Protein domains search pipeline

~3k datasets with assembled contigs

Inputs:

1) the (assembled) contigs
2) protein domain models
   - ✗ ~300k CDD, no PFAM, POGs, etc. to avoid later work in filtering overlapping hits
   - ✓ ~2k “viral”-enriched set from Rodney

Protein domains search pipeline

~3k datasets with assembled contigs

Inputs:
1) the (assembled) contigs
2) protein domain models
   - ✗ ~300k CDD, no PFAM, POGs, etc. to avoid later work in filtering overlapping hits
   - ✓ ~2k “viral”-enriched set from Rodney

?Dataset clustering?

~100k datasets with non-assembled reads

Inputs:
1) the (non-assembled) reads
2) protein domain models
   - but...how to connect?
     i. k-mers – build our own method?
     ii. min-hash, if it would work?

Protein domains search pipeline

~100k datasets with non-assembled reads

Inputs:

1) the (non-assembled) reads
2) protein domain models
   - but...how to connect?
     i. k-mers – build our own method?
     ii. min-hash, if it would work?

Raw reads

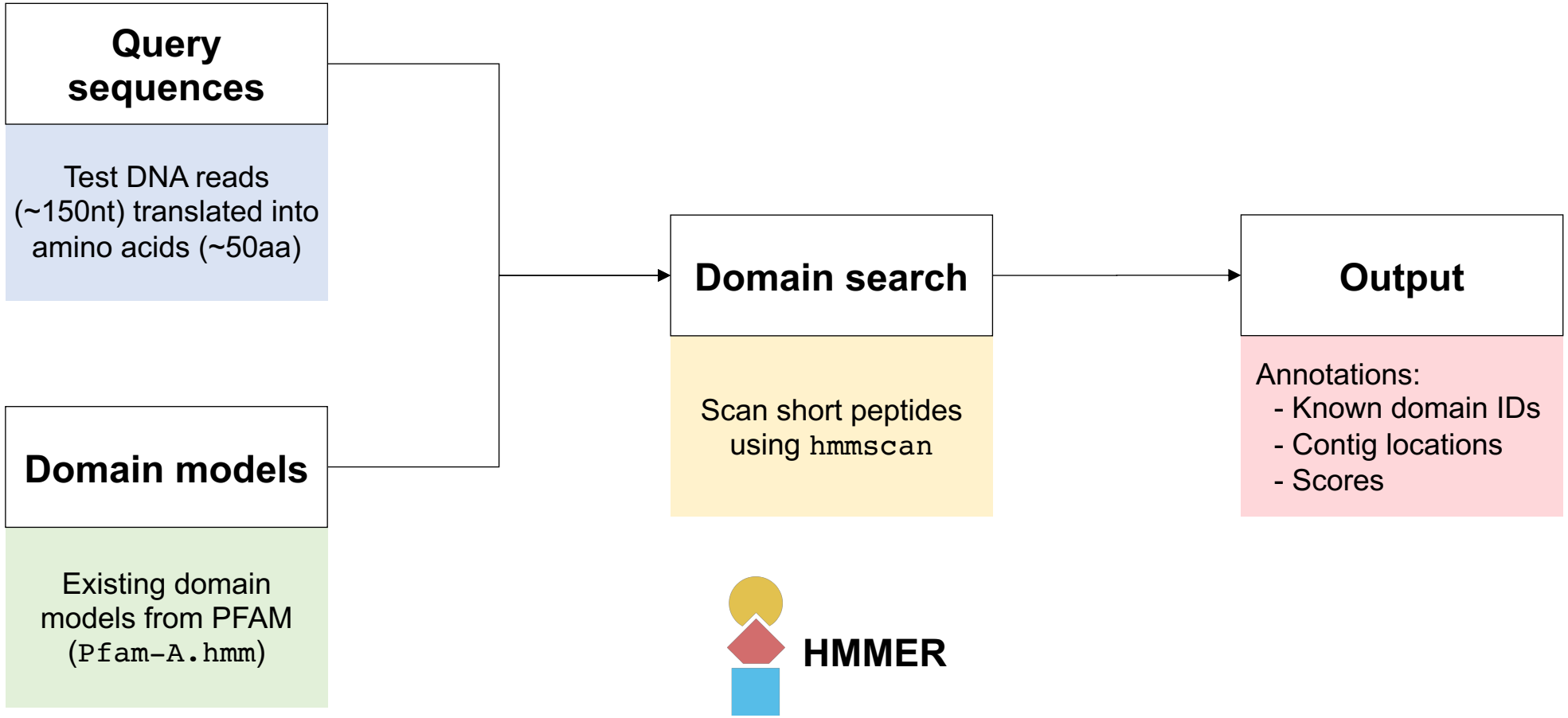
6frame TLN

Sketch query

Sketch against Rodney's Virus CDD

Compare w/ rpstbln output

Searching domains in translated raw reads



Domains found in small peptides (HHV-1)

Alpha_TIF(PF02232)
Collagen(PF01391)
DNA_pack_C(PF02499)
DNA_pack_N(PF02500)
DNA_pol_B_exo1(PF03104)
DNA_pol_B(PF00136)
dUTPase(PF00692)
Fusion_gly_K(PF01621)
Glycoprot_B_PH1(PF17416)
Glycoprot_B_PH2(PF17417)
Glycoprotein_B(PF00606)
GlyL_C(PF12524)
Herpes_env(PF01673)
Herpes_gE(PF02480)
Herpes_gI(PF01688)
Herpes_glycop_D(PF01537)
Herpes_glycop(PF01528)
Herpes_Helicase(PF02689)
Herpes_HEPA(PF03324)
Herpes_ICP4_C(PF03585)
Herpes_ICP4_N(PF03584)
Herpes_IE68(PF02479)
Herpes_MCP(PF03122)
Herpes_ori_bp(PF02399)
Herpes_teg_N(PF04843)

Herpes_TK(PF00693)
Herpes_U34(PF04541)
Herpes_UL1(PF05259)
Herpes_UL14(PF03580)
Herpes_UL16(PF03044)
Herpes_UL17(PF04559)
Herpes_UL20(PF04544)
Herpes_UL21(PF03252)
Herpes_UL24(PF01646)
Herpes_UL25(PF01499)
Herpes_UL3(PF03369)
Herpes_UL31(PF02718)
Herpes_UL33(PF03581)
Herpes_UL35(PF04496)
Herpes_UL36(PF03586)
Herpes_UL37_1(PF03970)
Herpes_UL4(PF03277)
Herpes_UL42(PF02282)
Herpes_UL43(PF05072)
Herpes_UL46(PF03387)
Herpes_UL49_2(PF04823)
Herpes_UL51(PF04540)
Herpes_UL52(PF03121)
Herpes_UL55(PF04537)
Herpes_UL56(PF04534)

Herpes_UL6(PF01763)
Herpes_UL7(PF01677)
Herpes_US12(PF05363)
Herpes_US9(PF06072)
Herpes_V23(PF01802)
Herpes_VP19C(PF03327)
HHV-1_VABD(PF16852)
Marek_A(PF02124)
NAD_binding_2(PF03446)
Peptidase_S21(PF00716)
Pkinase_Tyr(PF07714)
Pkinase(PF00069)
PP1c_bdg(PF10488)
PRTP(PF01366)
Ribonuc_red_lgC(PF02867)
UDG(PF03167)
UL11(PF11094)
UL45(PF05473)
US2(PF02476)
Viral_alk_exo(PF01771)
Viral_DNA_bp(PF00747)
XPG_I(PF00867)
zf-C3HC4_2(PF13923)
zf-C3HC4(PF00097)
zf-RING_2(PF13639)