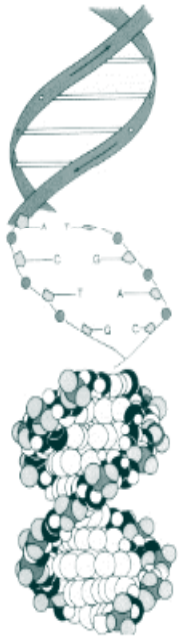


Use of Genomic Variants in i2b2

Lori Phillips MS
Simon Minovitsky
Igor Ratnere
Inna Dubchak Ph.D.
Isaac Kohane MD Ph.D.
Shawn Murphy MD, Ph.D.

Agenda

- Overview
  - **Requirements**
  - **Challenges**
  - **Ontological solutions**
  - **VISTA tools**

- Discussion

Requirements

- Ability to organize the variants for ease of navigation

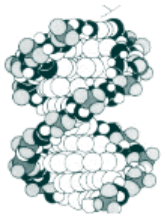
- Ability to query for the variant in the workbench
  - **Implication is that the identifier (basecode) for the variant does not change over time or is maintainable.**

- Ability to explore or annotate the variant within the workbench
  - **Implication is that we know enough about the variant so that it can be located in existing external genome browsers, analytical tools, etc**

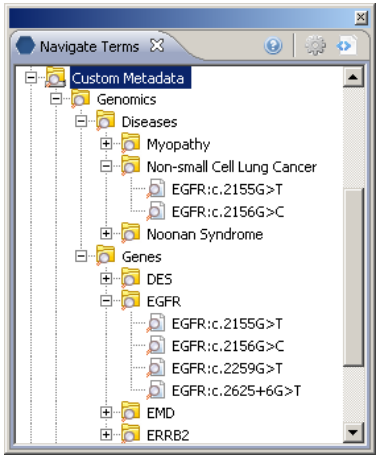
Challenges

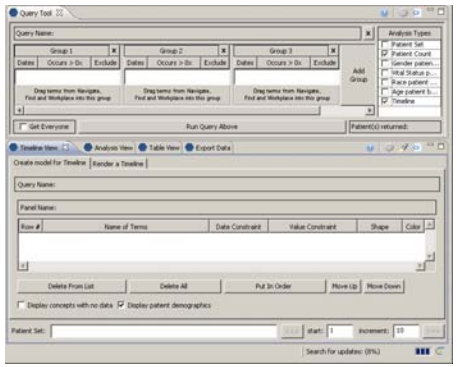
- Balancing the capabilities of multiple providers
  - **Genomic labs may report data differently**
- Maintainability
  - **Define the variant so it may be reliably identified over time**
- Balancing the needs of multiple consumers
  - **Needs may differ for geneticists vs physicians vs research scientists**

i2b2 Genomic vocabulary



Gather SNP data from genomic lab reporting system





Use vocabulary to query against CRC

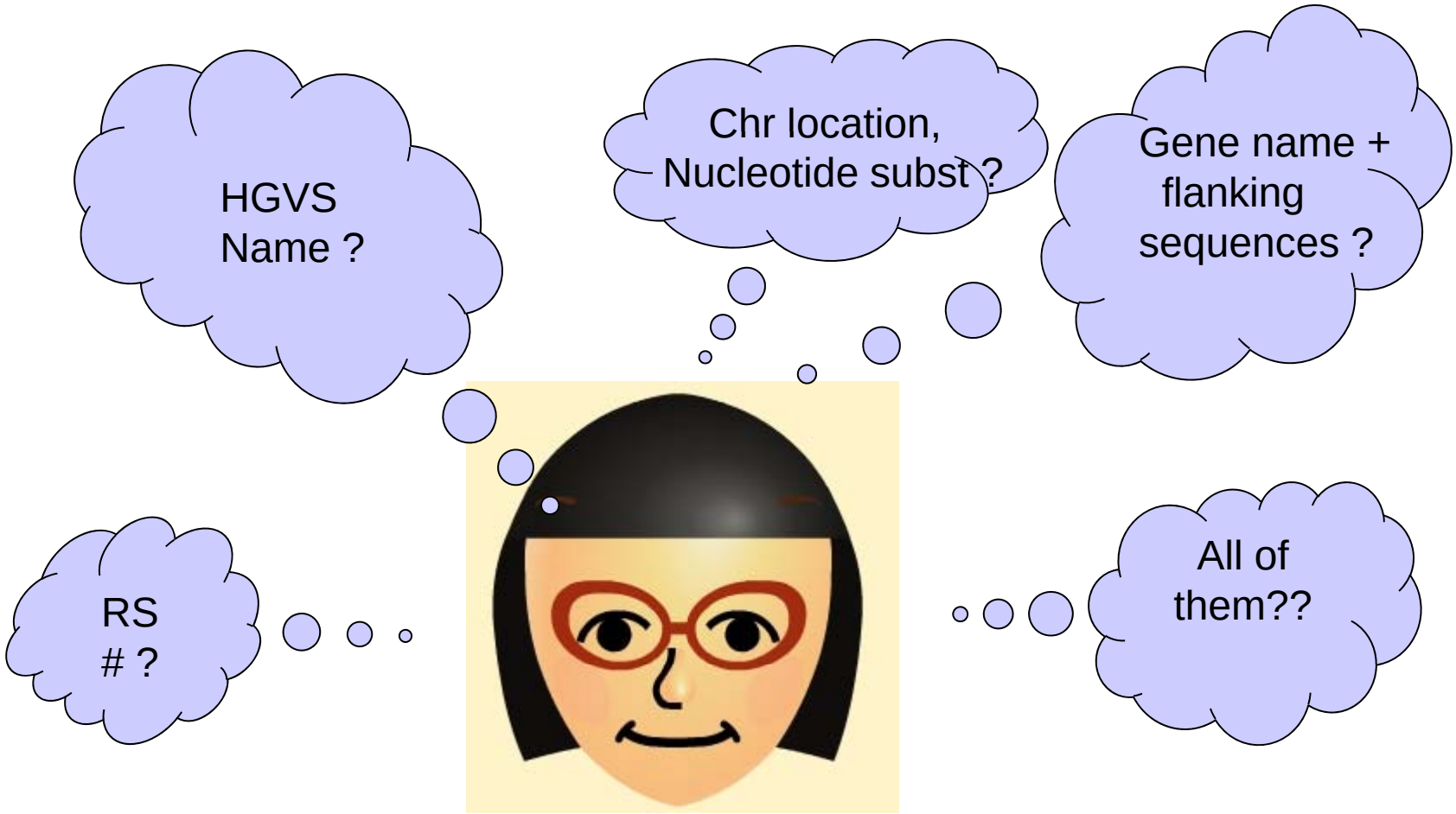
Gather SNP data from reference data

Use vocabulary to query VISTA annotations

Weighing the data provided by the lab source

- Gene location MYH7
- Flanking sequences
  - 5’ AGGCGCTAGAGAAGTCCGAGGCTC
  - 3’ CCGCAAGGAGCTGGAGGAGAAGAT
- Positional information c.2606
- Nucleotide substitution G>A
- Functional information p.Arg869His

How to (reliably) identify a genomic variant?



RS number

- Uniquely identifies a variant over time ….but….

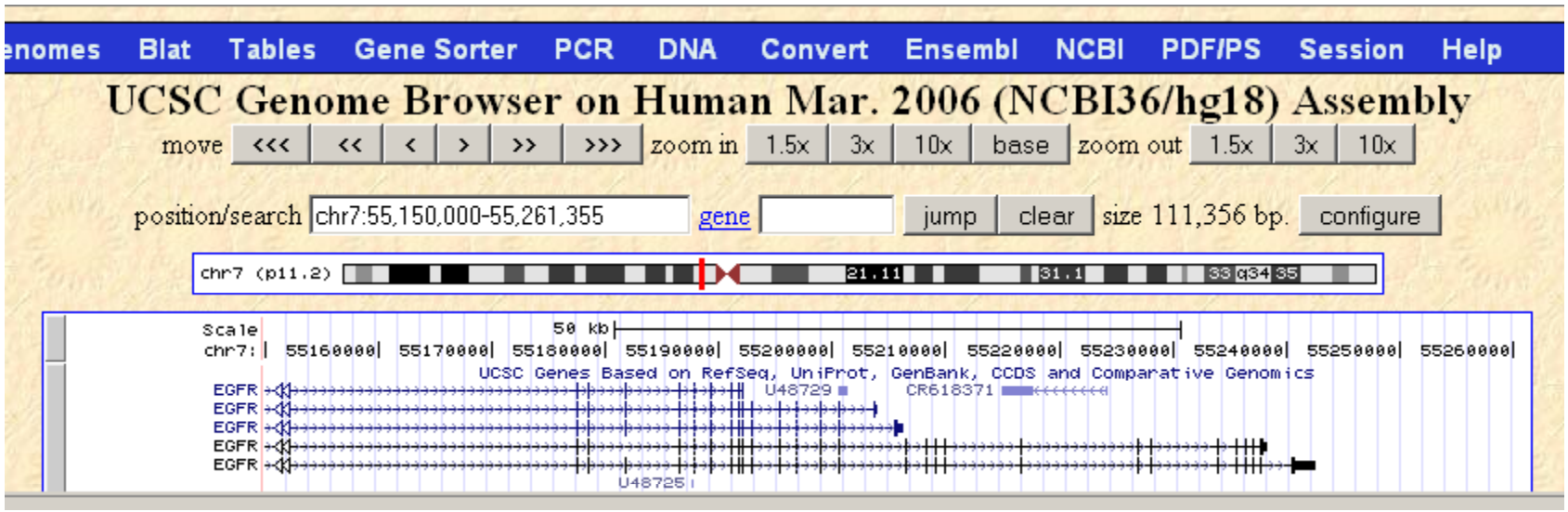
Reference SNP(refSNP) Cluster Report: rs28929495 **clinically associated**

RefSNP		Allele		HGVS Names
Organism:	human (Homo sapiens)	Variation Class:	SNP: single nucleotide polymorphism	NC_000007.12:g.55209201G>T
Molecule Type:	cDNA			NG_007726.1:g.159983G>A
Created/Updated in build:	125/132	RefSNP Alleles:	A/G/T	NG_007726.1:g.159983G>T
Map to Genome Build:	37.1	Ancestral Allele:	G	NM_005228.3:c.2155G>A
Citation:	PubMed	Clinical Association:	VarView OMIM	NM_005228.3:c.2155G>T
				NP_005219.2:p.Gly719Cys
				NP_005219.2:p.Gly719Ser
				NT_033968.6:g.4831076G>A
				NT_033968.6:g.4831076G>T

- Novel variants may not have rs number
  - User may not want to submit to dbSNP

Gene name + flanking sequences

- Not guaranteed if gene has several isoforms
  - EGFR



HGVS Name

- Uniquely identifies variant within a referenced and versioned accession and details the nucleotide substitution.

NM_005228.3:c.2155G>T

RefSeq accession

Coding DNA

Position

Nucleotide substitution

Is there a common denominator in all of this?

- Yes … all ultimately describe variant location on a chromosome.
- Nucleotide substitution defines the physical manifestation of the variant.

WE PROPOSE:

- HGVS name (n/t subst, positional info)
- Flanking sequences (a way to verify positional info)

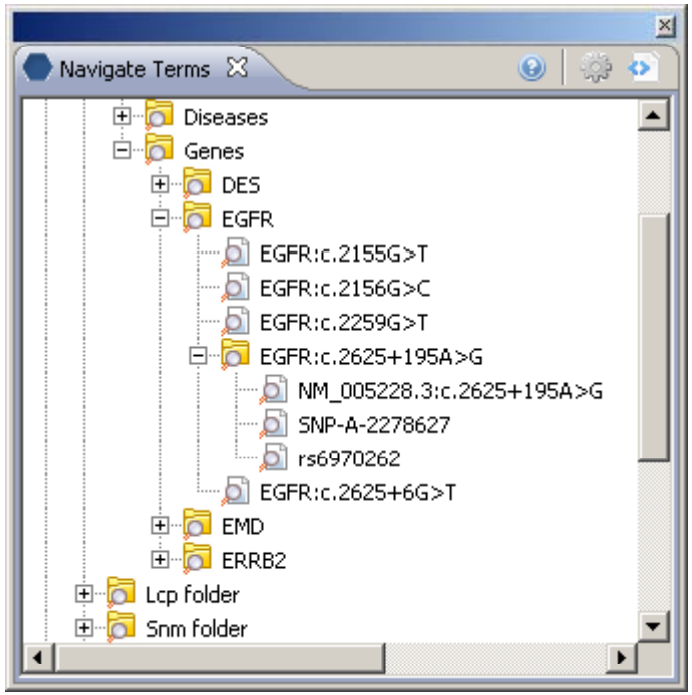
AS A WAY TO UNEQUIVOCALLY EQUATE TWO VARIANTS

- ACROSS DOMAINS
- ACROSS VERSIONS

GenomicMetadata record

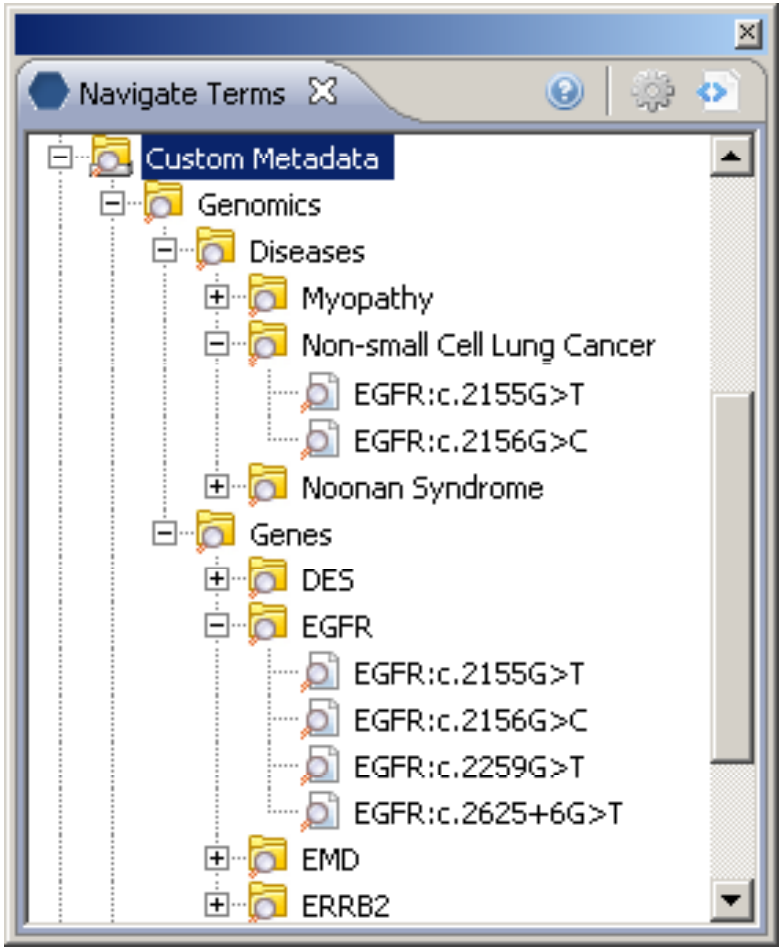
```
GenomicMetadata
  Version  1.0
  ReferenceGenomeVersion hg18
  SequenceVariant
    HGVSName  NM_0005228.3:c.2155G>T
    SystematicName c.2155G>T
    SystematicNameProtein  p.Glu719Cys
    AaChange missense
    DnaChange substitution
  SequenceVariantLocation
    GeneName EGFR
    FlankingSeq_5  GAATTCAAAAAGATCAAAGTGCTG
    FlankingSeq_3  GCTCCGGTGCGTTCGGCACGGTGT
    RegionType exon
    RegionName Exon 18
    Accessions
      Accession
        Name NM_005228
        Type mrna (NCBI)
      Accession
        Name NP_005219
        Type protein (NCBI)
      Accession
        Name NT_004487
        Type contig (NCBI)
  ChromosomeLocation
    Chromosome chr7
    Region 7p12
    Orientation +
```

Combining equivalent terms

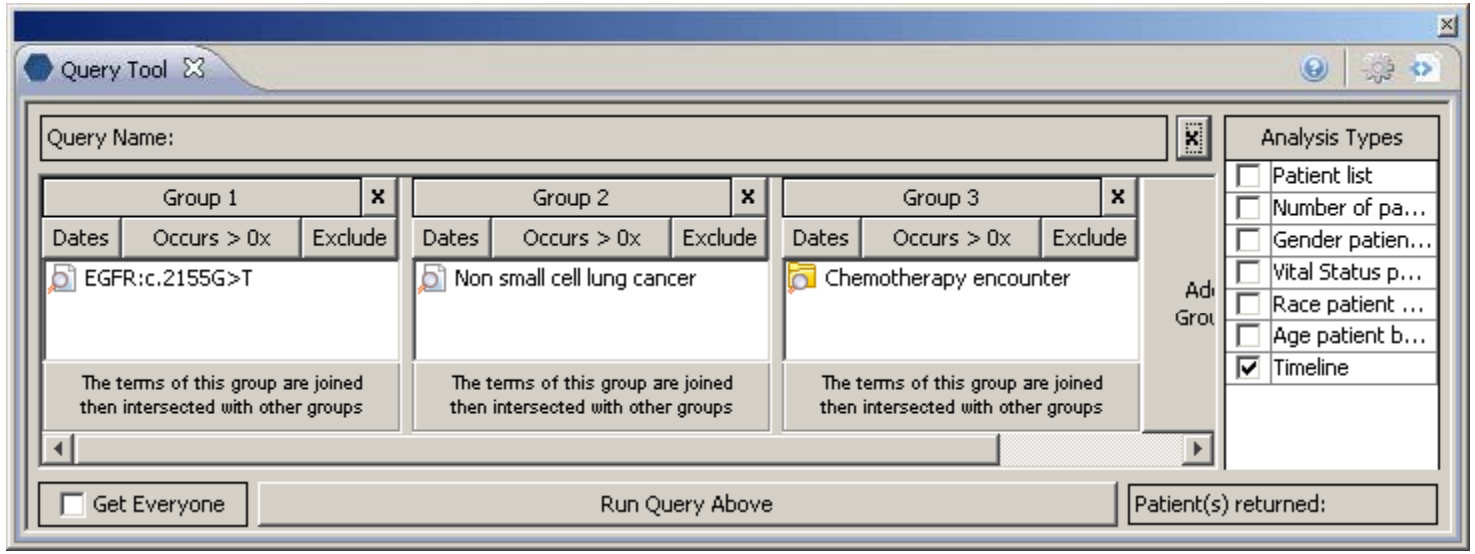


Organizational challenges

- By Disease?
- By Gene?



Translational query across 3 domains



Linking to external services

- Genome Browser
  - Requires chromosome location; reference genome
- PolyPhen (predicted functional effects)
  - Requires chromosome location; reference genome
  - RS number
  - Or HGVS name

VISTA Services

- Flankmap (location service)

  Converts several formats to a chromosome location on a reference genome

  - Gene/flanking sequence
  - Full HGVS notation
  - dbSNP rs number

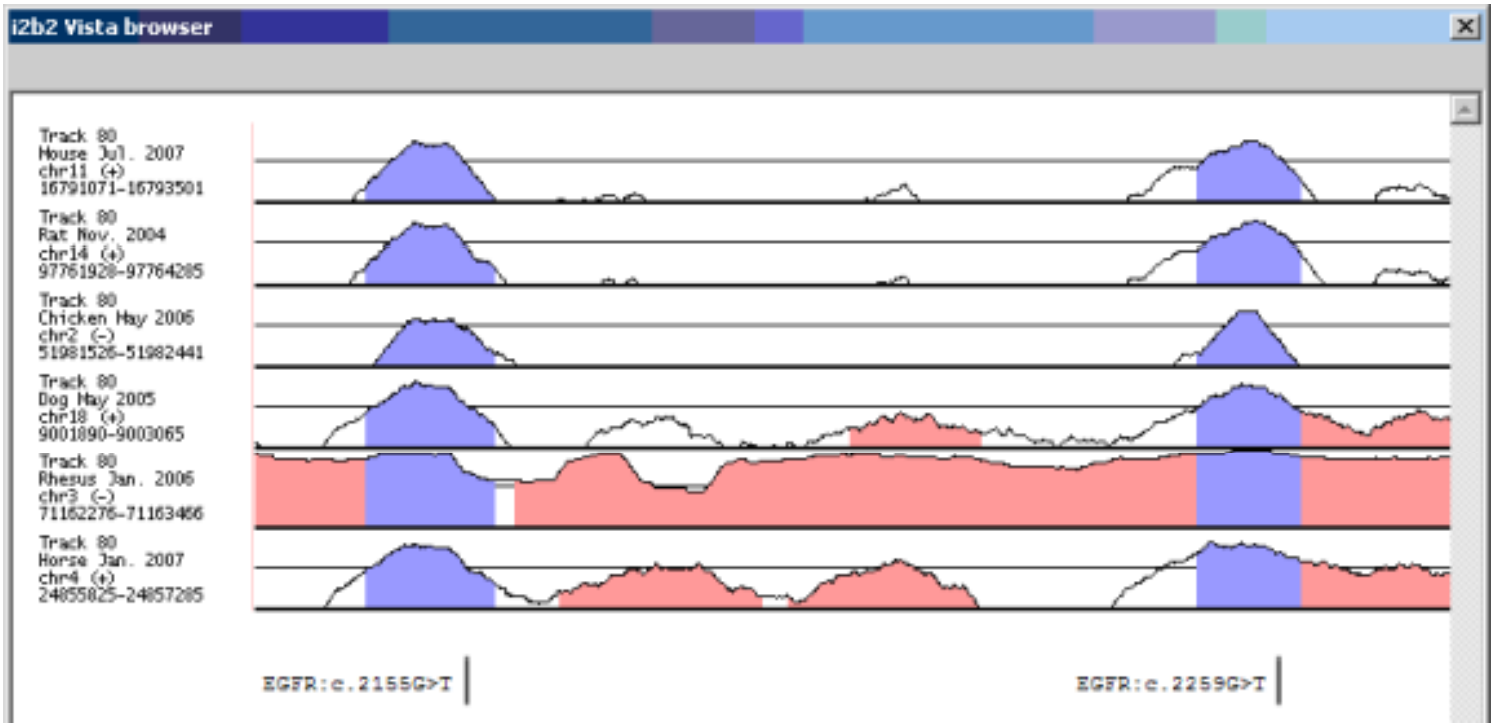
- Conservation plots
  - Based on location

VISTA workbench tools

Variant View

Select	SNP Name	Gene Name	Chromosome	Conservation Map	PolyPhen Score	Variant type
☑	c.2155G>T	EGFR	chr7		Probably damaging	Non-synonymous coding
☐	c.2156G>C	EGFR	chr7		Probably damaging	Non-synonymous coding
☑	c.2259G>T	EGFR	chr7			Synonymous coding
☐	c.2625+6G>T	EGFR	chr7			Non coding

Embedded VISTA browser



Acknowledgements

- VISTA team
  - Inna Dubchak
  - Simon Minovitsky
  - Igor Ratnere

THANK YOU