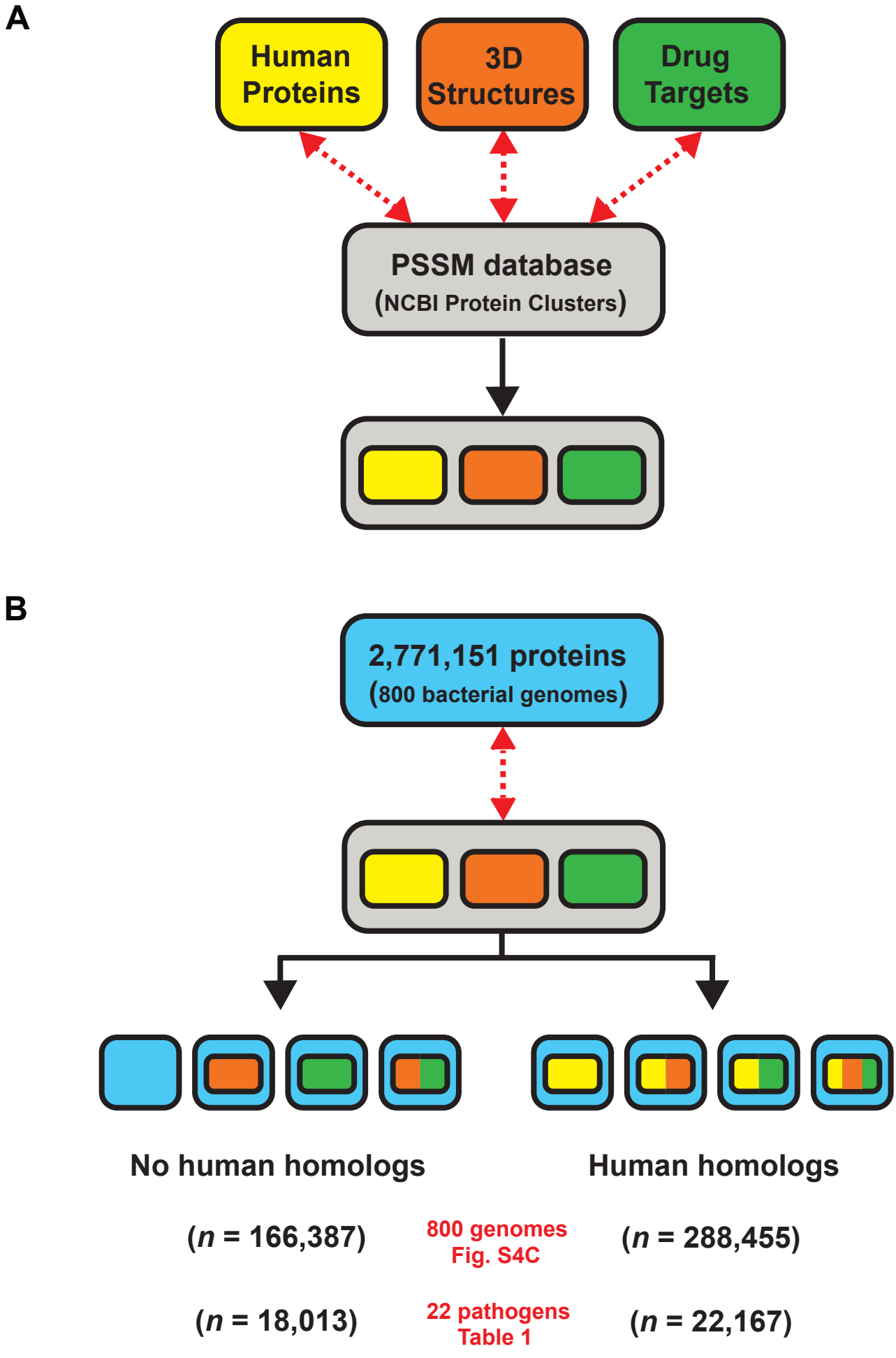
Fig. S4. Additional information supporting the use case "Application to Annotation Driven by Data Integration: Drug and Vaccine Targeting". (A, B) Experimental design for annotating bacterial genes with drug targeting attributes. Black arrows depict the direction of the workflow, while red dashed arrows illustrate reverse position specific blast (RPSBLAST) searches [1] (all significant matches with an E-value cutoff of 0.001). The process for annotating genes consists of two steps. (A) In the first step, proteins encoded within the human genome (yellow box, see NCBI ftp site: ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.3/, [2]), proteins with published 3-D structure information (orange box, [3]), and proteins from a database of annotated drug targets (green box, [4, 5]) were searched against a protein family database of position-specific scoring matrices (PSSMs, gray box, [6]). This resulted in the construction of a diverse set of proteins with highly conserved domains, with each protein containing at least one of the predefined drug targeting attributes. (B) In the second step, protein sequences encoded within 800 bacterial genomes (light blue box) were searched against the set of proteins created in A, matching bacterial proteins with proteins containing at least one of the predefined drug targeting attributes. Obtained bacterial proteins were primarily categorized into those without significant similarity to human proteins (36.6%) and those with significant matches to one or more domains within human proteins (63.4%). All proteins were then further distinguished as containing one, two or three of the predefined drug targeting attributes. (C) Attributes and distribution of the "reverse annotated" proteins across the genomes of 22 genera containing NIAID Category A, B and C priority microbial pathogens (800 total genomes). Target class information as follows: first column indicates whether a human protein homologue exists for the protein, with "H" indicating 'yes'; second column indicates whether a drug target homolog exists, with "A" indicating an approved drug target homolog and "D" indicating an experimental drug target homolog; third column indicates whether a 3-D structure exists for a protein homolog. Note: attributes of the "reverse annotated" proteins for an exemplar species from each genus (Table 1) provides an idea of the portion per genome of identified candidate drug targets.

1. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: **CDD: a database of conserved domain alignments with links to domain three-dimensional structure**. *Nucleic Acids Res* 2002, **30**(1):281-283.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.
4. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V *et al*: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs**. *Nucleic Acids Res* 2011, **39**(Database issue):D1035-1041.
5. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Res* 2008, **36**(Database issue):D901-906.
6. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins**. *Proc Natl Acad Sci U S A* 1987, **84**(13):4355-4358.


A
Human Proteins
3D Structures
Drug Targets
PSSM database
(NCBI Protein Clusters)
B
2,771,151 proteins
(800 bacterial genomes)
No human homologs
Human homologs
(n = 166,387)
800 genomes
Fig. S4C
(n = 288,455)
(n = 18,013)
22 pathogens
Table 1
(n = 22,167)

C

Target Class by Homology [1]			Organism Genus [2]																					
Human	Drug Target	3D	Bac	Bar	Bor	Bru	Bur	Cam	Chl	Clo	Cox	Ehr	Esc	Fra	Hel	Lis	Myc	Ric	Sal	Shi	Sta	Str	Vib	Yer
			3554	301	309	1323	4537	760	124	3469	48	82	1644	195	594	523	1141	722	751	2034	831	1092	3975	3523
		S	7511	603	560	2247	7612	1807	223	8569	117	171	2241	422	1398	1293	3320	1282	1098	2313	2039	2861	6354	5310
	A	S	1293	84	90	326	1282	384	52	1465	15	22	237	58	305	173	939	179	111	263	329	465	879	569
	D		5				14	2		6			24	1	2				8	11		3	39	44
	D	S	7696	432	516	2560	11663	1355	180	10959	88	107	2003	361	1113	1262	4260	820	1011	1998	1717	2604	8093	5027
H			1126	80	53	415	1153	182	28	1206	17	33	274	57	176	139	408	333	133	265	231	298	765	749
H		S	4734	659	837	1753	4866	1213	357	5132	138	285	1066	385	1090	789	2366	1799	530	1081	1535	1982	3704	2778
H	A		586	60	21	201	1145	89	3	331	21	30	174	86	51	62	278	171	76	184	148	82	292	293
H	A	S	21645	1801	1601	7299	25755	4403	911	23875	462	673	3496	1428	3730	3196	19974	4088	1696	3536	6039	8078	13822	9134
H	D		59	10	9	18	39	19	8	87	2	6	22	6	18	10	33	31	8	18	30	17	43	27
H	D	S	8858	836	1014	2729	8739	2095	495	10896	202	332	1748	642	1879	1515	5439	2058	813	1771	2552	3925	6828	4343

[1] N, protein has no significant similarity to a human protein; H, protein has significant similarity to a human protein; A, protein has a significant similarity to an approved drug target; D, protein has significant similarity to a drug target under experimental testing; S, protein has significant similarity to a protein with associated 3D structure in the Protein Data Bank.

[2] BAC, *Bacillus*; BAR, *Bartonella*; BOR, *Borrelia*; BRU, *Brucella*; BUR, *Burkholderia*; CAM, *Campylobacter*; CHL, *Chlamydophila*; CLO, *Clostridium*; COX, *Coxiella*; EHR, *Ehrlichia*; ESC, *Escherichia*; FRA, *Francisella*; HEL, *Helicobacter*; LIS, *Listeria*; MYC, *Mycobacterium*; RIC, *Rickettsia*; SAL, *Salmonella*; SHI, *Shigella*; STA, *Staphylococcus*; STR, *Streptococcus*; VIB, *Vibrio*; YER, *Yersinia.*