

1 ACEseq – allele specific copy number estimation from whole genome sequencing

2

3 Kortine Kleinheinz<sup>1,2</sup>, Isabell Bludau<sup>2,3</sup>, Daniel Hübschmann<sup>1,2,4</sup>, Michael Heinold<sup>1,2</sup>, Philip  
4 Kensche<sup>2</sup>, Zuguang Gu<sup>2,5</sup>, Cristina López<sup>6</sup>, Michael Hummel<sup>7</sup>, Wolfram Klapper<sup>8</sup>, Peter  
5 Möller<sup>9</sup>, Inga Vater<sup>10</sup>, Rabea Wagener<sup>6</sup>, ICGC MMML-Seq project, Benedikt Brors<sup>11</sup>, Reiner  
6 Siebert<sup>6</sup>, Roland Eils<sup>1,2,5\*</sup>, Matthias Schlesner<sup>2,12\*</sup>

7

8 1. Department for Bioinformatics and Functional Genomics, Institute for Pharmacy and Molecular  
9 Biotechnology (IPMB) and BioQuant, Heidelberg University, Germany

10 2. Division of Theoretical Bioinformatics (B080), German Cancer Research Center (DKFZ), Heidelberg,  
11 Germany

12 3. Present address: Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich,  
13 Switzerland.

14 4. Department of Pediatric Immunology, Hematology and Oncology, University Hospital Heidelberg, Germany

15 5. Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center (DKFZ),  
16 Heidelberg, Germany.

17 6. Institute of Human Genetics, University of Ulm and University Hospital of Ulm, Ulm, Germany

18 7. Institute of Pathology, Charité - University Medicine Berlin, Berlin, Germany.

19 8. Pathology, Haematopathology Section and Lymph Node Registry, Christian-Albrechts-University Kiel and  
20 University Hospital Schleswig-Holstein, Kiel, Germany

21 9. Institute of Pathology, University of Ulm and University Hospital of Ulm, Ulm, Ulm, Germany.

22 10. Institute of Human Genetics, Christian-Albrechts-University, Kiel, Germany

23 11. Division of Applied Bioinformatics (G200), German Cancer Research Center (DKFZ), Heidelberg, Germany.

24 12. Bioinformatics and Omics Data Analytics (B240), German Cancer Research Center (DKFZ), Heidelberg,  
25 Germany

26 \* corresponding authors

27

28 *ACEseq is a computational tool for allele-specific copy number estimation in tumor genomes*  
29 *based on whole genome sequencing. In contrast to other tools it features GC-bias correction,*  
30 *unique replication timing-bias correction and integration of structural variant (SV) breakpoints*  
31 *for improved genome segmentation. ACEseq clearly outperforms widely used state-of-the art*  
32 *methods, provides a fully automated estimation of tumor cell content and ploidy, and*  
33 *additionally computes homologous recombination deficiency scores.*

34

35 Copy number aberrations (CNAs) play an important role in tumorigenesis and are often used  
36 to subgroup cancer entities. Whole genome sequencing (WGS) identifies CNAs at  
37 unprecedented resolution, but poses challenges to CNA calling algorithms such as non-  
38 random errors and coverage biases<sup>1</sup>. Changing degrees of genomic complexity, tumor  
39 heterogeneity, varying tumor cell content (TCC) and aneuploidy are further challenges when  
40 analyzing tumor genomes.

41 Many modern tools combine tumor/control coverage ratios with B-allele frequencies (BAF) of  
42 heterozygous SNPs<sup>2,3</sup>. Some tools correct for GC bias, a major source of noise in the  
43 coverage signal<sup>4</sup>, and allow for the incorporation of SV breakpoints to assist segmentation<sup>5</sup>.  
44 However, to the best of our knowledge, none of the available tools provides all of the above-  
45 mentioned features.

46 Here, we present ACEseq, a tool to estimate absolute allele-specific copy numbers on WGS  
47 data. ACEseq involves coverage bias correction, genome segmentation allowing the  
48 incorporation of previously known breakpoints, TCC and ploidy estimation, and absolute  
49 allele-specific copy number calculation to enable fully automated CNA calling on cancer WGS  
50 data without prior information requirements.

51 The first step of ACEseq performs coverage bias correction, which significantly reduces noise  
52 levels (Figure 1). Noisy coverage profiles as depicted in Figure 1A cause over-segmentation  
53 and can mask CNAs. While GC bias correction greatly reduces noise, a remaining fluctuation  
54 of the signal is still observed in the shown sample (Figure 1B). This fluctuation can be  
55 attributed to replication timing coverage bias, which is particularly prominent in fast-replicating  
56 tumors<sup>6</sup> (Supplementary Figure 1). Due to cells in S-phase these samples show a higher  
57 average coverage in early replicating regions than late replicating regions, as the fraction of  
58 cells with already replicated DNA at early loci is higher. Correction for replication timing bias  
59 further smoothens the coverage profile considerably, enabling more robust genome  
60 segmentation in the next step (Figure 1C).

61 During the coverage bias correction steps ACEseq records statistical parameters that carry  
62 technical and biological information about the sequenced samples. Both bias correction steps  
63 are based on loess curve fitting (Supplementary Methods). Slope and curvature of the fitted  
64 GC curve (Figure 1D) indicate the magnitude of the bias. Differences in GC bias between a  
65 tumor sample and its matched healthy control indicated by these quality metrics likely affect  
66 sensitivity and specificity of the variant calling procedures for mutation types like insertions  
67 and deletions (INDELs) and single nucleotide variants (SNVs) due to differences in coverage<sup>1</sup>.  
68 The full width half maximum (FWHM) captures the evenness of coverage (Figure 1E,  
69 Supplementary Methods). For the vast majority of analyzed samples the FWHM decreases  
70 drastically with GC bias correction (Figure 1F,  $\phi_{\text{reduction}}=28\%$ ) and even further with additional  
71 replication-timing correction (Figure 1G,  $\phi_{\text{reduction}}=6\%$ ). The FWHM after bias corrections  
72 indicates remaining coverage fluctuations and hence serves as direct quality parameter for  
73 CNA calling. Notably, it also helps to assess the quality of sequencing libraries, a feature that  
74 has been used routinely by the International Cancer Genome Consortium PanCancer  
75 Analysis of Whole Genomes (ICGC PCAWG) community<sup>7</sup>. Experimental proof for the  
76 reliability of the slope from the loess curve fitted for replication-timing correction (Figure 1H)  
77 as estimator of the tumor proliferation rate could be demonstrated with KI-67 estimates,  
78 where we could show a significant correlation (n=147 germinal center derived B-cell  
79 lymphomas, p-value < 0.01, Figure 1I).

80 We often observed extremely noisy coverage profiles in matched controls from projects  
81 outside the ICGC MMML-Seq, possibly due to wrong handling of blood samples, preventing  
82 accurate copy number calls based on tumor/control ratios. For such samples ACEseq offers  
83 an option to replace the coverage signal from the matched control with an independent  
84 control whilst still maintaining the BAFs of the matched control. This control replacement  
85 option enables full analysis of these sample pairs including reliable discrimination between  
86 runs of homozygosity (ROH) in the germline and somatic loss of heterozygosity (LOH).  
87 Furthermore ACEseq can be run without matched control enlarging the spectrum of samples  
88 that can be processed.

89 Parallel to coverage correction SNPs are haplotype-phased to increase the sensitivity of  
90 allelic imbalance detection. Subsequently, ACEseq segments the genome based on changes  
91 in the BAF and tumor/control coverage ratio. Previously known breakpoints from SV calling  
92 algorithms such as DELLY<sup>8</sup> or SOPHIA (manuscript in preparation) can be incorporated.

93 Resulting raw genomic segments are clustered and merged when indicated to reduce  
94 oversegmentation (Supplementary Figure 2, Supplementary Methods).

95 Final segments are used for TCC, ploidy and copy number estimation. The final copy number  
96 data are visualized for inspection and validation of the analysis (Supplementary Figure 4).

97 Additionally, genomic measures well known to be significantly associated with homologous  
98 recombination (HR) defects are computed: the homologous recombination deficiency (HRD)-,  
99 the large scale transition (LST)<sup>9</sup>- and the telomeric allelic imbalances (TAI)<sup>10</sup>-score. A  
100 connection between these parameters and treatment response to platinum containing  
101 neoadjuvants and poly-ADP-ribose polymerase (PARP) inhibition has been recently  
102 indicated<sup>11-14</sup>.

103 To evaluate ACEseq's performance we compared it to ABSOLUTE<sup>3</sup> (SNP array) as well as  
104 TITAN<sup>15</sup> and THetA<sup>16</sup> (WGS) using 11 B-cell lymphoma samples from the ICGC MMML-seq  
105 project<sup>17</sup> selected based on the availability of SNP array data from germline DNA. First, we  
106 compared ploidy and TCC predictions of the tools (Supplementary Table 1). Fluorescence *in*  
107 *situ* hybridization (FISH) analyses were taken as gold standard for ploidy estimations. Here,  
108 ACEseq, TITAN and ABSOLUTE showed very similar concordance with FISH based ploidy  
109 assessments. Since no gold standard for TCC was available, a comparison with an  
110 orthogonal method was used: the median mutant allele frequencies (MAF) from somatic  
111 SNVs in diploid balanced regions (Supplementary Methods). While this measure can be  
112 affected by the presence of subclonal SNVs, it can be considered as a lower boundary for the  
113 true TCC. We observed that ACEseq was able to predict the TCC with highest accuracy  
114 compared to the other tools based on the number of samples deviating less than 10% from  
115 MAF-based estimates (Supplementary Table 1).

116 Next we compared fractions of the genome with copy number gain and loss (Figure 2). Most  
117 tools reported similar fractions of gains and losses with the exception of THetA. THetA  
118 deviated strongly from the other methods in several samples, probably due to strong  
119 differences in TCC estimations. TITAN and ABSOLUTE only deviated from the ACEseq  
120 results in one sample each. A further investigation of these revealed that sample 4121361  
121 was estimated at much higher TCC by ABSOLUTE, which requires a larger change in  
122 coverage for a segment to be called as gain. The other sample (4112512), called with higher  
123 fraction of gains by TITAN, was strongly affected by replication timing bias. Though ploidy and  
124 TCC were estimated at similar levels, the concordance of allelic as well as total copy number  
125 level was very low (Supplementary Table 2). TITAN and THetA showed a considerably higher

126 number of segments than ACEseq (factor 2-5x higher) for this particular sample, suggesting  
127 that replication timing-dependent coverage bias led to oversegmentation. Resulting small  
128 segments, reflecting peaks and valleys of the noisy raw coverage track, would then be  
129 assigned to different copy number states. Titan and THetA increased the fraction of amplified  
130 genome from less than 3 % as determined by ACEseq and ABSOLUTE to 29% and 42%,  
131 respectively. No indications for this reported increased fraction were found in the raw  
132 coverage data (Figure 1A). Furthermore, the sample's karyotype (46,X,-  
133 X,del(3)(p14p24),t(8;14)(q24;q32),+del(12)(q15)[10]/46,XX[1]), matched the ACEseq copy  
134 number predictions, demonstrating better estimates by ACEseq. Only ABSOLUTE showed  
135 results similar to ACEseq for this sample, though its TCC estimation was 68% below the  
136 ACEseq and the MAF-based estimate.

137 For a more detailed comparison we calculated the overall concordance of copy number calls  
138 for both total and allele-specific copy numbers (Supplementary Table 2). Strikingly, the  
139 highest concordance was observed between ACEseq and ABSOLUTE (average agreement  
140  $\emptyset=0.96$ ) emphasizing the robustness of the copy number calls as these methods use a  
141 different, independent data basis with WGS and SNP arrays, respectively. The concordance  
142 of ACEseq with the WGS-based methods was much smaller (average agreement: THetA  
143  $\emptyset=0.45$ , TITAN  $\emptyset=0.84$ ), though it increases to 0.91 for TITAN upon removal of the fast  
144 replicating sample 4112512 further confirming ACEseq copy number calls. Results from  
145 THetA differed substantially from ACEseq in 7 out of 11 samples, in which THetA reported  
146 much higher fractions of the genome as gained or lost. Again this is probably related to the  
147 strongly deviating TCC estimates and over-segmentation.

148 Overall, these results demonstrate the good performance of ACEseq in fully automated TCC  
149 and ploidy as well as allele-specific copy number estimation. ACEseq clearly benefits from its  
150 unprecedented integration of many and partially new features into a single tool. Even though  
151 ABSOLUTE and TITAN performed on a similar level for many of the samples they bear  
152 several shortcomings. ABSOLUTE always offered multiple TCC/ploidy solutions, reaching up  
153 to more than 40 possible solutions for one sample. The desired solution had to be extracted  
154 manually from an R-object and required further manual interaction. TITAN resulted in very  
155 good TCC estimation with the downside that the ploidy needs to be set in advance. Testing  
156 different ploidies requires multiple runs per sample. Additionally a strong replication timing  
157 bias caused problems for TITAN leading to over-segmentation and subsequently larger  
158 fractions of segments assigned as a gain or loss.

159 In conclusion, ACEseq provides a novel analysis platform for fully automated CNA calling on  
160 cancer WGS data without the requirement of prior information or any necessity for manual  
161 interference. By integrating GC and replication timing bias correction it improves  
162 segmentation and CNA calling performance compared to other tools. Importantly, it further  
163 provides quantitative metrics, which have been widely used for automatized quality control in  
164 large-scale pan cancer WGS projects. ACEseq is comprehensively documented under  
165 [aceseq.readthedocs.io](http://aceseq.readthedocs.io) and freely available at <https://github.com/eilslabs/ACEseqWorkflow>.

166

#### 167 Acknowledgement:

168 We thank P. Ginsbach and R. Drews for their contribution to establish impute2 and  
169 ABSOLUTE analyses, L. Fischer for improving the tool efficiency and Thomas Wolf for  
170 statistical support. This work was supported by the BMBF-funded Heidelberg Center for  
171 Human Bioinformatics (HD-HuB) within the German Network for Bioinformatics Infrastructure  
172 (de.NBI) (#031A537A, #031A537C), and the BMBF-funded projects ICGC MMML-Seq  
173 (#01KU1002A-J) and ICGC DE-MINING (#01KU1505E,G). Infrastructural support of the  
174 KinderKrebsInitiative Buchholz/Holm-Seppensen zu R.S. is gratefully acknowledged.

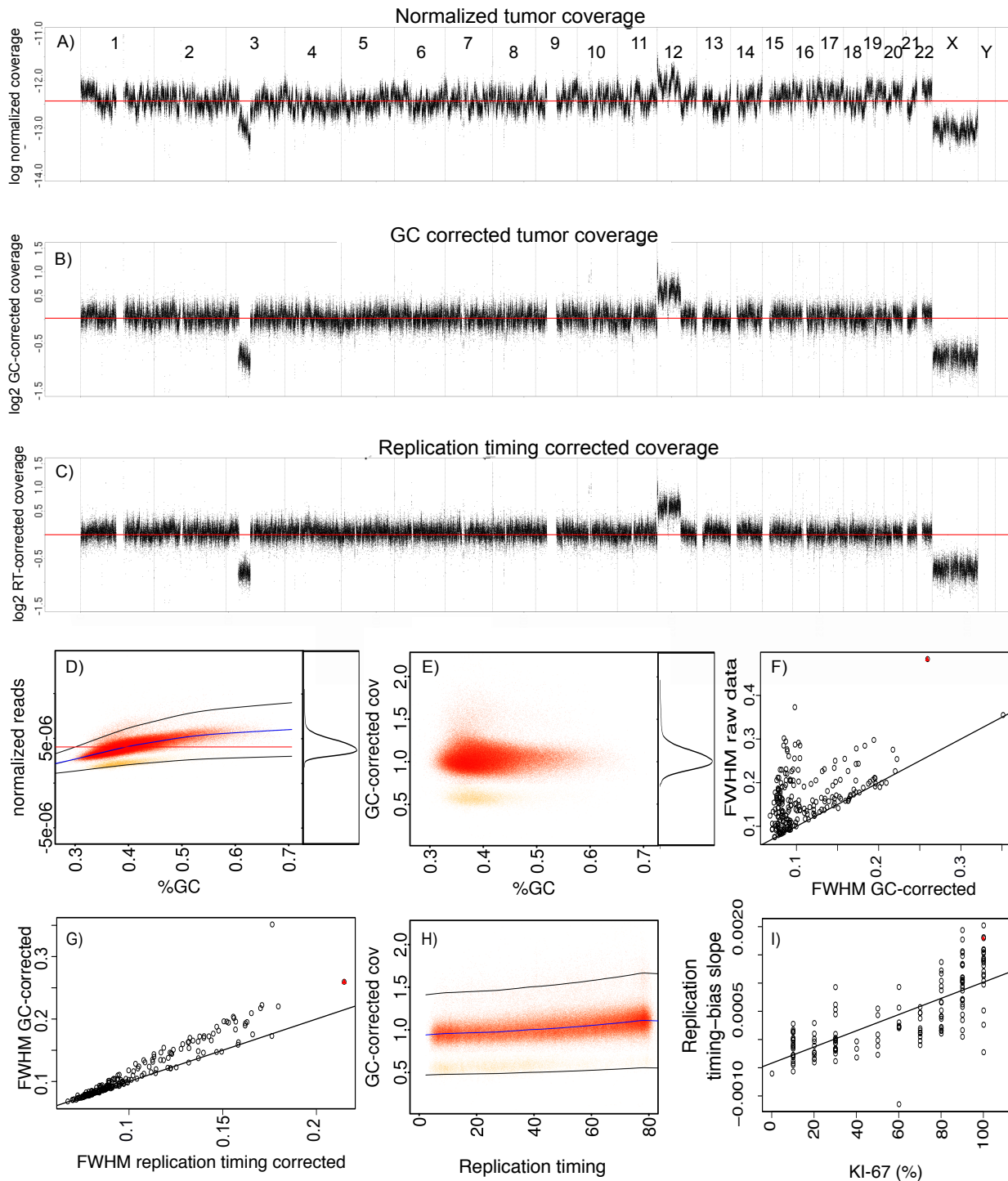
175

176

#### 177 Author contributions:

178 K.K, I.B, D.H, Z.G and M.S developed and implemented ACEseq. Integration into the Roddy  
179 pipeline framework was done by K.K, M.H. and P.K. The ICGC MMML Seq network  
180 coordinated by R.S. provided sequencing, FISH, SNP-array and pathology data. In particular,  
181 M.Hu., W.K., P.M., provided KI-67 measurements; C.L. performed FISH analyses and  
182 karyotyping, I.V. and C.L. provided SNP array data, R.W. coordinated quality control of  
183 sequencing data. K.K. analyzed data. R.E., R.S., B.B. and M.S. supervised the research. K.K.  
184 and M.S. wrote the manuscript. All authors provided feedback on the manuscript.

185



**Figure 1. GC- and replication-timing bias correction and QC.** Coverage profile for 10 kb windows before correction (A), after GC-bias correction (B), and after GC- and replication timing-bias correction (C). D: GC-bias correction curve fitted to the raw data shown in A, to firstly identify the main copy number state windows (red points) before a second curve is fitted used for correction of the bias. E: GC-corrected coverage distribution used for FWHM estimation. Comparison of FWHM estimates prior to and after GC bias (F) and replication-timing bias (G) correction for 219 lymphoma samples. H: Replication-timing bias correction curve estimating the replication speed. I: Comparison of KI-67 and estimated replication-timing slope. Sample 4112512 shown in panel A-E & H and is marked by a red triangle in F, G & I. RT: replication timing; Cov: coverage; FWHM: full width half maximum.

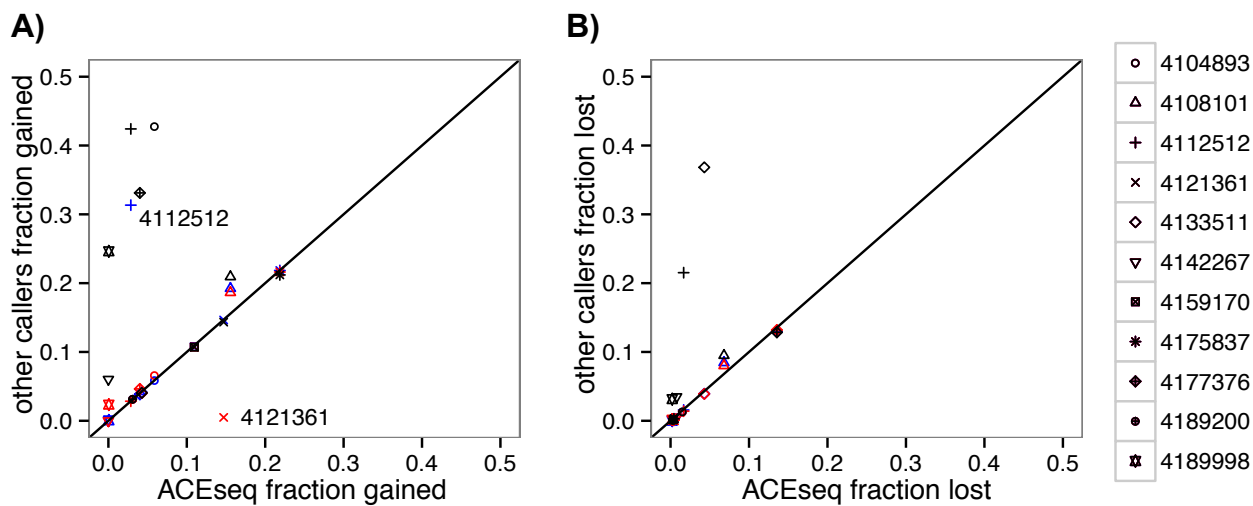


Figure 2: Comparison of gained (A) and lost (B) fraction of the genome for 11 B-cell lymphoma samples. The fraction called by ACeseq is compared to the other three callers. Two samples that deviate strongly for ABSOLUTE and TITAN are marked with the sample ID. TITAN: blue, THetA: black, ABSOLUTE: red.



- 186 1. Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer  
187 using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
- 188 2. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.*  
189 *U. S. A.* **107**, 16910–5 (2010).
- 190 3. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer.  
191 *Nat. Biotechnol.* **30**, 413–21 (2012).
- 192 4. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-  
193 throughput sequencing. *Nucleic Acids Res.* **40**, 1–14 (2012).
- 194 5. Chen, X. *et al.* CONSERTING: integrating copy-number analysis with structural-  
195 variation detection. *Nat. Methods* (2015). doi:10.1038/nmeth.3394
- 196 6. Koren, A. *et al.* Article Genetic Variation in Human DNA Replication Timing. 1015–1026  
197 (2014). doi:10.1016/j.cell.2014.10.025
- 198 7. Whalley, J. P. *et al.* Framework For Quality Assessment Of Whole Genome, Cancer  
199 Sequences. *bioRxiv* 1–27 (2017).
- 200 8. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-  
201 read analysis. *Bioinformatics* **28**, i333–i339 (2012).
- 202 9. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-  
203 like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* **72**, 5454–5462 (2012).
- 204 10. Birkbak, N. J. *et al.* Telomeric Allelic Imbalance Indicates Defective DNA Repair and  
205 Sensitivity to DNA-Damaging Agents Nicolai. *Cancer* **11**, 8274 (2011).
- 206 11. Wahle, M.-K. von *et al.* Intratumor Heterogeneity of Homologous Recombination  
207 Deficiency in Primary Breast Cancer. *Clin. Cancer Res.* **23**, 1193–1199 (2017).
- 208 12. Engert, F., Kovac, M., Baumhoer, D., Nathrath, M. & Fulda, S. Osteosarcoma cells with  
209 genetic signatures of BRCAness are susceptible to the PARP inhibitor talazoparib alone  
210 or in combination with chemotherapeutics. *Oncotarget* **8**, 48794–48806 (2016).
- 211 13. Melinda, L. T. *et al.* Homologous recombination deficiency (hrd) score predicts  
212 response to platinum-containing neoadjuvant chemotherapy in patients with triple-  
213 negative breast cancer. *Clin. Cancer Res.* **22**, 3764–3773 (2016).
- 214 14. Telli, M. L. *et al.* Phase II study of gemcitabine, carboplatin, and iniparib as neoadjuvant  
215 therapy for triple-negative and BRCA1/2 mutation-associated breast cancer with  
216 assessment of a tumor-based measure of genomic instability: PrECOG 0105. *J. Clin.*  
217 *Oncol.* **33**, 1895–1901 (2015).
- 218 15. Ha, G., Roth, A. & Khattra, J. TITAN: Inference of copy number architectures in clonal  
219 cell populations from tumor whole genome sequence data TITAN: Inference of copy  
220 number architectures in clonal cell populations from tumour whole genome sequence  
221 data. (2014). doi:10.1101/gr.180281.114
- 222 16. Oesper, L., Mahmoody, A. & Raphael, B. J. Inferring intra-tumor heterogeneity from  
223 high-throughput DNA sequencing data. *Lect. Notes Comput. Sci. (including Subser.*  
224 *Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **7821 LNBI**, 171–172 (2013).
- 225 17. Richter, J. *et al.* Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by  
226 integrated genome, exome and transcriptome sequencing. *Nat. Genet.* **44**, 1316–20  
227 (2012).