Analysis of protein-coding genetic variation in 60,706 humans

Lek et al. *Nature* 2016, 536(7616): 285-291.

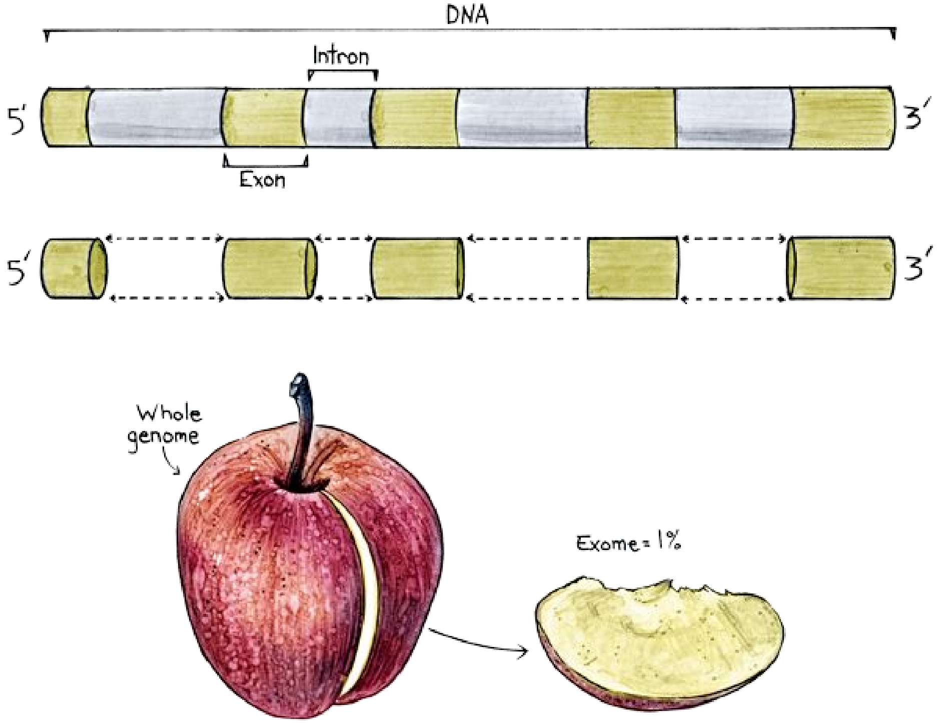
Min-Jung Wang, Mutong Zhao, Jiwon Lee, You Wu and Qianyu Yuan

March 6th

Background-Human exomes

Applications:

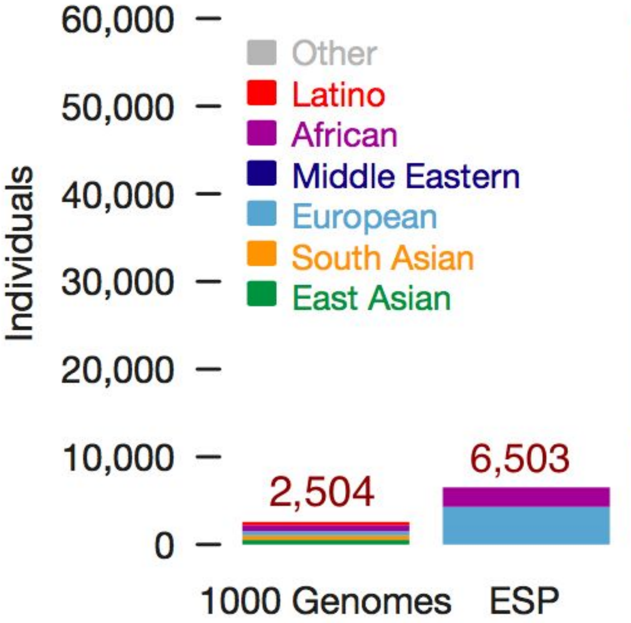
- Understanding of human population history
- Protein function
- Clinical interpretation of mendelian diseases



Background-Previous Datasets vs. ExAC

Limitations of previous datasets
(1000 Genomes Project, Exome Sequencing Project):

- Shallowly sequenced
- Not enough power for identification of protein truncating variation
- Lack of ethnic diversity

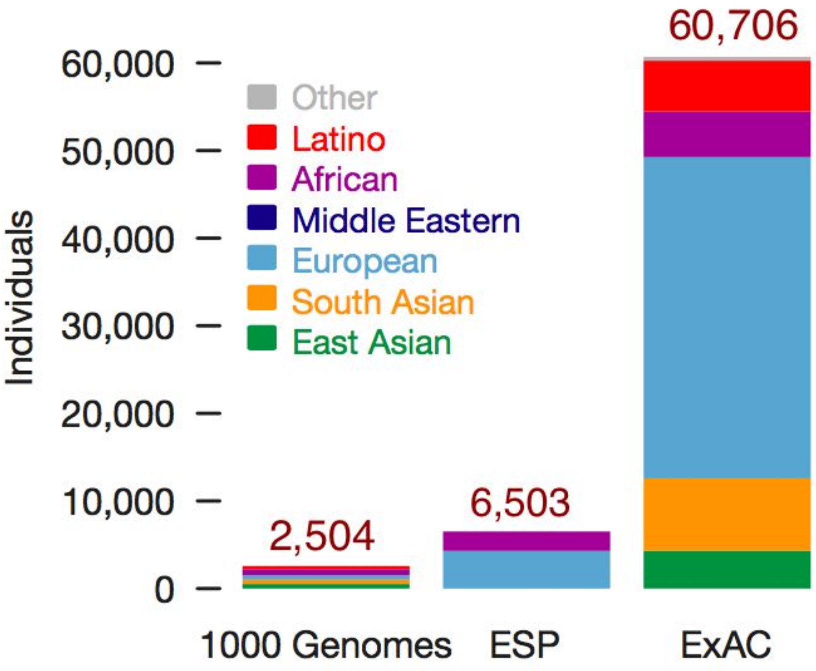


Background-Previous Datasets vs. ExAC

Limitations of previous datasets
(1000 Genomes Project, Exome Sequencing Project):

- Shallowly sequenced
- Not enough power for identification of protein truncating variation
- Lack of ethnic diversity

Exome aggregation consortium (ExAC)



Overview

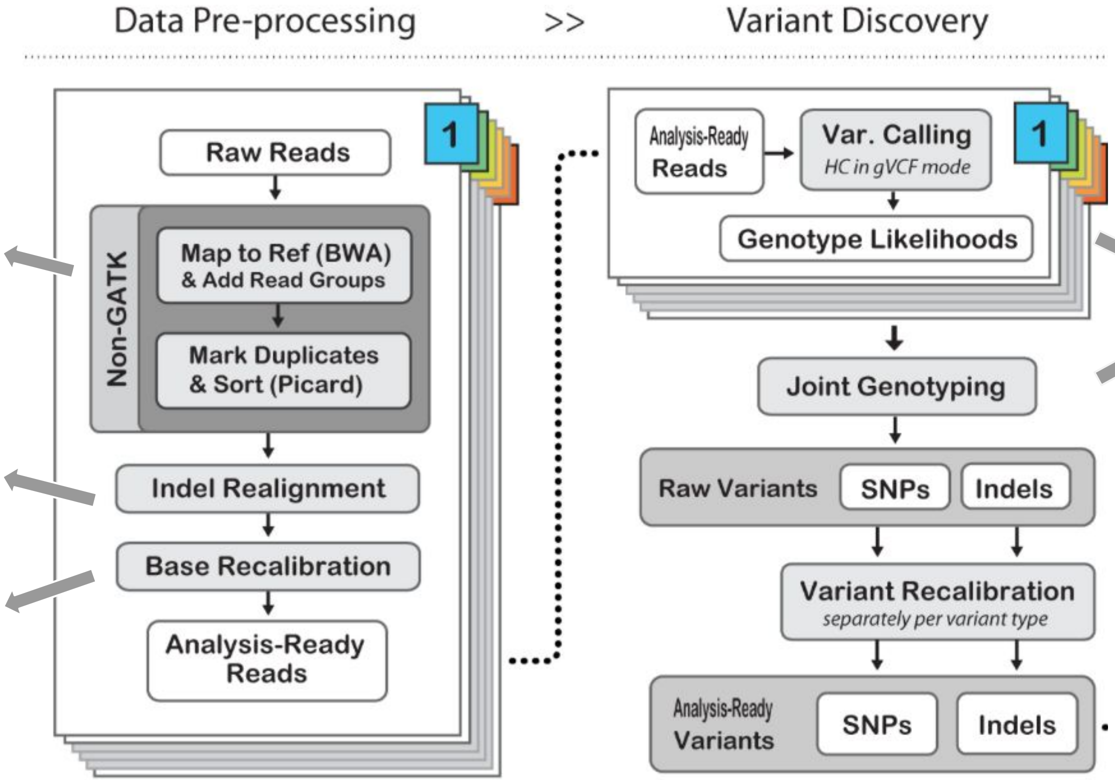
1. Methods
2. Results

   -Mutational recurrence

   -Variant deleteriousness and gene-level constraint

   -Variant interpretation in rare Mendelian diseases
3. Discussion
4. Implication and future direction

Methods



Ref: Human genome reference (hg19)

Clean up artifacts produced in initial mapping step

Yield more accurate base qualities

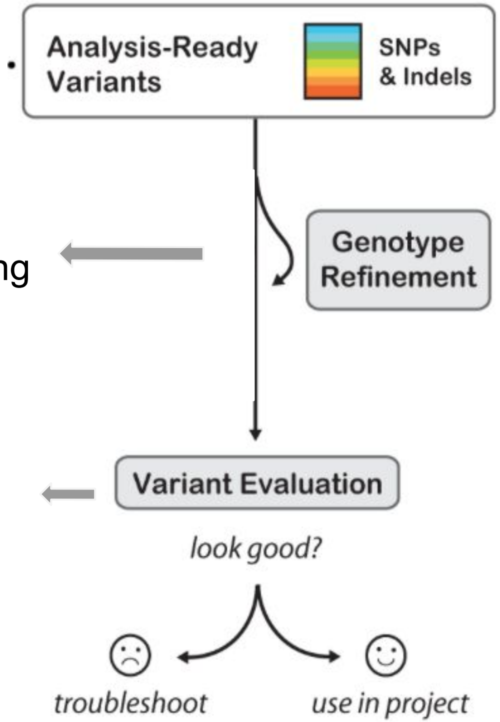
Maximize sensitivity

Maximize specificity/ Variant filtering

https://software.broadinstitute.org/gatk/

Methods

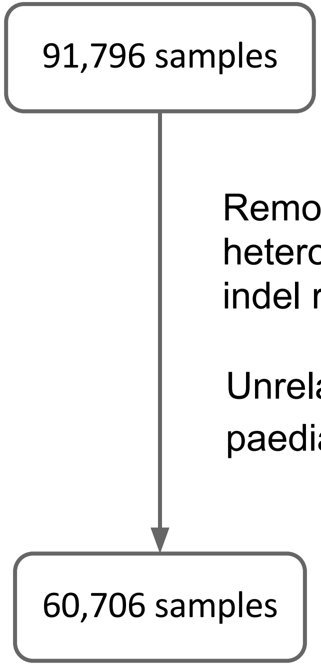
Quality assessment



GQ estimation
Transmission phasing

Compare the callset
to known resources

Sample filtering



Remove outliers (for TiTv, alternate heterozygous/homozygous ratio and indel ratio)

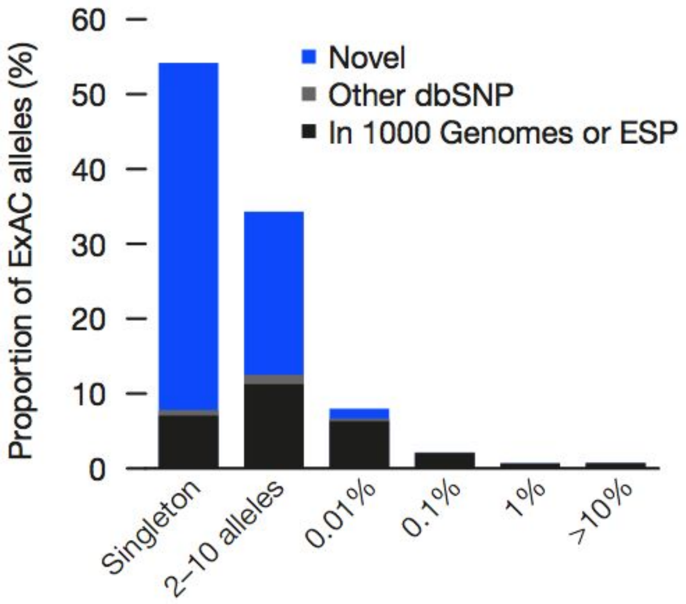
Unrelated adults without severe paediatric disease

https://software.broadinstitute.org/gatk/

Results

- Identified 10,195,872 candidate sequence variants
- After quality control, 7,404,909 high-quality variants left

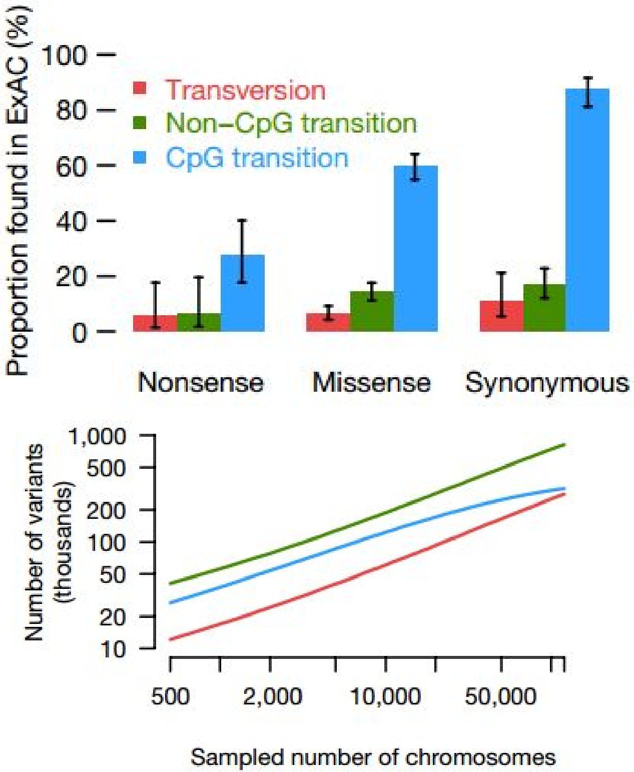
- Majority of genetic variants are rare and novel



Results - Mutational recurrence

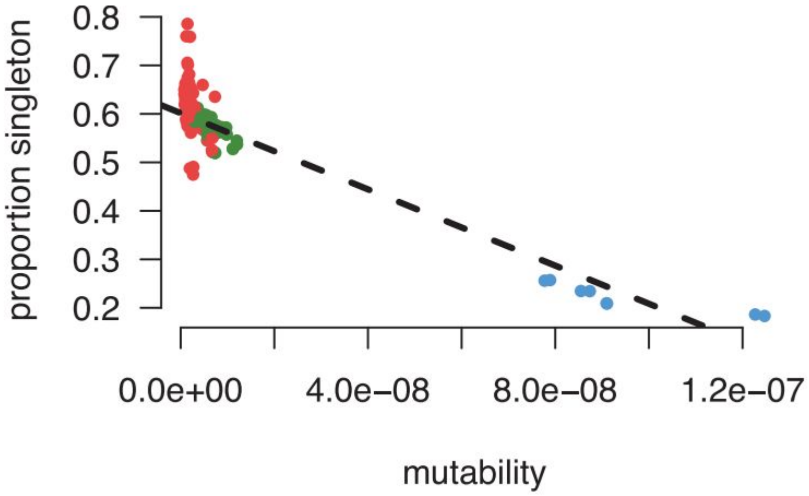
- **Definition:** multiple independent origins for same variant

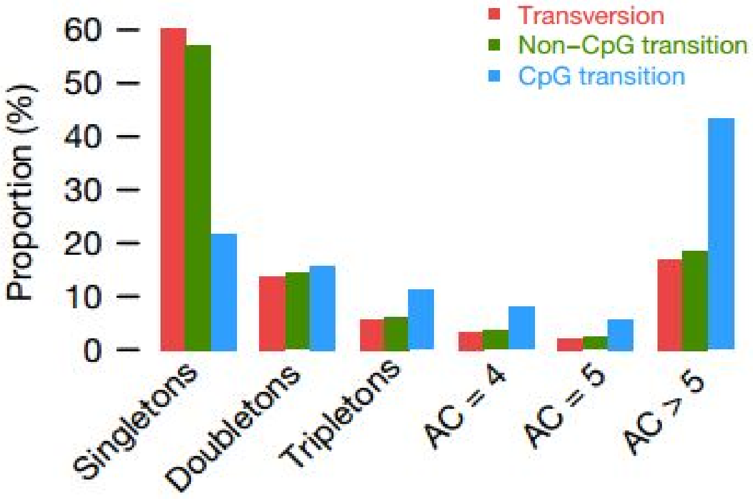
- Compared external data from parent-offspring trios to ExAC:
  - 43% of validated *de novo* synonymous variants found to recur in ExAC
- Number of observed unique CpG transitions:
  - Change in discovery rate from ~n=20,000
  - Large enough dataset, possible identification of all existing variations in a class



Results - Mutational recurrence

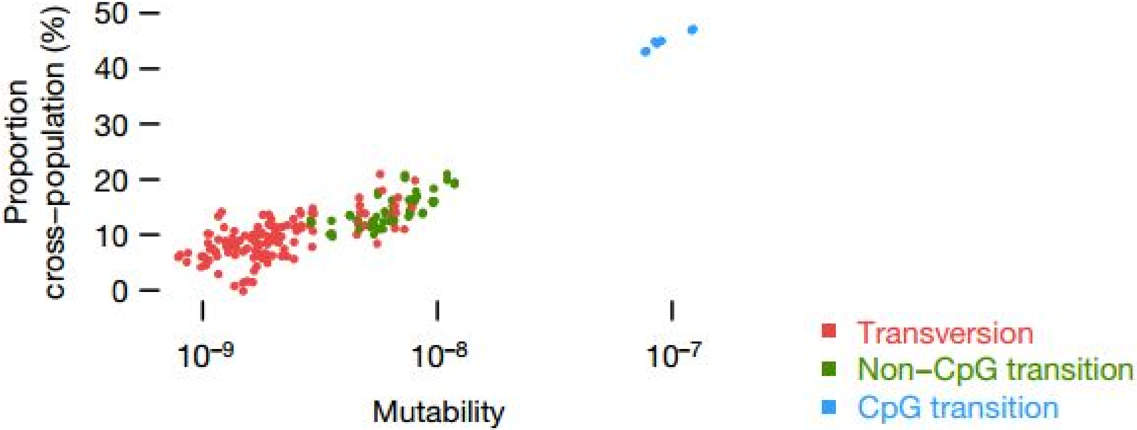
- Mutability: CpG transition >> Non-CpG transition > Transversion
  - Singletons (only one individual has it): more transversions, fewer CpG transitions
  - More mutable variant, more likely to be found more than once in ExAC
  - Higher allele frequencies may have many independent origins





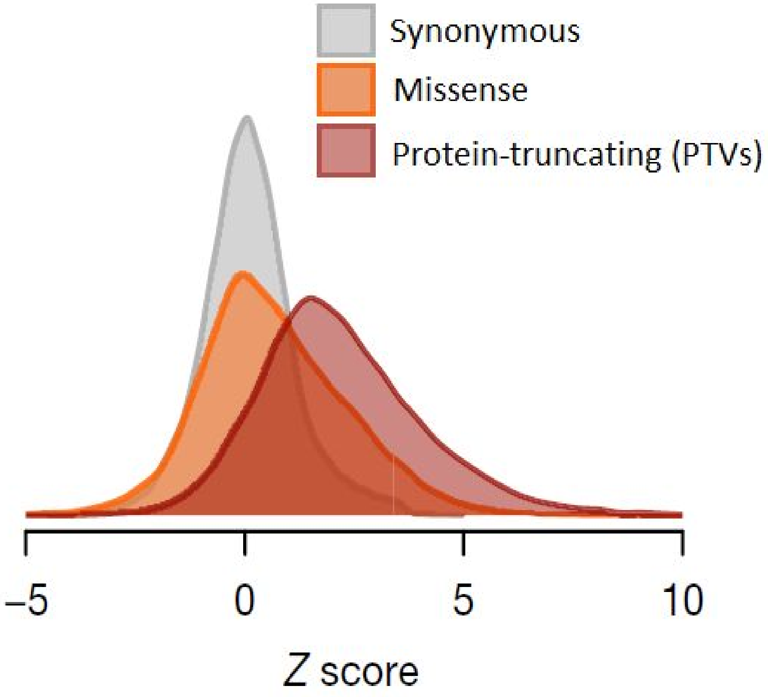
Results - Mutational recurrence

- Doubleton synonymous variants:
  - Independent mutational events
  - + corr. with being from different populations.
- Higher mutability = more cross-population discovery = higher mutational recurrence



Results - Variant deleteriousness and gene constraint

- **ExAC enables deep ascertainment of rare variation**
- **Quantification of gene-constraint to functional variation**
  - **Constraint Z-score** = Quantify extent of selection against functional classes of variation
  - Genes have highest intolerance to deleterious variation (missense, PTV)

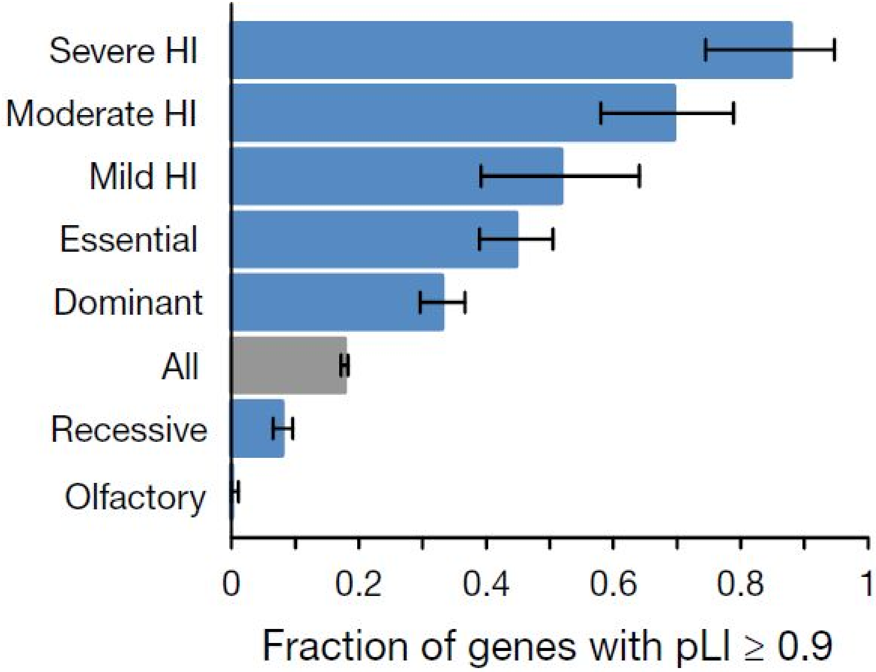


Results - Variant deleteriousness and gene constraint

Probability of being loss-of-function (LoF) intolerant (pLI)

- **pLI >=0.9** indicates high LoF-intolerance

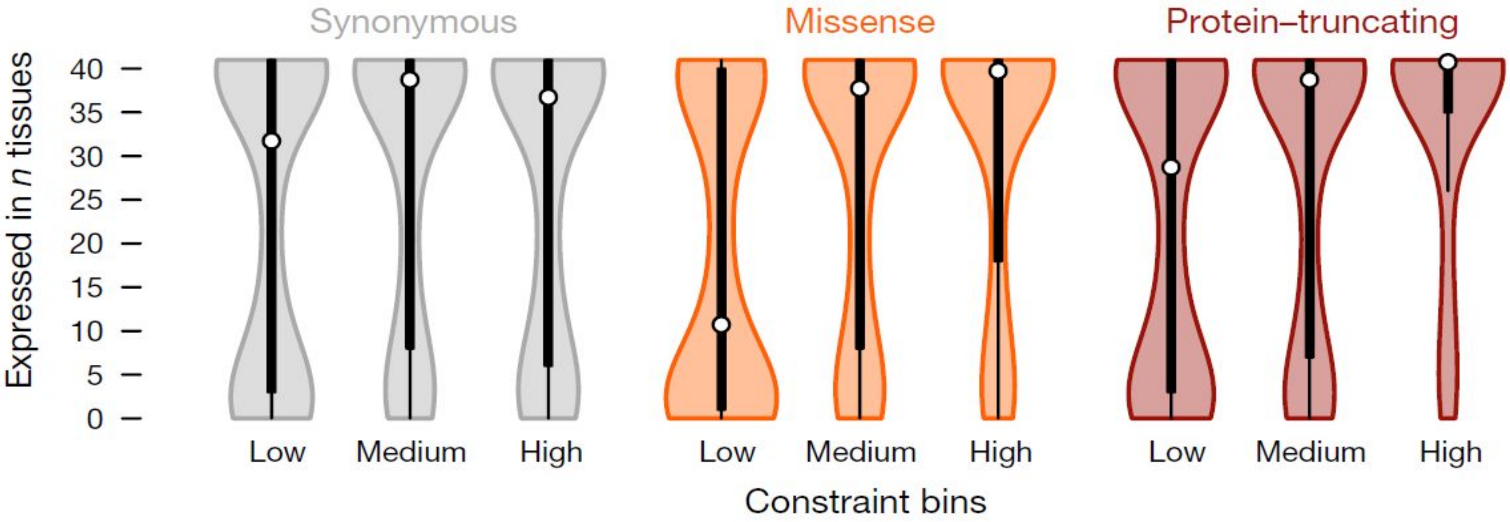
- Essential genes enriched for LoF-intolerant genes
- All known severe haploinsufficient (HI) genes are LoF-intolerant
- 72% of LoF-intolerant genes not implicated in human disease phenotypes



Results - Variant deleteriousness and gene constraint

Tissue expression by gene constraint to functional classes of variation

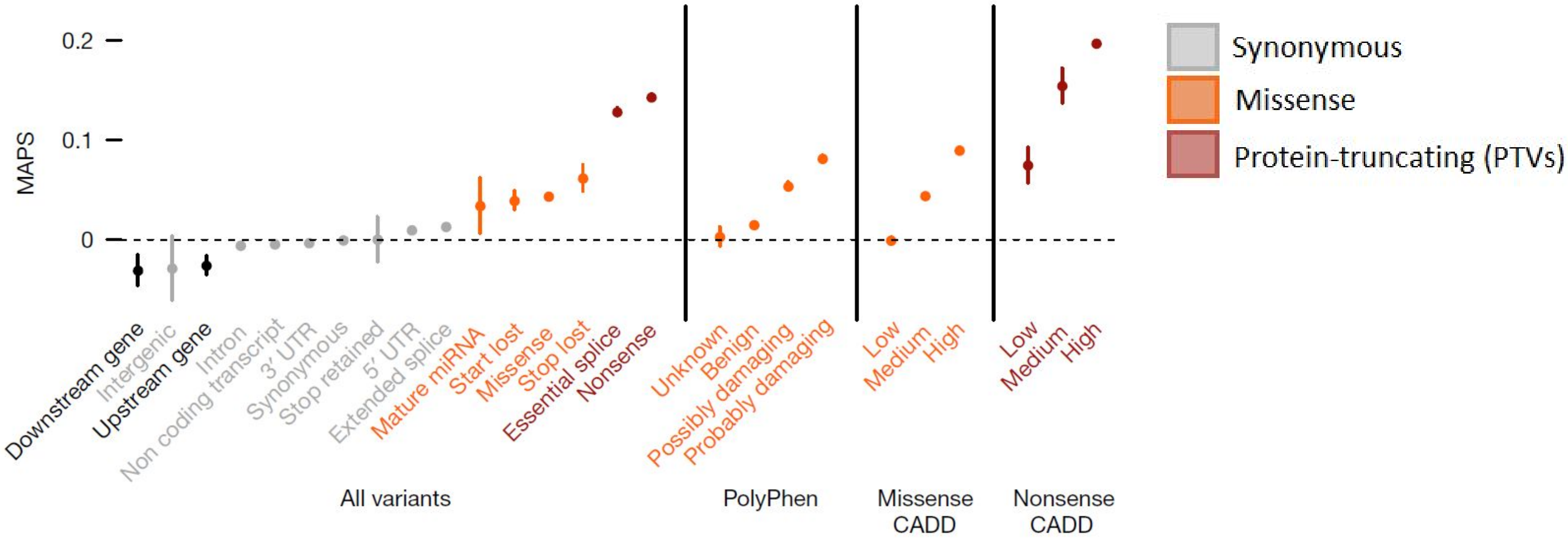
- Synonymous: no sig differences
- Highly missense & PTV-constrained genes tend to be expressed in more tissues



Results - Variant deleteriousness and gene constraint

Functional class of variation on MAPS (variant-level)

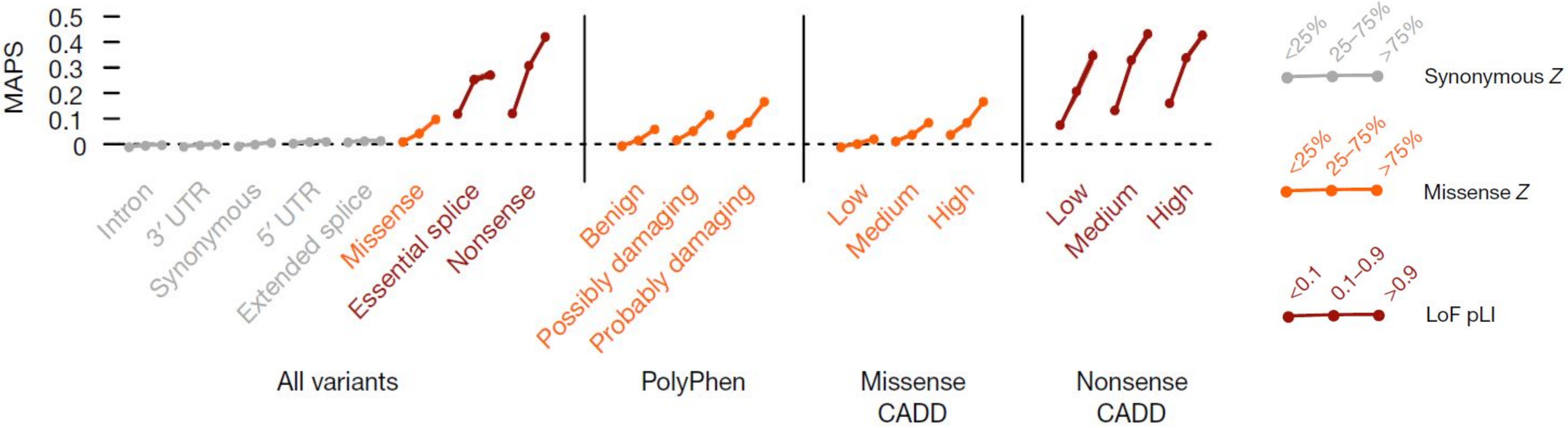
- Mutability-adjusted proportion singletons (MAPS)



Results - Variant deleteriousness and gene constraint

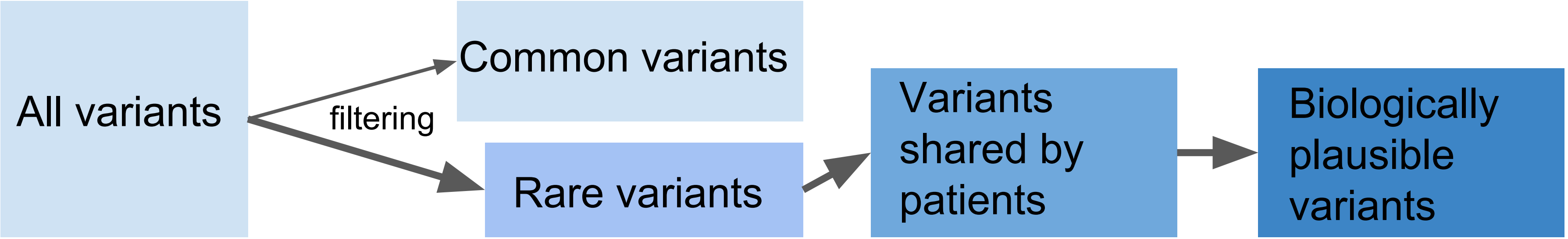
Variant-level by gene-level constraint on MAPS

- Nonsense and missense variants in LOF-intolerant genes are even more likely to be singletons
- Additional information on assessing pathogenicity



Results - Rare variants of rare disease

Identifying candidate variants causing rare Mendelian diseases

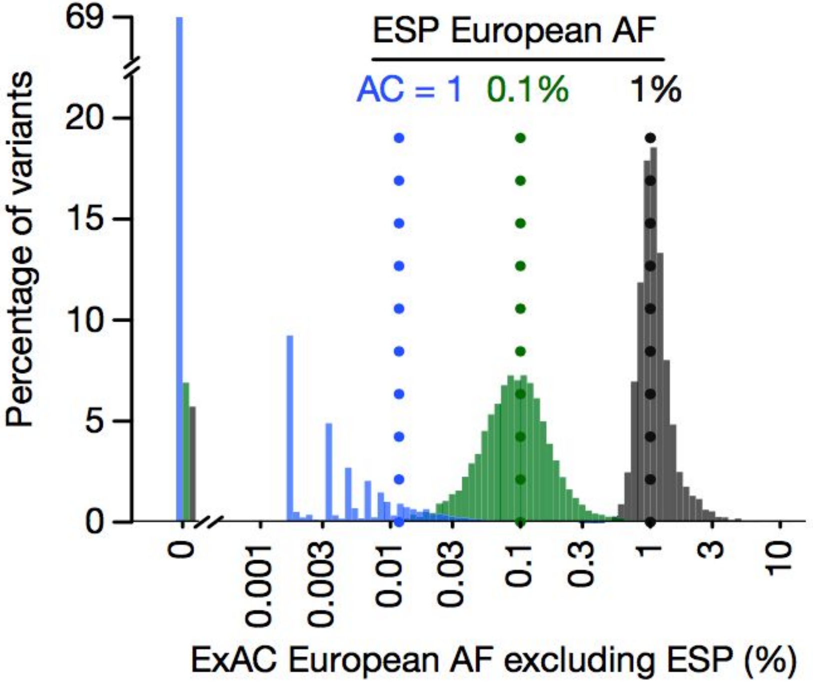


- Previous database was unreliable on very low allele counts
- ExAC has greater power to filter out common variants
- Most registered pathogenic variants had insufficient evidence

Results - Rare variants of rare disease

Previous database was unreliable on low allele frequency estimation

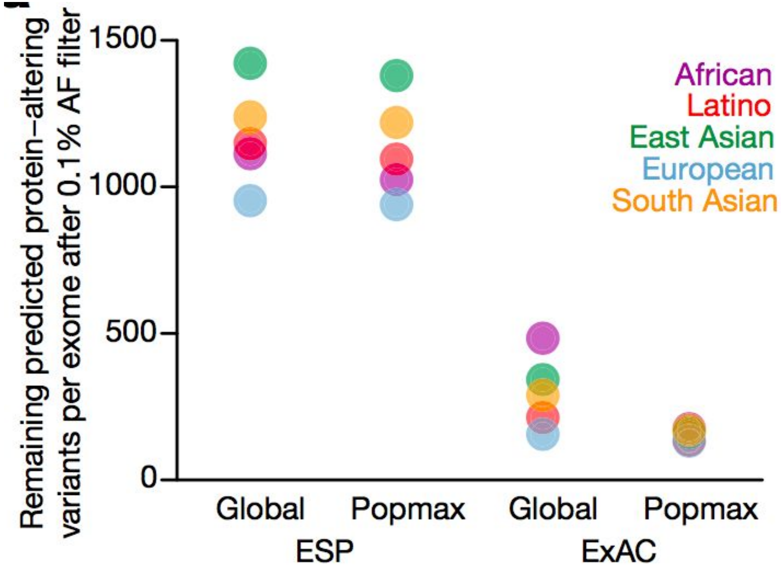
- The frequencies are more reliably estimated around 1%
- The distribution is wider at around 0.1%
- The estimations of even lower frequencies (AC=1) are very unstable



Results - Rare variants of rare disease

ExAC has greater power to remove common variants

- Predicted missense and protein-truncating variants in 500 randomly chosen ExAC individuals
- Filtering threshold: 0.1% allele frequency
- ExAC filter out more common variants than ESP
- Using Popmax AF resulted in fewer candidate variants than Global

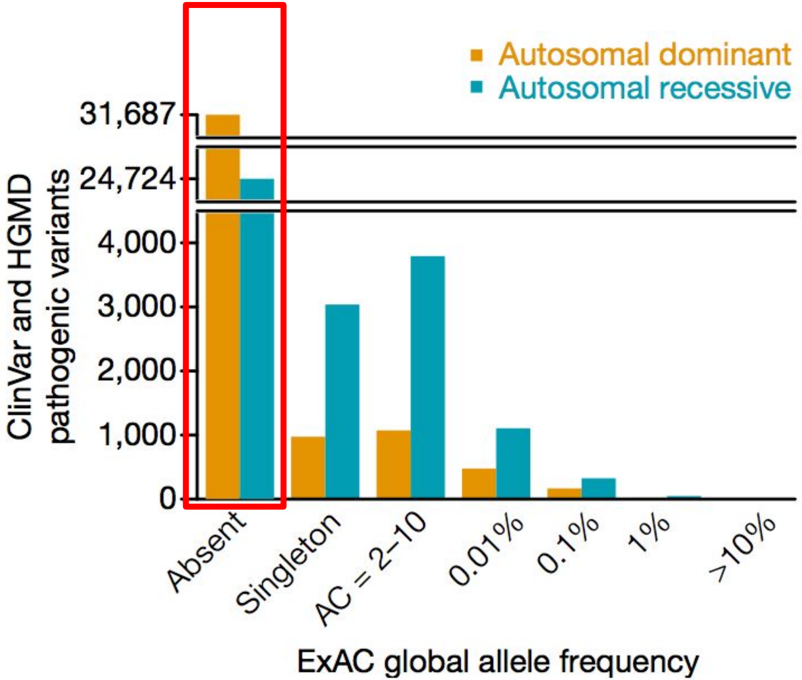


Popmax: the highest allele frequency in any one population

Results - Rare variants of rare disease

Allele frequency of ‘registered’ pathogenic variants

- Most ‘registered’ pathogenic variants in HGMD and Clinvar were absent in ExAC



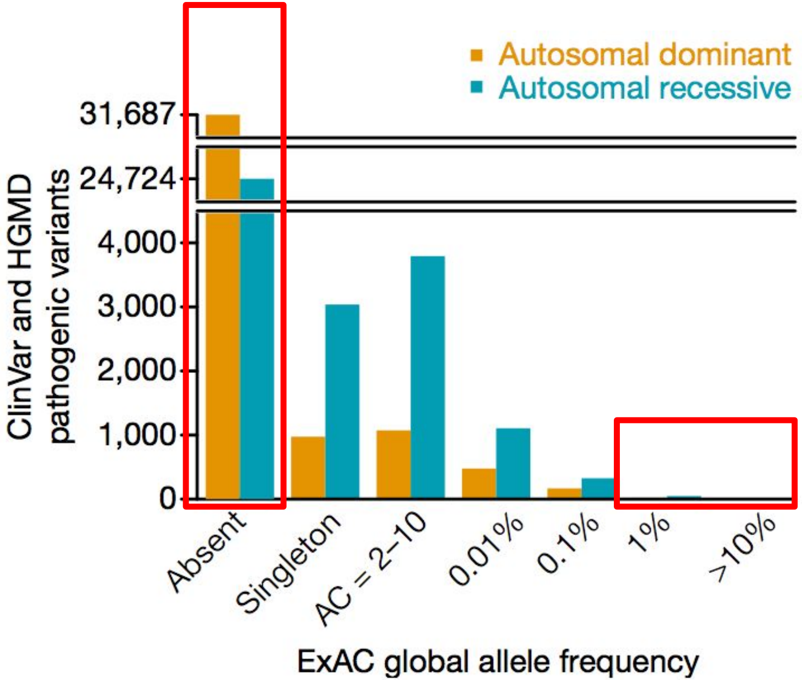
ClinVar https://www.ncbi.nlm.nih.gov/clinvar/

HGMD(Human Gene Mutation Database) http://www.hgmd.cf.ac.uk/ac/index.php

Results - Rare variants of rare disease

Allele frequency of ‘registered’ pathogenic variants

- Most ‘registered’ pathogenic variants in HGMD and Clinvar were absent in ExAC
- Some (~200) are implausibly frequent given the prevalence of rare disease in general population



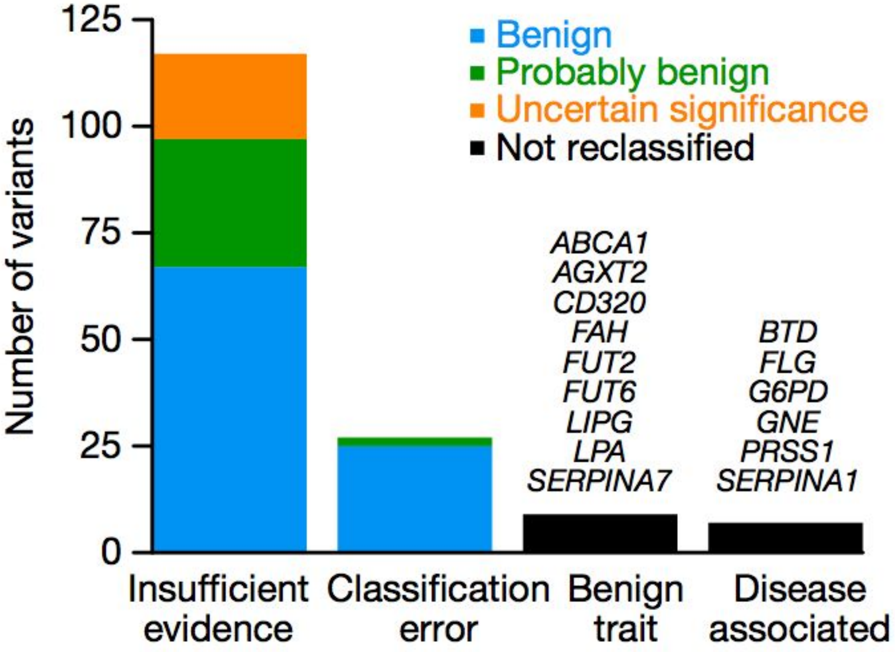
ClinVar https://www.ncbi.nlm.nih.gov/clinvar/

HGMD(Human Gene Mutation Database) http://www.hgmd.cf.ac.uk/ac/index.php

Results - Rare variants of rare disease

The implausibly high frequencies: insufficient evidence for pathogenicity

- Manually curated pathogenic information on the ‘implausibly high frequency’ variants
- Most had insufficient evidence of pathogenicity
- Some were wrongly classified as pathogenic
- Very few were confirmed as pathogenic

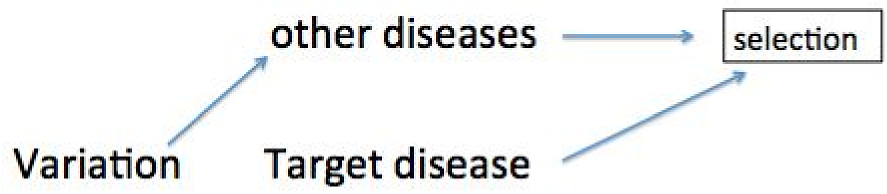


Discussion - Strengths

- Larger data set
- Better resolution
- Investigation of protein truncating variations and LoF-intolerant genes for the first time

Discussion - Limitations

- Control selection- possibly biased



- Middle Eastern and African populations are under-represented
- ExAC only contains data on exomes
- Detailed phenotype data are unavailable

Implications and Future Directions

- High quality and resolution, open-source database
  - Most comprehensive to date
  - Study of rare variants
- Saturation of genetic variation class
- Mendelian-gene discovery
- Reassessment of past studies
  - False positives or true associations?
- Precise diagnosis in rare disease patients

Implications and Future Directions

- Value of aggregating data
- Greater ethnic diversity
- Scale up to whole genome
- More sequenced exomes - next order of magnitude
- Link to phenotype data
  - Translation to biological and clinical understanding

Thank you!

Consortium/Cohort	Samples
1000 Genomes	1,851
Bulgarian Trios	461
GoT2D	2,502
Inflammatory Bowel Disease	1,675
Myocardial Infarction Genetics Consortium	14,622
NHLBI-GO Exome Sequencing Project (ESP)	3,936
National Institute of Mental Health (NIMH) Controls	364
SIGMA-T2D	3,845
Sequencing in Suomi (SISu)	948
Swedish Schizophrenia & Bipolar Studies	12,119
T2D-GENES	8,980
Schizophrenia Trios from Taiwan	1,505
The Cancer Genome Atlas (TCGA)	7,601
Tourette Syndrome Association International Consortium for Genomics (TSAICG)	297
Total	60,706

Supplementary Information Table 2. The Exome Aggregation Consortium (ExAC) sample numbers from each Consortia/cohort.

Population	**Male Samples**	**Female Samples**	**Total**
African/African American (AFR)	1,888	3,315	5,203
Latino (AMR)	2,254	3,535	5,789
East Asian (EAS)	2,016	2,311	4,327
Finnish (FIN)	2,084	1,223	3,307
Non-Finnish European (NFE)	18,740	14,630	33,370
South Asian (SAS)	6,387	1,869	8,256
Other (OTH)	275	179	454
Total	33,644	27,062	60,706

Supplementary Information Table 3. ExAC samples summarized by population and sex.