РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2018 106 934** (13) **A**

(51) МПК
G06F 19/22 (2011.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ЗАЯВКА НА ИЗОБРЕТЕНИЕ

(21)(22) Заявка: **2018106934, 25.08.2016**

Приоритет(ы):
(30) Конвенционный приоритет:
25.08.2015 US 62/209,858

(43) Дата публикации заявки: **26.09.2019** Бюл. № **27**

(85) Дата начала рассмотрения заявки PCT на национальной фазе: **26.03.2018**

(86) Заявка PCT:
US 2016/048768 (25.08.2016)

(87) Публикация заявки PCT:
WO 2017/035392 (02.03.2017)

Адрес для переписки:
190000, Санкт-Петербург, БОКС-1125

(71) Заявитель(и):
НАНТОМИКС, ЛЛС (US)

(72) Автор(ы):
САНБОРН, Джон Закари (US)

(54) СИСТЕМЫ И СПОСОБЫ ВЫСОКОТОЧНОГО ОПРЕДЕЛЕНИЯ ВАРИАНТОВ

(57) Формула изобретения

1. Способ in silico предсказания HLA-типа для пациента, включающий:

обеспечение референсной последовательности, которая включает набор последовательностей известных и различных аллелей HLA;

обеспечение набора ридов последовательностей пациента, причем по меньшей мере некоторые из указанных ридов последовательностей пациента включают последовательность, кодирующую пациент-специфический HLA;

разделение набора ридов последовательностей пациента на набор соответствующих множеств k-меров;

генерацию составного графа де Брёйна с применением референсной последовательности и указанного набора соответствующих множеств k-меров; и

ранжирование каждого из известных и различных аллелей HLA с применением составного показателя совпадения, который рассчитывается по соответствующим баллам указанного набора ридов последовательностей пациентов, и при этом в каждый балл вносят вклад k-меры, которые совпадают с соответствующими сегментами в указанных известных и различных аллелях HLA.

2. Способ по п. 1, характеризующийся тем, что указанная референсная последовательность включает аллели для по меньшей мере одного HLA-типа, которые имеют частоту аллеля, равную по меньшей мере 1%.

3. Способ по п. 1, характеризующийся тем, что указанная референсная последовательность включает по меньшей мере десять разных аллелей для по меньшей

RU 2018106934 A

RU 2018106934 A

мере одного HLA-типа.

4. Способ по п. 1, характеризующийся тем, что указанная референсная последовательность включает аллели для по меньшей мере двух других HLA-типов.

5. Способ по п. 1, характеризующийся тем, что HLA-тип представляет собой HLA-A-тип, HLA-B-тип, HLA-C-тип, HLA-DRB-1-тип и/или HLA-DQB-1-тип.

6. Способ по п. 1, характеризующийся тем, что указанный набор ридов последовательностей пациентов содержит по меньшей мере один из набора ридов секвенирования ДНК и ридов секвенирования РНК.

7. Способ по п. 1, характеризующийся тем, что указанные риды последовательностей пациента картированы на хромосому 6p21.3.

8. Способ по п. 1, характеризующийся тем, что указанные риды последовательностей пациента представляют собой риды секвенирования следующего поколения и дополнительно содержат метаданные.

9. Способ по п. 1, характеризующийся тем, что указанные риды последовательностей пациента имеют длину от 50 до 250 оснований.

10. Способ по п. 1, характеризующийся тем, что указанные k-меры имеют длину 10-20.

11. Способ по п. 1, характеризующийся тем, что указанные k-меры имеют длину от 5% до 15% длины рида последовательности пациента.

12. Способ по п. 1, характеризующийся тем, что указанный составной показатель совпадения представляет собой сумму всех баллов от указанного набора ридов последовательностей пациентов.

13. Способ по п. 1, характеризующийся тем, что указанный балл представляет собой значение, представляющее долю совпадающих k-меров от общего числа k-меров на рид последовательности пациента.

14. Способ по п. 1, дополнительно включающий этап идентификации лидирующего после ранжирования аллеля HLA как первого HLA-типа пациента.

15. Способ по п. 14, дополнительно включающий переранжирование оставшихся не лидирующих известных и различных аллелей HLA с применением откорректированного показателя совпадения для идентификации лидирующего после ранжирования с коррекцией аллеля HLA как второго HLA-типа пациента.

16. Способ по п. 15, характеризующийся тем, что откорректированный составной показатель совпадения вычисляют по соответствующим откорректированным баллам указанного набора ридов последовательностей пациентов.

17. Способ по п. 16, характеризующийся тем, что указанные откорректированные баллы вычисляют путем понижения веса k-мера, который совпадает с первым HLA-типом.

18. Способ по любому из предшествующих пунктов, характеризующийся тем, что указанная референсная последовательность включает аллели для по меньшей мере одного HLA-типа, которые имеют частоту аллеля, равную по меньшей мере 1%, или тем, что указанная референсная последовательность включает по меньшей мере десять разных аллелей для по меньшей мере одного HLA-типа, или тем, что указанная референсная последовательность включает аллели для по меньшей мере двух других HLA-типов.

19. Способ по любому из пп. 1-17, характеризующийся тем, что указанные k-меры имеют длину 10-20, или тем, что указанные k-меры имеют длину от 5% до 15% длины рида последовательности пациента.

20. Способ по любому из пп. 1-17, характеризующийся тем, что указанный составной показатель совпадения представляет собой сумму всех баллов от указанного набора ридов последовательностей пациентов, и/или тем, что указанный балл представляет

RU 2018106934 A

RU 2018106934 A

собой значение, представляющее долю совпадающих k-меров от общего числа k-меров на рид последовательности пациента.

21. Компьютерная система для in silico предсказания HLA-типа для пациента, включающая:

базу данных референсных последовательностей, в которой хранится референсная последовательность, которая включает набор последовательностей известных и различных аллелей HLA;

источник данных о последовательностях пациента, хранящий и обеспечивающий набор ридов последовательностей пациента, причем по меньшей мере некоторые из указанных ридов последовательностей пациента включают последовательность, кодирующую пациент-специфический HLA;

аналитическую систему, запрограммированную, чтобы

(i) разделять указанный набор ридов последовательностей пациентов на набор соответствующих множеств k-меров;

(ii) генерировать составной граф де Брёйна с применением референсной последовательности и указанного набора соответствующих множеств k-меров; и

(iii) ранжировать каждый из указанных известных и различных аллелей HLA с применением составного показателя совпадения, который вычисляется по соответствующим баллам указанного набора ридов последовательностей пациентов, при этом в каждый балл вносят вклад k-меры, которые совпадают с соответствующими сегментами в известных и различных аллелях HLA.

22. Компьютерная система по п. 21, характеризующаяся тем, что указанная референсная последовательность включает аллели для по меньшей мере одного HLA-типа, которые имеют частоту аллеля, равную по меньшей мере 1%, или тем, что указанная референсная последовательность включает по меньшей мере десять разных аллелей для по меньшей мере одного HLA-типа, или тем, что указанная референсная последовательность включает аллели для по меньшей мере двух других HLA-типов.

23. Компьютерная система по п. 21, характеризующаяся тем, что HLA-тип представляет собой HLA-A-тип, HLA-B-тип, HLA-C-тип, HLA-DRB-1-тип и/или HLA-DQB-1-тип.

24. Компьютерная система по п. 21, характеризующаяся тем, что указанный набор ридов последовательностей пациентов содержит по меньшей мере один из набора ридов секвенирования ДНК и ридов секвенирования РНК.

25. Компьютерная система по п. 21, характеризующаяся тем, что указанные риды последовательностей пациента картированы на хромосому 6p21.3.

26. Компьютерная система по п. 21, характеризующаяся тем, что указанные риды последовательностей пациента представляют собой риды секвенирования следующего поколения и дополнительно содержат метаданные, или тем, что указанные риды последовательностей пациента имеют длину от 50 до 250 оснований.

27. Компьютерная система по п. 21, характеризующаяся тем, что указанные k-меры имеют длину 10-20, или тем, что указанные k-меры имеют длину от 5% до 15% длины рида последовательности пациента.

28. Компьютерная система по п. 21, характеризующаяся тем, что указанный составной показатель совпадения представляет собой сумму всех баллов от указанного набора ридов последовательностей пациентов.

29. Компьютерная система по п. 21, характеризующаяся тем, что указанный балл представляет собой значение, представляющее долю совпадающих k-меров от общего числа k-меров на рид последовательности пациента.

30. Компьютерная система по п. 21, характеризующаяся тем, что аналитическая система дополнительно запрограммирована для того, чтобы идентифицировать

RU 2018106934 A

RU 2018106934 A

лидирующий после ранжирования аллель HLA как первый HLA-тип пациента.

31. Компьютерная система по п. 21, характеризующаяся тем, что аналитическая система дополнительно запрограммирована для того, чтобы переранжировать оставшиеся нелидирующие известные и различные аллели HLA с применением откорректированного показателя совпадения для идентификации лидирующего после ранжирования с коррекцией аллеля HLA как второго HLA-типа пациента.

32. Компьютерная система по п. 31, характеризующаяся тем, что аналитическая система дополнительно запрограммирована так, чтобы вычислять откорректированный составной показатель совпадения по соответствующим откорректированным баллам указанного набора ридов последовательностей пациентов.

33. Компьютерная система по п. 32, характеризующаяся тем, что аналитическая система дополнительно запрограммирована, чтобы вычислять откорректированные баллы путем понижения веса k-мера, который совпадает с первым HLA-типом.

34. Энергонезависимый компьютерочитаемый носитель, содержащий программные инструкции, обеспечивающие выполнение компьютерной системой, в которой база данных референсных последовательностей и источник данных последовательностей пациента информационно связаны с аналитической системой, способа, включающего следующие этапы:

передачу из базы данных референсной последовательности, в аналитическую систему, референсной последовательности, которая включает набор последовательностей известных и различных аллелей HLA;

передачу, из источника данных о последовательностях пациента, в аналитическую систему, набора ридов последовательностей пациента, причем по меньшей мере некоторые из указанных ридов последовательностей пациента включают последовательность, кодирующую пациент-специфический HLA;

разделение аналитической системой указанного набора ридов последовательностей пациентов на набор соответствующих множеств k-меров;

генерацию аналитической системой составного графа де Брёйна с применением референсной последовательности и указанного набора соответствующих множеств k-меров; и

ранжирование аналитической системой каждого из известных и различных аллелей HLA с применением составного показателя совпадения, который рассчитывается по соответствующим баллам указанного набора ридов последовательностей пациентов, при этом в каждый балл вносят вклад k-меры, которые совпадают с соответствующими сегментами в известных и различных аллелях HLA.

35. Компьютерочитаемый носитель по п. 34, характеризующийся тем, что указанная референсная последовательность включает аллели для по меньшей мере одного HLA-типа, которые имеют частоту аллеля, равную по меньшей мере 1%, или тем, что указанная референсная последовательность включает по меньшей мере десять разных аллелей для по меньшей мере одного HLA-типа, или тем, что указанная референсная последовательность включает аллели для по меньшей мере двух других HLA-типов.

36. Компьютерочитаемый носитель по п. 34, характеризующийся тем, что HLA-тип представляет собой HLA-A-тип, HLA-B-тип, HLA-C-тип, HLA-DRB-1-тип, и/или HLA-DQB-1-тип.

37. Компьютерочитаемый носитель по п. 34, характеризующийся тем, что указанный набор ридов последовательностей пациентов содержит по меньшей мере один из набора ридов секвенирования ДНК и ридов секвенирования РНК.

38. Компьютерочитаемый носитель по п. 34, характеризующийся тем, что указанные риды последовательностей пациента картированы на хромосому 6p21.3, или тем, что указанные риды последовательностей пациента представляют собой риды

RU 2018106934 A

RU 2018106934 A

секвенирования следующего поколения и дополнительно содержат метаданные, или, тем, что указанные риды последовательностей пациента имеют длину от 50 до 250 оснований.

39. Компьютерочитаемый носитель по п. 34, характеризующийся тем, что указанные k-меры имеют длину 10-20, или тем, что указанные k-меры имеют длину от 5% до 15% длины рида последовательности пациента.

40. Компьютерочитаемый носитель по п. 34, характеризующийся тем, что указанный составной показатель совпадения представляет собой сумму всех баллов от указанного набора ридов последовательностей пациентов.

41. Компьютерочитаемый носитель по п. 34, характеризующийся тем, что указанный балл представляет собой значение, представляющее долю совпадающих k-меров от общего числа k-меров на рид последовательности пациента.

42. Компьютерочитаемый носитель по п. 34, дополнительно включающий этап идентификации лидирующего после ранжирования аллеля HLA как первого HLA-типа пациента.

43. Компьютерочитаемый носитель по п. 42, характеризующийся тем, что способ дополнительно включает этап переранжирования остальных нелидирующих известных и различных аллелей HLA с применением откорректированного показателя совпадения для идентификации лидирующего после ранжирования с коррекцией аллеля HLA как второго HLA-типа пациента.

44. Компьютерочитаемый носитель по п. 43, характеризующийся тем, что откорректированный составной показатель совпадения вычисляют по соответствующим откорректированным баллам указанного набора ридов последовательностей пациентов.

45. Компьютерочитаемый носитель по п. 44, характеризующийся тем, что откорректированные баллы вычисляют путем понижения веса k-мера, который совпадает с первым HLA-типом.

RU 2018106934 A

RU 2018106934 A