Pervasive Chromosomal Instability and Karyotype Order During Tumour Evolution

Authors:

Thomas BK Watkins*[1], Emilia L Lim*[1,2], Marina Petkovic[†3], Sergi Elizalde[†4], Nicolai J Birkbak[1,5,6], Gareth A Wilson[1], David A Moore[2,7], Eva Grönroos[1], Andrew Rowan[1], Sally M Dewhurst[8], Jonas Demeulemeester[9,10], Stefan C Dentro[9,11,12], Stuart Horswell[13], Lewis Au[1,14], Kerstin Haase[9], Mickael Escudero[13], Rachel Rosenthal[1,2,15], Maise Al Bakir[1], Hang Xu[16], Kevin Litchfield[1], Wei Ting Lu[1], Thanos P Mourikis[17], Michelle Dietzen[17], Lavinia Spain[1,14], George D Cresswell[18], Dhruva Biswas[1,15], Philippe Lamy[5], Iver Nordentoft[5], Katja Harbst[19,20], Francesc Castro-Giner[21,22], Lucy R Yates[23,24], Franco Caramia[25], Fanny Jaulin[26], Cécile Vicier[27], Ian PM Tomlinson[28], Priscilla K Brastianos[29], Raymond J Cho[30], Boris C Bastian[30], Lars Dyrskjøt[5], Göran B Jönsson[19,20], Peter Savas[25,31], Sherene Loi[25,31], Peter J Campbell[23], Fabrice Andre[32], Nicholas M Luscombe[18,33,34], Neeltje Steeghs[35], Vivianne CG Tjan-Heijnen[36], Zoltan Szallasi[37,38,39], Samra Turajlic[1,14], Mariam Jamal-Hanjani[1,40], Peter Van Loo[9], Samuel F Bakhoum[41,42], Roland F Schwarz[3,43,44 #], Nicholas McGranahan[17 #], Charles Swanton[1,2,40 #]

* Joint first authors with equal contribution

† Joint second authors with equal contribution

Joint corresponding authors

1 Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, 1 Midland Rd, London, NW1 1AT, UK

2 Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, Paul O'Gorman Building, 72 Huntley Street, London, WC1E 6BT, UK

3 Berlin Institute for Medical Systems Biology, Max Delbrueck Center for Molecular Medicine, Berlin, Germany

4 Department of Mathematics, Dartmouth College, Hanover, New Hampshire, USA

5 Department of Molecular Medicine (MOMA), Aarhus University Hospital, Aarhus, Denmark

6 Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, Denmark

7 Department of Cellular Pathology, University College London Hospitals, London, UK

8 Laboratory for Cell Biology and Genetics, Rockefeller University, New York, NY, USA

9 Cancer Genomics Laboratory, The Francis Crick Institute, 1 Midland Rd, London, NW1 1AT, UK
10 Department of Human Genetics, University of Leuven, Herestraat 49, 3001 Leuven, Belgium
11 Oxford Big Data Institute, University of Oxford, Oxford, UK
12 Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton, UK
13 Department of Bioinformatics and Biostatistics, The Francis Crick Institute, London NW1 1AT, UK
14 Renal and Skin Units, the Royal Marsden Hospital NHS Foundation Trust, London SW3 6JJ, UK
15 Bill Lyons Informatics Centre, University College London Cancer Institute, Paul O'Gorman Building, 72 Huntley Street, London, WC1E 6BT, UK
16 Stanford Cancer Institute, Stanford, Palo Alto, CA 94304
17 Cancer Genome Evolution Research Group, University College London Cancer Institute, University College London, London, UK
18 Bioinformatics and Computational Biology Laboratory, The Francis Crick Institute, London, UK
19 Division of Oncology and Pathology, Department of Clinical Sciences Lund, Faculty of Medicine, Lund University, Scheelegatan 2, Medicon Village, 22185, Lund, Sweden
20 Lund University Cancer Centre, Lund University, Lund, Sweden
21 Department of Biomedicine, Cancer Metastasis Laboratory, University of Basel and University Hospital Basel, CH-4058, Basel, Switzerland
22 Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland.
23 Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK
24 Department of Clinical Oncology, Guy's and St Thomas' NHS Foundation Trust, London, SE1 9RT, UK
25 Division of Research, Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Victoria, Australia
26 INSERM U1279, Gustave Roussy, 114 rue Edouard Vaillant, 94805, Villejuif, France.
27 Department of Medical Oncology, Institut Paoli-Calmettes, Aix-Marseille University, Marseille, France.
28 Edinburgh Cancer Research Centre, IGMM, University of Edinburgh, Crewe Road South, Edinburgh EH4 2XU
29 Stephen E. and Catherine Pappas Center for Neuro-Oncology, Divisions of Hematology/Oncology and Neuro-Oncology, Departments of Medicine and Neurology, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, MA, 02114, USA.
30 Department of Dermatology, University of California, San Francisco, San Francisco, CA, USA
31 Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia
32 Gustave Roussy Department of Medical Oncology, Faculté de Medicine Paris-Sud XI, Université Paris-Saclay, Villejuif, France
33 UCL Genetics Institute, Department of Genetics, Evolution & Environment, University College London, UK
34 Okinawa Institute of Science & Technology, Okinawa, Japan
35 Department of Medical Oncology, School of GROW, Maastricht University Medical Center, Maastricht, the Netherlands

36 Department of Medical Oncology, Netherlands Cancer Institute, the Netherlands.
37 Danish Cancer Society Research Center, Copenhagen, Denmark
38 Computational Health Informatics Program, Boston Children's Hospital, USA
39 2nd Department of Pathology, SE-NAP Brain Metastasis Research Group, Semmelweis University, Budapest, Hungary
40 Department of Medical Oncology, University College London Hospitals, London NW1 2BU, UK
41 Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
42 Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
43 German Cancer Consortium (DKTK), partner site Berlin
44 German Cancer Research Center (DKFZ), Heidelberg

Abstract

Cancer chromosomal instability (CIN) results from dynamic changes to chromosome number and structure. The resulting diversity in somatic copy number alterations (SCNA) may provide the variation necessary for cancer evolution. Multi-sample phasing and SCNA analysis of 1421 samples from 394 tumours across 24 cancer types revealed ongoing CIN resulting in pervasive SCNA heterogeneity. Parallel evolutionary events, causing disruption to the same genes, such as *BCL9*, *ARNT*/*HIF1B*, *TERT* and *MYC*, within separate subclones were present in 35% of tumours. Most recurrent losses occurred prior to whole genome doubling (WGD), a clonal event in 48% of tumours. However, loss of heterozygosity at the human leukocyte antigen locus and loss of 8p to a single haploid copy recurred at significant subclonal frequencies, even in WGD tumours, likely reflecting ongoing karyotype remodeling. Focal amplifications affecting 1q21 (*BCL9*, *ARNT*), 5p15.33 (*TERT*), 11q13.3 (*CCND1*), 19q12 (*CCNE1*) and 8q24.1 (*MYC*) were frequently subclonal and exhibited an illusion of clonality within single samples. Analysis of an independent series of 1024 metastatic samples revealed enrichment for 14 focal SCNAs in metastatic samples, including late gains of 8q24.1 (*MYC*) in clear cell renal carcinoma and 11q13.3 (*CCND1*) in HER2-positive breast cancer. CIN may enable ongoing selection of SCNAs, manifested as ordered events, often occurring in parallel, throughout tumour evolution.

116

117

118

119 ***Introduction***

120 Chromosomal instability (CIN) results from the occurrence and tolerance of chromosome
121 segregation errors during cell division. CIN has been linked to poor prognosis[1-4] and leads to
122 somatic copy number alterations (SCNAs) which may act as a substrate for selection[5-7].

123

124 However, the prevalence of ongoing CIN later in tumour evolution[8] and the temporal order of
125 clonal and subclonal SCNAs in relation to whole genome doubling (WGD) events and
126 metastatic dissemination remains unclear.

127

128 ***Ongoing CIN and SCNA heterogeneity occur across cancer types***

129

130 We applied a multi-sample phasing SCNA analysis method (Figure S1,S2A,B, Methods 2.2-6)
131 to 1421 cancer samples from 394 patients across 24 cancer subtypes (range 2-16, median 3
132 samples/tumour Figure S3A,B, Table S1), to obtain SCNA heterogeneity at haplotype
133 resolution. We used MEDICC[9] to estimate copy number states of the most recent common
134 ancestor (MRCA) for each tumour - reflecting SCNAs acquired prior to subclonal
135 diversification. 1111 / 1421 samples were from treatment naive primary tumours, 51 were
136 from post-treatment primary tumours, 7 samples were obtained at local relapse, and 252
137 were of metastatic origin. In each case, there were at least two samples per tumour with 126
138 tumours having temporally separated samples.

139

140 To explore CIN during cancer evolution, we quantified the total proportion of the genome
141 affected by SCNAs and the proportion of clonal, early, SCNAs compared to subclonal, late,
142 SCNAs (Figure 1A-D). Clonal SCNAs were identified in every tumour (Figure 1C). 99% of
143 tumours (389/394) harboured at least one subclonal SCNA (Figure 1B). A median of 24% of
144 the genome was subject to clonal SCNAs and 17% subclonal SCNAs. 43% of tumours exhibited
145 >20% of the genome subject to subclonal SCNAs, highlighting that ongoing CIN is pervasive.
146 However, this is likely an underestimate of CIN as only a minority of each tumour is sequenced.

Consistent with this, we observed a significant correlation between the number of samples per tumour and SCNA heterogeneity (Figure S4). Moreover, triple-negative breast cancer (BRCA TN), esophageal adenocarcinoma (ESCA) and, clear cell renal cell carcinoma (KIRC) demonstrated a significant association between median purity (Figure 1E) and the proportion of the genome affected by subclonal SCNAs (Figure S5), indicating that purity may impede estimation of SCNA clonality.

The timing of SCNAs varied across cancers (Figure 1A-B, Figure S6). Despite a comparable total SCNA burden between lung adenocarcinoma (LUAD) and HER2-positive breast cancer (BRCA HER2+) (57% vs. 60%, P=0.81, ES=0.05), LUAD exhibited a larger proportion of the genome subject to clonal SCNA, whilst BRCA-HER2+ harboured a higher frequency of subclonal SCNA (LUAD: 28% vs BRCA HER2+: 44%, $P=8.1\times10^{-3}$, ES=0.59, analysis also controlled for sample number, see Figure S4B).

Consistent with increased proliferation in CIN tumours, total, clonal and subclonal SCNA burden correlated with increased cell cycle gene expression in 58 NSCLCs with RNA-sequencing and with increased mitotic index score in 84 NSCLCs with digitised diagnostic slides (Figure S7,S8 Methods 4.3-4, Table S2). Furthermore, in these 84 tumours, estimates of tumour size derived from diagnostic PET–CT scans were found to correlate with total and subclonal SCNA burden, however these associations did not remain significant when controlling sample number (Figure S9). Finally, anisonucleosis, a measure of variation in nuclear size (Methods 4.4), prognostic in NSCLC[10, 11], was associated with increased total and clonal SCNA, but not subclonal SCNAs (Figure S10, Table S2).

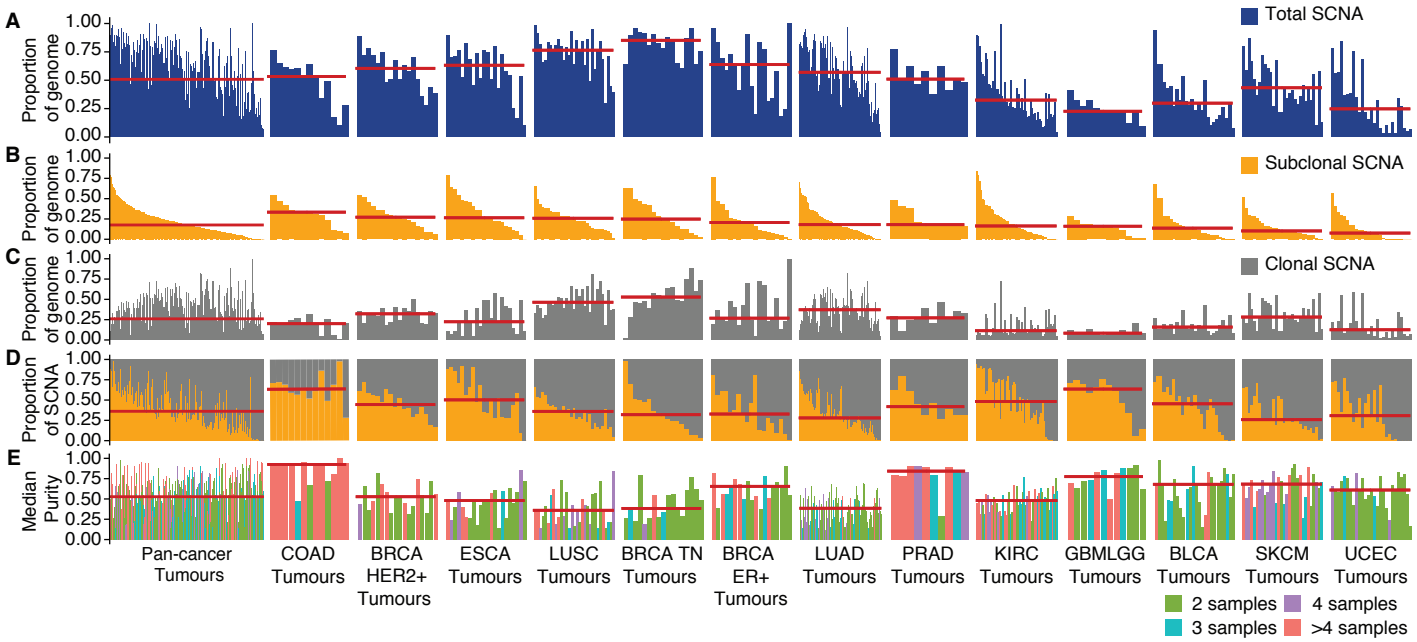
A
B
C
D
E
Proportion of genome
Proportion of SCNA
Median Purity
1.00
0.75
0.50
0.25
0.00
Total SCNA
Subclonal SCNA
Clonal SCNA
Pan-cancer Tumours
COAD Tumours
BRCA HER2+ Tumours
ESCA Tumours
LUSC Tumours
BRCA TN Tumours
BRCA ER+ Tumours
LUAD Tumours
PRAD Tumours
KIRC Tumours
GBMLGG Tumours
BLCA Tumours
SKCM Tumours
UCEC Tumours
2 samples
3 samples
4 samples
>4 samples

Figure 1 - **Overview of somatic copy number heterogeneity across cancer types**
A) For each tumour, the proportion of the genome that is affected by SCNA (both clonal and subclonal) is indicated. Cancer types examined with tumour samples from 10 or more patients included: colorectal adenocarcinoma (COAD, n=13), HER2+ breast cancer (BRCA HER2+, n=18), esophageal adenocarcinoma (ESCA, n=22), lung squamous cell carcinoma (LUSC, n=31), triple-negative breast cancer (BRCA TN, n=17), ER+ breast cancer (BRCA ER+, n=19), lung adenocarcinoma (LUAD, n=84), prostate adenocarcinoma (PRAD, n=10), clear cell renal cell carcinoma (KIRC, n=54), glioma (GBMLGG, n=12), bladder urothelial carcinoma (BLCA, n=26), melanoma (SKCM, n=30), and endometrial carcinoma (UCEC, n=27). Cancer types and tumours are ordered by the median percentage of the genome affected by subclonal SCNA - this order is maintained throughout the figure. Red lines indicate the median of the distribution. B-C) Barplots indicating the percentage of the genome affected by subclonal (B) and clonal (C) SCNA. D) The proportion of SCNA that are subclonal and clonal is displayed. Red line indicates median proportion of SCNA that are subclonal. E) The median purity and number of samples from each tumour.

55% of tumours exhibited whole genome doubling (WGD) (Methods 2.7), a clonal event in 87% of cases (Figure S11A). WGD was associated with an increased burden of clonal and subclonal gains and losses compared to non-WGD tumours (clonal $P=1.36\times10^{-34}$,ES=1.15; Subclonal $P=4.67\times10^{-9}$, ES=0.6, Figure S11B, Methods 2.3). Through multi-sample phasing we investigated mirrored subclonal allelic imbalance[3], resulting from SCNAs disrupting the same genomic region deriving from distinct haplotypes within separate tumour subclones (Methods 2.2-6). WGD tumours were enriched for mirrored subclonal allelic imbalance events compared to non-WGD tumours (Methods 2.2-2.6, $P=4.23\times10^{-7}$, ES=0.6, Figure S11C). In tumours with subclonal WGD, we observed a higher frequency of SCNAs in subclones affected by WGD compared to non-WGD sister clones ($P=9.5\times10^{-3}$, ES=0.59, paired t-test, Figure S11D), accounting for germline and prior somatic alterations as confounding variables.

Evolution of the SCNA landscape

To address whether the SCNA landscape is shaped by neutral evolution or selection, we considered whether the propensity for each chromosome arm to be gained or lost during tumour evolution was related to the density of tumour suppressor genes (TSGs) and oncogenes (OGs) encoded on each chromosome arm, as captured by the OG-TSG score[5]. Consistent with ongoing selection on cellular karyotypes, the OG-TSG score significantly correlated with the burden of arm-level alterations in the MRCA (Figure 2A) as well as subclonal arm-level alterations (Figure 2B and Figure S12A-C). No relationship between average clonal or subclonal chromosome copy number change and chromosome size was observed, suggesting SCNA detection is unlikely to contribute to this relationship (Figure S12D-G).

To understand subclonal SCNA dynamics within each tumour, we adapted our previous model that predicts population karyotypes over time [12, 13]. We used arm-level copy number profiles from each tumour's MRCAs as the starting point and compared how different iterations of the model predict the observed subclonal tumour karyotypes (Figure 2C, S13A,B, Methods 3.1-2). We compared three conditions; first, where karyotypes with higher oncogenic propensity or tumour suppressive propensity were favoured, or unfavoured, respectively, using the relative OG-TSG scores[5] (weighted model); second, where chromosome arms were treated equally (neutral model); and, third where OG-TSG scores were randomly permuted (scrambled model)) to achieve the same complexity as the weighted model. On average, the weighted model predicted the trajectory of subclonal SCNA more accurately, outperforming the two other models, as evidenced by significantly reduced deviance scores (Figure 2C,D, Figure S13C-G) irrespective of the rate of chromosome missegregation or the number of cell divisions (generations)(Figure S14).

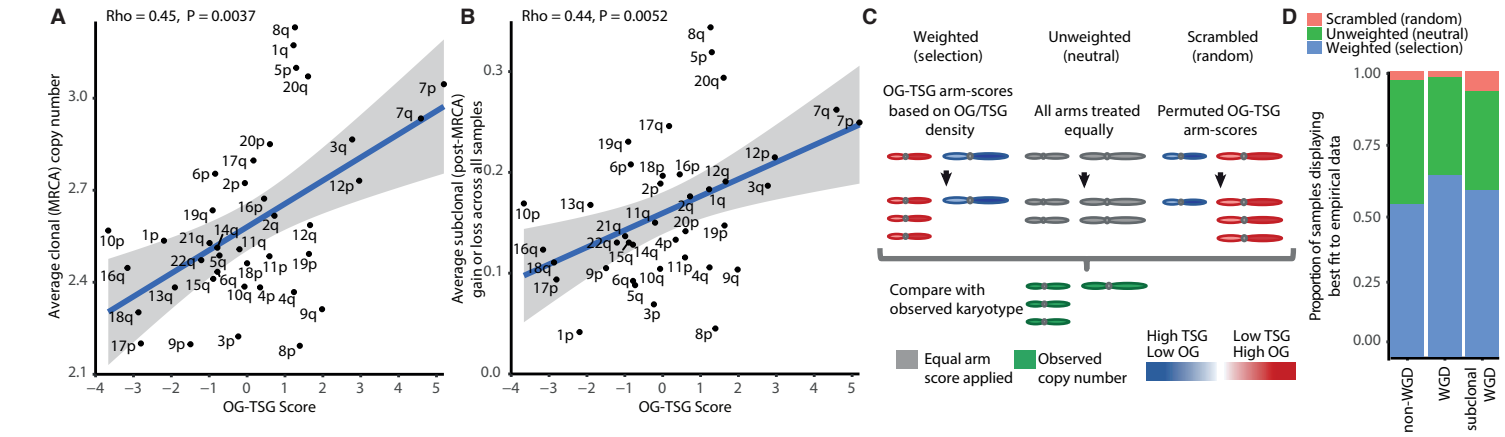
A
Rho = 0.45, P = 0.0037
Average clonal (MRCA) copy number
OG-TSG Score
B
Rho = 0.44, P = 0.0052
Average subclonal (post-MRCA) gain or loss across all samples
OG-TSG Score
C
Weighted (selection)
Unweighted (neutral)
Scrambled (random)
OG-TSG arm-scores based on OG/TSG density
All arms treated equally
Permuted OG-TSG arm-scores
Compare with observed karyotype
Equal arm score applied
Observed copy number
High TSG Low OG
Low TSG High OG
D
Scrambled (random)
Unweighted (neutral)
Weighted (selection)
Proportion of samples displaying best fit to empirical data
non-WGD
WGD
subclonal WGD

Figure 2 - Selection shapes the SCNA landscape

A) Scatter plot showing a positive correlation between average clonal copy number present in the MRCA and the OG-TSG score in n=394 tumours. The grey shaded area represents the 95% confidence interval. Rho and p are from a Spearman correlation test. B) Scatter plot showing a positive correlation between OG-TSG score and average change in SCNA (gain or loss) from MRCA in n=394 tumours. The grey shaded area represents the 95% confidence interval. Rho and P are from a Spearman correlation test. C) Schematic showing the three conditions under which karyotype evolution was modelled: chromosome arms incorporating OG-TSG scores (weighted model); chromosome arms treated equally (neutral model); OG-TSG scores randomly permuted (scrambled model). D) For each context (non-WGD n=194, WGD n=171, and subclonal WGD n=29), the percentage of tumours in which each model condition best recapitulates the empirically observed data is displayed in the bar chart.

Collectively, these data suggest that CIN enables ongoing selection driven by relative dosage imbalance of OGs and TSGs and that WGD may support ongoing genome remodeling later during tumour evolution, permitting further selection. However, the observed pattern of SCNA acquisition in 41% of our cohort in which the neutral or scrambled models outperform the weighted model might reflect neutral karyotype evolution or the need for cancer-type specific chromosome arm weightings [14, 15]. Notably, we see more evidence for subclonal selection in WGD tumours which may be consistent with WGD being a transformative event during subclonal evolution (Figure 2D, S13F,G)[12, 13].

Clonal SCNA recur across cancer types and losses are predominantly early

To decipher SCNA timing during evolution, we used GISTIC2.0 to identify recurrent SCNAs in at least two cancer types (Methods 2.12-14, Figure S15A-M, Table S3). We designed these as consensus peak regions and assigned each into distinct evolutionary timing categories: early, intermediate, or late (Figure 3A,B, Methods 2.15). SCNAs overlapping early peak regions may be implicated in tumourigenesis or result from specific constraints to tumourigenesis. SCNAs overlapping intermediate or late peak regions may be involved in tumour maintenance and progression. Recurrent clonal and subclonal arm-level gain or loss SCNAs for each cancer type were identified using permutations (Methods 2.8-9, Table S4).

We observed differences in evolutionary timing between peak regions associated with gains (gain-peaks) and those with losses (loss-peaks). Loss-peaks were significantly more likely to be early compared to gain-peaks ($P=6.8\times10^{-8}$, ES=0.57, Figure S16). Similarly, a higher proportion of recurrent arm-level losses were clonal compared to arm-gains ($P=2.8 \times 10^{-9}$, ES=0.77, Figure

S17). Gain peak regions were enriched for known OGs, while loss peak regions were enriched for known TSGs (Figure S18A). Early loss peak regions were also enriched for chromosomal fragile sites (Figure S18B), suggesting some may lack functional significance.

Clonal SCNA frequencies affecting early peak regions exceeded the frequency of clonal somatic driver point mutations and indels (Figure S19). The loss-peak 17p13.3−q11.2, encompassing *TP53*, was classified as early in 9/13 cancer types and only classified as late in KIRC (74% subclonal). In three cancer types (BRCA HER2+, LUSC and BRCA TN) >90% of cases exhibited clonal LOH at 17p13.1, suggesting loss is required for tumourigenesis. Across cancers, *TP53* LOH, was clonal in 90% of WGD cases in which it was observed, possibly permitting tolerance for WGD[16]. In KIRC, only 3p26.3-p12.1, encompassing *VHL*, was early (clonal LOH in 98% of KIRCs) (Figure S15H). Other high frequency clonal peaks within cancer types included, the gain-peak 17q12−q21.2, encompassing *ERBB2* in BRCA HER2+ (61% frequency), 3p LOH in LUSC (97% frequency), and gain of 7p11.2, encompassing *EGFR,* in LUAD (64% frequency).

We reasoned that loss occurring prior to WGD must lead to LOH with complete loss of the minor allele. Conversely, single losses occurring after WGD will not lead to LOH. On average, across the cohort, 92% of clonal losses overlapping early loss-peaks involved LOH, suggesting recurrent clonal loss events usually occur prior to WGD.

The timing of other peak-regions were promiscuous between cancer types (Figure 3B). For example, the loss-peak 4q35.2, encompassing *FAT1*, was early in BRCA TN (89% prevalence, 80% clonal), intermediate in BRCA ER+ (58% prevalence, 64% clonal) and late in BRCA HER2+ (61% prevalence, 73% subclonal) (Figure 3B).

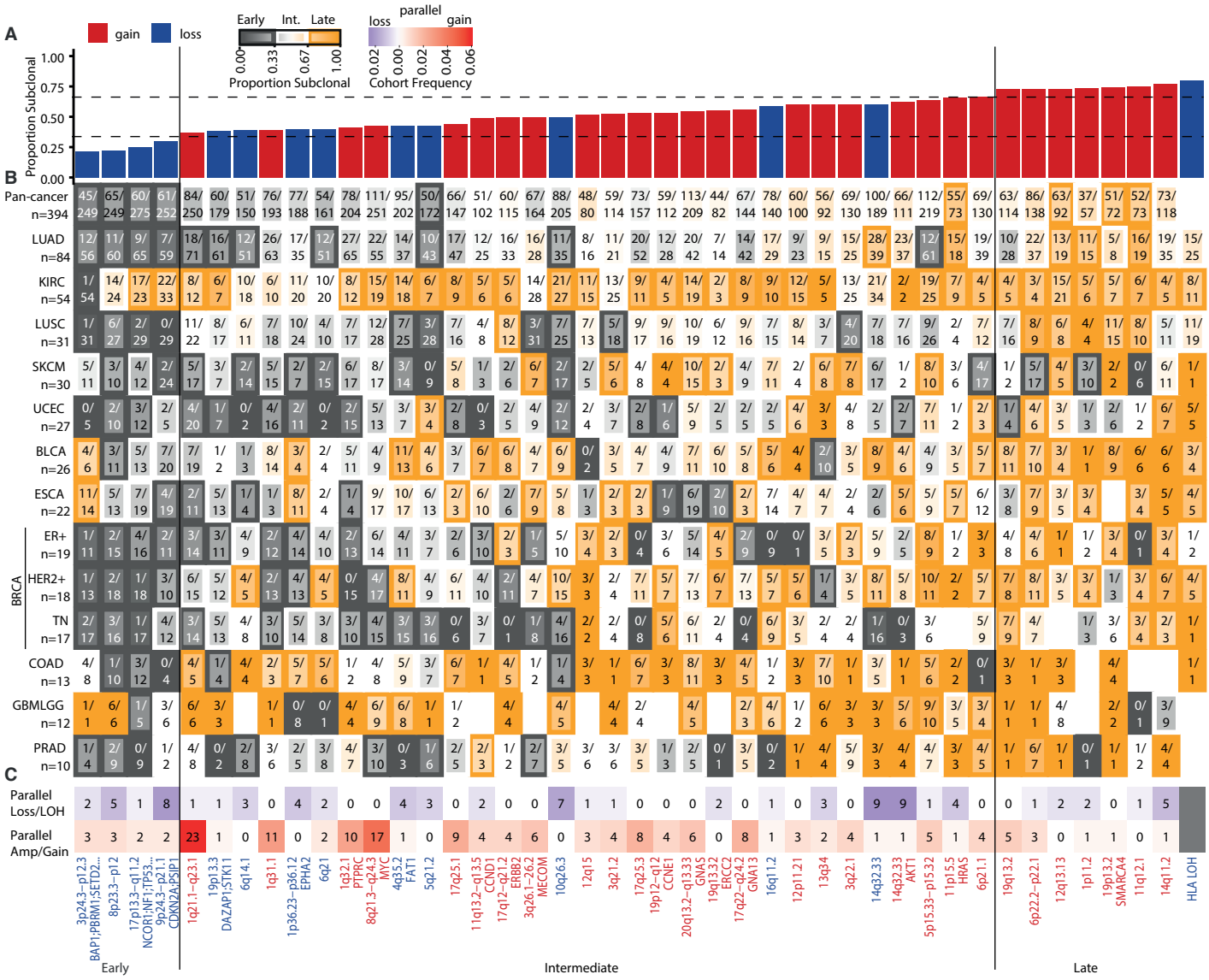
A
gain
loss
Early
Int.
Late
0.00
0.33
0.67
1.00
Proportion Subclonal
loss
parallel
gain
0.02
0.00
0.02
0.04
0.06
Cohort Frequency
Proportion Subclonal
1.00
0.75
0.50
0.25
0.00
B
Pan-cancer n=394
LUAD n=84
KIRC n=54
LUSC n=31
SKCM n=30
UCEC n=27
BLCA n=26
ESCA n=22
BRCA
ER+ n=19
HER2+ n=18
TN n=17
COAD n=13
GBMLGG n=12
PRAD n=10
C
Parallel Loss/LOH
Parallel Amp/Gain
3p24.3–p12.3 BAP1;PBRM1;SETD2...
8p23.3–p12
17p13.3–q11.2 NCOR1;NF1;TP53...
9p24.3–p21.1 CDKN2A;PSIP1
1q21.1–q23.1
19p13.3 DAZAP1;STK11
6q14.1
1q31.1
1p36.23–p36.12 EPHA2
6q21
1q32.1 PTPRC
8q21.3–q24.3 MYC
4q35.2 FAT1
5q21.2
17q25.1
11q13.2–q13.5 CCND1
17q12–q21.2 ERBB2
3q26.1–26.2 MECOM
10q26.3
12q15
3q21.2
17q25.3
19p12–q12 CCNE1
20q13.2–q13.33 GNAS
19q13.32 ERCC2
17q22–q24.2 GNA13
16q11.2
12p11.21
13q34
3q22.1
14q32.33 AKT1
5p15.33–p15.32
11p15.5 HRAS
6p21.1
19q13.2
6p22.2–p22.1
12q13.13
1p11.2
19p13.2 SMARCA4
11q12.1
14q11.2
HLA LOH
Early
Intermediate
Late

Figure 3 – Timing, Recurrence and Parallel Evolution of Subclonal SCNAs

A) Barplot of consensus peaks (Methods 2.13) ordered by median percentage of subclonal occurrence across cancer types. Bars representing gain-peaks are coloured in red and loss-peaks are coloured in blue. Vertical lines indicate separation of consensus peaks into pan-cancer categories of early, intermediate and late, according to tertiles of median proportion of SCNA that is subclonal (horizontal dashed lines). B) Heatmap of the percentage subclonal occurrence of all consensus peaks in each cancer type. Numerator within each cell indicates, in that cancer type, the total number of subclonal occurrences of that peak region and the denominator indicates the total number of both clonal and subclonal occurrences of that consensus peak in that cancer type. Shading of each cell in the heatmap indicates the percentage subclonal occurrence of a consensus peak within a cancer type with orange indicating higher subclonality and grey indicating higher clonality. The border of each cell indicates the classification of that consensus peak in a cancer type as either early (thick dark grey border), intermediate (no border) or late (thick dark orange border). C) Heatmap showing the number of instances of parallel evolution of loss/LOH (blue) and gain/amplification (red) affecting the consensus peak regions. Shading of each cell indicates the number of occurrences of parallel evolution and the number within the cell states the number of such parallel evolutionary events.

Evolution of Subclonal SCNAs

We next addressed which specific subclonal SCNAs are recurrent during tumour evolution.

The highest frequency gain-peaks, including 1q21.1-q21.3 (encoding *BCL9*, *ARNT/HIF1B*) and 5p15.33-p15.32 (encompassing *TERT*), varied in timing across cancers. For example, in LUAD, 80% of 5p15.33-p15.32 gains were clonal, while the majority were subclonal in KIRC (76% subclonal), BRCA ER+ (89% subclonal) and GBMLGG (90% subclonal) (Figure 3B). In LUSC, the timing of *TERT* gains was related to both its focality and amplitude; the majority of low-level gains (>ploidy & <2x ploidy, Methods 2.3) were both clonal and arm-level (8/14) while high-level *TERT* amplifications were often subclonal and focal (10/11). This may reflect augmentation of gene dosage during evolution, with low-level *TERT* gain selected clonally, followed by a high-level amplification selected in a subset of cancer cells later in tumour evolution.

The gain-peak of 19p12−q12 (encompassing *CCNE1*) was late or intermediate in 10/13 cancer types. High-level amplifications of *CCNE1* (>2x ploidy), previously associated with WGD [6, 17], occurred exclusively in WGD tumours. *CCNE1* amplification was subclonal in 9/20 tumours with clonal WGD, suggesting it may be selected both before and after WGD.

Parallel evolution of SCNA events, reflecting distinct subclones within individual tumours converging on a similar evolutionary solution, was observed in 139/394 (35%) tumours within the cohort (Figure 3C,S20). Allele-specific expression tracked parallel evolutionary events originating from distinct haplotypes in samples with matched multi-sample RNA-seq (Figure S21, rho = 0.89, $P=1.75\times10^{-15}$, Spearman correlation).

Consistent with positive selection, parallel gains were significantly more focal than non-parallel subclonal gains ($P=5.9\times10^{-3}$, ES=0.1). The most prominent parallel gains included those overlapping 1q21.3-q44 encompassing BCL9 and *ARNT/HIF1B,* 5p15.33 encompassing *TERT* and 8q24.1 encompassing *MYC* (Figure 3C, Figure S20). The most common parallel loss events included 14q (14q32.33/*ASPP1* and 14q11.2/*NDRG2*), 10q and 9p (Figure S20).

Subclonal LOH after a clonal WGD event occurs through more than one loss event of the same allele after the doubling event (Figure S22). The HLA locus (6p21.3) represented a clear peak of subclonal LOH in WGD samples, affecting 22% of the cohort, indicative of two loss events of the same alleles after genome doubling within the subclone (Figure S23). HLA LOH was prevalent as a subclonal event in KIRC, BRCA, BLCA, NSCLC-other, UCEC and ESCA (Figure S24, Methods 2.11) in addition to NSCLC as previously reported[18]. One exception was SKCM, which is characterised by high mutational burden and benefits from checkpoint inhibitor blockade [19]. SKCM exhibited the lowest frequency of HLA-LOH (0% clonal, 4% subclonal) in the cohort. 6p24.2 also harbours the melanoma metastasis gene *NEDD9*[20], identified as the most prevalent recurrent clonal arm-level gain event in SKCM, which may constrain subsequent HLA loss (Figure S15L).

In a diploid cancer cell, any loss results in LOH. If this cell doubles, the LOH will be maintained, with the remaining allele being duplicated, leading to a total copy number of two. Interestingly, in the case of clonal 8p23.3-p12 loss, we observed a peak region of haploid LOH in WGD tumours, with only a single copy (Figure S22). This haploid, single copy, LOH strongly suggests a loss event of one of the two remaining copies after WGD. Loss of 8p23.3-p12 was most prominent in breast cancer where it has been linked to a chromosome-dosage effect, influencing lipid metabolism and metastatic potential[21].

We next investigated whether we could identify an association between the presence of SCNA overlapping our consensus peaks and overall survival in cancer types matched to those in our multi-sample cohort from the TCGA (Figure S25). Few individual consensus SCNA were associated with survival, suggesting the binary presence or absence of individual SCNAs is rarely prognostic.

Late emerging subclones frequently seed metastases

Next, we explored associations of SCNAs with metastasis. Consistent with previous work[22], we observed a greater percentage of the genome affected by SCNAs in metastatic (n=137 patients) compared with primary tumour samples (n=373 patients) (Figure S26A, $P=1.5\times10^{-5}$, ES=0.4). This remained significant when controlling for cancer type and considering both paired and unpaired primary-metastasis tumour comparisons (Figure S26B) with LOH events displaying the greatest increase from primary to metastasis compared to gain or losses without LOH (Figure S26C). No significant increase in ploidy was observed between primary and metastatic samples in the cohort or in any individual cancer types examined.

Consistent with an evolutionary bottleneck, SCNAs were found to be more frequently clonal in metastatic compared to primary samples (Figure S26D). Indeed, in all 14 cases where we had multi-sample primary tumours and a matched metastatic sample, we identified SCNAs which were fully clonal in the metastasis and present as minor subclones within the primary. In all cancer types with multiple primary-metastatic pairs, in the majority of tumours, most LOH events were found to be shared between primary and metastatic samples, suggesting a relatively late divergence of the metastatic clone assuming LOH events occur at a constant rate throughout cancer evolution (Figure S27, Methods 4.9).

To evaluate the relative importance of specific SCNAs in metastasis, we focused on recurrent SCNAs and performed a combined analysis using both paired analyses on 118 matched primary-metastatic samples and unpaired analyses on 2631 TCGA primary samples, and 1024 Hartwig Medical Foundation metastatic samples[23] in the four cancer types with sufficient primary-metastatic pairs (BRCA HER2+, BRCA ER+, LUAD, KIRC). However, distinct patterns of

SCNA metastatic dissemination were observed in different cancer types. In BRCA ER+, BRCA HER2+ and LUAD, the majority of the arm-events that were enriched in metastasis relative to primary samples were early (Figure S28). Conversely, in KIRC, which had the lowest proportion of shared LOH between primary and metastatic samples, most recurrent arm-events enriched in metastatic samples were classified as intermediate or late events (Figure S28), suggesting these arm-events are associated with metastatic potential in a limited number of cells within the primary tumour.

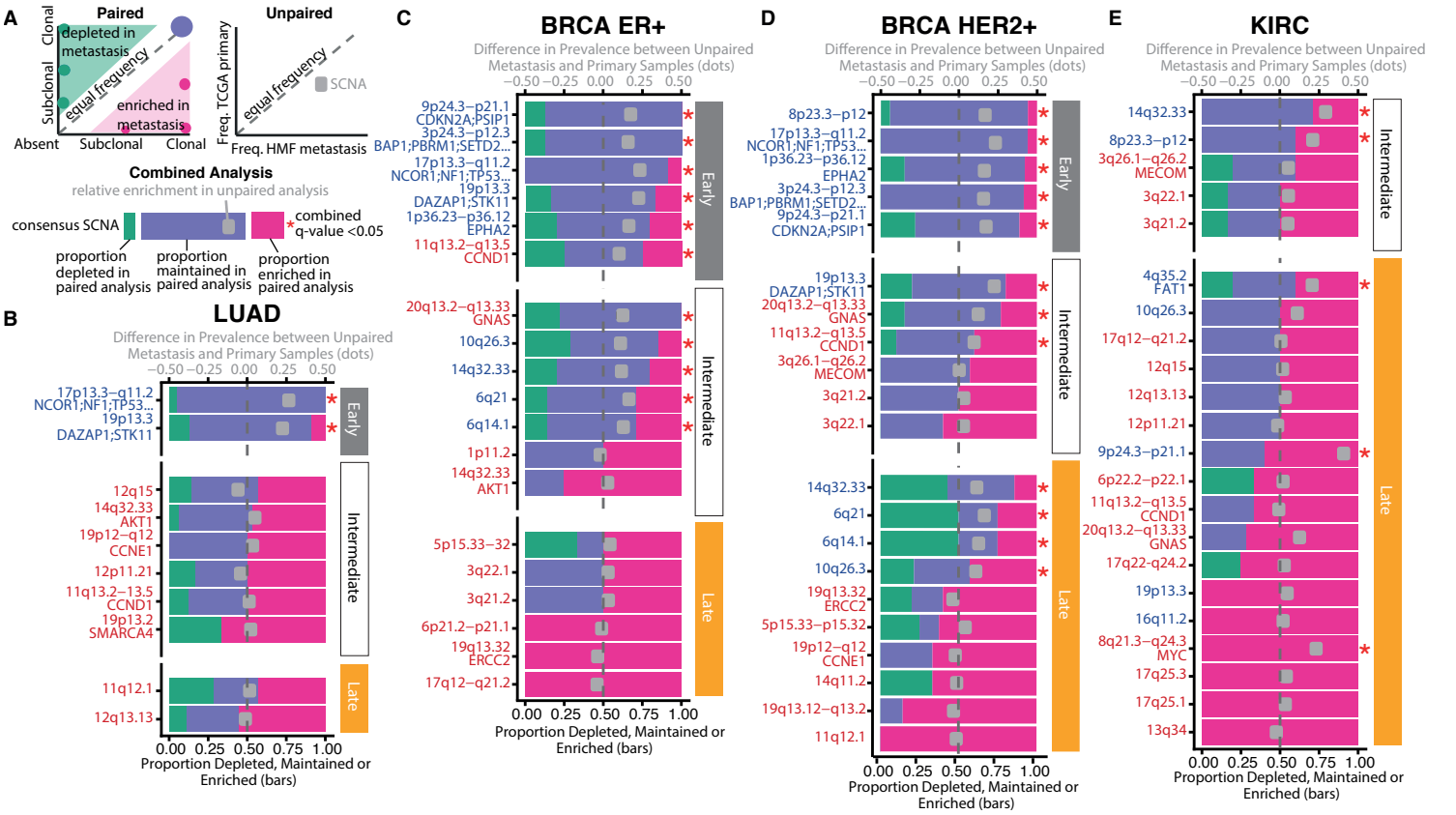
A
Paired
Clonal
Subclonal
depleted in metastasis
equal frequency
enriched in metastasis
Absent
Subclonal
Clonal
Unpaired
Freq. TCGA primary
equal frequency
SCNA
Freq. HMF metastasis
Combined Analysis
relative enrichment in unpaired analysis
consensus SCNA
combined q-value <0.05
proportion depleted in paired analysis
proportion maintained in paired analysis
proportion enriched in paired analysis
B
LUAD
Difference in Prevalence between Unpaired Metastasis and Primary Samples (dots)
−0.50 −0.25 0.00 0.25 0.50
17p13.3–q11.2
NCOR1;NF1;TP53...
19p13.3
DAZAP1;STK11
12q15
14q32.33
AKT1
19p12–q12
CCNE1
12p11.21
11q13.2–13.5
CCND1
19p13.2
SMARCA4
11q12.1
12q13.13
Early
Intermediate
Late
0.00 0.25 0.50 0.75 1.00
Proportion Depleted, Maintained or Enriched (bars)
C
BRCA ER+
Difference in Prevalence between Unpaired Metastasis and Primary Samples (dots)
−0.50 −0.25 0.00 0.25 0.50
9p24.3–p21.1
CDKN2A;PSIP1
3p24.3–p12.3
BAP1;PBRM1;SETD2...
17p13.3–q11.2
NCOR1;NF1;TP53...
19p13.3
DAZAP1;STK11
1p36.23–p36.12
EPHA2
11q13.2–q13.5
CCND1
20q13.2–q13.33
GNAS
10q26.3
14q32.33
6q21
6q14.1
1p11.2
14q32.33
AKT1
5p15.33–32
3q22.1
3q21.2
6p21.2–p21.1
19q13.32
ERCC2
17q12–q21.2
Early
Intermediate
Late
0.00 0.25 0.50 0.75 1.00
Proportion Depleted, Maintained or Enriched (bars)
D
BRCA HER2+
Difference in Prevalence between Unpaired Metastasis and Primary Samples (dots)
−0.50 −0.25 0.00 0.25 0.50
8p23.3–p12
17p13.3–q11.2
NCOR1;NF1;TP53...
1p36.23–p36.12
EPHA2
3p24.3–p12.3
BAP1;PBRM1;SETD2...
9p24.3–p21.1
CDKN2A;PSIP1
19p13.3
DAZAP1;STK11
20q13.2–q13.33
GNAS
11q13.2–q13.5
CCND1
3q26.1–q26.2
MECOM
3q21.2
3q22.1
14q32.33
6q21
6q14.1
10q26.3
19q13.32
ERCC2
5p15.33–p15.32
19p12–q12
CCNE1
14q11.2
19q13.12–q13.2
11q12.1
Early
Intermediate
Late
0.00 0.25 0.50 0.75 1.00
Proportion Depleted, Maintained or Enriched (bars)
E
KIRC
Difference in Prevalence between Unpaired Metastasis and Primary Samples (dots)
−0.50 −0.25 0.00 0.25 0.50
14q32.33
8p23.3–p12
3q26.1–q26.2
MECOM
3q22.1
3q21.2
4q35.2
FAT1
10q26.3
17q12–q21.2
12q15
12q13.13
12p11.21
9p24.3–p21.1
6p22.2–p22.1
11q13.2–q13.5
CCND1
20q13.2–q13.33
GNAS
17q22-q24.2
19p13.3
16q11.2
8q21.3–q24.3
MYC
17q25.3
17q25.1
13q34
Intermediate
Late
0.00 0.25 0.50 0.75 1.00
Proportion Depleted, Maintained or Enriched (bars)

Figure 4: **Analysis of consensus peak regions in metastatic LUAD, BRCA ER+, BRCA HER2+, and KIRC.** A) Schematic describing the paired (left graph), unpaired analysis (right graph), and combined (barplot below) of consensus peak regions. The schematic barplot summarises the left graph for each peak consensus region and indicates the proportion of paired primary-metastasis cases where a SCNA overlapping the consensus peak region was enriched (pink), depleted (green) or maintained (blue) in metastatic samples. These data were assessed using a two-sided binomial test. The grey square in the schematic bar plot indicates the difference between proportions of metastatic (Hartwig Medical Foundation) and primary (TCGA) samples that harbour the event in the unpaired primary-metastasis analysis (two-sided test of equal or given proportions) - a positive number indicates that the event was more prevalent in the metastatic (Hartwig Medical Foundation) samples, while a negative number indicates that the event was more prevalent in the primary (TCGA) samples. The red stars indicate if an event was significantly enriched in metastatic samples as determined by a combined analysis of paired (multi-sample) and unpaired (Hartwig Medical Foundation and TCGA) data using Fisher's method after multiple testing correction using the Benjamini-Hochberg method. The event timing classifications (Early, Intermediate or Late) were determined based on proportion of subclonal occurrence (Methods 2.15). Only losses (blue text) or gains (red text) which are either significant ($q<0.05$) or exhibit $\geq$ 40% enrichment are shown. We restricted our analysis to cancer subtypes with $\geq$ 10 primary-metastasis paired samples (LUAD paired n=28, unpaired n=844 TCGA; 315 Hartwig (B), BRCA ER+ paired n=17, unpaired n=1015 TCGA; 620 Hartwig (C), BRCA HER2+ paired n=13, unpaired n=1015 TCGA; 620 Hartwig (D), and KIRC paired n= 10, unpaired n=772 TCGA; 89 Hartwig (E)).

The early clonal loss-peak at chromosome 1p36.23−p36.12, which encompasses *EPHA2*, and early clonal loss-peak at chromosome 17p13.3-q11.2, encoding *TP53*, were found to be enriched in metastatic samples compared to primary samples in BRCA ER+ and BRCA HER2+ (Figure 4). In LUAD, we observed two early loss consensus peak regions significantly enriched in metastasis (17p13.3-q11.2 [TP53], and 19p13.3 encompassing *STK11*), consistent with these events engendering phenotypes permissive for dissemination early in tumour evolution.

In contrast, other consensus peak-regions enriched in metastases were classified as intermediate or late (Figure 4). Examples include losses of 14q32.33, 6q21 (encompassing *PRDM1*), 6q14.1 and 10q26.3 (encompassing *MGMT*) in BRCA HER2+, and losses of 4q35.2 (encompassing *FAT1*), 9p24.3-p21.1 and gain of 8q21.3-q24.3 in KIRC. Gain of 8q21.3-q24.3, encompassing *MYC*, was highly enriched in our combined analysis as well as exclusively identified in the metastatic samples of our matched primary-metastatic pairs in KIRC. Intriguingly, loss of 9p24.3-p21.1, which encompasses *CDKN2A*, was a late metastasis-associated event in KIRC, while in ER+ and HER2+ BRCA, where loss of 9p24.3-p21.1 was also significantly associated with metastasis, it was predominantly early. Similarly, 11q13.2-q13.5, which encompasses *CCND1*, was an early event in BRCA ER+ and intermediate in BRCA HER2+ and associated with metastasis in both cancer types.

Taken together, these highlight the importance of both early and ongoing SCNA acquisition during tumour evolution and their potential importance during the metastatic transition.

Discussion

Clonal and subclonal SCNAs are pervasive across cancers and show a propensity for order, potentially reflecting the ongoing optimization of fitness landscapes throughout cancer evolution. WGD is a transformative event in tumourigenesis, associated with clonal and subclonal SCNA acquisition. LOH events affecting tumour suppressor genes, including *TP53*, preceded WGD and recurrent gains (eg *CCNE1*) frequently followed WGD and were more likely subclonal.

The subclonal landscape of SCNA is sculpted by both positive and negative selection, as well as neutral evolution. In a minority of tumours, our results are consistent with subclonal karyotypic evolution may predominantly reflecting neutral growth[14, 15]. However, particularly in tumours with WGD, SCNA evolution was better recapitulated using models incorporating both positive and negative selection (Figure 2D). Positive selection was further evidenced by recurrent peaks of subclonal amplifications, enriched for established oncogenes, subclonal losses resulting in LOH, even after genome doubling, and parallel evolution of SCNAs. These data are consistent with previously documented parallel and convergent evolution of SCNAs[3, 24-26] Finally, recurrent focal subclonal SCNAs, encompassing oncogenic events including *CCND1* and *MYC* were enriched at metastatic sites suggesting a potential role in metastasis. Consistent with this, *MYC* was recently described as a SCNA driver of brain metastasis in LUAD[27]. While certain SCNA were enriched in metastasis, in the majority of tumours, most LOH events were shared between primary and metastatic samples, suggesting a late divergence of the metastatic clone, and or negative selection against extensive loss after the emergence of the MRCA[28].

Our work is not without limitations. Detection of recurrent SCNAs is not solely indicative of selection for functional advantage and may result from other processes driving tumour progression such as DNA repair dysfunction or the presence of adjacent fragile sites. Indeed,

552 the higher frequency of recurrent SCNA compared to driver point-mutations need not reflect
553 selection. However, we only found an association of fragile sites with early loss peak regions.
554 Extrachromosomal DNA may also contribute to the subclonal SCNA amplification events
555 observed[29]. The number of tumour samples, their sequencing depths and the lack of an
556 extensive paired primary-metastasis cohort or single cell sequencing analysis influence the
557 degree to which subclonal heterogeneity can be deciphered, suggesting the extent of diversity
558 is underestimated. The lack of uniform clinical data collection and central pathology review
559 prevented detailed analysis of clinically relevant parameters. We are endeavoring to address
560 these deficiencies to time metastatic dissemination events and clonal expansions within
561 TRACERx[3].
562
563 In conclusion, our work highlights the importance of ongoing chromosomal instability during
564 cancer evolution and metastasis. As our understanding of the propensity for different
565 chromosomes to mis-segregate[30] and extent to which chromosomal alterations may be
566 deleterious or advantageous to the cancer cell improves[31], it will be possible to refine the
567 parameters of selection models and improve the ability to detect novel SCNA drivers, which
568 may drive metastatic dissemination and death.
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583

References

1. McGranahan, N., et al., *Cancer chromosomal instability: therapeutic and diagnostic challenges. 'Exploring aneuploidy: the significance of chromosomal imbalance' review series.* EMBO Rep, 2012.
2. Schwarz, R.F., et al., *Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis.* PLoS Med, 2015. **12**(2): p. e1001789.
3. Jamal-Hanjani, M., et al., *Tracking the Evolution of Non-Small-Cell Lung Cancer.* N Engl J Med, 2017. **376**(22): p. 2109-2121.
4. Hieronymus, H., et al., *Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death.* Elife, 2018. **7**.
5. Davoli, T., et al., *Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome.* Cell, 2013. **155**(4): p. 948-62.
6. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-1140.
7. Turajlic, S., et al., *Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal.* Cell, 2018. **173**(3): p. 595-610 e11.
8. Bolhaqueiro, A.C.F., et al., *Ongoing chromosomal instability and karyotype evolution in human colorectal cancer organoids.* Nat Genet, 2019. **51**(5): p. 824-834.
9. Beerenwinkel, N., et al., *Cancer evolution: mathematical models and computational inference.* Syst Biol, 2014.
10. von der Thusen, J.H., et al., *Prognostic significance of predominant histologic pattern and nuclear grade in resected adenocarcinoma of the lung: potential parameters for a grading system.* J Thorac Oncol, 2013. **8**(1): p. 37-44.
11. Kadota, K., et al., *Comprehensive pathological analyses in lung squamous cell carcinoma: single cell invasion, nuclear diameter, and tumor budding are independent prognostic factors for worse outcomes.* J Thorac Oncol, 2014. **9**(8): p. 1126-39.
12. Laughney, A.M., et al., *Dynamics of Tumor Heterogeneity Derived from Clonal Karyotypic Evolution.* Cell Rep, 2015. **12**(5): p. 809-20.
13. Elizalde, S., A.M. Laughney, and S.F. Bakhoum, *A Markov chain for numerical chromosomal instability in clonally expanding populations.* PLoS Comput Biol, 2018. **14**(9): p. e1006447.
14. Sottoriva, A., et al., *A Big Bang model of human colorectal tumor growth.* Nat Genet, 2015. **47**(3): p. 209-16.
15. Williams, M.J., et al., *Identification of neutral tumor evolution across cancer types.* Nat Genet, 2016. **48**(3): p. 238-244.
16. Fujiwara, T., et al., *Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells.* Nature, 2005. **437**(7061): p. 1043-7.
17. Bielski, C.M., et al., *Genome doubling shapes the evolution and prognosis of advanced cancers.* Nat Genet, 2018. **50**(8): p. 1189-1195.
18. McGranahan, N., et al., *Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution.* Cell, 2017.
19. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
20. Kim, M., et al., *Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene.* Cell, 2006. **125**(7): p. 1269-81.

21. Cai, Y., et al., *Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism.* Cancer Cell, 2016. **29**(5): p. 751-766.
22. Bakhoum, S.F., et al., *Chromosomal instability drives metastasis through a cytosolic DNA response.* Nature, 2018. **553**(7689): p. 467-472.
23. Priestley, P., et al., *Pan-cancer whole-genome analyses of metastatic solid tumours.* Nature, 2019. **575**(7781): p. 210-216.
24. Lackner, C., et al., *Convergent Evolution of Copy Number Alterations in Multi-Centric Hepatocellular Carcinoma.* Sci Rep, 2019. **9**(1): p. 4611.
25. Jakubek, Y.A., et al., *Large-scale analysis of acquired chromosomal alterations in non-tumor samples from patients with cancer.* Nat Biotechnol, 2020. **38**(1): p. 90-96.
26. Zaccaria, S.R., B. J., *Characterizing the allele- and haplotype-specific copy number landscape of cancer genomes at single-cell resolution with CHISEL.* bioRxiv https://doi.org/10.1101/837195, 2019.
27. Shih, D.J.H., et al., *Genomic characterization of human brain metastases identifies drivers of metastatic lung adenocarcinoma.* Nat Genet, 2020. **52**(4): p. 371-377.
28. William Cross, M.M., Salpie Nowinski, George Cresswell, Abhirup Banerjee, Marc Williams, Laura Gay, Ann-Marie Baker, Christopher Kimberley, Hayley Davis, Pierre Martinez, Maria Traki, Viola Walther, Kane Smith, Giulio Caravagna, Sasikumar Amarasingam, George Elia, Alison Berner, Ryan Changho Choi, Pradeep Ramagiri, Ritika Chauhan, Nik Matthews, Jamie Murphy, Anthony Antoniou, Susan Clark, Jo-Anne Chin Aleong, Enric Domingo, Inmaculada Spiteri, Stuart AC McDonald, Darryl Shibata, Miangela M Lacle, Lai Mun Wang, Morgan Moorghen, Ian PM Tomlinson, Marco Novelli, Marnix Jansen, Alan Watson, Nicholas A Wright, John Bridgewater, Manuel Rodriguez-Justo, Hemant Kocher, Simon J Leedham, Andrea Sottoriva, Trevor A Graham, *Stabilising selection causes grossly altered but stable karyotypes in metastatic colorectal cancer.* bioRxiv, 2020. **doi: https://doi.org/10.1101/2020.03.26.007138**.
29. Turner, K.M., et al., *Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity.* Nature, 2017. **543**(7643): p. 122-125.
30. Worrall, J.T., et al., *Non-random Mis-segregation of Human Chromosomes.* Cell Rep, 2018. **23**(11): p. 3366-3380.
31. Lopez, S., et al., *Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution.* Nat Genet, 2020. **52**(3): p. 283-293.