

Combining phenotypic and genotypic data – issues and opportunities

Reinhard Simon, Lukas Mueller

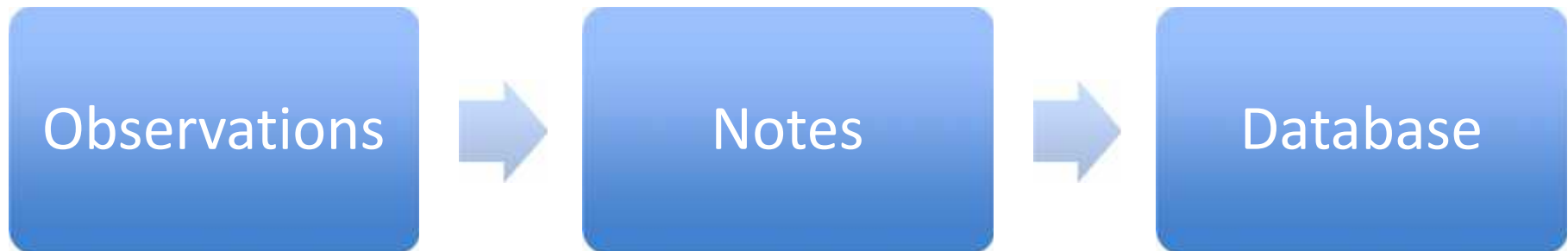
The Annual Sweetpotato Breeders Meeting,
Colline Hotel, Mukono, Uganda, June 2-5, 2015

The 'pull'

- Why an integrated database?
- Old issues
- New issues
- New options
- Example integrative uses

Why? - 1

- An information highway
 - for breeders
 - Or (thinking about our true clients)
 - to increase acceptability of new varieties
 - Monitoring KPI



Why? - 2

- As a 'memory'
 - Baseline data to
 - Monitor genetic gain
 - Track materials (Reference lists)
 - Track environmental conditions (climate change)
 - Re-use of raw data
 - Re-use of NIRS data when new calibrated compound becomes available
 - Re-use of SNP data when new effects are established



"Remember me, Mr. Schneider? Kenya, 1947. If you're going to shoot at an elephant, Mr. Schneider, you better be prepared to finish the job."

I don't want your calculations – I want your observations.

Isaac Newton

Why? - 3

- As a collaboration tool
- Who is doing what, when, where, how, why
 - Across institutions
 - Across borders
 - Across generations of breeders
 - Across crops
 - Across communities of practice (breeders, crop modelers, social scientists, nutritionists, economists, ecologists, phytopathologists, ..., stakeholders, donors)

Why? - 4

- As an integration platform
 - Linked data!
 - Different aspects of a
 - Variety
 - Breeding parent
 - Consumer community/market segment
 - Facilitate discoverability

Why? - 5

- As a base for data mining /market intelligence
- *The whole is more than the sum of the pieces.*
 - Shifts in variety performance (breakdowns)
 - epidemiology
 - Shifts in local consumer preferences
 - Regional production trends
 - Shifts in local/regional climate
 - ...

Old issues - 1

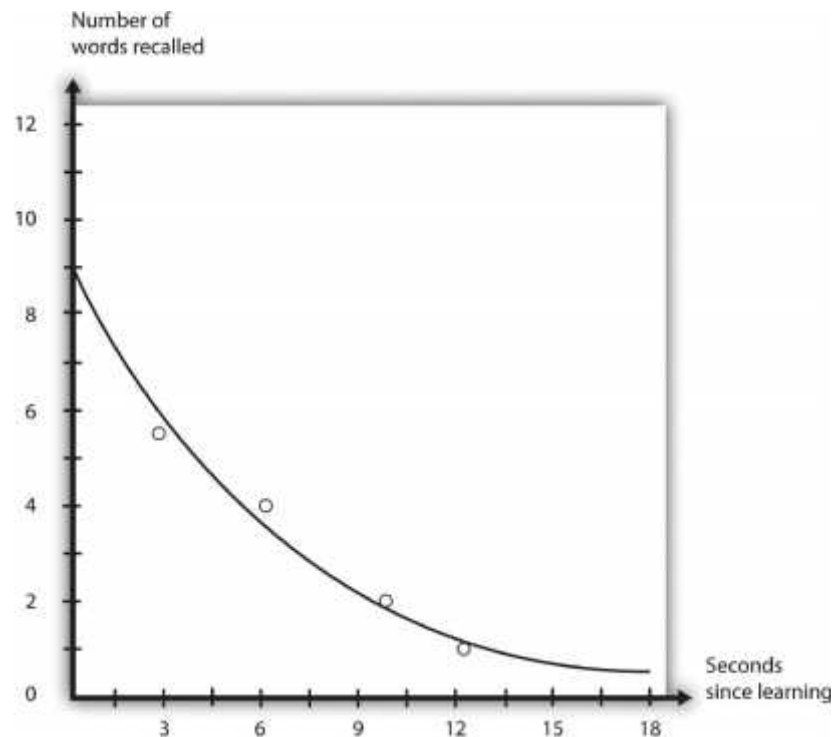
- Manage identity
 - $A = A?$
 - Pro-active: provide reference lists
 - Retro-active: Use of characterization and evaluation data to check if true-to-type

Old issues - 2

- Manage variation in data
 - Promote best practices
 - Standard concepts/ontologies
 - Standard measurements/protocols
 - Standard documentation/Meta-data
 - Standard structure/forms
 - The less variability between batches the higher the quality.
 - Consistency and completeness

Old issues - 3

- Information decay
 - The fresher the better: use data now!
 - If you don't use it, it will fade away.
 - Need to get data quickly in and out!



Old issues - 4

- Access and availability
 - No more 'data hugging' – open access

The wise man does not lay up his own treasures. The more he gives to others, the more he has for his own.

Lao Tzu

Old issues - 5

- Performance / responsiveness
 - Good physical infrastructure (server, bandwidth, ...)
 - Database user interface optimized for search

New issues - 1

- High-throughput technologies
 - Cheaper to re-do than to store
 - “Industrialization of biology”

New issues - 2

- High volume
 - Big data – although 'big' is a relative term (like NG)

Data tsunami today, tomorrow a storm in in a water glass.



New issues - 3

- High dimensionality: '-omics'
 - Genomics
 - Phenomics
 - Metabolomics
 - Transcriptomics
 - Proteomics
 - ...

New issues - 4

- Real-time: data anywhere anytime
 - Need to get better & faster at
 - Integration
 - Analysis
 - Decision making

New issues - 5

- Connecting scales
 - From SNPs to satellite imagery
 - From preferences to phenotype
 - From consumers profiles to choice of breeding parents

(New) options - 1

- The web is the way!
 - Connect ideas / information resources
 - Cloud computing
 - Operating system independent
 - Device independent
 - The browser is the new office

Options - 2

- Newer standards for concepts
 - Ontologies:
 - a basic tool for automated knowledge creation
 - Facilitates logical operations on words.

Whereof you cannot speak, thereof you must be silent.

Ludwig Wittgenstein

Options - 3

- Newer devices
 - Tablets, smartphones, ...
 - New ways of accessing, interacting, and creating information

Options - 4

- Newer user interaction paradigms
 - Based on navigation: zoom (drill), pan
 - Based on touch (point) gestures
 - Based on speech recognition

Options - 5

- Increasing automation of computation



Options - 5

- Increasing automation of computation

HOW DO YOU SOLVE THE ANOVA USING A SCIENTIFIC CALCULATOR?

Solution:

$$\text{BSS} = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} + \frac{(\sum X_4)^2}{n_4} - CF$$

$$= \frac{(37)^2}{7} + \frac{(57)^2}{7} + \frac{(26)^2}{7} + \frac{(36)^2}{7} - 869.19$$

$$= 195.57 + 464.14 + 96.57 + 185.14 - 869.14$$

$$\text{BSS} = 72.28$$

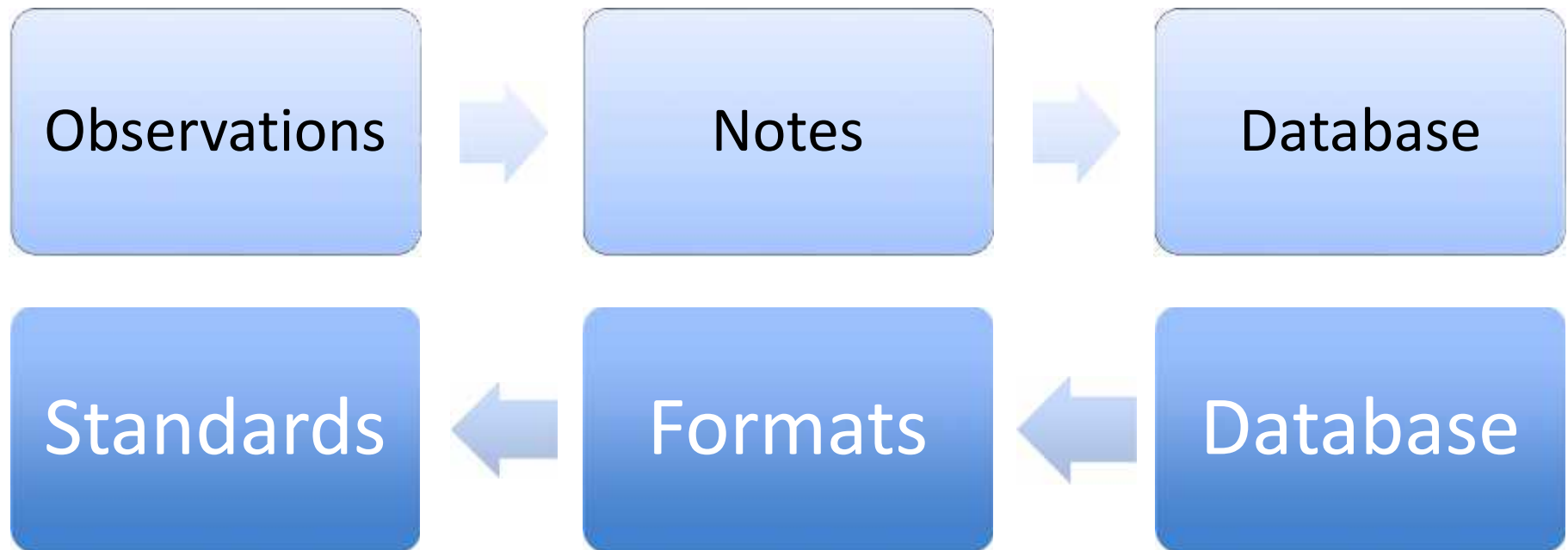
$$\text{WSS} = \text{TSS} - \text{BSS}$$

$$= 144.86 - 72.28$$

$$\text{WSS} = 72.58$$

How do we get there?

Data flow – give and take



Summary

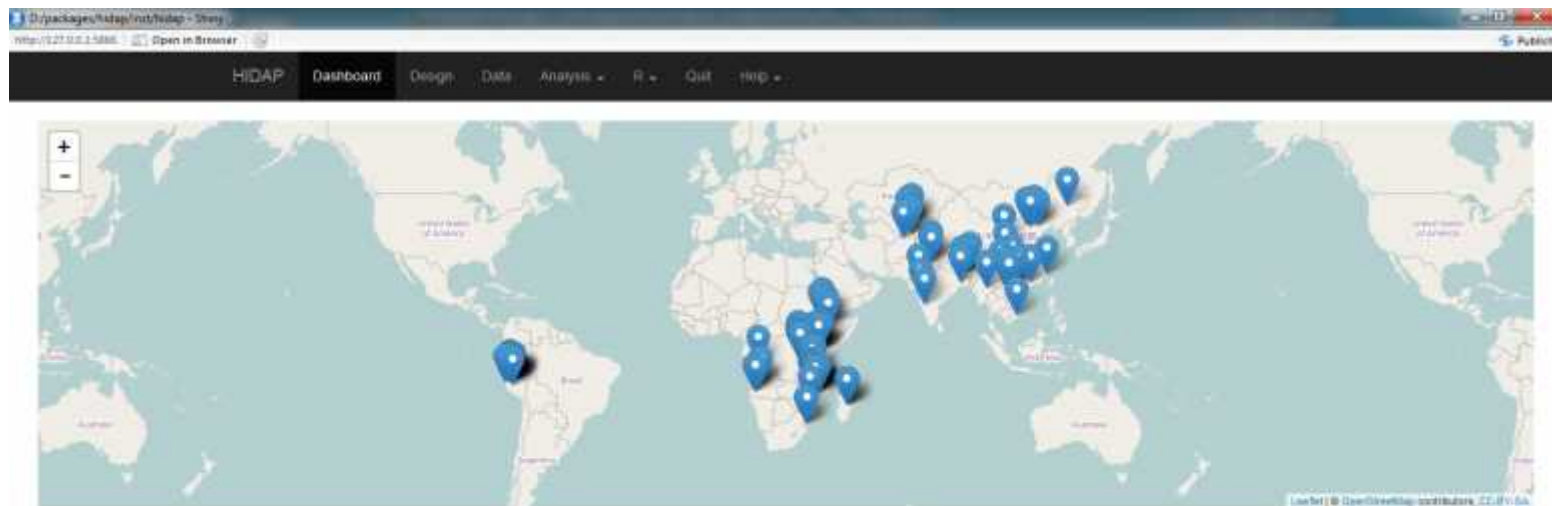
- **Why:**
 - information highway, memory, collaboration, integration, re-purposing
- **Old issues:**
 - Identity assurance, reduction of variability, Information decay, access, performance
- **New issues:**
 - High throughput, high volume, high dimensionality, real-time, connecting scales
- **Options**
 - Web 3.0, ontologies, devices, interaction paradigms, automation of analysis

A first stab at a web-based analytical platform – part I

- Based on open source statistical platform (R, agricolae package)
- + current web technologies (up-coming standards)
 - Means will be available with current browsers
 - API (a way to exchange data over the web between apps)
- Works both online and offline
- + Current databases (Lukas)

Examples - 1

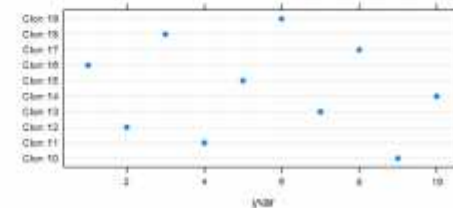
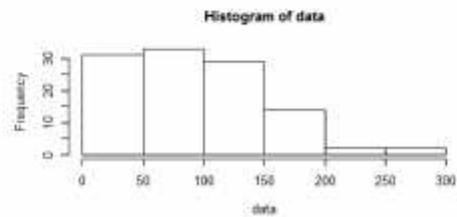
- High level information integration - dashboard



Site

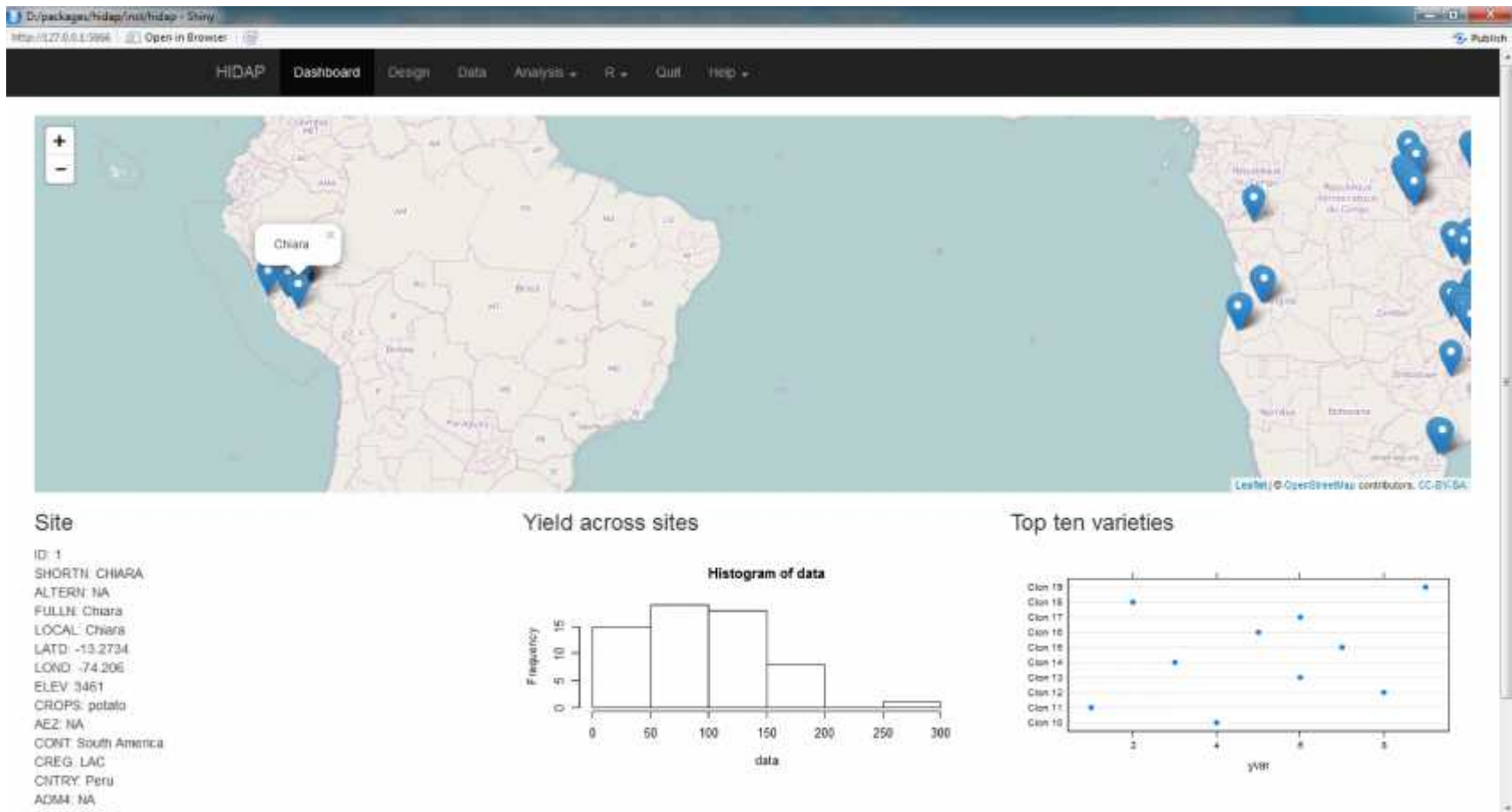
Yield across sites

Top ten varieties



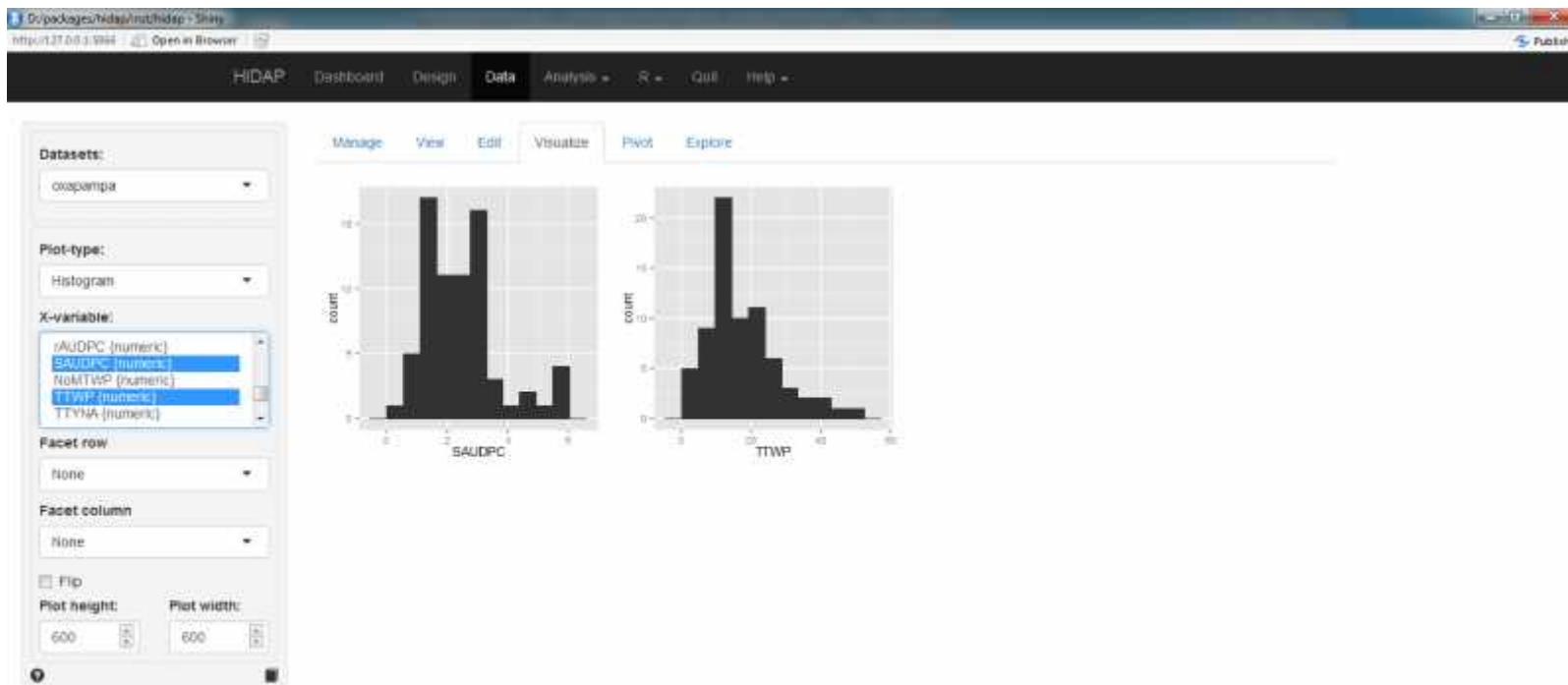
Examples - 2

- Linked graphs



Examples - 3

- Graphical exploration of data



Examples - 4

- Automated reports

```
1: |
2: * | { setopt,echo=FALSE, results='hide', message=FALSE}
3: | library(datacheck)
4: | library(etable)
5: | options(etable.type = 'html')
6: * |
7: |
8: | A vignette for "datacheck" (version 'r pkg_version' 'datacheck') |
9: |-----|
10: | Heinrich Simon, International Potato Center, Lima, Peru
11: |
12: | The library "datacheck" provides some simple functions to check the consistency of a dataset. It
13: | assumes data are available in tabular format - typically a csv file with
14: | objects or records in rows and attributes or variables in the columns.
15: |
16: | In a database setting the variables would be controlled by the database - at least
17: | conformance to types (character, numeric, etc) and allowed min/maximum values.
18: | However, often data are gathered in simple spreadsheets or are for other reasons
19: | without such constraints. Here, data constraints like allowed types or values, expected
20: | values and relationships can be defined using R commands and syntax. This allows such
21: | more flexibility and fine grained control. Typically it demands also a lot of domain
22: | knowledge from the user. It is therefore often useful to re-use such domain aware rule files
23: | across tables with similar content. Therefore this tool is forgiving if rules cannot be executed
24: | if a variable is not present in the table to be analyzed allowing the reuse of such rule files.
25: |
26: | Using the HTML interface
27: |-----|
28: |
29: | Use the following commands to copy some example files to your current working directory
30: | (uncomment the file.copy commands):
31: |
32: | {
33: |   etable = system.file("examples/soilexamples.csv", package="datacheck")
34: |   styles = system.file("examples/soil_rules.R", package="datacheck")
35: | }
36: | # Uncomment the next two lines
37: | # file.copy(etable, "soilexamples.csv")
38: | # file.copy(styles, "soil_rules.R")
```

Examples - 4

- Automated reports

The screenshot displays a Shiny web application interface. On the left, there is a code editor with R code. The code includes comments and commands for setting up example files and running a data check. On the right, there is a text area with instructions and a plot titled "Rules per variable".

```
22 IF a variable is not present in the table to be analyzed allowing the reuse of such rule files.
23
24 Using the HTML interface
25 -----
26
27 Use the following commands to copy some example files to your current working directory
  (uncomment the file.copy command):
28
29 [...]
29 enable = system.file("examples/soilexamples.csv", package="datacheck")
30 srules = system.file("examples/soil_rules.R", package="datacheck")
31
32 # Uncomment the next two lines
33
34 # file.copy(stable, 'soilexamples.csv')
35 # file.copy(srules, 'soil_rules.R')
36
37
38 Then type in the command 'run_datacheck()' in the R editor.
39
40 Use the upload buttons to load the respective files in your working directory.
41 Review the results.
42
43
44 Give the command line interface
```

Typically it demands also a lot of domain knowledge from the user. It is therefore often useful to re-use such domain aware rule files across tables with similar content. Therefore this tool is foregoing if rules cannot be executed if a variable is not present in the table to be analyzed allowing the reuse of such rule files.

Using the HTML interface

Use the following commands to copy some example files to your current working directory (uncomment the file.copy command):

Then type in the command `run_datacheck()` in the R editor.

Use the upload buttons to load the respective files in your working directory. Review the results.

Using the command line interface

Assuming you have copied the above mentioned files in your working directory proceed to read in the data.

You can inspect a graphical summary of rules per variable:

Rules per variable

Variable	Number of Rules
Average rules	2
pH	2
Longitude	2
ID	2
Altitude	2
Soil_texture	2
Clay	2
Lime	2
Sand	2
P	2
CaCO3	2
Conductivity	2
Adm3	2
Adm2	2
Adm1	2
Country	2
Latitude	2
Organic_matter	1

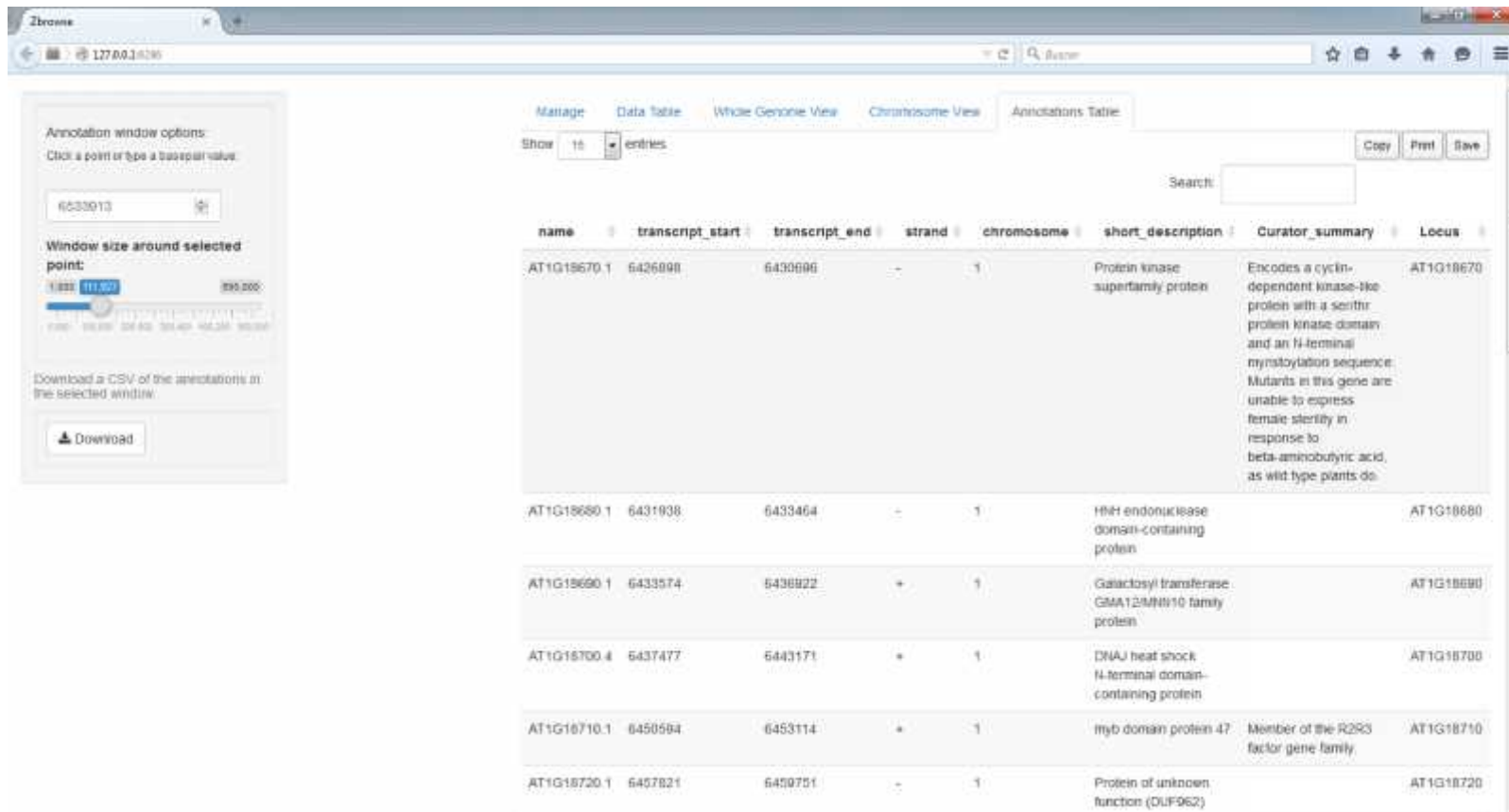
Examples - 5

- Zooming in: from QTLs to genes (Zbrowse)



Examples - 5

- Zooming in: from QTLs to genes (Zbrowse)



The screenshot shows the Zbrowse web interface. On the left, there are options for annotation window settings, including a search box with the value '6533013' and a window size slider around a selected point. Below this is a 'Download' button for a CSV of annotations. The main area displays a table of gene annotations with columns for name, transcript_start, transcript_end, strand, chromosome, short_description, Curator_summary, and Locus. The table lists several genes, with the first one, AT1G18670.1, having a detailed description in the Curator_summary column.

name	transcript_start	transcript_end	strand	chromosome	short_description	Curator_summary	Locus
AT1G18670.1	6426098	6430696	-	1	Protein kinase superfamily protein	Encodes a cyclin-dependent kinase-like protein with a serine protein kinase domain and an N-terminal myristoylation sequence. Mutants in this gene are unable to express female sterility in response to beta-aminobutyric acid, as wild type plants do.	AT1G18670
AT1G18680.1	6431938	6433464	-	1	HhH endonuclease domain-containing protein		AT1G18680
AT1G18690.1	6433574	6436822	+	1	Galactosyl transferase, GMA12/MNH10 family protein		AT1G18690
AT1G18700.4	6437477	6443171	+	1	DNAJ heat shock N-terminal domain-containing protein		AT1G18700
AT1G18710.1	6450594	6453114	+	1	Myb domain protein 47	Member of the R263 factor gene family.	AT1G18710
AT1G18720.1	6457821	6459751	-	1	Protein of unknown function (DUF962)		AT1G18720