

Multi-tiered genomic analysis of head and neck cancer ties *TP53* mutation to 3p loss

Andrew M Gross¹, Ryan K Orosco², John P Shen³, Ann Marie Egloff⁴, Hannah Carter³, Matan Hofree⁵, Michel Choueiri³, Charles S Coffey², Scott M Lippman^{2,3,6}, D Neil Hayes⁷, Ezra E Cohen^{3,6}, Jennifer R Grandis⁴, Quyen T Nguyen^{2,6} & Trey Ideker^{1,3,5,6}

Head and neck squamous cell carcinoma (HNSCC) is characterized by aggressive behavior with a propensity for metastasis and recurrence. Here we report a comprehensive analysis of the molecular and clinical features of HNSCC that govern patient survival. We find that *TP53* mutation is frequently accompanied by loss of chromosome 3p and that the combination of these events is associated with a surprising decrease in survival time (1.9 years versus >5 years for *TP53* mutation alone). The *TP53*-3p interaction is specific to chromosome 3p and validates in HNSCC and pan-cancer cohorts. In human papillomavirus (HPV)-positive tumors, in which HPV inactivates *TP53*, 3p deletion is also common and is associated with poor outcomes. The *TP53*-3p event is modified by mir-548k expression, which decreases survival further, and is mutually exclusive with mutations affecting RAS signaling. Together, the identified markers underscore the molecular heterogeneity of HNSCC and enable a new multi-tiered classification of this disease.

It is increasingly appreciated that the diversity of clinical outcomes in HNSCC is likely a reflection of the molecular heterogeneity of the tumor population^{1–3}. Previous studies have led to the identification of a variety of genes and other molecular features for stratifying HNSCC tumors^{4–8}, such as efforts to cluster gene expression profiles to define subtypes. To comprehensively define this heterogeneity in common tumor types, including in HNSCC, The Cancer Genome Atlas (TCGA) project has generated multi-tiered molecular profiles for over 7,000 tumors from affected individuals, providing an unprecedented opportunity to study the

complex interrelations among fundamentally different types of molecular events and clinical outcomes, such as survival time.

Here we have built on the infrastructure established by TCGA to systematically and transparently unravel these complex relationships for HNSCC. To this end, we obtained all available molecular and clinical data from TCGA (TCGA HNSCC working group, unpublished data) as of the Firehose run on 15 January 2014 (Online Methods and **Supplementary Table 1**). Five tiers of data—somatic mutations, chromosomal aberrations, mRNA expression, microRNA (miRNA) expression and clinical variables—were analyzed for a total of 378 HNSCC cases, resulting in the measurements of over 34,000 molecular or clinical values for each case (**Supplementary Fig. 1a**). All data processing and analysis have been documented in a series of IPython Notebooks⁹ (Online Methods). Because old age and HPV status are associated with distinct molecular profiles and clinical outcomes¹ (**Supplementary Fig. 2**), we focused our analysis on the 250 cases under 85 years of age with HPV-negative tumors and complete molecular profiles.

RESULTS

Identification of prognostic events in HNSCC

We first sought to distill the TCGA data set into a set of informative molecular and clinical events with potential relevance to cancer. First, individual somatic mutations and mRNA expression levels were integrated with knowledge of human molecular pathways to define aggregate ‘pathway-level events’ (Online Methods and **Supplementary Fig. 1b–e**). Second, both individual and pathway events were filtered to select those that occurred at high frequency (somatic mutations, chromosomal aberrations) or involved differential expression (mRNA and miRNA levels) in tumor versus normal tissue. The result of this analysis was a pool of 878 total events combined over all 5 tiers of data (**Supplementary Fig. 1a**).

Next, we screened for individual events within each data type that would be strongly predictive of survival, identifying 82 prognostic events out of the 878 total (**Fig. 1a** and **Supplementary Table 2**). Among somatic mutation events, *TP53* mutation was most strongly predictive overall, resulting in poor prognosis (hazard ratio (HR) = 2.9 ± 0.8 (standard error); Benjamini-Hochberg-corrected $P < 0.01$). As has been observed previously, survival outcomes were dependent on the *TP53* protein domain affected by the mutation or its predicted functional status¹⁰ (**Fig. 1b**). However, we found that cases with mutations predicted to be non-disruptive of function nonetheless had

¹Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, California, USA. ²Division of Head and Neck Surgery, University of California, San Diego, La Jolla, California, USA. ³Department of Medicine, University of California, San Diego, La Jolla, California, USA. ⁴Department of Otolaryngology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. ⁵Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA. ⁶Moore Cancer Center, University of California, San Diego, La Jolla, California, USA. ⁷Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. Correspondence should be addressed to T.I. (ideker@ucsd.edu).

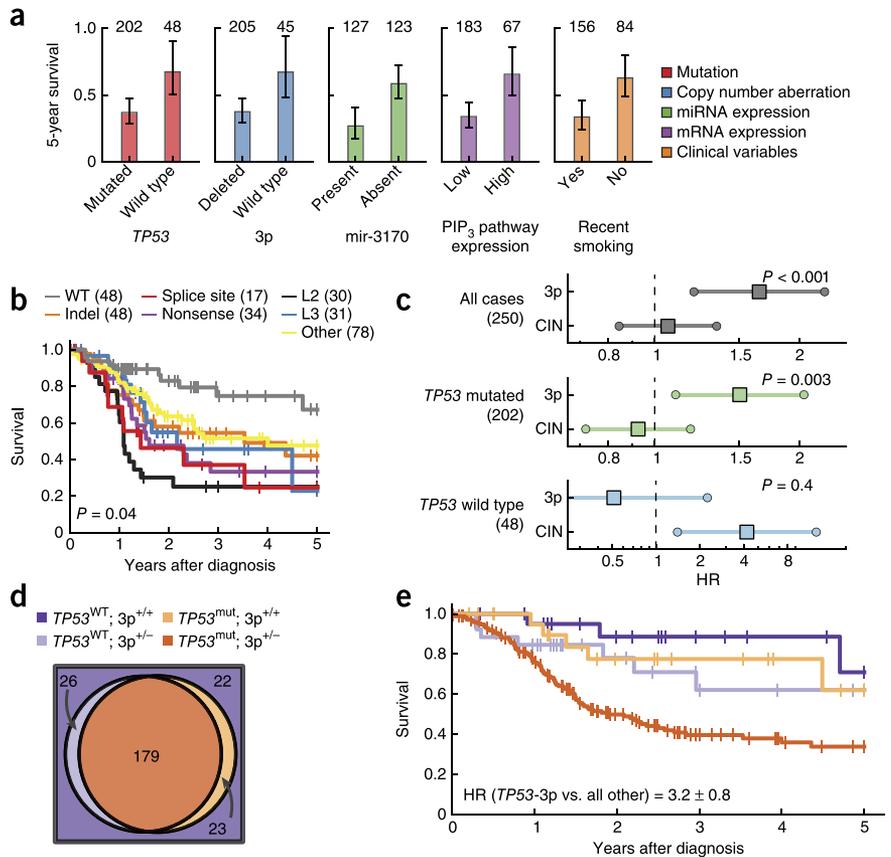
Received 20 March; accepted 10 July; published online 3 August 2014; doi:10.1038/ng.3051

Figure 1 Prognostic effects and co-occurrence of *TP53* and 3p events. **(a)** Five-year survival (error bars indicate 95% confidence intervals) for the most significant events of each data type. Numbers above the bars represent the number of cases with each event. **(b)** Comparison of 5-year survival for cases with different types of non-silent *TP53* mutation versus cases with wild-type (WT) *TP53*. L2 and L3 represent *TP53* binding domains. Numbers in parentheses represent the number of cases with a given mutation; cases with multiple *TP53* mutations are represented multiple times in this plot. The *P* value represents the results of a log-rank test for the *TP53* mutation types excluding wild-type *TP53*. **(c)** HRs for a multivariate Cox model fit with 3p deletion and global deletion rate (CIN) across different case sets (age covariate not shown; error bars indicate 95% confidence intervals; *P* values represent the significance of a likelihood ratio test for a model fit with and without 3p deletion). **(d)** Venn diagram showing co-occurrence of *TP53* mutation (mut) and deletions on the 3p chromosome. **(e)** Kaplan-Meier curves showing survival outcome for all combinations of 3p deletion and *TP53* mutation events (colors correspond to the case subsets in **d**).

worse prognosis than cases with wild-type *TP53* (HR = 2.2 ± 0.7 ; *P* = 0.03). Among copy number alterations, the most significant association with survival was with heterozygous chromosomal deletions on the 3p arm, which also led to very poor prognosis (HR = 3.5 ± 1.1 ; Benjamini-Hochberg-corrected *P* = 0.002; **Fig. 1a**). Further analysis of chromosome 3p showed that many cases had a deletion spanning a large fraction of the arm, with increasing frequency of deletion approaching a fragile site in the 3p14.2 region¹¹ (**Supplementary Fig. 3**). Although general chromosomal instability (CIN), as well as deletion of many individual chromosomal regions, has previously been implicated as diagnostic^{1,12} and prognostic^{7,13–15} markers, we found that the 3p event in particular was responsible for the majority of the effect on survival in comparison with global rates of gene deletion (**Fig. 1c**).

TP53 and 3p events co-occur and their combination predicts worse clinical outcome

It has previously been shown that genetic alterations often act by redundant or synergistic mechanisms to confer a growth advantage in the tumor^{16,17}. Under the hypothesis that individual events might act in concert, we next examined the 82 prognostic events for pairwise association across the case cohort. This analysis identified 33 pairs of events that were significantly co-occurring or mutually exclusive (Bonferroni-corrected *P* < 0.01; **Supplementary Table 3**). Among these paired events, a particularly striking finding was that mutation of *TP53* and deletion of 3p occurred very frequently together, in 179 of 250 HPV-negative tumors (**Fig. 1d** and **Table 1**). Although mutation of *TP53* has previously been associated with chromosomal instability¹, we found that *TP53* mutation was associated with 3p loss far more frequently than it was associated with deletions in other chromosomal regions (**Supplementary Fig. 4** and **Supplementary Tables 4–6**). Moreover, the combination of *TP53* and 3p events led to worse survival than was predicted by either event independently or additively (**Supplementary Fig. 5**). Thus, the synergistic interaction between *TP53* and 3p events, with respect to both co-occurrence and



survival, supports a clear molecular stratification of HNSCC tumors with and without this combination of events (**Fig. 1c–e**, **Online Methods**, **Supplementary Fig. 5** and **Supplementary Table 7**).

We found that the *TP53*-3p combination of events was associated with advanced tumor stage, although stratification remained prognostic at all stages (**Supplementary Fig. 6**). Furthermore, the prognostic effect could not be explained by clinical covariates alone and was particularly strong for smokers under 75 years of age (175 cases, the majority of the TCGA cohort) for whom HR was 5.1 for the *TP53*-3p event relative to individuals without this combination (**Online Methods** and **Supplementary Fig. 7**).

To explore whether the interaction between *TP53* mutation and 3p deletion could be replicated in new cases, we obtained 126 additional HPV-negative HNSCC samples that had been deposited in TCGA while our initial study was underway (not included in the 15 January 2014 Firehose run). Although these new cases did not yet have sufficient clinical follow-up data for survival analysis, we indeed observed the same high rate of co-occurrence of *TP53* mutation and 3p deletion (**Table 1**).

Table 1 Co-occurrence and survival interaction of *TP53* and 3p events

| Cohort | <i>n</i> | Co-occurrence of <i>TP53</i> and 3p events | | Survival interaction for <i>TP53</i> -3p versus <i>TP53</i> events | |
|--------------------------|----------|--|---------------------|--|-----------------------|
| | | OR | <i>P</i> | HR ^a | <i>P</i> ^a |
| TCGA (discovery) | 250 | 6.6 | $1 \times 10^{-4*}$ | 5.6 | 0.001 |
| Recent TCGA (validation) | 126 | 10 | 1×10^{-6} | ND | ND |
| UPMC (validation) | 48 | 2.5 | 0.2 | 6.3 | 0.01 |
| Pan-Cancer (validation) | 4,404 | 2.0 | 1×10^{-25} | 1.4 | 0.002 |

The *P* value indicated by an asterisk was Bonferroni corrected for test space. OR, odds ratio; HR, hazard ratio; ND, not determined.

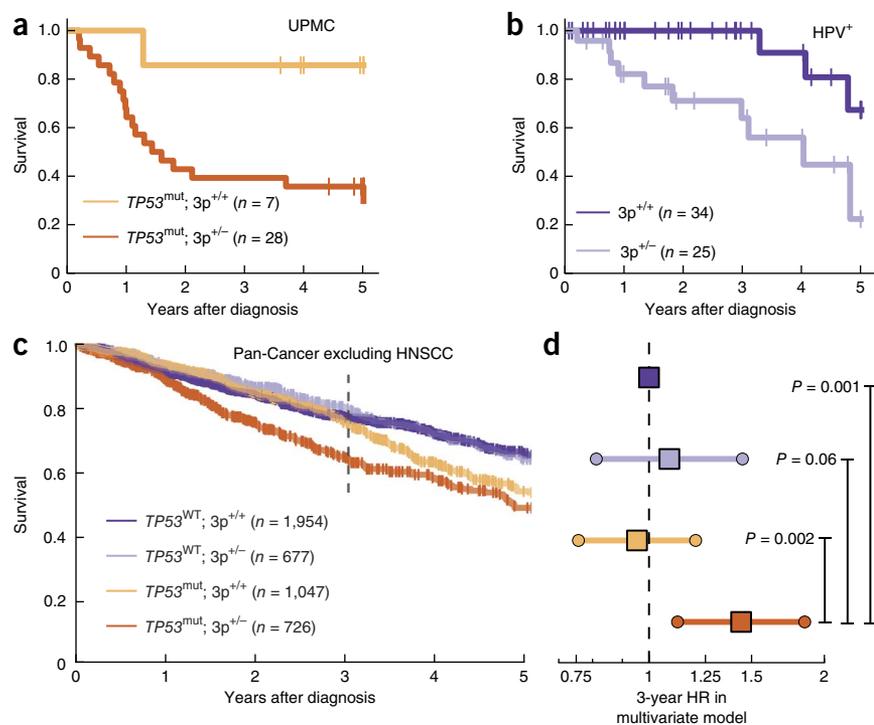
^aUnivariate model in cases under 75 years of age only.

Figure 2 Replication of *TP53*-3p association.

(a) Survival comparison of cases with a *TP53*-3p aggregate event with those with only *TP53* mutation in the independent UPMC cohort. (b) Loss of the 3p chromosomal arm is associated with lower survival rates in cases with HPV-positive tumors (TCGA and independent cohorts). (c) Assessment of 3p loss and *TP53* mutation association in the TCGA Pan-Cancer cohort (HNSCC excluded). (d) Corresponding HRs for a multivariate model of 3-year truncated survival (dashed line in c) when controlling for tissue type, age and stage covariates. Error bars indicate 95% confidence intervals.

We also analyzed clinical follow-up data for 48 HPV-negative HNSCC tumors from the University of Pittsburgh Medical Center³ for which exome sequencing and copy number profiles had previously been collected after surgery (UPMC cohort; **Supplementary Table 1**). We observed that, in this cohort, individuals whose tumors contained the *TP53*-3p aggregate event had substantially worse prognosis than individuals with *TP53* mutation alone, confirming the very large effect seen in the TCGA population (**Fig. 2a** and **Table 1**). *TP53* and 3p events also co-occurred in the UPMC cohort, although with a lower effect size than in the two TCGA cohorts (**Table 1**); we suspect that this difference in effect size is due to the much higher error rate of DNA sequencing in the earlier UPMC study, which resulted in false negative mutation calls (Online Methods).

We also sought evidence for the *TP53*-3p combination in cases with HPV-positive tumors, in which the *TP53* protein is inactivated via interaction with HPV viral proteins^{18,19}. Analysis of 59 HPV-positive



tumors from the TCGA and UPMC cohorts showed that *TP53* mutation was very rare in the presence of HPV (odds ratio (OR) = 0.01; $P = 1 \times 10^{-27}$ by Fisher's exact test), consistent with the expectation that the mutation confers little selective advantage once *TP53* is inactivated by HPV. Among the HPV-positive tumors, the 25 tumors with 3p deletion had significantly worse prognosis than the 34 without the 3p event (HR = 5.5 ± 2.6 ; $P = 0.004$). This finding lends further support for interaction between *TP53* and chromosome 3p with respect to survival and stratifies the growing population of affected individuals with HPV-positive tumors¹⁹ (**Fig. 2b**).

Another question was whether the *TP53*-3p interaction was specific to HNSCC or had broader support across diverse tissues. To address this issue, we performed a pan-cancer analysis based on all publicly available molecular data in TCGA (excluding HNSCC cases), covering 4,404 cases over an additional 17 cancer types²⁰ (Online Methods). Although these tissues were molecularly heterogeneous and presented with different patient outcomes (**Supplementary Fig. 8a-c**), we nonetheless found compelling evidence for both the co-occurrence and impact on survival of *TP53* mutation and 3p deletion in this broader cohort, even when tissue type, patient age and staging were accounted for (**Fig. 2c,d** and **Table 1**).

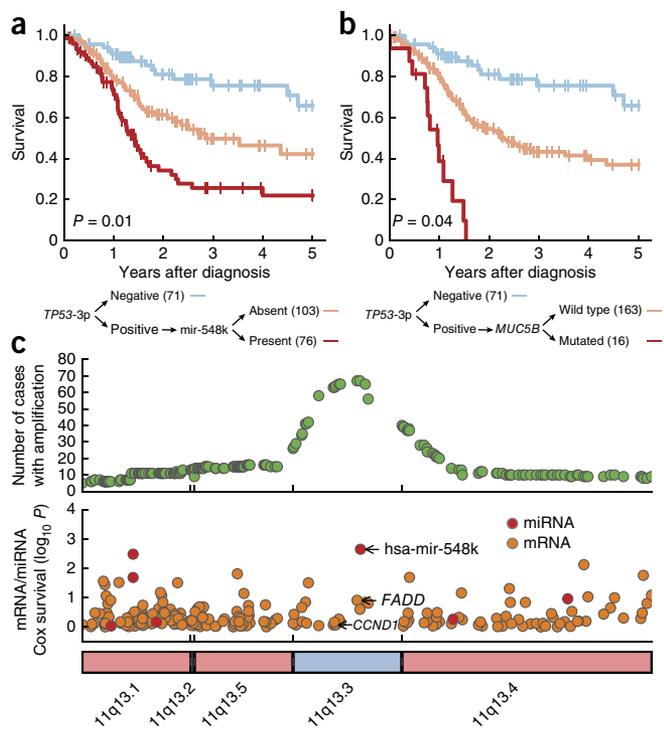


Figure 3 Characterization of molecular subtypes defined by the *TP53*-3p aggregate event. (a,b) Cases with the *TP53*-3p aggregate event can be further stratified by the presence of mir-548k (a) or *MUC5B* (b) events. (c) Frequency of high-gain amplification (top) and association with survival for gene or miRNA expression (bottom) along the 11q13 chromosomal segment. P values in a and b are Benjamini-Hochberg corrected for 1,008 events in a secondary prognostic biomarker screen (Online Methods). All survival associations are calculated by a likelihood ratio test with age and year of diagnosis used as covariates in the set of 179 cases with the *TP53*-3p event (*TP53*-3p-negative curves are shown for comparison but are not used in the computation).

Table 2 Mutual exclusivity of the TP53-3p aggregate event and gene mutations

| Cohort | n | TP53-3p event versus CASP8 mutation | | | TP53-3p event versus RAS signaling pathway ^a mutation | | |
|--------------------------|-----|--|-------|-----------------------|---|------|-----------------------|
| | | Number of cases mutated | OR | P | Number of cases mutated | OR | P |
| TCGA (discovery) | 250 | 21 | 0.13 | 3 × 10 ^{-3*} | 23 | 0.11 | 3 × 10 ^{-4*} |
| TP53-3p positive | 179 | 6 | | | 6 | | |
| TP53-3p negative | 71 | 15 | | | 17 | | |
| Recent TCGA (validation) | 126 | 20 | 0.038 | 7 × 10 ⁻⁸ | 21 | 0.86 | 5 × 10 ⁻⁶ |
| TP53-3p positive | 81 | 2 | | | 4 | | |
| TP53-3p negative | 45 | 18 | | | 17 | | |

P values indicated by an asterisk were Bonferroni corrected for a test space of 121 gene and pathway mutation events.
^aBioCarta SOS1-Mediated RAS Signaling Pathway (Reactome 524).

Characterization of subtypes defined by the combined TP53-3p event

Finally, we investigated whether the major subtypes defined by TP53 and 3p status (Fig. 1e) could be subdivided further by additional molecular markers (Online Methods). Indeed, we found that the 179 cases with the combined TP53-3p event were well stratified by the additional presence of the mir-548k miRNA (Fig. 3a and Supplementary Fig. 7c) or mutation of the MUC5B gene (Fig. 3b and Supplementary Fig. 7d), both of which were associated with worse prognosis. The gene encoding mir-548k is near CCND1 and FADD at 11q13.3, in a region that is commonly amplified in HNSCC¹⁴. Very recently, this miRNA has been shown to have oncogenic behavior in esophageal squamous cell carcinoma cell lines²¹. Although we found that 11q13.3 amplification was associated with survival to a lesser degree than mir-548k expression, the prognostic effect seemed to be specific to the expression of the miRNA (Fig. 3c and Supplementary Fig. 9).

Among cases lacking the TP53-3p event combination, we found strong enrichment for mutations affecting caspase-8 as well as RAS and components of RAS signaling (Table 2 and Supplementary Fig. 1b). These enrichments were replicated in the TCGA molecular validation cohort (Table 2). The mutual exclusivity of caspase-8 or RAS events with TP53-3p events provides further support for a TP53-3p-defined subtype, and it implicates alternative routes to tumor progression in the absence of a TP53-3p event.

DISCUSSION

As we approach a full inventory of driver events in cancer²², a key next step is to map and decode the complex network of interactions among individual events. Here such an analysis was performed to identify a definitive stratification of head and neck cancer on the basis of the largest tissue bank and data set in existence. We have shown that TP53 mutation, a well-studied driver event that leads to poor survival rates, is nearly always accompanied by specific loss of chromosome 3p (Fig. 1d and Table 1). As has been argued for other cancer-related mutations^{17,23}, the frequent co-occurrence of TP53 and 3p alterations implies a selective advantage of cells acquiring both genomic events. In this study, detection of the TP53-3p interaction was possible owing to the high prevalence of each event individually and the high (marginal) associations with survival rates.

Although our study focused almost entirely on a single compelling interaction, our full analysis uncovered 32 additional interactions in HNSCC that remain to be investigated (Supplementary Table 3). It is likely that this number is an underestimate, as low-frequency and/or non-prognostic events were not evaluated. As cancer cohorts become larger, analyses such as this will become more powered,

creating the opportunity to reevaluate the cancer landscape from the perspective of pairwise and, ultimately, higher-order interactions among events.

Our analysis identifies two distinct clinical and molecular paths to cancer in individuals with HPV-negative HNSCC. The first group of cases, characterized by TP53 mutation and loss of the 3p chromosome, is associated with advanced clinical stage and common risk factors such as smoking. This group tends to have very poor outcomes, even when evaluated independently of these risk factors (Supplementary Fig. 7). The second group of cases, lacking the TP53-3p combination

of events, is characterized by mutations affecting RAS signaling and caspase-8 (Table 2) and, ultimately, less aggressive tumors.

Further study is clearly warranted to elucidate the molecular underpinnings of these two groups of cases, with the goal of using such molecular stratification alongside clinical variables to inform patient treatment. Open questions relate to mechanism and the ordering of TP53 and 3p events. For example, what is the factor or factors encoded on chromosome 3p that are responsible for the interaction with TP53? Does one event necessarily precede the other, and is a particular order required for poor survival? It is plausible that genomic instability primed by TP53 mutation gives rise to the loss of the activity of a key factor encoded on chromosome 3p, but other scenarios are possible. Regardless, because the interaction of 3p with TP53 or HPV status is independent of tumor stage, treatment of patients with HNSCC might be modified to coincide with this specific molecular classification. In HPV-negative HNSCC, the need for patient-tailored treatment programs is especially great, as in the current era toxicity of existing regimens has been maximized without necessarily improving outcome in cancers.

Our results also underscore the importance and value of public efforts such as TCGA in gathering, organizing and distributing genomic data. Our work builds on the exemplary TCGA data collection and analysis pipeline²⁰ to integrate data across different measurement platforms, with the goal of finding higher-order interactions of molecular events. Following the example of TCGA, we have documented and made public all analyses conducted in this study, ranging from data download to processing, exploratory analyses, statistical modeling and visualization (Online Methods). With such a large and complex data set, transparency and reproducibility of analysis are essential to provide a clear understanding of the methodology and to allow for further mining of the results and extension to new data sets.

URLs. Study source code and analysis notebook repository, <https://github.com/theandygross/TCGA/>; Broad Firehose, <https://confluence.broadinstitute.org/display/GDAC/Home>; TCGA Data Portal, <https://tcga-data.nci.nih.gov/tcga>; UCSC Cancer Genomics Hub, <https://cghub.ucsc.edu/>.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The results published here are based upon data generated by the TCGA Research Network. We thank K. Messer and A. Tward for helpful discussions. We gratefully acknowledge support for this study from the US National Institutes of Health (P50 GM085764, U24 CA184427 and P41 GM103504 to T.I.; T32 DC000028 to R.K.O.; Burroughs Wellcome Fund CAMS to Q.T.N.; P50 CA097190 and The American Cancer Society to J.R.G.; K07 CA137140 to A.M.E.; DP5 OD017937-01 to H.C.). J.P.S. is supported in part by grants from the Marsha Rivkin Center for Ovarian Cancer Research and a Conquer Cancer Foundation of the American Society of Clinical Oncology Young Investigator Award.

AUTHOR CONTRIBUTIONS

A.M.G., R.K.O., Q.T.N. and T.I. conceived the study. A.M.G. carried out most analyses. R.K.O., J.P.S., M.C., C.S.C., E.E.C., S.M.L., Q.T.N. and D.N.H. provided expertise. M.H. and H.C. aided in bioinformatics analysis. A.M.E. and J.R.G. collected and compiled clinical follow-up data for the UPMC cohort. A.M.G. and T.I. wrote the manuscript with assistance from other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

1. Leemans, C.R., Braakhuis, B.J.M. & Brakenhoff, R.H. The molecular biology of head and neck cancer. *Nat. Rev. Cancer* **11**, 9–22 (2011).
2. Mroz, E.A. *et al.* High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma: Genetic Heterogeneity and HNSCC Outcome. *Cancer* **119**, 3034–3042 (2013).
3. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
4. Chung, C.H. *et al.* Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell* **5**, 489–500 (2004).
5. Walter, V. *et al.* Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. *PLoS ONE* **8**, e56823 (2013).
6. Pickering, C.R. *et al.* Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discov.* **3**, 770–781 (2013).
7. Temam, S. *et al.* Epidermal growth factor receptor copy number alterations correlate with poor clinical outcome in patients with head and neck squamous cancer. *J. Clin. Oncol.* **25**, 2164–2170 (2007).
8. Lui, V.W.Y. *et al.* Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. *Cancer Discov.* **3**, 761–769 (2013).
9. Perez, F. & Granger, B.E. IPython: a system for interactive scientific computing. *Comput. Sci. Eng.* **9**, 21–29 (2007).
10. Poeta, M.L. *et al.* TP53 mutations and survival in squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* **357**, 2552–2561 (2007).
11. Ohta, M. *et al.* The FHIT gene, spanning the chromosome 3p14.2 fragile site and renal carcinoma-associated t(3;8) breakpoint, is abnormal in digestive tract cancers. *Cell* **84**, 587–597 (1996).
12. Rosin, M.P. *et al.* Use of allelic loss to predict malignant risk for low-grade oral epithelial dysplasia. *Clin. Cancer Res.* **6**, 357–362 (2000).
13. Partridge, M., Emilion, G. & Langdon, J.D. LOH at 3p correlates with a poor survival in oral squamous cell carcinoma. *Br. J. Cancer* **73**, 366–371 (1996).
14. Meredith, S.D. *et al.* Chromosome 11q13 amplification in head and neck squamous cell carcinoma. Association with poor prognosis. *Arch. Otolaryngol. Head Neck Surg.* **121**, 790–794 (1995).
15. Partridge, M. *et al.* The prognostic significance of allelic imbalance at key chromosomal loci in oral cancer. *Br. J. Cancer* **79**, 1821–1827 (1999).
16. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).
17. Bredel, M. *et al.* A network model of a cooperative genetic landscape in brain tumors. *J. Am. Med. Assoc.* **302**, 261–275 (2009).
18. Thomas, M., Pim, D. & Banks, L. The role of the E6-p53 interaction in the molecular pathogenesis of HPV. *Oncogene* **18**, 7690–7700 (1999).
19. Marur, S., D'Souza, G., Westra, W.H. & Forastiere, A.A. HPV-associated head and neck cancer: a virus-related cancer epidemic. *Lancet Oncol.* **11**, 781–789 (2010).
20. Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
21. Song, Y. *et al.* Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* **509**, 91–95 (2014).
22. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
23. Xing, F. *et al.* Concurrent loss of the PTEN and RB1 tumor suppressors attenuates RAF dependence in melanomas harboring ^{V600E}BRAF. *Oncogene* **31**, 446–457 (2012).

ONLINE METHODS

Availability. All data retrieval and processing steps are documented in a series of IPython notebooks⁹ available along with source code online (see URLs). These notebooks provide fully executable instructions for the reproduction of the analyses and the generation of figures and statistics for this study.

Molecular data. Data were obtained from the TCGA Genome Data Analysis Center (GDAC) Firehose website (see URLs) using the `firehose_get` data-retrieval utility. All data were downloaded from the 15 January 2014 standard data and analyses runs unless otherwise specified. To maintain the coherency of the analysis across different data layers and cancer types, we used Level 3 normalized molecular data as the input to our analysis. The use of the GDAC pipeline is intended to make these results easy to update as more TCGA data become available.

For a number of pan-cancer samples, we generated mutation calls from TCGA aligned BAM files obtained from the UCSC Cancer Genomics Hub (see URLs). These calls were only used for cases with sequenced exome data that had yet to go through the Firehose processing pipeline. Somatic mutation calls were made by running the MuTect mutation calling program²⁴ and the Genome Analysis Toolkit (GATK) SomaticIndelDetector²⁵ function on targeted regions with default parameters. All steps for downloading and processing these data are documented in the analysis notebooks and accompanying software repository. All mutation calls generated for this analysis are included as **Supplementary Table 8**. Although these calls have yet to go through manual curation, we benchmarked this pipeline against TCGA working group mutation calls and found very high overlap with 94% sensitivity and 96% specificity.

Pathway data. Pathway data were downloaded from the Molecular Signatures Database²⁶ (mSigDB). Version 3 of the canonical pathway gene sets was used for this analysis.

Candidate biomarker construction. Mutation calls were extracted from the annotated MAF files obtained from Firehose and filtered to include only non-silent mutations. Each case was associated with a binary vector in which each position represented a gene; the position was set to 1 if the gene was observed to harbor one or more mutations in the case and was set to 0 otherwise. Mutation meta-markers were constructed by collapsing the genes within a pathway gene set via a Boolean or operator, such that the pathway was considered altered in a case if any of its genes had a mutation (**Supplementary Fig. 1b,c**). Pathway markers that were characterized by a single highly mutated gene or were highly correlated with mutation rate (Mann-Whitney U test $P < 0.01$) were filtered out.

Copy number aberrations were extracted from the GISTIC2 (ref. 27) processing pipeline included in the standard Firehose analysis run. For biomarker construction, data aggregated on significantly altered lesions (deemed significant at the default 99% confidence settings) were used.

mRNA and miRNA expression data were obtained from the Level 3 normalized gene-by-case matrices generated as part of the Firehose analysis pipeline. Data were \log_2 transformed. Genes and miRNAs were first filtered on the basis of differential expression, comparing the full set of tumor expression profiles with the 34 profiles available for matched normal tissue (t -test cutoff at $P < 0.01$). A pattern of background expression was estimated by taking the first principal component of non-differentially expressed genes or miRNAs. This background signal is meant to approximate the most common non-tumor-related variation in expression due to inherent properties of the cohort such as population substructure or tissue-specific changes in expression. Real-valued features with high correlation (Pearson's correlation $P < 1 \times 10^{-5}$) to this background expression pattern were filtered out. For the survival analysis, only the top 300 (of a possible 20,502) differentially expressed genes were included in the analysis to limit the burden of multiple-hypothesis correction (all 251 differentially expressed miRNAs were used).

Markers used in this analysis consisted of binary markers and continuous-valued markers. Binary markers were used when expression was only present (having more than 0.5 reads per million) in a moderate fraction of the cohort (between 20 cases and half of the cohort). Real-valued gene and miRNA expression levels were used for differentially expressed features not

assigned as binary markers. Gene expression meta-markers were constructed from the loading of the first principal component of the reduced gene-by-case matrix defined by each gene set. Because similarity of gene sets resulted in redundant gene expression meta-markers, marker pairs with high correlation (Spearman's $\rho > 0.7$) were reduced to a single informative marker by choosing the marker with the greatest differential expression. For the survival analysis, continuous-valued markers were transformed into binary events before testing by setting a threshold that minimized the difference in variance between the resulting two groups. This was used to capture the skew of the distribution and assign the cases on the tail of the expression distribution as having an expression event (**Supplementary Fig. 1e**).

Clinical data. Clinical data were downloaded directly from the TCGA Data Portal (see URLs). All outcomes reported relate to all-cause survival. Survival times were censored after 5 years to reduce the confounding effect of patient age. For **Figure 2d**, survival times were censored after 3 years to show the specific effect within this time window, but all other figures and all statistics cited in the paper used 5-year survival. Although data on comorbidity were limited for this cohort, from other studies, we could estimate the competing mortality within this time frame to be about 20% (refs. 28,29). We expect the actual effect of such confounding to be minimal, as separation in the survival curves that we observed generally occurred within the first 2 years, during which time we expect non-cancer-associated death rates to be much lower.

For the primary and secondary survival screens, clinical data with missing values were used, but statistics were only calculated on cases with data reported. In multivariate analysis (**Supplementary Fig. 7**), missing value indicators were used.

HPV status. HPV calls from sequencing data were obtained from the TCGA HNSCC analysis working group. Owing to the incompleteness of this data set, this information was supplemented with HPV status called from a PCR-based MassARRAY Assay diagnostic provided on the TCGA data portal for cases where sequence-based data were not available.

Prioritization of prognostic events. Feature selection was performed before prognostic event prioritization. Events were selected for which at least 5% of cases were assigned to each group.

Prognostic events (**Fig. 1a**) were prioritized via a likelihood ratio test comparing a Cox proportional hazards model³⁰ fit with a candidate biomarker and covariates against a null model fit with the covariates alone. Age and the binary variable of patient age > 75 were used as covariates (both age variables were used to model a nonlinear association of patient age with survival). A correction for multiple-hypothesis testing was employed that uses the method of Benjamini and Hochberg³¹ to control for the false discovery rate across the entire pooled space of tested features. After multivariate testing, a univariate log-rank test was assessed for each event, and features with high multivariate significance but low univariate significance ($P < 0.05$) were filtered from the pool of prognostic events.

As discussed in the text and in **Figure 3**, we conducted a second prognostic screen within the 179 cases with the *TP53-3p* aggregate event. For this analysis, feature construction was repeated, resulting in 1,008 candidate biomarkers (note that this number was higher than in the primary screen owing to more events passing the 5% threshold). During this secondary screen, we found that the year of diagnosis for a patient had a large impact on outcomes. For this reason, we included this variable as a covariate in this screen.

Statistical analysis of *TP53-3p* interaction on survival. To assess the role of an interaction term in a statistical model of patient outcomes, we performed leave-one-out cross-validation on a logistic regression model as shown in **Supplementary Figure 5**. To convert the survival data into a binary classification problem, we organized patients into two classes depending on whether they were surviving or deceased at t years after surgery. In this analysis, the ratio of deceased to surviving patients was artificially high owing to the ability to observe a death in a shorter follow-up period than the full time interval required to annotate a patient as surviving (i.e., the basis of the Cox censorship problem). To reduce this bias, we removed patients with an observed death but a time of surgery after a set year ($2013 - (t - 1)$). As the problem

was often unbalanced (the number of surviving patients differed from the number of deceased patients), reweighting was performed to give both classes equal weight. A multivariate Cox model fit to the most significant model is also shown in **Supplementary Table 7**.

University of Pittsburgh Medical Center cohort. The chromosomal status at 3p was estimated via the median copy number of the 12 genes on the 3p14.2 locus. Matched exome and copy number data were available for 48 of 63 cases with HPV-negative tumors. In preliminary analysis, we found the UPMC cohort to have a significantly lower overall mutation rate than the TCGA cohort, with a median of 73 mutations per case as compared to 104 mutations per case in TCGA (Mann-Whitney U test $P < 0.001$). This difference can likely be attributed to lower depth of coverage and/or less sophisticated variant calling techniques, as the UPMC study was one of the first large whole-exome molecular cohorts and predates TCGA data collection by about 2 years.

Pan-cancer analysis. Pan-cancer data were downloaded and processed in the same manner as the HNSCC cohort. The chromosomal status at 3p was estimated via the median copy number of the 12 genes on the 3p14.2 locus.

To limit the heterogeneity of the pan-cancer cohort such that differences in molecular characteristics could be assessed, we performed a number of preprocessing steps. This processing reduced the cohort from 7,081 to 4,404 cases appropriate for survival analysis through the following filters:

- (1) Only primary tumors were used for all cases; metastatic tumors were discarded.
- (2) Glioblastoma cases were excluded owing to the extremely low survival rate (5-year survival of 6%).

- (3) Diffuse large B cell lymphoma, kidney chromophobe, thyroid carcinoma and prostate adenocarcinoma cases were removed owing to extremely high rates of survival in the cohorts (84%, 86%, 90% and 96% for 5-year survival, respectively).
- (4) Adrenocortical carcinoma, esophageal carcinoma and pancreatic adenocarcinoma were excluded owing to low sample counts (14, 39 and 69 cases in each tissue, respectively).
- (5) Cases older than 85 years of age were excluded from the analysis to limit confounding from age (115 cases; HR = 2.2 ± 3).
- (6) Cases with high levels of residual tumor were excluded (66 cases; HR = 2.9 ± 0.5).
- (7) Stage IV cases were excluded (612 cases; HR = 2.0 ± 0.1).
- (8) To limit circularity, HNSCC cases were excluded from all pan-cancer calculations but remain in **Supplementary Figure 8** to allow for comparison to other tissue types.

24. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
25. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
26. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
27. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
28. Mell, L.K. *et al.* Predictors of competing mortality in advanced head and neck cancer. *J. Clin. Oncol.* **28**, 15–20 (2010).
29. Farshadpour, F. *et al.* Survival analysis of head and neck squamous cell carcinoma: influence of smoking and drinking. *Head Neck* **33**, 817–823 (2011).
30. Cox, D.R. *Analysis of Survival Data* (Chapman and Hall, London, New York, 1984).
31. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).