Supplementary Figure 1. The phylogeny of Case P7.
E1-E4 are samples from the esophagus collected at autopsy and E* is an endoscopy sample collected at diagnosis. We were unable to construct a single tree representation for this case, since E* carries mutations from 2 independent trees, represented by samples E1/E4 and E2, respectively. The most likely explanation is that P7 contains 2 independent tumors, both of which are represented subclonality in E*. However, E* had low tumor content (estimated 17%), yielding insufficient resolution to identify subclonality with confidence.
Supplementary Figure 2. Driver amplifications detected in primary, lymph node and distant metastatic sites. P=primary, L=lymph node, M=distant organ metastasis. Red indicates that at least one sample of the tissue type, including both H-WGS and S-WGS samples, contained a high-level amplification of the gene. White=no amplification identified in samples of the tissue type. Grey=no samples of the tissue type were sequenced. Although some CNAs were found across all tissue types within some patients, it is notable that every amplified gene identified was found in only a subset of samples (and in 17/20 cases in a subset of tissue types) in at least one patient, indicating substantial heterogeneity of CNAs and that selection of CNAs is ongoing during metastatic spread.
Supplementary Figure 3

(a) Simulated phylogeny. The number of mutations simulated was 10,000, spread across 6 clones/subclones. Each clone or subclone is shown as a separate oval, and samples in which each subclone is present are labelled with sample names S1 to S5. (b) The correct number of subclones (6) was identified in every simulation, despite incorporating up to 10% of mutations that violated the ISA. (c) Classification error in assigning mutations to clusters was very low and was unaffected by mutations violating the ISA. Number of simulations for each rate of violation of ISA, n = 20. Dark lines = median, Boxes = 25th and 75th quantiles, whiskers extend to the most extreme point within 1.5 × interquartile range of the box edge. (d) The error in estimating CCF rises as the number of mutations violating the ISA increases, but does not exceed 0.02. Number of simulations for each rate of violation of ISA, n = 20. Dark lines = median, Boxes = 25th and 75th quantiles, whiskers extend to the most extreme point within 1.5 × interquartile range of the box edge.
Supplementary Figure 4. Simulation to test sensitivity for detecting small clusters that invalidate the identification of diaspora. (a) Simulated phylogeny. The red branch indicates the branch that invalidates the stellar pattern expected for diaspora. (b) Number of ‘red’ branches detected, out of a total of 9. When the red branch contains less than 100 mutations, it is undetected. For 100 or more mutations most branches are detected, with all branches detected when the red branch contains 200 or more mutations. (c) Classification error is unaffected by the size of the red cluster (n=20 for no shared cluster, n=9 for all other boxes). Dark lines = median, Boxes = 25th and 75th quantiles, whiskers extend to the most extreme point within 1.5 × interquartile range of the box edge. (d) Error in CCF estimation is unaffected by the size of the red cluster (n=20 for no shared cluster, n=9 for all other boxes). Dark lines = median, Boxes = 25th and 75th quantiles, whiskers extend to the most extreme point within 1.5 × interquartile range of the box edge.
Supplementary Figure 5

A  Copy Number Profiles clustered across all samples in six cases

B  Amplifications seen in the primary tumor at diagnosis

Chr 12 P10 diagnostic biopsy

Chr 12 P3 diagnostic biopsy

C  Copy Number Alterations across all samples

|     | CDK6 | CKDN2A | CCND3 | EGFR | GATA4 | KRAS | MYC | MET | PRKC 1 | VEGFA |
|-----|------|--------|-------|------|-------|------|-----|-----|--------|-------|
| P1  | Diagnosis |       |       |       |       |      |     |     |        |       |
| P2  | Diagnosis |       |       |       |       |      |     |     |        |       |
| P3  | Diagnosis |       |       |       |       |      |     |     |        |       |
| P4  | Diagnosis |       |       |       |       |      |     |     |        |       |
| P5  | Diagnosis |       |       |       |       |      |     |     |        |       |
| P6  | Diagnosis |       |       |       |       |      |     |     |        |       |
| P8  | Diagnosis |       |       |       |       |      |     |     |        |       |
| P9  | Diagnosis |       |       |       |       |      |     |     |        |       |
| P10 | Diagnosis |       |       |       |       |      |     |     |        |       |

- **Amplification**
- **Deletion**
- **Unaltered**
Supplementary Figure 5. Copy Number Aberrations from 1x WGS
A. Copy number profiles clustered across all samples (n=248) for six cases at autopsy. Shallow whole genome sequencing was performed at an average depth of 1x to detect copy number changes across 248 samples at autopsy. Pearson correlation similarity matrix clustering was performed on all samples for each case (plotted against each other) with red indicating sample similarity (r=1) and blue indicating dissimilarity (r=-1). Sample sites used in this part of the study are shown in Table S10. The matrices for P1, P2 and P4 are shown. We have also included the number of copy number segments used to calculate the correlations, which varies between the cases but had a median of 141 (range 98-282).
B. Amplifications seen in the primary tumor at diagnosis. Copy number traces for Chromosome 12 are shown for P3 and P10 endoscopy samples at diagnosis. The chromosome position is on the X-axis and the segmented coverage on the Y-axis.
C. Copy Number Alterations across samples at Diagnosis and Death. The presence or absence of amplifications and deletions in specific genes known to be significant in EAC was analysed in eight cases where diagnostic and autopsy samples were available.
### Supplementary Table 1: Clinical demographics table for Surgical Cohort

| Case ID | Age at Diagnosis (years) | Treatment Status of Samples | Final Pre-treatment TNM Staging | Sampling Time-point | Survival (months) | Recurrence | Metastases sites sampled |
|---------|--------------------------|-----------------------------|---------------------------------|---------------------|-------------------|------------|-------------------------|
| S1      | 74.2                     | EOX                          | T3N1M0                          | Resection           | 14                | Not known  | Lymph node (1)          |
| S2      | 77.7                     | Naive                        | TxN0M0                          | Resection           | 20                | Yes (Liver)| Lymph node (1)          |
| S3      | 61.5                     | Naive                        | TxN2Mx                          | Laparoscopy         | 18                | DP         | Liver(1), Lymph Node(1) |
| S4      | 49.6                     | Lapatinib/ EOX               | T3N2M0                          | Resection/ Craniotomy* | 37                | Yes (Lungs, Brain) | Brain(2), Lymph Node(1) |
| S5      | 51.4                     | Naive/ EOX                   | T3N3M0                          | Endoscopy/ Craniotomy* | 10                | DP         | Brain(2), Lymph Node(1) |
| S6      | 67.6                     | Naive                        | TxN2 M1                         | Endoscopy           | 6                 | DP         | Lymph Node (1)          |
| S8      | 56.0                     | EOX                          | T3N1M0                          | Resection           | Alive             | Yes (mediastinal) | Lymph Node (1)          |
| S9      | 45.1                     | Naive                        | T3N2M1                          | Endoscopy           | Alive             | DP         | Lymph Node (1)          |

ECX: epirubicin, capecitabine and cisplatin. Not all samples had disease recurrence; those that had disease progression are marked with DP.
Supplementary Table 2: Clinical demographics table for Autopsy Cohort

| Case | Age (yrs) | Place of Death | Primary Tumor Location | Differentiation at Autopsy | Survival from diagnosis (months) | Metastatic Sites | Chemotherapy status |
|------|-----------|----------------|------------------------|---------------------------|---------------------------------|-----------------|---------------------|
| P1   | 45.1      | Hospice        | Type 2 GEJ             | Poorly differentiated      | 20                              | Pleural, Lung, Nodal, mediastinal | Treated          |
| P2   | 91.0      | Home           | Type 1 GEJ             | Poorly differentiated      | 12                              | Few Liver, Nodal          | Naive             |
| P3   | 82.1      | Hospice        | Type 1 GEJ             | Poorly differentiated      | 14                              | Few Liver, Nodal          | Naive             |
| P4   | 46.5      | Hospital       | Type 1 GEJ             | Poorly differentiated      | 5                               | Multiple Liver, Pancreas, Adrenal, Nodal | Treated          |
| P5   | 81.0      | Home           | Type 1 GEJ             | Poorly differentiated      | 28                              | Nodal              | Naive             |
| P6   | 65.4      | Hospice        | Type 1 GEJ             | Poorly differentiated      | 5                               | Multiple Liver, Pancreas, Adrenal, Nodal | Treated          |
| P7   | 71.6      | Hospice        | Type 1 GEJ             | Poorly differentiated      | 25                              | Primary, Nodal          | Treated          |
| P8   | 79.8      | Home           | Type 1 GEJ             | Poorly differentiated      | 30                              | Multiple Liver, Nodal     | Naive             |
| P9   | 44.9      | Hospital       | Type 1 GEJ             | Poorly differentiated      | 12                              | Multiple Liver, Nodal, Lung, Pleural, Serosal | Treated          |
| P10  | 78.5      | Home           | Type 1 GEJ             | Poorly differentiated      | 4                               | Liver, Nodal, Serosal, Pancreas | Treated          |

GOJ= gastro-esophageal Junction
**Supplementary Table 3: Detailed annotation of samples that underwent H-WGS as per Figure 1 and Extended Data 3 (P cases = Post mortem).**

| Case | Sample Abbreviation | Source | Case | Sample Abbreviation | Source |
|------|---------------------|--------|------|---------------------|--------|
| P1   | E*                  | Endoscopic biopsy (Esophagus) | P7   | E1                  | Esophagus |
|      | E**                 | Endoscopic biopsy (Esophagus) |      | E2                  | Esophagus |
|      | E1                  | Esophagus |      | E4                  | Endoscopic biopsy (Esophagus) |
|      | E2                  | Esophagus |      | E4                  | Esophagus |
|      | E3                  | Esophagus | P8   | E1                  | Esophagus |
|      | E4                  | Esophagus |      | E2                  | Esophagus |
|      | L1                  | Local lymph node (liver hilum) |      | E3                  | Esophagus |
|      | L2                  | Small bowel mesentery lymph node |      | E4                  | Esophagus |
|      | D1                  | Pleura |      | L1                  | Local lymph node (left gastric) |
| P2   | E1                  | Esophagus | L2   | Neck lymph node     |        |
|      | E2                  | Esophagus | L3   | Para-aortic lymph node |    |
|      | E3                  | Esophagus | D1   | Liver               |        |
|      | E4                  | Esophagus | P9   | E1                  | Esophagus |
|      | L1                  | Local lymph node (left gastric) |      | E2                  | Esophagus |
|      | L2                  | Neck lymph node |      | E3                  | Esophagus |
|      | D1                  | Liver |      | E4                  | Esophagus |
| P3   | E1                  | Esophagus | L1   | Local lymph node (left gastric) |    |
|      | E2                  | Esophagus | L2   | Neck lymph node     |        |
|      | E4                  | Esophagus | L3   | Para-aortic lymph node |    |
|      | L1                  | Local lymph node | D1   | Liver               |        |
|      | L2                  | Neck lymph node | D2   | Liver               |        |
|      | D1                  | Liver |      | D3                  | Right colon (serosa) |
| P4   | E1                  | Esophagus | D4   | Right pleura        |        |
|      | E2                  | Esophagus | D5   | Right lung          |        |
|      | E3                  | Esophagus | D6   | Left lung           |        |
|      | E4                  | Esophagus | B2   | Esophagus (Barrett’s) |    |
|      | L1                  | Local lymph node |      |                    |        |
|      | L2                  | Neck lymph node |      |                    |        |
| Case  | Sample Abbreviation | Source                      | Case  | Sample Abbreviation | Source                   |
|-------|---------------------|-----------------------------|-------|---------------------|--------------------------|
| P4 (contd) | L3                  | Para-aortic lymph node      | P10   | E1                  | Esophagus                |
|       | D1                  | Liver                       |       | E2                  | Esophagus                |
|       | D3                  | Adrenal                     |       | E3                  | Esophagus                |
|       | D2                  | Pancreas                    |       | E4                  | Esophagus                |
| P5    | E*                  | Esophagus                   | L1    | Local lymph node    |                          |
|       | E1                  | Esophagus                   | L2    | Neck lymph node     |                          |
|       | E2                  | Esophagus                   | L3    | Para-aortic lymph node |                      |
|       | E3                  | Esophagus                   | D1    | Liver               |                          |
|       | E4                  | Esophagus                   | D2    | Right hemidiaphragm |                          |
|       | L1                  | Local lymph node            | D3    | Falciformligament   |                          |
|       | L2                  | Local lymph node            | D4    | Pancreas            |                          |
| P6    | E1                  | Esophagus                   | D5    | Left hemidiaphragm  |                          |
|       | E2                  | Esophagus                   | D6    | Left colon (peritoneal) |                  |
|       | E3                  | Esophagus                   | D7    | Pelvis (peritoneal) |                          |
|       | E4                  | Esophagus                   | D8    | Right colon (peritoneal) |              |
|       | L1                  | Local lymph node            |       |                     |                          |
|       | L2                  | Neck lymph node             |       |                     |                          |
|       | L3                  | Para-aortic lymph node      |       |                     |                          |
|       | D1                  | Liver                       |       |                     |                          |
|       | D2                  | Adrenal                     |       |                     |                          |
Supplementary Table 4: Detailed annotation of samples that underwent H-WGS as per Figure 1 and Extended Data 3 (S cases=Surgery/ endoscopy)

| Case | Sample Abbreviation | Source                                      |
|------|--------------------|---------------------------------------------|
| S1   | E1                 | Esophagus                                   |
|      | L1                 | Local Lymph node (left gastric)             |
| S2   | E1                 | Esophagus                                   |
|      | L1                 | Local Lymph node (left gastric)             |
| S3   | E1                 | Esophagus                                   |
|      | E2                 | Esophagus                                   |
|      | L1                 | Local Lymph node (left gastric)             |
|      | D1                 | Liver                                       |
| S4   | E1                 | Esophagus                                   |
|      | E2                 | Esophagus                                   |
|      | E3                 | Esophagus                                   |
|      | L1                 | Local Lymph node (left gastric)             |
|      | L2                 | Local Lymph node (mediastinal)              |
|      | D1                 | Brain                                       |
|      | D2                 | Brain                                       |
| S5   | E1                 | Esophagus                                   |
|      | D1                 | Brain                                       |
|      | D2                 | Brain                                       |
| S6   | E1                 | Esophagus (Biopsy)                          |
|      | L1                 | Local Lymph Node                            |
|      | B1                 | Esophagus (Barrett’s)                       |
| S8   | E1                 | Esophagus                                   |
|      | L1                 | Local Lymph Node                            |
| S9   | E1                 | Esophagus                                   |
|      | L1                 | Local Lymph Node (left gastric)             |
Supplementary Table 5: Cancer cell fractions (CCFs) and number of Single Nucleotide Variants (SNVs) in clones and subclones in Figure 2. The samples in which each subclone occurs is shown, with the CCFs corresponding to it in the adjacent column. For example for case P1: E*, D1,, the CCF=0.52, 0.36 refers to E*=0.52, D1=0.52.. Associated number of SNVs in that subclone are shown. Please see excel spreadsheet

Supplementary Table 6: Cancer cell fractions (CCFs) and number of Single Nucleotide Variants (SNVs) in subclones in Extended Data Fig. 2 and 3. Please see excel spreadsheet.

Supplementary Table 7: Sample purities estimated using the Battenberg algorithm. Please see excel spreadsheet.

Supplementary Table 8: Confidence intervals of Cancer Cell Fractions. Each mutation cluster containing at least 1000 mutations or 1% of all mutations is shown, for each case. Confidence intervals (CIs) for each cluster were estimated as the 2.5% and 97.5% quantiles of the mean CCF of the mutations assigned to each cluster, obtained from Gibbs sampling of cluster positions from a Bayesian Dirichlet process. In individual samples, subclones corresponding to clusters are coloured dark green if they are clonal in that sample, light green if they are subclonal in that sample and white if they are absent from that sample. Clonal: upper CI > 1.0 or lower CI > 0.95; absent: lower CI ≤ 0.02; subclonal: 0.02 < lower CI ≤ 0.95. Five cases (P3, P4, P9, S2, S4) contained pairs of clusters whose CCF values could not be conformed to the sum or crossing rules. The corresponding CCFs of the corresponding clusters are shown in red and have been adjusted in order to fit all clusters to a single phylogeny. Please see excel spreadsheet.

Supplementary Table 9: Summary of wet lab validation.
Sheet 1: High depth verification exercise using Custom SNV panel in tissue samples (n=34). Each sample is listed along with the Pearson correlation between the 1x wgs and 50x wgs for subclone detection, the precision of 1xWGS and the number of mutations that were verified per sample.
Sheet 2: Exome sequencing of plasma samples (n=15). Each case is listed with the Pearson correlation between the 1x wgs and exome sequencing for subclone detection. Please see excel spreadsheet.
Supplementary Table 10: Distribution of sample sites for shallow whole genome sequencing (relating to Figure 4). Multiple metastases refer to multiple samples from a single metastatic site.

| Case ID | Primary | Lymph Nodes | Distant organ | Multiple metastases | Total Samples |
|---------|---------|-------------|---------------|---------------------|---------------|
| P1      | 6       | 7           | 11 (lung, pleural) | 2                   | 26            |
| P2      | 8       | 16          | -             | 8                   | 32            |
| P4      | 8       | 24          | 8 (liver)     | 8                   | 48            |
| P6      | 8       | 24          | 8 (liver)     | 8                   | 48            |
| P8      | 8       | 24          | 8 (liver)     | 8                   | 48            |
| P10     | 7       | 15          | 16 (8 liver, 8 serosal) | 8                   | 46            |

Supplementary Table 11: Table showing signature activity before and after the stellate pattern on the phylogenetic tree (as shown by red clone in Figure 2) and in truncal and branch clusters. The number of SNVs from each signature is shown, and the related accuracy. Please see excel spreadsheet.
Supplementary Table 12: Summary of clinical details of ctDNA from earlier time-points

| Case ID | Number of Samples | Time-points | Clinical Details | TP53 AF (%) from Digital PCR |
|---------|------------------|-------------|-----------------|-----------------------------|
| P1      | 1                | 6 months post-operative | 330 days from diagnosis, 172 days prior to radiological diagnosis of recurrence | 1.2 |
| P6      | 1                | Endoscopic Ultrasound (EUS) | 10 days from diagnosis, 139 days prior to radiological diagnosis of progression | 16.3 |
| P10     | 3                | Pre chemotherapy, chemotherapy end, follow up | 27, 48, 69 days from diagnosis | 57.6, 1.9, 48.9 |
| S3      | 1                | Clinic visit | 11 days prior to staging laparoscopy where metastasis diagnosed | NA as TP53 WT |
| S4      | 1                | Clinic- pre cycle 6 of chemotherapy and prior to detection of brain metastases on CT | 941 days from diagnosis, 116 days before radiological diagnosis of recurrence | 3.1 |
### Supplementary Table 13: Comparison of bioinformatic and biological parameters and clinical predictions for linear, parallel, neutral and diaspora models of metastases.

| MODEL PREDICTIONS                                                                 | LINEAR | PARALLEL | DIASPORA |
|----------------------------------------------------------------------------------|--------|----------|----------|
| Seeding before or after sub-clonal diversification of primary                    | After  | Before   | After    |
| Synchronous spread to multiple sites                                             | No     | Yes      | Yes      |
| Rapid establishment of multiple metastases (lack of signature 1 mutations)       | No     | No       | Yes      |
| Metastatic potential rare (~3 subclone) or common (many subclones)               | Rare   | Common   | Common   |
| Monoclonal / polyclonal seeding                                                  | Monoclonal | Either | Polyclonal |
| Metastasis-to-metastasis seeding?                                                | Likely | Unlikely | Unlikely |
| Tropism?                                                                         | No     | Yes      | No       |
| Majority of metastatic subclones in primary?                                     | No     | No       | Yes      |
| TNM staging appropriate?                                                          | Yes    | No       | No       |
| All subclones from primary identified in blood                                   | No     | Yes      | Yes      |
| Surgery + Neo-adjuvant chemotherapy sufficient?                                  | Yes    | Unlikely | Unlikely |
### Supplementary Table 14: TP53 Digital PCR Assay Probes

| Case ID(s) | Amino Acid Change | FAM prob | HEX           |
|------------|------------------|----------|---------------|
| P1         | p.R175H          | dhhsaCP2000105 | dhhsaCP2000106 |
| P6         | p.G245d          | dhhsaCP2500542 | dhhsaCP2500543 |
| P10        | p.R282W          | dhhsaCP2506902 | dhhsaCP2506903 |
| S4         | p.R273H          | dhhsaCP2000109 | dhhsaCP2000110 |

### Supplementary Table 15: Clonal structure composition: ISA Violation of Robustness Test

| Subclone number | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | Number of mutations |
|-----------------|----------|----------|----------|----------|----------|---------------------|
| 1               | 1.0      | 1.0      | 1.0      | 1.0      | 1.0      | 7000                |
| 2               | 1.0      | 0.0      | 0.0      | 0.0      | 0.0      | 200                 |
| 3               | 0.0      | 1.0      | 0.6      | 0.0      | 0.0      | 500                 |
| 4               | 0.0      | 0.0      | 0.4      | 0.7      | 0.2      | 1000                |
| 5               | 0.0      | 0.0      | 0.2      | 0.7      | 0.7      | 800                 |
| 6               | 0.0      | 0.0      | 0.0      | 0.3      | 0.3      | 500                 |

### Supplementary Table 16: Summary of samples and corresponding sequencing IDs uploaded to online data repository (EGA). Please see excel spreadsheet.
Supplementary Note

Library Construction and Sequencing

High Depth Whole Genome Sequencing (50X WGS)

A single library was created for each sample, and 100-125bp paired-end sequencing was performed under contract by Illumina (San Diego, USA,) to a typical depth of at least 50x, with 94% of the known genome being sequenced to at least 8x coverage while achieving a PHRED quality of at least 30 for at least 80% of mapping bases. QC metrics were computed on a per lane basis using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and in-house tools, enabling the identification of sequence reads that required trimming.

Shallow Whole Genome Sequencing (1X WGS)

A single library was created for each sample using the Thruplex DNA-seq kit (Rubicon Genomics, city, USA) workflow from up to 750ng of DNA. All tissue samples underwent sonication using the Covaris (Covaris Inc., City, USA) to 300 bp and fragment size analysis on the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA) prior to library construction. All frozen samples underwent 5 cycles of PCR and FFPE samples underwent between 5 and 10 cycles of amplification in order to achieve the minimum 10nm of library for sequencing. Quantification was performed using the Kapa Library Quantification Kit for NGS (Kapa Biosystems, city, USA) and each library underwent fragment size assessment using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA) in order to correct for fragment size. Up to 48 libraries per case were pooled equimolarly before undergoing clean up using Agencourt AMPure XP beads (Beckman Coulter Inc, city, USA) repeat qPCR and fragment size collection being undergoing PE 125 bp sequencing on the HiSeq 2500 in the Genomics core at the Cambridge Institute, CRUK.

Digital PCR

Digital PCR was performed on the QX200 Droplet Digital PCR system (Bio-Rad Laboratories, Berkley, USA). All reactions were performed to a total volume of 20μl, including 1ul (20x) of each wild type and mutant TP53 probe, 2x of supermix. HEX and FAM fluorophore tags were used to distinguish the wild type and mutant probes. All reactions were performed in triplicate, including a positive control, a water blank in each experiment and TP53 wild type
sample. A maximum of 50ng of DNA was inserted as per the manufacturer’s protocol. All probes (including those validated by the company) were first optimized using a gradient PCR reaction before selecting a final annealing temperature. Details of the TP53 exonic 65nt amplicon assays are provided in Supplementary Table 14.

**Mutation calling**

Sequence reads were aligned to the human reference genome (GRCh37) with BWA MEM v0.7.10 (Li and Durbin, 2009); the h37d5 decoy sequence used in the 1000 Genomes project (Genomes Project et al., 2012) was included in the reference genome. Alignments were sorted by genomic coordinate and PCR duplicates marked using Picard as previously described. Somatic single nucleotide variants (SNVs) were called using Strelka v1.0.13 (Saunders et al., 2012) and filtered as described previously. Structural variants (SVs) were called using Manta v0.27.1 (Chen et al., 2016) and filtered as described previously. Read counts in 15kb windows, corrected for sequence mapability, GC content, and filtered against a blacklist of problematic genomic regions, were obtained using QDNAseq v1.6.1. Allele read counts were tabulated for SNP loci taken from the 1000 Genomes project (phase 1 integrated SNP calls).

**Structural variant analysis**

Generalised Linear Models (GLMs) were used to test for a significant difference in either the number or type of structural variants. SV counts were found to be over-dispersed and a negative binomial, rather than Poisson, distribution was modelled, using the function glm.nb in the R package MASS. To compare the number of SVs, patient and sample type (primary or metastatic) were considered as independent variables, with sample type not leading to a significant effect on number of SVs (p=0.42). To test whether different sample types were associated with different types of SVs, negative binomial GLMs were constructed using sampletype and SVtype (insertion, deletion, duplication, inversion, interchromosomal, mobile element), with and without interaction between the variables. A chi-squared test indicated significant interaction (p=0.045), with more mobile elements identified in metastatic than primary samples (p=0.037).
Copy Number Identification and Haplotype Phasing

Segmental copy number information was derived for each sample using the Battenberg algorithm as previously described (Nik-Zainal et al., 2012). Briefly, the algorithm phases heterozygous SNPs with use of the 1000 genomes genotypes as a reference panel. The resulting haplotypes are corrected for occasional errors in phasing in regions with low linkage disequilibrium. After segmentation of the resulting b-allele frequency (BAF) values, t-tests are performed on the BAFs of each copy number segment to identify whether they correspond to the value resulting from a fully clonal copy number change. If not, the copy number segment is represented as a mixture of 2 different copy number states, with the fraction of cells bearing each copy number state estimated from the average BAF of the heterozygous SNPs in that segment.

Based on the above methods, copy number analysis was carried out on multiple samples in space and time. From the genome-wide variation in copy number, the similarity in copy number profiles was calculated for samples subject to 1X WGS at autopsy (Extended Data7A). Three cases (P1, P2, P4) clustered with 3 or more gross groups, whereas the remaining three cases (P6, P8, P10) demonstrated predominantly two main groups that clustered based on their copy number profile, resulting in a largely homogenous appearance for most of the samples in these cases (Extended Data7A). For P6 there were 2 samples that clustered separately from the rest of the samples at autopsy, based on common histological origin (adrenal gland). For P8 and P10 there were primary tumour samples (one in P8 and two in P10) that did not have metastatic potential, as correlated with the SNV subclonal data, which again resulted in predominantly two groups based on CNA profile.

We further analysed the presence or absence of copy number aberrations known to play a significant role in EAC (Secrier & Li et al. 2016, Weaver & Ross Innes et al. 2014, Nones et al. 2014, Dulak et al. 2013) at death and diagnosis (Extended Data7C). Focal amplifications covering KRAS in endoscopy samples from both P3 and P10 showed that these amplifications were not present in any of the samples subject to 50X sequencing, suggesting that they only occurred in one region of the primary tumour.

Copy Number Calling for Shallow Whole Genome Sequencing Analysis

The sequencing coverage of 1X WGS samples was summed in 15KB bins across the whole genome. Coverage values were normalized using qDNAseq (Scheinin et al., 2014) and regions with different copy number states were identified using the segmentation
algorithm multipcf, which is part of the R package *copynumber*, with a *gamma* parameter of 250.

Following segmentation, variable regions were identified as those containing at least 50 bins, i.e. at least 750KB in size, and with standard deviation across samples from the same patient ≥ 0.08. The similarity between related samples was then calculated as the Pearson correlation between the normalized copy numbers of the variable regions.

**Precision of Shallow Whole Genome Sequencing**

In order to assess the precision of 1x WGS, we performed targeted capture sequencing at high depth using a custom capture design using online software ‘SureSelect DNA Advanced Design Wizard’ from Agilent Technologies with a tiling density of 2x and balanced boosting. Both non-coding and coding mutations were used for the design, and the stringency was set on least, in order to ensure that non-coding mutations (in some cases occupying most clusters) were targeted. This design included 700 mutations across 14 subclones from 48 autopsy samples and one diagnostic FFPE sample from P10, and 900 mutations from P4 where an FFPE diagnostic sample was available. In two of the samples (P10 T4a and P10 lymph node 2 jejunum) we detected very few of the targeted mutations (0 and 1 respectively). The remaining samples yielded very high Pearson correlations between the allele frequencies of mutations assigned to each subclone from targeted high-depth sequencing and the fraction of mutations detected from 1x WGS (median 0.990, minimum 0.959). The fraction of mutations detected from 1x WGS that were validated by high-depth targeted resequencing, i.e. the precision, was 0.993, indicating a false positive rate of 0.7%.

For the plasma samples in Figure 5C and the FFPE diagnostic samples in Extended Data 6, we show only those subclones for which at least 1% of mutations were detected, above the false positive rate.

As further validation of 1x WGS of plasma samples, we carried out whole exome sequencing (WES) of plasma from 8 patients, sampled from autopsy and in five cases earlier time-points were also included. The correlation between the average WES allele frequencies of mutations assigned to each subclone and the proportion of mutations detected in 1x WGS was extremely high (Supplementary Table 15) and, interestingly, these correlations remained high when calculated just for subclones with allele frequencies below 2%.
Tests for Diaspora

We performed simulations in order to test the robustness of our identification of diaspora. In particular, we ascertained that our clustering algorithm is robust to

1. Mutations that violate the infinite sites assumption (ISA)
2. Subclones represented by a small number of mutations shared between elements of a diaspora.

Robustness Test 1: ISA Violation

In order to assess whether our clustering algorithm is robust to mutations occurring more than once within the same tumor, we generated data from a simulated tumor with the following properties:

- Sequencing depth drawn from a Poisson distribution with mean 55
- Tumor purity 55%
- 10,000 SNVs
- Five samples
- Clonal structure composed of truncal mutations plus four subclones arranged in a diaspora pattern and one subclone (subclone 6) that is a descendant of one of the diasporic subclones (subclone 5), illustrated in Supplementary Figure 3a, and described in Supplementary Table 15.

Purity, sequencing depths and number of SNVs were representative of values observed in real tumors and the diaspora was of similar size to that observed in real tumors, and contained a mixture of clonal and subclonal mutations; mutations shared between samples and unique to one site; samples that were clonal and others that were composed of multiple subclones.

Twenty simulations were performed for each proportion of mutations violating the ISA and robustness was assessed using 3 metrics: the number of subclones identified, the proportion of mutations incorrectly classified (excluding the mutations spiked into 2 subclones) and the mean error in CCF, calculated as the Euclidean distance between the vector of CCF values identified from clustering and the true CCF. In all simulation runs, the number of subclones was correctly identified as 6 (Supplementary Figure 3b). The proportion of mutations incorrectly assigned to a cluster was low (median 0.15%) and remained low across all simulations (Supplementary Figure 3c). There was an increase in
the error when estimating the CCF of subclones as the proportion of mutations violating
the ISA increased (Supplementary Figure 3d). However, CCF estimation appears to be
robust, since the median error level in the simulations with 10% mutations violating the ISA
was just 0.015.

Robustness Test 2: Diaspora Detection
A key feature of clonal diaspora is the appearance of a stellate pattern on the phylogenetic
tree, resulting from multiple branches derived from the founder clone. The identification of
diaspora would be invalidated by the occurrence of mutations shared between subclones
within a diaspora, since this would indicate that the single branch point identified should
be separated into multiple branch points. In order to assess our power to identify such
shared mutations, we performed further simulations in which an additional node was
inserted on the phylogenetic tree, representing mutations shared between 2 diaspora
subclones. For example, inserting a subclone that shares mutations between subclone 2
and subclone 3 results in the phylogeny shown in Supplementary Figure 4a (additional
subclone shown as a red line).

There are 10 possible pairs of subclones that may be combined. However, one of the pairs
– subclones 4 and 5 – is impossible, since subclone 5 is subclonal to subclone 4 and all cells
in subclone 5 already possess all mutations from subclone 4. Nine pairs of subclones were
therefore considered, and 10 simulations were performed for each pair, with the number
of mutations within the shared subclone varied between 25 and 250.

When the number of mutations in the shared subclone was less than 100, the shared
subclone was not identified (Supplementary Figure 4b). This number, corresponding to 1%
of all mutations, represents the lower limit of detection of such subclones. The number of
shared subclones detected rose from 7 out of 9 when the subclone carried 100 mutations,
to 9 out of 9 when the shared subclone carried 200 or more mutations, equivalent to 2% of
all mutations. The addition of the shared subclone did not lead to an increase in the
classification error (Supplementary Figure 4c) or in the error in estimating CCFs
(Supplementary Figure 4d) of the remaining subclones.

Copy number amplification identification
We looked for copy number amplifications in the following known driver genes: AKT1,
AKT2, AKT3, AR, AURKA, BRAF, CCND1, CCND2, CCND3, CCNE1, CDK4, CDK6, CRKL, EGFR, EPHA3, ERBB2, ERBB4, ESR1, FGFR1, FGFR2, FGFR4, FLT1, FLT4, GATA4, GATA6, HRAS, IGF1R, KDR, KIT, KRAS, MAP2K1, MAP2K2, MAPK1, MAPK3, MCL1, MDM2, MDM4, MET, MTOR, MYB, MYC, NFKBIA, NKX2-1, NRAS, PDGFRα, PDGFRβ, PIK3α, PIK3CB, PIK3R1, PRKCI, RAF1, VEGFA.

For 50X WGS samples, amplifications were called in any gene with a total copy number greater than or equal to 8. It was observed that, for some genes, some samples from a patient were called as amplified by this method while other samples were borderline, with total copy number between 6 and 8. For these genes, samples with total copy number between 6 and 8 were also called as amplified.

For 1X WGS samples, amplifications were called in any gene that was overlapped by a region with copy number 4 times greater than the average copy number for that sample. The normal contamination was highly variable across samples, reducing the sensitivity for detecting CNAs in some samples. To mitigate this effect, for genes that had been called as amplified in either a 50X WGS sample or another 1X WGS sample from the same patient, genes overlapped by regions that had copy number either greater than 3 times the genomic average or in the highest 5% of regions were also called as amplified.

**Selection Analysis**

The presence of positive selection was assessed using a dN/dS method (Martincorena Inigo, 2017). This identifies higher than expected occurrence of non-synonymous mutations, correcting for gene length, sequence composition, mutation signatures acting across patients and for the variation of the mutation rate along the genome.

**Mutational Signature Analysis**

Mutations were assigned to mutational signatures using previously described methods (Alexandrov et al., 2015). Mutational signatures provide another method to time events in the evolution of a cancer, particularly signature 1, which is due to enzymatic deamination of 5-methylcytosine, and causes mutations that accumulate linearly with age (Alexandrov et al., 2015). We thus evaluated the relative abundance of mutational signatures across all samples in a pooled analysis and used the number of mutations attributed to signature 1 to further understand the timing of events in Figure 3.

Mutations were classified as truncal or branch according to their position on the
Distant metastases all detected in one organ type
Separate subclones seeding different organs
Separate subclones seeding lymph nodes and distant organs
Metastases detected only in distant organs or only in lymph nodes

Tropism
We investigated whether there was evidence for tropism, presenting in any one of four ways:
- Distant metastases all detected in one organ type
- Metastases detected only in distant organs or only in lymph nodes
- Separate subclones seeding lymph nodes and distant organs
- Separate subclones seeding different organs

For six cases (S1, S2, S6, S8, S9, P5) we only sequenced one metastatic lesion and P7 included no metastases. These cases were therefore uninformative on the occurrence of tropism.

Just one of the remaining cases (P1) held strong evidence for tropism. For this cancer, separate subclones seeded the lymph nodes metastases and a pleural metastasis (Figure 2). 1X WGS confirmed the separation of lymph node and solid organ sites (Figure 4).

In the majority of the remaining cases (S4, P2, P3, P4, P6, P8, P9, P10) at least one subclone was found to have spread to both lymph node and distant organ sites, ruling out tropism.

We sequenced 1 lymph node and 1 liver metastases from S3. The 2 metastases had different subclonal origins, but the small number of samples does not enable us to ascertain whether this is the result of tropism or random seeding. Only 2 metastases were identified in S5, both in the brain. This is possible evidence for tropism, but the small number of samples does not enable us to ascertain whether these subclones may have also spread to other organs.