Dokumentacja programu PHASE\_KEY wersja 1.0

Jacek Nowak, Elżbieta Graczyk-Pol, Anna Marosz-Rudnicka

Czerwiec 2013

Adres do korespondencji:

Jacek Nowak

Zakład Immunogenetyki

Instytut Hematologii i Transfuzjologii

ul. Indiry Gandhi 14, 02-776 Warszawa

e-mail: szpik@ihit.waw.pl

<http://www.ihit.waw.pl/>

Spis treści

|  |  |
| --- | --- |
|  | Str. |
| Rozdział 1. O programie PHASE\_KEY v. 1.0 | 3 |
| Rozdział 2. Funkcje programu PHASE\_KEY | 3 |
| Rozdział 3. Instrukcja obsługi PHASE\_KEY | 4 |
| 3.1. Przygotowanie arkusza DANE w formacie Excel (populacja wzorcowa  o znanym pochodzeniu etnicznym i lokalizacji geograficznej). | 4 |
| 3.2. Przygotowanie arkusza TYPY w formacie Excel. | 4 |
| 3.3. Instalowanie bazy danych własnej populacji. | 4 |
| 3.4. Uruchamianie programu PHASE\_KEY | 5 |
| 3.5. Obsługa programu PHASE\_KEY | 5 |
| 3.6.Warunki instalacji PHASE i uzyskania licencji akademickiej z University  of Washington, Seattle, USA, na użytkowanie programu PHASE, | 5 |
| 3.7. Przetwarzanie danych za pomocą PHASE v. 2.1. | 6 |
| 3.8. Zwrotne przetwarzanie (odkodowywanie) danych PHASE | 6 |
| 3.9. Formularz wyniku | 8 |
| Rozdział 10. Piśmiennictwo | 9 |

Rozdział 1. O programie PHASE\_KEY v. 1.0

1.1. Autorzy oprogramowania: Jacek Nowak\*) i Jakub Stokwiszewski\*\*)

1.2. Program PHASE\_KEY v. 1.0 stworzono w 2006 roku w \*)Samodzielnej Pracowni Immunogenetyki Instytutu Hematologii i Transfuzjologii, przy współudziale \*\*)Zakładu Statystyki Państwowego Zakładu Higieny.

1.3.1. Program PHASE\_KEY jest realizacją jednego z zadań grantu naukowego pt.: Polimorfizm alleli HLA-A, B, Cw, DRB1 i DQB1 w polskiej populacji: wewnętrzne sprzężenia genowe i dystans genetyczny w stosunku do innych populacji. Program PHASE\_KEY był finansowany ze środków, przyznanych przez Komitet Badań Naukowych/Ministerstwo Nauki i Informatyzacji w latach 2004-2006.

1.3.2. Numer projektu badawczego KBN: 2 P05E 12626

1.3.3. Kierownik projektu: Dr n. med. Jacek Nowak

1.4. Kontakt:

Jacek Nowak

Zakład Immunogenetyki

Instytut Hematologii i Transfuzjologii

ul. Chocimska 5, 00-957 Warszawa

e-mail: szpik@ihit.waw.pl

<http://www.ihit.waw.pl/>

# Rozdział 2. Cel

Program PHASE\_KEY spełnia następujące funkcje:

2.1. Określanie segregacji haplotypów pacjenta na podstawie jego genotypu, poprzez integrację genotypów pacjenta i dawców z genotypami populacji wzorcowej (PHASE\_KEY) i przeprowadzenie populacyjnej segregacji haplotypów (oprogramowanie PHASE v. 2.1),

2.2. Predykcja alleli pominiętych podczas genotypowania pacjenta.

2.2. Przetwarzanie (kodowanie) immunogenetycznych danych genotypowych (HLA, mikrosatelity, SNP, zgodnych z aktualną nomenklaturą genetyczną, opublikowaną przez Komitet Nomenklatury przy WHO) zawartych w arkuszu Microsoft Excel na format wymagany przez program PHASE v. 2.1 oraz bezbłędne przetwarzanie zwrotne (odkodowywanie) tekstowych plików wynikowych PHASE 2.1 na nazwy alleli HLA, mikrosatelitów i SNP z zachowaniem prawidłowej segregacji haplotypów, częstości alleli i haplotypów i innych danych statystycznych wygenerowanych przez PHASE 2.1.

2.3. Określanie częstości genowych alleli i haplotypów oraz wybranych statystyk z nimi związanych, **nie określanych** przez PHASE 2.1.

2.4. Budowa tablic częstości par alleli różnych loci, zawierających:

2.4.1. Częstość genową (GF),

2.4.2. Wartość Delty (D)

2.4.3. Wartość względnej delty (RD),

2.4.4. Wartość p, istotność statystyczną delty (istotne p zaznaczono czcionką **bold**)

2.5. Program obsługuje loci typu M (HLA, mikrosatelity, polimorfizm pojedynczego nukleotydu [SNP] z substytucją 3 lub 4 zasad, jako alternatywa) oraz loci typu S (SNP z substytucją tylko 1 pary alternatywnych zasad).

# Rozdział 3. Instrukcja obsługi PHASE\_KEY

3.1. Przygotowanie arkusza DANE w formacie Excel

Uwaga: Arkusz DANE służy do podstawienia danych własnej populacji, zbliżonej etnicznie do badanego pacjenta, o ile nie chcemy skorzystać z populacji polskiej (populacja wzorcowa o znanym pochodzeniu etnicznym i lokalizacji geograficznej). W arkuszu DANE wbudowano ponadto klawisze UTWÓRZ PLIK oraz WCZYTAJ WYNIKI uruchamiające odpowiednie funkcje programu PHASE\_KEY.

3.1.1. Wiersz tytułowy arkusza Excel powinien zawierać nazwy kolumn:

3.1.1.1. kolumna A: Lp. (liczba porządkowa),

3.1.1.2. kolumna B: ID (identyfikator osoby badanej, ciąg do 10 znaków),

3.1.1.3. kolumna C: nazwa locus1\_1,

3.1.1.4. kolumna D: nazwa locus1\_2,

3.1.1.4. kolumna E: nazwa locus2\_1,

3.1.1.5. kolumna F; nazwa locus2\_2, itd.   
Liczba loci nie jest limitowana, aczkolwiek przeprowadzane testy obejmowały 5 lub 6 loci. Kolejność loci musi być zgodna z ich położeniem na chromosomie (np. dla układu HLA - A, C, B, itd.).

3.1.2. Wiersz 2 i kolejne zawierają dane genotypowe osób należących do populacji wzorcowej, w kolumnach zgodnych z p. 3.1.1.

3.1.3. Liczba osób losowej próby (populacji wzorcowej) jest dowolna (testowano liczebności 20-800 osób).

Uwaga: Dla loci typu M (typy loci omówiono w p. 3.2.2.) możliwe jest wpisanie 4 znakowych (high resolution) oznaczeń alleli (zalecane) lub 2 znakowych oznaczeń (low resolution). Zapisanie 2 znakowej wartości choćby tylko jednego allelu u jednej osoby populacji wzorcowej spowoduje, że wszystkie osoby posiadające allele tej grupy zostaną potraktowane jako typy 2 znakowe (np. wpisanie u trzech osób w locus A\* typów 0201, 0205 i 02 spowoduje skumulowanie tych trzech osób w grupie 02 locus A\*. Pozostałe grupy locus A\* i innych loci, pozostaną na poziomie 4 znakowym, o ile wszystkie osoby populacji wzorcowej zostały zbadane na poziomie 4 znaków, tj, high resolution). Nie jest możliwe wprowadzenie zapisu alleli o innej liczbie znaków niż 4 lub 2 dla loci typu M, lub dwóch jednoznakowych wartości dla loci typu S .

3.2. Przygotowanie arkusza TYPY w formacie Excel.

Uwaga: Arkusz TYPY zawiera informacje określające odległości pomiędzy genetycznymi loci i ich rodzaje. Informacje te są wykorzystywane przez program PHASE do wyliczeń częstości rekombinacji i ustalania segregacji haplotypów.

3.2.1. Pierwszy wiersz zawiera liczby określające położenie wszystkich loci (np. odległość środka genu lub użytj w badaniach sondy lub primera) od punktu odniesienia, wyrażona w bp. Dane nt. pozycji genów można uzyskać min. pod adresem internetowym: http://www.ncbi.nlm.nih.gov

3.2.2. Drugi wiersz zawiera typy loci. Dla loci wieloallelicznych (HLA, mikrosat.) – wpisać kod M w kolejnych komórkach, dla loci SNP wpisać kod S (tylko SNP z 2 alternatywnymi zasadami. W przypadku możliwości zamiennego występowania w danej pozycji 3 lub 4 różnych zasad użyć kodu M).

3.3. Dane arkuszy: DANE i TYP zastąpić własnymi w powyższym formacie.

3.4. Aby uruchomić program PHASE\_KEY należy zamknąć wszystkie otwarte arkusze Excel, uruchomić program pod nazwą „PHASE\_KEY\_1\_1.xls” koniecznie włączając makra. Następnie kliknąć przycisk „UTWÓRZ PLIK .inp” w arkuszu DANE.

Uwaga:

* Program PHASE\_KEY funkcjonuje prawidłowo (w ogóle działa), jeśli **nie są otwarte** inne pliki MS Excel. Należy zamknąć inne arkusze Excel przed uruchomieniem.
* Makra muszą być włączone w momencie uruchamiania PHASE\_KEY.

3.5. Postępować zgodnie z poleceniami programu PHASE\_KEY, wypełniając pojawiające się okienka:

3.5.1. Na pytanie „Czy dodać kolejny rekord?” należy odpowiedzieć „Tak” (tyle razy u ilu pacjentów lub dawców chcemy określić segregację haplotypów), nacisnąć OK

3.5.2. Podać unikalny identyfikator lub nazwisko i imię chorego1, u którego określamy segregację haplotypów, nacisnąć OK,

3.5.3. Podać nazwę, jaką ma przybrać tworzony plik tekstowy (ciąg do 10 znaków, np. <nazwapliku>), nacisnąć OK,

3.5.4. Podać pierwszy allel pierwszego locus, nacisnąć OK (dla loci typu M **tylko dwucyfrowy lub czterocyfrowy zapis genetyczny**),

3.5.5. Podać drugi allel pierwszego locus, nacisnąć OK (allele nieznane pomijać naciskając OK bez wpisywania ich wartości),

3.5.6. Podać pierwszy allel drugiego locus (uwaga na kolejność loci – tak jak na chromosomie, np A, C, B, a nie ABC), nacisnąć OK (loci nieznane pomijać naciskając OK bez wpisywania ich wartości w żadnym z dwóch okienek),

3.5.7. Podać pierwszy allel trzeciego locus, itd.

**UWAGA: Aby uzyskać propozycje wartości brakujących alleli (np. jeżeli locus nie było typowane, lub propozycje drugiego allelu w locus C typowanym serologicznie) należy podczas wypełniania okienek oznaczyć ‘missing value’, tj. brakującą wartość. W locus typu M należy wpisać ‘–1’ (minus jeden), a w locus typu S wpisać ‘?’ (znak zapytania).**

3.5.8. Po ukazaniu się podsumowania genotypu pacjenta1 należy potwierdzić poprawność wpisu, nacisnąć „Tak” lub „Nie”, (wybranie „Nie” umożliwi wpis od nowa),

3.5.9. Podać unikalny identyfikator lub nazwisko i imię chorego2, nacisnąć OK,

3.5.10. Podać pierwszy allel pierwszego locus, nacisnąć OK

3.5.11. Podać drugi allel pierwszego locus, nacisnąć OK , itd.

3.5.12. Potwierdzić poprawność wpisu ostatniego pacjenta,

3.5.13. Na pytanie „Czy dodać kolejny rekord?” odpowiedzieć „Nie”

3.5.14. W kolejno ukazujących się okienkach potwierdzić segregację których loci należy analizować.

UWAGA: Liczba analizowanych loci jest ograniczona zawartością arkusza DANE, ale może być mniejsza, jeśli oba okienka alleli w którymś locus pozostaną niewypełnione. Liczba analizowanych loci **musi** być taka sama dla wszystkich badanych pacjentów. Brakujące dane uzupełnić wpisując –1 (dla loci typu M) lub ? (dla loci typu S).

3.5.13. Po naciśnięciu OK w okienku potwierdzenia wpisu ostatniego pacjenta pojawia się okienko: Tworzenie pliku <nazwapliku.inp>, nacisnąć OK

3.5.14. Określić „Pod jaką nazwą zapisać plik \*.inp, (wpisać nazwę pliku bez rozszerzenia (**najlepiej 8 znaków**), a rozszerzenie .inp zostanie dopisane automatycznie).

3.5.15. Plik tekstowy w formacie PHASE o nazwie <nazwapliku.inp> jest zapisywany do katalogu, w którym znajduje się program PHASE\_KEY

3.6. Dalsze etapy zakładają, że użytkownik programu PHASE\_KEY v. 1.0 jest posiadaczem licencji akademickiej na użytkowanie programu PHASE, oraz samego programu PHASE v. 2.1, Oba programy należy umieścić w tym samym katalogu. Program PHASE oraz licencję można uzyskać pod adresem internetowym:   
<http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/PHASEv2.php>

**3.7. Przetwarzanie zakodowanych danych za pomocą PHASE v. 2.1.**

3.7.1. Uruchomić wiersz poleceń programu DOS lub Total Commander

3.7.2. Przejść do katalogu, w którym znajdują się oba programy (PHASE i PHASE\_KEY), poprzez wpisanie w wierszu poleceń odpowiedniej komendy. **Nazwa pliku .out powinna mieć również 8 znaków.** (Jeśli np. programy PHASE i PHASE\_KEY znajdują się w podkatalogu PHASE w katalogu Moje dokumenty, a wiersz poleceń w DOSie wygląda tak:

C:\>

to należy wpisać (treść wpisu wyróżniono boldem):

C:\>**cd Moje dokumenty\PHASE**(enter)

aby otrzymać na ekranie:

C:\Moje dokumenty\PHASE>

3.7.3 Po przejściu do odpowiedniego katalogu należy w wierszu poleceń dopisać polecenie (treść wpisu wyróżniono boldem):

C:\Moje dokumenty\PHASE>**Phase –f1 nazwapliku.inp nazwaplik2.out 200 1 200**(enter)

3.7.4. Program PHASE rozpocznie obliczenia mogące trwać kilka minut lub dłużej, zależnie od mocy obliczeniowej posiadanego komputera.

3.7.5. We wspólnym katalogu zostaną zapisane tekstowe pliki wynikowe o wspólnej nazwie <nazwapliku> lecz z rozszerzeniem .out\_freqs, .out\_pairs, .out\_probs i podobne, gotowe do odkodowania przez program PHASE\_KEY.

Uwaga: możliwe są modyfikacje wpisywanego polecenia (i modelowanie otrzymywanych wyników obliczeń PHASE), zgodnie z zasadami określonymi w „[Documentation for PHASE, version 2.1](http://www.stat.washington.edu/stephens/instruct2.1.pdf)” autorstwa M. Stephens’a, NJ. Smith’a i P. Donnely’ego z Department of Statistics, University of Washington, Seattle, USA.

**3.8. Zwrotne przetwarzanie (odkodowywanie) dawnych PHASE**

3.8.1. Kliknąć przycisk „WCZYTAJ WYNIKI” w arkuszu DANE programu PHASE\_KEY.

3.8.2. W pojawiającym się okienku wpisać nazwę pliku wytworzonego przez program PHASE 2.1, <nazwaplik2> (bez rozszerzenia) - taką jak w komendzie wiersza poleceń (p. 3.7.3.) i nacisnąć OK.

**3.9. WYNIK**

3.9.1.Stworzony zostanie plik .exe zawierający podsumowanie wyników określania segregacji haplotypów wszystkich badanych osób. Obok każdego rozpatrywanego haplotypu podano jego częstość genową. Pary haplotypów posegregowano od najbardziej do najmniej prawdopodobnej, a w ostatniej kolumnie umieszczono prawdopodobieństwo wystąpienia danej pary haplotypów u danej osoby. Poniżej wyniku znajduje się miejsce na komentarz immunogenetyczny.

3.9.2. Wynik można wydrukować zgodnie z obsługą programu EXCEL.

3.9.3. Załącznik Nr 1 zawiera przykładowy wynik analizy haplotypowej.

Rozdział 10. Piśmiennictwo

10.1. Stephens, M., Smith, N., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics, 68, 978--989.

10.2. Stephens, M., and Scheet, P. (2005). Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. American Journal of Human Genetics, 76:449-462.

10.3. Crawford et al (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. Nature Genetics, 36:700-706.

10.4. Li, N., and Stephens, M. (2003). Modelling Linkage Disequilibrium, and identifying recombination hotspots using SNP data. Genetics, 165:2213-2233.

10.5. Stephens, M., and Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. American Journal of Human Genetics, 73:1162-1169.

10.6. Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics 164: 1567-1587.

10.7. Stephens, M. and Donnelly, P. (2003). Ancestral inference in population genetics models with selection. Australian and New Zealand Journal of Statistics, 45 , 901--931.

Załącznik Nr 1

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Welcome to PHASE\_KEY**, PATIENTS Module  Haplotype Segregation Software | | | | | | | | | | | | | | | |
| Institute of Haematology and Blood Transfusion, Warsaw, Poland | | | | | | | | | | | | | | | |
| http://www.ihit.waw.pl/ | | | | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| SUMMARY RESULTS OF HAPLOTYPE SEGREGATION | | | | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Patient1: F. D. (chora) | | | | | | | | | | | | | | | |
| A | Cw | B | DRB1 | DQB1 | HF |  | A | Cw | B | DRB1 | DQB1 | HF |  | *P* |  |
| **0201** | **1203** | **3906** | **1601** | **0502** | **1,6%** | **,** | **3101** | **0501** | **5101** | **1101** | **0301** | **1,00%** | **,** | **0,424** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Patient2: F. J. (brat) | | | | | | | | | | | | | | | |
| A | Cw | B | DRB1 | DQB1 | HF |  | A | Cw | B | DRB1 | DQB1 | HF |  | *P* |  |
| **1101** | **0701** | **0801** | **0301** | **0201** | **0,00824** |  | **0201** | **0202** | **2705** | **0101** | **0501** | **0,00728** |  | **0,967** |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Population detailes: | | | | | | | | | | | | |  |  |  |
| Sample size: | | 200 | | | |  | Country: | | Poland | | | |  |  |  |
| Race: | | Caucasian/White | | | |  | Region: | | Whole country | | | |  |  |  |
| Ethnicity: | | Polish | | | |  | Bias: | | Unbiased healthy voluntears | | | |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| DETAILED RESULTS OF HAPLOTYPE SEGREGATION | | | | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| IND: F. D. | | |  |  |  |  |  |  |  |  |  |  |  |  |  |
| A | Cw | B | DRB1 | DQB1 | HF |  | A | Cw | B | DRB1 | DQB1 | HF |  | P |  |
| 0201 | 1203 | 3906 | 1601 | 0502 | 0,0162 | , | 3101 | 0501 | 5101 | 1101 | 0301 | 0,0102 | , | 0,424 |  |
| 0201 | 1203 | 5101 | 1601 | 0502 | 0,0074 | , | 3101 | 0501 | 3906 | 1101 | 0301 | 0,0029 | , | 0,294 |  |
| 0201 | 1203 | 5101 | 1101 | 0301 | 0,00236 | , | 3101 | 0501 | 3906 | 1601 | 0502 | 0,0020 | , | 0,207 |  |
| 0201 | 0501 | 5101 | 1101 | 0301 | 0,00153 | , | 3101 | 1203 | 3906 | 1601 | 0502 | 0,0011 | , | 0,047 |  |
| 0201 | 0501 | 3906 | 1601 | 0502 | 0,00022 | , | 3101 | 1203 | 5101 | 1101 | 0301 | 0,0005 | , | 0,015 |  |
| IND: F. J. | | |  |  |  |  |  |  |  |  |  |  |  |  |  |
| A | Cw | B | DRB1 | DQB1 | HF |  | A | Cw | B | DRB1 | DQB1 | HF |  | P |  |
| 1101 | 0701 | 0801 | 0301 | 0201 | 0,00824 |  | 0201 | 0202 | 2705 | 0101 | 0501 | 0,00728 |  | 0,967 |  |
| 1101 | 0202 | 2705 | 0101 | 0501 | 0,00005 |  | 0201 | 0701 | 0801 | 0301 | 0201 | 0,00842 |  | 0,033 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Comments: | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Częstości obu najbardziej prawdopodobnych haplotypów pacjentki D. F. są nadzwyczaj niskie (<0,5%). Brak przynajmniej 1 haplotypu o częstości >1% sugeruje b. małą szansę na znalezienie w pełni zgodnego dawcy w szerokiej rodzinie pacjentki. Wywiad potwierdzający istnienie pokrewieństwa małżeństw w rodzinie chorej mógłby zwiększyć szansę na znalezienie dawcy. |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |