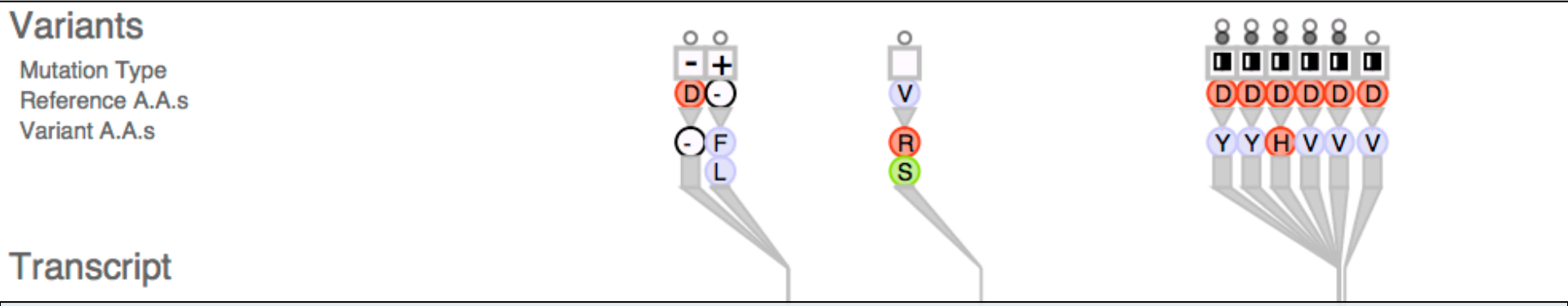


Variant View

Visualizing Sequence Variants in their Gene Context

Joel A. Ferstay[1]*, Cydney B. Nielsen[2], Tamara Munzner[1]

University of British Columbia[1], now at AeroInfo Systems/Boeing Canada*

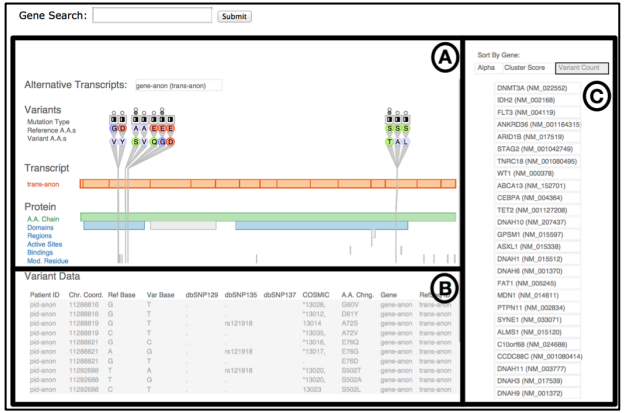
BC Cancer Agency[2]

Design study

- Real users, real tasks, real data
- Find out what users are doing
- Support with visualization tool
- Reflect and present guidelines

Analyst

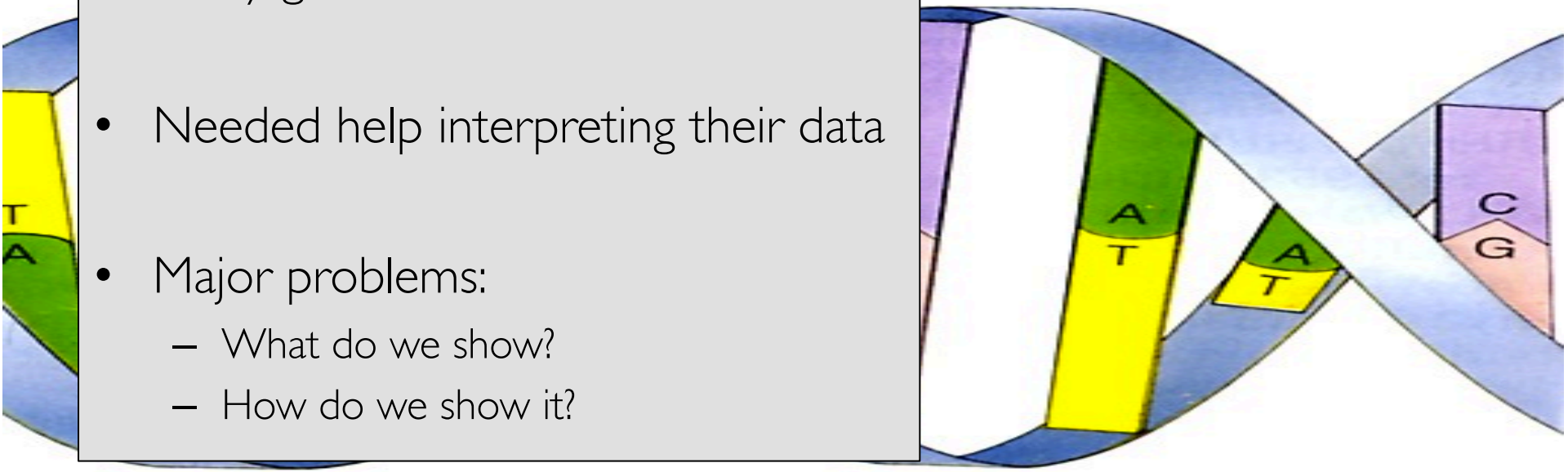
Tool



The Design Process

Collaborated with analysts at the Genome Sciences Centre

- Study genetic basis of leukemia
- Needed help interpreting their data
- Major problems:
  - What do we show?
  - How do we show it?



Design cycle

Interview

Design cycle

Interview

Data and Tasks

Design cycle

Interview

Data and Tasks

Create Data Sketch

Design cycle

Interview

Present Data Sketch

Data and Tasks

Create Data Sketch

Design cycle

Interview

Present Data Sketch

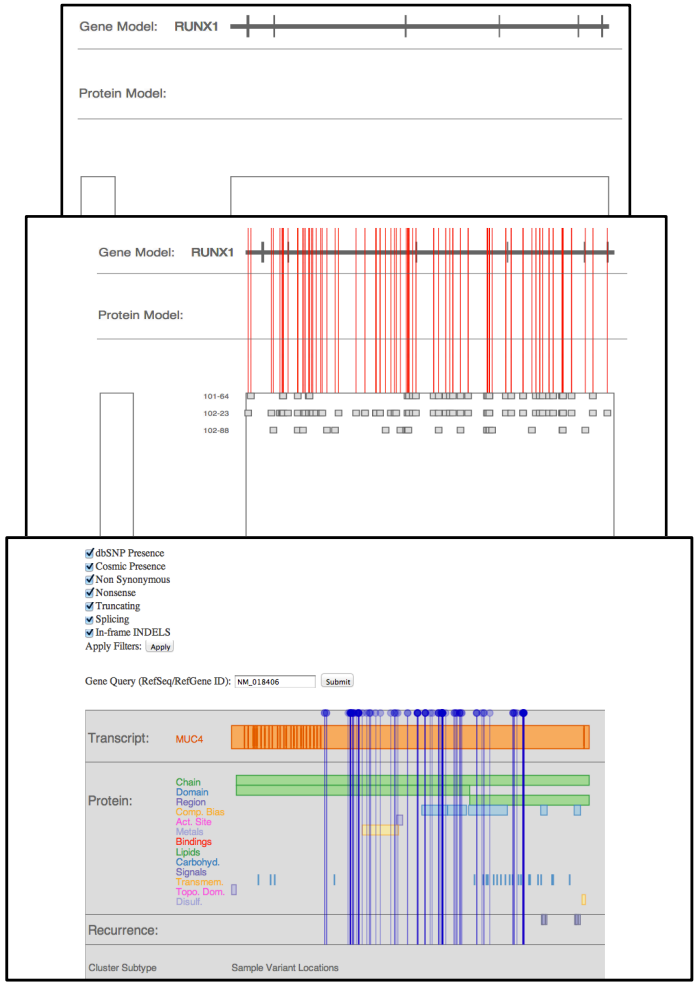
Data and Tasks

REPEAT

Create Data Sketch

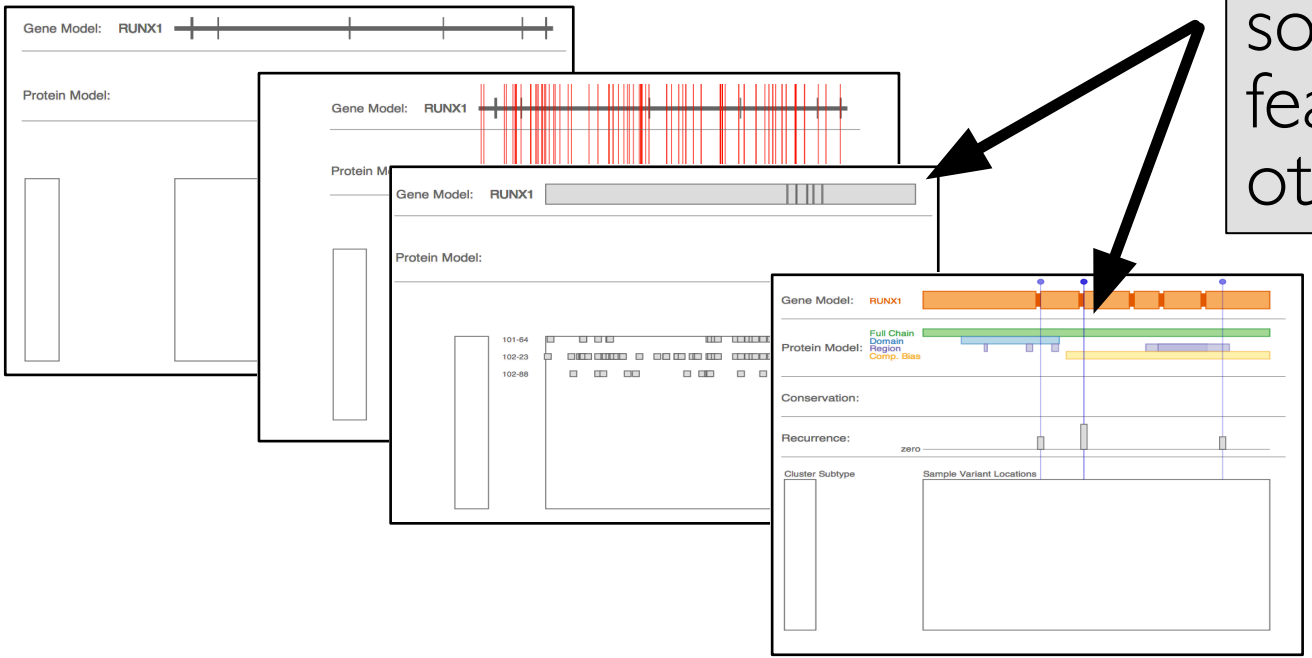
Data sketches

- Alternative to paper prototyping
- Load and show real data
- Beneficial when the data is complex



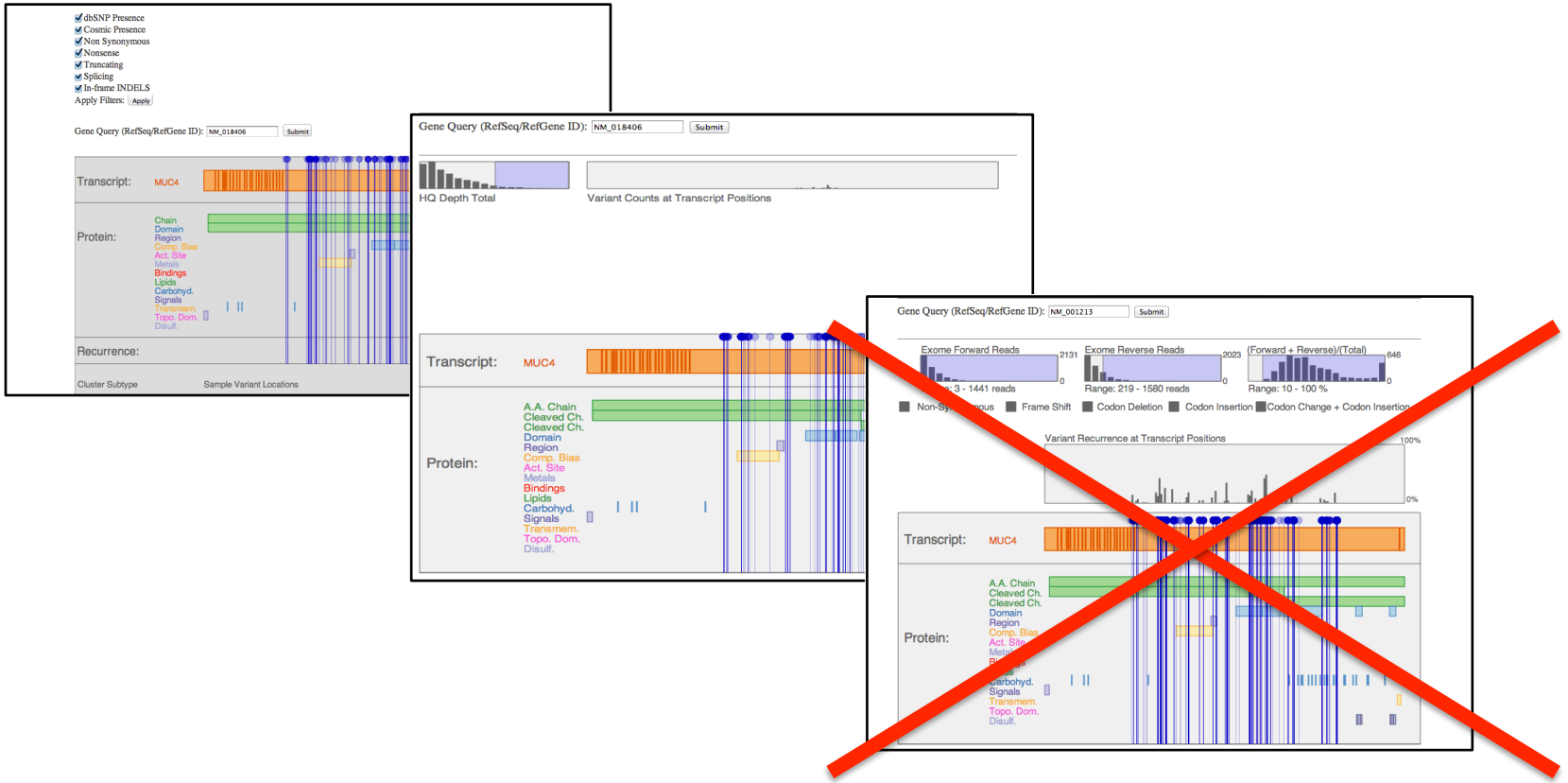
[Lloyd and Dykes, InfoVis 2011]

Can identify features of **interest** in data



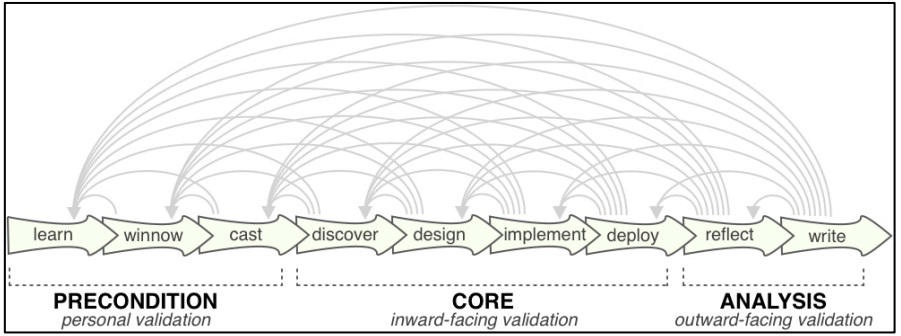
Emphasize some features over others

Can identify **design dead ends** early



Methods and durations

- Semi-structured interviews
  - 7 months
  - Once per week
  - One hour in duration
- Presented data sketches
  - 8 deployed over 5 months
  - Implemented with D3 toolkit
    [Bostock et al., InfoVis 2011]

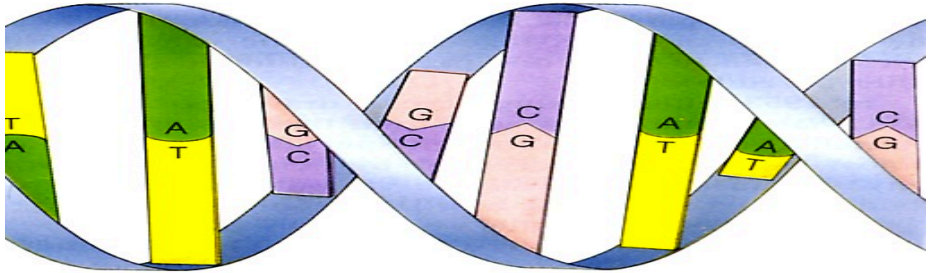


Design study methodology [Sedlmair et al., 2012]

Problem characterization: Data

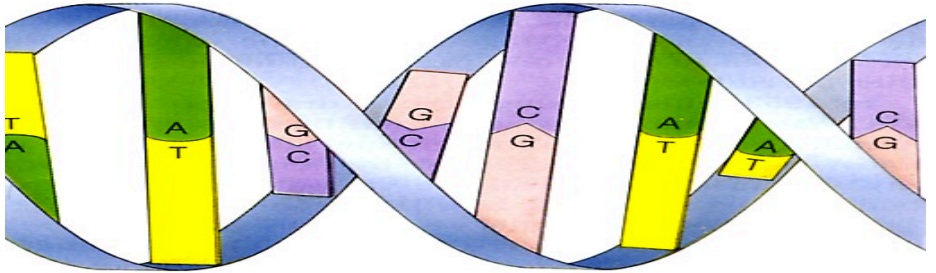
The data

- Data are **sequence variants**
  - Difference between reference genome and a given genome



The data

- Data are **sequence variants**
  - Difference between reference genome and a given genome



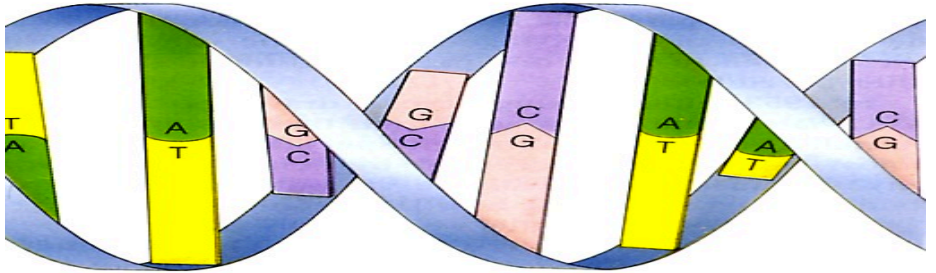
Reference Genome DNA: ATA TGA TCA ACA CTT

Sample 1 Genome DNA: ATA TGG TCA ATA CTT

Sample 2 Genome DNA: ATA TGA TGA ACA CCT

The data

- Data are **sequence variants**
  - Difference between reference genome and a given genome



Reference Genome DNA: ATA TGA TCA ACA CTT

Sample 1 Genome DNA: ATA TGG TCA ATA CTT

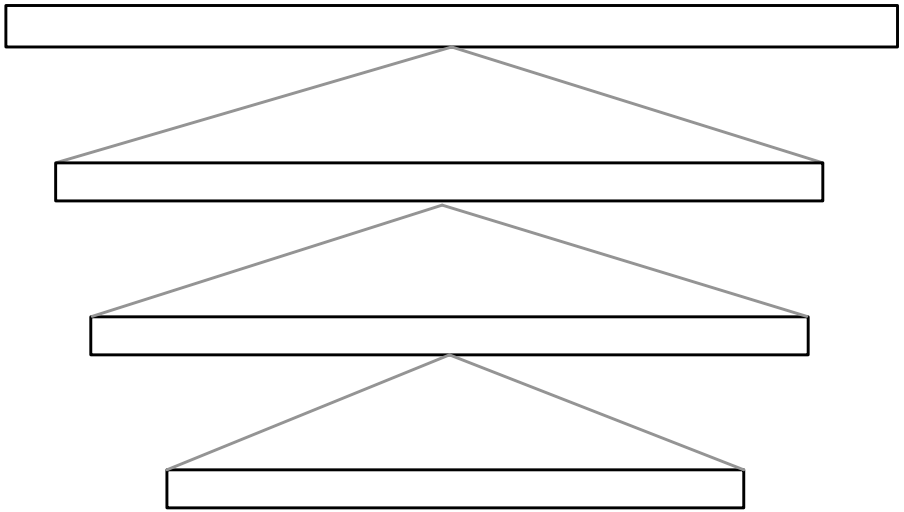
Sample 2 Genome DNA: ATA TGA TGA ACA CCT

Harmful?

Harmless?

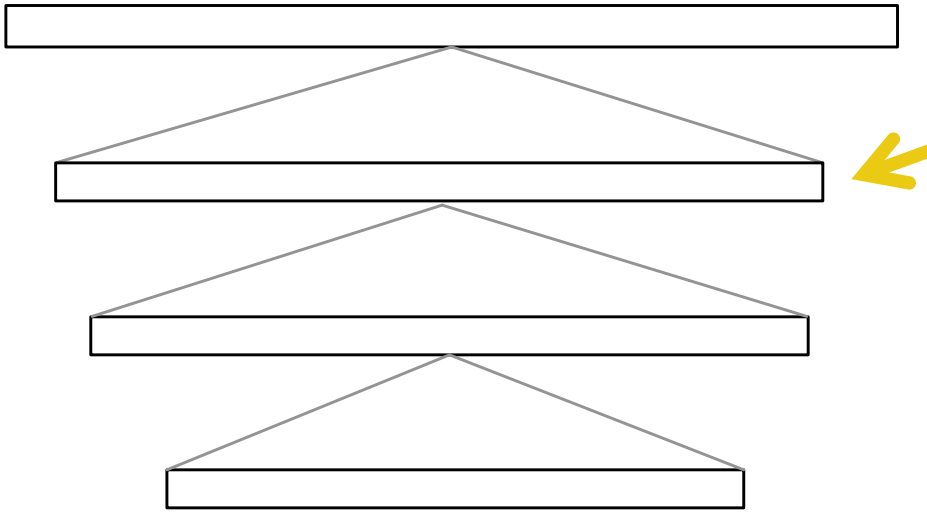
Multi-scale data

DNA of the organism



Genome	~3 billion nucleotides
Gene	~10,000 nucleotides
Exon	~100 nucleotides
Region	~50 amino acids

Multi-scale data



Genome

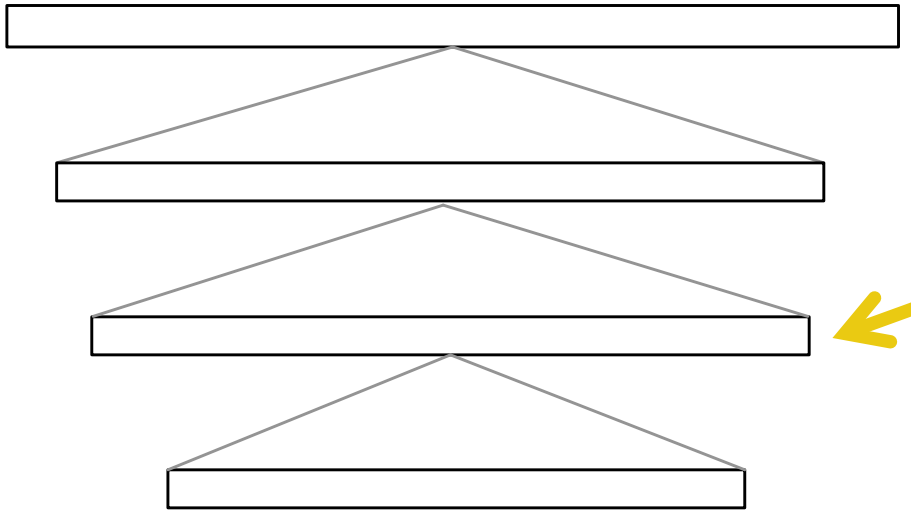
Genes are functional units

Gene ~10,000 nucleotides

Exon ~100 nucleotides

Region ~50 amino acids

Multi-scale data



Genome ~3 billion nucleotides

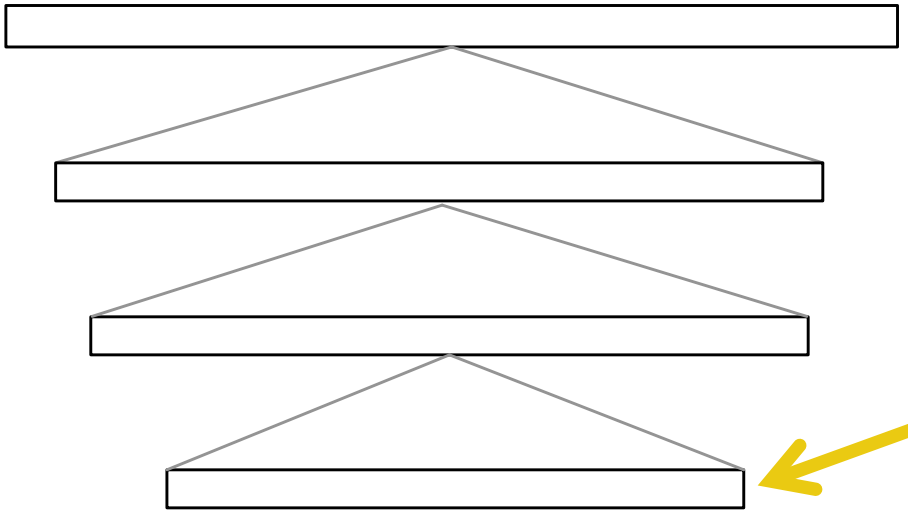
Gene 10,000 nucleotides

Exons code for proteins

Exon ~100 nucleotides

Region ~50 amino acids

Multi-scale data



Genome ~3 billion nucleotides

Gene ~10,000 nucleotides

Exo

Regions within proteins perform activities

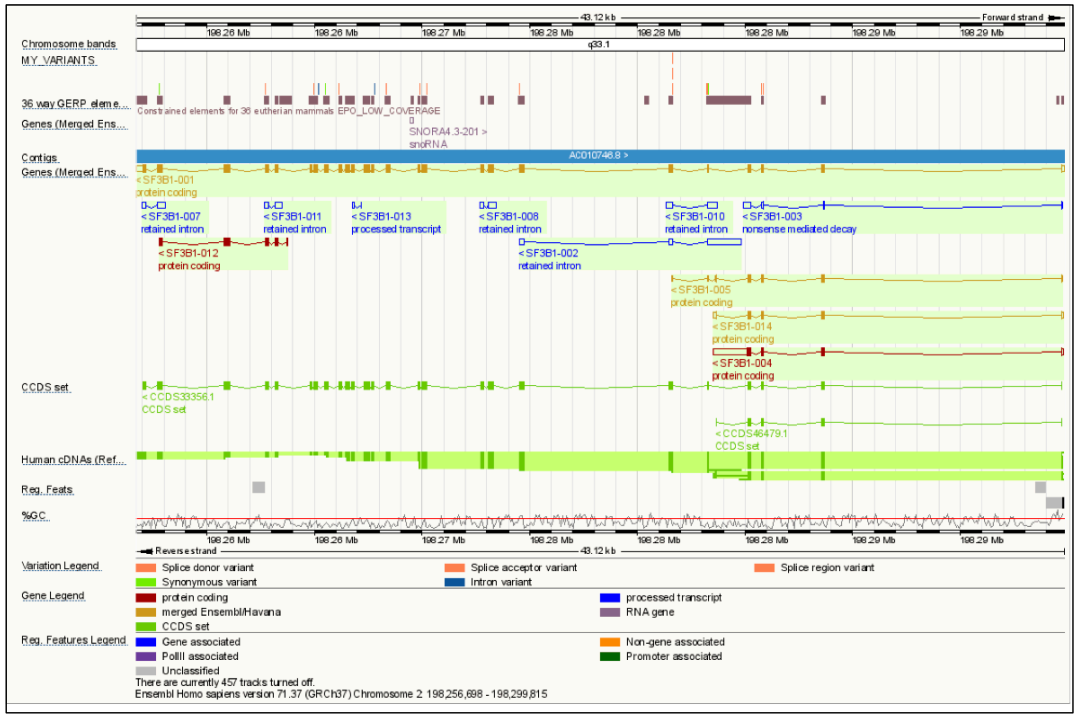
Region ~50 amino acids

Related work: Genome scale

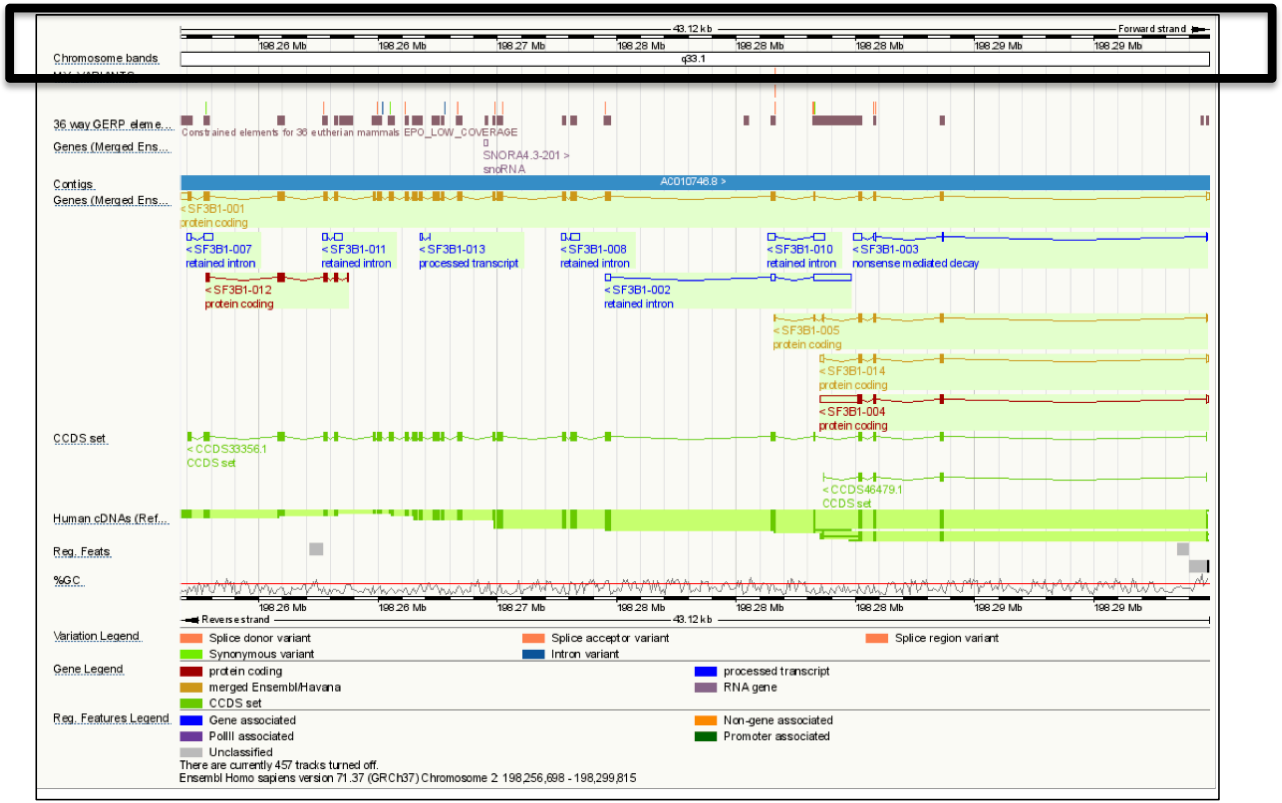
Ensembl genome browser

- Explore genome and variant data

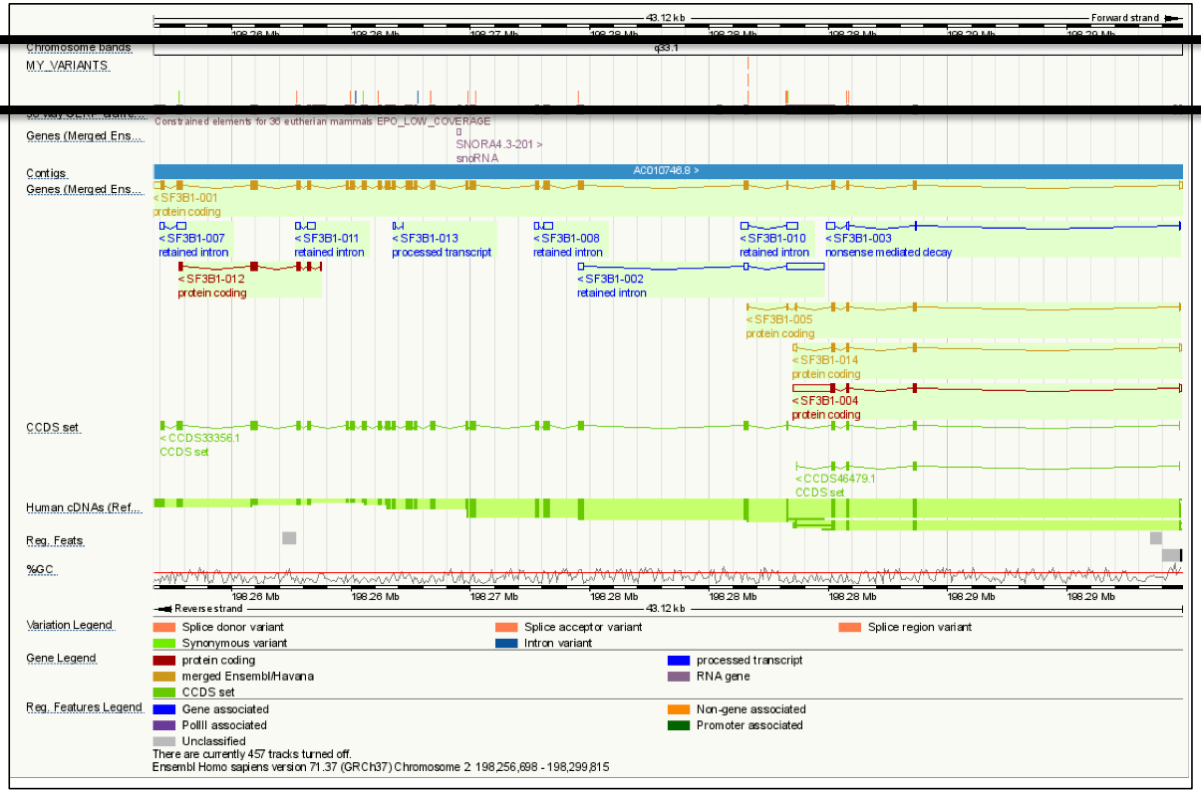
[Chen et al., BMC Genomics 2010]



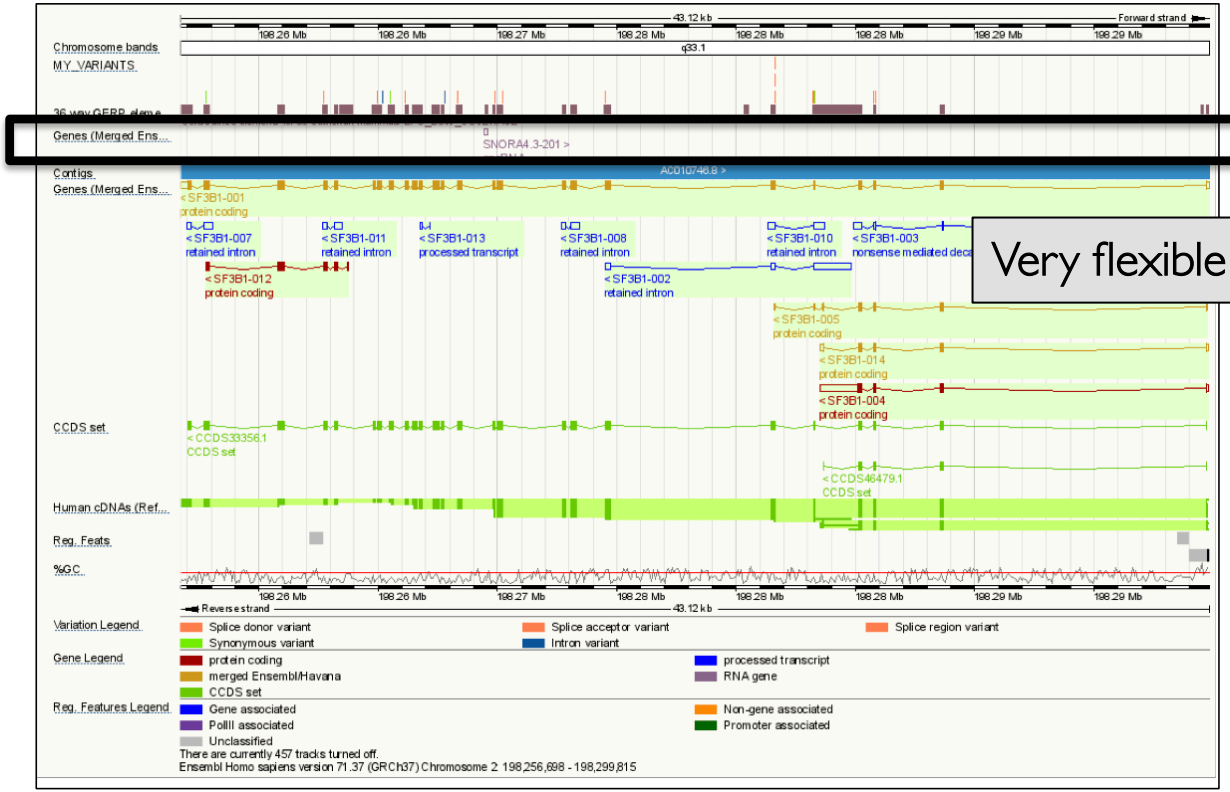
Genome scale shown at the top



User data is stacked in horizontal tracks



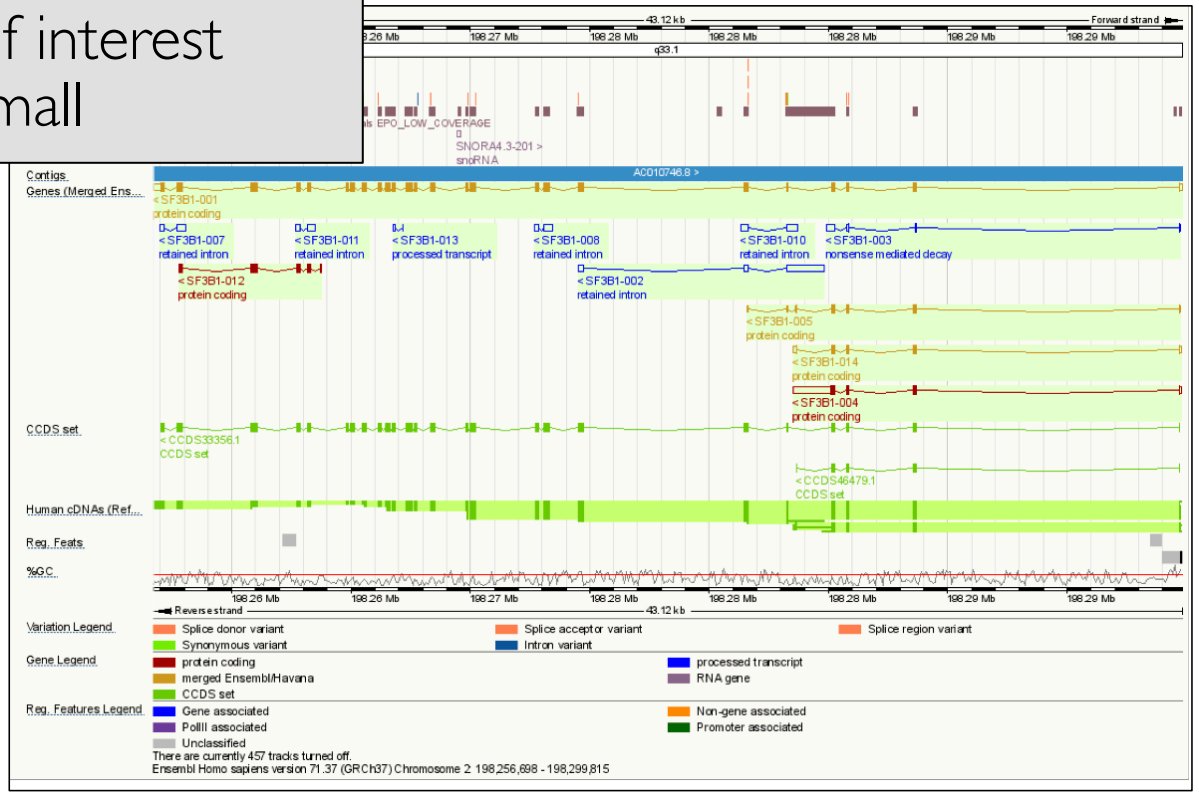
User data is stacked in horizontal tracks



Very flexible framework

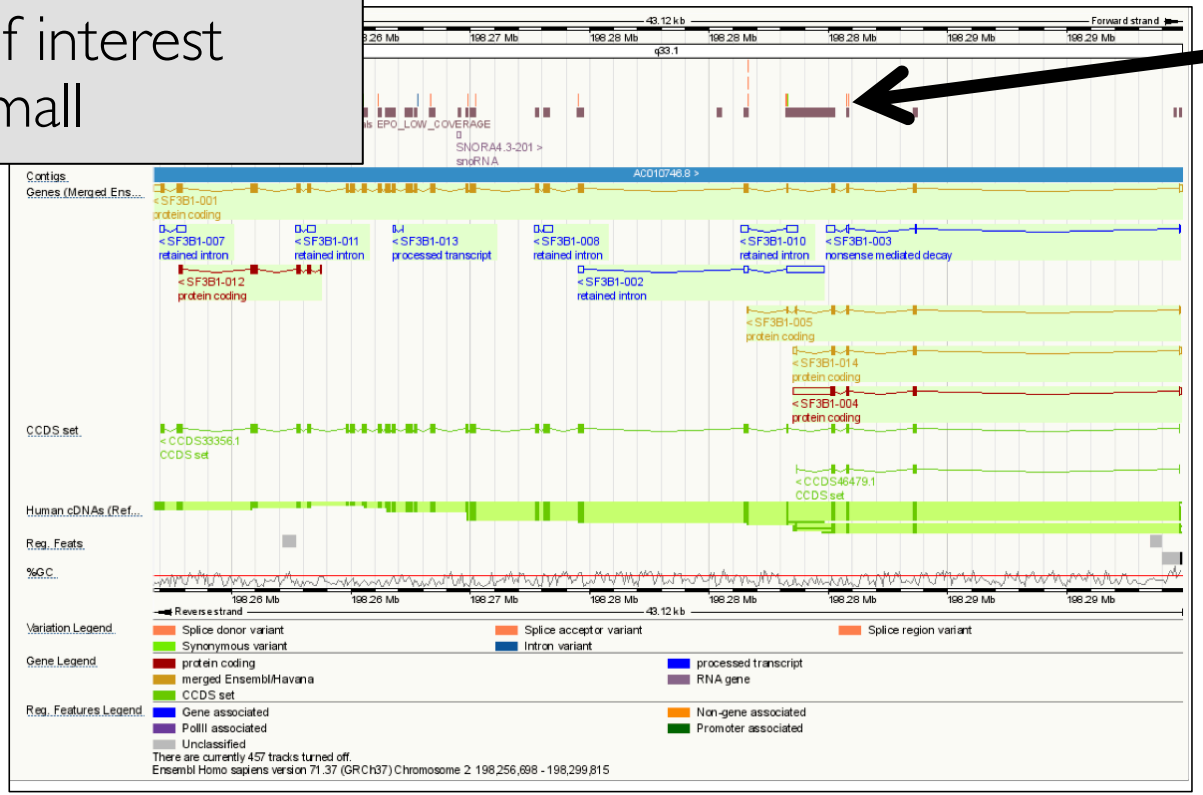
Problem with the genome scale

- Features of interest become small



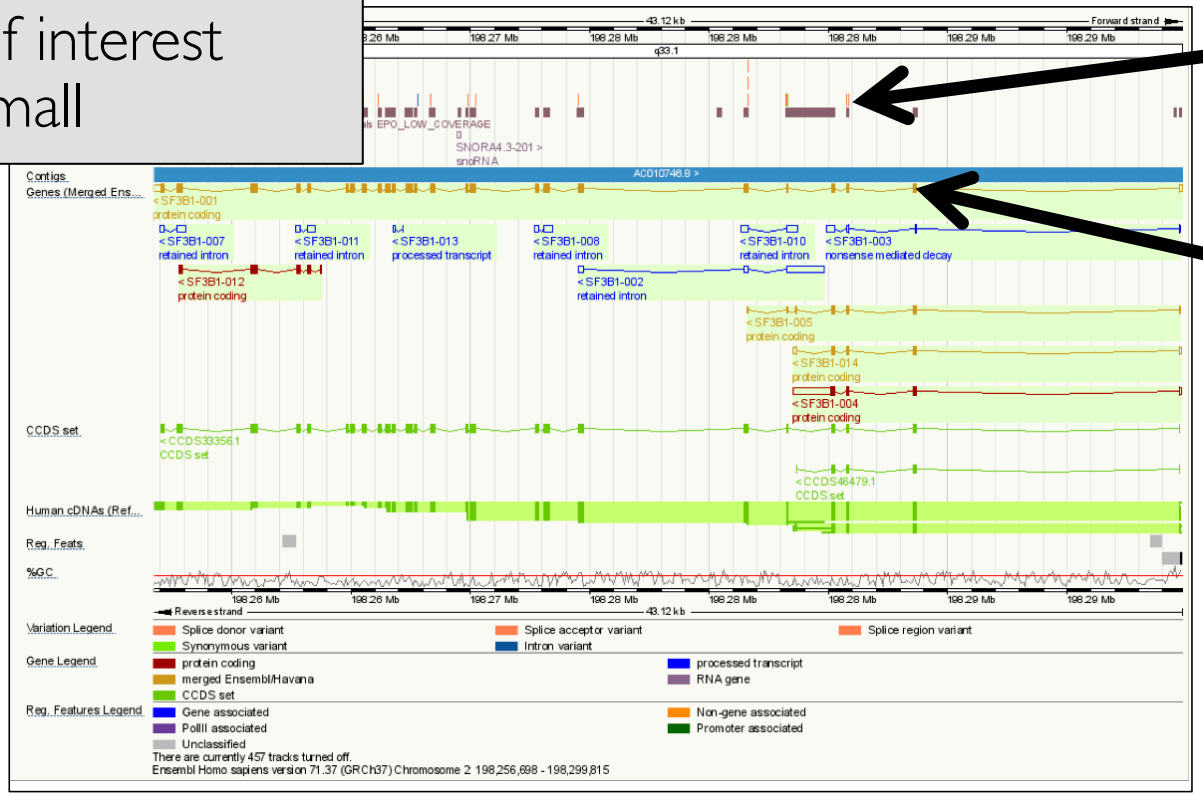
Problem with the genome scale

- Features of interest become small

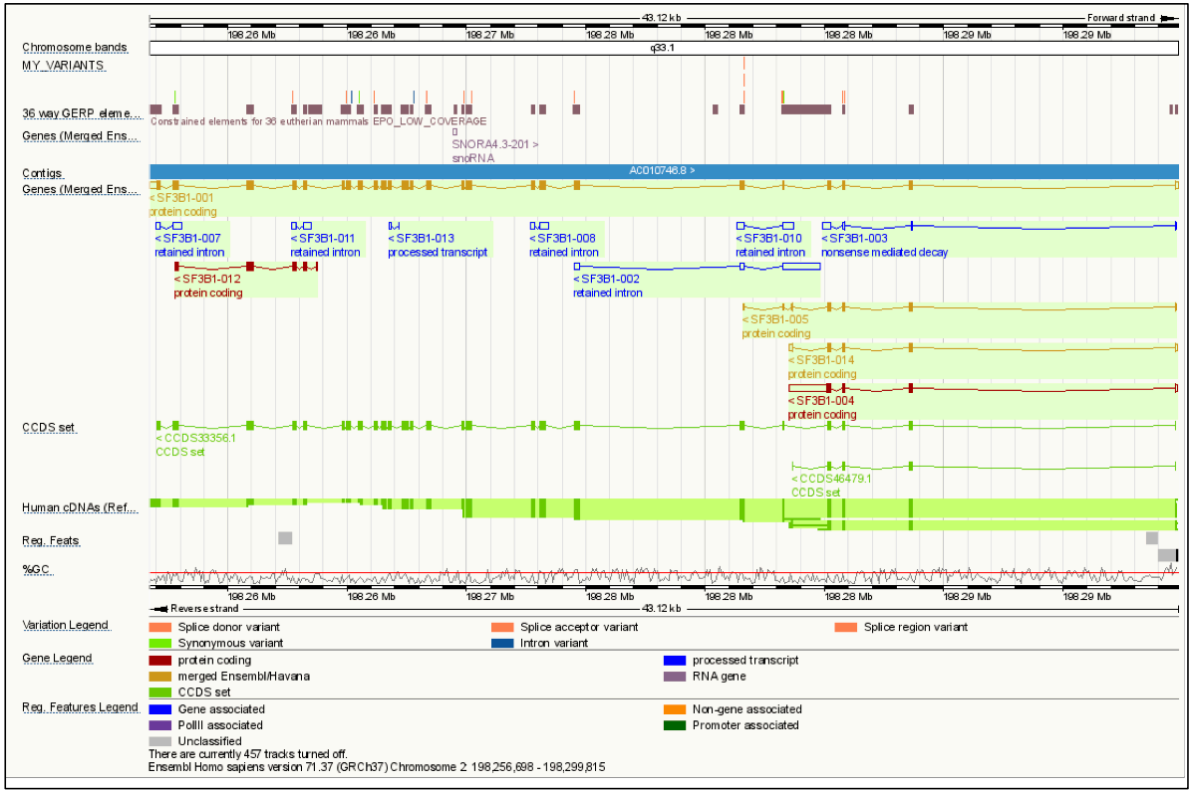


Problem with the genome scale

- Features of interest become small

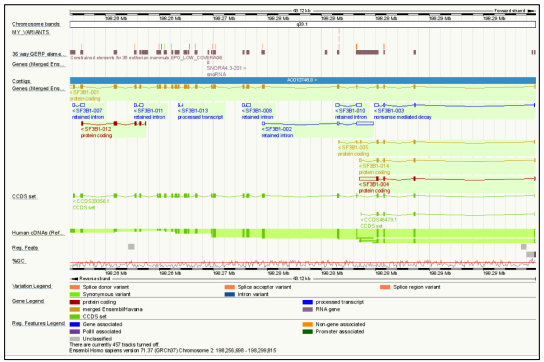


Analysts must pan and zoom



Analysts must pan and zoom

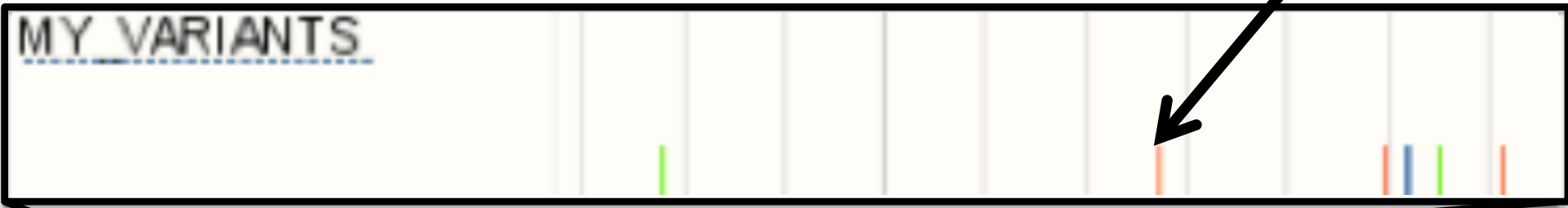
- Heavy interaction costs with zooming

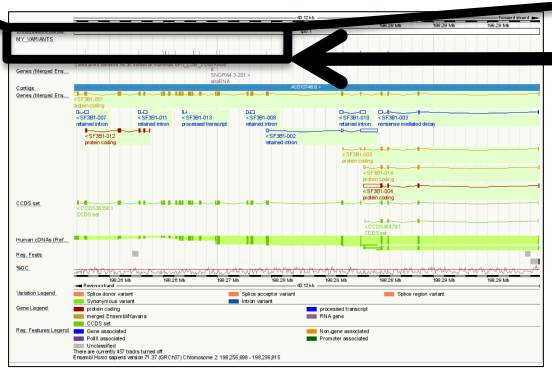


Analysts must pan and zoom

- Heavy interaction costs with zooming

Variant





Must know where to look

Raw variant data
(What they looked at before)

Raw data **variant-centric**

- Tabular data format

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

Raw data **variant-centric**

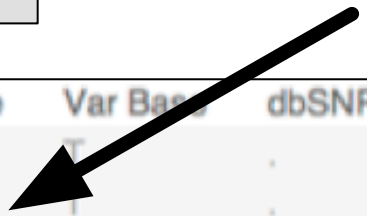
- Tabular data format

Variant row 1

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

Raw data **variant-centric**

- Tabular data format



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

Raw data **variant-centric**

- Tabular data format

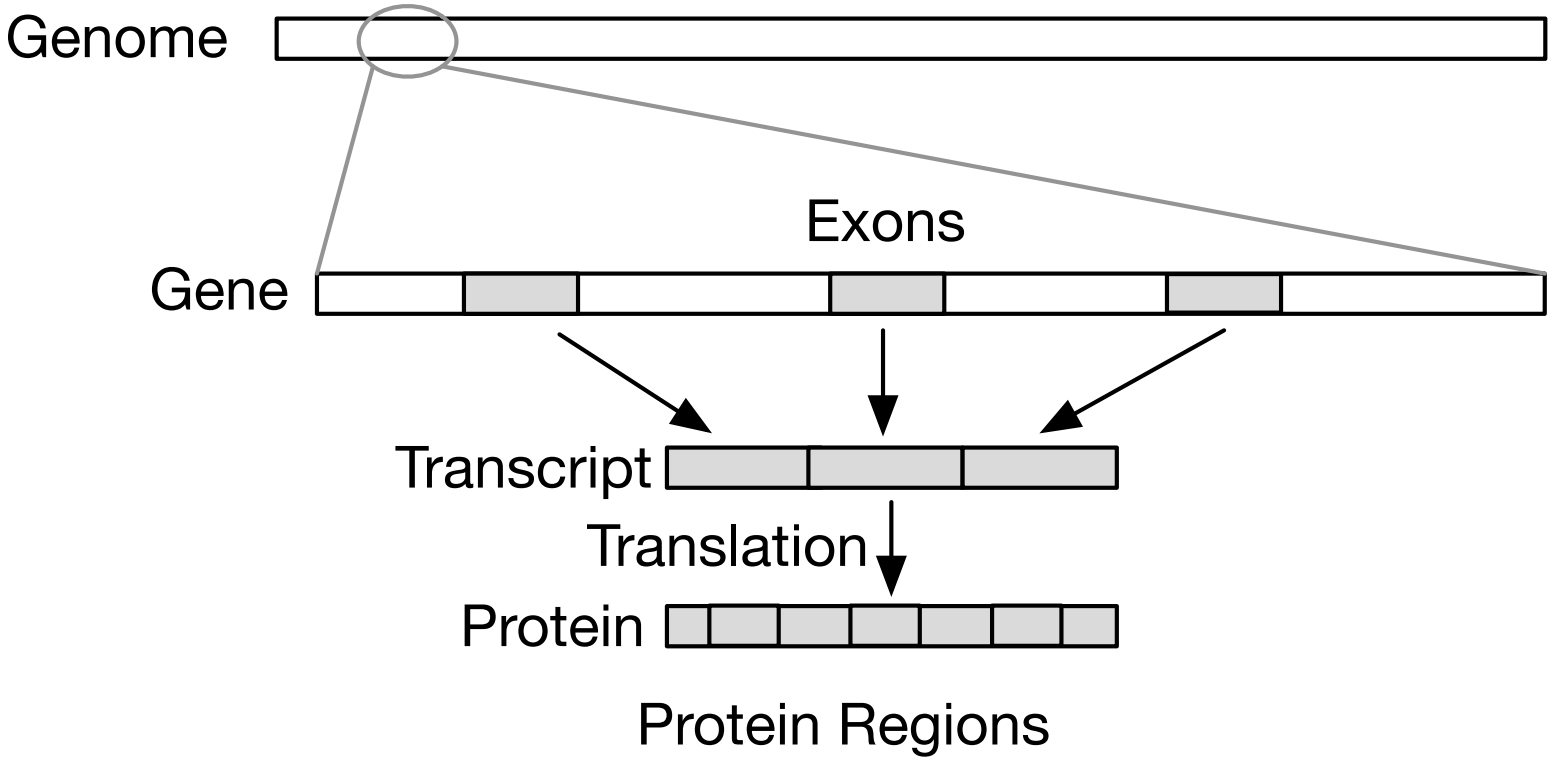
Variant row 2

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

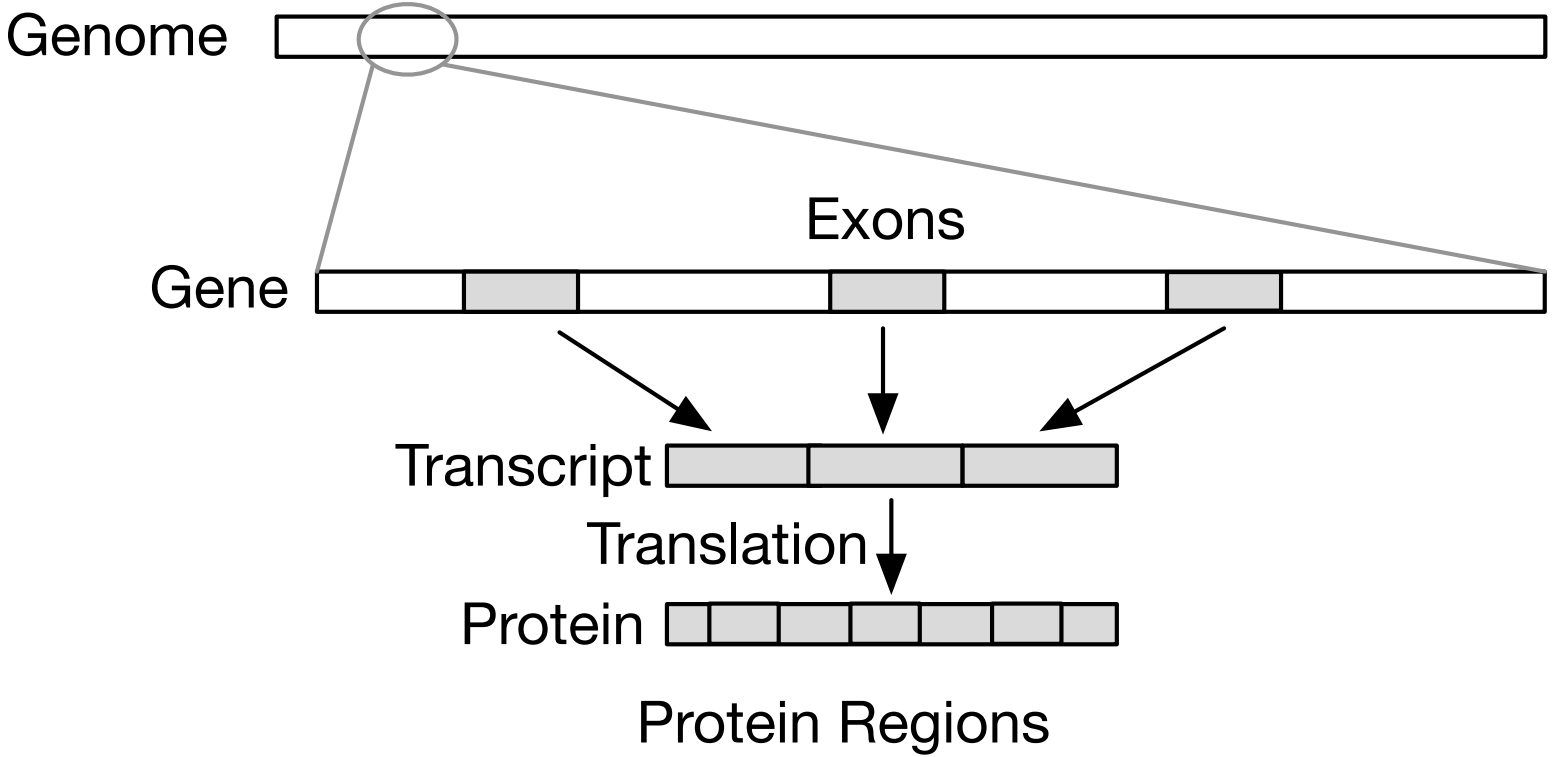
- Difficult to reason about variants without their biological context

Multiple biological levels/scales
What to show?

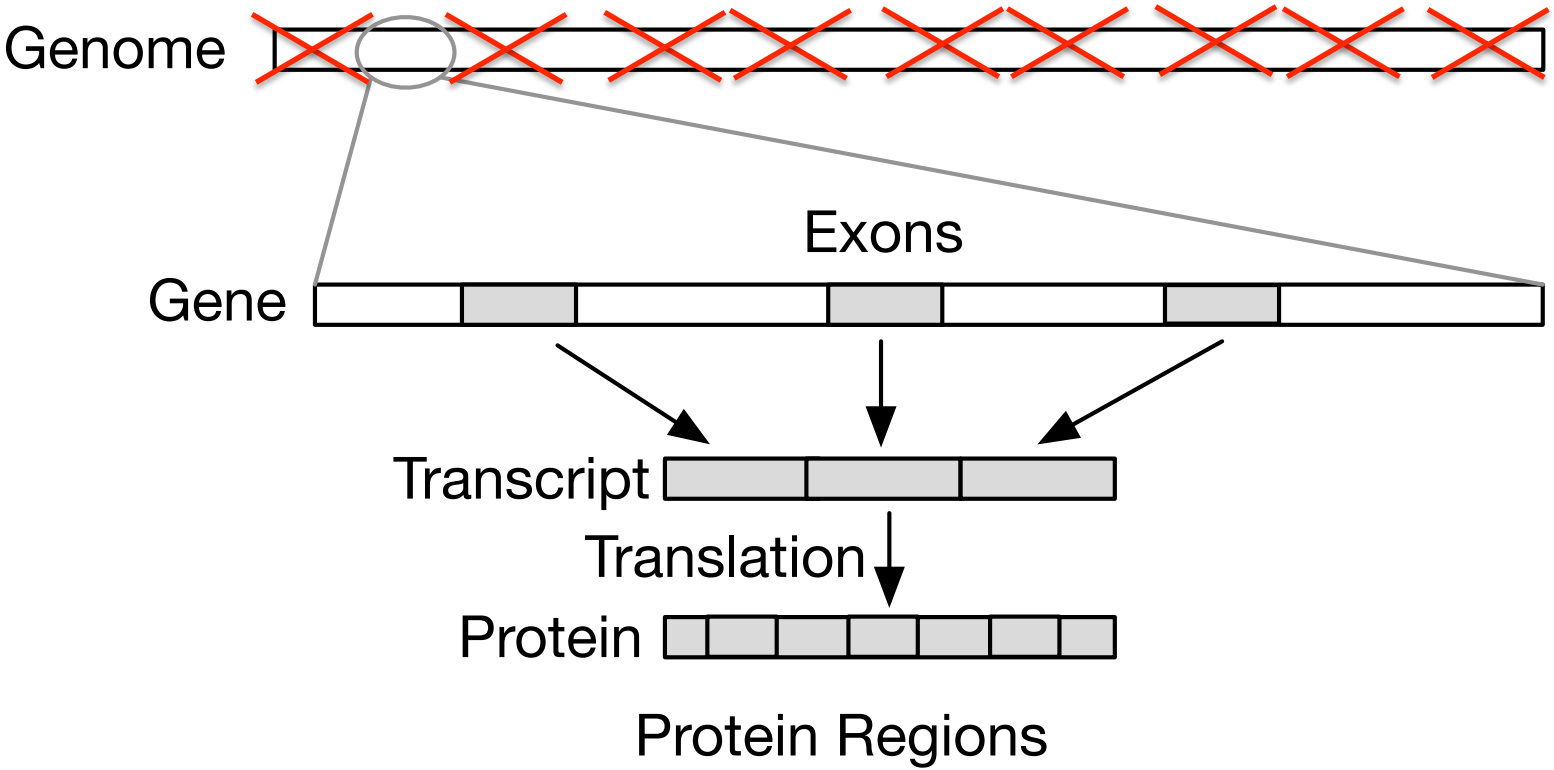
Many biological levels and scales



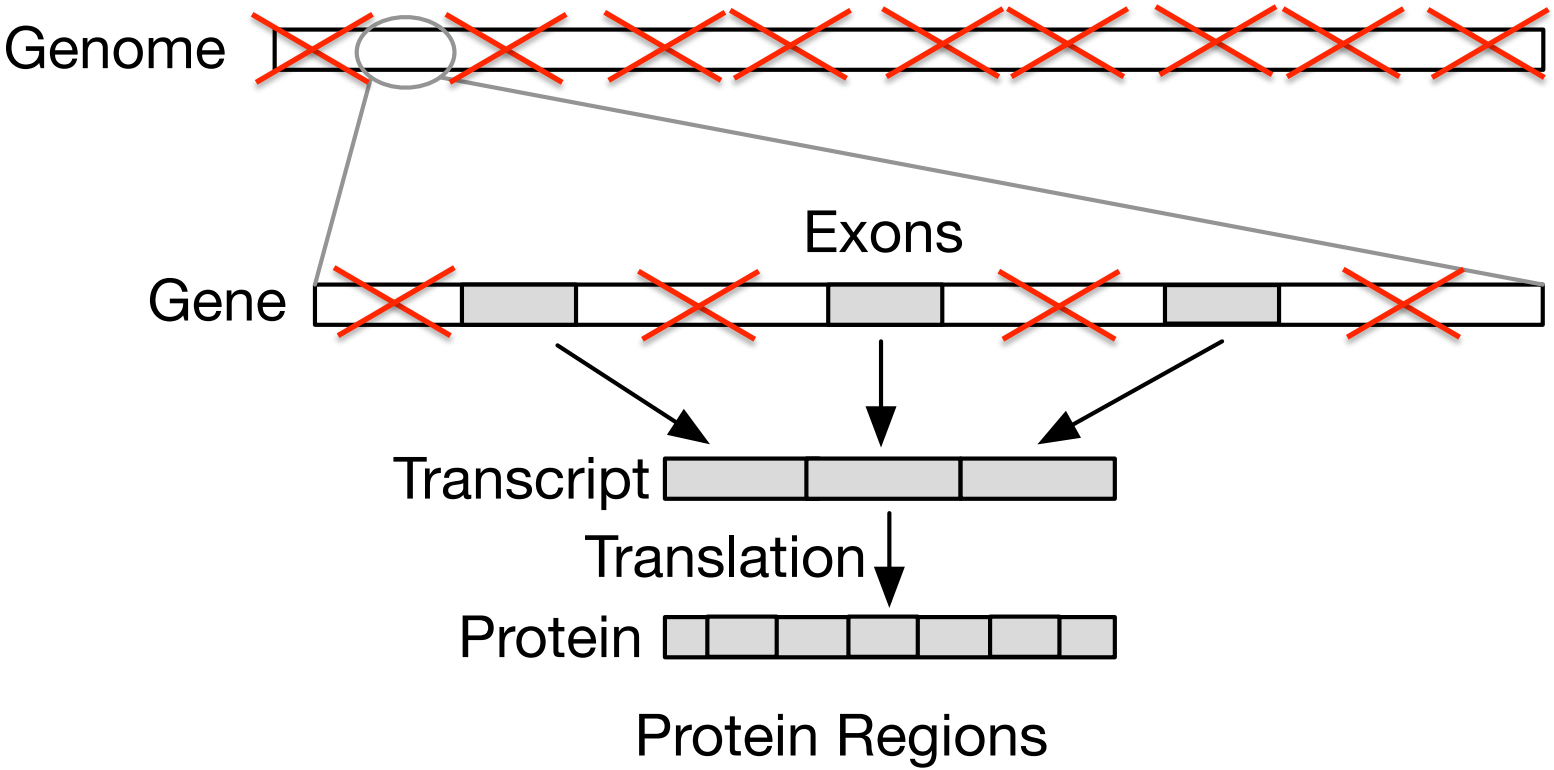
Only some levels and scales are beneficial for variant analysis



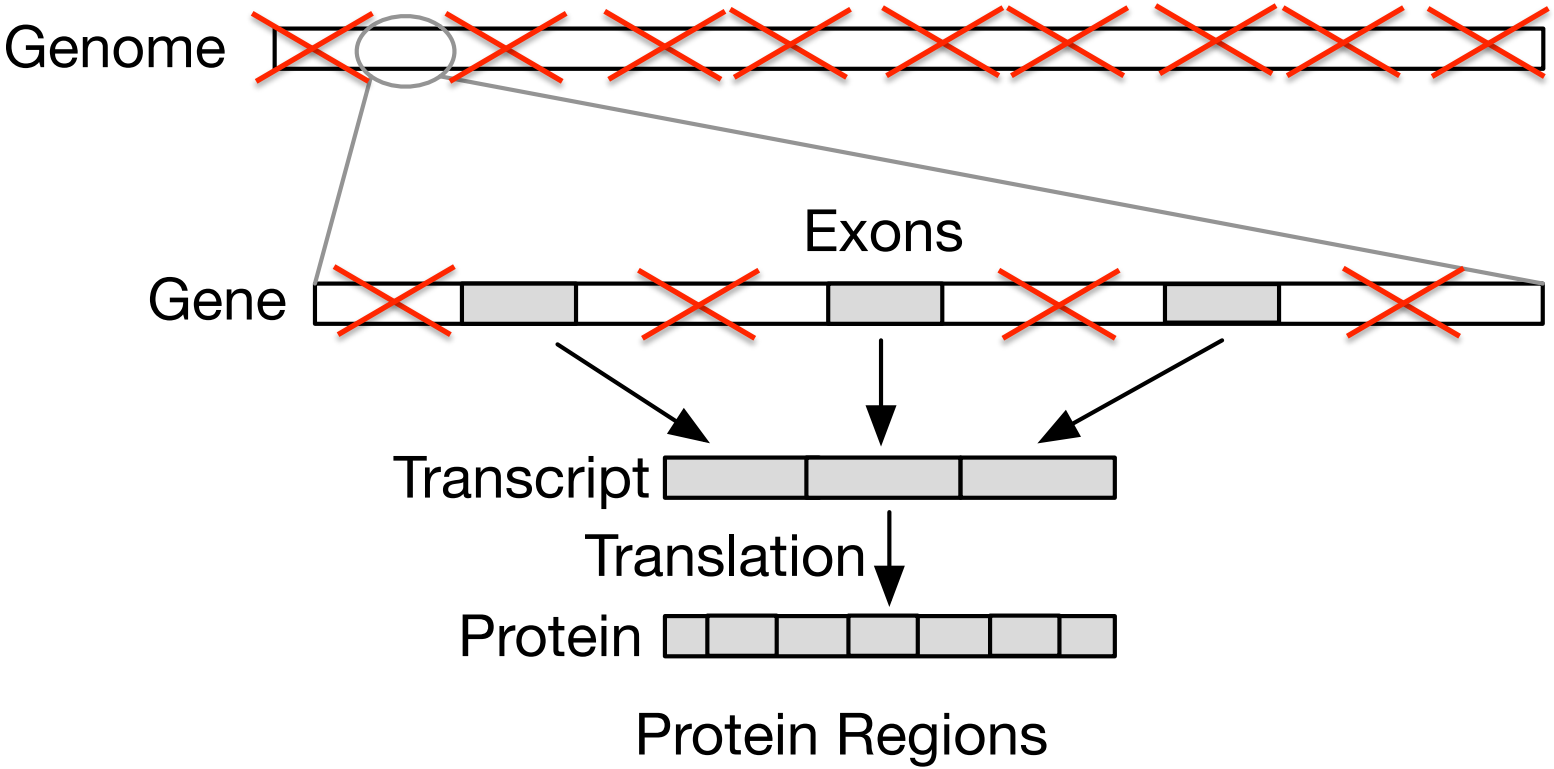
Filter out whole genome; keep genes



Filter out non-exon regions

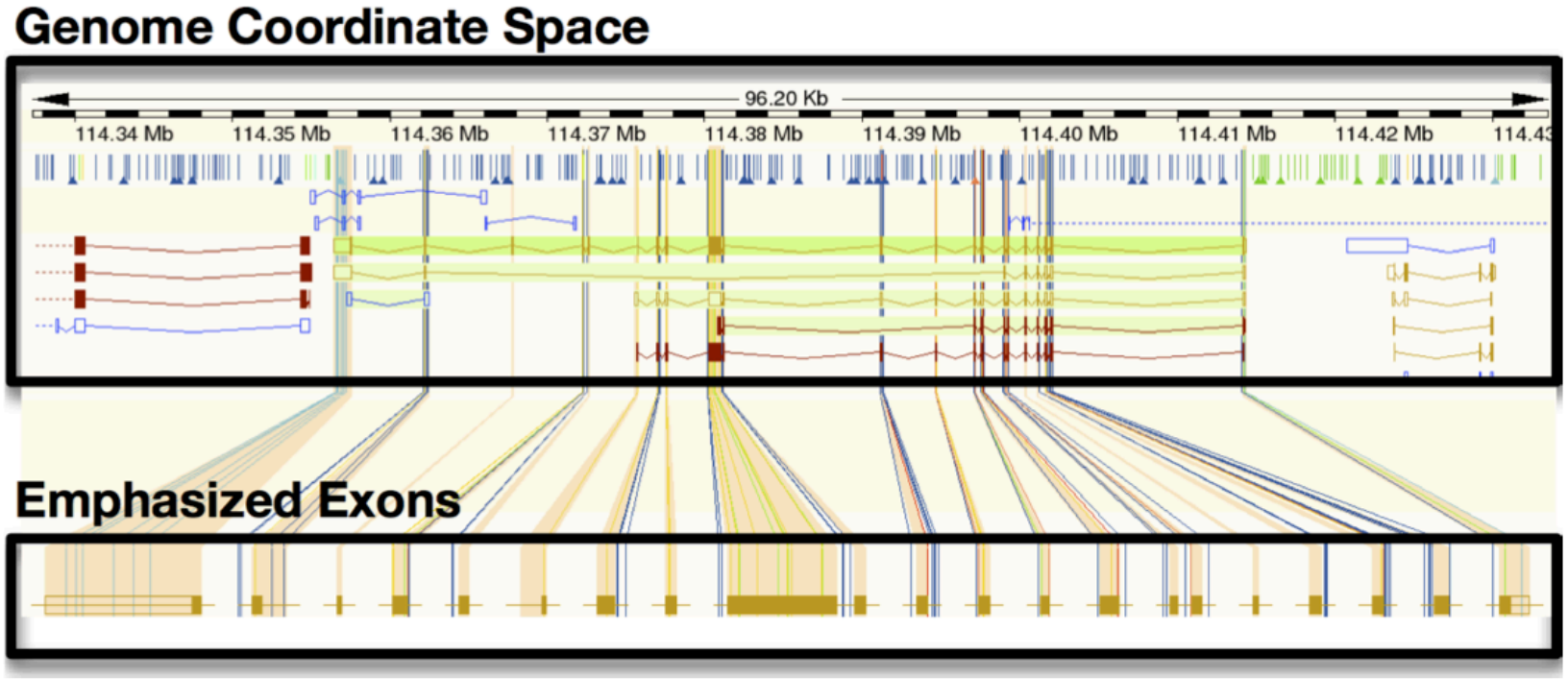


Left with a **filtered scope**



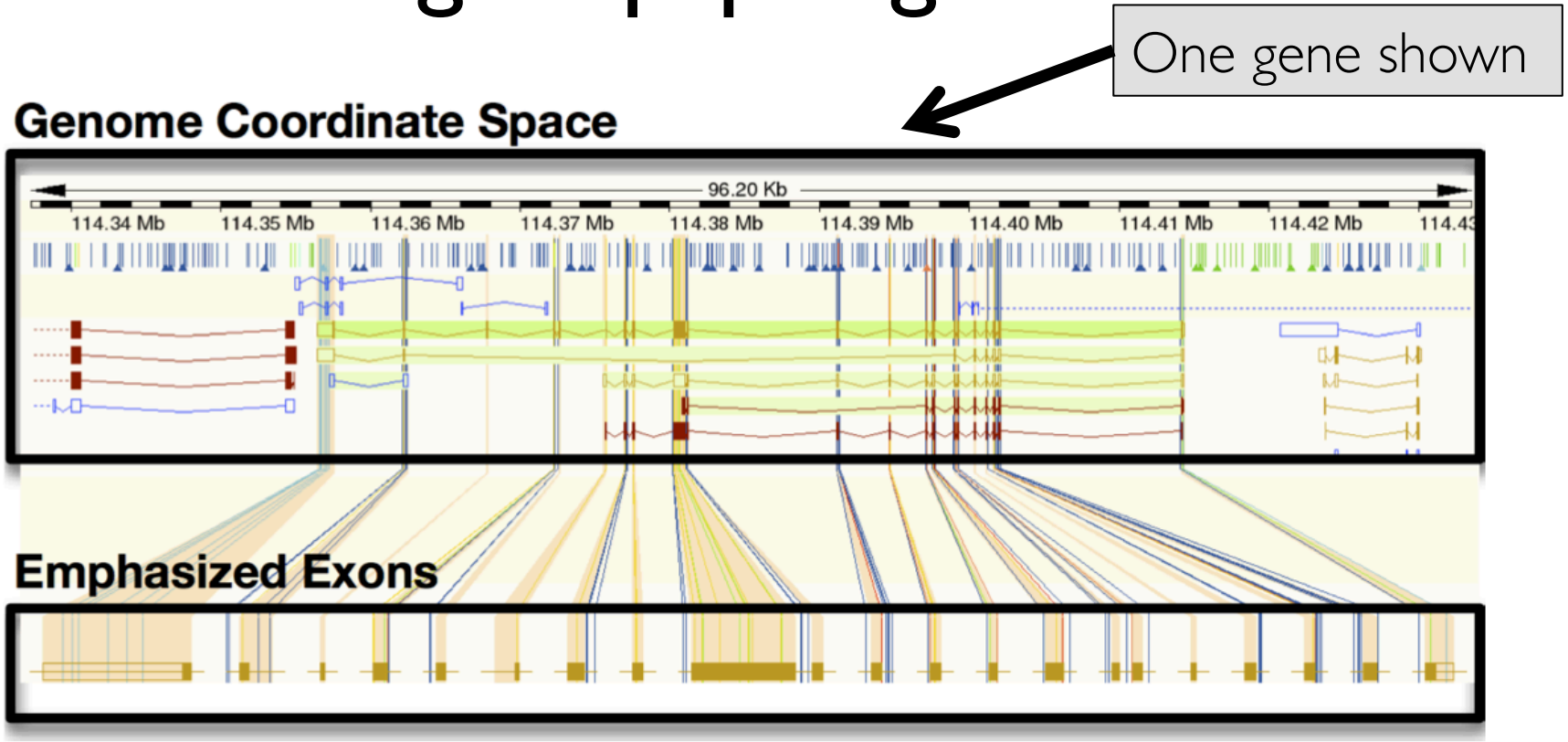
Related work:
Filtered scope

The Ensembl Variation Image



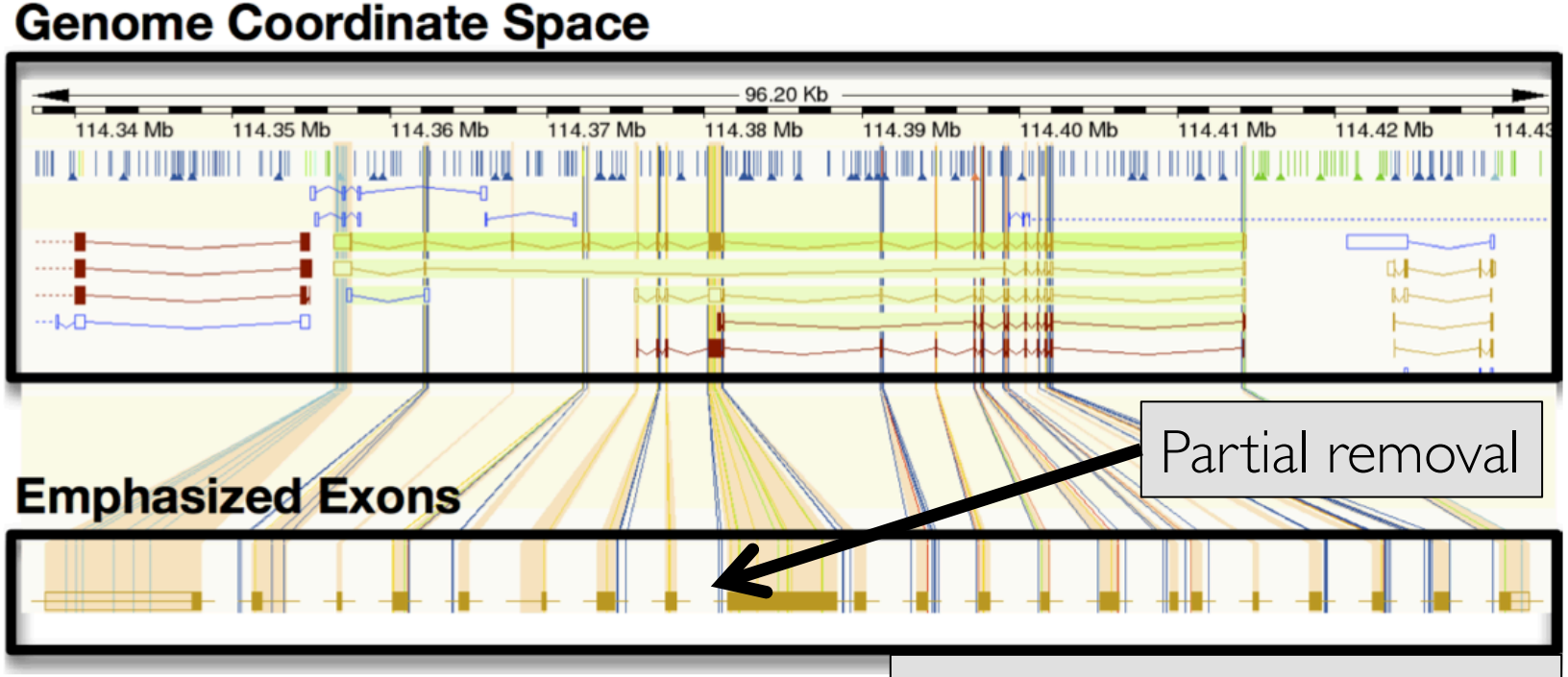
[Chen et al., BMC Genomics 2010]

First filtering step: per-gene view



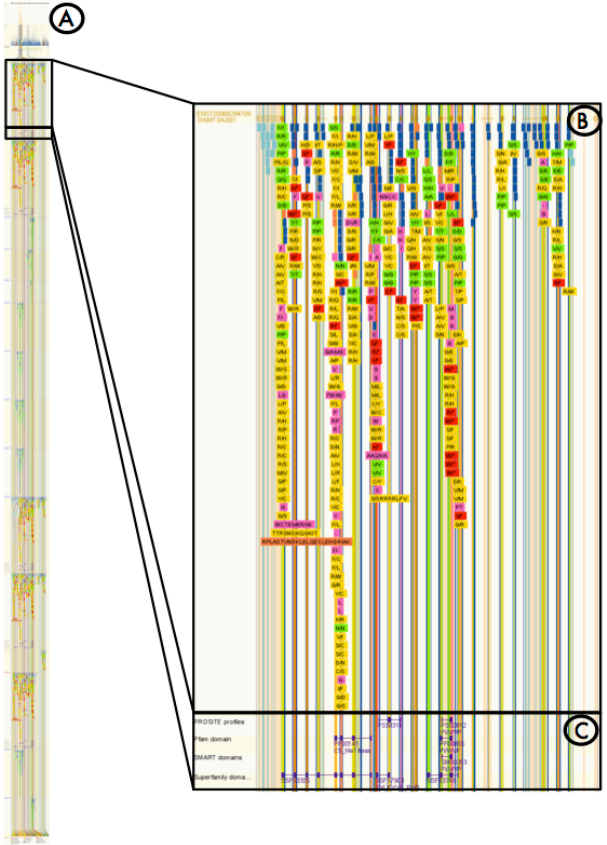
[Chen et al., BMC Genomics 2010]

Second filtering step: partial removal inter-exon regions



[Chen et al., BMC Genomics 2010]

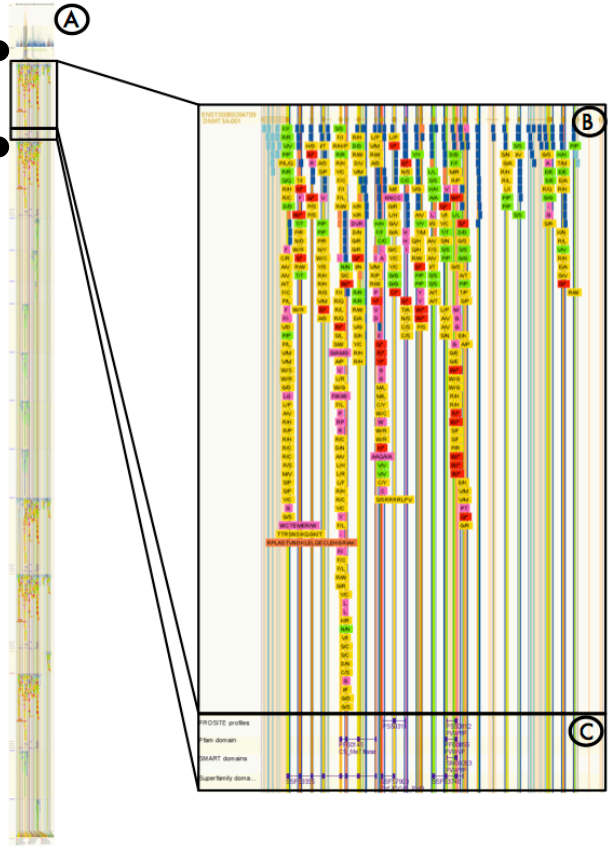
Problem: Extends multiple screens



[Chen et al., BMC Genomics 2010]

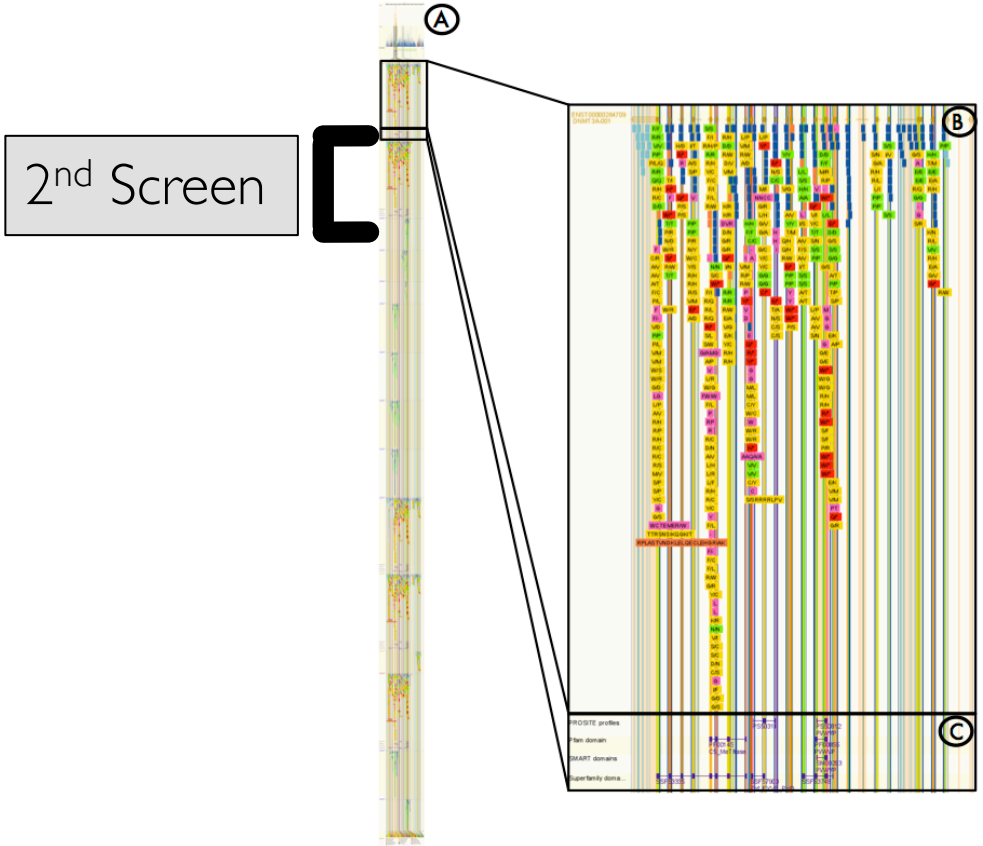
Problem: Extends multiple screens

1st Screen



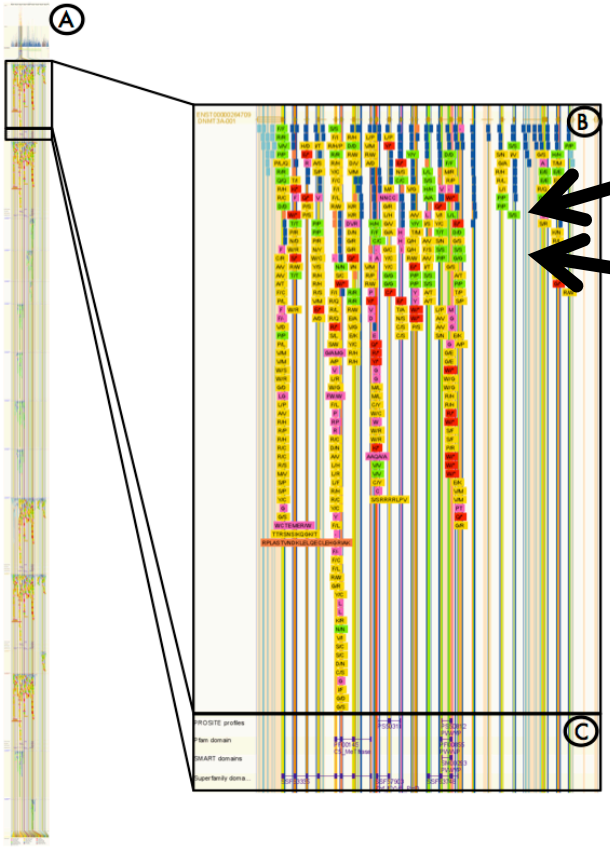
[Chen et al., BMC Genomics 2010]

Problem: Extends multiple screens



[Chen et al., BMC Genomics 2010]

Problem: Features of interest small



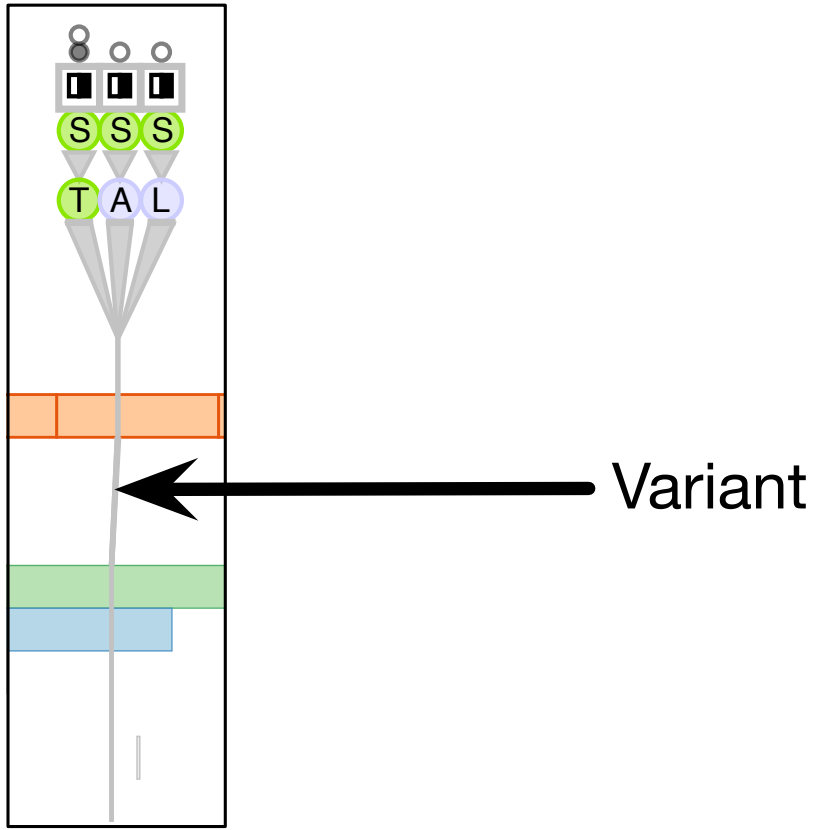
Exon regions small

Color coding difficult to see

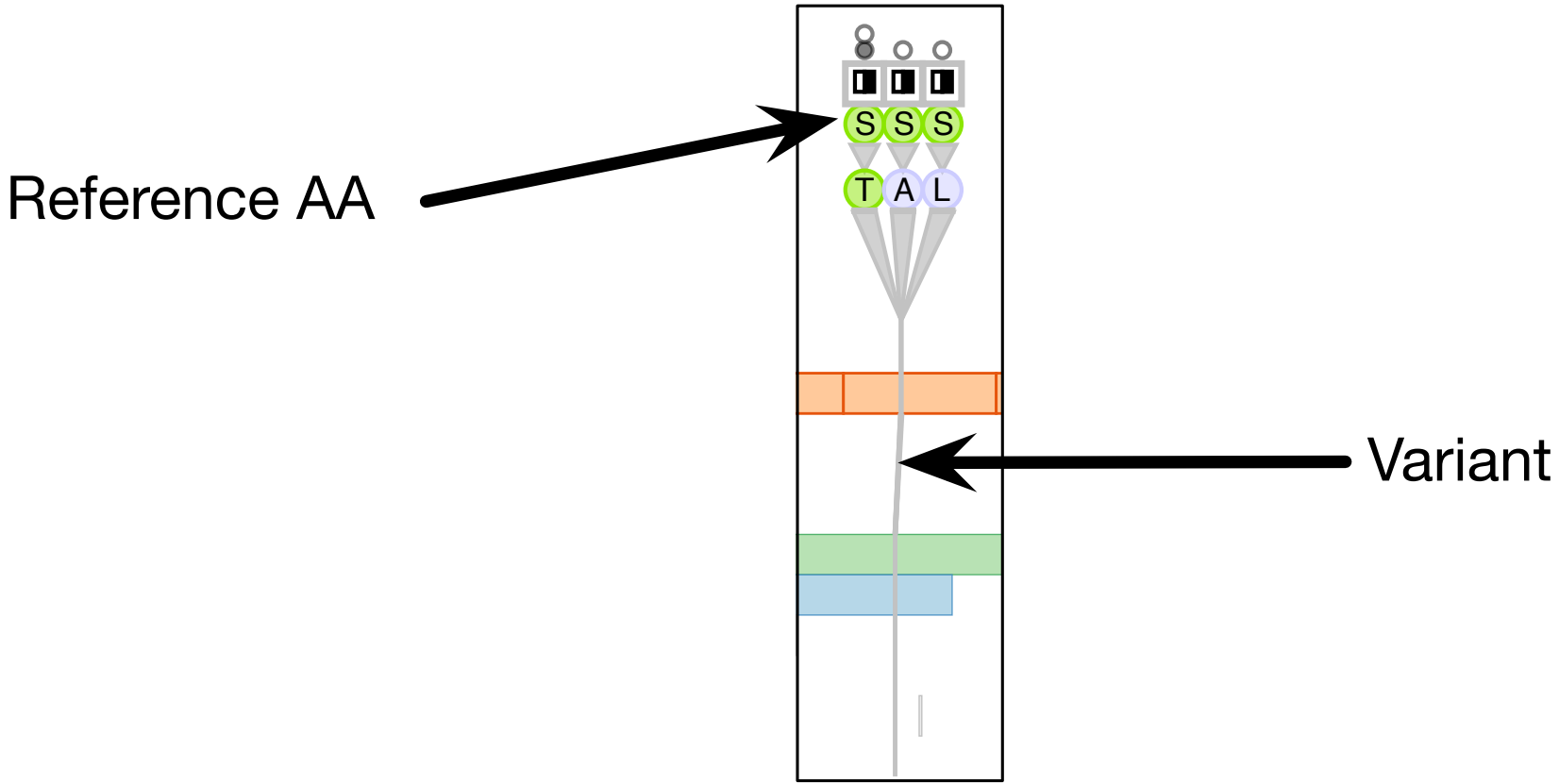
[Chen et al., BMC Genomics 2010]

Our Goal:
Show attributes necessary for **variant analysis**

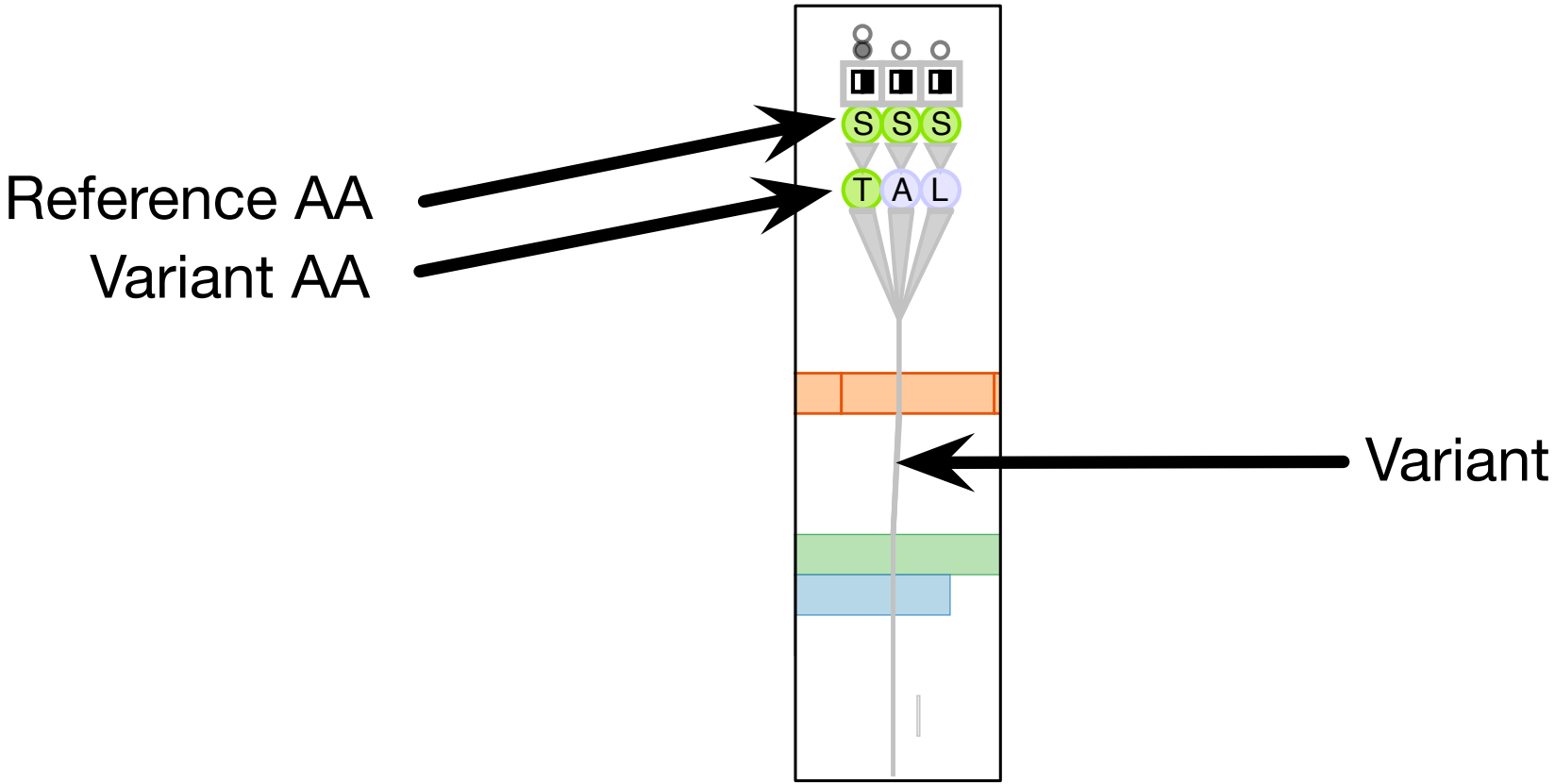
Use information-dense visual encoding



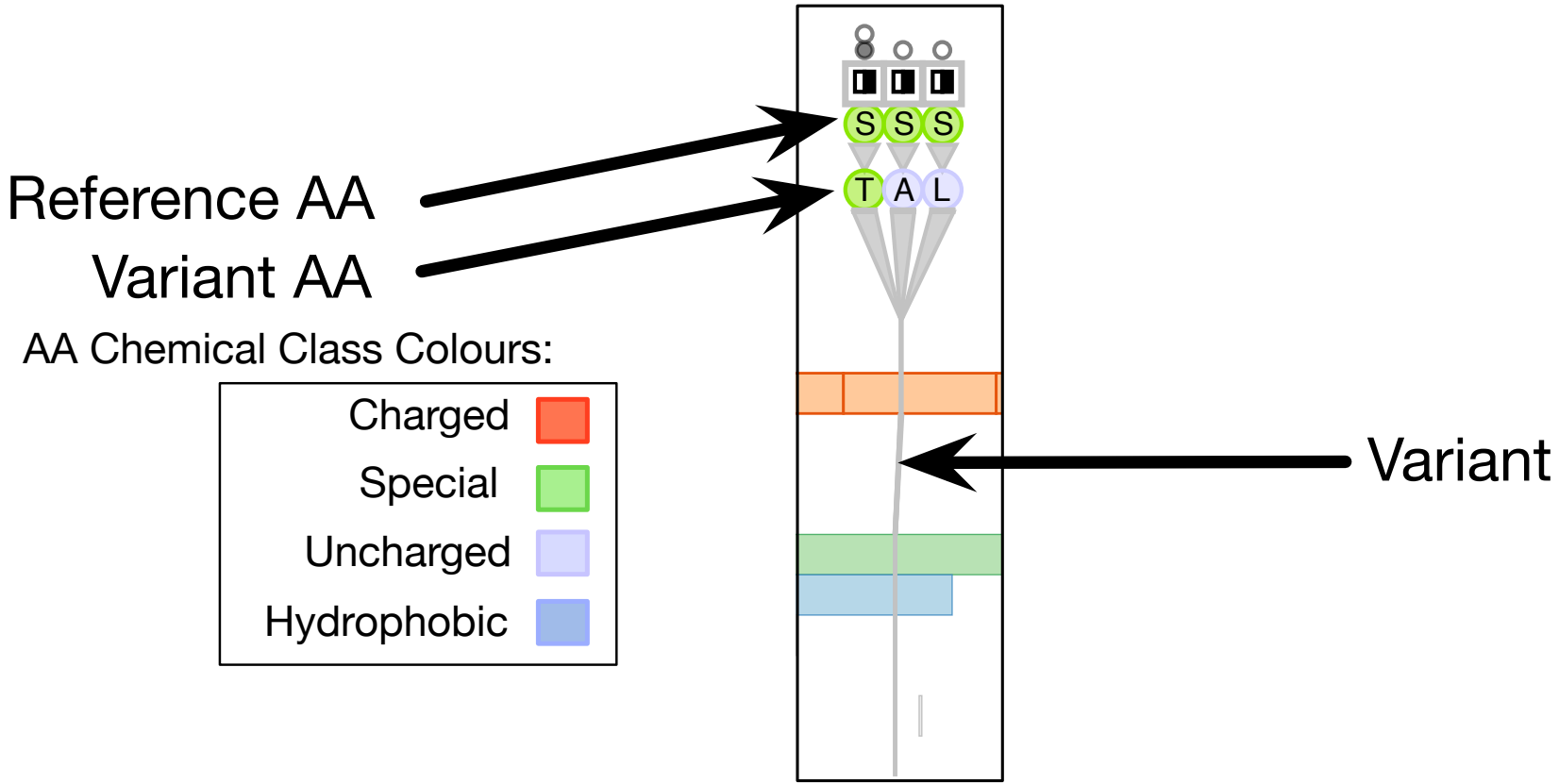
Use information-dense visual encoding



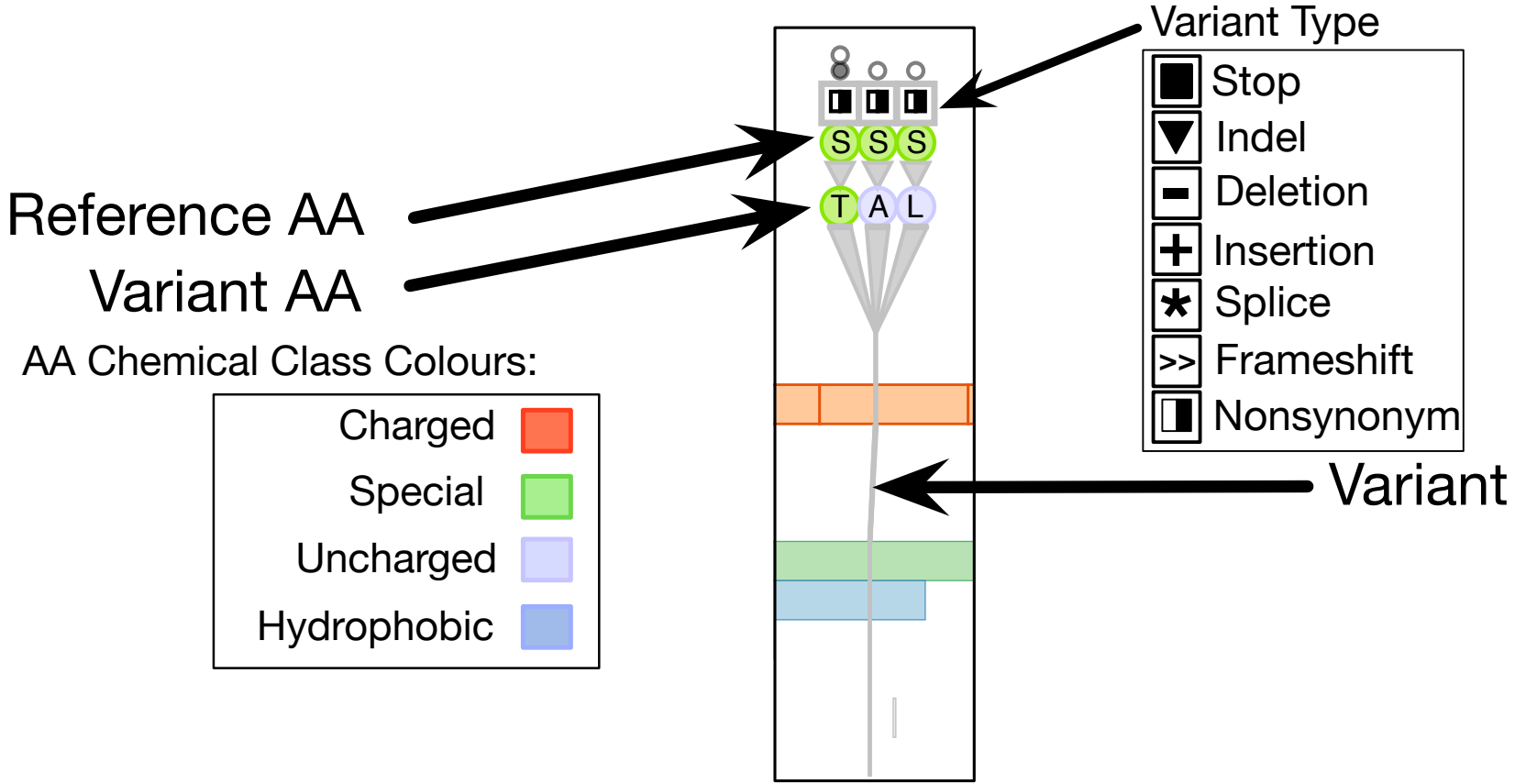
Use information-dense visual encoding



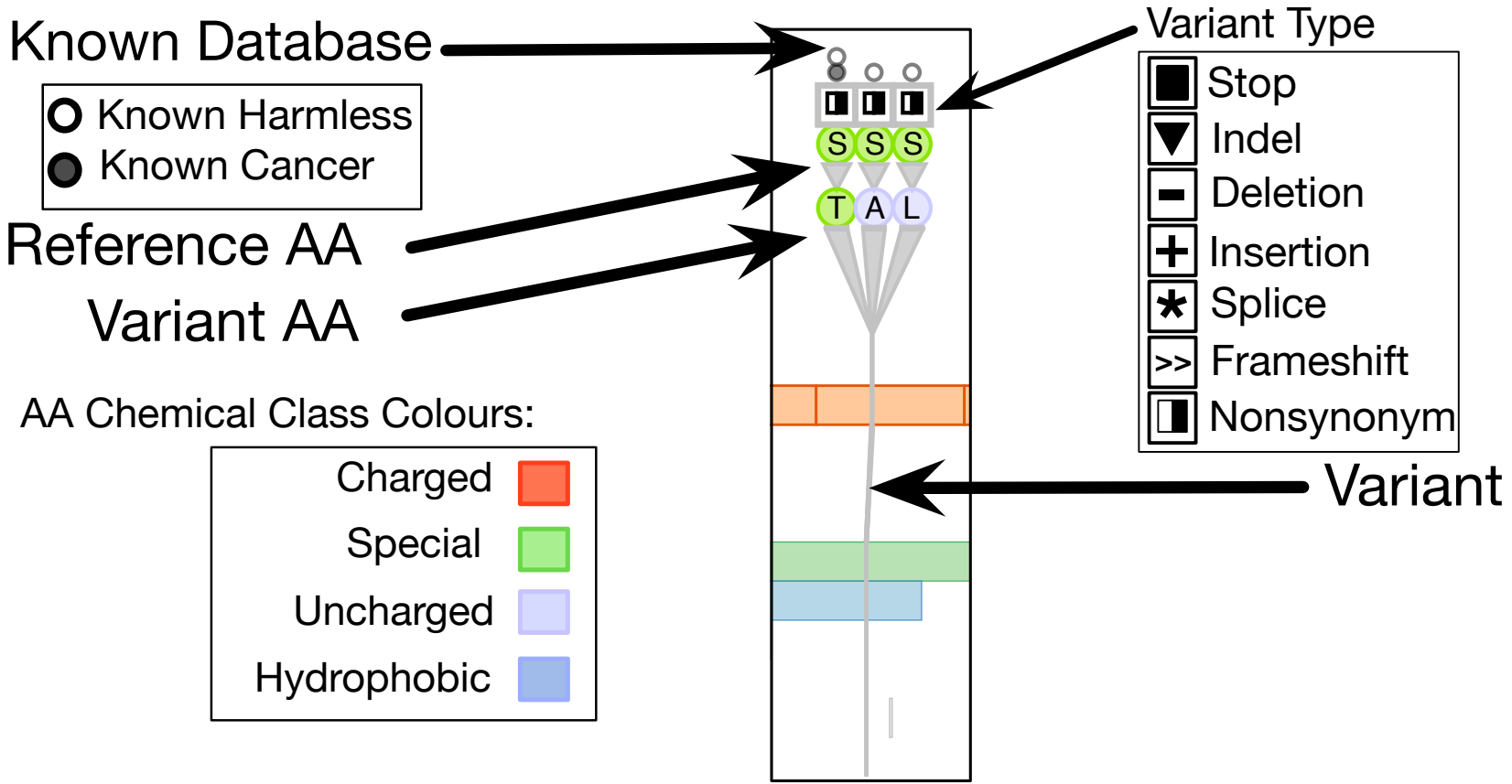
Use information-dense visual encoding



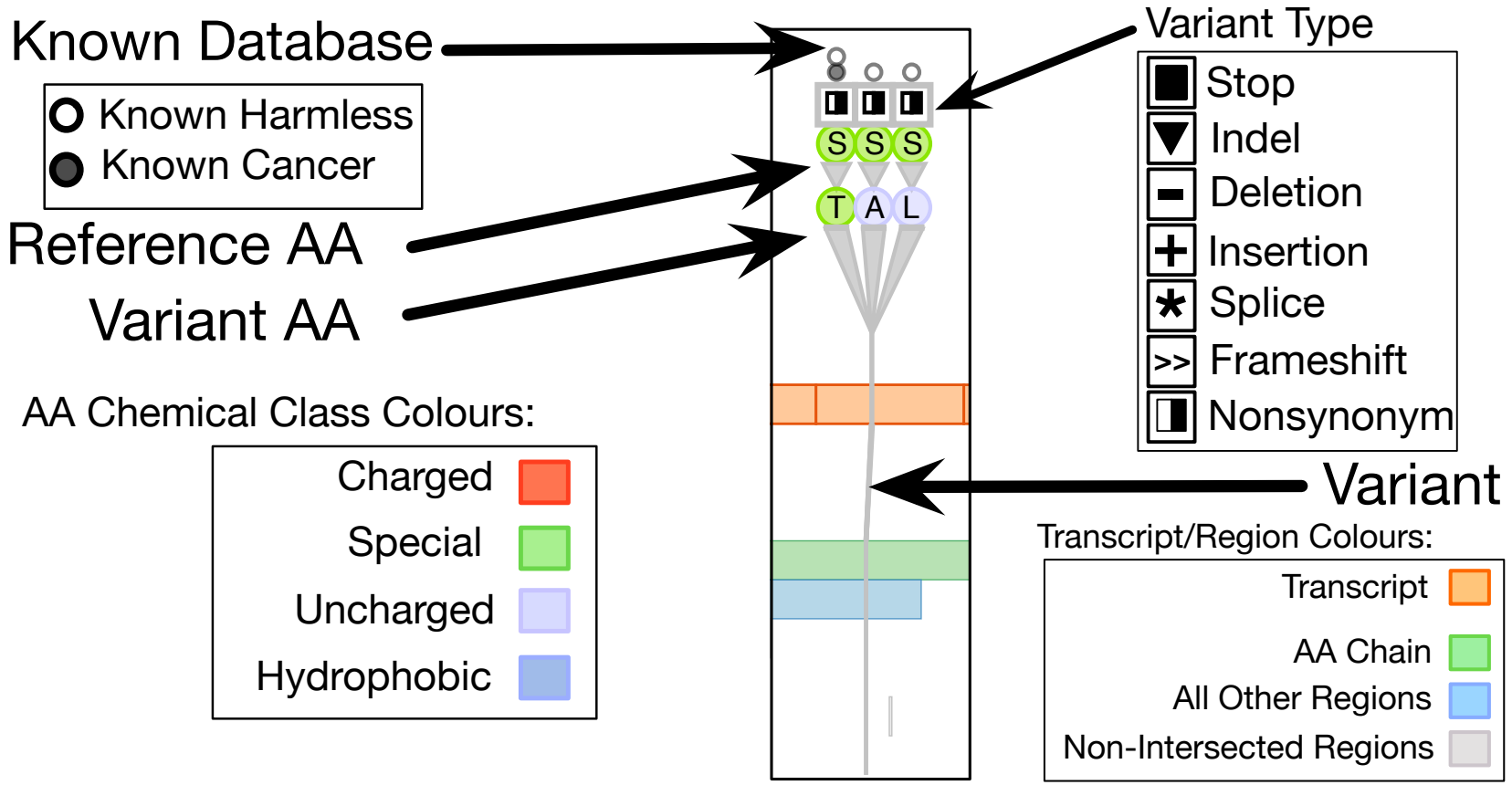
Use information-dense visual encoding



Use information-dense visual encoding



Use information-dense visual encoding



The tool:
Variant View

Variant View

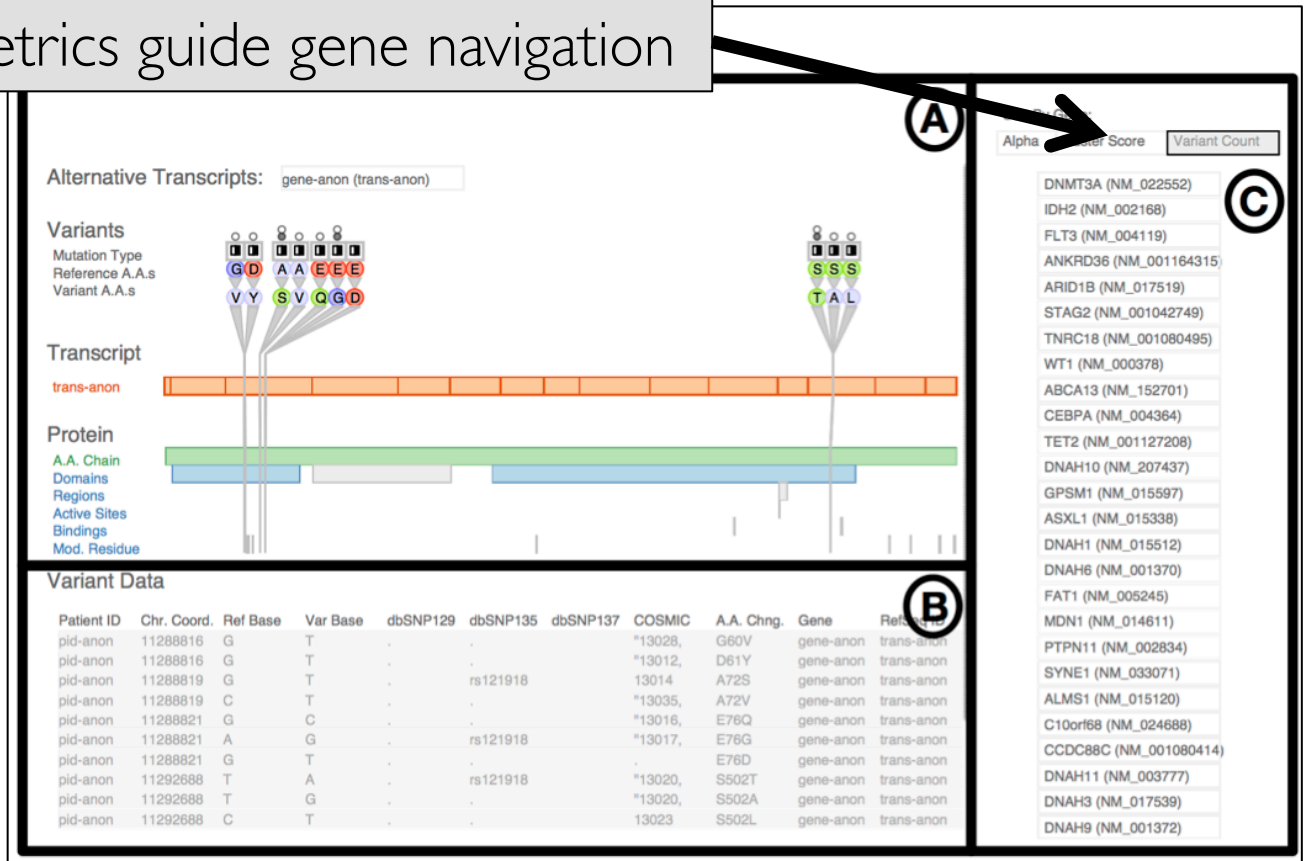
Variant View

Information-dense single gene view

Variant View

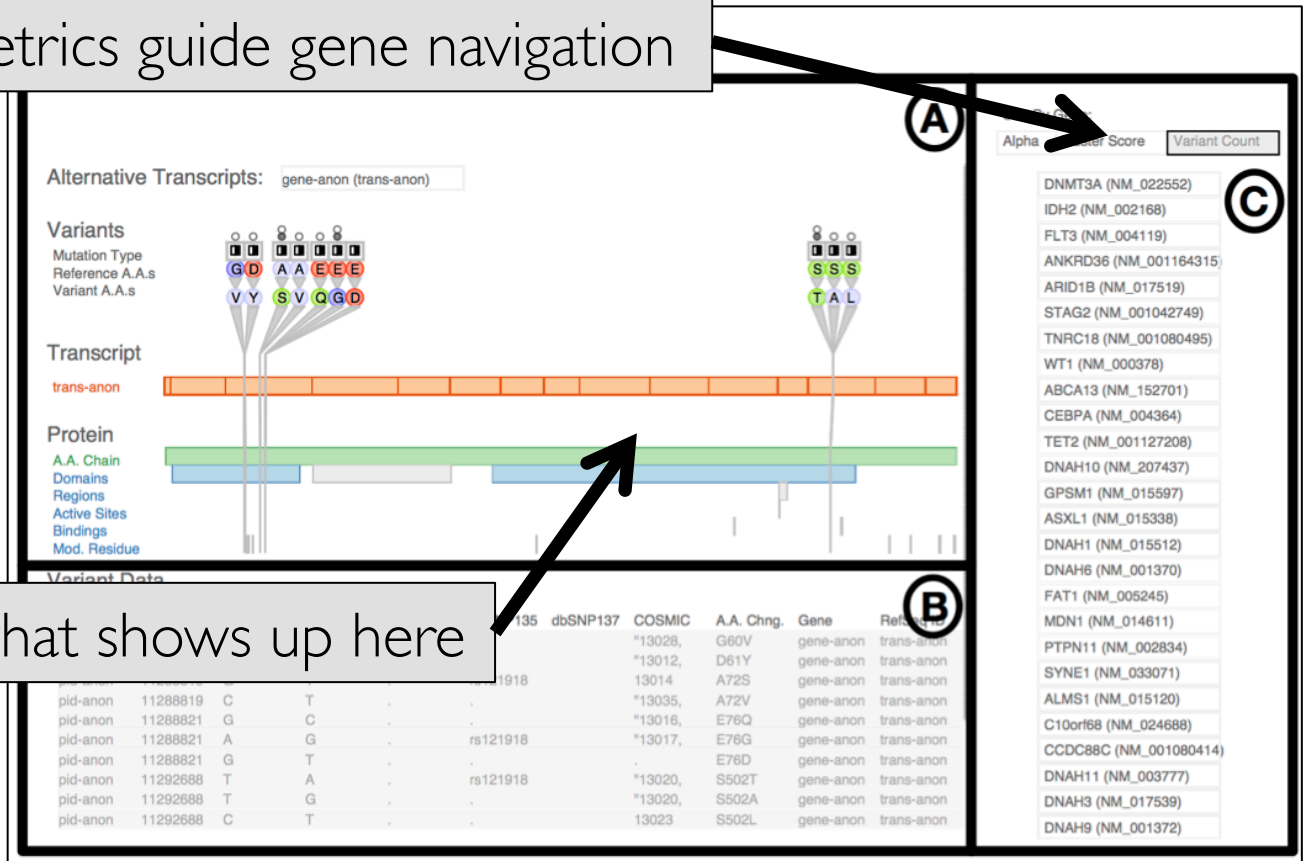
Variant View

Sorting metrics guide gene navigation



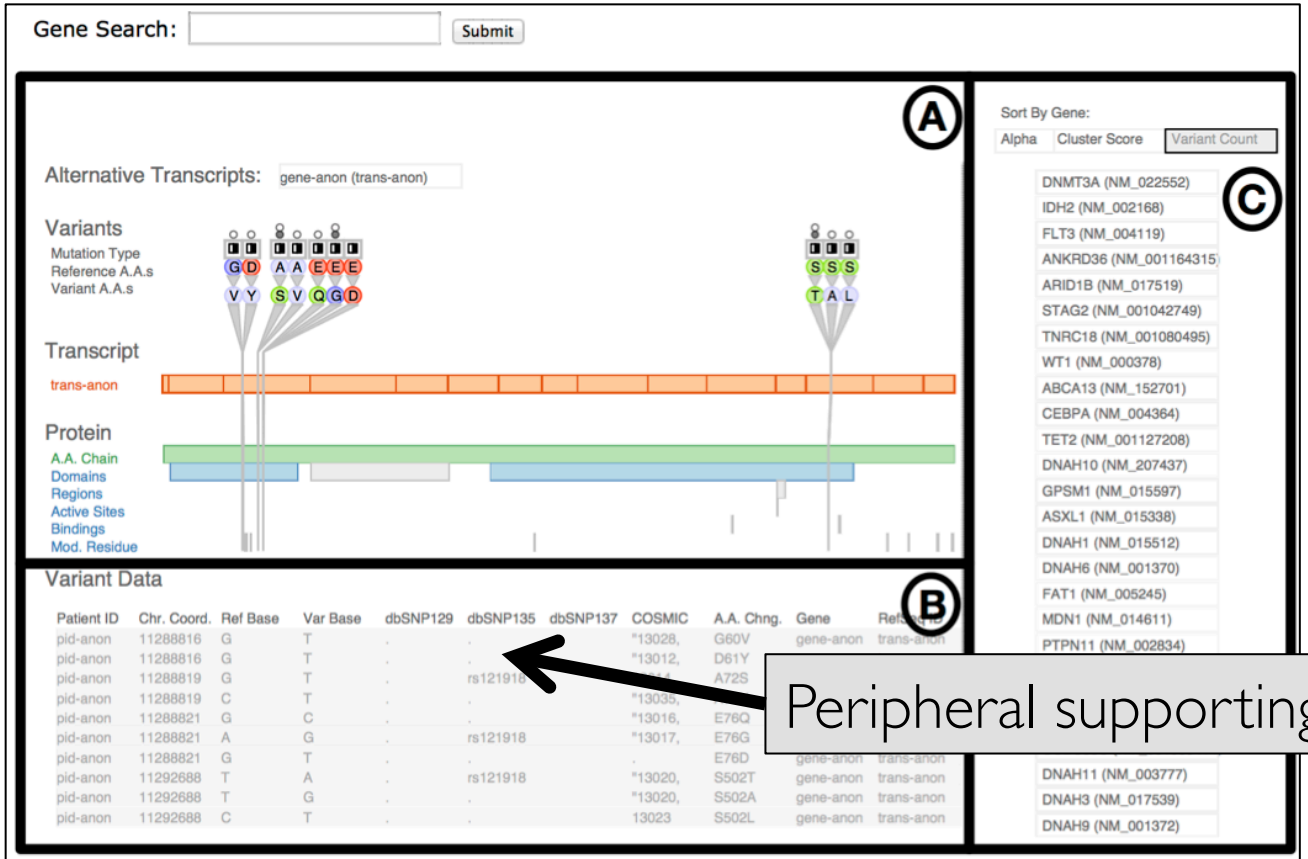
Variant View

Sorting metrics guide gene navigation



Control what shows up here

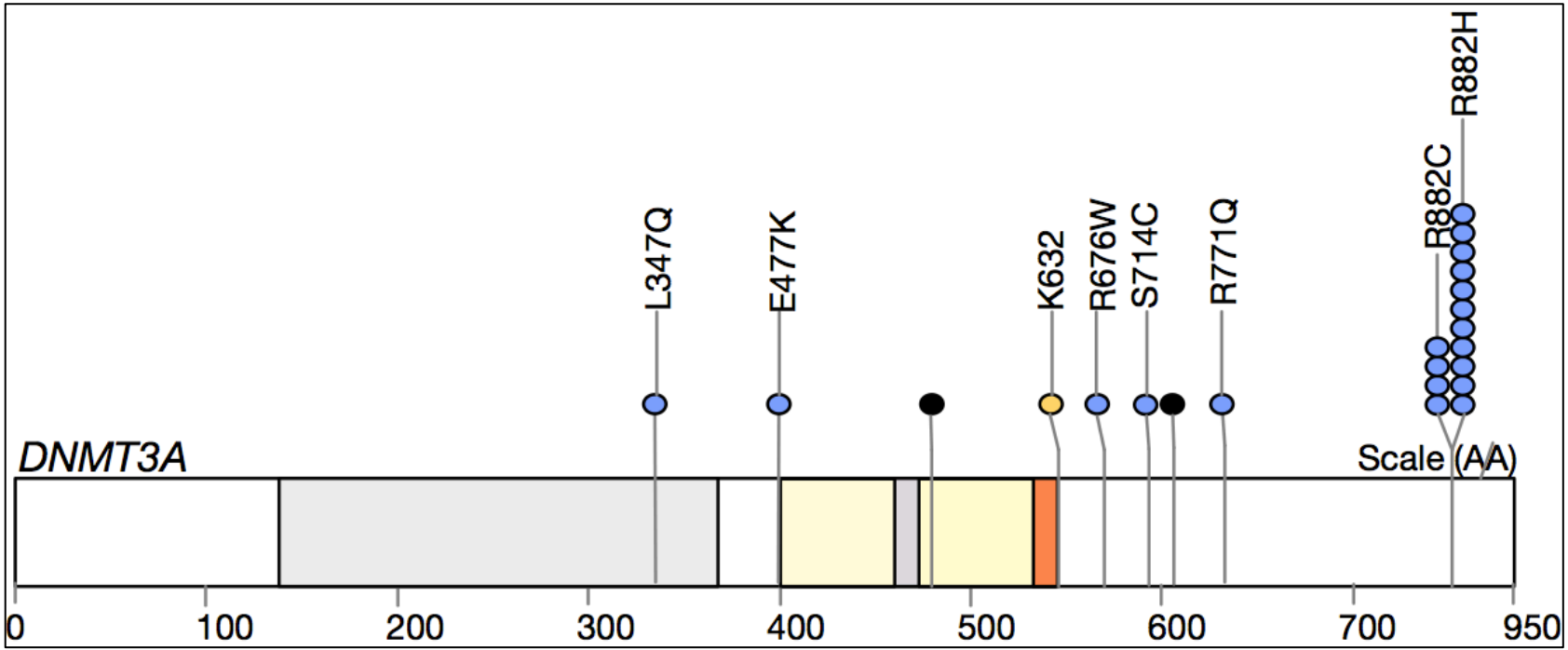
Variant View



Peripheral supporting data

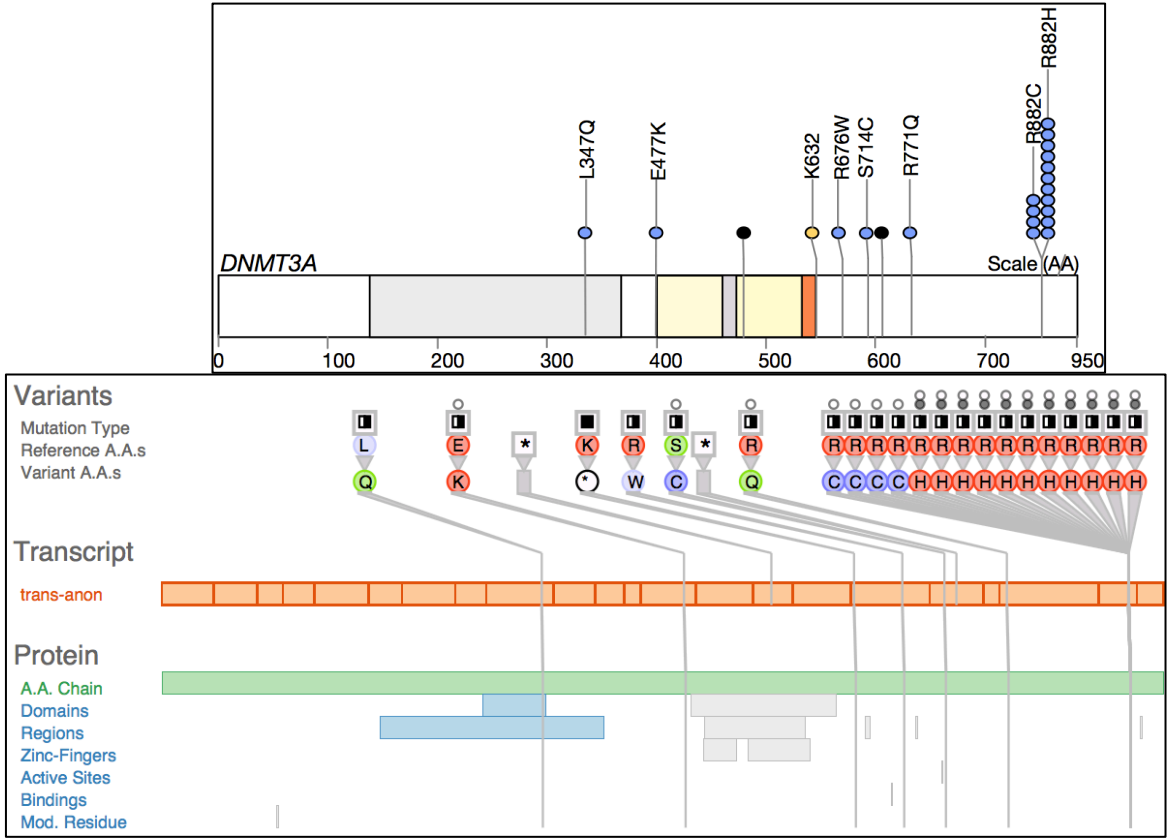
Related work:
Targeted for variant analysis

MuSiC variant visualization plot



[Dees et al., Genome Research 2012]

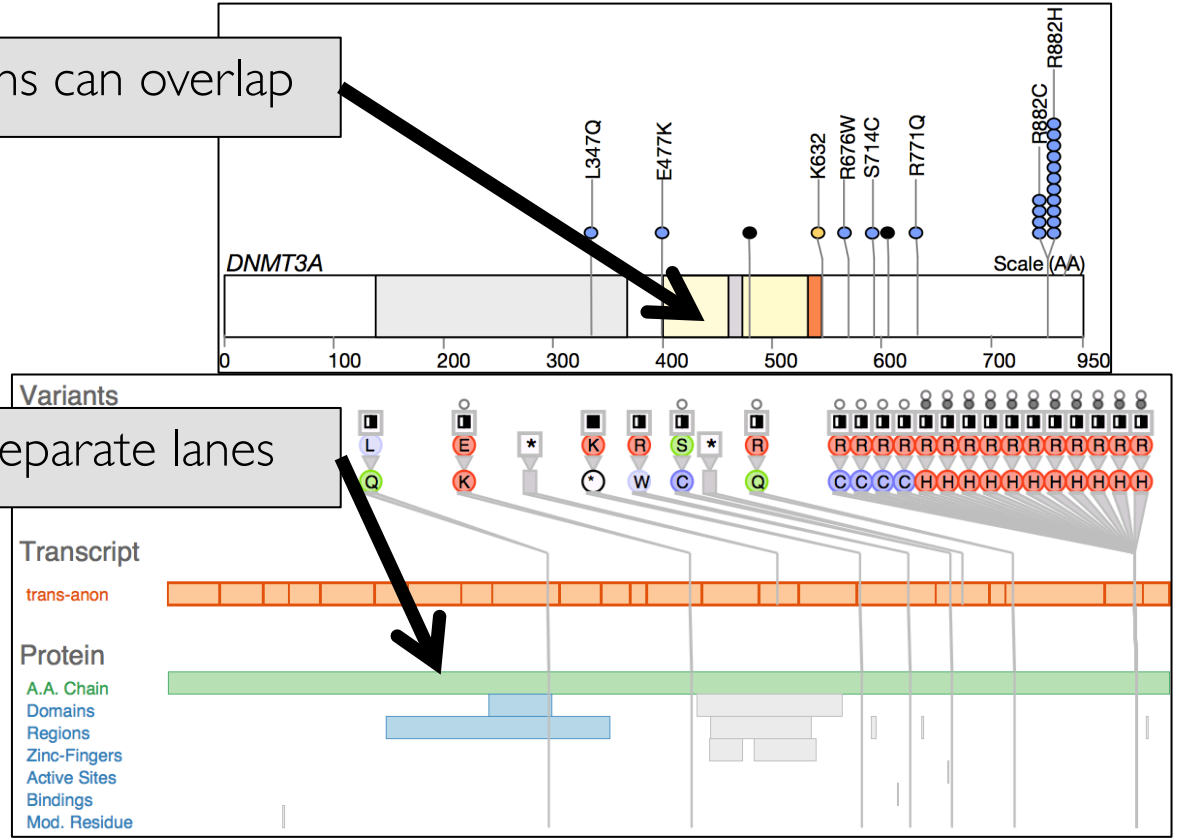
Side-by-side comparison



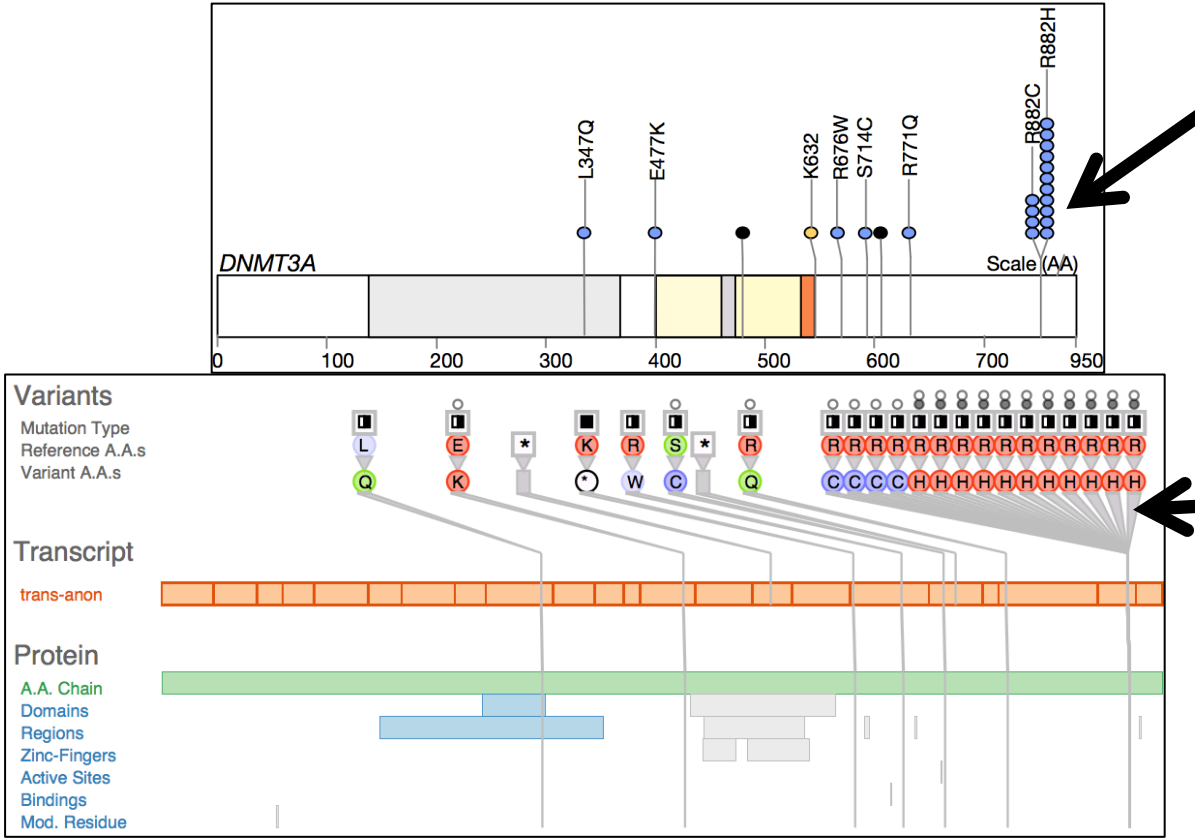
Side-by-side comparison

Protein regions can overlap

Regions get separate lanes



Side-by-side comparison



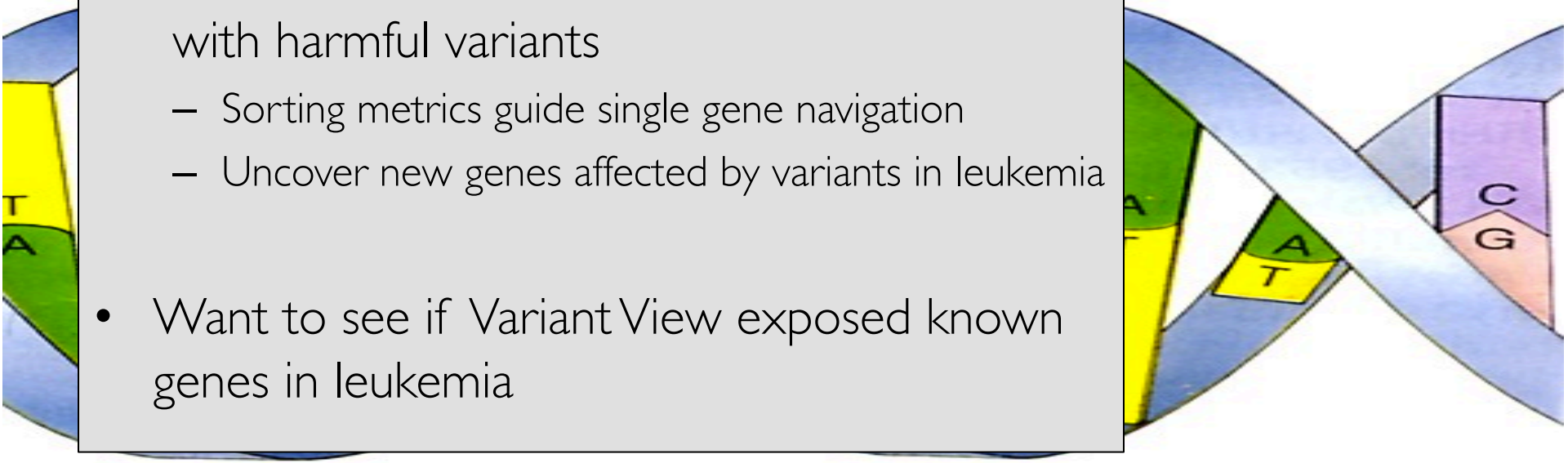
Many collocated variants

Large bloom of repeated elements: more salient

Driving biological tasks

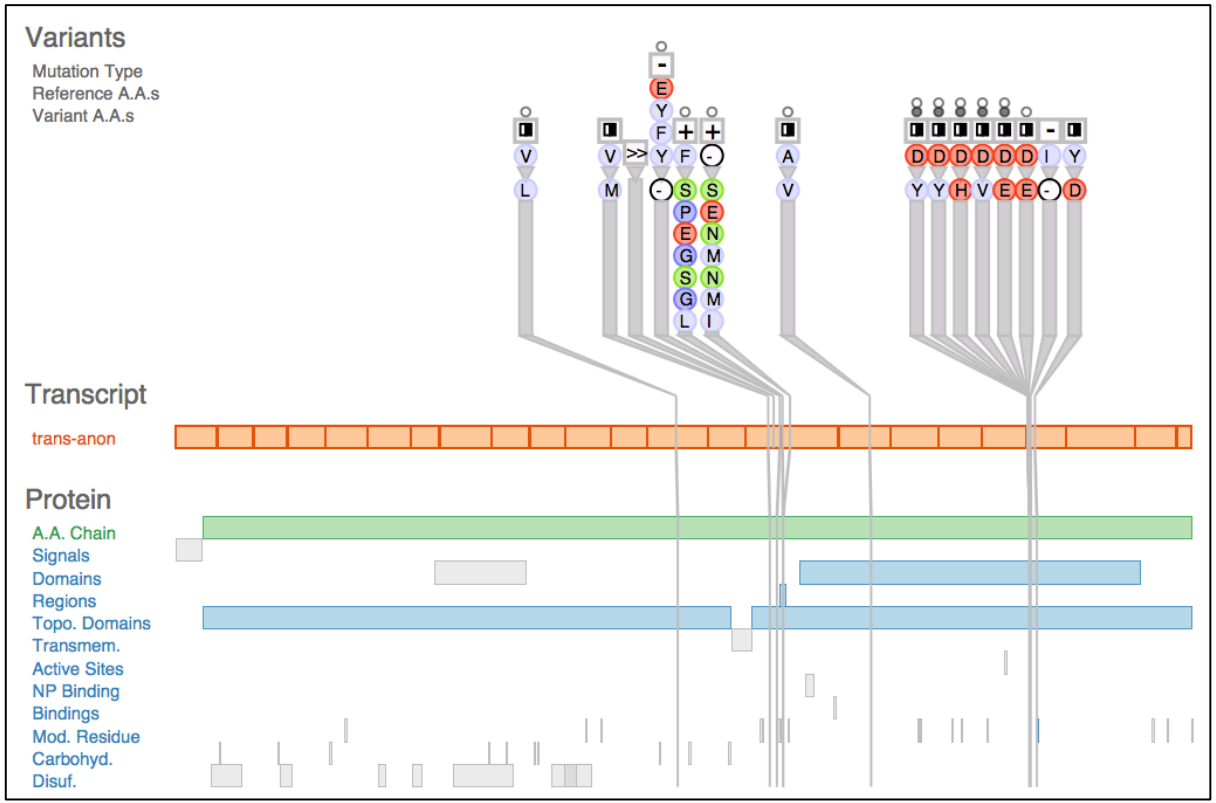
Task 1: Discover genes

- Tool originally designed to **discover genes** with harmful variants
  - Sorting metrics guide single gene navigation
  - Uncover new genes affected by variants in leukemia
- Want to see if Variant View exposed known genes in leukemia

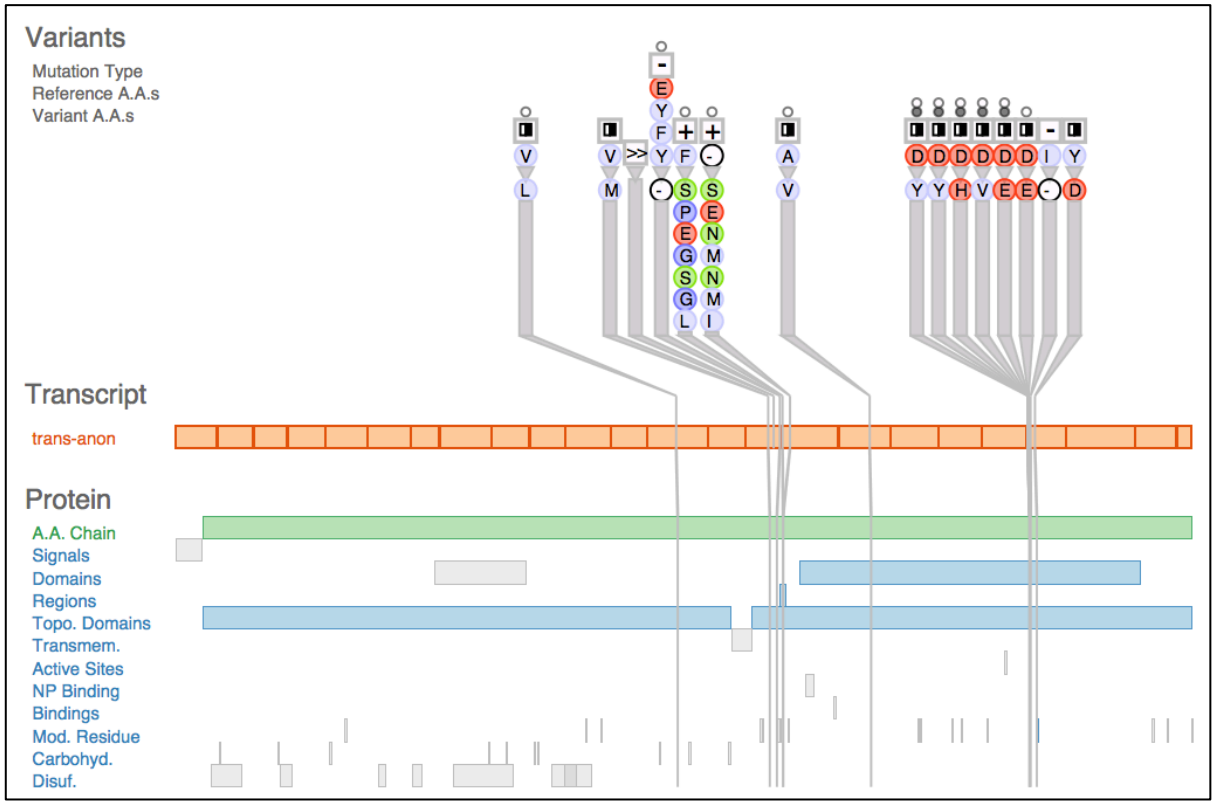



Variant_View_Video.mov
Discover Task
02:35
-01:22

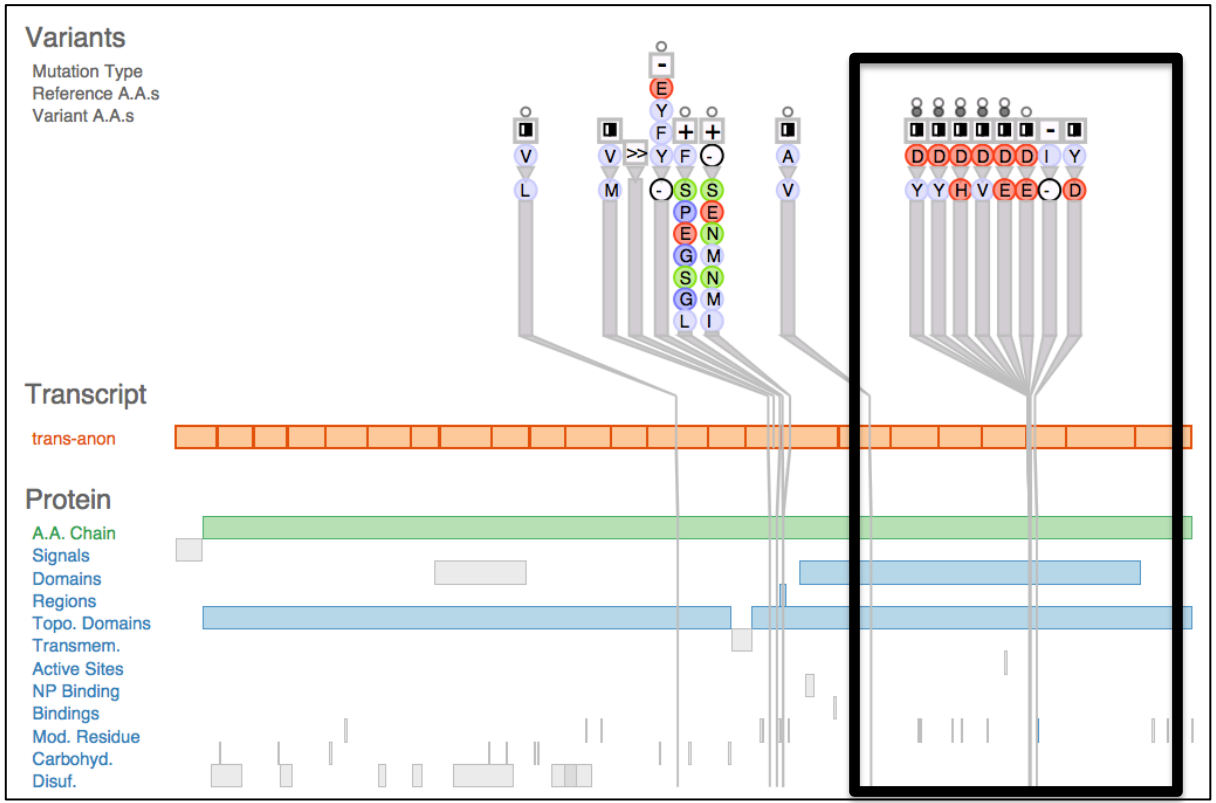
Sorting by derived metric revealed known leukemia genes



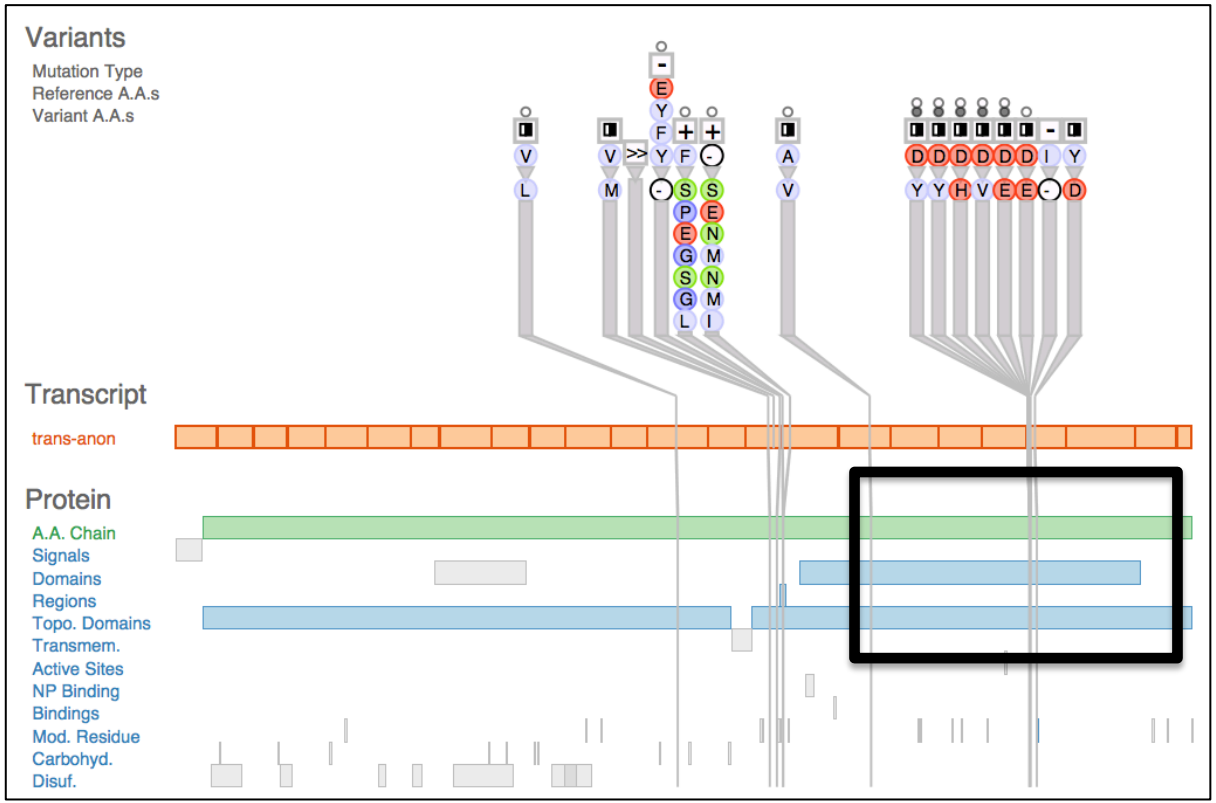
Highly scored gene by sorting metric



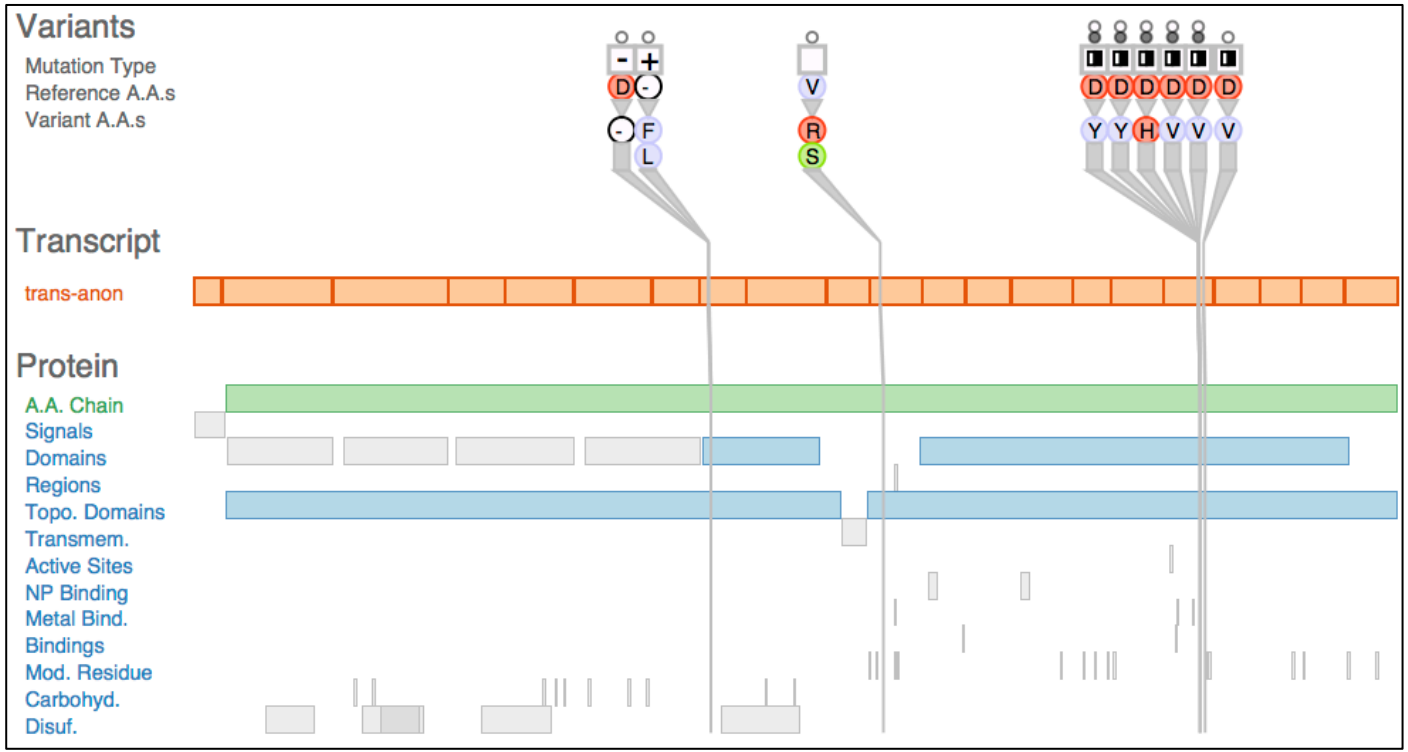
Visual inspection reveals collocation of variants



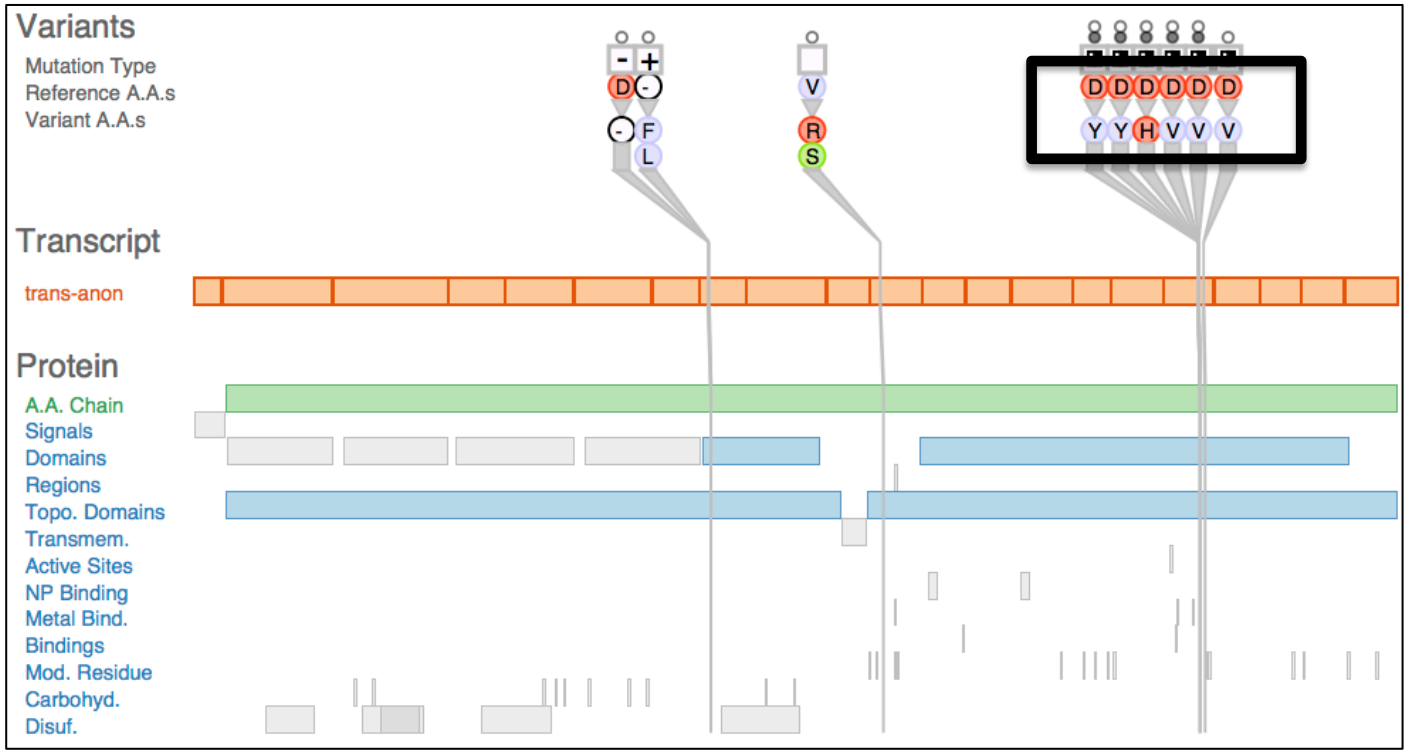
Several functional protein regions affected



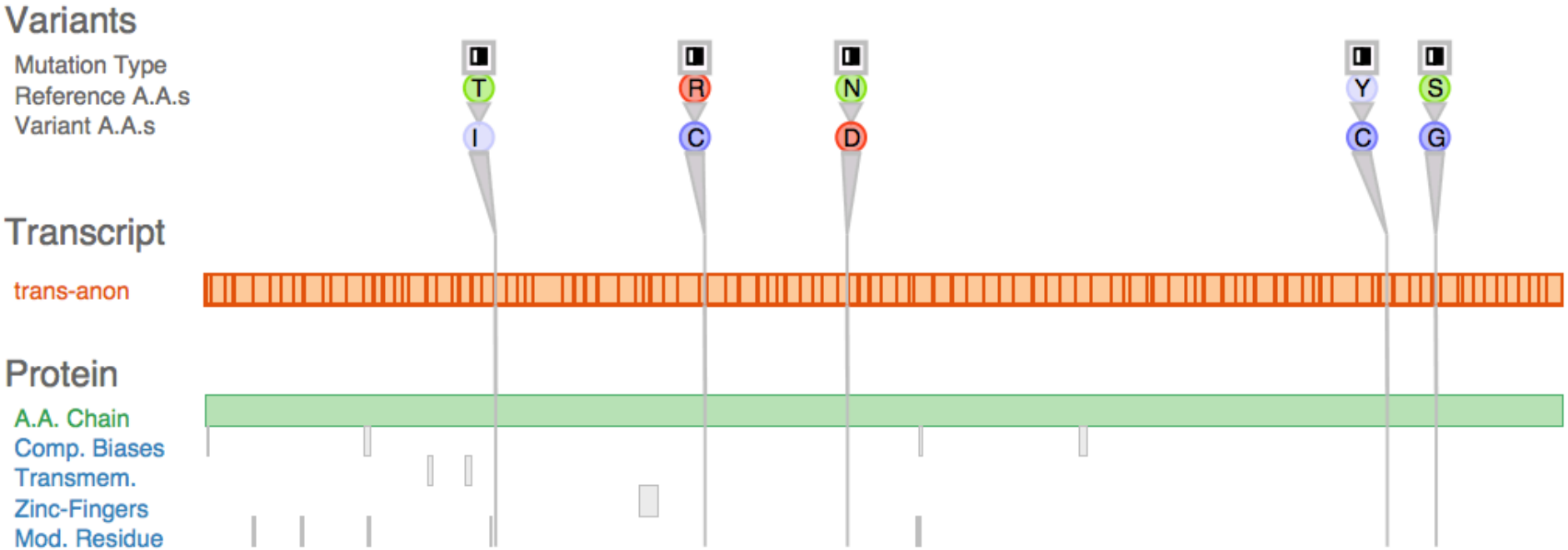
Highly scored by metric and not known



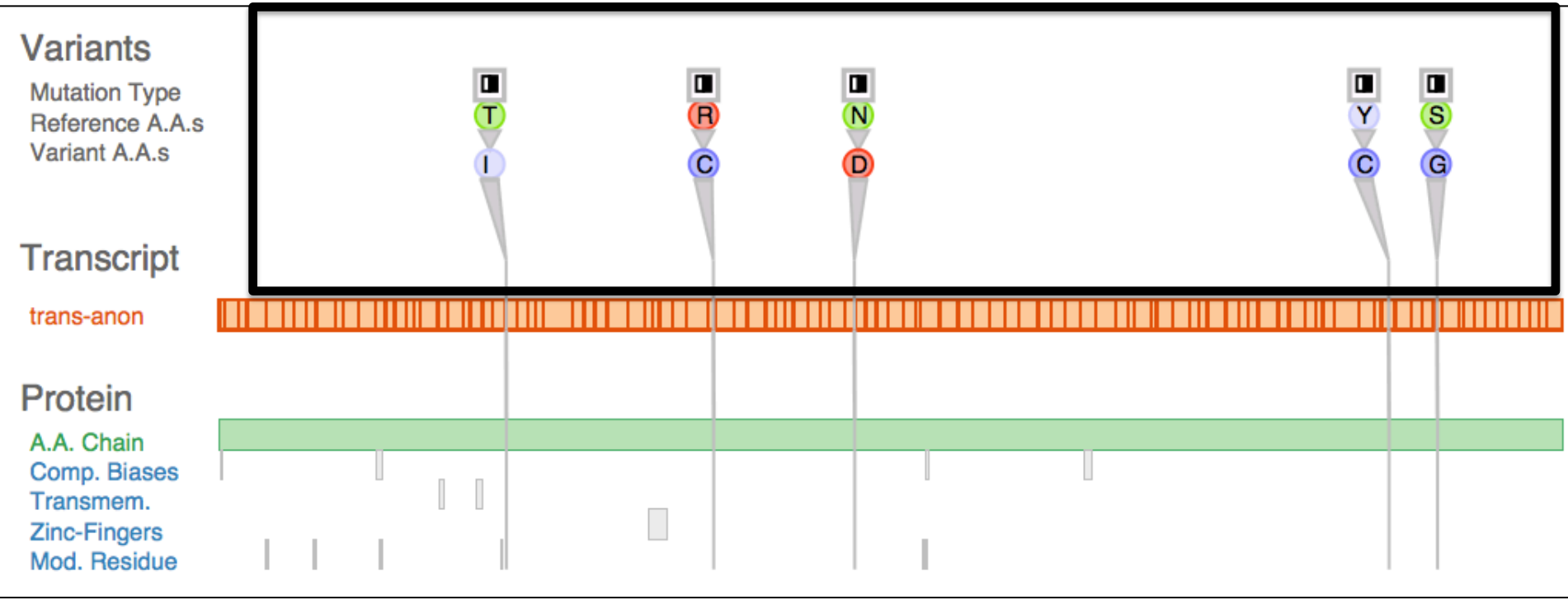
Protein chemical class change evident



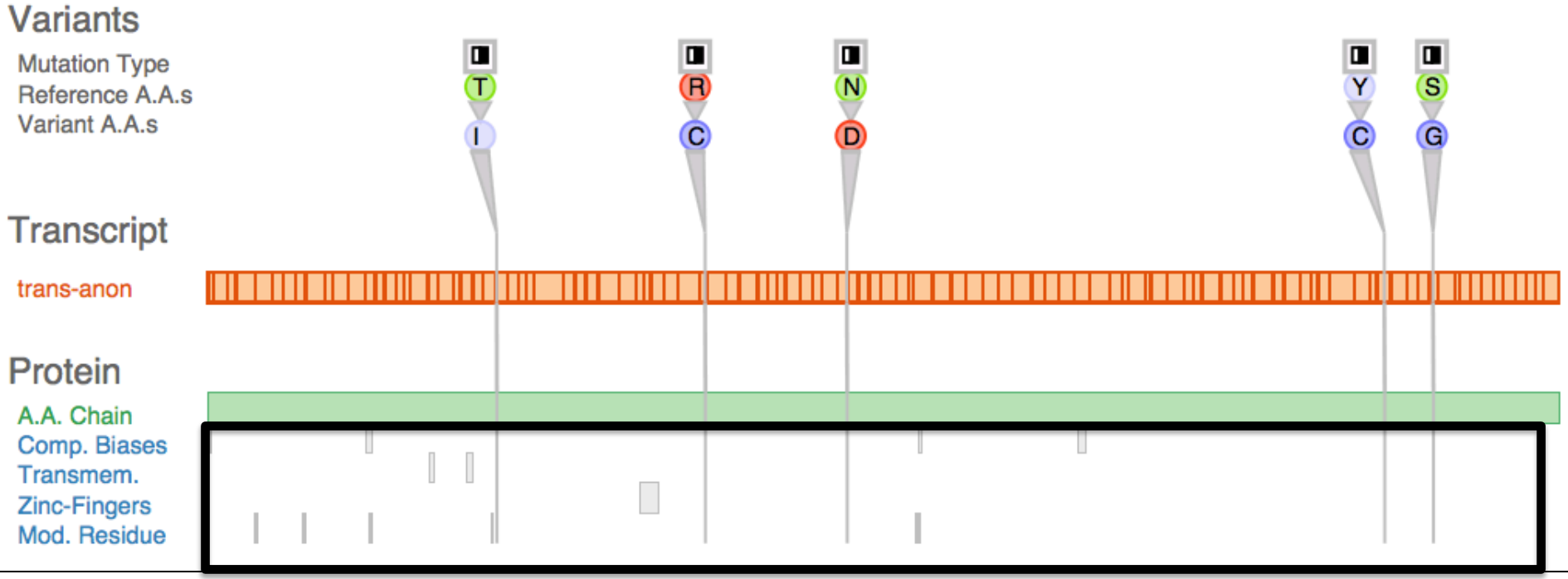
In contrast, low scoring gene



No collocation of variants

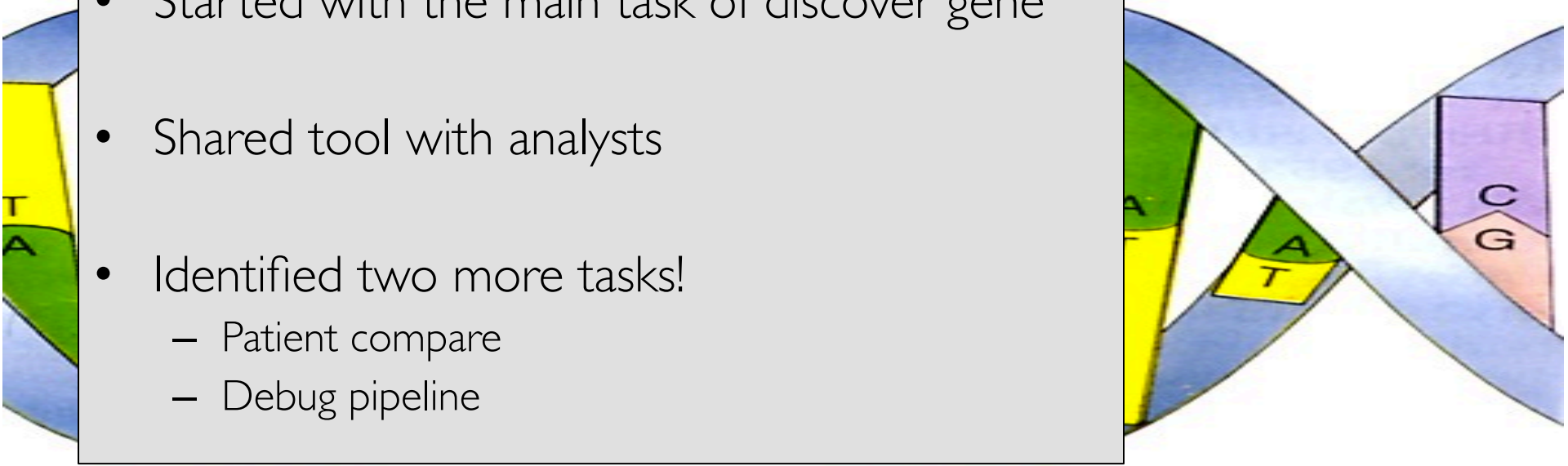


Mostly unaffected protein regions



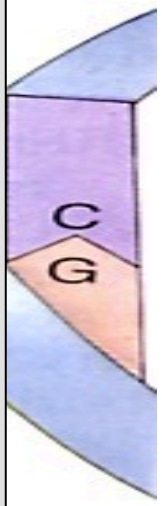
Variant tasks

- Started with the main task of discover gene
- Shared tool with analysts
- Identified two more tasks!
  - Patient compare
  - Debug pipeline



Task 2: Patient compare

- Clinical setting application
- Compare patient data to known harmful variants
- The **challenge**
  - Similarity is loosely understood rather than fully characterized
  - What constitutes a match requires visual inspection



Adapted Variant View with minimal changes

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	-

Navigate through patient data with list

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	-

Patient data emphasized with arrows

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	-

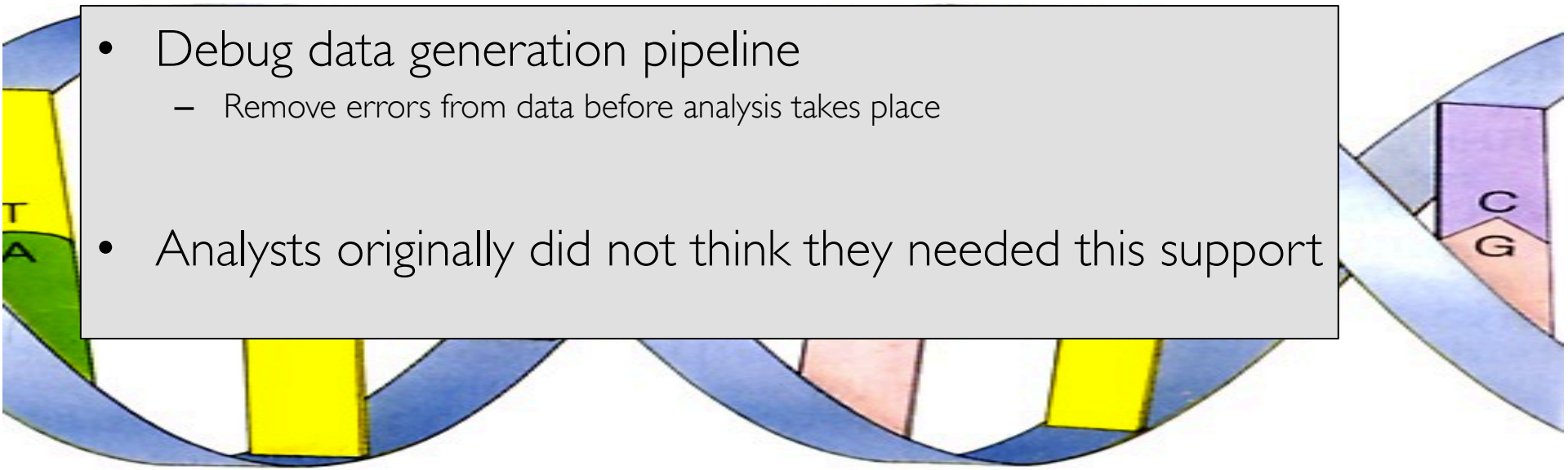
Patient has same harmful L to P mutation

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	-

These variants probably do not match

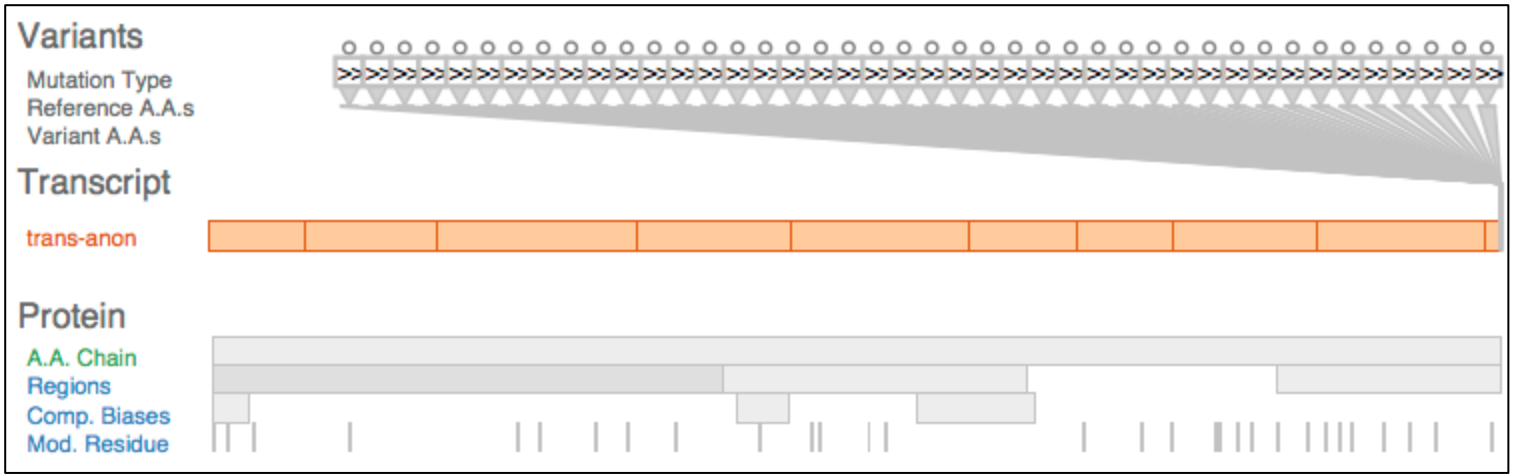
Variant ID	Chr. Coord.	Ref Base	Var Base	Effect Level	Effect Type	Gene Name	Trans. Name	Prot. Coord.
pid-anon	31022959	T	C	MODERATE	NON_SYNONY	gene-anon	trans-anon	L815P
pid-anon	31022959	T	C		NON_SYNONY	gene-anon	trans-anon	L815P
pid-anon	31023029	G	T		NON_SYNONY	gene-anon	trans-anon	K838N
pid-anon	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	-

Task 3: Debug pipeline



- Debug data generation pipeline
  - Remove errors from data before analysis takes place
- Analysts originally did not think they needed this support

Tool revealed errors in the data

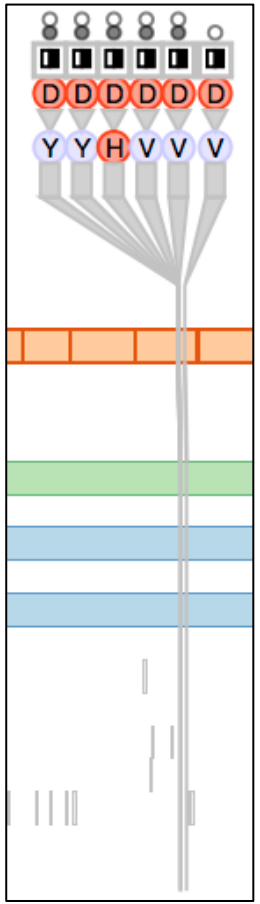


The tool exposed artifacts in the data that slid past at least two rounds of quality metric filtering … this type of problem would not have been caught by our previous, automated methods.

- Analyst 3

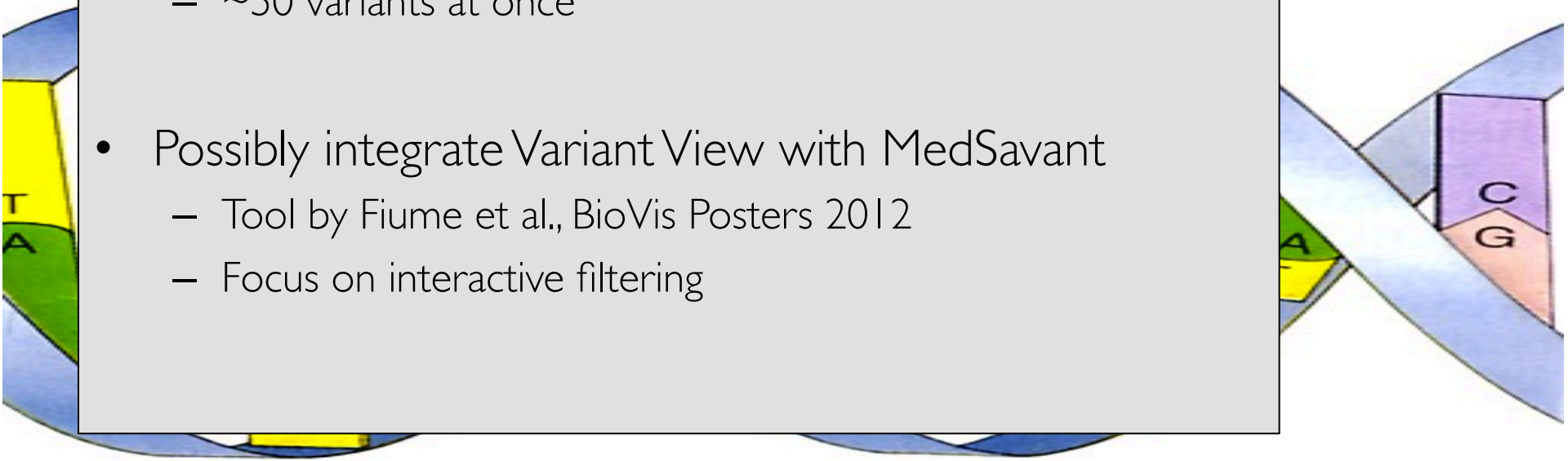
Conclusions

- Designed, implemented, and deployed tool for visual variant impact assessment
- Originally designed for Discover Genes task
  - Adapted to two others with minimal changes
- Methods
  - What to show
    - Filtering data scope
  - How to show it
    - Carefully selected visual encodings
- Goals
  - Navigation-free main overview at gene level
  - Reveal genes of interest; accomplished by sorting by new, derived metrics



Future work: Other use contexts

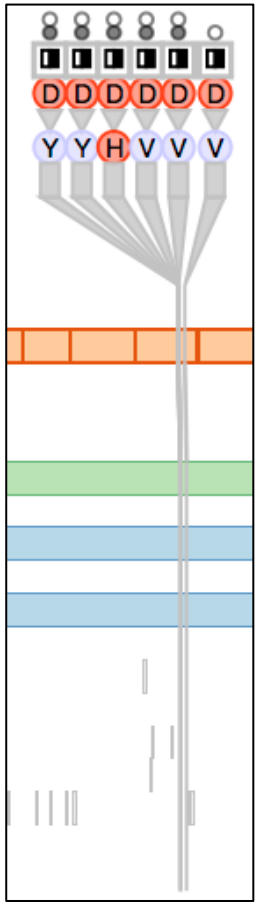
- Scaling up beyond the current design target
  - ~50 variants at once
- Possibly integrate Variant View with MedSavant
  - Tool by Fiume et al., BioVis Posters 2012
  - Focus on interactive filtering



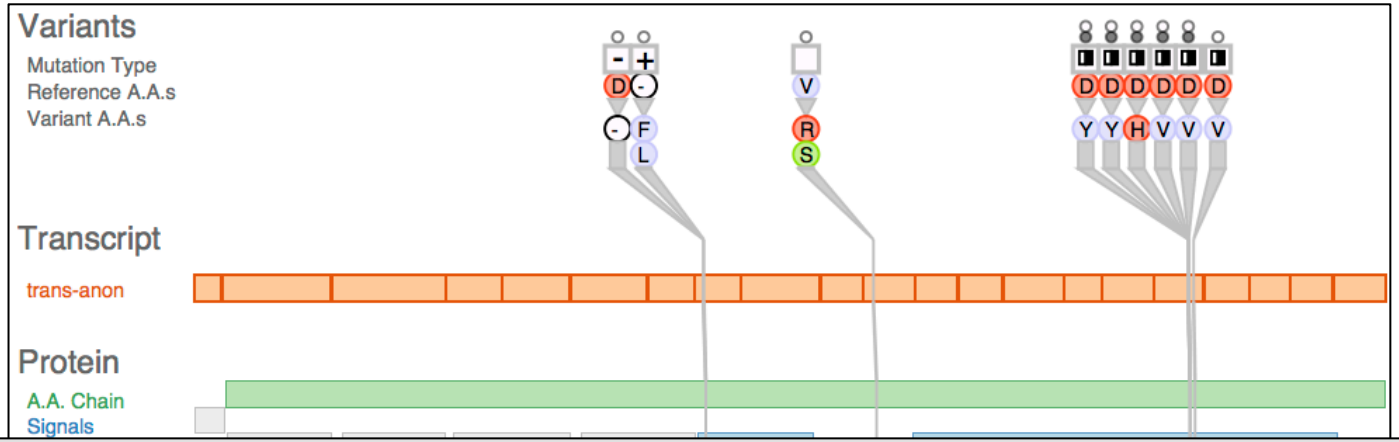
Acknowledgements

- Our collaborators at the GSC
  - Dr Aly Karsan
  - Rod Docking
  - Dr Linda Chang
  - Dr Gerben Duns
  - Simon Chang

- Funding
  - VIVA, AeroInfo Systems/Boeing Canada, MITACS



Questions?



Joel Ferstay: **joel.ferstay@gmail.com**

Paper Page:

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

Software Available as Open Source