Importing Next Generation Sequencing Data in LOVD 3.0

Jerry Hoogenboom, Ivar C. Lugtenburg, Ivo F.A.C. Fokkema
Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands

Introduction

Leiden Open Variation Database (LOVD)[1] is an open-source Locus Specific Database (LSDB) system. LSDBs are set up around a specific gene or disease to aid researchers by storing a patient's variants in a structured, searchable way. Various online data sources and publications are referenced to help identify possibly pathogenic variants.
The Human Genome Variation Society (HGVS) has provided a nomenclature[2] for the description of variants in LSDBs. The HGVS nomenclature has been created to avoid ambiguities in variant descriptions.

Next Generation Sequencing

LOVD 3.0, which entered the beta phase of development in January 2012, extends its abilities beyond that of a traditional LSDB by allowing storage of variants anywhere on the genome, including intergenic regions. This enables the ability to store Next Generation Sequencing (NGS) data.
NGS is becoming increasingly popular due to the rapidly increasing performance of sequencing platforms. A major advantage of NGS is the ability to screen all genes simultaneously, eliminating the need for selecting genes of interest beforehand, which can be a very laborious task.

Importing the data

NGS generates a huge amount of data — too much to enter into LSDBs manually. Therefore, LOVD 3.0 offers import functionality for two popular NGS file formats: Variant Call Format (VCF)[3] and SeattleSeq Annotation format. The Variant Call Format is the most popular file format currently in use for the description of variants found using NGS. The SeattleSeq Annotation file format is the output format of the SeattleSeq Annotation service[4], which combines a number of web resources to annotate variants in detail.

Making HGVS descriptions

```
##fileformat=VCFv4.1
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=SUBST,Number=0,Type=Flag,Description="Indicates that the variant is not an INDEL.">
##FILTER=<ID=q10,Description="Quality below 10">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=PL,Number=-1,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO            FORMAT      sample1
chrX    1000000 .       GA      G       20      PASS    INDEL;DP=15     GT:GQ:PL    0/1:75:219,0,247
chrX    1000003 .       AC      A       12      PASS    INDEL;DP=15     GT:GQ:PL    1/1:61:127,205,0
chrX    1000006 .       G       GA      8       q10     INDEL;DP=11     GT:GQ:PL    0/1:51:103,0,95
chrX    1000008 .       T       TC      17      PASS    INDEL;DP=13     GT:GQ:PL    1/1:62:130,120,0
chrX    1000010 .       CTTGG   CCAAG   13      PASS    INDEL;DP=13     GT:GQ:PL    0/1:70:201,0,234
chrX    1000016 .       AGG     A,AG    7       q10     INDEL;DP=10     GT:GQ:PL    1/2:56:250,245,198,210,0,193
chrX    1000020 .       T       G       20      PASS    SUBST;DP=14     GT:GQ:PL    0/1:80:252,0,249
chrX    1000022 .       A       C       19      PASS    SUBST;DP=15     GT:GQ:PL    1/1:72:210,201,0
```

Figure 1: An example VCF file. The data shown here is artificial.

POS	REF	ALT			
203	GA	G	⇒	Deletion	g.204del
402	AC	ATGC	⇒	Insertion	g.402_403insTG
514	ACT	ACTCT	⇒	Duplication	g.515_516dup
587	T	A	⇒	Substitution	g.587T>A
598	CTTGG	CCAAG	⇒	Inversion	g.599_601inv
623	AT	AGGC	⇒	Del & ins	g.624delinsGGC

Table 1: Constructing HGVS descriptions from VCF variant data.

Mapping variants to transcripts

Since VCF files do not include transcript-related data, LOVD 3.0 automatically maps the variants to transcripts. For this, data is fetched from a number of web services provided by the Mutalyzer[5] project.
Because the many queries that need to be done over the internet make the mapping process very slow, mapping is done in the background *after* the variants have been imported. Variants are mapped in small groups while users are browsing the database. Progress is visualised by means of a small progress meter in the footer of every page, as can be seen in Figure 5.

Finding variants in other LOVDs

As a side project, a service was created that allows users to search individual variants in all known public LOVDs. So, after having filtered the imported NGS data, users can click a 'Search' button for the potential variants of interest. LOVD will then open up a simple pop-up window listing all other LOVDs worldwide that share the same variant, so the user may find more information there. A screenshot of the pop-up window can be seen in Figure 6.

References

[1] Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, and den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat*, 32(5):557–63, May 2011.

[2] den Dunnen JT and Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Hum Mutat*, 15(1):7–12, Jan 2000.

[3] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, and Durbin R. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–8, Aug 2011.

[4] SeattleSeq Annotation server. `http://snp.gs.washington.edu/SeattleSeqAnnotation134/`.

[5] Wildeman M, van Ophuizen E, den Dunnen JT, and Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the MUTALYZER sequence variation nomenclature checker. *Hum Mutat*, 29(1):6–13, Jan 2008.

Screenshots

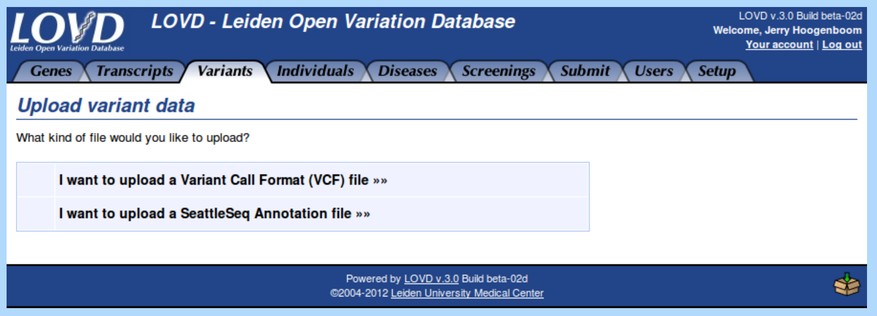

Figure 2: The user is asked which type of file they want to upload in the LOVD 3.0 submission process.

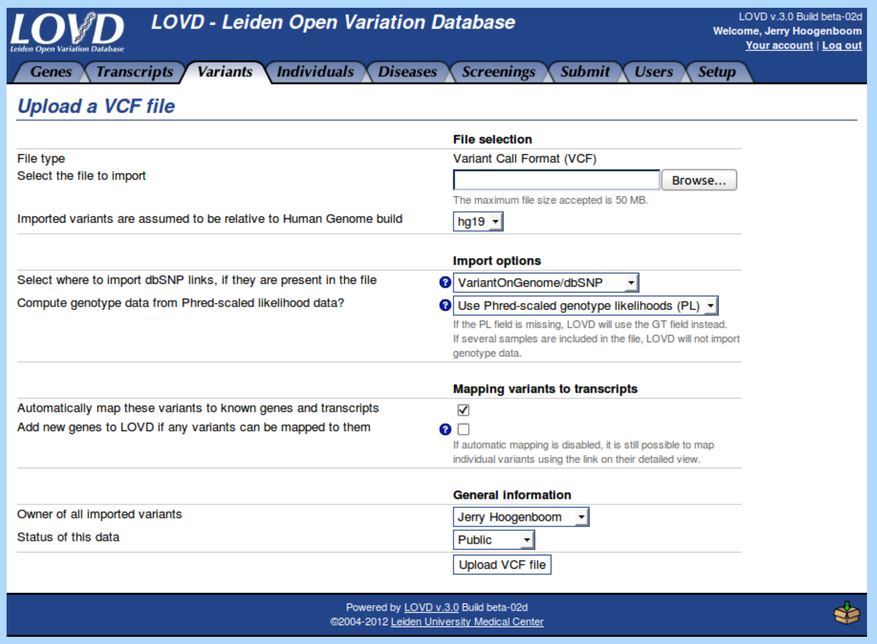

Figure 3: The VCF file upload form.

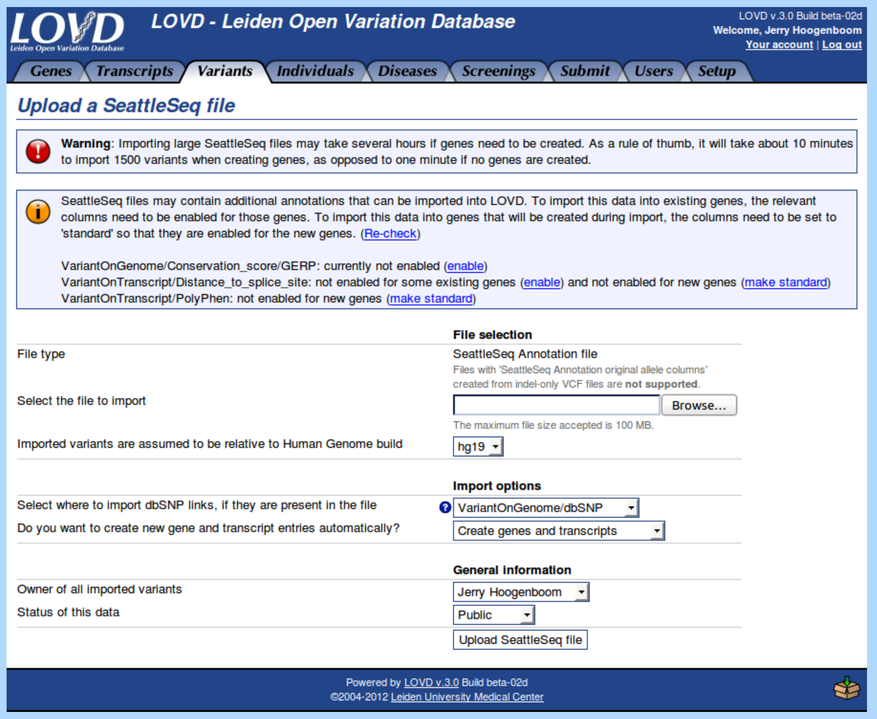

Figure 4: The SeattleSeq Annotation file upload form.

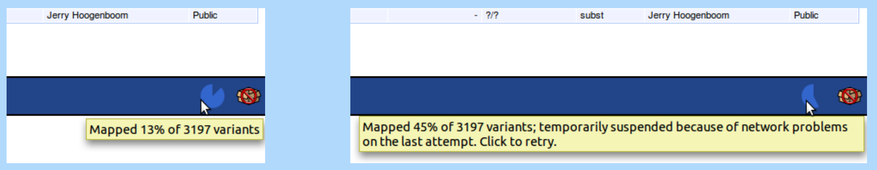

Figure 5: Variants imported from VCF files are mapped to transcripts automatically in the background.

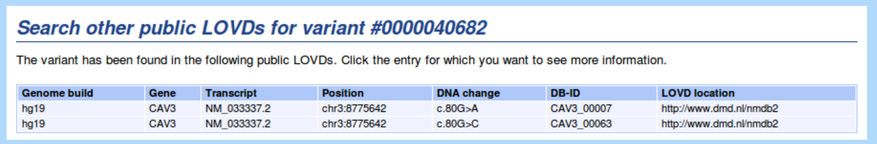

Figure 6: Individual variants of interest can be searched in other public LOVDs worldwide.