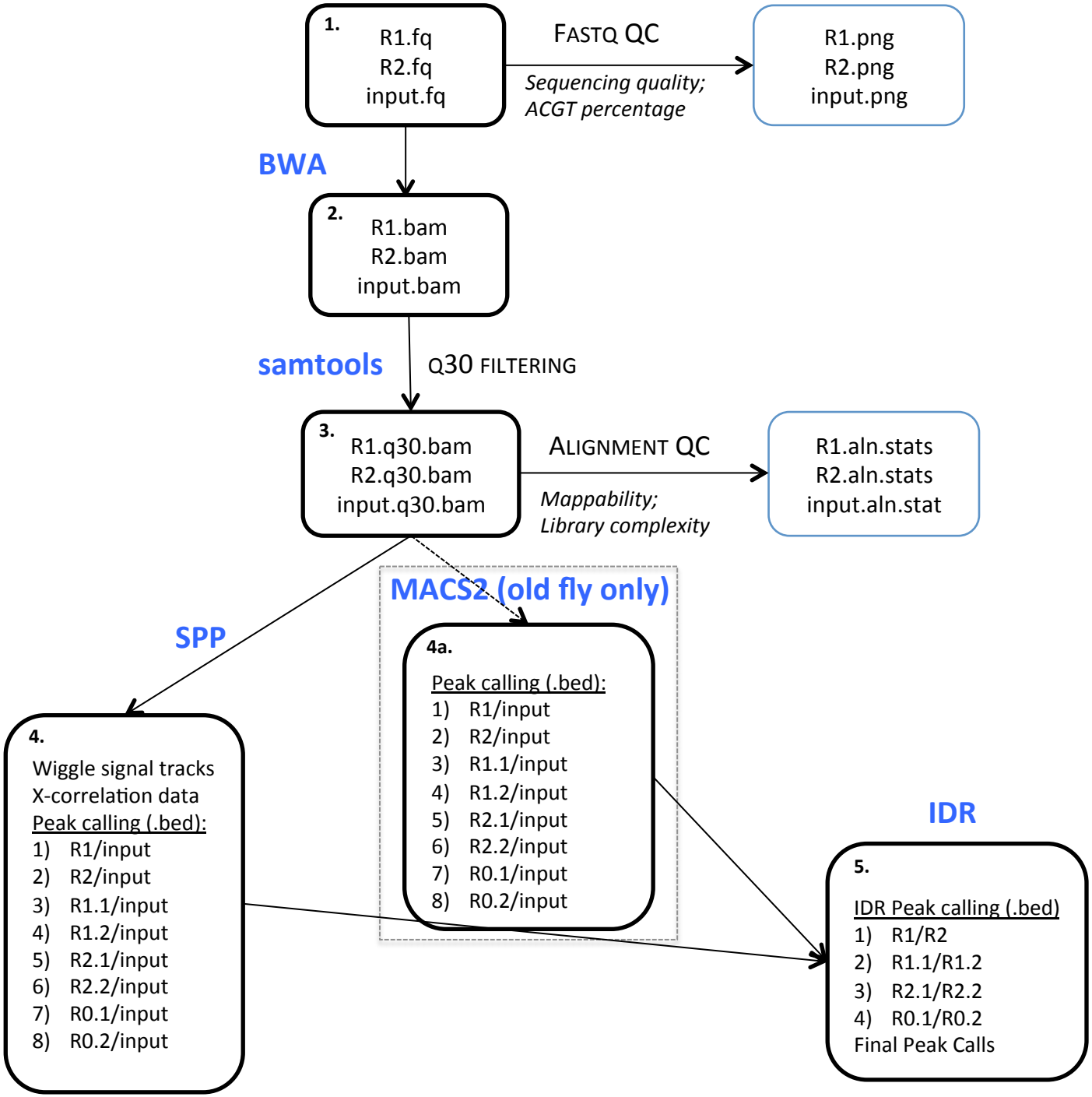
ChIP-seq Metrics – 2013 modENCODE (Bridge)

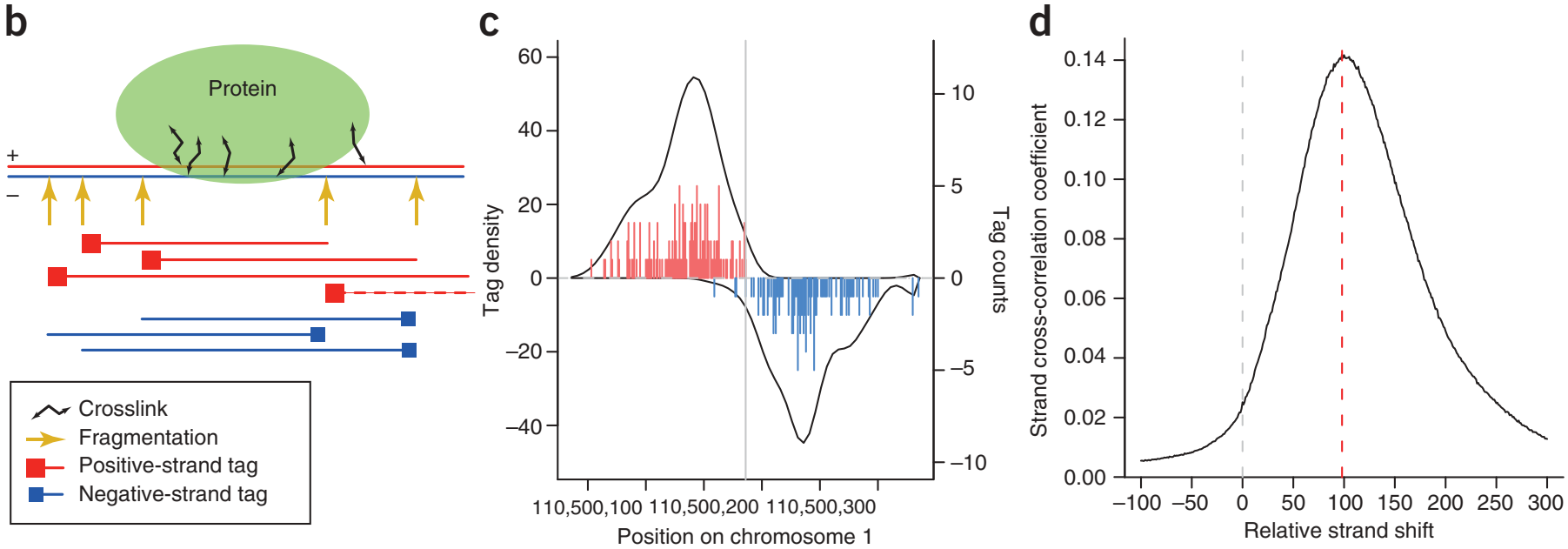
E. Jay Rehm
Lijia Ma
Kevin P. White
2.14.2013


1.
R1.fq
R2.fq
input.fq
FASTQ QC
Sequencing quality;
ACGT percentage
R1.png
R2.png
input.png
BWA
2.
R1.bam
R2.bam
input.bam
samtools
Q30 FILTERING
3.
R1.q30.bam
R2.q30.bam
input.q30.bam
ALIGNMENT QC
Mappability;
Library complexity
R1.aln.stats
R2.aln.stats
input.aln.stat
MACS2 (old fly only)
SPP
4a.
Peak calling (.bed):
1) R1/input
2) R2/input
3) R1.1/input
4) R1.2/input
5) R2.1/input
6) R2.2/input
7) R0.1/input
8) R0.2/input
4.
Wiggle signal tracks
X-correlation data
Peak calling (.bed):
1) R1/input
2) R2/input
3) R1.1/input
4) R1.2/input
5) R2.1/input
6) R2.2/input
7) R0.1/input
8) R0.2/input
IDR
5.
IDR Peak calling (.bed)
1) R1/R2
2) R1.1/R1.2
3) R2.1/R2.2
4) R0.1/R0.2
Final Peak Calls

Dataset Quality Metrics

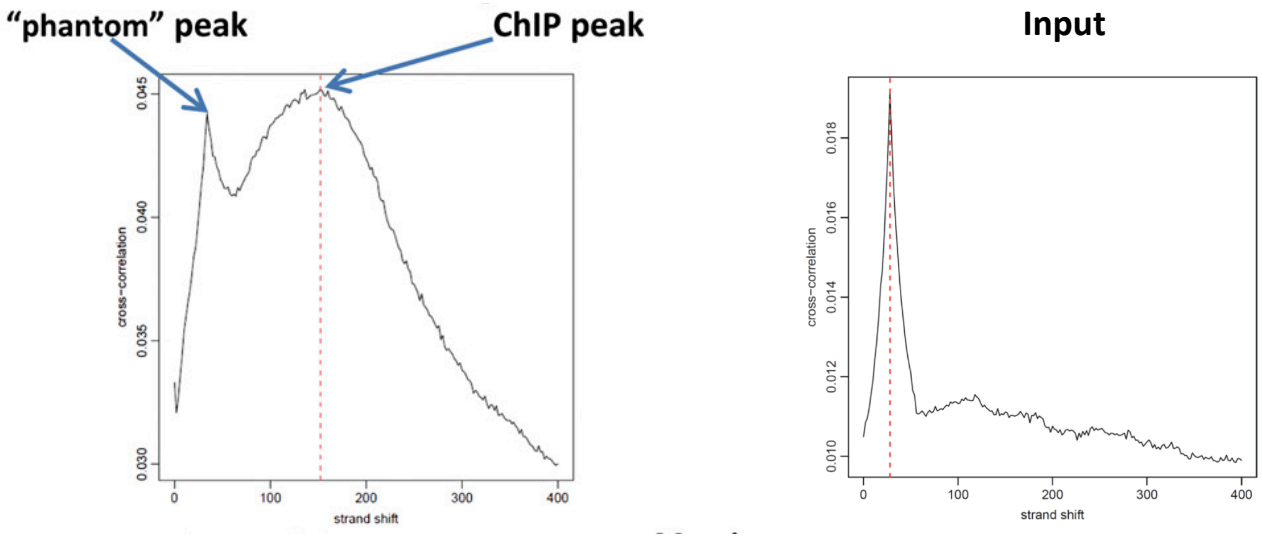
1. Consistency among/within replicates of a dataset
   - IDR Consistency
   - IDR Self-consistency
2. Individual replicate quality
   - Cross correlation analysis
3. Alignment stats
   - samtools

Cross Correlation

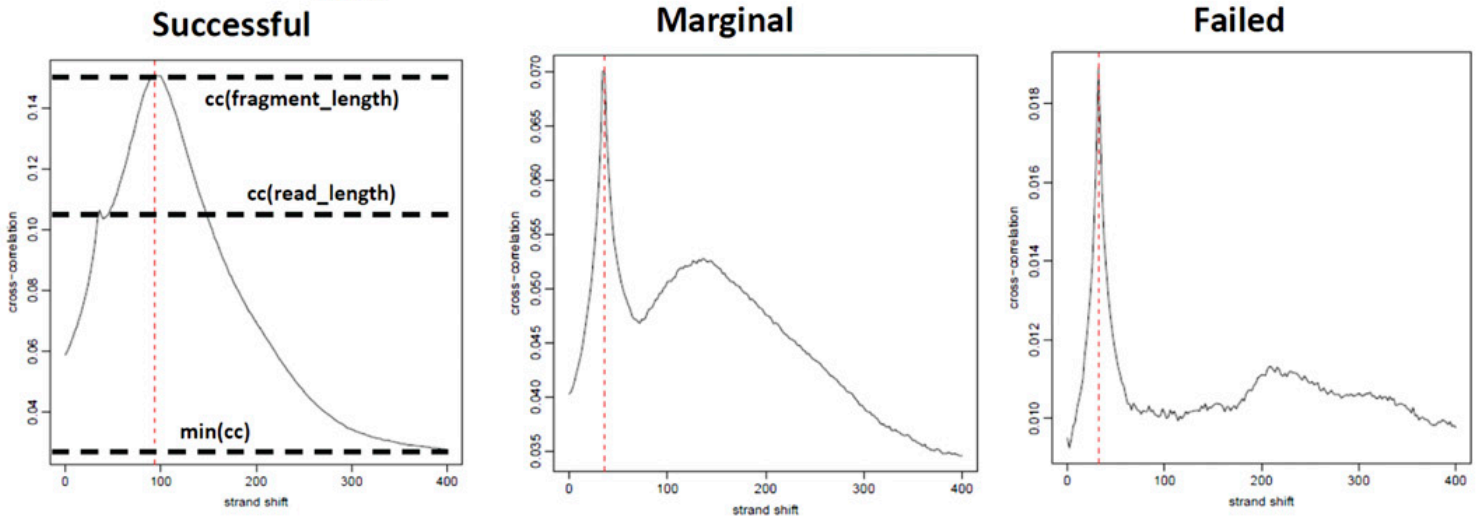


Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Nat Biotechnol.

E



G

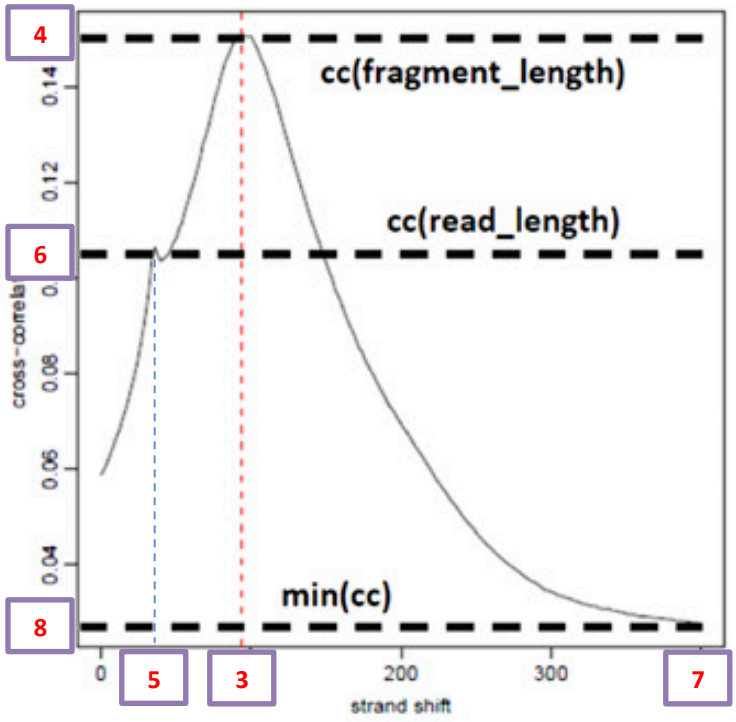


$$NSC = \frac{cc(fragment\ length)}{min(cc)} > 1.05 \quad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)} > 1$$

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. 2012. Genome Res.

SPP CC Outputs: 11 tab delimited columns

COL1: Filename
COL2: numReads: effective sequencing depth
COL3: estFragLen: fragment length (ChIP peak) cross-correlation peak(s)
COL4: corr_estFragLen: strand cross-correlation value(s) in decreasing order
COL5: phantomPeak: Read length/phantom peak strand shift
COL6: corr_phantomPeak: Correlation value at phantom peak
COL7: argmin_corr: strand shift at which cross-correlation is lowest
COL8: min_corr: minimum value of cross-correlation
COL9: Normalized strand cross-correlation coefficient (NSC) = COL4 / COL8
COL10: Relative strand cross-correlation coefficient (RSC) = (COL4 - COL8) / (COL6 - COL8)
COL11: QualityTag: Quality tag based on thresholded RSC (codes: -2:veryLow, -1:Low, 0:Medium, 1:High, 2:veryHigh)



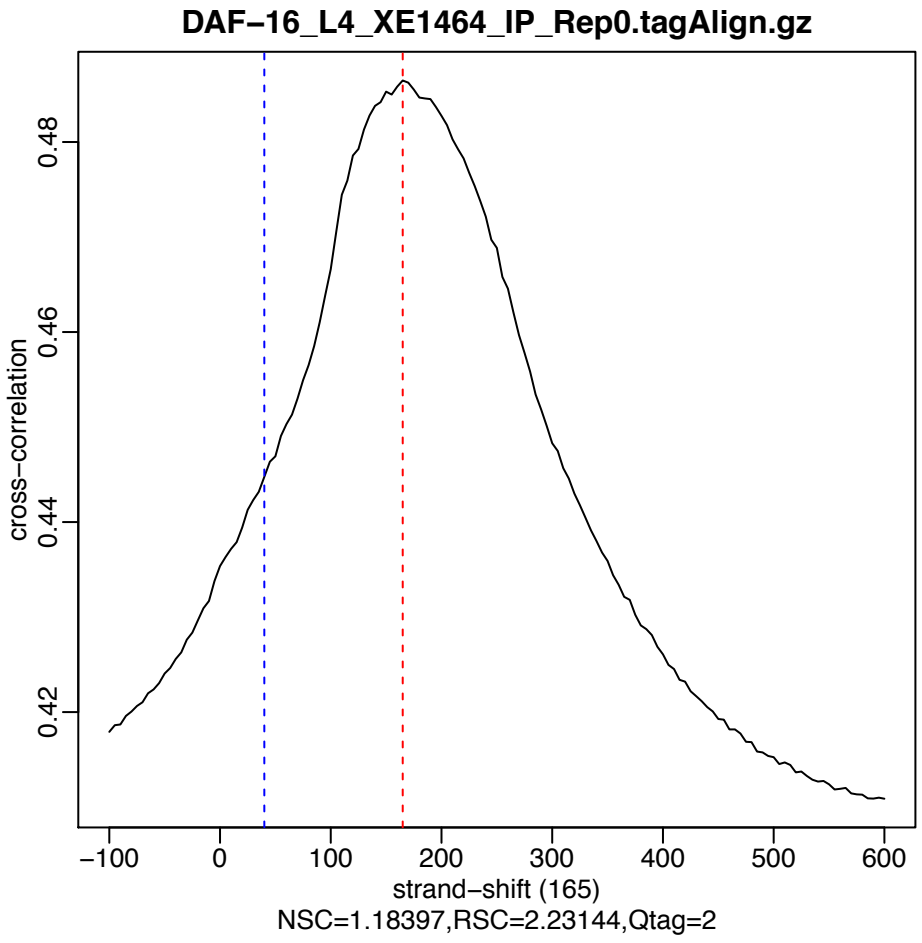
Note: RSC not helpful with worm datasets.

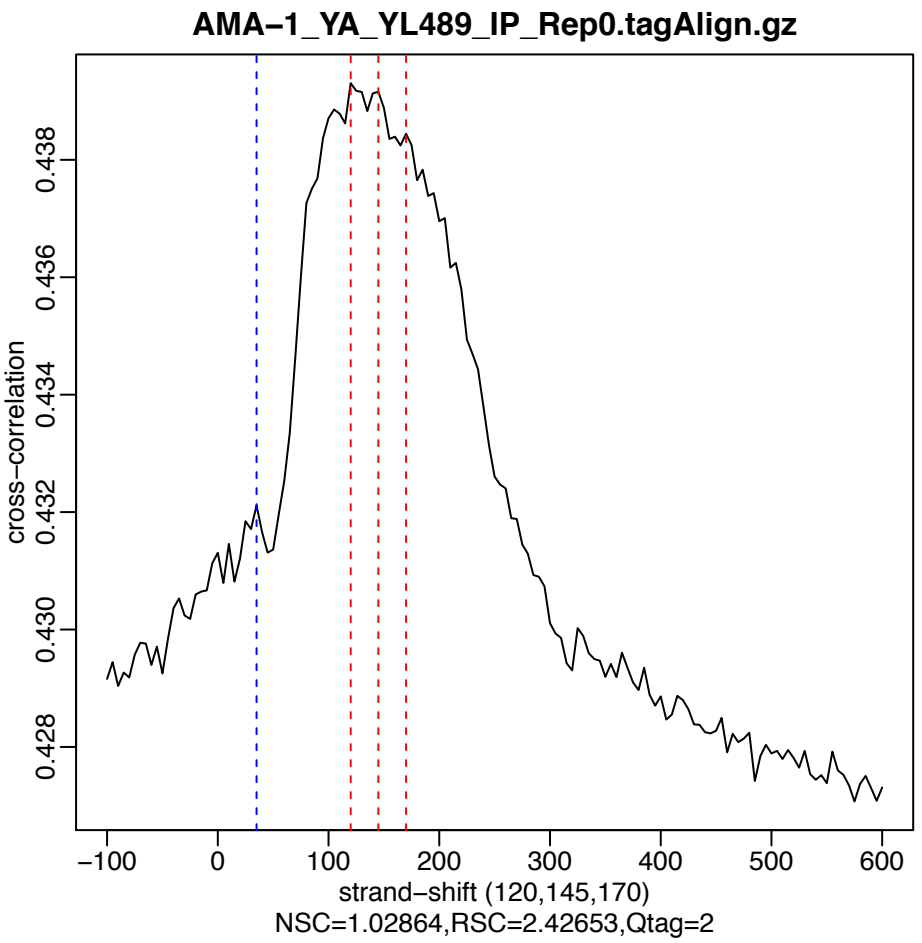
SPP CC Outputs: 11 tab delimited columns

COL1: Filename:	DAF-16_L4_XE1464_a-GFP_Rep0.tagAlign.gz
COL2: numReads:	8,721,920
COL3: estFragLen:	165
COL4: corr_estFragLen:	0.486472876
COL5: phantomPeak:	40
COL6: corr_phantomPeak:	0.4447579
COL7: argmin_corr:	600
COL8: min_corr:	0.4108828
COL9: (NSC) = COL4 / COL8:	1.18397
COL10: (RSC) = (COL4 - COL8) / (COL6 - COL8):	2.231438
COL11: QualityTag:	2

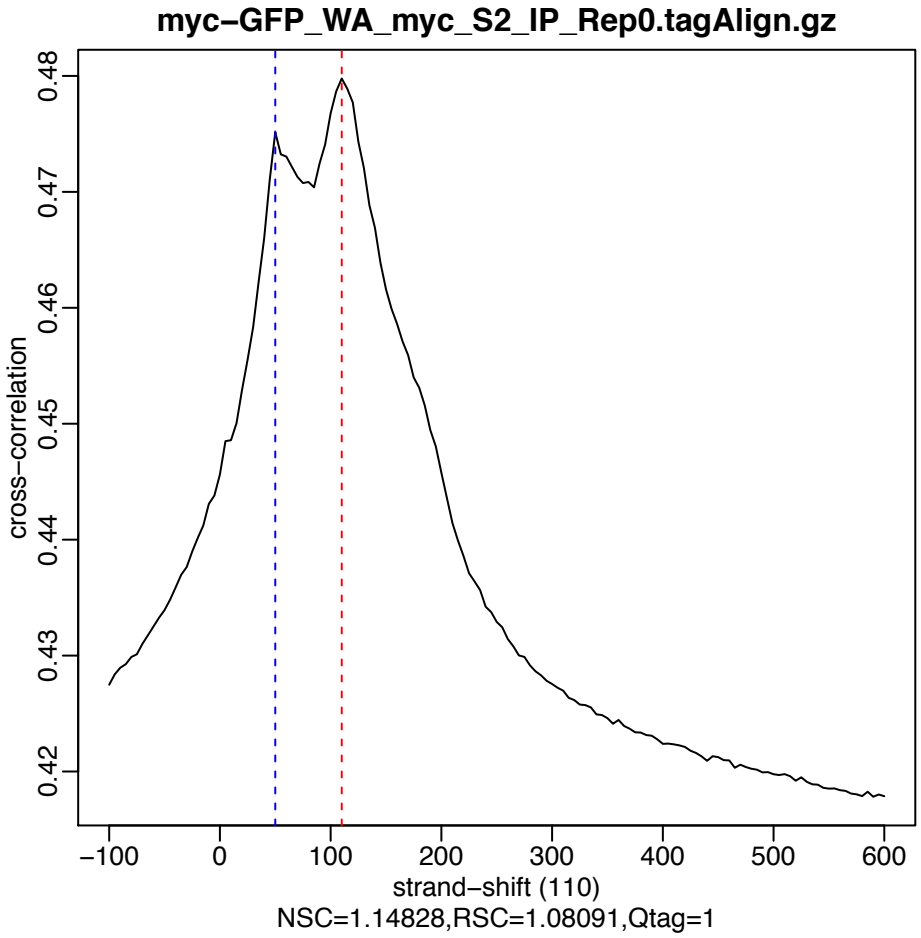
COL1: Filename:	AMA-1_YA_YL489_a-GFP_Rep0.tagAlign.gz
COL2: numReads:	4,919,928
COL3: estFragLen:	120, 145, 170
COL4: corr_estFragLen:	0.439305514381849, 0.439158556907702, 0.438443808017683
COL5: phantomPeak:	35
COL6: corr_phantomPeak:	0.432114
COL7: argmin_corr:	575
COL8: min_corr:	0.4270727
COL9: (NSC) = COL4 / COL8:	1.028643
COL10: (RSC) = (COL4 - COL8) / (COL6 - COL8):	2.426529
COL11: QualityTag:	2

Example CC Plots -- worm

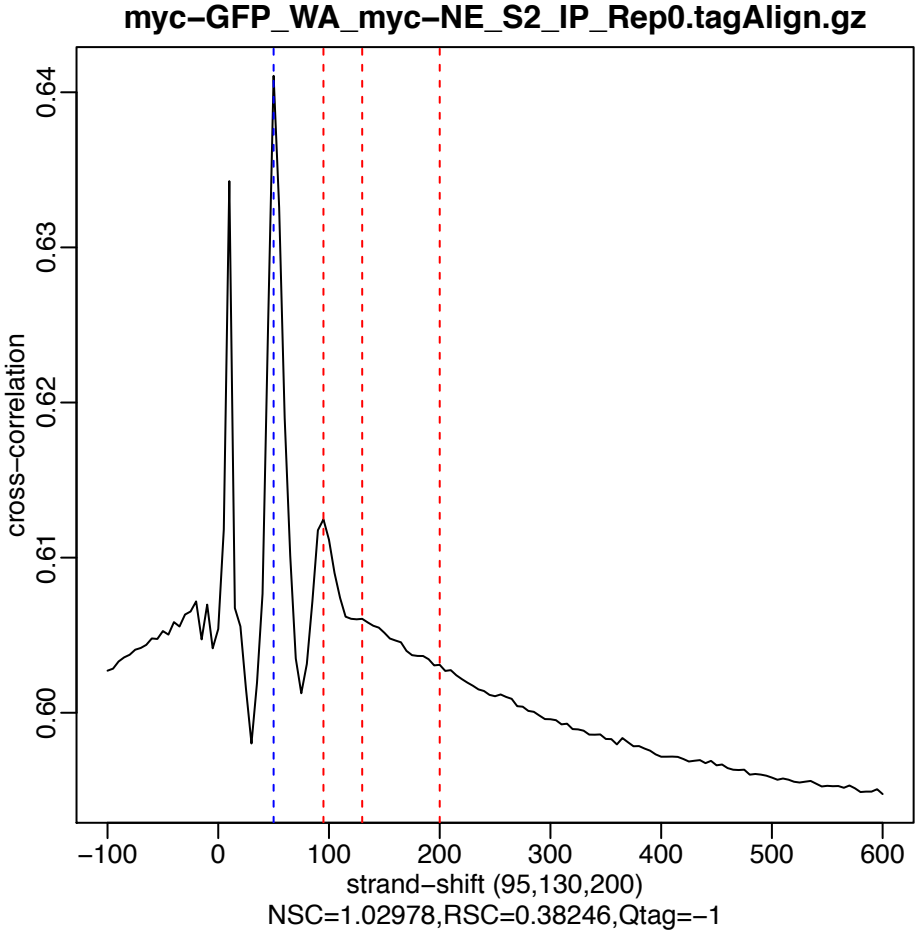




Example CC Plots -- fly

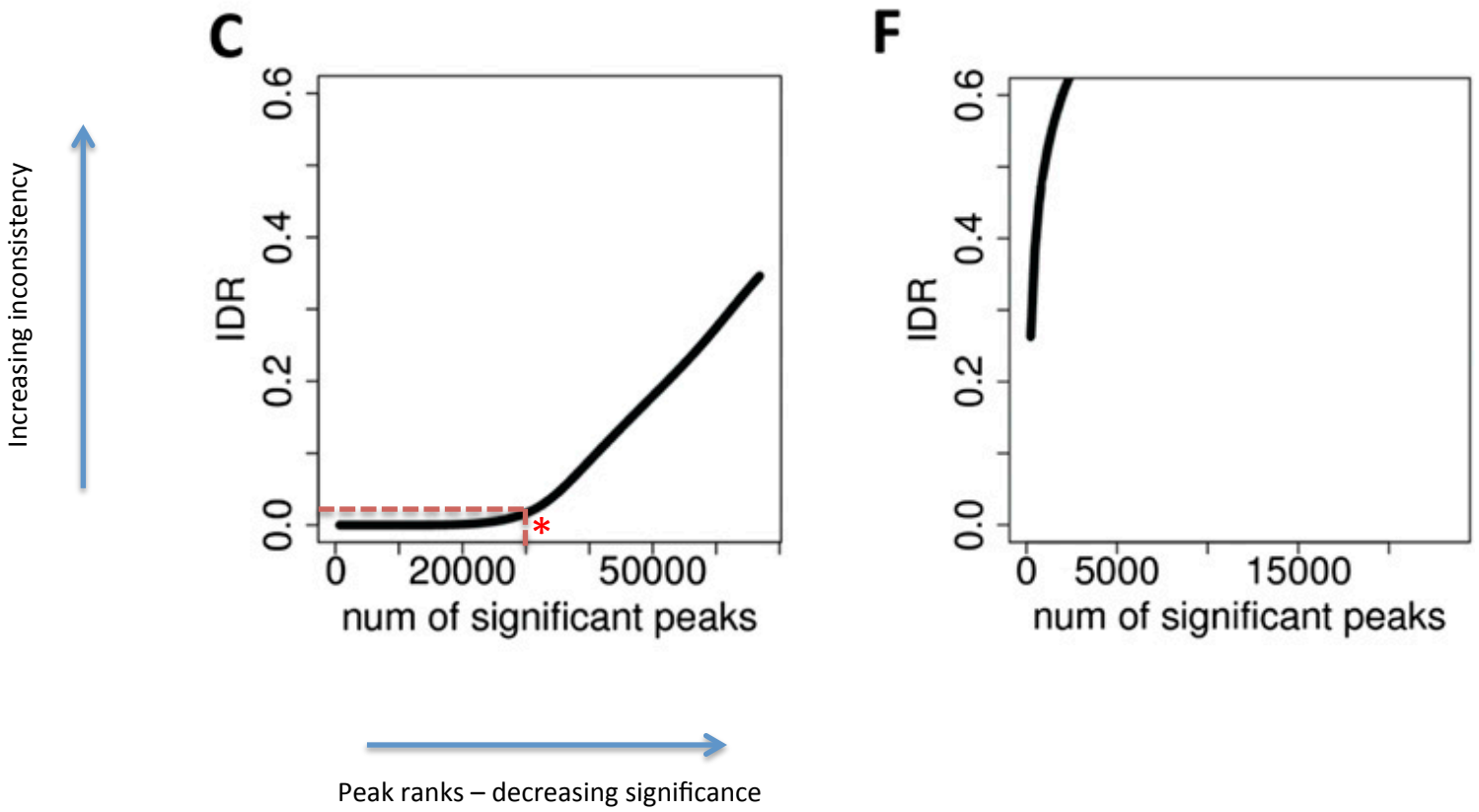


Nugen library prep



Nextera library prep

IDR Plots



Landt SG et al. 2012. Genome Res.

IDR

3 classes of consistency measured:

			IDR outputs	IDR thresholds
1.	Replicate consistency:	Rep1/Rep2	Nt	(.02)
2.	Replicate self-consistency:	Rep1_pr1/Rep1_pr2 & Rep2_pr1/Rep2_pr2	N1, N2	(.01)
3.	Pooled self-consistency:	Rep0_pr1/Rep0_pr2 (Rep0 = Rep1 U Rep2)	Np	(.01)

Low-stringency Peak calls (SPP set to 30,000):

(8 peak calls required for 2 replicates -- 11 for 3 replicates)

1) Rep1/Input
2) Rep2/Input
3) Rep1_pr1/Input
4) Rep1_pr2/Input
5) Rep2_pr1/Input
6) Rep2_pr2/Input
7) Rep0_pr1/Input
8) Rep0_pr2/Input

Flagging datasets for low consistency:

If **N2/N1 > 2** AND **Np/Nt > 2** → FLAGGED

or

If **N2/N1 >> 2** OR **Np/Nt >> 2** → FLAGGED

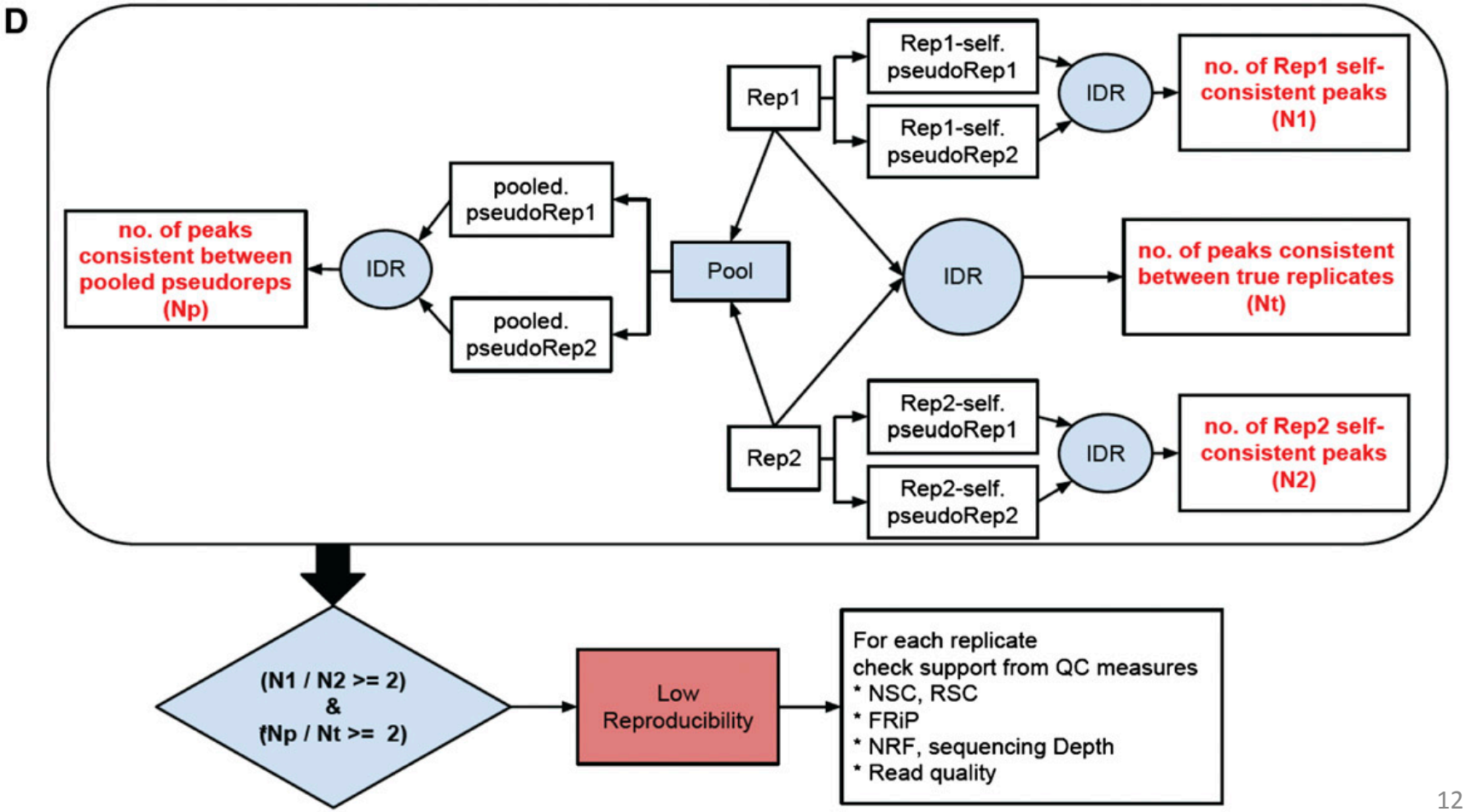
IDR

3 classes of consistency measured:

1. Replicate consistency: REP1/REP2
2. Replicate self-consistency: REP1_PR1/REP1_PR2 AND REP2_PR1/REP2_PR2
3. Pooled self-consistency: REP0_PR1/REP0_PR2 (REP0 = REP1 U REP2)

4 IDR outputs:

Nt
N1, N2
Np



Landt SG et al. 2012. Genome Res.

IDR outputs:

1. Replicate consistency: Nt "numPeaks_Rep1_Rep2"
2. Replicate self-consistency: N1, N2 "numPeaks_Rep1_pr" & "numPeaks_Rep2_pr"
3. Pooled self-consistency: Np "numPeaks_Rep0_pr"

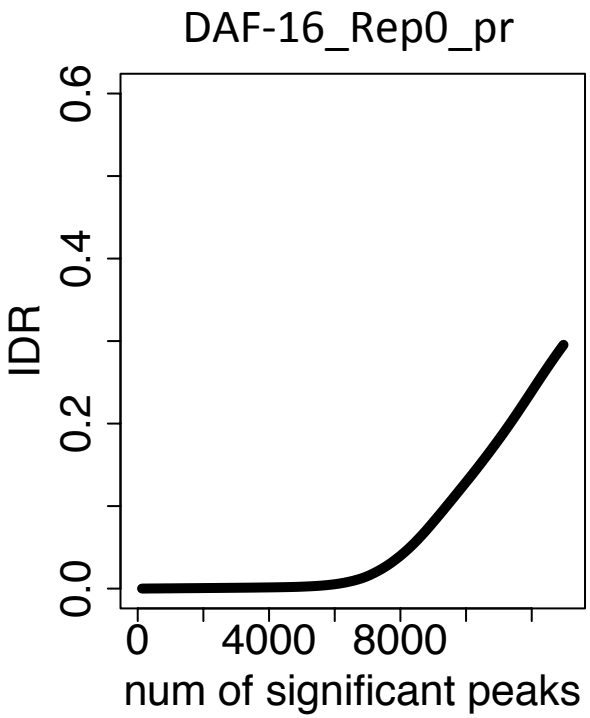
				Self-consistency flags	Np/Nt	Batch	Comments
IPs:	DAF-16_L4_XE1464_a-GFP_Rep1	DAF-16_L4_XE1464_a-GFP_Rep2				20130129	
Inputs:	DAF-16_L4_XE1464_Input_Rep0.tagAlign.gz					20130129	
numPeaks_Rep1_pr	5493					20130129	
numPeaks_Rep1_Rep2	**4092**					20130129	
numPeaks_Rep2_pr	5144					20130129	
numPeaks_Rep0_pr	6611				1.62	20130129	
optThresh	6611					20130129	
conThresh	4092					20130129	
IPs:	EFL-1_YA_YL479_a-GFP_Rep1	EFL-1_YA_YL479_a-GFP_Rep2				20130129	
Inputs:	EFL-1_YA_YL479_Input_Rep0.tagAlign.gz					20130129	
numPeaks_Rep1_pr	1471					20130129	
numPeaks_Rep1_Rep2	**1600**					20130129	
numPeaks_Rep2_pr	1064					20130129	
numPeaks_Rep0_pr	1886				1.18	20130129	
optThresh	1886					20130129	
conThresh	1600					20130129	
IPs:	AMA-1_YA_YL489_a-GFP_Rep1	AMA-1_YA_YL489_a-GFP_Rep2				20130129	
Inputs:	AMA-1_YA_YL489_Input_Rep0.tagAlign.gz					20130129	
numPeaks_Rep1_pr	183			Rep1_pr/Rep2_pr		20130129	double-flag
numPeaks_Rep1_Rep2	**169**					20130129	
numPeaks_Rep2_pr	407					20130129	
numPeaks_Rep0_pr	524				3.10	20130129	double-flag
optThresh	524					20130129	
conThresh	169					20130129	

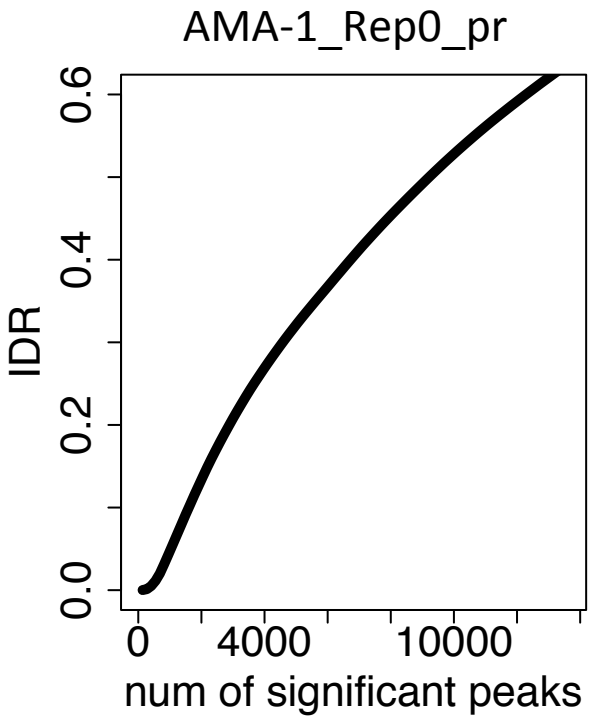
Conservative and Optimal final peak calls:

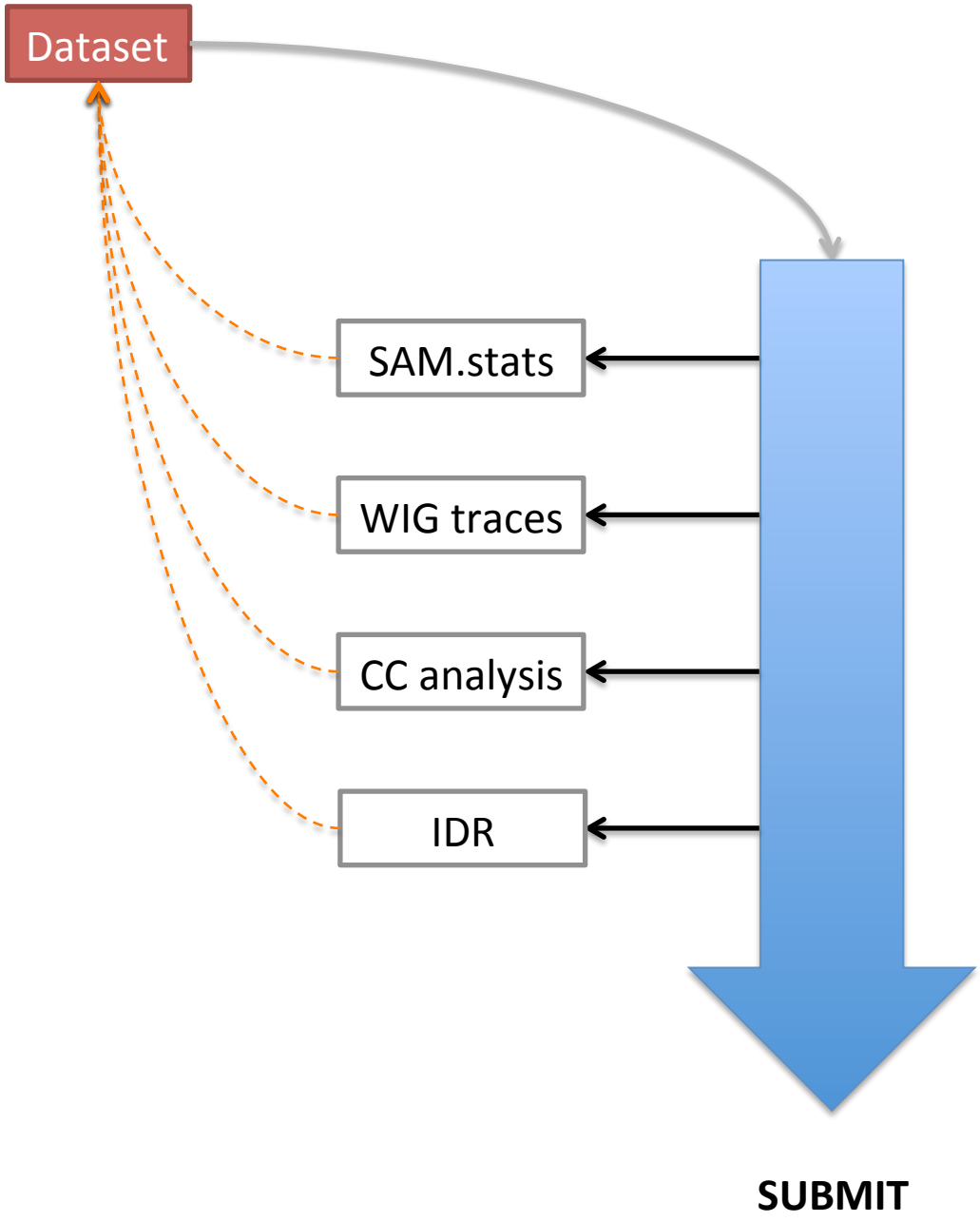
Optimal threshold = MAX[numPeaks_Rep1_Rep2, numPeaks_Rep0_pr]

Conservative threshold = numPeaks_Rep1_Rep2

				Self-consistency flags	Np/Nt	Batch	Comments
IPs:	DAF-16_L4_XE1464_a-GFP_Rep1	DAF-16_L4_XE1464_a-GFP_Rep2				20130129	
Inputs:	DAF-16_L4_XE1464_Input_Rep0.tagAlign.gz					20130129	
numPeaks_Rep1_pr	5493					20130129	
numPeaks_Rep1_Rep2	**4092**					20130129	
numPeaks_Rep2_pr	5144					20130129	
numPeaks_Rep0_pr	6611				1.62	20130129	
optThresh	6611					20130129	
conThresh	4092					20130129	
IPs:	AMA-1_YA_YL489_a-GFP_Rep1	AMA-1_YA_YL489_a-GFP_Rep2				20130129	
Inputs:	AMA-1_YA_YL489_Input_Rep0.tagAlign.gz					20130129	
numPeaks_Rep1_pr	183			Rep1_pr/Rep2_pr		20130129	double-flag
numPeaks_Rep1_Rep2	**169**					20130129	
numPeaks_Rep2_pr	407					20130129	
numPeaks_Rep0_pr	524				3.10	20130129	double-flag
optThresh	524					20130129	
conThresh	169					20130129	

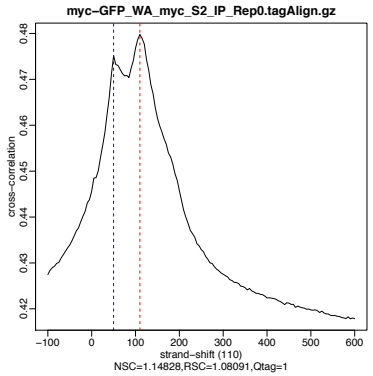




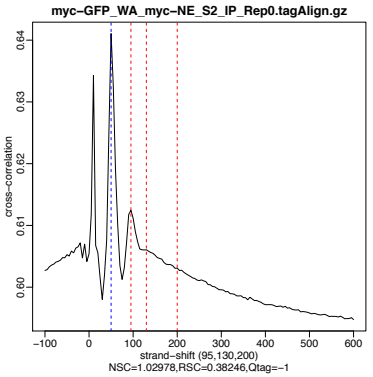

Dataset
SAM.stats
WIG traces
CC analysis
IDR
SUBMIT

A final word: Nextera and Nugen…

Nugen library prep



Nextera library prep



Both datasets pass IDR.

However, Nugen results in **2.5x** more peaks than Nextera

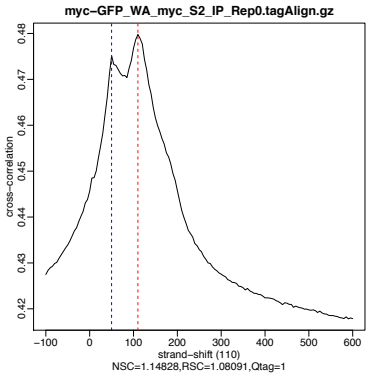
	SPP peak calls	MACS2 peak calls		Np/Nt (SPP)	Np/Nt (MACS2)
IDR-filtered SPP/MACS2 myc-GFP_WA_myc_S2 (nugen libraries:pooled ChIPs)					
IPs:	2012-1512_121109_SN1070_0094_BD1F1FACXX_6_sequence	2012-1513_121109_SN1070_0094_BD1F1FACXX_1_sequence			
Inputs:	myc-GFP_WA_myc_S2_Input_Rep0.tagAlign.gz				
numPeaks_Rep1_pr	3902	5781			
numPeaks_Rep1_Rep2	**3698**	**4766**			
numPeaks_Rep2_pr	4172	6582			
numPeaks_Rep0_pr	5364	6104		1.45	1.28
optThresh	5364	6104			
conThresh	3698	4766			
IDR-filtered SPP/MACS2 myc-GFP_WA_myc-NE_S2 (nextera libraries:pooled ChIPs)					
IPs:	2012-1451_121031_SN1070_0093_BD1F3RACXX_1_sequence	2012-1452_121031_SN1070_0093_BD1F3RACXX_1_sequence			
Inputs:	myc-GFP_WA_myc-NE_S2_Input_Rep0.tagAlign.gz				
numPeaks_Rep1_pr	391	2015			
numPeaks_Rep1_Rep2	815	**2468**			
numPeaks_Rep2_pr	402	2209			
numPeaks_Rep0_pr	705	2193			0.89
optThresh	815	2468			
conThresh	815	2468			

Nugen library prep

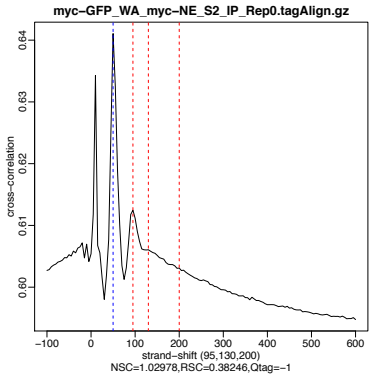
Nextera library prep

A final word: Nextera and Nugen…

Nugen library prep



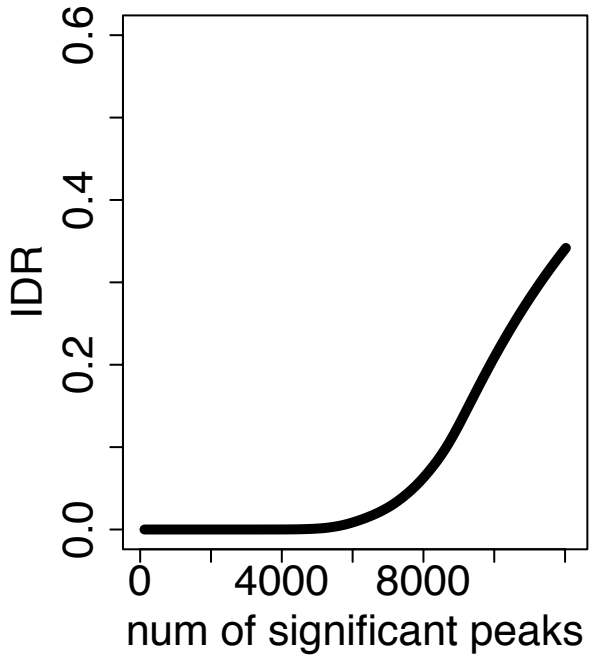
Nextera library prep



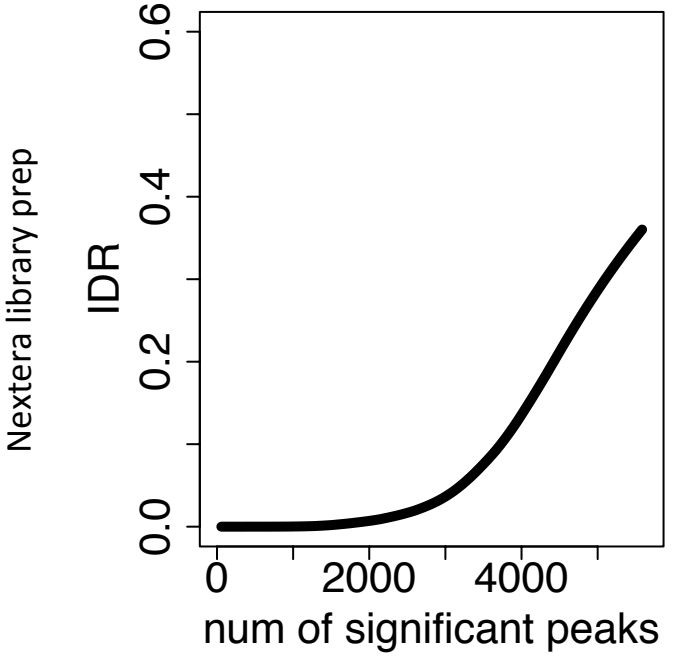
Both datasets pass IDR.

However, Nugen results in **2.5x** more peaks than Nextera

Nugen library prep



Nextera library prep



References

1. Anshul's document on Thresholding ChIPseq datasets using SPP and IDR: https://sites.google.com/site/anshulkundaje/projects/idr

2. Anshul's "phantompeakqualtools" documentation for using cross-correlation analysis to assess the quality of ChIP-seq datasets: https://code.google.com/p/phantompeakqualtools/

3. Official ENCODE quality metrics (not always appropriate for fly/worm): http://genome.ucsc.edu/ENCODE/qualityMetrics.html#chipSeq

1. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012 Sep;22(9):1813–31.

2. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics. Institute of Mathematical Statistics; 2011 Sep 1;5(3):1752–79.

3. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008 Dec;26(12):1351–9.

4. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet. 2012 Oct 23.