

trans-anon

# Variant View

Visualizing Sequence Variants in their Gene Context

Joel A. Ferstay<sup>1</sup>\*, Cydney B. Nielsen<sup>2</sup>, Tamara Munzner<sup>1</sup>

University of British Columbia<sup>1</sup>, now at AerolInfo Systems/Boeing Canada\*  
BC Cancer Agency<sup>2</sup>

# Design study

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



Analyst

Tool

A

Gene Search: [ ] Submit [ ]

Sort By Gene: Alpha Cluster Score Variant Count

Alternative Transcripts: gene-antron (trans-anon)

Variants

Mutation Type

Reference A.A.s

Variant A.A.s

V V S V G G T A L

Transcript

trans-anon

Protein

A.A. Chain

Domains

Active Sites

Bondings

Met. Residue

B

Variant List

Patient ID	Chr.	Coord.	Ref. Base	Var. Base	dbSNP129	dbSNP135	dbSNP137	COSMC	AA Chng.	Gene	Ref. Gene
pid-antron	11288616	G	T	-	rs121918	-	-	*13038	S69Y	gene-antron	trans-anon
pid-antron	11288616	G	T	-	-	-	-	*13012	A72S	gene-antron	trans-anon
pid-antron	11288616	G	T	-	-	-	-	*13026	A71Y	gene-antron	trans-anon
pid-antron	11288621	G	C	-	-	-	-	*13016	E79Q	gene-antron	trans-anon
pid-antron	11288621	G	C	-	-	-	-	*13017	E79Q	gene-antron	trans-anon
pid-antron	11292988	T	A	-	rs121918	-	-	*13020	E79W	gene-antron	trans-anon
pid-antron	11292988	T	A	-	-	-	-	*13020	S90T	gene-antron	trans-anon
pid-antron	11292988	T	G	-	-	-	-	*13020	S90A	gene-antron	trans-anon
pid-antron	11292988	C	T	-	-	-	-	*13033	S90L	gene-antron	trans-anon

C

Sort By Gene: Alpha Cluster Score Variant Count

DNAH1 (NM\_022502) DNAH1 (NM\_022502) DNAH1 (NM\_022502)  
ANKRD26 (NM\_0014419) ANKRD26 (NM\_0014419) ANKRD26 (NM\_0014419)  
ARD1B (NM\_017319) ARD1B (NM\_017319) ARD1B (NM\_017319)  
STAU2 (NM\_00104249) STAU2 (NM\_00104249) STAU2 (NM\_00104249)  
TNRC18 (NM\_001080498) TNRC18 (NM\_001080498) TNRC18 (NM\_001080498)  
WT1 (NM\_000379) WT1 (NM\_000379) WT1 (NM\_000379)  
ABCAT1 (NM\_152701) ABCAT1 (NM\_152701) ABCAT1 (NM\_152701)  
CEBPB (NM\_004364) CEBPB (NM\_004364) CEBPB (NM\_004364)  
DNAH1 (NM\_022502) DNAH1 (NM\_022502) DNAH1 (NM\_022502)  
DNAH1 (NM\_207437) DNAH1 (NM\_207437) DNAH1 (NM\_207437)  
GPM6 (NM\_015597) GPM6 (NM\_015597) GPM6 (NM\_015597)  
ASXL1 (NM\_015338) ASXL1 (NM\_015338) ASXL1 (NM\_015338)  
DNAH1 (NM\_015512) DNAH1 (NM\_015512) DNAH1 (NM\_015512)  
DNAH1 (NM\_001370) DNAH1 (NM\_001370) DNAH1 (NM\_001370)  
FAT1 (NM\_002459) FAT1 (NM\_002459) FAT1 (NM\_002459)  
MDN1 (NM\_014611) MDN1 (NM\_014611) MDN1 (NM\_014611)  
PRPF8 (NM\_003854) PRPF8 (NM\_003854) PRPF8 (NM\_003854)  
EYRD (NM\_003759) EYRD (NM\_003759) EYRD (NM\_003759)  
ALMS1 (NM\_015129) ALMS1 (NM\_015129) ALMS1 (NM\_015129)  
C10orf88 (NM\_024688) C10orf88 (NM\_024688) C10orf88 (NM\_024688)  
CCDC88C (NM\_024688) CCDC88C (NM\_024688) CCDC88C (NM\_024688)  
DNAH11 (NM\_001377) DNAH11 (NM\_001377) DNAH11 (NM\_001377)  
DNAH9 (NM\_017339) DNAH9 (NM\_017339) DNAH9 (NM\_017339)  
DNAH9 (NM\_01372) DNAH9 (NM\_01372) DNAH9 (NM\_01372)

# 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



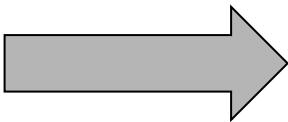
Data and Tasks



Create Data Sketch

# Design cycle

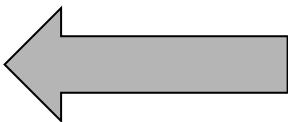
Interview



Data and Tasks



Present Data Sketch



Create Data Sketch

# Design cycle

Interview



REPEAT

Data and Tasks

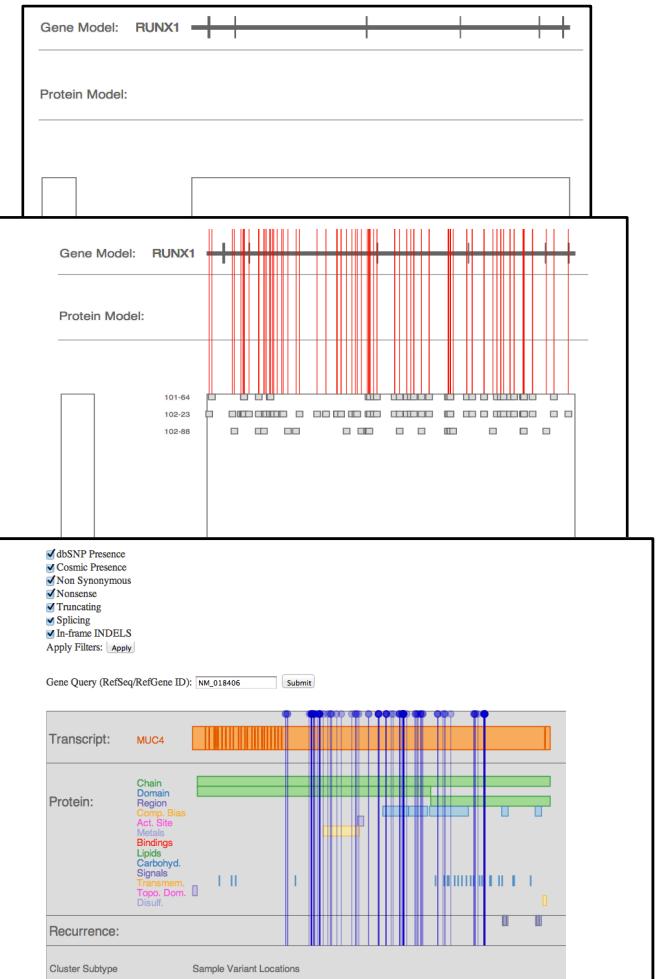


Present Data Sketch

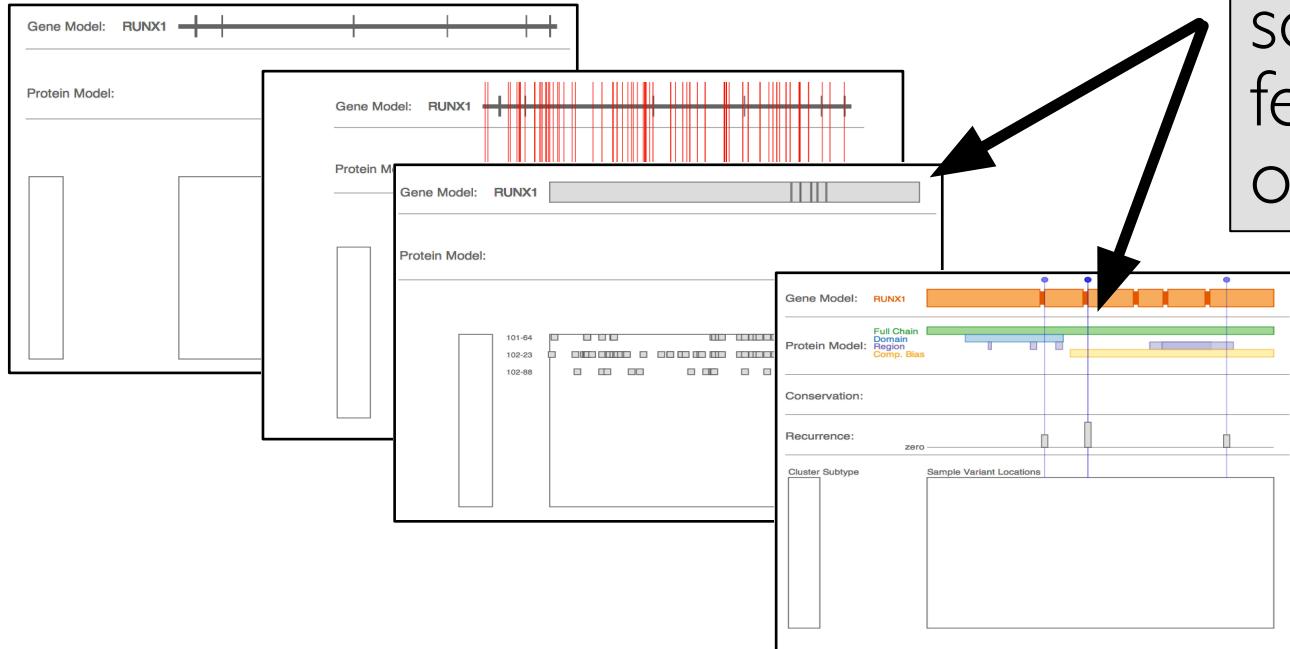
Create Data Sketch

# Data sketches

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

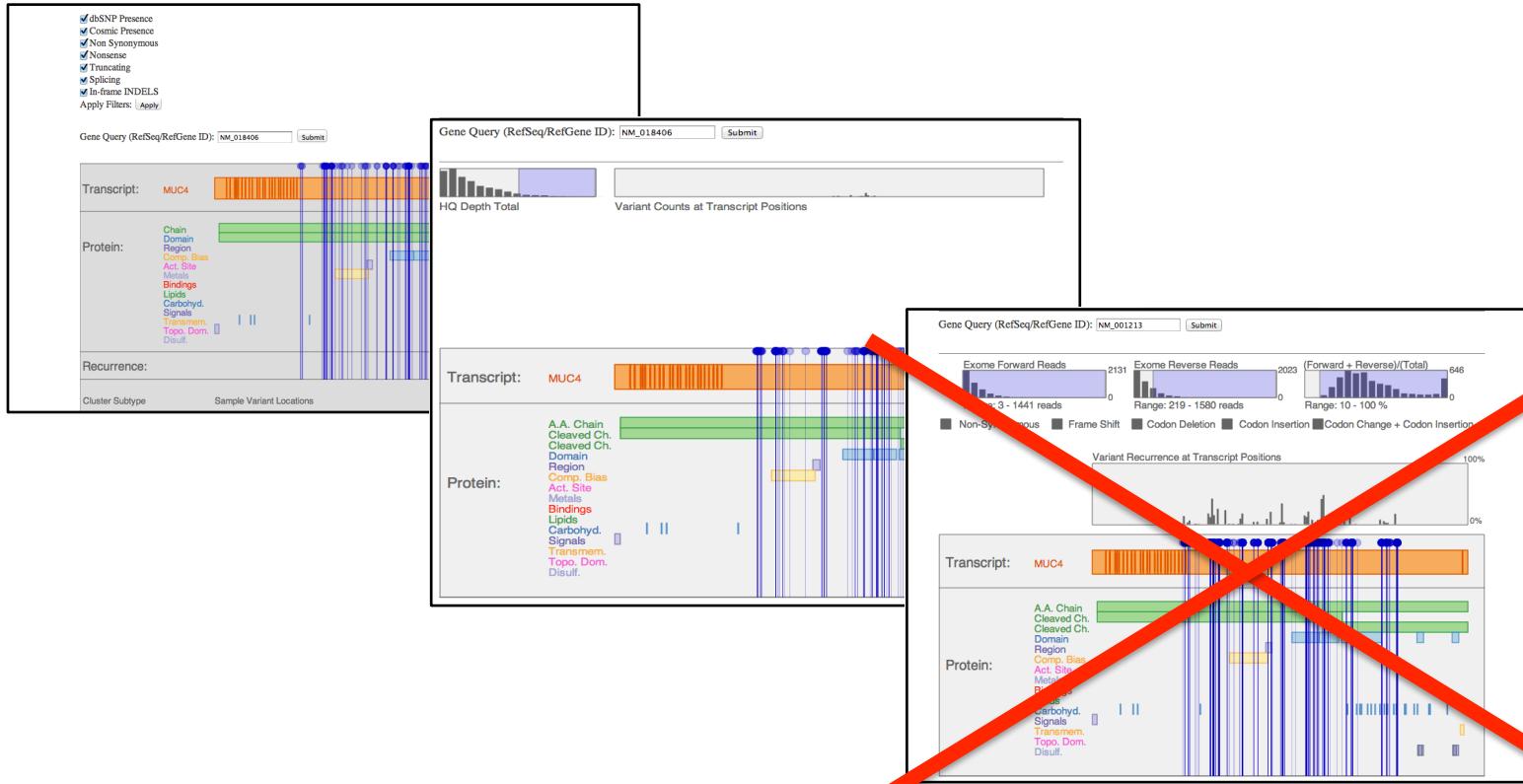


# Can identify features of **interest** in data



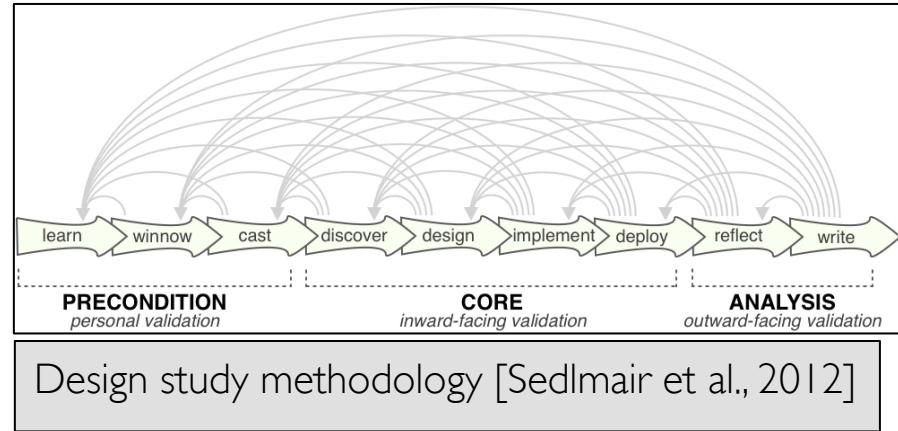
Emphasize some features over others

# Can identify design dead ends early



# Methods and durations

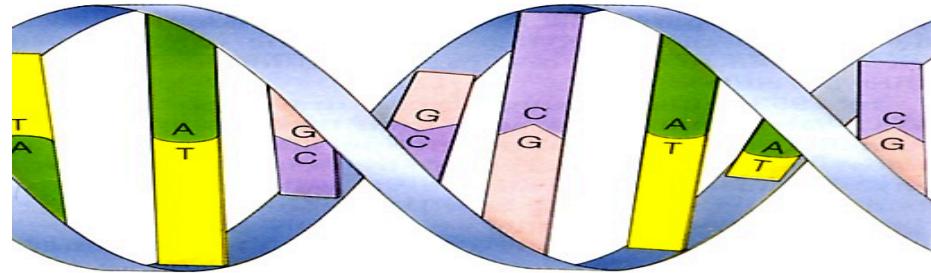
- Semi-structured interviews
  - 7 months
  - Once per week
  - One hour in duration
- Presented data sketches
  - 8 deployed over 5 months
  - Implemented with D3 toolkit  
[Bostock et al., InfoVis 2011 ]



# Problem characterization: Data

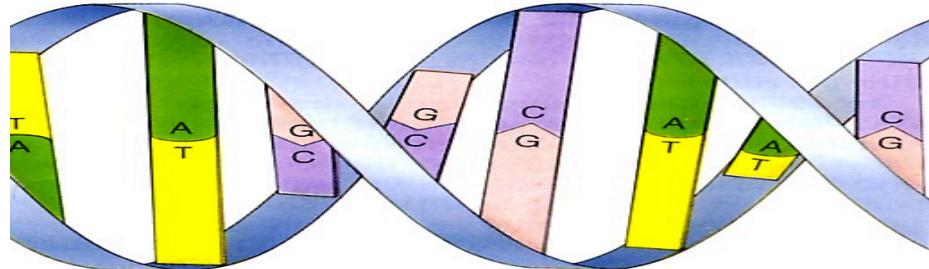
# The data

- 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



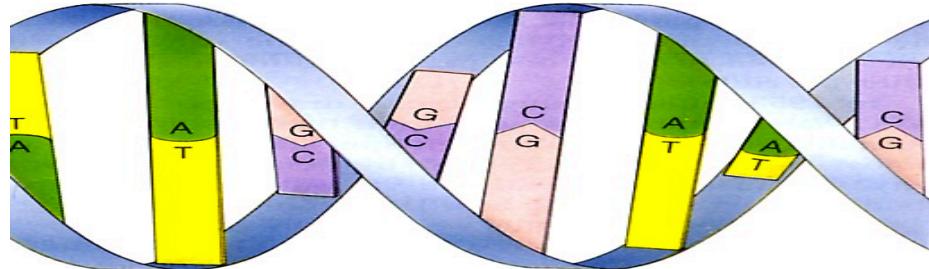
Reference Genome DNA: ATA TGA TCA ACA CTT

Sample 1 Genome DNA: ATA TGGT CA 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

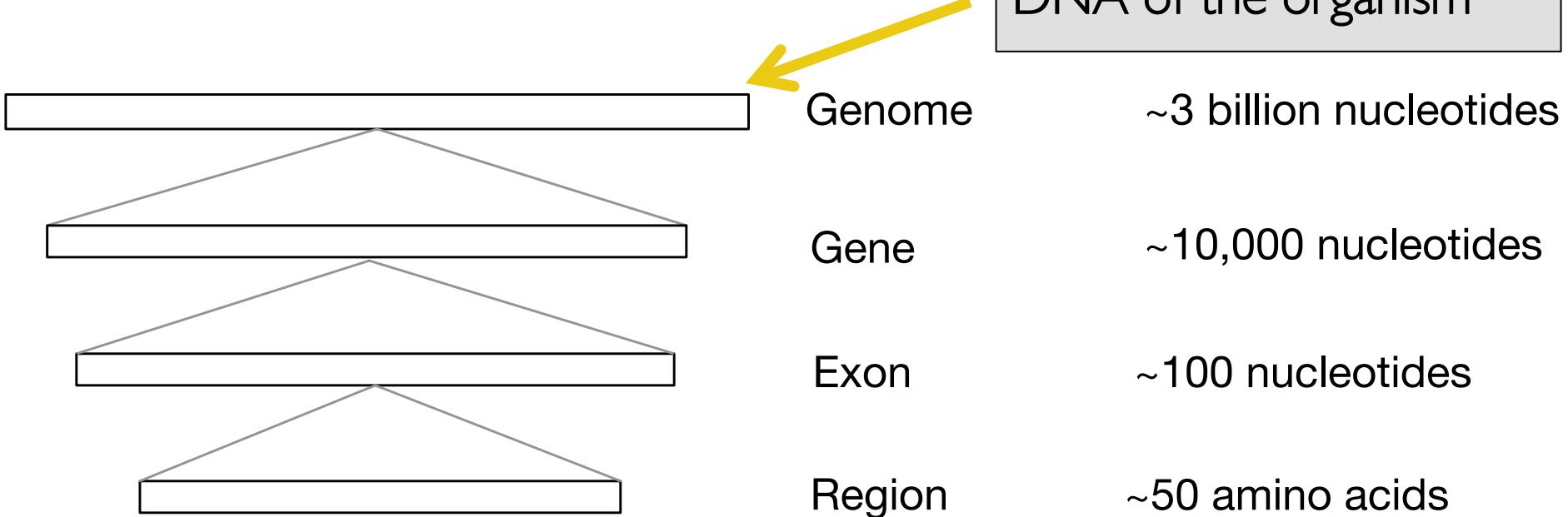
Sample 1 Genome DNA: ATA TGGT CA ATA CTT

Harmful?

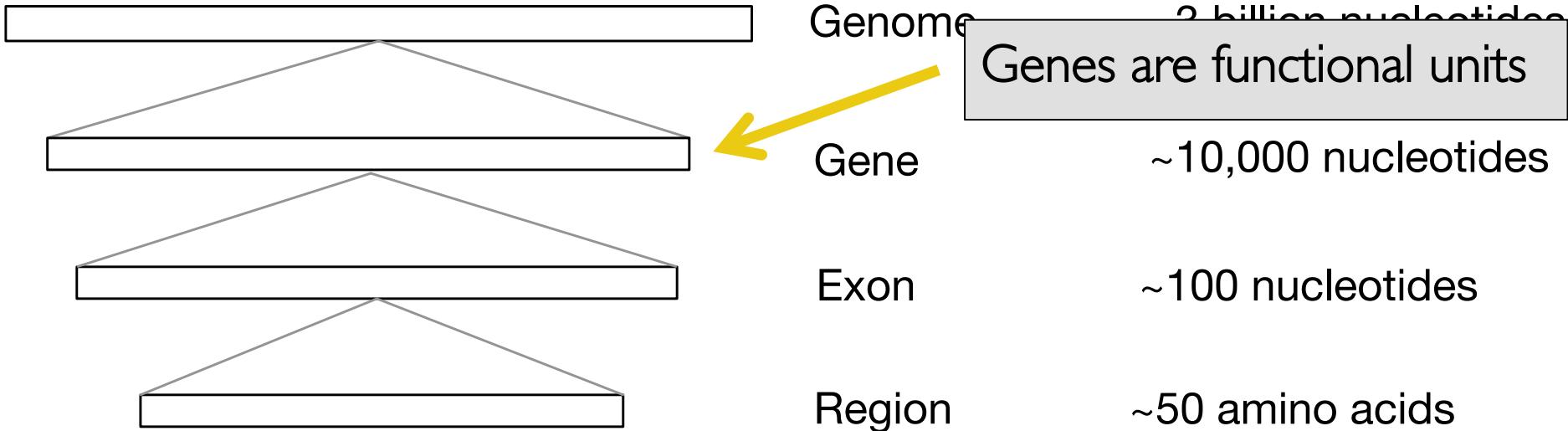
Sample 2 Genome DNA: ATA TGA TGA ACA CCT

Harmless?

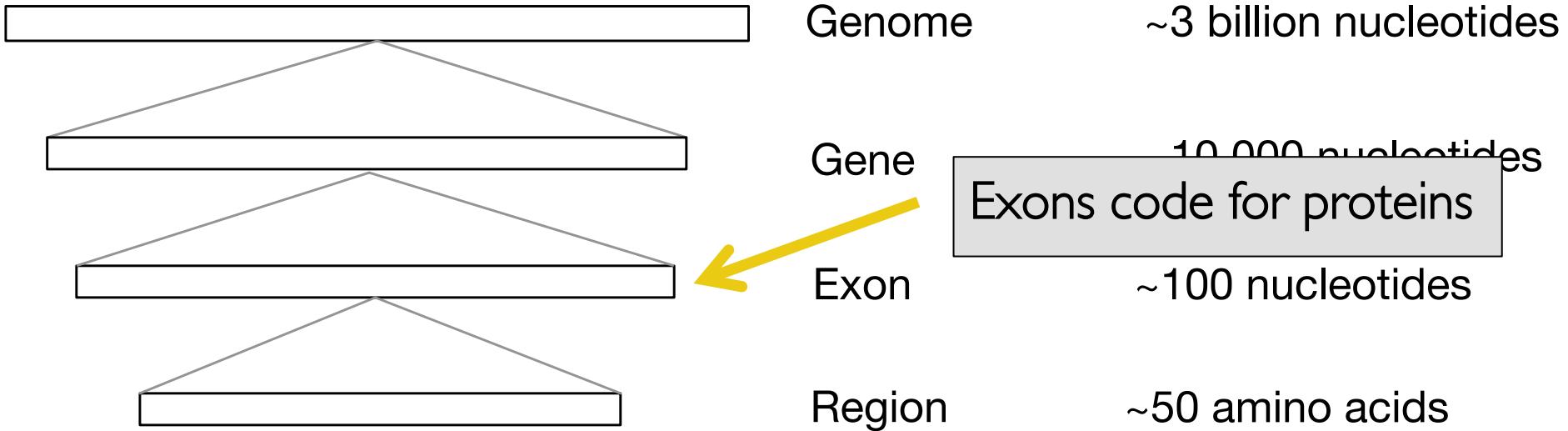
# Multi-scale data



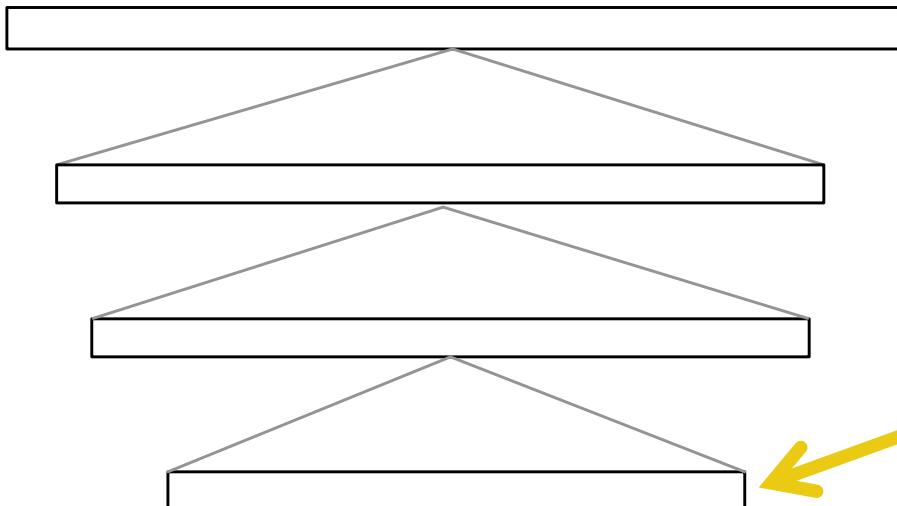
# Multi-scale data



# Multi-scale data



# Multi-scale data



Genome

~3 billion nucleotides

Gene

~10,000 nucleotides

Exon

Regions within proteins perform  
activities

Region

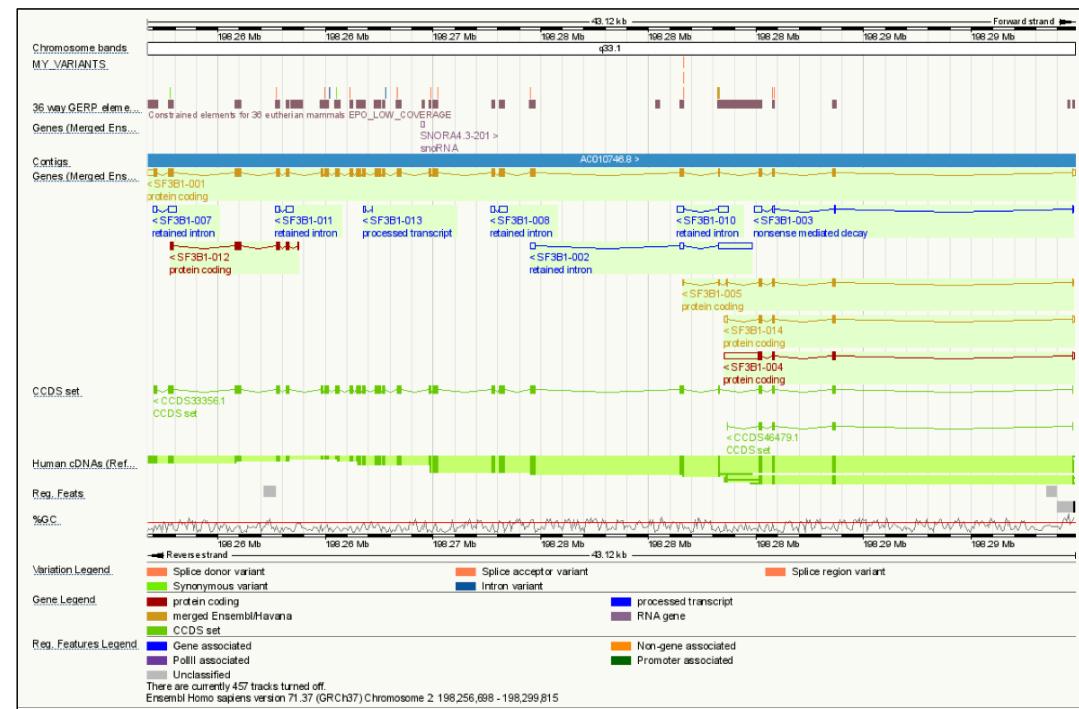
~50 amino acids

# Related work: Genome scale

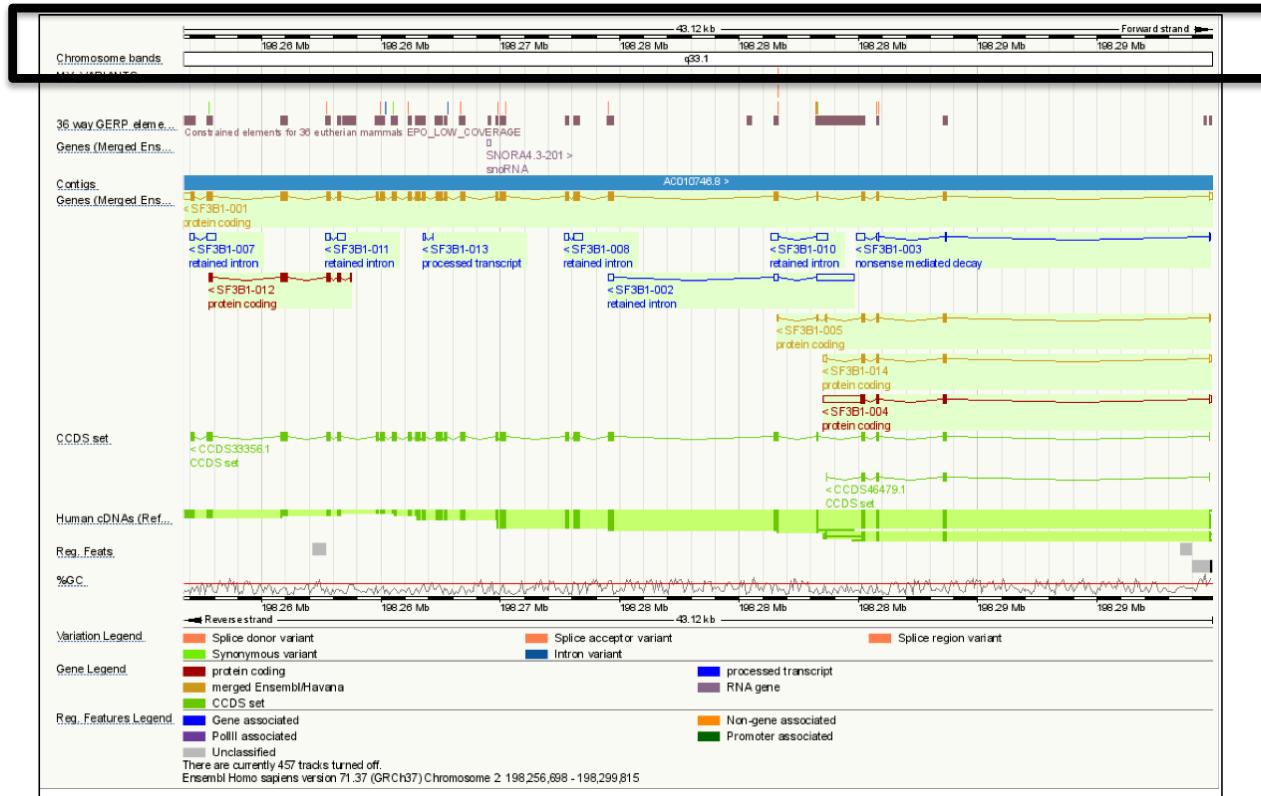
# Ensembl genome browser

- Explore genome and variant data

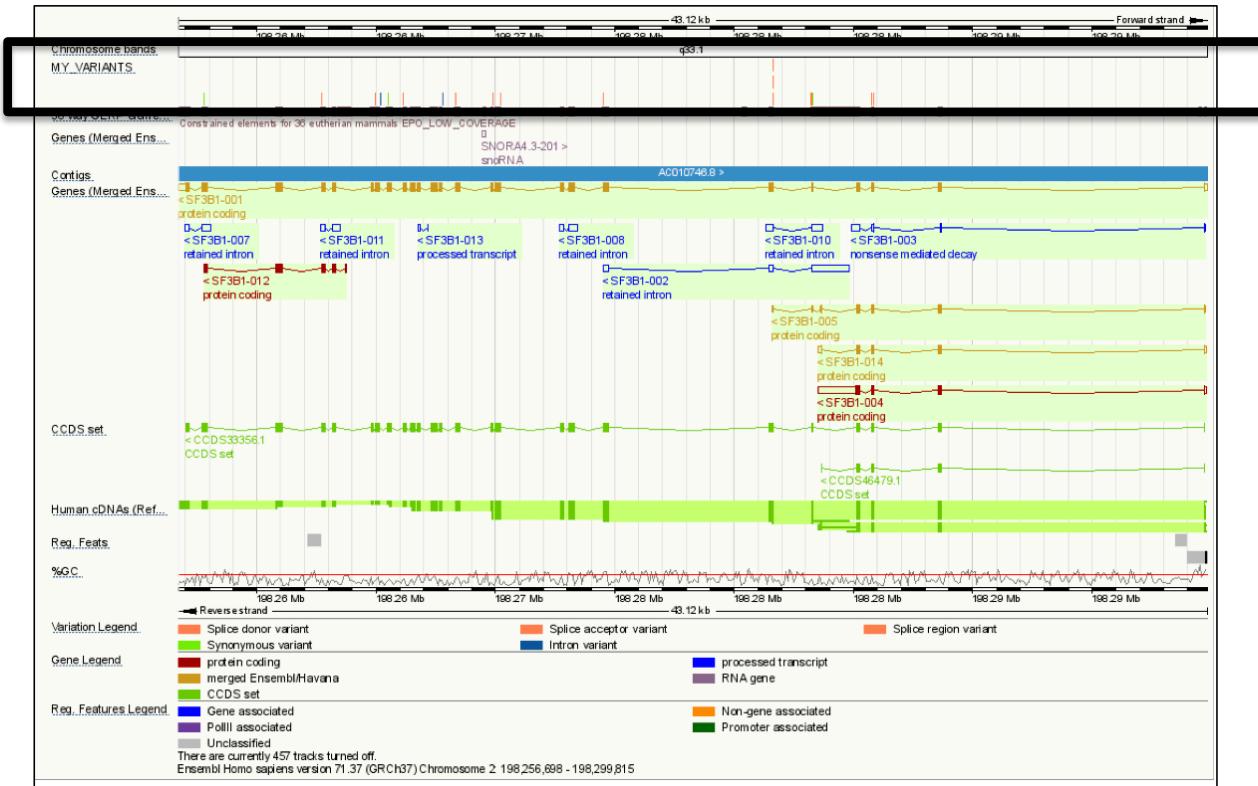
[Chen et al., BMC Genomics 2010]



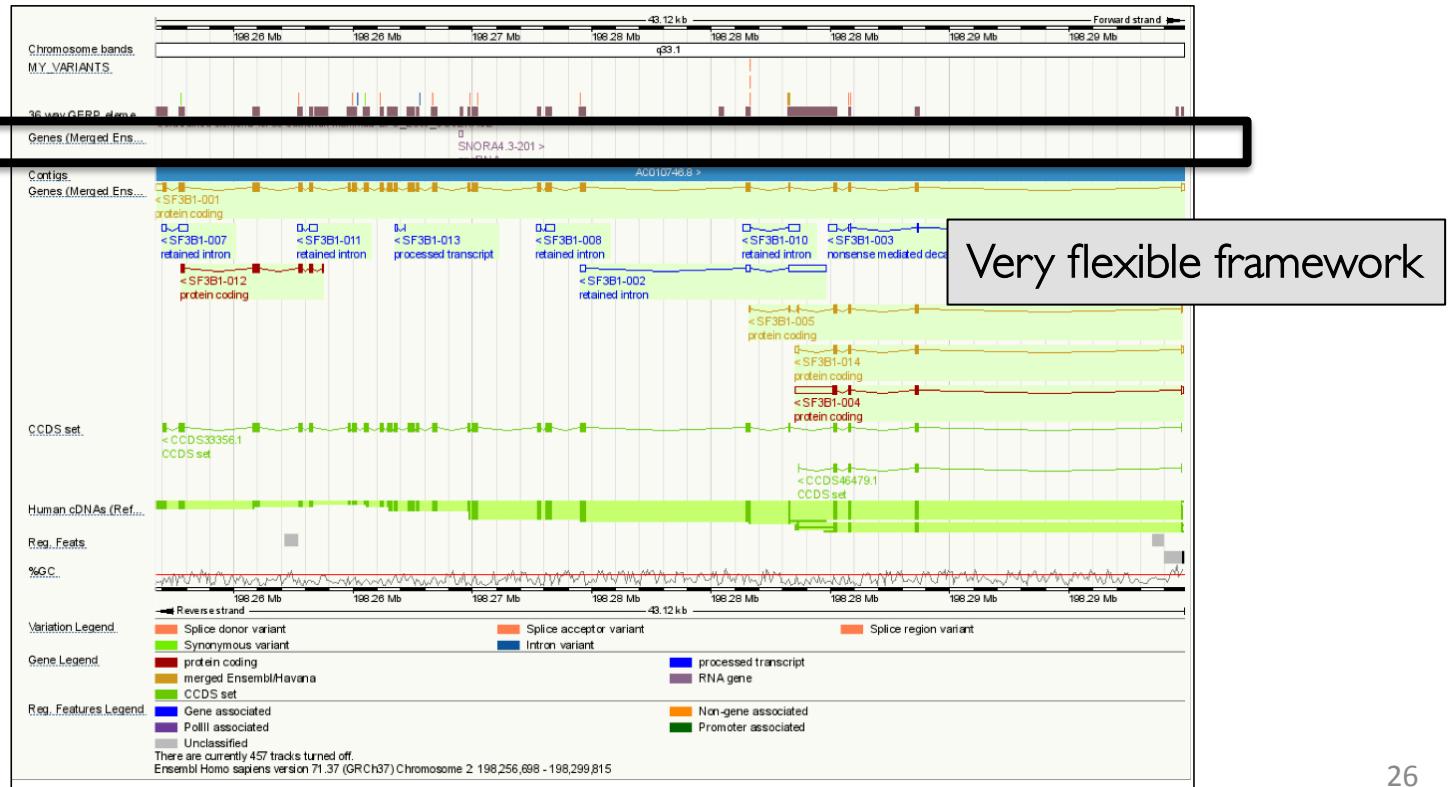
# Genome scale shown at the top



# User data is stacked in horizontal tracks

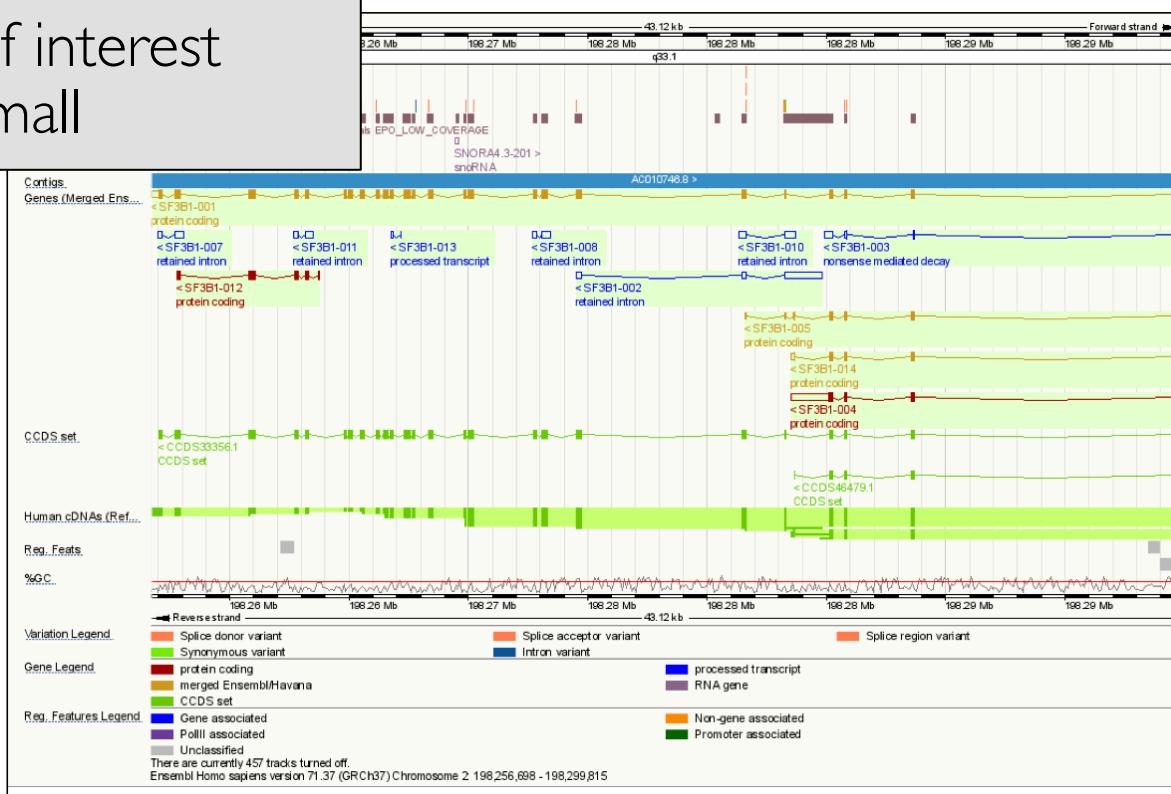


# User data is stacked in horizontal tracks



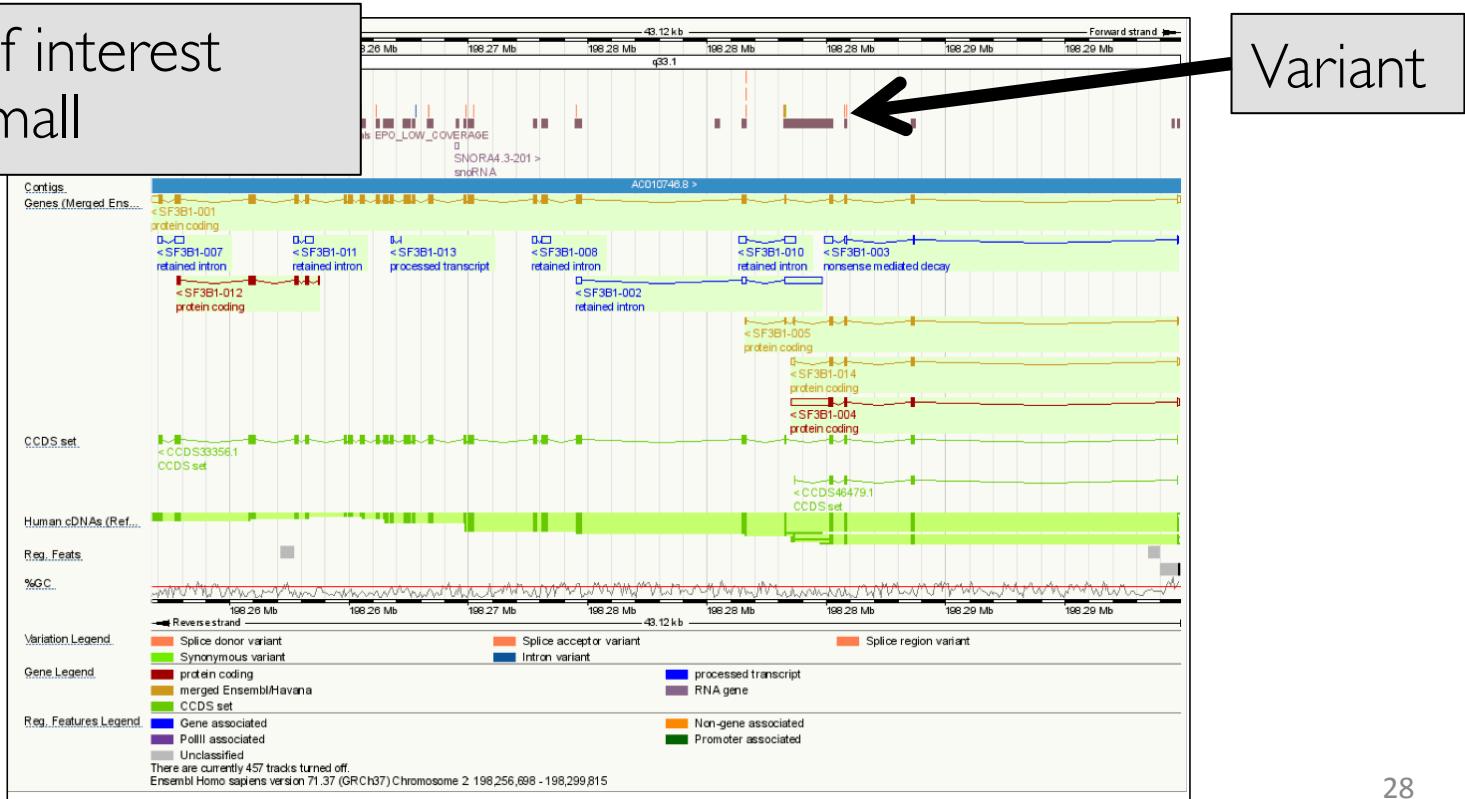
# Problem with the genome scale

- Features of interest become small



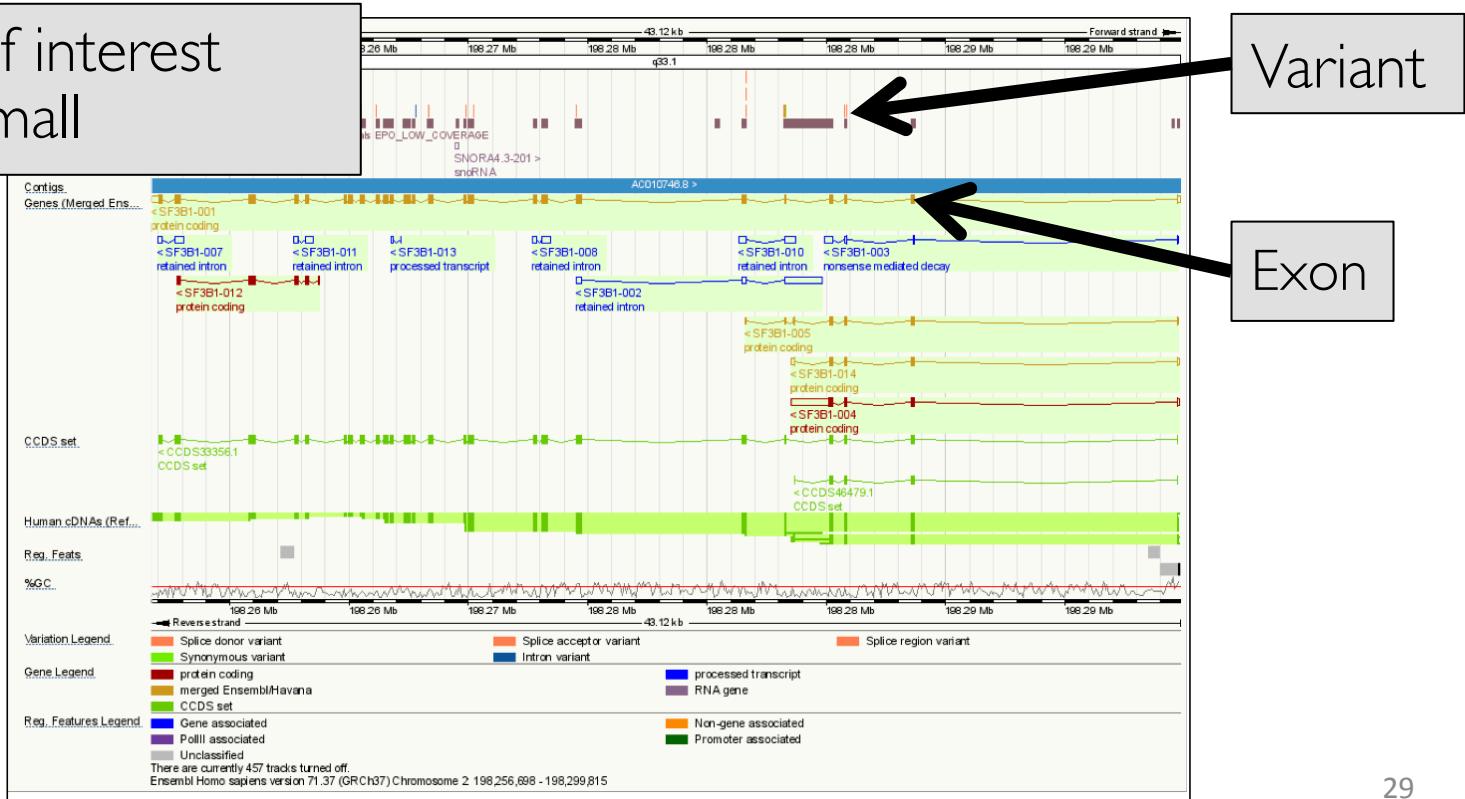
# Problem with the genome scale

- Features of interest become small

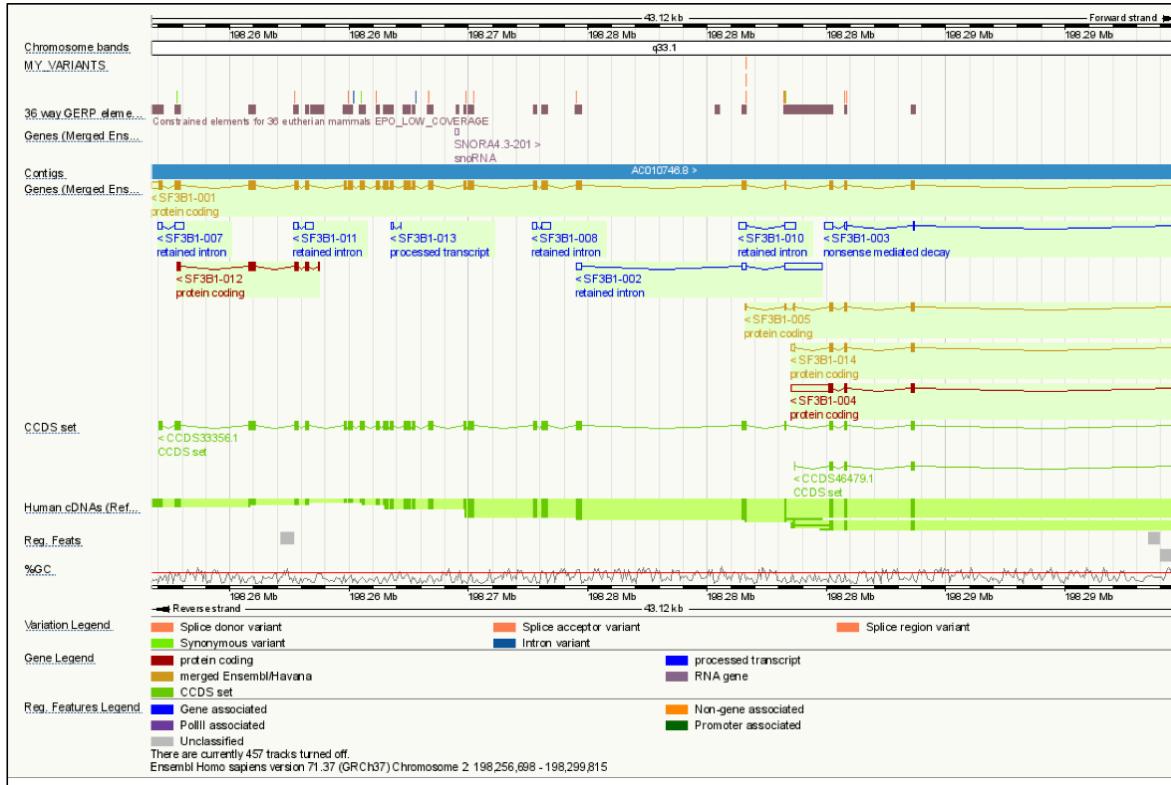


# Problem with the genome scale

- Features of interest become small

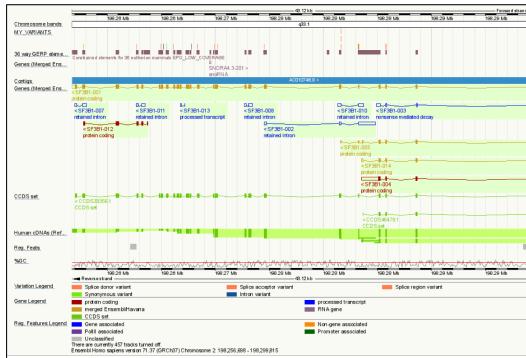


# Analysts must pan and zoom



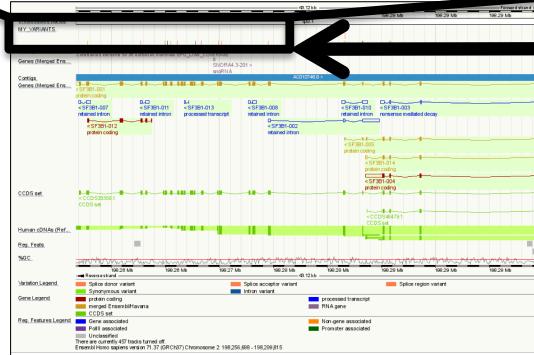
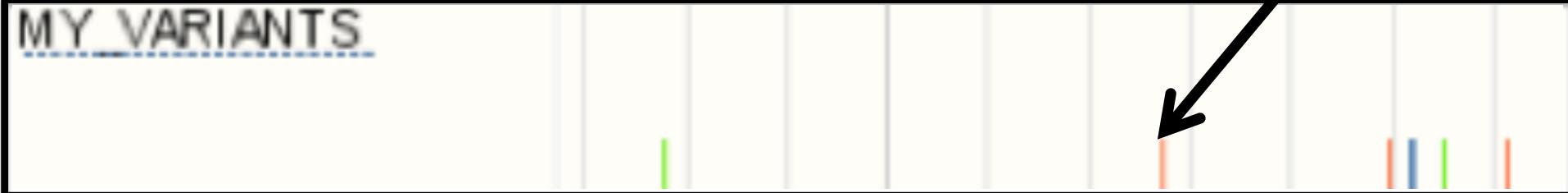
# Analysts must pan and zoom

- Heavy interaction costs with zooming



# Analysts must pan and zoom

- Heavy interaction costs with zooming



Raw variant data  
(What they looked at before)

# Raw data variant-centric

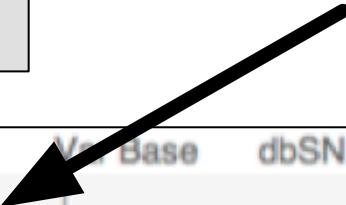
- Tabular data format

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	Alt 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	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

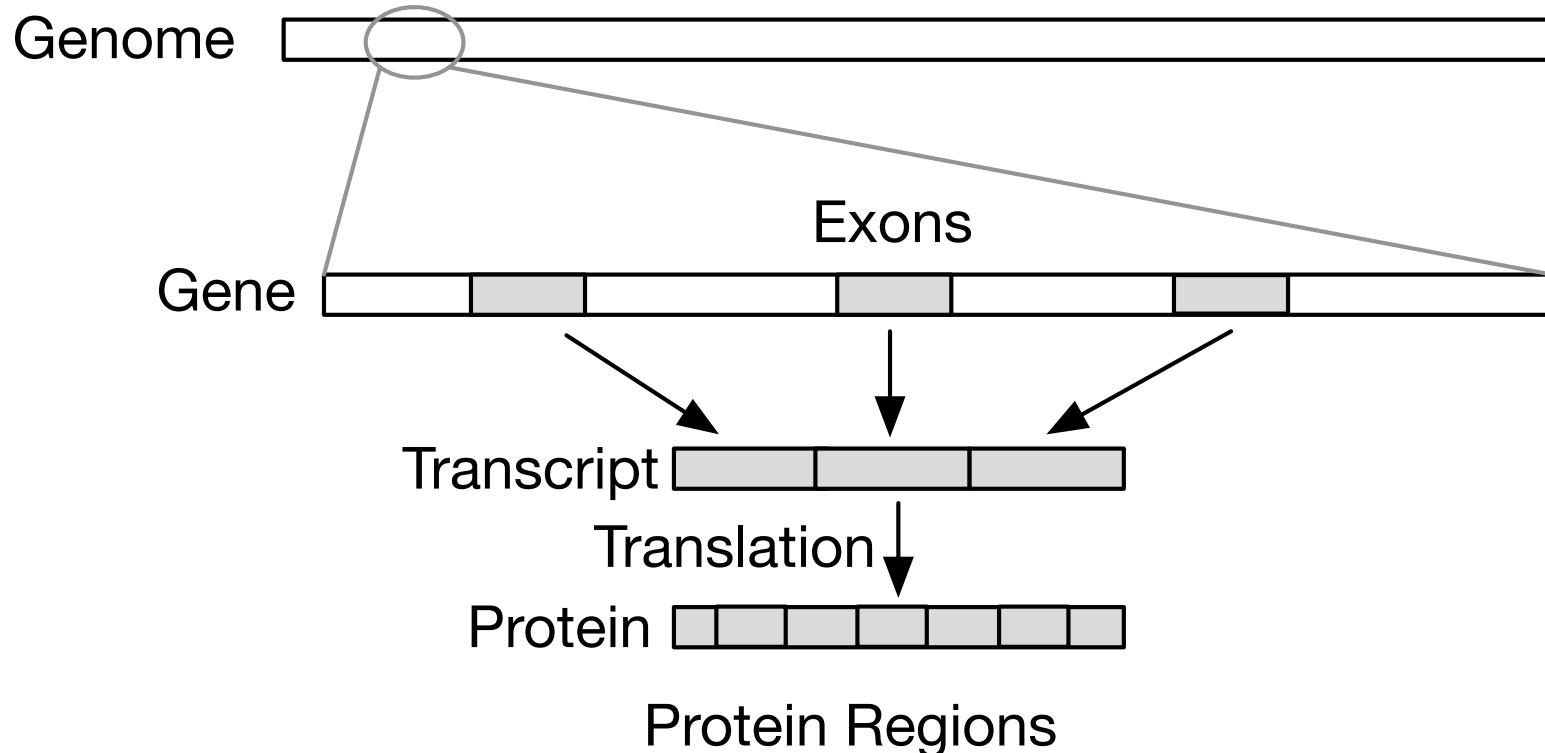
Patient ID	Chr. Coord.	Ref Base	Var Base	dbSNP129	dbSNP135	dbSNP137	COSMIC	A.A. Chng.
pid-anon	11288816	G	T	.	.	.	"13028,	G60V
pid-anon	11288816	G	T	.	.	.	"13012,	D61Y
pid-anon	11288810	G	T	.	rs121918	.	13014	A72G

- Difficult to reason about variants without their biological context

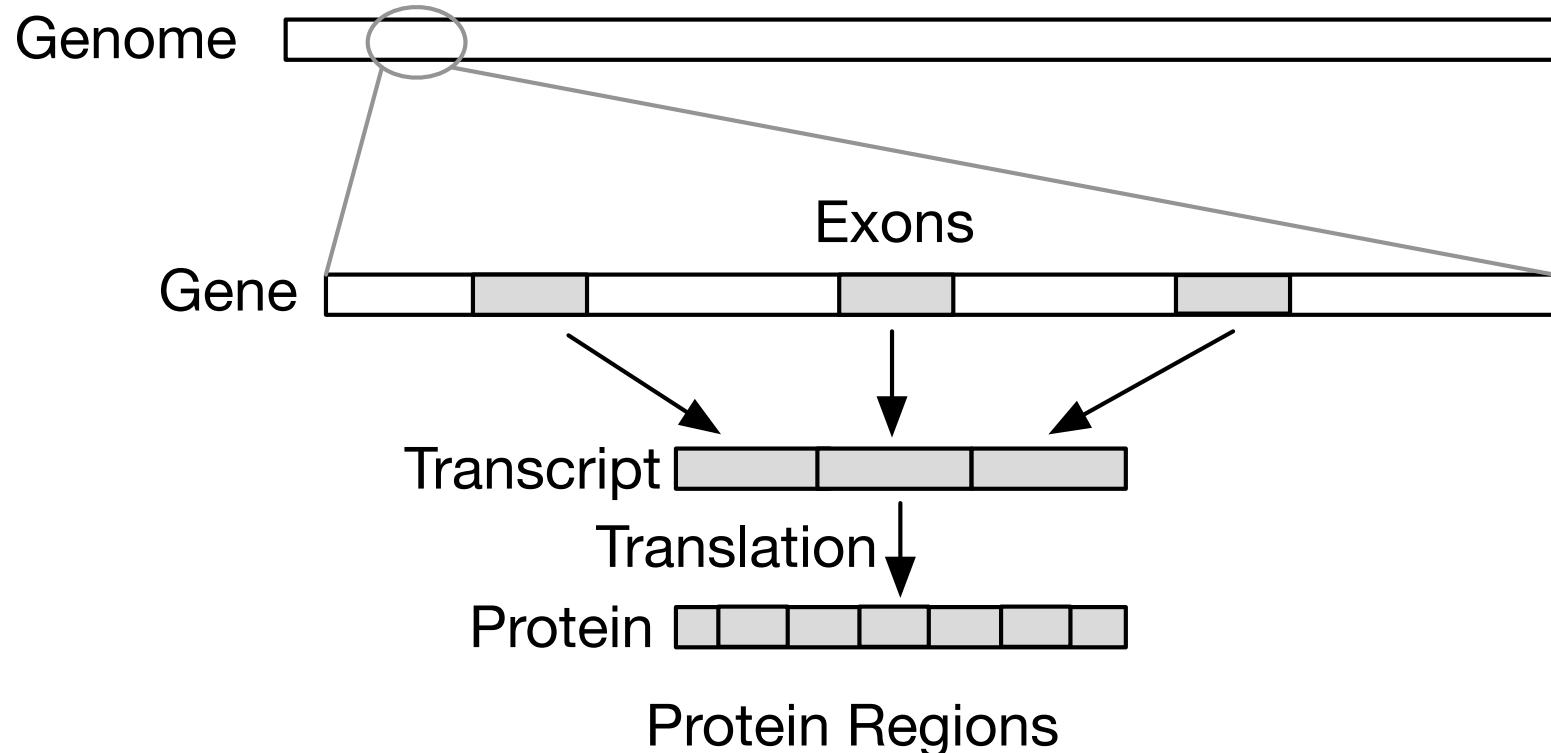
pid-anon	11288821	A	G	.	rs121918	"13017,	E76G
pid-anon	11288821	G	T	.	.	.	E76D
pid-anon	11292688	T	A	.	rs121918	"13020,	S502T
pid-anon	11292688	T	G	.	.	"13020,	S502A
pid-anon	11292688	C	T	.	.	13023	S502L

Multiple biological levels/scales  
What to show?

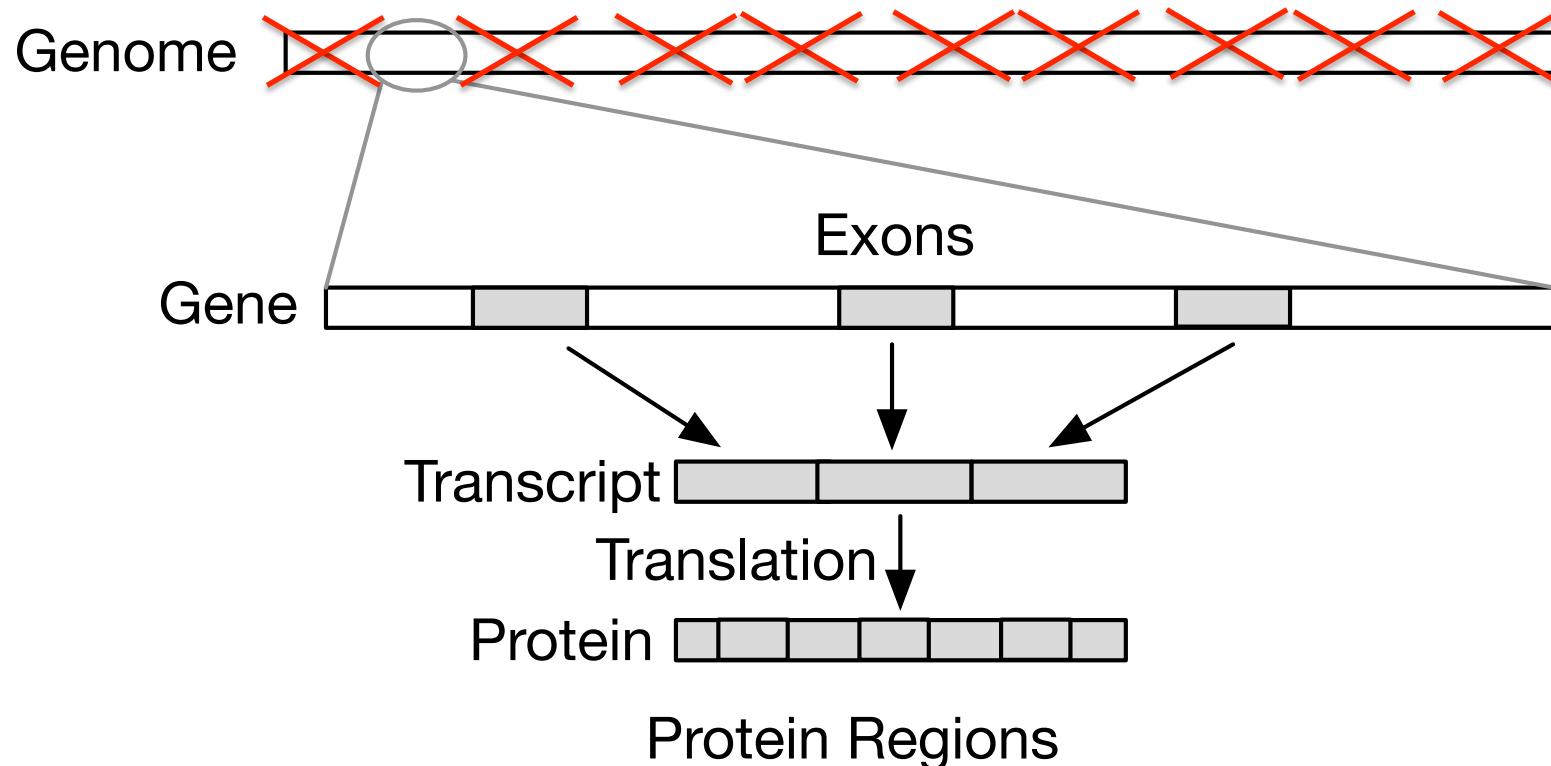
# Many biological levels and scales



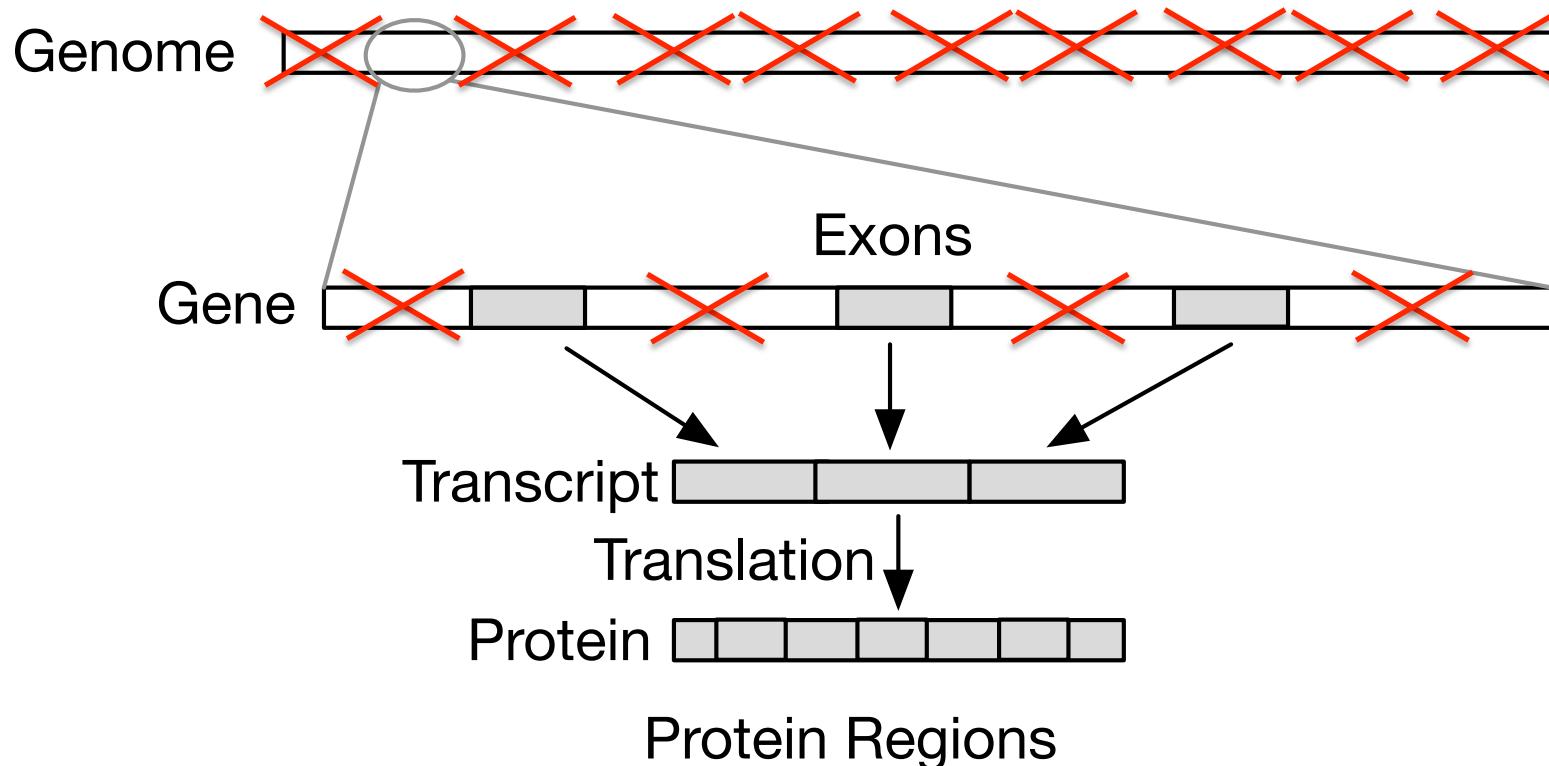
# Only some levels and scales are beneficial for variant analysis



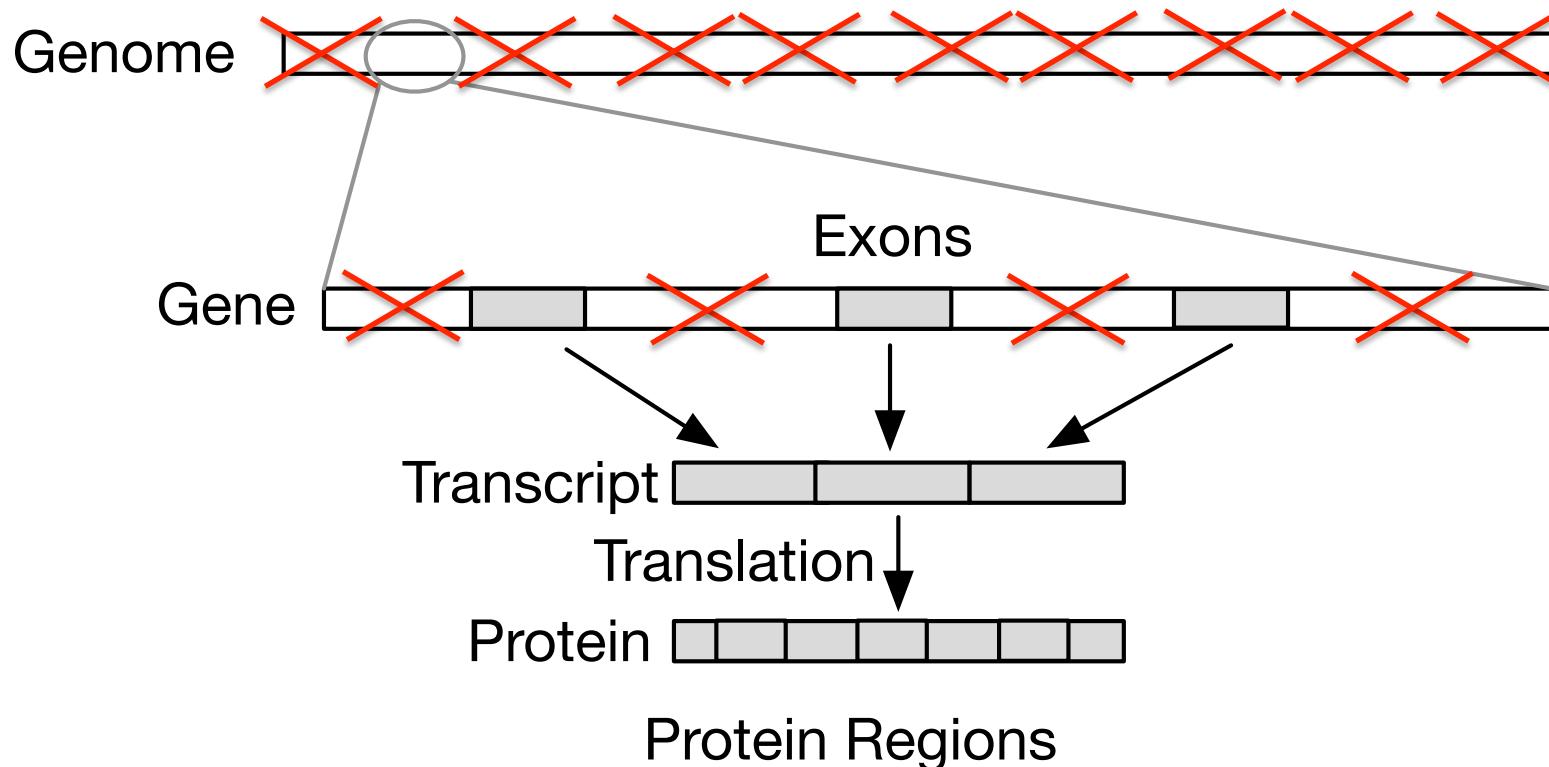
# Filter out whole genome; keep genes



# Filter out non-exon regions



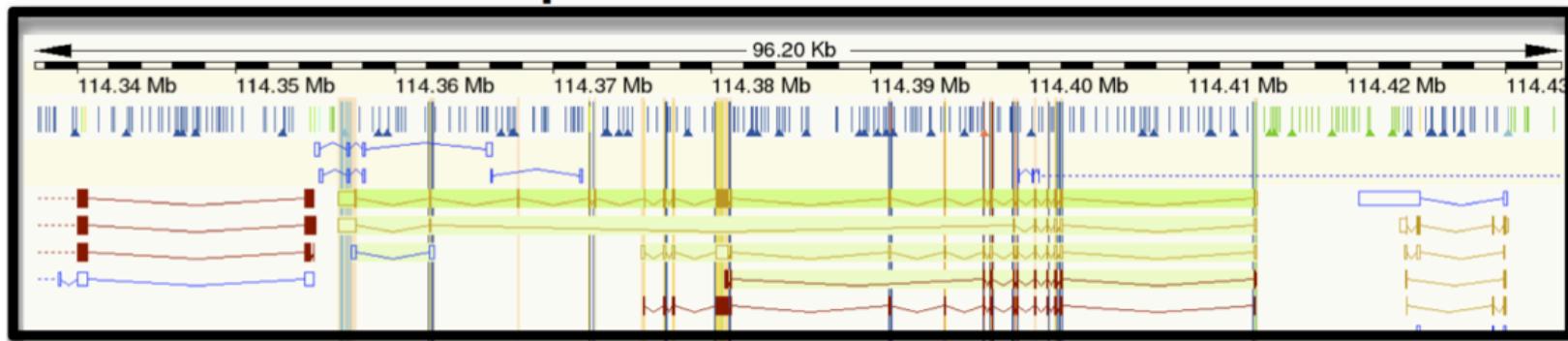
# Left with a **filtered scope**



# Related work: Filtered scope

# The Ensembl Variation Image

## Genome Coordinate Space



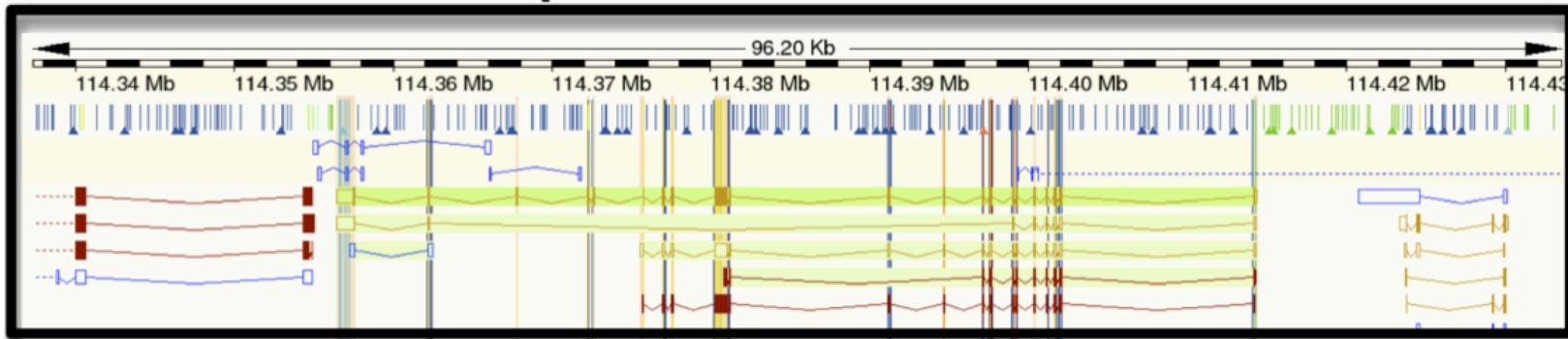
## Emphasized Exons



# First filtering step: per-gene view

One gene shown

Genome Coordinate Space

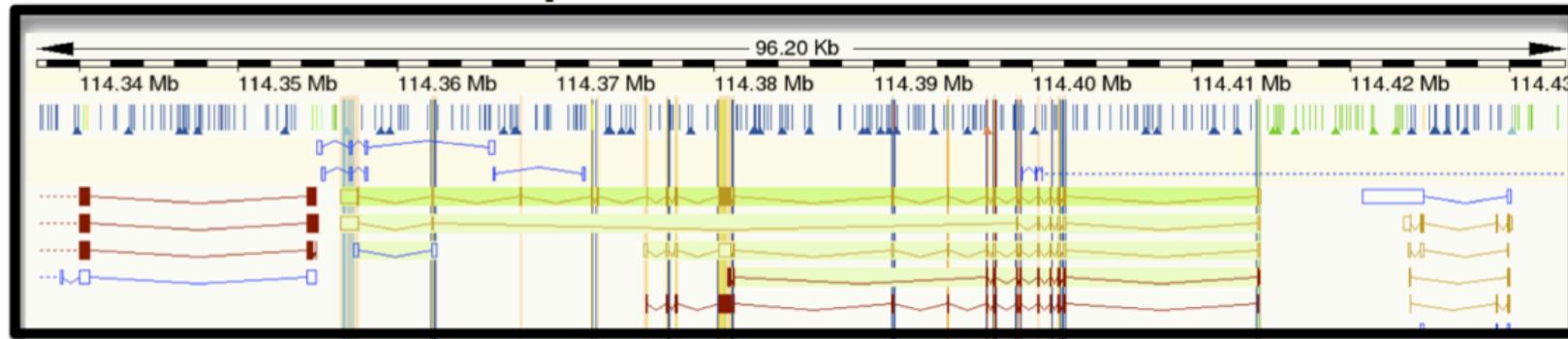


Emphasized Exons

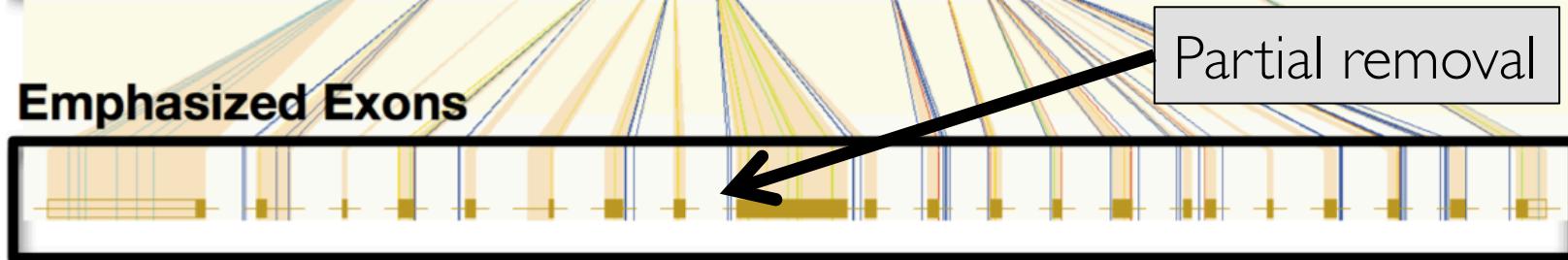


# Second filtering step: partial removal inter-exon regions

Genome Coordinate Space

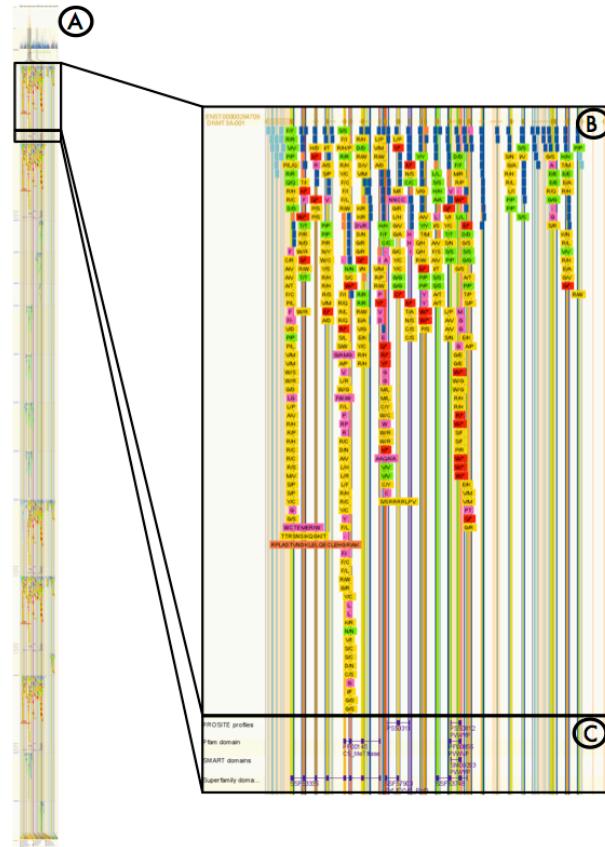


Emphasized Exons



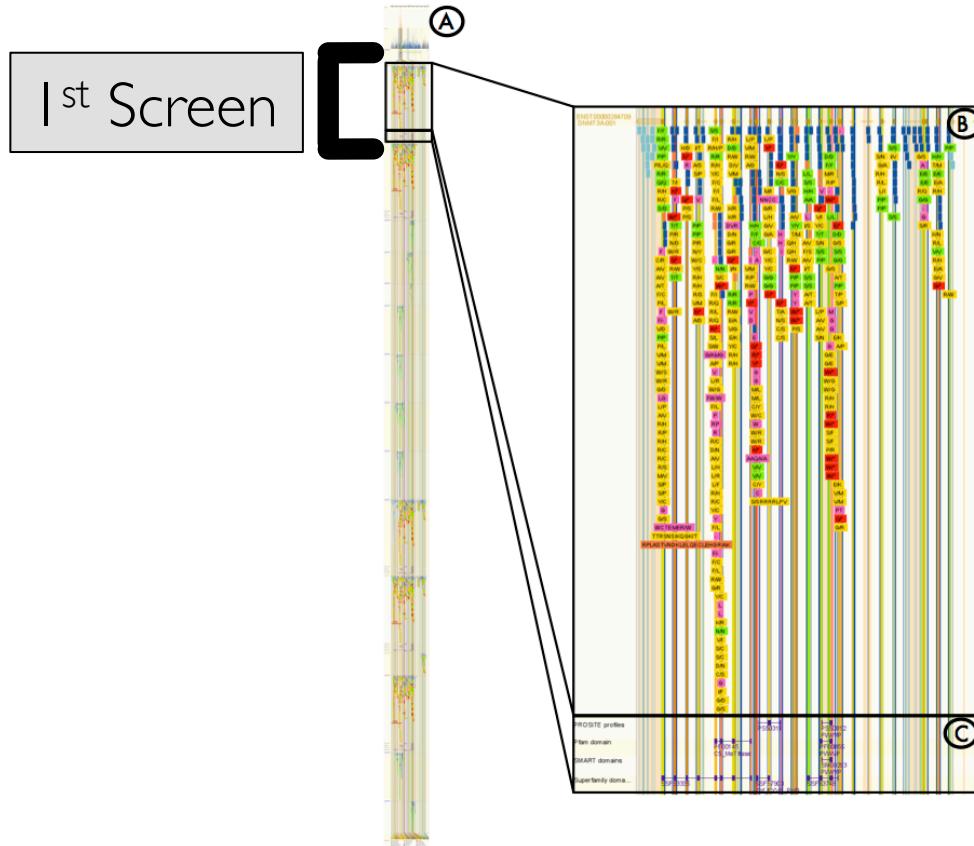
[Chen et al., BMC Genomics 2010]

# Problem: Extends multiple screens



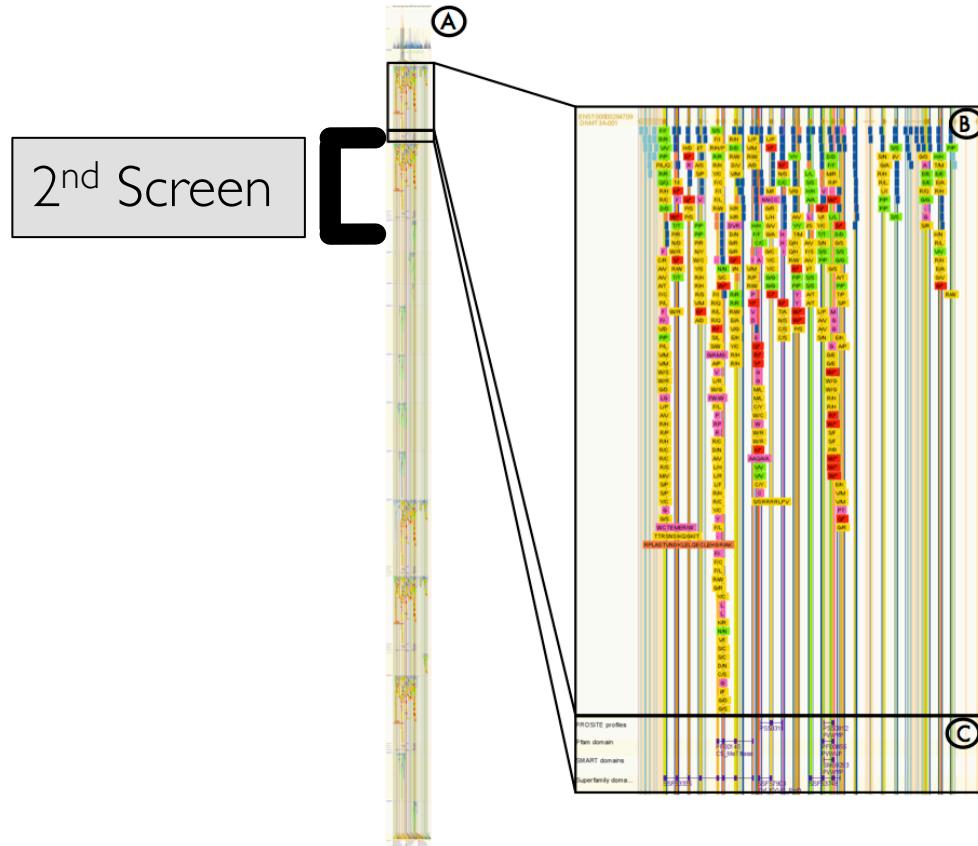
[Chen et al., BMC Genomics 2010]

# Problem: Extends multiple screens



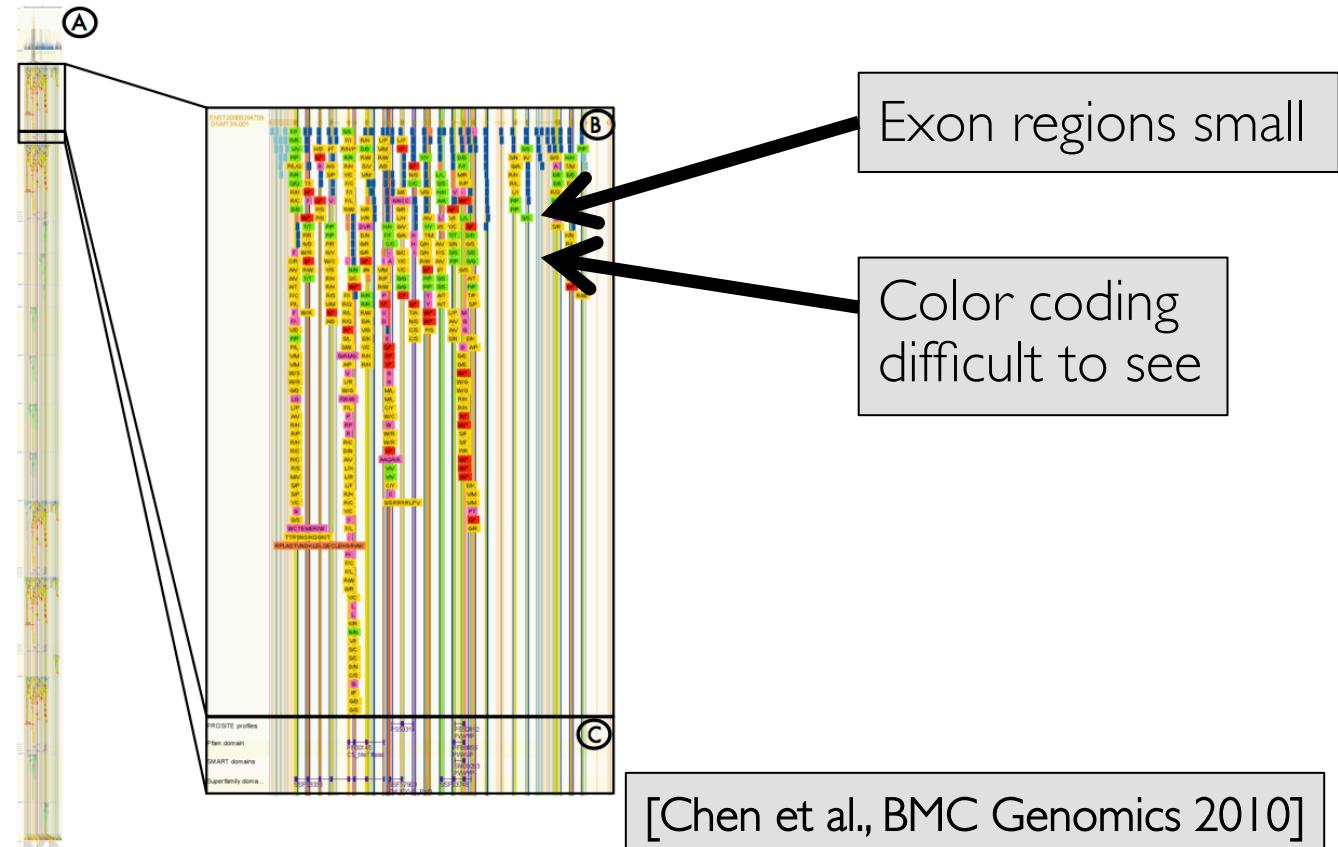
[Chen et al., BMC Genomics 2010]

# Problem: Extends multiple screens



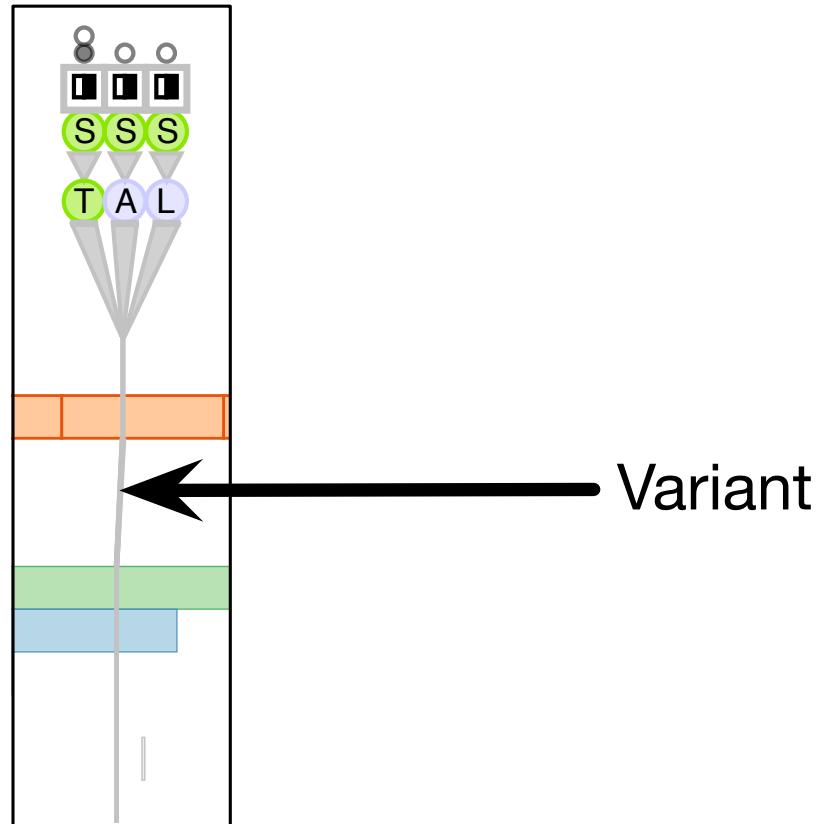
[Chen et al., BMC Genomics 2010]

# Problem: Features of interest small



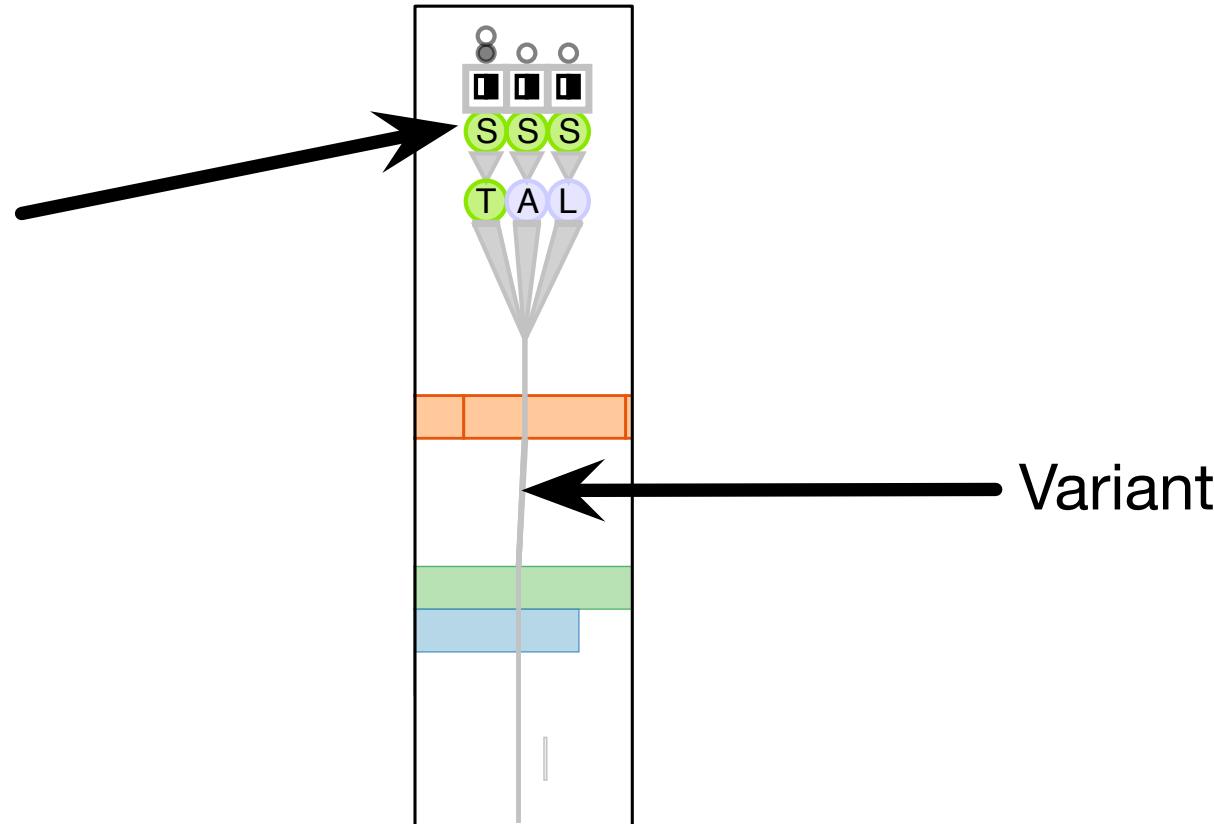
Our Goal:  
Show attributes necessary for **variant analysis**

# Use information-dense visual encoding



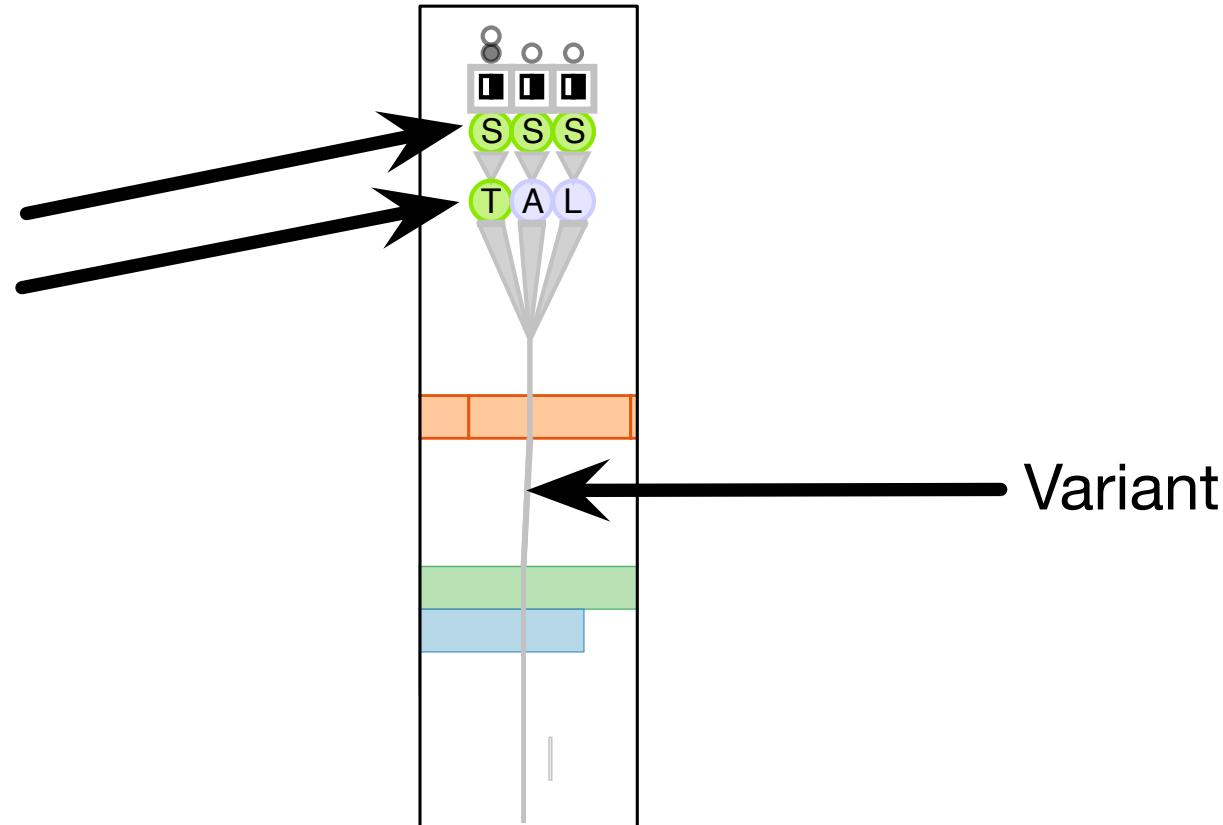
# Use information-dense visual encoding

Reference AA



# Use information-dense visual encoding

Reference AA  
Variant AA

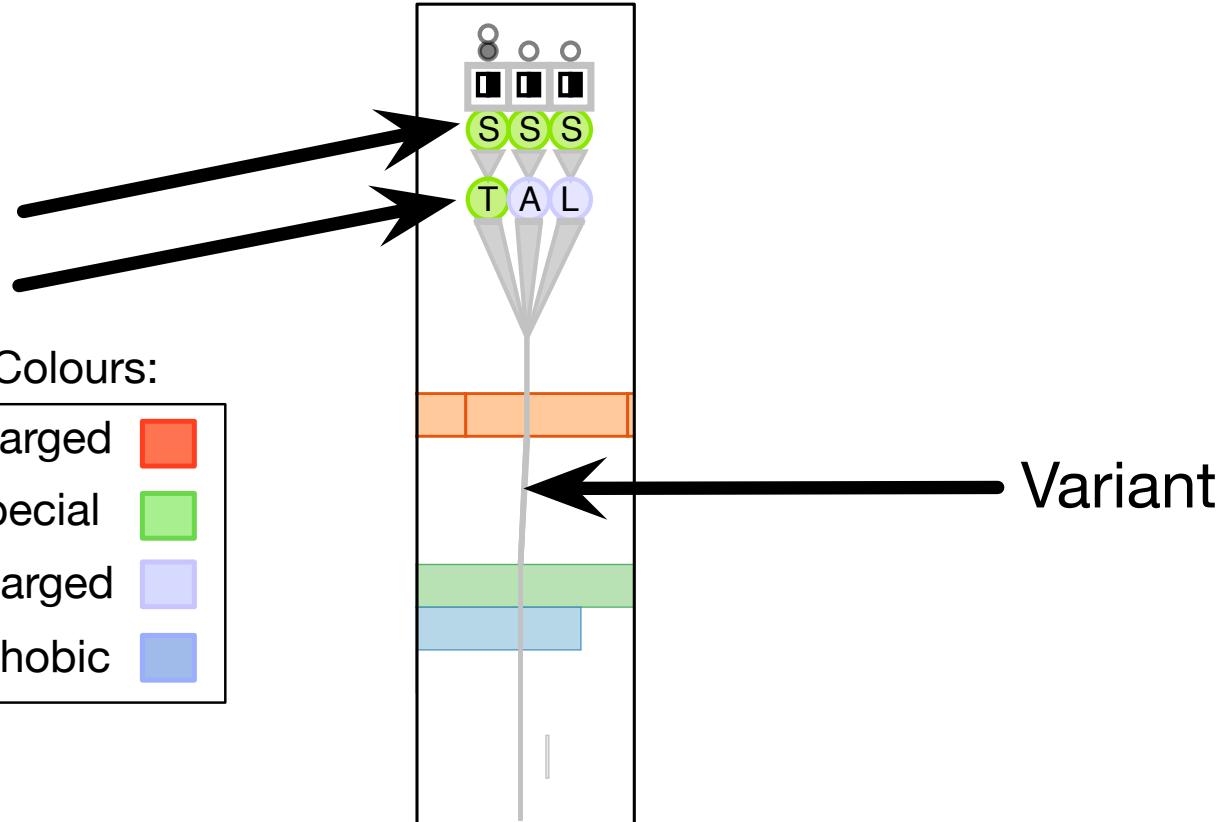


# Use information-dense visual encoding

Reference AA  
Variant AA

AA Chemical Class Colours:

Charged	
Special	
Uncharged	
Hydrophobic	

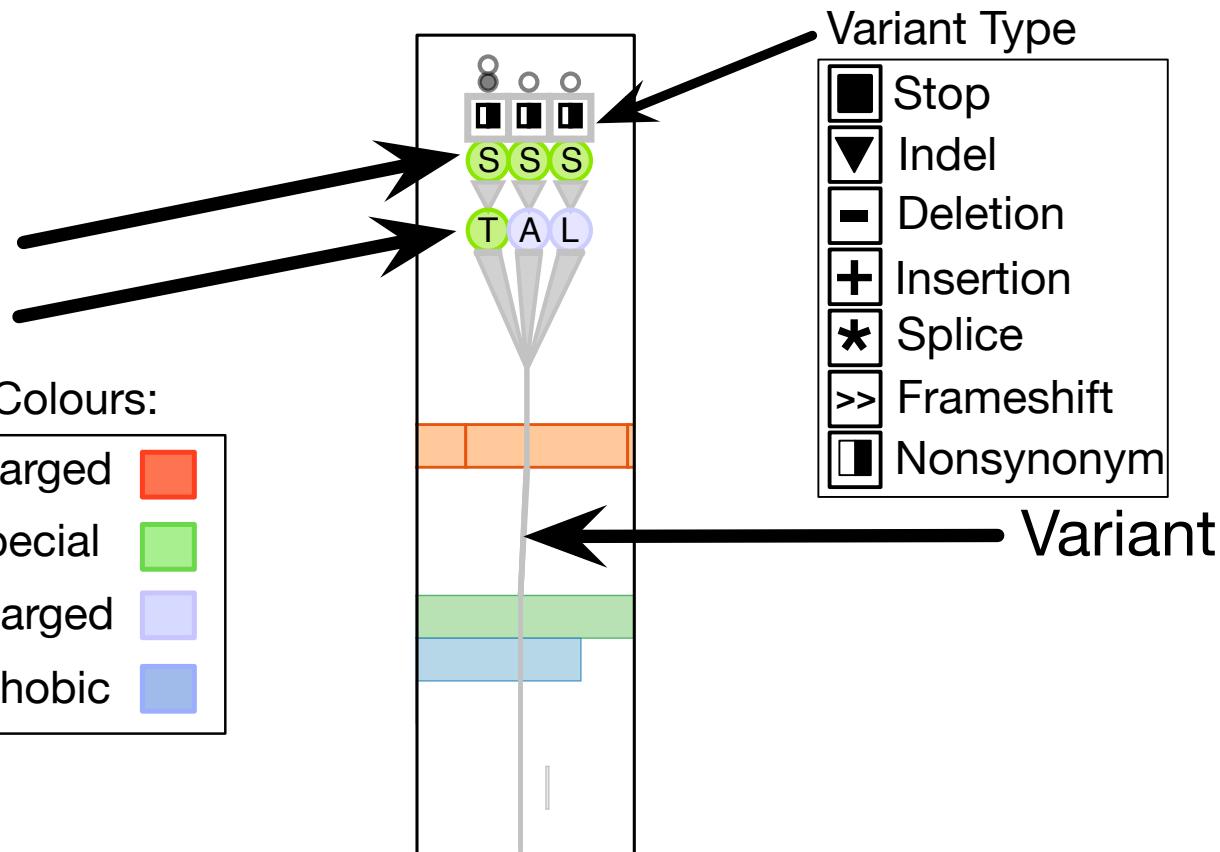


# Use information-dense visual encoding

Reference AA  
Variant AA

AA Chemical Class Colours:

Charged	
Special	
Uncharged	
Hydrophobic	



# Use information-dense visual encoding

Known Database

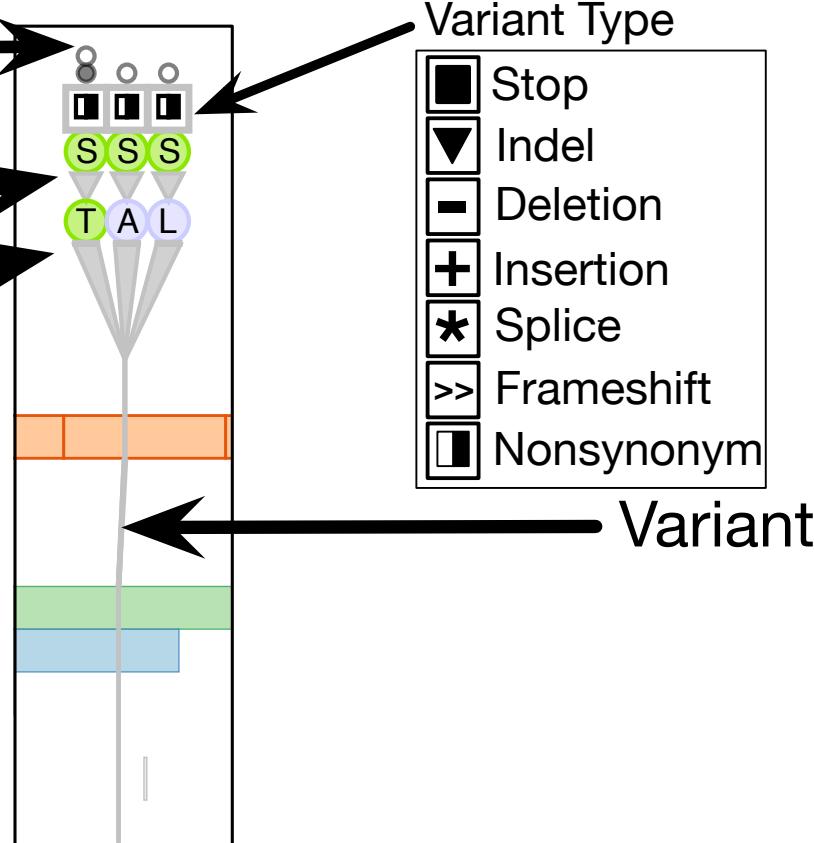
- Known Harmless
- Known Cancer

Reference AA

Variant AA

AA Chemical Class Colours:

Charged	
Special	
Uncharged	
Hydrophobic	



# Use information-dense visual encoding

Known Database

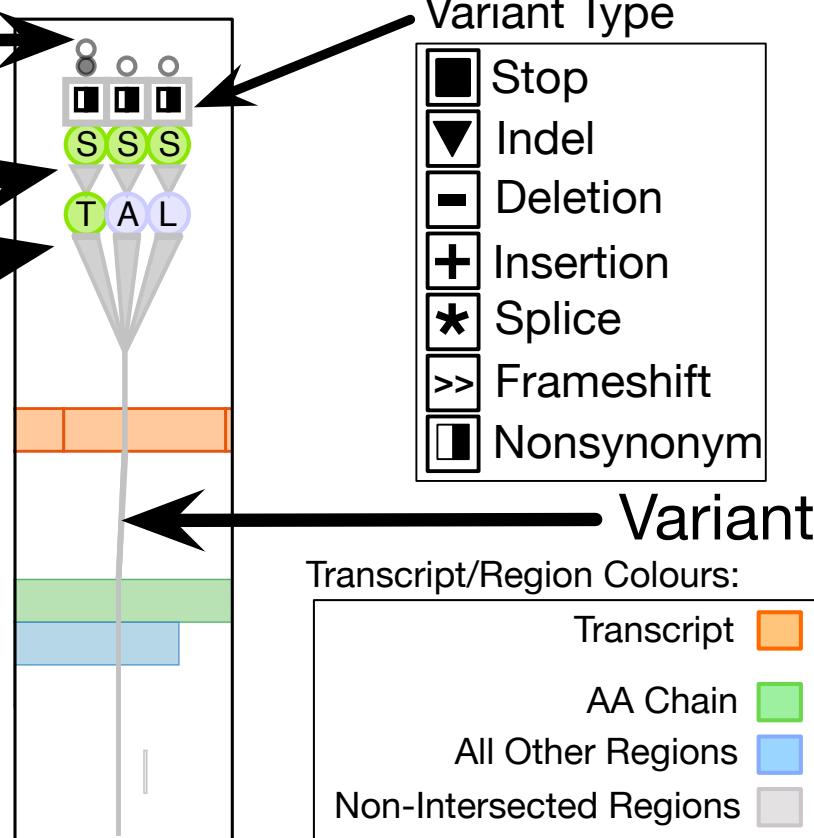
- Known Harmless
- Known Cancer

Reference AA

Variant AA

AA Chemical Class Colours:

Charged	■
Special	■
Uncharged	■
Hydrophobic	■



# The tool: Variant View

# Variant View

Gene Search:

Alternative Transcripts: gene-anon (trans-anon)

**A**

**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

**B**

**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

**C**

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

Information-dense single gene view

Gene Search:  Submit

Alternative Transcripts: gene-anon (trans-anon)

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

(A) (B) (C)

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

Information-dense single gene view

Gene Search:  Submit

Alternative Transcripts: gene-anon (trans-anon)

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
pid-anon	11288816	G	T	.	.	.	"13028, G60V D61Y	gene-anon
pid-anon	11288816	G	T	.	.	.	"13012, A72S	trans-anon
pid-anon	11288819	G	T	.	.	.	"13035, A72V	gene-anon
pid-anon	11288821	G	C	.	.	.	"13016, E76Q	trans-anon
pid-anon	11288821	A	G	.	rs121918	.	"13017, E76G	gene-anon
pid-anon	11288821	G	T	.	.	.	."	trans-anon
pid-anon	11292688	T	A	.	rs121918	.	"13020, S502T	gene-anon
pid-anon	11292688	T	G	.	.	.	"13020, S502A	trans-anon
pid-anon	11292688	C	T	.	.	.	13023, S502L	gene-anon
								trans-anon

Sort By Gene:  
Alpha Cluster Score Variant Count

(A) (C)

No need for pan and zoom

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\_0013270)

STRE1 (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: gene-anon (trans-anon)

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 ID
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

Panel A: Alpha Filter Score   Variant Count

Panel C: Gene List

- 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: gene-anon (trans-anon)

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

(A) Alpha Cluster Score Variant Count

(B)

	135	dbSNP137	COSMIC	A.A. Chng.	Gene	Ref
pid-anon	11288819	C	T	.	"13028,	G60V
pid-anon	11288821	G	C	.	"13012,	D61Y
pid-anon	11288821	A	G	.	"13017,	A72S
pid-anon	11288821	G	T	.	"13035,	A72V
pid-anon	11292688	T	A	.	"13016,	E76Q
pid-anon	11292688	T	G	.	"13017,	E76G
pid-anon	11292688	C	T	.	"13020,	E76D
					S502T	gene-anon
					S502A	trans-anon
					S502L	gene-anon
						trans-anon

(C)

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)

Control what shows up here

65

# Variant View

Gene Search:  Submit

Alternative Transcripts: gene-anon (trans-anon)

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		
pid-anon	11288819	G	T	.	rs121918	.	"13028,	A72S		
pid-anon	11288819	C	T	.	.	.	"13035,			
pid-anon	11288821	G	C	.	.	.	"13016,	E76Q		
pid-anon	11288821	A	G	.	rs121918	.	"13017,	E76G		
pid-anon	11288821	G	T	.	.	.	"13020,	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

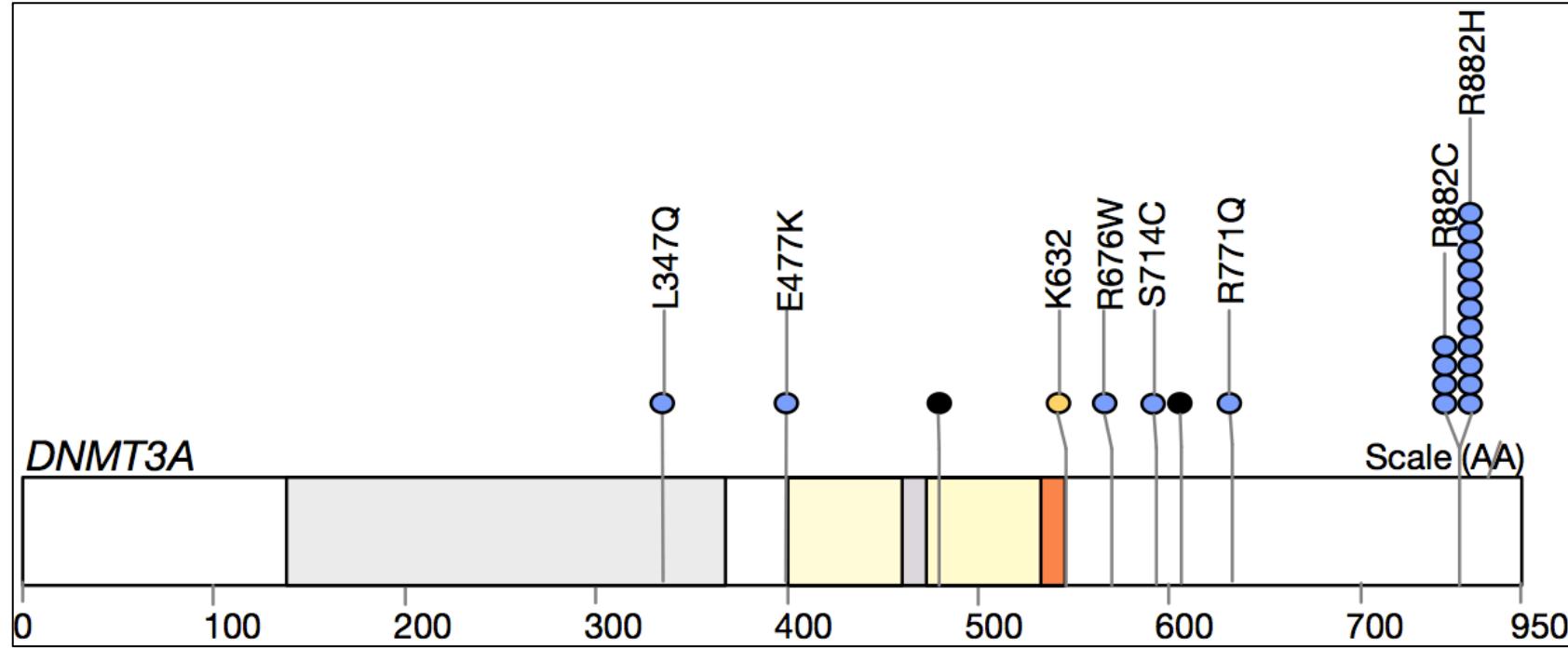
(A) (B) (C)

Peripheral supporting data

66

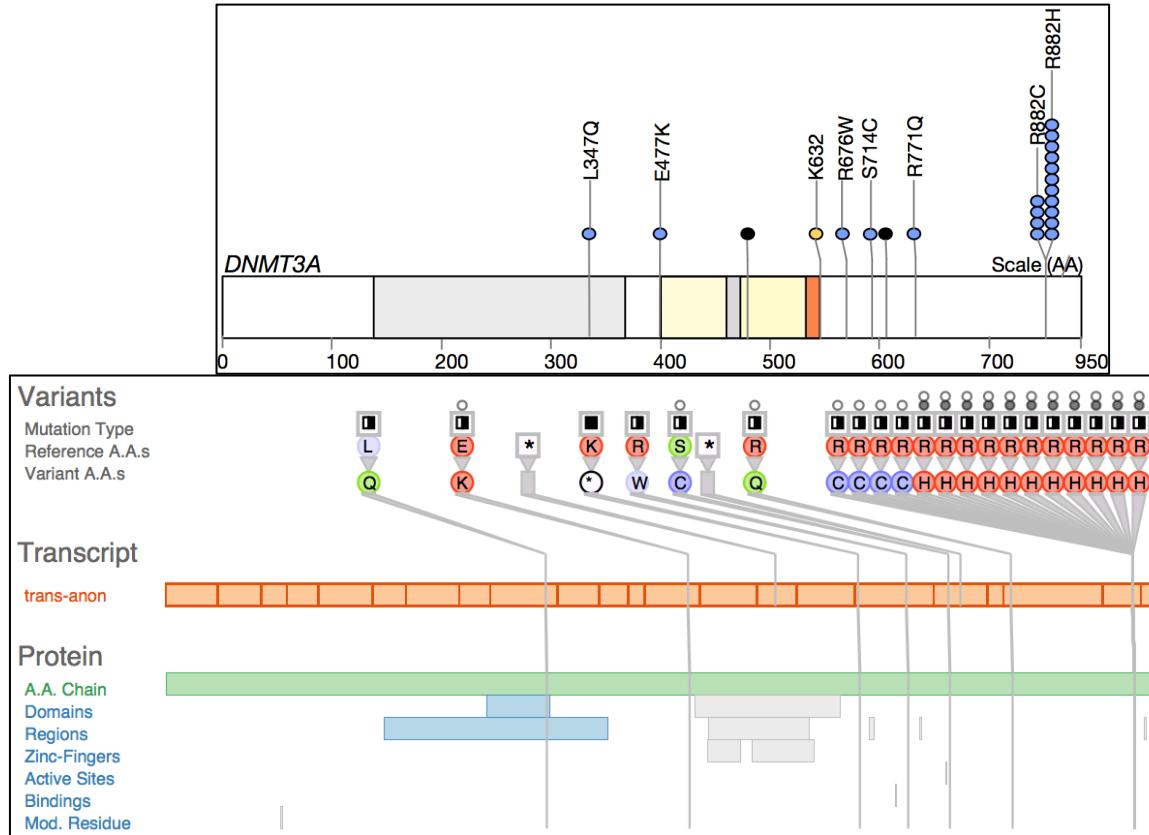
Related work:  
Targeted for variant analysis

# MuSiC variant visualization plot



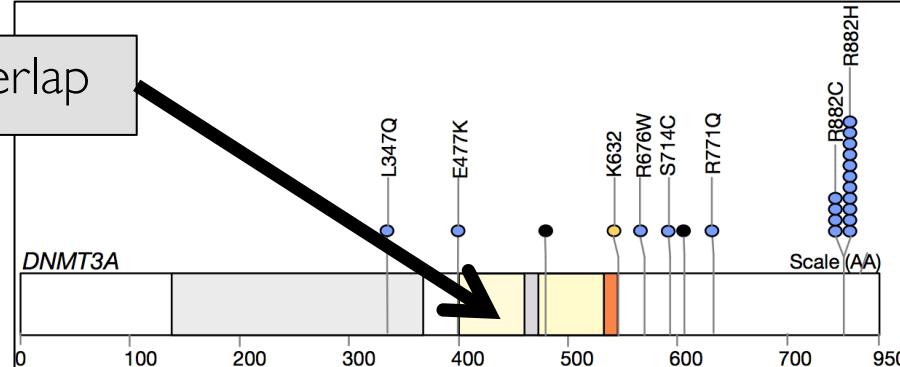
[Dees et al., Genome Research 2012]

# Side-by-side comparison



# Side-by-side comparison

Protein regions can overlap



Variants

Regions get separate lanes

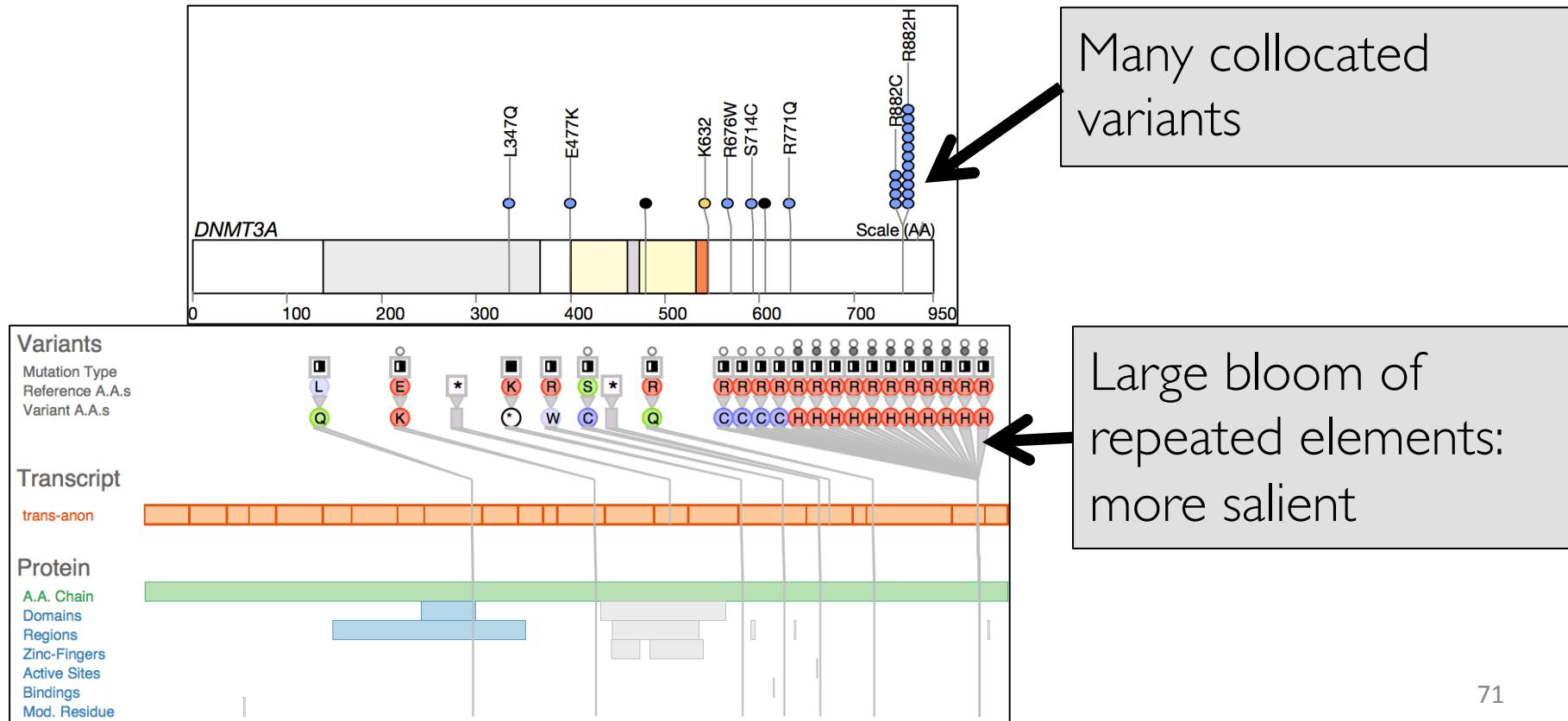
Transcript

trans-anon

Protein

A.A. Chain  
Domains  
Regions  
Zinc-Fingers  
Active Sites  
Bindings  
Mod. Residue

# Side-by-side comparison



# Driving biological tasks

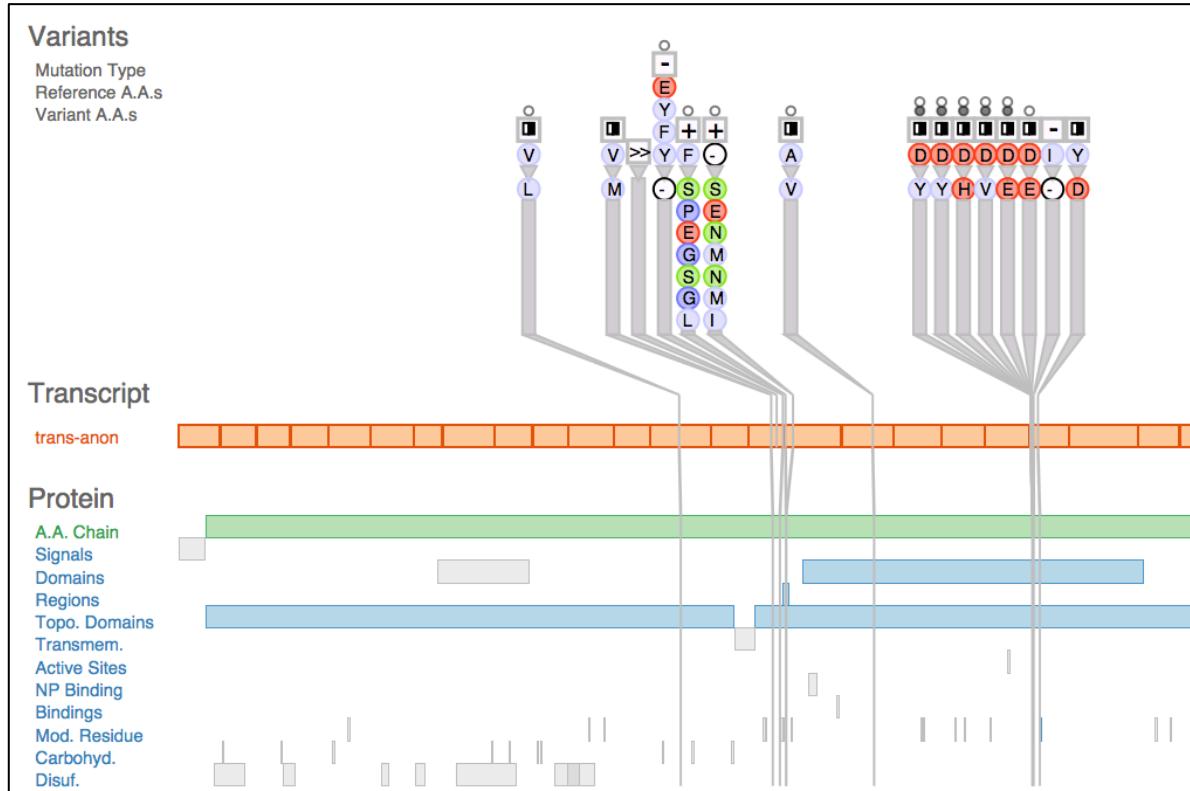
# Task I: Discover genes

- 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

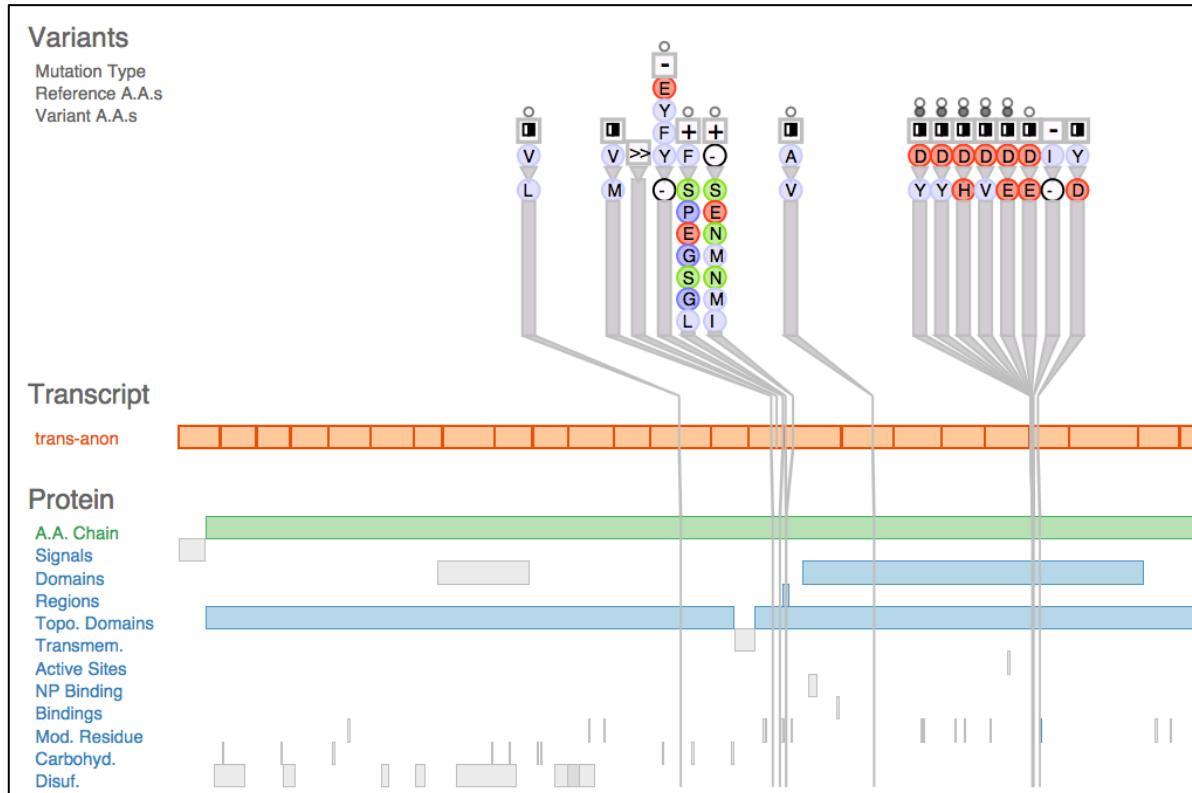
# Discover Task



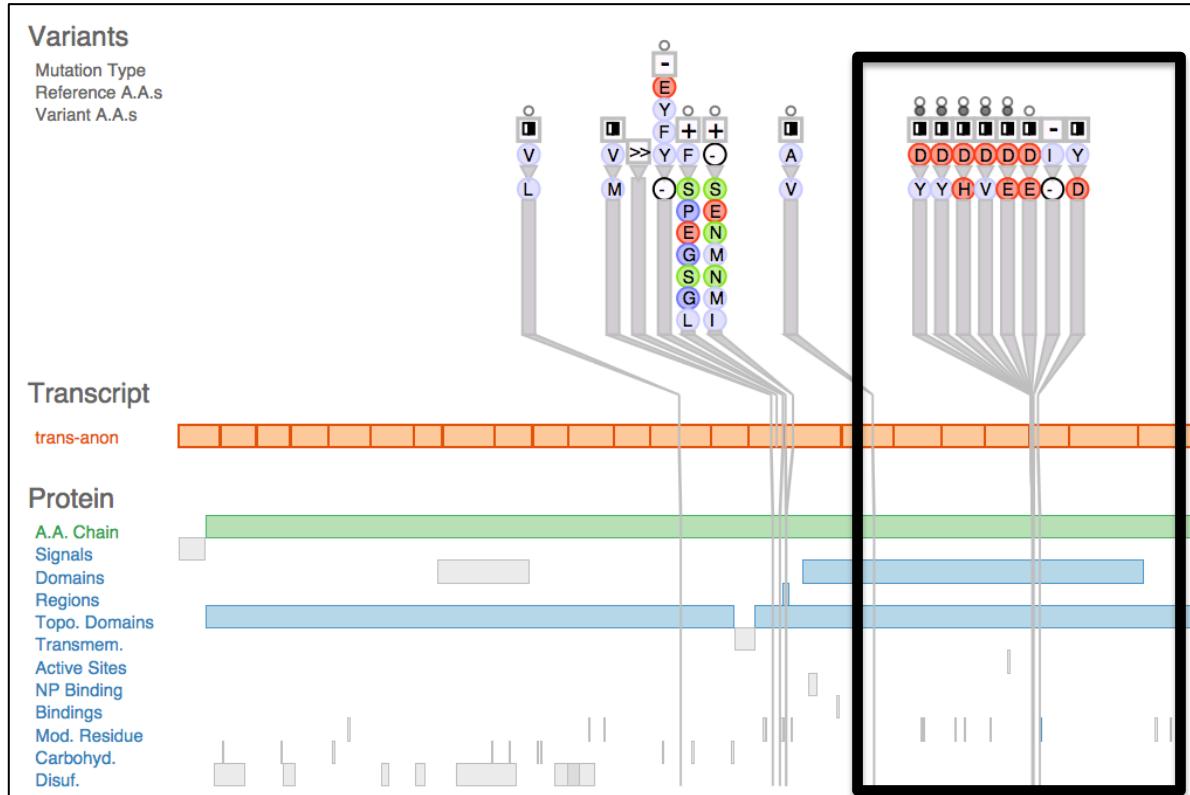
# Sorting by derived metric revealed known leukemia genes



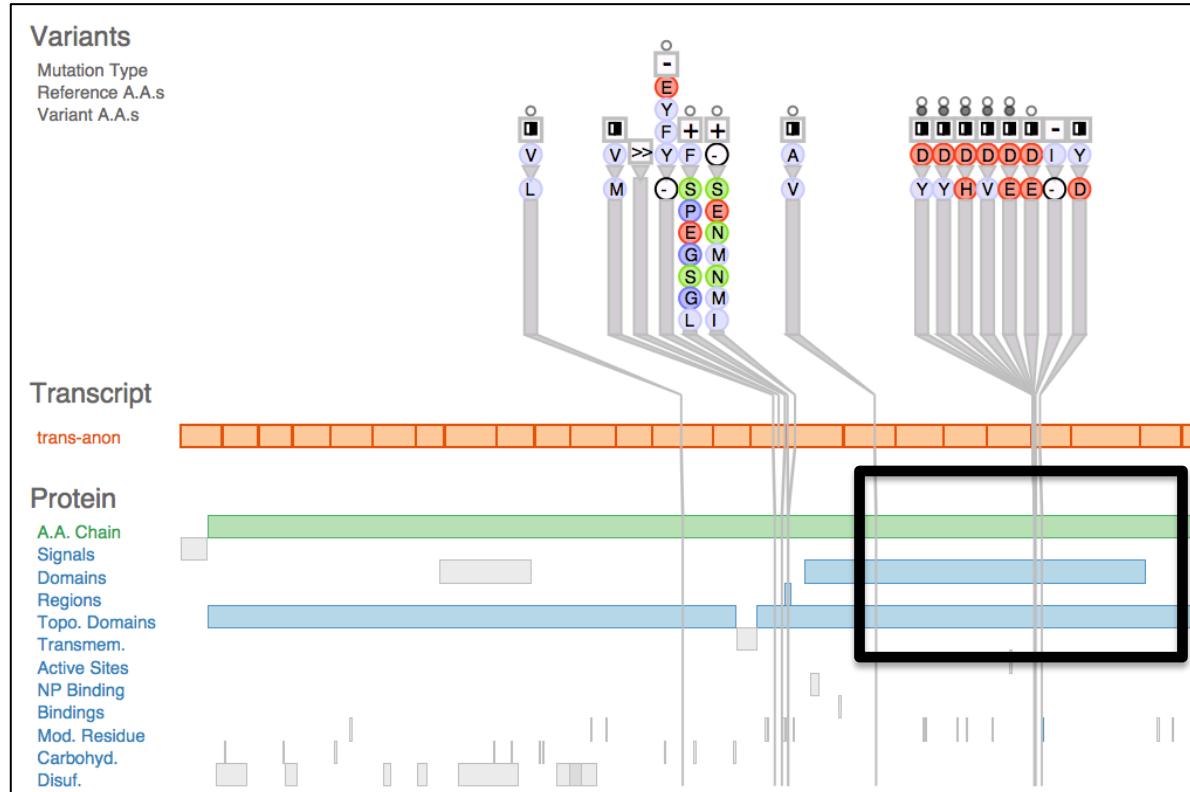
# Highly scored gene by sorting metric



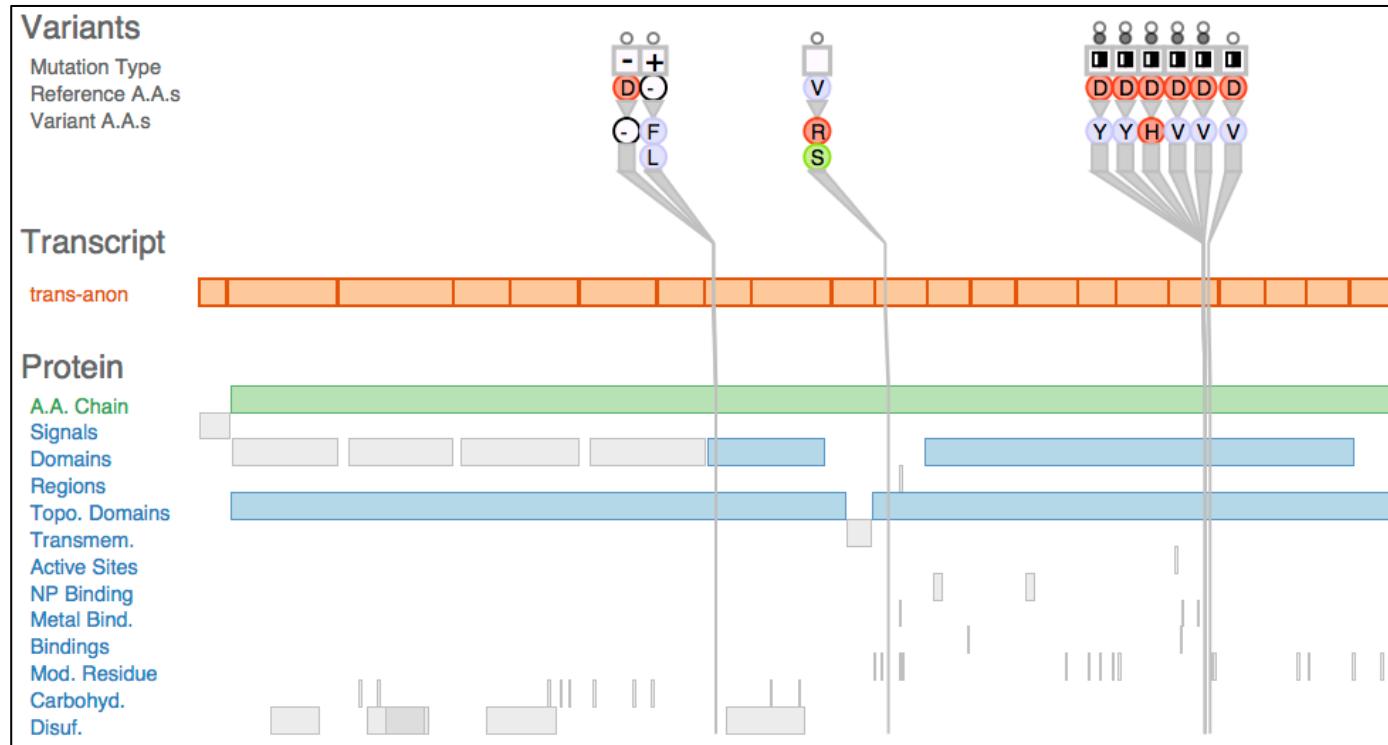
# Visual inspection reveals collocation of variants



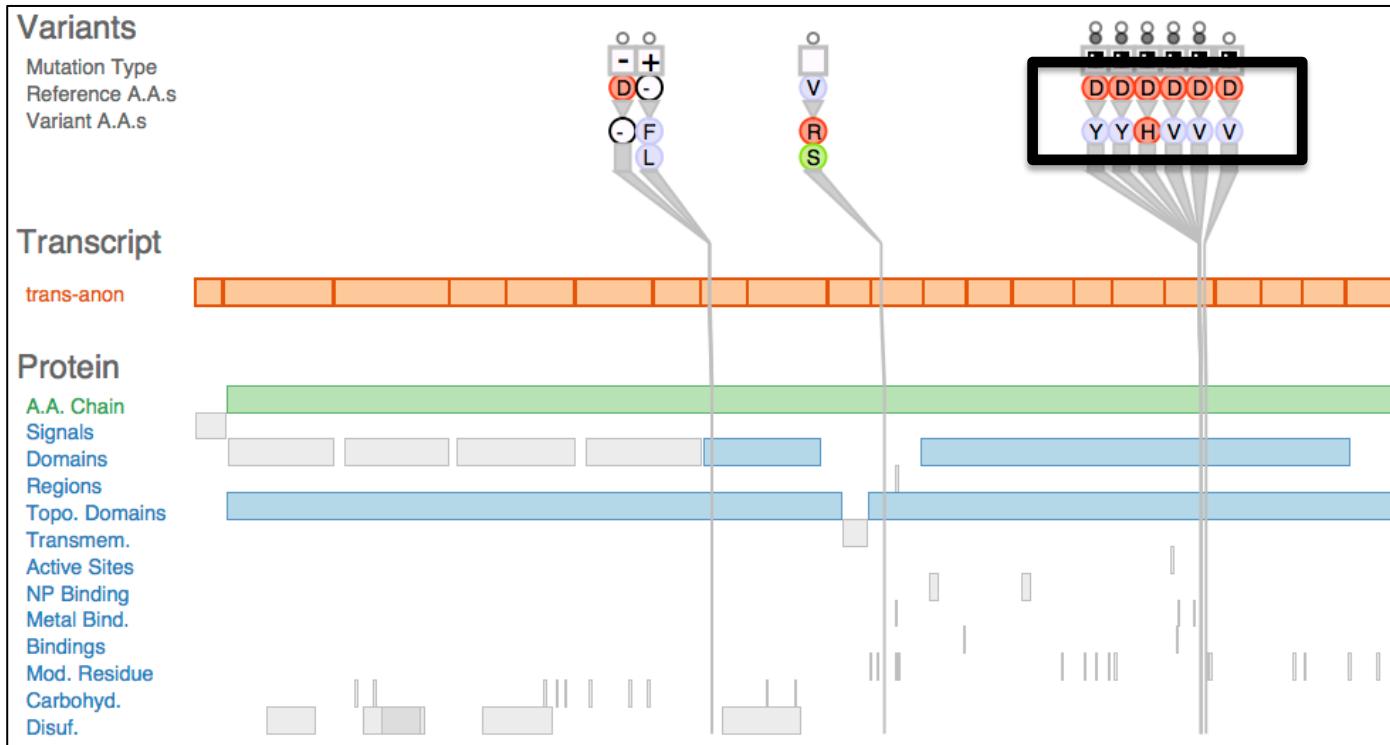
# Several functional protein regions affected



# Highly scored by metric and not known



# Protein chemical class change evident



# In contrast, low scoring gene

## Variants

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



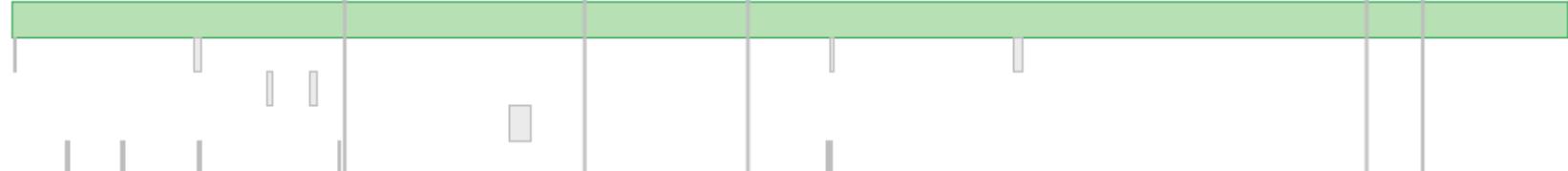
## Transcript

trans-anon

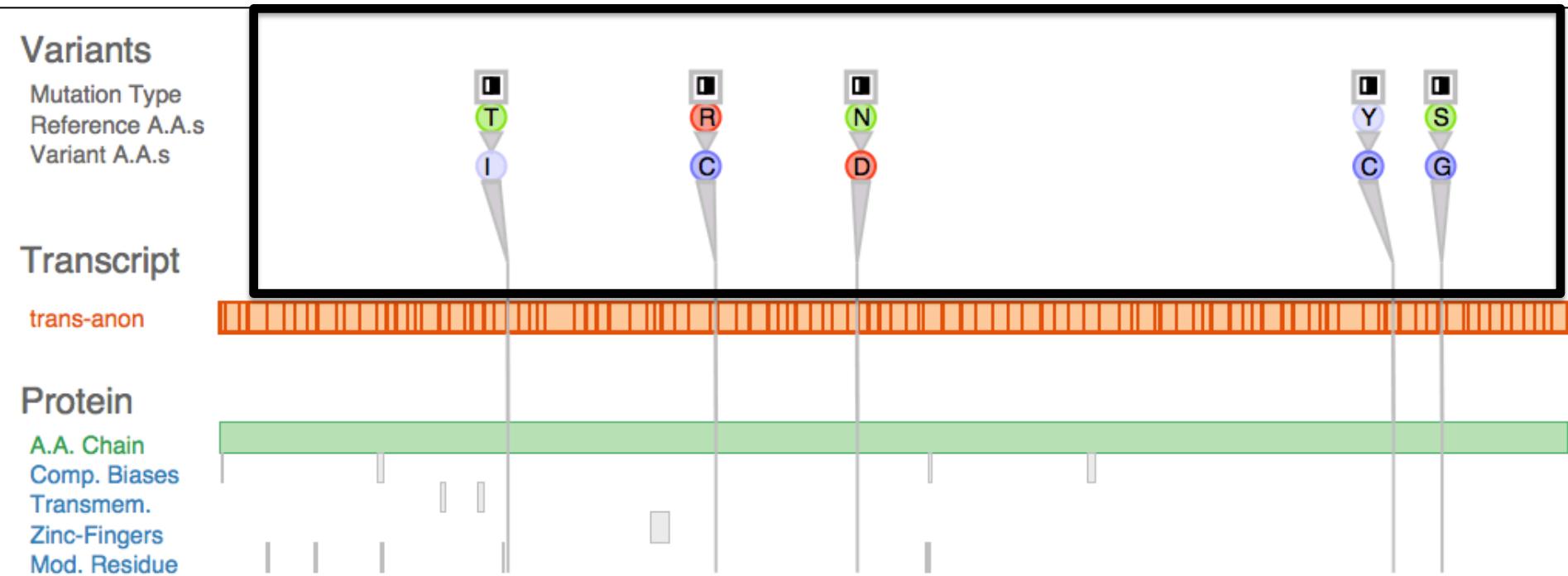


## Protein

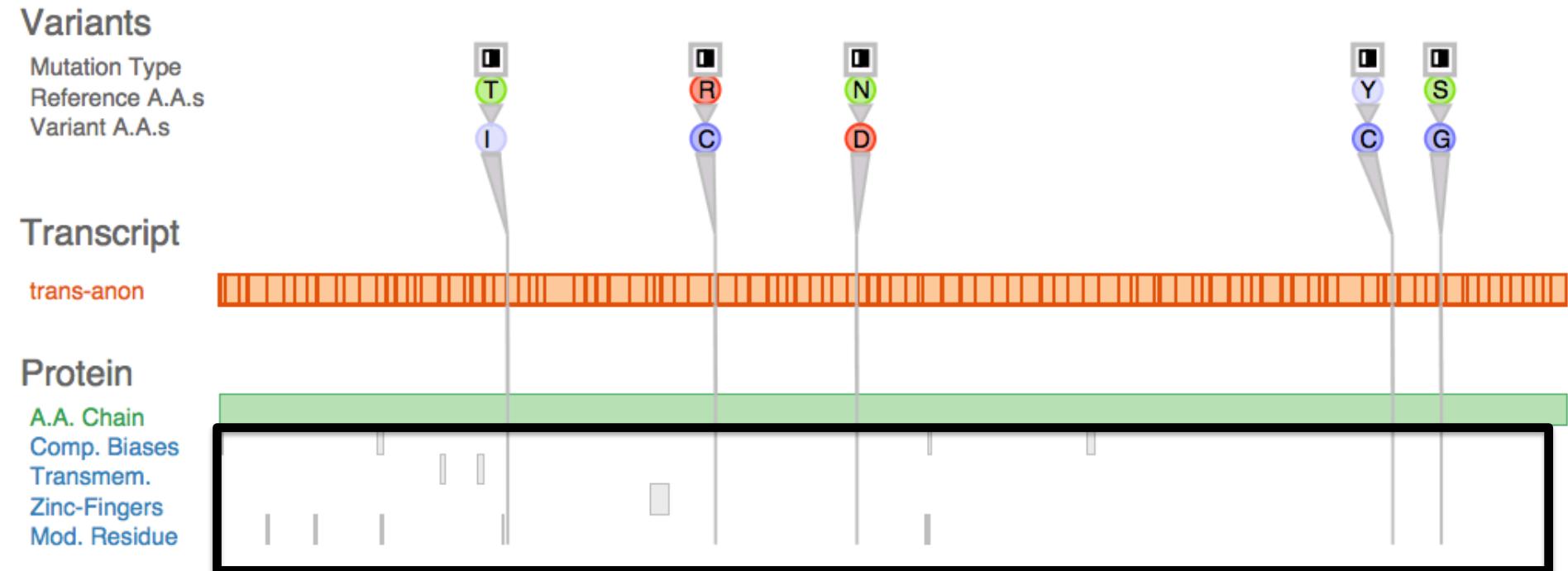
A.A. Chain  
Comp. Biases  
Transmem.  
Zinc-Fingers  
Mod. Residue



# 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



# Navigate through patient data with list

Select Patient: Patient 1 Submit  
Patient Genes: gene-anon Submit

Alternative Transcripts: gene-anon (trans-anon) gene-anon (trans-anon)

**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 data emphasized with arrows

Select Patient: Patient 1 Submit  
Patient Genes: gene-anon Submit

Alternative Transcripts: gene-anon (trans-anon) gene-anon (trans-anon)

Variants

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

Transcript

trans-anon

Protein

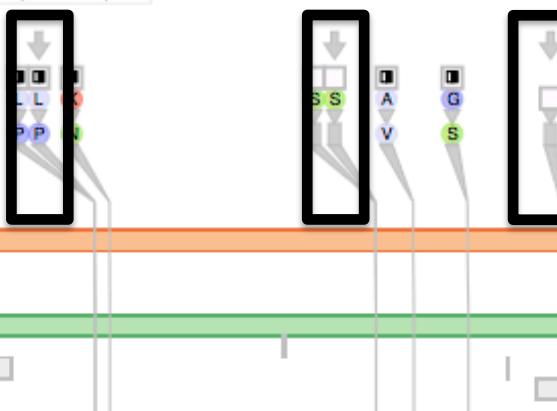
A.A. Chain  
Regions  
Comp. Biases  
Zinc-Fingers  
Mod. Residue

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



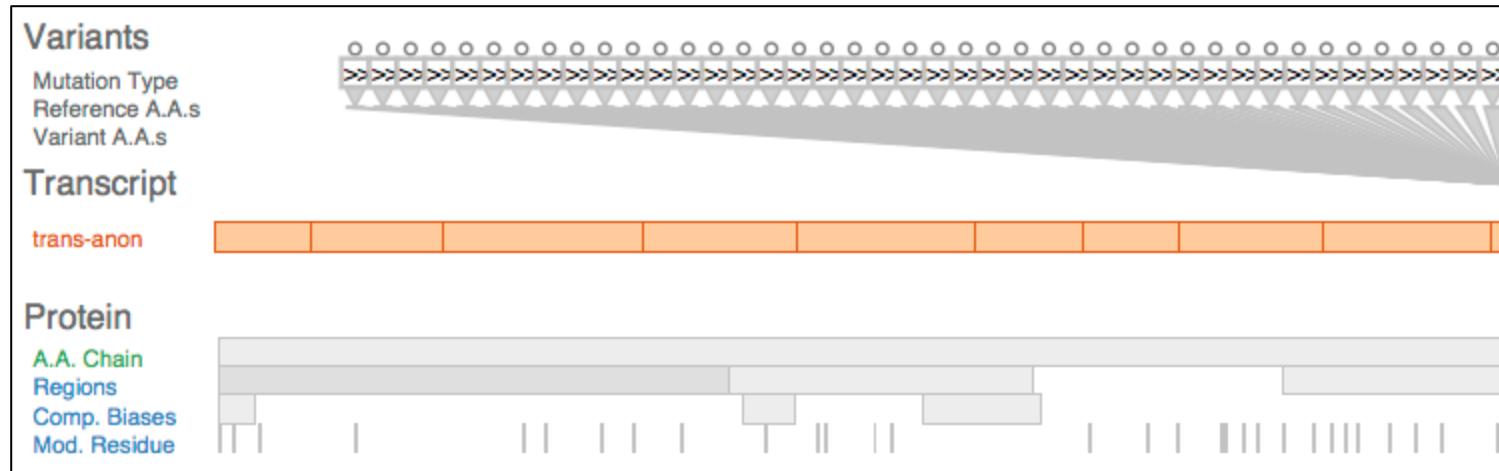
# These variants probably do not match



# 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

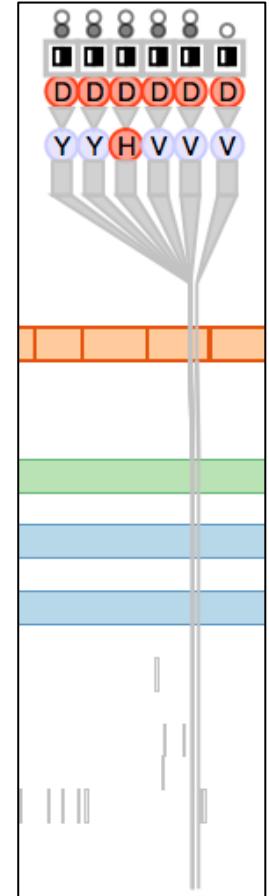


*The tool exposed artifacts in the data that slid past at least two rounds of quality metric filtering ... this type of problem would not have been caught by our previous, automated methods.*

- 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

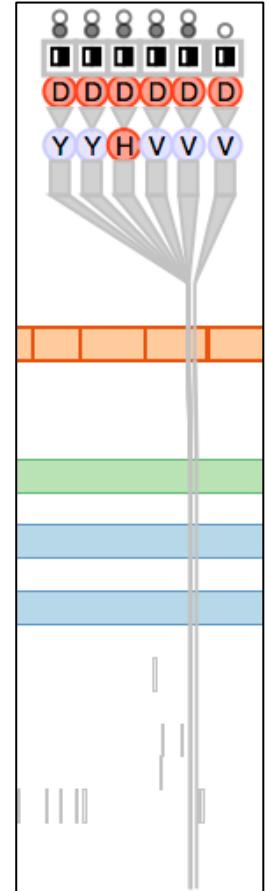


# Future work: Other use contexts

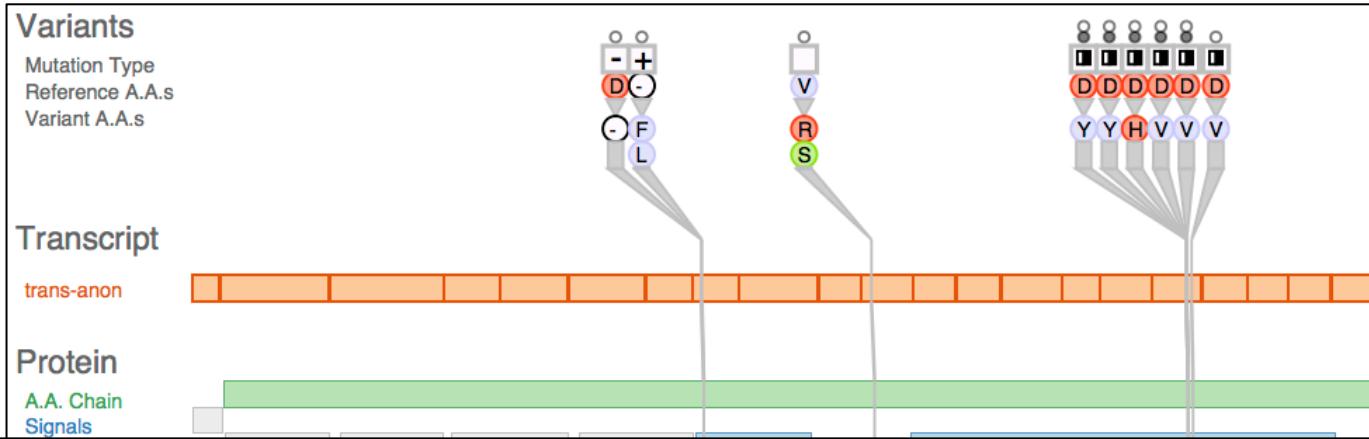
- Scaling up beyond the current design target
  - ~50 variants at once
- Possibly integrate Variant View with MedSavant
  - Tool by Fiume et al., BioVis Posters 2012
  - Focus on interactive filtering

# Acknowledgements

- Our collaborators at the GSC
  - Dr Aly Karsan
  - Rod Docking
  - Dr Linda Chang
  - Dr Gerben Duns
  - Simon Chang
- Funding
  - VIVA, AerolInfo Systems/Boeing Canada, MITACS



# Questions?



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

Paper Page:

**[http://www.cs.ubc.ca/labs/imager/tr/2013/  
VariantView/](http://www.cs.ubc.ca/labs/imager/tr/2013/VariantView/)**

Software Available as Open Source