OptiCADD

Путь в светлое будущее

Цель

A general framework for estimating the relative ... - NCBI - NIH

https://www.ncbi.nlm.nih.gov/pubmed/24487276 ▾ Перевести эту страницу

автор: M Kircher - 2014 - Цитируется: 2353 - Похожие статьи

2 февр. 2014 г. - Here we describe Combined Annotation-Dependent Depletion (**CADD**), a method for objectively integrating many diverse annotations into a ...

Цель

Обработка идет в один поток – долго и не задействует все ресурсы машины.

Цель

Обработка идет в один поток – долго и не задействует все ресурсы машины.

Требуется набор аннотаций размером более **100Gb** (оптимально около **200**).

Цель

Обработка идет в один поток – долго и не задействует все ресурсы машины.

Требуется набор аннотаций размером более **100Gb** (оптимально около **200**).

В облачной среде на каждый образец происходит инициализация машины – нужно скачать из S3 все данные аннотаций – долго.

Цель

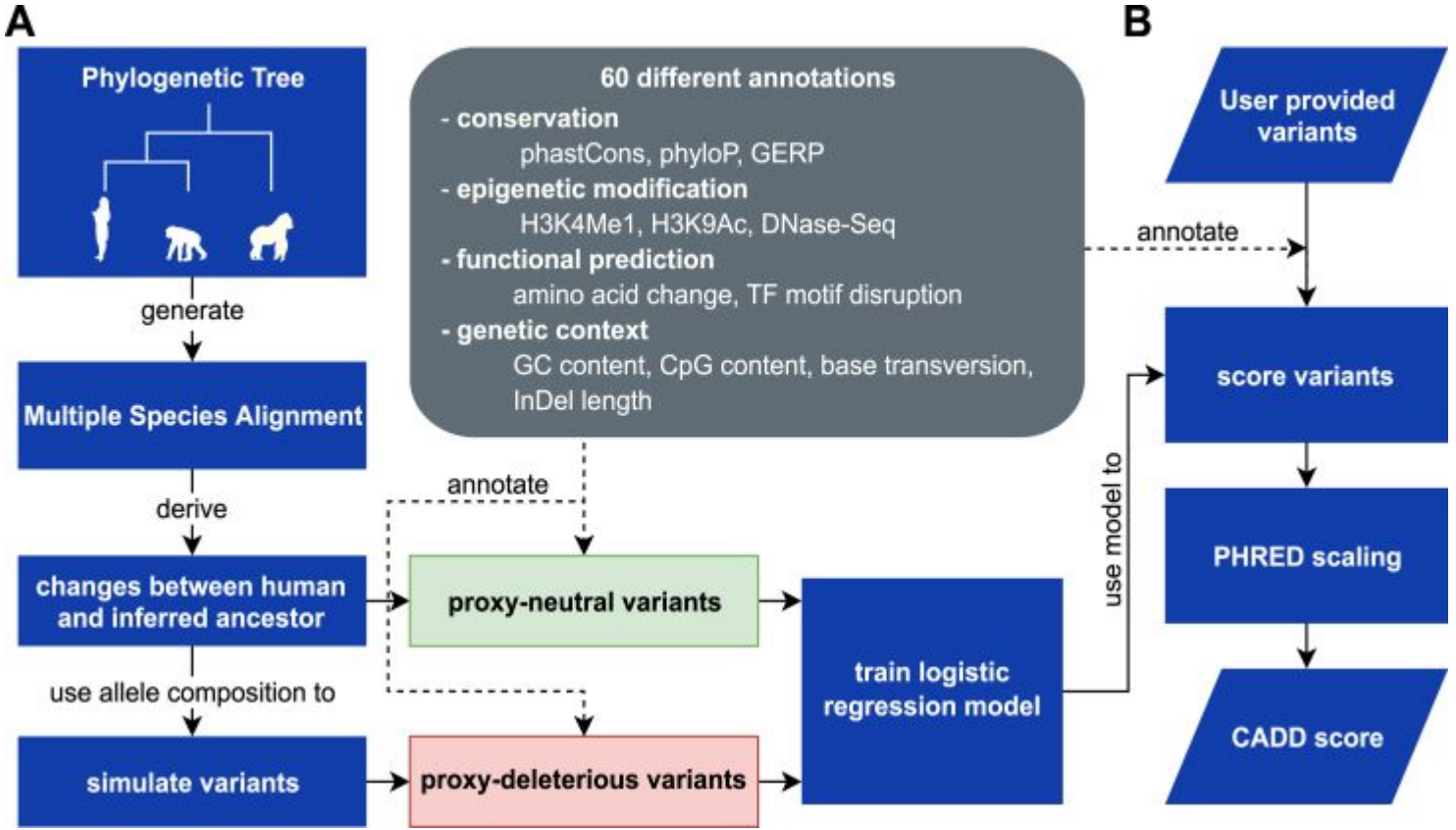
Обработка идет в один поток – долго и не задействует все ресурсы машины.

Требуется набор аннотаций размером более **100Gb** (оптимально около **200**).

В облачной среде на каждый образец происходит инициализация машины – нужно скачать из **S3** все данные аннотаций – долго.

Размещение аннотаций на общей файловой системе **NFS** создает проблемы с производительностью.

Цель


A
Phylogenetic Tree
generate
Multiple Species Alignment
derive
changes between human and inferred ancestor
use allele composition to
simulate variants
60 different annotations
- conservation
phastCons, phyloP, GERP
- epigenetic modification
H3K4Me1, H3K9Ac, DNase-Seq
- functional prediction
amino acid change, TF motif disruption
- genetic context
GC content, CpG content, base transversion, InDel length
annotate
proxy-neutral variants
proxy-deleterious variants
train logistic regression model
use model to
B
User provided variants
annotate
score variants
PHRED scaling
CADD score

Задачи

Разобраться в имеющихся алгоритмах и наборах данных, используемых утилитой CADD для оценки клинической значимости вариаций.

Задачи

Разобраться в имеющихся алгоритмах и наборах данных, используемых утилитой CADD для оценки клинической значимости вариаций.

Придумать, как оптимизировать доступ к данным для того чтобы обеспечить параллельную обработку большого числа образцов в облачной среде.

Задачи

Разобраться в имеющихся алгоритмах и наборах данных, используемых утилитой CADD для оценки клинической значимости вариаций.

Придумать, как оптимизировать доступ к данным для того чтобы обеспечить параллельную обработку большого числа образцов в облачной среде.

- Разбиение кода утилиты на параллельно выполняющиеся потоки.
- Перенос данных в TmpFS.
- Перенос данных из текстовых файлов в базу, способную обслуживать одновременно несколько работающих машин.

Методы оптимизации

BS - базовая версия с одним потоком :(

Методы оптимизации

BS - базовая версия с одним потоком :(

MP - Python-реализация **multiprocessing**

Методы оптимизации

BS - базовая версия с одним потоком :(

MP - Python-реализация **multiprocessing**

SPLIT - **BASH-обертка** для параллельного исполнения:

```
split -n `nproc --all` --additional-suffix=.vcf $filename && \
ls *.vcf | xargs -n 1 -P `nproc --all` ../CADD.sh -g GRCh37 -v v1.4
```

Методы оптимизации

BS - базовая версия с одним потоком :(

MP - Python-реализация **multiprocessing**

SPLIT - **BASH-обертка** для параллельного исполнения

SHM - использование **tmpfs** для хранения **prescored** данных

Методы оптимизации

BS - базовая версия с одним потоком :(

MP - Python-реализация **multiprocessing**

SPLIT - **BASH-обертка** для параллельного исполнения

SHM - использование **tmpfs** для хранения **prescored** данных

Сжатие путем поворота (**pivot**) данных

Методы оптимизации

BS - базовая версия с одним потоком :(

MP - Python-реализация **multiprocessing**

SPLIT - **BASH-обертка** для параллельного исполнения

SHM - использование **tmpfs** для хранения **prescored** данных

Сжатие путем поворота (**pivot**) данных

Вычленение **prescored exome** из **whole genome**

Методы оптимизации

BS - базовая версия с одним потоком :(

MP - Python-реализация **multiprocessing**

SPLIT - **BASH-обертка** для параллельного исполнения

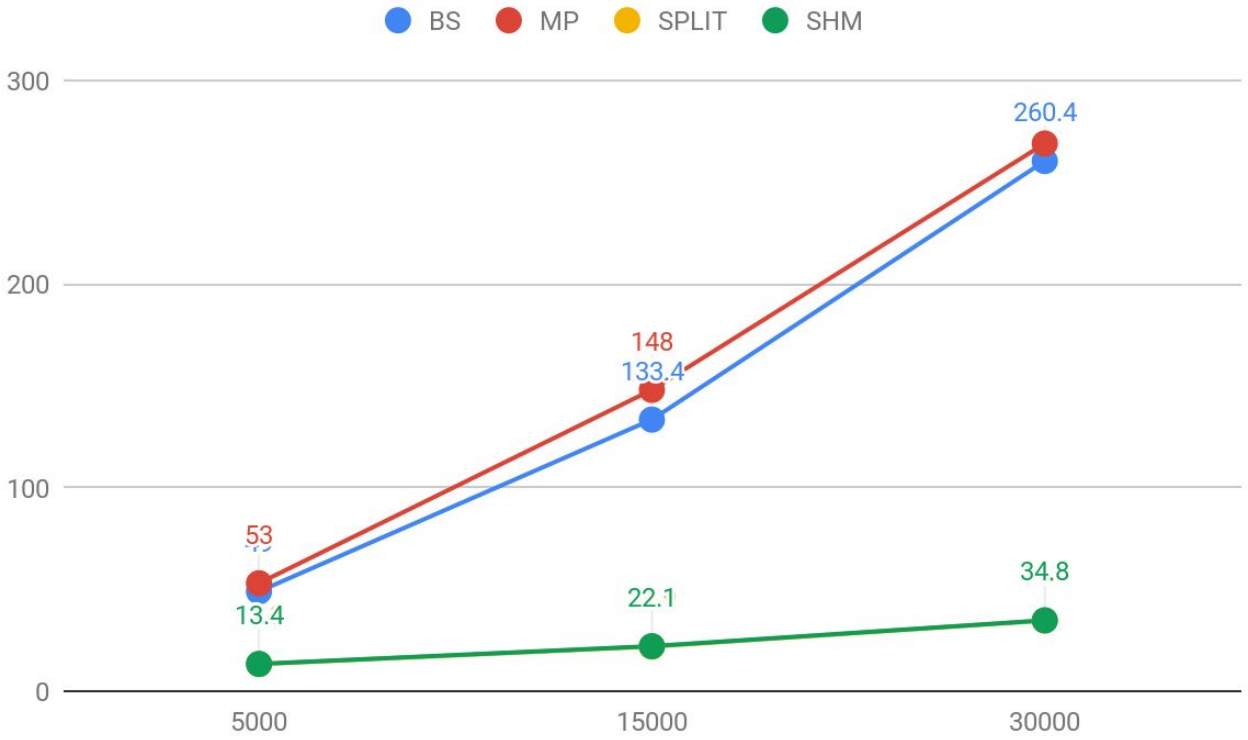
SHM - использование **tmpfs** для хранения **prescored** данных

Сжатие путем поворота (**pivot**) данных

Вычленение **prescored exome** из **whole genome**

Соревнование с **tabix** путем получения предварительной выборки

Методы оптимизации



Аналитическая оптимизация

Внутренняя структура хранения данных в **CADD TSV**:

1 10001 T A 0.118631 4.575

1 10001 T C 0.135541 4.848

1 10001 T G 0.111762 4.462

Преобразуется в:

1 10001 T 0.1,0.1,0.1 4.6,4.8,4.5

Аналитическая оптимизация

Extract **T**ransform **L**oad - центральная парадигма биоинформатики:

```
zcat whole_genome_SNVs.tsv.1.gz \
  | awk '{printf("%d\t%d\t%s\t%s\t%0.2f\t%0.2f\n",$1,$2,$3,$4,$5,$6)}' \
  | bedtools groupby -g 1,2,3 -c 5,6 -o collapse,collapse \
  | bgzip \
 > whole_genome_SNVs.tsv.1.gz.compressed.gz
```

Аналитическая оптимизация

6.2G whole_genome_SNVs.tsv.**chr1**.gz

1.5G whole_genome_SNVs.tsv.**chr1**.gz.**compressed**.gz

6.5G whole_genome_SNVs.tsv.**chr2**.gz

1.6G whole_genome_SNVs.tsv.**chr2**.gz.**compressed**.gz

5.4G whole_genome_SNVs.tsv.**chr3**.gz

1.3G whole_genome_SNVs.tsv.**chr3**.gz.**compressed**.gz

78G whole_genome_SNVs.tsv.gz

19.2G (est) whole_genome_SNVs.tsv.gz.**compressed**.gz (в четыре раза меньше!!!!)

Аналитическая оптимизация

Зачем нам все **prescored** значения, если к нам едут **exomes**?

78G whole_genome_SNVs.tsv.gz

\+ **tabix -R exome.bad $SCORE $INDEX** >

25G exome.tsv.gz (это можно еще ужать предыдущим методом!)

CADDaS

```yaml
openapi: 3.0.0
info:
  title: OptiCADD
  description: Multithread cloud-ready CADD as Service.
  version: 0.1
servers:
  - url: ws://api.cadd/v1
    description: Optional server description, e.g. Main (production) server
paths:
  /score:
    get:
      summary: Returns a CADD score.
      description: Recieving VCF string as input and returns CADD score.
      responses:
        '200':    # status code
          description: A JSON array of CADD scores
          content:
            application/json:
              schema:
                type: array
                items:
                  type: string
```

Итоги

- **BASH-обертка** - оптимальное решение для параллельного выполнения в как на многоядерной системе, так и в HPC среде
- Если только exome, то можно **78G whole genome -> 25G exome**
- **Pivot-сжатие** 1 к 4
- **1M VCF за 32:25 vs. НИКОГДА на оригинальной версии**

GitHub, TODO, etc

TODO:

Websockets
Cassandra
mailto: Rentzsch, Philipp (cadd-support@uw.edu)