Chapter 4

A rigorous quality control procedure

4.1 Introduction

During my projects I have developed a rigorous quality control (QC) procedure for copy number profiles. In the previous chapter there was always a ground-truth available for the simulated cases, but for real data that is not available. Furthermore, my experience of working with Battenberg and other somatic copy number callers across 1000s of cases has revealed that whole genome duplication calling and handling various types of noise on the input data are difficult problems. For the ICGC pan-cancer project we therefore developed a consensus copy number calling approach (described later in this thesis) that is robust against outlier calls. However, the effect can be mitigated for a single caller by a quality control review and refitting procedure.

The procedure consists of a series of QC failure criteria, which can be assessed using a series of figures and are described further below. Once a profile fails QC it must go through a refit procedure that either involves an automatic or a manual refit (see Section 4.3). The profile then either passes QC or it will fail again, resulting in another refit. Most profiles that require this procedure pass after one refit, but for some samples it is impossible to find a fit that does not violate any of the criteria highlighted below. In such a scenario the QC violation could be unexpected interesting biology and requires further investigation. Examples of such violations can be multi-focal tumours or cases with a pre-malignant lesion.

The QC procedure is based on the expectation that a cancer sample contains the clone (the most recent common ancestor) that contained SNVs and, depending on the cancer type, CNAs that are shared by all cancer cells and are therefore clonal. The expectation is that the sequencing data shows those clonal mutations by means of a clonal mutation cluster and a large proportion of clonal CNAs. Furthermore, clonal SNVs are carried by a number of chromosome copies, which distribution should roughly follow the proportion of the genome

covered by different copy number states. Those three expectations link the SNV and CNA data together and they should fit as a trio as they are different views of the same cancer.

In an ideal world, every copy number profile is backed by independent validation. However, for the data described in this thesis there is very little validation data available. The samples described also provide the most heterogeneous data set that Battenberg has seen to date, with samples sequenced on different platforms and protocols, at different time points, to different specifications and as part of different projects. The diversity of these data required curation to obtain information about data quality and method performance.

A series of metrics have been developed that aim to capture the QC metrics described below. However, at the time of writing there is no substantial analysis on those metrics available. With the ICGC pan-cancer consensus copy number profiles available however, it should now in theory be possible to create a set of metrics that capture every QC failure.

4.2 Quality control metrics

The quality of a subclonal architecture is addressed by inspection of the copy number profile, the subclonal reconstruction and the estimated copy number states from the SNV data. In general, most samples pass after the first fit, but the success rate can be variable depending on the type of cancer (biology) or the sequencing project (data or biology).

I use Fig. 4.1 for initial assessment of the criteria highlighted below. In general one expects the copy number profile to be without any of the *fail* criteria, for the copy number estimate of SNVs to show peaks at integers (i.e. if there are 2 copies of a particular allele, then in general it is expected that some mutations are present on two copies) and the mutation clustering should normally show a peak at 1 for the clonal cluster.

Fig. 4.1 The main quality control figure used to assess the copy number fit and subclonal reconstruction initially. The top figure contains the copy number profile with the major allele in orange and the minor allele in dark grey. The middle figure shows the copy number estimate of the SNV data (which is calculated through Eq. 2.16). The bottom figure shows the subclonal architecture for this tumour with the mutation data as a histogram in the background and the clustering result in the foreground as a purple line with a turquoise confidence interval. The vertical lines represent found cluster locations.

4.2.1 Large homozygous deletions

Large homozygous deletions are an instant QC fail. As previously discussed, it would be unexpected if a whole chromosome was lost completely. But it is not directly clear where the

cutoff lies for a homozygous deletion to be believable. To be conservative I flag homozygous or subclonal homozygous deletions of 10Mb or greater, which means they are clearly visible in the copy number figure. In some cases the homozygous deletion should then be accepted as real, after closer inspection.

The case shown in Fig. 4.2 contains two subclonal homozygous deletions, of which one covers the majority of chromosome 18. This example also shows that multiple QC metrics can be triggered as it also contains a large number of chromosomes with subclonal states that would become clonal by doubling. This profile therefore also triggers a failure for the metric described in section 4.2.2. The mutation copy number and CCF space do not trigger any failures. With a purity of 17% one expects the clonal peak to appear somewhat shifted because with the relatively low coverage a sizeable number of clonal mutations fall below the detection limit. The shift is therefore the result of the *winner's curse*, which is addressed in Chapter 6. The solution for this sample is to add a whole genome duplication, which makes the homozygous deletions become a mixture of 1+0 and 2+0 (Fig. 4.3).

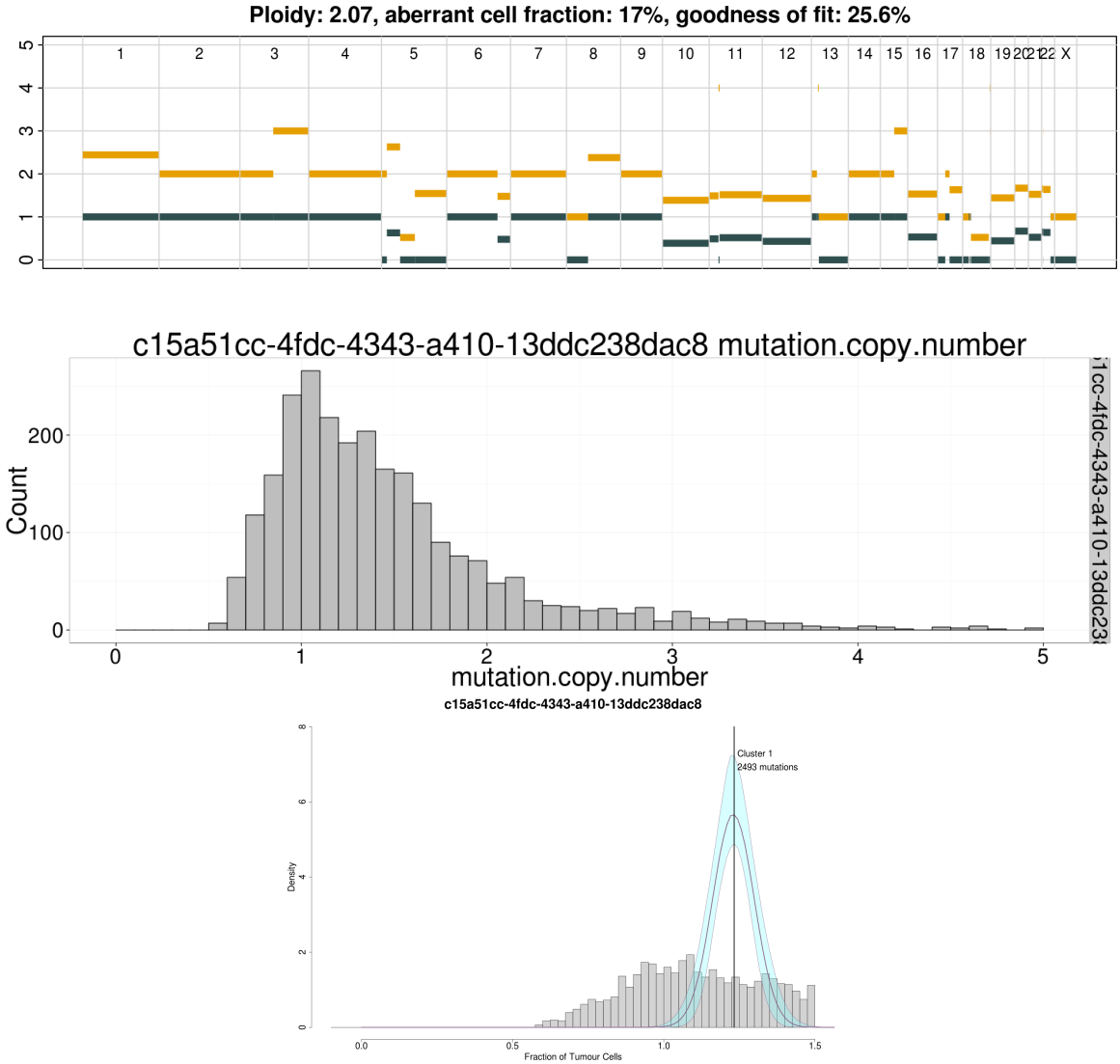

Fig. 4.2 QC failure case because of large (subclonal) homozygous deletions on chromosomes 5 and 18. This copy number profile also contains a number of subclonal copy number segments near 50% of tumour cells (detailed in Section 4.2.2).

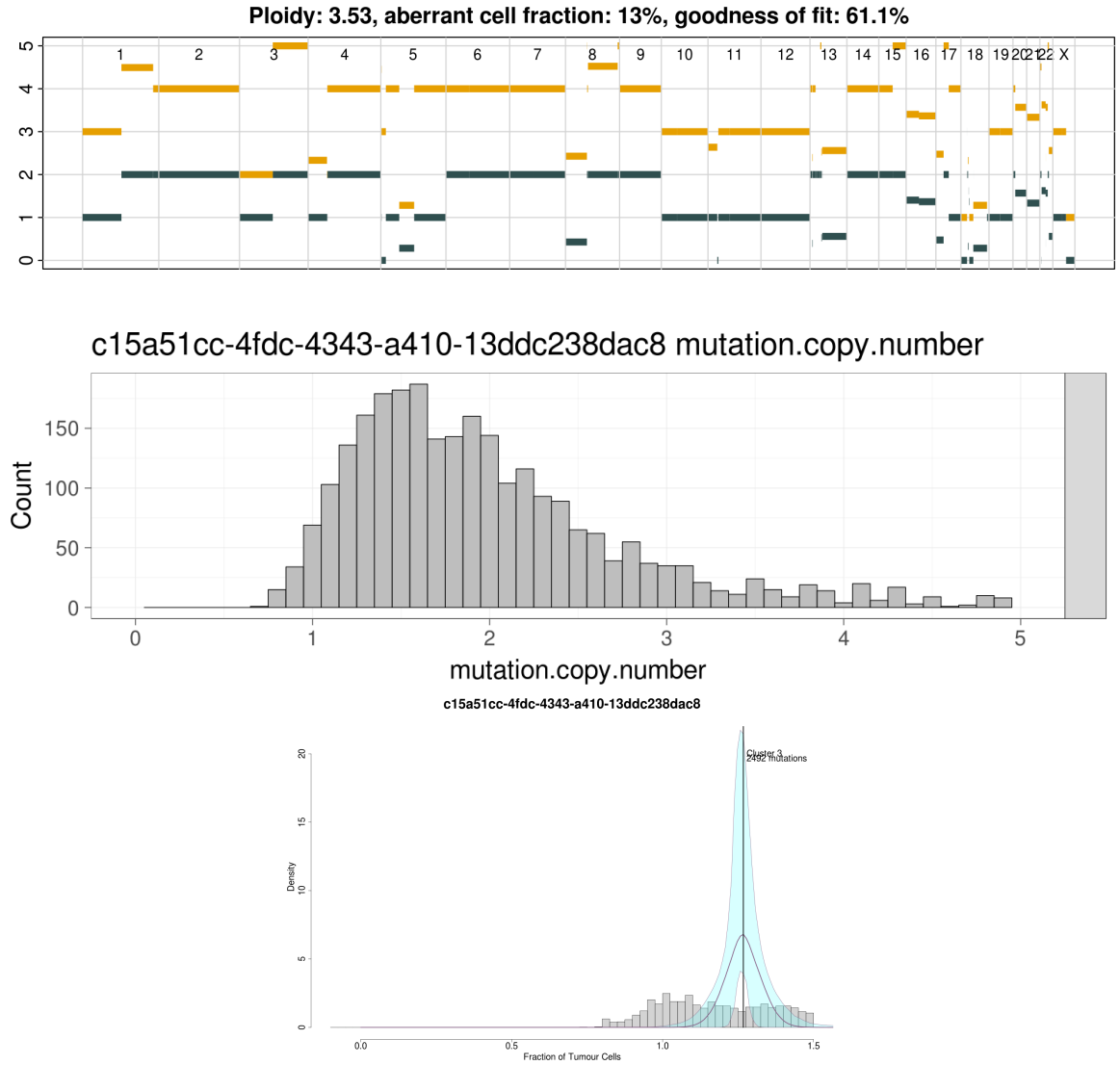

Fig. 4.3 The sublonal homozygous deletions on chromosomes 5 and 18 are resolved by adding a whole genome duplication (normal diploid 1+1 regions become 2+2). The subclonal segments near 50% tumour cells on chromosomes 1, 10, 12 and 19 become clonal. This tumour is of very low purity (13% of the sequencing sample contains tumour cells, according to this copy number fit), which in this case means a number of clonal SNVs fall below the detection limit. In the DPClust output figure (bottom) we therefore identify the clonal cluster shifted to a CCF higher than one (the shifting due to what is referred to as the *winner's curse*, which is briefly addressed in Chapter 6). This is a characteristic of the data and cannot be reverted by adjusting the copy number profile.

4.2.2 50% subclone in copy number

Beyond the subclonal homozygous deletions, Fig. 4.2 contains segments covering nearly the whole of a number of chromosomes that appear right in between two clonal states. The profile would initially fail and closer inspection of the detailed copy number figures that Battenberg produces reveals that the segments on chromosomes 11 and 16 have estimates close to 50% of tumour cells (Fig. 4.4). A whole chromosome arm or multiple large segments on different chromosomes is enough to trigger a fail. The next step is to refit the profile by doubling one of the clonal alterations (in this case for example the large segment on chromosome 13). A refit is accepted if the identified subclonal segments become clonal and no other failure criteria are triggered.

4.2.3 Empty odd numbered copy number state

A whole genome duplication can always be added to a copy number profile, and it produces an equally likely explanation of the data. But when a duplication too many is added it sometimes leaves a copy number state empty. In Fig. 4.5 there is no segment that takes on copy number state 1. Furthermore, in the mutation copy number figure there is no clear peak at 1 either. That suggests that either the whole genome duplication was the last event that became clonal, or that the duplication was added erroneously. Without further evidence it is not possible to distinguish between the former and latter, but the profile without a duplication provides a simpler explanation of the observed data and is considered a maximum parsimony explanation. The approach taken with samples reported in this thesis is that a whole genome duplication must be supported by clear evidence, preferably from the copy number and SNV data. The solution for this sample is therefore to halve the ploidy, as there is no information to support the duplication (Fig. 4.6).

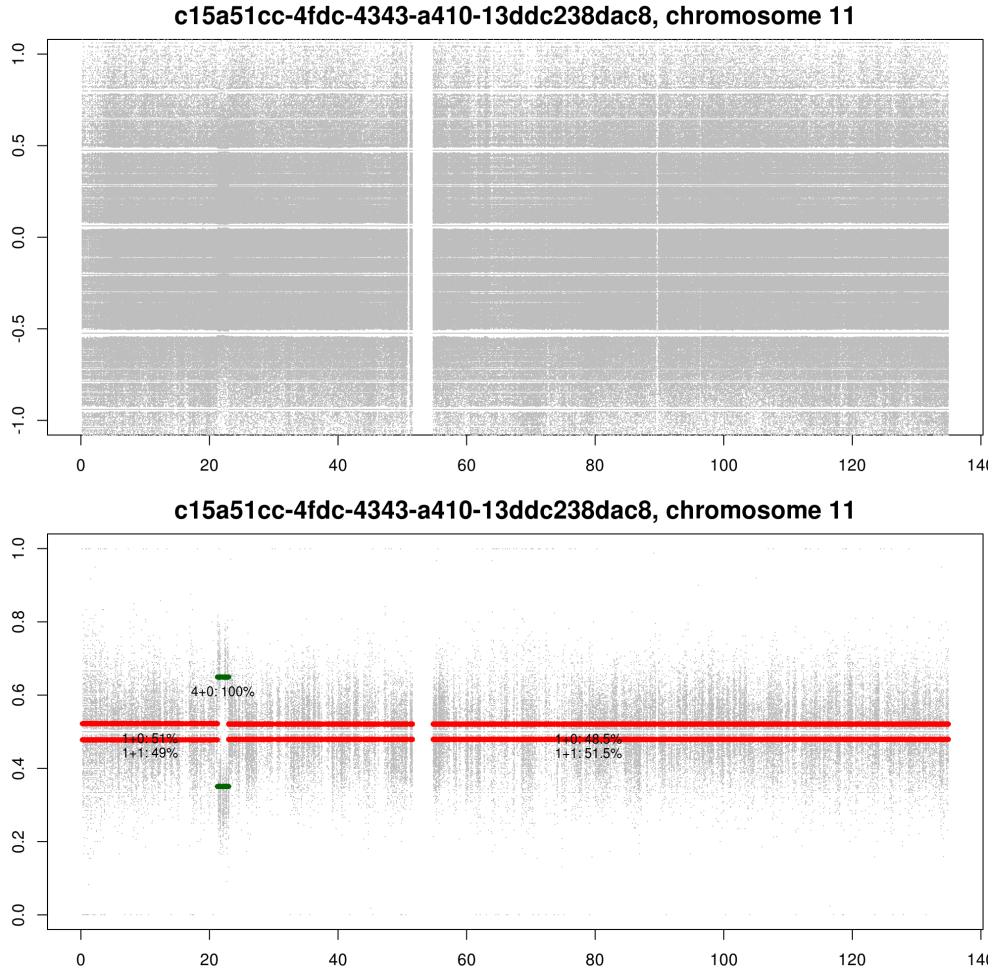

Fig. 4.4 Detailed copy number fit of chromosome 11. The top figure shows the relative copy number (logR), which is not informative for these purposes. The bottom figure contains the BAF and the copy number fit where subclonal copy number is plot by a red line. The subclonal segments on this chromosome are fit with CCF values close to 50% of tumour cells. This QC fail can be resolved by adding a whole genome duplication to this copy number profile.

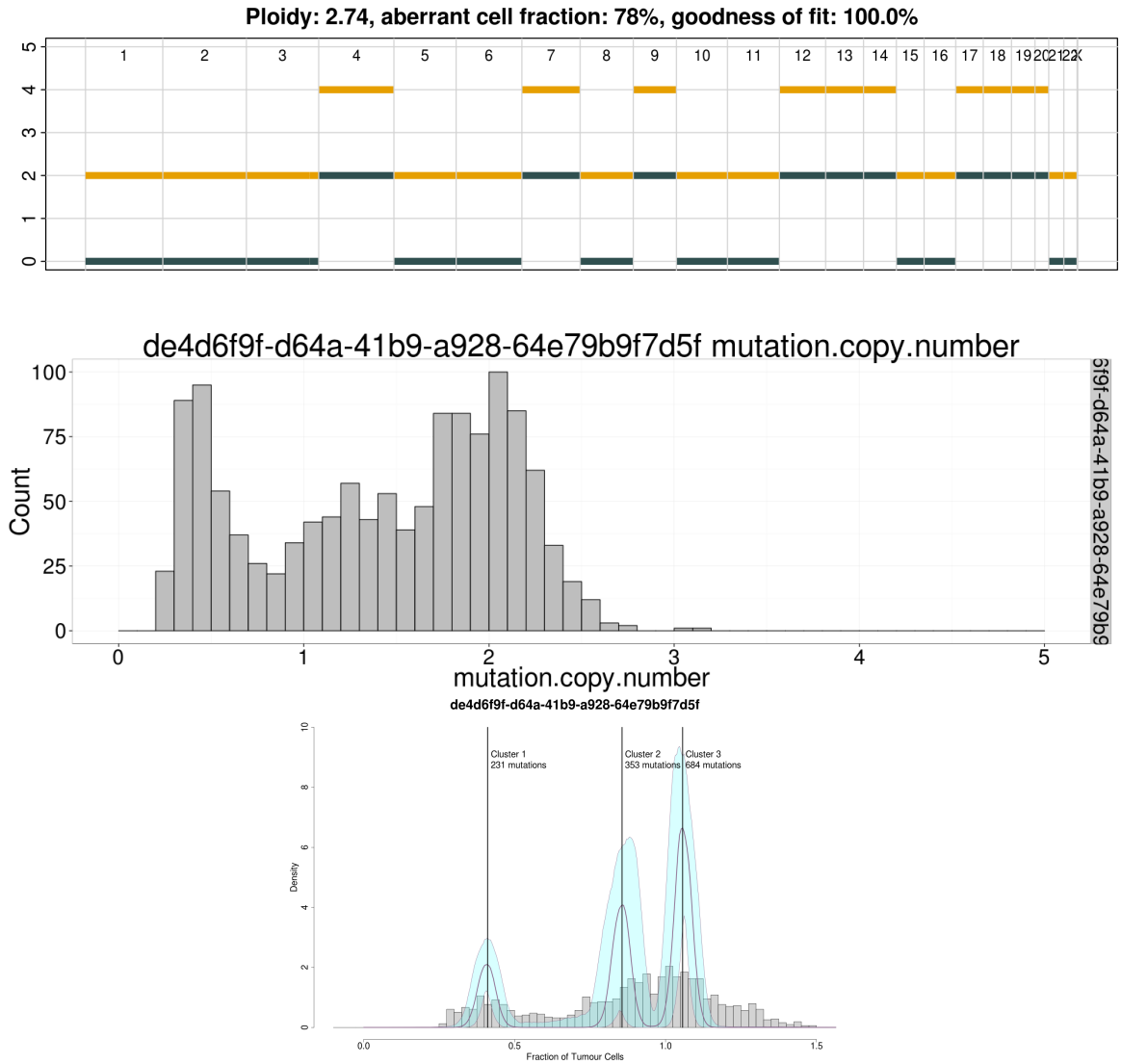

Fig. 4.5 This case fails QC because there are no segments fit with copy number state 1. In this scenario a whole genome duplication has been added that has not yielded an increase in the proportion of clonal copy number.

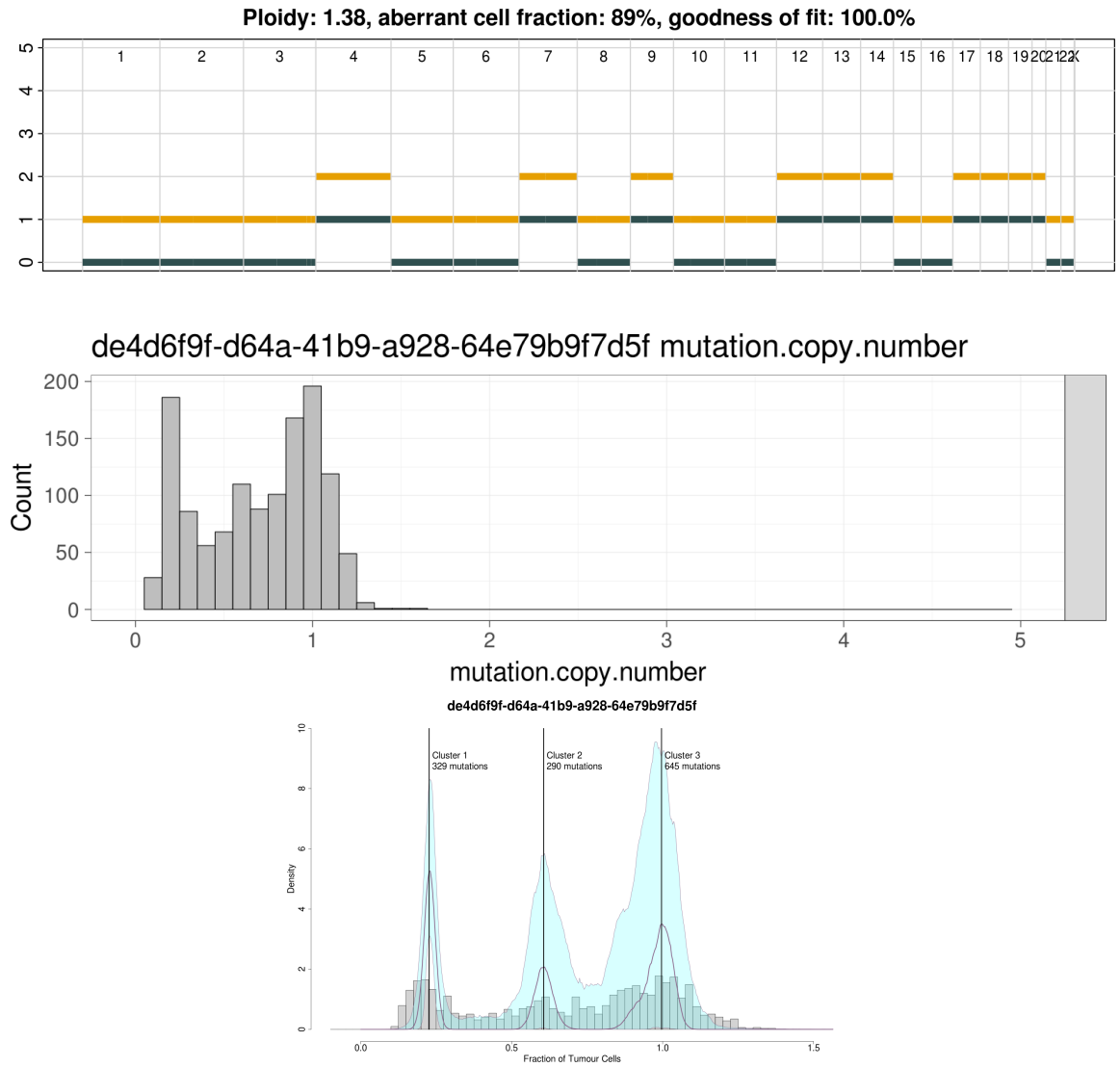

Fig. 4.6 The whole genome duplication is removed by refitting chromosome 1 with a 1+0 copy number state. This adjustment creates a clear mutation copy number peak at 1. Note that in the bottom figure Cluster 2 is not called at 50% of tumour cells and therefore does not fail that criteria (see Section 4.2.6).

4.2.4 No clonal copy number alteration

In some cases Battenberg does not find a solution with a single clonal copy number alteration. Battenberg requires at least one clonal alteration to estimate the purity, hence in cases like the one shown in Fig. 4.7 the purity estimate is incorrect. If there are no alterations in the profile then an alternative source must be used to estimate purity (for example from clonal SNVs). But in this case Battenberg has not been able to fit the segments on chromosomes 7 or 8 with a clonal copy number state, and in this scenario, the purity estimate is too high. This has affected the mutation copy number and CCF spaces by shifting the clonal peak to the left. There are two possible solutions for cases like this: force Battenberg to fit a selected segment with a particular copy number state, or obtain a purity estimate by other means (from SNVs in normal diploid 1+1 regions for example). The former approach focusses on the belief that there must be at least one clonal CNA, the latter on the belief that there might not be a single CNA.

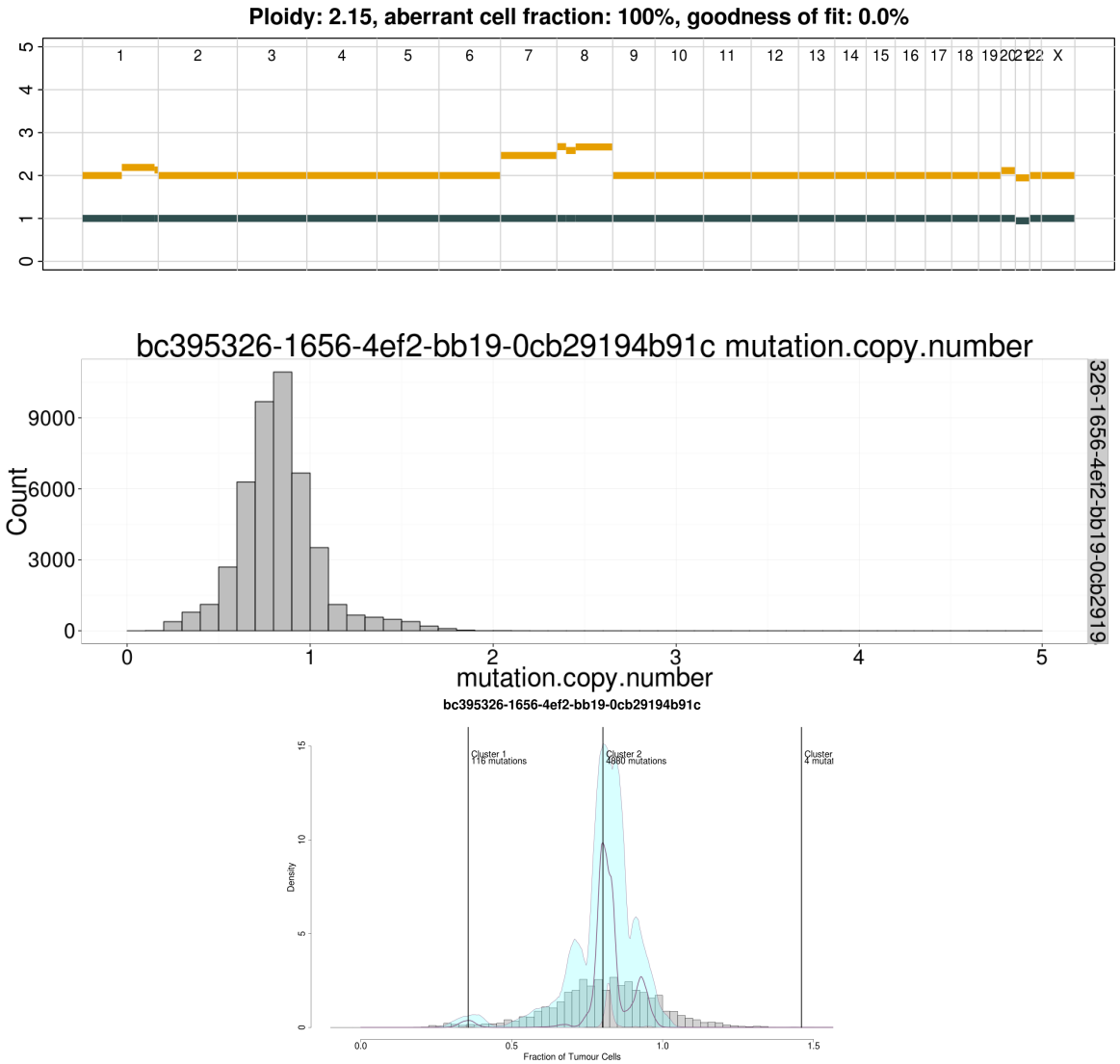

Fig. 4.7 Battenberg has not fit a segment with a clonal alteration. That means the purity estimate is most likely incorrect, supported by the shifted peaks in mutation copy number and CCF space.

4.2.5 Shifted clonal mutation cluster

The example highlighted in Fig. 4.7 would require attention due to its shifted clonal peak. In some cases a clonal peak can be shifted towards the right. That shift is caused by the winner's curse (see Chapter 6) where a number of clonal SNVs fell below the detection limit and that has caused the weight of the clonal distribution to shift. However, in cases where the clonal cluster is either seemingly fully sampled or the cluster is shifted to the left, the shift could be an indication that the purity estimate is incorrect. That can have different reasons for different profiles. For example, in Fig. 4.7 it is due to no segments being fit with a clonal alteration, but it can also be the effect of the wrong segment fit with a clonal state. In Fig. 4.7 fitting the altered segment on chromosome 1 as 2+1 will yield a very different purity and ploidy then fitting the chromosome 7 as 2+1. In cases of a shifted clonal cluster and no other QC violations the solution is often to look for an alternative clonal segment until the shift is resolved.

4.2.6 Mutation cluster at 50% of tumour cells

A clear mutation cluster at 50% of tumour cells can be a QC violation. But by chance one can observe a tumour with a subclone that takes up exactly half of the tumour. It is therefore not always clear whether a sample should pass or fail QC. In the example depicted in Fig. 4.8 there is a clear SNV cluster at 0.5 visible and the copy number profile also violates the subclone at 50% QC criterion (Section 4.2.2). Furthermore, DPClust finds an SNV cluster at a CCF of 1, but from the histogram it is not clearly there. The fact that Battenberg fits this profile with only chromosome 1q as a clonal alteration is suspicious as it would appear that chromosomes 6, 7, 8p, 8q, 11 and 18 could become clonal by 'stretching' the profile (i.e. 8p is 1+0, 8q is 5+1). That solution possibly shifts the 50% mutation cluster to become clonal. In such a scenario, a solution that yields a large proportion of the alterations as clonal is preferable as it one of the foundations upon which Battenberg is based. For the purpose of describing heterogeneity it would also lead to a conservative estimate if the clonal solution is incorrect. Fig. 4.9 shows that this provides a coherent fit that does not violate any of the criteria listed in this Chapter.

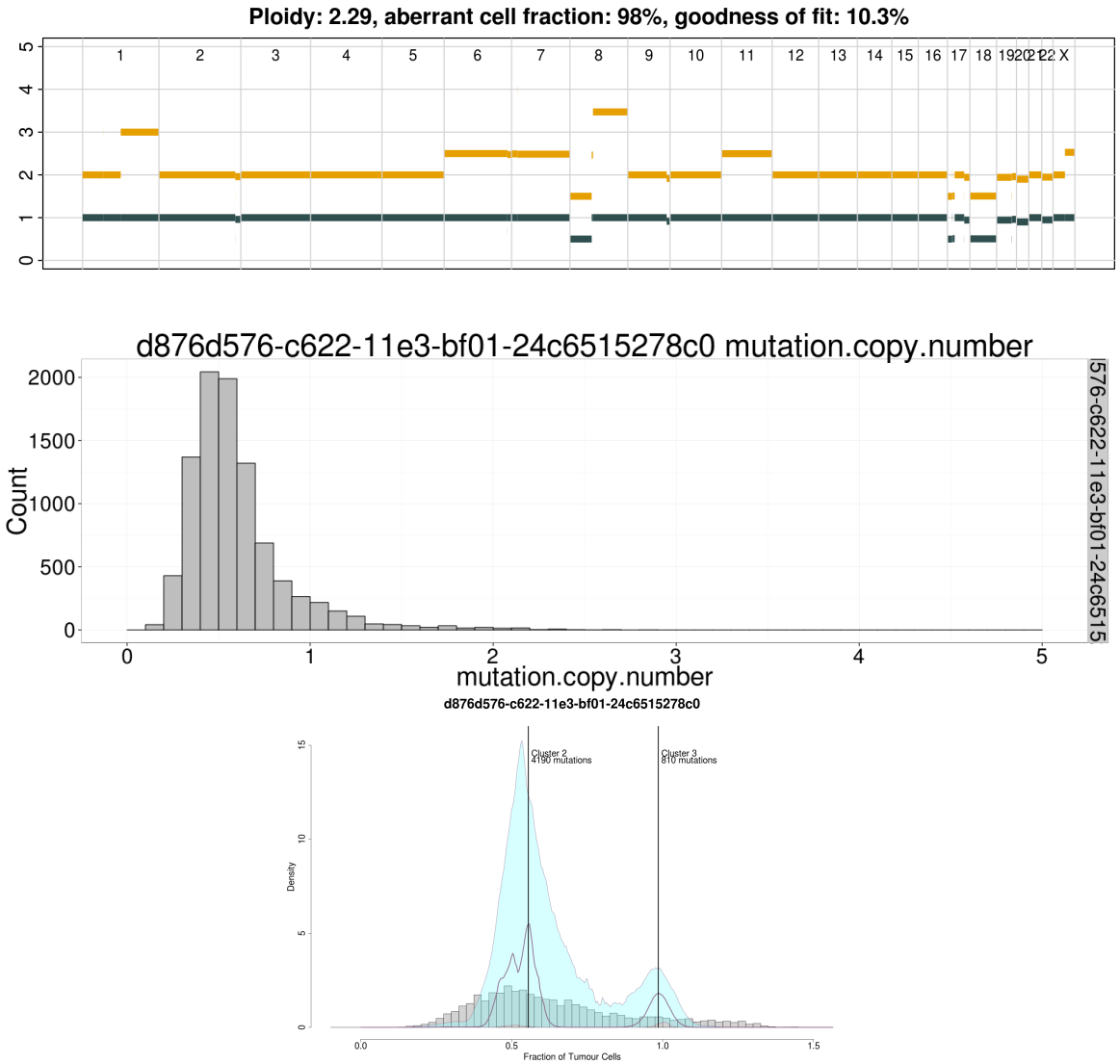

Fig. 4.8 Example of a case with a clear large mutation cluster at a CCF of about 0.5. In this particular scenario there are other QC violations, most notably the copy number segments in between two clonal states. The mutation cluster at 0.5 could be a sign that the ploidy needs doubling, in this case one could also try to choose another clonal segment over chromosome 1q.

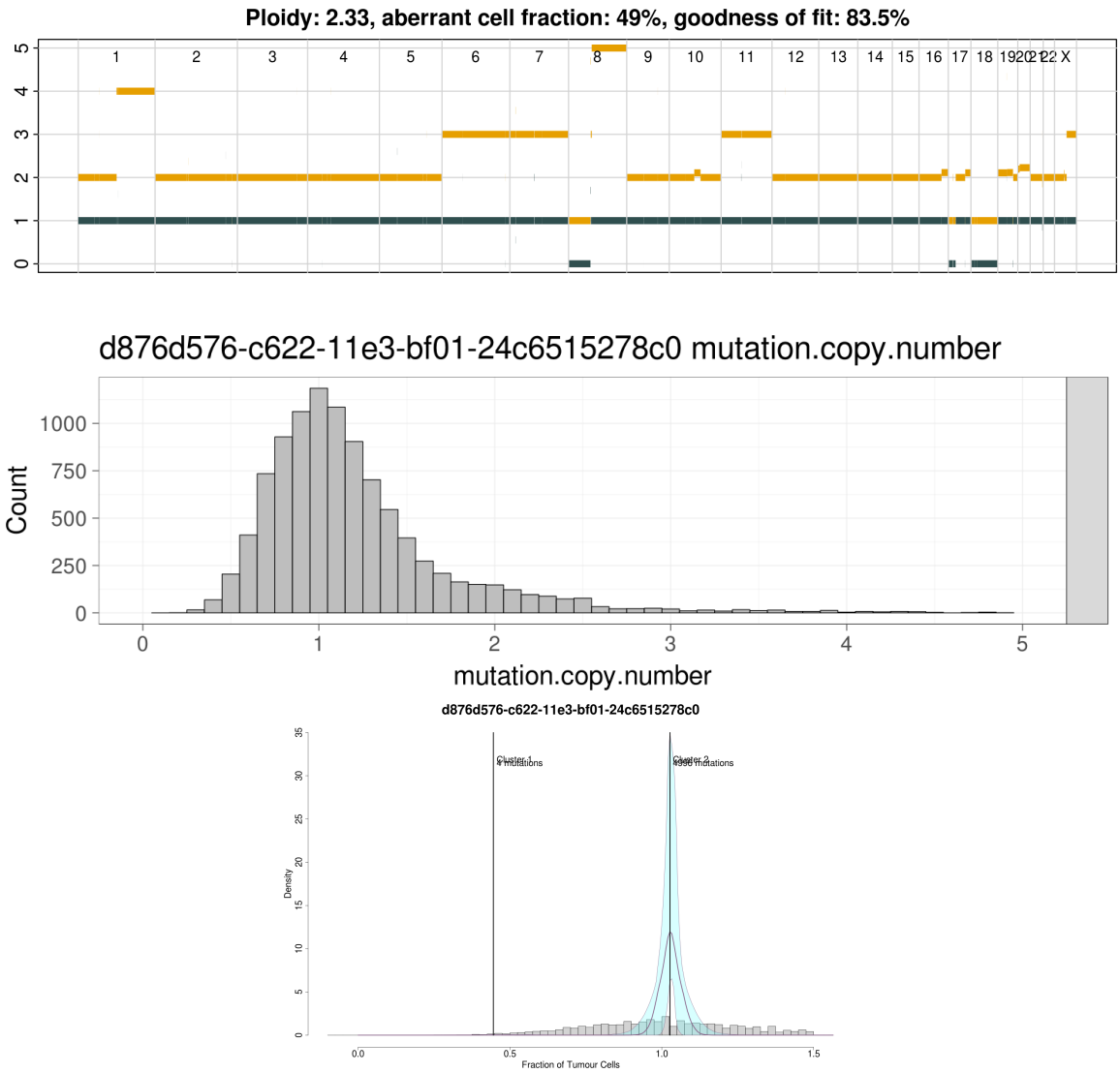

Fig. 4.9 The refit copy number profile does not contain any QC violations. The mutation cluster at 50% of tumour cells is removed, as are the copy number segments exactly in between two clonal states.

4.2.7 Empty mutation copy number state

In some cases the addition of an incorrect whole genome duplication or of an extra copy to some alleles can yield an empty mutation copy number state. Fig. 4.10 shows no clear peak at a mutation copy number state of 1, suggesting there are very few SNVs that are clonal and carried by a single chromosome copy. In this case the peak at mutation copy number 3 contains mutations on chromosome segments that are 3+0, whilst the peak at mutation copy number 2 contains mutations on the balanced 2+2 chromosomes. This particular example also contains an empty copy number state (see Section 4.2.3).

In this scenario the raw data can be equally likely explained by subtracting copies from every segment that does not have a copy number count of 0. One could refit with for example chromosome 2 as 2+0 or 1+0. This compresses the profile and will adjust the mutations with copy number states 2 and 3 to 1 and 2. That leads to a maximum parsimony explanation of the data as the additional duplication in the current profile does not allow for much more of the alterations to be explained as clonal and hence the data can be explained without the duplication. Fig. 4.11 shows the refit with chromosome 2 fit as 1+0, which yields a clean mutation copy number and DPClust figure and no QC criteria are violated.

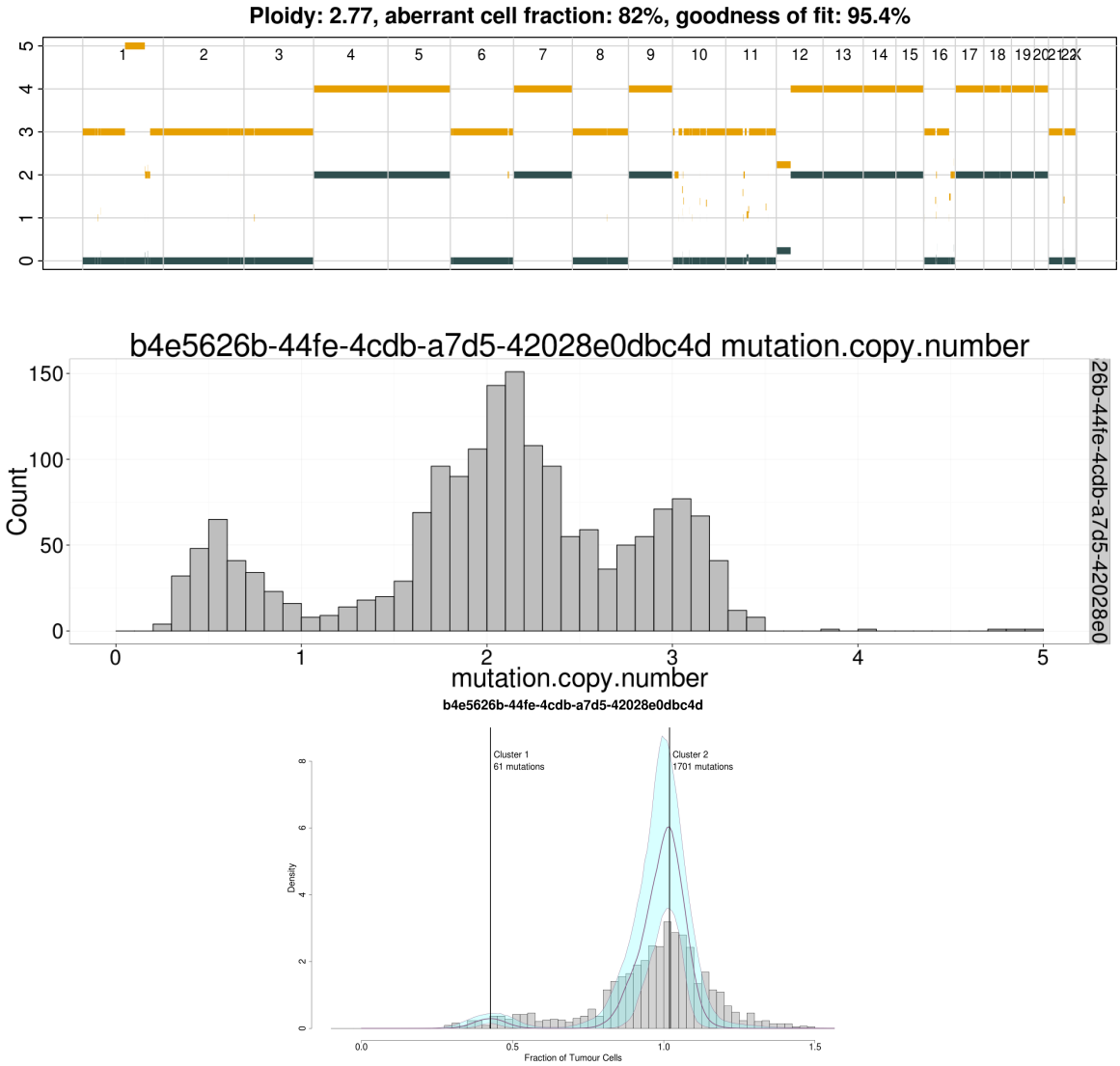

Fig. 4.10 An empty mutation copy number state can be an indication that additional chromosome copies have been added that do not help in explaining the largest possible proportion of the alterations as clonal. In this case there is no peak at mutation copy number 1, and the copy number profile also contains an empty state at 1.

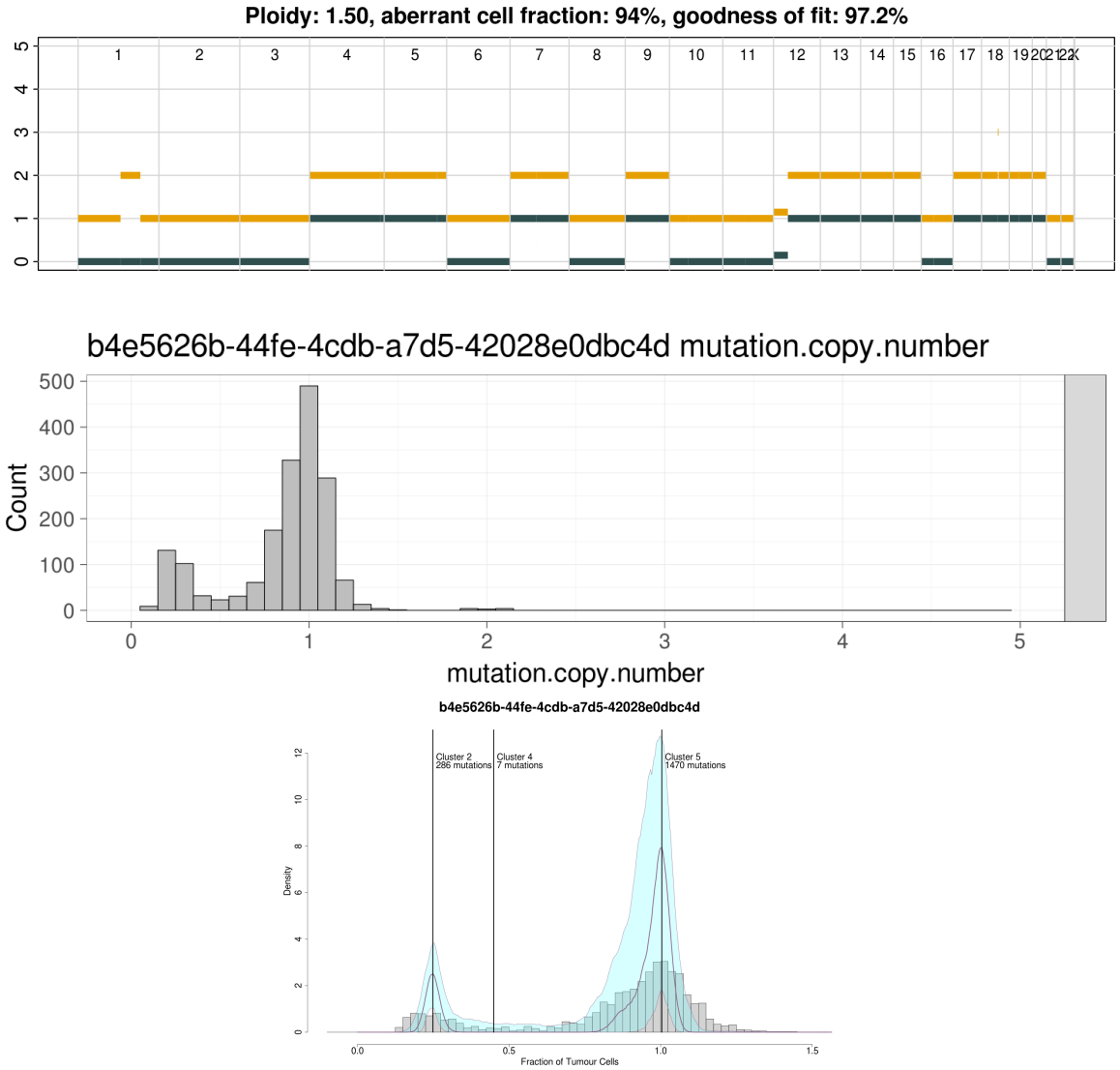

Fig. 4.11 A refit with chromosome 2 as 1+0 yields a clean CNA profile, mutation copy number figure and DPClust result.

4.3 Resolving a quality control failure

A QC fail can be resolved in multiple ways. An automatic procedure for Battenberg exists that obtains a purity estimate from SNVs, but leaves the ploidy unchanged. There is also a manual approach where one can request Battenberg to fit a copy number profile with a certain segment with a particular combination of major and minor allele states.

4.3.1 Automatic correction

An automatic correction is currently possible for cases where the purity should be estimated from another source because there are no clear clonal copy number alterations. A purity estimate can be obtained from SNVs in balanced copy number regions (preferably 1+1) where the VAF of clonal SNVs can be directly used. An estimate can be obtained by running DPClust in VAF space and to take the location of the SNV cluster closest to 1 and multiply it by 2 for a purity estimate. If SNVs in 2+2 regions are taken, the estimate should be adjusted appropriately. This approach does not make any adjustments to the ploidy and is therefore best suited for types of cancer with very quiet copy number profiles.

4.3.2 Manual correction

Manual correction is possible for copy number profiles that contain clonal alterations. One can hypothesize that a particular segment should have a particular combination of major and minor allele states, for example, chromosome 8p 1+0.

The ASCAT equations can be rewritten to obtain Eqs. 4.1 and 4.2 that convert the hypothesis into a suggested ρ and ψ parameter combination that Battenberg takes in when it fits a profile. Battenberg then skips the first fitting step to obtain an initial global fit, and it will start with the local optimisation to find the best solution.

In Eqs. 4.1 and 4.2, $n_{A,i}$, $n_{B,i}$ are the major and minor allele copy number states suggested for segment i and b_i and l_i are the BAF and logR respectively of segment i.

$$\rho = \frac{2b_i - 1}{2b_i - b_i(n_{A,i} + n_{B,i}) - 1 + n_{A,i}} \tag{4.1}$$

$$\psi = \frac{\rho(n_{A,i} + n_{B,i}) + 2 - 2\rho}{2^{l_i}} \tag{4.2}$$

After a refit suggestion, Battenberg produces a new profile, which is followed by a DPClust run and a new QC procedure.

4.4 Inventory of metric triggers in the PCAWG data set

I have attempted to incorporate the above metrics into a series of automated checks, through which a profile can be automatically flagged as either pass or fail. And applying these metrics to Battenberg profiles where no refitting has taken place can reveal how often these scenarios occur. I have therefore performed a rerun across all samples in the data set without refit suggestions of the copy number fitting pipeline, with the Battenberg version (2.2.5) that was used for PCAWG.

There are two main reasons why an additional run was required and these numbers could not be extracted from notes taken during the Battenberg PCAWG QC, or by simply comparing Battenberg with the PCAWG consensus. First, Battenberg has received a range of upgrades over the course of PCAWG. Most notably GC correlated wave correction and the inclusion of SV breakpoints. Both these additions were essential to increase performance on this heterogeneous data set. Second, a comparison against the PCAWG consensus profiles (detailed in section 6.2) would be imperfect as the PCAWG consensus consists of only clonal copy number states. This means that an unknown percentage of copy number segments is represented with a slightly different fit than the data suggests, which affects the calculated ploidy. A rerun of the exact same version of Battenberg with and without refitting is not affected by these downsides.

The comparison of two Battenberg runs reveals that refitting causes a discrepancy in either purity or ploidy in 15.2% of 2,748 samples for which output of both runs was available (Fig. 4.12). Nearly half of these are caused by the lack of a clonal copy number alteration (Table 4.1), while the other metrics trigger either between 20-30% or 10% or fewer cases. A total of 14 samples did not trigger any of the metrics. Manual inspection of the profiles revealed that in 7 cases the refit profile may be incorrect, as it triggered one or more metrics. The other 7 are the result of a bug that has been fixed at the time of writing, but was still prevalent in Battenberg version 2.2.5. This bug caused a slight discrepancy in the stored ploidy value, which could occasionally push subclonal gains into losses or vice versa.

It is often a single metric that causes the bulk of the triggers in a cancer type. Many cancer types with often quiet copy number profiles (pilo-astrocytoma, thyroid adenocarcinoma, benign bone cancers and AML) therefore often trigger the "No clonal CNA" metric. In this scenario one can use SNVs to estimate the purity and, without further evidence of a whole genome duplication available, set the ploidy to equal 2. It is for this scenario that

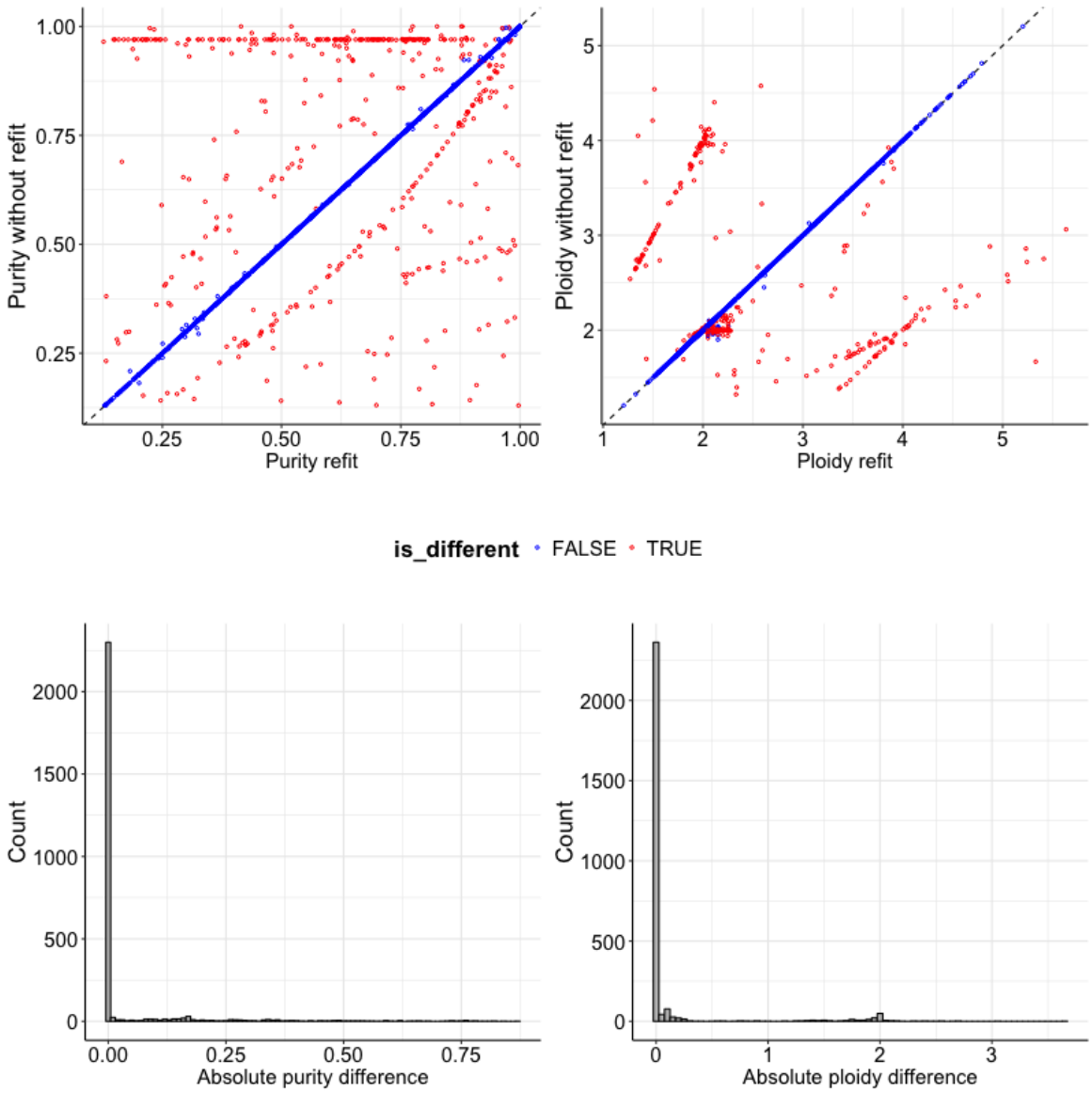

Fig. 4.12 Comparison of purity and ploidy values generated by Battenberg with and without refitting. The top figures compare the purity (left) and ploidy (right) where a discordant sample is coloured red. The bottom figures show a distribution of the difference between purity (left) and ploidy (right) between the two runs.

the automatic refitting pipeline was developed (see section 4.3.1), which should be a future extension of the Battenberg pipeline.

Pancreatic endocrine cancers often trigger scenario C (empty copy number state). This may also reflect the underlying biology. Pancreatic endocrine tumours often show very little subclonal copy number alterations, often contain whole chromosome LOH and (to a lesser extent) whole chromosome gains and relatively frequent whole genome duplications.

Metric	Num. cases	Frac. different	Frac. total
A. No clonal CNA	207	49.2%	7.4%
B. CNA subclone at 50%	108	25.77%	3.8%
C. Empty CN state	98	23.3%	3.5%
D. Shifted clone	88	20.9%	3.2%
E. SNV cluster at 50%	43	10.2%	1.5%
F. Large hom del	32	7.6%	1.1%

Table 4.1 Overview of QC metric triggers between Battenberg with and without refitting. Almost half the cases with a discrepancy in either purity or ploidy contain copy number profiles without a clonal CNA. Between 20-30% of cases trigger a CNA subclone at near 50% of tumour cells, an empty copy number state or a shifted clonal cluster. Around 10% of cases contain an SNV cluster near 50% of tumour cells or a large homozygous deletion.

Histology	A	B	C	D	E	F	Samples	Samples diff.	Frac. diff.
CNS-PiloAstro	64	2	3	2	1	2	88	69	0.78
Prost-AdenoCA	44	15	5	6	5	2	284	65	0.22
Liver-HCC	7	19	5	8	7	4	326	34	0.10
CNS-Medullo	4	8	12	12	4	6	139	31	0.22
Panc-Endocrine	0	2	26	1	2	2	85	27	0.31
Thy-AdenoCA	20	0	4	1	1	0	48	23	0.47
Lymph-BNHL	8	5	5	8	0	1	106	19	0.17
Kidney-RCC.clearcell	4	2	6	6	1	1	111	14	0.12
Lymph-CLL	8	2	7	4	2	0	90	14	0.15
Kidney-ChRCC	2	1	6	0	0	1	45	11	0.24
Bone-Benign	9	0	1	0	0	0	16	10	0.62
Myeloid-AML	7	2	0	0	0	0	16	7	0.43
Kidney-RCC.papillary	3	2	3	3	1	2	33	6	0.18
Bone-Osteosarc	2	3	1	0	2	0	38	6	0.15
Myeloid-MPN	4	0	1	1	1	0	45	5	0.11

Table 4.2 The number of samples triggering the six metrics, split per cancer type: A=No clonal CNA, B=CNA subclone at 50%, C=Empty CN state, D=Shifted clone, E=SNV cluster at 50%, F=Large hom del.

This means that for the Battenberg metric (the proportion of the genome that is fit with a clonal state) is often extremely similar between a profile with and without a whole genome duplication.

Furthermore, it is possible that some pancreatic cancers indeed contain an empty copy number state. Such a scenario can occur if the last copy number alteration to occur is a whole genome doubling. Sample SA570847, shown in Fig. 4.13, contains many SNVs on 1 and many SNVs on 2 chromosome copies. During the PCAWG expert panel review (see

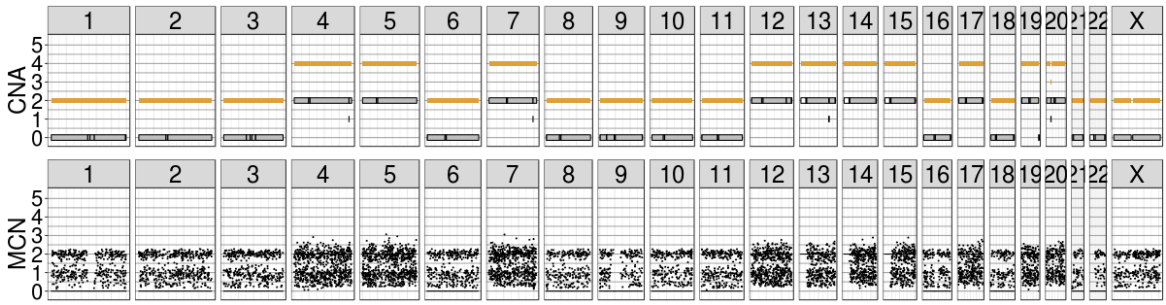

Fig. 4.13 The copy number profile of PCAWG tumour SA570847 (top), with the total copy number in orange and the minor allele in grey. The bottom figure shows mutation copy number (MCN, the raw estimated number of chromosome copies an SNV is carried by) for all SNVs detected in this tumour. SA570847 clearly shows a large number of SNVs on 1 and on 2 chromosome copies, justifying the addition of a whole genome doubling to the copy number profile, even though it leaves almost no segment at 1 chromosome copy.

section 6.2.6) we have occasionally allowed an empty copy number state based on convincing evidence of a genome doubling.

This example suggests that, even though these metrics capture the essence of what a manual QC captures, there can be exceptions due to specific characteristics of a particular type of cancer. It also highlights that a combination of metrics can sometimes lead to convincing evidence that contradicts a single metric. This is a sign that a combination of metrics could be a fruitful approach. It is unclear however, how to adequately weight multiple metrics against each other. A machine learning approach may be able to learn weights between the metrics by using the metrics from the PCAWG data set. This may be an interesting direction to explore in the future in order to further improve the metrics system.