Massively-parallel sequencing assists the diagnosis and guided treatment of cancers of unknown primary

Richard W Tothill[1,5,7], *Jason Li*[1], *Linda Mileshkin*[1,4], *Ken Doig*[1], *Terence Siganakis*[1], *Prue Cowin*[1], *Andrew Fellowes*[1], *Timothy Semple*[1], *Stephen Fox*[1], *Keith Byron*[2], *Adam Kowalczyk*[3], *David Thomas*[1], *Penelope Schofield*[1], *David D Bowtell*[1, 4, 5, 6, 7]

1. The Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia;

2. Healthscope Advanced Pathology, Clayton, VIC, Australia;

3. National (ICT) Australia, The University of Melbourne, Parkville, VIC, Australia

4. The Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, VIC, Australia

5. The Department of Pathology, University of Melbourne, Parkville, VIC, Australia

6. The Department of Biochemistry, University of Melbourne, Parkville, VIC, Australia

7. Corresponding authors.

richard.tothill@petermac.org

d.bowtell@petermac.org

Phone: +61 3 96561356

Fax: +61 3 96561414

Conflict of interest statement: The authors have no conflicts of interest to disclose

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/path.4251

Primary Genomic Data: Affymetrix SNP 6.0 array data is made available through NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/, GEO Accession ID: GSE49602 and Illumina short-read sequencing data made available through NCBI SRA (http://www.ncbi.nlm.nih.gov/sra, SRA Accession ID: SRP028343).

Abstract

The clinical management of patients with cancer of unknown primary (CUP) is hampered by the absence of a definitive site of origin. We explored the utility of massively-parallel (next-generation) sequencing for the diagnosis of a primary site of origin and for the identification of novel treatment options. DNA enrichment by hybridisation capture of 701 genes of clinical and/or biological importance, followed by massively-parallel sequencing, was performed on 16 CUP patients who had defied attempts to identify a likely site of origin. We obtained high quality data from both fresh-frozen and formalin-fixed paraffin-embedded samples, demonstrating accessibility to routine diagnostic material. DNA copy number obtained by massively-parallel sequencing was comparable to that obtained using oligonucleotide microarrays or quantitatively hybridized fluorescently tagged oligonucleotides. Sequencing to an average depth of 458-fold enabled detection of somatically acquired single nucleotide mutations, insertions, deletions and copy number changes, and measurement of allelic frequency. Common cancer causing mutations were found in all cancers. Mutation profiling revealed therapeutic gene targets and pathways in 12/16 cases, providing novel treatment options. The presence of driver mutations that are enriched in certain known tumour types, together with mutational signatures indicative of exposure to sunlight or smoking, added to clinical, pathological, and molecular indicators of likely tissue of origin. Massively-parallel DNA sequencing

can therefore provide comprehensive mutation, DNA copy number and mutational signature data that is of significant clinical value for a majority of CUP patients, providing both cumulative evidence for the diagnosis of primary site and options for future treatment.

Key Words

1. Cancer of unknown primary
2. Massively-parallel sequencing
4. Next-generation sequencing
5. Cancer diagnostic
6. Mutation profiling
7. Targeted therapy

Introduction

The starting point for the treatment of patients with metastatic cancer relies on first identifying the site of origin of the primary tumour, since such information provides important prognostic and predictive data. Cancers of unknown primary (CUP), in which the primary site cannot be identified, therefore pose a particular challenge for conventional approaches to patient treatment. CUP represents 2-5% of all cancer diagnoses and is the 4th highest cause of cancer related deaths worldwide[1]. There is therefore an urgent unmet need to improve the diagnosis and treatment of this disease.

Post-mortem autopsy can reveal a primary tumour in a majority of CUP cases[2], suggesting that failure to detect a primary tumour stems principally from limitations in conventional diagnostic methods. We[3] and others[4-7] have shown that gene-

expression profiling (GEP) can facilitate identification of the likely site of origin of CUP. Although GEP tests appear to have utility in diagnosis of site of origin of CUP, such tests can be inaccurate or inconclusive[8]. Evaluation of the performance of GEP assays is hampered by a lack of a definitive standard against which findings can be judged. For these reasons, a diagnosis of likely primary site is best obtained by the cumulative weight of clinical, pathological and molecular evidence, rather than relying on a single test.

Massively-parallel sequencing of tumour samples has been used to identify novel driver genes, investigate intra-tumoral heterogeneity, profile the mutational load of individual patient cancers, and guide therapeutic selection[9]. Sequencing of CUP tumours should provide insights into their biology, including testing the long held belief that these heterogeneous cancers share common biological properties[1]. Some driver mutations show a restricted distribution among cancer histotypes[10] and therefore mutation detection may add to other evidence suggesting a likely site of origin for individual CUP patients. The identification of actionable mutations complements considerations of anatomically-based therapy for CUP patients for whom a possible location for the primary has been determined by GEP or other assays. In those CUP patients where no site of origin can be found even after extensive clinical and molecular testing, sequencing may provide the only rational therapeutic approach available.

We explored the clinical merit of massively-parallel sequencing of 701 cancer-associated genes in a cohort of 16 CUP tumours where there was no conclusive site of origin. We derived both mutational and copy-number data by targeted capture and

sequencing of DNA from formalin fixed paraffin-embedded or fresh frozen tissues. We found that sequence information provided therapeutically useful information and facilitated the identification of probable site of origin in a substantial proportion of the patients.

Materials and Methods

Patient samples. Tumour specimens and germline DNA (blood) were collected from the Peter MacCallum Cancer Centre Tissue Bank and through referral from treating oncologists. Patient consent and Institutional Review Board approval were obtained according to the guidelines of the Australian National Health and Medical Research Council. CUP patient cases were selected retrospectively for the study based on established criteria[11, 12], where the primary tumour could not be identified beyond reasonable doubt following extensive clinical and histopathological review. Cases were evaluated using a combination of imaging modalities, endoscopy, immunohistochemistry and blood serum analysis. Details of clinical evaluation, diagnostic tests, patient treatment and outcome are provided in **Table 1** and **Supplementary Table S1**.

DNA Extraction: Genomic DNA was extracted using QIAamp DNA Blood Mini Kit (Qiagen, Germany) according to the manufacturer's protocol. For FFPE samples, a modified protocol was used where tissue was digested in Buffer ATL (Qiagen) containing proteinase K at 56°C for 3 days with daily proteinase K replacement. Purified DNA was quantified using a fluorometric assay (Quant-IT, Life Technologies, NY, USA).

Hybridisation capture and massively-parallel DNA sequencing: DNA capture bait libraries, complementary to gene targets, were designed using eArray (https://earray.chem.agilent.com/earray/) and manufactured for single sample hybridization capture (SureSelect, Agilent, CA, USA). Target DNA (0.3-1ug) was sheared using a focal acoustic device (Covaris, MA, USA) and then used for generating fragment libraries and hybridization capture, following the SureSelect recommended protocols. Ten indexed sample libraries were run per lane of an Illumina HiSeq2000 flowcell (paired-end 100bp) according to standard protocol (Illumina, CA, USA).

Variant and copy number detection: Sequence data was processed through the Illumina CASAVA software to split index reads and generate FASTQ data files. Data was quality checked using FASTQC program (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) and then aligned to the human genome (hg19 assembly) using BWA[13]. Local realignment around indels were performed using the GATK[14] software, and duplicate reads removed using Picard (http://picard.sourceforge.net). For tumour-normal paired samples, single nucleotide variants (SNVs) and indels were identified using the GATK Unified Genotyper, Somatic Indel Detector[15] and MuTect[16]. For the two tumours where no matching germline DNA was available, the GATK Unified genotyper was employed. Variants were annotated with information from Ensembl[17] Release 58, using the Ensembl Perl API including Variant Effect Predictor[18]. DNA copy number variation was estimated by computing median log-ratios in ~50bp windows using CONTRA[19] and then applying t-tests on the log-ratios of each gene against +/-0.2, with *p*-values adjusted using Benjamini-Hochberg method. Copy-number calls were filtered to

include only those calls where adjusted *p*-value was less than 0.05 and bin size (n) was greater than 10. A schematic representation of the sequence analysis pipeline is shown in **Supplementary Figure S1**. Versioning and software analysis parameters are detailed in **Supplementary Table S4**.

GEP Tumour Site of Origin Classifier: Detailed methods for construction of the GEP classifier have been described previously[3]. Briefly, total RNA was extracted from fresh frozen tissue specimens using phenol-chloroform and then column chromatography (RNeasy, Qiagen, CA). RNA from tumours was reverse transcribed, amplified and then labeled with Cyanine-5 flurophor. Labeled samples were co-hybridized to custom 10.5K feature cDNA arrays with Cyanine-3 labeled reference cDNA derived from a pool of 11 cell lines. Arrays were scanned for detection of fluorescent signal at gene features, converted to numerical data and normalised. Normalised data was used to classify samples using a support vector machine trained on a dataset of 229 tumour samples of known origin representing 13 tumour classes (breast, colorectal, gastric, lung, melanoma, mesothelioma, ovarian, pancreas, prostate, renal, SCC of skin or head and neck, testicular, uterine). A normalised score of between 0 and 100 was generated for each test sample from the SVM classification based on one-versus-rest classification. The margin between the first and second highest scoring tumour classes was used to assign a classification confidence measure (low, 0-25; medium, 25-50; high 50-100).

NanoString CNV Validation: DNA samples (~300ng) were analysed using the nCounter® Cancer CNV Assay (NanoString, WA, USA). Hybridisation, sample preparation, and scanning were performed according to the NanoString protocol. Copy-number was determined by normalizing raw count data against control probes targeting copy-number invariant regions and then against unmatched pooled normal tissue controls. Data from three probes targeting each gene was then averaged to give a single copy-number value per gene.

Copy-number arrays: Affymetrix SNP6.0 DNA copy number arrays were performed and analysed as previously described[20, 21](Affymetrix, CA, USA). All SNP CEL files were normalized in a single batch using the R package 'aroma.affymetrix' and segmented using the circular binary segmentation (CBS) algorithm to improve the signal to noise ratio. Matched normal tissue was not available for normalization, so the average signal from pooled male and female normal samples generated in our laboratory was used as a reference.

Results

Clinical profiles of CUP cohort

Sixteen patients were chosen for analysis, representing a spectrum of common CUP clinical presentations. Tumours were classified into four histological subgroups: adenocarcinoma, squamous cell cancer (SCC), small cell (neuroendocrine) and undifferentiated carcinoma. Most patients presented with multiple metastatic deposits, except for three SCC patients who had isolated lymphadenopathy of the head and neck or axillary node. All patients had undergone comprehensive histological and clinical workups, including immunohistochemistry (IHC), imaging modalities such as

CT, PET/CT as well as endoscopy (**Table 1** and **Supplementary Table S1**). Two patients with head and neck lymph node involvement appeared to be HPV-associated based on strong IHC p16-positivity. These were considered to be consistent with nodal metastases from an oropharyngeal tumour, although no primary could be identified. In two other cases potential primary sites became apparent several years after initial presentation (breast mass and pelvic mass from fallopian tube). Cell type-specific antibody staining (TTF1+) was suggestive of lung cancer in two cases, and an elevated serum CA125 levels suggested a gynaecological tract primary in two patients. A GEP-based classifier was applied to seven patients, enabling high confidence predictions in three instances (**Supplementary Table S2**). The classifier could not resolve three tumours displaying squamous differentiation. A low confidence prediction of breast cancer was made for one patient that was not consistent with clinicopathological evidence.

Profiling somatic mutations and copy-number alterations

Targeted hybridisation-capture and massively parallel sequencing was applied to tumour and, where available, matching germline DNA. The capture design of 701 genes was based on the human kinome, as these genes represent the largest single group for which there are targeted cancer therapeutics[22]. We interrogated the Cancer Gene Census[23], COSMIC[24], and other data sources to add additional cancer therapeutic targets and driver genes (**Supplementary Table S3**).

An average of 458-fold coverage was achieved for tumour and germline samples, with greater than 20-fold coverage in at least 97% of targeted bases (**Supplementary Figure S2**). A higher number of duplicate reads were observed for two of four FFPE

samples likely due to the smaller quantity and poorer quality of DNA extracted from archival specimens. However, as duplicate reads could be accurately removed from the paired-end sequence data this did not impact downstream analysis aside from a small reduction in read depth. Matching normal DNA was available for sequencing for 14/16 patients, allowing the use of a high accuracy somatic variant calling software (MuTect). For two samples without germline DNA, *post-hoc* filtering of variants was done using a reference set derived from combined germline variants called from 14 CUP cases to remove sequencing artefacts or common germline polymorphisms (*see* **Supplementary Table S5** for complete list of mutations).

High average read-depth enabled gene-specific DNA copy-number detection (**Supplementary Table S6**). The accuracy of copy-number detection was validated using independent techniques (Affymetrix SNP 6.0 array and/or NanoString) in 11 DNA samples showing a high level of concordance (**Figure 1** and **Supplementary Figures S3 and S4**).

Individual tumours displayed a diverse frequency of somatic mutations, many of which lay within known cancer genes (Figure 2). The frequency of mutations likely reflected both biological and technical influences. For example, one tumour (2864) had an exceptionally high number of single-nucleotide variants, likely explained by a somatically-acquired nonsense protein truncating mutation in *MLH1*, consistent with mis-match repair defect[25]. Some tumours displayed low maximum variant allele frequencies, suggesting correspondingly low tumour cellularity (**Supplementary Figure S5**). Although driver mutations could still be detected in such cases, low tumour cellularity impacted on the detection of copy number events.

Therapeutically actionable lesions

We first considered mutations, high-level copy-number gains or homozygous deletions in individual genes and pathways that represent potential targets for therapeutic treatment. Cases were assigned to three different categories based on prior evidence of drug efficacy according to tumour genotype. Category 1 included cases where there was strong clinical evidence supporting the efficacy of a drug based on the tumour genotype, category 2 where there was compelling pre-clinical evidence for efficacy of a drug based on tumour genotype and category 3 where a drug could be deployed based on a known gene to drug relationship, but where there is currently limited or inconclusive evidence supporting the efficacy of that drug in the context of the observed tumour genotype. Actionable mutations and copy-number aberrations were identified in 12 of 16 cases, with one case (1005) having two clinically relevant lesions (**Table 2**).

The majority of actionable mutations and copy-number alterations were identified in core mitogenic and cell growth pathways. Known hotspot point mutations in *PIK3CA*, *AKT1* and *KRAS* were identified in six samples. Two of three *PIK3CA* mutations represent a rare but recurrent mutation (E81K) of unknown functional significance, while the third *PIK3CA* mutation (E545K) and two *KRAS* mutations (G12C) are both well-known hotspot mutations[26, 27]. *AKT1* E17K is a known functional hotspot mutation in plecksterin homology domain and is a low frequency but relatively specific event in breast cancer[28]. Receptor tyrosine kinases (RTKs) that signal upstream of the core mitogenic pathways were altered in three cases. This included a mutation in the extracellular semaphorin domain of *MET* (R400S), a mutation in the

kinase domain of *FGFR3* (T742I) and high-level gain of *JAK2* that occurred concurrently with a *PIK3CA* mutation in the same sample. Other clinically actionable lesions included those involved in cell cycle (*CCND1*), DNA repair (*BRCA1*), cell fate (*PTCH1*) and metabolism (*IDH1*).

Diagnostic utility of mutation data

We next considered whether mutation analysis for next-generation sequencing could also have potential diagnostic utility and therefore assist in identifying cancer site of origin. Gene to cancer type relationships were systematically analysed for those cancer genes found to be mutated in the CUP samples by extracting the corresponding gene mutation frequencies observed in the COSMIC database[29](Figure 3A). The analysis of thousands tumours across major solid cancer types demonstrates that gene specific mutations could have diagnostic utility. Although many genes are found mutated in more than one cancer type the presence of a mutation in a tumour sample could theoretically assist restricting the differential diagnosis to one or a few sites.

We also investigated nucleotide-substitution patterns that are reflective of exposure to exogenous mutagens. Indeed, despite the small proportion of the genome captured, we were able to identify nucleotide substitution profiles in CUP samples associated with either UV-damage and tobacco smoking, consistent with skin and lung cancer respectively(**Figure 3B**). UV-induced DNA damage in skin cancers, such as melanomas, results in an exceptionally high number of C>T/G>A transitions and an increased frequency of variants at dipyrimidine bases[30]. Conversely, many lung cancers resulting from tobacco smoking, such as small cell carcinoma, exhibit a higher number of transversions of purine bases (C>A/G>T)[31]. While the smoking-

and UV-related signatures are not pathognomonic of a single tumor type[32], mutational profiles clearly have diagnostic utility, especially when this information is considered in light of a patient's history and the disease presentation.

The cumulative information obtained from gene-specific mutations and mutation profiles combined with clinical and pathological data of the respective cases was useful for narrowing the probable site of origin in several CUP patients (**Table 3**). For example, case 563 harboured mutations in *AKT* and *CDH1*, which are enriched in lobular breast cancer, thereby supporting the clinicopathologic evidence and GEP classification for this patient. Cases 1382 and 3282 harboured truncating and missense mutations, respectively, in *STK11*. *STK11* mutations are particularly common in NSCLC and their presence was consistent with TTF1 IHC positivity, a GEP classification of lung for patient 1382, and a nucleotide substitution profile reflecting smoking-associated DNA damage in both cases. The tumour from patient 168 had a hotspot mutation in *IDH1,* which occurs with high frequency in cholangiocarcinoma of intrahepatic origin[33]. This finding was consistent with the IHC profile (CK7+, CK20-) and presentation of an isolated segment four liver lesion in this patient. Finally, patient 3461 had a high-level *MYCL1* amplification. *MYCL1* amplifications occur frequently in small cell lung cancer and Merkel cell tumours[34]. The presence of a *MYCL1* amplification, together with a UV-associated mutation signature, was consistent with the clinical suspicion of a Merkel cell skin cancer. In summary, mutational profiling was helpful in the diagnosis of 11/16 cases. In some cases the mutation profile alone provided diagnostic evidence of cancer origin, where other diagnostic modalities had proven to be inconclusive.

Discussion

Recent studies have demonstrated the utility of massively-parallel sequencing for real-time analysis of patient samples and the subsequent deployment of targeted therapies[35, 36]. The technology has superior performance over multiallele-specific approaches and can be scaled to allow analysis of large gene panels in a single assay. Targeted sequence analysis provides a cost-effective strategy for clinical sequencing, enabling comprehensive analysis of clinically actionable gene sets to a very high read depth for sensitive variant calling and copy-number detection. Using a targeted panel, we identified mutations and copy-number alterations of therapeutic importance in the majority (12/16) of CUP samples. Although none would have dictated deployment of a currently approved therapeutic agent, for many of the patients mutation detection may have qualified them for inclusion in a clinical trial. The additional expansion of the gene panel to include gene fusions such as EML4-ALK, is both advisable and technically feasible using the approach described here[37].

Previous studies have reported a relatively low frequency of known cancer mutations across CUP cohorts when using more limited mutation detection panels[38, 39]. In our study, we identified known cancer driver lesions in all CUP cases profiled, which validated the importance of using large gene panels to detect uncommon gene mutations and the use of assays that can detect larger genomic alterations. Sequencing a broad gene panel in CUP could assist in the diagnosis of primary site where gene mutations occur in a cell type restricted manner and therefore including genes that are enriched in certain cancers is important, even if they are not therapeutically targetable. We showed that nucleotide substitution patterns associated with UV or smoking-related DNA damage can also be diagnostically informative. The identification of

such mutational signatures using a ~700 gene panel was possible for those tumours that have high mutation burden, however, the identification of other mutation signatures where the mutation frequency is lower is more restricted. It is therefore plausible that use of whole-exome or whole-genome sequencing may allow the classification of more cancer types based on known or currently unknown mutational profiles and this could be clinically useful [40, 41].

CUP cancers may share common biological properties that result in an early, aggressive and atypical metastatic spread. In this context, the reported high frequency of *MET* mutations in CUP[42] is intriguing. We observed a *MET* mutation in only one CUP tumour, which occurred in gene region coding for the extracellular domain and did not lie at a known mutation hotspot therefore is of unknown biological function. Given the heterogeneity of CUP, it is possible that our sample size was too limited to validate earlier findings. However, our systematic analysis of *MET* mutation frequency across the common tumours from the COSMIC database also shows that *MET* mutations arise in a cancer type restricted pattern. *MET* mutations occur at varying frequency across tumours of different cell type but appear absent from pancreatic cancer (0/474), a tumour type commonly implicated in CUP. It therefore seems likely that the mutation frequency of any gene within a CUP cohort can be influenced by the underlying representation of cancer types and therefore this should be considered for interpretation of future studies.

A majority of CUP patients had mutations for which there was clinical or pre-clincal data to suggest novel treatment options. Mutations in the PI3K or RAS pathways were seen in more than a third of patients analysed. An allosteric AKT1 inhibitor that

is effective against cells harbouring the E17K mutation has recently been described[43]. Phase I clinical trials deploying PI3K/AKT/mTOR inhibitors combined with chemotherapy have shown better response in patients harbouring PIK3CA mutations[44]. *KRAS*-mutant lung and low-grade ovarian cancers also have been showed to be responsive to the MEK inhibitor selumetinib[45, 46]. It is important to note that mutations in these signalling pathways can also confer resistance to both conventional and targeted agents. For example, activating *KRAS* mutations are associated with failure of anti-EGFR therapy in colorectal cancer[47]. The use of a multi-drug/multi-target regime is therefore likely to be a more effective strategy for treating solid cancers, especially when more than one oncogenic driver is implicated[48].

Mutations in receptor tyrosine kinases also appear relatively common in CUP and make effective targets for therapeutic inhibition. Mutation or amplification of *MET*, *FGFR3* or *JAK2* seen in some patients could theoretically direct the rational deployment of small molecular inhibitors crizotinib, BGJ398 and ruxolitinib, respectively[49-52]. A high-level gain of *CCND1* (cyclin D1) was identified in patient 1478 suggesting deregulation of the cell cycle control. Palbociclib has recently been recognized as a breakthrough therapy in breast cancer patients with CDK4/6 activation, protein partners of cyclin D1[53].

Alternative pathways for therapeutic intervention included those with mutated genes associated with DNA repair, cell fate and metabolism. A homozygous deletion of *BRCA1* was detected in patient 1698. Recent clinical trials show high response rates to PARP inhibitors in ovarian cancer patients with *BRCA1* mutation[54], with more

recent studies also suggesting activity in other tumour types[55]. A protein truncating *PTCH1* deletion in case 2864 indicated deregulation of hedgehog signalling pathway that may have conferred sensitivity to smoothened (SMO) inhibitor vismodegib[56]. Case 168 had a well-known hotspot mutation in *IDH1* known to cause deregulation of the Kreb-cycle and accumulation of the oncometabolite 2-hydroxyglutarate (2-HG)[57]. A recent study has reported the development of a mutation-specific *IDH1* inhibitor[58]. Interestingly, a concurrent *TET2* nonsense mutation was also identified in the same sample. These two mutational events are mutually exclusive in acute myeloid leukaemia, likely owing to the inhibitory effect of 2-HG on *TET2*[59]. This may suggest an evolutionary convergence on a common pathway in two sub-clones within case 168. It is conceivable that the co-occurrence of *IDH1* and *TET2* mutations in the same tumour may have implications for targeted treatment using an IDH1 inhibitor, as it would be expected that cells with inactivating *TET2* mutations could be insensitive to upstream IDH1 inhibition.

Although mutation profiling suggested a number of therapeutic options for CUP patients in our series, it is important to note that there are significant hurdles in translating mutational data to altered patient care. For example, even where there is a detailed understanding of the relationship between a given mutation and drug activity, extrapolation from one cancer type to another may not hold. Differences in the clinical activity of BRAF inhibitors in melanoma and colorectal cancer[60] demonstrate that cell lineage provides a molecular context that can influence the ability to attenuate the effects of a given driver mutation. Therefore, for CUP patients it will remain important to attempt to identify site of origin even where an actionable driver mutation is found, since lineage may affect the likelihood of response. The

heterogeneity of CUP has made the development of clinical trials problematic, reflected in the small number of phase III trials over the last decade. Histology-independent, aberration-specific clinical trials, so called "basket trials", represent potential new model to elucidate genetic biomarkers of drug response[61]. The pooling of data from individual patient studies[62] will also help provide an evidence base for the use of mutation data to guide novel treatment decisions.

With a median survival of between ~6-9 months, patients with a diagnosis of CUP have amongst the worst prognoses of any solid cancer[63]. Following the development of GEP for the identification of likely site of origin, massively-parallel DNA sequencing represents a further attempt to personalize the clinical management of CUP and improve outcomes. Although the number of patients analysed in this series was relatively modest, we established proof of principle in the use of massively-parallel DNA sequence to simultaneously obtain copy number and mutational data of therapeutic and diagnostic utility from fresh and FFPE samples. Based on this study, and our previous use of GEP classifiers, we suggest a flow chart for integrated clinical management of CUP patients (**Figure 4**) to be explored in future randomized trials.

Acknowledgments

We would like to thank the patients and clinicians who provided materials and information for the study.

Statement of author contributions

RT and DB conceived the study. RT, TSe and PC performed the experiments. JL, KD, TSi and AK analysed the data. SF, AF and KB provided pathological interpretation and material support. LM, PS and DT provided the clinical interpretation. RT and DB wrote the manuscript. All authors contributed to revision of the manuscript and approved the final version.

Funding

This study was supported by Cancer Australia

CUP patients were recruited through the Cancer 2015 study and supported by a Victorian Cancer Agency

Supplementary Files

Supplementary Tables S1-S5.xls

Table S1. Complete patient clinical summary data

Table S2. GEP results

Table S3. Targeted gene panel

Table S4. Analysis parameters

Table S5. Complete mutation results

Table S6. Complete CNV data from read-depth analysis with validation by Nanostring and Affymetrix SNP 6.0

Supplementary Figures.pdf

Figure S1. Sequencing analysis pipelines

Figure S2. Hybridisation capture performance summaries

Figure S3. Copy-number profiles from CUP samples

Figure S4. Correlation between orthogonal CNV data types

Figure S5. Tumour variant allele frequencies.

References

1. Pavlidis N, Pentheroudakis G. Cancer of unknown primary site. Lancet 2012; 379: 1428-1435.

2. Pentheroudakis G, Golfinopoulos V, Pavlidis N. Switching benchmarks in cancer of unknown primary: from autopsy to microarray. Eur J Cancer 2007; 43: 2026-2036.

3. Tothill RW, Kowalczyk A, Rischin D, et al. An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res 2005; 65: 4031-4040.

4. Monzon FA, Lyons-Weiler M, Buturovic LJ, et al. Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. J Clin Oncol 2009; 27: 2503-2508.

5. Greco FA, Spigel DR, Yardley DA, et al. Molecular profiling in unknown primary cancer: accuracy of tissue of origin prediction. Oncologist 2010; 15: 500-506.

6. Horlings HM, van Laar RK, Kerst JM, et al. Gene expression profiling to identify the histogenetic origin of metastatic adenocarcinomas of unknown primary. J Clin Oncol 2008; 26: 4435-4441.

7. Varadhachary GR, Spector Y, Abbruzzese JL, et al. Prospective gene signature study using microRNA to identify the tissue of origin in patients with carcinoma of unknown primary. Clin Cancer Res 2011; 17: 4063-4070.

8. Beck AH, Rodriguez-Paris J, Zehnder J, et al. Evaluation of a gene expression microarray-based assay to determine tissue type of origin on a diverse set of 49 malignancies. Am J Surg Pathol 2011; 35: 1030-1037.

9. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009; 458: 719-724.

10. Shepherd R, Forbes SA, Beare D, et al. Data mining using the Catalogue of Somatic Mutations in Cancer BioMart. Database (Oxford) 2011; 2011: bar018.

11. Massard C, Loriot Y, Fizazi K. Carcinomas of an unknown primary origin--diagnosis and treatment. Nat Rev Clin Oncol 2011; 8: 701-710.

12. Fizazi K, Greco FA, Pavlidis N, et al. Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2011; 22 Suppl 6: vi64-68.

13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25: 1754-1760.

14. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20: 1297-1303.

15. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011; 43: 491-498.

16. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013; 31: 213-219.

17. Flicek P, Ahmed I, Amode MR, et al. Ensembl 2013. Nucleic Acids Res 2013; 41: D48-55.

18. McLaren W, Pritchard B, Rios D, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 2010; 26: 2069-2070.

19. Li J, Lupat R, Amarasinghe KC, et al. CONTRA: copy number analysis for targeted resequencing. Bioinformatics 2012; 28: 1307-1313.

20. Cowin PA, George J, Fereday S, et al. LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin. Cancer Res 2012; 72: 4060-4073.

21. Gorringe KL, George J, Anglesio MS, et al. Copy number analysis identifies novel interactions between genomic loci in ovarian cancer. PLoS One 2010; 5.

22. Zhang L, Daly RJ. Targeting the human kinome for cancer therapy: current perspectives. Crit Rev Oncog 2012; 17: 233-246.

23. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer 2004; 4: 177-183.

24. Forbes SA, Tang G, Bindal N, et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res 2010; 38: D652-657.

25. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012; 487: 330-337.

26. Rudd ML, Price JC, Fogoros S, et al. A unique spectrum of somatic PIK3CA (p110alpha) mutations within primary endometrial carcinomas. Clin Cancer Res 2011; 17: 1331-1340.

27. Schubbert S, Shannon K, Bollag G. Hyperactive Ras in developmental disorders and cancer. Nat Rev Cancer 2007; 7: 295-308.

28. Carpten JD, Faber AL, Horn C, et al. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature 2007; 448: 439-444.

29. Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 2011; 39: D945-950.

30. Pleasance ED, Cheetham RK, Stephens PJ, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010; 463: 191-196.

31. Pleasance ED, Stephens PJ, O'Meara S, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 2010; 463: 184-190.

32. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature 2013.

33. Borger DR, Tanabe KK, Fan KC, et al. Frequent mutation of isocitrate dehydrogenase (IDH)1 and IDH2 in cholangiocarcinoma identified through broad-based tumor genotyping. Oncologist 2012; 17: 72-79.

34. Paulson KG, Lemos BD, Feng B, et al. Array-CGH reveals recurrent genomic changes in Merkel cell carcinoma including amplification of L-Myc. J Invest Dermatol 2009; 129: 1547-1555.

35. Wagle N, Berger MF, Davis MJ, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov 2012; 2: 82-93.

36. Tran B, Brown AM, Bedard PL, et al. Feasibility of real time next generation sequencing of cancer genes linked to drug response: results from a clinical trial. Int J Cancer 2013; 132: 1547-1555.

37. Lipson D, Capelletti M, Yelensky R, et al. Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. Nat Med 2012; 18: 382-384.

38. Hale KS, Wang H, Karanth S, et al. Mutation profling in patients with carcinoma of unknown primary using the Sequenom MassARRAY system. J Clin Oncol 30, 2012 (suppl; abstr 4131); 2012; 2012.

39. Ohta S, Cho Y, Shibata M, et al. Possibility of molecular targeting therapy for the treatment of cancer of unknown primary origin by analysis of intracellular signaling molecules. Exp Ther Med 2012; 3: 547-549.

40. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Deciphering signatures of mutational processes operative in human cancer. Cell Rep 2013; 3: 246-259.

41. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013; 499: 214-218.

42. Stella GM, Benvenuti S, Gramaglia D, et al. MET mutations in cancers of unknown primary origin (CUPs). Hum Mutat 2010:44-50.

43. Jo H, Lo PK, Li Y, et al. Deactivation of Akt by a small molecule inhibitor targeting pleckstrin homology domain and facilitating Akt ubiquitination. Proc Natl Acad Sci U S A 2011; 108: 6486-6491.

44. Janku F, Wheler JJ, Westin SN, et al. PI3K/AKT/mTOR inhibitors in patients with breast and gynecologic malignancies harboring PIK3CA mutations. J Clin Oncol 2012; 30: 777-782.

45. Janne PA, Shaw AT, Pereira JR, et al. Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. Lancet Oncol 2013; 14: 38-47.

46. Farley J, Brady WE, Vathipadiekal V, et al. Selumetinib in women with recurrent low-grade serous carcinoma of the ovary or peritoneum: an open-label, single-arm, phase 2 study. Lancet Oncol 2013; 14: 134-140.

47. Allegra CJ, Jessup JM, Somerfield MR, et al. American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. J Clin Oncol 2009; 27: 2091-2096.

48. Saini KS, Loi S, de Azambuja E, et al. Targeting the PI3K/AKT/mTOR and Raf/MEK/ERK pathways in the treatment of breast cancer. Cancer Treat Rev 2013.

49. Rodig SJ, Shapiro GI. Crizotinib, a small-molecule dual inhibitor of the c-Met and ALK receptor tyrosine kinases. Curr Opin Investig Drugs 2010; 11: 1477-1490.

50. Guagnano V, Kauffmann A, Wohrle S, et al. FGFR genetic alterations predict for sensitivity to NVP-BGJ398, a selective pan-FGFR inhibitor. Cancer Discov 2012; 2: 1118-1133.

51. Chase A, Bryant C, Score J, et al. Ruxolitinib as potential targeted therapy for patients with JAK2 rearrangements. Haematologica 2013; 98: 404-408.

52. Gozgit JM, Wong MJ, Moran L, et al. Ponatinib (AP24534), a multitargeted pan-FGFR inhibitor with activity in multiple FGFR-amplified or mutated cancer models. Mol Cancer Ther 2012; 11: 690-699.

53. Finn R. Results of a randomized phase 2 study of PD 0332991, a cyclin-dependent kinase (CDK) 4/6 inhibitor, in combination with letrozole vs letrozole alone for first-line treatment of ER+/HER2- advanced breast cancer (BC). Abstract. Publication Number S1-6. 2012 San Antonio Breast Cancer treatment of ER+/HER2-advanced breast cancer (BC). San Antonio Breast Cancer Symposium (SABCS); 2012; San Antonio, Texas, USA; 2012.

54. Fong PC, Boss DS, Yap TA, et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. N Engl J Med 2009; 361: 123-134.

55. Kaufman B, Shapira-Frommer R, Schmutzler RK, et al. Olaparib monotherapy in patients with advanced cancer and a germ-line BRCA1/2 mutation: An open-label phase II study. Journal of Clinical Oncology 2013; 31.

56. Sekulic A, Migden MR, Oro AE, et al. Efficacy and safety of vismodegib in advanced basal-cell carcinoma. N Engl J Med 2012; 366: 2171-2179.

57. Dang L, White DW, Gross S, et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature 2010; 465: 966.

58. Rohle D, Popovici-Muller J, Palaskas N, et al. An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells. Science 2013; 340: 626-630.

59. Figueroa ME, Abdel-Wahab O, Lu C, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell 2010; 18: 553-567.

60. Prahallad A, Sun C, Huang S, et al. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature 2012; 483: 100-103.

61. Sleijfer S, Bogaerts J, Siu LL. Designing transformative clinical trials in the cancer genome era. J Clin Oncol 2013; 31: 1834-1841.

62. Lillie EO, Patay B, Diamant J, et al. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med 2011; 8: 161-173.

63. Pavlidis N. Cancer of unknown primary: biological and clinical characteristics. Ann Oncol 2003; 14 Suppl 3: iii11-18.

64. Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. Science 2013; 339: 1546-1558.

Table 1. Summary of patient histories and clinicopathological features.

CUP Cases (*n*)	16
Male	7
Female	9
Risk factors	
Personal history of malignant or benign tumours	5
Family history of cancer	1
Smoking	9
Extensive sun exposure	2
Clinical presentation	
Liver	2
Gynaecological tract	2
Lung	2
Other extremity (e.g. limb)	1
Lymph node (pelvic, inguinal)	1
Lymph node (abdominal, para-aortic, mediastinal)	2
Lymph node (supraclavicular, axillary)	3
Lymph node (head and neck)	3
Bone	5
Histology	
Adenocarcinoma	4
Poorly Differentiated Carcinoma/Adenocarcinoma	7
SCC	4
Neuroendocrine/Small cell carcinoma	1
Investigations	
IHC and/or serum markers	14
Microarray gene-expression classifier (Tothill et al 2005)	7
CT, PET-CT, MRI, X-ray, Mammogram	15
Endoscopy or other invasive investigation	5

Table 2. Molecular-based therapeutic targets in cancers of unknown primary

	Actionable lesions	**Drugs**
Category 1		
1005	*PIK3CA p*.E545K VAF: 0.33	PIK3CA/AKT/mTORi (e.g. PX866 or temsirolimus)
1698	*BRCA1 Homozygous deletion*	PARPi (e.g. olaparib)
1382	*KRAS p*.G12C VAF:0.33	MEKi (e.g. selumetinib)
8593	*KRAS p*.G12C VAF:0.45	MEKi
2864	*PTCH1* p.S1203Afs*52 VAF:0.36	SMOi (e.g. vismodegib)
Category 2		
168	*IDH1 p.R*132L VAF: 0.09*	IDH1i (e.g. AGI-5198)
563	*AKT1 p*.E17K VAF: 0.51	AKTi (e.g. SC66)
Category 3		
1478	*CCND1* HLG	CDK4/CDK6i (e.g. palbociclib)
1184	*PIK3CA p*.E81K (VUS) VAF:0.23	PIK3CA/AKT/mTORi
1005	JAK2 *High level CN-gain*	JAKi (e.g. ruxolitinib)
91	*PIK3CA p*.E81K (VUS) VAF: 0.06*	PIK3CA/AKT/mTORi
3461	*FGFR3 p*.T742I (VUS) VAF:0.78	FGFRi (e.g. ponatinib)
3282	*MET p*.R400S (VUS) VAF:0.19	METi (e.g. crizotinib)

VUS, variant of unknown significance predicted to be damaging by variant predictor software SIFT/polyphen/Condel; CN-gain, copy-number gain. VAF; Variant allele frequency. * Tumour content likely to be low in these samples

Table 3. Cumulative evidence to support cancer site diagnosis

Case	Clinicopathological details	Genomic evidence	Likely Origin
563	F, 67yrs. Hx of OvCa. PD-AD in bone, liver and cervical node dissimilar to 1^{o} OvCa episode. IHC: CK7+CK20-, HER2–; Serum CA125 normal	**GEP**: BR (High) **Mut Genes**: *AKT1*, *CDH1*	Breast
1698	F, 37yrs. MD-AD in ovary, uterus, omentum and pleura. Mixed histological features including signet ring morphology. IHC: CK7+, CK20-; Serum CA125 +++	**GEP**: OV (High) **Mut Genes**: *TP53*, *BRCA1*	Ovary
1184	M, 66yrs. Hx of smoking, PD-SCC in bone and multiple nodal sites	**GEP**: LU (Low) **Mut Sig**: smoking	Lung
1382	F, 74yrs. Hx of KiCa and smoking. PD-AD in bone. IHC: Vim-, CK7+, CK20-, CEA+, ER/PR-, TTF1+	**GEP**: LU (High) **Mut Genes**: *STK11* **Mut Sig:** smoking	Lung
2864	M, 53yrs, Hx of skin lesions. PD-C with clear cell features in axillary node. IHC: CK7-, CK20-, HMWCK+, CEA-, TTF1-, AFP-, Vim+, Muc-	**GEP**: LU(Low). **Mut Genes**: *PTCH1* **Mut Sig**: UV	Skin
3282	M, 49yrs. Hx of smoking. PD-AD in bone. IHC: CK7+, CK20-, TTF1 weak +	**GEP**: N/T **Mut Genes**: *STK11* **Mut sig**: smoking	Lung
3461	F, 72yrs. Hx of BrCa, ColoCa, SkiCa (benign). UD-SmCC in inguinal node. IHC: AE1/AE3+, CK7-, CK20-, CEA-, CD56-, Chromogranin-, S100-, TTF1 -Synap+	**GEP**: N/T **Mut Genes**:*MYCL1* **Mut Sig** UV	Skin (MerkCa)
4413	F, 74yrs. Family Hx of BrCa. PD-AD with papillary features in para-aortic node and ovary. IHC: CK7+, CK20-, Serum CA125+++	**GEP**: BR (Low)	Unknown
8593	M, 79yrs. Hx of smoking. AD in subcutaneous lesion, bone, adrenal and lung. IHC: CK7+ CK20-, TTF1-, CEA+	**GEP**: N/T. **Mut sig**: smoking	Lung
11674	M, 64yrs, SCC in cervical node. IHC: P16+	**GEP**: N/T.	H&N
2406	F, 74yrs. Hx of smoking. AD in inguinal node. IHC: CK7+, CK20-, S100- HMB45- Serum CEA, CA15, CA15.3, CA19.9 normal	**GEP**: N/T	Unknown
1478	F, 77yrs. PD-SCC in cervical node and liver. IHC: CK7-, CK20-, ER-, PR-, thyroglobulin-, TTF1-	**GEP**: LU(Low)	Unknown
1005	M, 53yrs. Hx of smoking. SCC in cervical node. IHC: P16+	**GEP**: N/T	H&N
160	M, 81yrs. Hx of SkiCa (BCC). PD-C in subcutaneous lesion	**GEP**: N/T **Mut sig**: UV	Skin
91	F, 54yrs. Hx of smoking. MD-AD in omentum and multiple nodes. IHC: CK7+, CK20 weak+, CDX2-, Vim-, TTF1-	**GEP**: N/T **Mut sig**: smoking	Lung
168	F, 79yrs. PD-AD in liver and peritoneum. IHC: CK7+, CK20-, ER-, GCDFP-15-	**GEP**: N/T **Mut Genes**: *IDH1*	Billiary tract

F, female; M, male; yrs, years of age; Hx, history; OvCa, ovarian cancer; SkiCa, skin cancer; BrCa, breast cancer; KiCa, kidney cancer; LuCa, lung cancer; BCC, basal cell carcinoma; ColoCa, colorectal cancer; MerkCa, PD-AD, poorly differentiated adenocarcinoma; MD-AD, moderately differentiated adenocarcinoma; AD, adenocarcinoma; PD-C, poorly differentiated carcinoma; UD-SmCC, undifferentiated small cell carcinoma; PD-SCC, poorly differentiated squamous cell carcinoma; SCC,

squamous cell carcinoma; c/w, consistent with; IHC, immunohistochemistry; GEP, gene-expression profile, Mut, mutation; N/T, not tested; confidence level for GEP shown in parenthesis.

Figure 1. Validation of copy-number variations in a single fresh frozen clinical sample (1698) using three orthogonal platforms. Y-axis shows the log-ratio of copy-number change compared to a normal genome sample (unrelated) determined using read depth analysis (ReadDepth), Affymetrix SNP6.0 array (SNP6) and Nanostring. The chromosome position for each gene and data point is ordered along X-axis.

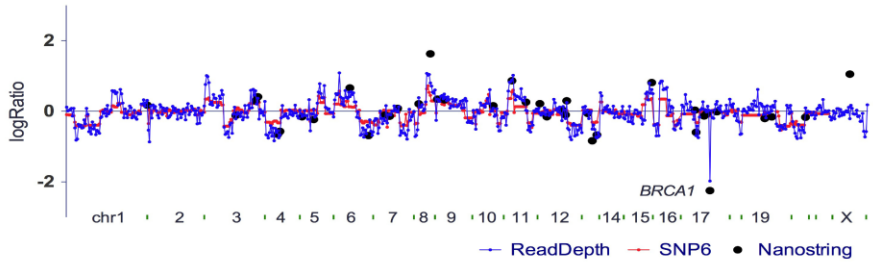

Figure 1

Figure 2. Mutation and copy-number analysis of 16 CUP cases A) Total number of subtle mutations and gene copy-number variations identified across all CUP samples. High-level gains denoted where log2-fold change >1. Homozygous deletions denoted where log2-fold change <-1. B) Known cancer genes harbouring functional mutations predicted to be non-synonymous, protein truncating or affect essential splice sites in addition to genes affected by high-level gain or homozygous deletion. The cancer gene list together with tumour suppressor gene (TSG) or oncogene (ONC) annotation shown in parenthesis was derived from Vogelstein *et al*[64].

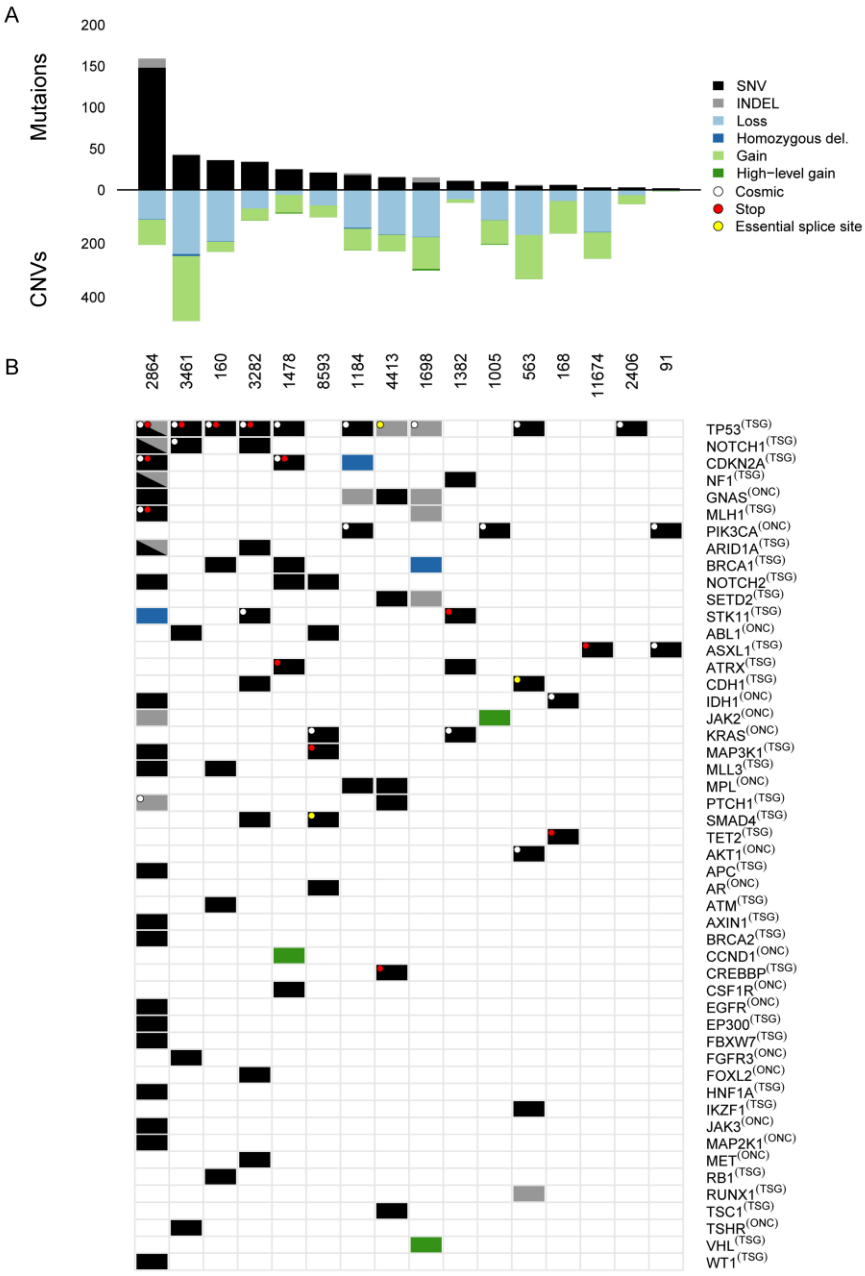

Figure 2

Figure 3. Gene mutation frequency in common cancer types and mutational signatures defining exogenous mutagenic exposures A) Mutation frequency in common epithelial derived cancers extracted from COSMIC database (v65). Data shown only for those genes harbouring mutations across the 16 CUP cohort. The heatmap mutation frequency has been capped at 10%. Data is not shown for cancer types and genes if less than 50 samples had been screened and is represented by pink shading B) Evidence of mutational signatures corresponding to an excessive exposure to UV light and tobacco smoke similar to that observed in melanoma and small-cell lung cancer (SCLC).

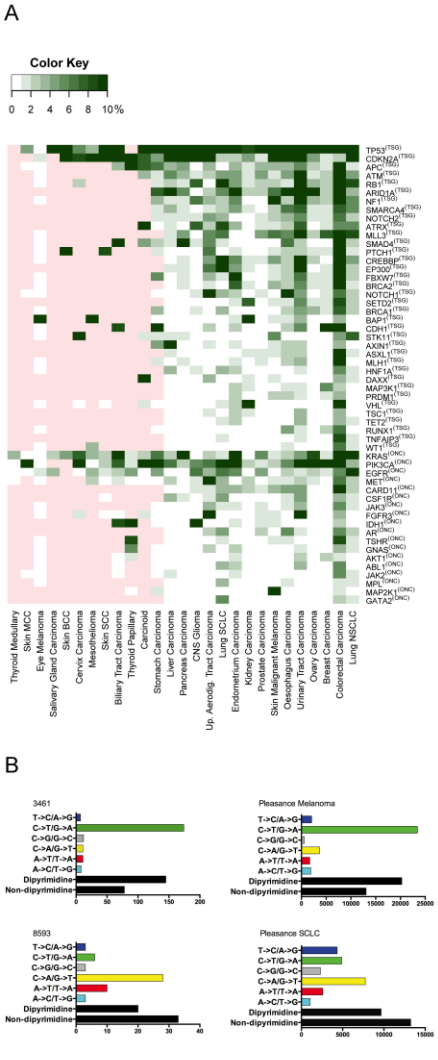

Figure 3

Figure 4. Integration of molecular tests into the clinical management of CUP. A patient is deemed to have cancer of unknown primary following thorough clinical and pathological investigation. Tissue biospecimens are collected from surgery or biopsy and blood is taken for germline testing. RNA and DNA is extracted from tissue and/or blood samples. Targeted DNA capture and sequencing is applied to tumour and blood DNA to identify somatic and germline variants and to detect mutational signatures associated with smoking or UV exposure. Gene-expression profiling (GEP) is performed on tumour RNA (microarray or other platform) and samples are classified using a multi-cancer type classifier. Cumulative evidence from GEP, cancer type specific mutations, mutation signatures and clinical data (disease presentation, immunohistochemistry (IHC) profile, family or patient history) is used to identify a likely site of origin. Strong molecular evidence of cancer's primary origin may be used to direct further clinical investigation to confirm a primary tumour (e.g. endoscopy). Any therapeutically actionable mutations or putative germline risk alleles should be validated using an orthogonal method before influencing patient management. The therapeutic approach chosen in an individual patient will be influenced by the strength of evidence implicating a likely tissue of origin, the nature of any actionable mutation(s) detected, the clinical condition of the patient and the types of conventional and targeted agents available.

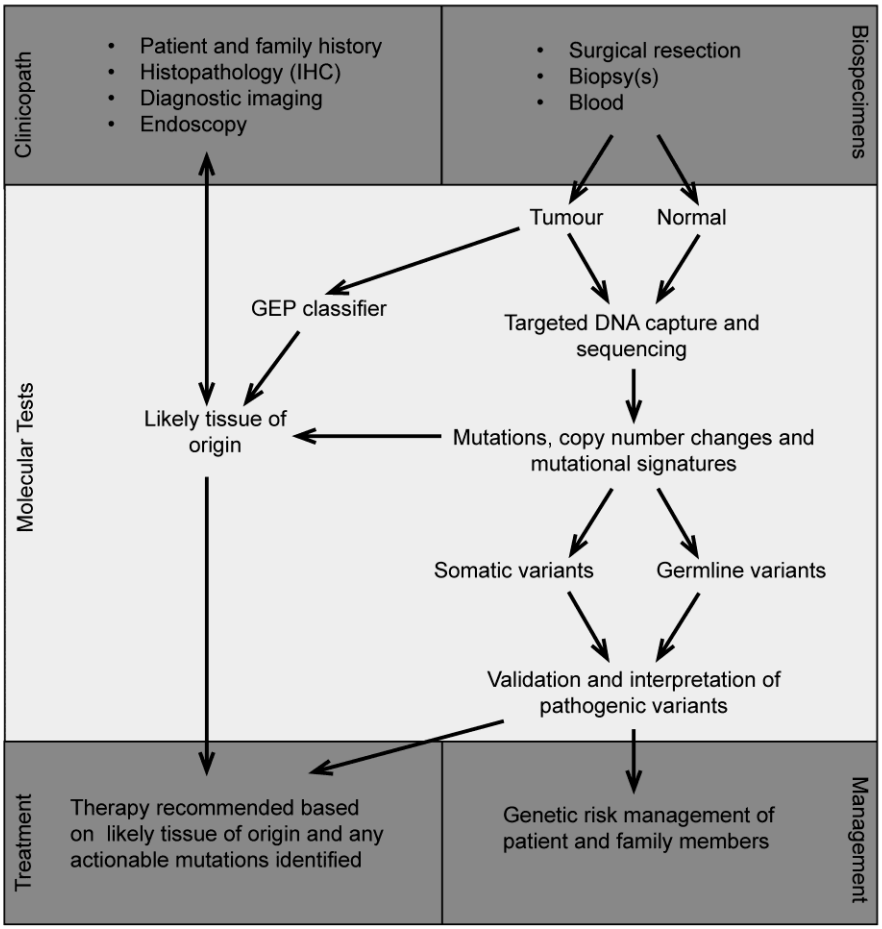

Figure 4