Page 12: [1] Deleted Author 3/20/16 4:55 PM

a

Page 12: [1] Deleted Author 3/20/16 4:55 PM

a

Page 12: [1] Deleted Author 3/20/16 4:55 PM

a

Page 12: [1] Deleted Author 3/20/16 4:55 PM

a

Page 12: [1] Deleted Author 3/20/16 4:55 PM

a

Page 12: [1] Deleted Author 3/20/16 4:55 PM

a

Page 12: [1] Deleted Author 3/20/16 4:55 PM

a

Page 12: [2] Deleted Author 3/20/16 4:55 PM

(PMID:26017442)

Page 12: [2] Deleted Author 3/20/16 4:55 PM

(PMID:26017442)

Page 12: [2] Deleted Author 3/20/16 4:55 PM

(PMID:26017442)

Page 12: [2] Deleted Author 3/20/16 4:55 PM

(PMID:26017442)

Page 12: [2] Deleted Author 3/20/16 4:55 PM

(PMID:26017442)

Page 12: [2] Deleted Author 3/20/16 4:55 PM

(PMID:26017442)

Page 12: [2] Deleted Author 3/20/16 4:55 PM

(PMID:26017442)

Page 12: [2] Deleted Author 3/20/16 4:55 PM

(PMID:26017442)

Page 12: [3] Deleted Author 3/20/16 4:55 PM

”.

Page 12: [3] Deleted Author 3/20/16 4:55 PM

”.

Page 12: [3] Deleted Author 3/20/16 4:55 PM

”.

Page 12: [3] Deleted Author 3/20/16 4:55 PM

”.

Page 12: [3] Deleted Author 3/20/16 4:55 PM

”.

Page 12: [3] Deleted Author 3/20/16 4:55 PM

”.

Page 14: [4] Deleted Author 3/20/16 4:55 PM

pol

Page 14: [4] Deleted Author 3/20/16 4:55 PM

pol

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [5] Deleted Author 3/20/16 4:55 PM

\cite{22955619, 25164757, 22125477}.

Page 14: [6] Deleted Author 3/20/16 4:55 PM

the

Page 14: [6] Deleted **Author** **3/20/16 4:55 PM**

the

Page 14: [6] Deleted **Author** **3/20/16 4:55 PM**

the

Page 14: [6] Deleted **Author** **3/20/16 4:55 PM**

the

Page 14: [6] Deleted **Author** **3/20/16 4:55 PM**

the

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [7] Deleted **Author** **3/20/16 4:55 PM**

to be updated to analyze

Page 22: [8] Deleted **Author** **3/20/16 4:55 PM**

(IDR)

Page 22: [8] Deleted **Author** **3/20/16 4:55 PM**

(IDR)

Page 22: [8] Deleted **Author** **3/20/16 4:55 PM**

(IDR)

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 22: [9] Deleted **Author** **3/20/16 4:55 PM**

and

Page 24: [10] Deleted **Author** **3/20/16 4:55 PM**

simpler

Page 24: [10] Deleted **Author** **3/20/16 4:55 PM**

simpler

Page 24: [10] Deleted **Author** **3/20/16 4:55 PM**

simpler

Page 24: [10] Deleted **Author** **3/20/16 4:55 PM**

simpler

Page 24: [11] Deleted **Author** **3/20/16 4:55 PM**

compare

Page 24: [11] Deleted **Author** **3/20/16 4:55 PM**

compare

Page 24: [11] Deleted **Author** **3/20/16 4:55 PM**

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [11] Deleted Author 3/20/16 4:55 PM

compare

Page 24: [12] Deleted Author 3/20/16 4:55 PM

, NAR 24335146

Page 24: [12] Deleted Author 3/20/16 4:55 PM

, NAR 24335146

Page 25: [13] Deleted Author 3/20/16 4:55 PM

Because of

Page 25: [13] Deleted Author 3/20/16 4:55 PM

Because of

Page 25: [13] Deleted Author 3/20/16 4:55 PM

Because of

Page 25: [13] Deleted Author 3/20/16 4:55 PM

Because of

Page 25: [13] Deleted Author 3/20/16 4:55 PM

Because of

Page 25: [13] Deleted Author 3/20/16 4:55 PM

Because of

Page 25: [13] Deleted Author 3/20/16 4:55 PM

Because of

Page 25: [14] Deleted Author 3/20/16 4:55 PM

in ENCODE

Page 25: [14] Deleted Author 3/20/16 4:55 PM

in ENCODE

Page 25: [14] Deleted Author 3/20/16 4:55 PM

in ENCODE

Page 25: [14] Deleted Author 3/20/16 4:55 PM

in ENCODE

Page 25: [14] Deleted Author 3/20/16 4:55 PM

in ENCODE

Page 25: [15] Deleted Author 3/20/16 4:55 PM

a

Page 25: [15] Deleted Author 3/20/16 4:55 PM

a

Page 25: [15] Deleted Author 3/20/16 4:55 PM

a

Page 25: [15] Deleted Author 3/20/16 4:55 PM

a

Page 25: [15] Deleted Author 3/20/16 4:55 PM

a

Page 25: [15] Deleted Author 3/20/16 4:55 PM

a

Page 25: [16] Deleted Author 3/20/16 4:55 PM

,

Page 25: [16] Deleted Author 3/20/16 4:55 PM

,

Page 25: [16] Deleted Author 3/20/16 4:55 PM

,

Page 25: [16] Deleted Author 3/20/16 4:55 PM

,

Page 25: [16] Deleted Author 3/20/16 4:55 PM

,

Page 26: [17] Deleted Author 3/20/16 4:55 PM

output

Page 26: [17] Deleted Author 3/20/16 4:55 PM

output

Page 26: [18] Deleted **Author** **3/20/16 4:55 PM**

pipeline for

Page 26: [18] Deleted **Author** **3/20/16 4:55 PM**

pipeline for

Page 26: [18] Deleted **Author** **3/20/16 4:55 PM**

pipeline for

Page 26: [19] Deleted **Author** **3/20/16 4:55 PM**

of John Stamatoyannopoulos, because

Page 26: [19] Deleted **Author** **3/20/16 4:55 PM**

of John Stamatoyannopoulos, because

Page 26: [19] Deleted **Author** **3/20/16 4:55 PM**

of John Stamatoyannopoulos, because

Page 26: [19] Deleted **Author** **3/20/16 4:55 PM**

of John Stamatoyannopoulos, because

Page 26: [19] Deleted **Author** **3/20/16 4:55 PM**

of John Stamatoyannopoulos, because

Page 26: [19] Deleted **Author** **3/20/16 4:55 PM**

of John Stamatoyannopoulos, because

Page 28: [20] Deleted **Author** **3/20/16 4:55 PM**

hindbrain development

Page 28: [20] Deleted **Author** **3/20/16 4:55 PM**

hindbrain development

Page 28: [20] Deleted **Author** **3/20/16 4:55 PM**

hindbrain development

Page 28: [20] Deleted **Author** **3/20/16 4:55 PM**

hindbrain development

Page 28: [21] Deleted **Author** **3/20/16 4:55 PM**

in 2010 (

Page 28: [21] Deleted **Author** **3/20/16 4:55 PM**

in 2010 (

Page 28: [21] Deleted **Author** **3/20/16 4:55 PM**

in 2010 (

Page 28: [21] Deleted **Author** **3/20/16 4:55 PM**

in 2010 (

Page 28: [21] Deleted **Author** **3/20/16 4:55 PM**

in 2010 (

Page 28: [21] Deleted **Author** **3/20/16 4:55 PM**

in 2010 (

Page 28: [21] Deleted Author 3/20/16 4:55 PM

in 2010 (

Page 28: [21] Deleted Author 3/20/16 4:55 PM

in 2010 (

Page 28: [21] Deleted Author 3/20/16 4:55 PM

in 2010 (

Page 28: [21] Deleted Author 3/20/16 4:55 PM

in 2010 (

Page 28: [21] Deleted Author 3/20/16 4:55 PM

in 2010 (

Page 28: [21] Deleted Author 3/20/16 4:55 PM

in 2010 (

Page 28: [21] Deleted Author 3/20/16 4:55 PM

in 2010 (

Page 28: [21] Deleted Author 3/20/16 4:55 PM

in 2010 (

Page 28: [22] Deleted Author 3/20/16 4:55 PM

(Ay 2015 26328929)):

Page 28: [22] Deleted Author 3/20/16 4:55 PM

(Ay 2015 26328929)):

Page 28: [22] Deleted Author 3/20/16 4:55 PM

(Ay 2015 26328929)):

Page 28: [22] Deleted Author 3/20/16 4:55 PM

(Ay 2015 26328929)):

Page 28: [23] Comment [11] William Stafford Noble 3/17/16 11:00 PM

[15] E. Yaffe and A. Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet, 43:1059–1065, 2011.

[16] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods, 9:999–1003, 2012.

[17] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics, 28(23):3131–3133, 2012.

[18] A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, and J. Mozziconacci. Normalization of a chromosomal contact map. BMC Genomics, 13:436, 2012.

[19] W. Li, K. Gong, Q. Li, F. Alber, and X.J. Zhou. Hi-Corrector: a fast, scalable and

memory-efficient package
for normalizing large-scale Hi-C data. Bioinformatics, 31(6):960–962, 2015.
[20] Ei-Wen Yang and Tao Jiang. GDNorm: An improved poisson regression model for reducing biases in Hi-C
data. In Proceedings of the 14th International Workshop of Algorithms in Bioinformatics, volume 8701 of
Lecture Notes in Computer Science, pages 263–280, Berlin, Heidelberg, 2014. Springer-Verlag.
[21] Yoli Shavit and Pietro Lio'. Combining a wavelet change point and the bayes factor for analysing chromoso-
mal interaction data. Mol. Biosyst., 10(6):1576–1585, 2014.

Page 28: [24] Deleted **Author** **3/20/16 4:55 PM**

about the data

Page 28: [24] Deleted **Author** **3/20/16 4:55 PM**

about the data

Page 28: [24] Deleted **Author** **3/20/16 4:55 PM**

about the data

Page 28: [24] Deleted **Author** **3/20/16 4:55 PM**

about the data

Page 28: [24] Deleted **Author** **3/20/16 4:55 PM**

about the data

Page 28: [24] Deleted **Author** **3/20/16 4:55 PM**

about the data

Page 28: [24] Deleted **Author** **3/20/16 4:55 PM**

about the data

Page 28: [24] Deleted **Author** **3/20/16 4:55 PM**

about the data

Page 28: [24] Deleted **Author** **3/20/16 4:55 PM**

about the data

Page 31: [25] Deleted **Author** **3/20/16 4:55 PM**

DAC's

Page 31: [25] Deleted **Author** **3/20/16 4:55 PM**

DAC's

Page 31: [25] Deleted **Author** **3/20/16 4:55 PM**

DAC's

Page 31: [25] Deleted **Author** **3/20/16 4:55 PM**

DAC's

Page 31: [25] Deleted **Author** **3/20/16 4:55 PM**

DAC's

Page 31: [25] Deleted Author 3/20/16 4:55 PM

DAC’s

Page 31: [25] Deleted Author 3/20/16 4:55 PM

DAC’s

Page 31: [25] Deleted Author 3/20/16 4:55 PM

DAC’s

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted Author 3/20/16 4:55 PM

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [26] Deleted **Author** **3/20/16 4:55 PM**

of

Page 31: [27] Deleted **Author** **3/20/16 4:55 PM**

,

Page 31: [27] Deleted **Author** **3/20/16 4:55 PM**

,

Page 31: [27] Deleted **Author** **3/20/16 4:55 PM**

,

Page 31: [27] Deleted **Author** **3/20/16 4:55 PM**

,

Page 31: [27] Deleted **Author** **3/20/16 4:55 PM**

,

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 31: [28] Deleted **Author** **3/20/16 4:55 PM**

This

Page 32: [29] Deleted **Author** **3/20/16 4:55 PM**

in

Page 32: [29] Deleted **Author** **3/20/16 4:55 PM**

in

Page 32: [29] Deleted **Author** **3/20/16 4:55 PM**

in

Page 32: [29] Deleted **Author** **3/20/16 4:55 PM**

in

Page 32: [29] Deleted **Author** **3/20/16 4:55 PM**

in

Page 32: [29] Deleted **Author** **3/20/16 4:55 PM**

in

Page 32: [29] Deleted **Author** **3/20/16 4:55 PM**

in

Page 32: [29] Deleted **Author** **3/20/16 4:55 PM**

in

Page 32: [30] Deleted **Author** **3/20/16 4:55 PM**

group

Page 32: [30] Deleted **Author** **3/20/16 4:55 PM**

group

Page 32: [30] Deleted **Author** **3/20/16 4:55 PM**

group

Page 32: [30] Deleted **Author** **3/20/16 4:55 PM**

group

Page 32: [31] Deleted **Author** **3/20/16 4:55 PM**

The

Page 32: [31] Deleted **Author** **3/20/16 4:55 PM**

The

Page 32: [32] Deleted **Author** **3/20/16 4:55 PM**

.

Page 32: [32] Deleted **Author** **3/20/16 4:55 PM**

.

Page 32: [32] Deleted **Author** **3/20/16 4:55 PM**

.

Page 32: [32] Deleted **Author** **3/20/16 4:55 PM**

.

Page 32: [33] Deleted **Author** **3/20/16 4:55 PM**

The Regulation

Page 32: [33] Deleted **Author** **3/20/16 4:55 PM**

The Regulation

Page 32: [34] Deleted **Author** **3/20/16 4:55 PM**

A key continuing role

Page 32: [34] Deleted Author 3/20/16 4:55 PM

A key continuing role

Page 32: [34] Deleted Author 3/20/16 4:55 PM

A key continuing role

Page 32: [34] Deleted Author 3/20/16 4:55 PM

A key continuing role

Page 32: [34] Deleted Author 3/20/16 4:55 PM

A key continuing role

Page 32: [34] Deleted Author 3/20/16 4:55 PM

A key continuing role

Page 32: [34] Deleted Author 3/20/16 4:55 PM

A key continuing role

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [35] Deleted Author 3/20/16 4:55 PM

The DAC will set up a

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted Author 3/20/16 4:55 PM

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted **Author** **3/20/16 4:55 PM**

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted **Author** **3/20/16 4:55 PM**

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted **Author** **3/20/16 4:55 PM**

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted **Author** **3/20/16 4:55 PM**

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted **Author** **3/20/16 4:55 PM**

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [36] Deleted **Author** **3/20/16 4:55 PM**

Another important responsibility of the DAC in writing consortium papers is connecting

Page 33: [37] Deleted **Author** **3/20/16 4:55 PM**

that

Page 33: [37] Deleted **Author** **3/20/16 4:55 PM**

that

Page 33: [37] Deleted **Author** **3/20/16 4:55 PM**

that

Page 33: [37] Deleted **Author** **3/20/16 4:55 PM**

that

Page 33: [37] Deleted **Author** **3/20/16 4:55 PM**

that

Page 33: [37] Deleted **Author** **3/20/16 4:55 PM**

that

Page 33: [37] Deleted **Author** **3/20/16 4:55 PM**

that

Page 33: [37] Deleted **Author** **3/20/16 4:55 PM**

that

Page 33: [37] Deleted **Author** **3/20/16 4:55 PM**

that

Page 35: [38] Deleted **Author** **3/20/16 4:55 PM**

data,

Page 35: [38] Deleted **Author** **3/20/16 4:55 PM**

data,

Page 35: [38] Deleted **Author** **3/20/16 4:55 PM**

data,

Page 35: [38] Deleted **Author** **3/20/16 4:55 PM**

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [38] Deleted Author 3/20/16 4:55 PM

data,

Page 35: [39] Deleted Author 3/20/16 4:55 PM

stay as truthfully as possible to the

Page 35: [39] Deleted Author 3/20/16 4:55 PM

stay as truthfully as possible to the

Page 35: [39] Deleted Author 3/20/16 4:55 PM

stay as truthfully as possible to the

Page 35: [39] Deleted Author 3/20/16 4:55 PM

stay as truthfully as possible to the

Page 35: [40] Deleted Author 3/20/16 4:55 PM

, ChIP-seq of

Page 35: [40] Deleted Author 3/20/16 4:55 PM

, ChIP-seq of

Page 35: [40] Deleted Author 3/20/16 4:55 PM

, ChIP-seq of

Page 35: [40] Deleted Author 3/20/16 4:55 PM

, ChIP-seq of

Page 36: [41] Deleted Author 3/20/16 4:55 PM

matrix

Page 36: [41] Deleted Author 3/20/16 4:55 PM

matrix

Page 36: [41] Deleted Author 3/20/16 4:55 PM

matrix

Page 36: [42] Deleted Author 3/20/16 4:55 PM

PMID: 22936248

Page 36: [42] Deleted Author 3/20/16 4:55 PM

PMID: 22936248

Page 36: [42] Deleted Author 3/20/16 4:55 PM

PMID: 22936248

Page 36: [42] Deleted Author 3/20/16 4:55 PM

PMID: 22936248

Page 36: [42] Deleted Author 3/20/16 4:55 PM

PMID: 22936248

Page 36: [42] Deleted Author 3/20/16 4:55 PM

PMID: 22936248

Page 36: [42] Deleted Author 3/20/16 4:55 PM

PMID: 22936248

Page 36: [43] Deleted Author 3/20/16 4:55 PM

it

Page 36: [43] Deleted Author 3/20/16 4:55 PM

it

Page 36: [43] Deleted Author 3/20/16 4:55 PM

it

Page 36: [43] Deleted **Author** **3/20/16 4:55 PM**

it

Page 36: [43] Deleted **Author** **3/20/16 4:55 PM**

it

Page 36: [43] Deleted **Author** **3/20/16 4:55 PM**

it

Page 36: [43] Deleted **Author** **3/20/16 4:55 PM**

it

Page 36: [43] Deleted **Author** **3/20/16 4:55 PM**

it

Page 36: [43] Deleted **Author** **3/20/16 4:55 PM**

it

Page 36: [43] Deleted **Author** **3/20/16 4:55 PM**

it

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [44] Deleted **Author** **3/20/16 4:55 PM**

lies in

Page 36: [45] Deleted **Author** **3/20/16 4:55 PM**

lab effect

Page 36: [45] Deleted Author 3/20/16 4:55 PM

lab effect

Page 36: [45] Deleted Author 3/20/16 4:55 PM

lab effect

Page 37: [46] Deleted Author 3/20/16 4:55 PM

Page 37: [46] Deleted Author 3/20/16 4:55 PM

Page 37: [46] Deleted Author 3/20/16 4:55 PM

Page 37: [47] Deleted Author 3/20/16 4:55 PM

Page 37: [47] Deleted Author 3/20/16 4:55 PM

Page 37: [47] Deleted Author 3/20/16 4:55 PM

Page 37: [47] Deleted Author 3/20/16 4:55 PM

Page 37: [47] Deleted Author 3/20/16 4:55 PM

Page 37: [47] Deleted Author 3/20/16 4:55 PM

Page 37: [47] Deleted Author 3/20/16 4:55 PM

Page 37: [48] Deleted Author 3/20/16 4:55 PM

compared with

Page 37: [48] Deleted Author 3/20/16 4:55 PM

compared with

Page 37: [48] Deleted Author 3/20/16 4:55 PM

compared with

Page 37: [48] Deleted Author 3/20/16 4:55 PM

compared with

Page 37: [48] Deleted Author 3/20/16 4:55 PM

compared with

Page 37: [48] Deleted Author 3/20/16 4:55 PM

compared with

Page 37: [48] Deleted Author 3/20/16 4:55 PM

compared with

Page 37: [49] Formatted Author 3/20/16 4:55 PM

Not Highlight

Page 37: [49] Formatted Author 3/20/16 4:55 PM

Not Highlight

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 37: [50] Deleted Author 3/20/16 4:55 PM

,

Page 38: [51] Deleted Author 3/20/16 4:55 PM

).

Page 38: [51] Deleted Author 3/20/16 4:55 PM

).

Page 38: [51] Deleted Author 3/20/16 4:55 PM

).

Page 38: [51] Deleted Author 3/20/16 4:55 PM

).

Page 38: [51] Deleted Author 3/20/16 4:55 PM

).

Page 38: [52] Deleted Author 3/20/16 4:55 PM

significant

Page 38: [52] Deleted Author 3/20/16 4:55 PM

significant

Page 38: [52] Deleted Author 3/20/16 4:55 PM

significant

Page 38: [52] Deleted Author 3/20/16 4:55 PM

significant

Page 38: [52] Deleted Author 3/20/16 4:55 PM

significant

Page 38: [52] Deleted Author 3/20/16 4:55 PM

significant

Page 38: [53] Deleted Author 3/20/16 4:55 PM

The

Page 38: [53] Deleted Author 3/20/16 4:55 PM

The

Page 38: [53] Deleted Author 3/20/16 4:55 PM

The

Page 38: [53] Deleted Author 3/20/16 4:55 PM

The

Page 38: [53] Deleted Author 3/20/16 4:55 PM

The

Page 38: [53] Deleted Author 3/20/16 4:55 PM

The

Page 38: [53] Deleted Author 3/20/16 4:55 PM

The

Page 38: [54] Deleted Author 3/20/16 4:55 PM

the

Page 38: [54] Deleted Author 3/20/16 4:55 PM

the

Page 39: [55] Deleted Author 3/20/16 4:55 PM

in

Page 39: [55] Deleted Author 3/20/16 4:55 PM

in

Page 39: [55] Deleted Author 3/20/16 4:55 PM

in

Page 39: [55] Deleted Author 3/20/16 4:55 PM

in

Page 39: [55] Deleted Author 3/20/16 4:55 PM

in

Page 39: [55] Deleted **Author** **3/20/16 4:55 PM**

in

Page 39: [55] Deleted **Author** **3/20/16 4:55 PM**

in

Page 39: [55] Deleted **Author** **3/20/16 4:55 PM**

in

Page 39: [55] Deleted **Author** **3/20/16 4:55 PM**

in

Page 39: [55] Deleted **Author** **3/20/16 4:55 PM**

in

Page 39: [55] Deleted **Author** **3/20/16 4:55 PM**

in

Page 39: [55] Deleted **Author** **3/20/16 4:55 PM**

in

Page 39: [56] Deleted **Author** **3/20/16 4:55 PM**

in an analogous way to

Page 39: [56] Deleted **Author** **3/20/16 4:55 PM**

in an analogous way to

Page 39: [56] Deleted **Author** **3/20/16 4:55 PM**

in an analogous way to

Page 39: [56] Deleted **Author** **3/20/16 4:55 PM**

in an analogous way to

Page 39: [56] Deleted **Author** **3/20/16 4:55 PM**

in an analogous way to

Page 39: [56] Deleted **Author** **3/20/16 4:55 PM**

in an analogous way to

Page 39: [57] Deleted **Author** **3/20/16 4:55 PM**

Thus for

Page 39: [57] Deleted **Author** **3/20/16 4:55 PM**

Thus for

Page 39: [57] Deleted **Author** **3/20/16 4:55 PM**

Thus for

Page 39: [57] Deleted **Author** **3/20/16 4:55 PM**

Thus for

Page 39: [57] Deleted **Author** **3/20/16 4:55 PM**

Thus for

Page 39: [57] Deleted **Author** **3/20/16 4:55 PM**

Thus for

Page 39: [57] Deleted Author 3/20/16 4:55 PM

Thus for

Page 39: [57] Deleted Author 3/20/16 4:55 PM

Thus for

Page 39: [57] Deleted Author 3/20/16 4:55 PM

Thus for

Page 39: [58] Deleted Author 3/20/16 4:55 PM

in

Page 39: [58] Deleted Author 3/20/16 4:55 PM

in

Page 39: [58] Deleted Author 3/20/16 4:55 PM

in

Page 39: [58] Deleted Author 3/20/16 4:55 PM

in

Page 39: [58] Deleted Author 3/20/16 4:55 PM

in

Page 39: [58] Deleted Author 3/20/16 4:55 PM

in

Page 39: [58] Deleted Author 3/20/16 4:55 PM

in

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 39: [59] Deleted Author 3/20/16 4:55 PM

One

Page 40: [60] Deleted Author 3/20/16 4:55 PM

(except that TF peaks are bigger, several hundred bps).

Page 40: [60] Deleted Author 3/20/16 4:55 PM

(except that TF peaks are bigger, several hundred bps).

Page 40: [60] Deleted Author 3/20/16 4:55 PM

(except that TF peaks are bigger, several hundred bps).

Page 40: [60] Deleted Author 3/20/16 4:55 PM

(except that TF peaks are bigger, several hundred bps).

Page 40: [60] Deleted Author 3/20/16 4:55 PM

(except that TF peaks are bigger, several hundred bps).

Page 40: [60] Deleted Author 3/20/16 4:55 PM

(except that TF peaks are bigger, several hundred bps).

Page 40: [60] Deleted Author 3/20/16 4:55 PM

(except that TF peaks are bigger, several hundred bps).

**Page 40: [60] Deleted	Author	3/20/16 4:55 PM**

(except that TF peaks are bigger, several hundred bps).

**Page 40: [60] Deleted	Author	3/20/16 4:55 PM**

(except that TF peaks are bigger, several hundred bps).

**Page 40: [60] Deleted	Author	3/20/16 4:55 PM**

(except that TF peaks are bigger, several hundred bps).

**Page 40: [60] Deleted	Author	3/20/16 4:55 PM**

(except that TF peaks are bigger, several hundred bps).

**Page 40: [60] Deleted	Author	3/20/16 4:55 PM**

(except that TF peaks are bigger, several hundred bps).

**Page 40: [60] Deleted	Author	3/20/16 4:55 PM**

(except that TF peaks are bigger, several hundred bps).

**Page 40: [61] Deleted	Author	3/20/16 4:55 PM**

Thus or RBP

**Page 40: [61] Deleted	Author	3/20/16 4:55 PM**

Thus or RBP

**Page 40: [61] Deleted	Author	3/20/16 4:55 PM**

Thus or RBP

**Page 40: [61] Deleted	Author	3/20/16 4:55 PM**

Thus or RBP

**Page 40: [61] Deleted	Author	3/20/16 4:55 PM**

Thus or RBP

**Page 40: [61] Deleted	Author	3/20/16 4:55 PM**

Thus or RBP

**Page 40: [61] Deleted	Author	3/20/16 4:55 PM**

Thus or RBP

**Page 40: [62] Deleted	Author	3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

**Page 40: [62] Deleted	Author	3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

**Page 40: [62] Deleted	Author	3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

**Page 40: [62] Deleted	Author	3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

**Page 40: [62] Deleted	Author	3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

**Page 40: [62] Deleted	Author	3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [62] Deleted **Author** **3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [62] Deleted **Author** **3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [62] Deleted **Author** **3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [62] Deleted **Author** **3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [62] Deleted **Author** **3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [62] Deleted **Author** **3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [62] Deleted **Author** **3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [62] Deleted **Author** **3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [62] Deleted **Author** **3/20/16 4:55 PM**

cell types. These datasets are at 40 kb

Page 40: [63] Deleted **Author** **3/20/16 4:55 PM**

above

Page 40: [63] Deleted **Author** **3/20/16 4:55 PM**

above

Page 40: [63] Deleted **Author** **3/20/16 4:55 PM**

above

Page 40: [63] Deleted **Author** **3/20/16 4:55 PM**

above

Page 40: [63] Deleted **Author** **3/20/16 4:55 PM**

above

Page 41: [64] Deleted **Author** **3/20/16 4:55 PM**

),

Page 41: [64] Deleted **Author** **3/20/16 4:55 PM**

),

Page 41: [64] Deleted **Author** **3/20/16 4:55 PM**

),

Page 41: [64] Deleted **Author** **3/20/16 4:55 PM**

),

Page 41: [64] Deleted **Author** **3/20/16 4:55 PM**

),

Page 41: [64] Deleted Author 3/20/16 4:55 PM

),

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [65] Deleted Author 3/20/16 4:55 PM

to minimizing batch effects that may be caused by differences among production centers.

Page 41: [66] Deleted Author 3/20/16 4:55 PM

We

Page 41: [66] Deleted Author 3/20/16 4:55 PM

We

Page 42: [67] Deleted Author 3/20/16 4:55 PM

require the integration of

Page 42: [67] Deleted Author 3/20/16 4:55 PM

require the integration of

Page 42: [67] Deleted Author 3/20/16 4:55 PM

require the integration of

Page 42: [67] Deleted Author 3/20/16 4:55 PM

require the integration of

Page 42: [67] Deleted Author 3/20/16 4:55 PM

require the integration of

Page 42: [67] Deleted Author 3/20/16 4:55 PM

require the integration of

Page 42: [68] Deleted Author 3/20/16 4:55 PM

elements

Page 42: [68] Deleted Author 3/20/16 4:55 PM

elements

Page 42: [68] Deleted Author 3/20/16 4:55 PM

elements

Page 42: [69] Deleted Author 3/20/16 4:55 PM

, 2006 16719718}. Active

Page 42: [69] Deleted Author 3/20/16 4:55 PM

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [69] Deleted **Author** **3/20/16 4:55 PM**

, 2006 16719718}. Active

Page 42: [70] Deleted **Author** **3/20/16 4:55 PM**

computational efforts in predicting enhancers by integrating

Page 42: [70] Deleted **Author** **3/20/16 4:55 PM**

computational efforts in predicting enhancers by integrating

Page 42: [70] Deleted **Author** **3/20/16 4:55 PM**

computational efforts in predicting enhancers by integrating

Page 42: [70] Deleted **Author** **3/20/16 4:55 PM**

computational efforts in predicting enhancers by integrating

Page 42: [70] Deleted **Author** **3/20/16 4:55 PM**

computational efforts in predicting enhancers by integrating

Page 42: [70] Deleted **Author** **3/20/16 4:55 PM**

computational efforts in predicting enhancers by integrating

Page 42: [70] Deleted **Author** **3/20/16 4:55 PM**

computational efforts in predicting enhancers by integrating

Page 42: [70] Deleted **Author** **3/20/16 4:55 PM**

computational efforts in predicting enhancers by integrating

Page 42: [71] Deleted Author 3/20/16 4:55 PM

defining

Page 42: [71] Deleted Author 3/20/16 4:55 PM

defining

Page 44: [72] Deleted Author 3/20/16 4:55 PM

such

Page 44: [72] Deleted Author 3/20/16 4:55 PM

such

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [73] Deleted Author 3/20/16 4:55 PM

further develop

Page 44: [74] Deleted Author 3/20/16 4:55 PM

During

Page 44: [74] Deleted Author 3/20/16 4:55 PM

During

Page 44: [74] Deleted Author 3/20/16 4:55 PM

During

Page 44: [74] Deleted Author 3/20/16 4:55 PM

During

Page 44: [74] Deleted Author 3/20/16 4:55 PM

During

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 44: [75] Deleted Author 3/20/16 4:55 PM

,

Page 45: [76] Deleted Author 3/20/16 4:55 PM

technique

Page 45: [76] Deleted Author 3/20/16 4:55 PM

technique

Page 45: [76] Deleted Author 3/20/16 4:55 PM

technique

Page 45: [76] Deleted Author 3/20/16 4:55 PM

technique

Page 45: [76] Deleted Author 3/20/16 4:55 PM

technique

Page 45: [76] Deleted Author 3/20/16 4:55 PM

technique

Page 45: [76] Deleted Author 3/20/16 4:55 PM

technique

Page 45: [76] Deleted Author 3/20/16 4:55 PM

technique

Page 45: [76] Deleted Author 3/20/16 4:55 PM

technique

Page 45: [77] Deleted Author 3/20/16 4:55 PM

silencers,

Page 45: [77] Deleted Author 3/20/16 4:55 PM

silencers,

Page 45: [77] Deleted Author 3/20/16 4:55 PM

silencers,

Page 45: [77] Deleted Author 3/20/16 4:55 PM

silencers,

Page 45: [77] Deleted Author 3/20/16 4:55 PM

silencers,

Page 45: [77] Deleted Author 3/20/16 4:55 PM

silencers,

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [78] Deleted Author 3/20/16 4:55 PM

a two-label annotation

Page 46: [79] Deleted Author 3/20/16 4:55 PM

.

Page 46: [79] Deleted Author 3/20/16 4:55 PM

.

Page 46: [79] Deleted Author 3/20/16 4:55 PM

.

Page 46: [79] Deleted Author 3/20/16 4:55 PM

.

Page 46: [79] Deleted Author 3/20/16 4:55 PM

.

Page 46: [79] Deleted Author 3/20/16 4:55 PM

.

Page 46: [80] Deleted Author 3/20/16 4:55 PM

,

Page 46: [80] Deleted Author 3/20/16 4:55 PM

,

Page 47: [81] Deleted Author 3/20/16 4:55 PM

here

Page 47: [81] Deleted Author 3/20/16 4:55 PM

here

Page 47: [81] Deleted Author 3/20/16 4:55 PM

here

Page 47: [81] Deleted Author 3/20/16 4:55 PM

here

Page 47: [81] Deleted Author 3/20/16 4:55 PM

here

Page 47: [81] Deleted Author 3/20/16 4:55 PM

here

Page 47: [81] Deleted Author 3/20/16 4:55 PM

here

Page 47: [82] Deleted Author 3/20/16 4:55 PM

,

Page 47: [82] Deleted Author 3/20/16 4:55 PM

,

Page 47: [82] Deleted Author 3/20/16 4:55 PM

,

Page 47: [82] Deleted Author 3/20/16 4:55 PM

,

Page 47: [82] Deleted Author 3/20/16 4:55 PM

,

Page 47: [82] Deleted Author 3/20/16 4:55 PM

,

Page 47: [82] Deleted Author 3/20/16 4:55 PM

,

Page 47: [82] Deleted Author 3/20/16 4:55 PM

,

Page 47: [82] Deleted Author 3/20/16 4:55 PM

,

Page 47: [82] Deleted **Author** **3/20/16 4:55 PM**

,

Page 47: [82] Deleted **Author** **3/20/16 4:55 PM**

,

Page 47: [82] Deleted **Author** **3/20/16 4:55 PM**

,

Page 47: [83] Deleted **Author** **3/20/16 4:55 PM**

sequencing

Page 47: [83] Deleted **Author** **3/20/16 4:55 PM**

sequencing

Page 47: [83] Deleted **Author** **3/20/16 4:55 PM**

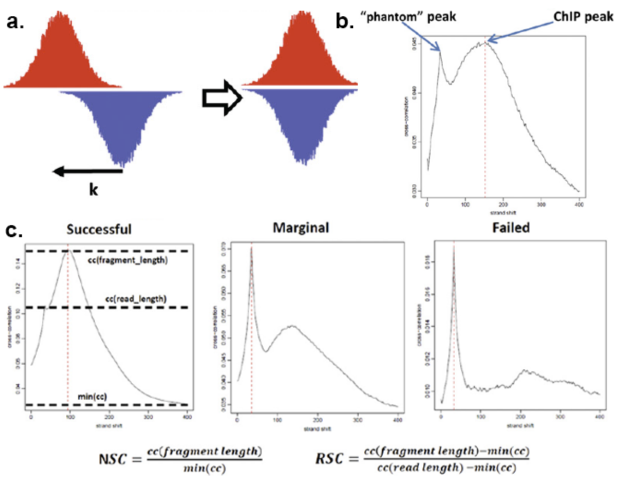
sequencing

Page 48: [84] Deleted **Author** **3/20/16 4:55 PM**

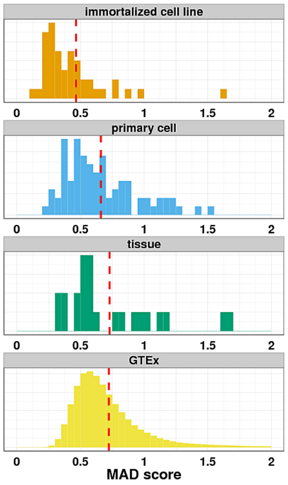
The DAC[1] recently evaluated correlation-based methods using the capture Hi-C data on GM12878 cells {Mifsud
2015 NG PMID: 25938943}

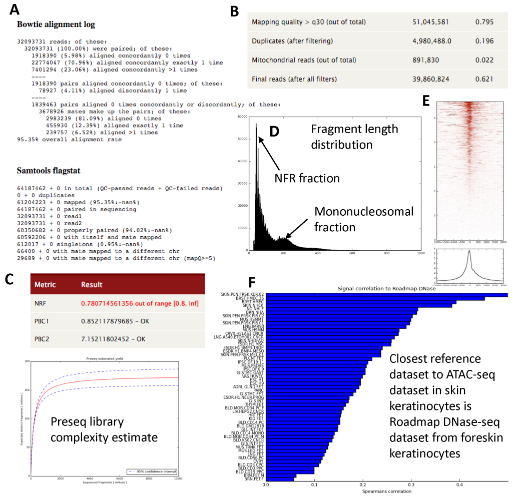
The DAC will continue to evaluate existing methods. In addition, we will also

Page 51: [85] Deleted **Author** **3/20/16 4:55 PM**



$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$





Bowtie alignment log

```
32093731 reads; of these:
  32093731 (100.00%) were paired; of these:
    1918390 (5.98%) aligned concordantly 0 times
    22774047 (70.96%) aligned concordantly exactly 1 time
    7401294 (23.06%) aligned concordantly >1 times
    ----
    1918390 pairs aligned concordantly 0 times; of these:
      78927 (4.11%) aligned discordantly 1 time
    ----
    1839463 pairs aligned 0 times concordantly or discordantly; of these:
      3678926 mates make up the pairs; of these:
        2983239 (81.09%) aligned 0 times
        455930 (12.39%) aligned exactly 1 time
        239757 (6.52%) aligned >1 times
95.35% overall alignment rate
```

Samtools flagstat

```
64187462 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
61204223 + 0 mapped (95.35%:-nan%)
64187462 + 0 paired in sequencing
32093731 + 0 read1
32093731 + 0 read2
60350682 + 0 properly paired (94.02%:-nan%)
60592206 + 0 with itself and mate mapped
612017 + 0 singletons (0.95%:-nan%)
66400 + 0 with mate mapped to a different chr
29689 + 0 with mate mapped to a different chr (mapQ>=5)
```

Mapping quality > q30 (out of total)	51,045,581	0.795
Duplicates (after filtering)	4,980,488.0	0.196
Mitochondrial reads (out of total)	891,830	0.022
Final reads (after all filters)	39,860,824	0.621

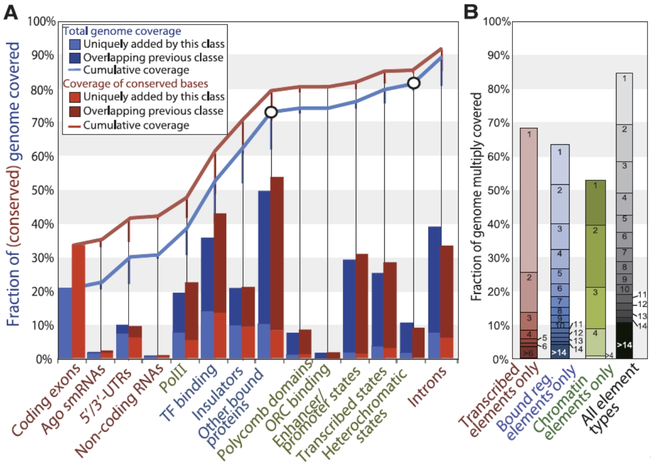
Metric	Result
NRF	0.780714561356 out of range [0.8, inf]
PBC1	0.852117879685 - OK
PBC2	7.15211802452 - OK

Closest reference dataset to ATAC-seq dataset in skin keratinocytes is Roadmap DNase-seq dataset from foreskin keratinocytes

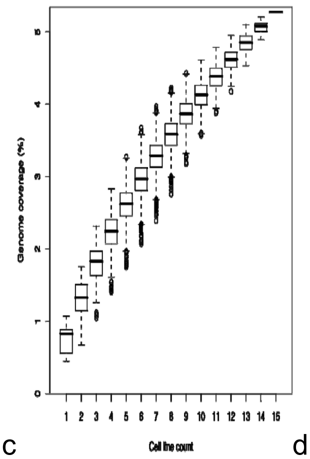
Page 51: [86] Comment [29] Zhiping Weng 3/21/16 2:55 AM

+purcaro@gmail.com Michael can you make one file out of the panels in this figure? Please make sure to preserve the quality. You need to put a "d" label in the middle panel, and change ABCDEF to efghij for the right panel. Thanks!

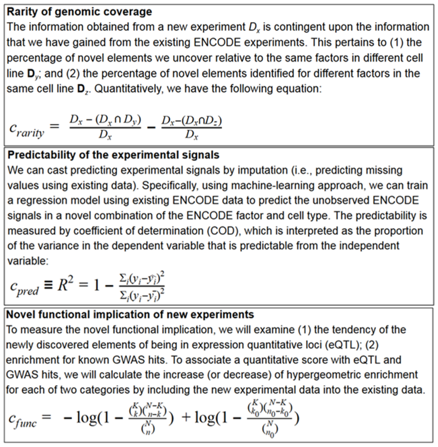
Page 54: [87] Deleted Author 3/20/16 4:55 PM



c



d



Rarity of genomic coverage

The information obtained from a new experiment D_x is contingent upon the information that we have gained from the existing ENCODE experiments. This pertains to (1) the percentage of novel elements we uncover relative to the same factors in different cell line $\mathbf{D}_y$; and (2) the percentage of novel elements identified for different factors in the same cell line $\mathbf{D}_z$. Quantitatively, we have the following equation:

$$c_{rarity} = \frac{D_x - (D_x \cap D_y)}{D_x} - \frac{D_x - (D_x \cap D_z)}{D_x}$$

Predictability of the experimental signals

We can cast predicting experimental signals by imputation (i.e., predicting missing values using existing data). Specifically, using machine-learning approach, we can train a regression model using existing ENCODE data to predict the unobserved ENCODE signals in a novel combination of the ENCODE factor and cell type. The predictability is measured by coefficient of determination (COD), which is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable:

$$c_{pred} \equiv R^2 = 1 - \frac{\Sigma_i (y_i - \hat{y}_i)^2}{\Sigma_i (y_i - \bar{y}_i)^2}$$

Novel functional implication of new experiments

To measure the novel functional implication, we will examine (1) the tendency of the newly discovered elements of being in expression quantitative loci (eQTL); (2) enrichment for known GWAS hits. To associate a quantitative score with eQTL and GWAS hits, we will calculate the increase (or decrease) of hypergeometric enrichment for each of two categories by including the new experimental data into the existing data.

$$c_{func} = -\log\left(1 - \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}\right) + \log\left(1 - \frac{\binom{K}{k_0}\binom{N-K}{n_0-k_0}}{\binom{N}{n_0}}\right)$$

Page 55: [88] Comment [39] William Stafford Noble 3/17/16 10:50 PM

[50] J. Edmonds. Matroids, submodular functions, and certain polyhedra. Combinatorial Structures and Their
Applications, pages 69–87, 1970.
´

[51] L. Lovasz. Submodular functions and convexity. In M. Grotchel A. Bachem and B. Korte, editors, Mathemat-
ical Programming – The State of the Art, pages 235–257. Springer-Verlag, 1983.
[52] A. Schrijver. Combinatorial Optimization. Springer, 2004.
[53] H. Narayanan. Submodular functions and electrical networks. Annals of Discrete Mathematics, 54, 1997.
´

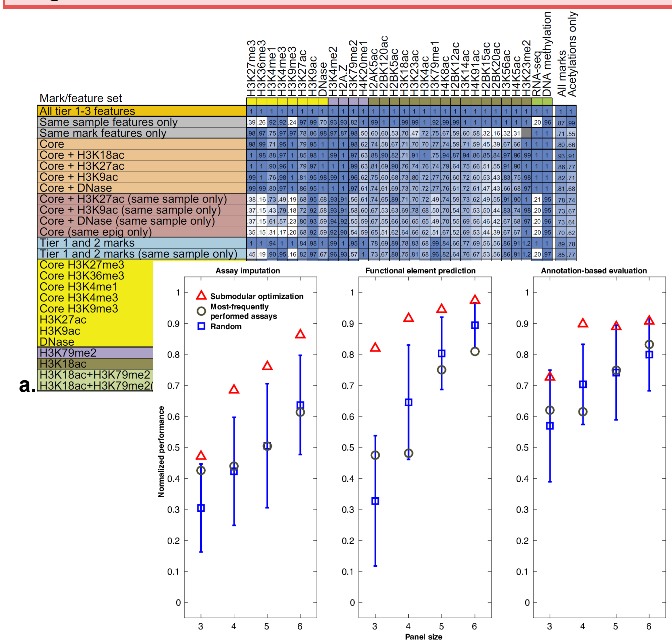
[54] G. Cornunejols, G. L. Nemhauser, and L. A. Wolsey. The uncapacitated facility location problem. In P.B.
Mirchandani and R.L. Franci, editors, Discrete Location Theory, chapter 3. Wiley/Interscience, New York,
1990.
[55] M. Narasimhan and J. Bilmes. A submodular-supermodular procedure with applications to discriminative
structure learning. In Uncertainty in Artificial Intelligence (UAI), Edinburgh, Scotland, July 2005. Morgan
Kaufmann Publishers.
[56] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian

processes: Theory,
efficient algorithms and empirical studies. The Journal of Machine Learning Research,
9:235–284, 2008.
[57] Y. Liu, K. Wei, K. Kirchhoff, Y. Song, and J. Bilmes. Submodular feature selection
for high-dimensional
acoustic score spaces. In Acoustics, Speech and Signal Processing (ICASSP), 2013
IEEE International
Conference on, pages 7184–7188. IEEE, 2013.
[58] Kai Wei, Yuzong Liu, Katrin Kirchhoff, Christopher Bartels, and Jeff Bilmes.
Submodular subset selection
for large-scale speech training data. In Acoustics, Speech and Signal Processing
(ICASSP), 2014 IEEE
International Conference on, pages 3311–3315. IEEE, 2014.

Page 55: [89] Comment [40] William Stafford Noble 3/17/16 10:49 PM

Here are these cites: [57] Y. Liu, K. Wei, K. Kirchhoff, Y. Song, and J. Bilmes.
Submodular feature selection for high-dimensional
acoustic score spaces. In Acoustics, Speech and Signal Processing (ICASSP), 2013
IEEE International
Conference on, pages 7184–7188. IEEE, 2013.
[62] B. Mirzasoleiman, A. Karbasi, R. Sarkar, and A. Krause. Distributed submodular
maximization: Identifying
representative elements in massive data. In Advances in Neural Information Processing
Systems, 2013.
[63] R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized
information maximization. In
Advances in Neural Information Processing Systems, 2010.
[64] M. Libbrecht, M. M. Hoffman, J. A. Bilmes, and W. S. Noble. Entropic graph-based
posterior regularization.
In Proceedings of the International Conference on Machine Learning, Lille, France,
2015.

Page 55: [90] Deleted Author 3/20/16 4:55 PM



Mark/feature set	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3	H3K27ac	H3K9ac	DNase	H3K4me2	H2A.Z	H3K79me2	H4K20me1	H2AK5ac	H2BK120ac	H2BK5ac	H3K18ac	H3K23ac	H3K4ac	H3K79me1	H4K8ac	H2BK12ac	H3K14ac	H4K91ac	H2BK15ac	H2BK20ac	H3K56ac	H4K5ac	H3K23me2	RNA-seq	DNA methylation	All marks	Acetylations only
All tier 1-3 features	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Same sample features only	.39	.26	.92	.92	.24	.97	.99	.70	.93	.93	.82	1	.99	1	1	.99	.99	1	.92	.99	.99	1	1	1	1	1	1	1	.20	.96	.87	.99
Same mark features only	.98	.97	.75	.97	.97	.78	.86	.98	.97	.87	.98	.50	.60	.60	.53	.70	.47	.72	.75	.67	.59	.60	.58	.32	.16	.32	.31		1	1	.71	.55
Core	.98	.99	.71	.95	1	.79	.95	1	1	1	.98	.62	.74	.58	.67	.71	.70	.70	.77	.74	.59	.71	.59	.45	.39	.67	.66	1	1	1	.80	.66
Core + H3K18ac	1	.98	.88	.97	1	.85	.98	1	.99	1	.97	.63	.88	.90	.82	.71	.91	1	.75	.94	.87	.94	.86	.85	.84	.97	.96	.99	1	1	.93	.91
Core + H3K27ac	1	1	.90	.96	1	.79	.97	1	1	1	.99	.63	.81	.69	.90	.76	.74	.76	.74	.79	.64	.75	.66	.51	.55	.91	.90	1	1	1	.86	.77
Core + H3K9ac	.99	1	.76	.98	1	.81	.95	.98	.99	1	.96	.62	.75	.60	.68	.73	.80	.72	.77	.76	.60	.72	.60	.53	.43	.83	.75	.98	1	1	.82	.71
Core + DNase	.99	.99	.80	.97	1	.86	.95	1	1	1	.97	.61	.74	.61	.69	.73	.70	.72	.77	.76	.61	.72	.61	.47	.43	.66	.68	.97	1	1	.81	.68
Core + H3K27ac (same sample only)	.38	.16	.73	.49	.19	.68	.95	.68	.93	.91	.56	.61	.74	.65	.89	.71	.70	.72	.49	.74	.59	.73	.62	.49	.55	.91	.90	1	.21	.95	.78	.74
Core + H3K9ac (same sample only)	.37	.15	.43	.79	.18	.72	.92	.58	.93	.91	.58	.60	.67	.53	.63	.69	.78	.68	.50	.70	.54	.70	.53	.50	.44	.83	.74	.98	.20	.95	.73	.67
Core + DNase (same sample only)	.37	.15	.61	.57	.23	.80	.93	.59	.94	.92	.55	.59	.67	.55	.66	.66	.65	.65	.49	.70	.55	.69	.56	.46	.42	.67	.68	.97	.20	.96	.73	.64
Core (same epig only)	.35	.15	.31	.17	.20	.68	.92	.59	.92	.90	.54	.59	.65	.51	.62	.64	.66	.62	.51	.67	.52	.69	.53	.44	.39	.67	.67	1	.20	.95	.70	.62
Tier 1 and 2 marks	1	1	.94	1	1	.84	.98	1	.99	1	.95	1	.78	.69	.89	.78	.83	.68	.99	.84	.66	.77	.69	.56	.56	.86	.91	1.2	1	1	.89	.78
Tier 1 and 2 marks (same sample only)	.45	.19	.90	.95	.16	.82	.97	.67	.96	.93	.57	1	.73	.67	.88	.75	.81	.68	.96	.82	.64	.76	.67	.53	.56	.86	.91	1.2	.20	.97	.85	.77

Page 65: [91] Deleted	Author	3/20/16 4:55 PM

We will adopt the current chairing structure of the AWG calls (where the leaders are appointed by the NHGRI).

Integration of production efforts**.** We will

Page 67: [92] Deleted	Author	3/20/16 4:55 PM

Old Section 3:

,,,,,keeppar The wealth of data from the ENCODE Project will lead to biological insights only if well-defined analyses are performed with appropriate computational methods in a reasoned priority. The large number of possible analyses could lead to confusion or even stagnation if not prioritized wisely. Thus, the design of the DAC has placed a strong emphasis on posing questions that are potentially the most informative and prioritizing the work flow. The AWG is responsible for selecting the questions to be addressed and solved. This will be done in consultation with the DAC, which will provide feedback on feasibility, resource requirements, time-lines, etc. The DCC and data producers also play important roles in selecting the questions. The AWG includes other members, such as the production PIs, bioinformatics members of the production groups, members of the U01 groups interested in participating in integrative analysis, and other informatics groups loosely associated with the Consortium. The DAC will work on analyses formulated by the AWG, coordinating closely with the DCC and data producers, and to some extent with the computational groups funded by the U01s. The types and details of the analyses will be determined during the course of the project, and we provide example analyses throughout this proposal.

,,,nokeep The interactions among AWG, DAC, DCC, production groups, and interested U01-funded analysis groups are best explained in the context of the entire work-flow for DAC, from experimental data through to analysis results. Much of the initial specific analysis of each dataset will occur by the appropriate production group, e.g., peak finding in ChIP-seq data and transcript identification in RNA-seq data. The DAC aims to establish specific contacts with the bioinformaticians in each production group and the DCC, in order to keep up-to-date with the nuances of the specific data types which can benefit integrative analysis. Furthermore, the DAC can assist the primary analysis in several ways: controlling data quality, maintaining data format consistency, and providing analysis pipelines.

,,,,,keeppar Moving beyond the primary data processing, the main source of work will be biological questions or analysis directions posed by the AWG. The DAC leadership committee (Profs. Weng, Kellis, and Gerstein) will assign them to a group for investigation and determination of whether there are existing methods that have both appropriate statistics and are scalable to genome-wide techniques. When an existing method is appropriate, the analysis will be run. Depending on the type of the analysis, the results will be reported either as data integrated into the genome browser and/or a new pipelined approach provided for regular use by all users, in particular other members of the ENCODE Consortium.

For some questions, new statistical methods are required as some existing methods do not scale for genome-wide datasets. We may be able to draw upon the strengths of some U01 groups. An example is Prof. Peter Bickel or Prof. David Gifford. In these cases we will invite these U01 groups to perform or participate in the analysis. For other questions, we will create a small group of DAC members to focus on the problem, in many cases starting with developing new statistical methods, followed by algorithmic/engineering analysis for genome-wide scalability. The statistical methods are likely to be novel applications of the broad collection of statistical and machine-learning tool kits, but the precise outline of what models to build and test requires a complex interaction between statisticians, bioinformaticians, and biologists. Once a new method has been successfully created and scales well, this method can then be appropriately pipelined. Some problems will require us to recruit expertise outside the Consortium. We have established a procedure to invite outside investigators to join the AWG and fund some of their efforts through the DAC (see the PI Leadership Plan).

We expect to quickly have more tasks than resources, in particular because we expect to have a steady flow of tasks to change existing ad hoc analysis schemes into methods that can either be pipelined or provided as end-user tools. As the number of questions posed by the AWG mounts, they will need prioritization. We will schedule our work in a completely transparent manner, with any priority disputes being resolved, and with the AWG having the final say on priorities. Figure XXX outlines the management of the prioritized list of tasks being performed by the DAC.

,,,,,keeppar As in the current phase of ENCODE, we will organize a weekly conference call for the entire AWG. The call at the beginning of each month will be open to any member of the Consortium and used to report progress from the previous month and discuss the prioritization of the active tasks. DAC members will provide estimates on the relative “cost” of each task, and the optimal groups to handle each task. New tasks are generated by the AWG and by DAC members suggesting new pipelining or engineering approaches to make existing ad hoc methods more robust. Although the DAC members can suggest tasks and obviously must be involved in the assessment of matching tasks to groups, we will strictly enforce the prioritization by the AWG to ensure that the DAC does the analysis the Consortium needs.

The DAC will ensure openness by inclusiveness in its working practices and a formal prioritization process from the AWG. To ensure inclusiveness, the DAC will conduct all meetings in formats that allow any ENCODE Consortium member to join and participate as equal members in the analysis. The progress of analysis will be posted on the ENCODE Wiki, again allowing complete access by other Consortium members. The progress of DAC analysis tasks will be reviewed and future priorities set in an open forum with the AWG and related ENCODE Consortium members, under the working principle that the AWG is the final arbiter of any priority dispute. DAC members will be available, given practicalities, for any AWG-proposed meeting. AWG phone calls will be chaired by the three members of the DAC leadership committee, with the three members changing their roles between incoming, outgoing, and incumbent chairs in a weekly rotation. This structure is modelled on the existing ENCODE PI calls and has proven to work effectively.

The members and leadership of the proposed DAC recognize that the proposed work flow and management is quite different than the usual mode of working on investigator-initiated projects. The DAC will be responsive to a wide range of inputs, with problems defined and prioritized by the AWG, close coordination with the DCC, and inclusion of other members of the ENCODE Consortium and others with needed expertise. We will organize annual data analysis workshops to bring together researchers from both inside and outside of the Consortium including the data producers, experts in data analysis, and experts in the biology of different classes of functional elements for face-to-face interactions to promote better data analysis. We expect that these interactions will be collegial and that the expertise and experience of the DAC members will be weighed favorably as the AWG establishes priorities. Indeed, this community-based approach to problem solving will be exciting and lead to insights that may not be obtained in studies by single investigators. Below we summarize the intended interactions between the DAC and various entities in and beyond the ENCODE Consortium:

AWG: DAC members are active participants of the AWG, and the goal of the DAC is to perform and facilitate the analyses defined by the AWG. The DAC will assist the AWG in defining and prioritizing the tasks, performing the tasks in the most efficient and thorough manner, reporting analysis progress, and disseminating analysis results.

Production groups and U01 groups: A challenge in the AWG will be to coordinate between the diversity of groups participating in analysis, and to ensure not only that each planned analysis is successfully completed, but also that there is minimal replication of effort between different groups. Achieving these goals will require constant communication between the different members of the AWG. DAC members will work closely with each production group and the U01 analysis groups to coordinate analysis goals, plan deliverables, establish milestones, and ensure that different groups build on each other's results. In particular, we will work with the PIs of the production groups who are ideally situated for understanding the subtleties of their datasets to establish optimal protocols, sequencing depth, and number of biological and technical replicates, and agree upon uniform processing pipelines based on the statistical expertise of the DAC and the AWG and the particularities of each data type. We will also work with the PIs of the production groups to establish goals for integrative analysis, and the specific integration plan for each type of data, interpret the analysis results, and adjust analysis goals based on their biological interpretation.

DCC: The DAC and DCC components of the EDCAC will work together on several issues, with five examples listed here. (1) Perform uniform primary processing of datasets and establish a single processing pipeline to be run by the DCC component on all datasets; (2) Establish uniform naming schemes for all datasets, a transparent directory structure, and ways to access the datasets; (3) Determine data processing priorities for data freezes and data quality standards for accepting and posting individual datasets; (4) Build a common computing platform for running common tools on all datasets; and (5) Disseminate the analysis results of the DAC.

NHGRI: On a regular basis, the DAC will update the ENCODE Consortium and the NHGRI on the progress towards completion of the comprehensive catalog of functional elements. We will also provide quarterly reports that will allow assessment of progress towards achieving DAC goals.

Additional informatics groups outside the Consortium: We will identify areas of expertise not represented in the AWG, and will invite outside investigators to join AWG analyses when necessary. We have specifically budgeted funding for the DAC to support some of these efforts. Candidates will be discussed among the DAC investigators and the AWG head, with the final decision made by the DAC leadership committee in consultation with the NHGRI.

[[cut this b/c it's in 1.7]] Other large consortia and outside production groups with complementary datasets: In addition to the ENCODE project, several large consortia are involved in systematic data generation activities involving the human genome, resulting in a wealth of functional information that would be of great value to ENCODE integrative analyses. We will work with the scientific leadership and analysis groups of each consortium to ensure that our analysis plans are synergistic, and to ensure mutual understanding of data use policies, embargo dates, and coordination in our published analyses. In addition to consortia, we will work with individual labs that systematically generate large-scale datasets of significant value to the ENCODE Project, based on the priorities set forth by the AWG, the ENCODE PIs and co-PIs, and the NHGRI.

3.2 Provide shared computational guidelines and infrastructure for data processing, common analysis tasks, and data exchange

To successfully accomplish the integrative analyses required in a consortium as diverse and complex as ENCODE, great care must be taken to ensure that in addition to uniform data production standards, uniform data processing standards are established and followed during each step of the analysis, in consultation with the AWG.

We have worked closely with the AWG and DCC to ensure that common analysis tasks are standardized. Such tasks include processing of RNA-seq datasets (short-read sequencing of mRNA), peak calling in ChIP-seq datasets (chromatin immunoprecipitation followed by deep sequencing), DNAme, and the identification of sequence motifs and transcription factor (TF) binding sites from TF ChIP-seq data. We will develop and evaluate different analysis methods for such frequently performed tasks, provide sound statistics for selecting among them, and work with the AWG to ensure uniformity in subsequent processing of each data type using the selected methods. The vast majority of ENCODE data is based on deep sequencing, a technology that only became widely practiced in the past few years. This presents potential biases and errors not yet completely understood, and thus we discuss plans for assessing the quality of ENCODE data. Lastly, we will work closely with the DCC to facilitate data import, access, and uniformity between ENCODE and the larger community. We will ensure that relevant public datasets are available in a common repository and in uniform formats to ENCODE members. We will also ensure that all analysis results by ENCODE members are shared with the larger community, in accordance with the software guideline.

Below, we describe two examples: RNA-seq data for transcript annotation, and ChIP-seq data for TF binding and chromatin marks.

References for DNA methyaltion section

1. Hansen, K.D., Langmead, B. & Irizarry, R.A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology* 13, R83 (2012).
2. Wu, H. et al. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic acids research* 43, e141 (2015).
3. Sun, D. et al. MOABS: model based analysis of bisulfite sequencing data. *Genome biology* 15, R38 (2014).
4. Lee, W. & Morris, J.S. Identification of Differentially Methylated Loci Using Wavelet-Based Functional Mixed Models. *Bioinformatics* (2015).
5. Dolzhenko, E. & Smith, A.D. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole genome bisulfite sequencing experiments. *BMC Bioinformatics* 15, 215 (2014).
6. Hebestreit, K., Dugas, M. & Klein, H.U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* 29, 1647-1653 (2013).
7. Saito, Y., Tsuji, J. & Mituyama, T. Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions. *Nucleic acids research* 42, e45 (2014).
8. Akalin, A. et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology* 13, R87 (2012).
9. Park, Y., Figueroa, M.E., Rozek, L.S. & Sartor, M.A. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* 30, 2414-2422 (2014).
10. Marx, V. Genetics: profiling DNA methylation and beyond. *Nature methods* 13, 119-122 (2016).
11. Ziller, M.J., Hansen, K.D., Meissner, A. & Aryee, M.J. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature methods* 12, 230-232, 231 p following 232 (2015).

We will keep close communication with the DCC to ensure timely and consistent processing of raw data through a uniform pipeline. The output of the pipeline will be cleansed and normalized, ready for downstream analyses. We will assemble a subgroup of DAC analysts responsible for various aspects of uniform data processing and the data transform that is necessary prior to the uniform data processing. In particular, the subgroup will include analysts who have been part of the uniform data processing during the current DAC and will be part of the next DAC. During the EDCAC kickoff meeting, these DAC analysts will meet with the appropriate DCC members to evaluate the status of uniform data processing pipelines. During the first three months of the project, all existing pipelines will be established and applied to all available datasets. This subgroup will be the point of contact with DCC on this issue. Whenever an analysis method has matured to the stage of a pipeline, it will be turned over to this subgroup to implement and apply to all existing data.

,,,,keeppar We will work closely with the AWG and DCC to ensure that all analysis results, similar to primary data, are available in uniform formats from all groups, and additionally that all metadata are stored, including versions of programs used to generate the data, which pipeline version was used, and that all tools and data processing methods remain available for the duration of the project. For each type of functional element, we will establish a unique and standardized representation, including for enhancers, promoters, transcripts, alternative splice forms, transcription start/end sites, genome segmentation etc. We will also make sure that all analyses done in parallel in human, mouse, fly, and worm use the same formats, parameters, and pipelines whenever possible, and that these are clearly documented to facilitate cross-species comparisons. To achieve this, we will establish a single coordinator within the DAC for each type of analysis who will be responsible for ensuring reproducibility and consistency within the Consortium.

All data exchange will also be stored at the DCC servers, and all results of validated analyses will be made public upon validation, following the ENCODE common Data Release Policy. Care will be taken to anticipate the versioning of ENCODE data generated for all three organisms between different genome builds. As the project progresses the data products generated earlier in the project will get updated to the most recent genome builds for each organism in a controlled fashion. This will either be done via simple liftOver or via uniform reprocessing. We also anticipate data freezes (approximately every six months) where the collective set of all primary processed data products generated up until that stage by the entire Consortium are frozen for more detailed integrative downstream analyses.

We will also store at the DCC all public datasets which have been reformatted for ENCODE use by members of the DAC or AWG to minimize duplication of efforts, and ensure consistency between different groups. The DCC will also serve as our intermediate repository for all datasets coming from the epigenome project, the 1000 Genomes Project, TCGA, GTEx, Brainspan, GEO, SRA, FlyBase, and WormBase. These clear guidelines and standards may appear cumbersome at first, but they will be invaluable in preparing the datasets for integration, and preparing the integrative analysis for publication. Lastly, all data shared with the larger community will be shared through the DCC servers and browsers, and through FlyBase, WormBase, and general genome browsers (such as NCBI, ENSEMBL, and NCBI) whenever possible.

,,,,,keeppar One of the main roles of the DAC is to facilitate the analysis and writing of integrative Consortium papers, specifically the paper that describes the Encyclopedia of human and mouse. In addition, the DAC will assist with other Consortium reports such as data and analysis standards documents. The leadership of these papers and documents naturally springs from the leadership of the AWG and involves the interplay of many different scientists, and the DAC is at the service of this leadership by The DAC will performing various integrative analyses and providing a technological infrastructure.

,,,,,keep In writing these papers the data from multiple techniques and approaches must be combined in a standardized fashion in order to maximize their utility. The readers of the papers should also be able to see how large-scale genomic datasets have biological and medical utility. Thus, it is essential for the papers to provide a clear linkage between the prose describing the results and the actual data and analyses done. Moreover, clearly connecting the genomics data in a particular freeze to the literature has many scholarly advantages in terms of time stamping, attribution, and future citation[117,118].

Aims 1-3 discussed the necessary calculations to enable these integrative analysis papers, and ,,,,,keeppar,keep here we focus on the infrastructure that the DAC will provide to enable these papers to be written. There are two aspects of this infrastructure: social and technological. In terms of social infrastructure, the DAC will organize and moderate phone calls as well as setting up and moderating meetings targeting particular analysis papers. For the conference calls and meetings we will handle creation of precise agendas and the recording of detailed minutes and action items.

,,,,,keeppar There are a number of things to bear in mind regarding setting up a technological infrastructure for enabling collaborative papers. First of all, there is a vast and rapidly evolving industry in developing social media and computational infrastructure for collaboration, including companies such as Google, Facebook, and Twitter. In this regard, the DAC will provide a gateway for the entire Consortium to make use of such computational tools and services. Second, many genomics consortia papers are extremely complex from the perspective of paper writing, often involving many hundreds of authors and tens if not hundreds of main figures and tables as well as supplementary exhibits. Simply keeping track of all figures and supplements is a non-trivial task.

The DAC will facilitate the use of appropriate technology to help with this. Currently most of the genomics papers written by the ENCODE Consortium make use of mailing lists and conference calls for many of these interactions. To handle manuscript preparation addition they make use of resources such as Wikis, Google Docs, and reference managers such as Endnote, Bookends, and Papers. We will continue to use these types of infrastructures, and the DAC will provide expertise and resources in these areas for future Consortium publications. However we will also be aware of newly evolving tools that are currently being developed, including improved solutions for tracking references and figures (e.g., Mendeley and BibTeX). Also, online file sharing services such as Dropbox and SugarSync can be utilized for distribution of files and figures; we also intend to investigate the use of fully-fledged open-source content management systems such as Drupal or Joomla. It is important to keep in mind that while a lot of these technologies might be appealing from a purely computational perspective, they might not be of as much use practically when employed by the various members of the Consortium, and thus we will have to keep track of their usefulness in terms of actual paper writing.

The writing of Consortium standards documents that focus on protocols and methodologies for data analyses and experimental procedures will proceed in a similar way as the writing of integrative Consortium papers. One difference, however, is that the role of the DAC for creating standards documents is sometimes to produce standardized datasets and standardized analyses that illustrate standard practices and provide a better third-party working-knowledge of the standard beyond that from individual production laboratories.

,,,,,keeppar A third aspect of the DAC in writing large Consortium papers and standards documents is carefully connecting the underlying data to the prose. One can conceptualize a big genomics "roll out" (publishing a number of genomics papers commenting on a single underlying data freeze) as a hierarchical information structure, designed to present the Consortium's genomic data and results in an organized fashion. The "main" integrative paper sits at the top, synthesizing everything broadly, which provides pointers to other high-profile companion papers and further, more detailed companions focusing on specific sub-analyses. Each of these individual papers, in turn, often refers to a huge amount of supplementary calculations and datasets. Some of these are in formal paper supplements while others are on project Web sites. Moreover, the datasets most referred to in the papers are usually not the actual raw data but subsidiary analysis products that summarize the data (e.g., peak call lists, transcript structures, and segmentations). At the bottom of the hierarchy is the actual underlying raw data (usually sequencing reads), stored in central repositories (such as the short-read archive). Given that the raw data files and, to some degree the analysis summaries, are usually huge and unwieldy, it makes most sense to approach the information in a particular freeze from the top down, starting with the papers (assuming everything is linked together correctly). In the future one may see some machine readable, “structured” versions of the text of the paper (i.e., the structured digital abstract and structured digital table[119,120], which allow authors to make this hierarchy and its linkages even more explicit.

,,,,,keeppar There is an even more detailed micro-structure to the hierarchy: All of the data tables and figures in each paper rely on a considerable chain of small specific analysis results as well as programming scripts. These, in turn, connect to specific versions of the overall analysis summaries put out on the project website. One of the roles of genome analysts is to link these all together and make sure that it is clear which version of a particular analysis result goes with which version of the underlying data and how these in turn link with a specific paper figure. This can often be done through making available the small subsidiary analysis files underlying each exhibit (such as networks connecting particular genomic entities or Excel files or R data frames) in an organized fashion. (See modencode.org/publications/integrative_worm_2010 and modencode.org/publications/integrative_fly_2010 for examples.) Often in the rush to publish a large manuscript these smaller files are neglected but they are essential for truly reproducible research[121]. The DAC will strive to make all of these available through the DAC and DCC. It will also push out larger analysis product datasets such as peak calls and segmentations regularly as part of data freezes and we will version these along with the underlying data from the production groups.

[[From old DAC4 grant (was in Aim 1 & moved by MG on 12-Mar)]]

We will create and analyze a meta-network composed of the TF regulatory network and protein-protein interactions integrated with the miRNA data. To integrate the miRNA data, the TargetScan software will be used to obtain the miRNA-TF edges in this network[98] by assessing complementarity of the miRNA seed region coupled with conservation information. This expanded, integrated network will be analysed using the tools and methodologies described below, allowing us to look for new types of motifs, regulatory patterns, and relationships that would likely remain undetected in individual analyses of the separate networks.

The networks will be analyzed using several methods to calculate key statistics. One such method is 'tYNA'[99], a web system developed to compare and mine multiple networks in order to identify cliques and motifs, as well as calculating statistics on a network. Statistics such as ‘eccentricity' and ‘betweeness' can help explain the connectivity and behavior of nodes in a network[100]. Eccentricity is defined as the maximum shortest path from a node to any other node in the network; this describes how a node interacts with all the other nodes it is connected to, i.e., a node with a small eccentricity is tightly connected to all nodes that it interacts with- including nodes to which it is not directly connected. Betweeness is defined as the number of shortest paths in the network that pass through a given node; this is a measure of a node’s centrality and is related to how involved a node is the communications between all other nodes in a network. These and other statistics can help in the definition and understanding of a particular network.

Connectivity statistics, in particular the difference between out- (O) and in- (I) degree, elucidate the direction of information flow in the network and can reveal hierarchical organization. In previous work, we employed a simple simulated annealing procedure to arrange TFs into discrete levels that maximize the number of edges propagating down from higher to lower levels. To complement these discrete level assignments we defined a continuous parameter h=(O-I)/(O+I), which can be interpreted as the height of a TF within the hierarchical structure. We plan to apply the same procedures to the transcriptional regulatory network in human, mouse, fly, and worm. Of particular interest is a comparison of the distribution of TFs between the different levels to determine, for example, the extent to which middle-level regulator nodes are conserved across different networks. Furthermore, it is important to evaluate the robustness of our results, as there exist several methods of constructing network hierarchies (e.g. breadth-first search) and we intend to assess the impact of various network construction methods on these results.

We have previously exploited the model of hierarchical organization by examining the degree of collaboration among different regulators[101-103]. This is essentially the ratio of the number of genes co-regulated by two regulators (from the same or different levels) to the union of their target genes summed over all such pairs of regulators from the two levels[102,103]. We found that in E. coli, yeast, and human the highest degree of collaboration is between regulators from the middle level, which is analogous to a corporate setting in which middle managers play an important organizational role. We plan to investigate the same arrangement using more comprehensive co-regulatory networks from human, mouse, worm, and fly. In addition to co-regulation we also studied the overlap between modules in terms of their position within the hierarchy. We defined a module as all accessible nodes downstream of a top regulator and investigated the overlap (share of regulators) between modules. We found that the modules in E. coli are more independent compared to those within the call-graph of the Linux kernel[104]. We intend to examine the module-overlap in human, mouse, worm and fly, and other eukaryotes compared to our previous observations in E. coli. We speculate that these networks of more complex eukaryotes will be more similar to the Linux call-graph.

Building upon the expertise in dynamic Bayesian network (DBN) in the Noble lab in close collaboration with the Weng lab, we propose to develop computational methods to identify TF binding sites in a generalized way. We will focus in particular on the problem of predicting the results of a ChIP-seq assay for binding of a sequence-specific transcription factor in formulating the problem as follows. Other ENCODE data types (e.g., chromatin accessibility measured by DNase-seq) can be easily incorporated into our computational framework. In a particular cell line or tissue, we are given several generic chromatin architecture datasets—e.g., DNase accessibility plus a variety of informative histone modification ChIP-seq assays—as well as the results of ChIP-seq experiments for a set of "training" TFs. We are then asked to predict what would happen if we were to run a ChIP-seq experiment for a given "test" TF, knowing only its binding affinity sequence motif, as determined from a ChIP-seq assay in another cell line.

This problem is akin to what is known in the field of speech recognition as "speaker adaptation." In that context, a system trained to produce a textual translation of a spoken utterance must learn to adapt to the peculiar characteristics of a new speaker. Similarly, our model will learn to adapt to the characteristics of a new TF, using methods developed in speech recognition. In particular, the model will predict the locations of in vivo binding events based upon features of the local chromatin architecture, while taking into account that some types of TF binding exhibit different local chromatin properties than others.

We will solve this problem using a combined classifier and DBN. The classifier (a support vector machine or a deep neural network) will learn to predict the binding of a given TF based on the observed local chromatin profile. We will train one such classifier for each TF in the training set. The scores produced by these classifiers will then be combined in the context of a DBN such as the one shown in Figure 6. In this model, MotifScore is an observed variable measuring how well the sequence starting at the current position matches the known motif of the target TF represented by its PSSM. The hidden variables state and Class indicate (respectively) whether a position is inside a binding site of the target TF and the identity of the training TF whose chromatin pattern is the closest match to those of the current position. In addition, the model contains one virtual evidence track for each of the *N* training TFs. Thus, V^t_i is a positional virtual evidence variable that tells how well the chromatin pattern associated with the sequence starting at position *i* matches the patterns learned for the training TF *t*. Because every virtual evidence track corresponds to exactly one class value, we introduce a binary variable I^t_i, indicating the corresponding virtual evidence track of each class value. The conditional probability of this indicator variable, $Pr[I^t_i = 1|class_i = c]$, equals 1 only if $c = t$, and is 0 otherwise. As mentioned above, the virtual evidence, $Pr[V^t_i = 1|I^t_i]$, is produced by the TF-specific classifier. Finally, the observed binary variable missing is used as a switching parent of the virtual evidence nodes to handle positions where the chromatin profile contains missing data.

The accuracy of this predictive model can be validated in a prospective fashion within the context of ENCODE, because the Consortium is constantly producing new data. Accordingly, we will make predictions for experiments that are currently in the pipeline, and directly measure the accuracy of our predictions once the data become available. In addition to providing a valuable resource for TFs that have not yet been fully characterized experimentally and shedding light on the relationship between local chromatin architecture and TF binding, this project will enable us to prioritize future experiments. By investigating the dependence of prediction accuracy upon properties of the cell line, available local chromatin data, and properties of the TF itself, we will be able to predict which entries in the 2D experimental matrix can be easily imputed and which entries will likely provide the most value in training future imputation models.

Using protein domains to identify and evaluate orthologs: In addition to phylogenetic approaches, we will use protein domain annotations to help map orthologous proteins and evaluate ortholog assignments by other methods. We will use the Proteome Folding Pipeline[105] to sequentially analyze primary, secondary, and tertiary structures of protein sequences encoded globally in these four genomes. The analysis of potential 3D structures can allow us to recognize very distant homologs, as structural relationships can often be identified well into the “twilight zone” of sequence identity (as low as ~15-25%). Protein domains and structure will also provide testable functional hypotheses at the molecular level for directed experimental studies. We will use these results to help evaluate ortholog assignments, and compare the domain content of orthologous proteins, especially with regard to functional differences between the species.

A unique opportunity with ENCODE comes from the parallel experiments in human and mouse. We will study the relative role of pre- and post-transcriptional regulation in the two organisms, the interplay between orthologous transcription factors and orthologous chromatin marks, and the expression patterns, onset stages, targets, and target expression for orthologous factors and miRNAs. We will also study patterns of alternative splicing and alternative polyadenylation for orthologous genes in corresponding stages, especially with respect to how they affect miRNA targeting and functional protein domains. More generally, we will compare the transcription factor and miRNA inventory of the two species, the relative distributions of target gene counts for orthologous regulators, and similarly the relative distribution and diversity of regulators bound for orthologous gene targets. We will also study whether discovered modules of biologically related genes in **Aim 3.5** correspond to orthologous modules across species and for significantly conserved modules we will study the properties of gained and lost components. Using network properties discovered in **Aim 4.3** we will compare the two species in terms of determining if similar network motifs are discovered in the regulatory networks of the two species, and the types of feedback loops that are found in mixed TF/miRNA regulatory networks.

Old Aim 1 text on integrating ENCODE with other consortia [[ANS on 03/15/2016]]

> Integrate and harmonize ENCODE with other datasets from all the other consortia - normalize or 4DN may have sophisticated pictures of chromatin that we (simplify) or variant kind of data
> IHEC, Roadmap - done
> GTex - Roderic, Manolis - Normalize data and harmonize eQTLs
> PsychENCODE, BrainSpan - MG/ZW are part of this and we need to normalize the data/consitency in pipelines. Also have eQTLs
> exRNA - Gerstein
> 1000G- MG - germline variants
> TCGA/ICGC/PCAWG - somatic variants + harmonizing RNA-seq

Integration of ENCODE with Roadmap Epigenomics data and international human epigenome consortium (IHEC) [[already done, Manolis]]

The NIH Common Fund Epigenomics Roadmap Project (http://www.roadmapepigenomics.org/) and the International Human Epigenome Consortium (http://www.ihec-epigenomes.org/) have generated rich maps of histone modifications, including high-resolution maps of more than 20 modifications in a small number of cell lines, maps of a few modifications in a large number of cell types, as well as maps of DNA methylation and DNA accessibility. As each of these data types are widely used in the ENCODE Consortium, we will work with the data coordination centers and informatics groups participating in these consortia to integrate relevant datasets into ENCODE pipelines for joint integrative analyses by the ENCODE AWG. For those cases in which we seek to integrate non-ENCODE datasets into our existing pipelines, we will ensure that such data meet the same quality standards as ENCODE datasets. To optimize the compatibility with published papers by each of these consortia, we will use the processed files from the investigators as much as possible. However, when necessary, we will reprocess the primary data according to the ENCODE uniform pipelines to ensure that data quality standards are met. [2][3][4]

Members of the DAC have been highly involved in the analysis of the Roadmap Epigenomics consortium datasets, and its continuation, the international human epigenome consortium (IHEC). In particular, we have jointly re-processed all raw datasets from ENCODE and the [5][6][7][8]

***GTEx - Inter-relating ENCODE annotations with large-scale RNAseq & eQTLs [[Roderic, Manolis, normalize, harmonize eQTLs]]

Other sources of complementary, large-scale human data include the NIH GTEx (Genotype-Tissue Expression) Project (http://genome.gov/gtex), the NIMH Brainspan Project (http://brainspan.org), and the NCI Cancer Genome Atlas (TCGA) Project (http://cancergenome.nih.gov). Over 1,200 data samples from primary tissues have already been collected and analyzed during the initial stages of the GTEx project. We propose combining the large amount of population-specific lymphoblastoid genotype and mRNA expression data in ENCODE with the GTEx samples to improve the identification of expression quantitative trait loci (eQTL). In addition, compilation and comparison of DNA methylation, mRNA/miRNA expression, as well as genotype data from ENCODE, GTEx, and Brainspan, will permit a robust analysis of the expression landscape of the human brain and provide a valuable resource for other investigators given the scarcity of such large-scale, high-quality datasets.

***PsychENCODE/BrainSpan [[Gerstein, Weng, normalize, look consistency in pipelines, connect w eQTLs]]

We will compare parameters like thresholds for calling peaks in ChIP-seq data, look at discrepancies between the two projects and calibrate parameters to reduce biases due to cross-project analyses. In addition to comparing the parameters in the uniform processing pipelines, we will also directly compare the annotated genomic regions or transcripts called by these pipelines. For example, we will identify the brain cell types studied by ENCODE and other consortia, match them with the most appropriate datasets in psychENCODE, and investigate whether the corresponding pipelines in the two consortia have detected a similar set of genomic elements. While performing this comparison, we will take into account the differences in cell sources and the inherent variation among biological replicates, and focus on the regions and transcripts deemed most significant by either or both pipelines. If we identify major differences, we will investigate whether they are due to the underlying raw data, or the differences in the pipelines.

***exRNA [[Gerstein, shorten, different types]]

Moreover, Gerstein lab has considerable expertise in developing standardized pipelines and quality control metrics for RNA-seq and evaluating them in many consortia including the Extracellular RNA Communication Consortium (ERCC). The lab has developed a custom pipeline developed for the analysis of small exRNA-seq data for the Extracellular RNA Communication Consortium (ERCC). The Gerstein Lab is in charge of the DIAC of the exRNA consortium to develop the standardized RNA-seq pipelines for the analysis of exRNAs.

***1000 Genomes & Germline Variants [[Gerstein]]

We aim to integrate different non-coding annotations with the variants identified by the 1000 Genomes Consortium. In this way, the non-coding annotation of each variant will be available to all users of the 1000 Genomes data and should serve as a valuable resource for the genetics community to identify the causal variant in GWAS. It will also provide easy access to downstream analyses for all the users. For instance, it will enable the calculation of selection pressure on variants in different annotation classes. The integration of ENCODE data with 1000 Genomes variants (SNPs, small insertions and deletions, and large structural variants) will be provided in the annotated Variant Call Format (VCF; http://www.1000genomes.org/node/101) These VCF files will contain the annotation of each genomic variant, including presence in non-coding RNA, TF peaks, TF motifs, and pseudogenes. Furthermore, SNPs which exhibit allele-specific behavior will also be identified with a distinct tag in the VCF files.

***TCGA/ICGC/PCAWG - Somatic Variants & Cancer Functional Genomics Data [[Shirley, Gerstein, somatic variants & also harmonizing RNA-seq]]

Finally, the integration of the large volume of ENCODE cancer data (from both primary tissue samples and immortalized cell lines) with over 500 clinical samples in TGCA will enable analyses of somatic mutations, mRNA/miRNA expression, copy number, and DNA methylation on an unprecedented scale.

***4D Nucleome Consortium[9] - [[Noble, More elaborate chromatin annotations]]

Recently, Dr. Noble and and Dr. Jay Shendure, in collaboration with five other UW investigators, were awarded a U54 center grant as part of the NIH 4D Nucleome Consortium. With Bing Ren, Dr. Noble is co-chairing the steering committee for the 4D Nucleome Nuclear Organization and Function Interdisciplinary Consortium. Dr. Noble is thus in an excellent position to ensure coordination between the ENCODE and 4D Nucleome consortia.

We have already developed one computational method that will facilitate integration between ENCODE and 4DN data. Graph-based regularization (GBR) is a principled method for making use of Hi-C data during the semi-automated annotation of the genome 41 (); 64 (). GBR expresses a pairwise prior that encourages certain pairs of genomic loci to receive the same label in a genome annotation. We used GBR to exploit Hi-C data during genome annotation by encouraging positions that are close in 3D to occupy the same type of domain. Using this approach, we produced a model of chromatin domains in eight human cell types, thereby revealing the relationships among known domain types. Through this model, we identified clusters of tightly regulated genes expressed in only a small number of cell types, which we term “specific expression domains.” We also found that domain boundaries marked by promoters and CTCF motifs are consistent between cell types even when domain activity changes. During the next phase of the ENCODE project, we will deploy GBR in Segway, thereby making use of Hi-C data being generated both within ENCODE and by the 4D Nucleome Consortium.

We will also explore the use of Hi-C and other ENCODE data types to assist in 3D structural inference. One particularly valuable, orthogonal type of data that the 4D Nucleome Consortium will be generating is microscopy data. For example, oligopainting data can be used to visualize different types of topologically associated domains 65 (). We will use such data in combination with Hi-C data generated by ENCODE in a hierarchical extension of our existing Pastis 3D structure inference algorithm 13 (). In general, we expect within-domain contact probabilities to exhibit different scaling properties relative to genomic distance 66 (). Accordingly, a series of hyperparameters corresponding to domains of different sizes will adjust the Poisson distribution in the Pastis model, or the Poisson rates might be sampled from a prior distribution derived from the microscopy data. Additional constraints will come from ENCODE CTCF and PolII ChIP-seq data, which empirical evidence based on Hi-C suggests tend to bind to regions on the outside of the 3D structure 9 ().[10][11]

2.XXX. Data integration for defining protein-coding and non-coding genes and transcripts. [12]

To characterize coding and non-coding transcripts, we will integrate RNA-seq data, promoter-associated chromatin marks, transcription elongation-associated chromatin marks, and comparative genomics of related species.

Transcriptional evidence: Several types of data generated by ENCODE provide transcriptional evidence[56]: RNA-seq can be used to derive the structure and level of transcripts, CAGE determines the precise positions of the 5’ ends of transcripts, and RNA-PET (also called diTAGs) provides connectivity for the 5’ and 3’ ends of transcripts. Moreover, ENCODE has generated transcript data with high-density (5 bp) tiling DNA microarrays for some cell lines.

Chromatin evidence: We have found that distinct combinations of chromatin marks and chromatin accessibility are associated with promoter regions, transcribed regions, and transcription termination regions. Surprisingly, internal exon-intron boundaries are also associated with distinct chromatin marks and nucleosome positioning biases[57], suggesting that these can be used as an additional line of evidence in defining transcript boundaries, especially with respect to low-expression genes for which the transcriptional evidence may be weaker. We have recently used such marks to discover more than a thousand novel long intergenic non-coding RNAs (lincRNAs) in mouse[58] and more recently in human[24].

Comparative evidence: The Kellis lab identified evolutionary signatures that are uniquely associated with each class of functional elements[59], with protein-coding genes showing distinct patterns of codon substitution frequencies and reading-frame conservation, non-coding RNAs showing compensatory changes and silent GU-involving substitutions, miRNAs showing a distinct conservation profile (high conservation in the star and mature arm and lower conservation in loop and flanking regions), and other structural conservation properties. We used these signatures to reveal at least 30 non-coding genes in the fly genome using modENCODE transcript data and comparative genomics of 12 Drosophila species[60]. We will apply a similar approach to human and mouse ENCODE data.

Combinations of features: We will combine these transcriptional, chromatin, and comparative features to distinguish different classes of genes in a machine learning framework. We will use Support Vector

Machines to classify each transcript into coding or non-coding. We will also apply a previously developed Conditional Random Field (CRF) framework to predict coding and non-coding exons and transcripts from the combined evidence. CRFs are graphical probabilistic models similar to Hidden Markov Models (HMMs) but they allow much richer feature sets due to their discriminative training nature. Lastly, we will associate non-coding RNAs with precursors of miRNAs, piRNAs, and other classes of small RNAs using short-RNA sequencing results.

Identification of non-coding RNA genes: We will utilize an integrative machine learning approach for identifying novel non-coding RNA (ncRNA) genes[17]. The method is based on support vector classification to classify non-coding RNAs from other elements like coding sequences and UTRs. For each non-coding RNA, a set of features are computed which are then used for training. These features comprise the quantities that are chosen to maximize the discriminatory power of the machine-learning algorithm. For example, high short RNA-seq expression levels, high secondary structure stability and conservation, medium nucleotide conservation, and low amino acid conservation are the expected characteristics for the non-coding RNAs. The machine learning algorithm enables combining these features for prediction of the new ncRNAs in a formal manner. The known non-coding RNAs are utilized as the training set for the method. We will then apply the method on the remaining unannotated parts of the genome to discover novel non-coding RNA genes.

2.XXX Correlating gene expression with TF binding and histone modifications: Transcriptome monitoring by RNA-seq can provide accurate genome-wide estimates of the steady-state abundance of transcripts. Such estimates are essential to fully understand the pathways involved in RNA biogenesis. A number of genetic and epigenetic factors cooperate to determine the abundance of a specific RNA species in a particular cellular compartment: 1. signals in genomic DNA and in intermediate RNA molecules, e.g., TF binding sites, splicing regulatory sites, and polyadenylation signals; 2. the structure and status of chromatin; and 3. the abundance and concentration of the regulatory molecules—including both proteins and RNAs themselves. The relative contribution of each of these factors and their mode of cooperation, are largely unknown. The goal of the ENCODE Project is to simultaneously monitor many of these factors across multiple cell conditions and types.

Given the gene expression data from RNA-seq and the TF binding or histone modification data from ChIP-seq experiments, the Gerstein and Weng labs have investigated their relationship in a quantitative fashion[5,61]. We have previously constructed TF models and histone modification models in different species from yeast to human[18,62,63]. Our results indicate that both TF binding and histone modifications are predictive of gene expression levels in a position dependent manner, and either account for at least 50% of gene expression level variation. The histone modification model accurately predicts gene expression in a wide chromatin region from the promoter to the transcribed regions, whereas the TF model achieves high accuracy only in a narrow chromatin region around transcription start sites (TSSs) of genes (Figure 4). Our study also indicates that TF and histone modifications are highly coordinated during transcriptional regulation and a combination of TF and histone modification signals does not further improve prediction accuracy relative to using each alone.

Based on the CpG content, promoters of genes can be divided into high CpG (HCP) and low CpG (LCP). Interestingly, we find that expression levels of HCPs are easier to predict than those of LCPs[5]. More detailed analysis indicates that the relative importance of TFs and histone modifications in the models are different between the two promoter categories. These results suggest different regulatory mechanisms between HCPs and LCPs. The prediction accuracy of TF and histone modification models to some extent reflects the quality of the expression data. For example, the models achieve significantly higher accuracy for expression measured by RNA-seq than by microarray. In addition, the models can also be used to predict the expression levels of non-coding RNAs (miRNAs) with fairly high accuracy. In the future, we suggest using the TF and histone modification models as a benchmark to understand transcription regulation.

Predicting alternative splicing using chromatin modifications and TF binding. We propose to investigate how inclusion levels of alternative exons covariate with histone modifications and TF binding, which will help us understand how these factors cooperate to modulate the specific abundance of RNA splice variants in the cell. Sophisticated probabilistic models have been recently developed to successfully predict tissue specific exonic inclusion[64]. However, the splicing code delineated in such a way includes hundreds of variables, many of which are concomitant but unlikely to be mechanistically involved in splicing. Within the current phase of ENCODE the Guigo lab has used statistical models to explore the relationship between levels of histone modifications and inclusion of exons. We found that, even when controlled for gene expression, some histone marks (e.g., H3K9ac and H3K36me3) are consistently significant (albeit weak) predictors of exon inclusion, and we built a predictive model of exon inclusion by combining the signals of these marks[25,28] (Figure 5). By computing a measure of splicing completion on RNA-seq datasets obtained in different sub-cellular compartments, we found strong evidence that splicing in the human genome occurs predominantly during transcription[29], providing a molecular justification to the increasing evidence connecting chromatin with splicing[65]. We propose to further develop and apply statistical methods to the data obtained through the ENCODE Project, as well as to other publicly available data, to identify novel genetic and epigenetic factors involved in RNA processing, splicing in particular. We propose to:

1. Identify and characterize histone modifications that play a role in the regulation of alternative splicing, using correlation and modeling analysis.

2. Apply methods used to investigate the distribution of TF binding sites in promoter regions to investigate their distribution in exons and exon-intron boundaries. We will further correlate exon skipping events, as measured by RNA-seq, with differential binding of TFs.

3. Identify novel regulators of splicing by searching for co-variation between the levels of exon inclusion (or alternative splicing events in general) and gene expression. This approach will uncover protein coding genes not yet known to play a role in splicing regulation, and more interestingly, lncRNAs that may act as splicing regulators. The RNA expression levels of protein coding genes are only a proxy for protein expression—the latter is the effector of the biological function. In contrast, expression levels of lncRNAs are almost the actual physiological levels. Inferences based on co-variation between exon inclusion and abundances of lncRNAs are therefore more likely to be biologically meaningful.

4. We will investigate the relationship between RNA expression levels, protein expression levels, and protein binding activity—both to DNA and RNA—for transcription and splicing regulators, and determine how they relate to cellular abundance of the regulated RNAs.

2.XXX Associate TF/miRNA with their targets:

Given a high-quality annotation of enhancers, promoters, insulators, protein-coding genes (including all transcription factors, TFs), microRNAs, and other non-coding RNAs, we will work to piece together regulatory networks based on condition-specific TF binding, conserved TF motif instances in active promoter and enhancer regions, and conserved miRNA motif instances in 3'UTRs. We previously showed that comparative genomics of the 12 Drosophila genomes enables the high-accuracy definition of regulator targets, and that combining comparative genomics with condition-specific TF binding led to a further increase in signal[67].

Binding vs. regulation: We will use condition-specific changes in gene expression levels of TFs, miRNAs, and their targets to distinguish 'biochemical' binding from 'biological' transcriptional regulation. We previously showed that bound motifs associated with expression changes are under stronger selective pressure across related species, and show stronger functional enrichments[67]. This suggests that we can use positive or negative expression changes to further annotate our inferred transcriptional networks with activation, repression, or simply binding edges, when no effect is detectable. We have further characterized the chromatin and sequence context of each of these three classes of edges, to search for predictive features of each class that could lead to new insights on the logic of gene regulation[68].

2.XXX miRNA regulation: A complete view of mixed TF/miRNA pre- and post-transcriptional networks requires accurate promoter and enhancer annotation for miRNA precursors, and accurate tissue- and stage-specific 3'UTRs for all miRNA target genes. Kellis led the fly modENCODE integrative paper and generated TF/miRNA regulatory networks[60] (Figure 7) and we will perform similar analysis on human and mouse ENCODE data. We will use chromatin marks to annotate miRNA promoters based on genomic proximity and tissue-correlated activity. We will use the longest 3'UTR from updated gene models combined with multi-species sequence alignments based on the most up-to-date genome alignments with all sequenced species to identify evolutionarily conserved 3'UTR target sites. We will search for and re-evaluate previously-discovered 3'UTR motifs that were not associated with miRNAs to determine whether we can now identify putative trans-acting agents that may mediate regulation through these sites. We will also evaluate the potential for other non-coding genes to serve as targets for miRNA regulation. Lastly, we will study the effect of alternative polyadenylation sites on miRNA regulatory networks. We have identified many such sites within protein-coding transcripts resulting in loss of key functional miRNA binding sites, which may play key roles in altering regulatory relationships during development and proliferation. We will look for anti-correlations between miR NAs and their predicted targets, and assess whether transcripts with short vs. long 3'UTR isoforms which exclude or include conserved miRNA binding sites, differ in expression in the presence of cognate miRNAs. This may have major implications for human gene regulation as well, as the Burge and Sharp labs reported that proliferating cells, and by extension cancer cells, preferentially shift towards shorter transcript isoforms that escape miRNA regulation[70].