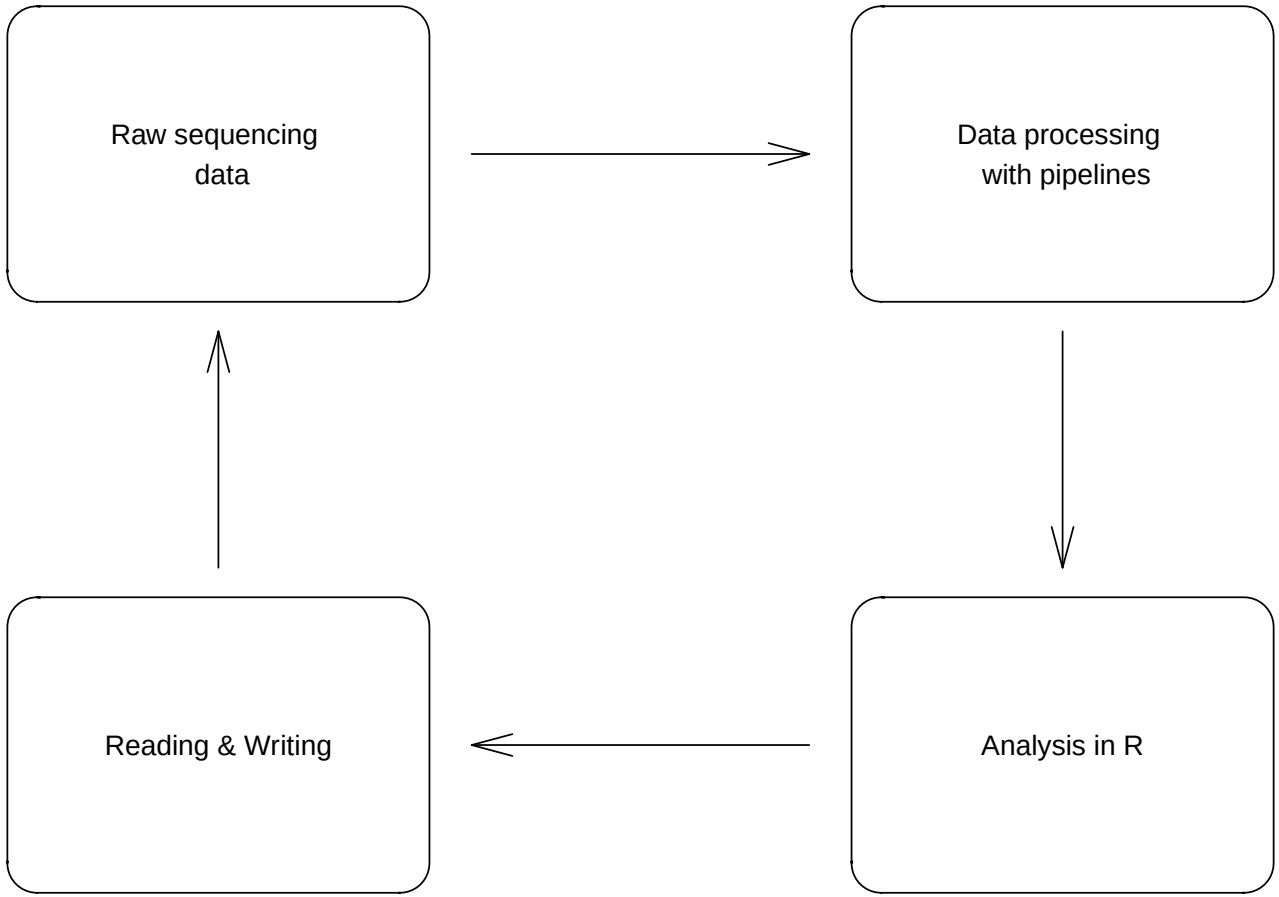
Exploring genome variation using GEMINI framework in R

Sergey Naumenko, CCM Bioinformatics Training series

November 30, 2016

‘Lazy bioinformatics’ approach



Only possible for simple analyses

- Variants in whole exome sequencing
- Diff expression and variants in RNA-seq

What is database?

- Database is a collection of related dataframes (tables);
- GEMINI uses SQLite database which does not require database server or database engine;
- database slang:
  - dataframe = table
  - column = field
  - select * from [table_name] where [condition];
- Web Interface to plot database schemas: `http://ondras.zarovi.cz/sql/demo/?keyword=default`

Gemini database schema

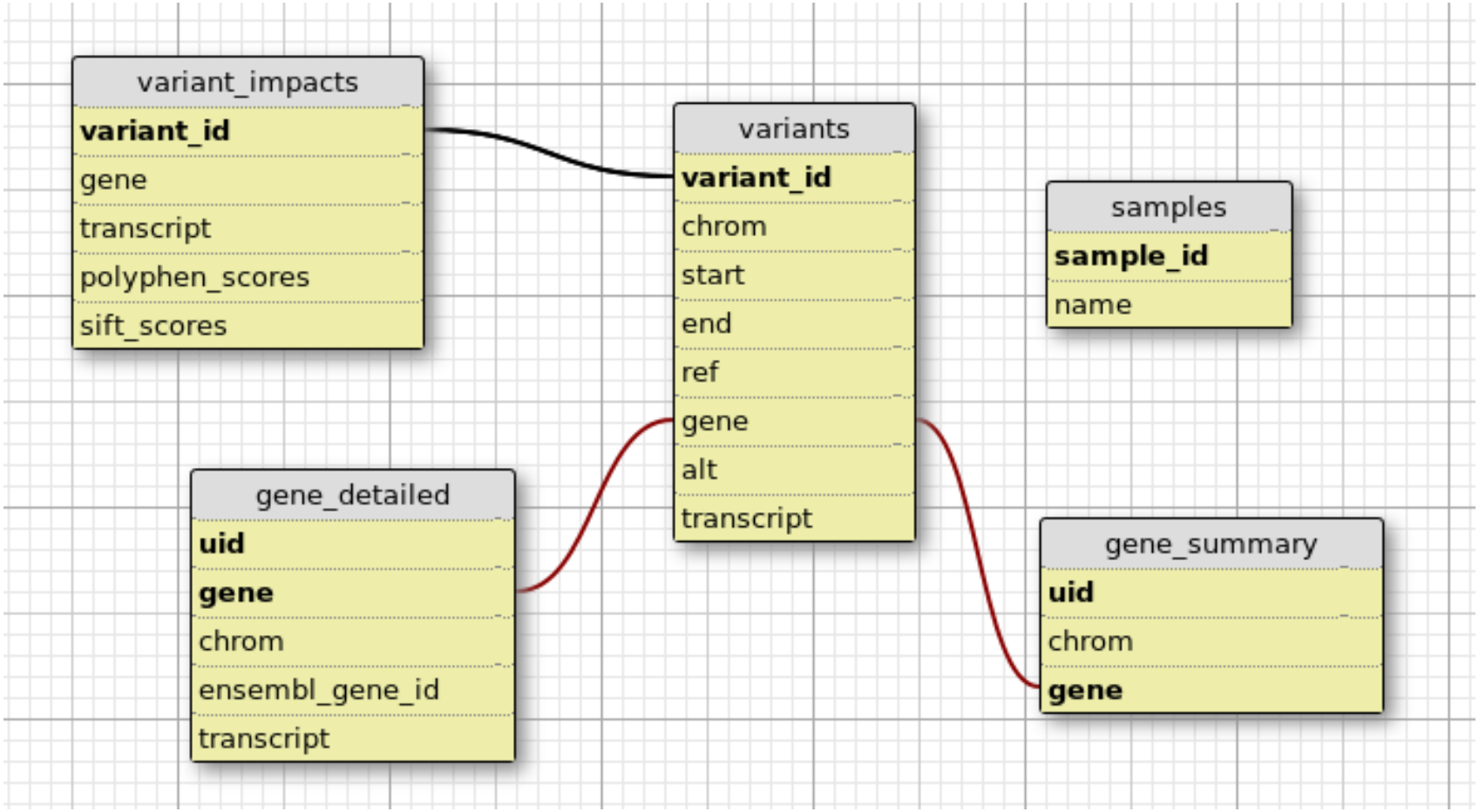


Figure 1: `http://gemini.readthedocs.io/en/latest/content/database_schema.html`

How to load your variants into the GEMINI database?

- Ask your bioinformatics provider to send results in the GEMINI format;
- Ask CCM to convert vcf2gemini;
- Convert
  - gemini load (linux);
  - vcf2db.py
- Use bcbio pipeline system – it outputs GEMINI and vcf.

Bcbio pipelines

- open source, community supported
- developed mainly by Harvard T.Chan Medical School Bioinformatics Core
- validated
- well documented
- Pipelines

c..variant.calling. . . .cancer. . . .structural.variant.calling. . .

variant calling
cancer
structural variant calling
RNA-seq
smallRNA-seq
ChIP-seq
- Resources +
`https://bcbio-nextgen.readthedocs.io/en/latest/` +
`https://github.com/chapmanb/bcbio-nextgen` +
`https://bcb.io`

Example queries

select gene,transcript from gene_detailed where gene=‘AGRN’;
select variant_id,gene,transcript from variant_impacts where variant_id=2;

Additional columns not present in GEMINI

- UCSC hyperlink
- OMIM gene description (registration required)
- Orphanet status
- exac pLi score
- exec missense score
- phastcons score
- imprinting status
- information from HGMD database (public version is unusable, pro version is pricy)
- pseudoatosomal gene

Slide with R Output

```
kable(summary(cars))
```

speed	dist
Min. : 4.0	Min. : 2.00
1st Qu.:12.0	1st Qu.: 26.00
Median :15.0	Median : 36.00
Mean :15.4	Mean : 42.98
3rd Qu.:19.0	3rd Qu.: 56.00
Max. :25.0	Max. :120.00

Slide with Plot

```
plot(pressure)
```