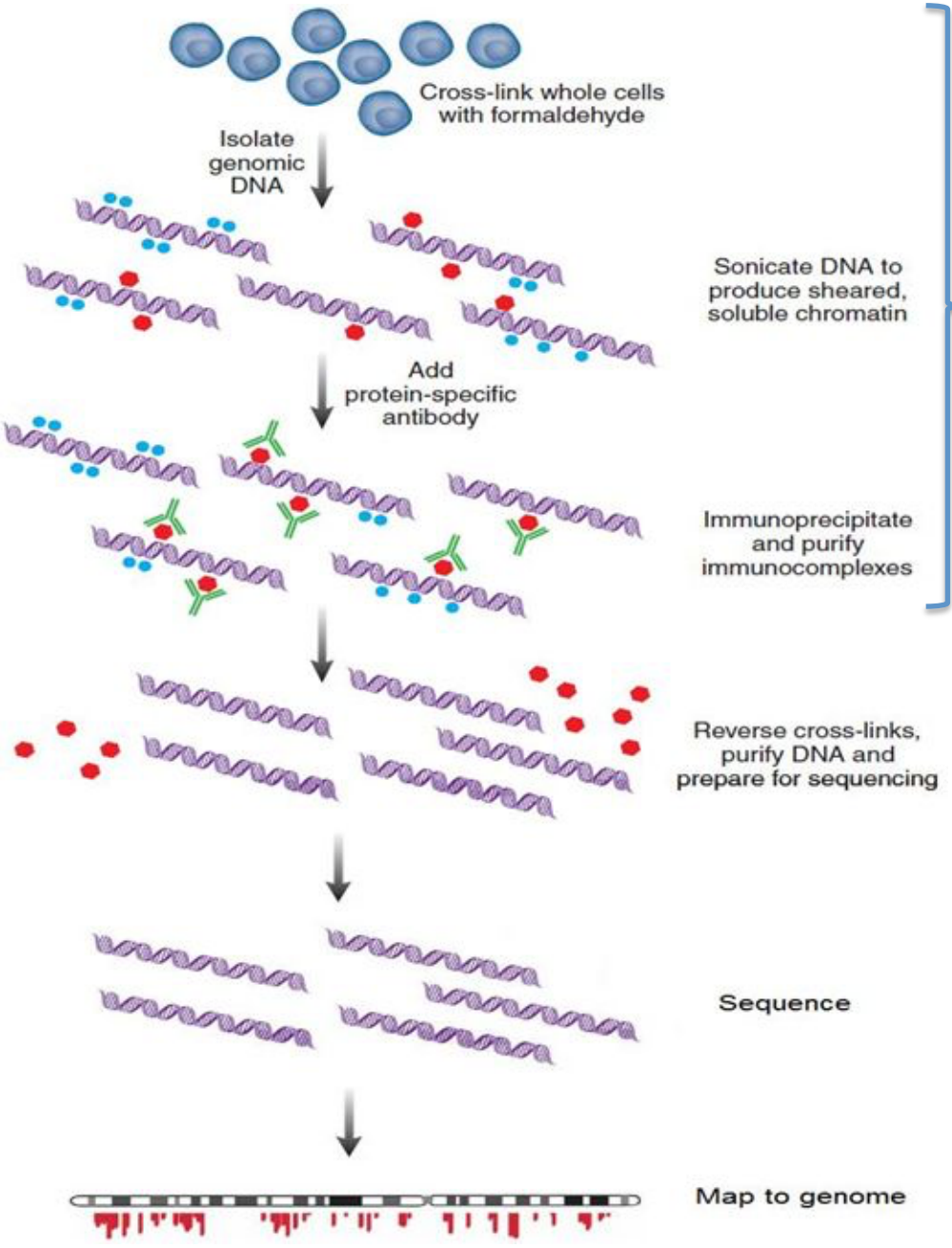
Updates on modERN

Jinrui Xu

06/15/2017

IP in ChIP-Seq:

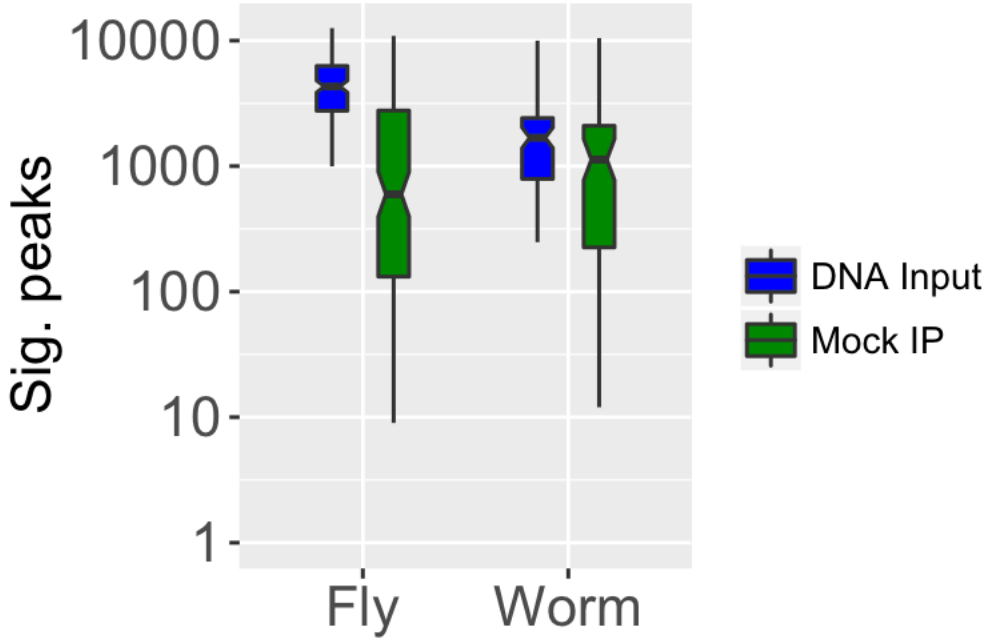
Controlled by DNA Input



Controlled by Mock IP

Sig. peaks using SPP pipeline

- 99 fly and 80 worm TFs with DNA input and mock IP



Two possible explanations

- Reduction of sig. peaks using mock IP
  - Pros: Mock IP controls more steps, thus removes more spurious peaks than DNA input
    - E.g. spurious peaks caused by antibody, if any.
  - Cons: Mock IP has large technical noise, thus reduce *bona fide* peaks using IDR

Mock IP removes both spurious and *bona fide* peaks

Combined method

- To use mock IP, meanwhile alleviate its noise
  - Use DNA input to find candidate peaks
  - Use both DNA input and mock IP to score the peaks for IDR to identify sig. peaks

Scoring model

- Each ChIP-seq has four sets of reads from:
  - IP (a), its DNA input (a'), mock IP (b) and its DNA input (b')
- For each peak region, random variables R_a , $R_{a'}$ R_b and $R_{b'}$ indicate the four numbers of normalized reads

$$\theta_a = \frac{R_a}{R_a + R_{a'}} \sim Beta(r_a + 1, r_{a'} + 1)$$

$$\theta_b = \frac{R_b}{R_b + R_{b'}} \sim Beta(r_b + 1, r_{b'} + 1)$$

, where r represents a realization of R

- For a *bona fide* peak,
  - IP has higher read enrichment than DNA input

$$\theta_a = \frac{R_a}{R_a + R_{a'}} > 0.5$$

  - IP has higher read enrichment than mock IP

$$\theta_a > \theta_b$$

- Thus, prob. of being *bona fide*:

$$P(\theta_a > 0.5 \; and \; \theta_a > \theta_b)$$
$$= \int_{\theta_b > 0}^{\theta_a} \int_{\theta_a > 0.5}^{1} \theta_a^{r_a} (1 - \theta_a)^{1 - r_a} \theta_b^{r_b} (1 - \theta_b)^{1 - r_b} d\theta_a d\theta_b$$

- To calculate the prob.

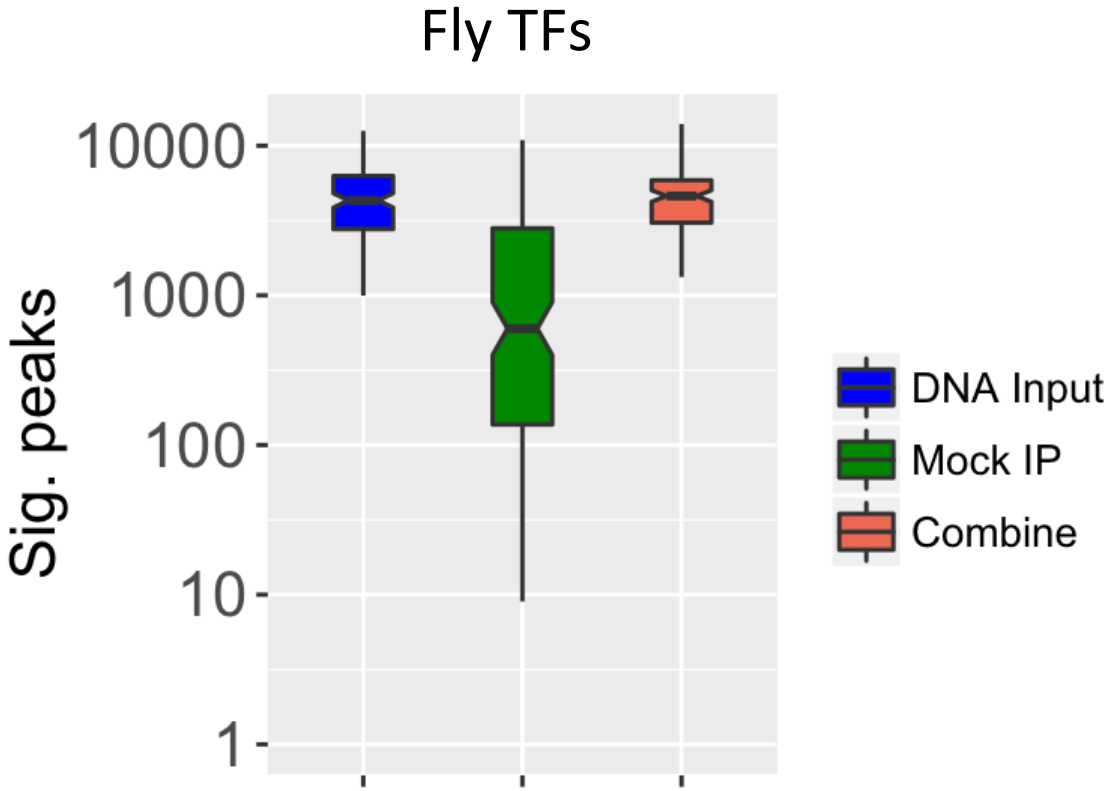
$$P = \int_{\theta_b>0}^{\theta_a} \int_{\theta_a>0.5}^{1} \theta_a^{r_a} (1-\theta_a)^{1-r_a} \theta_b^{r_b} (1-\theta_b)^{1-r_b} d\theta_a d\theta_b$$

  - No analytical solution

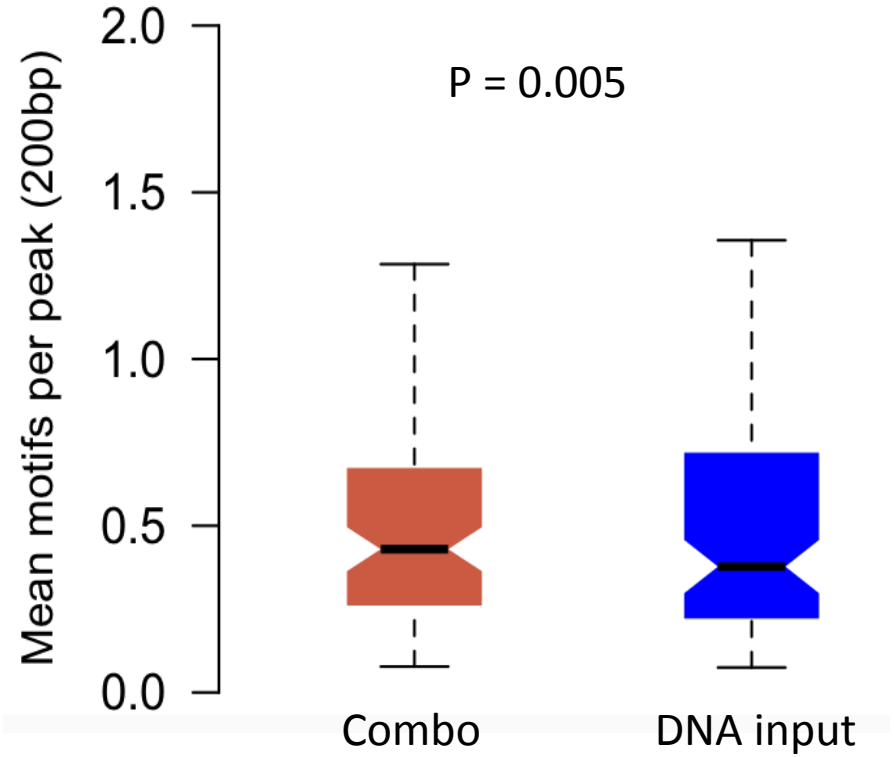
  - *Monte Carlo* method: randomly sample θ_a and θ_b 1 million times, respectively

$$P \approx \frac{\#\ of\ times\ \theta_a > 0.5\ and > \theta_b}{10^6}$$

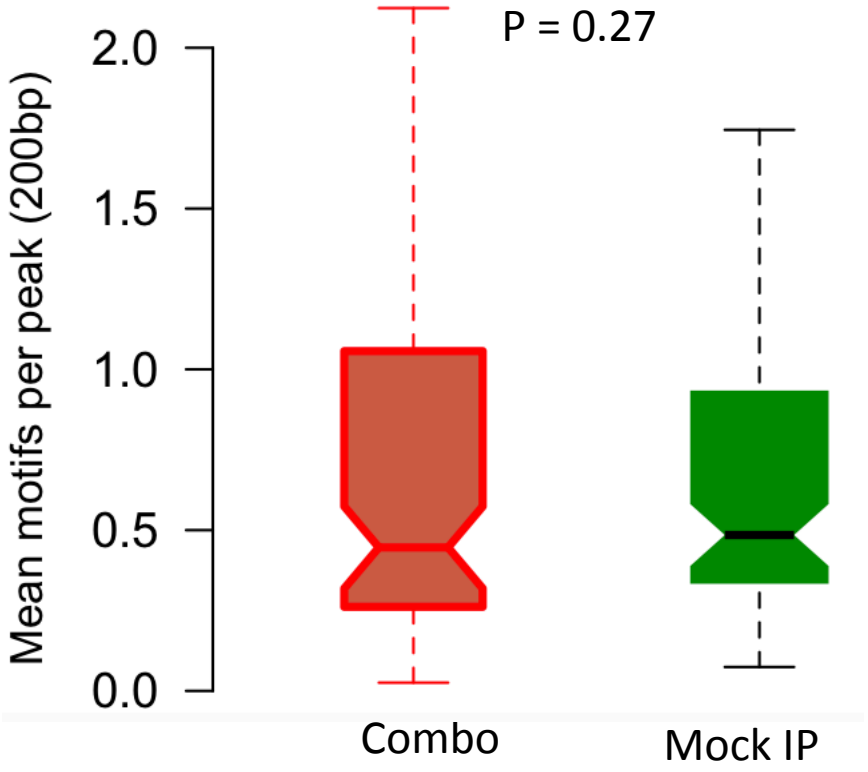
- Using the prob., similar number of significant peaks are identified as using DNA input



- Higher motif enrichment using combined method than using DNA input



- Similar motif enrichment as using mock IP (for same number of sig. peaks)



Future works

- More motif enrichment and expression analyses
  - To compare peak qualities

- Discrepancy among peaks using DNA input, mock IP and combined method
  - To find genome features associated with such

- To optimize the code and/or use parallel computing
  - The combined method is time consuming, due to random sampling to solve the integral