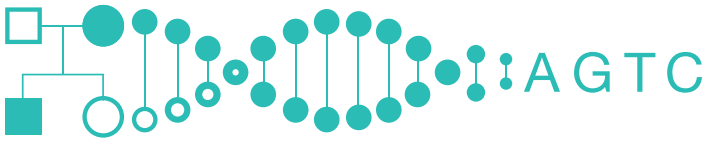
The Genetic Basis of Mendelian Conditions: Discoveries, Challenges, and Opportunities



Centers for Mendelian Genomics

NHGRI Future Opportunities for Genome Sequencing and Beyond

July 28-29, 2014, Rockville, MD

HISTORCIAL CONTEXT OF GENE DISCOVERIES FOR MENDELIAN CONDITIONS

Introduction

From its conception, it was anticipated that a major benefit of the Human Genome Project (HGP) would be new understanding of human disease. Part of this promise has been realized, with identification of the consequence of germline mutation for more than 2,700 protein-coding genes in humans. These mutations directly link altered protein to human phenotype, and have transformed understanding of the basic biology of every organ system as well as complex integrated physiology and metabolism. These discoveries have defined new diagnostic, preventive, and therapeutic strategies for a growing and broad range of both rare and common diseases.

These advances notwithstanding, far more discovery lies ahead than behind. The HGP and subsequent annotation efforts have established that there are ~20,000 protein-coding genes in humans. Sequencing across the phylogenetic tree has demonstrated that nearly all of these genes are conserved across 500 million years of vertebrate evolution, leaving little doubt that mutation of virtually every non-redundant gene will have phenotypic consequence, either constitutively or in response to specific environmental challenges. The continued description of new Mendelian conditions (MCs) by segregation patterns, as well as the advent of genotype-driven discovery of new MCs, supports this expectation.

The deep understanding of human biology and disease, as well as the often dramatic differences in physiology of humans compared to model organisms, argues that primary mutation discovery in humans will remain crucial to progress. Having defined a bounded set of protein-coding genes, determining the phenotypic consequence of their mutation in humans represents an attainable and vital goal, dramatically fueled by saltatory changes that have made the production and analysis of exome sequencing (ES) data rapid and inexpensive. Each success will be of enduring value, defining the diagnostic, preventive and therapeutic opportunities for the resulting diseases as long as *Homo sapiens* persist.

Additionally, it is anticipated that there will likely also be many non-coding mutations that have large phenotypic effects. To date, the vast majority of those that have been identified have been in elements that regulate expression of protein-coding genes, and have been presaged by discovery of coding region mutations. The continued reduction in cost of whole genome sequencing affords the opportunity for unbiased searches for such high-impact non-coding mutations. MCs in which mutations have not been discovered in coding regions and splice sites will be ideal candidates for such studies.

The value of studying Mendelian conditions

The vast majority of diseases studied to date are rare MCs. Rare diseases are arbitrarily defined as those diagnosed in <200,000 individuals in the United States or < 1/2000 individuals in Europe. Per this definition, >8,000 rare conditions have been reported and the majority of these are MCs and congenital anomalies. While each MC is uncommon, collectively they occur in ~0.4% of live births, and if all congenital anomalies are included, ~8% of live births have a genetic disorder recognizable by early adulthood. This translates to ~8 million children born each year with a serious genetic condition. In the U.S., they collectively affect more than 25 million people.

Collectively, genetic disorders are associated with high morbidity, mortality, and economic burden in both pediatric and adult populations. Birth defects are the most common cause of death in the first year of life, and each year, more than 3 million children under the age of five years die from a birth defect and a similar number survive with substantial morbidity. More than a decade ago, each child with a genetic syndrome was estimated to cost the healthcare system a total of 5 million dollars over their lifetime.

From a translational perspective, understanding the genetic basis of a MC can provide a means for direct diagnosis, carrier status determination, prenatal diagnosis, population screening and, in many instances, for more effective use of therapeutic approaches. Yet MCs remain challenging to diagnose. In a general clinical genetics setting, the diagnostic rate is ~50%. Across a broader range of rare diseases, the rate of diagnosis is even poorer. For example, in the NIH Undiagnosed Disease Program, the diagnostic rate was, despite state-of-the-art evaluations, only 34% in adults and 11% in children. The latter is noteworthy because ~50% of all rare diseases present in childhood. Inability to make a diagnosis limits the capacity to provide anticipatory guidance, counsel about recurrence risk, and treat with targeted therapeutics.

Our understanding of biological pathways underlying development, metabolism, and disease is largely imperfect. Yet discoveries of the genes underlying MCs have overwhelmingly made the largest contribution to our knowledge of gene function, gene regulation, development, and metabolism. Much of what is known about the relationship between gene function and human phenotypes is based on the study of rare variants underlying MCs. As discussed in greater detail

below, this understanding of biological pathways and gene function obtained by studying MCs has directly supported the development of important therapeutics for many common diseases.

The scope of Mendelian conditions

Our knowledge of the scope and phenotypic diversity of MCs is becoming increasingly sophisticated, but formidable gaps remain. Specifically, it is challenging to define the actual number of existing MCs, to delineate new MCs, to distinguish novel MCs from known MCs, and to develop metrics to assess the relationships and diversity of phenotypes caused by variants in the same gene. To date, ~7,417 rare MCs (7,296 monogenic; 121 chromosomal duplications/deletions) have been delineated and ~300 new MCs are delineated each year per data from OMIM (July 2014). However, delineation of new MCs is rate-limited by a worldwide lack of infrastructure, resources, and expertise to comprehensively evaluate families. Based on studies in mice, the majority of mice with loss-of-function (LOF) mutations compatible with survival to birth are associated with a recognizable phenotype and LOF of ~30% of genes results in embryonic lethality. **While variants in every human gene may not result in a MC, it is likely that mutation(s) in the vast majority of non-redundant human protein-coding genes result in a rare disease phenotype (Figure 1A), most of which have yet to be recognized, much less characterized.**

Mendelian gene discovery efforts to date

The first successful efforts to identify genes for MCs required extensive prior knowledge of disease biology including the identity of the affected protein. Discovery of the gene for chronic granulomatous disease in 1986 demonstrated that mapping followed by sequencing of genes proximate to the linkage peak offered a promising alternative, and over the next ten years, 42 genes for MCs were found via positional cloning. The ensuing two decades witnessed a steady accumulation of genes discovered for MCs by a combination of positional cloning and candidate gene approaches. However, increasingly it became clear that gene identification for most MCs was intractable using these approaches. Gene discovery strategies based on exome and whole genome sequencing (ES/WGS) introduced powerful alternatives that were agnostic to both biology and mapping data (**Figure 1B**). Since the first successful use of ES to find the gene underlying a MC in 2010, ES/WGS has proved to be a disruptive technology that has rapidly accelerated the pace of discovery of genes underlying MCs. As of July 2014, ~2,776 genes underlying ~3,593 MC have been discovered (**Figure 1A**) but the **genes underlying another 3,703 MCs, at a minimum, remain to be discovered.**

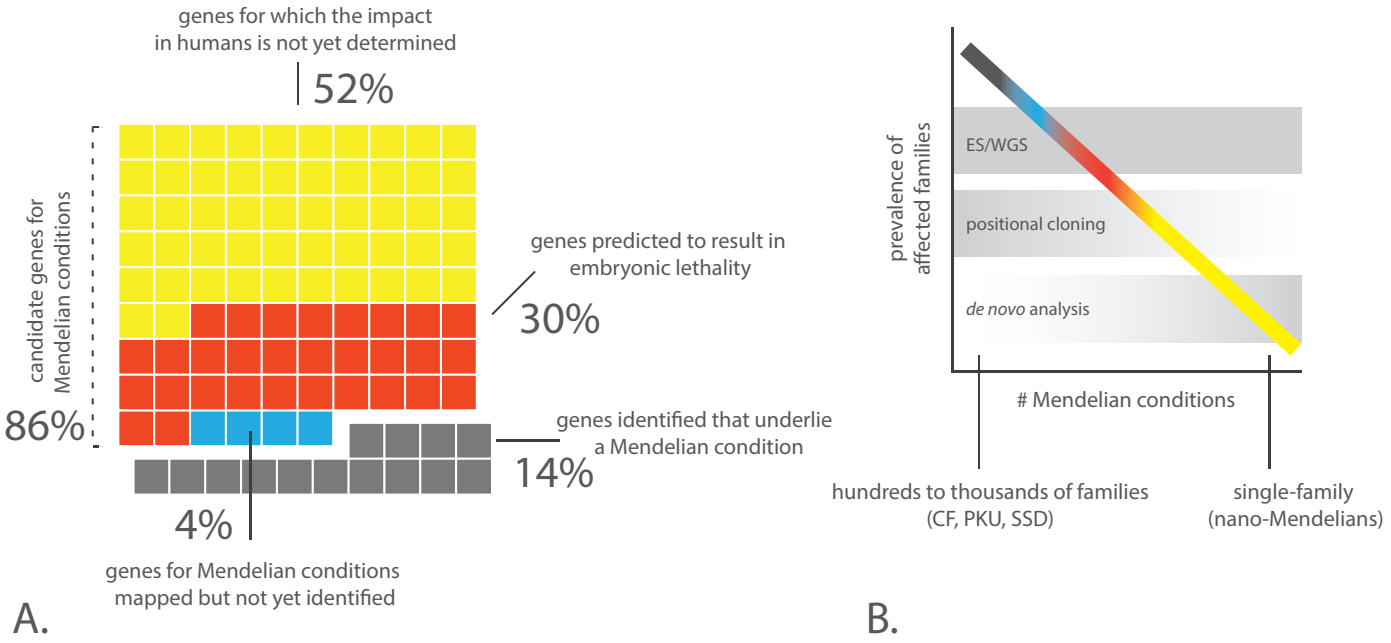


Figure 1. Relationship between human protein-coding genes and Mendelian conditions. A) Of approximately 18,910-20,345 protein-coding genes predicted to exist in the human genome, variants that cause MCs have been identified in ~2,776 (~14.2%; grey squares). Genes for ~783 MCs (~3.94%; blue squares) have been mapped, but not yet identified. Based on analysis of knockout mouse models, LOF variants in up to a maximum of ~30% of genes (~5,960; red squares) could result in embryonic lethality in humans. Note that the consequences of missense variants in these genes could be different. For a minimum of ~52% of genes (~10,330; yellow squares) the impact in humans has not yet been determined.

Collectively, ~17,090 genes remain candidates for MCs. **B)** MCs for which the underlying gene(s) have been identified (grey) typically occur in tens to thousands of families. MCs for which genes have been mapped but not identified (orange) often are less common. MCs caused by variants in the genes in which null alleles result in embryonic lethality in mouse models (red) or for which the impact in humans is unknown (yellow) are predicted to be very rare, found perhaps in only a single or few families (i.e., nano-Mendelians). Positional cloning (e.g., linkage mapping) approaches are most useful for identifying genes underlying MCs that are common whereas analysis of trios for *de novo* variants is perhaps the most effective strategy to identify causal variants for very rare MCs. Exome and/or genome sequencing can be exploited for gene discovery across the entire prevalence spectrum of MCs.

Relationship of Mendelian conditions to complex traits

Genes that cause monogenic subsets of otherwise complex phenotypes rarely explain much of the genetic variance of common diseases, but the identification of genes for MCs is often highly relevant to our understanding of the more general mechanisms of common, complex health-related traits. Indeed, nearly 20% of genes for MCs are found within GWAS signals, GWAS signals are enriched for genes that underlie MCs, and the fraction of genes that underlie a MC found in GWAS signals is positively correlated with the strength of association (**Figure 2**). Widespread co-morbidity among MCs and complex diseases suggests that rare, large effect variants in genes that underlie MCs play a role in complex disease. Importantly, the ultimate biological or therapeutic value of a variant, gene, or pathway discovered to underlie a MC may have little relation to the frequency of the MC. A classic example in cardiovascular disease research is the identification of the genetic basis of rare, monogenic forms of atherosclerotic disease that led to critical insights into the relevance of lipid transport for heart disease. In turn, these findings have led to the development of new therapies (e.g., statins) for common, complex cardiovascular diseases by targeting the implicated genes and pathways. Many additional drugs for common diseases that are based on gene discoveries for MCs are in advanced clinical trials (e.g., orexin antagonists for sleep, BACE inhibitors for Alzheimer's, PCSK9 monoclonal antibodies to lower LDL levels).

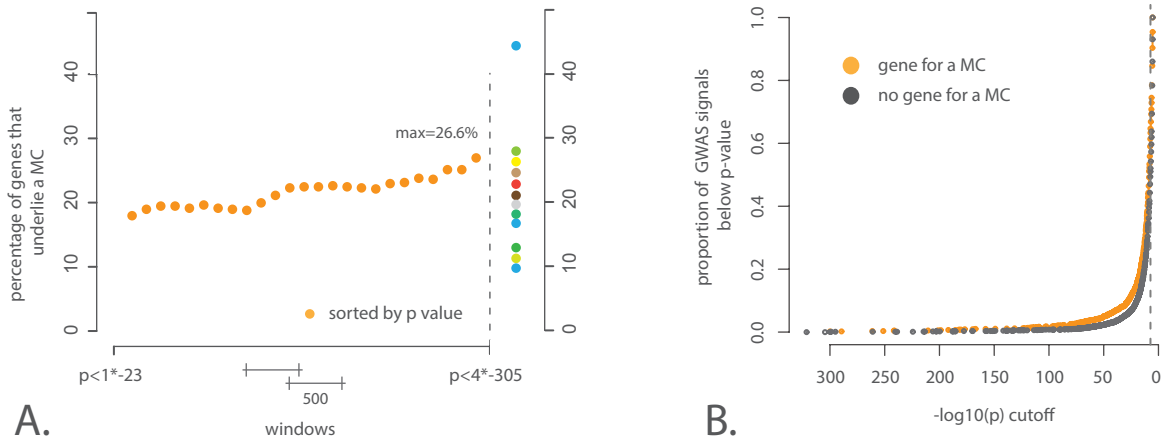


Figure 2. Relationship between GWAS signals and genes for Mendelian conditions. A) Plot of the fraction of genes underlying MCs found in GWAS signals sorted by p-value in windows of 500 (orange dots) and increments of 100. Approximately 26.6% of genes in GWAS signals with the top 500 lowest p-values underlie a MC. In contrast, only 14.2% of genes overall are known to underlie a MC suggesting that GWAS signals are more likely to be found in genes for MCs. Varied colored dots represent the percentage of genes underlying a MC in GWAS signals underlying different phenotypic categories and range from 13% for respiratory traits to 44% for musculoskeletal traits; mental health 17%; cardiovascular 27%; pharmacogenetics 28%. **B)** Cumulative plot of the proportion of GWAS signals in which a gene for a MC is found (orange dots) and GWAS signals in which a gene for a MC is not found (grey dots). At virtually every p-value, a higher proportion of GWAS signals overlap genes for MCs.

ESTABLISHMENT OF THE CENTERS FOR MENDELIAN GENOMICS (CMG)

Widespread, convenient, and cost-effective use of ES/WGS to find genes underlying MCs posed a number of major challenges when the technology was first introduced. Moreover, achieving the goal of solving most, if not all, MCs requires an unprecedented degree of cooperation and coordination among clinicians and scientists worldwide. Accordingly, efforts were initiated to establish the collaborative framework and physical infrastructure necessary to undertake large-scale gene discovery for MCs. The largest of these efforts were the NIH Centers for Mendelian Genomics (CMGs) and Finding Of Rare disease GEnes (FORGE) Canada, both established in 2011, and the Wellcome Trust Deciphering Developmental Disabilities (WTDDD). Other major efforts established to date at lesser scale and not coordinated nationally included programs in the Netherlands and China.

The CMG consists of three centers: (1) **University of Washington Center for Mendelian Genomics** (coordinating center)-PIs-Mike Bamshad, Deborah Nickerson, and Jay Shendure; (2) **Baylor-Hopkins Center for Mendelian Genomics**-PIs- Jim Lupski and David Valle; and the **Yale Center for Mendelian Genomics**-PIs-Mark Gerstein, Murat Gunel, Rick Lifton, and Shrikant Mane.

AIMS OF THE CMG

The CMG have two major aims, both of which require establishing substantial infrastructure and networks with clinicians and researchers worldwide as well as substantial sequencing scale: (1) to assess the genetic basis of ~1,000 MCs, most of which are very rare and have proved intractable to conventional gene discovery approaches and (2) to develop new technology and approaches for discovering the genetic basis of MCs and rare diseases in general and generate public resources that can be leveraged by the biomedical community at large to facilitate investigator-initiated gene discovery activities, studies of gene function, and the clinical translation / interpretation of the human genome; and to lead and coordinate U.S. efforts with international large-scale MC gene discovery projects.

MAJOR ACCOMPLISHMENTS AND ACTIVITIES OF THE CMG

Discovery

To date, 15,790 samples from 6,421 families representing 673 known and 760 novel MCs (621 novel phenotypes and 139 phenotype expansions (i.e., expansion of clinical features for a known MC)) have been assessed in partnership with 568 investigators from >255 institutions in 36 countries (i.e., 1 of every 5 countries in the world). Sixty percent of countries, 32% of institutions and 20% of investigators are located outside of North America, Europe, or Australia (**Figure 3**). Exome and whole genome data have been produced for 11,801 and 60 samples respectively, and about half of these sequences will be deposited in dbGaP without usage restrictions. Additionally, candidate gene data from all samples will be made available with a Mendelian browser, through ClinVar, and via a new track on the UCSC browser. Among 142 R01 applicants to NIH in 2013-2014 to use ES for rare disease gene discovery, 10% of awardees and 55% of non-awardees are collaborating with the CMG. **Accordingly, the CMG have broadly empowered the entire international rare-disease research community in the U.S. and worldwide.**

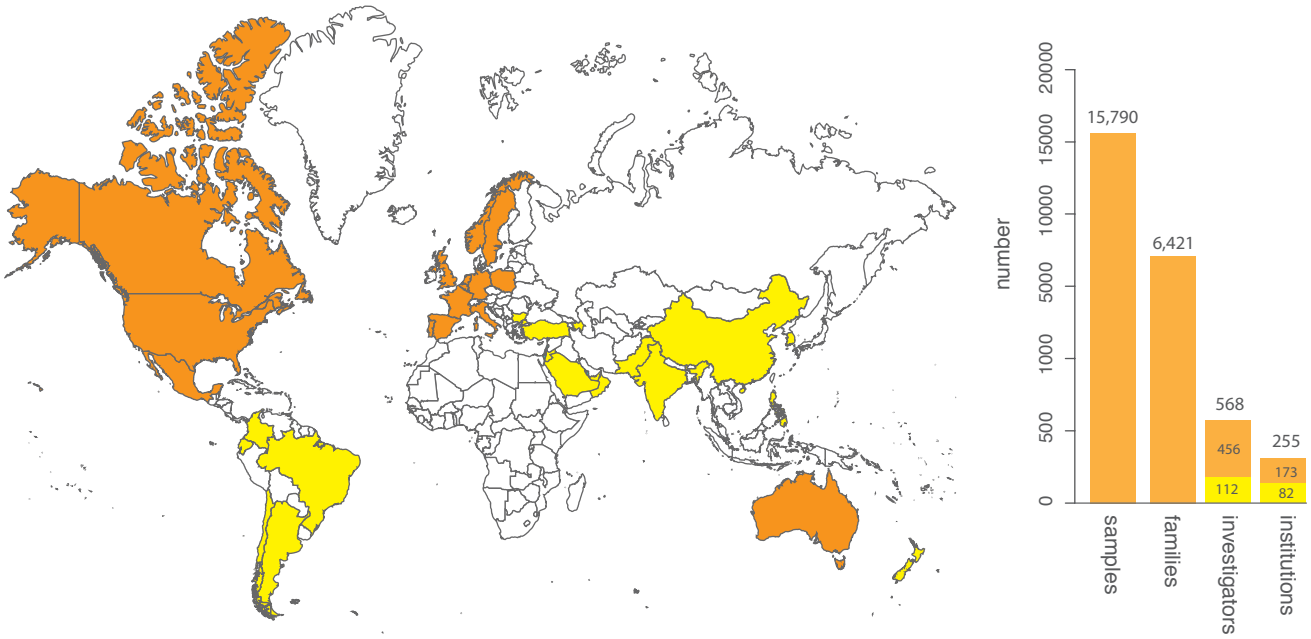


Figure 3. Worldwide interactions with the Centers for Mendelian Genomics (CMGs). The CMG have collected 15,790 samples, from 568 investigators representing 255 institutions in 36 countries or one of every five countries in the world. Approximately 60% (n=20) of these countries (yellow) are located outside of North America, Europe, or Australia (orange).

To assess the success of finding causal genes for MCs (i.e., solve rates) it is critical to apply objective metrics. To date, reported solve rates are hard to interpret or to use for comparisons across different contexts (i.e., clinical service

vs. research). Because of its simplicity, "solve rate" has been reported as the proportion of families in which a causal variant for a MC is identified, or, alternatively, as the proportion of phenotypes for which the gene is identified. The former definition isn't particularly useful on its own as one could have a high solve rate by only sequencing families diagnosed with disorders with a single known gene, while the latter has multiple interpretations because the same phenotype can be caused by variants in multiple genes, variants in a single gene can cause multiple phenotypes, and causal variants often cannot be identified in all families.

We developed three complementary metrics for "solve rate" and applied them to the CMG discovery projects using strict criteria for variant causality and clear definitions of novel vs. known phenotype and novel gene for a MC. The overall diagnostic rate, defined as the proportion of families for which a causal variant was identified, was 0.26. This is comparable to diagnostic rates achieved by clinical ES even though the conditions studied by the CMG are biased against phenotypes expected to be explained by a gene known to underlie a MC. Overall, a total of 515 genes were, to date, discovered to underlie a MC. Of these 515 discoveries, 286 were for either 1) a gene not previously known to underlie a MC that was found to underlie a known MC or a novel MC, 2) a gene previously known to cause a MC that explained a different known MC or novel MC, or 3) a gene found to underlie a phenotype not previously described for the MCs the gene explained and consequently expanding the the clinical features of the MC (i.e., phenotype expansion). Two-hundred and twenty nine discoveries were for a gene previously known to underlie the MC studied; the vast majority of these MCs were categorically-defined MCs (e.g., non-syndromic hearing loss, asphyxiating thoracic dysplasia, oculocutaneous albinism) that exhibit high genetic heterogeneity. Accordingly, the solve rate, defined as the ratio of causal genes discovered to MCs studied, is 0.58 genes / MC studied. Analysis of solve rates by mode of inheritance used to model segregation in a subset of 200 MCs provides further resolution about the types of MCs that have been solved and the challenges that remain. Solve rates ranged from 0.52 (dominant) to 1.0 (X-linked) (**Figure 4A**); for comparison, if a causal gene was identified for every phenotype, this ratio would approach its maximum value of one. Gene discoveries for consanguineous families were frequently complicated by locus heterogeneity and extreme rarity consistent with a lower than anticipated solve rate (0.81). Lastly, the novel discovery rate, or proportion of MCs solved in which the gene was newly discovered to underlie a MC or the condition itself was novel/unexplained was 0.49 (**Figure 4B**).

The criteria that we used to estimate solve rates are conservative, in large part because of the requirement that either causal variants in the same gene had to be identified in two or more families with the same MC or if a putatively causal variant was found in only one family, both high-confidence mapping and compelling functional data (e.g., recapitulation of the phenotype in an animal model) were required. Given that many of the MCs studied are very rare and additional families are not yet available for study, the net effect was a 1.4-fold reduction in the number of MCs we considered solved. Yet, our results in these families portend genuine gene discoveries for MCs. Accordingly, we also calculated solve rates with relaxed criteria. Overall, genes for a total of 704 genes for MCs were discovered. Of these 704 discoveries, 457 were for either 1) a gene not previously known to underlie a MC that was found to underlie a known MC or a novel MC, 2) a gene previously known to cause a MC that explained a different known MC or novel MC, or 3) a gene found to underlie a phenotype not previously described for the MCs the gene explained and consequently expanding the the clinical features of the MC (i.e., phenotype expansion). Two hundred and forty-seven discoveries were for a known gene and MC (i.e., a gene previously known to underlie the MC studied). The overall solve rate was 0.65 genes / MC studied.

CMG-facilitated discoveries have been reported in 95 publications over the past two and a half years.

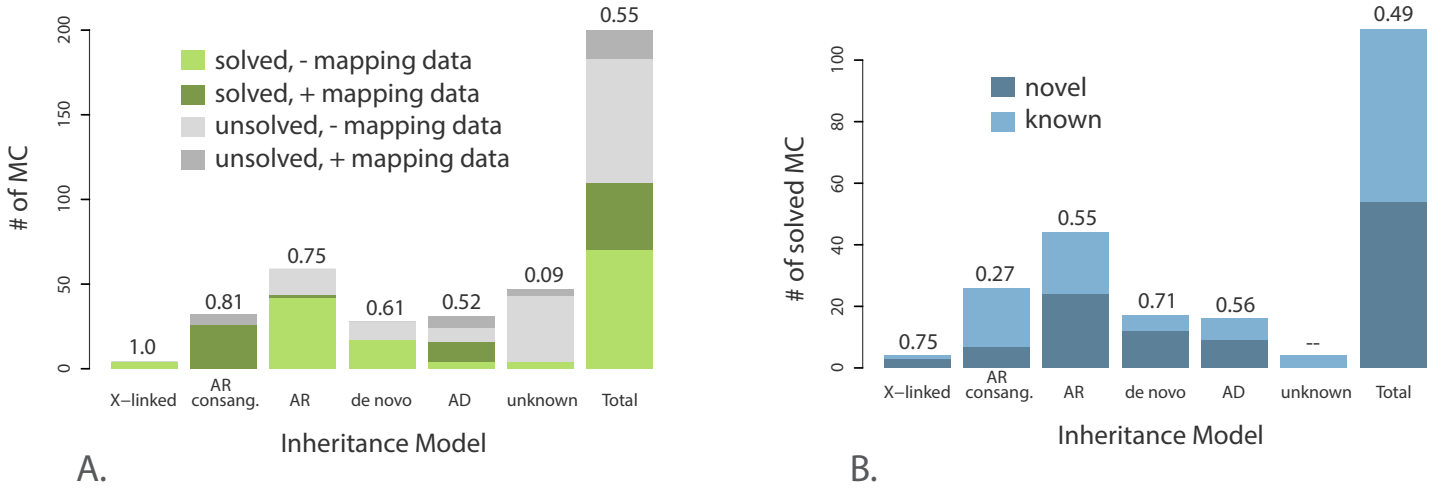


Figure 4. Solve rates (i.e., gene discovery for Mendelian conditions) under different models of inheritance for a subset of MCs studied by the CMG. A) The mean number of genes discovered per MC studied is 0.55 (110/200). If MCs with an unknown mode of inheritance are excluded, the mean rises to 0.69 (106/153). **B)** Of the genes for MCs found, 0.49 are novel (i.e., not previously found to underlie a MC).

Impact of CMG accomplishments on clinical care and the scientific community

The translation and impact of CMG efforts on clinical care has been immediate and substantial. Approximately 621 novel MCs have been delineated and the clinical features of 139 known MCs have been expanded (i.e., phenotypic expansion). The net effect of delineating new MCs, identifying genes for MCs, and phenotypic expansion of MCs will be a much richer understanding of the genetic and allelic architecture of MCs, an ability to distinguish among similar phenotypes, and improved ability to diagnose disease. Collectively, this is already having dramatic clinical implications. For example, in the past two years the CMGs have discovered four new genes underlying distal arthrogryposis syndromes, defined two new MCs classified as distal arthrogryposis caused by genes previously associated with a MC, and expanded the phenotypes associated with variants in the gene, *TNNI2*, known to cause distal arthrogryposis (**Figure 5**). Accordingly, the diagnostic rate for families with distal arthrogryposis syndromes has substantially increased and three novel mechanisms, quite different from those previously known to cause distal arthrogryposis, have been discovered.

Across CMG discoveries, it is commonplace for variants in the same gene to cause very different phenotypic effects—a gene property for which we suggest the term, phenotropy, specifically defined as the spectrum of phenotypes associated with variants in a given gene. This raises the interesting question: what is the underlying biological basis for the number of phenotypes that can be produced by variation in a particular gene? **One of the major long term deliverables of gene discovery efforts for MCs at scale will be the addition of genes underlying at least several thousand known MCs, an unknown but predicted several thousand yet-to-be delineated MCs, and a catalogue of phenotropy for many, if not most, human genes.**

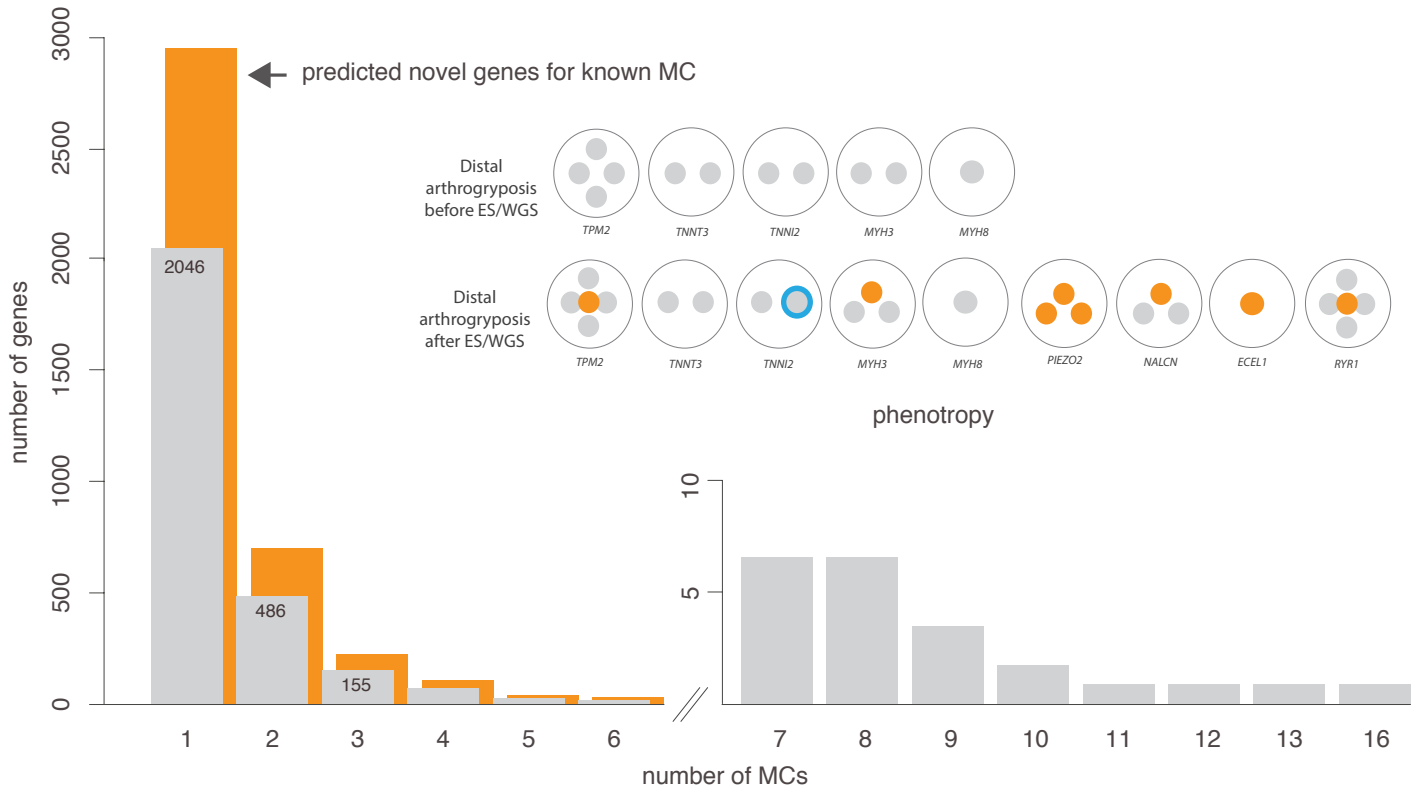


Figure 5. Plot of the number of number of MCs caused by each gene known to cause one or more MC (grey bars / circles). Approximately 26% of genes for MCs cause two or more MCs up to a maximum of 16 MCs caused by variants in the same gene (*COL2A1*). Orange bars represent predicted novel genes to be discovered for known MCs. Inset, conceptual construct of the spectrum of different MCs caused by variants in the same gene or phenotropy for genes (large open circles) and known phenotypes (grey circles) broadly classified as distal arthrogryposis syndromes. ES/WGS has enabled discovery of new genes (*PIEZO2*, *NALCN*) not previously associated with a MC (orange circles), new MCs (orange circles) in genes (*ECEL1*, *RYR1*) previously associated with a different MC (grey circles), and expansion of known phenotypes (blue border).

CMG discoveries have had a considerable impact on overall clinical diagnostic rates for MCs (i.e., the fraction of cases of a MC in which a causal variant can be identified). Diagnostic rates for MCs are in general, poor. For example, of 292 well characterized MCs for which clinical testing is available per GeneReviews (July, 2014), a causal variant can be identified only in 52% of cases overall (**Figure 6**). The diagnostic rate varies widely per MC and is inversely correlated with the level of genetic heterogeneity (**Figure 6 inset**). These observations are consistent with diagnostic rates of 25-75% for many major categories of inherited conditions (e.g., kidney disease, cardiomyopathy, seizures, etc.) and partly explain the diagnostic rate of ~20-30% reported by clinical ES providers. Based on discoveries by the CMGs and others, the diagnostic rate of ES is increasing (i.e., from 25-28% at Baylor between 2012 and 2014). Interestingly, variants in genes identified since 2012 as underlying MCs represent ~25% of positive findings in these clinical diagnostic efforts. This result emphasizes the ongoing importance of disease gene discovery. **Together, a catalogue of MCs, phenotropy, and of causative variants or MENDELOME will be the cornerstone of the MEDICAL GENOME.**

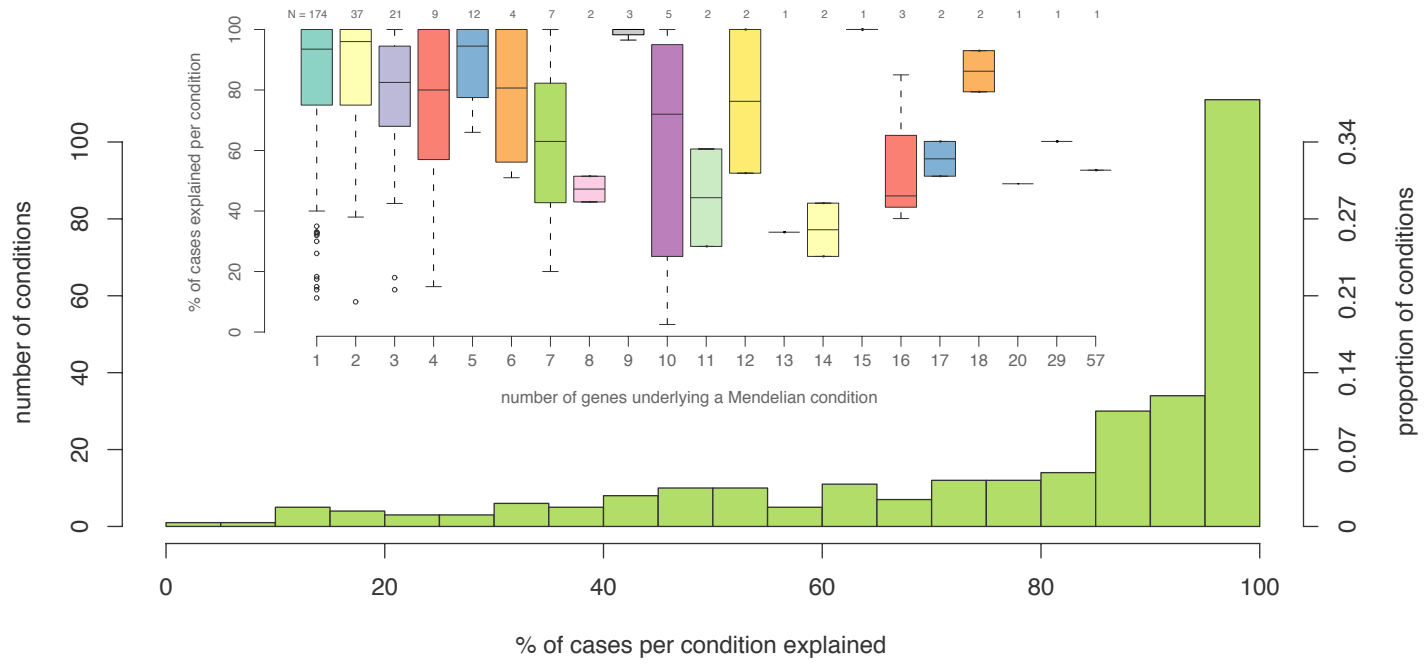


Figure 6. Clinical diagnostic rates of Mendelian conditions for which the gene(s) has been identified. Histogram of the number / proportion of MCs (Y-axis) vs. the percentage of cases explained per condition (X-axis) for 292 MCs in GeneReviews for which one or more causal genes have been identified. Collectively, of 292 MCs, a causal variant can be identified in only ~52% of cases overall. Inset, boxplot of percentage of cases in which a causal variant can be identified (i.e., diagnostic rate) per condition (Y-axis) for which one or more causal genes (X-axis) have been identified. The diagnostic rate per condition is inversely correlated with the level of genetic heterogeneity.

Development of new therapeutics to address common diseases that constitute major public health problems is hamstrung by ignorance of the fundamental biology underlying disease pathogenesis. As a consequence, 90% of drugs entering human clinical trials fail, commonly due to lack of efficacy and/or unanticipated mechanism-based adverse effects. **Families with rare MCs segregating mutations that have large effects to promote or protect from common disease in humans can directly establish the causal relationship of genes and pathways to common diseases and identify targets likely to have large beneficial effects and few mechanism-based adverse effects when manipulated.** For example, Mendelian forms of high and low blood pressure are due to mutations that cause respective increases or decreases in renal salt reabsorption and net salt balance; these have identified promising new targets such as the K+ channel ROMK that are now in clinical trials and have provided the scientific basis for public health efforts in more than 30 countries to reduce heart attacks, strokes, and mortality by modest reduction in dietary salt intake.

Use of other approaches, such as identification of common variants with small effects, is less effective at facilitating drug development. Of ~348 proteins linked to a human gene and specifically targeted by current therapeutics, 42.5% encode a gene underlying a MC whereas only 28.2% of proteins targeted by current therapeutics are found within GWAS signals (closest downstream and upstream genes counted per intergenic signal and all overlapping genes counted per coding signal). Accounting for the over-representation of MC genes in GWAS signals, 27.3% of proteins targeted are only encoded by a gene for a MC, while 13.6% of proteins targeted are found only in a GWAS signal. Moreover, the

fraction of therapeutics approved (42.5%) vs. in clinical trials (32.8%) that encodes a gene for a MC is higher suggesting that drug targets associated with an underlying gene for a MC more often receive FDA approval. No such relationship is observed for genes found within GWAS signals (28.2% FDA approved vs. 29.4% in clinical trials). Accordingly, use of information about whether a target protein is encoded by a gene underlying a MC is a potentially powerful way to stratify drugs candidates for development. To this end, CMG discoveries have added several hundred putative starting points for the development of targeted therapeutics.

Beyond disease gene discovery, the CMG have also developed and disseminated both methods and tools to: (1) enable large-scale inventory and indexing of MCs (PhenoDB), (2) facilitate candidate gene comparisons among investigators (GeneMatcher), (3) reconstruct extended pedigree relationships from genotype data (PRIMUS), (4) perform quality control of variant data (VAT), and (5) make computational predictions of the impact of sequence variants including non-coding regions (ALOFT, CADD, FunSeq).

SUMMARY

Achieving NHGRI's goal of understanding the genetic basis of inherited disease will ultimately require discovering the gene underlying virtually every MC. While great progress has been made toward identifying the genetic basis of MCs, the genes underlying more than half of all known MCs (i.e., 3,703) have not yet been discovered despite ~19% (696) of these MCs being mapped (~80% with robust linkage data). An even greater number of MCs are predicted to exist, and to this point, over each of the past several years, ~300 new MCs have been delineated per review of the biomedical literature by OMIM. Most of these "unsolved" MCs are very rare, exhibit high locus heterogeneity, and / or are caused by *de novo* mutations of the germline or somatic cells. Sequencing technologies and analytical approaches, developed in part by the CMGs, have now matured to the point to make highly successful gene discovery at scale to solve all MCs tractable and cost-effective. To this end, NHGRI and the CMGs are uniquely positioned to coordinate and stimulate the burgeoning international effort to identify the genetic basis of all MCs even as the number recognized increases each year.

The resultant MEDICAL GENOME generated, in large part, by the discovery of genes for all MCs will be the foundation of clinical genomic medicine of the future. As of today and for the foreseeable future, the vast majority of genetic diagnostic tests, primary and incidental results returned to families, and results that guide clinical management in children and adults are based on discoveries of genes underlying MCs. The rapidly growing catalogue of phenotypes and causative variants for MCs immediately: (1) makes it possible for clinical laboratories to rapidly improve diagnostic testing for rare diseases; (2) advances the development and re-purposing of better therapeutics; and (3) empowers clinicians to improve the care and management of individuals with a very wide spectrum of conditions. Accordingly, solving most, if not all MCs, should transform the care of families with rare diseases.

Of broader interest to the biomedical research community, discovery of genes for all MCs will: (1) inform the scientific community about gene function; (2) provide a pathway for developing better strategies to find genetic and environmental modifiers underlying common diseases; and (3) enable understanding of the functional and phenotypic consequences of non-coding variation. Indeed, it is becoming increasingly clear that, in the absence of deep phenotypic characterization, genomic information even from a large number of individuals is of limited value. Moreover, what can be understood about gene function and phenotypic consequences is proportional to the breadth and diversity of conditions studied. **To this end, it should be expected that the full power of human genomics to explain and understand fundamental biological processes and transform medical care in general will be more readily realized by discovering the genetic basis of all MCs rather than large-scale sequencing of a handful of common diseases.**

FUTURE DIRECTIONS

Over the next five years, there will be a growing need for (1) **keystone projects** (e.g., delineate and deeply phenotype all MCs, catalogue all validated large-effect variants underlying MCs, etc.) to empower the biomedical community and (2) **large-scale rare disease genomics projects** to understand the genomic basis of all MCs and rare diseases. Together these projects will generate a complete catalogue of the phenotypic characteristics of all rare diseases including all MCs and the relationships that distinguish them; identify, index, and warehouse all causal variants linked to all MCs and the range of different phenotypes observed within each individual MC (**MENDELOME**); and pioneer new approaches and analytical methods that can be exploited at scale to discover genes underlying MCs and rare disorders caused by non-traditional modes of inheritance (e.g., oligogenic, *de novo*, somatic mosaicism, uniparental disomy, etc.) and understand their overall contribution to the burden of human disease.