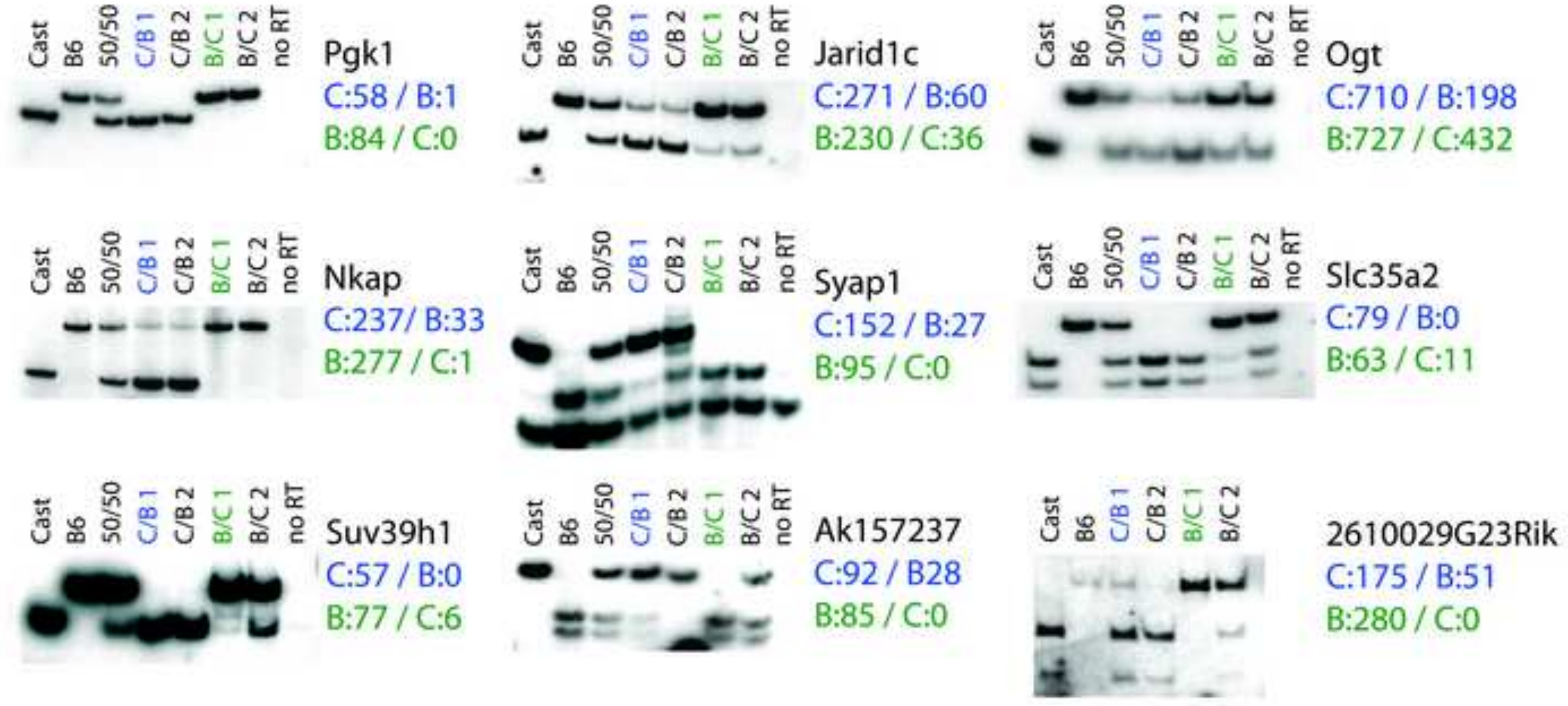
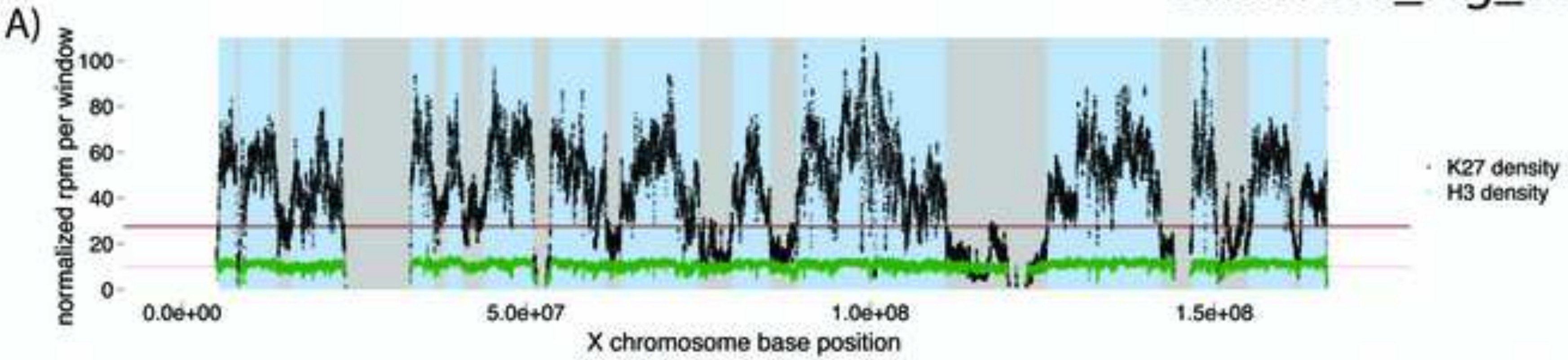
Supplemental Figure_S1

Calabrese_Fig_S1



Supplemental Figure_S2

Calabrese_Fig_S2



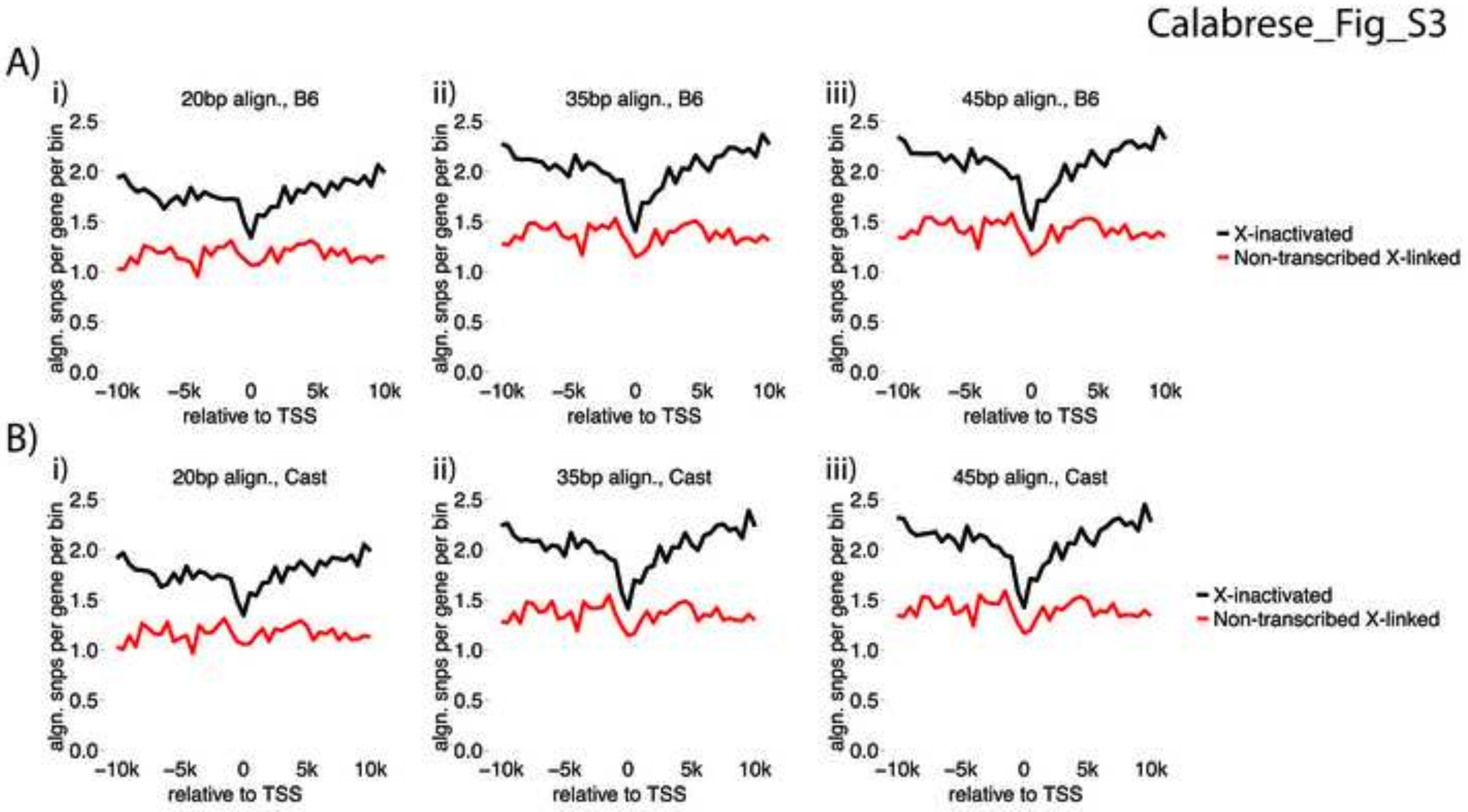
B)

P/V #	Start	End	Length	Alignability	Prop. LINE
peak1	5300000	7900000	2600000	0.8491	0.2391
valley1	7900000	8450000	550000	0.4197	0.4959
peak2	8450000	14000000	5550000	0.8389	0.2051
valley2	14000000	15750000	1750000	0.7439	0.4986
peak3	15750000	23250000	7500000	0.7706	0.3712
valley3	23250000	33250000	10000000	0.0996	0.516
peak4	33250000	36750000	3500000	0.7468	0.2521
valley4	36750000	38250000	1500000	0.774	0.49
peak5	38250000	40750000	2500000	0.8476	0.2747
valley5	40750000	43750000	3000000	0.7601	0.5153
peak6	43750000	51000000	7250000	0.8333	0.2871
valley6	51000000	53500000	2500000	0.2989	0.3444
peak7	53500000	61600000	8100000	0.8128	0.3363
valley7	61600000	63750000	2150000	0.7431	0.5095
peak8	63750000	75100000	11350000	0.7509	0.3333
valley8	75100000	80000000	4900000	0.716	0.4883
peak9	80000000	85500000	5500000	0.8472	0.3033
valley9	85500000	89350000	3850000	0.6877	0.4864
peak10	89350000	111000000	21650000	0.7952	0.363
valley10	111000000	125750000	14750000	0.6394	0.4429
peak11	125750000	142250000	16500000	0.8274	0.2912
valley11	142250000	146750000	4500000	0.4037	0.3341
peak12	146750000	150250000	3500000	0.8092	0.3051
valley12	150250000	155000000	4750000	0.758	0.3905
peak13	155000000	161450000	6450000	0.8595	0.2256
valley13	161450000	162500000	1050000	0.7712	0.4185
peak14	162500000	166450000	3950000	0.8792	0.2065

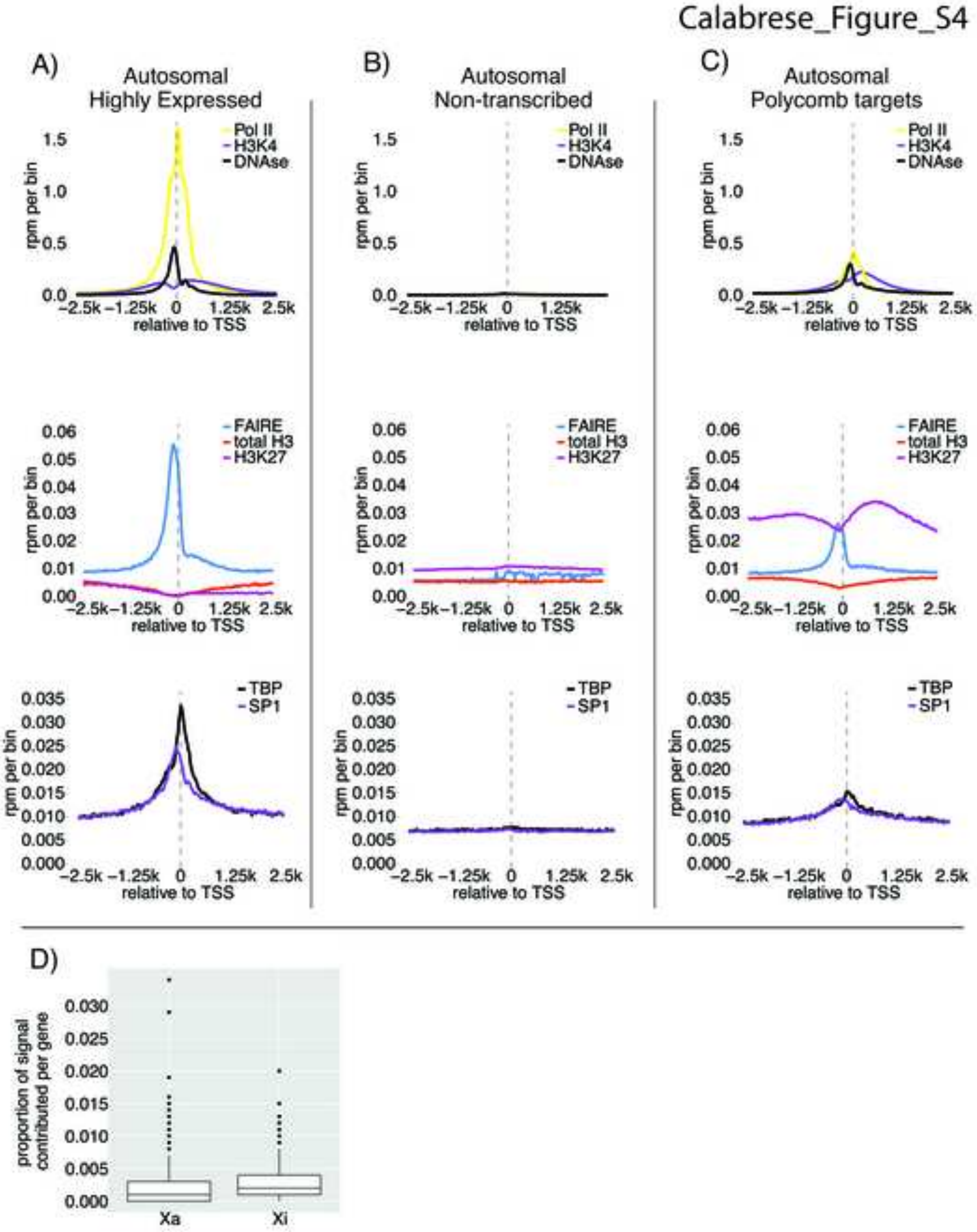
C)

Figure 3 probe ID:	Ube1x	15L	83H	78L	Chic1	87L	Rnf12	116L	160H	152L
start	20224883	14720724	83580255	78014372	100537038	86998950	101145921	115875046	159658600	152749993
stop	20264898	14919689	83773430	78205725	100577585	87167836	101187238	116079262	159867586	152931686
length	40015	198965	193175	191353	40547	168886	41317	204216	208986	181693
% align	92	80	85	74	92	78	94	74	82	75
% LINE	5	55	30	54	15	51	7	56	27	56
K27me3 (rpk)	513	185	410	95	638	101	574	45	368	113
% cells w/2 dots	91	94	96	95	90	94	98	100	100	100
BACPAC ID	G135-P65743A11	RP23-204P18	RP23-351J22	RP23-267N5	G135-P66518D5	RP23-368B24	G135-P605237C7	RP23-468D22	RP23-133E13	RP23-259M13

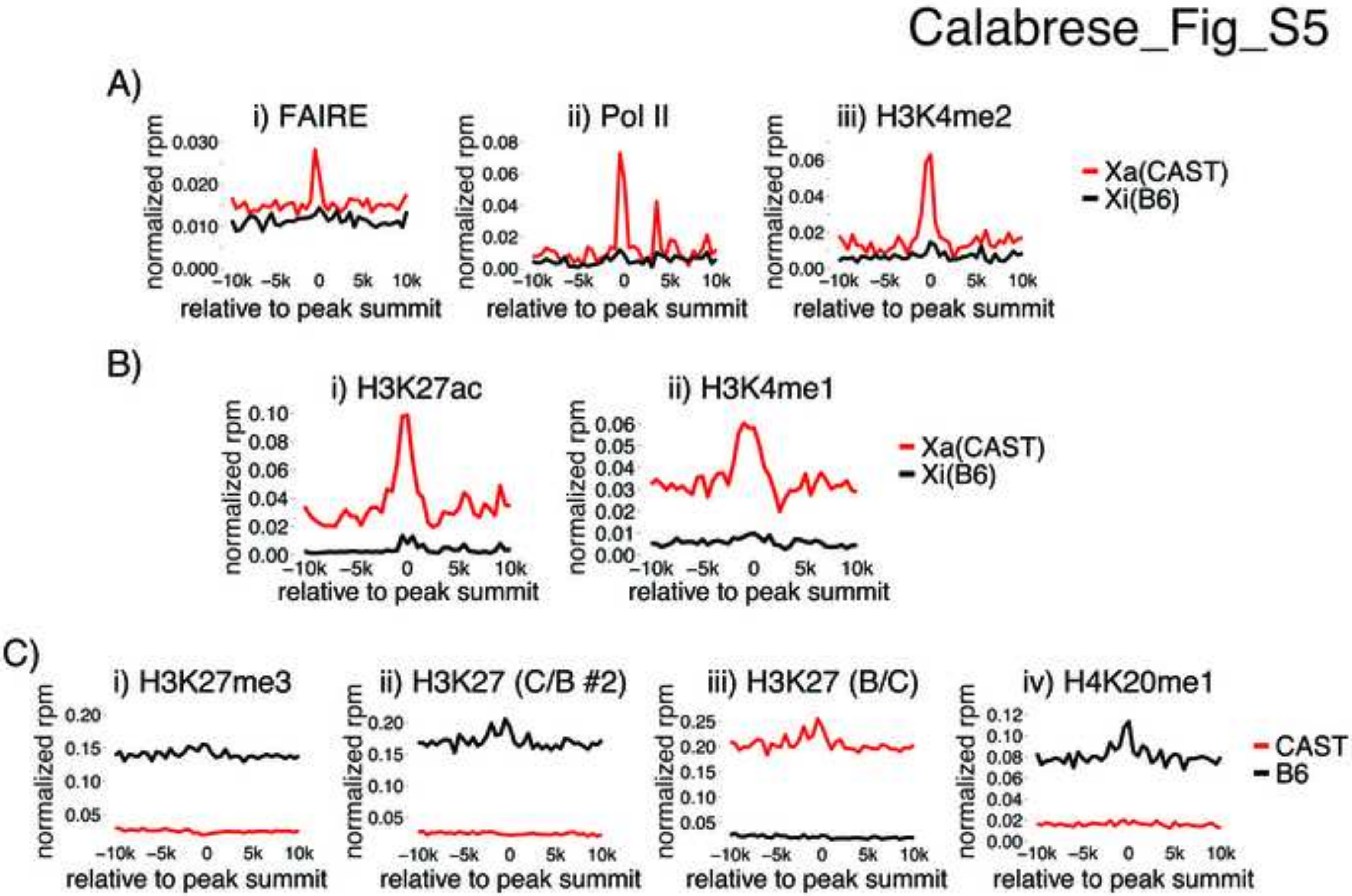
Supplemental Figure_S3



Supplemental Figure_S4

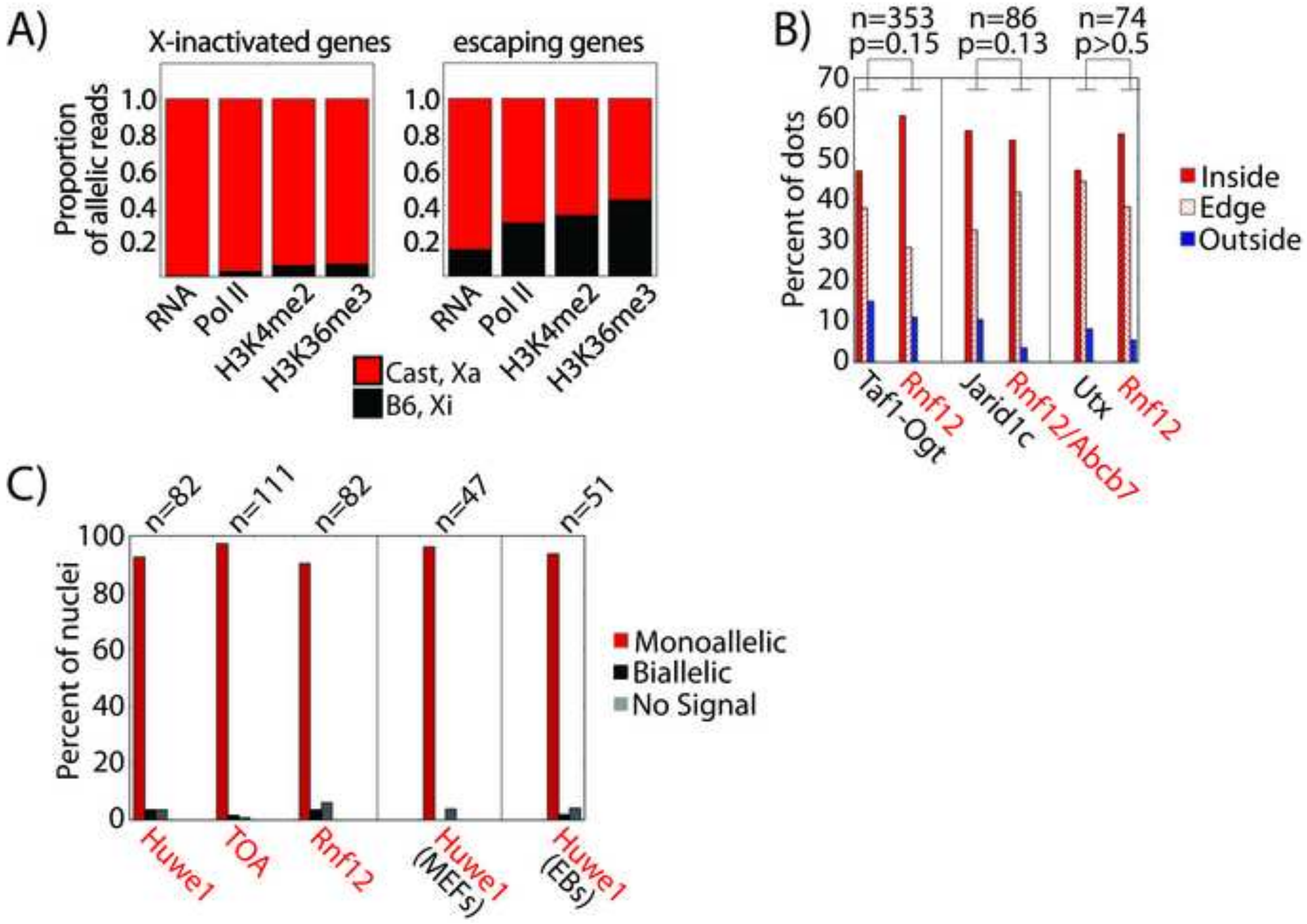


Supplemental Figure_S5



Supplemental Figure_S6

Calabrese_Fig_S6



Determination of XCI Using Allele-Specific RNA-seq

For gene i, let the number of allele-specific RNA-seq reads mapped to the inactivated/activated chromosomes be $n_{i,0}$ and $n_{i,1}$, and let $n_i \equiv n_{i,0} + n_{i,1}$. We first model $n_{i,0}$ by a binomial distribution:

$$p(n_{i,0}|n_i, p_i) = \binom{n_i}{n_{i,0}} p_i^{n_{i,0}} (1 - p_i)^{n_i - n_{i,0}},$$

where p_i indicates the expected proportion of reads from the inactivated chromosome. We further assume that p_i follows a mixture of two beta distributions:

$$f(p_i) = \pi_{i0} f_0(p_i; \alpha_0, \beta_0) + (1 - \pi_{i0}) f_1(p_i; \alpha_1, \beta_1), \tag{1}$$

where $f_0(p_i; \alpha_0, \beta_0)$ and $f_1(p_i; \alpha_1, \beta_1)$ are two beta distributions for inactivated genes and genes that escape inactivation, respectively, and α_0, β_0, α_1, and β_1 are the unknown parameters to be estimated. Known inactivated genes, such as Rnf12, has p_i approaching 0. Therefore, in general, p_i's from $f_0(p_i; \alpha_0, \beta_0)$ are small (e..g, $<$ 0.01), reflecting possible sequencing errors. π_{i0} is the prior probability that gene i is inactivated. We integrate out p_i to obtain the posterior distribution of $n_{i,0}$ in terms of α_0, β_0, α_1, and β_1.

$$p(n_{i,0}|n_i, \alpha_0, \beta_0, \alpha_1, \beta_1) = \int p(n_{i,0}|n_i, p_i) f(p_i) dp_i = \pi_{i0} h_{i0} + (1 - \pi_{i0}) h_{i1}$$

where h_{i0} and h_{i1} are two beta-binomial distributions

$$h_{i0} = \binom{n_i}{n_{i,0}} \frac{B(n_{i,0} + \alpha_0, n_i - n_{i,0} + \beta_0)}{B(\alpha_0, \beta_0)}$$
$$h_{i1} = \binom{n_i}{n_{i,0}} \frac{B(n_{i,0} + \alpha_1, n_i - n_{i,0} + \beta_1)}{B(\alpha_1, \beta_1)}$$

and $B(\alpha, \beta)$ is beta function with parameters α and β. Beta-binomial distribution is a generalization of binomial distribution to allow extra variance, which has been used to model RNA-seq data before [Pickrell et al., 2010]. In this study, the extra variability comes from the fact that each gene has its own proportion of reads escaping inactivation.

For each read, we can obtain a base-calling quality score at the SNP location. We model the prior probability one gene escapes inactivation by a logistic regression with two predictors: the total number of escaping reads and the summation of quality scores of these reads (denoted by q_i):

$$\log\left(\frac{\pi_{i0}}{1-\pi_{i0}}\right) = b_0 + b_1 n_{i,0} + b_2 q_i, \tag{2}$$

where b_0, b_1, and b_2 are regression coefficients to be estimated.

Now we have finished the model setup and there are altogether seven parameters to be estimated: α_0, α_1, β_0, β_1, b_0, b_1, and b_2. We estimated these parameters by Maximum Likelihood approach using Expectation-Maximization (EM) algorithm [Dempster et al., 1977]. For the robustness of the algorithm and based on the prior belief that most of genes are inactivated, we impose an extra restriction that $\pi_{i0} \geq 0.2$. This is equivalent to adding a large penalty $\lambda I_{\pi_0<0.2}$ to the likelihood, where λ is an arbitrary large positive number and $I_{\pi_0<0.2}$ is an indicator function which equals to 1 if $\pi_0 < 0.2$ and 0 otherwise. To maximize this alternative likelihood, we simply maximize the original likelihood and set π_{i0} to be 0.2 if its estimate is smaller than 0.2. Our final results remain similar for any π_{i0} cutoff from 0.05 to 0.3. Given the parameter estimates from the EM algorithm, we can estimate the posterior probability that one gene is inactivated by

$$\hat{\tau}_{i0} = \frac{\hat{\pi}_{i0}\hat{h}_0}{\hat{\pi}_{i0}\hat{h}_{i0} + \hat{\pi}_{i1}\hat{h}_{i1}},$$

where the hat sign $\hat{}$ indicates the estimate of the corresponding parameter. We then assign one gene as activated or inactivated based on $\hat{\tau}_{i0}$. Note that $\hat{\tau}_{i0}$ can also be interpreted as local False Discovery Rate (FDR) [Efron et al., 2001]. If we claim one gene is activated when $\hat{\tau}_{i0} \leq \tau_C$, then the overall FDR is $\sum_i \hat{\tau}_{i0} I_{\hat{\tau}_{i0}\leq\tau_C} / \sum_i I_{\hat{\tau}_{i0}\leq\tau_C}$, where $I_{\hat{\tau}_{i0}\leq\tau_C}$ is an indicator function, which equals to 1 if $\hat{\tau}_{i0} \leq \tau_C$, and 0 otherwise.

References

[Dempster et al., 1977] Dempster, A., Laird, N., Rubin, D., et al., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1):1–38.

[Efron et al., 2001] Efron, B., Tibshirani, R., Storey, J., and Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**(456):1151–1160.

[Pickrell et al., 2010] Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J., Stephens, M., Gilad, Y., and Pritchard, J., *et al.*, 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**(7289):768–772.