**GO và KEGG**

**1️⃣ Vì sao sau RNA-seq cần phân tích GO và KEGG?**

**RNA-seq cho bạn danh sách các gene khác biệt biểu hiện (DEGs) giữa các nhóm điều kiện (ví dụ tế bào GES-1 nhiễm H. pylori vs control).**

* **Danh sách này thường rất dài (hàng trăm đến hàng nghìn gene).**
* **Nếu chỉ nhìn danh sách gene, khó hiểu bức tranh sinh học tổng thể.**

**👉 GO và KEGG giúp gom nhóm và diễn giải danh sách gene theo chức năng sinh học và đường truyền tín hiệu.**

**2️⃣ GO (Gene Ontology)**

* **Mục tiêu: Gom các gene theo chức năng đã biết.**
* **Ba nhóm chính:**
  + **BP (Biological Process): Các quá trình sinh học (viêm, apoptosis, tín hiệu miễn dịch…).**
  + **MF (Molecular Function): Chức năng phân tử (gắn DNA, hoạt tính kinase, vận chuyển ion…).**
  + **CC (Cellular Component): Vị trí phân tử trong tế bào (màng tế bào, nhân, ti thể…).**
* **Ứng dụng sau RNA-seq:**
  + **Hiểu nhóm gene DEGs tham gia vào quá trình nào.**
  + **Gợi ý cơ chế bệnh học hoặc tác động sinh học của điều kiện thí nghiệm.**
  + **Chọn nhóm gene tiêu biểu để phân tích sâu (qPCR, WB…).**

**3️⃣ KEGG (Kyoto Encyclopedia of Genes and Genomes)**

* **Mục tiêu: Gom gene vào đường truyền tín hiệu (pathway) hoặc con đường chuyển hóa.**
* **Ví dụ: NF-κB signaling, MAPK signaling, apoptosis, metabolism…**
* **Ứng dụng sau RNA-seq:**
  + **Xác định con đường tín hiệu bị kích hoạt hoặc ức chế.**
  + **Liên kết kết quả với các cơ chế bệnh học đã biết (ví dụ: H. pylori kích hoạt NF-κB, MAPK…).**
  + **Dự đoán tác động toàn hệ thống → hỗ trợ lựa chọn thí nghiệm xác nhận (Western blot các protein chủ chốt, luciferase assay…).**

**4️⃣ Ứng dụng tổng thể trong nghiên cứu**

**Sau RNA-seq, GO và KEGG:**

* **📌 Tóm tắt nhanh ý nghĩa sinh học của DEGs.**
* **📌 Ưu tiên hóa các con đường quan trọng để kiểm chứng.**
* **📌 Đề xuất cơ chế mới cho bài báo hoặc đề tài.**
* **📌 So sánh với nghiên cứu trước (ví dụ: xem H. pylori ở Việt Nam có kích hoạt cùng pathway như chủng Đông Á khác không).**
* **📌 Định hướng biomarker hoặc mục tiêu điều trị (targeted therapy, ức chế pathway đặc hiệu).**

**Hai nhóm chính: Over-representation Analysis (ORA) vs Functional Class Scoring (FCS)**

**(A) ORA – Over-Representation Analysis**

* **Nguyên lý:**
  + **Lấy danh sách DEGs (gene thay đổi đáng kể, thường lọc theo |log2FC| & p/FDR).**
  + **So sánh xem các GO terms / pathways nào giàu gene DEGs hơn mong đợi ngẫu nhiên.**
* **Phân tích thống kê: Thường dùng Fisher’s Exact test hoặc Hypergeometric test.**
* **Phần mềm phổ biến:**
  + **clusterProfiler::enrichGO() / enrichKEGG() (R)**
  + **DAVID, Enrichr, WebGestalt**
* **Đầu vào: Danh sách gene DEGs + tập gene nền (background / universe).**
* **Ưu điểm: Nhanh, dễ, phù hợp khi số DEGs rõ ràng.**
* **Nhược điểm: Bỏ qua thông tin của các gene không đạt ngưỡng DE nhưng có xu hướng đồng biến đổi.**

**(B) FCS – Functional Class Scoring**

* **Tiêu biểu: GSEA (Gene Set Enrichment Analysis)**
* **Nguyên lý:**
  + **Không lọc ra DEGs.**
  + **Dùng toàn bộ danh sách gene được xếp hạng theo mức độ thay đổi (log2FC, stat score…).**
  + **Xem các gene thuộc GO term/pathway tập trung ở đầu hoặc cuối danh sách hay phân tán ngẫu nhiên.**
* **Thống kê: Kolmogorov–Smirnov-like statistic → Normalized Enrichment Score (NES).**
* **Phần mềm phổ biến:**
  + **GSEA desktop (Broad Institute)**
  + **clusterProfiler::GSEA()**
* **Ưu điểm: Không phụ thuộc ngưỡng cắt DEGs, nhạy với thay đổi nhẹ nhưng đồng hướng.**
* **Nhược điểm: Phức tạp hơn, cần dữ liệu đầy đủ và sắp xếp.**

**2️⃣ Các biến thể và phương pháp nâng cao**

**Ngoài ORA và GSEA, có một số biến thể:**

* **Camera / Roast (limma): Kết hợp tính toán tương quan giữa các gene.**
* **GSVA (Gene Set Variation Analysis): Xác định mức hoạt động pathway từng mẫu (không chỉ giữa 2 nhóm).**
* **Pathview + KEGGprofile: Visual hóa pathway KEGG kèm log2FC từng gene.**

**3️⃣ Ứng dụng thực tế**

* **ORA: Khi bạn có danh sách DEGs rõ ràng → dùng để tìm GO/KEGG quan trọng nhất.**
* **GSEA: Khi số gene DEGs ít, hoặc muốn khai thác tín hiệu pathway tiềm ẩn (gene thay đổi nhẹ nhưng đồng hướng).**
* **GSVA: Khi muốn so sánh mức độ hoạt hóa pathway giữa các mẫu/lô (ví dụ từng mẫu GES-1 nhiễm các chủng H. pylori khác nhau).**

**1️⃣ Gene Ontology (GO)**

**a. Định nghĩa Gene Ontology (GO)**

* **Gene Ontology (GO)** là một hệ thống phân loại và mô tả chuẩn hóa về **chức năng của gen và sản phẩm gen** (mRNA, protein).
* GO ra đời nhằm **thống nhất cách mô tả chức năng gen giữa các loài** và tạo điều kiện cho phân tích dữ liệu sinh học quy mô lớn.
* GO được phát triển bởi **Gene Ontology Consortium (GOC)** và duy trì cập nhật liên tục.

**b. Ba nhánh chính của GO**

GO chia mọi chức năng gen thành **3 ontology chính**:

1. **Biological Process (BP)**
   * Mô tả **quá trình sinh học** mà gen tham gia.
   * Thường là một chuỗi các sự kiện có trật tự liên quan đến nhiều thành phần phân tử.
   * Ví dụ:
     + *Apoptotic process* (GO:0006915)
     + *Cell proliferation* (GO:0008283)
     + *Inflammatory response* (GO:0006954)
2. **Molecular Function (MF)**
   * Mô tả **hoạt động phân tử** của sản phẩm gen (protein hoặc RNA).
   * Ví dụ:
     + *ATP binding* (GO:0005524)
     + *DNA-binding transcription factor activity* (GO:0003700)
     + *Protein kinase activity* (GO:0004672)
3. **Cellular Component (CC)**
   * Mô tả **vị trí/cấu trúc** trong tế bào nơi sản phẩm gen tồn tại hoặc hoạt động.
   * Ví dụ:
     + *Nucleus* (GO:0005634)
     + *Cytosol* (GO:0005829)
     + *Plasma membrane* (GO:0005886)

**c. Cấu trúc của GO**

* GO có dạng **cây phân cấp Directed Acyclic Graph (DAG)**:
  + Các **GO term** có quan hệ *cha–con* (parent–child).
  + Ví dụ: *MAPK cascade* là con của *signal transduction*.
* Mỗi GO term có:
  + **Tên (Name)** – mô tả ngắn gọn
  + **GO ID** – mã duy nhất (ví dụ: GO:0000165)
  + **Định nghĩa (Definition)** – mô tả chi tiết
  + **Quan hệ (Relationships)** – *is\_a*, *part\_of*, *regulates*...

**d. Ứng dụng của GO trong phân tích dữ liệu**

* **Chú giải gen**: Liệt kê các chức năng, vị trí, quá trình của gen trong dữ liệu RNA-seq.
* **Phân tích enrichment (GO enrichment)**:
  + Đầu vào: danh sách gen quan tâm (ví dụ gen DE trong RNA-seq)
  + So sánh với toàn bộ gen nền (background)
  + Kết quả: GO terms nào **giàu (enriched)** trong danh sách gen hơn mức kỳ vọng thống kê.
  + Ví dụ: Nếu nhóm gen tăng biểu hiện giàu GO term *inflammatory response* (GO:0006954) → gợi ý đáp ứng viêm.

**🔍 1. Một gene có thể gán vào rất nhiều GO term**

* Mỗi GO term mô tả một khía cạnh riêng: **BP**, **MF**, **CC**.
* Một gene có nhiều chức năng hoặc tham gia nhiều quá trình → được gán vào nhiều term.
* Ví dụ: **TP53**
  + BP: *DNA damage response*, *apoptotic process*, *cell cycle arrest*…
  + MF: *DNA binding*, *transcription factor binding*, *protein kinase regulator activity*…
  + CC: *nucleus*, *cytosol*, *chromatin*…  
    → Riêng TP53 có thể nằm trong **hàng chục GO term**.

**🔍 2. Số gene người có trong GO annotation**

* Theo **Gene Ontology Annotation (GOA) for Homo sapiens**, số gene được gán GO terms:
  + **~19.500 gene protein‑coding** (trong Ensembl hoặc UniProtKB)
  + Kèm một số lượng nhỏ non‑coding RNA cũng có annotation (ncRNA, miRNA host…).
* Tức là **hầu hết gene người đều có ít nhất 1 GO term**.

**🔍 3. Vì sao có tới ~18.000 GO term**

* Tổng số GO term toàn hệ thống: ~44.000 (cho mọi loài).
* Chỉ **~17.900–18.000 term thực sự được sử dụng để annotate gene người**.
* Mỗi term có thể chứa từ vài gene → hàng ngàn gene.  
  Và **1 gene nằm trong nhiều term**, nên tổng số association (gene–term pairs) rất lớn (hàng trăm nghìn).

**🔍 4. Tra số gene người trong GO terms**

* Xem **GOA Human file** tại EBI QuickGO → hiện có > 600.000 annotation (gene–term pairs).
* Nghĩa là **~19.000 gene** người được phân loại vào **~18.000 term**, tạo ra **~600.000 mối quan hệ**.

💡 Tóm lại:

* **Số gene người có annotation GO**: ~19.000.
* **Số GO term áp dụng cho người**: ~18.000.
* **Một gene thường nằm trong nhiều GO term**, nên số lượng association rất lớn.

Trang chủ của **Gene Ontology (GO)** là:

🌐 [**http://geneontology.org**](http://geneontology.org)

Từ đây bạn có thể truy cập toàn bộ thông tin GO terms, annotation, và tìm gene.

**🔍 1. Tra một gene nằm trong các GO term nào**

Cách dễ nhất là dùng **AmiGO** hoặc **QuickGO**:

* **AmiGO 2** (trang chính thức của GO):  
  🔗 http://amigo.geneontology.org/amigo  
  👉 Gõ tên gene (ví dụ *TP53*, *EGFR*, hoặc Ensembl ID) → hiện tất cả GO terms mà gene được gán, kèm loại (BP, MF, CC) và bằng chứng.
* **QuickGO** (EBI):  
  🔗 https://www.ebi.ac.uk/QuickGO/  
  👉 Gõ tên gene hoặc UniProt ID → xem danh sách GO term kèm link sang term chi tiết.

**🔍 2. Tra một GO term xem có gene nào trong đó**

* Vào **AmiGO 2** → tìm GO term (ví dụ *GO:0008284 positive regulation of cell proliferation*).  
  Trong phần **Annotations** sẽ liệt kê tất cả gene được gán vào term đó, lọc theo loài (*Homo sapiens*).
* Hoặc **QuickGO**:
  + Gõ ID GO (ví dụ GO:0008284).
  + Chọn tab **Gene products** → lọc *Homo sapiens*.
  + Sẽ thấy danh sách gene trong term này.

💡 **Tip nhanh:**  
Nếu muốn lấy danh sách lớn (ví dụ tất cả GO terms và gene người trong đó) → tải file **GOA Human**:  
🔗 GOA Human GAF file  
File này chứa toàn bộ mapping giữa GO term ↔ gene người.

**2️⃣ Kyoto Encyclopedia of Genes and Genomes (KEGG)**

**a. Định nghĩa KEGG**

* **KEGG** là một cơ sở dữ liệu lớn về **mạng sinh học và con đường sinh học (pathways)**, được phát triển bởi **Kanehisa Laboratories**.
* KEGG tập trung vào việc mô tả **mối quan hệ giữa gen/protein và các chức năng sinh học** thông qua:
  + **Pathway maps** (bản đồ con đường)
  + **Mạng tương tác phân tử**
  + **Các module chức năng**

**b. Nội dung chính của KEGG**

KEGG bao gồm nhiều phần, quan trọng nhất cho phân tích RNA-seq là:

1. **KEGG Pathway**
   * Các **bản đồ con đường** mô tả các phản ứng sinh hóa, tín hiệu, chuyển hóa...
   * Mỗi con đường có **ID riêng** (ví dụ: hsa04010 cho *MAPK signaling pathway* ở người).
2. **KEGG BRITE**
   * Hệ thống phân loại phân cấp (hierarchical classification) về protein, hợp chất, bệnh lý, thuốc...
3. **KEGG Modules**
   * Các khối chức năng sinh học cụ thể (ví dụ: một phần của con đường chuyển hóa).

**c. Ứng dụng KEGG trong nghiên cứu**

* **Phân tích enrichment KEGG** (KEGG enrichment):
  + Đầu vào: danh sách gen DE (hoặc tất cả gen kèm mức biểu hiện)
  + So sánh xem con đường nào có **nhiều gen DE hơn mức ngẫu nhiên**
  + Kết quả: Xác định **các con đường sinh học bị ảnh hưởng**
  + Ví dụ: Trong H. pylori nhiễm GES‑1, KEGG thường thấy *MAPK signaling*, *NF‑κB pathway*, *Gastric acid secretion*...

**d. So sánh GO và KEGG**

| **Tiêu chí** | **GO** | **KEGG** |
| --- | --- | --- |
| Mục tiêu | Mô tả chức năng gen (BP, MF, CC) | Mô tả con đường sinh học và mạng |
| Dạng dữ liệu | Các **GO terms** | Các **pathway maps** |
| Đầu ra phân tích | Danh sách GO enriched | Danh sách pathways enriched |
| Ứng dụng | Chú giải chức năng, phân tích enrichment | Phân tích con đường, tác động sinh học |

**1. Hiện nay KEGG có bao nhiêu Con đường cho Người?**

* Theo thống kê cập nhật từ KEGG tính đến **ngày 28–30 tháng 7 năm 2025**, số bản đồ **KEGG PATHWAY** (tham khảo toàn cầu và cho từng loài) là **580**. Trong đó, bản đồ dành riêng cho **Homo sapiens (prefix hsa) là khoảng 280 đường** [KEGG+12KEGG+12BioMed Central+12](https://www.genome.jp/kegg/docs/statistics.html?utm_source=chatgpt.com).
* Một cộng đồng báo cáo cụ thể cho đường dành riêng cho người là **283 bản đồ hsa** (human-specific pathway maps) [Biostars](https://www.biostars.org/p/9080/?utm_source=chatgpt.com).

📌 **Tóm tắt**:

* Tổng số **khoảng 580 KEGG pathway maps**,
* Trong đó **~280–283 là bản đồ dành riêng cho người (hsa)**.

**🧬 2. Danh sách các gen trong pathway có theo phiên bản không? Hiện có bao nhiêu?**

* Mỗi pathway dành cho người (hsaXXXXX) được xây dựng bằng cách **chuyển đổi nhóm orthologs (KO) sang gene Entrez cụ thể của người**. Do đó, **danh sách gene tương ứng cho mỗi con đường được cập nhật theo phiên bản KEGG mới nhất** [KEGG](https://www.kegg.jp/kegg/download/Readme/README.pathway?utm_source=chatgpt.com)[KEGG](https://www.kegg.jp/kegg-bin/show_organism?menu_type=pathway_maps&org=hsa&utm_source=chatgpt.com).
* KEGG cập nhật thường xuyên: tính đến phiên bản tháng 7‑2025, số gene nhân bản trong các loài là hàng trăm nghìn, và **Entrez gene map hiện đại bao gồm cả gene protein‑coding, enzyme, gene liên quan**. Ước tính là **~27,700 KO groups** và tương ứng hàng chục ngàn gene mỗi loài [KEGG](https://www.genome.jp/kegg/docs/statistics.html?utm_source=chatgpt.com)[rdrr.io](https://rdrr.io/bioc/mirIntegrator/man/kegg_pathways.html?utm_source=chatgpt.com).

📌 **Tóm tắt**:

* Danh sách gene theo pathway là **phiên bản theo ngày cập nhật KEGG**, có thể khác giữa các bản (ví dụ hsa00010 năm 2020 vs 2025).
* Ví dụ: pathway “Glycolysis” giờ có thể đổi danh sách gene khi cơ sở KEGG cập nhật chính thức.

**🔗 3. Mối quan hệ với phân tích enrich: vì sao số gene thực tế nhỏ?**

Dựa trên câu trước bạn hỏi:

* **Dữ liệu input** của bạn có ~6467 gene hợp lệ.
* Nhưng chỉ có ~2626 gene được dùng trong phân tích KEGG vì:
  + Chỉ gene nằm trong **khoảng 280–283 KEGG pathways** đã được annotate.
  + Gene nằm trong pathway quá nhỏ ( <10 gene) hoặc quá lớn (>500 gene) bị loại theo cấu hình.

**✅ Kết luận nhanh:**

| **Tiêu chí** | **Giá trị** |
| --- | --- |
| Tổng số KEGG pathway maps | 580 maps |
| Bản đồ dành cho người (hsa) | ~280–283 maps |
| Gene total mapped trong database Kegg human | hàng chục ngàn, phụ thuộc phiên bản cập nhật |
| Các gene đủ điều kiện phân tích enrich | ví dụ 2626 trong trường hợp của bạn |
|  |  |