BlockBuster2D: chimera reads clustering

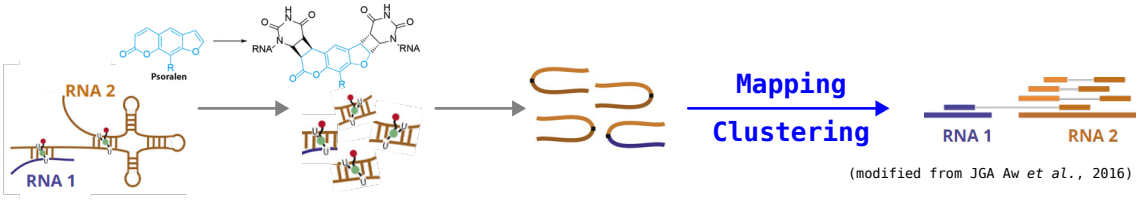
Hua-Ting Yao

Theoretical Biochemistry Group (TBI), University of Vienna

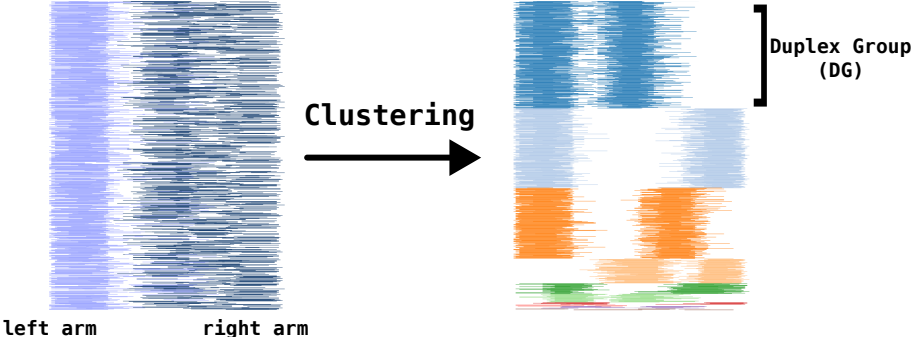
Vienna, Austria

Katrin Gutenbrunner

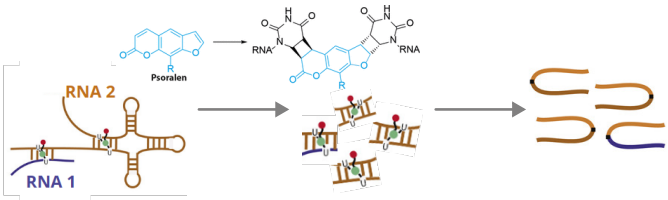
Cluster chimeric reads



(modified from JGA Aw *et al.*, 2016)



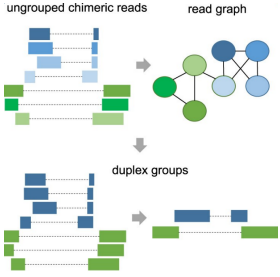
Cluster chimeric reads



Mapping

Clustering

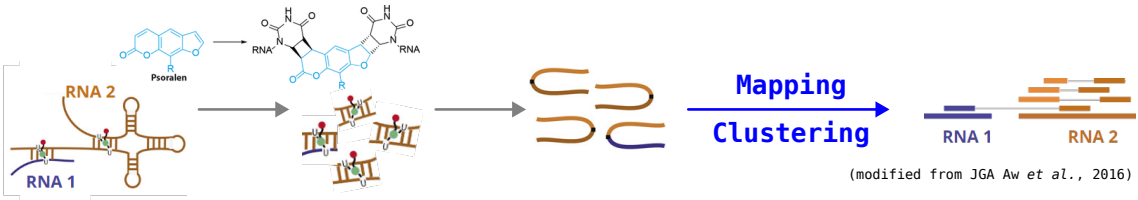
(modified from JGA Aw *et al.*, 2016)



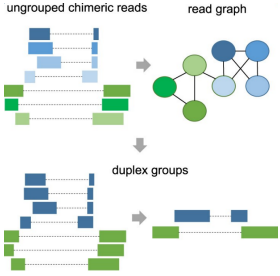
(E. Semenchenko *et al.*, 2025)

- Experimental protocols: COMRADES, PARIS, SPLASH . . .
- Clustering methods: CRSSANT, RNAnue, DuplexDiscovereR . . .
- DG ranking: support read count, hybrid prediction . . .

Cluster chimeric reads



(modified from JGA Aw *et al.*, 2016)

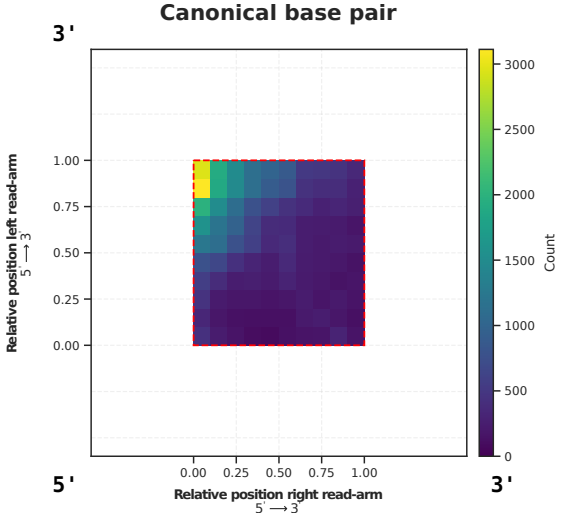


(E. Semenchenko *et al.*, 2025)

- Experimental protocols: COMRADES, PARIS, SPLASH . . .
- Clustering methods: CRSSANT, RNAnue, DuplexDiscovereR . . .
- DG ranking: support read count, hybrid prediction . . .

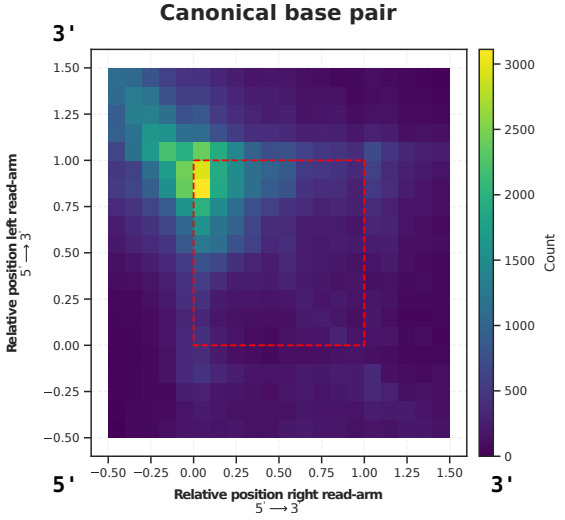
How to locate base pair in DG?

Base pair location in chimeric reads are strongly shifted (18s rRNA as example)



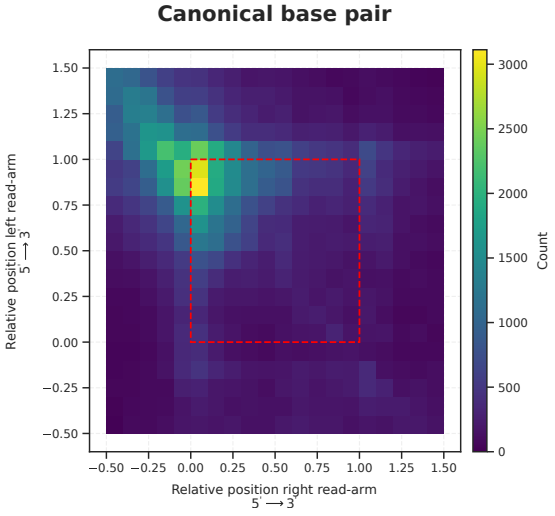
- 18s structure from 6YBS and 6YBW
- Found at least one canonical base pair in 24% chimeric reads

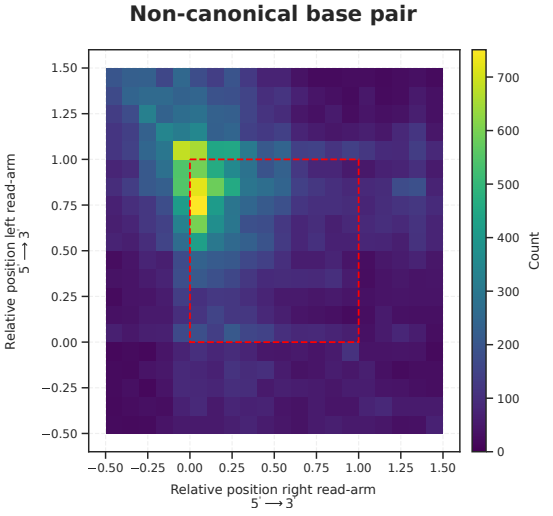
Base pair location in chimeric reads are strongly shifted (18s rRNA as example)



- 18s structure from 6YBS and 6YBW
- Found at least one canonical base pair in 24% chimeric reads
- Found at least one canonical base pair in 52% chimeric reads + extended regions

Base pair location in chimeric reads are strongly shifted (18s rRNA as example)

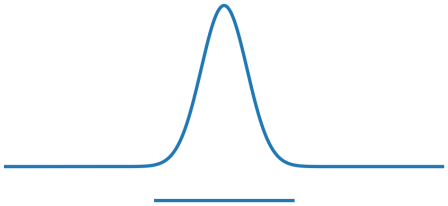




- 18s structure from 6YBS and 6YBW

BlockBuster in 1D

- Identify miRNA (off)target blocks
- Each read is presented by a Gaussian distribution



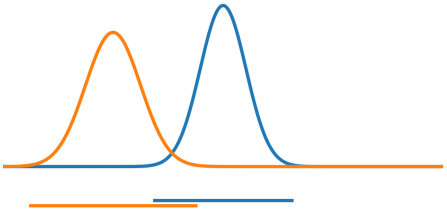
UGCCUAUUACCAGCGAAUGGACGUGUGGCGCACAGUUGAACGUAGACUAGAUAGUAUCAUAACGUAAACAGUAUGUACUAUGUAUCA

Reference

(D. Langenberger *et al.*, 2009)

BlockBuster in 1D

- Identify miRNA (off)target blocks
- Each read is presented by a Gaussian distribution



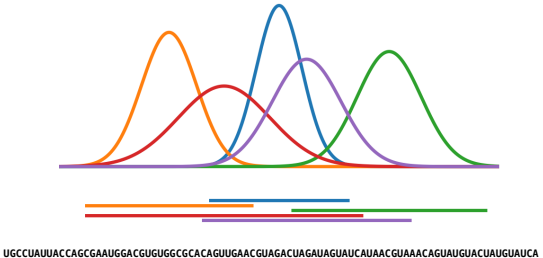
UGCCUAUUACCAGCGAAUGGACGUGUGGCGCACAGUUGAACGUAGACUAGAUAGUAUCAUAACGUAAACAGUAUGUACUAUGUAUCA

Reference

(D. Langenberger *et al.*, 2009)

BlockBuster in 1D

- Identify miRNA (off)target blocks
- Each read is presented by a Gaussian distribution

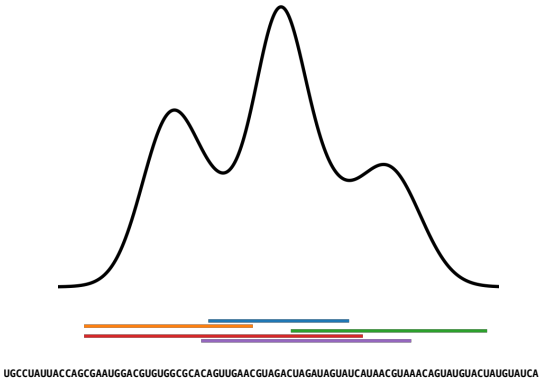


Reference

(D. Langenberger *et al.*, 2009)

BlockBuster in 1D

- Identify miRNA (off)target blocks
- Each read is presented by a Gaussian distribution

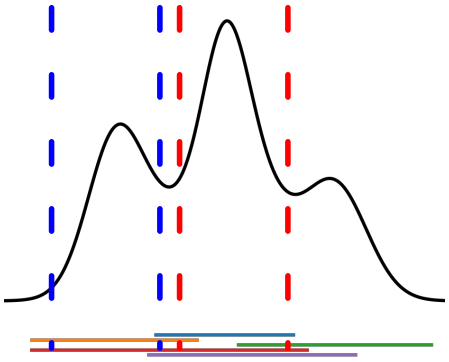


Reference

(D. Langenberger *et al.*, 2009)

BlockBuster in 1D

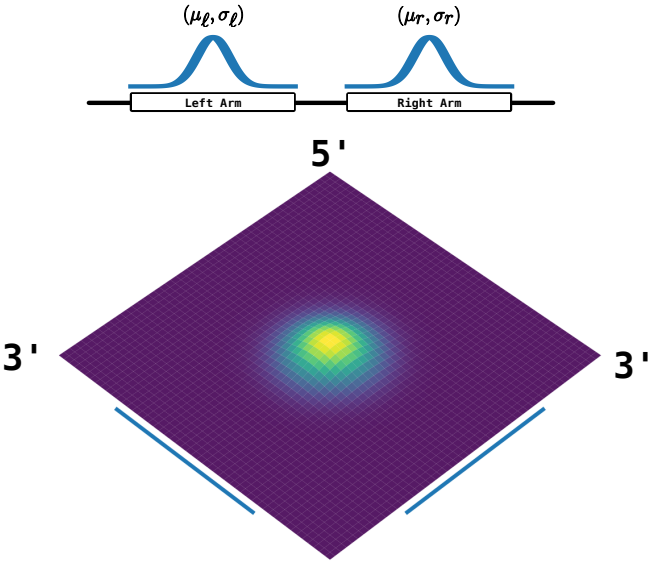
- Identify miRNA (off)target blocks
- Each read is presented by a Gaussian distribution



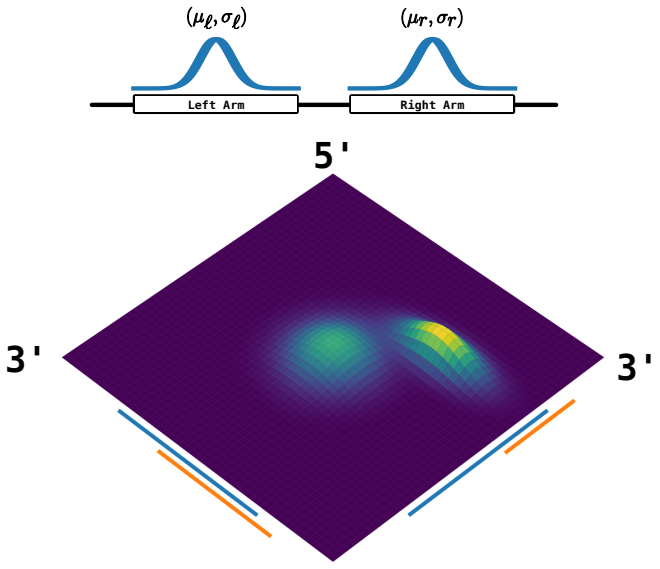
Reference

(D. Langenberger *et al.*, 2009)

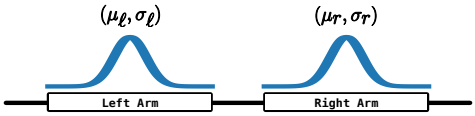
BlockBuster 2D: Let's add one dimension

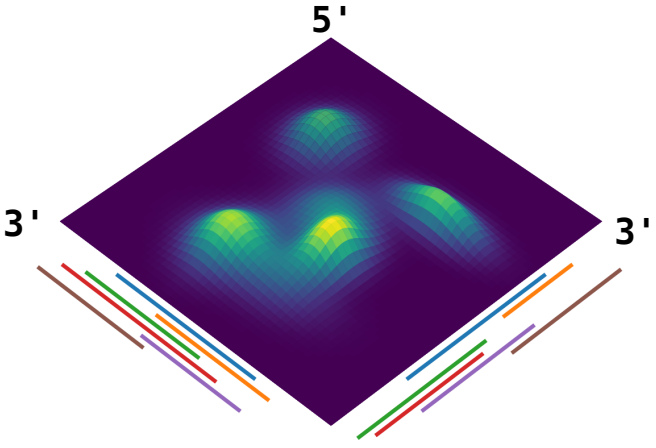


BlockBuster 2D: Let's add one dimension



BlockBuster 2D: Let's add one dimension



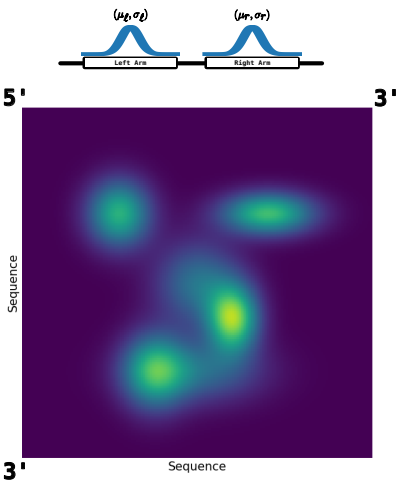


BlockBuster 2D: Cluster extraction

1. Determine highest peak (p_x, p_y) from density function f
2. Include in the block B, all chimeric reads (ℓ, r) s.t.

$$p_x \in [\mu_\ell - (\sigma_\ell + \delta_x), \quad \mu_\ell + (\sigma_\ell + \delta_x)]$$
$$p_y \in [\mu_r - (\sigma_r + \delta_y), \quad \mu_r + (\sigma_r + \delta_y)]$$

δ_x, δ_y: standard deviation of all μ_ℓ, μ_r added in B



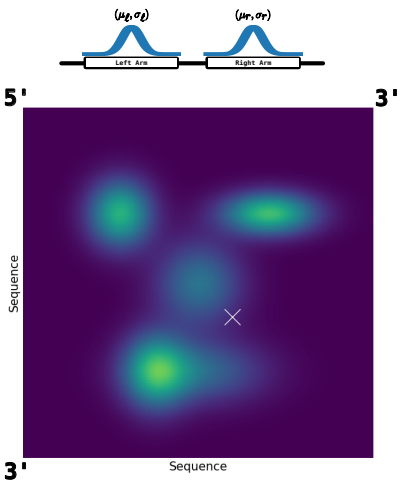
BlockBuster 2D: Cluster extraction

1. Determine highest peak (p_x, p_y) from density function f
2. Include in the block B, all chimeric reads (ℓ, r) s.t.

$$p_x \in [\mu_\ell - (\sigma_\ell + \delta_x), \quad \mu_\ell + (\sigma_\ell + \delta_x)]$$
$$p_y \in [\mu_r - (\sigma_r + \delta_y), \quad \mu_r + (\sigma_r + \delta_y)]$$

   δ_x, δ_y: standard deviation of all μ_ℓ, μ_r added in B

3. Update density function $f \leftarrow f - f_B$
4. Repeat till $f = 0$



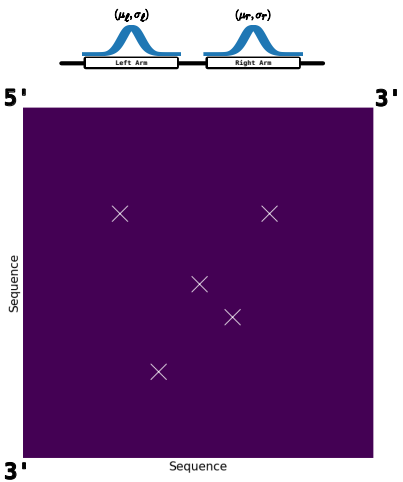
BlockBuster 2D: Cluster extraction

1. Determine highest peak (p_x, p_y) from density function f
2. Include in the block B, all chimeric reads (ℓ, r) s.t.

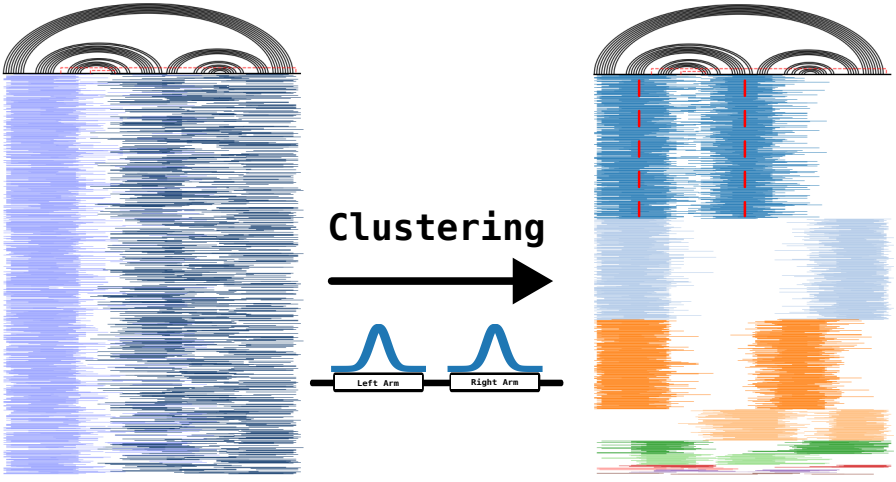
$$p_x \in [\mu_\ell - (\sigma_\ell + \delta_x), \quad \mu_\ell + (\sigma_\ell + \delta_x)]$$
$$p_y \in [\mu_r - (\sigma_r + \delta_y), \quad \mu_r + (\sigma_r + \delta_y)]$$

   δ_x, δ_y: standard deviation of all μ_ℓ, μ_r added in B

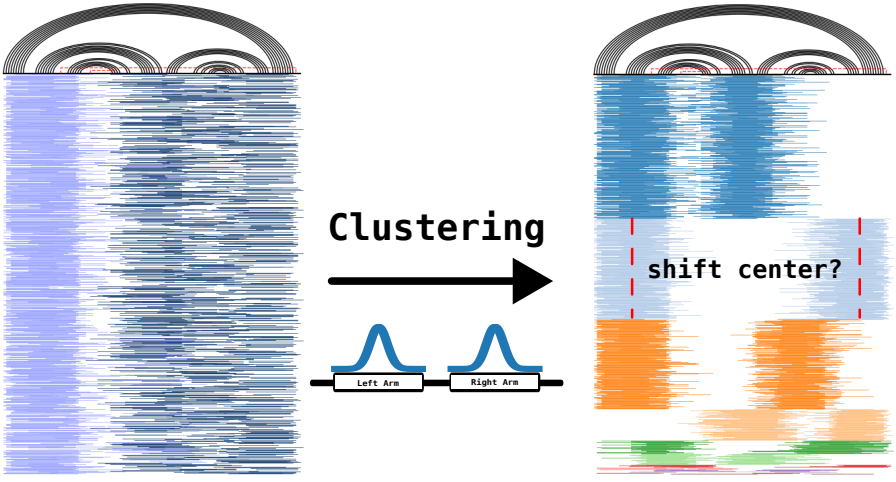
3. Update density function $f \leftarrow f - f_B$
4. Repeat till $f = 0$



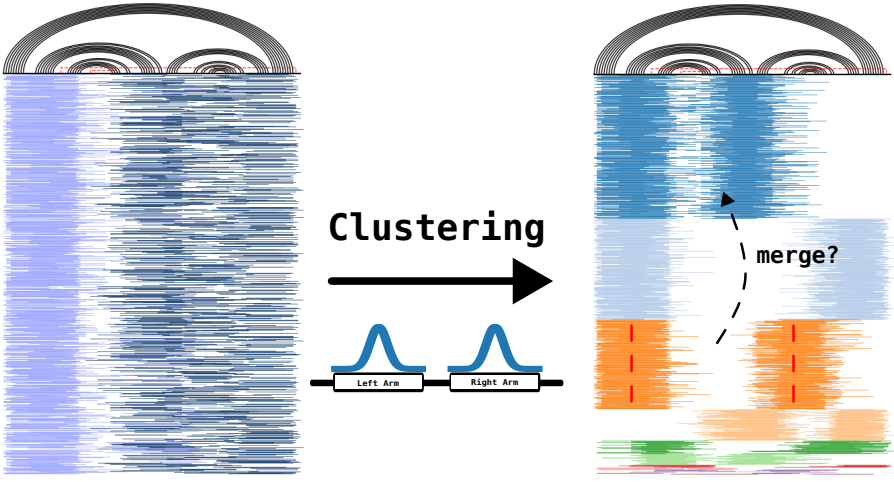
Case study: 5s RNA (PARIS, 8BGU)



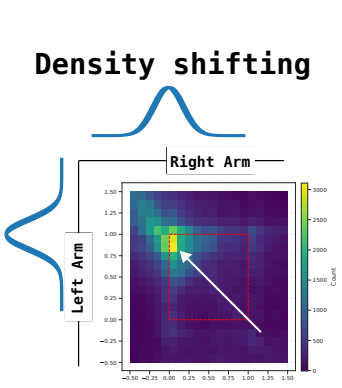
Case study: 5s RNA (PARIS, 8BGU)

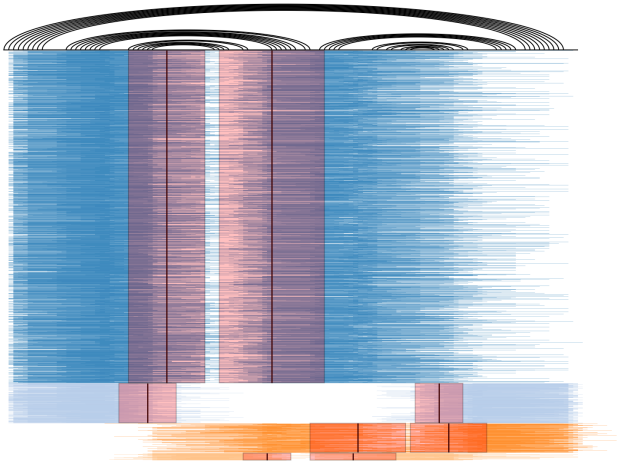


Case study: 5s RNA (PARIS, 8BGU)



Case study: 5s RNA (PARIS, 8BGU)

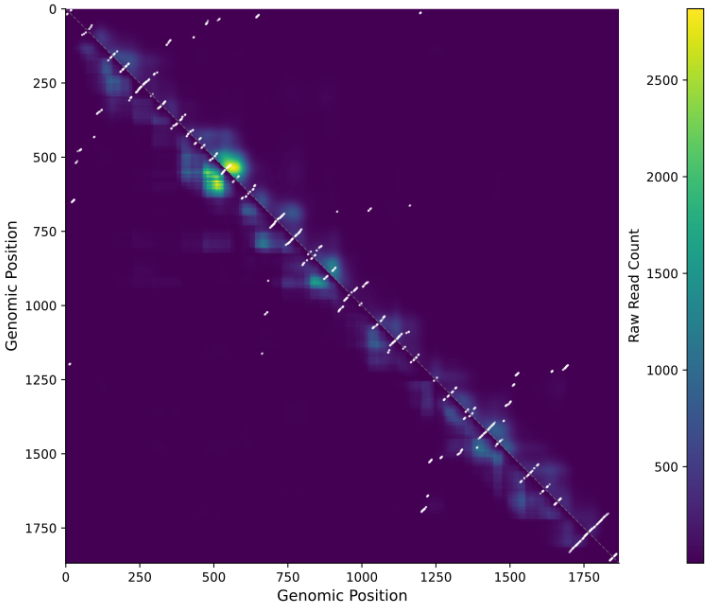




Density shifting solves some problems, but not all . . .

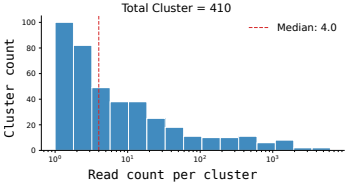
Case study: 18s RNA

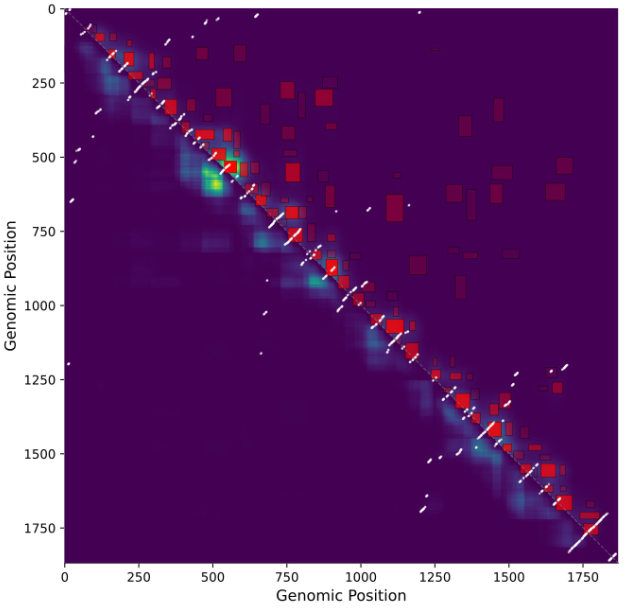
- reference (6YBS & 6YBW)
- white: canonical base pairs
- lower: Chimeric reads heatmap
- upper: Read density function



Case study: 18s RNA

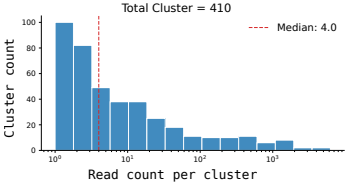
- reference (6YBS & 6YBW)
- white: canonical base pairs
- lower: Chimeric reads heatmap
- upper: Clustered DGs



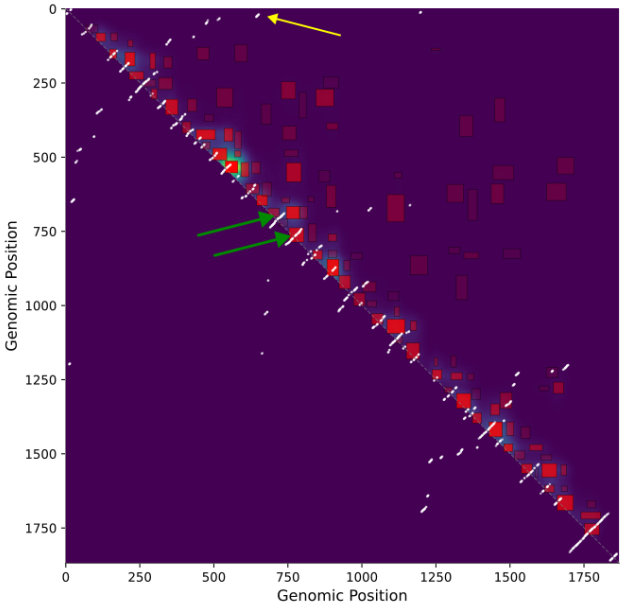


Case study: 18s RNA

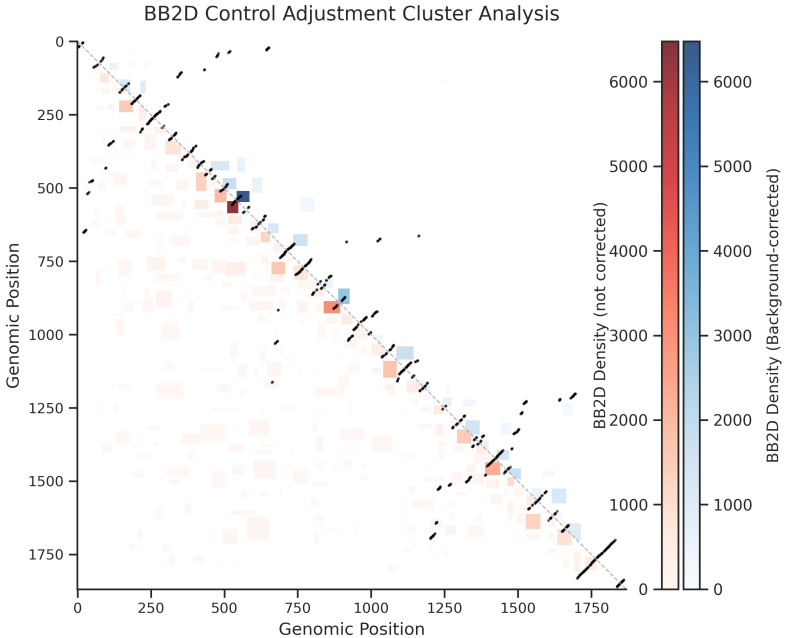
- reference (6YBS & 6YBW)
- white: canonical base pairs
- lower: Chimeric reads heatmap
- upper: Clustered DGs



Structure prediction is still challenging



Case study: 18s RNA (Background noise correction)

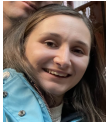


Conclusion & Perspective

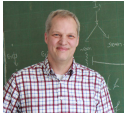
- Base pair location is bias
  → Protocal dependent?
  → How to locating base pair from crosslinking data
- Base pair is close to resulting duplex group
  → Can using empirical base pair distribution help?
  → Still need some effort for parameter calibration for fine tuned analysis
- Extension for PK identification and RNA-RNA interaction
  → combined with RNAnue or DuplexDiscovereR for fast screening

Acknowledgment

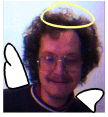
Katrin Gutenbrunner



Paulína Holotová



Ivo Hofacker
Denis Skibinski
Thomas Spicher
Leonhard Sidl
Yingjie Pan
Maximilian Faissner
Cristian A. Velandia-Huerto
⋮



Peter Stadler
Sarah von Löhneysen
Maria Waldl
Ronny Lorenz
⋮

Mimi & Maomao