Webová aplikácia na analýzu genetických údajov

Projekt pozostáva z webového API vytvoreného pomocou Java Spring Boot, ktoré umožňuje nahrávanie VCF súborov, ich anotáciu a následné ukladanie relevantných informácií do relačnej databázy. **API** funguje nasledovne: Používateľ nahraje originálny VCF súbor, ktorý obsahuje informácie o genetických variantoch. Následne je súbor anotovaný nástrojom **snpEff**, ktorý k jednotlivým variantom pripojí dôležité informácie o ich potenciálnom biologickom vplyve. Po anotácii sa výsledný súbor uloží a následne sa parsuje. API umožňuje filtrovať výsledky podľa konkrétnych atribútov (napríklad pomocou filter), čo výrazne uľahčuje prácu s veľkými objemami dát. V budúcnosti plánujem pridať aj možnosť odosielania dopytov do databázy ClinVar pre konkrétne varianty, ak si to používateľ vyžiada a integráciu s AI.

(na konci je uvedený príklad anotácii a vracania dát v JSON)

Použité technológie a architektúra

Na realizáciu projektu som využil nasledujúce technológie:

- **Java Spring Boot:** Pre vytvorenie robustného a rozšíriteľného webového API.
- **HTSJDK:** Knižnica na parsovanie VCF súborov, ktorá zabezpečuje správne načítanie genetických dát.
- **snpEff:** Nástroj na anotáciu genetických variantov, ktorý priraďuje každému variantu informácie o jeho potenciálnom biologickom význame.
- **JpaRepository:** Uľahčuje prácu s entitami a interakciu s relačnou databázou.
- **PostgreSQL Docker:** Na spustenie testovacej databázy využívam PostgreSQL v Docker kontejnery, čo mi umožňuje rýchle a jednoduché nasadenie a testovanie databázového prostredia.

Architektúra aplikácie je založená na princípe **controller-service-repository**, kde:

- **Controller:** Spracováva prichádzajúce HTTP požiadavky a odosiela odpovede.
- **Service:** Obsahuje obchodnú logiku – od parsovania VCF súborov, cez volanie snpEff až po filtrovanie dát.
- **Repository:** Zabezpečuje komunikáciu s databázou a ukladanie údajov prostredníctvom JpaRepository.

Databázový model

V databáze sú navrhnuté tri tabuľky. V kóde sú explicitne reprezentované dve entity:

- **VariantEntity:** Uchováva základné informácie o genetických variantoch (napríklad chromozóma, pozícia a podobne).
- **AnnotationEntity:** Uchováva anotácie, ktoré boli priradené ku každému variantu.

Tretia tabuľka slúži na uchovávanie zoznamu alternatívnych allel, ktoré sú prepojené s identifikátorom každého variantu. Tento prístup umožňuje efektívne ukladanie a neskoršie spracovanie oboch typov dát – pôvodných informácií z VCF súboru aj výsledných anotácií.

Implementácia a fungovanie API

Po nahratí VCF súboru prebieha spracovanie nasledovne:

1. **Anotácia variantov:** Nástroj snpEff pripojí k jednotlivým genetickým variantom dôležité informácie o ich funkčnom význame.
2. **Ukladanie a parsovanie anotovaného súboru:** Po anotácii sa súbor uloží a následne parsuje, aby sa do databázy uložili údaje z oboch zdrojov – originálne aj anotované.
3. **Sprístupnenie dát cez API:** Výsledné dáta sú prístupné vo formáte JSON, pričom používateľ môže využiť filter na vyhľadávanie konkrétnych variantov.

Príklad vstupného VCF súboru pre anotáciu:

```
##PEDIGREE=<Derived=Patient_01_Somatic,Original=Patient_01_Germline>
#CHROM   POS      ID       REF      ALT      QUAL     FILTER   INFO     FORMAT   Patient_01_Germline
        Patient_01_Somatic
1        69091    .        A        C,G      .        PASS     AF=0.1122        GT       1/0      2/1
1        69849    .        G        A,C      .        PASS     AF=0.1122        GT       1/0      2/1
1        69511    .        A        C,G      .        PASS     AF=0.3580        GT       1/1      2/2
```

Príklad VCF súboru po anotácii:

```
##PEDIGREE=<Derived=Patient_01_Somatic,Original=Patient_01_Germline>
##SnpEffVersion="5.2e (build 2024-10-27 15:58), by Pablo Cingolani"
##SnpEffCmd="SnpEff  GRCh37.75
/var/folders/cf/fyybggx953q388h7zttr09kh0000gn/T/original3918953506663575129.vcf -o vcf "
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation |
Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID | Transcript_BioType | Rank |
HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance |
ERRORS / WARNINGS / INFO' ">
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this
variant. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_gene |
Percent_of_transcripts_affected'">
##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for
this variant. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_gene |
Percent_of_transcripts_affected'">
#CHROM   POS      ID       REF      ALT      QUAL     FILTER   INFO     FORMAT   Patient_01_Germline
        Patient_01_Somatic
1        69091    .        A        C,G      .        PASS
        AF=0.1122;ANN=G|start_lost|HIGH|OR4F5|ENSG00000186092|transcript|ENST00000335137|protein_
coding|1/1|c.1A>G|p.Met1?|1/918|1/918|1/305||,C|initiator_codon_variant|LOW|OR4F5|ENSG00000186092
|transcript|ENST00000335137|protein_coding|1/1|c.1A>C|p.Met1?|1/918|1/918|1/305||;LOF=(OR4F5|ENSG
00000186092|1|1.00)      GT       1/0      2/1
1        69849    .        G        A,C      .        PASS
        AF=0.1122;ANN=A|stop_gained|HIGH|OR4F5|ENSG00000186092|transcript|ENST00000335137|protein
_coding|1/1|c.759G>A|p.Trp253*|759/918|759/918|253/305||,C|missense_variant|MODERATE|OR4F5|ENSG00
000186092|transcript|ENST00000335137|protein_coding|1/1|c.759G>C|p.Trp253Cys|759/918|759/918|253/
305||    GT       1/0      2/1
1        69511    .        A        C,G      .        PASS
        AF=0.3580;ANN=C|missense_variant|MODERATE|OR4F5|ENSG00000186092|transcript|ENST0000033513
7|protein_coding|1/1|c.421A>C|p.Thr141Pro|421/918|421/918|141/305||,G|missense_variant|MODERATE|O
R4F5|ENSG00000186092|transcript|ENST00000335137|protein_coding|1/1|c.421A>G|p.Thr141Ala|421/918|4
21/918|141/305|| GT       1/1      2/2
```

V anotovanej verzii VCF súboru nájdeme navyše hlavičkové informácie generované nástrojom snpEff, ktoré obsahujú verziu a príkaz na spustenie. Dôležitou časťou je rozšírený `INFO` atribút, ktorý obsahuje podrobné funkčné anotácie (identifikované pomocou `ANN`) a informácie o stratách funkcie (`LOF`). Tieto dáta slúžia ako základ pre následné spracovanie a ukladanie do databázy.

Príklad JSON odpovede cez API:

```
[
```

```
{
    "id": 7,
    "chrom": "1",
    "pos": 69091,
    "ref": "A",
    "alts": [
        "C",
        "G"
    ],
    "lof": "(OR4F5|ENSG00000186092|1|1.00)",
    "annotations": [
        {
            "id": 23,
            "alternativeAllele": "G",
            "effect": "start_lost",
            "impact": "HIGH",
            "geneName": "OR4F5"
        },
        {
            "id": 24,
            "alternativeAllele": "C",
            "effect": "initiator_codon_variant",
            "impact": "LOW",
            "geneName": "OR4F5"
        }
    ]
},
{
    "id": 8,
    "chrom": "1",
    "pos": 69849,
    "ref": "G",
    "alts": [
        "A",
        "C"
    ],
    "lof": null,
    "annotations": [
        {
            "id": 25,
            "alternativeAllele": "A",
            "effect": "stop_gained",
            "impact": "HIGH",
            "geneName": "OR4F5"
        },
        {
            "id": 26,
            "alternativeAllele": "C",
            "effect": "missense_variant",
            "impact": "MODERATE",
            "geneName": "OR4F5"
        }
    ]
},
{
    "id": 9,
    "chrom": "1",
    "pos": 69511,
    "ref": "A",
    "alts": [
        "C",
        "G"
    ],
    "lof": null,
    "annotations": [
        {
            "id": 27,
            "alternativeAllele": "C",
            "effect": "missense_variant",
            "impact": "MODERATE",
            "geneName": "OR4F5"
        },
        {
            "id": 28,
            "alternativeAllele": "G",
            "effect": "missense_variant",
            "impact": "MODERATE",
            "geneName": "OR4F5"
        }
    ]
}
```