



Tutorial

Visualize Variants on Protein Structure

June 27, 2019

— Sample to Insight —

Visualize Variants on Protein Structure

Once variants have been detected in a sample, the challenge of interpretation remains. The first task is to separate disease-causing variants from a potentially large background of neutral variants. The second, greater challenge, is to understand the mechanism of disease action.

In this tutorial, we will address both of these problems using the **Link Variants to 3D Protein Structure** tool. This tool encourages interactive analysis, with insights limited only by the range of external knowledge that the user can bring to bear. This tutorial will provide some general suggestions for how an analysis may proceed.

Chronic myeloid leukaemia (CML) is associated with the formation of a kinase Bcr-Abl fusion protein. In 2001 the pharmaceutical company Novartis released *imatinib* as an inhibitor of this kinase. Imatinib dramatically increased CML survival rates, and was hailed by Time magazine as a "magic bullet" against cancer. However, despite its initial success, it soon became clear that some patients had imatinib-resistant forms of CML. Additional inhibitors have since been released, and a variant may confer drug resistance to some or more of these.

The provided data set contains several variants in the Abl gene, of which some are associated with drug resistance, whereas others are common variants from the Hapmap and 1000 Genomes projects. We will aim to separate the benign and clinically relevant variants, and infer the mechanism of resistance.

For this tutorial, you need to be working with CLC Genomics Workbench. The analyses carried out in this tutorial include:

- Linking variants to 3D protein structures
- Inspection of model quality
- Distinguishing harmful and neutral variants
- Investigating the mechanism of variant action

Generating links to 3D protein structure

Database of protein sequences with known 3D structure The **Link Variants to 3D Protein Structure** tool uses a database of protein sequences with known 3D structures. This database must be downloaded the first time the tool is run. To download the database, use the Launch button () to find the tool **Download 3D Protein Structure Database** (). Select a download location from the drop-down menu (a default is supplied) and click **Finish**.

CDS and reference sequence Before getting started you will need to download a reference genome and CDS track using the Reference Data Manager (1) found in the upper right corner of the Workbench (figure 1). Under the **Download Genome** tab (2), select the **Homo sapiens - hg19** data (3). Choose to "Download genome sequence" (4) and check the "Genome Annotations" item (5) to get annotation tracks for hg19. Click on **Download** (6). You can check the Download process at the bottom of the wizard.

The CDS and reference sequence are now saved in the CLC_References | Genomes | Homo_sapiens_hg19 folder accessible from the Navigation Area.

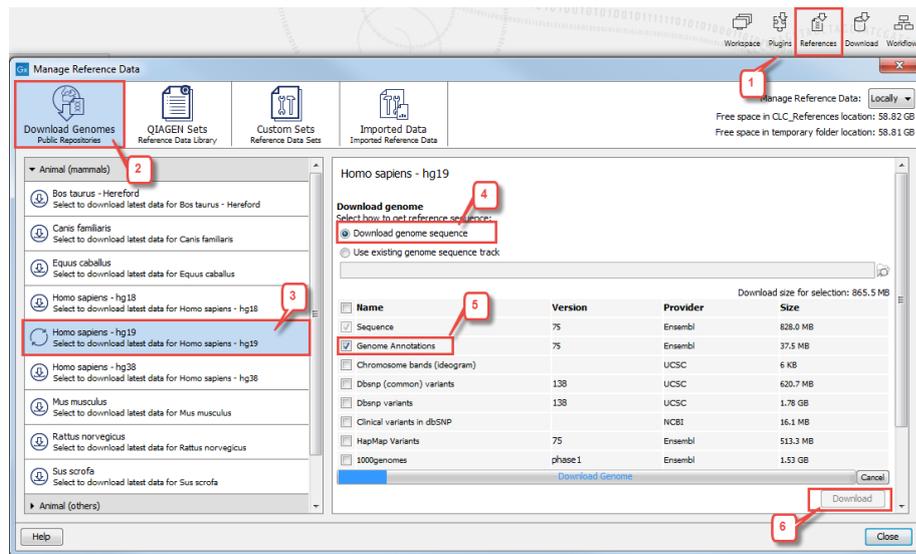


Figure 1: Open the Reference Data Manager to download the relevant reference databases.

Data specific to the tutorial

1. Download the example data from http://resources.qiagenbioinformatics.com/testdata/visualize_variant.zip.
2. Import the Abl variants tracks and the ABL1model10PK file using the standard import option:

File | Import | Standard import

3. Create a new folder for this tutorial and click **Finish**.

You are now ready to run the tool.

1. Use the Launch button (🚀) to find the tool **Link Variants to 3D Protein Structure** (🧬). If you are connected to a server, you will first be asked where you want to run the analysis.
2. In the next wizard step you will be asked for an input file. Select the variant track Abl variants from the downloaded tutorial data (figure 2). Click **Next**.

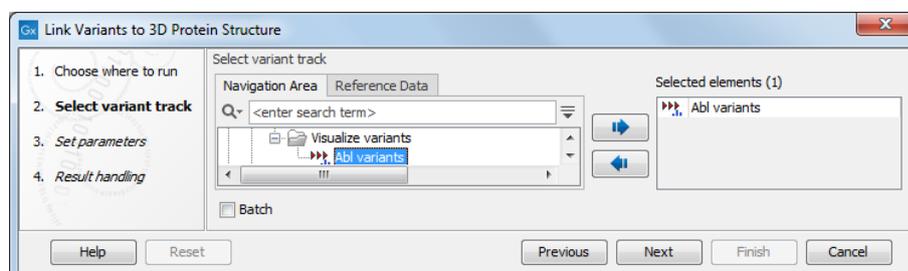


Figure 2: Select the variant track holding the variants that you would like to visualize on 3D protein structures.

3. In the next wizard step, you must provide a CDS track and the reference sequence track hg19 (figure 3). You can find these in the CLC_References in the Navigation Area.

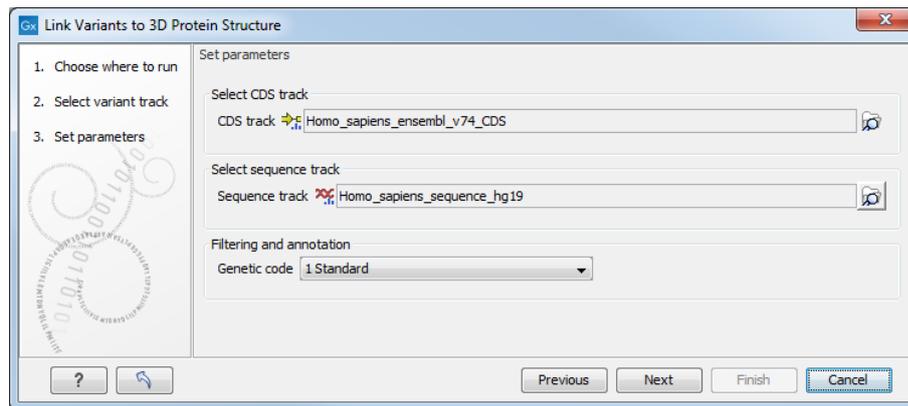
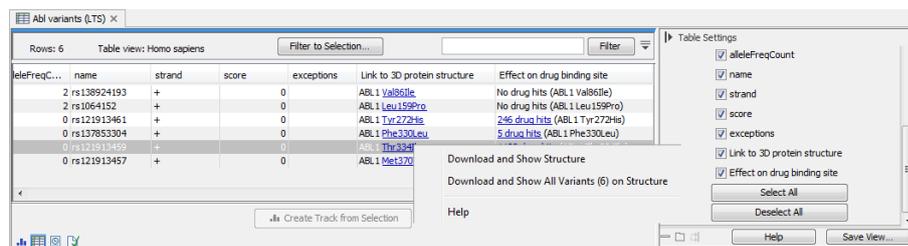


Figure 3: Select CDS and reference sequence.

- In the last wizard, choose where you would like to save the data, and click **Finish**.

As output, the tool produces a new variant track `Abl variants (LTS)`. You can click on the table icon found in the lower left corner of the View Area to shift to table view and notice that this file has an additional column called "Link to 3D Protein Structure". This column contains links to the modeled structures (figure 4).



alleleFreqC...	name	strand	score	exceptions	Link to 3D protein structure	Effect on drug binding site
2	rs138924193	+		0	ABL1 Val86Ile	No drug hits (ABL1 Val86Ile)
2	rs1064152	+		0	ABL1 Leu159Pro	No drug hits (ABL1 Leu159Pro)
0	rs121913461	+		0	ABL1 Tyr272His	246 drug hits (ABL1 Tyr272His)
0	rs137853304	+		0	ABL1 Phe330Leu	5 drug hits (ABL1 Phe330Leu)
0	rs121913457	+		0	Thr334Ile	
0	rs121913457	+		0	ABL1 Met370Thr	

Figure 4: The output of the Link Variants to 3D Protein Structure tool. The table view is selected by the button indicated by the arrow. A new column contains clickable links to 3D protein structures.

Variants conferring drug resistance in Chronic Myeloid Leukemia

An overview of the variants is provided in the following table. We will fill in this table as the tutorial progresses.

Variant	Description
Val86Ile	
Leu159Pro	
Tyr272His	
Phe330Leu	
Thr334Ile	
Met370Thr	

Select one of the variants by clicking on one of the links found in the Link to 3D Protein Structure column and select **Download and Show All Variants on Structure** as shown in figure 4. After a few seconds of modeling, a 3D view of the variant will open (figure 5).

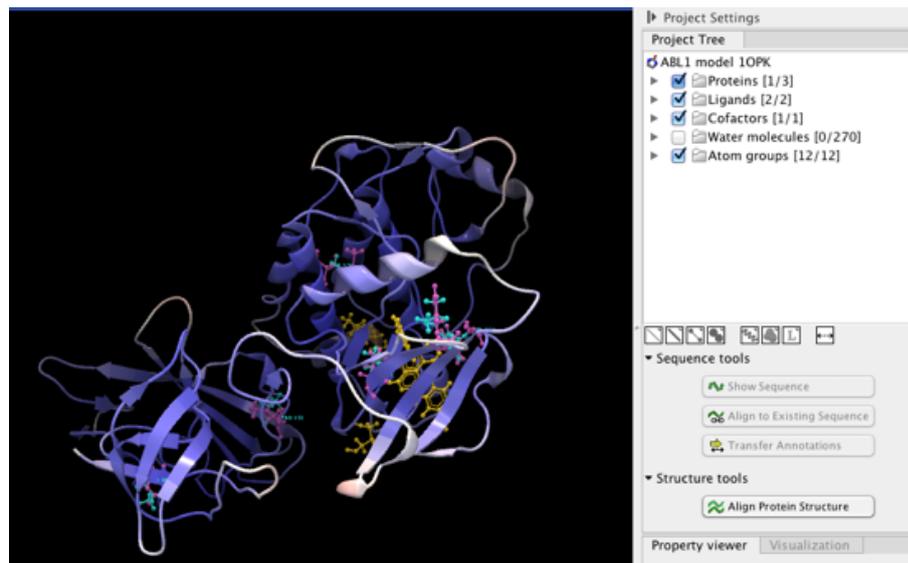


Figure 5: Initial view of six modeled variants.

Model quality We first assess whether the model is of sufficiently high quality. The initial view of the protein structure is colored blue in high-quality regions, and red in low-quality regions. Regions can be red either because the underlying structure data is itself of low quality, or because the protein structure on which the model was based has too dissimilar a sequence from the gene product for us to be certain where a variant lies on the structure. The Abl model is mostly blue, with some white loops, so we can be confident in continuing the analysis. If the variant had been located in a red region, or near to a gap, we would not have attempted to proceed with the analysis. Examples of bad models are shown in figure 6.



Figure 6: Examples of bad models. The model on the left is entirely red and contains many gaps. It is not suitable for analysis. The model on the right has a good blue region, but the variant lies in a red region of higher positional uncertainty.

The modeling process used by **Link Variants to 3D Protein Structure** has been designed to clearly show regions where the model quality is expected to be poor, and not to attempt to infer structure in regions where there is no supporting experimental evidence.

Viewing a variant We would now like to take a closer look at the variant. To do this we will use the **Project Tree** palette found in the **Side Panel**. In the **Project Tree** of the 3D view, the

category "Atom groups" contains two entries for each variant shown on the structure: one entry for the reference and one for the variant. Double-click on an entry to zoom the 3D view to the variant (figure 7). The initial view will show a set of atoms in magenta (the reference), and a set in cyan (the variant). These colors and styles can be changed using the buttons in the **Side Panel** (red rectangle in figure 7). For more details please see http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Customizing_visualization.html.

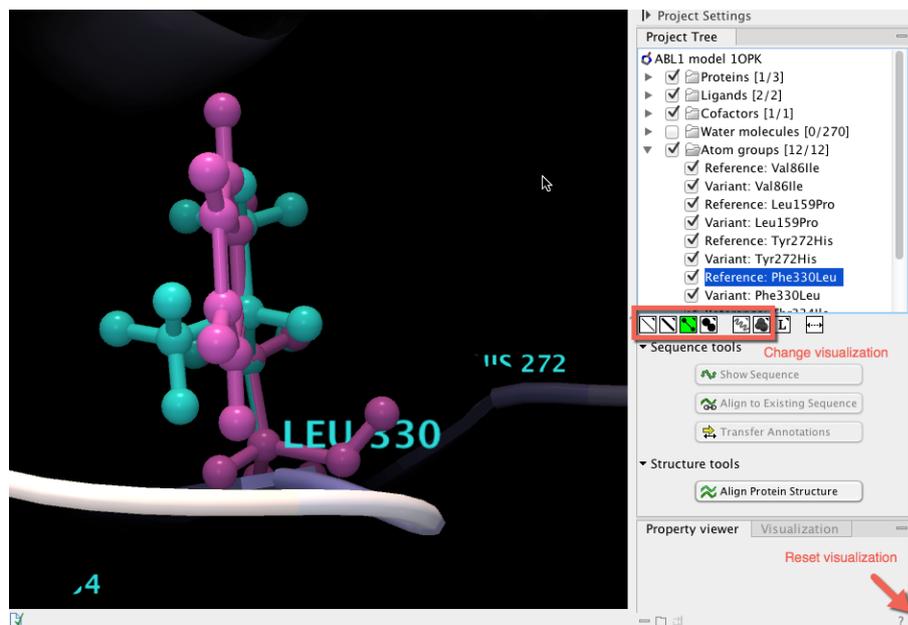


Figure 7: Zoomed view of a variant selected from the "Atom groups" in the Side Panel. The visualization can be changed using the buttons in the red rectangle, and reset using the button in the bottom right corner.

The initial view settings are saved on the Molecule Project as "Initial visualization", and can always be reapplied from the **Save View** menu (☰) found in the bottom right corner of the **Side Panel** (red arrow in figure 7).

Inspecting variants: location and clashes To understand the impact of a variant, it is useful to know the function of the protein region in which it lies. A good starting point is often the paper that describes the experimental structure on which the variant is modeled. If such a paper exists, it will be cited in the **History** (📄) at the bottom left part of the view. A search online for the relevant citation in this case retrieves the paper [http://dx.doi.org/10.1016/S0092-8674\(03\)00194-6](http://dx.doi.org/10.1016/S0092-8674(03)00194-6).

In this example, we know the effect of the variants we are interested in: they should confer drug resistance. This suggests that they are located close to the drug binding pocket, which is here occupied by the inhibitor, P166326 (found as P16 A 2 in the "Ligands" category of the **Project Tree**). There are only two variants close to this pocket, Tyr272His and Thr334Ile, and these seem the most natural candidates for further investigation.

Two of the other variants, Val86Ile and Leu159Pro do not lie on the kinase domain and so seem unlikely to contribute to drug resistance. These two are the variants taken from the HAPMAP and 1000 Genomes sets, and have no known clinical significance.

For now, we will focus on the variants in the binding pocket. The most obvious sign of a damaging

variant is that it introduces a clash.

In figure 8 we see that an atom in the variant Ile 334 appears to clash with the inhibitor. We confirm this by selecting the folder **Atom groups** in the **Project Tree**, right-clicking the green button beneath the **Project Tree**, and selecting "Color by Temperature". In this scheme, clashing atoms will be bright red (figure 9). Note that atoms may be bright red for other reasons, most commonly if the backbone is also red, but this is not the case here. The clash suggests that the variant Thr334Ile prevents the inhibitor from binding in the correct conformation, and is likely to confer drug resistance.

What is a clash? A clash occurs when atoms are too close together. The exact definition of "too close" depends on the type of atoms involved. For example, a pair of hydrogen atoms can sit more closely than a pair of carbon atoms.

What can I infer from a clash? A clash indicates that the local region of the model is incorrect. Clashes in a binding site may suggest disruption of binding, whereas clashes within a protein suggest that the backbone of the real protein product may be slightly different than shown.

When are clashes expected? Atoms coordinated with metal ions will often have clashes because there is a bond between them, pulling them closer than they would otherwise be.

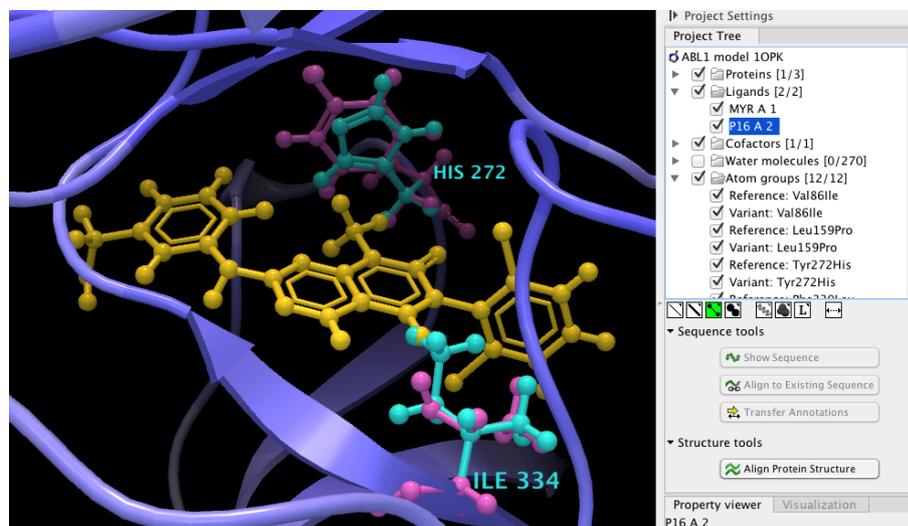


Figure 8: Variants around the binding site (viewed by double-clicking the ligand P 16 A 2).

Inspecting variants: interactions A summary of our current knowledge of the Abl variants is provided in the next table. We have identified two variants around the binding site, but have so far only suggested a mechanism of drug resistance for one of them. We will now investigate whether Tyr272His might also be involved in drug resistance.

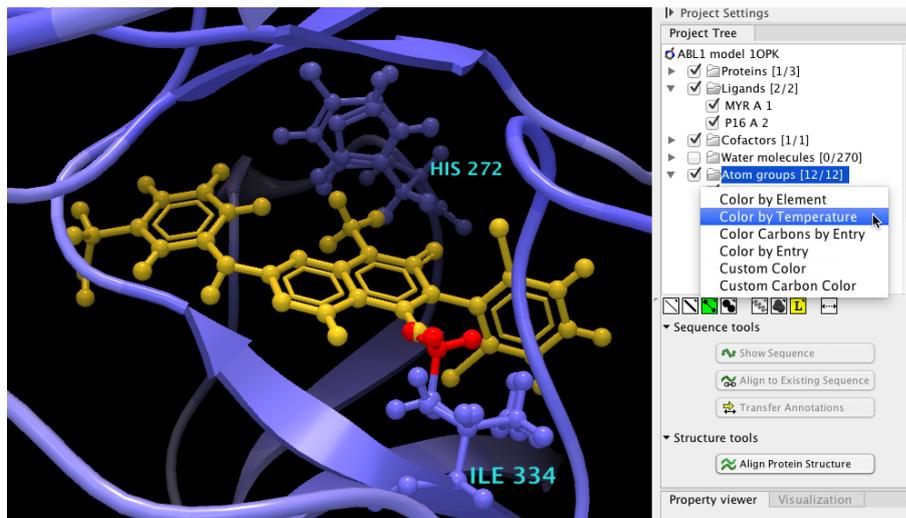


Figure 9: The same view as in figure 8, but now colored by temperature to show clashes in red.

Variant	Description
Val86Ile	Not in kinase domain.
Leu159Pro	Not in kinase domain.
Tyr272His	Present in binding site.
Phe330Leu	
Thr334Ile	Present in binding site. Drug resistance.
Met370Thr	

We know that Tyr272His does not contain any clashes, and this is to be expected – this variant replaces a larger ring with a smaller one. However, it is possible that this change has removed some chemical interactions that affect ligand binding. To test this, we zoom in on the variant by double-clicking its atom group, and change the color scheme to "Color by Element" (figure 10). Surrounding atoms in the rest of the protein are shown by changing the visualization of the protein as in figure 11, these atoms are also shown in the "Color by Element" scheme.

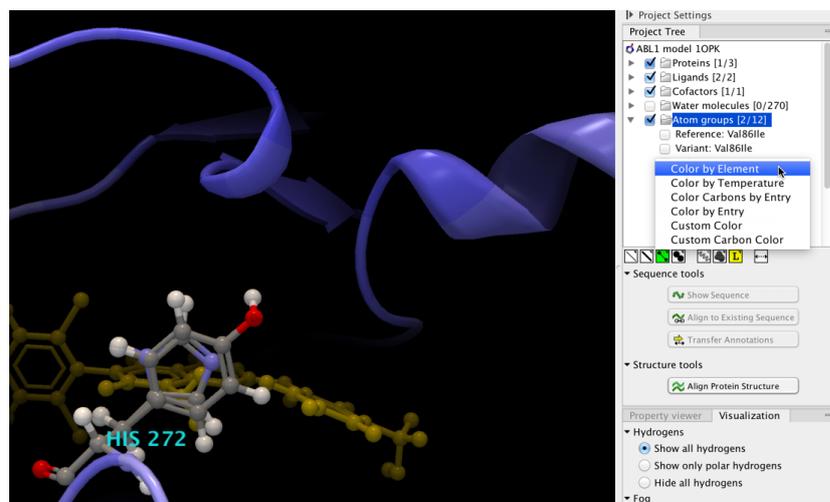


Figure 10: Zoomed in view of variant Tyr272His with element color scheme applied.

The principal difference introduced by the variant is the removal of an oxygen atom (red) and its associated hydrogen. This hydroxyl group had the potential to form stabilising hydrogen bonds

Tutorial

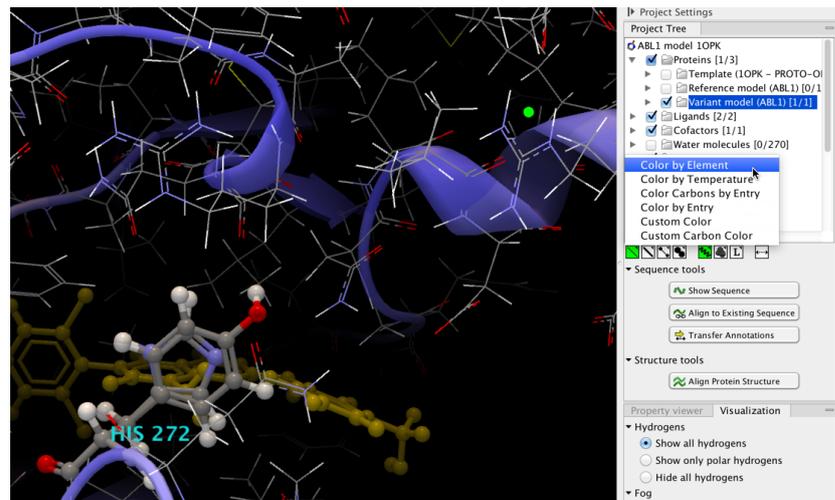


Figure 11: Atoms around the variant Tyr272His with element color scheme applied.

with some nearby oxygen (red) and nitrogen (blue) containing groups. Figure 12 shows a clearer view of these interactions. These hydrogen bonds would act to draw together three distinct regions of the structure to form the ligand binding site. It is possible that disruption of these would make the binding site less rigid and affect the energetics of drug binding.

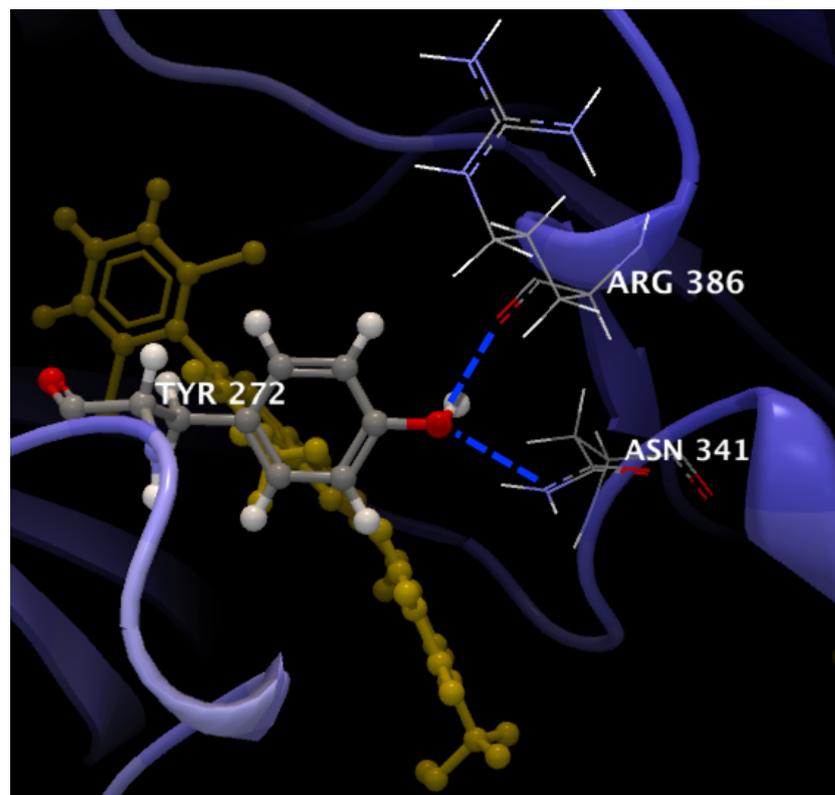


Figure 12: Zoomed in view of possible hydrogen bond interactions in the reference but not in the variant Tyr272His. This figure was made using a user defined atom group. The dashed blue lines have been added by hand.

The remaining variants The same techniques discussed above can be applied to the two remaining variants. The variant, Phe330Leu leads to no obvious difficulties, even though it is annotated as deleterious in the ClinVar database. One possibility is that it affects a different protein conformation from that shown in the model. Alternatively it may be that the effect of the variant is small – a suggestion supported by literature observations that increases in drug dosing are sufficient to overcome this resistance.

By contrast Met370Thr introduces a clash, a substantial change in side chain volume, and a change in the chemical properties of the side chain. Although these factors all suggest a deleterious effect, there is no obvious link to drug resistance; for example the clash may be resolved in the protein by local changes in the backbone without necessarily affecting the binding site.

A final series of conclusions are shown in the next table, where an additional column has been added to mark if a variant is known to confer drug resistance.

Variant	Description	Drug resistance?
Val86Ile	Not in kinase.	No
Leu159Pro	Not in kinase.	No
Tyr272His	Present in binding site. Drug resistance.	Yes
Phe330Leu	Appears benign.	Yes
Thr334Ile	Present in binding site. Drug resistance.	Yes
Met370Thr	Appears deleterious. Possible drug resistance.	Yes

Avenues for further analysis and conclusion

We have now concluded the analysis for this tutorial. In this final section we sketch out suggestions for a deeper analysis in the case when related protein sequences and structures are known.

The kinase protein in this tutorial has several known inhibitors and two principal conformations (active and inactive). It is useful to see a variant in as many of these contexts as possible, for example because a variant may confer selective drug resistance. A simple way to do this is via the **Align Protein Structure** tool in the **Side Panel**. In figure 13, a structure with bound imatinib in a different conformation has been structurally aligned. The variant Thr334Ile also appears to disrupt the binding of imatinib – a conclusion that is supported by clinical data.

Another line of evidence for whether a variant is tolerated is to look at the degree of conservation at the affected position, and at whether the alternative amino acid is present in an alignment of homologous proteins. Conserved positions are unlikely to tolerate substitutions. The **Align to Existing Sequence** tool in the **Side Panel** can take a set of sequences or alignments, and turn them into a linked alignment that displays both conservation information and a sequence logo. Selecting a residue in the linked alignment selects the corresponding atoms in the 3D view (figure 14).

In this tutorial we have examined a hand-picked set of six variants in the Abl gene for resistance to drugs used in the treatment of Chronic Myeloid Leukemia. By examining the location of the variants within the gene product, we quickly identified two promising resistance candidates, and decided not to focus on two other variants that have no known association with drug resistance.

Plausible mechanisms of resistance were found for both candidates. In one case a clash was

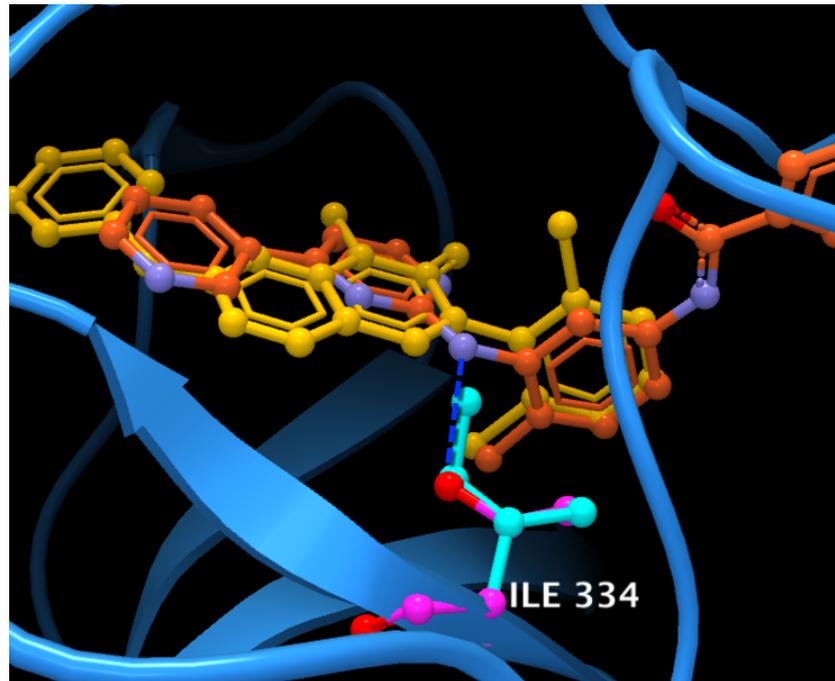


Figure 13: A structure alignment suggests that the Thr334Ile mutation probably also confers drug resistance to imatinib (dark orange). The steric clash discussed earlier remains, and a potential hydrogen bond to imatinib is removed (blue dashed line).

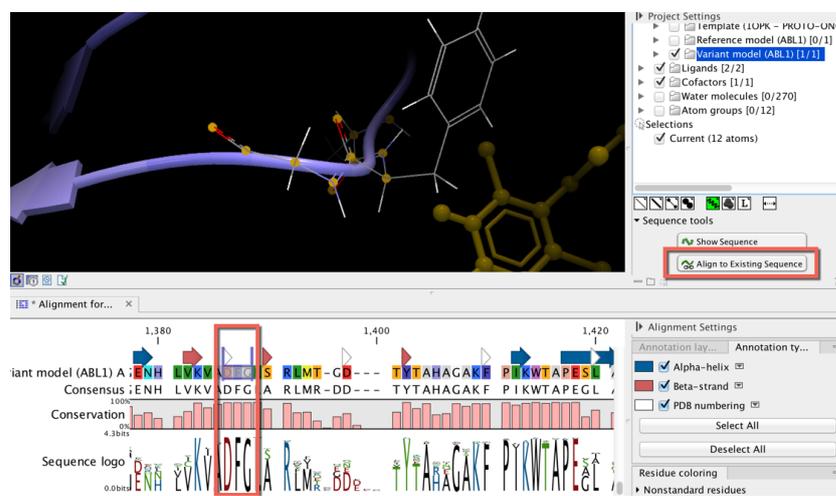


Figure 14: A linked sequence alignment can be used to view conservation information for a protein structure. Here the conserved DFG motif is selected in the alignment, leading to the corresponding atoms being shown in the 3D view.

detected with a co-crystallised drug, and in the other the loss of hydrogen bonds may lead to a less defined binding pocket. Of the remaining variants, one appeared benign despite appearing in the ClinVar database, and the other appeared deleterious, but with no clear link to drug binding.

Suggested further analyses included comparison of multiple related structures and sequences. Ultimately the scope of the interactive analysis available with the **Link Variants to 3D Protein Structure** tool is determined by the individual researcher.