Psychiatric Gene-Epigene Annotation

General Workflow

This project provides a comprehensive bioinformatics pipeline for annotating gene lists derived from epigenetic studies, specifically targeting psychiatric research. The final output is a multi-layered dataset (`annotated_genes.xlsx`), with the `augmented` sheet serving as the combined genetic-epigenetic annotation

1. Core annotation

- Gene metadata: resolving gene identifiers (Symbols, LOC IDs) to standardized HGNC, Entrez, and Ensembl IDs via NCBI Datasets.
- Functional synthesis: aggregating biological summaries and "FUNCTION" comments from NCBI, UniProt, and HGNC into a unified functional description.

2. Evidence Integration

Psychiatric associations: identifying gene-disease links across four major evidence streams
1. GWAS Catalog: significant SNP-trait associations filtered for psychiatric phenotypes
2. Harmonizome: Cross-database (CTD, GAD, etc.) disease associations enriched for mental health keywords
3. PubMed Literature: Automated mining of high-confidence gene-literature links via MeSH terms and title-based genetic tagging
4. DisGeNET: Integration of curated psychiatric disease-gene associations, providing both association scores and supporting evidence

3. Epigenetic Context & Augmentation

- CpG-to-Gene mapping: explicitly links CpG probe IDs and chromosomal coordinates back to the annotated gene list
- EWAS Atlas integration: pulling trait associations directly associated with the CpG probes to provide a primary epigenetic signal.
- Broad association mapping: expanding the dataset beyond psychiatric traits to include a broad-spectrum view of all known gene-disease associations (via DisGeNET), allowing for a comprehensive analysis of pleiotropy and non-psychiatric context.

Detailed workflows

Gene Annotation

▶ HGNC & NCBI Metadata

NCBI Datasets (REST API) & **HGNC** (REST API):
unique `ewas_res_groupsig_128.xlsx` gene list → base annotation

1. Input resolution
    - Accepts gene symbols or LOC IDs

- LOC IDs (e.g., `LOC101926933`) resolved via NCBI Gene ID
- Symbols resolved via NCBI symbol search (restricted to human)

2. Metadata extraction
   - NCBI: Retrieves HGNC `symbol`, `name`, `hgncID`, `entrezID`, `ensemblID`, `gene_type`, `synonyms`, UniProt accession as `uniprot`
   - HGNC Fallback: If NCBI lacks an Ensembl ID, queries HGNC using the HGNC ID to fill the gap
   - creates `annotated_genes.xlsx` with above columns

▶ Function Aggregation

UniProt (REST API); **HGNC** (REST API):
Gene symbol → enriched annotation

1. Fetching
   - UniProt: Queries API using the `uniprot` accession from NCBI. Extracts "FUNCTION" comments, entry status (Reviewed/Unreviewed), and protein existence evidence (n/a, 1–5)
   - HGNC: Queries API; attempts to find "curator_summary", "gene_group", or "name" to serve as a functional description
2. Aggregation & Deduping
   - Combines text from NCBI (Summary), UniProt, and HGNC
   - Normalizes text (lowercase, whitespace) to detect duplicates automatically
   - Format: Returns a multi-line string labeling the source and collapsing duplicates in `function` column
   - Example:

```
[SOURCE: NCBI] Protein coding gene...
[SOURCE: UniProt] same as NCBI
[SOURCE: HGNC] n/a
```

3. Columns Added: `uniprot_review`, `protein_existence`, `function`

▶ GWAS Catalog (Psychiatric)

GWAS Catalog (Data Downloads):
`gwas-catalog-associations_ontology-annotated.tsv` → `misc/gene_psych_gwas.tsv`

1. Pre-processing
   - Filters EBI GWAS Catalog associations for psychiatric keywords (see *Appendix A*)
   - Explodes multi-gene entries (comma-separated MAPPED_GENE)
   - Aggregates traits, studies, and PMIDs per gene
2. Attachment
   - Joins the pre-computed psychiatric GWAS table to the main gene list by gene symbol
   - Logic: Left join on standardized symbol
3. Columns Added: `gwas_assoc_count`, `gwas_traits`, `gwas_labels`, `gwas_pmids`, `gwas_efo_uris`, `gwas_study_accessions`

▶ Harmonizome (Psychiatric)

Harmonizome (REST API):
Gene Symbol → Psychiatric Associations (`annotated_genes.xlsx`)

1. Query all associations for a gene symbol via `download/associations` REST endpoint
2. Filtering
    - Datasets: Restricts to key disease datasets (*CTD, DisGeNET, GAD, GWASdb, DISEASES*)
    - Keywords: Filters associations for psychiatric terms (see *Appendix A*)
3. Columns added: `harmonizome_count` , `harmonizome_terms` , `harmonizome_datasets`

▶ PubMed Literature

NCBI E-utilities (API):
Entrez ID / Symbol → Mental Health Literature Hits (`annotated_genes.xlsx`)

1. Identification
    - Primary: `elink` (Gene ID → PubMed ID) for high-confidence links
    - Secondary: `esearch` (Symbol in Title/Abstract/MeSH) if Entrez ID is missing
2. Filtering
    - Retrieves full XML for identified PMIDs stored in `pubmed_pmid`
    - Mental Health Filter: Keeps papers matching specific MeSH terms or Title keywords (see *Appendix B*), lists term hits in `pubmed_terms`
    - Genetic Tagging: Flags papers as "Genetic" if they contain terms like "Polymorphism", "GWAS", "Variant"
3. Formatting
    - Generates a brief summary string for quick review in `pubmed_brief` column
    - Example:

```
Schizophrenia genetics 2024 PMID: 12345 Genetic
Depression study 2023 PMID: 67890 Not Genetic
```

4. Columns added: `pubmed_count` , `pubmed_genetic_count` , `pubmed_pmid` , `pubmed_terms` , `pubmed_brief`

▶ DisGeNET

`disgenet2r` :
gene symbol → Gene-Disease Associations (`data/disgenet_gda.csv`)
gene symbol → Gene-Evidence Associations (`data/disgenet_gea.csv`)

1. Filter `disgenet_gea.csv` by column `diseaseClasses_UMLS_ST` == "Mental or Behavioral Dysfunction (T048)"
2. For every gene in `annotated_genes.xlsx` :
    1. `disgenet_psych_diseases` column:
        - Group `disease_name` values in `disgenet_gea.csv`
        - Logic:
            - `score` consistency across entries; else "error"

        - Sort unique disease names by score
    - Format: `[Disease Name], [Score]`
        - Separator: `;` + newline
        - *Example*:

```
Schizophrenia, 0.7;
Bipolar Disorder, 0.5
```

2. `disgenet_evidence` column:
    - Separate `disease_name` (rows in `disgenet_gea.csv`)
    - Logic:
        - Polarity: "Positive", "Negative", or "NAPolarity" (for NA)
        - Reference: if `reference_type` == "PMID" → use `reference` ID; else `reference("NA")(source, associationType)`
        - Sort by (1) descending disease score, then (2) polarity (+ > na > -) and chronological publication
    - Format: `[Disease], [Polarity], [Year], [Ref]`
        - *Example*:

```
Schizophrenia, Positive, 2006, 2489764;
Schizophrenia, NAPolarity, 2004, NA(CLINVAR,
GeneticVariation);
Schizophrenia, Negative, 2005, 99348737;
Bipolar Disorder, Positive, 2015, 9398387
```

> [Note:] *The `disgenet_diseases` column represents a broad extraction of all associated traits, bypassing the mental health filter applied to the `disgenet_psych_diseases` column.*

Epigene Annotation

▶ UCSC Screen

UCSC Genome Browser (REST API):
CpG coordinates (`ewas_res_groupsig_128.xlsx`) → Track annotations (`data/ewas_ucsc_annotated.xlsx`)

1. Confirm coordinates resolve to the correct CpG ID via `snpArrayIllumina850k` track
2. **Track Queries**:
    1. Direct Overlap: query `clinvarMain` at exact CpG coordinates to identify clinical significance or mutations (unlikely)
    2. Neighborhood: query `gwasCatalog` within a 5kb window (±5000bp) to identify nearby disease-associated SNPs
3. Generate flattened hits written to Excel file with separate sheets for each track (`clinvarMain` , `gwasCatalog`)

- Headers determined by resulted json fields

▶ EWAS Atlas

NGDC EWAS Atlas (REST API):
CpG probe ID → trait associations (`data/ewas_atlas.csv`)

1. Query API per CpG `probeID` in `ewas_res_groupsig_128.xlsx` :
   - Resolve symbols from relatedTranscription as `genes`
   - Capture cpgIsland status as `cpg_island`
   - Flatten `associationList` , containing `trait` , `correlation` , `rank` , `pmid`
2. Methodological refining
   - Rank scoring: calculate `rank_score = rank / total_associations` (3 decimal places). If rank is missing, empty value for output readability
   - Correlation (methylation) mapping:
     - `pos` ↔ `hyper`
     - `neg` ↔ `hypo`
     - `NA` ↔ `NR`
   - Unmapped gene gap: cross-reference all genes in `ewas_atlas.csv` against validated `symbol` and `synonyms` columns from `annotated_genes.xlsx` *and* the nearest gene (`unique_gene_name`) from `ewas_res_groupsig_128.xlsx`
3. Data integration logic
   - Formatting: traits are aggregated into a single cell, separated by `; \n` , in the format: `[trait], [rank_score], [correlation], [pmid]` , sorted by descending rank score, then alphabetically by trait
   - Columns: results partitioned into `ewas_atlas_traits` , `ewas_unmapped_gene` (established symbols), and `ewas_unmapped_regions` (decimal-named clone/LncRNA loci; "AP003039.3" vs. "NTM"). Established symbols table can be found in Table 1, while uncharacterized regions are in *Appendix C* Table C.1
   - Example `ewas_atlas_traits` output

```
preterm birth, 0.9, hyper, 1239847;
allergies, 0.8, hypo, 10923874;
Alzheimer, 0.8, NR, 019834
```

▶ Augmentation

EWAS Atlas & **DisGeNET**:
`annotated_genes.xlsx` → `annotated_genes.xlsx[ sheets = 'annotated_genes', # original 'ewas_atlas', # copy of data/ewas_atlas.csv 'augmented', # NEW: combined sheet 'ewas_res_groupsig_128' ]`

1. Comprehensive join
   - Joins the provided `ewas_res_groupsig_128.xlsx` chromosomal coordinates and probe IDs to new augmented sheet
   - Logic: right join provided CpG list to ensure every probe is retained regardless of gene annotation

    - Primary join on the `input` column to handle gene name discrepancies between raw data (nearest genes) and validated symbols (via Entrez → HGNC queries in above section)
2. EWAS Atlas enrichment
    - Cross-references CpG probe IDs with the EWAS Atlas dataset to identify additional trait associations and unmapped genes/loci
    - Logic: empty associations are left as nulls to maintain clarity in the sheet, associations without a rank, and subsequently a rank score, are given a placeholder score of `0.000`
3. Broad-spectrum DisGeNET: pulls all gene-disease associations without psychiatric filtering to provide broad phenotypic context in the `disgenet_disease` column
4. Columns added: `cpg`, `cpg_chr`, `cpg_start`, `cpg_end`, `ewas_atlas_traits`, `ewas_unmapped_gene`, `ewas_unmapped_regions`, `disgenet_disease`

Findings

DISTAL GENE-CPG UNMAPPED ASSOCIATIONS

A critical finding during the augmentation phase was the identification of CpG-gene associations via EWAS Atlas that were previously unmapped to genes from proximity identification. Established CpG-gene associations (e.g., GSE1, CNTD2) were recorded and displayed in Table 1; uncharacterized long non-coding RNA (lncRNA) loci or genomic regions named using the Human Genome Project clone system (e.g., `AP003039.3`) can be found in *Appendix C*.

A primary example is *cg02255242*, associated with the gene *FMN1*. The probe *cg02255242* is located at `chr15:33128710-33128712`, approximately 70.9kb from the FMN1 transcription start site (`33057746`). Despite this distance, EWAS Atlas identifies a validated association from a single study characterized by 100% hypermethylation in the gene body. According to EWAS Atlas, FMN1 is broadly implicated in the epigenetic landscape, associated with 42 unique probes and 36 traits; notably, these include high-priority conditions such as Alzheimer's disease (ranked 3rd; 5 associations) and Mild Cognitive Impairment (tied for 1st; 6 associations).

Table 1: Unmapped Gene Against Known Synonyms and Originating Input

CpG	EWAS Atlas Genes	Unmapped Genes	Synonym	Nearest Gene, Original Input
cg15035382	CNTD2	CNTD2		CNTD2
cg26456563	CNTD2	CNTD2		CNTD2
cg10957166	AC084018.1;RHOF;RP11-347I19.7;RP11-347I19.8	RHOF		LOC338799
cg24315757	MAGI1;MAGI1-AS1;MAGI1-IT1	MAGI1-IT1		MAGI1-AS1, MAGI1
cg10638439	HCP5;MICA;Y_RNA	Y_RNA		MICA
cg10260205	PCBP3;PRED62	PRED62		PCBP3
cg25857471	DPCR1;SFTA2	SFTA2		DPCR1

CpG	EWAS Atlas Genes	Unmapped Genes	Synonym	Nearest Gene, Original Input
cg02255242	FMN1	FMN1		
cg16185115	GSE1	GSE1		
cg04732357	ANKRD36C	ANKRD36C		
cg21088344	SLFN12L	SLFN12L		
cg16910670	NADSYN1	NADSYN1		
cg13796823	SAMD3;TMEM200A	SAMD3		
cg13796823	SAMD3;TMEM200A	TMEM200A		

Genes (*Unmapped Genes*) extracted from EWAS Atlas by provided CpG ProbeID via REST API were cross-referenced against known gene aliases from previous annotation (*Synonym*) and initial suspected genes by proximity(*Nearest Gene*, *Original Input*). Underlined genes (*SFTA2*, *FMN1*, *GSE1*, *ANKRD36C*, *SLFN12L*, *NADSYN1*, *SAMD3*, *TMEM200A*) are not listed in error (*CNTD2*, *RHOF*), possible noise (*MAGI1-IT1*, *Y_RNA*), or deprecated (*PRED62*), and thus serve as a potentially significant expansions for exploration. Genes with cross-referenced synonyms accounted for in `annotated_genes.xlsx` (*WBSCR17*, *DPCR1*, *PRAMEF23*, *B3GNTL1*, *RARS*, *FAM134B*) were removed, while the *Synonym* variable remained to guide interpretation

MISCELLANEOUS FINDINGS

Although selection was based solely on genomic proximity while exploring the UCSC screening data (`data/ewas_ucsc_annotated.xlsx`), the two closest traits of 765 interestingly ranged from a classical biological phenotype to an educational attainment proxy, namely the highest mathematics course completed (see below).

Table 2: Proximal Traits Identified via UCSC Screening

CpG ID	Distance (bp)	PubMed ID	Trait	Region	Genes
cg06941159	6	30038396	Highest math class taken (MTAG)	16p11.2	Intergenic
cg20910361	8	36224396	Height	4q31.21	

Top associations identified by genomic proximity within 10bp to the query CpG sites using UCSC screening data (`data/ewas_ucsc_annotated.xlsx`). Traits range from behavioral proxies (*Highest math class taken*) to physical phenotypes (*Height*). *Distance* represents the offset in base pairs from the CpG site. *Genes* are noted as *Intergenic* when the CpG falls outside defined gene bodies within the specified *Region*.

Because multiple genes are now linked to the same CpG, filtering by CpG in `annotated_genes.xlsx, sheet = 'augmented'` can yield interesting, albeit expected results. Genes *UGTA10* and *UGT1A8* are both close enough to *cg00922271* to be listed as a unique gene, and their disease association profiles are expectedly similar from DISGENET data (see below).

Table 3: DisGeNET Disease Profiles for Co-Located Genes

CpG	Symbol	DisGeNET Diseases (Score)

CpG	Symbol	DisGeNET Diseases (Score)
cg00922271	UGT1A10	Gilbert Disease, 0.75; Increased bilirubin level (finding), 0.65; Crigler Najjar syndrome, type 1, 0.6; Crigler-Najjar syndrome, 0.55; BILIRUBIN, SERUM LEVEL OF, QUANTITATIVE TRAIT LOCUS 1, 0.4; Crigler Najjar syndrome, type 2, 0.4; GILBERT SYNDROME, SUSCEPTIBILITY TO, 0.4; Lucey-Driscoll syndrome (disorder), 0.4
cg00922271	UGT1A8	Diarrhea, 0.5; Gilbert Disease, 0.45; Crigler Najjar syndrome, type 2, 0.4; BILIRUBIN, SERUM LEVEL OF, QUANTITATIVE TRAIT LOCUS 1, 0.4; Crigler-Najjar syndrome, 0.4; Lucey-Driscoll syndrome (disorder), 0.4; Crigler Najjar syndrome, type 1, 0.4; Increased bilirubin level (finding), 0.4; Diarrheal disorder, 0.4; GILBERT SYNDROME, SUSCEPTIBILITY TO, 0.4

Comparison of selected disease association profiles for *UGT1A10* and *UGT1A8*, two genes linked to the same CpG probe (*cg00922271*) in the augmented dataset (`annotated_genes.xlsx`). Disease associations were retrieved from DisGeNET and ordered by association score. The overlap in phenotypes (e.g., *Gilbert Disease*, *Crigler-Najjar syndrome*) highlights the functional similarity of these clustered genes.

Discussion

This research involves the synthesis of multi-omic data from disparate public repositories. While every effort was made to ensure accuracy and completeness, several methodological considerations and limitations must be acknowledged to guide interpretations.

The bioinformatics landscape is characterized by high volatility in dataset availability and API accessibility. During the course of this project, several key resources (e.g., specific web-endpoints for the NHGRI-EBI GWAS Catalog) became unavailable or shifted to API-only access. This volatility extends to gene identifiers themselves; for instance, the unmapped gene *PRED62* (associated with `cg10260205`) could not be indexed via HGNC or Entrez. Manual tracing through the ExPheWas Browser identified it as Ensembl ID *ENSG00000268040*, which was revealed to be a deprecated identifier archived in February 2014. Additionally, automated extraction logic occasionally yielded "ghost" associations: `cg10957166` returned a multitude of genes (e.g., *AC084018.1*; *RHOF*; *RP11-347I19.7*; *RP11-347I19.8*) via the EWAS Atlas logic, yet manual validation via the browser confirmed zero associated genes. Consequently, the results presented here represent a specific point-in-time snapshot, and future reproduction efforts may encounter similar discrepancies.

Beyond temporal volatility, the associations identified via the EWAS Atlas, GWAS Catalog, and PubMed mining are primarily descriptive. While a high correlation or frequent literature co-occurrence suggests a potential biological link, these metrics do not imply direct mechanistic causation. Epigenetic signals (CpG methylation) and

genomic variants (SNPs) often act in complex regulatory networks where proximity to a gene's transcription start site does not always equate to functional regulation.

This distinction is underscored by the discovery of distal associations, like *FMN1*, which demonstrates that relying solely on "nearest gene" mapping can overlook biologically relevant signals. For example, `cg02255242` is located at `chr15:33128710-33128712`, approximately 70.9kb from the *FMN1* start site (`33057746`). Despite this significant distance, EWAS Atlas contains association of our probe to *FMN1* from 1 study, which showed 100% hypermethylation in the gene body. More importantly, *FMN1* is associated with 42 probes and 36 traits via EWAS Atlas, including high-priority conditions like Alzheimer's disease (ranked 3rd; 5 associations) and Mild Cognitive Impairment (tied for 1st; 6 associations). This finding effectively transforms our approach from proximity scanning to capturing genes with identifiable, validated EWAS-associated links.

Similar interpretative caution is needed for DisGeNET association scores and the `rank_score` metric utilized in the EWAS Atlas integration. The `rank_score` is restricted to associations where an explicit rank was reported in the source study, which excludes contextual information for traits with large total association counts but unreported ranks. A pertinent example is the aforementioned `cg02255242`: while it has a hypermethylated association with "infertility" in PMID 25753583, it lacks a rank among the study's 2,751 associations, leaving it without a score in our augmented dataset.

These evidentiary constraints are further influenced by statistical thresholds. To facilitate a broad-spectrum view, associations were included even when p-values exceeded traditional significance thresholds; in the current augmented set, only 11 hits were below 1e-04, and only one was below 1e-01. While this decision broadens the capture to include potentially relevant traits (e.g., Alzheimer's, diabetes), it introduces the risk that traits of lower clinical relevance to this study (e.g., bariatric surgery, Gulf War illness) could be misinterpreted without strict statistical context.

Search parameters were also shaped by dataset accessibility and bias. Early gene annotations utilized a specific set of psychiatric keywords (see *Appendix A*). Manual exploration revealed that broader terms (e.g., "Diseases of mental health") might yield additional hits in repositories like the Harmonizome. Furthermore, the decision to rely on HGNC as a fallback mechanism for gene resolution introduced specific extraction errors. For instance, *CNTD2* was incorrectly resolved to *ARAP* (aliases: *CENTD2*, *cnt-d2*) via Entrez, whereas HGNC correctly identifies it as *CCNP* (aliases: *CNTD2*). While HGNC is authoritative for established genes, it fails on loci like *LOC338799* (requiring prefix removal processing) and the aforementioned, deprecated *PRED662*. The decision to prioritize capture via Entrez without a secondary verification step compromised specificity.

Finally, a gap analysis reveals that a non-zero number of genes exhibit significant psychiatric associations in DisGeNET but remain absent from the GWAS Catalog or Harmonizome results. This indicates a potential refinement in earlier extraction logic and that, while the pipeline is effective at capturing well-established links, the "dark matter" of gene-disease associations requires further investigation. Conversely, there exists expected gene-disease associations that are not indexed in DisGeNET. While the curated database serves as a high-fidelity filter, distinguishing established links from low-confidence predictions, this exclusivity introduces a bias against less-characterized or complex genomic entities. Specifically, a subset of genes (`misc/disgenet_missing_associations.csv`) returned no data during the enrichment phase. This void arises from two distinct scenarios: genes that are indexed but lack associations within the specific curated subset, and those that are entirely absent from the DisGeNET index (e.g., *KLRC4-KLRK1*). Crucially, the current extraction logic does not differentiate between these two states, effectively flattening a database gap and a biological null result into a single error class. This data requires further exploration but suggests the pipeline's reliance on the

curated DisGeNET subset may be overly conservative, potentially discarding novel candidates that simply lack the historical literature density required for curation.

Immediate Data Validation Required

- Rectify Entrez-HGNC resolution: the reliance on Entrez for gene annotation and resolution (e.g., *CNTD2* resolving to *ARAP*) requires a new verification layer. Future logic must confirm Entrez results against HGNC rather than accepting the first Entrez result as absolute.
- Investigate ghost associations: the anomaly with `cg10957166` (yielding *AC084018.1*, *RHOF*, etc., despite no browser evidence), while seemingly isolated, must be investigated to determine if this is an API parsing error or a database inconsistency

Future Strategic Improvements

- Reprocess and annotate newly-identified CpG-gene associations to increase granularity and accuracy of data, and potentially illuminate clinically-relevant psychiatric associations.
- Adjusting the 5kb window in the UCSC neighboring GWAS screening to better capture distal regulatory elements.
- Direct processing of raw summary statistics from curated Psychiatric Genomics Consortium (PGC) studies to enhance annotation accuracy beyond literature-level metadata. Downloadable study data are available, categorized by psychiatric condition, but require more complex data processing than I felt was relevant.

Appendix

A. Psychiatric Keywords (GWAS & Harmonizome)

Specific keywords used to filter trait and disease labels across the GWAS Catalog and Harmonizome datasets; matches are case-sensitive and inclusive of substrings

▶ Keywords by Category

```
# core psych disorders
"schizophrenia",
"psychosis",
"psychotic",
"bipolar",
"mania",
"mood disorder",
"depression",
"depressive",
"major depressive",
"unipolar",
"dysthymia",

# neurodevelopmental / autism / adhd
"autism",
"asperger",
"pervasive developmental",
"adhd",
"attention deficit",
"hyperactivity",
```

```
    "conduct disorder",
    "oppositional defiant",
    "disruptive behavior",

    # anxiety / ocd / ptsd / stress
    "anxiety",
    "panic disorder",
    "agoraphobia",
    "social phobia",
    "obsessive-compulsive",
    "obsessive compulsive",
    "ocd",
    "post-traumatic stress",
    "posttraumatic stress",
    "ptsd",
    "stress-related",
    "stress related",

    # eating / substance
    "eating disorder",
    "anorexia nervosa",
    "bulimia",
    "binge eating",
    "substance use",
    "substance-use",
    "substance abuse",
    "drug dependence",
    "alcohol dependence",
    "alcohol-use disorder",
    "nicotine dependence",

    # other psych/neuropsychiatric baskets
    "personality disorder",
    "somatoform",
    "tic disorder",
    "tourette",
    "neuropsychiatric",
    "mental disorder",
    "mental or behavioural",
    "mental or behavioral"
```

B. Mental Health MeSH, Title, & Genetic Terms (PubMed)

Medical Subject Headings (MeSH) and fallback title keywords used to identify relevant literature from gene queries via NCBI E-utilities and classify as genetic

▶ MeSH Terms by Category

```
MESH_TERMS: List[str] = [
# core mood disorders
"Depressive Disorder",
```

```
    "Depressive Disorder, Major",
    "Bipolar Disorder",
    "Cyclothymic Disorder",
    # psychotic disorders
    "Schizophrenia",
    "Psychotic Disorders",
    "Schizoaffective Disorder",
    # anxiety, trauma, stress
    "Anxiety Disorders",
    "Panic Disorder",
    "Phobic Disorders",
    "Obsessive-Compulsive Disorder",
    "Stress Disorders, Post-Traumatic",
    # neurodevelopmental
    "Autism Spectrum Disorder",
    "Attention Deficit Disorder with Hyperactivity",
    "Intellectual Disability",
    # substance use
    "Substance-Related Disorders",
    "Alcohol-Related Disorders",
    "Opioid-Related Disorders",
    "Cocaine-Related Disorders",
    # self-harm/suicide
    "Suicide",
    "Suicidal Ideation",
    # broad
    "Mental Disorders",
    "Mental Health",
    ]
```

▶ Title Keywords

```
    TITLE_TERMS: List[str] = [
    "schizophrenia",
    "bipolar",
    "depressi",     # depression, depressive
    "mania",
    "psychosis",
    "psychotic",
    "autism",
    "asperger",
    "adhd",
    "attention-deficit",
    "anxiety",
    "panic disorder",
    "ptsd",
    "post-traumatic",
    "suicide",
    "suicidal",
    "substance use",
```

```
    "substance-use",
    "addiction",
    "alcohol use",
    "alcohol-use",
    "mental disorder",
    "mental health",
    ]
```

▶ Genetic Keywords

```
GENETIC_MESH_TERMS: List[str] = [
"Polymorphism, Genetic",
"Genetic Variation",
"Genome-Wide Association Study",
"Genotype",
"Genetic Predisposition to Disease",
]
```

C. Uncharacterized Genomic Regions (EWAS Atlas)

Detailed list of CpG sites associated with uncharacterized loci, including long non-coding RNA (lncRNA) and genomic regions named using the Human Genome Project clone system (e.g., RP11, AC nomenclature)

Table C.1: Uncharacterized Genomic Regions (EWAS Atlas)

CpG ID	Associated Loci (Atlas)
cg01267120	AL157871.2, RP11-638I2.6
cg01334824	RP11-100M12.3
cg02282594	RP11-298I3.1
cg02780130	RP11-283I3.2
cg02834909	RP11-316F12.1
cg04070200	AC007879.5
cg05406088	RP11-321F6.1
cg05715076	RP11-543C4.1
cg05860956	RP11-283I3.2
cg06941159	RP11-22P6.2
cg07252486	AP003039.3
cg09535960	RP11-177H13.2

CpG ID	Associated Loci (Atlas)
cg09911534	RP11-622C24.2
cg10075163	RP11-283I3.2
cg10957166	AC084018.1, RP11-347I19.7, RP11-347I19.8
cg11340603	RP1-78O14.1
cg12658972	RP11-95P2.3
cg12790145	RP11-283I3.2
cg13303179	RP11-316F12.1
cg13689053	AP000688.14
cg14545602	AC005498.3
cg15501526	RP11-526P5.2
cg15587955	RP11-1080G15.1
cg16469117	AC007092.1
cg17240725	AC073846.5
cg20188212	RP11-283I3.2
cg23102195	CTD-2269E23.4
cg23184739	AC110781.3
cg24468780	RP11-112J1.2
cg24947255	XXbac-BPG181B23.6
cg25329573	RP11-283I3.2

List of CpG sites associated with uncharacterized loci, including long non-coding RNA (lncRNA) and genomic regions named using the Human Genome Project clone system