**Archer Fusion Full Results to cBioPortal Structural Variants (SV)**

**Version**: v2

**Script**: archer\_fusions\_to\_cbio\_sv\_v2.py

This README explains how the converter parses Archer Analysis Full Results Summary text files (\*.full\_results.txt) and produces a cBioPortal Structural Variant TSV (data\_sv.tsv). It documents field-to-field mappings, transcript selection rules, fallbacks, and usage instructions.

**1) Overview**

Archer full results contain two layers of fusion information:

* **FCA** (Fusion Candidate Annotation): high-level fusion records, gene pair, evidence flags, and per-transcript annotations (FCA\_{n}\_ANNOTATION\_1-#, FCA\_{n}\_ANNOTATION\_2-#).
* **FC** (Fusion Candidate): per-isoform evidence such as junction/spanning counts and genomic annotation coordinates (FC\_{n}\_EITHER\_R1\_OR\_R2, FC\_{n}\_BOTH\_R1\_AND\_R2, FC\_{n}\_GENOMIC\_ANNOTATION\_#).

The converter:

1. Enumerates FCA indices in a file (each is a fusion candidate).
2. For each fusion, selects the best per-transcript annotation on both 5′ and 3′ sides using the rule **NM\_ > XM\_ > NR\_ > none** (details in section 4 below).
3. Parses exon/intron + number, chromosome, and breakpoint position from the chosen annotations.
4. Resolves read support by following FCA\_…\_SUPPORTING\_ISOFORM\_ID\_\* → FC\_\* counts.
5. If any coordinate is missing, falls back to FC\_\*\_GENOMIC\_ANNOTATION\_\* values.
6. Writes one SV row per supporting isoform (or per FCA if no FC IDs are present; counts default to 0).

Note: Site\*\_Ensembl\_Transcript\_Id is intentionally left blank. The annotations may contain RefSeq IDs (NM/XM/NR), but they are not placed in Ensembl transcript columns. The Site\*\_Ensembl\_Transcript\_Id column name in cBioPortal is Ensembl-oriented (ENST…). Archer provides RefSeq (NM\_…, XM\_…). I left Ensembl blank rather than inserting a non-Ensembl ID. If you’d like, I can either (a) place the RefSeq IDs there (most sites do this and it works fine), or (b) keep them only in Site\*\_Description, or (c) do a RefSeq→Ensembl crosswalk (this will need a mapping file).

**2) Input & Output**

Input: one or more Archer \*.full\_results.txt files (tab-separated key/value lines).

Output: a single tab-delimited data\_sv.tsv with cBioPortal SV columns (see mapping below).

Suggested companion file (place alongside data\_sv.tsv when loading a study, as per cBioPortal documentation):

meta\_structural\_variants.txt

--------------------------------

cancer\_study\_identifier: <your\_study\_id>

genetic\_alteration\_type: STRUCTURAL\_VARIANT

datatype: SV

stable\_id: structural\_variants

show\_profile\_in\_analysis\_tab: true

profile\_name: Structural variants

profile\_description: Structural Variant Data

data\_filename: data\_sv.tsv

**3) Field Mapping (Archer to cBioPortal SV)**

| **cBioPortal SV column** | **Populated from Archer** | **Notes** |
| --- | --- | --- |
| Sample\_Id | SAMPLE\_NAME | Use your sample barcodes if they differ from Archer names. |
| SV\_Status | constant SOMATIC | Adjust if you maintain germline context elsewhere. |
| Site1\_Hugo\_Symbol | first gene in FCA\_{n}\_GENES\_UNIQUE | 5′/left gene. |
| Site2\_Hugo\_Symbol | second gene in FCA\_{n}\_GENES\_UNIQUE | 3′/right gene. |
| Site1\_Region | from chosen FCA\_{n}\_ANNOTATION\_1-# | Exon or Intron. |
| Site1\_Region\_Number | from chosen FCA\_{n}\_ANNOTATION\_1-# | numeric index. |
| Site1\_Chromosome | from chosen FCA\_{n}\_ANNOTATION\_1-#(breakpoint end) | If missing, fallback to FC\_\*\_GENOMIC\_ANNOTATION\_\*. |
| Site1\_Position | from chosen FCA\_{n}\_ANNOTATION\_1-#(breakpoint end) | Prefer the second coordinate if present. |
| Site1\_Description | full chosen annotation string (side 1) | Preserves raw Archer text for traceability. |
| Site2\_Region | from chosen FCA\_{n}\_ANNOTATION\_2-# | Exon or Intron. |
| Site2\_Region\_Number | from chosen FCA\_{n}\_ANNOTATION\_2-# | numeric index. |
| Site2\_Chromosome | from chosen FCA\_{n}\_ANNOTATION\_2-# | Fallback to FC\_\*\_GENOMIC\_ANNOTATION\_\*if needed. |
| Site2\_Position | from chosen FCA\_{n}\_ANNOTATION\_2-# | Prefer breakpoint end. |
| Site2\_Description | full chosen annotation string (side 2) | Preserves raw Archer text. |
| Site1\_Ensembl\_Transcript\_Id | **blank** | Requested behavior (RefSeq IDs not placed here). |
| Site2\_Ensembl\_Transcript\_Id | **blank** | Same as above. |
| Site2\_Effect\_On\_Frame | FCA\_{n}\_HAS\_INFRAME\_TRANSLATION | In-frame / Out-of-frame. |
| NCBI\_Build | command-line option | Default GRCh37; set GRCh38 if hg38. |
| Class | derived | Translocation (different genes) or Intragenic (same gene). |
| Tumor\_Split\_Read\_Count | FC\_\*\_EITHER\_R1\_OR\_R2 | Junction reads (R1 or R2). |
| Tumor\_Paired\_End\_Read\_Count | FC\_\*\_BOTH\_R1\_AND\_R2 | Spanning fragments (paired-end). |
| Annotation | concatenation of both chosen FCA annotation strings | For human-readability in tables. |
| RNA\_Support | Yes if split reads > 0 else No | Archer fusion discovery is RNA-driven. |
| DNA\_Support | default No | Set to Yes only if independent DNA evidence is integrated. |
| Other counts (\*Read\_Count, \*Variant\_Count) | left blank | Not supplied by Archer full results summary. |

**Fallback rule for coordinates:** if a side’s chromosome/position is still missing after FCA parsing, the converter inspects FC\_\*\_GENOMIC\_ANNOTATION\_\* (ordered) and picks a compatible entry (prefers matching chromosome; otherwise first valid GA). The **second coordinate** in GA (chr:start:end:strand) is used as the breakpoint.

**4) “Best transcript” selection (NM\_ > XM\_ > NR\_ > none)**

Archer often reports multiple per-transcript annotations for each side of a fusion:

* 5′ side: FCA\_{n}\_ANNOTATION\_1-1, FCA\_{n}\_ANNOTATION\_1-2, …
* 3′ side: FCA\_{n}\_ANNOTATION\_2-1, FCA\_{n}\_ANNOTATION\_2-2, …

Each value is a compact string, e.g.:

GENE(+)[optional NM\_/XM\_/NR\_] | exon:6 | chr17:41610300,chr17:41620000

Scoring & choice

1. Score each candidate string: NM\_ = 3, XM\_ = 2, NR\_ = 1, none = 0.
2. Pick the highest score; if ties, the earliest key (Archer’s natural order).
3. Parse region/number/chr/pos from the chosen string.

Rationale

* NM\_ (curated coding RefSeq) yields canonical exon numbering and more stable frame context.
* If no NM\_, XM\_ (model coding) or NR\_ (non-coding) still provide usable breakpoints.
* If no transcript token appears, we still parse region/coordinates when present.

Notes

* If a single annotation contains multiple transcript tokens (e.g., NM\_.../XM\_...), it still receives the NM\_ scorebecause NM\_ is present.
* The selection is performed independently for 5′ and 3′ sides.
* We leave Ensembl transcript columns blank but retain the full raw string in Site\*\_Description and Annotation for traceability.

5) Evidence gating & class

* --strong-only: include only FCA\_{n}\_STRONG\_EVIDENCE\_ABERRATION = TRUE.
* Without this flag, all candidates are included. If no FC IDs are present, we still emit a row with counts = 0—useful for downstream review.
* Class: Translocation when partner genes differ; Intragenic otherwise.
* Frame: Site2\_Effect\_On\_Frame from FCA\_{n}\_HAS\_INFRAME\_TRANSLATION (In-frame / Out-of-frame).

**6) Running the script**

**Requirements**

* **Python**: 3.8+ (tested on 3.9–3.12)
* **Dependencies**: standard library only (re, csv, argparse, pathlib, typing) — **no external packages** required.

**Basic usage**

Python3 archer\_fusions\_to\_cbio\_sv\_v2.py \

-o data\_sv.tsv \

IGM\_PBCUKS-0DZY4T\_20240801.full\_results.txt

Example:

python3 archer\_fusions\_to\_cbio\_sv\_v2.py -o test3\_sv\_v2.tsv IGM\_PBBWED-0DIW74\_20230228.full\_results.txt

**Use GRCh38**

Python3 archer\_fusions\_to\_cbio\_sv\_v2.py \

-o data\_sv.tsv \

--ncbi-build GRCh38 \

sample1.full\_results.txt sample2.full\_results.txt

**Strong-evidence only**

Python3 archer\_fusions\_to\_cbio\_sv\_v2.py \

-o data\_sv.tsv \

--strong-only \

sample.full\_results.txt

**Exit output**

Wrote <N> rows to data\_sv.tsv

**Loading into cBioPortal**

1. Place data\_sv.tsv and meta\_structural\_variants.txt in your study folder.
2. Ensure Sample\_Id values match the study’s clinical samples.
3. Import with your usual cBioPortal study loading workflow.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Updated v3

archer\_fusions\_to\_cbio\_sv\_v3.py

**What’s new**

* **Default filters (stricter by default):**
  + Intronic events are excluded unless you pass --keep-intron.
  + Out-of-frame events are excluded unless you pass --keep-out-of-frame.
* **New column:** Percent Coverage  
  The script looks for the most relevant Archer metric available and uses the first present value (in this order):
  + FCA\_{idx}\_PERCENT\_FUSION
  + FCA\_{idx}\_PERCENT\_GSP2\_COVERAGE
  + {FC\_n}\_PERCENT\_FUSION
  + {FC\_n}\_PERCENT\_RNA\_READS  
    If none are present, it leaves the field blank.

**Usage examples**

# Default (filters intronic + out-of-frame)

python archer\_fusions\_to\_cbio\_sv\_v3.py -o data\_sv.tsv IGM\_PBCUKS-0DZY4T\_20240801.full\_results.txt

# Include intronic events too

python archer\_fusions\_to\_cbio\_sv\_v3.py -o data\_sv.tsv --keep-intron IGM\_PBCUKS-0DZY4T\_20240801.full\_results.txt

# Include out-of-frame events too

python archer\_fusions\_to\_cbio\_sv\_v3.py -o data\_sv.tsv --keep-out-of-frame IGM\_PBCUKS-0DZY4T\_20240801.full\_results.txt

# Include both intronic and out-of-frame events

python archer\_fusions\_to\_cbio\_sv\_v3.py -o data\_sv.tsv --keep-intron --keep-out-of-frame IGM\_PBCUKS-0DZY4T\_20240801.full\_results.txt