Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes

Tae-Min Kim,1 Ruibin Xi,1,2 Lovelace J. Luquette,1 Richard W. Park,1 Mark D. Johnson,3 and Peter J. Park1,4,5,6

1Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA; 2School of Mathematical Sciences and Center for Statistical Science, Peking University, 100871 China; 3Department of Neurosurgery, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA; 4Children’s Hospital Informatics Program, Boston, Massachusetts 02115, USA; 5Division of Genetics, Brigham and Women’s Hospital, Boston, Massachusetts 02115, USA

A large database of copy number profiles from cancer genomes can facilitate the identification of recurrent chromosomal alterations that often contain key cancer-related genes. It can also be used to explore low-prevalence genomic events such as chromothripsis. In this study, we report an analysis of 8227 human cancer copy number profiles obtained from 107 array comparative genomic hybridization (CGH) studies. Our analysis reveals similarity of chromosomal arm-level alterations among developmentally related tumor types as well as a number of co-occurring pairs of arm-level alterations. Recurrent (“pan-lineage”) focal alterations identified across diverse tumor types show an enrichment of known cancer-related genes and genes with relevant functions in cancer-associated phenotypes (e.g., kinase and cell cycle). Tumor type-specific (“lineage-restricted”) alterations and their enriched functional categories were also identified. Furthermore, we developed an algorithm for detecting regions in which the copy number oscillates rapidly between fixed levels, indicative of chromothripsis. We observed these massive genomic rearrangements in 1%-2% of the samples with variable tumor type-specific incidence rates. Taken together, our comprehensive view of copy number alterations provides a framework for understanding the functional significance of various genomic alterations in cancer genomes.

Cancer genomes harbor various somatic forms of genetic alterations ranging from nucleotide-level changes (e.g., nucleotide substitutions and small insertions/deletions) (Greenman et al. 2007) to large chromosomal events (e.g., translocations and copy number alterations) (Albertson et al. 2003; Mitelman et al. 2007). As a comprehensive catalog of tumor-related chromosomal alterations can help identify genomic features with potential clinical benefits (Chin and Gray 2008; Meyerson et al. 2010), collecting and profiling tumor samples using genome-wide, high-throughput platforms have been major efforts during the last decade. The resulting accumulation of cancer genome studies has provided new mechanistic insights that aid in a more systematic understanding of human cancer and in identification of molecular targets for cancer therapy (Chin et al. 2011; Hanahan and Weinberg 2011). However, due to the intrinsic complexity and heterogeneity of human cancer genomes, there is still a large number of unresolved issues. For example, a majority of cancer-related genomic alterations are believed to be “passengers” that arise as a by-product during cancer genome evolution, without obvious advantage for the affected clones. The discrimination of such passenger alterations from “driver” alterations that contribute to tumor initiation and/or progression remains a difficult task.

In the past few years, array-based comparative genomic hybridization (array-CGH) has become a dominant tool for genome-wide detection of copy number changes in cancer (Pinkel et al. 1998; Snijders et al. 2001). It has been applied to a wide range of tumor types, with notable success in subtype classification and biomarker screening (Albertson and Pinkel 2003; Pinkel and Albertson 2005). Most array-CGH studies performed so far, however, focus on a specific tumor type with a limited number of samples. Some efforts have been made to construct a large-scale array-CGH database (Baudis and Cleary 2001; Beroukhim et al. 2010; Cao et al. 2010), with indications that a large compilation of cancer genomes may be advantageous in distinguishing driver alterations from passenger events. In particular, Beroukhim et al. (2010) reported analysis of ~3000 cancer genome profiles that resulted in identification of numerous focal somatic alterations across tumor types, corroborating that a large database can help identify recurrent focal alterations that are likely to be oncogenic drivers. That analysis, however, was based on a single array-CGH platform from Affymetrix with 250K probes (Beroukhim et al. 2010). Considering the growing number of cancer data sets in public databases and continual improvement in platforms, it is imperative that the research community should take advantage of the statistical power and a wide range of tumor types that come with a larger compendium.

In this study, we present extensive computational analysis of 8227 copy number profiles gathered from 107 array-CGH human cancer studies. The data were collected from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/), a public repository of microarray data sets (Barrett et al. 2009). Among the array-CGH studies available in GEO, we collected those profiled by high-resolution, oligonucleotide-based platforms (>100K probes) from two commercial vendors (Affymetrix and Agilent). After normalization and application of segmentation algorithms, the arm-level alteration frequencies across the samples were investigated for lineage-specific patterns in major tumor types and for concordant relationships between alterations. We also delineated the minimal common regions (MCRs) across the entire data set.
and in a given tumor type ("pan-lineage" and tumor type-specific MCRs, respectively). MCR (sometimes called MAR for "minimally altered region") is defined as the minimal region of amplifications or deletions representing a common genomic alteration across the examined cancers (Santarius et al. 2010). MCRs were interpreted by their association with known cancer-related genes (Putreal et al. 2004) and Gene Ontology (GO) functional categories (Ashburner et al. 2000). Finally, we performed a survey of chromothripsis, a massive genomic rearrangement event that is thought to occur in a small fraction of tumors (Stephens et al. 2011). Overall, our study presents a global view of copy number alterations and demonstrates the utility of a large-scale copy number database in prioritizing candidate biomarkers and in exploring unique patterns of chromosomal imbalances such as chromothripsis.

Results

A compendium of copy number profiles from human cancer genomes

We searched the GEO database to obtain data from high-resolution, array-CGH platforms for human cancer samples (Fig. 1A). We focused on the five platforms comprising more than 100K oligonucleotide probes from two commercial vendors (Agilent and Affymetrix): Agilent 244K and Affymetrix 100K, 250K, 500K, and SNP6.0. The number of probes ranged from 115,417 (Affymetrix 100K) to 1.8 million (Affymetrix SNP6.0). From the 107 array-CGH studies based on these platforms, we collected a total of 8227 cancer genome copy number profiles after removing normal controls or duplicates. The sample numbers are shown for individual tumor types and array-CGH platforms in Figure 1B. The frequencies of individual tumor types in our data set were compared with the tumor incidence rates (Supplemental Fig. S1). A substantial level of correlation ($r^2 = 0.44$) was observed, attributable to tumor types with higher incidence rates that are also well-represented in databases, such as breast and lung cancers. The detailed information on the data sets is available in Supplemental Tables S1 and S2. For subsequent analysis, we used the log2 ratio profiles as available in the GEO database for the Agilent 244K platforms ($n = 1750$) and derived log2 ratios (tumor/reference) using HapMap population data as a universal reference for the Affymetrix platform ($n = 6477$) (see Methods for segmentation steps and additional processing). The segmentation profiles of all tumor samples and other analysis files are available at http://compbio.med.harvard.edu/metaCGH/.

In our data set, we observed an average of 79.3 gains and 80.9 losses per sample, involving 9.7% and 11.6% of the reference genome, respectively. Among the five platforms, Affymetrix SNP6.0 (190.5 gains and 233.6 losses per sample) and Agilent 244K platform (108.5 gains and 107.4 losses per sample) showed a substantially higher number of alterations compared to Affymetrix 100K–500K platforms, with smaller alteration sizes on average (Supplemental Fig. S2). The small size and relative abundance of alterations in these platforms may be due to the higher sensitivity with increased resolution of the platforms (Affymetrix SNP6.0) or higher signal-to-noise ratio of longer oligonucleotides (60-bp oligonucleotides of the Agilent platform compared to Affymetrix 25-bp oligonucleotides). It may also be associated with hyper-segmentation, which has been previously observed in regions with extreme copy numbers due to different attenuation curves of neighboring probes (Beroukhim et al. 2007; The Cancer Genome Atlas Research Network 2008).

Chromosomal arm-level alterations in human cancer genomes

We defined chromosomal arm-level alteration as a single alteration or an aggregate of alterations that encompass >50% of a chromosomal arm. The arm-level alteration frequency measured across the entire data set ($n = 8227$) (Fig. 2A) highlights the frequent gains at 1q, 5p, 7p, 8q, and 20q along with frequent losses at 4q, 6q, 8p, 13q, and 17p (for per-platform alteration frequencies, see Supplemental Fig. S3). The arm-level alteration frequencies were converted into chromosomal size-adjusted $Z$ scores as described previously (Beroukhim et al. 2010) to show the extent of deviation from the background alteration rate (Supplemental Fig. S4). The distribution of alteration frequencies (negative correlation of $r = -0.617$ between arm-level gain and loss frequencies) (Fig. 2A) suggests that arm-level alteration-frequent chrosomes tend to be preferentially gained or lost, but rarely both (Beroukhim et al. 2010).

To investigate tumor-type specificity of arm-level gains or losses, we measured the arm-level alteration frequencies separately for the 19 tumor types that have more than 100 samples per tumor type. Hierarchical clustering segregates three main clusters of tumor types with similar developmental origins (Fig. 2B). In one cluster, five hematologic-origin tumors were observed together. A second cluster contained most of the epithelial tumors. A third cluster contained the four tumors arising from neuroepithelial origin (medulloblastoma, neuroblastoma, glioma, and melanoma) and two others. These clusters emphasize the strong relationship between embryogenesis and the post-embryonic development of human cancers at a molecular level. The similarity between sarcoma and renal cell carcinoma and their presence in the neuroepithelial cluster were unexpected. This pattern was previously reported in another large-scale study (Beroukhim et al. 2010) and was present even after the removal of the platform used in that study (Affymetrix 250K/StyI) (Supplemental Fig. S5), but its biological significance is not yet clear.

Figure 1. Compilation of large-scale cancer genome copy number profiles. (A) A schematic of data collection is shown. Five high-resolution, array-CGH platforms used are listed with the corresponding GEO accession ID (GPL) and the number of associated samples. (B) Major tumor types (>100 samples for each type) are shown with their sample numbers with respect to the five platforms. (MDS) Myelodysplastic syndrome, (MM) multiple myeloma, (MPD) myeloproliferative disorder.
Next, we examined the extent of co-occurrence for arm-level alteration pairs. We distinguished concordant arm-level alteration pairs (gain-gain and loss-loss) (Fig. 3A) from discordant ones (gain-loss and loss-gain) (Fig. 3B). Among the potential concordant chromosomal arm pairs, we observed a cluster of pairs between short, gene-rich chromosomes, such as 16p, 17p/q, 19p/q, 20q, and 22q (gene density > 10 genes/Mb) (arrow indicated in Fig. 3A). We also observed two clusters of significant discordant pairs between short, gene-rich chromosomes listed above and gene-poor chromosomes such as 4q, 5p, 6q, 13q, and 18q (marked by two solid arrows in Fig. 3B).

The landscape of focal recurrent alterations across diverse tumor types

Genomic regions frequently altered across diverse tumor types are of primary interest, as they have elevated the likelihood of containing driver alterations. To identify significantly recurrent focal MCRs, we used the GISTIC algorithm to summarize multiple profiles and assign statistical significance (Beroukhim et al. 2007; Mermel et al. 2011). In this method, the average magnitude of copy number alteration (versus frequency alone) is used as a score, and a permutation-based test is used for estimating statistical significance. Across the entire data set, we identified a total of 94 amplification and 71 deletion MCRs (Supplemental Table S3). We termed these 165 recurrent alterations as pan-lineage MCRs.

Pan-lineage amplification and deletion MCRs comprise 0.69% and 0.50% of the reference genome, encompassing coding sequences of 421 and 156 known genes (out of 20,234 autosomal RefSeqs), respectively. Known cancer-related genes were significantly overrepresented in these MCRs (17 out of 264 autosomal cancer consensus genes; Fisher’s exact test \( P = 0.0015 \) (Futreal et al. 2004). Ten (MYCN, PDGFRB, EGFR, FGFRI, WHSC1L1, MYC, HOXC13, CDK4, NTRK3, and ERBB2) and seven cancer-related genes (CDKN2A, PTEN, ATM, ERC1, FOXO1, TP53, and TCF3) were associated with pan-lineage amplification and deletion MCRs, respectively. These 17 genes and their significance (GISTIC \( Q \)-values) are shown in Figure 4A.

GO analysis showed that 15 and 35 out of 1004 GO functional categories were significantly enriched in amplification and deletion MCRs, respectively.
MCRs with known cancer genes. Corresponding functional categories remained significant after removal of related GO functions in parentheses. An asterisk indicates that the corresponding functional categories remained significant after removal of MCRs with known cancer genes.

Figure 4. Recurrent chromosomal alterations across diverse tumor types. (A) The significance of recurrent amplification (left) and deletion (right) as measured by the GISTIC algorithm (GISTIC Q-value; log-scaled) is plotted across the genome. Seventeen known targets (cancer consensus genes) are shown at the corresponding peaks. (B) The GO categories significantly enriched in pan-lineage amplification and deletion MCRs are shown. Similar functional categories (e.g., “kinase” or “kinase activity”) are grouped, and the representative function is shown with the number of related GO functions in parentheses. An asterisk indicates that the corresponding functional categories remained significant after removal of MCRs with known cancer genes.

deletion MCRs (FDR < 0.01) (Fig. 4B; Supplemental Table S4), respectively. Among the 15 functions enriched in amplification MCRs, seven had functional annotations related to “protein tyrosine kinase.” “Cell cycle/apoptosis”-related functional categories comprised 10 out of 35 functions enriched in deletion MCRs. After excluding the peaks that contain cancer consensus genes, “microtubule,” “lyase,” and “transporter” (with amplification) and “cell junction/adhesion” and “metabolism” categories (with deletion) remained significant (Supplemental Table S4).

To identify MCRs that may have arisen due to increased local genomic fragility rather than selective advantage or functional significance, we compared the pan-lineage MCRs with 37 known fragile regions available in the literature (Bignell et al. 2010). Two (FR66/H6p21 and FR113/E13q22) and three fragile sites (FR33B/3p14, FR4F/4q22, and FR10D/10q21) were associated with amplification and deletion MCRs, respectively. In addition, when we measured the overlap between pan-lineage MCRs and known germline copy number variations (CNVs) obtained from the Database of Genomic Variants (DGV) (http://projects.tcgca.ca/variation/), eight MCRs showed substantially elevated CNV density (>100 CNVs/Mb; mean of all MCRs was 36.2 CNVs/Mb), indicative of potential germline origins for these recurrent alterations. The detailed information about the overlap with known fragile regions and CNVs is available in Supplemental Table S3.

Since our compendium is composed of genomic profiles from multiple array-CGH platforms, the reduction of potential platform biases may enhance the true biological signal. This is a challenging problem, especially because we do not have the same samples profiled on multiple platforms and the probe characteristics are variable across platforms. In our attempt to mitigate the impact of the platform-specific effects, we employed a linear mixed model on the profiles after they have been segmented (see Methods). The adjusted profiles show an overall improvement in the extent of overlap between the genomic peaks identified in each platform, compared to those from unadjusted profiles (Supplemental Fig. S6).

Tumor type-specific alterations and functional association map

Our cancer genome database can be used to identify tumor type-specific alterations (“lineage-restricted” alterations) that may be important in a specific cellular context. We used the copy number profiles adjusted for platform effects to call tumor type-specific peaks in each of the 19 tumor types (those with >100 samples per tumor type). As we expected, the set of cancer genes observed in tumor type-specific alterations represents a mixture of cancer genes with broad tumorigenic potential such as CDKN2A, PTEN, ER1C, PDGFRB, MYC, and HOXC13 (for those in pan-lineage alterations) and CCND1, FGFR3, PAX5, and R1B (not in pan-lineage alterations; all observed in five or more tumor types), as well as those with lineage-restricted functionality. Known cancer genes observed in tumor type-specific alterations with the significance of enrichment are listed in Table 1. Only four cancer type-specific alterations showed significant enrichment for known cancer genes (P < 0.05): deletion peaks of breast, leukemia, colorectal, and ovarian cancers. Although we observed some well-recognized pairs of cancer types and copy number altered genes such as breast cancer-ERBB2 (Slamon et al. 2001) and high-grade glioma-EGFR (The Cancer Genome Atlas Research Network 2008), it is likely that some genes whose oncogenic behavior is lineage-dependent are not on the list of known targets due to the incompleteness of this list. For example, although not listed as a cancer consensus gene, microphthalmia-associated transcription factor MITF (3p14.1), known as a master regulator of melanocyte development and a melanoma-specific oncogene (Garraway et al. 2005), was only included in the lineage-specific alterations of melanoma.

One of the 17 prostate cancer-specific deletions was observed within the intergenic region between ERG and TMPRSS2 loci. The average copy number profile of 372 prostate cancer genomes shows that the majority of genomic deletions involving the ERG locus in prostate cancer have ERG and TMPRSS2 loci at their 5’ and 3’ breakpoints, respectively (Supplemental Fig. S7). This deletion is known to give rise to the TMPRSS2-ERG gene fusion in prostate cancer (Kumar-Sinha et al. 2008), and the observed pattern is consistent with the “fusion breakpoint principle” of unbalanced translocation events (Wang et al. 2009). The co-occurrence of intragenic deletion breakpoints of ERG and TMPRSS2 loci was significant (n = 45 out of 372 prostate cancers; P = 9.1 × 10⁻²⁵ by Fisher’s exact test).

To enhance biological interpretation of the tumor type-specific alterations and their associated genes, we measured the enrichment with respect to GO categories (Fig. 5). These maps show significant gene sets (based on tumor type-specific amplifications/deletions) and GO categories as nodes and significant overlaps between them as edges. Figure 5A lists the 31 gene sets showing significant enrichment with amplification peaks of nine tumor types. Genomic loci encoding kinase and signaling molecules are hotspots of amplification, especially in breast cancer, but are also common in other tumor types. Some enrichment can be explained by alterations on a few genomic loci affecting adjacent genes with similar functions (i.e., gene clusters). For example, gene clusters of chemokine ligands (CCL on 17q12) and fibroblast growth factors (FGF on 11q13) are responsible for the enrichment of “chemokine activity/immune” in colorectal cancers and “growth factor activity/signaling” in lung cancers, respectively (for details, see Supplemental Table S5). However, some enrichment such as “kinase activity” of breast cancer (20 genes on nine different chromosomal arms) and “MAP kinase activity” of renal cell carcinoma (MAPK9 on 5q35, MAPK11 and MAPK12 on 22q13, and MAPK15 on 8q24) are sug-
gestive of functionally coordinated genomic alterations in the corresponding tumor types (Cooper et al. 2007). Tumor type-specific deletions of many tumor types (11 out of 19 tumor types examined) were linked to cell cycle and kinase-related gene sets, and they often share the enriched functions, especially for cell cycle- and adhesion-related functions (Fig. 5B). Some functional categories, such as genes encoding mitochondrial membrane components (“mitochondrial” in cervical cancer), DNA repair enzyme machineries (“transcription/DNA” in colorectal cancer) and metalloprotease (“enzyme/metabolism” in melanoma) highlight the relatively unique functionality driven by the genomic deletion in the corresponding tumor types.

### Large-scale survey of genomic hallmark for chromothripsis

Chromothripsis refers to a genomic instability-generating phenomenon in which tens to hundreds of chromosomal rearrangements occur in a “one-off” cellular event (Stephens et al. 2011). This has been observed so far in cancer cell lines and sarcomas (Stephens et al. 2011), multiple myeloma (Magrangeas et al. 2011), and colorectal cancers (Kloosterman et al. 2011). The proposed mechanism for chromothripsis is a massive fragmentation of one or a few chromosomes, followed by rejoining of the fragments (Stephens et al. 2011). During this rearrangement, the fragments can be lost or retained, which will appear as a series of copy number losses and gains, respectively, along the chromosome. To identify copy number changes associated with chromothripsis, we developed a statistical method that measures the extent of structured oscillations in segmented copy ratios and the deviation from the expected distribution of the segment sizes per chromosome (see Methods and Supplemental Fig. S8). Using this method, we performed a large-scale survey of chromothripsis in our compendium and observed that 124 samples (1.5% out of 8227 samples) may harbor genomic evidence of chromothripsis. We find that the expected distribution of the segment sizes per chromosome is not uniform, indicating that chromothripsis is a rare genomic event. This is consistent with previous observations that chromothripsis is a rare phenomenon in human cancer genomes.

### Table 1. Tumor type-specific alterations and associated cancer genes

| Tumor type | Alteration | Peaks | Cancer genes | P-value |
|------------|------------|-------|--------------|---------|
| Breast     | Amp 91     | ERBB2, FUS, HSP90A1, TPTT, LMO2, CRTC1, ELL, RECQL4, MYC, FLT4 | 0.209  |
| Lung       | Del 34     | HRAS, PTEN, MAP2K4, STK11, CDKN2A, RB1, FGFR1OP, MLLT4, CBAF2T3 | 1.1 \times 10^{-5} |
| Leukemia   | Amp 87     | BCR, TSC2, PDGFRB, CDK4, CBAF2T3 | 0.960  |
| LYMPHOMA   | Del 61     | PTEN, ACSL6, JAK2, SMARCBL1, CDKN2A, ETV6, ER1, RB1, CD73, DD10X | 1.3 \times 10^{-5} |
| MDS        | Amp 147    | HOX13, HOX14, TSC2, RET, HOX9, CYLD, PDGFRB, TIE2, HOX4A1, CDX4, ETV1, NOTCH1, CBAF2T3 | 0.596  |
| Glioma     | Amp 168    | PTEN, FGFR3, PDGFRB, FANCA, ASPP1, CDKN1, CD4K, FOXO1, EGFR, CHIC2, MYCN, STK11, CRTC1, RECQL4, PAX1, NOTCH1 | 0.458  |
| Medulloblastoma | Amp 161 | HOX13, HOX14, FGFR3, HOX9, ACSL6, CDKN1, HOX11, LMO1, MYC, PAX5 | 0.824  |
| Colorectal | Del 168    | NUP214, DDB2, ATR, DDT3, TSC2, FGFR3, SDC2, PDGFRB, STK11, NSF1, ZNF384, NCOA2, ASPP1, HRAS, MLLT11, MYC, BRAF, CBAF2T3 | 0.117  |
| Ovarian    | Amp 40     | FGFR3, PTT1, CEBP8 | 0.780  |
| Melanoma   | Amp 66     | PTEN, RB1, TCF3, CDKN2A, RAD51L1 | 0.038  |
| Renal      | Amp 44     | ALB2, SEPT5, CDX2, BX1, FANCA, CDKN2, ER1, TSC3, FLT3 | 0.493  |
| Sarcoma    | Amp 35     | FGFR1, WHSC1L1, ELL, MYC, MYCN, NTRK3 | 0.145  |
| Hepatocellular | Del 17 | PTEN, ATM, CDKN2A | 0.833  |
| Esophageal | Amp 52     | FGFR1, WHSC1L1, ELL, MYC, EGFR, CEBP8, CRTC1 | 0.594  |

Nineteen tumor types are shown with the sample size in parentheses. The cancer consensus genes in each tumor type-specific amplification (Amp) and deletion (Del) peak are shown in order of significance of the corresponding peaks (the number of peaks are separately shown). P-value is the significance of observing no less than the number of target genes by Fisher’s exact test.
tate cancer), PTEN signatures of localized chromothripsis involving chromothripsis events in chromosome 2. In addition, we observed MYCN amplification, the other two cases (GSM333824 and GSM314024) some-wide chromothripsis that may be independent of the (Fig. 6B). Although the first case (GSM313805) showed a chromothripsis in chromosome 2 have amplifications in ample, all of the three candidate neuroblastoma cases showing chromothripsis in chromosome 2 have amplifications in chromosome 2. For example, all of the three candidate neuroblastoma cases showing chromothripsis in chromosome 2 have amplifications in chromosome 2. The hematologic malignancies showed lower frequencies compared to those of epithelial origin, with multiple myeloma (1.9%) and leukemia (0.3%) as the most and least frequent in this group. Among the 19 tumor types examined (>100 supporting samples per tumor type), renal cell and hepatocellular carcinomas showed no evidence of chromothripsis. Initial reports estimated the prevalence of chromothripsis to be 2%–3% cell line data) and 2.0% in multiple myeloma (eight out of 391 multiple myeloma cases). We observed a number of examples in which localized chromothripsis events involved known cancer-related genes. For example, the first case (GSM313805) showed a chromosome-wide chromothripsis that may be independent of the MCN amplification, the other two cases (GSM333824 and GSM314024) have MCV amplification that is embedded within the localized chromothripsis events in chromosome 2. In addition, we observed signatures of localized chromothripsis involving EGFR (chr7; prostate cancer), PTEN (chr10; prostate cancer), and CCND1 (chr11; esophageal cancer) (Fig. 6C).

Discussion
This study presents a meta-analysis of a compendium of copy number profiles for more than 8000 human cancer genomes. Two types of recurrent alterations, chromosomal arm-level and focal MCRs, were analyzed separately across the entire data set and within individual tumor types. Our results show that hierarchical clustering of arm-level alteration frequencies can largely segregate the tumor types according to their developmental lineages (e.g., hematologic, epithelial, and neuroepithelial clusters) (Fig. 2B). Although clustering of multiple tumor types based on developmental lineages was demonstrated using a large-scale transcriptome data set (Ramawamy et al. 2001), our results are suggestive of a substantial embryological influence on the pattern of large chromosomal alterations that arises during development.

In a pairwise correlation analysis between arm-level alterations, we observed significant concordant pairs between small, gene-rich chromosomes. The higher contact probability between these chromosomes as measured in a long-range interaction map (Lieberman-Aiden et al. 2009) raises the intriguing possibility that these chromosomes share physical domains in the nucleus and that this physical proximity may be responsible for the observed copy number changes. Some of the discordant pairs (gain-loss) appear to be examples of functional synergism between known gene targets, e.g., gain of 7p and loss of 10q associated with potential EGFR gain and PTEN loss (marked by a dotted arrow in Fig. 3B; von Deimling et al. 1992). However, we also observed clusters of discordant pairs between gene-rich and gene-poor chromosomes. Although speculative, one hypothesis is that the gain of a driver alteration in a gene-rich chromosome is followed by dosage-compensating losses of gene-poor chromosomes, or vice versa. Restoration of copy number states by the original allele may be preferred (e.g., copy number-neutral loss of heterozygosity) (Makishima and Maciejewski 2011), but gene-poor chromosomes may be substituted in order to minimize the perturbation of essential genes. Biological significance of these concordant/discordant pairs of chromosomal changes needs to be investigated further, as previously explored for glioblastoma and hematologic malignancies (Bredel et al. 2009; Klijn et al. 2010).

MCRs identified from the entire data set (pan-lineage MCRs) showed a significant enrichment for known cancer-related genes as well as with genomic loci encoding kinases (amplification) and cell-cycle/apoptosis-related molecules (deletion). This suggests that a selective advantage during clonal evolution of tumor cells can be largely derived from common genomic alterations. We also observed that “microtubule” and “transporter” (in amplification MCRs) along with “cell junction” and “metabolism” (in deletion MCRs) are overrepresented GO categories in pan-lineage MCRs. The enrichment of these molecular functions remained significant after removal of recurrent alterations containing known targets (i.e., cancer consensus genes). Microtubules constituting the mitotic spindle in dividing cells have been major targets of chemotherapeutic agents such as taxane (Dumontet and Jordan 2010). The role of genomic dosage imbalances in microtubule-encoding genes is less well-established in tumorigenesis; however, point mutations have been shown to influence the drug susceptibility to microtubule-binding agents (Giannakakou et al. 1997; Kavallaris et al. 2001). Further investigation is needed to evaluate the re-
The prevalence of chromothripsis and examples of local chromothripsis involving known cancer genes. (A) The prevalence of chromothripsis measured across different tumor types is shown with 95% confidence intervals. The number of samples showing genomic evidence of chromothripsis is shown in parentheses with the total sample number associated with the tumor type. A dashed line indicates the average frequency across the entire data set. (B) Three neuroblastoma cases with evidence of chromothripsis on chromosome 2 are shown. Arrows indicate the MYCN locus, and insets show a more detailed pattern of copy number changes around the locus. (C) Three examples of local chromothripsis involving known cancer genes of EGFR, PTEN, and CCND1 loci are shown, with the log2 ratios at the cancer gene loci in parentheses.
with high CAAI had little evidence of an alternating pattern in the copy number.

While we have assembled a large compendium of array-CGH profiles, not all subtypes are present for each tumor type, and data for some tumors are gathered from only one or two studies. Thus, the interpretation of some tumor type-specific results should be tempered until verified using additional data sets. Importantly, although researchers are encouraged to make their data and metadata available in public databases once published, the currently available data sets represent only a fraction of the profiles generated by the community. An increase in the data deposition rate into public databases would facilitate meta-analyses such as this one as well as reanalyses for confirming specific hypotheses, and this would be critical for making efficient use of limited resources for cancer research. Another caveat related to our study is that this analysis only focused on copy number profiles, whereas known oncogenes or tumor suppressor genes can be activated or inhibited by several alternative mechanisms (Chin et al. 2011). Thus, integrative, multidimensional analysis using different genomic data types will be critical. The potential of such analysis was recently demonstrated, for example, in identifying novel cancer genes (Akavia et al. 2010) and cancer classification (Kim et al. 2011).

The use of additional data sources (e.g., mRNA and microRNA expression, somatic mutation, and promoter methylation), especially from large-scale cancer genome projects such as the TCGA consortium, will enable these integrative, multidimensional analyses (The Cancer Genome Atlas Research Network 2008, 2011).

Methods

Compendium of human cancer genome copy number profiles

We collected human cancer array-CGH data sets from a public microarray database (GEO: http://www.ncbi.nlm.nih.gov/geo/) (Barrett et al. 2009). We limited our analysis to studies using high-resolution, array-CGH platforms containing more than 100K oligonucleotide probes from two commercial vendors (Agilent and Affymetrix). For the Agilent platform, we used the GEO platform ID of GPL4091 and GPL9128 (both 244K probes). For the Affymetrix platform, we used GPL2004/2005 (100K probes), GPL3718/3720 (500K probes), and GPL6801 (Affymetrix SNP6.0; 1.8 million probes). For the 100K Affymetrix platform, we only used the paired data (i.e., a sample is profiled by both GPL2004/50K-HindIII and GPL2004/50K-XbaI to achieve 100K resolution). For the 500K platforms (GPL3718/250K-NspI and GPL3720/250K-StyI), we included samples genotyped either by GPL3718 or GPL3720 and designated them as “Affymetrix 250K.” For five platforms (Agilent 244K and Affymetrix 100K, 250K, 500K, and SNP6.0), we collected a total of 107 GEO studies. The GEO accession of 107 studies and related information is available in Supplemental Table S1. In this study, we only collected the genomic profiles corresponding to tumor genomics, removing control profiles. We also searched for duplicate samples (most of which are dye-swap cases) or samples that were included in multiple GEO data sets. Although we have done our best, these are sometimes difficult to identify, and it is still possible that our data set contains biological or technical duplicates across different data sets. Detailed information for the final 8227 tumor samples is available in Supplemental Table S2.

Data processing and segmentation

For Agilent platforms, we downloaded the probe-level log2 ratio profiles of individual samples from the GEO database. For the Affymetrix platform, we used the CRMA algorithm (Bengtsson
et al. 2009) for probe summarization and normalization to obtain probe-level intensity values from individual CEL files. We also processed the normal HapMap population CEL files (available in http://www.affymetrix.com). The average intensity values of the HapMap population were calculated for individual probes and used as a reference to calculate tumor/reference log2 ratios. We ensured that the genomic coordinates used in this study are hg18/build 36 using UCSC Genome Browser liftover tools (Fujita et al. 2011). Segmentation was performed using the Circular Binary Segmentation (CBS) algorithm available as an R package (Olshen et al. 2004).

The segment values were median-centered by extracting the median of autosomal segment values per sample. For each study, we calculated MAD (median absolute deviation) for the 50th percentile of autosomal segments whose absolute log2 ratios were close to zero and rescaled the segment values of the samples in the corresponding study. The resulting segment values were further rescaled so that the standard deviation of autosomal segment values of the entire data set was equal to one. We verified that our preprocessing step of median-centering does not impact the analysis results by comparing the results obtained with and without the matched controls samples in the GBM and OV TCGA data sets (data not shown). However, for cancer types with a substantial fraction of the genome altered, it is conceivable that median-centering can cause a slight bias in segmentation calls. The processed segmentation profiles and associated analysis files are available at http://compbio.med.harvard.edu/metaCGH/.

Analysis of chromosomal arm-level alterations

We defined alterations as segments with a predefined threshold for the rescaled segment values (>0.2 and <−0.2 for copy number gain and loss, respectively). In this study, we defined chromosomal arm-level alterations as a single alteration or an aggregate of alterations that exceeds half of the size of the corresponding chromosomal arm. To take into account the background arm-level alteration frequency, we converted the arm-level alteration frequency into size-adjusted Z scores as described previously (Beroukhim et al. 2010). The expected frequency of gain and loss was determined by linear regression, and the Z score was calculated using the normal approximation to the binomial distribution. For hierarchical clustering, tumor type-specific arm-level alteration frequencies were calculated as arm-level gain minus arm-level loss frequencies for 19 tumor types supported by more than 100 samples per tumor type. Hierarchical clustering was performed using 1 – Pearson correlation as the distance with average linkage. In concordance analysis, four possible co-occurrence scenarios between arm-level alterations were separated into concordant (gain/gain and loss/loss) and discordant pairs (gain/loss and loss/gain). The significance of gain/gain co-occurrence between two chromosomal arms of A and B was calculated using Fisher’s exact test:

\[
P = 1 - \frac{1}{\sum_{i=0}^{n_{AB} - 1} \frac{n_{nA} \binom{nA}{i}}{\binom{nA + nB}{i}} \binom{nB + nC}{nA + nB + nC - i}}
\]

where \(n_{AB}\), \(n_A\), and \(n_B\) represent the number of samples showing gain for both chromosomal arms of A and B (\(n_{AB}\)), chromosomal arm A (\(n_A\)), and chromosomal arm B (\(n_B\)), and the total number of samples (\(N\)), respectively. To account for multiple test adjustment, we permuted the calls of arm-level alterations in each sample across the entire data set. For each permuted data set, we calculated the significance of co-occurrence for all possible chromosomal arm pairs to find the minimal \(P\)-value. This permutation was repeated 10,000 times separately for gain-gain, loss-loss, and gain-loss pairs. The empirical \(P\)-value for each arm-level alteration pair was computed as the proportion of permutations whose minimum \(P\)-value is smaller than the observed \(P\)-value of the corresponding alteration pair.

Pan-lineage alterations and functional analysis

To identify MCR, we used the peak-calling algorithm of GISTIC (Beroukhim et al. 2007). The algorithm calculates the probe-level sum of log2 ratios above or below a given threshold (amplification and deletion, separately) across the samples. Then, the algorithm determines significantly altered regions as sets of consecutive probes with a predefined significance threshold (Q < 0.25). From significantly altered regions, it identifies the highest scoring peak corresponding to the MCR. Because the peak or MCR can be displaced from true targets due to neighboring passenger events or noise (Beroukhim et al. 2007), we allowed for 100 kb of confidence intervals at both flanking regions of the identified peak. To suppress the known germline alterations of the HapMap population, we filtered probes corresponding to 2333 autosomal HapMap CNVs obtained from two previous studies using Affymetrix 500K Early Access and SNP6.0 platforms (Redon et al. 2006; McCarroll et al. 2008). For known cancer-related genes, we used 264 autosomal cancer consensus genes available in the literature (Futreal et al. 2004). The significance of enrichment of 165 pan-lineage MCRs with known cancer-related genes was calculated by Fisher’s exact test. We also selected 421 and 156 genes associated with pan-lineage amplification and deletion MCRs, respectively, and measured the enrichment with GO functional categories (http://www.broadinstitute.org/gsea/msigdb/index.jsp; c5 GO categories) (Subramanian et al. 2005). The significance of enrichment was calculated by Fisher’s exact test and corrected for multiple hypothesis testing using the Benjamini-Hochberg false discovery rate. Thirty-seven common fragile regions were obtained elsewhere (Bignell et al. 2010). We also downloaded a list of hg18-compatible CNVs from a public database (variation.hg18.v10.nov.2010; http://projects.tcag.ca/variation/). A total of 57,706 CNVs were collected from 17 studies (>1000 CNVs per study), and regional CNV density was calculated for each pan-lineage MCR.

Removal of platform-specific biases

We employed a linear mixed effect model to remove the potential platform biases. A natural idea is to model the intensity value of each probe in the platforms to adjust for the local bias. However, as probe sets are different across platforms, a probe in one platform may not exist in another platform. We will thus have many missing values for these platform-specific probes. A solution to this “missing-value problem” is to use intensity values of the nearby probes, but this will require determining the size of the neighborhood of the probe, and this choice may have an important influence on the model fitting and bias removal. In this paper, we choose to perform the bias removal based on the segmentation data. Suppose that N samples were profiled on L platforms, and each sample has been processed with a segmentation algorithm. For clarity, we only consider segmentations for one chromosome in the discussion. The segmentation of one sample will correspond to one set of breakpoints. Collect the breakpoints of all samples and denote the breakpoints as \(b_1 < b_2 < \cdots < b_n\). Given the kth interval \((b_k, b_{k+1})\) (\(k = 1, \ldots, B - 1\)) and the ith sample profiled on the ith platform (call this sample \(S_i\); \(i = 1, \ldots, I\); \(l = 1, \ldots, L_i\); \(l = 1, \ldots, L_i\); \(n = N\)), there will be a unique segment from the segmentation of the sample \(S_i\) overlapping with the interval \((b_k, b_{k+1})\). Denote the “segment” value (i.e., the mean of probe intensity values in the corresponding segment) of this segment as \(y_{i,k,l}\). We fit a linear mixed effect model with \(y_{i,k,l}\) as the response variable and the platform as a random effect. Since the tumor type may be an important factor influencing
the values of $y_{i,k,l}$ we also include the tumor type as a predictor (fixed effect) in the model. Assume that there are $T$ types of tumors. Given a sample $S_{ij}$, let $X_{i,j} = (X_{i,j}^{1}, \ldots, X_{i,j}^{T-1}) \in \mathbb{R}^{T-1}$ be the vector such that $X_{i,j}^{t} = 1$ if the sample $S_{ij}$ belongs to the $t$th tumor type ($t = 1, \ldots, T - 1$) and 0 for all the other $t$. Then, we have the following linear mixed effect model:

$$y_{i,k,l} = \alpha_{0} + \beta_{k} X_{i,j} + b_{k} + \epsilon_{i,k,l},$$

$$b_{k} \sim \mathcal{N}(0, \sigma_{b}^{2}), l = 1, \ldots, L,$$

$$\epsilon_{i,k,l} \sim \mathcal{N}(0, \sigma_{\epsilon}^{2}).$$

where $b_{k}$ corresponds to the random platform effect, $\beta_{k}$ is the fixed tumor type effect, and $\alpha_{0}$ is the intercept coefficient in the linear model. For each $k$, we then use the restricted maximum likelihood estimation (RMLE) to estimate the parameters in the above model. In particular, we can get an estimate $b_{k}$ of $b_{k}$. From this estimate, we can get the platform effect-corrected value

$$\tilde{y}_{i,k,l} = y_{i,k,l} - b_{k}.$$ Given a segment $I$ in the segmentation of the sample $S_{ij}$, suppose that it overlaps with $m$ intervals like $(b_{k}, b_{k+1})$. Assume that these intervals are the $k_{1}, \ldots, k_{m}$th intervals and that $w_{k_{1}}, \ldots, w_{k_{m}}$ is the corresponding overlapping length with the segment $I$. We use $\sum_{m}^{w_{k_{1}}} w_{k_{m}} / \sum_{m}^{w_{k_{m}}} w_{k_{m}}$ as the platform effect-corrected intensity value for the segment $I$ of the sample $S_{ij}$.

### Chromothripsis

To identify the genomic signatures representing chromothripsis, we focused on chromosomes with at least 10 alterations. Furthermore, the breakpoints in chromothripsis should occur randomly (Stephens et al. 2011). Thus, the sizes of neighboring segments in a chromothripsis region should be roughly the same or at least at the same order of magnitude. If the sizes of neighboring segments are very different from each other, we would expect that those are not from chromothripsis. Therefore, we designed a score to measure the “extremeness” of the neighboring segments in a copy number profile. Suppose that a chromosome has $n$ segments and their corresponding sizes are $s_{1}, s_{2}, \ldots, s_{n}$. If the sizes of the segments and $i + 1$ are very different, we would have that $R_{i} = \left| \log(2) s_{i+1} / s_{i} \right|$ is large. The median $R$ of the $R_{i}$’s would be a good measure of the extremeness of the neighboring segments. To account for the number of segments in a copy number profile, we normalized the score $R$ by its expected value $R_{e}$ under the hypothesis that the breakpoints are randomly distributed in the chromosome, i.e., we use $S = R / R_{e}$ as the score. We used $S \leq 2.0$ to filter out the profiles that are unlikely to be involved with chromothripsis.

### Acknowledgments

This work was supported by grants from the National Institutes of Health to P.J.P. (R01GM082798, U24CA144025) and by grants from the National Institutes of Health to M.D.J. (NIH Director’s New Innovator Award [DP2OD002319]; NINDS NS062219-01).

### References

Akvava UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Moses E, Garraway LA, Pe’er D. 2010. An integrated approach to uncover drivers of cancer. Cell 143: 1005–1017.

Albertson DG, Pinkel D. 2003. Genome microarrays in human genetic disease and cancer. Hum Mol Genet 12: R145–R152.

Albertson DG, Collins C, McCormick F, Gray JW. 2003. Chromosome aberrations in solid tumors. Nat Rev 34: 369–376.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.

Barrett T, Diep DB, Willhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, et al. 2009. NCBI GEO: Archive for high-throughput functional genomic data. Nucleic Acids Res 37: D885–D890.

Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh J, Linhart D, Vivanco I, Lee JC, Huang JJ, Alexander S, et al. 2007. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. Proc Natl Acad Sci 104: 20007–20012.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, et al. 2010. The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.

Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen L, Beare D, Latimer C, et al. 2010. Signatures of mutation and selection in the cancer genome. Nature 463: 893–898.

Bredel M, Schultens DM, Harsh GR, Bredel C, Chandler JP, Renfrow JJ, Yadav AK, Vogel H, Scheck AC, Tibshirani R, et al. 2009. A network model of a cooperative genetic landscape in brain tumors. JAMA 302: 261–275.
Large-scale analysis of human cancer genomes

The Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061–1068.

The Cancer Genome Atlas Research Network. 2011. Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615.

Cao Q, Zhou M, Wang X, Meyer CA, Zhang Y, Chen Z, Li C, Liu XS. 2010. CaSNP: A database for interrogating copy number alterations of cancer genome from SNP array data. Nucleic Acids Res 39: D968–D974.

Chin L, Gray JW. 2008. Translating insights from the cancer genome into clinical practice. Nature 452: 553–563.

Chin L, Hahn WC, Getz G, Meyerson M. 2011. Making sense of cancer therapeutics. Nat Rev Drug Discov 10: 795–803.

Dumontet C, Jordan MA. 2010. Microtubule-binding agents: A dynamic field of cancer therapeutics. Nat Rev Drug Discov 9: 790–805.

Garraway LA, Widlund HR, Rubin MA, Getz G, Berger AJ, Ramaswamy S, Chen H, Teague J, Butler A, Stevens C, et al. 2007. Patterns of somatic mutation number changes using a genome-wide scoring approach. Cell 129: 27–40.

Klijn C, Bot J, Adams DJ, Reinders M, Wessels L, Jonkers J. 2010. Identification of networks of co-occurring, tumor-related DNA copy number variations. Cell 144: 557–568.

Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. 2008. Recurrent gene fusions and their association with survival in breast cancer. J Clin Oncol 26: 3413–3422.

Klijn C, Bot J, Adams DJ, Reinders M, Wessels L, Jonkers J. 2010. Identification of networks of co-occurring, tumor-related DNA copy number variations. Cell 144: 557–568.

Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. 2008. Recurrent gene fusions and their association with survival in breast cancer. J Clin Oncol 26: 3413–3422.

Lieberman-Aiden E, van Berkum NL, Williams I, Imakaev M, Raganay C, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293.

Mangraunes F, Avet-Loiseau H, Munshi NC, Minvielle S. 2011. Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. Blood 118: 675–678.

Makishima H, Maciejewski JP. 2011. Pathogenesis and consequences of uniparental disomy in cancer. Clin Cancer Res 17: 3913–3923.

Martin TA, Jiang W. 2009. Loss of tight junction barrier function and its role in cancer metastasis. Biochim Biophys Acta 1788: 872–891.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysocker A, Shapiro MH, de Bakker PI, Maller JB, Kirby A, et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40: 1166–1174.

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 12: R41.

Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11: 685–696.

Mittelman F, Johansson B, Mertens F. 2007. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 7: 233–245.

Oshen AB, Venkatraman ES, Luotto R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557–572.

Pinkel D, Albertson DG. 2005. Array comparative genomic hybridization and its applications in cancer. Nat Genet Suppl 37: S11–S17.

Pinkel D, Segraves R, Sundar D, Clark S, Poole J, Kowbel D, Collins C, Kuo W, Chen C, Zhai Y, et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207–211.

Ritchie A, Troge J, Grubor V, et al. 2006. Novel patterns of genome amplification in malignant melanoma. Nature 436: 117–122.

Shapero MH, de Bakker PI, Maller JB, Kirby A, et al. 2008. Integrated genome-wide expression profiles. Nat Genet 40: 1166–1174.

Shapero MH, Díaz-Rowe AJ, Talbot J, Carneiro M, et al. 2009. An integrative approach to large-scale analysis of human cancer genomes. Cancer Res 69: 861–867.

Shapero MH, de Bakker PI, Maller JB, Kirby A, et al. 2008. Integrated genome-wide expression profiles. Nat Genet 40: 1166–1174.

Received March 7, 2012; accepted in revised form October 17, 2012.