

C C A T T 0 1 0 0 0
G A G G A 0 1 1 0 1
G A A T T 0 0 1 1 0
A C A A G 0 0 1 0 0
T A C C A 0 0 1 1 0
T T A C A 0 1 0 0 0
A C C T C 0 0 0 1 0
A A G G A 0 0 0 0 0
G A T G A 0 1 1 0 0
T A G A T 0 0 1 0 0
G A T G A 1 0 1 0 0
T G T A G 1 0 0 0 0
T A G T A 0 0 0 0 0
G A T A T 1 0 0 0 0
G A G T G 1 1 0 0 0
A G A T T 1 1 0 0 0
G A G T A 1 1 0 0 0
T G A T G 1 1 0 0 0
A T T A G 1 1 0 0 0
T A G A T 1 1 0 0 0
G A G A 1 1 0 0 0
G T A 1 1 0 0 0
G A T 1 1 0 0 0
T A G 1 1 0 0 0
A G 1 1 0 0 0
G A 1 1 0 0 0
A 1 1 0 0 0
T 1 1 0 0 0

Case Study

Sequencing-based genetic testing for rare diseases

March 11, 2008



Sequencing-based genetic testing for rare diseases

For diagnosing genetic diseases, sequencing-based methods have been used for years, but still there are a lot of challenges in terms of data analysis since the sequencing data has to be analyzed and interpreted to present a diagnose.

BioGlobe GmbH, a leading Germany-based medical genetics company, performs diagnostic tests for a large number of rare genetic diseases. The company offers its services to hospitals and provides all the laboratory expertise and scientific knowledge required to deliver a diagnose to the hospital's clinical staff. This case study uses BioGlobe as an example of how CLC bio's *DNA Workbench* supports a genetic testing work flow.

Cystic fibrosis - an example

As an example to illustrate this work flow, we describe the genetic test for *Cystic fibrosis*. There are different methods to test for cystic fibrosis. MLPA (Multiplex Ligation-dependent Probe Amplification) can be used for detecting gross deletions or insertions, and OLA (Oligonucleotide Ligation Assay) is used to test for predefined known mutations. An alternative to OLA is sequencing which will detect all point mutations in the sequenced regions. This example described the sequencing-based test.

Cystic fibrosis is caused by mutations in the *cystic fibrosis transmembrane conductance regulator* gene (CFTR). The protein product of this gene works as an ion channel across the cell membrane, where it among other things regulates osmosis and other cell membrane channels. The CFTR gene is large (188,703 base pairs) with 26 exons (see figure 1). More than 1,000 mutations have been reported in the CFTR gene, but not all are associated with the cystic fibrosis disease.

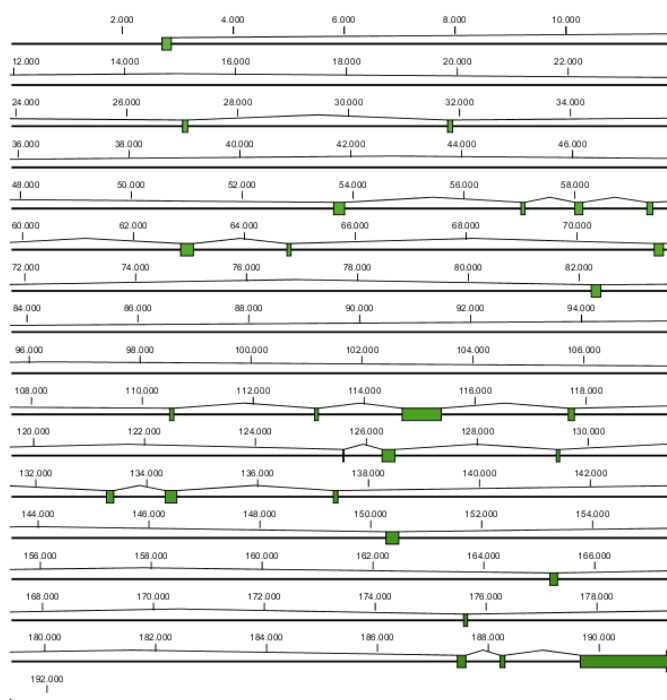


Figure 1: The exon structure of the CFTR gene as annotated on the sequence downloaded from NCBI.

At BioGlobe, a number of regions across six different exons are sequenced to capture the mutations known to cause over 99% of the cystic fibrosis cases. Using information in online databases, variations in the sequenced data are evaluated, and subsequently a diagnosis is produced.

Figure 2 gives an overall impression of the work flow for performing a sequence-based test for cystic fibrosis. The first part of the work flow is only performed the first time a test is conducted - once the reference sequence has been annotated and the primers have been designed, they can be re-used for future tests. Only occasional minor updates to the reference sequence are needed.

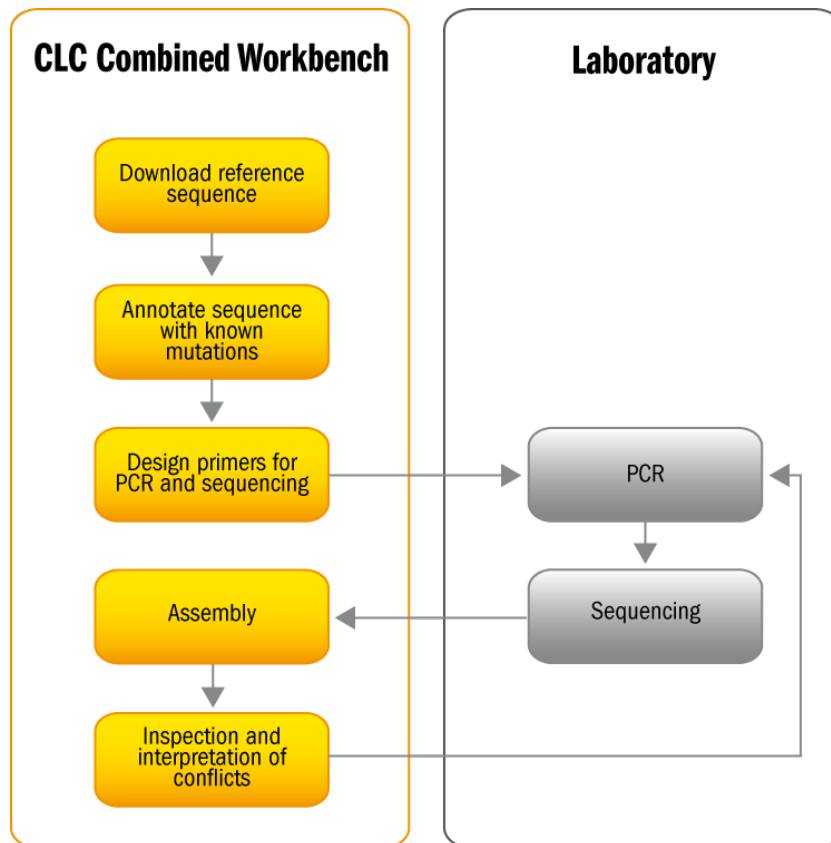


Figure 2: Sequence-based testing work flow. Note that the top part only is performed once - the primers and reference sequence can readily be used again for the next sample to be tested.

Preparing reference sequence and designing primers

Before sequence-based diagnostic can be performed, a reference sequence needs to be prepared. In the CLC DNA Workbench, the genomic sequence for the *cftr* gene is downloaded from GenBank. The sequence is then manually annotated with the most important mutations. Across the six exons, 17 mutations are annotated. A number of online databases are used:

- The NCBI SNP database, dbSNP
- NCBI reference genomic chromosome assemblies
- Expasy(SwissProt)

- OMIM
- HGMD
- A cystic fibrosis-specific mutation database at <http://www.genet.sickkids.on.ca/cftr>.

Using the *SNP Annotation Using BLAST* functionality of the Workbench, the sequence is also automatically annotated with SNPs from dbSNP, the SNP database at NCBI. The rest of the mutations are manually annotated.

Once the sequence is annotated, the next step is to design PCR primers for the regions that should be amplified. This work is guided by the annotations - both to make sure that the most important mutations are included in the amplicons, and to make sure that the primers do not anneal to a region that could possibly include a mutation.

The CLC DNA Workbench makes it possible to evaluate the quality of the primers and set criteria for melting temperature. For PCR primers, it calculates pairs of primers with minimal risk of pair annealing. The primer binding sites are annotated on the sequence, and the primers are saved ready for ordering.

Figure 3 shows an excerpt of the annotated sequence. The light-red annotations are the binding sites of the PCR primers (called *CFTR_EX07_F* and *CFTR_EX07_R*), and the dark-red annotations are the binding sites for the sequencing primers (called *CFTR_EX07_SF* and *CFTR_EX07_SR*). The sequencing starts a while before the exon (indicated by the green mRNA annotation) because the quality of the sequencing data is typically lower at the ends of the sequencing read.

This amplicon contains three known mutations (the light-blue annotations) which have been manually annotated.

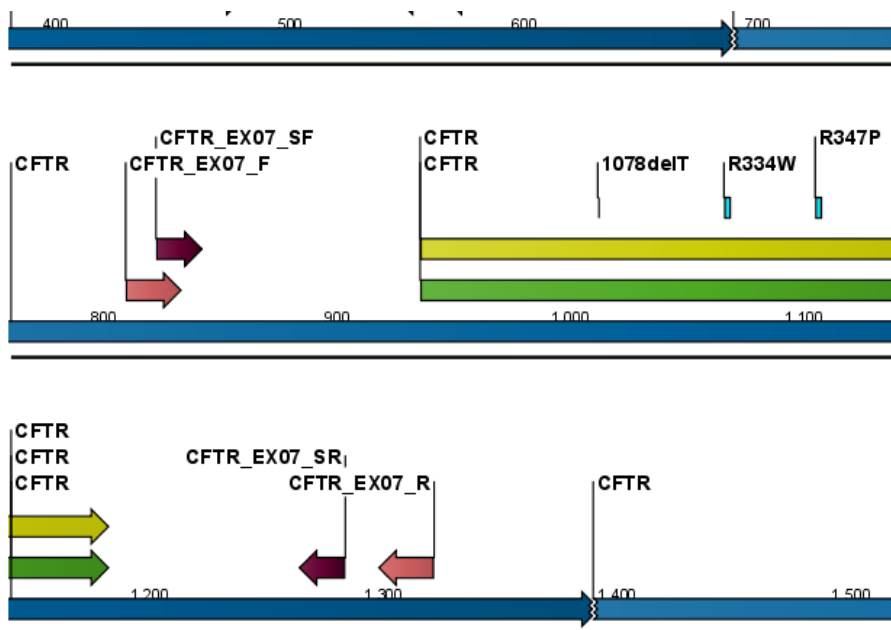


Figure 3: An excerpt of the annotated reference sequence. Note that the sequence is wrapped, the first line starting at around position 400, the second line continues at around position 750 etc.

Sequencing and assembly

In the lab, the PCR and sequencing processes are performed, and the sequencing data are exported in a scf format and imported into the CLC DNA Workbench. A reference assembly is conducted using the same sequence, now annotated with both mutations and primer binding sites.

The reference assembly automatically trims the sequencing data and produces an alignment displaying the chromatogram traces as shown in figure 4.

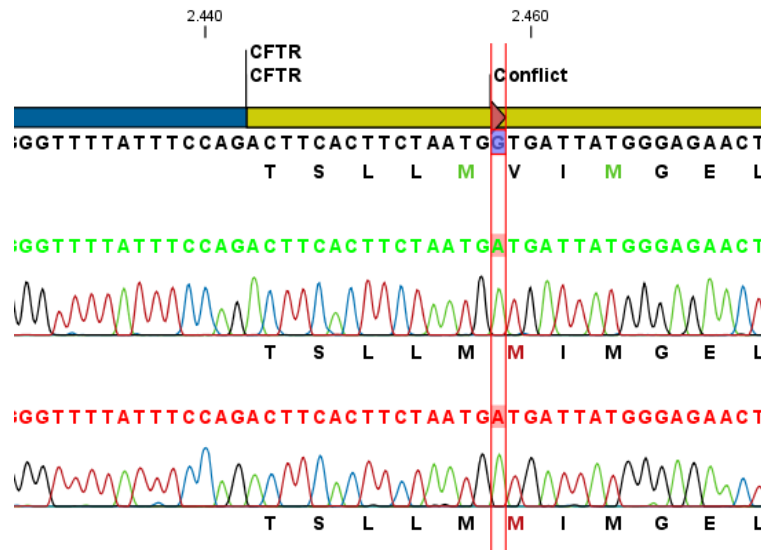


Figure 4: A forward read (shown in green) and a reverse read (shown in red). Both reads have an A in this position, whereas the reference has a G. The translation shows that this mutation will substitute Valine for Methionine. In this example, there is no known mutation annotated.

The assembly is then inspected by stepping through the positions where there are conflicts between the sequencing reads and the reference sequence. Each conflict is evaluated aided by the chromatogram traces and the mutation annotations. If the mutation in the sequencing data matches the annotation on the reference sequence, it is noted for the final report.

Mutations are also evaluated in terms of their effect on the protein product. "Silent" mutations not resulting in an amino acid substitution are shown in yellow, whereas frameshift mutations (caused by indels) and non-synonymous substitutions are shown in red (methionine in figure 4).

For documentation, the whole process of inspecting the contig is documented in the history which is saved together with the contig file. This means that it is possible to retrace the evaluation process later on. Snapshots of selected parts of the contig are saved as pdf documents which are used in the final report.

Conclusion

The integrated nature of the CLC DNA Workbench provides excellent support for the BioGlobe work flow. Integrating sequence annotation, primer design, assembly and inspection makes it faster to get a result for the test, and the built-in documentation and data handling elevates the quality control level.

This means that the researchers at BioGlobe can focus on adding value for customers through their high level of knowledge about the interpretation of the test results.

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.



See <http://creativecommons.org/licenses/by-nc-nd/2.5/> for more information on how to use the contents.