# Wzorce różnic DNA w całym genomie trzech populacji ludzkich

David A. Hinds, Laura L. Stuve, Geoffrey B. Nilsen,
Eran Halperin Eleazar Eskin, Dennis G. Gallinger,
Kelly A. Frazer, David R. Cox

5 czerwca 2006

**Pojedyncze różnice w sekwencjach DNA są genetyczną podstawą ludzkiej różnorodności. Scharakteryzowaliśmy pełnogenomowe wzorce typowych różnic w ludzkim DNA przez genotypowanie 1586383 jedno-nukleotydowych polimorfizmów (single-nucleotide polymorphism - SNP) u 71 amerykanów o pochodzeniu europejskim, afrykańskim i azjatyckim. Nasze wyniki wskazują że wśród zebranych SNP najbardziej typowe wariacje genomu wynikają z braku równowagi powiązań (ang. linkage disequilibrium), tj. korelacji między typowymi allelami SNP. Obserwujemy silną korelację pomiędzy rozszerzonymi regionami nie równoważności powiązań a funkcjonalnymi elementami genomu. Dzięki opracowanym danym stworzyliśmy narzędzie, które pozwoli odpowiedzieć na powstające pytania, związane ze znaczeniem przypadkowości różnic w ludzkim DNA na tle złożonych cech ludzkich, jak i badać naturę różnorodności genetycznej w populacjach, jak i między nimi.**

Polimorfizmy jedno-nukleotydowe (SNP) są najczęstszym rodzajem różnorodności DNA ludzkiego genomu. Szacuje się, że u całej populacji ludzkiej (1) istnieje ok 7 milionów typowych SNP o małych częstotliwościach alleli (minor allele frequency - MAF) większych od 5%. Najbardziej typowe SNP znajdują się w większości głównych populacji, aczkolwiek częstotliwości którychkolwiek alleli mogą się różnić w zależności od populacji (2). Pozostaje 4 miliony innych SNP o MAF pomiędzy 1 a 5%. Ponadto, jest jeszcze niezliczona ilość bardzo rzadkich różnic pojedynczych zasad, które występują zazwyczaj tylko u jednego osobnika.

Związek między wariacjami DNA a ludzkimi różnicami fenotypicznymi (jak wzrost, kolor oczu, odporność) jest trudny do znalezienia. Chociaż istnieją badania wskazujące że zarówno typowe SNP jak i rzadkie różnice współtworzą znaną różnorodność złożonych cech ludzkich (3, 4), to stosunek typowych do rzadkich różnic pozostaje nie zbadany. Struktura różnorodności genotypów między populacjami i ich powiązanie z różnorodnością fenotypiczną jest nie jasne. Podobnie, wpływ na cechy ludzkie wariacji DNA, które przyczyniają się do modyfikacji struktury białek poprzez zamianę aminokwasów, kontra wariantów modyfikujących wzorce ekspresji genów bez zmiany na strukturę białek, pozostaje nie znane. W pewnych przypadkach, zagadnienia te były badane dla małych przedziałów, lecz analizy pełnego genomu nie były możliwe.

Badania powiązań całego genomu, mające na celu zidentyfikowanie alleli wpływających na złożone cechy na potrzeby medycyny są obecnie prowadzone na podzbiorach najczęstszych SNP, a więc opierają się na założeniu że chore allele powinny być skorelowane z nimi. Chociaż badania wykazały że zmiany występujące w bliskim sąsiedztwie (fizycznie) są często mocno powiązane, to struktura tych powiązań, lub nie równoważności powiązań (linkage disequilibrium - LD) jest złożona i różni się w zależności od obszaru genomu, oraz typu populacji (5, 6). Do badania zależności, niezbędny jest wybór podzbioru najbardziej informatywnych SNP, by zapewnić silną podstawę do oszacowania znaczenia przypadkowości różnic DNA dla złożonych cech ludzkich. Mimo że duża część wszystkich typowych ludzkich SNP jest dostępna w publicznych bazach danych, to brak informacji na temat częstotliwości ich alleli oraz struktury zależności wśród, oraz pomiędzy populacjami ludzkimi spowodowały, że wybór optymalnego podzbioru był trudny.

Przebadaliśmy częstotliwości alleli SNP oraz wzorce LD 1586383 SNP równomiernie rozłożonych w ludzkim genomie u niezależnych osobników, amerykanów o pochodzeniu europejskim, afrykańskim i azjatyckim. Naszym głównym celem było stworzenie źródła do dalszych badań nad strukturą ludzkiej różnorodności genetycznej oraz jej związku z różnicami fenotypicznymi.

**Gęsta mapa SNP**. By scharakteryzować panel markerów informatywnych w skali badań całego genomu, wybraliśmy łącznie 2384494 SNP typowe dla jednostek o różnym pochodzeniu (7). Zidentyfikowaliśmy większość(69%) SNP poprzez array-based resequencing 24 próbek DNA ludzi o różnym pochodzeniu (5). Te SNP zostały dodane do SNP wybranych z publicznych baz danych aby osiągnąć bardziej równomierny rozkład w genomie człowieka. Dalsze informacje dot. doboru SNP zostały umieszczone w materiałach dostępnych online (7). Opracowaliśmy 49 macierzy oligonukleotydowych wysokiej rozdzielczości do genotypowania wybranych SNP (8 i 9) oraz około 300000 par primerów

do reakcji polimerazy (PCR), pokrywających wybrane SNP ze średnią ośmiu SNP na pojedynczy region amplifikacji PCR. Amplikony (amplicons) miały średnią długość 9kb i pokrywały 92% dostępnego genomu człowieka. Średnio 6250 amplikonów pochodzących od jednej jednostki zostało pobranych a następnie zchybrydyzowanych do pojedynczej macierzy oligonukleotydowej, tworząc genotypy dla ok 48000 SNP.

Określiliśmy genotyp 71 niezależnych osobników z trzech populacji: 24 amerykanów europejskiego pochodzenia, 23 afrykańskiego i 24 chińskiego (Han Chinese) z okolic Los Angeles. Te 71 osób nie było powiązane z poprzednimi, wykorzystanymi do wyznaczenia SNP. Próbki DNA zostały wybrane z Coriell Cell Repositories'Human Variation Collection, korzystając także z szacunków populacji próbek. Uwzględniliśmy politykę Coriell dotyczącą badań DNA na podstawie próbek z nie anonimowych populacji.

Każde SNP zostało ocenione wg. zestawu metryk, zbieżne z jakością genotypu na naszej platformie i dane dla mało odpowiadających SNP zostały na tym etapie odrzucone. Metryki te zawierały ocenę odwołań, liczbę klastrów genotypu, uwzględniały istnienie prawie dokładnie odpowiadających sekwencji SNP w innych miejscach genomu, obecność innych znanych SNP w najbliższych sekwencjach, oraz zgodność z równoważnikiem Hardy Weinberga. Testy z równoważnikiem Hardy Weinberga pozwalają bardzo skutecznie identyfikować pewne typy artefaktów (10), aczkolwiek, ponieważ testy te wykorzystano do kontroli jakości, zebrane dane dot. genomów nie nadają się do badania biologicznie interesujących prawdziwych odchyleń od równoważnika Hardy Weinberga. Dalsze informacje na temat kontroli jakości genotypów znajdują się w materiałach online (7).

Podzbiór 1586383 SNP został z powodzeniem przebadany wg. wymienionych kryteriów jakości, tak że każde dwie allele odnaleziono co najmniej raz wśród 71 osób. Łącznie, ponad 112 milionów pojedynczych genomów zostało zweryfikowane pod kątem tych SNP. Dla 64% nie było brakujących genotypów, z czego 94% miało mniej niż 5% brakujących danych. Szczegóły analiz SNP zostały umieszczone w bazie danych SNP National Center for Biotechnology Information (NCBI) (sbSNP, build 123, numer dostępowy ss23145044 do ss24731426). Genotypy 156757 SNP dla dziewięciu europejczyków, które były częścią tego projektu zostały wcześniej określone przez International HapMap Project, korzystając z różnych narzędzi genotypowania (11). Nasze dane 1,6 miliona genotypów są w 99,54% są zgodne z tymi z projektu HapMap. Rozkład niezgodnych genotypów jest mało przypadkowy, jedynie 0,3% SNP odpowiada za 50% niezgodności, a szacujemy że 90% SNP z pełnych danych nie ma w ogóle błędnych genotypów. Szczególnie analizy haplotypu zyskają na takim rozkładzie błędów, ponieważ do właściwego wnioskowania wymagają one właśnie spójności genotypu w ramach dużych grup sąsiadujących markerów.

Rozkład 1.6 miliona wysokiej jakości genotypów SNP (tabela S1) jest zbliżony do omawianej wcześniej mapy liczącej 1.42 miliona SNP. Ponad 95% genomu jest w przedziałach inter-SNP krótszych niż 50kb, a ponad dwie trzecie pokrywają przedziały inter-SNP do 10kb (tabela S2). Średnia odległość między sąsiednimi SNP to 1871 par zasad. Mimo że powtarzających się elementów nie jest wiele, w naszym zbiorze znalazło się 269611 SNP w obszarach powtarzających się, gdzie sekwencje skrajne mogły być unikalnie zmapowane. Dla 735094 SNP (46%) w rejonów kodujących geny, przyjmujemy że należą do 10kb transkrybowanych przedziałów należących do 22904 genów kodujących białka, wg. 3 wydania komentarzy NCBI (build 34). Co najmniej 1 SNP występuje w 78% wszystkich transkrypcji. Biorąc pod uwagę także 10kb regiony upstream i downstream, 93% genów kodujących białka zawiera co najmniej jedno SNP. Łącznie 20165 SNP (1.3%) występuje w sekwencjach kodujących aminokwasy a 9370 z nich nie jest równoważne, a więc prowadzi do zmian aminokwasów (tabela S3). Mimo że nasz dobór SNP nie jest całkowicie przypadkowy, to podzbiór ten ma całkiem równomierny rozkład w całym genomie ludzkim, z uwzględnieniem wspomnianych genów kodujących białka, jak i biorąc pod uwagę fizyczne odległości.

**Typowe SNP w trzech populacjach**. Tabela 1 przedstawia efekty próby wyznaczenia zbioru typowych SNP, które mają znaczenie nie zależnie od pochodzenia osobnika. Większość z 1586383 SNP o genotypach wysokiej jakości ma charakter polimorficzny w każdej z trzech próbnych populacji tego badania. Dziewięćdziesiąt cztery procent (tj. 1483594 SNP) ma po dwie allele w próbie o pochodzeniu afrykańskim, 81% (1286277 SNP) w próbie o pochodzeniu europejskim oraz 74% (1168029 SNP) w azjatyckim. W każdej populacji, większość segregujących SNP ma MAF ponad 10%, w przedziale między 68% dla próby afrykańskiej a 57% dla azjatyckiej. Tylko 263029 z 1586383 SNP (17%) ma MAF poniżej 10% w badanych próbach. Rozkłady MAF prób trzech populacji są bardzo podobne, szczególnie między europejczykami i Azjatami, oraz z uwzględnieniem trochę wyższej częstotliwości występowania rzadszych alleli z populacją afrykańską (S1). Zgodnie z poprzednimi badaniami (2, 13), największą różnorodność genetyczną zaobserwowaliśmy u osobników o pochodzeniu afrykańskim. Nasze założenia dotyczące SNP nie pozwalają stawiać bardziej złożonych wniosków z uwagi na charakter rozkładu częstotliwości alleli SNP w różnych populacjach.

Mimo że małe rozmiary prób wykluczają jakiekolwiek twierdzenia dotyczące całkowitego braku pewnych alleli w populacji, zaobserwowaliśmy że 291012 SNP (18%) występowało tylko w jednej z populacji ("prywatne SNP"). Większość z tych prywatnych SNP (75%) występowało w próbie o pochodzeniu afrykańskim, aczkolwiek w każdej z prób jakieś wystąpiły (tabela 1). Mimo że prywatne SNP mają na ogół niższe MAF od pozostałych, to w naszym zbiorze znacząca grupa ma typowe: 106404 (czyli 37%) ma $MAF > 0.10$.

By skwantyfikować różnice genetyczne wśród oraz pomiędzy populacjami, dla każdego SNP wyliczyliśmy $F_{ST}$ w

każdej parze populacji, jak i wartość łączoną dla wszystkich trzech populacji (14) $F_{ST}$ wyznacza różnicę genetyczną pomiędzy populacjami jako ułamek całkowitej różnicy genetycznej (fraction of the total genetic variance). Ponieważ afro-amerykanie stanowią dosyć zmieszaną populacją, z silnymi wpływami (admixture) genów europejskich (15), szacunki $F_{ST}$ w porównaniach z tą grupą będą bardziej zmienne, jednak ogólnie powinny odpowiadać nieznacznie zaniżonemu wynikowi dla próby rdzennych Afrykanów. Rozkłady $F_{ST}$ dla par Afrykanów i europejczyków, oraz europejczyków i azjatów są bardzo zbliżone (S2). Wnioski te są zgodne z dotychczasowymi badaniami (16, 1), wykazującymi że większość typowych wariacji DNA jest współdzielona przez wszystkie populacje, a jedynie częstotliwości alleli się różnią.

Markery z dużą wariancją między populacjami są szczególnie przydatne przy mapowaniu wpływów genetycznych między populacjami, powodujących różnice fenotypiczne (18). Mapowanie wpływów dostarcza informacji o korelacji stosunkowo długich obszarów w populacjach które ostatnio ulegały mieszaniu się, dzięki czemu można zidentyfikować różnice i genetyczne przyczyny tych różnic, w stosunku do rodowitych populacji. Jest to uwarunkowane dryfowaniem genów (genetic drift), lub lokalną naturalną selekcją. Technika ta wymaga doboru ograniczonej liczby markerów identyfikujących informacje rasowe ("ancestry-informative markers"). Wskazanie przez nas dużej liczby takich markerów likwiduje jedną z głównych barier przed stosowaniem tej obiecującej ale jeszcze mocno nie przetestowanej metody.

**Ewidencja doboru naturalnego między populacjami.** Wykazuje się, że dobór naturalny zniekształca obserwowany rozkład $F_{ST}$ między genomami ludzkimi i duże wartości $F_{ST}$ mogą wskazywać kandydatów szczególnie ulegających lokalnemu doborowi (13, 19). Jeżeli to jest prawda, wówczas duże wartości $F_{ST}$ powinny być obserwowane w pobliżu obszarów genetycznych o dużym znaczeniu funkcjonalnym. Przyjrzeliśmy się rozkładowi $F_{ST}$ zarówno dla genowych i nie genowych SNP, kodujących i nie kodujących, oraz równoważnych i nie równoważnych. Przeprowadziliśmy analizę w podzbiorach SNP pogrupowanych wg. MAF, co pozwoliło skutecznie zaobserwować części różnic między populacjami o SNP o takiej samej łącznej wariacji genetycznej (S3). Typowe SNP w obszarach genowych mają nieco wyższe wartości $F_{ST}$ w porównaniu do obszarów nie genowych o tym samym MAF (analiza wariancji (ANOVA), $P = 1.8 \times 10^{-4}$), oraz typowe kodujące regiony mają nieco wyższe $F_{ST}$ od nie kodujących SNP (ANOVA, $1.1 \times 10^{-4}$). Nie zauważyliśmy istotnej różnicy $F_{ST}$ pomiędzy równoważnymi i nie równoważnymi SNP, ale nasz poziom istotności był stosunkowo ograniczony przez małą liczność próby oraz spodziewane korelacje między SNP w tych samych transkrypcjach. Wyniki te są zgodne z wpływem doboru lokalnego na zmianę rozkładu $F_{ST}$ w okolicach obszarów funkcyjnych. Aczkolwiek, ponieważ rozkłady $F_{ST}$ w obszarach genowych i nie genowych SNP, są bardzo zbliżone, duże wartości $F_{ST}$ jako jedyne, wydają się być słabym dowodem doboru.

Podobną analizę wykonaliśmy także by sprawdzić, czy jest powiązanie między prywatnymi SNP a elementami funkcyjnymi. Grupując wg. MAF, nie zauważyliśmy różnic frekwencji prywatnego SNP, między kodującymi i nie kodującymi SNP (S4). To oznacza, że SNP odpowiedzialne za występowanie doboru lokalnego w analizie $F_{ST}$ na ogół są prywatne i segregują różne populacje. Chociaż znane są przypadki powiązań alleli SNP typowych dla danej populacji z różnicami fenotypicznymi (20-22), nasze wyniki każą raczej przypuszczać, że większość funkcjonalnych różnic genetycznych nie jest typowa dla poszczególnych populacji.

**Struktura korelacji typowych SNP.** Warianty DNA bezpośrednio sąsiadujące ze sobą w chromosomie na ogół są powiązane. Powiązania te znane są jako nie równoważność powiązań (linkage disequilibrium - LD). LD jest skutkiem złożenia procesów, jak mutacje, dobór naturalny, dryf genetyczny. Początkowo może obejmować na bardzo długie obszary genomu, jednak stale ulega rozpadowi wskutek rekombinacji. Obserwowana struktura LD w dowolnym przedziale genomu zależy więc od historii wzajemnych oddziaływań demograficznych, zdarzeń losowych oraz stałych zależności funkcyjnych. Istnieje kilka miar, wyznaczających LD pomiędzy parami SNP; my użyliśmy $r^2$, tj. kwadrat korelacji współczynnika dla tablicy 2 na 2 frekwencji haplotypów (23).

Do zidentyfikowania koszy typowych SNP o bardzo silnym LD, takich że każdy kosz ma jeden "tag SNP" z $r^2$ o wartości co najmniej 0.8, oraz większej od $r^2$ każdego innego SNP z tego kosza (24) użyliśmy modyfikacji poprzednio opisanego algorytmu. Ten algorytm "zachłanny" kolejno wyznacza największe możliwe podzbiory o takich własnościach z listy dostępnych SNP, następnie usuwając wykorzystane SNP z listy w następnej iteracji. Analizując zredukowany zbiór tagów SNP, koszt genotypowania w badaniu powiązań może być znacząco zmniejszony, nie tracąc przy tym możliwości odkrycia nietypowych powiązań w pełnym zbiorze SNP. W przeciwieństwie do bloków haplotypów, będących grupami ciągów SNP, poszczególne SNP tworzące kosz mogą być kojarzone z SNP należącymi do innych koszy.

Tabela 2 podsumowuje charakterystykę koszy na przestrzeni genomu, wyłączając chromosom Y, dla każdej z trzech prób populacji. Skupiliśmy się na analizie typowych SNP z $MAF > 10\%$, ponieważ szacunki LD dla wariantów o mniejszym MAF nie mają znaczenia w przypadku małego rozmiaru próby (23). Mimo że większość koszy LD zawierała tylko jedno SNP, były one niewielką częścią wszystkich SNP, z których większość jest silnie powiązana z wieloma innymi SNP. W danych europejczyków, 52.2% z 293.677 koszy zawierało 1 SNP, chociaż stanowiły one jedynie 15.5% z 991185 typowych SNP. Znacząca część wszystkich SNP nadawała się na tagi dzięki dużym wartościom $r^2$ z każdym innym elementem kosza, co oznacza że kosze były silnie ze sobą powiązane. Dla próby afrykańskiej, powstało zdecydowanie mniej koszy o dużej liczbie SNP (wykres 1). Należy jednak pamiętać że badana struktura LD, opiera się jedynie na

25% wszystkich typowych SNP w genomie. Chociaż rozmiary dłuższych przedziałów LD powinny być stosunkowo silne do naszego niepełnego stwierdzenia, proporcje wszystkich typowych SNP w wysokim LD do pozostałych SNP mogą być znacznie zaniżone na podstawie naszych danych.

**LD i elementy funkcyjne.** Zaobserwowaliśmy silne powiązanie między rozszerzonymi przedziałami LD a funkcyjnymi cechami genomu (tabela 3). W dużych koszach większość stanowiły genowe SNP (test trendu, $P \approx 0$), a w regionach genowych, kodujące SNP były znacząco częstsze od nie kodujących (test trendu, $P = 1.9 \times 10^{-26}$). W dużych koszach było także więcej nie równoważnych SNP (test trendu, $P = 5.3 \times 10^{-4}$). Wynik ten jest zgodny z hipotezą o powiązaniach między doborem a pewnymi obszarami rozszerzonego LD (25, 26) oraz przypuszczeniami że niektóre regiony genowe rozszerzonego LD mogą odgrywać znaczącą rolę determinującą podstawy różnic fenotypicznych między ludźmi.

Zidentyfikowaliśmy pięć koszy liczących ponad 200 SNP każdy, oraz 17 przedziałów genowych zawierających kosze, rozciągniętych na ponad 1000kb w jednej lub kilku populacjach (tabele S4 i S5). Niektóre z tych dużych koszy pokrywały podobnie duże geny. Kosz o największej liczbie SNP był w chromosomie 17 mapy europejskiej i miał nietypowy wzór wariacji, z dwoma wcześniej notowanymi haplotypami rozciągniętymi na 518 SNP i pokrywającymi długość 800kb (27). Rzadszy haplotyp występował u 25% członków próby europejskiej i u 9% próby afrykańskiej, a był nie obecny u azjatyckiej. Ten kosz zawiera gen dla białka tau powiązanego z mikrotubulami, którego mutacje powodują wiele chorób neurologicznych; gen kodujący *protease*, podobną do *presenilins*, którego mutacje powodują chorobę Alzheimera; oraz gen receptora hormonu wydzielającego kortykotropinę (*corticotropin*), który pośredniczy w imunologicznym, endokrynologicznych, autonomicznych i zachowawczych reakcjach na stres (przypisy 27-29).

**Globalne wzorce LD.** Rozkład SNP i LD na przestrzeni całego ludzkiego genomu przedstawia rysunek 2, w dokładniejszej postaci dostępny online. Górna ścieżka przedstawia stosunkową jednorodność pokrycia analizowanymi SNP, niezależnie od przedziałów centrometrycznych i telemetrycznych heterochromatyn. Środkowa ścieżka przedstawia część typowych SNP, w wysokim LD z co najmniej jednym innym SNP. W większości regionów, zauważyliśmy dużą nadmiarowość u próby europejskiej i azjatyckiej, oraz nieznacznie mniejszą u próby afrykańskiej. Dolna ścieżka przedstawia część typowych SNP, które trafiły do stosunkowo dużych koszy LD w poszczególnych populacjach. Ścieżka ta ma wyraźną strukturę w skali milionów par zasad. Ogólnie, wszystkie trzy populacje są zbliżone, jednak występują także przedziały o znaczących rozbieżnościach.

Przeprowadzona analiza całego genomu odsłania fakt, że duże struktury LD są skorelowane z dużymi różnicami rekombinacyjnymi, co jest zgodne z dotychczasowymi odkryciami dotyczącymi pojedynczych chromosomów (30). W szczególności, regiony o bardzo silnym LD są w większości skupione w regionach o niskiej rekombinacji (wykres S5). To powiązanie dużych struktur LD ze stopniem rekombinacji oraz odkrycie że regiony rozszerzonego LD zawierają ślady doboru, dostarcza silnych dowodów na rzecz hipotezy że struktury LD ludzkiego genomu spełniają istotne funkcje i nie są prostym wynikiem losowych dryftów genetycznych czy procesów demograficznych.

**Podzbiory SNP z większością typowych wariacji.** Jako że tylko część typowych SNP w ludzkiej populacji została do dziś scharakteryzowana, badanie powiązań oparte na znanych podzbiorach SNP polega na założeniu że nie odkryte, powiązane z chorobami warianty będą w skorelowane z allelami analizowanych SNP. Statystycznie szansa na znalezienie nie analizowanej, chorobowej alleli nie bezpośrednio skorelowanej z allelami analizowanych SNP jest uzależniona od $r^2$. Dokładniej, szansa na znalezienie nie bezpośredniego powiązania wśród $N$ osobników is równoważna szansie wykrycia jej bezpośrednio wśród $Nr^2$ osobników (31). Faktyczna szansa na znalezienie pewnego przypadkowego wariantu zależy od rodzaju akcji tego wariantu, penetracji, oraz szczegółów samych badań. Tak więc $r^2$ może być odpowiedzią tylko na prostsze pytanie, jaki jest koszt wynikający z rozmiaru próby, w poprawnie przeprowadzonych badaniach, dotyczących nie bezpośrednio analizowanych przypadkowych wariantów.

By określić szanse znalezienia nie poznanego, wariantu chorobotwórczego przy dotychczasowym zbiorze SNP, wzięliśmy także pod uwagę fakt, że genomy jednostek o pochodzeniu europejskim oraz afrykańskim z naszych prób były sekwencjonowane także w ramach SeattleSNPs Program for Genomic Applications (PGA) (32). Dla tych osobników, wyniki dostarczają kompletnych danych nt. różnic genetycznych w sekwencjonowanych regionach, co pozwala szacować odsetek wszystkich wariacji zawartych w naszym zbiorze SNP. Co więcej, dane te pozwalają określić pokrycie sekwencjonowanymi SNP w miejscach, które bezpośrednio nie podlegały analizie.

Przebadaliśmy dane dla 16601 wariantów sekwencji rozpoznanych w 152 genach, z czego 2465 należało do naszego zbioru SNP. Ta zbieżność naszych danych genotypów oraz danych PGA dla wymienionych 2465 SNP wynosi 99.2%. Nasz zbiór SNP zawierał $\sim 24\%$ wszystkich SNP z MAF $\geqslant 10\%$ dla tych 152 genów w próbie afrykańskiej i europejskiej. SNP o niskim MAF występowały stosunkowo rzadko w porównaniu z danymi PGA, ponieważ nasze SNP poszukiwano w sekwencjach z mniejszej liczby niepowtarzalnych chromosomów. Te raczej rzadkie warianty zaliczają się do stosunkowo małych grup w różnorodności nukleotydów. W danych PGA, dla europejczyków, 45% SNP ma MAF $< 10\%$, ale zalicza się do jedynie 15% różnic nukleotydów, dla próby afrykańskiej, 58% SNP ma MAF $< 10\%$ a przypadają na 23% różnorodności nukleotydów.

Tabela 4 ilustruje średnie $r^2$ oraz odsetek wartości $r^2$, przekraczających progi, dla SNP PGA z najbardziej sko-

relowanym SNP w tym samym regionie, który został dodany do naszego zbioru SNP. Wyniki te wskazują że nawet przy dosyć wymagającym założeniu współczynnika $r^2 > 0.8$, nasz zbiór SNP zapewnia $\approx 73\%$ typowych wariacji w próbie europejskiej, oraz $\approx 54\%$ w próbie afrykańskiej. Wartości te są zbliżone do przewidywanych wcześniej, o ile 2,7 miliona SNP z publicznych baz danych zostałoby poddane analizie genotypu (przypis 17). Ta analiza pozwala określić bardzo ostrożne dolne ograniczenie na pokrycie, ponieważ traktuje SNP poniżej współczynnika $r^2 = 0.8$ jako całkowicie nie poznane i nie nagradza pokrycia przekraczającego ten współczynnik. Zakładając mniej wymagające założenie, że $r^2 > 0.5$, pokrycie zwiększyłoby się do 86% w próbie europejskiej i do 71% w próbie afrykańskiej. Skośność rozkładu $r^2$ dla dużych wartości jest widoczna w wartościach średnich 0.84 próby europejskiej i 0.72 próby afrykańskiej. Wartości te sprawiają szczególne wrażenie, gdy pamiętać, że nie badaliśmy 75% wszystkich SNP w tych przedziałach.

Wybór jednego znacznika SNP z każdego kosza LD dla prób trzech populacji dostarczył 296313 spośród 991398 SNP segregowanych w próbie europejskiej (30%), 256766 z 909824 w próbie azjatyckiej (28%), oraz 540533 spośród 1083638 w próbie afrykańskiej (50%). Gdy tagi SNP z próby europejskiej i afrykańskiej zostały wykorzystane do oszacowania typowej wariacji w danych PGA, dla MAF $> 10\%$, liczba wszystkich pewnych typowych wariacji zmniejszyła się bardzo nieznacznie, w porównaniu do wyznaczenia, gdy użyto pełne zbiory typowych SNP (tabela 4). Numery tych tagów SNP są mniejsze niż do tej pory przypuszczano, przy podobnych strategiach wyboru (24), chociaż nie udało nam się osiągnąć 100% pokrycia, jak w tamtej pracy. Chociaż wybieranie podzbiorów SNP opierając się na powiązaniach koszy redukuje koszt genotypowania w przypadku ogólnych badań pełnego genomu we wszystkich populacjach, dane wskazują że nawet podejmując próbę wyznaczenia takich tagów SNP, wyczerpujące badanie powiązań w całym genomie wymaga u każdego przypadku genotypowania setek tysięcy SNP.

**Struktura blokowa haplotypów.** Mapy LD oraz mapy haplotypów przedstawiają różne aspekty lokalnej struktury wariacji genetycznych. Genetyczna architektura pewnego fenotypu określi, która reprezentacja jest najbardziej przydatna do identyfikacji wariantów funkcjonalnych (33). Równolegle z testami LD, użyliśmy programu HAP (34), by wyznaczyć haplotypy z diploidu naszych danych genotypowych. Te zrekonstruowane haplotypy podzieliliśmy na bloki o ograniczonej różnorodności, osobno dla każdej populacji. Te bloki zostały zdefiniowane jako zbiory SNP, dla których co najmniej 80% wyznaczonych haplotypów mogłoby zostać zgrupowane w typowe wzorce występujące w co najmniej 5%.

Tabela 5 podsumowuje strukturę trzes wynikowych map haplotypów dla całego genomu, nie licząc chromosomu Y. Statystyki mapy haplotypu na przestrzeni wszystkich trzech populacji przedstawiają się całkiem podobnie do map LD, przy czym zdecydowanie więcej bloków pochodzi z mapy próby afrykańskiej, niż z europejskiej czy azjatyckiej. Liczby SNP potrzebnych do reprezentowania typowych wzorców haplotypów były zbliżone do liczb tagów SNP zidentyfikowanych w mapach LD. Spora część koszy LD o dwóch lub więcej SNP przecinała granice bloków haplotypu, wahając się od 33% w mapie azjatyckiej do 48% w mapie afrykańskiej.

Struktura koszy SNP w regionie genu CFTR w chromosomie 7 (rysnek 3) demonstruje pewne różnice między kozami LD a mapami bloków haplotypu, ilustruje także że mogą istnieć znaczące różnice populacji w lokalnej strukturze mapy. W tym przedziale, mapy LD europejczyków i Afrykanów mają podobną złożoność, z wieloma nakładającymi się koszami, natomiast mapa azjatycka jest zdominowana przez dwa rozłączne kosze silnie skorelowanych SNP. Odwrotnie, punkt przecięcia w okolicy pozycji 116790 kb jest współdzielony w afrykańskiej i azjatyckiej mapie LD, ale jest podtrzymywany przez wiele zgrupowań LD w mapie europejskiej. Wszystkie trzy mapy haplotypu, współdzielą ten punkt. Chociaż mapa afrykańska zawiera wiele innych bardziej wyjątkowych bloków haplotypu, niż mapy dla pozostałych dwóch populacji.

**Typowe wariacje genetyczne a nasze zdrowie.** Nasze skupienie na typowych wariacjach genetycznych opiera się na istotnej motywacji. Typowe wariacje zaliczają się do większej części różnorodności nukleotydów ludzkich niż rzadkie wariacje oraz łatwiejsze w badaniach eksperymentalnych. Z tego samego powodu, typowe wariacje reprezentują większą część różnorodności fenotypicznej niż rzadkie, tak więc typowe warianty są cenniejsze z punktu widzenia diagnostyki i interwencji. Ostatecznie, wykrywanie i charakteryzowanie skutków rzadkich wariantów wymaga bardzo dużych rozmiarów próby, by osiągnąć wyniki o istotnym znaczeniu statystycznym. Nie ma wątpliwości że rzadkie warianty odgrywają znaczącą rolę w etiologii pospolitych chorób, ale podążanie za typowymi, łatwiejsze także ze względów technologicznych.

Wiele częstych schorzeń, jak choroby układu krążenia, czy choroby psychiatryczne, są skutkiem przeplatania się czynników genetycznych i środowiskowych. Niezmienny charakter genomu oraz lepsza dostępność sekwencji genetycznych umożliwiają coraz większe wysiłki nad zdefiniowaniem genetycznych podstaw różnych zagrożeń ludzkich. Jednym z podejść zmierzającym do zidentyfikowania takich genetycznych współczynników ryzyka jest case-control association study, gdzie w grupie osobników ze stwierdzoną chorobą, zaobserwowano zwiększoną częstotliwość pewnych genetycznych zmian w stosunku do grupy kontrolnej. Wiele współczynników ryzyka dla wybranych chorób zostało właśnie w ten sposób wyznaczone (przypisy 3, 4, 35, 36). Wyniki tych badań sugerują, że wiele genów, rozrzuconych po całym genomie, jako całość wpływają na łączną genetyczną podatność na dane zagrożenie, przy czym każdy pojedynczy gen ma znaczenie nie większe niż kilka procent (37). Praktyka z badań typu case-control study, wskazuje że przeprowadzone na

próbie 1000 osób mogą dostarczyć odpowiednio mocnych podstaw do zidentyfikowania genów odpowiadających co najwyżej za tylko kilka procent całego genetycznego wpływu na jakieś schorzenie, nawet zakładając bardzo rygorystyczny poziom istotności na etapie testowania dużych liczb typowych wariantów DNA (37). Stosowanie się do takich praktyk w połączeniu ze szczegółowym opisem typowych wariacji DNA zawartych tutaj, może być możliwe identyfikowanie zbioru głównych genetycznych czynników ryzyka, wpływających na występowanie schorzeń, lub sposób postępowania w przypadku ich zajścia. O ile wiedza o pojedynczym genetycznym czynniku ryzyka rzadko może pomóc w diagnozie zachorowania, lub sposobu postępowania, to wiedza o dużej części głównych genetycznych czynników ryzyka, może być od razu wykorzystana, pozwalając dostosować znane sposoby postępowania do indywidualnych potrzeb pacjenta, bez wiedzy o mechanizmach prowadzących od różnic genetycznych do różnych objawów.

W analizach przyjęliśmy reprezentację danych, włączając zarówno analizy par LD jak podejście oparte na haplotypach, która była wg. nas najbardziej użyteczna przy ogólnej charakterystyce. Skupiliśmy si na analizach LD parami, ponieważ dostarczają stosunkowo łatwych danych do wyznaczania pokrycia oraz zasobów informacyjnych dla różnych zbiorów SNP. Optymalna reprezentacja wariacji genetycznych nadal pozostaje tematem aktywnych badań. Chociaż określiliśmy przykładowe mapy haplotypu ludzkiego genomu w trzech badanych populacjach, najbardziej odpowiednia reprezentacja danych silnie zależy od specyfiki pytań, na jakie szukamy odpowiedzi. Istnieje wiele map różnorodności genetycznej ludzi, wszystkie dostosowane do specyficznych celów.

**Publiczny dostęp do danych.** Zaimplementowaliśmy generyczną przeglądarkę genomu (38) pod adresem http://genome.perlegen.com, pozwalającą przeglądać SNP, LD oraz haplotypy omawiane w artykule, dane te są także dostępne w redakcji *Science*. Więcej szczegółowych analiz haplotypu jest dostępnych na http://research.calit2.net/hap/wgha oraz poprzez dbSNP. Dane zawarte tutaj znacząco zwiększają liczbę SNP scharakteryzowanych dla wielu populacji. Dla porównania, publiczna baza dbSNP build 122 zawierała mapy pozycji ponad 8.1 miliona ludzkich SNP, z czego tylko dla 797000 dostępne były częstotliwości, w większości z jednej populacji, a genotypy dla jedynie 210000 SNP. Nasze dane wzbogacą także wyniki międzynarodowego projektu HapMap (11), przez udostępnienie danych o wielu SNP od niewielkiej liczby osób.

Powyższa praca pozwala kontynuować szczegółowych analiz struktury genetycznej różnorodności ludzi w skali całego genomu. Przebadaliśmy genetyczną różnorodność jednostek z trzech różnych populacji o różnych historiach oraz opisaliśmy ogólne różnice wewnątrz populacji, jak i pomiędzy nimi. Ponieważ te próby nie pokrywają całej genetycznej różnorodności badanych populacji, na podstawie zgromadzonych danych nie można więc odpowiedzieć na bardzie szczegółowe pytania odot. genetycznej struktury (39). Jednak publiczny dostęp do zawartych danych umożliwi szerokie badania szczegółowych przypadków, zarówno w kwestii różnic genetycznych wśród ludzi, jak i genetycznych podstaw ludzkich różnic fenotypicznych.