CGAS（叶绿体基因组分析套件）：面向比较叶绿体基因组学的自动化 Python 流程

Abdullah[1,2*], 闫如山[1,2], 田晓轩[1,2*]

[1]天津中医药大学现代中药创制全国重点实验室

[2]现代中医药海河实验室

Abdullah, Rushan Yan, Xiaoxuan Tian. 2026. CGAS (Chloroplast Genome Analysis Suite): An Automated Python Pipeline for Comprehensive Comparative Chloroplast Genomics. ***iMetaOmics*** 3: e70093. https://doi.org/10.1002/imo2.70093

挑战

叶绿体基因组片段分析工作流程

Multiple Tools

Fragmented workflow requiring many independent tools

Manual Formatting

Complex data formatting and conversion steps

Complex Pipelines

Significant bioinformatics expertise required

耗时·难以复制·技术壁垒高

亮点

Chloroplast Genome Analysis Suite (CGAS)

Integrated Platform for Comparative Chloroplast Genomics

Inputs

Raw Sequencing Reads

Assembled Chloroplast Genomes

Single or Multiple Species

CGAS

Genome Processing

- Raw Reads Quality Analysis
- Chloroplast Genome Assembly
- Coverage depth analysis
- Annotation
- Annotations Standardization

Comparative Chloroplast Genomics

- Gene Content & Structural Variation
- SNP detection
- Intron Analysis with Complete Statistics
- Nucleotide Diversity (π) Analysis
- SSRs Detection with Genomic & Functional Localization
- Codon Usage & Amino Acid Bias

Phylogenomics

- Feature-Level Extraction
- Concatenated Matrix Construction
- Phylogenetic Inference

Outputs

Submission-Ready Files

Comparative Feature Matrices

Publication-Ready Tables

High-Resolution Figures

Phylogenetic Trees

CGAS模块组成

三个主要分析阶段

1

Phase 1: Preparation

Genome assembly, Quality control,
Annotations and NCBI submission

2

Phase 2: Comparative Analysis

Comprehensive
chloroplast genomics

3

Phase 3: Phylogenetic Analysis

Evolutionary
relationships

阶段一：准备模块

Chloroplast Genome Analysis Suite (CGAS)

Unified Pipeline for Comparative Chloroplast Genomics

Input Data

Modules 1-4 (Preparation)
• FASTQ files (raw sequencing reads) • Paired-end/Single-end

Modules 5-14 (Analysis & Phylogeny)
• Annotated GenBank files (.gb, .gbk) • Multiple genomes (batch processing)

Phase 1: Preparation Modules (Run Individually)

1 Assembly & QC
- Quality Control (fastp)
- Adapter Trimming
- Genome Assembly (GetOrganelle)
- Coverage Analysis

Output: Assembled genomes

2 Annotation
- Plastome Gene Annotation
- PGA Pipeline
- Reference-guided Annotation

Output: GenBank files

3 Gene Comparison
- Normalize Gene Names
- Standardize Annotations
- Cross-genome Consistency

Output: Normalized GenBank

4 Format Conversion
- Convert to NCBI Format
- Generate FASTA + TBL files
- Submission-ready

Output: NCBI submission files

Output: Assembled genomes → GenBank files → Normalized annotations → Annotation validation and NCBI submission files

阶段二: 比较基因组分析模块

Phase 2: Main Analysis Modules

(Run as Pipeline: cgas 5,6,7,8,9,10,11,12,13)

5 Gene Comparative Analysis

Compare gene content across species
Identify conserved and unique genes
Gene presence/absence matrix

Output: Comparative tables

6 Gene Content Tables

Publication-ready Word documents
Formatted gene tables
Journal-ready format

Output: .docx tables

7 Genome Structure

LSC/SSC/IR region analysis
Functional gene classification
GC content distribution

Output: Structure data

8 Codon Usage (RSCU)

Relative synonymous codon usage
Codon bias patterns
R-based visualization

Output: RSCU data + plots

9 Amino Acid Analysis

Amino acid composition
Patterns across genomes
R-based visualization

Output: AA data + plots

10 SNP Analysis

Single nucleotide polymorphisms
Substitution detection
Quality visualization

Output: SNP data + plots

11 Intron Analysis

Gene and tRNA intron detection
Intron-exon structure
Comprehensive comparison

Output: Intron data

12 SSR Analysis

Simple sequence repeat detection
Genomic/functional region mapping
High-quality visualization

Output: SSR data + plots

13 Nucleotide Diversity

Pi (π) diversity assessment
Region-wise analysis
High-quality visualization

Output: Diversity plots

Pipeline Command: cgas --modules 5,6,7,8,9,10,11,12,13
All 9 modules run together for comprehensive comparative analysis

Note: Modules 8, 9, 10, 12, 13 include R-based visualizations (optional, requires R installation)

阶段三: 系统发育分析模块

14

Module 14: Phylogenetic Analysis (Run Separately)

Phylogenetic matrix construction • MACSE/MAFFT alignment • IQ-TREE inference

Outgroup support • Complete genome or gene-specific phylogenies

Output: Phylogenetic trees (.treefile) + alignment matrices

Preparation Analysis Phylogeny

生成可直接发表的结果

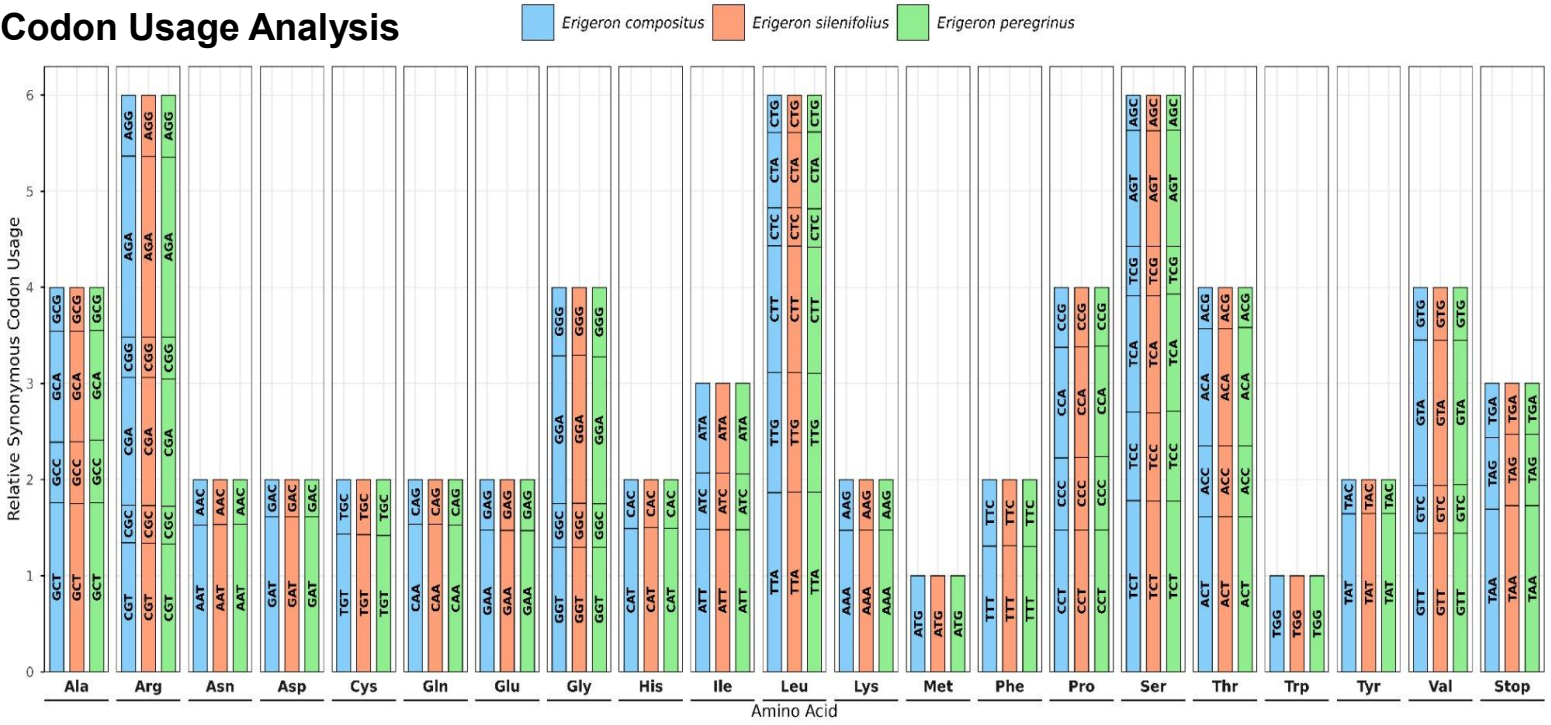
Structured Tables

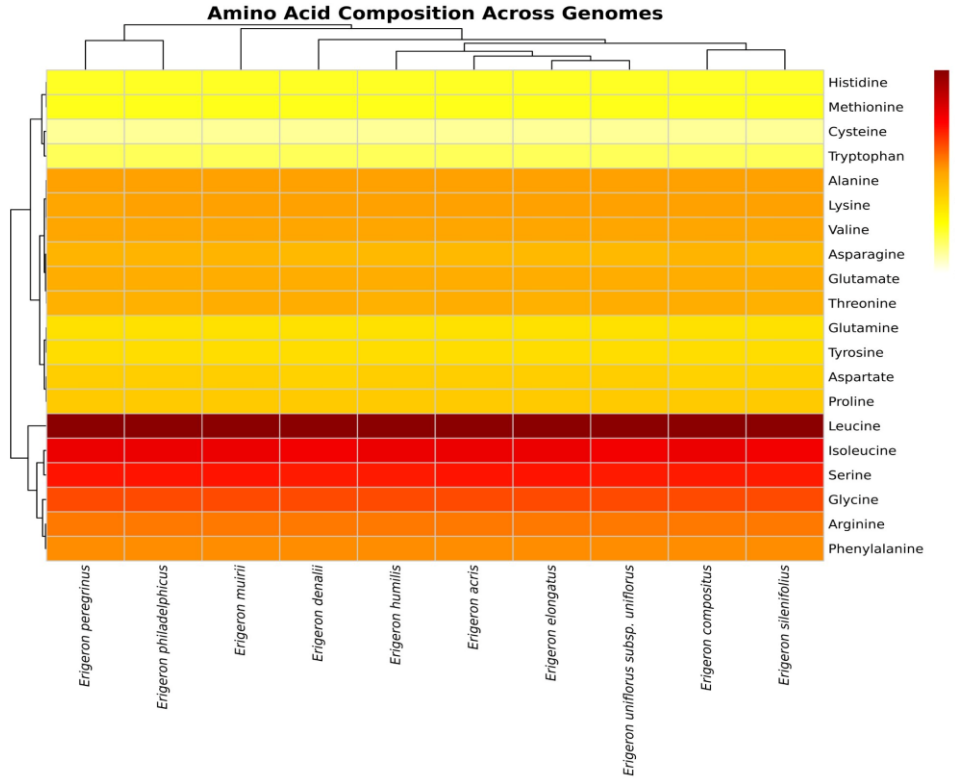
Gene content
Comparative analysis
Statistical summaries

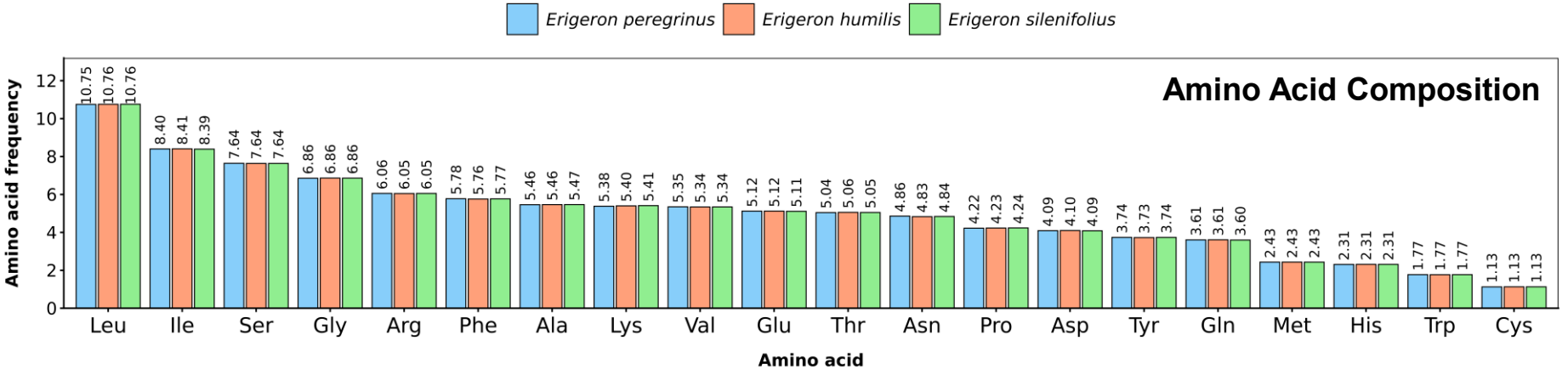
Visualizations

SNP plots
SSR distributions
Diversity analyses

部分结果示例


Codon Usage Analysis
Erigeron compositus
Erigeron silenifolius
Erigeron peregrinus
Relative Synonymous Codon Usage
Amino Acid
Ala
Arg
Asn
Asp
Cys
Gln
Glu
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Ser
Thr
Trp
Tyr
Val
Stop


Amino Acid Composition Across Genomes
Histidine
Methionine
Cysteine
Tryptophan
Alanine
Lysine
Valine
Asparagine
Glutamate
Threonine
Glutamine
Tyrosine
Aspartate
Proline
Leucine
Isoleucine
Serine
Glycine
Arginine
Phenylalanine
Erigeron peregrinus
Erigeron philadelphicus
Erigeron muirii
Erigeron denalii
Erigeron humilis
Erigeron acris
Erigeron elongatus
Erigeron uniflorus subsp. uniflorus
Erigeron compositus
Erigeron silenifolius


Erigeron peregrinus
Erigeron humilis
Erigeron silenifolius
Amino Acid Composition
Amino acid frequency
Amino acid
Leu 10.75 10.76 10.76
Ile 8.40 8.41 8.39
Ser 7.64 7.64 7.64
Gly 6.86 6.86 6.86
Arg 6.06 6.05 6.05
Phe 5.78 5.76 5.77
Ala 5.46 5.46 5.47
Lys 5.38 5.40 5.41
Val 5.35 5.34 5.34
Glu 5.12 5.12 5.11
Thr 5.04 5.06 5.05
Asn 4.86 4.83 4.84
Pro 4.22 4.23 4.24
Asp 4.09 4.10 4.09
Tyr 3.74 3.73 3.74
Gln 3.61 3.61 3.60
Met 2.43 2.43 2.43
His 2.31 2.31 2.31
Trp 1.77 1.77 1.77
Cys 1.13 1.13 1.13

总结

简化叶绿体比较基因组学分析

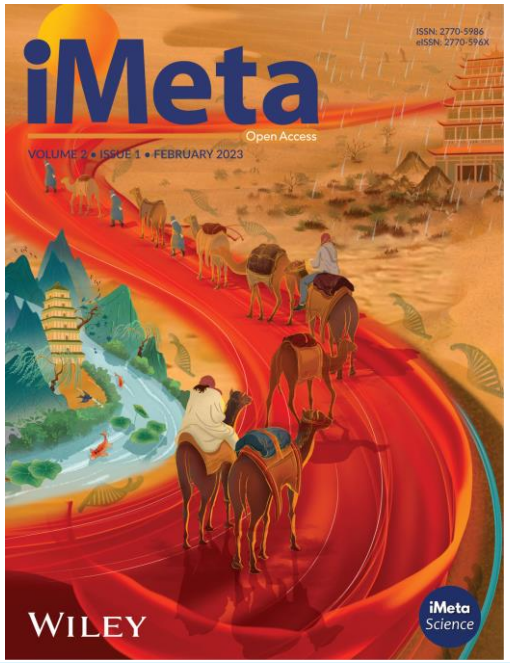
加速叶绿体基因组学研究

- ❑ 集成 14 个模块
- ❑ 生成可直接用于发表的结果
- ❑ 自动化分析流程
- ❑ 可重复的研究流程
- ❑ **GitHub:** github.com/abdullah30/Chloroplast-Genome-Analysis-Suite-CGAS

Abdullah, Rushan Yan, Xiaoxuan Tian. 2026. CGAS (Chloroplast Genome Analysis Suite): An Automated Python Pipeline for Comprehensive Comparative Chloroplast Genomics. ***iMetaOmics*** 3: e70093. https://doi.org/10.1002/imo2.70093

iMeta(宏): 整合宏组学重新认识生物和医学

iMeta WILEY



iMeta*(宏)**期刊是由宏科学、千名华人科学家和威立共同出版，对标Cell***的生物/医学类综合期刊，主编刘双江和傅静远教授，欢迎高影响力的研究、方法和综述投稿，重点关注生物技术、大数据和组学等前沿交叉学科。已被SCIE、PubMed等收录，最新IF 33.2，位列全球SCI期刊第65位(前千分之三)，中国第5位，微生物学研究类全球第一，中科院生物学双1区Top。外审平均21天，投稿至发表中位数87天。
子刊***iMetaOmics*** (宏组学)、***iMetaMed*** (宏医学)定位IF>10和15的生物、医学综合期刊，欢迎投稿!

主页: http://www.imeta.science
出版社: https://wileyonlinelibrary.com/journal/imeta

iMeta: https://wiley.atyponrex.com/journal/IMT2
投稿: iMetaOmics: https://wiley.atyponrex.com/journal/IMO2
iMetaMed: https://wiley.atyponrex.com/journal/IMM3

office@imeta.science
imetaomics@imeta.science

宣传片

iMeta

更新日期
2025/7/6