Exploring structural variation in the tomato genome with JBrowse

Richard Finkers, Wageningen UR Plant Breeding

Richard.Finkers@wur.nl; @rfinkers

Version 1.0, December 2013
# Table of Contents

Table of Contents ......................................................................................................................... 3

Chapter 1: Background ................................................................................................................. 4
  150+ tomato genome re-sequencing project ............................................................................... 4
  Aim ............................................................................................................................................... 4
  Accessions .................................................................................................................................... 4
  Heinz as a reference genome ....................................................................................................... 5
  Mapping en variant calling procedure ....................................................................................... 6
  JBrowse Genome browser .......................................................................................................... 7
  JBrowse generics ......................................................................................................................... 7
  JBrowse track types ................................................................................................................... 9
  The 150+ Tomato genome JBrowse .............................................................................................. 10

Chapter 2: JBrowse tutorial ........................................................................................................... 11
  Navigating JBrowse .................................................................................................................. 11
  Exploring read mapping results ................................................................................................ 12
  Exploring variant calls .............................................................................................................. 13
  Combine, Extract, and Export variant call data ........................................................................ 13

Chapter 3: Explanation of fields in SAM and VCF file’s ............................................................... 15
  Information fields, stored in bam files ....................................................................................... 15
  Fixed fields .................................................................................................................................. 15
  Genotype fields .......................................................................................................................... 16
  Information fields, stored in vcf files ....................................................................................... 17

References ..................................................................................................................................... 18
Chapter 1: Background

In this tutorial, we will explore SNP and (small) INDEL variation in data obtained from the 150+ tomato genome re-sequencing project (http://www.tomatgenome.net). The major goals are:

1. Learn about the available re-sequenced tomato, and its wild relatives, resources.
2. Understand the implications of genome re-sequencing approaches.
3. Learn how-to explore structural variation in the tomato clade

150+ tomato genome re-sequencing project

Two factors are essential for continued improvement of crop species by plant breeding: tools to identify adequate genetic variation, and technology to efficiently (re)combine useful alleles in new breeding lines. Material from wild relatives, ancestors and landraces held in germplasm collections of crop species contains an underexploited wealth of genetic variation, and will therefore offer a useful gene pool to cope with existing and new breeding challenges.

Exploiting wild and early domesticated resources has the potential to genetically enrich extant crops with alleles that can improve traits that have recently become important in the face of new challenges and requirements regarding climate change, sustainable production and a growing demand for more and better food. Once adequate genetic variation has been identified, the efficiency and success rate of breeding programs that make use of it can be greatly increased by DNA based selection of lines and markers associated with traits of interest.

Aim

The aim of the 150 Tomato Genome re-sequencing project is to reveal and explore the genetic variation available in tomato. Tomato has been selected as target crop because it is economically one of the most important crop species for the Dutch breeding industry, and is one the most important vegetables globally. However, since the tomato shows only limited genetic diversity in commercial breeding lines, valuable alleles will be available in wild tomato relatives. Since breeding and selection was targeted at only a narrow range of desirable agricultural traits, also old breeding material could be source of interesting alleles that have been lost during domestication.

Accessions

In order to identify the sequence diversity within tomato, 83 genotypes including 10 old varieties, 43 land races and 30 wild accessions will be sequenced (Figure 1). Ten accessions of S. lycopersicum var. lycopersicum and S. lycopersicum var. cerasiforme will be selected that represent the maximum range of expected genetic variation. Forty-three landraces will be selected based upon the analysis performed within the EU-SOL project, or made available by the industrial parties. The selection will include cherry, beef, round, momotaro (pink tomato) and heirloom types. A further thirty accessions will be selected from 10 Solanum section Lycopersicon species besides S. lycopersicum. These wild tomato species represent the full range of expected genetic variation around S. lycopersicum that can still be used as potential breeding material. From three wild species new reference genomes...
will be composed. Finally, 60 F$_8$ individuals will be selected from an
$S$. pinnellifolium RIL population for low depth sequencing. This set of
sequenced individuals provides a first step in tomato research to develop
tools and pipelines for “genotyping-by-sequencing” approaches, and to study
recombination events at the sequence level.

(re-)sequencing collection

Heinz as a reference genome
The reference genome for tomato cv. Heinz has been published (Sato et al.,
2012). The development of a reference genome is a tedious task, which will
take a lot of effort and requires sequencing of a lot of different libraries, at
sufficient depth. Also, more than one sequencing platform is often used for the
development of a de novo genome assembly.
The availability of a good reference genome, now means it becomes feasible
to unlock the diversity in a wider set of accessions, with relative ease. The
approach taken is mapping of reads to the reference genome instead of the
tedious tasks of construction a genome de novo. Short read (2x 100bp PE)
technology, such as the Illumina Hiseq 2000, is cost efficient methods, which
produce sufficient data for such an effort. However, it is important to realize
that this also has impact on the interpretation of your data. Issue’s that might
arise include:

1. Genomes might include regions, not present in the reference genome.
   This information will be lost. This issue already arises with relatively
   small INDELS (>20bp; depending on the settings).
2. The reference genome is the best possible prediction, however, a
   reference genome assembly also does contain many (small) assembly
   errors.
3. A re-sequenced genome might contain sequence inversions, as compared to the reference genome. Standard read mapping approaches will not detect this type of structural variation. For many studies, the outlined issues are not important. However, it is good to realize that a reference genome re-sequence mapping approach does have it limitations.

Figure 2: Outline of the process of variant calling (SNP & INDEL variation). Image taken from http://www.broadinstitute.org/gatk/about/#typical-workflows

Mapping en variant calling procedure
Raw sequencing reads are usually obtained from the sequencer in the fastq format. The pre-processing steps include quality control and quality filtering. Once these steps are completed, reads can be mapped against a genome reference with tools such as BWA (Li & Durbin, 2010), SOAP3 (Liu et al., 2012). SNP and INDEL variants can subsequently be called with tools such as SAMTOOLS (Li et al., 2009) or GATK (McKenna et al., 2010). Both tools will write the results in the so-called Variant Call Format (VCF), a text file that will contain information about all detected variants. These files can subsequently be processed with analysis tools, such as VCFTOOLS (Danecek et al., 2011) or visualized with tools such as the Integrative Genomics Viewer (Thorvaldsdóttir, Robinson, & Mesirov, 2012) or JBrowse (Westesson, Skinner, & Holmes, 2013). The outlined process give a condensed view on the steps required obtaining a set of raw variant files. To obtain high-quality variant files, the process is more complex. Figure 2 outlines steps necessary to high-quality variant files. Within the 150+ tomato genome re-sequencing project, we have thus far developed only the raw variant files. The read mapping and variant calling process takes approximately one month, on a 48CPU machine (or two years on a single core). Each additional step, easily adds one-two month to the process.
**JBrowse Genome browser**

A VCF file tends to become tediously large. For example, the VCF file containing variant data for 84 cultivated tomato accessions, and its wild relatives, is over 250 GB in size. Only few tools are effectively able to handle visualization of the data, in such a file.

**JBrowse generics**

In this tutorial, we will have a look at one of these tools, namely: JBrowse (Skinner, Uzilov, Stein, Mungal, & Holmes, 2009; Westesson et al., 2013). JBrowse is a genome browser and can be accessed via the web. The reason to choose JBrowse is because it scales relatively well and allows visualization of the diverse set of types of tracks (Figure 3).

---

**Figure 3**: Layout of JBrowse displaying different types of tracks.
Figure 4: Different sections of the JBrowse screen. The top section contains features navigational controls and shows the coordinates of the current zoom. The current coordinates box doubles as a search box, in which, for example, Solyc gene identifiers or marker names can be entered. The right site of the window shows the currently selected tracks. The remainder of the screen visualizes the data.

The JBrowse screen is organized in different sections (Figure 4). The top section of the screen can be used to navigate through the data, and to start search queries for specific regions, either by coordinates or text based searches on, for example, gene ID or marker name. Selected tracks are shown on the right. The rest of the screen is reserved for visualization of the data.

Figure 5: Select Tracks window.

After clicking on the Select tracks button (top right of the screen), a window will appear in which you can select which tracks you would like to visualize (Figure 5). A set of filters is shown on the left of the screen (e.g. Elevation or
Mating system; exact contend depends on the configuration of the browser. Cultivar names and accession ID’s can be queried using the “contains text” field in the top-middle of the page. Accessions can be selected by clicking the select box in front of the ID field.

**JBrowse track types**

JBrowse is capable to show different types of information. Some examples were already shown in Figure 3 and include:

- Reference sequence track (FASTA file)
- Gene model track (GFF file)
- Repetitive element track (GFF file)
- Marker track (GFF file)
- Assembly track (GFF file)
- Variant track (VCF file)
- Sequence coverage track (BAM file)
- Sequence alignment track (BAM file)

Additional types of tracks are possible, but currently not configured within the 150+ tomato genome re-sequence browser. The majority of the configured tracks have context-dependent pop-ups, which can be accessed by using the right mouse button click (Figure 6). The functionality of the pop-up differs per track and can include links to external databases, highlight the selected elements, open screens, which contain in-depth information of the selected data type (Figure 7).

*Figure 6: Most elements on the JBrowse have context depended pop-ups, which provide the opportunity to obtain additional information about the data.*
Figure 7: Examples of additional information, which can be obtained for sequence variants, mapped reads, or gene models respectively.

The 150+ Tomato genome JBrowse

The 150+ tomato genome JBrowse contains the following types of information:

- Reference sequence of *S. lycopersicum* cv. Heinz, version SL2.40 (Sato et al., 2012)
- Genome annotation of *S. lycopersicum* cv. Heinz (ITAG, version 2.30)
  - Gene models
  - SGN unigene assemblies ([http://solgenomics.net](http://solgenomics.net))
  - SGN markers
  - Repetitive elements
  - Positions of assembled contigs on the tomato pseudo molecules.
- SNP and INDELs, detected in the 84 cultivars, re-sequenced within the 150+ tomato genome re-sequencing project.
- SNP and INDELs (separate tracks) of *S. lycopersicum* cv. Micro tom (Kobayashi et al., 2013).
- SNP and INDELs (separate tracks) of *S. lycopersicum* accessions, sequenced by INRA France (Causse et al., 2013).
- Sequencing coverage for selected accessions.
- Sequencing reads (BAM files), for selected accessions.

The browser is configured to allow selection of accessions using one or more criteria. These criteria can include species, genebank, year of collecting, country of origin, elevation, and/or mating system. Text based search functionality is available to quickly search for a specific accession or cultivar. The 150+ tomato genome re-seq JBrowse is accessible at: [http://www.tomatogenome.net/VariantBrowser](http://www.tomatogenome.net/VariantBrowser). The data can be accesses after accepting the data usage policy.
Chapter 2: JBrowse tutorial
This tutorial consists out of a number of questions, which aims to lead you through the basics of JBrowse. Feel free to try-out your own queries and receive feedback on the topics you need.

Navigating JBrowse
Browse to the website http://www.tomatogenome.net and find the Variant browser in the menu. After accepting the data usage policy, JBrowse will be shown. Find the Select tracks button (Right-top of the screen), and click on this. This will open a window in which specific track can be selected. Please select the tracks with the following Id:

1) Gene models
2) Assembly
3) 001
4) 001 Coverage

After you finish your selection, click once more on the Select tracks button (Now on the Left-top of the screen). All selected tracks should now appear on screen. Each track can be identified because of the box, which is shown on the right site of the screen. Note: You can click-drag each box, to change the order of the tracks.

Question: Can you explain what each of these tracks represents?

Basic navigation can be done either via the mouse (hold the right mouse button and move the mouse) or via the navigation bar at the top of the screen.

Exercise: Select chromosome 5 from the dropdown box (SL2.40ch05). Zoom in to a region around 50.000.000 bp (go to these coordinates on the lower of the two coordinate bars and hold the mouse while selecting a window in this region). Also try the other buttons (e.g. zoom in, move left) from the navigation bar.

Exercise: The tomato genome annotation contains gene(models), which are annotated with Solyc IDs. For example, the Solyc ID for a gene explaining a QTL for fruit weight in tomato (Frary, 2000) is Solyc02g090730. To quickly navigate to this gene, you can enter the Solyc ID in the text box (Figure 8).

Note: The SolycID is followed by an additional number (Solyc02g090730.2.1 in this example), which indicate the version number of the annotation.

Exercise: You will now see that the browser zooms in to the gene model. To look at this gene in more detail, you have two options:
1. Right click on the gene model. This will show a pop-up menu
2. Left click on the gene model. This will execute a search query with the name of the gene model on the solgenomics network.

Exercise: The search box can be used to search other features of the annotation. For example, you can search for the marker TG25.
Note: The SGN marker track opens automatically.
Figure 8: Specific gene models can be searched by entering their Solyc ID in the search box.

**Exercise**: You might have noticed by now, that after completion of your search, the search box reverts back to show the coordinates of the selected region. You can use this functionality as well to zoom-in into a specific region. Enter for example: SL2.40ch02:18729100..18737555.

**Question**: Look at the coverage track, can you explain what is happening here? It might also be instrumental to investigate the information in the Assembly, repetitive elements (and after zoom-in further, the sequence track).

**Exploring read mapping results**

The tomato 150+ JBrowse is configured to contain read mapping information for a number of individuals. The visualization of the reads can provide detailed insight in the mapping, and also can allow investigation of potential errors. For this exercise, we will explore some this data.

Please select the region: SL2.40ch06:6599639..6599794
Please select the tracks with the following Id:
1) 001 Coverage
2) 001 Alignment

If you look at the track containing the sequence alignments (001 BAM) you see a graphical representation of the reads. The light red reads are mapped to the forward strand and the blue reads are mapped to the reverse strand.

**Question**: What is the average sequence coverage in this region?

**Question**: Do the sequences contain errors? If yes, are they random errors or systematic errors?

You can hover over the individual reads. A pop-up will appear, which shows the ID of the individual sequence. Left click on a sequence, which contains two (or more) errors and select the view details window.

**Question**: What are the quality scores of the bases, which are potential sequencing errors?

Now, select the region: SL2.40ch02:18803688..18805261
Ask yourself the same questions as previously (Starting from the question about the average coverage, on the previous page).

**Question:** Many reads in this region are bright red. Can you figure out at least two reasons why a read might be colored red?

**Question:** In contradiction to the previous region, this region contains bases, which show alternative nucleotides with good quality scores. Can you distinguish erroneous base calls from true sequence variants?

**Question:** Do you consider the identified SNPs to be in a heterozygous or homozygous state?

**Exploring variant calls**

By now, we have seen sequence reads mapped to the reference sequence, and indication of structural variation. However, it is impossible to visually call all variants. Again, there is software to aid us in this process and the results are stored in the variant call format (VCF). For all accessions, of the 150+ tomato genome re-sequencing project, VCF files are available and were used to integrate SNP and INDEL variation in this browser.

Use the select tracks option to select the variant track for moneymaker (track ID: 001).

**Question:** Does the variants called in the visualized region corresponds to the variants you detected by eye?

**Question:** Are these variants called homozygous or heterozygous? How can you quickly see this information in the variant track?

**Exercise:** Select the gene with the ID: Solyc04g010310. What can we learn about the distribution of the variants in this gene?

**Exercise:** Click on a variant and view its details. What is your interpretation of the values given in the different fields?

Please select the tracks with the following Id:

1) Gene models
2) 001
3) 072

**Question:** Can you explain the difference in the number of variants between track 001 and track 072?

**Combine, Extract, and Export variant call data**

**Exercise:** In the file menu, select the Add sequence search track. Type ATG in the search for box and press Search. Can you explain what happens?
Zoom in / out and explore the pop-up of the different elements.

**Exercise:** Make sure that you are again focused on the gene: Solyc04g010310. From the file menu, add a combination track. First, drag the lane 001 to the combination track. Secondly, drag the lane 072 to the combination track. From the menu, select XOR and press combine tracks. Zoom in to the region: SL2.40ch04:3636489..3636588

**Question:** Can you explain what is happening?

**Question:** Try the same with the other options (intersection, union, etc.). Explain the different options.

**Exercise:** JBrowse offers the possibility to save different types of data to the computer. For the different types of tracks, find the save track data button. Try to export sequence information (FASTA format), variant data (any format) and variants differing between two select individuals (any format).

This concludes the guided part of the tutorial. Please feel free to explore the other information, which is available via this JBrowse and repeat exercise of try to give an interpretation of the information yourself.
Chapter 3: Explanation of fields in SAM and VCF file's

Information fields, stored in bam files

Fixed fields

Fixed fields is a collection of tags, usually present in a VCF file.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Name of the read</td>
</tr>
<tr>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>Overall score of the mapping</td>
</tr>
<tr>
<td>Position</td>
<td>Position on the reference sequence</td>
</tr>
<tr>
<td>AM</td>
<td>The smallest template-independent mapping quality of segments in the rest</td>
</tr>
<tr>
<td>CIGAR</td>
<td>CIGAR string</td>
</tr>
<tr>
<td>MD</td>
<td>String for mismatching positions</td>
</tr>
<tr>
<td>MQ</td>
<td>Mapping quality of the mate/next segment</td>
</tr>
<tr>
<td>NM</td>
<td>Edit distance to the reference, including ambiguous bases but excluding clipping</td>
</tr>
<tr>
<td>RG</td>
<td>Read group</td>
</tr>
<tr>
<td>SM</td>
<td>Template-independent mapping quality</td>
</tr>
<tr>
<td>X0</td>
<td>Number of best hits</td>
</tr>
<tr>
<td>X1</td>
<td>Number of suboptimal hits found by BWA</td>
</tr>
<tr>
<td>XG</td>
<td>Number of gap extensions</td>
</tr>
<tr>
<td>XM</td>
<td>Number of mismatches in the alignment</td>
</tr>
<tr>
<td>XO</td>
<td>Number of gap opens</td>
</tr>
<tr>
<td>XT</td>
<td>Type: Unique/Repeat/N/Mate-sw</td>
</tr>
<tr>
<td>XA</td>
<td>Alternative hits; format (chr, pos, CIGAR, NM;)</td>
</tr>
</tbody>
</table>
**Genotype fields**

Genotype fields are optionally and are not always (nowadays usually they are), present in a VCF file. These fields are of particular importance, for VCF files in which variant data is stored for more than one genotype (e.g. the track with the name: 000, which contains variant calls for the 84 accessions of the tomato 150+ genome re-sequencing project.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
<td>Genotype; encode as allele values separated by either “/” or “</td>
</tr>
<tr>
<td>DP</td>
<td>Read depth at this position for this sample</td>
</tr>
<tr>
<td>GQ</td>
<td>Conditional genotype quality, encoded as a phred quality.</td>
</tr>
<tr>
<td>AD</td>
<td>Allelic depths for the ref and alt alleles in the order listed</td>
</tr>
<tr>
<td>PL</td>
<td>The phred-scaled genotype likelihoods rounded to the closest integer</td>
</tr>
</tbody>
</table>
### Information fields, stored in vcf files

<table>
<thead>
<tr>
<th>Tag</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>INS/DEL/DUP/INV/CNV</td>
</tr>
<tr>
<td>Score</td>
<td>Quality estimate</td>
</tr>
<tr>
<td>Description</td>
<td>Human readable description of the variant</td>
</tr>
<tr>
<td>Position</td>
<td>Position on the reference sequence</td>
</tr>
<tr>
<td>Length</td>
<td>Length of the variant</td>
</tr>
<tr>
<td>AC1</td>
<td>Allele count in genotypes, for each ALT allele, in the same order as listed</td>
</tr>
<tr>
<td>AF1</td>
<td>Allele frequency for each ALT allele in the same order as listed: use this when estimated from primary data, not called genotypes</td>
</tr>
<tr>
<td>DP</td>
<td>Number of reads covering or bridging POS.</td>
</tr>
<tr>
<td>DP4</td>
<td>Number of 1) forward ref alleles; 2) reverse ref; 3) forward non-ref; 4) reverse non-ref alleles, used in variant calling. Sum can be smaller than DP because low-quality basis are not counted.</td>
</tr>
<tr>
<td>FQ</td>
<td>Consensus quality. If positive, FQ equals the phred-scaled probability of there being two or more different alleles. If negative, FQ equals the minus phred-scaled probability of all chromosomes being identical. Notably, given one sample, FQ is positive at hets and negative at homs.</td>
</tr>
<tr>
<td>MQ</td>
<td>RMS mapping quality</td>
</tr>
<tr>
<td>VDB</td>
<td>Variant distance bias checks if variant bases occur at random positions in the aligned portion of the reads. It is useful mainly for RNA-seq reads which are aligned against a genomic reference sequence. Higher values indicate higher likelihoods that the variant is distributed within the reads randomly.</td>
</tr>
<tr>
<td>Alternative alleles</td>
<td>Description of the alternate allele</td>
</tr>
<tr>
<td>Description</td>
<td>Description of the variant</td>
</tr>
<tr>
<td>Reference allele</td>
<td>Description of the reference allele</td>
</tr>
</tbody>
</table>
References


