

I-BIX-DAT

Data Integration and Interaction Networks

Assignment

Dr Tomasz Kurowski t.j.kurowski@cranfield.ac.uk

14 February 2025

www.cranfield.ac.uk



You will implement a REST API service for storing and sharing variant data for eukaryotic genomes.



You will develop:

 A relational database (SQLite) capable of storing variant data.

2. A tool for parsing annotated VCF files to populate the database.



You will develop:

- A REST API server deployable on Node.js and capable of serving the following requests:
- Listing all VCF samples (genomes) loaded into the database.
- Listing the number of variants (SNPs, InDels, or both) contained in each sample, grouped by chromosome.
- Listing genes impacted by high impact variants in a specified chromosomal region of a specific sample.
- Listing samples which contain variants that impact a specified gene.
- Visualizing chromosomes with high impact variants marked & labelled with the names of the genes they impact.



You will be using VCF data with variants contained in resequenced tomato genomes.

Three files (subsets of variants from three different cultivars) which you will need to process are on Canvas. They come from the following publication:

Aflitos et al., Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. Plant J. 2014 Oct;80(1):136-48. doi: 10.1111/tpj.12616.

```
##fileformat=VCFv4.1
##samtoolsVersion=0.1.14 (r933:170)
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=M0,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the site allele frequency of the first ALT allele">
##INFO=<ID=CI95,Number=2,Type=Float,Description="Equal-tail Bayesian credible interval of the site allele frequency at the 95% level">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the nonRef allele frequency in group1 samples being larger (,smaller) than in group2.">
##INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P-value for testing the association between group1 and group2 samples.">
##INFO=<ID=OCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">
##INFO=<ID=RP,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="# high-quality bases">
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled <u>strand bias P-value"></u>
##FORMAT=<ID=PL,Number=-1,Type=Integer,Description="List of Phred-scaled genotype likelihoods, number of values is (#ALT+1)*(#ALT+2)/2">
##SnpEffVersion="3.2 (build 2013-05-23), by Pablo Cingolani"
##SnpEffCmd="SnpEff -no-upstream -no-downstream -ud 0 -csvStats Slyc2.40 /home/assembly/tomato150/reseg/mapped/Heinz/RF 001 SZAXPI008746-45.vcf.gz "
##INFO=<ID=EFF,Number=.,Type=String,Description="Predicted effects for this variant.Format: 'Effect ( Effect Impact | Functional Class | Codon Change | Amino Acid cha
##reference=S lycopersicum chromosomes.2.50.fa
#CHROM POS
                                                                   FILTER INFO
                      ID
                                 REF
                                             ALT
                                                        OUAL
                                                                                          FORMAT RF 001
SL2.50ch01
                      20288
                                                                   222.0 .
                                                                                          DP=25;AF1=1;CI95=1,1;DP4=0,0,15,8;MQ=60;FQ=-96;EFF=INTERGENIC(MODIFIER||||||||1)
                                                                                                                                                                                                                        GT:PL:DP:GO
                                                                                                                                                                                                                                               1/1:255,6
                                                        T
SL2.50ch01
                                                                                                                 INDEL;DP=17;AF1=1;CI95=1,1;DP4=0,0,5,9;MQ=58;FQ=-76.5;EFF=INTERGENIC(MODIFIER||||||
                      57154
                                            ATTTTTTTTT
                                                                   ATTTTTTTTTT
                                                                                          40.5
                                                                                                                                                                                                                                                           G
SL2.50ch01
                                                                                                                 INDEL;DP=23;AF1=1;CI95=1,1;DP4=0,0,11,8;MQ=60;FQ=-91.5;EFF=INTERGENIC(MODIFIER||||
                      79378
                                             ATTTTTTTTTT
                                                                   ATTTTTTTTTTT 117.0
                                                                                                                 INDEL;DP=21;AF1=1;CI95=1,1;DP4=0,0,7,10;MQ=59;FQ=-85.5;EFF=INTERGENIC(MODIFIER|||||||
SL2.50ch01
                      107776 .
                                             GAAAAAAAA
                                                                   GAAAAAAAAA
                                                                                          55.5
                                                                                                                                                                                                                                                           G
                     118005 .
                                            CAAAAAAAAA
                                                                   CAAAAAAAAA
                                                                                                                 INDEL;DP=31;AF1=1;CI95=1,1;DP4=0,0,16,9;MQ=42;FQ=-110;EFF=INTERGENIC(MODIFIER|||||||
SL2.50ch01
                                                                                          46.5
SL2.50ch01
                      146150 .
                                             TAAAAAAAAA
                                                                    TAAAAAAAAA
                                                                                          55.5
                                                                                                                 INDEL;DP=41;AF1=1;CI95=1,1;DP4=0,0,15,19;MQ=59;FQ=-137;EFF=INTERGENIC(MODIFIER|||||
                     169786 .
                                            GAAAAAAAAA
                                                                   GAAAAAAAAAA
                                                                                          60.5
                                                                                                                 <u>INDEL;DP=23;AF1=1;CI95</u>=1,1;DP4=0,0,8,12;MQ=60;FQ=-94.5;EFF=INTERGENIC(MODIFIER|||||||||1)</u>
SL2.50ch01
                                                                                          DP=18;AF1=1;CI95=1,1;DP4=0,0,17,0;MQ=60;FQ=-78;EFF=INTERGENIC(MODIFIER|||||||||1)
                                                                                                                                                                                                                                               1/1:208,5
SL2.50ch01
                      226278 .
                                                                   175.0 .
                                                                                                                                                                                                                        GT:PL:DP:GO
                                                        Α
                                                                                          INDEL;DP=32;AF1=0.5;CI95=0.5,0.5;DP4=5,2,6,5;MQ=57;FQ=63.5;PV4=0.64,1,1,1;EFF=INTERGENIC(MODIFIER||||||||1)
SL2.50ch01
                      265708 .
                                                        AAAAC
                                                                   183.0 .
                                                                                                                 INDEL;DP=14;AF1=1;CI95=1,1;DP4=0,0,5,4;MQ=52;FQ=-61.5;EFF=INTRON(MODIFIER|||||Solyc01g005390.2|||
SL2.50ch01
                      276241 .
                                            ATTTTTT
                                                                   ATTTTTTTT
                                                                                          46.5
SL2.50ch01
                                                                                          DP=31;AF1=1;CI95=1,1;DP4=0,0,26,4;MQ=60;FQ=-117;EFF=INTERGENIC(MODIFIER|||||||||1)
                                                                                                                                                                                                                        GT:PL:DP:GO
                      294797 .
                                                                   222.0 .
                                                                                                                                                                                                                                               1/1:255,9
SL2.50ch01
                      295720 .
                                                        G
                                                                    222.0 .
                                                                                          DP=32;AF1=1;CI95=1,1;DP4=0,0,13,18;MQ=58;FQ=-120;EFF=INTERGENIC(MODIFIER||||||||1)
                                                                                                                                                                                                                        GT:PL:DP:GQ
                                                                                                                                                                                                                                               1/1:255,9
SL2.50ch01
                      296620 .
                                                                    222.0
                                                                                          DP=28;AF1=1;CI95=1,1;DP4=0,0,14,11;MQ=51;FQ=-102;EFF=INTERGENIC(MODIFIER||||||||1)
                                                                                                                                                                                                                        GT:PL:DP:GQ
                                                                                                                                                                                                                                               1/1:255,7
                                                        G
                                                                                                                            INDEL;DP=6;AF1=1;CI95=0.5,1;DP4=0,0,0,3;MQ=29;FQ=-43.5;EFF=INTERGENIC(MODIFIER|||||||||||
                                                                               CATATATAT
SL2.50ch01
                      297210 .
                                            CATATATATATATATAT
                                                                                                      32.4 .
```



The variants have been annotated using SnpEff. Note that these are the older "classic" SnpEff annotations, which use an "EFF" field instead of an "ANN" field. See the documentation:

https://pcingola.github.io/SnpEff/snpsift/filter/

Example:

SL2.50ch09 71507830 . ATTTTTTTT ATTTTTTTT 62.5 . INDEL;DP=30;AF1=1;CI95=1,1;DP4=0,0,14,15;MQ=60;FQ:Solyc09g092400.1.1|4|1),SPLICE_SITE_ACCEPTOR(HIGH|||||Solyc09g092400.1|||mRNA:Solyc09g092400.1.1|5|1) GT:PL:DP:GQ 1/1:103,87,0:29:99



Homozygous (1/1) **HIGH** impact InDel variant affecting the Solyc09g092400.1 gene



You will need to design and implement a schema (i.e., write an SQL script which will create your tables) capable of storing all the data that you need in an appropriate form. An "appropriate" schema should make it easy to execute the queries you need – don't make it too complicated!

The SQL database schema creation script is part of the deliverables.

An ER diagram for the schema would be useful for visualising it in your technical document.



Parsing inputs and populating the database

You will need to implement a data import tool (using the programming language of your choice) which will populate the database with the data from the VCF files.

The data import tool must be included among the deliverables.

The database file must also be included, and it must contain the full data from all three datasets. If you decide to leave any data out, justify it in the documentation.



You will implement a REST API server which can be deployed on localhost and accessed via HTTP (use curl for testing!).

The NPM package with your code must be included in the deliverables.

Do not include the node_modules directory in your submission – any dependencies should listed be in package.json.



You are responsible for designing reasonable API endpoints for each of the possible requests, as well as for the formatting of the output.

The outputs shown in the following slides are <u>examples</u>. If you decide a different format is more appropriate, use it. Make sure you justify your design choices in the documentation.

Each of the API endpoints should be documented.



Listing all the VCF datasets (genomes) loaded into the database

Example output:

```
[
"RF_001",
"RF_041",
"RF_090"
```

Listing the number of variants (SNPs, InDels, or both) contained in each genome, grouped by chromosome

Example output: number of InDels for RF_001:

```
"SL2.50ch01": 1152,
"SL2.50ch02": 483,
"SL2.50ch03": 729,
"SL2.50ch04": 734,
"SL2.50ch05": 508,
"SL2.50ch06": 666,
"SL2.50ch07": 557,
"SL2.50ch08": 427,
"SL2.50ch09": 602,
"SL2.50ch10": 479,
"SL2.50ch11": 645,
"SL2.50ch12": 799
```



Listing genes impacted by moderate / high impact variants in a specified chromosomal region of a specific sample

Example output: genes affected by high impact variants within the first 20 megabases of chromosome 3 in RF_041:



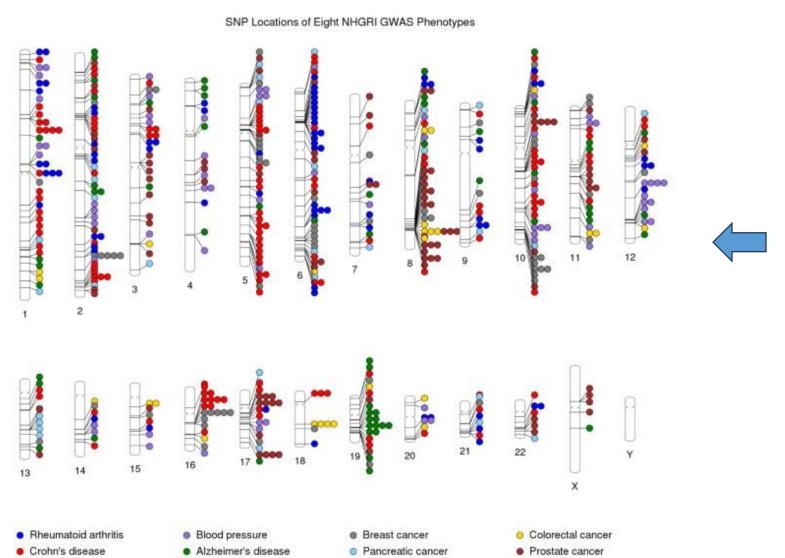
Listing samples which contain variants that impact a specified gene

Example output: samples which contain a variant in the Solyc01g056370.2 gene:

```
[
"RF_041",
"RF_090"
]
```



Visualizing chromosomes with high impact variants marked & labelled with the names of the genes they impact



Example of a (human!) chromosome visualisation.



You can download the full files, as well as others from the same dataset, at the Sol Genomics Network website:

- https://solgenomics.sgn.cornell.edu/organism/Solanum_lycopersicum/tomato_150
- https://solgenomics.net/ftp/genomes/tomato_150/150_VCFs_2.50/

They could prove useful for stress-testing your application and developing extra features.



Marking scheme

Specification	Marks
Software (70%)	
Functional SQLite database (database creation SQL script must be delivered).	10
Tool for processing VCF files exists and data from the three data sets from Canvas were placed in the database	10
REST API can list datasets	5
REST API can report the numbers of variants	10
REST API can list genes affected by moderate / high impact variants	10
REST API can list samples with variants in a specific gene	5
REST API can visualise chromosomes with high impact variants & impacted gene labels	5
Clean, well-structured and commented code	5
Extra functionalities	5
Stability	5
Documentation (30%)	
Technical documentation - relevance, conciseness, accuracy	15
User manual	10
Flowcharts, diagrams, ER diagram of database	5



Deadline: 22 February 2025 at 23:59 (Full-time)

8 March 2025 at 23:59 (Part-time)

Deliverables:

- Archive containing the REST API server project.
- Archive containing the parser tool (if separate from REST API).
- SQLite database file.
- SQL script for database creation.
- Documentation: Technical document and user manual.