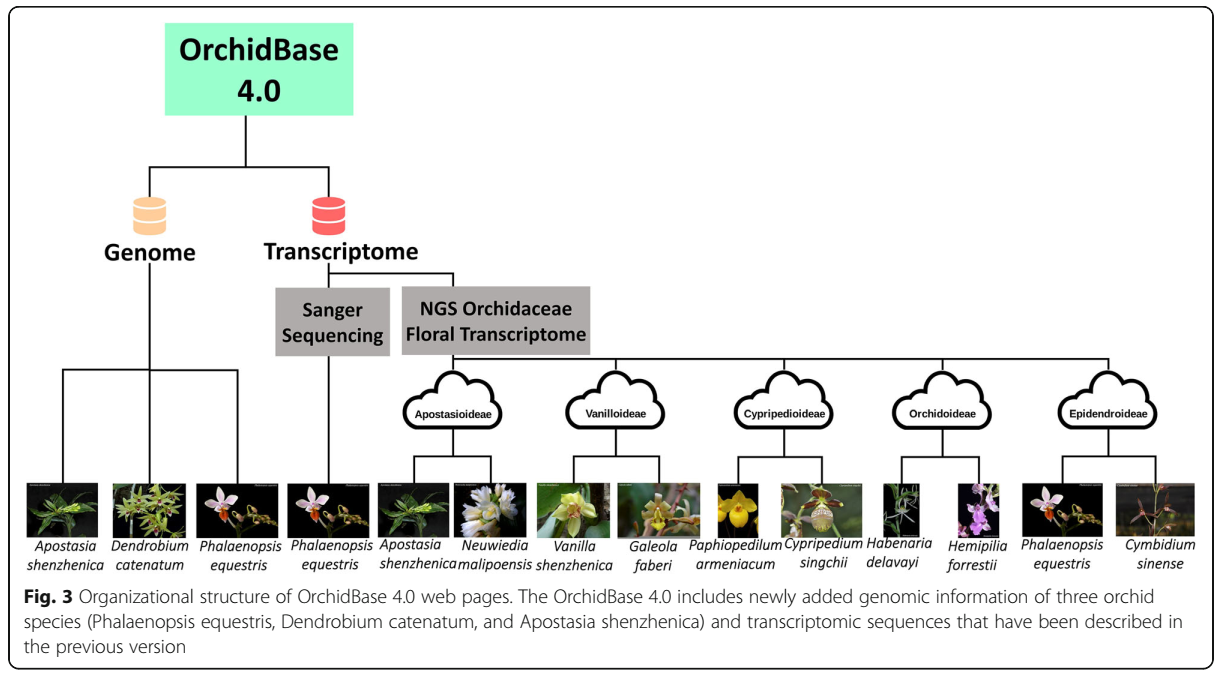
Cơ sở dữ liệu số Orchidbase là một cơ sở dữ liệu số về lan Hồ điệp (Phalaenopsis). Đây là cơ sở dữ liệu các trình tự EST dựa trên web, cung cấp thông tin tích hợp về EST từ lan hồ điệp. OrchidBase hiện chứa 84.617 trình tự được phiên mã, trong đó có 76.116 đơn lẻ. Bên cạnh cách tiếp cận chung là cung cấp kho dữ liệu của EST, OrchidBase cũng được tích hợp thông tin toàn diện, bao gồm các cụm thông tin, các chú thích, bản thể gen (GO) và các con đường trao đổi chất liên quan dựa trên các tìm kiếm tương tự BLAST. Thông tin có thể được truy xuất bằng tìm kiếm văn bản, công cụ truy vấn hoặc tìm kiếm bằng công cụ BLAST (Fu và cộng sự, 2011).

Phần mềm cơ sở dữ liệu OrchidBase bao gồm một hệ thống cơ sở dữ liệu hợp tác, một ứng dụng Windows thực hiện phân tích trình tự, một ứng dụng web tự động thực thi nhiều truy vấn cơ sở dữ liệu và thực hiện một yêu cầu bằng các Dịch vụ Web giao thức truy cập đối tượng đơn giản. OrchidBase được thiết kế để lưu trữ và khám phá số lượng lớn EST, là các thông tin sinh học phức tạp. Có thể truy cập và tìm kiếm thông tin trình tự và chú thích tương ứng trong mỗi thư viện được xây dựng riêng biệt từ các thư viện cDNA khác nhau thông qua ứng dụng web (Fu và cộng sự, 2011)..

Từ phần mềm cơ sở dữ liệu Orchidbase ban đầu vào năm 2011 chỉ với việc lưu trữ và phân tích các trình tự EST trên loài lan Hồ điệp, hiện nay đã phát triển thành phần mềm Orchidbase 4.0 (Hsiao và cộng sự, 2021) với nhiều thông tin khác như biểu hiện gen, các chú thích gen, các con đường trao đổi chất cụ thể và được sơ đồ hóa, . . .trong đó gồm hai cấu phần chính là genome và transcriptome trên nhiều đối tượng hoa lan khác nhau (Hình 1).



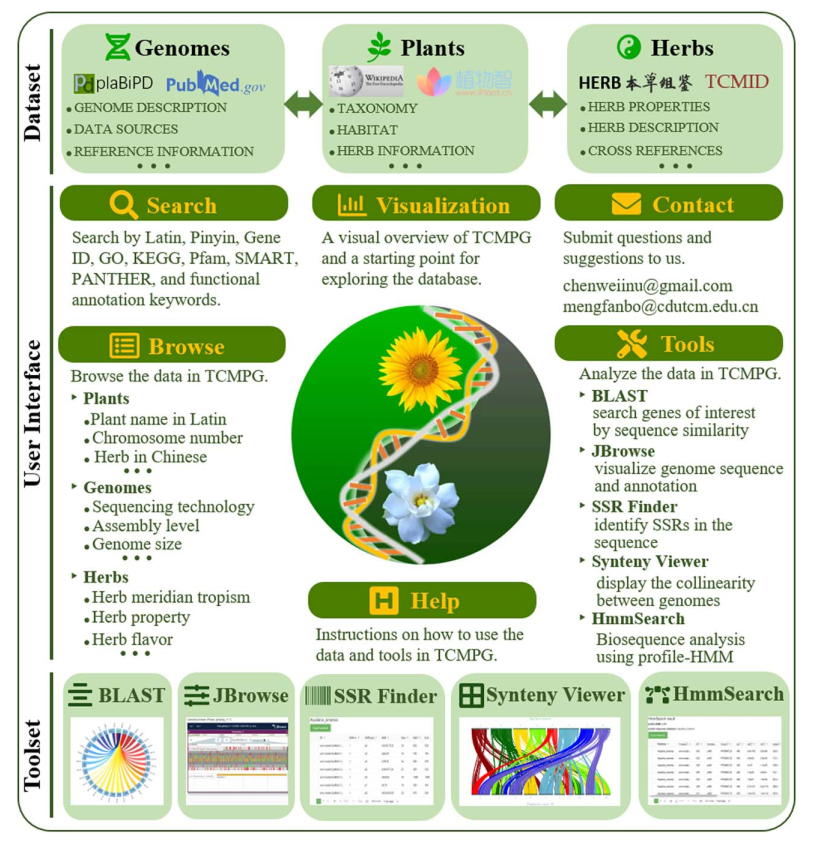
Đây là cơ sở dữ liệu số về các thông tin di truyền ở mức độ phân tử có thể tiến hành truy xuất, phân tích, so sánh với các cơ sở dữ liệu khác.

# Cơ sở dữ liệu TCMPG (Traditional Chinese medicine Plant Genome database)

là một cơ sở dữ liệu số tích hợp để lưu trữ các bộ gen rải rác của cây thuốc. TCMPG hiện bao gồm 160 cây thuốc, 195 bộ gen tương ứng và 255 loại thuốc thảo dược. Thông tin chi tiết về các loài thực vật, bộ gen và thuốc thảo dược cũng được tích hợp vào TCMPG. Các công cụ phân tích bộ gen phổ biến được nhúng trong TCMPG để tạo thuận lợi cho việc phân tích có hệ thống các cây thuốc. Chúng bao gồm BLAST để xác định các chỉnh hình từ các cây khác nhau, SSR Finder để xác định các lần lặp lại trình tự đơn giản, JBrowse để duyệt các bộ gen, Synteny Viewer để hiển thị các khối tổng hợp giữa hai bộ gen và HmmSearch để xác định các miền protein. TCMPG sẽ được cập nhật liên tục bằng cách tích hợp dữ liệu và công cụ mới để phân tích bộ gen so sánh và chức năng (Meng và cộng sự, 2022).

Trong TCMPG, dữ liệu bộ gen của một cây thuốc bao gồm bộ gen tham chiếu và định dạng tính năng chung (GFF3), trình tự mã hóa (CDS) và tệp trình tự protein (PEP). Trong đó, dữ liệu bộ gen cây thuốc có sẵn được xác định thông qua bốn bước. Trong bước đầu tiên, các tên và bài viết bằng tiếng Latinh được thu thập về tất cả các loài thực vật đã được xuất bản, được giải trình tự từ plaBiPD (//www.plabipd.de/index.ep). Trong bước thứ hai, các thông tin về các loại thảo mộc từ TCMID và HERB được tải xuống. Trong bước thứ ba, tên Latin của các loài thực vật được sử dụng làm tham chiếu để thu được giao điểm của dữ liệu thu được trong hai bước trên. Trong bước thứ tư, dữ liệu bộ gen được xác định bằng cách quản lý thủ công dựa trên các liên kết được đưa ra trong bài báo. Nếu có nhiều phiên bản cho một cây thuốc, tất cả các phiên bản đã được thu thập. Theo đó, đã thu được 195 bộ dữ liệu gen chất lượng cao của 160 cây thuốc.

Trong TCMPG, để làm phong phú dữ liệu liên quan đến các loài thực vật, các loài cây thuốc được kiểm tra phân loại của từng loài tại Wikipedia và tải xuống hình ảnh đại diện của hầu hết các loài. Ngoài ra, số lượng nhiễm sắc thể của mỗi loài được lấy từ CCDB, thông tin sự phân bố của hầu hết các loài ở Trung Quốc được thu thập từ iPlant ([www.iplant.cn](http://www.iplant.cn)). Ngoài ra, chi tiết từng bài báo về bộ gen, bao gồm tiêu đề, ngày xuất bản, tạp chí và PMID được ghi lại. Đối với mỗi bộ gen, thông tin từ các bài báo được thu thập theo cách thủ công, bao gồm kích thước bộ gen, mức độ lắp ráp và số lượng gen dự đoán. Các thông tin cơ bản và chức năng của cơ sở dữ liệu TCMPG được mô tả qua hình 2:



Hình ảnh cho thấy các trường dữ liệu khác nhau được thiết thập trong TCMPG từ phân loại, vùng sinh sống, thông tin dược liệu, đặc tính dược chất, các tính chất bộ gen, … được mô tả cụ thể và có thể sử dụng và truy xuất.

Từ ví dụ về hai cơ sở dữ liệu đã được xây dựng thành công này, việc xây dựng một cơ sở dữ liệu số về nguồn gen lan rừng khu vực Tây Nam Bộ, Tây Nguyên và Nam Trung Bộ với các thông tin về hình ảnh cây, hoa (nếu có), các đặc tính hình thái, vùng phân bố, các dữ liệu về sinh trưởng trong điều kiện bảo quản chuyển vị (ex situ) trong nhà lưới và in vitro, các thông tin di truyền về trình tự DNA của từng mẫu thu thập cũng như tình trạng sống của mẫu thu thập là khả thi và có thể được ứng dụng cho mục tiêu bảo tồn và khai thác một cách hiệu quả nguồn gen này đồng thời có thể dễ dàng chia sẻ thông tin với các đơn vị khác một cách thuận lợi.