elastic{on} 2017

Integrating Human Genetic Data to Help Drive Drug Discovery

Dan Myung
Bhasker Bokuri

MERCK
Be well

March 9, 2017

Agenda

- About Us
- Genetics for Drug Discovery
- Our Journey
- The Future

MERCK

Be well

We are a global healthcare company with a 125-year history of working to make a difference in global health.

BUSINESSES

Pharmaceuticals, Vaccines, Biologics and Animal Health

About Us

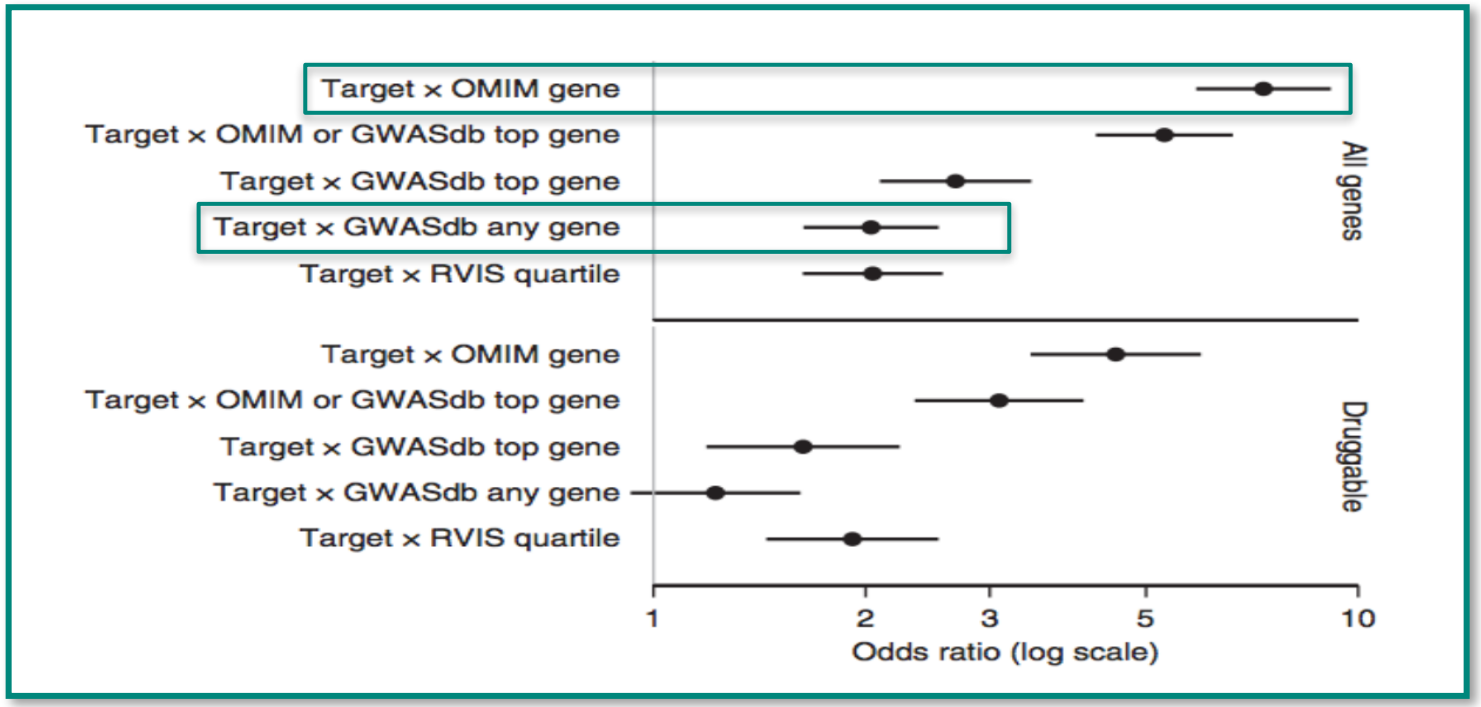
- Scientific Computing for Merck Research Laboratories
  - Engineering resources for early discovery research areas
    - Translational Medicine (Genetics & Pharmacogenomics)
    - Chemistry and Pharmacology
    - Modeling, Simulation & Applied Mathematics
    - Scientific Information Management
  - We build stuff!

We Need a Better Way to Predict Efficacy and Safety Much Earlier in the Drug Development Process

Portfolios of Drug Targets With Human Genetic Support: 2-Fold Higher Probability of Success

The Support of Human Genetic Evidence for Approved Drug Indications

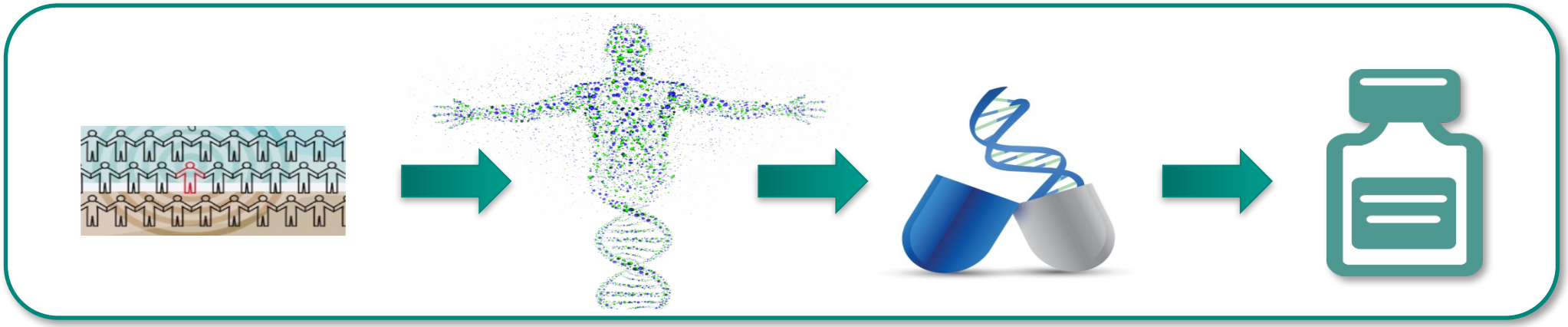
Nelson MR; Tipney H; Painter JL; Shen J; Nicoletti P; Shen Y; Floratos A; Sham PC; Li MJ; Wang J; Cardon LR; Whittaker JC; Sanseau P.



Nelson MR, et al. *Nat Genet.* 2015;47(8):856-860.

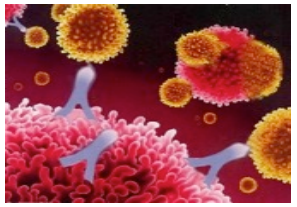
Summary

- Use genetics to drive drug development→2x improvement in POS of bringing a drug to market
- Could translate to reducing the cost of drug development
- …and reduce drug prices, making cutting edge medications available to those who need it



Scientific Approach

Identify area of unmet medical need



Immunology

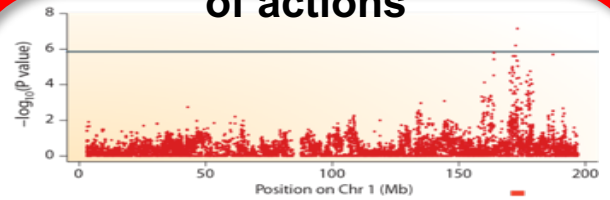
Cancer

Cardiometabolic disease

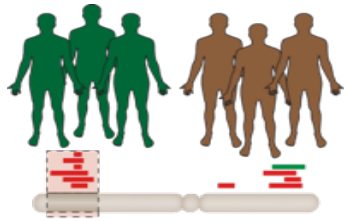
Neuropathology

Mine genetic datasets and pathways to generate target hypotheses

Pathways and mechanism of actions



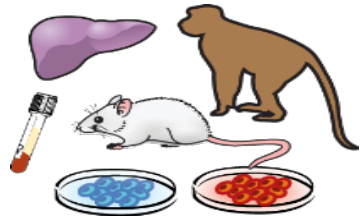
GWAS candidates



Genotype-phenotype correlation

Combine computational approaches and functional biology to determine highest POS targets

Comprehensive 'omics data integration



Experimentation in physiologically and genetically relevant in vivo/in vitro models

Leverage clinical data for human validation

Clinical biomarkers

Electronic medical records

Patient response information

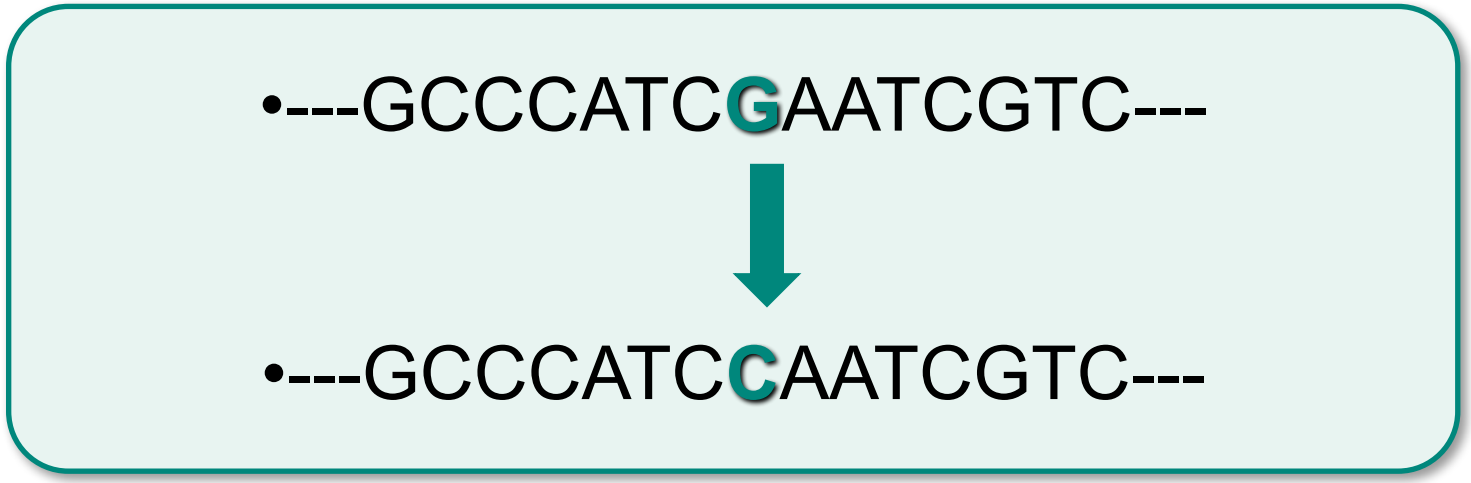
Find and Filter

Brief Genetics Primer

Definitions

- Genomics: Genes of organism, sequences, and its information
  - ~19,000-20,000 protein coding genes in human genome, 23 chromosomes
  - Genes encoded nucleotides using A-C-G-T
- Genetics: The study of the effect that genes have on an organism
- Genotype: The set of genes an organism carries (“code”)
- Phenotype: The observable characteristics

SNP – Single-Nucleotide Polymorphism



Definition

- Single base change in genetic sequence
- 1 SNP per 1000 base pairs
- 3-4 million in the genome
- Each SNP has a set of alleles (usually 2)
- Basic unit of variation in our work

Really Really Simplified Code Analogy

```
$ git tag –a b38 –m “Final genome build 38” [1]
23 chromosomes changed, 3.2 mega-basepairs changed
 $ git checkout -b dan_b38
 $ git merge mom/mom dad/dad
$ make && make install && mvn test
# asthma PASS
# lactose_tolerance FAIL
# total_cholesterol WARN
…
$ git blame
(mom)
(dad)
(mom)
```

```
$ git diff b38

+a
-g
+c
-t
+aa
…
```

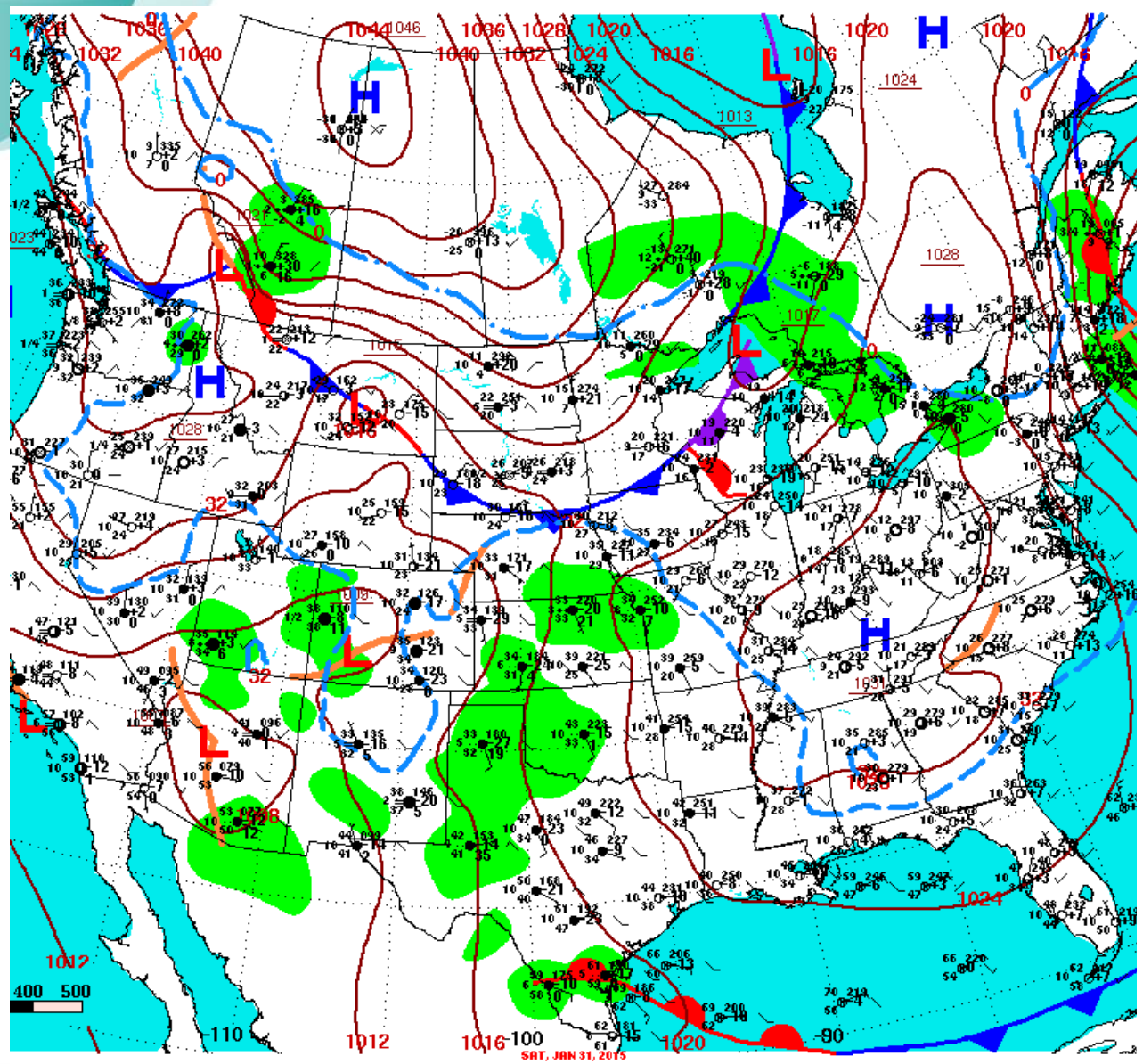
SNP

[1] https://en.wikipedia.org/wiki/Reference_genome

SNP Importance

- SNPs are codified variations in genotype that could result in observable variations in phenotype
  - eg: Hair color, susceptibility of cancer, heart and other diseases
- SNP prevalence frequencies vary within and among populations
- Understanding SNPs may help understand causes of disease and possible approaches to drug identification
- SNP to phenotype observations studies are underway
  - Identification, location, and nomenclature issues abound!
  - GWAS – Genome Wide Association Studies
    - Survey of gene/variant activity for disease association
  - eQTL – Expression Quantitative Trait Loci
    - Exploration of variant affecting surrounding gene(s) expressions

Basic units…



Surface Weather Map and Station Weather at 7:00 A.M. E.S.T.

https://upload.wikimedia.org/wikipedia/commons/a/a7/2015-01-31_Surface_Weather_Map_NOAA.png

Location
- Latitude, longitude, altitude

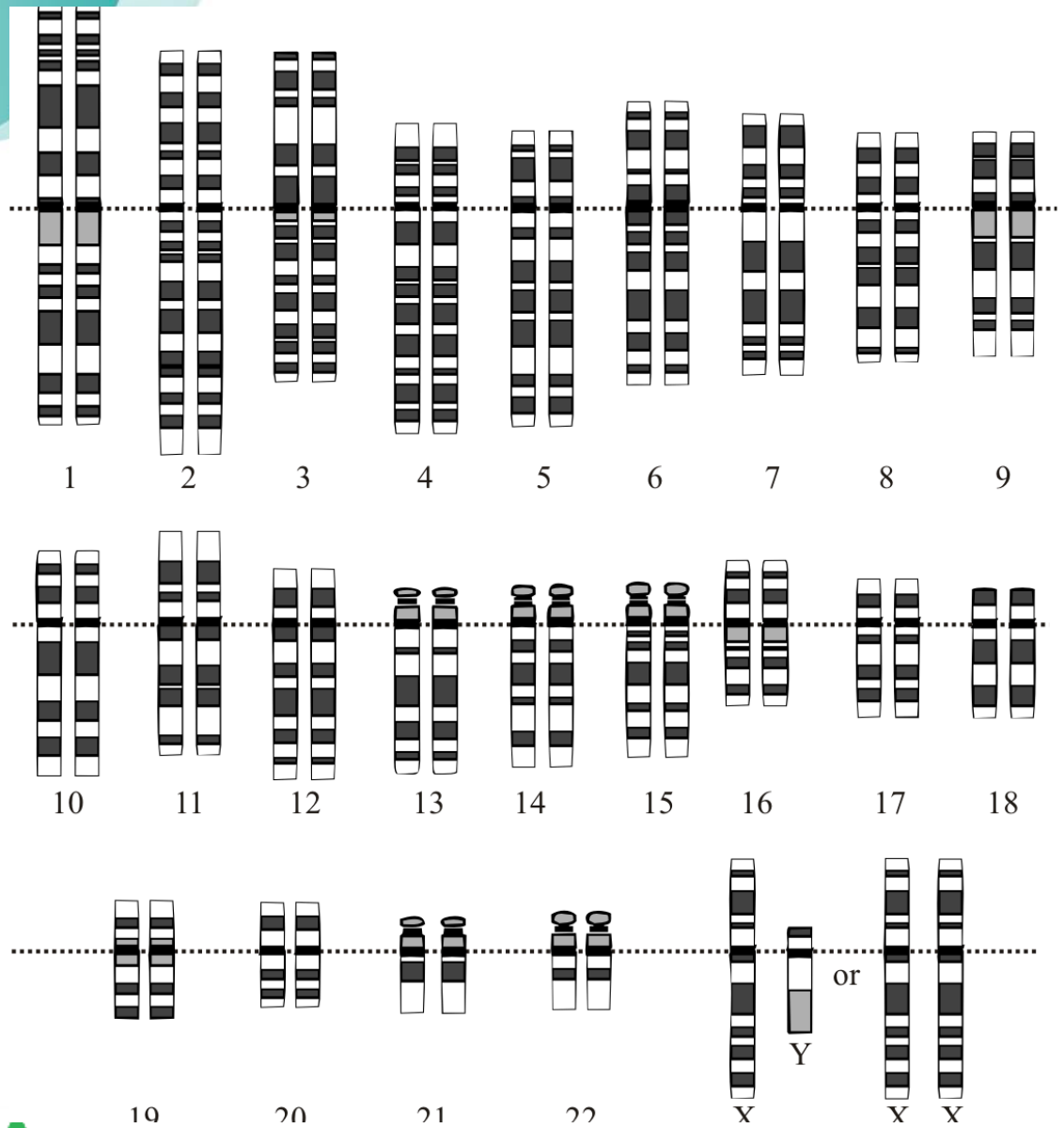
More Locations
- Addresses
- Intersection
- Regions
- Business Name
- Geographic Boundaries

Sensors
- Pressure, wind
- Temperature, humidity
- Radar, satellite

Observations
- Trips, Traffic
- Tweets
- Check-ins
- Reviews
- Photos

Predictions
- Hurricanes, tornados
- El Niño
- Crop outputs
- Wildlife Migrations
- Traffic impacts

Our Basic Unit



https://commons.wikimedia.org/wiki/File:Karyotype.png

Location
- Chromosome, Position

Location/Region Identifiers
- SNP (variant, rsid, HGVS, dbsnpid)
- Gene (Ensembl, Entrez, HGNC)

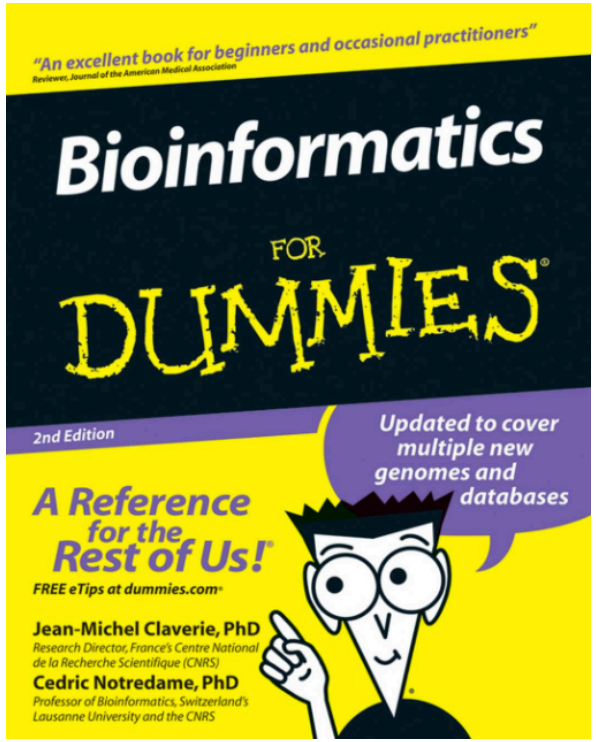
Additional Properties & Observations
- Gene Expression
- Epigenetics

Aggregate Observations
- Linkage Disequilibrium
- Population Allele Frequency
- Expression Quantitative Trait Loci (eQTL)
- Genome Wide Association Study (GWAS)
- Phenome Wide Association Studies (PheWAS)

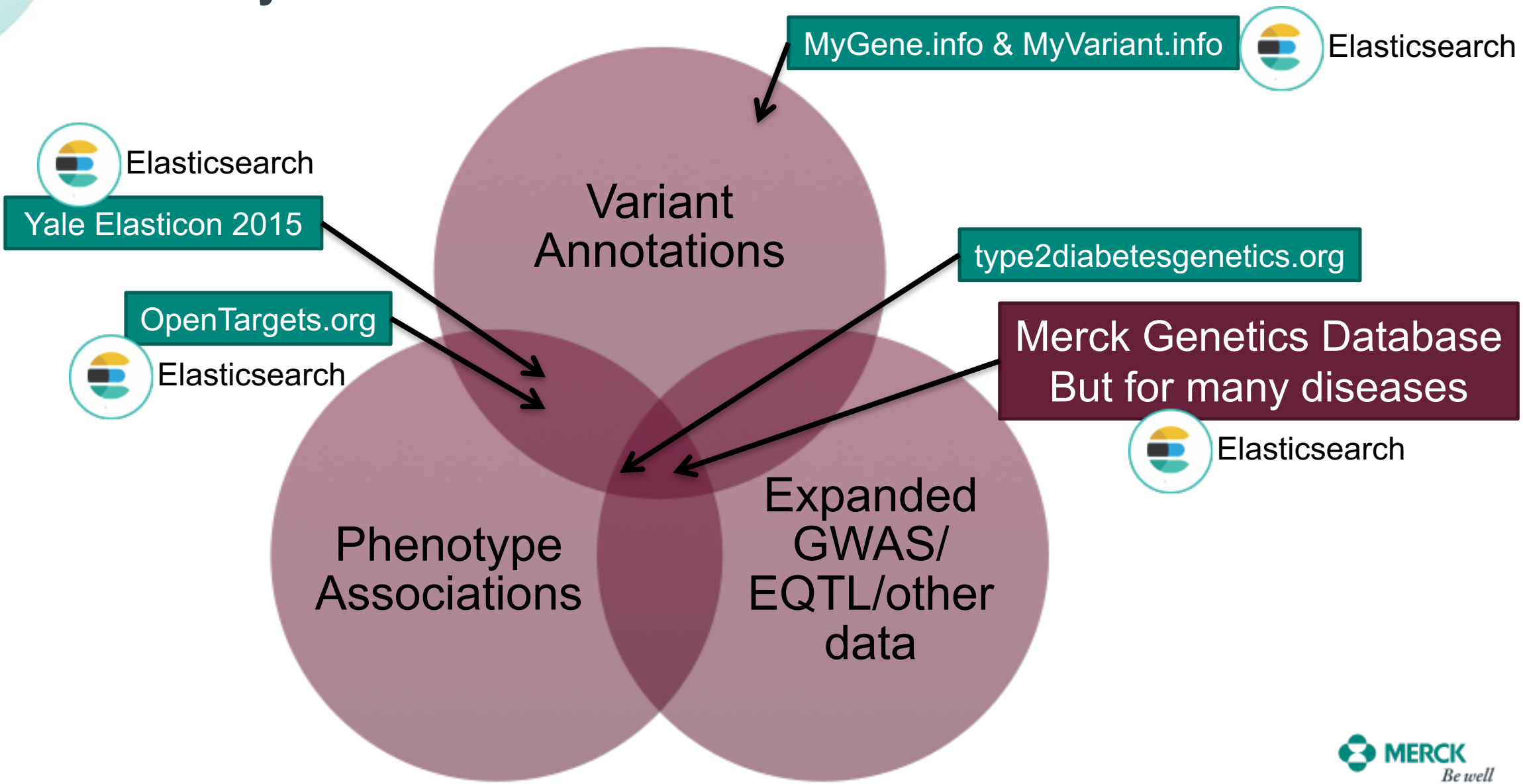
Predictions/Insights
- Identify disease causal genes/effects
- Reveal unknown biological mechanisms

The Need

- It's wild
  - Standards, methods, references
- It's tedious
  - FTP, gunzip, grep/cut/awk
  - RDBMS to Excel
  - Files in random folders everywhere
- It's slow
  - Mile long sql statements
  - RDBMS scale/variety woes
  - Basic question turnaround time days/weeks



Industry Case Studies



Data Harmonization

- Goal:
  - Annotated Variant (SNP) is the single unified identifier
  - Map diverse observations (eQTL, GWAS) to our variant doc
- Additional Guidelines
  - Normalize like data fields to common nomenclature
  - Any unusual fields prefix with ‘X_’

Doc Structure

Variant

Lat/Long/*region*/Identifier

_parent

Observations & Aggregations

- GWAS/PheWAS
- eQTL
- Allelefreq
- Genotype
- Expression

_id: b37-7-55154975-AAC-A


```
"dbsnp_id": "rs199536192",
"b38": {
  "ref": "AAC",
  "alt": "A",
  "chrom": "7",
  "pos": 55087282,
  "hgvs": "b38:7:g.55087284_55087285delCA"
},
"b37": {
  "ref": "AAC",
  "alt": "A",
  "chrom": "7",
  "pos": 55154975,
  "hgvs": "b37:7:g.55154977_55154978delCA"
},
"b36": {
  "ref": "AAC",
  "alt": "A",
  "chrom": "7",
  "pos": 55122469,
  "hgvs": "b36:7:g.55122471_55122472delCA"
},
"entrez_gene_symbol": "EGFR",
"hgnc_id": "HGNC:3236",
"ensembl_gene_id": "ENSG00000146648",
```

Doc Examples

```
{
    "_type": ”genotype”,
    "_parent": "b37-4-80989630-C-T",
    "filter": "PASS",
    "hap2": [
      417,
      976,
      1188, …
    ],
    "call": {
      "HG01272": "0|1",
      "HG03091": "0|1",
      …
    },
    "format": "GT",
    "hap1": [
      1091
    ],
    "study": "1000Gph3",
    "qual": "100"
}
```

```
{
    “_type”: “eqtl”,
    "_parent": "b37-4-80989630}C-T",
    "expr_id": "ENSG00000169174",
    "gene_id": "PCSK9",
    "beta": 0.3987,
    "gene_ensembl": "ENSG00000169174",
    "cis_trans": "cis",
    "p_value": 0.000002545,
    "study_id": "GTEx_liver_cis",
    "study": "GTEx",
    "fdr": 0.0003283,
    "effect_allele": "unk",
    "sub_study": "liver",
    "t_stat": 4.925
}
```

```
{
    “_type:” “gwas”,
    "_parent": "b37-4-80989630-C-T",
    "p_value": 0.82,
    "beta": 0.0072,
    "study_id": ”GWASCatalog+Heights",
    "sub_study": ”Heights",
    "study": ”GWAS",
    "effect_allele_freq": 0.308,
    "se": 0.038
```

656,938,422 hits

Allele Frequencies

_type:allelefreq AND freq:[0 TO 0.5]

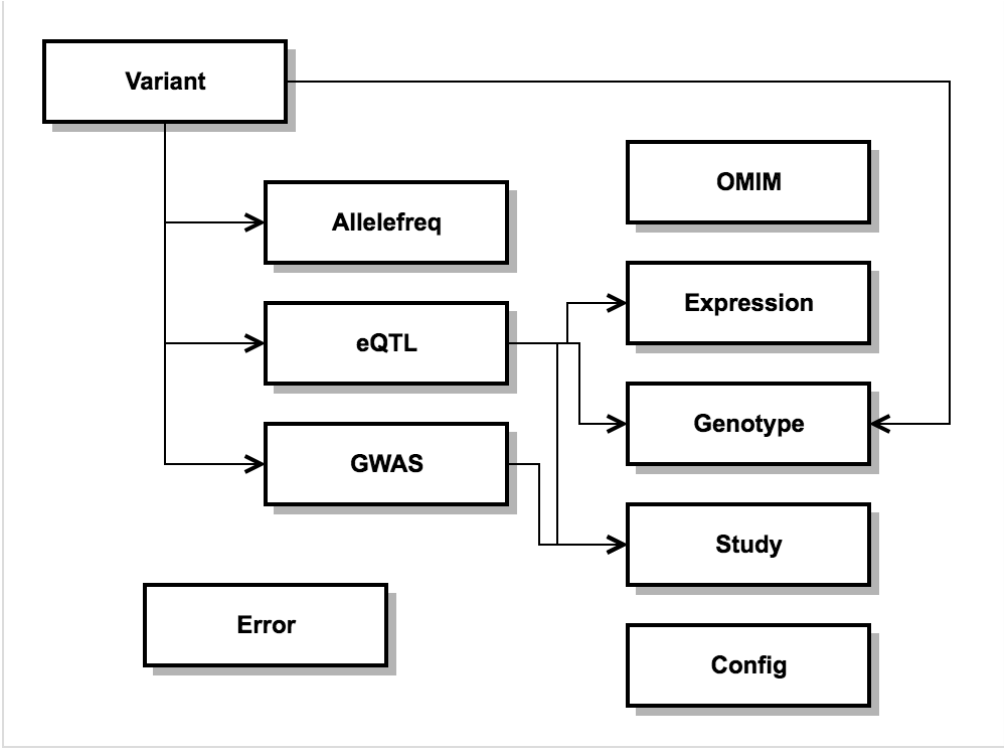
pop	freq	ac	an
1000Gph3/EUR_TSI	0.5	107	214
1000Gph3/AMR_MXL	0.5	64	128
1000Gph3/AMR_PEL	0.5	85	170
1000Gph3/EUR_IBS	0.5	107	214
1000Gph3/SAS_GIH	0.5	103	206
1000Gph3/SAS_ITU	0.5	102	204
1000Gph3/EUR_IBS	0.5	107	214
1000Gph3/EUR_FIN	0.5	99	198
1000Gph3/EUR_CEU	0.5	99	198
1000Gph3/SAS_PJL	0.5	96	192
1000Gph3/AFR_YRI	0.5	108	216
1000Gph3/EAS_CDX	0.5	93	186
1000Gph3/EUR	0.5	503	1,006
1000Gph3/AFR_LWK	0.5	99	198
1000Gph3/AFR	0.5	661	1,322

Mapping Our Mapping Journey

- Version 1 (ES 2.4)
  - Single variant document type with all data includes (Nested schema)
  - Slow indexing, Fast queries
  - Natural way of consuming
    - “Give me a variant with every observation we have”
- Version 2
  - Variant parent with child types
  - Fast indexing, slow `has_parent` and `has_child` queries
  - Two step query workaround
    - get parent ids then filter children by _parent term
- Version 3-4:
  - Predefined data types, selective `index: false`, 10Tb to 1Tb!

Mapping Journey cont'd

- Version 3-4 (current):
  - ES 5.1.1 upgrade
    - no more `_parent`, `has_parent` hack, but performance improved
  - High cost of data retrieval vs actual query cost



Data Usage Story

- Aggregations underutilized
  - Users want to see every point (10k+)
  - With table of all doc data!
  - Edge of page vs scroll experience for ad-hoc chart requests
- Un-indexed data used for other calculations
  - Painless scripting candidates
  - Pairwise variant calculations using genotype data

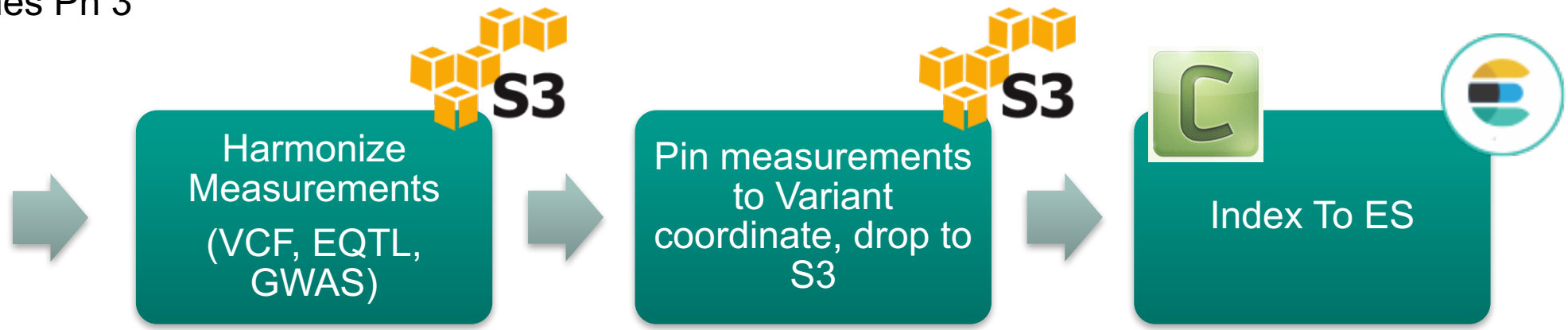
Data Loading

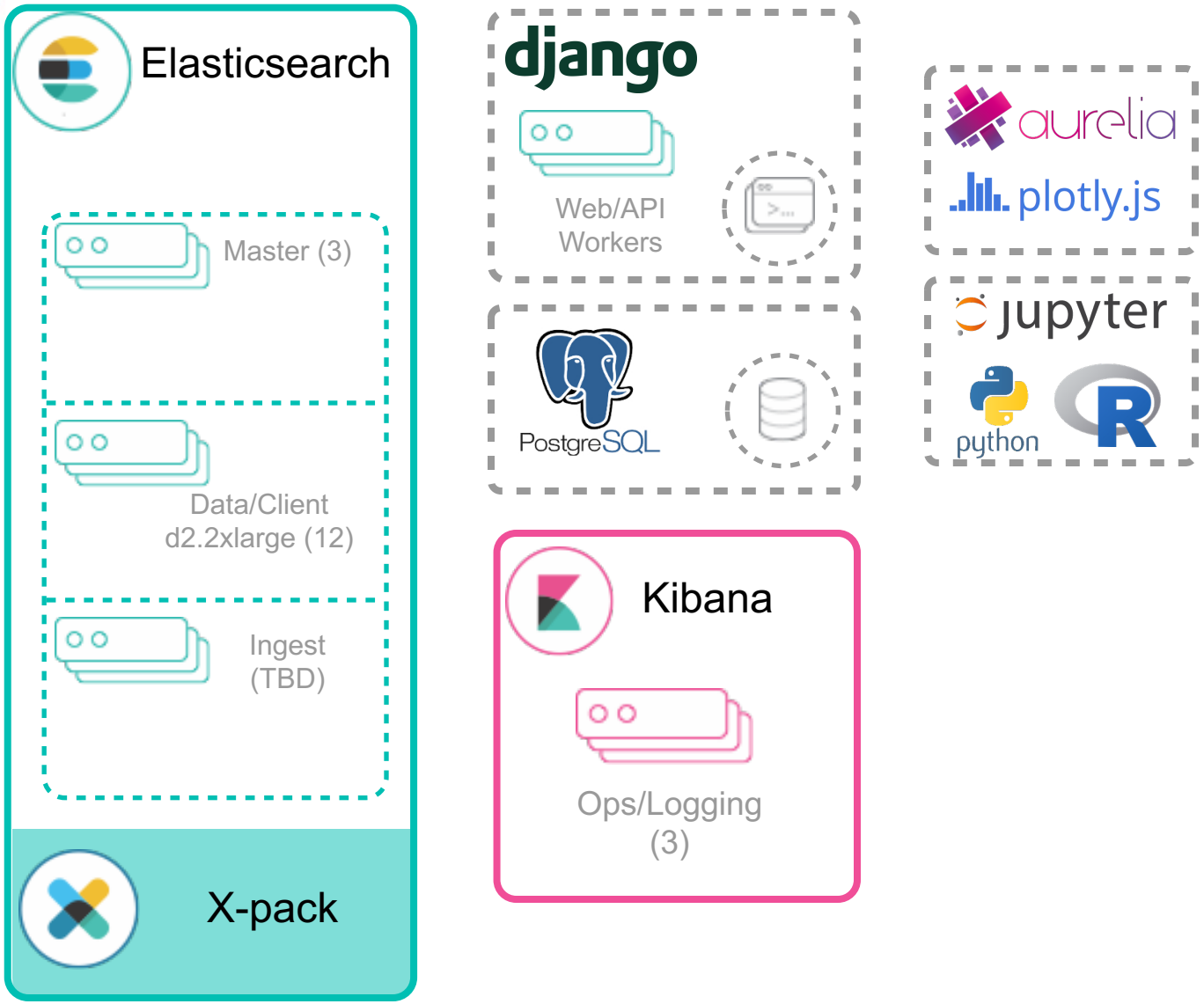
IGSR 1000 Genomes Ph 3

GTEx Portal

ExAC

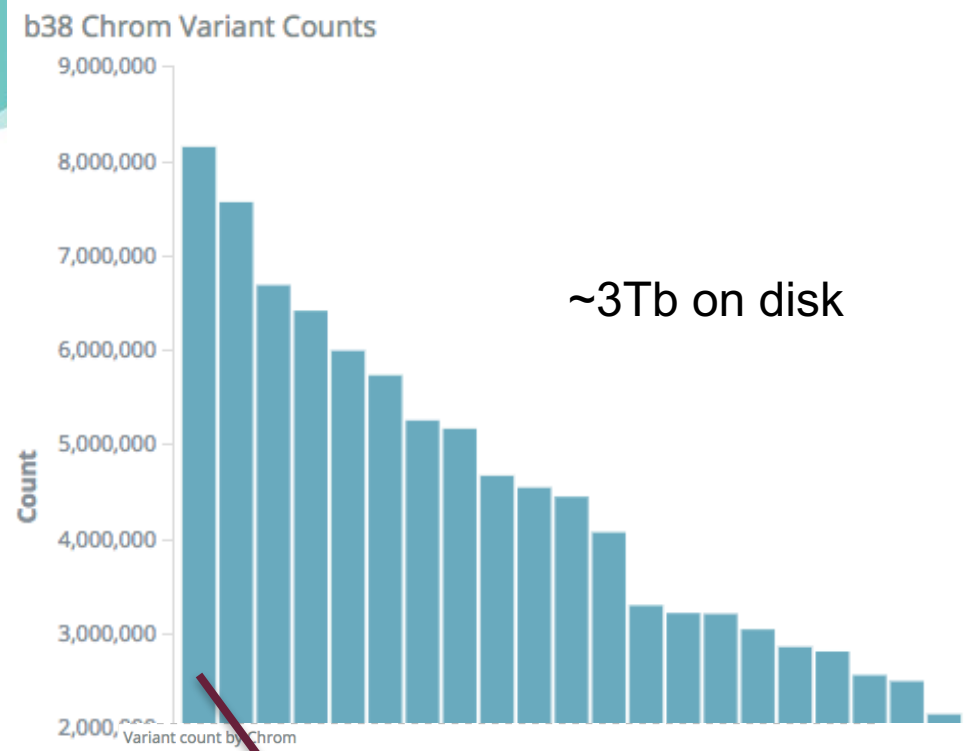
…and more





Current Stats

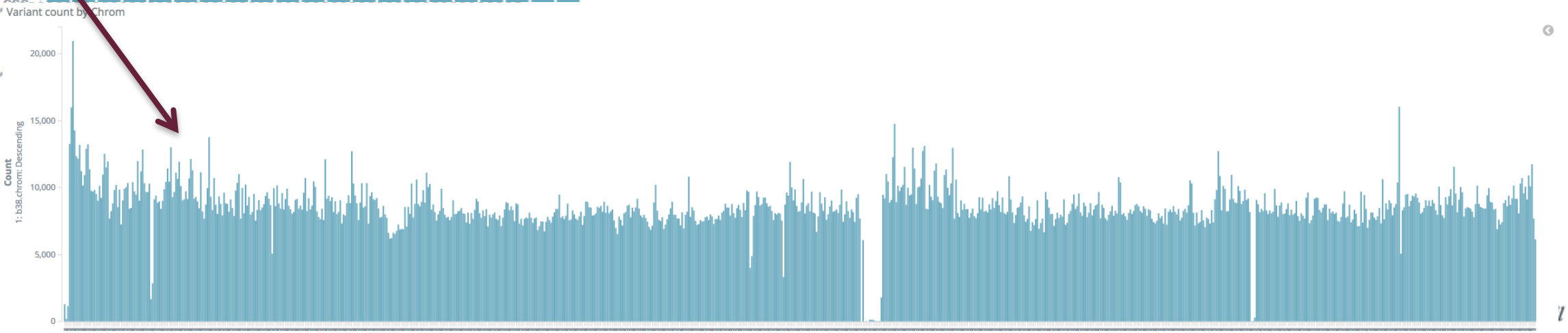
Expecting to 2-3x by year's end

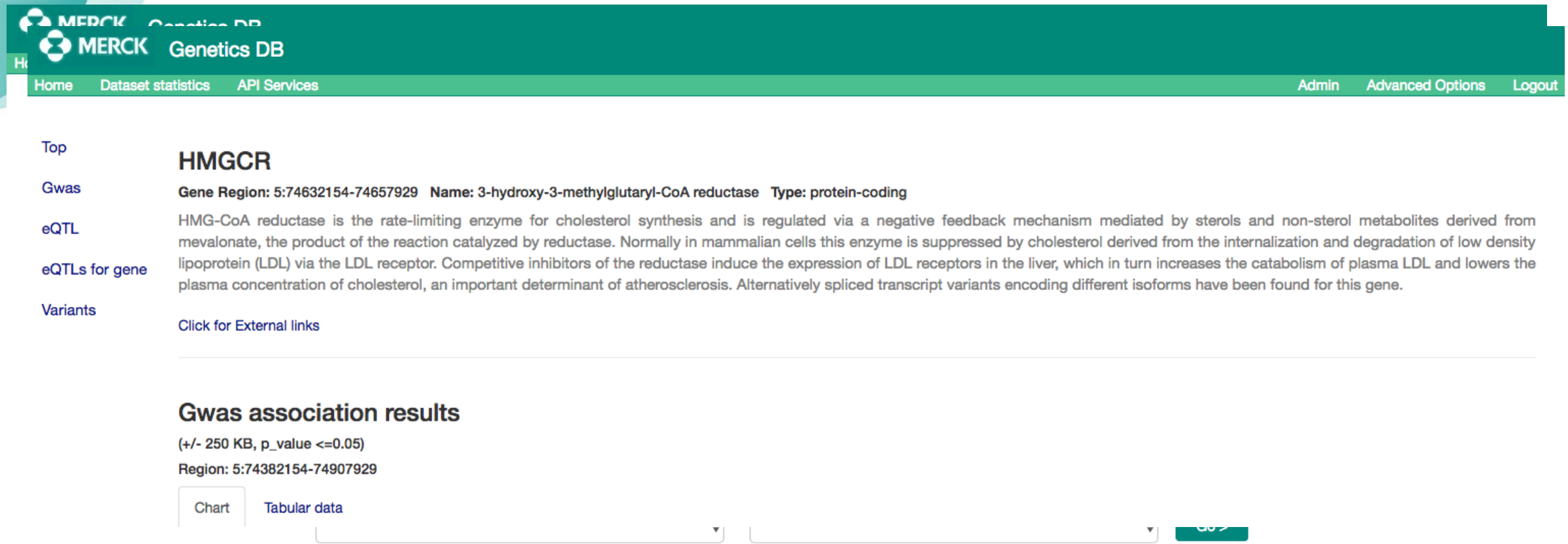


~3Tb on disk

Gendb Types Breakdown

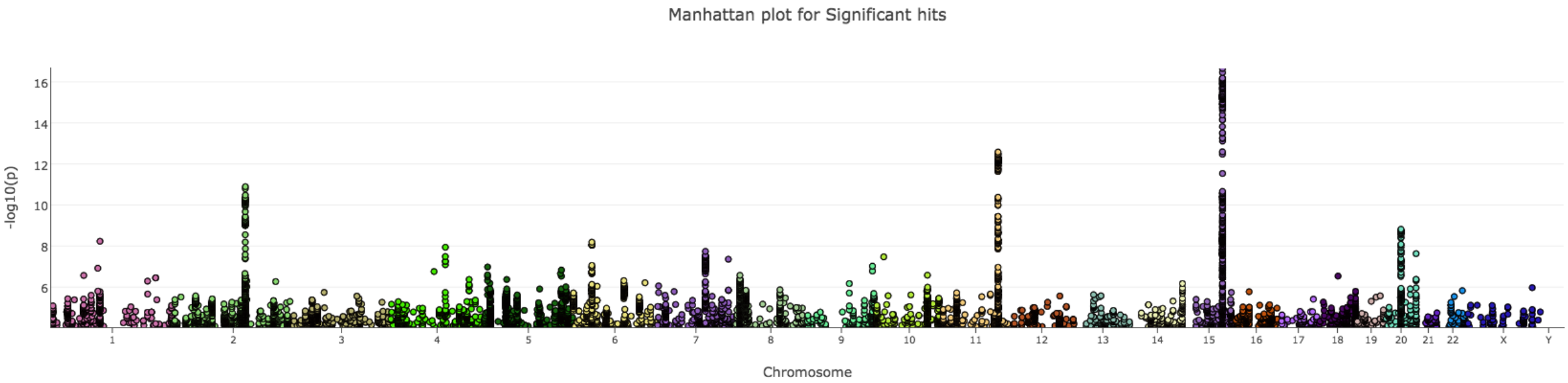
_type: Descending	Count
gwas	998,911,076
allelefreq	730,551,258
genotype	252,626,342
eqtl	114,946,978
variant	99,460,062
expression	1,376,228
omim	24,740
study	5,553



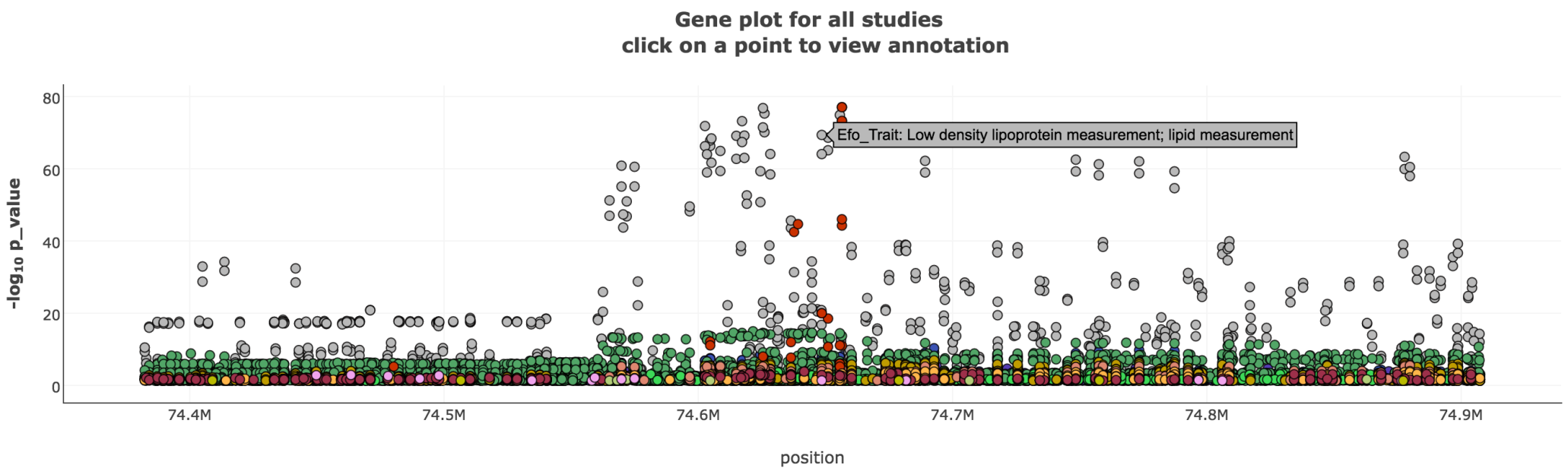


Genome-wide view

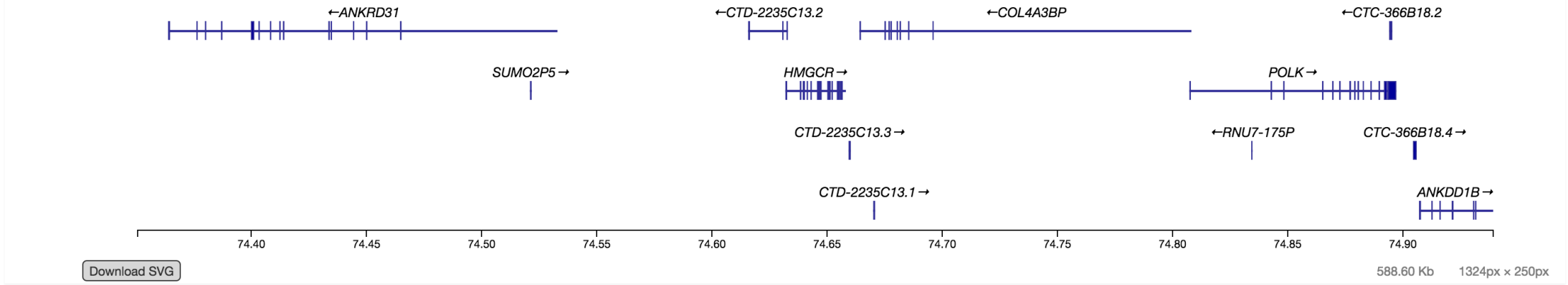
High level variant activity for a given phenotype by chromosome and position



GWAS

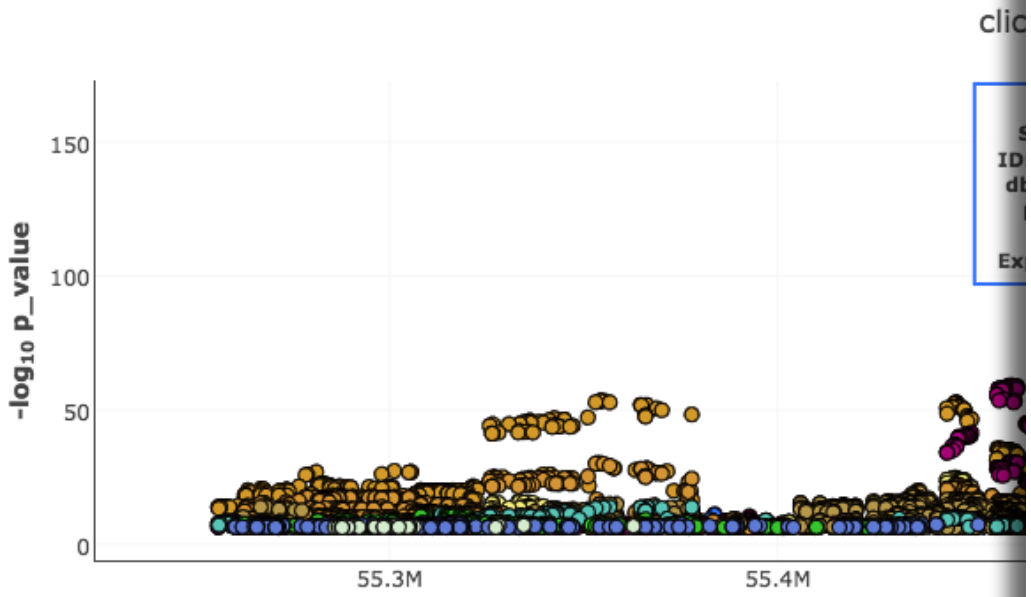


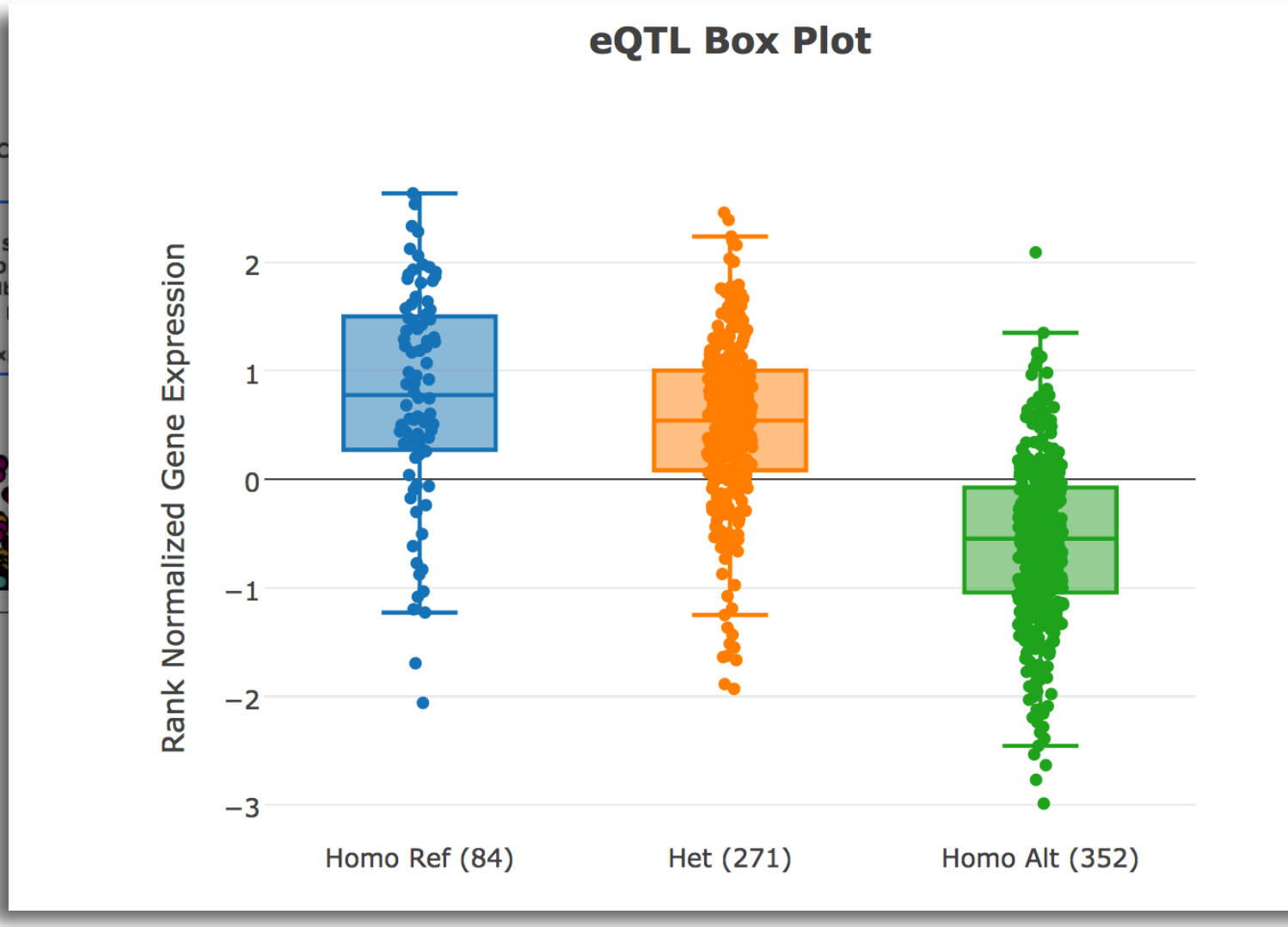
Needs expert interpretation!

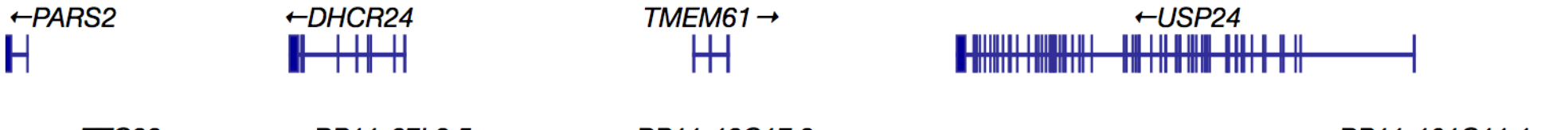


EQTL

Needs expert interpretation!







Conclusion

- A first iteration!
  - Variant identification and harmonization
  - Key data sets layered on top
  - **Basic** visualizations – days, now one-click
- The Future
  - Self service data loading, expanded types and observations
  - Denormalize more fields into child docs
  - Enable additional analytics
    - Apply new statistical tests
    - Deeper Utilization of Elasticsearch features (Scripting, Graph)

Thank You!

Daniel Myung (daniel.myung@merck.com)
Bhasker Bokuri (bhasker_bokuri@merck.com)

Special thanks
Jason Hughes, PhD, Dan Chang, PhD – MRL IT Informatics and MRL GpGx