



# Spicy Barbeques

Axel Mosig

Joint work with Sonja Prohaska

UNIVERSITÄT LEIPZIG



# Overview



I.

**Discovering cis-regulatory modules** using `bbq` and other footprinting tools

II.

**Weighted Barbeques and other Variants**

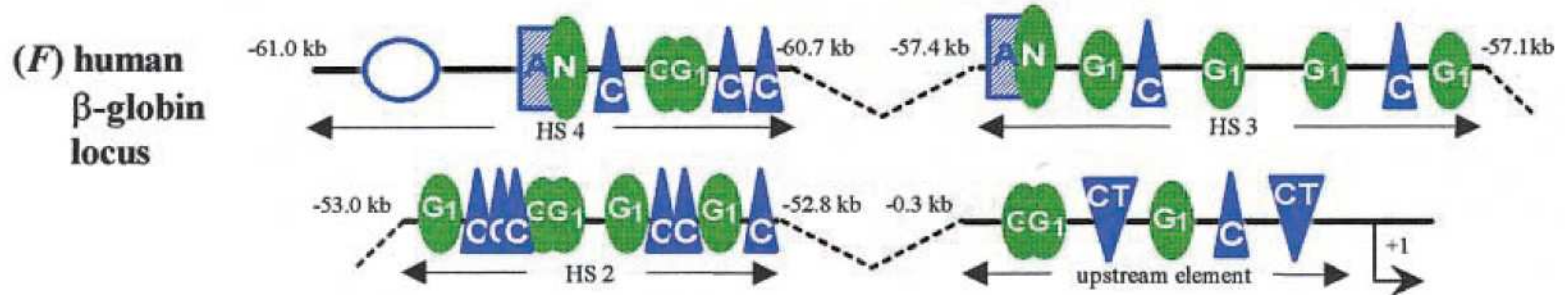
III.

**Fair Barbeques and Complexity**



# Discovering CRMs

- Genes are regulated by transcription factor binding sites
- Binding sites responsible for a single gene occur **clustered**, but may be **shuffled** (Ludwig *et al.* 2000):



(Arnone and Davidson)

- We often have candidates for binding sites
- Find binding sites that occur as clusters

# The bbq approach

- **Given:**

- $n$  candidate binding sites (nucleotide sequences)

$s_1, \dots, s_n$

E.g.:  $s_1 = \text{Meis}$ ,  $s_2 = \text{Pbx-Hox1-5}$ , ...

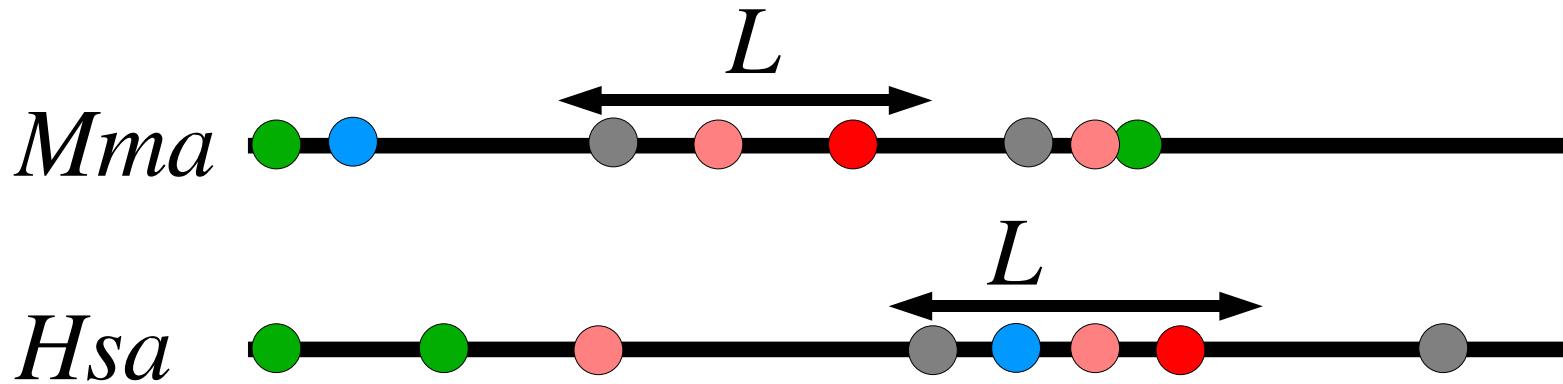
- Genomes  $T_1, \dots, T_K$

E.g.:  $T_1 = \text{Mma}$ ,  $T_2 = \text{Hsa}$ , ...)

- Cluster Length  $L$  (e.g.  $L = 200$ )

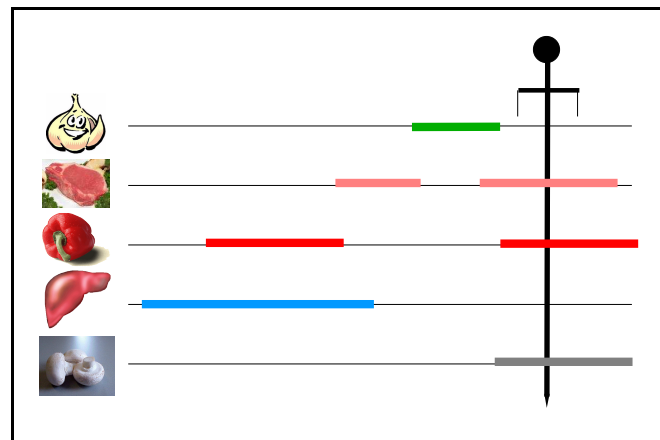
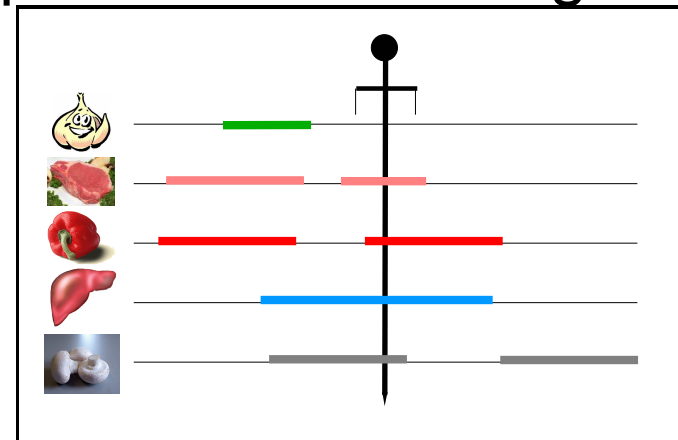
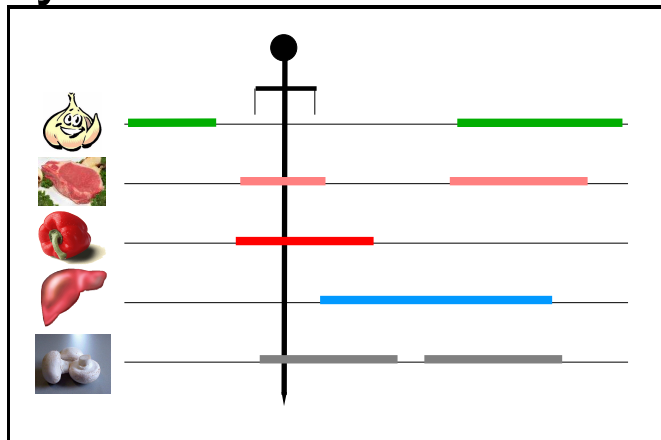
- **Question:** What is the largest possible selection  $S$  of binding sites such that all binding sites in  $S$  occur within an interval of length  $L$  on each  $T_i$ ?

# The bbq approach



# The bbq approach

- In terms of stabbing features: We want to serve as many common features as possible to all our guests



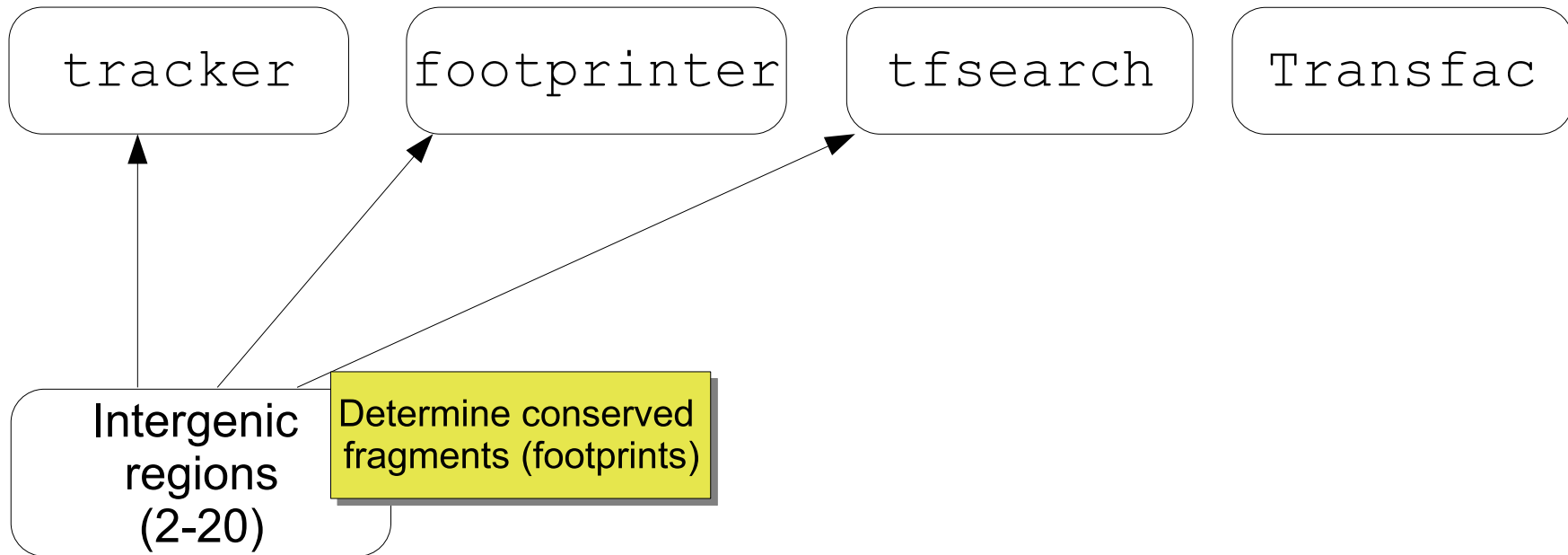
# Application scenario



Intergenic  
regions  
(2-20)

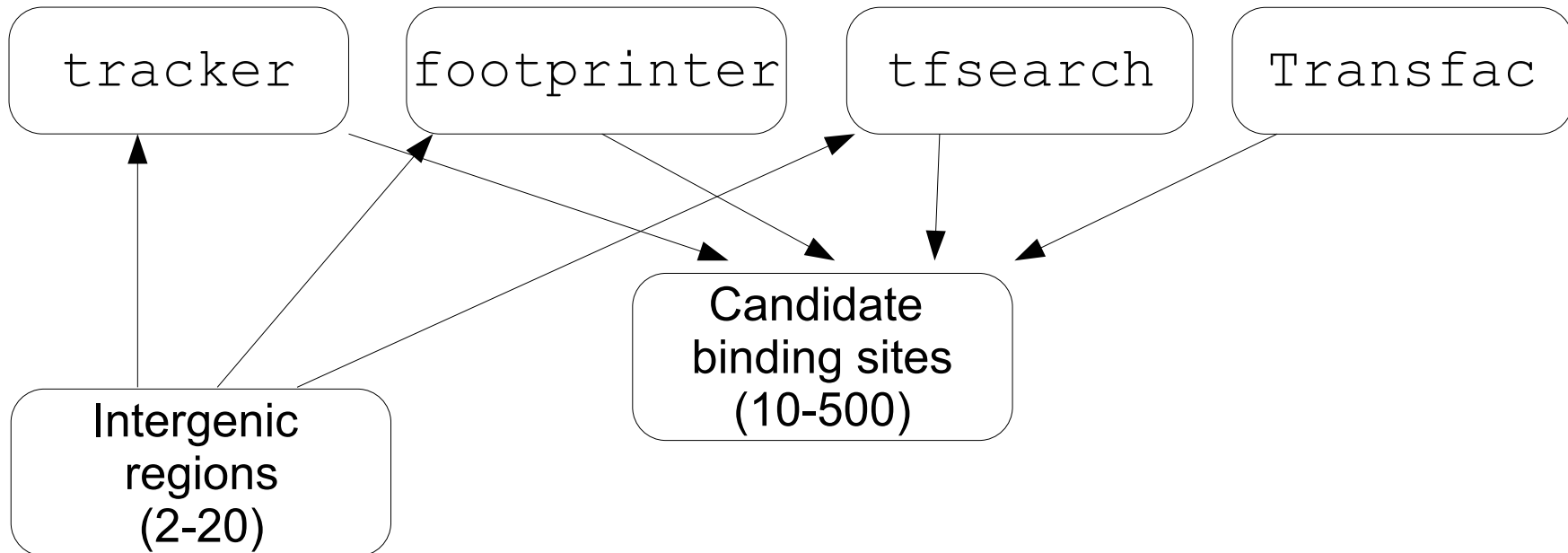


# Application scenario

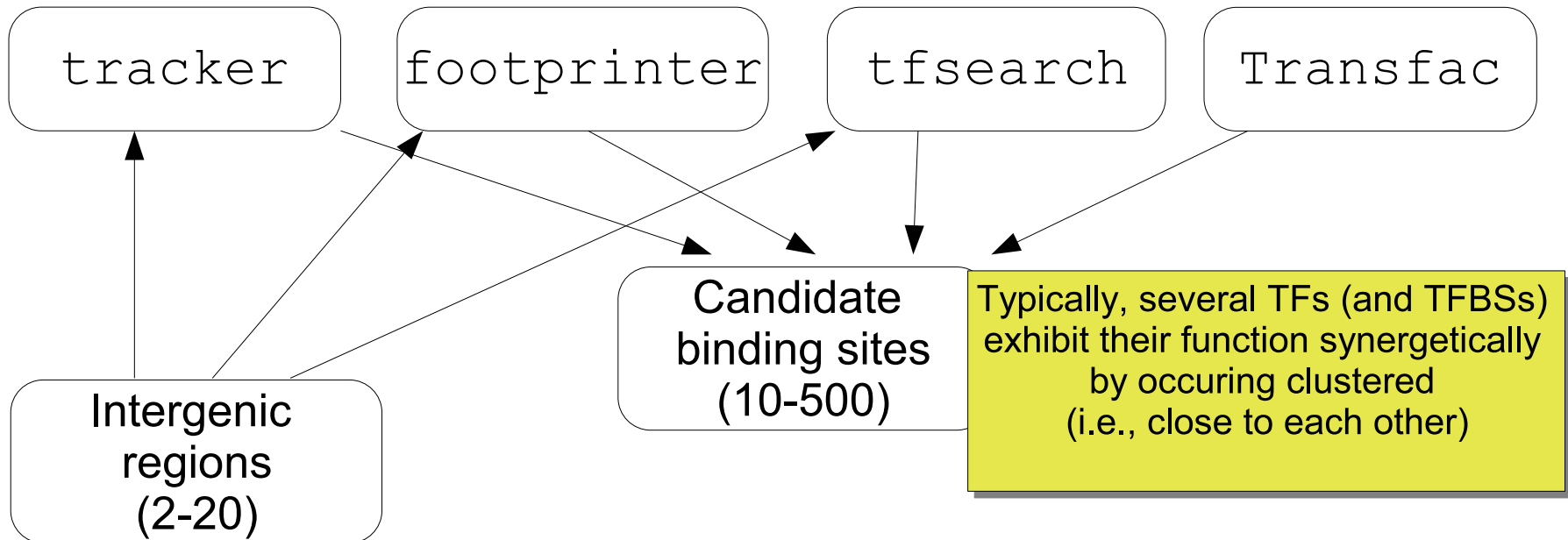




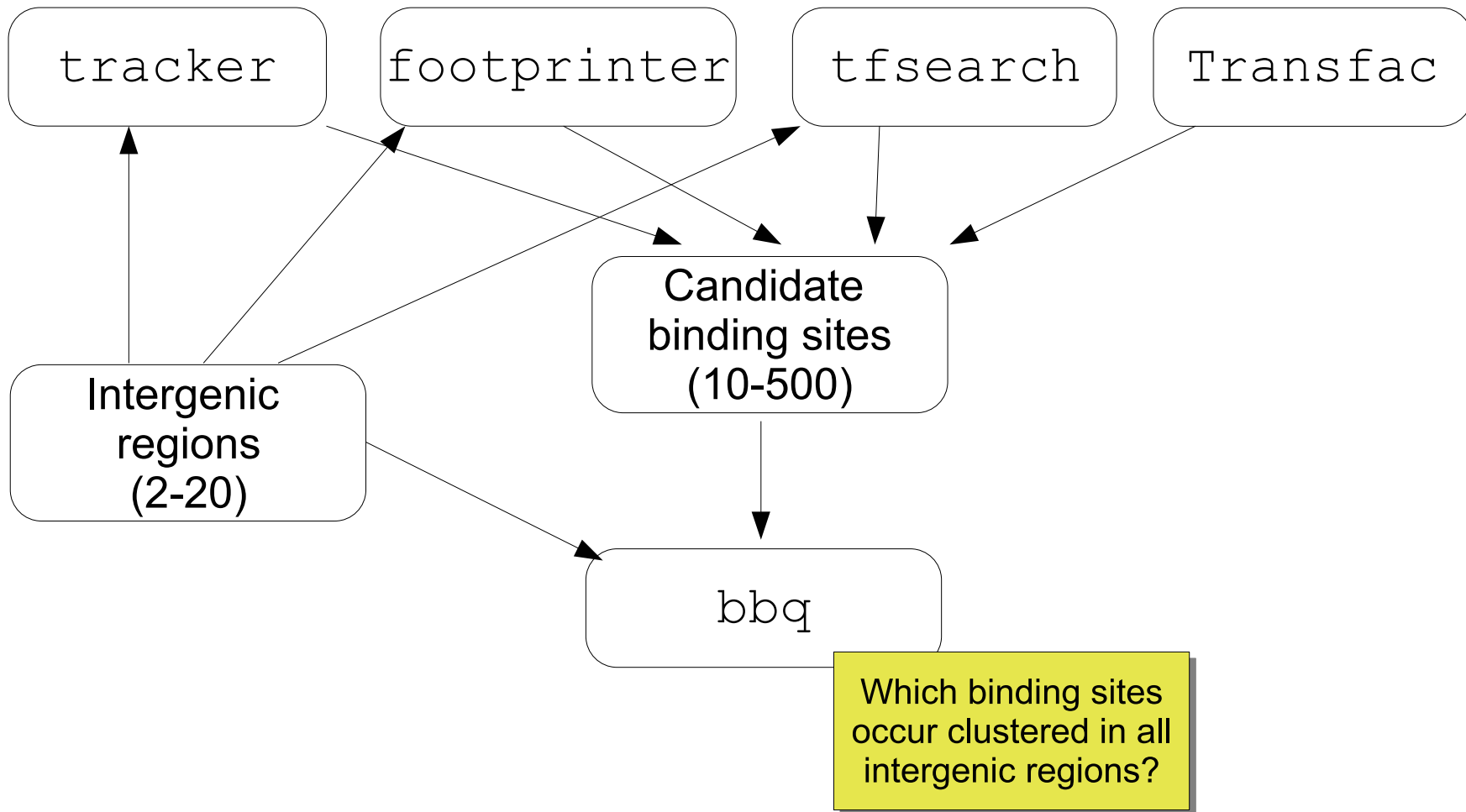
# Application scenario



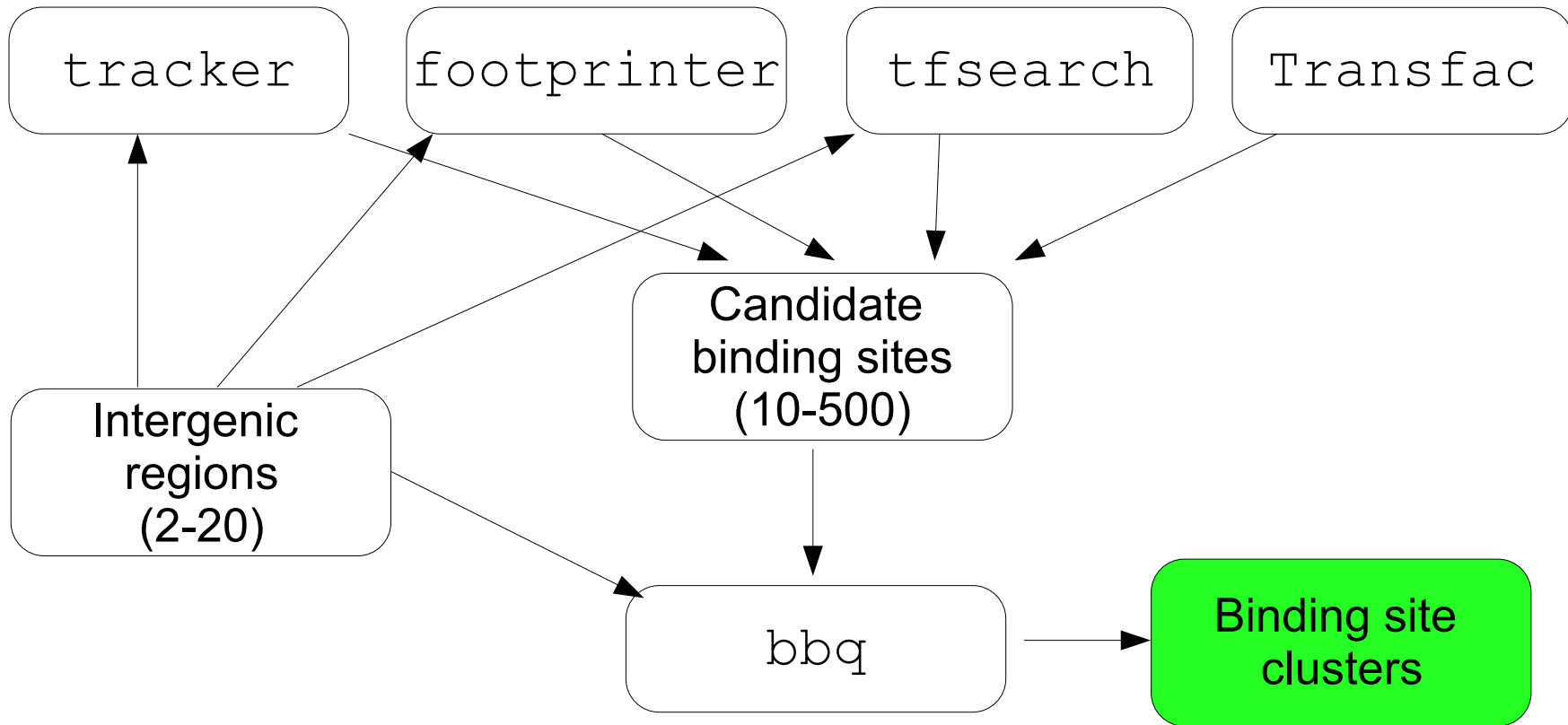
# Application scenario



# Application scenario



# Application scenario



# Weighting schemes

- Straightforward problem setting: Stab maximum number of features
- Refined problem:
  - Assign a weight to each binding site occurrence
  - Maximize the sum of all weights rather than the number of BSs
- $\rightsquigarrow$  what are reasonable weights?

# *p*-value based weighting



- Given genome sequence  $T$
- for each pair  $\alpha, \beta \in \{A, C, G, T\}$  determine:
  - how often does  $\alpha\beta$  occur as a subsequence of  $T$ ?
- $\rightsquigarrow$  dinucleotide-based Markov Model  $M$
- $\rightsquigarrow$  for each candidate binding site  $s$ , obtain probability:
  - $p_M(s) :=$  probability of  $s$  being produced by  $M$



# *p*-value based weighting



- We want to have weights rather than probabilities...
- $w(s) := -\log p_M(s)$
- implemented in `bbq`
- allow certain number of mismatches in the Markov Model



# Other weighting schemes

- Other possibilities for reasonable weights:
  - based on  $f_{s,T} :=$  number of occurrences of  $s$  in  $T$
  - when using Transfac: use position-weight-matrices rather than a fixed string  $s$ 
    - ↪ “occurrences” of a PWM yield a weight as well
- not (yet?) implemented



# Further options supported by bbq

- **weighted** and unweighted optimization
- **grouping**: treat several binding sites as one group;  
↔ maximize number of groups instead of number of BSs
- Maximize (weighted) **multiset intersections** instead of set intersections
- Compute **suboptimal** solutions:
  - best  $h$  solutions or
  - all solutions exceeding threshold weight  $\theta$
- **3 different algorithms**

# Is the best barbeque fair?



- Consider the following **optimal solution** for a barbeque instance:

$C_{\text{Peter}} = \{\text{Beef, Onion, Mushroom, Green Pepper, Pork, Liver, Cucumber, Salmon}\}$

$C_{\text{Sonja}} = \{\text{Beef, Onion, Mushroom}\}$

$C_{\text{Konstantin}} = \{\text{Beef, Onion, Mushroom, Salmon}\}$

- $B := C_{\text{Peter}} \cap C_{\text{Sonja}} \cap C_{\text{Konstantin}}$ ;

$|B| = 3$

- **...is this fair!?**



# Is the best barbeque fair?



- In terms of binding sites:
  - too many “irrelevant” binding sites in a cluster might **disturb function**
  - if we allow no “irrelevant” binding sites, we **miss significant clusters!**
- ↪ introduce parameter  $\delta$  and find best barbeque  $B$  satisfying

$$|C_i \setminus B| \leq \delta$$

for all  $i$ .

- $\delta = 2$  reasonable choice
- $\delta$  small ↪ **computational advantage!**




# How hard is barbeque optimization?

- Decision version of best barbeque problem is NP-complete
- ⇒ **no polynomial-time** algorithm unless  $P = NP$
- 3 algorithmic variations:
  - Exponential in  $K$  (num. of gen. seq.)
  - Exponential in  $m$  (num. of cand. binding sites)
  - **Exponential in  $\delta$**
- $\delta$  is a “hidden” parameter that is usually **small!**

# How hard is barbeque optimization?

- **NP-complete** in general
- “well-behaved” parameters (e.g.,  $\delta$ )
- ↪ **parameterized complexity**
- Can we find good approximations?
- ↪ solution computed is (provably!) only a constant factor worse than optimal solution
- ↪ **structural complexity** (MAX-SNP-hardness, ... ?)
- possibly interesting for some **future work**



---

THX1E6

