Which STACKS parameters minimise genotyping error

Emma Carroll PhD
Host: Prof. Oscar Gaggiotti

Marie Skłodowska-Curie Actions

newton
international fellowships

University of St Andrews | FOUNDED 1413

Genotyping error

Implications for many types of studies

- false departure from HWE (Xu et al 2002)

- overestimate inbreeding (Gomes et al 1999)

- impact reliability of population structure and demographic history inferences (Miller et al 2002; Pool et al 2010)

Standard genetic studies

- Recognised problem e.g. chimpanzee paternity errors in Gagneux et al (1997)

- Standard measures to minimise error (e.g. Bonin et al 2004; Morin et al 2010)

- Standard to rerun 10% of samples in microsatellite studies to estimate genotyping error rate

Measuring genotyping error in (dd)RAD

Varied parameters:

-m: minimum number of identical, raw reads required to create a stack: 2, 5, 10

-M: number of mismatches allowed between loci when processing a single individual: 2, 4

-n: number of mismatches allowed between loci: 2, 4

- depth per locus: varied 10x; 20x; 30x

Measuring genotyping error in (dd)RAD

Analysis set up:

- replicate samples plus 'topped up' to 10 samples per pop (n=55)
- run denovo_map.pl; export SNPs found in >75% samples

Measuring error rates:

- SNP error rate: proportion of SNP mismatches between replicate pairs

    all samples: n=15

    high-quality repeats n=6

    low-quality repeats, n=9

- missing loci: proportion of missing loci per replicate pair

    high-quality repeats only

Measuring genotyping error in (dd)RAD

Error rates pretty low

-Overall: ranges from 0.06 – 0.34 % per SNP

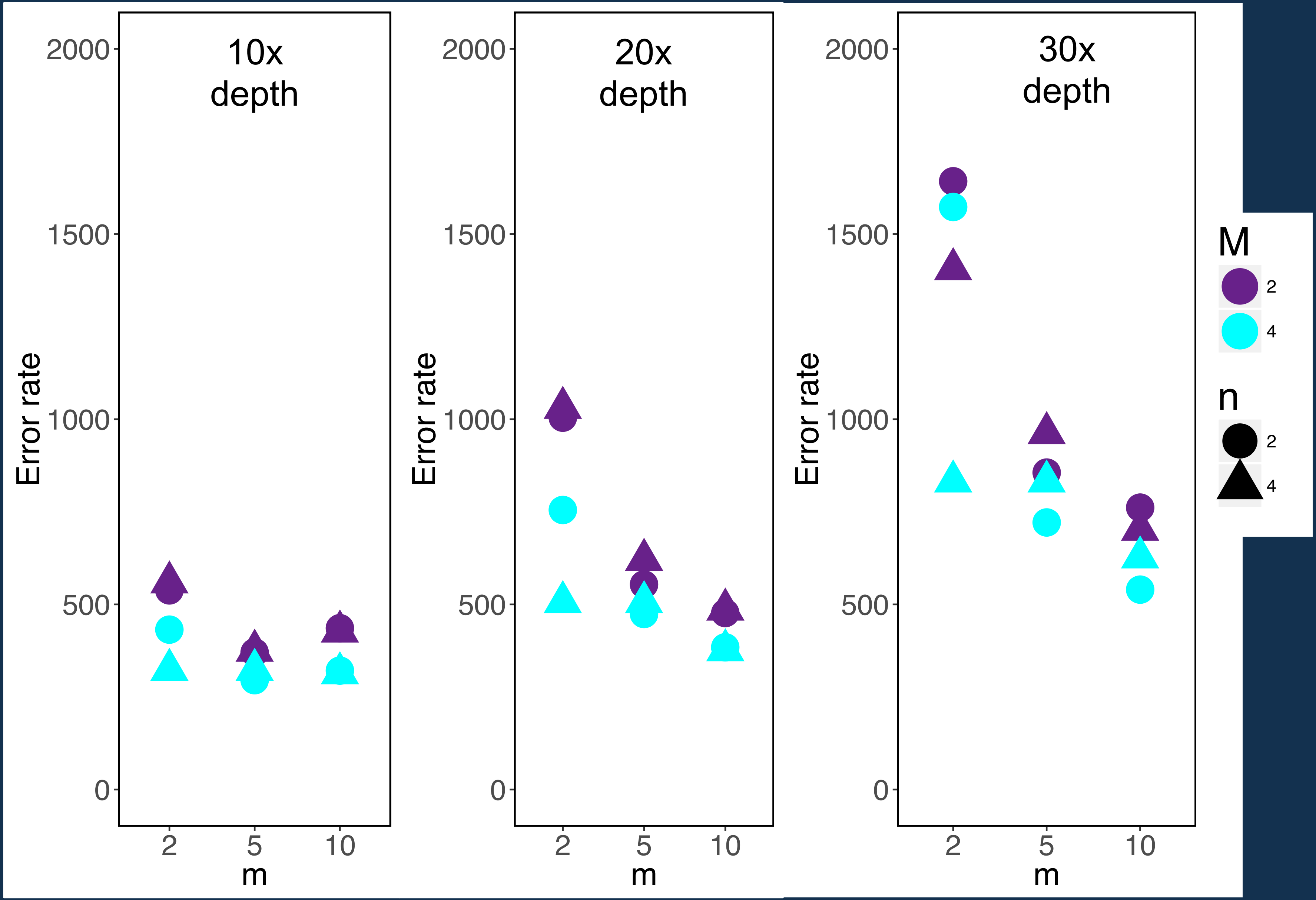
-High quality samples: 0.06 – 0.29% per SNP

-Low quality samples: 0.12 – 0.75% per SNP

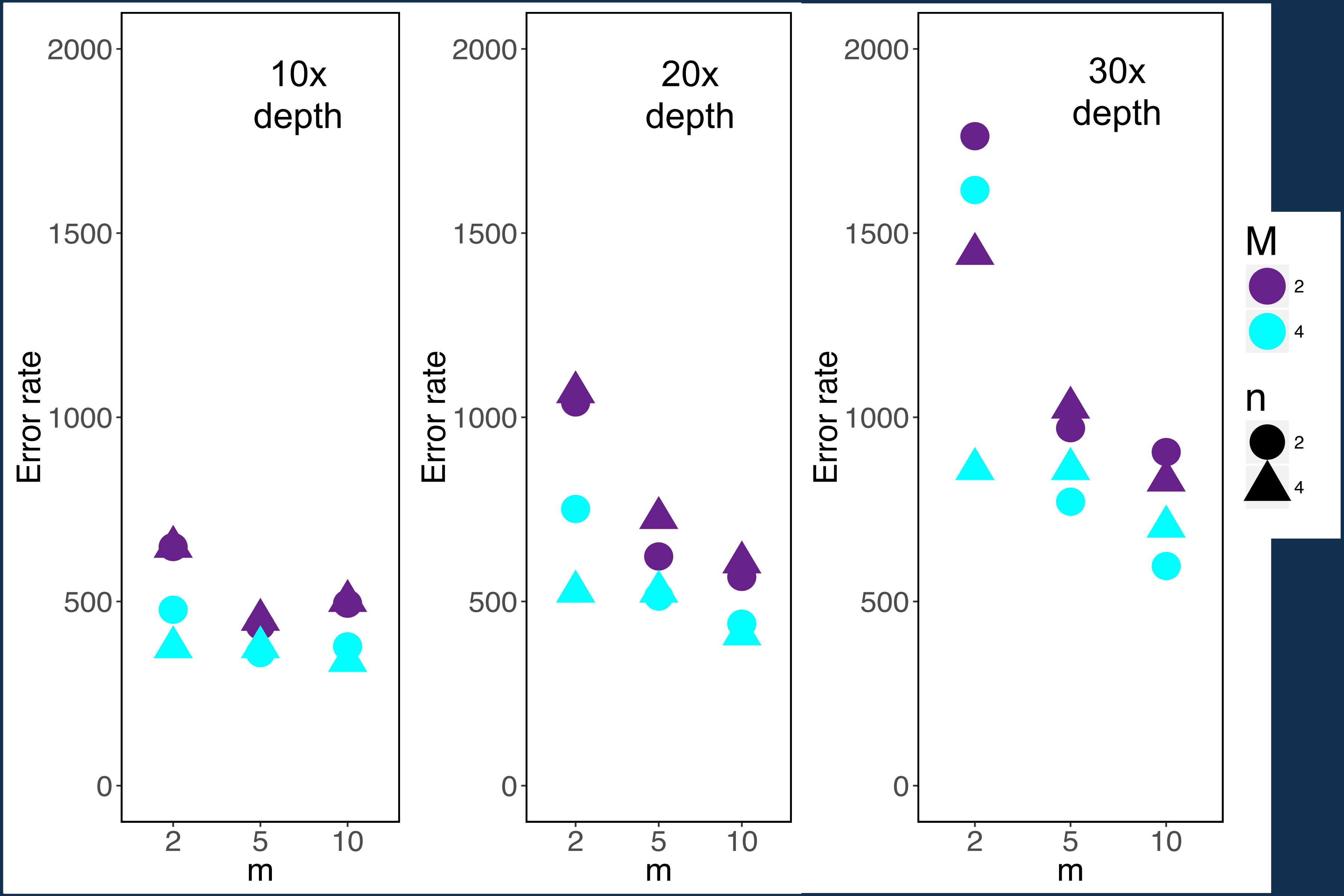
- On average, drop out (one allele per SNP match) 36 x more common than outright error: PCR bias between alleles? PCR error?

-To better visualise these: report average number of SNPs per error (100/per SNP error rate)

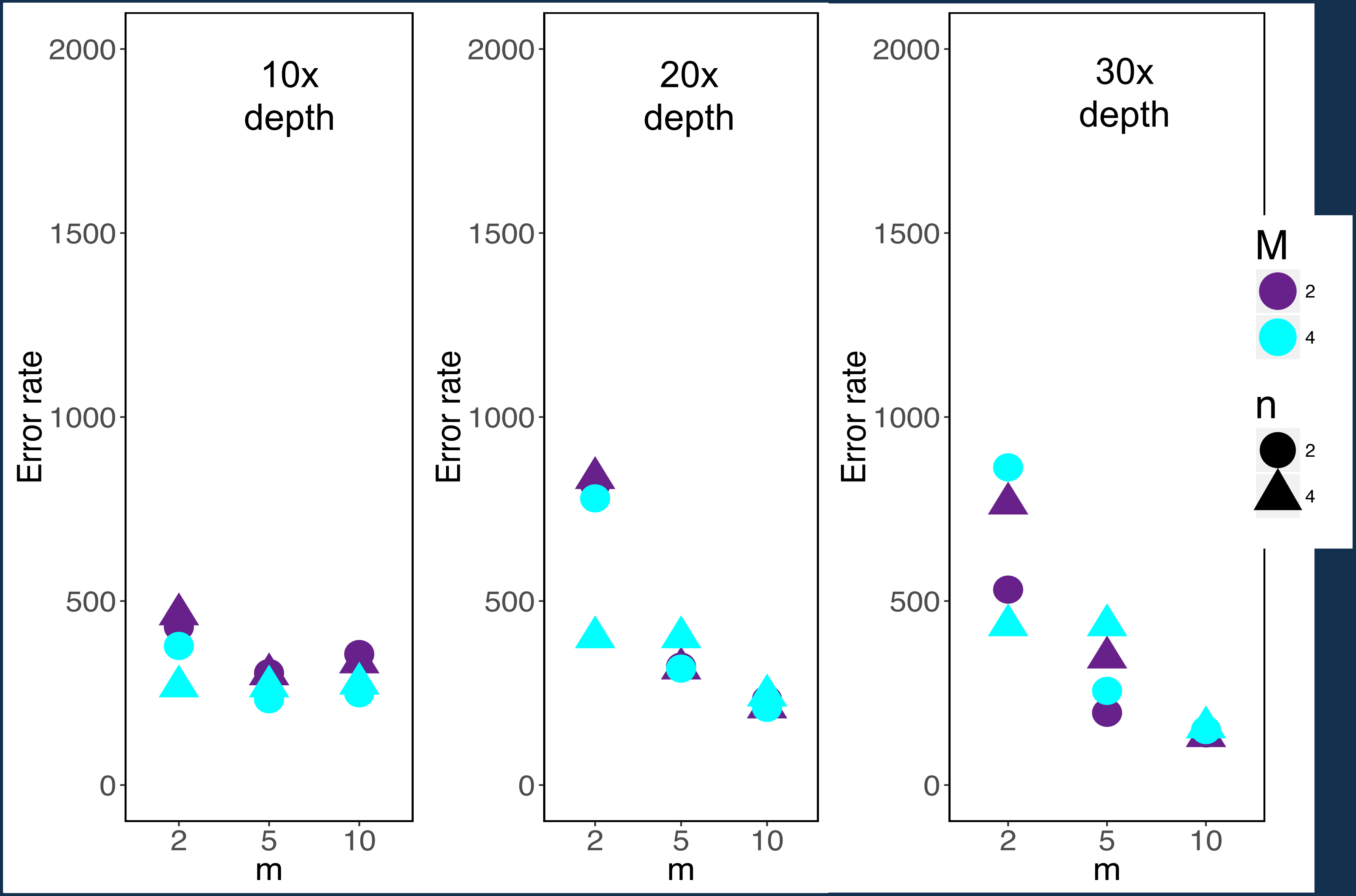
Error rate – Overall



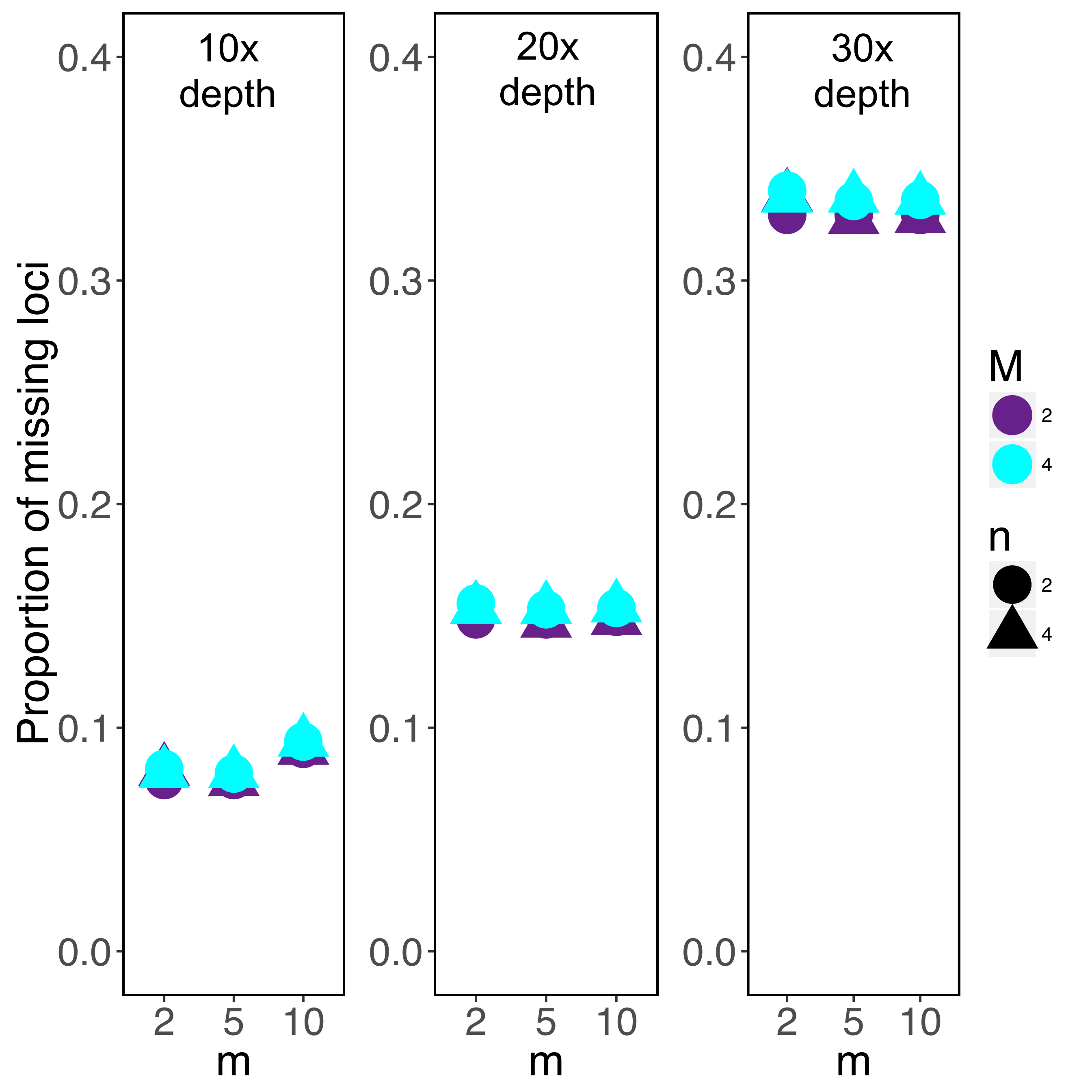
Error rate – High Quality Samples



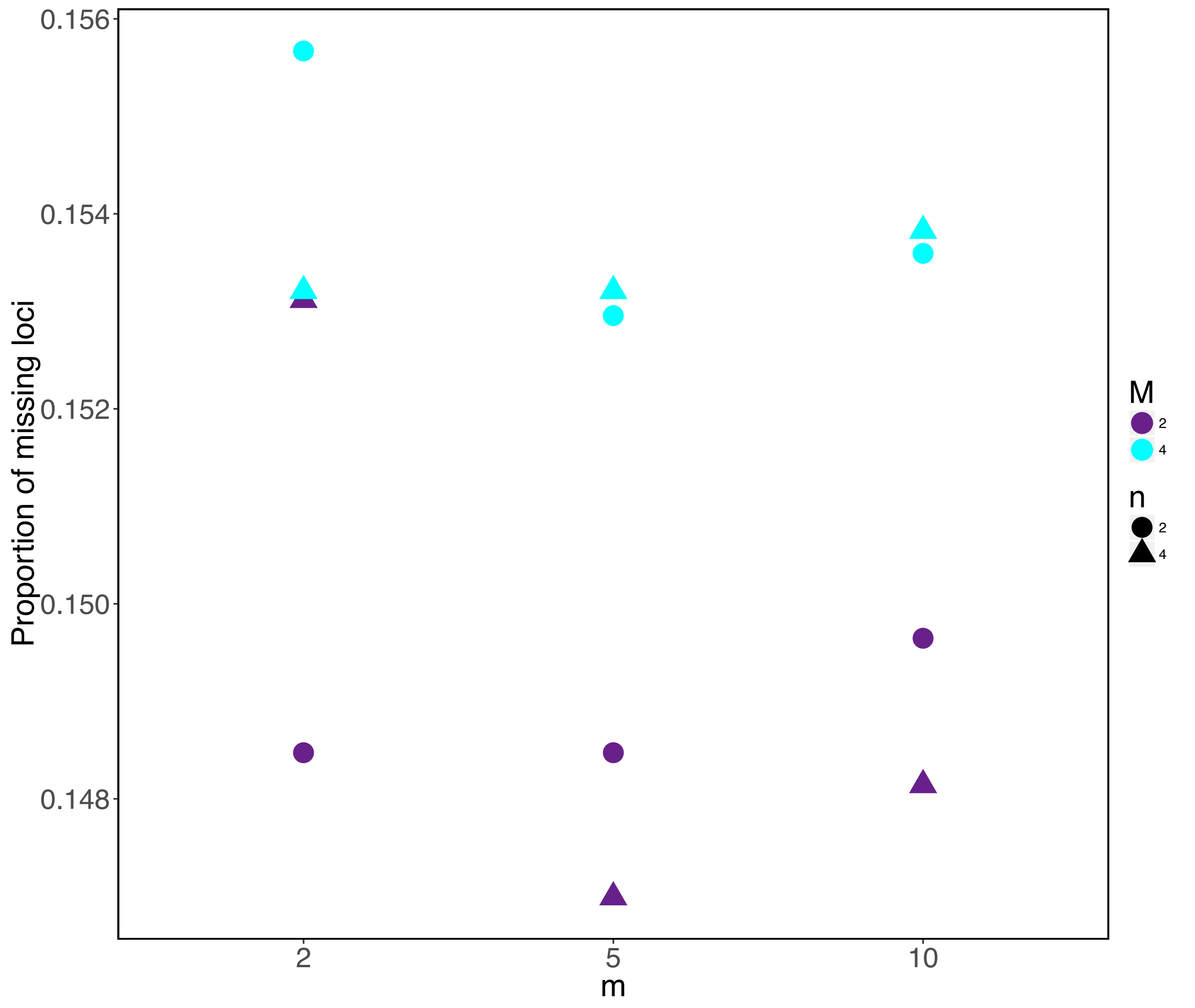
Error rate – Low Quality Samples



Missing loci rate – All



Missing loci – 20x



M=2
n=4
lower proportion
of missing error

lower n
?splitting loci

Minimising genotyping error

Read depth seems to have more impact on missing loci than STACKS parameters (within limits)

-m = 2 – suprisingly, seems to have lowest error rate

-M = 2 – overall, lower missing loci % and error rate

-n = 4 – reduces missing loci % c/w –n =2

Minimising genotyping error

In the literature

- Mastretta-Yanes et al (2015): Varied STACKS parameters to estimate error rate of ddRAD in 11 replicate samples
- SNP error rate 2-12%
- also found trade-off between error rates and missing loci proportion
- Fountain et al (2016): varied quality score used to clean raw reads (process_radtags) and sequence depth of loci used in analysis
- estimated error by looking at departure from Mendelian inheritance in 16 mother-offspring sloth pairs
- error rates declined with depth (10-13 fold decline 5x to 30x)
- ref genome better than denovo