The American Journal of Human Genetics, Volume *105*

Supplemental Data

Mendelian Gene Discovery: Fast and Furious with No End in Sight

Michael J. Bamshad, Deborah A. Nickerson, and Jessica X. Chong

Supplemental Materials and Methods

Table of Contents

Supplemental Figures

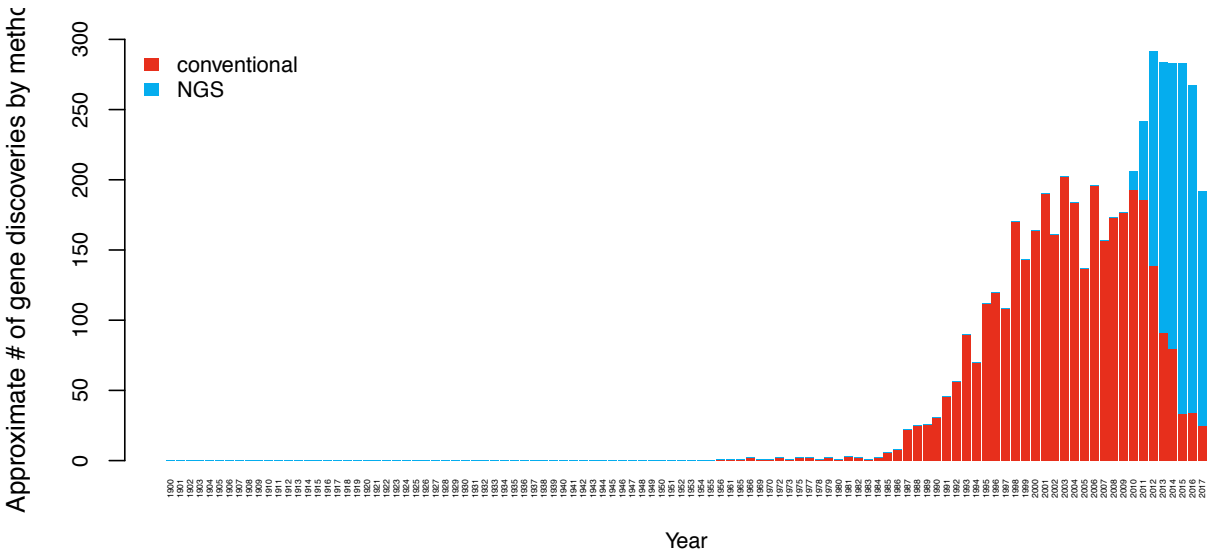

Figure S1. Estimated number of gene discoveries per year since 1900.
From 1900 to 1986, a handful of new MCs were characterized each year, and even fewer underlying genes were discovered. Beginning with the introduction of positional cloning in 1986, gene discovery for MCs accelerated greatly.

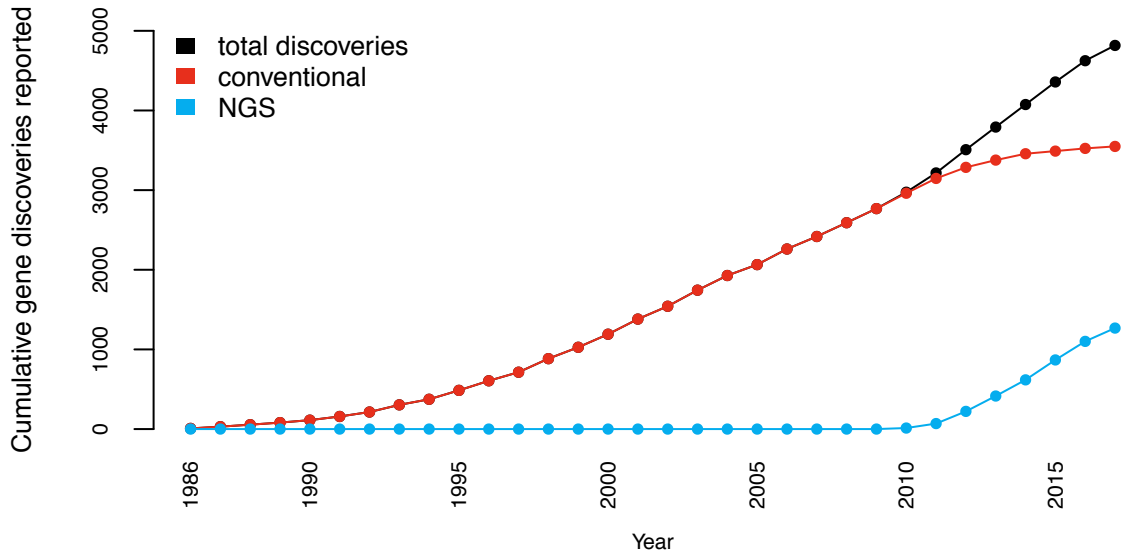

Figure S2. Cumulative estimated number of gene discoveries per year since 1986.
NGS-based approaches (primarily ES) have led to ~36% (1,268 / 3,549) of all reported Mendelian gene discoveries.

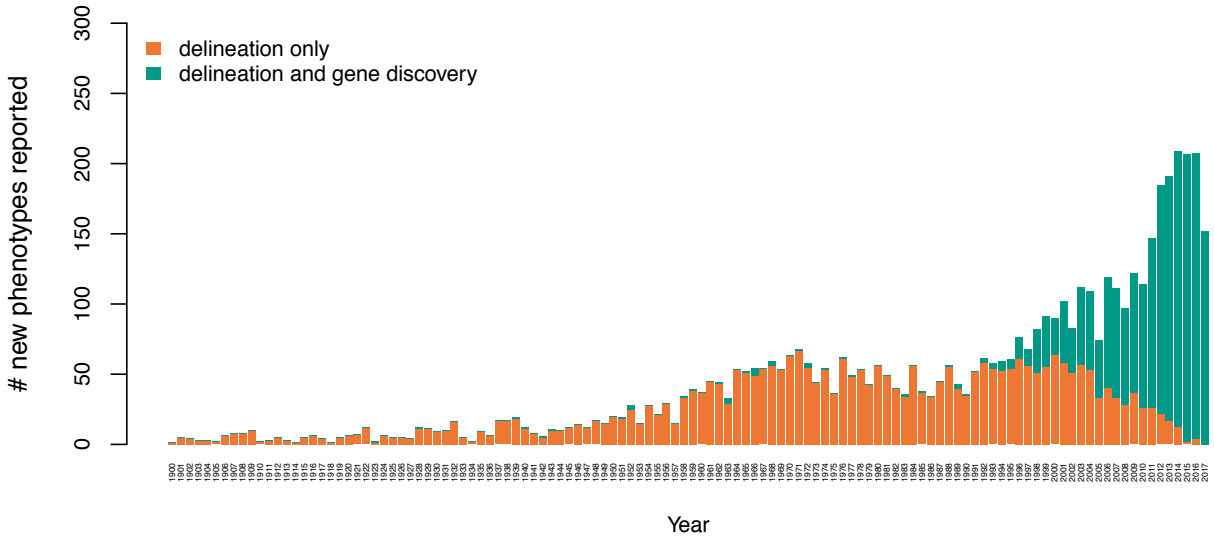

Figure S3. Estimated number of delineated syndromes per year since 1900.
Historically, particularly prior to the introduction of positional cloning in 1986, all or nearly all syndrome delineations were phenotype-driven. Classical syndrome delineation (orange) is phenotype-driven and proceeds by identifying multiple individuals with overlapping phenotypes, and then discovering the underlying gene. In contrast, in genotype-driven syndrome delineation (teal), the underlying (candidate) gene is discovered in an individual with a new phenotype, then additional individuals with overlapping phenotype are identified on the basis of the shared gene.

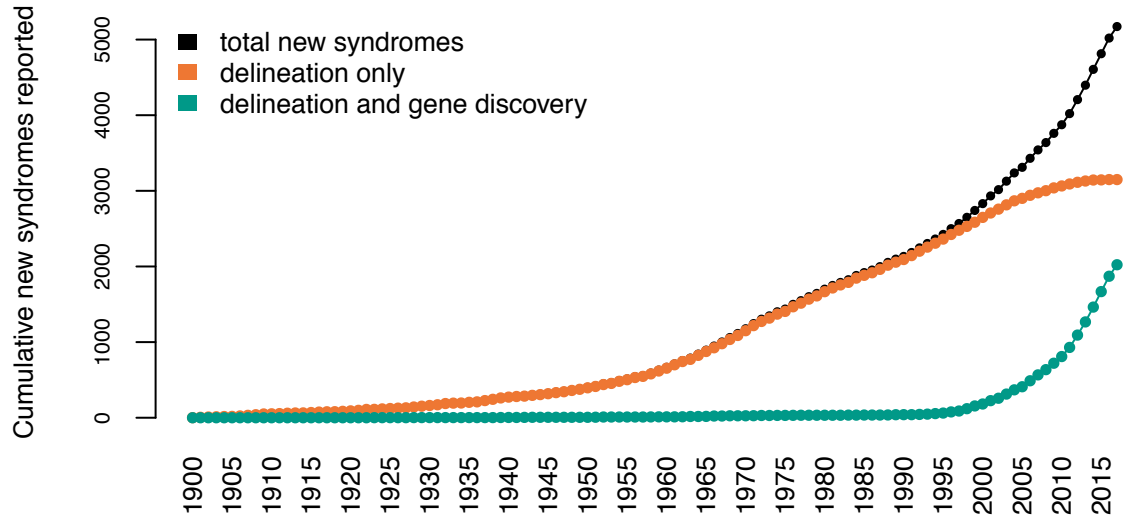

Figure S4. Cumulative estimated number of delineated syndromes per year since 1900.
In total, genotype-driven syndrome delineation has led to the description of 2,023 MCs vs. 3,149 MCs described via phenotype driven delineation. Ultimately most MCs will be ascertained via genotype-driven delineation.

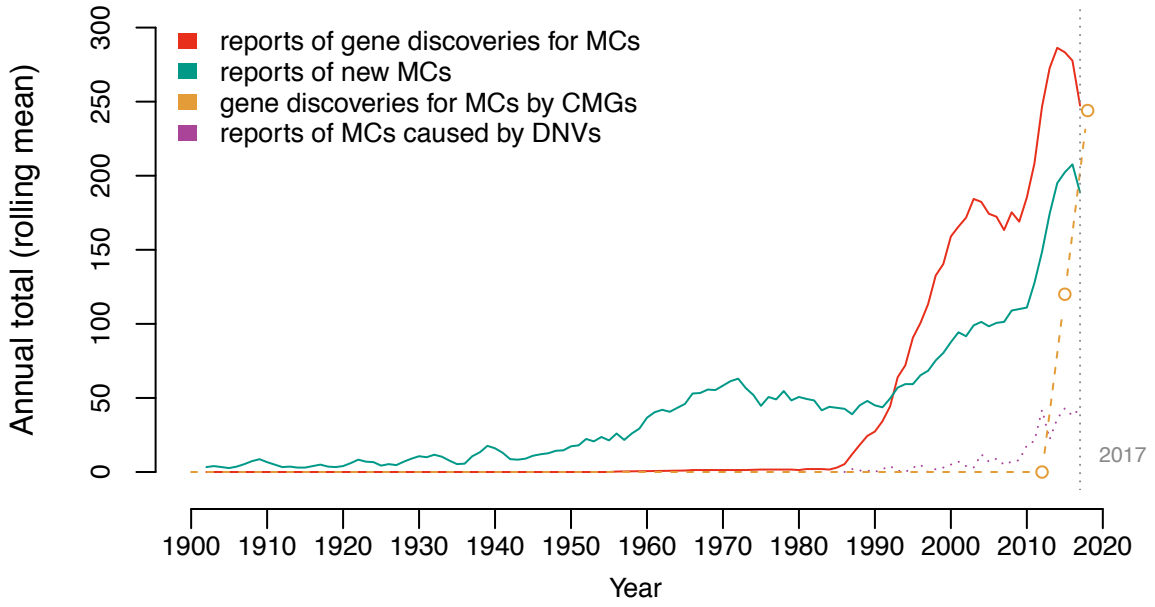

Figure S5. Approximate rates of reported gene discoveries for Mendelian conditions, delineation of Mendelian conditions, gene discoveries caused primarily by de novo variants, and unpublished discoveries by the Centers for Mendelian Genomics over time (1900-2017).

This graph illustrates trends in reported (i.e., published) delineations of new MCs, including the so-called “Golden Age” of syndrome delineation in the 1970s, leading to a peak throughout that decade. It also shows the impact of technical and methodological advances that fueled gene discovery, namely the impact of positional cloning in 1986, development of dense, genome-wide linkage maps in the early 1990s, and increasing knowledge via the Human Genome Project (1990-2001) of the physical location and sequence content of genes. The latter two made it far easier to locate and sequence candidate genes of interest, which facilitated genotype-driven syndrome delineation even prior to the introduction of ES-based approaches. Linkage maps and sequencing the human genome, made it possible to more efficiently identify and sequence candidate genes from the same pathway/gene family as a known gene in a cohort of affected individuals not explained by the known gene. The introduction of NGS completed the shift to genotype-driven delineation. The pace of discovery of MCs that are seemingly caused mostly/entirely by de novo variants took off after microarrays and then NGS made large-scale detection of DNVs possible, nevertheless, these MCs account for only a minor fraction of all discoveries each year.

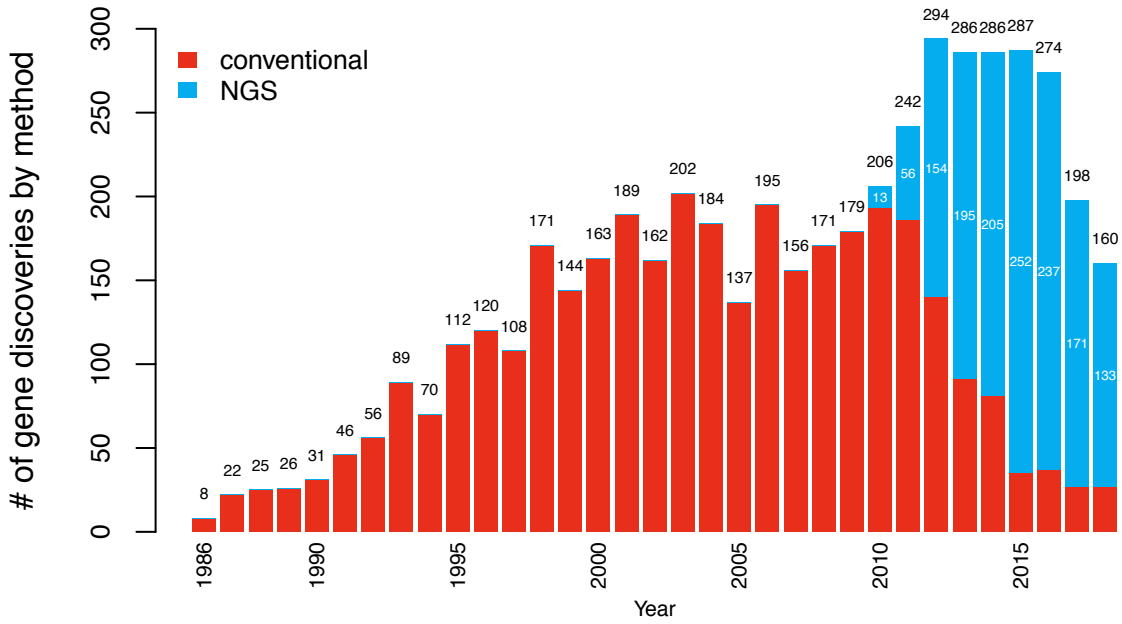

Figure S6. Approximate number of gene discoveries per year for MCs made by ES/NGS versus conventional approaches (including data through the end of 2018).
This graph is identical to Figure 1B except that it includes reports of gene discoveries through the end of 2018 (as cataloged in OMIM as of May 16, 2019). OMIM is still curating the literature for gene discoveries published in 2018 so a small incremental increase in the 2018 totals is expected as OMIM's curation efforts lag by roughly ~6 months (personal communication, A. Hamosh).

Methods

Analyses based on OMIM data

All analyses based on OMIM are limited to the text and data recorded in the database's phenotype and gene entries as of February 15, 2019, with the exception of Figure S6 [data downloaded May 16, 2019]. Therefore, estimated rates of gene discovery should be interpreted as a reflection of the rate at which OMIM curates publications of gene discoveries. OMIM's curation is a manual, human-driven process and thus not able to identify all newly-published gene discoveries within a fixed length of time post-publication. Furthermore, no mechanism yet exists through which one can directly measure the rates of unpublished discoveries being made (e.g. manuscripts in preparation, matches made via MatchMaker Exchange or other matchmaking efforts, or discoveries within a single research group). In order to adjust for lag time in curation of published gene discoveries by OMIM, the entries assessed in all analyses were limited to those with estimated year of discovery or delineation of 2017 or prior with the exception of Figure 2 (estimation of number of undiscovered Mendelian genes), which used all entries as of the date of download.

As shown by a recent analysis[1], older legacy entries in OMIM are enriched for MCs that are not well-established or supported; nevertheless, some legacy entries still appear to describe novel, unexplained conditions, so we continue to include all such entries in these analyses.

Search phrases and patterns were determined after reviewing >50 sample OMIM entries for common word/phrase usage. Code for analysis and generating figures is available at https://github.com/jxchong/mendelian_commentary.

Inferring the year of gene discovery

Estimated year and method (next-generation sequencing/exome sequencing/genome sequencing vs. conventional methods) of gene discovery were extracted from OMIM as previously described[2].

Inferring the year of and approach to syndrome delineation

The estimated year and approach to (genotype-driven vs. phenotype-driven) of syndrome delineation were extracted from text analysis of OMIM as follows. For each OMIM phenotype entry in the downloadable file "OMIM.txt.gz", the earliest year listed in the "Clinical Features" section was obtained by searching for in-text citations that matched patterns such as "(19XX)", "(2XXX)", "In 19XX, McKusick et al.", "McKusick et al. described in 2XXX", etc. We assumed that the earliest year detected in the Clinical Features section would correspond to either the earliest case report of an individual with the associated MC or the actual publication of the syndrome delineation. If the estimated year of gene discovery was greater than the estimated year of syndrome delineation, we classified the delineation as phenotype-driven; if the estimated year of gene discovery was equal or prior to the year of syndrome delineation, we classified the delineation as genotype-driven.

Inferring the mode of inheritance of MCs

The mode of inheritance for each MC was obtained from multiple places in the OMIM database as not all entries have an official mode of inheritance listed (in Clinical Synopsis:

Inheritance and/or genemap2.txt's phenotype name). The "Phenotypes" column in the "genemap2.txt" downloadable file was searched for case-insensitive matches to "autosomal dominant" (AD), "autosomal recessive" (AR), and "X-linked" and each match was recorded as a mode of inheritance for the corresponding MIM phenotype entry. These data were combined with modes of inheritance listed in the "Clinical Synopsis: Inheritance section of "OMIM.txt.gz." Additionally, the "Clinical Synopsis: Miscellaneous" section was searched for the presence of the phrase "de novo."

We used additional criteria to narrow down modes of inheritance because some entries lack a designation in Clinical Synopsis or genemap2 Phenotypes column. We designated phenotypes as likely to be inherited if the genemap2 phenotype name, Clinical Features, Mapping, or Molecular Genetics sections of the entry contained one of the following phrases: "x-generation," but excluding the phrase "next-generation sequencing" (e.g. "3-generation pedigree" or "across four generations"), "linkage analysis", "linkage mapping", "lod score", "lod (" [e.g., lod (3.17)], "point linkage" (e.g. 2-point or multipoint linkage), or "linkage study". We assumed that any autosomal dominant and X-linked entries that mentioned multi-generation pedigrees and/or linkage analysis were likely describing a MC that is at least somewhat frequently inherited by an affected child from an affected parent.

We designated phenotype entries as likely to be de novo if the entry contained the phrase "de novo" in a number of different sections of the OMIM entry (TEXT [introductory summary], Molecular Genetics, Clinical Synopsis: Inheritance, or Clinical Synopsis: Miscellaneous); the entry was also listed as autosomal dominant or X-linked (consistent with an MC that could be caused by de novos in many/most affected individuals); and the phenotype was not categorized as likely to be inherited. This enabled us to count conditions that are likely caused by de novo variants and are likely not compatible (i.e. phenotype too severe) with being transmitted from an affected parent to affected child.

Not all Mendelian gene discoveries have been cataloged in OMIM – in particular, genes that were discovered via statistical enrichment/association studies, were published with little or no phenotypic details, and no follow-up papers with more detailed phenotype data have been published (i.e., the resulting syndrome has yet to be delineated) are typically not included. Because most Mendelian gene discoveries discovered via contemporary enrichment studies are likely to be de novo, we attempted to assess the proportion of such discoveries that are likely to be unrepresented in OMIM. In 2017, the DDD study published 14 genes that achieved genome-wide significance in their de novo enrichment analysis that they considered to not have been previously associated with developmental disorders (DD) with compelling evidence. Of the 14 genes, nine had entries in OMIM (64%) and were successfully flagged as de novo according to our criteria, while five (~36%) did not have an OMIM phenotype entry for a DD (*GNAI1, CNOT3, MSL3, KCNQ3* (in OMIM but not with a DD phenotype), *TCF20)*. If we use this to crudely approximate the number of MCs typically caused by de novo variants, that are not listed in OMIM, and were discovered via statistical analyses of ES/GS/NGS in a large cohort study, then potentially a total of 565 de novo entries might exist (292 existing de novo entries discovered 2010-2017/0.64 + 109 entries discovered prior to 2010). This is probably a gross overestimate, however, as currently, the vast majority of MCs caused by de novo variants are not identified solely by large cohort studies that only report limited phenotypic data (i.e. most such discoveries are also delineated in detail in a separate publication), and most large-scale de novo enrichment-based studies to date each identified a limited number of statistically significant novel Mendelian genes. Thus we feel confident that as of when these analyses were conducted, most MCs discovered and delineated in a traditional gene discovery publication would be included in OMIM.

Even if this higher estimate is correct, the % of phenotypes caused by de novo variants would only be ~12% overall and up to ~19% of all discoveries between 2010-2018.

Estimated number of remaining "unsolved" Mendelian conditions

We created a set of genes depleted of certain functional classes of variation in ExAC/gnomAD by selecting four complementary measurements of constraint – Constrained Coding Regions (CCRs) [3], Nonsense-Mediated Decay escape intolerance (NMD-)[4], loss-of-function observed/expected upper bound (LOEUF) fraction[5], and missense observed/expected[5].

CCRs are designed to detect extremely constrained regions within genes (e.g., binding pocket or functional domains when the rest of the gene can tolerate variation). NMD- genes are relatively depleted for protein truncating variants that are predicted to escape nonsense-mediated decay due to their location near the 3' end of the gene and are potential candidate genes that may cause disease via gain of function. LOEUF is an updated successor score to the ExAC pLI score (probability of loss of function intolerance) that detects genes that exhibit a deficit of predicted loss of function variation and are thus likely to be haploinsufficient. While an updated missense constraint-specific score has not yet been described by the gnomAD consortium, the same expected/observed upper fraction metric is available for missense variation.

We designated a gene as being "supported by human data" if the gene was included in any of the following gene sets:

(1) >90%ile of CCRs (as advised by the authors);
(2) in the top 1,996 ranked NMD- gene list;
(3) in the top 40%ile of LOEUF; or
(4) in the top 20%ile of missense observed/expected scores.

The 40%ile cutoff was chosen for LOEUF because the gnomAD manuscript demonstrates that the enrichment for known Mendelian genes is similar for the 0-40%iles for LOEUF (~20-25% of genes in each decile). The fraction that are known Mendelian genes begins to decrease at the 50% decile, so we chose the 40%ile as a conservative cutoff. Because the missense observed/expected score has not yet been fully characterized by the gnomAD consortium, we chose the 20%ile as an informal cutoff that replicates the recall of LOEUF -- ~22% of the genes in the 0-20%ile of the missense observed/expected metric are known Mendelian genes. These cutoffs are still conservative underestimates of the number of genes with evidence for constraint according to these metrics.

We designated a human gene as being "supported by mouse data" if at least one abnormal phenotype was identified in at least one mutant mouse strain for that gene's mouse ortholog. We downloaded "HMD_HumanPhenotype.rpt" from http://www.informatics.jax.org/downloads/reports/index.html#pheno on March 4, 2019. We considered abnormal phenotypes to be any entry, including lethality, in the Mammalian Phenotype column of this file except MP:0003012 (no phenotypic analysis) and MP:0002873 (normal phenotype).

Supplemental References

1. Hartley, T., Balcı, T.B., Rojas, S.K., Eaton, A., Canada, C., Dyment, D.A., and Boycott, K.M. (2018). The unsolved rare genetic disease atlas? An analysis of the unexplained phenotypic descriptions in OMIM®. Am J Med Genet *178*, 458–463.

2. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., Mcmillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet *97*, 199–215.

3. Havrilla, J.M., Pedersen, B.S., Layer, R.M., and Quinlan, A.R. (2018). A map of constrained coding regions in the human genome. Nat Genet *75*, 1–12.

4. Coban Akdemir, Z., White, J.J., Song, X., Jhangiani, S.N., Fatih, J.M., Gambin, T., Bayram, Y., Chinn, I.K., Karaca, E., Punetha, J., et al. (2018). Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. Am J Hum Genet *103*, 171–187.

5. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv.