The CMap Assembly Editor

Faga, B[1], Carmichael, L[2], Belter, E[2], Minx, P[2], Stein, L[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor NY
[2]Washington University, St. Louis MO

Abstract

With the advent of low-coverage sequencing and exotic sequencing technologies it is becoming increasingly necessary to use all available genome mapping data, including optical, physical and genetic maps, during the assembly and finishing phases of a sequencing project. The CMap Assembly Editor (CMAE), a desktop application, is being developed to assist in visualizing and editing large scale sequence assemblies for the maize sequencing project. Using the CMap comparative mapping database, CMAE allows sequence assemblies to be superimposed on top of diverse other types of mapping data, allowing the finisher to view assemblies in the context of a cascade of mapping data at a variety of resolutions.¬† For example, the editor can show sequence contigs aligned to fingerprinted physical map contigs, which are aligned in turn to genetic maps. Correspondence links between the different objects indicate the quality of the assembly and highlight possible mis-assemblies. The editor will then allow mis-assembled contigs to be split, merged or moved, or the troubled contigs can be exported to a more specialized program.

Introduction

The CMap Assembly Editor (CMAE) is being developed to aid in visualizing and modifying the large scale sequence assemblies being produced by the maize sequencing project. CMAE will allow a finisher to view the sequence assembly as it relates to other data such as genetic or physical maps. The finisher will also be able to drill down to the sequence contig level to see interesting read pairs. This will allow the finisher to visualize mis-assemblies and make appropriate modifications. Then when finished, the modifications can be exported to an external program to update the users in-house data.

Usage

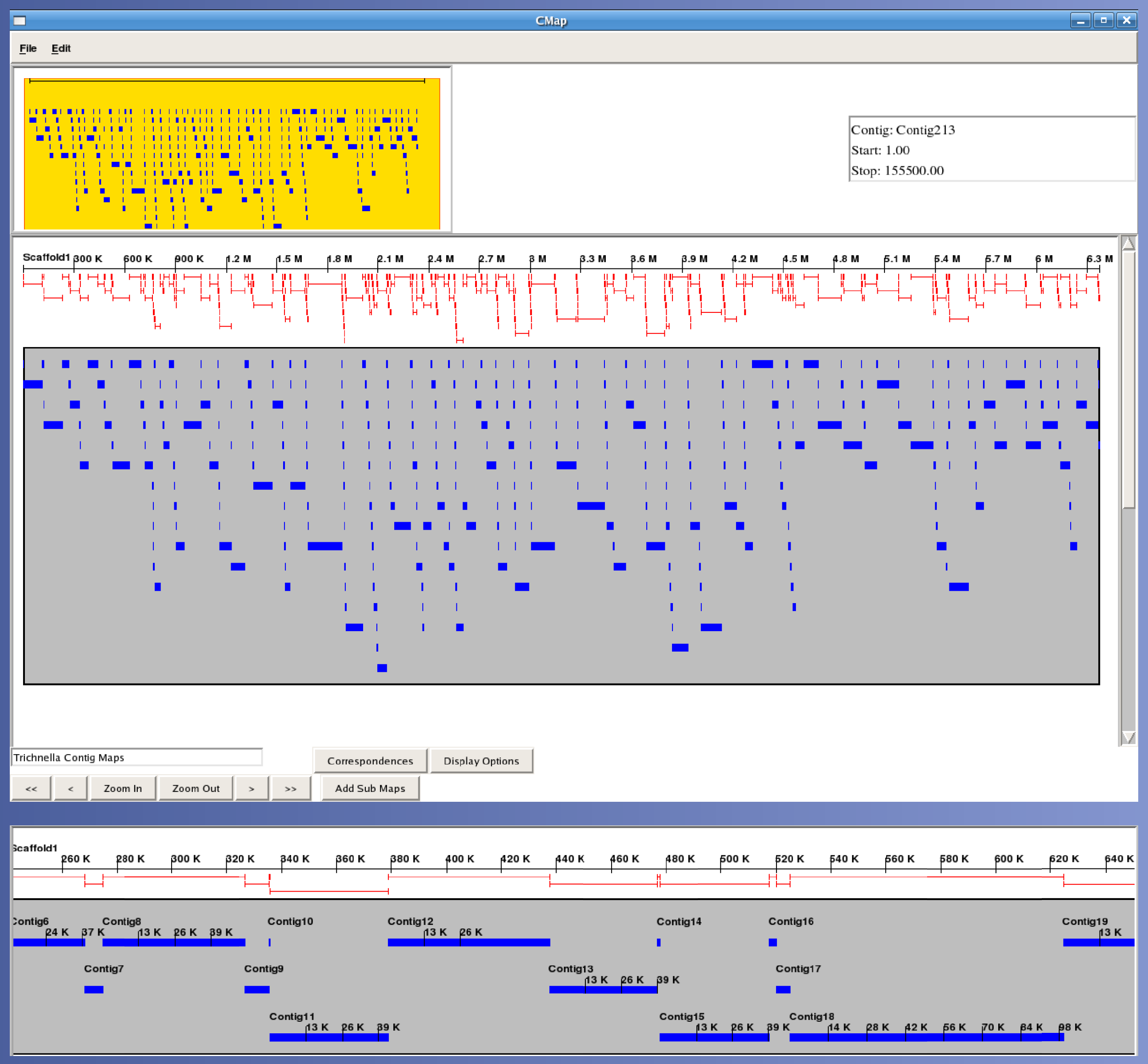

Contig: Contig213
Start: 1.00
Stop: 155500.00

Figure 1: Maps are presented in a tiered structure where sub-maps are aligned to parent maps. A simple case using the Trichnella spirilis assembly is shown in Figure 1a. Scaffolds of sequence contigs are displayed at the top with the contigs aligned below as sub-maps. Figure 1b displays a zoomed in view. Correspondences (not shown) between the contigs represent read pairs from the same clone that span contigs. This allows the finisher to see potential mis-assemblies.

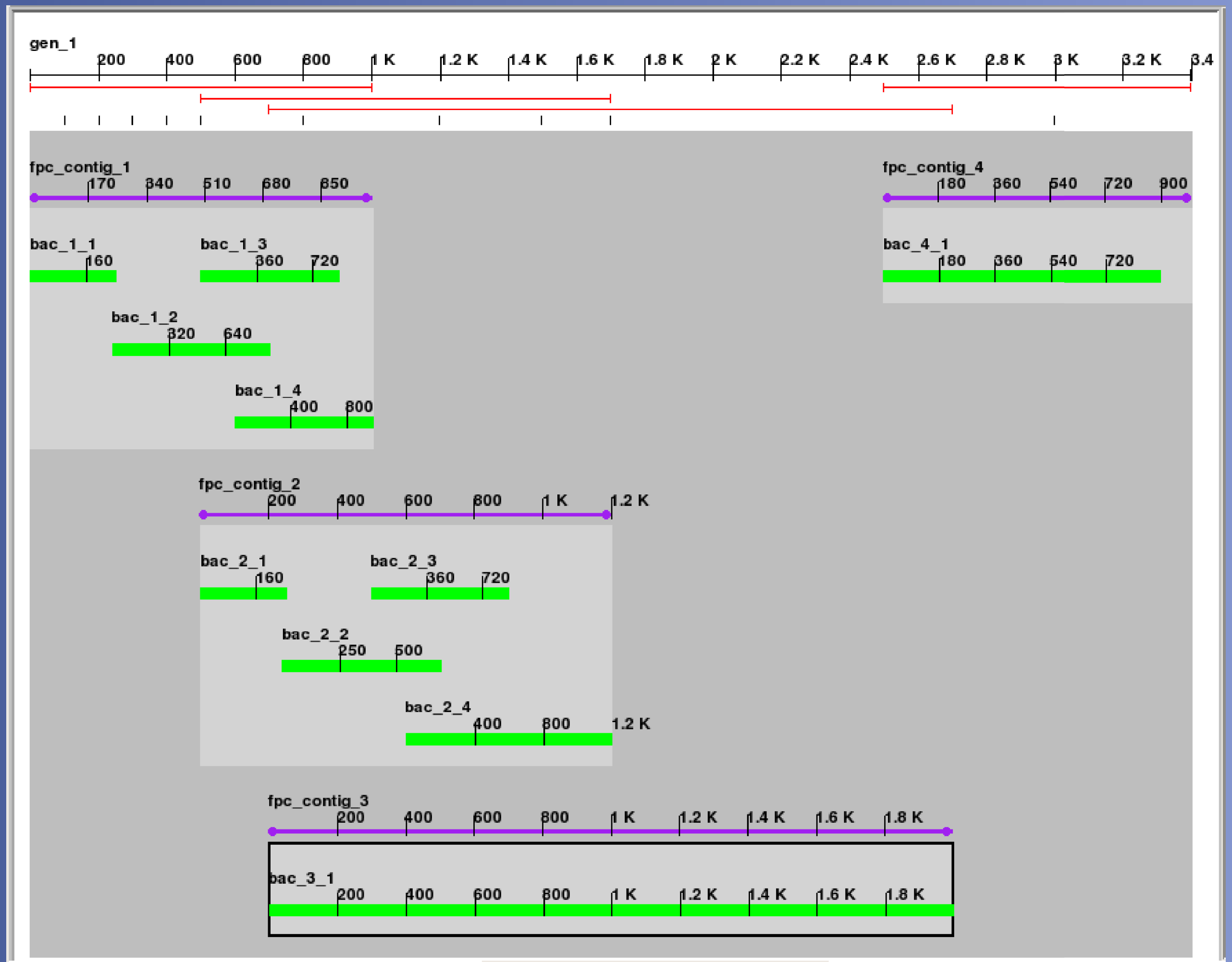


Figure 2: Other potential uses include displaying a genetic map with fingerprint contigs aligned. Then sequenced BACs can be aligned to the fingerprint contigs. All of this data can then be viewed together. Figure 2 shows this view using hypothetical data.

Moving Maps

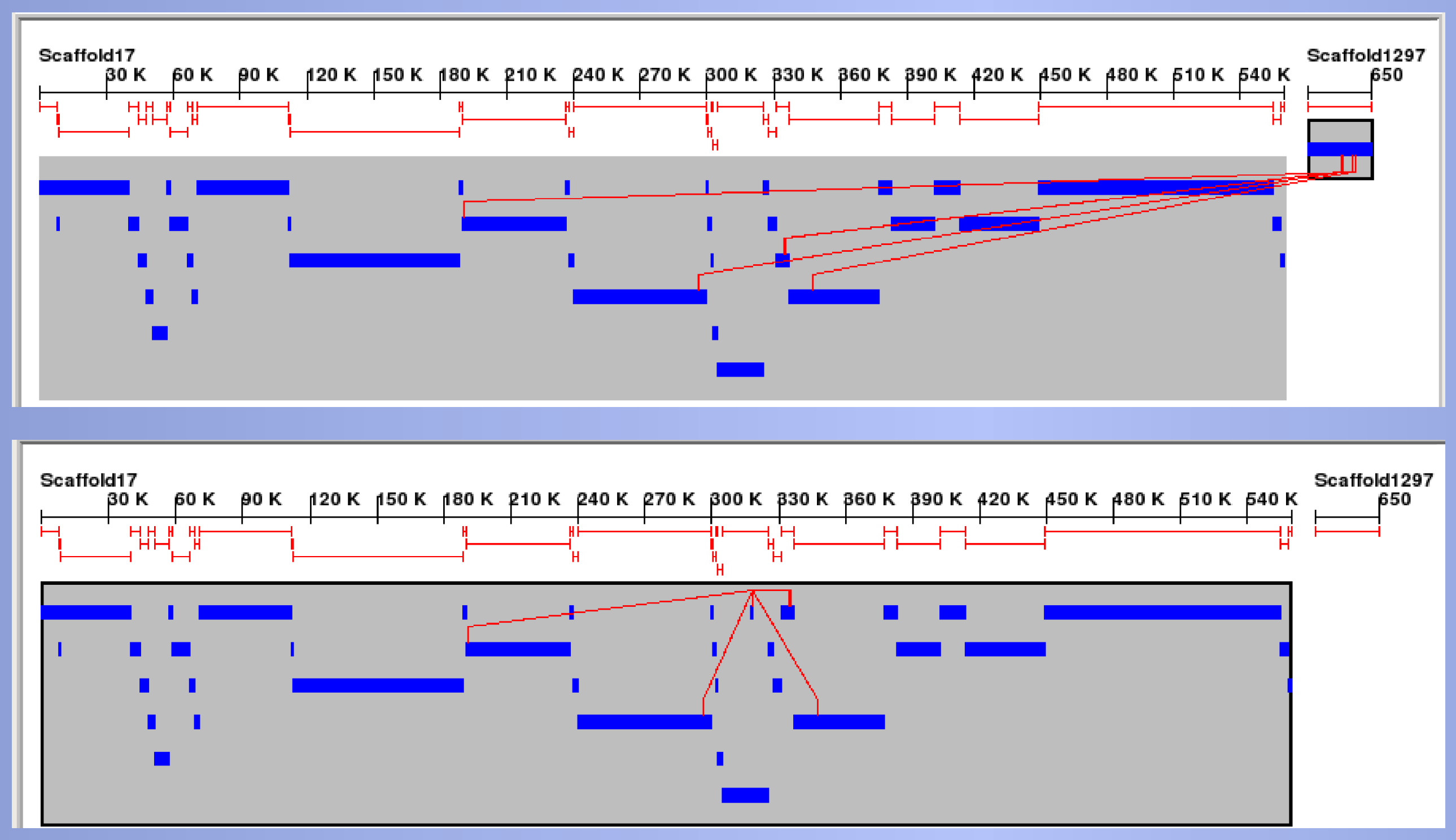
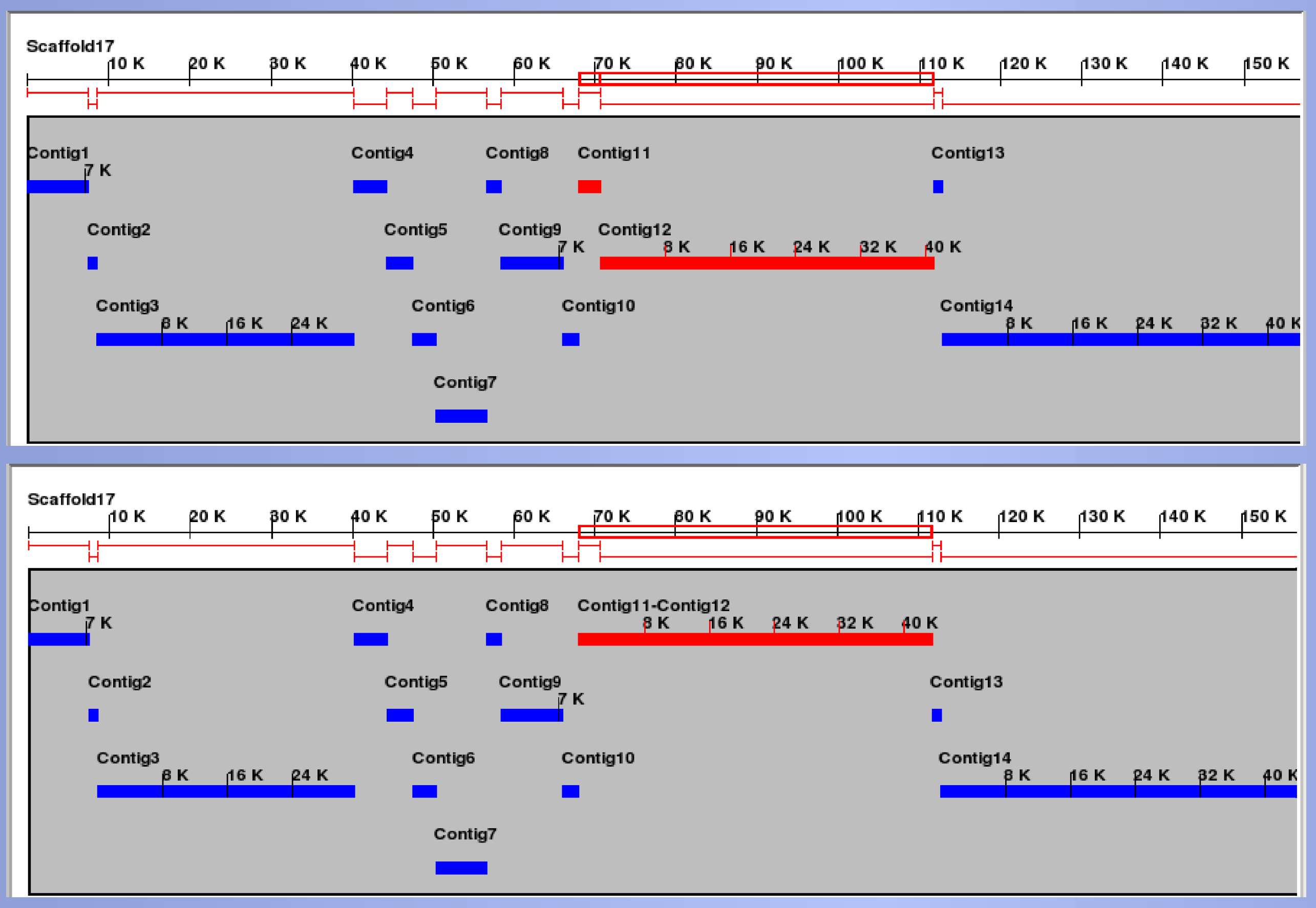


Figure 3: If a map has been determined to have been placed on the wrong parent such as a contig placed in the wrong assembly, that map can be moved to a new parent. The result of this move can be viewed. Figure 3a shows a contig that could potentially be placed on a different scaffold based on correspondences. The result of moving the map is shown in Figure 3b. Then the finisher can decide to save the move in the CMap database and export it to another program to affect the local data storage system.

Merge/Split Maps



CMAE allows a finisher to break a mis-assembled contig or merge adjoining contigs. Figure 4a and 4b show the before and after of merging a contigs. The effect can then be viewed and if the finisher decides to keep the change, it can be saved in the CMap database and exported to another program to affect the local data storage system.

Database Integration at Washinton University

The Genome Sequencing Center at Washington University will use the CMap editor and custom software to view and edit large, whole-genome assemblies. A round-trip pipeline between the Center's LIMS database and the CMap database is being implemented that will allow changes in one database to be automatically reflected in the other. In addition, custom software modules will be used to launch various viewers and analytical programs directly from the CMap interface.

Plug-ins

In order to facilitate integration of CMAE with the in-house data systems of the users, CMAE includes a GBrowse style plugin system. There are various hooks in the CMAE code where second or third party plug-ins can be attached. The plug-ins are passed object that will be most useful to them but are also given access to all of the objects in the program if a programmer wants more control. An example plug-in could insert a button in the right click menu to email a list of selected maps to another finisher to look at.

Plug-ins are stored with the application, so there is no security risk of a remote data source adding plug-ins.

Other Features

A remote server can provide data to CMAE, using the web-based CMap. This server can be configured to grant access to the application and also allow remote commits. If the remote server is password protected, CMAE will ask the finisher for a password. This is useful to allow off-site finishers the ability access to the data.

When an interesting section of the assembly is detected, a script can write an XML file to direct CMAE to view a specified set of maps. This can be used to assign various trouble regions to individual finishers.

CMAE provides the ability to undo and redo a change. This lets the finisher try a change to see how it looks without being forced to commit it.

Leveraging the CMap API

CMAE takes advantage of the CMap code base which was designed to store and quickly display correspondences between features. CMAE uses the CMap database for storing biological data. The data structure of CMap is generic as to allow for disparate biological maps to be linked with correspondences and displayed together. This reuse of code allows for quicker development by accessing the CMap application programming interface (API).

The CMap API can be used by other programs to make integration with other data storage systems easier. The CMap API allows a program to create any of the CMap data types such as maps, features and correspondences. Data can also be modified or removed.

About CMap

CMap is a web-based program used to view comparisons between various biological maps such as genetic maps and sequence maps. The strength of CMap lies in the generic database which holds a minimum amount of data to display a map with features. These features can be then linked with correspondences. This allows CMap to display disparate types of maps together in one view.

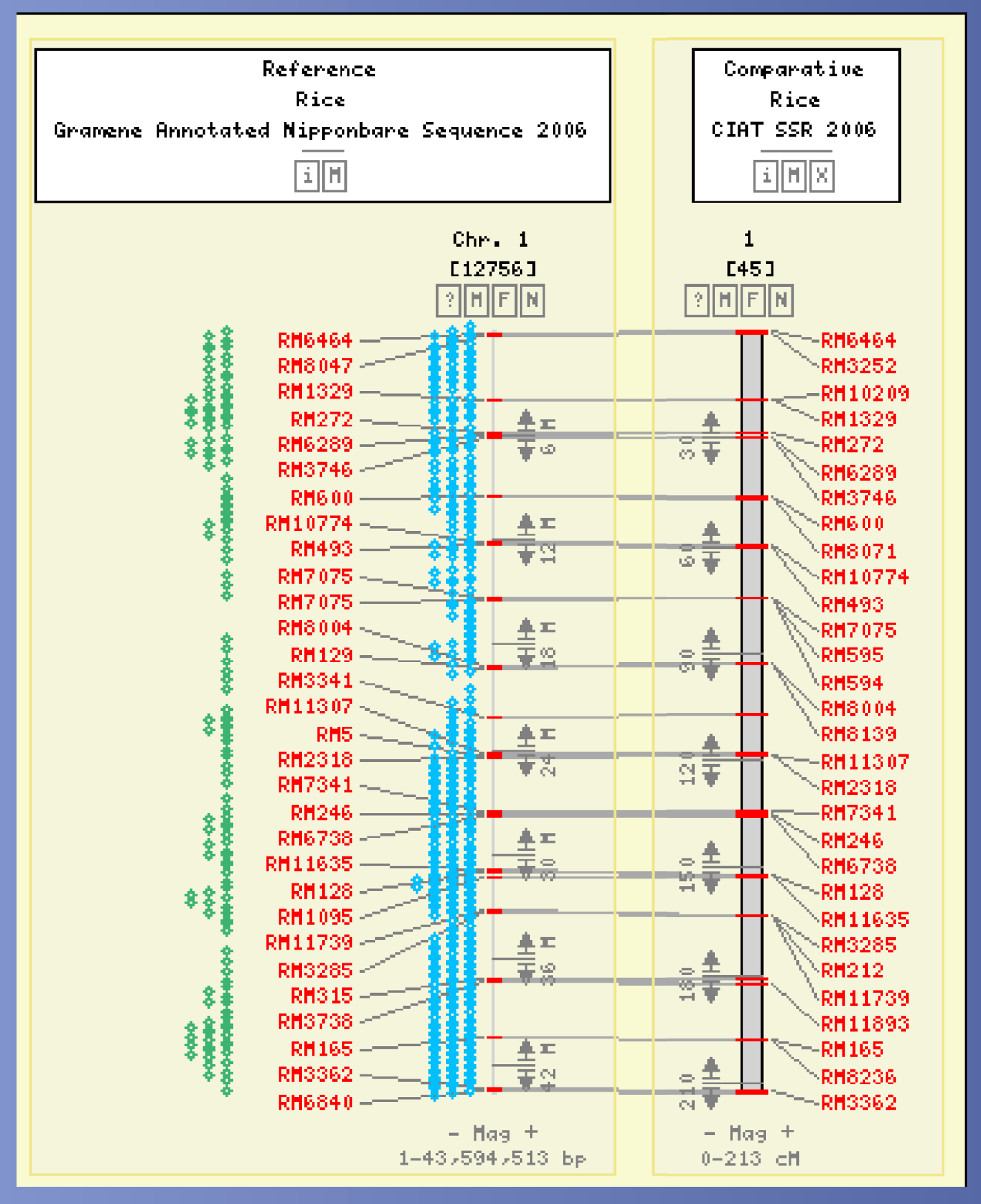


Figure 5 is an example of CMap displaying different types of data together which was taken from the Gramene Project website (www.gramene.org). In it, the sequence assembly of chromosome 1 is compare to a genetic map of chromosome1.