DATA NOTE

Whole-genome sequencing of nine esophageal adenocarcinoma cell lines [version 1; referees: 3 approved]

Gianmarco Contino¹, Matthew D. Eldridge², Maria Secrier², Lawrence Bower², Rachael Fels Elliott¹, Jamie Weaver¹, Andy G. Lynch², Paul A.W. Edwards³, Rebecca C. Fitzgerald¹

¹Medical Research Council (MRC) Cancer Unit, University of Cambridge, Cambridge, UK
²Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
³Department of Pathology, University of Cambridge, Cambridge, UK

Abstract

Esophageal adenocarcinoma (EAC) is highly mutated and molecularly heterogeneous. The number of cell lines available for study is limited and their genome has been only partially characterized. The availability of an accurate annotation of their mutational landscape is crucial for accurate experimental design and correct interpretation of genotype-phenotype findings. We performed high coverage, paired end whole genome sequencing on eight EAC cell lines—ESO26, ESO51, FLO-1, JH-EsoAd1, OACM5.1 C, OACP4 C, OE33, SK-GT-4—all verified against original patient material, and one esophageal high grade dysplasia cell line, CP-D. We have made available the aligned sequence data and report single nucleotide variants (SNVs), small insertions and deletions (indels), and copy number alterations, identified by comparison with the human reference genome and known single nucleotide polymorphisms (SNPs). We compare these putative mutations to mutations found in primary tissue EAC samples, to inform the use of these cell lines as a model of EAC.

This article is included in the Data: Use and Reuse collection.
Corresponding authors: Gianmarco Contino (gc502@mrc-cu.cam.ac.uk), Rebecca C. Fitzgerald (RCF29@MRC-CU.cam.ac.uk)

Competing interests: No competing interests were disclosed.

How to cite this article: Contino G, Eldridge MD, Secrèr M et al. Whole-genome sequencing of nine esophageal adenocarcinoma cell lines [version 1; referees: 3 approved] F1000Research 2016, 5:1336 (doi: 10.12688/f1000research.7033.1)

Copyright: © 2016 Contino G et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: This work was funded by an MRC Programme Grant to R.C.F. and a Cancer Research UK grant to PAWE. The pipeline for mutation calling is funded by Cancer Research UK as part of the International Cancer Genome Consortium. G.C. is a National Institute for Health Research Lecturer as part of a NIHR professorship grant to R.C.F. AGL is supported by a Cancer Research UK programme grant (C14303/A20406) to Simon Tavare and the European Commission through the Horizon 2020 project SOUND (Grant Agreement no. 633974).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

Esophageal adenocarcinoma (EAC), including cancers of the gastro-esophageal junction, represent a substantial health concern in Western countries due to its increasing incidence and poor prognosis. To date, there are no widely accepted animal models for EAC and a limited number of cell lines are all that are available for in vitro functional studies. Recent genome-wide sequencing projects have shown that EAC is one of the most highly mutated solid cancers with a high degree of heterogeneity (Dulak et al., 2013; Weaver et al., 2014). In addition to point mutations there are also widespread copy number alterations with evidence of catastrophic events such as chromothripsis and bridge fusion breakages in about one-third of cases (Nones et al., 2014). An accurate annotation of the mutational landscape of available EAC cell lines is therefore crucial for optimal experimental design, interpretation of genotype-phenotype data and to analyse drug sensitivities. We selected eight EAC cell lines—ESO26, ESO51, FLO-1, JH-EsoAd1, OACM5.1 C, OACP4 C, OE33, SK-GT—4—the identities of which have been verified by short tandem repeat (STR) analysis, p53 mutation and xenograft histology against the original tumors (Boonstra et al., 2010), and one esophageal high grade dysplasia (CP-D) cell line. We performed high-coverage paired-end whole genome sequencing and aligned the sequence data to the human reference genome in order to detect single nucleotide variants, indels and copy number alterations.

Materials and methods

Ethics

Cell lines

All cell lines were from a certified source (Table 1) and verified in house for >90% match with publicly reported STR profiles. Cell lines were mycoplasma tested and grown in standard conditions reported in cell repositories indicated in Table 1. Matched germline DNA was not available.

Library preparation, sequencing and QC

Genomic DNA was prepared from cultured cells with AllPrep-DNA/RNA Mini Kit (Qiagen) according to manufacturer’s instructions. A single library was created for each sample, and 90-bp paired-end sequencing was performed at Beijing Genomic Institute (BGI, Guangdong, China) according to Illumina (CA, USA) instructions to a typical depth of 30x, with 94% of the known genome being sequenced to at least 10x coverage and achieving a Phred quality of 30 for at least 80% of mapping bases. FastQC 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to assess the quality of the sequence data. Additional alignment, duplication and insert size metrics quality metrics are reported in Supplementary material 7. Sequence reads were mapped to the human reference genome (Ensembl GRCh37, release 84) using BWA 0.5.9 (Li, 2009), sorted into genome coordinate order and duplicates marked using Picard 1.105 (FixMateInFormation and MarkDuplicates tools respectively, http://broadinstitute.github.io/picard/). Original BAM files are available in the European Bioinformatics Institute (EBI) repository (project: PRJEB14018; sample accessions: ERS1158075-ERS1158083).

Mutation calling

GATK v3.2.2 (Broad Institute, MA, USA) was used to call and filter single nucleotide and indel variants compared to the reference genome. In brief, the steps run were as follows: 1) local realignment of reads to correct misalignments around indels using GATK RealignerTargetCreator and IndelRealigner tools; 2) recalibration of base quality scores using GATK BaseRecalibrator tool; 3) SNP and indel calling using GATK HaplotypeCaller which determines haplotype by re-assembly within regions determined to be active, i.e. where there is evidence for a variation, and uses a Bayesian approach to assign genotypes. Hard filters were applied to the resulting call set using recommendations available from the GATK documentation (https://www.broadinstitute.org/gatk) to generate a high-confidence set of SNV and indel calls. These were analyzed with Ensembl Variant Effect Predictor (release 75, http://www.ensembl.org/info/docs/tools/vep/index.html) to annotate with genomic features and consequences of protein coding regions (Supplementary material 4). For the purposes of the analysis, all variants with global minor allele frequency (GMAF) >0.0014 described in the 1000 Genomes project were separated out as likely germline polymorphisms (The 1000 Genomes Project Consortium et al., 2012) according to the criteria adopted in the Cosmic Cell Lines Project (Wellcome Trust Sanger Institute, Cambridge). Further, we removed all SNPs that have a minor allele frequency in the DBSNP (Ensembl v.58) and variants with a frequency ≥0.00025 in the ESP6500 (NHLBI GO Exome Sequencing Project, released June 20th 2012). A full list of the filtered variants is available in Supplementary material 4 and Supplementary material 6.

Copy number assessment

Copy number (CN) analysis was carried out using Control-FREEC (Boeza et al., 2012). Control-FREEC computes and segments CN profiles and is capable of characterizing over-diploid genomes, taking into consideration the CG-content and mapability profiles to normalize read count in the absence of a control sample. Ploidy in each cell line was assessed interactively with the Crambed app v.2.0 according to the methods described by Lynch (2015).

Dataset validation

Whole genome sequencing

We identified a median of 1.3 × 10⁹ variants across all 9 cell lines (range 105,487–151,879: Figure 1a, Table 2, Supplementary material 3, Supplementary material 4). We found that 1.5% of the variants were in coding regions; additionally, 4% fell in surrounding gene regions (i.e. regulatory as defined in Zerbino et al. (2015), upstream and downstream regions), 41% in introns and 23% in intergenic regions. Among the variants in the coding sequence, the majority, 57.4%, were in the UTR regions, followed by exonic missense and synonymous variants (21% and 11% respectively (Figure 1, Table 2, Supplementary material 3, Supplementary material 4). The number of variations identified in the high-grade dysplasia CP-D line was not significantly lower to the median of other EAC cell lines, consistent with the finding that such pre-malignant lesions have already accumulated many SNVs (Weaver et al., 2014).
Table 1. Characteristics and clinico-pathological features of the EAC cell lines. Verified origin identifies cell lines whose pathological origin from EAC has been verified in Boostra et al., 2010.

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Alternative Names</th>
<th>Age</th>
<th>Sex</th>
<th>Ethnicity</th>
<th>Histology</th>
<th>Date Derived</th>
<th>Stage</th>
<th>Ploidy</th>
<th>Commercial Availability</th>
<th>Verified origin</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP-D</td>
<td>CP-18821</td>
<td>Adult</td>
<td>M</td>
<td>Caucasian</td>
<td>hTERT immortalized oesophageal HGD</td>
<td>1995</td>
<td>HGD</td>
<td>hypohypotetraploid</td>
<td>ATCC</td>
<td>Palanca-Wessels et al., 1998</td>
<td></td>
</tr>
<tr>
<td>ESO26</td>
<td>56</td>
<td>M</td>
<td>Caucasian</td>
<td>GOJ adenocarcinoma</td>
<td>2000</td>
<td>Stage IV</td>
<td>hypodiploid (1.8)</td>
<td>Public Health England –Culture Collection</td>
<td>YES</td>
<td>Boostra et al., 2010</td>
<td></td>
</tr>
<tr>
<td>ESOS51</td>
<td>74</td>
<td>M</td>
<td>Caucasian</td>
<td>Distal Oesophageal Adenocarcinoma</td>
<td>2000</td>
<td>Stage IV</td>
<td>hypotripleoid (2.75)</td>
<td>Public Health England –Culture Collection</td>
<td>YES</td>
<td>Boostra et al., 2010</td>
<td></td>
</tr>
<tr>
<td>FLO-1</td>
<td>68</td>
<td>M</td>
<td>Caucasian</td>
<td>Distal Oesophageal Adenocarcinoma</td>
<td>1991</td>
<td>hypodiploid (1.9)</td>
<td>Public Health England –Culture Collection</td>
<td>YES</td>
<td>Hughes et al., 1997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JH-EsoAd1</td>
<td>JHAD1</td>
<td>66</td>
<td>M</td>
<td>Caucasian</td>
<td>Moderately to poorly differentiated Oesophageal Adenocarcinoma</td>
<td>1997</td>
<td>Stage IIA (T3 N0 M0)</td>
<td>triploid</td>
<td>No, due to be deposited in ATCC</td>
<td>YES</td>
<td>Alvarez et al., 2008</td>
</tr>
<tr>
<td>OACM5.1 C</td>
<td>47</td>
<td>F</td>
<td>Caucasian</td>
<td>Lymph node metastases of Distal Oesophageal Adenocarcinoma</td>
<td>2001</td>
<td>Stage IV</td>
<td>hypodiploid</td>
<td>Public Health England –Culture Collection</td>
<td>YES</td>
<td>de Both et al., 2001</td>
<td></td>
</tr>
<tr>
<td>OACP4 C</td>
<td>55</td>
<td>M</td>
<td>Caucasian</td>
<td>Gastric cardia adenocarcinoma</td>
<td>2001</td>
<td>Stage IV</td>
<td>Aneuploidy (53–57 chromosomes)</td>
<td>Public Health England –Culture Collection</td>
<td>YES</td>
<td>de Both et al., 2001</td>
<td></td>
</tr>
<tr>
<td>OE33</td>
<td>JROECL33</td>
<td>73</td>
<td>F</td>
<td>Caucasian</td>
<td>Distal Oesophageal Adenocarcinoma</td>
<td>1993</td>
<td>Stage IIA</td>
<td>hypotetraploid (3.5)</td>
<td>Public Health England –Culture Collection</td>
<td>YES</td>
<td>Rockett et al., 1997</td>
</tr>
<tr>
<td>SK-GT-4</td>
<td>83</td>
<td>M</td>
<td>Caucasian</td>
<td>Distal Oesophageal Adenocarcinoma</td>
<td>1989</td>
<td>Stage IIB</td>
<td>Aneuploidy (mode 59 chromosomes, SK)</td>
<td>Public Health England –Culture Collection</td>
<td>YES</td>
<td>Altorki et al., 1993</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. Distribution of detected variants and coding sequence consequences (mean percentage value). A) Bar chart showing the distribution of called variants across various regions of the genome as indicated; B) Details of the coding sequence variants identified by the Variant Effect Predictor (Ensembl) expressed as a mean percentage value of all cell lines (values were not statistically different among samples).

OACP4C and ESO26 showed the smallest and largest number of variants, respectively. (Figure 1, Table 2).

A limitation of this study is represented by the lack of an available normal counterpart. In order to overcome this problem, in addition to the GATK calling pipeline we have applied a series of filters according to the criteria reported in methods and derived the 1000 Genomes Project (The 1000 Genomes Project Consortium et al., 2012), DBSNP (Ensembl v.58) and ESP6500 (released June 20\textsuperscript{th} 2012). This approach reduced the number of variants by an order of magnitude from the original GATK pipeline (from a median of 4.1×10\textsuperscript{6} to 1.3×10\textsuperscript{5}). Yet, the abundance of called variants compared to a range of 4,8×10\textsuperscript{3}-6×10\textsuperscript{4} reported in human EAC (Weaver et al., 2014), may indicate that a proportion of the variants called in our final annotation are of germline origin. Also, additional mutations may have accumulated in vitro. A comprehensive annotation of the coding sequence variants identified is reported in Supplementary material 3 and Supplementary material 4.

Analysis of putative EAC driver genes

In order to investigate how closely cell lines reflect the spectrum of mutations observed in human specimens we analysed the mutational landscape of known cancer and putative EAC driver genes and compared to the previously reported mutation rate (Dulak et al., 2013; Weaver et al., 2014; Figure 2b & 2c). 69% of EACs have TP53 mutations (Weaver et al., 2014), while all cell lines carried at least one deleterious TP53 mutation. A SMAD4 mutation was present in 2 of 9 cell lines, ESO26 and JH-EsoAd, consistent with the 13% observed in EAC (Weaver et al., 2014). We were not able to identify mutations in ARID1A (affected by UTR variants in 1 of 9 cell lines) that is reportedly mutated in about 10% of cases of EAC specimens. Only some of the missense variants in the genes shown in Figure 2b resulted in known pathogenic mutations (i.e. TP53, PIK3CA, and TLR4). Other genes harboured benign or likely benign variants and/or variants with uncertain functional significance.

We expanded our analysis to other cancer genes of potential relevance to OAC. We identified a pathogenic KRAS mutation in SKGT4, and a missense mutation of uncertain significance in MET (OE33), EGFR (CP-D, ESO26, JH-EsoAd1). Among DNA repair genes all cell lines carry benign missense variants of ATM and missense variants of uncertain significance in BRCA2. MSH2 is affected by a missense variant in SKGT4, splice site variants in CP-D, JH-EsoAd1, and UTR variants in ESO51 and OACP4C (Supplementary material 3, Supplementary material 4, Supplementary material 6). Copy number analysis (Supplementary material 1, Supplementary material 2) identified recurrent amplifications in ERBB2, MYC, MET and SEMA5A, and deletions in SMAD4, CDKN2A, CCDC102B and SMARCA4.
Table 2. Detailed distribution of identified variants for each cell lines. Absolute number, median, median absolute deviation and range interval are listed for each category of mutation according to Variant Effect Predictor classification (Ensembl).

<table>
<thead>
<tr>
<th>Coding variants (type)</th>
<th>CP-D</th>
<th>ESO26</th>
<th>ESO51</th>
<th>FLO-1</th>
<th>JH-EsoAD1</th>
<th>OACM5.1</th>
<th>OACP4C</th>
<th>OE33</th>
<th>SK-GT-4</th>
<th>Median</th>
<th>Median Absolute Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTR</td>
<td>5 prime UTR</td>
<td>229</td>
<td>301</td>
<td>262</td>
<td>191</td>
<td>206</td>
<td>264</td>
<td>229</td>
<td>216</td>
<td>305</td>
<td>229</td>
<td>33</td>
<td>191 305</td>
</tr>
<tr>
<td>3 prime UTR</td>
<td>979</td>
<td>1097</td>
<td>1002</td>
<td>926</td>
<td>929</td>
<td>1026</td>
<td>848</td>
<td>986</td>
<td>1113</td>
<td>996</td>
<td>57</td>
<td>848</td>
<td>1113</td>
</tr>
<tr>
<td>Start/Stop</td>
<td>initiator codon</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>stop lost</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>stop retained</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>stop gained</td>
<td>10</td>
<td>14</td>
<td>17</td>
<td>16</td>
<td>14</td>
<td>17</td>
<td>9</td>
<td>14</td>
<td>24</td>
<td>14</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>Missense</td>
<td>missense</td>
<td>385</td>
<td>496</td>
<td>497</td>
<td>436</td>
<td>435</td>
<td>481</td>
<td>431</td>
<td>446</td>
<td>454</td>
<td>446</td>
<td>15</td>
<td>385 497</td>
</tr>
<tr>
<td>Splice Sites</td>
<td>splice acceptor</td>
<td>4</td>
<td>11</td>
<td>7</td>
<td>8</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>4   11</td>
</tr>
<tr>
<td></td>
<td>splice donor</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>10</td>
<td>6</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>18</td>
<td>6</td>
<td>1</td>
<td>5   18</td>
</tr>
<tr>
<td></td>
<td>splice region</td>
<td>105</td>
<td>113</td>
<td>107</td>
<td>92</td>
<td>96</td>
<td>95</td>
<td>83</td>
<td>103</td>
<td>102</td>
<td>102</td>
<td>6</td>
<td>83   113</td>
</tr>
<tr>
<td>Frameshift INDEL</td>
<td>frameshift</td>
<td>42</td>
<td>52</td>
<td>41</td>
<td>45</td>
<td>34</td>
<td>34</td>
<td>49</td>
<td>46</td>
<td>54</td>
<td>45</td>
<td>4</td>
<td>34   54</td>
</tr>
<tr>
<td>In Frame INDEL</td>
<td>inframe deletion</td>
<td>11</td>
<td>10</td>
<td>15</td>
<td>18</td>
<td>15</td>
<td>14</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>3</td>
<td>10   20</td>
</tr>
<tr>
<td></td>
<td>inframe insertion</td>
<td>10</td>
<td>17</td>
<td>19</td>
<td>8</td>
<td>14</td>
<td>10</td>
<td>11</td>
<td>8</td>
<td>16</td>
<td>11</td>
<td>3</td>
<td>8   19</td>
</tr>
<tr>
<td>Synonymous</td>
<td></td>
<td>199</td>
<td>278</td>
<td>284</td>
<td>259</td>
<td>221</td>
<td>283</td>
<td>202</td>
<td>208</td>
<td>242</td>
<td>242</td>
<td>36</td>
<td>199 284</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Non coding variants (regions)</td>
<td>Gene boundaries</td>
<td>downstream</td>
<td>19197</td>
<td>20411</td>
<td>18927</td>
<td>18009</td>
<td>17711</td>
<td>19363</td>
<td>16202</td>
<td>18463</td>
<td>20318</td>
<td>18927</td>
<td>918 16202 20411</td>
</tr>
<tr>
<td></td>
<td>upstream</td>
<td>19197</td>
<td>20761</td>
<td>19332</td>
<td>18122</td>
<td>18196</td>
<td>20182</td>
<td>16825</td>
<td>18944</td>
<td>21239</td>
<td>19197</td>
<td>1001</td>
<td>16825 21239</td>
</tr>
<tr>
<td>Intergenic</td>
<td></td>
<td>29694</td>
<td>38091</td>
<td>34040</td>
<td>31999</td>
<td>27269</td>
<td>31875</td>
<td>21550</td>
<td>32985</td>
<td>33380</td>
<td>31999</td>
<td>2041</td>
<td>21550 38091</td>
</tr>
<tr>
<td>Introns</td>
<td></td>
<td>55372</td>
<td>61682</td>
<td>56671</td>
<td>54869</td>
<td>5163</td>
<td>56193</td>
<td>43210</td>
<td>55945</td>
<td>61374</td>
<td>55945</td>
<td>1076</td>
<td>43210 61682</td>
</tr>
<tr>
<td>Non-coding transcripts</td>
<td>Mature miRNA</td>
<td>8</td>
<td>13</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>4   13</td>
</tr>
<tr>
<td></td>
<td>non-coding transcript</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0   2</td>
</tr>
<tr>
<td></td>
<td>non coding transcript exon</td>
<td>2149</td>
<td>2200</td>
<td>2116</td>
<td>1868</td>
<td>1920</td>
<td>2113</td>
<td>1811</td>
<td>2095</td>
<td>2310</td>
<td>2113</td>
<td>87</td>
<td>1811 2310</td>
</tr>
<tr>
<td>Regulatory regions</td>
<td>TF binding site</td>
<td>404</td>
<td>453</td>
<td>469</td>
<td>431</td>
<td>413</td>
<td>500</td>
<td>408</td>
<td>440</td>
<td>486</td>
<td>440</td>
<td>29</td>
<td>404 500</td>
</tr>
<tr>
<td></td>
<td>regulatory region</td>
<td>4667</td>
<td>5863</td>
<td>5301</td>
<td>4686</td>
<td>4512</td>
<td>5011</td>
<td>3582</td>
<td>4778</td>
<td>6158</td>
<td>4778</td>
<td>266</td>
<td>3582 6158</td>
</tr>
<tr>
<td></td>
<td></td>
<td>132674</td>
<td>151879</td>
<td>139131</td>
<td>132006</td>
<td>123179</td>
<td>137498</td>
<td>105487</td>
<td>135718</td>
<td>147631</td>
<td>135718</td>
<td>3712</td>
<td>105487 151879</td>
</tr>
</tbody>
</table>
This sequencing data will enable the research community to undertake and interpret further analyses (reviewed in Supplementary material 5) and to inform the use of these cell lines as a model of EAC. Our data highlight the need to develop additional in vitro models that have a germline reference genome to identify clearly the somatic changes (Gazdar et al., 1998). A larger number of cell lines might also more closely recapitulate the range of mutations observed in human disease.

Data availability
BAM files are available at the European Nucleotide Archive (ENA, EMBL-EBI, www.ebi.ac.uk/ena, Study PRJEB14018). Accession numbers: CP-D ERS1158083; SK-GT-4 ERS1158082; OE33 ERS1158081; OACP4 C ERS1158080; OACM5.1 ERS1158079; JH-EsoAd1 ERS1158078; FLO-1 ERS1158077; ES051 ERS1158076; ES026 ERS1158075.

Author contributions
GC collected and analysed the data, ME, AGL, MS and LB carried out bioinformatic analysis, RFE and JW contributed to STR analysis and DNA preparation, RCF, PAWE and GC conceived the study and wrote the manuscript. RCF and PAWE obtained funding for the study.
Supplementary material

1. **A** Copy Number Alteration of EAC cell lines according to ploidy shown by FREEC plots (loss, normal, and gain are indicated in blue, green and red, respectively). Genes annotated in red are the genes of the Cancer Genes Cosmic Census that fall in the amplified regions defined as copy number ≥5 for diploid and ≥7 for triploid and tetraploid cell lines. Genes annotated in blue are genes of the Cancer Genes Cosmic Census that fall in deleted regions with CN ≤1.

2. **B** Tables reporting all the genes of the Cancer Genes Cosmic Census that falls in deleted or amplified regions according to FREEC. Cell lines are shown in the following order: 1) CP-D, 2) ESO26, 3) ESO51, 4) FLO-1, 5) JH-EsoAd1, 6) OACM5.1 C, 7) OACP4 C, 8) OE33, 9) SK-GT-4.

3. **FREEC output of CNV by chromosome of the analysed cell lines.** CNV of each cell line is indicated by chromosome consistently to known ploidy and in silico verification with the Crambled App (Lynch et al., 2015).

4. Effect Predictor Analysis annotated VCF files of GAKT called variants for CP-D, ESO26, ESO51, FLO-1, JH-EsoAd1, OACM5.1 C, OACP4 C, OE33, SK-GT-4 are available for download at the EMBL-EBI European Variation Archive (EVA, http://www.ebi.ac.uk/eva/) under the study PRJEB14018.

5. **Filtered variants:** 1) CP-D, 2) ESO26, 3) ESO51, 4) FLO-1, 5) JH-EsoAd1, 6) OACM5.1 C, 7) OACP4 C, 8) OE33, 9) SK-GT-4.

6. **Publicly Available datasets for analysed cell lines.** For each cell line, currently available datasets from COSMIC, the Broad-Novartis Cancer Cell Line Encyclopaedia, and GEO (Gene Expression Omnibus) are listed.

7. **Alignment, duplication and insert size metrics for each cell line.**

References


Open Peer Review

Current Referee Status: ✔ ✔ ✔

Version 1

Referee Report 21 July 2016

doi:10.5256/f1000research.7571.r14746

Marnix Jansen
Barts Cancer Institute - a Cancer Research UK Centre of Excellence, Barts and The London School of Medicine and Dentistry, London, EC1M 6BQ, UK

In this study Contino present their WGS analysis of 9 (verified) oesophageal adenocarcinoma cell lines. This is an adequate platform to present these data and the fact that the authors make all raw BAM files easily accessible to the community means that this study is particularly valuable to colleagues looking to contrast cell lines with particular genomic aberrations or different neo-antigenic burdens. Such studies always come with the known caveats of in vitro selection and the authors rightfully acknowledge this. As expected, the study in large part confirms earlier large scale sequencing studies of primary material. The lack of a patient-specific reference control means that the impact of more subtle genomic abnormalities in for example regulatory regions remain difficult to study. Nonetheless this work represents a valuable addition to previously published datasets and the authors are to be commended for publishing this analysis. The paper is terse and I enjoyed reading this study.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 11 July 2016

doi:10.5256/f1000research.7571.r14843

Claire Palles, Laura Chegwidden
Oxford Centre for Cancer Gene Research, Wellcome Trust Centre for Human Genetics, Oxford, UK

The authors have performed whole genome sequencing of eight esophageal adenocarcinoma cell lines and one esophageal high grade dysplasia cell line to an average depth of 30x. The authors have made the BAM and VCF files available through the EBI repository and this will be an excellent resource for researchers working on this cancer. We feel the methods used are appropriate and most of the analyses described are informative. We do however have a few suggestions for the authors to address, these are listed below:

Dataset validation WGS section:
1. Clarify the % of variants that fall in each sequence context, coding, intronic, regulatory, intergenic. We assume this should sum to 100%.

2. In the next sentence there is a “in front of the 21% and 11% respectively”

3. Table 1: ploidy state of CP-D, should this be hypotetraploid?

4. Paragraph 2 of this section: Change 4,8x10^3 to 4.8x10^3

5. MuTect was used as variant caller in the Dulak paper and SomaticSniper was used in the Weaver paper. The authors should explain that they can’t use a somatic variant caller as these require a "normal" sample and also that application of a different caller for this cell line project may also make comparisons with the Dulak and Weaver papers less powerful.

Analysis of putative EAC driver genes:
1. There isn’t an ARID1A UTR variant shown for any of the cell lines in Figure 2b yet the authors mention 1 of the 9 cell lines has such a variant in the text.

On a related note we think the authors should consider the relevance of including UTR and synonymous changes in figure2b. We don’t think that these are considered in the Dulak and Weaver papers and are, as far as we understand, unlikely to be functional.

2. Second sentence of the second paragraph needs clarifying. Presumably missense mutations were found in MET and EGFR? IH-EsoAd1 should be JH-EsoAd1 in the same sentence.

3. Authors should make more of the fact that they have sequenced whole genomes whereas the COSMIC cell line project has only sequenced cell line exomes. The authors could perhaps highlight the useful extra data that is available from this sequencing effort, such as identification of mutations in putative regulatory regions and germline variants. Both classes of variants will be of interest to researchers working on understanding the genetics of oesophageal adenocarcinoma and wishing to identify appropriate cell models to work with.

**Competing Interests:** No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 05 July 2016
doi:10.5256/f1000research.7571.r14325

Ian Beales
School of Health Policy and Practice, University of East Anglia, Norwich, Norfolk, UK

The authors have examined the DNA sequences of 8 oesophageal adenocarcinoma cells lines and one high-grade dysplasia cell line. The authors should be congratulated for tackling this important unmet need in oesophageal cancer research and publishing these important findings in such an accessible manner. As the authors state, oesophageal adenocarcinoma seems to be one of the cancers carrying the most mutations, and although several cell lines, including those utilized in this study are commonly used for
laboratory studies, there has never been a systemic study of the genetic abnormalities in these cells lines. The data in this study does fill that important gap, allowing comparisons between them and the cancer in vivo.

The methods are appropriate for the study and well-described and the abstract accurately represents the contents of the study. The results are appropriately and clearly presented. The conclusions appear to be sound based on the data presented and most importantly the paper provides the data to enable other researchers to build on these data and hopefully further refine laboratory models for oesophageal adenocarcinoma.

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.