# Applications in Computational Biology

Stephen J Eglen
University of Cambridge

“Mathematics Is Biology’s next microscope, only better;

(J Cohen, 2004 PLOS Biol)

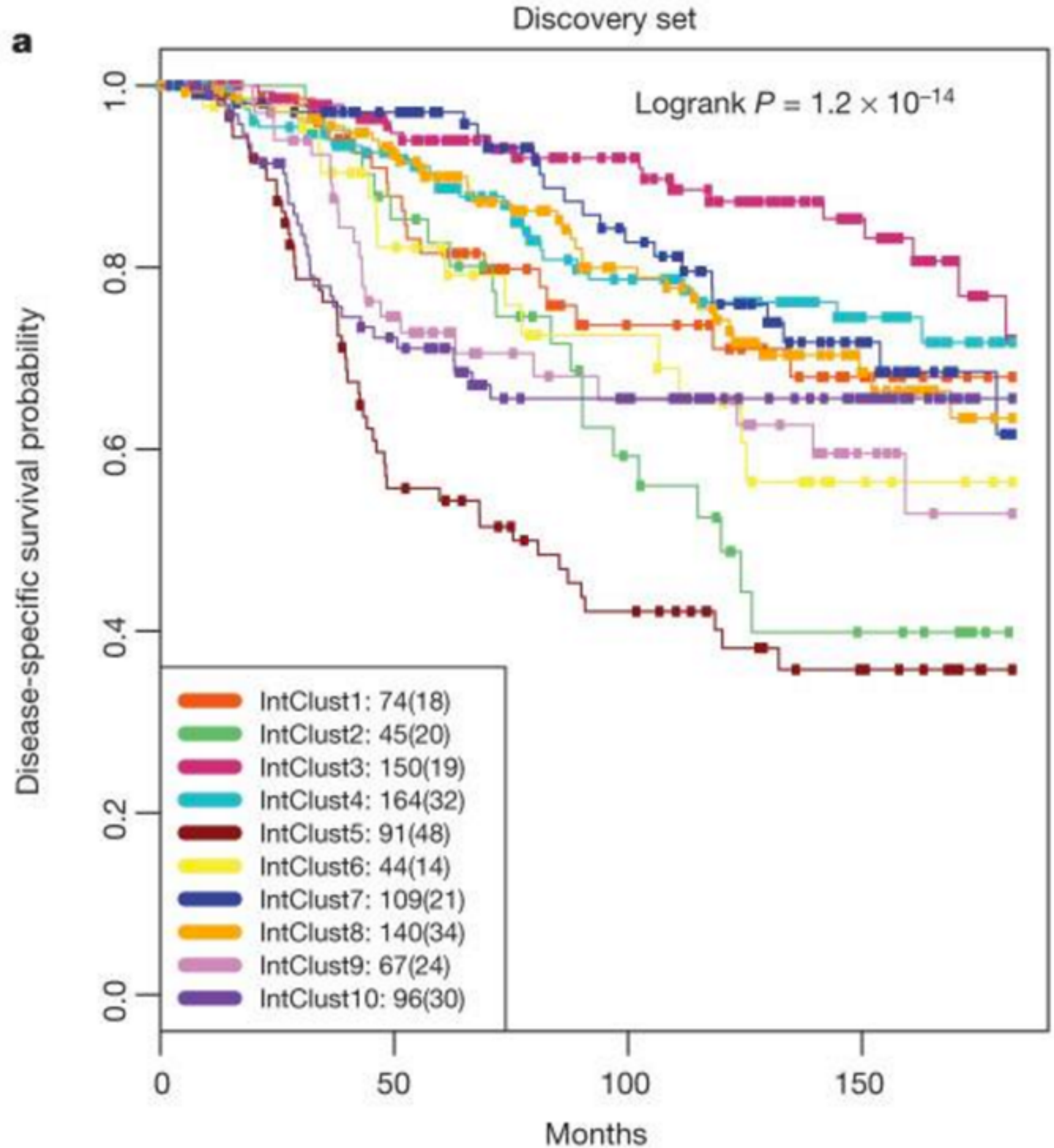
# Cambridge Computational Biology Institute (CCBI)

- Launched in 2003, with support across University.
- Wellcome Trust PhD programme in Mathematical Genomics and Medicine
- MPhil programme in Computational Biology
- Key faculty: Prof. Simon Tavaré (Cancer Research Institute), Dr Gos Micklem (Genetics), Dr Julia Gog (DAMTP)
- http://www.ccbi.cam.ac.uk

# Classification of breast cancers

Curtis C et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486:346–352.

How to move beyond morphological markers (and two key markers) used currently in clinics?

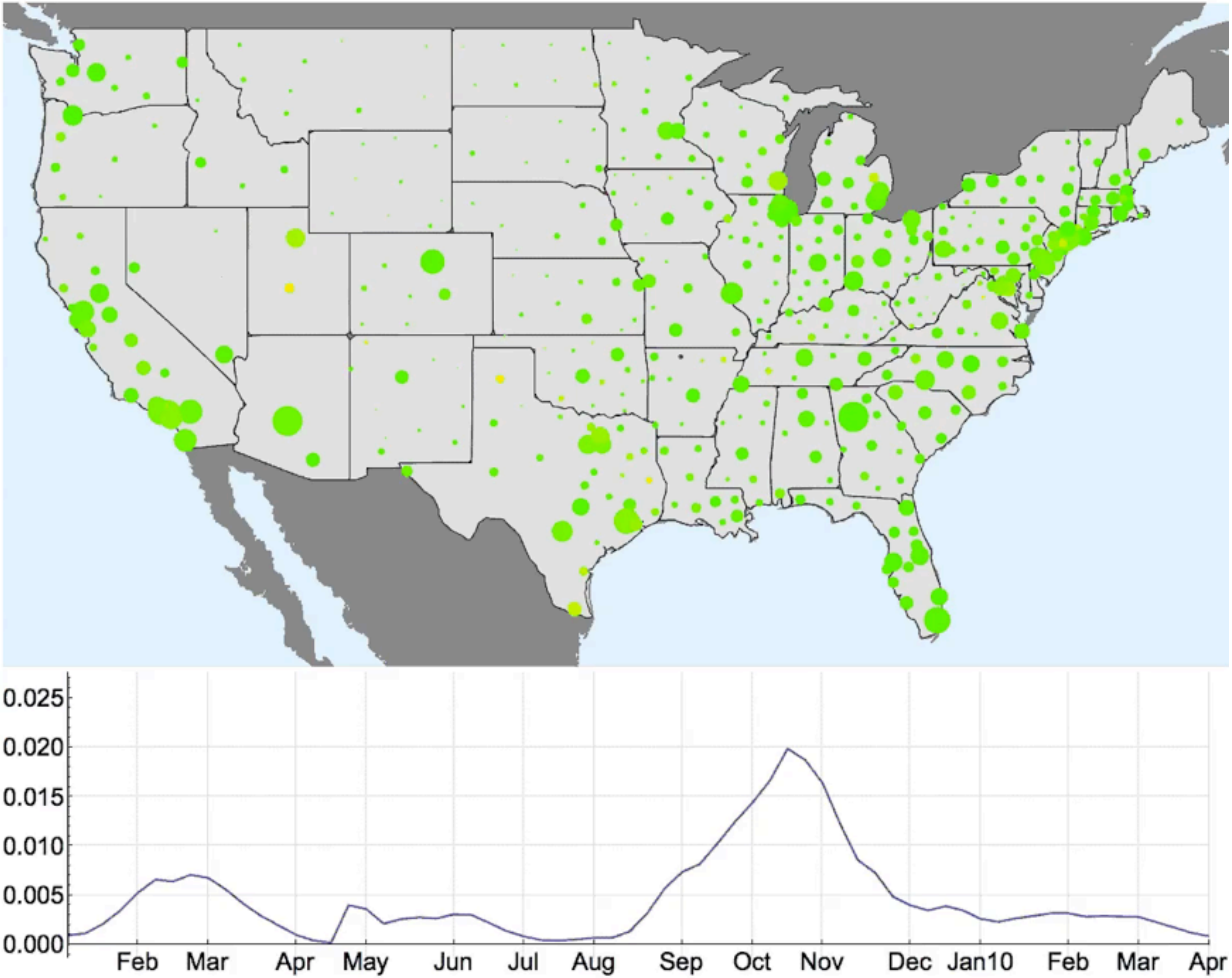


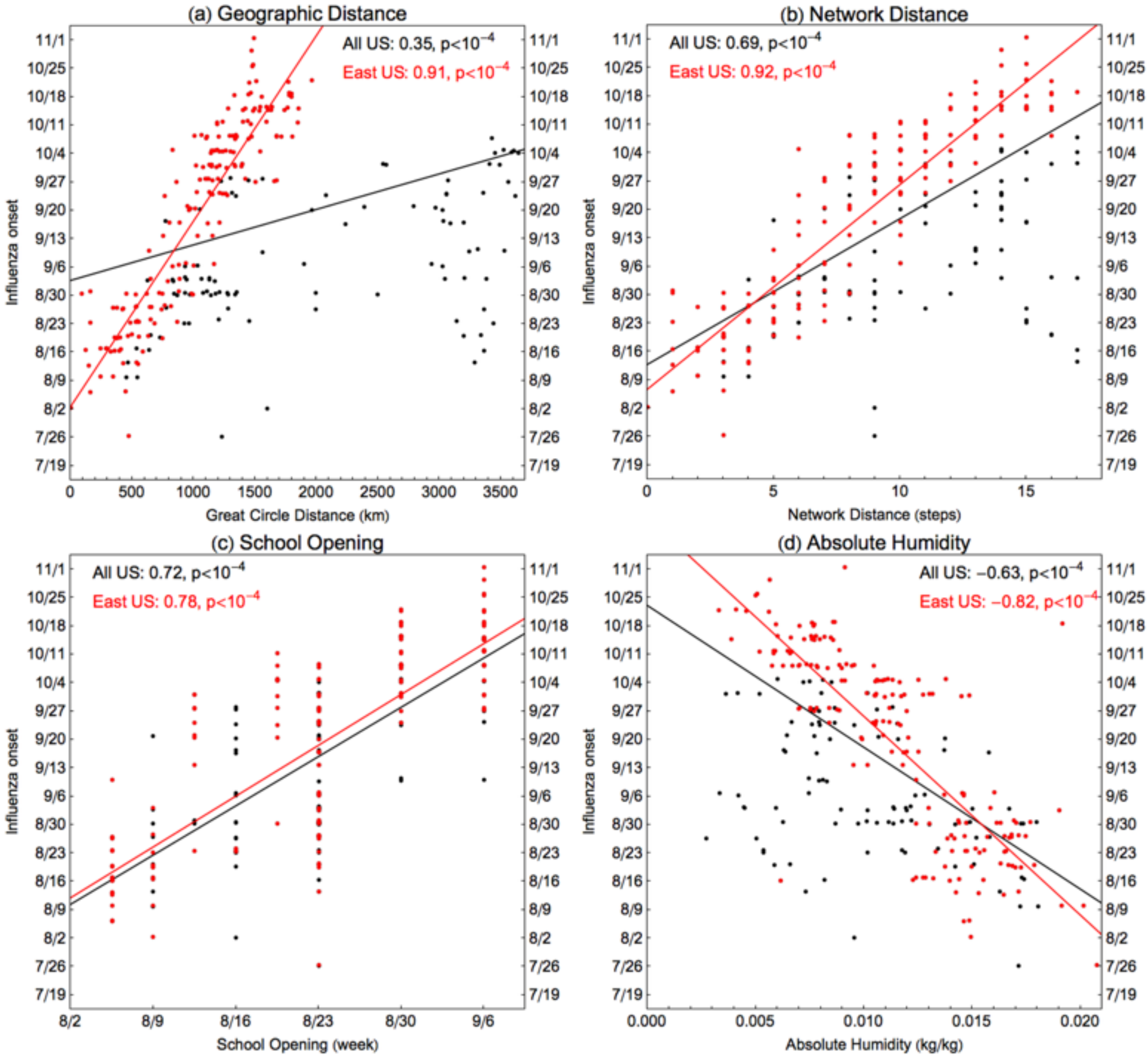
a
Discovery set
Logrank $P = 1.2 \times 10^{-14}$
Disease-specific survival probability
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
Months
IntClust1: 74(18)
IntClust2: 45(20)
IntClust3: 150(19)
IntClust4: 164(32)
IntClust5: 91(48)
IntClust6: 44(14)
IntClust7: 109(21)
IntClust8: 140(34)
IntClust9: 67(24)
IntClust10: 96(30)

# Disease dynamics: flu spread

Gog JR, Ballesteros S, Viboud C, Simonsen L, Bjornstad ON, Shaman J, Chao DL, Khan F, Grenfell BT (2014)
Spatial Transmission of 2009 Pandemic Influenza in the US.
PLoS Comput Biol 10:e1003635.

# Influenza-like illness:  4 Jan 2009




(a) Geographic Distance
All US: 0.35, p<10^−4
East US: 0.91, p<10^−4
Influenza onset
Great Circle Distance (km)
(b) Network Distance
All US: 0.69, p<10^−4
East US: 0.92, p<10^−4
Influenza onset
Network Distance (steps)
(c) School Opening
All US: 0.72, p<10^−4
East US: 0.78, p<10^−4
Influenza onset
School Opening (week)
(d) Absolute Humidity
All US: −0.63, p<10^−4
East US: −0.82, p<10^−4
Influenza onset
Absolute Humidity (kg/kg)

For each of i=1..271 cities, prob of outbreak given by:

$$P_i(T) = \left(1 - e^{-\lambda_i(T)}\right) \prod_{t=1}^{T-1} e^{-\lambda_i(t)}$$

$$\lambda_i(t) = h_i^{\phi} e^{\omega x_i} \left( \beta_0 N_i^{\alpha} + \beta_s I_i + (\beta_d + \beta_{ds} I_i) N_i^{\mu} \frac{\sum_{j \in \Lambda} N_j^{\nu} d_{i,j}^{-\gamma}}{\left[ \sum_{j \neq i} N_j^{\nu} d_{i,j}^{-\gamma} \right]^{\varepsilon}} \right)$$

# Most parsimonious model

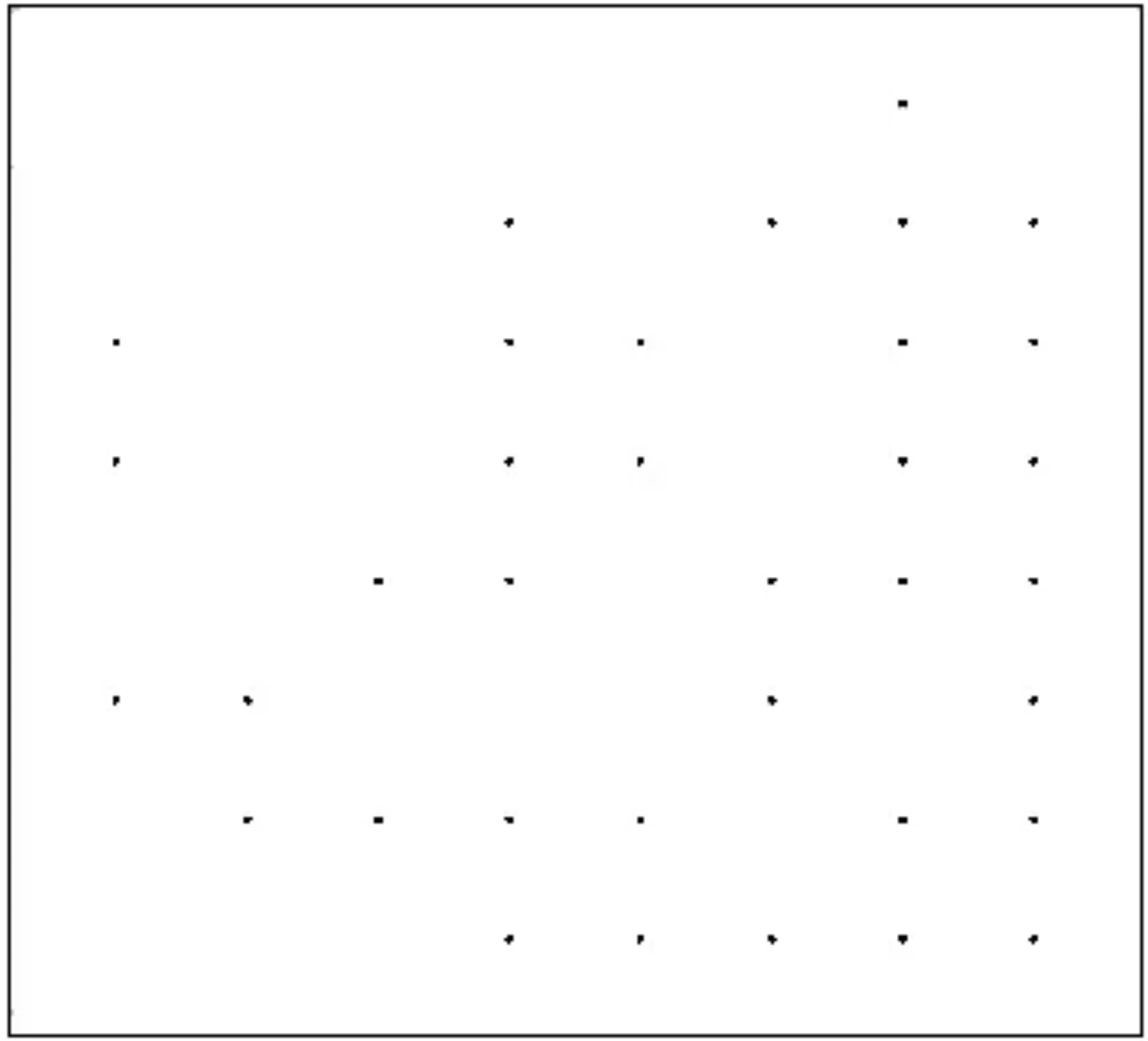
| Parameter | Description | Units | Value | 95% Confidence interval |
|---|---|---|---|---|
| $\beta_0$ | Background transmission rate | $(\Delta t)^{-1}$ | 0.0013 | 0.0004-0.0028 |
| $\beta_d$ | Spatial transmission coefficient | $(\Delta t)^{-1}(km)^{1-\varepsilon}$ | 0.84 | 0.26-2.3 |
| $\beta_{ds}$ | Boost to $\beta_d$ when schools are open | $(\Delta t)^{-1}(km)^{1-\varepsilon}$ | 3.0 | 1.4-6.3 |
| $\mu$ | Exponent of dependence on recipient population size | none | 0.27 | 0.11-0.44 |
| $\gamma$ | Exponent of distance in gravity model kernel | none | 2.6 | 2.3-2.8 |
| $\varepsilon$ | Strength of density normalisation | none | 0.87 | 0.80-0.94 |

$$\lambda_i(t) = \beta_0 + (\beta_d + \beta_{ds} I_i) N_i^{\mu} \frac{\sum_{j \in \Lambda} d_{i,j}^{-\gamma}}{\left[\sum_{j \neq i} d_{i,j}^{-\gamma}\right]^{\varepsilon}}$$

# Detecting correlations

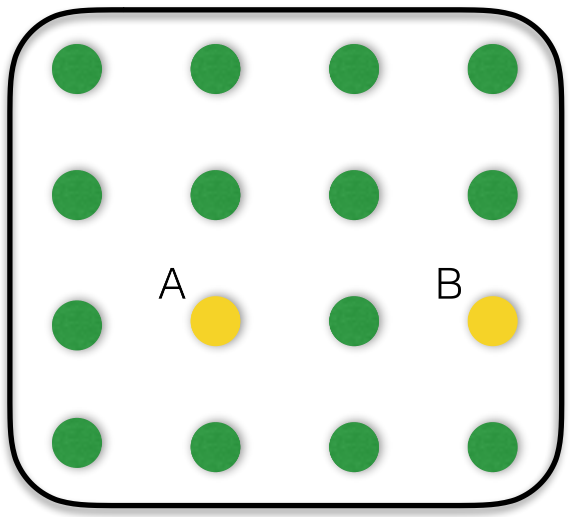
Catherine Cutts (J Neurosci 2014)

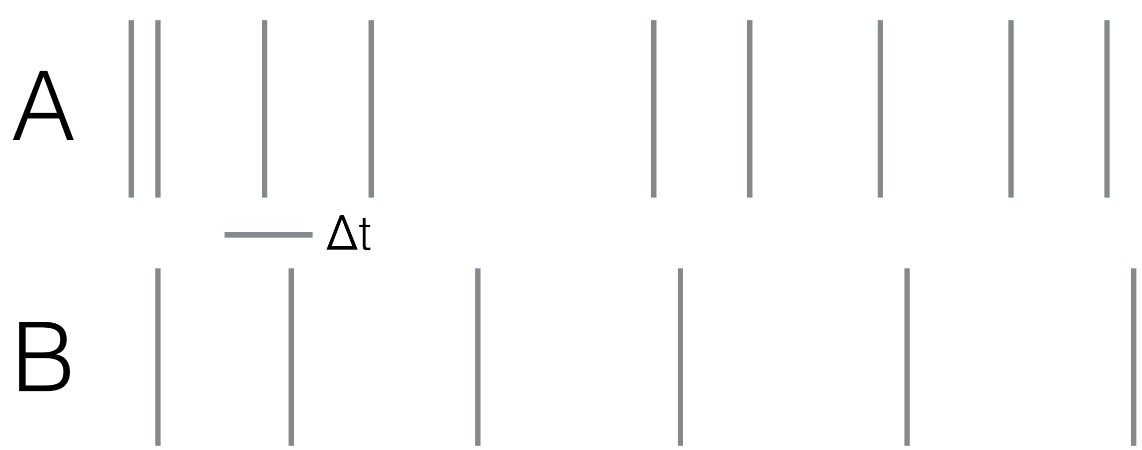
400.0



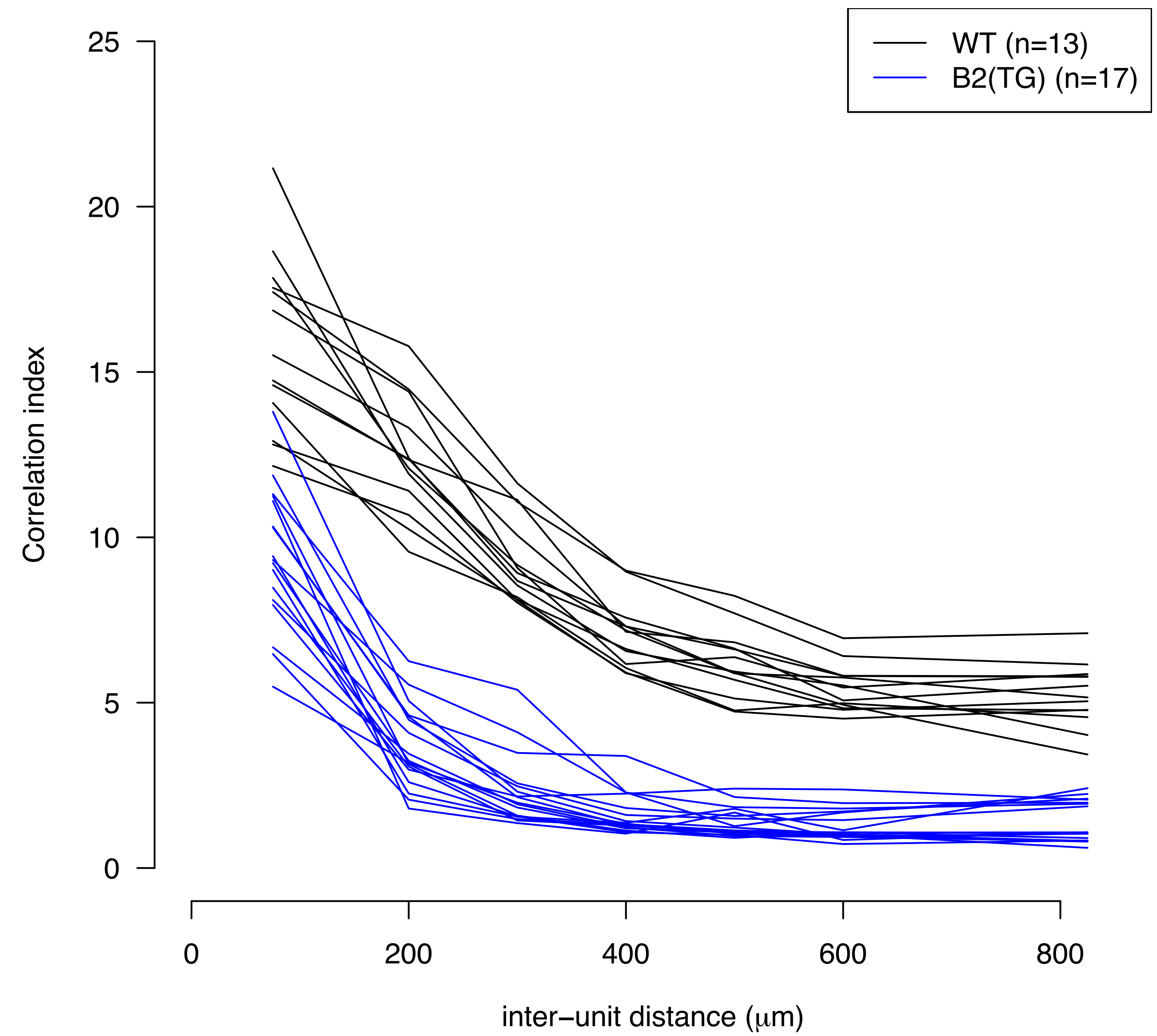
(Demas et al. 2003, P11 mouse)

# Correlation index




$$c = \frac{N_{A,B[-\Delta t,+\Delta t]}T}{N_A N_B 2\Delta t}$$




(Wong et al. 1993; Xu et al. 2011)

# Measuring correlation

**Distance measures and cost functions**

1 Victor and Purpura (1997)
2 ISI-distance (Kreuz et al., 2007a)
3 Hunter-Milton similarity (Hunter and Milton, 2003)
4 Van Rossum (2001)
5 SPIKE (Kreuz et al., 2013)

**Cross-correlation based**

6 Coincidence index (Pasquale et al., 2008)
7 Altered Coincidence index*
8 Cross correlation coefficient (Pasquale et al., 2008)
9 Schreiber et al. (2003) similarity coefficient
10 Altered Schreiber et al. similarity coefficient*
11 Kerschensteiner and Wong (2008) cross-correlation
12 Jimbo and Robinson index (Jimbo et al., 1999)

**Synchrony not from cross-correlation**

13 Correlation index (Wong et al., 1993)
14 Activity pair (Eytan et al., 2004)
15 Unitary events analysis (Grün et al., 2002)
16 Event synchronization (Kreuz et al., 2007b)*
17 Joris et al. (2006) correlation index

**Information theory**

18 Mutual information (Li, 1990)
19 Mutual information with smoothing*

**Measures from shot-noise process**

20 Coherence (at zero) (Eggermont, 2010)
21 Spike count correlation (Eggermont, 2010)
22 Smoothed spike count correlation (Kruskal et al., 2007)*
23 Spike count covariance (Eggermont, 2010)

**Measures assuming a marked point process**

24 Stoyan's $K_{mm}$ function (Stoyan and Stoyan, 1994)
25 Isham's mark correlation function (Isham, 1985)
26 Ripley's $K_{mm}$ function (Ripley, 1976)
27 Simpson (1949) index
28 Simpson (1949) index no correction
29 Stoyan's mark covariance function (Stoyan, 1984)
30 Mark variogram (Cressie, 1993)
31 Mark covariance function (Cressie, 1993)
32 Mark conditional expectation ($E$; Schlather et al., 2004)
33 Mark conditional variance ($V$; Schlather et al., 2004)
34 Mark conditional standard deviation (Schlather et al., 2004)

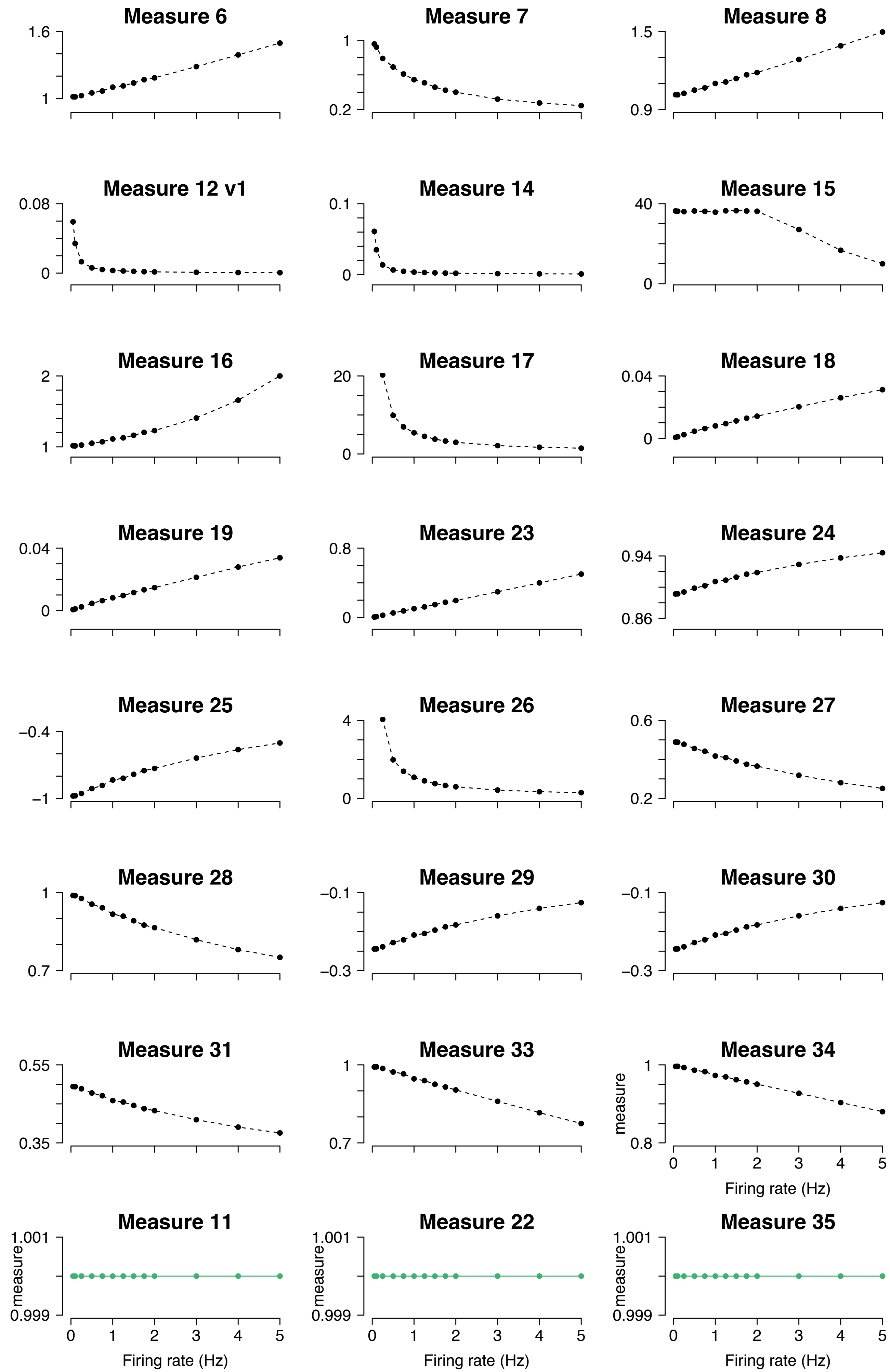
# Which method is best?

- 34 measures in literature + 1 from us => 35.
- Phase 1: six necessary properties:

1. Symmetric
2. Robust to variations in firing rate
3. Robust to amount of data
4. Bounded [-1, +1]
5. Robust to variations in bin width ($\Delta t$)

Twenty-one measures rejected as they depend on firing rate.

(autocorrelation of Poisson trains)



# Short-list from Phase 1

$$r(\mathbf{A},\mathbf{B}) = \frac{\sum_{i=1}^{N}(\mathbf{A}_i - \overline{\mathbf{A}})(\mathbf{B}_i - \overline{\mathbf{B}})}{\sqrt{\sum_{i=1}^{N}(\mathbf{A}_i - \overline{\mathbf{A}})^2}\sqrt{\sum_{i=1}^{N}(\mathbf{B}_i - \overline{\mathbf{B}})^2}},$$

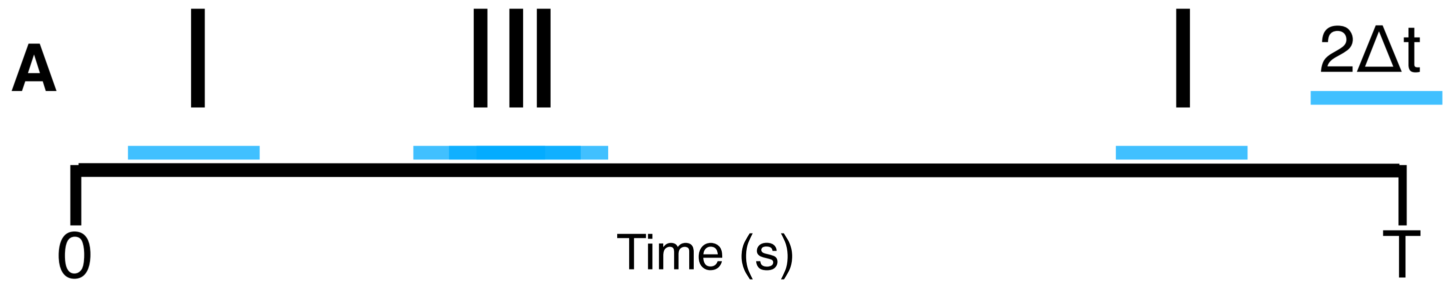
1. Spike count correlation coefficient

2. Kerschensteiner and Wong (2008)

3. Kruskal et al. (2007)

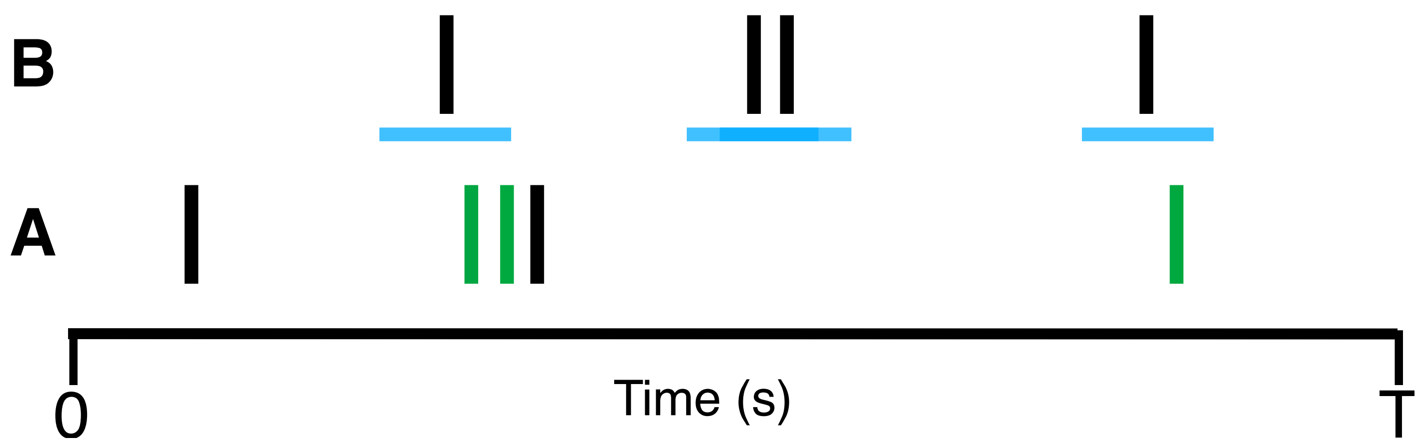
4. Spike time tiling coefficient

**$T_A$:** the proportion of total recording time which lies within $\pm\Delta t$ of any spike from A. **$T_B$** calculated similarly.




$T_A$ is given by the fraction of the total recording time (black) which is covered (tiled) by blue bars. Here $T_A$ is 1/3.

**$P_A$:** the proportion of spikes from A which lie within $\pm\Delta t$ of any spike from B. **$P_B$** calculated similarly.




$P_A$ is the number of green spikes in A (3) divided by the total number of spikes in A (5). Here $P_A$ is 3/5.

$$TC = \frac{1}{2}\left(\frac{P_A - T_B}{1 - P_A T_B} + \frac{P_B - T_A}{1 - P_B T_A}\right)$$

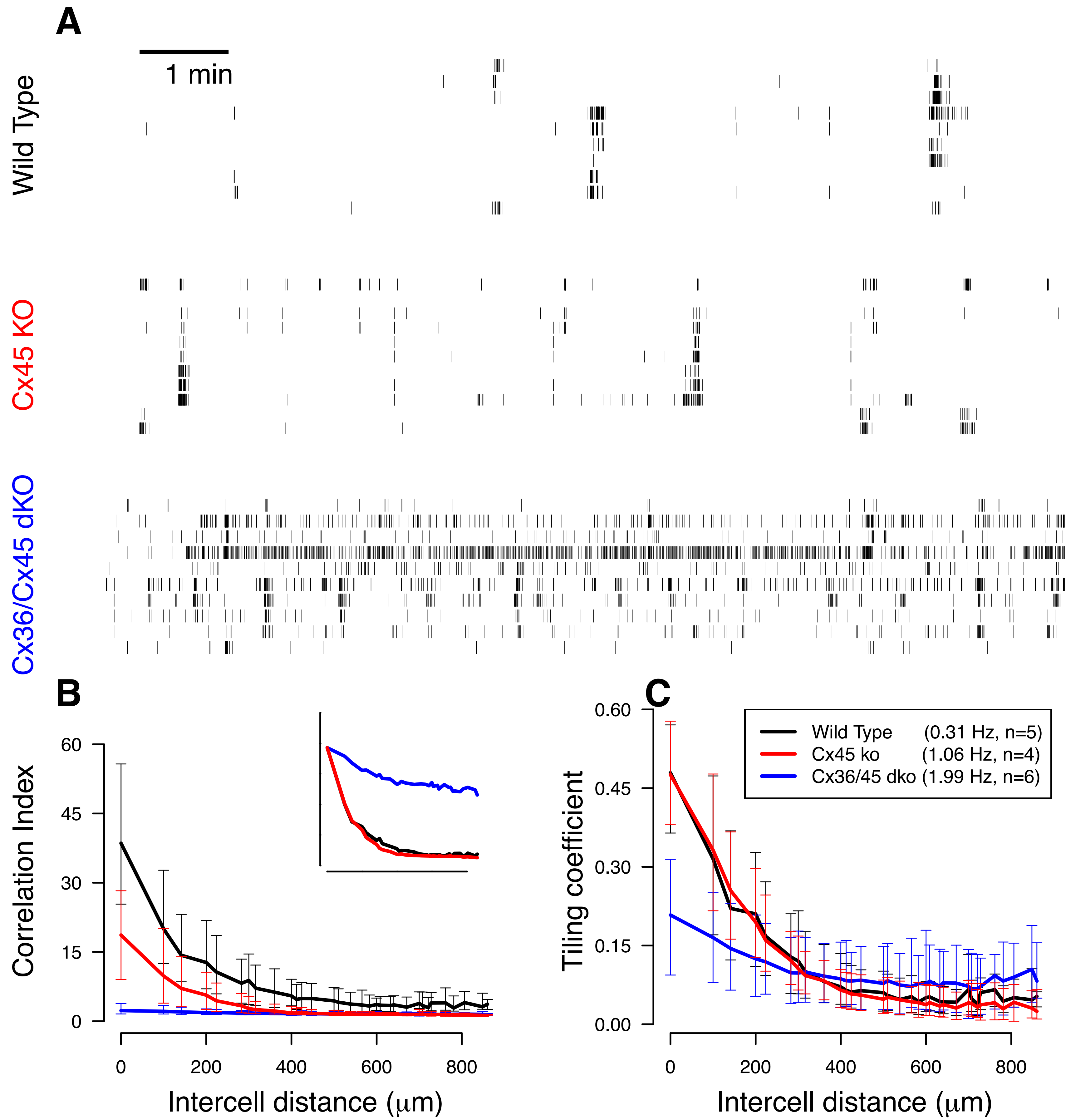
# Phase 2: Desirable properties

## Desirable properties:

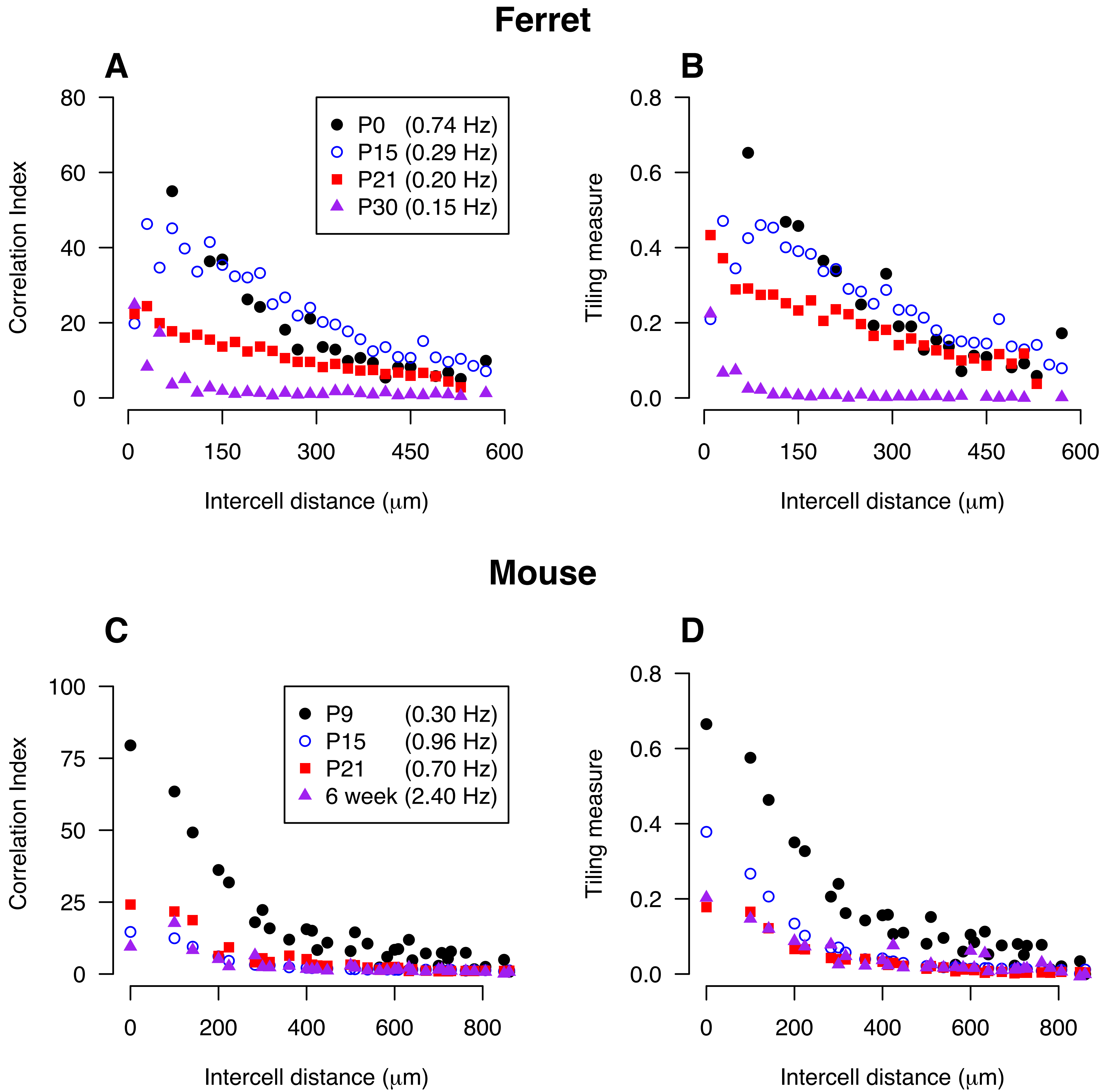
- D1: Ignore periods when both neurons are inactive.
- D2: minimal assumptions on structure.
- D3: aside from Δt, minimise number of parameters

## Four methods:

- Kerschensteiner and Wong correlation (D1, D2)
- Tiling coefficient (D1, D2, D3)
- Spike count correlation (D2, D3)
- Kruskal et al. binless correlation measure (D2, D3)




(Blankenship et al. 2011)



Ferret
A
80
60
40
20
0
Correlation Index
P0 (0.74 Hz)
P15 (0.29 Hz)
P21 (0.20 Hz)
P30 (0.15 Hz)
0
150
300
450
600
Intercell distance (μm)
B
0.8
0.6
0.4
0.2
0.0
Tiling measure
Intercell distance (μm)
Mouse
C
100
75
50
25
Correlation Index
P9 (0.30 Hz)
P15 (0.96 Hz)
P21 (0.70 Hz)
6 week (2.40 Hz)
200
400
600
800
Intercell distance (μm)
D
Tiling measure
Intercell distance (μm)

# Summary

- Correlation measure already applied in recent papers.  Code available, easy to use.

- Too specific to retinal waves (Eisenman et al, J Neurophys 2015)?

- Bayesian hierarchical modelling framework for comparing correlation curves on Biorxiv…

# Acknowledgments

**Funding**: EPSRC, Wellcome Trust, USIAS, BBSRC

**Data/code**: http://damtp.cam.ac.uk/user/eglen

See also posters by:

Stephen Kissler (disease dynamics)

Ellese Cotterill (burst analysis)

Linda Bowns (motion detection).