Comprehensive analysis of copy number aberrations in microsatellite stable colon cancer in view of stromal component

M. Henar Alonso[1,2], Susanna Aussó[1,2], Adriana Lopez-Dóriga[1,2], David Cordero[1,2], Elisabet Guinó[1,2], Xavier Solé[1,2], Mercè Barenys[2,3,4], Javier de Oca[2,4,5], Gabriel Capella[2,4,6], Ramón Salazar[2,4,7], Rebeca Sanz-Pamplona[1,2*], Victor Moreno[1,2,4*].

1 Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO), and CIBERESP, Barcelona, Spain

2 Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain.

3 Gastroenterology Service, Hospital de Viladecans, Barcelona, Spain.

4 Department of Clinical Sciences, Faculty of Medicine, University of Barcelona (UB), Barcelona, Spain

5 Department of General and Digestive Surgery, Bellvitge University Hospital, Barcelona, Spain.

6 Hereditary Cancer Program. Catalan Institute of Oncology (ICO) and CIBERONC, Barcelona, Spain.

7 Oncology Service, Catalan Institute of Oncology (ICO) and CIBERONC, Barcelona, Spain

* Corresponding authors:

Victor Moreno v.moreno@iconcologia.net

Rebeca Sanz-Pamplona rebecasanz@iconcologia.net

Av. Gran Vía 199-203. L'Hospitalet de Llobregat, 08908 Barcelona, Spain.

Telephone: +34 93 260 71 86 / Fax: +34 93 260 71 88

Running title

Copy number aberrations in colon cancer

Contributions

MHA, RSP and VM conceived the study and wrote the manuscript. MHA, RSP, SA, ALD, DC and XS performed bioinformatics analyses. MHA and EG performed statistical analyses. MB, JdO, GC and RS recruited patients, collected and revised clinical data and follow-up. All authors revised and contributed to the final version of the manuscript.

Funding

This work was supported by the Catalan Institute of Oncology and the Instituto de Salud Carlos III, co-funded by FEDER funds –a way to build Europe– (grants PI08-1635, PS09-1037, PI11-1439, PI14-00613), and CIBERESP (grant CB07/02/2005). Also by the Catalan Government DURSI (grant 2014SGR647). This work was supported by COST Action BM1206.

KEYWORDS

Colon cancer, Copy number aberrations, Gene expression, Stroma, Molecular subtypes

ABSTRACT

Background: Somatic copy number aberrations (CNA) are common acquired changes in cancer cells playing an important role in the progression of colon cancer (CRC). This study aimed to perform a characterization of CNA and their impact in gene expression.

Methods: CNA were inferred from SNP array data in a series of 99 CRC. CNA events were calculated and used to assess the association between copy number dosage, clinical and molecular characteristics of the tumours, and gene expression changes. All analyses were adjusted for the quantity of stroma in each sample, that was inferred from gene expression data.

Results: High heterogeneity among samples was observed, the proportion of altered genome ranged between 0.04 and 26.6%. Recurrent CNA regions with gains were frequent in chromosomes 7p, 8q, 13q, and 20 while 8p, 17p, and 18 cumulated loses. A significant positive correlation was observed between the number of somatic mutations and total CNA (Spearman r=0.42, *P*=0.006). Approximately 37% of genes located in CNA regions changed their level of expression, and the average partial correlation (adjusted for stromal content) with copy number was 0.54 (inter-quartile range 0.20 to 0.81). Altered genes showed enrichment in pathways relevant for colorectal cancer. Tumours classified as CMS2 and CMS4 by the consensus molecular subtyping showed higher frequency of CNA. Loses of one small region in 1p36.33, with gene *CDK11B,* were associated with poor prognosis. More than 66% of the recurrent CNA were validated in the TCGA data when analysed with the same procedure. Also 79% of the genes with altered expression in our data were validated in the TCGA.

Conclusion: Though CNA are frequent events in MSS CRC, few focal recurrent regions were found. These aberrations have strong effects on gene expression and contribute to deregulate relevant cancer pathways. Due to the diploid nature of stromal cells, it is important to consider the purity of tumour samples to accurately calculate CNA events in CRC.

INTRODUCTION

Colorectal cancer (CRC) is the third most commonly diagnosed cancer in the world and the second leading cause of cancer deaths in both men and women. CRC's incidence trend is still increasing in most countries (Ferlay *et al*, 2013). At a molecular level, CRC is a complex disease involving different alterations (An *et al*, 2015). Large chromosomal aberrations have been described in colon tumours, with recurrent gains in chromosome arms 7p, 8q, 13q and 20 and losses in 8p, 17p and 18 (Ashktorab *et al*, 2010; Brosens *et al*, 2011; Goossens-Beumer *et al*, 2015; Meijer *et al*, 1998). Tsafrir et al. showed that tumour copy number aberrations (CNA) may lead to changes in gene expression relevant in colorectal carcinogenesis (Tsafrir *et al*, 2006). In particular, genes in amplified chromosome regions (7p, 8q, 13q, and 20q) usually were over-expressed and genes in regions with chromosome losses (1p, 4, 5q, 8p, 14q, 15q, and 18) were under-expressed. These aberrations can lead to the silencing or amplification of tumour suppressor genes, oncogenes, or non-coding RNAs that modify the expression of genes. Some examples of the relevance of CNA in CRC are losses of chromosome 17p, which contains tumour suppressor genes *TP53* and *MAP2K4* (Han *et al*, 2013); and gains in 7q31 associated with *WNT2* overexpression which alters *Wnt* signalling activation (Wang *et al*, 2016). Gains in 20q have been studied in more detail because they are associated with poor prognosis in CRC (Hidaka *et al*, 2000;Wang *et al*, 2016). This amplification is correlated with the overexpression of *TPX2* and *AURKA* genes (Sillars-Hardebol *et al*, 2012;Wang *et al*, 2016) both involved in processes that promote colorectal adenoma to carcinoma progression, cell viability, and the anchorage-independent growth and invasion processes. This region also harbours *C20orf24* (20q11.23), *ADRM1* (20q13.33), *TCFL5* (20q13.33), *PLCG1* (20q12) and *TH1L* (20q13.32), genes that have been highlighted for their importance in chromosomal instability and adenoma to carcinoma progression (Ali Hassan *et al*, 2014; Loo *et al*, 2013; Sokolova *et al*, 2016). However, the relationship between CNA and gene expression is complex and still not completely defined. Also, difficulties in the methodology to define CNAs from SNP arrays may explain some of the heterogeneity in the results reported so far.

CRC tumours have been classically classified into microsatellite instable (MSI), derived from deficient DNA mismatch repair machinery which leads to hyper-mutated tumours, and microsatellite stable (MSS), also referred to as chromosomal instable tumours (CIN). MSS tumours often show CNA (Brosens *et al*, 2011; The Cancer Genome Atlas Network, 2012; Trautmann *et al*, 2006; Xie *et al*, 2012) and follow the classical adenoma-to-carcinoma progression model (Brosens *et al*, 2011). Recently, consensus molecular subtypes of CRC have been defined by means of non-supervised classification techniques using gene expression data (Guinney *et al*, 2015). This classification establishes four major subtypes (CMS1-4) with specific molecular characteristics. CMS1 (14% of CRC) comprises tumours associated to a MSI phenotype and to immune pathways activation. This subtype usually has the best prognosis. CMS2 (41% of CRC) is characterized by high CIN and strong *WNT*/*MYC* pathways activation. CMS3 (8% of CRC) show low CIN, but are generally *KRAS* mutant and have activated pathways related to energy metabolism. Finally, CMS4 (20% of CRC) show up-regulation of *TGF-β* signalling and have been associated with the worst survival and poor response to chemotherapy. Some controversy exists around whether tumours of CMS4 subtype exhibit a mesenchymal phenotype or are enriched in the stromal component, since genes up-regulated in this subtype are mainly expressed by stromal cells rather than by epithelial cells (Isella *et al*, 2015). Indeed, this is an important issue to consider in copy number analysis, since the diploid nature of stromal cells amalgamated within the tumour bulk could mask real CNA changes in cancer epithelial cells.

In this study we have performed a detailed characterization of CNA in stage II, MSS colon tumours, taking into account the quantity of diploid stromal cells, which was estimated for each tumour sample. Moreover, we have explored the relation of these aberrations with gene expression changes and characteristics of the tumours such as molecular subtyping and prognosis, aiming to decipher the complex biology underlying colon cancer.

METHODS

Patients and samples

Tumour tissues and their paired adjacent normal mucosa from 100 stage II, MSS colon cancer patients have been molecularly profiled to obtain data on copy number, gene expression, methylation and somatic mutations (with exome sequencing in a subset of 42 samples; Colonomics project: www.colonomics.org; NCBI BioProject PRJNA188510). All patients were treated with radical surgery, did not receive adjuvant therapy and have been followed-up a minimum of three years (Supplementary Table 1). Adjacent normal tissue was dissected at pathology from the proximal tumour resection margin with a minimum distance of 10 cm to the tumour lesion. All patients were recruited at the Bellvitge University Hospital (Spain) between 1998 and 2002, provided written informed consent and the hospital Ethics Committee approved the protocol with reference PR074/11.

Copy number and gene expression data on 222 colon tumours and 22 normal adjacent tissues from The Cancer Genome Atlas (TCGA) repository (The Cancer Genome Atlas Network, 2012) were downloaded and used as a validation dataset. These tumours were selected because gene expression was available on the Agilent array platform, equivalent to our setting, that was convenient to estimate the stromal content. To use the maximum sample size, both colon and rectal samples and diverse stages at diagnosis were included (Supplementary Table 1).

Copy Number Analysis

CNA were inferred from the analysis of Affymetrix Genome-Wide Human SNP Array 6.0 genotyping arrays (Carter, 2007; Eldai *et al*, 2013). This array includes probes for the detection of over 906,600 SNPs and an additional 946,000 non-polymorphic oligonucleotides for the assessment of copy number variation. The average inter-marker distance was less than 700 bp. Affymetrix Power Tools (Version 1.16.1) software was used (Eckel-Passow *et al*, 2011), with default parameters, to assess a quantitative locus-level copy number estimate

(CNE) for each tumour sample, as $CNE = \log_2\left(\frac{\hat{\theta}_{ij}}{\hat{\theta}_{Rj}}\right)$, where $\hat{\theta}_{ij}$ was the normalized intensity at probe *j* for sample *i* and $\hat{\theta}_{Rj}$ was a reference intensity at probe *j*, typically representing the mean diploid signal derived from the average pool of normal mucosa samples.

Segmentation

A segmentation algorithm was applied to split the set of ordered locus-specific CNE into regions of adjacent elements that had similar CNE. Each region was assigned a unique value that represented the average CNE of segment. Segmentation was performed for each sample in three steps: 1) Normalization: a smoothing spline was fitted to the raw data and used to normalize the distribution of samples CNE; 2) Raw partition: the *Vega* R package (Morganella ***et al***, 2010) was used to locate change-points in CNE patterns that would split each chromosome into discrete segments. 3) Consolidation: a t-test was used to compare the CNE values between consecutive regions and those with similar CNE values (p-value<0.0001) were merged.

Tumours with high stroma content, which is assumed to be diploid, could bias the CNA measure in cancer cells due to a masking effect. For this reason, the threshold value to identify a CNA from the CNE was defined as a function of the estimated proportion of stroma in the tumour. The stromal proportion in each sample was calculated with the *ESTIMATE* R package from gene expression data (Yoshihara ***et al***, 2013). A hierarchical cluster analysis was used to group the tumour samples into four clusters reflecting their different levels of stromal content, and a varying cut-off was assigned to each cluster: ±0.5 for low stromal, =±0.4 for medium-low stromal, ±0.3 for medium-high stromal, and ±0.2 for high stromal (Supplementary Figure 1). CNE that exceeded these cut-offs were considered aberrations (gains or losses). The proportion of altered genome was estimated for each tumour by summing the length of regions with CNA.

Data availability

Segmented data for each sample is freely available to download at the project website: https://www.colonomics.org/data and the raw data have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession number (EGAS00001002453).

Recurrent aberrations in chromosomal segments

The next step was the alignment of segments for all tumours to identify those with recurrent aberrations. The first step was to calculate minimal recurrent regions based on segmentation for each sample. The minimum recurrent regions were small regions with at least 5% of individuals with CNA. These were the smallest units of analysis used in this study. Focal regions with CNA were also calculated. Focal regions were defined as a set of minimal recurrent regions consecutive with the same CNA sign. Note that two consecutive regions with similar CNA may exist but defined by different tumours contributing to each minimal region. This may dilute further associations analysed at this level of aggregation.

Focal regions were analysed with the GAIA package (Morganella *et al*, 2011; Yuan *et al*, 2012), that allowed using different cut-offs according to the proportion of stroma. GAIA uses a statistical framework based on a conservative permutation test that estimates the null probability distribution of CNA based on the observed data. A stringent false discovery rate (FDR<1e-5) was used to identify focal regions.

Finally, the third level of analysis considered broad events. A broad event was defined as a gain or loss of more than 50% of a chromosome arm. A permutation test was performed to detect these changes. In this test, each sample was randomly assigned a CNA status by chromosome and arm to define the null distribution. From that, the significance of observing a loss of more than 50% of a chromosome arm was calculated.

Molecular characterization of the tumour samples

The *CMSclassifier* R package was used to classify our samples into the four CRC consensus molecular subtypes (CMS), using a random forest approach (Guinney ***et al***, 2015). Tumour CIMP classification was derived from the methylation status in CpG islands of genes *MLH1, RUNX3, CACNA1G, IGF2, NEUROG1, SOCS1, CRABP1* and *CDKN2A*. CIMP high cut-off was set at >= 6/8 methylated promoters, CIMP low was defined as the presence of 1/8 to 5/8 methylated markers, and No CIMP as 0/8 methylated markers (Ogino ***et al***, 2007). The frequency of somatic mutations located in coding regions was assessed for a subset of 42 samples with whole exome sequencing results (Sanz-Pamplona et al, 2015). Briefly, Genomic DNA from the set of 42 adjacent tumour paired samples was sequenced in the National Center of Genomic Analysis (Barcelona, Spain; CNAG) using the Illumina HiSeq-2000 platform. Exome capture was performed with the commercial kit Sure Select XT Human All Exon 50MB (Agilent). Tumour exomes were sequenced at 60X coverage and exomes from adjacent tissues were sequenced at 40X. Bowtie 2.0 software was used to align sequences over the human reference genome. Variant calling was executed with GATK software, and low-quality variants (mapping quality below 30, read depth below 10 or frequency < 10%) were discarded. Germline variants were also removed, that is, variants that were present in normal adjacent paired sequence for each tumour and variants reported in the 1000G project. Finally, variants were annotated using the SeattleSeq Variant Annotation web tool. Mutation data is freely available to download at the project website: https://www.colonomics.org/data.

Association of CNA with clinical and molecular features and prognosis

Non-parametric tests were used to assess the association between clinical or molecular features and the proportion of altered genome or minimal recurrent regions. Separate analyses for gains and losses were also performed.

The Kaplan-Meier method was used to estimate disease free survival curves for each CNA state in a specific region. A total of 21 progression events had been observed in the sample with a minimum follow-up of three years (median 5 years). Multivariate proportional hazards models were used to assess CNA gains and losses as independent prognostic predictors, adjusted for age, sex, tumour location and the proportion of stroma. Only minimal recurrent regions were analysed. False discovery rate (FDR) was used to control for multiple testing in all analyses when minimal recurrent regions were explored.

Association of CNA with gene expression changes

Gene expression data, assessed by Affymetrix Human Genome U219 expression array has been previously analysed (Sanz-Pamplona *et al*, 2014). A unique expression value for each gene was estimated from multiple probes using the first principal component to capture maximal variability among probes. The analyses were focused on gene expression changes between tumour and paired adjacent normal. Thus, gene expression differences were analysed in relation to CNA status (loss, diploid, gain) with linear models, adjusted for age, sex, tumour location and stromal content. Also, partial Pearson correlation was calculated to assess adjusted correlations between the quantitative CNE at each region and gene expression changes. These analyses were restricted to 14,654 genes (out of 18,902 annotated in the microarray) that had enough variability among samples (standard deviation>0.2). Two analyses (*cis* and *trans*) were performed. The former only interrogated genes located within each minimum recurrent region (FDR<0.05). The later assessed associations of all CNA with all genes, except for genes located in chromosomes X and Y,

and Bonferroni method was used to adjust for multiple comparisons. These analyses were replicated using TCGA data.

Functional analysis

A functional analysis was performed to characterise the list of genes showing significant associations with CNA. The *Sigora* R package was used, which focuses on genes or gene-pairs that are (as a combination) specific to a single pathway (Foroushani et al, 2013). This analysis was restricted to genes with a strong association to CNA (FDR<0.05 and r^2>0.33). Also, transcription factor (TF) enrichment was assessed using Fisher's exact test.

RESULTS

Proportion of altered genome. Correlation with clinical and molecular characteristics

CNA events were detected in all samples, with a range of altered genome between 0.04 and 26.6% (Figure 1A). Despite a homogeneous group of MSS stage II tumours were analysed, high variability among samples was observed. Interestingly, 10% of the tumours had less than 0.1% of the genome altered. A more detailed analysis, dividing CNA between gains and losses, revealed tumours that essentially only showed either gains or losses. The proportion of gained genome ranged between 0.007 and 10.4% (Figure 1B) whereas the proportion of lost genome ranged between 0.02 and 16.9% (Figure 1C). The proportion of altered genome was independent of age, sex, site location, progression, CIMP, consensus molecular subtype (CMS) and stromal infiltration (Table 1). The latter is not strange, since the proportion of altered genome at each tumour was adjusted by the proportion of stroma. However, a significant positive correlation was observed between the number of somatic mutations and total CNA (Spearman r=0.42, *P*=0.006) (Supplementary Figure 2). This association, which analysis was restricted to the 42 samples with whole-exome sequencing available, was weaker and no longer significant when gains and losses were analysed separately.

Minimal recurrent regions. Correlation with clinical and molecular characteristics

A total of 26,423 segments with CNA (10,777 gains and 15,646 losses) were identified. The median number of altered segments per tumour was 53 gains, and 118 losses. These segments were transformed into 13,279 minimal recurrent regions (MRR), defined as CNA segments shared by at least 5 samples (5%) (Supplementary Table 2). Figure 2 shows the chromosomal distribution and frequency of the MRR (both gains and losses). It should be noted that 54% of these regions were located in recurrent regions already described in CRC (recurrent gains in chromosome arms 7p, 8q, 13q and 20, and recurrent losses in 8p, 17p and 18). Interestingly, 116 of these MRR were shared by more than 50% of the samples (Table 2 shows a summary of these regions). Only three of these regions included genes: *GSTM1* in 1p13.3*, SIRPB1* in 20p13 and *ADAM5/ADAM3A* in 8p11. The median number of samples per MRR was 8 (inter-quartile range 6 to 66). This small number of affected samples at each segment limited the power to detect associations with clinical variables. Indeed, no relevant association between MRR and any clinical characteristic was found (FDR>0.05). The association of all MRR with prognosis was also evaluated. After correction for multiple testing (FDR<0.05), only one region in 1p36.33 (chr1:1,627,906-1,628,405) was found to be statistically associated with disease free survival (*P*=0.00002). Tumours losing these region (n=6) showed poor prognosis in comparison with diploid ones (Supplementary Figure 3). Gene *CDK11B* is located within this region.

Minimal recurrent regions. Correlation with gene expression

CIS analysis. Only one third (4,292/13,279) of the MRR contained genes. Some large genes were included in more than one region. The linear models for the 2,168 genes included in such regions revealed that 785 of them (36.2% in 545 MRR) showed a significant relationship between the differences of expression and the CNA state at FDR<0.05 (Supplementary Table 3). The median of the partial Pearson correlation coefficient between the differences of expression and CNE of these 785 genes was of 0.54 (inter-quartile range 0.20 to 0.81), indicating that CNA explained a large fraction of the variability in gene expression changes between normal and tumour tissues. To note, potential stromal

contamination was considered in these analyses by adjusting each test for the stromal content of each sample. Interestingly, 64 out of the 785 differentially expressed genes showed a partial correlation with CNE higher than 0.7 (more that 50% of variance explained; see some examples in Figure 3A-D).

As expected, these genes were mainly located on chromosomes 6, 7, 8, 13, 17, 18 and 20, because these are the region's most often showing CNA (Table 3 and Figure 3E). Also unsurprisingly, CNA gains were associated with higher gene expression and CNA losses were associated with lower gene expression levels. This happened in 236 genes located in gained regions and 30 genes located in lost regions, respectively . Also, the expression of genes located in regions in which both losses and gains had been observed (n=266) showed a good correlation with the quantitative CNE in each tumour, which can be interpreted as proportional to the average number of DNA copies. There were 117 MRR in which more than one gene showed significant changes in gene expression (20.6% of 567 MRR with >1 gene). Reinforcing the idea that copy number alterations are associated with changes in gene expression, 78% (92 of 117) of these regions showed half or more of the genes with altered expression. Specifically, 38 of them (32%) were MRR in which all included genes had consistent significant changes in expression.

A functional analysis was performed with the 325 genes whose changes in expression were strongly associated to CNA (FDR<0.05 and r^2>0.33). A significant enrichment was identified in "*Colorectal cancer pathway*". This analysis also identified a significant enrichment of pathways related to "*RNA degradation*", "*Endocytosis*", "*Basal transcription factors*" or "*Glycerophospholipid biosynthesis*"; among others (Supplementary Table 4 and Supplementary Table 5).

***Trans analysis*.** Under the hypothesis that CNA could also have long distance effects (*trans*) on gene expression, due to regulatory effects, the association between CNA and gene

expression changes of all annotated genes was evaluated. This analysis explored 15,225,697 relationships (the expression of 14,654 variable genes and 13,279 MRR; *cis* relationships previously analysed were excluded). Of them, only 191 were significant after Bonferroni correction (*P*<3.3e-9), involving 42 genes and 168 MRR (Figure 3E). All relationships were between genes and regions located in different chromosomes. Unexpectedly, 105 out of the 168 MRR (62.5%) did not contain genes, pointing to regulatory elements different from transcription factor activity. The remaining CNA regions (n=63) included 53 genes. We tested if these genes were predominantly transcription factors, but only 4 of them (*GATA3, ST18, PRDM6, ZNF641*) had this function, while we expected 10% by chance alone (Supplementary Table 6).

Focal regions and broad events. Correlation with clinical and molecular characteristics

Larger *focal regions* were identified from specific MRR, defined as consecutive minimal recurrent regions with the same CNA status (but possibly different quantitative CNE value). If one of these regions involved more than 50% of a chromosome arm, it was defined as a *broad event*.

From 13,279 MRR, 353 focal regions were found (97 focal gains and 256 focal losses; Figure 3E; Supplementary Table 7). The median number of samples with some aberration in these focal regions was 11 (inter-quartile range 8 to 14). The focal regions represented 12.5% of the altered genome (9.8% in lost focal regions, 2.7% in gained focal regions). These focal regions were enriched in genes, since 26% of the total number of human genes was in these CNA regions (16% in lost focal regions, 10% in gained focal regions). However, no significant associations were found between the average CNE in these focal regions and any of the clinical characteristics explored, including prognosis.

Five recurrent broad regions were identified: gains in 8q (6% of the samples), 13q (7% of the samples), 20p (6% of the samples) and 20q (24% of the samples), and losses in 8p (7% of the samples). No significant association between recurrent broad CNA and clinical characteristics or prognosis was found, except for gains in 20q, which were related to the number of somatic mutations (*P*=0.00006). Other classically altered regions in CRC were also detected, but at lower frequency: 7p gain (n=4), 17p loss (n=4), and 18q loss (n=3). Indeed, if less astringent criteria were used to detect broad regions, more altered tumours emerged (Supplementary Table 8).

Validation in TCGA data

To assess the consistency of our findings, a validation was performed using the TCGA dataset comprising 222 CRC tumours. To ensure a comparable data, the same pipeline of analysis used in our samples was followed starting from the raw TCGA data. In agreement with our results, the range of altered genome was 0.038 to 28.1% (0.004-12.1% gains and 0.025-22.5% losses). Unexpectedly, the proportion of altered genome in the TCGA showed a significant negative association with the number of somatic mutations (Spearman r= -0.15, P=0.03). Nevertheless, when only MSS samples were considered, this negative correlation changed and a non-significant positive correlation between the number of mutations and the proportion of lost genome emerged (Spearman r=0.14, P=0.08). This change in correlation derives from the fact that MSI tumours are hyper-mutated and usually diploid. Interestingly, a strong association between CMS and the proportion of altered genome was found in TCGA data. Subtypes CMS2 and CMS4 accumulated higher levels of chromosomal alterations than CMS1 and CMS3 (Supplementary Figure 4).

A total of 8771 MRR (66% of 13279) were validated in TCGA samples, and the agreement was very high for MRR altered in more than 50% of the samples (Table 2). 4105 MRR were identified in TCGA that we had not previously observed in our data. If only stage II and III

MSS tumours were considered, the percentage of MRR validated in TCGA increased to 69% (n=51) and 68% (n=44) respectively. Finally, when comparing samples from different stages in TCGA dataset, 72% of MRR from stage II tumours were found in stage III tumours.

Regarding the association with gene expression, it should be noted that the TGCA dataset only analysed 22 normal tissues, thus, the tumour-normal changes have been estimated respect to the average expression of these normal in an unpaired analysis. Also, only 631 out of the 785 significant genes were found in the TCGA validation dataset. For this subset, 79% (496 genes) of our gene expression-CNA associations were replicated, thus confirming that expression levels of such genes were in part explained by CNA in colon cancer (Table 3, Supplementary Table 3).

The *trans* validation was performed in 127 out of the 191 associations because 19 genes and 45 MRR were not found in the TCGA. From these, only 64 relationships (50%) were confirmed in the TCGA dataset, which indicated that some of our findings could be spurious even though we used Bonferroni correction to protect from false positive findings.

Concerning focal and broad events, 51% of the focal regions were validated. Surprisingly, almost all were lost regions. Only 10 out of 97 focal gained regions were validated and all broad regions were validated.

DISCUSSION

This comprehensive analysis confirms that CNA are frequent in MSS colon tumours, and probably induce relevant changes in gene expression that alter key cancer pathways. Even though all analysed samples were MSS, stage II colon tumours, a high heterogeneity in CNA among them has been observed, both in the percentage of altered genome and the location of the CNA.

The percentage of altered genome ranged from 0.04 to 26.6% (mean 2.6%). This percentage, validated in TCGA data when the same methodology was used to define CNA, is lower than the reported in previous studies (Brosens et al, 2011; Trautmann et al, 2006; Xie et al, 2012). A probable reason is the rigorous cut-off used in our analysis, selected in such a way to reduce the number of false positive CNA that could attenuate the associations with gene expression. However, the frequency of recurrent CNA regions found in our study is consistent with previous reports, with gains in chromosomes 7, 8, 13, and 20, and losses in chromosomes 8, 17 and 18 (Ashktorab et al, 2010; Brosens et al, 2011; Tsafrir et al, 2006; Xie et al, 2012).

In this study we have paid special attention to the stromal content of tumours. An initial analysis that used a fixed cut-off (±0.4) for all samples revealed a strong association between CNA and the proportion of stroma on the tumours. Specifically, tumours with molecular subtype CMS4, that are characterized by high stromal content (Calon *et al*, 2015; Isella *et al*, 2015), also showed a reduced frequency of CNA. Therefore, and since stromal cells are diploid, we thought that this result could be a biased estimation of CNA in tumours with high stromal content due to a dilution effect. Moreover, since other studies have described and association of CNA with poor prognosis (Andersen *et al*, 2011; Kurashina *et al*, 2008; Orsetti *et al*, 2014), it seemed paradoxical that the CMS4 subtype that has poor prognosis was the less altered subtype. Based on this observation, we adjusted the cut-off to define a CNA as a function of the stromal content of the sample. After this correction, which

we consider less biased, no significant associations were observed between CNA and stromal infiltration. Though not statistically significant, differences were observed per molecular subtypes. CNA were more frequent in CMS4 and CMS2 tumours than in CMS1 and CMS3 (Supplementary Figure 5). In the analysis of the TCGA data we found a statistically significant CNA enrichment in CMS2 and CMS4 tumours. This result agrees with the reported in the study describing the molecular characteristics of CRC consensus molecular subtypes (Guinney *et al*, 2015). To note, the subtype CMS4, that had the least CNA when stromal component was not considered, emerged as the subtype with more CNA changes when adjusting for stromal content. This observation should be taken into account when interpreting the differences in CNA among tumours with diverse proportion of stroma in studies that have not adjusted this effect.

Most MRR identified in our tumours were validated in the TCGA dataset, confirming the validity of our analysis. Furthermore, this percentage was high when only MSS stage II tumours were used for validation purposes. Regarding focal regions, it is interesting to note that almost all lost regions were validated in TCGA data whereas only 10% of gains were validated. This result suggests that a higher heterogeneity in gained events across patients exists whereas lost events are prone to be more recurrent. Also, all described broad events were validated in TCGA data, confirming their validity.

Aberrations in copy number are relevant for the consequences in gene dosage that may produce. This can have a direct effect on the protein levels of genes located in regions with CNA or mediated through modifications in regulatory elements. As expected, we have observed that a large fraction of expression changes in colon tumours can be explained by changes in CNA in the regions where these genes are located. Also, when a region contains multiple genes, most of them change their level of expression in a similar pattern. Nevertheless, though frequent, this is not a general mechanism of gene expression alteration in tumours, since not all genes located in CNA regions change their levels of expression

between normal and tumour samples. Indeed, almost 15% of genes were located in CNA regions but only 36% of them changed their level of expression between normal and tumour tissues in a way that might be causal. As expected, most expression changes directly followed the change in gene dosage (though some non-significant exceptions have been observed, possibly due to multiple comparisons). Indeed, this relation has been widely described in CRC (Tsafrir et al, 2006; Sillars-Hardebol et al, 2012; Wang et al, 2016).

Gene expression regulation is complex. In addition to these direct relationships, *trans* associations among CNA and gene expression were also found. We hypothesized that transcription factors located in CNA regions could explain changes in level of expression of genes located in distant regions of the genome. However, only 3% of such genes are known transcription factors (and we expected 10%). What is more, 105 out of 168 CNA regions implicated in *trans* relationship did not contain genes, thus alternative regulation mechanisms, possibly involving enhancers, methylation or non-coding RNAs must be involved in these long distance effects of CNA in gene expression changes. It is reassuring that most *cis* (79%) and some *trans* (50%) relationships were validated using TCGA public data.

We also assessed the association of CNA with clinical and molecular parameters. We found that tumours with higher number of CNA also exhibited higher number of somatic mutations (though this association was restricted to 42 tumours with exome data). Since only MSS tumours have been included in this analysis, we could hypothesize that the inverse relationship between CNA and mutational load previously described only emerged when MSS tumours were compared with hyper mutant MSI tumours. Indeed, this inverse relationship was observed in the TCGA validation dataset, which included MSI tumours. When only MSS tumours were considered, in line with our results, the trend is towards a positive correlation of aberrations (CNA and somatic mutations). Interestingly, these CNA are likely to be segment losses, which might be related to a requirement of double hit for many mutations to be active.

Specific CNA have been previously suggested as prognostic biomarkers (Brosens *et al*, 2011; Wang *et al*, 2016; Zhang *et al*, 2015). In our data, we have only found one region in 1p36.33 significantly associated with prognosis when multiple comparisons were considered. Tumours with this region lost showed worse prognosis than diploid tumours. This region contains *CDK11B* gene, which encodes for a cyclin-dependent kinase that plays multiple roles in cell cycle progression and apoptosis regulation. Thus, we hypothesize that in a subset of colon tumours *CDK11B* could act as a tumour suppressor gene. However, due to the small number of cancer recurrence events in our study (21 out of 99 patients) we cannot exclude the possibility that this region was associated with prognosis just by chance. This result could not be validated in TCGA data because the follow-up information of the individuals has poor quality, so it deserves further study.

Although originally developed to assess genetic diversity, genotyping arrays have emerged as a useful technology to identify regions with CNA. It is particularly important to highlight the high rates of false positive focal regions that can result by using these high-throughput techniques. For this reason, we have used a conservative and variable threshold according to the proportion of stroma for each sample. The selection of a method to call CNA regions represents a great challenge because there are many available, usually with little experimental validation, and the results are not necessarily consistent. So far, little work has been deserved to compare results obtained through different methods among them (Koike *et al*, 2011; Morganella *et al*, 2010). After exploring diverse software tools, we selected a method that provided more precise results when focal CNA regions were visually inspected. Also, the results obtained regarding the frequency and chromosomal distribution of CNA were similar to previously reported for CRC using different methods, thus reassuring the validity of our approach (Morganella *et al*, 2010; Morganella *et al*, 2011; Rueda & Diaz-Uriarte, 2010). Smaller regions with CNA observed in multiple samples help to better identify potential causal genes behind the observed associations. Larger focal regions, as identified

by the GAIA software, paradoxically decrease the power to detect associations with clinical variables, because the enlarged region usually combines samples with heterogeneous CNA.

In conclusion, this comprehensive analysis has shown that CNA are highly frequent and heterogeneous events in MSS stage II colon tumours. The variation of gene expression between tumour tissues and their paired adjacent normal mucosa was explained by CNA on 36% of the genes affected by this type of aberrations, and genes often altered belong to key cancer pathways. These altered genes by CNA represent 5% of the total number of genes expressed in the colon.

Also, from a methodological perspective, we have found that the proportion of tumour stroma may bias the estimation of CNA. To avoid this effect, an adjusted cut-off definition proportional to the estimated stromal content produced more accurate results.

Acknowledgments

The authors would like to thank Isabel Padrol, Pilar Medina and Carmen Atencia for their technical assistance. The “Xarxa de Bancs de Tumors de Catalunya” sponsored by “Pla Director d'Oncologia de Catalunya (XBTC)”, the ICOBiobanc and PLATAFORMA BIOBANCOS PT13/0010/0013 helped with sample collection.

Conflict of interest

The authors declare no conflict of interest.

REFERENCES

Ali Hassan NZ, Mokhtar NM, Kok Sin T, Mohamed Rose I, Sagap I, Harun R, Jamal R (2014) Integrated analysis of copy number variation and genome-wide expression profiling in colorectal cancer tissues. *PloS one* **9**(4)**:** e92553

An N, Yang X, Cheng S, Wang G, Zhang K (2015) Developmental genes significantly afflicted by aberrant promoter methylation and somatic mutation predict overall survival of late-stage colorectal cancer. *Scientific reports* **5:** 18616

Andersen CL, Lamy P, Thorsen K, Kjeldsen E, Wikman F, Villesen P, Oster B, Laurberg S, Orntoft TF (2011) Frequent genomic loss at chr16p13.2 is associated with poor prognosis in colorectal cancer. *International journal of cancer* **129**(8)**:** 1848-58

Ashktorab H, Schaffer AA, Daremipouran M, Smoot DT, Lee E, Brim H (2010) Distinct genetic alterations in colorectal cancer. *PloS one* **5**(1)**:** e8879

Brosens RP, Belt EJ, Haan JC, Buffart TE, Carvalho B, Grabsch H, Quirke P, Cuesta MA, Engel AF, Ylstra B, Meijer GA (2011) Deletion of chromosome 4q predicts outcome in stage II colon cancer patients. *Cellular oncology* **34**(3)**:** 215-23

Calon A, Lonardo E, Berenguer-Llergo A, Espinet E, Hernando-Momblona X, Iglesias M, Sevillano M, Palomo-Ponce S, Tauriello DV, Byrom D, Cortina C, Morral C, Barcelo C, Tosi S, Riera A, Attolini CS, Rossell D, Sancho E, Batlle E (2015) Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nature genetics* **47**(4)**:** 320-9

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature genetics* **39**(7 Suppl)**:** S16-21

Eckel-Passow JE, Atkinson EJ, Maharjan S, Kardia SLR, de Andrade M (2011) Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform. *BMC Bioinformatics* **12**(1)**:** 220

Eldai H, Periyasamy S, Al Qarni S, Al Rodayyan M, Muhammed Mustafa S, Deeb A, Al Sheikh E, Afzal M, Johani M, Yousef Z, Aziz MA (2013) Novel genes associated with colorectal cancer are revealed by high resolution cytogenetic analysis in a patient specific manner. *PloS one* **8**(10)**:** e76251

Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, Comber H, Forman D, Bray F (2013) Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *European journal of cancer* **49**(6)**:** 1374-403

Foroushani AB, Brinkman FS, Lynn DJ (2013) Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures. *PeerJ* **1:** e229

Goossens-Beumer IJ, Oosting J, Corver WE, Janssen MJ, Janssen B, van Workum W, Zeestraten EC, van de Velde CJ, Morreau H, Kuppen PJ, van Wezel T (2015) Copy number alterations and allelic ratio in relation to recurrence of rectal cancer. *BMC genomics* **16:** 438

Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, De Sousa E Melo F, Missiaglia E, Ramay H, Barras D, Homicsko K, Maru D, Manyam GC, Broom B, Boige V, Perez-Villamil B, Laderas T, Salazar R, Gray JW, Hanahan D, Tabernero J, Bernards R, Friend SH, Laurent-Puig P, Medema JP, Sadanandam A, Wessels L, Delorenzi M, Kopetz S, Vermeulen L, Tejpar S (2015) The consensus molecular subtypes of colorectal cancer. *Nature Medicine* **21**(11)**:** 1350-1356

Guo W, Pylayeva Y, Pepe A, Yoshioka T, Muller WJ, Inghirami G, Giancotti FG (2006) Beta 4 integrin amplifies ErbB2 signaling to promote mammary tumorigenesis. *Cell* **126**(3)**:** 489-502

Han SW, Kim HP, Shin JY, Jeong EG, Lee WC, Lee KH, Won JK, Kim TY, Oh DY, Im SA, Bang YJ, Jeong SY, Park KJ, Park JG, Kang GH, Seo JS, Kim JI, Kim TY (2013) Targeted sequencing of cancer-related genes in colorectal cancer using next-generation sequencing. *PloS one* **8**(5)**:** e64271

Hidaka S, Yasutake T, Takeshita H, Kondo M, Tsuji T, Nanashima A, Sawai T, Yamaguchi H, Nakagoe T, Ayabe H, Tagawa Y (2000) Differences in 20q13.2 copy number between colorectal cancers with and without liver metastasis. *Clinical cancer research : an official journal of the American Association for Cancer Research* **6**(7)**:** 2712-7

Isella C, Terrasi A, Bellomo SE, Petti C, Galatola G, Muratore A, Mellano A, Senetta R, Cassenti A, Sonetto C, Inghirami G, Trusolino L, Fekete Z, De Ridder M, Cassoni P, Storme G, Bertotti A, Medico E (2015) Stromal contribution to the colorectal cancer transcriptome. *Nature genetics* **47**(4)**:** 312-9

Koike A, Nishida N, Yamashita D, Tokunaga K (2011) Comparative analysis of copy number variation detection methods and database construction. *BMC genetics* **12:** 29

Kurashina K, Yamashita Y, Ueno T, Koinuma K, Ohashi J, Horie H, Miyakura Y, Hamada T, Haruta H, Hatanaka H, Soda M, Choi YL, Takada S, Yasuda Y, Nagai H, Mano H (2008) Chromosome copy number analysis in screening for prognosis-related genomic regions in colorectal carcinoma. *Cancer science* **99**(9)**:** 1835-40

Loo LW, Tiirikainen M, Cheng I, Lum-Jones A, Seifried A, Church JM, Gryfe R, Weisenberger DJ, Lindor NM, Gallinger S, Haile RW, Duggan DJ, Thibodeau SN, Casey G, Le Marchand L (2013) Integrated analysis of genome-wide copy number alterations and gene expression in microsatellite stable, CpG island methylator phenotype-negative colon cancer. *Genes, chromosomes & cancer* **52**(5)**:** 450-66

Meijer GA, Hermsen MA, Baak JP, van Diest PJ, Meuwissen SG, Belien JA, Hoovers JM, Joenje H, Snijders PJ, Walboomers JM (1998) Progression from colorectal adenoma to carcinoma is associated with non-random chromosomal gains as detected by comparative genomic hybridisation. *J Clin Pathol* **51**(12)**:** 901-9

Morganella S, Cerulo L, Viglietto G, Ceccarelli M (2010) VEGA: variational segmentation for copy number detection. *Bioinformatics* **26**(24)**:** 3020-7

Morganella S, Pagnotta SM, Ceccarelli M (2011) Finding recurrent copy number alterations preserving within-sample homogeneity. *Bioinformatics* **27**(21)**:** 2949-56

Ogino S, Kawasaki T, Kirkner GJ, Kraft P, Loda M, Fuchs CS (2007) Evaluation of markers for CpG island methylator phenotype (CIMP) in colorectal cancer by a large population-based sample. *The Journal of molecular diagnostics : JMD* **9**(3)**:** 305-14

Orsetti B, Selves J, Bascoul-Mollevi C, Lasorsa L, Gordien K, Bibeau F, Massemin B, Paraf F, Soubeyran I, Hostein I, Dapremont V, Guimbaud R, Cazaux C, Longy M, Theillet C (2014) Impact of chromosomal instability on colorectal cancer progression and outcome. *BMC cancer* **14:** 121

Rueda OM, Diaz-Uriarte R (2010) Finding Recurrent Copy Number Alteration Regions A Review of Methods. *Current Bioinformatics* **5**(1)**:** 17

Sanz-Pamplona R, Berenguer A, Cordero D, Molleví DG, Crous-Bou M, Sole X, Paré-Brunet L, Guino E, Salazar R, Santos C, de Oca J, Sanjuan X, Rodriguez-Moranta F, Moreno V

(2014) Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. *Molecular Cancer* **13:** 46

Sanz-Pamplona R, Lopez-Doriga A, Pare-Brunet L, Lazaro K, Bellido F, Alonso MH, Ausso S, Guino E, Beltran S, Castro-Giner F, Gut M, Sanjuan X, Closa A, Cordero D, Moron-Duran FD, Soriano A, Salazar R, Valle L, Moreno V (2015) Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **21**(20)**:** 4709-18

Sillars-Hardebol AH, Carvalho B, Tijssen M, Belien JA, de Wit M, Delis-van Diemen PM, Ponten F, van de Wiel MA, Fijneman RJ, Meijer GA (2012) TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression. *Gut* **61**(11)**:** 1568-75

Sokolova V, Crippa E, Gariboldi M (2016) Integration of genome scale data for identifying new players in colorectal cancer. *World journal of gastroenterology* **22**(2)**:** 534-45

The Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**(7407)**:** 330-7

Trautmann K, Terdiman JP, French AJ, Roydasgupta R, Sein N, Kakar S, Fridlyand J, Snijders AM, Albertson DG, Thibodeau SN, Waldman FM (2006) Chromosomal instability in microsatellite-unstable and stable colon cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **12**(21)**:** 6379-85

Tsafrir D, Bacolod M, Selvanayagam Z, Tsafrir I, Shia J, Zeng Z, Liu H, Krier C, Stengel RF, Barany F, Gerald WL, Paty PB, Domany E, Notterman DA (2006) Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer research* **66**(4)**:** 2129-37

Wang H, Liang L, Fang JY, Xu J (2016) Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers. *Oncogene* **35**(16)**:** 2011-9

Xie T, G DA, Lamb JR, Martin E, Wang K, Tejpar S, Delorenzi M, Bosman FT, Roth AD, Yan P, Bougel S, Di Narzo AF, Popovici V, Budinska E, Mao M, Weinrich SL, Rejto PA, Hodgson JG (2012) A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. *PloS one* **7**(7)**:** e42001

Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, Mills GB, Verhaak RGW (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications* **4**

Yuan X, Zhang J, Zhang S, Yu G, Wang Y (2012) Comparative analysis of methods for identifying recurrent copy number alterations in cancer. *PloS one* **7**(12)**:** e52516

Zhang J, Yan B, Spath SS, Qun H, Cornelius S, Guan D, Shao J, Hagiwara K, Van Waes C, Chen Z, Su X, Bi Y (2015) Integrated transcriptional profiling and genomic analyses reveal RPN2 and HMGB1 as promising biomarkers in colorectal cancer. *Cell & bioscience* **5:** 53

TABLES

Table 1: Association between the percentage of altered genome and clinical characteristics.

Table 2: Summary of minimal recurrent regions (MRR) with more than 50% of the samples altered in the CLX dataset.

Table 3: Chromosomal distribution of genes related to CNA and validation in TCGA data.

FIGURE LEGENDS

Figure 1: Distribution of clinical characteristics according to the proportion of altered genome

The histogram represents the proportion of altered genome by sample (purple: gains, red: losses). In the lower part, the clinical characteristics of the individuals are represented: sex (blue: female, red: male), age (sliding scale from white: minimum to brown: maximum), tumour location (light green: left, dark green: right), development of metastases (light pink: no, dark pink: yes), CIMP (white: no, green: CIMP low, blue: CIMP high), number of mutations (sliding scale from white: 0 to dark blue: maximum), proportion of stroma (light green: low, dark green: high), molecular subtype (yellow: CMS1, blue: CMS2, plink: CMS3, green: CMS4). **A.** All CNA. **B.** Gains. **C.** Losses.

Figure 2: Frequency of CNA by chromosome

Each graph represents a chromosome with chromosomal position in the X-axis. Y-axis displays the percentage of tumour with gains (>0, purple) or losses (<0, red). The height of the bar is proportional to the number of samples showing the CNA change. Dashed lines represent the frequency (black: 5%, green: 20% red: 50%).

Figure 3: Relationship between gene expression and CNA

A-D: Boxplots showing examples of gene expression changes based on CNA levels. Spearman correlation and FDR p value are shown. “L” indicates number of individuals losing the region whereas “G” indicates number of individuals gaining the region. **E.** Circus plot of

CNA recurrent regions and their association with changes in gene expression. Outer circle shows ideograms of the chromosomes. Inner circles show, in order, focal regions (gains in purple, losses in red), broad events, and genomic location of significant associations between CNA and the difference in expression between tumour and adjacent normal (blue) in *cis* analysis. The central arcs indicate genomic locations with significant *trans* associations between CNA and changes of gene expression.

SUPPLEMENTARY FILES

Supplementary Table 1: Baseline characteristics of tumours of CLX and TCGA datasets.

Supplementary Table 2: List of minimal recurrent region (MRR) in CLX and TCGA datasets.

Supplementary Table 3: List of 785 significant genes obtained in the *cis* analysis.

Supplementary Table 4: Significant pathways in which CNA genes are involved.

Supplementary Table 5: Categorization of CNA genes into significant pathways.

Supplementary Table 6: List of transcription factors identified in the *trans* analysis.

Supplementary Table 7: Number of recurrent focal regions in all samples by chromosome.

Supplementary Table 8: Number of individuals altered by percentage of altered chromosomal arm.

Supplementary Figure 1: Steps to calculate cut-off values based on the estimated stroma. The threshold value to define a CNA was defined variable for each tumour, as a function of its estimated stromal proportion. Tumours with a lower proportion of stroma were assigned a higher cut-off of the CNE; and tumours with a higher proportion of stroma were assigned a lower cut-off of the CNE. A hierarchical cluster analysis was used to group the tumour samples into four clusters reflecting their different levels of stromal content. Dendrograms generated from the estimate of the stromal proportion were shown in **A** (CLX dataset) and **B** (TCGA dataset). The cut-off value to call a CNA was determined by the histogram of CNE, such that both the left and right tail areas cover 5, 10, 20 and 30 percent

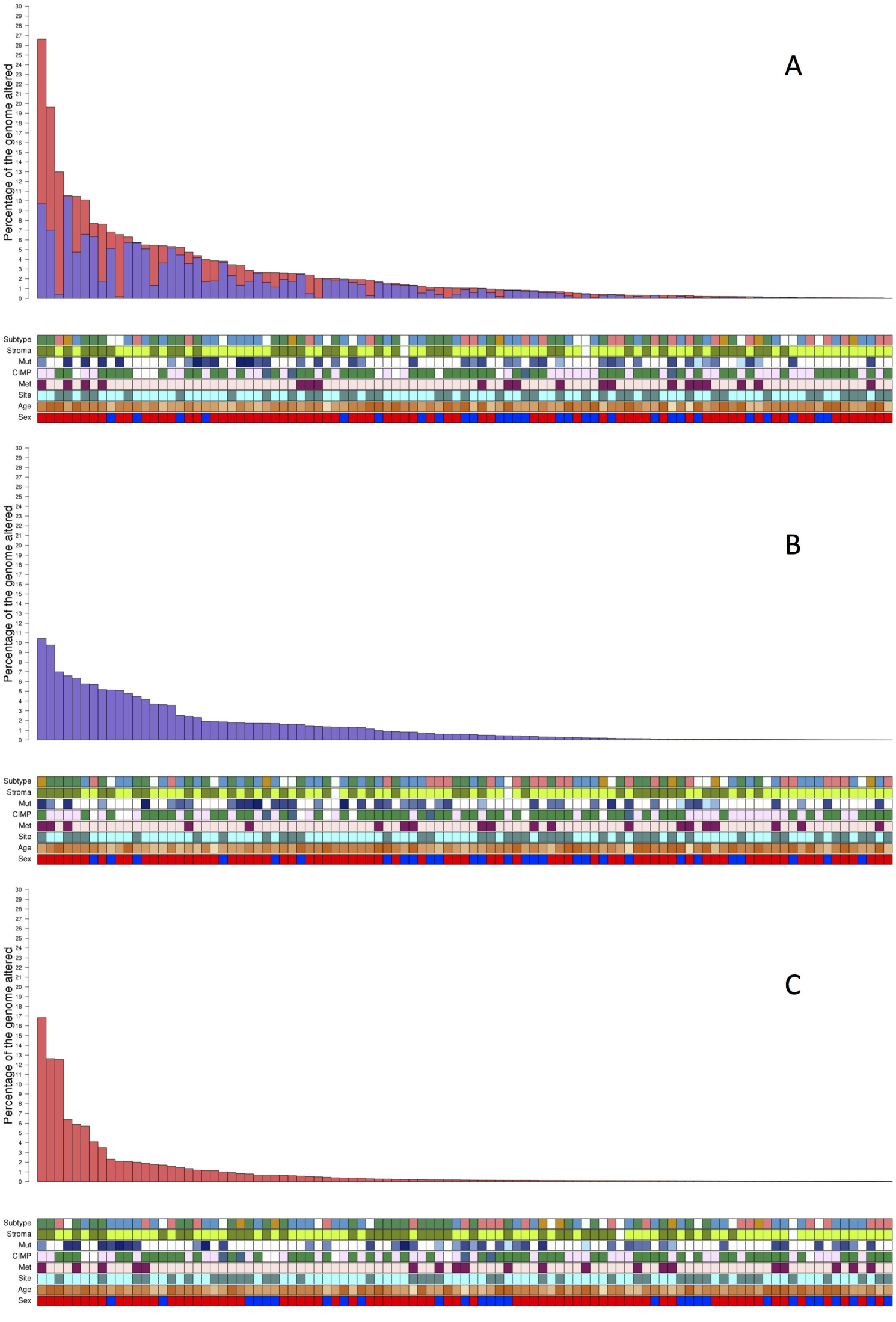
of the whole distribution. The CNE values of each segment were represented on the Y axis. The distribution of the different quartiles of the read depth signal was represented by boxplots in CLX dataset (**C**) and TCGA dataset (**D**). A log2 ratio cut-off was assigned to each stroma cluster: very low stromal (cut-off=±0.5), relatively low stromal (cut-off=±0.4), relatively high stromal (cut-off=±0.3), and high stromal (cut-off=±0.2). The estimated stromal proportion was represented on the Y axis. The cut-off of the CNE were represented in the different boxes in CLX dataset (**E**) and TCGA dataset (**F**).

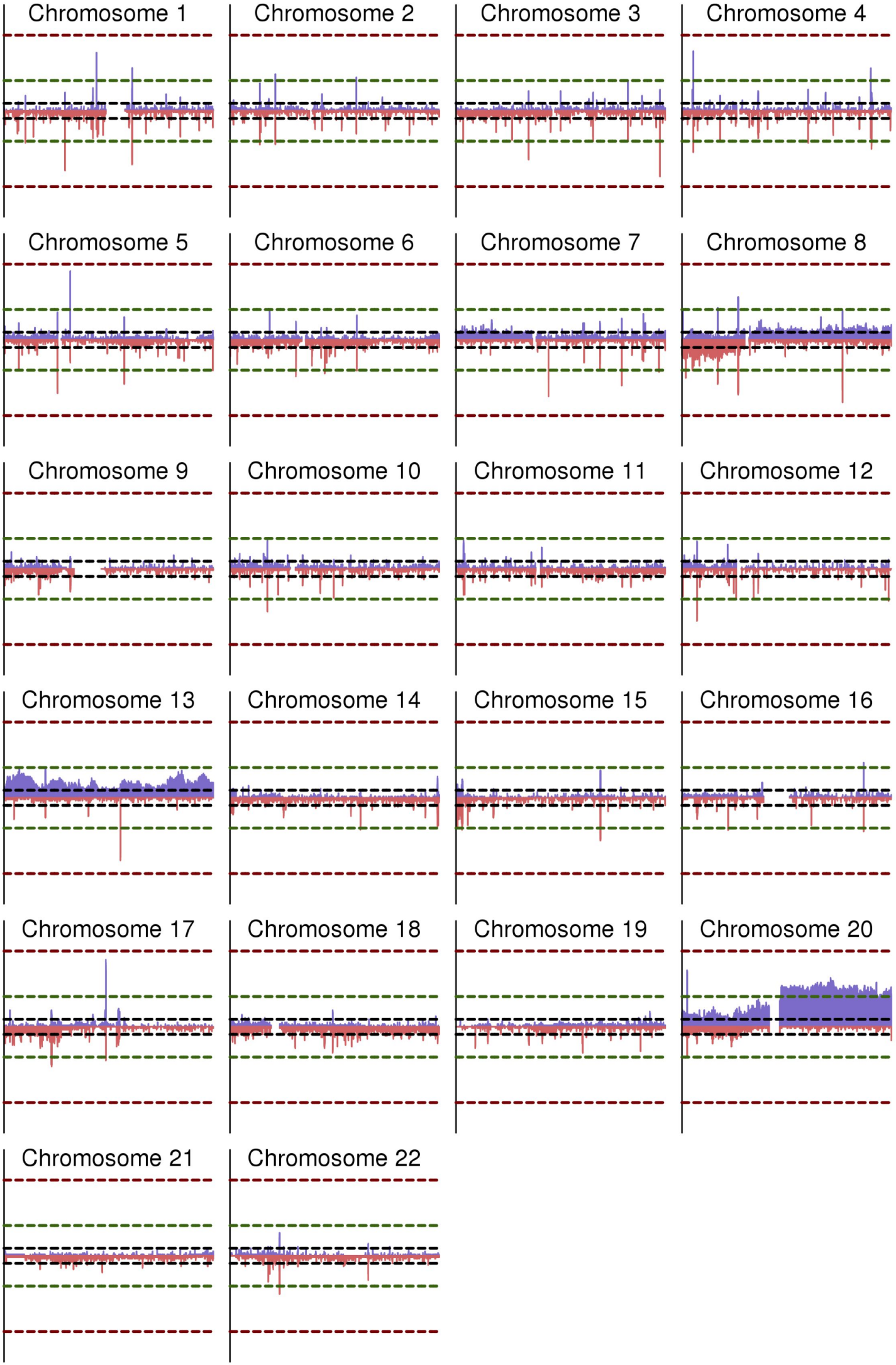
Supplementary Figure 2: Relationship between the proportion of altered genome and the number of mutations in CLX dataset. The p-values for the Kruskal-Wallis test are shown for all CNA **(A)**, gains (**B)** and losses (**C)**.

Supplementary Figure 3: Disease-free survival curves for region in 1p36.33 according to CNA in CXL dataset. Cox p value is showed.

Supplementary Figure 4: Relationship between the proportion of altered genome and the molecular subtype in TCGA dataset. The p-values for the Kruskal-Wallis test are shown for all CNA **(A)**, gains (**B)** and losses (**C)**.

Supplementary Figure 5: Boxplots showing proportion of altered genome according to tumour stromal content (low/high) and consensus molecular subtype (CMS1-4) when fixed and variable cut-offs were used to define CNA in CLX dataset. **(A)** Relationship between the proportion of altered genome and tumour stromal content with a fixed cut-off point of CNA. **(B)** Relationship between the proportion of altered genome and tumour stromal content with a variable cut-off point of CNA. **(C)** Relationship between the proportion of altered genome and CMS with a fixed cut-off point of CNA. **(D)** Relationship between the proportion of altered genome and CMS with a variable cut-off point of CNA.


A
B
C
Percentage of the genome altered
Subtype
Stroma
Mut
CIMP
Met
Site
Age
Sex


Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
Chromosome 13
Chromosome 14
Chromosome 15
Chromosome 16
Chromosome 17
Chromosome 18
Chromosome 19
Chromosome 20
Chromosome 21
Chromosome 22

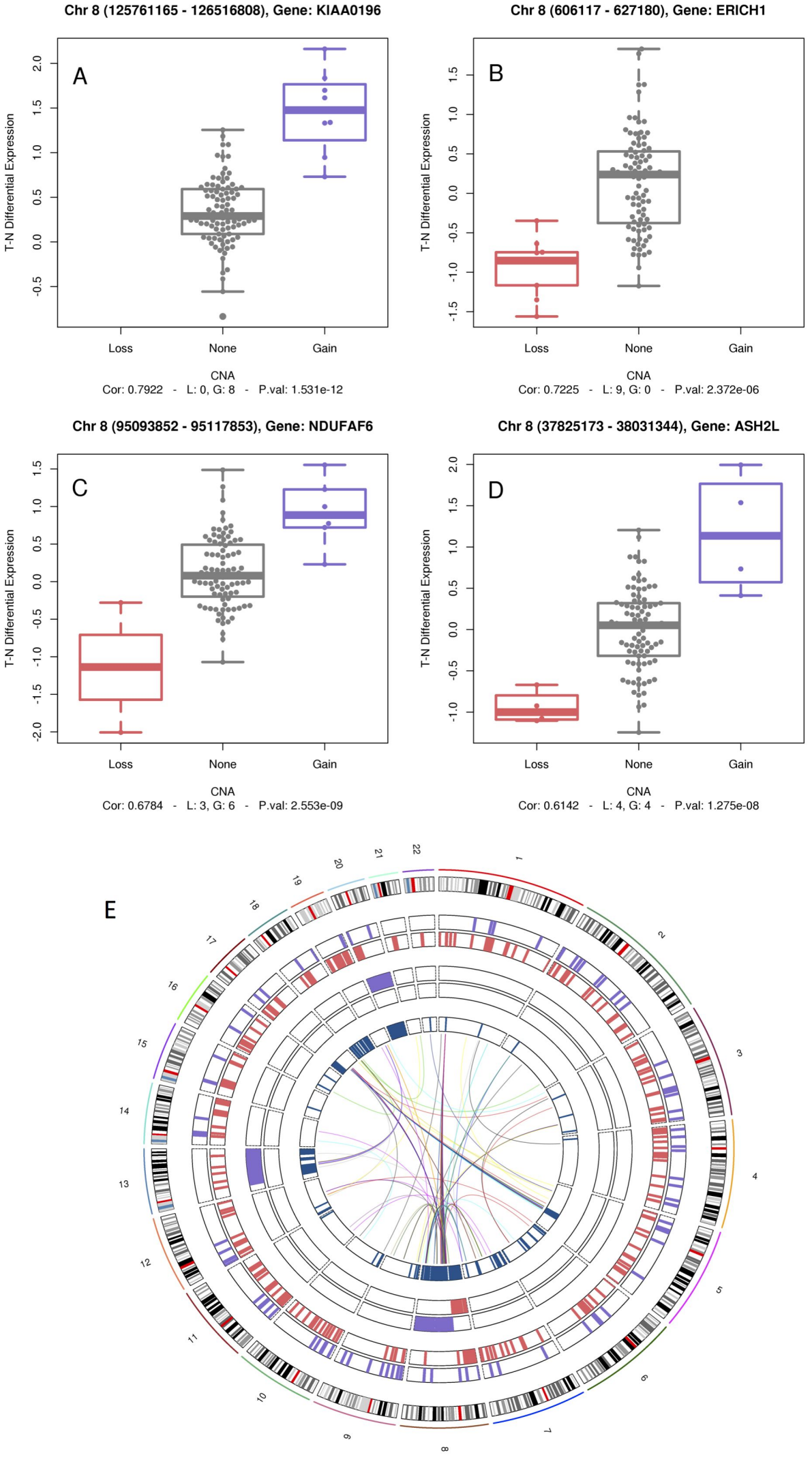

Chr 8 (125761165 - 126516808), Gene: KIAA0196
A
T-N Differential Expression
Loss
None
Gain
CNA
Cor: 0.7922 - L: 0, G: 8 - P.val: 1.531e-12
Chr 8 (606117 - 627180), Gene: ERICH1
B
T-N Differential Expression
Loss
None
Gain
CNA
Cor: 0.7225 - L: 9, G: 0 - P.val: 2.372e-06
Chr 8 (95093852 - 95117853), Gene: NDUFAF6
C
T-N Differential Expression
Loss
None
Gain
CNA
Cor: 0.6784 - L: 3, G: 6 - P.val: 2.553e-09
Chr 8 (37825173 - 38031344), ASH2L
D
T-N Differential Expression
Loss
None
Gain
CNA
Cor: 0.6142 - L: 4, G: 4 - P.val: 1.275e-08
E

Table 1: Association between the percentage of altered genome and clinical characteristics

		Total		Losses		Gains	
		Median*	***P***	**Median***	***P***	**Median***	***P***
Sex	**Female**	0.76	0.11	0.15	0.06	0.43	0.37
	Male	1.56		0.21		0.73	
Age	**<65**	1.07	0.53	0.19	0.7	0.6	0.18
	[65-72)	1.06		0.16		0.73	
	[72-78)	1.92		0.29		1.33	
	>=78	0.63		0.15		0.26	
	Spearman	-0.08	0.42	0.008	0.94	-0.18	0.07
Site	**Left**	1.19	0.24	0.18	0.55	0.84	0.14
	Right	0.84		0.16		0.32	
CIMP	**No CIMP**	0.66	0.86	0.19	0.54	0.35	0.78
	CIMP Low	1.05		0.15		0.56	
	CIMP High	1.7		0.5		1.16	
Molecular subtype	**CMS1**	0.54	0.06	0.12	0.07	0.15	0.03
	CMS2	1.24		0.18		0.82	
	CMS3	0.39		0.14		0.3	
	CMS4	1.62		0.25		1.24	
Number of Mutations	**<60**	0.21	0.0006	0.13	0.02	0.08	0.01
	[60-120)	0.87		0.15		0.78	
	[120-150)	1.04		0.15		0.74	
	>=150	5.48		2.1		1.76	
	Spearman	0.42	0.006	0.3	0.06	0.26	0.09
Proportion of stroma	**Low**	0.87	0.07	0.15	0.06	0.4	0.02
	High	1.69		0.22		1.33	
	Spearman	0.25	0.02	0.24	0.02	0.26	0.01
Progression	**No**	1.02	0.73	0.2	0.48	0.59	0.67
	Yes	0.78		0.13		0.43	

* Median percentage of altered genome. Except for rows with Spearman in which r is shown
P values are unadjusted for multiple comparisons.

Table 2: Summary of minimal recurrent regions (MRR) with more than 50% of the samples altered of the CLX dataset

Chromosome	Cytoband	Start	End	Number of MRR contained	% median samples altered in CLX	Number of MRR contained in TCGA	% samples altered in TCGA	Genes included in the region
1	1p31.1	72771355	72779078	5	52	4	54	
1	1p13.3	110233315	110234397	3	51	3	63	GSTM1
1	1q21.3	152555783	152586104	10	51	10	65	
3	3q29	192880714	192882890	5	56	6	74	
4	4p16.1	10214160	10234595	16	63	27	73	
5	5p11	46271918	46273489	2	52	3	55	
5	5q11.2	57326015	57333502	16	59	22	52	
8	8p11.22	39246663	39386952	22	64	25	47	ADAM5, ADAM3A
8	8q23.2	112295069	112295247	2	56	3	67	
12	12p13.31	9637897	9690962	5	53	7	61	
13	13q21.31	72479535	72480543	1	57	1	51	
17	17q21.2	39423091	39430518	12	58	15	46	
20	20p13	1561568	1582194	17	54	20	51	SIRRPB1

Table 3: Chromosomal distribution of genes related to CNA and validation in TCGA data

Chr	Number of genes	Genes in CNA region	Significant Genes*	Validation in TCGA data	Validation in MSS and stage II in TCGA data
1	1542	44 (2.85 %)	2	1	0
2	1013	23 (2.27 %)	1	0	0
3	847	49 (5.8 %)	3	0	0
4	606	20 (3.3 %)	1	0	0
5	680	64 (9.4 %)	12	0	0
6	812	105 (13 %)	30	1	0
7	733	150 (20.5 %)	18	0	10
8	528	518 (98.1 %)	282	220	134
9	599	27 (4.5 %)	4	3	1
10	579	35 (6.04 %)	3	0	0
11	896	36 (4.02 %)	1	0	0
12	779	39 (5 %)	17	5	5
13	268	268 (100 %)	115	91	74
14	496	17 (3.43 %)	0	0	0
15	487	17 (3.49 %)	1	0	0
16	679	32 (4.71 %)	2	1	0
17	921	142 (15.42 %)	36	5	1
18	210	97 (46.19 %)	35	5	2
19	1033	30 (2.9 %)	0	0	0
20	416	412 (99.04 %)	221	164	135
21	156	17 (10.9 %)	0	0	0
22	374	26 (6.9 %)	1	0	1
Total	**14654**	**2168 (14.8 %)**	**785**	**488**	**363**

* FDR was used to identify significant associations