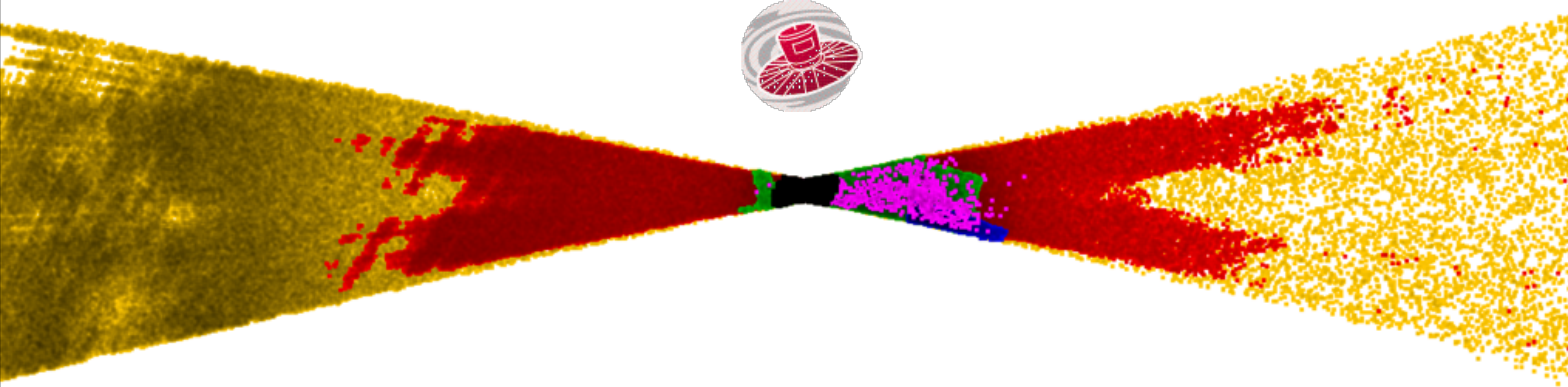


Crossmatching variable objects with the Gaia data

Lorenzo Rimoldini, Krzysztof Nienartowicz, Maria Süveges, Jonathan Charnas, Leanne P. Guy, Grégory Jevardat de Fombelle, Berry Holl, Isabelle Lecoeur-Taïbi, Nami Mowlavi, Diego Ordóñez-Blanco, and Laurent Eyer

ADASS XXVI, Trieste (Italy)
16-20 Oct. 2016

UNIVERSITÉ DE GENÈVE

Old topic, new shoes

Crossmatch of celestial objects:

- 🕐 Since a long time ago
- 📚 Useful to combine information:
  - Different wavebands
  - Epochs (time series)
  - Results (periodicity, classification, etc.)

Simple **positional** method (e.g. the nearest neighbour within a few arcsec, possibly dropping cases with multiple neighbours for safety):

- 😀 Many correct matches
- 😟 An embarrassing number of incorrect matches

Problem:

- Ground-based positional uncertainties can be large
- Proper motion (not always available)
- Gaia may split some blended sources
- Survey-specific artefacts (spurious sources)
- Variability signal (e.g. eclipsing binaries, long period variables)
- Ever growing number of sources

Solution: an “intelligent” crossmatch (AI)

Before any crossmatch, verify the Equinox of the reference systems and the corresponding Epochs (if proper motion is not negligible)!

Crossmatch by supervised classification

Machine learning in the **Variability Processing and Analysis** of Gaia data:

- Crossmatch* (Richards et al. 2012)
- Variability detection
- (Multi-)periodicity identification
- Classification (variability types)
- and more…

Crossmatch: typical task of a binary classifier (match, non-match)

- Automate decisions we would do by visual inspections (with multiple sources of information)
- Apply to millions of objects

*Not related to the crossmatch in the Gaia Archive (http://archives.esac.esa.int/gaia/)

Classifier pros

- Variety of **attributes**: position, mean photometry, colours, time-series features, catalog attributes, etc.
- Better than a **single** multi-dimensional metric, because:
  - Robust to inaccurate components
  - It does not have to depend on (often imperfect) uncertainties
  - It adapts to the data, not theoretical expectations
- Different photometric bands can be compared **directly** without a-priori **transformations** (ingredients included as attributes)
- If **mix** of similar/dissimilar features:
  - Train as a **match** (if dissimilar features are not relevant)
  - Train as a **non-match** (e.g., no interest in an eclipsing binary without eclipse)
- Recover matches with low **positional accuracy** or significant **proper motion** without knowledge of positional errors or models of the object motion
- It returns a **score** of crossmatch reliability

Classifier cons

🏃 Results depend on **training**

- Proper training (see later)
- Check misclassifications
- Iterate

🏃🏃 Every survey is unique (attributes, bands, sampling): train a **separate** classifier for each catalog

🏃🏃🏃 **Multiple** classifiers (training sets) per survey:

- Select easy matches first
- Dedicated classifier(s) for the **difficult** cases

⌛ **Time** ~ 1 day / catalog

(for < 1000 targets, visual confirmation of matches is faster)

9 STEPS TO CROSSMATCH WITH A CLASSIFIER

1. Define the purpose

Classifier adapted to purpose:

- ✨ **Training** classification of variables
  - Signal shape is relevant (no eclipsing binary without eclipse)
  - Match probability > 0.5
- 🌐 **Completeness**
  - May limit to position, mag, color: no dependence on signal shape or time series sampling
  - Match probability < 0.5

2. Find neighbours

- Neighbours within 5 arcsec (or more)
    - Positional accuracy
    - Proper motion
- Database queries (PostgreSQL, Q3C spatial indexing)

Koposov & Bartunov, ADASS XV

3. Compute attributes

Attributes for targets and all their neighbours:

- Angular separation
- Magnitude (difference)
- Color (difference)
- Number of observations
- Amplitude
- Various statistics
- Correlations
- Parameters on folded light-curves (if periods are known)
- Survey attributes
- etc…

4. Select training set sources

For both **match** and **non-match** classes

A. Good **representation** of variability types, colours, magnitudes, sampling, artefacts, separation distances, data quality

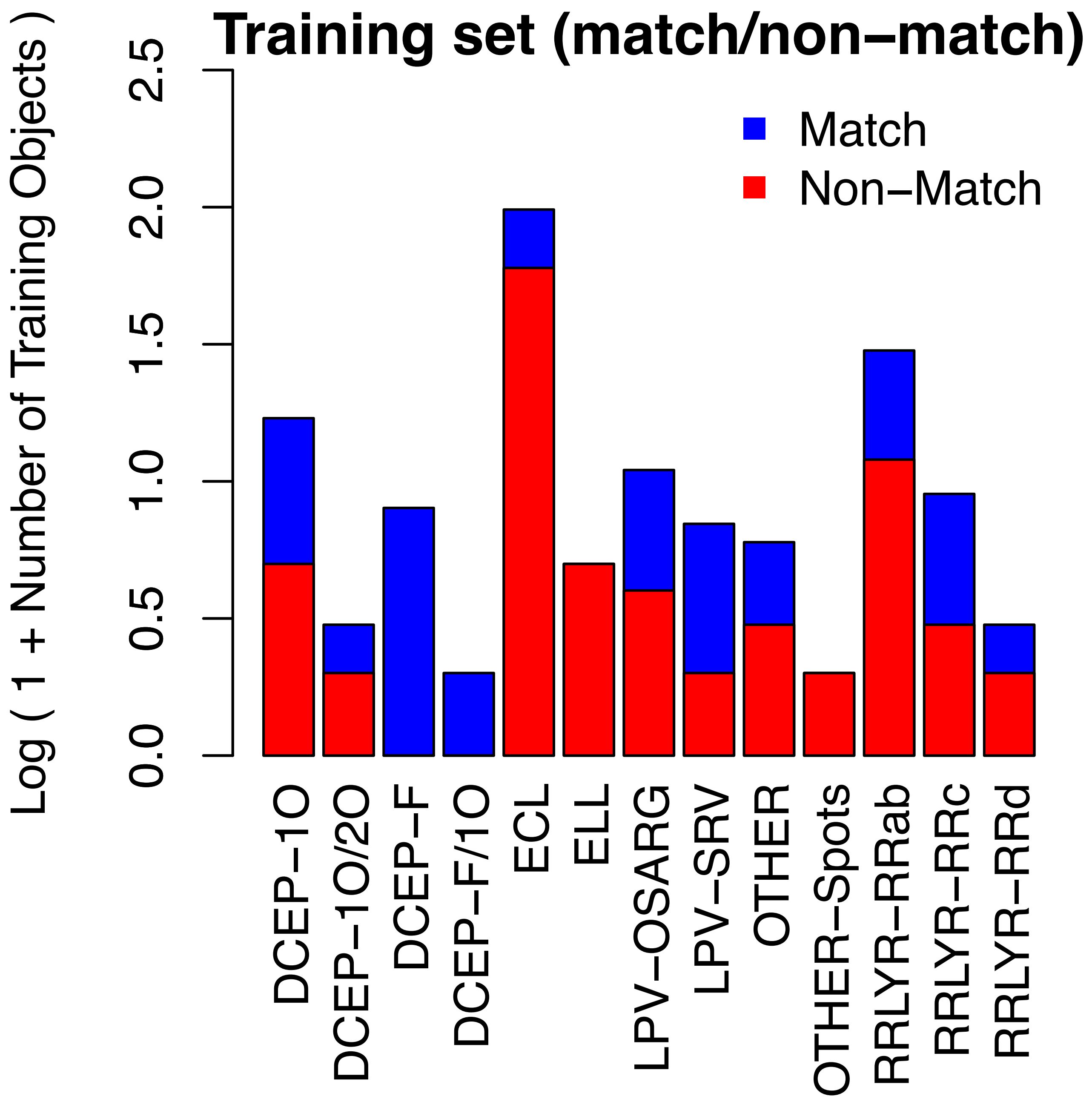
B. Not only the obvious cases: teach the classifier as many **challenging** decisions as possible

C. Embed **all the reasons** which drive decisions during visual inspections

D. Check **misclassifications** (false positives/negatives) and improve their correct representation until they are in the “grey region” (acceptable mistakes)

4. Select training set sources

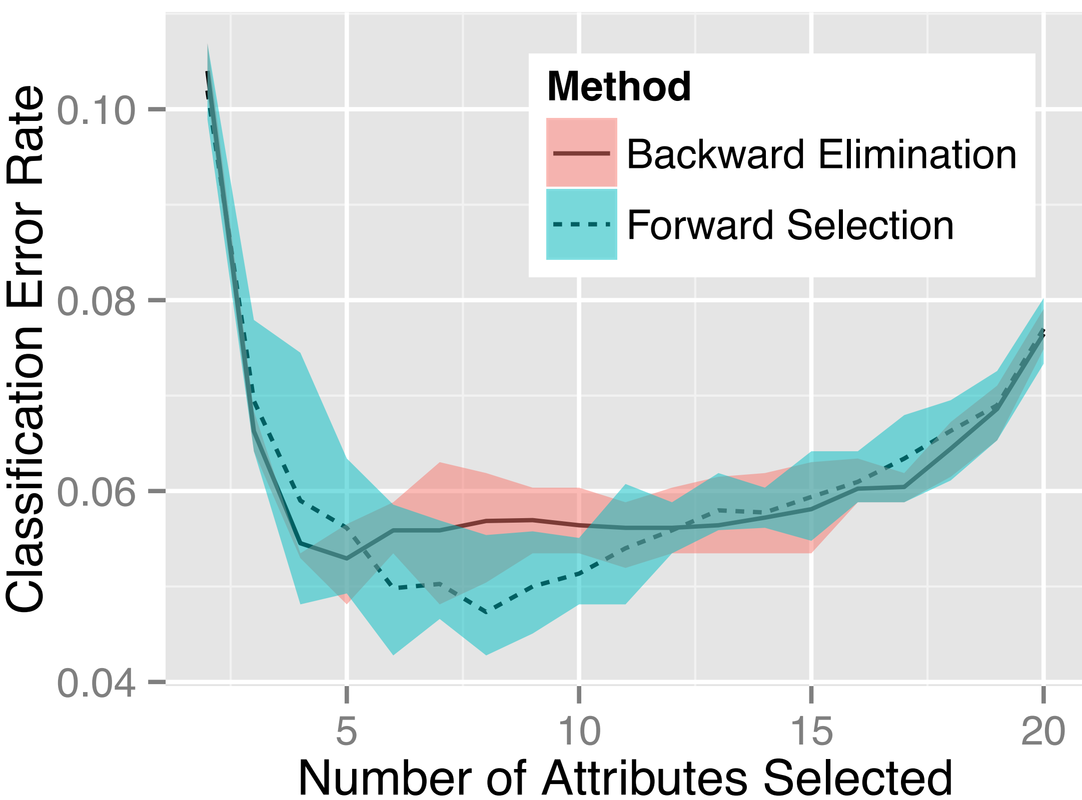
Visualise light curves of
targets to crosscmatch
vs
all neighbours
(folded by period if known)

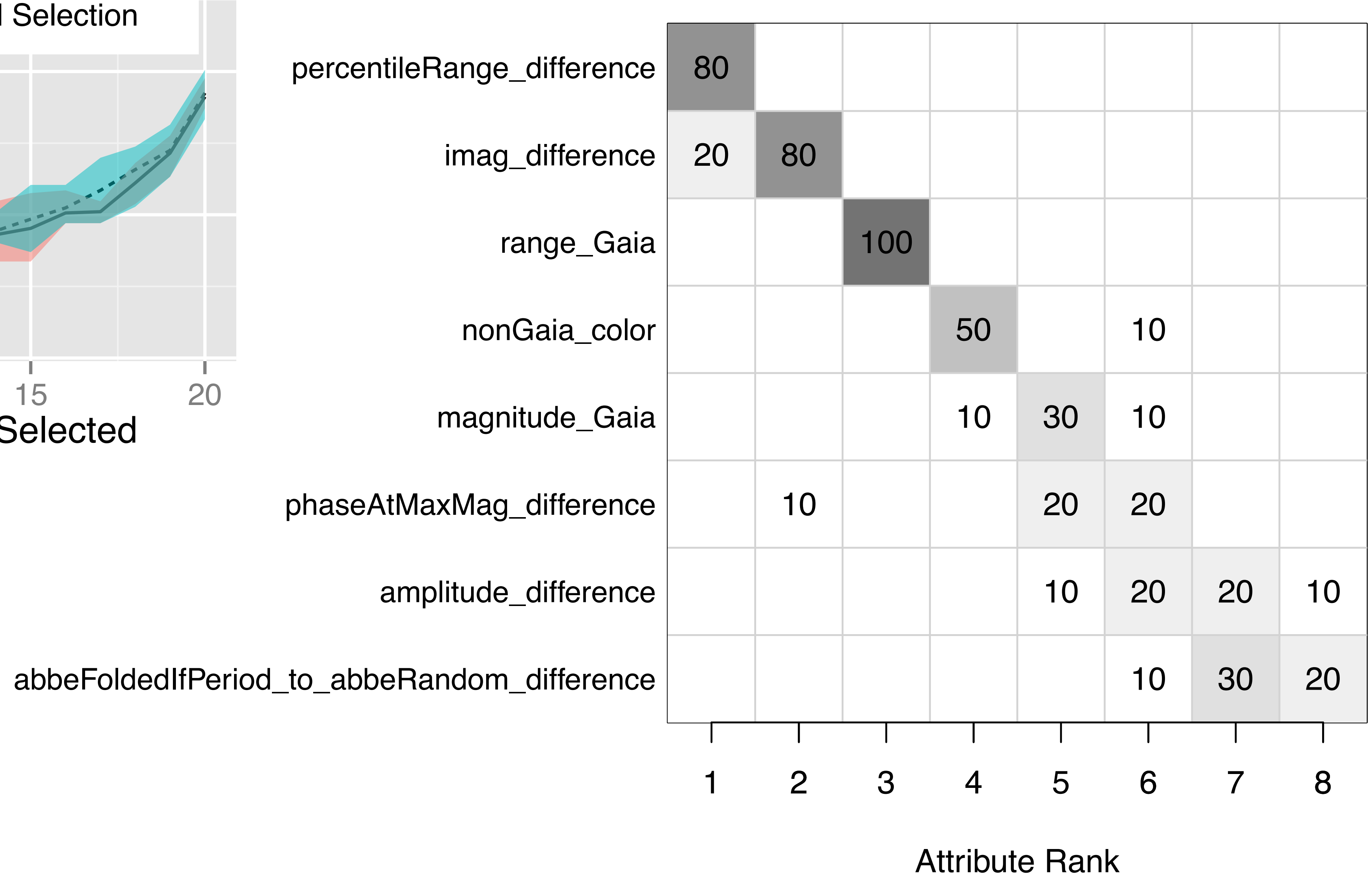


5. Optimise classifier

- Select useful (not just important) attributes
- Optimise classifier (tuning parameters)
- Assess classifier (confusion matrix)

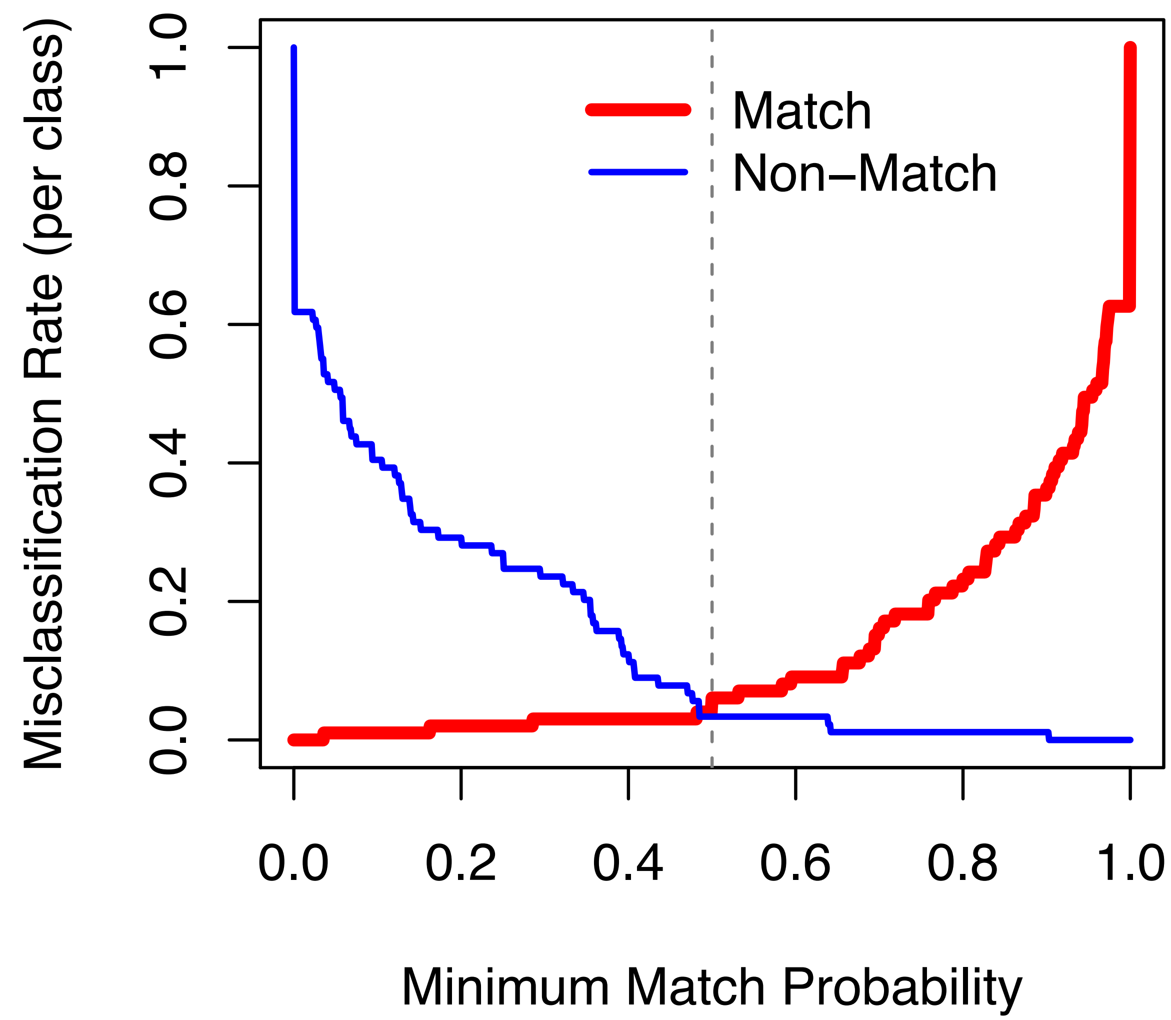
Attribute selection





Classifier assessment

# obj./class		match	non_match	
99	match	96	4	%
89	non_match	3	97	%
	Contamination	3	4	%



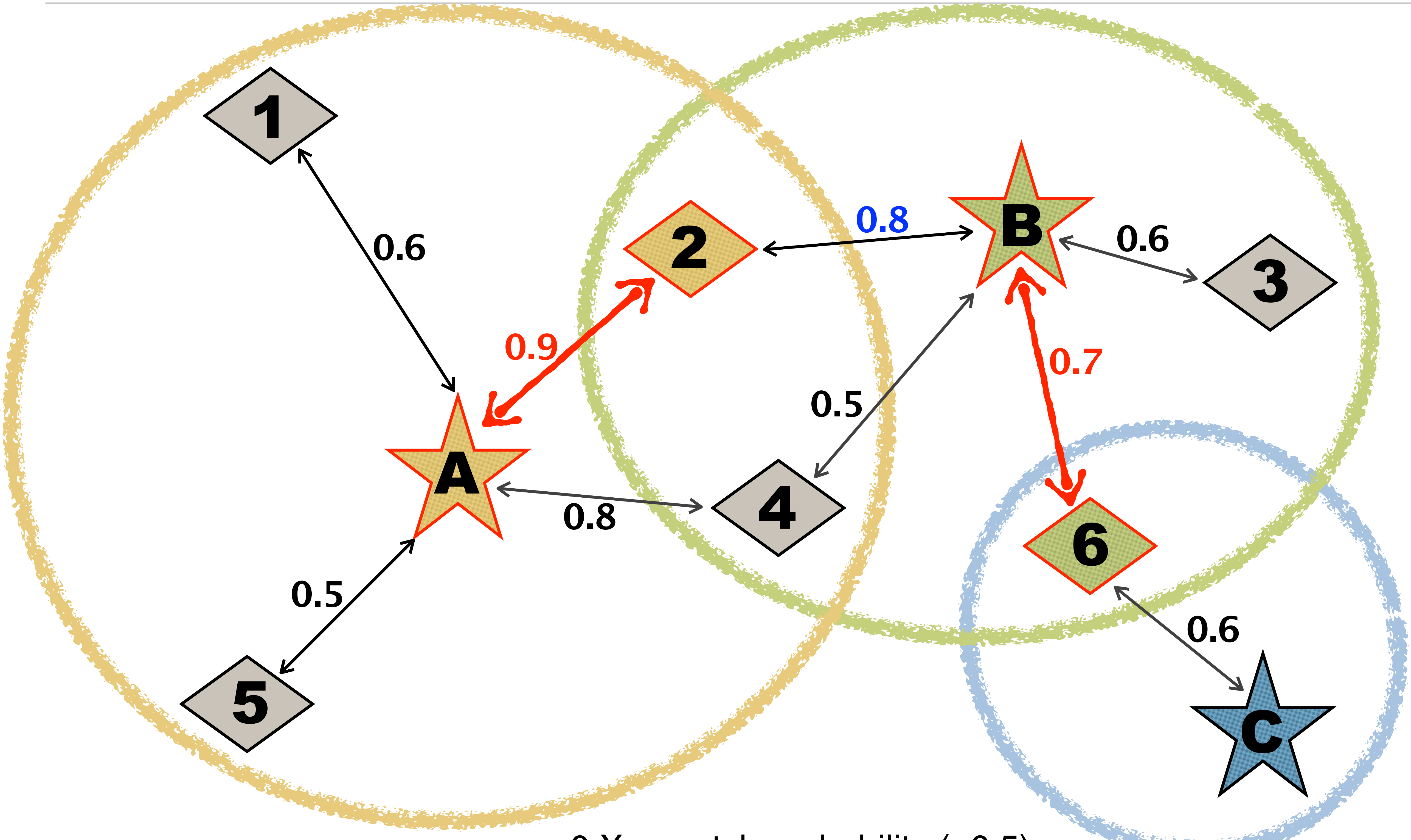
6. Classify (match / non-match)

- Techniques for **missing** attribute values
- **Predict** on data to crossmatch
- Assumed **only one** match for each target and **vice-versa**

Multiple matches per target?
Take the one with the **highest** probability

Among the **selected** matches, if more than one is associated with the **same target**, different options possible (e.g., the highest probability first)

Matches in crowded fields



0.X = match probability (>0.5)

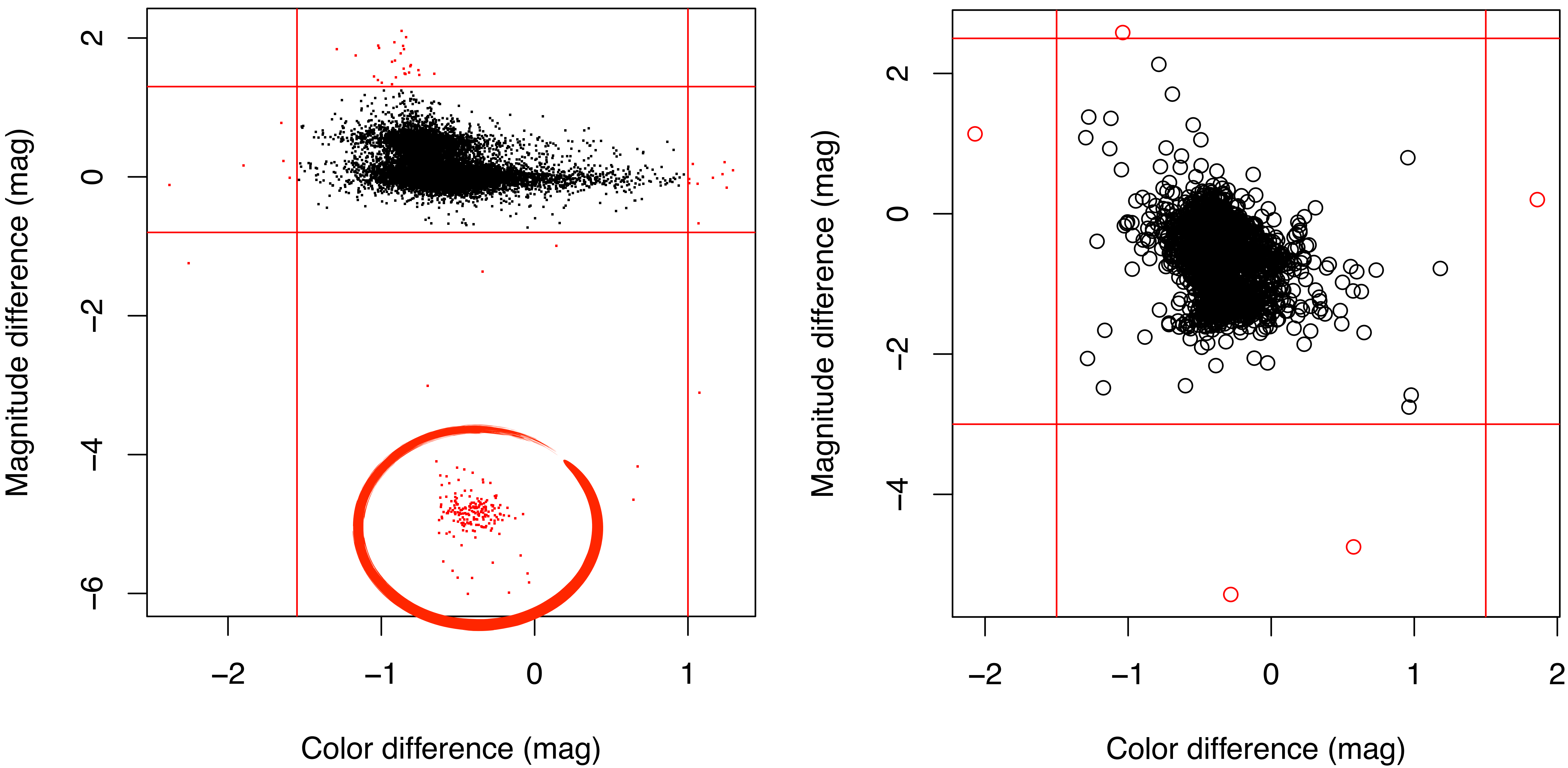
7. Assess results

Verify:

- Prediction statistics
- Low-probability matches
- Low-probability non-matches
- Farthest matches
- Nearest non-matches

- Feed incorrect classifications back to the training set
- Iterate steps 4 to 7 (until misclassifications are acceptable)

8. Sanity checks



9. Difficult cases

Repeat steps 4 to 8:

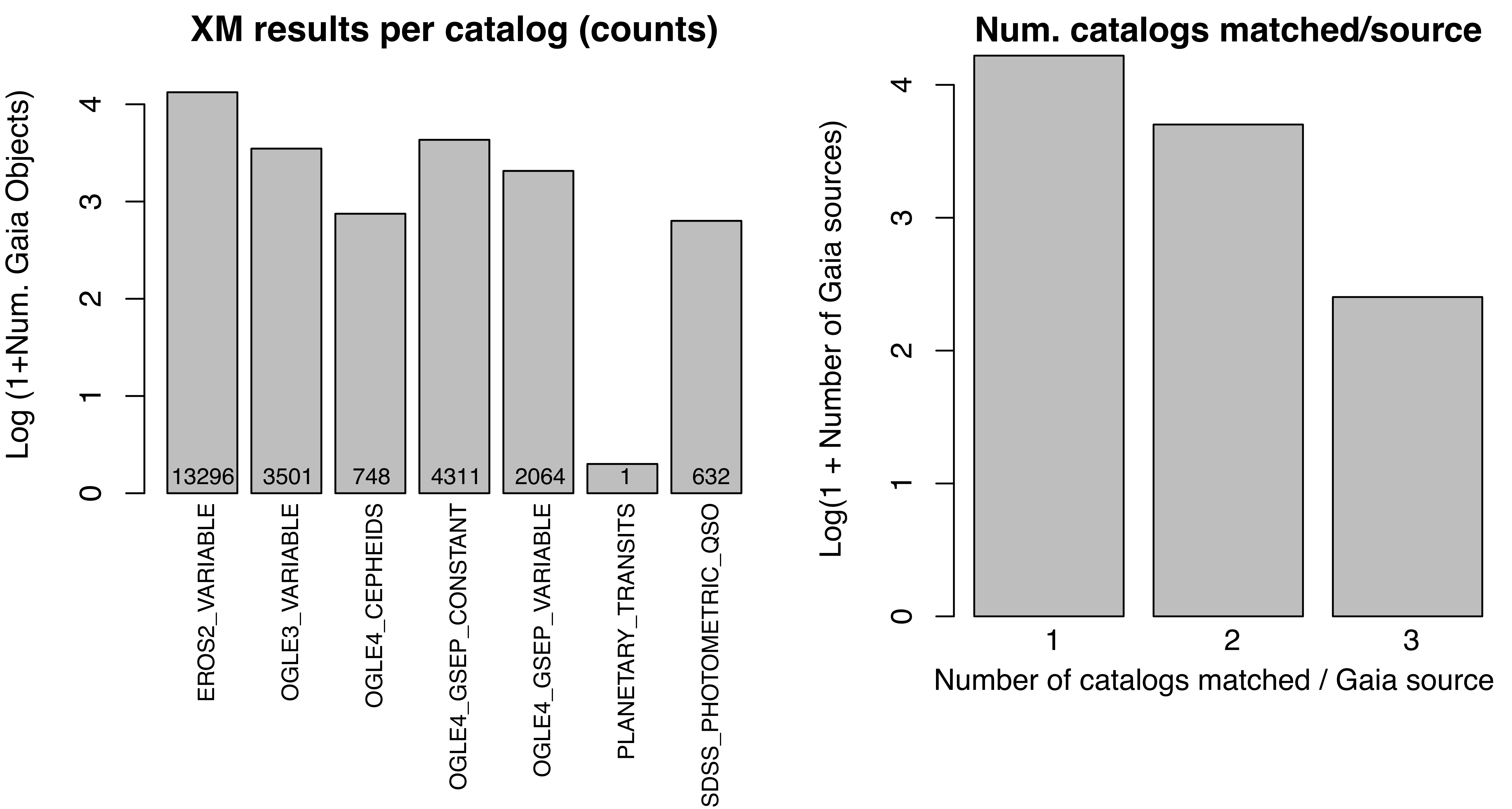
- Train additional classifier(s) **dedicated** to difficult cases (if needed)
- Reclassify low-probability matches and all non-matches

Surveys crossmatched with Gaia

Mostly around the **South Ecliptic Pole** (near the Large Magellanic Cloud):

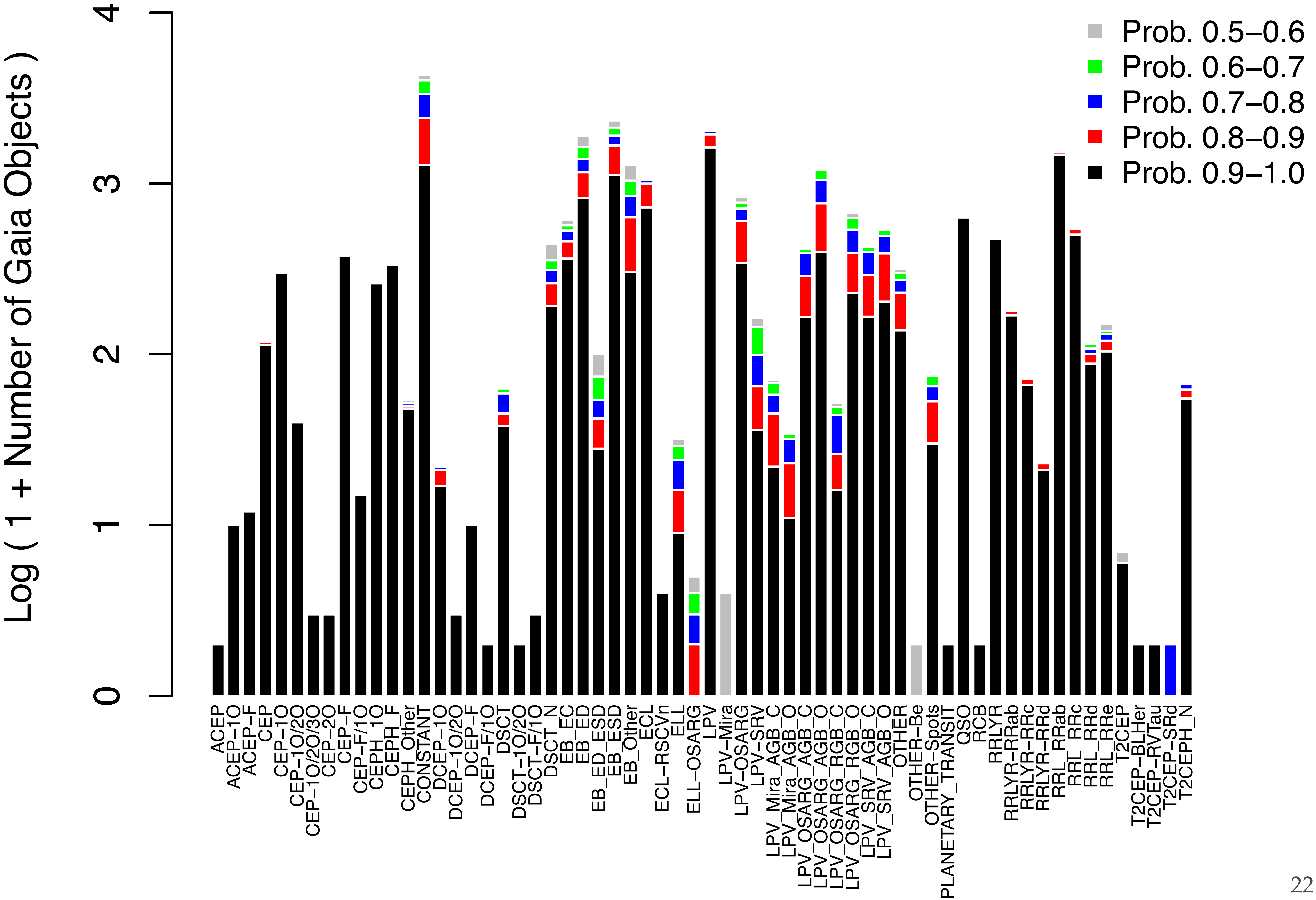
- The **OGLE4 GSEP** variable stars (Soszynski+ 2012)
- The **OGLE4 GSEP** constant star candidates (OGLE4/GSEP/maps)
- The **OGLE4** Cepheids (Soszynski+ 2015)
- The **OGLE3** variable stars (Udalski+ 2008)
- The **EROS2** periodic variable stars (Kim+ 2014)
- High-confidence (99%) SDSS photometric **quasar** candidates with radio and/or X-ray association (in the Half Million Quasar catalog, Flesch 2015)
- Confirmed **planetary transits** (Southworth, as of Aug. 2015)

Surveys crossmatched with Gaia



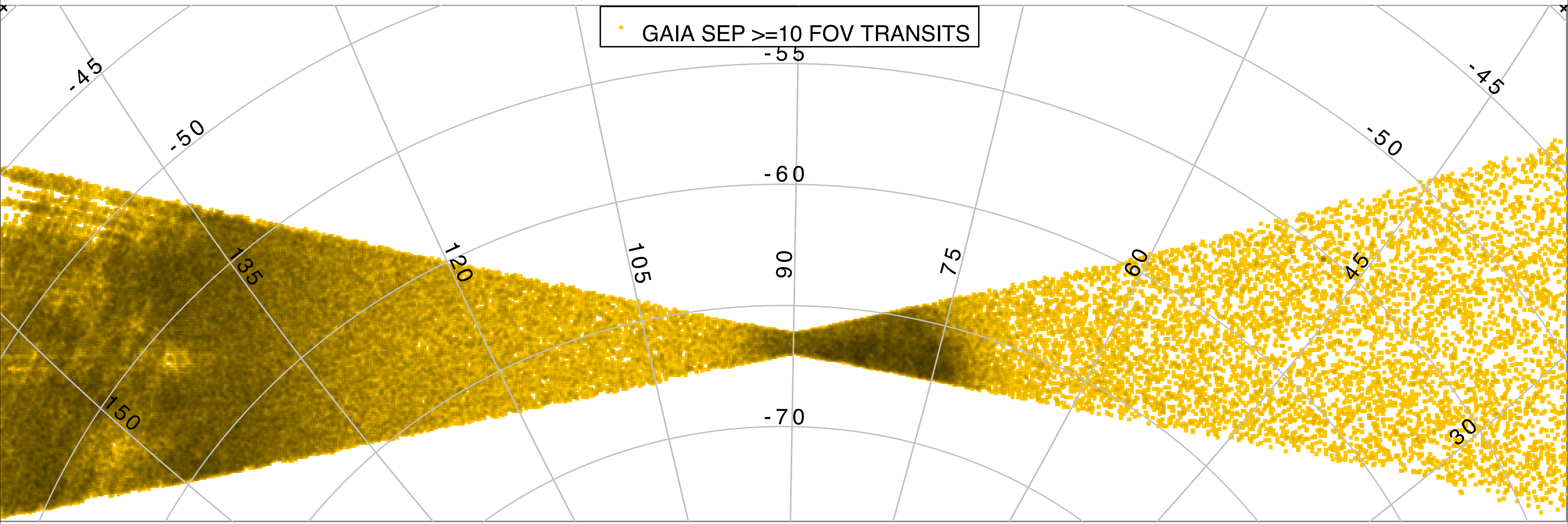
Crossmatch with Gaia sources sampled by at least 10 Field-of-View transits in the G band

XM results per type and match probability



Gaia near the South Ecliptic Pole (SEP)

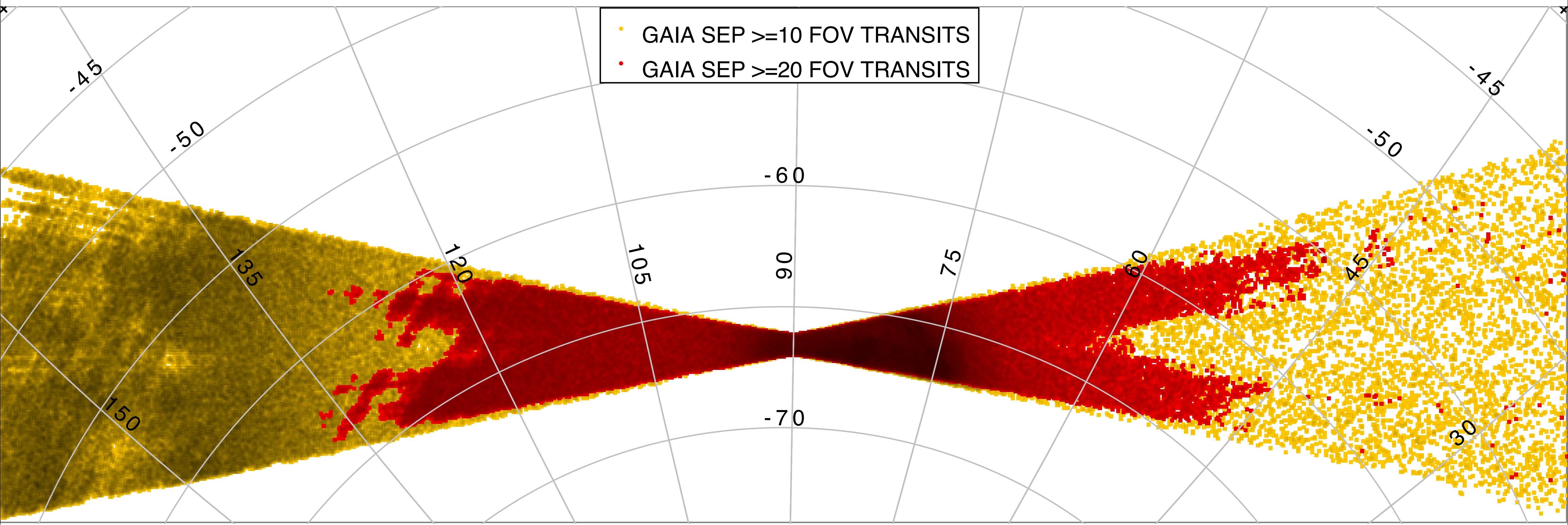
[preliminary data, subset of data release 1]



Equatorial Coordinates (deg)

Gaia near the South Ecliptic Pole (SEP)

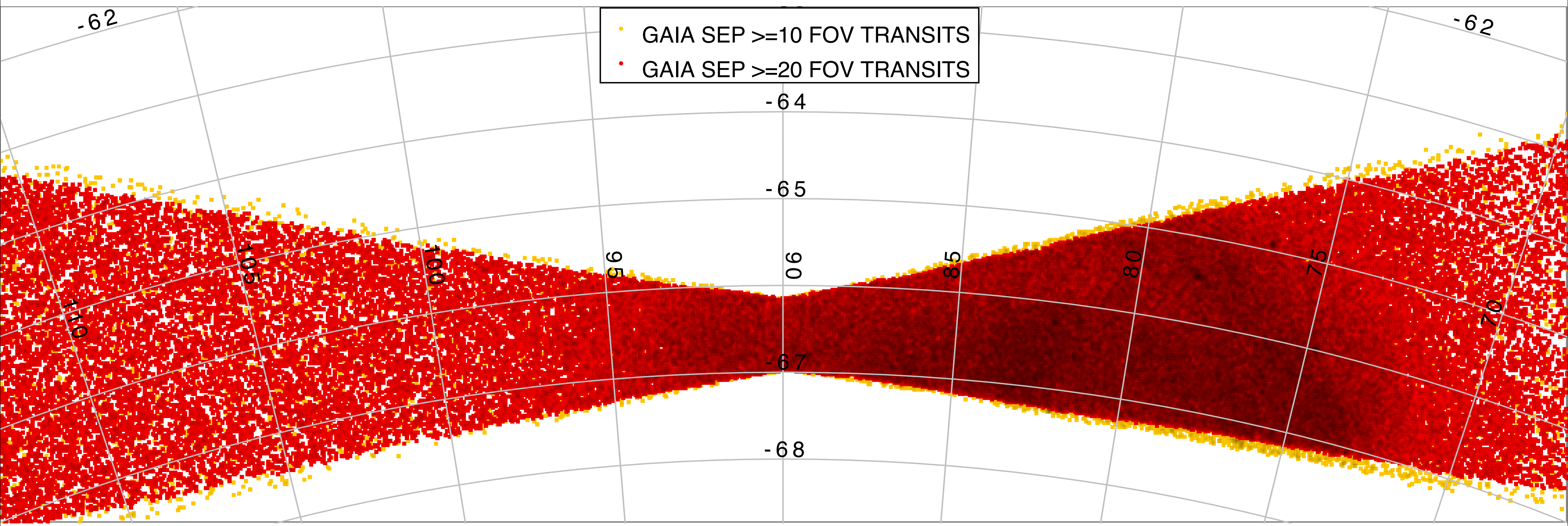
[preliminary data, subset of data release 1]



Equatorial Coordinates (deg)

Gaia near the South Ecliptic Pole (SEP)

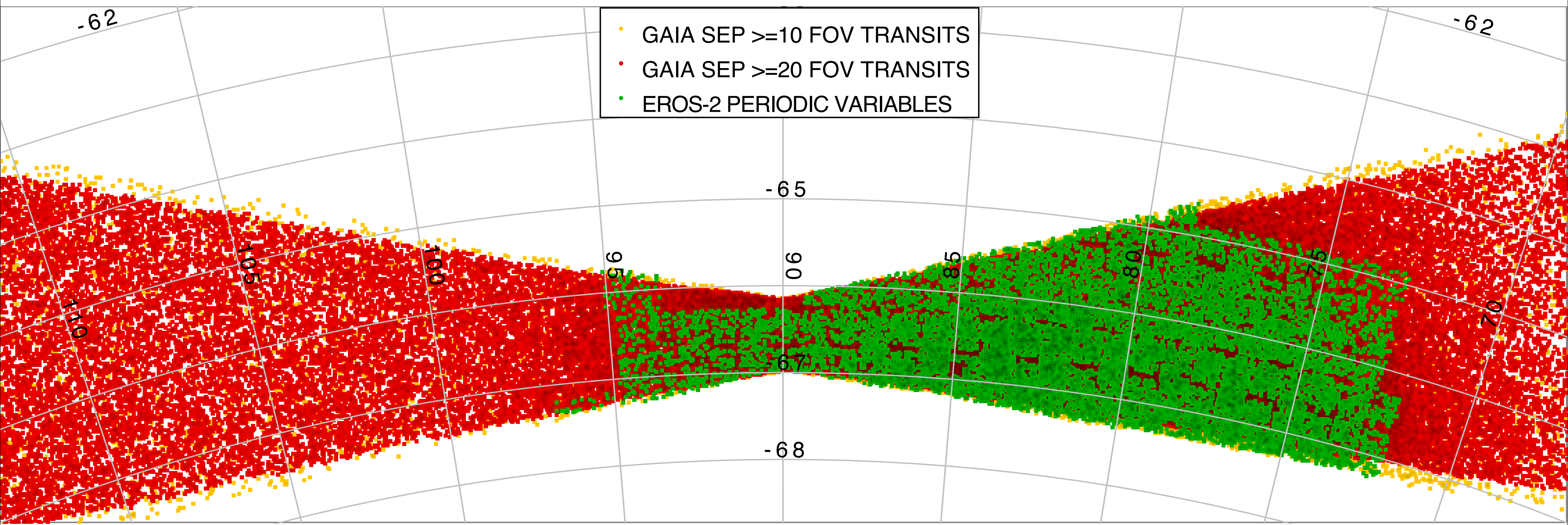
[preliminary data, subset of data release 1]



Equatorial Coordinates (deg)

Matches near the South Ecliptic Pole (SEP)

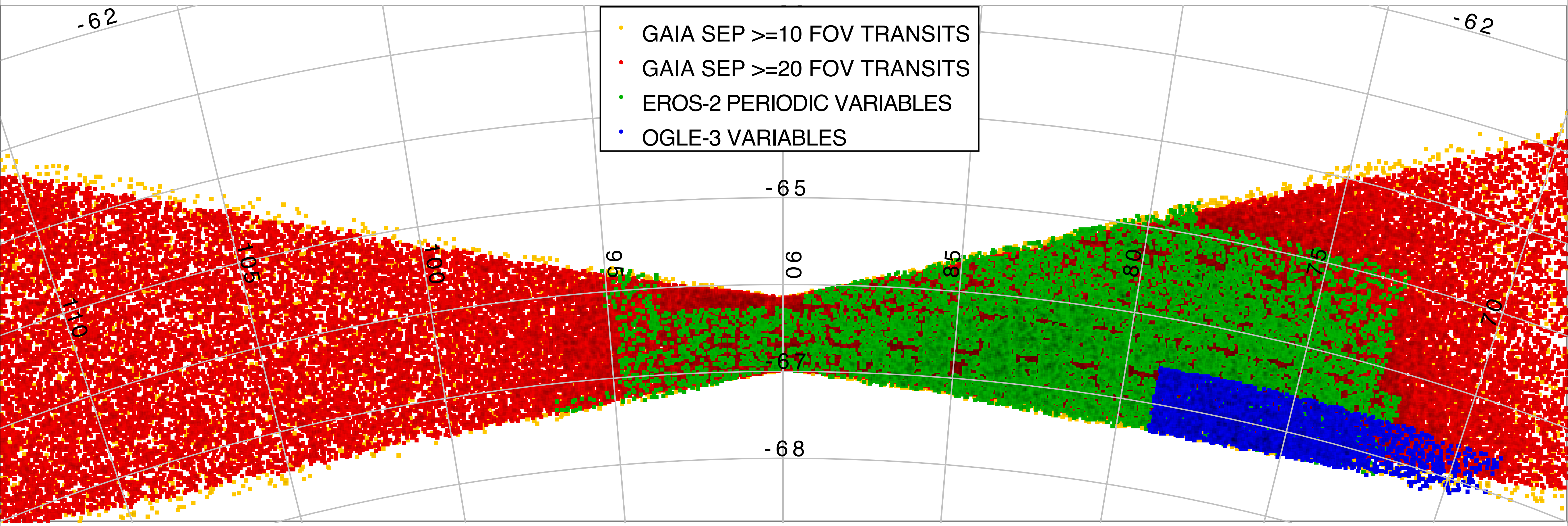
[preliminary data, subset of data release 1]



Equatorial Coordinates (deg)

Matches near the South Ecliptic Pole (SEP)

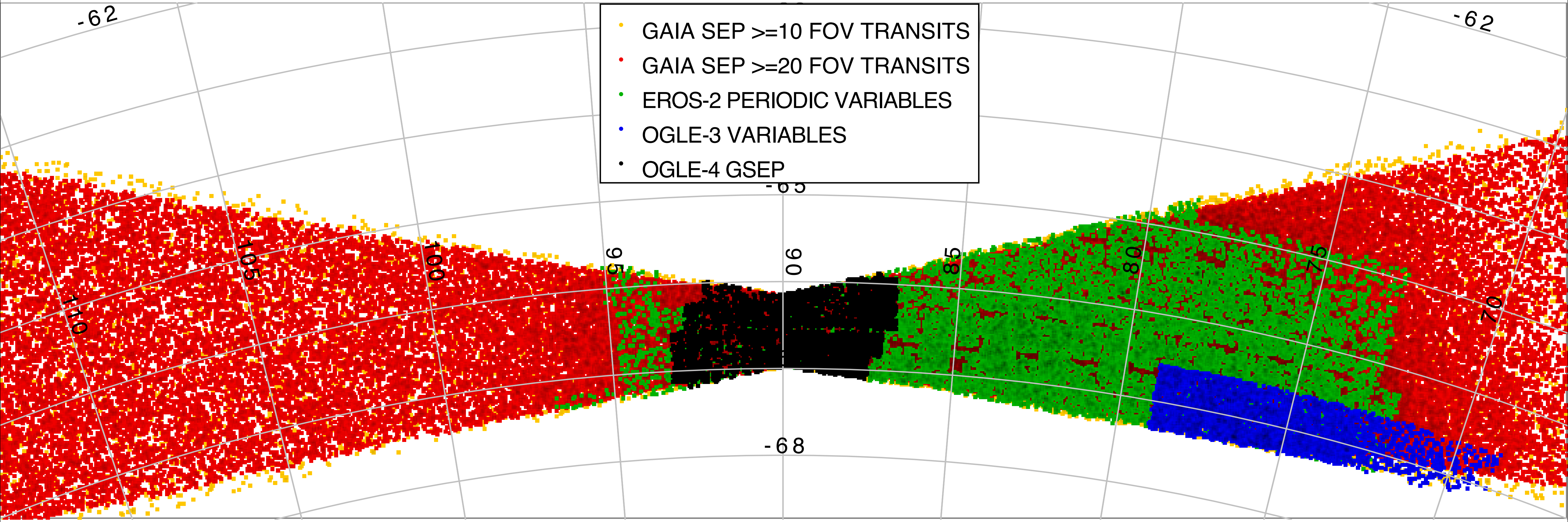
[preliminary data, subset of data release 1]



Equatorial Coordinates (deg)

Matches near the South Ecliptic Pole (SEP)

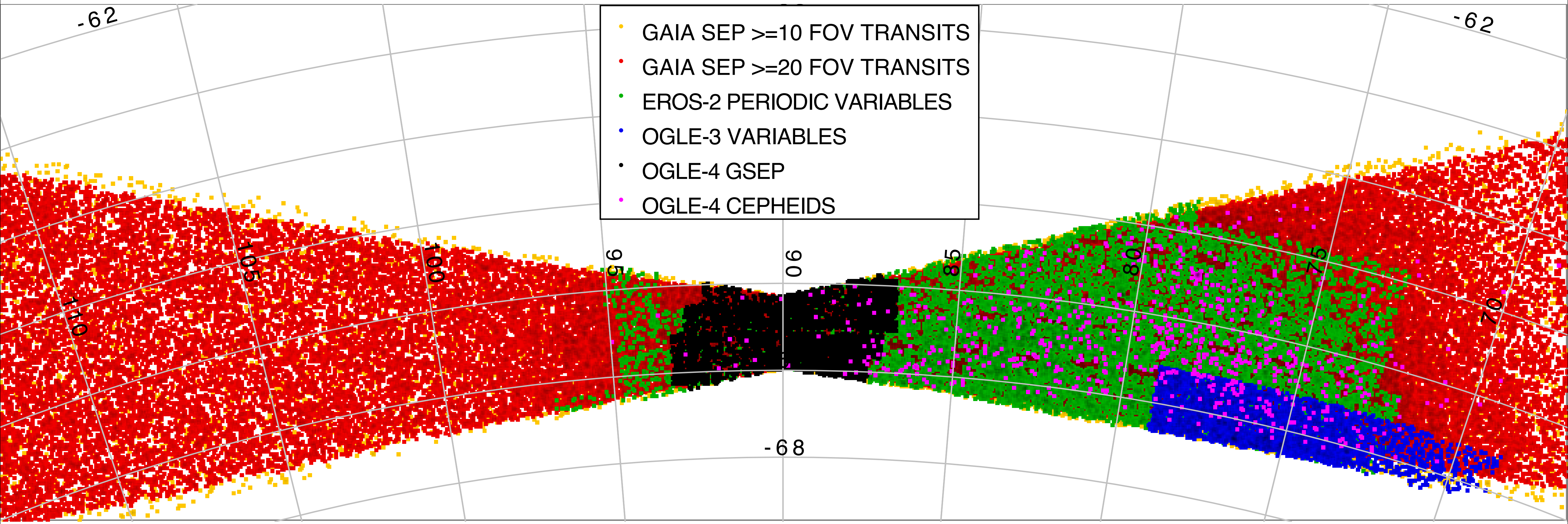
[preliminary data, subset of data release 1]



Equatorial Coordinates (deg)

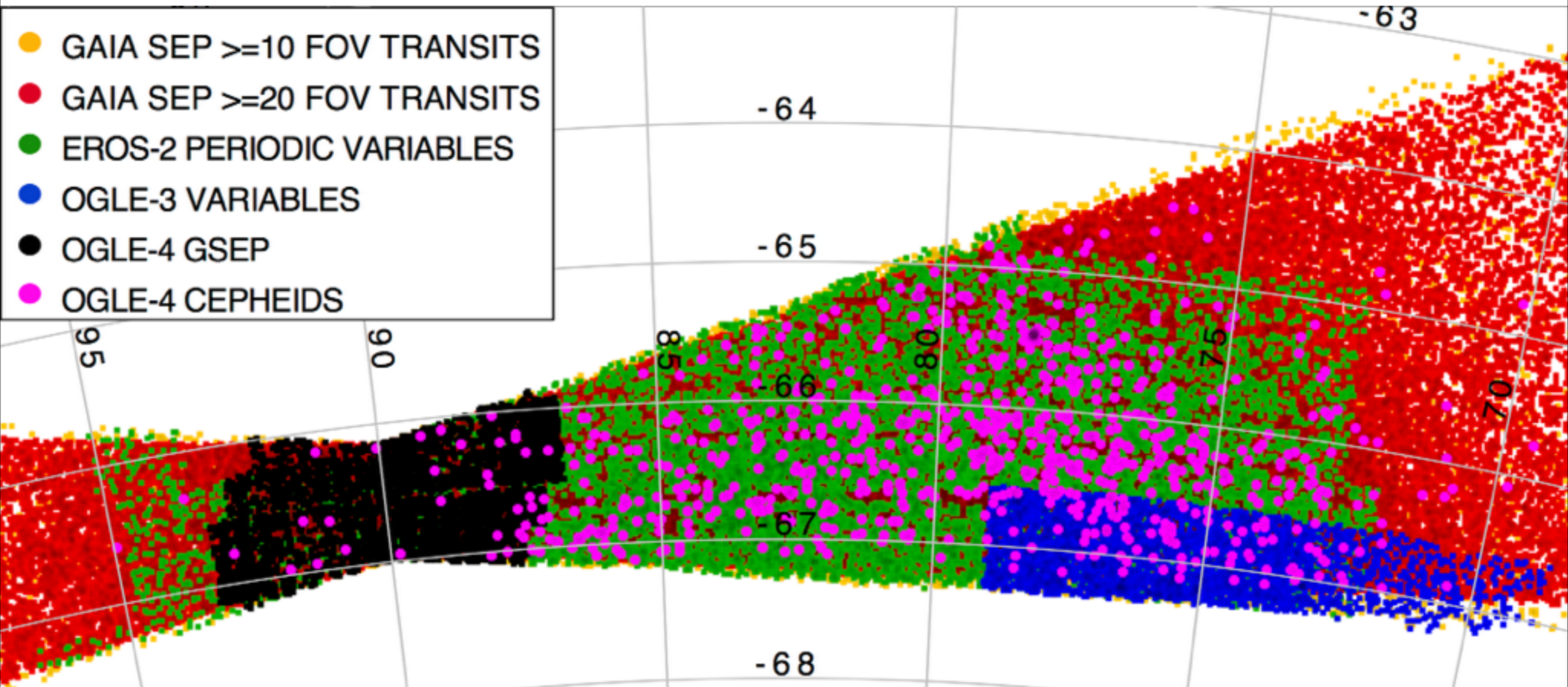
Matches near the South Ecliptic Pole (SEP)

[preliminary data, subset of data release 1]



Equatorial Coordinates (deg)

Variable star matches used in the Gaia data release 1



Equatorial Coordinates (deg)

Eyer et al. (submitted)