Make tables with code

How to make fewer mistakes and have more fun

Institute of Virology seminar ✶ 25.09.2023 ✶ Terry Jones

How to publish fewer errors

Using a specific example to illustrate a more general theme

Making data tables? WTF?

Context, motivation, importance.

The many shortcomings of the typical approach.

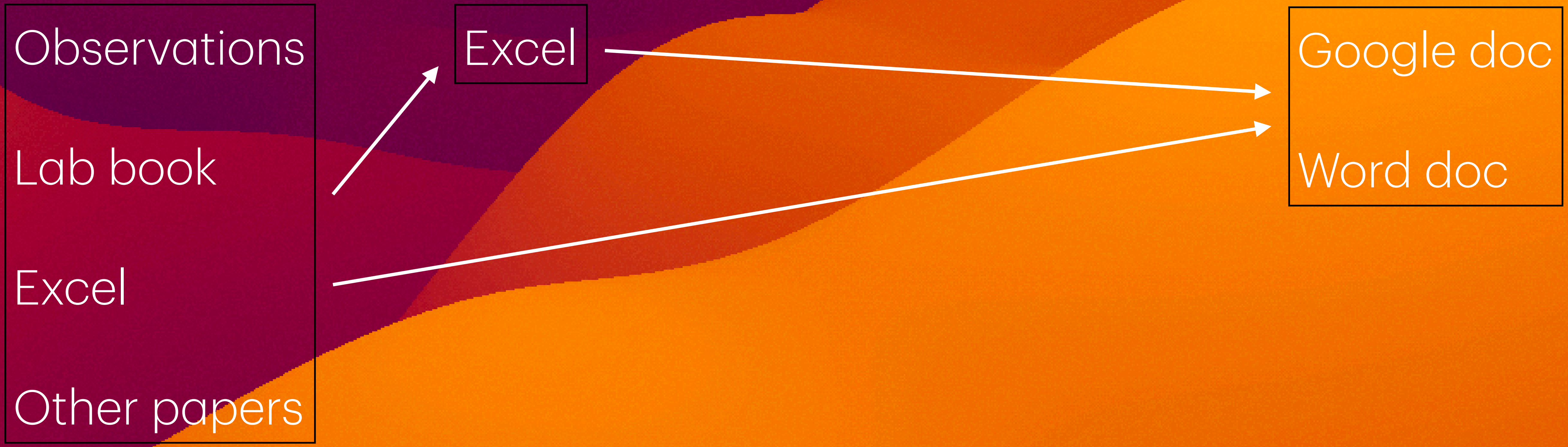
All these problems can be easily fixed.

You can then easily do so much more.

Finally: swagger like a programmer.

The typical path to tables

How we all did it at some point, and most of you still do



Observations

Lab book

Excel

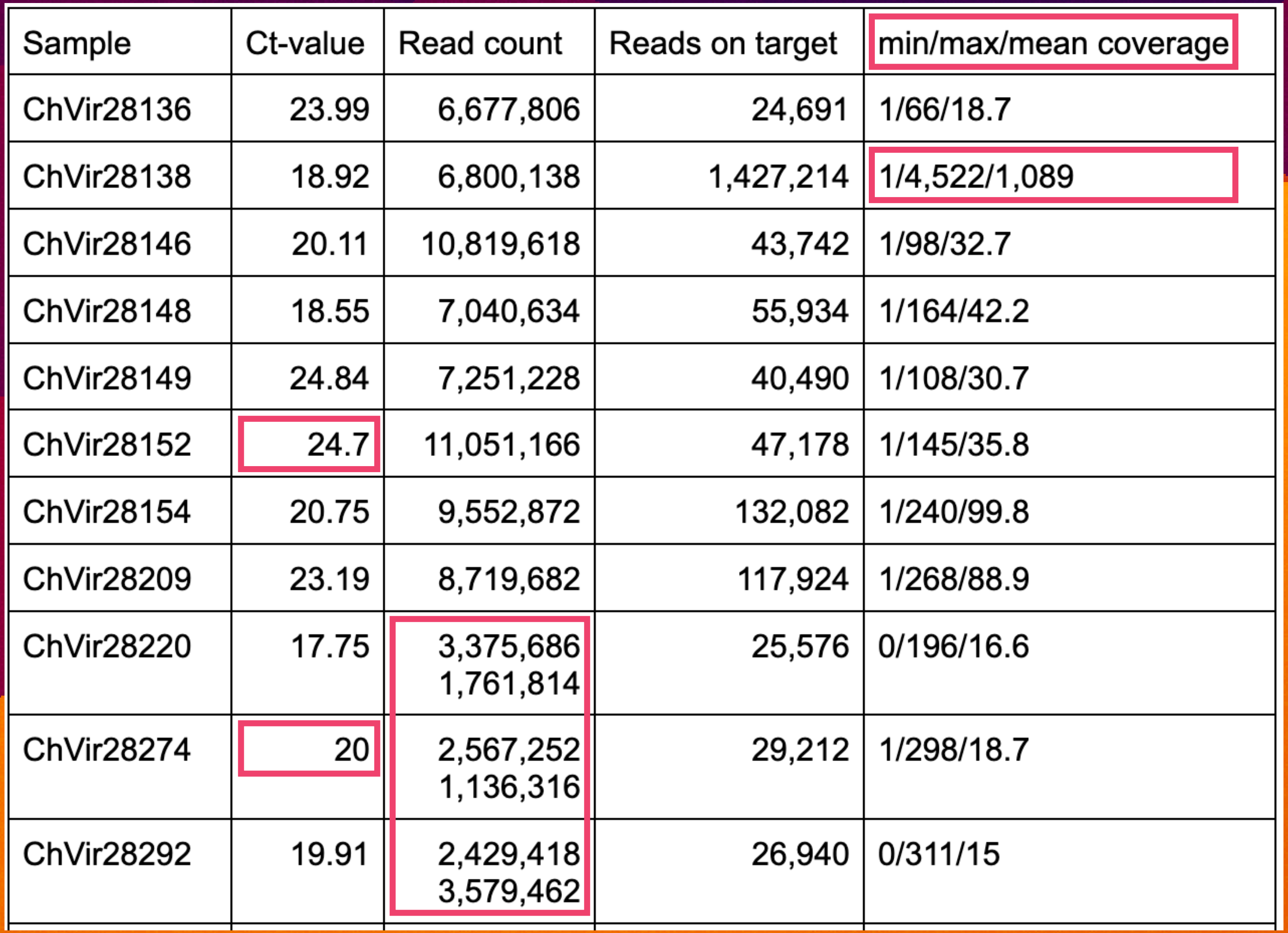
Other papers

Excel

Google doc

Word doc

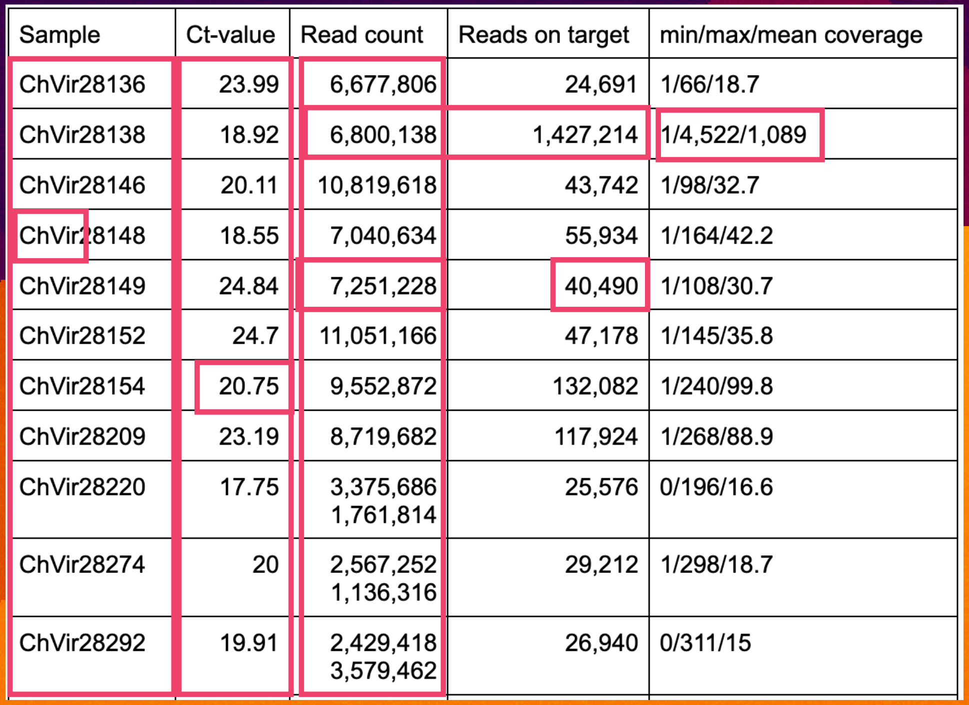
Part of the original table



Sample	Ct-value	Read count	Reads on target	min/max/mean coverage
ChVir28136	23.99	6,677,806	24,691	1/66/18.7
ChVir28138	18.92	6,800,138	1,427,214	1/4,522/1,089
ChVir28146	20.11	10,819,618	43,742	1/98/32.7
ChVir28148	18.55	7,040,634	55,934	1/164/42.2
ChVir28149	24.84	7,251,228	40,490	1/108/30.7
ChVir28152	24.7	11,051,166	47,178	1/145/35.8
ChVir28154	20.75	9,552,872	132,082	1/240/99.8
ChVir28209	23.19	8,719,682	117,924	1/268/88.9
ChVir28220	17.75	3,375,686 1,761,814	25,576	0/196/16.6
ChVir28274	20	2,567,252 1,136,316	29,212	1/298/18.7
ChVir28292	19.91	2,429,418 3,579,462	26,940	0/311/15

Things I can't (easily) check

All of which I have to re-check if the table changes!

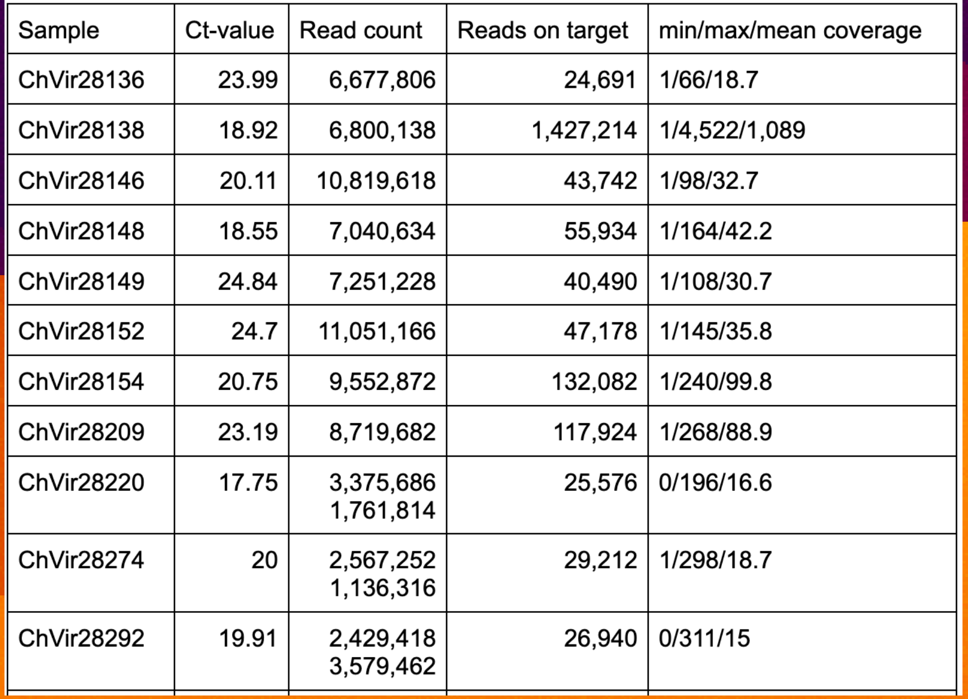


Sample	Ct-value	Read count	Reads on target	min/max/mean coverage
ChVir28136	23.99	6,677,806	24,691	1/66/18.7
ChVir28138	18.92	6,800,138	1,427,214	1/4,522/1,089
ChVir28146	20.11	10,819,618	43,742	1/98/32.7
ChVir28148	18.55	7,040,634	55,934	1/164/42.2
ChVir28149	24.84	7,251,228	40,490	1/108/30.7
ChVir28152	24.7	11,051,166	47,178	1/145/35.8
ChVir28154	20.75	9,552,872	132,082	1/240/99.8
ChVir28209	23.19	8,719,682	117,924	1/268/88.9
ChVir28220	17.75	3,375,686 1,761,814	25,576	0/196/16.6
ChVir28274	20	2,567,252 1,136,316	29,212	1/298/18.7
ChVir28292	19.91	2,429,418 3,579,462	26,940	0/311/15

- The number of rows in the table.
- That all sample ids in our study are in the table.
- That there are no duplicate ids.
- That the sample ids appear in sorted order.
- That various fields are numeric or have expected text (e.g., the ChVir prefix).
- That read counts are always non-negative integers.
- That the min. reads is less than the mean and that the mean is less than the max.
- That the number of reads on target is less than the total number of reads.
- That the Ct values are reasonable.

Things I can't (easily) change

Some I can do in Excel, but with less flexibility



Sample	Ct-value	Read count	Reads on target	min/max/mean coverage
ChVir28136	23.99	6,677,806	24,691	1/66/18.7
ChVir28138	18.92	6,800,138	1,427,214	1/4,522/1,089
ChVir28146	20.11	10,819,618	43,742	1/98/32.7
ChVir28148	18.55	7,040,634	55,934	1/164/42.2
ChVir28149	24.84	7,251,228	40,490	1/108/30.7
ChVir28152	24.7	11,051,166	47,178	1/145/35.8
ChVir28154	20.75	9,552,872	132,082	1/240/99.8
ChVir28209	23.19	8,719,682	117,924	1/268/88.9
ChVir28220	17.75	3,375,686 1,761,814	25,576	0/196/16.6
ChVir28274	20	2,567,252 1,136,316	29,212	1/298/18.7
ChVir28292	19.91	2,429,418 3,579,462	26,940	0/311/15

- The number of decimal places shown.
- Display the total read count.
- Add more metadata (e.g., sample dates).
- Split min/max/mean into three columns.
- Change the order of the min/max/mean column to be min/mean/max.
- Only show rows where the read count is in a certain range.
- Sort the table on read count (or any other column or column combination).
- Produce alternate versions of the table to distribute in other ways (HTML, CSV, etc).
- Produce alternate non-table summaries. E.g., plots.

Time for a demo

Using a text editor and the command line

- Checking data by writing tests.
- Changing your table: extreme (and easy) flexibility with code.
- Producing more outputs.

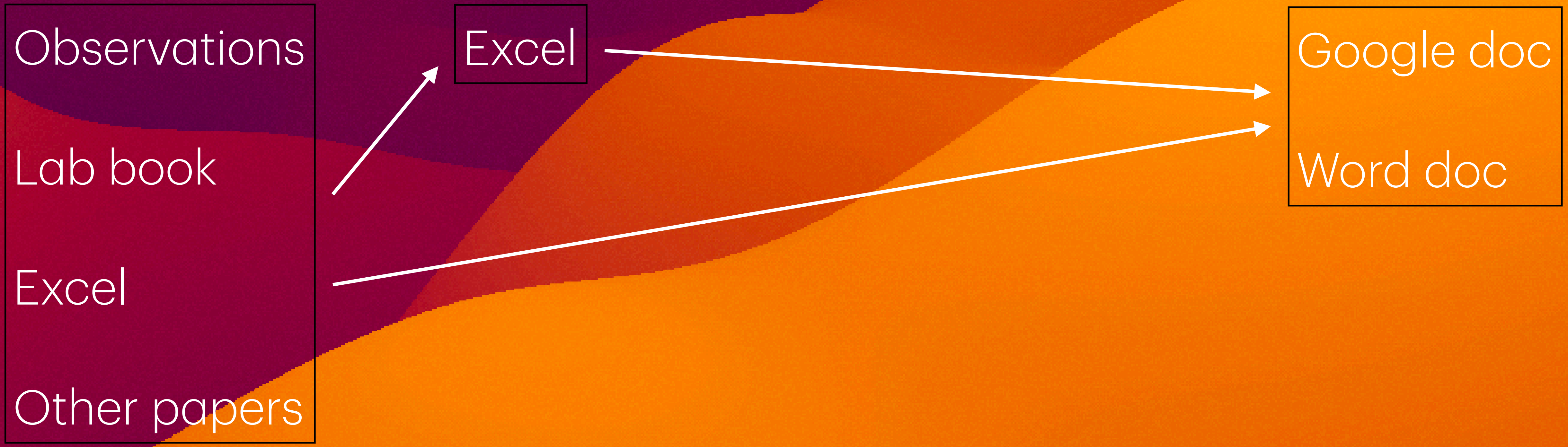
What you should do instead

Six simple steps to enduring glory

- Don’t paste your data directly into a document.
- Instead, put them into a text file.
- Write a simple loop to print the table.
- Write tests.
- Modify your table printing in any way(s) you like.
- Import your table in Excel, copy it, paste it into your document.

The typical path to tables

How we all did it at some point, and most of you still do



Observations

Lab book

Excel

Other papers

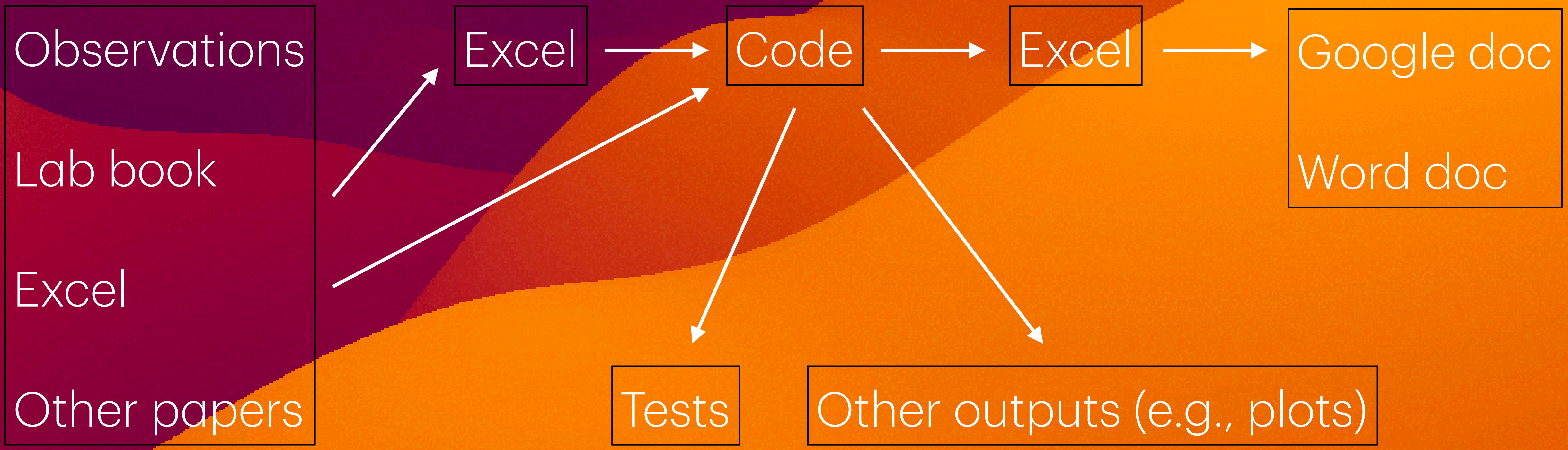
Excel

Google doc

Word doc

A better path to tables

Add intermediate processing with code



Observations

Lab book

Excel

Other papers

Excel

Code

Excel

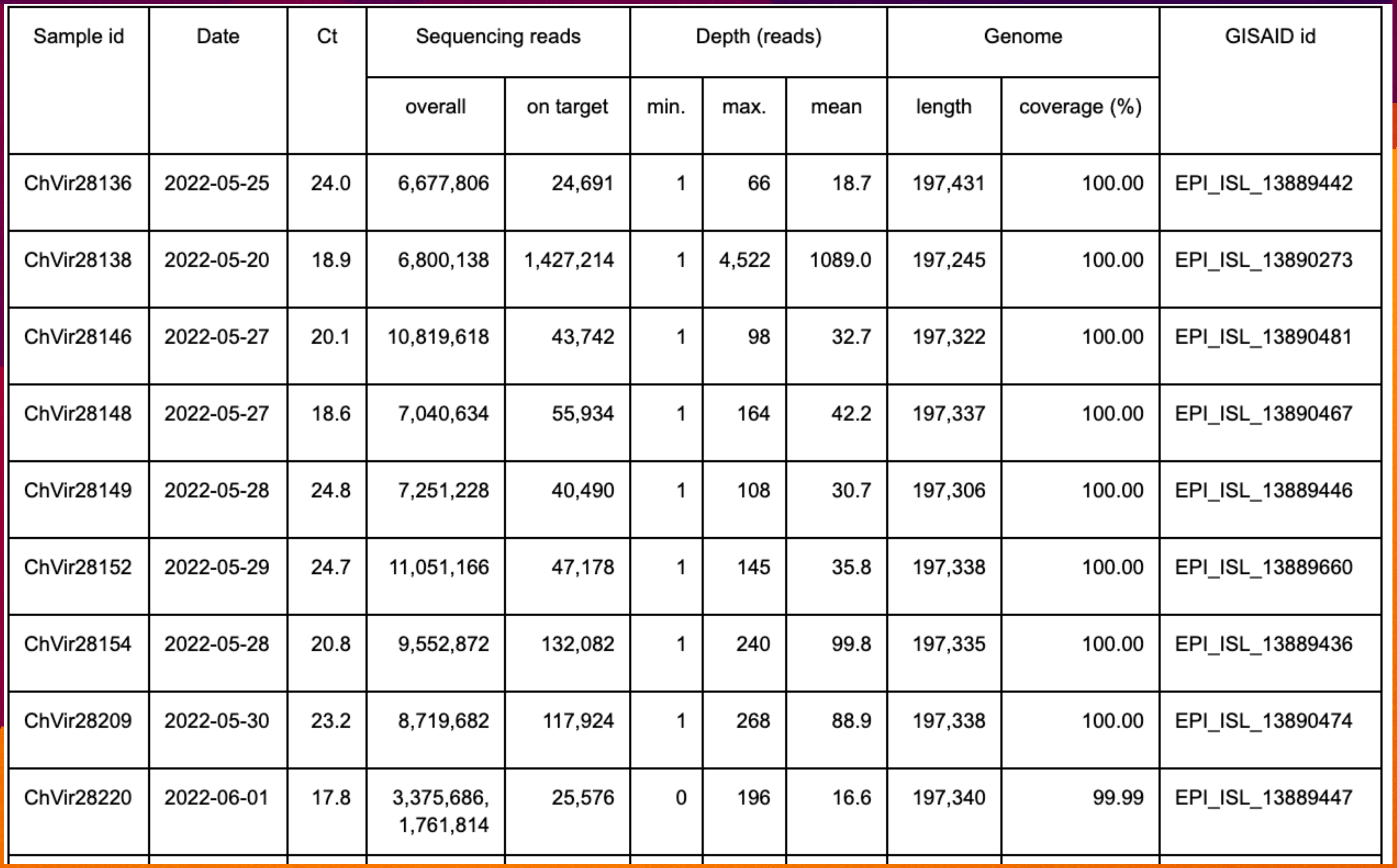
Google doc

Word doc

Tests

Other outputs (e.g., plots)

The finished product



Sample id	Date	Ct	Sequencing reads		Depth (reads)			Genome		GISAID id
			overall	on target	min.	max.	mean	length	coverage (%)	
ChVir28136	2022-05-25	24.0	6,677,806	24,691	1	66	18.7	197,431	100.00	EPI_ISL_13889442
ChVir28138	2022-05-20	18.9	6,800,138	1,427,214	1	4,522	1089.0	197,245	100.00	EPI_ISL_13890273
ChVir28146	2022-05-27	20.1	10,819,618	43,742	1	98	32.7	197,322	100.00	EPI_ISL_13890481
ChVir28148	2022-05-27	18.6	7,040,634	55,934	1	164	42.2	197,337	100.00	EPI_ISL_13890467
ChVir28149	2022-05-28	24.8	7,251,228	40,490	1	108	30.7	197,306	100.00	EPI_ISL_13889446
ChVir28152	2022-05-29	24.7	11,051,166	47,178	1	145	35.8	197,338	100.00	EPI_ISL_13889660
ChVir28154	2022-05-28	20.8	9,552,872	132,082	1	240	99.8	197,335	100.00	EPI_ISL_13889436
ChVir28209	2022-05-30	23.2	8,719,682	117,924	1	268	88.9	197,338	100.00	EPI_ISL_13890474
ChVir28220	2022-06-01	17.8	3,375,686, 1,761,814	25,576	0	196	16.6	197,340	99.99	EPI_ISL_13889447

What this buys you

Solving the problems of checking and changing

You get to check everything.

You get extreme flexibility.

Confidence.

A foundation for doing even more (with code).

Your next steps

To making fewer mistakes and having more fun

Give this a try!

See `https://github.com/VirologyCharite/make-tables-with-code`

Come ask us for help.