Why don’t we agree? Studying influenza with RNA interference

Michael Newton
UW Madison

Buehler-Martin Lecture
University of Minnesota Statistics
From a project on influenza biology with:

Qiuling He
Lin Hao
Mark Craven
Paul Ahlquist

Thanks Christina Kendziorski...
What human genes does influenza virus co-opt during its life cycle?
a bit about flu
Experimental Method: RNAi

RNA interference is a process within living cells that moderates the activity of their genes.

Fire and Mello, 2006, Nobel Prize
a bit about RNAi
RNA interference =
Or, more simply, ...

RNA interference =

genome-wide...
cell

expression level

target gene
add siRNA

+
add siRNA

+ cell

expression level

target gene
add siRNA

+ 

expression level

target gene

cell
add siRNA

+○

cell

expression level

target gene
add siRNA

+  

expression level

target gene

cell
add siRNA

+<

expression level

target gene
cell

expression level

target gene
cell

expression level

target gene

phenotype of interest changes
explain phenotype
Issues
Involvement: gene may not affect phenotype
Efficiency: knockdown may not be complete
Accessibility: something blocks the phenotype

a. no expression in these particular cells
**Accessibility:** something blocks the phenotype

**b. Redundency/masking**
Accessibility: something blocks the phenotype

cytotoxicity
Off target effects

Expression level

Off target g2 involved

Target g1 not involved
Measurement error
Measurement error

expression level

non-involved gene

false positive
Measurement error
Measurement error

expression level

involved gene

false negative
one siRNA

+ knock down

no knock down
one siRNA

on target

off target

knock down

on target

off target

no knock down
one siRNA

[Diagram]

- On target
- Off target

Knock down

No knock down

Measurement

Error

Error

Error

Error
Meta analysis of four recent studies

**Drosophila RNAi screen identifies host genes important for influenza virus replication**
Linhui Hao1,2,4, Akira Sakurai2,4,6, Tokiko Watanabe3, Ericka Sorensen1, Chairul A. Nidom5,6, Michael A. Newton4, Paul Ahlquist1,2 & Yoshihiro Kawaoka1,5,8,9

**The IFITM Proteins Mediate Cellular Resistance to Influenza A H1N1 Virus, West Nile Virus, and Dengue Virus**
Abraham L. Bross1,2,4,6, I-Chueh Huang3,6, Yair Benita3,6, Sinu P. John1,10, Manoj N. Krishnan,6, Eric M. Feasley,1 Bethany J. Ryan,1 Jessica L. Weyer,1 Louise van der Wyden,2 Erol Fikrig,6,7 David J. Adams,4 Ramnik J. Xavier6,7 Michael Farzan,2,5 and Stephen J. Elledge1,6

**Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication**
Alexander Karlas1,*, Nikolaus Machuy1,*, Yujin Shin1, Klaus-Peter Pleissner2, Anita Artarini1, Dagmar Heuer1, Daniel Becker1, Hany Khalili1, Lesley A. Ogilvie1, Simone Hess†, André P. Mäurer1, Elke Müller1†, Thorsten Wolff1, Thomas Rudel† & Thomas F. Meyer1

**Human host factors required for influenza virus replication**
Renate König1,*, Silke Stertz1,*, Yingyao Zhou1, Atsushi Inoue1, H.-Heinrich Hoffmann5, Suchita Bhattacharyya7, Judith G. Alamares1, Donna M. Tscherné1, Mia B. Ortigoza1, Yuhong Liang1, Qinshan Gao1, Shane E. Andrews1, Sourav Bandyopadhyay6, Paul De Jesus1, Buu P. Tu1, Lars Pache1, Crystal Shih1, Anthony Orth7, Ghislain Bonamy7, Loren Miraglia1, Troy Ideker1, Adolfo Garcia-Sastre1,5,6, John A. T. Young7, Peter Palese1,2, Megan L. Shaw1,4, & Sumit K. Chanda1,*
Data

results (gene lists) from 4 two-stage RNAi studies
Data

results (gene lists) from 4 two-stage RNAi studies

1. detection
2. confirmation
Data results (gene lists) from 4 two-stage RNAi studies

1. detection
2. confirmation

e.g., one gene

<table>
<thead>
<tr>
<th></th>
<th>DL-1</th>
<th>U2OS</th>
<th>A549_{DE}</th>
<th>A549_{US}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection Screen</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Confirmation Screen</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pattern Code</td>
<td>11</td>
<td>10</td>
<td>00</td>
<td>00</td>
</tr>
</tbody>
</table>
### Detection and Confirmation Patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>DL-1</th>
<th>U2OS</th>
<th>A549&lt;sub&gt;DE&lt;/sub&gt;</th>
<th>A549&lt;sub&gt;US&lt;/sub&gt;</th>
<th># genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21,016</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>127</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>79</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>80</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>81</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G = 22000</td>
</tr>
</tbody>
</table>
Agreement among studies is low.

Among the 614 genes confirmed by at least one study:
Is the limited overlap due more to false positive or false negative factors?
6.7% average pairwise overlap of confirmed gene lists is significantly higher than expected by chance
Modeling approach

\[ P_\pi = \text{Prob (gene shows detection/confirmation pattern } \pi \text{ )} \]

\[ n_\pi = \# \text{ (gene shows detection/confirmation pattern } \pi \text{ )} \]

Likelihood

\[ L = \prod_\pi P_{\pi}^{n_\pi} \]

<table>
<thead>
<tr>
<th>( \pi )</th>
<th>DL-1</th>
<th>U2OS</th>
<th>A549_{DE}</th>
<th>A549_{US}</th>
<th>( n_\pi )</th>
<th>( P_\pi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21,016</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>127</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>79</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>80</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>81</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G = 22000</td>
</tr>
</tbody>
</table>
Data and latent variables

\[ D_{g,s} = 1 \text{ [gene } g \text{ detected in study } s \text{]} \]

\[ C_{g,s} = 1 \text{ [gene } g \text{ confirmed in study } s \text{]} \]

\[ I_g = 1 \text{ [ } g \text{ involved in flu} \text{]} \]

\[ A_{g,s} = 1 \text{ [ } g \text{ accessible in } s \text{]} \]

\[ T_{g,s} = \#\{\text{involved, accessible off targets, study } s, \text{ target } g\} \]
Detection screen

pool of 4 distinct siRNA’s per target gene

one phenotype call $D_{g,s}$
Confirmation screen

\[ C_{g,s,1} + C_{g,s,2} + C_{g,s,3} + C_{g,s,4} \]

\[ C_{g,s} = 1 \left( \sum_{k} C_{g,s,k} \geq 2 \right) \]
Figure 2. Plate diagram for statistical model
on target knock down model

\[ U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1) \]
on target knock down model

\[ U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1) \]
on target knock down model

\[ U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1) \]
on target knock down model

\[ U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1) \]
on target knock down model

\[ U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1) \]
on target knock down model

\[ U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1) \]
on target knock down model

\[ U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1) \]
on target knock down model

\[ U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1) \]
on target knock down model

\[ U_1, U_2, U_3, U_4 \sim_{iid} \text{Uniform}(0, 1) \]

knock down effect if

\[ \prod_{j} U_j < \omega \]

threshold parameter

\[
\text{Prob: } 1 - G_4 [\log(1/\omega)]
\]
off target knock down model

t = \#\{involved, accessible off targets\}

\(U_1, U_2, \ldots, U_t \sim \text{i.i.d. Uniform}(0, 1)\)

e.g. \(t = 3\)

\[\begin{align*}
U_1 &< \omega \\
U_2 &< \omega \\
U_3 &< \omega
\end{align*}\]

knock down effect if any \(U_j < \omega\)

Prob: \(1 - (1 - \omega)^t\)
\[ 1 - G_4 \left[ \log(1/\omega) \right] \]

\[ 1 - (1 - \omega)^4 \]
detection screen: \( D_{g,s} \mid A_{g,s} = a, I_g = i, T_{g,s} = t \)

\[
\begin{align*}
1 - \{ G_4 [\log(1/\omega)] \}^{\alpha i} & \quad \text{knock down} \\
1 - (1 - \omega)^t & \quad \text{measurement} \\
\{ G_4 [\log(1/\omega)] \}^{\alpha i} & \quad \text{no knock down}
\end{align*}
\]
confirmation screen

\[
C_{g,s} = 1 \left[ \sum_k C_{g,s,k} \geq 2 \right]
\]
confirmation screen: 

\[ C_{g,s,k} \mid A_{g,s} = a, I_g = i, T_{g,s} = t \]

\[
\begin{align*}
1 - (1 - \omega)^{ai} & \quad \text{knock down} \\
1 - (1 - \frac{\omega}{4})^t & \quad \text{measurement} \\
1 - \omega & \quad \text{no knock down} \\
\end{align*}
\]
On the number of off targets per siRNA

- very limited data
- libraries overlap among 4 studies
one target gene

K siRNA’s available to all studies (on average)

siRNA’s

involved

off targets

$T_g$
one target gene

study s uses 4 siRNA’s

siRNA’s

involved, accessible off targets

\[ T_{g,s} \]
\( T_g = \#\{\text{involved off-targets, target } g \} \)

\( T_g \sim \text{Poisson}(K\theta \nu) \)

\( T_{g,s} | T_g = t \sim \text{Binomial} \left( t, \frac{4\gamma_s}{K} \right) \)  

\( T_{g,s,k} | T_{g,s} = t \sim \text{Binomial} \left( t, \frac{1}{4} \right) \)
Parameters

\( \theta \) = proportion of genome involved in influenza virus replication
\( \alpha \) = false positive measurement error
\( \beta_s \) = false negative measurement error, study \( s \)
\( \gamma_s \) = rate at which genes are accessible, study \( s \)
\( \omega \) = knockdown efficiency per siRNA
\( \nu \) = average number of off-targets per siRNA.
The log-likelihood is

$$L = \log \text{Prob}_{\text{data}} = \sum_{v} \log P^{v}$$

where pattern probabilities \(\{P^{v}\}\) are defined by a smaller number of parameters through a stochastic model of genome-wide RNAi. The model is hierarchical and is specified using latent random effects:

$$I_g \sim \text{Bernoulli}(\theta)$$

$$A_{g,s} \sim \text{Bernoulli}(\gamma_s)$$

$$T_g \sim \text{Poisson}(K\theta\nu)$$

$$T_{g,s} | [T_g = t] \sim \text{Binomial}(t, 4\gamma_s / K)$$

$$D_{g,s} | [I_g = i, A_{g,s} = a, T_{g,s} = t] \sim \text{Bernoulli} \left[ 1 - \beta_s + (\alpha + \beta_s - 1) \left[ G_{4,1}(-\log \omega) \right]^{ai} \left(1 - \omega\right)^{t} \right]$$

$$C'_{g,s,k} | [I_g = i, A_{g,s} = a, T_{g,s} = t] \sim \text{Bernoulli} \left[ 1 - \beta_s + (\alpha + \beta_s - 1)(1 - \omega)^{ai} \left(1 - \omega / 4\right)^{t} \right]$$
Calculating pattern probabilities:  \[ P_\pi = \sum_{i,a} P_\pi(i, a) \]

\[ P_\pi(i, a) = P(\pi|I_g = i, \{A_{g,s}\}_{s=1}^4 = a). \]

\[
\begin{align*}
P_\pi(i, a) &= \sum_{t=0}^{\infty} P(T_g = t) \ P(\pi|I_g = i, \{A_{g,s}\}_{s=1}^4 = a, T_g = t) \\
&= \sum_{t=0}^{\infty} \text{Po}(t) \ \prod_{s=1}^{4} P(\pi_s|I_g = i, A_{g,s} = a_s, T_g = t) \quad \pi = \bigcap_s \pi_s. \\
&= \sum_{t=0}^{\infty} \text{Po}(t) \ \prod_{s=1}^{4} \sum_{u=0}^{t} B_s(t, u) \ P(\pi_s|I_g = i, A_{g,s} = a_s, T_{g,s} = u) \\
&= \sum_{t=0}^{\infty} \text{Po}(t) \ \prod_{s=1}^{4} \sum_{u=0}^{t} B_s(t, u) \ Q_{s,i,a_s,u}
\end{align*}
\]
A key to simplifying the computation further is to recognize that with respect to the count variable
sub-pattern probabilities are:

\[ Q_{s,i,a}^{'} = P \left( \pi_s | I_g = i, A_{g,s} = a_s, T_{g,s} = u \right) \]

**Lemma:**

\[
Q_{s,i,a}^{'} = \sum_{p=0}^{1} \sum_{q=0}^{4} b_{s,p,q} \left( \xi_1^p \xi_2^q \right)^u
\]

\[ \xi_1 = 1 - \omega \]
\[ \xi_2 = 1 - \frac{\omega}{4} \]
\[
\sum_{u=0}^{t} B_s(t, u) Q_{s,i,a_s,u} = \sum_{p=0}^{1} \sum_{q=0}^{4} b_{s,p,q} \sum_{u=0}^{t} (\xi_1^p \xi_2^q)^u B_s(t, u)
\]
\[
= \sum_{p=0}^{1} \sum_{q=0}^{4} b_{s,p,q} \left( 1 - \frac{4\gamma_s}{K} + \frac{4\gamma_s}{K} (\xi_1^p \xi_2^q) \right)^t
\]
\[
= \sum_{p=0}^{1} \sum_{q=0}^{4} b_{s,p,q} e_{s,p,q}^t,
\]
\[
e_{s,p,q} = 1 - \frac{4\gamma_s}{K} + \frac{4\gamma_s}{K} (\xi_1^p \xi_2^q)
\]
\[ P_\pi(i, a) = \sum_{t=0}^{\infty} \text{Po}(t) \prod_{s=1}^{4} \sum_{p=0}^{1} \sum_{q=0}^{4} b_{s,p,q} e_{s,p,q} t \]

\[ = \sum_{t=0}^{\infty} \text{Po}(t) \sum_{p_1=0}^{1} \sum_{p_2=0}^{1} \sum_{p_3=0}^{1} \sum_{p_4=0}^{1} \sum_{q_1=0}^{4} \sum_{q_2=0}^{4} \sum_{q_3=0}^{4} \sum_{q_4=0}^{4} \left( \prod_{s=1}^{4} b_{s,p,s,q} \right) \sum_{t=0}^{\infty} \text{Po}(t) \left( \prod_{s=1}^{4} e_{s,p,s,q} \right)^t \]

\[ = \sum_{p_1=0}^{1} \sum_{p_2=0}^{1} \sum_{p_3=0}^{1} \sum_{p_4=0}^{1} \sum_{q_1=0}^{4} \sum_{q_2=0}^{4} \sum_{q_3=0}^{4} \sum_{q_4=0}^{4} \left( \prod_{s=1}^{4} b_{s,p,s,q} \right) \exp \left\{ -K \theta \nu \left( 1 - \prod_{s} e_{s,p,s,q} \right) \right\} \]
Model fitting

• 12 parameters
• numerical (\textit{nlminb} in R) (\textit{point estimation})
• extensive code testing
• MCMC (Bayes under flat prior) (\textit{induced parameters and prediction})
MCMC looks good
Posterior predictive checks look good
Cross validation

Table S3-5: Predicted number of extra genes confirmed by a 4th study based on modeling the other three studies.

<table>
<thead>
<tr>
<th>Leave Out</th>
<th>Predicted Additional</th>
<th>95% Prediction Interval</th>
<th>Observed Additional</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL-1</td>
<td>133</td>
<td>(76, 207)</td>
<td>136</td>
</tr>
<tr>
<td>U2OS</td>
<td>128</td>
<td>(75, 199)</td>
<td>114</td>
</tr>
<tr>
<td>A549US</td>
<td>144</td>
<td>(89, 212)</td>
<td>188</td>
</tr>
<tr>
<td>A549DE</td>
<td>156</td>
<td>(80, 284)</td>
<td>131</td>
</tr>
</tbody>
</table>
## Point estimates

<table>
<thead>
<tr>
<th></th>
<th>MLE</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta$</td>
<td>0.128</td>
<td>0.128</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.083</td>
<td>0.164</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.340</td>
<td>0.400</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.312</td>
<td>0.375</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.038</td>
<td>0.125</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.063</td>
<td>0.072</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.094</td>
<td>0.107</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>0.113</td>
<td>0.127</td>
</tr>
<tr>
<td>$\gamma_4$</td>
<td>0.084</td>
<td>0.095</td>
</tr>
<tr>
<td>$\omega$</td>
<td>0.809</td>
<td>0.902</td>
</tr>
<tr>
<td>$\nu$</td>
<td>0.000</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Density of False Discovery Rate (FDR) and False Non-Discovery Rate (FNDR)

\[ P(I_g = 0|C_{g,s} = 1) \]

\[ P(I_g = 1|C_{g,s} = 0) \]

Density of False Positive Rate (FP) and False Negative Rate (FN)

\[ P(C_{g,s} = 1|I_g = 0) \]

\[ P(C_{g,s} = 0|I_g = 1) \]
Predicted number of confirmed genes

Cumulative number of confirmed genes (square root scale)

Number of studies (log scale)

1. DL-1
2. U2OS
3. A549DE
4. A549US

95% confidence band of predicted number of confirmed genes
50% confidence band of predicted number of confirmed genes
95% confidence interval of number of involved genes

+1 study
+10 studies
+20 studies
+40 studies
What about low estimated off target rate ??

Kulkarni et al. 2006, Nat. Meth.
- computational predictions
- overestimate
where the last line comes from the moment generating function of a Negative Binomial distribution.

Finally, the pattern probability $P_{\pi}(i, a)$ is obtained by summing over the 25 states of $i$ and $a$, as indicated previously. This provides a route to computing all 81 multi-study pattern probabilities required for likelihood evaluation.

$$P_{\pi}(i, a) = \sum_{t=0}^{\infty} \text{Po}(t) \prod_{s=1}^{4} \sum_{p=0}^{1} \sum_{q=0}^{4} b_{s,p,q} e_{s,p,q} t$$

$$= \sum_{t=0}^{\infty} \text{Po}(t) \sum_{p_1=0}^{1} \sum_{p_2=0}^{1} \sum_{p_3=0}^{1} \sum_{p_4=0}^{1} \sum_{q_1=0}^{1} \sum_{q_2=0}^{1} \sum_{q_3=0}^{1} \sum_{q_4=0}^{1} \left( \prod_{s=1}^{4} b_{s,p_s,q_s} \right) \left( \prod_{s=1}^{4} e_{s,p_s,q_s} \right) t$$

$$= \sum_{p_1=0}^{1} \sum_{p_2=0}^{1} \sum_{p_3=0}^{1} \sum_{p_4=0}^{1} \sum_{q_1=0}^{1} \sum_{q_2=0}^{1} \sum_{q_3=0}^{1} \sum_{q_4=0}^{1} \left( \prod_{s=1}^{4} b_{s,p_s,q_s} \right) \sum_{t=0}^{\infty} \text{Po}(t) \left( \prod_{s=1}^{4} e_{s,p_s,q_s} \right)^t$$

$$= \sum_{p_1=0}^{1} \sum_{p_2=0}^{1} \sum_{p_3=0}^{1} \sum_{p_4=0}^{1} \sum_{q_1=0}^{1} \sum_{q_2=0}^{1} \sum_{q_3=0}^{1} \sum_{q_4=0}^{1} \left( \prod_{s=1}^{4} b_{s,p_s,q_s} \right) \exp \left\{ -K \theta \nu \left( 1 - \prod_{s=1}^{4} e_{s,p_s,q_s} \right) \right\}$$

very hard to extend to Negative Binomial
Sensitivity analysis

• large $K \iff$ independent studies
• simpler likelihood
• separate implementation with Negative Binomial gives essentially the same fits
Profile analysis

Profile log likelihood

max_[other parameters] log Prob[ data given off–target rate ]

Detection/confirmation rate

Prob[ detected ]

Prob[ confirmed if detected ]

Four study average

Error rate

FDR = Prob[ not involved if confirmed ]

FNDR = Prob[ involved if not confirmed ]

Off target rate per siRNA
Summary

1. gene-level agreement among studies
2. functional category analysis
3. protein interaction analysis