modERN call

White Lab

2016-10-13

modERN *D. mel.* ChIP-seq datasets

253 lines complete data sets:

- ab, **Abd-B**, achi, acj6[!], ADD1-mimic, Antp-mimic, Atf-2, Atf3, az2, Bab2, bcd, br, brk, bsh, Bteb2, btn, **cad**, Camta-mimic, cato, CG10274, CG10462, CG10565, CG10631, CG10654, CG11398, CG11723, CG11762, CG11902, CG12104, CG12155, CG12236, CG12744, CG12942, CG13123, CG13624, CG13775, CG14965, CG15073, CG1602, CG1620, CG1647, CG16815, CG16863, CG1792, CG180111, CG1832, CG18476, CG18764, CG2116, CG2120, CG30403, CG30431, CG3065, CG31388, CG31627, CG3163, CG32206, CG32264, CG33213, CG3838, CG3919, CG3995, CG4282, CG4318, CG4402(aka CG34406), CG4424, CG4617, CG4820, CG4854, CG5204, CG5245, CG6765, CG6792(aka Plzf), CG7045, CG7368, CG7556, CG7786, CG8089, CG8319, CG8944, CG9305, CG9609, CG9727, CG9876, CHES-1-like, chif*, chn, Chrac-16, **cic**, Clk, cnc, corto, crc-mimic, crebA-mimic, crg-1, **crp**[!], cyc, da, **dac***, Dad, **Dfd**, Dif, Dip3, disco, dl*, dm, dpn(mimic), dsf, dsx, E(bx), **EcR**, Eip75B-MiMIC, Eip78C, Eip93F, Elba2, emc, ems, en, ERR, esn, E(spl)m3, E(spl)mγ-HLH, Ets21C, Ets65A(ets3), Ets97D, E(var)3-9, **eve**, **exd**, ey, eyg, E(z), fkh, FoxP, foxo(MiMIC), fru-mimic, ftz-f1, fu2, GATAd, gcm2, gfzf, grh, **grn**, gro, **h**, her, HLH54F, HLHm7, HmgD, Hmx-MiMIC, Hnf4, Hr38, Hr39, **Hr4**[!], Hr46, Hr51, **Hr78**, Hr83, Hr96, hsf, ind, insv, jim, **jing**, Jra*, jumu-mimic, kay, kn, Kr, lbe, lilli , **lola***, Lpt, luna-mimic, lz, Mad, maf-s, mam-mimic, Max, med, meics, Mes4, Met, Mio, Mnt, mod(mdg4), myb, **N**, NC2alpha, NC2beta, net Neu2, NK7.1, **nmo**, odd, OdsH[!], org-1, ovo, p53-mimic, pb, pdm3, **pdp1**(mimic), pho, Pif1A, Pif1B, pita, pnt-MiMIC, psq[!], pum-mimic, Rel, repo, REPTOR, rgr, sage, salr, sens, shn, side-mimic, sima, six4, slou, slp2, Smox, Sox102F, Sox14, Sox15, Sry-delta, **Stat92E**, Su(H), su(Hw), su(var)2-10-RH, Su(var)3-7, sv, svp, tai, TFAM, tin, **tio**, tj, **tll**, toe, topi, trh, trl-mimic, tup, tx, Usf, **usp**, vfl, vri, Vsx2, woc, Xbp1, YL-1, Zfh2, ZnT49B.

- **14 already submitted by modENCODE**
- 239 / 300 for modERN
  - 29 / 90 for FY (+26)

* multiple isoforms run
! Multiple time-points collected
XX: verifying data with rerun.
204: Released by DCC
0: Ready for release by DCC
4: Ab validation required for release

Current *D. mel.* ChIP-seq

27 new lines being revalidated.

MiMIC being expanded for ChIP-seq (CG16779, CG9727, snoo, dsx†)

87 tagged lines being expanded for ChIP-seq
Target Stages:

Embryo: ac, ash1‼, ATbp, Beaf-32, bigmax, BtbVII, caup, Cdc5, CG10543‼, CG1233, CG13894, CG14962, CG15011, CG15514, CG15601, CG15812, CG17181, CG17568, CG18599, CG3281, CG45071, CG5180, CG6254, CG6654, CG6683‼, CG6813, CG7056, CG7271^, CG7928, CG7987, CG8281, CG8359, CG9797, CG9817‼, CG9948, CrebB‼, CTCF, D19B, dalao, E(spl)m5^, E(spl)m-beta, fd96Cb‼, Fer1, Fer2, Fer3, Hand, Hey, HmgZ‼, kni, lab, l(3)neo38, mor, MTF-1‼, nau, Nfl‼, oc, run, sc, sry-Beta^, trem, unpg, Vsx1, Z, Zif,.

W3L: dmrt93B, HLH106§.

WPP: CG14655, CG17803, CG6276, CG6808, CG6854, CG8145, CG8301, CG9139(Rabex-5), D1, salr‼, Sp1‼.

Ad: CG7839‼

AM: CG11617, CG15710, CG33017, CG8216.

AF: CG14711, CG17802, CG2678, CG30403, CG8159, pad, phol.

Repeat: Abd-B(for sue), cic, insv, jing, nmo-small.

38: expanding
06: Chromatin extracted
08: IPed
21: Librared
- 12: In seq queue
- 01: Processing
- *08*: Awaiting reps or recollecting

*: Probably need to recollect at better time point
†: Collect at different time-point if fail
‼: failed once
^: tight expression

C. elegans ChIP-seq

- ~160 datasets

  AHR-1, ALY-1, B0035.1, B0261.1, B0310.2, BLMP-1, C04F5.9, C06A8.2, C08G9.2, CEH-14, CEH-18, CEH-2, CEH-24, CEH-31, CEH-32, CEH-34, CEH-36, CEH-48, CEH-9, CEH-90, CEY-2, CHD-7, CHE-1, COG-1, DAF-16*, DAO-5, DIE-1, DMD-4, DPL-1, DSC-1, DUXL-1, DVE-1!, EFL-1*, EGL-13, ELT-1, ELT-2!, ELT-4, ETS-4, ETS-7, F08F3.9, F10B5.3, F13H6.1!, F22D6.2, F23B12.7, F37D6.2, F49E8.2!, F52B5.7, F55B11.4, FAX-1, FKH-3, FKH-4, FKH-6, FKH-8, GMEB-2, HIF-1, HIM-1*, HLH-12, HLH-15, HLH-30!, HLH-4, HLH-6-R, HLH-8, HMBX-1, HMG-11, HND-1, IRX-1, K09A11.1, LET-607, LIM-6, LIN-40, LIR-3!, LSY-12, LSY-27, MADF-10, MEC-3, MED-1, MEL-28, MES-2, MES-4, MLS-2, MXL-1, NFYA-1!, NHR-102, NHR-179, NHR-20, NHR-232, NHR-25, NHR-43, NHR-47, NHR-48, NHR-71, NHR-80, NHR-85, NHR-90*!, NPAX-4, ODD-2, PAG-3!, POP-1, PQM-1, RBR-2, REC-8*, REF-2, RNT-1!, RPC-1!, SDC-2, SDZ-38, SMA-3, SMA-9, SNPC-4*, SNU-23, SOX-4, SPR-1, ~~SPR-4~~, SWSN-7, SYD-9, T02C12.2, T07F8.4, T26A5.8, TBX-2!, TBX-7, TBX-9, TTX-3, UNC-120, UNC-130!, UNC-3, UNC-42, UNC-86!, WAGO-9, XBP-1, XND-1, Y116A8C.19, Y22D7AL.16, Y53C12C.1, ZFP-2, ZIP-5, ZK185.1, ZTF-11, ~~ZTF-16~~, ZTF-18

- ~15 in UofC pipeline

* Multiple lines
! Multiple time-points collected
109: Released by DCC

19: Ready for release
03: incomplete registration
20: need Ab characterization

goatV IP with *wt* and nlsGFP

wt Exp. 1	
Rep2_pr	3823
Rep2_Rep1	5927
Rep2_Rep3	5517
Rep1_pr	3752
Rep1_Rep3	5832
Rep3_pr	3697
Rep0_pr	5819
optThresh	5927
conThresh	5927

wt Exp. 2	
Rep2_pr	4961
Rep2_Rep1	5728
Rep2_Rep3	6571
Rep1_pr	2784
Rep1_Rep3	5293
Rep3_pr	4570
Rep0_pr	7022
optThresh	7022
conThresh	6571

nlsGFP Exp. 1	
Rep2_pr	6222
Rep2_Rep3	8920
Rep2_Rep1	9000
Rep3_pr	6000
Rep3_Rep1	8732
Rep1_pr	6235
Rep0_pr	8013
optThresh	9000
conThresh	9000

nlsGFP Exp. 2	
Rep2_pr	5898
Rep2_Rep3	6707
Rep2_Rep1	9294
Rep3_pr	3986
Rep3_Rep1	6366
Rep1_pr	7169
Rep0_pr	9230
optThresh	9294
conThresh	9294

	Total Peaks, Exp1	Total Peaks, Exp2	Exp1 ∩ Exp2	Pooled Biorep. IDR Exp1 vs Exp2
Embryonic wt IP	5466!	6243!	4446	8450
Embryonic nls-GFP	8468!	8488!	6824	9878

! After merging overlapping peaks

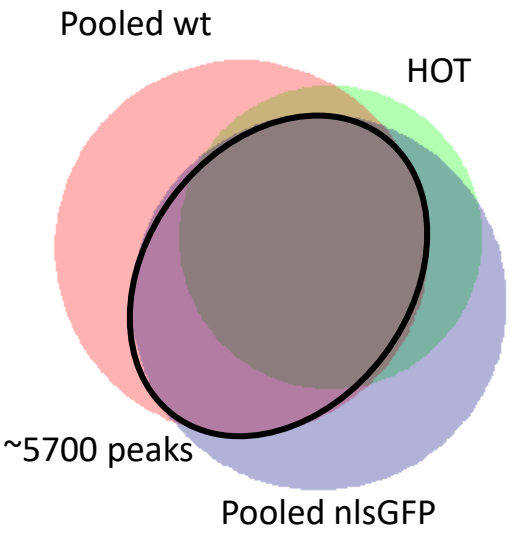
Better set of peaks?

goatV IP with *wt* and nlsGFP

How do both attempts to call harmonized peaks for both experiments compare?

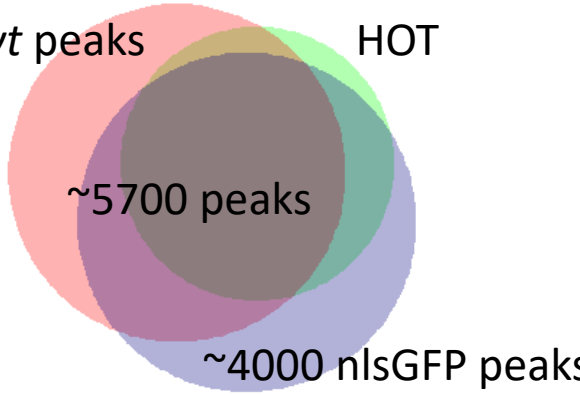
% of peaks in dataset (column) overlapping with peaks in second dataset (row)

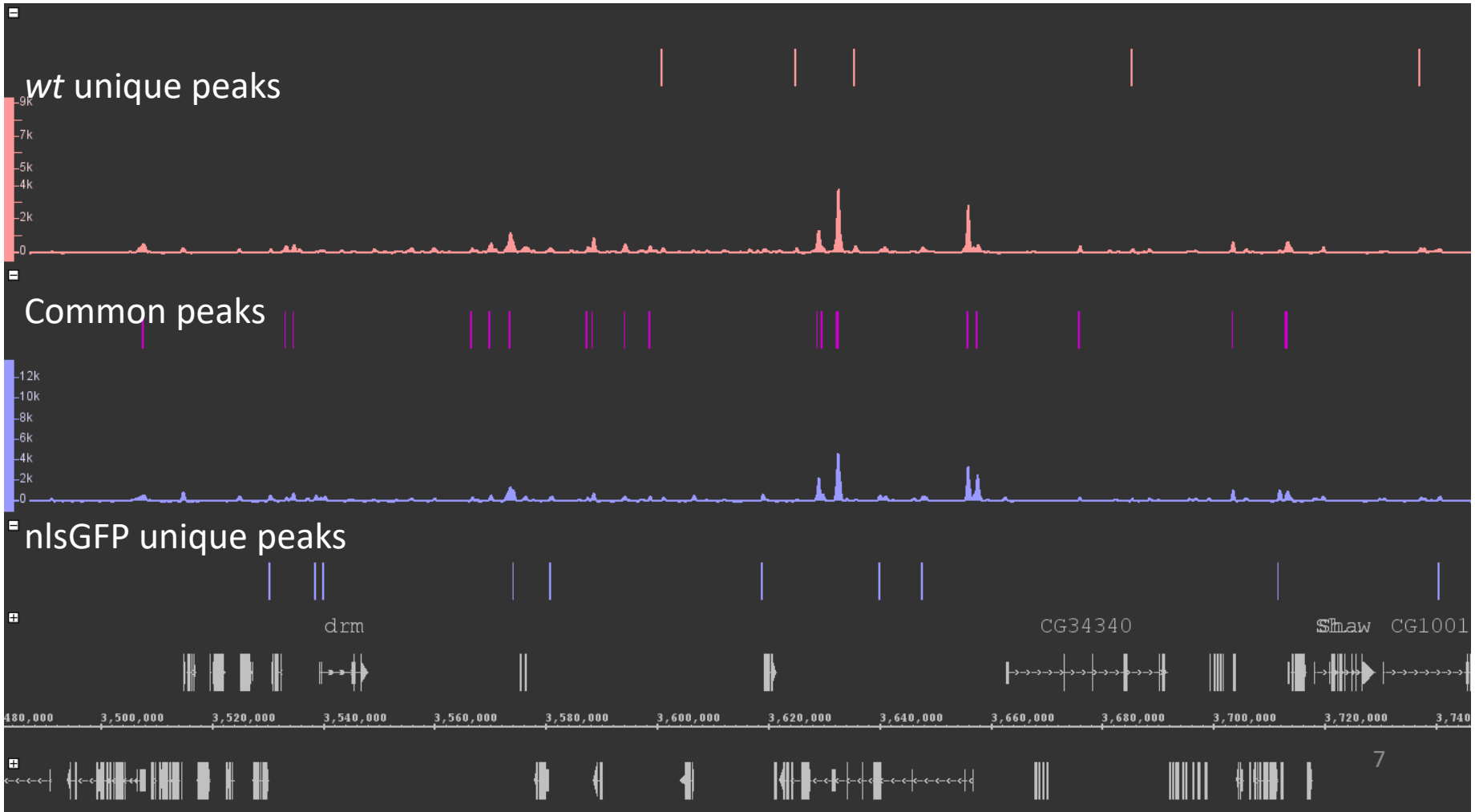
	HOT regions	wt ∩	*wt* Pooled BioRep. IDR	nlsGFP ∩	nlsGFP Pooled BioRep. IDR
Total Peaks	5564	4446	8450	6824	9878
HOT regions	-	82%	59%	82%	60%
wt ∩	61%	-	52%	52%	40%
wt pooled	81%	98%	-	70%	59%
nlsGFP ∩	88%	77%	55%	-	64%
nlsGFP pooled	91%	88%	68%	93%	-



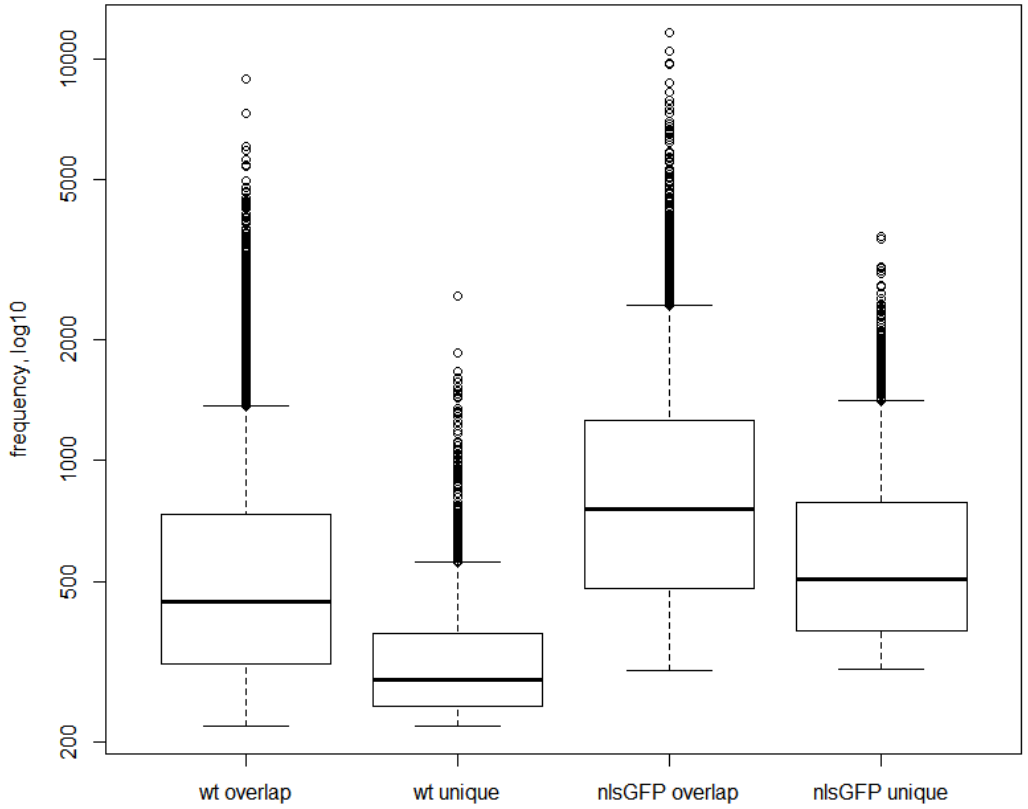
- With the more conservative method intersection method, still not 100% overlap with pooled peaks (yellow).
- Pooled peaks calls overlap better with HOT regions (green)
- Unfounded to expect wt to be a subset of nlsGFP? (purple)

Pooled nlsGFP and *wt* Wigs





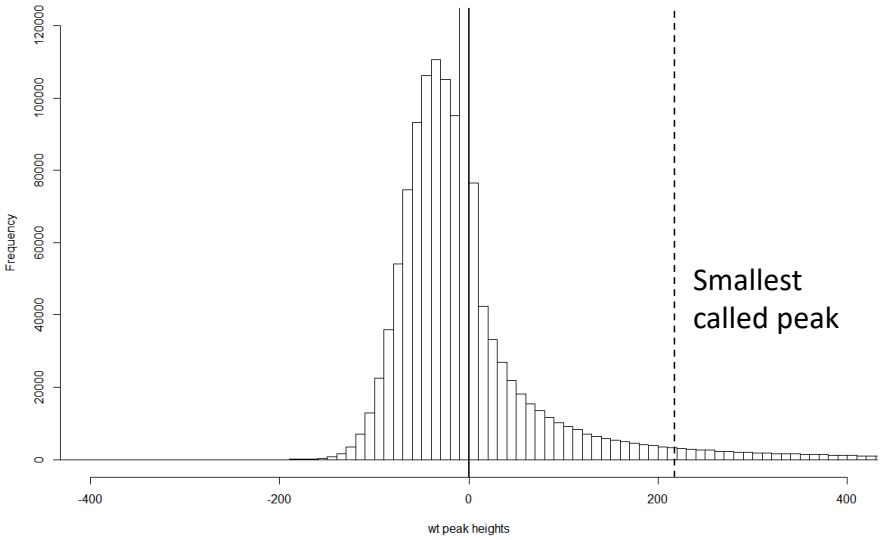
Peak intensities



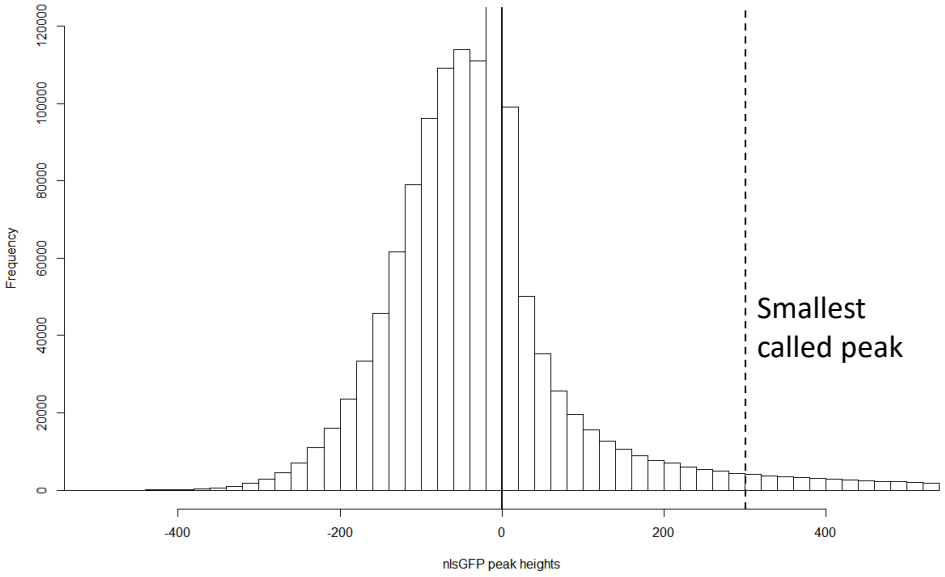
Intensities of peaks unique to *wt* or nlsGFP datasets are on average smaller than peaks seen in both datasets

Distributions of Peak Heights

Pooled *wt*



Pooled nlsGFP



- IDR threshold too loose at 0.01?
- Concede that small peaks in *wt* control wont always be observed in actual datasets?
  - Will removing these small peaks from actual data increase false negative calls?