The American Journal of Human Genetics, Volume *103*

Supplemental Data

Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes

Jung Hoon Son, Gangcai Xie, Chi Yuan, Lyudmila Ena, Ziran Li, Andrew Goldstein, Lulin Huang, Liwei Wang, Feichen Shen, Hongfang Liu, Karla Mehl, Emily E. Groopman, Maddalena Marasa, Krzysztof Kiryluk, Ali G. Gharavi, Wendy K. Chung, George Hripcsak, Carol Friedman, Chunhua Weng, and Kai Wang

Figure S1: A comparison of two sets of manually compiled phenotype terms in prioritization of disease genes with causal variants. We obtained phenotype terms from either review of the entire EHRs by an expert (Expert Chart Review) or review of a single clinical note by an expert (Expert Single Genetics Note Extraction), then analyzed the terms by Phenolyzer and assessed where the genes with causal variants rank among all candidate genes. Expert Single Genetics Note Extraction has comparable performance with Expert Chart Review.

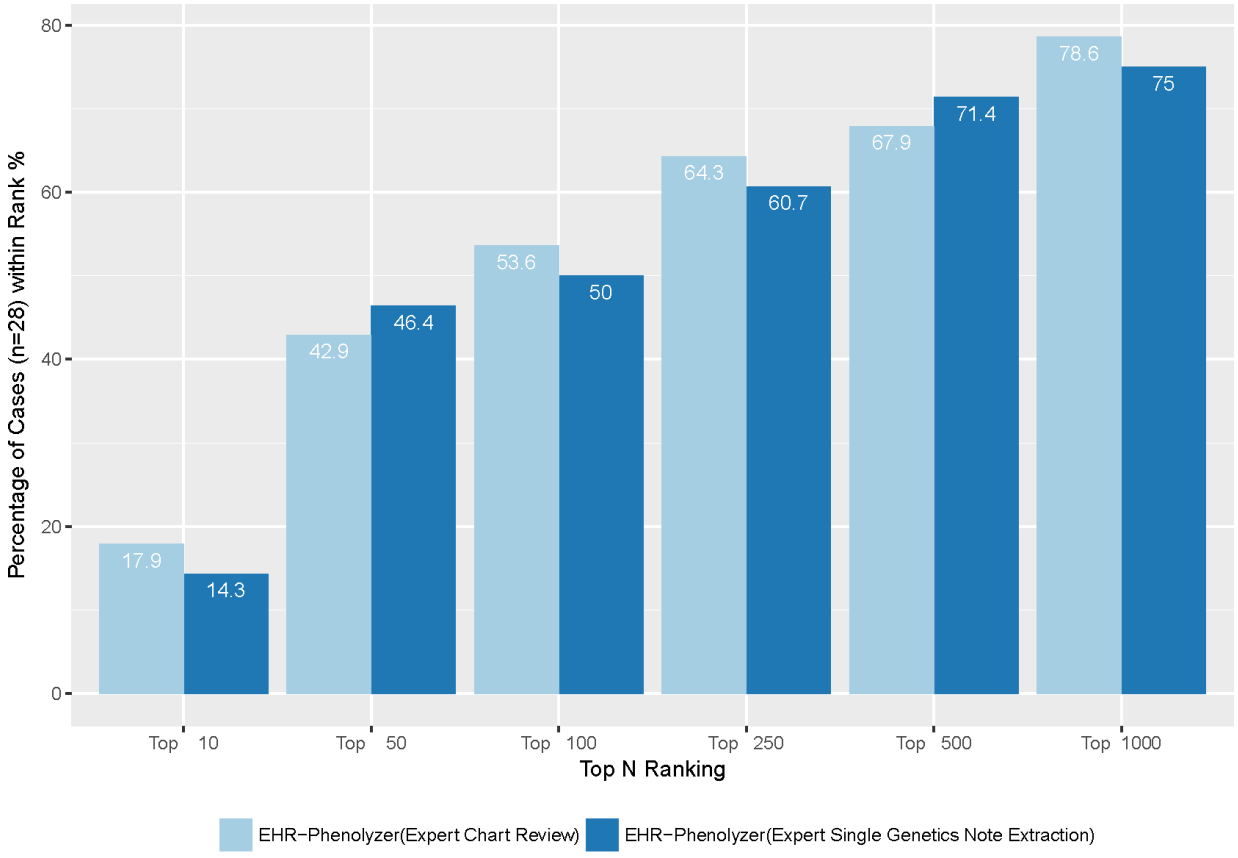


Figure S2: Focused analysis on ~5000 OMIM genes (as opposed to all ~20,000 genes in the human genome) to assess the performance of the three methods in diagnostic settings. We note that two positive diagnosis involve genes (*MYH10* and *NAA15*) that are not yet documented in OMIM for Mendelian diseases.

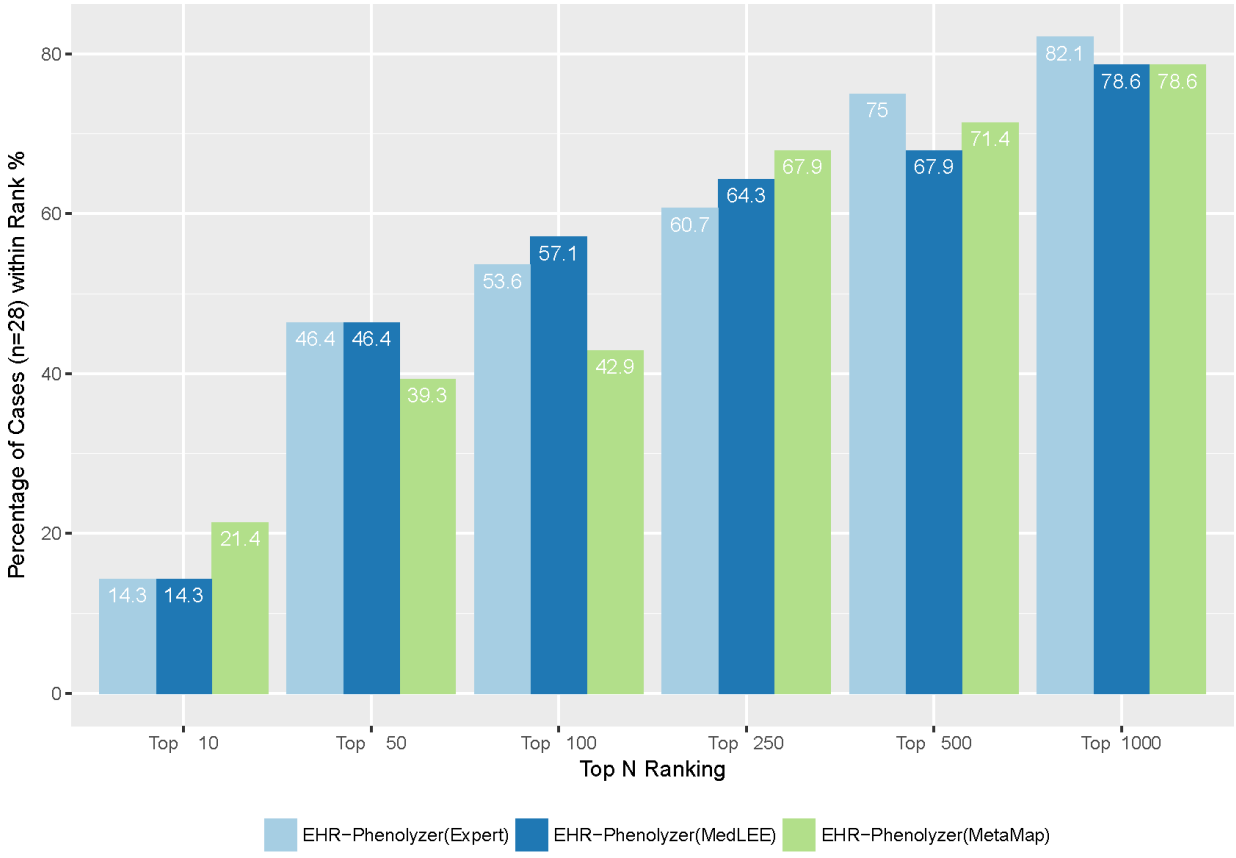
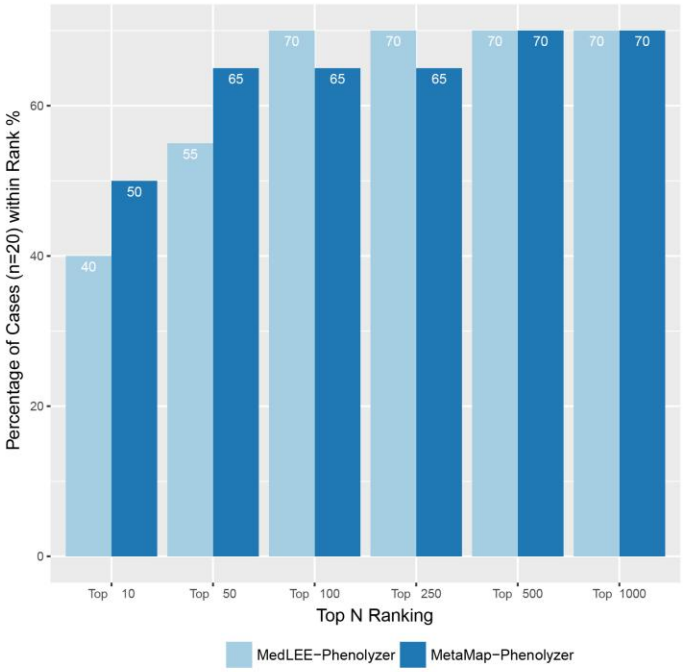


Figure S3. Analysis of 20 individuals affected with chronic kidney disease (CKD) who had a positive genetic diagnosis from whole-exome sequencing. (A) EHR-Phenolyzer (based on both MedLEE and MetaMap) can accurately rank the genes with causal variants within top 10 for nearly half of the affected individuals. (B) Phenotype-based hierarchical clustering of 20 individuals with CKD (row: phenotypes, column: affected individuals with diagnostic genes). Notes: only affected individuals with the genes ranked within top 50 and the phenotypes found in at least two affected individuals but not all were used in the clustering analysis.

A



B

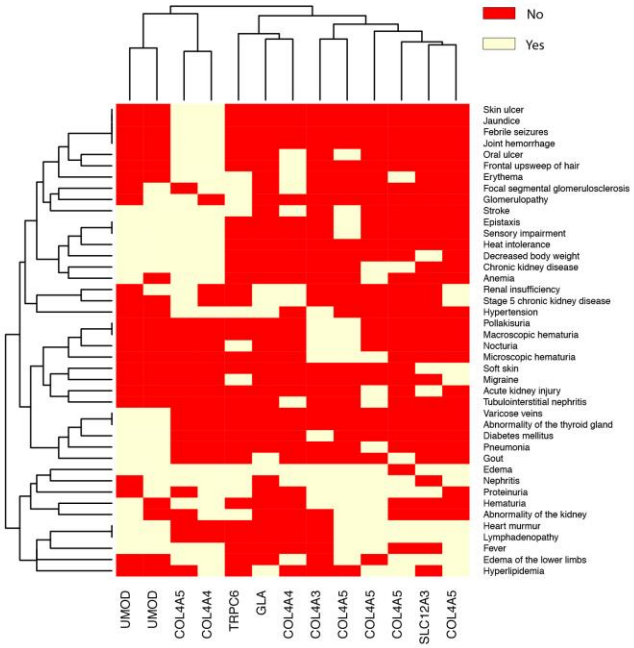


Figure S4. Expert and MedLEE methods perform much better than MetaMap in an individual with generalized seizure. (A) Only one phenotype term is shared among the phenotype terms extracted by the three methods. (B) Expert and MedLEE rank genes with causal variants much better (#1, #18) than MetaMap (#118). (C, D) The network of connected candidate genes and phenotype terms extracted by expert (C) or by MedLEE (D). The gene *SCN1A* carries a causal mutation. The size of the circle around gene name is correlated with the Phenolyzer ranking.

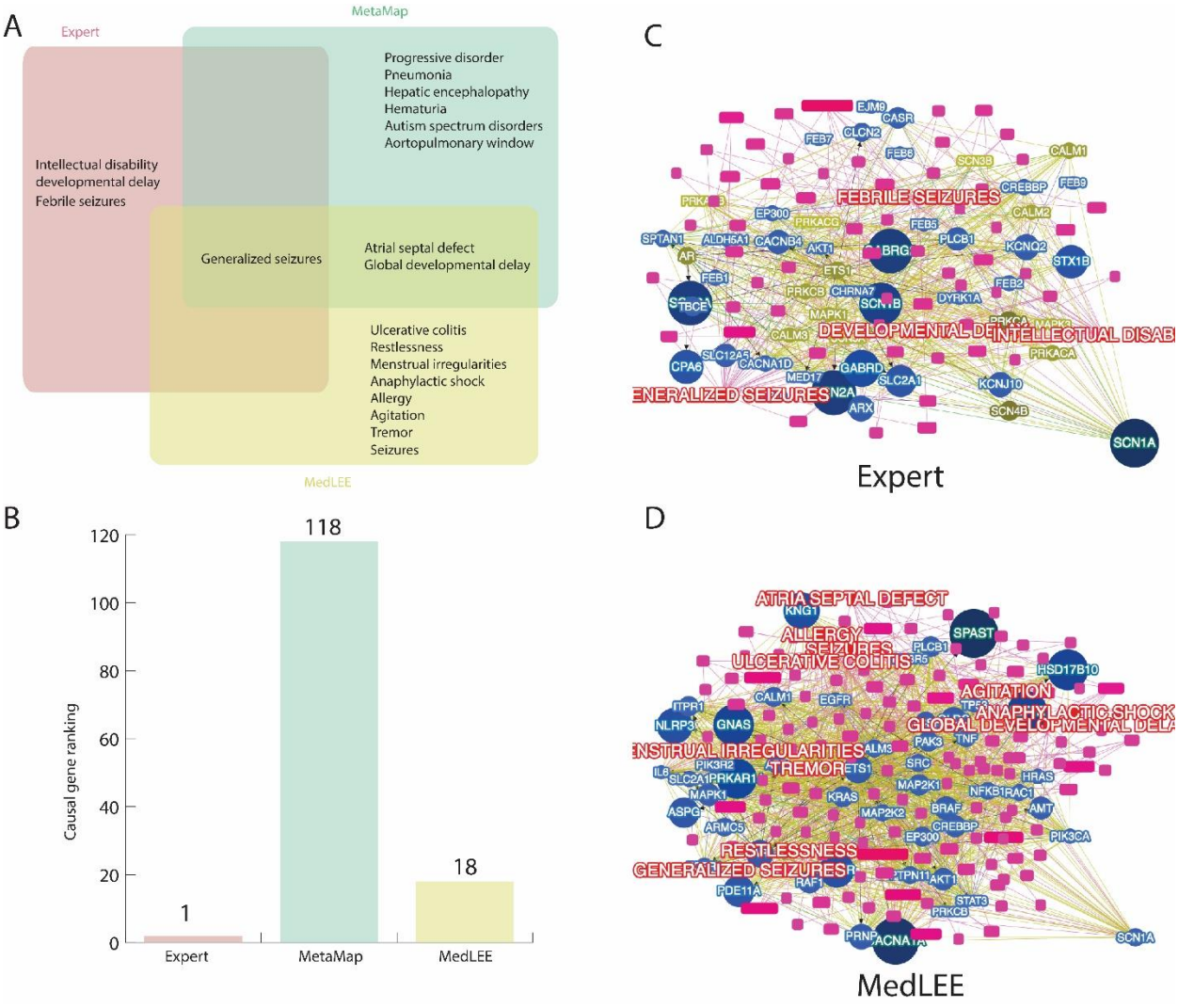


Figure S5. Correlation between the ranks of genes with causal variants based on expert and the ones based on two natural language processing methods, including MedLEE (A) and MetaMap (B). The results with log2 transformed ranks are illustrated in C and D, respectively. Each dot in the plot represents one individual, and in total there are 38 individuals (Cohort 1 and 2 combined). Pearson's correlation coefficient *r* and the *p* values are provided for each plot.

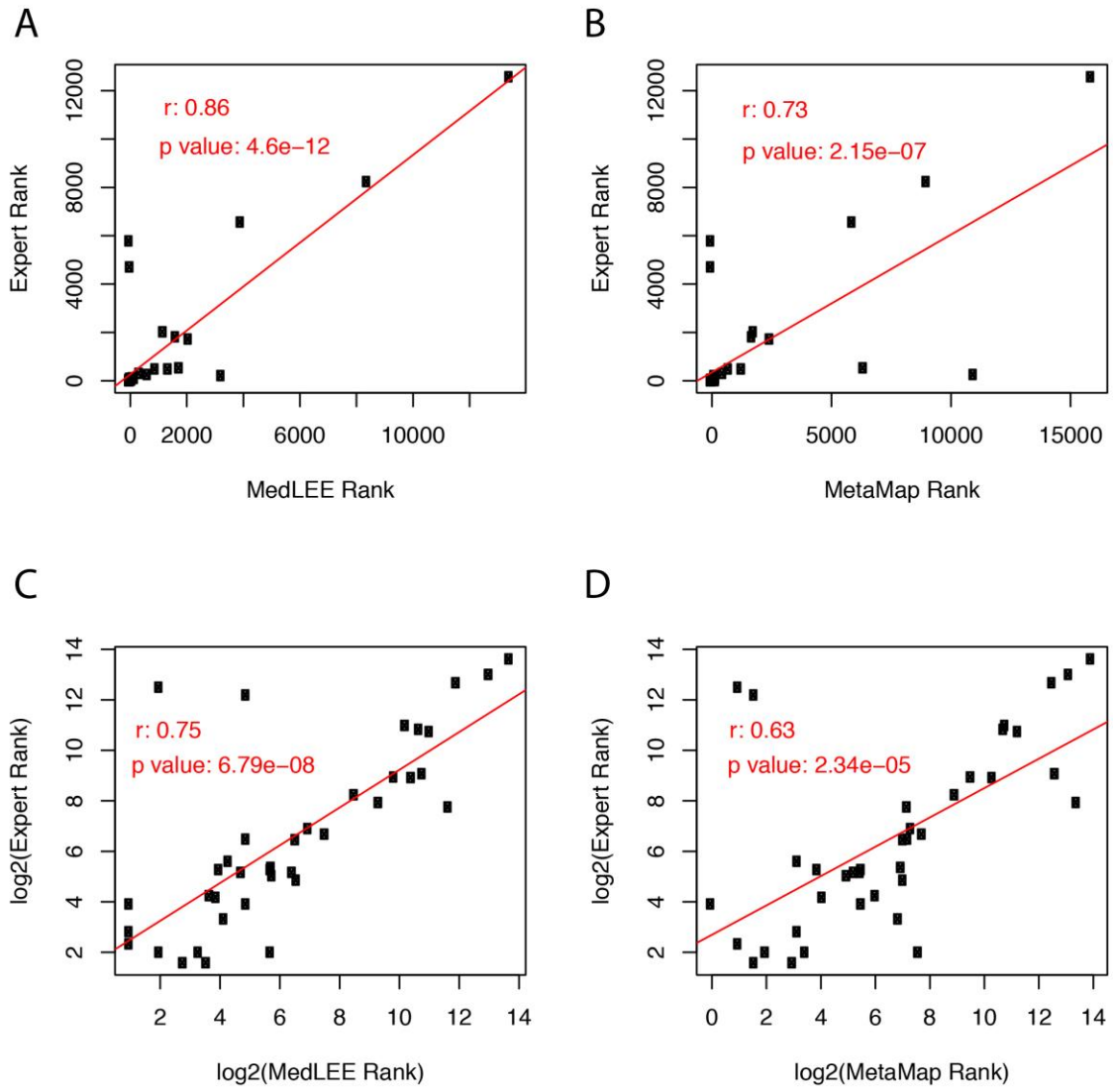


Table S1. A comparison of the "indication" shown in genetic diagnostic reports and the clinical phenotypes inferred from EHR by two algorithms, for 12 affected individuals in cohort 3.

Table S2: Raw phenotype description and NLP-extracted phenotype terms for an individual with confirmed diagnosis of KBG syndrome and with a *de novo* mutation in *ANKRD11* as the causal variant.

Clinical notes	The proband was born to a non-consanguineous couple, who had an unremarkable pregnancy history; however, at birth a large fontanel was reported. Parents and siblings were healthy and no significant family history was reported (Figure 1). The proband met all developmental milestones, except crawling, up to his first epileptic episode, which occurred at three years of age. After this episode, he lost all speech, began exhibiting autistic behavior, and also started to have frequent generalized tonic-clonic seizures. Over time, tonic, atonic, mild clonic, complex partial, myoclonic and gelastic seizures were reported in the proband. Other developmental skills, including throwing a ball, responding to his name, feeding himself with utensils and self-care skills were lost by 4-years of age. No significant conductive hearing loss, heart abnormalities or delayed bone age were found in the proband at that age. The proband was evaluated (by G.J.L.) at eleven years of age. He presented with several neurological and craniofacial abnormalities including epilepsy, ventriculomegaly, relative macrocephaly, prominent forehead, low hairline, thick eyebrows, wide-set eyes, macrodontia of upper central incisors, and full lips. Hand and foot abnormalities included clinodactyly of the fifth digit, bilateral single transverse palmar creases, brachydactyly and flat feet (Figure 2). He also had a diagnosis of cerebral folate deficiency due to the presence of folate receptor autoantibodies.
MetaMap phenotype	Abnormality of the nervous system Bilateral single transverse palmar creases Brachydactyly Cerebral folate deficiency Clinodactyly of the 5th finger Conductive hearing loss Congenital heart defects Delayed bone age Epilepsy Hypermethioninemia Large fontanelle Macrodontia Myoclonic seizures Prominent forehead Relative macrocephaly Seizures, tonic-clonic Thick vermilion border Ventricular dilatation Widely-spaced maxillary central incisors
Gene ranking (MetaMap)	6
MedLEE phenotype	Abnormality of the foot Abnormality of the hand Abnormality of the uterus Autism Autistic behavior Brachydactyly syndrome Clinodactyly of the 5th finger Clonus Finger clinodactyly Gelastic seizures Generalized seizures Generalized tonic-clonic seizures Hyperhistidinemia Incisor macrodontia Increased head circumference Large fontanelles Macrodontia Malnutrition Megalencephaly Myoclonus Pes planus Prominent forehead Relative macrocephaly

	Seizures Thick eyebrow Ventriculomegaly
Gene ranking (MedLEE)	24

Table S3: Raw phenotype description and NLP-extracted phenotype terms for two affected individuals (brother and sister) with confirmed diagnosis of Sanfilippo syndrome and with compound heterozygous mutations in the *NAGLU* gene as the causal variants.

Affected Individual	1
Clinical notes	Individual II-1 is a 10 year old boy. He was born at term with normal birth parameters and good APGAR scores (9/10/10). The neonatal period was uneventful, and he had normal motor development during early childhood: he began to look up at 3 months, sit by himself at 5 months, stand up at 11 months, walk at 13 months, and speak at 17 months. He attended a regular kindergarten, without any signs of difference in intelligence, compared to his peers. Starting at age 6, the parents observed ever increasing behavioral disturbance for the boy, manifesting in multiple aspects of life. For example, he can no longer wear clothes by himself, cannot obey instruction from parents/teachers, can no longer hold subjects tightly in hand, which were all things that he could do before 6 years of age. In addition, he no longer liked to play with others; instead, he just preferred to stay by himself, and he sometimes fell down when he walked on the stairs, which had rarely happened at age 5. The proband continued to deteriorate: at age 9, he could not say a single word and had no action or response to any instruction given in clinical exams. Additionally, rough facial features were noted with a flat nasal bridge, a synophrys (unibrow), a long and smooth philtrum, thick lips and an enlarged mouth. He also had rib edge eversion, and it was also discovered that he was profoundly deaf and had completely lost the ability to speak. He also had loss of bladder control. The diagnosis of severe intellectual disability was made, based on Wechsler Intelligence Scale examination. Brain MRI demonstrated cortical atrophy with enlargement of the subarachnoid spaces and ventricular dilatation (Figure 2). Brainstem evoked potentials showed moderate abnormalities. Electroencephalography (EEG) showed abnormal sleep EEG.
MetaMap phenotype	Cerebral atrophy Deafness Drowsiness Hypoacusis Synophrys Thick vermilion border Urinary incontinence Ventricular dilatation
Gene ranking (MetaMap)	201
MedLEE phenotype	Abnormality of the uterus Brain atrophy Depressed nasal bridge Falls Hearing impairment Intellectual disability Intellectual disability, profound Intellectual disability, severe Severe hearing impairment Smooth philtrum Thick vermilion border Ulcerative colitis Urinary incontinence Ventriculomegaly
Gene ranking (MedLEE)	795
-	-
Affected individual	2
Clinical notes	Individual II-2 is a 9 years old girl. She was born at term, also with normal birth parameters. She began to stand at 11 months, walk with aid at 13 months, and speak at 17 months. At age 5, she was just like other children of similar age, with the ability to dress and sing, and count by herself. Starting at 6 years of age, she began to show regression of developmental patterns: she could not dress by herself anymore, and could not express even a single sentence or count numbers. Clinical examination revealed a coarse face with low anterior and posterior hairlines, prominent frontal bossing, thick eyebrows, synophrys (unibrow), hypertelorism, and thick lips. Growth parameters were normal. Her clinical course was also severe, with progressive neurodegeneration, behavioral problems (including hyperactivity, impulsivity,

	obstinacy, anxious behaviors and autistic-like behaviors), and hearing loss. The diagnosis of severe intellectual disability was made, based on Wechsler Intelligence Scale examination. Measuring activities of daily living showed extreme disability. Brain MRI demonstrated cortical atrophy with enlargement of the subarachnoid spaces and ventricular dilatation (Figure 2). Brainstem evoked potentials showed moderate abnormalities. EEG recording showed abnormal sleep EEG, just like her brother's manifestation.
MetaMap phenotype	Cerebral atrophy Coarse facial features Developmental regression Drowsiness Electroencephalogram abnormal Frontal bossing Hypertelorism Progressive neurologic deterioration Severe hearing impairment Synophrys Thick eyebrow Thick vermilion border Ventricular dilatation
Gene ranking (MetaMap)	42
MedLEE phenotype	Abnormality of the uterus Anxiety Autism Brain atrophy Coarse facial features Developmental regression Frontal bossing Hearing impairment Hyperkinesis Hypertelorism Intellectual disability Intellectual disability, profound Intellectual disability, severe Neurodegeneration Progressive neurologic deterioration Synophrys Thick eyebrow Thick vermilion border Ulcerative colitis Ventriculomegaly
Gene ranking (MedLEE)	50