Интерпретация данных NGS: сложности и пути решения

Никитин А.Г.
ФГБУ ФНКЦ ФМБА России
Гордиев М.Г.
ГАУЗ «РКОД МЗ РТ»

Уровни анализа данных NGS

- Первый уровень – сырые данные и файлы FASTQ
- Второй уровень – поиск вариантов и аннотация
- Третий уровень – интерпретация и принятие решений

Первый уровень – сырые данные и файлы FASTQ

- Проблема хранения:
  - Таргетная панель – 200-500 Мб
  - Экзом – 8 Гб
  - Геном – 90 Гб
- Институт Броада генерирует геном раз в 12 минут – это примерно 4000 Тб в год, общий объем хранилища больше хранилища Фейсбука для фотографий

Второй уровень – поиск вариантов и аннотация

- Проблема вычислительных мощностей:
  - Таргетная панель – 30 мин
  - Экзом – 3-4 часа
  - Геном – около суток
- Выбор алгоритмов – критерий скорость, а не качество
- Массовые рутинные исследования требуют соответствующей инфраструктуры, сейчас значительная часть выполняется «на коленке»

Третий уровень – интерпретация и принятие решений

- Проблема баз данных:
  - Отсутствуют данные о встречаемости вариантов в популяции
  - Отсутствует доступ к платным базам данных зарубежных компаний (HGDM, CentoMD и др.)
  - Слабая применимость чужих данных к некоторым российским популяциям
  - Фрагментарные и разрозненные данные собственных исследований

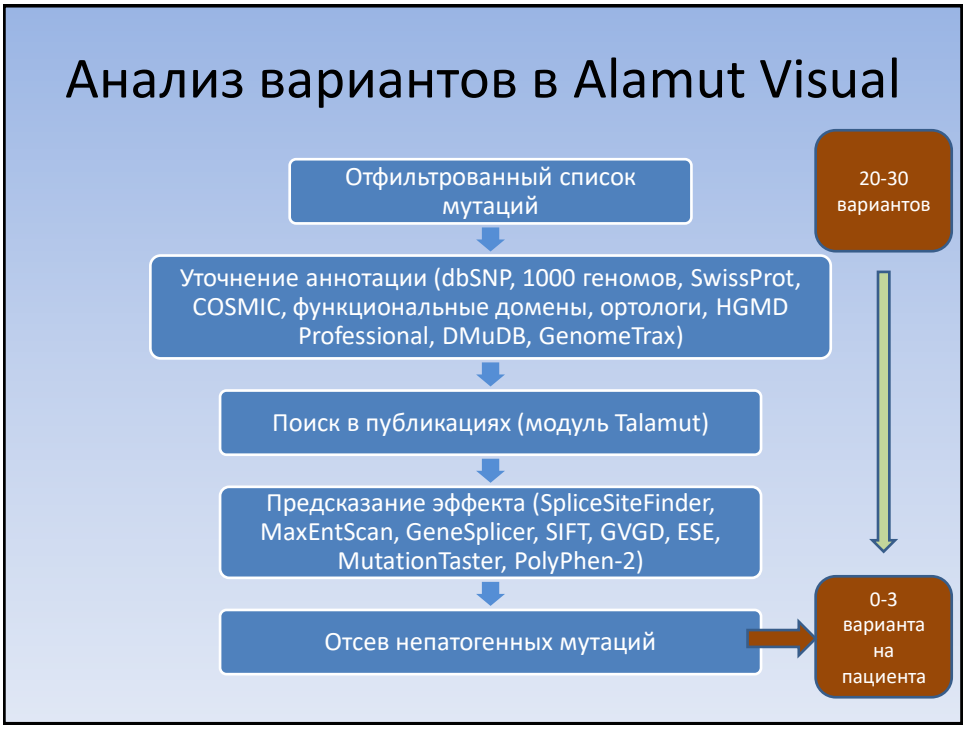
Проблемы внедрения NGS

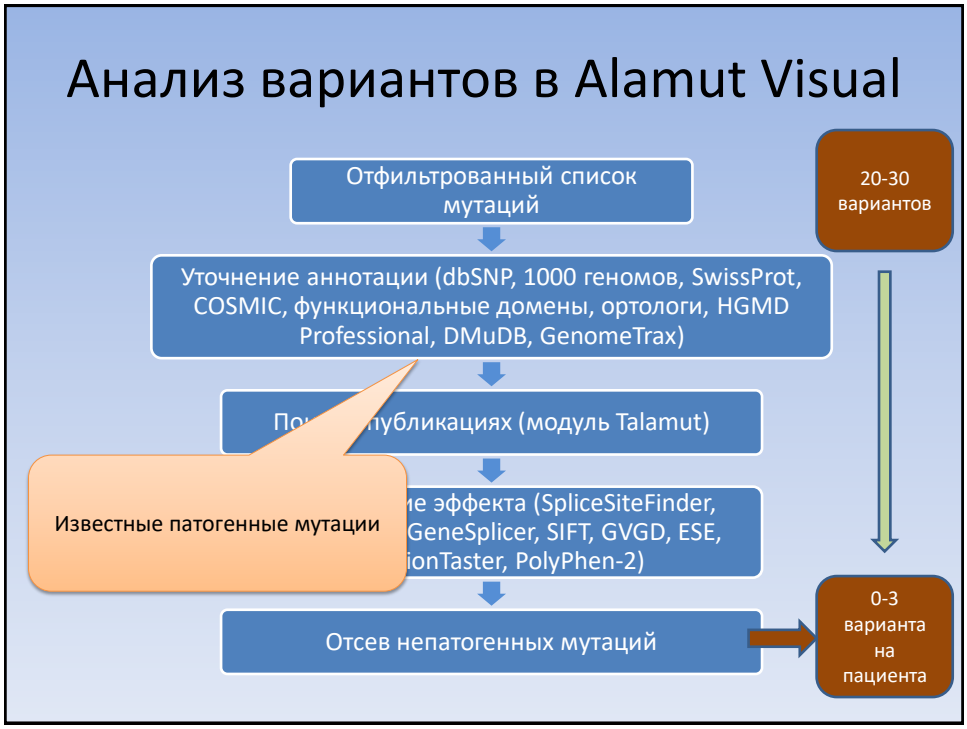
Проблемы внедрения NGS

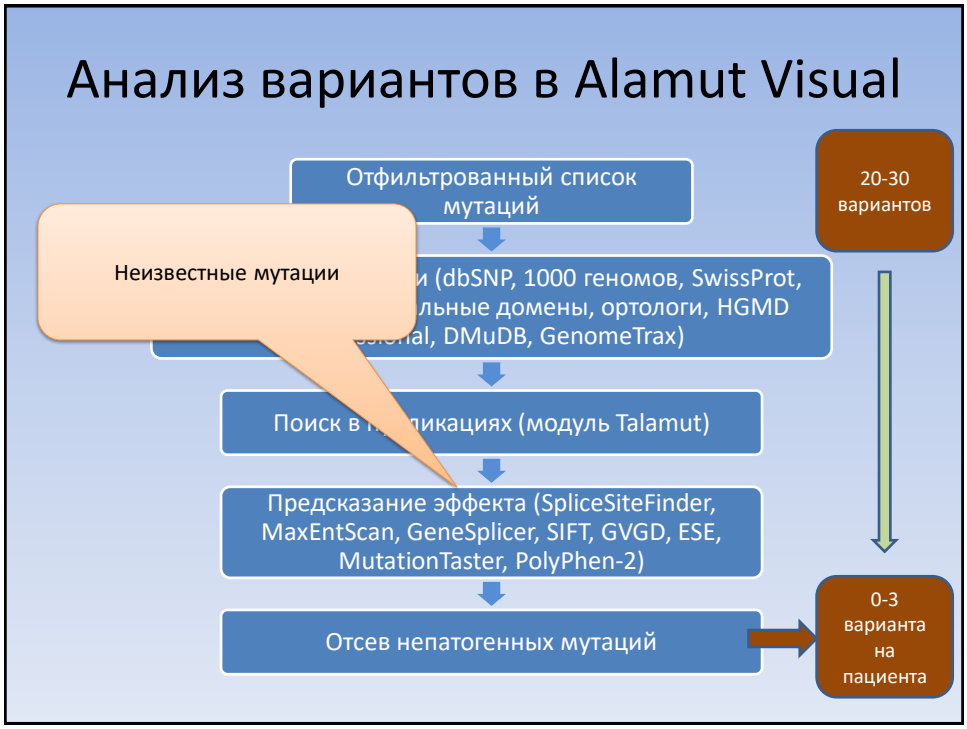
Поиск вариантов

Анализ вариантов в Alamut Batch

Анализ вариантов в Alamut Visual




Анализ вариантов в Alamut Visual
Отфильтрованный список мутаций
20-30 вариантов
Уточнение аннотации (dbSNP, 1000 геномов, SwissProt, COSMIC, функциональные домены, ортологи, HGMD Professional, DMuDB, GenomeTrax)
Известные патогенные мутации
0-3 варианта на пациента
Отсев непатогенных мутаций


Анализ вариантов в Alamut Visual
Отфильтрованный список мутаций
20-30 вариантов
Неизвестные мутации
Предсказание эффекта (SpliceSiteFinder, MaxEntScan, GeneSplicer, SIFT, GVGD, ESE, MutationTaster, PolyPhen-2)
0-3 варианта на пациента
Отсев непатогенных мутаций

Проблемы интерпретации

HGMD Professional, CentoMD, ClinVar и др.

- Источник справочной информации о публикациях, в которых мутация признавалась патогенной
- Достоверность записей различается между базами данных

Популяционные частоты

- Частота варианта в общей популяции – ключевой критерий для клинической интерпретации
- Частота аллели выше ожидаемой для заболевания – аргумент в пользу непатогенности варианта

Популяционные частоты

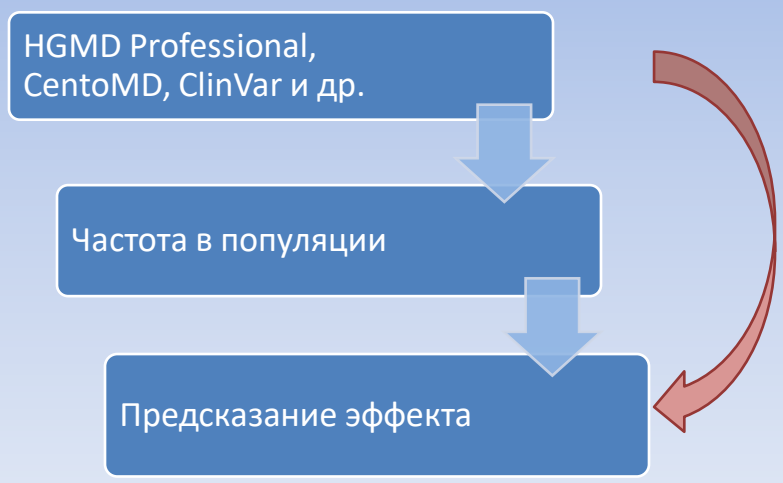
- 2012 г., Exome Aggregation Consortium (*ExAC*)
  - 60 706 экзомов
  - 26 PI, 22 проекта
- 2017 г., Genome Aggregation Database (*gnomAD*)
  - 123 136 экзомов, 15 496 геномов
  - 108 PI, 47 проектов
- 2017 г., UK Biobank
  - 337 000 образцов

Популяционные частоты

- 2018 г., Российские геномы (Санкт-Петербург)
  - ~200 из запланированных 3000 геномов
- 2018 г., 2000 экзомов (НИИ ФХМ)
  - ~100 из запланированных 2000 экзомов

Информация о популяционных частотах в РФ отсутствует!

Проблемы интерпретации



Популяционные частоты

Ген	Мутация	*OR*
BRCA1	c.5382insC	197,85
BRCA1	c.T300G	141,94
BRCA1	c.2080delA	258,06
STK11	c.1117C>T p.Pro373Ser	9 образцов, 0 в gnomAD
APC	c.6497G>A p.Arg2166Gln	9 образцов, 0 в gnomAD

Популяционные частоты

Ген	Мутация	*OR*
BRCA1	c.5382insC	197,85
BRCA1	c.T300G	141,94
BRCA1	c.2080delA	258,06
STK11	c.1117C>T p.Pro373Ser	9 образцов, 0 в gnomAD
APC	c.6497G>A p.Arg2166Gln	9 образцов, 0 в gnomAD

Все предикторы предсказывают патогенную мутацию

Все предикторы предсказывают полиморфизм

Популяционные частоты

- На 1000 образцов – 86 мутаций, встречающихся более двух раз в нашей выборке и не встречающихся в 130 тыс. gnomAD
- 52 мутации, встречающиеся более двух раз и имеющие OR > 20 (популяция MAX из gnomAD)

Как интерпретировать?

Если анализировать небольшую выборку

VUS???		
APC	c.6497G>A p.Arg2166Gln	1 пациент

Полиморфизм		
STK11	c.1117C>T p.Pro373Ser	1 пациент

Популяционные частоты

Ген	Мутация	*OR*
STK11	c.1117C>T p.Pro373Ser	9 образцов, 0 в gnomAD
APC	c.6497G>A p.Arg2166Gln	9 образцов, 0 в gnomAD

Обязателен анализ фенотипа!

- Только крупные выборки (больше 500-600 образцов) позволяют сделать обоснованные выводы о патогенности неизвестных мутаций
- Использование только предикторов патогенности приводит как к ложноположительным, так и к ложноотрицательным результатам

Повышение надежности

- Объем выборки образцов не менее 500
- Строгие критерии отбора
- Собственная контрольная группа
- Семейный анамнез
- Справочная информация из баз данных

- Рост размера выборки с 200 до 1000 образцов позволил реклассифицировать 60% вариантов с неизвестной значимостью

Повышение надежности

- Планируемый размер выборки – 5000 образцов из более чем 10 регионов, что будет крупнейшей коллекцией просеквенированных наследственных раков в России

ГАУЗ «Республиканский клинический онкологический диспансер МЗ РТ»

"Казанский (Приволжский) федеральный университет»

ФНКЦ специализированных видов медицинской помощи и медицинских технологий ФМБА России

Спасибо за внимание!