Human sequence variation and our diseases

International Commission on the Clinical Use of Human Germline Genome Editing
US National Academy of Science
Washington, DC
August 13, 2019

Aravinda Chakravarti, PhD
New York University School of Medicine
aravinda.chakravarti@nyulangone.org

Outline

1. Patterns of human sequence variation
2. Genomics of Mendelian disorders
3. Genomics of multifactorial disorders

Patterns of human sequence variation

Types of sequence variation

SNVs (single nucleotide variants): 88%

AAGTCGATTGACCGAATTAATTAATTGCGGT

AAGTCGATTGATCGAATTAATTAATTGCGGT

1 in 1,000 bases or 3 million differences in a pair of genomes

INDELs (insertions/deletions <50nt), small CNVs (copy number variants): 7%

AAGT- GATTGACCGAATTAATTAATTGCGGT

AAGTCGATTGACCG ------ AATTAATTGCGGT

Inversions, segmental rearrangements: 5%

1) Short read sequencing technology misses many variants;
2) There are many structurally complex loci across the genome.

Sequence variation in 53,831 humans (TOPMed): 410m variants*

*European, African, Asian & Native American ancestry; Hispanic & non-Hispanic ethnicity; ~40% individuals admixed

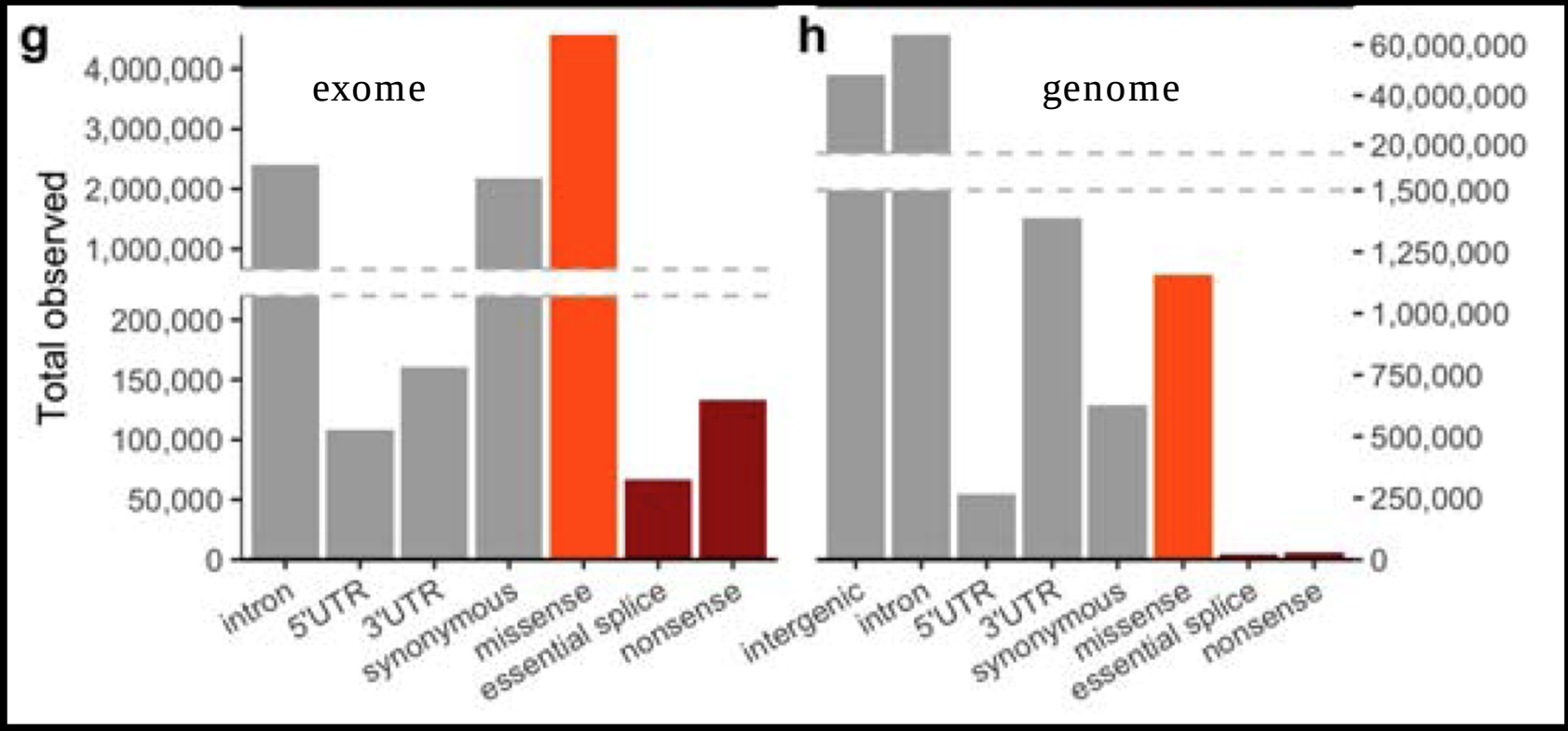
Variant type	Total #	# per individual
single nucleotide variants	381,343,078	3,383,710
insertions/deletions	28,980,753	183,759
coding: synonymous	1,525,971	11,073
non-synonymous	3,172,551	10,875
stop/± 2nt splice	105,042	456
frameshift	113,805	127
in-frame deletions	55,806	99

~50% of variation is unique to individuals

Taliun et al *bioRxiv* 2019

Sequence variation in 141,456 exomes and genomes (gnomAD): 270m variants*

*diverse ethnicities and ancestries; includes admixed individuals



443,769 loss-of-function (protein) variants in 16,694 genes…
genes vary considerably in tolerance to such variants

Karczewski *bioRxiv* 2019

Genomics of Mendelian disorders

The molecular basis of rare and severe disorders

- 5,048 of 7,926 Mendelian phenotypes have known genes (*Mendelian Inheritance in Man*);
- Exome/genome sequencing (22,742 families) revealed 1,430 known genes, 470 phenotypic expansions and 1,767 new genes (*Centers for Mendelian Genomics)*;
- 70% of known Mendelian phenotypes have associated genes and 24% of these are pleiotropic;
- 20% of human genes have an associated Mendelian disorder and 30% likely involve embryonic lethality.

1) ~50% of causal variants can be recognized today;
2) functional analysis of pathogenic alleles in a cellular context is severely lagging;
3) deleterious alleles can be beneficial in some contexts (Fy^0 protection for *P vivax*).

Chong et al *Am J Hum Genet* 2015; Posey et al *Genet in Med* 2019

Clinical genetics of rare and severe disorders

- In clinical genetic testing laboratories, the fractions of autosomal dominant/recessive and X-linked dominant/recessive cases are 53, 31, 3 and 13%, respectively

- ...but, exome/genome sequencing of 6,040 families (proband, parents) with an undiagnosed rare disorder reveals 46% are from *de novo* coding mutations, 3.6% are recessive but rising to 31% in inbred families (*Deciphering Developmental Disorders*).

The 50% 'gap' in undiagnosed disorders is owing to our inability to recognize all coding pathogenic alleles or they are non-coding, have reduced penetrance or are polygenic (synthetic lethals).

Farwell et al *GIM* 2014; Retterer e al *GIM* 2015;Martin et al *Science* 2018

Mutation-Selection balance

- Pathogenic alleles in humans are rare and at equilibrium between their elimination by natural selection and occurrence through *de novo* mutations (DNM, μ);

- autosomal/X-linked recessive/dominant disorders with large fitness reductions have *incidence* proportional to the mutation rate (μ) of the underlying gene (3-fold difference);

- $\mu = 1.29 \times 10^{-8}$ /base/generation in humans but depends on gene sequence, particularly CpG transitions, and increases with paternal (1.47 DNM/yr) and less so with maternal (0.37 DNM/yr) age.

1) $\mu \sim 5 \times 10^{-5}$ /gene (median) and varies 100-fold;
2) μ depends on demographic/reproductive/environmental factors.

Jonsson et al *Nature* 2017

Pathogenic alleles of high frequency

- Some pathogenic alleles have arisen to high frequency (>1%) in some populations;

- the usual cause is genetic drift in geographic, cultural or religious isolates with or without consanguinity;

- less frequently, mutant alleles have heterozygote advantage (fitness advantage over either homozygote, e.g., malarial protection in sickle cell trait);

Genetic testing together with community education and counseling have reduced/eliminated some disorders in some populations (β-thalassemia in Cyprus, Sardinia; Tay-Sachs disease in the Ashkenazim; several recessive disorders in the Orthodox Jewish community: Dor Yeshorim).

Cao & Kan *Cold Spr Harb Pers* 2013

Genetic modifiers of pathogenic alleles

- Evolutionary arguments suggest that alleles at genes that can reduce the pathogenicity of a disease allele or delay its onset will be positively selected;

- genetic 'modifiers' are expected but contrary to model organisms can be heterogeneous or polygenic in humans;

- this is one genetic scenario by which complex inheritance can arise.

1) Pathogenic alleles identified in "affected" families have high penetrance (an ascertainment bias);
2) the same allele found by screening may have lower penetrance ...genetic background effect.

Haldane *J Genet* 1941; Van Driest et al *JAMA* 2016

Implications for families

- Most Mendelian disorders are dominant, usually arise from *de novo* mutations and do not recur in a family except when a parent is a gonadal mosaic (unfortunately recognized through recurrence);

- Recessive disorders are infrequent in outbred but frequent in inbred families and can be recognized through segregating pathogenic variants, as in some dominant families;

- X linked recessives are somewhat intermediate because 1/3 of all affected males are from *de novo* mutations with recurrence from segregating variants and some gonadal mosaics.

1) Many families can benefit from sequence-based diagnosis;
2) Despite the *ClinVar* database assessing variant pathogenicity is challenging...many variants of unknown significance.

Disease prevention

- With increased sequencing in affected families and in the general population, plus education and counseling, many cases of single gene high penetrance disorders can be avoided;

- couples at high-risk today choose prenatal testing, selective termination, adoption, use of sperm/egg or embryo donors, *in vitro* fertilization followed by prenatal genetic testing (IVF/PGT) of embryos to mitigate risk;

- couples with 100% risk in their children are exceptional but are rare: expected frequency of q^4 for a recessive and ~$2q^2$ for a dominant disease (q, mutant allele frequency), respectively, or ~1 per 10^{10} couples...unless q is very high.

Lander et al *Nature* 2019

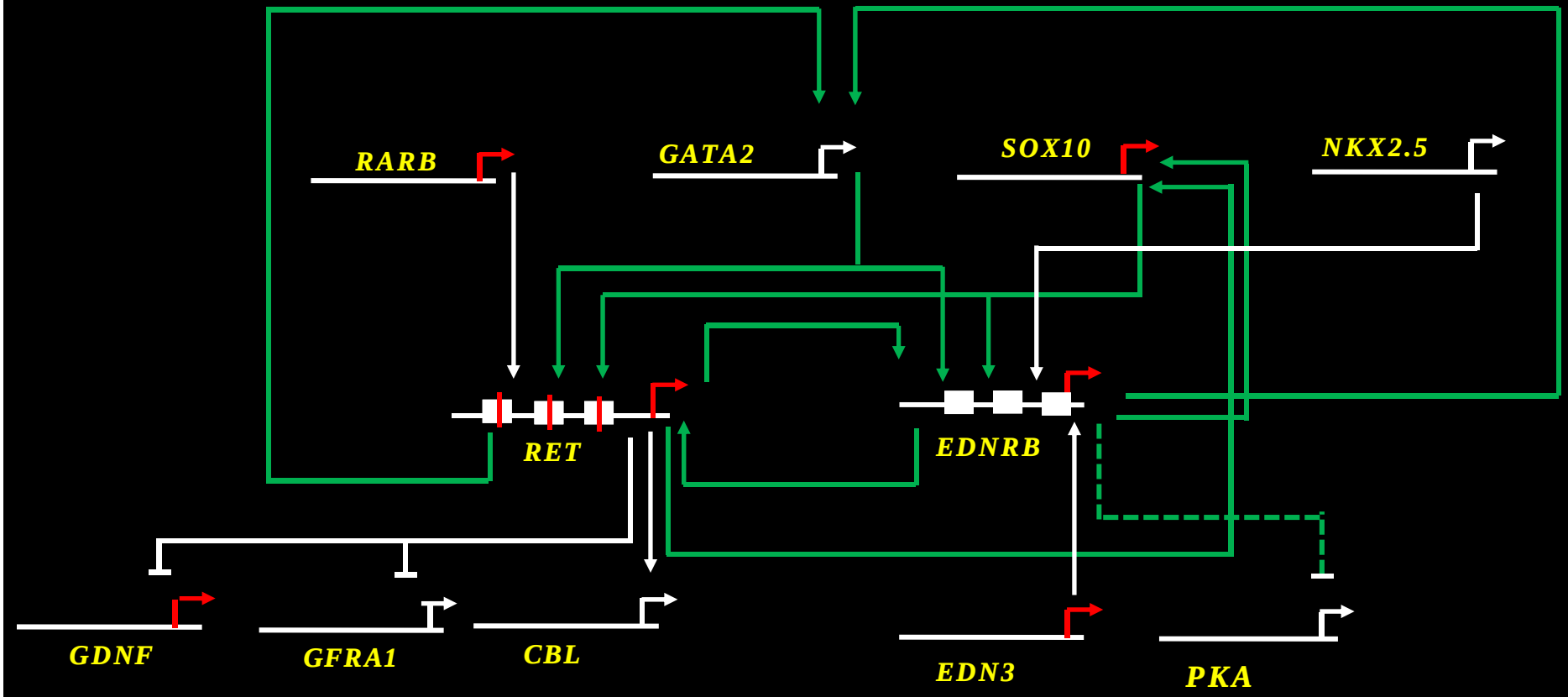
Genomics of multifactorial disorders

Genomic properties of complex phenotypes

- Multigenic inheritance is the rule rather than the exception;
- great progress in mapping complex trait loci (~72,000 replicated associations from >1,000 traits) (GWAS and sequencing phenotypic extremes/diseases);
- additive contributions from 100s to 1000s of common variants with small-to-modest effects allow polygenic risk prediction in those highly susceptible or protected;
- functional bases of these disorders are largely unknown...although regulatory defects are suspected, and demonstrated in some cases, this is an emerging but important area.

Buniello et al *NAR* 2019

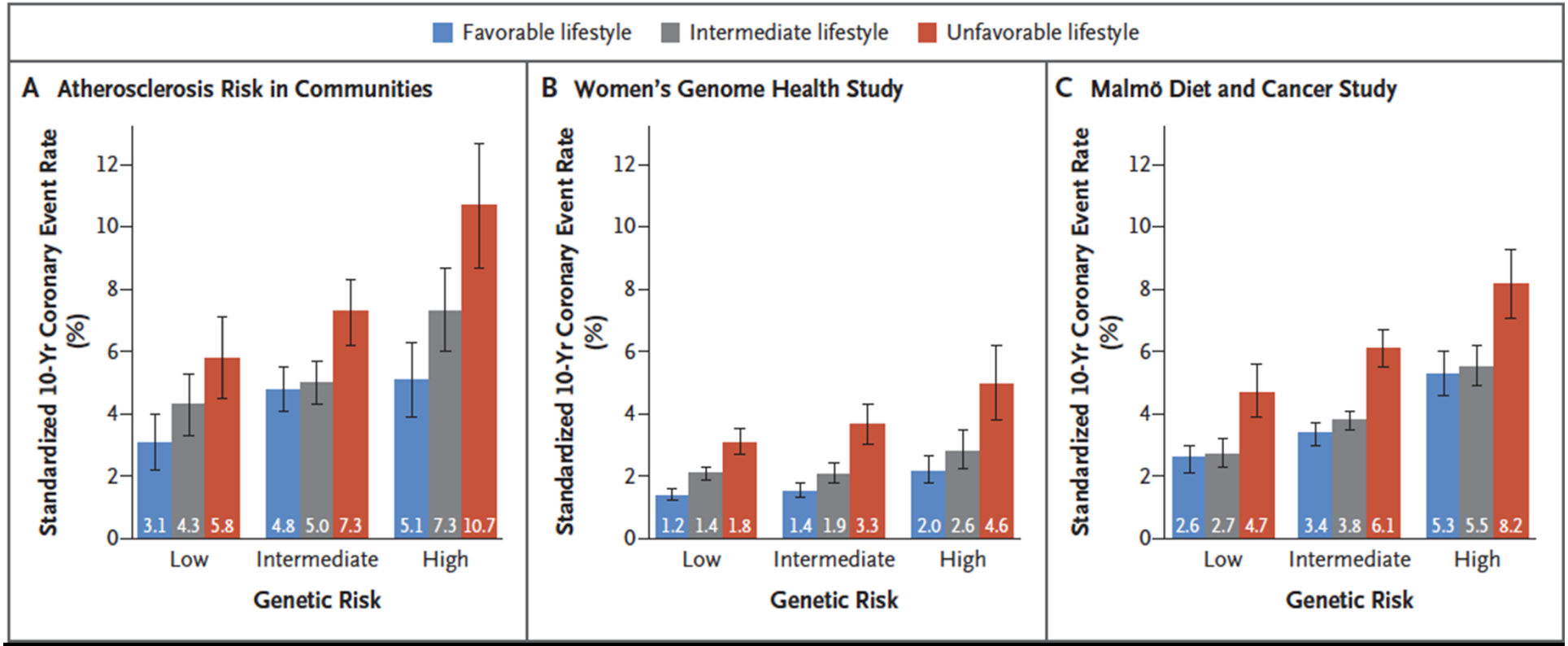
Genes for a complex disease do interact in the relevant cell-type: Hirschsprung disease



HSCR associated mutation/enhancer variant
Genetic interactions

Chatterjee et al *Cell* 2016,
Hum Mol Genet 2019

Gene-environment interactions may be pervasive: coronary artery disease



55,685 subjects across 3 studies; 50 genome-wide significant CAD GWAS SNPs; 4 healthy lifestyle factors from the AHA

Khera et al *NEJM* 2016

Challenges in elucidating complex disorders

- Despite progress, we have not achieved complete genetic, epigenetic and environmental dissection of even one phenotype;

- individual causal variants/genes are neither necessary nor sufficient for trait expression...mechanisms for explaining this feature and the disease are largely absent;

- gene interactions are important but unknown;

- environmental interactions can occur through the same transcriptions factors (thalidomide embryopathy by inactivating SALL4), chromatin regulators (O2 sensing by KDM6A) or enhancers (effects of 2-OGDD dioxygenase on *HIF1*α responsive elements) variant in disease.

Disease prevention

- Variants/genes for complex disorders have largely been used to identify individuals at high risk from single variants or polygenic scores (coronary artery disease, breast cancer, Alzheimer's, etc.);

- despite many genes contributing to high risk, drugs against single proteins (HMG CoA reductase, PCSK9, etc.) can reduce risk because its effect is usually larger than a common variant at the same gene;

- lifestyle changes are also necessary to reduce risk;

- gene editing in a complex disease is unpredictable because of the multiplicity of factors but likely to have a minor effect because the allelic effect is small.

Thank you!!