## **Entity Categories**

Below are representative **single-level** categories for the domain of plant GWAS. They are deliberately broad enough to capture most relevant concepts without fragmenting into hierarchical sub-types.

1. **Organism**
   * For referencing living organisms at the species level (e.g., *Arabidopsis thaliana*, *Oryza sativa*).
2. **Accession / Line**
   * For specific lines or accessions of a species in GWAS contexts (e.g., “Col-0”, “Ts-1”).
3. **Population / Group**
   * For groups of organisms or references to demography (e.g., “F2 populations”, “Southeast Asian populations”).
4. **Gene**
   * For named genes (e.g., *AtHKT1;1*, *HAC1*).
5. **Allele / Variant**
   * For specific alleles or genetic variants (e.g., “T at Chr4:6392276”, “SNP ID: rsXXXX”, “QTL region”).
6. **Genomic Region**
   * *Definition*: For references to specific segments of DNA or RNA, such as exons, introns, promoter regions, or other defined loci (not the entire gene name, which goes under *Gene*).
   * *Examples*: “exon 3 of AtHKT1;1,” “5’ UTR region of gene X,” “chromosomal segment on Chr4: 10,000–20,000 bp.”
7. **Protein / Polypeptide**
   * *Definition*: For references to proteins, polypeptide chains, or protein complexes.
   * *Examples*: “Rubisco large subunit,” “predicted HAC1 polypeptide,” “an enzyme subunit responsible for arsenate reduction.”
8. **Amino Acid**
   * *Definition*: For individual amino acids or amino acid residues, rather than entire polypeptides.
   * *Examples*: “glutamate,” “Arg99,” “methionine.”
9. **Tissue / Plant Structure**
   * For anatomical references within the plant (e.g., “leaf”, “root”, “shoot”)
10. **Trait / Phenotype**
    * For measured or observed characteristics (e.g., “leaf Na+ concentration”, “shoot arsenic accumulation”, “grain yield”).
11. **Biological Process**
    * For broad processes or functional concepts (e.g., “adaptation”, “accumulation”, “photosynthesis”, “flowering time”).
12. **Chemical Substance**
    * For specific chemicals or ions (e.g., “Na+”, “Inorganic arsenic”, “NADPH”).
13. **Environmental Factor**
    * For environmental variables, conditions, or gradients (e.g., “soil salinity”, “high temperature”, “Bangladesh field conditions”).
14. **Study / Analysis**
    * For references to a research study, analysis, or experiment (e.g., “our GWA study”, “the current analysis”).
15. **Tool / Method**
    * For instruments, assays, or protocols (e.g., “Affymetrix SNP-tiling array”, “arsenate reductase assay”).
16. **Measurement / Statistical Result**
    * For reported values, p-values, or measured statistics (e.g., “2,325 mg Na+ kg-1”, “p-value = 2e-16”, “QTL effect size”).
17. **Concept / Abstract Entity**
    * For non-physical, abstract notions not captured by the other categories (e.g., “balance”, “population structure”, “adaptive cline”).

## **Relation (Predicate) Types**

Below is a streamlined set of **relation types** (predicates) commonly needed to express GWAS-related statements. In practice, you might refine or expand them for more nuance, but the goal is to keep them as **multi-purpose** and **domain-relevant** as possible.

1. **IS\_A**
   1. **Definition**: Captures a definitional or ontological statement—“X is (a type of) Y.”
   2. **When to Use**:
      1. To state that an entity (e.g., a variant) is of a particular subtype, class, or has a specific identity.
      2. E.g. “HAC1Kr-0 **is** a loss-of-function allele,” “Col-0 **is** an accession of A. thaliana.”
   3. **Examples**:
      1. “[HAC1Kr-0] IS\_A [loss-of-function allele].”
      2. “[F2 population] IS\_A [subset of broader population].” (Alternative to using *PART\_OF*, depending on how you want to model it.)
2. **ASSOCIATED\_WITH**
   1. **Definition**: A non-causal statement of connection—entities occur together or are observed in tandem, **without** specifying causation.
   2. **When to Use**:
      1. Any mention that X is linked to, correlated with, or otherwise frequently co-occurs with Y, but does not imply either partial or direct causality.
   3. **Examples**:
      1. “[Allele X] ASSOCIATED\_WITH [Trait Y].”
      2. “[Increased leaf Na+] ASSOCIATED\_WITH [saline conditions].”
3. **CAUSES**
   1. **Definition**: A strong causal relationship—“X directly brings about Y.”
   2. **When to Use**:
      1. If the text explicitly states or very strongly implies direct causation (e.g., “knockout of this gene **causes** seed lethality”).
   3. **Examples**:
      1. “[Allele X] CAUSES [reduced salt uptake].”
      2. “[High arsenic exposure] CAUSES [increased cancer risk].”
4. **CONTRIBUTES\_TO**
   1. **Definition**: A partial or weaker causal relationship—“X plays a role in Y” but does not singlehandedly cause Y.
   2. **When to Use**:
      1. If the text uses language like “leads to,” “influences,” “affects,” or “contributes to.”
   3. **Examples**:
      1. “[QTL region] CONTRIBUTES\_TO [higher arsenate accumulation].”
      2. “[Temperature stress] CONTRIBUTES\_TO [reduced grain yield].”
5. **LOCATED\_IN**
   1. **Definition**: Describes the physical or genomic location of an entity.
   2. **When to Use**:
      1. For geographic locations (e.g., “Population X LOCATED\_IN region Y”) or genomic coordinates (e.g., “QTL peak LOCATED\_IN chromosome 4 region”).
   3. **Examples**:
      1. “[QTL peak] LOCATED\_IN [Chr4: 10,000–20,000 bp].”
      2. “[Southeast Asian lines] LOCATED\_IN [Bangladesh fields].”
6. **PART\_OF**
   1. **Definition**: A hierarchical or compositional relationship.
   2. **When to Use**:
      1. For subsets, membership, or references to being part of a larger whole.
   3. **Examples**:
      1. “[337 accessions] PART\_OF [the 349 total set].”
      2. “[Leaf] PART\_OF [plant shoot].”
7. **HAS\_MEASUREMENT**
   1. **Definition**: Links an entity (often a genotype, accession, or trait) to a measured value or statistic.
   2. **When to Use**:
      1. When a paper says “Accession X shows 2,325 mg/kg sodium,” or “p-value = 2e-16.”
   3. **Examples**:
      1. “[Accession X] HAS\_MEASUREMENT [2,325 mg Na+ kg–1].”
      2. “[Genome-wide study] HAS\_VALUE [p = 1e–8].”
8. **ENCODES**
   1. **Definition:** Captures a direct gene → protein relationship, i.e., “Gene X encodes Protein Y.”  
      **When to Use:**
      1. If the text explicitly states or strongly implies that a gene produces or specifies a particular protein or polypeptide.
   2. **Examples:**
      1. “[Gene X] ENCODES [Protein Y].”
      2. “ACR2 ENCODES an arsenate reductase.”
9. **REGULATES**
   1. **Definition:** Covers regulatory relationships, e.g., “Regulatory element X modulates Gene Y” or “Protein X upregulates Trait Y.”
   2. **When to Use:**
      1. If the text indicates that an entity (gene, protein, promoter region, etc.) modulates, influences, or controls another entity (another gene, a protein’s activity, or a phenotypic trait).
   3. **Examples:**
      1. “[Protein X] REGULATES [Trait Y].”
      2. “[Promoter region Z] REGULATES [Gene Y].”
10. **IDENTIFIED\_IN**
    1. **Definition**: Links discoveries or findings to specific studies, populations, or contexts.
    2. **When to Use**:
       1. If the text says “This QTL was **identified in** a GWA study of 337 lines.”
    3. **Examples**:
       1. “[QTL for leaf Na+] IDENTIFIED\_IN [F2 population].”
       2. “[Significant SNP] DETECTED\_IN [genetic screen].”
11. **USED\_IN**
    1. **Definition**: Describes which tools or methods are used in a study or analysis.
    2. **When to Use**:
       1. If the text says “We used an Affymetrix array for genotyping.”
    3. **Examples**:
       1. “[Affymetrix array] USED\_IN [genotyping 337 A. thaliana accessions].”
       2. “[arsenate reductase assay] APPLIED\_IN [phenotyping experiments].”
12. **NOT\_ASSOCIATED\_WITH**
    1. **Definition**: Captures an explicit *negative* statement that X does **not** correlate with, influence, or otherwise affect Y.
    2. **When to Use**: Use this relation when the text clearly states that an entity X is *not* linked to, has no observed effect on, or does not alter entity Y—i.e., the absence of correlation or influence is *explicitly* mentioned.
    3. **Examples**:
       1. “[Allele X] NOT\_ASSOCIATED\_WITH [Trait Y].”
       2. “[T DNA insertion alleles of ACR2] NOT\_ASSOCIATED\_WITH [arsenic homeostasis in A. thaliana].”