Case Study: Visualization of annotated DNA sequences

Tim Peeters[1], Mark Fiers[2], Huub van de Wetering[1], Jan-Peter Nap[2], and Jarke J. van Wijk[1]

[1] Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, Eindhoven, The Netherlands
[2] Applied Bioinformatics, Plant Research International, Wageningen University and Research Centre, Wageningen, The Netherlands

Abstract
DNA sequences and their annotations form ever expanding data sets. Proper explorations of such data sets require new tools for visualization and analysis. In this case study, we have defined the requirements for a visualization tool for annotated DNA sequences. We have implemented these requirements in a new and flexible tool for browsing and comparing annotated DNA sequences interactively and in real-time. The use of standard information visualization techniques, such as linked windows, perspective walls, and smooth interaction, enables genome researchers to obtain better insight in large DNA data sets in an effective, efficient, and attractive way.

Categories and Subject Descriptors (according to ACM CCS): J.3 [Computer Applications]: Biology and genetics

1. Introduction

Ongoing DNA research results in massive amounts of DNA sequence and additional knowledge on these sequences. Such knowledge is stored as *annotations* on a DNA sequence. An example of an annotation is: the sequence formed by the nucleotides at the positions 2000 up to 7000 establishes a gene. DNA sequences and their annotations can form very large data sets. The *Arabidopsis thaliana* (thale cress, a weed and model plant species) genome, for example, consists of an estimated 125,000,000 nucleotides of sequence and the public set of annotations [SZG*02] used in this article contains over 400,000 annotations. Sequences and their annotations have to be explored for research purposes. Comparisons between subsequences within and between different data sets may infer biological relevance and/or importance from DNA sequences. The mere size of the data sets requires proper tools for this analysis, which are not readily available.

This case study has been a close collaboration between plant bioinformaticians and computer scientists. The former supplied domain knowledge and data from own research as well as public databases [SZG*02]. The latter supplied knowledge on the concepts and application of information visualization. The joint efforts have resulted in a new, interactive, and real-time visualization tool for exploration and comparison of annotated DNA sequences. It facilitates the analysis of annotated genome and is called DNAvis.

2. Data sets

A DNA data set $\mathcal{D} = (\mathcal{S}, \mathcal{A})$ consists of a DNA sequence $\mathcal{S}$ and a collection $\mathcal{A}$ of annotations on $\mathcal{S}$. The sequence $\mathcal{S}$ is a string of letters from the alphabet $\{A, T, C, G, N\}$. The symbols A, T, C, and G represent the building blocks of the DNA molecule, the nucleotides adenine, thymine, cytosine, and guanine, respectively. The symbol N indicates a position in the DNA sequence where the actual nucleotide is unknown. DNA is double stranded. A given DNA sequence represents the forward strand and defines the other, so-called reverse strand, with the nucleotides replaced by their complements as follows: $A \rightarrow T$, $T \rightarrow A$, $C \rightarrow G$, $G \rightarrow C$, and $N \rightarrow N$. An interval $[i..j]$ of positions in $\mathcal{S}$ is called a *region* on $\mathcal{S}$. The DNA sequences are read from files in FASTA format [FAS].

An annotation $\alpha \in \mathcal{A}$ can be represented by the tuple

$$\alpha = (type, source, strand, regions, properties)$$

where the attributes are specified as follows:

type : *string* identifies the type, for example "Gene".
source : *string* identifies the source, for example, a public database or the program used to predict this annotation.
strand $\in \{$'+', '−', '.'$\}$ indicates whether the annotation is located on the forward strand ('+'), the reverse strand ('−'), or both strands (or is unknown) ('.').
regions defines a non-empty set of regions on $\mathcal{S}$. Although many annotations have only one region, multiple and overlapping regions per annotation are allowed. The anno-
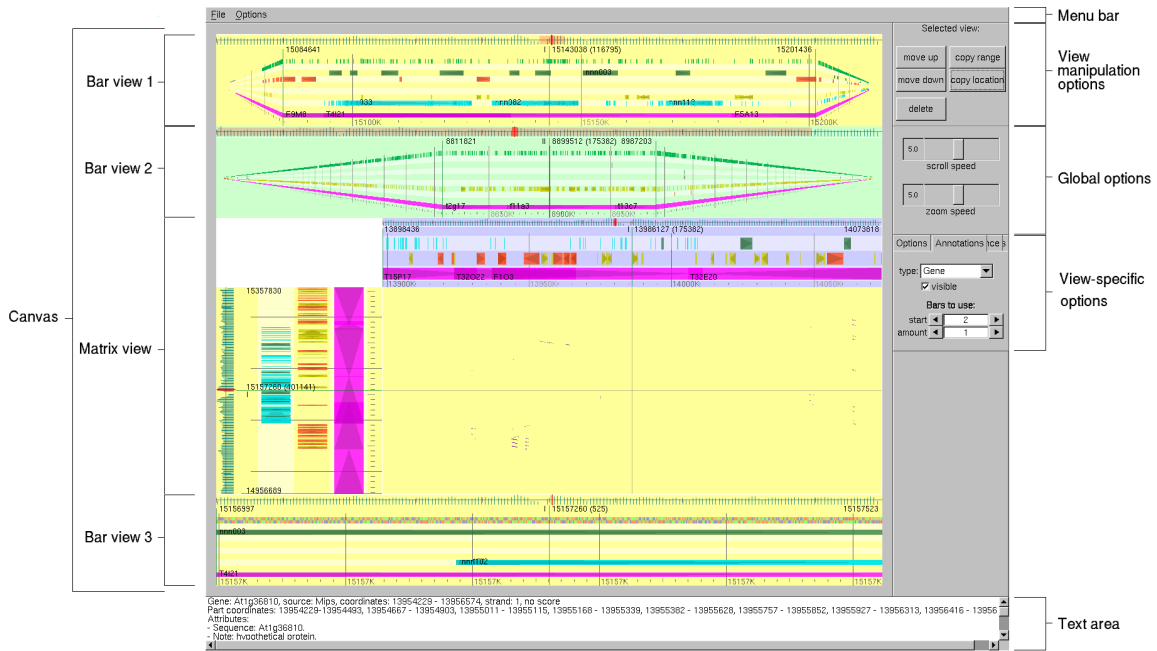
Figure 1: *A screen shot of DNAVis with notes added to identify user interface elements. DNAVis has a configurable number of views. Here four views are shown. Each bar view shows information on a region of DNA sequence and matrix views on a combination of two regions.*

tation type may dictate the number of regions. The *extent* of an annotation α is the region [*min..max*] where *min* is the minimum of all start positions and *max* the maximum of all end positions of the *regions* of α.

properties is a set of name-value pairs representing properties that are not contained in the standard attributes.

Annotations are supplied as files in GFF format [GFF]. There are many types of annotations. Three annotation types that show the large variation in annotation types are gene, miRNA_candidate, and similarity. Annotations of type "Gene" indicate the involvement of this region of DNA in transcription and translation (genes). Genes may have introns and exons. Exons are defined by the *regions* of the annotation. Sequences inside the extent of the annotation that are not exons, are introns. The name of the gene, e.g. "At1g01330", is provided as a property. The annotation type "miRNA_candidate" indicates a sequence that could belong to a novel class of non-coding regulatory elements in the cell that are associated with the regulation system known as RNA interference [BB03]. The miRNA_candidate annotation identifies two regions that represent the arms of the supposed precursor miRNA molecule. An annotation of sequence S_1 with type "Similarity", and *regions* attribute $\{r_1\}$ specifies that the region r_1 of S_1 is similar to a region r_2 of a sequence S_2. Similarity is a precomputed and flexible criterion, it can involve similarity on nucleotide level, but also

be based on expression information. The region r_2 and an identifier for S_2 are stored in the *properties* attribute.

3. Requirements

On the basis of ongoing discussions between the partners in this case study, we have defined twelve requirements for the functionality of a visualization tool for annotated genome sequences. The requirements can be divided in four different categories that concern the type of data sets included (I-III), the desired level of detail (IV-VII), the degree of interaction (VIII+IX), and the possibility for detailed and multi data set comparisons (X-XII).

Data set requirements. First of all we require the tool to be able to handle the following:

 I sequences of arbitrary length,
 II sets with an arbitrary number of annotations, and
III user-defined data sets.

Level of detail requirements. In addition, the (relative) locations and details of annotations should be shown; both are defined by the annotation's *regions*. The absolute start and end positions of the regions should be shown, but also their positions relative to regions in the same or different annotations. It should also be possible to view the nucleotides of the DNA sequence itself. For detailed visualizations, only a small part of a large sequence can be shown. Therefore,
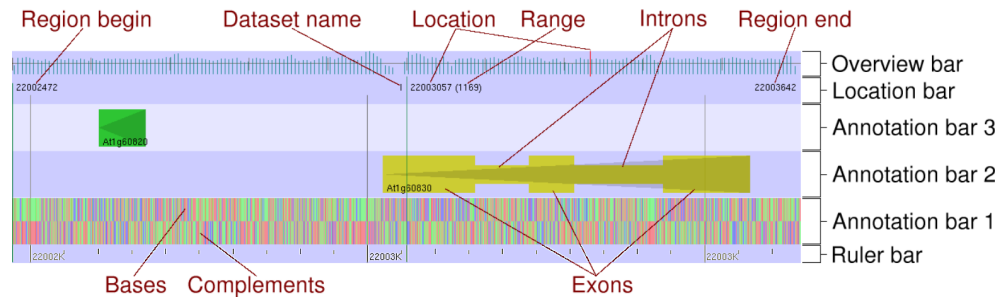
Figure 2: *A bar view showing a region of* Arabidopsis thaliana, *chromosome 1, with two annotations: At1g60820 representing a 'tRNA' annotation and At1g60830 representing a 'Gene' annotation with multiple exons and introns. See color section.*

it is important to provide context by showing the position of this part relative to the whole sequence. Hence, the following features should be included in the visualization:

  IV DNA nucleotides,
   V (attributes of) annotations,
  VI relative positions of annotations, and
 VII an overview.

     Interaction requirements. Navigation through a DNA data set by zooming and panning should be real-time, interactive, and continuous [LH02, WH00]. Furthermore, it must be possible to interactively specify which (types of) annotations are shown in which way and to re-organize the visualization such that it is the most effective for the desired analysis. This defines another two requirements:

 VIII real-time zooming & panning and
  IX flexible and responsive visualization configuration.

     Comparison requirements. To see recurrent features in DNA sequences, it must be possible to compare (sub)sequences at different levels (nucleotides, annotations, or larger) and from different data sets. Comparison of the following features between any subsequence of one or more data sets is therefore required:

   X DNA nucleotides,
  XI annotations, and
 XII global structures in data sets.

4. DNAVis

Having defined our requirements for a visualization tool for annotated DNA sequences, we have evaluated various existing tools and methods. Without showing all details, we conclude that none of the currently available tools or methods, either web-based or locally installed, complies with all requirements defined [Pee03]. Remarkable is that most of these tools miss the real-time zooming & panning (VIII) and the responsive user controlled visualization (IX) . They are mostly weak in the comparison requirements (X - XII), especially if it concerns the relationship between different levels of detail of a single data set. For example, the popular and

very useful Generic Genome Browser [SMS*02] does offer semantic zooming and annotation comparison, but zooming and panning is not real-time (VIII) and the browser does not support DNA sequence comparisons (X).

     Therefore, we have developed a DNA visualization tool that fulfills all defined requirements. DNAVis is written in C++ and runs on both Windows and Linux platforms. Figure 1 shows a screen shot of DNAVis. The largest part of the screen is taken by a canvas. The canvas contains one or more views. In each view one or more data sets are visualized. Figure 1 shows four views: three bar views and one matrix view. The nucleotide and annotation visualization is described in section 4.1. The bar view is described in section 4.2. Section 4.3 outlines the benefits of multiple bar views. The matrix view is described in section 4.4.

4.1. Nucleotides and annotation visualization

Nucleotides and annotations of a DNA sequence are shown by both icons and text. These elements are ordered along a horizontal axis, the so-called *sequence axis*. This allows stacking of multiple axes, economical use of screen space, and rendering of text in natural read order. The use of semantic zooming [BH94] accomplishes that some attributes are only rendered if space permits. This improves both the rendering speed and the clarity of the image.

     Representations of the nucleotides are drawn along the sequence axis. Each nucleotide is rendered as a colored rectangle with the corresponding letter, *A*, *C*, *T*, or *G*, inside. Figure 2 shows both a DNA sequence and its complement. If space becomes limited, first the letters and upon further limitation the colored rectangles disappear.

     Annotations are also displayed along a sequence axis by visualizing several of their attributes (see figure 2):

- The *type* is visually represented by a configurable color.
- The *regions* of an annotation (e.g. the exons of genes) are represented with thick rectangles.
- The *extent* of an annotation, spanning all its regions, is shown as a thin rectangle connecting all thick rectangles.
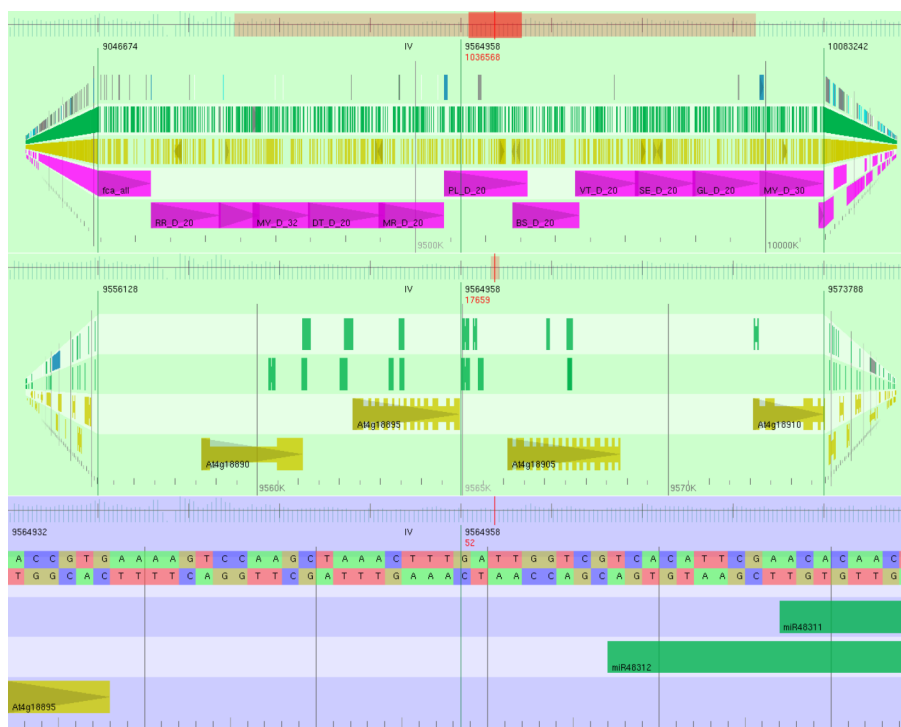
Figure 3: *Three views that show the same location (around the 10 millionth nucleotide) on chromosome 4 of* Arabidopsis *but each on a different level of detail. All views have multiple bars with different layouts for the annotations. The top view shows approximately 1 million nucleotides of the data set. The center view shows a region of 17659 nucleotides and the bottom view shows the actual nucleotides and their complements on a sequence region of 52 nucleotides.*

- The *name* of the annotation is shown inside its extent.
- The forward and reverse *strand* are drawn as transparent gray triangles pointing to the right or the left, respectively.

Using semantic zooming, if an annotation is reduced to only a few pixels its strand and subdivision in regions are not drawn. If the user clicks on an annotation, all textual information for that annotation is shown in the text area (see Figure 1). This allows analysis of exact regions or additional properties that are not visualized.

4.2. Bar view

A bar view shows information of a single data set on a horizontal sequence axis; the vertical axis is used for stacking information. To do this, each bar view is partitioned into *bars*, each of which is a horizontal strip that covers the view's full width (see Figure 2). A bar view and each of its bars visualize the same region *r* of the same data set. Information on this region and data set are shown in three separate bars:

1. The *overview bar* represents the whole data set by a histogram showing the density of the annotations in the data set. It highlights the current location and region with a red line and a transparent red rectangle.

2. The *location bar* displays the name of the data set and gives the exact position and size of the region.

3. The *ruler bar* shows ruler-like vertical lines at the bottom of the view indicating the scale of the visualization. This scale indication automatically adapts to the range used.

Annotations are displayed on *annotation bars*. The user can configure the rendering of these bars per bar view by setting the number of annotation bars and assigning any range of annotation bars to each of the annotation types. Hence, multiple types of annotations can be shown on one annotation bar and annotations of one type can be spread over multiple annotation bars. The former can be useful to save space, especially if the concerned annotations do not overlap, the latter is useful if annotations of one type do overlap. In Figure 3 an example is shown where genes and miRNAs each have a single bar in the top view and are each spread over two bars in the middle view.

Any bar view can use a perspective wall [MRC91]. By allowing parts of a bar view to have less detail perspective walls can show more of a data set at once and consequently improve context awareness. Figures 1 and 3 show bar views with perspective walls. The interaction and navigation implemented in a bar view with respect to panning and zoom-
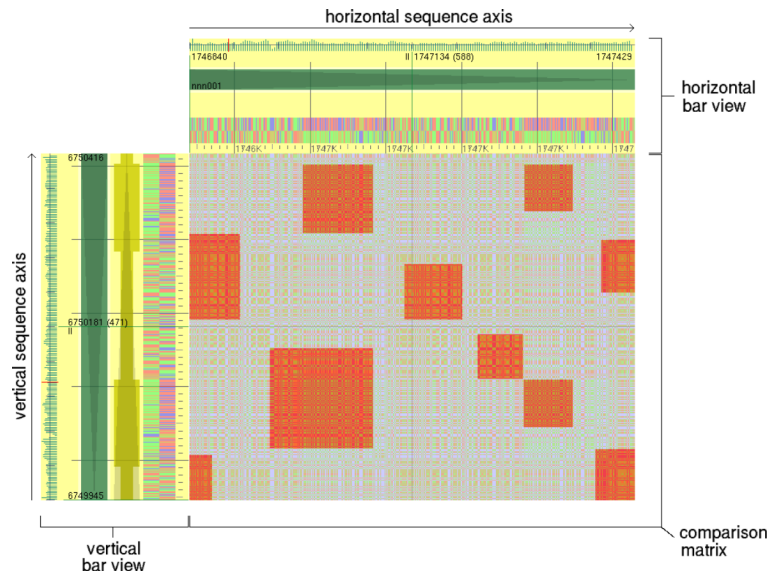
Figure 4: *A matrix view comparing two different regions of the* Arabidopsis *genome. The large dark rectangles in the comparison matrix indicate a precalculated large scale similarity between the two regions. On a smaller scale the results of a nucleotide comparison, which is calculated on-the-fly, is shown.*

ing is continuous. Horizontal dragging of the bar leads to panning left or right, vertical dragging to zooming in or out.

4.3. Multiple bar views

The DNAVis canvas can contain multiple views (see Figure 1) which the user can add and remove. Each view has independent settings for the data set(s) it displays, the current region(s) it shows, the number of annotation bars, the distribution of annotations as well as the DNA sequence over the annotation bars. Moreover, views can be linked [JBO94]. When panning *n* nucleotides to the left or right in the selected view, all linked views follow independent of their location and level of detail. Also when zooming in or out the linked views zoom with the same factor. The range and/or the location of a view can be copied to the views linked to it. This facilitates browsing and comparing multiple data sets at the same level of detail as well as browsing the same data set at multiple levels of detail (See Figure 3).

4.4. Matrix view

For comparing data sets DNAVis also has a matrix view (see Figure 4). This view is used for visualizing precalculated similarity annotations and allows on the highest level of detail direct comparison of sequence data. The matrix view consists of a comparison matrix, a horizontal bar view, and a vertical bar view. These bar views indicate the compared regions of the data sets. Inside the comparison matrix, similarities between these regions are shown as color-coded rectangles. Regions with a similarity annotation are colored red.

If the level of detail is high enough, identical nucleotides in the compared regions are shown as orange rectangles, resulting in a dot plot [CH93]. Complementary nucleotides in these regions are green.

Because nucleotide comparisons do not overlap and they are drawn slightly transparent over higher-level similarities, all three types of similarities can be shown in one image. Figure 4 shows a matrix view comparing two parts of *Arabidopsis thaliana* chromosome 2. Navigating the two data sets and their comparison can be done in two ways: by dragging and zooming of the comparison matrix or by manipulating the bar views. The user can interactively browse through the comparison and change the level of detail in a continuous way. Finally, the matrix view can be combined and linked with other views.

5. Results

The DNAVis tool offers a fast and highly flexible method to visualize large genomes and their annotations. The visualization features accomplished contribute to a better analysis and interpretation of biological relevance. Different types of annotations can be generated, for example for miRNA candidates, possible chromatin boundaries and/or existing splice variants. Visual inspection with DNAVis makes it easier to relate these annotations to position and underlying DNA sequence. This allows investigating common biological questions about chromosomal distribution as well as the relationship between annotations.

For example, the miRNA candidate Mir10733 (un-
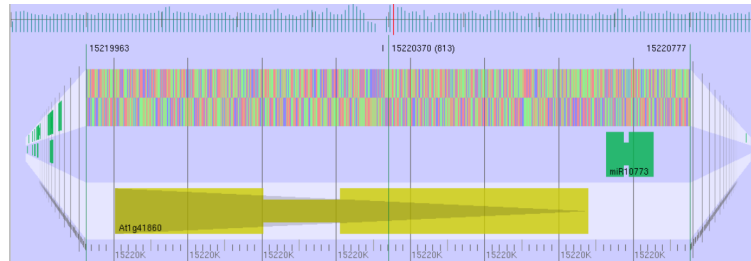
Figure 5: *A miRNA candidate shown to be directly downstream of a gene.*

published data) is located directly downstream of gene At1g41860 (Figure 5). This may indicate that the expression of At1g41860 affects expression of the target genes of Mir10733. In this way, the visualization accomplished by DNAVis helps to generate new hypotheses relating position to function that can be investigated. Notably simultaneously viewing multiple areas of different genomes supports comparative studies on genome organization and activity. For example, incompletely sequenced genomes of bacteria may be compared to fully sequenced genomes to identify potential regions or genes of interest. A current disadvantage of DNAVis is that all annotations must be supplied in advance in GFF format. When coupled to a database, future versions of DNAVis may deal in a more friendly and flexible way with annotations. This allows direct error correction in genome annotations, without compromising performance. The ease of visualization is then helpful in improving annotations of genome data. Particularly exciting is the future inclusion of gene expression data as annotation. This poses the additional challenge of visualizing tissue and/or organ specific expression, or expression that is developmentally controlled. Easy browsing a genome for genes expressed only in root, or only during leaf development helps biologists even more to explore structure and function of a genome.

Finally, DNAVis meets the requirements specified in section 3 but it is to be expected that further usage of the tool will generate additional requirements, triggered by additional annotation types or new biological insight in DNA.

References

[BB03] BARTEL B., BARTEL D.: MicroRNAs: At the root of plant development. *Plant Physiology 132* (2003), 709–717. 2

[BH94] BEDERSON B. B., HOLLAN J. D.: Pad++: A zooming graphical interface for exploring alternate interface physics. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST'94)* (1994), ACM Press, pp. 17–26. 3

[CH93] CHURCH K. W., HELFMAN J. I.: Dotplot: A program for exploring self-similarity in millions of lines of text and code. *Journal of Computational and Graphical Statistics 2*, 2 (1993), 153–174. 5

[FAS] FASTA file format. tigrblast.tigr.org /web-hmm /fasta.html. Valid February 2004. 1

[GFF] General feature format. www.sanger.ac.uk /Software/formats/GFF. Valid February 2004. 2

[JBO94] JACOBSON A. S., BERKIN A. L., ORTON M. N.: LinkWinds: interactive scientific data analysis and visualization. *Commun. ACM 37*, 4 (1994), 42–52. 5

[LH02] LORAINE A. E., HELT G. A.: Visualizing the genome: techniques for presenting human genome data and annotations. *BMC Bioinformatics 3*, 1 (2002), 19. 3

[MRC91] MACKINLAY J. D., ROBERTSON G. G., CARD S. K.: The perspective wall: detail and context smoothly integrated. In *Proceedings of ACM CHI'91* (1991), pp. 173–179. 4

[Pee03] PEETERS T.: *Interactive visualization of annotated DNA sequences.* Master's thesis, Technische Universiteit Eindhoven, 2003. 3

[SMS*02] STEIN L., MUNGALL C., SHU S., CAUDY M., MANGONE M., DAY A., NICKERSON E., STAJICH J., HARRIS T., ARVA A., LEWIS S.: The generic genome browser: a building block for a model organism system database. *Genome Res. 12*, 10 (Oct. 2002), 1599–1610. 3

[SZG*02] SCHOOF H., ZACCARIA P., GUNDLACH H., LEMCKE K., RUDD S., KOLESOV G., ARNOLD R., MEWES H., MAYER K.: MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res. 30*, 1 (2002), 91–93. 1

[WH00] WICKENS C. D., HOLLANDS J. G.: *Engineering Psychology and Human Performance.* Prentice Hall, 2000. 3