This document has been created as input for the upcoming biomarker evaluation meeting with ▮▮▮▮▮ where this initial candidate list will be discussed. After the evaluation meeting Euretos will undertake a detailed analysis of selected markers. The following research question is addressed:

Find potential novel biomarkers for Sarcoidosis that are secreted (i.e. easily accessible)

Approach

The approach leverages the Euretos Knowledge Platform (EKP), where over 175 data sources are consulted simultaneously. The platform generates for each criterion gene lists that meet the criterion. These gene sets are then overlapped to find those gene candidates meeting the highest amount of criteria.

Result

- **13929 potential markers** (i.e. meeting at least one criterion) were analysed
- **Four main groups of criteria were evaluated**:
    1. *Healthy Expression (13847 genes)*: RNA expression in the lung
    2. *Differential Expression (242 genes):* in healthy versus Sarcoidosis tissue
    3. *Associations with disease specific characteristics (684 genes)* in particular genes associated with innate immunity and t-cell activation
    4. *Associations with Sarcoidosis (307 genes):* Any gene associated directly in literature with Sarcoidosis.
- **3 genes were found that meet criteria in all 4 categories and are also secreted:**
    - 5 genes were found that meet criteria in all four groups: ▮▮▮▮▮, ▮▮▮▮▮, ▮▮▮▮▮, ▮▮▮▮▮, ▮▮▮,
    - Of these 5 genes, 3 are also secreted: ▮▮▮▮, ▮▮▮▮, ▮▮▮.
    - Two genes, ▮▮▮▮ and ▮▮▮, are also overexpressed and highly associated with Sarcoidosis

In the following section these potential markers will be further outlined. The attached Excel sheet provides this overview including all the gene sets for the various criteria.

The following 5 genes meet all the initial requirement. Three of these are also secreted:

Gene	Classification	Localisation
▮▮▮▮▮▮	Chemokine Plasma proteins Signaling molecule Cytokine Predicted secreted proteins Cancer-related genes	signalling secreted
▮▮▮▮▮▮	Interleukin superfamily Signaling molecule Cytokine Cancer-related genes	signalling
▮▮▮▮▮▮	Chemokine Plasma proteins Signaling molecule Cytokine Predicted secreted proteins Cancer-related genes	signalling secreted
▮▮▮▮▮▮	Predicted membrane proteins Plasma proteins Cancer-related genes	transmembrane
▮▮▮▮▮▮	Defense/immunity protein Predicted membrane proteins Major histocompatibility complex antigen Predicted secreted proteins	secreted

For the three secreted genes the following detailed information is available:

Of these 3 markers, especially ▮▮▮▮▮ and ▮▮▮▮▮,look promising as they are highly overexpressed, a high disease association score (see addendum) and have limited interactions with other genes. The ▮▮▮▮▮ gene also has a relatively high normal expression in the lung.

|	Name	Differential Expression	Baseline expression value	Genetic interactions	Involvement in other diseases	Functional association
☐	▮▮▮▮▮	6.9	53.33	22	83	95.1
☐	▮▮▮▮▮	6.6	18.28	21	85	95.1
☐	▮▮▮▮▮	-4.5	11	5	125	93.61

Next Steps

These markers will be discussed during the biomarker survey meeting. During this meeting the model will be refined based on expert feedback after which the key markers will be selected for further, more detailed analysis of the molecular mechanism.
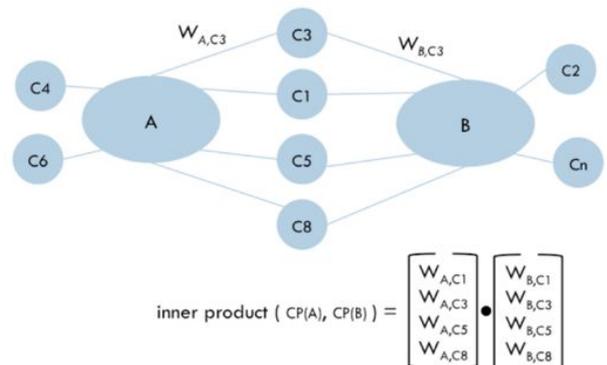
Addendum - Disease association score

An objective measure comparing the number and specificity of indirect relations between two (*related or unrelated*) biological concepts against a reference set of related concepts of the same type. The resulting score is the is percentile rank against the frequency distribution of a reference set.

$$CP(A)= \begin{bmatrix} W_{A,C1} \\ W_{A,C3} \\ W_{A,C2} \\ W_{A,C5} \\ W_{A,C8} \\ \vdots \\ W_{A,Cn} \end{bmatrix}$$

The Disease Association Score ranks two concepts (*whether they have a direct relation or not*) in terms of their indirect relations, against a reference set of concepts of the same semantic category for which direct relations do exist. It provides an objective measure of the level of mutual information two concepts share. In cases where no direct relation exists the Disease Association Score act as a predictive statistical function as it compares the level of mutual information of both existing and non existing relations.
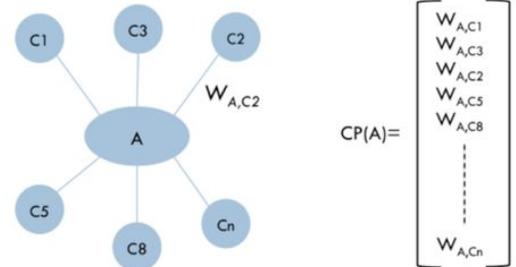
The Disease Association Score uses the methodology of "concept profile analysis" as developed at the Leiden University Medical Centre (LUMC), and described in scientific publications (eg van Haagen HH et al ; " Novel protein-protein interactions inferred from literature context" , 2009).

$$\text{inner product } (\, CP(A), CP(B) \,) = \begin{bmatrix} W_{A,C1} \\ W_{A,C3} \\ W_{A,C5} \\ W_{A,C8} \end{bmatrix} \cdot \begin{bmatrix} W_{B,C1} \\ W_{B,C3} \\ W_{B,C5} \\ W_{B,C8} \end{bmatrix}$$

The score is calculated as follows:

- Selected Concepts: The two selected biological concepts for which the functional association will be calculated. Each of the concepts belongs to a **semantic category**.
- Reference set: set of existing relations in the dataset where the concepts in the relationship have the same **semantic category** assignment as the Selected Concepts that are being compared.
- Weight: distance score between 2 related concepts, indicate similarity (represented by the Jaccard index) of the connected concepts based on context in the graph environment.
- Concept profile: Summation of all weights associated to a single concept
- Concept profile analysis score: the inner product of the weights associated with relations that the Selected Concept have in common.
- Functional Association Score: the percentile rank of the concept profile analysis score on the 2 concepts in the hypothesis, compared to the frequency distribution of the concept profile analysis scores of the reference sets