Figure S1. Panel (a) primary and (b) metastatic tumour stained with H&E. Panels (c-j) show the metastatic tumour IHC. The pattern did not differ from the primary (not shown) (c) ER (d) PR (e) E-cadherin (f) CK5/6 (g) EGFR (h) ERBB2 (i) INSR (j) Ki67. Magnification 200x (a,b,j), 400x other panels.
**Figure S2.** Schematic diagram of analysis workflow depicting from top to bottom the analytical steps and resulting inferences. The blue boxes show the biological features inferred from the aligned sequence reads through bioinformatic approaches. The yellow boxes show the complimentary experimental assays used to validate these features. The red boxes show features inferred from two different data sources.
Figure S3. Chromosome 17 and 19 copy number in the metastatic and primary tumour.

Panels (A) and (D) sequence based HMM predicted copy number, the top track in each panel is the primary tumour and bottom the metastasis. The high-level amplicons positions for MAP2K3 and INSR are shown boxed. Blue represents neutral copy number, red a gain and green a predicted loss. Panels (B) primary tumour and (C) metastasis show FISH for MAP2K3, centromere in green, MAP2K3 in red. MAP2K3 was not amplified in the primary tumour (figure S6). Panels (E) primary tumour and (F) metastasis show FISH for INSR, centromere in green, INSR in red. The INSR amplicon was present in the primary and metastasis (FISH). Panel (G) schematic of the INSR locus with RNA-seq read abundance showing in light blue over the INSR exons (black boxes). Read pairs supporting alternative splicing are linked by red lines. Exon 11 shows significantly reduced read representation.
**Figure S4.** Alignment showing deletion allele in ERBB2 in an additional cancer. The multiple alignment shows the relationship between a novel ERBB2 deletion allele and similar deletions observed in lung cancers for EGFR.
Figure S5. Extracted regions of Sanger traces for somatic mutation positions. Representative four colour traces representing somatic mutation positions analysed in the metastatic, primary and germline are shown. The full Sanger (.ab1 format) traces are available in the supplemental trace data files.
Figure S6. Schematic of the canonical estrogen signaling pathway (adapted from the Ingenuity database), with genes exhibiting alternative splicing events in green circles or diamonds. Solid lines/arrows represent direct functional relationships.
Figure S7. Cumulative frequency distribution plot showing the coverage of transcriptome sites defined as edited in normal tissues.

The data plot shows along the x-axis the read depth interval and along the y-axis, the cumulative fraction of the data. The data are 7844 positions from the study by Li et al (2009) survey of editing in normal tissues by deep sequencing, that are also present in the lobular transcriptome. The distribution shows that over 70% of the positions were covered at 10 reads or more and so could have detected editing, if it was occurring at high frequency. In fact, only 576 of these positions had any evidence for a non-reference base (i.e. a high-quality mismatch), though the majority (338) of these were consistent with an A->I edit.
Figure S7. Cumulative frequency distribution plot showing the coverage of transcriptome sites defined as edited in normal tissues. The data plot shows along the x-axis the read depth interval and along the y-axis, the cumulative fraction of the data. The data are 7844 positions from the study by Li et al (2009) survey of editing in normal tissues by deep sequencing, that are also present in the lobular transcriptome. The distribution shows that over 70% of the positions were covered at 10 reads or more and so could have detected editing, if it was occurring at high frequency. In fact, only 576 of these positions had any evidence for a non-reference base (i.e. a high-quality mismatch), though the majority (338) of these were consistent with an A->I edit.

Figure S8. Receiver operator characteristic curves (true positive rate (TPR)) vs false positive rate (FPR)) for SNVMix predictions using CRLMM high confidence predictions of Affymetrix SNP 6.0 high density genotyping array results as ground truth. (A) ROC considering a true positive (TP) as a prediction of heterozygous (AB) or homozygous non-reference position (BB) in the WGSS library. The AUC was 0.99 for 14649 positions for which there was at least 1 read in the WGSS and a high confidence (>0.99) CRLMM genotype call. (B) ROC for the WTSS library for 9688 positions matching the same criteria as (A). The grey dashed horizontal line indicates the FPR=0.01 used to determine the threshold (0.53) at which to ‘call’ an SNV. (C) ROC for WGSS where we considered (TP) to be positions called as homozygous for the reference allele in the Affymetrix chip. This resulted in a threshold of 0.99 to call a homozygous reference (aa) position, used as a threshold in the prediction of RNA editing and extinction events.

Figure S9. Effect of aligned sequence depth on the rate of new allele discovery and transcript coverage. A) Number of novel SNVs predicted by SNVMix against randomly sampled subsets of reads for WGSS (blue squares). B) The number of validated SNVs that appeared in the randomly sampled subsets (red triangles). The shape of the fitted curves shows how the rate of discovery for both predicted and confirmed SNVs is attenuated with additional reads. C) Number of novel SNVs predicted by SNVMix against randomly sampled subsets of reads for WTSS (red circles). D) Number of genes sequenced to a depth of 1x. Curves were fit to the data using a non-linear least squares estimate of the parameters of the logistic function using nls and SSlogis in R.
Figure S7. Cumulative frequency distribution plot showing the coverage of transcriptome sites defined as edited in normal tissues. The data plot shows along the x-axis the read depth interval and along the y-axis, the cumulative fraction of the data. The data are 7844 positions from the study by Li et al (2009) survey of editing in normal tissues by deep sequencing, that are also present in the lobular transcriptome. The distribution shows that over 70% of the positions were covered at 10 reads or more and so could have detected editing, if it was occurring at high frequency. In fact, only 576 of these positions had any evidence for a non-reference base (i.e. a high-quality mismatch), though the majority (338) of these were consistent with an A->I edit.

Figure S8. Receiver operator characteristic curves (true positive rate (TPR)) vs false positive rate (FPR)) for SNVMix predictions using CRLMM high confidence predictions of Affymetrix SNP 6.0 high density genotyping array results as ground truth. (A) ROC considering a true positive (TP) as a prediction of heterozygous (AB) or homozygous non-reference position (BB) in the WGSS library. The AUC was 0.99 for 14649 positions for which there was at least 1 read in the WGSS and a high confidence (>0.99) CRLMM genotype call. (B) ROC for the WTSS library for 9688 positions matching the same criteria as (A). The grey dashed horizontal line indicates the FPR=0.01 used to determine the threshold (0.53) at which to ‘call’ an SNV. (C) ROC for WGSS where we considered (TP) to be positions called as homozygous for the reference allele in the Affymetrix chip. This resulted in a threshold of 0.99 to call a homozygous reference (aa) position, used as a threshold in the prediction of RNA editing and extinction events.

Figure S9. Effect of aligned sequence depth on the rate of new allele discovery and transcript coverage. A) Number of novel SNVs predicted by SNVMix against randomly sampled subsets of reads for WGSS (blue squares). B) The number of validated SNVs that appeared in the randomly sampled subsets (red triangles). The shape of the fitted curves shows how the rate of discovery for both predicted and confirmed SNVs is attenuated with additional reads. C) Number of novel SNVs predicted by SNVMix against randomly sampled subsets of reads for WTSS (red circles). D) Number of genes sequenced to a depth of 1x. Curves were fit to the data using a non-linear least squares estimate of the parameters of the logistic function using nls and SSlogis in R.
Supplemental table legends

Table S1. Metastatic Transcriptome (WTSS-PE) coverage by RNA-seq reads.
The WTSS-PE library coverage of genome features such as exons, exon junctions and genes is shown. *A gene or exon is considered “observed” if at least 3 RNA-seq reads unambiguously mapped within it. Expression level was calculated as number of reads supporting each transcript position and the number of genes with >1.0 >10.0 and >100.0 reads per transcript nucleotide are shown. The median read coverage of all genes, exons or exon junctions considered observed is shown in the last column.

Table S2. Single nucleotide variants predicted by SNVMix for the WGSS-PE, WTSS-PE and WGSS-PRI libraries. The ‘notes’ column is broken down as follows: genomic position (NCBI build 36.1), Ensembl gene id (v51), strand (-1 or 1), reference base, non-reference base, codon position affected (1,2 or 3), [number of reference reads, reference base, number of non-reference reads, non-reference base, SNVMix p(aa), SNVMix p(ab), SNVMix p(bb), SNVMix call (1=aa,2=ab or 3=bb), reference codon, variant codon, reference amino acid, variant amino acid, mutation type (CODING=non synonymous, or SYNONYMOUS), effect on polarity, effect on charge, Grantham distance.

Table S3. Confirmed germline insertions and deletions validated using PCR cloning and Sanger sequencing.

Table S4. Confirmed germline inversions validated using PCR cloning and Sanger sequencing. Validation was attempted for all candidate inversions. All successful PCR amplicons were sequenced and cases in which the sequence supported a true inversion are included. All inversions were validated in the germline exclusively or in both the tumour and germline.

Tables S5a and S5b. Predicted high level amplicons in WGSS-PE indicating the gene(s) harboured, genomic coordinates (NCBI build 36.1), the frequency found in Chin et al cell lines and the frequency found in Chin et al (2007) tumors. Table S13 shows FISH validated regions, Table S14 contains the HMM copy number states and segment medians.

Table S6a. Somatic mutation frequency positions in primary tumour.
Table shows the position, number of reference allele calls, reference allele, number of nonreference allele calls, non-reference allele and the last column indicates whether the position assayed is the SNV or the flanking position.

Table S6b. Somatic mutation frequency positions in the metastatic tumour.
Column headings as for primary S6a.
Table S7. Allele frequency measurements full table. Green shaded positions represent germline alleles – all 36 control germline alleles are shown. Somatic mutations not present in the primary are shaded blue. First column, genome positions as chr:coordinate, second column locus symbol, third column reference base, fourth column non-reference base, fifth column primary read depth (number of reads), sixth column primary non-ref base ratio, seventh column primary Binomial exact p-value, eighth column status of allele, ninth column metastasis read depth, tenth column metastasis ratio, eleventh column metastasis p-value, twelfth column copy number state (loss, neutral, gain, amplified, high-level amplified)

Table S8. Alternative splicing. Predicted alternative intron usage determined from WTSS-PE reads aligned to exon junctions. All introns involved in at least one significant alternative splicing event are included. The number of reads supporting the use of the various alternative introns is shown (column 4) as well as the number of skipped Ensembl exons (column 5) where appropriate. An event is flagged as cryptic if one end of the intron does not terminate at an annotated Ensembl exon. For convenience the alternative exon events are also annotated in the supplemental data .bed file, which can be uplinked to a UCSC browser instance for browsing.

Table S9. Biased Allelic Expression Table. Heterozygous genomic (WGSS-PE) positions (either confirmed novel or dbSNP heterozygous loci) showing statistically significant unequal allelic expression in WTSS-PE inducing non-synonymous changes coding changes.

Table S10. Transcript expression table. The expression of each gene is summarized in various ways. The total number of reads indicates how many unambiguously aligned reads mapped to exons of this gene. The latter columns reflect the total bases sequenced from this gene (total coverage) and the corrected coverage after compensating for gene length.

Table S11. Candidate RNA edits. This table shows positions where a non-reference variation was called (Methods Summary) in the cDNA but not in the corresponding genome position. The first and second columns report chromosome and position. Third column genome strand, fourth column gene symbol, fifth column gene description, sixth column reference base, seventh column reads supporting ref base, eighth column non ref base, ninth column reads supporting non-ref base, tenth column effect of edit. Last column indicates whether the position is annotated in dbSNP or is a novel cDNA prediction (note of course, that none of the SNP positions showed the SNP in the lobular genome). Whether non-canonical events represent errors in transcription is unknown.

Table S12. RNA edits confirmed in the lobular transcriptome by Sanger sequencing. The first column shows the locus symbol and in brackets the expression level from RNA-seq (Table S10). The second column shows the position validated by Sanger amplicons sequencing as edited in the cDNA but not the genome as chr:pos, based on hg18 reference. Positions marked ** are non-synonymous.