INTEGRATIVE ANALYSIS OF TWO CELL LINES DERIVED FROM A NON-SMALL-LUNG CANCER PATIENT – A PANOMICS APPROACH

OLEG MAYBA[1*], FLORIAN GNAD[1*], MICHAEL PEYTON[4*], FAN ZHANG[3], KIMBERLY WALTER[2], PAN DU[1], MELANIE A. HUNTLEY[1], ZHAOSHI JIANG[1], JINFENG LIU[1], PETER M. HAVERTY[1], ROBERT C. GENTLEMAN[1], RUIQIANG LI[3], JOHN D. MINNA[4], YINGRUI LI[3], DAVID S. SHAMES[2], ZEMIN ZHANG[1#]

Departments of [1]Bioinformatics and Computational Biology and [2]Development Oncology Diagnostics, Genentech, Inc., South San Francisco, CA 94080, USA

[3]BGI-Shenzhen, Shenzhen 518083, China

[4]Hamon Center for Therapeutic Oncology Research, UT-Southwestern Medical Center, Dallas, TX 75390, USA

[*] These authors contributed equally to this work.

[#] To whom correspondence should be addressed: Zemin Zhang (zemin@gene.com)

Cancer cells derived from different stages of tumor progression may exhibit distinct biological properties, as exemplified by the paired lung cancer cell lines H1993 and H2073. While H1993 was derived from chemo-naive metastasized tumor, H2073 originated from the chemo-resistant primary tumor from the same patient and exhibits strikingly different drug response profile. To understand the underlying genetic and epigenetic bases for their biological properties, we investigated these cells using a wide range of large-scale methods including whole genome sequencing, RNA sequencing, SNP array, DNA methylation array, and de novo genome assembly. We conducted an integrative analysis of both cell lines to distinguish between potential driver and passenger alterations. Although many genes are mutated in these cell lines, the combination of DNA- and RNA-based variant information strongly implicates a small number of genes including *TP53* and *STK11* as likely drivers. Likewise, we found a diverse set of genes differentially expressed between these cell lines, but only a fraction can be attributed to changes in DNA copy number or methylation. This set included the ABC transporter *ABCC4*, implicated in drug resistance, and the metastasis associated *MET* oncogene. While the rich data content allowed us to reduce the space of hypotheses that could explain most of the observed biological properties, we also caution there is a lack of statistical power and inherent limitations in such single patient case studies.

1. Introduction

Cancer arises as a result of genomic or epigenomic alterations that change a wide range of cellular processes, leading to uncontrolled tumor cell proliferation and other tumor-specific characteristics (1). Cytotoxic agents and targeted therapies have been developed to treat cancer patients. However, one major challenge during treatment is the potential development of drug resistance (2). Lung cancer, the leading cause of cancer-related death (3), is one of the most heterogeneous of cancer types in terms of underlying molecular characteristics and therapy response. It is biologically and clinically important to understand the underlying genetic lesions influencing cancer cell behaviors such as differential drug response. Recent advances in high-throughput sequencing allow the elucidation of genomewide patient-specific molecular profiles that reveal individual tumor drivers and form the basis for personalized treatments (4). However, most

identified genetic variation is usually difficult to interpret, as the vast majority of alterations are passenger mutations. In addition, not all genomic features can be obtained by a single technology. Integrative approaches have the potential to capture the combination of patient-specific characteristics on various levels for a better understanding and targeting of the molecular basis of specific cancers – a rising field termed "panomics". It is however not clear if comprehensive and deep analyses of a small number of patients, or single patients, might reveal new insights of the genetic basis of patient phenotype.

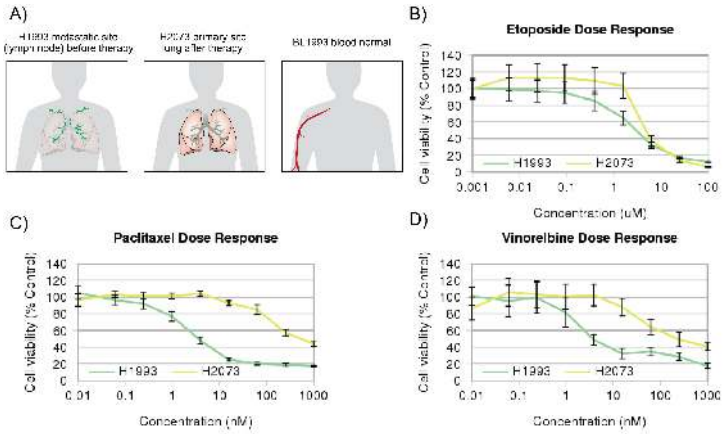


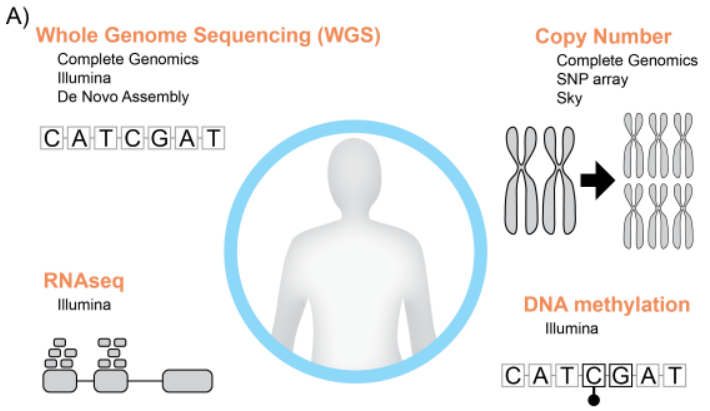
Figure 1. Sample description and cytotoxic drug resistance of H2073

A) Cell line H1993 was derived from a metastatic site in patient's lymph nodes, while H2073 originated from the primary lung tumor after treatment with cisplatin and etoposide. BL1993 was derived from lymphoblastoid cells of the same patient, thus representing the matched normal blood sample. (**B-D**) In comparison to H1993, H2073 cells show higher viability upon treatment with etoposide, paclitaxel and vinorelbine. Error bars indicate standard deviation.

In this study, we performed a wide spectrum of genomic analyses to study a lung cancer patient, who underwent chemotherapy but relapsed with tumor regrowth at the primary site. Two cell lines were derived from this patient: one from a lymph node metastasis isolated prior to chemotherapy, and the other from the lung tumor regrowth months after chemotherapy. Although derived from the same individual, these two cell lines have distinct drug response profiles. To understand the underlying genetic basis for their phenotypic differences, we performed whole genome sequencing, transcriptome sequencing, SNP array, DNA methylation array, and de novo whole genome assembly to thoroughly interrogate genetic and epigenetic events. We conducted an integrative analysis of both cell lines and constructed a model that might explain the development of the patient's cancer and drug resistance after chemotherapy.

2. Sample Description, Drug Response and Screening Overview

Cell line H1993 was derived from the lymph nodes of a 47 year old female Caucasian with history of smoking and diagnosed with non-small cell lung cancer in 1988 (Figure 1A). After treatment with cisplatin and etoposide, H2073 was derived from the resected lung tumor of the same patient. We performed drug response studies as previously described (5). As expected, H2073 shows resistance to etoposide (Figure 1B). Interestingly, the spectrum of drug resistance of H2073 cells encompasses a broader range of therapeutics including paclitaxel and vinorelbine (Figures 1C-D), which target mitotic division.



B)

		H1993	H2073	BL1993
WGS coverage	Complete Genomics	36.4x	60.2x	-
	Illumina	31.3x	37.1x	17.9x
RNAseq reads	Illumina	170×10^6	119×10^6	-

C)

	H1993	H2073
SNVs (total)	3,352,205	3,454,322
Non-synonymous somatic mutations	362	355
Indels (total)	553,115	637,227
Somatic indels in gene coding regions	46	95
Somatic indels in gene coding regions identified by de novo assembly only	1	5
Genes with significantly higher expression	855	1,668
Expressed genes with copy number amplification	100	114
Expressed genes with decreased DNA methylation	68	325
Genes with significantly higher expression and copy number amplification	46	39
Genes with significantly higher expression and decreased DNA methylation	43	146
Interchromosomal structural variations	123	1186
Interchromosomal structural variations (in exonic regions)	87	109
Intrachromosomal structural variations	383	13917
Intrachromosomal structural variations (in exonic regions)	177	1159

Figure 2. Integrative analysis of H1993 and H2037: a panomics approach

A) We applied an integrative analysis of H1993 and H2073 based on whole genome sequencing, RNA sequencing, DNA methylation quantification and copy number investigation. **B**) The genome of each cell line was sequenced at minimum 30x coverage. **C**) The panomics approach allowed us to analyze the landscape of single nucleotide variants, indels, differential gene expression, copy number changes, and structural variations. The numbers of detected aberrations are shown for these two cell lines.

To elucidate the development of the patient's cancer and to understand the drug resistance after chemotherapy, we applied an integrated analysis of somatic exonic mutations, messenger RNA sequencing, DNA copy number, and promoter DNA methylation (Figure 2A).

Whole genome sequencing (WGS) of both cell lines was conducted on two independent platforms: Complete Genomics (CG) and Illumina, to a minimum depth of 30x (Figure 2B). In addition, we constructed DNA libraries with variable insert sizes for both cell lines, performed Illumina-based paired-end sequencing, and used the resulting reads for de novo genome assembly, in order to identify genomic features missed by reference-based approaches. We also carried out Illumina WGS on DNA isolated from BL1993, a lymphoblastoid cell line from the same patient, representing the matched normal blood sample.

To identify genes differentially expressed between H1993 and H2073, we collected RNA-Seq data in 3 replicates. DNA methylation was measured by Illumina Infinium array, and copy number analysis with the Illumina OMNI 2.5M SNP array, processed by a modified version of the PICNIC algorithm, as previously described (6, 7). Results are summarized in Figure 2C.

3. Mutation Landscapes and the Identification of Expressed Variants

Somatic mutations were identified by comparing the variant calls in H1993 and H2073 with BL1993. We selected non-synonymous mutations with a minimum support of five reads and excluded known germline variants from a variety of sources (see Methods). Any variants listed in COSMIC database of somatic mutations in cancer (8) were retained. This resulted in 313 somatic non-synonymous single-base substitutions in common between H1993 and H2073, of which 290 were missense mutations, 21 resulted in stop gain, and 2 resulted in stop loss. Consistent with the patient's smoking history, we observed an enriched fraction of C:G > A:T transversions, the smoking-related mutation signature, in the tumor-specific variants (data not shown).

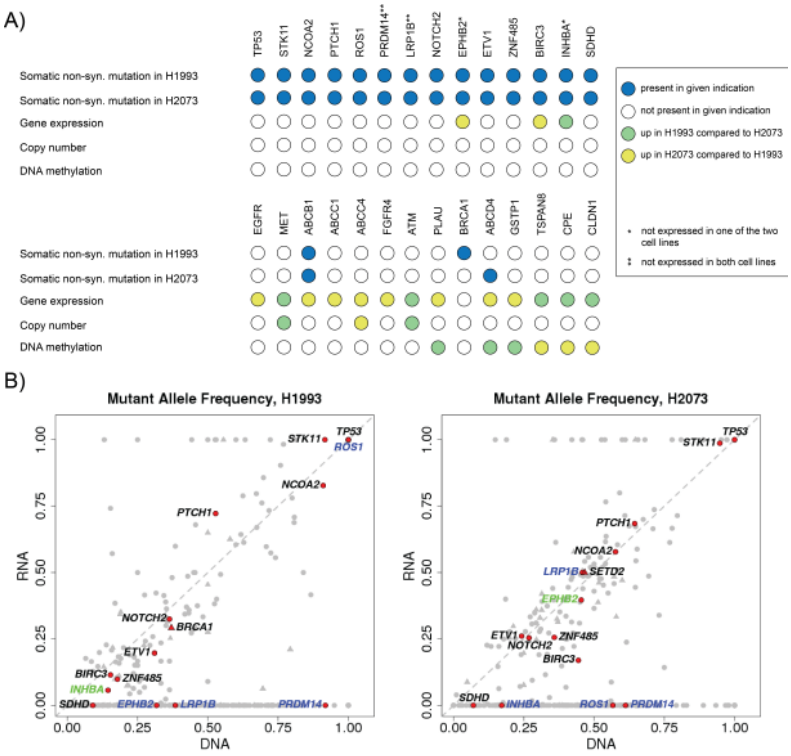


Figure 3. Genomic landscapes and pathway alterations of H1993 and H2073

A) Multiple cancer related genes were somatically mutated in both cell lines (upper panel) or differentially expressed between the cell lines (lower panel). **B**) Integrating gene expression and focusing on instances of high mutant allele frequency enabled us to substantially reduce the set of candidate drivers. Known cancer related genes are highlighted. Genes with low expression (<0.5 RPKM) in both cell lines are shown in blue, while genes with low expression in one cell line are shown in green. Triangles indicate cell line-specific mutations, while circles correspond to common mutations.

Both cell lines harbor non-synonymous mutations in genes known to be altered in lung cancer, including *TP53, STK11, EPHB2, LRP1B, INHBA*, *ZNF458,* and *PRDM14* (Figure 3A). Other somatically mutated cancer genes, which are listed in the Cancer Gene Census (CGC) (9), include *NOTCH2, BIRC3, PTCH1, ETV1, ROS1, SDHD* and *NCOA2*. To prioritize these putative drivers, we used RNA-Seq to eliminate genes with little or no expression (RPKM<0.5). This expression based filtering reduced the number of common mutations from 313 to 106 (96 missense, 10 stop gain), eliminating a large fraction of candidate genes at the risk of possibly discarding low expressed drivers (Figure 3B).

We further hypothesized that the mutant alleles for driver mutations should be selected for, leading to higher mutant allele frequencies for driver genes. We then assessed mutant allele frequencies in DNA and RNA data and grouped the mutations into three frequency classes (Figure

3B): high (>0.9, class 1), medium (0.3 to 0.9, class 2), and low/inconsistent (class 3). Mutations at loci with a total DNA or RNA read coverage < 10 were also assigned to class 3. Class 1 comprised only 10 genes, 8 of which had stop gain or missense mutations that were predicted to be deleterious (10) based on Polyphen2 (11) and SIFT (12) calculations. In this reduced set of candidate drivers were tumor protein 53 (*TP53*) and serine/threonine kinase 11 (*STK11*, also known as *LKB1*), the two most significantly mutated tumor suppressors in lung cancer (13). Both mutations were observed in regions with loss of heterozygosity. The homozygous *TP53* missense mutation C242W was also observed in other cancer types including breast (14) and stomach (15) cancer, while the homozygous stop gain mutation on position 199 within the kinase domain of *STK11* has been previously reported in other lung cancer samples (16). Thus, integrating WGS and RNA-Seq data on the two cell lines allowed us to reduce a set of non-synonymous mutations to two likely drivers of oncogenesis in this patient.

While 106 non-synonymous mutations in expressed genes were common to both cell lines, 20 and 22 were specific to H1993 and H2073, respectively. These included Cancer Gene Census genes *SETD2* (class 2) in H2073, and *BRCA1* (class 2) in H1993. Inactivation of *BRCA1* is associated with tumor aggressiveness and invasion (17), consistent with the metastatic state of H1993. None of the cell line specific mutations was assigned to class 1. Overall, the limited difference between H1993 and H2073 mutation profiles indicates that unique point mutations are unlikely to explain the phenotypical variations between them.

Among 138 somatic coding indels detected in either cell line, 7 affected Cancer Gene Census genes. All of these were cell line-specific, with frame-shifting indels observed in genes *SF3B1*, *BMPR1A*, and *GPHN* in H1993, and in genes *JUN*, *MLL3*, *NR4A3* in H2073. We also observed an in-frame insertion in gene *MLL2* in H2073. It is unclear what role, if any, these mutations may play in the observed phenotypic differences between the two cell lines. While histone methyltransferases *MLL2* and *MLL3* have been linked to *TP53*-mediated DNA damage response pathway (18, 19), our cell lines exhibited lack of a functional copy of *TP53*, rendering any additional mutations to this pathway inconsequential.

4. Differentially Expressed Genes and the Relationship with DNA Changes

Our RNA-Seq analysis identified 2,523 differentially expressed genes between H1993 and H2073 (Figure 4A), of which 1,668 (67%) were over-expressed in H2073. Classical markers for epithelial/mesenchymal status, including *CDH1, CDH2, VIM* and *FN1*, were not consistently differentially expressed between the two cell lines, suggesting that the observed differences between the primary and the metastatic cell line were not due to epithelial-to-mesenchymal transition.

The large number of differentially expressed genes also suggests that most of these expression changes are downstream effects of the causal events. We hypothesized that the primary expression differences should have certain degree of genetic or epigenetic basis. We therefore focused on differentially expressed genes that can be directly attributed to changes in copy number or DNA methylation state. We found that 39 out of 1,668 (2.3%) genes overexpressed in H2073 are in regions amplified in H2073 relative to H1993 (ploidy adjusted CN fold change >=2). Ploidy adjustment was carried out because H1993 is mostly tetraploid, while H2073 has average ploidy

between 2 and 3, consistent with cytogenetic results (data not shown). Similarly, we observed that 46 out of 885 (5.4%) genes overexpressed in H1993 are in genomic regions amplified in H1993 relative to H2073 (Figure 4B).

Overall, regions amplified in H1993 and H2073 contained 100 and 114 expressed genes, respectively, out of which 46 (46%) and 39 (34%) were overexpressed according to our cutoffs, exhibiting higher rate of overexpression events than non-amplified regions (Figure 4C, Fisher exact test p-value $<7x10^{-9}$ for both cell lines). In total, we identified seven amplified regions in either cell line longer than 1 Mb, six of which (three in each cell line) accounted for 82 out of 85 differentially expressed genes with underlying CN changes. One of the H2073 amplicons included transporter gene *ABCC4*, previously implicated in drug resistance and showing 3-fold overexpression in H2073. The region on chromosome 7, highly amplified (>10 copies) in H1993, contained oncogene *MET* (Figure 5A), which is known to be involved in tumor cell invasion and metastasis (20). We found *MET* to be 7-fold overexpressed in H1993, consistent with the metastatic character of H1993. The dependence of H1993 on MET is confirmed by its low viability in the presence of MET inhibitors (Figure 4D). Another highly amplified genomic region was located on chromosome 11 and contained the oncogene *ATM* (4-fold overexpression in H1933), which was also reported to promote metastasis (21).

Comparing the two cell lines further, we found that 427 genes expressed in at least one cell line showed differentially methylated regions (DMRs) within 2kb of their transcription start site (TSS). Out of 1,668 genes overexpressed in H2073, 166 (9.9%) contained DMRs (Figure 4B). In 146 cases (82%), the extent of methylation was higher in H1993, consistent with down-regulation of

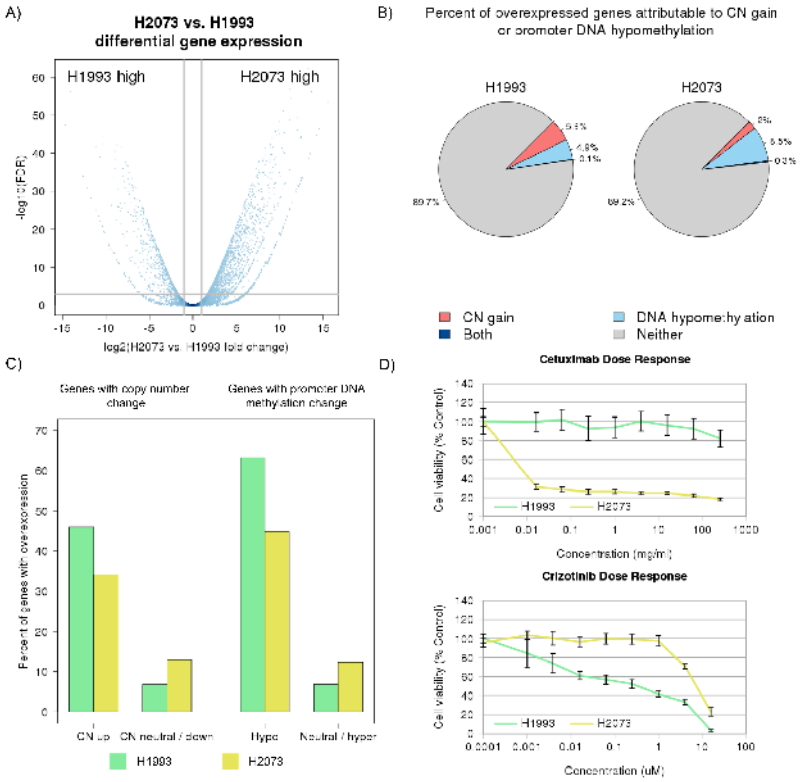


Figure 4. Differential gene expression analysis between H1993 and H2073

A) Volcano plot illustrating fold changes and false discovery rates for all human genes as calculated by differential gene expression analysis. **B**) Percentage of overexpressed genes with significant copy number gain or DNA hypomethylation. **C**) The sets of expressed genes with copy number amplification or promoter DNA hypomethylation were enriched for overexpressed genes. **D**) Treating both cell lines with an EGFR inhibitor Cetuximab reveals lower viability of H2073 in comparison to H1993. Treating the two cell lines with a MET inhibitor Criozotinib reveals lower viability in H1993. Error bars indicate standard deviation.

expression by hypermethylation. In comparison, 61 out of 885 (6.9%) genes overexpressed in H1993 contained DMRs within 2kb of TSS, with 43 (70%) exhibiting higher methylation in H2073. In total, hypomethylated DMRs were associated with 68 and 325 genes in H1993 and H2073, respectively, out of which 43 (63.2%) and 146 (44.9%) showed overexpression, exhibiting higher rate of overexpression events than hypermethylated or non-differentially methylated regions (Figure 4C, Fisher exact test p-values < $2x10^{-32}$ for both cell lines). Several of the genes with overexpression and promoter DNA hypomethylation in H2073 have been implicated in apoptosis evasion and drug resistance, including *PLAU* (Figure 5B), *SNCG, BNIP3, GSTP1, ETS1*, and *MSLN*. Interestingly, we found the binding partners *PLAU* and *PLAUR* to be overexpressed in H2073, suggesting co-regulation of their expression. Binding of *PLAU* to *PLAUR* can activate the *ERK* pathway and contribute to cancer development (22).

Genes overexpressed and hypomethylated in H1993 included the metastasis effectors *RAB25*, *TSPAN8*, and *CPE*, as well as *CLDN1*, whose up-regulation has been associated with cisplatin sensitivity (23), consistent with cisplatin resistance in H2073. Overall, 10.8% of genes overexpressed in H2073 and 10.3% of genes overexpressed in H1993 are associated with either differential DNA methylation or copy number re-arrangements. Thus, the integration of these two additional data types allowed us to substantially reduce the number of candidate drivers, while possibly omitting driver genes activated via alternative mechanisms.

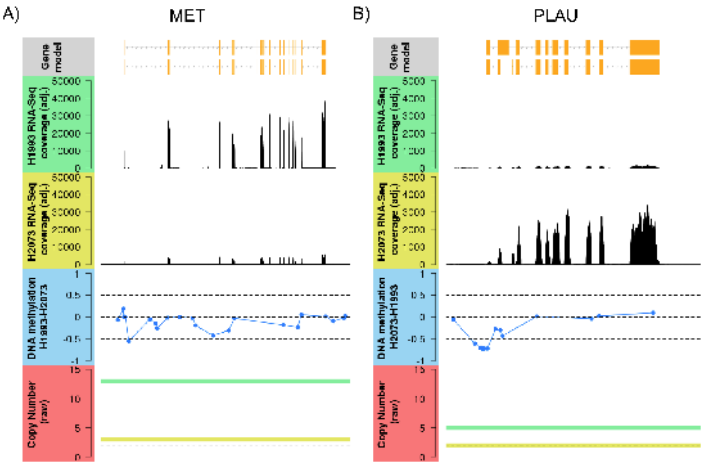


Figure 5. Overexpression of MET in H1993 is associated with copy number gain (A), while overexpression of PLAU in H2073 is associated with decreased promoter methylation (B).

The panels show gene structure (top, individual transcript isoforms), expression normalized by sequencing depth (H1993: second from the top, H2073: third from the top), difference in DNA methylation (second from bottom, dashed lines correspond to differences of 0.5,0, and -0.5), and raw copy number (bottom, green line: H1993, yellow line : H2073, dashed black like: CN=2 (baseline)).

Guided by our discovery of the amplification of transporter gene *ABCC4* in the drug-resistant cell line H2073, we tested for differential expression of other transporter genes. While one transporter gene, *ABCB10*, showed overexpression in H1993, several others were overexpressed in H2073 and are known to play a role in drug resistance. We found that the multi-drug resistance transporter *MDR1/ABCB1* was expressed in H2073 but almost absent from H1993. Both *ABCC4* and *ABCC1*, also implicated in drug resistance, were also at least 3-fold overexpressed in H2073 (24, 25). Furthermore we found 9-fold higher expression of *FGFR4* in H2073. A recent study reported that inhibition of *FGFR* reverses *ABCB1*-mediated drug resistance in cancer (26). Overall, these results suggest an efflux-based drug resistance mechanism developed

by H2073, which involves *ABCC4*, *ABCB1*, and possibly other transporter proteins that were not over-expressed in H2073 based on our cutoffs.

Integrating information on changes in DNA copy number and methylation allowed us to reduce a large set of differentially expressed genes 10-fold to candidate drivers with clear underlying mechanism of differential expression. Close examination of these candidate drivers revealed a number of genes overexpressed in H1993 and known to be involved in metastasis. This allowed us to construct a drug resistance model for H2073. However, this reductionist approach has its limitations, as not all meaningful differential expression can be attributed to a change in either DNA copy number or methylation. As an example, the expression of the well-known cancer gene *EGFR* is 8-fold higher in H2073 than in H1993, and the dependence of H2073 on *EGFR* for survival and proliferation is strongly suggested by its higher sensitivity to *EGFR* inhibitors (Figure 4D). However, the observed overexpression of *EGFR* was not associated with either a copy number change or differential promoter DNA methylation in this study. It is likely that other types of genetic or epigenetic alterations, such as histone mark changes, are responsible for the observed *EGFR* expression change but are not captured by our existing assays.

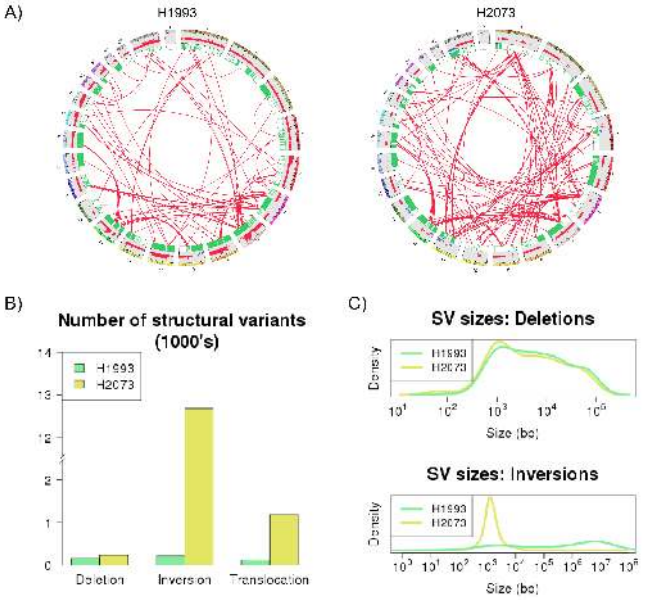


Figure 6. Structural variations in H1993 and H2073

A) Illustration of genomic alterations in H1993 and H2073 using Circos plots. Candidate interchromosomal structural variations identified by the Complete Genomics Platform are shown as red lines. Copy number changes detected by Illumina SNP arrays are illustrated as bar plots. Loss of heterozygosity regions are shown in green. **B**) Structural variations, in particular smaller inversions (**C**), were more frequent in the cell line derived after chemotherapy (H2073)

5. Structural Variation Analysis

Based on WGS by the Complete Genomics platform, we observed 164 large deletions (50bp-100kb), 219 inversions, and 123 translocations in H1993, supported by at least 5 reads (Figure 6A-B). H2073 showed substantially more structural variants, with 237 large deletions, 13680 inversions, and 1186 translocations. This significant increase in the number of structural variants, in particular short inversions (Figure 6C), might be due to the stress imposed on the cell by the chemotherapy (27). This is consistent with the fact that H1993 was derived from tumor cells prior to chemo-treatment, while H2073 was derived afterward and therefore is chemo-resistant.

6. De Novo Genome Assembly Reveals Additional Variant Information

To discover genomic alterations that might be missed by standard WGS analysis, we performed de novo assembly of H1993 and H2073 genomes, based on paired-end Illumina sequences. The insert

size ranged from 200bp to 40kb, in order to aid longer range DNA assembly. The resulting assembled sequences span 2.96 (H1993) and 2.89 (H2073) Gb, including 2.29 and 2.48 Gb of fully resolved (non-gapped) sequence. The N50 values were 1.9 Mb and 1.26 Mb, respectively, reflecting a large portion of the sequence in scaffolds of substantial (>1Mb) size.

We aligned the assembled sequences to the reference genome to identify insertions or deletions, which may have been missed by resequencing-based approaches. We identified 2 insertions and 3 deletions that were exclusively detected by the assembly approach and that affected exons (Table 1). These indels ranged in size from 51 to 123 bp, indicating the utility of the assembly approach in detecting medium size indels, that are not short enough to be detected by most resequencing-based indel callers, but are not long enough to be detected by the copy number or structural variation analyses. We note that the observed frame-shifting deletion in TSPAN8 in H2073 may have contributed to its lower expression in that cell line, alongside the hypermethylation component, described above.

Table 1. Assembly-specific exonic indels.

Indel Type	Coordinate	Length (bp)	Affected gene	Cell line
Deletion	Chr1:7,889,973-7,890.026	54	PER3	Both
Deletion	Chr2:27,324,254-27,324,304	51	CGREF1	H2073
Insertion	Chr12:71,523,133-71,523,134	109	TSPAN8	H2073
Deletion	Chr14:104,645,583-104,645,705	123	KIF26A	H2073
Insertion	Chr20:62,196,017-62,196,018	57	PRIC285	H2073

7. Conclusions

The expansion of high-throughput assays for analyzing cellular states has provided new opportunities for integrative analyses. Here we used several genome-scale analyses of 2 cancer cell lines to ask whether we could better explain their observed biological similarities and differences. Perhaps the most significant challenge in interpreting genomic data is to pinpoint the most relevant genomic changes from a large collection of data points, and the panomics approach by definition epitomizes this problem. While it might be practically impossible to achieve statistical significance for such panomics approaches, we believe that prior knowledge and logical combination of different data could dramatically reduce the search space and propose biologically meaningful models.

In this study, while variant analysis revealed more than 300 non-synonymous mutations, combining this analysis with expression data reduced the number of candidate drivers 3-fold. Integrating allele frequencies on both DNA and RNA levels further reduced the focal set to 8 homozygously mutated genes, including likely drivers *TP53* and *STK11*. Similarly, while expression analysis alone revealed thousands of differentially expressed genes between the two cell lines, only a small fraction of such genes were associated with the underlying genetic and epigenetic changes. Among these small number of genes, *MET* was present in a highly amplified region and showed 7-fold overexpression in H1993, and *ABCC4* was amplified and overexpressed in the drug resistant cell line H2073. Although we could not exclude other genomic changes that

might also explain the phenotypic differences between these two cell lines, our integrated analyses readily produced a working model that is consistent with our knowledge of the samples.

It is worth noting that although H1993 and H2073 have been independently cultured ex vivo for decades, they show remarkable similarity and display largely overlapping point mutations. This shows that any new mutations acquired during the cell culturing steps are at the minimum if they exist. This finding boosts the validity of these cell lines as stable model systems for cancer studies. From the technology point of view, our de novo assembly of both cell lines revealed a number of additional insertions and deletions, missed by the reference-based assembly. Only 5 of these altered protein coding regions, indicating that reference-based assembly captures most of the actionable variants.

It should also be noted that this study is exploratory by nature. With such small sample size, the statistical power is nonexistent, so it is currently impossible to draw any causal relationships with any confidence. This approach should be viewed as a hypothesis generating method. Alternatively, this approach can be viewed as a "hypothesis-supporting method". Our current knowledge of lung cancer and drug resistance has led us to propose genes like *EGFR, MET*, and *ABCC4* as functionally relevant culprits in these cell lines, but an improved knowledge in the field might implicate a different set of genes. It is therefore necessary to view the panomics data with a grain of salt, as the interpretation of these data can be influenced by the current biological knowledge. Nevertheless, the maturation of this field will enhance our ability to better analyze panomics data, as no single assay can provide a full picture of the cell state or to point in the direction of possible therapeutic actions.

8. Materials and Methods

8.1. *Whole Genome Sequencing and Variant Calling*

Whole genome sequencing (WGS) of H1993 and H2073 was performed by Complete Genomics, as described (7). Independently, WGS of H1993, H2073, and BL1993 was performed by Illumina sequencing (100bp paired-end reads), using libraries with insert sizes of 200, 500, 2000, 5000, 10000, 20000, and 40000 bp. Reads where aligned to reference human genome (hg19) using BWA (28). Single nucleotide variants (SNV) and indels were called by the Complete Genomics WGS processing pipeline. Several variant callers were applied to Illumina WGS data. We used SOAPsnp (29) to identify germline SNVs in all 3 cell lines, and VarScan (30) to identify cell line-specific SNVs in every possible 2-cell line comparison. Only variants supported by 5 or more reads and separated by 10 or more base pairs from the nearest variant were retained. Somatic mutations were identified by requiring that no variant-supporting reads be detected in BL1993 WGS. Unless the variant was listed in COSMIC database of cancer mutations, we further required that it was covered by 10 or more reads in BL1993, and not present in dbSNP (v.132) (31) or among variants from 1000 Genomes Project (32), 6515 normal exomes published by NHLBI (33), or 69 normal genomes sequenced by Complete Genomics and made available to the public (34). We used Dindel (35) to identify germline indels and GATK (36) to identify cell line-specific indels. Indels were declared somatic if no compatible indel was detected in BL1993 by Dindel, and if the indel was not part of a set of known normal indels obtained from 1000 Genomes Project

and 69 publicly available Complete Genomics sequenced normal genomes. Structural variants were obtained from the Complete Genomics pipeline.

8.2. *Copy Number Analysis*

H1993 and H2073 cell lines were assayed with Illumina OMNI 2.5M SNP array and processed with a modified version of PICNIC (7). When calculating copy number fold change between the two cell lines, adjustment was made for average cell line ploidy. This was calculated as the average copy number per base pair, and was 3.9 for H1993 and 2.4 for H2073.

8.3. *Messenger RNA Sequencing*

Three temporally separate, standard RNA-seq library preparations and subsequent sequencing data were collected for each of the two cell lines. One of the libraries for each cell line was sequenced on an Illumina GAII, while the remaining libraries were sequenced on an Illumina HiSeq machine. The resulting RNA-seq data was filtered for read quality, ribosomal RNA contamination, and then aligned to the human reference genome (NCBI Build 37) using the GSNAP alignment tool (37). Alignments were permitted a maximum of 3 mismatches per 75 base pair sequence and used the following GSNAP parameters: "-M 2 -n 10 -B 2 -i 1 -N 1 -w 200000 -E 1 --pairmax-rna=200000". These steps, and the downstream processing of the resulting alignments to obtain read counts and RPKMs per gene (over coding exons of RefSeq gene models) per replicate are implemented in the Bioconductor package, HTSeqGenie (v 3.10.0) (38).

We compared the gene expression profiles of the two cell lines using the gene count data described above, and the Bioconductor package edgeR (39). Each of the three temporally separate RNA-seq libraries per cell line was used as biological replicates for dispersion estimates within edgeR. Genewise exact tests for differential gene expression were performed, and resulting summary statistics reported. We used the cutoffs of FDR<0.001, fold change > 2, and RPKM >=0.5 (in cell line with overexpression) to declare differential expression.

8.4. *DNA Methylation Analysis*

DNA methylation was measured by Illumina Infinium Human Methylation 450K BeadChips and preprocessed using the Bioconductor lumi package (40). Methylation status was measured in beta-values ranging from 0 (unmethylated) to 1 (methylated). A probe was considered to show significant methylation change, if the difference between H1993 and H2073 beta-values was larger than 0.5. Nearby (less than 2kb) differentially methylated probes were merged into differentially methylated regions (DMR). Final differential methylation calls were based on DMRs near gene transcription starting site (TSS) (2kb upstream of TSS or overlapping with the first exon of the gene) with a minimum of two supporting probes.

8.5. *Data Access*

The results of Complete Genomics WGS and copy number SNP array assays have been previously published (7). The remaining data will be made available to the public and the repository location and accession information can be obtained from the authors.

Acknowledgments

We thank Jens Reeder and Gregoire Pau for development of the transcriptome sequencing analysis pipeline, Allison Bruce for help with the design of figure graphics, and Gerard Manning for critical review of this manuscript.

References

1. Vogelstein B, Kinzler KW. 2004. *Nat Med* 10: 789-99
2. Longley DB, Johnston PG. 2005. *J Pathol* 205: 275-92
3. Siegel R, Naishadham D, Jemal A. 2012. *CA Cancer J Clin* 62: 10-29
4. Roychowdhury S, Iyer MK, Robinson DR, et al. 2011. *Sci Transl Med* 3: 111ra21
5. Gandhi J, Zhang J, Xie Y, Soh J, Shigematsu H, et al. 2009. *PLoS One* 4: e4576
6. Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, et al. 2012. *Genome Res* 22: 593-601
7. Liu J, Lee W, Jiang Z, Chen Z, Jhunjhunwala S, et al. 2012. *Genome Res* 22: 2315-27
8. Forbes SA, Bhamra G, Bamford S, et al. 2008. *Curr Protoc Hum Genet* Chapter 10: Unit 10 1
9. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. 2004. *Nat Rev Cancer* 4: 177-83
10. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. 2013. *BMC Genomics* 14 Suppl 3: S7
11. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, et al. 2010. *Nat Methods* 7: 248-9
12. Kumar P, Henikoff S, Ng PC. 2009. *Nat Protoc* 4: 1073-81
13. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, et al. 2008. *Nature* 455: 1069-75
14. Chevillard S, Lebeau J, Pouillart P, de Toma C, et al. 1997. *Clin Cancer Res* 3: 2471-8
15. Kubicka S, Claas C, Staab S, Kuhnel F, Zender L, et al. 2002. *Dig Dis Sci* 47: 114-21
16. Imielinski M, Berger AH, Hammerman PS, Hernandez B, et al. 2012. *Cell* 150: 1107-20
17. Albiges L, Andre F, Balleyguier C, Gomez-Abuin G, et al. 2005. *Ann Oncol* 16: 1846-7
18. Guo C, Chang CC, Wortham M, Chen LH, et al. 2012. *Proc Natl Acad Sci U S A* 109: 17603-8
19. Lee J, Kim DH, Lee S, Yang QH, Lee DK, et al. 2009. *Proc Natl Acad Sci U S A* 106: 8513-8
20. Jeffers M, Rong S, Vande Woude GF. 1996. *J Mol Med (Berl)* 74: 505-13
21. Sun M, Guo X, Qian X, Wang H, Yang C, et al. 2012. *J Mol Cell Biol* 4: 304-15
22. Nguyen DH, Hussaini IM, Gonias SL. 1998. *J Biol Chem* 273: 8502-7
23. Fortier AM, Asselin E, Cadrin M. 2013. *J Biol Chem* 288: 11555-71
24. Nath S, Daneshvar K, Roy LD, Grover P, Kidiyoor A, et al. 2013. *Oncogenesis* 2: e51
25. Oprea-Lager DE, Bijnsdorp IV, RJ VANM, AJ VDE, et al. 2013. *Anticancer Res* 33: 387-91
26. Patel A, Tiwari AK, Chufan EE, et al. 2013. *Cancer Chemother Pharmacol* 72: 189-99
27. Portugal J, Mansilla S, Bataller M. 2010. *Curr Pharm Des* 16: 69-78
28. Li H, Durbin R. 2009. *Bioinformatics* 25: 1754-60
29. Li R, Li Y, Fang X, Yang H, Wang J, et al. 2009. *Genome Res* 19: 1124-32
30. Koboldt DC, Chen K, Wylie T, Larson DE, et al. 2009. *Bioinformatics* 25: 2283-5
31. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. 2001. *Nucleic Acids Res* 29: 308-11
32. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. 2012. *Nature* 491: 56-65
33. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, et al. 2013. *Nature* 493: 216-20
34. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. 2010. *Science* 327: 78-81
35. Albers CA, Lunter G, MacArthur DG, McVean G, et al. 2011. *Genome Res* 21: 961-73
36. McKenna A, Hanna M, Banks E, Sivachenko A, et al. 2010. *Genome Res* 20: 1297-303
37. Wu TD, Nacu S. 2010. *Bioinformatics* 26: 873-81
38. Pau G, Reeder J. 2012. *R package* R package version 3.10.0
39. Robinson MD, McCarthy DJ, Smyth GK. 2010. *Bioinformatics* 26: 139-40
40. Du P, Kibbe WA, Lin SM. 2008. *Bioinformatics* 24: 1547-8