from three continents (120 European (CEU), 120 African (YRI) and 180 Asian from Japan and China (JPT + CHB)).

In our analysis of HapMap2, we first implemented two widely used tests that detect recent positive selection by finding common alleles carried on unusually long haplotypes[2]. The two, the Long-Range Haplotype (LRH)[8] and the integrated Haplotype Score (iHS)[9] tests, rely on the principle that, under positive selection, an allele may rise to high frequency rapidly enough that long-range association with nearby polymorphisms—the long-range haplotype[8]—will not have time to be eliminated by recombination. These tests control for local variation in recombination rates by comparing long haplotypes to other alleles at the same locus. As a result, they lose power as selected alleles approach fixation (100% frequency), because there are then few alternative alleles in the population (Supplementary Fig. 2 and Supplementary Tables 1–2).

We next developed, evaluated and applied a new test, Cross Population Extended Haplotype Homozogysity (XP-EHH), to detect selective sweeps in which the selected allele has approached or achieved fixation in one population but remains polymorphic in the human population as a whole (Methods, and Supplementary Fig. 2 and Supplementary Tables 3–6). Related methods have recently also been described[10–12].

Our analysis of recent positive selection, using the three methods, reveals more than 300 candidate regions[1] (Supplementary Fig. 3 and Supplementary Table 7), 22 of which are above a threshold such that no similar events were found in 10 Gb of simulated neutrally evolving sequence (Methods). We focused on these 22 strongest signals (Table 1), which include two well-established cases, *SLC24A5* and *LCT*[2,5,13], and 20 other regions with signals of similar strength.

The challenge is to sift through genetic variation in the candidate regions to identify the variants that were the targets of selection. Our candidate regions are large (mean length, 815 kb; maximum length, 3.5 Mb) and often contain multiple genes (median, 4; maximum, 15). A typical region harbours ~400–4,000 common SNPs (minor allele frequency >5%), of which roughly three-quarters are represented in current SNP databases and half were genotyped as part of HapMap2 (Supplementary Table 8).

We developed three criteria to help highlight potential targets of selection (Supplementary Fig. 1): (1) selected alleles detectable by our tests are likely to be derived (newly arisen), because long-haplotype tests have little power to detect selection on standing (pre-existing) variation[14]; we therefore focused on derived alleles, as identified by comparison to primate outgroups; (2) selected alleles are likely to be highly differentiated between populations, because recent selection is probably a local environmental adaptation[2]; we thus looked for alleles common in only the population(s) under selection; (3) selected alleles must have biological effects. On the basis of current knowledge, we therefore focused on non-synonymous coding SNPs and SNPs in evolutionarily conserved sequences. These criteria are intended as heuristics, not absolute requirements. Some targets of selection may not satisfy them, and some will not be in current SNP databases. Nonetheless, with ~50% of common SNPs in these populations genotyped in HapMap2, a search for causal variants is timely.

We applied the criteria to the regions containing *SLC24A5* and *LCT*, each of which already has a strong candidate gene, mutation and trait. At *SLC24A5*, the 600 kb region contains 914 genotyped

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA. [2]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [4]Department of Biology, MIT, Cambridge, Massachusetts 02139, USA. [5]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. [6]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.
†Lists of participants and affiliations appear at the end of the paper.

©2007 Nature Publishing Group

Table 1 | The twenty-two strongest candidates for natural selection

Region	Chr:position (MB, HG17)	Selected population	Long Haplotype Test	Size (Mb)	Total SNPs with Long Haplotype Signal	Subset of SNPs that fulfil criteria 1	Subset of SNPs that fulfil criteria 1 and 2	Subset of SNPs that fulfil criteria 1, 2 and 3	Genes at or near SNPs that fulfil all three criteria
1	chr1:166	CHB + JPT	LRH, iHS	0.4	92	39	30	2	*BLZF1, SLC19A2*
2	chr2:72.6	CHB + JPT	XP-EHH	0.8	732	250	0	0	
3	chr2:108.7	CHB + JPT	LRH, iHS, XP-EHH	1.0	972	265	7	1	*EDAR*
4	chr2:136.1	CEU	LRH, iHS, XP-EHH	2.4	1,213	282	24	3	*RAB3GAP1, R3HDM1, LCT*
5	chr2:177.9	CEU,CHB + JPT	LRH, iHS, XP-EHH	1.2	1,388	399	79	9	*PDE11A*
6	chr4:33.9	CEU,YRI, CHB + JPT	LRH, iHS	1.7	413	161	33	0	
7	chr4:42	CHB + JPT	LRH, iHS, XP-EHH	0.3	249	94	65	6	*SLC30A9*
8	chr4:159	CHB + JPT	LRH, iHS, XP-EHH	0.3	233	67	34	1	
9	chr10:3	CEU	LRH, iHS, XP-EHH	0.3	179	63	16	1	
10	chr10:22.7	CEU, CHB + JPT	XP-EHH	0.3	254	93	0	0	
11	chr10:55.7	CHB+JPT	LRH, iHS, XP-EHH	0.4	735	221	5	2	*PCDH15*
12	chr12:78.3	YRI	LRH, iHS	0.8	151	91	25	0	
13	chr15:46.4	CEU	XP-EHH	0.6	867	233	5	1	*SLC24A5*
14	chr15:61.8	CHB + JPT	XP-EHH	0.2	252	73	40	6	*HERC1*
15	chr16:64.3	CHB + JPT	XP-EHH	0.4	484	137	2	0	
16	chr16:74.3	CHB + JPT, YRI	LRH, iHS	0.6	55	35	28	3	*CHST5, ADAT1, KARS*
17	chr17:53.3	CHB + JPT	XP-EHH	0.2	143	41	0	0	
18	chr17:56.4	CEU	XP-EHH	0.4	290	98	26	3	*BCAS3*
19	chr19:43.5	YRI	LRH, iHS, XP-EHH	0.3	83	30	0	0	
20	chr22:32.5	YRI	LRH	0.4	318	188	35	3	*LARGE*
21	chr23:35.1	YRI	LRH, iHS	0.6	50	35	25	0	
22	chr23:63.5	YRI	LRH, iHS	3.5	13	3	1	0	
		Total SNPs		16.74	9,166	2,898	480	41	

Twenty-two regions were identified at a high threshold for significance (Methods), based on the LRH, iHS and/or XP-EHH test. Within these regions, we examined SNPs with the best evidence of being the target of selection on the basis of having a long haplotype signal, and by fulfilling three criteria: (1) being a high-frequency derived allele; (2) being differentiated between populations and common only in the selected population; and (3) being identified as functional by current annotation. Several candidate polymorphisms arise from the analysis including well-known *LCT* and *SLC24A5* (ref. 2), as well as intriguing new candidates.

SNPs. Applying filters progressively (Table 1 and Fig. 1a–d), we found that 867 SNPs are associated with the long-haplotype signal, of which 233 are high-frequency derived alleles, of which 12 are highly differentiated between populations, and of which only 5 are common in Europe and rare in Asia and Africa. Among these five SNPs, there is only one implicated as functional by current knowledge; it has the strongest signal of positive selection and encodes the A111T polymorphism associated with pigment differences in

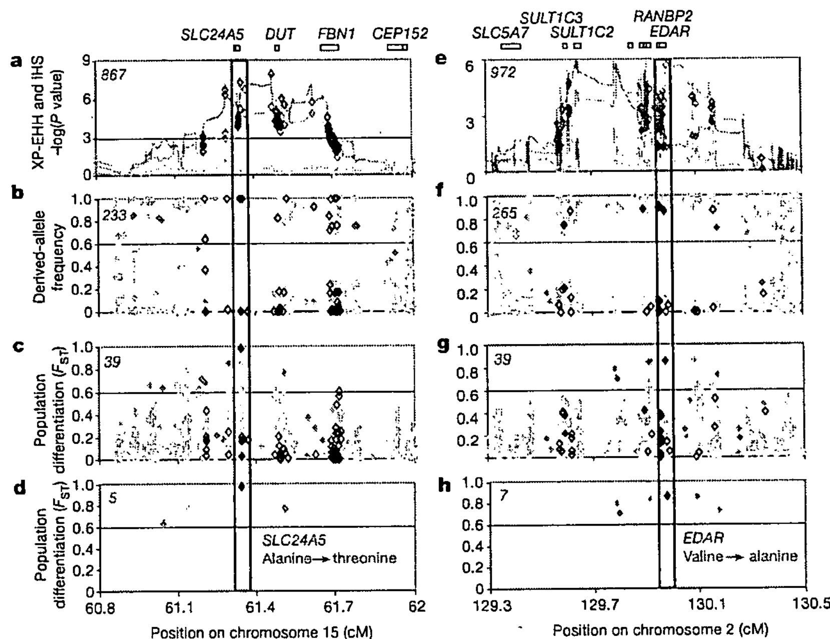

Figure 1 | Localizing *SLC24A5* and *EDAR* signals of selection. **a–d**, *SLC24A5*. **a**, Strong evidence for positive selection in CEU samples at a chromosome 15 locus: XP-EHH between CEU and JPT + CHB (blue), CEU and YRI (red), and YRI and JPT + CHB (grey). SNPs are classified as having low probability (bordered diamonds) and high probability (filled diamonds) potential for function. SNPs were filtered to identify likely targets of selection on the basis of the frequency of derived alleles (**b**), differences between populations (**c**) and differences between populations for high-frequency derived alleles (less than 20% in non-selected populations) (**d**). The number of SNPs that passed each filter is given in the top left corner in red. The threonine to alanine candidate polymorphism in *SLC24A5* is the clear outlier. **e–h**, *EDAR*. **e**, Similar evidence for positive selection in JPT + CHB at a chromosome 2 locus: XP-EHH between CEU and JPT + CHB (blue), between YRI and JPT + CHB (red), and between CEU and YRI (grey); iHS in JPT + CHB (green). A valine to alanine polymorphism in *EDAR* passes all filters: the frequency of derived alleles (**f**), differences between populations (**g**) and differences between populations for high-frequency derived alleles (less than 20% in non-selected populations) (**h**). Three other functional changes, a D→E change in *SULT1C2* and two SNPs associated with *RANBP2* expression (Methods), have also become common in the selected population.

©2007 Nature Publishing Group

humans and thought to be the target of positive selection[5]. Our criteria thus uniquely identify the expected allele.

At the *LCT* locus, we found similar degrees of filtration. Within the 2.4 Mb selective sweep, 24 polymorphisms fulfil the first two criteria (Table 1, and Supplementary Fig. 4), with the polymorphism thought to confer adult persistence of lactase among them. However, this SNP was only identified as functional after extensive study of the *LCT* gene[15]. Thus *LCT* shows both the utility and the limits of the heuristics.

Given the encouraging results for *SLC24A5* and *LCT*, we performed a similar analysis on all 22 candidate regions (Table 1). Filtering the 9,166 SNPs associated with the long-haplotype signal, we found that 480 satisfied the first two criteria. We identified 41 out of the 480 SNPs (0.2% of all SNPs genotyped in the regions) as possibly functional on the basis of a newly compiled database of polymorphisms in known coding elements, evolutionarily conserved elements and regulatory elements (Methods; B.F., unpublished), together containing ~ 5.5% of all known SNPs.

Eight of the forty-one SNPs encode non-synonymous changes (Table 1 and Supplementary Table 9). Apart from the well-known case of *SLC24A5*, they are found in *EDAR*, *PCDH15*, *ADAT1*, *KARS*, *HERC1*, *SLC30A9* and *BLFZ1*. The remaining 33 potentially functional SNPs lie within conserved transcription factor motifs, introns, UTRs and other non-coding regions.

To identify additional candidates, we reversed the process by taking non-synonymous coding SNPs with highly differentiated high-frequency derived alleles; these SNPs comprise a tiny fraction of all SNPs and have a higher a priori probability of being targets of selection. Of the 15,816 non-synonymous SNPs in HapMap2, 281 (Supplementary Table 10) have both a high derived-allele frequency (frequency >50%) and clear differentiation between populations (F_{ST} is in the top 0.5 percentile). We examined these 281 SNPs to identify those embedded within long-range haplotypes[16], and identified 26 putative cases of positive selection. These include the eight non-synonymous SNPs identified in the genome-wide analysis above.

Interestingly, analysis of the top regions and the non-synonymous SNPs together revealed three cases of two genes in the same pathway both having strong evidence of selection in a single population.

In the European sample, there is strong evidence for two genes already shown to be associated with skin pigment differences among humans. The first is *SLC24A5*, described above. We further examined the global distribution (Fig. 2) and the predicted effect on protein activity of the *SLC24A5* A111T polymorphism (Supplementary Fig. 5, 6). The second, *SLC45A2*, has an important role in pigmentation in zebrafish, mouse and horse[4]. An L374F substitution in *SLC45A2* is at 100% frequency in the European sample, but absent in the Asian and African samples. A recent association study has shown that the Phe-encoding allele is correlated with fair skin and non-black hair in Europeans[4]. Together, the data support *SLC45A2* as a target of positive selection in Europe[10,17].

In the African sample (Yoruba in Ibadan, Nigeria), there is evidence of selection for two genes with well-documented biological links to the Lassa fever virus. The strongest signal in the genome, on the basis of the LRH test, resides within a 400 kb region that lies entirely within the gene *LARGE*. The LARGE protein is a glycosylase that post-translationally modifies α-dystroglycan, the cellular receptor for Lassa fever virus (as well as other arenaviruses), and the modification has been shown to be critical for virus binding[3]. The virus name is derived from Lassa, Nigeria, where the disease is endemic, with 21% of the population showing signs of exposure[18]. We also noted that the *DMD* locus is on our larger candidate list of regions, with the signal of selection again in the Yoruba sample. *DMD* encodes a cytosolic adaptor protein that binds to α-dystroglycan and is critical for its function. We hypothesize that Lassa fever created selective pressure at *LARGE* and *DMD*[12]. This hypothesis can be tested by correlating the geographical distribution of the selected haplotype with endemicity of the Lassa virus, studying infection of genotyped cells *in vitro*, and searching for an association between the selected haplotype and clinical outcomes in infected patients.

In the Asian samples, we found evidence of selection for non-synonymous polymorphisms in two genes in the ectodysplasin (EDA) pathway, which is involved in development of hair, teeth and exocrine glands[6]. The genes are *EDAR* and *EDA2R*, which encode the key receptors for the ligands EDA A1 and EDA A2, respectively. Notably, the EDA signalling pathway has been shown to be under positive selection for loss of scales in multiple distinct populations of freshwater stickleback fish[19]. A mutation encoding a V370A substitution in *EDAR* is near fixation in Asia and absent in Europe and Africa (Fig. 1e–h). An R57K substitution in *EDA2R* has derived-allele frequencies of 100% in Asia, 70% in Europe and 0% in Africa.

The *EDAR* polymorphism is notable because it is highly differentiated between the Asian and other continental populations (the 3rd most differentiated among 15,816 non-synonymous SNPs), and also within Asian populations (in the top 1% of SNPs differentiated between the Japanese and Chinese HapMap samples). Genotyping of the *EDAR* polymorphism in the CEPH (Centre d'Etudie du Polymorphisme Humain) global diversity panel[20] shows that it is at high but varying frequency throughout Asia and the Americas (for example, 100% in Pima Indians and in parts of China, and 73% in Japan) (Fig. 2, and Supplementary Fig. 7). Studying populations like the Japanese, in which the allele is still segregating, may provide clues to its biological significance.

EDAR has a central role in generation of the primary hair follicle pattern, and mutations in *EDAR* cause hypohidrotic ectodermal

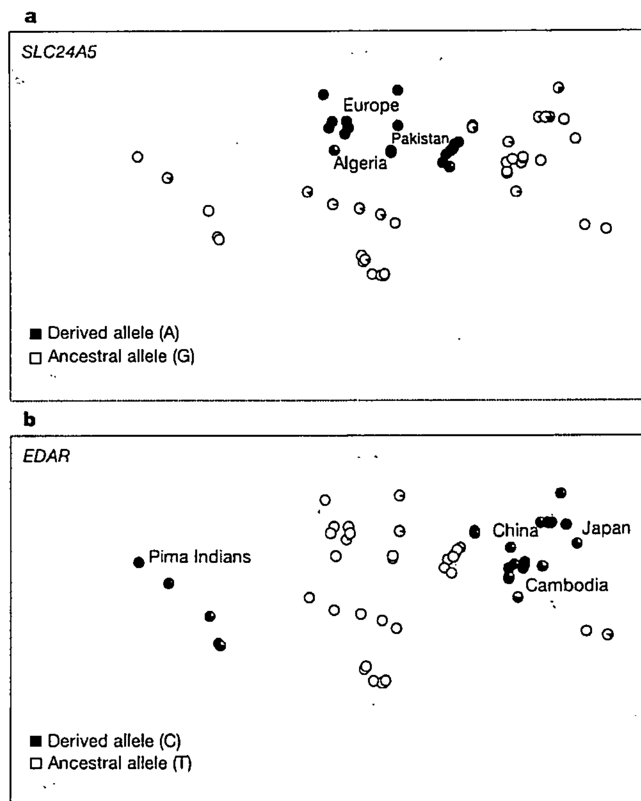

Figure 2 | Global distribution of *SLC24A5* A111T and *EDAR* V370A. Worldwide allele-frequency distributions for candidate polymorphisms with the strongest evidence for selection[20]. **a**, *SLC24A5* A111T is common in Europe, Northern Africa and Pakistan, but rare or absent elsewhere. **b**, *EDAR* V370A is common in Asia and the Americas, but absent in Europe and Africa.

©2007 Nature Publishing Group

dysplasia (HED) in humans and mice, characterized by defects in the development of hair, teeth and exocrine glands[6]. The V370A polymorphism, proposed to be the target of selection, lies within *EDAR*'s highly conserved death domain (Supplementary Fig. 8), the location of the majority of *EDAR* polymorphisms causing HED[21]. Our structural modelling predicts that the polymorphism lies within the binding site of the domain (Fig. 3).

Our analysis only scratches the surface of the recent selective history of the human genome. The results indicate that individual candidates may coalesce into pathways that reveal traits under selection, analogous to the alleles of multiple genes (for example, *HBB*, *G6PD* and *DARC*) that arose and spread in Africa and other tropical populations as a result of the partial protection they confer against malaria[2,12]. Such endeavours will be enhanced by continuing development of analytical methods to localize signals in candidate regions, generation of expanded data sets, advances in comparative genomics to define coding and regulatory regions, and biological follow-up of promising candidates. True understanding of the role of adaptive evolution will require collaboration across multiple disciplines, including molecular and structural biology, medical and population genetics, and history and anthropology.

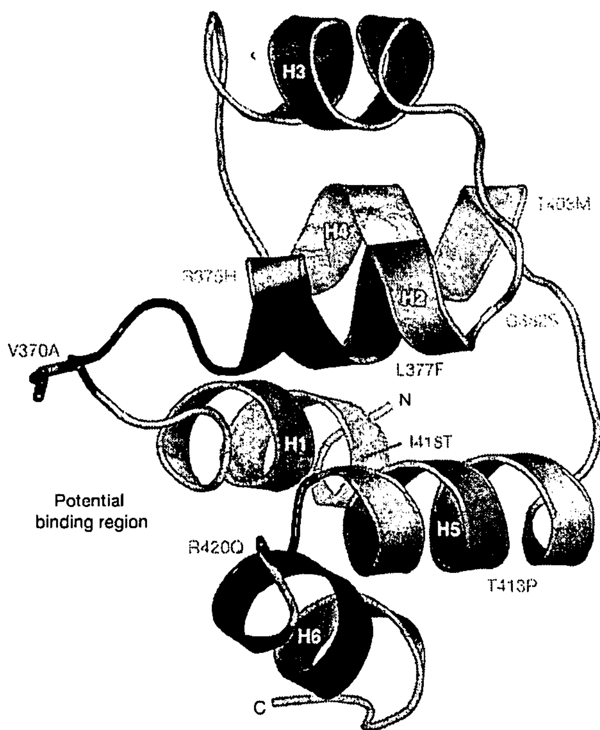

Figure 3 | Structural model of the EDAR death domain. Ribbon representation of a homology model of the EDAR death domain (DD), based on the alignment of the EDAR DD amino acid sequence (EDAR residues 356–431), with multiple known DD structures. The helices are labelled H1 to H6. Residues in blue (the H1–H2 and H5–H6 loops, residues 370–376 and 419–425, respectively) correspond to the homologous residues in Tube that interact with Pelle in the Tube-DD–Pelle-DD structure[24]. These EDAR-DD residues therefore form a potential region of interaction with a DD-containing EDAR-interacting protein, such as EDARADD. The V370A polymorphic residue (red) is located prominently within this potential binding region in the H1–H2 loop. Seven of the thirteen known mis-sense mutations in EDAR that lead to hypohidrotic ectodermal dysplasia (HED) in humans are located in the EDAR-DD: the only four mutations in EDAR that lead to the dominant transmission of HED (green) and three recessive mutations (yellow)[21]. Four of these mutations, R375H, L377F, R420Q and I418T are located in the vicinity of the predicted interaction interface.

METHODS SUMMARY

Genotyping data. Phase 2 of the International Haplotype Map (HapMap2) (www.hapmap.org) contains 3.1 million SNPs genotyped in 420 chromosomes in 3 continental populations (120 European (CEU), 120 African (YRI) and 180 Asian (JPT+CHB))[1]. We further genotyped our top HapMap2 functional candidates in the HGDR-CEPH Human Genome Diversity Cell Line Panel[20].

LRH, iHS and XP-EHH tests. The Long-Range Haplotype (LRH), integrated Haplotype Score (iHS) and Cross Population EHH (XP-EHH) tests detect alleles that have risen to high frequency rapidly enough that long-range association with nearby polymorphisms—the long-range haplotype—has not been eroded by recombination; haplotype length is measured by the EHH[8,9]. The first two tests detect partial selective sweeps, whereas XP-EHH detects selected alleles that have risen to near fixation in one but not all populations. To evaluate the tests, we simulated genomic data for each HapMap population in a range of demographic scenarios—under neutral evolution and twenty scenarios of positive selection—developing the program Sweep (www.broad.mit.edu/mpg/sweep) for analysis. For our top candidates by the three tests, we tested for haplotype-specific recombination rates and copy-number polymorphisms, possible confounders.

Localization. We calculated F_{ST} and derived-allele frequency for all SNPs within the top candidate regions. We developed a database for those regions to annotate all potentially functional DNA changes (B.F., unpublished), including non-synonymous variants, variants disrupting predicted functional motifs, variants within regions of conservation in mammals and variants previously associated with human phenotypic differences, as well as synonymous, intronic and untranslated region variants.

Structural model. We generated a homology model of the EDAR death domain (DD) from available DD structures using Modeller 9v1 (ref. 22). The distribution of conserved residues, built using ConSurf[23] with an EDAR sequence alignment from 22 species, shows a bias to the protein core in helices H1, H2 and H5, supporting our model.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 August; accepted 13 September 2007.

1. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* doi:10.1038/nature06258 (this issue).
2. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* 312, 1614–1620 (2006).
3. Kunz, S. *et al.* Posttranslational modification of α-dystroglycan, the cellular receptor for arenaviruses, by the glycosyltransferase LARGE is critical for virus binding. *J. Virol.* 79, 14282–14296 (2005).
4. Graf, J., Hodgson, R. & van Daal, A. Single nucleotide polymorphisms in the *MATP* gene are associated with normal human pigmentation variation. *Hum. Mutat.* 25, 278–284 (2005).
5. Lamason, R. L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 310, 1782–1786 (2005).
6. Botchkarev, V. A. & Fessing, M. Y. Edar signaling in the control of hair follicle development. *J. Investig. Dermatol. Symp. Proc.* 10, 247–251 (2005).
7. The International Haplotype Map Consortium. A haplotype map of the human genome. *Nature* 437, 1299–1320 (2005).
8. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837 (2002).
9. 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).
10. Kimura, R., Fujimoto, A., Tokunaga, K. & Ohashi, J. A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS ONE* 2, e286 (2007).
11. Tang, K., Thornton, K. R. & Stoneking, M. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* 5, e171 (2007).
12. Williamson, S. H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS Genet.* 3, e90 (2007).
13. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74, 1111–1120 (2004).
14. Teshima, K. M., Coop, G. & Przeworski, M. How reliable are empirical genomic scans for selective sweeps? 16, 702–712 *Genome Res.* (2006).
15. Kuokkanen, M. *et al.* Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. *Gut* 52, 647–652 (2003).
16. Miller, R. G. *Simultaneous statistical inference* XVI 299 (Springer, New York, 1981).
17. Soejima, M., Tachida, H., Ishida, T., Sano, A. & Koda, Y. Evidence for recent positive selection at the human *AIM1* locus in a European population. *Mol. Biol. Evol.* 23, 179–188 (2006).
18. Richmond, J. K. & Baglole, D. J. Lassa fever: epidemiology, clinical features, and social consequences. *Br. Med. J.* 327, 1271–1275 (2003).
19. Colosimo, P. F. *et al.* Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* 307, 1928–1933 (2005).

©2007 Nature Publishing Group