

# Lawrence Berkeley National Laboratory

## LBL Publications

### Title

Not Created Equal Towards comprehensive citation capture and classification at the US DOE Joint Genome Institute

### Permalink

<https://escholarship.org/uc/item/9mf1s1z7>

### Author

Byers, Neil

### Publication Date

2023-06-08

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed



# Not Created Equal

*Towards comprehensive citation capture and classification at the US DOE Joint Genome Institute*

—

**Neil Byers**

**Data Scientist / Impact Analyst**

—

Bibliometrics & Research Impact Conference  
Ottawa, ON - June 8, 2023

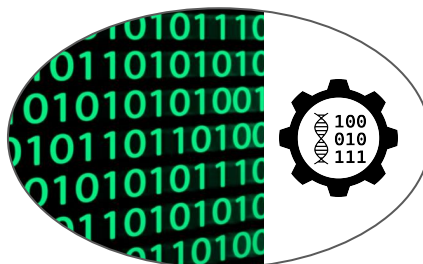
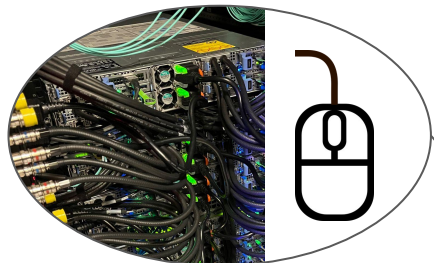
# Outputs & Impacts



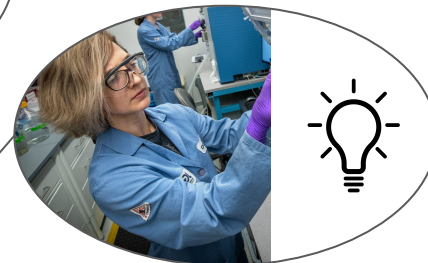
# Outputs & Impacts

Data

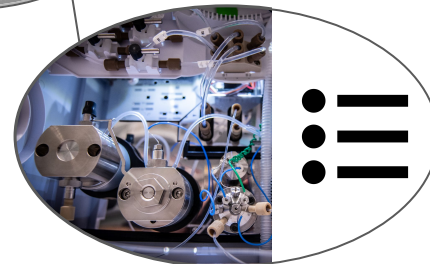
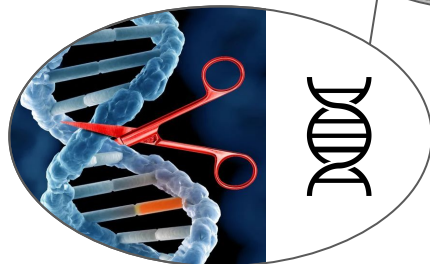
Systems &  
Tools



Discoveries &  
Findings



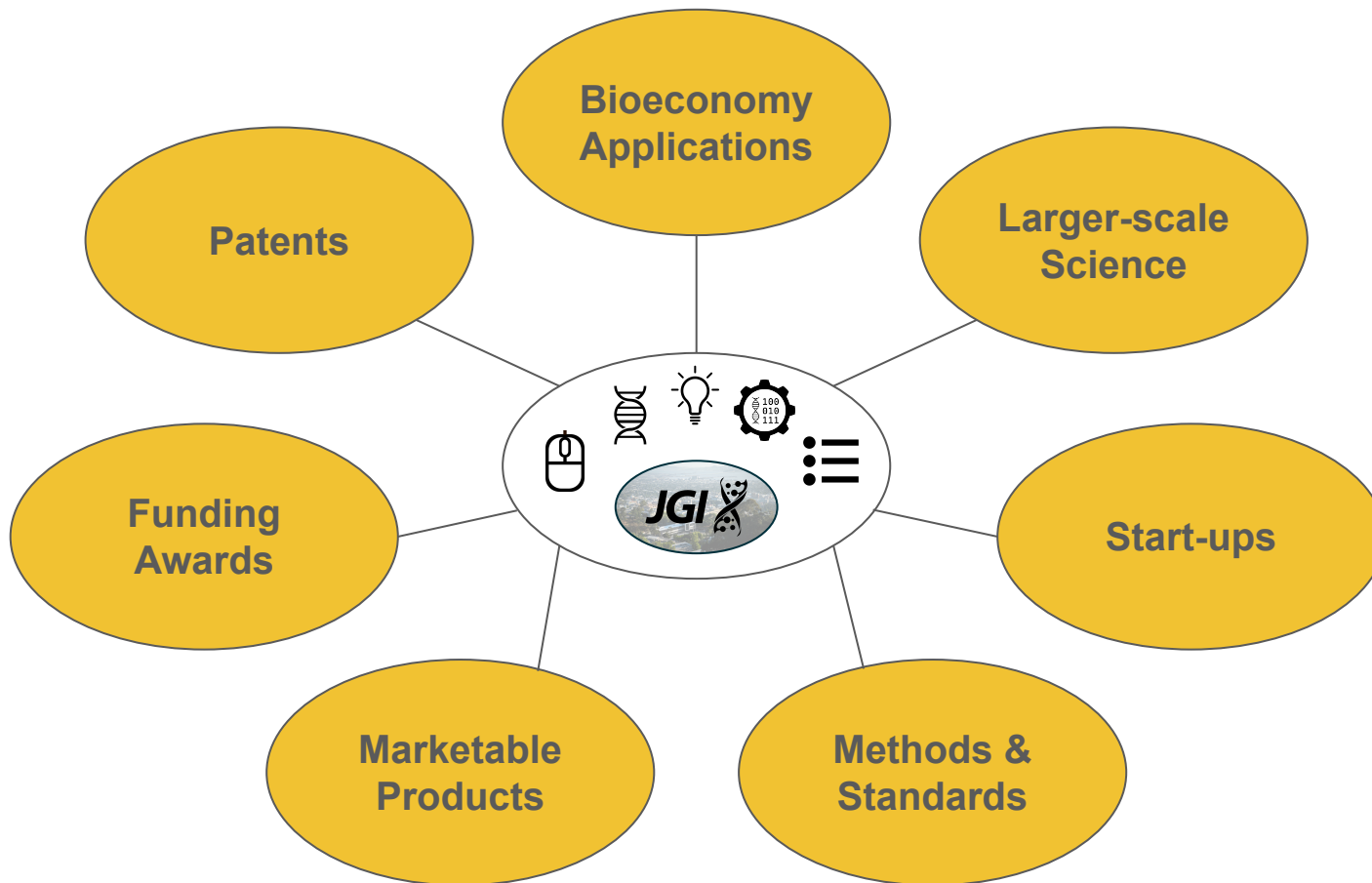
Genetic  
Materials



Technologies &  
Protocols

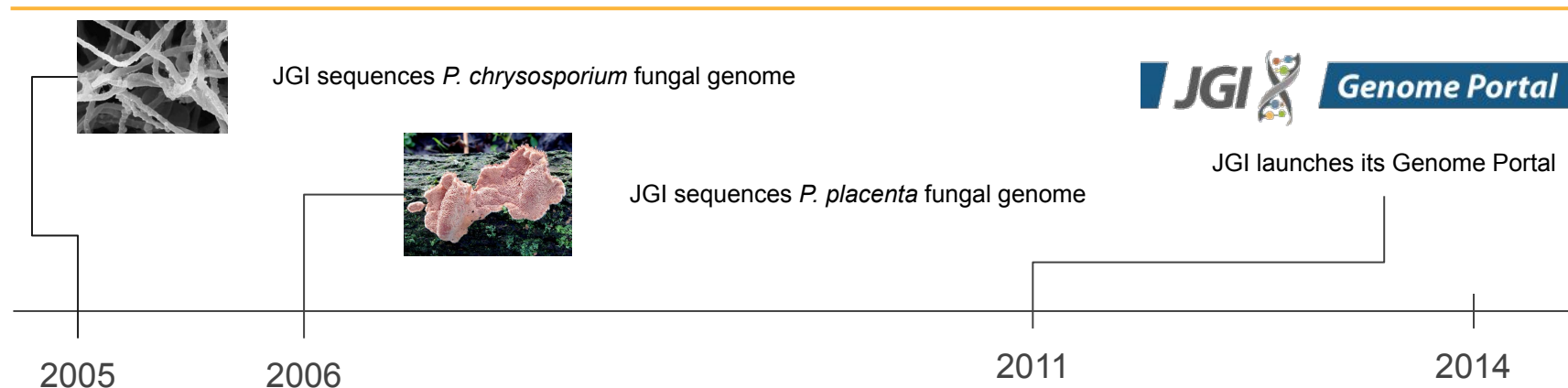


# Outputs & Impacts



# Impact Narrative Examples

## How does 'impact' begin?



Berkeley  
UNIVERSITY OF CALIFORNIA

**I** UNIVERSITY OF  
ILLINOIS  
URBANA-CHAMPAIGN

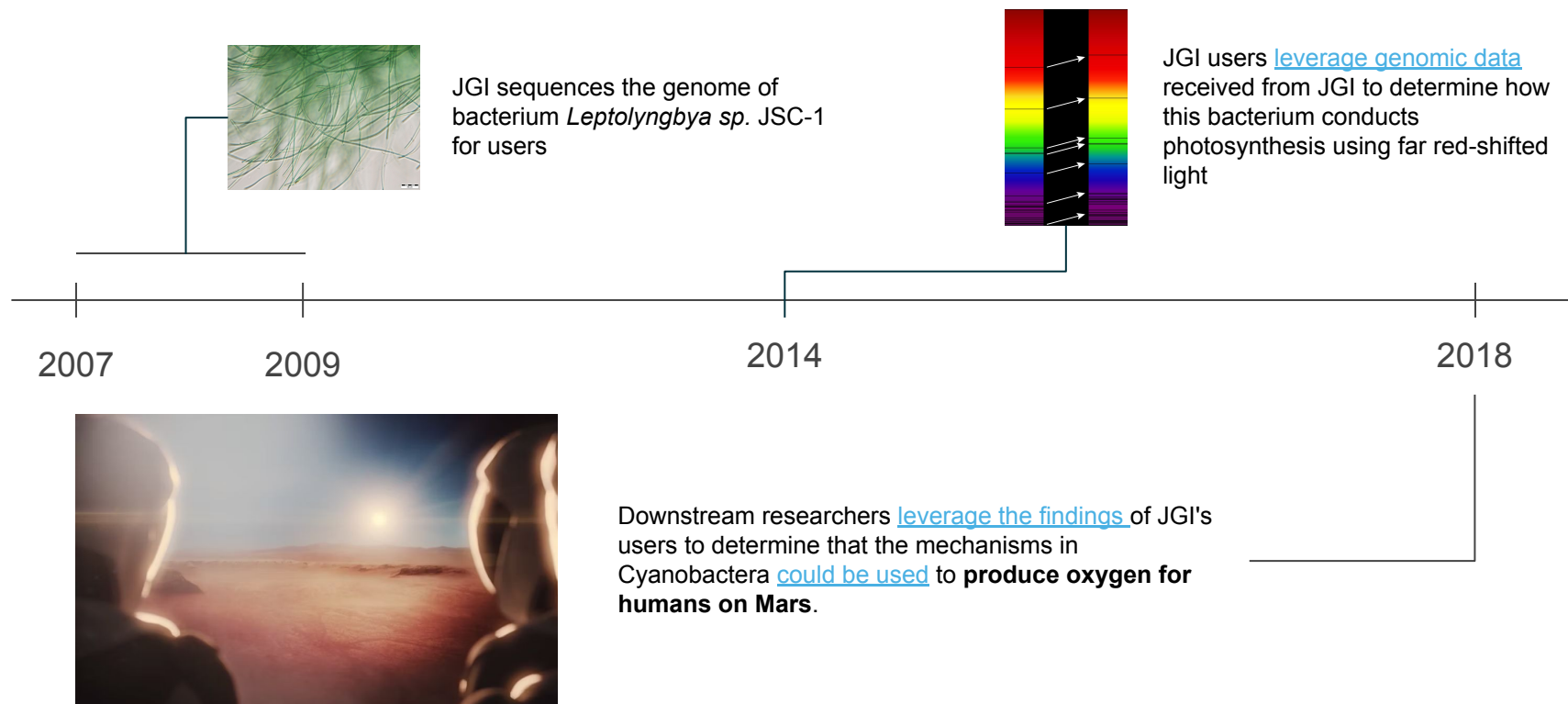


Using data from the above genomes retrieved from JGI's Genome Portal, university and BP researchers [patent processes](#) to improve cellular sugar transportation for **production of biofuels** like ethanol.



# Impact Narrative Examples

## How does 'impact' begin?



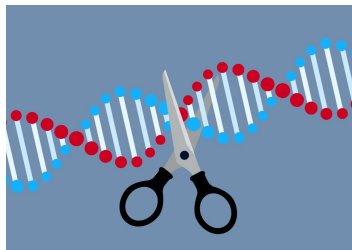


# Impact Narrative Examples

## One cautionary tale



JGI authors [publish a review](#) synthesizing community findings surrounding CRISPR as a defense against viruses



Researchers from universities in Sweden, Germany, and Austria cite the JGI review in [a study](#) detailing mechanisms involved with CRISPR and anti-viral immunity

2008

2011

2013

2018



Bayer CropScience



This CRISPR study is cited by over 500 patents by inventors from 70 organizations, including the large companies at left



**A**

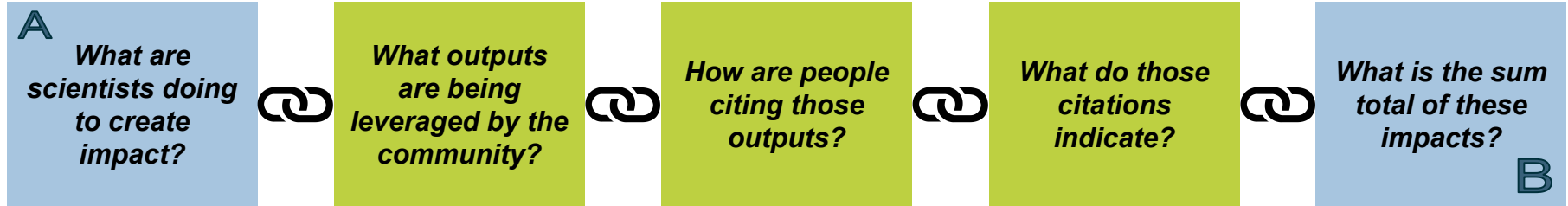
***What are  
scientists doing  
to create impact?***



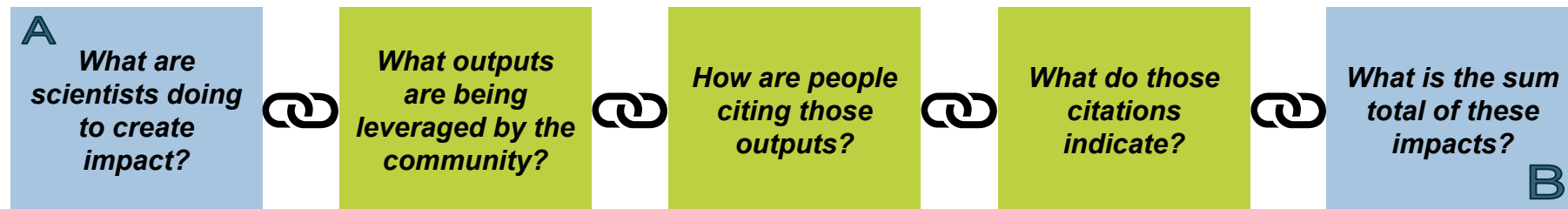
**B**

***What is the sum  
total of these  
impacts?***

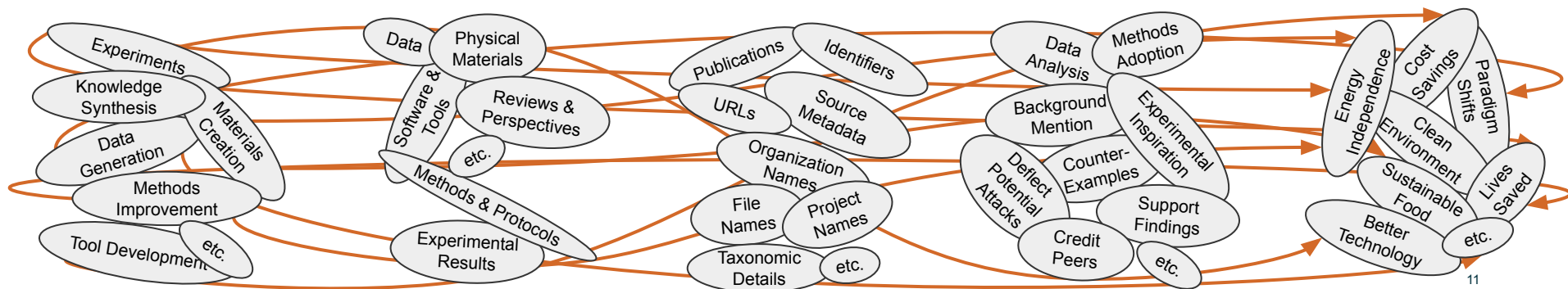
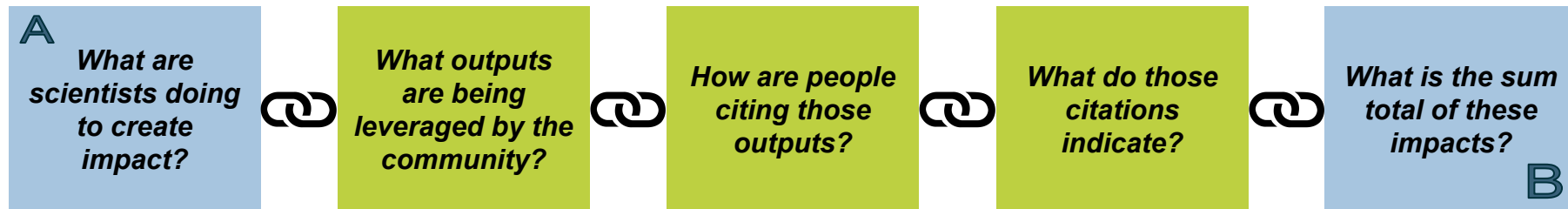
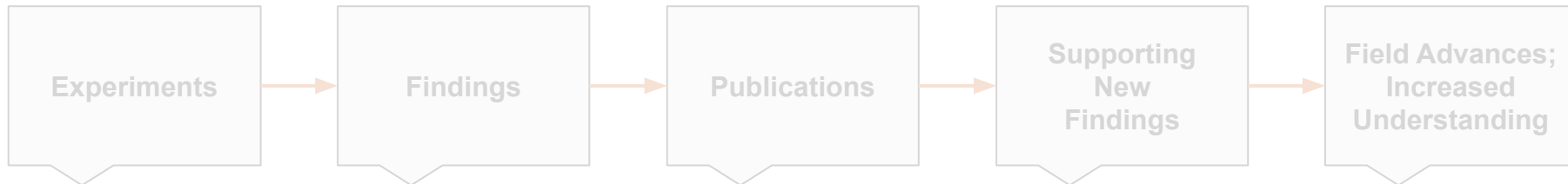
# Getting from A to B



# Getting from A to B



# A Messy Reality

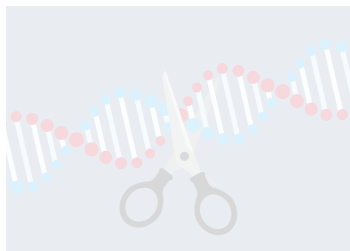


# Impact Narrative Examples

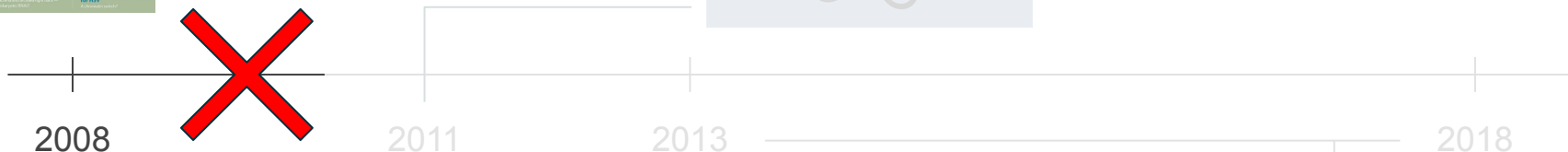
## One cautionary tale



JGI authors [publish a review](#) synthesizing community findings surrounding CRISPR as a defence against viruses



Researchers from universities in Sweden, Germany, and Austria cite the JGI review in [a study](#) detailing mechanisms involved with CRISPR and anti-viral immunity

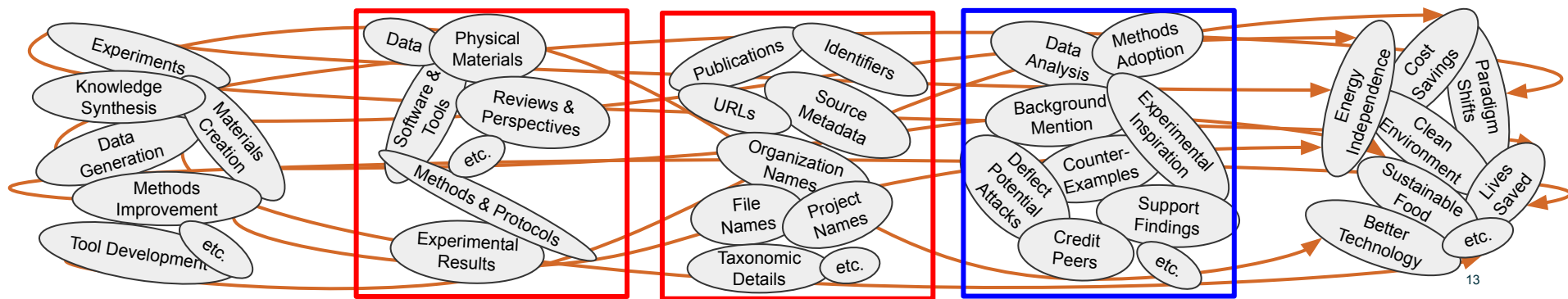
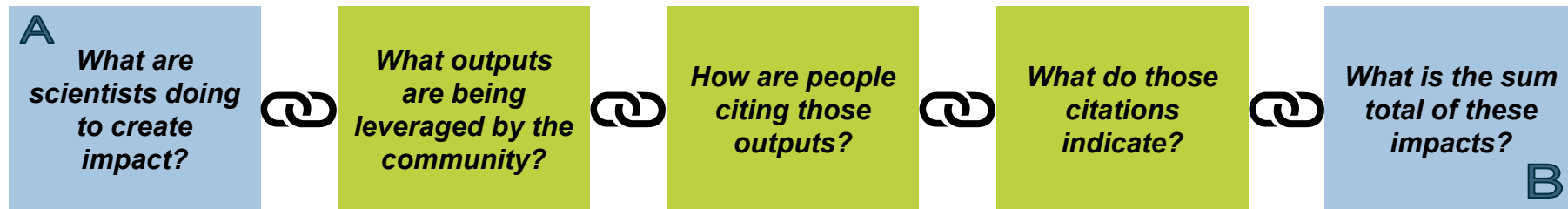


Bayer CropScience



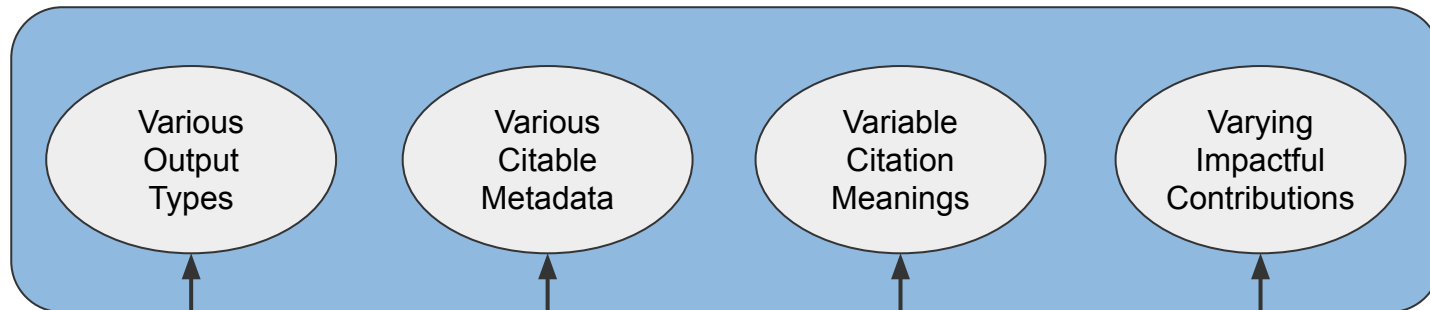
This CRISPR study is cited by over 500 patents by inventors from 70 organizations, including the large companies at left

# A Messy Reality



# Examples: 1KFG and Soybean

**1KFG**  
~  
**1800**  
citations



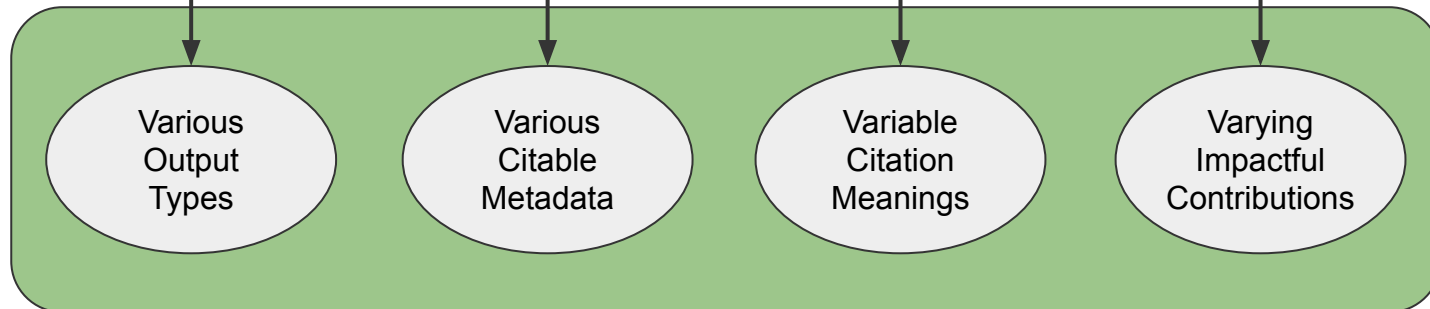
Includes...

Can be  
cited via...

Shows  
influence via...

Is important  
because of...

**Soybean**  
~  
**5000**  
citations





# How can JGI be cited?

*JGI Outputs*

*Citable Metadata*

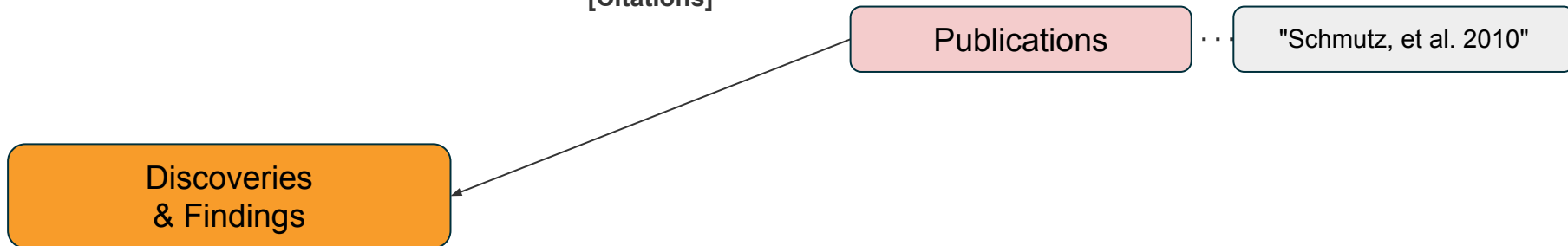
*Examples*

[Citations]

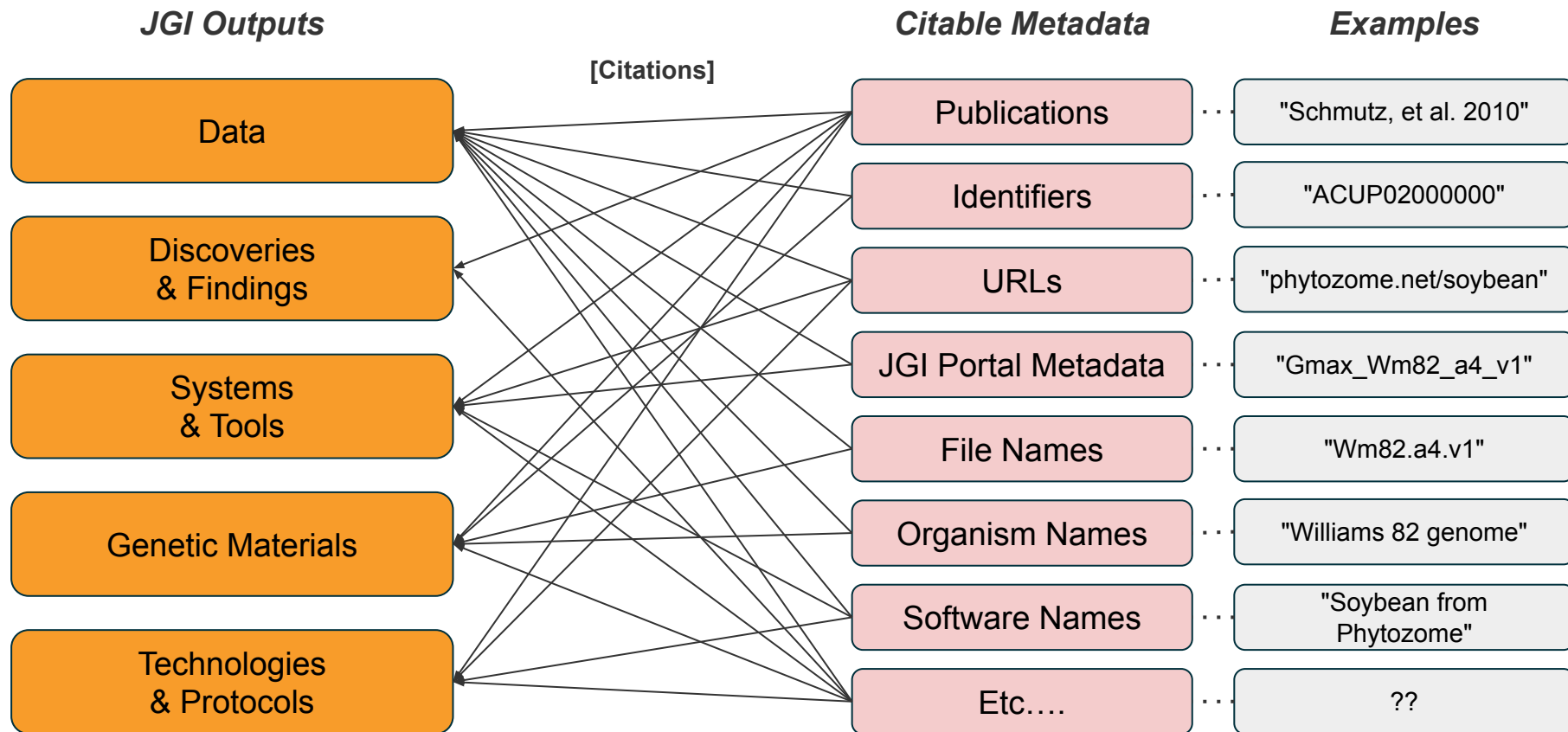
Publications

"Schmutz, et al. 2010"

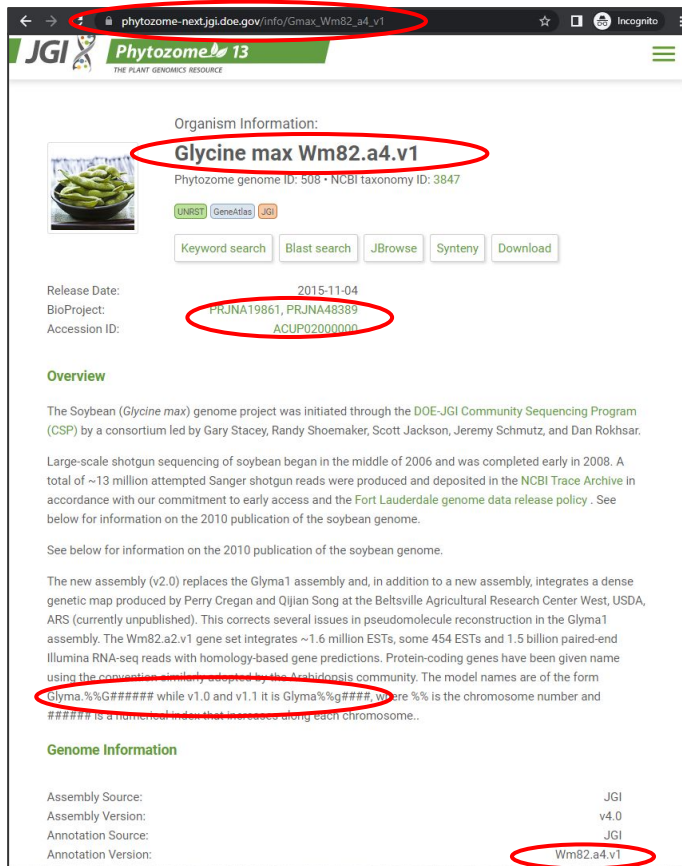
Discoveries  
& Findings



# How can JGI be cited?



# How can JGI be cited?



phytozome-next.jgi.doe.gov/info/Gmax\_Wm82\_a4\_v1

**Glycine max Wm82.a4.v1**

Phytozome genome ID: 508 • NCBI taxonomy ID: 3847

UNRST GeneAtlas JGI

Keyword search Blast search JBrowse Synteny Download

Release Date: 2015-11-04  
BioProject: PRJNA19861, PRJNA48389  
Accession ID: ACUP02000000

**Overview**

The Soybean (*Glycine max*) genome project was initiated through the DOE-JGI Community Sequencing Program (CSP) by a consortium led by Gary Stacey, Randy Shoemaker, Scott Jackson, Jeremy Schmutz, and Dan Rokhsar.

Large-scale shotgun sequencing of soybean began in the middle of 2006 and was completed early in 2008. A total of ~13 million attempted Sanger shotgun reads were produced and deposited in the NCBI Trace Archive in accordance with our commitment to early access and the Fort Lauderdale genome data release policy. See below for information on the 2010 publication of the soybean genome.

See below for information on the 2010 publication of the soybean genome.

The new assembly (v2.0) replaces the Glyma1 assembly and, in addition to a new assembly, integrates a dense genetic map produced by Perry Cregan and Qijian Song at the Beltsville Agricultural Research Center West, USDA, ARS (currently unpublished). This corrects several issues in pseudomolecule reconstruction in the Glyma1 assembly. The Wm82.a2.v1 gene set integrates ~1.6 million ESTs, some 454 ESTs and 1.5 billion paired-end Illumina RNA-seq reads with homology-based gene predictions. Protein-coding genes have been given name using the convention similarly adopted by the Arabidopsis community. The model names are of the form Glyma.g##### while v1.0 and v1.1 it is Glyma.g###, where % is the chromosome number and ##### is a numerical index that increases along each chromosome.

**Genome Information**

Assembly Source: JGI  
Assembly Version: v4.0  
Annotation Source: JGI  
Annotation Version: Wm82.a4.v1

DB Xrefs:

- BioProject: PRJNA19861
- BioProject: PRJNA48389
- JGIAP: 1297380
- JGISP: 1145325
- JGISP: 1031115
- JGISP: 1031106
- JGISP: 1127719
- JGISP: 1047865
- WGS: ACUP02000000

## Reference Publication(s)

- Valliyodan, B., Cannon, S. B., Bayer, P. E., Shu, S., Brown, A. V., Ren, L., ... Nguyen, H. T. (2019). Construction and comparison of three reference-quality genome assemblies for soybean. *The Plant Journal*, 100(5), 1066–1082. <https://doi.org/10.1111/tpj.14500>

## Related Publications

- Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., ... Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278), 178–183. <https://doi.org/10.1038/nature08670>

# How can JGI be cited?

← → mycosm.jgi.doe.gov/Pcit129764/Pcit129764.home.html ☆ Incognito

**JGI MycoCosm** THE FUNGAL GENOMICS RESOURCE  
JGI HOME GENOME PORTAL MYCOCOSM PHYCOCOSM LOGIN

Home • **Phyllosticta citrichinaensis CBS 129764 v1.0**

SEARCH BLAST BROWSE ANNOTATIONS MCL CLUSTERS SYNTENY DOWNLOAD INFO HOME

This genome was sequenced as part of the JGI CSP "1KFG - Deep Sequencing of Ecologically-relevant Dikarya (CSP 1974) and more specifically as a part of the Dothideomycetes Sequencing Project, which seeks to densely sample members of a diverse lineage of saprotrophic, endophytic and pathogenic fungi to examine functional diversity of fungi with a shared evolutionary history.

*Phyllosticta* is an Ascomycete fungus in the Dothideomycetes clade. *Phyllosticta* spp. have globally been recorded as endophytes, plant pathogens and saprobes from a wide range of plant hosts. Several *Phyllosticta* species have been isolated from *Citrus* spp. worldwide. Some of these are causal agents of impactful diseases such as citrus black spot and tan spot, subjected to phytosanitary legislation in the EU and the U.S.A. *Phyllosticta citrichinaensis* was isolated from leaves and fruits of mandarins, pomeloes, oranges and lemon. This taxon caused minor irregular spots or freckles, showing a weak virulence. Considering their economic impact, whole genome sequences for all the species associated with citrus plants are needed to improve our understanding of the differences in pathogenicity and evolutionary separation. These data will also allow for the development of robust DNA barcodes for quick detection and will facilitate further research on this important *Citrus* pathogenic and non-pathogenic species.

Researchers who wish to publish analyses using data from unpublished CSP genomes are respectfully required to contact the PI and JGI to avoid potential conflicts on data use and coordinate other publications with the CSP master paper(s).

**Genome Reference(s)**

Please cite the following publication(s) if you use the data from this genome in your research:

Buijs VA, Groenewald JZ, Haridas S, LaButti KM, Lipzen A, Martin FM, Barry K, Grigoriev IV, Crous PW, Seidl MF  
Enemy or ally: a genomic approach to elucidate the lifestyle of *Phyllosticta citrichinaensis*. G3 (Bethesda). 2022 May 6;12(5):. doi: 10.1093/g3journal/jkac061

**Phyllosticta citrichinaensis CBS 129764 v1.0**

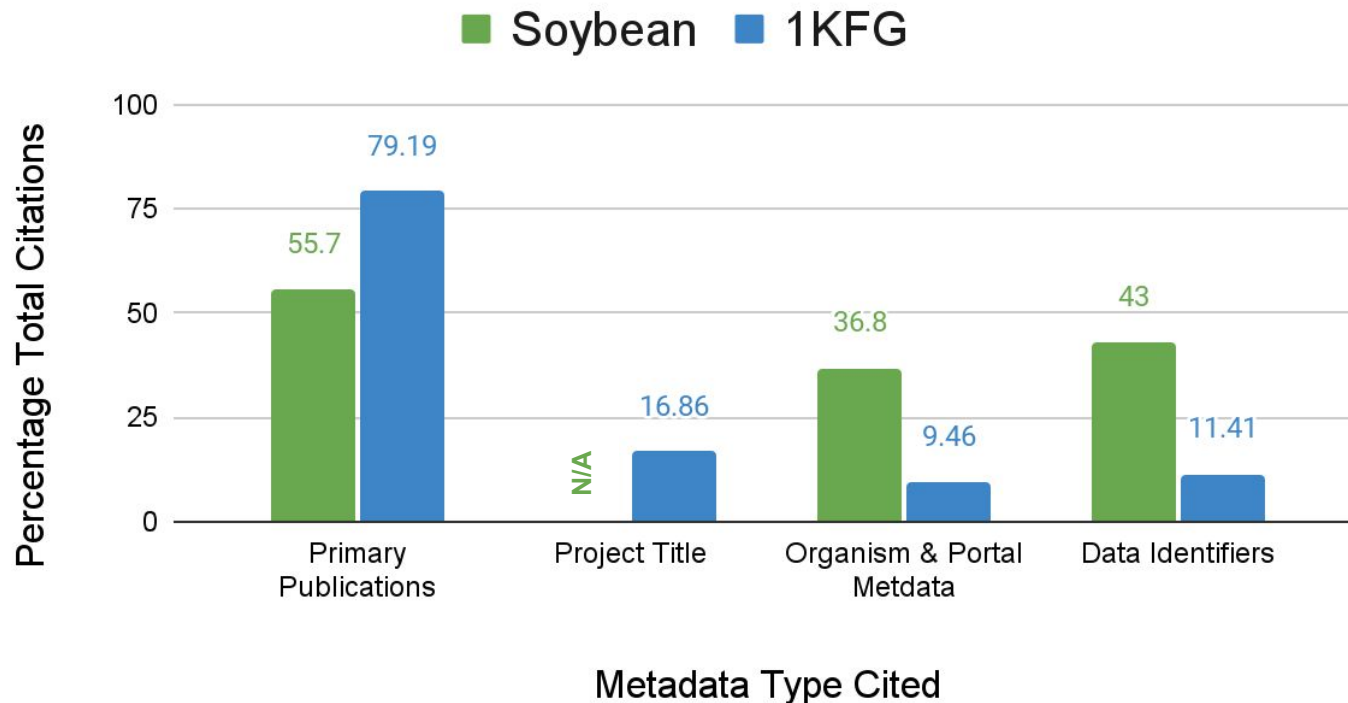
MYCOCOSM	DOWNLOAD	STATS	HELP
<b>Project name:</b>	Phyllosticta citrichinaensis CBS 129764 Annotated Standard Draft ( Project ID: 1249048 )		
<b>Product:</b>	Fungal Annotation		
<b>Proposal Name:</b>	1KFG: Deep Sequencing of Ecologically-relevant Dikarya (Proposal ID: 1974)		
<b>Project PI:</b>	Francis Michel Martin		
<b>User Program:</b>	CSP		
<b>Program Year:</b>	2016		
<b>Scientific Program:</b>	Fungal		
<b>Related Projects:</b>	FD 1248996; SP 1249049; SP 1249048; AP 1248999; AP 1274600; AP 1248997; AP 1248998		
<b>Release Date:</b>	2020-10-21		

**The data on the next page is public. Please cite:**

Buijs VA, Groenewald JZ, Haridas S, LaButti KM, Lipzen A, Martin FM, Barry K, Grigoriev IV, Crous PW, Seidl MF  
Enemy or ally: a genomic approach to elucidate the lifestyle of *Phyllosticta citrichinaensis*. G3 (Bethesda). 2022 May 6;12(5):. doi: 10.1093/g3journal/jkac061

# How can JGI be cited?

## Citation Source Percentages: Soybean vs. 1KFG



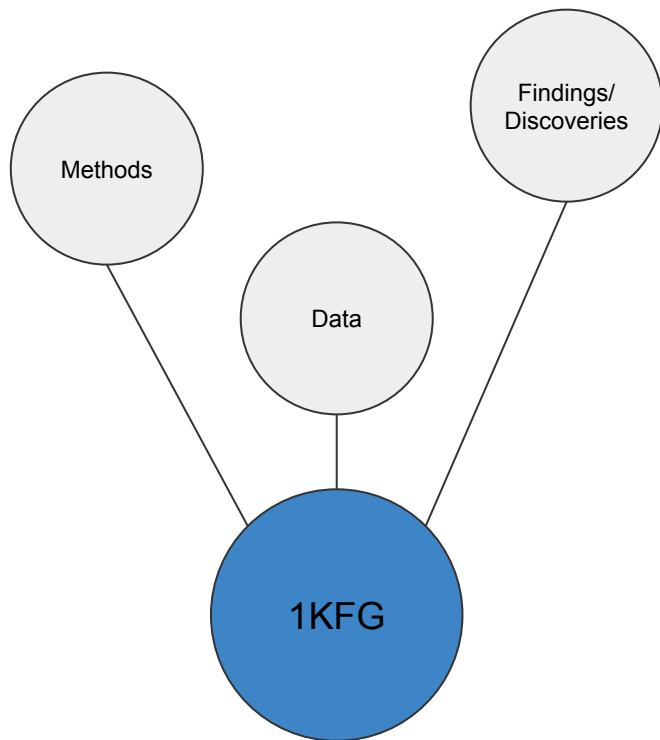
*Soybean Citations:*

*~5000*

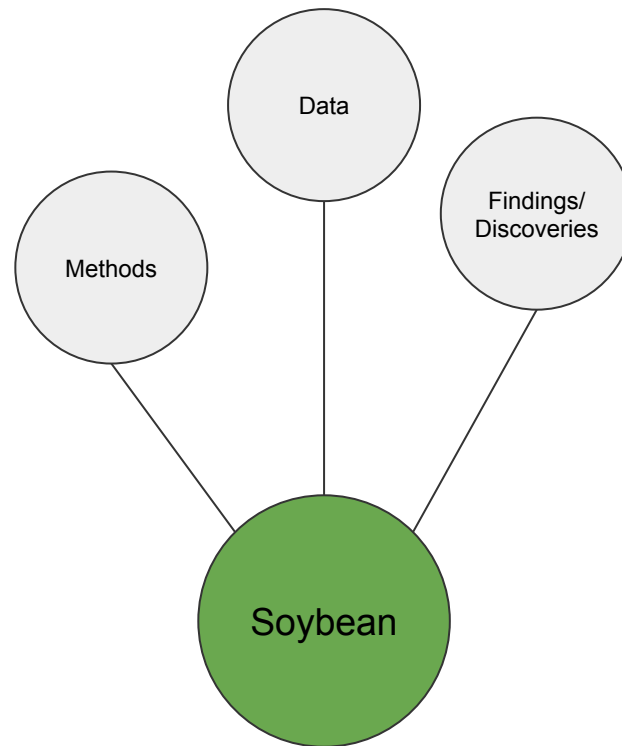
*1KFG Citations*

*~1800*

# What's being cited and for what?

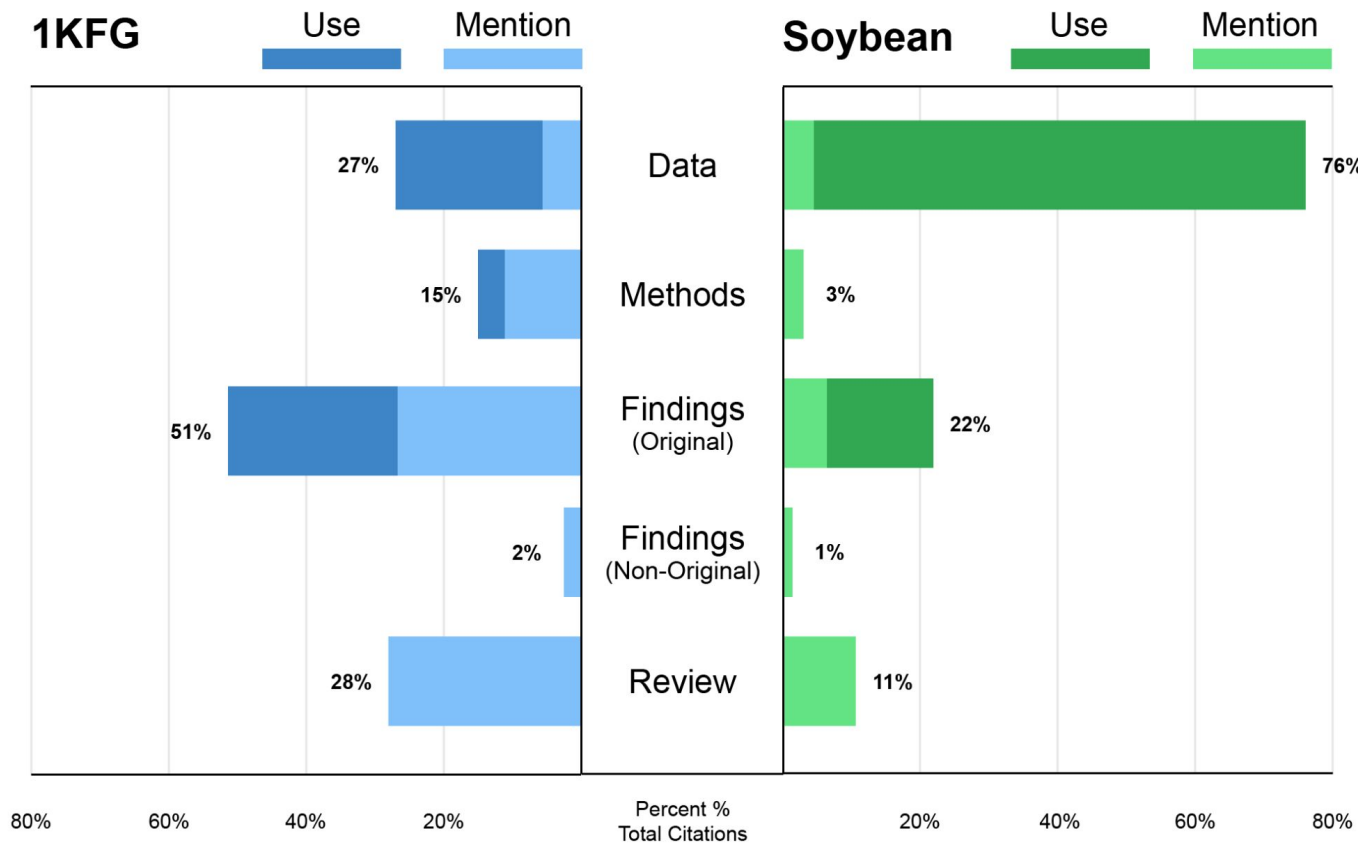


~1800 Citations



~5000 Citations

# What's being cited and for what?

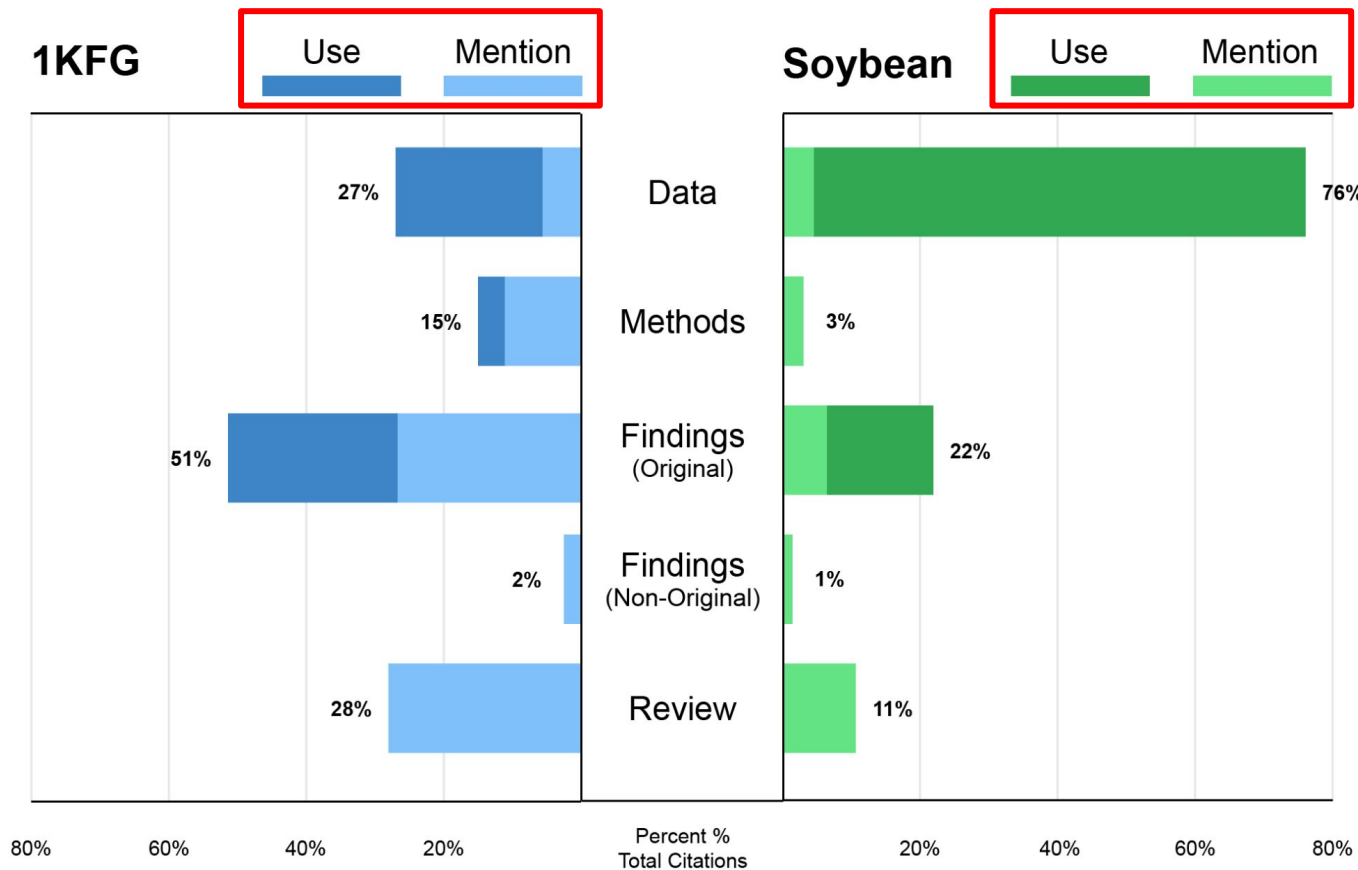


*Soybean Citations:*  
~5000

—  
*1KFG Citations*  
~1800



# What's being cited and for what?

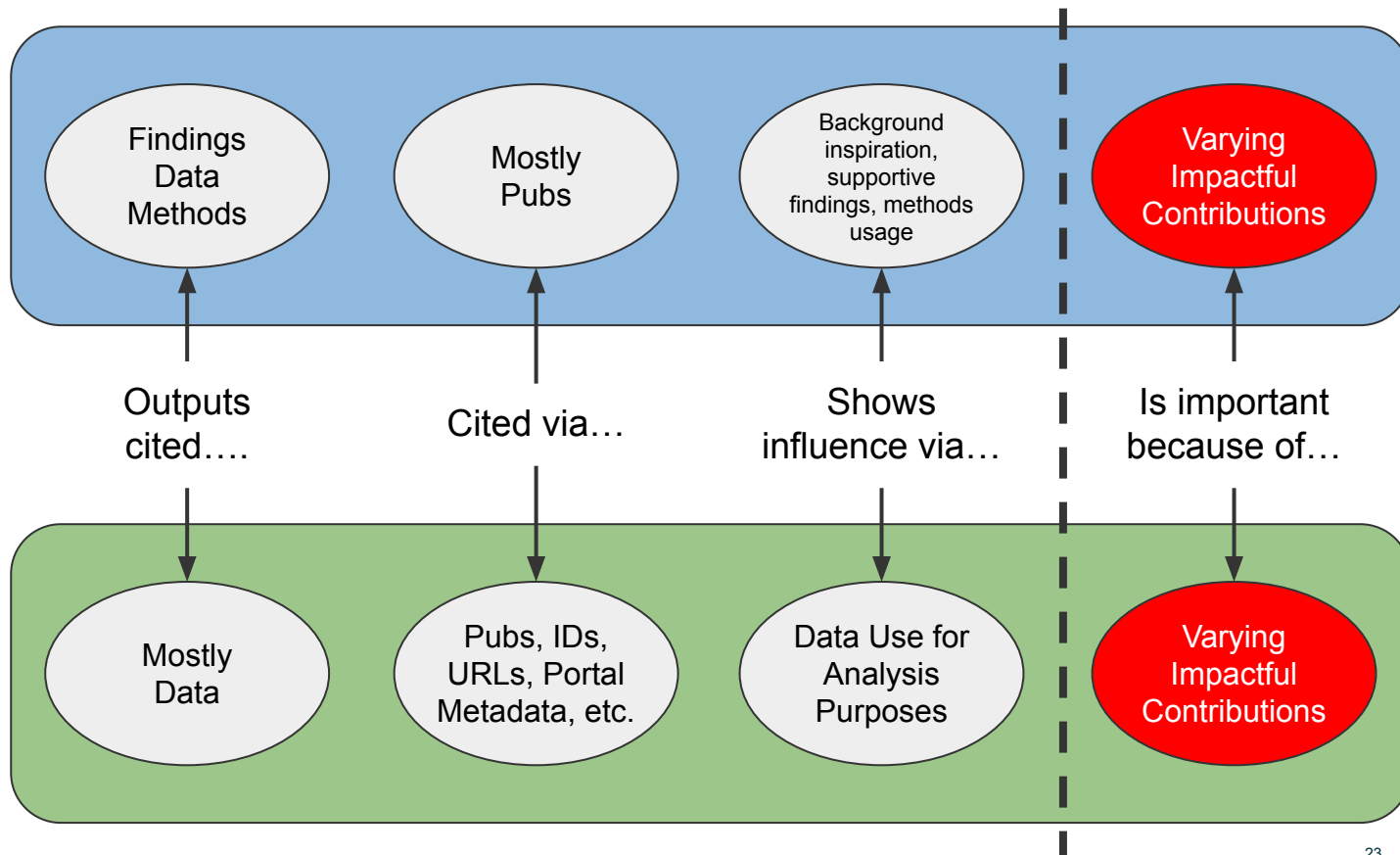


*Soybean Citations:*  
**~5000**

—  
*1KFG Citations*  
**~1800**

# Examples: 1KFG and Soybean

**1KFG**  
~  
**1800**  
citations



**Soybean**  
~  
**5000**  
citations

# Tying it all together

What are scientists doing to create impact?



What outputs are being leveraged by the community?



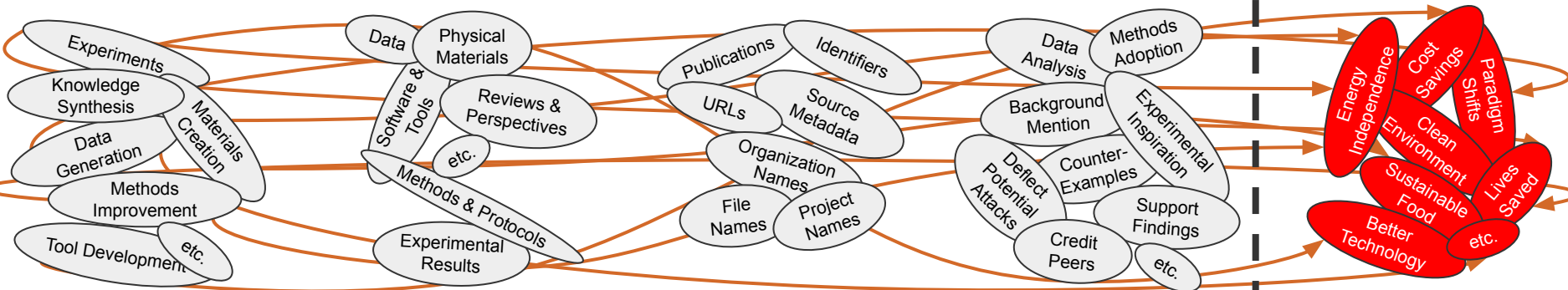
How are people citing those outputs?



What do those citations indicate?



What is the sum total of these impacts?



Scientists don't just run experiments

Scientists produce more than just scientific results

These outputs can be cited in many different ways

These citations indicate varying usages and influence

Paths from citation to impact are not clear or uniform

- *Scientists don't just run experiments*
- *Scientists produce more than just scientific results*
- *These outputs can be cited in many different ways*
- *These citations indicate varying usages and influence*
- *Paths from citation to impact are not clear or uniform*

- 1. Publications are not outputs in themselves, but rather representations of widely variable outputs**
- 2. Citation metrics don't just gloss over how outputs influence downstream studies, but also which outputs are doing the influencing**
- 3. Comprehensive citation pictures require:**
  - a. Cataloging of all institutional outputs
  - b. Cataloging of all citable metadata
  - c. Scalable and reproducible citation classification schemas
- 4. All of the above will require custom tailoring for most organizations**
- 5. Even comprehensive capture doesn't get us all the way to tangible impact narratives**

# Thank You!



U.S. DEPARTMENT OF

**ENERGY**

Office of  
Science



**BERKELEY LAB**

Bringing Science Solutions to the World



**UNIVERSITY  
OF  
CALIFORNIA**

