

1 **Polycomb-associated and Trithorax-associated developmental conditions –**
2 **phenotypic convergence and heterogeneity**

3

4

Alice Smail^{1,2}, Eema Jawahiri¹, Kate Baker^{1,3*}

5

6 1. MRC Cognition and Brain Sciences Unit, University of Cambridge, UK

7 2. Department of Medical & Molecular Genetics, Kings College London, UK

8 3. Department of Medical Genetics, University of Cambridge, UK

9

10 * Corresponding author: kate.baker@mrc-cbu.cam.ac.uk

11

12 Funding support – UKRI/MRC (MC_UU_00030/3 to KB)

13

14 Word count 3718

15

16

17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Abstract

Polycomb group (PcG) and Trithorax group (TrxG) complexes represent two major components of the epigenetic machinery. This study aimed to delineate phenotypic similarities and differences across developmental conditions arising from rare variants in PcG and TrxG genes, using data-driven approaches.

462 patients with a PcG or TrxG-associated condition were identified in the DECIPHER dataset. We analysed Human Phenotype Ontology (HPO) data to identify phenotypes enriched in this group, in comparison to other monogenic conditions within DECIPHER. We then assessed phenotypic relationships between single gene diagnoses within the PcG and TrxG group, by applying semantic similarity analysis and hierarchical clustering. Finally, we analysed patient-level phenotypic heterogeneity in this group, irrespective of specific genetic diagnosis, by applying the same clustering approach. Collectively, PcG/TrxG diagnoses were associated with increased reporting of HPO terms relating to integument, growth, head & neck, limb and digestive abnormalities. Gene group analysis identified three multi-gene clusters differentiated by microcephaly, limb/digit dysmorphologies, growth abnormalities and atypical behavioural phenotypes. Patient-level analysis identified two large clusters differentiated by neurodevelopmental abnormalities and facial dysmorphologies respectively, as well as smaller clusters associated with more specific phenotypes including behavioural characteristics, eye abnormalities, growth abnormalities and skull dysmorphologies. Importantly, patient-level phenotypic clusters did not align with genetic diagnoses. Data-driven approaches can highlight pathway-level and gene-level phenotypic convergences, and individual-level phenotypic heterogeneities. Future studies are needed to understand the multi-level mechanisms contributing to both convergence and variability within this population, and to extend data collection and analyses to later-emerging health characteristics.

Keywords: Polycomb, Trithorax, epigenetic, Mendelian, phenotypes, hierarchical clustering

48
49

Introduction

50 Chromatin remodelling is a fundamental epigenetic mechanism that orchestrates development and
51 maintains the stability of differentiated tissues. Polycomb group (PcG) and Trithorax group (TrxG) multi-
52 subunit complexes have key roles in chromatin remodelling, and act antagonistically to regulate
53 chromatin accessibility by inducing the repression and activation of gene expression respectively ^{1,2}. PcG
54 complexes mediate transcriptional repression by promoting nucleosomal compaction, as well as
55 deterring TrxG action ^{1,3,4}. TrxG complexes (including COMPASS methyltransferases and BAF (SWI/SNF)
56 remodelling complexes) maintain active gene expression and oppose PcG action by relaxing
57 nucleosomal compaction ^{1,3,4}. A tightly regulated balance of PcG and TrxG activity is thus necessary for
58 the precise timing of gene expression during mammalian development; in particular, the PcG-TrxG
59 system has a vital role in neurogenesis, by mediating the balance of neural progenitor cell differentiation
60 and self-renewal ⁵.

61 Multiple developmental conditions arise from pathogenic variants in PcG and TrxG genes, forming a
62 subset of Mendelian Disorders of Epigenetic Machinery (MDEMs) ⁶. Whilst individually rare, PcG and TrxG-
63 related conditions collectively account for approximately 8% of monogenic developmental disorders in
64 the DECIPHER database ⁷. These conditions are phenotypically heterogeneous, with variability within and
65 across different genetic diagnoses. It has been suggested that while MDEMs are monogenic, they are
66 conceptually similar to complex disorders, with epigenomic dysregulations contributing to an array of
67 different phenotypes and severities via downstream pathways ⁸. Patients with different PcG and TrxG-
68 related conditions can share overlapping phenotypic characteristics: for example, broad similarities can
69 be observed between BAF- and COMPASS-complex associated conditions ⁹⁻¹⁴. It is plausible that
70 phenotypic similarities across conditions are due to epigenomic dysregulation at shared genomic loci
71 and at similar developmental stages, leading to downstream effects on gene expression that affect
72 convergent developmental pathways. However, there are also phenotypic divergences within this group:
73 for example, opposing features relating to growth have been observed amongst PcG-related conditions
74 arising from variants in PRC1, PRC2 and PR-DUB complex genes ^{2,5}. Phenotypic descriptions of each
75 MDEM in isolation may not fully capture the extent of overlap or distinctiveness between conditions, and
76 may underplay the extent of variability within each condition. A phenotypic evidence base that accounts
77 for this complexity could improve post-diagnostic counselling and stimulate hypothesis driven, clinically
78 relevant research.

79 The current paper reports data-driven analysis, aiming to describe the phenotypic commonalities and
80 heterogeneities of PcG and TrxG-related conditions. Our first objective was to identify phenotypes that
81 occur at elevated frequencies across the population of patients with PcG and TrxG-associated conditions,
82 in comparison to other monogenic conditions that are present in the DECIPHER dataset. If phenotypic
83 enrichments exist for this broad group of conditions, this reinforces the evidence for pathway-level
84 convergence in developmental mechanisms and clinical needs. Our second objective was to map gene-
85 level heterogeneity within PcG and TrxG-related conditions. To do this, we investigated whether there are
86 phenotypic clusters of PcG and TrxG-associated conditions when patient data are grouped by affected
87 gene. If gene-driven clusters are present, this supports the recognition of clinical syndromes aligning with
88 variants across several specific genes, and may point to shared epigenomic dysregulations. Our third
89 objective was to map patient-level heterogeneity to explore whether there are phenotypic clusters of
90 patients with PcG and TrxG-associated conditions; the results of this final analysis may support or
91 challenge the alignment between MDEM-associated phenotype co-occurrences and variants in specific
92 genes.

93

94 **Materials and Methods**

95 *Gene List Curation*

96 Gene Ontology (GO) annotations were used to collate a comprehensive PcG and TrxG gene list
97 (Supplementary Figure 1A), following a similar approach to Ciptasari and van Bokhoven¹⁵. GO Cellular
98 Component terms selected to represent PcG complexes were 'PRC1 complex' (GO:0035102), 'ESC/E(Z)
99 complex' (synonymous with 'PRC2 complex'; GO:0035098) and 'PR-DUB complex' (GO:0035517). To
100 gather terms related to SWI/SNF and COMPASS complexes, several child terms of 'SWI/SNF
101 superfamily-type complex' (GO:0070603) and three terms that correspond to the three subtypes of
102 COMPASS complex were selected. After collating 114 genes using relevant GO terms, a further 21 genes
103 that are likely to possess PcG/TrxG involvement were added based on literature reviews, and 6 genes
104 that lack direct or specific PcG/TrxG involvement were removed (Supplementary Table 1). This resulted in
105 a final list of 129 PcG/TrxG genes.

106 *Cohort and phenotypes curation*

107 The DECIPHER database is a repository of genotypic and Human Phenotype Ontology (HPO) standardised
108 phenotypic data deposited by clinical geneticists and laboratory scientists^{7,16}. The DECIPHER open

109 access dataset was filtered using the pre-defined PcG/TrxG gene list, to identify all pathogenic or likely
110 pathogenic sequence variants located in a PcG-related or TrxG-related gene (Supplementary Figure 1B).
111 506 variants that were present across 499 patients were identified; importantly, all patients with more
112 than one variant had variants present in the same gene. Patients with 0 HPO terms were removed
113 (n=37), resulting in a cohort of 462 patients with pathogenic or likely pathogenic variants across 38
114 PcG/TrxG genes. 99 patients had PcG variants, 359 patients had TrxG variants, and 4 patients had
115 variants in *HCFC1* which has roles in both PcG (PR-DUB accessory subunit) and TrxG (SET1A/B-COMPASS
116 complex subunit). All patients with autosomal variants were heterozygous (n=447), while all patients with
117 X-linked variants were hemizygous (n=15). The majority (81%) of patients in this cohort had a *de novo*
118 variant, while 13% had unknown inheritance and 6% had a maternally or paternally inherited variant. To
119 obtain a comparison cohort, the remainder of the DECIPHER dataset was filtered to include unique
120 patients with pathogenic or likely pathogenic sequence variants, and patients with 0 HPO terms were
121 removed (n=5070).

122 *Group-level HPO enrichment analysis*

123 To address the first study objective, the HPO terms reported for each patient were propagated using the
124 Python package PyHPO, so that broader 'top-level/parent' phenotypes could be compared across groups,
125 in addition to more specific 'lower-level/child' phenotypes¹⁶. The median number of reported HPO terms,
126 and top-level HPO terms for each patient in the PcG/TrxG and comparison cohorts were compared via
127 Mann-Whitney U rank tests. We then compared the frequency of propagated HPO terms between the
128 PcG/TrxG and comparison cohorts, using a two-proportions z-test with Benjamini-Hochberg correction for
129 multiple testing (Figure 1a). First, we examined whether any top-level HPO terms were enriched in the
130 PcG/TrxG population: for this analysis the top-level term 'Abnormality of the musculoskeletal system'
131 was split into the descendent terms 'Abnormality of the skeletal system', 'Abnormality of the
132 musculature' and 'Abnormality of connective tissue', which resulted in a total of 25 top-level HPO terms
133 (Supplementary Table 2). Second, we examined lower-level term enrichments.

134 *Gene-level HPO analysis*

135 Figure 1b details the approach taken to examine gene-associated phenotypes within PcG- and TrxG-
136 related conditions. This involved grouping patients by genetic condition, identifying a set of
137 representative HPO terms of each gene group, and computing semantic similarity scores, before
138 performing cluster analysis to identify phenotypic similarity between gene groups. To identify

139 'representative' HPO terms for each gene group, unique propagated HPO terms were collated, and
140 depending on gene group size, HPO terms present in at least 2 patients or 20% of patients were retained
141 (Supplementary Table 3). At this stage, each gene group contained a variable number of HPO terms, with
142 some groups containing higher or lower number of terms. This could impact the cluster analysis
143 outcomes by biasing the gene groups with similar numbers of HPO terms to artefactually cluster
144 together. To reduce this bias, the number of HPO terms per group were approximately equalised. Based
145 on the interquartile range of the number of terms across all gene groups, the cut off for the minimum
146 and maximum number of terms retained per gene group were determined (i.e., removing outliers). The
147 HPO terms retained within the set limit were prioritised based on their information content (IC)^{17,18}. The
148 IC of HPO terms is a measure of the rarity/ specificity of a term within a database, where the IC of term t
149 is given by, $IC(t) = -\log p(t)$, and $p(t)$ is the probability of occurrence of t. Thus, terms with higher IC reflect
150 more specific and clinically informative terms. As such, in this analysis, terms with higher IC were
151 prioritised to be retained within the IQR set limit of number of terms.

152 Next, the semantic similarity scores and pairwise gene group similarity scores for each pair of gene
153 groups was computed, using the graph IC method followed by the funSimAvg method¹⁸. This approach
154 takes into account all common ancestors of pairs of terms. The resulting similarity matrix was used as
155 input for clustering. The method of cluster analysis chosen for this study was agglomerative hierarchical
156 clustering using complete linkage; hierarchical clustering does not require a pre-specified number of
157 clusters, which is useful for this study, as the number of different phenotypic profiles across the cohort
158 was not known before clustering. Clustering was carried out using scikit-learn and SciPy packages in
159 Python, and number of clusters was selected via visual inspection of the dendrogram¹⁹. To illustrate
160 phenotype sharing between genes which leads to clustering, networks of semantic similarity were
161 visualised using Cytoscape. The method used to select HPO terms for network visualisation is detailed in
162 Supplementary Table 4: in brief, terms were included if present in at least 4 gene groups, terms with
163 listed descendent terms were excluded, and some terms were merged to provide clearer visualisation.

164 *Patient-level HPO analysis*

165 Figure 1c outlines the second approach for identifying phenotypic clusters, which involved directly
166 computing the semantic similarity between patients, prior to performing hierarchical clustering. Raw
167 HPO terms associated with each patient were collated. Modifier terms as well as terms with descendent
168 terms were removed (i.e. parent term was removed if its child term is present). Similar to the gene-level

169 pre-processing approach, the number of HPO terms associated with each patient were approximately
170 equalised by identifying the upper outlier term frequency ($Q3 + 1.5 \times IQR$) across all patients and using
171 this as an upper limit with terms prioritised by highest IC. The semantic similarity between each pair of
172 patients was then computed, before performing hierarchical clustering as described previously. Following
173 clustering, HPO terms associated with each patient were propagated and the proportion of patients with
174 each term in each cluster was compared to the proportion of patients with each term outside of each
175 cluster, using Fisher's exact test with a Benjamini-Hochberg correction. HPO terms that were significantly
176 enriched ($p_{adj} < 0.05$) in each cluster were identified.

177

178 **Results**

179 *Group-level HPO enrichment analysis*

180 The number of HPO terms per patient for the PcG- and TrxG-associated conditions and comparison
181 cohorts are displayed in Supplementary Figure 2. There is a significant distribution shift ($p=1.51e-10$)
182 towards a higher number of reported HPO terms for patients with PcG- and TrxG-associated conditions
183 (median = 7) relative to the remainder of the DECIPHER dataset (median = 6). Similarly, using top-level
184 HPO terms as a proxy for the number of organ systems affected, there is a significant shift ($p=1.51e-13$)
185 towards a higher number of top-level HPO terms across patients with PcG- and TrxG-associated
186 conditions (median = 5) relative to the remainder of the DECIPHER dataset (median = 4). Together, these
187 findings indicate an increased number of phenotypes per patient, and a larger breadth of affected
188 organs/organ systems, in patients with PcG- and TrxG-associated conditions compared to other
189 monogenic conditions.

190 Using propagated terms for each patient, the percentage occurrence of top-level HPO terms among
191 patients with PcG and TrxG-associated conditions was compared to the percentage occurrence of each
192 term across the remainder of the DECIPHER dataset. Five top-level terms were significantly increased at
193 $p_{adj} < 0.005$ among patients with PcG and TrxG-associated conditions (Figure 2, Supplementary Table 5):
194 'Abnormality of the integument' ($p_{adj}=2.04e-09$), 'Growth abnormality' ($p_{adj}= 5.44e-08$), 'Abnormality of
195 head or neck' ($p_{adj}=1.33e-06$) 'Abnormality of the digestive system' ($p_{adj}= 4.04e-06$), 'Abnormality of
196 limbs' ($p_{adj}=0.0012$). 'Abnormality of the nervous system' was also marginally increased at $p_{adj}=0.019$.
197 No top-level terms were significantly reduced in the PcG/TrxG group. The distribution of each top-level
198 term across each genetic condition is illustrated in Supplementary Figure 4; it can be observed that a

199 substantial proportion of patients with each PcG and TrxG-associated condition have abnormalities of the
200 head or neck, integument, limbs and nervous system, while growth abnormalities and digestive system
201 abnormalities are more unevenly distributed.

202 Supplementary Table 6 lists specific HPO terms which were significantly increased or decreased at
203 $p_{\text{adj}} < 0.05$, and that have a percentage difference of at least $\pm 3\%$ between the PcG/TrxG group and
204 comparison group. 66 descendent terms of the 6 significantly enriched top-level terms were also
205 significantly enriched in the PcG/TrxG group, and 2 descendent terms were significantly reduced
206 ('Seizure' and 'Abnormality of skull size'; Figure 2). 5 additional terms not descended from enriched top-
207 level terms were significantly increased in the PcG/TrxG group. In summary, despite the genetic
208 heterogeneity of PcG-related and TrxG-related conditions, there are detectable differences between this
209 amalgamated population and the broader developmental disorders population: we observed differences
210 in the median number of reported phenotypes, and differences across broad and specific phenotype
211 frequencies.

212 *Gene-level HPO analysis: semantic similarity, clustering and network analysis*

213 Representative HPO terms for each genetic diagnosis are listed in Supplementary Table 3. The
214 computed semantic similarity matrix was used to perform hierarchical clustering (Figure 3A), resulting in
215 three multi-gene clusters and 2 stand-alone gene groups. We examined clusters for potential bias arising
216 from HPO term frequencies across each cluster (Supplementary Figure 3). Network analysis was applied
217 to visualise the broad shared phenotypic characteristics of each cluster – for visualisation purposes,
218 networks are presented separately for "Abnormalities of the nervous system" (Figure 3B) and descendent
219 terms of "Head and neck abnormality", "Abnormality of limbs", "Abnormality of the integument" and
220 "Growth abnormality" (Figure 3C). Broadly, it can be observed that genes within cluster 1 share growth
221 abnormalities (including terms relating to overgrowth and restricted growth) and atypical behaviour;
222 genes within cluster 3 share abnormalities of brain morphology, particularly microcephaly; genes within
223 clusters 2 and 3 share limb abnormalities and facial dysmorphology. There was no clear pattern of
224 PcG/TrxG complex membership aligning with phenotypic cluster membership i.e. COMPASS and PcG
225 associated genes were distributed across all three clusters, while BAF complex genes were clustered into
226 clusters 2 and 3.

227 *Patient-level HPO analysis: phenotypic similarity, clustering and cluster comparisons*

228 To identify clusters of patients with phenotypic similarity within the PcG- and TrxG-associated conditions
229 population, irrespective of specific genetic diagnosis, hierarchical clustering was performed on the
230 unpropagated HPO terms for each individual patient. Following inspection of the dendrogram (Figure 4A),
231 the population of patients was divided into 12 clusters, varying in size between 3 and 111 patients; this
232 number of clusters was selected as it provides good insight into sub-clusters across the population,
233 without separating larger clusters into smaller very similar clusters. Supplementary Figure 3 shows the
234 HPO term numbers across each cluster. The phenotypic profile of each of the 12 clusters was then
235 analysed using propagated patient-level terms, with significantly enriched terms identified at $p_{adj} < 0.05$
236 (Supplementary Table 7). Figure 4B summarises the two most significant HPO terms for each cluster.
237 Cluster 1 contains the largest number of patients (n=111) and is enriched for a collection of
238 neurodevelopmental abnormalities, particularly neurological speech impairment, in addition to abnormal
239 muscle tone. Cluster 2 (n=91) is characterised by various abnormal facial morphologies, and Cluster 3
240 (n=48) has a higher proportion of patients with behavioural characteristics, including autistic behaviour.
241 Cluster 4 (n=48) is characterised by abnormalities of the eye, while Cluster 5 (n=32) contains patients
242 with growth abnormalities – both undergrowth (short stature, growth delay) and overgrowth (tall stature,
243 overgrowth). Cluster 6 (n=30) is associated with skull dysmorphologies, particularly microcephaly. The
244 remaining clusters contained 10 or fewer patients and were each defined by more specific HPO terms.

245 Supplementary Figure 5 illustrates the distribution of cluster memberships for patients within each PcG-
246 and TrxG gene diagnosis group. It can be observed that the majority of gene groups are phenotypically
247 heterogeneous, and that phenotype clusters are distributed across genes. There are some exceptions,
248 namely CHD8 (>50% patients within Cluster 11), SETD1A (>50% patients within Cluster 1) and BCOR
249 (>50% patients within Cluster 12). However the absolute number of patients with each of these genetic
250 diagnoses is low, hence this variation in proportional cluster membership may occur by chance.

251 **Discussion**

252 This paper presents data-driven analyses of the phenotypic landscape of PcG- and TrxG-associated
253 conditions. These analyses provide important insights into clinical convergences and divergences within
254 this population, relevant to diagnosis, post-diagnostic counselling and management. Moreover, this study
255 highlights opportunities and limitations of HPO computational methods for investigating disorder-
256 associated pathways defined by convergent gene function.

257 We first assessed group-level differences in HPO term frequencies between PcG/TrxG-associated
258 conditions and other monogenic developmental disorders. PcG/TrxG-associated conditions shared multi-
259 system characteristics, with significantly increased frequency of abnormalities of growth, limbs, digestive
260 system, integument and head or neck. More specific HPO terms reported in >20% of PcG/TrxG
261 individuals include abnormal skin or hair morphology, abnormal ocular or oral morphology, and
262 abnormal hands or upper limbs. These significant enrichments are useful phenotypic markers for
263 confirming pathogenicity of PcG/TrxG variants. On the other hand, occurrences of each specific
264 characteristic affect a minority of the population, emphasising the variability of this population, and need
265 to assess combinations of phenotypes rather than each in isolation.

266 94% of patients with PcG/TrxG variants are reported to have nervous system abnormality, a modest
267 elevation compared to other conditions (89%). 'Global developmental delay' (54% cases, 46% controls)
268 and 'Neurological speech impairment' (25% cases, 19% controls) were also enriched, but terms relating
269 to severity of developmental delay did not differ between groups. However, the reporting of specific
270 cognitive and behavioural characteristics is sparse within DECIPHER, and HPO terminology is not
271 designed for quantification of neurodevelopmental differences. 'Abnormal cerebral white matter
272 morphology' is an enriched phenotype within the case group highlighting convergence on structural brain
273 connectivity; seizures are reported at lower frequency (10% cases, 20% controls). Given the prevalence
274 and variability of neurodevelopmental phenotypes within the PcG/TrxG group, investigation of multi-level
275 mechanisms linking epigenetic regulation to brain development and cognitive vulnerabilities are
276 warranted.

277 Building on group-level analysis of phenotype enrichments, our next goal was to assess phenotypic
278 variability within PcG/TrxG-related conditions. We observed three clusters of genes which share broad
279 top-level terms but differ in frequency of lower-level terms. Genes within clusters show multi-phenotype
280 similarity to their within-cluster neighbours, but sparse connectivity to genes in other clusters. The
281 phenotypes driving cluster separation within this dataset are likely to be partially influenced by the
282 reporting biases of DECIPHER. However, these results are potentially hypothesis-setting for mechanistic
283 studies: for example, the high phenotypic similarity between conditions associated with BAF, COMPASS
284 and PcG subunits may reflect convergent impacts on gene expression⁸. However we also observed
285 contrasting phenotypic profiles associated with functionally similar genes (for example, KDM6A and
286 KDM6B); these differences may be influenced by variant-specific, cell-specific or timing-specific effects.

287 The pooled phenotypic profile for each gene does not take into account patient-level heterogeneity²⁰.
288 Therefore, we carried out a second clustering analysis including all patients with at least 4 reported HPO
289 terms. We identified 12 phenotypically-defined clusters, which do not align with genetic diagnoses.
290 Smaller clusters are driven by low prevalence phenotypes, whereas larger clusters are characterised by
291 broad and frequent terms, with more subtle between-cluster differences. Phenotype cluster
292 memberships may be influenced by unanalysed factors, such as the age of patients at phenotyping, and
293 variable reporting by clinical geneticists, developmental paediatricians or neurologists. Importantly, gene
294 groups were distributed across multiple clusters, and the majority of clusters were genetically
295 heterogeneous, suggesting that phenotyping is not overly influenced by prior expectation of gene-specific
296 features.

297 These results reinforce the complexity of MDEMs and the potential for early developmental mechanisms
298 which cascade to system-level consequences that are not predictable by gene alone. However,
299 limitations of our analyses include variable patient numbers across MDEMs in DECIPHER (more than
300 50% of the total group have one of four diagnoses), making it difficult to gauge and compare the
301 phenotypic spectrum of every condition. There are disparities in phenotyping detail across the DECIPHER
302 dataset - while some patients have an extensive list of fine-grained HPO terms, others have a single
303 broad term, such as 'Global developmental delay'; this may reflect the reality of a patient's phenotype, or
304 variations in phenotyping standards. However, we mitigated this bias by prioritising terms with higher IC
305 (and therefore specificity), and application of a term equalisation step in our pre-processing to mitigate
306 potential biases arising from disparities in HPO term numbers. Additionally, facial phenotypes and other
307 specific morphological features can differentiate disorders from each other²¹. The fact that not every
308 patient has a comprehensive list of specific phenotypes means that subtle dysmorphic differences will
309 not contribute to our clustering analyses.

310 Our analyses focused on a subset of MDEMs, and future studies may be more informative by taking a
311 more inclusive strategy to gene selection; we focused on the well-defined PcG and TrxG-associated
312 complexes, while there is uncertainty as to what constitutes an 'epigenetic' protein involved indirectly in
313 chromatin structural regulation. Despite this, there remains some ambiguity in our defined gene list: for
314 example, while USP7 interacts with PRC1 variants, it may not strictly be considered a PcG protein²².
315 Several genes have additional roles beyond the PcG-TrxG system, for example KDM2B functions as a
316 demethylase at several histone sites, as well as mediating ubiquitination²³. There are also limitations

317 arising from the classification of variants: this study relies on clinical reporting, and variants that may
318 only partially explain phenotype were included. Additionally, this study did not take into account variant
319 consequence; gain-of-function and loss-of-function variants may have different phenotypic
320 consequences^{24,25}.

321 One motivation for elucidating convergent clinical characteristics is that emerging treatment strategies
322 may apply to a larger cohort beyond single genes. Ciptasari and van Bokhoven suggest that
323 transcriptomic convergences across different MDEMs could be harnessed to develop effective
324 interventions¹⁵. Drugs targeting epigenetic mechanisms or downstream gene expression have the
325 potential to be efficacious in treating MDEMs⁶. HDAC inhibitors normalised H3K4 methylation and
326 neurogenesis in a mouse model of Kabuki Syndrome²⁶. CRISPR-based methods may also be applicable
327 to MDEMs²⁷. However, a major challenge is that therapies need to be administered at an appropriate
328 developmental stage. Epigenetic interventions and CRISPR-based therapies risk introducing off-target
329 changes⁶. The current study indicates that patient-relevant symptom domains are highly variable, hence
330 target outcomes and therapeutic strategies will need to weigh-up potential benefits and risks on an
331 individual basis.

332 The data-driven approach of this paper lays the foundation for future work disentangling MDEM-
333 phenotype relationships at scale. Future phenotyping research and clinical initiatives should move
334 beyond congenital characteristics to encompass health-related outcomes across the lifespan.

335 **Data Availability Statement**

336 The dataset analysed during the current study is available in the DECIPHER open access database
337 <https://www.deciphergenomics.org/>

338 **Code Availability**

339 Code used for this project is available at github.com/alicesmail12/HPOAnalysis.

340 **References**

- 341 1. Schuettengruber B, Bourbon HM, Di Croce L, Cavalli G: Genome Regulation by Polycomb and
342 Trithorax: 70 Years and Counting. *Cell* 2017; **171**: 34-57.
343
- 344 2. Doyle LA, Unlu Bektas F, Chatzantonaki E, Repton C, Derrien A, Illingworth RS: RINGs, DUBs and
345 Abnormal Brain Growth-Histone H2A Ubiquitination in Brain Development and Disease.
346 *Epigenomes* 2022; **6**.
347
- 348 3. Kuehner JN, Yao B: The Dynamic Partnership of Polycomb and Trithorax in Brain Development
349 and Diseases. *Epigenomes* 2019; **3**: 17-24.

- 350
351 4. Piunti A, Shilatifard A: Epigenetic balance of gene expression by Polycomb and COMPASS
352 families. *Science* 2016; **352**: aad9780.
353
354 5. Bolicke N, Albert M: Polycomb-mediated gene regulation in human brain development and
355 neurodevelopmental disorders. *Dev Neurobiol* 2022; **82**: 345-363.
356
357 6. Fahrner JA, Bjornsson HT: Mendelian disorders of the epigenetic machinery: postnatal
358 malleability and therapeutic prospects. *Hum Mol Genet* 2019; **28**: R254-R264.
359
360 7. Foreman J, Perrett D, Mazaika E, Hunt SE, Ware JS, Firth HV: DECIPHER: Improving Genetic
361 Diagnosis Through Dynamic Integration of Genomic and Clinical Data. *Annu Rev Genomics Hum*
362 *Genet* 2023; **24**: 151-176.
363
364 8. Luperchio TR, Boukas L, Zhang L et al: Leveraging the Mendelian disorders of the epigenetic
365 machinery to systematically map functional epigenetic variation. *Elife* 2021; **10**.
366
367 9. Tsurusaki Y, Okamoto N, Ohashi H et al: Mutations affecting components of the SWI/SNF
368 complex cause Coffin-Siris syndrome. *Nat Genet* 2012; **44**: 376-378.
369
370 10. Machol K, Rousseau J, Ehresmann S et al: Expanding the Spectrum of BAF-Related Disorders: De
371 Novo Variants in SMARCC2 Cause a Syndrome with Intellectual Disability and Developmental
372 Delay. *Am J Hum Genet* 2019; **104**: 164-178.
373
374 11. Kosho T, Okamoto N, Coffin-Siris Syndrome International C: Genotype-phenotype correlation of
375 Coffin-Siris syndrome caused by mutations in SMARCB1, SMARCA4, SMARCE1, and ARID1A. *Am*
376 *J Med Genet C Semin Med Genet* 2014; **166C**: 262-275.
377
378 12. Sheppard SE, Campbell IM, Harr MH et al: Expanding the genotypic and phenotypic spectrum in a
379 diverse cohort of 104 individuals with Wiedemann-Steiner syndrome. *Am J Med Genet A* 2021;
380 **185**: 1649-1665.
381
382 13. Faundes V, Newman WG, Bernardini L et al: Histone Lysine Methylases and Demethylases in the
383 Landscape of Human Developmental Disorders. *Am J Hum Genet* 2018; **102**: 175-187.
384
385 14. Koemans TS, Kleefstra T, Chubak MC et al: Functional convergence of histone
386 methyltransferases EHMT1 and KMT2C involved in intellectual disability and autism spectrum
387 disorder. *PLoS Genet* 2017; **13**: e1006864.
388
389 15. Ciptasari U, van Bokhoven H: The phenomenal epigenome in neurodevelopmental disorders.
390 *Hum Mol Genet* 2020; **29**: R42-R50.
391
392 16. Kohler S, Gargano M, Matentzoglou N et al: The Human Phenotype Ontology in 2021. *Nucleic*
393 *Acids Res* 2021; **49**: D1207-D1217.
394
395 17. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM: Semantic similarity in biomedical ontologies.
396 *PLoS Comput Biol* 2009; **5**: e1000443.
397
398 18. Deng Y, Gao L, Wang B, Guo X: HPOSim: an R package for phenotypic similarity measure and
399 enrichment analysis based on the human phenotype ontology. *PLoS One* 2015; **10**: e0115692.
400
401 19. Pedregosa F VG, Gramfort A, Michel V, Thirion B, Grisel O, et al.: Scikit-learn: Machine Learning in
402 Python. . *Journal of Machine Learning Research* 2011; **12**: 2825-2830. .
403
404 20. Rots D, Jakub TE, Keung C et al: The clinical and molecular spectrum of the KDM6B-related
405 neurodevelopmental disorder. *Am J Hum Genet* 2023; **110**: 963-978.
406
407 21. Bogershausen N, Wolnik B: Mutational Landscapes and Phenotypic Spectrum of SWI/SNF-
408 Related Intellectual Disability Disorders. *Front Mol Neurosci* 2018; **11**: 252.
409

- 410 22. Bracken AP, Brien GL, Verrijzer CP: Dangerous liaisons: interplay between SWI/SNF, NuRD, and
411 Polycomb in chromatin regulation and cancer. *Genes Dev* 2019; **33**: 936-959.
412
- 413 23. Kang JY, Kim JY, Kim KB *et al*: KDM2B is a histone H3K79 demethylase and induces
414 transcriptional repression via sirtuin-1-mediated chromatin silencing. *FASEB J* 2018; **32**: 5737-
415 5750.
416
- 417 24. Heyn P, Logan CV, Fluteau A *et al*: Gain-of-function DNMT3A mutations cause microcephalic
418 dwarfism and hypermethylation of Polycomb-regulated regions. *Nat Genet* 2019; **51**: 96-105.
419
- 420 25. Tatton-Brown K, Seal S, Ruark E *et al*: Mutations in the DNA methyltransferase gene DNMT3A
421 cause an overgrowth syndrome with intellectual disability. *Nat Genet* 2014; **46**: 385-388.
422
- 423 26. Nothof SA, Magdinier F, Van-Gils J: Chromatin Structure and Dynamics: Focus on Neuronal
424 Differentiation and Pathological Implication. *Genes (Basel)* 2022; **13**.
425
- 426 27. Hilton IB, D'Ippolito AM, Vockley CM *et al*: Epigenome editing by a CRISPR-Cas9-based
427 acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol* 2015; **33**:
428 510-517.
429
430
- 431

432

Acknowledgments

433 This work was supported by UKRI/MRC funding (grant number MC_UU_00030/3). This study makes use
434 of data generated by the DECIPHER community. A full list of centres who contributed to the generation of
435 the data is available from <https://deciphergenomics.org/about/stats> and via email from
436 contact@deciphergenomics.org. DECIPHER is hosted by EMBL-EBI and funding for the DECIPHER project
437 was provided by the Wellcome Trust [grant number WT223718/Z/21/Z]. Those who carried out the
438 original analysis and collection of the data bear no responsibility for the further analysis or interpretation
439 of the data.

440

Author Contribution Statement

441 AS and KB conceived the study. AS carried out data curation and analysis. EJ provided methodological
442 advice. KB and AS wrote the manuscript, and all authors reviewed and approved the final version.

443

Ethical Approval

444 The ethical framework for DECIPHER is provided at
445 https://www.deciphergenomics.org/files/pdfs/decipher_ethical_framework.pdf

446

Competing Interests

447 Nil to declare

448

449 **Figure legends**

450 **Figure 1 - Overview of analysis methods**

451 **Figure 2 - Group-level HPO analysis results.**

452 The proportion of each enriched ($p_{adj} < 0.05$) propagated HPO term among patients with PcG and TrxG-
453 associated conditions is indicated by the y-axis, relative to the remainder of DECIPHER (x-axis). Terms
454 with at least a $\pm 3\%$ change between the PcG/TrxG group and remainder of DECIPHER are shown.
455 Significantly enriched top-level terms are labelled in bold; also labelled are the two terms with the largest
456 positive percentage difference and the two terms with the largest negative difference between the
457 groups.

458 **Figure 3 - Gene-level clustering results.**

459 A) Dendrogram of gene-level clustering output. B) Network illustration of gene relationships within
460 clusters, based on nervous system abnormalities present in at least 4 gene groups. C) Network
461 illustration of gene relationships within clusters, based on abnormalities of physical development present
462 in at least 4 gene groups.

463 **Figure 4 - Patient-level clustering results.**

464 A) Patient-level dendrogram. B) Within-cluster proportional occurrence of top-level terms that were
465 increased across the population as a whole. C) Within-cluster proportional occurrence of 2 representative
466 HPO terms for each of the 12 clusters, which were identified using Fisher's exact test.

Group Analysis

462 patients with P/LP variants
across 38 TrxG & PcG genes



5070 patients with P/LP variants
across 1037 other genes



1. Propagate HPO terms



2. Compare total number of terms per patient
Mann-Whitney U test was used to compare number of raw terms & number of top-level terms per patient.



3. Identify enriched terms

Terms with a significantly different proportion between the two groups were identified using a two-proportions z-test with BH correction.

Gene-Level Analysis

462 patients with P/LP variants
across 38 TrxG & PcG genes



1. Propagate HPO terms

Propagated terms were used for this analysis, as patients in one group may not have the exact same phenotype (ie Generalised hypotonia & Axial hypotonia), but they may have the same parent phenotype (ie Hypotonia).



2. Identify Gene Groups

Group patients by affected gene & identify representative HPO terms for each group by retaining:

Small Groups (n<6)



Terms present in at least 2 patients

Larger Gene Groups (n≥6)

Terms present in >20% of patients



3. Approximately equalise HPO terms

Equalise groups with $n > Q3 + (IQR * 1.5)$ terms, prioritising terms by IC, and remove groups with $n < Q1$ terms.



28 gene groups (443 patients)



4. Compute pairwise semantic similarity

Performed using GraphIC & funSimAvg methods to compute term-term and group-group similarity.



5. Hierarchical clustering

Using agglomerative clustering. A dendrogram was used to select an appropriate number of clusters (n=5).



Patient-Level Analysis

462 patients with P/LP variants
across 38 TrxG & PcG genes



1. Remove patients with few terms (n<4)



399 patients



2. Approximately equalise HPO terms

Equalise groups with $n > Q3 + (IQR * 1.5)$ terms, prioritising terms by IC.



3. Compute pairwise semantic similarity

Performed using GraphIC & funSimAvg methods to compute term-term and patient-patient similarity.



4. Hierarchical clustering

Using agglomerative clustering. A dendrogram was used to select an appropriate number of clusters for further analysis (n=12).



5. Identify terms enriched in each cluster

Using Fisher's exact test (inside cluster vs outside) with BH correction.





