Next-generation sequencing in Charcot–Marie–Tooth disease: opportunities and challenges

Menelaos Pipis, Alexander M. Rossor, Matilde Laura and Mary M. Reilly*

MRC Centre for Neuromuscular Diseases, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, UK.

*e-mail: m.reilly@ucl.ac.uk

Charcot–Marie–Tooth disease and the related disorders hereditary motor neuropathy and hereditary sensory neuropathy, collectively termed CMT, are the commonest group of inherited neuromuscular diseases, and they exhibit wide phenotypic and genetic heterogeneity. CMT is usually characterized by distal muscle atrophy, often with foot deformity, weakness and sensory loss. In the past decade, next-generation sequencing (NGS) technologies have revolutionized genomic medicine, and as these technologies are being applied to clinical practice, they are changing our diagnostic approach to CMT. In this Review, we discuss the application of NGS technologies, including disease-specific gene panels, whole-exome sequencing (WES), whole-genome sequencing (WGS), mitochondrial sequencing and high-throughput transcriptome sequencing, to the diagnosis of CMT. We discuss the growing challenge of variant interpretation and consider how the clinical phenotype can be combined with genetic, bioinformatic and functional evidence to assess the pathogenicity of genetic variants in patients with CMT. WGS has several advantages over the other techniques that we discuss, which include unparalleled coverage of coding, non-coding and intergenic areas of both nuclear and mitochondrial genomes, the ability to identify

structural variants and the opportunity to perform genome-wide dense homozygosity mapping. We propose an algorithm for incorporating WGS into the CMT diagnostic pathway.

Introduction

Charcot–Marie–Tooth disease (CMT) encompasses a group of inherited peripheral neuropathies that are heterogeneous in terms of their phenotypic features, modes of inheritance and causative gene mutations. Epidemiological studies in Western Norway in the late 1960s estimated a population prevalence of 1:2,500 for CMT[1]. However, more recent population-based studies, using a range of different methodologies, have reported minimum prevalence ranging from 1:5,000 to 1:10,000 across different, mainly European populations[2]. A higher prevalence of CMT was recorded in Norway (1:1,250) than in the other populations that were studied.

The 'classic' CMT phenotype is typically a length-dependent motor and sensory neuropathy characterized by distal weakness, sensory loss and a high incidence of foot deformities such as pes cavus. Most patients with the classic phenotype present with symptoms in the first or second decade of life and both the weakness and sensory loss slowly progress and worsen over time[3]. However, certain genetic subtypes such as CMT2A, which is caused by mutations in *MFN2* (Supplementary Table 1) can present in infancy or early childhood with a more severe phenotype than classic CMT, resulting in patients becoming non-ambulant by the third decade of life[4].

During the past century, advances in neurophysiology and neuropathology have allowed the classification of patients with CMT in two broad groups: CMT1, in which the pathology is primarily demyelinating with upper limb motor nerve conduction velocities (MNCVs) of <38m s^{-1} and CMT2, in which the pathology is predominantly

axonal with MNCVs of >38m s^{-1} (Supplementary Table 1)[5]. The term 'intermediate CMT' is also used when the upper limb MNCVs are between 25 and 45 m s^{-1}. In recent years, the 'axonal CMT' group has been expanded to include not only CMT2, which has both motor and sensory involvement, but also motor-predominant and sensory-predominant forms of CMT (hereditary motor neuropathy (HMN) and hereditary sensory neuropathy (HSN), respectively) (FIG. 1). Further sub-classification of all types of CMT is based on inheritance pattern and genetic cause (FIG. 1; Supplementary Table 1). CMT, HMN and HSN are often collectively referred to as 'CMT and related disorders', but throughout this Review, we use CMT to refer to all three conditions.

The classification of rare diseases, such as CMT, has long been dominated by clinical evaluation of a phenotype, but advances in molecular genetics, often driven by next-generation sequencing (NGS) technologies, have brought new options for disease re-classification[6]. A phenotype-centric classification that also incorporates neurophysiological features (such as slow or normal MNCVs) is important for the diagnosis of CMT, not only because phenotyping precedes genotyping, but also because phenotypic information can be used to guide the choice of gene panels and the interpretation of NGS data. Nonetheless, when a gene containing a causative mutation is identified, the CMT classification system must include this new information, as treatments currently in development are mostly gene-specific.

In this Review, we discuss how knowledge of the genetic heterogeneity of CMT is expanding, partly owing to the increased use of NGS, and how this new information is shaping our diagnostic approach. We also examine how advances in NGS technologies are being translated into clinical practice for patients with CMT, highlight the accompanying challenge of data interpretation, and suggest practical guidelines

for contemporary CMT diagnostics. In addition to CMT, around 150 complex or multisystem inherited neurological disorders are characterized by neuropathy (for example, in association with cerebellar ataxia and pyramidal features in Friedreich ataxia), but this group of inherited peripheral neuropathies is beyond the remit of this Review[7].

CMT genetics and diagnostic challenges

Advances in genetic diagnosis, most notably the widespread use of NGS, have revealed the vast genetic heterogeneity of CMT[8], with >100 genes containing causative mutations described to date (FIG. 1; Supplementary Table 1)[9]. The commonest subtype of CMT is CMT1A, which accounts for >60% of genetically diagnosed cases of CMT. CMT1A is caused by a 1.4 Mb duplication in the short arm of chromosome 17 (17p), which contains 9 genes, including the peripheral myelin protein 22 gene (*PMP22*). A deletion of the same 1.4 Mb section causes another inherited neuropathy, hereditary neuropathy with pressure palsies (HNPP), highlighting the importance of the PMP22 protein expression level for peripheral nerve function. Other common CMT-associated genes include *GJB1* (CMTX1), *MFN2* (CMT2A) and *MPZ* (CMT1B). Three independent population studies, each assessing >700 patients with CMT, showed that collectively these five subtypes account for >90% of genetically diagnosed CMT cases[8,10,11]. However, only ~60% of the total number of patients who are given a clinical diagnosis of CMT also receive a genetic diagnosis[8,10–12]. Patients with CMT1 are the most likely to receive a genetic diagnosis (>85% of patients), followed by those with HSN (30–40%), CMT2 (25–35%) and HMN (15–25%)[8,10,12].

It is important to highlight that the cohorts used in these large population studies came from a heterogeneous population with mainly European ancestry. The epidemiology described above would not necessarily be mirrored in other ethnic populations, most notably those from the African continent, which remain understudied[13] and are underrepresented in large population databases[14]. Furthermore, the genetic epidemiology of CMT, like that of any rare disease, is skewed in bottlenecked populations [G], genetically isolated ethnic groups with high rates of endogamy [G] and ethnic groups with founder mutations [G]. For example, in the bottlenecked Finnish population[15], a founder mutation in *GDAP1* is more prevalent than CMT1A in some areas[16]. In the European Roma ethnic minority, private founder mutations in *NDRG1* (c.631C>T, Arg148X) and *HK1* (g.9712G>C) account for most cases of autosomal recessive CMT4, which in turn accounts for 30–50% of all cases of CMT in the Roma population[17]. These special circumstances should be taken into consideration when reviewing patients from specific ethnic backgrounds, and the diagnostic approach should be tailored accordingly, as is discussed later in this Review.

This genetic heterogeneity within and between different populations, along with the inter-familial and intra-familial phenotypic heterogeneity that many variants also exhibit, is one of the challenges to accurate diagnosis of CMT. Aside from the increased incidence of autosomal recessive forms of CMT in populations and ethnic groups with high rates of consanguinity [G], as discussed above, so-called 'sporadic' cases, in which the inheritance pattern is difficult to define, are seen in some small families, particularly in the UK. Awareness is also increasing that early-onset forms of CMT can be either *de novo* dominant or autosomal recessive, and that late-onset forms, which were traditionally thought to be autosomal dominant (perhaps with

reduced penetrance **[G]**), can be autosomal recessive (for example, recessive mutations in *MME* cause late-onset CMT2)[18,19]. In addition to the above difficulties, the main challenge in the diagnosis of CMT is to determine whether identified genetic variants are pathogenic. This problem is common in the diagnosis of inherited conditions, and in the case of CMT is made more difficult by the lack of easily applicable functional assays for the identified genes.

Opportunities of NGS in CMT

NGS, also referred to as massively parallel or high-throughput sequencing, is the collective term that is used to describe the different technologies that allow simultaneous genetic sequencing of large amounts of DNA: the whole genome, the whole exome (covering only protein-coding sequences) or only specific targeted genes (for example, a CMT gene panel)[12].

In Europe and North America, and increasingly in other regions, NGS has overtaken targeted Sanger sequencing **[G]** as the method of choice for genetic diagnosis in patients with CMT. As NGS can sequence billions of DNA fragments simultaneously, it substantially reduces the cost and time taken to sequence genetic material compared with Sanger sequencing methods. This allows the quick and relatively cheap sequencing of multiple loci in the genome, which is necessary because of the genetic heterogeneity of CMT. NGS also minimizes the need for fragment cloning during gene hunting studies. Targeted gene panels constitute the most commonly used NGS approach in clinical CMT practice, with whole-exome sequencing (WES) and whole-genome sequencing (WGS) being used less frequently. The only exception is for patients with autosomal dominant CMT1, in whom techniques such as multiple ligation-dependent probe amplification **[G]** (MLPA) continue to be the first-line

investigation to diagnose the chromosome 17p duplication that causes CMT1A[8], although it is perhaps only a matter of time before WGS can reliably detect such structural variants[20].

Targeted gene panels

Gene-specific panels employ capture kits that only target and sequence the regions of the genome that contain known CMT-associated genes. These panels can be limited to specific types of CMT (for example, demyelinating or axonal forms[12]) with each panel containing the 10–15 genes most commonly associated with that subtype, although clinical diagnostic laboratories are now moving towards the use of a single unified CMT panel containing 56–197 genes[21–25]. This change of approach has mainly been prompted by the substantial intergenic and intragenic phenotypic heterogeneity of CMT, whereby mutations in different genes can independently cause the same phenotype. In cases of autosomal dominant demyelinating CMT, our current practice is to screen for the 17p duplication using MLPA before proceeding to a CMT targeted gene panel.

Phenotypic overlap between axonal CMT and ataxias, distal myopathies and hereditary spastic paraplegias (HSPs) is increasingly recognized. For example, mutations in several genes that have been linked to HSP, including *BSCL2*, *REEP1* and *SPG11,* can manifest as a pure peripheral neuropathy that is initially diagnosed as CMT or distal HMN (dHMN); the upper motor neuron signs that would suggest a diagnosis of HSP may or may not emerge later in the disease course. Similarly, some genetic ataxias, such as leukoencephalopathy with brainstem and spinal cord involvement and lactate elevation (LBSL), which is caused by recessive mutations in *DARS2,* can present with a neuropathy before the development of cerebellar signs (or

might present with only minor cerebellar signs). Therefore, if the results from the CMT gene panel are negative, we might proceed to testing a HSP or ataxia gene panel depending on the phenotype. Finally, differentiating a distal myopathy from a distal motor neuropathy can be challenging as the two pathologies can coexist, for example, in the case of *HSPB1* or *HSPB8* mutations. Therefore, we would advocate testing for selected myopathy-associated genes, including *HSPB1*, *HSPB8*, *BAG3*, *LAMA2*, *LMNA*, *MTTK*, *ETFDH* and *CNTNAP1*, in patients with dHMN and associated myopathic features.

Even with larger unifying gene panels, this targeted approach minimizes background genetic noise and provides a finite number of genetic variants for interpretation. Gene panels provide excellent capture of the intended CMT-associated gene regions, thus minimizing false negatives, with persistently uniform coverage and high read depths **[G]**[21–23,25]. The diagnostic hit rate for CMT gene panels ranges from 18–31% depending on the CMT cohort, its demographic background, the sequencing platform and the number of genes included[21–25].

Whole-exome sequencing

In WES, gene-specific oligonucleotide baits capture the majority of the coding regions of the genome, which are subsequently amplified and sequenced. Genetic variation that is predicted to be protein-altering is then identified. In the past decade, WES has been a very popular research tool and has been the main driver for identifying novel CMT-associated genes, as it enables the sequencing of genes that have never been associated with CMT or, indeed, other Mendelian diseases. A shorter version of WES, referred to as clinical exome or mini-exome sequencing, sequences the ~5,000 genes that are currently associated with monogenic Mendelian disease. Both WES and

clinical exome sequencing produce short-read data, and as WES only captures and sequences 1–2% of the entire genome, it has a high average read depth[26] and high resolution for detecting single nucleotide variants (SNVs). Each WES returns ~90,000 SNVs and small insertion or deletion mutations (indels), which require interpretation to determine their relevance. Even though high-quality WES can capture and sequence ~96% of the exome, the inability to offer complete coverage of protein-coding regions and the accompanying unacceptable false-negative rate have limited its widespread use as a first-tier test in clinical settings.

In the CMT field, WES is a valuable research tool, with independent groups reporting diagnostic rates of 19–45% in individuals with CMT or complex neuropathy for whom previous genetic testing was negative[27–30]; this diagnostic yield is similar to that achieved with WES in other Mendelian disorders[31]. The diagnostic yield of WES can reach 45% when combined with phenotype-enriched analysis for certain subtypes of CMT, such as motor-predominant CMT2 or dHMN[32].

Aside from the identification of SNVs and indels, WES data have aided the identification of other forms of genetic variation in patients with CMT. Structural variants (SVs; defined as genomic variants >50 bp) consist of two broad types: balanced or copy-neutral rearrangements, such as inversions and translocations, and unbalanced or copy number variants (CNVs), such as duplications, deletions and repeat element insertions. The high average read depth achieved by WES, coupled with a computational analysis of read depth deviation, can identify CNVs, including the 17p rearrangement in CMT1A and HNPP[33].

Whole-genome sequencing

Since the commercialization of NGS technologies and the integration of gene panels into diagnostic services, the idea that WGS could provide a single genetic test for CMT, and inherited conditions in general, has been appealing. During WGS, the entire genome is fragmented and sequenced, thus enabling near-complete coverage of the protein-coding genome (98.4–100% coverage of NCBI reference sequence database genes depending on read depth), unlike WES (up to 96%), despite the latter having a higher average read depth[26,34,35]. Furthermore, because WGS covers almost the complete genome, the technique is able to detect SVs with breakpoints outside exonic regions, such as those in intronic or intergenic areas, which is not possible with other methods[20].

A small exploratory study of WGS by two commonly used sequencing platforms in 12 individuals[36] initially led to concerns about the incomplete coverage of 10–19% of disease-associated genes. However, subsequently many independent groups have successfully used WGS to diagnose rare Mendelian diseases[20,37–39] including CMT[40–42], and some have advocated its use as a first-tier genetic test[43,44].

WGS on a single germline sample will produce ~5,000,000 variants, of which 30,000 are in coding regions; therefore, robust filtering strategies are necessary for the efficient interpretation of this data. The application of WES or WGS trios, which include sequencing data from the affected proband and unaffected parents, as opposed to proband-only testing, allows more stringent filtering of variants and leads to a tenfold reduction in the number of flagged variants that require interpretation[45]. Therefore, parental phenotyping, when available, is critical in WES or WGS trio-based studies, especially as the presence of subclinical peripheral neuropathy in a parent would

change the inheritance pattern and, hence, bias variant analysis. Furthermore, trio-based studies allow the identification of *de novo* variants, which carry the highest predictive value and diagnostic yield, and in cases of compound heterozygous variants enable the determination of phase **[G]** (*cis* or *trans*) if the variants are present in the parental samples.

WGS is beginning to provide insight into the mutational burden of complex SVs in CMT, which, beyond the common 17p rearrangement, remain largely unknown. For example, through the use of traditional single nucleotide polymorphism (SNP) linkage analysis followed by WGS, a novel interchromosomal translocation (78 kb duplication) was identified in a large family with X-linked CMT (CMTX3)[41] and an intrachromosomal inverse translocation (1.35 Mb duplication) was identified in a family with autosomal dominant dHMN[42]. SVs are common in the healthy population, and identification of disease-causing SVs in the absence of segregation data from large families with multiple affected members is challenging. Nevertheless, with the establishment of large disease-specific databases of WGS data, such as the CMT cohort in the UK-wide 100,000 Genomes Project (which uses PCR-free WGS — see below)[46] and the GENESIS CMT platform hosted by the University of Miami[47], future disease enrichment analyses of WGS data are likely to result in the discovery of an increasing number of disease-causing SVs in patients with CMT.

PCR-free WGS, in which no PCR-based amplification of genomic DNA is performed during library preparation, has an average read depth less than half that of WES or PCR-based WGS, but does offer specific advantages. First, the lack of PCR amplification reduces the random insertion of errors by DNA polymerase enzymes over long sequence reads. This feature is important, as during variant calling **[G]** these errors would be flagged as false positives. Second, the technique helps to overcome

the low coverage of GC-rich regions by WES and PCR-based WGS, which results from poor PCR amplification. Therefore, PCR-free WGS provides more complete coverage of genomic areas with high GC-content, which include the first coding exons of many CMT-related genes[26]. Third, PCR-free WGS can detect copy-neutral SVs such as inversions and balanced translocations, as the ratio of genomic material is accurately preserved during amplification and sequencing.

Genome-wide association studies (GWAS) apply high-throughput technologies to genotype common and rare SNPs across the entire genome of selected case and control individuals. These studies then test for statistically significant differences in the frequency of one or more sets of alleles between the case and control populations. GWAS are primarily used to identify loci that predispose to specific human diseases and in this capacity are also used in the quest for genetic modifiers. Similar to other monogenic Mendelian conditions, CMT1A shows considerable phenotypic variability between affected individuals, and the concept of genetic modifiers that can affect clinical phenotypes through changes in the penetrance, dominance, expressivity and pleiotropy [G] of Mendelian traits (in this case, the 17p duplication) is being increasingly considered to help explain this variability[48]. In simple terms, an allele that resides at an independent genetic locus and is not pathogenic for the disease might modify the phenotype caused by the pathogenic variant. This phenomenon could lead to variation in age of onset, severity and pace of progression of the disease.

Identifying genetic modifiers is challenging as it often requires large sample sizes, but this problem can be overcome with the use of extreme phenotype sampling (EPS) — a technique that focuses on alleles that tend to associate with one phenotypic extreme — rather than random sampling[49]. This approach was used in a well-designed genome-wide genotyping study on a large sample of 330 deeply phenotyped patients

with CMT1A who were of European ancestry[50]. This study identified the *SIPA1L2* gene as a genetic modifier of ankle dorsiflexion weakness, a hallmark feature of CMT. Furthermore, the study highlighted the importance of multicentre collaboration for the acquisition of genomic data, as well as the necessity for standardized phenotyping in the context of powering of sample sizes in rare disease association studies. Building on the hypothesis that SNPs in microRNAs (miRNA) have an indirect translational effect on mRNA expression, another study identified the miRNA miR-149 as a potential genetic modifier of the CMT1A phenotype[51]. Identifying such modifier allele associations is important not only because it can help decipher the phenotypic variability in CMT but also because it could highlight potential therapeutic targets.

Mitochondrial genome sequencing

Mutations in the mitochondrially encoded gene *MT-ATP6* can occasionally present as CMT2[52] and, therefore, mitochondrial genome sequencing should be considered for inclusion in the diagnostic evaluation of patients with CMT. Each nucleated cell in the body contains hundreds to thousands of mitochondria, with larger numbers being observed in higher-energy-demanding cells such as myocytes and hepatocytes. Each mitochondrion contains multiple copies of mitochondrial DNA (mtDNA), which is distinct from the single copy of nuclear DNA that is present in each nucleated cell. Recent advances in the sensitivity of mtDNA sequencing now mean that a low number of mitochondrial genomes containing a variant (known as low heteroplasmy) can be detected in DNA derived from blood[53]. This capability is important for the diagnosis of CMT, because even when the mtDNA mutation is present in disease-relevant tissue (for example, nerves in the case of CMT) at a sufficient heteroplasmy level to cause pathology, the heteroplasmy level in blood can be low. Traditionally, PCR-based NGS technologies such as WES were not thought to be useful to screen for mtDNA variants,

as mtDNA had not intentionally been targeted for capture and amplification during NGS. However, we now know that owing to the presence of multiple mtDNA copies per cell, NGS in the form of WES does in fact provide reasonable coverage of the mtDNA, even without mtDNA-specific amplification, and might be useful for screening mtDNA on a research basis[54]. PCR-free WGS can also consistently achieve >1,000-fold coverage across the entire mitochondrial genome and can reliably detect >2% heteroplasmy when a linked analysis of nuclear and mtDNA sequence data is employed[43].

Transcriptome sequencing

The use of NGS in CMT has traditionally focused on screening for variants in the coding regions of genes, and this approach has a reasonable hit rate in terms of identifying the pathogenic variant. However, variants in non-coding regions of CMT-associated genes might also be pathogenic. For example, variants in the 5' untranslated region (5' UTR) or promoter region of a gene might impair gene transcription, leading to haploinsufficiency **[G]**, such as in *GJB1* (causing CMTX1)[55]. Furthermore, variants in the 3' UTR of *GJB1*[55,56] could impair stability of the mRNA transcript, resulting in reduced GJB1 protein translation[57]. In our case series, variants in the 5' and 3' UTRs of *GJB1* accounted for 11.4% of all cases of CMTX1[55]. Therefore, it is important that CMT gene panels now include coverage of these non-coding regions, which are not covered by WES or traditional CMT gene panels.

Although WGS is proving to be useful for detecting non-coding variants, the high frequency of such variants makes interpretation of the data challenging in a diagnostic setting. In an effort to aid the interpretation of variants in non-coding DNA, research groups have used NGS of RNA (termed high-throughput transcriptome sequencing)

as a complementary tool alongside WES and/or WGS. This technique allows the identification of splice variants and novel transcripts, and the quantification of levels of gene expression. The presence of splice variants and novel transcripts can be used to validate coding and non-coding variants in known disease-related genes that have been identified by WES or WGS. This ability is relevant to CMT, as both intronic and coding synonymous variants [G] in *MPZ* and *SH3TC2* have been shown to cause CMT1B[58,59], Dejerine–Sottas syndrome[60] and CMT4C[61] by disrupting the normal splicing of pre-mRNA.

In rare muscle disorders, transcriptome sequencing on disease-relevant tissue (muscle) has been shown to improve the diagnostic rate by facilitating interpretation of the coding and non-coding germline variants identified by WES or WGS[62]. In contrast to muscle diseases, RNA cannot be easily accessed from disease-relevant tissue in CMT, as RNA extraction from the nucleus-containing dorsal root ganglion and/or anterior horn cell bodies[63] is not practical *in vivo.* Transcriptome analysis of easily obtained tissues such as whole blood and/or skin fibroblasts is hindered by the expression patterns of CMT-associated genes, many of which are specific either to neurons (for example, *NEFL*) or to Schwann cells (for example, *MPZ*)[64]. Sufficient RNA for transcriptome analysis might be obtained from the Schwann cells of dermal myelinated fibres in skin biopsies[65]; however, this is not possible for neurons. Some researchers have performed transcriptome sequencing on neural crest cells differentiated from induced pluripotent stem cells (iPSCs), which were generated from peripheral blood monocytes of healthy individuals and patients with CMT1[66] and compared the gene expression profile between the two cohorts. Thus, high-throughput transcriptome sequencing in CMT is currently an area in development and is yet to enter routine clinical practice.

Challenges of using NGS in CMT

The increasing uptake of NGS has resulted in the generation of an increasing amount of sequence data. As discussed above, both WES and WGS produce an abundance of genetic variants and the task is to establish which of these variants are responsible for the disease phenotype. Thus, the diagnostic bottleneck has shifted from the identification of genetic variants to their interpretation. The challenges and approaches to evaluating genetic variants in CMT apply to both the interpretation of novel mutations in known CMT-associated genes and the identification of novel CMT-associated genes.

Novel variants in CMT-associated genes

When evaluating the pathogenicity of novel variants in known CMT-associated genes, several criteria are commonly assessed. First, is the patient's phenotype consistent with the gene that carries the candidate variant? Second, have previous publications identified the variant as disease-causing? Third, what is the population allele frequency (AF; **Box 1**) of the variant? Fourth, is the variant heterozygous, compound heterozygous or homozygous in the context of the gene's known inheritance pattern and the family pedigree? Fifth, does the variant segregate with the observed phenotype? Sixth, what is the in silico predicted functional effect and evolutionary conservation pattern of the variant? Last, is any experimental evidence available on the effect of the candidate variant on the structure and function of the resulting protein[67]?

The aggregated set of data can then be formalized using the American College of Medical Genetics and Genomics (ACMG) criteria[68]. These guidelines state that variants of uncertain significance should not be used in clinical decision-making, but

they also recognize that the weight assigned to certain criteria can vary by disease and gene. For example, in CMT the availability of phenotype–genotype correlations allows investigators to use deep phenotyping as a criterion for evaluating the pathogenicity of novel variants. The weight assigned to this criterion might depend on how many specific phenotypic features the patient exhibits.

Population allele frequency

The population AF of a variant refers to the proportion of alleles in that population (general or control population) that carry the variant. Population genetics stipulate that the AF of a fully penetrant pathogenic variant cannot be higher than the prevalence of the disease it causes in that population. Therefore, we would expect that a fully penetrant pathogenic variant should not be present or be extremely rare in a control population. However, a common problem that arises when evaluating a novel variant in a known CMT-associated gene is the determination of this acceptable or maximum credible population AF in large control population variant databases such as the Exome Aggregation Consortium (ExAC) and Genome Aggregation Database (GnomAD)[14]. For example, in the ExAC database, 11 of 87,342 alleles (AF 0.00013) are predicted to contain the 17p duplication that causes CMT1A, and in GnomAD, the most common pathogenic mutation in *TTR* (Val50Met), which causes familial amyloid polyneuropathy, is present in 26 of 251,462 alleles (AF 0.0001). These examples show that known pathogenic variants causing CMT may be present in a control population, albeit with rare AF, and hence the maximum credible population AF of pathogenic variants should be determined and can be used to filter novel variants in known CMT-associated genes.

To determine this maximum credible population AF, a statistical framework suggested by Whiffin and colleagues[69] is very useful. Taking into account the genetic architecture of the second commonest subtype of CMT, CMTX1[8,70], which shows a dominant pattern of inheritance, we would allow a maximum credible population AF of 3.52 x 10^{-6} for any pathogenic variant in known dominant CMT-associated genes; this AF would equate to 3 of the ~280,000 alleles currently in GnomAD. Any novel variant that is more frequent than this would be deemed too common and thus discarded (BOX 1). Similarly, using the most common subtype of recessive CMT, CMT4C[8], as a reference, we would allow a maximum population AF of 0.000667 (210 in ~280,000 alleles) for any pathogenic variant in known recessive CMT-associated genes.

Phenotype–genotype associations

CMT classically causes a length-dependent mixed motor and sensory neuropathy — a phenotype that is common to many genetic subtypes of the disease (Supplementary Table 1). Nevertheless, there are many clinical features that can help us to differentiate between the various genetic subtypes, which is important when determining the pathogenicity of novel variants (FIG. 2).

A useful approach to defining the phenotype in a patient with CMT is to first establish the age of onset and whether the neuropathy is predominantly motor, sensory or both. The pattern of involvement of different regions of the body should also be described. Although the pattern of neuropathy is usually length-dependent, beginning in the lower limbs and extending over time to involve the upper limbs, pathogenic variants in a small number of genes, including *GARS* and *BSCL2*, cause an upper limb length-dependent neuropathy. Neurophysiology remains an essential component of the clinical assessment of CMT and can determine whether conduction velocity is slowed,

which would indicate a demyelinating neuropathy. A lack of slowing of conduction velocity would indicate an axonal neuropathy.

Once the nature of the peripheral neuropathy has been defined, additional clinical features might give important clues to the genetic diagnosis and aid variant interpretation. Examples include the presence of cognitive impairment (implicating *DNMT1* variants if progressive and associated with deafness), spasticity (suggesting *BSCL2* or *KIF5A* variants), optic atrophy (suggesting *MFN2* variants) and prominent vocal cord involvement (suggesting *TRPV4, SLC5A7* or *DCTN1* variants).

In certain circumstances, ancillary tests can also facilitate variant interpretation. For example, in a congenital form of HMN with selective involvement of the quadriceps but sparing and hypertrophy of the adductors, caused by dominant mutations in *BICD2* or *DYNC1H1*, thigh muscle MRI can be used to detect this pattern of selective muscle involvement[71,72]. Some forms of CMT also have characteristic features that can be observed in nerve biopsy samples and show a strong correlation with mutations in certain genes predominantly associated with demyelinating CMT. For example, mutations in *SBF1* and *SBF2* are associated with focally folded myelin in nerve biopsy samples[73]. Despite these strong correlations, a nerve biopsy is not recommended in the initial diagnostic evaluation of a patient with CMT, as the prevalence of CMT subtypes with such specific neuropathological features is too low to justify carrying out the procedure routinely.

Identification of new phenotypic features that are considered to 'expand the phenotype' of rare diseases is commonplace, but caution should be exercised in the case of CMT as individual patients with genetic variation in more than one CMT-associated gene have been described. A population study of 787 patients with CMT

identified at least 11 patients with multiple CMT subtypes (CMT1A/1E, CMT1E/1B, CMTX1/1B or CMT1A/1C), and as not all patients were tested for multiple variants, this percentage could be higher[11]. Other examples include the coexistence of the 17p duplication with two *LITAF* variants in a patient with severe CMT1[74], and the presence of both *MFN2* and *GDAP1* variants in a patient with severe CMT2[75]. Furthermore, a recent retrospective analysis of 2,076 patients with a rare genetic diagnosis found that that 5% of patients had between two and four diagnoses[76], suggesting that the co-occurrence of more than one independent genetic diagnosis in a single individual is more common than previously thought. In relation to these observations, there are also reports of patients with neuromuscular weakness who had molecular diagnoses of both CMT and myotonic or muscular dystrophy[77,78]. Often termed 'double trouble', this genetic mechanism, which is different to digenic inheritance, might account for a proportion of cases at the severe end of the phenotypic spectrum. Moreover, it challenges the conventional 'search and stop' approach when interpreting genetic variation, in which the quest to identify the pathogenic variants causing a Mendelian disease in an individual stops once one pathogenic variant that explains the phenotype is found.

De novo variants

De novo variants arise spontaneously and are more likely to be pathogenic than are inherited variants, as by definition they are exposed to less evolutionary pressure. Nevertheless, we should not assume that all de novo mutations are causal for an observed phenotypic trait. Statistical algorithms that analyse data from family-based studies consisting of WGS trios estimate that each human genome harbours ~73 de novo germline SNVs and 3–9 de novo germline indels per generation, with at least one of these de novo mutations occurring in the exome[79,80]. The rate of de novo

mutations is influenced by many factors, including the presence of low-copy repeat (LCR) sequences (increases the likelihood of de novo occurrence of SVs in that genomic area) and a high density of CpG dinucleotides (increases the likelihood of de novo occurrence of SNVs). De novo CNVs at the 17p locus that cause CMT1A or HNPP have an increased chance of occurring owing to nonallelic homologous recombination between LCRs that are present at this locus. Furthermore, certain CMT-causing SNVs that frequently occur de novo in unrelated families, such as p.Arg252Trp in *MORC2*[81] and p.Arg94Gln/Trp in *MFN2*[82], occur in cytosine and guanine nucleotides that reside in a CpG-sequence.

Predicted loss-of-function variants

Loss-of-function (LoF) variants are abundant in the general population, and the genome of a healthy individual is estimated to carry ~100 of these variants[83]. Nonetheless, biallelic LoF variants in recessive genes are always prioritized for analysis owing to the predicted protein knockdown. However, heterozygous LoF variants in dominant CMT-associated genes should also be considered for analysis in the context of the gene's known mechanism for inducing pathology and its tolerance to haploinsufficiency. For example, heterozygous LoF variants in *MPZ* have been described as pathogenic[84], whereas those in the *NEFL* gene seem to cause no phenotype[85,86] and no change in neurophysiology[86] as the gene tolerates haploinsufficiency.

False positives and false negatives

Technical errors, such as sequencing errors by the DNA polymerase or erroneous reading of fluorescently tagged nucleotide bases owing to relative changes in their fluorescent signals, can be introduced at various steps of the NGS process. These

errors can lead to false-positive variant calling and, thus, the necessity to confirm candidate variants by Sanger sequencing[87].

Complex genomic regions (for example, GC-rich regions, AT-rich regions and homopolymer stretches) and mosaicism are often the culprits of false-negative base calling. A large study of 20,000 hereditary-cancer NGS panels estimated a false-negative rate of 2.2% of Sanger sequencing-confirmed variants[87]. On the other hand, in individuals with false-negative Sanger sequencing results, NGS can be used (in combination with linkage analysis) to identify causative variants, as illustrated by one of our patients, who had dHMN and a familial pathogenic *HSPB1* variant (Pro182Ala). This variant was initially missed by Sanger sequencing owing to a *cis* insertion of 4 bp in the intron immediately upstream of the pathogenic variant, resulting in preferential amplification of the normal allele[88].

Somatic mosaicism is rarely reported in CMT, but offspring with CMT due to variants in *MFN2*, *PMP22*, *GJB1* or *MPZ* and concurrent parental mosaic mutations, which cause milder phenotypes have been reported[89]. A sequencing trace **[G]** showing low levels of the mutant allele in the parent, which may be falsely recorded as negative, often gives a clue to possible mosaicism. Mosaicism might then need to be confirmed by proving that the mutation is present in the affected offspring and by showing differential genomic expression of the mutant allele in different parental tissues.

Novel genes in CMT

NGS technologies have substantially increased the number of patients with CMT who receive a genetic diagnosis. Despite these advances, a notable proportion of patients with a clinical diagnosis of CMT remain genetically undiagnosed and are presumed to

have as yet unidentified genetic mutations or genomic rearrangements. Patients with CMT2 are particularly affected, with >60% remaining without a diagnosis[12,24].

Traditional genetic tools such as linkage analysis on large pedigrees and homozygosity mapping on large or small pedigrees — especially those with a degree of consanguinity — are invaluable in the quest for novel genes with mutations that cause CMT. We increasingly observe that newly discovered CMT-associated genes, such as *MME*[19], *ATP1A1*[90] and *SCO2*[91], are expressed across many tissues. Gene hunting studies using WES and WGS data become more challenging when such widely expressed genes are considered as candidate novel CMT-associated genes, alongside axon-specific or Schwann cell-specific genes. Nonetheless, in the past 5 years, the number of published reports of novel CMT-associated genes in single families or sporadic cases has increased substantially. However, the pathogenicity of the genetic mutation in such cases remains uncertain, and additional genetic and functional evidence of pathogenicity would be required to correctly interpret these findings. This uncertainty is important to consider as some of the genes included in commercially available CMT panels were identified in single families or sporadic cases. In some cases, the original families in which a novel CMT gene was identified were subsequently discovered to harbour a pathogenic mutation in another known CMT-associated gene. For example, variants in *PNKP* have recently been reported to cause CMT2B2, which was previously associated with genetic variants in *MED25*[92].

A 'genome-first' diagnostic pathway

NGS is revolutionizing genomic medicine, and we propose a shift in the CMT diagnostic pathway towards a genome-first approach. In this approach, for patients with a non-CMT1 phenotype, molecular investigations would start with WGS. For

patients with a CMT1 phenotype, stand-alone testing would be used to exclude the presence of a 17p rearrangement before performing WGS. Following the successful introduction of the 100,000 Genomes Project in the UK, our diagnostic practice in CMT is shifting towards adopting WGS, coupled with analysis of a virtual panel of 200 genes, as a first-line investigation.

A genome-first approach has multiple advantages over the existing diagnostic pathway, for example, the ability to reliably sequence across all of the loci that contribute to the genetic heterogeneity of CMT, including both the nuclear and mitochondrial genomes and more reliable coverage of GC-rich regions[26,37].

The superior coverage performance of WGS as compared with WES places it on a par with the NGS gene panels that are currently used in the clinical setting. Moreover, WGS can detect coding, splice site and non-coding SNVs more reliably than WES[39] and can also identify large balanced and unbalanced SVs (including those >500 kb), which have occasionally been shown to cause CMT[41,42]. Furthermore, in the case of single recessive variants, WGS allows one to screen for compound heterozygous SVs, such as deletions, without needing to proceed to MLPA.

Homozygosity mapping, usually performed using high-density microsatellite [G] or SNP arrays, identifies genomic runs of homozygosity [G] (ROH) and is an effective molecular technique for mapping recessive alleles in consanguineous families, inbred minority populations or families from isolated geographical regions. Short ROH (<4 Mb) are ubiquitous in outbred populations, whereas multiple large ROH (usually >10 Mb) across different chromosomes are indicative of consanguinity[93,94]. One or two large ROH on the same chromosome, especially if telomeric, would indicate uniparental disomy [G]. In families with suspected autosomal recessive CMT,

homozygosity mapping has been applied to WES data to detect ROH, and has identified recessive variants with a sensitivity of 84% and a specificity of 82% using high-density SNP arrays as a reference[95]. In comparison with WES, WGS offers genome-wide genotyping of a larger number of SNPs and can identify and finely map large and small ROH with a greater sensitivity[94].

Ongoing efforts, which have been profoundly transformed by NGS technologies, aim to further our understanding of and improve diagnosis in rare inherited diseases. These efforts include studying the coding and non-coding transcriptome[62,96,97] and mapping the human epigenome[98], that is, the set of non-covalent modifications of genetic material, such as nucleotide methylation and histone modification, that affect gene expression. The availability of banked WGS data will allow re-analysis as new knowledge of variants and genotype–phenotype associations comes to light.

Understandably, the cost of routinely carrying out WGS is a concern for any clinical diagnostic service. However, the cost of WGS is decreasing over time, and small studies into the health economics of WES and WGS usage in a clinical setting illustrate that these approaches could represent a cost-effective use of national health-care resources[99,100]. On a case-by-case basis, the cost of employing WGS to diagnose common pathogenic variants in *MFN2* or *GJB1* might be difficult to justify, but the overall cost-effectiveness of a genome-wide approach in the CMT cohort as a whole should be considered. To date, 40% of patients with CMT remain without a genetic diagnosis, despite having undergone multiple genetic investigations including individual gene testing, microarray comparative genomic hybridization **[G]**, gene-specific panels and/or WES. Furthermore, a considerable proportion of patients who eventually receive a genetic diagnosis of CMT undergo more than one genetic test to reach that diagnosis. Therefore, a genome-first approach is likely to reduce the

cumulative costs of repeated genetic testing and partly offset the cost of using WGS to diagnose genetically 'straightforward' cases of CMT. Along these lines, in one study, an exome-first approach that was proposed as a diagnostic pathway in a cohort of patients with CMT (after exclusion of the 17p rearrangement) was deemed to be cost-effective overall[101]. Despite this evidence, we appreciate that the availability of and access to WGS facilities varies greatly between health-care settings, and we suspect that most investigators will not be able use WGS as a first-line test, at least in the short term. Where possible, WES can be performed instead of WGS; however, the investigators would need to take into consideration the lower diagnostic yield of WES and its incomplete coverage of protein-coding regions, which were both discussed earlier in this Review.

Interpreting the identified variants in a genome-first approach will be challenging but can be aided by applying phenotype-guided virtual gene panels to the WGS data (FIG. 3), which will increase the diagnostic yield and reduce both background genetic variation (or noise) and the likelihood of secondary findings [G]. Furthermore, this filtering approach will completely bypass the need to continually update and validate the CMT gene panel testing kit, as any new CMT-associated genes will simply be added to the virtual gene panel.

The genome-first diagnostic approach that we have suggested takes into consideration the genetic epidemiology of a heterogeneous CMT population. Therefore, a more tailored approach would be warranted when studying individuals with CMT from specific ethnic populations in which the CMT epidemiology can be skewed. This adaptation of the diagnostic approach might reduce the time taken to achieve a genetic diagnosis. For example, in consanguineous families with multiple affected siblings and unaffected parents, it may be appropriate to proceed directly to

testing a panel of common recessive CMT-associated genes, such as *NDRG1*, which should be checked for genetic variation in Roma families from Eastern Europe.

Conclusions

The introduction of NGS in the form of gene-specific panels has revolutionized the genetic diagnostic pathway in patients with CMT. The genetic heterogeneity observed in CMT means that each patient with a clinical diagnosis of CMT needs to have a group of CMT-associated genes checked in one go, and this standardized approach has enabled more patients with CMT to receive a genetic diagnosis than ever before. Although WES has proven to be a reliable research tool for discovering novel CMT-associated genes, WGS holds the most promise for a holistic genetic approach. Among other virtues, WGS offers near-complete coverage of the exome, sequences more reliably than other techniques through complex genomic regions, is able to detect balanced and unbalanced SVs and is suitable for searching and identifying genetic modifiers that may at some point act as therapeutic targets.

As NGS becomes cheaper and more accessible, we are now moving towards the use of WGS as the first-line diagnostic test in the evaluation of patients with CMT (excluding CMT1A). Although this approach is likely to increase the diagnostic rate and further our understanding of the genetic basis of CMT, it also requires specific skills and expertise for the interpretation of the multiple variants observed in known CMT-associated genes. The burden of variant interpretation is considerable, and robust filtering strategies that use all the available clinical, genetic and bioinformatic information are required to curate variants. Our knowledge and handling of genomic data acquired through WGS is improving; however, the next big step will be a move towards functional genomics, whereby the integration of genomic, transcriptomic,

proteomic and metabolomic data will enable us to construct a more accurate model of human disease.

1.   Skre, H. Genetic and clinical aspects of Charcot-Marie-Tooth's disease. *Clin. Genet.* **6**, 98–118 (1974).

2.   Barreto, C. L. S. *et al.* Epidemiologic Study of Charcot-Marie-Tooth Disease: A Systematic Review. *Neuroepidemiology* **46**, 157–165 (2016).

3.   Reilly, M. M., Murphy, S. M. & Laura, M. Charcot-Marie-Tooth disease. *J. Peripher. Nerv. Syst.* **16**, 1–14 (2011).

4.   Feely, S. M. E. *et al.* MFN2 mutations cause severe phenotypes in most patients with CMT2A. *Neurology* **76**, 1690–1696 (2011).

5.   Harding, A. E. & Thomas, P. K. The clinical features of hereditary motor and sensory neuropathy types I and II. *Brain* **103**, 259–280 (1980).

6.   Mathis, S. *et al.* Charcot-Marie-Tooth diseases: An update and some new proposals for the classification. *J. Med. Genet.* **52**, 681–690 (2015).

7.   Rossor, A. M. *et al.* Peripheral neuropathy in complex inherited diseases: an approach to diagnosis. *J. Neurol. Neurosurg. Psychiatry* **88**, 846–863 (2017).

8.   Fridman, V. *et al.* CMT subtypes and disease burden in patients enrolled in the Inherited Neuropathies Consortium natural history study: a cross-sectional analysis. *J. Neurol. Neurosurg. Psychiatry* **86**, 873–878 (2015).

9.   Rossor, A. M., Tomaselli, P. J. & Reilly, M. M. Recent advances in the genetic neuropathies. *Curr. Opin. Neurol.* **29**, 537–548 (2016).

10.   Murphy, S. M. *et al.* Charcot-Marie-Tooth disease: Frequency of genetic

subtypes and guidelines for genetic testing. *J. Neurol. Neurosurg. Psychiatry* **83**, 706–710 (2012).

11. Saporta, A. S. D. *et al.* Charcot-marie-tooth disease subtypes and genetic testing strategies. *Ann. Neurol.* **69**, 22–33 (2011).

12. Rossor, A. M., Polke, J. M., Houlden, H. & Reilly, M. M. Clinical implications of genetic advances in Charcot-Marie-Tooth disease. *Nat. Rev. Neurol.* **9**, 562–571 (2013).

13. Landrum, M. J. *et al.* ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

14. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

15. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proc. Natl. Acad. Sci. U S A* **111**, E455–E464 (2014).

16. Marttila, M. *et al.* Molecular Epidemiology of Charcot-Marie-Tooth Disease in Northern Ostrobothnia, Finland: A Population-Based Study. *Neuroepidemiology* **49**, 34–39 (2017).

17. Gabrikova, D. *et al.* Founder mutations in NDRG1 and HK1 genes are common causes of inherited neuropathies among Roma/Gypsies in Slovakia. *J. Appl. Genet.* **54**, 455–460 (2013).

18. Lupo, V. *et al.* Characterising the phenotype and mode of inheritance of patients with inherited peripheral neuropathies carrying MME mutations. *J. Med. Genet.* **55**, 814–823 (2018).

19. Higuchi, Y. *et al.* Mutations in MME cause an autosomal-recessive Charcot-

Marie-Tooth disease type 2. *Ann. Neurol.* **79**, 659–672 (2016).

20. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).

21. Antoniadi, T. *et al.* Application of targeted multi-gene panel testing for the diagnosis of inherited peripheral neuropathy provides a high diagnostic yield with unexpected phenotype-genotype variability. *BMC Med. Genet.* **16**, 10.1186/s12881-015-0224–8 (2015).

22. Nam, S. H. *et al.* Identification of Genetic Causes of Inherited Peripheral Neuropathies by Targeted Gene Panel Sequencing. *Mol. Cells* **39**, 382–388 (2016).

23. Lupo, V. *et al.* Assessment of Targeted Next-Generation Sequencing as a Tool for the Diagnosis of Charcot-Marie-Tooth Disease and Hereditary Motor Neuropathy. *J. Mol. Diagn.* **18**, 225–234 (2016).

24. Yoshimura, A. *et al.* Genetic profile and onset features of 1005 patients with Charcot-Marie-Tooth disease in Japan. *J. Neurol. Neurosurg. Psychiatry* **90**, 195–202 (2018).

25. Wang, W. *et al.* Target-enrichment sequencing and copy number evaluation in inherited polyneuropathy. *Neurology* **86**, 1762–1771 (2016).

26. Meienberg, J., Bruggmann, R., Oexle, K. & Matyas, G. Clinical sequencing: is WGS the better WES? *Hum. Genet.* **135**, 359–362 (2016).

27. Choi, B. O. *et al.* Exome sequencing is an efficient tool for genetic screening of Charcot-Marie-Tooth disease. *Hum. Mutat.* **33**, 1610–1615 (2012).

28. Drew, A. P. *et al.* Improved inherited peripheral neuropathy genetic diagnosis

by whole-exome sequencing. *Mol. Genet. Genomic Med.* **3**, 143–154 (2015).

29. Gonzaga-Jauregui, C. *et al.* Exome Sequence Analysis Suggests that Genetic Burden Contributes to Phenotypic Variability and Complex Neuropathy. *Cell Rep.* **12**, 1169–1183 (2015).

30. Hartley, T. *et al.* Whole-exome sequencing is a valuable diagnostic tool for inherited peripheral neuropathies: Outcomes from a cohort of 50 families. *Clin. Genet.* **93**, 301–309 (2018).

31. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).

32. Bansagi, B. *et al.* Genetic heterogeneity of motor neuropathies. *Neurology* **88**, 1226–1234 (2017).

33. Jo, H. Y. *et al.* Application of whole-exome sequencing for detecting copy number variants in CMT1A/HNPP. *Clin. Genet.* **90**, 177–181 (2016).

34. Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U S A* **112**, 5473–5478 (2015).

35. Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A. & Gilissen, C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum. Mutat.* **36**, 815–822 (2015).

36. Dewey, F. E. *et al.* Clinical Interpretation and Implications of Whole-Genome Sequencing. *JAMA* **311**, 1035–1045 (2014).

37. Carss, K. J. *et al.* Comprehensive Rare Variant Analysis via Whole-Genome

Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. *Am. J. Hum. Genet.* **100**, 75–90 (2017).

38. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* **19**, 253–268 (2018).

39. Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* **47**, 717–726 (2015).

40. Lupski, J. R. *et al.* Whole-Genome Sequencing in a Patient with Charcot-Marie-Tooth Neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).

41. Brewer, M. H. *et al.* Whole Genome Sequencing Identifies a 78 kb Insertion from Chromosome 8 as the Cause of Charcot-Marie-Tooth Neuropathy CMTX3. *PLoS Genet.* (2016). doi:10.1371/journal.pgen.1006177

42. Drew, A. P., Cutrupi, A. N., Brewer, M. H., Nicholson, G. A. & Kennerson, M. L. A 1.35 Mb DNA fragment is inserted into the DHMN1 locus on chromosome 7q34–q36.2. *Hum. Genet.* **135**, 1269–1278 (2016).

43. Raymond, F. L., Horvath, R. & Chinnery, P. F. First-line genomic diagnosis of mitochondrial disorders. *Nat. Rev. Genet.* **19**, 399–400 (2018).

44. Minoche, A. E. *et al.* Genome sequencing as a first-line genetic test in familial dilated cardiomyopathy. *Genet. Med.* **21**, 650–662 (2018).

45. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).

46. Turnbull, C. *et al.* The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* **361**, k1687 (2018).

47.   Gonzalez, M. *et al.* Innovative Genomic Collaboration Using the GENESIS (GEM.app) Platform. *Hum. Mutat.* **36**, 950–956 (2015).

48.   Nadeau, J. H. Modifier genes in mice and humans. *Nat. Rev. Genet.* **2**, 165–174 (2001).

49.   Barnett, I. J., Lee, S. & Lin, X. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genet. Epidemiol.* **37**, 142–151 (2013).

50.   Tao, F. *et al.* Variation in SIPA1L2 Is Correlated with Phenotype Modification in Charcot-Marie-Tooth Disease Type 1A. *Ann. Neurol.* **85**, 316–330 (2019).

51.   Nam, S. H. *et al.* Association of miR-149 polymorphism with onset age and severity in Charcot–Marie–Tooth disease type 1A. *Neuromuscul. Disord.* **28**, 502–507 (2018).

52.   Pitceathly, R. D. *et al.* Genetic dysfunction of MT-ATP6 causes axonal Charcot-Marie-Tooth disease. *Neurology* **79**, 1145–1154 (2012).

53.   Tang, S. *et al.* Transition to Next Generation Analysis of the Whole Mitochondrial Genome: A Summary of Molecular Defects. *Hum. Mutat.* **34**, 882–893 (2013).

54.   Picardi, E. & Pesole, G. Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat. Methods* **9**, 523–524 (2012).

55.   Tomaselli, P. J. *et al.* Mutations in noncoding regions of GJB1 are a major cause of X-linked CMT. *Neurology* **88**, 1445–1453 (2017).

56.   Chen, D. H. *et al.* An 8-generation family with X-linked Charcot–Marie–Tooth: Confirmation Of the pathogenicity Of a 3′ untranslated region mutation in GJB1

and its clinical features. *Muscle Nerve* **57**, 859–862 (2018).

57. Chatterjee, S. & Pal, J. K. Role of 5'- and 3'-untranslated regions of mRNAs in human diseases. *Biol. Cell* **101**, 251–262 (2009).

58. Corrado, L. *et al.* A novel synonymous mutation in the MPZ gene causing an aberrant splicing pattern and Charcot-Marie-Tooth disease type 1b. *Neuromuscul. Disord.* **26**, 516–520 (2016).

59. Crehalet, H. *et al.* U1 snRNA mis-binding: A new cause of CMT1B. *Neurogenetics* **11**, 13–19 (2010).

60. Taioli, F. *et al.* Dejerine-Sottas syndrome with a silent nucleotide change of myelin protein zero gene. *J. Peripher. Nerv. Syst.* **16**, 59–64 (2011).

61. Laššuthová, P. *et al.* High frequency of SH3TC2 mutations in Czech HMSN I patients. *Clin. Genet.* **80**, 334–345 (2011).

62. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).

63. Sapio, M. R., Goswami, S. C., Gross, J. R., Mannes, A. J. & Iadarola, M. J. Transcriptomic analyses of genes and tissues in inherited sensory neuropathies. *Exp. Neurol.* **283**, 375–395 (2016).

64. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

65. Li, J. *et al.* Skin biopsies in myelin-related neuropathies: bringing molecular pathology to the bedside. *Brain* **128**, 1168–1177 (2005).

66. Kitani-Morii, F. *et al.* Analysis of neural crest cells from Charcot-Marie-Tooth

disease patients demonstrates disease-relevant molecular signature. *Neuroreport* **28**, 814–821 (2017).

67. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).

68. Richards, S. *et al.* Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

69. Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.* **19**, 1151–1158 (2017).

70. Panosyan, F. B. *et al.* Cross-sectional analysis of a large cohort with X-linked Charcot-Marie-Tooth disease (CMTX1). *Neurology* **89**, 927–935 (2017).

71. Rossor, A. M. *et al.* Phenotypic and molecular insights into spinal muscular atrophy due to mutations in BICD2. *Brain* **138**, 293–310 (2015).

72. Scoto, M. *et al.* Novel mutations expand the clinical spectrum of DYNC1H1-associated spinal muscular atrophy. *Neurology* **84**, 668–679 (2015).

73. Houlden, H., King, R. H., Wood, N. W., Thomas, P. K. & Reilly, M. M. Mutations in the 5' region of the myotubularin-related protein 2 (MTMR2) gene in autosomal recessive hereditary neuropathy with focally folded myelin. *Brain* **124**, 907–915 (2001).

74. Meggouh, F. *et al.* Early onset neuropathy in a compound form of Charcot-Marie-Tooth disease. *Ann. Neurol.* **57**, 589–591 (2005).

75. Anghelescu, C. *et al.* Targeted exomes reveal simultaneous MFN2 and

GDAP1 mutations in a severe Charcot-Marie-Tooth disease type 2 phenotype. *Eur. J. Neurol.* **24**, e15–e16 (2017).

76. Posey, J. E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N. Engl. J. Med.* **376**, 21–31 (2017).

77. Hodapp, J. A. *et al.* Double Trouble in Hereditary Neuropathy: Concomitant Mutations in the PMP-22 Gene and Another Gene Produce Novel Phenotypes. *Arch. Neurol.* **63**, 112–117 (2006).

78. Schreiber, O. *et al.* Facioscapulohumeral muscular dystrophy and Charcot-Marie-Tooth neuropathy 1A - evidence for ' double trouble' overlapping syndromes. *BMC Med. Genet.* **14**, 92 (2013).

79. Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* **6**, 5969 (2015).

80. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).

81. Sevilla, T. *et al.* Mutations in the MORC2 gene cause axonal Charcot-Marie-Tooth disease. *Brain* **139**, 62–72 (2016).

82. Verhoeven, K. *et al.* MFN2 mutation distribution and genotype/phenotype correlation in Charcot-Marie-Tooth type 2. *Brain* **129**, 2093–2102 (2006).

83. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).

84. Sanmaneechai, O. *et al.* Genotype-phenotype characteristics and baseline natural history of heritable neuropathies caused by mutations in the MPZ gene.

Brain **138**, 3180–3192 (2015).

85. Abe, A. *et al.* Neurofilament light chain polypeptide gene mutations in Charcot–Marie–Tooth disease: nonsense mutation probably causes a recessive phenotype. *J. Hum. Genet.* **54**, 94–97 (2009).

86. Yum, S. W., Zhang, J., Mo, K., Li, J. & Scherer, S. S. A Novel Recessive NEFL Mutation Causes a Severe, Early-Onset Axonal Neuropathy. *Ann Neurol* **66**, 759–770 (2009).

87. Mu, W., Lu, H. M., Chen, J., Li, S. & Elliott, A. M. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. *J. Mol. Diagn.* **18**, 923–932 (2016).

88. Rossor, A. M. *et al.* Pilot phenotype and natural history study of hereditary neuropathies caused by mutations in the HSPB1 gene. *Neuromuscul. Disord.* **27**, 50–56 (2017).

89. Schon, K. *et al.* Mosaicism for a pathogenic MFN2 mutation causes minimal clinical features of CMT2A in the parent of a severely affected child. *Neurogenetics* **18**, 49–55 (2017).

90. Lassuthova, P. *et al.* Mutations in ATP1A1 Cause Dominant Charcot-Marie-Tooth Type 2. *Am. J. Hum. Genet.* **102**, 505–514 (2018).

91. Rebelo, A. P. *et al.* SCO2 mutations cause early-onset axonal Charcot-Marie-Tooth disease associated with cellular copper deficiency. *Brain* **141**, 662–672 (2018).

92. Leal, A. *et al.* The polynucleotide kinase 3′-phosphatase gene (PNKP) is involved in Charcot-Marie-Tooth disease (CMT2B2) previously related to

MED25. *Neurogenetics* **19**, 215–225 (2018).

93. Kirin, M. *et al.* Genomic runs of homozygosity record population history and consanguinity. *PLoS One* **5**, e13996 (2010).

94. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: Windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).

95. Kancheva, D. *et al.* Novel mutations in genes causing hereditary spastic paraplegia and Charcot-Marie-Tooth neuropathy identified by an optimized protocol for homozygosity mapping based on whole-exome sequencing. *Genet. Med.* **18**, 600–607 (2016).

96. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, 951–969 (2015).

97. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* **19**, 535–548 (2018).

98. Rivera, C. M. & Ren, B. Mapping Human Epigenomes. *Cell* **155**, 39–55 (2013).

99. Schwarze, K., Buchanan, J., Taylor, J. C. & Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* **20**, 1122–1130 (2018).

100. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: https://www.genome.gov/sequencingcostsdata. (Accessed: 10th July 2019)

101. Walsh, M. *et al.* Diagnostic and cost utility of whole exome sequencing in

peripheral neuropathy. *Ann. Clin. Transl. Neurol.* **4**, 318–325 (2017).

Acknowledgements

Author contributions

All authors researched data for this article and provided substantial contributions to discussions of the content, writing the article and reviewing and/or editing the manuscript before submission.

Competing interests

The authors declare no competing interests.

Publisher's note

Supplementary information

Key Points

- In Charcot–Marie–Tooth disease (CMT), next-generation sequencing (NGS) technologies are applied in the form of specific CMT-associated gene panels, whole-exome sequencing, whole-genome sequencing (WGS), mitochondrial sequencing and high-throughput transcriptome sequencing.

- Interpretation of NGS-derived variants is challenging owing to the high volume of returned variants and the phenotypic and genetic heterogeneity of CMT.

- Setting a maximum credible population allele frequency for pathogenic variants in dominant and recessive CMT-associated genes is crucial for efficiently filtering NGS-derived variants.

- A genome-first approach utilizing WGS has multiple advantages over other genetic tests for the diagnosis of CMT and should be combined with virtual gene panels to achieve the optimum balance between improved diagnostic yield and burden of variant analysis.

Box 1: Maximum credible population AF of CMT-associated variants

When assessing candidate variants for rare diseases, crude filters for minor allele frequency (AF) have historically been set at <1% for recessive and <0.1% for dominant traits. However, this filtering strategy still leaves a myriad of candidate variants, and a more thoughtful approach is needed to determine the maximum credible population AF of pathogenic variants in large population databases.

To address this problem, Whiffin and colleagues[69] developed a statistical framework for predicting this threshold, taking into account the genetic architecture of the disease in question. The genetic architecture of a disease comprises its prevalence in a population, its genetic heterogeneity (the maximum proportion of the disease caused by variants in a single gene), its allelic heterogeneity (the commonest pathogenic allele within a disease-related gene) and its penetrance. In cystic fibrosis, for example, the genetic heterogeneity is 1 as the disease is caused by genetic variation in one gene *(CFTR)*, and the allelic heterogeneity is 0.7 as the F509del variant within *CFTR* is responsible for 70% of the disease.

We applied this model to the commonest subtypes of CMT and set the population prevalence of CMT at 1:2,500. In CMTX1, which is caused by X-linked dominant *GJB1* variants and accounts for 10.7% of all genetically diagnosed cases of CMT (genetic heterogeneity), the commonest pathogenic variant accounts for 7.5% of all cases (allelic heterogeneity). The penetrance was set at 0.5 to account for the male–female difference in phenotypic expression and we calculated a maximum credible AF of 3.52 x 10^{-6} for pathogenic variants. As illustrated below, the commonest pathogenic variants in the commonest dominant subtypes of CMT are not present in GnomAD[14] more than once, thus respecting the predicted maximum credible AF of 3.52 x 10^{-6},

which translates to a predicted maximum GnomAD allele count (AC) of 3. By extrapolation, this filtering threshold would apply to all dominant CMT-associated genes, as they are rarer and account for far fewer cases of CMT than these subtypes that we used here as examples.

Gene characteristic	GJB1	MFN2	MPZ	SH3TC2
Inheritance pattern	X-linked dominant	Autosomal dominant	Autosomal dominant	Autosomal recessive
Genetic heterogeneity	10.7%[8]	7%[8]	6.7%[8]	1.4%[8]
Allelic heterogeneity	7.5% (Arg15Gln)*	14% (Arg94Gln/Trp)[§]	8.5% (His39Pro)[84]	23% (Arg954X)[§]
Penetrance	0.5[‡]	0.8	0.8	0.95
Predicted maximum credible AF	3.52×10^{-6}	2.45×10^{-6}	1.4×10^{-6}	0.000667
Predicted maximum GnomAD AC in ~280,000 alleles	3	2	2	210
Observed GnomAD AC	0	1	0	206

*The commonest pathogenic *GJB1* variant is Arg15Gln[70]. [‡]Set at 0.5 due to sex-specific phenotype. [§]Unpublished observations .

Figure 1. Venn diagrams showing CMT-associated genes by broad CMT phenotype. a | Autosomal and X-linked dominant genes. **b** | Autosomal and X-linked recessive genes. The Charcot–Marie–Tooth disease (CMT) phenotype is shown in a coloured box and the genes associated with each phenotype are contained within the oval of the corresponding colour. Genes that have been linked to more than one phenotype are illustrated in the overlap fields of phenotypes, for example, *MORC2, HSPB8* and *HSPB1* are in the CMT2/hereditary motor neuropathy (HMN) overlap field and *RAB7A* is in the CMT2/hereditary sensory neuropathy (HSN) overlap field. CMT subtypes with intermediate motor nerve conduction velocities are categorized in the CMT1/CMT2 overlap. X-linked dominant genes (part **a**; for example, *GJB1*) are associated with a phenotype in the hemizygous or heterozygous state, whereas X-linked recessive genes (part **b**; for example, *AIFM1*) are associated with a phenotype in the hemizygous or homozygous state. 17p, 1.4 Mb duplication in the short arm of chromosome 17.

Figure 2. Deep CMT phenotype–genotype correlations. Charcot–Marie–Tooth disease (CMT)-associated genes (according to the Online Mendelian Inheritance in Man (OMIM) and OrphaNet databases) are grouped according to phenotype terms described by the Human Phenotype Ontology and other commonly used clinical terms. **a** | Phenotypes relating to different regions of the body. **b** | Phenotypes relating to the head and neck region. Genes on a red background are associated with a dominant inheritance pattern (autosomal or X-linked), genes on a blue background are associated with a recessive inheritance pattern (autosomal or X-linked) and genes on a green background are associated with a mitochondrial inheritance pattern. Complex neurological syndromes with neuropathy have not been included in this figure, neither

have common terms such as areflexia, pes cavus and scoliosis. Phenotypic terms have been included if they have been previously described in at least two unrelated families. Image of nerve biopsy showing focally folded myelin adapted with permission from REF[73].

Figure 3. Proposed CMT diagnostic pathway integrating whole-genome sequencing. Initially, the clinical and neurophysiological assessment of the patient helps decide the Charcot–Marie–Tooth disease (CMT) phenotype, which guides the subsequent diagnostic pathways. If a patient presents with a classic CMT1 phenotype, testing for the 17p rearrangement needs to be carried out before proceeding to whole-genome sequencing (WGS). Sequenced data from both nuclear and mitochondrial genomes are then put through bioinformatic analysis and filtered for rare deleterious variants. The established phenotype of the patient (determined by neurophysiological studies, ancillary investigations and deep phenotyping) guides the application of relevant virtual gene panels to the filtered data, yielding candidate variants. These variants are interpreted with the aid of existing, published phenotype–genotype correlations and if a pathogenic variant is identified, a genetic diagnosis is confirmed. If no candidate variants or pathogenic variants are identified, a complex neuropathy virtual gene panel[7] is applied to the filtered genetic variants and any structural variants are analysed. If this approach does not yield a genetic diagnosis, the WGS data are analysed on a research basis to try and identify novel causes of CMT. If a genetic diagnosis is achieved, the details of the pathogenic variant(s) are entered into public domain variant databases or are published to enrich the existing phenotype–genotype correlations and aid analysis of future cases. HPO, Human Phenotype Ontology.

Glossary terms

Bottlenecked populations

Populations that at some point suffered a steep decline in size and subsequently recovered from a smaller pool of individuals.

Endogamy

Marriage within a community or restricted population.

Founder mutation

A mutation that occurred spontaneously in an ancestral allele at some point in the past and has been inherited by individuals in subsequent generations. In ethnically or geographically restricted populations that do not outbreed, the founder mutation will be observed at an increased frequency and might even be private to that population.

Consanguinity

Mating of closely blood-related individuals such as first or second cousins.

Penetrance

The frequency of individuals with an expressed phenotype among carriers of a genetic mutation. If some individuals with the genetic mutation never express a phenotype, the disease is described as having incomplete or reduced penetrance.

Sanger sequencing

Also referred to as first-generation sequencing, Sanger sequencing is the standard sequencing biochemistry described by Frederick Sanger's group in 1977 and was used to sequence the reference human genome. The technique uses fluorescently

labelled DNA chain terminators in the form of dideoxynucleotide triphosphates, which are randomly incorporated into chains of sequenced DNA during the polymerase chain reaction and arrest the elongation of the DNA sequence chain. These chains of DNA are subsequently aligned according to length and the fluorescent signal is read one nucleotide at a time, thus yielding a sequence read.

Multiple ligation-dependent probe amplification

A molecular genetic technique in which multiple pairs of oligonucleotide probes are used to hybridize to specific genomic sites. Each pair of probes is designed to hybridize to immediately adjacent sites and, once this is done, each pair of probes is ligated into a single fragment that is unique in length. Successfully ligated fragments are amplified by a polymerase chain reaction, separated according to size, detected and quantitatively analysed by comparison to reference values. Genomic sites with single nucleotide polymorphisms, point mutations and copy number variants would interfere with the ligation and/or proportional amplification of fragments and, therefore, will be highlighted during the fragment detection and analysis step.

Read depths

The number of times a specific nucleotide or base pair has been sequenced and read in a single sequencing experiment or series of experiments.

Phase

The position of two variants in a set of alleles in relation to one another. If the variants are on the same allele they are in '*cis* phase' and if they are on opposite alleles they are in '*trans* phase'.

Variant calling

The step in next-generation sequencing data analysis during which the sequenced data are reviewed or 'queried' for genetic variation compared with the reference genome, and variations at the nucleotide base level or over longer DNA sequences are identified and marked (or called).

Pleiotropy

A genetic principle describing the variety of phenotypic features in affected individuals with a specific gene mutation. A genetic modifier (which is different to the causal genetic mutation) can influence which phenotypic features manifest in the affected individual.

Haploinsufficiency

A phenomenon that occurs when the functional loss of one of two alleles of a specific gene causes a reduction in the amount of gene product, usually by 50%. Depending on the gene product and its function in the cell or tissue in which it is expressed, haploinsufficiency can lead to a disease state. If no disease ensues, the cell, tissue or organism harbouring the heterozygous allele loss is said to tolerate haploinsufficiency.

Coding synonymous variants

Single nucleotide changes in the protein-coding DNA that do not alter the amino acid in the translated protein.

Sequencing trace

The colour-coded peak chart, also known as the electropherogram, that is produced as a consequence of the read-out of the fluorescent signal in a Sanger sequencing reaction.

Microsatellite array

An array-based molecular genetic technique that uses specific oligonucleotide probes to genotype specific short repetitive DNA sequences (referred to as microsatellites) that are present at particular loci across the genome. The number of repeats in these microsatellites varies between individuals, so the unique combination of a set of repeat sequences can be used as a genetic tracker.

Runs of homozygosity

Contiguous genomic regions that are homozygous across all base pairs in an individual. This phenomenon occurs when the transmitted maternal and paternal alleles are identical and would have been inherited from a common ancestor at some point in the past.

Uniparental disomy

A phenomenon that occurs when offspring inherit two copies of a chromosome or part of a chromosome from one parent and no copies from the other parent. Uniparental disomy usually occurs as a random event during the stage of meiosis in gametogenesis.

Microarray comparative genomic hybridization

An array-based molecular genetic technique that compares the genome of interest to the reference genome for duplications or deletions of genetic material, also known as copy number variants (CNVs). This technique can detect CNVs as small as 50 kb.

Secondary findings

Genetic variants which may be of significance, that are identified in a patient but that are unrelated to the primary diagnostic question or the reason for which the sequencing data were generated. They may be discovered additionally (if there was intentional opportunistic screening e.g. for variants in cancer genes) or incidentally (if they were not sought).