

# Variant View

*Visualizing Sequence Variants in their Gene Context*

**Tamara Munzner**

Department of Computer Science  
University of British Columbia

*Oxford University*

*2 July 2014, Oxford UK*

<http://www.cs.ubc.ca/~tmm/talks.html#oxford14>

# Variant View: Visualization *Design Study*

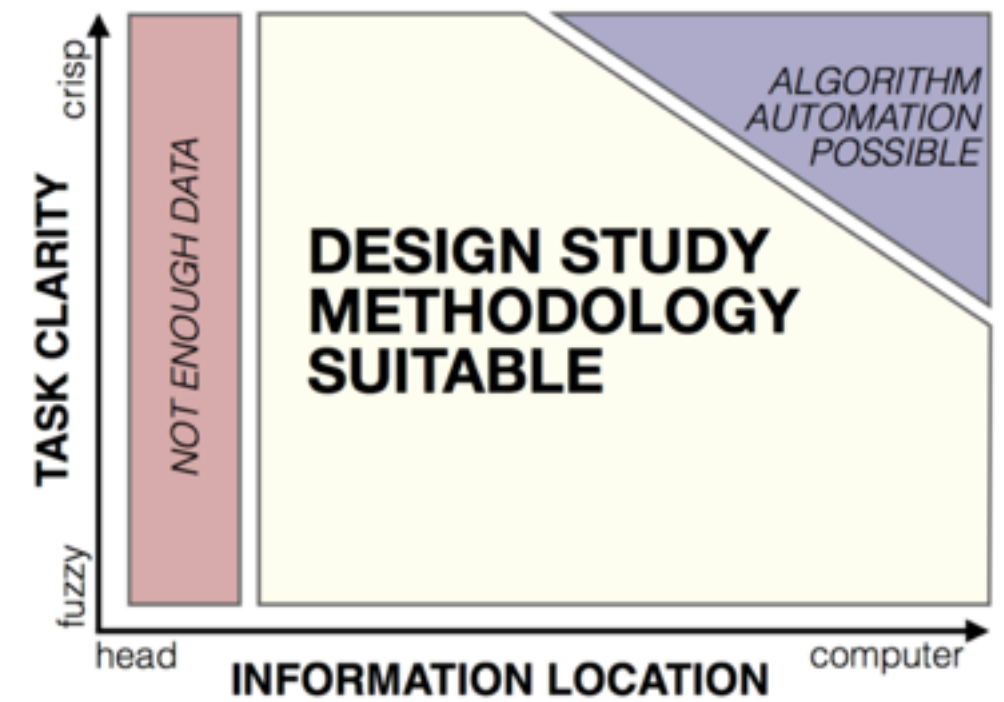
- a specific **real-world** problem
  - real users and real data,
  - collaboration is (often) fundamental
- **design** a visualization system
  - implications: requirements, multiple ideas
- **validate** the design
  - at appropriate levels
- **reflect** about lessons learned
  - transferable research: improve design guidelines for vis in general
    - confirm, refine, reject, propose

more at:

A Nested Model of Visualization Design and Validation.  
Munzner. *IEEE TVCG* 15(6):921-928, 2009 (Proc. InfoVis 2009).

more at:

The Nested Blocks and Guidelines Model.  
Meyer, Sedlmair, Quinan, Munzner. *Information Visualization Journal*, 2014,  
to appear.



# Design Study Methodology

*Reflections from the Trenches and from the Stacks*

**joint work with:**

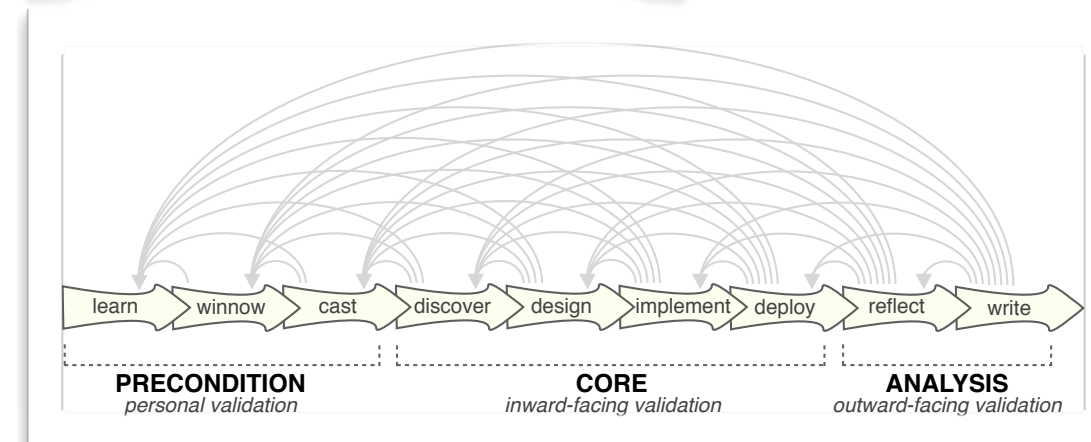
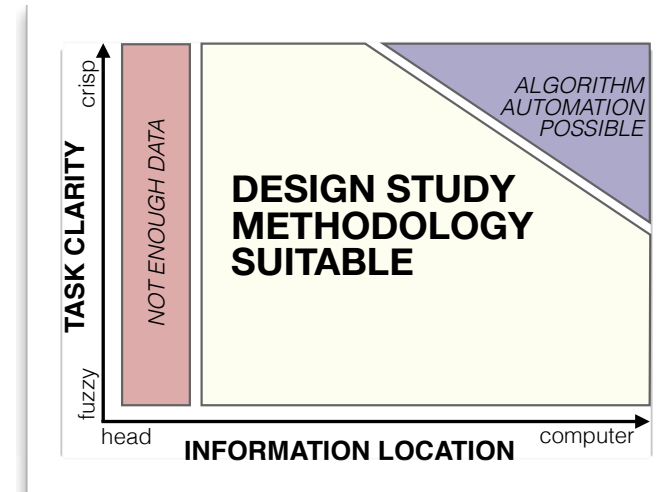
Michael Sedlmair, Miriah Meyer

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

Design Study Methodology: Reflections from the Trenches and from the Stacks.  
Sedlmair, Meyer, Munzner. *IEEE TVCG* 18(12): 2431-2440, 2012 (Proc. InfoVis 2012).

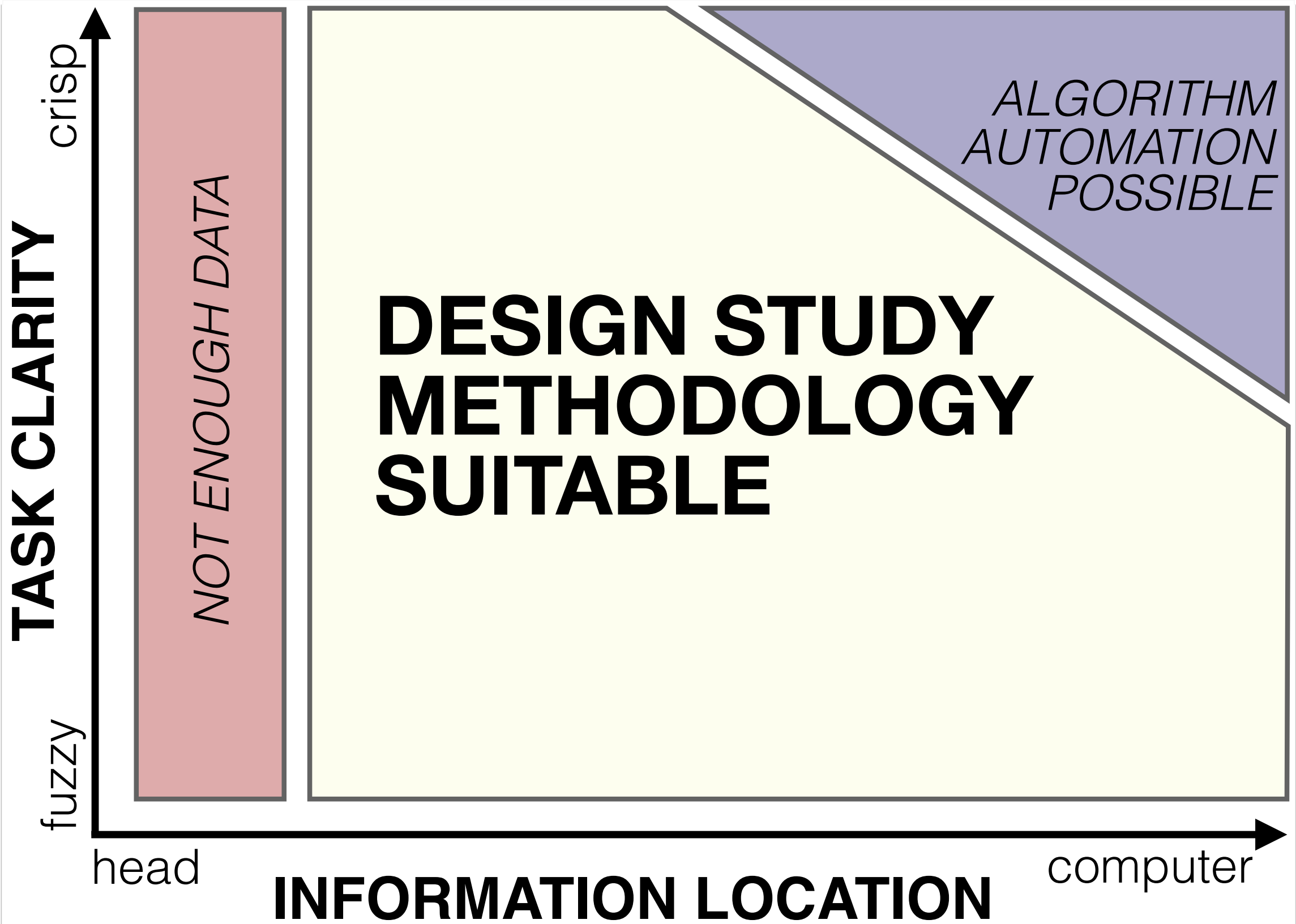
# How To Do Design Studies

- definitions
- 9-stage framework
- 32 pitfalls and how to avoid them



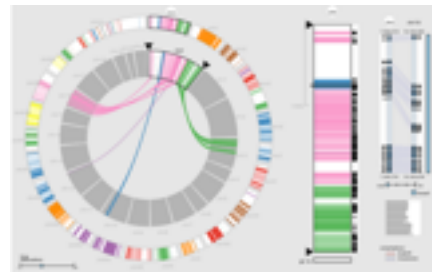
|      |   |         |
|------|---|---------|
| PF-1 | premature advance: jumping forward over stages            | general |
| PF-2 | premature start: insufficient knowledge of vis literature | learn   |
| PF-3 | premature commitment: collaboration with wrong people     | winnow  |
| PF-4 | no real data available (yet)                              | winnow  |
| PF-5 | insufficient time available from potential collaborators  | winnow  |
| PF-6 | no need for visualization: problem can be automated       | winnow  |
| PF-7 | researcher expertise does not match domain problem        | winnow  |
| PF-8 | no need for research: engineering vs. research project    | winnow  |
| PF-9 | no need for change: existing tools are good enough        | winnow  |

# When To Do Design Studies





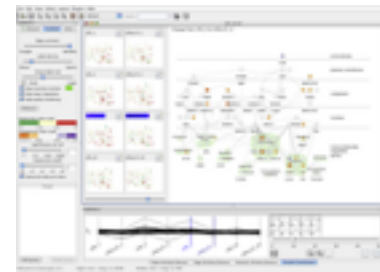
# Design Studies: Lessons learned after 21 of them



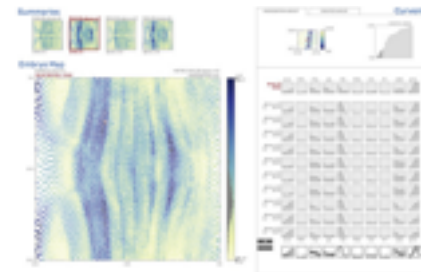
*MizBee*  
genomics



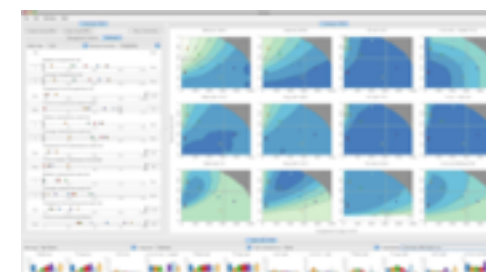
*Pathline*  
genomics



*Cerebral*  
genomics



*MulteeSum*  
genomics



*Vismon*  
fisheries management



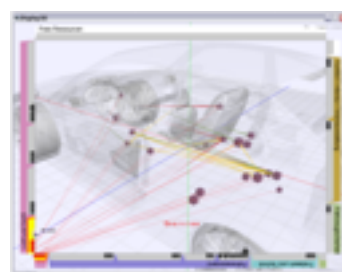
*QuestVis*  
sustainability



*WiKeVis*  
in-car networks



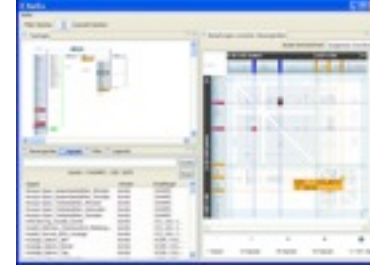
*MostVis*  
in-car networks



*Car-X-Ray*  
in-car networks



*ProgSpy2010*  
in-car networks



*ReEx*  
in-car networks



*Cardiogram*  
in-car networks



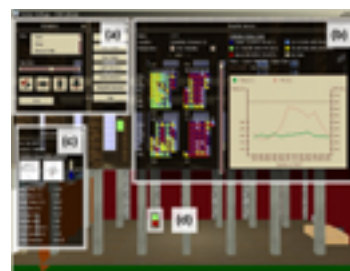
*AutobahnVis*  
in-car networks



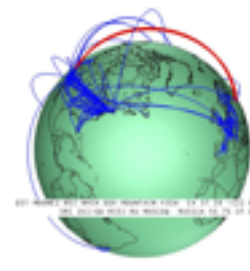
*VisTra*  
in-car networks



*Constellation*  
linguistics



*LibVis*  
cultural heritage



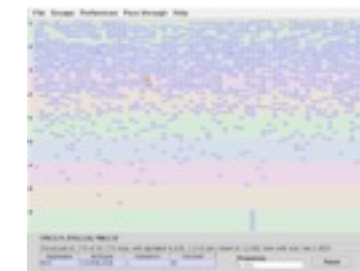
*Caidants*  
multicast



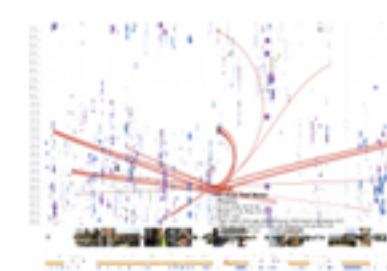
*SessionViewer*  
web log analysis



*LiveRAC*  
server hosting



*PowerSetViewer*  
data mining

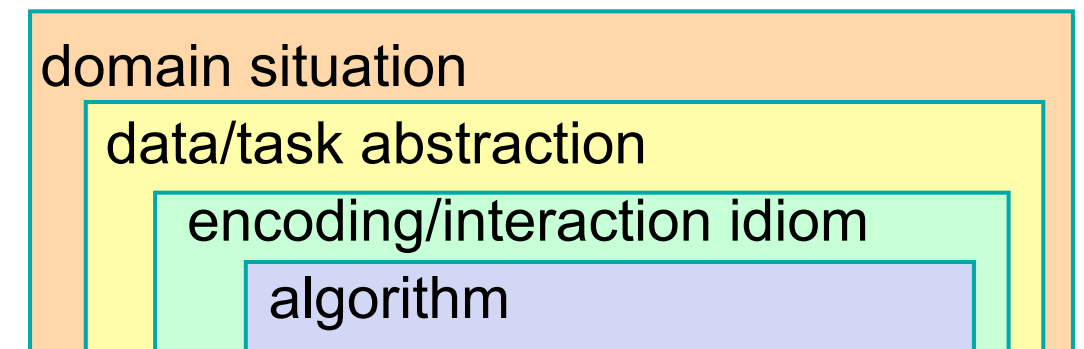


*LastHistory*  
music listening

- 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



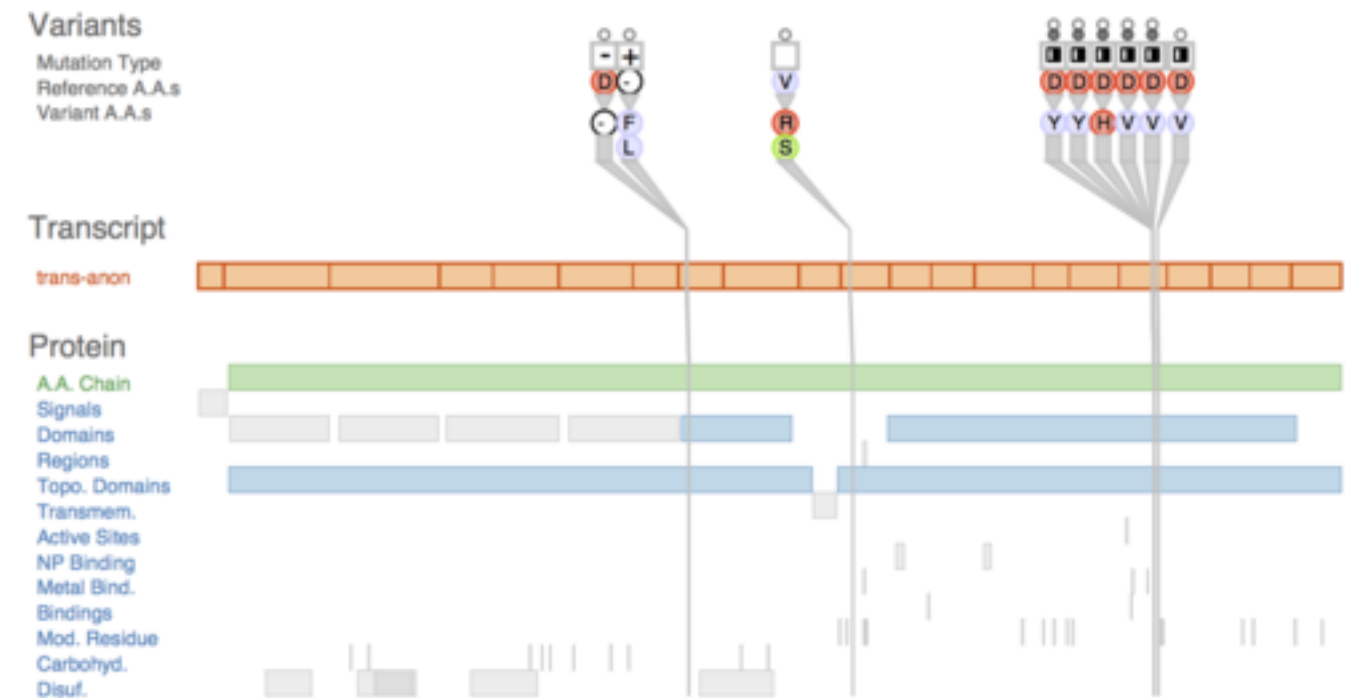
# Variant View

## Visualizing Sequence Variants in their Gene Context

**joint work with:**

Joel Ferstay, Cydney Nielsen

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

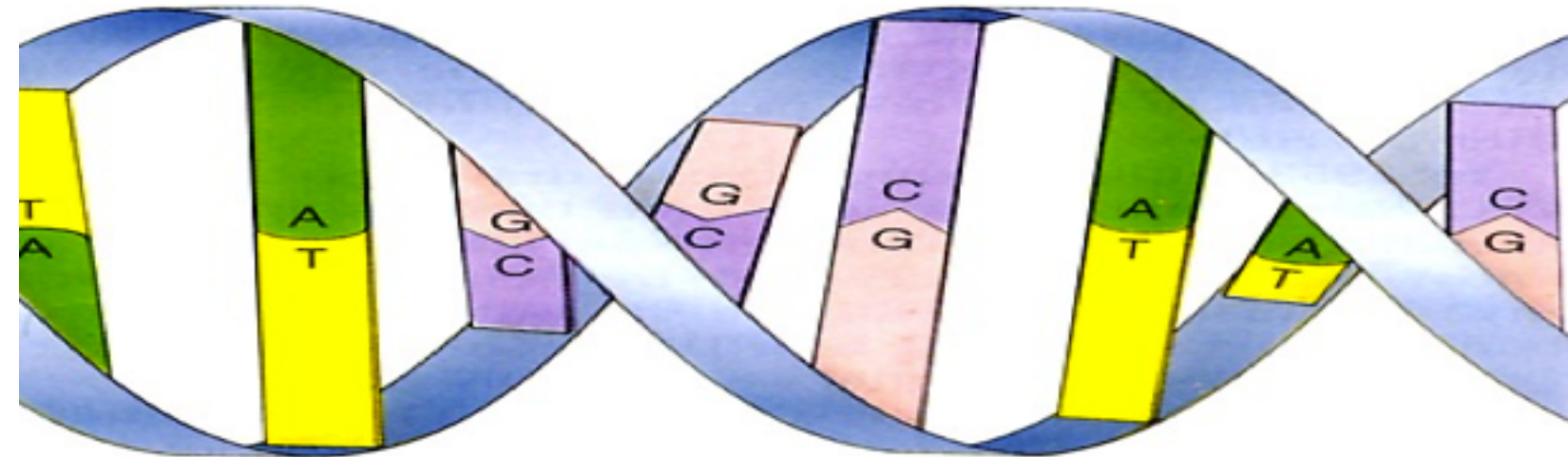


Variant View: Visualizing Sequence Variants in their Gene Context.  
Ferstay, Nielsen, Munzner. *IEEE TVCG* 19(12): 2546-2555, 2013 (Proc. InfoVis 2013).



# 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 **ATA** CTT

Sample 2 Genome DNA:      ATA TGA **TGA** ACA **CCT**

Harmful?

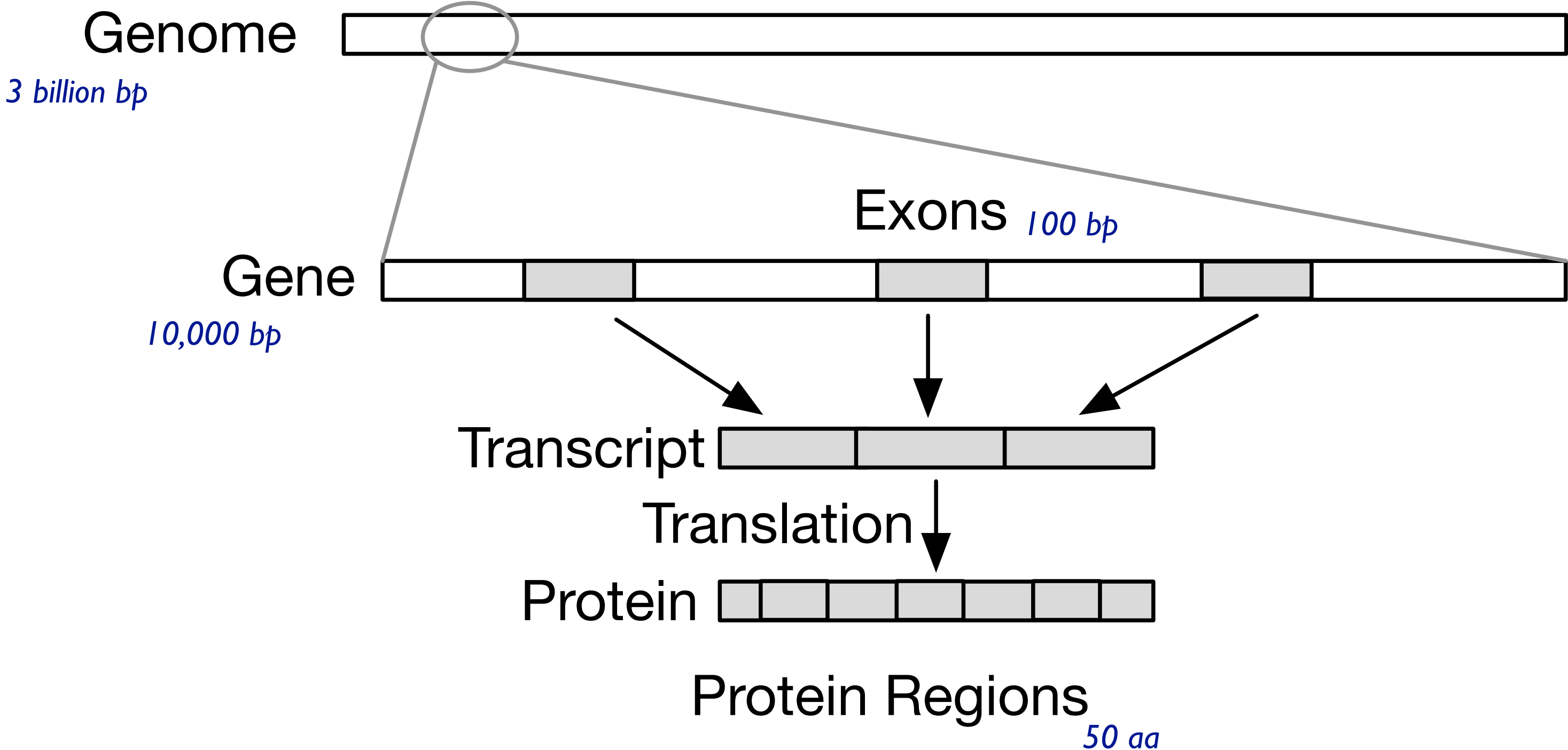
Harmless?

# Cancer Research

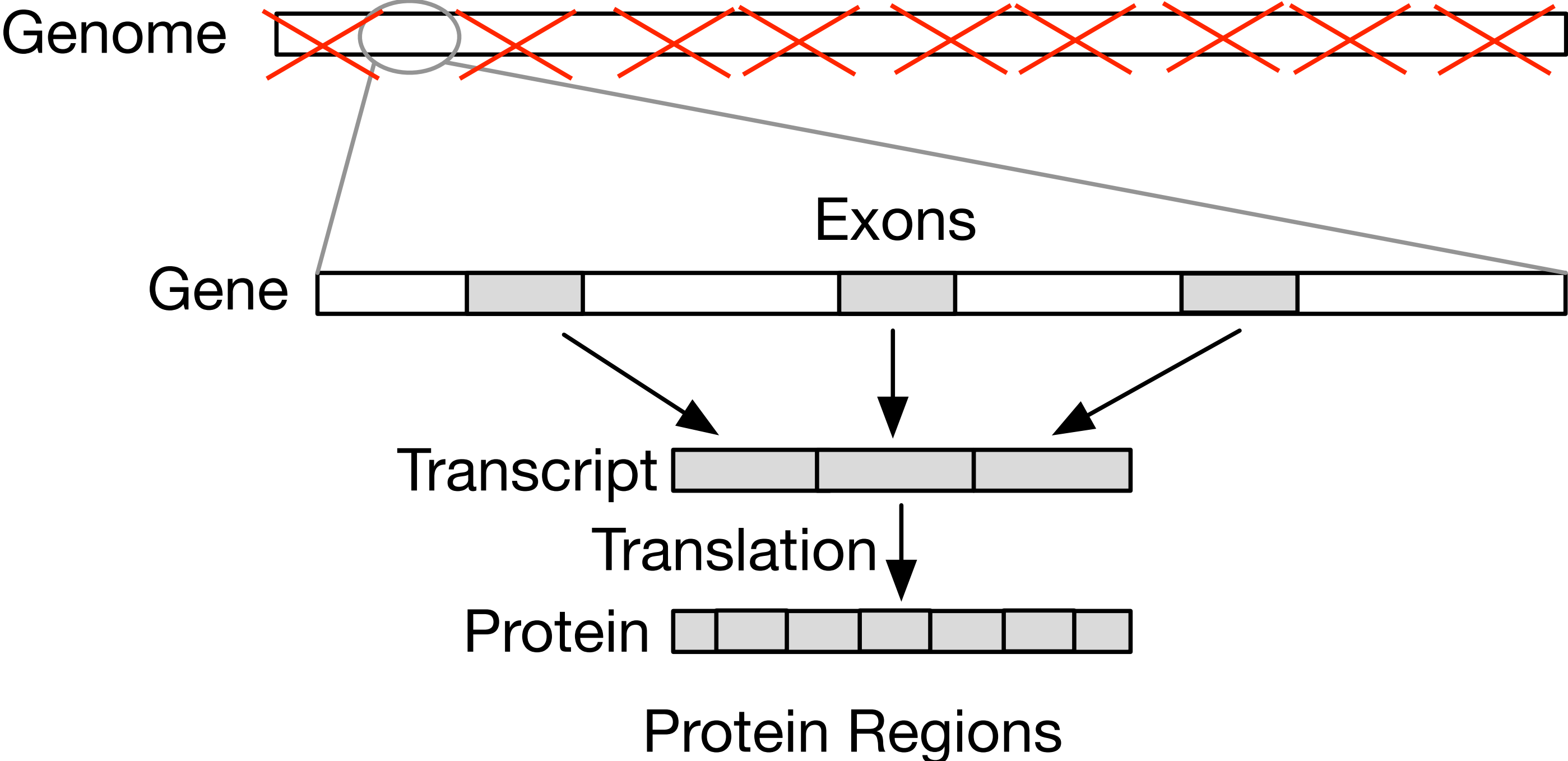
- collaboration with analysts at BC Genome Sciences Center
  - studying genetic basis of leukemia
- driving task
  - discover new candidate genes with harmful variants
- two big questions
  - what to show
    - data abstraction
    - challenge: enormous range of scales in the data
  - how to show it
    - visual encoding idiom

# Abstractions

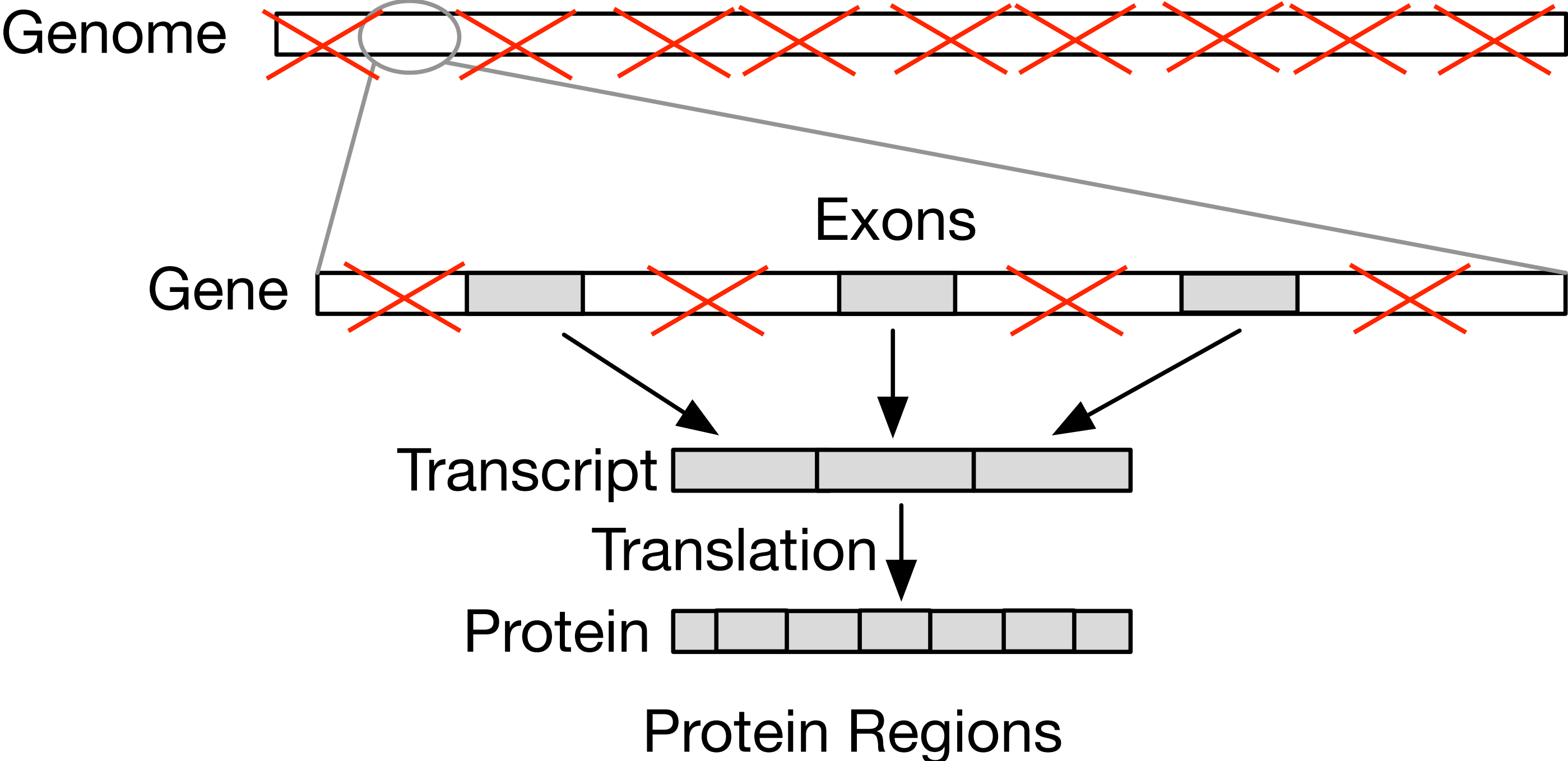
# Data: Filtering to relevant biological levels and scales



# Filter out whole genome; keep genes

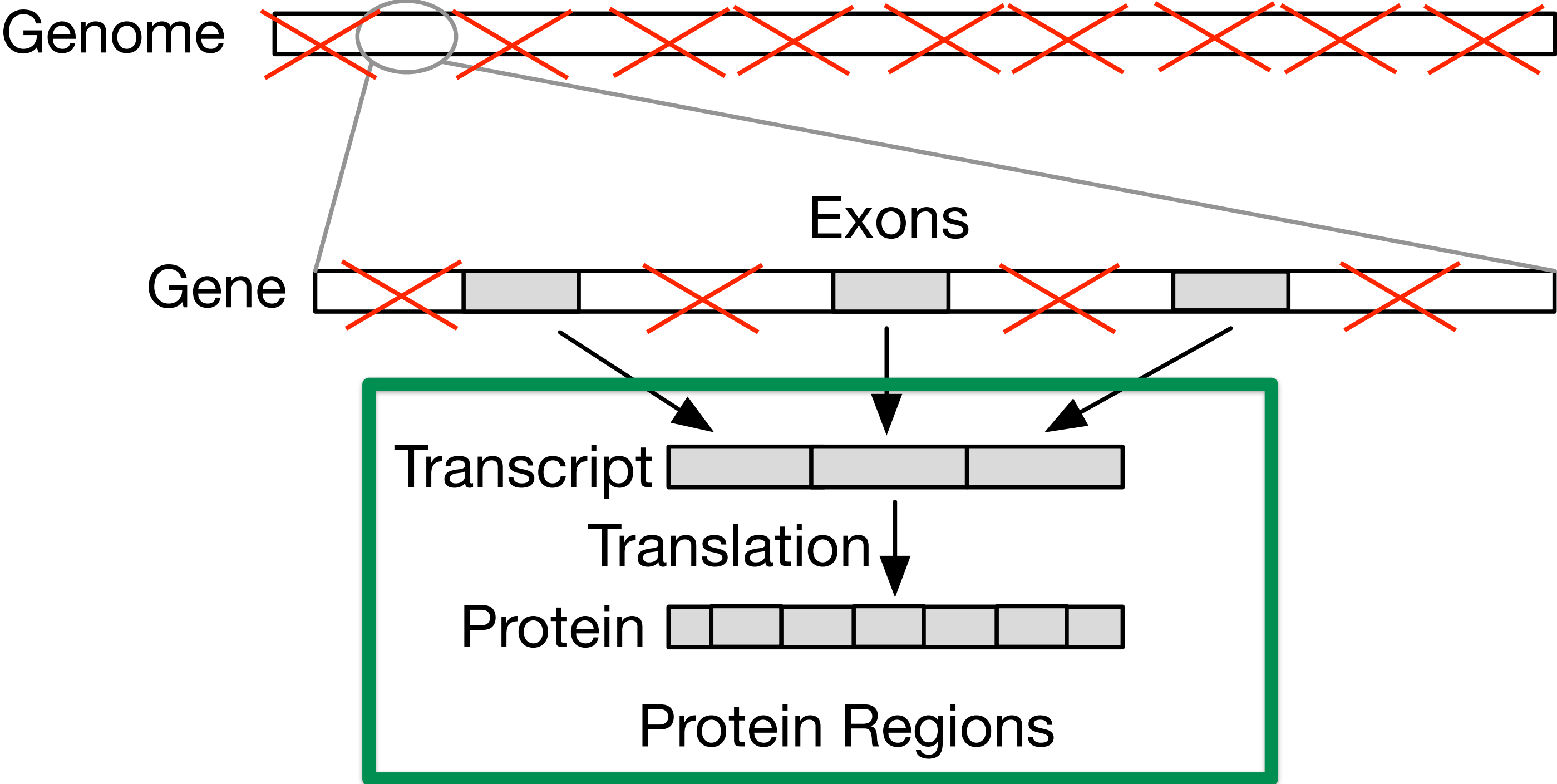


# Filter out non-exon regions



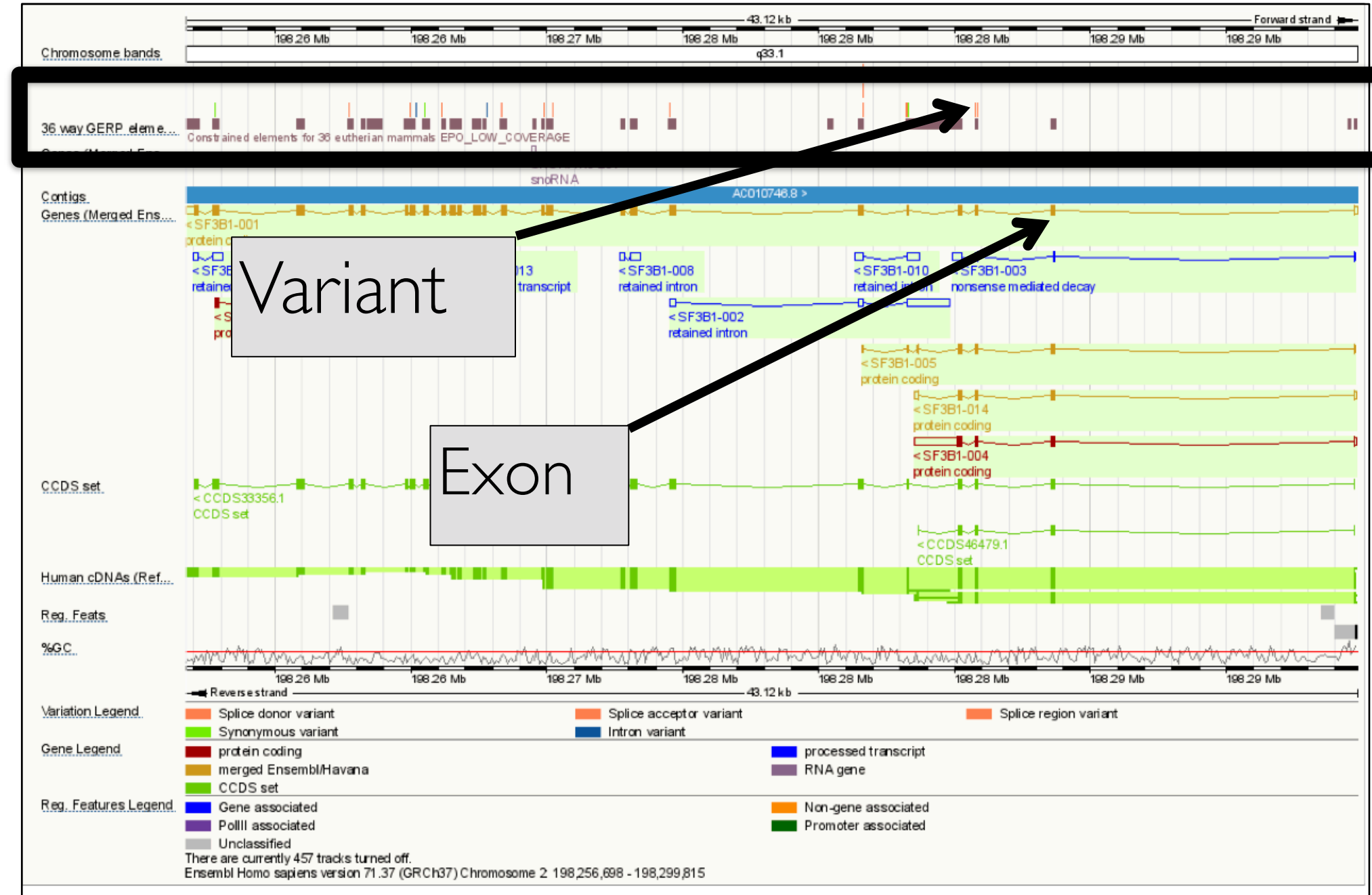


# Data abstraction: highly filtered scope



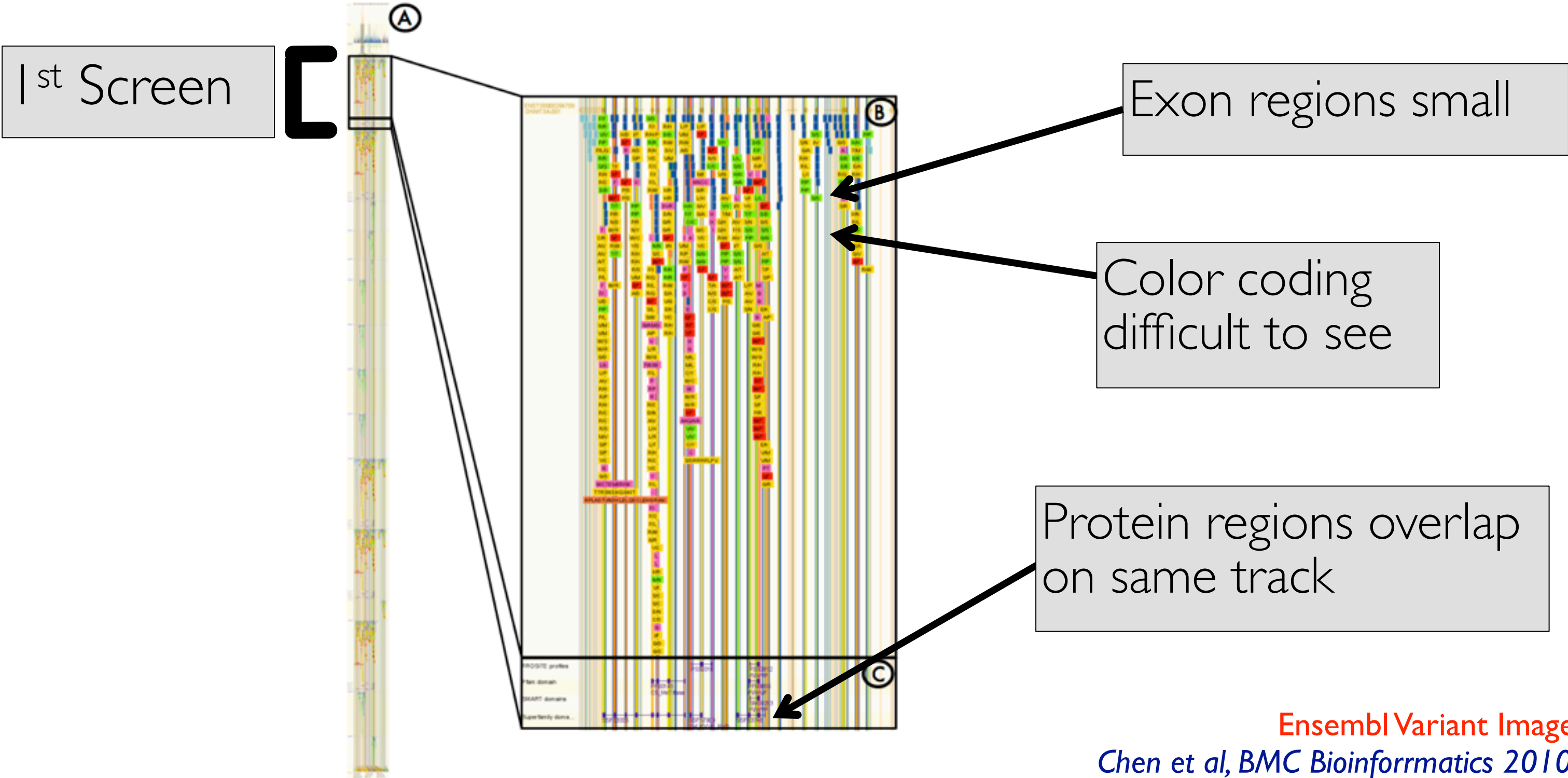
# Dominant paradigm: genome browsers

- 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  
Chen et al, BMC Bioinformatics 2010.

# Features of interest small even in variant-specific view



# Idioms

# Variant View

Gene Search:

Alternative Transcripts:

**Variants**

Mutation Type  
Reference A.A.s  
Variant A.A.s

**Transcript**

trans-anon

**Protein**

A.A. Chain  
Domains  
Regions  
Active Sites  
Bindings  
Mod. Residue

Sort By Gene:

Alpha Cluster Score **Variant Count**

- DNMT3A (NM\_022552)
- IDH2 (NM\_002168)
- FLT3 (NM\_004119)
- ANKRD36 (NM\_001164315)
- ARID1B (NM\_017519)
- STAG2 (NM\_001042749)
- TNRC18 (NM\_001080495)
- WT1 (NM\_000378)
- ABCA13 (NM\_152701)
- CEBPA (NM\_004364)
- TET2 (NM\_001127208)
- DNAH10 (NM\_207437)
- GPSM1 (NM\_015597)
- ASXL1 (NM\_015338)
- DNAH1 (NM\_015512)
- DNAH6 (NM\_001370)
- FAT1 (NM\_005245)
- MDN1 (NM\_014611)
- PTPN11 (NM\_002834)
- SYNE1 (NM\_033071)
- ALMS1 (NM\_015120)
- C10orf68 (NM\_024688)
- CCDC88C (NM\_001080414)
- DNAH11 (NM\_003777)
- DNAH3 (NM\_017539)
- DNAH9 (NM\_001372)

**Variant Data**

| Patient ID | Chr. Coord. | Ref Base | Var Base | dbSNP129 | dbSNP135 | dbSNP137 | COSMIC  | A.A. Chng. | Gene      | Ref. Gene  |
|------------|-------------|----------|----------|----------|----------|----------|---------|------------|-----------|------------|
| pid-anon   | 11288816    | G        | T        | .        | .        | .        | *13028, | G60V       | gene-anon | trans-anon |
| pid-anon   | 11288816    | G        | T        | .        | .        | .        | *13012, | D61Y       | gene-anon | trans-anon |
| pid-anon   | 11288819    | G        | T        | .        | rs121918 | .        | 13014   | A72S       | gene-anon | trans-anon |
| pid-anon   | 11288819    | C        | T        | .        | .        | .        | *13035, | A72V       | gene-anon | trans-anon |
| pid-anon   | 11288821    | G        | C        | .        | .        | .        | *13016, | E76Q       | gene-anon | trans-anon |
| pid-anon   | 11288821    | A        | G        | .        | rs121918 | .        | *13017, | E76G       | gene-anon | trans-anon |
| pid-anon   | 11288821    | G        | T        | .        | .        | .        | .       | E76D       | gene-anon | trans-anon |
| pid-anon   | 11292688    | T        | A        | .        | rs121918 | .        | *13020, | S502T      | gene-anon | trans-anon |
| pid-anon   | 11292688    | T        | G        | .        | .        | .        | *13020, | S502A      | gene-anon | trans-anon |
| pid-anon   | 11292688    | C        | T        | .        | .        | .        | 13023   | S502L      | gene-anon | trans-anon |

# Variant View

Information-dense single gene view

Gene Search:

Alternative Transcripts:

**Variants**

Mutation Type  
Reference A.A.s  
Variant A.A.s

**Transcript**

trans-anon

**Protein**

A.A. Chain  
Domains  
Regions  
Active Sites  
Bindings  
Mod. Residue

**Variant Data**

| Patient ID | Chr. Coord. | Ref Base | Var Base | dbSNP129 | dbSNP135 | dbSNP137 | COSMIC  | A.A. Chng. | Gene      | Ref. Gene  |
|------------|-------------|----------|----------|----------|----------|----------|---------|------------|-----------|------------|
| pid-anon   | 11288816    | G        | T        | .        | .        | .        | *13028, | G60V       | gene-anon | trans-anon |
| pid-anon   | 11288816    | G        | T        | .        | .        | .        | *13012, | D61Y       | gene-anon | trans-anon |
| pid-anon   | 11288819    | G        | T        | .        | rs121918 | .        | 13014   | A72S       | gene-anon | trans-anon |
| pid-anon   | 11288819    | C        | T        | .        | .        | .        | *13035, | A72V       | gene-anon | trans-anon |
| pid-anon   | 11288821    | G        | C        | .        | .        | .        | *13016, | E76Q       | gene-anon | trans-anon |
| pid-anon   | 11288821    | A        | G        | .        | rs121918 | .        | *13017, | E76G       | gene-anon | trans-anon |
| pid-anon   | 11288821    | G        | T        | .        | .        | .        | .       | E76D       | gene-anon | trans-anon |
| pid-anon   | 11292688    | T        | A        | .        | rs121918 | .        | *13020, | S502T      | gene-anon | trans-anon |
| pid-anon   | 11292688    | T        | G        | .        | .        | .        | *13020, | S502A      | gene-anon | trans-anon |
| pid-anon   | 11292688    | C        | T        | .        | .        | .        | 13023   | S502L      | gene-anon | trans-anon |

Sort By Gene:  
Alpha Cluster Score Variant Count

- DNMT3A (NM\_022552)
- IDH2 (NM\_002168)
- FLT3 (NM\_004119)
- ANKRD36 (NM\_001164315)
- ARID1B (NM\_017519)
- STAG2 (NM\_001042749)
- TNRC18 (NM\_001080495)
- WT1 (NM\_000378)
- ABCA13 (NM\_152701)
- CEBPA (NM\_004364)
- TET2 (NM\_001127208)
- DNAH10 (NM\_207437)
- GPSM1 (NM\_015597)
- ASXL1 (NM\_015338)
- DNAH1 (NM\_015512)
- DNAH6 (NM\_001370)
- FAT1 (NM\_005245)
- MDN1 (NM\_014611)
- PTPN11 (NM\_002834)
- SYNE1 (NM\_033071)
- ALMS1 (NM\_015120)
- C10orf68 (NM\_024688)
- CCDC88C (NM\_001080414)
- DNAH11 (NM\_003777)
- DNAH3 (NM\_017539)
- DNAH9 (NM\_001372)



# Variant View

Gene Search:

Alternative Transcripts:

**Information-dense single gene view**

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

**Transcript**  
trans-anon

**Protein**  
A.A. Chain  
Domains  
Regions  
Active Sites  
Bindings  
Mod. Residue

**Variant Data**

| Patient ID | Chr. | Coord. | Ref Base | Var Base | dbSNP129 | dbSNP135 | dbSNP137 | COSMIC  | A.A. Ch | Gene      | Transcript |
|------------|------|--------|----------|----------|----------|----------|----------|---------|---------|-----------|------------|
| pid-anon   | 11   | 28816  | G        | T        | .        | .        | .        | *13028, | G60V    | gene-anon | trans-anon |
| pid-anon   | 11   | 28816  | G        | T        | .        | .        | .        | *13012, | D61Y    | gene-anon | trans-anon |
| pid-anon   | 11   | 28819  | G        | T        | .        | rs121918 | .        | 13014   | A72S    | gene-anon | trans-anon |
| pid-anon   | 11   | 28819  | C        | T        | .        | .        | .        | *13035, | A72V    | gene-anon | trans-anon |
| pid-anon   | 11   | 28821  | G        | C        | .        | .        | .        | *13016, | E76Q    | gene-anon | trans-anon |
| pid-anon   | 11   | 28821  | A        | G        | .        | rs121918 | .        | *13017, | E76G    | gene-anon | trans-anon |
| pid-anon   | 11   | 28821  | G        | T        | .        | .        | .        | .       | E76D    | gene-anon | trans-anon |
| pid-anon   | 11   | 292688 | T        | A        | .        | rs121918 | .        | *13020, | S502T   | gene-anon | trans-anon |
| pid-anon   | 11   | 292688 | T        | G        | .        | .        | .        | *13020, | S502A   | gene-anon | trans-anon |
| pid-anon   | 11   | 292688 | C        | T        | .        | .        | .        | 13023   | S502L   | gene-anon | trans-anon |

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

- DNMT3A (NM\_022552)
- IDH2 (NM\_002168)
- FLT3 (NM\_004119)
- ANKRD36 (NM\_001164315)
- ARID1B (NM\_017519)
- STAG2 (NM\_001042749)
- TNRC18 (NM\_001080495)
- WT1 (NM\_000378)
- ABCA13 (NM\_152701)
- CEBPA (NM\_004364)
- TET2 (NM\_001127208)
- DNAH10 (NM\_207437)
- GPSM1 (NM\_015597)
- ASXL1 (NM\_015338)
- DNAH1 (NM\_015512)
- DNAH6 (NM\_001370)
- SYNET (NM\_033071)
- ALMS1 (NM\_015120)
- C10orf68 (NM\_024688)
- CCDC88C (NM\_001080414)
- DNAH11 (NM\_003777)
- DNAH3 (NM\_017539)
- DNAH9 (NM\_001372)

**No need for pan and zoom**

# Variant View

Sorting metrics guide gene navigation

Alternative Transcripts:

**Variants**

Mutation Type  
Reference A.A.s  
Variant A.A.s

**Transcript**

trans-anon

**Protein**

A.A. Chain  
Domains  
Regions  
Active Sites  
Bindings  
Mod. Residue

**Variant Data**

| Patient ID | Chr. Coord. | Ref Base | Var Base | dbSNP129 | dbSNP135 | dbSNP137 | COSMIC  | A.A. Chng. | Gene      | Ref. Gene  |
|------------|-------------|----------|----------|----------|----------|----------|---------|------------|-----------|------------|
| pid-anon   | 11288816    | G        | T        | .        | .        | .        | *13028, | G60V       | gene-anon | trans-anon |
| pid-anon   | 11288816    | G        | T        | .        | .        | .        | *13012, | D61Y       | gene-anon | trans-anon |
| pid-anon   | 11288819    | G        | T        | .        | rs121918 | .        | 13014   | A72S       | gene-anon | trans-anon |
| pid-anon   | 11288819    | C        | T        | .        | .        | .        | *13035, | A72V       | gene-anon | trans-anon |
| pid-anon   | 11288821    | G        | C        | .        | .        | .        | *13016, | E76Q       | gene-anon | trans-anon |
| pid-anon   | 11288821    | A        | G        | .        | rs121918 | .        | *13017, | E76G       | gene-anon | trans-anon |
| pid-anon   | 11288821    | G        | T        | .        | .        | .        | .       | E76D       | gene-anon | trans-anon |
| pid-anon   | 11292688    | T        | A        | .        | rs121918 | .        | *13020, | S502T      | gene-anon | trans-anon |
| pid-anon   | 11292688    | T        | G        | .        | .        | .        | *13020, | S502A      | gene-anon | trans-anon |
| pid-anon   | 11292688    | C        | T        | .        | .        | .        | 13023   | S502L      | gene-anon | trans-anon |

Sort By Gene:

- DNMT3A (NM\_022552)
- IDH2 (NM\_002168)
- FLT3 (NM\_004119)
- ANKRD36 (NM\_001164315)
- ARID1B (NM\_017519)
- STAG2 (NM\_001042749)
- TNRC18 (NM\_001080495)
- WT1 (NM\_000378)
- ABCA13 (NM\_152701)
- CEBPA (NM\_004364)
- TET2 (NM\_001127208)
- DNAH10 (NM\_207437)
- GPSM1 (NM\_015597)
- ASXL1 (NM\_015338)
- DNAH1 (NM\_015512)
- DNAH6 (NM\_001370)
- FAT1 (NM\_005245)
- MDN1 (NM\_014611)
- PTPN11 (NM\_002834)
- SYNE1 (NM\_033071)
- ALMS1 (NM\_015120)
- C10orf68 (NM\_024688)
- CCDC88C (NM\_001080414)
- DNAH11 (NM\_003777)
- DNAH3 (NM\_017539)
- DNAH9 (NM\_001372)

# Variant View

Sorting metrics guide gene navigation

Alternative Transcripts:

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

**Transcript**  
trans-anon

**Protein**  
A.A. Chain  
Domains  
Regions  
Active Sites  
Bindings  
Mod. Residue

**Variant Data**

| Gene      | RefSeq ID  | dbSNP    | COSMIC  | A.A. Chng. | Gene      | RefSeq ID  |
|-----------|------------|----------|---------|------------|-----------|------------|
| gene-anon | trans-anon | rs121918 | *13028, | G60V       | gene-anon | trans-anon |
| gene-anon | trans-anon | rs121918 | *13012, | D61Y       | gene-anon | trans-anon |
| gene-anon | trans-anon | rs121918 | 13014   | A72S       | gene-anon | trans-anon |
| gene-anon | trans-anon | rs121918 | *13035, | A72V       | gene-anon | trans-anon |
| gene-anon | trans-anon | rs121918 | *13016, | E76Q       | gene-anon | trans-anon |
| gene-anon | trans-anon | rs121918 | *13017, | E76G       | gene-anon | trans-anon |
| gene-anon | trans-anon | rs121918 | .       | E76D       | gene-anon | trans-anon |
| gene-anon | trans-anon | rs121918 | *13020, | S502T      | gene-anon | trans-anon |
| gene-anon | trans-anon | rs121918 | *13020, | S502A      | gene-anon | trans-anon |
| gene-anon | trans-anon | rs121918 | 13023   | S502L      | gene-anon | trans-anon |

Sort By Gene:  
Alpha | **Cluster Score** | Variant Count

- DNMT3A (NM\_022552)
- IDH2 (NM\_002168)
- FLT3 (NM\_004119)
- ANKRD36 (NM\_001164315)
- ARID1B (NM\_017519)
- STAG2 (NM\_001042749)
- TNRC18 (NM\_001080495)
- WT1 (NM\_000378)
- ABCA13 (NM\_152701)
- CEBPA (NM\_004364)
- TET2 (NM\_001127208)
- DNAH10 (NM\_207437)
- GPSM1 (NM\_015597)
- ASXL1 (NM\_015338)
- DNAH1 (NM\_015512)
- DNAH6 (NM\_001370)
- FAT1 (NM\_005245)
- MDN1 (NM\_014611)
- PTPN11 (NM\_002834)
- SYNE1 (NM\_033071)
- ALMS1 (NM\_015120)
- C10orf68 (NM\_024688)
- CCDC88C (NM\_001080414)
- DNAH11 (NM\_003777)
- DNAH3 (NM\_017539)
- DNAH9 (NM\_001372)

**A** (Sort By Gene)  
**B** (Variant Data)  
**C** (Gene List)

Control what shows up here

# Variant View

Gene Search:

Alternative Transcripts:

**Variants**

Mutation Type  
Reference A.A.s  
Variant A.A.s

**Transcript**

trans-anon

**Protein**

A.A. Chain  
Domains  
Regions  
Active Sites  
Bindings  
Mod. Residue

**Variant Data**

| Patient ID | Chr. | Coord. | Ref Base | Var Base | dbSNP129 | dbSNP135 | dbSNP137 | COSMIC  | A.A. Chng. | Gene      | Ref. Gene  |
|------------|------|--------|----------|----------|----------|----------|----------|---------|------------|-----------|------------|
| pid-anon   | 11   | 288816 | G        | T        | .        | .        | .        | *13028, | G60V       | gene-anon | trans-anon |
| pid-anon   | 11   | 288816 | G        | T        | .        | .        | .        | *13012, | D61Y       | gene-anon | trans-anon |
| pid-anon   | 11   | 288819 | G        | T        | .        | rs121918 | .        | *13014, | A72S       | gene-anon | trans-anon |
| pid-anon   | 11   | 288819 | C        | T        | .        | .        | .        | *13035, | E76D       | gene-anon | trans-anon |
| pid-anon   | 11   | 288821 | G        | C        | .        | .        | .        | *13016, | E76Q       | gene-anon | trans-anon |
| pid-anon   | 11   | 288821 | A        | G        | .        | rs121918 | .        | *13017, | E76G       | gene-anon | trans-anon |
| pid-anon   | 11   | 288821 | G        | T        | .        | .        | .        | .       | E76D       | gene-anon | trans-anon |
| pid-anon   | 11   | 292688 | T        | A        | .        | rs121918 | .        | *13020, | S502T      | gene-anon | trans-anon |
| pid-anon   | 11   | 292688 | T        | G        | .        | .        | .        | *13020, | S502A      | gene-anon | trans-anon |
| pid-anon   | 11   | 292688 | C        | T        | .        | .        | .        | 13023   | S502L      | gene-anon | trans-anon |

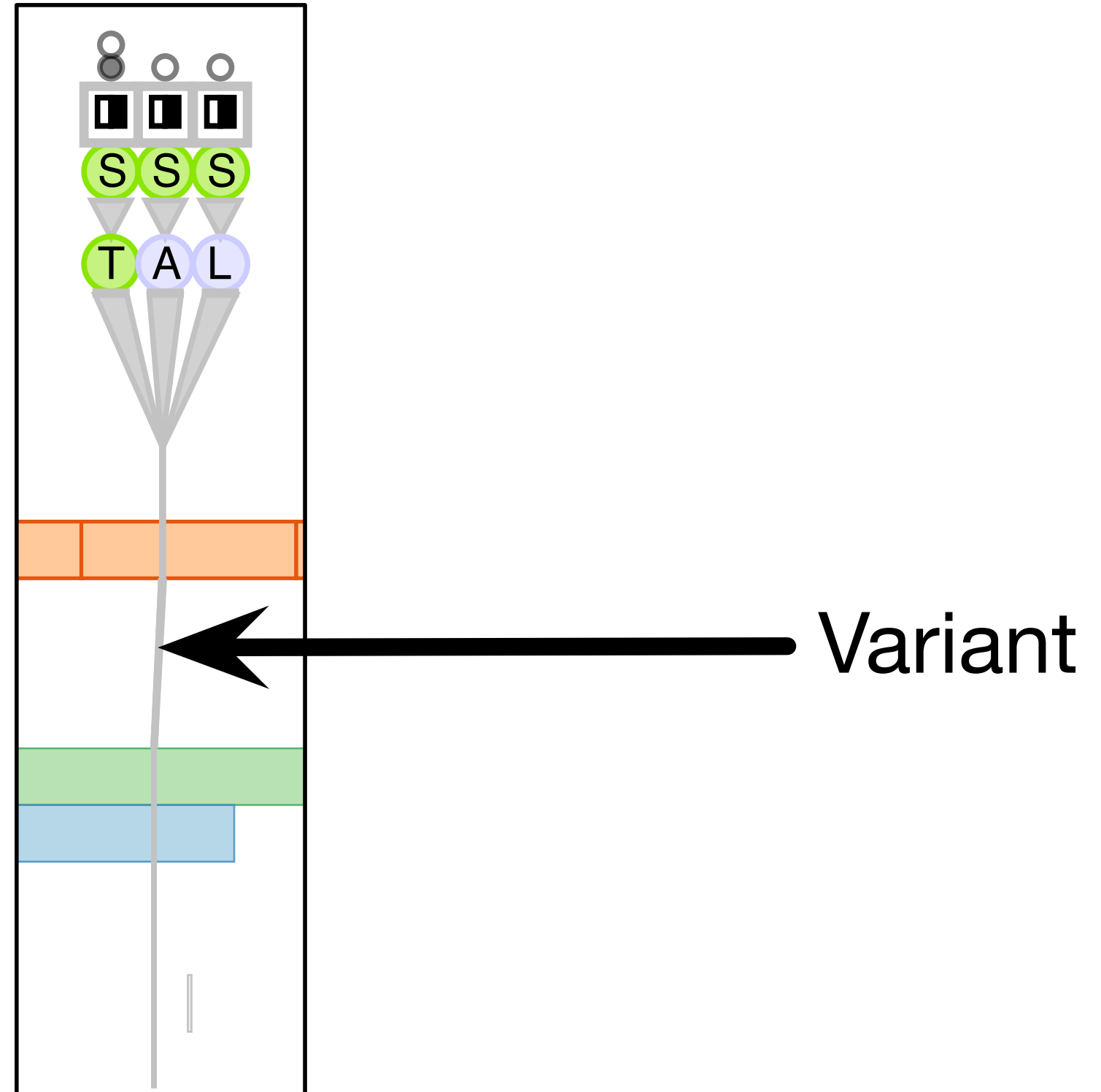
Sort By Gene:  
Alpha Cluster Score Variant Count

- DNMT3A (NM\_022552)
- IDH2 (NM\_002168)
- FLT3 (NM\_004119)
- ANKRD36 (NM\_001164315)
- ARID1B (NM\_017519)
- STAG2 (NM\_001042749)
- TNRC18 (NM\_001080495)
- WT1 (NM\_000378)
- ABCA13 (NM\_152701)
- CEBPA (NM\_004364)
- TET2 (NM\_001127208)
- DNAH10 (NM\_207437)
- GPSM1 (NM\_015597)
- ASXL1 (NM\_015338)
- DNAH1 (NM\_015512)
- DNAH6 (NM\_001370)
- FAT1 (NM\_005245)
- MDN1 (NM\_014611)
- PTPN11 (NM\_002834)
- DNAH11 (NM\_003777)
- DNAH3 (NM\_017539)
- DNAH9 (NM\_001372)

Peripheral supporting data

# Design information-dense visual encoding

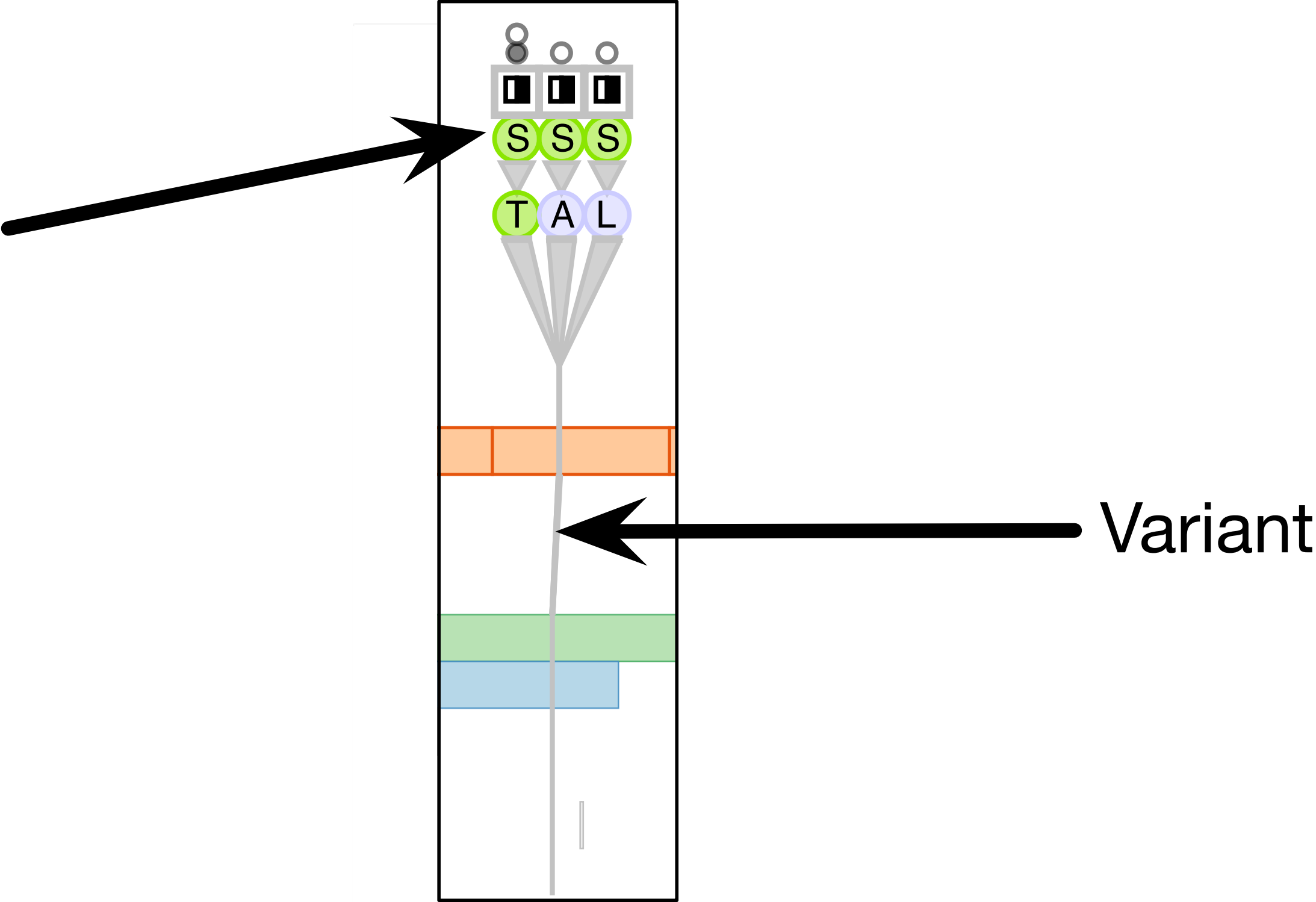
- show all attributes necessary for variant analysis
  - match salience with importance for analysis task
- variant not just a thin line!
- emphasize with high salience
  - collocated variants fan out at top
  - grey variant vertical stroke intersects horizontal colored protein regions





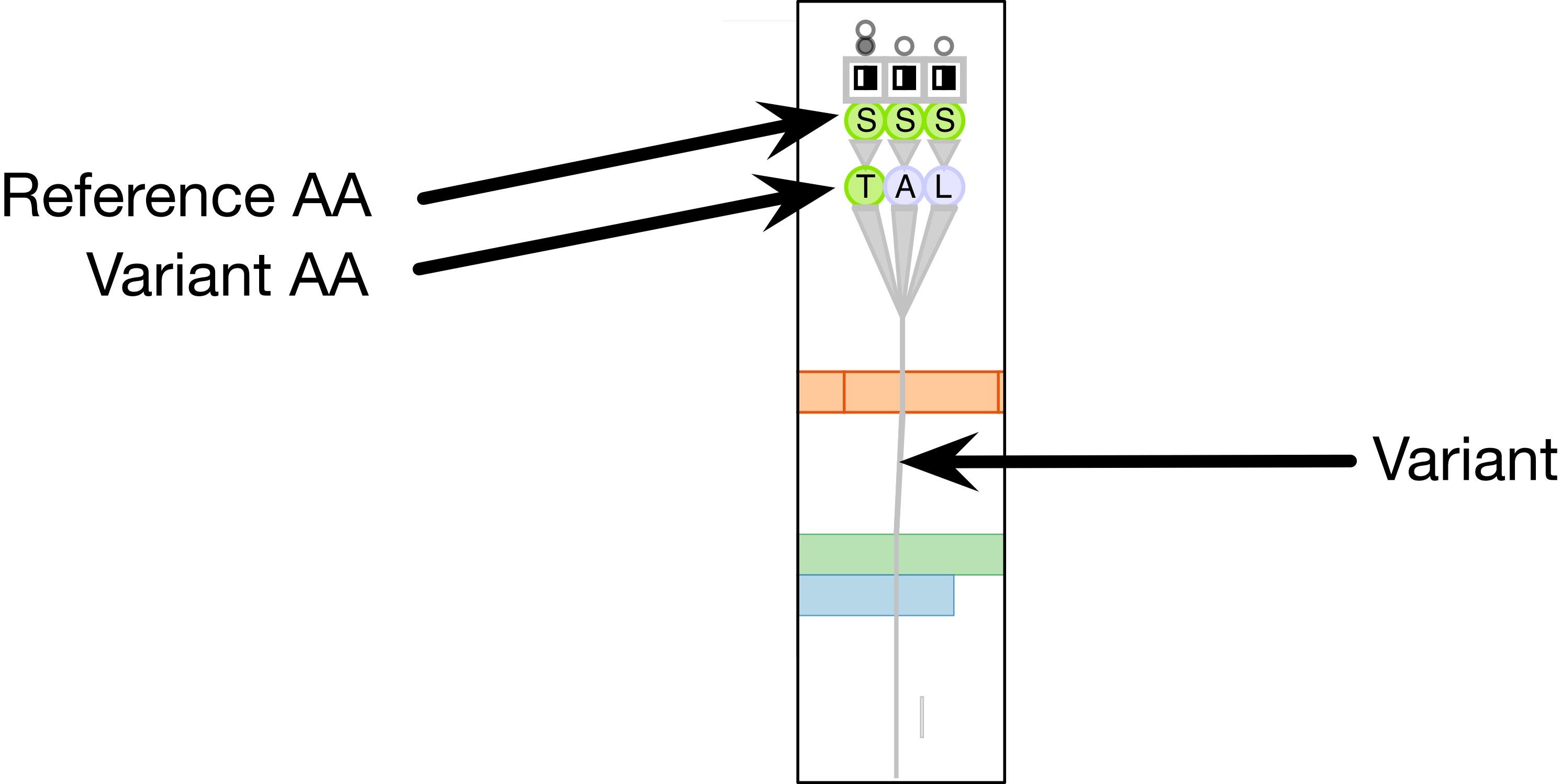
# Design information-dense visual encoding

Reference AA





# Design information-dense visual encoding



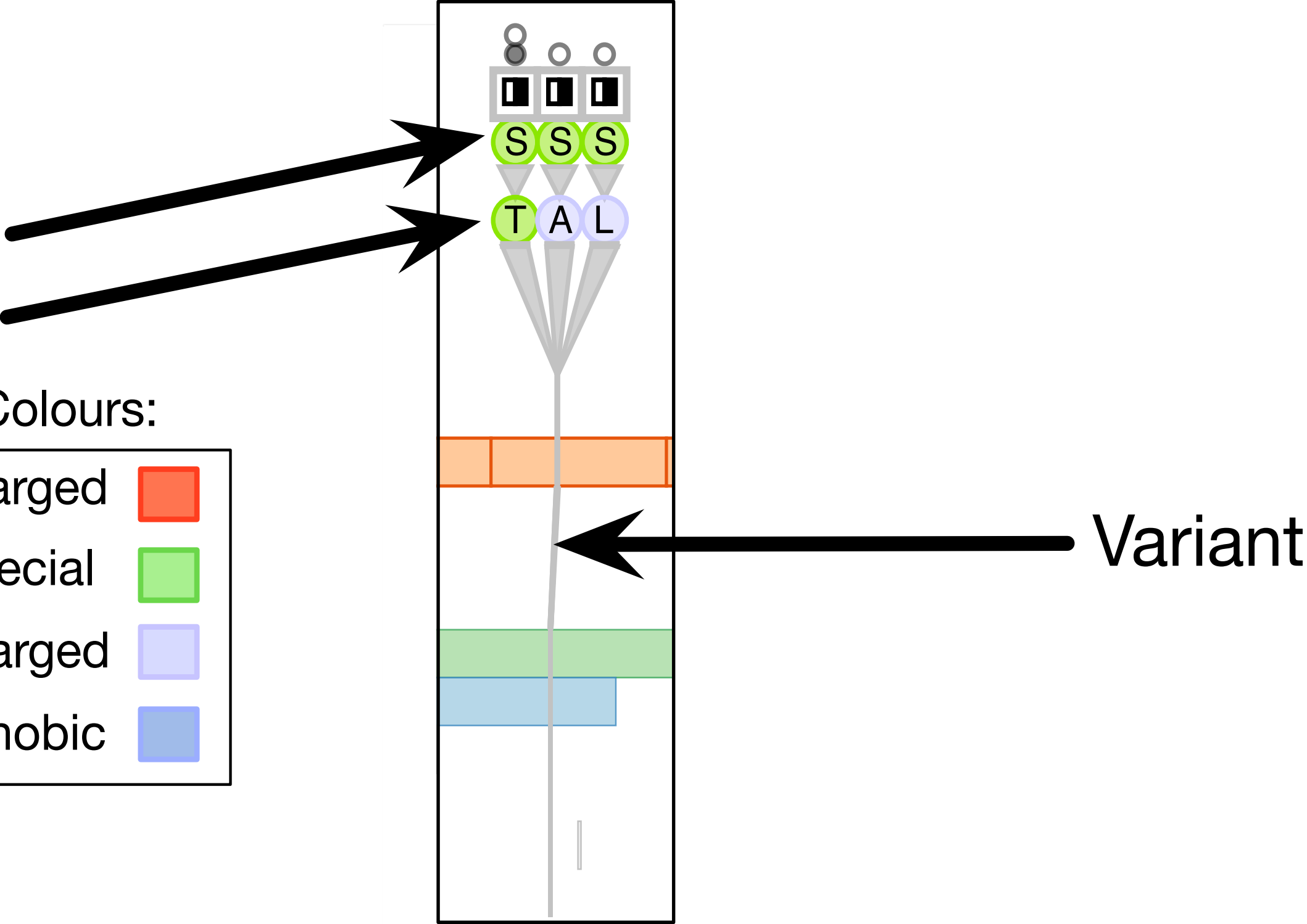
# Design information-dense visual encoding

Reference AA

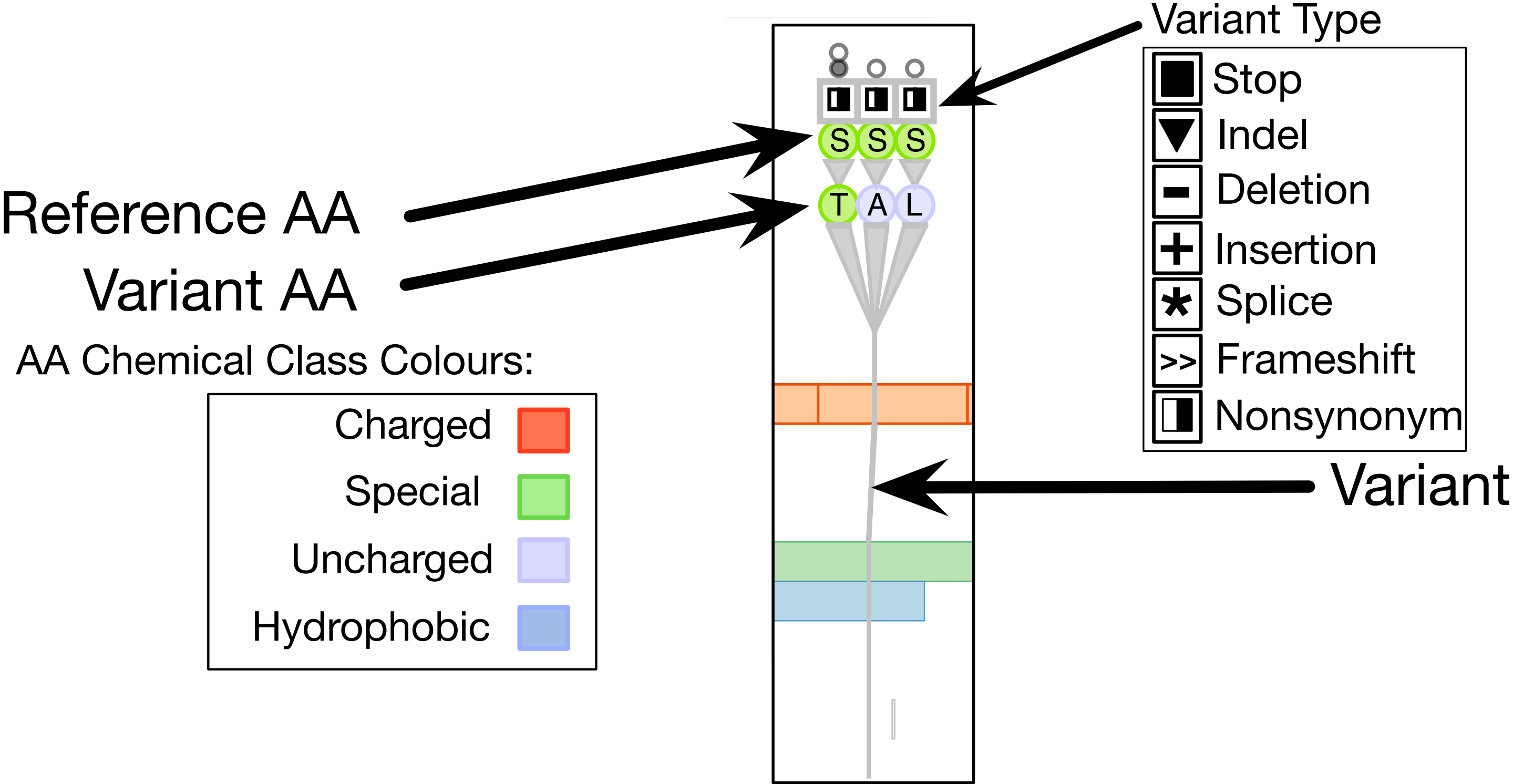
Variant AA

AA Chemical Class Colours:

|             |  |
|-------------|--|
| Charged     | <span style="color: red;">■</span>       |
| Special     | <span style="color: green;">■</span>     |
| Uncharged   | <span style="color: lightblue;">■</span> |
| Hydrophobic | <span style="color: blue;">■</span>      |



# Design information-dense visual encoding



# Design information-dense visual encoding

Known Database

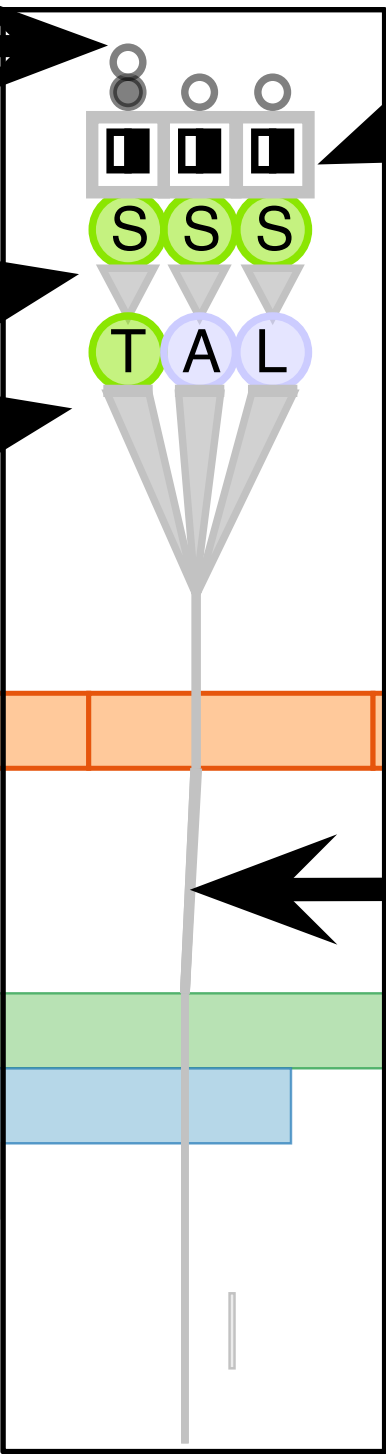
- Known Harmless
- Known Cancer

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 ■

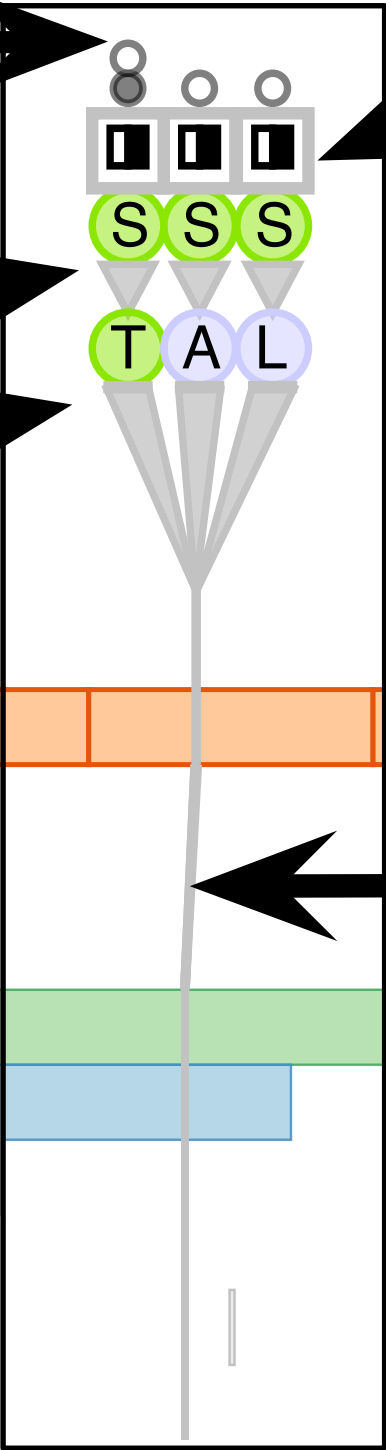
Variant Type

- Stop
- ▼ Indel
- ▬ Deletion
- ⊕ Insertion
- \* Splice
- >> Frameshift
- ▣ Nonsynonym

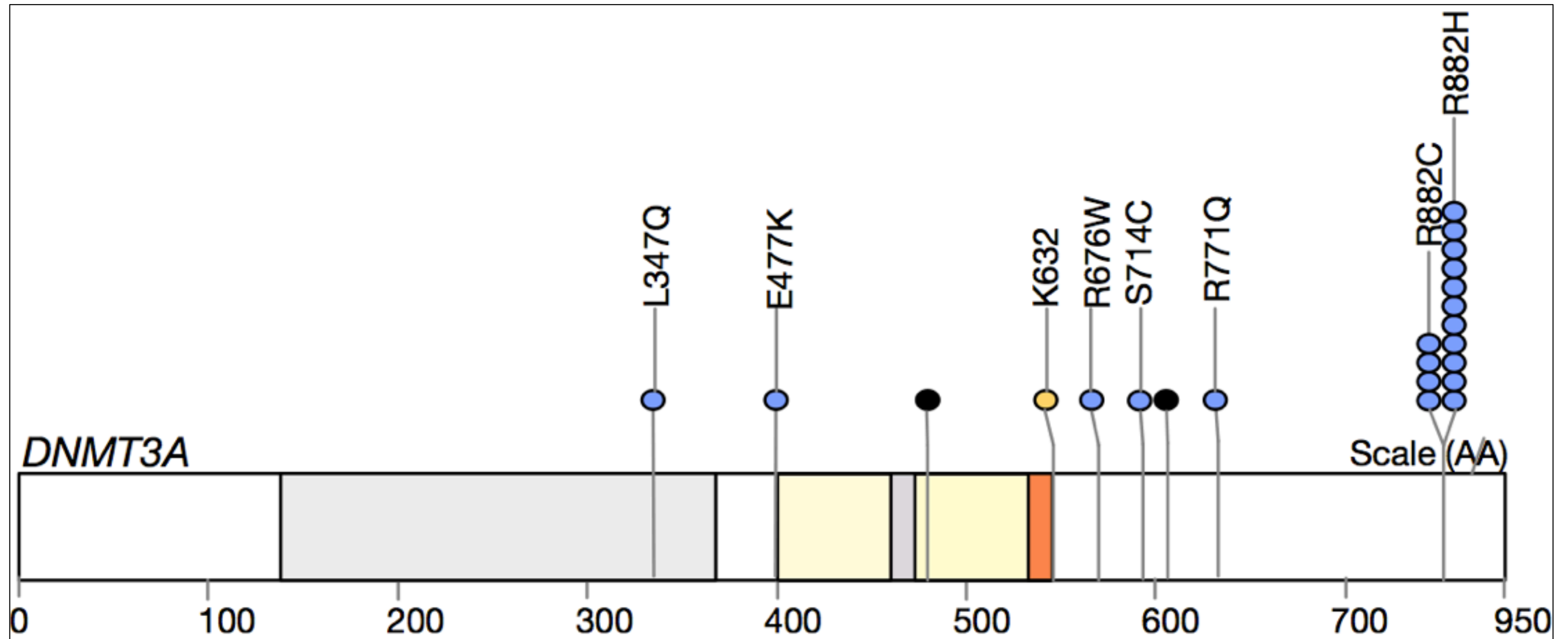
Variant

Transcript/Region Colours:

- Transcript ■
- AA Chain ■
- All Other Regions ■
- Non-Intersected Regions ■

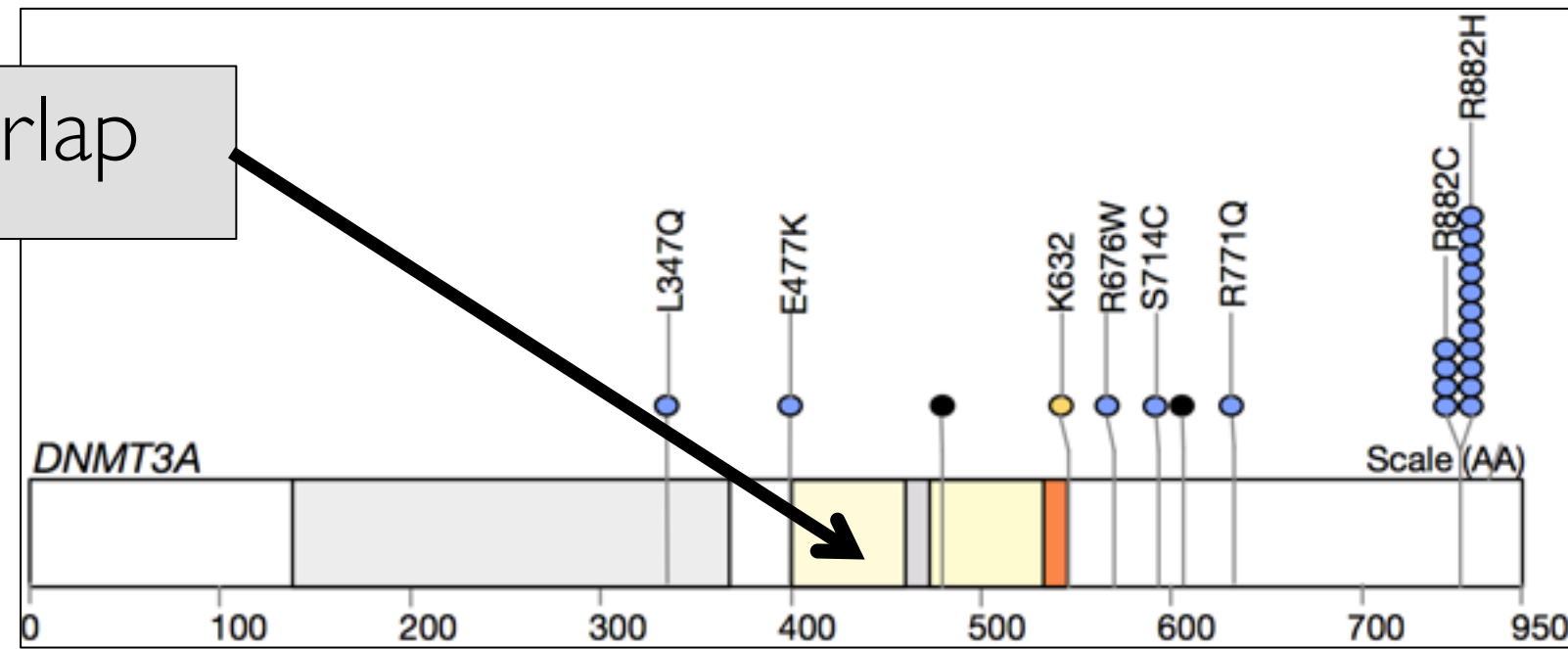


# Previous work targeted at variant analysis: MuSiC

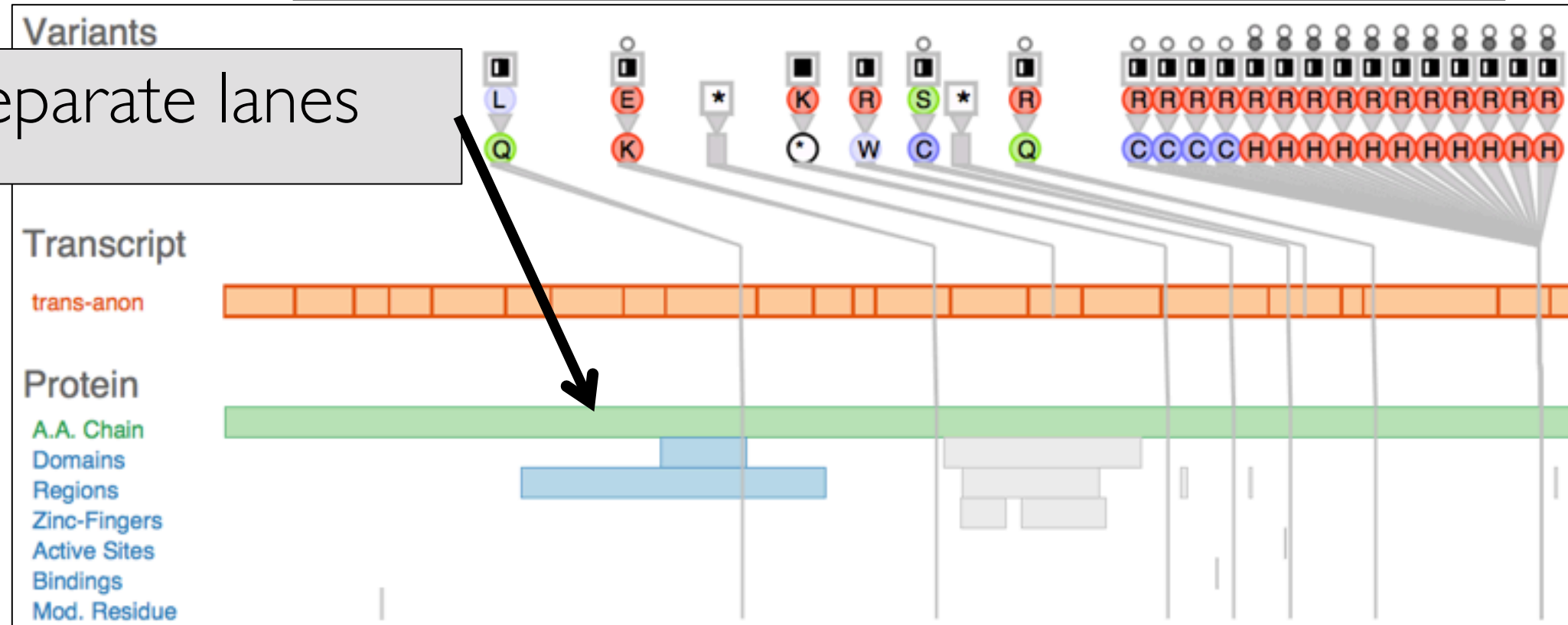


# Side-by-side comparison: MuSiC vs Variant View

Protein regions can overlap

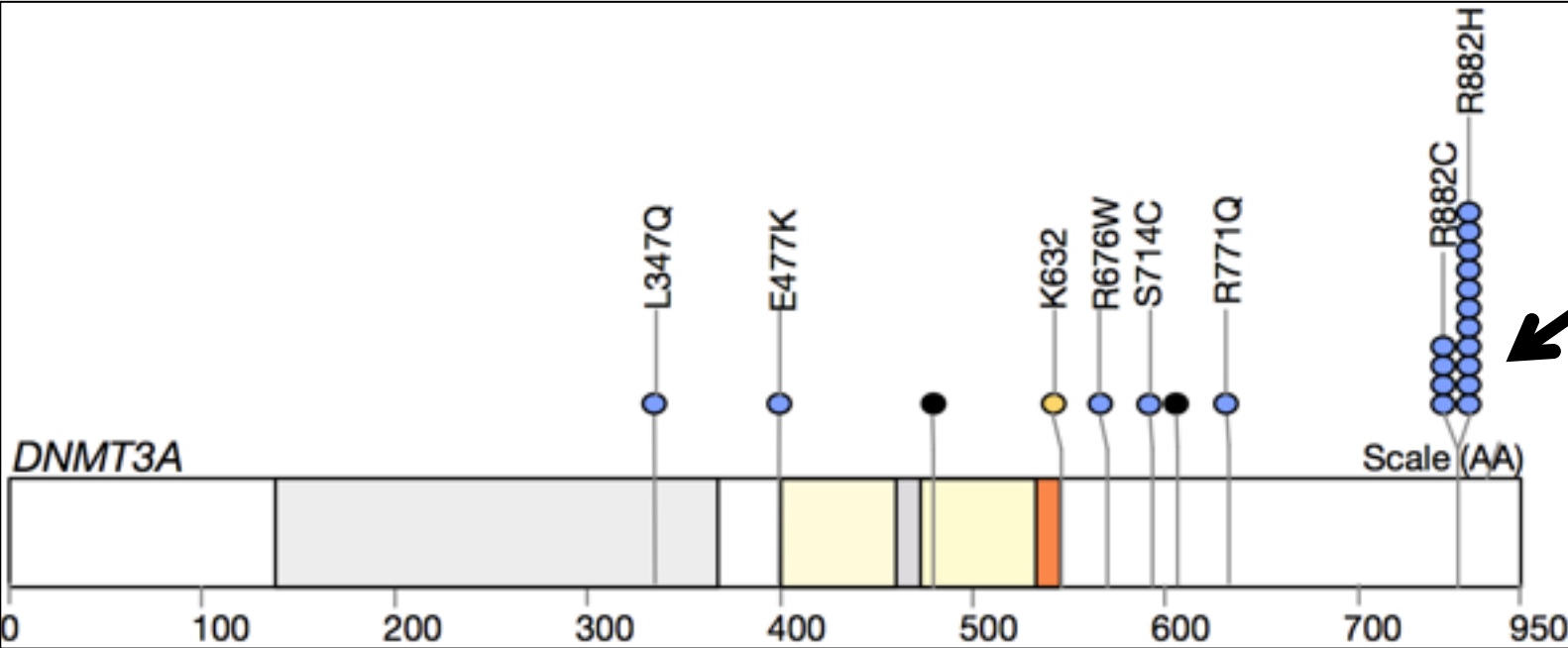


Regions get separate lanes

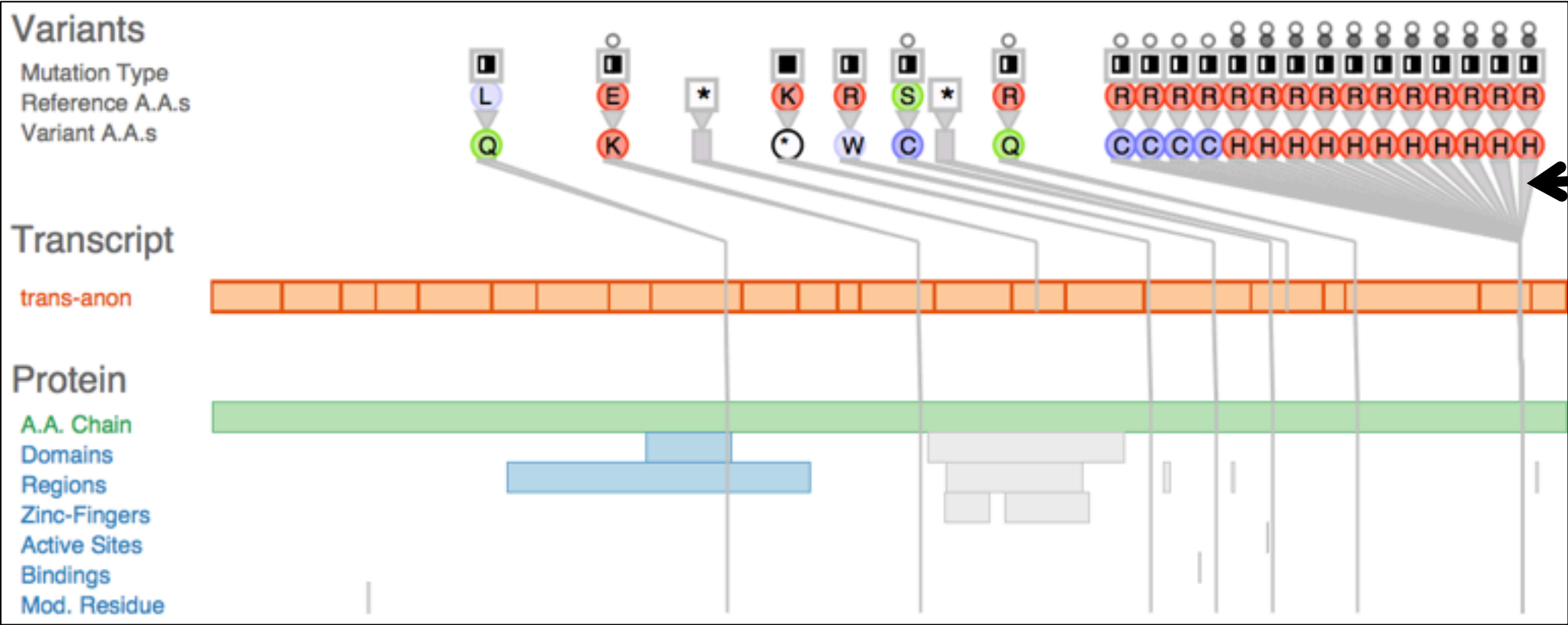




# Side-by-side comparison: MuSiC vs Variant View



Many collocated variants



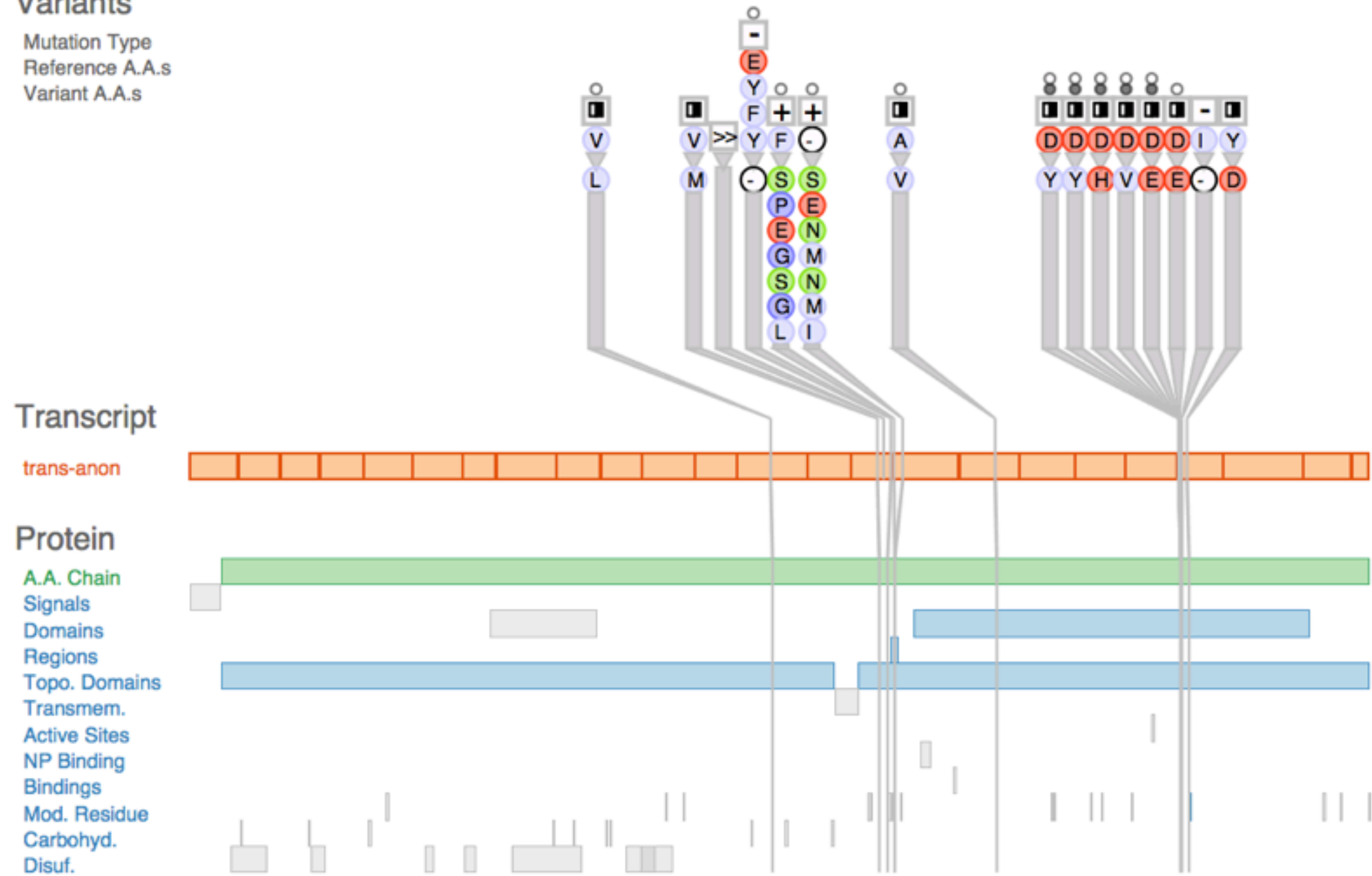
Large bloom of repeated elements: more salient

# Results

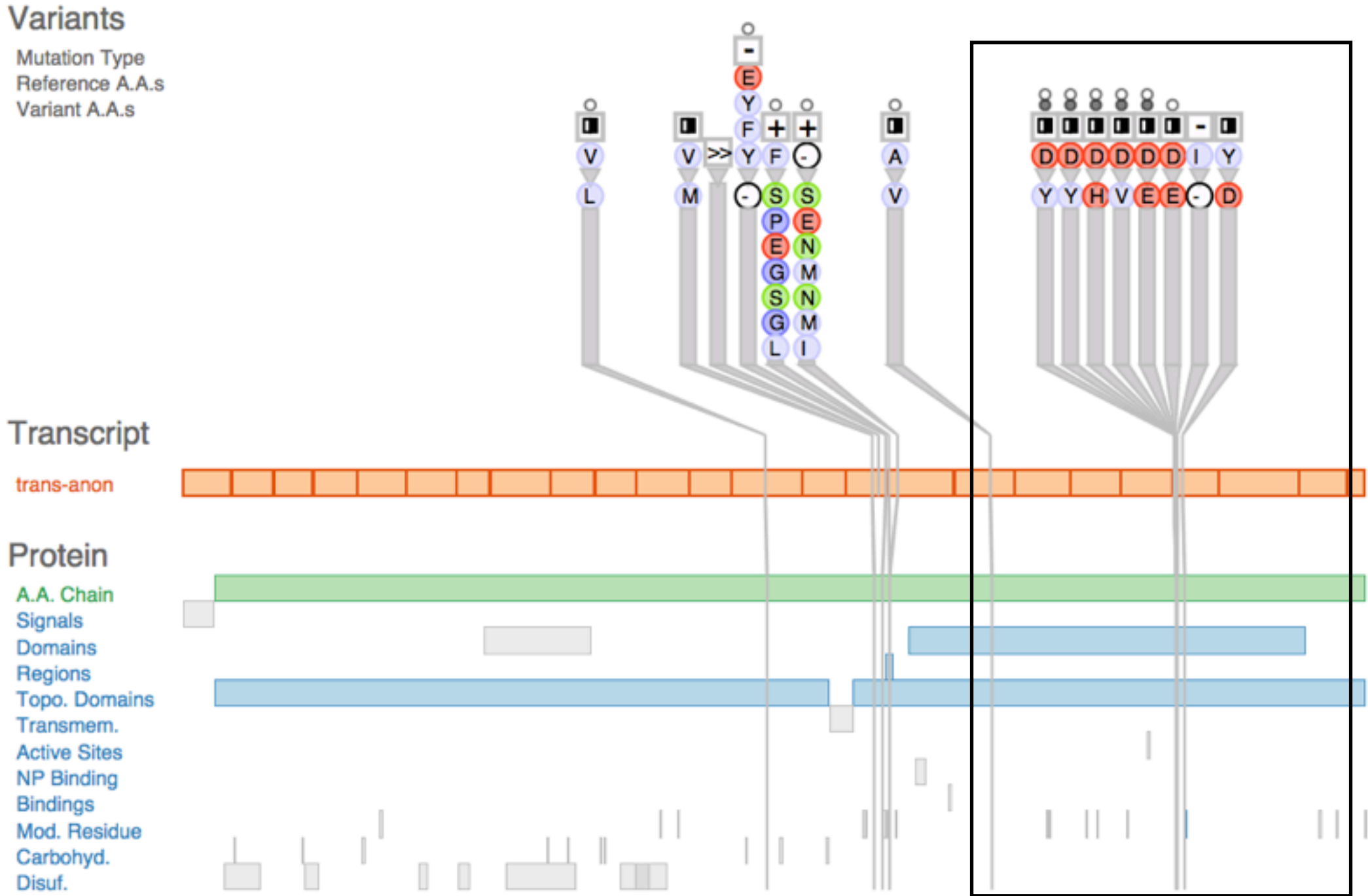
# Verify known leukemia gene: Highly scored by sorting metric

## Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



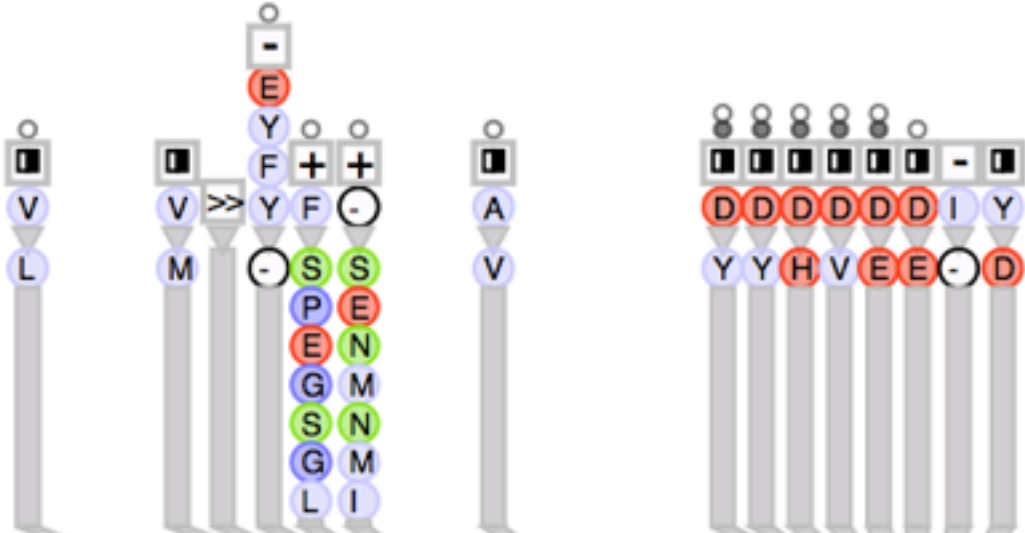
# Visual inspection reveals collocation of variants



# Several functional protein regions affected

## Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



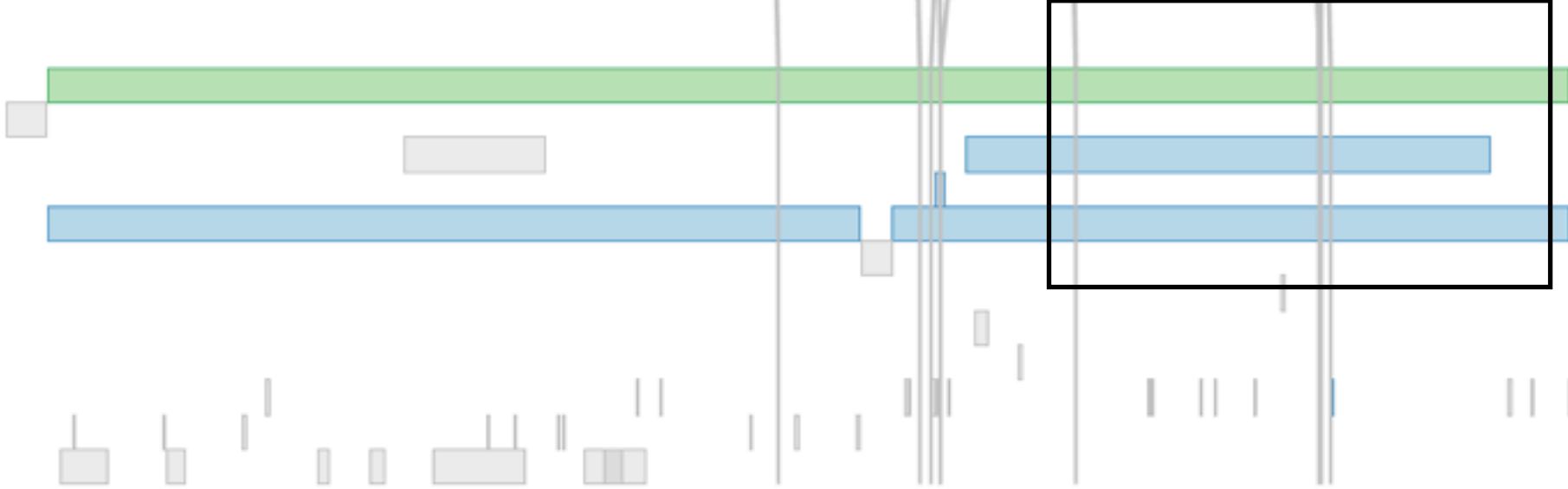
## Transcript

trans-anon



## Protein

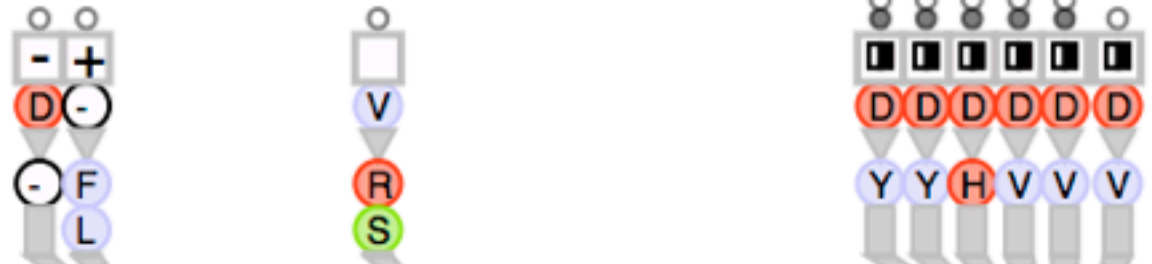
A.A. Chain  
Signals  
Domains  
Regions  
Topo. Domains  
Transmem.  
Active Sites  
NP Binding  
Bindings  
Mod. Residue  
Carbohyd.  
Disuf.



# Highly scored by metric: not previously known, good candidate

## Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



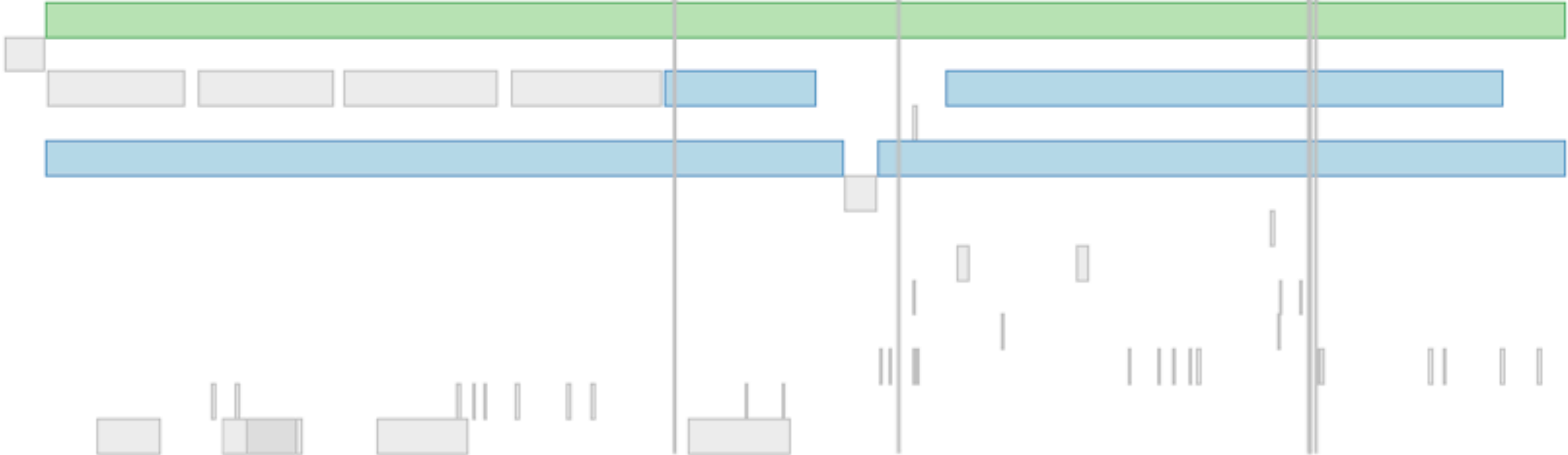
## Transcript

trans-anon



## Protein

A.A. Chain  
Signals  
Domains  
Regions  
Topo. Domains  
Transmem.  
Active Sites  
NP Binding  
Metal Bind.  
Bindings  
Mod. Residue  
Carbohyd.  
Disuf.



# Protein chemical class change evident

## Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



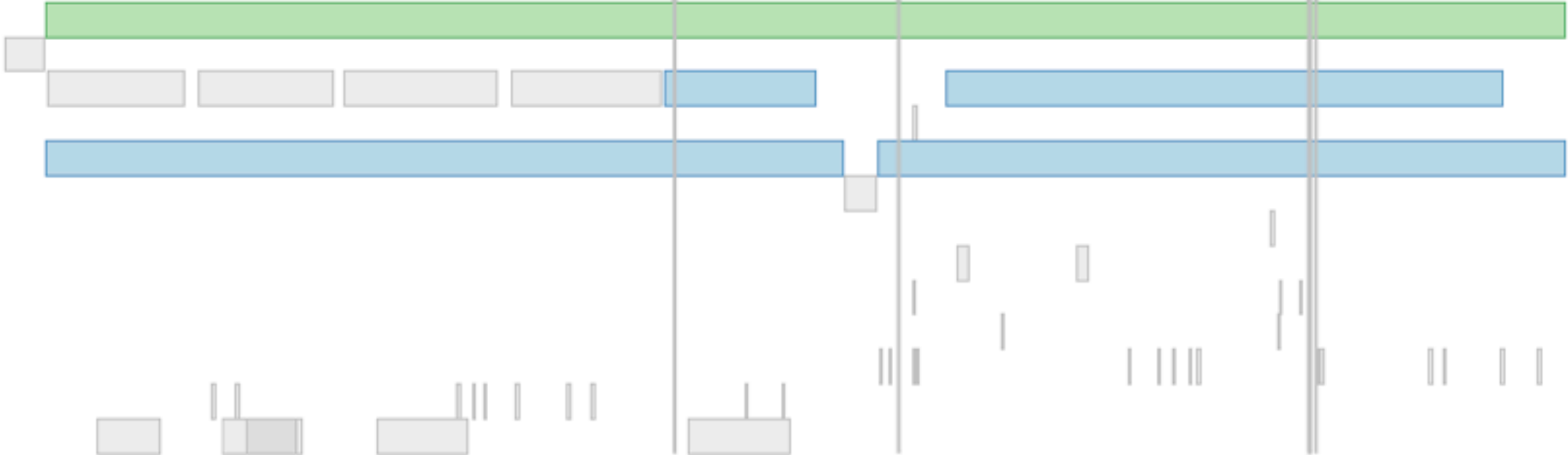
## Transcript

trans-anon



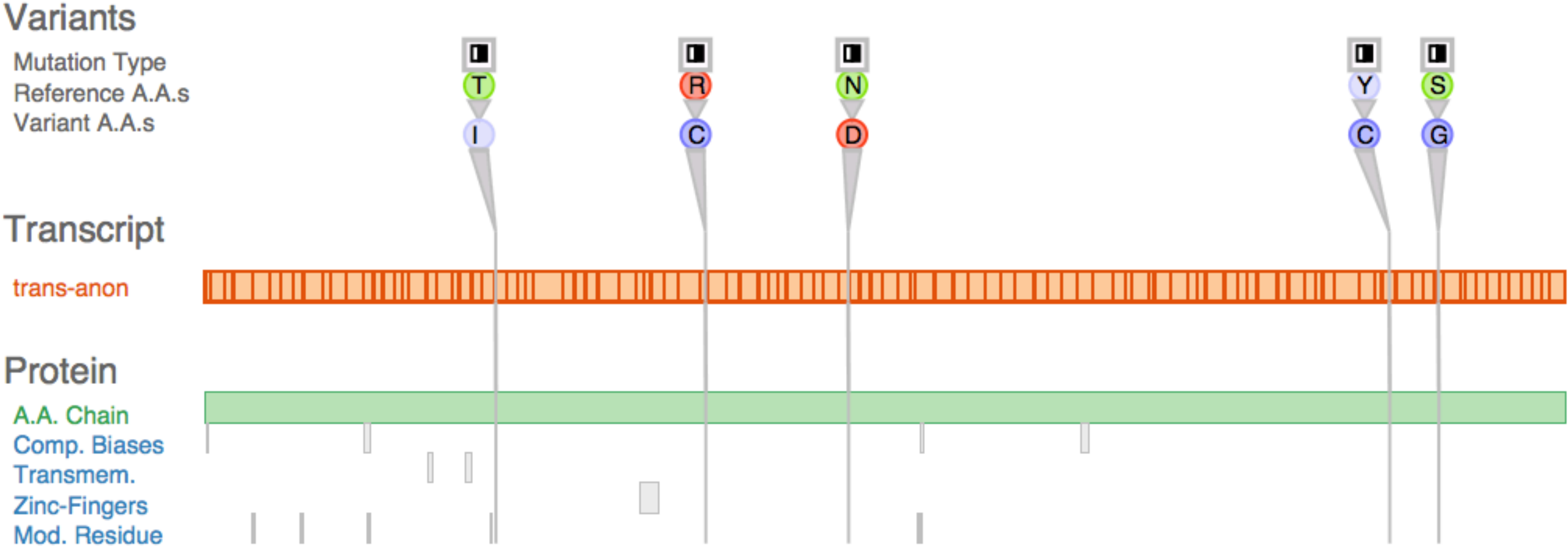
## Protein

A.A. Chain  
Signals  
Domains  
Regions  
Topo. Domains  
Transmem.  
Active Sites  
NP Binding  
Metal Bind.  
Bindings  
Mod. Residue  
Carbohyd.  
Disuf.





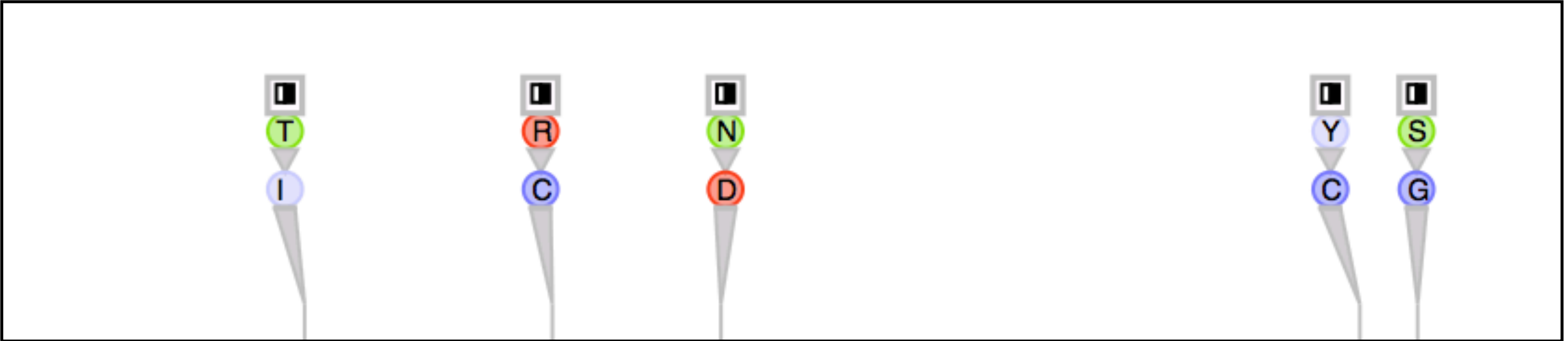
# In contrast, low scoring gene



# No collocation of variants

## Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



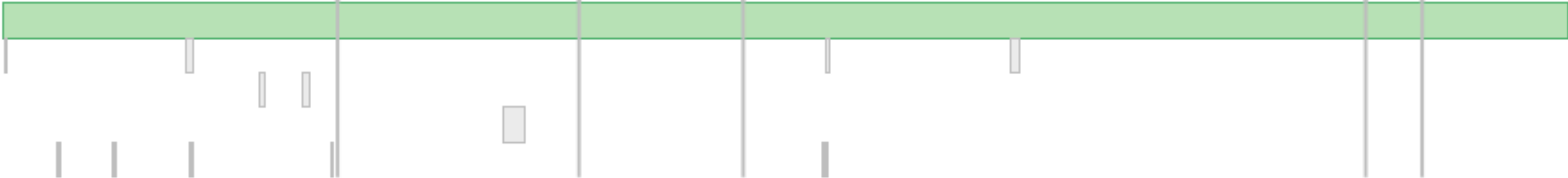
## Transcript

trans-anon



## Protein

A.A. Chain  
Comp. Biases  
Transmem.  
Zinc-Fingers  
Mod. Residue



# Mostly unaffected protein regions

## Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



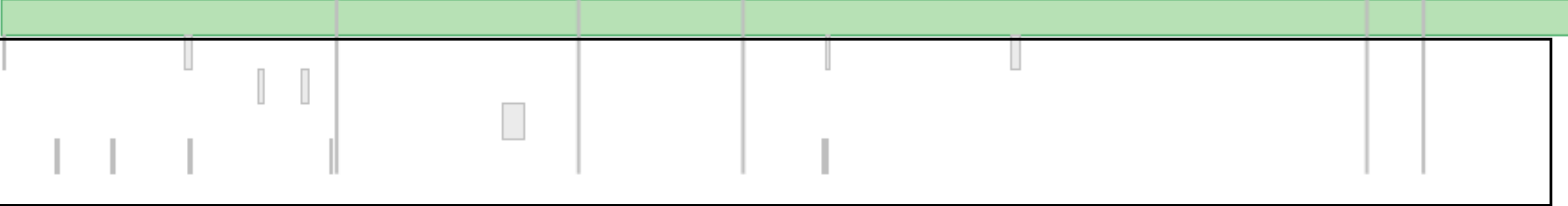
## Transcript

trans-anon



## Protein

A.A. Chain  
Comp. Biases  
Transmem.  
Zinc-Fingers  
Mod. Residue



# 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

Select Patient:

Patient Genes:

Alternative Transcripts:

**Variants**

Mutation Type  
Reference A.A.s  
Variant A.A.s

**Transcript**

trans-anon

**Protein**

A.A. Chain  
Regions  
Comp. Biases  
Zinc-Fingers  
Mod. Residue

**Variant Details**

| Variant ID | Chr. Coord. | Ref Base | Var Base | Effect Level | Effect Type | Gene Name | Trans. Name | Prot. Coord. |
|------------|-------------|----------|----------|--------------|-------------|-----------|-------------|--------------|
| pid-anon   | 31022959    | T        | C        | MODERATE     | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31022959    | T        | C        |              | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31023029    | G        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | K838N        |
| pid-anon   | 31024274    | T        | C        | LOW          | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024274    | T        | C        |              | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024450    | C        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | A1312V       |
| pid-anon   | 31024704    | G        | A        |              | NON_SYNONY  | gene-anon | trans-anon  | G1397S       |
| pid-anon   | 31025163    | A        | G        | MODIFIER     | UTR_3_PRIM  | gene-anon | trans-anon  | -            |

**Comparison Modes**

Show Patient Data Only

Show Patient + Neighborhood

# Navigate through patient data with list

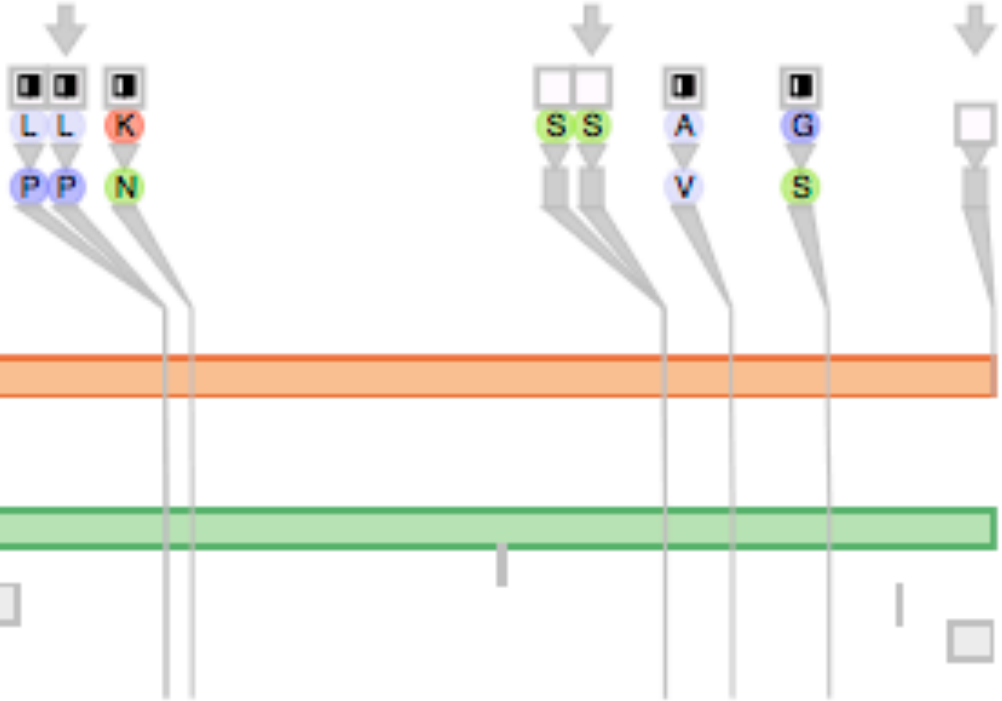
Select Patient:

Patient Genes:

Alternative Transcripts:

### Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



### Transcript

trans-anon

### Protein

A.A. Chain  
Regions  
Comp. Biases  
Zinc-Fingers  
Mod. Residue

### Variant Details

| Variant ID | Chr. Coord. | Ref Base | Var Base | Effect Level | Effect Type | Gene Name | Trans. Name | Prot. Coord. |
|------------|-------------|----------|----------|--------------|-------------|-----------|-------------|--------------|
| pid-anon   | 31022959    | T        | C        | MODERATE     | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31022959    | T        | C        |              | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31023029    | G        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | K838N        |
| pid-anon   | 31024274    | T        | C        | LOW          | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024274    | T        | C        |              | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024450    | C        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | A1312V       |
| pid-anon   | 31024704    | G        | A        |              | NON_SYNONY  | gene-anon | trans-anon  | G1397S       |
| pid-anon   | 31025183    | A        | G        | MODIFIER     | UTR_3_PRIM  | gene-anon | trans-anon  | -            |

### Comparison Modes

- Show Patient Data Only
- Show Patient + Neighborhood

# Patient data emphasized with arrows

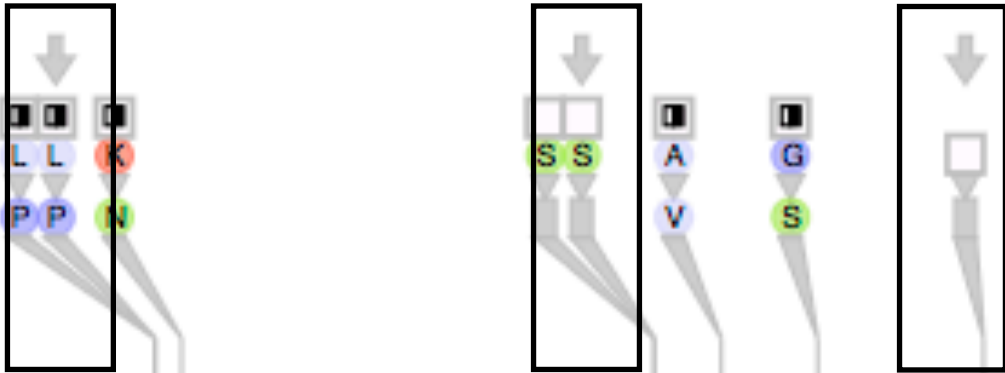
Select Patient:

Patient Genes:

Alternative Transcripts:

### Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



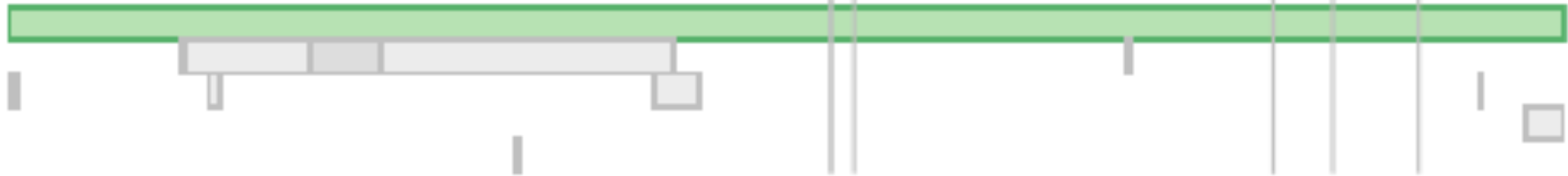
### Transcript

trans-anon



### Protein

A.A. Chain  
Regions  
Comp. Biases  
Zinc-Fingers  
Mod. Residue



### Variant Details

| Variant ID | Chr. Coord. | Ref Base | Var Base | Effect Level | Effect Type | Gene Name | Trans. Name | Prot. Coord. |
|------------|-------------|----------|----------|--------------|-------------|-----------|-------------|--------------|
| pid-anon   | 31022959    | T        | C        | MODERATE     | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31022959    | T        | C        |              | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31023029    | G        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | K838N        |
| pid-anon   | 31024274    | T        | C        | LOW          | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024274    | T        | C        |              | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024450    | C        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | A1312V       |
| pid-anon   | 31024704    | G        | A        |              | NON_SYNONY  | gene-anon | trans-anon  | G1397S       |
| pid-anon   | 31025163    | A        | G        | MODIFIER     | UTR_3_PRIM  | gene-anon | trans-anon  | -            |

### Comparison Modes

- Show Patient Data Only
- Show Patient + Neighborhood



# Patient has same harmful L to P mutation

Select Patient:

Patient Genes:

Alternative Transcripts:

### Variants

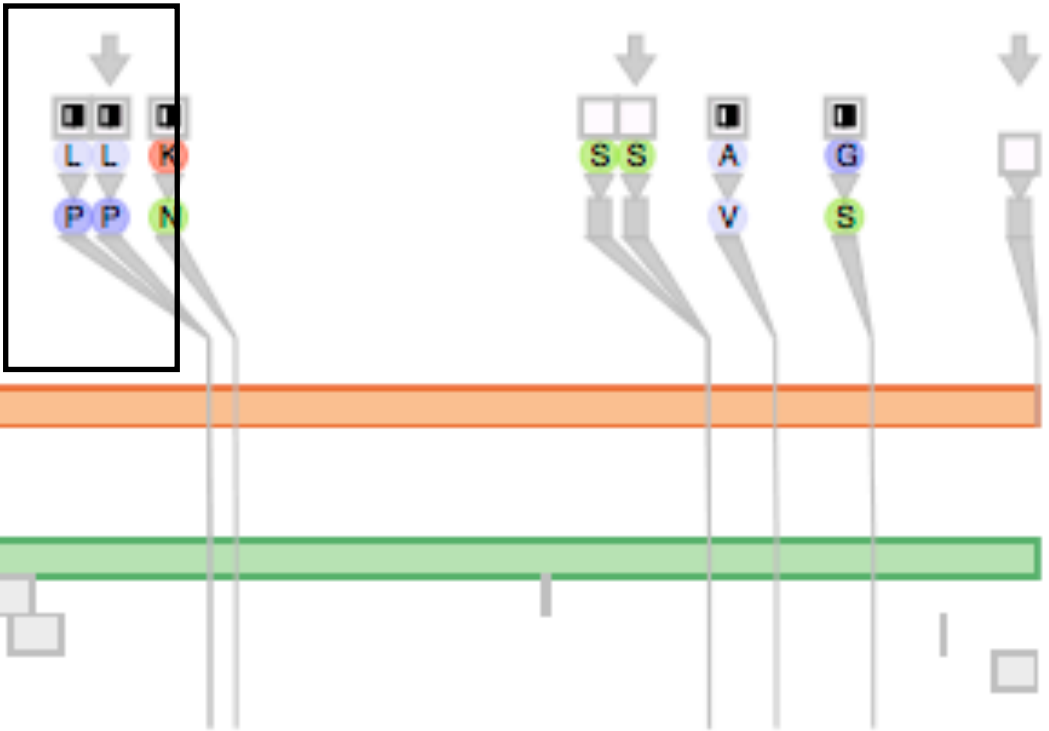
Mutation Type  
Reference A.A.s  
Variant A.A.s

### Transcript

trans-anon

### Protein

A.A. Chain  
Regions  
Comp. Biases  
Zinc-Fingers  
Mod. Residue



### Variant Details

| Variant ID | Chr. Coord. | Ref Base | Var Base | Effect Level | Effect Type | Gene Name | Trans. Name | Prot. Coord. |
|------------|-------------|----------|----------|--------------|-------------|-----------|-------------|--------------|
| pid-anon   | 31022959    | T        | C        | MODERATE     | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31022959    | T        | C        |              | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31023029    | G        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | K838N        |
| pid-anon   | 31024274    | T        | C        | LOW          | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024274    | T        | C        |              | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024450    | C        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | A1312V       |
| pid-anon   | 31024704    | G        | A        |              | NON_SYNONY  | gene-anon | trans-anon  | G1397S       |
| pid-anon   | 31025183    | A        | G        | MODIFIER     | UTR_3_PRIM  | gene-anon | trans-anon  | -            |

### Comparison Modes

- Show Patient Data Only
- Show Patient + Neighborhood

# Nonmatching variants

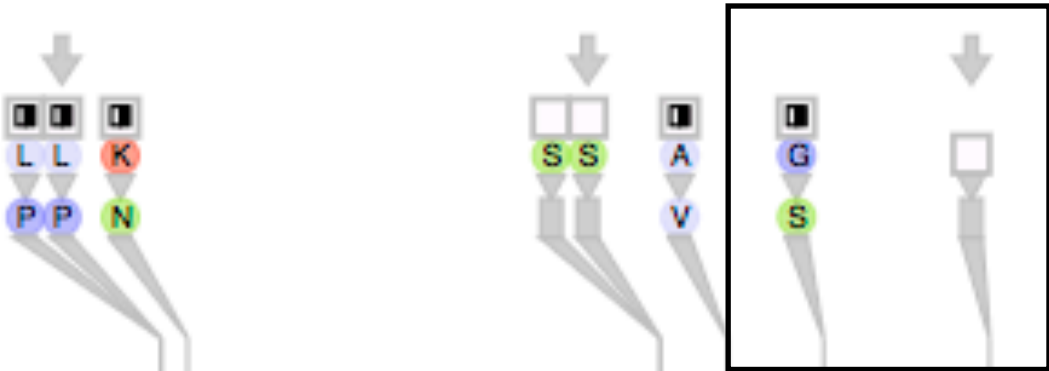
Select Patient:

Patient Genes:

Alternative Transcripts:

### Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



### Transcript

trans-anon

### Protein

A.A. Chain  
Regions  
Comp. Biases  
Zinc-Fingers  
Mod. Residue

### Variant Details

| Variant ID | Chr. Coord. | Ref Base | Var Base | Effect Level | Effect Type | Gene Name | Trans. Name | Prot. Coord. |
|------------|-------------|----------|----------|--------------|-------------|-----------|-------------|--------------|
| pid-anon   | 31022959    | T        | C        | MODERATE     | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31022959    | T        | C        |              | NON_SYNONY  | gene-anon | trans-anon  | L815P        |
| pid-anon   | 31023029    | G        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | K838N        |
| pid-anon   | 31024274    | T        | C        | LOW          | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024274    | T        | C        |              | SYNONYMOUS  | gene-anon | trans-anon  | S1253        |
| pid-anon   | 31024450    | C        | T        |              | NON_SYNONY  | gene-anon | trans-anon  | A1312V       |
| pid-anon   | 31024704    | G        | A        |              | NON_SYNONY  | gene-anon | trans-anon  | G1397S       |
| pid-anon   | 31025183    | A        | G        | MODIFIER     | UTR_3_PRIM  | gene-anon | trans-anon  | -            |

### Comparison Modes

- Show Patient Data Only
- Show Patient + Neighborhood

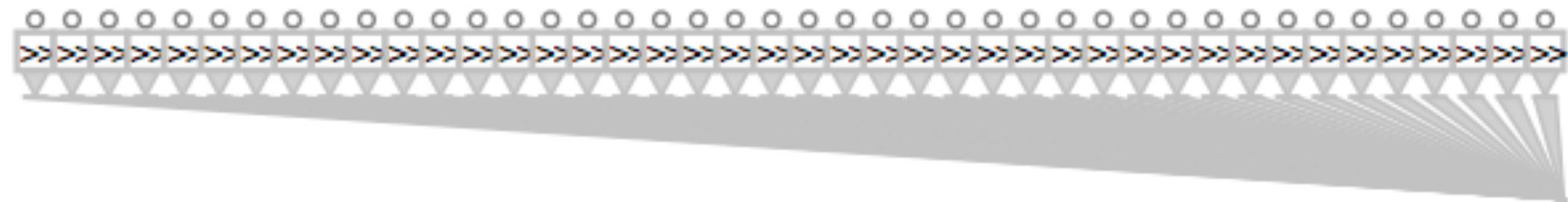
# Additional tasks

- task 3: debug pipeline
  - data cleansing before analysis
  - analysts originally thought pipeline fully debugged
    - no perceived need for vis support

# Tool revealed errors in the data

## Variants

Mutation Type  
Reference A.A.s  
Variant A.A.s



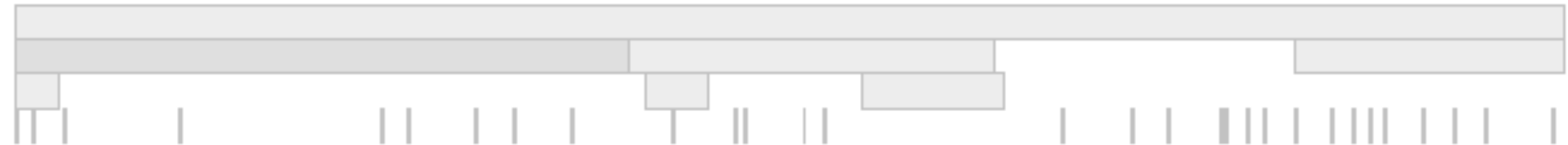
## Transcript

trans-anon



## Protein

A.A. Chain  
Regions  
Comp. Biases  
Mod. Residue



*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.*

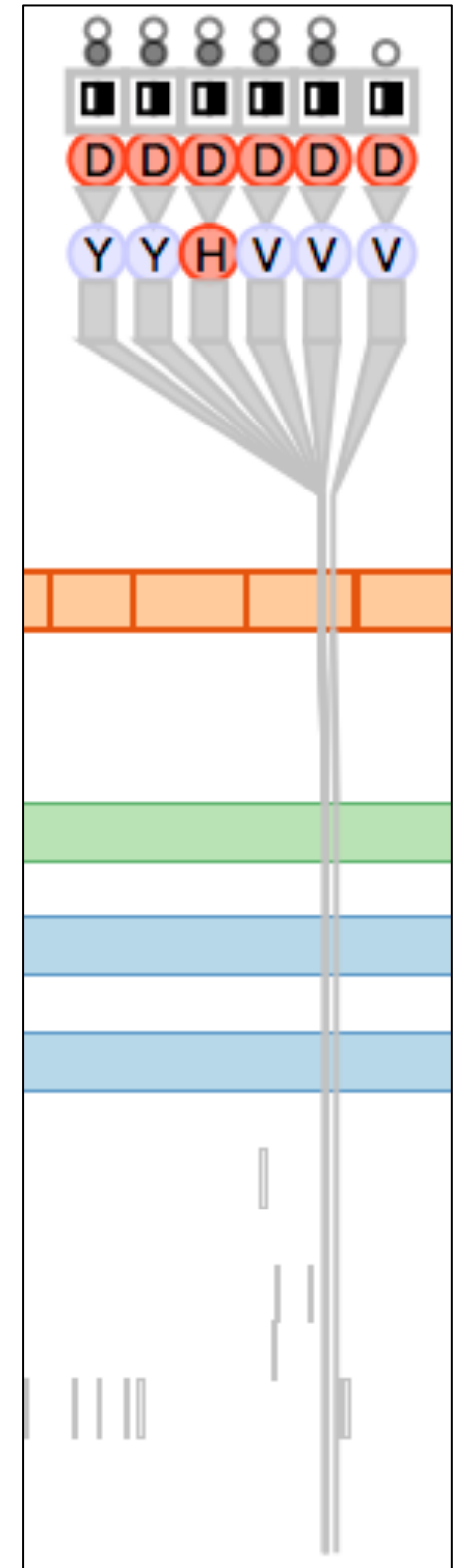
**- Analyst 3**

# Reflections: vis design guidelines

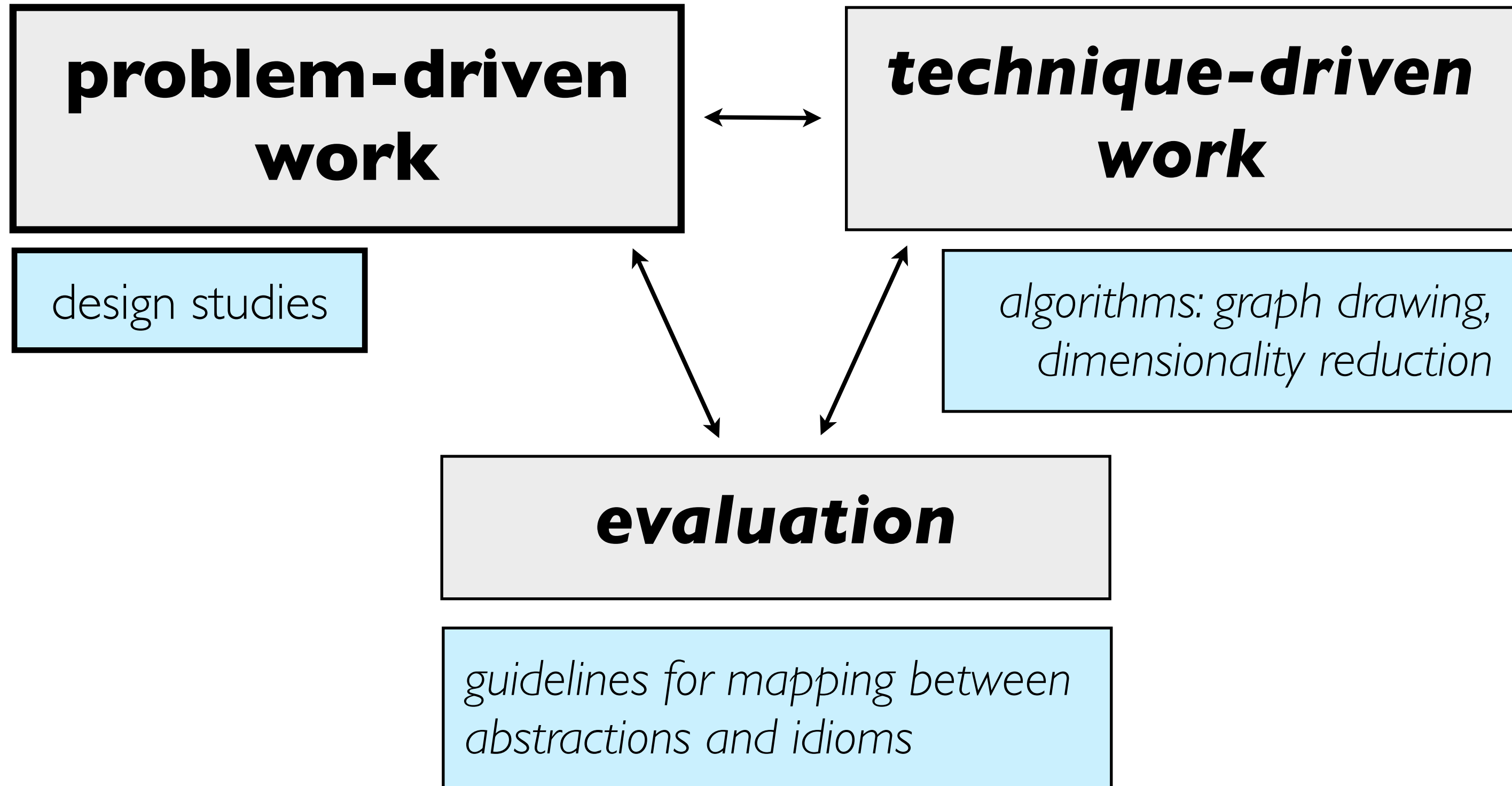
- transferrable to other domains
  - specialize first, generalize later
    - good for domains where with complex, multi-scale data
    - difficult to judge a priori which design elements will generalize
  - high-level considerations
    - identifying scales of interest
    - what to visually encode 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



# Research Interests



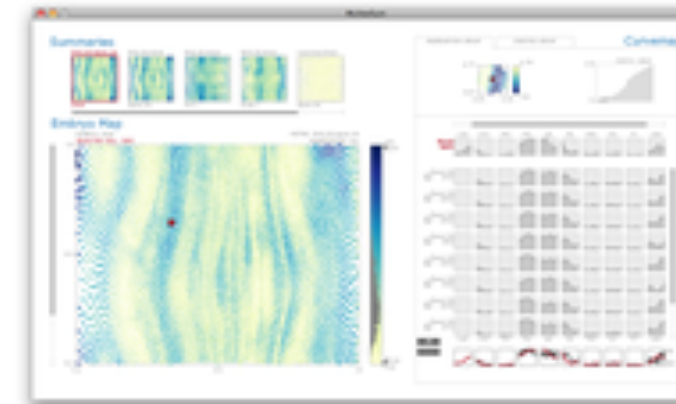
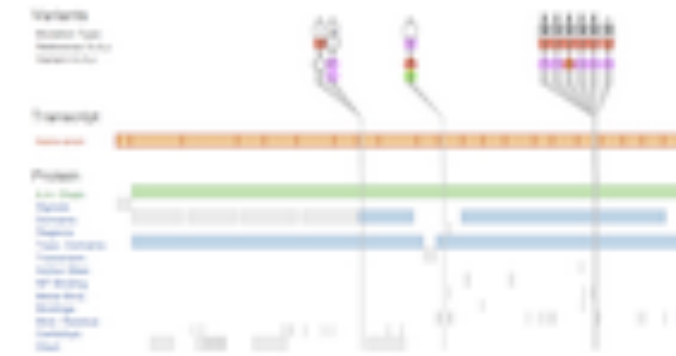


# Many Flavors of Work

- Problem-driven work
  - design studies
- Technique-driven work
  - algorithms, systems
- Evaluation
  - lab/field/data studies
- Theoretical foundations
  - models

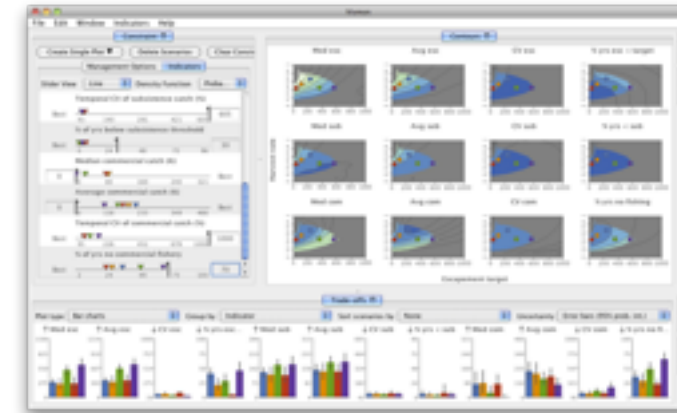
# Design Studies: Biology Domain

- Variant View: gene sequence variants
- MulteeSum, Pathline: comparative functional genomics
- MizBee: comparative genomics (synteny)
- Cerebral: gene expression + interaction network



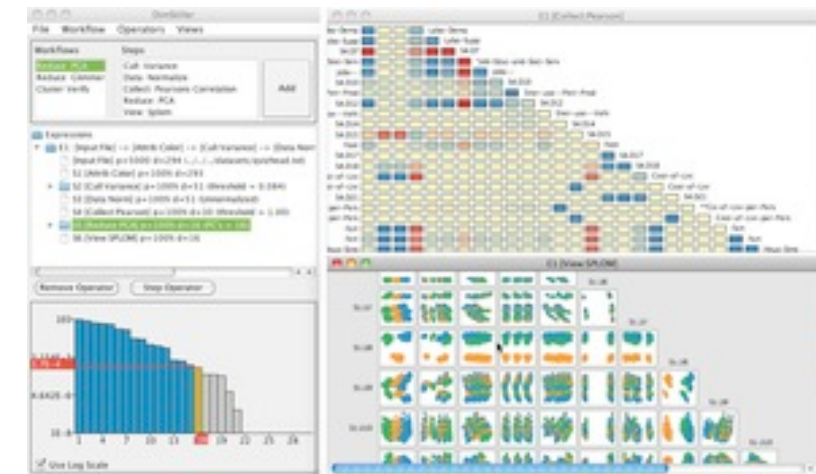
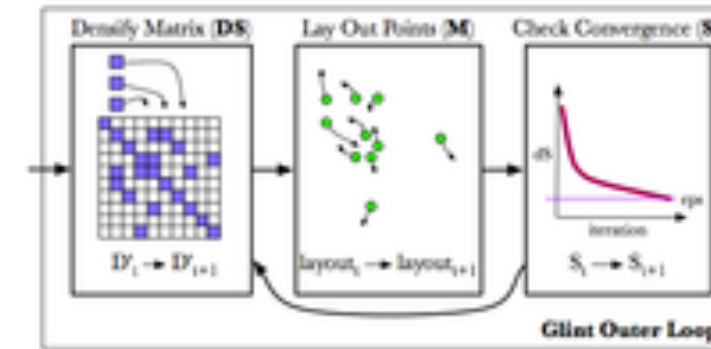
# Design Studies: Other Domains

- RelEx: automative networks
- Vismon: fisheries simulation/mgmt
- LiveRAC: large-scale system monitoring
- SessionViewer: web logs



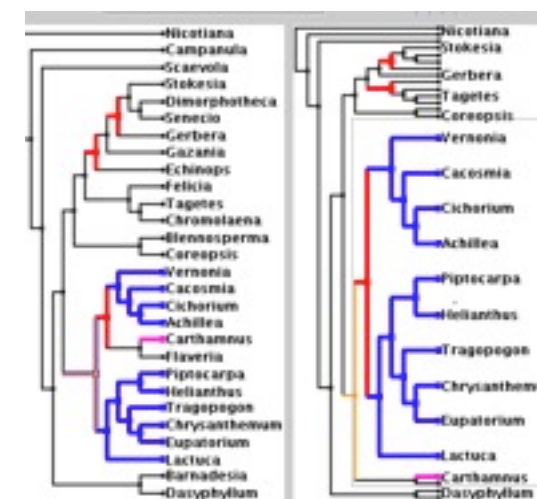
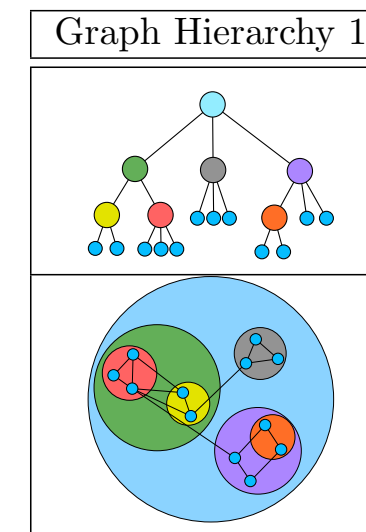
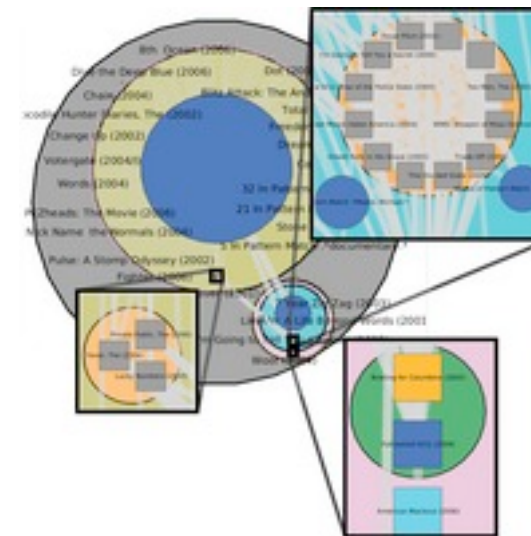
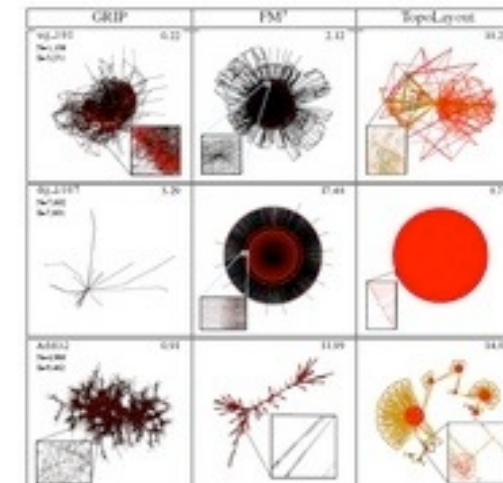
# Techniques: Dimensionality Reduction

- Glint: costly distance functions
- DimStiller: visual dimensional analysis and reduction toolkit
- Glimmer: GPU accelerated MDS



# Techniques: Graphs/Trees

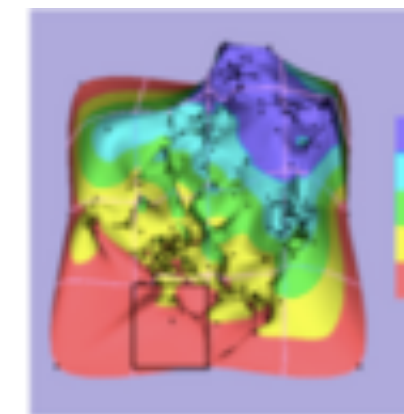
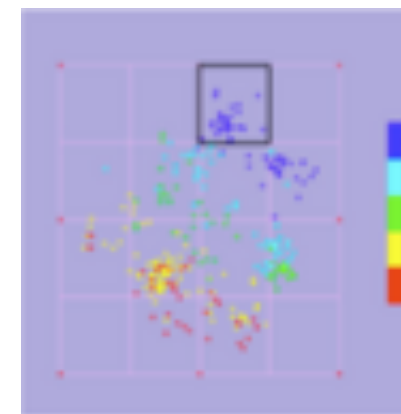
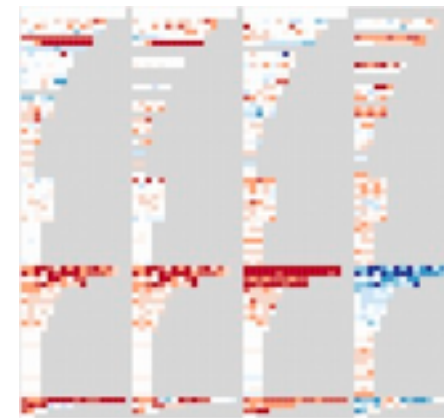
- general multilevel/compound graphs
  - layout
    - TopoLayout
  - interaction
    - Grouse
    - GrouseFlocks
    - TugGraph
- evolutionary tree comparison
  - TreeJuxtaposer





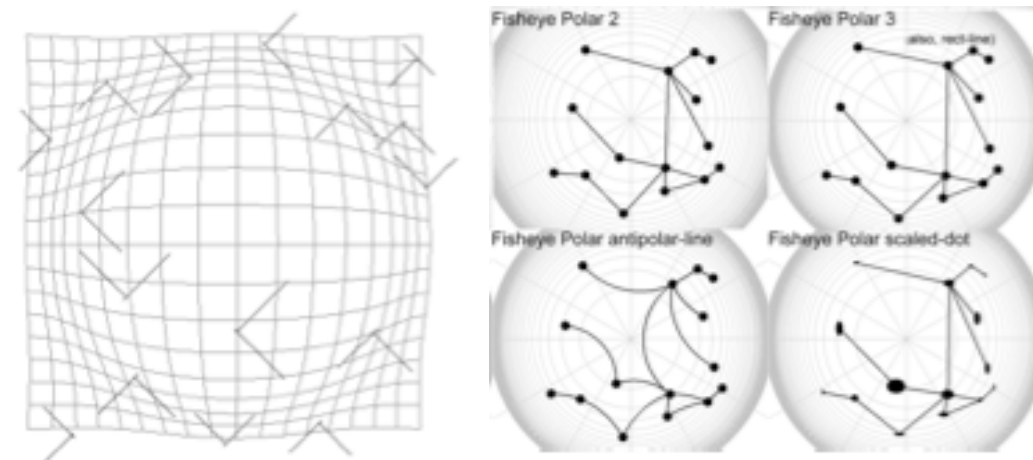
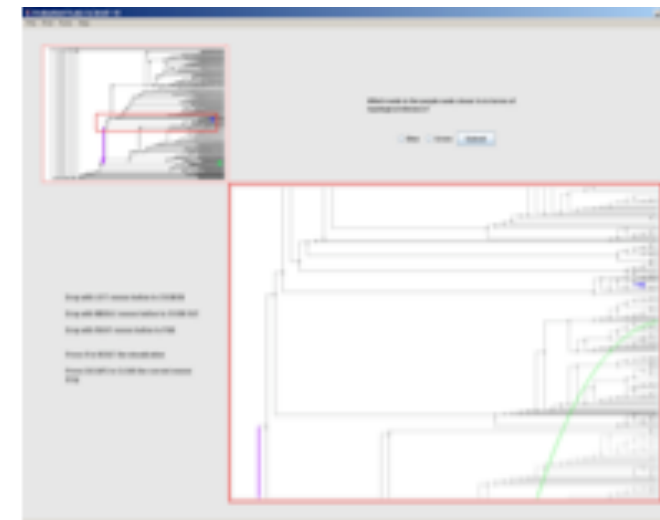
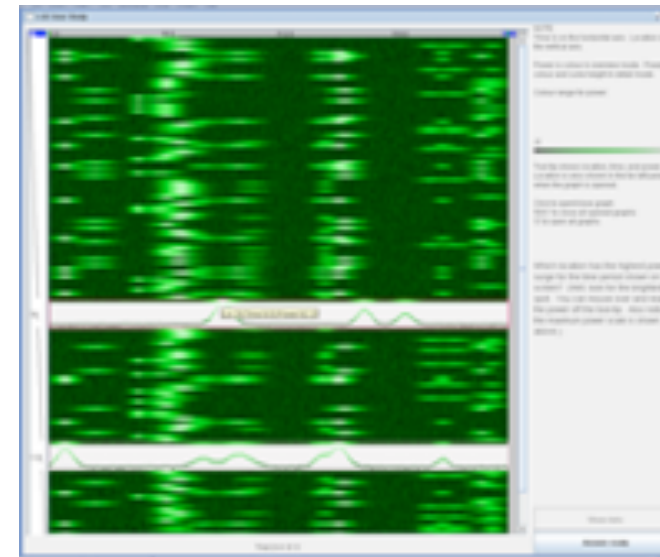
# Evaluation: Dimensionality Reduction

- guidance on scatterplot/DR choices
- taxonomy of cluster separation factors
- 2D points vs 3D landscapes



# Evaluation: Focus+Context

- overviews: separate vs. integrated views
- navigation: stretch and squish vs. pan/zoom navigation
- impact of distortion on visual search, visual memory





# Further Information

- further info

- <http://www.cs.ubc.ca/~tmm/talks.html#oxford14> (this talk, and many others)
- <http://www.cs.ubc.ca/group/infovis> (papers, software, videos)
- <http://www.cs.ubc.ca/~tmm/courses/infovis/book> (book: to appear)
  - Visualization Analysis and Design. Munzner. AK Peters, 2014

- paper & open source software download

- <http://www.cs.ubc.ca/labs/imager/tr/2013/VariantView/>

- 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

