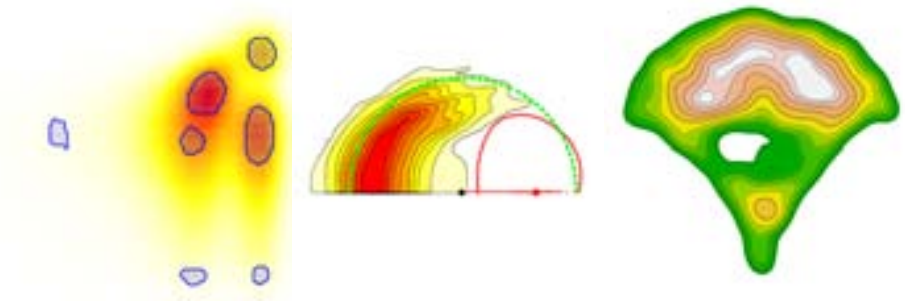
Quantitative statistical analysis of bioloical experimental data



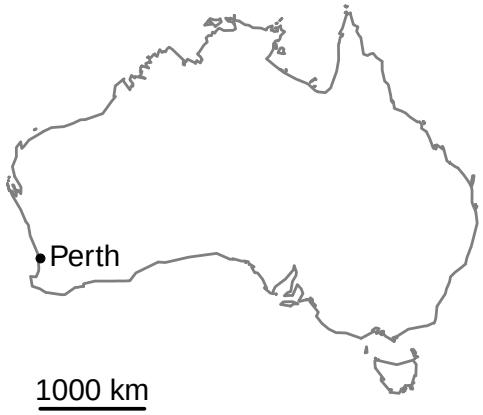
Tarn Duong

Molecular Mechanisms of Intracellular Transport Laboratory (Bruno Goud), Institut Curie, Paris

8 March 2010

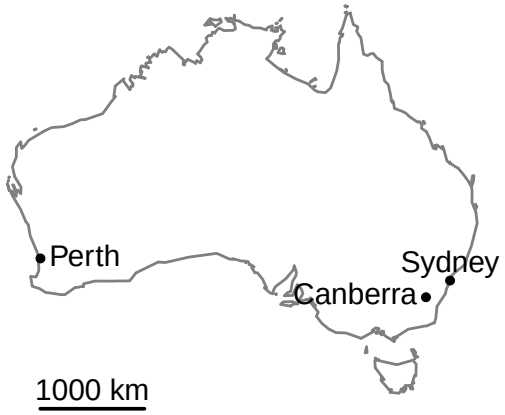
Brief CV

1994–1998 BSc (Math & Comp Science), Perth, Univ. West. Australia ($\sim$ Montpellier 2)



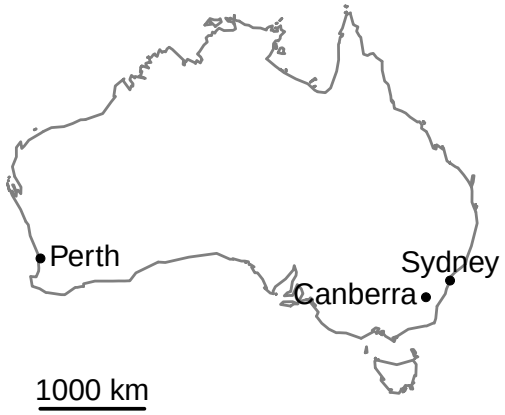
Brief CV

1994–1998	BSc (Math & Comp Science), Perth, Univ. West. Australia ($\sim$ Montpellier 2)
1999–2000	Aust. Bureau of Statistics, Canberra & Sydney, Australia ($\sim$ INSEE)
2001–2004	Ph.D. (Statistics), Perth, Univ. West. Australia
2005	Lecturer, Macquarie Univ., Sydney ($\sim$ Paris 8)
2005–2007	Post-doc, Univ. New South Wales, Sydney ($\sim$ Paris 6/7)



Brief CV

1994–1998	BSc (Math & Comp Science), Perth, Univ. West. Australia ($\sim$ Montpellier 2)
1999–2000	Aust. Bureau of Statistics, Canberra & Sydney, Australia ($\sim$ INSEE)
2001–2004	Ph.D. (Statistics), Perth, Univ. West. Australia
2005	Lecturer, Macquarie Univ., Sydney ($\sim$ Paris 8)
2005–2007	Post-doc, Univ. New South Wales, Sydney ($\sim$ Paris 6/7)
2007–2009	Post-doc, C. Zimmer Group, Institut Pasteur, Paris
2010–present	Post-doc, B. Goud Laboratory, Institut Curie, Paris



Brief CV

1994–1998	BSc (Math & Comp Science), Perth, Univ. West. Australia ($\sim$ Montpellier 2)
1999–2000	Aust. Bureau of Statistics, Canberra & Sydney, Australia ($\sim$ INSEE)
2001–2004	Ph.D. (Statistics), Perth, Univ. West. Australia
2005	Lecturer, Macquarie Univ., Sydney ($\sim$ Paris 8)
2005–2007	Post-doc, Univ. New South Wales, Sydney ($\sim$ Paris 6/7)
2007–2009	Post-doc, C. Zimmer Group, Institut Pasteur, Paris
2010–present	Post-doc, B. Goud Laboratory, Institut Curie, Paris

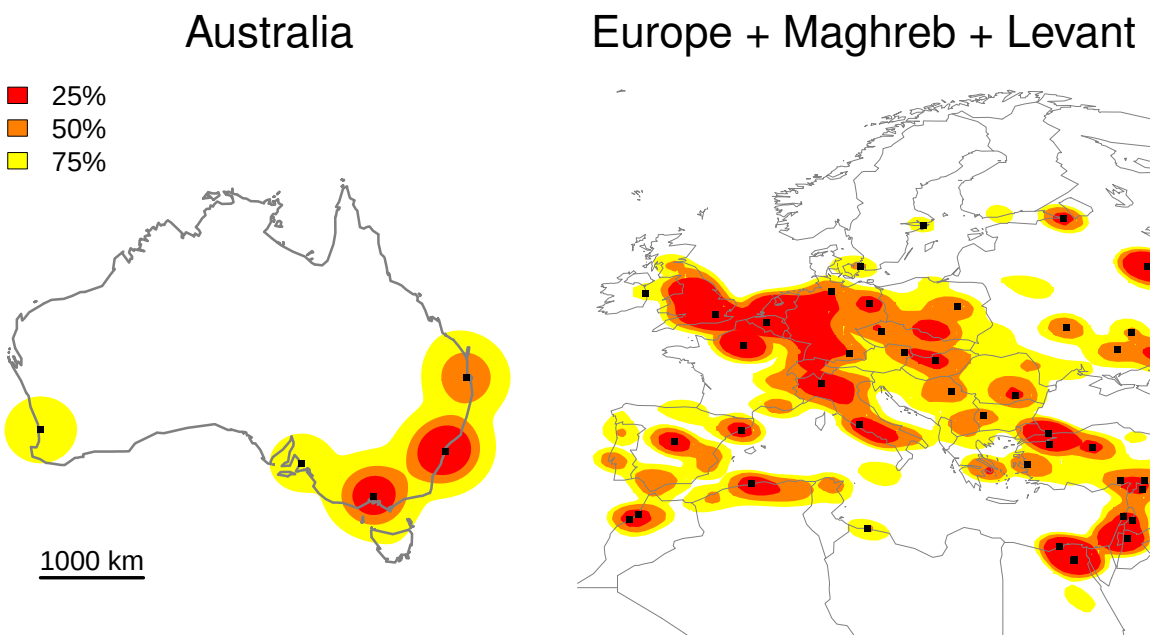
Spatial density (individual towns)

Australia

Europe + Maghreb + Levant

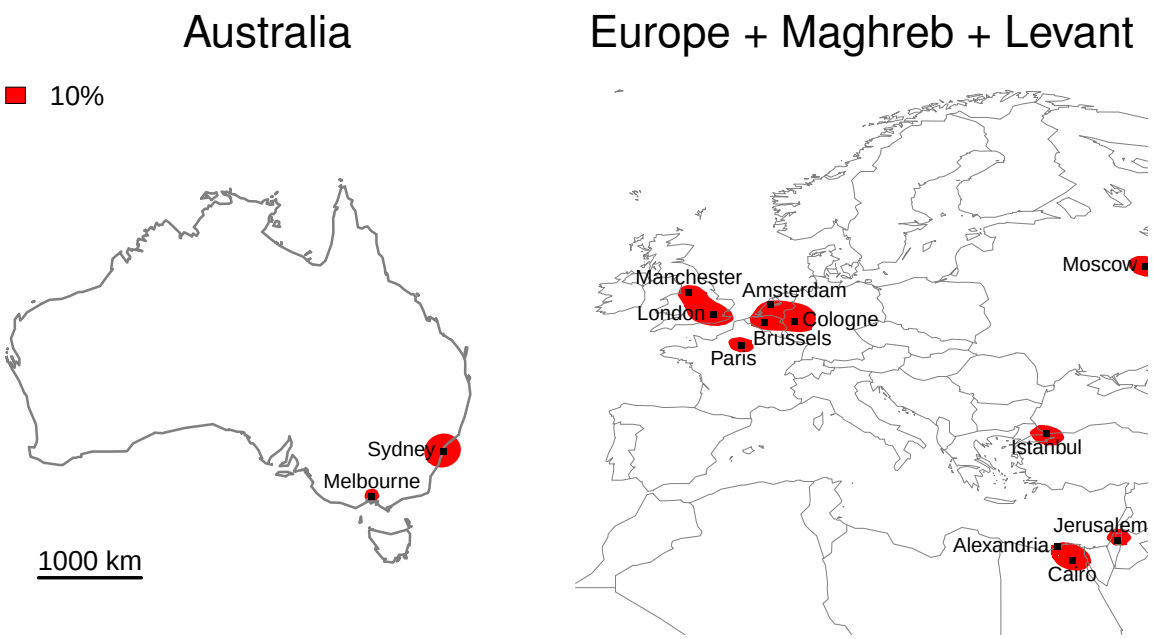
Each dot = 10 000 people

Spatial density (overall population)



Each dot = city $\geq$ 1 000 000 people

Spatial density (overall population)



Some philosophy of mathematics

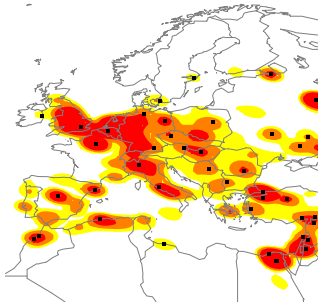
System descriptions fall into two main, complementary categories

Eulerian, particle following

- Builds system behaviour from aggregating individual particle behaviour
- Requires accurate information of all individual particles

Lagrangian, population based

- Focuses on aggregated system behaviour
- Gives less accurate knowledge of individual particles



Data smoothing

Converting point clouds to smooth density functions

$n = 28\,943$
2D co-ordinates

Qualitative
$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n$

Data smoothing

Converting point clouds to smooth density functions

$n = 200$
2D co-ordinates

Qualitative
$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n$

Data smoothing

Converting point clouds to smooth density functions

$n = 200$
2D co-ordinates

Kernels

Qualitative
$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n$

$K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_1), \ldots, K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_n)$

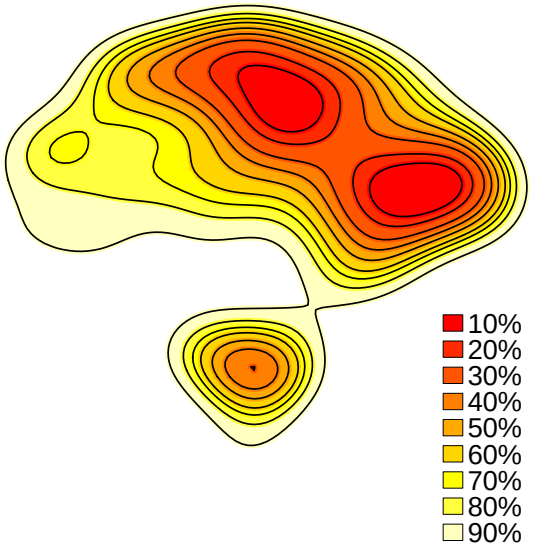
Data smoothing

Converting point clouds to smooth density functions

$n = 200$
2D co-ordinates

Kernels

Kernel density estimate



Qualitative
$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n$

$K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_1), \ldots, K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_n)$

Quantitative
$\hat{f}_{\mathbf{H}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$

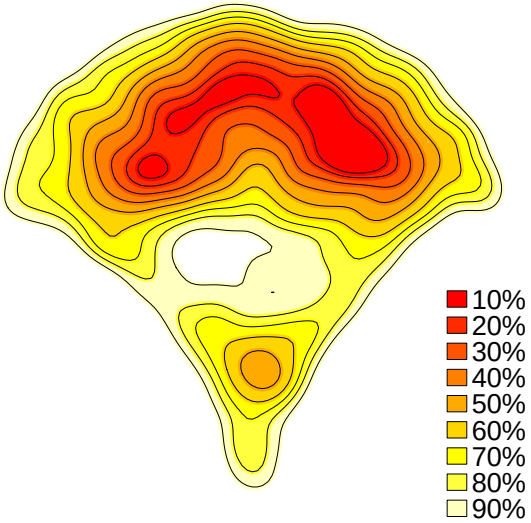
Data smoothing

Converting point clouds to smooth density functions

$n = 28\,943$
2D co-ordinates

Qualitative
$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_n$

Kernel density estimate



Quantitative
$\hat{f}_{\mathbf{H}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$

(Schauer, Duong, Bleakley, Bardin, Brito, Bornens & Goud, *Nature Meth*, revised)

Smoothing parameter estimation

- Estimating smoothing parameter matrix $\mathbf{H}$ is most important factor
- Target (unknown) optimal smoothing parameter: $\mathbf{H} \stackrel{\text{def}}{=} \text{argmin}_{\mathbf{H}} \text{OPT}(\mathbf{H})$
- Estimate: $\hat{\mathbf{H}} = \text{argmin}_{\mathbf{H}} \widehat{\text{OPT}}(\mathbf{H})$
- Convergence: $\hat{\mathbf{H}} \to \mathbf{H}$?

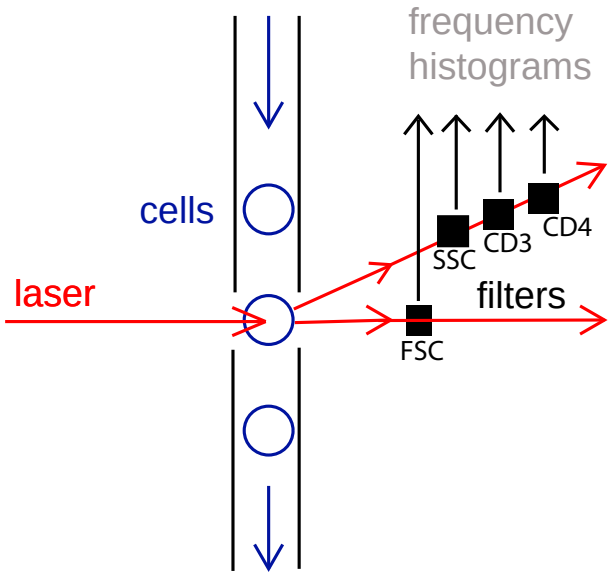
Smoothing parameter estimation

- Estimating smoothing parameter matrix $\mathbf{H}$ is most important factor
- Target (unknown) optimal smoothing parameter: $\mathbf{H} \stackrel{\text{def}}{=} \text{argmin}_{\mathbf{H}} \text{OPT}(\mathbf{H})$
- Estimate: $\hat{\mathbf{H}} = \text{argmin}_{\mathbf{H}} \widehat{\text{OPT}}(\mathbf{H})$
- Convergence: $\hat{\mathbf{H}} \to \mathbf{H}$ at relative rate n^{α} if we can show that

$$\begin{aligned}
\text{MSE}(\hat{\mathbf{H}}) &\stackrel{\text{def}}{=} \mathbb{E}[\text{vec}(\hat{\mathbf{H}} - \mathbf{H})\,\text{vec}(\hat{\mathbf{H}} - \mathbf{H})^T] \\
&= \mathbb{E}[(\partial/\partial\,\text{vec}\,\mathbf{H})(\widehat{\text{OPT}} - \text{OPT})(\mathbf{H})]\,\mathbb{E}[(\partial/\partial\,\text{vec}\,\mathbf{H})(\widehat{\text{OPT}} - \text{OPT})(\mathbf{H})]^T \\
&\quad + \text{Var}[(\partial/\partial\,\text{vec}\,\mathbf{H})(\widehat{\text{OPT}} - \text{OPT})(\mathbf{H})] \\
&= O(n^{2\alpha})(\text{vec}\,\mathbf{H})(\text{vec}^T\,\mathbf{H}).
\end{aligned}$$

(Duong & Hazelton, *J. Nonparametric Stat.*, 2003; Duong & Hazelton, *J. Multivariate Analysis*, 2005 ; Duong & Hazelton, *Scandinavian J. Stat.*, 2005)

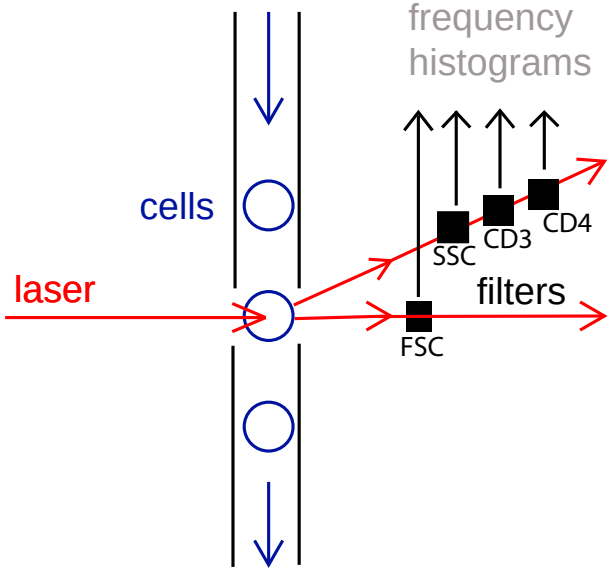
Automatic gating for flow cytometry (FACS) data



Schematic for flow cytometer machine

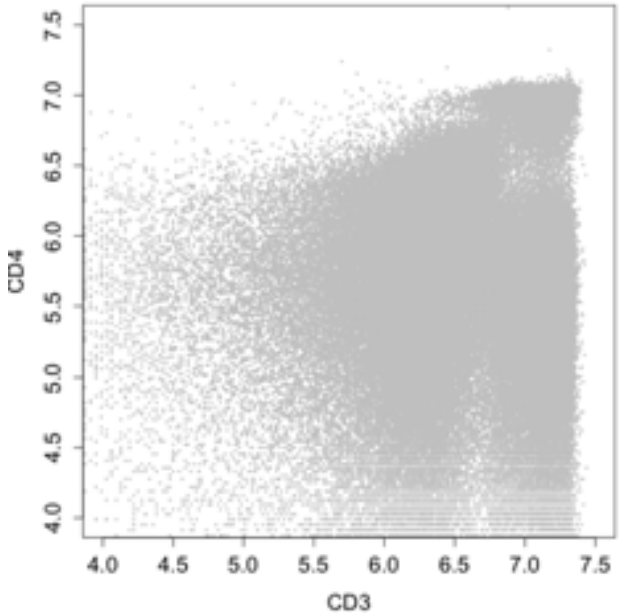
Automatic gating for flow cytometry (FACS) data

How to choose sub-populations of interest for further analysis from $\sim$ 100 000 cells?



Schematic for flow cytometer machine

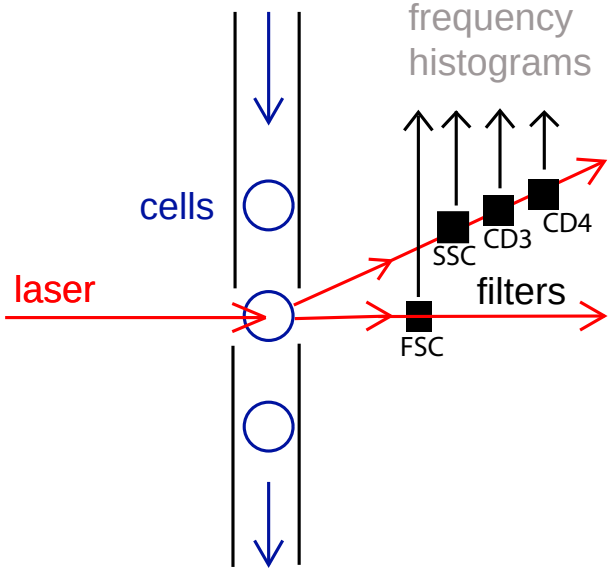
$n = 146\,740$ (CD3, CD4)



2D fluoresence hisotgrams

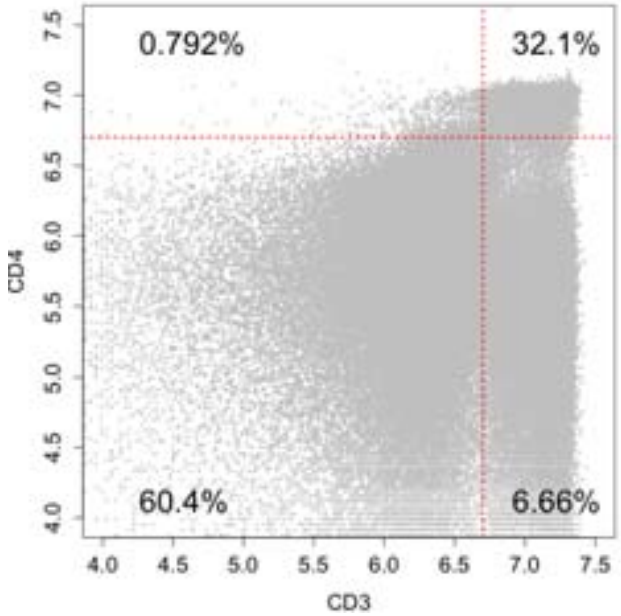
Automatic gating for flow cytometry (FACS) data

How to choose sub-populations of interest for further analysis from $\sim$ 100 000 cells?



Schematic for flow cytometer machine

$n = 146\,740$ (CD3, CD4)

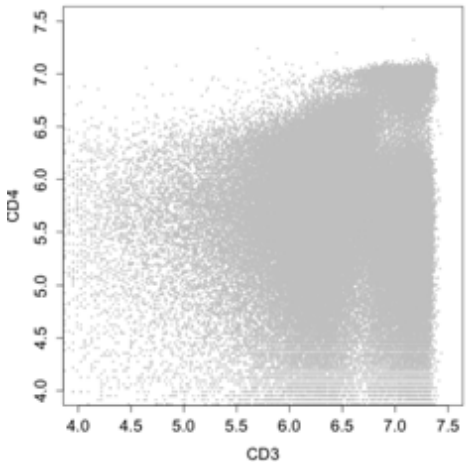


2D fluoresence hisotgrams

- Manual gates: rectangular gates chosen subjectively by eye, informed by experience
- Not reproducible (even by same person)
- Rectangular gates do not correspond naturally to sub-populations
- Automatic, data-shaped shaped gates

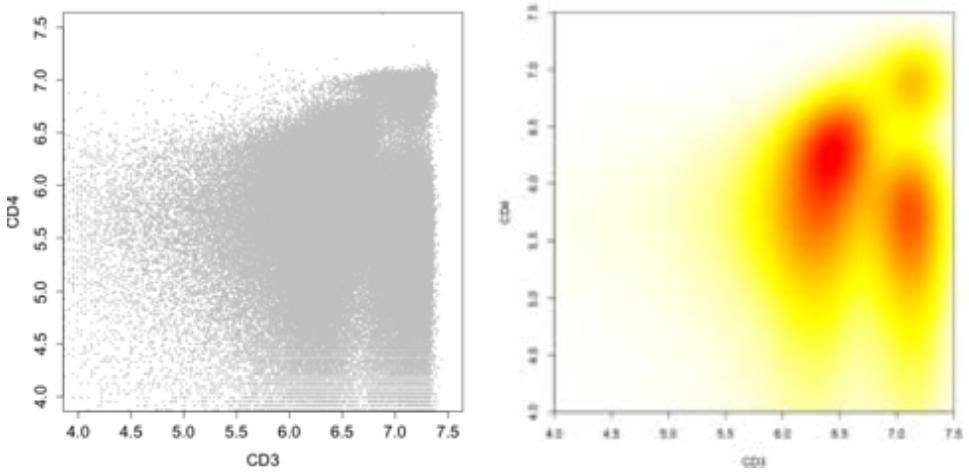
Significant curvature regions

- Sub-population $\stackrel{\text{def}}{=}$ region with high local density f $\stackrel{\text{def}}{=}$ modal region
- Modal region $\stackrel{\text{def}}{=}$ $\{\boldsymbol{x} : \mathrm{D}^2 f(\boldsymbol{x})$ is negative definite$\}$ where $\mathrm{D}^2 f$ is the Hessian matrix of second order partial derivatives of f



Significant curvature regions

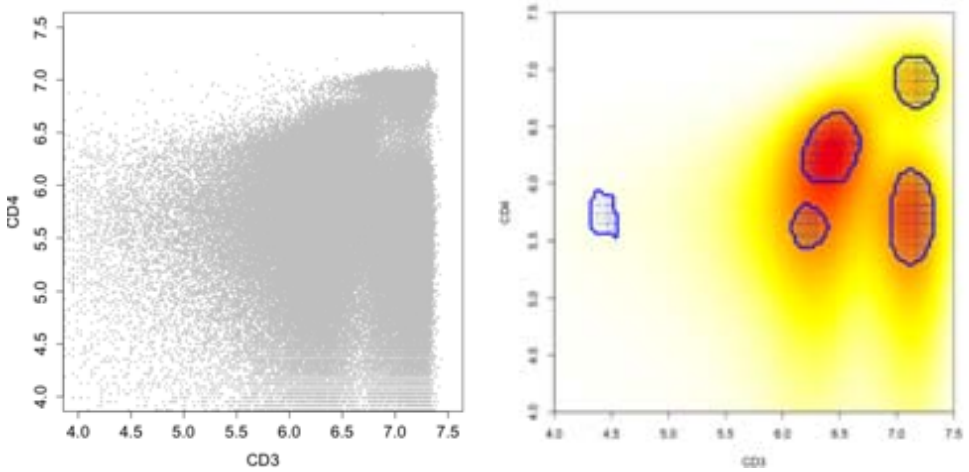
- Sub-population $\stackrel{\text{def}}{=}$ region with high local density f $\stackrel{\text{def}}{=}$ modal region
- Modal region $\stackrel{\text{def}}{=}$ $\{\boldsymbol{x} : \mathrm{D}^2 f(\boldsymbol{x})$ is negative definite$\}$ where $\mathrm{D}^2 f$ is the Hessian matrix of second order partial derivatives of f



- Convert data point cloud to kernel density estimate $\hat{f}_{\mathbf{H}}$

Significant curvature regions

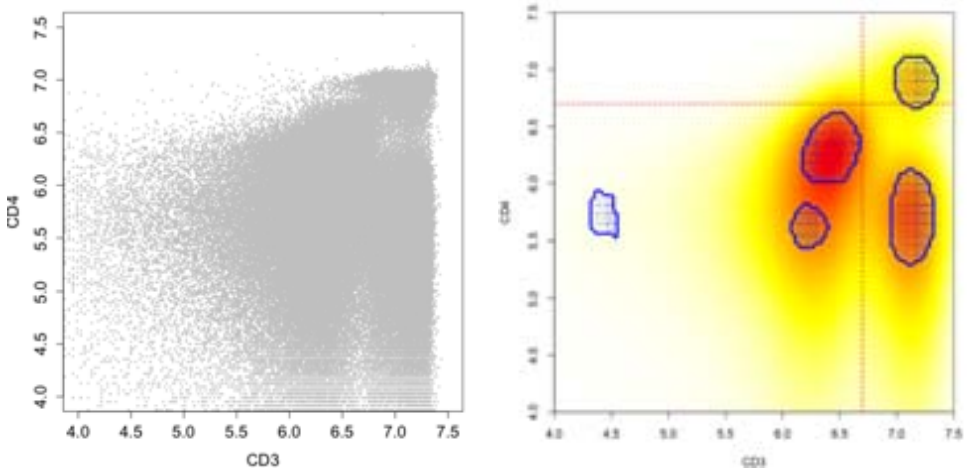
- Sub-population $\stackrel{\text{def}}{=}$ region with high local density f $\stackrel{\text{def}}{=}$ modal region
- Modal region $\stackrel{\text{def}}{=}$ $\{\boldsymbol{x} : \mathrm{D}^2 f(\boldsymbol{x})$ is negative definite$\}$ where $\mathrm{D}^2 f$ is the Hessian matrix of second order partial derivatives of f



- Convert data point cloud to kernel density estimate $\hat{f}_{\mathbf{H}}$
- Modal region estimate = significant curvature region = $\{\boldsymbol{x}$: reject $H_0 : \|\mathrm{D}^2 \hat{f}_{\mathbf{H}}(\boldsymbol{x})\|^2 = 0$ and $\mathrm{D}^2 \hat{f}_{\mathbf{H}}(\boldsymbol{x})$ is positive definite$\}$
- Null distribution of $\|\hat{\boldsymbol{\Sigma}}_{\mathbf{H}}(\boldsymbol{x})^{-1/2} \operatorname{vec} \mathrm{D}^2 \hat{f}_{\mathbf{H}}(\boldsymbol{x})\|^2$ is approx $\chi^2(4)$ (chi-squared distn with 4 d.f.)
(Cowling, Duong, Koch & Wand, 2008, Comp. Stat. Data Analysis)

Significant curvature regions

- Sub-population $\stackrel{\text{def}}{=}$ region with high local density f $\stackrel{\text{def}}{=}$ modal region
- Modal region $\stackrel{\text{def}}{=}$ $\{\boldsymbol{x} : \mathrm{D}^2 f(\boldsymbol{x})$ is negative definite$\}$ where $\mathrm{D}^2 f$ is the Hessian matrix of second order partial derivatives of f



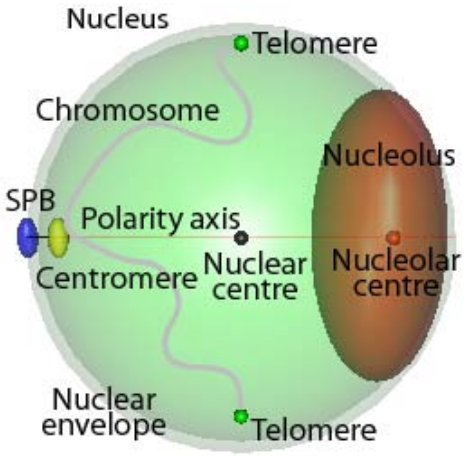
- Convert data point cloud to kernel density estimate $\hat{f}_{\mathbf{H}}$
- Modal region estimate = significant curvature region = $\{\boldsymbol{x}$: reject $H_0 : \|\mathrm{D}^2 \hat{f}_{\mathbf{H}}(\boldsymbol{x})\|^2 = 0$ and $\mathrm{D}^2 \hat{f}_{\mathbf{H}}(\boldsymbol{x})$ is positive definite$\}$
- Null distribution of $\|\hat{\boldsymbol{\Sigma}}_{\mathbf{H}}(\boldsymbol{x})^{-1/2} \operatorname{vec} \mathrm{D}^2 \hat{f}_{\mathbf{H}}(\boldsymbol{x})\|^2$ is approx $\chi^2(4)$ (chi-squared distn with 4 d.f.)
(Cowling, Duong, Koch & Wand, 2008, Comp. Stat. Data Analysis)

Spatial organisation of genomic DNA inside cell nuclei

What is the relationship between spatial location of genomic loci and their function?

For *Saccaromyces cerevisiae* yeast

- Polarity axis: Spindle Pole Body (SPB) (MTOC) - Nuclear centre - Nucleolar centre
- Single nucleolus mostly excludes genomic DNA
- SPB embedded in nuclear envelope
- Chromosome attached at centromere, centromere attached to SPB via microtubule
- Telomeres (chromosome extremities) preferentially localise at nuclear envelope



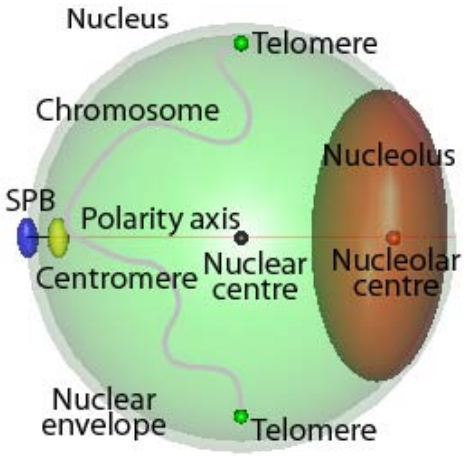
Schematic for single chromosome inside yeast nucleus

Spatial organisation of genomic DNA inside cell nuclei

What is the relationship between spatial location of genomic loci and their function?

For *Saccaromyces cerevisiae* yeast

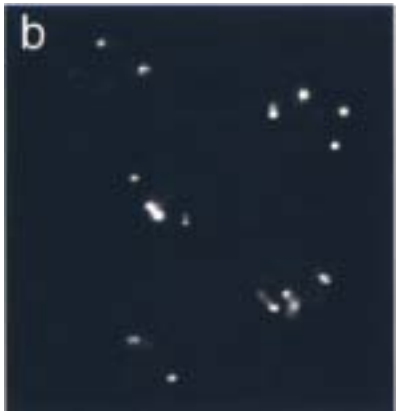
- Polarity axis: Spindle Pole Body (SPB) (MTOC) - Nuclear centre - Nucleolar centre
- Single nucleolus mostly excludes genomic DNA
- SPB embedded in nuclear envelope
- Chromosome attached at centromere, centromere attached to SPB via microtubule
- Telomeres (chromosome extremities) preferentially localise at nuclear envelope
- GAL1 gene moves to nuclear periphery during transcription (Cabal et al, *Nature*, 2006)
- Genes genomically close to telomeres when localised at nuclear periphery tend to be silenced (Hediger et al, *Current Biol*, 2002)
- and have highest DNA repair efficiency (Thérizols et al, *JCB*, 2005)



Schematic for single chromosome inside yeast nucleus

Sub-telomeric foci in yeast

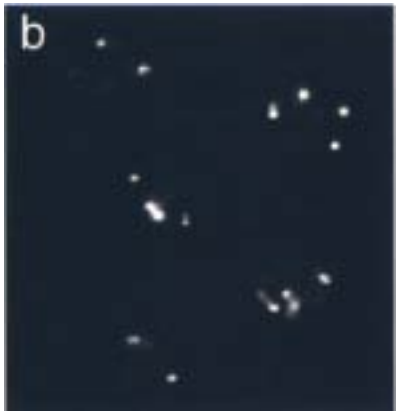
- Rap1 staining of 32 telomeres show approx 2 to 8 dots in vitro
- Spatial proximity of sub-telomeres implies sub-telomeres form foci/clusters
  - fixation shrinks cells thus reducing spatial distances
  - Rap1 binds to sites other than telomeres
  - Not all Rap1 is bound to chromosome
  - Not all Rap1 foci are detected
  - Not all telomeres are bound to Rap1



(Gotta et al, *JCB*, 1996, Fig. 7)

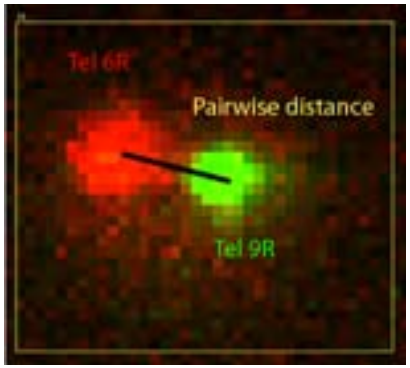
Sub-telomeric foci in yeast

- Rap1 staining of 32 telomeres show approx 2 to 8 dots in vitro
- Spatial proximity of sub-telomeres implies sub-telomeres form foci/clusters
  - fixation shrinks cells thus reducing spatial distances
  - Rap1 binds to sites other than telomeres
  - Not all Rap1 is bound to chromosome
  - Not all Rap1 foci are detected
  - Not all telomeres are bound to Rap1



(Gotta et al, *JCB*, 1996, Fig. 7)

- Ideal: in vivo analysis of 32 telomeres each simultaneously stained in a different colour
- In vivo tagging limited to 2 simultaneous colours (red, green) → pairwise distances



Re-sampling analysis for sub-telomeric foci

- Pairwise distance data for Tel6R and 20 other telomeres
- Probabilistic composition of sub-telomeric foci

Data

6R2L	6R2R	...	6R16L
0.718	1.348		1.780
1.870	1.479		1.480
1.400	1.266		0.709
0.851	1.372	...	1.490
1.220	1.852		1.520
0.274	0.620		1.460
		⋮	

Re-sampling analysis for sub-telomeric foci

- Pairwise distance data for Tel6R and 20 other telomeres
- Probabilistic composition of sub-telomeric foci

Data

6R2L	6R2R	...	6R16L
0.718	1.348		1.780
1.870	1.479		1.480
$\boxed{1.400}$	1.266		0.709
0.851	$\boxed{1.372}$	...	1.490
1.220	1.852		$\boxed{1.520}$
0.274	0.620		1.460
		⋮	

Re-sampled theoretical cell

6R2L	6R2R	...	6R16L
1.400	1.372	...	1.520

Re-sampling analysis for sub-telomeric foci

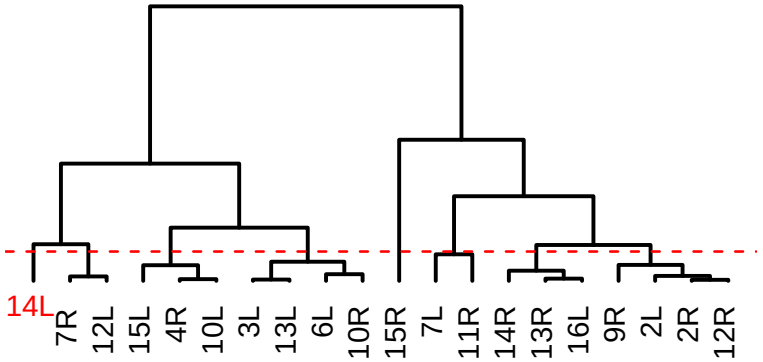
- Pairwise distance data for Tel6R and 20 other telomeres
- Probabilistic composition of sub-telomeric foci

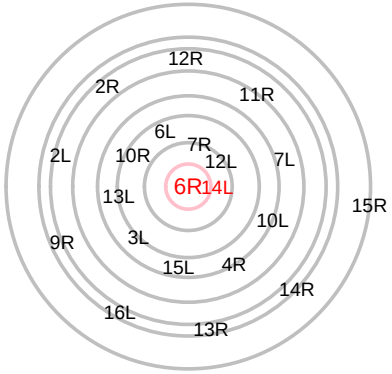
Data

6R2L	6R2R	...	6R16L
0.718	1.348		1.780
1.870	1.479		1.480
1.400	1.266		0.709
0.851	1.372	...	1.490
1.220	1.852		1.520
0.274	0.620		1.460
		⋮	

Re-sampled theoretical cell

6R2L	6R2R	...	6R16L
1.400	1.372	...	1.520





Re-sampling analysis for sub-telomeric foci

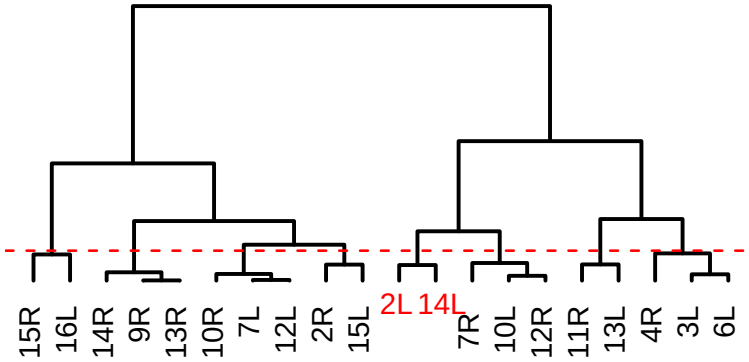
- Pairwise distance data for Tel6R and 20 other telomeres
- Probabilistic composition of sub-telomeric foci

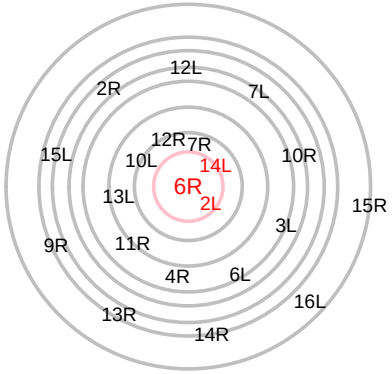
Data

6R2L	6R2R	...	6R16L
0.718	1.348		1.780
1.870	1.479		1.480
1.400	1.266		0.709
0.851	1.372	...	1.490
1.220	1.852		1.520
0.274	0.620		1.460
		⋮	

Re-sampled theoretical cell

6R2L	6R2R	...	6R16L
0.274	1.348	...	1.780





Re-sampling analysis for sub-telomeric foci

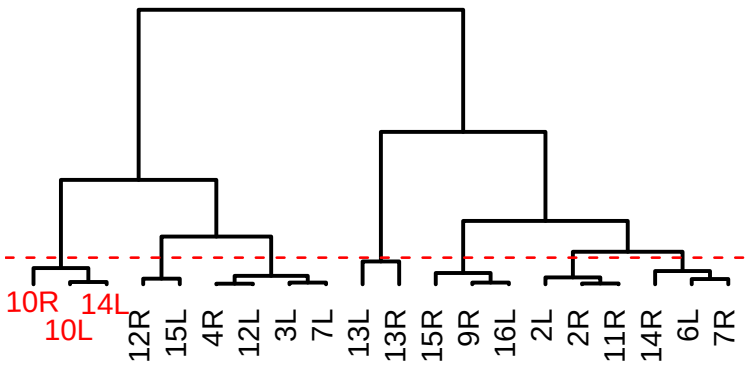
- Pairwise distance data for Tel6R and 20 other telomeres
- Probabilistic composition of sub-telomeric foci

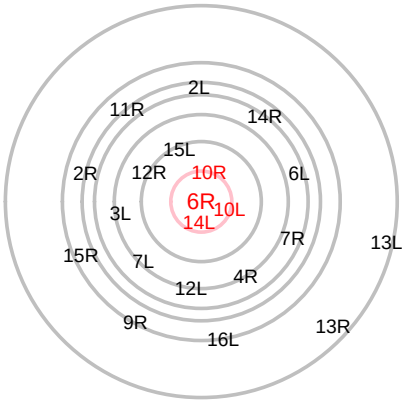
Data

6R2L	6R2R	. . .	6R16L
0.718	1.348		1.780
1.870	1.479		$\boxed{1.480}$
1.400	$\boxed{1.266}$		0.709
0.851	1.372	. . .	1.490
$\boxed{1.220}$	1.852		1.520
0.274	0.620		1.460
		⋮	

Re-sampled theoretical cell

6R2L	6R2R	. . .	6R16L
1.220	1.266	. . .	0.709

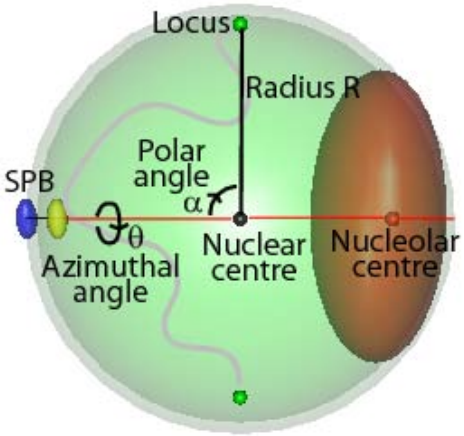




- Sub-telomeric foci are transient and dynamic (space and time)
  (Thérizols, Duong, Dujon, Zimmer & Fabre, *PNAS*, 2010)

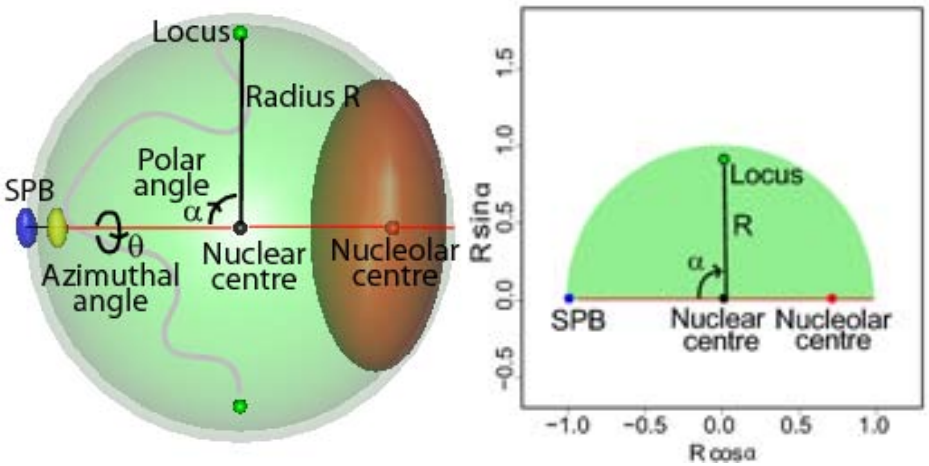
Locus density maps

- Nuclear landmarks (SPB, nuclear centre, nucleolar centre) lie on polarity axis → unable to uniquely specify 3D location of locus



Locus density maps

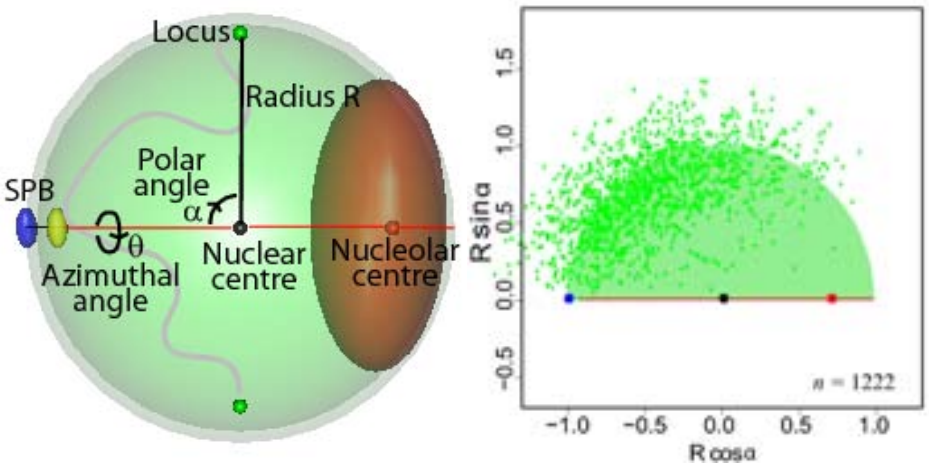
- Nuclear landmarks (SPB, nuclear centre, nucleolar centre) lie on polarity axis → unable to uniquely specify 3D location of locus



- 2D cylindrical projection: radius R and polar angle α known, but azimuthal angle (angle of rotation about polarity axis) unknown
  (Berger, Cabal, Fabre, Duong, Buc, Nehrbass, Olivo-Marin, Gadal & Zimmer, *Nature Meth*, 2008)

Locus density maps

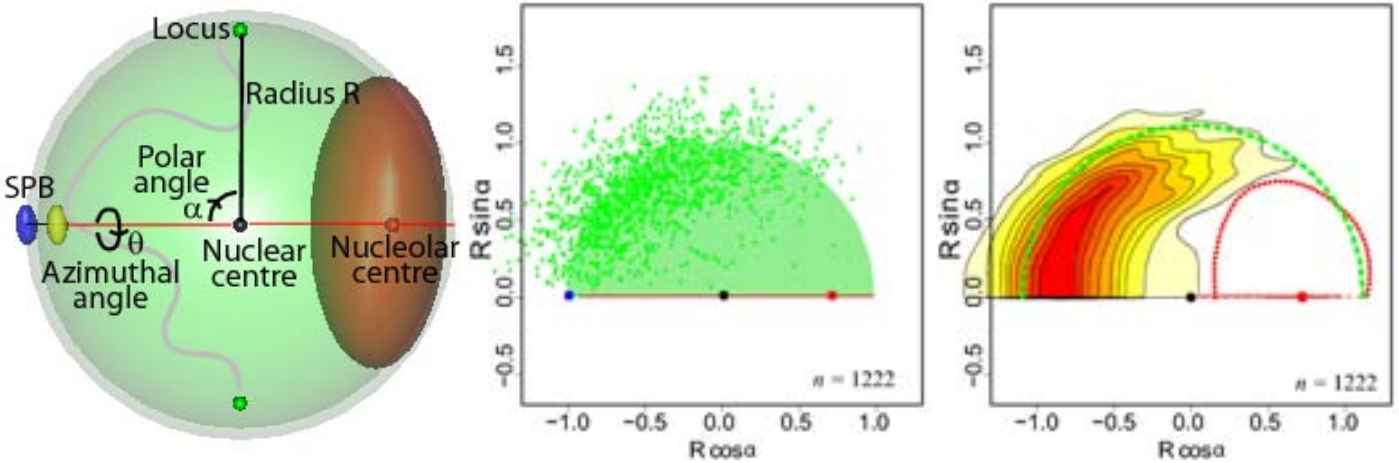
- Nuclear landmarks (SPB, nuclear centre, nucleolar centre) lie on polarity axis → unable to uniquely specify 3D location of locus



- 2D cylindrical projection: radius R and polar angle α known, but azimuthal angle (angle of rotation about polarity axis) unknown
  (Berger, Cabal, Fabre, Duong, Buc, Nehrbass, Olivo-Marin, Gadal & Zimmer, *Nature Meth*, 2008)

Locus density maps

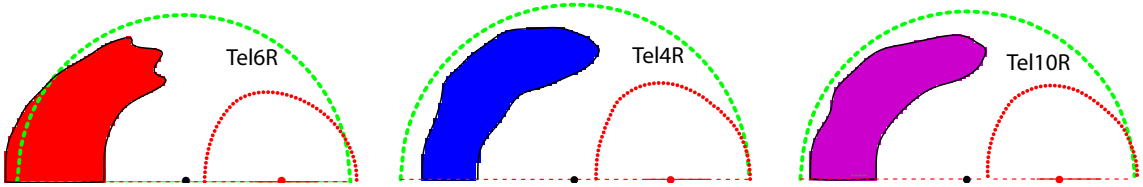
- Nuclear landmarks (SPB, nuclear centre, nucleolar centre) lie on polarity axis → unable to uniquely specify 3D location of locus



- 2D cylindrical projection: radius R and polar angle α known, but azimuthal angle (angle of rotation about polarity axis) unknown
  (Berger, Cabal, Fabre, Duong, Buc, Nehrbass, Olivo-Marin, Gadal & Zimmer, *Nature Meth*, 2008)

3D telomere reconstruction (work in progress) (1)

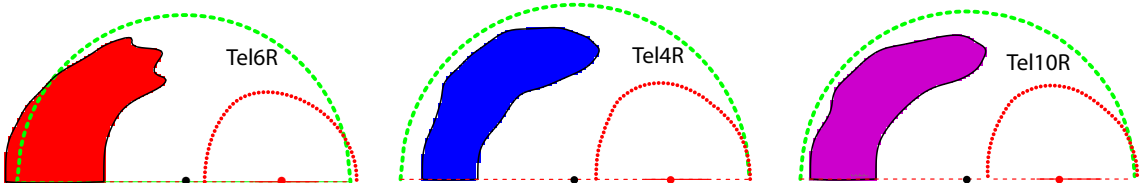
- Due to lack of identifiable rotation angle θ, overlapping locus maps do not imply co-localisation



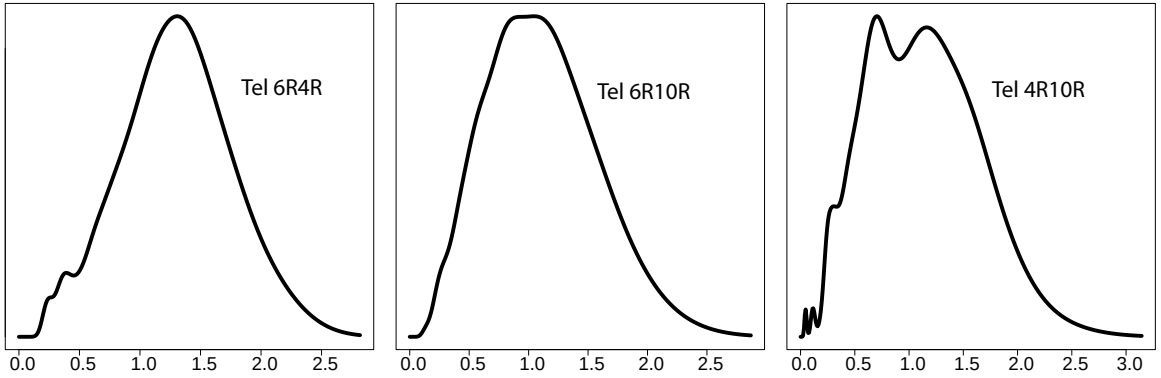
- Use pairwise distance data from telomeric foci experiments

3D telomere reconstruction (work in progress) (1)

- Due to lack of identifiable rotation angle θ, overlapping locus maps do not imply co-localisation

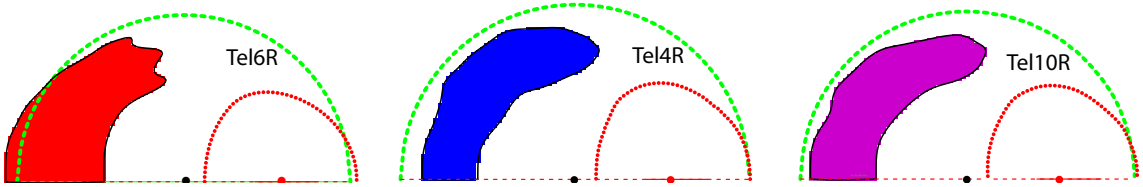


- Use pairwise distance data from telomeric foci experiments

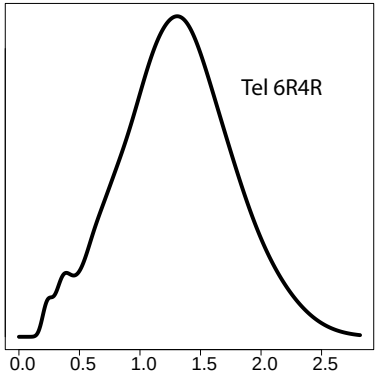


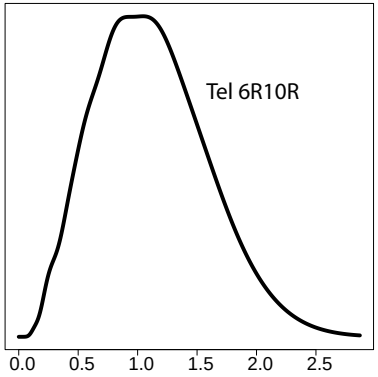
3D telomere reconstruction (work in progress) (1)

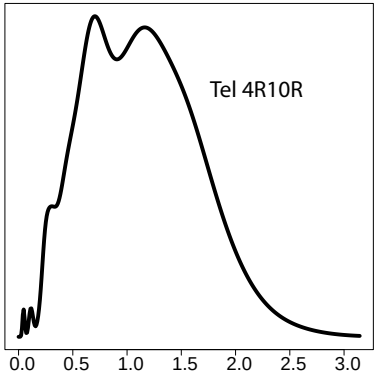
- Due to lack of identifiable rotation angle θ, overlapping locus maps do not imply co-localisation

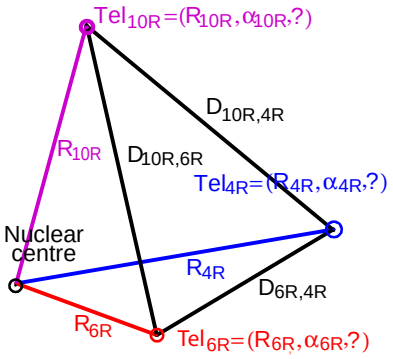


- Use pairwise distance data from telomeric foci experiments









3D telomere reconstruction (work in progress) (2)

- Solve for unkownn angles θ_{6R}, θ_{4R} and θ_{10R}

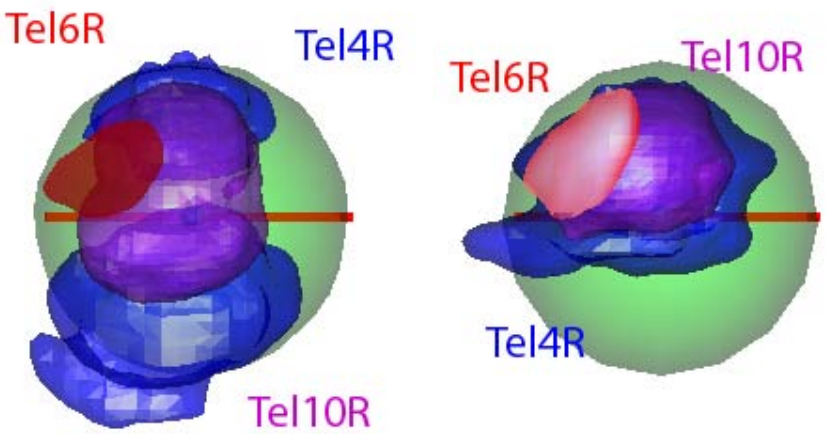
$$R_{6R}R_{4R}\sin\alpha_{6R}\sin\alpha_{4R}\cos\theta_{6R}\cos\theta_{4R} + R_{6R}R_{4R}\sin\alpha_{6R}\sin\alpha_{4R}\sin\theta_{6R}\sin\theta_{4R}$$
$$= \frac{1}{2}(R_{6R}^2 + R_{4R}^2 - D_{6R,4R}^2 - 2R_{6R}R_{4R}\sin\alpha_{6R}\sin\alpha_{4R})$$
$$R_{6R}R_{10R}\sin\alpha_{6R}\sin\alpha_{10R}\cos\theta_{6R}\cos\theta_{10R} + R_{6R}R_{10R}\sin\alpha_{6R}\sin\alpha_{10R}\sin\theta_{6R}\sin\theta_{10R}$$
$$= \frac{1}{2}(R_{6R}^2 + R_{10R}^2 - D_{6R,10R}^2 - 2R_{6R}R_{10R}\sin\alpha_{6R}\sin\alpha_{10R})$$
$$R_{4R}R_{10R}\sin\alpha_{4R}\sin\alpha_{10R}\cos\theta_{4R}\cos\theta_{10R} + R_{4R}R_{10R}\sin\alpha_{4R}\sin\alpha_{10R}\sin\theta_{4R}\sin\theta_{10R}$$
$$= \frac{1}{2}(R_{4R}^2 + R_{10R}^2 - D_{4R,10R}^2 - 2R_{4R}R_{10R}\sin\alpha_{4R}\sin\alpha_{10R})$$

3D telomere reconstruction (work in progress) (2)

- Solve for unkownn angles θ_{6R}, θ_{4R} and θ_{10R}

$$R_{6R}R_{4R}\sin\alpha_{6R}\sin\alpha_{4R}\cos\theta_{6R}\cos\theta_{4R} + R_{6R}R_{4R}\sin\alpha_{6R}\sin\alpha_{4R}\sin\theta_{6R}\sin\theta_{4R}$$
$$= \frac{1}{2}(R_{6R}^2 + R_{4R}^2 - D_{6R,4R}^2 - 2R_{6R}R_{4R}\sin\alpha_{6R}\sin\alpha_{4R})$$
$$R_{6R}R_{10R}\sin\alpha_{6R}\sin\alpha_{10R}\cos\theta_{6R}\cos\theta_{10R} + R_{6R}R_{10R}\sin\alpha_{6R}\sin\alpha_{10R}\sin\theta_{6R}\sin\theta_{10R}$$
$$= \frac{1}{2}(R_{6R}^2 + R_{10R}^2 - D_{6R,10R}^2 - 2R_{6R}R_{10R}\sin\alpha_{6R}\sin\alpha_{10R})$$
$$R_{4R}R_{10R}\sin\alpha_{4R}\sin\alpha_{10R}\cos\theta_{4R}\cos\theta_{10R} + R_{4R}R_{10R}\sin\alpha_{4R}\sin\alpha_{10R}\sin\theta_{4R}\sin\theta_{10R}$$
$$= \frac{1}{2}(R_{4R}^2 + R_{10R}^2 - D_{4R,10R}^2 - 2R_{4R}R_{10R}\sin\alpha_{4R}\sin\alpha_{10R})$$

- Result



(Duong & Zimmer, in preparation)