LEIDEN UNIVERSITY MEDICAL CENTER

# Mutalyzer 2.0

**Jeroen F. J. Laros**
**Department of Human Genetics**
**Center for Human and Clinical Genetics**

# Introduction

A curational tool for *Locus Specific Mutation Databases* (LSDBs).

- Variant nomenclature checker applying *Human Genome Variation Society* (HGVS) guidelines.
- Position conversion.
  - Genomic to transcript coordinates and vice versa.
  - World Wide Querying Service in LOVD 2.0.
  - Descriptions on multiple transcripts in LOVD 3.0.
- Gives information about affected proteins.
- ...

## HGVS nomenclature

Genomic orientated:

`AL449423.14:g.[65449_65463del;65564T>C]`

Transcript orientated:

`AL449423.14(CDKN2A_v001):c.[5A>G;106_120del]`

- `AL449423.14` – reference sequence.
- `CDKN2A_v001` – transcript variant 1 of gene CDKN2A.
- `c.[5A>G;106_120del]`
  - A *substitution* at position 5 counting from the start codon.
  - A *deletion* from position 106 to position 120.

# Name checker: disambiguation

We observe a change from `ggatcatcg` to `ggatcatcatcg`.

```
123456789
ggatcatcg
  ↑   ↑   ↑
```

Which can be described as an insertion of `atc` at three places:

- `g.2_3insatc`
- `g.5_6insatc`
- `g.8_9insatc`

# Name checker: disambiguation

We observe a change from `ggatcatcg` to `ggatcatcatcg`.

```
123456789
ggatcatcg
   ↑  ↑
```

. . . or an insertion of `tca` at two places:

- `g.3_4instca`
- `g.6_7instca`

# Name checker: disambiguation

We observe a change from `ggatcatcg` to `ggatcatcatcg`.

```
123456789
ggatcatcg
    ↑   ↑
```

. . . or an insertion of `cat` at two places:

- `g.4_5ins`**`cat`**
- `g.7_8ins`**`cat`**

## Name checker: disambiguation

We observe a change from `ggatcatcg` to `ggatcatcatcg`.

```
123456789
ggatcatcg
        ↑
```

The only correct one is the one on the 5' end.

- `g.8_9insatc`

# Name checker: disambiguation

We observe a change from gctccggccagg to gctggccggagg.

```
                111
123456789012
gctccggccagg
```

We can describe it as follows:

- g.2_11inv
- g.4_9delinsGGCCGG

```
                111
123456789012
gctccggccagg
```

But the correct way is:

- g.4_9inv

## Name checker: correction

Other pitfalls:

- A deletion-insertion can actually be:
  - An inversion `2_3delACinsGT` ⇒ `2_3inv`
  - An insertion `2delTinsTAA` ⇒ `2_3insAA`
  - A substitution `2delAinsT` ⇒ `2A>T`
  - A deletion `2_3delTAinsA` ⇒ `2del`
- An inversion can actually be a substitution.
- An insertion can actually be a duplication.
- A variant can have no effect (`2_5invACGT`, `2A>A`, etc.).

## Name checker: workflow

`NM_002001.2:c.[12_14del;102G>T]`

1. Parse the variant description (previous presentation).
   - Reference sequence e.g., `NM_002001.2`.
   - Coordinate system (`c.`, `g.`, `n.`, ...).
   - List of variants (`12_14del`, `102G>T`).
2. Download the reference sequence.
3. Check the variants to the reference sequence.
   - Is there an `G` at position `c.102`?
4. Mutate the reference sequence.
5. Predict the variant protein when applicable.
6. ...

## Name checker

After a description is checked, other useful information is returned.

- Overview of the change on DNA level.
- A genomic description.
- A description on all affected transcripts.
- Description of affected proteins.
- Visualisation of the original and affected protein with changes highlighted.
- Exon and CDS start / stop information.
- Effects on restriction sites.

# Name checker: demo

AL449423.14(CDKN2A_v001):c.247_250delinsCTTT

http://www.mutalyzer.nl/2.0/check

## Position converter

In Next Generation Sequencing we often encounter chromosomal coordinates.

LSDB's usually use transcripts.

The position converter:

- Works on both hg18 (NCBI Build 36.1) and hg19 (GRCh37).
- Works in both ways:
  - `NM_003002.2:c.274G>T` to `NC_000011.9:g.111959695G>T`.
  - `chr11:g.111959695G>T` to `NM_003002.2:c.274G>T`.
- Can be used to *lift over* from hg18 to hg19 and vice versa.

## Other functionalities

Other functionalities of Mutalyzer 2.0 include:

- Syntax checking.
- SNP conversion (from dbSNP rsId to HGVS notation).
- Name generator (to help people that don't use the HGVS notation that often).
- GenBank uploader (to make your own reference sequences).
  - Automatically uses the correct strand when a HGNC gene symbol is used.
- Recently added functionality for the LRG (Locus Reference Genomic) reference files.

## Other functionalities

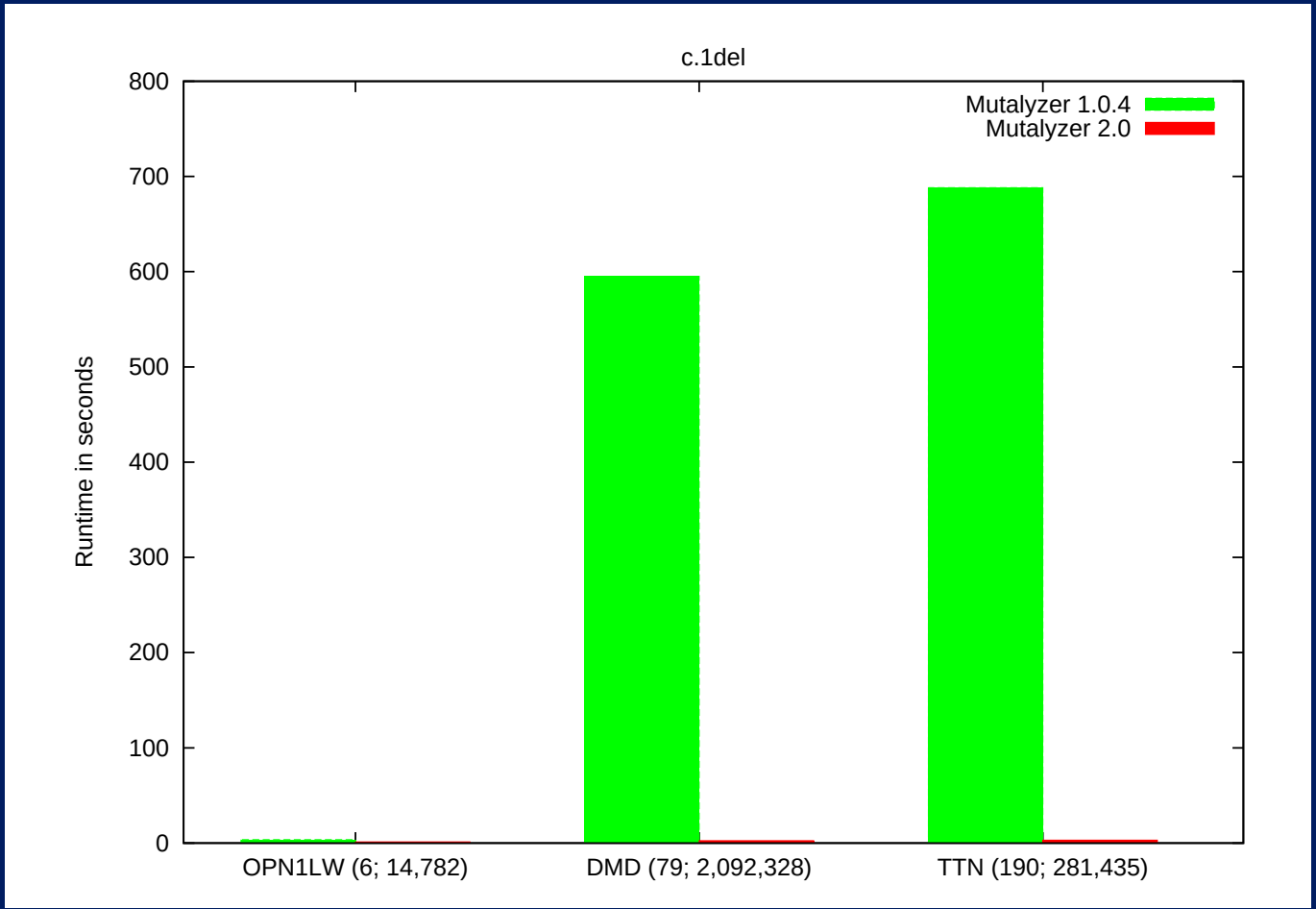
For a large number of checks, there are other interfaces.

- Batch interfaces (upload a table, receive the result by mail):
  - Name checker.
  - Syntax checker.
  - Position converter.
- Programmatic acces via a SOAP webservice (use from your own scripts).
  - Currently 11 functions available.
    - Position conversion.
    - Mutate a reference sequence.
    - Retrieve all transcripts in a range of a chromosome.
    - ...

# Comparison to the old version (1.0.4)

| | Mutalyzer 1.0.4 | Mutalyzer 2.0 |
|---|---|---|
| Disambiguation | ± | ++ |
| Complex variants | −− | ++ |
| Protein description | ± | + |
| Up / downstream descriptions | −− | ++ |
| Comprehensible error messages | − | ++ |
| Using a protein reference | ± | −− |
| Batch checkers | ± | ++ |
| GenBank uploader | + | ++ |
| Position conversion | −− | ++ |
| Programmatic access | −− | ++ |
| Other organisms / organelles | ± | ++ |

# Comparison to the old version (1.0.4): runtime




A $229\times$ speedup was measured (from almost $12min$ to about $3s$).

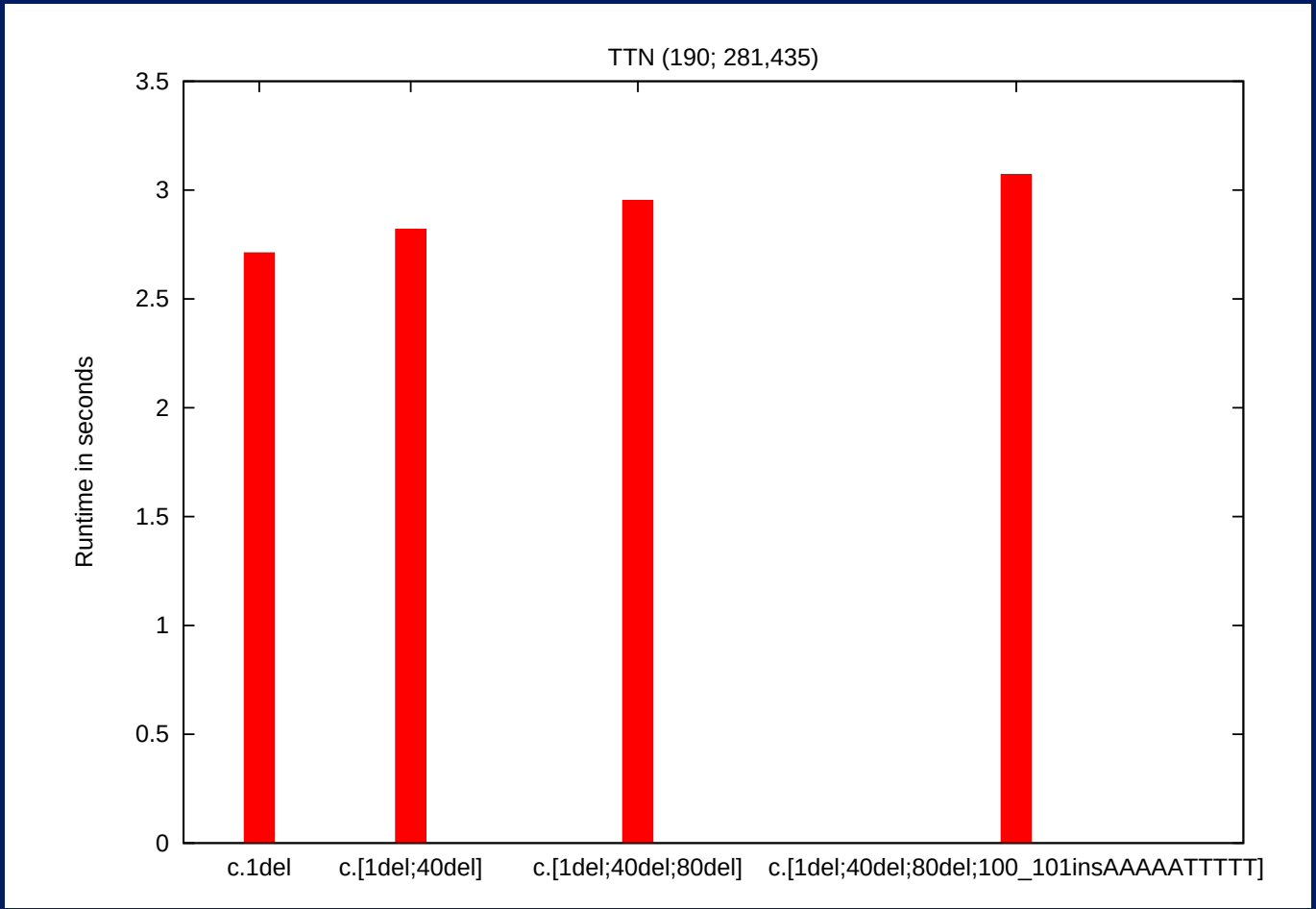
LUMC

# Comparison to the old version (1.0.4): code

| | Mutalyzer 1.0.4 | Mutalyzer 2.0 |
|---|---|---|
| Total (lines) | 7,752 | 11,396 |
| Total (bytes) | 365,736 | 390,316 |
| Minimised (lines) | 5,102 | 4,320 |
| Minimised (bytes) | 232,611 | 156,803 |
| Percentage of code (lines) | 66% | 38% |
| Percentage of code (bytes) | 64% | 42% |

The total amount of *source code* in Mutalyzer 2.0 is 107% of that in Mutalyzer 1.0.4, but the amount of *program code* is only 67%.

# Scalability: runtime with increasing complexity




The overhead ($\pm 2.5s$) is due to loading the reference sequence.

## Conclusions and further research

- Connection to LOVD via webservices.
- Development of a batch interface for the SNP converter.
- Using protein reference sequences.
- Connection to SVEP.
  - Splice prediction.
  - Alternative start.
  - Branch sites.
  - Transcription factors binding sites.
  - Protein effect prediction.
  - ...

# Questions?

Acknowledgements

Gerben Stouten
Gerard Schaafsma
Ivo Fokkema
Jacopo Celli
Johan den Dunnen
Peter Taschner

http://www.mutalyzer.nl/