Mining 3 - Rare variants & ExAC

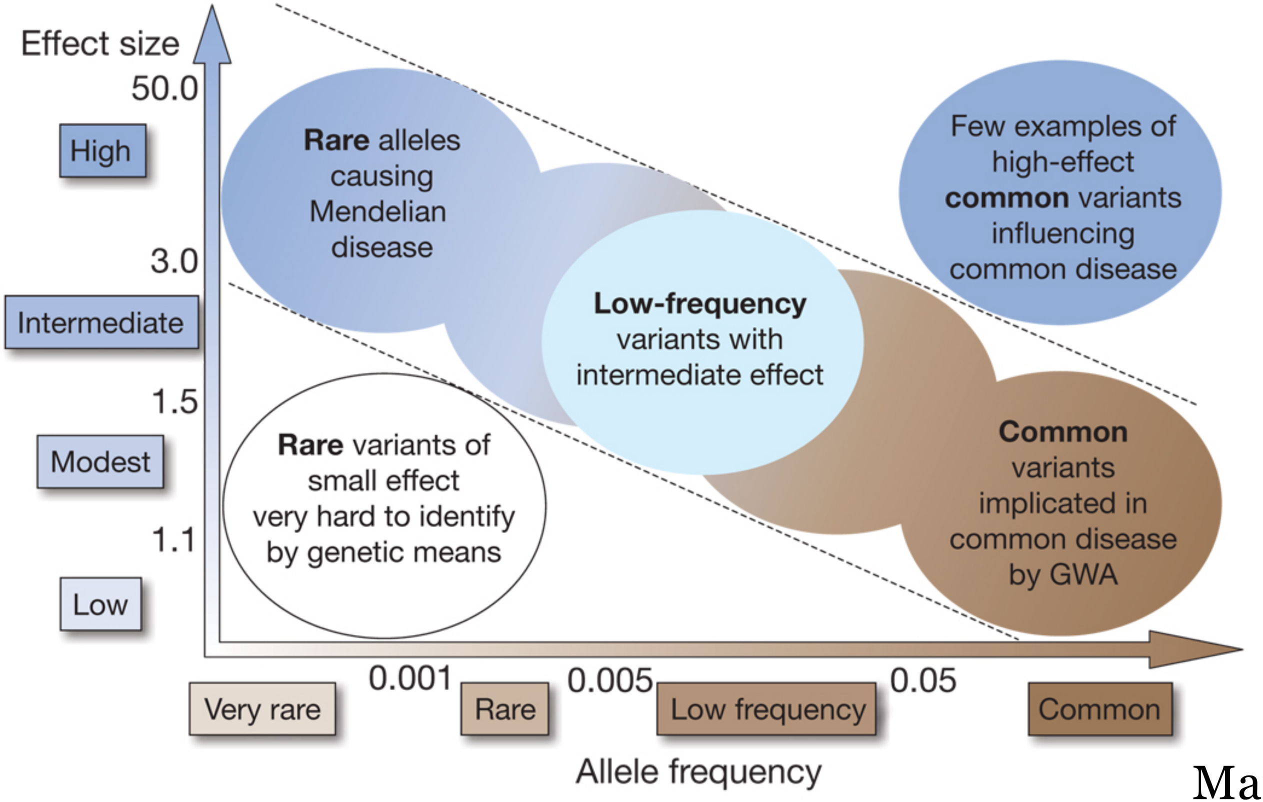
CBB752 Biomedical Data Science: Mining and Modeling

Monkol Lek, February 21st 2018

Overview

- Rare variants and ExAC
- Gene Constraint
- Beyond ExAC – gnomAD project
- Big data and scaling up analysis

Rare variants in human populations

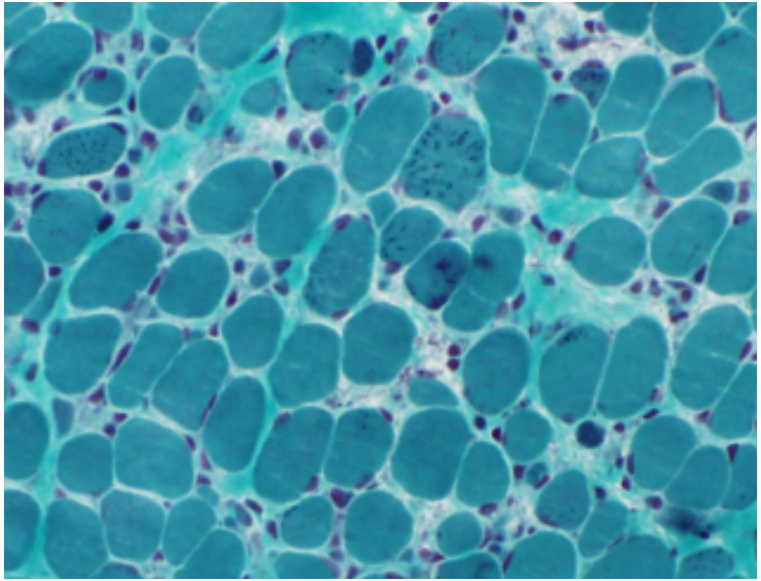

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect **common** variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Manolio *et. al.* Nature 2009

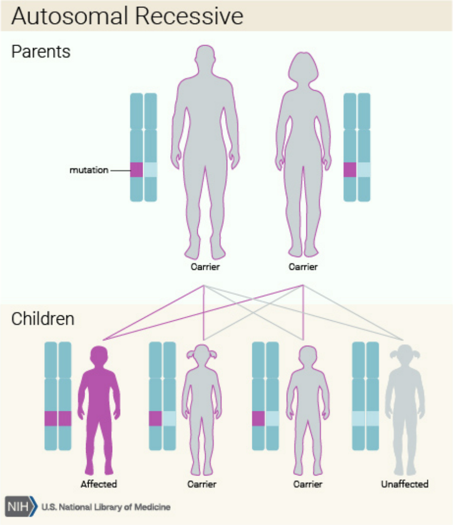
Each rare disease patient is a genetic mystery

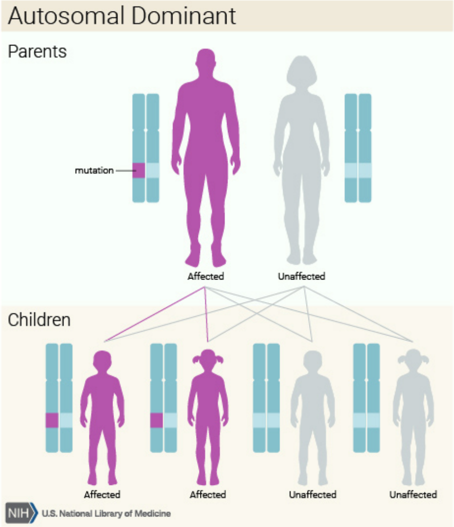
Which of the 3-4 million variants cause disease?

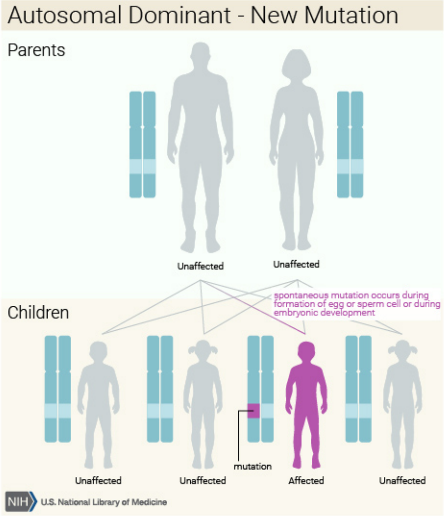
Nemaline Myopathy Family



Inheritance patterns

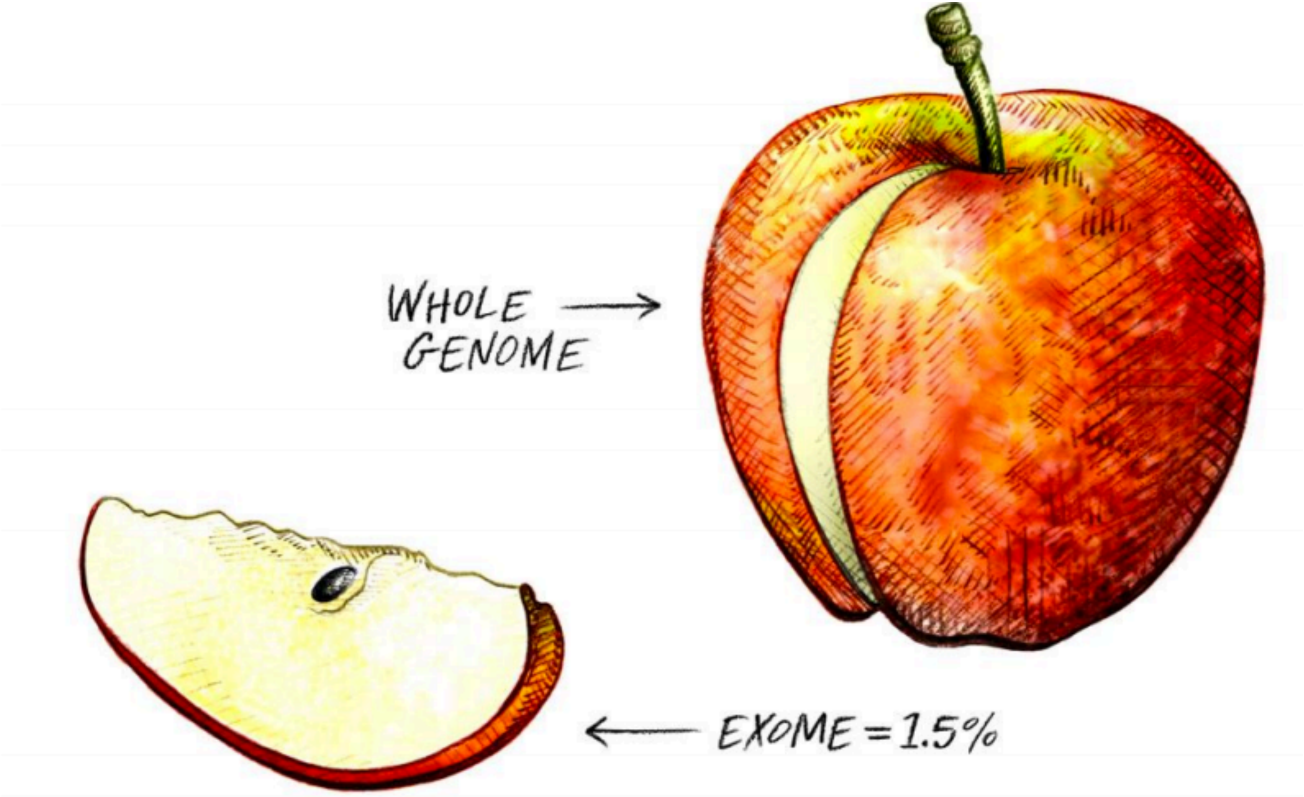




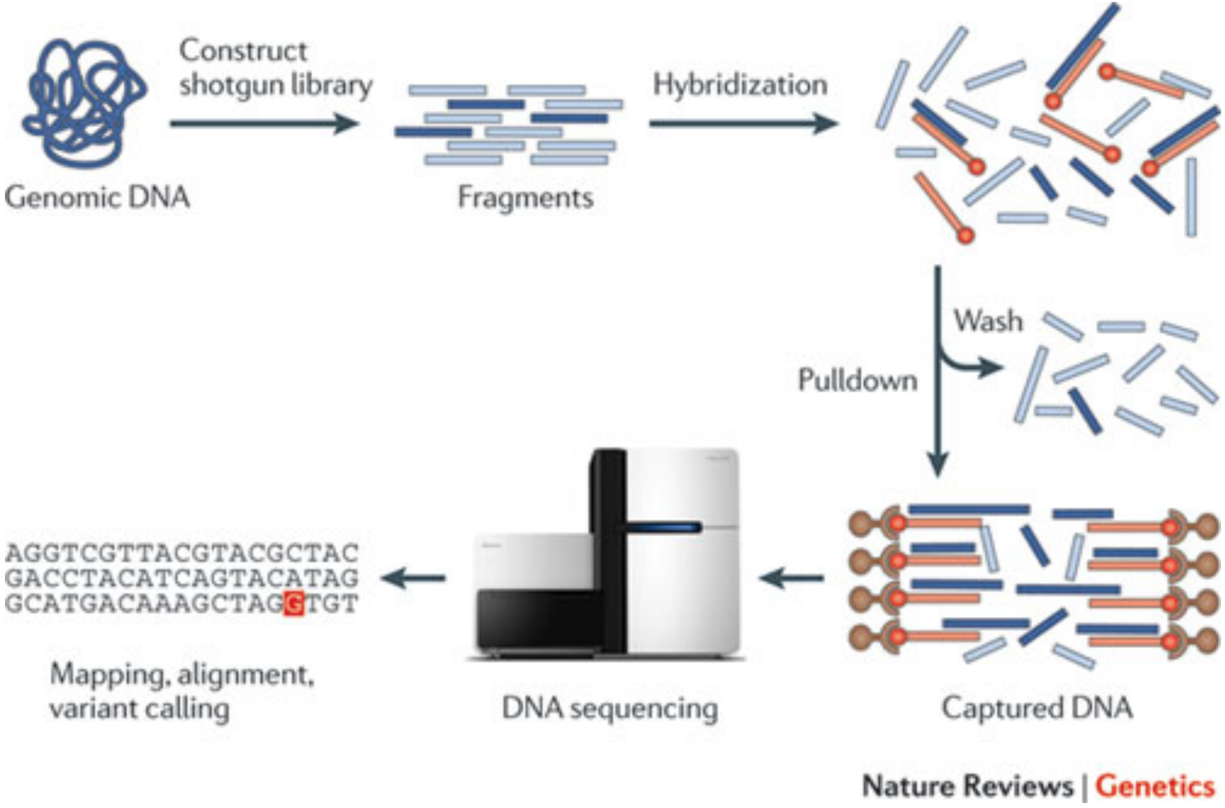

Autosomal Dominant - New Mutation
Parents
Unaffected
Unaffected
spontaneous mutation occurs during formation of egg or sperm cell or during embryonic development
Children
mutation
Unaffected
Unaffected
Affected
Unaffected
NIH U.S. National Library of Medicine

https://ghr.nlm.nih.gov/primer/inheritance/inheritancepatterns

Exome Sequencing



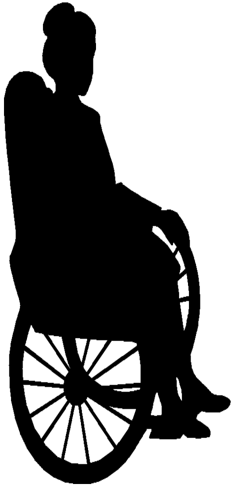
Exome Sequencing



Genetic variants discovered in a typical patient exome

Number of Missense or protein-truncating variants: 11,928

Making sense of one patient exome requires tens of thousands of exomes

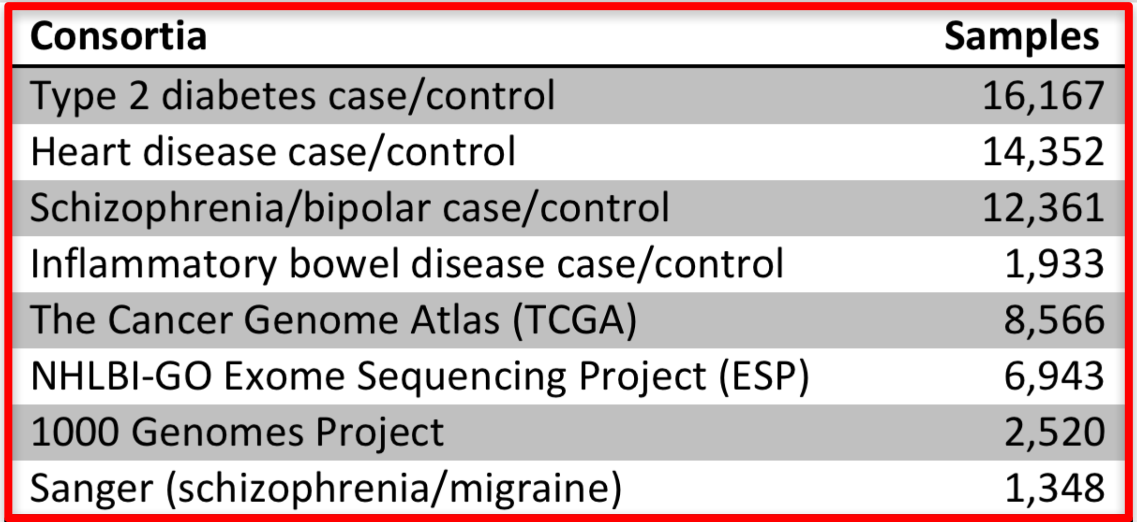


vs

Approximately 500,000 exomes have been sequenced

But these data has been **siloed by project** and **inconsistently processed**

Exome Aggregation Consortium (ExAC): aggregating and calling 92,000 exomes

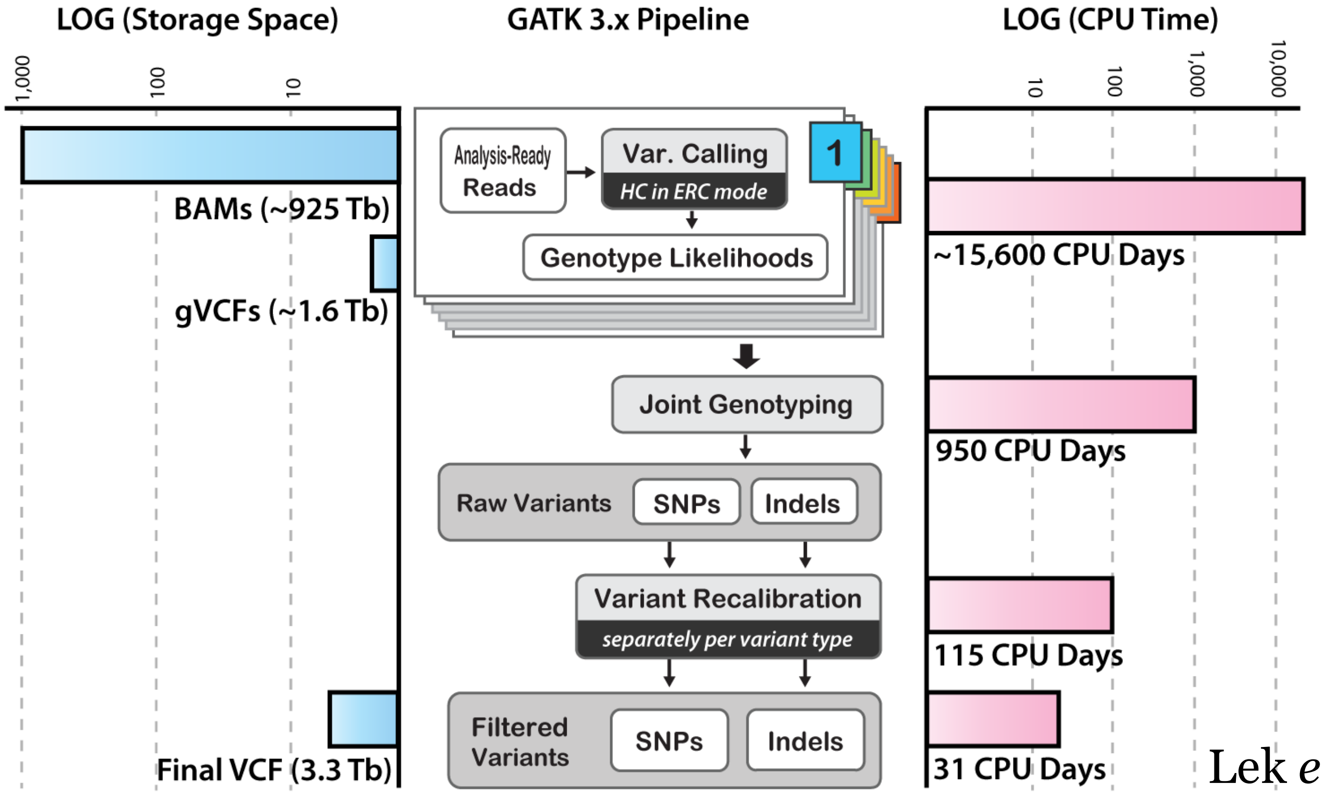


Consortia	Samples
Type 2 diabetes case/control	16,167
Heart disease case/control	14,352
Schizophrenia/bipolar case/control	12,361
Inflammatory bowel disease case/control	1,933
The Cancer Genome Atlas (TCGA)	8,566
NHLBI-GO Exome Sequencing Project (ESP)	6,943
1000 Genomes Project	2,520
Sanger (schizophrenia/migraine)	1,348

Subset of **60,706 “reference” samples**:

- high-quality exomes
- unrelated individuals
- consent for public data sharing
- free of **known** severe pediatric disease

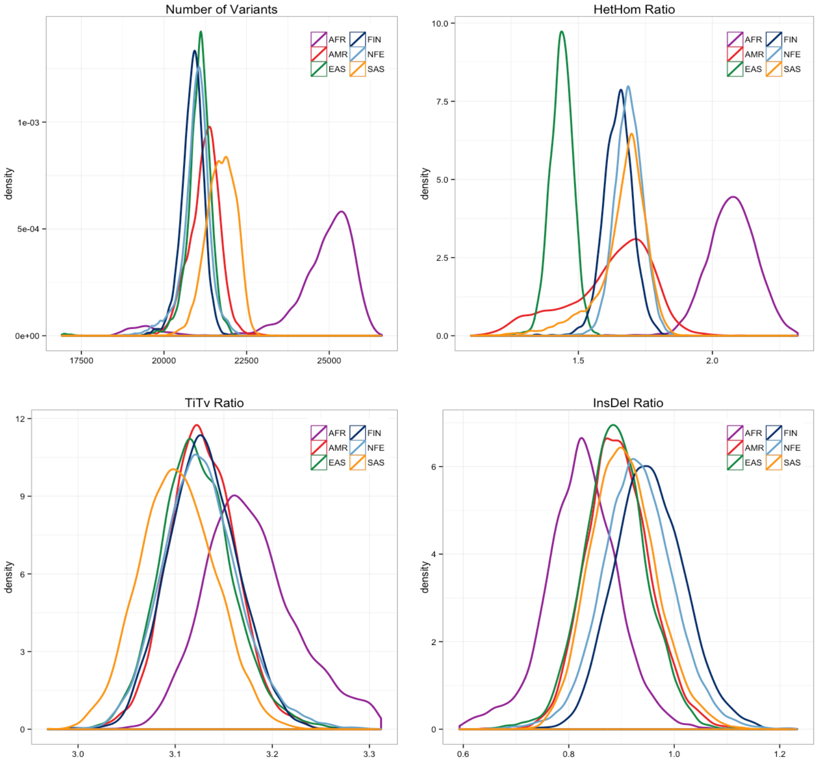
Scalable pipeline for joint discovery



Lek *et. al.* Nature 2016

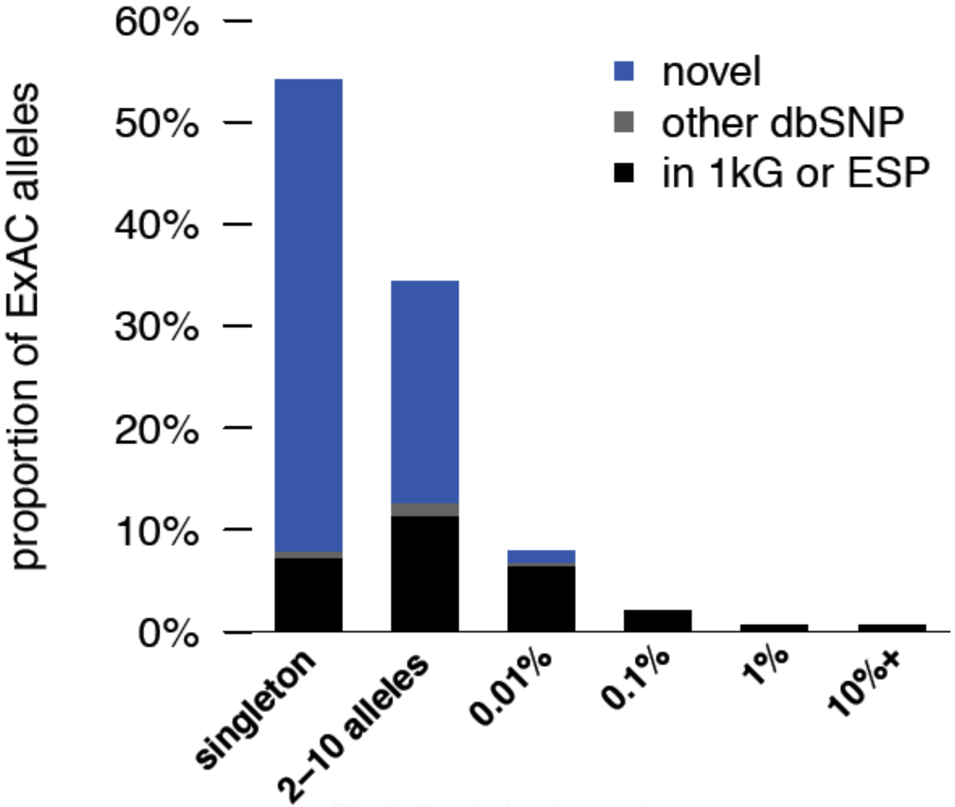
How do we ensure we didn’t produce junk?

Sample Quality Control



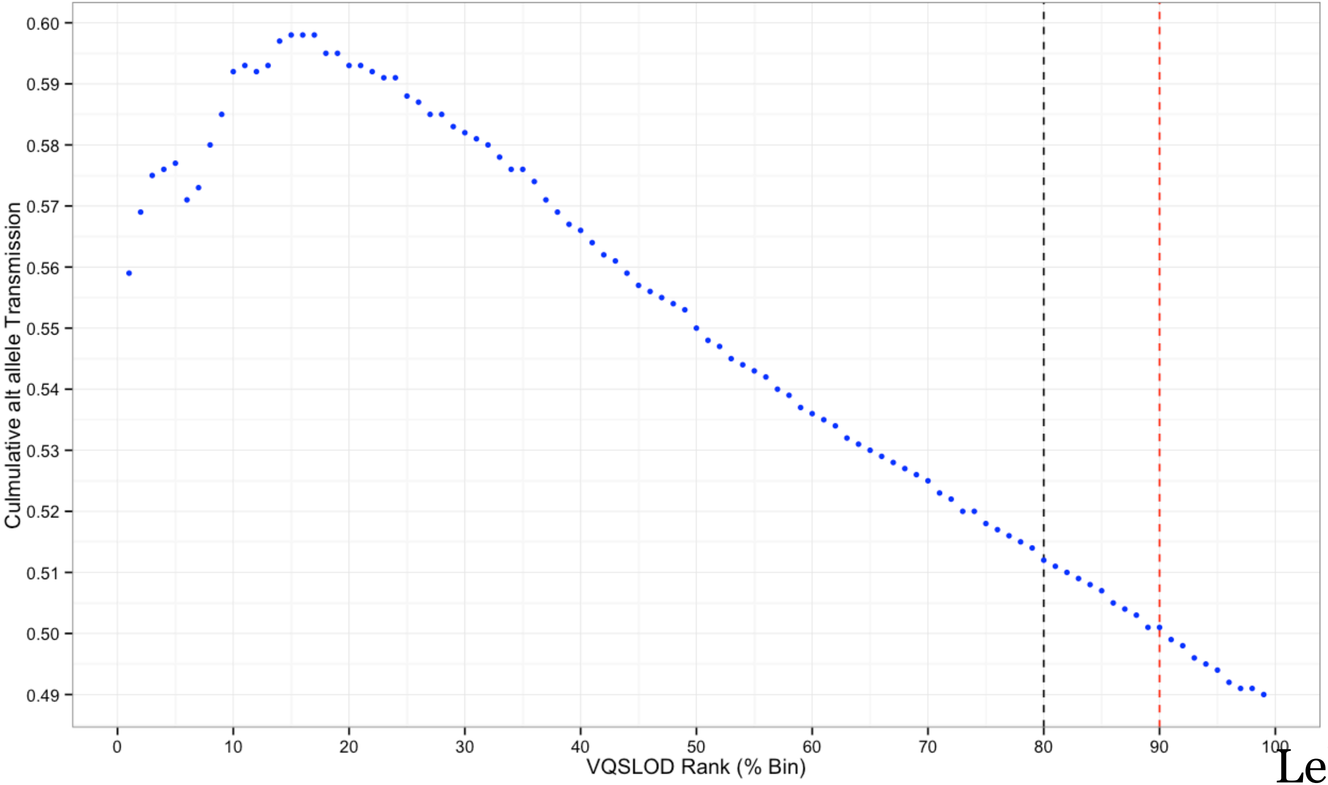
Lek *et. al.* Nature 2016

The majority of variants discovered are ultra rare novel variants



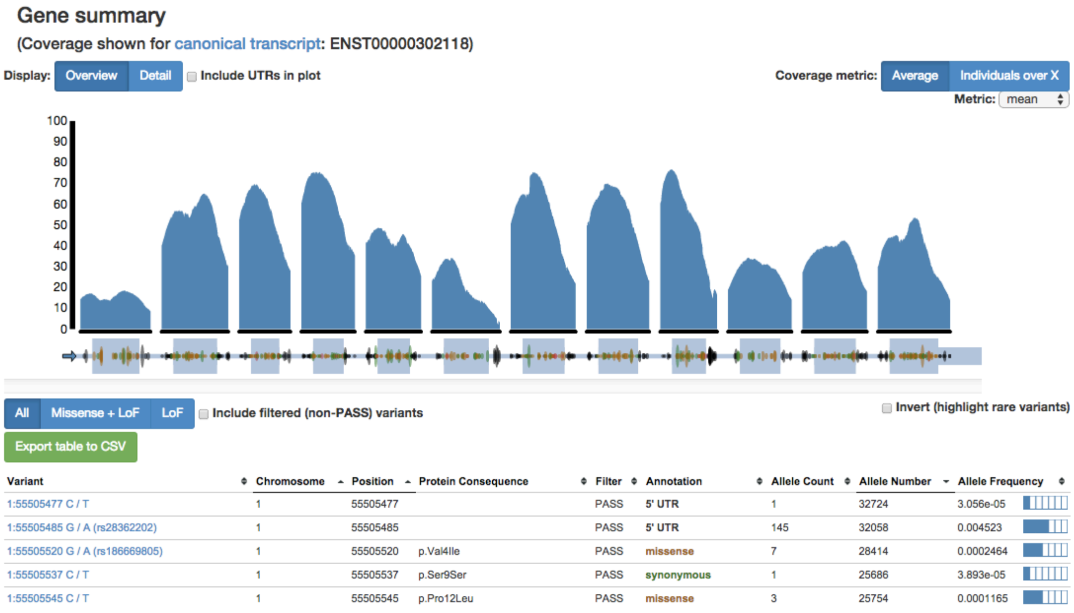
Lek *et. al.* Nature 2016

The majority of variants discovered are ultra rare novel variants



Lek *et. al.* Nature 2016

ExAC Browser (http://exac.broadinstitute.org)



Gene summary

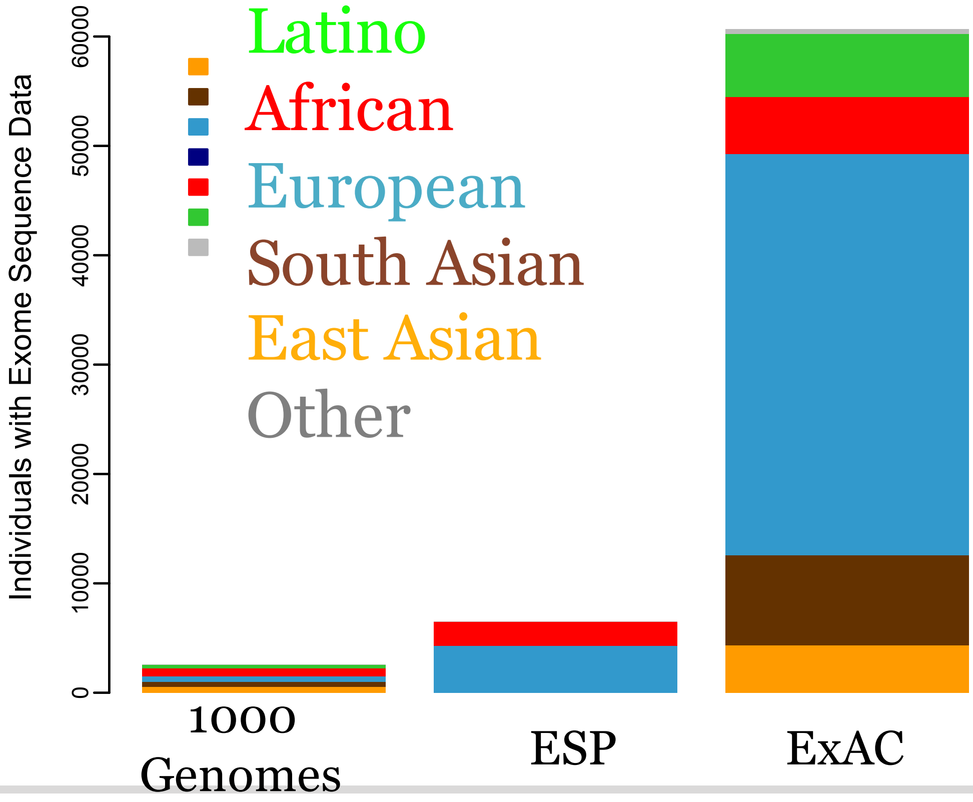
(Coverage shown for canonical transcript: ENST00000302118)

Variant	Chromosome	Position	Protein Consequence	Filter	Annotation	Allele Count	Allele Number	Allele Frequency
1:55505477 C / T	1	55505477		PASS	5' UTR	1	32724	3.056e-05
1:55505485 G / A (rs28362202)	1	55505485		PASS	5' UTR	145	32058	0.004523
1:55505520 G / A (rs186669805)	1	55505520	p.Val4Ile	PASS	missense	7	28414	0.0002464
1:55505537 C / T	1	55505537	p.Ser9Ser	PASS	synonymous	1	25686	3.893e-05
1:55505545 C / T	1	55505545	p.Pro12Leu	PASS	missense	3	25754	0.0001165

>12 million page views

>70,000 page view/week

Unprecedented size and diversity



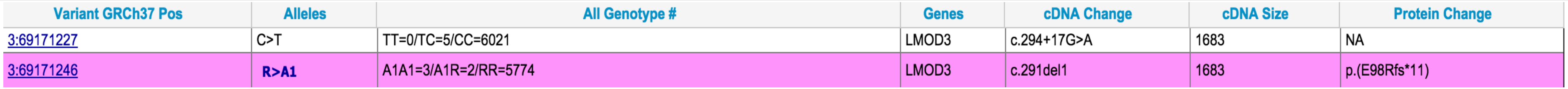
Lek *et. al.* Nature 2016

Rare genetic variants discovered in a typical patient exome

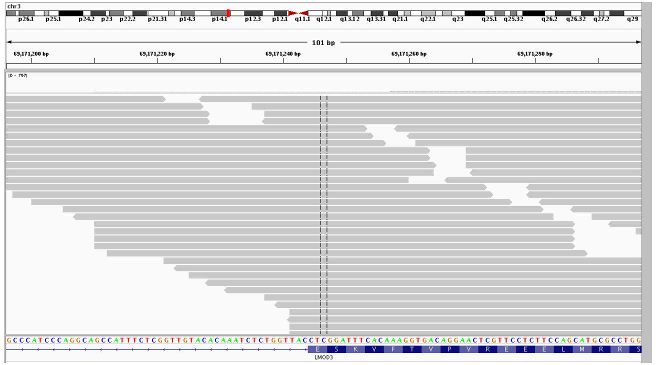
Number of Missense or protein-truncating variants: 11,928

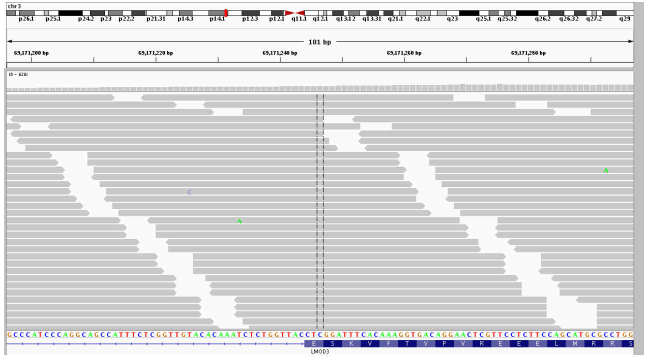
Number of **Rare** Missense or protein-truncating variants: 132

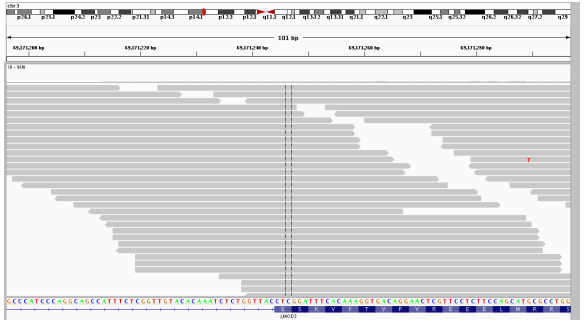
Empowering novel disease gene discovery



Variant GRCh37 Pos	Alleles	All Genotype #	Genes	cDNA Change	cDNA Size	Protein Change
3:69171227	C>T	TT=0/TC=5/CC=6021	LMOD3	c.294+17G>A	1683	NA
3:69171246	R>A1	A1A1=3/A1R=2/RR=5774	LMOD3	c.291del1	1683	p.(E98Rfs*11)

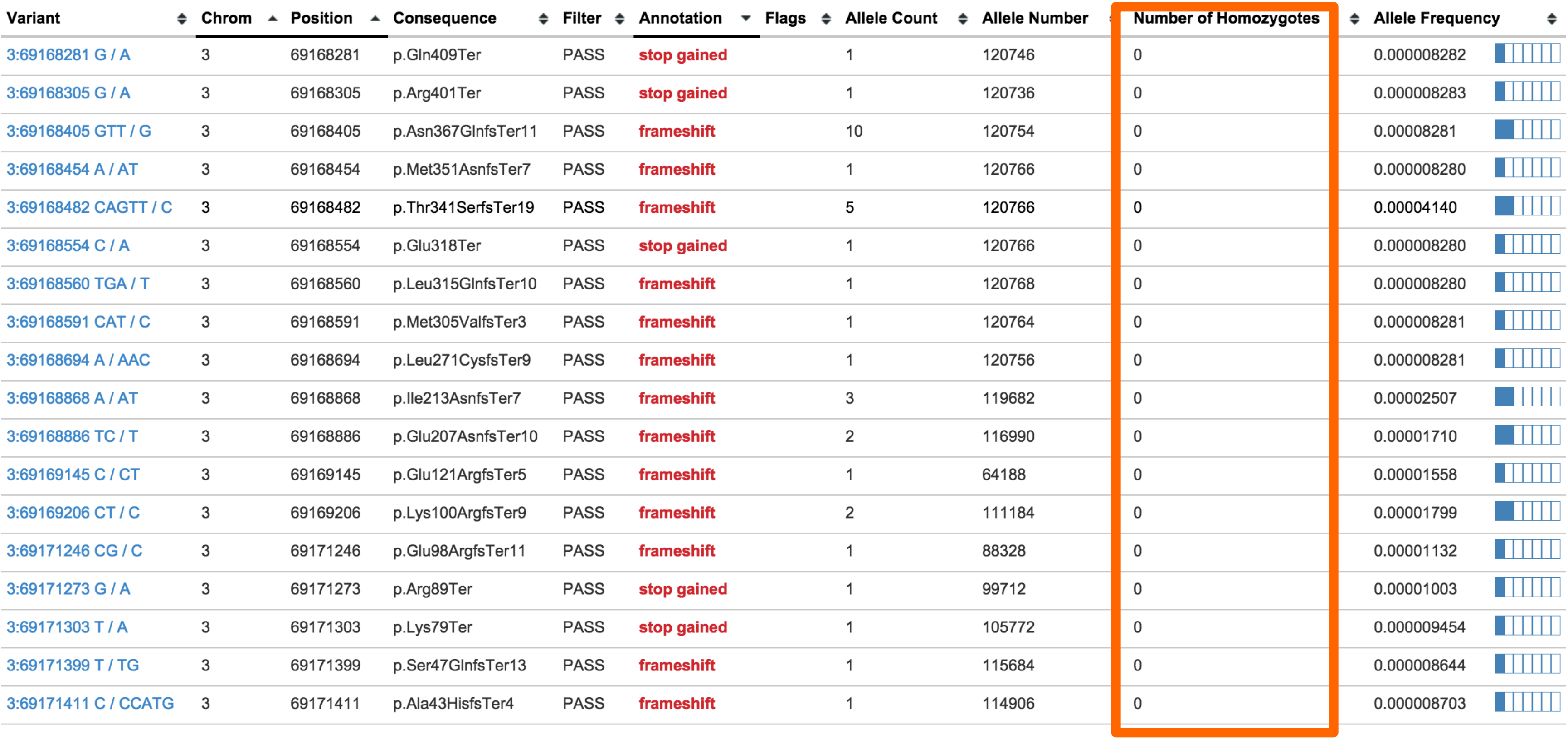






LMOD3: 5 homozygous frameshift variants across 2 different variant sites

LMOD3: Protein truncating variants are rare with no homozygous individuals



Variant	Chrom	Position	Consequence	Filter	Annotation	Flags	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
3:69168281 G / A	3	69168281	p.Gln409Ter	PASS	stop gained		1	120746	0	0.000008282
3:69168305 G / A	3	69168305	p.Arg401Ter	PASS	stop gained		1	120736	0	0.000008283
3:69168405 GTT / G	3	69168405	p.Asn367GlnfsTer11	PASS	frameshift		10	120754	0	0.00008281
3:69168454 A / AT	3	69168454	p.Met351AsnfsTer7	PASS	frameshift		1	120766	0	0.000008280
3:69168482 CAGTT / C	3	69168482	p.Thr341SerfsTer19	PASS	frameshift		5	120766	0	0.00004140
3:69168554 C / A	3	69168554	p.Glu318Ter	PASS	stop gained		1	120766	0	0.000008280
3:69168560 TGA / T	3	69168560	p.Leu315GlnfsTer10	PASS	frameshift		1	120768	0	0.000008280
3:69168591 CAT / C	3	69168591	p.Met305ValfsTer3	PASS	frameshift		1	120764	0	0.000008281
3:69168694 A / AAC	3	69168694	p.Leu271CysfsTer9	PASS	frameshift		1	120756	0	0.000008281
3:69168868 A / AT	3	69168868	p.Ile213AsnfsTer7	PASS	frameshift		3	119682	0	0.00002507
3:69168886 TC / T	3	69168886	p.Glu207AsnfsTer10	PASS	frameshift		2	116990	0	0.00001710
3:69169145 C / CT	3	69169145	p.Glu121ArgfsTer5	PASS	frameshift		1	64188	0	0.00001558
3:69169206 CT / C	3	69169206	p.Lys100ArgfsTer9	PASS	frameshift		2	111184	0	0.00001799
3:69171246 CG / C	3	69171246	p.Glu98ArgfsTer11	PASS	frameshift		1	88328	0	0.00001132
3:69171273 G / A	3	69171273	p.Arg89Ter	PASS	stop gained		1	99712	0	0.00001003
3:69171303 T / A	3	69171303	p.Lys79Ter	PASS	stop gained		1	105772	0	0.000009454
3:69171399 T / TG	3	69171399	p.Ser47GlnfsTer13	PASS	frameshift		1	115684	0	0.000008644
3:69171411 C / CCATG	3	69171411	p.Ala43HisfsTer4	PASS	frameshift		1	114906	0	0.000008703

ExAC Principal Investigators

Daniel MacArthur
David Altshuler
Diego Ardissino
Michael Boehnke
Mark Daly
John Danesh
Roberto Elosua
Gad Getz
Christina Hultman
Sekar Kathiresan
Markku Laakso
Steven McCarroll
Mark McCarthy
Ruth McPherson
Benjamin Neale
Aarno Palotie
Shaun Purcell
Danish Saleheen
Jeremiah Scharf
Pamela Sklar
Patrick Sullivan
Jaakko Tuomilehto
Hugh Watkins

Analysis

Monkol Lek
Konrad Karczewski
Eric Minikel
Kaitlin Samocha
Anne O'Donnell
James Ware
Beryl Cummings
Andrew Hill
Menachem Fromer
Doug Ruderfer

Production

Monkol Lek
Eric Banks
Timothy Fennell
Ryan Poplin
Fengmei Zhao

Website

Konrad Karczewski
Brett Thomas
Ben Weisburd

Contributing cohorts

1000 Genomes
Bulgarian Trios
Finland-United States Investigation of NIDDM Genetics (FUSION)
GoT2D
Inflammatory Bowel Disease
METabolic Syndrome In Men (METSIM)
Myocardial Infarction Genetics Consortium:

- Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
- Ottawa Genomics Heart Study
- Pakistan Risk of Myocardial Infarction Study (PROMIS)
- Precocious Coronary Artery Disease Study (PROCARDIS)
- Registre Gironi del COR (REGICOR)

NHLBI-GO Exome Sequencing Project (ESP)
National Institute of Mental Health (NIMH) Controls
SIGMA-T2D
Sequencing in Suomi (SISu)
Swedish Schizophrenia & Bipolar Studies
T2D-GENES
Taiwanese Trios
The Cancer Genome Atlas (TCGA)
Tourette Syndrome Association International Consortium for Genomics (TSAICG)

Broad Genomics and Data Sciences Platforms

What can you do with a large collection of rare protein-coding variation?

Gene Constraint

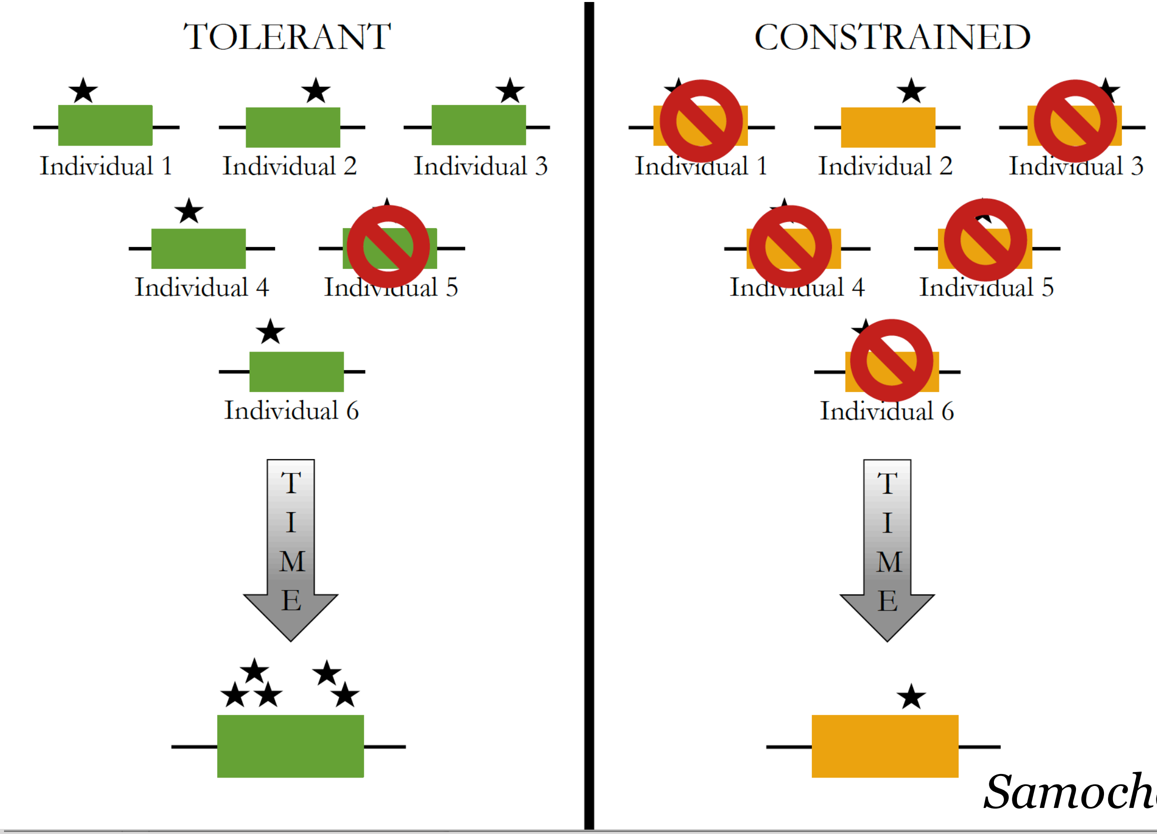
Kaitlin Samocha

Samocha et. al. Nat. Genet. 2014

Lek et. al. Nature 2016

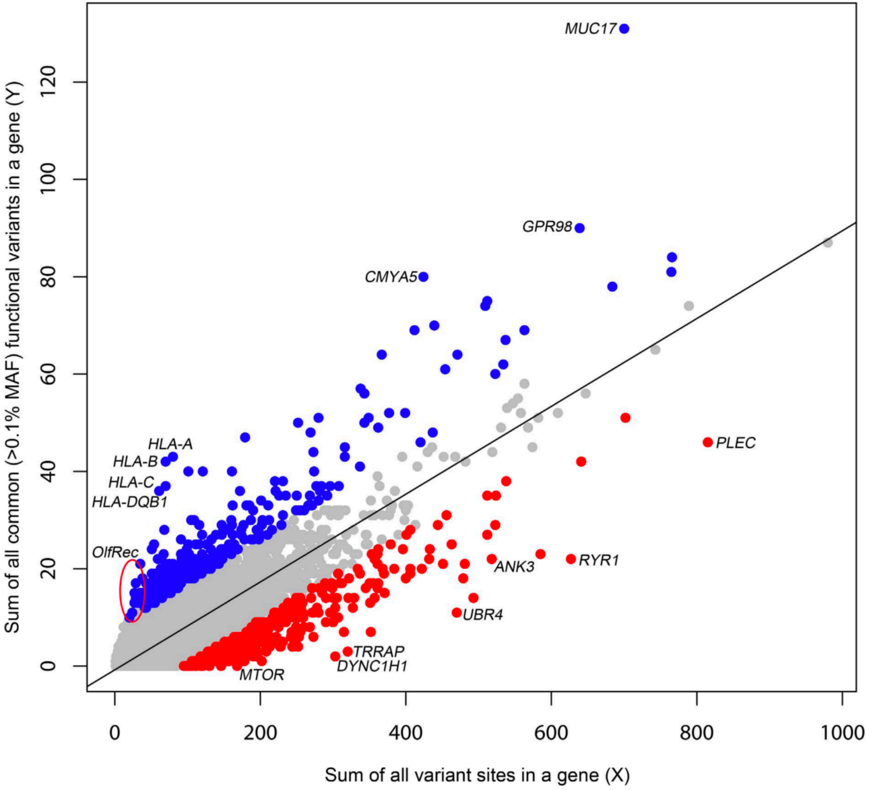
Samocha et. al. BioRxiv 2017

Constrained genes accumulate less variation



Samocha et. al. Nat. Genet. 2014

Residual Variation Intolerance Score (RVIS)

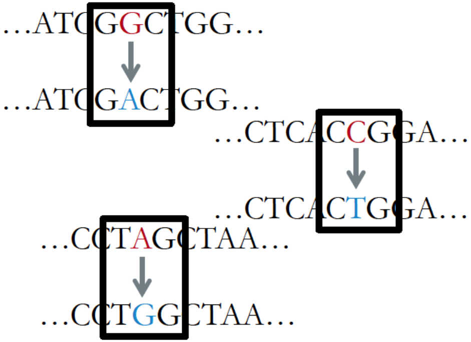


Top 2% most intolerant

Top 2% most tolerant

Petrovski et. al. PLoS 2013

Sequence context model of mutation



…ATCGGCTGG…

…ATCGACTGG…

…CTCACCGGA…

…CTCACTGGA…

…CCTAGCTAA…

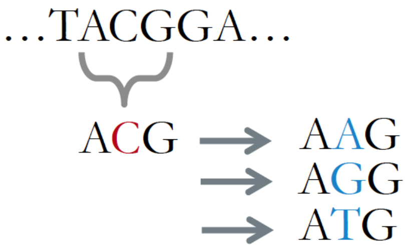
…CCTGGCTAA…

Change	Probability
AAA → ACA	*a*
AAA → AGA	*b*
AAA → ATA	*c*
AAC → ACC	*d*
AAC → AGC	*e*
AAC → ATC	*f*

…

Created a mutation rate table

Used the sequence to determine each gene's probability of mutating



…TACGGA…

ACG → AAG

→ AGG

→ ATG

Per gene:

p(synonymous)

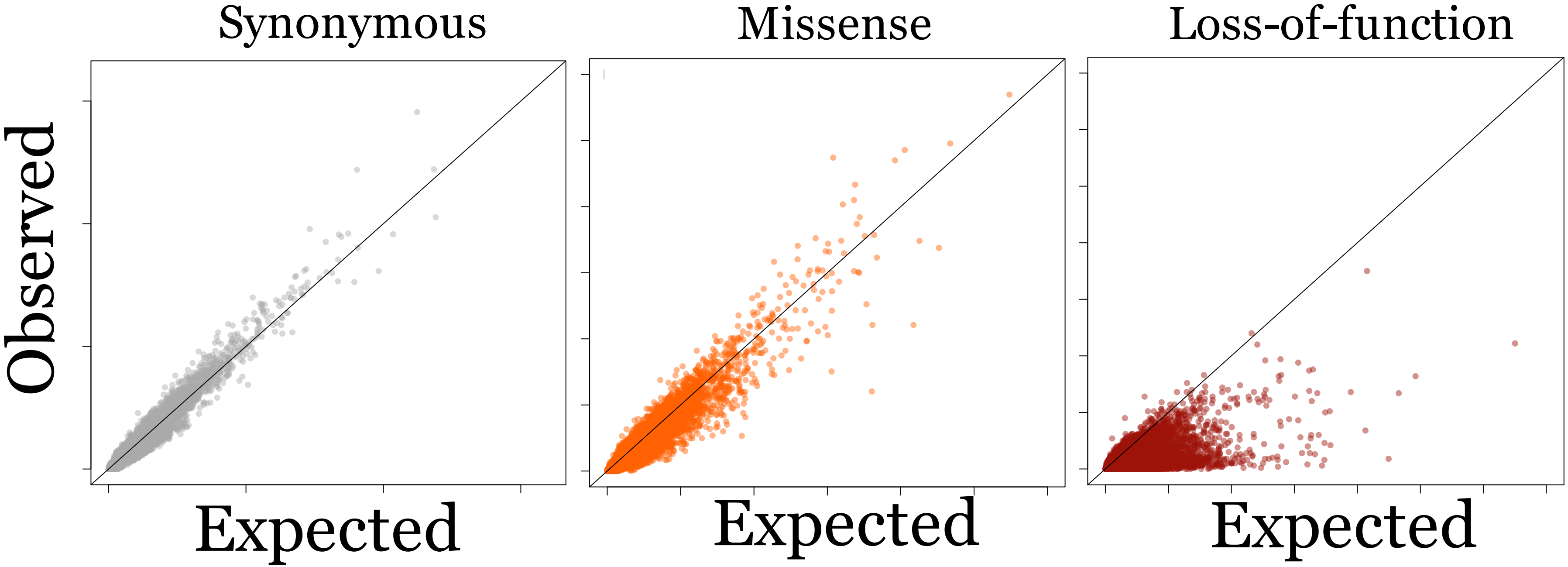
p(missense)

p(nonsense)

p(splice site)

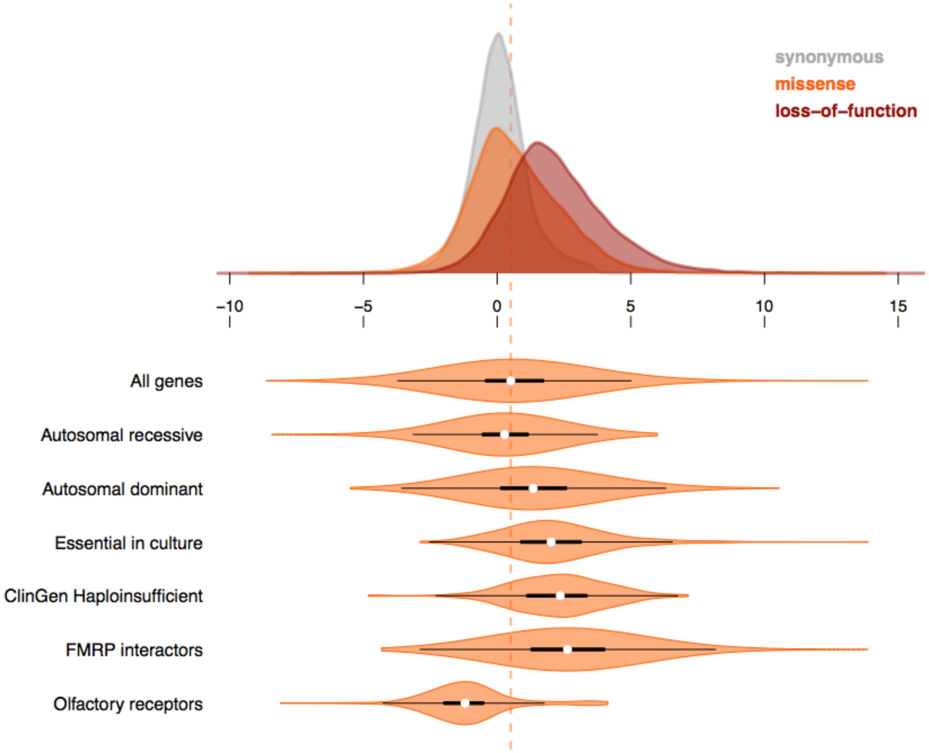
Samocha et. al. Nat. Genet. 2014

Identify genes with significant depletion of variation



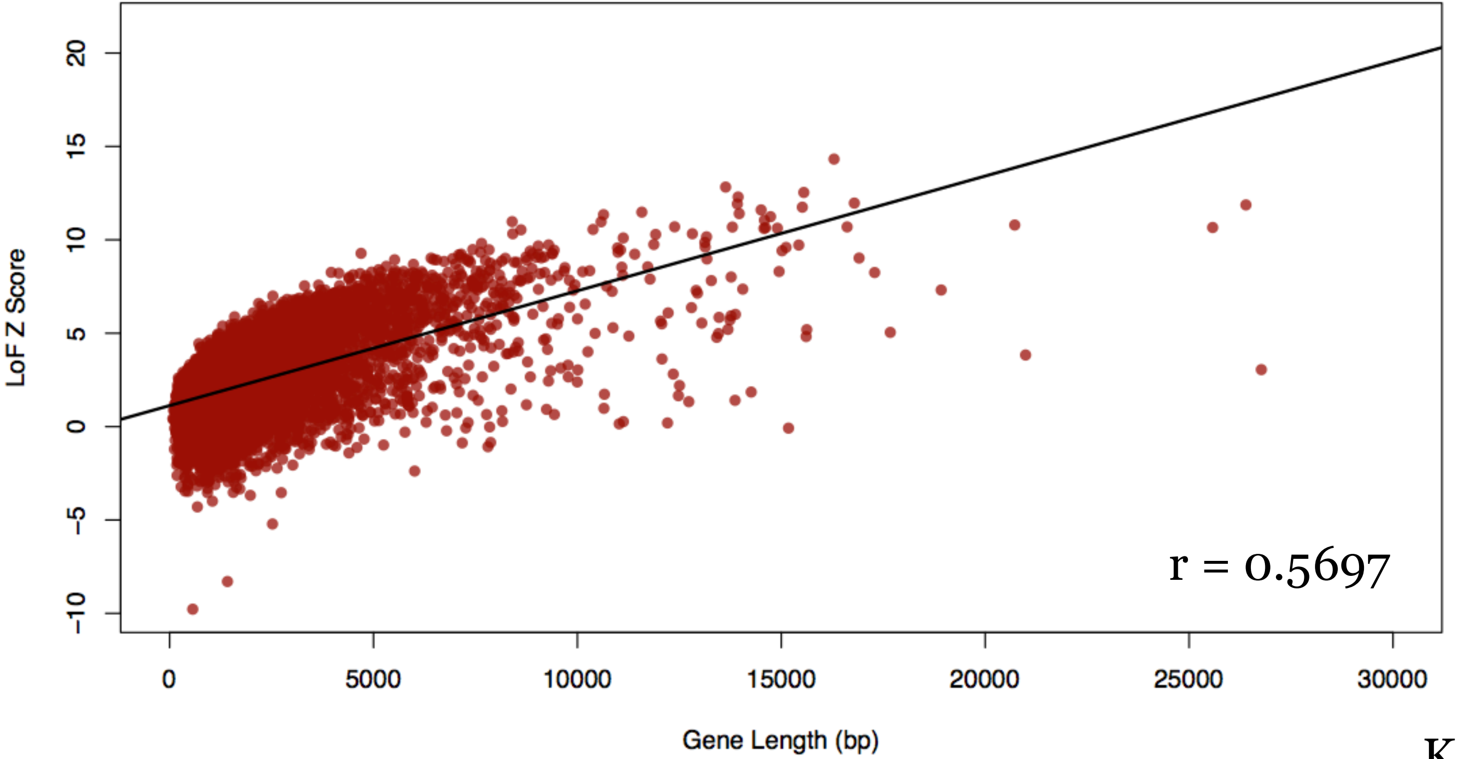
Kaitlin Samocha

Distribution of missense Z scores for gene sets



Kaitlin Samocha

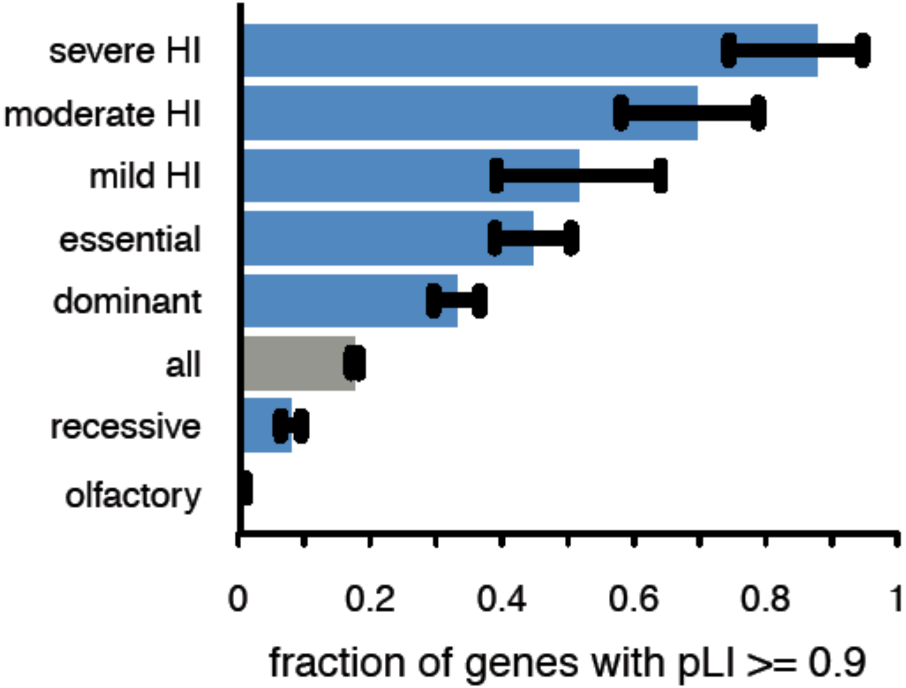
Correlation between gene length and LoF z-score



r = 0.5697

Kaitlin Samocha

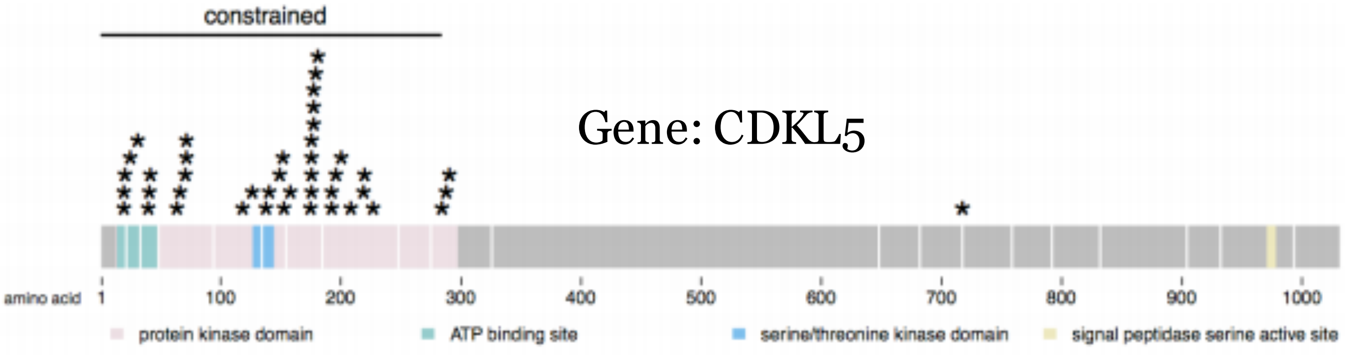
pLI: Identifying genes that are intolerant to loss of function variation



Lek *et. al.* Nature 2016

Weakness of current constraint models

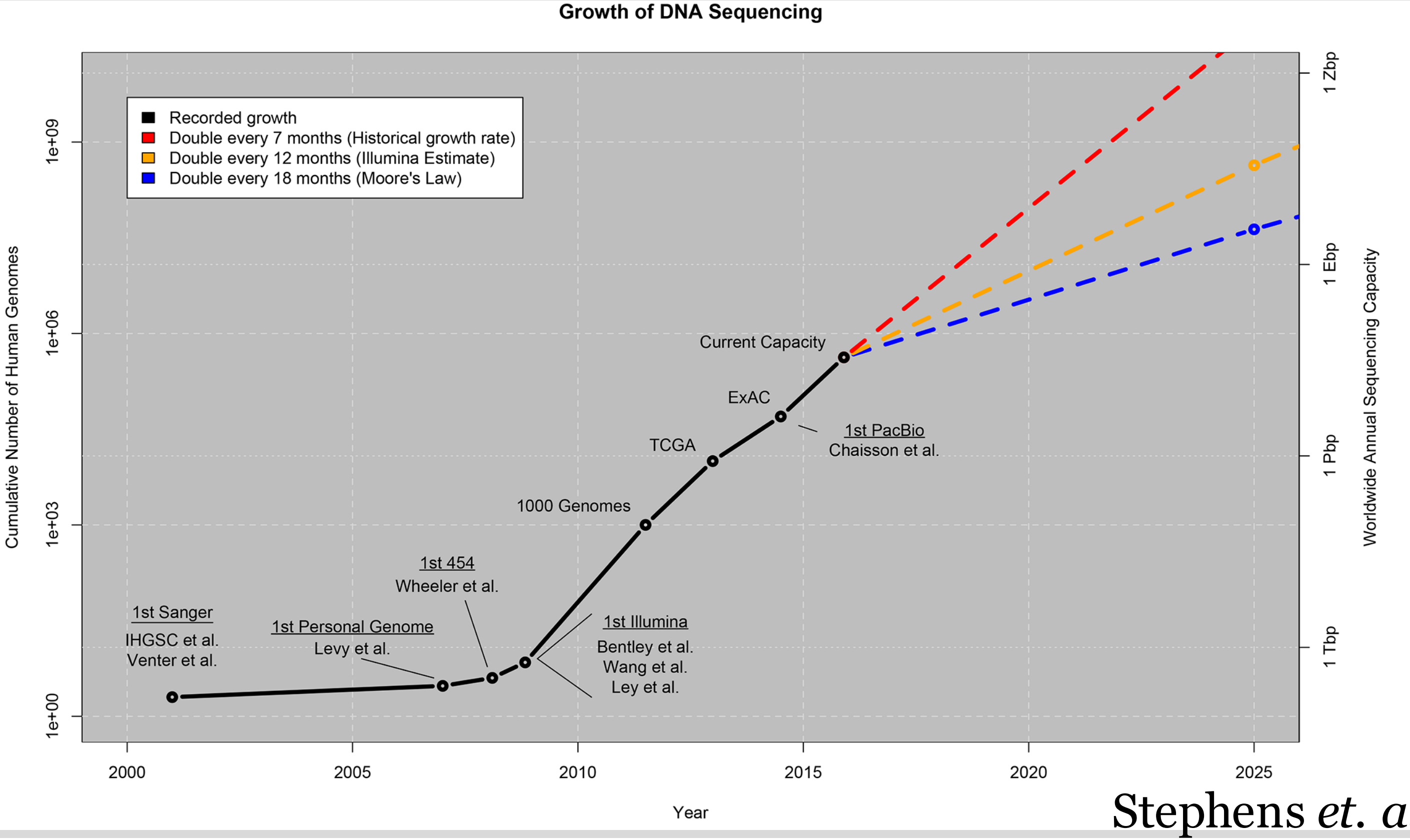
- Insertions/Deletions are currently not modelled
- Tri-nucleotide mutation model vs other models (e.g. heptamer model)
- Only includes coding variants
- Not all genes are represented due to coverage
- Genes may only have particular regions that are constrained



Gene: CDKL5

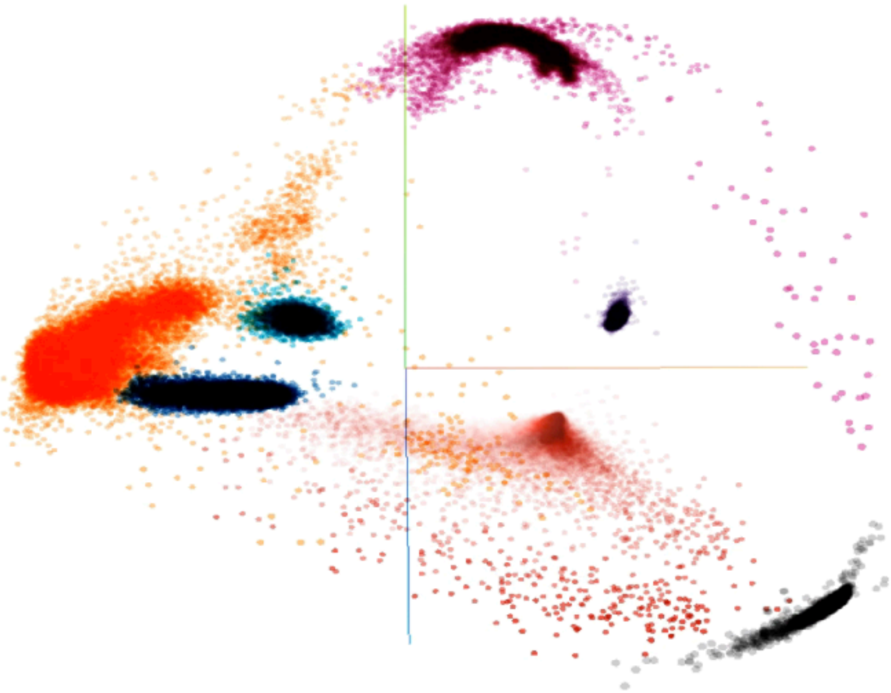
Samocha et. al. BioRxiv 2017

ExAC and Beyond



Stephens *et. al.* PLoS Biology 2015

Genome Aggregation Database (gnomAD) http://gnomad.broadinstitute.org

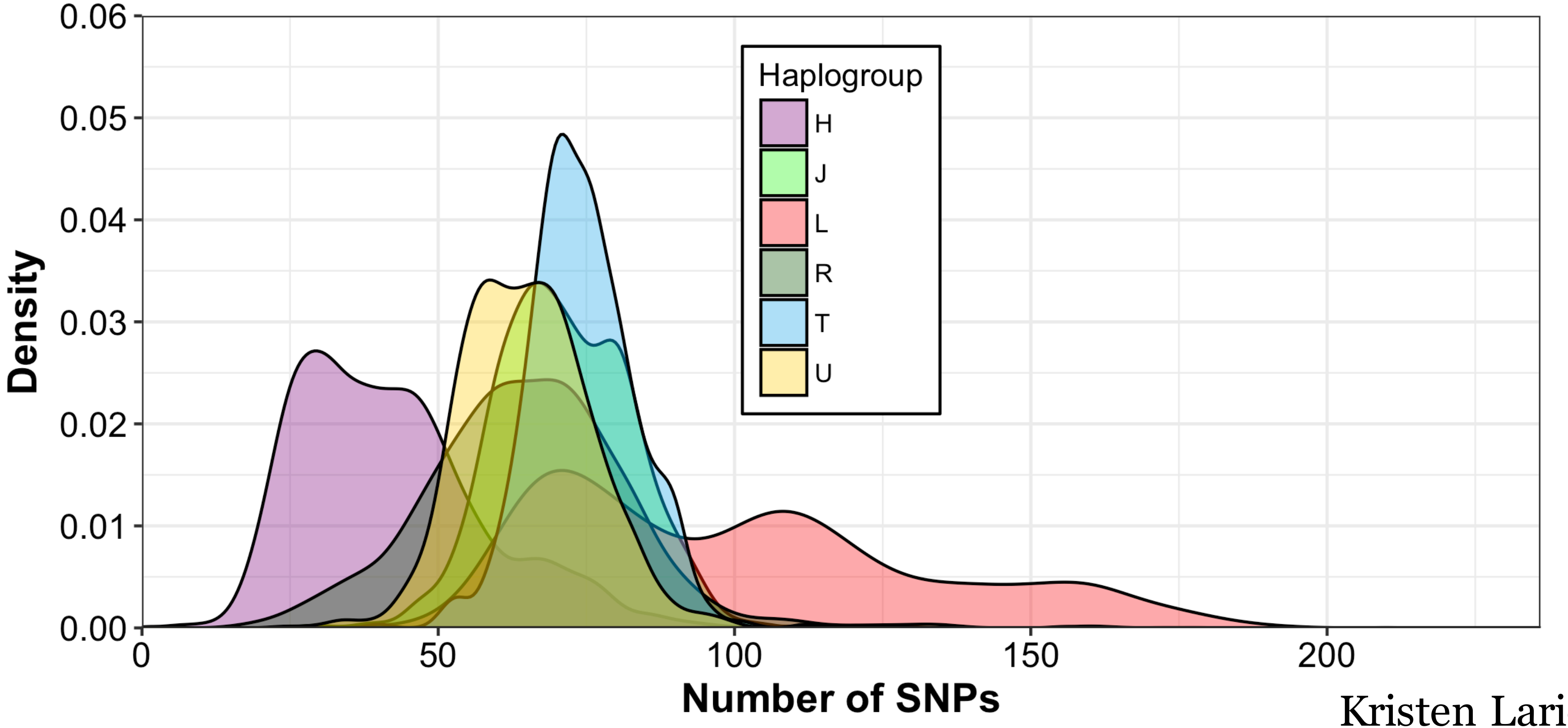


Released ASHG 2016

- Over twice as large
  - 126,216 exomes
  - 15,136 genomes
- Non-coding variants from genomes
- Higher resolution population clustering
- Improved filtering strategy

Konrad Karczewski & Laurent Francioli

gnomAD: Mitochondrial reference panel



Kristen Laricchia

Big Data: The need to scale analysis

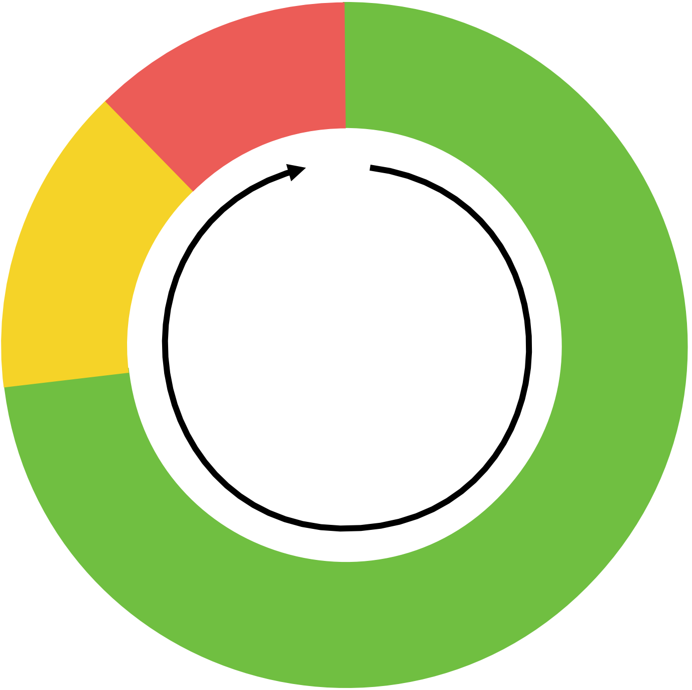
Big Data: The need to scale analysis

Hail Team
Cotton Seed
Tim Poterba

https://hail.is

Hail usage:
UK Biobank
Swedish Schizophrenia & Bipolar exomes
gnomAD

Computational Experiments



Science

Implementation

Runtime

Hail Team

Tools used for ExAC analysis

PLINK

Eigensoft

GATK

KING

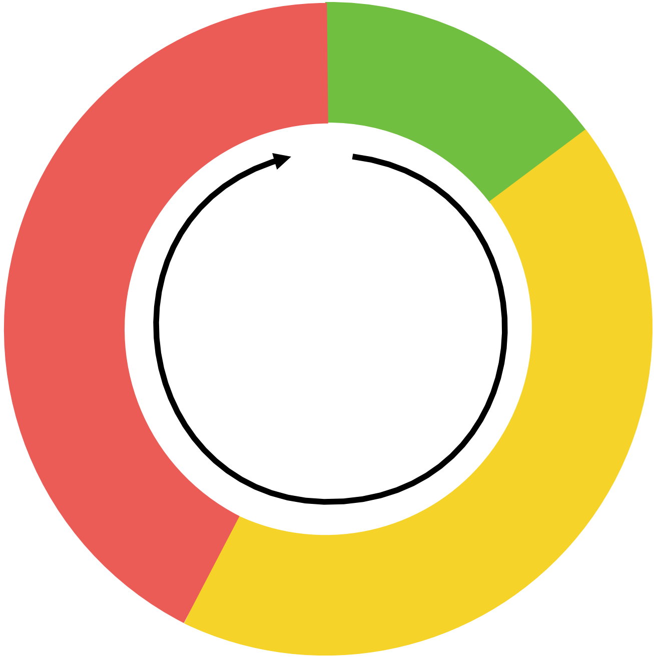
VerifyBAMID

python

R

Hail Team

Failure to scale: ExAC and other projects



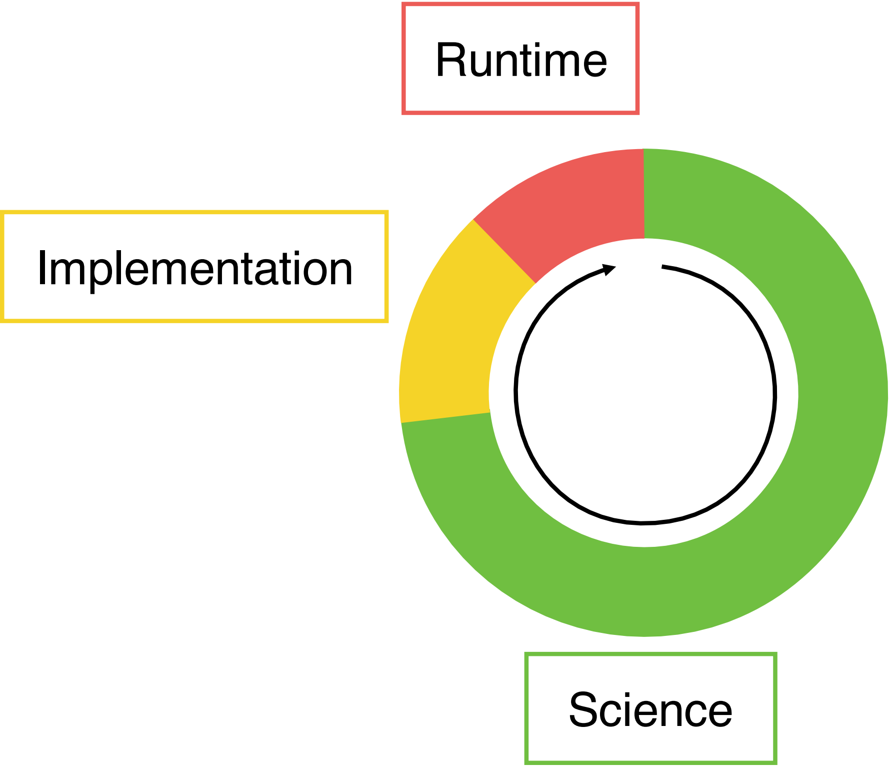
Science

Implementation

Runtime

Hail Team

Ideal computational analysis framework



🟩 Powerful and flexible

🟨 Domain-relevant, easy to use

🟥 Fast and scalable

Hail Team

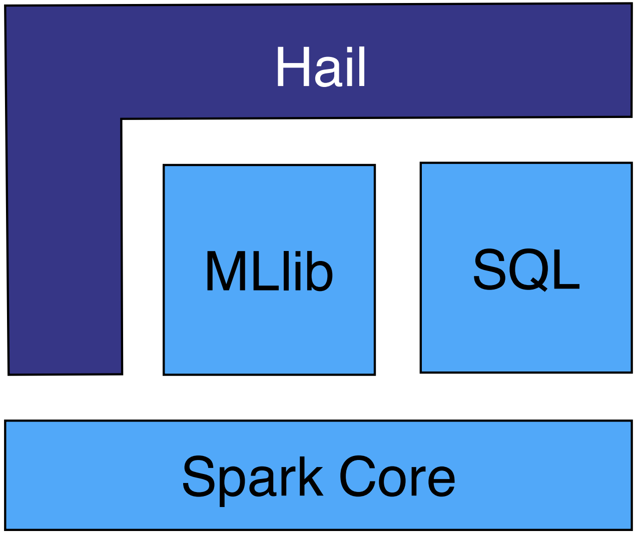
Hail: Scaling Genetic Data Analysis

hail

- genomic data ETL
- high-level APIs for multi-dimensional data query
- stats and ML methods
- Scala, Python

Spark

- scalability
- high-level programming APIs
- linear algebra, MLlib
- Scala, Python, R

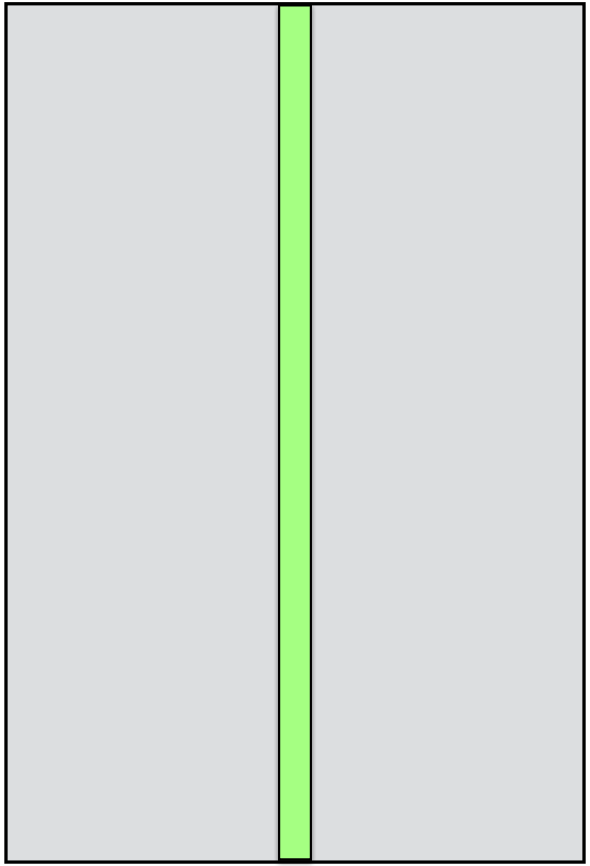


mongoDB Solr cassandra APACHE KUDU Parquet hadoop

Hail Team

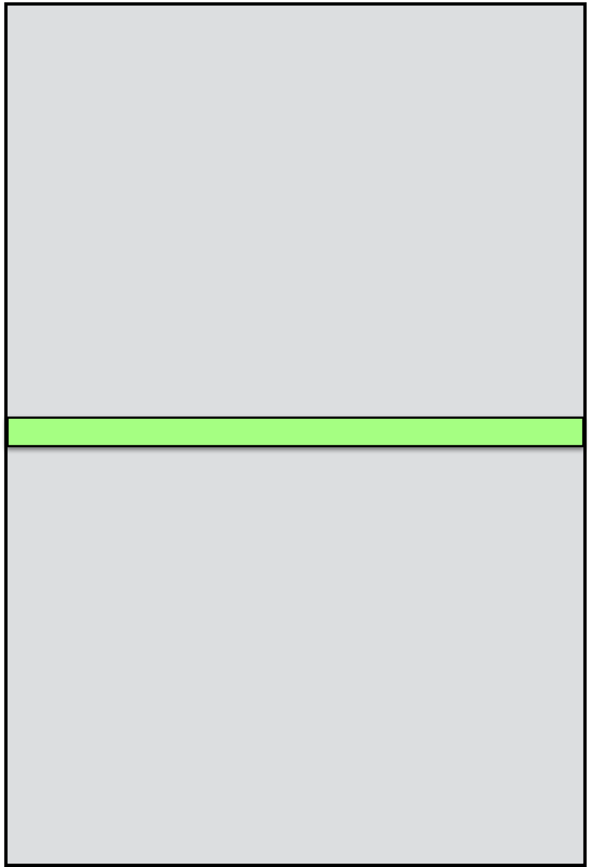
Individual ID

“NA12878”



Individual ID

“NA12878”



Genomic Locus

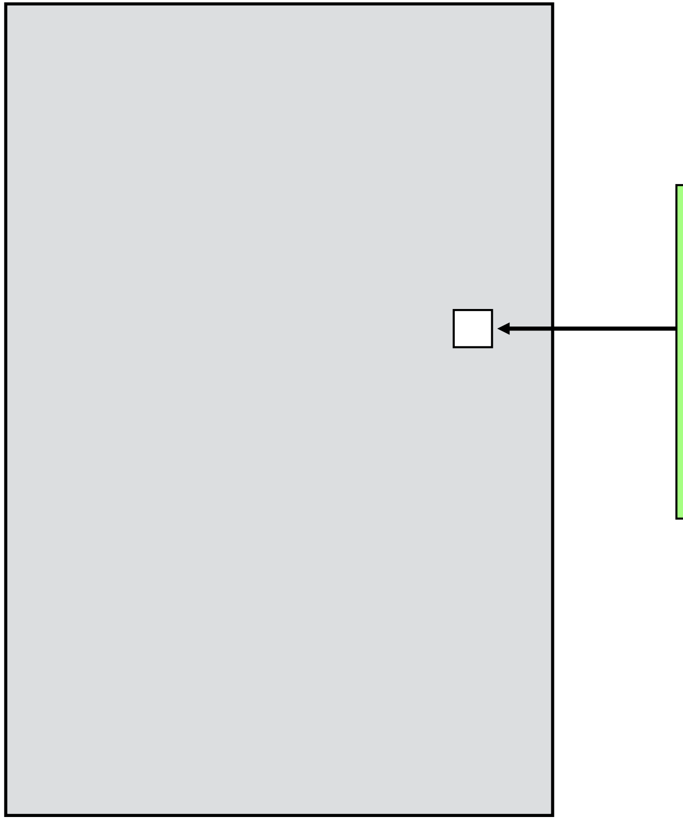
```
{
  “chromosome”: 1,
  “position”: 16123092,
  “reference”: “A”,
 “alternate”: “T”
}
```

Individual ID

“NA12878”

Genomic Locus

```
{
  “chromosome”: 1,
  “position”: 16123092,
  “reference”: “A”,
  “alternate”: “T”
}
```



Genotype

```
{
  “call”: “A/T”,
  “reads”: [10, 8],
  “quality”: 43,
  “p”: [43, 0, 52]
}
```

Individual ID

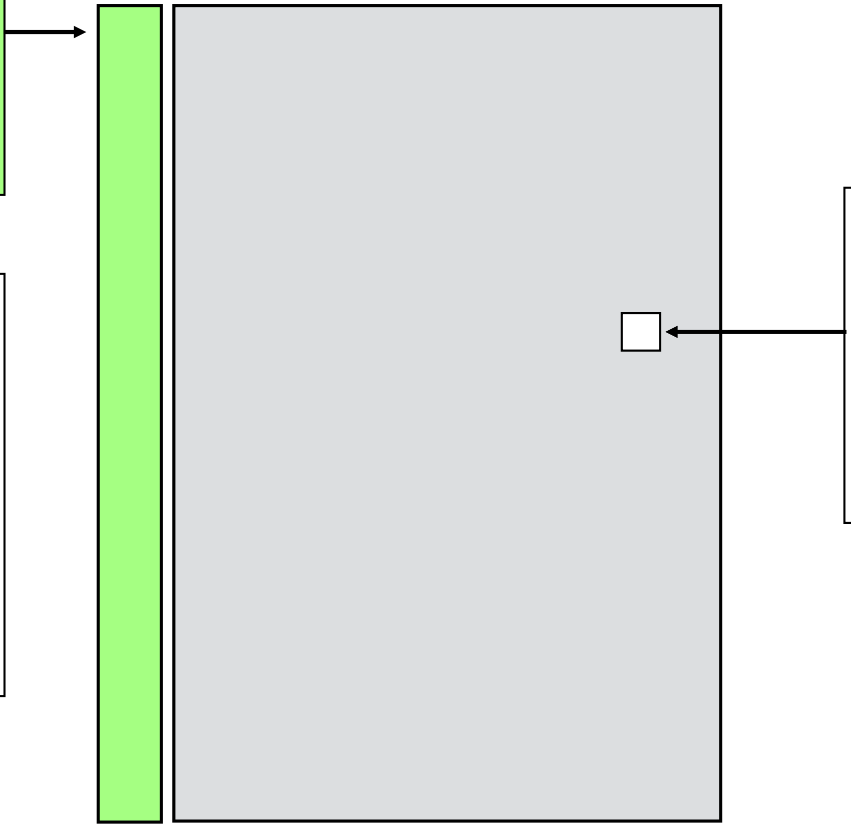
“NA12878”

Locus-indexed table

```
{
  “gene”: “SHH”,
  “pred_impact”: “high”,
  “pop_frequency”: 0.102
}
```

Genomic Locus

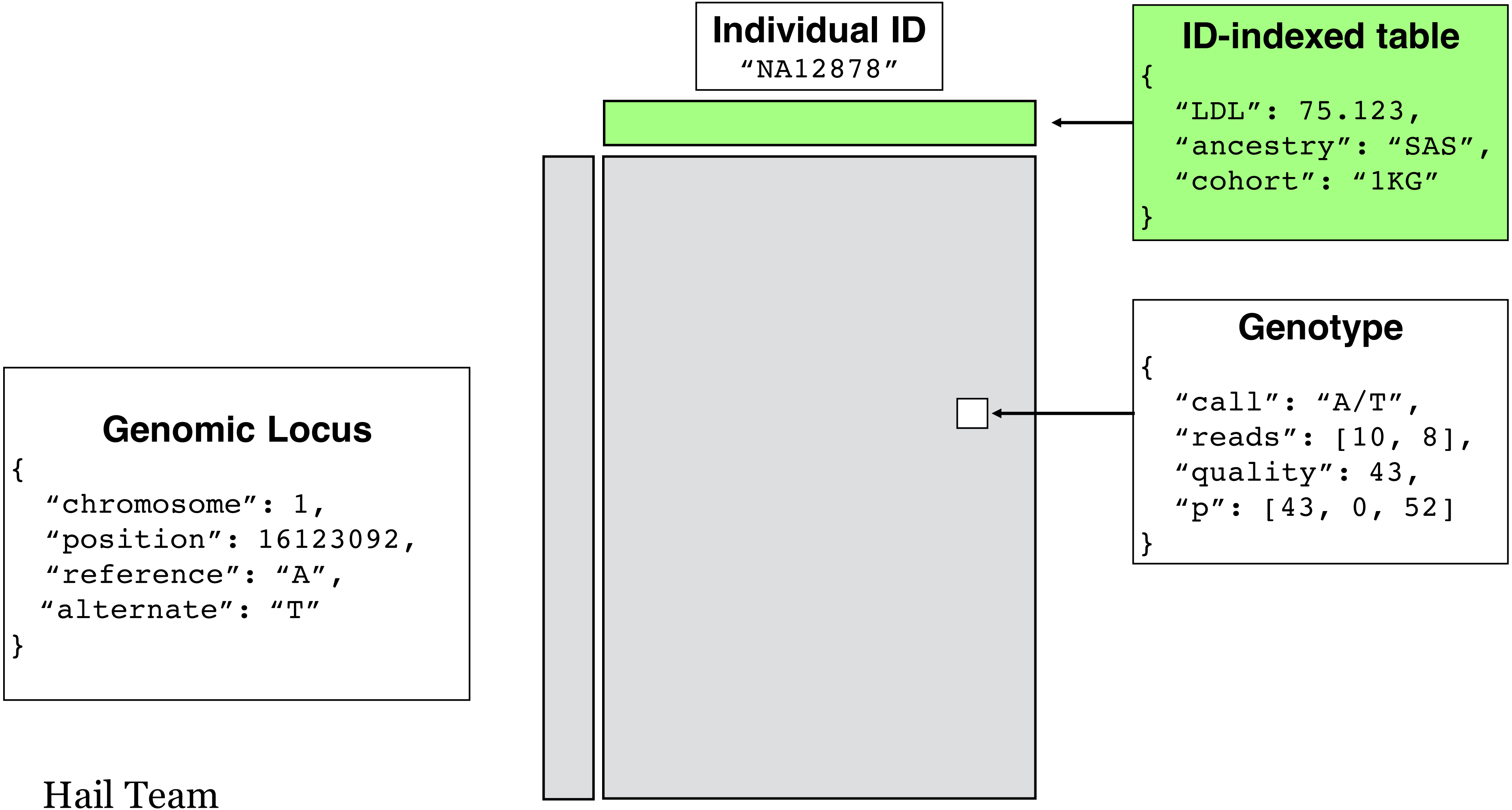
```
{
  “chromosome”: 1,
  “position”: 16123092,
  “reference”: “A”,
 “alternate”: “T”
}
```



Genotype

```
{
  “call”: “A/T”,
  “reads”: [10, 8],
  “quality”: 43,
  “p”: [43, 0, 52]
}
```

Hail Team



Individual ID

“NA12878”

ID-indexed table

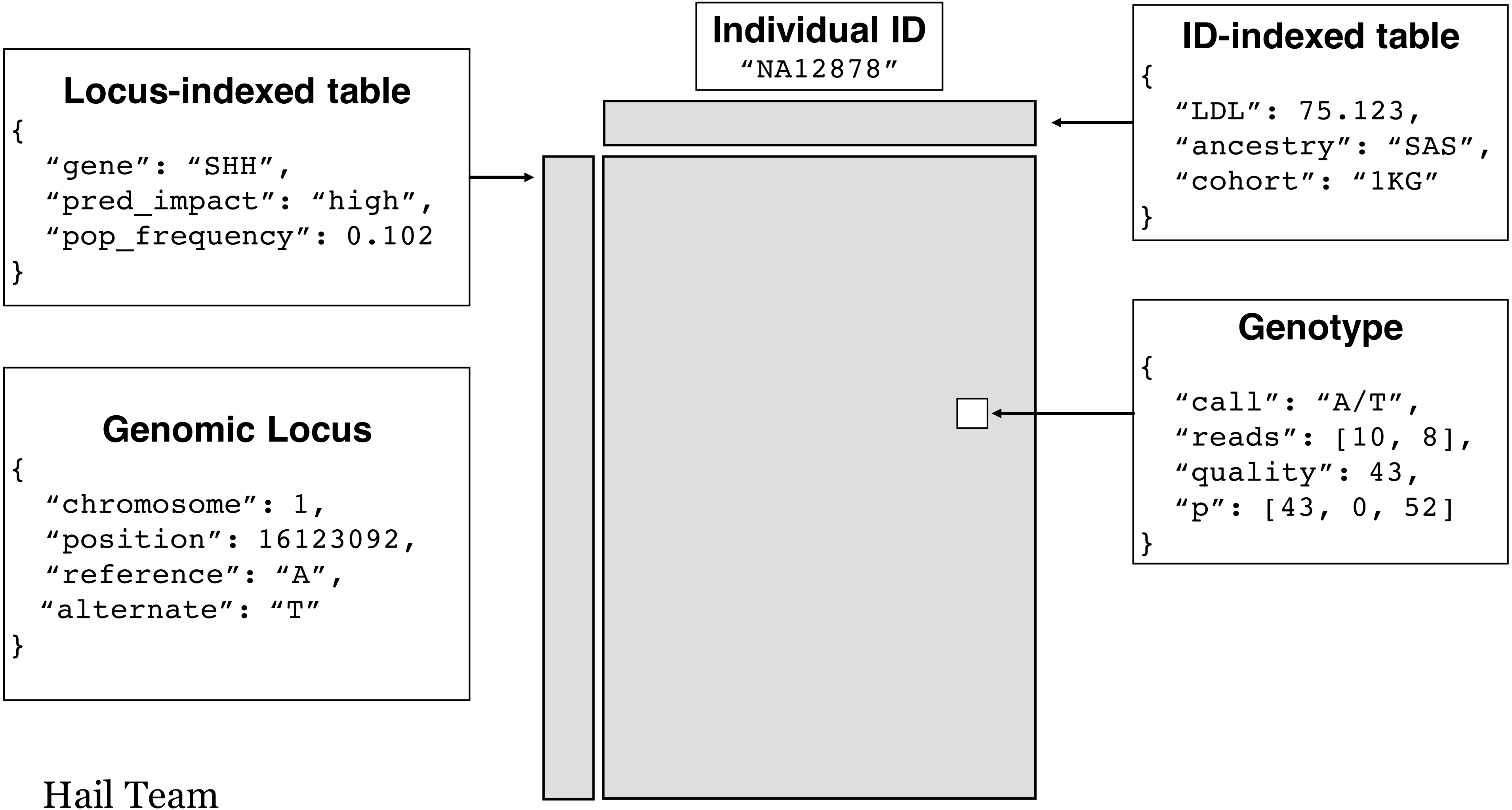
```
{
  “LDL”: 75.123,
  “ancestry”: “SAS”,
  “cohort”: “1KG”
}
```

Genotype

```
{
  “call”: “A/T”,
  “reads”: [10, 8],
  “quality”: 43,
  “p”: [43, 0, 52]
}
```

Genomic Locus

```
{
  “chromosome”: 1,
  “position”: 16123092,
  “reference”: “A”,
 “alternate”: “T”
}
```



Individual ID

“NA12878”

ID-indexed table

```
{
  “LDL”: 75.123,
  “ancestry”: “SAS”,
  “cohort”: “1KG”
}
```

Locus-indexed table

```
{
  “gene”: “SHH”,
  “pred_impact”: “high”,
  “pop_frequency”: 0.102
}
```

Genotype

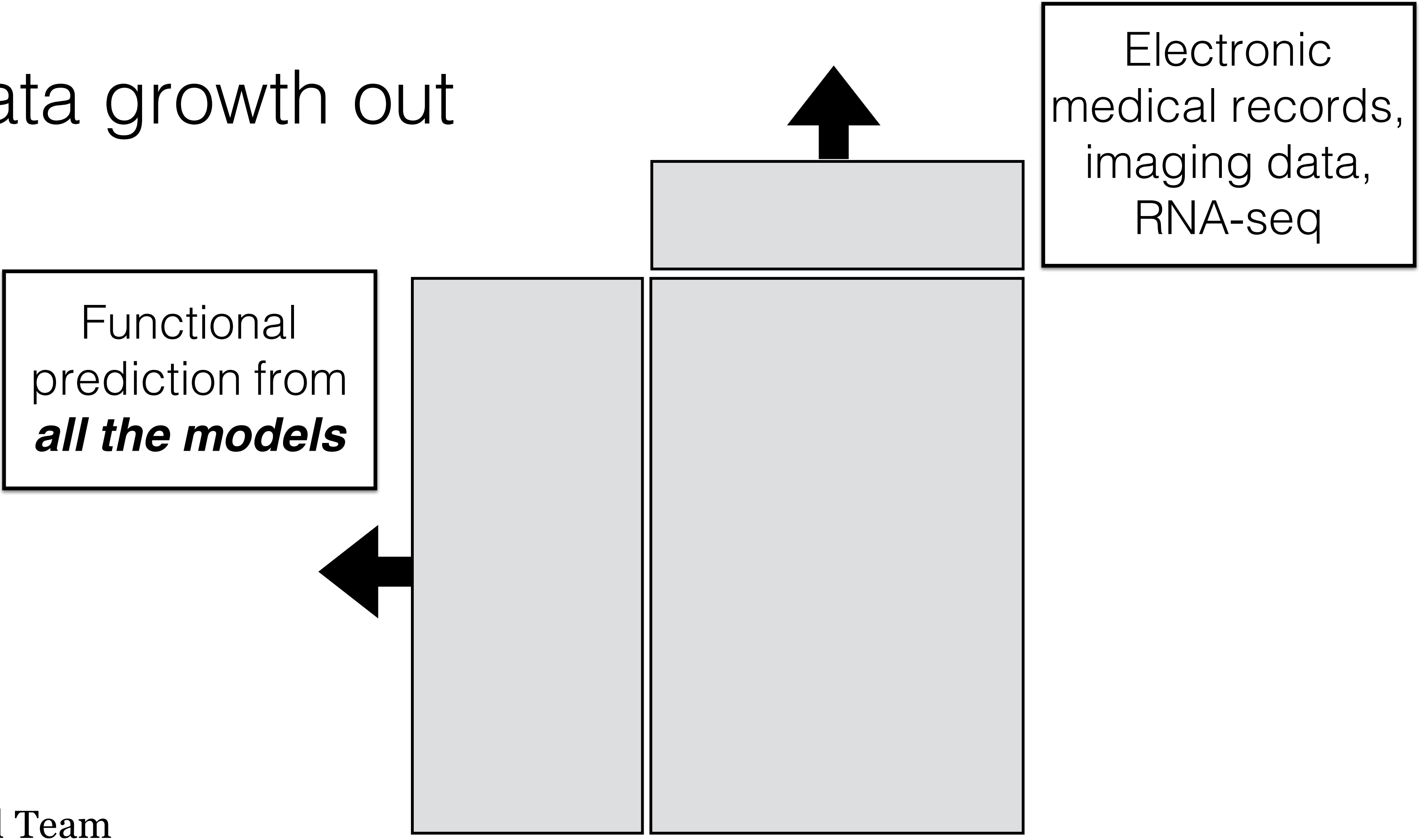
```
{
  “call”: “A/T”,
  “reads”: [10, 8],
  “quality”: 43,
  “p”: [43, 0, 52]
}
```

Genomic Locus

```
{
  “chromosome”: 1,
  “position”: 16123092,
  “reference”: “A”,
 “alternate”: “T”
}
```

Hail Team

Data growth out

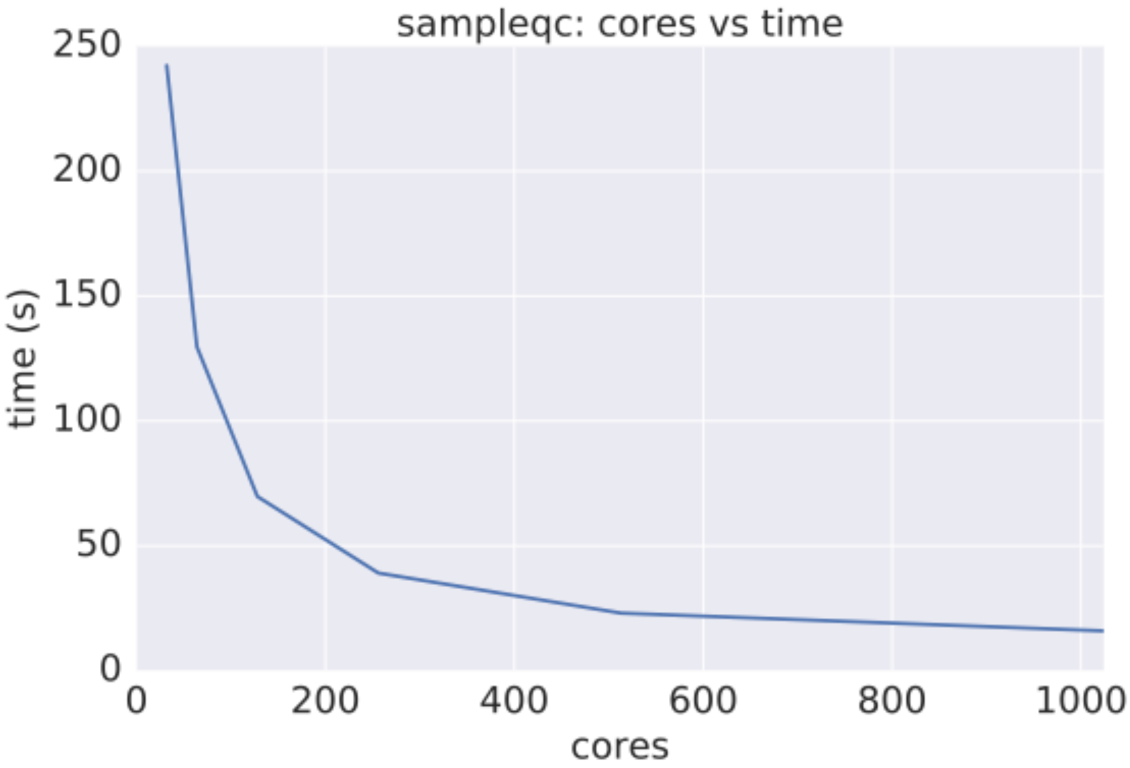


Electronic medical records, imaging data, RNA-seq

Functional prediction from ***all the models***

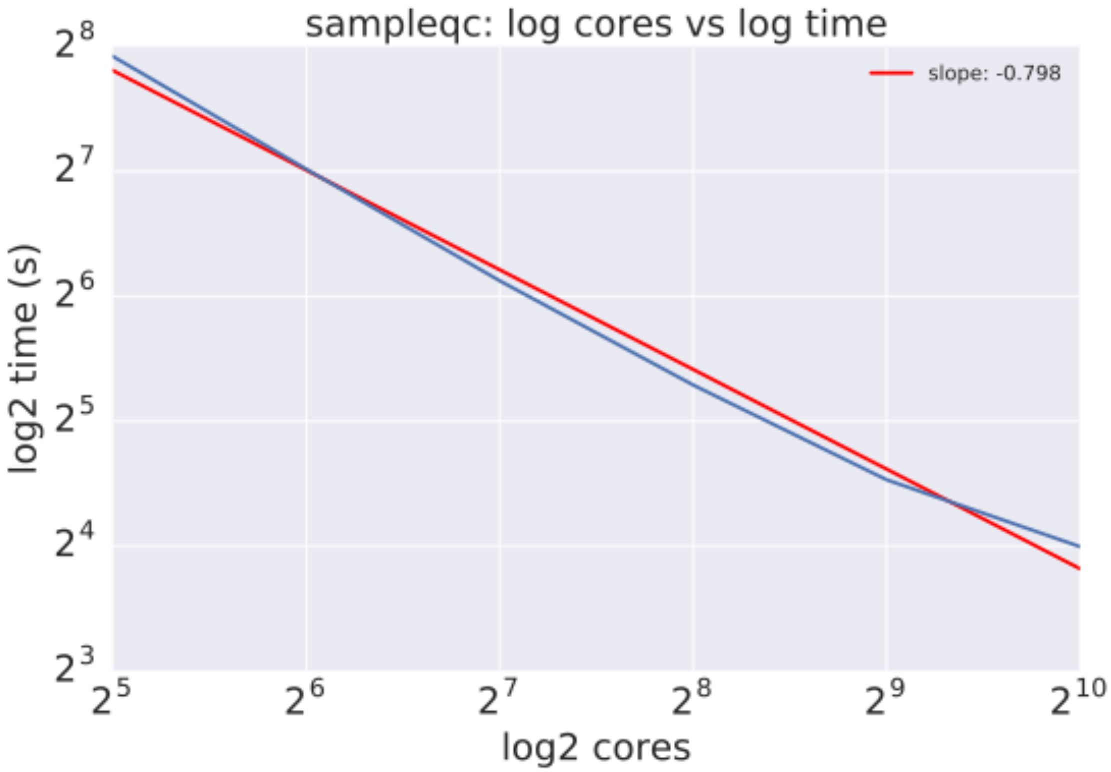
Hail Team

More cores, less time



Hail Team

More cores, less time



Hail Team

Thanks! Any Questions?