A national diagnostic framework for patients with ultra-rare disorders: molecular genetic findings using phenotypic and sequencing data

Peter Krawitz (peter.krawitz@gmail.com)
Universität Bonn https://orcid.org/0000-0002-3194-8625

Article

Keywords:

Posted Date: March 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1416633/v1

Abstract

Most individuals with rare diseases first contact primary care physicians. Although efficient diagnostic routines exist for a subset of rare diseases, ultra-rare entities often require expert clinical knowledge or comprehensive genetic diagnostics, which poses structural challenges to public healthcare systems. To address these challenges, a novel structured diagnostic concept based on the presence of multidisciplinary expertise at centers for rare diseases (CRDs) that have been established at German university hospitals in recent years, was evaluated in a prospective study (TRANSLATE-NAMSE). Between January 2018 and December 2020, 5652 patients were enrolled in the study and were comprehensively assessed by multidisciplinary teams (MDTs) at ten CRDs. Exome sequencing (ES) was initiated for 268 adult and 1309 pediatric patients and partially complemented by additional molecular tests. Conclusive diagnoses were established in 494 individuals, covering 400 diagnostic-grade genes, suggesting ultra-rare disorders were enriched in this cohort. In addition, we describe 56 novel gene-phenotype associations, mainly in individuals with neurodevelopmental delay. A subcohort of 211 individuals was analyzed with the artificial intelligence–based PEDIA protocol, which integrates next-generation phenotyping on medical imaging and sequencing data. With the entire cohort data, we developed a tool to predict the diagnostic yield from the clinical features of a patient if advanced molecular testing strategies are applied.

Full Text

According to a recent analysis of the Orphanet database, rare diseases affect about 3 to 6% of the population, with a genetic cause assumed in 72%[1]. Although rare diseases thus represent a significant burden on the population, the average time to diagnosis for individuals with rare diseases is still around six years, and only a minority of patients receive a confirmatory molecular test result[2,3]. In Germany, a “national action alliance for people with rare diseases” (NAMSE) was established in 2010 with the aim of shortening the diagnostic odyssey of patients.

Because rare diseases often have a genetic etiology, genetic testing is a key element to speed up the time to diagnosis. In Germany, a statutory health insurance covers the healthcare costs of around 90% of the population, and genetic testing can be performed by academic or private laboratories. Until 2020, the reimbursement scheme allowed physicians to request chromosome analyses, comparative genomic hybridization arrays, and sequencing of single genes or small gene panels. This system has proven effective for some of the more common rare diseases and the ones that are caused by variants in well-defined molecular pathways (Supplemental Material). However, in total, these disorders only represent a small subset of all rare disorders. For highly heterogeneous phenotypes, such as neurodevelopmental delay, which has more than a thousand known disease genes to date, exome sequencing (ES) has already been shown to be more efficient than gene panels[4,5]. Due to structural differences between healthcare systems internationally, these results are not directly transferable to Germany, where ES can only be performed upon individual case applications. Therefore, the three-year prospective TRANSLATE-NAMSE study investigated the diagnostic yield of ES, its consequences on time to diagnosis, and cost-effectiveness as key parameters within the German national healthcare system. For this purpose, a

prospective cohort of patients suspected of having a genetically caused rare disease, was recruited at CRDs at German university hospitals and was analyzed using ES in four diagnostic centers. Here, we report on the molecular genetic findings of this study, which yielded a particularly high proportion of ultra-rare diagnoses and novel candidate disease genes. Participants of the study were also informed about the possibility of participating in an additional study (PEDIA-study) that involved the application of next-generation phenotyping technology, which uses artificial intelligence to analyze medical imaging data and to interpret genetic variants[6–8]. The results support the TRANSLATE-NAMSE framework that (i) enables the identification of ultra-rare disorders in a timely manner on a national level, and (ii) allows for the remaining undiagnosed patients to be incorporated into suitable research projects on an international level.

Results

Phenotypic characterization of the study cohort. In total, 5652 individuals with a suspected rare disorder were enrolled in TRANSLATE-NAMSE by CRDs at ten German university hospitals over a period of three years (2018–2020). MDTs including medical genetic and domain-specific clinical expertise evaluated information from the patients' health records and family histories and then made recommendations on how to proceed for each individual.

Here, we report on the patients (268 adults and 1309 children, total: 1577) who underwent ES following the recommendation of the MDT. Clinical features of these patients were encoded using the Human Phenotype Ontology (HPO) terminology, resulting in an average of five HPO terms per patient (Fig. 1a, Supplemental Fig. 1)[9]. In addition, a subset of 211 individuals consented to the analysis of their portrait photos by an artificial intelligence tool (PEDIA subset).

On the basis of the leading presenting symptom, each case was assigned to one of five major disease groups (Fig. 1b). The majority of children presented with neurodevelopmental disorders (51%), and the majority of adults with neurological or neuromuscular disorders (41%). Smaller proportions of cases presented with organ malformation, endocrine/metabolic, and immune/hematologic disorders. This is comparable to other large cohorts of undiagnosed patients[10]. However, the challenge of these assignments to disease groups can be illustrated by a comparison of the annotated phenotypic features: the correlation of higher-order HPO terms between the disease groups is considerable, indicating a high phenotypic overlap (Fig. 1c, Supplemental Material). This can also be visualized in the clinical feature space in which individuals are positioned according to their original HPO annotations: while most patients of the same disease group are close together, their clusters partially overlap (Fig. 1c). For instance, many patients with neuromuscular or neurodevelopmental disorders are phenotypically often so similar that an assignment to a disease group seems rather arbitrary. We therefore analyzed the diagnostic yield not only per disease group, but also based on phenotypic features.

Diagnostic yield and molecular findings. All lab results including rare, potentially causative variants were analyzed and discussed by the MDTs in context of the presenting phenotypes and facial dysmorphic

features. All such variants were classified according to standard guidelines[11,12], and their allelic contribution to disease was assessed.

In total, a molecular diagnosis could be established in 494 patients (36%) because pathogenic or likely pathogenic variants were found that explained the phenotype fully or partially. The diagnostic yield was 5 percentage points higher for children (32%) than for adults (27%), most probably indicating the higher likelihood of a monogenic disorder if the age of onset is early in life (Fig. 2a).

Patients with a neurodevelopmental disorder were more than twice as likely to receive a molecular diagnosis than patients with disorders of, for example, the endocrine system (Fig. 2a). However, since assignment to disease groups can be ambiguous, we also analyzed the influence of all phenotypic features on the diagnostic yield in a multivariate regression analysis.

Least absolute shrinkage and selection operator (LASSO) analysis yielded “dysfunction of higher cognitive abilities”, “hematological abnormalities”, and “ataxia” as very influential parameters for identifying a disease gene (Fig. 2b). Although our model was trained on the TRANSLATE-NAMSE cohort, it was also validated on an independent cohort with comparable results. We, therefore, made the model available as a web service that can be used to estimate the diagnostic yield of genetic testing given the phenotypic features of a patient (http://tnamse.de/)[13].

The diagnostic yield in patients that agreed to an evaluation of their molecular and clinical data, including analysis of portrait images by artificial intelligence, was 42%. Although this is substantially higher than for the remaining TRANSLATE-NAMSE cohort, this is most probably explained by an ascertainment bias, since patients with facial dysmorphism were more likely to participate in the PEDIA subset[14]. However, of note is the high sensitivity of the fully automated prioritization pipeline that lists the correct disease gene among the top ten suggestions in four out of five cases[7] (Fig. 2c). The support of the PEDIA workflow could further be increased by including gestalt scores of a recent algorithmic update that focuses on ultra-rare phenotypes[8]. The AI support not only speeds up data analysis but also yields additional evidence for variant classification, particularly if the gestalt is quantified as highly similar in a phenotype of high distinctiveness[11,15].

For 18 cases that were classified as uncertain or unsolved after initial ES, functional assays such as analysis of the methylome (n = 4), transcriptome (n = 11), or proteome (n = 3), were conducted for further classification. Proteome analyses were particularly informative in three cases where this strategy was used, highlighting the importance of variant validation strategies in diagnostics (Supplementary Case Reports)[16–18]. Epigenetic signatures could clarify the status of *de novo* missense variants as likely to be benign or pathogenic, as exemplified by a case with a missense variant in *KMT2D (*Supplementary Case Reports)[19,20]

Mode of inheritance and recessive disease burden. In accordance with previous reports on comparable cohorts, 214 (44%) of the solved cases were due to *de novo* variants (Fig. 3a). In three families,

establishing the diagnosis was particularly challenging due to mosaicism (Supplemental Material). In one of these families, the same pathogenic variant in *PUF60* was identified as the cause of developmental delay in two affected brothers. Because the variant was not detectable in the exome data of either parent, the presence of the presumable gonadal mosaicism could only be suggested to the special family history. Previously reported proportions of parental mosaicism below 1% should therefore only be regarded as a lower bound[21–23].

The second-largest fraction of diagnoses was due to an autosomal recessive (AR) mode of inheritance, with autozygosity being an important covariate. We computed the homozygosity and used a threshold of 2% to assign patients to a group of high (n = 126) or low (n = 262) autozygosity[24]. Although there was no significant difference in the diagnostic yield between the groups (low autozygosity 29% vs. high autozygosity 28%), the composition of the modes of inheritance differed considerably (Fig. 3b); the relative contribution of homozygous variants was considerably higher in the high autozygosity group (44%) than in the low autozygosity group (3%) (OR 15.9). In contrast, the proportion of *de novo* variants contributing to disease was 50% in the low autozygosity group compared with 22% in the high autozygosity group (OR 2.3).

Because the *de novo* mutation rate depends on parental age but not on autozygosity, the disease prevalence attributable to such variants should be comparable in both groups and can be used for normalization (Fig. 3c). For an inbreeding coefficient above 2%, this suggests a recessive disease burden seven times that for those with lower inbreeding coefficients, which is in agreement with previous reports [24–26].

In eight individuals, representing roughly 2% of all solved cases, we reported two molecular diagnoses of distinct or overlapping disease phenotypes in a single family, which is in agreement with earlier reports[27]. This group was also enriched for high autozygosity and recessive disorders, which also concurs with earlier reports (Supplemental Table dual diagnosis)[28].

In addition to the relatedness of two healthy parents, a more accurate estimate for the disease risk in offspring needs to incorporate the number of heterozygous pathogenic variants in recessive genes, which can vary considerably depending on demographics[29–32]. We found that 89 of the 116 variants that we reported in recessive disease genes, would also have been classified as pathogenic if they were identified in healthy individuals[33]. That also means that ES as an expanded carrier screen would have pointed to an elevated risk in 77% of the couples in our cohort that had an offspring affected by a recessive disease[34].

Novel diagnostic-grade genes and candidates. For all 494 individuals with a molecular diagnosis, we reported in total 546 distinct pathogenic or likely pathogenic variants in 364 different diagnostic-grade genes (DGGs) (Supplemental Material). We estimated the incidences of the associated disorders and tracked the years in which they were first described. As a proxy for incidence, we ordered all known DGGs according to the number of case submissions in ClinVar and plotted the number of variants in the TRANSLATE-NAMSE cohort corresponding to these genes (Fig. 4a). The first quartile contains clinical

reports of 24 patients in 10 different DGGs, whereas the last quartile features 113 DGGs, with most diagnosis in the latter group corresponds only to a single patient. In comparison to other cohorts of comparable size, this distribution is shifted to the right, suggesting a significant enrichment for ultra-rare disorders in the TRANSLATE-NAMSE cohort[1,5,35,36] (Supplemental Material). Almost half of the diagnoses that we established have only become possible in the last decade (Fig. 4b), demonstrating the huge progress in medical genetics due to high-throughput sequencing and emphasizing the importance of data reanalysis[37,38]. Cases in which no diagnosis could be established in a known DGG were included in national and international studies for the discovery of novel disease etiologies via the MatchMaker Exchange (MME) Network[39,40]. Variants with a high likelihood of being disease-causing, for example those with loss of function or high pathogenicity scores or that arose *de novo*, were shared through MME to identify similar patients[41,42]. In 64 cases, we identified indications for novel disease relationships in 55 genes, most of them related to a neurodevelopmental phenotype. Of this set, 23 candidates achieved medium evidence and 32 high evidence and are currently under further investigation. Fifteen genes have subsequently reached DGG status. We briefly describe a few exemplary cases below, most of which have already been published in detail elsewhere[43–52], and provide a comprehensive list of all patients in the Supplemental Material. In *SMARCA5*, a gene coding for a chromatin remodeler, we identified *de novo* variants in two patients with neurodevelopmental delay and similar dysmorphic features. The phenotype-gene association was strengthened by the identification of ten additional cases from other cohorts, and rescue experiments with wildtype transcripts in *Drosophila* suggested a hypomorphic effect of the variants[47]. In another individual with learning disabilities, autism, dystonia, and intention tremor, we identified a *de novo* missense variant in *KCNN2*, a small-conductance calcium-activated potassium channel. Interestingly, a preexisting rat model with missense substitution identical to that found in the affected individual partially mirrors this phenotype with abnormal locomotor activity and tremor. The identification of nine additional individuals from other cohorts and results of functional analysis of the variants on channel function established *KCNN2* as a dominant disease gene for a neurodevelopment movement disorder[50]. A recognizable syndrome with multiorgan manifestations could be delineated for biallelic truncating variants in *MAPKAPK5* (MAPK-activated protein kinase 5). Patients presented with severe developmental delay, variable brain anomalies, congenital heart defects, facial dysmorphism, and a distinct type of synpolydactyly with an additional hypoplastic digit between the fourth and fifth digits of hands and/or feet[43]. Aside from novel neurodevelopmental disorders, *OAS1*, which encodes a type I interferon–induced, intracellular double-stranded RNA (dsRNA) sensor that is required for antiviral defense, was established as a DGG. We identified a *de novo* gain-of-function variant that caused dsRNA-independent activity of OAS1 in a patient with immunodeficiency, pulmonary alveolar proteinosis, and phospholipid accumulation. The hyperinflammatory disorder was cured by allogeneic hematopoietic cell transplantation[45]. A homozygous frameshift variant in *CYHR1*, encoding a protein with to-date-unknown function but high cerebral and cerebellar expression, was identified in a girl from non-consanguineous parents who presented with global developmental delay and hyperreflexia. Interestingly, the variant was inherited from the unaffected father and was homozygous in the patient due to paternal uniparental isodisomy of chromosome 8 as confirmed by exome-wide SNP analysis. RNA sequencing revealed

significantly lower *CYHR1* gene expression, highlighting *CYHR1* as an excellent candidate gene for AR non-syndromic intellectual disability.

In comparison with pathogenic variants in previously known DGGs, there was a higher proportion of missense variants in our candidate gene set, most likely because classification is more challenging (Fig. 4c). In addition to standard pathogenicity scoring approaches, we therefore also made use of the structural predictions that became recently available from AlphaFold 2 and found high conservation of amino acids that were closer than six Angstroms to the mutated site, potentially indicating a higher likelihood of pathogenicity (Supplemental Material)[53].

Diagnoses with causal therapeutic implications. For five patients in the TRANSLATE-NAMSE cohort with a molecular diagnosis personalized treatments, or therapies directed against the mechanism of the disease could be initiated.[54] A patient with metachromatic leukodystrophy (MLD) due to pathogenic variants in arylsulfatase alpha (*ARSA*) was treated with autologous CD34 + cells that were transduced *ex vivo* using a lentiviral vector encoding *ARSA*[55]. The gene therapeutic approach with atidarsagene autotemcel has been authorized by EMA in the EU since 17 December 2020. A patient with pyruvate dehydrogenase E1-alpha deficiency due to a *de novo* variant in *PDHA1* and another patient with GLUT1-deficiency due to pathogenic variants in *SLC2A1* were treated with a ketogenic diet. In a patient with cerebral creatine deficiency syndrome 1, due to a misssense substitution in *SLC6A8*, supplementation with creatine was started. In a patient with congenital disorder of glycosylation of type IIc, due to a homozygous missense variant in *SLC35C1* the fucosylation deficiency was treated by oral fucose supplementation[56].

Discussion

The International Rare Disease Research Consortium has set a goal that, by 2027, all patients who come to medical attention with a suspected rare disease should be diagnosed within one year if their disorder is known in the medical literature [57]. The reduction of the time to diagnosis from the current six years to less than one year is particularly important because of increasing number of approved therapies for rare diseases that produce better patient outcomes with earlier treatment[58]. This will require frameworks in the healthcare system that are dedicated to patients with rare diseases, and evaluating such a concept was the very objective of TRANSLATE-NAMSE. We found that a combination of clinical assessment by an MDT and subsequent ES is well suited to accelerate the time to diagnosis, and patients with ultra-rare genetic disorders benefited particularly from exome trio sequencing, confirming findings from other healthcare systems[4,36,59,60]. In our cohort, molecular diagnoses also resulted in a change of clinical management with a causal or even curative approach to therapy

The large number of individuals with variants associated with a novel gene-disease association (12% of solved cases) highlights that the concept of a clear separation between diagnostics and research cannot and should not be applied for patients with ultra-rare disorders.

Despite the described benefits of the advanced multi-omics approaches, it cannot be ignored that they are costly. Therefore, these tests also compete with other analyses in a complex healthcare system, and estimating their efficacy in establishing a diagnosis has to be considered, too. In our study, we showed how artificial intelligence–powered next-generation phenotyping could become an integral part of the diagnostic workup and contribute to efficient data analysis. For all currently undiagnosable individuals, it is also important that they can enter globally coordinated research initiatives if they wish. With TRANSLATE-NAMSE, we established a framework in Germany that is suitable to organize this contribution on the national level, and we hope to grow this network in the future.

Wir können nicht forschung von research diagnostik trennen. Trios sind für konsanguine wichtig. ¼ haben de novo

Online content and code availability

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgments, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at http://tnamse.de/.

References

1. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur. J. Hum. Genet. **28**, 165–173 (2020).
2. Blöß, S. *et al.* Diagnostic needs for rare diseases and shared prediagnostic phenomena: Results of a German-wide expert Delphi survey. PLoS One **12**, e0172532 (2017).
3. Boycott, K. M. *et al.* International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. Am. J. Hum. Genet. **100**, 695–705 (2017).
4. Stark, Z. *et al.* Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement. Genet. Med. **19**, 867–874 (2017).
5. 100,000 Genomes Project Pilot Investigators *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N. Engl. J. Med. **385**, 1868–1880 (2021).
6. Gurovich, Y. *et al.* Identifying facial phenotypes of genetic disorders using deep learning. Nat. Med. **25**, 60–64 (2019).
7. Hsieh, T.-C. *et al.* PEDIA: prioritization of exome data by image analysis. Genet. Med. **21**, 2807–2814 (2019).
8. Hsieh, T.-C. *et al.* GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. Nat. Genet. (2022) doi:10.1038/s41588-021-01010-x.
9. Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. **83**, 610–615 (2008).
10. Wise, A. L. *et al.* Genomic medicine for undiagnosed diseases. Lancet **394**, 533–540 (2019).

11. 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).
12. Miller, D. T. *et al.* ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. **23**, 1381–1390 (2021).
13. Peng, C. *et al.* CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. NAR Genom Bioinform **3**, (2021).
14. Pantel, J. T. *et al.* Efficiency of Computer-Aided Facial Phenotyping (DeepGestalt) in Individuals With and Without a Genetic Syndrome: Diagnostic Accuracy Study. J. Med. Internet Res. **22**, e19263 (2020).
15. Tavtigian, S. V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet. Med. (2018) doi:10.1038/gim.2017.210.
16. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. **9**, (2017).
17. Murdock, D. R. *et al.* Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. J. Clin. Invest. (2020) doi:10.1172/JCI141500.
18. Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat. Med. **25**, 911–919 (2019).
19. Aref-Eshghi, E. *et al.* Genomic DNA Methylation Signatures Enable Concurrent Diagnosis and Clinical Genetic Variant Classification in Neurodevelopmental Syndromes. Am. J. Hum. Genet. **102**, 156–174 (2018).
20. Mirza-Schreiber, N. *et al.* Blood DNA methylation provides an accurate biomarker of KMT2B-related dystonia and predicts onset. Brain (2021) doi:10.1093/brain/awab360.
21. Cao, Y. *et al.* A clinical survey of mosaic single nucleotide variants in disease-causing genes detected by exome sequencing. Genome Med. **11**, 48 (2019).
22. Gambin, T. *et al.* Low-level parental somatic mosaic SNVs in exomes from a large cohort of trios with diverse suspected Mendelian conditions. Genet. Med. **22**, 1768–1776 (2020).
23. Wright, C. F. *et al.* Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. Nat. Commun. **10**, 2985 (2019).
24. Martin, H. C. *et al.* Quantifying the contribution of recessive coding variation to developmental disorders. Science **362**, 1161–1164 (2018).
25. Fridman, H. *et al.* The landscape of autosomal-recessive pathogenic variants in European populations reveals phenotype-specific effects. *Cold Spring Harbor Laboratory* 2020.11.16.384206 (2020) doi:10.1101/2020.11.16.384206.
26. Hu, H. *et al.* Genetics of intellectual disability in consanguineous families. Mol. Psychiatry (2018) doi:10.1038/s41380-017-0012-2.

27. Posey, J. E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. N. Engl. J. Med. **376**, 21–31 (2017).
28. Mitani, T. *et al.* High prevalence of multilocus pathogenic variation in neurodevelopmental disorders in the Turkish population. Am. J. Hum. Genet. (2021) doi:10.1016/j.ajhg.2021.08.009.
29. Gao, Z., Waggoner, D., Stephens, M., Ober, C. & Przeworski, M. An estimate of the average number of recessive lethal mutations carried by humans. Genetics **199**, 1243–1254 (2015).
30. Narasimhan, V. M. *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. Science **352**, 474–477 (2016).
31. Chakraborty, R. & Chakravarti, A. On consanguineous marriages and the genetic load. Hum. Genet. **36**, 47–54 (1977).
32. La Rocca, L. A. *et al.* Drop of Prevalence after Population Expansion: A lower prevalence for recessive disorders in a random mating population is a transient phenomenon during and after a growth phase. *bioRxiv* 2021.09.29.462290 (2021) doi:10.1101/2021.09.29.462290.
33. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. Nucleic Acids Res. **48**, D835–D844 (2020).
34. Antonarakis, S. E. Carrier screening for recessive disorders. Nat. Rev. Genet. **20**, 549–561 (2019).
35. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature **586**, 757–762 (2020).
36. Turro, E. *et al.* Whole-genome sequencing of patients with rare diseases in a national health system. Nature **583**, 96–102 (2020).
37. Wright, C. F. *et al.* Evaluating variants classified as pathogenic in ClinVar in the DDD Study. Genet. Med. **23**, 571–575 (2021).
38. Wright, C. F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. Genet. Med. (2018) doi:10.1038/gim.2017.246.
39. Philippakis, A. A. *et al.* The Matchmaker Exchange: a platform for rare disease gene discovery. Hum. Mutat. **36**, 915–921 (2015).
40. Sobreira, N. L. M. *et al.* Matchmaker Exchange. Curr. Protoc. Hum. Genet. **95**, 9.31.1–9.31.15 (2017).
41. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. Nature **581**, 434–443 (2020).
42. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. Nat. Genet. **46**, 944–950 (2014).
43. Horn, D. *et al.* Biallelic truncating variants in MAPKAPK5 cause a new developmental disorder involving neurological, cardiac, and facial anomalies combined with synpolydactyly. Genet. Med. **23**, 679–688 (2021).
44. Tan, N. B. *et al.* Recurrent de novo missense variants in GNB2 can cause syndromic intellectual disability. J. Med. Genet. (2021) doi:10.1136/jmedgenet-2020-107462.

45. Magg, T. *et al.* Heterozygous OAS1 gain-of-function variants cause an autoinflammatory immunodeficiency. Sci Immunol **6**, (2021).
46. den Hoed, J. *et al.* Mutation-specific pathophysiological mechanisms define different neurodevelopmental disorders associated with SATB1 dysfunction. Am. J. Hum. Genet. **108**, 346–356 (2021).
47. Li, D. *et al.* Pathogenic variants in *SMARCA5*, a chromatin remodeler, cause a range of syndromic neurodevelopmental features. Sci Adv **7**, (2021).
48. Thaventhiran, J. E. D. *et al.* Whole-genome sequencing of a sporadic primary immunodeficiency cohort. Nature **583**, 90–95 (2020).
49. Brugger, M. *et al.* A homozygous truncating variant in CCDC186 in an individual with epileptic encephalopathy. Ann Clin Transl Neurol **8**, 278–283 (2021).
50. Mochel, F. *et al.* Variants in the SK2 channel gene (KCNN2) lead to dominant neurodevelopmental movement disorders. Brain **143**, 3564–3573 (2020).
51. Stenton, S. L. *et al.* Impaired complex I repair causes recessive Leber's hereditary optic neuropathy. J. Clin. Invest. **131**, (2021).
52. Strande, N. T. *et al.* Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am. J. Hum. Genet. **100**, 895–906 (2017).
53. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. Nature **596**, 583–589 (2021).
54. Bick, D. *et al.* An online compendium of treatable genetic disorders. Am. J. Med. Genet. C Semin. Med. Genet. (2020) doi:10.1002/ajmg.c.31874.
55. Capotondo, A. *et al.* Safety of arylsulfatase A overexpression for gene therapy of metachromatic leukodystrophy. Hum. Gene Ther. **18**, 821–836 (2007).
56. Feichtinger, R. G. *et al.* A spoonful of L-fucose-an efficient therapy for GFUS-CDG, a new glycosylation disorder. EMBO Mol. Med. **13**, e14332 (2021).
57. Austin, C. P. *et al.* Future of Rare Diseases Research 2017–2027: An IRDiRC Perspective. Clin. Transl. Sci. **11**, 21–27 (2018).
58. Tambuyzer, E. *et al.* Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. Nat. Rev. Drug Discov. **19**, 93–111 (2020).
59. Retterer, K. *et al.* Clinical application of whole-exome sequencing across clinical indications. Genet. Med. **18**, 696–704 (2016).
60. Kingsmore, S. F. *et al.* A Randomized, Controlled Trial of the Analytic and Diagnostic Performance of Singleton and Trio, Rapid Genome and Exome Sequencing in Ill Infants. Am. J. Hum. Genet. **105**, 719–733 (2019).

Figures

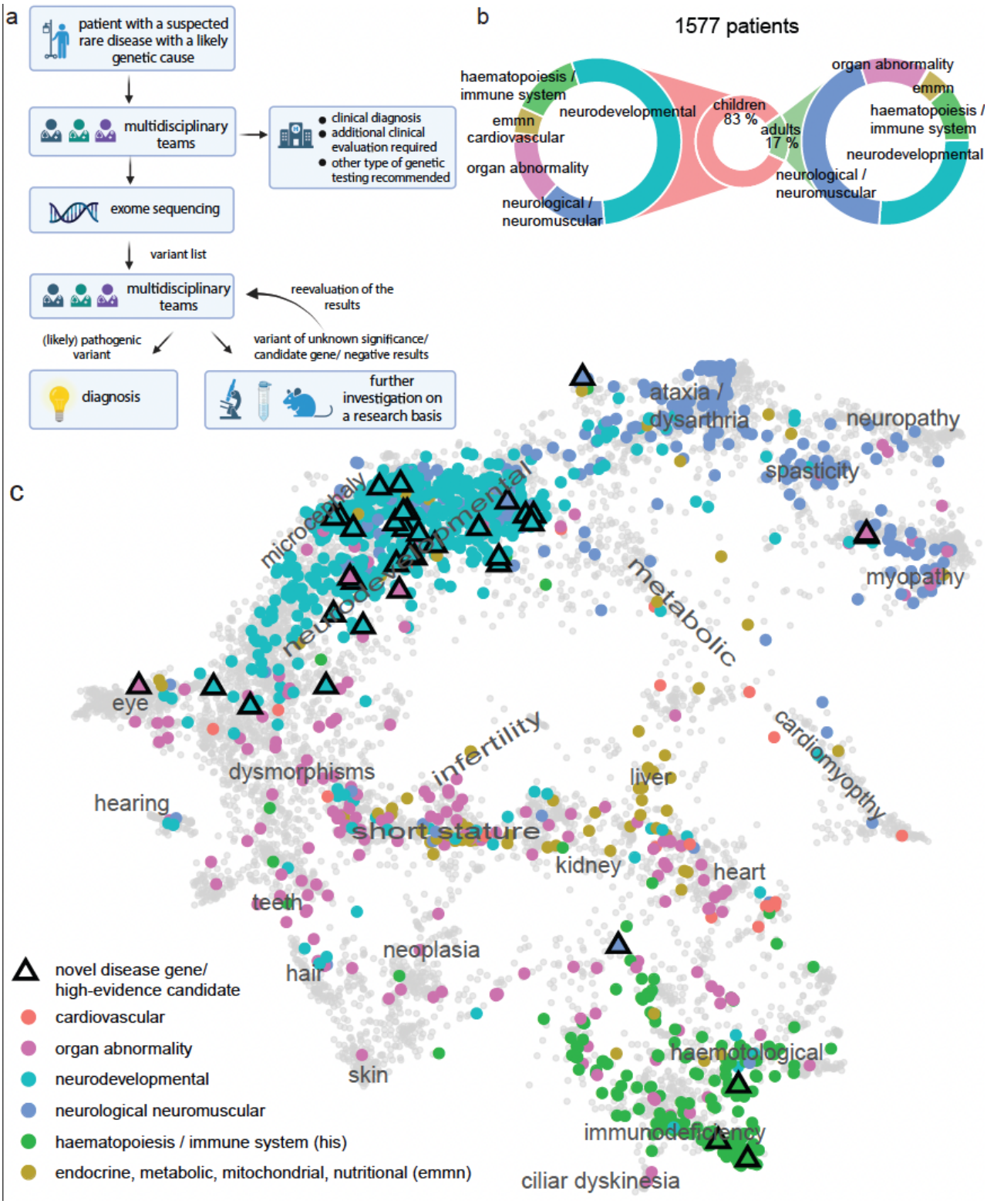

Figure 1

Workflow in the TRANSLATE-NAMSE project and phenotypes in which exome sequencing was conducted. a) Patients with suspected rare diseases were presented to a multidisciplinary team (MDT) and deeply phenotyped using terminology of the Human Phenotype Ontology (HPO). If a genetic diagnosis was assessed as likely, exome sequencing (ES) was initiated. Likewise, the MDT evaluated the molecular findings and could order additional analysis for variants of unknown significance or variants in

potentially novel disease genes. b) ES was performed primarily in children. The indications for exome sequencing in children were primarily neurodevelopmental disorders, whereas neurological/neuromuscular disorders were the predominant indication in adults. Cardiovascular, endocrine, metabolic mitochondrial, nutritional (emmn) and hematopoiesis and immune system (his) were the smallest disease categories. c) The phenotypic similarities of the patients which are encoded in their HPO terms, are visualized with UMAP. The clinical phenotype space is initially defined by all OMIM diseases using their HPO annotations (grey dots in the background). Each individual is color-coded by the disease-group according to the leading symptom. Note the overlap of patients in the neurodevelopmental and neuromuscular groups (blue and purple clusters) indicating high phenotypic similarity. This also means a definite assignment to a group is impossible. The patients contributing to the identification of a novel gene-phenotype association of high evidence are represented by triangles.

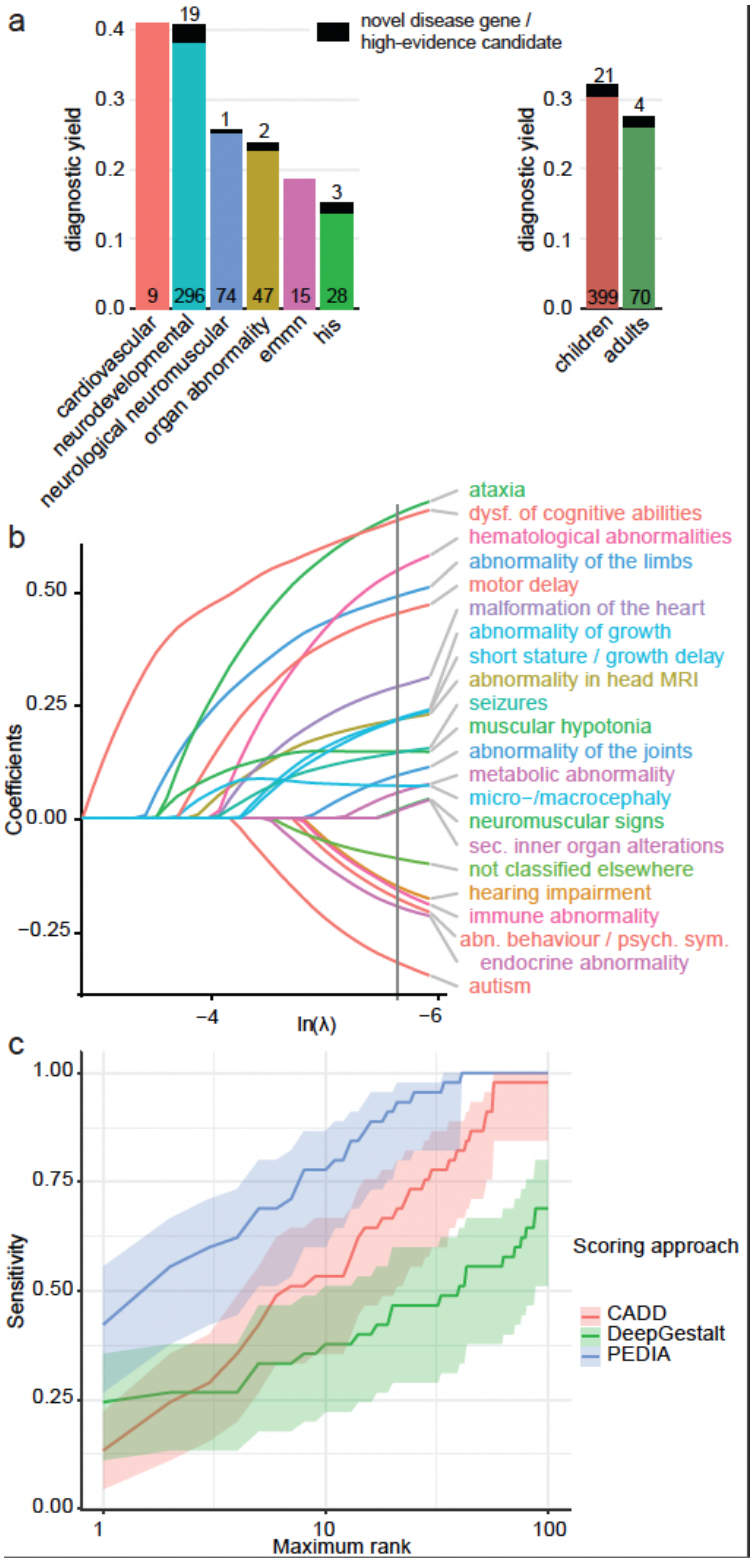

Figure 2

Machine learning identifies features relevant to the diagnostic yield and can support variant prioritization. a) The diagnostic yield differs by disease category and age group (adult/child). For all disease categories except cardiovascular the diagnostic yield could be increased due to novel gene-phenotype associations (dark colored tip of the bar). b) The coefficient paths of regression analysis using the Least Absolute Shrinkage and Selection Operator (LASSO) are shown. The more to the left [lower ln(λ)] a coefficient path

starts to deviate from the x-axis, the more informative the corresponding feature is for predicting the diagnostic yield. Features with positive coefficients increase the diagnostic yield, in contrast to features with negative coefficients whose presence makes it less likely that a monogenic cause can be found. Here, e.g. dysfunction of higher cognitive abilities and ataxia are clinical features that are relatively predictive of an increased diagnostic yield (clinical features are colored according to their higher-order HPO groups, see Supplemental Material and Methods for details). (c) Three approaches for variant prioritization were tested in solved cases of the PEDIA cohort (n=88): a molecular pathogenicity score (CADD), a gestalt score from portrait analysis (DeepGestalt), and a combination of molecular and phenotypic information (PEDIA). All DGGs were ordered by the respective variant prioritization method and the proportion of cases detected with the correct DGG (sensitivity) is shown as a function of the number of DGGs considered beginning at the top score. Note that bootstrapped confidence intervals are indicated by lighter shades around the lines.

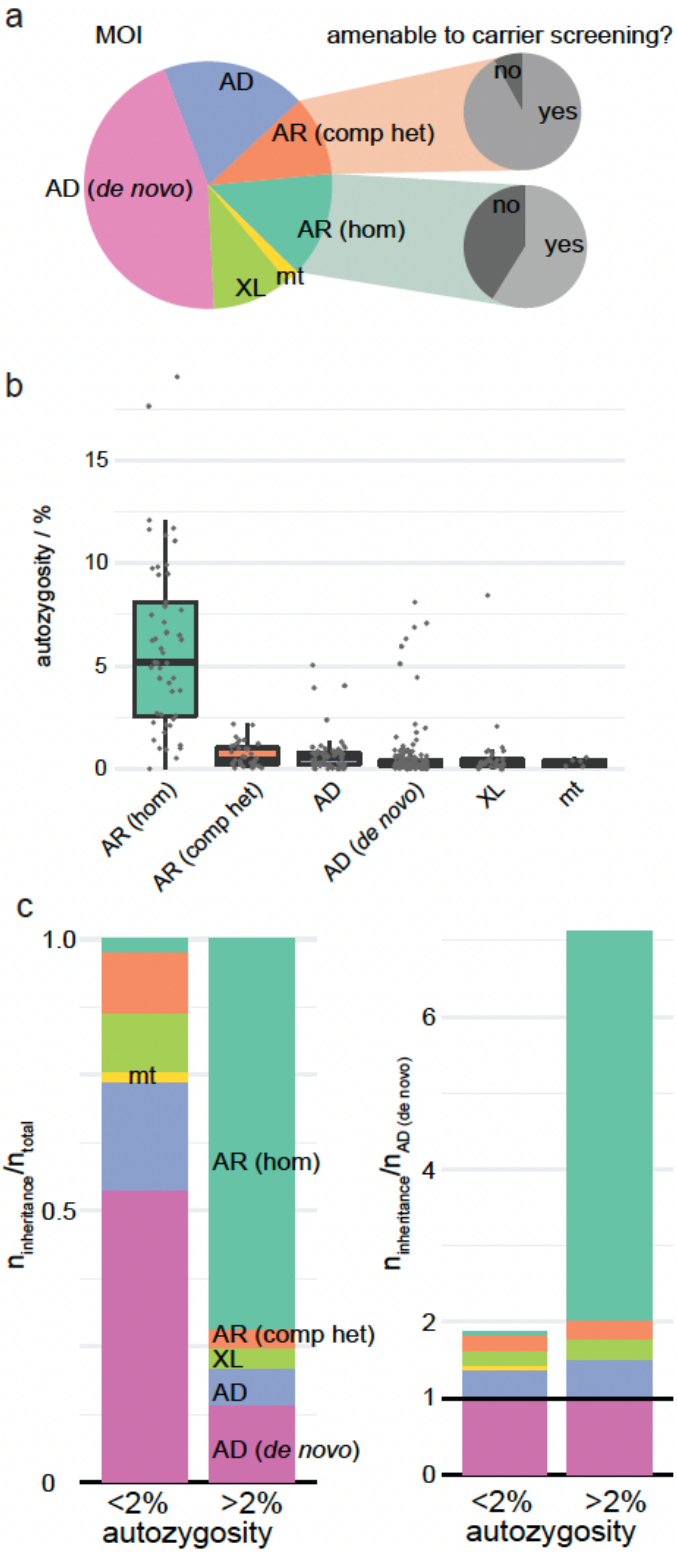

Figure 3

Mode of inheritance and disease burden depend on autozygosity. a) Pie chart showing the distribution of modes of inheritance (MOI) in all solved cases. Disease-causing variants most frequently occurred *de novo*. At least 73% of all autosomal recessive cases could have been identified by extended carrier screening (slice) b) Box plots of autozygosity for each MOI. Individuals are additionally shown as gray dots. The autozygosity is considerably increased in individuals with autosomal recessive inheritance and

homozygous variants. However, several cases with high autozygosity were also caused by AD (de novo) and AR (comp het) (c) Bar graphs illustrating MOI in individuals with low (<2%) and high (>2%) autozygosity. On the right, the AD *de novo* rate has been used for normalization. Individuals with high autozygosity have a higher relative burden of recessive diseases, mainly due to homozygous pathogenic variants. AD: autosomal dominant inheritance, variant inherited or unknown origin; AD (*de novo*): autosomal dominant inheritance with *de novo* variant; AR (comp het): autosomal recessive inheritance with compound heterozygous variants; AR (hom): autosomal recessive inheritance with homozygous variant; mt: mitochondrial inheritance; XL: X-linked inheritance.

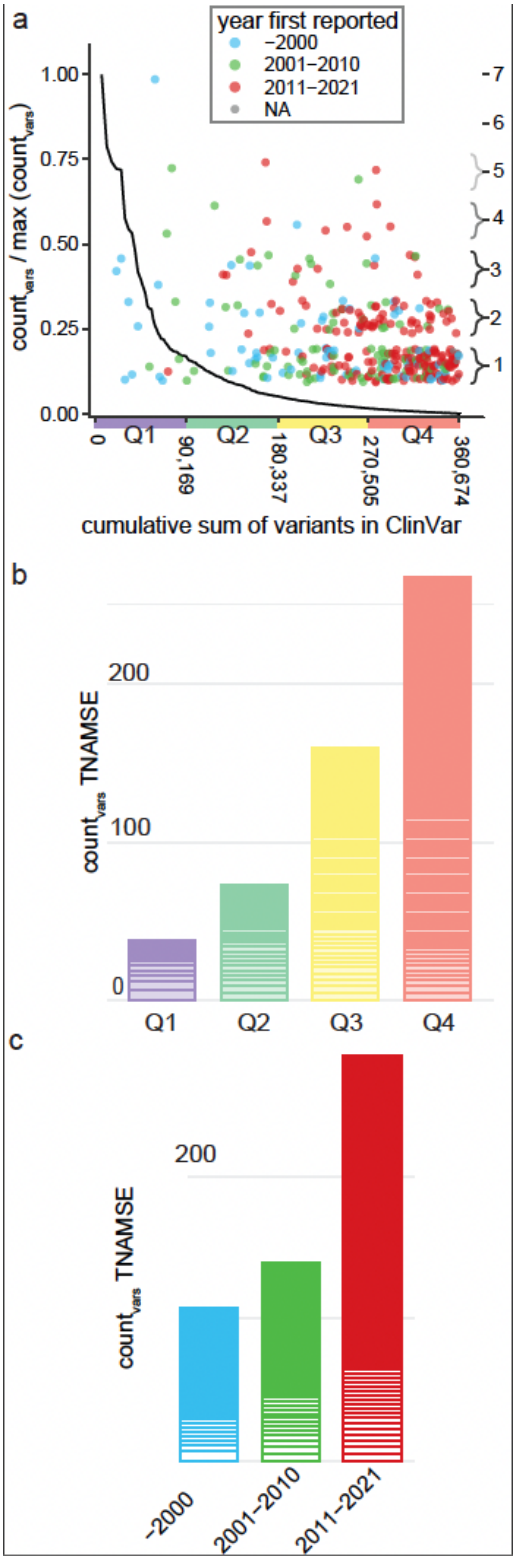

Figure 4

The variants reported in TRANSLATE-NAMSE cause mainly ultrarare disorders that were associated with a gene in the last decade. a) The normalized number of variants per gene in ClinVar or TRANSLATE-NAMSE was plotted against the cumulative number of pathogenic/likely pathogenic variants in ClinVar. The black line corresponds to the genes in ClinVar and shows the arrangement of genes in the graph - from left to right, the number of variants per gene decreases. Genes with diagnostic variants in

TRANSLATE-NAMSE were plotted as dots and were colored according to the year in which the gene was first described as a disease gene. (b) Variant counts in TRANSLATE-NAMSE in genes with high (Q1) to low (Q4) variant counts per gene in ClinVar. The genes in Q1-Q4 each cover approximately 1/4 of the likely or confirmed pathogenic variants in ClinVar, as shown in a) on the x-axis. Note that variants in the same genes are grouped in horizontal blocks. c) Bar graph showing the number of variants relative to the time interval in which the gene was first described as a disease gene.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AuthorArrangerTNAMSE.docx
- TNAMSESupplementalMaterial.docx
- SupplementalMaterialallcases.xlsx
- DGGCandidatesMainSupplement.xlsx
- flatKrawitzepc.pdf
- flatKrawitzrs.pdf