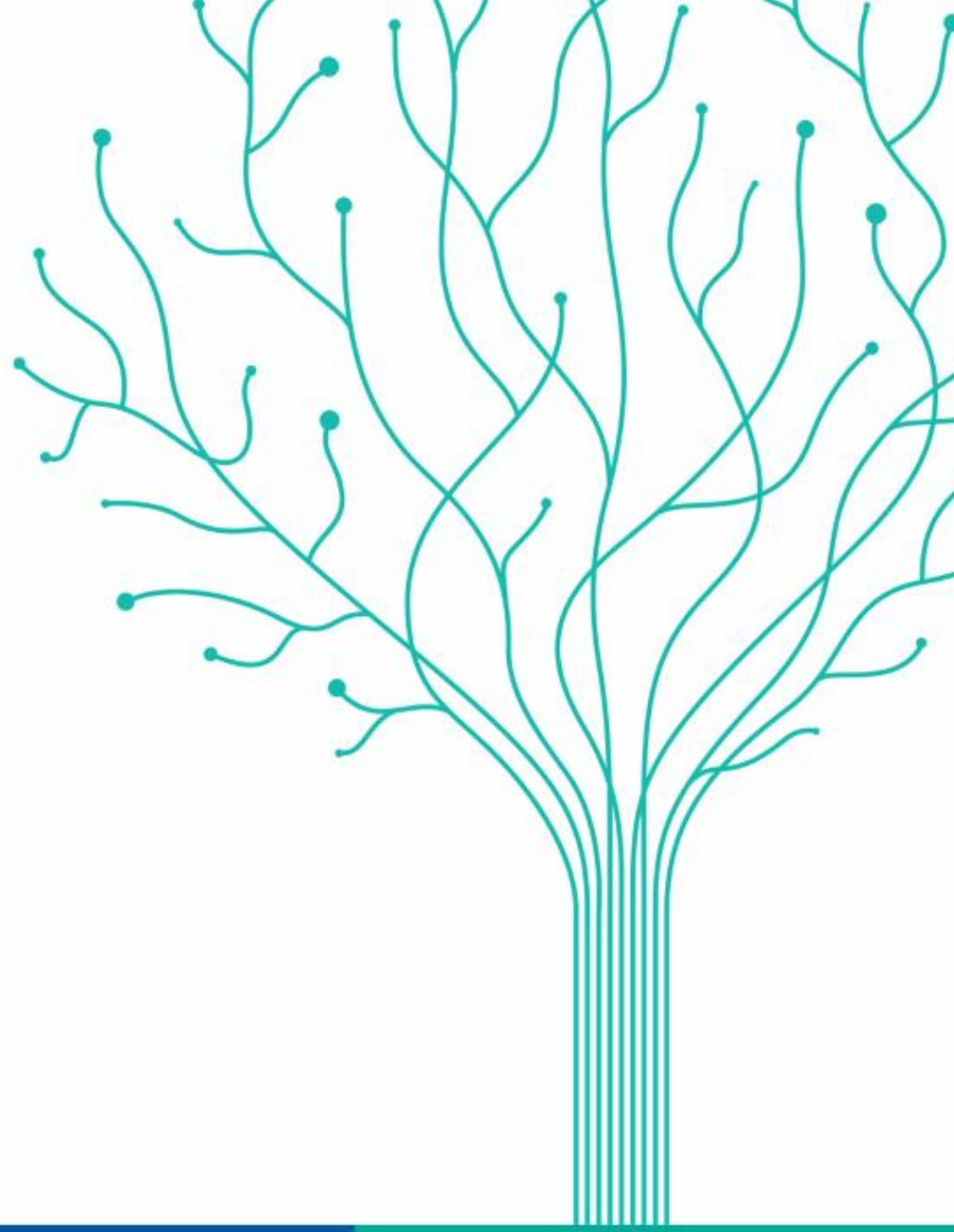
future inside Genetico®

Биоинформатический анализ данных секвенирования нового поколения

Дарья Хмелькова

руководитель направления по онкогенетике

руководитель отдела биоинформатики

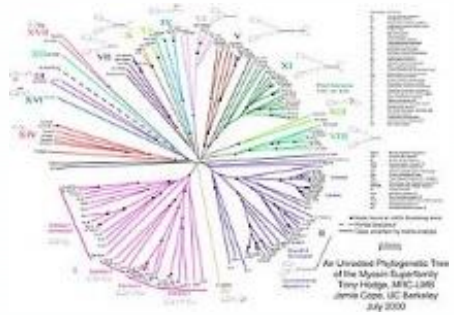


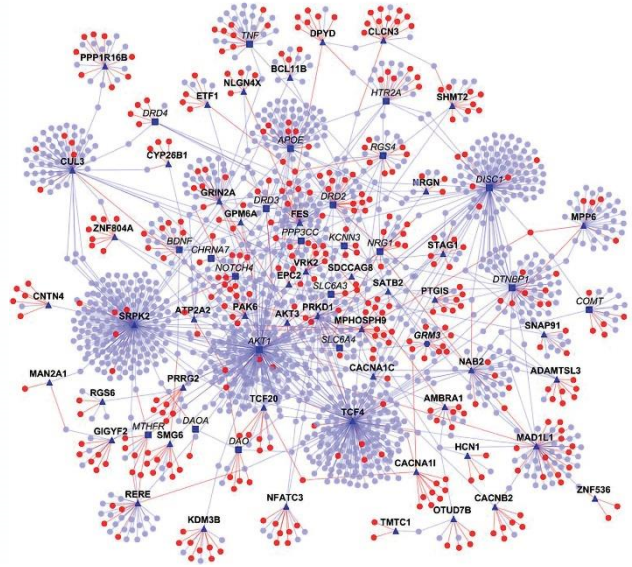
Биоинформатика

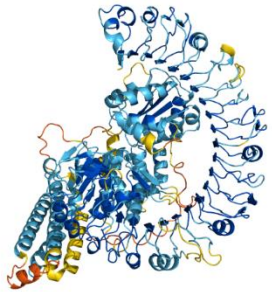
Биоинформатика — междисциплинарная область, позволяющая применять вычислительные методы и подходы к анализу и интерпретации больших объемов биологических данных.

- Получение
- Анализ
- Организация
- Хранение
- Визуализация

данных







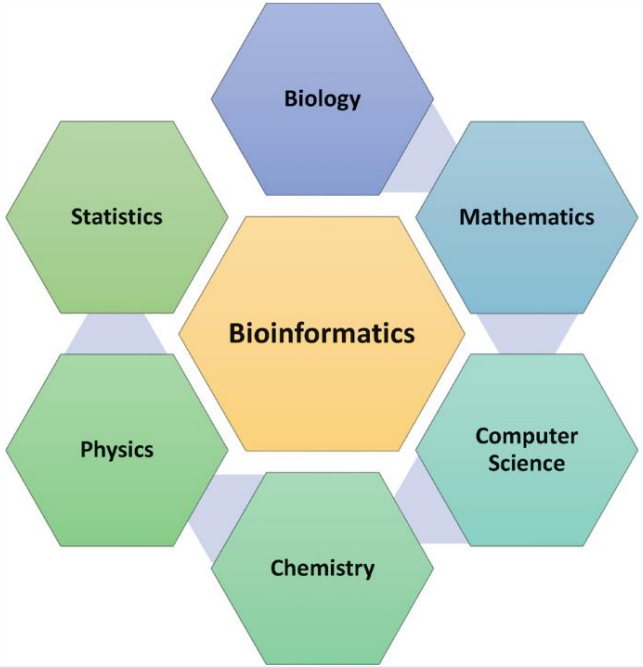
```
A5ASC3.1   14 SIKLWPPSQTTRLLLVERMANNLST..PSIFTRK..YGSLSKEEAARENAKQIEEVACSTANQ.....HYEKEPDGDGGSAVQLYAKECSKLILEVLK 101
B4F917.1   13 SIKLWPPSESTRIMLVDRMTNNLST..ESIFSRK..YRLLGKQEAHENAKTIEELCFALADE.....HFREEPDGDGSSAVQLYAKETSKMMLEVLK 100
A9S1V2.1   23 VFKLWPPSQGTREAVRQKMALKLSS..ACFESQS..FARIELADAQEHARAIEEVAFGAAQE......ADSGGDKTGSAVVMVYAKHASKLMLETLR 109
B9GSN7.1   13 SVKLWPPGQSTRLMLVERMTKNFIT..PSFISRK..YGLLSKEEAEEDAKKIEEVAFAAANQ.....HYEKQPDGDGSSAVQIYAKESSRLMLEVLK 100
Q8H056.1   30 SFSIWPPTQRTRDAVVRRLVDTLGG..DTILCKR..YGAVPAADAEPAARGIEAEAFDAAAA..SGEAAATASVEEGIKALQLYSKEVSRRLLDFVK 120
Q0D4Z3.2   44 SLSIWPPSQRTRDAVVRRLVQTLVA..PSILSQR..YGAVPEAEAGRAAAAVEAEAYAAVTES.SSAAAAPASVEDGIEVLQAYSKEVSRRLLELAK 135
B9MVW8.1   56 SFSIWPPTQRTRDAIISRLIETLST..TSVLSKR..YGTIPKEEASEASRRIEEEAFSGAST.......VASSEKDGLEVLQLYSKEISKRMLETVK 141
Q0IYC5.1   29 SFAVWPPTRRTRDAVVRRLVAVLSGDTTTALRKRYRYGAVPAADAERAARAVEAQAFDAASA....SSSSSSSVEDGIETLQLYSREVSNRLLAFVR 121
A9NW46.1   13 SIKLWPPSESTRLMLVERMTDNLSS..VSFFSRK..YGLLSKEEAAENAKRIEETAFLAAND.....HEAKEPNLDDSSVVQFYAREASKLMLEALK 100
Q9C500.1   57 SLRIWPPTQKTRDAVLNRLIETLST..ESILSKR..YGTLKSDDATTVAKLIEEEAYGVASN.......AVSSDDDGIKILELYSKEISKRMLESVK 142
Q2HRI7.1   25 NYSIWPPKQRTRDAVKNRLIETLST..PSVLTKR..YGTMSADEASAAAIQIEDEAFSVANA.......SSSTSNDNVTILEVYSKEISKRMIETVK 110
Q9M7N3.1   28 SFKIWPPTQRTREAVVRRLVETLTS..QSVLSKR..YGVIPEEDATSAARIIEEEAFSVASV.ASAASTGGRPEDEWIEVLHIYSQEIXQRVVESAK 119
Q9M7N6.1   25 SFSIWPPTQRTRDAVINRLIESLST..PSILSKR..YGTLPQDEASETARLIEEEAFAAAGS.......TASDADDGIEILQVYSKEISKRMIDTVK 110
Q9LE82.1   14 SVKMWPPSKSTRLMLVERMTKNITT..PSIFSRK..YGLLSVEEAEQDAKRIEDLAFATANK.....HFQNEPDGDGTSAVHVYAKESSKLMLDVIK 101
Q9M651.2   13 SIKLWPPSLPTRKALIERITNNFSS..KTIFTEK..YGSLTKDQATENAKRIEDIAFSTANQ.....QFEREPDGDGGSAVQLYAKECSKLILEVLK 100
B9R748.1   48 SLSIWPPTQRTRDAVITRLIETLSS..PSVLSKR..YGTISHDEAESAARRIEDEAFGVANT.......ATSAEDDGLEILQLYSKEISRRMLDTVK 133
```

Биоинформатика: применение

- Генетика и геномика
- Транскриптомика
- Протеомика
- Филогенетика
- Метагеномика
- Метаболомика
- Молекулярное моделирование
- Анализ сигнальных путей в клетке
- Персонализированная медицина
- Разработка лекарственных препаратов
- Генная терапия
- Биотехнология
- Микробиология
- Ветеринария
- Сельское хозяйство

Биоинформатика: история

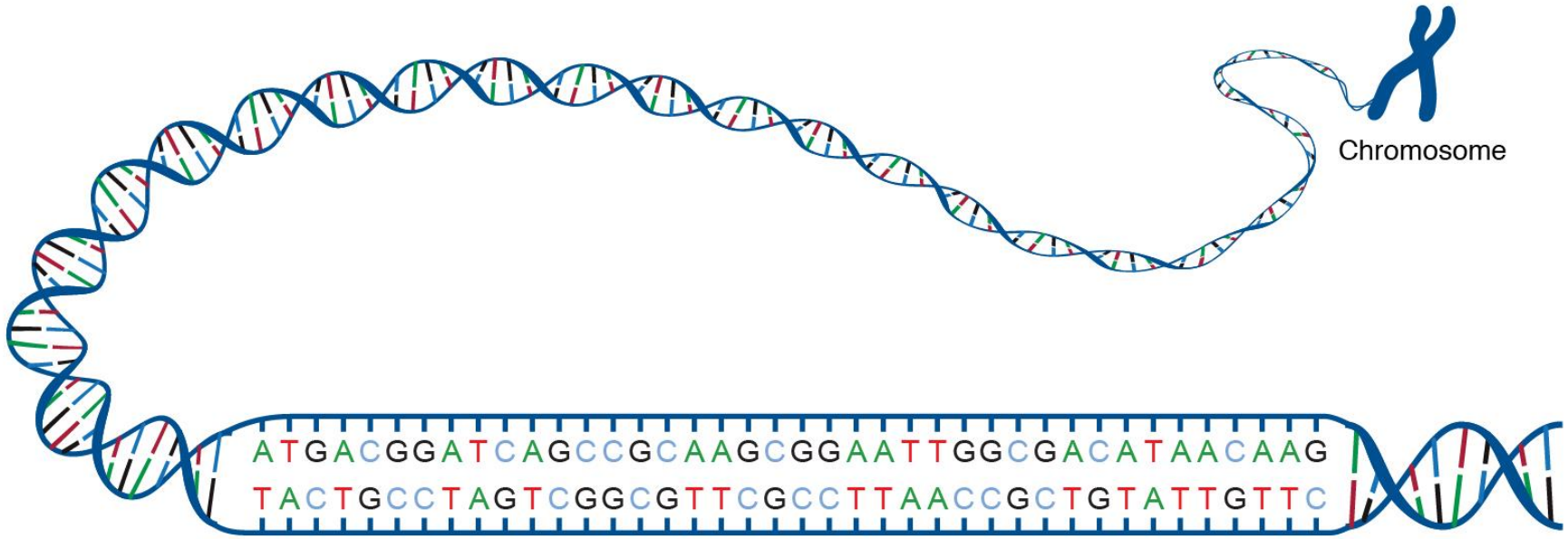
- Термин введен в 1970 г. *(Paulien Hogeweg & Ben Hesper)*
- Толчок к бурному развитию: проект «Геном человека» *(1990-2003)*
- Развитие и удешевление методов секвенирования ДНК -> дальнейший рост



Метод NGS

Секвенирование

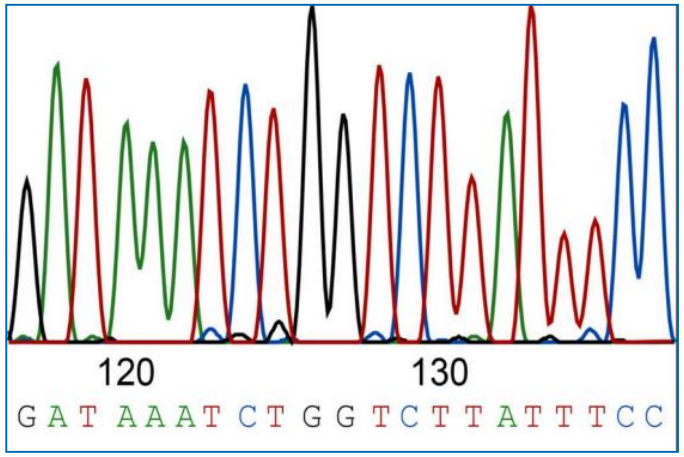
Секвенирование – определение последовательности нуклеотидов ДНК/РНК



https://www.genome.gov/genetics-glossary/acgt

Секвенирование по Сэнгеру

Определение последовательности нуклеотидов в определенном гене или локусе хромосомы.



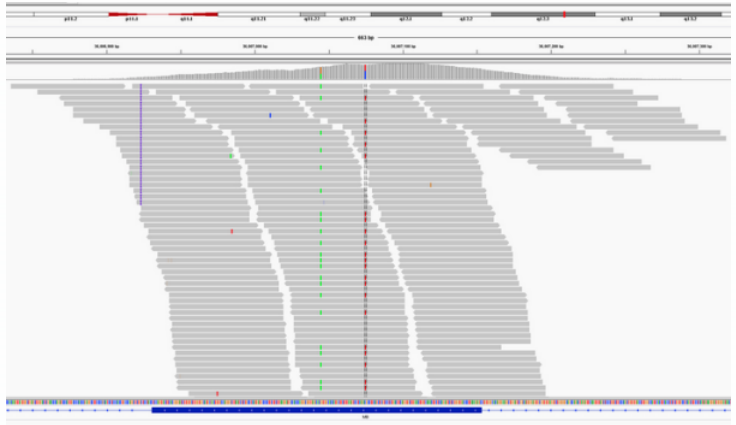
Исследуется последовательность ДНК определенных генов или локусов хромосом в клетке/ткани.

500-1000 п.о.

Высокопроизводительное секвенирование

Секвенирование нового поколения *(next generation sequencing, NGS)*
= *massive parallel sequencing* = *high throughput sequencing*

Определение последовательности нуклеотидов ДНК/РНК в рамках панели генов, полного экзома, генома или транскриптома.



Высокопроизводительное секвенирование

- **Секвенирование второго поколения**:
  - Illumina
  - Ion Torrent
  - BGI
- **Секвенирование третьего поколения**:
  - Pacific Biosciences
  - Oxford Nanopore

Высокопроизводительное секвенирование

- **Секвенирование второго поколения**:
  - Illumina
  - Ion Torrent
  - BGI

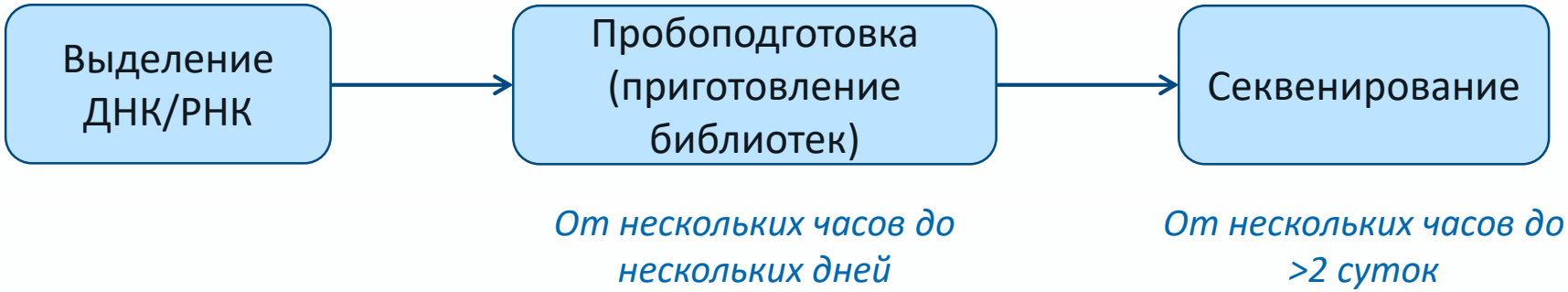
100-300 п.о.

- **Секвенирование третьего поколения**:
  - Pacific Biosciences
  - Oxford Nanopore

10 000-300 000 п.о.

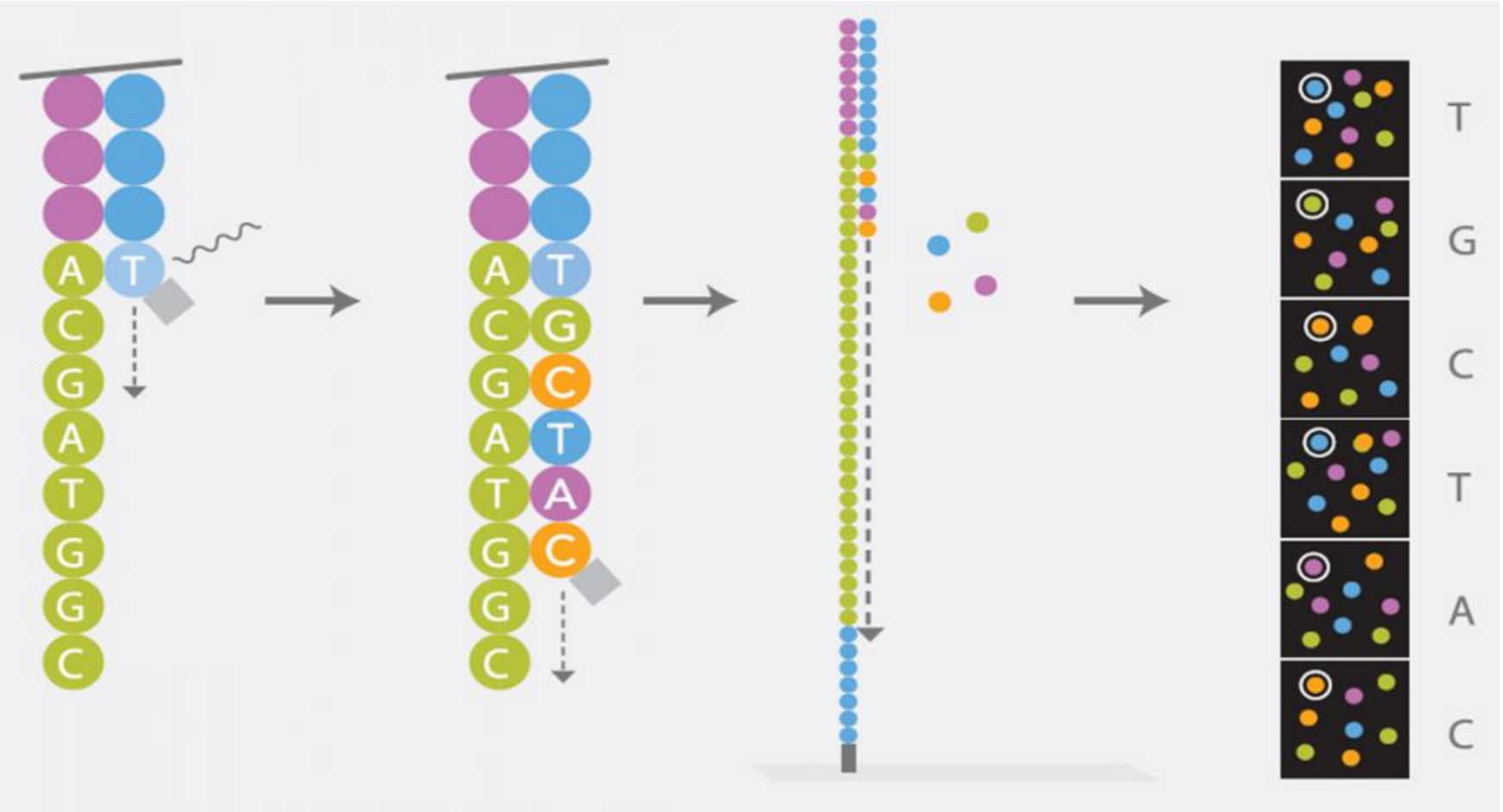
NGS
(этапы лабораторного процесса)*

NGS – next generation sequencing – секвенирование нового поколения



* Illumina

NGS
(секвенирование)



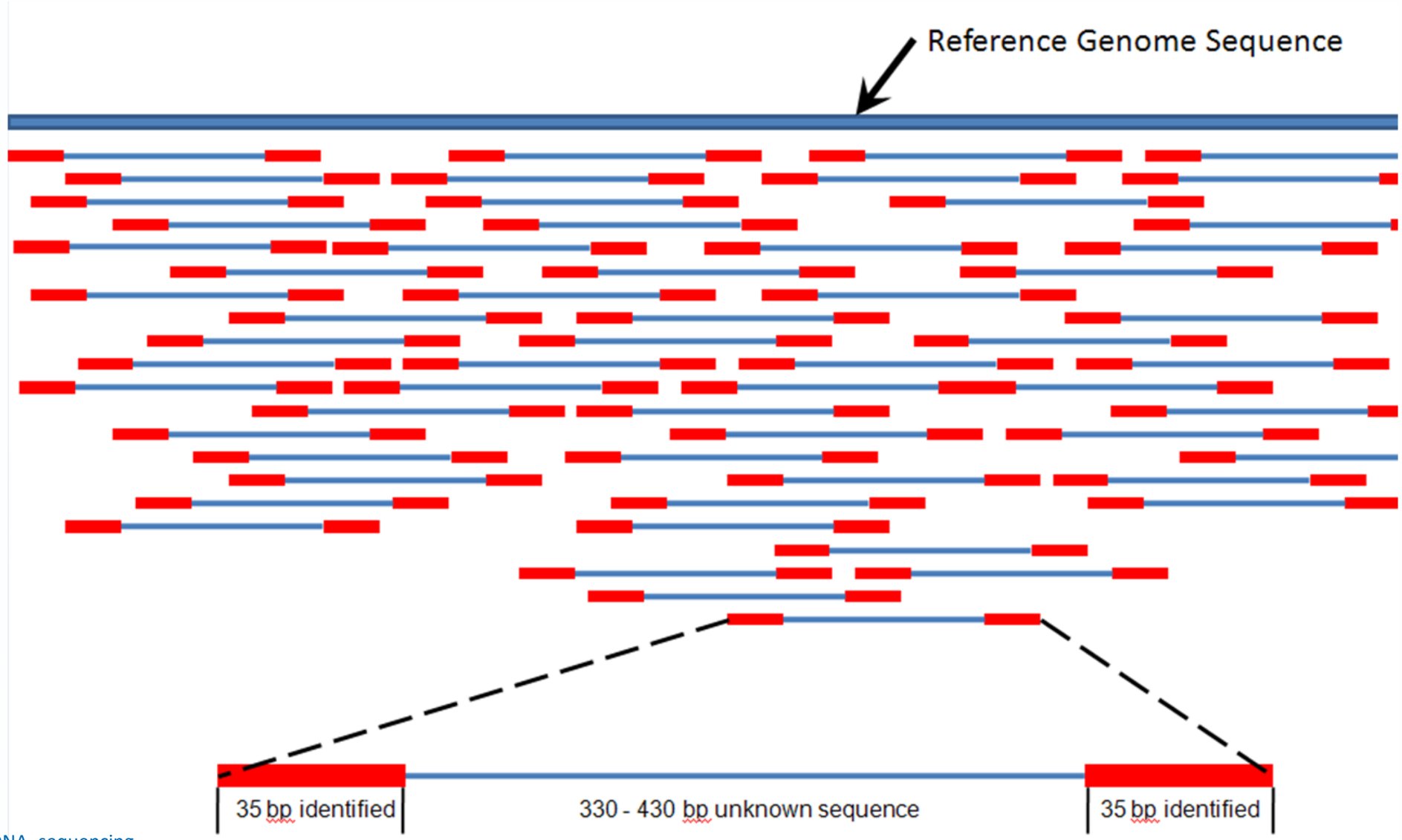
Риды *(reads)* = прочтения – последовательности ДНК, «прочитанные» в секвенаторе

Нуклеотиды:
Аденозин (A)
Цитозин (C)
Гуанин (G)
Тимин (T)

Принцип комплементарности – способность азотистых оснований образовывать связи A-T и C-G

https://www.cegat.de/en/services/next-generation-sequencing/

NGS



https://en.wikipedia.org/wiki/DNA_sequencing

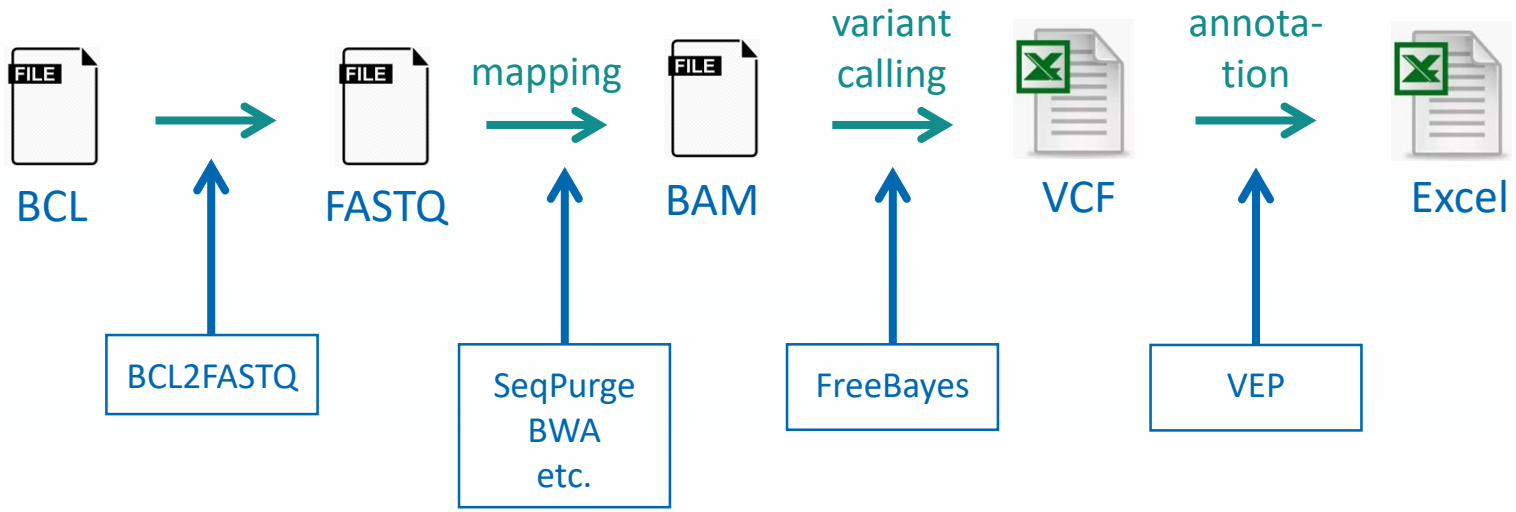
Применение NGS

- **Геномика**
  - Сборка геномов *de novo*
  - Ресеквенирование
- **Транскриптомика (RNA-Seq)**
  - Экспрессия генов
  - Фьюжны
  - Сплайсинг
- **Эпигенетика**
  - Метилирование
  - Изучение хроматина (ATAC-Seq, Hi-C)
  - Анализ модификаций гистонов (ChIP-Seq)
- **Single cell sequencing**

Биоинформатический анализ данных NGS

Общая схема пайплайна

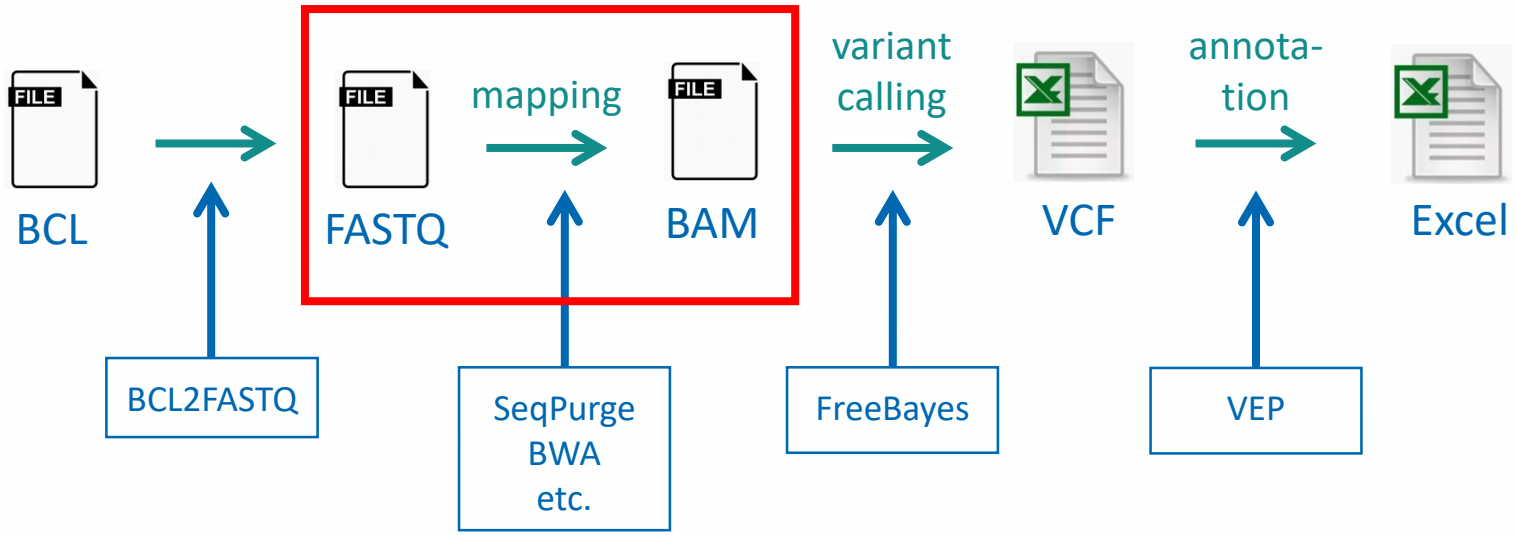
Пайплайн – это цепочка программ, берущих на вход одни файлы, производящих с ними определенные манипуляции и выдающих на выходе другие файлы



SeqPurge – удаление адапторов
BWA – картирование ридов на референс
FreeBayes – variant calling
VEP – аннотация VCF-файла

Общая схема пайплайна

Пайплайн – это цепочка программ, берущих на вход одни файлы, производящих с ними определенные манипуляции и выдающих на выходе другие файлы



SeqPurge – удаление адапторов
BWA – картирование ридов на референс
FreeBayes – variant calling
VEP – аннотация VCF-файла

Характеристика файлов

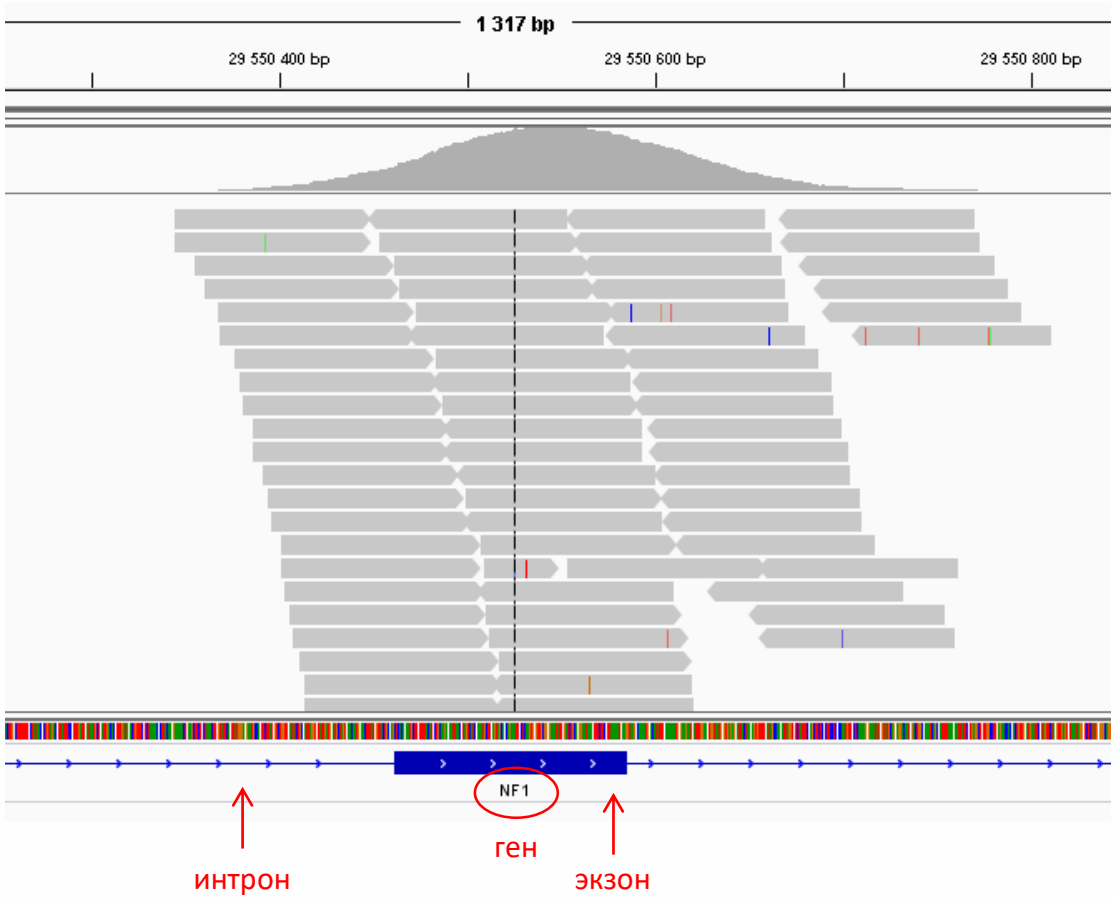
BCL – binary base call. Формат сырых данных, выдаваемых секвенаторами Illumina

FASTQ – текстовый файл, содержащий последовательности коротких фрагментов ДНК (ридов) и соответствующие показатели качества.

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Характеристика файлов

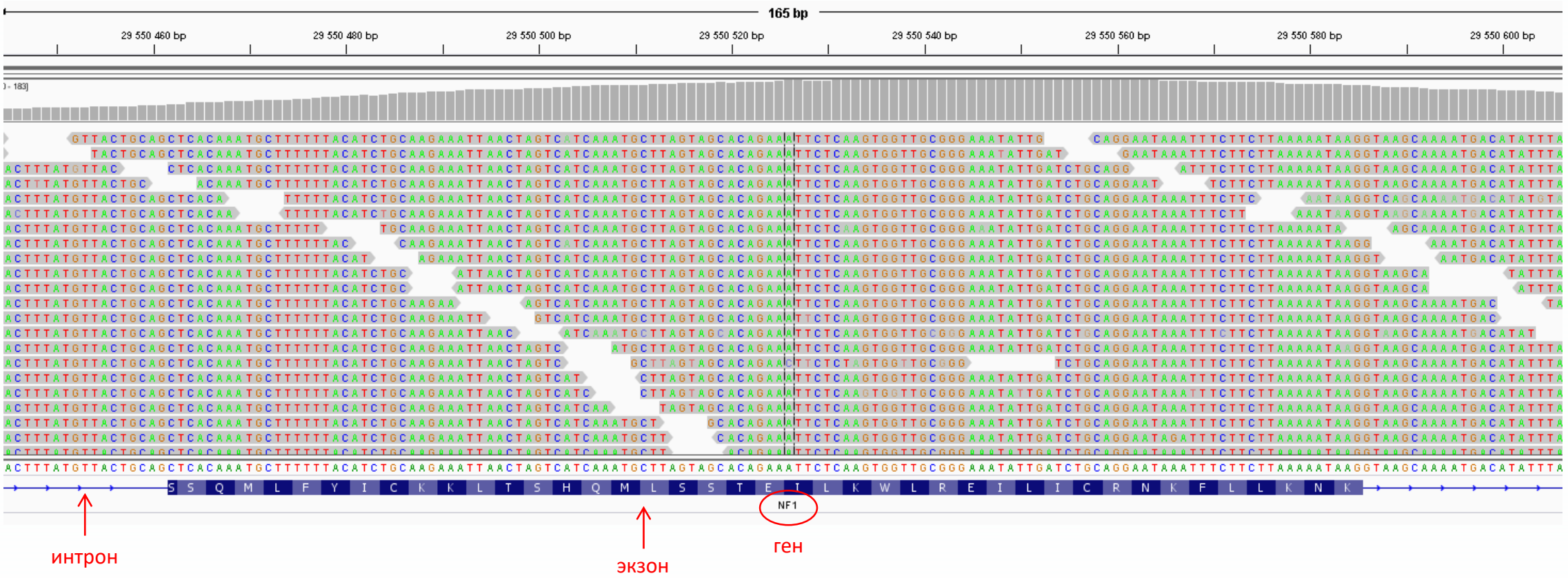
BAM – файл, содержащий риды, картированные на референс



«Маппинг» *(mapping)* = картирование – определение положения рида на геноме

Референсный геном – усредненный геном

Картирование («маппинг») ридов



Биоинформатика: true/false

1) Мою задачу кто-то пробовал решать до меня

Биоинформатика: true/false

1) Мою задачу кто-то пробовал решать до меня

   Огромное количество готовых инструментов, подходов и баз данных

Биоинформатика: true/false

1) Мою задачу кто-то пробовал решать до меня

   Огромное количество готовых инструментов, подходов и баз данных

2) Сейчас найду готовый инструмент и применю

Биоинформатика: true/false

1) Мою задачу кто-то пробовал решать до меня

   Огромное количество готовых инструментов, подходов и баз данных

2) Сейчас найду готовый инструмент и применю

   Только часть из существующих инструментов валидирована и общепризнана

Биоинформатика: true/false

1) Мою задачу кто-то пробовал решать до меня

   Огромное количество готовых инструментов, подходов и баз данных

2) Сейчас найду готовый инструмент и применю

   Только часть из существующих инструментов валидирована и общепризнана

Из 10 биоинформатических программ от силы 5 удастся установить

Биоинформатика: true/false

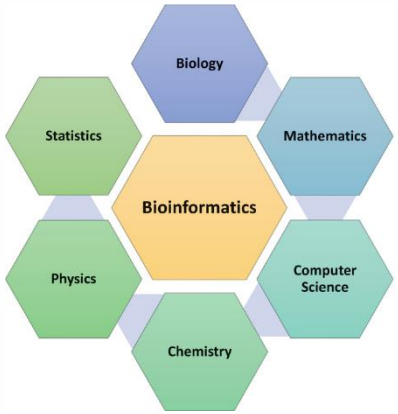
1) Мою задачу кто-то пробовал решать до меня

   Огромное количество готовых инструментов, подходов и баз данных

2) Сейчас найду готовый инструмент и применю

   Только часть из существующих инструментов валидирована и общепризнана

- **Требуются специальные знания и навыки для применения**



Биоинформатика: true/false

1) Мою задачу кто-то пробовал решать до меня

   Огромное количество готовых инструментов, подходов и баз данных

2) Сейчас найду готовый инструмент и применю

   Только часть из существующих инструментов валидирована и общепризнана

- **Требуются специальные знания и навыки для применения**
- **Обучение биоинформатике**

Биоинформатика: true/false

1) Мою задачу кто-то пробовал решать до меня

   Огромное количество готовых инструментов, подходов и баз данных

2) Сейчас найду готовый инструмент и применю

   Только часть из существующих инструментов валидирована и общепризнана

- **Требуются специальные знания и навыки для применения**
- **Обучение биоинформатике**
- **Коммерческие решения**

Биоинформатика: true/false

1) Мою задачу кто-то пробовал решать до меня

   Огромное количество готовых инструментов, подходов и баз данных

2) Сейчас найду готовый инструмент и применю

   Только часть из существующих инструментов валидирована и общепризнана

- **Требуются специальные знания и навыки для применения**
- **Обучение биоинформатике**
- **Коммерческие решения**
- **Услуги по биоинформатике**

future inside Genetico®

Спасибо за внимание!

future inside Genetico®

Спасибо за внимание!

Метод NGS: pros and cons

Преимущества	Недостатки
1) Высокая производительность	1) Сложность анализа данных
2) Высокая чувствительность, специфичность и точность	2) Большие объемы данных
3) Сравнительно невысокие требования к количеству биоматериала	3) Сравнительно высокие требования к качеству ДНК/РНК
4) Возможность решения большого круга задач	4) Срок выполнения
5) Возможность переанализа данных	5) Стоимость

Ограничения метода NGS

1) Неравномерное покрытие
2) Артефакты секвенирования
3) Мутации в областях повторяющихся последовательностей ДНК (экспансии повторов и т.д.)
4) Мутации в областях генов, гомологичных псевдогенам
5) Мутации в GC-богатых участках
6) CNV (вариации числа копий, *copy number variations*)
7) Структурные перестройки (инверсии, транслокации)
8) Мозаицизм