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

Michael J. Bamshad,[1,2,3,*] Deborah A. Nickerson,[2,3] and Jessica X. Chong[1,3]

Gene discovery for Mendelian conditions (MCs) offers a direct path to understanding genome function. Approaches based on next-generation sequencing applied at scale have dramatically accelerated gene discovery and transformed genetic medicine. Finding the genetic basis of ~6,000–13,000 MCs yet to be delineated will require both technical and computational innovation, but will rely to a larger extent on meaningful data sharing.

Most of what we understand about how the human genome encodes function and what constitutes a causal variant has been motivated by gene discovery for Mendelian conditions (MCs).[1] Indeed, the vast majority of variants of known function in the genome underlie MCs, and study of MCs is currently the gold standard for adjudicating variants of unknown significance (VUSs). While new computational strategies as well as technologies (e.g., multiplexed assays for variant effects) and biological models that can be scaled to assess the impact of every possible variant offer an unprecedented opportunity to explore genome function,[2] it is the study of natural genome variation in humans when manifested by MCs that still provides the most efficient and putatively cost-effective path to link genotype with human phenotype. Moreover, this path leads directly to development and testing of new preventive, diagnostic, and treatment strategies for rare diseases (e.g., cystic fibrosis transmembrane regulator modulators).[3] So it is not surprising that the overwhelming majority of genetic diagnostic tests, results returned to families, and results that inform reproductive options, guide clinical management, and enable selection of therapeutics are based on discoveries of the genes underlying MCs.

Prior to 2010, gene discovery was driven by positional cloning, which requires information about the genomic location and function of a candidate gene, the phenotype, or both, limiting its effectiveness. Introduction of computational approaches based on exome sequencing (ES) that required neither was a disruptive innovation that replaced not only positional cloning, but virtually all incumbent approaches to gene discovery.[4–6] Accordingly, thousands of MCs that had been intractable to conventional gene discovery approaches for various reasons suddenly became solvable using ES. The impact has been stunning.[7,8]

Rapid adoption of ES, and approaches using next-generation sequencing (NGS) in general, to identify genes associated with MCs (1) markedly accelerated the rate of novel gene discovery for MCs (i.e., a gene not previously known to underlie an MC [novel gene] or a gene found to underlie a novel condition or known but unexplained MC); (2) enabled identification of >1,000 new MCs; (3) replaced "phenotype-driven" with "genotype-driven" syndrome delineation; (4) led to the deconstruction of heuristic phenotypic classes (e.g., developmental disorders, autism, epilepsy, congenital heart defects) into separate and often distinct MCs with otherwise low clinical recognizability (LCR); and (5) expanded our understanding the phenotypic effects of thousands of genotypes and MCs. Summed across all genes underlying MCs discovered in the past decade, application of ES and NGS has rapidly advanced our knowledge of genome function, transformed the clinical practice of genetic medicine and challenged our understanding of fundamental concepts in human genetics (e.g., risk, penetrance, variable expressivity, etc.). However, gene discovery for MCs risks becoming a victim of its own success: There is perception in some circles that the pace of discovery is leveling off or even declining, the number of unsolved MCs is small, the remaining MCs are unlikely to be solvable by existing ES-based approaches, and/or many of the remaining MCs will be solved in the course of clinical diagnostic testing alone. We offer a different perspective.

From 1900 to 1950, a handful of new MCs were characterized each year (Figures 1A–1C, Figure S1, see Supplemental Methods in Supplemental Data). In the 1950s, the rate at which new MCs were delineated increased coincident with the emergence of the disciplines of medical and biochemical genetics and dysmorphology, reaching a peak in the 1970s. Despite the growing number of rare MCs cataloged, a relatively small number (i.e., ~40) of genes underlying MCs were known prior to the introduction of positional cloning in 1986.[9] Subsequently, both the rate of MC delineation and the rate of reports (i.e., publications) of discovery

[1]Department of Pediatrics, University of Washington, Seattle, WA 98195, USA; [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [3]Brotman-Baty Institute for Precision Medicine, Seattle, WA 98195, USA
*Correspondence: mbamshad@uw.edu
https://doi.org/10.1016/j.ajhg.2019.07.011.

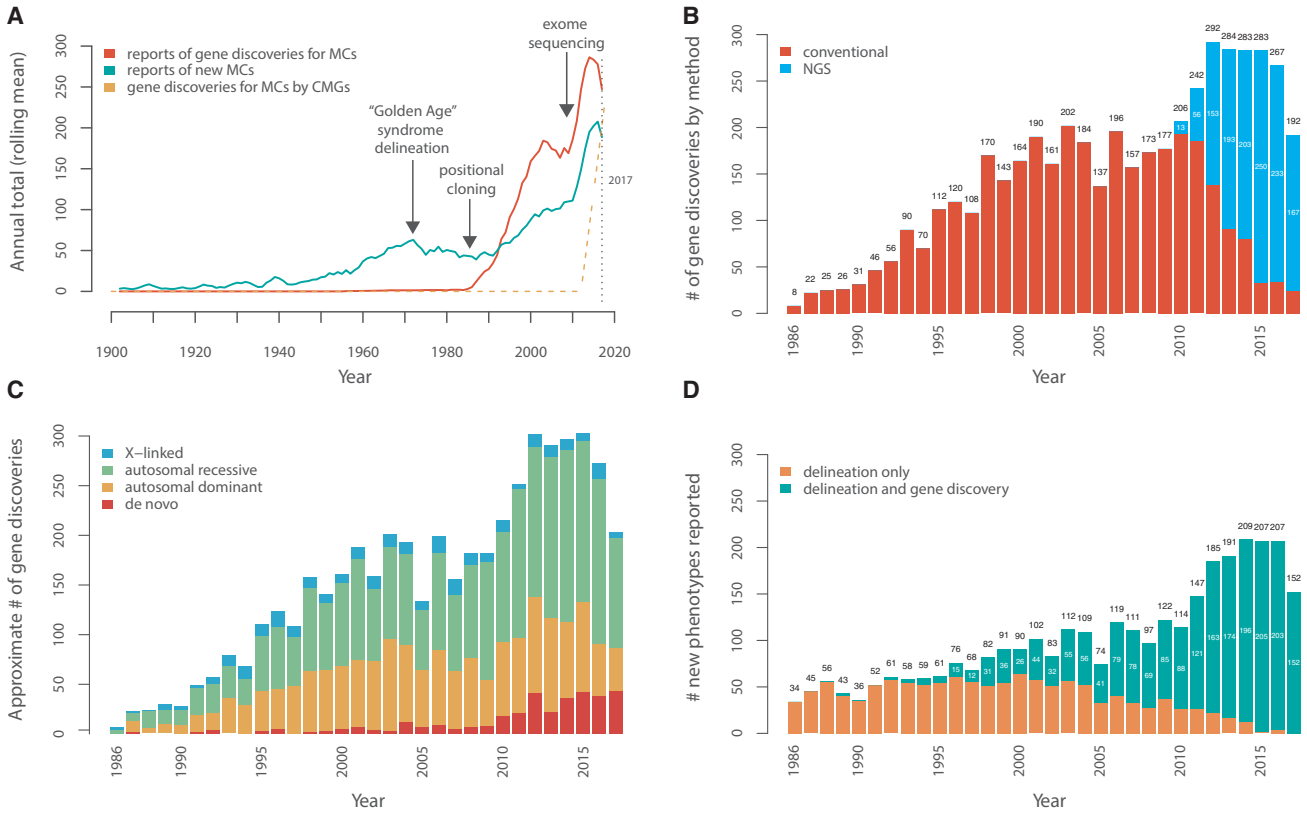


Figure 1. Annualized Metrics of Gene Discovery for Mendelian Conditions

(A) Approximate rates of reported gene discoveries for Mendelian conditions (MCs) and of delineation of MCs over time (1900–2017). 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. These data also show the impact of technical and methodological advances that fueled gene discovery, namely the impact of positional cloning in 1986; the development of dense, genome-wide linkage maps in the early 1990s; and increasing knowledge, gained via the Human Genome Project (1990–2001), of the physical locations and sequence content of genes. Rates shown for gene discoveries and syndrome delineations reflect publications, not unpublished discoveries or syndrome delineations, as recorded in OMIM and extracted by text analysis. The dashed line represents the number of genes for MCs reported discovered by the Centers of Mendelian Genomics, most of which are unpublished.

(B) Approximate number of gene discoveries per year for MCs made by exome sequencing (ES) and next-generation sequencing (NGS) versus conventional approaches. Following the introduction of positional cloning in 1986 and of ES in 2010, there were rapid increases in the rate of gene discovery for MCs. Each new approach made gene discovery possible for MCs that had otherwise been intractable to prior approaches, and this added to the baseline rate of gene discovery. Since 2010, NGS-based approaches (blue) have been used to make nearly all gene discoveries for MCs compared to conventional approaches (red).

(C) Approximate number of gene discoveries per year for MCs by mode of inheritance. The estimated proportion of gene discoveries for MCs due to *de novo* variants (DNVs, red) has increased since 2010 as NGS made routine identification of such variants possible (see Supplemental Methods in Supplemental Data). However, the proportions of gene discoveries for autosomal recessive (green), dominant (orange), and X-linked (blue) MCs each continue to be equal to or greater than the number of discoveries for MCs due to *de novo* (red) variants. Moreover, until 2010, the vast majority of gene discoveries for MCs were for inherited conditions (~97% before 2010; ~89% from 2010–2016; and ~79% in 2017), so still, most MCs known to date (~90%–93%) are predominately due to inherited variants. MCs assessed as being attributable to DNVs were excluded from the autosomal dominant and X-linked groups and vice versa, and MCs attributed to both dominant and recessive variants are not shown. Modes of inheritance were inferred by text analysis of OMIM entries.

(D) Impact of ES and NGS on the rate and method of syndrome delineation. Classical syndrome delineation (orange) is phenotype-driven and proceeds by ascertaining multiple individuals with overlapping clinical findings and then identifying of the underlying gene. In contrast, for genotype-driven syndrome delineation (teal), persons with overlapping clinical findings are identified only after discovery that they share pathogenic variants in the same candidate gene. Introduction of NGS-based approaches rapidly extinguished phenotype-driven syndrome delineation, and as of 2017, new MCs have been reported only after discovery of the underlying gene. MCs for which the gene was discovered in the same year as the first publication of data from an individual with the MC were categorized as genotype-driven; MCs for which data on the first individual with the MC was published one year or more prior to gene discovery were categorized as phenotype-driven (see Supplemental Methods in Supplemental Data).

of genes associated with MC increased steeply (Figure 1A). Specifically, between 1986 and 1997, the number of MCs delineated and the number of reported discoveries of genes underlying MCs increased annually by ~3 ($p = 5.9 \times 10^{-4}$) and ~10 ($p = 1.2 \times 10^{-6}$) per year, respectively. However, between 1998 and 2010, prior to the introduction of ES in 2009[6] and its application toward gene discovery in 2010,[4,10] the rate of reports of gene discoveries for MCs had plateaued (Figure 1A). After 2010, the number of MCs delineated and the number

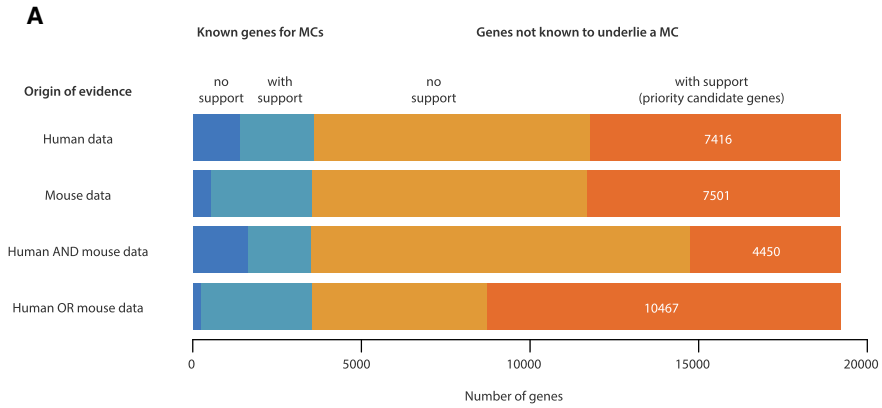
of discoveries of genes underlying MCs reported each year markedly increased by ~19 ($p = 0.006$) and ~14 ($p = 0.06$), respectively, per year through 2015. The impact has been rapid and profound. NGS-based approaches (primarily ES) have led to ~36% (1,268/3,549) of all reported Mendelian gene discoveries (Figures S1 and S2), and by the end of 2017, the majority (87%) of reported gene discoveries were made via NGS-based approaches (Figure 1C).

The annual number of MCs for which the genetic basis is reported peaked between 2012 and 2015 and declined slightly each year thereafter (Figures 1A and 1B), suggesting that perhaps the underlying rate of gene discovery is declining as well. To distinguish whether discovery trends parallel reporting trends, we reviewed annualized totals of novel gene discoveries, both published and unpublished, publicly reported by the National Institutes of Health (NIH) Centers for Mendelian Genomics (CMG). By 2015, the CMGs made 419 novel gene discoveries and published 93, a ratio of gene discoveries to published reports of 4.5.[8] By the end of 2018, the CMGs had made 1,937 discoveries, an increase from ~120 discoveries per year to ~244 per year, but reported only 287 (6.7 gene discoveries per reported discovery).[7] In other words, despite a drop in publication rate, the rate of discovery has continued to increase. This reporting delay, of obvious concern, has multiple explanations, but is due in part to investigators spending time to ascertain additional affected families, deeply characterize phenotypes and delineate new MCs (an obligatory consequence of the shift to genotype-driven delineation), and generate functional data to establish causality, link variants to function and outcome, and leverage high-impact publications. Whether the pace of discovery of investigators collaborating with CMGs reflects the worldwide experience of all investigators is unclear, but the total number of novel gene discoveries published by the CMGs represents about one of every six publications of novel gene discoveries; this suggests that these data seem a reasonable estimator of discovery trends. Public reporting of numbers of novel genes underlying MCs discovered each year by other large-scale programs would help to validate these results.

Online Mendelian Inheritance in Man (OMIM) and Orphanet both include only several hundred MCs for which the underlying gene is still unknown,[11] so it is often alleged that at the current pace of discovery, the genetic basis of nearly all MCs will be identified within the next ten years or so. However, the pervasive use of ES and NGS to identify the genetic basis of MCs has also accelerated the pace of novel MC delineation (Figure 1D). Historically, delineation of new MCs has been phenotype driven. That is, a person, persons, or a family with a recognizable but heretofore unreported pattern of phenotypic findings was ascertained, clinical characterization of additional persons with an overlapping pattern of findings established the canonical phenotype, and subsequently the underlying genetic basis of the canonical phenotype was sought. In contrast, new MCs are now delineated only after discovery of their genetic basis (i.e., delineation is genotype-driven), that is the rates of syndrome delineation and gene discovery have become inextricably linked, with >80% of the novel gene discoveries reported each year representing genes for newly described MCs (Figure 1D and Figure S3). The historical totals of MCs described by genotype-driven ($n = 2,023$) versus phenotype driven ($n = 3,149$) delineation are approaching equality (Figure S4), and ultimately most MCs will be ascertained via genotype-driven delineation.

What is the source of these new MCs? Foremost, NGS of large numbers of persons with a condition representative of a phenotypic class has enabled delineation, based on the underlying gene responsible, of hundreds of new LCR-MCs (e.g., intellectual disability, developmental disorders[12]). This splitting of phenotypic classes into separate LCR-MCs is, and will continue to be, a primary source of newly delineated MCs and novel disease-associated gene discoveries. Moreover, even many MCs considered to be of high clinical recognizability (HCR) (e.g., Brachmann-De Lange [MIM: 122470], Noonan [MIM: 163950], and Kabuki [MIM: 147920]) are being found to be comprised of multiple MCs caused by variants in several genes. For some such HCR-MCs (e.g., arguably, Coffin-Siris [MIM: 135900]), the canonical phenotype and distribution of phenotypic effects are, in large part, indistinguishable across different causal genes, although they may eventually be resolvable from one another by deep phenotyping. In other words, we suggest that genetic heterogeneity in an MC is often a reflection of our general lack of knowledge of gene/genotype-phenotype relationships. Thus, we predict that as our understanding of the phenotypic effects of variants which cause MCs improves, fewer and fewer MCs will be considered genetically heterogeneous. However, while burden analyses of increasingly larger cohorts of proband-parent trios diagnosed by phenotypic class (e.g., autism, congenital heart defects) are predicted to increase the number of MCs found to be caused by DNVs,[13] there may be diminishing returns as in many, if not most, persons categorized by rare disease phenotypic class; the condition is likely oligogenic or polygenic rather than an MC.[14]

While phenotypic classes might be a rich source of new MCs, how many as-of-yet unknown MCs might exist? Catalogues of hundreds of well-established, unexplained MCs, loci for many of which have been mapped, demonstrate that the opportunity for discoveries among inherited MCs remains high. Moreover, MCs with or without a known gene are still almost entirely ascertained from populations of European or Middle Eastern ancestry. To what extent this has limited our scope of knowledge of MCs in general is unclear but should be empirically assessed.



B

Known genes for MCs		Genes not known to underlie a MC	
no support	with support	no support	with support
1393	2180	8259	7416
533	2986	8174	7501
1623	1896	11225	4450
249	3270	5208	10467

Figure 2. Estimated Number of Genes for Mendelian Conditions

(A) Priority candidate genes for Mendelian conditions (MCs) yet to be discovered or delineated can be identified based on detecting a deficit of variation in healthy human controls due to purifying selection (i.e., human data) or the existence of at least one mutant mouse line with an abnormal phenotype (i.e., mouse data). Each stacked bar illustrates, from left to right, the number of human genes known to underlie an MC but not supported by a given type of evidence (dark blue); known to underlie an MC and supported by evidence (light blue); not known to underlie an MC (i.e., novel) and not supported by evidence (burnt orange); and not known to underlie an MC but evidence suggests that it does and it is therefore a priority candidate gene (bright orange) supported by evidence. Selecting the intersection of genes supported by both mouse and human data yields 4,450 priority candidate genes that are likely to underlie one or more novel MCs.

(B) Alternatively, the union of genes supported by either mouse or human data suggests there are at least 10,467 priority candidate genes likely to underlie one or more new MCs. Both of these estimates are probably conservative for several reasons: ~25% of genes underlie two or more distinct MCs (adjusting for this yields ~6,100–14,400 potential novel MCs), mutant phenotypes for >12,000 mouse genes have not yet been assessed, and current human constraint metrics still lack power to detect constraint in ~30% of all genes and are underpowered to detect constraint against homozygous loss of function. (See Supplemental Data for details.)

Accordingly, there is widespread interest in prioritizing efforts to characterize MCs and their underlying genetic basis in under-represented populations, with particular emphasis on surveying population isolates and populations with high levels of consanguinity that are typically not included in large-scale efforts to discover genes for MCs. A dedicated large-scale effort would require extensive infrastructure, coordination, governance, and resources, but the return on investment could be substantial.

Orthogonal evidence from humans and mice suggests that there are conservatively at least twice as many MCs that have yet to be delineated as there are known MCs (Figure 2 and see Supplemental Methods in Supplemental Data). Analysis of large databases of human coding variation (e.g., gnomAD) using metrics (e.g., Constrained Coding Regions,[15] LOEUF,[16] missense OEUF,[16] and nonsense-mediated decay escape rank[17]) that assess the depletion of classes of functional variation in specific regions or genes identifies 9,596 genes under strong selective constraint that are therefore priority candidates for MCs, 77% (i.e., 7,416) of which would be novel genes for MCs (Figure 2). Furthermore, of the 10,487 mouse genes in the Mouse Genome Database (MGI)[18] that are each linked to at least one non-lethal phenotype in a mutant strain, the human orthologs for 72% (i.e., 7,501 priority candidate genes) have yet to be shown to underlie an MC (Figure 2). Taken together, mouse- and human-supported data yield a total of 13,737 priority candidate genes for MCs, 78% (10,467) of which would be novel (Figure 2). Even under the more conservative assumption that a gene must be considered a priority candidate in both human and mouse, there are 6,346 genes predicted to underlie an MC, of which 4,450 (70%) would be novel (Figure 2). Accordingly, if we assume that each candidate gene underlies a single MC, there are ~1.5–3 times as many novel genes (4,450–10,467) for MCs yet to be discovered as there are genes (3,519) known already to underlie an MC. If we extrapolate that the same proportion of these genes underlie multiple MCs as is the case for known genes for MCs (i.e., 16% underlie two MCs, 4.7% underlie three, 1.8% underlie four, etc.), we predict that at a minimum, ~6,100–14,400 MCs remain to be discovered. And these figures are still a considerable underestimate of the number of unsolved MCs because we did not account for the fact that mutant phenotypes for over half (~12,000) of all protein-coding mouse genes have yet to be assessed. Moreover, we used conservative cutoffs for defining constrained genes, and current human constraint metrics both lack power to detect constraint in ~30% of all human genes and are underpowered to detect constraint against homozygous loss of function[16] mutations. For example, our analysis identifies 1,393 known genes for MCs that are not constrained in humans. The majority (>75%) of these genes underlie MCs that are inherited in an autosomal recessive pattern.

The widespread use of ES in general, and in diagnostic settings in particular, has highlighted the contribution of *de novo* variants (DNVs) to risk of rare disease in general, and especially of MCs that markedly reduce fitness.

Yet our analysis of OMIM suggests that each year prior to and since 2010, the majority (~80%–90%; see Supplemental Materials in Supplemental Data) of discoveries of genes underlying MCs are for MCs that are typically inherited in an autosomal recessive, dominant, or X-linked pattern rather than due entirely to DNVs (Figure 1C and Figure S5). A similar estimate (76%) is obtained from analysis of discoveries of genes for MCs identified in the course of diagnostic testing via ES (K. Retterer, GeneDx, *personal communication*). In diagnostic settings, variants in novel candidate genes for MCs are usually resolved as pathogenic via collaboration with dedicated discovery efforts by research programs. Such successful informal collaboration between industry and academic gene discovery efforts represents an opportunity that could be, if not should be, leveraged at scale for mutual benefit (i.e., to further accelerate gene discovery and in turn increase diagnostic rates). More importantly, such a collaborative effort across diagnostic labs and researchers could translate into a big windfall for families with rare diseases.

Even with an accelerating pace of gene discovery for MCs, efforts to find a gene underlying an MC are successful only about half the time. This observation underscores the reality that there are myriad factors limiting the rate of MC gene discovery using current ES- and NGS-based approaches. Such limiting factors include: (1) inability to robustly predict the impact of missense, synonymous, intronic, splice, and non-coding variants; (2) limited access to high-throughput functional validation of candidate variants; (3) much slower co-evolution of the infrastructure and regulatory framework necessary to share genomic data openly and at scale worldwide (hundreds of putative gene discoveries are unreported); (4) technical limitations of approaches based on ES (e.g., identifying indels, copy number variants, repeat expansions, structural variants, etc.); and (5) the challenges (insufficient resources, lack of organized efforts) of ascertaining and deeply phenotyping families with high priority candidate genotypes. The cost and impact of overcoming each limitation varies substantially, but to what extent and under what circumstances remains a topic of intense investigation both in the public and private sectors.

Many of the efforts to further improve the success of MC gene discovery have focused on application of new sequencing technologies and variant calling and/or annotation. Whole-genome sequencing (WGS), in particular, has been considered by many to be the logical tool to supplant ES for MC gene discovery. Yet to date, WGS has, after excluding coding, near-splice site, or structural variants (SVs) overlapping known MC-associated genes, yielded few discoveries of novel genes or loci underlying MCs.[19] This is due in part to limited availability and utility of annotations for untranslated regions, enhancers, insulators, silencers, RNA genes, and microRNAs as well as callers for SVs and repeat expansions. However, it also underscores the observation that the vast majority of known MCs are caused predominantly by coding variants with large effect sizes, and both ES and NGS gene panels already cover such coding regions robustly. Indeed, with the exception of repeat expansions, pathogenic non-coding variants have been reported for only 156 MCs and of these conditions, 150 (~96%) of the genes discovered were found via variants accessible to ES or microarrays.[19]

There are a handful of examples of successful gene discovery using WGS to identify pathogenic non-coding variants when ES failed, and these can inform judicious use of WGS for discovery. In virtually every case, the search space was reduced to a small fraction of the genome via linkage analysis, homozygosity mapping, or identification of shared chromosomal microdeletions or duplications. For a small number of recessive MCs, identification of only one protein-coding variant led to the search for a non-coding variant *in trans* via WGS.[20–24] Additionally, most non-coding pathogenic variants identified to date are small to moderate-size (e.g., multiple nucleotides) deletions, insertions, mobile element, or repeat expansions and contractions that remove or alter the sequence of a large portion or all of a regulatory element,[9–13] duplicate the element in its entirety,[25] or translocate it out of its normal sequence context.[26,27] Enrichment for non-coding SNVs in regulatory regions has been detected in, for example, developmental disorders[28] and autism,[29] but proving the pathogenicity for any one specific SNV is challenging because most are unique and alter non-overlapping bases.

Use of transcriptome sequencing (RNA-seq) to identify abnormally spliced transcripts and/or assess transcript abundance facilitates identification of non-coding variants, deep intronic splice variants,[16–18] and to a lesser extent, synonymous variants with unexpected effects on splicing. However, while useful for diagnosis or validating effects of variants detected by ES or WGS, successful applications of RNA-seq to novel gene discovery for MCs are currently constrained by lack of knowledge of and/or access to disease-relevant tissue or the expense of creating transdifferentiated cell lines[19,20] from affected individuals and controls.[30] Thus, while WGS may be technologically superior to ES at detecting non-coding and structural variants and to RNA-seq at highlighting pathogenic variants that alter splicing or transcript abundance, there is little evidence to date that the predicted thousands of currently undiscovered MCs will be even largely caused by non-coding and/or deep intronic variants that can only be detected by widespread application of RNA-seq or WGS.

Perhaps the principal bottleneck to discovering genes underlying MCs is the lack of meaningful sharing at scale of genetic data and phenotypic information from families with a known or suspected novel MC. Millions of people with rare diseases, particularly

children, have undergone targeted genetic testing, and ES and/or WGS has been performed on hundreds of thousands of them.[7,31,32] Yet most of these results are buried in medical records, proprietary or restricted-access databases, and scientific papers, and most are difficult to access, much less leverage for gene discovery. In the U.S., institutions that participate in research and/or clinical care, including diagnostic labs, must comply with federal regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the Common Rule, which place boundaries, as well as protections, on use of patient and research participant data. The ambiguity of these boundaries can make navigating regulatory and privacy issues surrounding data sharing challenging and expensive, but sharing of de-identified data (e.g., candidate gene and non-identifying phenotypic information) among researchers, clinicians, and scientists in academics and industry is relatively straightforward. Over the past decade, a growing number of web-based platforms and databases to support data sharing have been developed and linked to one another via a federated network called the MatchMaker Exchange (MME)[33] in order to facilitate matching of candidate genes. Matching within and between nodes of the MME has facilitated hundreds of discoveries of novel genes underlying MCs.

But such sharing doesn't happen nearly as often as it could and should. In some instances, it is fear of noncompliance with HIPAA or the Common Rule and the risk of fines, suspension, etc., a lack of awareness of the power of sharing, or inaccessibility to platforms for sharing. In other cases, the intangible incentives to share are offset by concern that sharing might result in losing priority to publish or diminish competitiveness for grant funding, which in turn could adversely affect professional recognition and career advancement. Lately, sharing has also been threatened by attempts to monetize the discovery process by, for example, commercial start-ups and advocacy groups who generate or aggregate data and then market it for profit or fundraising. Moreover, matching on candidate genes without additional data (e.g., phenotype, mode of inheritance, variant) is increasingly inefficient because even nowadays, most matches are false positive matches. This problem will only worsen as the number of candidate genes shared across MME approaches all human genes. Finally, matching does not ensure public reporting, much less timely reporting, of the results as demonstrated by the increasing ratio of discoveries to publications in the CMGs, so discoveries can remain unknown for years and the information unable to be used by diagnostic labs and clinicians.

Families with MCs are arguably eager to share health data[34,35] if it can improve their care or the care of other families with the same condition, and when patients share their own health data online, HIPAA and the Common Rule do not apply. However, use of MME is restricted to clinicians and researchers who are often disincentivized to share data with one another, or who deprioritize it due to time constraints and the perception that is unlikely to be of benefit, and are even less likely to share it publicly. Over the past several years, families have increasingly turned to social media to circumvent the obstacles that limit data sharing by clinicians and researchers and to advertise their child's health information and candidate genes to the public at large to make themselves more discoverable. This approach has led to some notable successes that are widely cited in the popular press.[36] However, most efforts to use social media to facilitate case-matching fail. Some families are unable to gain the attention of suitable researchers and clinicians, and others lack the expertise to prioritize the information that should be shared, releasing non-standardized health and genetic information that cannot be easily compared or interpreted. Newer family facing platforms (e.g., MyGene2) aim to increase patient control over their data and create a public knowledge base of variant data linked to rare disease phenotypes in order to promote and facilitate data sharing directly from families while still allowing researchers and clinicians to share de-identified data.

Use of ES- and NGS-based strategies coupled with phenotype-driven delineation of MCs has brought us within reach of identifying genes for all known MCs that remain unsolved. But importantly, it has also revealed that the majority of MCs that exist likely have not yet been delineated because they are likely not recognizable as discrete entities by commonly employed clinical phenotyping approaches. Indeed, genotype-driven delineation of MCs has rekindled an emphasis on the need for deep-phenotyping in families if we are to achieve the goal of understanding genome function and more importantly, its links to human disease. Moreover, barring some currently unknown or unexpected biological mechanism that underlies the majority of MCs yet to be delineated, technical innovations will continue to yield only marginal improvements in rates of gene discovery. A deeper and more sustained impact on gene discovery for MCs will likely require a far broader commitment to more open, simpler, and more meaningful data sharing among all stakeholders in research and clinical care worldwide, as well as identifying resources to support a worldwide infrastructure to ascertain, sequence, and phenotype families with a broad range of clinical findings. The return on investment is nothing short of a keystone in the foundation of precision genomic medicine.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.07.011.

Acknowledgments

We thank all of the families, clinicians, and investigators for their participation and

support. We thank Kati Buckingham, John Carey, Katrina Dipple, Ada Hamosh, Colby Marvin, Jay Shendure, and Kathryn Shively for helpful discussion. This work was supported in part by grants from the National Human Genome Research Institute (NHGRI) and National Heart, Lung, and Blood Institute (NHLBI) grant HG006493 (to the University of Washington Center for Mendelian Genomics). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NHGRI and NHLBI or of the National Institutes of Health.

Declaration of Interests

The authors declare no competing interests.

Web Resources

Code for OMIM analysis and figures, https://github.com/jxchong/mendelian_commentary

MatchMaker Exchange, https://www.matchmakerexchange.org

Mouse Genome Database, http://www.informatics.jax.org

MyGene2, https://mygene2.org/MyGene2/

Online Mendelian Inheritance in Man, http://www.omim.org

References

1. Antonarakis, S.E., and Beckmann, J.S. (2006). Mendelian disorders deserve more attention. Nat. Rev. Genet. *7*, 277–282.
2. Starita, L.M., Islam, M.M., Banerjee, T., Adamovich, A.I., Gullingsrud, J., Fields, S., Shendure, J., and Parvin, J.D. (2018). A multiplex homology-directed DNA repair assay reveals the impact of more than 1,000 BRCA1 missense substitution variants on protein function. Am. J. Hum. Genet. *103*, 498–508.
3. Ramsey, B.W., Davies, J., McElvaney, N.G., Tullis, E., Bell, S.C., Dřevínek, P., Griese, M., McKone, E.F., Wainwright, C.E., Konstan, M.W., et al.; VX08-770-102 Study Group (2011). A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. N. Engl. J. Med. *365*, 1663–1672.
4. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30–35.
5. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet. *42*, 790–793.
6. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.
7. Posey, J.E., O'Donnell-Luria, A.H., Chong, J.X., Harel, T., Jhangiani, S.N., Coban Akdemir, Z.H., Buyske, S., Pehlivan, D., Carvalho, C.M.B., Baxter, S., et al.; Centers for Mendelian Genomics (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet. Med. *21*, 798–812.
8. 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.; Centers for Mendelian Genomics (2015). The genetic basis of Mendelian phenotypes: Discoveries, challenges, and opportunities. Am. J. Hum. Genet. *97*, 199–215.
9. Collins, F.S. (1995). Positional cloning moves from perditional to traditional. Nat. Genet. *9*, 347–350.
10. Vissers, L.E.L.M., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm for mental retardation. Nat. Genet. *42*, 1109–1112.
11. Hartley, T., Balcı, T.B., Rojas, S.K., Eaton, A., Canada, C.R., 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. C. Semin. Med. Genet. *178*, 458–463.
12. Deciphering Developmental Disorders Study. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.
13. Jin, S.-C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang, W., Sierant, M.C., et al. (2017). Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat. Genet. *49*, 1593–1601.
14. Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al.; Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium; BUPGEN; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium; and 23andMe Research Team (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. *51*, 431–444.
15. Havrilla, J.M., Pedersen, B.S., Layer, R.M., and Quinlan, A.R. (2019). A map of constrained coding regions in the human genome. Nat. Genet. *51*, 88–95.
16. 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. 10.1101/531210.
17. 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.; Baylor-Hopkins Center for Mendelian Genomics (2018). Identifying genes whose mutant transcripts cause dominant disease traits by potential gain-of-function alleles. Am. J. Hum. Genet. *103*, 171–187.
18. Bult, C.J., Blake, J.A., Smith, C.L., Kadin, J.A., Richardson, J.E.; and Mouse Genome Database Group (2019). Mouse Genome Database (MGD) 2019. Nucleic Acids Res. *47* (D1), D801–D806.
19. Smedley, D., Schubach, M., Jacobsen, J.O.B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N.L., McMurry, J.A., et al. (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. Am. J. Hum. Genet. *99*, 595–606.
20. LaCroix, A.J., Stabley, D., Sahraoui, R., Adam, M.P., Mehaffey, M., Kernan, K., Myers, C.T., Fagerstrom, C., Anadiotis, G., Akkari, Y.M., et al.; University of Washington Center for Mendelian Genomics (2019). GGC repeat expansion

and Exon 1 methylation of XYLT1 is a common pathogenic variant in Baratela-Scott Syndrome. Am. J. Hum. Genet. *104*, 35–44.

21. Karolak, J.A., Vincent, M., Deutsch, G., Gambin, T., Cogné, B., Pichon, O., Vetrini, F., Mefford, H.C., Dines, J.N., Golden-Grant, K., et al. (2019). Complex Compound inheritance of lethal lung developmental disorders due to disruption of the TBX-FGF pathway. Am. J. Hum. Genet. *104*, 213–228.
22. Wu, N., Ming, X., Xiao, J., Wu, Z., Chen, X., Shinawi, M., Shen, Y., Yu, G., Liu, J., Xie, H., et al. (2015). TBX6 null variants and a common hypomorphic allele in congenital scoliosis. N. Engl. J. Med. *372*, 341–350.
23. Albers, C.A., Paul, D.S., Schulze, H., Freson, K., Stephens, J.C., Smethurst, P.A., Jolley, J.D., Cvejic, A., Kostadima, M., Bertone, P., et al. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat. Genet. *44*, 435–439, S1–S2.
24. Wieczorek, D., Newman, W.G., Wieland, T., Berulava, T., Kaffe, M., Falkenstein, D., Beetz, C., Graf, E., Schwarzmayr, T., Douzgou, S., et al. (2014). Compound heterozygosity of low-frequency promoter deletions and rare loss-of-function mutations in TXNL4A causes Burn-McKeown syndrome. Am. J. Hum. Genet. *95*, 698–707.
25. Ngcungcu, T., Oti, M., Sitek, J.C., Haukanes, B.I., Linghu, B., Bruccoleri, R., Stokowy, T., Oakeley, E.J., Yang, F., Zhu, J., et al. (2017). Duplicated enhancer region increases expression of CTSB and segregates with keratolytic winter erythema in South African and Norwegian families. Am. J. Hum. Genet. *100*, 737–750.
26. Brewer, M.H., Chaudhry, R., Qi, J., Kidambi, A., Drew, A.P., Menezes, M.P., Ryan, M.M., Farrar, M.A., Mowat, D., Subramanian, G.M., et al. (2016). Whole genome sequencing identifies a 78 kb insertion from chromosome 8 as the cause of charcot-marie-tooth neuropathy CMTX3. PLoS Genet. *12*, e1006177.
27. Spielmann, M., Brancati, F., Krawitz, P.M., Robinson, P.N., Ibrahim, D.M., Franke, M., Hecht, J., Lohan, S., Dathe, K., Nardone, A.M., et al. (2012). Homeotic arm-to-leg transformation associated with genomic rearrangements at the PITX1 locus. Am. J. Hum. Genet. *91*, 629–635.
28. Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., and Hurles, M.E. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. Nature *555*, 611–616.
29. Turner, T.N., and Eichler, E.E. (2019). The role of de novo noncoding regulatory mutations in neurodevelopmental disorders. Trends Neurosci. *42*, 115–127.
30. Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V.A., Avsec, Ž., Herzog, M., Bader, D.M., Prokisch, H., and Gagneur, J. (2018). OUTRIDER: A statistical method for detecting aberrantly expressed genes in RNA sequencing data. Am. J. Hum. Genet. *103*, 907–917.
31. Stark, Z., Dolman, L., Manolio, T.A., Ozenberger, B., Hill, S.L., Caulfied, M.J., Levy, Y., Glazer, D., Wilson, J., Lawler, M., et al. (2019). Integrating genomics into healthcare: A global responsibility. Am. J. Hum. Genet. *104*, 13–20.
32. GeneDx announces completion of 100,000 exome sequences. https://www.globenewswire.com/news-release/2018/06/12/1520222/0/en/GeneDx-Announces-Completion-of-100-000-Exome-Sequences.html
33. Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M., Brunner, H.G., Buske, O.J., Carey, K., Doll, C., et al. (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. Hum. Mutat. *36*, 915–921.
34. Lambertson, K.F., Damiani, S.A., Might, M., Shelton, R., and Terry, S.F. (2015). Participant-driven matchmaking in the genomic era. Hum. Mutat. *36*, 965–973.
35. Burstein, M.D., Robinson, J.O., Hilsenbeck, S.G., McGuire, A.L., and Lau, C.C. (2014). Pediatric data sharing in genomic research: attitudes and preferences of parents. Pediatrics *133*, 690–697.
36. Might, M., and Wilsey, M. (2014). The shifting model in clinical diagnostics: how next-generation sequencing and families are altering the way rare diseases are discovered, studied, and treated. Genet. Med. *16*, 736–737.

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

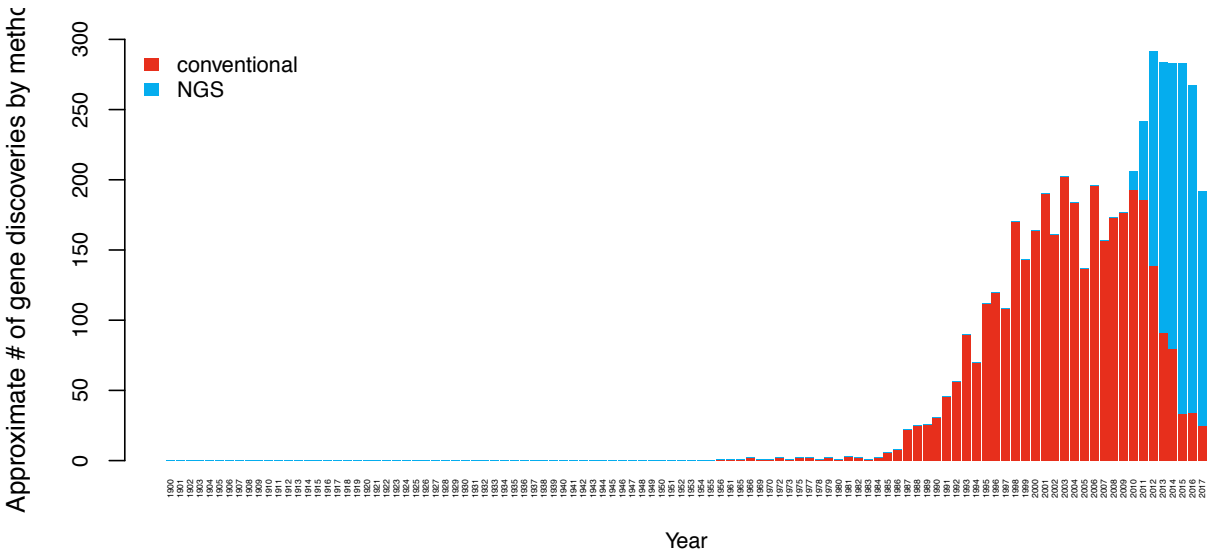


Figure S1. Estimated number of gene discoveries per year since 1900.
From 1900 to 1986, a handful of new MCs were characterized each year, and even fewer underlying genes were discovered. Beginning with the introduction of positional cloning in 1986, gene discovery for MCs accelerated greatly.

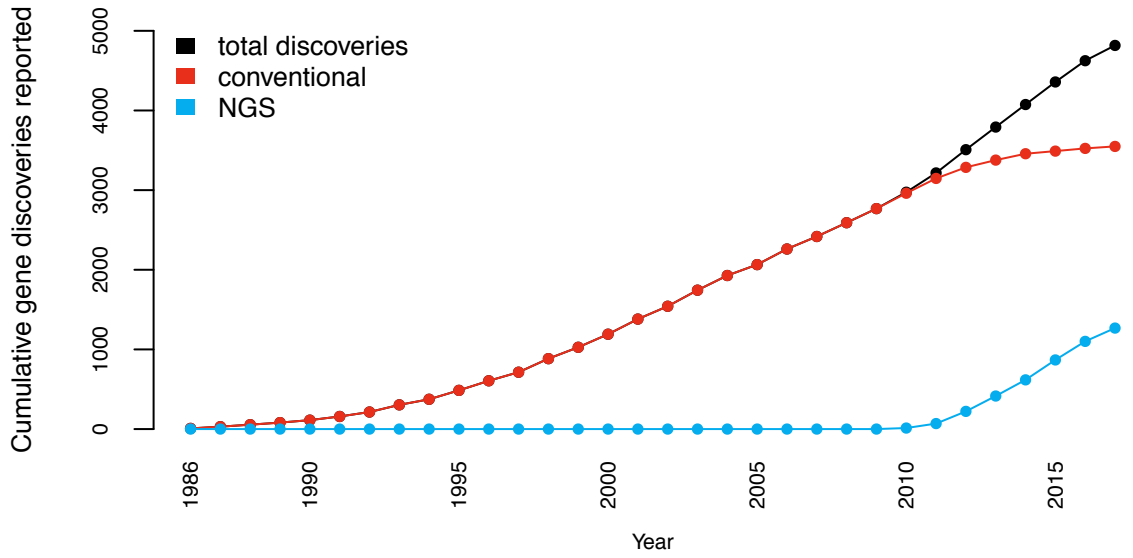


Figure S2. Cumulative estimated number of gene discoveries per year since 1986.
NGS-based approaches (primarily ES) have led to ~36% (1,268 / 3,549) of all reported Mendelian gene discoveries.

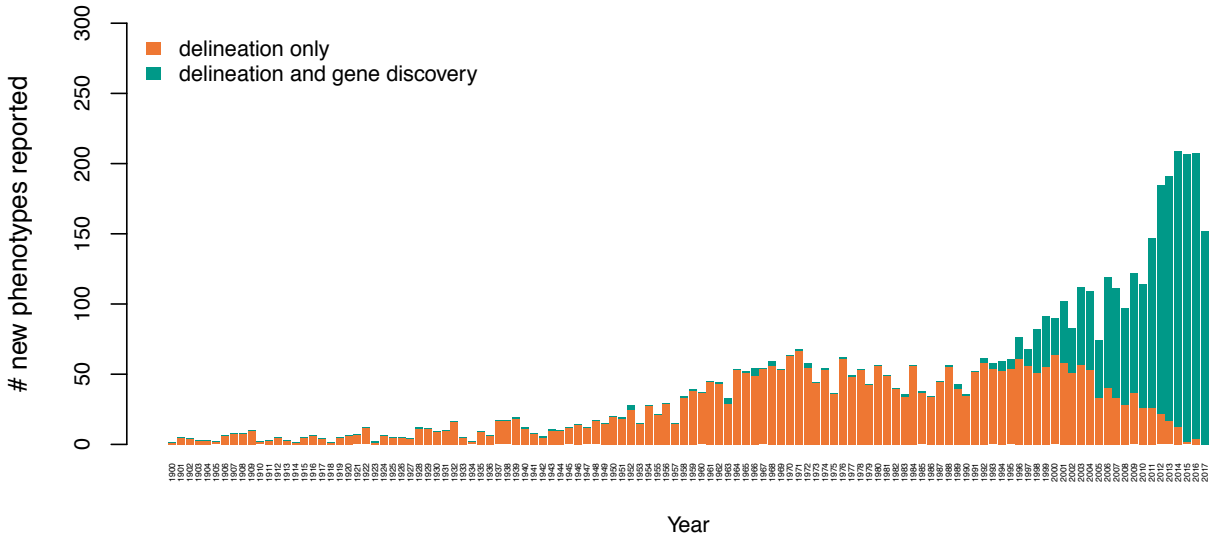


Figure S3. Estimated number of delineated syndromes per year since 1900.
Historically, particularly prior to the introduction of positional cloning in 1986, all or nearly all syndrome delineations were phenotype-driven. Classical syndrome delineation (orange) is phenotype-driven and proceeds by identifying 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.

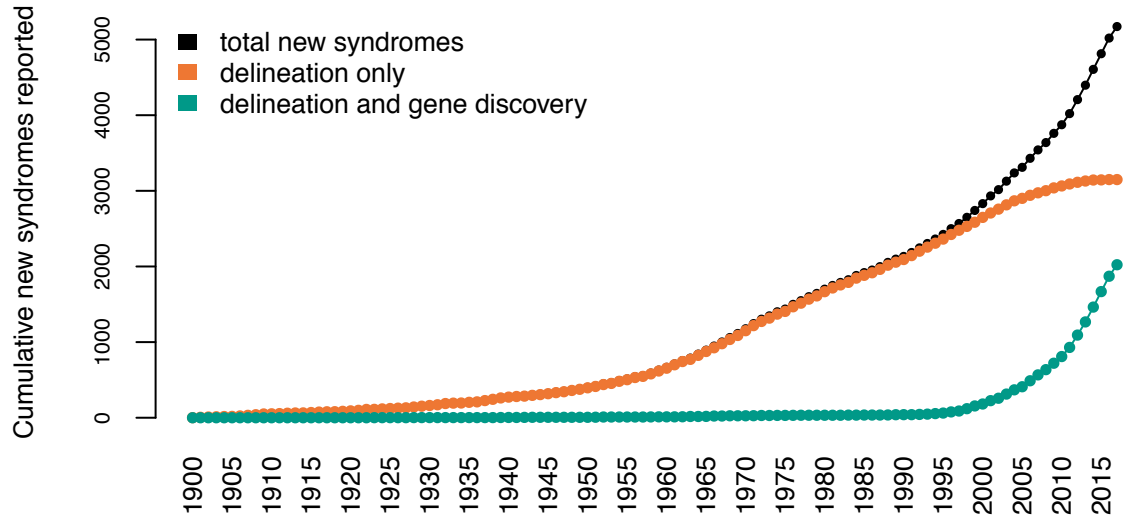


Figure S4. Cumulative estimated number of delineated syndromes per year since 1900.
In total, genotype-driven syndrome delineation has led to the description of 2,023 MCs vs. 3,149 MCs described via phenotype driven delineation. Ultimately most MCs will be ascertained via genotype-driven delineation.

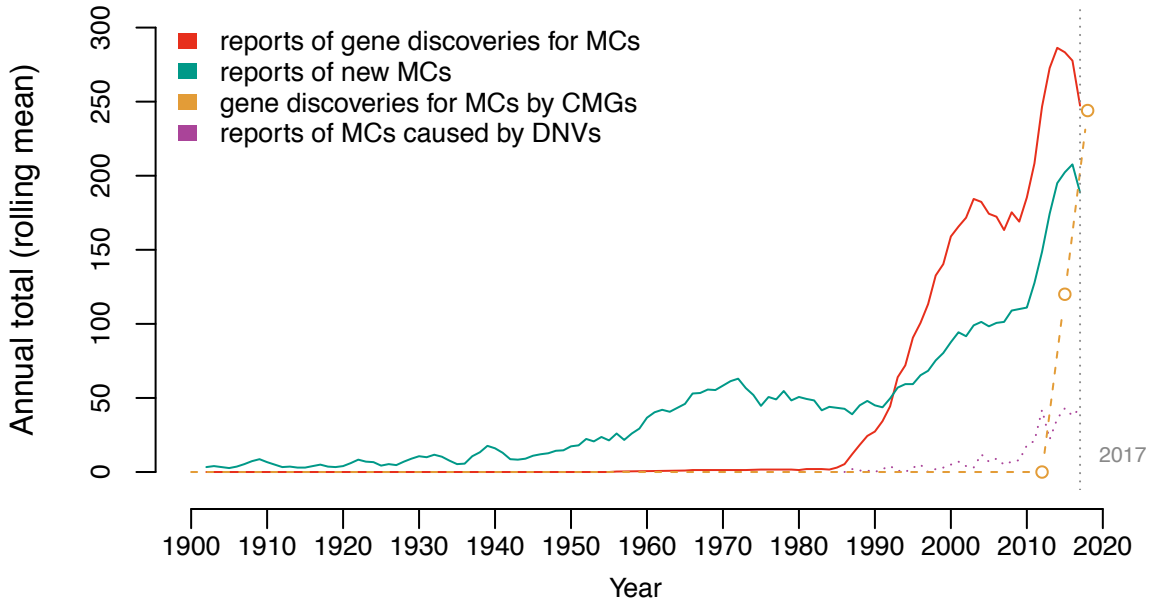


Figure S5. Approximate rates of reported gene discoveries for Mendelian conditions, delineation of Mendelian conditions, gene discoveries caused primarily by de novo variants, and unpublished discoveries by the Centers for Mendelian Genomics over time (1900-2017).

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.

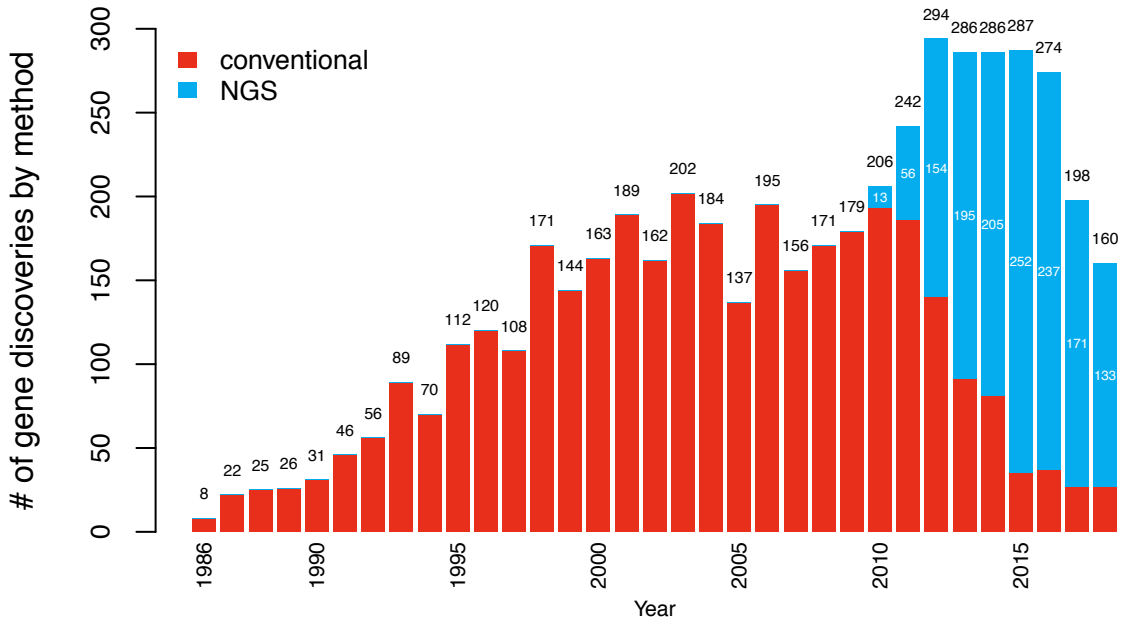


Figure S6. Approximate number of gene discoveries per year for MCs made by ES/NGS versus conventional approaches (including data through the end of 2018).
This graph is identical to Figure 1B except that it includes reports of gene discoveries through the end of 2018 (as cataloged in OMIM as of May 16, 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.