Variant View

Visualizing Sequence Variants in their Gene Context

Tamara Munzner
Department of Computer Science
University of British Columbia

BioIT World, Data Visualization and Exploration Tools Track

30 April 2014, Boston MA

http://www.cs.ubc.ca/~tmm/talks.html#bioit14
Variant View: Visualization Design Study

• a specific real-world problem
  – real users and real data,
  – collaboration is (often) fundamental

• design a visualization system
  – implications: requirements, multiple ideas

• validate the design
  – at appropriate levels

• reflect about lessons learned
  – transferable research: improve design guidelines for vis in general
    • confirm, refine, reject, propose

more at:

more at:
Design Study Methodology

Reflections from the Trenches and from the Stacks

joint work with:
Michael Sedlmair, Miriah Meyer

http://www.cs.ubc.ca/labs/imager/tr/2012/dsm/

Design Study Methodology: Reflections from the Trenches and from the Stacks.
How To Do Design Studies

• definitions

• 9-stage framework

• 32 pitfalls and how to avoid them
When To Do Design Studies

TASK CLARITY

not enough data

DESIGN STUDY METHODOLOGY SUITABLE

ALGORITHM AUTOMATION POSSIBLE

INFORMATION LOCATION

head

crisp

computer

fuzzy
Design Studies: Lessons learned after 21 of them

- MizBee genomics
- Pathline genomics
- Cerebral genomics
- MulteeSum genomics
- Vismon fisheries management
- QuestVis sustainability
- WiKeVis in-car networks
- MostVis in-car networks
- Car-X-Ray in-car networks
- ProgSpy2010 in-car networks
- RelEx in-car networks
- Cardiogram in-car networks
- AutobahnVis in-car networks
- VisTra in-car networks
- Constellation linguistics
- LibVis cultural heritage
- Caidants multicast
- SessionViewer web log analysis
- LiveRAC server hosting
- PowerSetViewer data mining
- LastHistory music listening

• commonality of representations cross-cuts domains!
Abstractions and Idioms

• abstractions
  – translate from specifics of domain to vocabulary of vis
    • task abstraction: why they’re looking at it
    • data abstraction: what to draw
  – transform data into form useful for task at hand
    • don't just draw what you're given; decide what is the right thing!

• idioms
  – visual encoding idiom: how to draw
  – interaction idiom: how to manipulate

• focus today: two mappings
  – from domain to abstraction
  – from abstraction to idiom

A Nested Model of Visualization Design and Validation.
Variant View

Visualizing Sequence Variants in their Gene Context

joint work with:
Joel Ferstay, Cydney Nielsen

http://www.cs.ubc.ca/labs/imager/tr/2012/VariantView/

Variant View: Visualizing Sequence Variants in their Gene Context.
Sequence Variant Definition

- Sequence variants
  - Difference between reference and given genome

Reference Genome DNA: ATA TGA TCA ACA CTT

Sample 1 Genome DNA: ATA TG\_G TCA AT\_A CTT

Sample 2 Genome DNA: ATA TGA TGA ACA C\_C T

Harmful? Harmless?
Cancer Research

• collaboration with analysts at BC Genome Sciences Center
  – studying genetic basis of leukemia

• driving task
  – discover new candidate genes with harmful variants

• two big questions
  – what to show
    • data abstraction
    • challenge: enormous range of scales in the data
  – how to show it
    • visual encoding idiom
Abstractions
Data: Filtering to relevant biological levels and scales

Genome

3 billion bp

Gene

10,000 bp

Exons

100 bp

Transcript

Translation

Protein

Protein Regions

3 billion bp

10,000 bp

100 bp

50 aa
Filter out whole genome; keep genes
Filter out non-exon regions
Data abstraction: highly filtered scope
Dominant paradigm: genome browsers

- **strengths**: flexible and powerful
  - horizontal tracks: user data
  - shared coordinate system: genome coordinates (bp)

- **problems**
  - tiny features of interest spread out across large extent
    - must zoom far in to inspect known feature, then zoom out and pan to locate next
    - high cognitive load for interaction
    - must already know where to look

representative example: Ensembl

Features of interest small even in variant-specific view

1st Screen

Exon regions small

Color coding difficult to see

Protein regions overlap on same track

Ensembl Variant Image
Idioms
Variant View
Variant View

Information-dense single gene view
Variant View

Information-dense single gene view

No need for pan and zoom
Variant View

Sorting metrics guide gene navigation
Variant View

Sorting metrics guide gene navigation

Control what shows up here
Peripheral supporting data
Design information-dense visual encoding

• show all attributes necessary for variant analysis
  – match salience with importance for analysis task

• variant not just a thin line!

• emphasize with high salience
  – collocated variants fan out at top
  – grey variant vertical stroke intersects horizontal colored protein regions
Design information-dense visual encoding
Design information-dense visual encoding

Reference AA

Variant AA

Variant
### Design information-dense visual encoding

#### Reference AA vs. Variant AA

**AA Chemical Class Colours:**
- Charged
- Special
- Uncharged
- Hydrophobic

#### Variant Data

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Chr. Coord.</th>
<th>Ref Base</th>
<th>Var Base</th>
<th>dbSNP129</th>
<th>dbSNP135</th>
<th>dbSNP137</th>
<th>COSMIC</th>
<th>A.A. Chng.</th>
<th>Gene</th>
<th>RefSeq ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>pid-anon</td>
<td>11288816</td>
<td>G</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288819</td>
<td>G</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288821</td>
<td>G</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11292688</td>
<td>T</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11292688</td>
<td>C</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Mutation Type**
- Reference A.A.s
- Variant A.A.s

**Transcript**
- active

**Protein**
- A.A. Chain
- Domains
- Regions
- Active Sites
- Bindings

**Mod. Residue**
- Variant Data

**Variant Data**
- Patient ID
- Chr. Coord.
- Ref Base
- Var Base
- dbSNP129
- dbSNP135
- dbSNP137
- COSMIC
- A.A. Chng.
- Gene
- RefSeq ID
Design information-dense visual encoding

<table>
<thead>
<tr>
<th>Reference AA</th>
<th>Variant AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA Chemical Class Colours:</td>
<td></td>
</tr>
<tr>
<td>Charged</td>
<td>Uncharged</td>
</tr>
<tr>
<td>Special</td>
<td></td>
</tr>
</tbody>
</table>

Variant Type
- Stop
- Indel
- Deletion
- Insertion
- Splice
- Frameshift
- Nonsynonym

**Gene Search:**

- **Submit**
- **Sort By Gene:**
  - Alpha
  - Cluster Score
  - Variant Count

**Variants**

<table>
<thead>
<tr>
<th>Mutation Type</th>
<th>Reference A.A.s</th>
<th>Variant A.A.s</th>
<th>Transcript</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>trans-anon</td>
<td></td>
</tr>
</tbody>
</table>

**Domains**

<table>
<thead>
<tr>
<th>Active Sites</th>
<th>Bindings</th>
</tr>
</thead>
</table>

**Mod. Residue**

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Chr. Coord.</th>
<th>Ref Base</th>
<th>Var Base</th>
<th>dbSNP129</th>
<th>dbSNP135</th>
<th>dbSNP137</th>
<th>COSMIC</th>
<th>A.A. Chng.</th>
<th>Gene</th>
<th>RefSeq ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>pid-anon</td>
<td>11288816</td>
<td>G</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288819</td>
<td>G</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>rs121918</td>
<td>13012</td>
<td>G60V</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288819</td>
<td>G</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>rs121918</td>
<td>13014</td>
<td>A72S</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288821</td>
<td>G</td>
<td>C</td>
<td>.</td>
<td>.</td>
<td>rs121918</td>
<td>13016</td>
<td>E76Q</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288821</td>
<td>A</td>
<td>G</td>
<td>.</td>
<td>rs121918</td>
<td>13017</td>
<td>E76G</td>
<td>gene-anon</td>
<td>trans-anon</td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288821</td>
<td>G</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>rs121918</td>
<td>13018</td>
<td>E76D</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
<tr>
<td>pid-anon</td>
<td>11292688</td>
<td>T</td>
<td>A</td>
<td>.</td>
<td>rs121918</td>
<td>13020</td>
<td>S502T</td>
<td>gene-anon</td>
<td>trans-anon</td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11292688</td>
<td>T</td>
<td>G</td>
<td>.</td>
<td>rs121918</td>
<td>13020</td>
<td>S502A</td>
<td>gene-anon</td>
<td>trans-anon</td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11292688</td>
<td>C</td>
<td>T</td>
<td>13023</td>
<td></td>
<td></td>
<td></td>
<td>S502L</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
</tbody>
</table>

**Alternative Transcripts:**

<table>
<thead>
<tr>
<th>gene-anon (trans-anon)</th>
<th></th>
</tr>
</thead>
</table>

**Variant Data**

<table>
<thead>
<tr>
<th>pid-anon</th>
<th>11288816</th>
<th>G</th>
<th>T</th>
<th>.</th>
<th>.</th>
<th>rs121918</th>
<th>13012</th>
<th>G60V</th>
<th>gene-anon</th>
<th>trans-anon</th>
</tr>
</thead>
<tbody>
<tr>
<td>pid-anon</td>
<td>11288819</td>
<td>G</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>rs121918</td>
<td>13014</td>
<td>A72S</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288821</td>
<td>G</td>
<td>C</td>
<td>.</td>
<td>.</td>
<td>rs121918</td>
<td>13016</td>
<td>E76Q</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288821</td>
<td>A</td>
<td>G</td>
<td>.</td>
<td>rs121918</td>
<td>13017</td>
<td>E76G</td>
<td>gene-anon</td>
<td>trans-anon</td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11288821</td>
<td>G</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>rs121918</td>
<td>13018</td>
<td>E76D</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
<tr>
<td>pid-anon</td>
<td>11292688</td>
<td>T</td>
<td>A</td>
<td>.</td>
<td>rs121918</td>
<td>13020</td>
<td>S502T</td>
<td>gene-anon</td>
<td>trans-anon</td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11292688</td>
<td>T</td>
<td>G</td>
<td>.</td>
<td>rs121918</td>
<td>13020</td>
<td>S502A</td>
<td>gene-anon</td>
<td>trans-anon</td>
<td></td>
</tr>
<tr>
<td>pid-anon</td>
<td>11292688</td>
<td>C</td>
<td>T</td>
<td>13023</td>
<td></td>
<td></td>
<td></td>
<td>S502L</td>
<td>gene-anon</td>
<td>trans-anon</td>
</tr>
</tbody>
</table>
Design information-dense visual encoding

Known Database
- Known Harmless
- Known Cancer

Reference AA

Variant AA

AA Chemical Class Colours:
- Charged
- Special
- Uncharged
- Hydrophobic

Variant Type
- Stop
- Indel
- Deletion
- Insertion
- Splice
- Frameshift
- Nonsynonym

Variant
Design information-dense visual encoding

Known Database
- Known Harmless
- Known Cancer

Reference AA
Variant AA

AA Chemical Class Colours:
- Charged
- Special
- Uncharged
- Hydrophobic

Variant Type
- Stop
- Indel
- Deletion
- Insertion
- Splice
- Frameshift
- Nonsynonym

Transcript/Region Colours:
- Transcript
- AA Chain
- All Other Regions
- Non-Intersected Regions
Previous work targeted at variant analysis: MuSiC

[Dees et al., Genome Research 2012]
Side-by-side comparison: MuSiC vs Variant View

Protein regions can overlap

Regions get separate lanes
Side-by-side comparison: MuSiC vs Variant View

Many collocated variants

Large bloom of repeated elements: more salient
Results
Verify known leukemia gene: Highly scored by sorting metric
Visual inspection reveals collocation of variants
Several functional protein regions affected
Highly scored by metric: not previously known, good candidate
Protein chemical class change evident
In contrast, low scoring gene
No collocation of variants
Mostly unaffected protein regions
Additional tasks

• task 2: compare patients
  – clinical setting application
  – compare patient data to known harmful variants
  – challenge
    • similarity is loosely understood rather than fully characterized
    • visual inspection for what constitutes a match
Adapted Variant View with minimal changes
Navigate through patient data with list
Patient data emphasized with arrows
Patient has same harmful L to P mutation
Additional tasks

• task 3: debug pipeline
  – data cleansing before analysis
  – analysts originally thought pipeline fully debugged
    • no perceived need for vis support
The tool exposed artifacts in the data that slid past at least two rounds of quality metric filtering … this type of problem would not have been caught by our previous, automated methods.

- Analyst 3
Reflections: vis design guidelines

• transferrable to other domains
  – specialize first, generalize later
    • good for domains where with complex, multi-scale data
    • difficult to judge a priori which design elements will generalize
  – high-level considerations
    • identifying scales of interest
    • what to visually encode directly vs what to support through interaction
    • when (and how) to eliminate navigation
Conclusions

• visual variant impact assessment
  – designed, implemented, and deployed tool for
• originally designed for Discover Genes task
  – adapted to two others with minimal changes
• features
  – navigation-free main overview at gene level
  – reveal genes of interest through sorting by new derived metrics
• major considerations
  – what to show
    • filtering data scope
  – how to show it
    • carefully selected visual encodings
Further Information

• paper page

• open source software download

• further info
  – http://www.cs.ubc.ca/~tmm/talks.html#bioit14 (this talk, and many others)
  – http://www.cs.ubc.ca/group/infovis (papers, software, videos)

• acknowledgements
  – funding: Vancouver Institute for Visual Analytics (VIVA), Aeroinfo/Boeing, Mitacs
  – collaborators at the GSC
    • Dr. Aly Karsan
    • Rod Docking
    • Dr. Linda Chang
    • Dr. Gerben Duns
    • Simon Chang