# Bioinformatics in Drug Discovery and Development

Lessons Learned

MD Healy

February 2018

# Acknowledgments

"Our success is defined by our contributions to the success of others." –Eleanor Roosevelt

- BMS Drug Discovery
- BMS Microbiology
- BMS Clinical Development
- BMS Pharmaceutics
- BMS HTS Group
- Ploss Lab at Princeton
- Yale Human Genetics
- Yale Medical Informatics
- Regeneron Translational Genetics
- Regeneron Analytical Genetics
- Geisinger Health System
- Regeneron Infectious TFA
- Regeneron Immune TFA
- Regeneron CVD TFA
- UMD Collaborators
- *Thousands of volunteer subjects*

# Sample accomplishments

Built and maintained tools for mining Microbial and Viral Genomes

Led first NGS Pilot project at BMS, including experimental design and building analysis pipeline from scratch

Helped maintain databases and tools for Gene Ontology and Biological Pathway Enrichment Analysis

Target Dossiers for: Infectious Diseases, Immunology, Oncology, Neuroscience, CVD, & Genetically Defined Diseases

Analyzed Drug Resistance Mutations in Viral Target Genes from Early Discovery through Pre-Clinical studies and Clinical Trials.

Mined large-scale databases of Human Genetics results to understand the genetics of common diseases

Helped design, and was main analyst for, many Transcriptional Profiling studies

Leveraged OMICS and Biological knowledge to understand Genetics results

# My proudest achievements

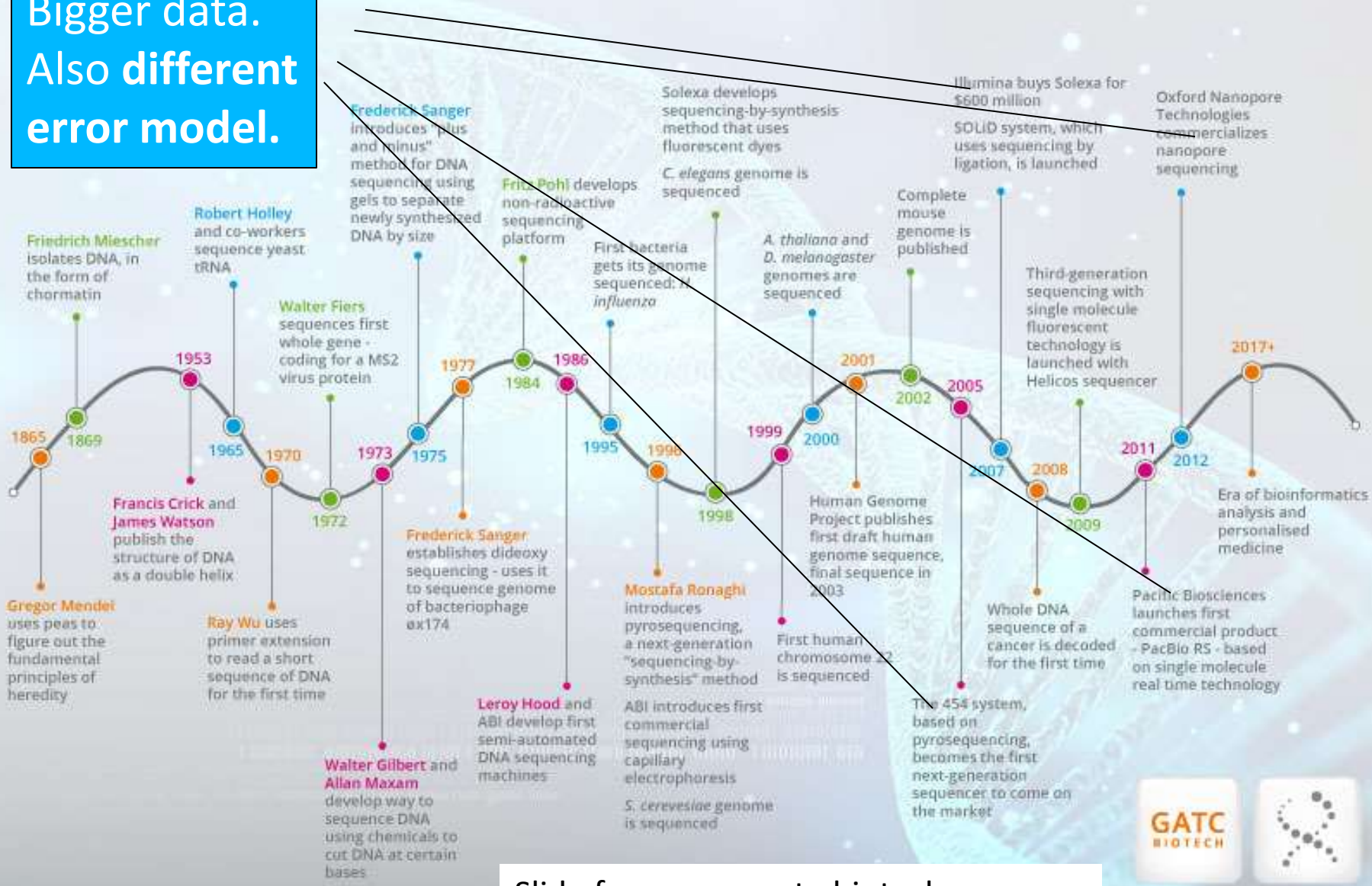
- Played a key role in getting **four new drugs on the market** and several others into human clinical trials
- Was key member of cross-functional teams that accomplished urgent data analyses for FDA submissions for drugs either on the market or in late-stage trials
- Identified novel Methyltransferase Motif found in gene conserved by all Gram-positive bacteria
- Was part of team that found novel mechanism of action for inhibiting Influenza Virus replication

# Outline

- Evolution of sequencing and how tools have adapted
- Impact of Genomics on Virology pipeline at BMS
  - Cross-functional collaborations
  - Resistance mutations from Early Discovery to The Clinic
  - Transition from Sanger to NGS
- Genomics Impact in *all* Disease Areas at BMS and RGC
  - Integrating huge range of internal and public data types
  - Genetics and Structural Biology play complementary roles
    - Each helps us to understand the other
  - Checkpoints and Pathways
- Key lessons learned

# A Brief History of DNA sequencing

Bigger data.
Also **different error model.**




Slide from www.gatc-biotech.com

GATC BIOTECH

Your sequencing services provider

# What makes a great tool

Bonus points if tool is easy to install and configure

- Plays nicely with other tools
  - Human-readable HTML or text formats
  - Tabular output, any columns, any order, any delimiters
  - Can it be a **drop-in replacement for a existing tool?**
  - Lends itself to parallel processing
- Easy for humans to use interactively
  - Well-chosen defaults
  - Beginner vs Expert modes
  - Canned Examples
  - Extensible Help System
- Helps user transition to commandline
  - Same executable senses its environment, runs as CGI or as commandline
  - Web output always includes commandline to generate it

Having Web/GUI interface include the equivalent commandline in its output is great for several use cases: helping a new user learn the commandline, saving a pipeline developer some time, and **capturing what was done for Electronic Lab Notebooks**

# Analysis changes as tech does

Short reads
⇒ambiguous mapping

Sanger: still "Gold Standard"

454: medium read length, more indels, *hotspots*

ONT & PacBio: very long reads with high error rates

Ion torrent: short reads, high indels

Illumina: short reads, high quality, mostly miscalls

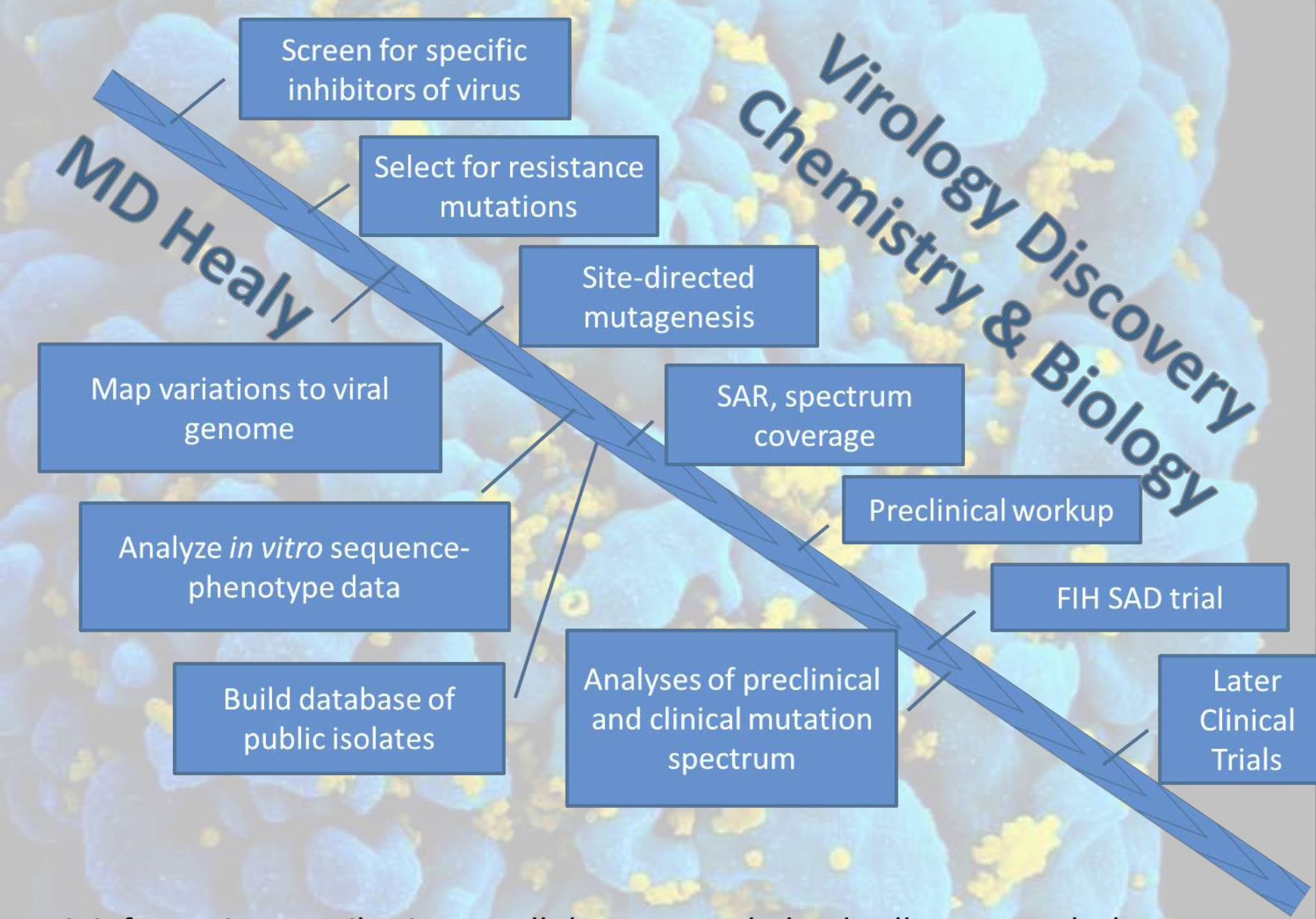
Genotyping Chips: lower cost, higher throughput, but imputation needed for coverage and integration with other data. *Need population-specific reference genome/exome data for imputation*. Need for larger sample sizes to find rare variants has led to current **resurgence of interest in chips**.

# Major challenges of GWAS today

- It's possible much of the "dark matter" heritability not found in GWAS to date comes from *very rare variants*
- Therefore, need to study more people. Bigger cohorts, more diverse cohorts, founder populations
- With good reference sequences and huge populations, chips and imputation can be powerful
- Population structure complicates analysis
- Founder populations have huge linkage blocks
- Must use OMICs, Biology, etc., to narrow it down!

# Virology Pipeline Impact

- Early Discovery
  - ID target (viral genetics)
  - Validate target
  - Resistance
  - Public sequence mutation spectrum
- Full Discovery
  - SAR/resistance
  - Transcriptional profiling of CPDs
  - Tox signals
  - OMICs studies on animal models
  - Continue monitoring public clinical sequences
- Translation to Clinic
  - Resistance, resistance, resistance
  - Tox, tox, tox
  - *Do we see same mutations in clinic as we saw before?*




Bioinformatics contributions at all these stages helped colleagues make key decisions in discovery and development of drugs for Hepatitis B, Hepatitis C, and HIV

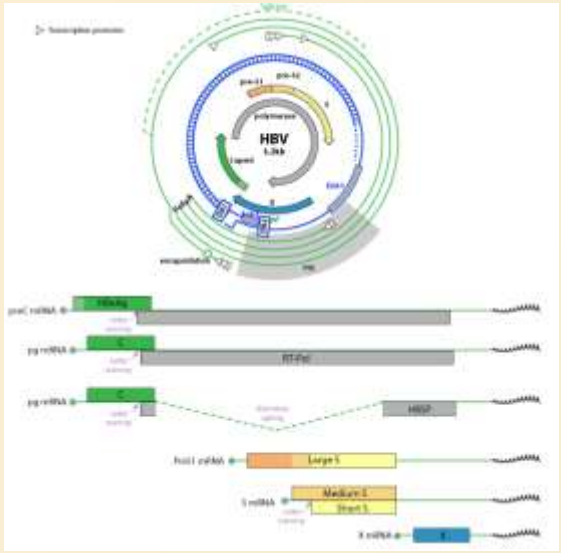
# My biggest program impact

## Virology Resistance Mutations

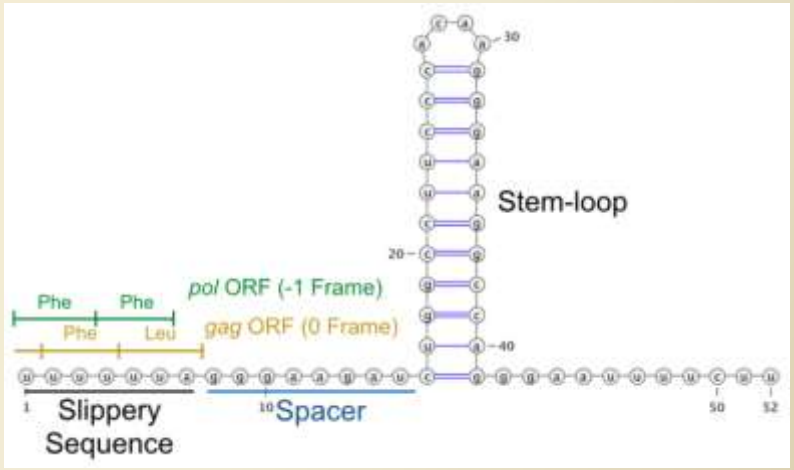
1. Essential for moving multiple drugs through the pipeline, to the clinic, and to market
2. I built tools to integrate public, BMS *in vitro*, and BMS clinical sequences for a complete picture of the resistance mutation spectrum
3. Led first BMS 454 pilot including building analysis pipeline. Later helped define BMS standards for reporting Illumina clinical seqs

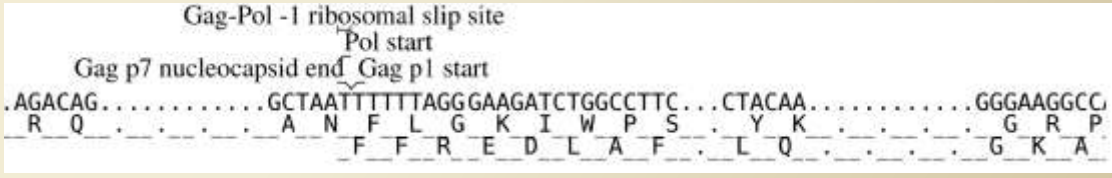
# Viruses: very strange transcripts

Hepatitis B Virus: circular genome; ribosomal slippage allows bypassing stop codon making longer proteins




HIV: ribosomal slippage creates overlapping reading frames with different translations of same NT

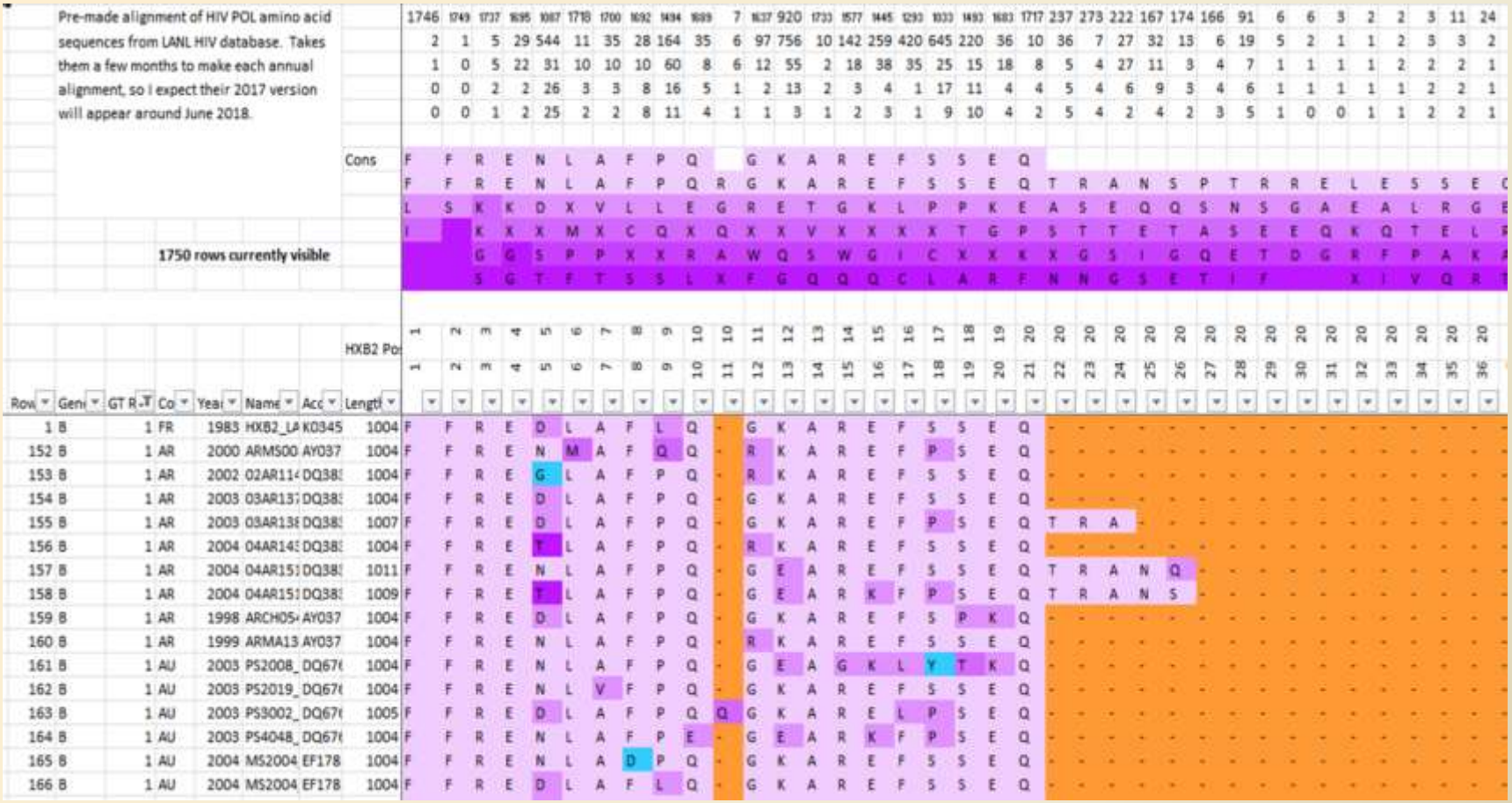



Standard alignment tools often get confused; Frameshift or Exonerate algorithm can use reference protein sequences to get codons of desired protein sequence

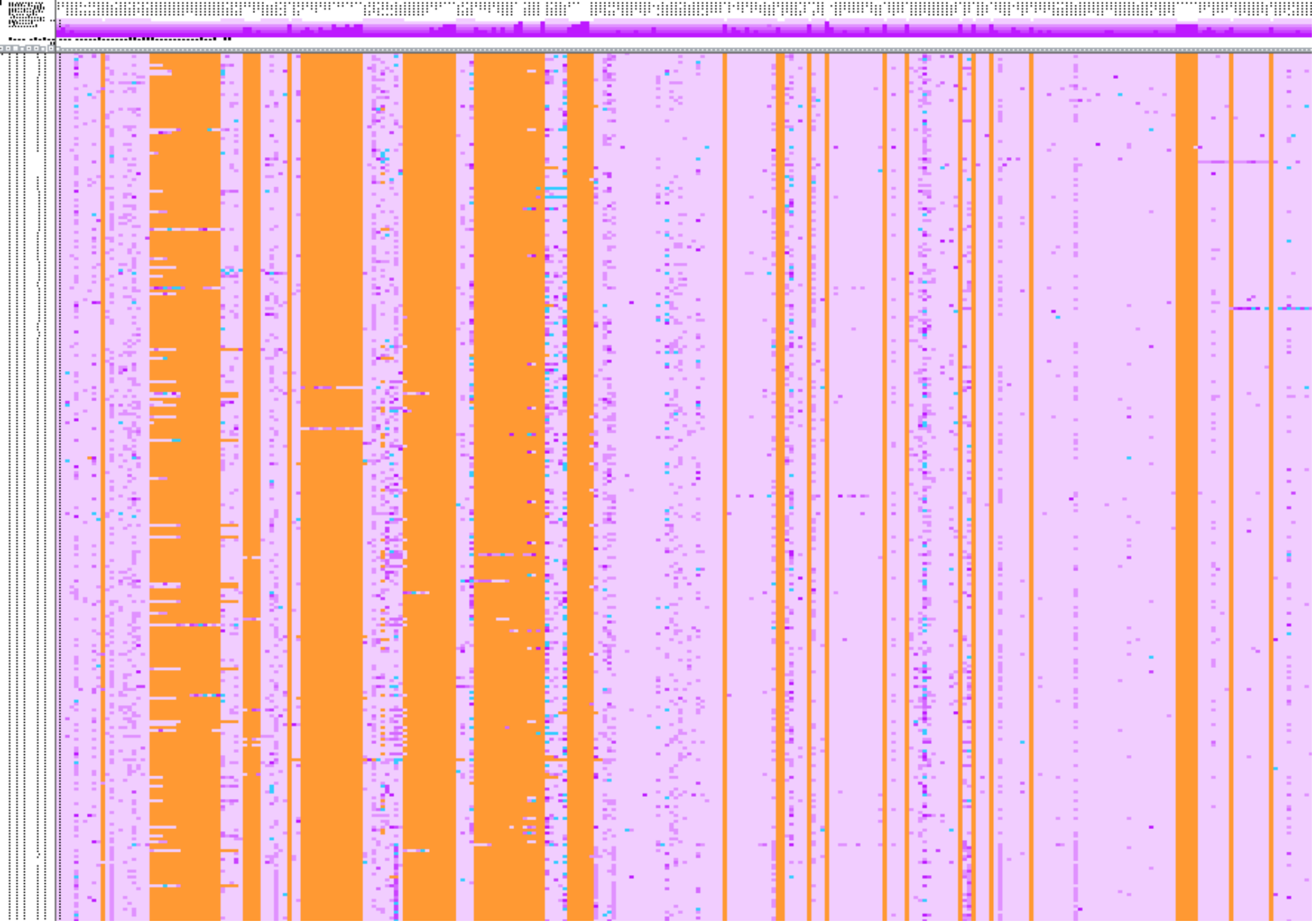
# Interactive Sequence Viewer

## HIV POL from LANL




I have uploaded a working example of my tool to GitHub. You can find at https://goo.gl/Gdsp5t

User can change which sequences are displayed by clicking on the Filtering arrows. **The profiles at the top and the color coding change dynamically as the user changes the displayed sequences by changing the filter settings.**



Zooming out, color coding enables user to see large-scale patterns

# Sanger/Affy → to NGS at BMS

**Key issues with NGS include:**

**Shorter read: lose linkage information**

**Clinical validation for Virology**

**Sanger mutation assays already approved**

**Huge data volumes.**

- **Alignment of reads to the reference**
  - **Accuracy vs speed**
  - **Handling reads that map to multiple sites**
  - **Transcript models: Ensembl vs RefSeq**
- **Normalization of NGS data still evolving**
  - **The field has yet to agree on metrics/tools**
  - **Operational implication #1: re-analyze same data**
  - **Operational implication #2: Keep raw data!**

# Impact on *many* Disease Areas

- Early Discovery
  - Known bio consistent with proposed MOA?
  - Off-target?
  - What do most need to know now and how can genomics help?
- Full Discovery
  - Models behave as expected?
  - What do the inevitable surprises mean?
  - What do most need to know now and how can genomics help?
- Translation to Clinic
  - Do tox and efficacy results exactly match our expectations?
  - *Of course not! But why not? Showstopper or not?*

# Data Sources Used

- Sequence/Genome Databases
  - NCBI, Ensembl, SwissProt, internal BMS sequences
- Structural Biology
  - PDB, literature, BMS internal/purchased structure data
- Transcriptional Profiling
  - NCBI UNIGENE & GEO, EBI ArrayExpress, BMS XPRESS database
- Pathway databases
  - KEGG, MetaCyc, GeneGo, WikiPathways, MetaCore
- OMIM and BMS in-house databases
  - Multiple Sequence Alignments are the single most useful type of sequence analysis, in my experience

# Data sources, cont.

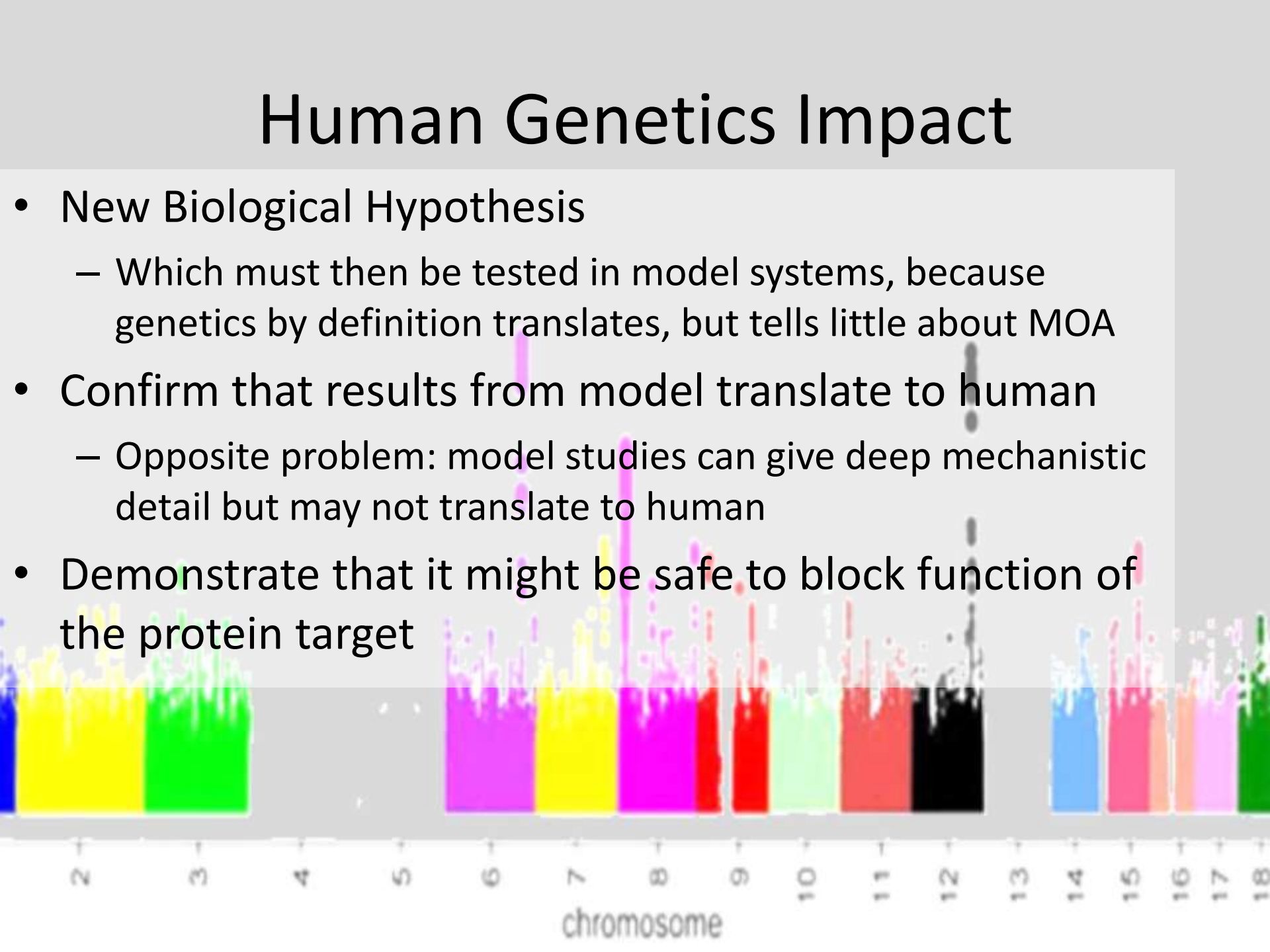
- Expression by Tissue and Disease State
  - UNIGENE, EMBL Expression Atlas, GEO, BMS Control Cell Lines, BioGPS, GTEx, etc.
- Biological Pathways
  - Upstream and downstream
  - Overlay profiling data on pathway diagrams
- Patent sequences
- Literature survey
  - Time-consuming but vital
  - Working Groups sometimes missed key findings

# Data sources, cont.

- Tool Strains
  - Knockouts, Conditional KO, Overexpression, Underexpression
- Tool Compounds
  - Phenotypes evoked, *how specific are they*
- Orthologs and Paralogs
  - Is binding pocket or Antibody Epitope conserved in models?
  - Possibilities for off-target binding
  - Do we believe the orthology calls!?
  - Homologene, InParanoid, Ensembl Genes are helpful tools
  - But no substitute for careful inspection of phylogenetic trees

# Human Genetics Impact

- New Biological Hypothesis
  - Which must then be tested in model systems, because genetics by definition translates, but tells little about MOA
- Confirm that results from model translate to human
  - Opposite problem: model studies can give deep mechanistic detail but may not translate to human
- Demonstrate that it might be safe to block function of the protein target



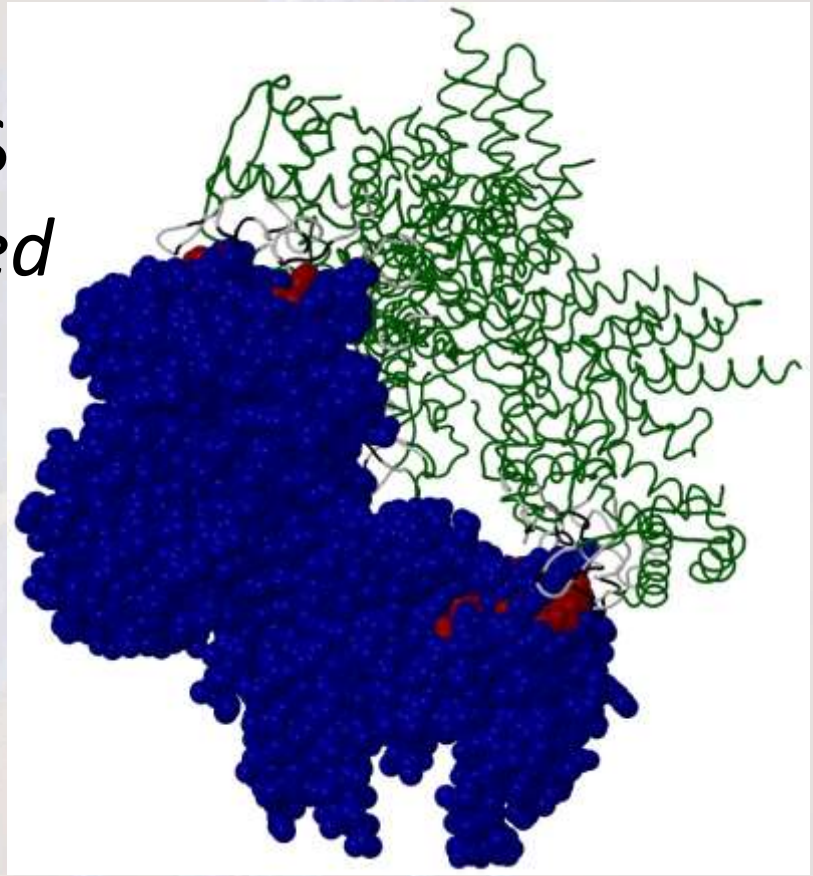
# Structures help interpret genetics

Never forget: a molecule is a 3-dimensional object. Genetics, Genomics, Transcriptomics, etc., are indirect ways to learn what molecules are doing.

Percent Amino Acid Identity and other Sequence Alignment Scores are indirect proxies for the much more complicated physics of molecules binding to each other.

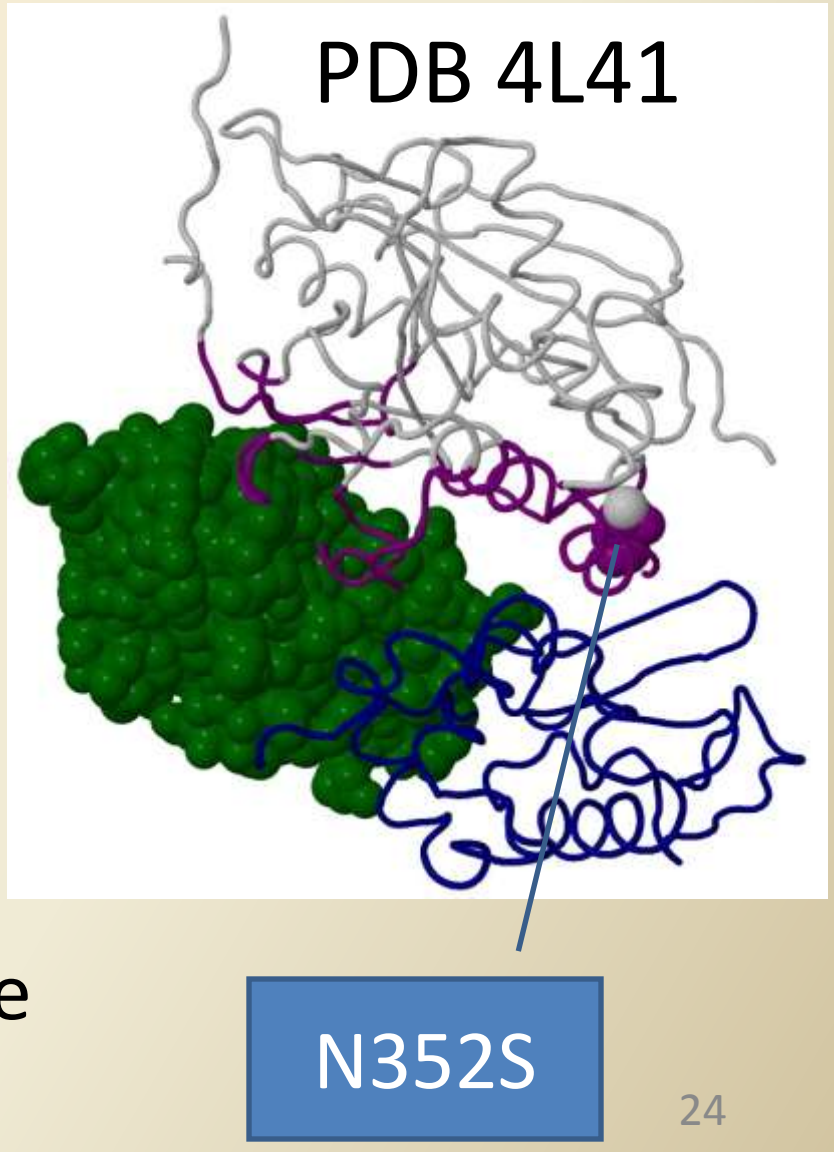
# Influenza NP: a novel MOA

- PDB structure from our PNAS paper shows *tetramer induced by our compound*
- Only the monomer structure was known before our work
- Red highlights resistance mutations, which puzzled us because there were two distinct places on flat hydrophilic surface



# β-1,4-galactosyltransferase 1: novel LDL association

- Presented at ASHG 2017
- Structure: complex with two lactalbumin molecules
- 4725 Amish exomes: N352S in B4GALT1 ≈6%; very rare in other pops
- Assoc ↓ LDL, Fibrinogen,
- ↑ AST, Creatinine Kinase
- Near both active site and one of the lactalbumins



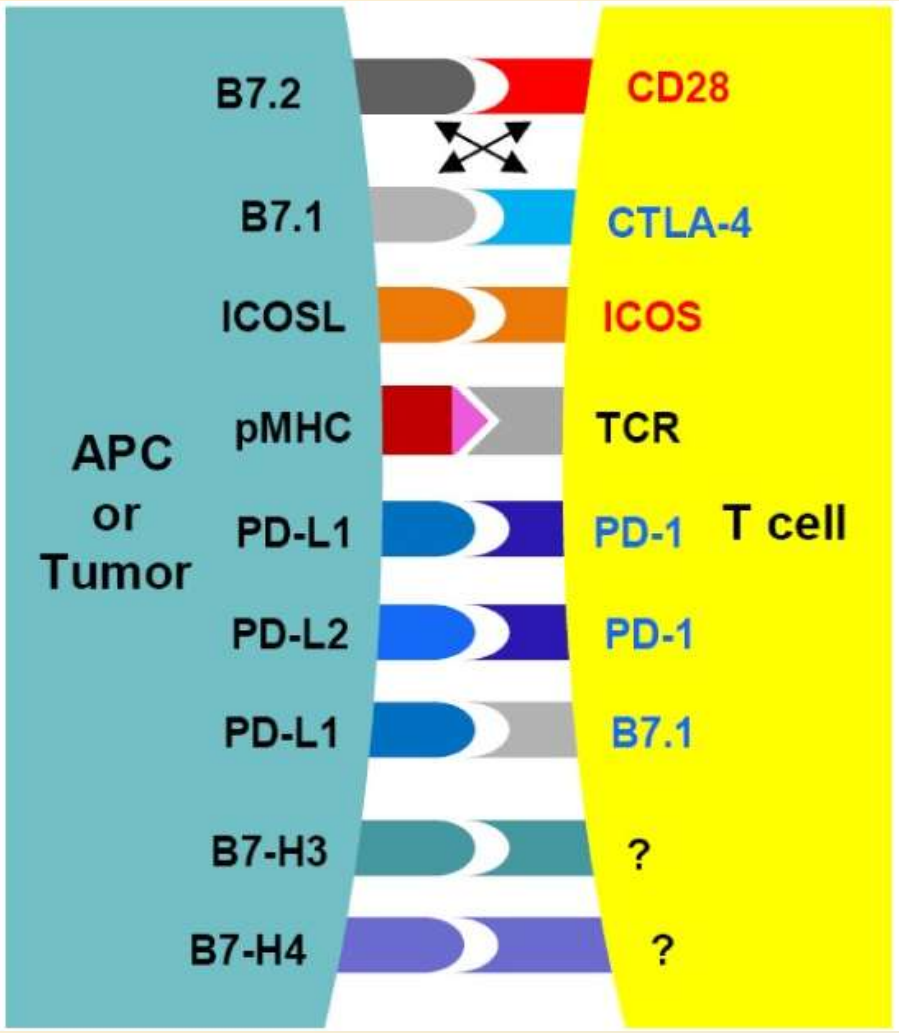
# Not every clever idea pans out

1. Common cold is caused by many different viruses that have evolved the same habit: cause a mild, self-limiting, upper respiratory infection
2. Symptoms are almost entirely due to immune response, which we do not in fact need to clear the virus
3. We did transcriptional profiling of human cells infected with a wide spectrum of viruses that often cause colds, along with related viruses that have different tissue tropisms
4. Hoped we'd find a signature specific to the *not needed* immune response to cold viruses. But we got signatures reflecting viral taxonomy.

# An idea that *did* work well

1. Immune modulation for Virology
   - Major Immuno-virology effort
   - One part was leveraging BMS Immuno-Oncology assets for chronic viral infections
2. Small molecule immune modulators
   - T-Cells activated by beads coated stimulatory and co-stimulatory signals
   - HTS screen sought compounds that increased T-cell response to stimulation *but had minimal effect* by themselves

# Key Immune Checkpoint Signals




Marketed BMS products:

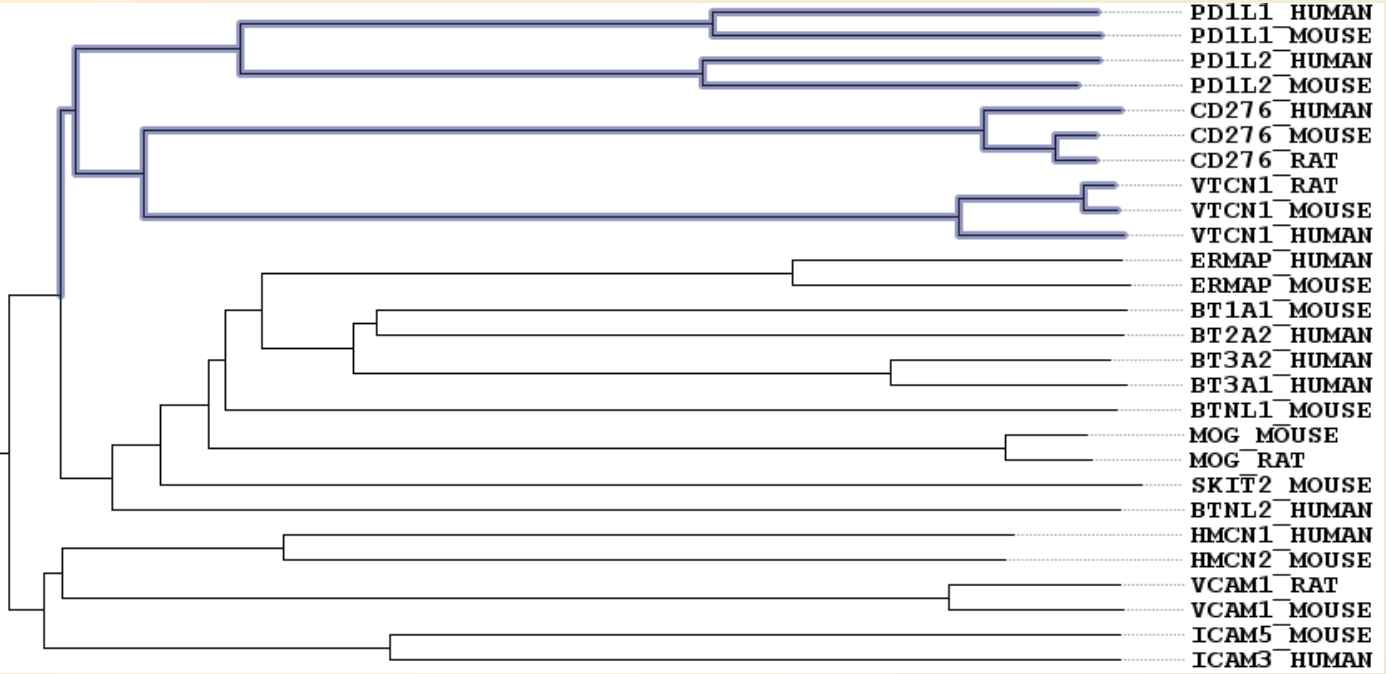
Ipilimumab (Yervoy)
Increases T-cell activation by blocking CTLA-4 signaling

Nivolumab (Opdivo)
Increases T-cell activation by blocking PD1 signaling

Abatacept (Orencia)
belatacept (Nulojix)
Decrease T-cell activation via CTLA-4 stimulation

- Leung & Suh, 2014 http://www.ncbi.nlm.nih.gov/pubmed/25550693

# PDL1/L2 phylogenetic tree

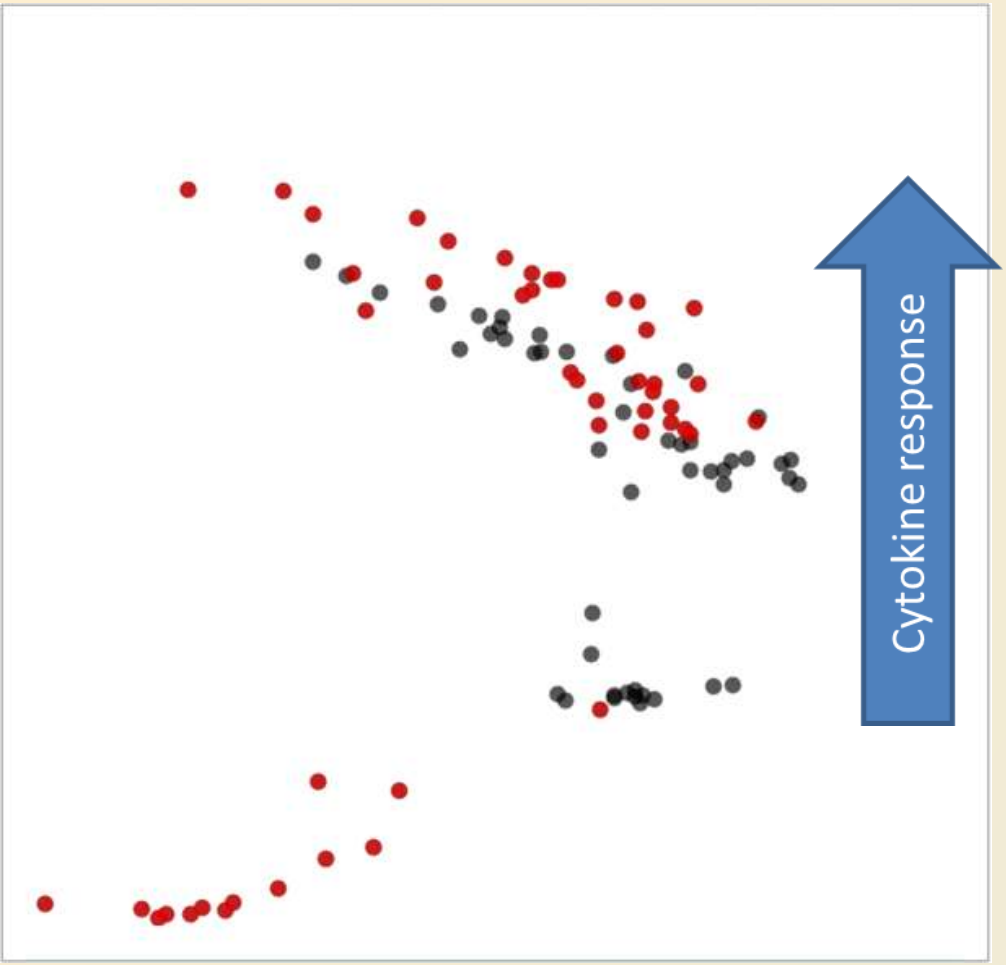



Phylogenetic analysis of model organism sequences, to verify orthology calls. In the case of PD1, PDL1, and PDL2 the orthology calls are simple and clearly correct, but it is good practice always to check.

For some large immune-related gene families, orthology versus paralogy is **much more challenging to call with confidence.**

Note on gene names used in various databases:
PDCD1 = PD1 = CD279 PD1L2 = PDL2 = PDCD1LG2 = CD273
PD1L1 = CD274 = B7-H1 = B7DC = PDCD1LG2 CD276 = B7H3 VTCN1 = B7H4

# Small Molecule Immunomodulation



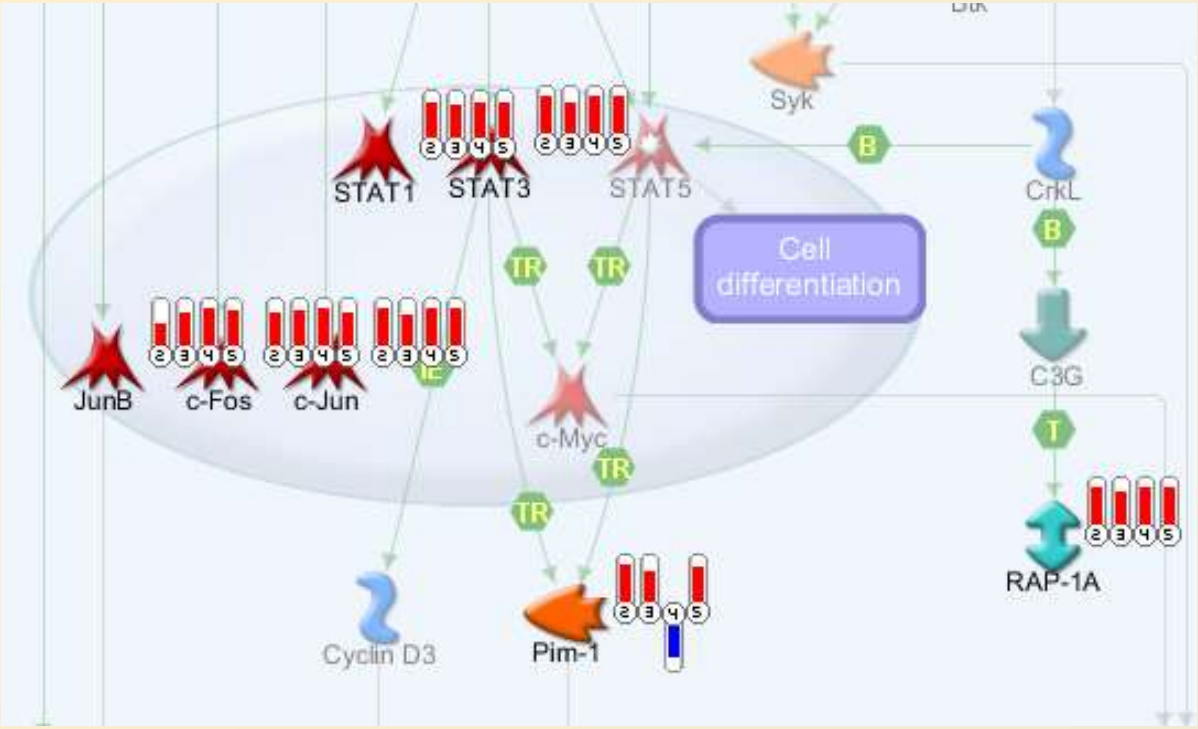

Compound synergizes with cytokine stimulation of primary immune cells **but has little effect alone**

Dose-response Affymetrix study: CD3/CD28 stimulation of primary T-cells +/- cpd

Color = timepoint

**PCA Plot of Affy Data: point = sample**

# Pathway Tools for MOA




Transcriptional profiling study: treated T-cells with compounds that activate latent Human Immunodeficiency Virus.

Above is from GeneGo, a major commercial pathway tool. Bar charts depict responses to several different treatments. MDH was the primary analyst on this project and co-designed the study.

The main algorithm we used to identify key pathways was Gene Set Enrichment Analysis and related algorithms such as the hypergeometric test.

# Key Lessons

- The fundamentals of statistics and sequence analysis do not change, but how they are applied must evolve as wet-lab technology does: newer technology yields much bigger volumes of data but also *different patterns of errors*
- Algorithms useful if and only if they yield biological insights
- People find surprising uses for a really great tool
  - Often these uses entail combining with other tools
  - Users will vary widely in their level of experience
  - Therefore, tools must support multiple use cases
    - Helpful for interactive users
    - Convenient for pipeline builders