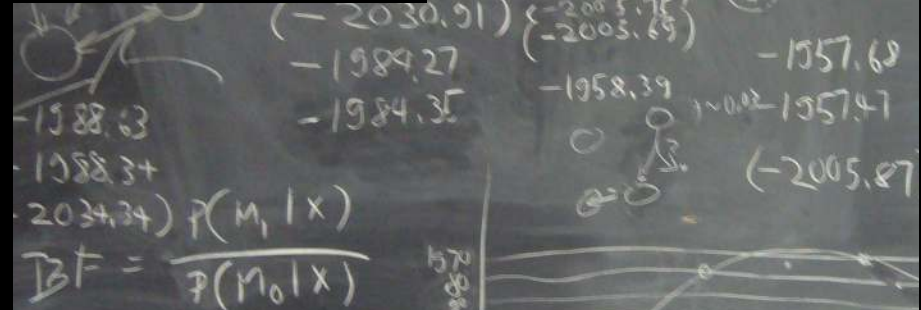


Sample size, model choice, and parallel runs

Peter Beerli

Department of Scientific Computing
Florida State University, Tallahassee



Wednesday, February 6, 13

1

Overview

1. Sample size
2. Model parameters
3. How to reduce parameters
4. Bayes Factors (practical)
5. Replication and parallel runtime

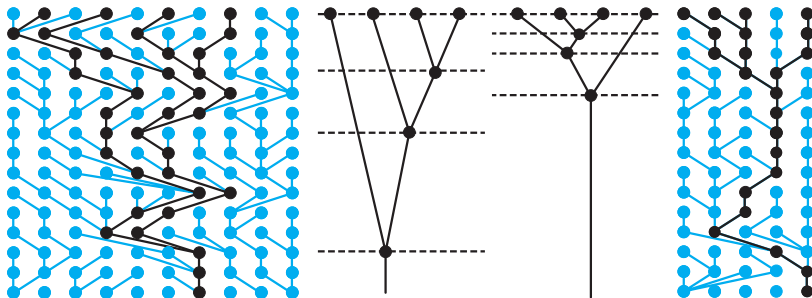
2 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

2

Population size mantra

Coalescence



3 of 33 – ©2012 Peter Beerli

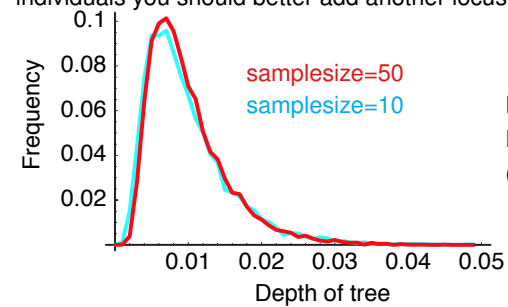
Wednesday, February 6, 13

3

Required samples is small

Single population

- ◆ The time to the most recent common ancestor is robust to different sample sizes.
- ◆ Simulated sequence data from a single population have shown that after 8 individuals you should better add another locus than more individuals.



Felsenstein (2005)
Pluzhnikov and Donnelly
(1996)

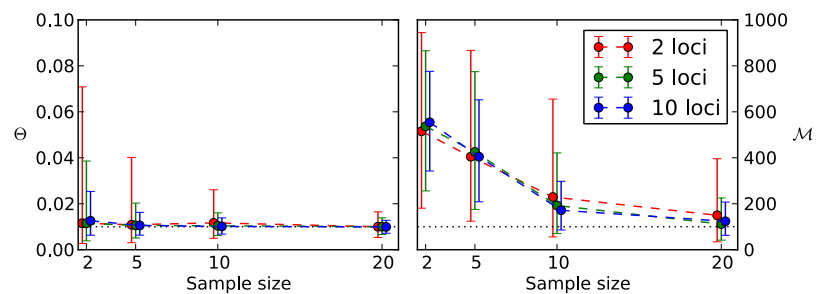
4 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

4

Required samples is small

Multiple populations



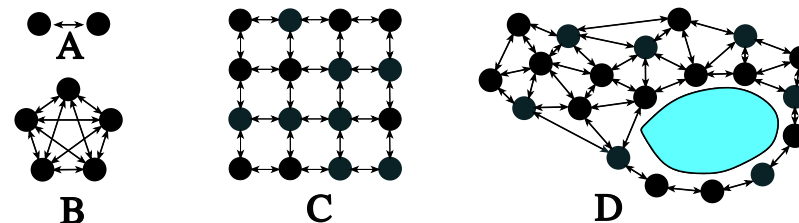
Medium variability DNA dataset: Mutation-scaled population size Θ and mutation-scaled migration rate M versus sample size for 2, 5, and 10 loci. The true $\Theta_T = 0.01$ is marked with the dotted gray line; $M = 100$

5 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

5

Population models

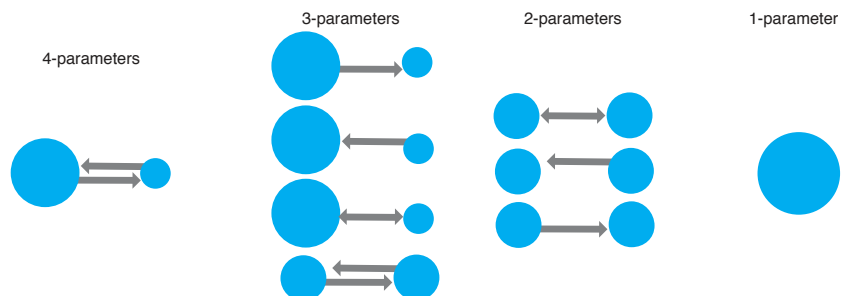


6 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

6

Migration model specification

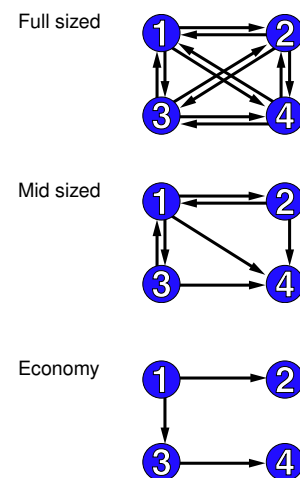


7 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

7

Migration model specification

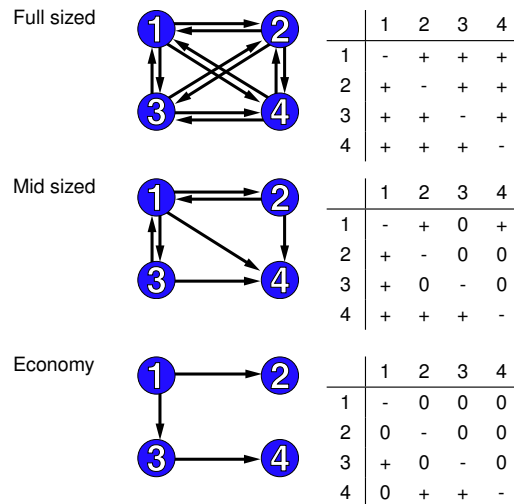


8 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

8

Migration model specification

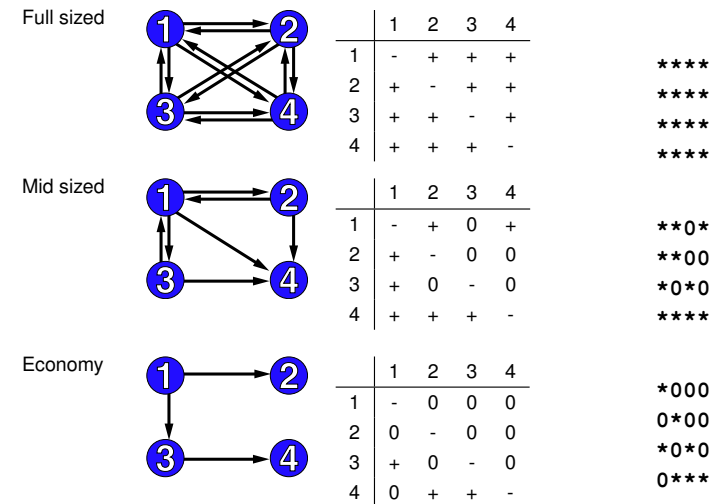


9 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

9

Migration model specification



10 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

10

Model comparison

With a criterium such as likelihood we can compare nested models. Commonly we use a likelihood ratio test (LRT) or Akaike's information criterion (AIC) to establish whether phylogenetic trees are statistically different or mutation models have an effect on the outcome, etc.

Kass and Raftery (1995) popularized the **Bayes Factor** as a Bayesian alternative to the LRT.

$$p(M_1|D) = \frac{p(M_1)p(D|M_1)}{p(D)} \quad \frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)}{p(M_2)} \times \frac{p(D|M_1)}{p(D|M_2)}$$

$$BF = \frac{p(D|M_1)}{p(D|M_2)} \quad LBF = 2 \ln BF = 2 \ln \left(\frac{p(D|M_1)}{p(D|M_2)} \right)$$

The magnitude of BF gives us evidence against hypothesis M_2

$$LBF = 2 \ln BF = z \quad \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

11 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

11

Model comparison

With a criterium such as likelihood we can compare nested models. Commonly we use a likelihood ratio test (LRT) or Akaike's information criterion (AIC) to establish whether phylogenetic trees are statistically different or mutation models have an effect on the outcome, etc.

Kass and Raftery (1995) popularized the **Bayes Factor** as a Bayesian alternative to the LRT.

Posterior Density

Prior

Likelihood

$$p(M_1|D) = \frac{p(M_1)p(D|M_1)}{p(D)} \quad \frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)}{p(M_2)} \times \frac{p(D|M_1)}{p(D|M_2)}$$

$$BF = \frac{p(D|M_1)}{p(D|M_2)} \quad LBF = 2 \ln BF = 2 \ln \left(\frac{p(D|M_1)}{p(D|M_2)} \right)$$

The magnitude of BF gives us evidence against hypothesis M_2

$$LBF = 2 \ln BF = z \quad \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

12 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

12

A total of 70 individuals from 7 populations analyzed for 377 microsatellite loci:
Mutation model is Brownian motion approximation to the single-step mutation model

Structured populations

Model selection

Model order and probability using Bayes factors

1. all other models: 0.0
Minimal model 1.0
- 2.
- 3.
- 4.
- 5.
- 6.
- 7.

Reanalysis of data from Rosenberg et al. Science 2001

21 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

21



Wednesday, February 6, 13

22

Run time concerns

MCMC

MCMC works perfectly fine when run infinitely long. It is rather difficult to know when the (finite) run has converged and is sampling from the distribution of interest **and** is reaching all important parts. Several methods are used to improve convergence and sampling:

- ◆ Improve the proposal procedure
- ◆ Use Metropolis-coupled MCMC to improve finding peaks in the distribution.
- ◆ Program optimization can improve runtime considerably.
- ◆ Run several analyses in parallel



23 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

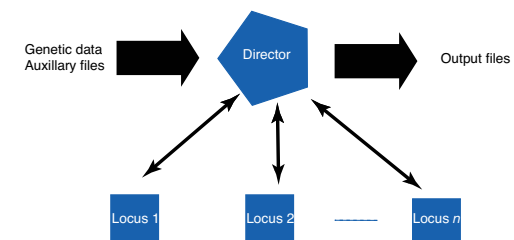
23

Embarrassingly parallel computation

MIGRATE

Each locus is completely independent, therefore can run on a different computer. Embarrassingly simple parallel computing can be done by splitting up data set and gathering “results” from individual nodes by “hand”. This gets really tedious with 100+ loci.

MIGRATE uses a more sophisticated strategy (MPI) and can use a cluster of (loosely) connected computer nodes. With more loci than nodes a load balancing scheme is used.



Beerli (2004)

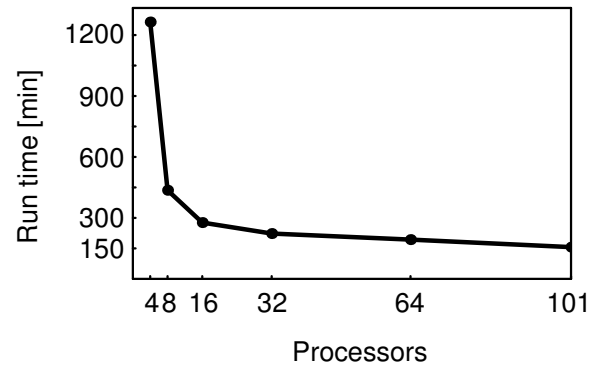
24 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

24

Speed up

Estimation of 9 parameters in a 3 population migration model using data from a total of 100 loci, distributed over 4, 8, 16, 32, 64, 101 computer nodes.



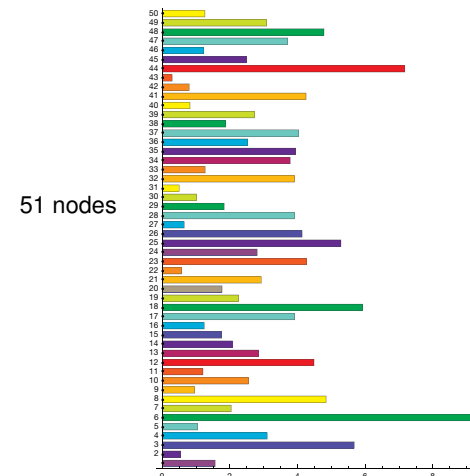
Beerli (2004) Effect of unsampled populations on the estimation of population sizes and migration rates ... 25 of 33 - ©2012 Peter Beerli

Wednesday, February 6, 13

25

Speed up

Speed of total run depends on the "slowest" locus (here out of 50)



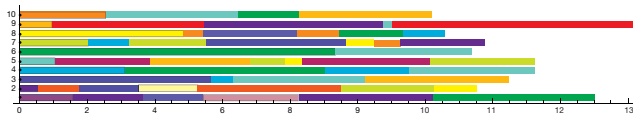
26 of 33 - ©2012 Peter Beerli

Wednesday, February 6, 13

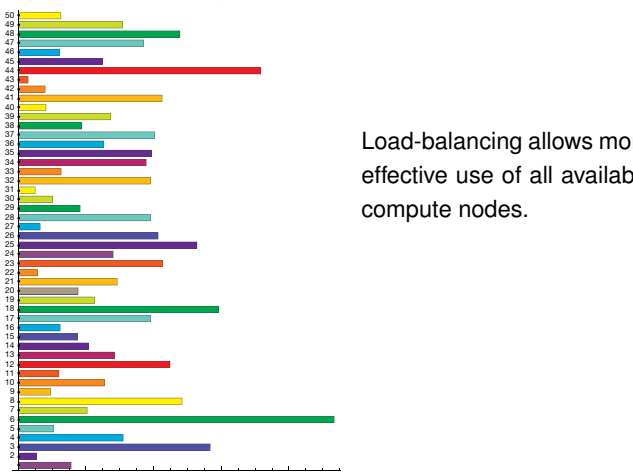
26

Speed up

11 nodes



51 nodes



Load-balancing allows more effective use of all available compute nodes.

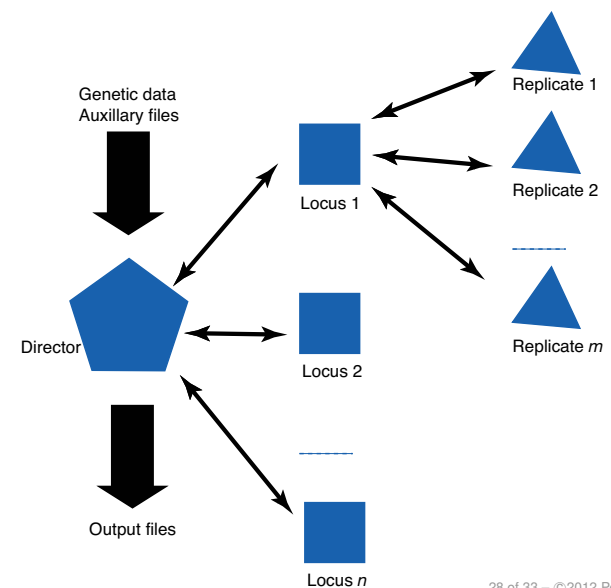
27 of 33 - ©2012 Peter Beerli

Wednesday, February 6, 13

27

Speed up even more?

MIGRATE



MIGRATE 2.2 (2007)

28 of 33 - ©2012 Peter Beerli

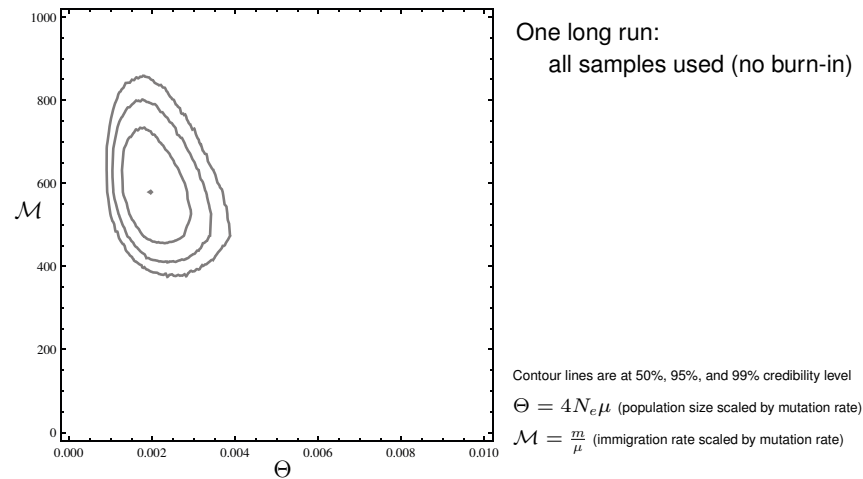
Wednesday, February 6, 13

28

Run time versus accuracy

One long run

Posterior density for a 2-parameter model (population size and gene flow) A run for 50×10^6 steps (sampling 3 quantities: 2 parameters and genealogies) took about 20 hours.



29 of 33 – ©2012 Peter Beerli

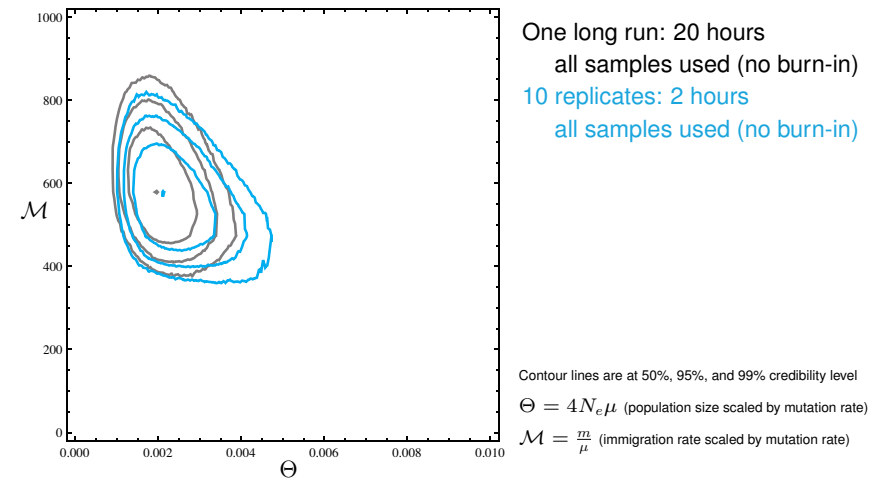
Wednesday, February 6, 13

29

Run time versus accuracy

10 replicated runs

Posterior density for a 2-parameter model (population size and gene flow) 10 runs each for 5×10^6 steps took about 2 hours.



30 of 33 – ©2012 Peter Beerli

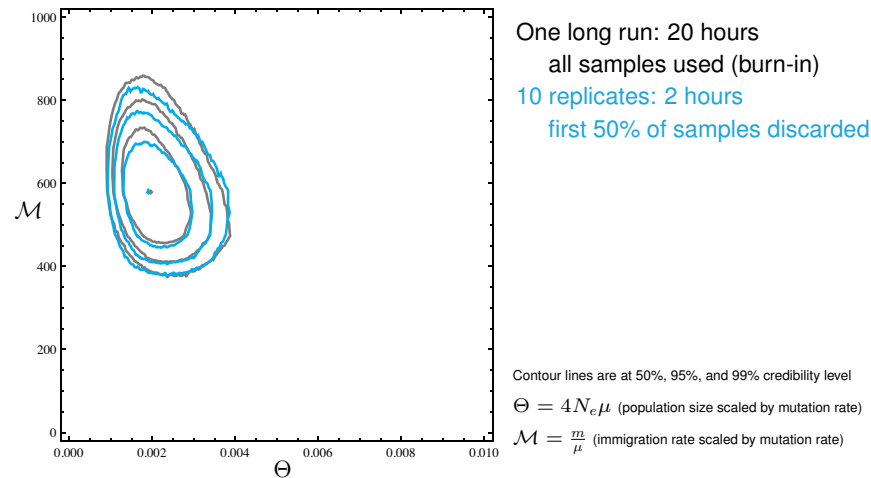
Wednesday, February 6, 13

30

Run time versus accuracy

10 replicated runs*

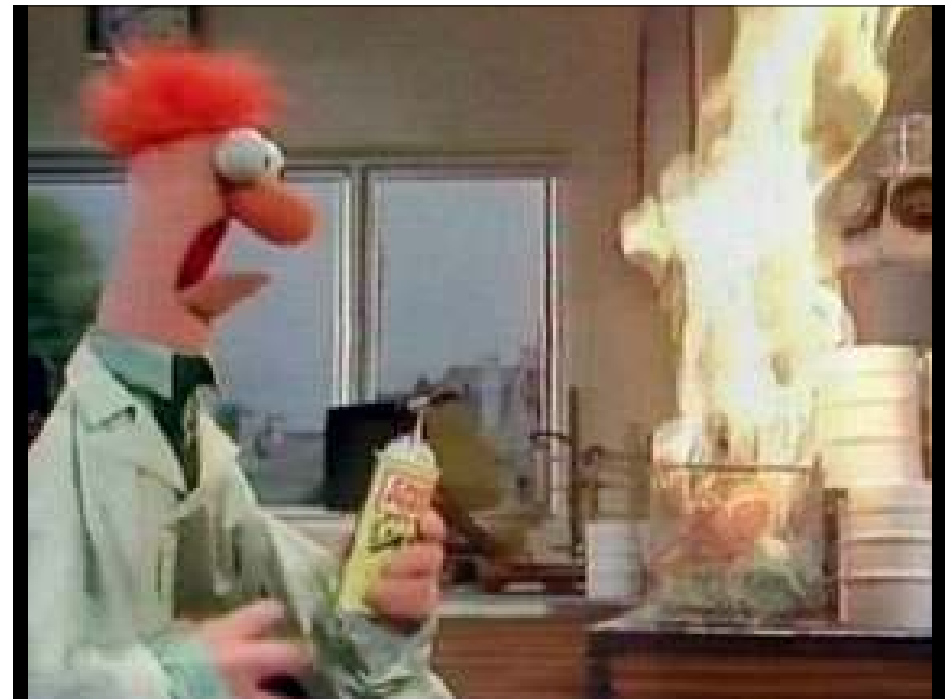
Posterior density for a 2-parameter model (population size and gene flow) 10 runs each for 5×10^6 steps took about 2 hours.



31 of 33 – ©2012 Peter Beerli

Wednesday, February 6, 13

31



Wednesday, February 6, 13

32



*Learn a computer scripting language today to
be ready for tomorrow, the parallel genome
sequencing revolution has begun.*