Combining phenotypic and genotypic data – issues and opportunities

Reinhard Simon, Lukas Mueller

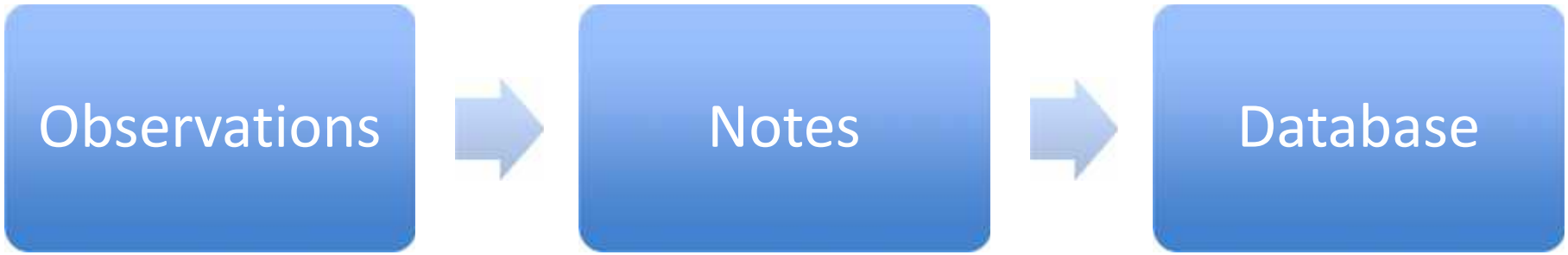
The Annual Sweetpotato Breeders Meeting,
Colline Hotel, Mukono, Uganda, June 2-5, 2015

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

"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."

- 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

I don't want your calculations – I wan't 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, nutrionists, 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

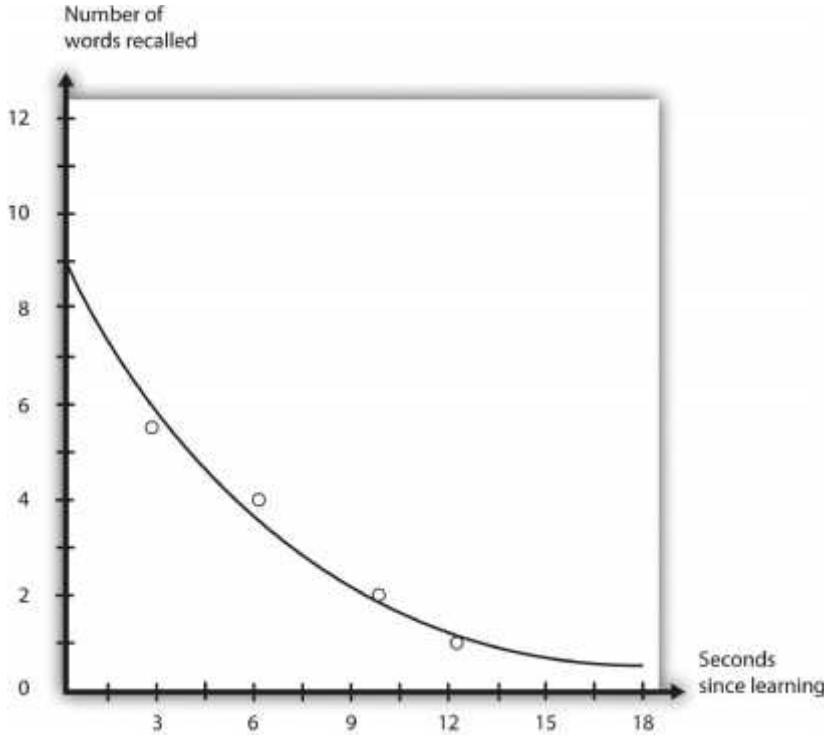
- Manage identity
  - A = A?
  - Pro-active: provide reference lists
  - Retro-active: Use of characterization and evaluation data to check if true-to-type

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

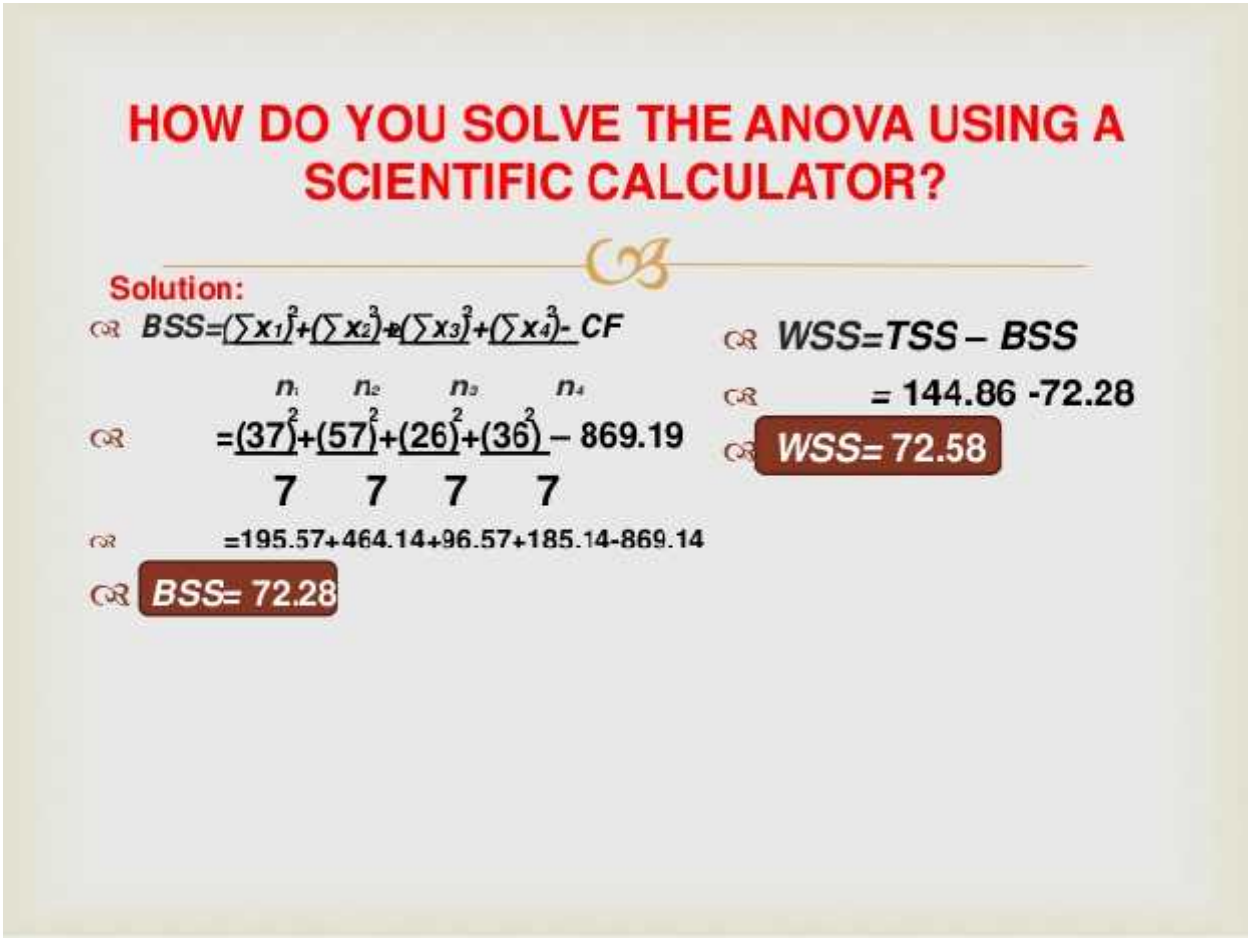
- Newer user interaction paradigms
  - Based on navigation: zoom (drill), pan
  - Based on touch (point) gestures
  - Based on speech recognition

Options - 5

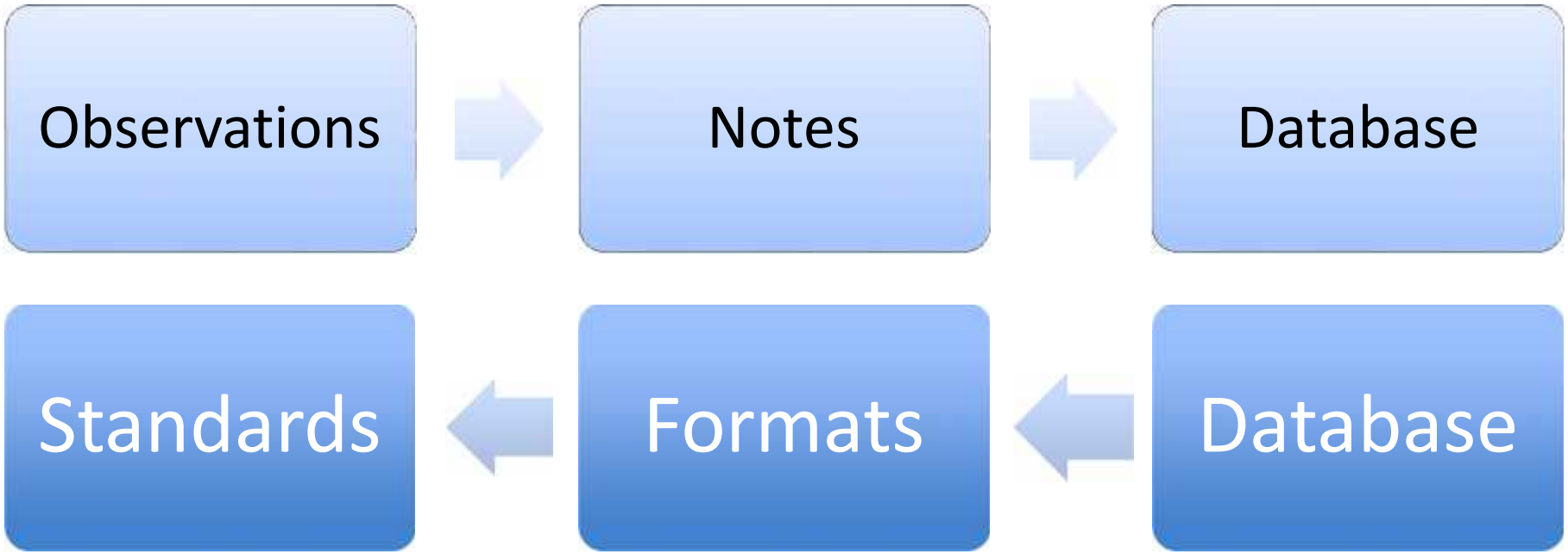
- Increasing automation of computation

Options - 5

- Increasing automation of computation



How do we get there?
Data flow – give and take



Summary

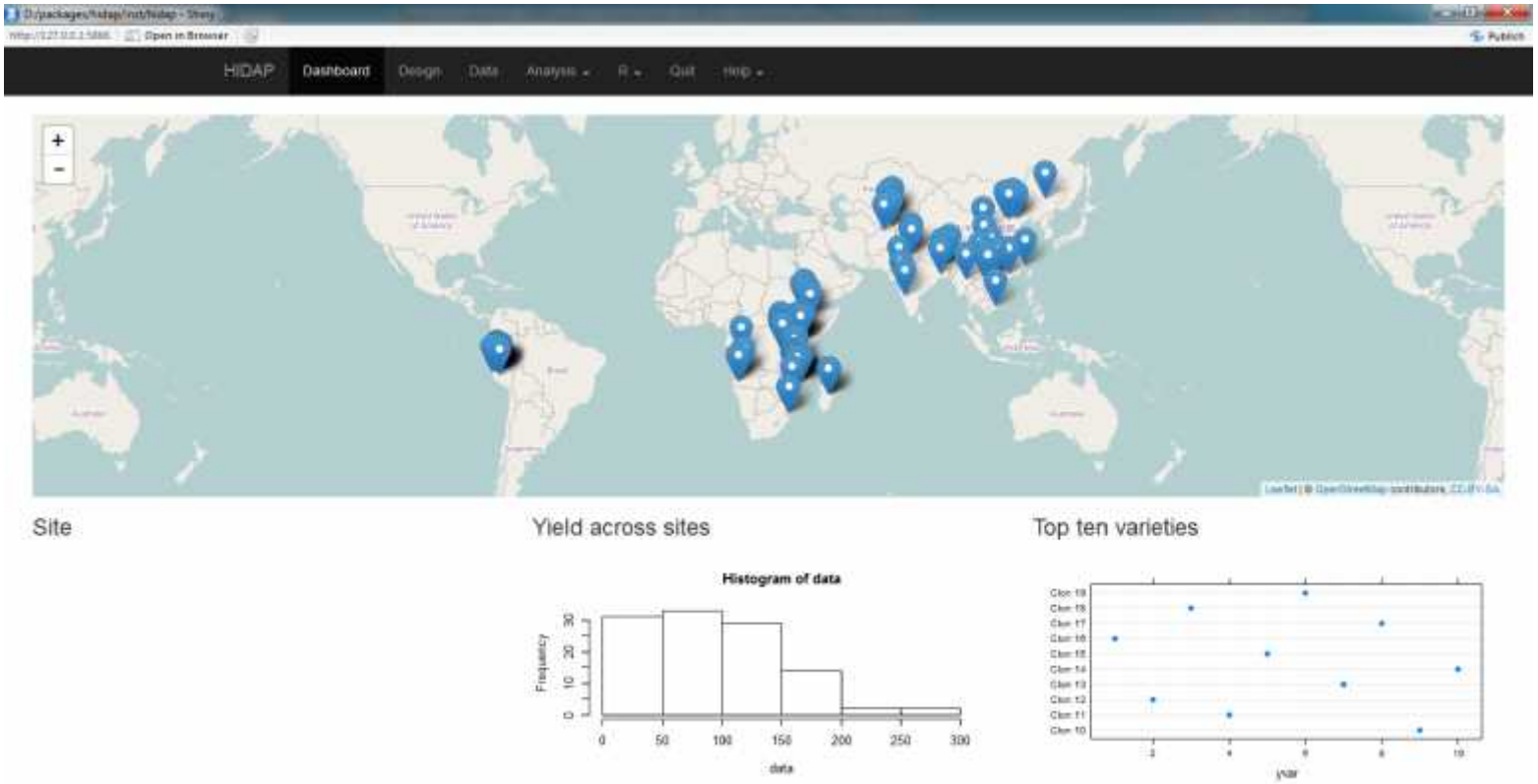
- Why:
  - information highway, memory, collaboration, integration, re-purposing
- Old issues:
  - Identity assurance, reduction of variability, Information decay, access, performance
- New issues:
  - High throughput, high volume, high dimensionality, real-time, connecting scales
- Options
  - Web 3.0,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

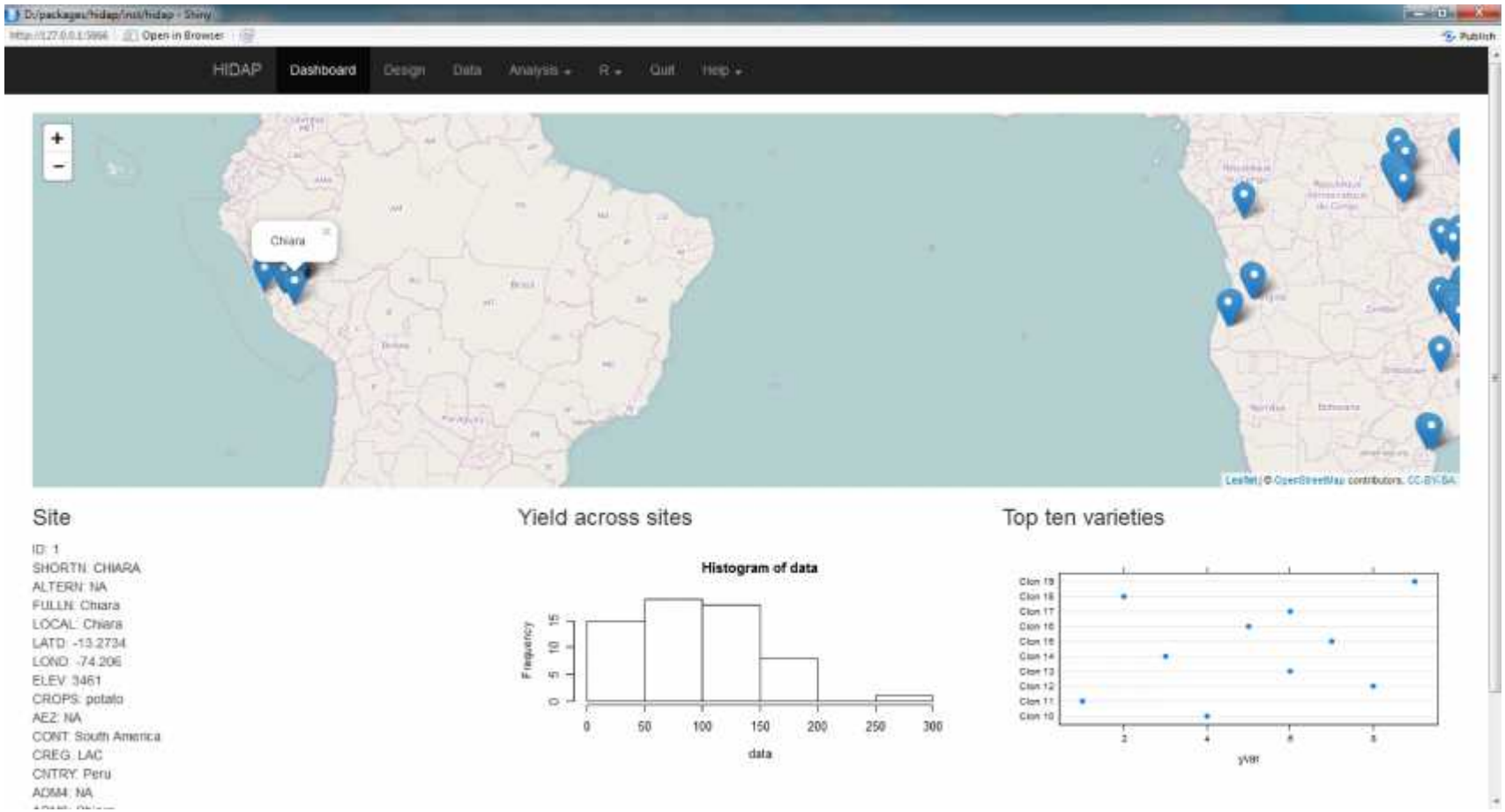
Examples - 1

- High level information integration - dashboard



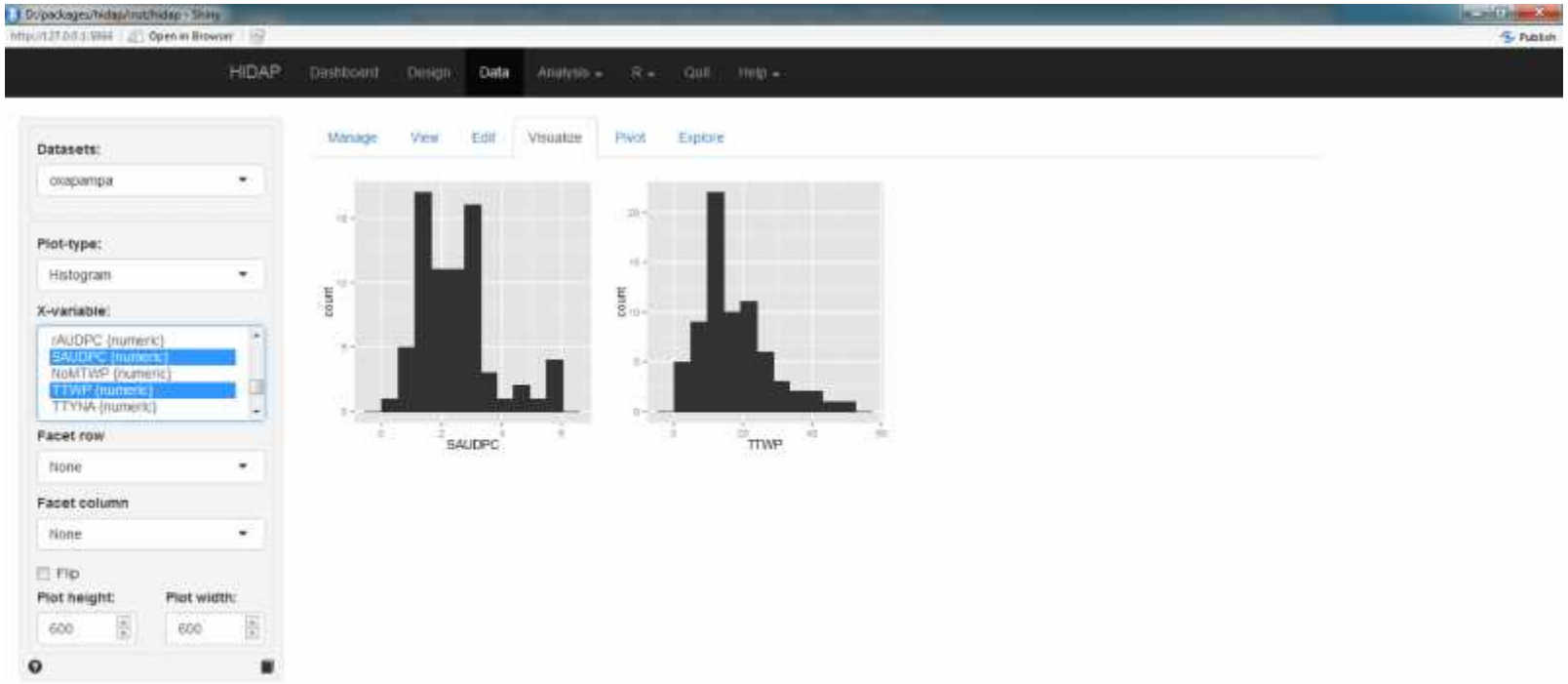
Examples - 2

- Linked graphs



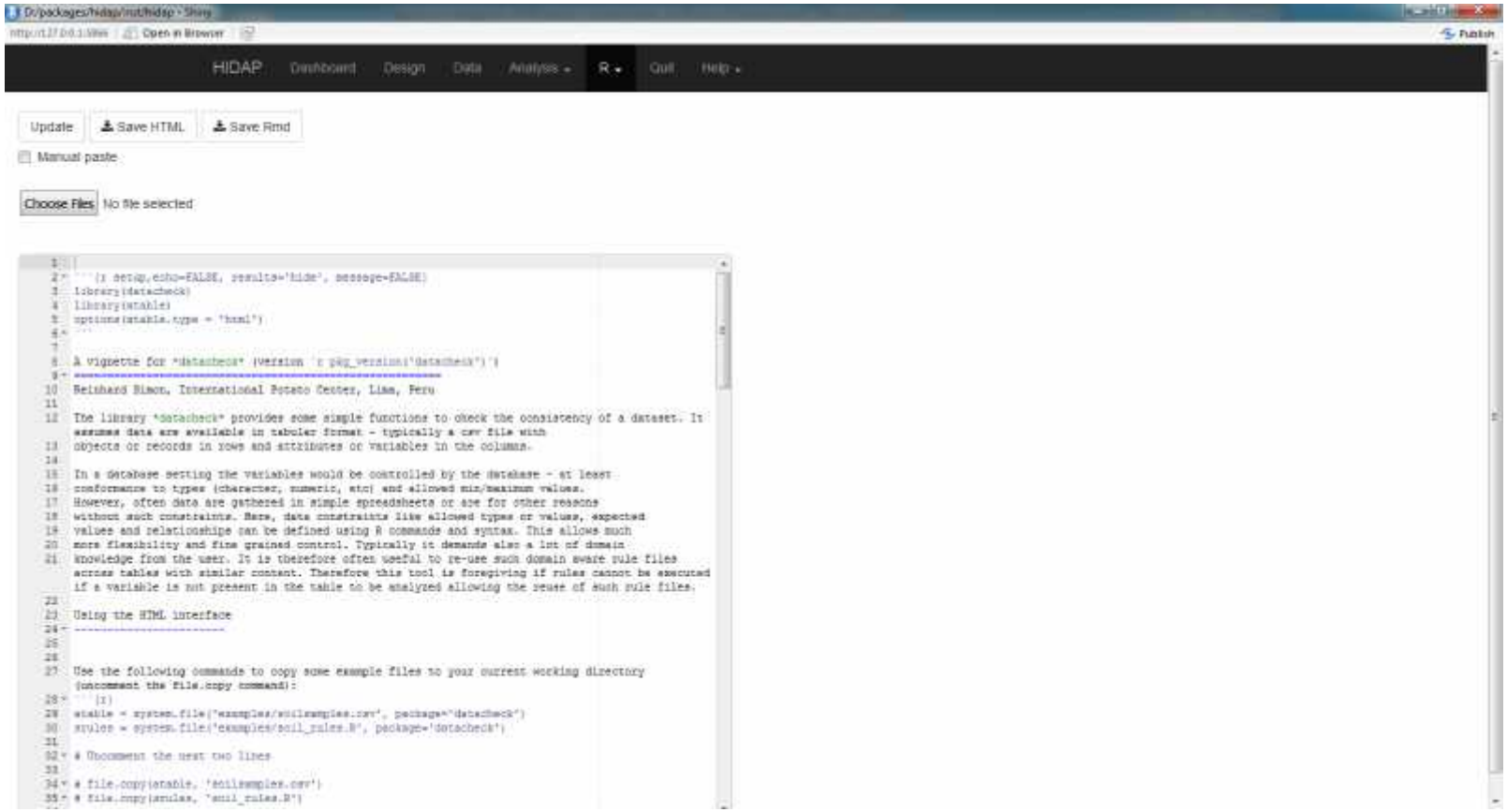
Examples - 3

- Graphical exploration of data



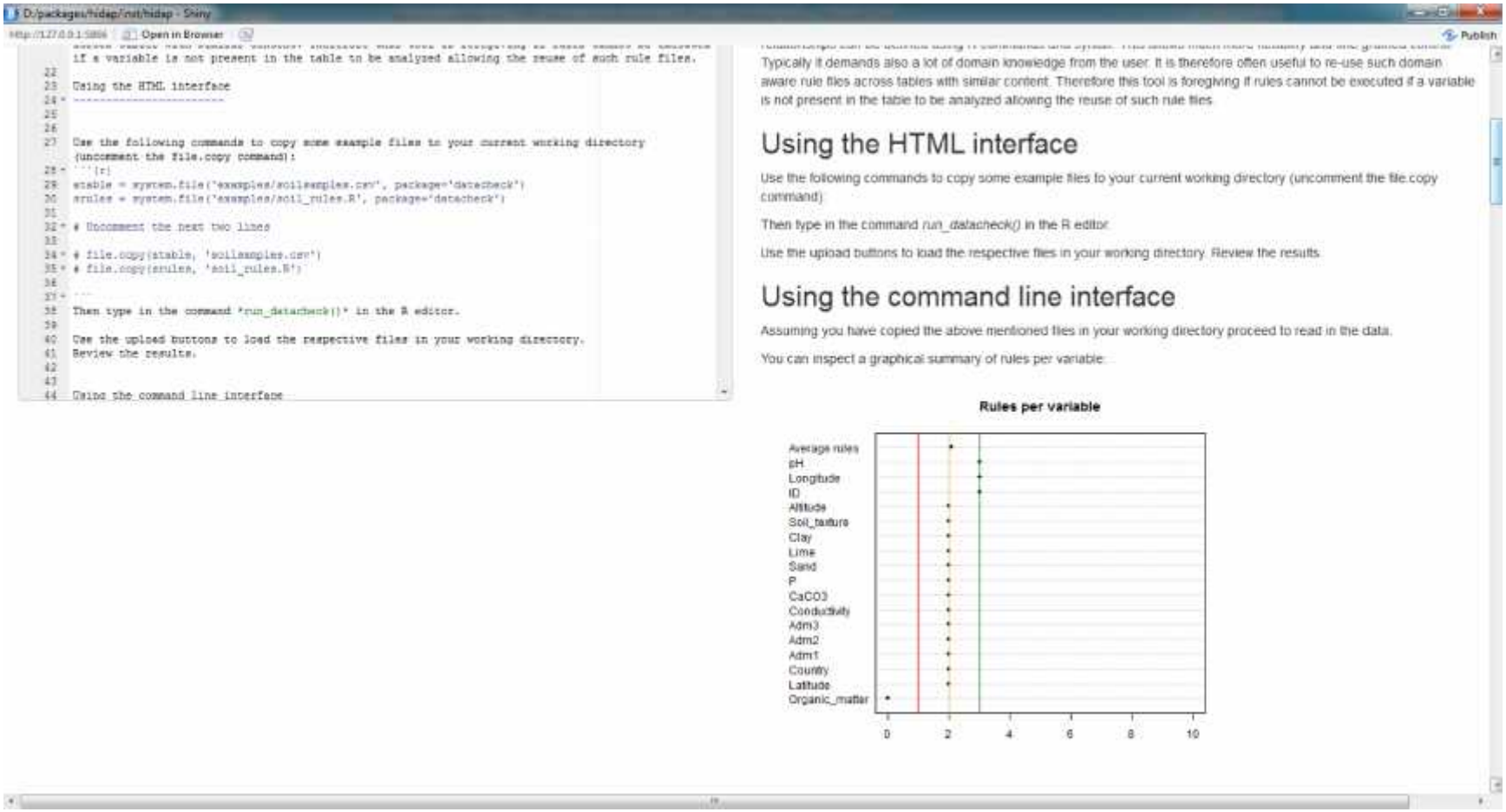
Examples - 4

- Automated reports



Examples - 4

- Automated reports



Examples - 5

- Zooming in: from QTLs to genes (Zbrowse)

Examples - 5

- Zooming in: from QTLs to genes (Zbrowse)