# Hypothesis formation

## Pre-assumed knowledge about hkp

1. Stable expression: stably expressed across various cell types, tissue and conditions 🡪 often used as normalization references (B)
2. Cellular maintenance function: core fundamental cellular biological functions, i.e. crucial for maintaining basic cellular functionality, e.g., transcription, translation, metabolism, cell cycle regulation, etc. (N)
3. **Essentiality:** essential for cellular survival 🡪 KO leads to lethality/significant dysfunction (H)
4. **Conservation (across species): highly conserved across species** (due to fundamental biological roles) (E and an additional one)
5. Commonly used hkgs: ACTB, GAPDH, HPRT1, B2M (B, C, F)

## Recent Findings and Uncertainties

1. HKPs not stable expressed:
   1. Large-scale transcriptomics datasets (GTEx V8, microarray meta-analyses) showed significant variation between tissues, conditions and individuals (F)
   2. Inter-individual variation is another source to consider (same species, different subject samples with expression profile fluctuation based on age, gender, disease state and genetic background)
2. Cellular maintenance function: the assumption seems to hold throughout the hkg/hkp studies 🡪 can be re-confirmed with a gene/protein list and functional enrichment
3. HKPs not (always) essential:
   1. KO without causing cell death
   2. Certain genes/proteins are tissue-selective (tailored to biological context)
4. HKPs not (all) highly conserved: we might not consider this criterion in depth since we focus on human HKPs, not across species

## Hypotheses and data analysis plan

### Define HKP criteria

Based on the research on 4 assumed criteria of HKGs/HKPs, we propose these criteria:

* **Functionality:**
  + This criterion holds up well. Most HKPs are involved in basic cellular processes (translation, transcription, metabolism, etc.).
  + Functional enrichment can help validate this as one approach to the “expected” HKGs/HKPs list. We expect the central nodes in functional interaction networks (central hub genes) to be overrepresented among HKPs.
* **Essentiality:**
  + Many HKPs are essential, but exceptions exist (e.g., tissue-specific enrichment).
  + We suggest computing a degree of essentiality, in which we extract the essential rate of each gene/protein being KO from public databases (e.g., DepMap of CRISPR KO essential screens on cancer cell lines).
  + This could serve either as a strict criteria (when we predefined a fixed threshold of essentiality) to be a HKG/HKP or a source of analysis after we define a list of HKGs then define an data-driven threshold
* **Stable expression:**
  + This criteria has been rejected in many publications and proven as dataset/context-dependent
  + We think we will do data analysis first then investigate different tissues/contexts to try to remove that enriched expression
  + This aims to find a set of HKPs actually being stably expressed despite tissue/context, or at least within a given range of variability, allowing mild fluctuation only
  + We might also consider the variation between individuals
* Conservation: not really considered in this study

### Data analysis plan

We aim to do data analysis partially depending on the predefined criteria above and then do data-driven refinement iteratively to output a reasonable HKGs/HKPs list.

* Stable expression:
  + Identifying universally expressed genes based on both RNA and protein
    - Normalize RNA-seq & proteomics data
    - Choose a level of stable expression, e.g., based on mean expression across tissues, standard deviation (low variability suggests stability), Gini coefficient (inequality of expression), etc.
  + Filtering out tissue-specific gene expression
    - Extract tissue-enriched profile from HPA (or re-analysis to extract)
    - Remove genes that are highly enriched in one or a few tissues
    - Compare against cell line data (to exclude culture-induced expression biases)
  + (Suggested by GPT) RNA-protein correlation expression
    - Spearman/Pearson correlation between RNA and protein levels
    - Identify genes that maintain consistent trends at both RNA and protein levels
    - Investigate outliers (genes with high RNA but low protein, indicating post-transcriptional regulation)
  + Output: a list of potential HKGs/HKPs and the threshold of stable expression for chosen HKGs/HKPs
* Functionality:
  + For the potential list of stable expression analysis
    - GO enrichment
    - Pathway analysis to determine overrepresented pathways
    - Network analysis to find key regulatory hubs (central genes with high connectivity). We could rank by network connections, a higher rank might be more reliable as HKPs (only hypothesize here)

🡪 see what pathways expected HKPs to be enriched, if they are related to expected functions of cellular processes

* + For another approach:
    - Choose certain cellular maintenance pathways assumed to have many HKPs involved (e.g., ribosome, glycolysis, transcriptional regulation)
    - Extract all genes in those pathways (e.g., from KEGG, Reactome, etc.) and maybe rank them in some ways related to HKPs (e.g., more central pathways participated?)

🡪 obtain an independent functional reference set of genes/proteins

* + We do overlap analysis, comparison, refinement, data-driven based, etc.
  + Output: a potential list of HKGs/HKPs and the function-enriched analysis
* Essentiality
  + Check the degree of essentiality of the potential list on DepMap
  + Define a threshold of essentiality for HKGs/HKPs
  + Output: a potential list of HKGs/HKPs and the essentiality degree threshold
* Cross-validation
  + With GTEx data
  + With the gene list of other publications

# Literature review

## What is hkg

**Main:** challenge the tradition criteria of the hkp 🡪 stress the importance of our research

1. **Constant Expression Level Across Tissues and Conditions**:
   * **Pre-assumption**: HKGs are uniformly expressed at consistent levels across all tissues and conditions.
   * **Authors’ Stance**: **Disagreed**. The authors argue that no gene exhibits truly constant expression across all contexts. They present data showing variability in expression levels, even for traditionally considered HKGs, depending on specific cellular states or conditions.
2. **Essential for Basic Cellular Functions**:
   * **Pre-assumption**: HKGs are vital for fundamental cellular processes like metabolism, transcription, or translation, and thus are indispensable for cell survival.
   * **Authors’ Stance**: **Partially Agreed**. While many HKGs contribute to essential cellular functions, the authors caution against overgeneralization. They emphasize that some genes traditionally labeled as HKGs may not strictly meet this criterion in specialized or unique contexts.
3. **Universally Expressed Across Species**:
   * **Pre-assumption**: HKGs should be conserved and expressed across a wide variety of species, reflecting their fundamental biological importance.
   * **Authors’ Stance**: **Disagreed**. They highlight examples where certain HKGs exhibit species-specific expression or conservation patterns, challenging the universality assumption.
4. **High and Stable Expression Levels**:
   * **Pre-assumption**: HKGs maintain high and stable expression levels, making them reliable normalization controls in experiments like qPCR or RNA-seq.
   * **Authors’ Stance**: **Disagreed**. The authors provide evidence that expression levels of HKGs can fluctuate and may not always be ideal as normalization controls, particularly in diseased or highly variable samples.
5. **Low Expression Variability**:
   * **Pre-assumption**: HKGs exhibit minimal variability in their expression across populations or experimental conditions.
   * **Authors’ Stance**: **Disagreed**. They show data indicating that expression variability can occur due to technical noise, biological heterogeneity, or changes in regulatory mechanisms.
6. **Involved in Core Metabolic Pathways**:
   * **Pre-assumption**: HKGs participate in core metabolic pathways essential for the survival and function of cells.
   * **Authors’ Stance**: **Partially Agreed**. While this is true for some HKGs, they argue that the definition should not be limited to metabolic genes alone and should consider other types of functional roles.
7. **Unregulated or Constitutively Expressed**:
   * **Pre-assumption**: HKGs are not subject to regulation and are constitutively expressed in all conditions.
   * **Authors’ Stance**: **Disagreed**. The authors demonstrate that many genes traditionally classified as HKGs can be subject to regulation, particularly in stress, development, or pathological states.

## **Human Housekeeping Genes Revisited**

* **Objective**: Re-examines previously identified housekeeping genes (HKGs) and their stability across tissues.
* **Key Findings**:
  1. Identified 3,688 human HKGs with low variability across 16 normal tissues.
  2. Some traditional HKGs (e.g., GAPDH, ACTB) show tissue-specific variations.
  3. HKGs tend to have **shorter introns and compact gene structures**, possibly for transcriptional efficiency.

## State-of-the-Art Housekeeping Proteins for Quantitative Western Blotting: Revisiting the First Draft of the Human Proteome (for cross validation)

* **Objective**: challenge the original criteria and suggest 20 stably-expressed proteins
* **Key Points**:
  + Traditional HKPs, such as GAPDH, β-Tubulin, and β-Actin, have been widely used as loading controls due to their presumed consistent expression across various tissues.
  + However, empirical evidence indicates that the expression levels of these proteins can vary significantly between different tissues and under various experimental conditions, challenging their reliability as universal loading controls.
  + The study analyzed data from ProteomicsDB, encompassing 16,857 liquid chromatography-tandem mass spectrometry datasets from 27 human tissues, to identify more stable and ubiquitously expressed proteins suitable as ILCs.
* **Conclusions**:
  + The authors proposed a list of 20 proteins that exhibited consistent and ubiquitous expression across the analyzed human tissues, suggesting them as more reliable ILCs for quantitative Western blotting.
  + Notably, commonly used HKPs like GAPDH, β-Tubulin, and β-Actin were excluded from this list due to their variable expression patterns.
  + The study underscores the importance of selecting appropriate reference proteins tailored to specific experimental contexts to ensure accurate and reproducible data.

## HRT Atlas v1.0 Database: human vs mouse hkg

* **Refined Definition of HKGs**:
  + The authors redefine HKGs as single constitutive genes expressing at least one protein-coding transcript with non-zero expression and low variability across tissues.
  + They introduce the term "reference transcript" for qPCR normalization, highlighting the importance of transcript-level analysis over gene-level metrics.
* **Data Sources**:
  + The study utilized RNA-seq data from GTEx (11,281 human samples from 52 tissues) and ARCHS4 (507 mouse samples from 14 tissues).
  + Criteria for data inclusion included high sequencing depth, paired-end read libraries, and mRNA enrichment protocols.
* **Improved Detection Criteria**:
  + HKGs must meet strict thresholds:
    - * Non-zero expression (RPKM > 1) in all tissues.
      * Low variability (standard deviation of log2 RPKM < 1).
      * Maximum fold change (MFC) < 2.
  + This approach reduces false positives and identifies HKGs more robustly.
* **Candidate Reference Transcripts**:
  + The authors developed a ranking algorithm to identify the most stable transcripts for tissue-specific qPCR normalization.
  + Reference transcripts were selected based on stability metrics (e.g., mean RPKM, standard deviation, and MFC).

## Housekeeping protein-coding genes interrogated with tissue and individual variations (same as 1 but contain a gene list – for CV)

* **Housekeeping Gene Stability**: While many genes are stably expressed across tissues, this stability does not necessarily imply essentiality.
* **Gene Essentiality**: Essential genes tend to have lower variability in expression (as measured by the Gini coefficient), but not all stably expressed genes are essential.
* **Functional Role**: Housekeeping genes are enriched in cellular maintenance pathways but differ between normal and cancer cells.
* **Evolutionary Conservation**: The study reveals that individual essential and stably expressed genes can vary between species, though the biological pathways they participate in remain conserved.
* **Practical Implications**: A refined list of housekeeping genes across multiple species is provided, which can serve as a more accurate reference for normalization in transcriptomics studies.

## Housekeeping Genes as Internal Standards: Use and Limits (1999)

**Main Focus:**

* This study examines the use of housekeeping genes (HKGs) as internal standards for RNA quantification in experiments.
* The authors discuss limitations in assuming stable expression of common HKGs like GAPDH, β-actin, and 18S rRNA under all conditions.

**Key Findings:**

* Many commonly used HKGs vary significantly across tissues and experimental conditions (e.g., immune stimulation, cell cycle changes).
* Three case studies demonstrate how HKG stability depends on experimental context, warning against using a single HKG for normalization.
* Recommendation: Using multiple HKGs or rRNA (18S/28S) as a reference is more reliable.

## A Compendium of Gene Expression in Normal Human Tissues (2001)

**Main Focus:**

* This study analyzes 7,000 genes across 19 normal human tissues using microarrays to create a reference for baseline gene expression.
* It aims to define tissue-specific and housekeeping genes (HKGs).

**Key Findings:**

* 451 genes were expressed in all 19 tissues, identified as housekeeping genes.
* Expression levels varied significantly across tissues, contradicting the idea that all HKGs have stable expression.
* Even among HKGs, there were distinct tissue-specific patterns, showing that some genes previously considered universal may be biased toward certain tissues.

## Further Understanding Human Disease Genes by Comparing with Housekeeping Genes and Other Genes (2006)

**Main Focus:**

* This study compares disease genes, housekeeping genes (HKGs), and other genes to understand their distinct evolutionary and functional properties.

**Key Findings:**

* Housekeeping genes evolve slower than both disease-related and other genes, suggesting strong purifying selection.
* Housekeeping genes tend to be essential, but not all essential genes are housekeeping genes.
* Disease genes have intermediate levels of evolutionary conservation, differing from both highly essential genes and non-essential genes.
* Some HKGs are involved in disease pathways, showing that gene function is not strictly divided into categories.

## Proteomic Signatures of 16 Major Types of Human Cancer Reveal Universal and Cancer-Type-Specific Proteins for the Identification of Potential Therapeutic Targets (for CV?)

**Main Focus:** important to think of when using as reference: study focused on cancer tissues, but the final list was checked up and ensured that they were stably expressed in both cancer and normal tissue (2384 proteins)

**Key Findings:**

* **Extensive Protein Mapping:** The analysis mapped a total of 8,527 proteins across various cancers, including brain, breast, lung, liver, and others.
* **Tissue-Enriched Proteins:** Identified 2,458 proteins with tissue-enriched expression patterns, providing insights into tissue-specific biology.
* **Therapeutic Targets:** The study identified 1,139 proteins as potential therapeutic targets and observed 21 cancer/testis antigens, which could be relevant for developing targeted therapies.

## Identification of Human Housekeeping Genes and Tissue-Selective Genes by Microarray Meta-Analysis

**Objective:** To identify HKGs and tissue-selective genes by analyzing a large collection of microarray datasets.

**Methodology:**

* Compiled 1,431 samples across 43 normal human tissues from 104 microarray datasets.
* Developed a novel method to improve gene expression assessment, demonstrating that more than ten samples are needed to robustly identify the protein-encoding transcriptome of a tissue.

**Key Findings:**

* Identified 2,064 HKGs and 2,293 tissue-selective genes.
* Functional enrichment analysis revealed that HKGs are mainly involved in fundamental cellular functions, while tissue-selective genes are closely related to functions and diseases specific to their tissue of origin.

## Systematic Identification of Human Housekeeping Genes Possibly Useful as References in Gene Expression Studies (CV)

**Objective:** To systematically identify HKGs that could serve as reliable reference genes in gene expression studies.

**Methodology:**

* Utilized advancements in the quality and completeness of human expression microarray data and their statistical analysis.
* Presented a general framework for choosing reference genes suitable for gene expression studies on normal human tissues and organs.

**Key Findings:**

* Addressed previous assumptions and provided an approach based on the Transcriptome Mapper (TRAM) tool, which overcomes issues associated with cross-platform analysis, such as probe assignment to locus, intra- and inter-sample normalization, and scaled quantile statistics.

## Evidence Based Selection of Housekeeping Genes (for CV)

* **Novel Candidate HKGs:** The study identified new candidate housekeeping genes, such as RPS13, RPL27, RPS20, and OAZ1, which demonstrated enhanced stability across diverse cell types and experimental conditions.
* **Limitations of Common HKGs:** None of the traditionally used housekeeping genes appeared in the top 50 most stable genes identified in this analysis, highlighting their variability and potential unsuitability for normalization purposes.
* **Cross-Species Validation:** The study extended its analysis to 2,543 diverse mouse gene array samples and confirmed the enhanced stability of the novel candidate housekeeping genes in another mammalian species, suggesting their broader applicability.

## Identification of New Reference Genes with Stable Expression Patterns for Gene Expression Studies Using Human Cancer and Normal Cell Lines

**Note:** highly relevant since it involved HPA data

**Objective:** To identify novel reference genes with stable expression across human cancer and normal cell lines, improving gene expression normalization.

**Methodology:**

* 12 candidate reference genes were tested in 13 cancer (e.g., HeLa, MCF-7, A549) and 7 normal cell lines (e.g., HEK293, MRC-5).
* New genes (SNW1, CNOT4) were proposed based on Human Protein Atlas data.

**Key Findings:**

* CNOT4 was highly stable across multiple cell lines.
* Traditional HKGs (e.g., GAPDH, ACTB) showed variability, confirming concerns about their reliability.

## A Comprehensive Functional Analysis of Tissue Specificity of Human Gene Expression (CV)

**Objective:** Identify and analyze housekeeping genes and tissue-specific genes across various human tissues.

**Key Findings:**

* Identified 2,374 housekeeping genes expressed across all examined tissues.
* Housekeeping genes were enriched in vital processes such as oxidative phosphorylation, ubiquitin-dependent proteolysis, translation, and energy metabolism.
* Tissue-specific genes were associated with specialized functions pertinent to each tissue type.