Reverse genomics predicts function of human conserved non-coding elements

Supporting Material – Marcovitz et al.

Supplementary tables

Supplementary table S1
A summary of genome assemblies for 19 extensively phenotyped placental mammals, and Platypus.

Supplementary table S2
Species in the Morphobank pheno-matrix (O'leary et al.) currently without whole-genome sequences. The list is sorted on the number of phenotypic traits score in Morphobank.

Supplementary table S3
List of 496 unique traits from Morphobank (O'leary et al.), with trait loss or modification in at least two independent lineages. Trait identifier and description is given in columns 1 and 2. The state of each trait and its modified state (e.g., "presence", "absence") are specified in columns 3 and 4, respectively. Columns 5-7 provides comma delimited lists for species (by UCSC genome assembly abbreviations), with conserved, unknown or modified/lost phenotypic state, respectively.

Supplementary table S4
List of 266,115 CNEs conserved in human with genomic coordinates (chr, start, end) in GRCh37/hg19. A column of 1/0 indicates whether each CNE is conserved across at least 7 species, and is next MGI/HPO annotated genes.

Supplementary table S5
List of 2,077 independently lost CNEs (IL-CNEs), within 100 kb or less from genes annotated by either HPO or MGI. Genomic coordinates for the CNEs (GRCh37/hg19) are specified in columns 2-4, and the neighboring gene and distance (bp) are specified in columns 5 and 6, respectively. A comma delimited list of species (by UCSC genome assembly abbreviations) conserving the CNE is given in column 7, and a similar list is provided for species with CNE loss and with unknown CNE state (columns 9 and 8, respectively).

Supplementary table S6
A table of 2,759 unique associations between Morphobank traits and IL-CNEs. GRCh37/hg19 Genomic coordinates of the IL-CNEs are given in a BED format (columns 3-5). Implicated gene and CNE distance to the transcription start site are specified in columns 6 and 7, respectively.

Supplementary table S7

A table of 183 unique associations between Morphobank traits and IL-CNEs that yield contextual closed loops with HPO and MGI ontologies (at least 4 matching ontology terms). GRCh37/hg19 Genomic coordinates of the IL-CNEs are given in BED format (columns 3-5). Implicated gene and CNE distance to the transcription start site are specified in columns 6 and 7, respectively. Identifiers of the ontology terms in closed loops are given in column 8.

Supplementary figures

Supplementary figure S1

Manual assessment of mapping accuracy between Morphobank textual trait descriptions and gene annotation terms from HPO and MGI. Our method automatically detects shared informative keywords between phenotypes and ontology terms and thus, might be challenged by factors such as keywords complexity and length or keywords ambiguity (e.g., "*neck*" may correctly map head and neck phenotypes to terms like "*Abnormality of head and neck*" or incorrectly to "*Abnormality of the femoral neck*"). (A) DAG (directed acyclic graphs) subsets are shown for HPO (left) and MGI (right) ontologies with the highest level terms for "teeth" and "brain" highlighted in green and red, respectively. (B) 11 unique trait-CNE "closed loop" associations with teeth related phenotypes from O'leary et al, showing 11 implicated CNEs near 9 genes. We manually confirmed that 10 of the 11 CNEs are near teeth related genes (8/9) annotated by at least one of the three highest level teeth ontological terms from either HPO or MGI. (C) A summary table with accuracy computed manually for various trait categories. Accuracy is defined as the fraction of CNEs (and genes) that are in correct "closed loop." For example, for teeth "closed loop" associations, CNEs mapping accuracy is 10/11 (90.9%), and gene accuracy is 8/9 (88.9%).