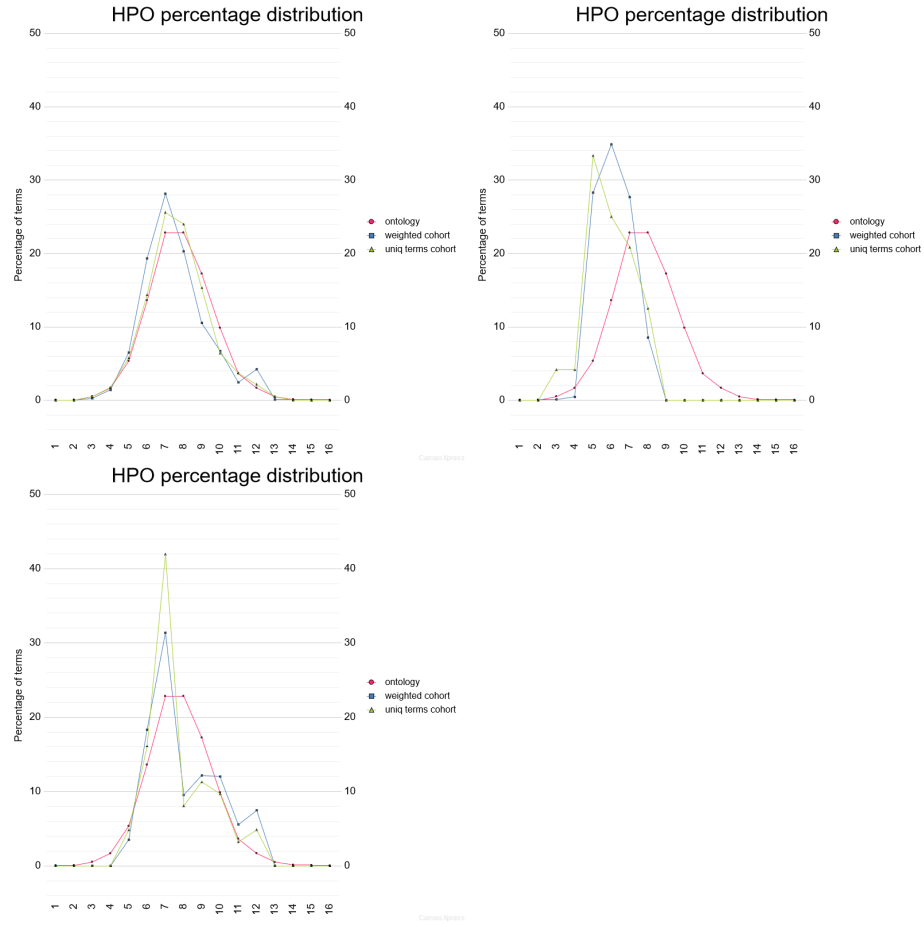


Supplementary Table S2: Top five patient clusters from the ID/MCA cohort using the Naïve method, including HPO terms and their frequencies for each cluster

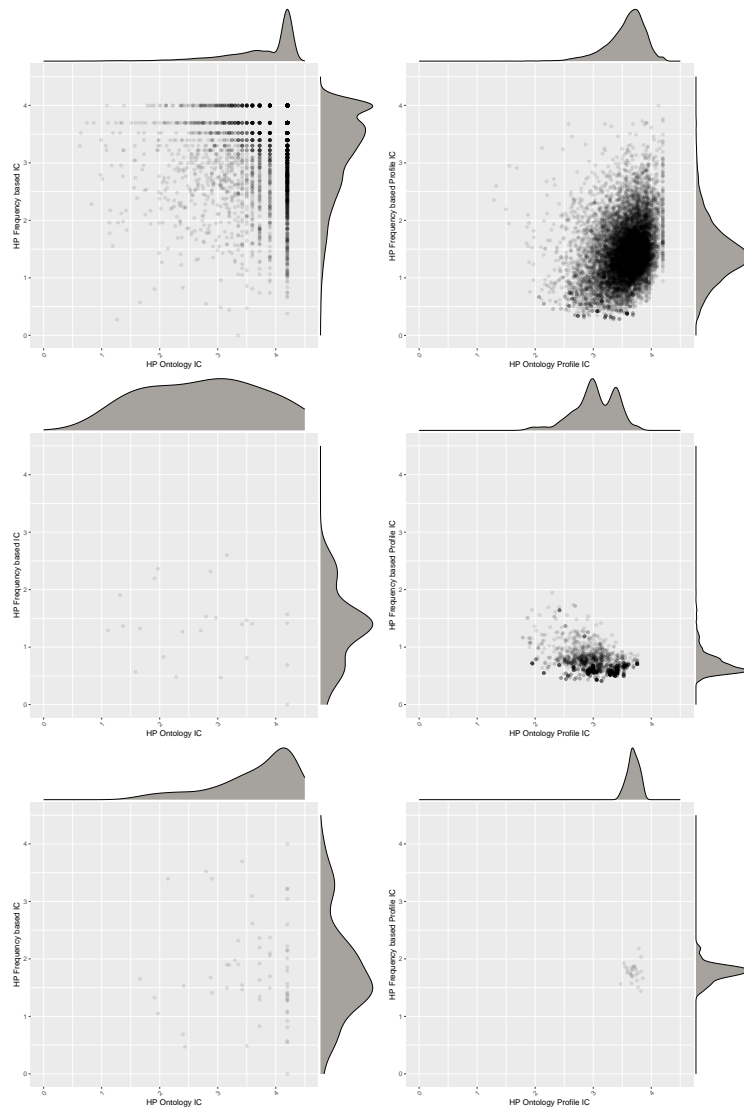
Cluster ID	Patients	Phenotypes	Frequency (%)
0	688	Cognitive impairment	100.0
1	137	Cognitive impairment, Intellectual disability, mild	100.0, 100.0
2	106	Cognitive impairment, Autistic behavior	100.0, 100.0
3	259	Cognitive impairment, Behavioral abnormality	100.0, 100.0
4	255	Cognitive impairment, Intellectual disability, moderate	100.0, 100.0

Supplementary Table S3: Top five most frequent HPO terms in the DECIPHER, ID/MCA and PMM2-CDG datasets. Patients with two or less HPO terms in their profile were removed

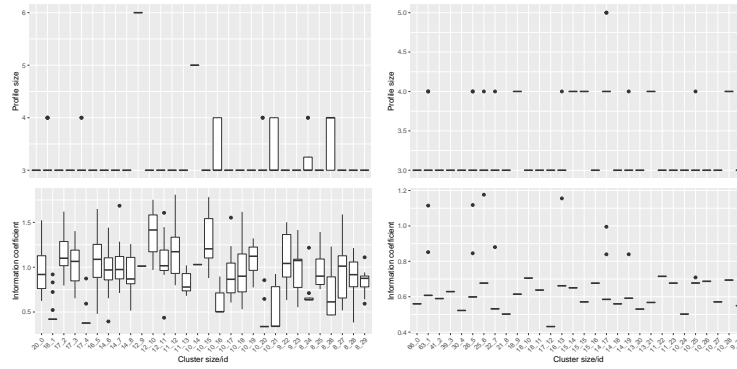
	DECIPHER	%	ID/MCA	%	PMM2-CDG	%
1	Intellectual disability	42.593	Cognitive impairment	93.426	Cerebellar atrophy	100.0
2	Delayed speech and language development	18.872	Intellectual disability, mild	29.244	Upslanted palpebral fissure	88.888
3	Microcephaly	14.754	Intellectual disability, severe	28.933	High, narrow palate	85.185
4	Hypotonia	14.621	Short stature	27.173	Strabismus	81.481
5	Global developmental delay	14.031	Seizure	26.501	Anteverted nares	74.074



Supplementary Figure S1: HPO terms percentage distribution plots for A) DE-CIPHER, B) ID/MCA and C) PMM2-CDG cohorts. Green curves represent unique HPO terms used to describe the cohort, blue curves represent the frequency of each term and pink curves are the number of terms included in each HPO level. Patients with two or less HPO terms in their profile were removed from cohort.



Supplementary Figure S2: Information content (IC) distribution for A) DECIPHER, B) ID/MCA and C) PMM2-CDG cohorts. Left figures correspond to HPO terms within HPO and right figures to phenotype profiles in each cohort. To compare the cohorts a general IC table was computed for the “Frequency based IC” axis with all patients used in the three cohorts with at least three assigned HPO terms. Patients with two or less HPO terms in their profile were removed from cohort.



Supplementary Figure S3: Information content (IC) distribution for the top 30 clusters calculated with Cohort Analyzer. Upper figures: HPO profile size for each cluster. Lower figures: IC distribution. A) DECIPHER and B) ID/MCA. Patients with two or less HPO terms in their profile were removed from cohort.