Adding complexity...

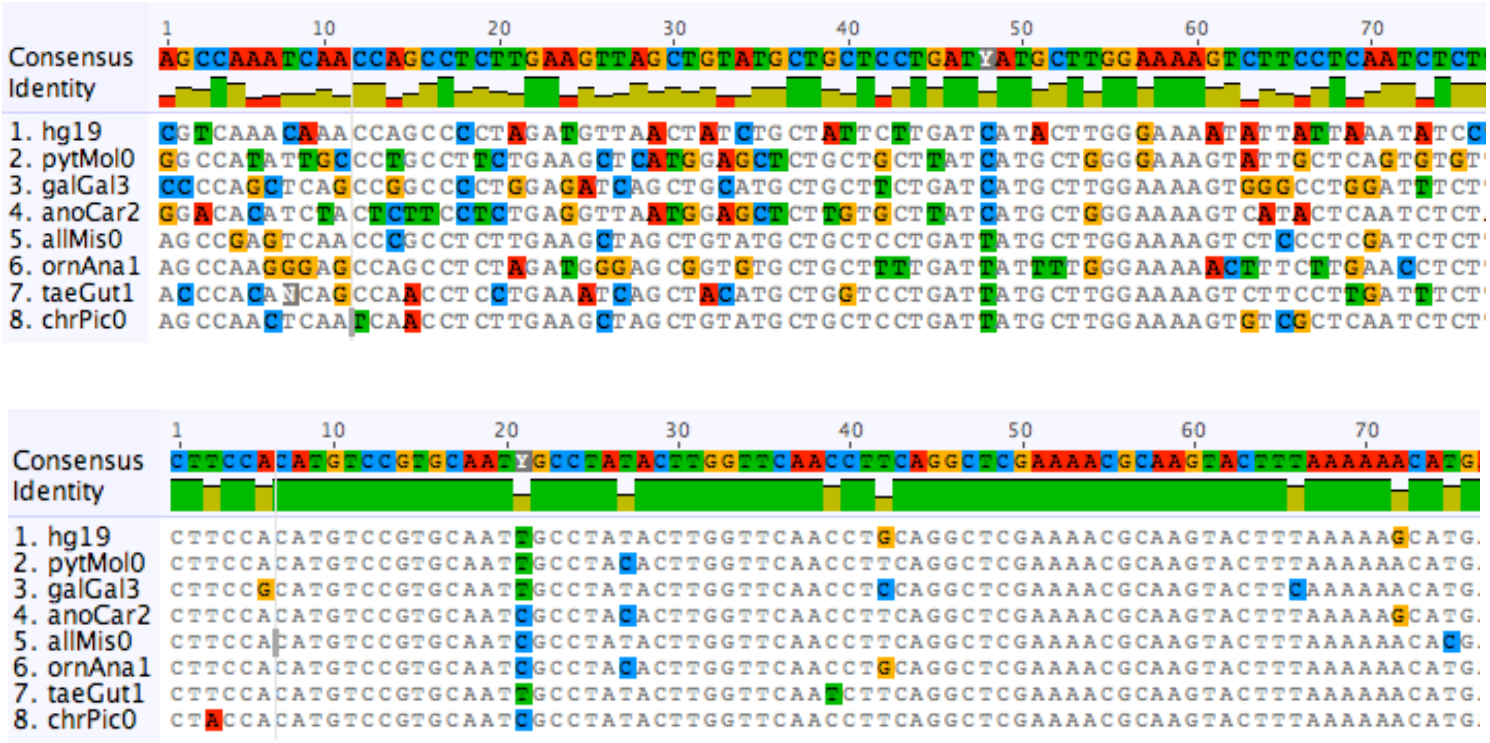
Heterogeneous substitution processes, partitioned models, and mixture models

Bob Thomson

Outline for this session

- Process heterogeneity
- Partitioned-data models
- Challenges & Difficulties
- Tutorial and Activity

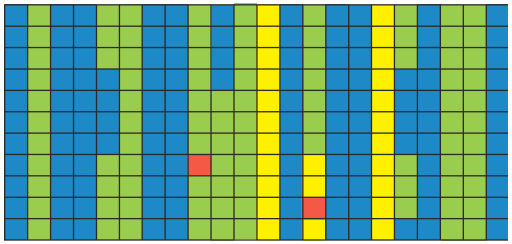
Heterogeneity

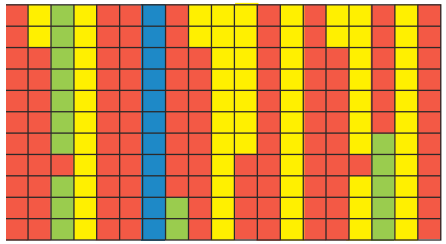


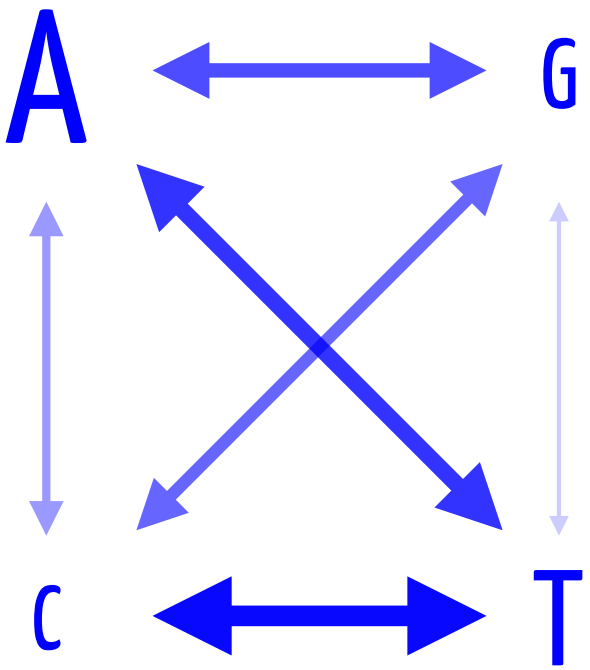
Different Genes, Different Processes

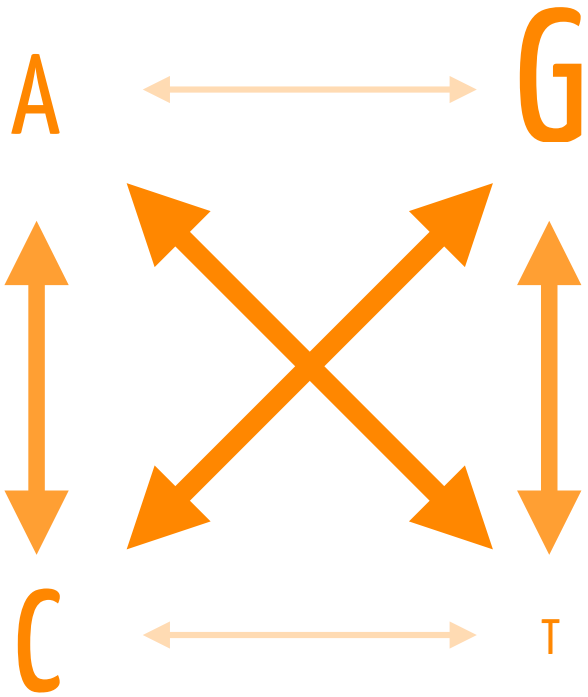
A key distinction...

- Process heterogeneity is distinct from rate heterogeneity





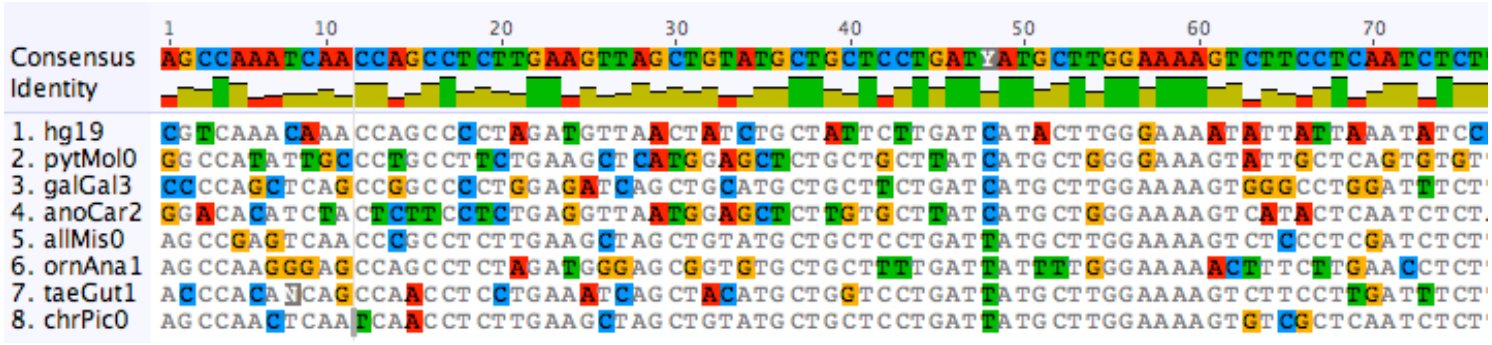


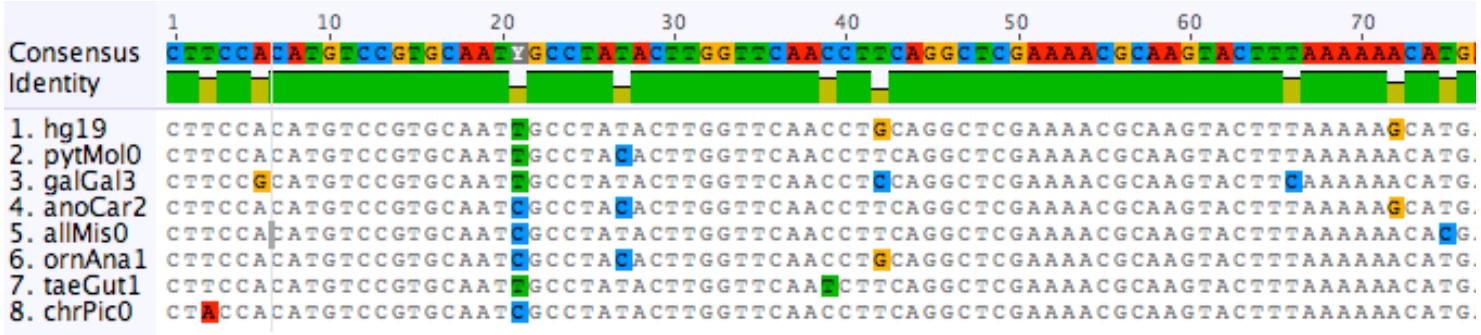


stolen from Jeremy

A key distinction…

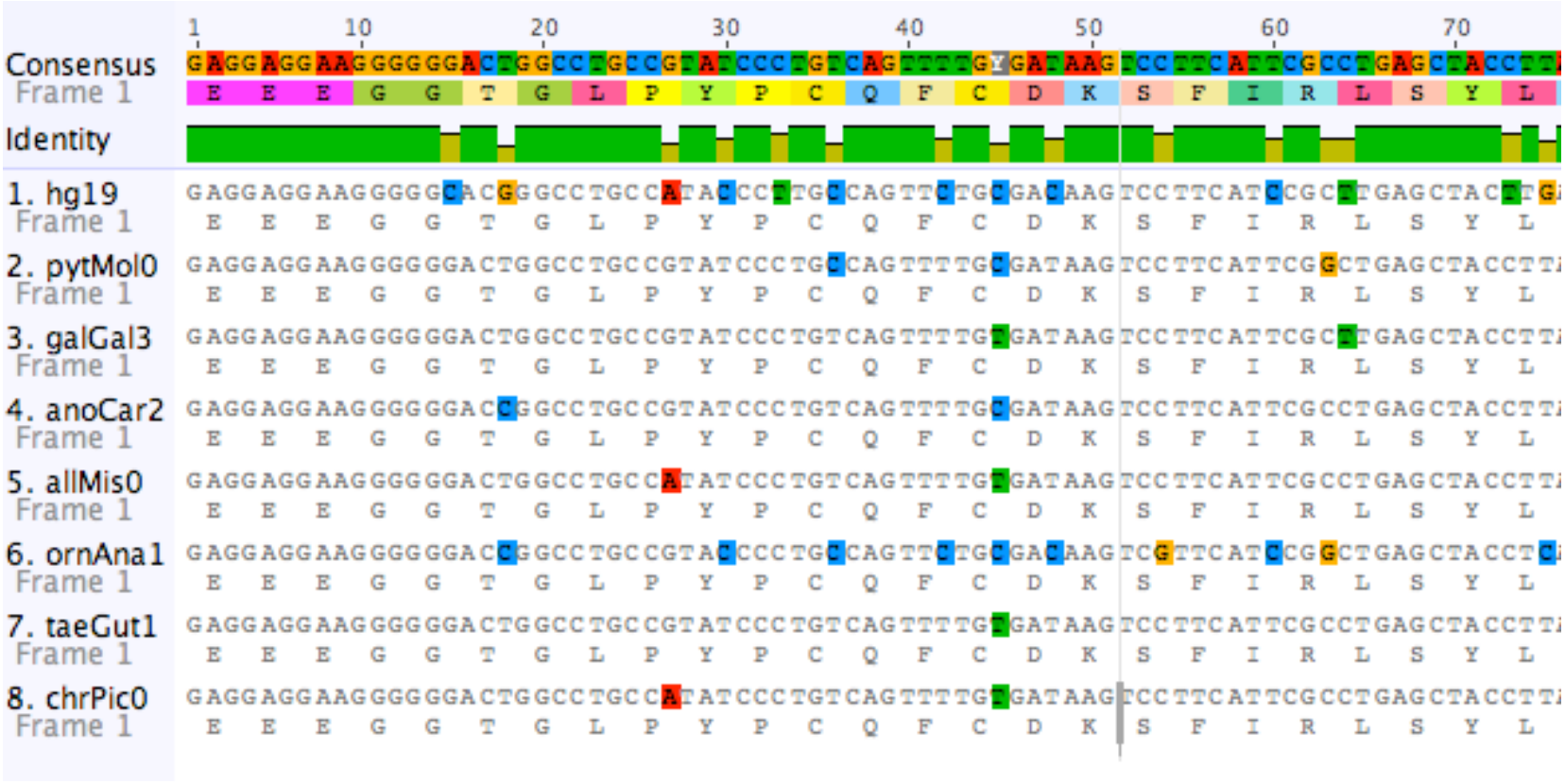
- Process heterogeneity is distinct from rate heterogeneity





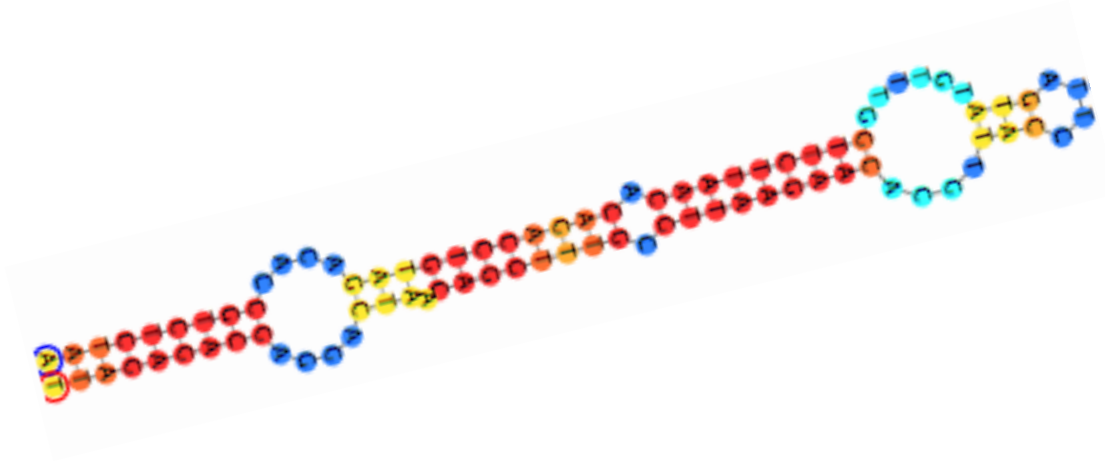
Different Genes, Different Processes

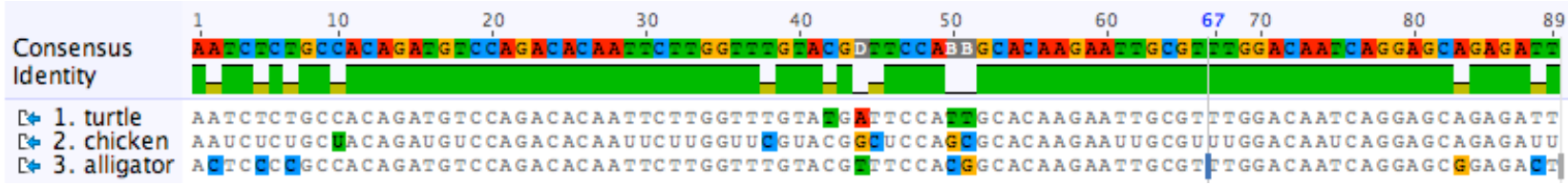
Heterogeneity



Single Gene, Different Processes

Heterogeneity





Single Gene, Different Processes

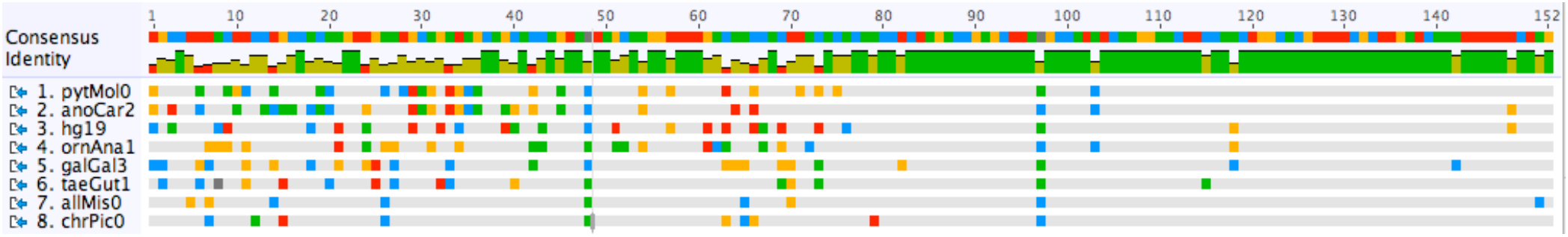
Nucleotide Substitution Models

- Recall (from yesterday) CTMC models for nucleotide substitution make several simplifying assumptions:
  - the rate of the substitution process is constant across sites
  - the nature of the substitution process is constant cross sites
  - sites are independant

For instance, the GTR:

$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$

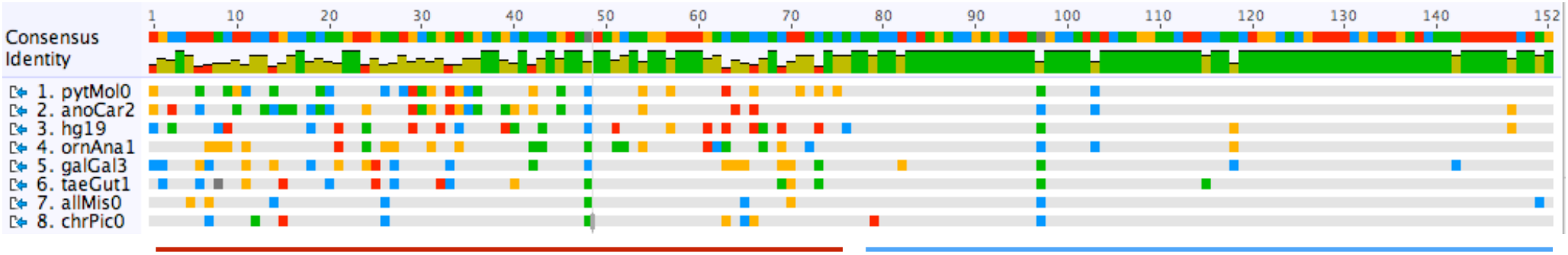
Process Heterogeneity



Class 1

Class 2

Process Heterogeneity



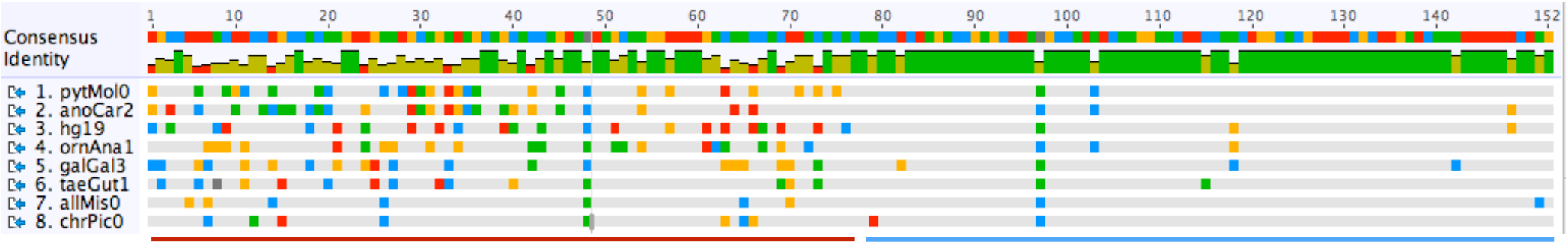
Class 1

Class 2

$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$

non-partitioned model

Accommodating Heterogeneity



Class 1

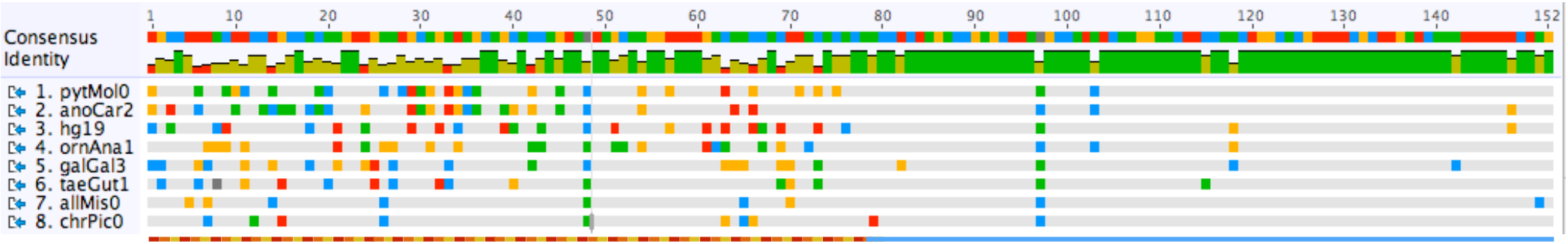
$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$

Class 2

$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$

Partitioned model

Accommodating Heterogeneity



$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$ Class 1

$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$ Class 4

$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$ Class 2

$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$ Class 3

more highly partitioned model

Accommodating Heterogeneity

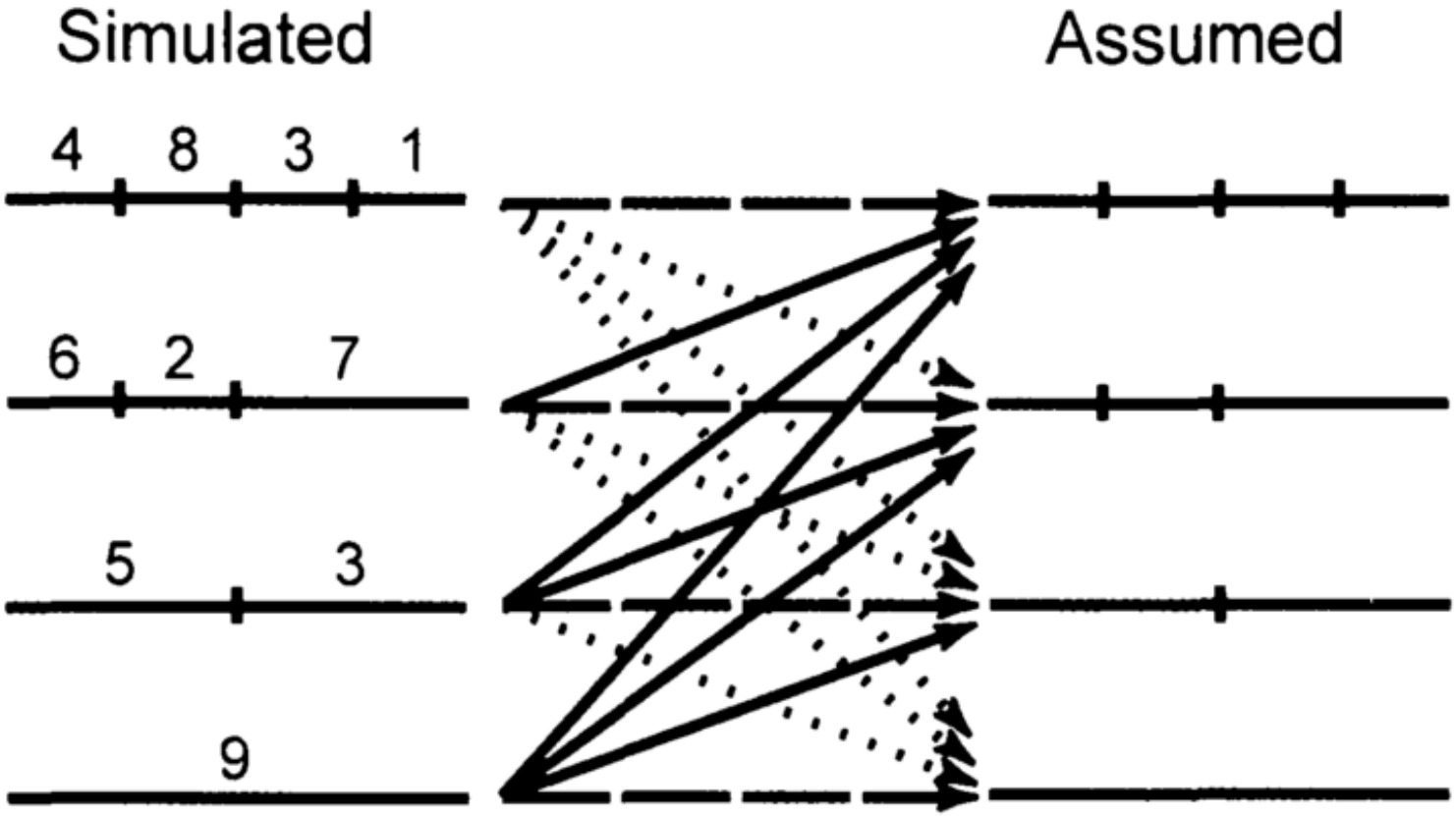
$$L(\tau, \boldsymbol{\nu}, \boldsymbol{\Phi}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i | \tau, \boldsymbol{\nu}, \boldsymbol{\Phi})$$

- Typically (but not necessarily):
  - τ, $\boldsymbol{\nu}$ - shared among data partitions
  - $\boldsymbol{\Phi}$ -independent for each data partition

$$\boldsymbol{r} = (r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct}, r_{gt})$$
$$\boldsymbol{\pi} = (\pi_a, \pi_c, \pi_g, \pi_t)$$
$$\alpha$$

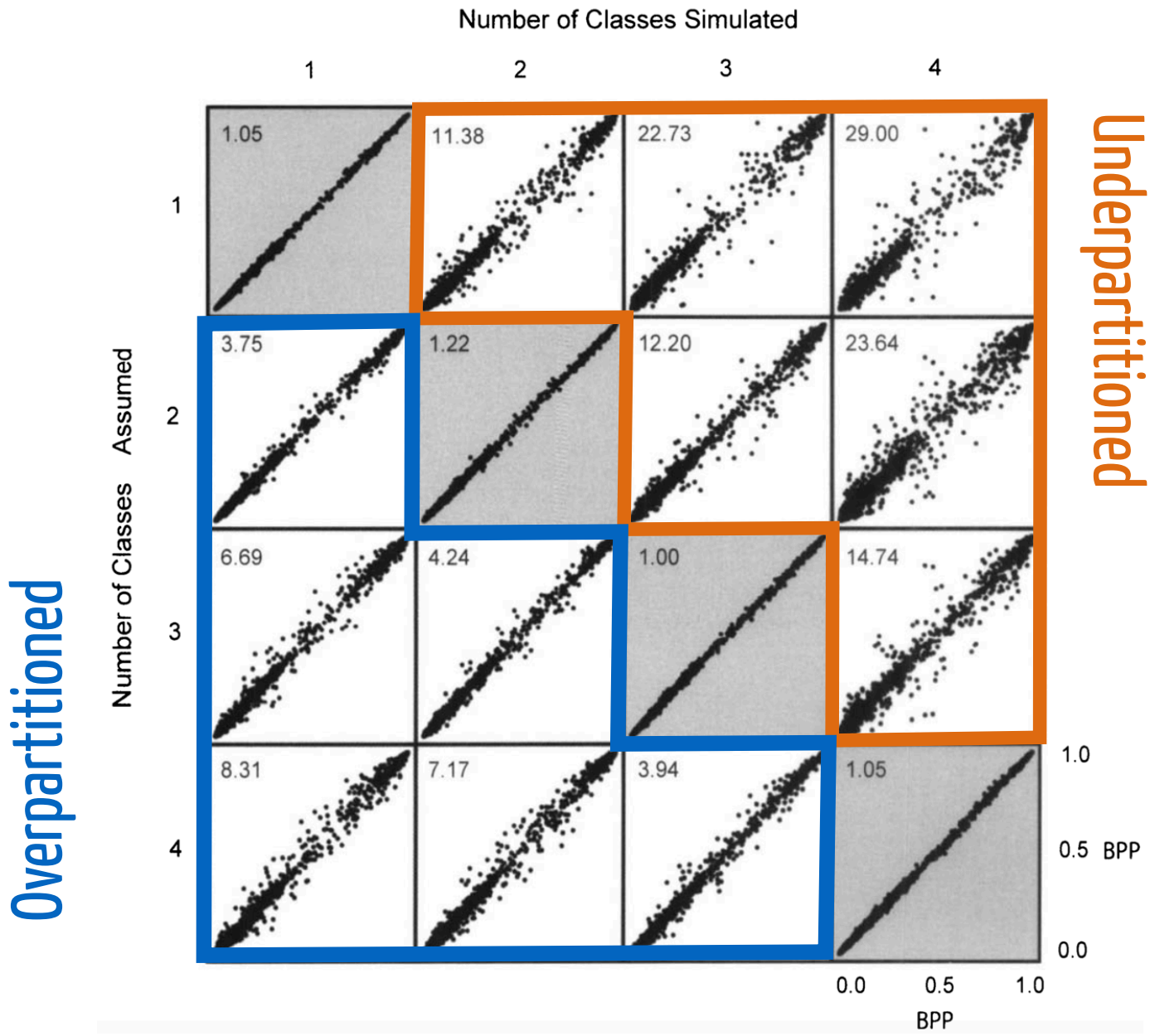
How does this influence inference?

Importance of accommodating process heterogeneity



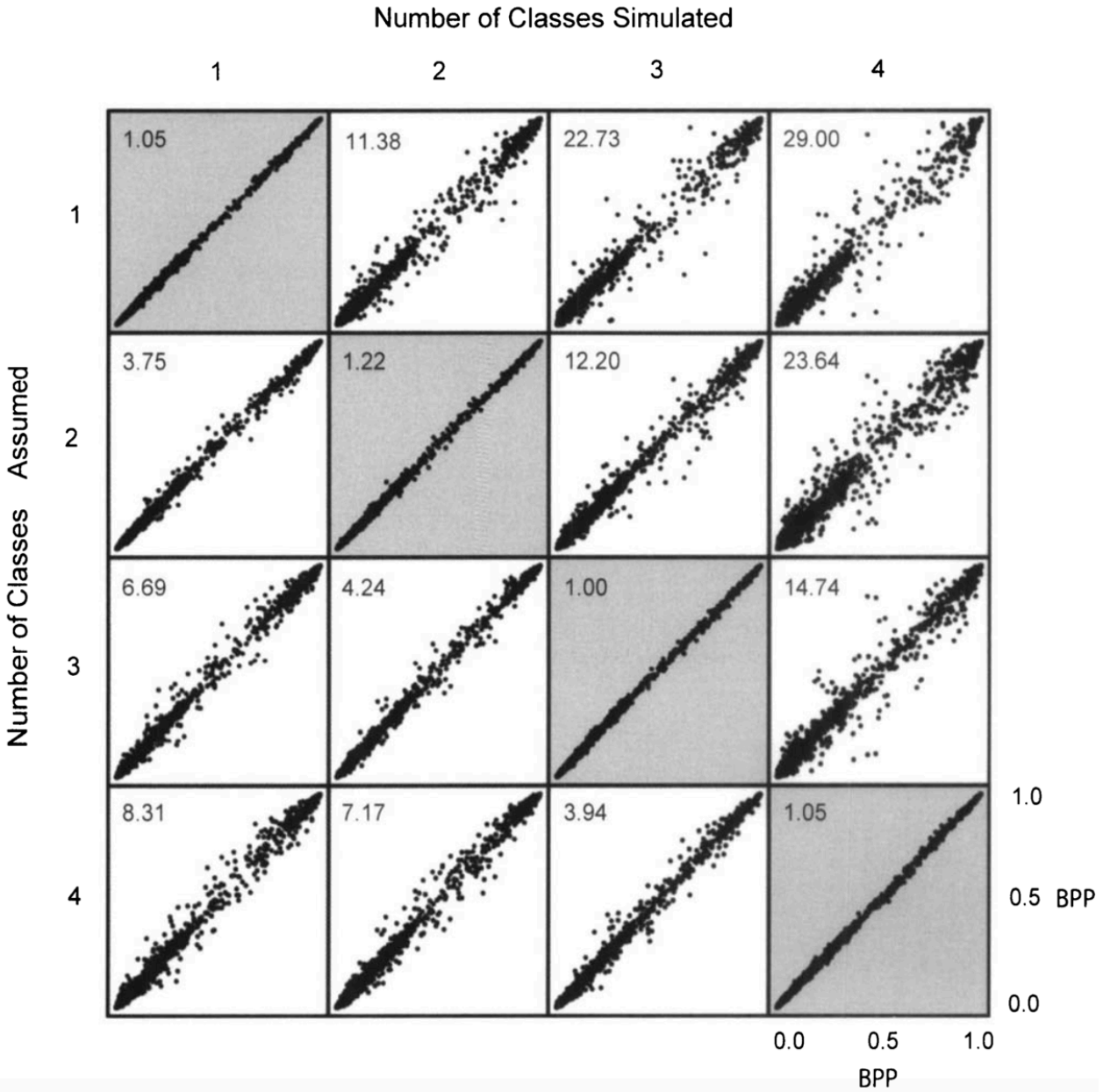
Brown & Lemmon 2007

Importance of accommodating process heterogeneity



Brown & Lemmon 2007

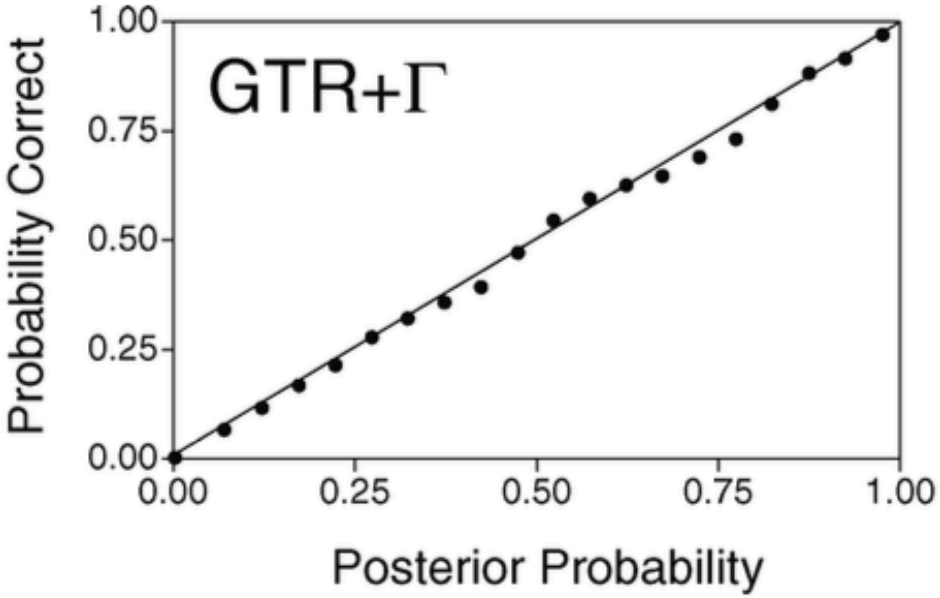
Importance of accommodating process heterogeneity

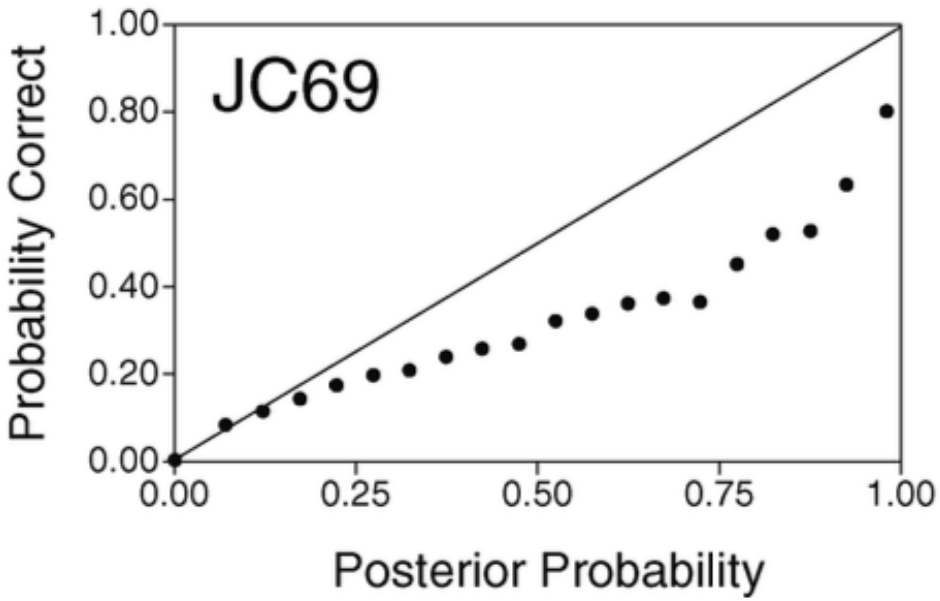


Brown & Lemmon 2007

Underparameterization is often worse than overparameterization

Generating model is GTR+G

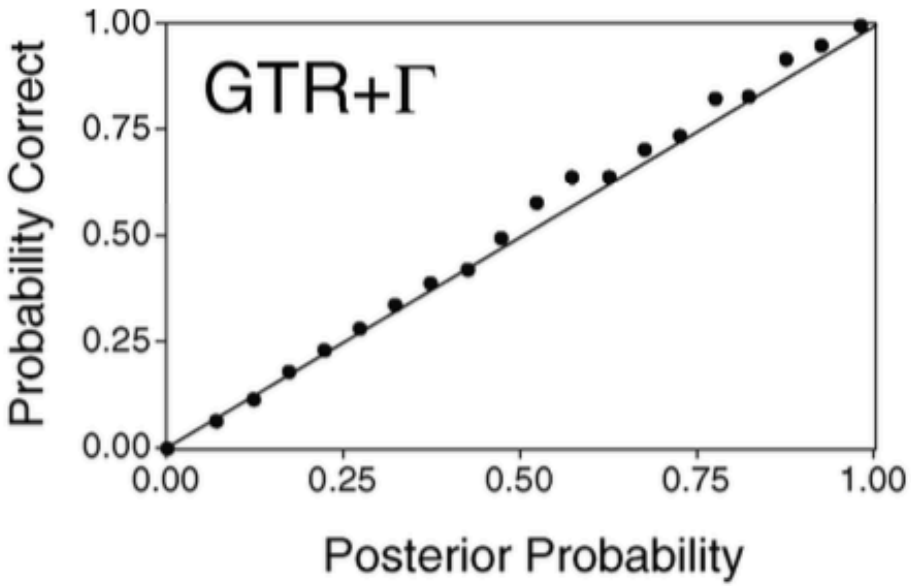


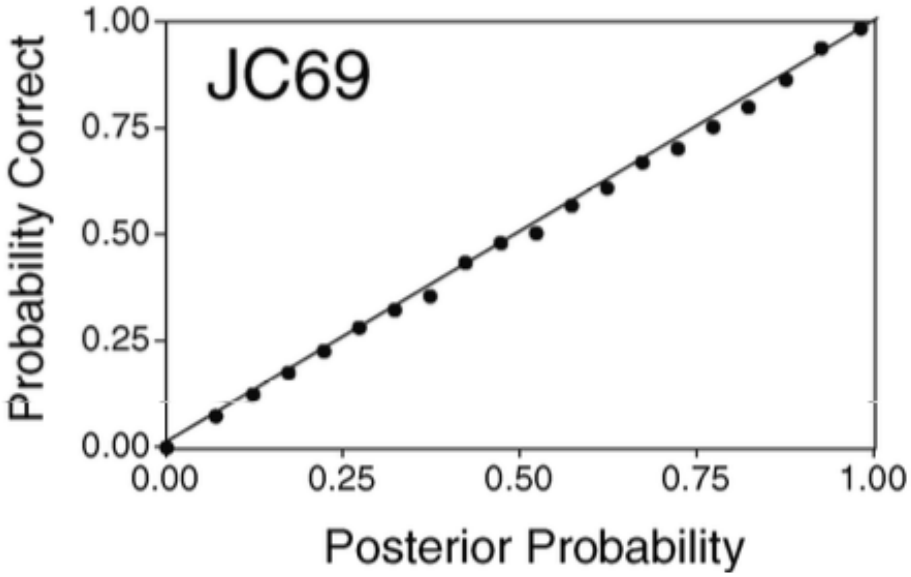


Huelsenbeck & Rannala 2004

Underparameterization is often worse than Overparameterization

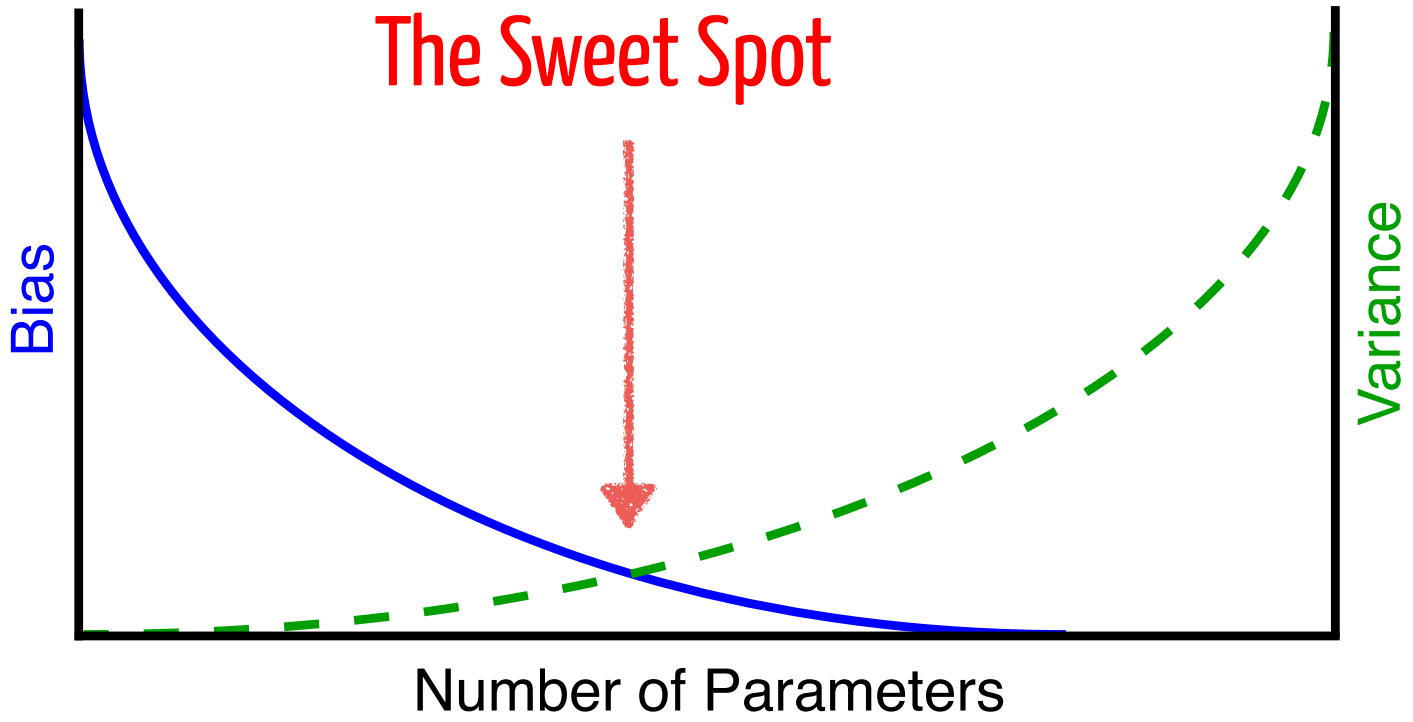
Generating model is JC



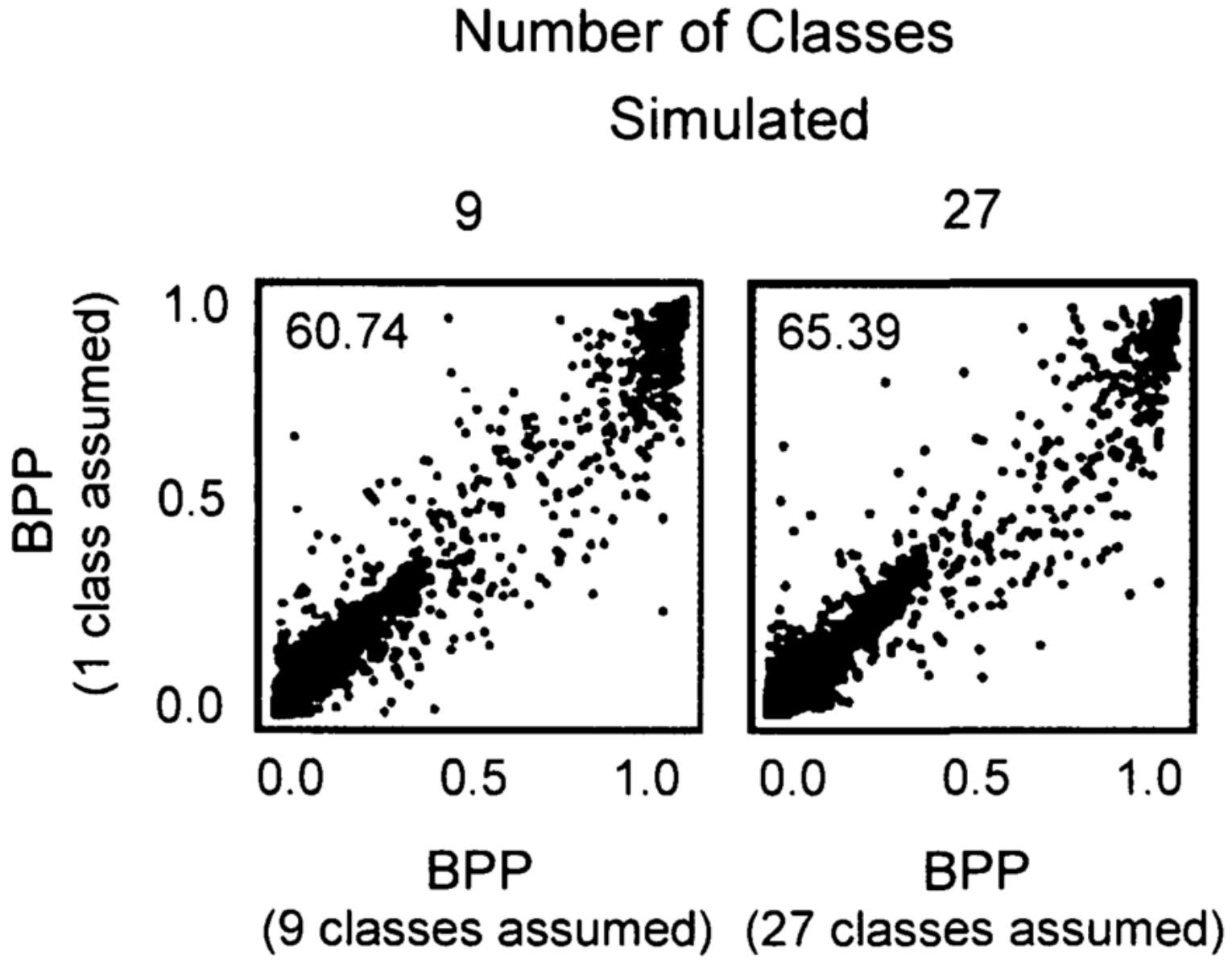


Huelsenbeck & Rannala 2004

From earlier today...



Importance of accommodating process heterogeneity



Brown & Lemmon 2007

Take homes

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors
- this effect is worse in large (more heterogenous) datasets

Importance of accommodating process heterogeneity

Data Set	Data Type[a]	Taxa	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.
Anderson_2013	M	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)
Brown_2012	N	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)
Devitt_2013	M	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)
Fong_2012	N	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)
Kang_2013a	N	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kang_2013b	M	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)
Lartillot_2012	N	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)
Leavitt_2013	M	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA
Li_2008	N	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)
Seago_2011	M	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)
Sharanowski_2011	N	139	3,982	11	Braconidae	Wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)
Siler_2013	M,N	61	2,697	7	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)
Tolley_2013	M	203	5,054	16	Chamaeleonidae	Chameleons	Tolley et al. (2013b)	Tolley et al. (2013a)
Unmack_2013	M	139	6,827	25	Melanotaeniidae	Rainbowfish	Unmack et al. (2013b)	Unmack et al. (2013a)
Wainwright_2012	N	188	8,439	30	Acanthomorpha	Fishes	Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn, Eytan, et al. (2012)	Wainwright et al. (2012)
Ward_2010	N	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)

34 real datasets

Kainer and Lanfear 2015

Importance of accommodating process heterogeneity

Data Set	Data Type[a]	Taxa	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.
Anderson_2013	M	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)
Brown_2012	N	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)
Devitt_2013	M	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)
Fong_2012	N	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)
Kang_2013a	N	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kang_2013b	M	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)
Lartillot_2012	N	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)
Leavitt_2013	M	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA
Li_2008	N	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)
Seago_2011	M	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)
Sharanowski_2011	N	139	3,982	11	Braconidae	Wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)
Siler_2013	M,N	61	2,697	7	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)
Tolley_2013	M	203	5,054	16	Chamaeleonidae	Chameleons	Tolley et al. (2013b)	Tolley et al. (2013a)
Unmack_2013	M	139	6,827	25	Melanotaeniidae	Rainbowfish	Unmack et al. (2013b)	Unmack et al. (2013a)
Wainwright_2012	N	188	8,439	30	Acanthomorpha	Fishes	Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn, Eytan, et al. (2012)	Wainwright et al. (2012)
Ward_2010	N	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)

34 real datasets

4 partitioning schemes

1. unpartitioned
2. a priori by feature
3. chosen using AIC
4. chosen using BIC

Kainer and Lanfear 2015

Importance of accommodating process heterogeneity

Data Set	Data Type[a]	Taxa	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.
Anderson_2013	M	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)
Brown_2012	N	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)
Devitt_2013	M	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)
Fong_2012	N	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)
Kang_2013a	N	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kang_2013b	M	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)
Lartillot_2012	N	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)
Leavitt_2013	M	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA
Li_2008	N	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)
Seago_2011	M	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)
Sharanowski_2011	N	139	3,982	11	Braconidae	Wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)
Siler_2013	M,N	61	2,697	7	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)
Tolley_2013	M	203	5,054	16	Chamaeleonidae	Chameleons	Tolley et al. (2013b)	Tolley et al. (2013a)
Unmack_2013	M	139	6,827	25	Melanotaeniidae	Rainbowfish	Unmack et al. (2013b)	Unmack et al. (2013a)
Wainwright_2012	N	188	8,439	30	Acanthomorpha	Fishes	Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn, Eytan, et al. (2012)	Wainwright et al. (2012)
Ward_2010	N	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)

34 real datasets

4 partitioning schemes

Simple

1. unpartitioned
2. a priori by feature
3. chosen using AIC
4. chosen using BIC

more complex

Kainer and Lanfear 2015

Importance of accommodating process heterogeneity

Data Set	Data Type[a]	Taxa	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.
Anderson_2013	M	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)
Brown_2012	N	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)
Devitt_2013	M	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)
Fong_2012	N	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)
Kang_2013a	N	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kang_2013b	M	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)
Lartillot_2012	N	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)
Leavitt_2013	M	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA
Li_2008	N	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)
Seago_2011	M	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)
Sharanowski_2011	N	139	3,982	11	Braconidae	Wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)
Siler_2013	M,N	61	2,697	7	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)
Tolley_2013	M	203	5,054	16	Chamaeleonidae	Chameleons	Tolley et al. (2013b)	Tolley et al. (2013a)
Unmack_2013	M	139	6,827	25	Melanotaeniidae	Rainbowfish	Unmack et al. (2013b)	Unmack et al. (2013a)
Wainwright_2012	N	188	8,439	30	Acanthomorpha	Fishes	Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn, Eytan, et al. (2012)	Wainwright et al. (2012)
Ward_2010	N	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)

34 real datasets

4 partitioning schemes

1. unpartitioned
2. a priori by feature
3. chosen using AIC
4. chosen using BIC

Analyze

Kainer and Lanfear 2015

Importance of accommodating process heterogeneity

Data Set	Data Type[a]	Taxa	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.
Anderson_2013	M	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)
Brown_2012	N	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)
Devitt_2013	M	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)
Fong_2012	N	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)
Kang_2013a	N	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kang_2013b	M	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)
Lartillot_2012	N	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)
Leavitt_2013	M	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA
Li_2008	N	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)
Seago_2011	M	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)
Sharanowski_2011	N	139	3,982	11	Braconidae	Wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)
Siler_2013	M,N	61	2,697	7	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)
Tolley_2013	M	203	5,054	16	Chamaeleonidae	Chameleons	Tolley et al. (2013b)	Tolley et al. (2013a)
Unmack_2013	M	139	6,827	25	Melanotaeniidae	Rainbowfish	Unmack et al. (2013b)	Unmack et al. (2013a)
Wainwright_2012	N	188	8,439	30	Acanthomorpha	Fishes	Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn, Eytan, et al. (2012)	Wainwright et al. (2012)
Ward_2010	N	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)

34 real datasets

4 partitioning schemes

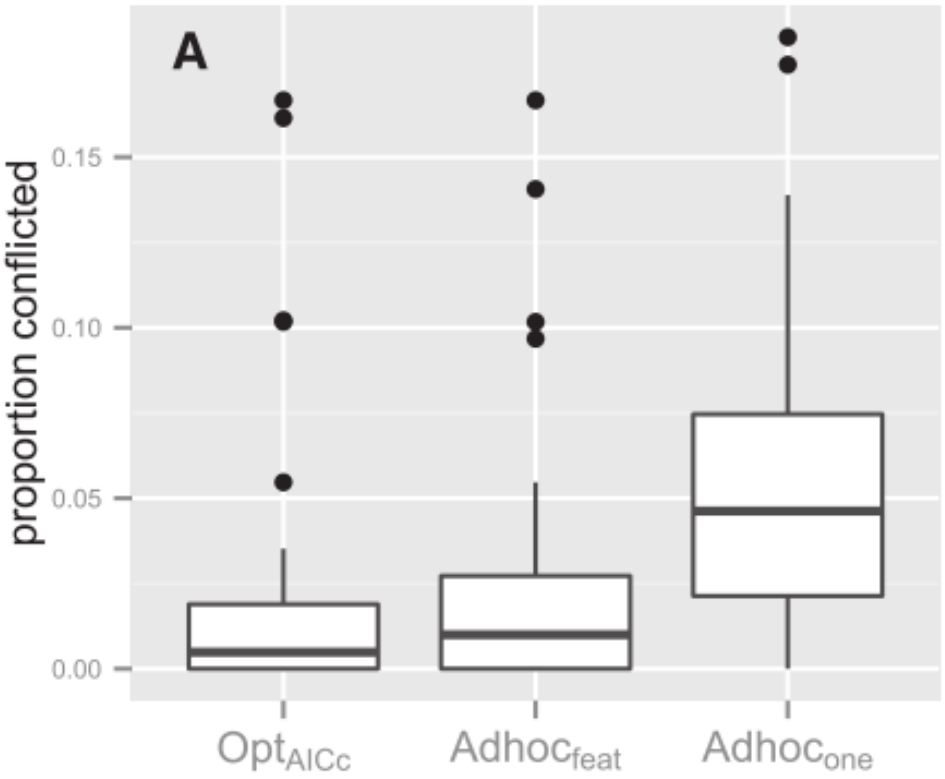
1. unpartitioned
2. a priori by feature
3. chosen using AIC
4. chosen using BIC

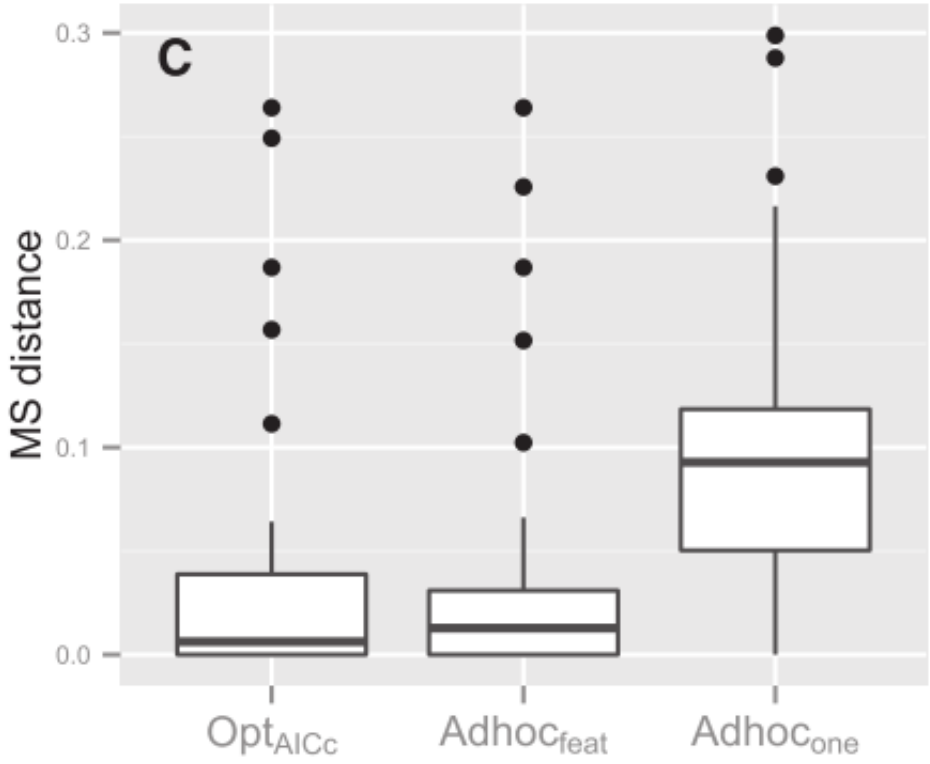
Analyze

Ask what changes

Kainer and Lanfear 2015

Importance of accommodating process heterogeneity

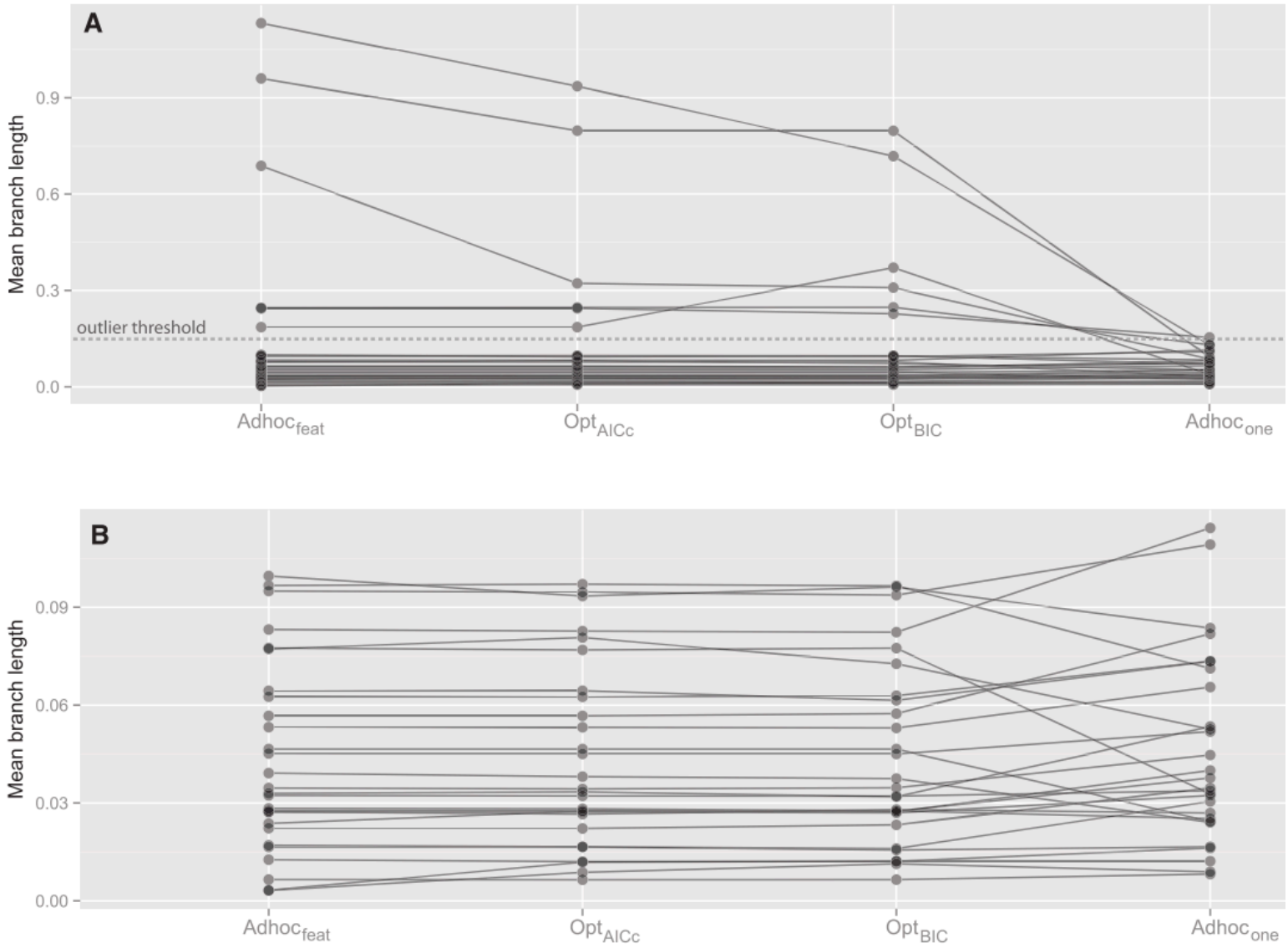




Topology changes

Kainer and Lanfear 2015

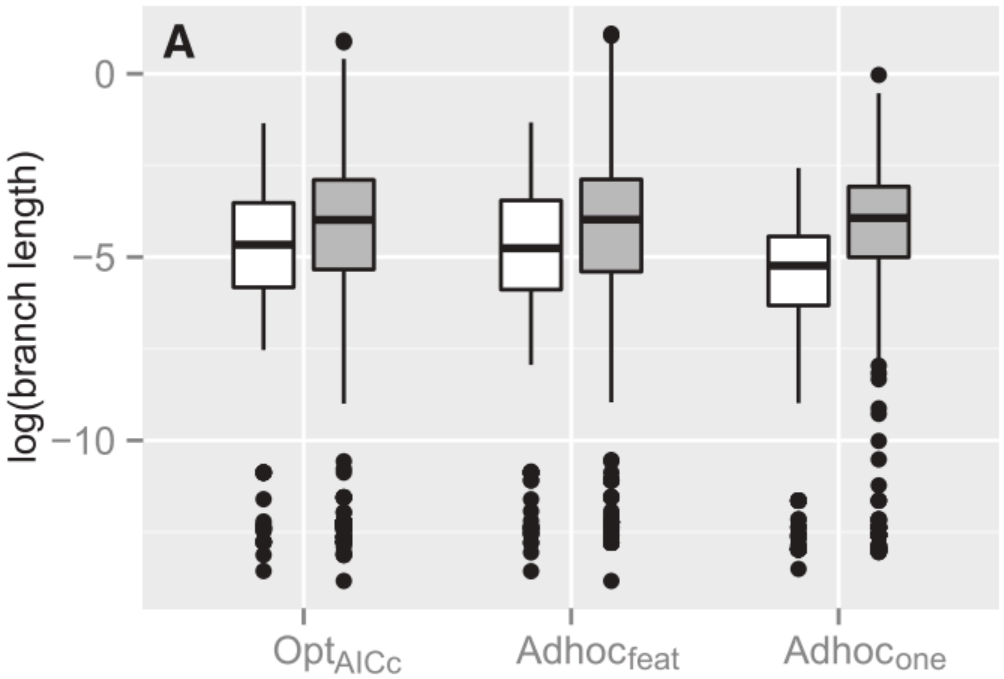
Importance of accommodating process heterogeneity

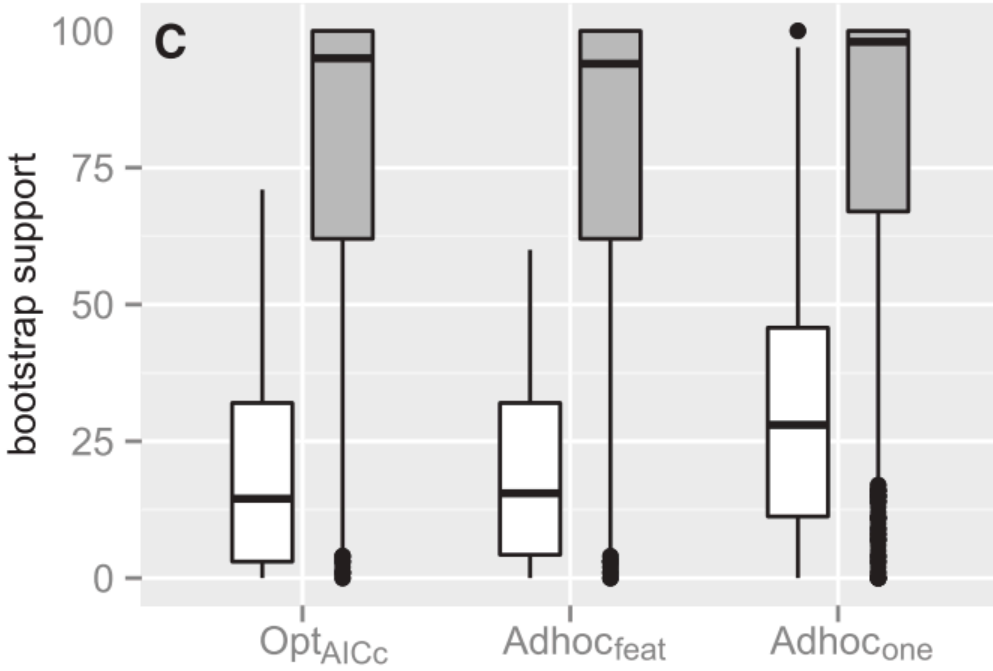


Branch lengths also change

Kainer and Lanfear 2015

Importance of accommodating process heterogeneity





Nodes subtended by short branch lengths and low support disproportionately effected

Kainer and Lanfear 2015

Take homes

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors
- this effect is worse in large (more heterogenous) datasets

Take homes

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors, **and branch lengths, and topologies**
- this effect is worse in large (more heterogenous) datasets

Take homes

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors, **and branch lengths, and topologies**
- this effect is worse in large (more heterogenous) datasets
- this effect is worse for difficult to resolve nodes

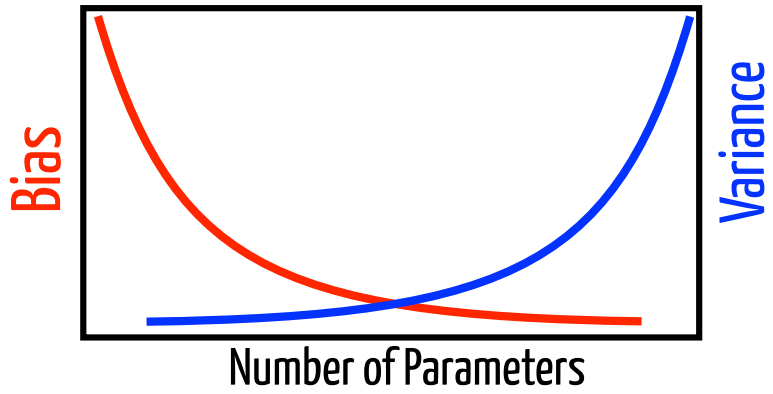
Take homes

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors, and branch lengths, and topologies
- **this effect is worse in large (more heterogenous) datasets**
- **this effect is worse for difficult to resolve nodes**

Precise conditions of 'Big Data' studies

How do we select a partitioning model?

- This is a similar problem to what we covered this morning
  - Seek to capture the relevant variation in the data, balancing the bias-variance tradeoff

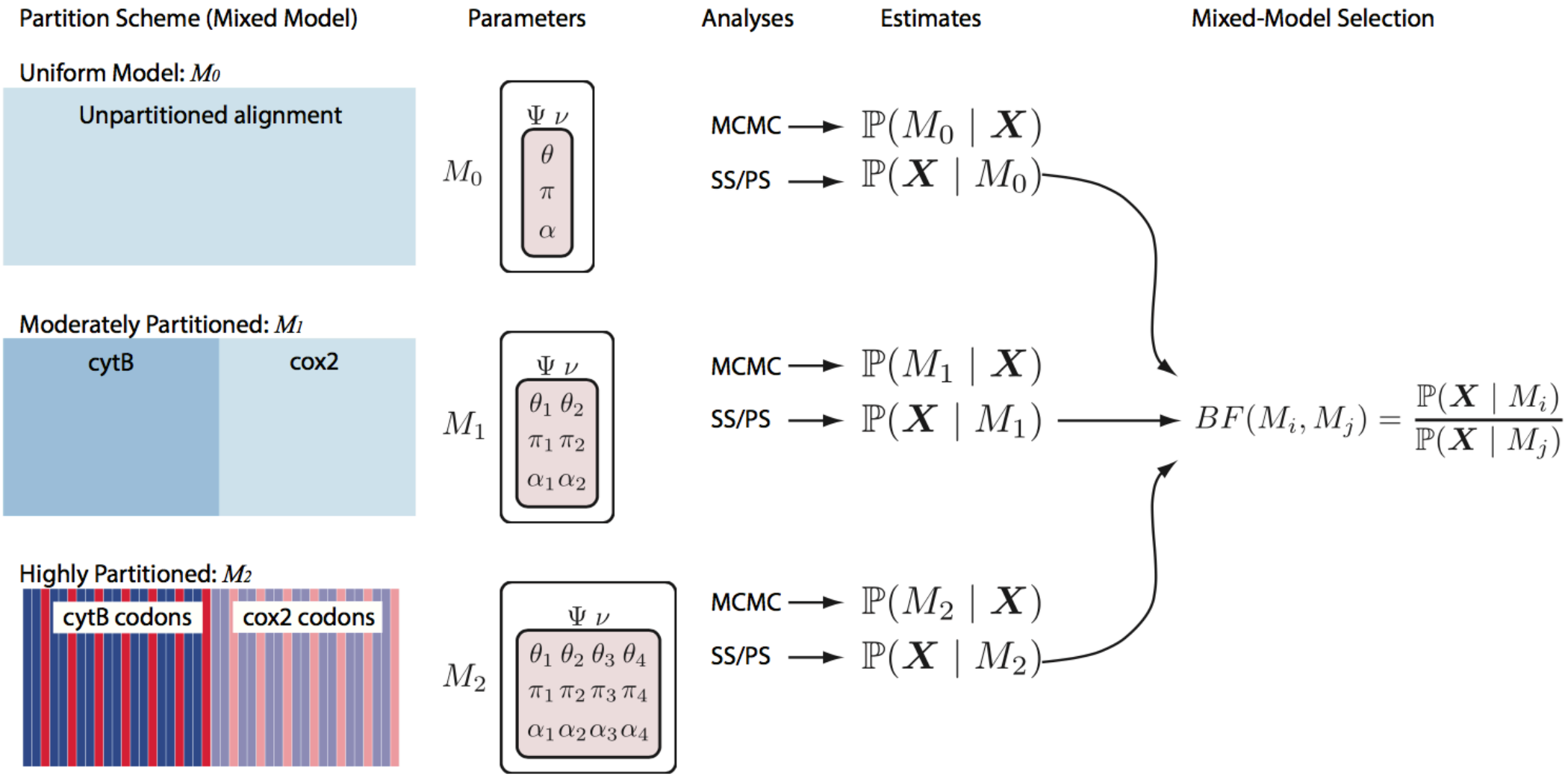


Extreme Underpartitioning

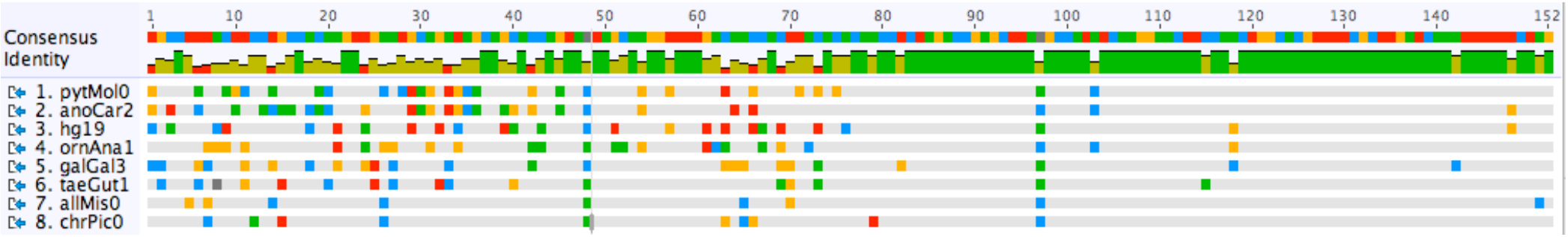
Extreme Overpartitioning

- We can use similar tools as before

How do we select a partitioning model?



Scaling up



- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	**203**

Scaling up

- Many possible partition models even for ‘small data’.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	**203**

Scaling up

- Many possible partition models even for ‘small data’.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	**203**

Two issues:

1. This counts only possibilities where the partition model is the same for all parameters

Scaling up

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	203

Phylogenetic Model Parameters

Topology τ

Branch lengths $\boldsymbol{\nu}$

Exchangeability rates $\boldsymbol{r} = (r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct}, r_{gt})$

Base frequencies $\boldsymbol{\pi} = (\pi_a, \pi_c, \pi_g, \pi_t)$

ASRV Gamma Shape α

Scaling up

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	203

Phylogenetic Model Parameters

Topology τ

Branch lengths $\boldsymbol{\nu}$

Shared across partitions

Exchangeability rates $\boldsymbol{r} = (r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct}, r_{gt})$

Base frequencies $\boldsymbol{\pi} = (\pi_a, \pi_c, \pi_g, \pi_t)$

ASRV Gamma Shape α

Scaling up

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	203

Phylogenetic Model Parameters

Topology τ

Branch lengths $\boldsymbol{\nu}$

Shared across partitions

Exchangeability rates $\boldsymbol{r} = (r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct}, r_{gt})$

Base frequencies $\boldsymbol{\pi} = (\pi_a, \pi_c, \pi_g, \pi_t)$

ASRV Gamma Shape α

Independent for each partition

Scaling up

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	203

Phylogenetic Model Parameters

Shared across partitions

Topology τ

Branch lengths $\boldsymbol{\nu}$

Exchangeability rates $\boldsymbol{r} = (r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct}, r_{gt})$

Base frequencies $\boldsymbol{\pi} = (\pi_a, \pi_c, \pi_g, \pi_t)$

ASRV Gamma Shape α

Should we consider other partial partition models?

If so, **many** more models are possible.

Scaling up

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	203

Phylogenetic Model Parameters

Topology τ

Branch lengths $\boldsymbol{\nu}$

Shared across partitions

Exchangeability rates $\boldsymbol{r} = (r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct}, r_{gt})$

Base frequencies $\boldsymbol{\pi} = (\pi_a, \pi_c, \pi_g, \pi_t)$

ASRV Gamma Shape α

Should we consider other partial partition models?

If so, **many** more models are possible.

Scaling up

- Many possible partition models even for ‘small data’.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	**203**

Two issues:

2. This counts only partition models for classes that we thought to write down a priori.

Scaling up

- Many possible partition models even for ‘small data’.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	**203**

Like many problems in phylogenetics, this doesn’t scale easily

Scaling up: potential solutions

- **Model selection:** select a single ‘best fitting’ model by comparing many possible alternatives

Scaling up: potential solutions

- **Model selection:** select a single 'best fitting' model by comparing many possible alternatives
  - estimate marginal likelihood and use **Bayes Factors** to do model selection
  - **AIC** or **BIC** model selection (e.g., Lanfear et al. 2012, Lanfear et al. 2017)

Scaling up: potential solutions

- **Model selection:** select a single 'best fitting' model by comparing many possible alternatives
  - estimate marginal likelihood and use **Bayes Factors** to do model selection
  - **AIC** or **BIC** model selection (e.g., Lanfear et al. 2012, Lanfear et al. 2017)

General approach: Calculate likelihood of data under a particular model, penalize for model complexity via the prior (marginal likelihood) or some penalty term (AIC or BIC). Compare.

Model Selection and Model Averaging

- **Model Selection:** Choose the model that best fits the data.
  - The data are random variables, the parameters are (or are not) random variables, but **the model is fixed.**

Model Selection and Model Averaging

- **Model Selection:** Choose the model that best fits the data.
  - The data are random variables, the parameters are (or are not) random variables, but **the model is fixed.**
- **Model Averaging:** Consider alternative models in proportion to their probability
  - The **model itself is a random variable** with associated uncertainty, so we account for this.

Scaling up: potential solutions

- **Model Averaging:** Treat the partition-model itself as a random variable and use MCMC to marginalize over possible partition models (and other parameters of the joint phylogenetic model)
  - Moore et al. 2014 - AutoParts
  - Wu et al. 2013 - substBMA (addon for BEAST 2)
  - RevBayes

Scaling up: potential solutions

- **Model Averaging:** Treat the partition-model itself as a random variable and use MCMC to marginalize over possible partition models (and other parameters of the joint phylogenetic model)
  - Moore et al. 2014 - AutoParts
  - Wu et al. 2013 - substBMA (addon for BEAST 2)
  - RevBayes

General approach: Model the number of partitions and the assignment of sites or classes to those partitions using the Dirichlet Process

Dirichlet Process

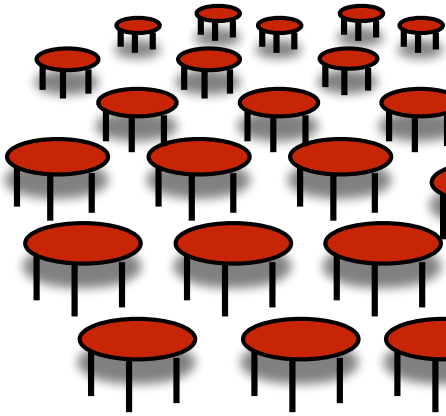
- A stochastic process that allows us to describe the prior probability of a mixture model for the **number of partitions** and the **assignment of data to those partitions**

Dirichlet Process

- A stochastic process that allows us to describe the prior probability of a mixture model for the **number of partitions** and the **assignment of data to those partitions**
- More simply: A prior probability distribution for clustering problems
  - How many clusters are there?
  - Which observations belong to which clusters?

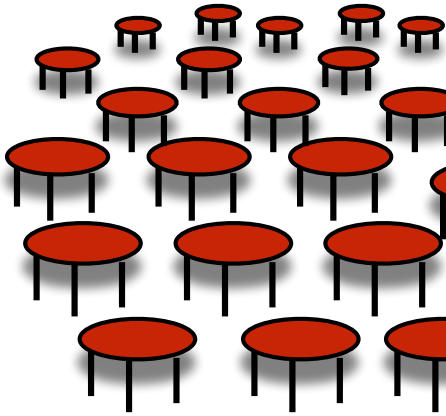
‘The Chinese Restaurant Process’

- Imagine a restaurant with an infinite number of tables. Customers walk in one at a time and choose a table to sit down at:



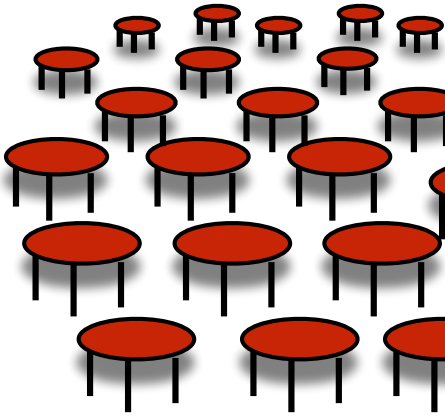
‘The Chinese Restaurant Process’

- Imagine a restaurant with an infinite number of tables. Customers walk in one at a time and choose a table to sit down at:
  - The first customer always chooses the first table.



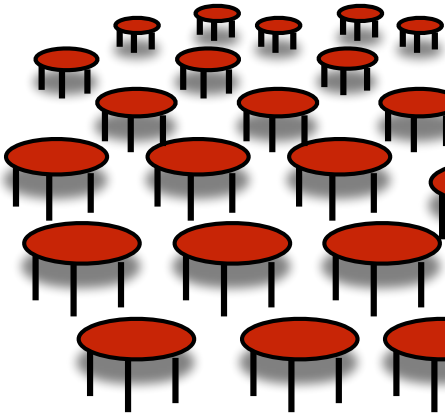
‘The Chinese Restaurant Process’

- Imagine a restaurant with an infinite number of tables. Customers walk in one at a time and choose a table to sit down at:
  - The first customer always chooses the first table.
  - The nth customer chooses the first unoccupied table with probability $\frac{\alpha}{n-1+\alpha}$, and an occupied table with probability $\frac{c}{n-1+\alpha}$, where c is the number of people already at that table.



‘The Chinese Restaurant Process’

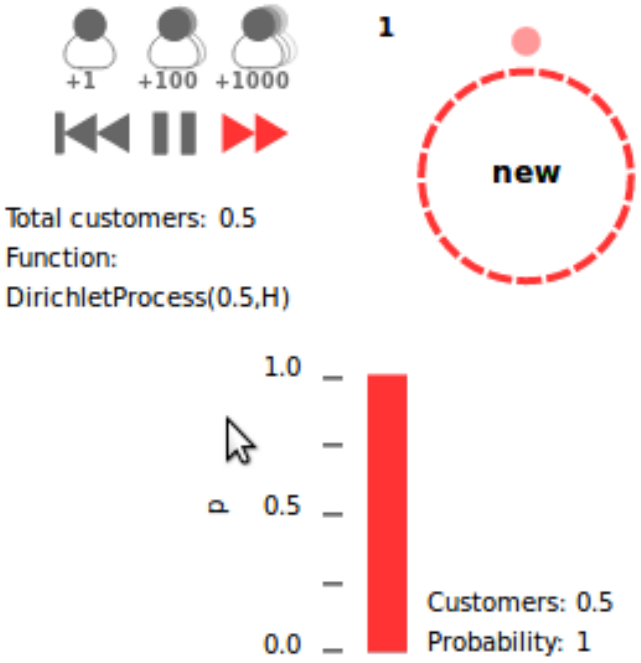
- Imagine a restaurant with an infinite number of tables. Customers walk in one at a time and choose a table to sit down at:
  - The first customer always chooses the first table.
  - The nth customer chooses the first unoccupied table with probability $\frac{\alpha}{n-1+\alpha}$, and an occupied table with probability $\frac{c}{n-1+\alpha}$, where c is the number of people already at that table.



Large alpha - many tables in use

Popular tables stay popular

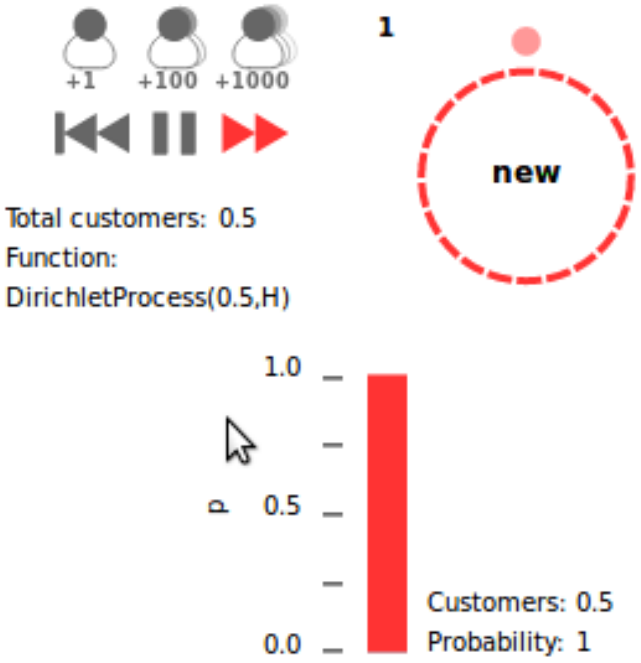
‘The Chinese Restaurant Process’



Sorts customers into tables

http://topicmodels.west.uni-koblenz.de/ckling/tmt/crp.html

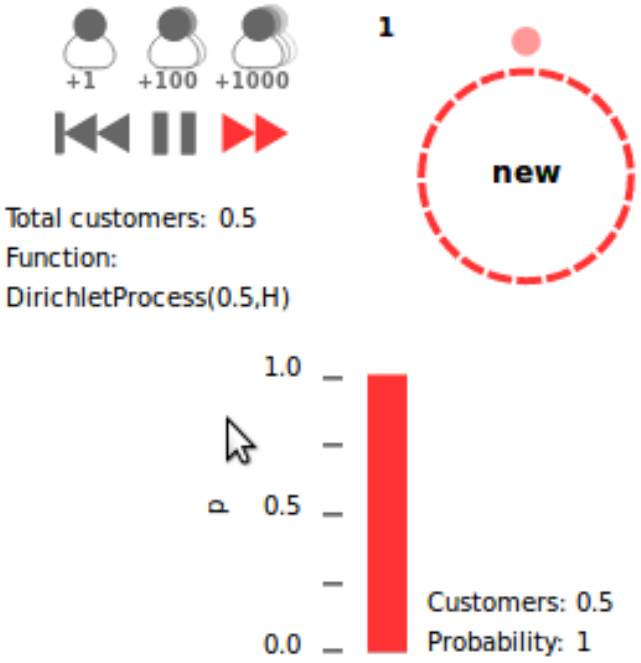
‘The Chinese Restaurant Process’



Sorts customers into tables
Sorts observations into clusters

http://topicmodels.west.uni-koblenz.de/ckling/tmt/crp.html

‘The Chinese Restaurant Process’



Sorts customers into tables
Sorts observations into clusters
Sorts alignment classes into partitions

http://topicmodels.west.uni-koblenz.de/ckling/tmt/crp.html

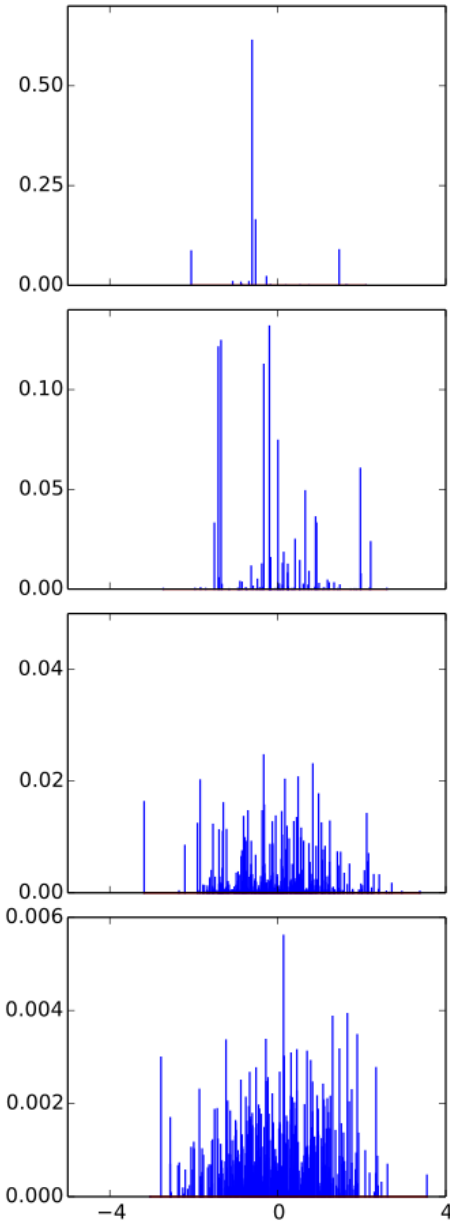
Dirichlet Process Prior

$\alpha = 1$

$\alpha = 10$

$\alpha = 100$

$\alpha = 1000$



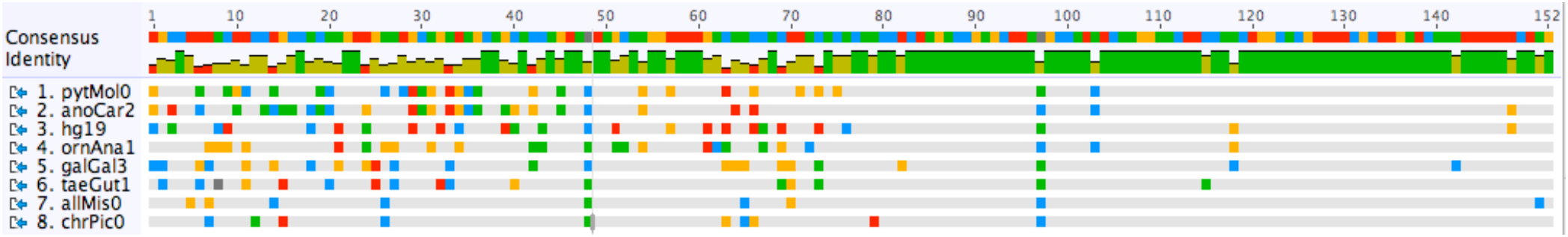
High Prior Probability for **few partitions**

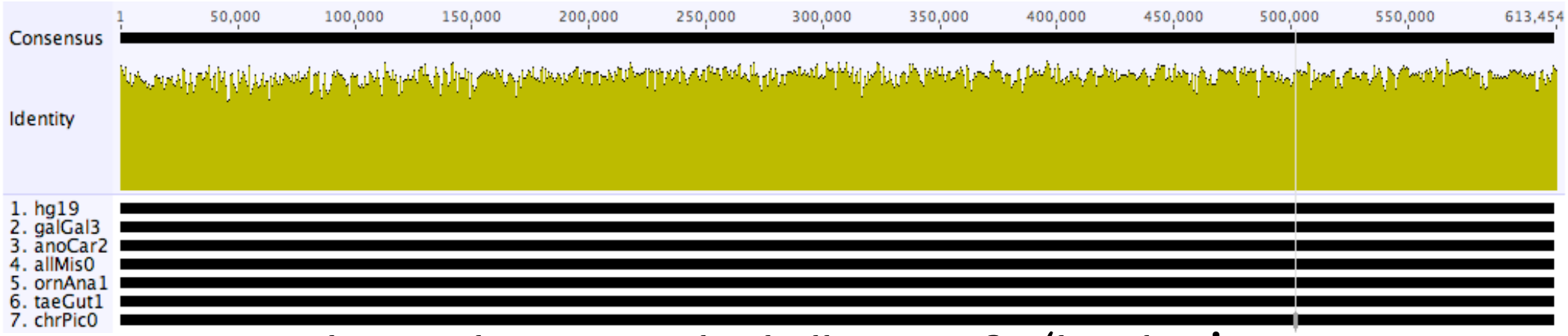
High Prior Probability for **many partitions**

new cluster: $\frac{\alpha}{n-1+\alpha}$

'occupied' cluster: $\frac{c}{n-1+\alpha}$

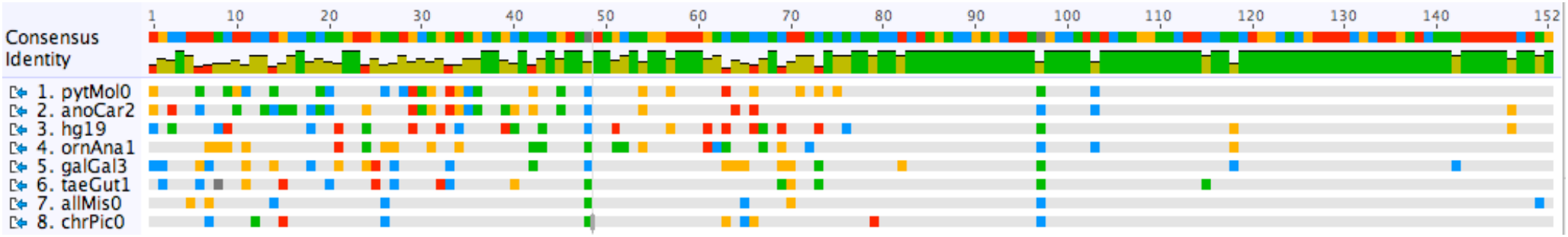
Scaling up

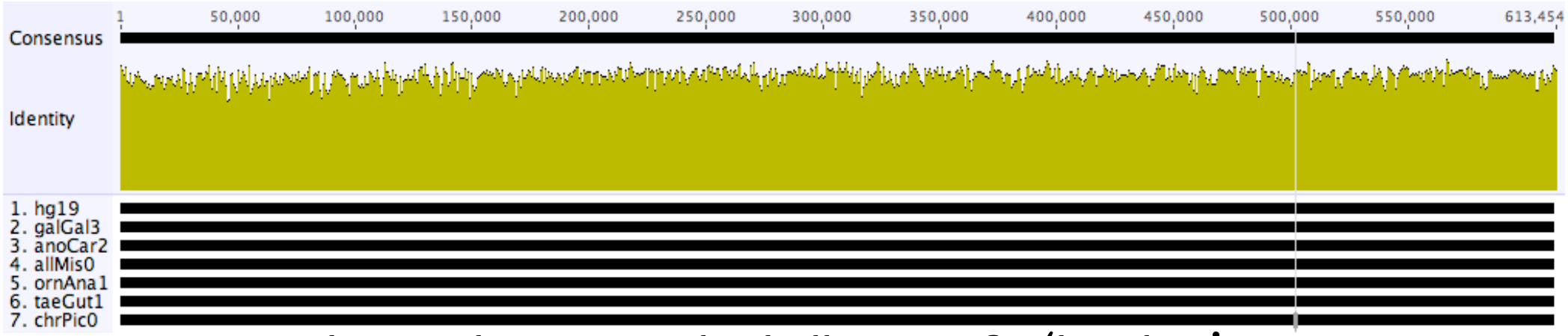




This can be extremely challenging for ‘big data’.

Scaling up



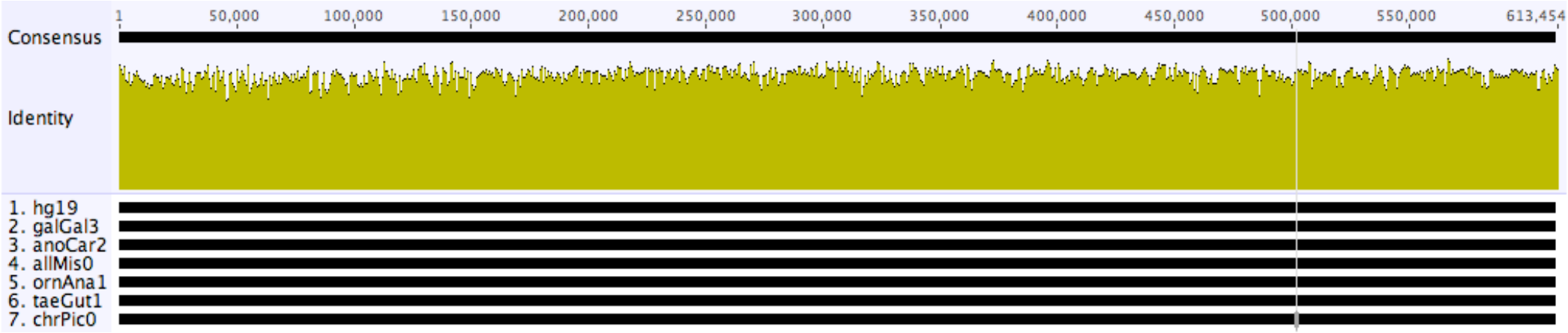


This can be extremely challenging for 'big data'.

e.g. 100 classes:

4758539127676483365879076884138720782636366968682561146661633463755911449789244262267272404421775630695355788256075 1 partition models

Scaling up



Scaling up may also mean that we need to relax this assumption

Phylogenetic Model Parameters

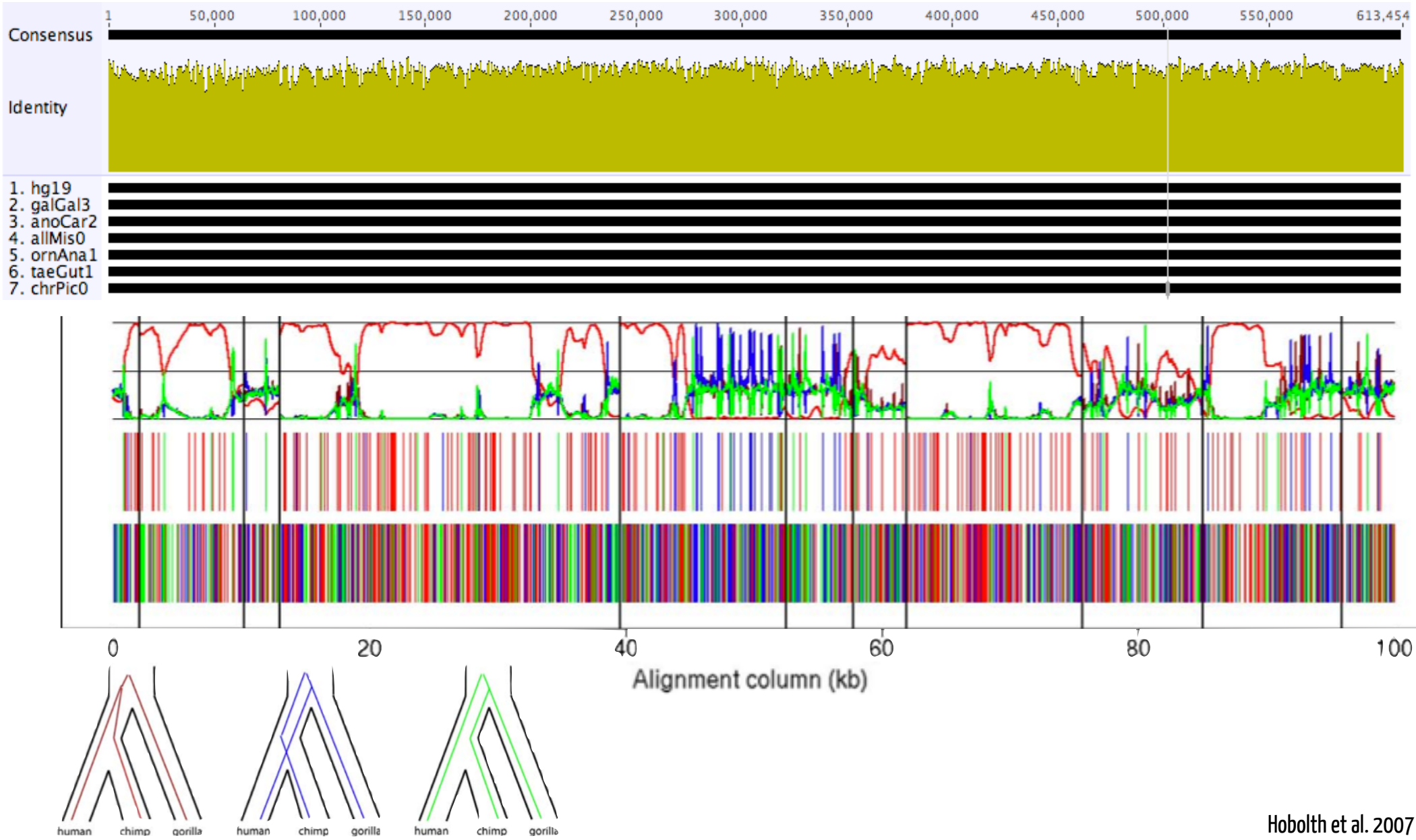
Topology	τ
Branch lengths	$\boldsymbol{\nu}$

Shared across partitions

Exchangeability rates	$\boldsymbol{r} = (r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct}, r_{gt})$
Base frequencies	$\boldsymbol{\pi} = (\pi_a, \pi_c, \pi_g, \pi_t)$
ASRV Gamma Shape	α

Independent for each partition

Scaling up



Hobolth et al. 2007

Tutorial

- Partition models in RevBayes
  - **set up** partitioned model
  - run **mcmc**
  - stepping stone integration for **marginal likelihood estimation**

Exercise

- I have supplied data for four 'loci' (sim_locus1.nex,...sim_locus4.nex)
- Choose (**and sign up for**) one partition model of the form (12,34 or 1,2,34 or 1234 or 1,2,3,4)
- GTR + Gamma model for all partitions
- Settings for stepping stone integration:
  - 50 power posteriors (cats=50)
  - burnin=500, tuning interval=200
  - 2000 generations per power posterior
  - add your estimate to the spreadsheet