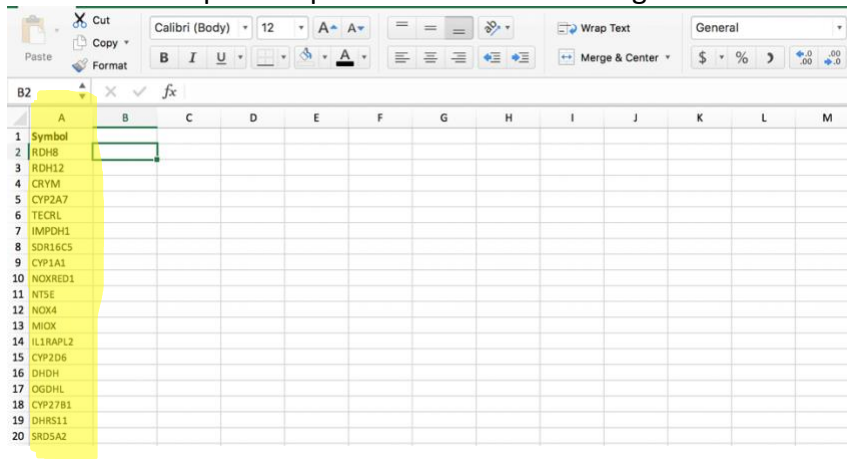


Combining Excel Sheets for Data Analysis – Written July 21, 2020

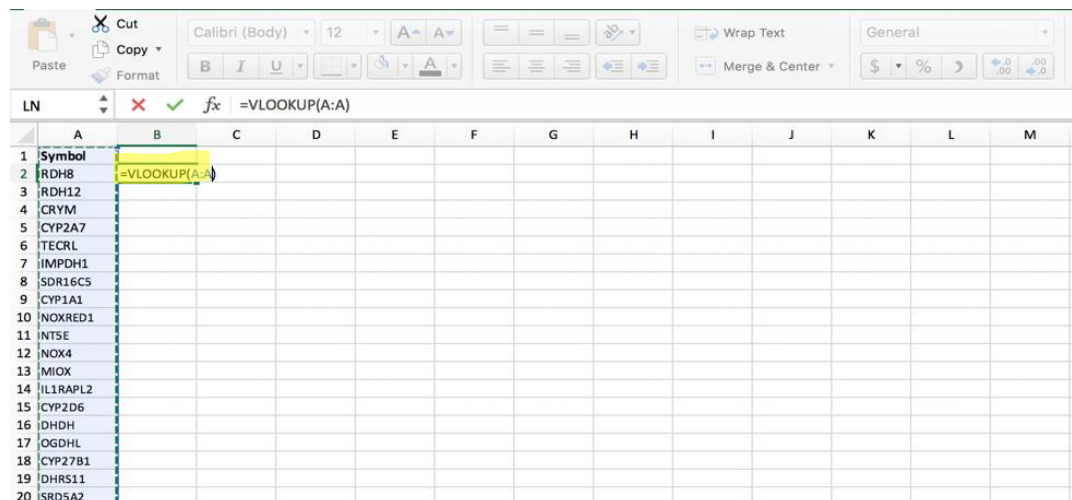
Combining excel sheets is a useful tool for bioinformatics. This guide will show you how to combine a gene list with a database and combine two sheets with multiple columns to pull-out relevant data.

Combining gene lists with a database

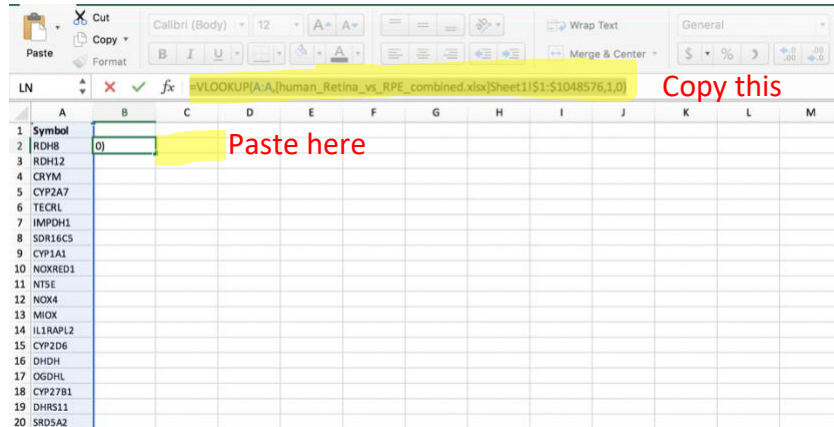
1. Before you begin, make sure both your database and gene list Excel sheets are open. Your database sheet should be formatted so that the data you are searching for is in the first column. For example, I am interested in searching for gene symbols in the database, so I made gene symbols column A.
2. In a new excel sheet, copy and paste your gene list into column A. This first column will be used to pull out pertinent data from a larger source.



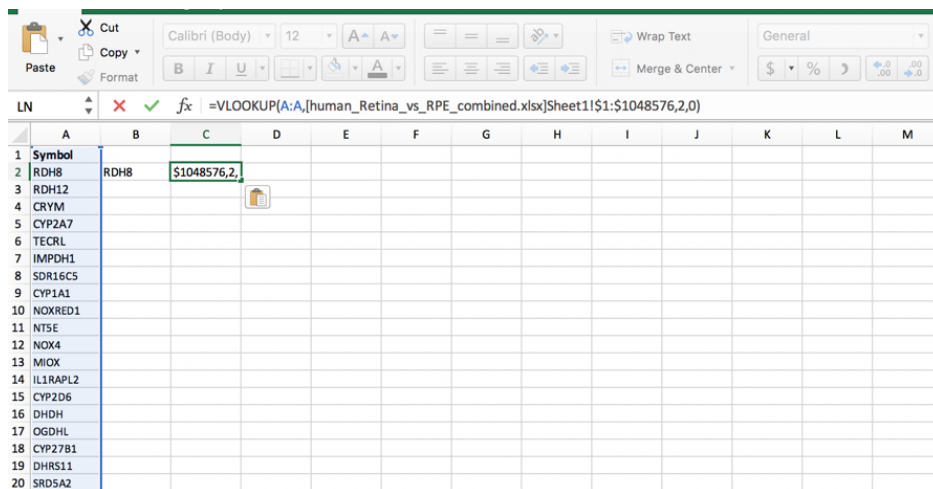
3. In column B type `=VLOOKUP` (or you can select VLOOKUP from the drop-down menu). Select column A so that it becomes highlighted. The formula bar should now read `=VLOOKUP(A:A)`.



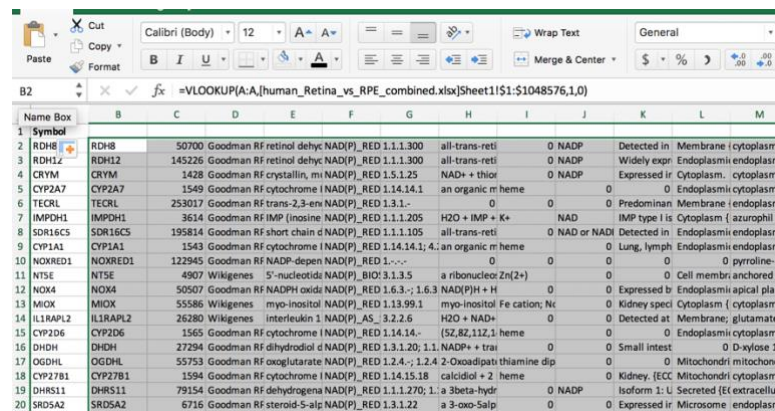
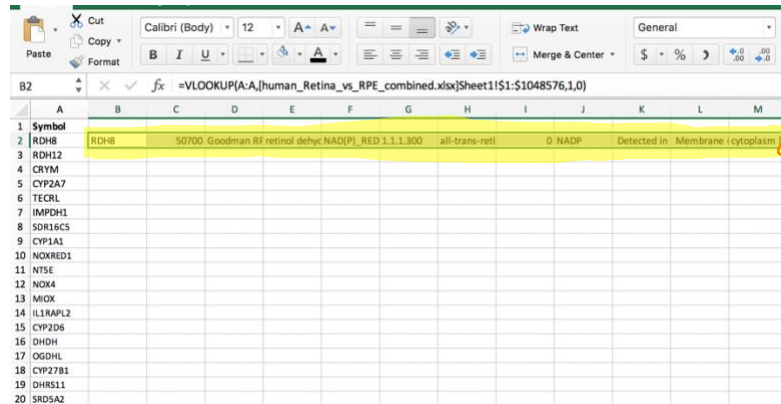
- We will repeat this formula across row 2 to combine relevant information from the database with our gene list. Copy the formula from column B and paste it in the equation bar in the next box in column C.



- Since now we are interested in data from the second column in the database, we will change 1,0 to **2,0** so that the formula bar reads =VLOOKUP(A:A, *name of your database sheet*,**2,0**). Hit enter/return. Continue pasting the formula down the row, adjusting for column number. If you aren't interested in data from a specific column of the database, you can skip that number in the formula. For example, let's say column 3 in the database contains species name and column 4 contains gene ID. If we are interested in gene ID and not species name, we can make the formula =VLOOKUP(A:A, *name of your database sheet*,**4,0**) so that species name is omitted from our combined sheet.

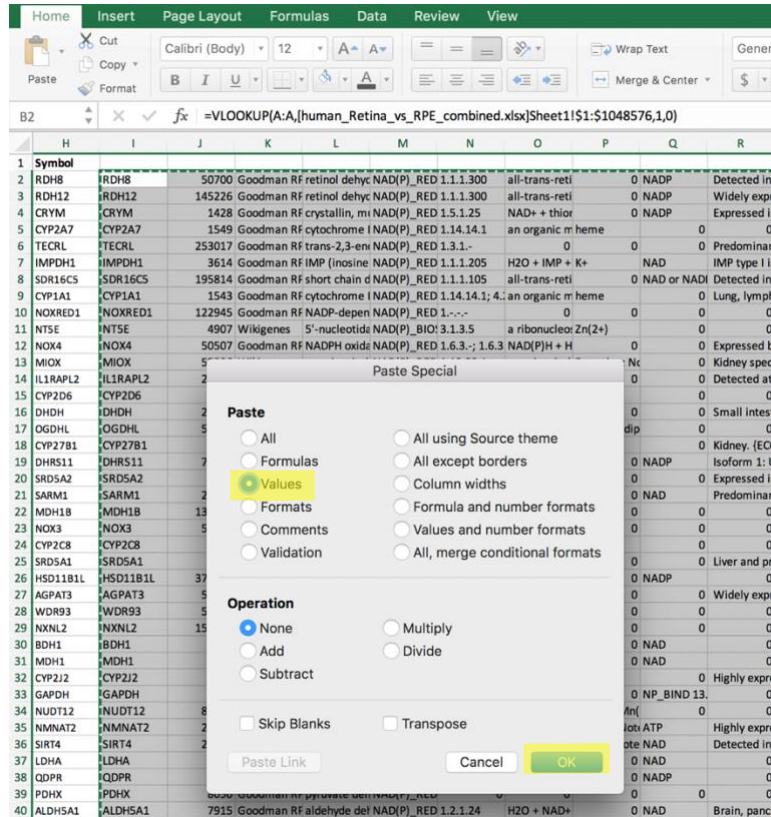


- Once you have the formulas for all your columns of interest, highlight the formula containing row. Click the black plus sign in the right lower corner of your selection and drag until you reach the end of your gene list. This action will automatically copy the formulas for each row.



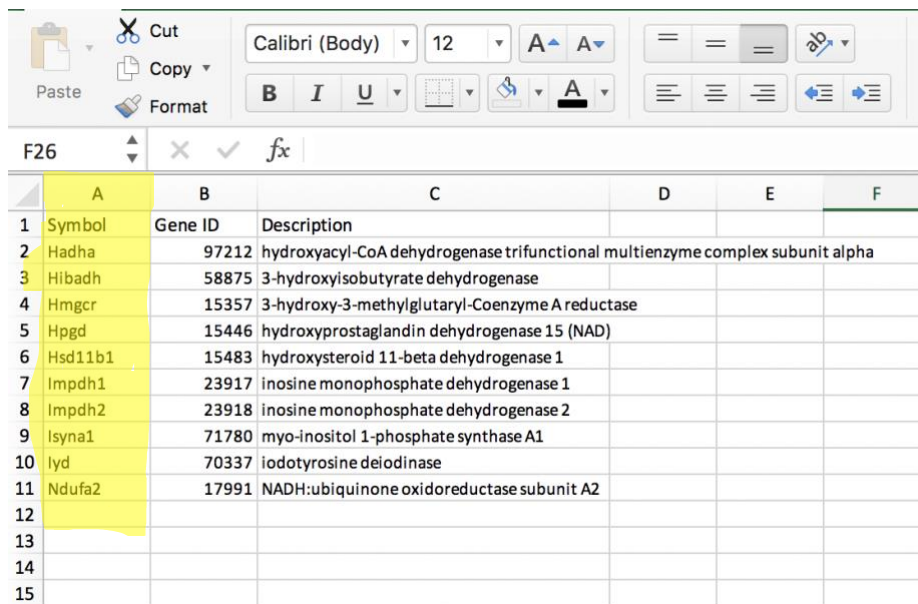
- Now, information from the database is combined with your gene list. If a row in your combined sheet says **#N/A** this may mean the information is not in the database (i.e. the gene is not in the database). Always double check the database by using the find tool in excel for your potential missing gene (control + F).

- Once your combined sheet looks the way you like, select the sheet and copy it. Control + click (or use the top menu go to edit>paste special) and select values only. This will allow you to filter and sort your sheet.



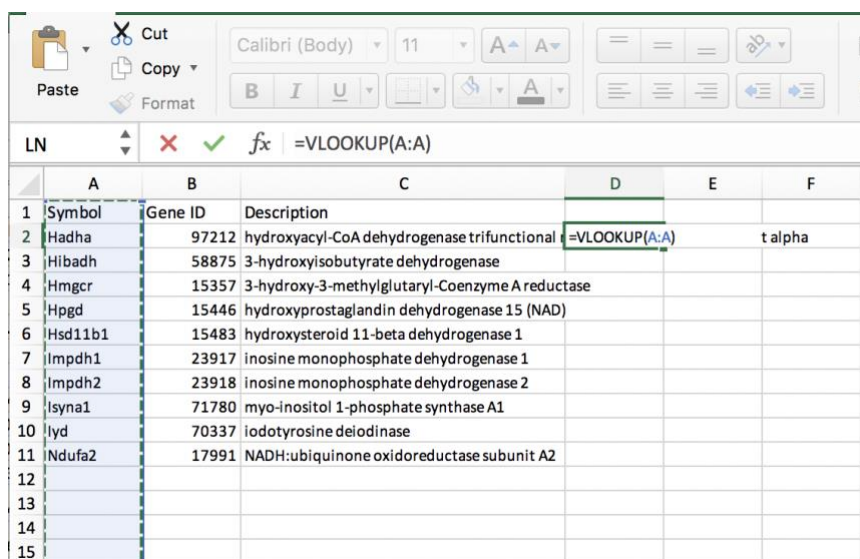
Combining sheets with multiple columns

1. Set up your data as before, with what you are searching for in column A. Below you see my data. This time, instead of just a gene list comprised of one column, my data has multiple columns.



	A	B	C	D	E	F
1	Symbol	Gene ID	Description			
2	Hadha	97212	hydroxyacyl-CoA dehydrogenase trifunctional multienzyme complex subunit alpha			
3	Hibadh	58875	3-hydroxyisobutyrate dehydrogenase			
4	Hmgcr	15357	3-hydroxy-3-methylglutaryl-Coenzyme A reductase			
5	Hpgd	15446	hydroxyprostaglandin dehydrogenase 15 (NAD)			
6	Hsd11b1	15483	hydroxysteroid 11-beta dehydrogenase 1			
7	Impdh1	23917	inosine monophosphate dehydrogenase 1			
8	Impdh2	23918	inosine monophosphate dehydrogenase 2			
9	Isyna1	71780	myo-inositol 1-phosphate synthase A1			
10	Iyd	70337	iodotyrosine deiodinase			
11	Ndufa2	17991	NADH:ubiquinone oxidoreductase subunit A2			
12						
13						
14						
15						

2. In the next free column (column D in this case) type **=VLOOKUP** (or you can select VLOOKUP from the drop-down menu). Select column A so that it becomes highlighted. The formula bar should now read **=VLOOKUP(A:A)**.



	A	B	C	D	E	F
1	Symbol	Gene ID	Description			
2	Hadha	97212	hydroxyacyl-CoA dehydrogenase trifunctional multienzyme complex subunit alpha	=VLOOKUP(A:A)		t alpha
3	Hibadh	58875	3-hydroxyisobutyrate dehydrogenase			
4	Hmgcr	15357	3-hydroxy-3-methylglutaryl-Coenzyme A reductase			
5	Hpgd	15446	hydroxyprostaglandin dehydrogenase 15 (NAD)			
6	Hsd11b1	15483	hydroxysteroid 11-beta dehydrogenase 1			
7	Impdh1	23917	inosine monophosphate dehydrogenase 1			
8	Impdh2	23918	inosine monophosphate dehydrogenase 2			
9	Isyna1	71780	myo-inositol 1-phosphate synthase A1			
10	Iyd	70337	iodotyrosine deiodinase			
11	Ndufa2	17991	NADH:ubiquinone oxidoreductase subunit A2			
12						
13						
14						
15						

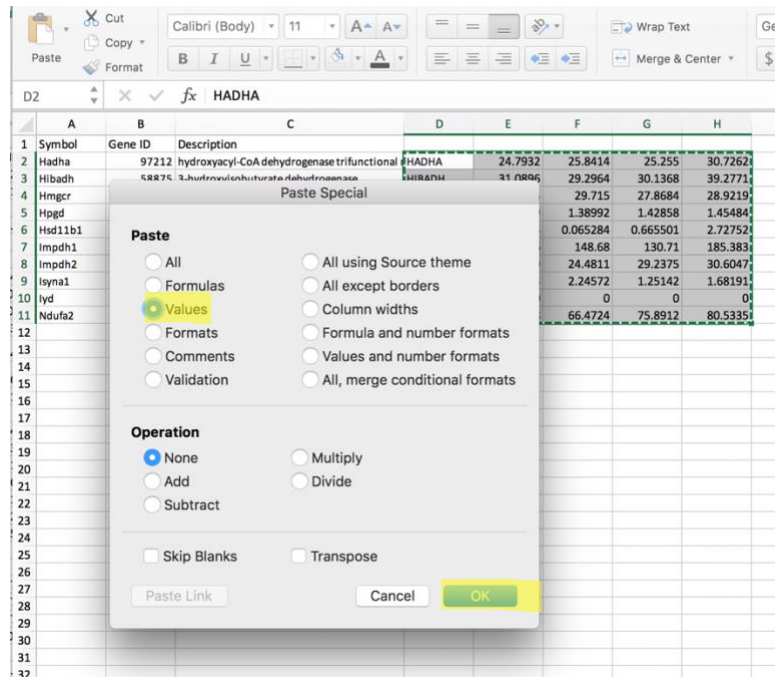
- Next, we will use our database excel sheet. Click the square icon in the left upper corner to highlight the entire sheet. The formula bar should now read **=VLOOKUP(A:A, name of your database sheet)**.

LN	A	B	C	D	E	F	G	H	I
1	Symbol	RETINA_MAI	RETINA_MAI	RETINA_MAI	RETINA_MAI	RETINA_NAS	RETINA_NAS	RETINA_NAS	RETINA_NAS
2	AASS	2.64457	4.37614	3.01459	4.12927	2.60464	4.36765	4.36492	6.38531
3	ABCC4	1.12387	1.75	0.517182	1.4342	1.09939	0.912626	0.908203	0.918654
4	ADH1A	0	0	0	0	0	0	0.0397028	0
5	ADH1B	0.548409	0.258844	0.113098	0.284036	0.577682	0.139915	0.0420607	0.141861
6	ADH1C	0	0.0636251	0	0	0	0.0638822	0.202614	0
7	ADH4	0	0	0	0.0525494	0.0606815	0	0.058362	0.146099
8	ADH5	44.2364	38.4765	46.8828	43.6464	35.9244	33.5086	38.2223	34.9814
9	ADH6	0.0325261	0.0337073	0	0	0.0298723	0	0	0.137436
10	ADH7	0.0622777	0	0.0242901	0	0	0.0431161	0.0546767	0
11	ADHFE1	3.55453	7.15227	3.90703	5.84323	4.65213	7.26958	6.27453	8.50871
12	AFMID	4.65715	2.93316	3.25128	3.72629	3.28853	2.08379	3.08794	2.51183
13	AGPAT3	63.9932	67.3111	70.2397	64.2906	78.7508	72.0983	72.8659	56.171
14	AGPAT4	1.71719	1.40104	1.17242	1.53086	0.990251	1.22065	1.23535	1.76151
15	AHCY	32.457	30.1983	32.9806	33.824	30.8339	26.5717	35.6773	28.5235

- Return to your data excel sheet. Since I am interested in pulling out genes from the gene list that are also in the database I will type **,1,0** in the formula bar so that it now reads **=VLOOKUP(A:A, name of your database sheet,1,0)**. Hit enter/return.

LN	A	B	C	D	E	F
1	Symbol	Gene ID	Description			
2	Hadha	97212	hydroxyacyl-CoA dehydrogenase trifunctional	=VLOOKUP(A:A,[database.xlsx]Sheet1!\$1:\$1048576,1,0)		
3	Hibadh	58875	3-hydroxyisobutyrate dehydrogenase			
4	Hmgcr	15357	3-hydroxy-3-methylglutaryl-Coenzyme A reductase			
5	Hpgd	15446	hydroxyprostaglandin dehydrogenase 15 (NAD)			
6	Hsd11b1	15483	hydroxysteroid 11-beta dehydrogenase 1			
7	Impdh1	23917	inosine monophosphate dehydrogenase 1			
8	Impdh2	23918	inosine monophosphate dehydrogenase 2			
9	Isyna1	71780	myo-inositol 1-phosphate synthase A1			
10	lyd	70337	iodotyrosine deiodinase			
11	Ndufa2	17991	NADH:ubiquinone oxidoreductase subunit A2			
12						
13						
14						
15						

7. Once your combined sheet looks the way you like select the sheet and copy it. Control + click (or use the top menu and go to edit>paste special) and select values only. This will allow you to filter and sort your sheet.



If you have any questions or want clarification, please email me at sherfinski@marshall.edu