Exploring k-mer spectra analysis for read-level sample clustering in *Brachypodium*

Benjamin Y.H. Bai*

BIOL3157: Advanced Studies Extension 3, Borevitz Lab

Introduction

Experimental design should account for the structure and range of input genetic diversity. An awareness of population structure is important, especially in genome-wide association studies (GWAS), where it is a major confounding factor [1]. Spurious associations between traits and quantitative trait loci (QTLs) can occur by shared evolutionary history rather than any functional link. Genome duplications are especially problematic, as homeologous variation between subgenomes will be associated with all the traits of that individual. GWAS also has low power to detect significant associations when traits are controlled by many QTLs with multiple low frequency alleles–this often occurs with a widely-distributed sample [2].

Methods to control population structure include population restructuring by crossing, or restricting the range of diversity in the analysis to a local sample. There is greater power at the local scale, as the limited genetic diversity results in fewer traits controlled by low-frequency alleles [2]. It is also more reliable to perform genotype imputation on samples with limited diversity [3]. All these techniques assume that samples have been clustered by some measure of genetic distance, and the broad-scale structure of the population is known.

Traditional SNP-based clustering methods fail for non-model organisms where there is no reliable reference. Even with reference-free methods (TASSEL UNEAK [4], STACKS [5]) based on restriction-enzyme techniques such as double digest restriction site-associated DNA sequencing (RAD-seq) and genotyping-by-sequencing (GBS) [6], the focus on SNPs means analysis relies on differences at the haplotypic level. When up to 90% data can be missing at SNP loci due to low coverage [7], samples are extremely diverged, or whole-genome duplications are present, haplotype comparisons become uninformative or impossible as the level of diversity in the experiment increases.

Comparing the k-mer spectra between samples allows us to operate at the read level, where even diverged samples show similarity. We explored the potential for read-level clustering in *Brachypodium* using properties of the k-mer frequency spectrum, focusing of distinguishing samples of differing ploidy, knowledge that is directly applicable to some reference-free SNP-calling techniques [8]. *Brachypodium* is ideal for this investigation due to both due to its importance as a model organism, and the presence of the autopolyploid *B. hybridum* ($2n = 30$), which is derived from the diploids *B. distachyon* ($2n = 10$) and *B. stacei* ($2n = 20$) [9].

The k-mer spectra of simulated whole genome sequencing (WGS), real WGS, and real GBS read datasets of known ploidy were observed, in an attempt to find peak structures corresponding to homeologous variation between subgenomes, although we were hampered by low coverage and high sequencing error rates. We also hierarchically-clustered the profiles of high abundance k-mers in the GBS samples, obtaining moderate separation between ploidies.

***Contact**: u5205339@anu.edu.au **Dates**: Sem. 2, 2014 **Supervisor**: Kevin Murray/Assoc. Prof. Justin Borevitz **Course**: BIOL3157 - ASE3: Genomic classification of *Brachypodium*.

Materials and Methods

Sequence data and data simulation

We used real WGS paired-end libraries from *B. distachyon* (sample code `Bdis05`) and *B. hybridum* (sample code `7424.2.73049.ACAGTG`), and 24 samples from two sets of B. distachyon GBS libraries with sample codes: `bd1` (diploid) and `bd6` (hexaploid). The diploid samples were from recombinant inbred lines; the hexaploid samples come from a wild population where selfing is prevalent; in both cases, low heterozygosity is expected. Library sizes ranged from 0.7 M to 4 M read pairs.

`wgsim` [10] was used to simulate 100 bp paired-end read libraries from the *B. distachyon* reference genome (annotation release version 1.2 [11], repeat-masked using MIPS [12]). Sequencing error was left at the default of 0.02, a conservative value for common next-generation platforms [13]. The polymorphism rate was set at 0.005, a rate that has been detected in *Brachypodium* resequencing projects [14].

For the diploid control, 5 M (million) read pairs were obtained by merging two sets of 2.5 M reads from independent simulations. This library size is towards the high end of the GBS libraries we wish to analyse. For instance, the mean library size of the `bd1` GBS samples was 1.16 M read pairs, with a maximum size of 5.89 M pairs. To simulate a polyploid read set partially derived from a subgenome with markedly differing composition, we raised the polymorphism rate to 0.05 for half of the reads.

K-mer counting

K-mer hashing and abundance counting was performed using `khmer` [15]. It uses a probabilistic data structure called a Count-Min sketch, where the false positive rate–chance of identifying a k-mer that is not present in a sample as present–increases with the number of unique k-mers, and decreases with the specified memory usage [16]. We aimed for false positive rates of less than 0.01.

Hashing was performed at $k = 25$ for simulated and real WGS datasets, with a specified memory usage of 8 GB per sample, giving a false positive rate of 0.001-0.004 (3 s.f.). Hashing was performed at $k = 12$ for the GBS datasets, with a specified memory usage of 400 Mb per sample, giving a false positive rate of 0.000-0.000 (3 s.f.).

High abundance k-mer profiles

We examined the repeat profiles of twelve samples from each GBS library by counting high-abundance k-mers. After choosing (arbitrarily) the sample with the largest library size as the reference, we ranked its k-mers by abundance, and selected a set of high abundance k-mers as the reference set. The abundances of the reference set were determined across all samples, then all counts were normalised by the respective library sizes. Hierarchical clustering was performed on the resulting counts matrix using the Euclidean distance measure (`http://stat.ethz.ch/R-manual/R-patched/library/stats/html/dist.html`).

Results and Discussion

Simulating subgenomes in GBS-scale read datasets

Figure 1 shows the k-mer frequency spectra for three simulated WGS datasets. In a typical WGS k-mer spectrum, we would expect a $1X$ peak centered at the k-mer coverage C_k, representing k-mers that occur once in the genome [17]. C_k is calculated via equations 1 and 2, where n is the number of reads, L is the read length (bp), N is the genome size (bp), C is the per-base coverage, and k is the k-mer length.

$$C = nL/N \tag{1}$$

$$C_k = C(L - k + 1)/L \tag{2}$$

The curve for the simulated polyploid was lower than the curve for the diploid, as the higher polymorphism rate generates more unique k-mers (486 M compared to 439 M) leading to lower frequencies. The apparent lack of peak structure compared to a typical WGS k-mer spectrum is attributed to the low $C_k = n(L - k + 1)/N = 10^7(100 - 25 + 1)/272Mb = 2.79$. Any $1X$ peak centered at 2.79 is swamped by the error peak, as are any homeologous subpeaks of lower

abundance. Zooming in on the region of the curve at the expected abundance is uninformative (Fig. 1b).

Sequencing error has the general effect of shifting the spectrum leftwards [17], but makes little difference in this case. Figure 1 (black line) also shows the resulting spectrum when sequencing error is set to 0.00, which has higher frequencies than either of the other simulations due to the smaller error peak, but still no $1X$ peak structure. Further simulation at higher coverage ($> 30X$) is required as a positive control, but discerning the presence of subgenomes using peak structure at GBS coverage levels is difficult regardless, even with a perfect sequencing platform.

A significant proportion of unique k-mers had high abundance, which may correspond to unmasked repetitive elements or common sequence motifs.

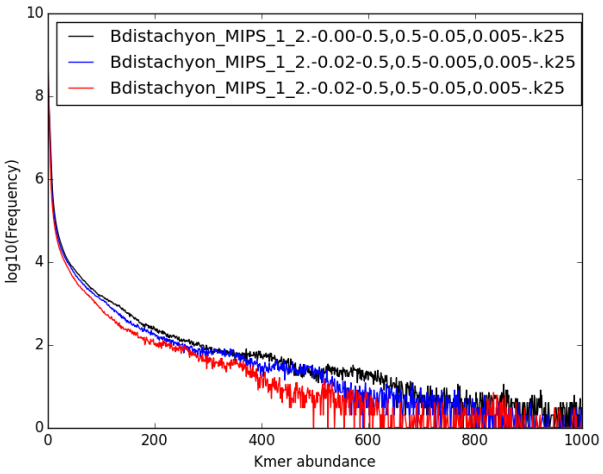
Real WGS and GBS datasets

We also analysed real WGS data from samples of known ploidy. Sample code `Bdis05-1` was a diploid *B. distachyon*, with 9.7 M read pairs and $C_k = 5.42$. Sample code `7424.2.73049.ACAGTG` was an allotetraploid *B. hybridum*. The larger genome size of 1.265 pg/2C [9] means 6 M read pairs yield only $C_k = 1.47$.

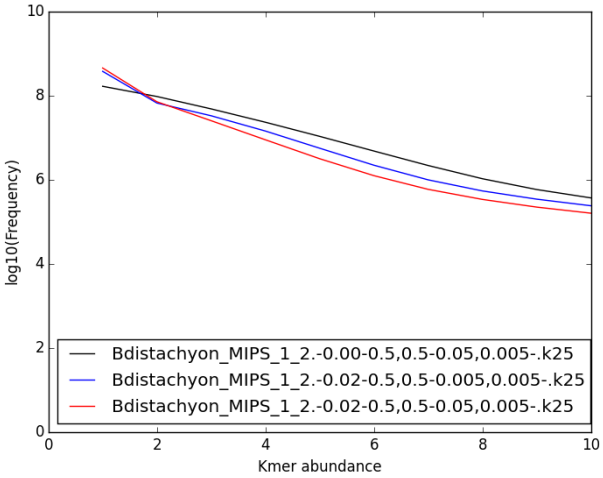
Once again, there is no peak structure near the expected coverage (Fig. 2b). Peaks occurring at very high abundance ($> 500X$) despite low coverage probably represent LTR retrotransposons [11] and contaminating organelle DNA. The coverage difference between the samples means the curves in Figure 2 are not directly comparable.

From the `bd1` and `bd6` GBS datasets, we selected pairs of samples with comparable library sizes. The largest of these pairs had approx. 4 M read pairs. Since the value of $C_k$does not change substantially with varying k (Fig. 3), we hashed at $k = 12$ for improved memory usage. As there are far fewer 12-mers than 25-mers, we could achieve a negligible false positive rate with only 400 Mb per sample, compared to the 8 GB required to hash the 25-mers in 5 M simulated reads (see Materials and Methods).

The `bd6` library contained the greater number of unique k-mers (6.7 M vs. 6.2 M of a possible $4^{12} =$ 16.8 M), as polyploids tend to have homeologous variation between their subgenomes. Thus `bd6` produced

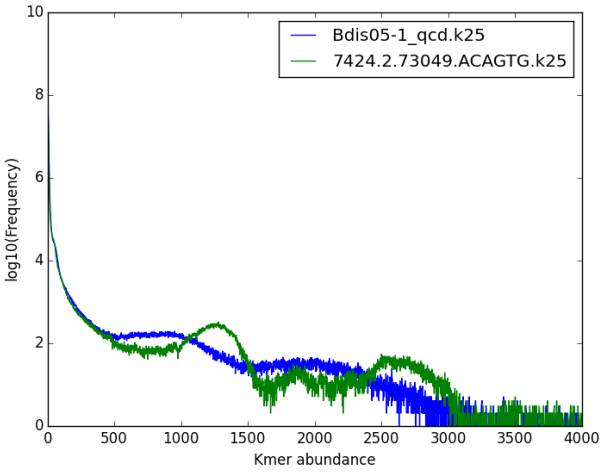


(a)

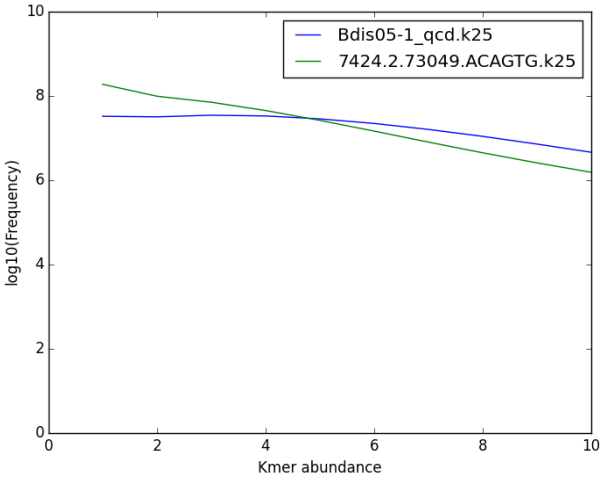


(b)

Figure 1: K-mer frequency spectra ($k = 25$) for three sets of 5 M read-pair WGS data simulated from the *B. distachyon* reference, with varying sequencing error and polymorphism rates. K-mer abundance is the number times a k-mer occurs in the reads; frequency is the total occurrence count over all unique k-mers of that abundance. The naming format for the legend is `Bdistachyon_MIPS_1_2.-<e>-0.5,0.5-<r1>,<r2>-.k25`, where `<e>` is the sequencing error rate, and `(<r1>, <r2>)` are the subgenome polymorphism rates. Figure 1b shows the region of Figure 1a around the expected $C_k =$ 2.79.



(a)



(b)

Figure 2: K-mer frequency spectra ($k = 25$) for real WGS data. `Bdis05-1`: diploid *B. distachyon* (9.7 M read pairs, $C_k = 5.42$). `7424.2.73049.ACAGTG`: (6 M read pairs, $C_k = 1.47$). Figure 2b shows the region of Figure 2a around the expected k-mer coverage.

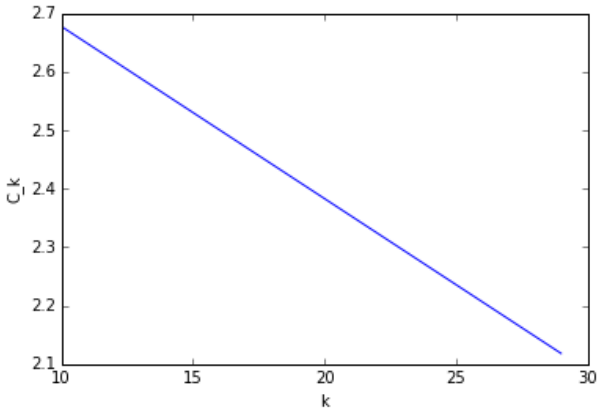


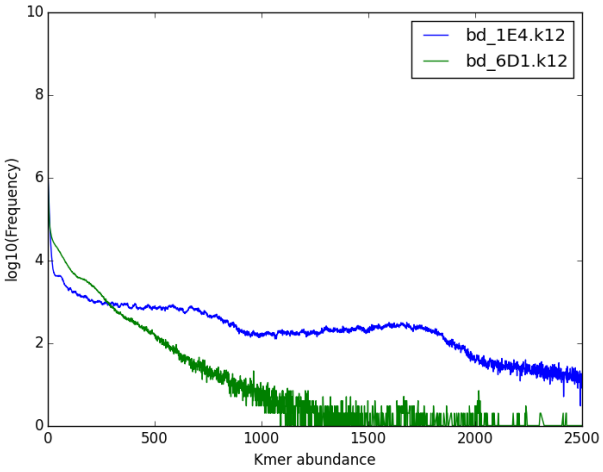
Figure 3: Plot of expected k-mer coverage C_k of $n = 4$ M read pairs ($L = 100bp$) over the $N = 272Mb$ *B. distachyon* vs. choice of k. Even at $k = 10$, C_k remains low enough for the $1X$ to be confounded by the error peak.

the lower curve at high abundances, with a corresponding higher proportion of low abundance k-mers (Fig. 4). The analysis of GBS data suffers from the same low $C_k = 2.24$, compounded by high error rate due to PCR in library preparation [18, 6], so looking at peak structure is once again impractical (Fig. 4b).

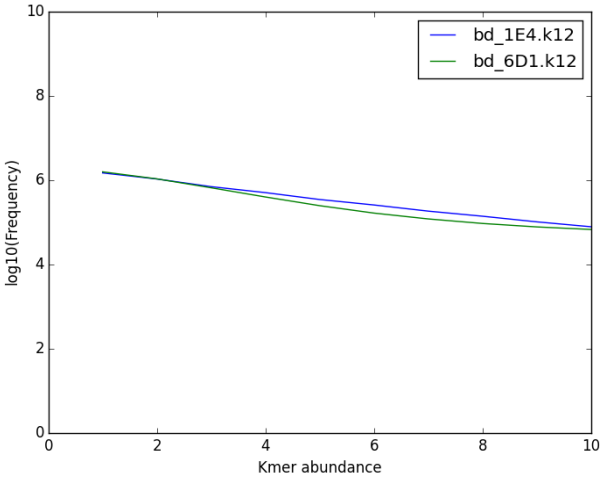
Analysing the k-mer repeat profile in GBS datasets

For such low coverage libraries, signal separating samples of differing ploidy is most readily discernible at the high abundance end of the spectra (Fig. 1a, 2a, 4a). The high abundance k-mers in GBS libraries are more likely to represent repetitive sequences derived from reads clustered around restriction sites rather than organelle DNA, thus their counts may be directly related to the nuclear genotype of the sample.

The most abundant 12-mers are likely to be low-complexity elements such as homopolymer repeats shared by many samples. To produce a more informative counts matrix, we discarded the 100 most abundant 12-mers, then used the next 200 as the reference set. Eighty-three percent of the reference 12-mers were present in the samples on average. The dendro-



(a)



(b)

Figure 4: K-mer frequency spectra ($k = 25$) for real WGS data. `Bdis05-1`: diploid *B. distachyon* (9.7 M read pairs, $C_k = 5.42$). `7424.2.73049.ACAGTG`: (6 M read pairs, $C_k = 1.47$). Figure 4b shows the region of Figure 4a around the expected k-mer coverage.

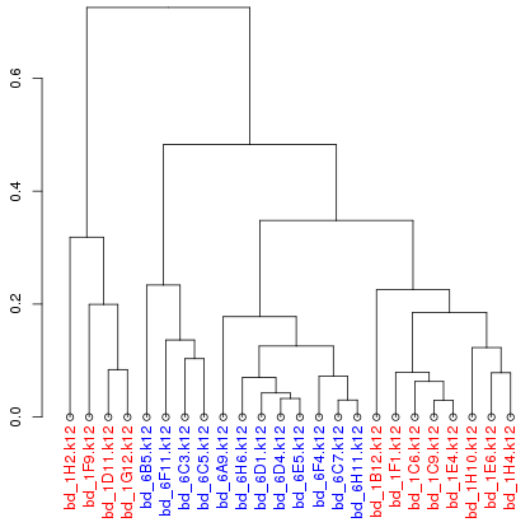


Figure 5: Dendrogram from hierarchical clustering of the normalised counts of 200 high abundance 12-mers over 24 diploid (`bd1`) and hexaploid (`bd6`) GBS read datasets.

gram from hierarchical clustering is shown in Figure 5. Two-thirds of the samples clustered by ploidy as expected, but two outgroups consisting of four samples from each ploidy group are observed. Interpretation is difficult without any knowledge of the biology behind these sample codes. It is possible that 83% mean presence is still too high to distinguish between ploidy groups reliably. One possibility is to look at medium abundance k-mers where count variation between samples is more likely.

Conclusions

GBS data is inherently error prone, low coverage, and non-uniformly distributed across the genome. Even with no sequencing error, low C_k makes it extremely difficult to differentiate GBS sample ploidy based on peak structure. K-mer spectra typically become useful only at $> 30X$ coverage, where there is sufficient distinction from the error peak. Conversely, the small

datasets allow for fast and memory efficient k-mer hashing.

Higher abundance k-mer profiles show more promise. In WGS datasets, high abundance organelle DNA peaks are significant (Fig. 2). This is not an issue for GBS datasets; in fact, high abundance k-mer counts may correlate with the presence-absence genotype of the chosen restriction sites. Experimentation with different k-mer reference sets drawn from different quantiles of the abundance distribution may lead to lower mean presence proportions, hopefully producing a more informative distance matrix.

Outgroups in the dendrogram need to be checked against sample biology. With further information about the identity of the various samples, we can also attempt to distinguish technical replicates, clonal groups, and geographic population structure.

Software and data sources

Pipelining, data analysis and plotting was done in Bash, Python 2.7, and R 3.0. All scripts are available under the the MIT License at `https://github.com/digitase/brachy_khmer`, and a usage demo can be found in the documentation at the same location.

- The `khmer` version used was `1.2-rc2-6-gc5dee21`, cloned from `https://github.com/ged-lab/khmer`.
- We used `wgsim` 0.3.0 from `samtools` library.
- Reference genome, WGS and GBS datasets were kindly provided by the Borevitz Lab (`http://borevitzlab.anu.edu.au/`), Research School of Biology, Australian National University.
- Computational resources were provided by the Borevitz Lab and their collaborators, as well as the Genome Discovery Unit, John Curtin School of Medical Research, Australian National University.

Acknowledgments

I would like to thank the members of the Borevitz Lab, especially Kevin Murray, for his expertise and patient guidance.

References

[1] Xu, H. and Shete, S. (2005). 'Effects of population structure on genetic association studies.' *BMC genetics* **6**(Suppl 1):S109.

[2] Brachi, B., Morris, G.P., and Borevitz, J.O. (2011). 'Genome-wide association studies in plants: the missing heritability is in the field.' *Genome Biol* **12**(10):232.

[3] Marchini, J. and Howie, B. (2010). 'Genotype imputation for genome-wide association studies.' *Nature Reviews Genetics* **11**(7):499–511.

[4] Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S., and Costich, D.E. (2013). 'Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol.' *PLoS genetics* **9**(1):e1003215.

[5] Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., and Cresko, W.A. (2013). 'Stacks: an analysis tool set for population genomics.' *Molecular ecology* **22**(11):3124–3140.

[6] Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). 'A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.' *PloS one* **6**(5):e19379.

[7] Fu, Y.B. (2014). 'Genetic Diversity Analysis of Highly Incomplete SNP Genotype Data with Imputations: An Empirical Assessment.' *G3: Genes| Genomes| Genetics* **4**(5):891–900.

[8] Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. (2012). 'Double digest RADseq: an inexpensive method for de

novo SNP discovery and genotyping in model and non-model species.' *PloS one* **7**(5):e37135.

[9] López-Alvarez, D., López-Herranz, M.L., Betekhtin, A., and Catalán, P. (2012). 'A DNA barcoding method to discriminate between the model plant Brachypodium distachyon and its close relatives B. stacei and B. hybridum (Poaceae).' *PloS one* **7**(12):e51058.

[10] Li, H. (2011). 'wgsim-Read simulator for next generation sequencing.'

[11] Vogel, J.P., Garvin, D.F., Mockler, T.C., Schmutz, J., Rokhsar, D., *et al.* (2010). 'Genome sequencing and analysis of the model grass Brachypodium distachyon.' *Nature* **463**(7282):763–768.

[12] Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H., and Spannagl, M. (2013). 'MIPS PlantsDB: a database framework for comparative plant genome research.' *Nucleic acids research* **41**(D1):D1144–D1151.

[13] Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). 'A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.' *BMC genomics* **13**(1):341.

[14] Catalan, P., Chalhoub, B., Chochois, V., Garvin, D.F., Hasterok, R., *et al.* (2014). 'Update on the genomics and basic biology of Brachypodium: International Brachypodium Initiative (IBI).' *Trends in Plant Science* **19**(7):414 – 418. URL http://www.sciencedirect.com/science/article/pii/S1360138514001198.

[15] Crusoe, M.R., Edvenson, G., Fish, J., Howe, A., McDonald, E., *et al.* (2014). 'The khmer software package: enabling efficient sequence analysis.' .

[16] Zhang, Q., Pell, J., Canino-Koning, R., Howe, A.C., and Brown, C.T. (2013). 'These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure.' *arXiv preprint arXiv:1309.2975* .

[17] Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., and Fan, W. (2013). 'Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects.' *arXiv preprint arXiv:1308.2012* .

[18] Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009). 'Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes.' *Nature methods* **6**(4):291–295.