Spatial and temporal diversity in genomic instability processes defines lung cancer evolution

Elza C. de Bruin[1]†, Nicholas McGranahan[2,3]†, Richard Mitter[2,]†, Max Salm[2]†, David C. Wedge[4]†, Lucy Yates[4,5,] ‡, Mariam Jamal-Hanjani[1,] ‡, Seema Shafi[1], Nirupa Murugaesu[1], Andrew J. Rowan[2], Eva Grönroos[2], Madiha A. Muhammad[1], Stuart Horswell[2], Marco Gerlinger[2], Ignacio Varela[6], David Jones[4], John Marshall[4], Thierry Voet[4,7], Peter Van Loo[4,7], Doris M Rassl[8], Robert C Rintoul[8], Sam M. Janes[9], Siow-Ming Lee[1,10], Martin Forster[1,10], Tanya Ahmed[10], David Lawrence[10], Mary Falzon[10], Arrigo Capitanio[10], Timothy T. Harkins[11], Clarence C. Lee[11], Warren Tom[11], Enock Teefe[11], Shann-Ching Chen[11], Sharmin Begum[2], Adam Rabinowitz[2], Benjamin Phillimore[2], Bradley Spencer-Dene[2], Gordon Stamp[2], Zoltan Szallasi[12,13], Nik Matthews[2], Aengus Stewart[2], Peter Campbell[4], Charles Swanton[1,2]*

Affiliations:

[1]Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, UK.

[2]Cancer Research UK London Research Institute, UK,

[3] Centre for Mathematics & Physics in the Life Science & Experimental Biology (CoMPLEX), University College London, UK.

[4]Wellcome Trust Sanger Institute, UK.

[5]University of Cambridge, Cambridge, UK.

[6]Instituto de Biomedicina y Biotecnología de Cantabria (CSIC-UC-Sodercan), Departamento de Biología Molecular, Universidad de Cantabria, Santander, Spain.

[7]Department of Human Genetics, University of Leuven, Leuven, Belgium.

[8]Papworth Hospital NHS Foundation Trust, Cambridge UK.

[9]Lungs for Living Research Centre, University College London, UK.

[10]University College London Hospitals, UK.

[11]Thermo Fisher Scientific, CA, USA.

[12]Technical University of Denmark, Denmark.

[13]Children's Hospital Informatics Program, Harvard Medical School, Boston, MA, USA.

† These authors contributed equally

‡ These authors contributed equally

*To whom correspondence should be addressed. E-mail: Charles.swanton@cancer.org.uk

Abstract:

Spatial and temporal dissection of the genomic changes occurring during the evolution of human non-small cell lung cancer (NSCLC) may help elucidate the basis for its dismal prognosis. We sequenced 25 spatially distinct regions from seven operable NSCLCs and found evidence of branched evolution, with driver mutations arising before and after subclonal diversification. There was pronounced intra-tumor heterogeneity in copy number alterations, translocations, and mutations associated with APOBEC cytidine deaminase activity. Despite maintained carcinogen exposure, tumors from smokers showed a relative decrease in smoking-related mutations over time, accompanied by an increase in APOBEC-associated mutations. In tumors from ex-smokers, genome-doubling occurred within a smoking-signature context before subclonal diversification, suggesting that a long period of tumor latency had preceded clinical detection. The regionally separated driver mutations, coupled with the relentless and heterogeneous nature of the genome instability processes are likely to confound treatment success in NSCLC.

Main Text:

Lung cancer is the leading cause of cancer-related mortality (*1, 2*). Understanding the pathogenesis and evolution of lung cancer may lead to greater insight into tumor initiation and maintenance, and guide therapeutic interventions. Previous work characterizing the genome of non-small cell lung cancer (NSCLC) has demonstrated that NSCLC genomes exhibit hundreds of non-silent mutations together with copy number aberrations and genome doublings (*3*-*9*). Although subclonal populations have been identified within single biopsies (*9*), the extent of genomic diversity within primary NSCLCs remains unclear. Moreover, while both exogenous mutational processes, such as smoking (*10*-*12*) and endogenous processes, such as up-regulation of APOBEC cytidine deaminases (*13*-*15*), have been found to contribute to the large mutational burden in NSCLC, the temporal dynamics of these processes and their contribution to driver somatic aberrations over time remains unknown.

To investigate lung cancer evolution, we performed multi-region whole-exome and/or whole-genome sequencing (M-seq WES/WGS) on a total of 25 tumor regions, collected from seven NSCLC patients who underwent surgical resection prior to receiving adjuvant therapy. The major NSCLC histological subtypes, including adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC), were represented (table S1). Sequencing of tumor and normal DNA to mean coverage depths of 107x and 54x for M-WES and M-WGS respectively (table S2), identified 1884 non-silent and 76,129 silent mutations (*16*).

To evaluate the intra-tumor heterogeneity of non-silent mutations, we classified each mutation as ubiquitous (present in all tumor regions) or heterogeneous (present in at least one, but not all regions). Spatial intra-tumor heterogeneity was identified in all seven NSCLCs with a median of 30% heterogeneous mutations (range 4-63%, Fig. 1A and fig. S1). In the adenosquamous tumor L002, heterogeneous mutations separated concordant with LUAD (regions R1-R2) or LUSC (regions R3-R4) histopathologies (Fig. S2). Patients L003 and L008 each presented with two tumors in separate lobes of the lung. Whilst M-seq WES revealed 74% ubiquitous mutations in L008, only a single mutation (EGFR^{L858R}) was detected in both tumors for L003 (Fig. 1A). Given that EGFR^{L858R} is a highly recurrent mutation (*17*) and also that no silent mutations were shared, we conclude these tumors were of independent clonal origin, converging upon identical oncogenic events.

To resolve the extent of genomic diversity in NSCLC and infer the ancestral relationships between tumor regions, we estimated the proportion of tumor cells within each region harboring each mutation (*16, 18*). Almost all ubiquitous mutations (>99%) were classified as fully clonal within each region. Moreover, in most regions, the majority of heterogeneous/branched mutations were clonal and thus present in all cells within the region (Figs. 1B and fig. S3). However, certain regions displayed considerable subclonal diversity. For example, >75% of heterogeneous mutations present in L004 R5 were subclonal and this region consisted of two distinct subclonal populations (Fig 1B). The subclonal structure of each tumor region was then used to construct phylogenetic trees, using both Maximum Parsimony and the Unweighted Pair Group Methods (UPGMA), and also accounting for regional copy number losses resulting in shared truncal mutations becoming private to one region (*16*) (fig. S4). Notably, all seven NSCLCs showed evidence of branched tumor evolution (fig. S5).

We next evaluated the regional heterogeneity of potential NSCLC driver mutations, classified into three categories based on current evidence supporting driver mutation status (*16*). Every tumor showed evidence for ubiquitous as well as heterogeneous driver mutations, many of which were clonally dominant in a subset of tumor regions while entirely absent in others (Fig. 1B, fig. S3 and table S3). Importantly, the probability of missing a category 1 “high confidence” driver gene by analyzing a single region for each tumor was on average 42% (range 0-67%), and 83% (range 67-100%) for all potential driver genes (category 1-3), highlighting the potential limitations of single-biopsy approaches. Nevertheless, category 1-2 driver mutations were significantly more often truncal compared to mutations in non-driver genes in our M-seq analysis (P=0.04). Consistent with these data, in the TCGA cohort previously reported driver genes (*5, 19, 20*) were significantly enriched for clonal mutations (fig. S6, P<0.001). These data indicate that in NSCLC most known driver mutations occur early in tumor evolution.

To determine the intra-tumor heterogeneity of copy number aberrations, we estimated integer DNA copy numbers for each tumor region (*7, 16, 21, 22*). A large fraction of the genome had undergone alterations in all tumors, and genomic profiles were more similar within tumors than between different tumors (fig. S7). To evaluate the spatial heterogeneity of potential tumor driver copy number aberrations, we explored the regional distribution of chromosomal segments identified as recurrently gained or lost in TCGA LUAD or LUSC tumors. Most segments were identified as aberrant in at least one tumor region and many recurrent gains and losses were found to be heterogeneous in at least one tumor (Fig 2A). For example, in L001, a focal EGFR amplification as well as deletions of chromosomal segments harboring CDKN2A and PTEN was observed in all regions, whereas in L008 we observed heterogenous copy number losses involving CDKN2A and PTEN. In support of copy number aberrations occurring later in tumor development, we also identified subclonal copy number aberrations within tumor regions. For instance, over 15% of the genome in region R1 of L008 was subject to subclonal copy number alterations (fig. S8). Consistent with evidence of subclonal copy number aberrations, centromeric FISH analyses confirmed numerical chromosomal diversity within individual tumor regions (fig. S9), suggesting chromosomal instability may provide a substrate for subclonal competition.

The high coverage M-seq WGS (mean 96x) for L002 and L008 enabled us to investigate the regional separation of large-scale genomic events in these samples. For the adenosquamous tumor L002, we identified 30 structural variants, most of which were found either in the LUAD region R1 or the LUSC region R3 but not both (table S4), suggesting they occurred after diversification into two distinct histological subtypes (Fig. 2B). By contrast, for L008, 48 of the 52 identified structural variants were shared between the two tumor regions from different lobes of the lung (Fig 2B). Intriguingly, in L008 "chains" of translocations with highly clustered breakpoints were found between chromosomes 14 and 17 as well as chromosomes 17 and 19 (fig. S10 and table S4), disrupting the *FANCM* and *NF1* tumor suppressor genes. Breakpoint homology profiling suggests involvement of either non-homologous or alternative end-joining (*23, 24*), indicative of double-strand break events. This lesion pattern is consistent with chromoanagenesis (*25*), and indicates a punctuated evolution pattern where multiple oncogenic events may occur simultaneously (*26*).

Four tumors displayed evidence for whole genome-doubling events (*16*). In three tumors (L001, L004, and L008), the genome-doubling event was shared across every tumor region, occurring prior to diversification with the majority of truncal mutations (84-88%) present at ploidy ≥2, indicative of a large mutational burden prior to genome doubling. In one tumor, L002, the majority of heterogeneous mutations were also present at ploidy ≥2, indicative of two independent genome-doubling events, one in the LUAD region and one in the LUSC region (fig. S11). Notably, every truncal driver mutation likely occurred prior to genome doubling.

To further explore the dynamics of the mutational processes shaping lung cancer genomes over time, the spectra of point mutations in each tumor were temporally dissected. Early (truncal) mutations likely reflect processes involved prior to and during tumor initiation and early development, while late (branched) mutations reveal mutational processes shaping the genome during tumor maintenance and progression, including those contributing to intra-tumor heterogeneity. For L002 we analyzed regions R1 and R3 separately allowing comparisons of LUAD and LUSC histologies within the same tumor.

In all tumors we observed statistically significant shifts in the mutation spectra over time (Fig. 3A; *P*<0.05 all cases). Furthermore, every tumor exhibited a statistically significant decrease in the proportion of C>A transversions in late compared to early mutations (Fig. 3A, *P*<0.05), although this was more pronounced in the LUAD cases (mean odds ratio: LUAD 3.13 (range 2.07-5.55), LUSC 1.34 (range 1.21-1.46)). Since C>A transversions are associated with the mutagenic effects of tobacco smoke (*12*), a decrease in the proportion of C>A transversions indicates a relative decrease in the mutational burden attributable to smoking during LUAD development, in both ex-smokers and current-smokers.

To validate these observations in a larger NSCLC cohort, mutations in TCGA LUAD and LUSC samples were temporally dissected (*16*). Consistent with our M-seq analyses, both TCGA LUAD and LUSC smokers and ex-smokers exhibited a decrease in the proportion of C>A transversions in late mutations (Fig. 3B; LUAD current-smokers *P*<0.0001; ex-smokers *P*<0.0001; never-smokers *P*=0.147; LUSC current-smokers *P*=0.003; ex-smokers *P*<0.0001; never-smokers *P*=0.673). Similarly, the least pronounced decrease was observed in LUSC current-smokers; 25% of LUSC displayed no decrease in C>A transversions, compared to less than 10% in LUAD. The mutational footprint of smoking exhibits a strand bias with C>A transversions accumulating preferentially on the transcribed strand (*10*, *12*). Both LUAD and LUSC ex-smokers revealed a statistically significant decrease in strand bias (*10*, *12*) in late compared to early C>A transversions (LUAD *P* = 0.00354; LUSC *P* = 0.046), consistent with smoking having left a footprint on these genomes but no longer being active. Conversely, no statistically significant difference was observed between early and late mutations in current-smokers (LUAD *P* = 0.23; LUSC *P* = 0.22).

In the majority of M-seq tumors, the decreased proportion of C>A mutations was accompanied by an increase in C>T and C>G mutations at TpC sites, indicative of APOBEC cytidine deaminase activity (*13-15*). Mutations consistent with APOBEC-mediated mutagenesis were more pronounced on the branches compared to the trunk in 4/5 LUAD M-seq samples (Fig. 3C). On average ~~19~~31% (~~8~~8-4~~3~~1%) of non-silent branch mutations occurred in an APOBEC mutation context compared to ~~8~~11% (~~5~~7-~~-~~1~~3~~6%) of truncal non-silent mutations. Branched driver mutations *PIK3CA*, *EP300*, *TGFBR1* and *AKAP9* harbored mutations in an APOBEC context indicating a possible functional impact of APOBEC activity upon subclonal expansion. Likewise, TCGA LUAD tumors with detectable APOBEC mutational signatures showed significant enrichment in late compared to early APOBEC mutations (fig. S12, *P* <0.001) and ~~18~~20% of subclonal driver mutations were found to occur in an APOBEC context, compared to ~~9~~11% of clonal driver mutations. However, for TCGA LUSC tumors with detectable APOBEC mutational signatures, temporal dissection of APOBEC mutations did not reveal such a clear trend (fig. S12), indicating potential differences in the temporal dynamics of APOBEC mediated mutagenesis between histological subtypes. In addition to temporal heterogeneity, spatial heterogeneity in both the proportion of APOBEC associated mutations (Fig 3D,E), as well as APOBEC mRNA expression was observed in the M-seq tumors (fig. S13).

To gain a deeper understanding of NSCLC evolution, we focused on the two tumors with high coverage M-seq WGS and temporally placed the genomic instability processes relative to the emergence of the most-recent common ancestor (Fig. 4). In L002, a current-smoker, tobacco carcinogens played a significant role early in tumor development, with C>A transversions representing 39% of truncal mutations (Fig 4A). Early mutations included multiple driver genes, such as *TP53* and *CHD8*. Upon diversification into a LUAD subclone and a LUSC subclone, copy number alterations (fig. S7) and driver mutations were acquired independently in both subclones, such as a stopgain mutation in the tumor suppressor gene *FAT1* on the LUSC branch, and mutations affecting *TGFBR1*, *ZFHX4*, *ARHGAP35* and *PTPRD* in the LUAD region. APOBEC-associated mutations were elevated specifically in the LUAD region, including the driver mutations in *TGFBR1* and *PTPRD*, and the highest *APOBEC3B* mRNA expression was detected in this region (fig. S13).

L008 also revealed truncal C>A transversions and spatial heterogeneity in APOBEC enrichment, with a more pronounced APOBEC signature in the tumor of the middle lobe compared to the upper lobe (Fig 4B). In L008 we gained further temporal resolution by exploring the mutations before and after the truncal genome-doubling event, revealing a tobacco smoke signature of C>A transversions in over 30% of truncal mutations both before and after doubling, and only in 21% and 9% of heterogeneous mutations in the two regions R1 and R3 from separate lobes of the lung. Since L008 ceased smoking over 20 years prior to surgery (table S1), these data suggest the genome-doubling event occurred within a smoking carcinogenic context over 20 years ago. Similarly, the genome-doubling event in ex-smoker L001 also appeared to occur prior to smoking cessation over 20 years prior to surgery (fig S14). These data suggest a prolonged tumor latency period following genome doubling prior to clinical detection in NSCLC.

Through sequencing multiple surgically resected tumor regions we were able to unravel both the extent of genomic heterogeneity and the evolutionary history of seven NSCLCs. Unlike clear cell renal cell carcinoma (ccRCC)(*27, 28*), we find that known driver mutations typically occur early in NSCLC development and the majority of high confidence driver events are fully clonal, conceivably explaining the progression free survival benefits associated with NSCLC oncogenic driver targeting (*29*). However, like ccRCC(*27, 28*), heterogeneous driver mutations and/or recurrent copy number aberrations were present in all tumors and many heterogeneous mutations gave the "illusion of clonality" being present in all cells from certain regions but undetectable within other regions. Notably, although our multi-regional sampling approach allowed us to evaluate spatial heterogeneity, we estimate only a small part of the entire tumor was sampled (on average <5%), suggesting we might be underestimating the full extent of heterogeneity in these tumors.

Conceivably, intra-tumor heterogeneity may compromise the ability of a single biopsy to define all driver events comprehensively for optimal tumor control. For instance, L008 presented with an activating *BRAF* (G469A) mutation (*30*) in all regions and an activating *PIK3CA* (E542K) mutation (*31*) only in region R3. Thus, a biopsy taken from R3 might suggest treatment with an inhibitor of the PI3K/mTOR signaling axis and combination therapy. Conversely, a single biopsy from any other region would suggest treatment with a BRAF inhibitor, for which the tumor cells from R3 might be resistant due to the *PIK3CA* mutation (*32*).

Our study also sheds light on the divergent genomic instability processes involved in NSCLC evolution and their dynamics over time. Evidence for spatial diversity in genomic instability processes suggests that opportunities to exploit such mechanisms therapeutically may be limited in this disease(*33*). In three tumors we detected genome-doubling events occurring prior to subclonal diversification but after acquisition of driver mutations, consistent with findings in colorectal cancer that genome doubling may accelerate cancer genome evolution (*34*). The relationship of chromosomal instability with drug resistance and early tumor recurrence (*35, 36*) suggest targeting truncal driver events may be compromised by the initiation of chromosomal instability later in tumor evolution. These results, coupled with the observation that NSCLC tumors may have prolonged latency periods, support continued efforts to optimize methods for earlier detection

Unexpectedly, after detectable subclonal diversification, we find that even in tumors that are continuously subject to the mutagenic insults of tobacco smoke, an additional genomic instability process, such as APOBEC-mediated mutagenesis, often contributes to tumor progression. A large proportion of subclonal driver mutations were found to occur in an APOBEC context suggesting the differences in mutation spectra over time and space may reflect the activity of the process generating the mutations as well as the selective advantage of the acquired mutations.

The presence of subclonal, regionally separated driver events coupled with the relentless and dynamic nature of genomic instability processes observed in this study, highlight the therapeutic challenges associated with NSCLC. Engaging an adaptable immune system may present a tractable approach to manage the dynamic complexity in NSCLC(*37*). Longitudinal studies will be required to decipher drivers of subclonal expansion, identify the origins of subclones contributing to metastatic recurrence and resolve the evolutionary principles that underpin the dismal outcome associated with this disease.

Figure Legends:

Fig. 1. Intratumor heterogeneity of somatic mutations.

A) Heatmaps show the regional distribution of all non-silent mutations; presence (blue) or absence (grey) of each mutation is indicated for every tumor region. Cartoons depict the location of each tumor. Column next to heatmap shows the intratumor heterogeneity; mutation present in all regions (blue), in more than one, but not all (yellow) or in one region (red). Mutations are ordered on tumor driver category with categories 1-3 indicated in the right column in black, dark grey and light grey respectively (details in table S3). Total number of non-silent mutations (n) is provided for each tumor. In L001, * this mutation is additional to the germ-line MEN1 mutation.

B) 2D-dirichlet plots show the cancer cell fraction of the mutations in all regions of tumors L004; increasing intensity of red indicates the location of a high posterior probability of a cluster. In region R5 the majority of heterogenous mutations are subclonal and a cluster of mutations with a cancer cell fraction below 1 can be observed.

Fig. 2. Intratumor heterogeneity of chromosomal alterations.

A) Distribution of potential tumor driver copy number alterations is indicated for each tumor region. The upper heatmaps show the regional distribution of recurrently amplified (left) or deleted (right) chromosomal segments based on TCGA LUAD data, and the lower heatmaps show the regional distribution of recurrently amplified or deleted chromosomal segments based on TCGA LUSC data. For each region gain (red) and loss (blue) was determined relative to the mean ploidy.

B) Circos plots depicting inter- and intrachromosomal translocations as well as deletions and insertions for regions R1 and R3 for L002 (upper) and L008 (lower); shared events are indicated in blue, events private to region R1 are indicated in red and private to region R3 in green. The outer circle represent the integer copy number data for R1 and the inner circle for R3 for each tumor sample, copy number segments with an integer value greater than mean ploidy are in red and those less than mean ploidy in blue.

Fig. 3. Temporal and spatial dissection of mutation spectra in LUAD and LUSC samples.

A) Pie chart showing the fraction of early mutations (trunk) and late mutations (branch) accounted for by each of the six mutation types in all M-seq samples.

B) Beeswarm plots showing the fraction of early mutations and late mutations accounted for by each of the six mutation types in every TCGA ex-smoker or current-smoker with both early and late mutations. Significance is indicated.

C) Barplot showing APOBEC mutation enrichment odds ratio for early (trunk, blue bars) and late (branch, red bars) mutations for M-seq samples. The APOBEC signature encompasses C>T and C>G mutations in a TpC context (*16*). 95% confidence intervals for Fisher’s exact test are indicated.

D,E) Stacked barplots showing three mutation types (C>A; C>G and C>T) at all sixteen possible trinucleotide contexts for L002 **(D)** and L008 **(E)**. For both samples, trunk mutations and branch mutations from two regions are depicted.

Fig. 4. A model of the evolutionary history of NSCLC.

A, B) Evolutionary histories of L002 (**A**) and L008 (**B**) are depicted. Genomic instability processes defining NSCLC evolution have been placed on their phylogenetic trees. Driver mutations occurring in an APOBEC context are highlighted with a blue dashed box, and those occurring in a smoking context with a grey dashed box. In each case, the timing of genome doubling events is indicated with an arrow.

References and Notes:

1. World Health Organization, http://www.who.int/cancer/en/, (2013).
2. R. Siegel, D. Naishadham, A. Jemal, Cancer statistics, 2013. *CA: a cancer journal for clinicians* **63**, 11-30 (2013).
3. H. Tanaka *et al.*, Lineage-specific dependency of lung adenocarcinomas on the lung development regulator TTF-1. *Cancer research* **67**, 6007-6011 (2007).
4. B. A. Weir *et al.*, Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893-898 (2007).
5. L. Ding *et al.*, Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-1075 (2008).
6. Z. Kan *et al.*, Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869-873 (2010).
7. S. L. Carter *et al.*, Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413-421 (2012).
8. T. I. Zack *et al.*, Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134-1140 (2013).
9. R. Govindan *et al.*, Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121-1134 (2012).
10. E. D. Pleasance *et al.*, A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190 (2010).
11. W. Lee *et al.*, The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).
12. G. P. Pfeifer, P. Hainaut, On the origin of G --> T transversions in lung cancer. *Mutat Res* **526**, 39-43 (2003).
13. S. A. Roberts *et al.*, An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**, 970-976 (2013).

14. M. B. Burns, N. A. Temiz, R. S. Harris, Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat Genet* **45**, 977-983 (2013).
15. L. B. Alexandrov *et al.*, Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).
16. Detailed information on methods is available on Science online.
17. S. V. Sharma, D. W. Bell, J. Settleman, D. A. Haber, Epidermal growth factor receptor mutations in lung cancer. *Nat Rev Cancer* **7**, 169-181 (2007).
18. N. Bolli *et al.*, Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun* **5**, 2997 (2014).
19. M. Imielinski *et al.*, Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-1120 (2012).
20. M. S. Lawrence *et al.*, Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
21. P. Van Loo *et al.*, Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915 (2010).
22. S. Nik-Zainal *et al.*, The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
23. A. Malhotra *et al.*, Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res* **23**, 762-776 (2013).
24. S. F. Bunting, A. Nussenzweig, End-joining, translocations and cancer. *Nat Rev Cancer* **13**, 443-454 (2013).
25. C. Z. Zhang, M. L. Leibowitz, D. Pellman, Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes Dev* **27**, 2513-2530 (2013).
26. S. C. Baca *et al.*, Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677 (2013).
27. M. Gerlinger *et al.*, Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* **46**, 225-233 (2014).
28. M. Gerlinger *et al.*, Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892 (2012).
29. M. Reck, D. F. Heigener, T. Mok, J. C. Soria, K. F. Rabe, Management of non-small-cell lung cancer: recent developments. *Lancet* **382**, 709-719 (2013).
30. H. Davies *et al.*, Mutations of the BRAF gene in human cancer. *Nature* **417**, 949-954 (2002).
31. S. Kang, A. G. Bader, P. K. Vogt, Phosphatidylinositol 3-kinase mutations identified in human cancer are oncogenic. *Proc Natl Acad Sci U S A* **102**, 802-807 (2005).
32. H. Shi *et al.*, Acquired Resistance and Clonal Evolution in Melanoma during BRAF Inhibitor Therapy. *Cancer Discov* **4**, 80-93 (2014).
33. C. J. Lord, A. Ashworth, The DNA damage response and cancer therapy. *Nature* **481**, 287-294 (2012).
34. S. M. Dewhurst *et al.*, Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov* **4**, 175-185 (2014).
35. R. Sotillo, J. M. Schvartzman, N. D. Socci, R. Benezra, Mad2-induced chromosome instability leads to lung tumour relapse after oncogene withdrawal. *Nature* **464**, 436-440 (2010).
36. A. Lee *et al.*, Chromosomal Instability Confers Intrinsic Multi-Drug Resistance. *Cancer Res* **71**, 1858-1870 (2011).

37. C. C. Soria JC, Bahleda R, et al. , Clinical activity, safety and biomarkers of PD-L1 blockade in non-small cell lung cancer (NSCLC). *European Cancer Congress*, (2013).
38. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
39. D. C. Koboldt *et al.*, VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576 (2012).
40. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
41. L. Ding *et al.*, Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506-510 (2012).
42. K. Ye, M. H. Schulz, Q. Long, R. Apweiler, Z. Ning, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
43. K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
44. X. Liu, X. Jian, E. Boerwinkle, dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat* **32**, 894-899 (2011).
45. K. Nakamura *et al.*, Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**, e90 (2011).
46. K. J. McKernan *et al.*, Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**, 1527-1541 (2009).
47. E. Papaemmanuil *et al.*, Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med* **365**, 1384-1395 (2011).
48. I. Varela *et al.*, Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539-542 (2011).
49. G. Schwarz, Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464 (1978).
50. K. Tamura *et al.*, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739 (2011).
51. R. McLendon *et al.*, Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, (2008).
52. The Cancer Genome Atlas Research Network, Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525 (2012).
53. T. Davoli *et al.*, Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948-962 (2013).
54. G. Nilsen *et al.*, Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
55. J. Wang *et al.*, CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* **8**, 652-654 (2011).
56. K. Chen *et al.*, TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. *Genome Res* **24**, 310-317 (2014).
57. W. J. Kent, BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

58. L. Yang *et al.*, Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919-929 (2013).

59. M. Krzywinski *et al.*, Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).

60. P. J. Stephens *et al.*, The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-404 (2012).

61. H. Bengtsson, P. Wirapati, T. P. Speed, A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics* **25**, 2149-2156 (2009).

62. H. Bengtsson, P. Neuvial, T. P. Speed, TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics* **11**, 245 (2010).

63. M. Ortiz-Estevez, A. Aramburu, H. Bengtsson, P. Neuvial, A. Rubio, CalMaTe: a method and software to improve allele-specific copy number of SNP arrays for downstream segmentation. *Bioinformatics* **28**, 1793-1794 (2012).

64. C. Yau *et al.*, A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol* **11**, R92 (2010).

Acknowledgments:

E.B. is a Rosetrees Trust fellow, M.J.H. has a Cancer Research UK fellowship, Nu.M. received funding from the Rosetrees Trust, M.G. is funded by the UK Medical Research Council, I.V. is funded by Spanish Ministerio de Economía y Competitividad subprograma Ramón y Cajal, R.C.R. and D.M.R. are part-funded by the Cambridge Biomedical Research Centre and Cancer Research UK Cancer Centre, PVL is a postdoctoral researcher of the Research Foundation – Flanders (FWO), SMJ is a Wellcome Senior Fellow in Clinical Science, C.S. is a senior Cancer Research UK clinical research fellow and is funded by Cancer Research UK, the Rosetrees Trust, EU FP7 (projects PREDICT and RESPONSIFY, ID:259303), the Prostate Cancer Foundation, and the Breast Cancer Research Foundation.

This research is supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. We would like to thank Servicio Santander Supercomputación for their IT support. The results published here

are in part based upon data generated by the Cancer Genome Atlas pilot project established by the NCI and NHGRI (information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at hhttp://cancergenome.nih.gov/. The data were retrieved through dbGaP authorization (Accession No. phs000178.v5.p5).

Sequence data has been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession numbers EGAS00001000809 and EGAS00001000840.