Article:

Wood, HM, Conway, C, Daly, C et al. (13 more authors) (2015) The clonal relationships between pre-cancer and cancer revealed by ultra-deep sequencing. Journal of Pathology, 237 (3). 296 - 306. ISSN 0022-3417

https://doi.org/10.1002/path.4576

The clonal relationships between pre-cancer and cancer revealed by ultra-deep sequencing

Henry M Wood[1*], Caroline Conway[1,4], Catherine Daly[1], Rebecca Chalkley[1], Stefano Berri[1,5], Burcu Senguven[1,6], Lucy Stead[1], Lisa Ross[1], Philip Egan[1], Preetha Chengot[2], Jennifer Graham[2], Neeraj Sethi[1], Thian K Ong[3], Alec High[2], Kenneth MacLennan[1,2], Pamela Rabbitts[1].

1 Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, LS9 7TF, UK

2 St James's Institute of Oncology, St James's University Hospital, Leeds, LS9 7TF, UK

3 Leeds Dental Institute, Leeds General Infirmary, Leeds, LS2 9LU, UK

4 Now at School of Biomedical Sciences, University of Ulster, Coleraine, Northern Ireland, BT52 1SA, UK

5 Now at Illumina UK Ltd., Chesterford Research Park, Saffron Walden, CB10 1XL, UK.

6 Now at Department of Oral Pathology, Faculty of Dentistry, Gazi University, Turkey.

***** Correspondence to:**

***Email:** h.m.wood@leeds.ac.uk, +44 (0)113 2064070*

Disclosure Declaration

None of the authors of this work has any conflict of interest.

Abstract

The study of the relationships between pre-cancer and cancer and identification of early driver mutations is becoming increasingly important as the value of molecular markers of early disease and personalised drug targets is recognised, especially now the extent of clonal heterogeneity in fully invasive disease is being realised. It has been assumed that pre-cancerous lesions exhibit a fairly passive progression to invasive disease; the degree to which they too are heterogeneous is unknown. We performed ultra-deep sequencing of thousands of selected mutations together with copy number analysis from multiple, matched pre-invasive lesions, primary tumours and metastases from five patients with oral cancer, some with multiple primary tumours presenting either synchronously or metachronously, totalling 75 samples. This allowed the clonal relationships between the samples to be observed for each patient.

We expose for the first time the unexpected variety and complexity of the relationships between this group of oral dysplasias and their associated carcinomas, and ultimately, the diversity of processes by which tumours are initiated, spread and metastasise.

Instead of a series of genomic precursors of their adjacent invasive disease, we have shown dysplasia to be a distinct dynamic entity, refuting the belief that pre-cancer and invasive tumours with a close spatial relationship always have linearly-related genomes. We show that oral pre-cancer exhibits considerable sub-clonal heterogeneity in its own right, that mutational changes in pre-cancer do not predict the onset of invasion, and that the genomic pathway to invasion is neither unified nor predictable.

Keywords: Pre-cancer, oral cancer, tumour progression, clonal evolution.

Introduction

Genomic changes occurring consistently in tumours provide the opportunity to develop disease markers for diagnosis, prognosis and prediction of treatment response, and targets for drug treatment. The international efforts to catalogue genomic changes in tumours [1] are testament to the near universal acceptance of this prospect. Identifying which changes occur early in development has always been recognised as of value, increasingly so now that the genomic heterogeneity of fully invasive tumours is being revealed [2-5].

Carcinomas (especially those with squamous cell histology) often have a recognised pre-cancerous stage, epithelial dysplasia. This provides an opportunity for direct identification of the earliest genomic changes, although the genomes of pre-cancerous lesions receive little attention, most probably due to their small size and the difficulty of obtaining tissue. However, oral cancer is an excellent source of pre-cancerous lesions because its associated dysplasia can be manifest in macroscopically visible lesions including erythroplakia and leukoplakia and is almost always present adjacent to fully-malignant carcinoma within the field of tumour resection. This has enabled several studies to compare the genomes of unrelated pre-invasive lesions and carcinomas [6,7], as well as a few comparing lesions from the same patients [8-10]. Copy number aberrations (CNAs) and microsatellite loci have been linked with early disease and progression..

An additional feature of oral cancer is that the disease is often multifocal. Clinical presentation of lesions can be both synchronously and metachronously and lesions are frequently clonally related [11]. In the present study we took advantage of this to examine five patients with between 11 and 22 samples each. Rather than examine candidate genes, we used high-throughput sequencing to produce an unbiased catalogue of genomic changes for each patient, enabling those changes occurring at a pre-invasive stage, representing the earliest events, to be clearly identified.

An immediate outcome of our analysis was that the distribution of mutations between the various lesions was unpredictable, indicating that many sub-clones co-existed even at the earliest stages of disease. By identifying classes of co-occurring mutations within each patient's group of lesions, we were able to chart the chronology of genomic changes and from this the natural history of each patient's disease.

Materials and methods

Samples

Patients were recruited from the Oral and Maxillofacial Surgery Outpatients' clinic of Leeds General Infirmary and informed consent obtained (ethics REC ref no 07/Q1206/30 and 08/H1306/127). Fresh frozen samples and blood were collected in clinic or from the operating theatre. Formalin-fixed, paraffin-embedded (FFPE) blocks were retrieved from the hospital archives. Following sectioning and dissection where appropriate, DNA was prepared using commercial kits (Qiagen, City, State/Province, Country).

Selection of mutations for analysis

Mutations were selected for each patient based on whole genome (fresh samples) and exome (FFPE samples) sequencing from between three and five samples per patient. Where more than 8μg DNA was available, sequencing was carried out by Complete Genomics and analysed using their proprietary pipeline [12]. For samples with less than 8μg DNA, whole genomes were sequenced using an Illumina HiSeq 2000, to 30X coverage. Reads were aligned to the human genome (hg19) using BWA 0.5.9 [13], processed using the GATK pipeline 1.3.8 [14], and mutations called using SNVmix2 0.11.8 [15,16]. Exome capture of FFPE samples and matching blood was performed using the Sureselect Human All Exome kit (Agilent). An average of 50X coverage was obtained using an Illumina HiSeq2500. Reads were aligned to the human genome, processed using GATK, and mutations called using Varscan2 2.3.5 [17] in somatic mode under default parameters. Due to the mix of sequencing platforms used, these results were only used to nominate mutations for further analysis, and were not used to compare samples directly.

Haloplex custom capture kits (Agilent) were designed to capture the nominated mutations for validation and direct comparison between patients. 10,000 putative somatic variants per patient were chosen for validation and clonal analysis. The selection was enriched for genes and spanned a range of mutant allele frequencies. Additionally, the potential effects of mutations were assessed using the Variant Effect Predictor [18] and Chasm 3.0 [19]. PCR primers for the 30 variants per patient deemed to have the most likely carcinogenic properties were designed using Primer3 2.2.3 [20] to increase the chances that the most informative mutations would be sequenced at high depth.

Deep sequencing of selected variants:

Libraries were prepared from the Haloplex kits for all samples and sequenced on an Illumina HiSeq 2500 to an average of 600X coverage. The PCR-captured variants were sequenced on an Illumina MiSeq to at least 3000X coverage. All reads were aligned to the human genome, processed using GATK, and mutations called using Varscan2 in somatic mode under default parameters.

Only variants that were completely absent in the normal blood sample, had a depth of at least 50X in all samples and which were on the original list used for Haloplex/PCR design were considered for clonal analysis. These mutations had, therefore, all been called by one sequencing method and then validated by another method in at least two samples. This ensured that no results were based on *de novo* mutation calls, minimising the chance of results being biased by any FFPE damage, or other sequencing or mutation-calling artefact.

Low coverage copy number sequencing and viral detection

Copy number data was obtained from all samples using previously published protocols [21]. DNA was sequenced on an Illumina HiSeq 2500 to around 0.15X coverage, and aligned to the human genome using BWA. Copy number data was produced using CNAnorm 1.3.5 [22] under default parameters with manual correction to allow for heterogeneous regions of copy number change, using a control sample generated by pooling reads from 20 British Caucasian individuals from the 1000 Genomes Project [23]. Breakpoints were called using DNAcopy 1.36.0 [24].

The same sequence data was also aligned to all known viral sequences to detect any Human Papillomavirus (HPV) infection [25].

Clonal reconstruction (see also supplementary information online)

Mutations were assigned to classes, and clonal relationships assigned, using methods similar to those in Newburger *et al.* [26], based on patterns of co-occurrence. The presence or absence of each mutation or copy number breakpoint across all samples was noted and combined to form a text string, *e,g.* 11011 or 01101 for two mutations across five samples. The strings for all events were compared, and identical strings grouped together to form classes. These classes represented detectable sub-clones. Classes supported by only one point mutation were discarded because of the increased chance of error due to miscalling, unless that point mutation was captured both by haloplex and PCR.

The cellular frequencies of copy number events were estimated for each sample using CNAnorm. Cellular frequencies of point mutations were estimated using PyClone 0.12.7 [27] taking into account epithelial cell content, overall ploidy and local copy number. Co-occurring classes were examined to ascertain if any two classes were present in over 50% of cells in a sample. This would indicate, by the pigeonholing principle, that there must be some cells containing both classes, therefore, the classes must be part of the same sequential lineage.

Each sample had a large number (14-89) of private mutations, only found in that sample and mostly at low frequencies. These were not considered for clonal analysis and had a higher chance of being spurious because they were not validated in more than one sample.

Results

Sequencing summary **(see also supplementary information online)**

We examined the clonal relationships between samples for five patients. Detailed clinical and sample information for each patient is given in the supplement, along with sequencing metrics for each sample. C.G > T.A mismatch rates, a known proxy for FFPE damage [28], were measured and found to vary very little between fresh and fixed samples. Between 837 and 2039 putative somatic mutations were validated out of 10000 in at least one sample per patient. Between 544 and 1258 of these were re-sequenced in enough depth (>50X) in all samples to be considered for clonal analysis, together with copy number data. Between 17 and 32 classes of somatic events were segregated according to patterns of co-occurrence (Table 1). Lists of point mutations, CNAs and classes for the five patients denoted as PG008, PG055, PG030, PG136 and PG019 are shown respectively in Supplementary Tables S6 and S7; S8 and S9; S10 and S11; S12 and S13; and S14.

Patients' summary

Each patient is described in brief below and in more detail in **supplementary information online.**.

Patient PG008 (Figure 1) had a left ventral tongue lesion. The HGD sample shared only one CNA with the rest of the lesion. A small proportion of the cells in the LGD sample contained 17 mutations (class 2, including a *TP53* mutation) that were ubiquitous throughout the carcinoma samples. From this single clonal origin, the carcinoma exhibited spatial heterogeneity.

Patient PG055 (Figure 2) had an HPV-driven lesion in the floor of mouth. The normal samples exhibited no viral infection. One marginal dysplasia sample shared only two CNAs with the rest of the lesion. The remaining samples shared 172 mutations (class 2). This common ancestor split into two sub-clones. Class 3 was only found in dysplasia samples. Class 4 was found in all carcinoma samples and as a sub-clonal fraction in the part of the dysplasia nearest the carcinoma. All of the carcinoma samples contained the class 6 events not found in any dysplasia samples. Most of the dysplasia samples were therefore dominated by sub-clones not found in the carcinoma.

Patient PG030 (Figure 3) presented with a right retromolar lesion, a left floor of mouth lesion and right-sided neck metastases. The disease appears to have originated in the right lesion. *TP53* was amongst the earliest genes mutated (class 2), in all neoplastic samples apart from one dysplasia. The biggest split in samples was between classes 3 and 4. Class 3 was absent in the left lesion. Class 4 was absent in the dysplasia furthest from the left lesion, but ubiquitous on the left. Both were present throughout the right-sided carcinoma and metastases, indicating mixed populations. The fact that this mixing was replicated in the metastases indicates that the metastatic event involved a population of cells, rather than just one founder. As every right-sided invasive sample contained both sub-clones, it is possible that they depended on each other in some way, cooperating to maintain the invasive phenotype. It is unclear if class 7, only found in invasive samples, arose in cells with class 3, 4 or 5 mutations, or represented a mixture of sub-clones descended from a combination of them. It was never found with only one of its potential ancestors (which were themselves not always found together), suggesting that one sub-clone alone was not enough to maintain invasion. Every left-sided sample contained class 6, which was only found in one dysplasia sample on the right side.

Patient PG136 (Figure 4) had a right lateral tongue lesion with nodal metastases. Unusually, this patient had no history of alcohol or tobacco use, or evidence of viral infection. Very few point mutations were present at high cellular frequencies, and none were shared between samples. Only copy number data is presented here. Some common events were found in adjacent normal tissue, while not appearing in all dysplasia samples, suggesting that the histological definition of normal did not always match the genomic profile. Most samples contained events at both clonal and sub-clonal cellular frequencies, indicating considerable heterogeneity, with clonal events in some samples appearing to be sub-clonal in others. There did however appear to be a dominant population at high cellular frequencies in all the SCC samples, appearing at class 5. The only homogenous samples were the biopsy, whose small size may have reduced the diversity sampled, and the metastases, indicating that only some of the multiple sub-clones detected had been able to successfully metastasise.

Patient PG019 (Figure 5) presented with multiple lesions over a 19-year period. All samples, including the earliest normal tissue, contained the 35 class 1 mutations, including a *TP53* mutation. We detected a cumulative increase in mutations from 1992 to 2007, when an additional *TP53* mutation (class 4) coincided with an increase in the frequency of lesions arising. A clone (class 5) first detected in 2008 was seen in the 2009 carcinoma samples, but not the 2009 adjacent dysplasia and normal samples, suggesting that the 2009 lesion had a mixed clonal ancestry. This could have occurred through multiple colonisations of the region, or through colonisation by a mixed population of cells. In the 2009 dysplasia and normal samples, and all subsequent samples, we only detected the pre-2008 classes. Mutations first detected in the 1992 and 2009 carcinomas (classes 2 and 6), were never detected subsequently, indicating successful surgical removal.

Heterogeneity of carcinomas

Clonal variation between and within carcinoma samples from the same lesion was detected in all five patients. The two patients with multiple lesions also demonstrated variation between lesions. Patients PG008, PG055 and PG136 showed the simplest pattern, whereby a single founder sub-clone for all carcinoma samples was detected. This then diverged into a more heterogeneous population in an approximately spatial pattern. Patient PG019 showed differences between pre-2007 lesions and post-2007 lesions. There was some evidence that the 2009 lesion was founded by a mixed population of sub-clones. Patient PG030 was the most complex. Every carcinoma in the right lesion appeared to contain the same mixed population of sub-clones, with no detectable spatial variation. The left lesion contained a different mixture of sub-clones, related to a subset of those from the right, with additional spatial variation detected.

Heterogeneity of dysplasias

Patient PG019 displayed variation between different lesions, but none was detectable within any one individual lesion. For the remaining four patients, sub-clonal variation was detected between dysplastic samples from the same lesion. The variation was not a case of simple progression towards carcinoma, with different samples capturing different stages along that progression. A number of dysplasia-only sub-clones were also detected, most notably in patients PG055 (class 3) and PG136 (classes 3 and 4).

Progression from normal to dysplasia

Patients PG008 and PG055 showed clear distinction between normal and neoplastic samples. PG030 and PG136 showed overlap in the genomic profiles of some normal and some dysplasia samples, hinting that histologically normal cells may have harboured dysplastic mutations. PG019 normal samples all contained the major sub-clone detected in their adjacent dysplasia. Only two low-grade dysplasia samples were collected, and they both contained mutations that were absent from nearby high-grade dysplasia samples but were however found in carcinoma, suggesting that histological grade and genomic profiles did not closely correlate.

Progression from dysplasia to carcinoma

The carcinoma samples from PG008 and PG055 both showed a clear monoclonal origin, which was detectable in a small population of cells in one dysplasia sample, and was itself descended from the common ancestor of all the dysplasia samples. PG136 was too heterogeneous to confidently predict the origin of the invasive clone. Most PG019 carcinomas were indistinguishable from their associated dysplasia. The exception was the 2009 lesion. The carcinoma appeared to be closely related to the 2008 dysplasia, while the 2009 dysplasia was more closely related to the 2007 disease. The 2009 carcinoma was therefore not descended from its adjacent dysplasia. The right lesion of PG030 appeared to contain two distinct populations. Descendants of both of these were found in every carcinoma sample, indicating some kind of clonal cooperation. The left lesion was descended from just one of the right-sided dysplastic sub-clones, not from a right-sided carcinoma sample. Once this sub-clone had arrived on the left, the subsequent dysplasia and carcinoma samples were indistinguishable.

Metastasis

Two patients presented with metastases. The PG030 primary carcinoma exhibited a mixed population of at least two distinct sub-clones. This heterogeneity was maintained in the metastases, suggesting a population of cells metastasising. The PG136 primary lesion was even more heterogeneous, but the metastases showed a clear monoclonal origin. This could imply selection of a metastatic clone, or it is also possible that the heterogeneity in PG030 was a result of clonal cooperation, and was thus necessary to found a metastasis, whereas the PG136 heterogeneity was possibly merely a product of constant accumulation of passenger events and was not necessary to maintain.

Discussion

This study has revealed in unprecedented detail the clonal relationships between pre-cancer and cancer from the same patients, charting the chronological order of genomic changes as lesions developed.

It has been generally believed that the relationship between dysplasia and carcinoma is a relatively simple one, and that in comparison with the carcinoma, the dysplasia is fairly passive, with mutations gradually accumulating until some of the cells become invasive. Our results expose the scale of variety and the complexity of relationships between dysplasia and invasive oral carcinoma. One patient (PG008) displayed classical progression, with gradual accumulation of mutations leading to invasion, whereas other patients showed considerable sub-clonal heterogeneity within the dysplasia itself, with dysplastic sub-clones independently acquiring mutations in several oncogenes without becoming invasive. PG019 and PG030 showed evidence of dysplastic cells leaving the primary tumour and seeding fresh lesions and of mutationally-similar cells having invasive or non-invasive phenotype depending on their surroundings. PG030 also showed potential cooperation between sub-clones to maintain a lesion.

Study design

When studying the relationship between dysplasias and carcinoma, the usual method is to examine groups of both types of sample [6,7]. The number of studies of dysplasia and carcinoma in the same patient are very few. They tend to examine candidates such as *TP53* or microsatellites and use limited sample numbers, so the full extent to which different morphological grades, or different regions of the same lesion are similar or different is not fully exposed [8,9].

By combining point mutation and CNA data, and using multiple samples per patient, we have been able not only to show the relationships between different cell populations within and between lesions, but also to associate tumour progression in each patient with specific genomic changes and infer the processes by which each lesion was initiated, developed and then spread.

We chose to use the recently described method [3-5] whereby initial sequencing results are used to nominate a selected group of mutations for much deeper sequencing. Our clonal relationships were entirely derived from the capture and CNA results, with no further input from our initial sequencing experiments. There have been a number of informatics approaches published [27,29,30] to infer sub-clonal architecture in tumour samples. These cluster mutations using accurately measured cellular frequencies. However, in common with another recent study [26], we found that our mutations did always split into multiple distinct clusters. As for that project and one other [2], we had the advantage of working with large numbers of samples per patient. Therefore we were able to adopt a similar approach and classify mutations by patterns of co-occurrence, relying on direct observations of mutation presence or absence, rather than inferring relationships from cellular frequencies which are themselves inferred.

TP53

We have listed all the mutations validated in our samples (Supplementary Tables S6, S8, S10, S12, S14). Some genes were mutated in more than one patient, but not in an early clone shared between many samples. The only gene that was frequently mutated at an early stage of development was *TP53*.

It has long been suspected that the *TP53* gene is important in head and neck SCC, and that it is often mutated in dysplasia as well as carcinoma [31-33]. This was confirmed by a recent exome sequencing study, which showed that *TP53* was by far the most commonly mutated gene and that P53 inactivation, either by mutation or HPV infection, was almost universal for this disease [34]. Our results reiterate that finding, and confirm that where present (by mutation in three patients and with HPV infection in one), it was always amongst the earliest changes detected. In PG019, a second subsequent *TP53* mutation coincided with more aggressive disease, making it likely that in conbination the two mutations totally ablated P53 protein.

The natural history of head and neck SCC

The frequent observation of the close physical association of dysplasia and carcinoma in head and neck SCC is generally believed to be indicative of a sequential developmental relationship [35-38].

In this study, the dysplasia samples all shared a common ancestor with the nearby SCC, but were not necessarily directly ancestral to it. The SCC samples were usually descended from a minor sub-clone from within one of the dysplasia samples. In the case of PG030 and PG019, SCC samples were descended from a relatively distant dysplasia sample, rather than the adjacent one.

All patients exhibited some level of dysplastic heterogeneity, which was particularly pronounced in patients PG055 and PG136. These samples showed evidence that the dysplasias developed considerable clonal variation before the invasive clone emerged, and have possibly continued to evolve since [39].

Many different pathways capable of generating HNSCC have been proposed, and associated with tobacco, alcohol use and viral infection [34,40-42]. Our five patients all developed their histologically similar tumours in different ways. PG136 appeared almost devoid of point mutations, so it is not known how the disease was triggered. PG055 was fuelled by viral infection. PG008 either had a very inactive, or short pre-cancerous phase, with most mutations occurring from either shortly before invasion, or afterwards.

Both lesions from PG030 were mixed populations, with all samples containing multiple sub-clones, leading to the possibility that they depended on each other to maintain an invasive phenotype, as has recently been described in breast cancer [43]. The left lesion was seeded by a pre-cancerous sub-clone from the right lesion. PG019 showed a slow disease progression, with gradual increase in aggressiveness as mutations accumulated, and eventual spread of mutated cells throughout the oral cavity.

Patterns of tumour evolution

Six possible different patterns of tumour initiation and progression have been inferred from our observations. These processes are summarised in Figure 6. Most patients exhibited more than one of these processes as their disease spread to a different part of the oral cavity or metastasised. These only represent the patterns inferred from these five patients, rather than any general rule. The fact that six patterns were inferred from only five patients suggests that applying any general rules to HNSCC development is a very simplistic approach.

The simplest process is characterised by local initiation and progression (Figure 6A). Initial mutations in normal tissue give rise to dysplasia containing one or more sub-clones. In one of these sub-clones, further mutations give rise to an invasive clone. This was observed in PG008, PG055 and PG136. Slightly more complex is the possibility of cooperating sub-clones (Figure 6B). PG030 invasive samples all contained multiple sub-clones that were also seen in non-invasive samples, sometimes in a majority of cells. It is possible that the sub-clones depend on each other to maintain an invasive phenotype. New SCCs were also seeded by distant dysplasias (Figure 6C) either spreading through the epithelium [11] or by detaching and alighting in the new position [44]. Most of the PG019 SCCs and the left lesion from PG030 developed this way. The PG019 2009 SCC was founded by new cells (from the 2008 lesion) seeding an area that has already been colonised by the 2007 clone (Figure 6D).

Two types of nodal metastasis were seen. In PG030, the sub-clonal variation of the primary tumour was maintained in the metastases (Figure 6E), whereas in PG136, a heterogeneous primary tumour gave rise to a single metastatic clone (Figure 6F).

Conclusions

The work described here exposes the diversity of processes by which histologically similar tumours are initiated, develop and spread, and the hitherto unknown complexity in the relationships between pre-invasive and invasive disease. We show that dysplasia has considerable sub-clonal variation, and that a carcinoma is often descended from a minority sub-clone in just part of that dysplasia.

Acknowledgements

This work was supported by Yorkshire Cancer Research [L341PG] and the Betty Woolsey Endowment.

Data access

Sequence data from this study has been deposited in the European Nucleotide Archive, accession number PRJEB6588.

Author contributions

HW, CC, SB, LS and PR designed the study. CC, RC, BS, LR, PE, PC, JG, NS, TO and AH collected patient material and clinical information. BS, PC, JG, AH, KM provided pathological examination. HW, CC, CD, RC, BS, LR, and PE did laboratory work. HW, CC, SB, LS and PR analysed the data. HW and PR wrote the paper. All authors edited and approved the manuscript.

References

1. Collins FS, Barker AD. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Scientific American* 2007; **296**: 50-57.
2. Gerlinger M, Horswell S, Larkin J*, et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* 2014; **46**: 225-233.
3. Nik-Zainal S, Alexandrov LB, Wedge DC*, et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* 2012; **149**: 979-993.
4. Nik-Zainal S, Van Loo P, Wedge DC*, et al.* The life history of 21 breast cancers. *Cell* 2012; **149**: 994-1007.
5. Walter MJ, Shen D, Ding L*, et al.* Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med* 2012; **366**: 1090-1098.
6. Mehanna HM, Rattay T, Smith J*, et al.* Treatment and follow-up of oral dysplasia - a systematic review and meta-analysis. *Head & neck* 2009; **31**: 1600-1609.
7. Zhang L, Poh CF, Williams M*, et al.* Loss of heterozygosity (LOH) profiles--validated risk predictors for progression to oral cancer. *Cancer Prev Res (Phila)* 2012; **5**: 1081-1089.
8. Califano J, Westra WH, Meininger G*, et al.* Genetic progression and clonal relationship of recurrent premalignant head and neck lesions. *Clin Cancer Res* 2000; **6**: 347-352.
9. Tsui IF, Garnis C, Poh CF. A dynamic oral cancer field: unraveling the underlying biology and its clinical implication. *The American journal of surgical pathology* 2009; **33**: 1732-1738.
10. Noutomi Y, Oga A, Uchida K*, et al.* Comparative genomic hybridization reveals genetic progression of oral squamous cell carcinoma from dysplasia via two different tumourigenic pathways. *J Pathol* 2006; **210**: 67-74.
11. Tabor MP, Brakenhoff RH, Ruijter-Schippers HJ*, et al.* Multiple head and neck tumors frequently originate from a single preneoplastic lesion. *Am J Pathol* 2002; **161**: 1051-1060.
12. Drmanac R, Sparks AB, Callow MJ*, et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**: 78-81.
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 60.
14. McKenna A, Hanna M, Banks E*, et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297-1303.
15. Shah SP, Morin RD, Khattra J*, et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 2009; **461**: 809-813.
16. Goya R, Sun MG, Morin RD*, et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* 2010; **26**: 730-736.
17. Koboldt DC, Zhang Q, Larson DE*, et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012; **22**: 568-576.

18. McLaren W, Pritchard B, Rios D*, et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010; **26**: 2069-2070.
19. Wong WC, Kim D, Carter H*, et al.* CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* 2011; **27**: 2147-2148.
20. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* 2007; **23**: 1289-1291.
21. Wood HM, Belvedere O, Conway C*, et al.* Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res* 2010; **38**.
22. Gusnanto A, Wood HM, Pawitan Y*, et al.* Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* 2012; **28**: 40-47.
23. Genomes Project C, Abecasis GR, Altshuler D*, et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061-1073.
24. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 2007; **23**: 657-663.
25. Conway C, Chalkley R, High A*, et al.* Next-generation sequencing for simultaneous determination of human papillomavirus load, subtype, and associated genomic copy number changes in tumors. *J Mol Diagn* 2012; **14**: 104-111.
26. Newburger DE, Kashef-Haghighi D, Weng Z*, et al.* Genome evolution during progression to breast cancer. *Genome Res* 2013; **23**: 1097-1108.
27. Roth A, Khattra J, Yap D*, et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* 2014.
28. Yost SE, Smith EN, Schwab RB*, et al.* Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res* 2012.
29. Jiao W, Vembu S, Deshwar AG*, et al.* Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics* 2014; **15**: 35.
30. Fischer A, Vazquez-Garcia I, Illingworth CJ*, et al.* High-definition reconstruction of clonal composition in cancer. *Cell reports* 2014; **7**: 1740-1752.
31. Waridel F, Estreicher A, Bron L*, et al.* Field cancerisation and polyclonal p53 mutation in the upper aero-digestive tract. *Oncogene* 1997; **14**: 163-169.
32. Rowley H, Sherrington P, Helliwell TR*, et al.* p53 expression and p53 gene mutation in oral cancer and dysplasia. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery* 1998; **118**: 115-123.
33. Shahnavaz SA, Regezi JA, Bradley G*, et al.* p53 gene mutations in sequential oral epithelial dysplasias and squamous cell carcinomas. *J Pathol* 2000; **190**: 417-422.
34. Stransky N, Egloff AM, Tward AD*, et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* 2011; **333**: 1157-1160.
35. Ho MW, Risk JM, Woolgar JA*, et al.* The clinical determinants of malignant transformation in oral epithelial dysplasia. *Oral oncology* 2012; **48**: 969-976.
36. Ho MW, Field EA, Field JK*, et al.* Outcomes of oral squamous cell carcinoma arising from oral epithelial dysplasia: rationale for monitoring premalignant oral lesions in a multidisciplinary clinic. *The British journal of oral & maxillofacial surgery* 2013; **51**: 594-599.
37. Brennan M, Migliorati CA, Lockhart PB*, et al.* Management of oral epithelial dysplasia: a review. *Oral surgery, oral medicine, oral pathology, oral radiology, and endodontics* 2007; **103 Suppl**: S19 e11-12.
38. Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer* 1953; **6**: 963-968.
39. Johnson CE, Gorringe KL, Thompson ER*, et al.* Identification of copy number alterations associated with the progression of DCIS to invasive ductal carcinoma. *Breast cancer research and treatment* 2012; **133**: 889-898.
40. Bhattacharya A, Roy R, Snijders AM*, et al.* Two distinct routes to oral cancer differing in genome instability and risk for cervical node metastasis. *Clin Cancer Res* 2011; **17**: 7024-7034.
41. Koch WM, Lango M, Sewell D*, et al.* Head and neck cancer in nonsmokers: a distinct clinical and molecular entity. *The Laryngoscope* 1999; **109**: 1544-1551.
42. Chung CH, Parker JS, Karaca G*, et al.* Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell* 2004; **5**: 489-500.
43. Cleary AS, Leonard TL, Gestl SA*, et al.* Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature* 2014; **508**: 113-117.

44. Pipinikas CP, Kiropoulos TS, Teixeira VH*, et al.* Cell migration leads to spatially distinct but clonally related airway cancer precursors. *Thorax* 2014; **69**: 548-557.

Table 1: Patient summary: Brief clinical details are shown for each patient, the numbers of somatic mutations validated, and the number of mutations sequenced to at least 50X coverage in all samples used to generate classes, as well as the number of derived classes.

Patient	Main clinical features	Risk Factors	TMN status	Samples	Somatic mutations	Shared somatic mutations	Validated somatic mutations	Mutations used in analysis	Classes
PG008	Single tumour	Alcohol, tobacco	pT2N0M0	10	FR-D 166185 FR-C 17433	3656	837	543	17
PG019	Multiple tumours over 20 year period.	Previous smoker	pT1N0M0 (first tumour)	18	FR-1-D 134164 FR-2-D 186770 FR-C 23140	57477	2039	795	22
PG030	Two synchronous tumours. Neck metastases.	Alcohol, tobacco	pT2N2bM0 (right) pT2N0M0 (left)	22	L-FR-C 213601 R-FR-C 12791	1161	1454	921	32
PG055	Single tumour	HPV	pT1N0M0	11	FR-D 106499 FR-C 102403	34263	1593	1038	19
PG136	Single tumour. Neck metastases.	None	pT2N2bM0	14	FR-1-D 910626 FR-1-C 14841	1598	905	683	24

Figure legends

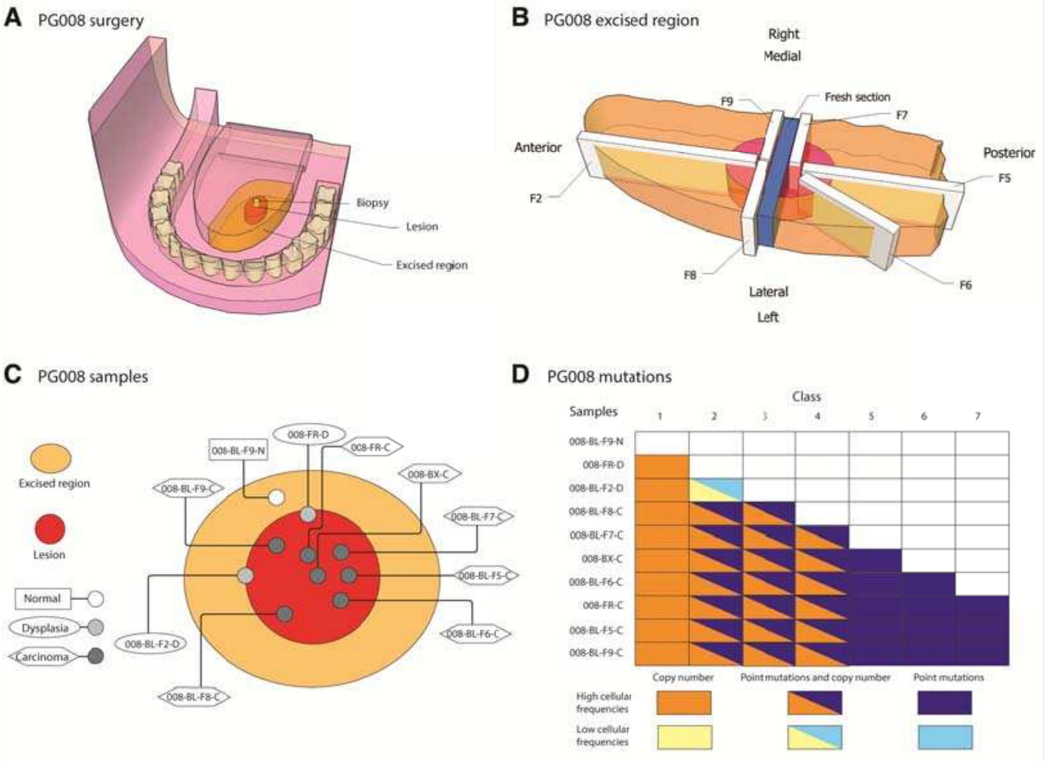

Figure 1: The position of the lesion from patient PG008 (a), the areas of the excised region from which samples were taken (b), the approximate positions of the samples (c) and how the samples relate to the mutation classes (d). In fig 1a, the initial biopsy is shown in yellow, the region of abnormal tissue is shown in red and the surgically excised region is shown in orange. In fig 1b, the excised region is shown in isolation allowing the visualization of the positions of the fresh specimen (blue) and fixed blocks (white). The abnormal tissue is shown in red. In fig 1c, all the samples analysed are shown. See table 1 for a block to sample conversion key. In fig 1d, filled rectangles indicate the presence of a class of mutations. The colours indicate whether the class is supported by point mutations, copy number or both. The shade indicates whether the class is found at high or low cellular frequencies. Classes found in only one sample are not shown.

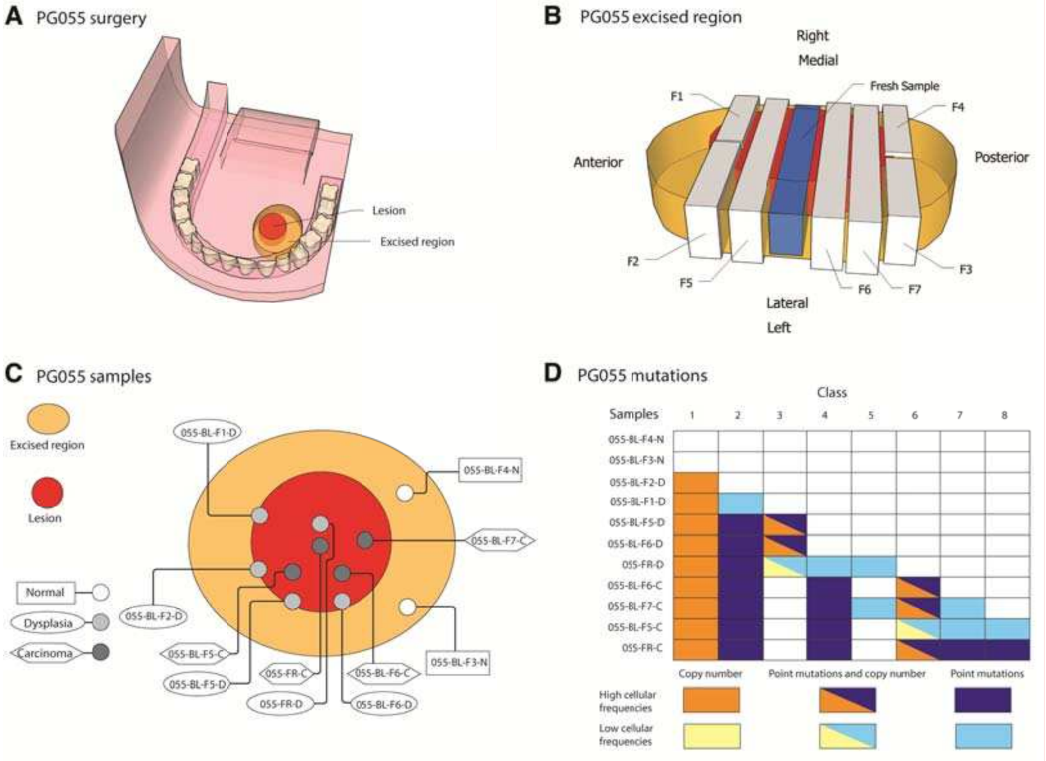

Figure 2: The position of the lesion from patient PG055 (a), the areas of the excised region from which samples were taken (b), the approximate positions of the samples (c), and how the samples relate to the classes of mutations (d). The distal portion of the tongue is not shown in fig 2a to aid visibility.

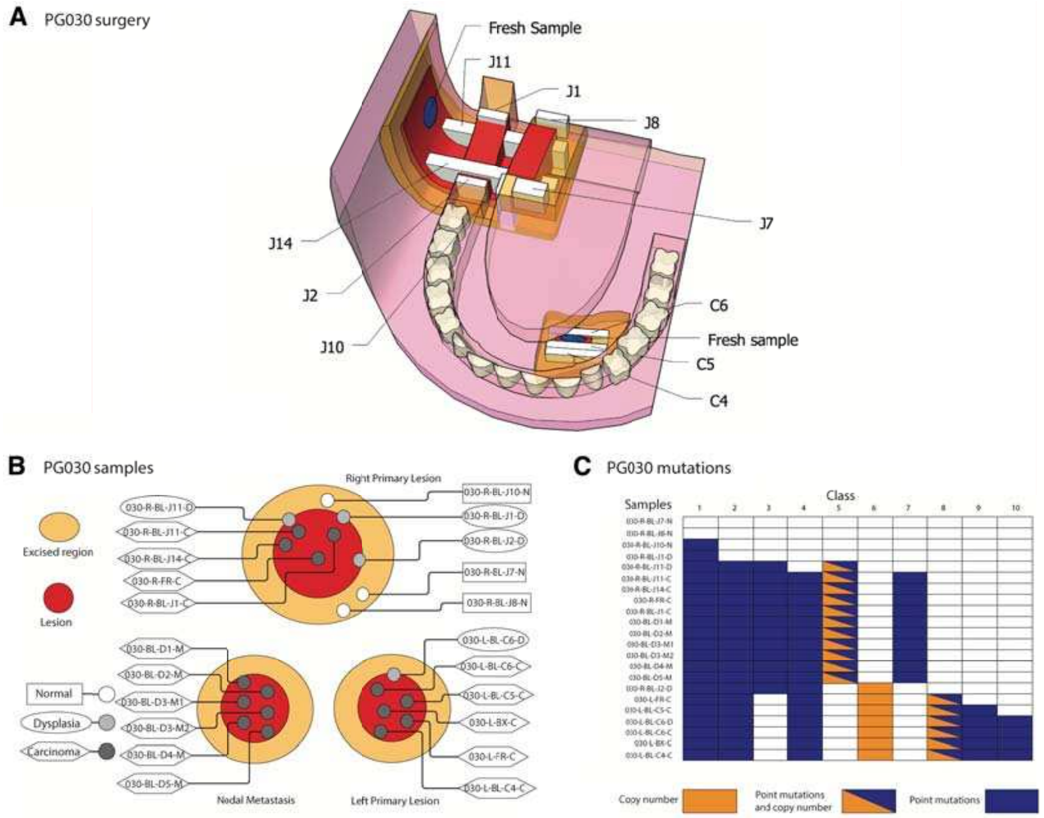

Figure 3: The position of the lesion from patient PG030 and areas of the excised region from which samples were taken (a), the approximate positions of the samples (b) and how the samples relate to the classes of mutations (c). This patient had two separate synchronous lesions. The nodal metastasis is not shown in fig 3a.

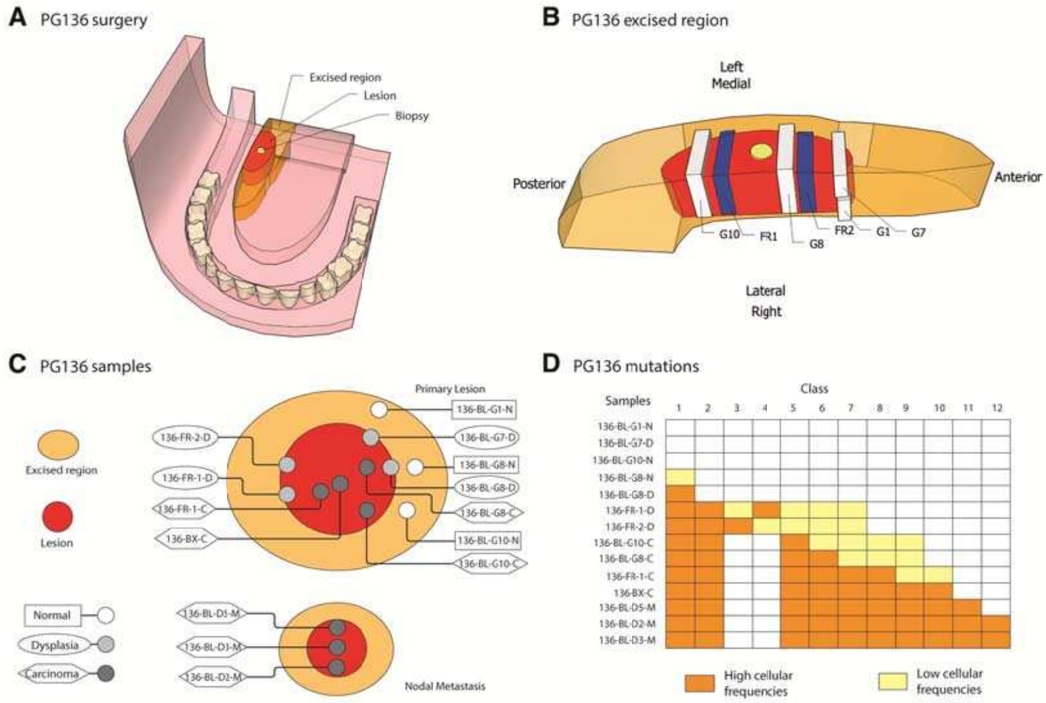

Figure 4: The position of the lesion from patient PG136 (a), the areas of the excised region from which samples were taken (b), the approximate positions of the samples (c) how the samples relate to the classes of mutations (d). The nodal metastasis is not shown in fig 4a/b.

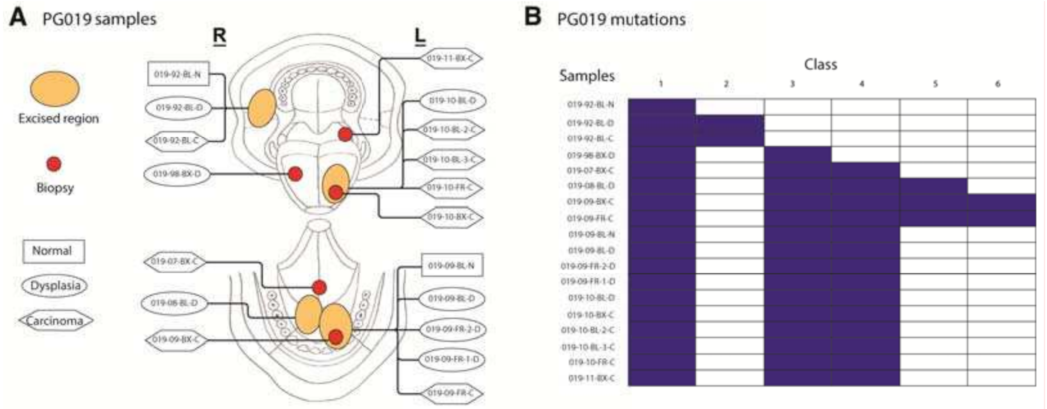

Figure 5: The approximate positions of the samples from patient PG019 (a) and how the samples relate to the classes of mutations (b). The year of surgery is indicated by the second set of digits in the sample name, e.g.. 019-**92**-BL-N is a sample taken from 1992.

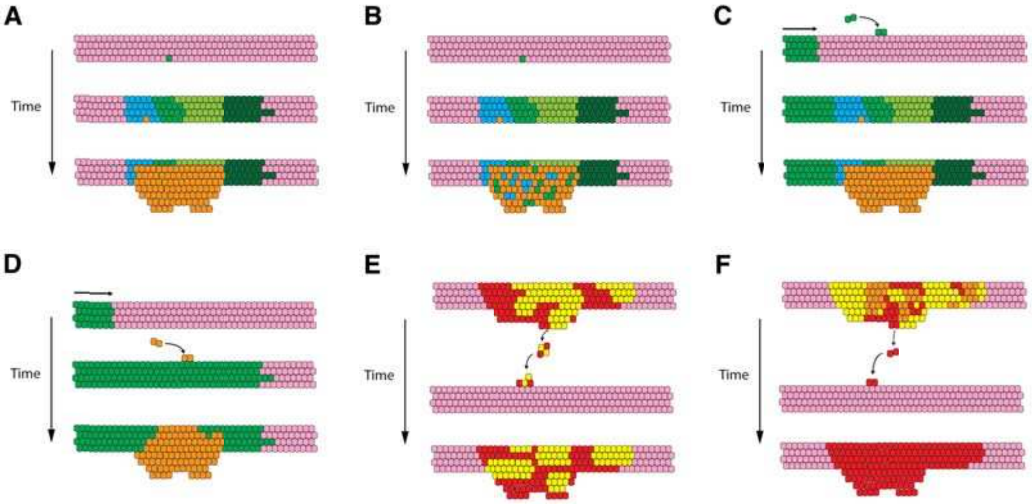

Figure 6: The six patterns observed by which tumours were initiated. For all panels, normal cells are shown in pink, and different clones of dysplasia and carcinoma are shown in other colours. Clonal diversity arising later in the subsequent carcinoma is not shown: (a) The tumour develops entirely *in situ*. A (green) dysplasia clone arises in normal tissue and develops into a multiclonal dysplasia in which a carcinoma clone (orange) arises. (b) Similar to (a), but the invasive carcinoma maintains a population of the pre-invasive sub-clones. (c) Dysplasia cells (green) from elsewhere travel to a new area of normal tissue, either through lateral spread or discrete seeding, and develop into a multiclonal dysplasia in which a carcinoma clone (orange) arises. (d) Dysplasia cells (green) spread to a new area of normal tissue and develop into a new dysplasia. Later, carcinoma cells (orange) from a separate lesion arrive and develop alongside the existing dysplasia. (e) A well developed tumour sheds multiple cells, which invade a fresh patch of normal tissue, maintaining clonal diversity. (f) A well developed tumour sheds cells, which invade a fresh patch of normal tissue. Only one, or a few clones from the original lesion contribute to the new lesion, so diversity is reduced.