The American Journal of Human Genetics, Volume *107*

Supplemental Data

Semantic Similarity Analysis Reveals Robust Gene-Disease Relationships in Developmental and Epileptic Encephalopathies

Peter D. Galer, Shiva Ganesan, David Lewis-Smith, Sarah E. McKeown, Manuela Pendziwiat, Katherine L. Helbig, Colin A. Ellis, Annika Rademacher, Lacey Smith, Annapurna Poduri, Simone Seiffert, Sarah von Spiczak, Hiltrud Muhle, Andreas van Baalen, NCEE Study Group, EPGP Investigators, EuroEPINOMICS-RES Consortium, Genomics Research and Innovation Network, Rhys H. Thomas, Roland Krause, Yvonne Weber, and Ingo Helbig

Supplemental Figures

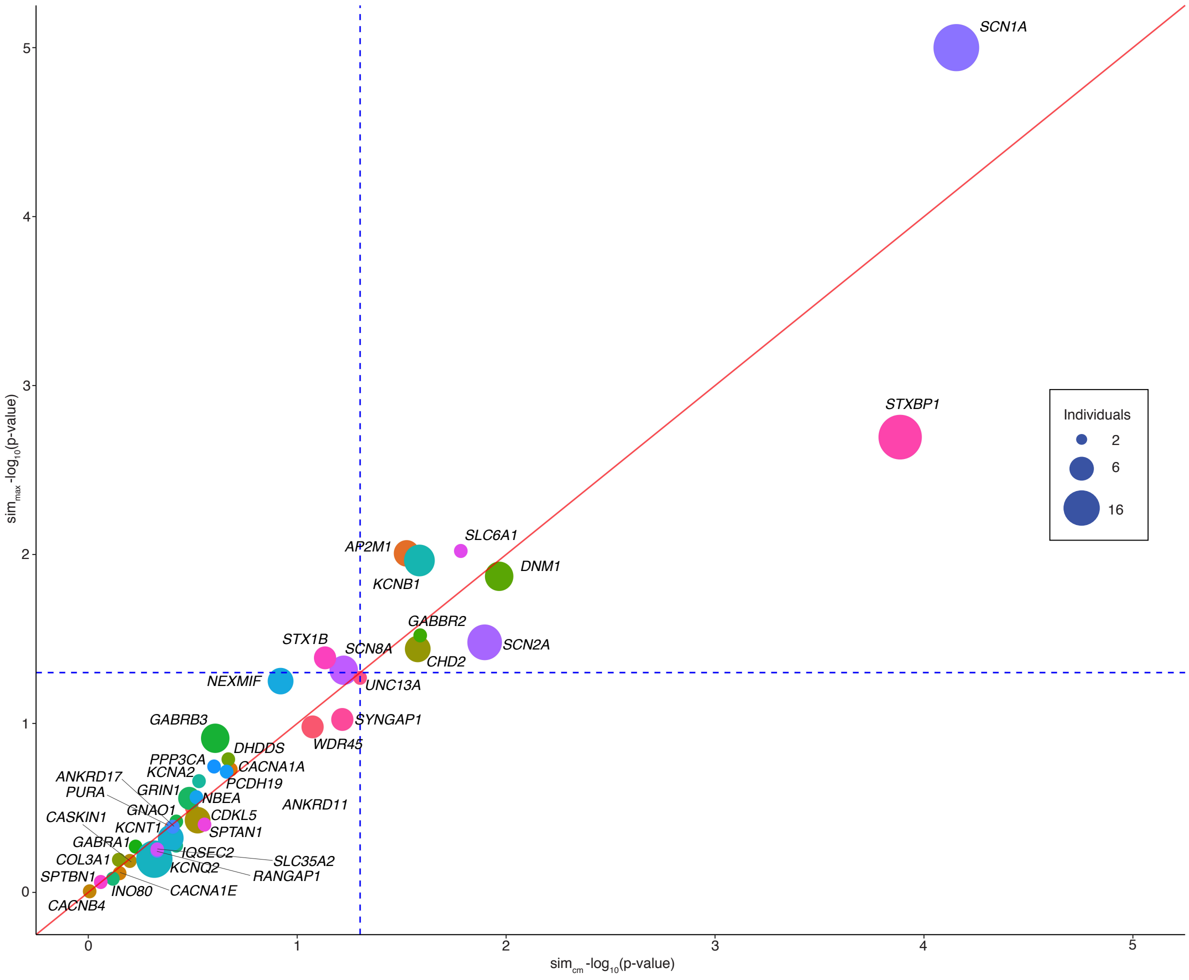


Figure S1. Comparison between sim_{cm} and sim_{max} algorithms

Gene-specific phenotypic similarity between both algorithms is correlated. Point size signifies the number of individuals with a de novo mutation in a specific gene, blue lines denote the -$\log_{10}$(0.05) threshold.

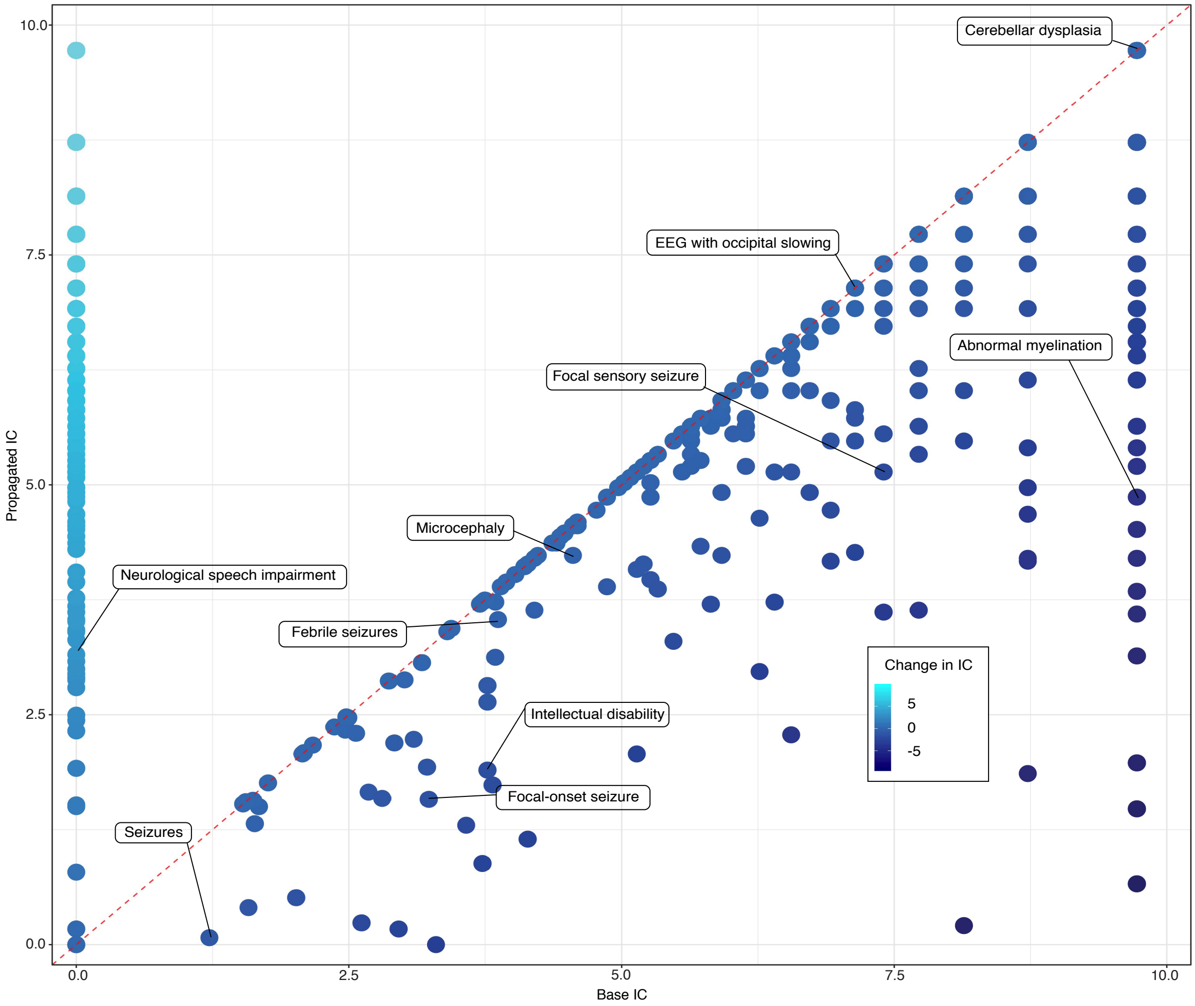


Figure S2. Information Content (IC) of HPO terms before and after propagation

Propagation of assigned HPO terms refers to the addition of all higher-level HPO terms within the ontology. The plot compares Information Content (IC) of all HPO terms in the cohort before and after propagation. Since IC is defined as the -$\log_2$ of the term frequency within the cohort, IC of specific terms after propagation either remains constant or decreases. A decrease in IC is observed if a specific HPO terms becomes more frequent due to the propagation of child terms. In addition, after propagation a significant number of HPO terms are generated that were previously not assigned (see HPO terms with x=0). Overall, propagation significantly adjusts the frequency of HPO terms in the cohort, specifically for higher-level terms that are only assigned infrequently and therefore appear artificially rare.

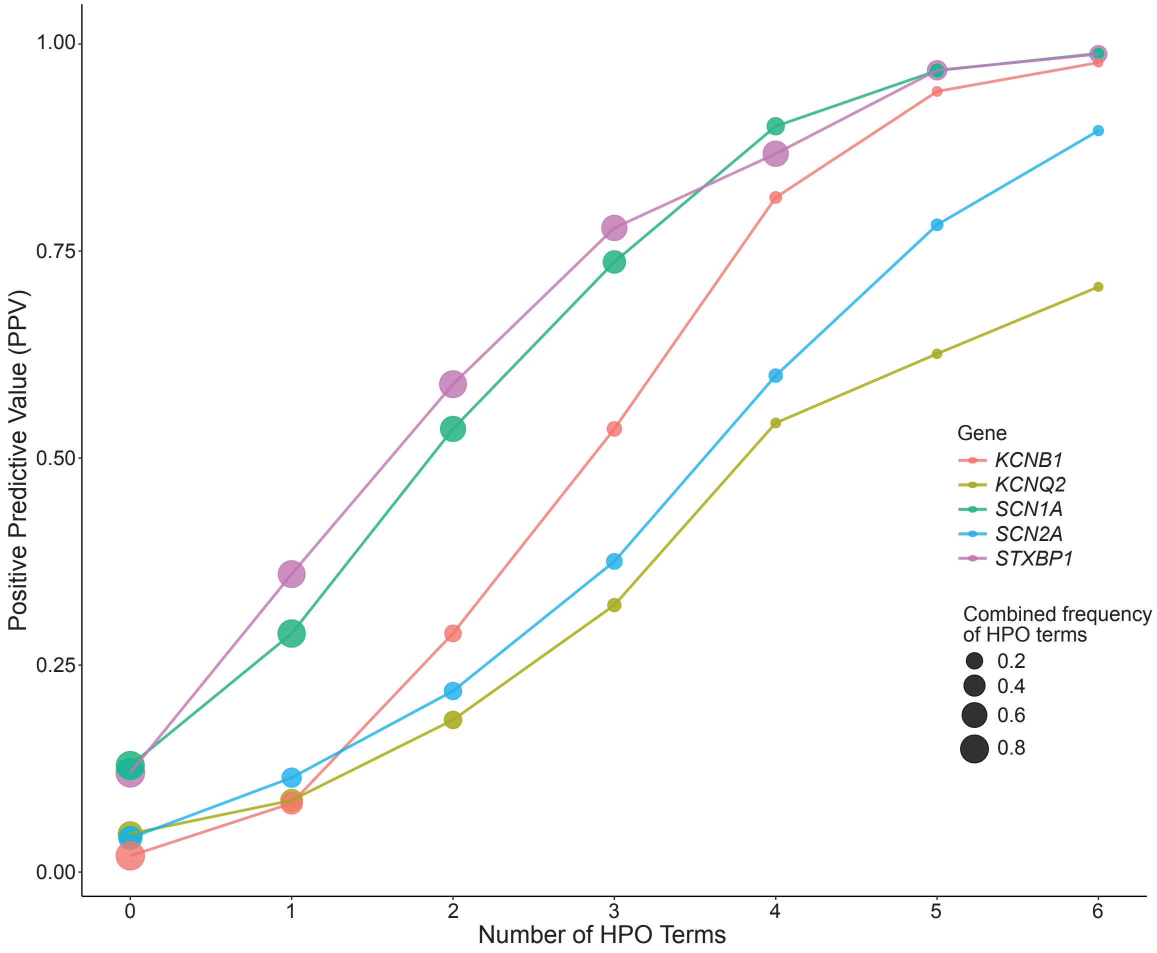
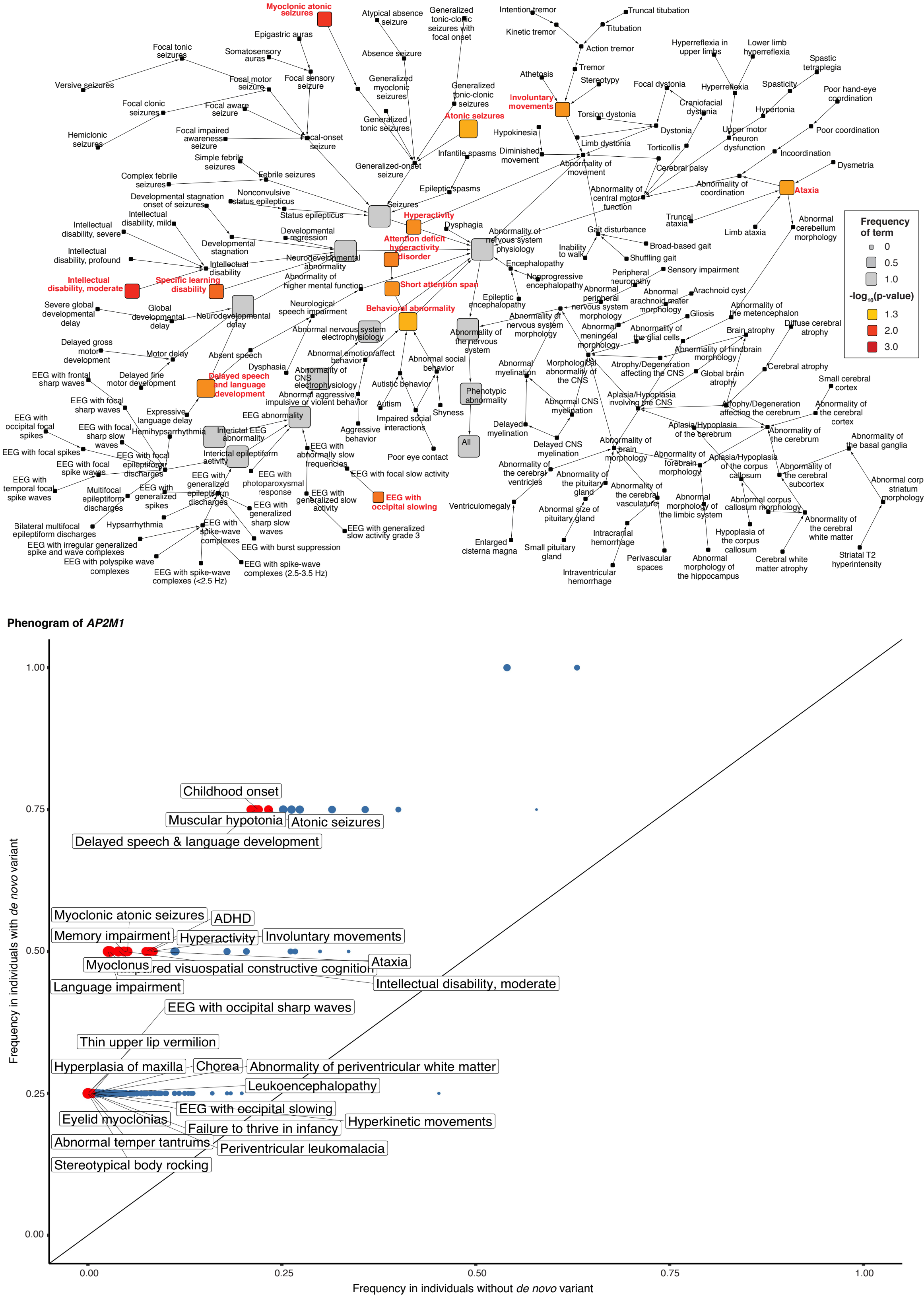


Figure S3. Growth of positive predictive value (PPV) with the addition of HPO terms This figure displays the subsequent growth of the PPV after the addition of HPO terms to the five genetic etiologies with the fastest growth rate. Each color represents a different gene, and the size of the dots indicates the combined frequency of the HPO terms up to that point in that particular gene. *KCNB1, SCN1A,* and *STXBP1* require just 5 terms or less to reach a PPV of at least 80%.

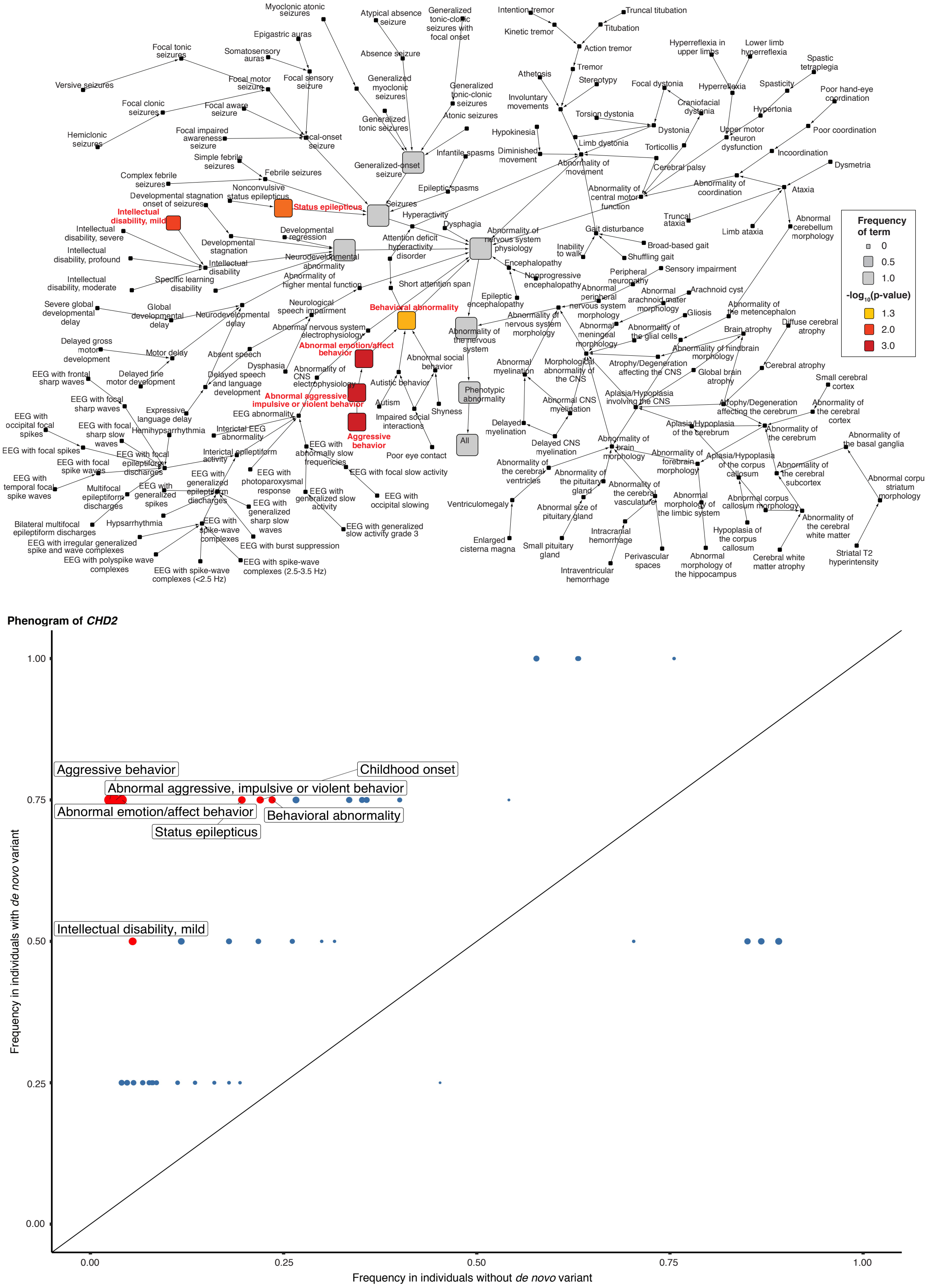
Figure S4 A



Phenotree of *AP2M1*

Phenogram of *AP2M1*

Figure S4 B

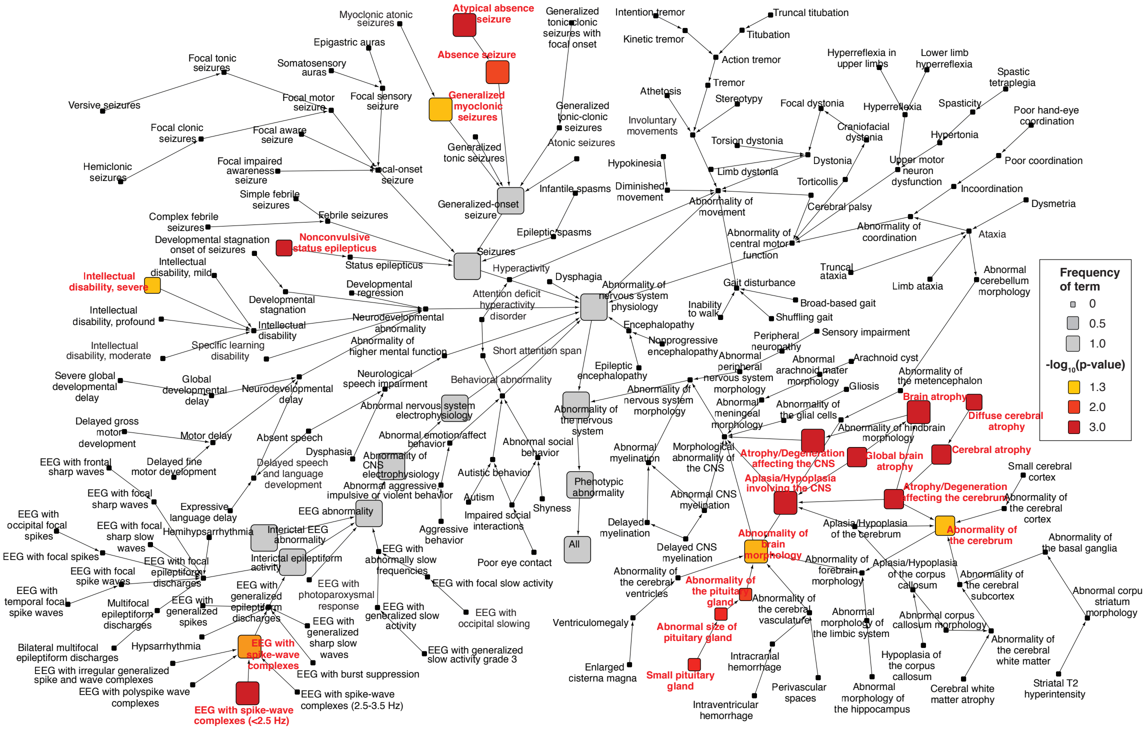


Phenotree of *CHD2*

Phenogram of *CHD2*

Figure S4 C

Phenotree of *DNM1*



Atypical absence seizure
Myoclonic atonic seizures
Generalized tonic-clonic seizures with focal onset
Intention tremor
Truncal titubation
Titubation
Kinetic tremor
Action tremor
Epigastric auras
Absence seizure
Focal tonic seizures
Somatosensory auras
Hyperreflexia in upper limbs
Lower limb hyperreflexia
Spastic tetraplegia
Tremor
Athetosis
Stereotypy
Focal dystonia
Hyperreflexia
Spasticity
Poor hand-eye coordination
Versive seizures
Focal motor seizure
Focal sensory seizure
Generalized myoclonic seizures
Generalized tonic-clonic seizures
Involuntary movements
Craniofacial dystonia
Hypertonia
Focal clonic seizures
Focal aware seizure
Atonic seizures
Torsion dystonia
Dystonia
Upper motor neuron dysfunction
Poor coordination
Hemiclonic seizures
Focal impaired awareness seizure
Focal-onset seizure
Generalized tonic seizures
Hypokinesia
Limb dystonia
Torticollis
Incoordination
Simple febrile seizures
Infantile spasms
Diminished movement
Abnormality of movement
Dysmetria
Complex febrile seizures
Febrile seizures
Generalized-onset seizure
Cerebral palsy
Abnormality of coordination
Developmental stagnation onset of seizures
Nonconvulsive status epilepticus
Epileptic spasms
Abnormality of central motor function
Ataxia
Intellectual disability, mild
Seizures
Status epilepticus
Hyperactivity
Truncal ataxia
Abnormal cerebellum morphology
Intellectual disability, severe
Developmental regression
Dysphagia
Gait disturbance
Limb ataxia
Developmental stagnation
Attention deficit hyperactivity disorder
Abnormality of nervous system physiology
Inability to walk
Broad-based gait
Intellectual disability, profound
Intellectual disability
Neurodevelopmental abnormality
Shuffling gait
Encephalopathy
Intellectual disability, moderate
Specific learning disability
Abnormality of higher mental function
Nonprogressive encephalopathy
Peripheral neuropathy
Sensory impairment
Short attention span
Abnormal peripheral nervous system morphology
Abnormal arachnoid mater morphology
Arachnoid cyst
Severe global developmental delay
Global developmental delay
Neurodevelopmental delay
Neurological speech impairment
Behavioral abnormality
Epileptic encephalopathy
Abnormality of the metencephalon
Gliosis
Abnormal nervous system electrophysiology
Abnormality of nervous system morphology
Abnormal meningeal morphology
Abnormality of the glial cells
Brain atrophy
Diffuse cerebral atrophy
Delayed gross motor development
Motor delay
Absent speech
Abnormal emotion/affect behavior
Abnormality of the nervous system
Abnormality of hindbrain morphology
Dysphasia
Abnormal social behavior
Abnormal myelination
Morphological abnormality of the CNS
Atrophy/Degeneration affecting the CNS
Global brain atrophy
Cerebral atrophy
EEG with frontal sharp waves
Delayed fine motor development
Delayed speech and language development
Abnormality of CNS electrophysiology
Autistic behavior
Small cerebral cortex
Abnormal aggressive, impulsive or violent behavior
Phenotypic abnormality
Aplasia/Hypoplasia involving the CNS
Atrophy/Degeneration affecting the cerebrum
Abnormality of the cerebral cortex
EEG with focal sharp waves
Expressive language delay
Autism
Abnormal CNS myelination
EEG with occipital focal spikes
EEG with focal sharp slow waves
EEG abnormality
Impaired social interactions
Shyness
Delayed myelination
Aplasia/Hypoplasia of the cerebrum
Abnormality of the cerebrum
Hemihypsarrhythmia
Interictal EEG abnormality
Aggressive behavior
All
EEG with focal spikes
EEG with focal epileptiform discharges
Interictal epileptiform activity
EEG with abnormally slow frequencies
Poor eye contact
Delayed CNS myelination
Abnormality of brain morphology
Abnormality of forebrain morphology
Aplasia/Hypoplasia of the corpus callosum
Abnormality of the basal ganglia
EEG with focal spike waves
EEG with generalized epileptiform discharges
EEG with photoparoxysmal response
EEG with focal slow activity
Abnormality of the cerebral ventricles
Abnormality of the pituitary gland
Abnormality of the cerebral subcortex
EEG with temporal focal spike waves
Multifocal epileptiform discharges
EEG with generalized spikes
EEG with generalized slow activity
EEG with occipital slowing
Abnormality of the cerebral vasculature
Abnormal morphology of the limbic system
Abnormal corpus callosum morphology
Abnormal corpus striatum morphology
EEG with generalized sharp slow waves
Ventriculomegaly
Abnormal size of pituitary gland
Abnormality of the cerebral white matter
Bilateral multifocal epileptiform discharges
Hypsarrhythmia
EEG with spike-wave complexes
EEG with generalized slow activity grade 3
Intracranial hemorrhage
Hypoplasia of the corpus callosum
EEG with irregular generalized spike and wave complexes
EEG with burst suppression
Enlarged cisterna magna
Small pituitary gland
Perivascular spaces
Abnormal morphology of the hippocampus
Cerebral white matter atrophy
Striatal T2 hyperintensity
EEG with polyspike wave complexes
EEG with spike-wave complexes (<2.5 Hz)
EEG with spike-wave complexes (2.5-3.5 Hz)
Intraventricular hemorrhage
Frequency of term
0
0.5
1.0
$-\log_{10}$(p-value)
1.3
2.0
3.0

Phenogram of *DNM1*

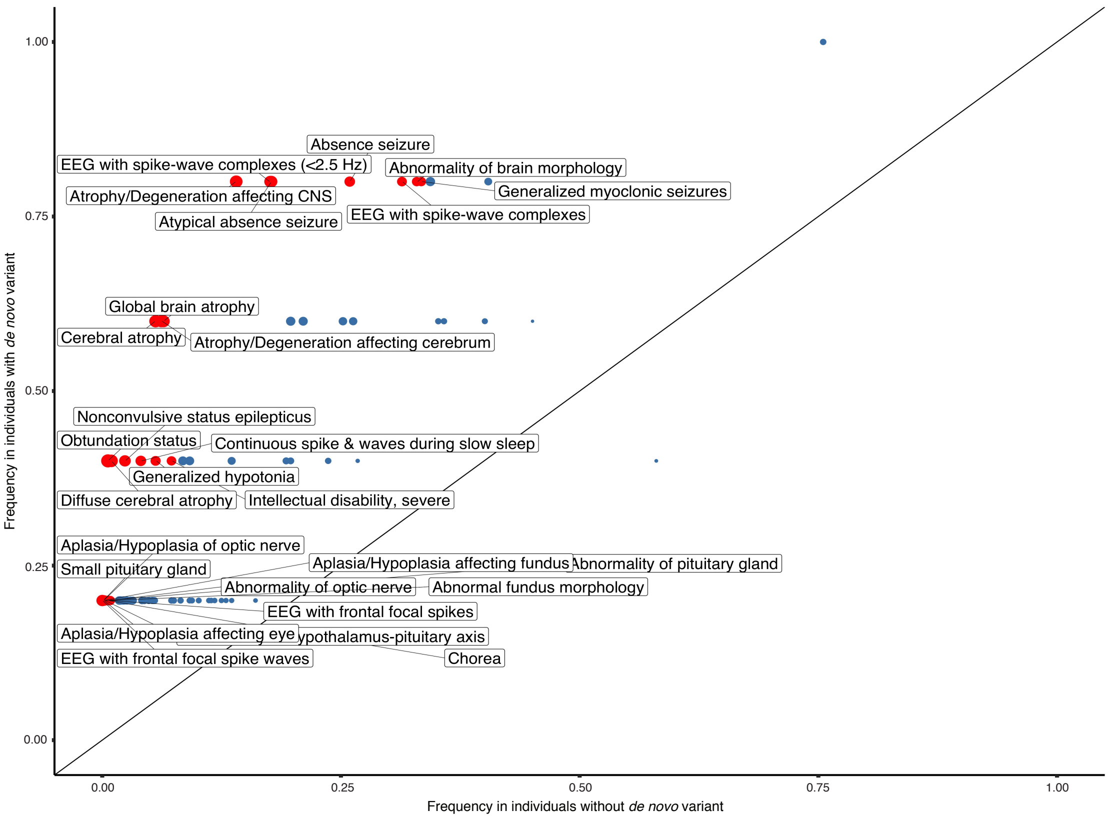
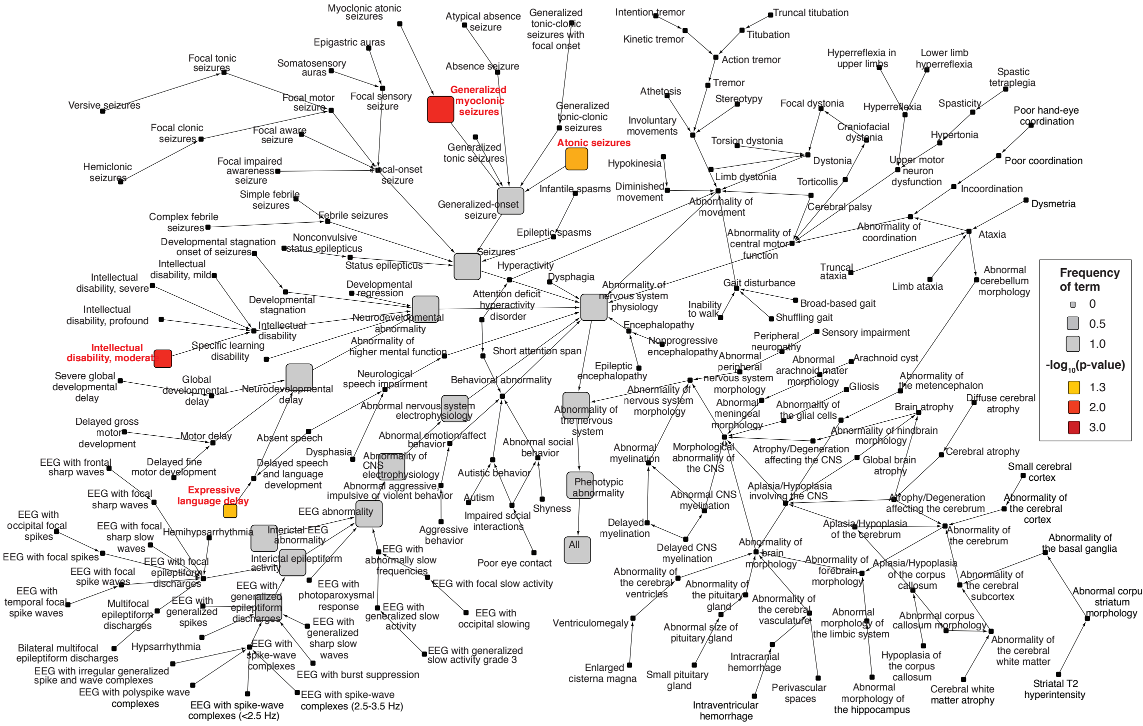


Figure S4 D

Phenotree of *NEXMIF*


Myoclonic atonic seizures
Atypical absence seizure
Generalized tonic-clonic seizures with focal onset
Intention tremor
Truncal titubation
Kinetic tremor
Titubation
Epigastric auras
Focal tonic seizures
Somatosensory auras
Absence seizure
Action tremor
Hyperreflexia in upper limbs
Lower limb hyperreflexia
Spastic tetraplegia
Generalized myoclonic seizures
Tremor
Athetosis
Versive seizures
Focal motor seizure
Focal sensory seizure
Stereotypy
Focal dystonia
Hyperreflexia
Spasticity
Poor hand-eye coordination
Generalized tonic-clonic seizures
Involuntary movements
Craniofacial dystonia
Focal clonic seizures
Focal aware seizure
Atonic seizures
Torsion dystonia
Hypertonia
Generalized tonic seizures
Dystonia
Hemiclonic seizures
Focal impaired awareness seizure
Focal-onset seizure
Hypokinesia
Upper motor neuron dysfunction
Poor coordination
Limb dystonia
Torticollis
Infantile spasms
Diminished movement
Incoordination
Simple febrile seizures
Generalized-onset seizure
Abnormality of movement
Cerebral palsy
Dysmetria
Complex febrile seizures
Febrile seizures
Epileptic spasms
Abnormality of central motor function
Abnormality of coordination
Ataxia
Nonconvulsive status epilepticus
Developmental stagnation onset of seizures
Seizures
Status epilepticus
Intellectual disability, mild
Hyperactivity
Truncal ataxia
Limb ataxia
Abnormal cerebellum morphology
Intellectual disability, severe
Dysphagia
Abnormality of nervous system physiology
Gait disturbance
Developmental regression
Attention deficit hyperactivity disorder
Broad-based gait
Developmental stagnation
Inability to walk
Intellectual disability, profound
Neurodevelopmental abnormality
Shuffling gait
Intellectual disability
Encephalopathy
Sensory impairment
Peripheral neuropathy
Intellectual disability, moderate
Specific learning disability
Abnormality of higher mental function
Nonprogressive encephalopathy
Short attention span
Abnormal peripheral nervous system morphology
Abnormal arachnoid mater morphology
Arachnoid cyst
Epileptic encephalopathy
Severe global developmental delay
Global developmental delay
Neurodevelopmental delay
Neurological speech impairment
Behavioral abnormality
Abnormality of nervous system morphology
Gliosis
Abnormality of the metencephalon
Abnormal meningeal morphology
Abnormality of the glial cells
Diffuse cerebral atrophy
Abnormal nervous system electrophysiology
Abnormality of the nervous system
Brain atrophy
Delayed gross motor development
Abnormal emotion/affect behavior
Abnormality of hindbrain morphology
Motor delay
Absent speech
Abnormal social behavior
Abnormal myelination
Morphological abnormality of the CNS
Atrophy/Degeneration affecting the CNS
Cerebral atrophy
Dysphasia
Abnormality of CNS electrophysiology
Global brain atrophy
EEG with frontal sharp waves
Delayed fine motor development
Delayed speech and language development
Autistic behavior
Small cerebral cortex
Abnormal aggressive, impulsive or violent behavior
Phenotypic abnormality
Abnormal CNS myelination
Aplasia/Hypoplasia involving the CNS
Atrophy/Degeneration affecting the cerebrum
Abnormality of the cerebral cortex
EEG with focal sharp waves
Expressive language delay
Autism
Shyness
EEG with occipital focal spikes
EEG abnormality
Impaired social interactions
Delayed myelination
Aplasia/Hypoplasia of the cerebrum
Abnormality of the cerebrum
EEG with focal sharp slow waves
Hemihypsarrhythmia
Interictal EEG abnormality
Aggressive behavior
All
Abnormality of the basal ganglia
EEG with focal spikes
EEG with focal epileptiform discharges
Interictal epileptiform activity
EEG with abnormally slow frequencies
Poor eye contact
Delayed CNS myelination
Abnormality of brain morphology
EEG with focal spike waves
EEG with photoparoxysmal response
EEG with focal slow activity
Abnormality of the cerebral ventricles
Abnormality of the pituitary gland
Abnormality of forebrain morphology
Aplasia/Hypoplasia of the corpus callosum
Abnormality of the cerebral subcortex
EEG with temporal focal spike waves
Multifocal epileptiform discharges
EEG with generalized spikes
EEG with generalized epileptiform discharges
EEG with generalized slow activity
EEG with occipital slowing
Abnormality of the cerebral vasculature
Abnormal corpus striatum morphology
EEG with generalized sharp slow waves
Ventriculomegaly
Abnormal size of pituitary gland
Abnormal morphology of the limbic system
Abnormal corpus callosum morphology
Hypsarrhythmia
Abnormality of the cerebral white matter
Bilateral multifocal epileptiform discharges
EEG with spike-wave complexes
EEG with generalized slow activity grade 3
Intracranial hemorrhage
Hypoplasia of the corpus callosum
EEG with irregular generalized spike and wave complexes
EEG with burst suppression
Enlarged cisterna magna
Small pituitary gland
Perivascular spaces
Abnormal morphology of the hippocampus
Cerebral white matter atrophy
Striatal T2 hyperintensity
EEG with polyspike wave complexes
EEG with spike-wave complexes (<2.5 Hz)
EEG with spike-wave complexes (2.5-3.5 Hz)
Intraventricular hemorrhage
Frequency of term
0
0.5
1.0
$-\log_{10}$(p-value)
🟨 1.3
🟥 2.0
🟥 3.0

Phenogram of *NEXMIF*

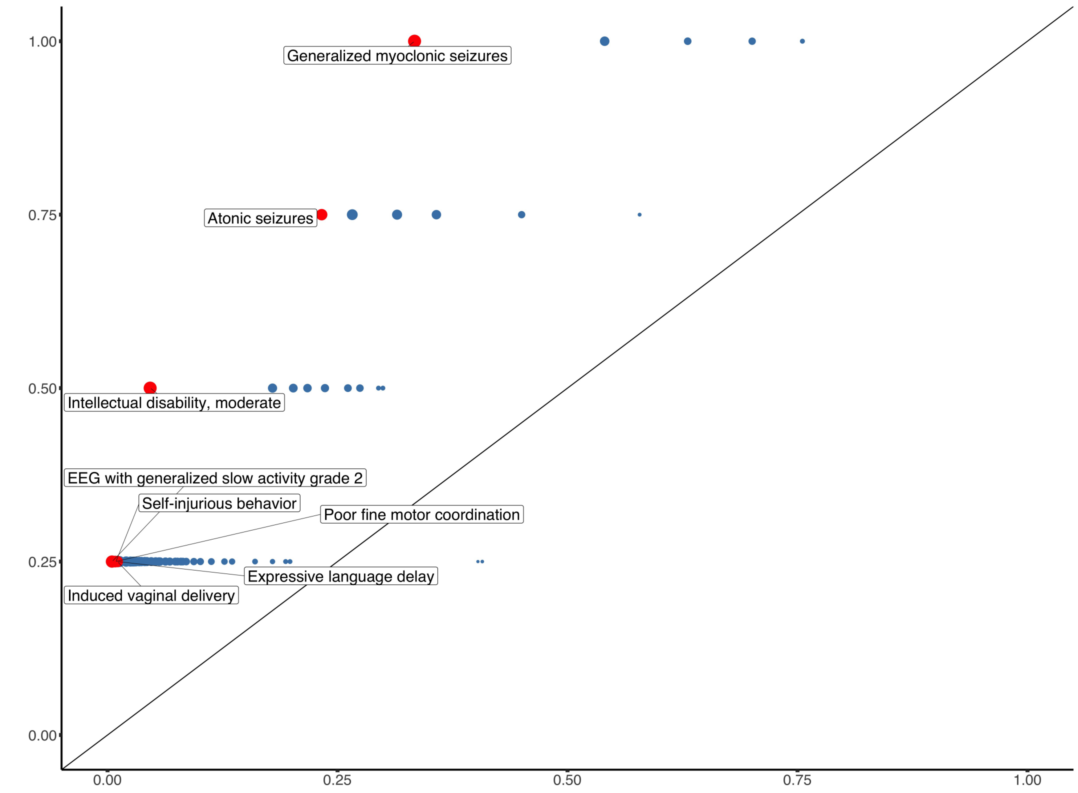
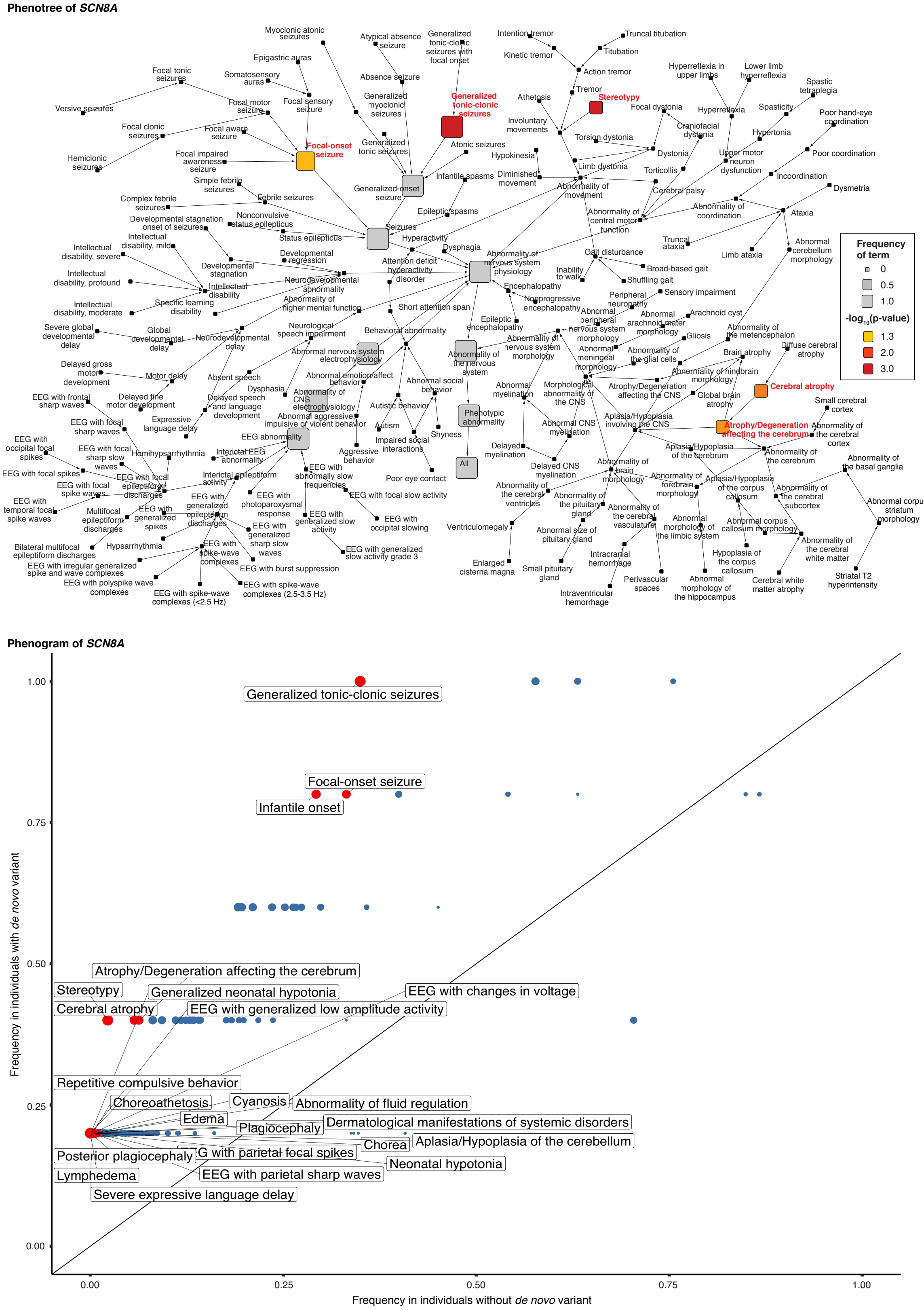


Figure S4 E



Phenotree of *SCN8A*

Phenogram of *SCN8A*

Figure S4 F

Phenotree of *STX1B*

Phenogram of *STX1B*

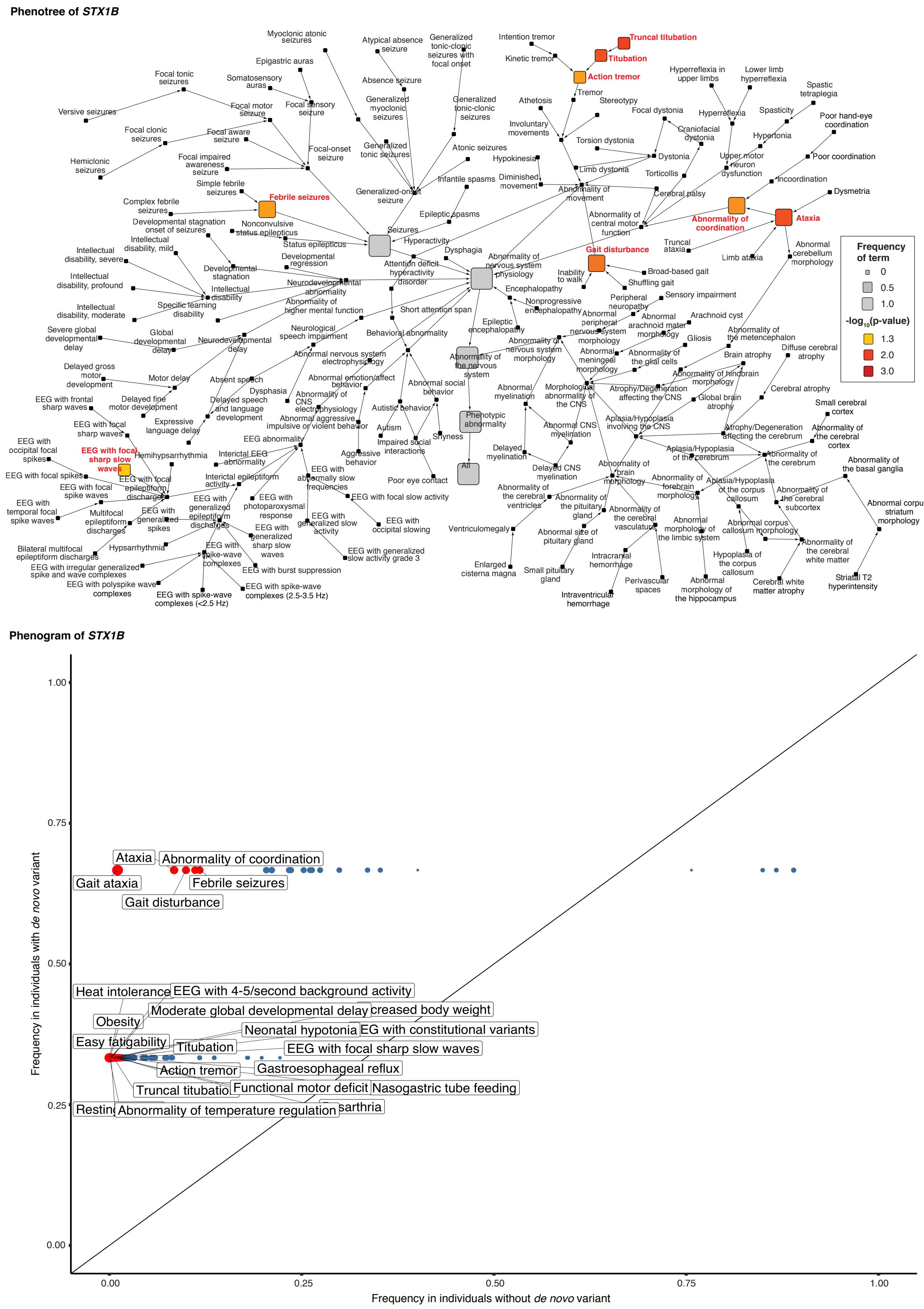


Figure S4. Phenograms and Phenotrees of six genetic etiologies with 20 de novo variants

(A)-(F) The graphs display frequencies of HPO terms in AP2M1, DNM1, CHD2, SCN8A, STX1B, and NEXMIF compared to the overall cohort. Red dots indicate significant associations ($p<0.05$) between HPO terms and specific genes, the size of the dot denotes the degree of significance displayed as $-\log_{10}$(p-value). Significant associations present in 15% of individuals or more with a specific gene are labeled. Parent terms that displayed redundant information were removed.

Supplemental Note

Comparison of sim_{max} and sim_{cm} algorithms

In our study, we used two similarity algorithms to assess the phenotypic relatedness between individuals, the sim_{max} and sim_{cm}. In a previous study, we have used the sim_{max} algorithm to provide evidence for the role of *AP2M1* in human disease. While conceptually related, both algorithms emphasize different features of the HPO and the following hypothetical example demonstrate the emphasis that both algorithms provide to different features in the dataset. For the calculation of Information Content (IC) and similarity, the observed values in our dataset of 846 individuals is used.

Assignment of HPO terms for two individuals

For our example, we assume that two individuals (P_1, P_2) are assigned the following HPO terms (**Table S7**). The assigned HPO terms are "base" HPO terms, e.g. inclusion of higher-level, ancestral HPO terms through propagation has not yet been performed.

HPO terms assigned in individual P_1	**HPO terms assigned in individual P_2**
Focal aware seizure (HP:0002349)	Focal-onset seizure (HP:0007359)
Focal clonic seizure (HP:0002266)	Neurodevelopmental delay (HP:0012758)
Mild global developmental delay (HP:0011342)	
Delayed speech and language development (HP:0000750)	

Table S7. Assigned HPO terms for two hypothetical individuals to demonstrate the differences between the sim_{max} and sim_{cm} algorithms.

Relative position of HPO terms within the HPO tree

Figure S5 shows the relative position of the HPO terms within the overall ontological tree. From this illustration, it is apparent that some of the HPO terms assigned to both individuals refer to related concepts within the HPO. For example, “Focal aware seizure” (HP:0002349) and “Focal clonic seizure” (HP:0002266) assigned in individual P_1 are both child terms of “Focal-onset seizure” (HP:0007359) assigned in individual P_2. It can be seen from Figure S5 and Figure S6 that P_1 is assigned more specific terms and that P_2 is assigned higher-level terms.

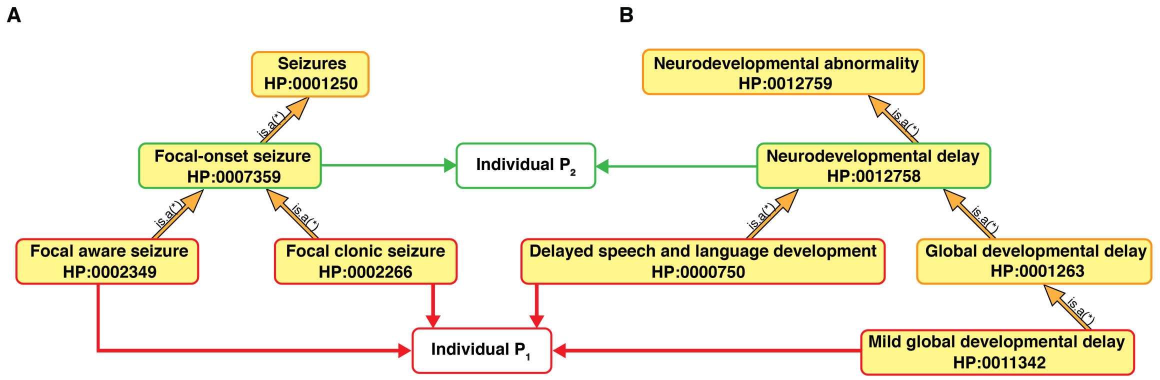


Figure S5. Structure of the HPO with two subbranches with superimposed terms that are assigned to both individuals P_1 (red) and P_2 (green).

Information content for assigned and propagated terms

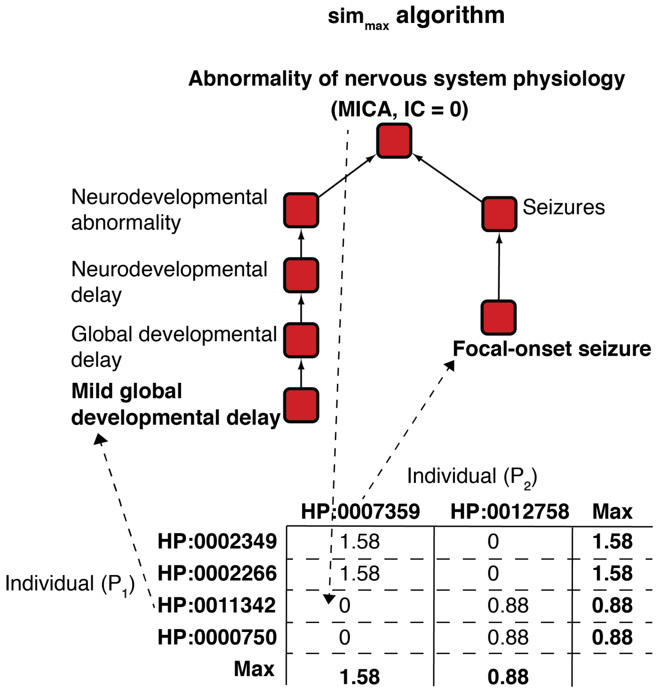
Based on the structure of the HPO, inclusion of higher-level HPO terms generates a list of extended phenotypic terms for both individuals as shown in Table S8. Information Content (IC) is generated as the -$\log_2$ of the term frequency and the initially assigned (“base”) terms are labelled in red (pat1) and green (pat2). With decreasing specificity of the terms within the structure of the HPO, terms become more frequent and the IC decreases.

P_1 propagated HPO terms	**P_2 propagated HPO terms**
All (HP:0000001; IC=0)	All (HP:0000001; IC=0)
Phenotypic abnormality (HP:0000118; IC=0)	Phenotypic abnormality (HP:0000118; IC=0)
Abnormality of the nervous system (HP:0000707; IC=0)	Abnormality of the nervous system (HP:0000707; IC=0)
Delayed speech and language development **(HP:0000750; IC=2.19)**	Seizures (HP:0001250; IC=0.08)
Seizures (HP:0001250; IC=0.08)	**Focal-onset seizure** **(HP:0007359; IC=1.58)**
Global developmental delay (HP:0001263; IC=1.32)	Abnormality of nervous system physiology (HP:0012638; IC=0)
Focal clonic seizures (HP:0002266; IC=3.64)	**Neurodevelopmental delay** **(HP:0012758; IC=0.88)**
Focal aware seizure **(HP:0002349; IC=5.55)**	Neurodevelopmental abnormality (HP:0012759; IC=0.66)
Focal-onset seizure **(HP:0007359; IC=1.58)**	
Focal motor seizure (HP:0011153; IC=2.64)	
Mild global developmental delay **(HP:0011342; IC=5.92)**	
Abnormality of nervous system physiology (HP:0012638; IC=0)	
Neurodevelopmental delay (HP:0012758; IC=0.88)	
Neurodevelopmental abnormality (HP:0012759; IC=0.66)	

Table S8. Propagated HPO terms for both P_1 and P_2 with the initially assigned terms in bold and red (P_1) or green (P_2). The highest-level terms in the HPO ("All"; HP:0000001 and Phenotypic abnormality"; HP:0000118) have an Information Content of zero, indicating that these terms are present in all individuals.

Phenotypic similarity assessed by the sim$_{max}$ algorithm

An intuitive way to conceptualize the sim$_{max}$ algorithm is to place the assigned HPO terms for P_1 and P_2 on a 2 x 2 matrix with the rows representing the terms assigned to P_1 and the columns representing the terms assigned to P_2 (**Figure S6**).



sim$_{max}$ algorithm

Abnormality of nervous system physiology
(MICA, IC = 0)
Neurodevelopmental abnormality
Neurodevelopmental delay
Global developmental delay
Mild global developmental delay
Seizures
Focal-onset seizure
Individual (P_2)
Individual (P_1)

	HP:0007359	HP:0012758	Max
HP:0002349	1.58	0	**1.58**
HP:0002266	1.58	0	**1.58**
HP:0011342	0	0.88	**0.88**
HP:0000750	0	0.88	**0.88**
Max	**1.58**	**0.88**	

$$\text{Similarity score (sim}_{max}\text{) for Individual } P_1 \text{ and } P_2 = \frac{(1.58 + 1.58 + 0.88 + 0.88) + (1.58 + 0.88)}{2}$$

Similarity score (sim$_{max}$) for Individual P_1 and P_2 = 3.69

Figure S6. Calculating the sim$_{max}$ score for individuals P_1 and P_2 using the assigned phenotypes from Table S7 including "Focal aware seizure" (HP:0002349), "Focal clonic seizure" (HP:0002266),"Mild global developmental delay" (HP:0011342) for P_1, "Delayed speech and language development" (HP:0000750) and "Focal-onset seizure" (HP:0007359) and "Neurodevelopmental delay" (HP:0012758) for P_2. The sim$_{max}$ algorithm assesses the Most Informative Common Ancestor (MICA) for each term combination and sums up the row-wise (P_1) and column-wise (P_2) maxima, thereby determining the similarity of P_1->P_2 and P_2->P_1. For the final similarity score, both the row-wise and column-wise similarity are averaged.

Phenotypic similarity assessed by the sim_{cm} algorithm

In contrast to determining the MICA for each term combination, the sim_{cm} algorithm assesses the Information Content of all HPO terms shared by P_1 and P_2 using the propagated HPO dataset. This is shown in **Table S9**, which is derived from **Table S8**.

P_1 propagated HPO terms	P_2 propagated HPO terms	Overlap
All (HP:0000001; IC=0)	**All (HP:0000001; IC=0)**	**All (HP:0000001; IC=0)**
Phenotypic abnormality (HP:0000118; IC=0)	**Phenotypic abnormality (HP:0000118; IC=0)**	**Phenotypic abnormality (HP:0000118; IC=0)**
Abnormality of the nervous system (HP:0000707; IC=0)	**Abnormality of the nervous system (HP:0000707; IC=0)**	**Abnormality of the nervous system (HP:0000707; IC=0)**
Delayed speech and language development (HP:0000750; IC=2.19)		
Seizures (HP:0001250; IC=0.08)	**Seizures (HP:0001250; IC=0.08)**	**Seizures (HP:0001250; IC=0.08)**
Global developmental delay (HP:0001263; IC=1.32)		
Focal clonic seizures (HP:0002266; IC=3.64)		
Focal aware seizure (HP:0002349; IC=5.55)		
Focal-onset seizure (HP:0007359; IC=1.58)	**Focal-onset seizure (HP:0007359; IC=1.58)**	**Focal-onset seizure (HP:0007359; IC=1.58)**
Focal motor seizure (HP:0011153; IC=2.64)		
Mild global developmental delay (HP:0011342; IC=5.92)		
Abnormality of nervous system physiology (HP:0012638; IC=0)	**Abnormality of nervous system physiology (HP:0012638; IC=0)**	**Abnormality of nervous system physiology (HP:0012638; IC=0)**
Neurodevelopmental delay (HP:0012758; IC=0.88)	**Neurodevelopmental delay (HP:0012758; IC=0.88)**	**Neurodevelopmental delay (HP:0012758; IC=0.88)**
Neurodevelopmental abnormality (HP:0012759; IC=0.66)	**Neurodevelopmental abnormality (HP:0012759; IC=0.66)**	**Neurodevelopmental abnormality (HP:0012759; IC=0.66)**
Total Similarity (adding IC values for overlapping HPO terms)		**0 + 0 + 0 + 0.08 + 1.58 + 0 + 0.88 + 0.66 = 3.2**

Table S9. Calculating the sim_{cm} score from the overlap of propagated HPO terms between P_1 and P_2. Given that both HPO terms initially assigned to P_2 were ancestral terms of the four HPO terms assigned to P_1, both assigned terms for P2 are part of the overlapping group of HPO terms. However, as the propagation also includes higher level terms, HPO terms including “Seizures” (HP:0001250; IC=0.08) and “Neurodevelopmental abnormality” (HP:0012759; IC=0.66) contribute to the final score, even though these terms had not been initially assigned. This makes this similarity measure vulnerable to changes in the overall granularity of the HPO within specific sub-branches.

Factors affecting similarities assessed by the $\mathrm{sim}_{\mathrm{max}}$ and $\mathrm{sim}_{\mathrm{cm}}$ algorithm

Effect of annotation density

In our example, P_1 was assigned two specific focal seizure terms, including "Focal aware seizure" (HP:0002349) and "Focal clonic seizure" (HP:0002266), whereas P_2 was only assigned a single higher-level HPO term for focal seizures, namely "Focal-onset seizure" (HP:0007359). Within the $\mathrm{sim}_{\mathrm{max}}$ algorithm, both focal seizures types (HP:0002349, HP:0002266) contribute to the overall similarity, whereas the $\mathrm{sim}_{\mathrm{cm}}$ algorithm would only capture the IC of the more general focal seizure term (HP:0007359). For example, if one specific focal seizure term were to be removed from P1, $\mathrm{sim}_{\mathrm{max}}$ would decrease, whereas $\mathrm{sim}_{\mathrm{cm}}$ would remain the same. Likewise, if another specific focal seizure term were to be added, $\mathrm{sim}_{\mathrm{max}}$ would increase, while $\mathrm{sim}_{\mathrm{cm}}$ would remain constant. **The $\mathrm{sim}_{\mathrm{max}}$ algorithm increases similarity with the addition of assigned HPO terms and, as a distinct property from the $\mathrm{sim}_{\mathrm{cm}}$ algorithm, increases similarity when multiple child terms are annotated.** Accordingly, $\mathrm{sim}_{\mathrm{max}}$ is affected by the annotation density of the assigned HPO terms, whereas $\mathrm{sim}_{\mathrm{cm}}$ removes this effect as HPO terms are de-duplicated after propagation and only overlapping ancestral terms are considered.

Effect of HPO granularity

In our example, the $\mathrm{sim}_{\mathrm{cm}}$ algorithm included higher-level terms in the assessment of similarity that are dependent on the structure of the HPO. The $\mathrm{sim}_{\mathrm{max}}$ algorithm is in principle independent of the overall structure of the overall ontology, as it only assesses the Information Content of the Most Information Common Ancestors (MICA), independent of how deep these ancestral terms are located within the HPO tree. However, the $\mathrm{sim}_{\mathrm{cm}}$

algorithm includes the information content of all propagated HPO terms and is therefore dependent on the local of the assigned terms within the HPO structure. For example, if another redundant HPO term would be placed between “Neurodevelopmental delay” (HP:0012758) and “Neurodevelopmental abnormality” (HP:0012759) that is equivalent to “Neurodevelopmental delay” (HP:0012758), this new, redundant term would also have an IC of 0.88. Such a “spacer term” could be introduced based on theoretical considerations on how to structure phenotypes or novel disease classifications suggested by professional organizations that aim at providing a higher granularity for phenotype assignment within the HPO in future studies. However, such a new, interspersed term would increase the overall similarity assessed through the sim_{cm} algorithm. The results of the sim_{max} algorithm remain unchanged. Likewise, if a term within the HPO structure is considered redundant and is removed, the sim_{cm} algorithm would generate a lower similarity. In our example, this would be the hypothetical situation in which “Neurodevelopmental delay” (HP:0012758) and “Neurodevelopmental abnormality” (HP:0012759) are collapsed into a single HPO term. **Accordingly, the sim_{cm} algorithm is dependent on the overall granularity of the HPO with higher granularity in branches with commonly assigned terms generating higher similarities.** However, the sim_{max} algorithm is not affected by the granularity of the HPO structure itself.

Simulating the recognition of Rett Syndrome

In the introduction of our manuscript, we used the clinical recognition of Rett Syndrome in 1954 as an example of how distinct clinical features may significantly stand out sufficiently to be recognized. We attempted to simulate this historical example by adding six hypothetical individuals with Rett Syndrome to our cohort with various combinations of clinical features that include three different scenarios (**Table S10**).

Scenario 1 (n=1 term)	**Scenario 2 (n=2 terms)**	**Scenario 4 (n=4 terms)**
Stereotypical hand wringing (HP:0012171)	Stereotypical hand wringing (HP:0012171)	Stereotypical hand wringing (HP:0012171)
	Developmental regression (HP:0002376)	Developmental regression (HP:0002376)
		Absent speech (HP:0001344)
		Apraxia (HP:0002186)

Table S10. Combination of HPO terms in simulated individuals with Rett Syndrome. For the similarity analysis, the existing frequencies of these phenotypes in the cohort were used. The phenotype "Stereotypical hand wringing" (HP:0012171) had not been assigned in the cohort and assigned an Information Content of 8.7 for the analysis based on the estimated frequency of 2/846. The IC for "Stereotypical hand wringing" (HP:0012171) was kept constant for n=2, n=4, n=6 patients. For the three other HPO terms ("Developmental regression" HP:0002376, "Absent speech" HP:0001344, "Apraxia" HP:0002186), the existing frequencies in the cohort of 846 individuals was used.

For the simulation, we assessed the combination of one term (Scenario 1), two terms (Scenario 2), and four terms (Scenario 3) in n=2, n=4, and n=6 individuals, using the sim$_{max}$ algorithm. **Table S11** shows the results obtained for the nine combinations of HPO term numbers and number of individuals.

Scenario	Number of individuals	Number of HPO terms	Median similarity using sim_{max}	p-value
Scenario 1	2	1	8.7	0.283
Scenario 1	2	2	11.6	0.180
Scenario 1	2	4	19.9	0.056
Scenario 2	4	1	8.7	0.192
Scenario 2	4	2	11.6	0.087
Scenario 2	4	4	19.9	**0.010**
Scenario 3	6	1	8.7	0.138
Scenario 3	6	2	11.6	**0.045**
Scenario 3	6	4	19.9	**0.002**

Table S11. Simulation of phenotypic similarity for Rett Syndrome using combinations of number of individuals (n=2, n=4, n=6) and number of HPO terms (n=1, n=2, n=4). With increasing number of individuals and HPO terms, the phenotypic similarity between individuals becomes significant.

Examining all combinations of number of individuals and number of terms, we observe that phenotypic significance is achieved once more terms or more individuals are included. For n=6 individuals, the phenotypic similarity is significant for n=2 terms (p=0.05) and n=4 HPO terms (p=0.002). The terms assigned to individuals are “Stereotypical hand wringing” (HP:0012171), “Developmental regression” (HP:0002376), “Absent speech” (HP:0001344), and “Apraxia” (HP:0002186). This hypothetical example highlights that the phenotypic similarity approach used in our study can recapitulate the clinical recognition of specific phenotypes such as Rett Syndrome. Mapping the overall phenotypic similarity onto the overall distribution of median phenotypic similarities in n=6 individuals demonstrates how an increasing number of Rett Syndrome-related HPO terms results in increasing phenotypic similarity that shifts to the right with the addition of more phenotypic terms (**Figures S7**).

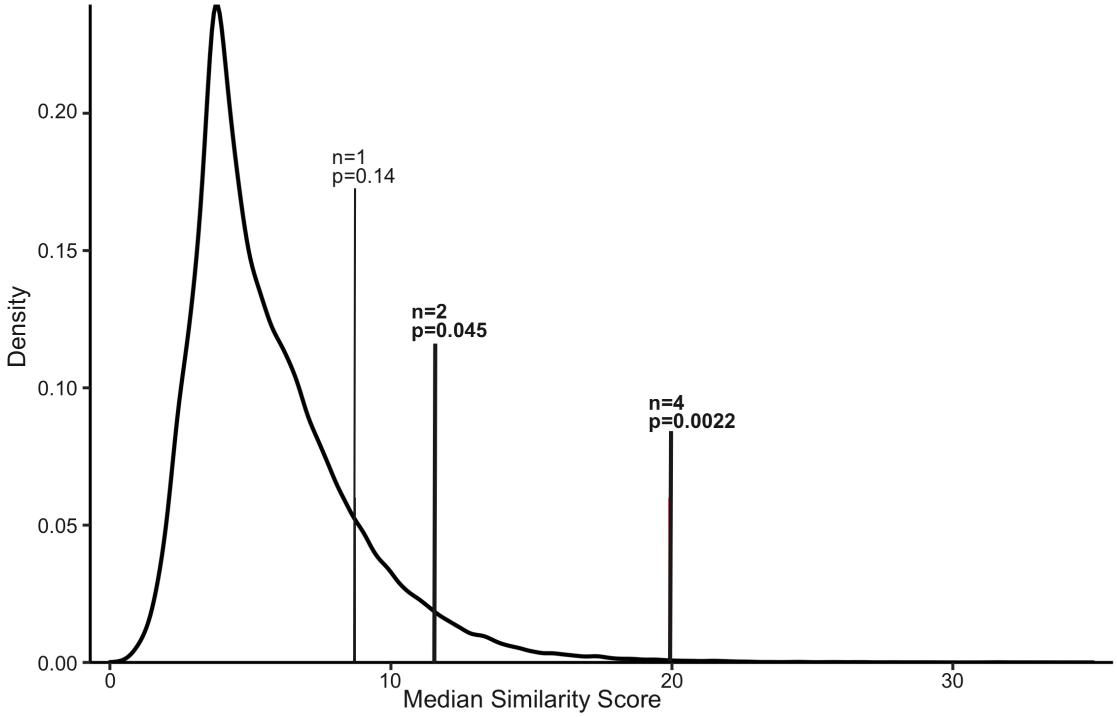


Figure S7. Simulation of phenotypic similarity for Rett Syndrome for n=6 individuals with n=1, n=2, and n=4 phenotypic terms. The curve indicates the distribution of median similarity scores for n=6 individuals within the cohort of 846 individuals included in the current study. With increasing number or Rett Syndrome-related HPO terms, the median phenotypic similarity between six simulated individuals with Rett Syndrome increases and moves further to the right of the curve. A p-value of 0.0022 indicates that a median similarity of 19.9 or higher is only observed in 220/100,000 randomly assessed combinations of six individuals in the cohort of 846 individuals, thereby generating an exact p-value of 0.0022. This example highlights the ability of phenotypic similarity approaches to recapitulate historical examples where constellations of phenotypic features were correctly mapped to a common genetic etiology.

Gene	PPV	Terms (n)	Cumulative frequency	Individuals with etiology	HPO ID	HPO term	Freq.
DNM1	0.80	4	0.41	5	HP:0012444	Brain atrophy	0.80
					HP:0007367	Atrophy/Degeneration affecting the CNS	0.80
					HP:0002977	Aplasia/Hypoplasia involving the CNS	0.80
					HP:0010847	EEG with spike-wave complexes (<2.5 Hz)	0.80
KCNB1	0.81	5	0.07	6	HP:0011442	Abnormality of central motor function	0.83
					HP:0011443	Abnormality of coordination	0.50
					HP:0000729	Autistic behavior	0.50
					HP:0000708	Behavioral abnormality	0.67
					HP:0000234	Abnormality of the head	0.50
SCN1A	0.90	5	0.23	16	HP:0002373	Febrile seizures	0.81
					HP:0002069	Generalized tonic-clonic seizures	0.94
					HP:0003593	Infantile onset	0.81
					HP:0010850	EEG with spike-wave complexes	0.75
					HP:0011153	Focal motor seizure	0.50
STXBP1	0.87	5	0.63	14	HP:0002167	Neurological speech impairment	0.86
					HP:0000750	Delayed speech and language development	0.86
					HP:0001263	Global developmental delay	1.00
					HP:0011446	Abnormality of higher mental function	0.86
					HP:0012758	Neurodevelopmental delay	1.00
AP2M1	0.86	6	0.18	4	HP:0001252	Muscular hypotonia	0.75
					HP:0000750	Delayed speech and language development	0.75
					HP:0011463	Childhood onset	0.75
					HP:0010819	Atonic seizures	0.75
					HP:0000708	Behavioral abnormality	0.75
					HP:0003808	Abnormal muscle tone	0.75
CHD2	0.84	6	0.12	4	HP:0002133	Status epilepticus	0.75
					HP:0011463	Childhood onset	0.75
					HP:0000708	Behavioral abnormality	0.75
					HP:0001249	Intellectual disability	0.75
					HP:0002373	Febrile seizures	0.50
					HP:0002123	Generalized myoclonic seizures	0.75
GABRB3	0.85	7	0.03	5	HP:0100022	Abnormality of movement	0.80
					HP:0003593	Infantile onset	0.80
					HP:0010847	EEG with spike-wave complexes (<2.5 Hz)	0.60
					HP:0000708	Behavioral abnormality	0.60
					HP:0001298	Encephalopathy	0.40
					HP:0001263	Global developmental delay	0.80
					HP:0007270	Atypical absence seizure	0.40
KCNT1	0.87	7	0.01	4	HP:0012444	Brain atrophy	0.50
					HP:0007367	Atrophy/Degeneration affecting the CNS	0.50
					HP:0007359	Focal-onset seizure	0.75
					HP:0002977	Aplasia/Hypoplasia involving the CNS	0.50
					HP:0002060	Abnormality of the cerebrum	0.50
					HP:0010841	Multifocal epileptiform discharges	0.50
					HP:0100547	Abnormality of forebrain morphology	0.50
SCN2A	0.90	7	0.04	8	HP:0000729	Autistic behavior	0.50
					HP:0001252	Muscular hypotonia	0.62
					HP:0002011	Morphological abnormality of the CNS	0.75
					HP:0012639	Abnormality of nervous system morphology	0.75
					HP:0003808	Abnormal muscle tone	0.62
					HP:0011804	Abnormal muscle physiology	0.62
					HP:0003011	Abnormality of the musculature	0.62

SCN8A	0.84	7	0.14	5	HP:0002069	Generalized tonic-clonic seizures	1.00
					HP:0003593	Infantile onset	0.80
					HP:0007359	Focal-onset seizure	0.80
					HP:0002384	Focal impaired awareness seizure	0.60
					HP:0002133	Status epilepticus	0.60
					HP:0001252	Muscular hypotonia	0.60
					HP:0410280	Pediatric onset	1.00
CDKL5	0.86	8	0.04	4	HP:0011097	Epileptic spasms	1.00
					HP:0011196	EEG with focal sharp waves	0.75
					HP:0003593	Infantile onset	0.75
					HP:0002521	Hypsarrhythmia	0.75
					HP:0012469	Infantile spasms	0.75
					HP:0007270	Atypical absence seizure	0.50
					HP:0010819	Atonic seizures	0.50
					HP:0002121	Absence seizure	0.50
KCNQ2	0.84	8	0.01	9	HP:0001298	Encephalopathy	0.56
					HP:0001263	Global developmental delay	0.78
					HP:0010818	Generalized tonic seizures	0.56
					HP:0001252	Muscular hypotonia	0.44
					HP:0000152	Abnormality of head or neck	0.33
					HP:0012759	Neurodevelopmental abnormality	0.89
					HP:0012758	Neurodevelopmental delay	0.78
					HP:0001288	Gait disturbance	0.22
NEXMIF	0.89	8	0.08	4	HP:0002123	Generalized myoclonic seizures	1.00
					HP:0010819	Atonic seizures	0.75
					HP:0001249	Intellectual disability	0.75
					HP:0010850	EEG with spike-wave complexes	0.75
					HP:0012758	Neurodevelopmental delay	1.00
					HP:0011446	Abnormality of higher mental function	0.75
					HP:0007270	Atypical absence seizure	0.50
					HP:0011196	EEG with focal sharp waves	0.50

Table S12. HPO terms required to reach a positive predictive value (PPV) of at least 80% for genetic etiologies with more than three individuals in the cohort

Genetic etiologies with more than three patients in the cohort were found to require 4-8 terms to reach a PPV of at least 80%. HPO terms were selected by taking terms that were had the highest odds ratio (i.e. the most common term within those individuals with the genetic etiology as compared to the rest of the cohort). Frequency of each term within individuals with the genetic etiology is displayed.