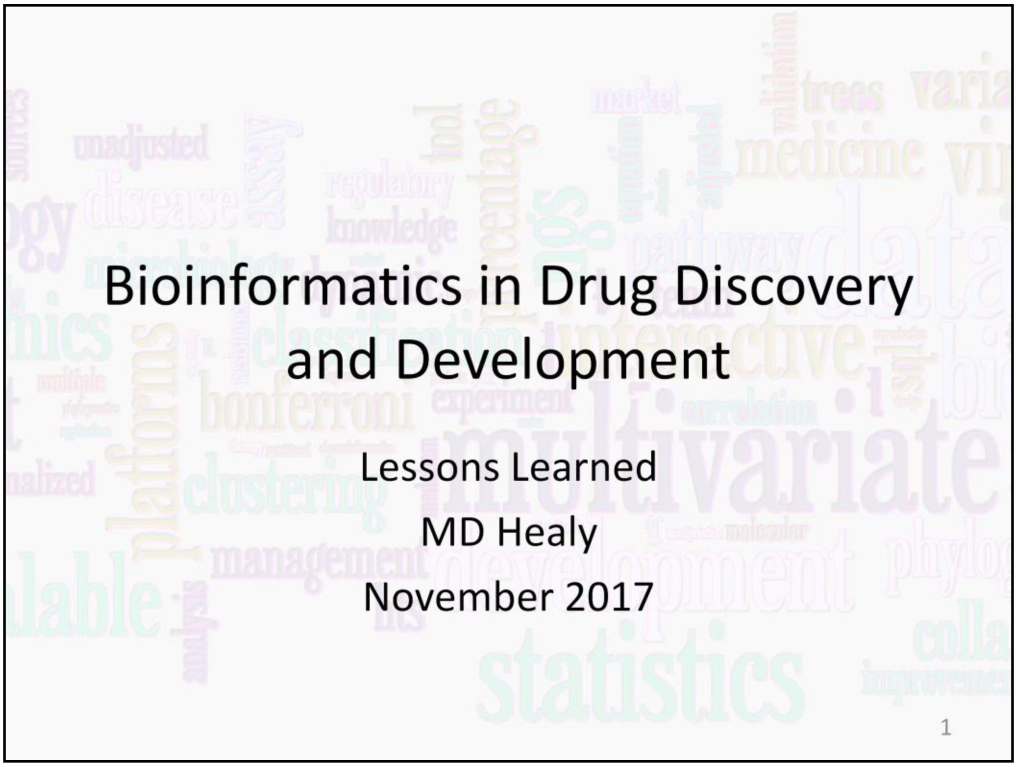


Bioinformatics in Drug Discovery and Development

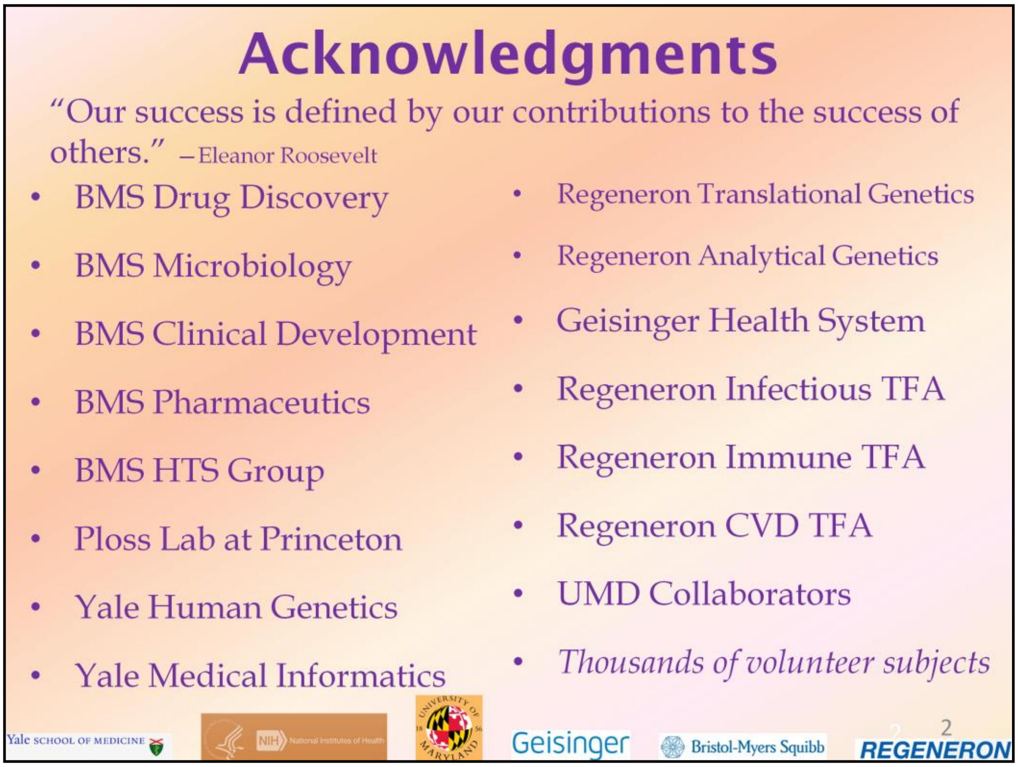
Lessons Learned

MD Healy

November 2017

1

Thanks for inviting me to talk about my research and what I’ve learned from it. Most of the work I’ll present has been published; in just a few cases I’ll have to leave out details that haven’t been published.



Acknowledgments

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

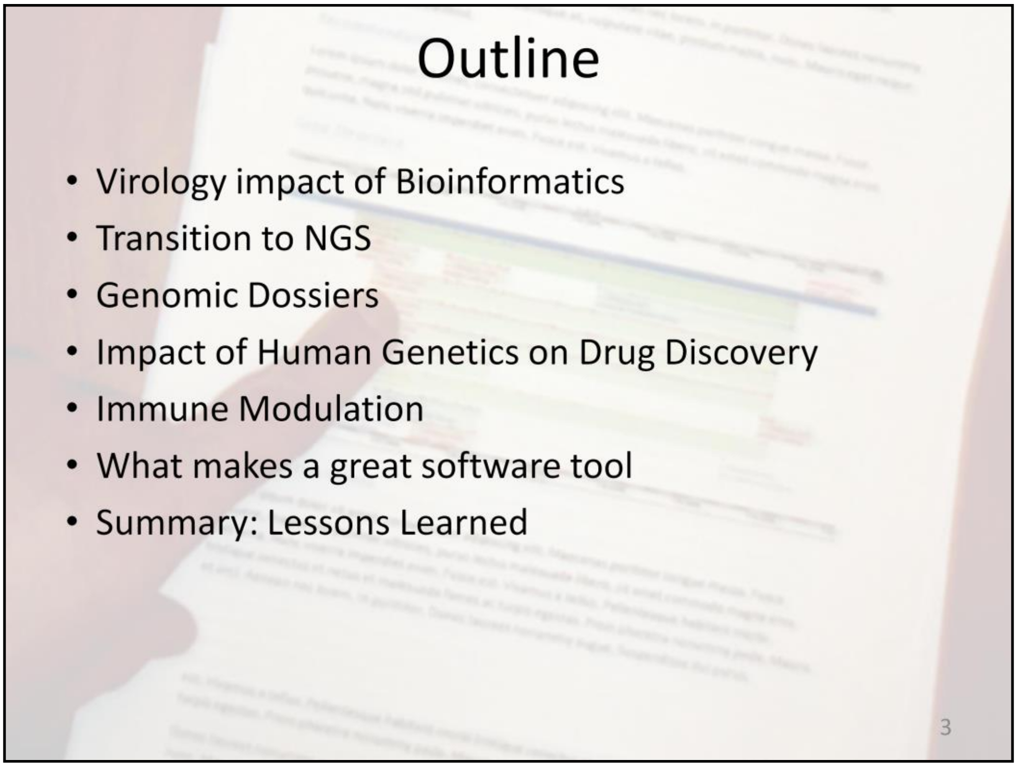
- BMS Drug Discovery
- BMS Microbiology
- BMS Clinical Development
- BMS Pharmaceutics
- BMS HTS Group
- Ploss Lab at Princeton
- Yale Human Genetics
- Yale Medical Informatics
- Regeneron Translational Genetics
- Regeneron Analytical Genetics
- Geisinger Health System
- Regeneron Infectious TFA
- Regeneron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

Yale SCHOOL OF MEDICINE NIH National Institutes of Health UNIVERSITY OF MARYLAND Geisinger Bristol-Myers Squibb REGENERON

2

Thanks for inviting me to speak! While I briefly introduce myself, I'll leave this slide up. I know people usually put acknowledgements at the end, but I want to emphasize the many great people with whom I have collaborated, because Drug Discovery and Development truly is a team sport. Those of you who have seen my CV will have noticed that my patent and my publications have lots of coauthors. I left the wet lab quite a while ago, which means nothing I do can make a difference to patients UNLESS IT HELPS WET LAB AND CLINICAL TEAMS DO THEIR JOBS. Therefore, the single most important part of my job has always been listening to them so as to learn what they need.

In my years at BMS, nearly all the Genomics and Bioinformatics people were in NJ. We never had more than a few people in CT, and sometimes I was the only Bioinformatics person in CT. I've worked on small molecule and biological Discovery and Development programs, and been on Biomarker Working Groups, in many TAs including Neuroscience, Genetically-Defined Diseases, Cardiovascular, Immunology, and Oncology, but for many of my years at BMS a big chunk of my time was basically "whatever Virology needs from Genomics," which covered a lot of ground! Of course I can hardly know everything in such a wide range of fields, but I do know which questions to ask and where to start seeking answers.



Outline

- Virology impact of Bioinformatics
- Transition to NGS
- Genomic Dossiers
- Impact of Human Genetics on Drug Discovery
- Immune Modulation
- What makes a great software tool
- Summary: Lessons Learned

3

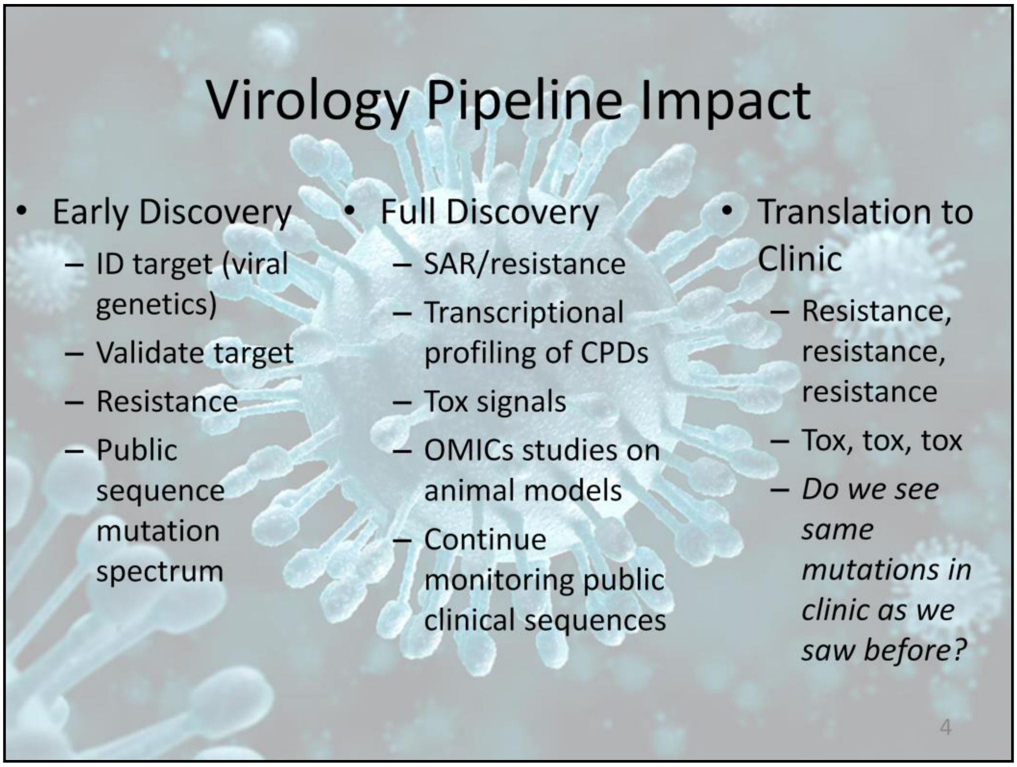
My father's dissertation adviser always said in a talk you should "tell 'em what you're gonna tell 'em, then tell 'em, then tell 'em what you just told them." So here's what I'm gonna tell you…

General points not given specific mention elsewhere that I'll verbally mention now:

1.When designing any OMICS study, whether experimental, computational, or both, **we must ask what decision are we trying to make and how will each possible result affect that decision**? If the answer is, the decision will be the same either way, then why are we planning to do the study?

2.We always want to maximize statistical power, but dollars are never infinite. At BMS we often spoke of "cost-disciplined science." Which endpoints do we capture? Which combinations of experimental factors do we include? How deeply do we analyze each experimental unit? For instance Genometry L1000 vs Affy Chips vs RNA-Seq. SNP chips vs Exome Sequencing vs Genome Sequencing. I've heard from several sources that many organizations are returning to SNP chips for GWAS because for rare variants they need huge numbers of subjects and cannot afford to sequence them all!

3.Failure to anticipate future uses of clinical samples can cause big issues if the consent forms signed by your volunteers don't allow for something you later want to do. Even if they are willing to sign additional consent forms, the logistics of sending somebody to get their signatures will cost money and time.



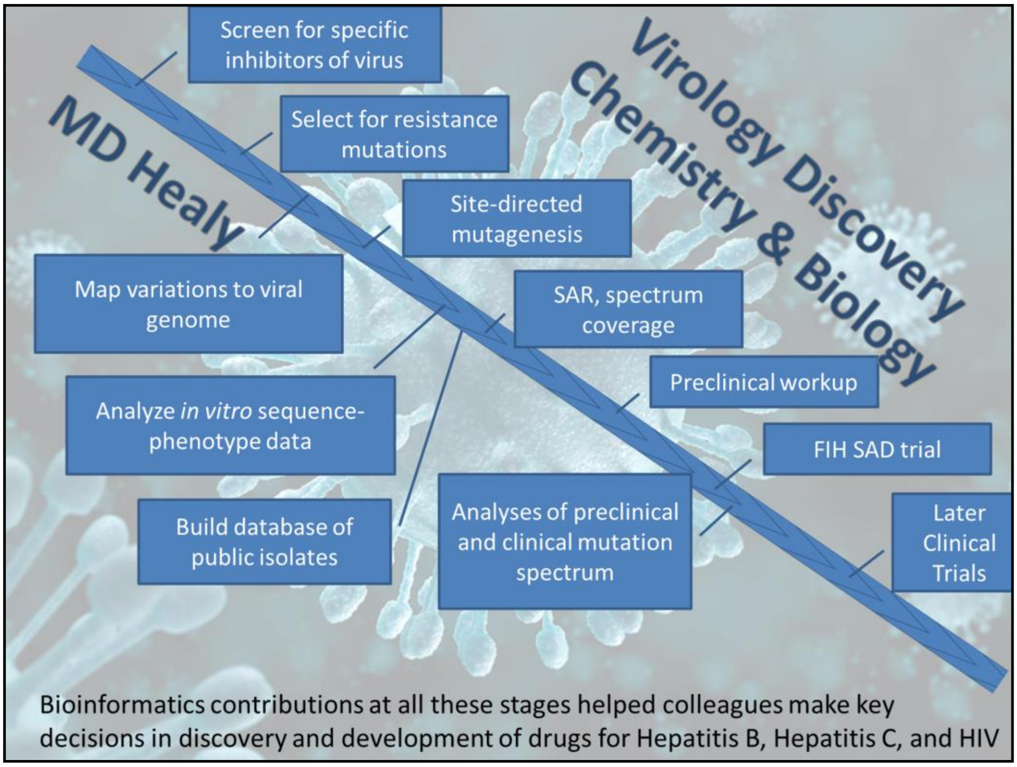
Virology Pipeline Impact

- 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?*

“Forward” starts from phenotype, tries to work out genetics.
“Reverse” starts from genetics, tries to work out phenotypic consequences

Virology approach often begins with a phenotypic HTS screen to get small molecule selective inhibitors of the targeted virus. Then we used hit compounds to select for resistant viruses. By sequencing the resistant viruses, we formed hypotheses about target, which we confirmed by site-directed mutagenesis.

Do you call that “forward” or “reverse” or what?



Virology Discovery Chemistry & Biology

MD Healy

Screen for specific inhibitors of virus

Select for resistance mutations

Site-directed mutagenesis

SAR, spectrum coverage

Preclinical workup

FIH SAD trial

Later Clinical Trials

Map variations to viral genome

Analyze *in vitro* sequence-phenotype data

Build database of public isolates

Analyses of preclinical and clinical mutation spectrum

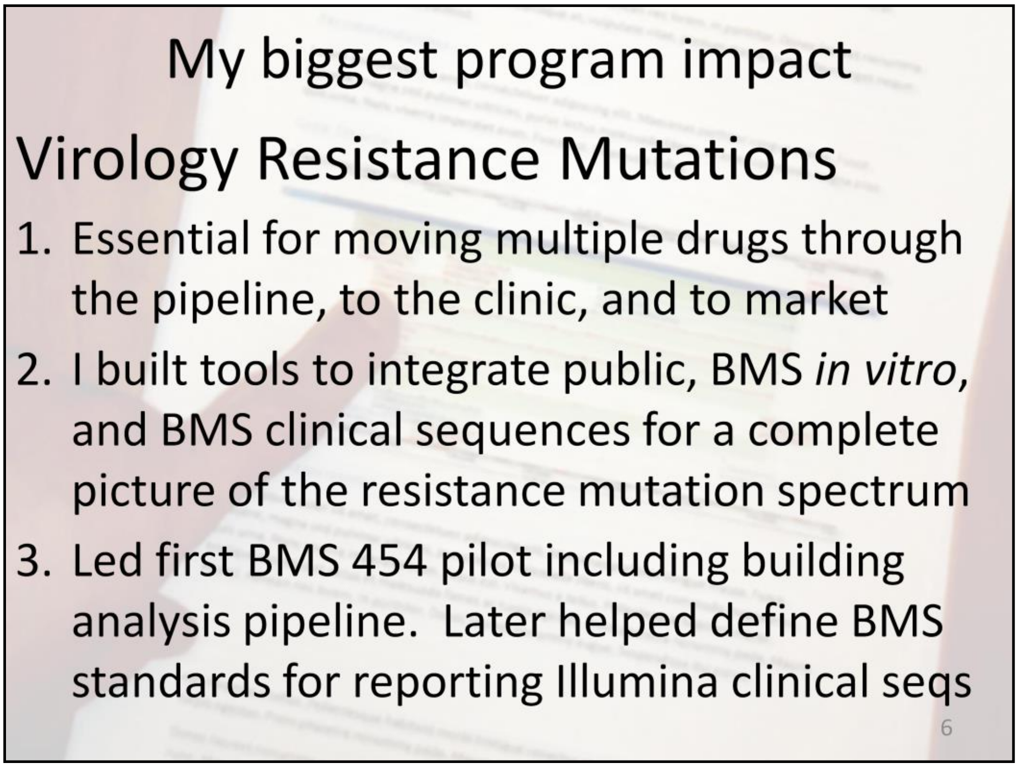
Bioinformatics contributions at all these stages helped colleagues make key decisions in discovery and development of drugs for Hepatitis B, Hepatitis C, and HIV

Bioinformatics Impact on Virology Programs
Background image: HCV

Key decisions included: which mutations to require compounds must cover and when criteria for progressing to next stage had been met. Some analyses were done under very tight timelines because the go/no-go decision for the next clinical trial hinged on the results.

Marketed drugs on which my work has had a significant impact include: ATRIPLA® (efavirenz, emtritricabine, & tenofovir fixed-dose combination for HIV), BARACLUDE® (entecavir for Hepatitis B), DAKLINZA™ (daclatasvir for Hepatitis C), SUSTIVA® (efavirenz for HIV), & SUNVEPRA™ (asunaprevir for Hepatitis C). Several of these have sold well over a Billion dollars per year.

Investigational drugs now in the clinic on which my work has had an impact include: an HIV Maturation Inhibitor, an HIV Attachment Inhibitor, and an HCV NS5B Non-Nucleoside inhibitor (beclabuvir, *now approved in Japan*) (also other HIV compounds that are close to first-in-human trials). Following its recent decision to end all Virology Discovery Research at the end of 2015 and to close the entire Wallingford site by the end of 2018, BMS has sold its entire portfolio of Investigational HIV compounds to ViiV in a transaction that could exceed Two Billion Dollars if certain milestones are met.

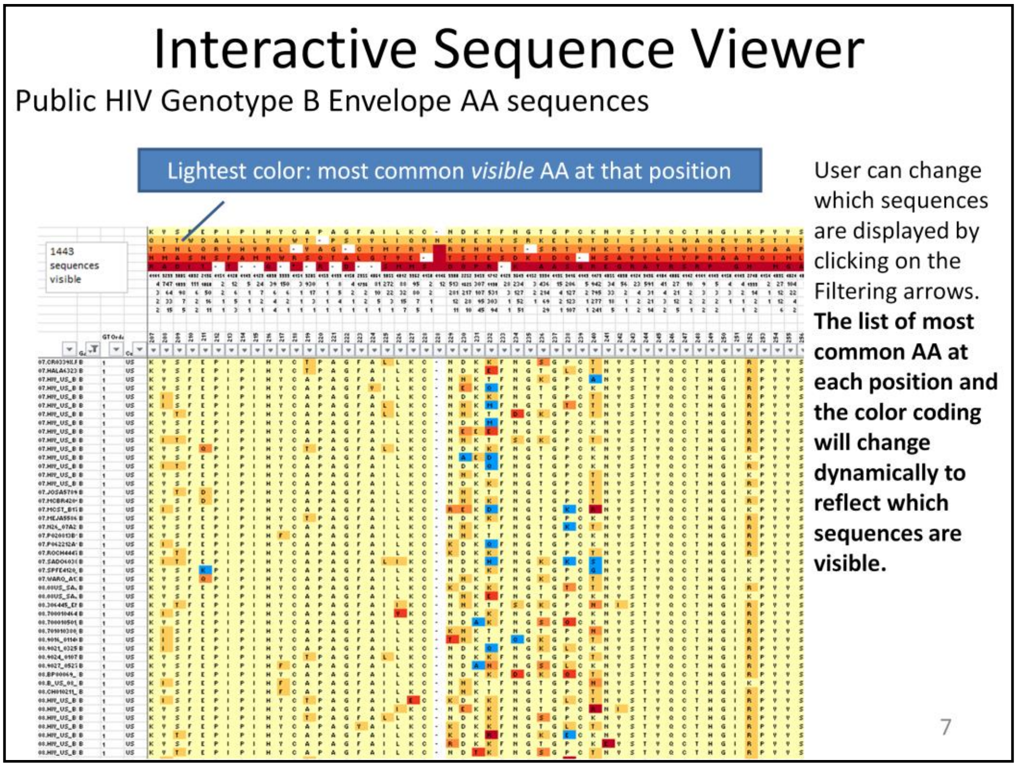


My biggest program impact

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

I'll say more about 454 a few slides later...



Interactive Sequence Viewer

Public HIV Genotype B Envelope AA sequences

Lightest color: most common *visible* AA at that position

User can change which sequences are displayed by clicking on the Filtering arrows. **The list of most common AA at each position and the color coding will change dynamically to reflect which sequences are visible.**

7

Example of how I had impact on Virology: by making tools like this one for them

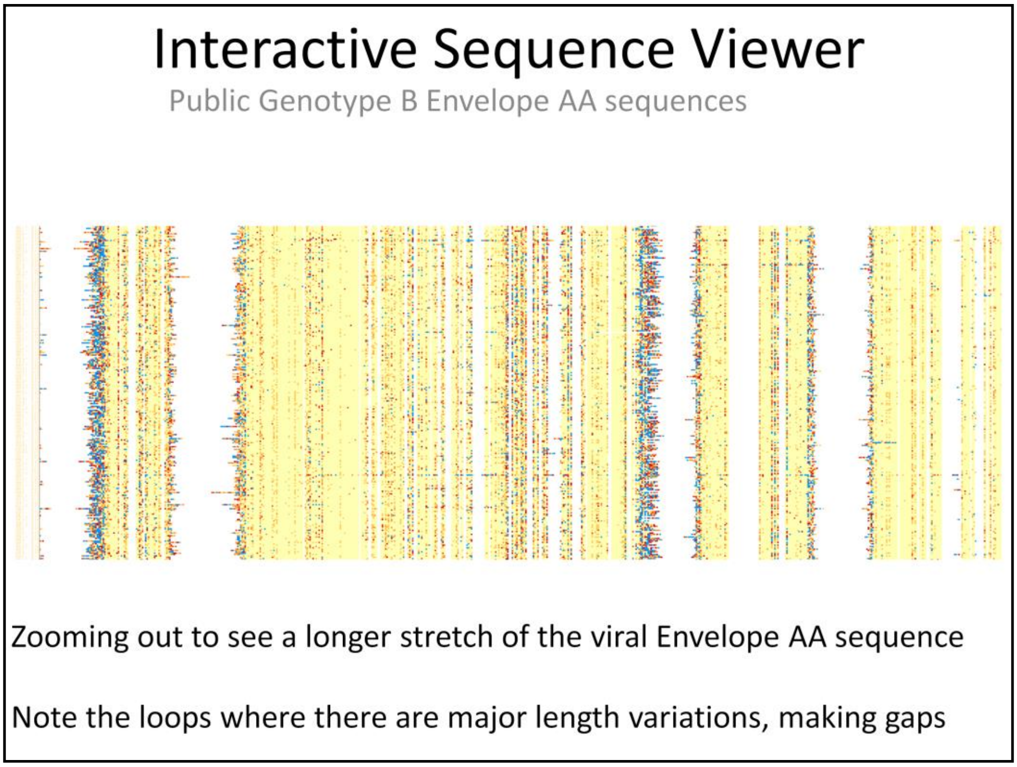
I invented this approach because existing tools didn’t meet their needs.

Wet-lab biologist users **love** these interactive Excel viewers, because they are all comfortable using MS Excel to filter and view data. This example uses public sequences; I also made many of these with BMS Preclinical and Clinical sequences. **And for Human genetics I built a different type of Interactive Excel tool, to facilitate exploring genetic associations with *Interactive Manhattan plots*.**

Analysis pipeline in brief:

1.Translated alignments to reference AA sequences using a frame-maintaining algorithm such as FrameSearch, Estwise, or Exonerate
2.Checked alignments and fixed blunders (such as overlapping viral ORFs confusing the aligner). **Handled mixture NT as ambiguous codons with slashes: UNIQUE value-add**
3.Generated profiles and formatted color-coded reports

I made similar tools for visualizing Pre-Clinical and Clinical sequence data. **I also did various count-based analyses** such as Hypergeometric (Fisher Exact and variants thereof) analyses looking for resistance associations. We usually accounted for multiple comparisons using the Benjamini-Hochberg FDR procedure. We also tried Random Forests and a few other approaches; since the main conclusions were quite similar I preferred to stick with the simpler approaches. Also, it’s difficult to find genetic interactions unless you have a really large number of subjects so that every possible combination is present in your data.



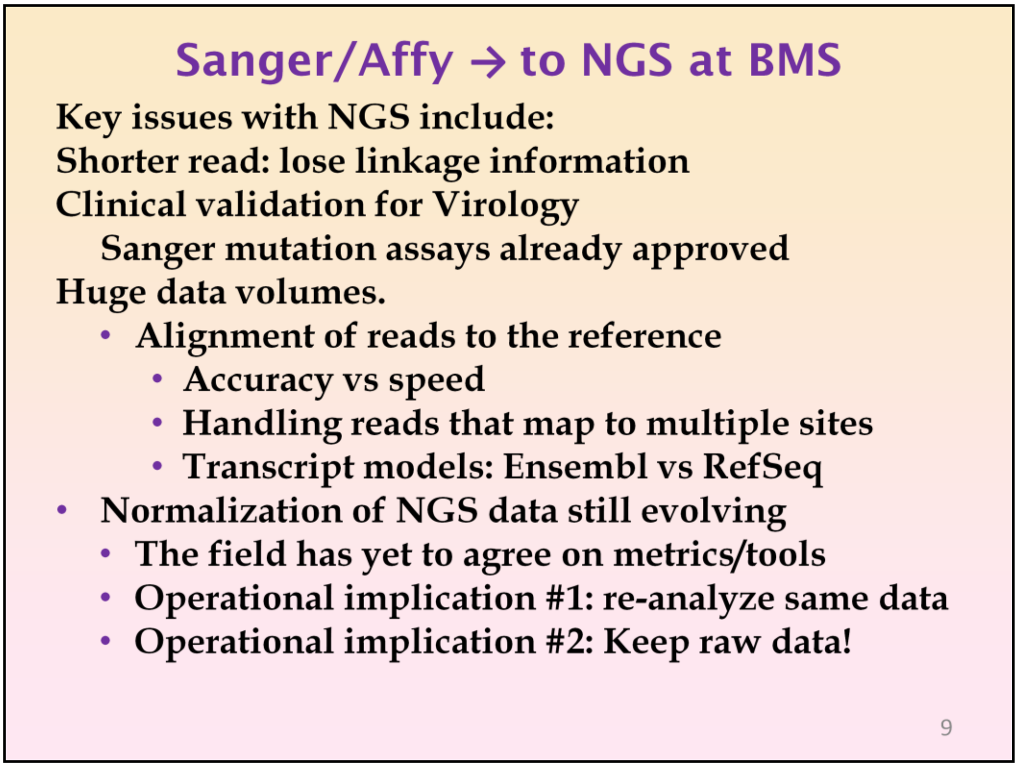
Interactive Sequence Viewer

Public Genotype B Envelope AA sequences

Zooming out to see a longer stretch of the viral Envelope AA sequence

Note the loops where there are major length variations, making gaps

Note the loops where GP160 has major length variations



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!

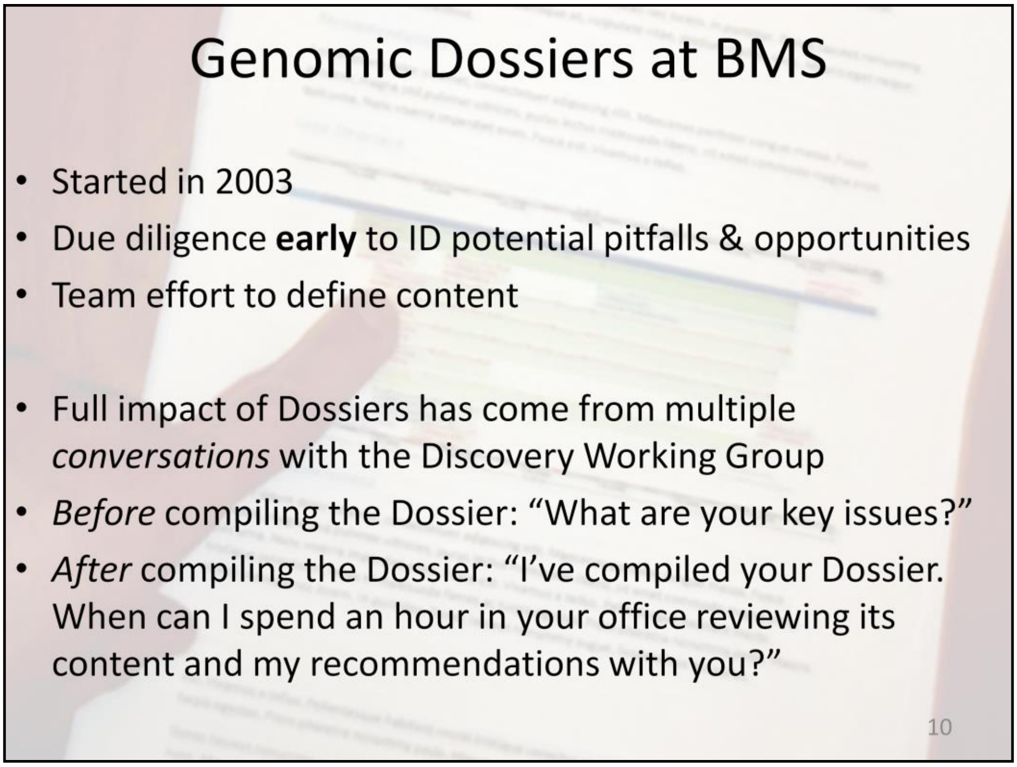
9

NCBI RefSeq strives to avoid calling splice variants that aren't real, but might miss real variants
Ensembl strives to avoid missing any real splice variants, but may call some that aren't real
Sometimes good to align reads with BOTH transcriptome models, in order to compare the results

When I took the lead in making the first BMS Pilot tests of then-new 454 Next-Gen Sequencing. our conclusion at that time was NGS wasn't yet ready for use with our Virology Clinical programs. More recently, I worked with BMS Virology Clinical colleagues to negotiate the reporting standards for NGS data on clinical samples with our vendors. The different data characteristics and error models of NGS have both advantages and disadvantages relative to Sanger for Virology Clincal samples. Biggest advantage: quantitative data on variant frequencies. Also it can be easier aligning NGS reads than Sanger reads in highly variable regions. Biggest disadvantage: short reads cannot find long-range linkages between viral polymorphisms; therefore some clonal Sanger work still needed!

A major issue of concern to FDA and which I had flagged in my analysis of the BMS data is **primer set bias**

NGS is not a drop-in replacement for Sanger. Different read length, and also different error models. I leveraged my expertise with **reading-frame-preserving alignment algorithms** when I built the analysis pipeline for BMS Pilot study of 454 because I knew that technology had a high indel rate.

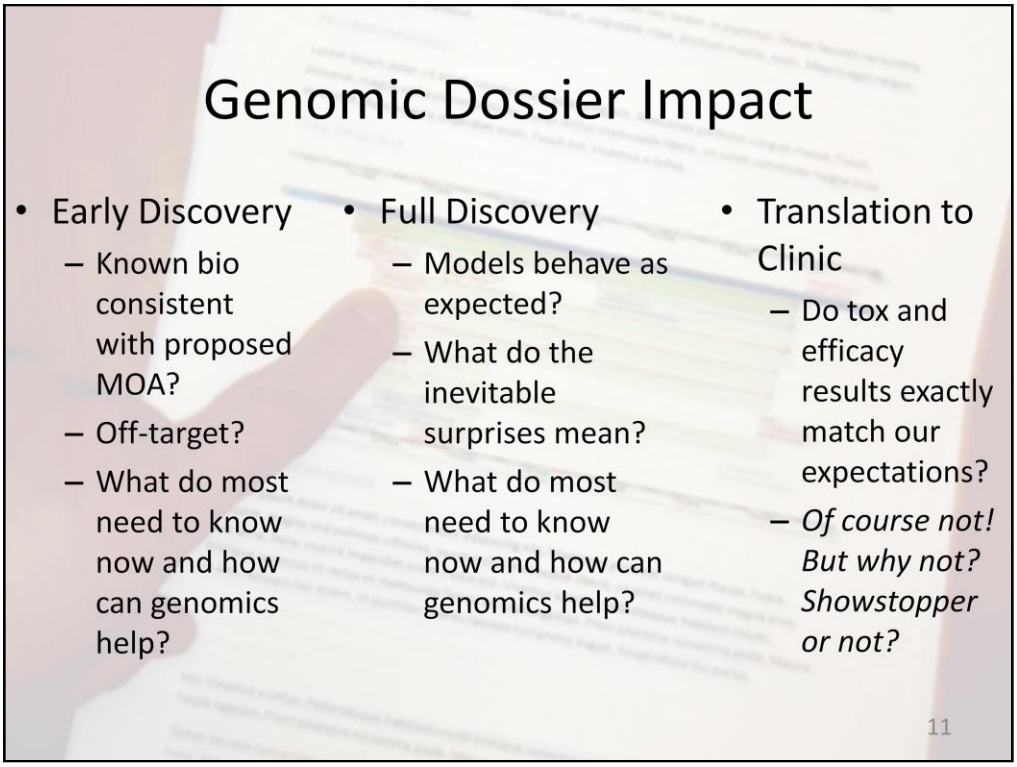


Genomic Dossiers at BMS

- Started in 2003
- Due diligence **early** to ID potential pitfalls & opportunities
- Team effort to define content

- Full impact of Dossiers has come from multiple *conversations* with the Discovery Working Group
- *Before* compiling the Dossier: “What are your key issues?”
- *After* compiling the Dossier: “I’ve compiled your Dossier. When can I spend an hour in your office reviewing its content and my recommendations with you?”

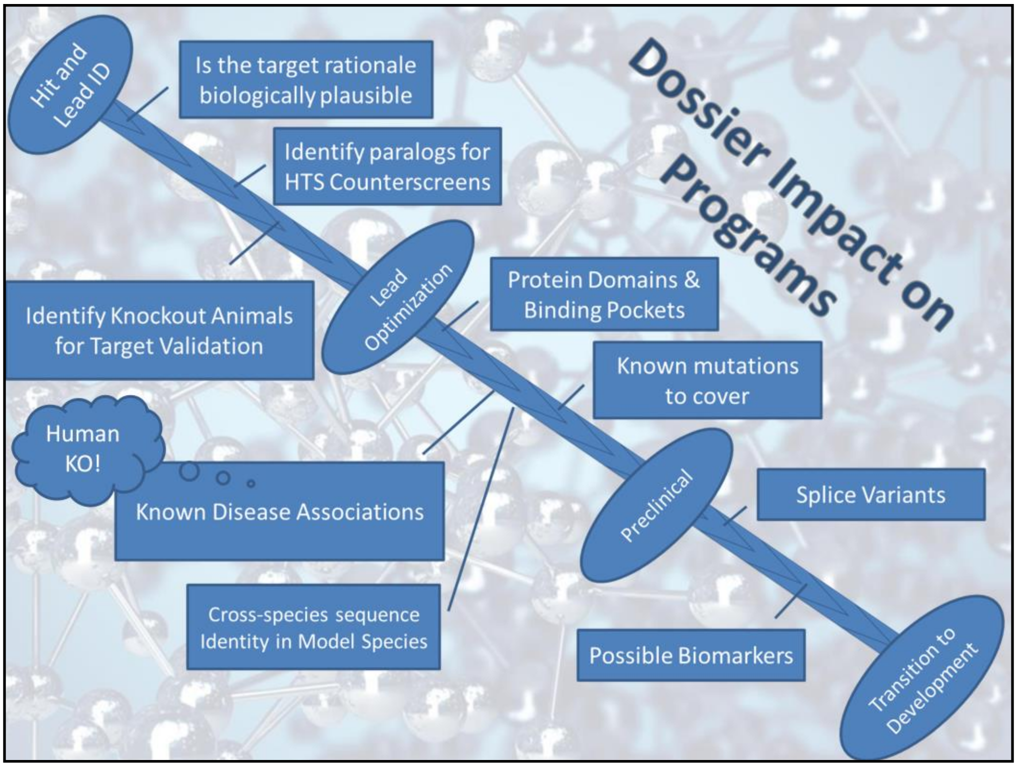
Dossier as product versus dossier as process!



Genomic Dossier Impact

- 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?*

Dossier as product versus dossier as process!



Dossier Impact on Programs

Hit and Lead ID

Is the target rationale biologically plausible

Identify paralogs for HTS Counterscreens

Identify Knockout Animals for Target Validation

Lead Optimization

Protein Domains & Binding Pockets

Known mutations to cover

Human KO!

Known Disease Associations

Preclinical

Splice Variants

Cross-species sequence Identity in Model Species

Possible Biomarkers

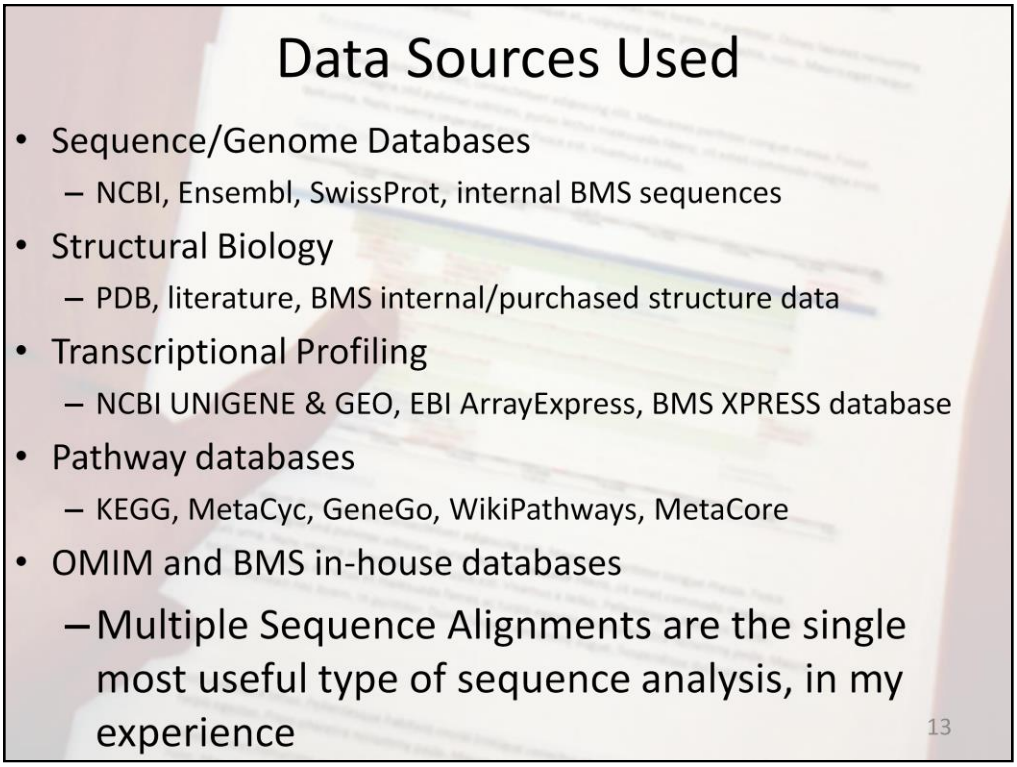
Transition to Development

Outside of Virology, where Bioinformatics has always been deeply involved in all stages through the clinic, Bioinformatics historically has focused mostly on Discovery. More recently with the advent of Precision Medicine, Bioinformatics has come to have more involvement with Development for areas outside of Virology.

Every human target at BMS since 2003 has had a Dossier. I'd estimate I personally did around 40 of them over the years, representing perhaps 15% of all Dossiers compiled at BMS from 2003 through 2015, some for programs with which I otherwise had limited involvement. For Virology, we did not use the Dossier format; instead I was embedded in all the Working Groups as a full member of the teams. The Dossier format was something of an impedance mismatch with Viral targets because it was designed for mammalian targets.

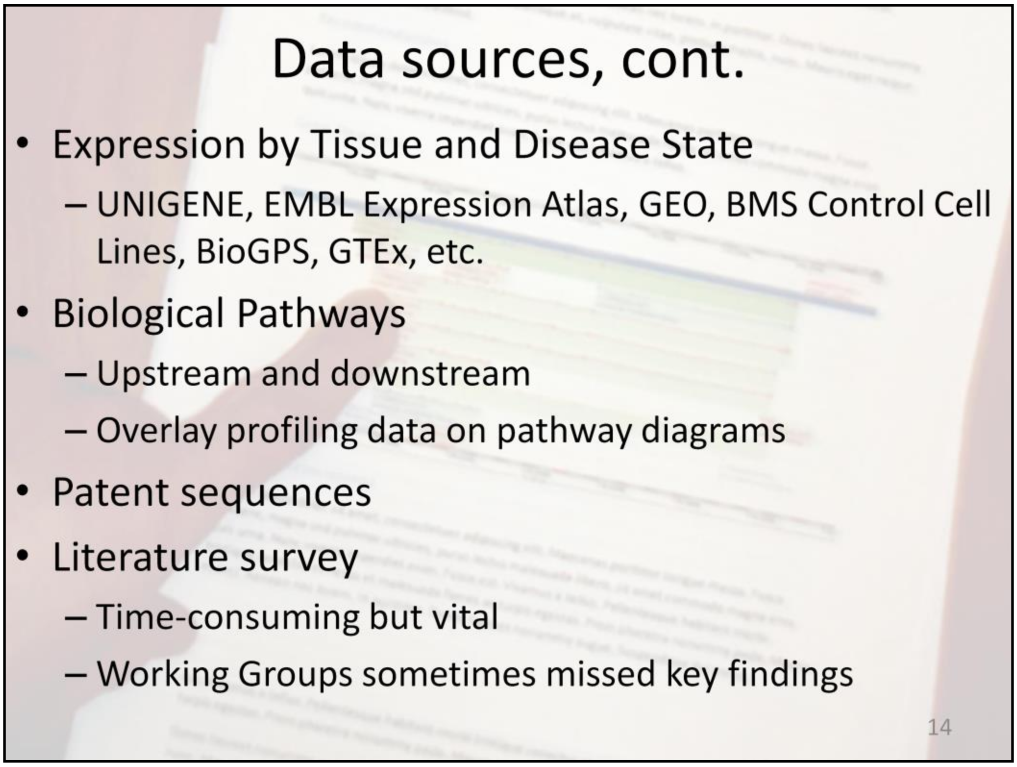
HTS Counterscreens: for small-molecule discovery, we typically began with a High Throughput Screen. The Counterscreen was to filter out compounds that weren't selective. Paralogs represent potential off-target activity.

One of the main things I've learned about Human Genetics is how Human Genetic data, especially homozygous LOF ("human KO") can *complement* model system and cell culture OMICS data.



Data Sources Used

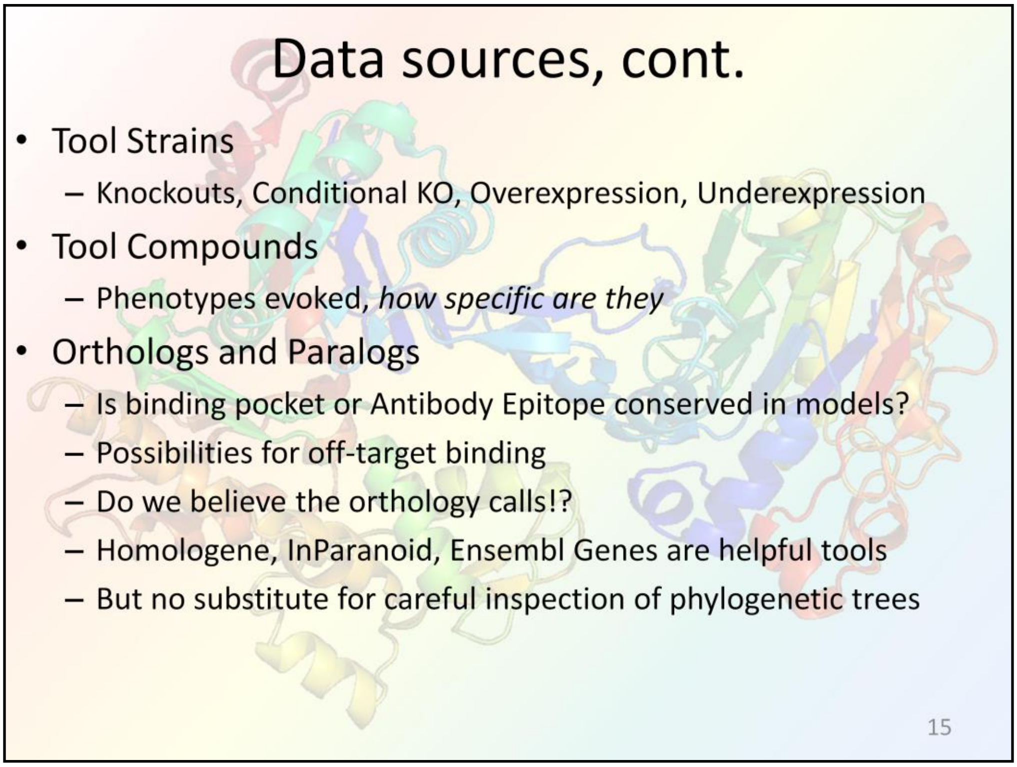
- Sequence/Genome Databases
  - NCBI, Ensembl, SwissProt, internal BMS sequences
- Structural Biology
  - PDB, literature, BMS internal/purchased structure data
- Transcriptional Profiling
  - NCBI UNIGENE & GEO, EBI ArrayExpress, BMS XPRESS database
- Pathway databases
  - KEGG, MetaCyc, GeneGo, WikiPathways, 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

Patent analysts told Bioinformatics, “only state facts such as that several sequences similar to this gene were found in these patents; leave interpretation to **us**”

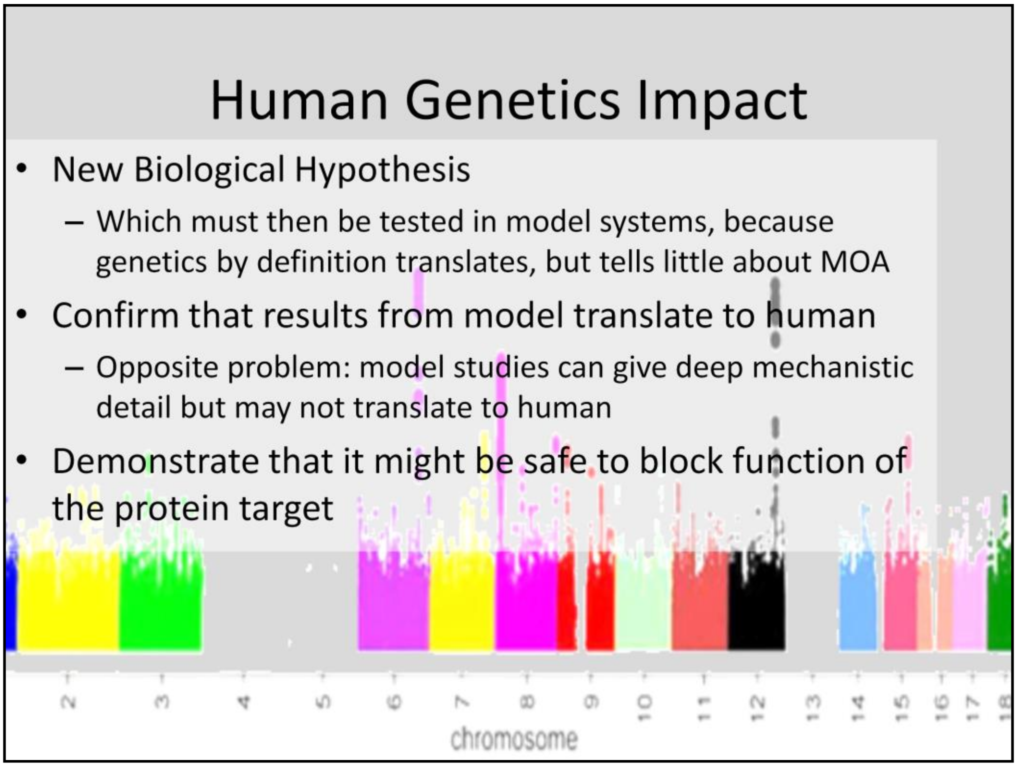


Data sources, cont.

- 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

15

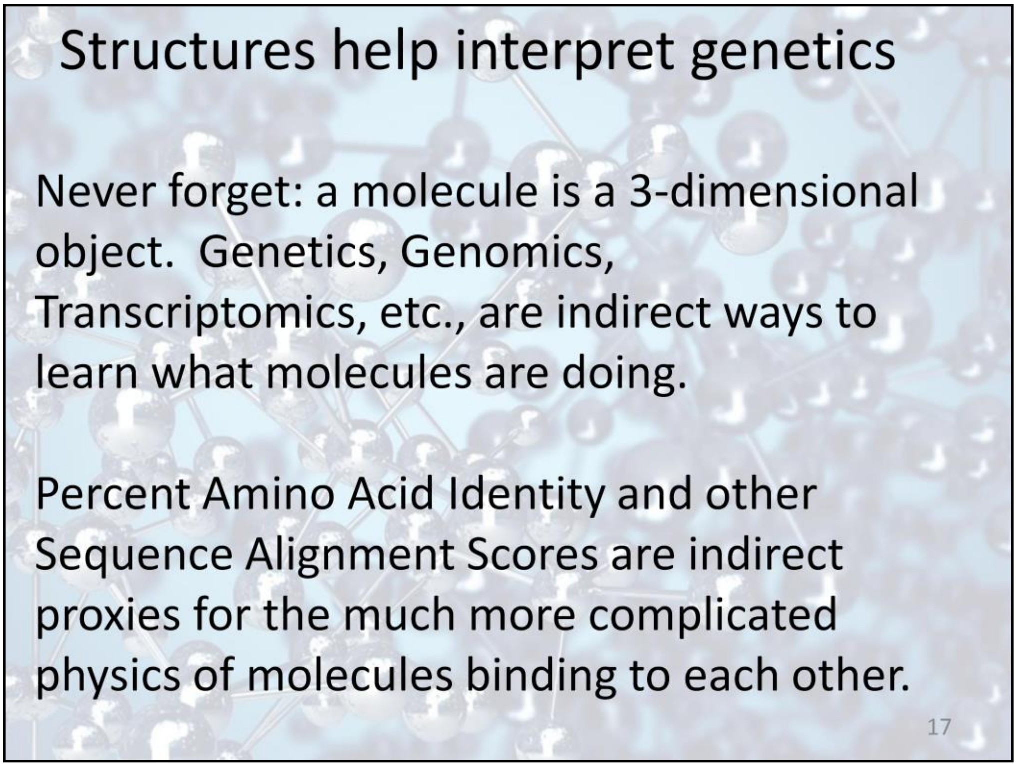
Usually the Working Group has identified proteins for use in their HTS Screening tree based on the literature. Generally, they do a reasonably good job this way, but fairly frequently the Bioinformatics analyst compiling the Genomic Dossier will identify several related proteins that should be added to the screening tree.



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

I made many Manhattan Plots of Genetic Association results, but all of them are proprietary to my employers, so the background of this slide is one I took from Wikipedia.



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.

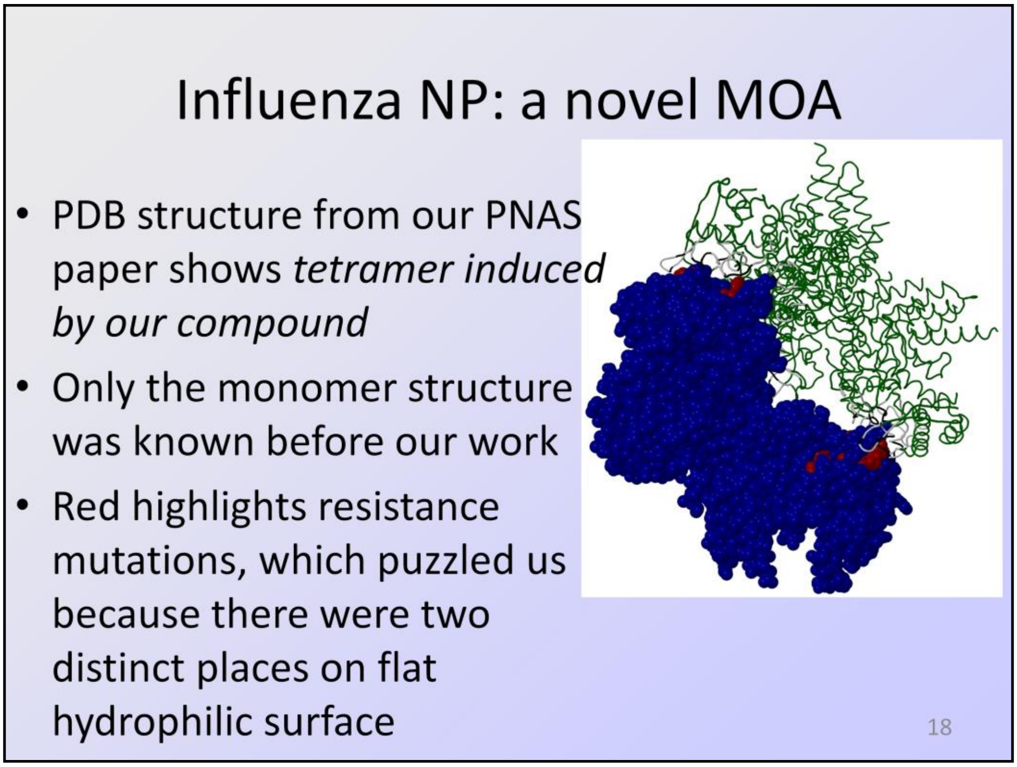
Percent Amino Acid Identity and other Sequence Alignment Scores are indirect proxies for the much more complicated physics of molecules binding to each other.

Outside of Virology, where Bioinformatics has always been deeply involved in all stages through the clinic, Bioinformatics historically has focused mostly on Discovery. More recently with the advent of Precision Medicine, Bioinformatics has come to have more involvement with Development for areas outside of Virology.

Every human target at BMS since 2003 has had a Dossier. I'd estimate I personally did around 40 of them over the years, representing perhaps 15% of all Dossiers compiled at BMS from 2003 through 2015, some for programs with which I otherwise had limited involvement. For Virology, we did not use the Dossier format; instead I was embedded in all the Working Groups as a full member of the teams. The Dossier format was something of an impedance mismatch with Viral targets because it was designed for mammalian targets.

HTS Counterscreens: for small-molecule discovery, we typically began with a High Throughput Screen. The Counterscreen was to filter out compounds that weren't selective. Paralogs represent potential off-target activity.

One of the main things I learned when I did Human Genetics is that Human Genetic data, especially homozygous LOF ("human KO") is how those data *complement* model system and cell culture OMICS data.



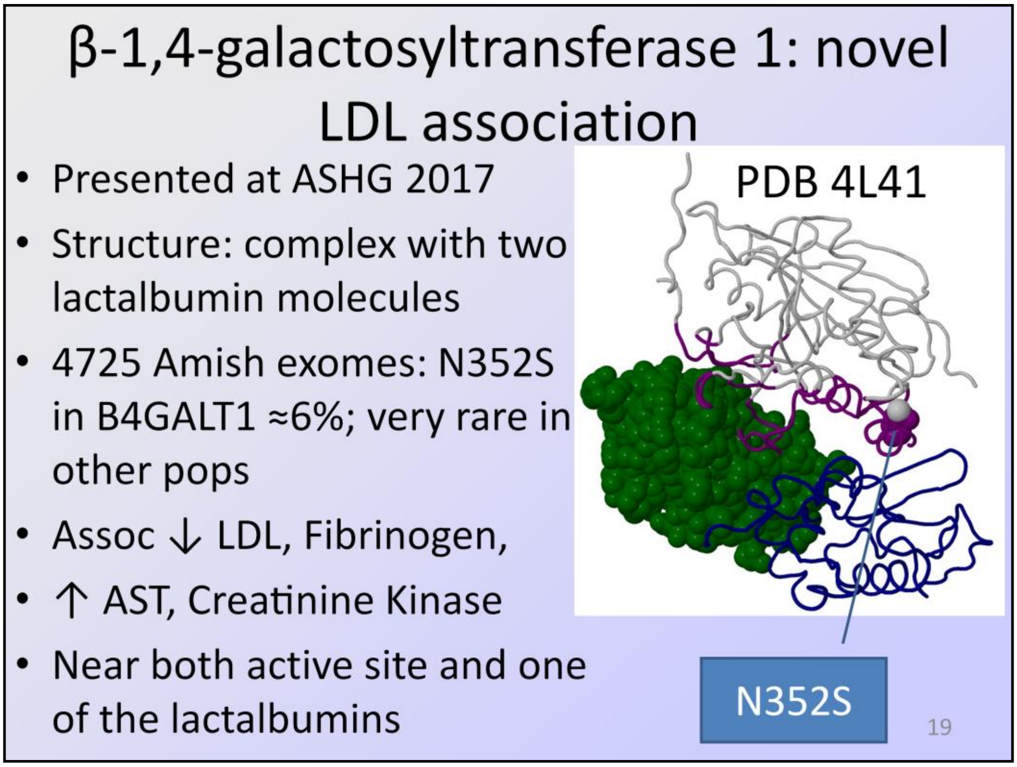
Influenza NP: a novel MOA

- 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

Emphasize: we didn’t know it was a tetramer until our colleagues in New Jersey made crystals, took them to a beamline, and built a 3d model

Here, I show two of the four subunits in space-filling style and two in trace style; I like to combine these views because each gives me different insights. Space fill helps me visualize binding pockets and interactions, while trace style helps me visualize structural features like α-helices and β-sheets. Also, only in the trace view can one see how the amino acid sequence is folded into a three-dimensional structure.

We used the same basic reverse genetics approach by which we got multiple molecules to the clinic (of which four currently on the market and others still in clinical trials). Make viral construct that has Luciferase and GFP markers added, and similar construct from a related virus. Run HTS screen with the virus of interest, and counterscreen with the related virus to ID specific inhibitors. Select for resistant strains. Map resistance hotspots. Confirm resistance substitutions by site-directed mutagenesis. Use identified targets for binding assays and SAR. Check frequency of resistance substitutions in public sequence databases and select strains to include in assays so we dial-in broad-spectrum coverage of the variation that is out there. Classic result: mutations conferring resistance to a compound cluster in a hydrophobic pocket. But we got two areas of resistance mutations on a flat hydrophilic surface of the protein!?!? So our Structural Biology colleagues made crystals, took them to a beamline, and so forth. **Novel MOA: two molecules of our compound glue two protein molecules together in antiparallel dimer which further aggregates into higher-order complexes**, thus depriving the virus of a protein it needs to replicate.



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

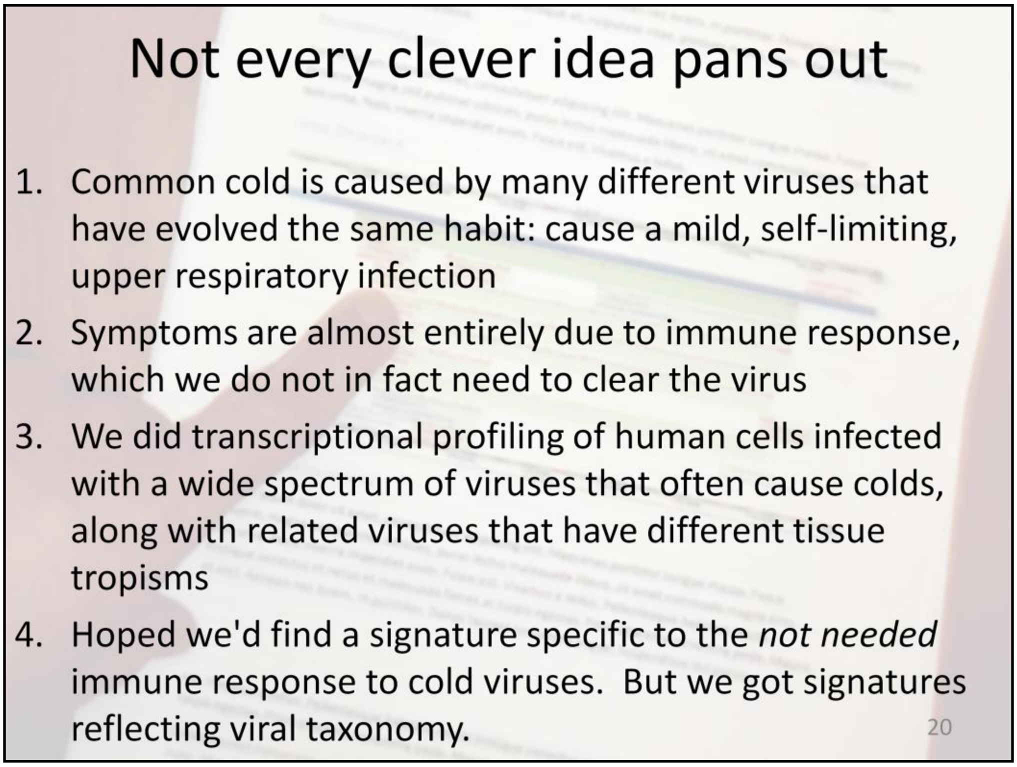
Here, I show one of the lactalbumin molecules in space-filling style and one in trace style; since this PDB structure has only one β-1,4-galactosyltransferase 1 molecule, I chose to show it in trace mode in order to see structural features like α-helices and β-sheets.

PDB 4L41

Human Lactose synthase: A 2:1 complex between human alpha-lactalbumin and human beta1,4-galactosyltransferase

One Lactalbumin chain in space fill view, one Lactalbumin chain in trace view. B4GALT1 chain in trace view, except ASN353 (N352) in space fill view. Purple color highlights residues of B4GALT1 that are within 10 Angstroms of a Lactalbumin chain.

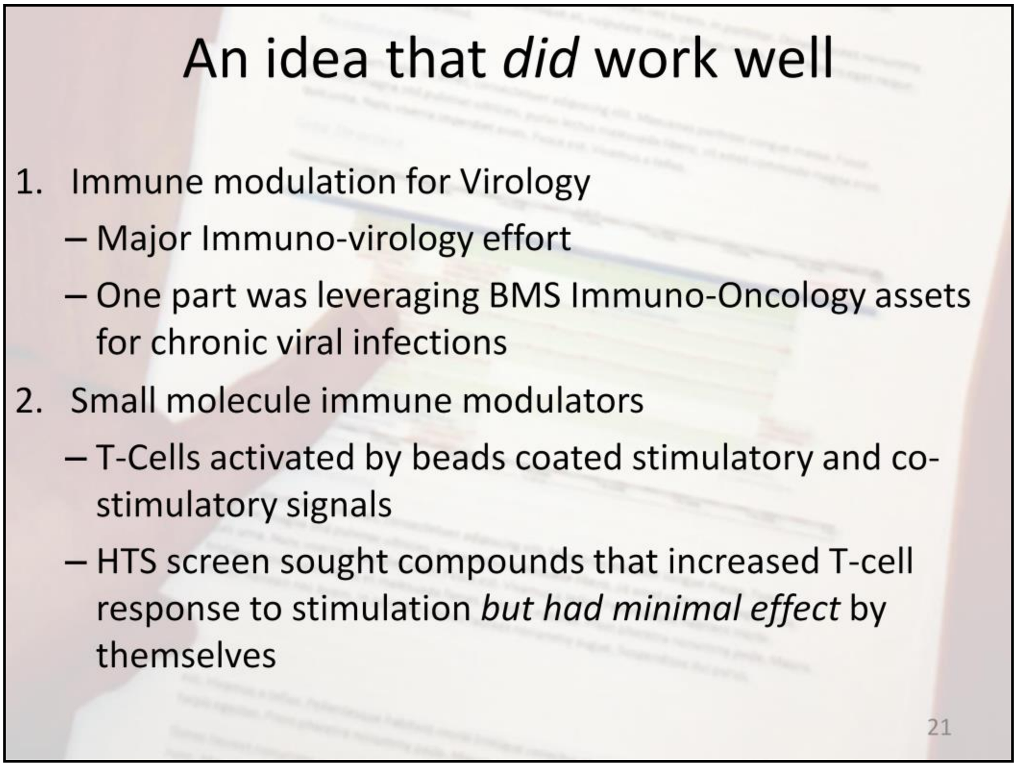
This mutation seems benign: the 13 N352S homozygotes are all apparently-healthy adults. About 15 years ago a 50AA c-terminal truncation of B4GALT1 was found to cause a severe multisystem disorder through glycosylation defects. We therefore think this mutation is likely a partial loss-of-function in this essential protein. Mechanism by which reduced glycosylation leads to lower LDL cholesterol is currently under active study at both Regeneron and University of Maryland.



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.

notes

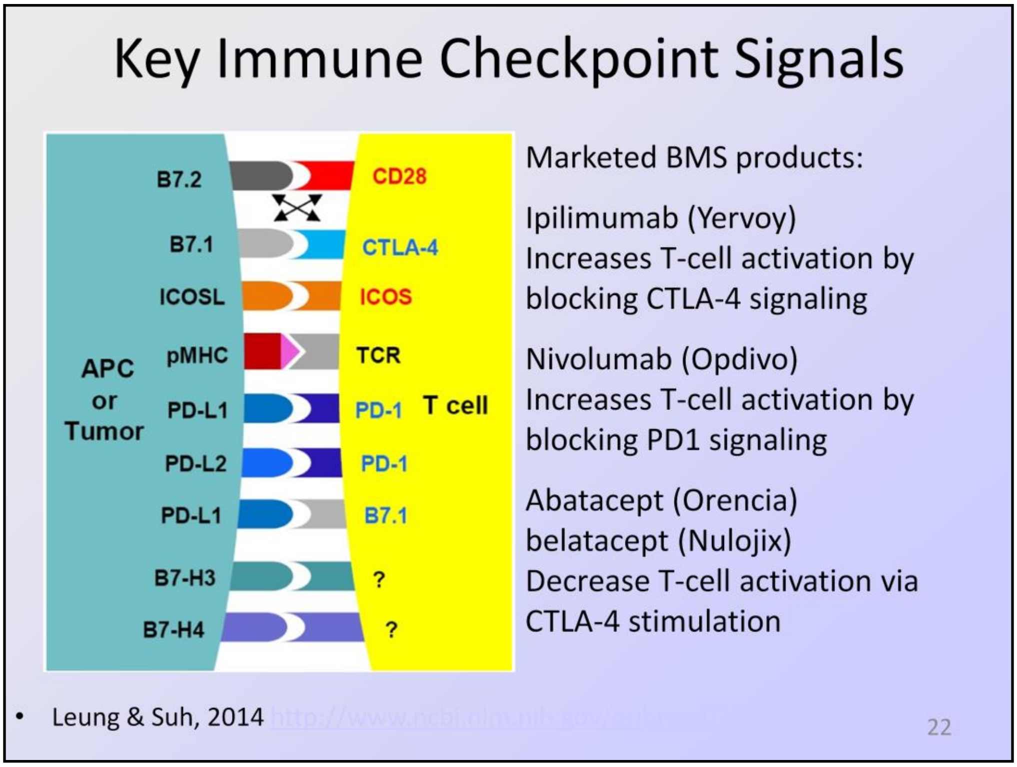


An idea that *did* work well

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

21

After BMS ended Virology Discovery research, the chemical matter from these efforts was transferred to Oncology. I hear they have continued to work on small-molecule approaches to immuno-modulation, but I know nothing about those efforts.



Key Immune Checkpoint Signals

Marketed BMS products:

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

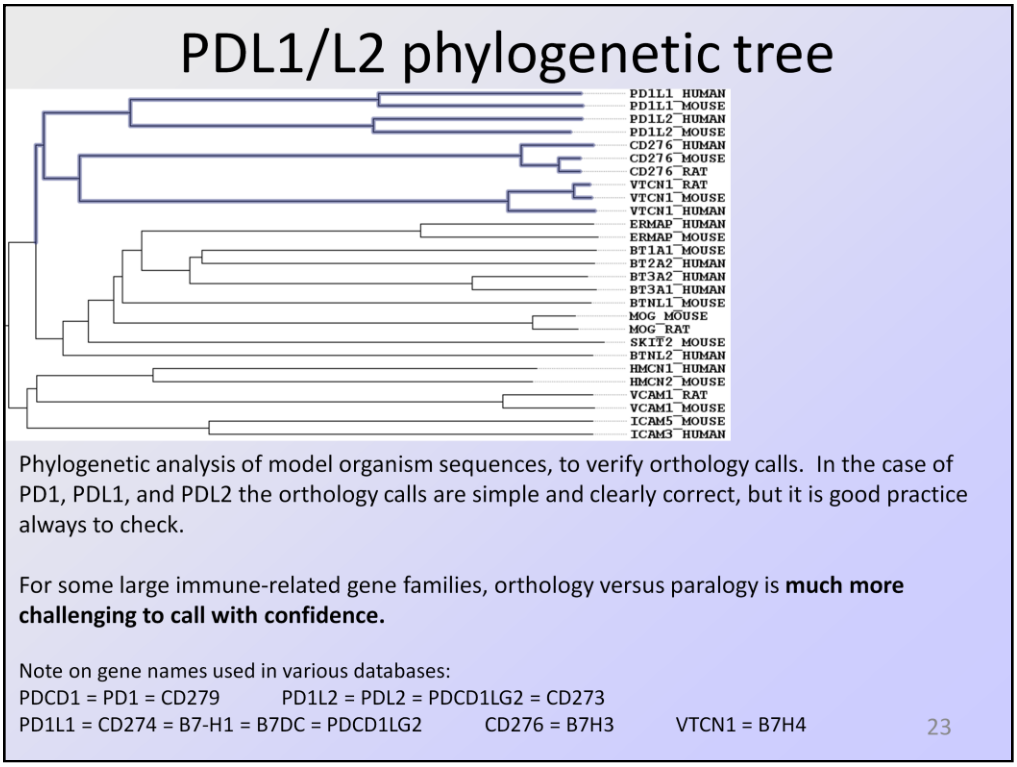
Nivolumab (Opdivo)
Increases T-cell activation by blocking PD1 signaling

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

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

Sets context of immune checkpoint signaling at BMS before I dive into PD1 Signaling.

Ipi and Nivo on market for cancer
Orencia sold for rheumatoid arthritis
Nulojix sold for organ transplantation



PDL1/L2 phylogenetic tree

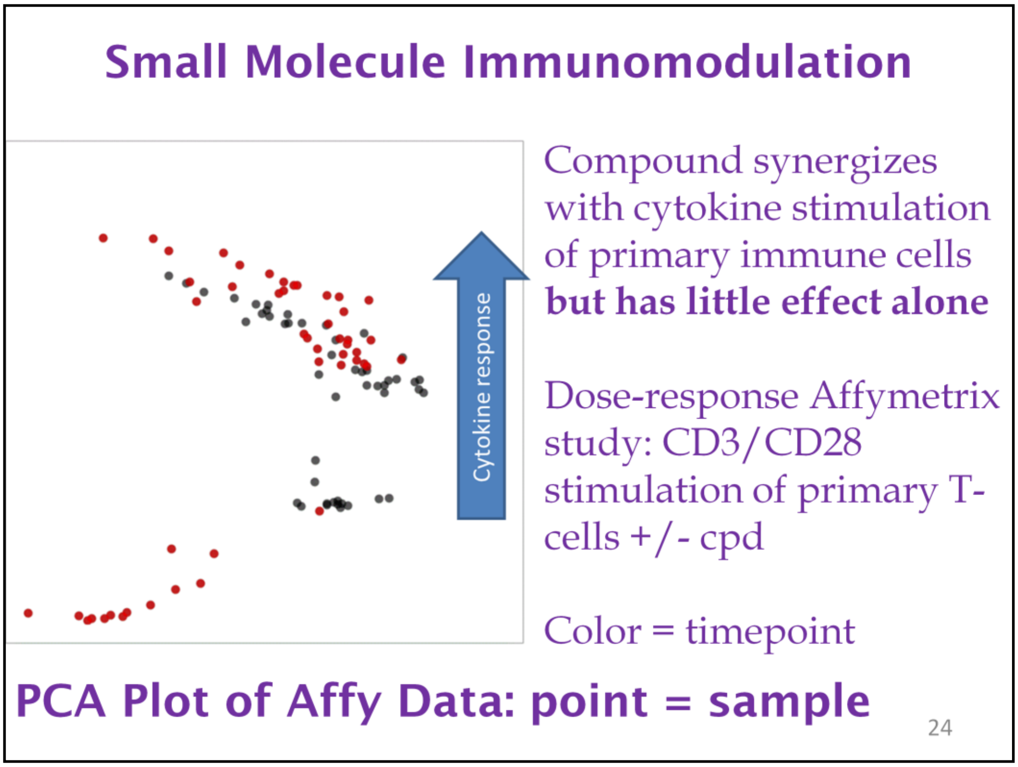
Phylogenetic analysis of model organism sequences, to verify orthology calls. In the case of PD1, PDL1, and PDL2 the orthology calls are simple and clearly correct, but it is good practice always to check.

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

To make this tree. First did NCBI Blast search. Specified Human, Rat, and Mouse as species; SwissProt as database (NCBI Blastable SP has only canonical isoforms), and 500 for number of hits to return. Then used ClustalX to make NJ tree (slow/accurate method for pairwise alignments) and multiple sequence alignment. Total time for the Clustal job with 500 sequences was 2 hours and 13 minutes, of which 1:33 for the all-against-all pairwise alignments to build the NJ tree and 40 minutes for the progressive MSA using that tree. Here is shown just a small portion of that tree.

Using 500 Blast hits to build our tree is an effective brute-force means of (1) being sure we got all close paralogs to our genes of interest and (2) being confident about rooting the section of the tree that we are examining. Expending a couple hours of CPU time avoids manual effort of selecting paralogs and outgroup from which to build a smaller tree. The human can do other things while the CPU runs this analysis.



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

24

I've heard that BMS Immuno-Oncology has continued exploring the small-molecule approaches to Immunomodulation that we pioneered in our Immuno-Virology work, but since BMS ended all Virology work at the end of 2015 I know almost nothing about what they are doing in that area now.

In addition to antibodies targeting PD1 and PDL1, BMS Virology had major efforts to find small molecules that would modulate immune responses to chronic infection. I was primary analyst for this OMICS study of some compounds that my colleagues discovered with a high-throughput screen followed by some Medicinal Chemistry to develop SAR. The basic idea of this study was a team effort, but the specific experimental design used was mostly my work.

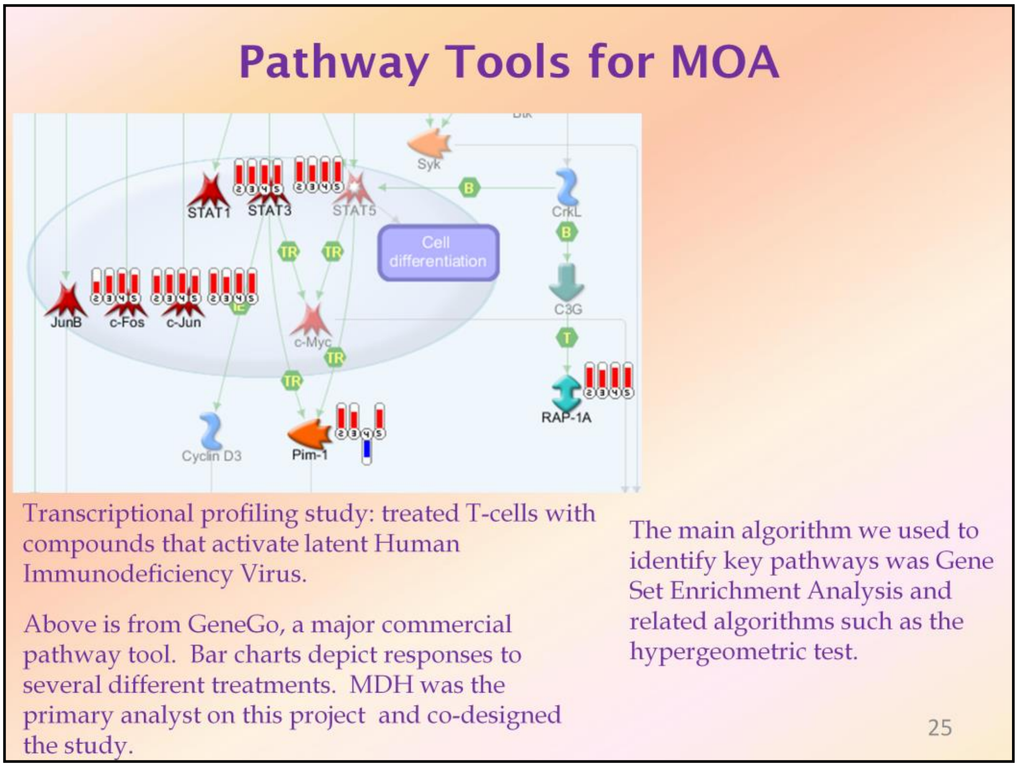
HTS screen looked for compounds that synergized with CD3 stimulation of PD1-expressing T-cells **but had little phenotypic effect in the absence of CD3.** One hit compound looked especially promising, and became the focus of an early discovery program with Chemists making analogs and Biologists assaying them for SAR.

These data came from an Affy study investigating the MOA of our lead compound. We used beads coated with **activating** antibodies to CD3 and CD28 and to CD3 plus CD28 to stimulate CD8 T-cells, with and without the compound, at varying doses.

Each point is a sample. Some samples got only cytokine stimulation; some samples got only CPD; some got both. **Compound alone look just like untreated cells. Low level of cytokine stimulation plus CPD looks like higher level of cytokine stimulation. Do not see compound-specific trend in PCA plot.**

I did all statistical analysis. Used Spearman's Rho to look for nonlinear correlations with dose; also used BMS in-house Scanning Dose Response Modeling tools. Used commercial tools (such as the commercial tool GeneGo) and in-house tools to do Pathway and Gene Set analysis of the genes that I found related to response. I found a clear immune response signature.

Based on transcriptional profiling, Luminex Cytokine profiling, and other data, we believe our compounds bind a Target close to **where PD1 signaling blocks T-Cell Receptor Signaling.**



Pathway Tools for MOA

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

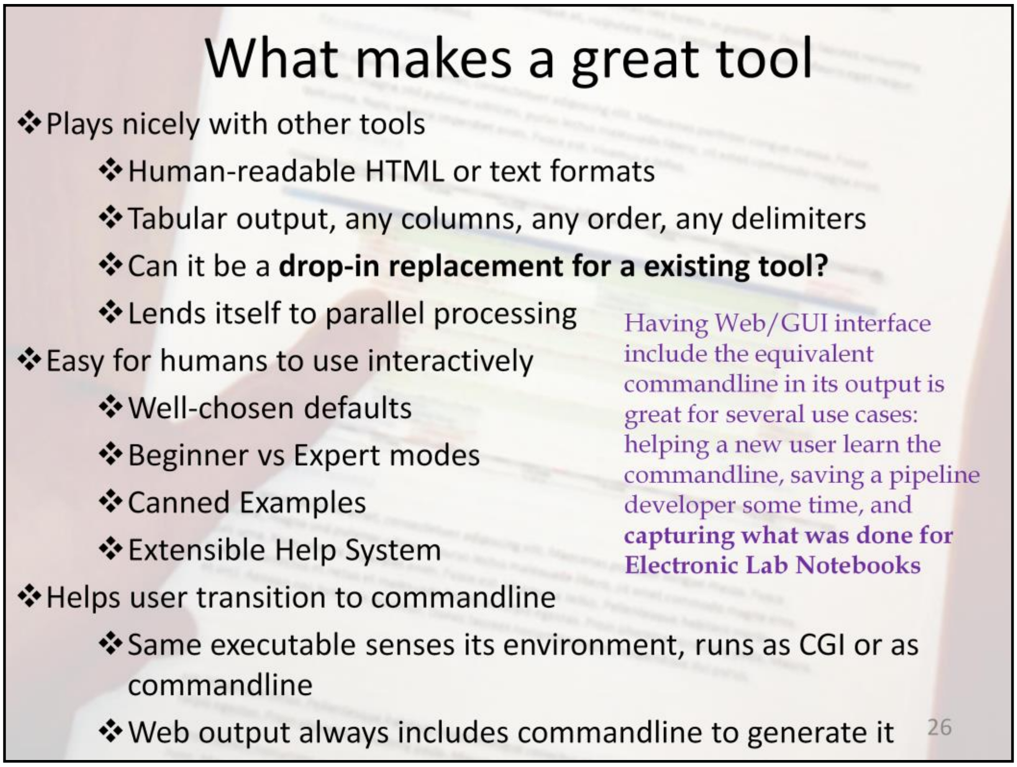
Above is from GeneGo, a major commercial pathway tool. Bar charts depict responses to several different treatments. MDH was the primary analyst on this project and co-designed the study.

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

25

Our approach to HIV cure research had two prongs, because HIV is a *retrovirus* which means the integrated viral genome can sit undetected and transcriptionally silent in resting memory T-Cells for many years. Stop antiviral drugs, and sooner or later one of those resting cells will wake up and re-infect the patient. So immunomodulation alone cannot flush out the viral reservoir from hiding. In the field, people talk of the “kick and kill” strategy. You’ve just seen an example of the “kill” part: increase immune response so it clears out any cells expressing the viruses. The “kick” part means getting those latent viral genomes expressed so there are antigens for the immune system to detect. If HIV cure therapy using this approach works, probably the patients will also need to continue standard antiviral drugs during the “kick and kill” treatment, in order to prevent the viruses that get flushed out of hiding from infecting healthy T-Cells.

One challenge for using Pathway Databases for fields other than Oncology is that in my experience most of them, both commercial and public domain, are rather Oncology-centric.



What makes a great tool

- 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**

People find exciting new ways to use any great tool. Which means (1) a great tool finds new users who need to get up to speed on it as quickly as possible and (2) expert users will want to combine the tool with other tools. So a good tool has both a clean interactive user interface (web or commandline) **and** lends itself to building pipelines. For instance, if there’s an existing tool that does a similar job, try to make your input and output formats as similar as possible to that older tool. And every tool should have features like customizeable output formats and support for dividing a job into parallel batches without needing a wrapper.

Examples of tools that allow capturing Web/GUI actions as commands include: tools built by my colleague Charles Tilford at BMS, TimeLogic DeCypher system, and R Commander (which, by the way, does play nicely with the more famous R-Studio; I like to install both myself).

Great example of a tool that has very flexible output and features to facilitate running on a compute farm is the alignment tool exonerate from EBI, which is an important part of the ENSEMBL pipeline.

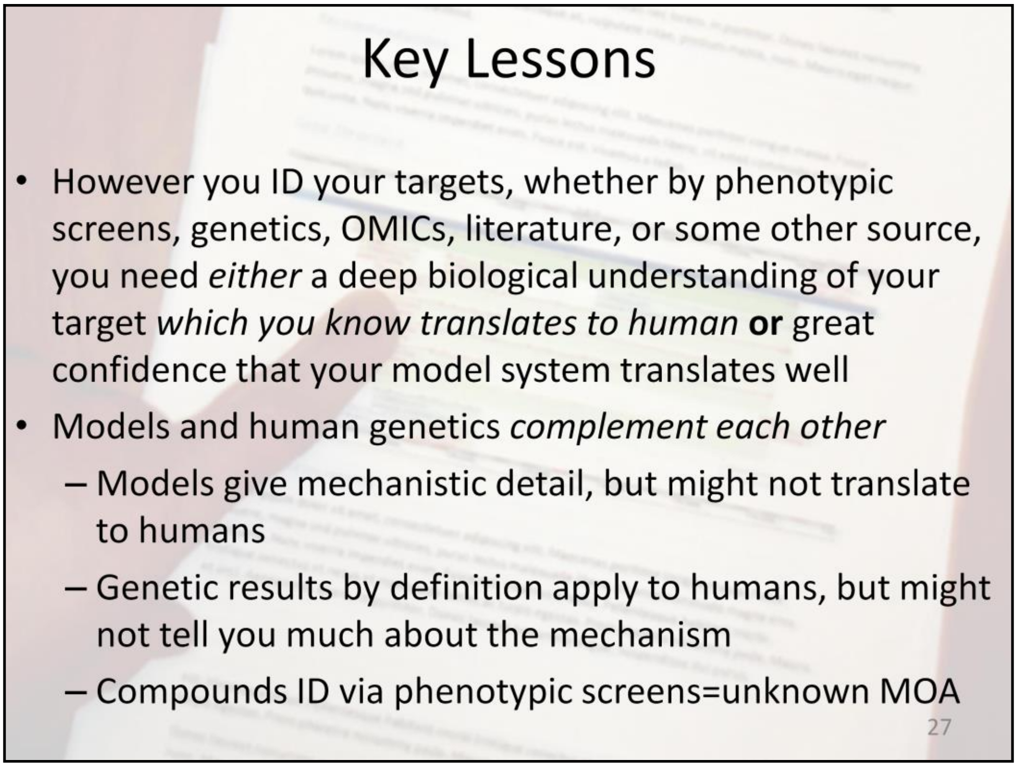
Translated alignment tools that maintain the correct reading frame are very useful for EST-to-genome and EST-to-protein alignments. ENSEMBL team at EBI has therefore developed one of my favorite alignment tools, EXONERATE:

Flexible both regarding alignment models and regarding output formats

Fast and scalable
Lends itself to high-throughput analysis pipelines
Easy to install and configure

Ability to handle indels comparable to that of Framesearch (on which I wrote the chapter for Current Protocols in Bioinformatics) without needing FPGA hardware. Much better for frameshifts than BLAST (on which I wrote the chapter for Current Protocols in Human Genetics), and almost as fast as BLAST on a single CPU – and more easily scalable to a farm than BLAST.

For MSA, I like Clustal, Muscle, and MACSE (slower but handles indels)



Key Lessons

- However you ID your targets, whether by phenotypic screens, genetics, OMICs, literature, or some other source, you need *either* a deep biological understanding of your target *which you know translates to human* **or** great confidence that your model system translates well
- Models and human genetics *complement each other*
  - Models give mechanistic detail, but might not translate to humans
  - Genetic results by definition apply to humans, but might not tell you much about the mechanism
  - Compounds ID via phenotypic screens=unknown MOA