IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

HGNC Gene Symbol Ambiguity: Census and CRO Matching Policy

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-11-24 · hgnc gene symbol ambiguity · gene symbol to hgnc id mapping · hgnc multi-symbol checker · ambiguous gene symbols · gene symbol aliases and previous symbols · excel gene name conversion errors · cro data delivery gene identifiers · ensembl id vs gene symbol · gene symbol normalization

Original article: https://intuitionlabs.ai/articles/hgnc-gene-symbol-joins-ambiguity-cro-policy



In this article

1. Executive Summary
2. Introduction and Background
3. Identifiers, Symbols, Aliases and Withdrawn Records
4. Reproducible Census of Symbol Ambiguity
5. Matching Policy and Worked Examples
6. CRO Delivery Schema and Acceptance Fixtures
7. Release Comparison and Rerun Decision
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

A human gene symbol is a label that the HUGO Gene Nomenclature Committee (HGNC) can change, share or retire, while the HGNC identifier (HGNC ID) is the stable key. HGNC states that its IDs "are stable and do not change if the gene symbol changes" (pmc.ncbi.nlm.nih.gov) and advises quoting the ID "to avoid ambiguity" (hgnc.genenames.org). This report converts that guidance into an acceptance policy for Contract Research Organization (CRO) gene lists and for joins between transcriptomics and target-evidence tables, backed by a reproducible census.

The census parsed the official complete-set file, which HGNC says is refreshed on Tuesdays and Fridays (hgnc.genenames.org) and whose storage object shows a last-modified value of 23 November 2026 (a Monday, so that timestamp can differ from the nominal Tuesday/Friday schedule): **45,470** approved records, **103,661** distinct case-sensitive tokens drawn from the approved symbol, alias and previous-symbol fields, and **2,129** tokens (**2.05%**) that point to more than one HGNC ID (storage.googleapis.com). The count is a property of a deliberately permissive all-field lookup, not an error rate in anyone's data. It was identical in the 6 October 2026 monthly archive yet **276** tokens changed candidate sets between the two releases, so a headline count cannot decide whether to rerun a join.

The census splits cleanly by policy. Searching approved symbols only produces **0** ambiguous tokens. Searching all three fields case-sensitively leaves **2,129**; folding case, which the checker does by default (www.genenames.org), raises that to **2,331**. An approved-symbol-first rule leaves **1,443** unresolved but silently hides a competing alias or previous-symbol claim for **686** tokens, which is **32.2%** of the 2,129. The official checker's own example shows the pattern: ACS3 is "a previous symbol of TWIST1, and an alias of ACSL3" (www.genenames.org).

The recommended contract has three outcomes. **Automatic:** an exact approved symbol with no competing claim (44,784 tokens), or a single-candidate alias or previous symbol logged with its matched field (56,687). **Review:** multiple candidates, an approved symbol that is also another gene's alias, a case-only match, a withdrawn symbol or a date-like string (2,190 vocabulary tokens plus input-level cases). **Reject or route:** gene

names (the checker "can only be used for searching lists of gene symbols, not gene names" (www.genenames.org)) and non-human species, which HGNC asks to be identified by taxonomy ID (pmc.ncbi.nlm.nih.gov).

Each ledger row must retain the original input, species, matched field, all candidate IDs, selected ID, mapping release (file, date, SHA-256) and reviewer decision. Between the 2026-10-06 and 2026-11-23 files, 8 approved symbols changed and one single-candidate token pointed to a different gene, a merge documented in the withdrawn file. HGNC IDs "do not change unless the gene structure undergoes extreme alteration" (pmc.ncbi.nlm.nih.gov). An identifier match says nothing about target validity or clinical relevance.

Introduction and Background

Biotech discovery-informatics teams routinely accept gene lists from a CRO, then join them to expression matrices, target-evidence tables and knowledge graphs. A join on a symbol string is only as reliable as the symbol vocabulary behind it. HGNC describes itself as "the sole authority for approving human gene symbols and corresponding descriptive gene names" (pmc.ncbi.nlm.nih.gov), and it warns that "sometimes an approved symbol can be used to denote a different gene in the literature" (hgnc.genenames.org). Symbols "are constantly updated through the discoveries and re-identification of genes" (f1000research.com). The practical question for a data owner is narrower than nomenclature theory: which incoming labels can be mapped without a human, which need review, and does a new vocabulary release force a rerun?

This report answers with an **ambiguity census** and a **matching-policy table**, not a directory of gene-name converters. It treats three things separately: the stable identifier, the current display symbol, and earlier or alternative labels. It is written for discovery-informatics leads, biomarker analysts, scientific data engineers and AI platform owners. The analysis below stands on public HGNC files and documentation.

Scope and limits

The census covers human genes in the HGNC complete set. It does not measure how many symbols in any pharma dataset are wrong. It does not assess biology: an identifier match shows that two records name the same HGNC entry, not that a gene is a valid target or clinically actionable. No public CRO-specific gene-identifier delivery standard was found among the sources opened for this report, so the acceptance contract in later sections is an analyst synthesis built on HGNC's documented behavior.

The HGNC home page opened on 24 November 2026 says that download files, archives included, "are now in a publicly accessible Google Storage Bucket" (www.genenames.org). The pages opened do not state a migration date, so none is asserted here.

Identifiers, Symbols, Aliases and Withdrawn Records

Four labels that are often conflated

- **HGNC ID:** every gene with an approved symbol "also has a unique HGNC ID in the format HGNC:number" (hgnc.genenames.org). IDs "are associated with the gene sequence and do not change unless the gene structure undergoes extreme alteration" (pmc.ncbi.nlm.nih.gov).

- **Approved symbol:** the current display label. HGNC guidelines say each gene "is assigned a unique symbol, HGNC ID and descriptive name" (pmc.ncbi.nlm.nih.gov), yet symbol changes do happen, and "it is essential for databases to track current approved symbols using HGNC IDs" (pmc.ncbi.nlm.nih.gov).
- **Alias symbol:** an additional label added by an HGNC editor. The same label can be an alias of one gene and the approved symbol of another. HGNC's own example is *GRID1* (HGNC:4575), which "has the symbol alias GluD1 and there is a separate gene with the approved gene symbol" GLUD1 (pmc.ncbi.nlm.nih.gov).
- **Previous symbol:** a former approved symbol. The checker defines it as a symbol where "the input was previously an approved gene symbol, but the gene has since been updated with the approved symbol shown" (www.genenames.org).
- **Withdrawn record:** a retired entry. The checker returns a match type for an input that "was an approved symbol for a gene that has since been withdrawn" (www.genenames.org), and the Withdrawn option is "selected by default and will return hits that have been withdrawn within our database" (www.genenames.org).

HGNC's nomenclature guidance adds that "Wherever possible we avoid reusing symbols from" withdrawn records (pmc.ncbi.nlm.nih.gov), but the qualifier "wherever possible" matters: a pipeline cannot assume that a withdrawn symbol never reappears.

Symbol, name and species are separate inputs

The checker "can only be used for searching lists of gene symbols, not gene names" (www.genenames.org). A CRO column headed "Gene" may contain symbols, full names, Ensembl IDs or free text, so the first acceptance test is classifying the input field. Species is the second test. HGNC recommends "citing the NCBI taxonomy ID" with symbols to distinguish homologous genes (pmc.ncbi.nlm.nih.gov), and mouse nomenclature follows different capitalization rules: the Mouse Genome Informatics guidelines refer to "the rule of uppercase first letter and lowercase remaining letters" for gene symbols (www.informatics.jax.org). Nomenclature authorities treat uniqueness as the main purpose of a symbol: "The primary purpose of a gene or locus name and symbol is to be a unique identifier" (www.informatics.jax.org). A mouse symbol such as Trp53 should therefore never be sent to HGNC as though it were human.

What the files contain

HGNC publishes the complete set and a separate withdrawn file as tab-separated and JSON downloads on Google Cloud Storage (hgnc.genenames.org). Statistics and download files "are updated on each Tuesday and Friday" (hgnc.genenames.org). Monthly archives "are updated on the first week of the month on a Tuesday or Friday" (hgnc.genenames.org), and "the quarterly files are not deleted currently so for snapshots older than 365 days please use the quarterly files" (hgnc.genenames.org). HGNC also invites authors to run lists through the multi-symbol checker to learn "whether the symbols in your list are HGNC approved, non-approved synonyms/aliases, previous symbols or completely unmatched to our database" (blog.genenames.org). It also publishes symbol-change files that "show a list of symbol changes for the span of a month or a quarter of a year" (pmc.ncbi.nlm.nih.gov).

One reconciliation finding matters for parsers. Records can be approved or withdrawn, yet all **45,470** rows of the 2026-11-23 complete set carry the status Approved. Withdrawn records live only in the separate file, which holds **5,293** rows: **1,803** "Entry Withdrawn" and **3,490** "Merged/Split" (storage.googleapis.com, storage.googleapis.com). A join that reads only the complete set will therefore never see a withdrawn ID.

Reproducible Census of Symbol Ambiguity

Method

The census was re-run on 24 November 2026 against the file at the HGNC storage path, last modified 23 November 2026, 17,037,207 bytes, SHA-256 `46b83736ec351f15e22331575dce2f08416c5233765b26a08c65e6b42e56b651` (storage.googleapis.com). The algorithm:

```
import csv, collections
rows = csv.DictReader(open(path, newline="", encoding="utf-8"), delimiter="\t")
cand = collections.defaultdict(set)          # token -> set of HGNC IDs
for r in rows:
    if r["status"] != "Approved": continue
    for field in ("symbol", "alias_symbol", "prev_symbol"):
        for tok in r[field].split("|"):      # pipe-separated multi-values
            if tok.strip(): cand[tok.strip()].add(r["hgnc_id"])
multi = {t for t, ids in cand.items() if len(ids) > 1}
print(len(cand), len(multi))                 # 103661 2129
```

Empty values are excluded and repeats of the same ID are deduplicated by the set. Case-insensitive runs key on `tok.upper()`. The same function was applied to the 2026-10-06 monthly archive and the 2025-10-07 quarterly archive for the release comparisons below.

Parser tests for quoted multi-values

The delimiter inside alias and previous-symbol cells is a pipe and the file is tab-separated, so a comma-oriented CSV reader would be wrong; RFC 4180 notes that "Each record is located on a separate line, delimited by a line" break (www.rfc-editor.org), which is why a split record breaks naive readers. The only quoted cells found in the current file are in the MANE Select column (for example `"ENST00000263100.8|NM_130786.4"`); no alias or previous-symbol cell contained a quote character. Python's `csv` module returned 45,470 rows with both minimal and no quoting. The same RFC says fields containing line breaks, double quotes or commas "should be enclosed in double-quotes" (www.rfc-editor.org), and Python's documentation says a file passed to the reader "should be opened with" `newline=''` (docs.python.org).

The older archive is less forgiving. The 2025-10-07 quarterly archive parsed to 44,539 rows, but two were fragments of 8 fields that lacked an `hgnc_id` (the records for HGNC:39 and HGNC:30 were split across two physical lines). It also carried 54 columns against 53 in 2026 files, because `location_sortable` is absent from the newer releases. A safe loader should therefore reject any row whose `hgnc_id` does not consist of HGNC: followed only by digits and should compare the header to a stored schema before every run.

A second parser trap sits in the data rather than the format. The vocabulary contains the tokens **NaN** (an alias of SCN11A, HGNC:10583) and **NA** (a previous symbol of XK, HGNC:12811) in the same file. pandas documents that "values recognized as missing (see" its `na_values` and `keep_default_na` settings are read as NaN (pandas.pydata.org), and its default list includes the string "NaN" (pandas.pydata.org). Loading gene columns as strings with default NA handling disabled is therefore part of the contract.

Results

Table 1 below reports the census under five lookup policies, each with its denominator.

Lookup policy (case-sensitive unless stated)	Distinct tokens	Tokens with more than one HGNC ID	Share
Approved symbol only	45,470	0	0.00%
Alias plus previous symbol	58,892	1,489	2.53%
All three fields	103,661	**2,129**	**2.05%**
All three fields, case folded to upper case	103,308	**2,331**	2.26%
All three fields, protein-coding genes only	62,475	1,723	2.76%

Source: own calculation from the HGNC complete set, 2026-11-23, using the code above.

Approved symbols are unique, as the guidelines require ("Each gene is assigned a unique symbol, HGNC ID and descriptive name" (pmc.ncbi.nlm.nih.gov)). Ambiguity enters through the other two fields and through collisions between fields. Folding case adds 202 ambiguous tokens (a **9.5%** increase over 2,129), and the distinct token count drops by 353 because labels that differ only by capitalization merge. The protein-coding row restricts both candidates and tokens to genes whose locus group is protein-coding (19,299 records), a common focus of CRO gene lists.

Matching Policy and Worked Examples

Where the 2,129 tokens come from

Table 2 classifies the 2,129 ambiguous tokens by which fields contribute candidates.

Fields contributing candidates	Tokens	Example
Alias only	1,011	ASP (7 candidate genes)
Alias and previous	352	P1 (3 candidates)
Approved and alias	457	ACAT1, CCR9
Previous only	80	ACSM2
Approved and previous	224	ERF, AK3
Approved, alias and previous	5	NAT1

Source: own calculation, same file. 1,840 tokens (86.4%) have exactly two candidates; the maximum is 11.

The 686 tokens in the three approved-plus-other rows (457, 224 and 5) are the dangerous subset. An approved symbol is the string a vendor, a CRO or a database export is most likely to use, and for these tokens another gene also claims the same string as an alias or former symbol. HGNC's own commentary on this pattern is that gene confusion arises "because an approved symbol clashes with an alias" (blog.genenames.org). The tafazzin gene (HGNC:11577) is a documented case: HGNC updated its approved symbol in March 2021 (blog.genenames.org), and in the current file TAZ is a previous symbol of TAFAZZIN and an alias of WWTR1 (HGNC:24042).

ACS3 as the reference example

The official checker help states the case in one line: "For example ACS3 will have two results as it is a previous symbol of TWIST1, and an alias of ACSL3" (www.genenames.org). The current file agrees: ACS3 appears in the previous-symbol field of TWIST1 (HGNC:12428) and the alias field of ACSL3 (HGNC:3570) (storage.googleapis.com). The withdrawn file adds a third lens, because it lists ACS3 as record HGNC:128, "Merged/Split", pointing to TWIST1. The checker's rule that multiple hits "will appear in sequential rows in the results table" (www.genenames.org) means a tool that keeps only the first row would pick whichever gene sorts first, and a downstream join would be wrong for one of the two genes with no warning.

Precedence choices and what each costs

- **Approved-symbol-first:** resolves 686 collision tokens to the approved gene, leaves 1,443 ambiguous, and hides the competing claim. It is acceptable only if the ledger still stores both candidates. For OXR1, it selects HGNC:15822 and drops HCRTR1 (HGNC:4848), for which OXR1 is an alias.
- **All-field, return all candidates:** matches the checker's documented behavior and surfaces all 2,129 tokens, with the cost of a larger review queue.
- **Case-sensitive default:** the checker's "default option is case insensitive but you can change this to perform a case sensitive search" (www.genenames.org). A pipeline must record which was used. The GluD1 example shows why: exact lookup returns GRID1 (HGNC:4575), while a case-folded lookup also returns GLUD1 (HGNC:4335).
- **Exact characters:** the help states that "there must be an exact character match between the input and matched term" (www.genenames.org), so OCT7 maps to POU3F2 as an alias but OCT-7 does not match. Do not strip punctuation to rescue an input silently.

The checker help does not say how duplicated input symbols are reported, so a CRO pipeline should not infer that behavior. The reference implementation instead keys output on a row ID, so two identical labels in different rows produce two ledger rows.

Policy classes over the whole vocabulary

Applying the policy below to every one of the 103,661 tokens gives the distribution in *Table 3*.

Class	Rule	Tokens	Automatic action	Review evidence kept
AUTO_APPROVED	One ID, matched an approved symbol, no competing claim	44,784	Map	Release, ID, symbol
AUTO_FLAGGED_ALIAS	One ID, matched an alias only	41,597	Map, flag	Matched field, release
AUTO_FLAGGED_PREVIOUS	One ID, matched a previous symbol only	15,090	Map, flag	Matched field, old and new symbol
REVIEW_MULTI_CANDIDATE	More than one ID, no approved match	1,443	Hold	All candidate IDs and fields
REVIEW_APPROVED_VS_ALIAS_PREV	Approved match plus another gene's alias or previous	686	Hold	Both claims, source context
REVIEW_WITHDRAWN_COLLISION	One current ID but the withdrawn file points elsewhere	61	Hold	Withdrawn status and merge target

Source: own calculation from the complete set and withdrawn file. Tokens that match the same ID in several fields are counted once, in the highest-precedence class.

The table shows that **2.1%** of the vocabulary (2,190 of 103,661 tokens) needs a human under this strict policy, while **97.9%** can map automatically with the matched field recorded. These are vocabulary shares and not list shares: a real CRO list is dominated by approved symbols, and the review rate depends on the list. A **flagged** automatic class is deliberately separate from the clean class so a release comparison can target alias and previous-symbol mappings first.

CRO Delivery Schema and Acceptance Fixtures

Input and output contract

For every delivered list the CRO supplies one row per measured feature with a stable row ID and the original label. The receiving pipeline returns one ledger row per input row, in the original source order (the checker's default output also follows submission order (www.genenames.org)), plus one candidate row per HGNC ID considered. Required ledger fields:

- **row_id and source_order:** integers from the delivery, never regenerated.
- **input_label and input_field_type:** the verbatim string and whether it is a symbol, a name, an Ensembl or Entrez ID, or unknown.
- **species and taxonomy_id:** 9606 for human; any other value routes to the relevant species authority.
- **matched_field:** approved, alias, previous, withdrawn, case-fold or none.
- **candidate_hgnc_ids:** all IDs, pipe-separated, never truncated.
- **selected_hgnc_id and selected_symbol_at_release:** blank unless automatic or reviewer-approved.
- **mapping_release:** file name, last-modified date and SHA-256 of the complete set and the withdrawn file.
- **policy_version, match_class, reviewer, reviewer_decision, decision_date.**

HGNC advises "quoting the HGNC ID for each gene to avoid ambiguity" (hgnc.genenames.org), and its blog guidance is to "Reference our stable HGNC IDs to remove ambiguity when referring to genes that have been previously published with confusing alias symbols" (blog.genenames.org). The contract therefore makes the ID the delivered join key and treats the symbol as an attribute that can be refreshed.

Acceptance fixtures

Table 4 lists fixtures that every implementation should pass. Outputs come from the author's reference implementation run against the 2026-11-23 files. Only the census snippet in the Method section is shown in this article; the classifier, the release diff and the fixture runner behind Tables 3 to 6 are not published here, so readers can reproduce the headline census counts but not those tables. Treat the figures in Tables 3 to 6 as the author's own calculations that have not been independently reproduced; check any count that matters to your own policy against the current HGNC files before relying on it. All counts in this report are tied to that release and will differ in later files; the input SHA-256 is given in the Method section so readers can confirm they are using the same file.

Fixture	Input (species)	Expected class	Candidates (ID / symbol)	Why
Approved unique	TP53 (human)	AUTO_APPROVED	HGNC:11998 / TP53	Exact approved match, no competing claim
Alias unique	OCT7 (human)	AUTO_FLAGGED_ALIAS	HGNC:9215 / POU3F2	One alias hit; flagged
Previous unique	MARCH1 (human)	AUTO_FLAGGED_PREVIOUS	HGNC:26077 / MARCHF1	One former symbol
Alias versus previous	ACS3 (human)	REVIEW_MULTI_CANDIDATE	HGNC:12428 / TWIST1; HGNC:3570 / ACSL3	The checker's own example
Approved versus alias	OXR1 (human)	REVIEW_APPROVED_VS_ALIAS_PREV	HGNC:15822 / OXR1; HGNC:4848 / HCRTR1	First-hit selection would hide HCRTR1
Withdrawn, no target	A12M1 (human)	REVIEW_WITHDRAWN	none (HGNC:1, Entry Withdrawn)	Record retired without a successor
Withdrawn, merged	A1S9T (human)	AUTO_FLAGGED_PREVIOUS	HGNC:12469 / UBA1	Merge target equals the single candidate
Wrong species	Trp53 (Mus musculus)	NOT_HUMAN	none	Route to mouse authority
Wrong species label	Trp53 (human)	UNMATCHED	none	Differs from TP53 by letters, not only case, so no hit even after case folding; probably a mouse symbol, route by species
Unmatched punctuation	OCT-7 (human)	UNMATCHED	none	Exact characters required

Fixture	Input (species)	Expected class	Candidates (ID / symbol)	Why
Case change	tp53 (human)	REVIEW_CASE_ ONLY	HGNC:11998 / TP53	Match only after case folding
Date-prone symbol	SEPT2 (human)	REVIEW_MULTI_ CANDIDATE	HGNC:7729 / SEPTIN2; HGNC:15848 / SEPTIN6	Previous symbol plus alias; Excel can convert labels of this form to dates (genomebiology.biomedcentral.com)

Source: own calculation with the reference implementation; ACS3 and OCT7 versus OCT-7 match the official help.

Two fixtures need explanation. The tp53 row is deliberately a review item rather than an automatic map: the checker folds case by default, but a ledger that cannot tell exact matches from case-folded ones loses evidence. Where a team prefers automation, it can promote a case-only match to a flagged automatic class only when the folded token has exactly one candidate, which holds for **44,740** of the 45,470 lowercased approved symbols; **723** (1.59%) fold to two or more genes and 7 remain exact tokens.

Spreadsheet damage

Excel conversion is a delivery risk in its own right. A programmatic screen of 35,175 supplementary Excel files, "finding 7467 gene lists attached to 3597 published papers" (genomebiology.biomedcentral.com), estimated that 19.6% of articles with such lists were affected by gene-name errors (genomebiology.biomedcentral.com), and the authors described the cause as Excel "inadvertently converting gene symbols to dates and floating-point numbers" (genomebiology.biomedcentral.com). A later scan of 11,117 publications with gene lists found errors in 30.9% (3,436) of them (pmc.ncbi.nlm.nih.gov) and concluded that "the proportion of papers with Excel gene lists affected by errors remained stable over this period" (pmc.ncbi.nlm.nih.gov). The two rates come from different detection software and are not directly comparable. Microsoft documents an option to store "values with continuous letters and numbers as text, instead of converting to a date" (support.microsoft.com), but notes that data with a space or other characters, such as JAN 1 or JAN-1, "may still be treated as dates" (support.microsoft.com), and that the settings "don't directly affect data imported by using Power Query" (support.microsoft.com).

Repair tools take different policies. HGNChelper "identifies HGNC human or MGI mouse gene symbols which are outdated or may have been mogrified by Excel or other" spreadsheet programs (cran.r-project.org), and its documented output for the damaged label 1-Mar is the two-gene string "MARCHF1 /// MTARC1" (cran.r-project.org), which a pipeline must treat as two candidates. The R package hgnc instead requires that values "must match once and only once in the" source column (cran.r-project.org). The ledger must record which policy produced a mapping.

Review checklist

Before a CRO list enters a join, the reviewer confirms each of the following and records the answer in the ledger:

- **Input type:** every column is symbols, not names, Ensembl IDs or free text.
- **Species:** the taxonomy ID is 9606, or the row is routed to another authority.
- **Spreadsheet damage:** no value is a date, a five-digit serial number or a floating-point number.
- **Case:** case-only matches are listed separately from exact matches.
- **Candidates:** every multi-candidate row shows all IDs and the field that matched each.
- **Withdrawn:** every withdrawn hit shows its status and merge targets.
- **Release:** file names, dates and hashes for both HGNC files are recorded.
- **Decision:** a named reviewer and date accompany each manual selection.

Release Comparison and Rerun Decision

What changes between releases

A symbol vocabulary is not static, and HGNC publishes monthly and quarterly archives so that a past state can be reproduced (hgnc.genenames.org). The ledger's `mapping_release` field exists so that two runs can be compared token by token. *Table 5* summarizes two comparisons computed with the same census function.

Measure	2026-10-06 to 2026-11-23	2025-10-07 to 2026-11-23
Approved records, old to new	45,233 to 45,470	44,537 to 45,470
HGNC IDs added / removed	239 / 2	942 / 9
Removed IDs present in withdrawn file	2 of 2	9 of 9
Approved symbol changed on a retained ID	8	119
Old symbol kept as previous symbol	7 of 8	119 of 119
Tokens whose candidate-ID set changed	276	1,519
Tokens newly ambiguous (multi-ID)	0	18
Tokens no longer ambiguous	0	3
Single-ID tokens now pointing to a different ID	1	5
Old tokens that no longer match anything	3	28

Source: own calculation from the archived complete sets (storage.googleapis.com, storage.googleapis.com), the current complete set and the withdrawn file. The 2025 archive's two fragment rows were excluded.

The two columns tell different stories. The multi-ID count is **2,129** in both the 2026-10-06 and 2026-11-23 files, which would suggest nothing changed. Yet 276 tokens had a different candidate set, and one token (SLC41A3-AS1) moved from HGNC:55559 to HGNC:40244, a merge recorded in the withdrawn file. Over the longer interval of about 13.5 months, **18** tokens became ambiguous and **5** unique tokens silently re-pointed. Examples with reasons:

- **PKD1:** HGNC:9008 only in 2025-10-07; now also an alias on PRKD1 (HGNC:9407). A previously clean approved symbol became an approved-versus-alias collision.
- **IPW:** HGNC:6109 (old approved symbol) to HGNC:37462 (SNHG14), listed as "Merged/Split" in the withdrawn file.
- **RSC1A1:** HGNC:10458 to HGNC:24578 (DDI2), also a documented merge.
- **AATK:** the approved symbol on HGNC:21 became LMTK1 (symbol changed 2026-07-09), and AATK is retained as a previous symbol, so the ID did not change.

Decision rule

Because the ID is stable and symbols are attributes, the rerun rule is mechanical:

1. **Header or schema changed** (as with `location_sortable`): rerun the loader tests before anything else.
2. **A selected ID is absent from the new complete set:** look it up in the withdrawn file; a merge target requires a reviewer decision before the join is rerun.
3. **A stored token's candidate set differs between releases:** re-adjudicate that row only; unchanged rows keep their decisions.
4. **Only the display symbol changed on a retained ID:** refresh the symbol column; no join rerun is needed when the join key is the HGNC ID.

A join keyed on symbol strings cannot detect rule 3 at all, which is the practical argument for storing the ID and the matched token together. HGNC advises tracking symbols through the ID.

Data Analysis and Evidence

Review workload depends on the list, not the vocabulary

Table 6 computes workload for lists with a declared composition, using the all-field policy and the 2026-11-23 file.

Declared input list	List size	Tokens needing review or hold	Share
Every approved symbol, upper case	45,470	686	1.51%
Approved symbols of protein-coding genes only	19,299	533	2.76%
Every distinct alias token	43,490	1,825	4.20%
Every distinct previous-symbol token	15,763	661	4.19%
Every approved symbol, lower-cased	45,470	723	1.59%

Source: own calculation, same file. Review here means more than one candidate across the three fields, or a case fold that lands on more than one gene; unmatched and withdrawn inputs are additional.

Alias and previous-symbol lists carry roughly three times the ambiguity of approved-symbol lists (4.2% against 1.5%). A CRO that delivers approved symbols is cheaper to review than one that delivers literature aliases, which is a reason to put the symbol type in the delivery specification. In the protein-coding subset, 533 of

19,299 approved symbols (2.76%) are also another gene's alias or former symbol. This is a separate count from the 1,723 of 62,475 tokens (2.76%) in Table 1: the two denominators differ (approved symbols versus all-field tokens), and the matching percentages are a coincidence of rounding (2.762% and 2.758% respectively). In all, **3,860** of 45,470 records (8.49%) are a candidate for at least one ambiguous token.

Calculate workload from the real list with a single pass: classify each input row, count rows in review classes, then add unmatched and withdrawn rows. Review workload is then the number of rows in review, unmatched and withdrawn classes divided by the number of rows in the list.

Why identifiers other than HGNC do not remove the problem

Other namespaces are stable for different reasons and have their own caveats. Ensembl says stable identifiers "should continue to refer to the same genomic features" (www.ensembl.org) and recommends authors keep note of them: "We recommend noting down the stable identifiers of your feature of interest in your publications" (www.ensembl.org). Gene IDs carry a version that "increments when the set of transcripts linked to a gene changes" (www.ensembl.org), though versions do not change "when there are changes to the annotation linked to an entity" (www.ensembl.org). Ensembl's gene-name page says the gene ID "is stable from release to release, and does not change" (www.ensembl.org). The GENCODE FAQ shows a sharp edge for joins: pseudoautosomal-region genes on chromosome Y were given IDs made by replacing the first zero of the Ensembl ID with the letter R (www.gencodegenes.org), so ENSGR strings will not match ENSG strings.

NCBI Gene makes the same point from the other direction: "Gene does not enforce uniqueness in preferred symbols" (www.ncbi.nlm.nih.gov), so "please consider use of the GeneID rather than a symbol as the stable identifier of a gene" (www.ncbi.nlm.nih.gov), and a replacement for a discontinued GeneID is found through a history file that "can be used to determine if there is a replacement GeneID" (www.ncbi.nlm.nih.gov).

The HGNC file itself shows why HGNC ID is a better pivot than Ensembl ID for human gene lists. **2,659** of 45,470 records (5.85%) have no Ensembl gene ID, including 43 protein-coding records, 904 have no Entrez ID, and three Ensembl IDs are shared by two HGNC records each (storage.googleapis.com). The Ensembl IDs in the HGNC file carry no version suffix, so versioned IDs from an upstream GTF must be stripped first and the version retained as a separate column.

Reading the evidence together

- HGNC IDs are stable, and symbols are mutable labels; both statements come from HGNC itself (pmc.ncbi.nlm.nih.gov).
- Ambiguity is a property of the lookup policy: 0 tokens under approved-only, 2,129 under all-field case-sensitive, 2,331 under case-folded.
- The headline count was constant across a seven-week interval while 276 candidate sets changed.
- Spreadsheet damage is a recurrent delivery risk, measured at 30.9% of papers with Excel gene lists in one scan using that study's own detection method (pmc.ncbi.nlm.nih.gov).

Implications and Future Directions

For teams building target-evidence graphs, the practical change is to store three fields where one used to be: the HGNC ID as key, the symbol as a time-stamped attribute, and the input label as provenance. This costs little at load time and avoids a class of silent join errors. AI platform owners can apply the same ledger to model outputs: when a retrieval or language-model pipeline emits gene symbols, they should pass through the identical policy before joining to experimental data.

Three open questions remain:

- **Delivery rates:** how often CRO deliveries contain ambiguous tokens has not been measured here, only the vocabulary was. Teams should run the census function on their own historical deliveries.
- **Checker behavior:** duplicate-input handling is not described in the help page and should be tested before it is relied upon.
- **Package defaults:** the hosting change is recent, and the CRAN hgnc package documents that its import function by default uses "the URL corresponding to the latest version" (cran.r-project.org) while it can list "previous versions of these data" for pinning (cran.r-project.org). Any pipeline should set the file explicitly and record its hash.

HGNC's REST service, described as "a convenient and quick way of searching and fetching data from our database" (hgnc.genenames.org), supports spot checks, but bulk adjudication is better done against a pinned file so that results are repeatable. HGNC also changes symbols case by case, since it "considers the naming of each and every gene on a case-by-case basis" (pmc.ncbi.nlm.nih.gov), so rerun checks are a permanent operating cost. A pinned file also guards against parser surprises such as pandas reading "NaN" as missing (pandas.pydata.org), and against assuming that a GeneID is current when "a replacement GeneID" may exist (www.ncbi.nlm.nih.gov). The same logic holds for other species: the mouse nomenclature guidelines describe a symbol's purpose as "a unique identifier" (www.informatics.jax.org), so each species should be matched against its own authority.

Frequently Asked Questions (FAQs)

How do gene symbol joins fail?

Joins fail when the first alias hit is selected, when Excel has converted a symbol to a date (genomebiology.biomedcentral.com), when a resource does not enforce uniqueness ("Gene does not enforce uniqueness in preferred symbols" (www.ncbi.nlm.nih.gov)), when a withdrawn symbol is missed because the complete set holds only approved rows, or when a release changes a token's candidates.

How does gene symbol to HGNC ID mapping work?

Match the token against approved, previous and alias fields, keep all candidate IDs, then select one only under a declared policy. The checker help says its results table includes "the Input search term, the Match Type, the Approved Name, HGNC ID and chromosomal Location" (www.genenames.org); store the ID, not the symbol.

Are aliases and previous symbols safe to map automatically?

Only when a single gene claims the token and the matched field is recorded. In the 2026-11-23 file, 56,687 tokens meet this test and 2,190 do not.

Can the multi-symbol checker take gene names?

No, only symbols are accepted. Symbols "can be separated by commas or new lines" (www.genenames.org), and output can be returned as a "tab delimited text file" (www.genenames.org).

Should Ensembl IDs replace symbols?

They should be carried as secondary attributes with versions retained. Ensembl IDs are stable by design (www.ensembl.org), but 5.85% of HGNC records have none, pseudoautosomal-region genes on chromosome Y use an altered "R" form (www.gencodegenes.org), and resources assign IDs by different rules.

Why does Excel change gene names to dates?

Excel interprets some strings as dates or numbers, an effect documented in published supplementary files (genomebiology.biomedcentral.com). Microsoft documents that disabling the relevant option stores such values "as text, instead of converting to a date" (support.microsoft.com), but the setting does not apply to Power Query imports (support.microsoft.com). Deliver text files and read gene columns as strings.

Do I need to rerun a join after every HGNC release?

Not always. Apply the four-step decision rule: loader schema, absent IDs, changed candidate sets, then display-symbol refresh only.

Conclusion

A reliable gene-symbol join separates three things that a symbol column blurs: the stable HGNC ID, the current approved symbol and the earlier or alternative labels that circulate in vendor files and papers. Ambiguity is a function of lookup policy: no approved symbol is ambiguous on its own, 2,129 of 103,661 tokens become ambiguous when approved, alias and previous fields are searched together, and 2,331 when case is ignored. Approved-symbol-first selection removes 686 of those from view without resolving them, so it must never run without a ledger.

The acceptance contract asks a CRO or internal team to deliver and retain the original input, species, matched field, all candidate IDs, selected ID, mapping release and reviewer decision for every row, with HGNC IDs as the join key. A new vocabulary release then becomes a bounded diff: the 2026-10-06 to 2026-11-23 comparison touched 276 candidate sets and one single-candidate re-point. Recalculate the census on every release adopted, because the count of 2,129 is a property of one file and one method, not a constant.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://pmc.ncbi.nlm.nih.gov/articles/PMC9825485

2. https://hgnc.genenames.org/useful/instructions-to-authors
3. https://hgnc.genenames.org/download
4. https://storage.googleapis.com/public-download-files/hgnc/tsv/tsv/hgnc_complete_set.txt
5. https://www.genenames.org/help/multi-symbol-checker
6. https://pmc.ncbi.nlm.nih.gov/articles/PMC7494048
7. https://f1000research.com/articles/9-1493
8. https://www.genenames.org
9. https://www.informatics.jax.org/mgihome/nomen/gene.shtml
10. https://blog.genenames.org/hgnc/2019/12/20/Multi-symbol-checker
11. https://storage.googleapis.com/public-download-files/hgnc/tsv/tsv/withdrawn.txt
12. https://www.rfc-editor.org/rfc/rfc4180
13. https://docs.python.org/3/library/csv.html
14. https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html
15. https://blog.genenames.org/hgnc/2021/10/21/The-risks-of-using-unapproved-gene-symbols
16. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7
17. https://pmc.ncbi.nlm.nih.gov/articles/PMC8357140
18. https://support.microsoft.com/en-us/excel/set-automatic-data-conversions
19. https://cran.r-project.org/web/packages/HGNChelper/vignettes/index.html
20. https://cran.r-project.org/web/packages/hgnc/refman/hgnc.html
21. https://storage.googleapis.com/public-download-files/hgnc/archive/archive/monthly/tsv/hgnc_complete_set_2026-10-06.txt
22. https://storage.googleapis.com/public-download-files/hgnc/archive/archive/quarterly/tsv/hgnc_complete_set_2025-10-07.txt
23. https://www.ensembl.org/info/genome/stable_ids/index.html
24. https://www.ensembl.org/info/genome/genebuild/gene_names.html
25. https://www.gencodegenes.org/pages/faq.html
26. https://www.ncbi.nlm.nih.gov/books/NBK3840
27. https://hgnc.genenames.org/help/rest

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.