Anotação comparativa do gênero *Pinus* com o gênero *Eucalyptus*

Daiane Rigoni

Orientador: Prof. Dr. Dieval Guizelini

Introdução

Setor de Florestas Plantadas: desenvolvimento econômico, social e ambiental;

2023: 10,14 milhões de ha (IBÁ, 2024);

Eucalipto: 7,7 milhões de hectares (~76%);

Pinus: 1,6 milhão de hectares (~16%);

Brasil: líder mundial em produtividade florestal;

1,2% PIB;

Sequestro de CO_2, conservação (APP e RL).

Introdução

- Materiais genéticos superiores: agregar valor;
- Crescente disponibilidade de sequências;
- Desafio: anotação;
- Ferramentas genômicas: genes responsáveis por características de interesse: crescimento, formação da madeira, teor nutritivo, resistência a doenças, tolerância à estresses bióticos e abióticos.

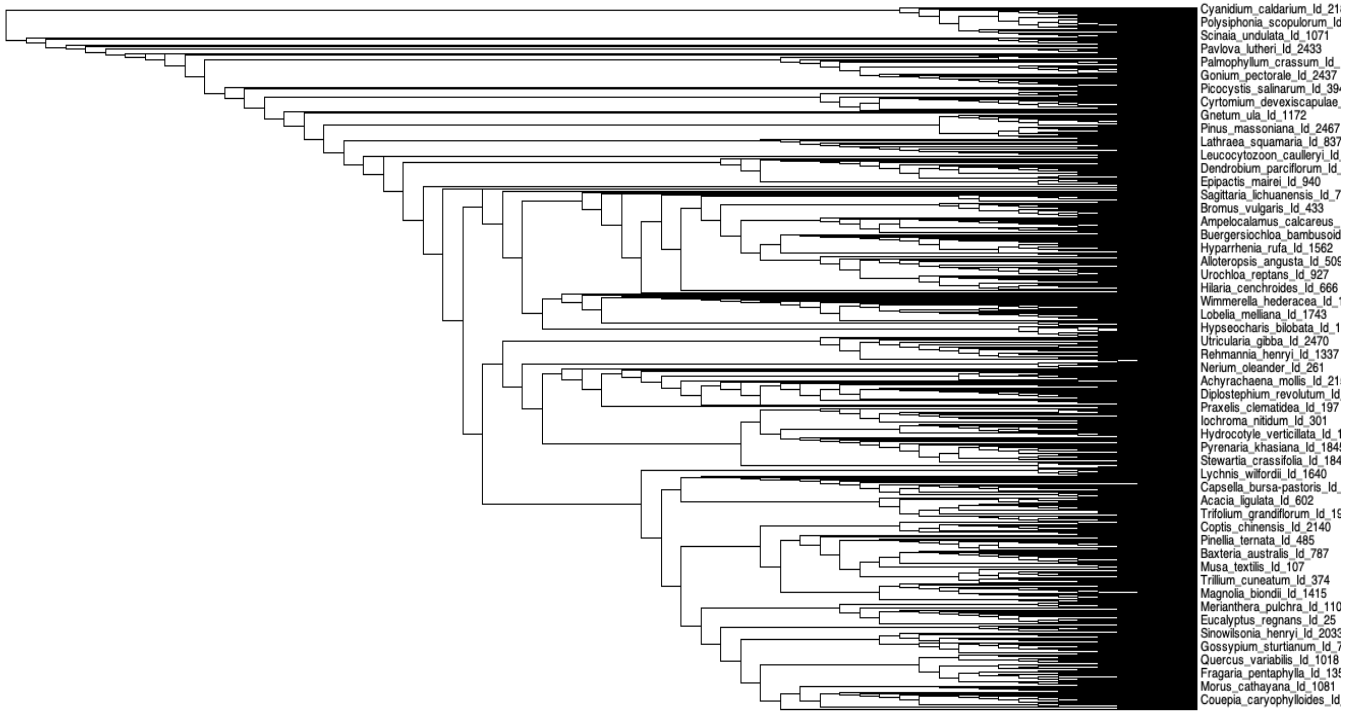
GAP

- Anotação fraca do Pinus;
- Descobrir, mapear e buscar o entendimento de genes e regiões genômicas de várias características disponíveis de *Eucalyptus* para elucidar aspectos correspondentes para o gênero Pinus.
- Identificação de genes conservados, duplicações e adaptações.

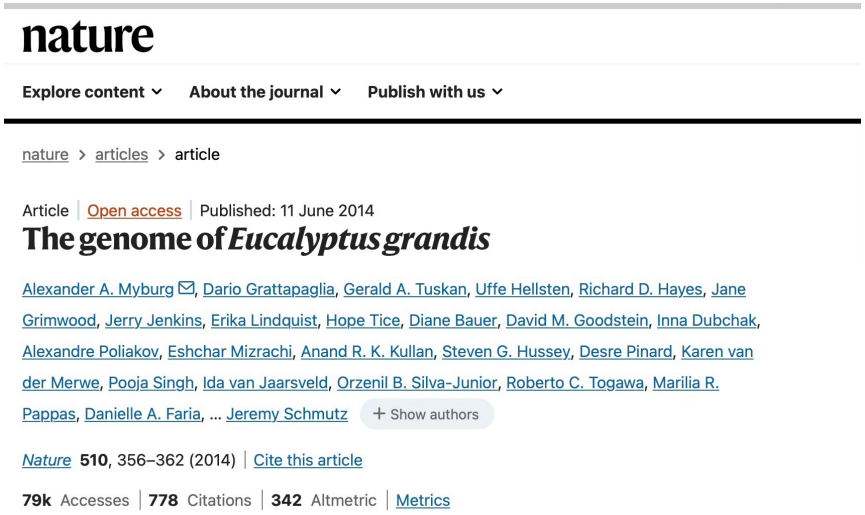
Objetivo

Avaliar e comparar a anotação estrutural e funcional do gênero *Eucalyptus* para o gênero *Pinus* visando mapear e identificar genes e regiões genômicas.

Árvore filogenética



Material e Métodos



nature

Explore content | About the journal | Publish with us

nature > articles > article

Article | Open access | Published: 11 June 2014

The genome of *Eucalyptus grandis*

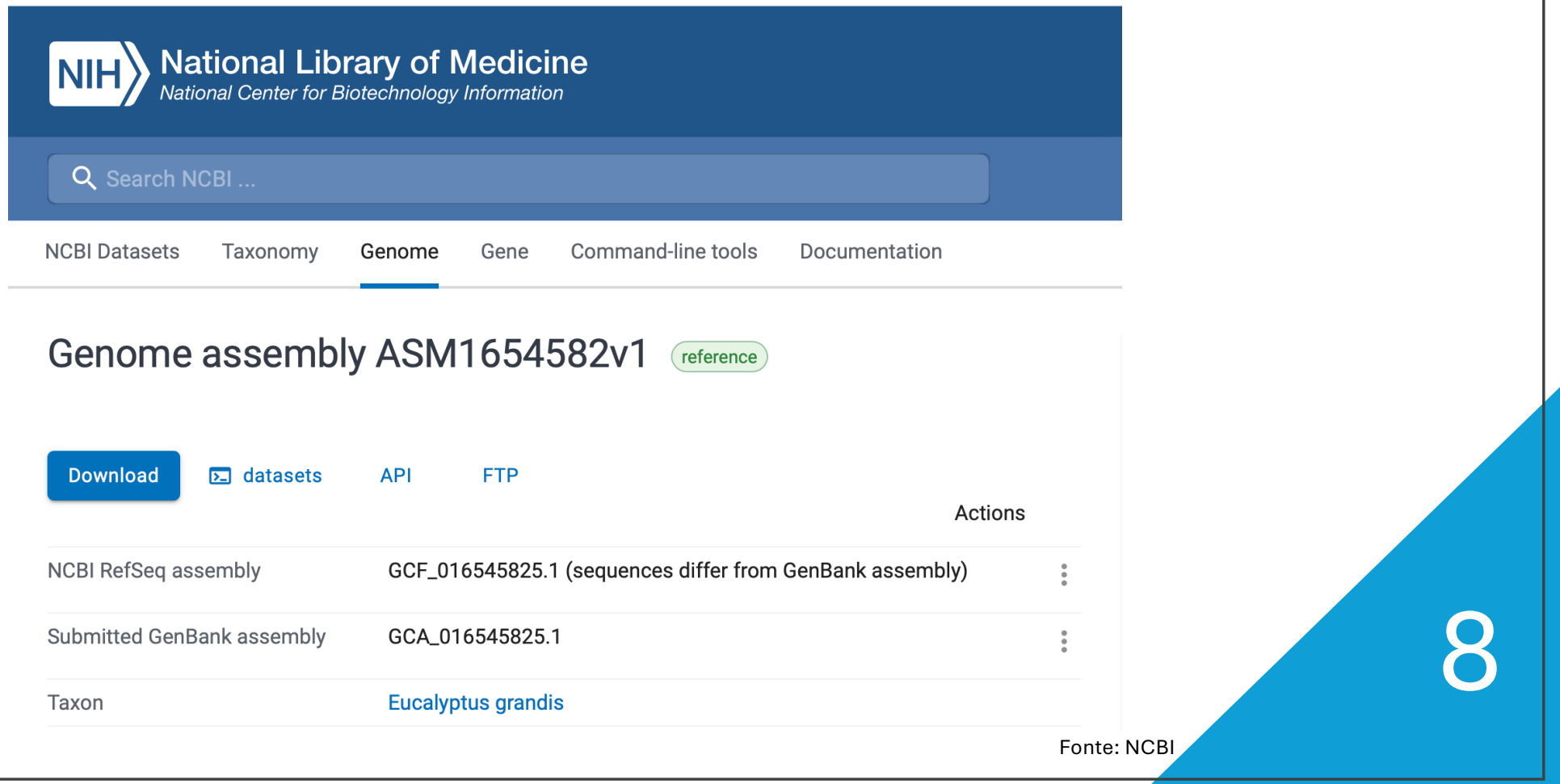
Alexander A. Myburg, Dario Grattapaglia, Gerald A. Tuskan, Uffe Hellsten, Richard D. Hayes, Jane Grimwood, Jerry Jenkins, Erika Lindquist, Hope Tice, Diane Bauer, David M. Goodstein, Inna Dubchak, Alexandre Poliakov, Eshchar Mizrachi, Anand R. K. Kullan, Steven G. Hussey, Desre Pinard, Karen van der Merwe, Pooja Singh, Ida van Jaarsveld, Orzenil B. Silva-Junior, Roberto C. Togawa, Marilia R. Pappas, Danielle A. Faria, ... Jeremy Schmutz + Show authors

Nature **510**, 356–362 (2014) | Cite this article

79k Accesses | **778** Citations | **342** Altmetric | Metrics

Genoma do *E. grandis*

Projeto Genolyptus: genoma completo de *E.grandis* (Nature, 2014)


National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
Genome assembly ASM1654582v1
reference
Download
datasets
API
FTP
Actions
NCBI RefSeq assembly
GCF_016545825.1 (sequences differ from GenBank assembly)
Submitted GenBank assembly
GCA_016545825.1
Taxon
Eucalyptus grandis

Fonte: NCBI

Assembly statistics

	RefSeq	GenBank
Genome size	615.9 Mb	615.9 Mb
Total ungapped length	615.7 Mb	615.7 Mb
Number of chromosomes	11	11
Number of organelles	1	1
Number of scaffolds	35	35
Scaffold N50	58.5 Mb	58.5 Mb
Scaffold L50	5	5
Number of contigs	1,747	1,747
Contig N50	614.3 kb	614.3 kb
Contig L50	286	286
GC percent	39.5	39.5
Genome coverage	25x	25x
Assembly level	Chromosome	Chromosome
View sequences	view RefSeq sequences	view GenBank sequences

Fonte: NCBI

Quality analysis

BUSCO analysis (4.0.2)

C:98.2%[S:91.5%,D:6.6%],F:0.4%,M:1.4%,n:2326
eudicots_odb10 (2326)

Chromosomes

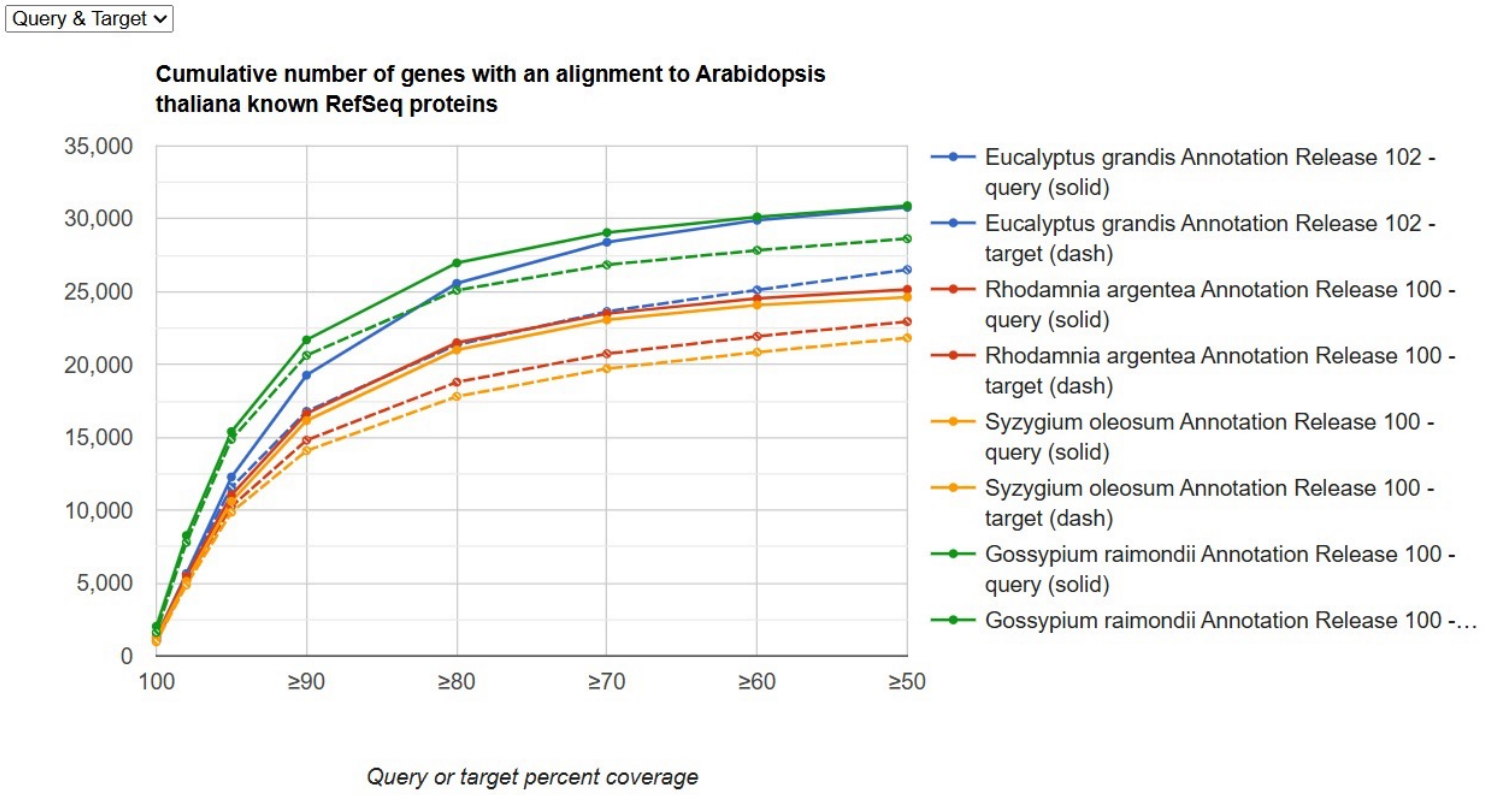
Fonte: NCBI

View chromosomes from: GenBank sequence | RefSeq sequence

Download

Chromosome	GenBank	RefSeq	Size (bp)	GC content (%)	Unlocalized count	Action
1	CM028308.1	NC_052612.1	53.951.317	39,5	0	⋮
2	CM028309.1	NC_052613.1	58.486.256	39,5	0	⋮
3	CM028310.1	NC_052614.1	72.470.685	39	0	⋮
4	CM028311.1	NC_052615.1	39.120.311	39	0	⋮
5	CM028312.1	NC_052616.1	62.829.834	39	0	⋮
6	CM028313.1	NC_052617.1	57.658.180	39,5	0	⋮
7	CM028314.1	NC_052618.1	60.484.081	39	0	⋮
8	CM028315.1	NC_052619.1	74.572.912	39,5	0	⋮
9	CM028316.1	NC_052620.1	44.060.374	39,5	0	⋮
10	CM028317.1	NC_052621.1	38.819.584	39,5	0	⋮
11	CM028318.1	NC_052622.1	50.013.132	39,5	0	⋮
Pltd	CM028319.1	n/a	160.269	36,5	0	⋮

Fonte: NCBI


Query & Target
Cumulative number of genes with an alignment to Arabidopsis thaliana known RefSeq proteins
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
100
≥90
≥80
≥70
≥60
≥50
Query or target percent coverage
Eucalyptus grandis Annotation Release 102 - query (solid)
Eucalyptus grandis Annotation Release 102 - target (dash)
Rhodamnia argentea Annotation Release 100 - query (solid)
Rhodamnia argentea Annotation Release 100 - target (dash)
Syzygium oleosum Annotation Release 100 - query (solid)
Syzygium oleosum Annotation Release 100 - target (dash)
Gossypium raimondii Annotation Release 100 - query (solid)
Gossypium raimondii Annotation Release 100 -…

Fonte: NCBI

INVESTIGATION
HIGHLIGHTED ARTICLE

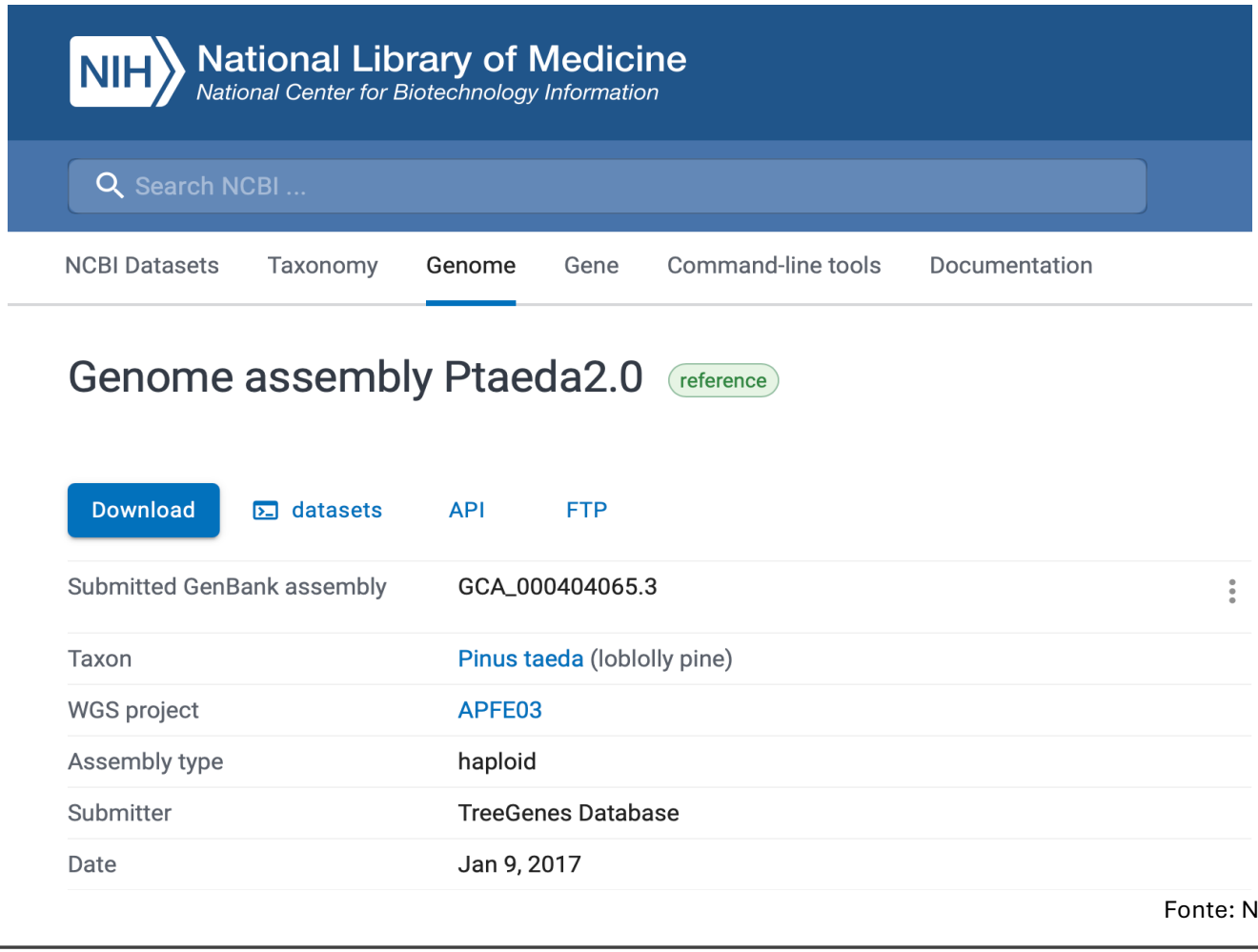
Sequencing and Assembly of the 22-Gb Loblolly Pine Genome

Aleksey Zimin,*,[1] Kristian A. Stevens,[†,1,2] Marc W. Crepeau,[†] Ann Holtz-Morris,[‡] Maxim Koriabine,[‡] Guillaume Marçais,* Daniela Puiu,[§] Michael Roberts,* Jill L. Wegrzyn, Pieter J. de Jong,[‡] David B. Neale,** Steven L. Salzberg,[§] James A. Yorke,*,[††] and Charles H. Langley[†]**

*Institute for Physical Sciences and Technology and [††]Departments of Mathematics and Physics, University of Maryland, College Park, Maryland 20742, [†]Department of Evolution and Ecology and **Department of Plant Sciences, University of California, Davis, California 95616, [‡]Children's Hospital Oakland Research Institute, Oakland, California 94609, [§]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University, Baltimore, Maryland 21205

Genoma do *Pinus taeda*

Genetics, 2014.

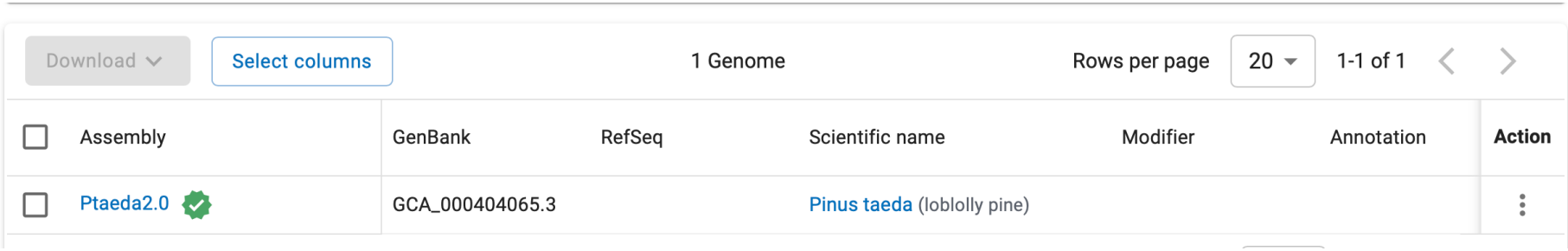

National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
Genome assembly Ptaeda2.0
reference
Download
datasets
API
FTP
Submitted GenBank assembly
GCA_000404065.3
Taxon
Pinus taeda (loblolly pine)
WGS project
APFE03
Assembly type
haploid
Submitter
TreeGenes Database
Date
Jan 9, 2017

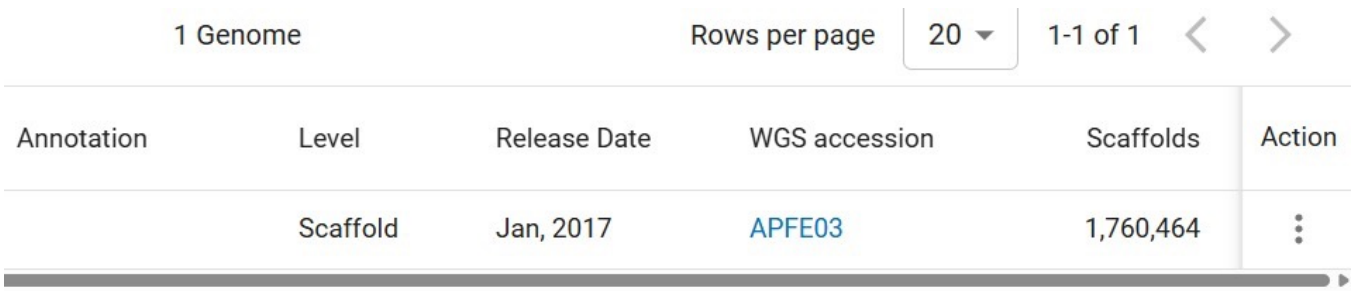
Fonte: NCBI

Assembly statistics

	GenBank
Genome size	22.1 Gb
Total ungapped length	20.5 Gb
Number of scaffolds	1,760,464
Scaffold N50	107 kb
Scaffold L50	55,536
Number of contigs	2,724,159
Contig N50	28.1 kb
Contig L50	195,676
GC percent	37.5
Assembly level	Scaffold

Fonte: NCBI

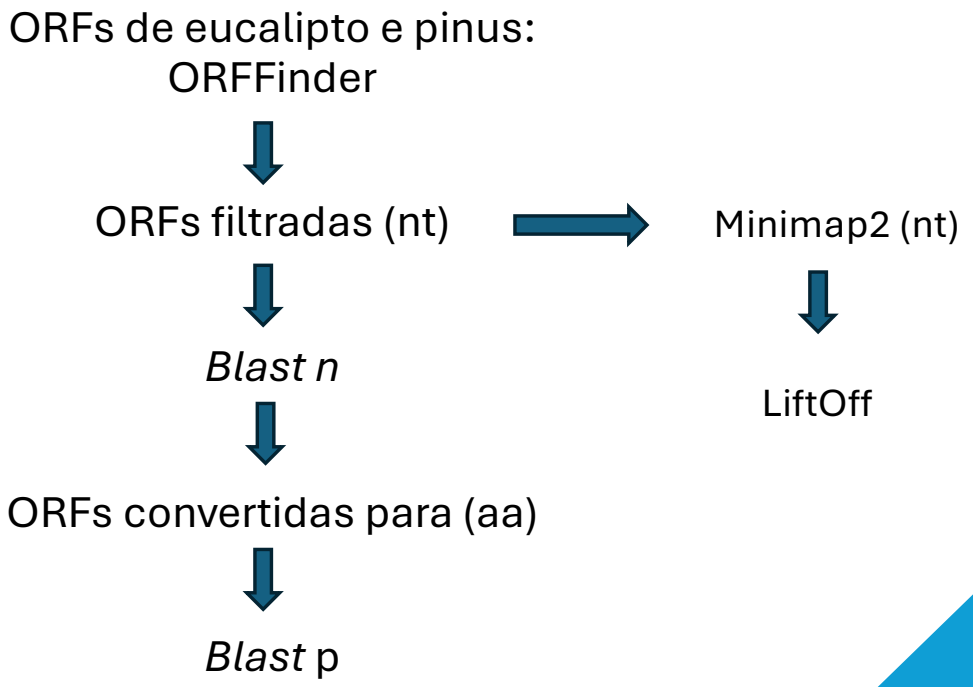




Contig sort - reorganizar os *scaffolds* por cromossomos.

Fonte: NCBI

Análises


ORFs de eucalipto e pinus:
ORFFinder
ORFs filtradas (nt)
Minimap2 (nt)
LiftOff
Blast n
ORFs convertidas para (aa)
Blast p

Análises

- SweeP: árvores internas para mitocôndria, cloroplasto e cromossomos de eucalipto.
- Identificar genes próximos ou duplicados – indicação de genes conservados.
- Árvores de mitocôndria e cloroplasto- controle interno
- Árvores dos cromossomos- todos contra todos

Resultados

Número de ORFs (nt):

```
root@asm1:/home/daiane# cat ORFs_euca_final.fa | grep ">" | wc -l
379486
root@asm1:/home/daiane# cat ORFs_pinus_final.fa | grep ">" | wc -l
11633556
```

Resultados

```
daiane@asm1:~/ncbi-blast-2.16.0+$ less top10_ids.txt

ORF_1711714     3099
ORF_2679755     3069
ORF_7386192     2757
ORF_11073875    2700
ORF_10722655    2646
ORF_9659038     2574
ORF_3391147     2532
ORF_5990318     2532
ORF_11476439    2352
ORF_9198000     2343
~
```

ORFs filtradas: ≥70% identidade e ≥70% cobertura

Top 10- *Blast n*

1 gene: PREDICTED: Eucalyptus grandis DExH-box ATP-dependent RNA helicase DExH12 (LOC104448968), mRNA

Resultados

- Número de ORFs convertidas para sequências de proteínas:

```
daiane@asm1:/opt/orf_x_aa$ grep -c "^>" ORFs_pinus_final_protein.fa
11550679
```

Perda de 82.877 ORFs por problemas de códons incompatíveis;

Motivo mais frequente: presença de N nas sequências.

Resultados

- *Blast p* das ORFs do pinus contra 44.965 sequências do proteoma do eucalipto : 83.238.047 alinhamentos;

Resultados

- Alinhamentos relacionados a15 proteínas do eucalipto:

id=1002203_lcl|NC_052621.1_prot_XP_010033053.2_2203
id=1003323_lcl|NC_052621.1_prot_XP_010034004.2_3323
id=101646_lcl|NC_052612.1_prot_XP_010046757.1_1646
id=1102838_lcl|NC_052622.1_prot_XP_039160155.1_2838
id=201617_lcl|NC_052613.1_prot_XP_010039798.2_1617
id=204323_lcl|NC_052613.1_prot_XP_010044938.1_4323
id=204697_lcl|NC_052613.1_prot_XP_010045349.2_4697
id=303859_lcl|NC_052614.1_prot_XP_010026838.2_3859
id=600582_lcl|NC_052617.1_prot_XP_010060264.2_582
id=700721_lcl|NC_052618.1_prot_XP_039174107.1_721
id=800544_lcl|NC_052619.1_prot_XP_039155478.1_544
id=800949_lcl|NC_052619.1_prot_XP_010041028.2_949
id=803806_lcl|NC_052619.1_prot_XP_010024814.2_3806
id=900747_lcl|NC_052620.1_prot_XP_010027997.1_747
id=902333_lcl|NC_052620.1_prot_XP_039158489.1_2333

Resultados

- Parâmetro utilizado: limite máximo de 10 alinhamentos /*query*;
- Gerar um banco de (aa) para cada cromossomo e aumentar para 500 alinhamentos/*query* para refinar o resultado – parcialmente
- Blast parece não ser viável: *splicing* está sendo ignorado.

Resultados

- *Minimap2*: ferramenta indicada para alinhamento em eucariotos (nt);
- 1.042.889 sequências ORFs do pinus alinhadas (presença de muitos “N”)
- Análise foi refeita – interpretação dos resultados – a fazer
- Alinhamentos resultantes: ferramenta *Liftoff*: mapeamento da anotações de genes entre montagens de genomas – a fazer

Resultados

- SweeP a partir do proteoma do eucalipto: para mitocôndria, cloroplasto e cromossomos - permite identificar genes próximos (ou duplicados) dentro de cada molécula anotada do eucalipto

- Mitocôndria- 74 genes anotados
- Cloroplasto – 38 genes anotados

Resultados

NC_052612_cromossomo_01_arvore_nj.tree

```
id=103420_lcl|NC_052612.1_prot_XP_039155991.1_3420 [gene=LOC120287299]
[db_xref=GeneID\:120287299] [protein=E3 ubiquitin-protein ligase PUB23-like]
[protein_id=XP_039155991.1] [location=complement\(join\(47366882..47367228\,47367339..47367807\)
\)] [gbkey=CDS]:0.53958,
(
id=103421_lcl|NC_052612.1_prot_XP_010024581.2_3421 [gene=LOC104415053]
[db_xref=GeneID\:104415053] [protein=E3 ubiquitin-protein ligase PUB24]
[protein_id=XP_010024581.2] [location=complement\(47385986..47387206\)] [gbkey=CDS]:0.35023,
id=103422_lcl|NC_052612.1_prot_XP_010024590.2_3422 [gene=LOC104415061]
[db_xref=GeneID\:104415061] [protein=E3 ubiquitin-protein ligase PUB24]
[protein_id=XP_010024590.2] [location=complement\(47395198..47396415\)] [gbkey=CDS]:0.34556)
Branch 950:0.35082)
Branch 1655:0.14452,
(
id=102828_lcl|NC_052612.1_prot_XP_010055882.2_2828 [gene=LOC104444001]
[db_xref=GeneID\:104444001] [protein=F-box protein At2g41170] [protein_id=XP_010055882.2]
[location=complement\
(join\(42206883..42207070\,42207393..42207467\,42207594..42207766\,42208131..42208210\,42208386.
.42209057\)\)] [gbkey=CDS]:0.76877,
id=103799_lcl|NC_052612.1_prot_XP_010055111.2_3799 [gene=LOC104443429]
[db_xref=GeneID\:104443429] [protein=bifunctional L-3-cyanoalanine synthase/cysteine synthase
1\, mitochondrial] [protein_id=XP_010055111.2]
[location=join\(51379020..51379205\,51379833..51379942\,51380074..51380131\,51380220..51380483\,
51380576..51380713\,51380817..51380868\,51381034..51381113\,51381229..51381288\,51381419..513814
99\,51381699..51381797\)] [gbkey=CDS]:0.76462)
Branch 2567:0.028681)
Branch 3407:0.015924,
(
```

Resultados

- SweeP do proteoma do pinus com base nas ORFs correlacionadas com os genes do eucalipto - a fazer

Resultados

- Propor o conjunto de genes comum aos dois gêneros – a fazer

Próximas etapas

- Blast com o proteoma de cada cromossomo do eucalipto, aumentando para 500 alinhamentos/*query*;
- Interpretar os resultados de alinhamento com o *Minimap2* e aplicar a ferramenta *Liftoff*;
- SweeP do proteoma do pinus com base nas ORFs correlacionadas com os genes do eucalipto;
- Propor o conjunto de genes comum aos dois gêneros.

Cronograma

- Finalizar análises: outubro 2025
- Interpretação dos resultados: outubro/novembro 2025
- Escrever dissertação: novembro 2025
- Defesa: dezembro 2025 ou fevereiro 2026.

Agradecimentos

Ao Programa de Pós-graduação em Bioinformática da UFPR;

Ao Prof. Dr Dieval Guizelini;

Aos Professores e secretária do programa;

Aos colegas.