UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1072

Copy Number Analysis of Cancer

MARKUS MAYRHOFER

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2015

ISSN 1651-6206
ISBN 978-91-554-9175-8
urn:nbn:se:uu:diva-244361

Dissertation presented at Uppsala University to be publicly examined in BMC E10:1307-1309, BMC, Husargatan 3, Uppsala, Friday, 17 April 2015 at 13:00 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Professor Simon Tavaré (University of Cambridge).

Abstract

Mayrhofer, M. 2015. Copy Number Analysis of Cancer. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 1072. 42 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9175-8.

By accurately describing cancer genomes, we may link genomic mutations to phenotypic effects and eventually treat cancer patients based on the molecular cause of their disease, rather than generalizing treatment based on cell morphology or tissue of origin.

Alteration of DNA copy number is a driving mutational process in the formation and progression of cancer. Deletions and amplifications of specific chromosomal regions are important for cancer diagnosis and prognosis, and copy number analysis has become standard practice for many clinicians and researchers. In this thesis we describe the development of two computational methods, TAPS and Patchwork, for analysis of genome-wide absolute allele-specific copy number per cell in tumour samples. TAPS is used with SNP microarray data and Patchwork with whole genome sequencing data. Both are suitable for unknown average ploidy of the tumour cells, are robust to admixture of genetically normal cells, and may be used to detect genetic heterogeneity in the tumour cell population. We also present two studies where TAPS was used to find copy number alterations associated with risk of recurrence after surgery, in ovarian cancer and colon cancer. We discuss the potential of such prognostic markers and the use of allele-specific copy number analysis in research and diagnostics.

Keywords: chromosomes, oncology, bioinformatics

Markus Mayrhofer, Department of Medical Sciences, Akademiska sjukhuset, Uppsala University, SE-75185 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-554-9175-8
urn:nbn:se:uu:diva-244361 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-244361)

I find my courage where I can, but I take my weapons from science.

-xkcd

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Rasmussen M, Sundström M, Kultima HG, Botling J, Micke P, Birgisson H, Glimelius B, Isaksson A: Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. Genome Biol 2011, 12:R108.

II Skirnisdottir I, Mayrhofer M, Rydåker M, Åkerud H, Isaksson A: Loss-of-heterozygosity on chromosome 19q in early-stage serous ovarian cancer is associated with recurrent disease. BMC Cancer 2012, 12:407.

III Mayrhofer M, DiLorenzo S, Isaksson A: Patchwork: allele-specific copy number analysis of whole genome sequenced tumor tissue. Genome Biol 2013, 14:R24.

IV Mayrhofer M, Kultima HG, Birgisson H, Sundström M, Mathot L, Edlund K, Viklund B, Sjöblom T, Botling J, Micke P, Påhlman L, Glimelius B, Isaksson A: 1p36 deletion is a marker for tumour dissemination in microsatellite stable stage II-III colon cancer. BMC Cancer 2014, 14:872.

Contents

Abbreviations

BAF	B allele frequency
CBS	Circular binary segmentation
DNA	Deoxyribonucleic acid
FFPE	Formalin-fixed paraffin-embedded
HMM	Hidden Markov model
LOH	Loss of heterozygosity
OR	Odds ratio
RR	Relative risk
SNP	Single nucleotide polymorphism

Introduction

A disease of the genome

Cancer is a genetic disease where multiple inherited or acquired mutations lead to loss of growth control and subsequent clonal expansion of cells[1,2]. It may begin as a benign lesion such as a colonic adenoma, in which epithelial cells lining the colon or rectum wall grow abnormally but slowly over years or decades. As the clonal growth of cells increases in size, so does the probability that mutations that further promote growth are sustained in some cell, giving rise to faster-growing subclones. Although the specific sequence of mutations would be unique in each case, the evolutionary process makes it a matter of time before mutations promoting angiogenesis, invasive growth, metastasis and other malignant traits have occurred[3]. If any potential mutations would confer resistance to therapy, the probability that such mutations eventually exist among the tumour cells may be significant. This partially explains why cancers that initially respond to chemotherapy may develop resistance after some time[4]. It is not feasible to detect all mutations present in a tumour cell population (without sequencing every single cell). However, as the mutations that caused the cancer should be present in the majority of the tumour cells, they can be identified in DNA from a tumour sample and taken into account for diagnosis and selection of treatment[5].

The most important difference between somatic and germline mutations is that germline mutations may be passed down to future generations and all their cells, while somatic mutations affect only the cell in which they occur and its relatively few descendants[6]. Some number of somatic mutations is likely to be present in any given cell in the body, but most would be unique to a minor lineage of cells and unlikely to be observed unless the cell grows into a large mass of near-identical cells from which DNA can be extracted and analysed. Benign and malignant tumours are examples of such clonal growths. A tumour does not fill an essential function in the body and does not pass mutations to future generations. Therefore their tolerance for mutations is relatively large[5]. Cancer genomes have typically sustained thousands of point mutations (alteration of one or a few nucleotides) and up to hundreds of deleted, duplicated or translocated chromosomal segments[7]. Deletions and duplications may have altered the number of copies for thousands of genes, some of which contribute to cancer development through dosage alteration or loss of functional transcripts[8].

Many types of cancer may be considered an unfortunate effect of a somewhat steady rate of random somatic mutation, the vast number of cells in our tissues, and our long lives. In our different tissues, certain combinations of mutations have the potential to deregulate key cellular pathways controlling DNA integrity, growth rate and proliferation and transform the normal phenotype of the cell into one with the hallmarks of cancer, as described by Hanahan and Weinberg[3]. Only a limited set of mutations actually contributes to the transformation, most are passenger events that were simply tolerated by the cell. Therefore no two cancers are likely to have the same or even similar sets of mutations. Decades of search for the specific mutations that cause cancer – the driver mutations – have evolved into an explosion of new information with the emergence of genome sequencing and microarray technology. Large collaborative initiatives such as the Cancer Genome Atlas have systematically catalogued mutations in various types of cancer and their impact on the disease[7,9]. The emerging consensus is that while each cancer is genetically unique, it exhibits a finite set of mutational and phenotypic traits that are meaningful to consider at diagnosis and during treatment.

Chromosomal alterations

With the discovery of the Philadelphia chromosome in patients with chronic myeloid leukaemia in 1960, specific chromosomal alterations were proven to play a role in cancer[10]. The invention of chromosome banding technology in 1970 enabled more detailed characterization of both structural and numerical (copy number) alterations, transforming cancer cytogenetics into a powerful diagnostic tool[11]. Numerous or complex chromosomal alterations were found to generally indicate malignancy, while benign tumours were associated with fewer and simpler alterations. Chromosome analysis is now a common diagnostic tool, and is used both to confirm suspected diagnoses and to resolve tumours without a distinct morphology[12].

Even for cancer types lacking clearly distinguishing chromosomal alterations, patterns of alterations are associated with tumour subgroups. Some subgroups stretch across cancer types with different tissue of origin[7]. Specific amplifications and deletions are prognostic markers in a range of cancers and sometimes predictive of sensitivity to specific treatment. Markers for poor prognosis include *NMYC* amplification and 11q deletion in neuroblastoma[13,14] and 1p and 16q loss of heterozygosity (LOH) in Wilm's tumour[15]. Chronic lymphocytic leukaemia is subdivided into five prognostic subgroups based on deletions and duplications[16]. In medulloblastoma, chromosomal alterations can be used to indicate poor, intermediate or excellent prognosis[17]. Targeted inhibitors are available for cancers with certain phenotypes such as *ERBB2*[18] overexpression, with the overexpression mainly caused by

focal copy number amplification of the gene[19]. Knowledge about the impact of cancer genomics on patient progression and survival is growing rapidly and it is possible that genomic profiling will eventually be performed at diagnosis for all types of cancer.

Copy number analysis

Analysis of structural and copy number alterations at low genomic resolution can be performed directly on cells using fluorescence-based karyotyping[20]. This has the advantage of being a single-cell analysis of the absolute number of copies per selected genomic region and cell, and allows detection of balanced translocations. Extracted DNA allows for high genome-wide resolution using either comparative genomic hybridization to microarrays or, more recently, deep sequencing of the DNA[21]. Microarrays are popular in both research and diagnostics due to their cost efficiency, high genomic resolution and relatively simple processing of the data, but cannot detect balanced translocations[22]. Second-generation DNA sequencing provides genome-wide copy number information in addition to single-nucleotide and structural alterations, and may eventually obsolete microarrays[21,23].

Analysis of tumour DNA is complicated by multiple potential problems[24]. Not all cells in a tumour are part of the clonal growth is its cause. Stromal cells are an example of cells with normal genomes that are prevalent inside solid tumours, causing the somatic mutations to be present in some fraction of the extracted DNA. The average copy number or ploidy (i.e. the amount of DNA per cell) may deviate from 2N in the cancer cells and may be difficult or impossible to estimate from extracted DNA alone. The sampled cancer cells may include subclones with a partially different set of somatic mutations[25]. Computational methods capable of handling these problems use SNP allele frequencies in addition to relative DNA abundance to estimate both the absolute number of copies per cell and the fraction of cells with the alteration.

Current developments in cancer genomics

Sequencing of the complete human genome at the turn of the 21st century[26,27] started a new era in cancer genomics marked by systematic hypothesis-generating research in a field that had mostly been hypothesis-driven[28]. Many systematic searches for recurrent (and thereby implicated as driver) mutations in cancer have appeared since shortly thereafter[29,30], and large collaborative efforts such as the Cancer Genome Atlas and the International

Cancer Genome Consortium have undertaken the monumental task of cataloguing recurrent mutations in all common types of cancer[31].

Mining this wealth of genetic data has led to many new insights into cancer-related mutational patterns and cellular pathways, most of which are common to multiple cancer types[7,9]. Cancer types with large amounts of somatic point mutations tend to have a lower frequency of copy number alteration, and vice versa[7], supporting the role of copy number alteration as a driving mutational process. Whole or partial genome doubling has been shown to cause hyperploidy in a range of cancer types[9,24].

Deep sequencing of cancer genomes has made it possible to define the order of mutational events in detail, shedding light on the long path – up to many decades – from the first driver mutation to detection of disease[32,33]. Although mutational events can be expected to occur randomly based on background mutation rate or exposure to DNA-damaging agents, a single event can lead to large numbers of mutations occurring near-simultaneously, some of which may contribute to cancer progression[34–36].

Continuous mutation throughout tumour development and progression inevitably leads to some heterogeneity within the tumour cell population[37], with different subclones harbouring a partially different set of mutations. Such subclones are not always spatially segregated or separated relative to the rest of the tumour, and may even depend on one another for growth[38].

Cancer types have traditionally been separated based on morphology and tissue of origin. The high level of similarity in the mutational patterns of such cancer types has motivated research into individualized treatment based on genetic profiling rather than subgrouping within types of cancer[39]. Genomic factors regardless of cancer type are systematically being tested as predictors for response to treatment[40], with the hope of establishing databases capable of indicating a first line of treatment for any tumour based on its genome.

Analysis of copy number alterations has been essential for many of these discoveries in cancer genomics and will remain an integral part of future cancer research and diagnostics[22]. New computational methods push the boundaries of what can be achieved with available technology and are desired by a large community of researchers and clinicians.

Project Aims

With this PhD project our primary aims were to 1) develop freely available solutions for allele-specific copy number analysis of cancer, suitable for samples with some tumour cell heterogeneity, unknown average ploidy and substantial normal cell admixture, and 2) demonstrate their use in research, taking advantage of the allele-specific information.

A secondary aim was collaboration with other researchers and clinicians to continuously improve on our solutions and their appeal to potential users.

Method: basics and development

Our method development in this field began with the need to understand and analyse cancer samples at Uppsala Array Platform. As we explored microarray-based analysis of cancer genomes we realized that the best way to learn about these complex genomes is to be involved in both development and application of new bioinformatic methods. After developing our first computational method we found that the general solution was equally applicable to DNA sequencing, the output of which could be transformed to closely resemble microarray data. Being involved in methods development has given us access to computational tools that we fully understand and can adapt to new technologies and circumstances immediately when needed.

We developed two methods for copy number analysis of cancer samples and made them freely available as R software. TAPS (for *Tumour Aberration Prediction Suite*, Paper I) is used with SNP microarray data. It takes normalized data from any high-resolution SNP array as input and generates visualizations and estimates of absolute allele-specific copy numbers in the cancer cell fraction. It can also be used to compare groups of samples. Patchwork (Paper III) does the same for whole genome sequenced cancer samples and can be run with or without a patient-matched normal control. While TAPS and Patchwork attempt to solve the same problem, they use very different technologies with different advantages and limitations. They are both designed to operate in two steps. First the data is normalized and visualized without imposing mathematical models or other preconceptions of the relationship between observations and copy numbers. The user may then provide guidance before the copy number estimates are finalized, which is an optional second step. This approach allows the analysis to benefit from the user's own expertise and prior knowledge of the samples, and makes the method flexible for samples and circumstances unexpected at the time of method development, including non-cancer applications.

Patchwork (Paper III) has much in common with TAPS (Paper I), but is designed to use whole genome sequencing data rather than microarray data for the copy number analysis. Most aspects of the analysis presented here apply similarly for both methods. Patchwork takes sequence reads mapped to the human reference genome and a list of detected variants as input. The number of mapped reads is summarized for fixed windows and normalized to produce coverage ratios that closely resemble microarray data. As we

perform most of our copy number analysis with microarrays, examples in this section are based on microarray data.

Copy number analysis based on relative DNA abundance

DNA microarray technology is based on hybridization of fluorescence-labelled single-strand DNA fragments to complementary sequence "probes" attached to a glass plate. The amount of DNA hybridized to each probe, for which the corresponding position on the reference genome is known, can then be assessed using a scanner. Due to imperfect probe sensitivity and specificity, a change in DNA abundance along the reference genome changes the intensity measured on the microarray by less[41], leading to a non-linear relationship between sample DNA abundance and microarray probe intensity.

Whole genome sequencing can be used to assess deviations in copy number throughout the genome by aligning sequence reads to a reference genome and summarizing read coverage over genomic segments. If e.g. one copy of an autosome is missing in an otherwise normal genome, read coverage along that chromosome of the reference genome can be expected to be only 50% of that of other autosomes. The DNA extraction and library preparation required for sequencing is subject to numerous sources of systematic bias that subsequently lead to variations in sequence depth, but most of these may be normalized for[42–44]. Therefore, the expected normalized sequence coverage along the reference genome is linearly proportional to the corresponding DNA abundance in the sample.

Both individual probe intensities and the sequence read coverage along the reference genome are subject to some stochastic sampling of DNA fragments. This leads to substantial random variation, but average intensities or sequence coverage along the reference genome represent reliable measurements of DNA abundance, and thereby the local number of DNA copies per cell (Figure 1).

The genome-wide intensity or coverage data are normally compared to their own average or median and sometimes to control samples, and logarithmized to produce the "log-ratio". A log-ratio of 0 – or a signal ratio of 1 – therefore indicates equal DNA abundance to its own average or median. A simple duplication of a chromosomal segment in a normal diploid genome would raise the DNA abundance by 50%, but as noted above the effect on microarray signal ratio is lower. In Figure 1, duplications reach a log-ratio of $0.35 \approx \log_2(1.28)$.

It is convenient to divide the genome into segments of equal copy number, defining break points based on statistical processing of the log-ratio.

Most methods available for segmentation are either based on Hidden Markov Models (HMMs)[45] or Circular Binary Segmentation (CBS)[46]. HMMs use a fixed set of expected log-ratio levels or states, defining break points that optimize the adherence of each segment to one of the expected states. CBS selects break points recursively, seeking to minimize log-ratio variation within each segment. The main differences for the user are that HMMs are less prone to producing false segment breaks if the sample fits the selected states well, while CBS is computationally demanding but not biased towards a pre-defined set of log-ratio states[47]. HMMs are popular in constitutional cytogenetics where samples are generally homogeneous and near-diploid, simplifying the expected log-ratio of copy number states. CBS is preferred in cancer studies where the log-ratio associated with specific copy numbers cannot be known in advance, as discussed below.

By selecting probes that contain SNPs, and including both alleles on the array, one can estimate not only copy number alterations but also SNP genotypes (Figure 1). By combining total and allele-specific DNA abundance (B allele frequency, BAF), copy number can be estimated based on the two homologous copies that would normally be present per cell. For example, a total of two copies could consist of two non-identical homologous copies, or of two identical copies of the same homolog in case of copy-neutral LOH. In the latter case, all SNPs would appear homozygous (Figure 1, part of chromosome 4). In case of amplification to a total of three copies, one could distinguish duplication of one homolog to 3 copies: 2 of the "major allele" and 1 of the "minor allele" (Figure 1 chromosomes 7-8), from triplication of one homolog with deletion of the other (3 copies: 3 of the major allele and 0 of the minor allele). The log-ratio of these 3-copy alternatives would be the same, but BAF would look very different.

As we began our method development, estimating absolute copy numbers was largely confined to visual inspection of these data tracks, often complicated by aneuploidy, tumour cell heterogeneity and admixture of normal cells.

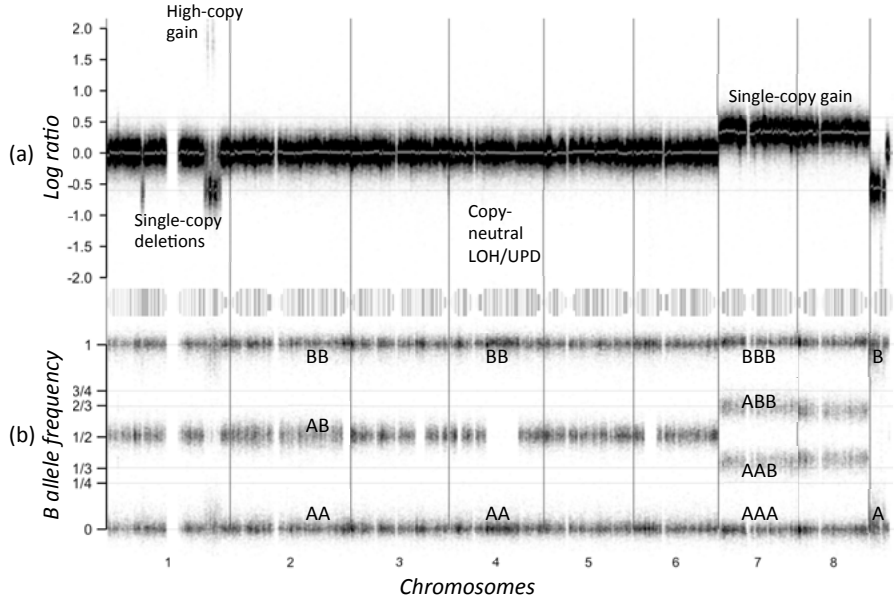

Figure 1. Copy number alterations revealed by microarray. (**a**) Millions of probes (black dots) measure DNA abundance throughout the human reference genome. Average log-ratio of chromosomal segments (grey) can be used to estimate the number of copies per cell. Copy number alterations can be detected as significant increases or decreases in segmented log-ratio. (**b**) Many probes are SNP sites with both alleles present on the microarray, allowing the relative abundance of each allele to be estimated as B allele frequency. Homozygous SNPs appear near 0 or 1, and heterozygous SNPs appear near 0.5 for normal diploid genomic segments. Copy number alteration often results in unequal number of copies of the maternal and paternal homolog, so that each heterozygous SNP appears above or below 0.5 (chromosomes 7-8). Parts of chromosomes 2, 3, 4 and 6 are affected by loss of heterozygosity, as indicated by local absence of heterozygous SNPs.

Aneuploidy and DNA content per cell

The technologies discussed here are applied to DNA extractions with no remaining information on the actual amount of DNA that each cell contained. A fixed amount of DNA is hybridized to the array, or a fixed coverage is sequenced, and differences between samples in total hybridization or coverage are normalized for so that each sample receives a median log-ratio of 0. For diploid and near-diploid genomes, the corresponding median copy number of the genome will be near 2, simplifying identification of gain and loss of copies as significant deviations from a log-ratio of 0 (Figure 1). Unfortunately cancer genomes are not always diploid (2N) or near diploid. The genomes of cancers where copy number alteration is common may range from about 1.5N to well above 4N in average ploidy, with great variation in the copy numbers of individual chromosomes[24]. Log-ratio or BAF do not

indicate absolute number of copies on their own, but if a certain total copy number is assumed, the BAF can indicate the copy number of each allelic copy or homolog $H_{major\text{-}allele}$ and $H_{minor\text{-}allele}$. For chromosomes 7-8 in Figure 1, the implied allele-specific copy number is 3(2+1): 3 copies total, 2 copies of the most abundant and 1 copy of the least abundant homologous copy of the DNA.

We define allelic imbalance as a function of the copy number of each homolog $H_{major\text{-}allele}$ and $H_{minor\text{-}allele}$ as shown in Formula 1, and estimate it in TAPS as shown in Formula 2. Rather than qualitatively classifying each SNP as germline heterozygous or homozygous, clustering of segment-wise abs(BAF-0.5) on two means, BAF_{het} and BAF_{hom}, is used to estimate the imbalance of heterozygous SNPs for each genomic segment.

$$\text{Allelic imbalance}_{theoretical} = \frac{H_{\text{major-allele}} - H_{\text{minor-allele}}}{H_{\text{major-allele}} + H_{\text{minor-allele}}} \qquad (1)$$

$$\text{Allelic imbalance}_{measured} = \frac{\text{BAF}_{\text{het}}}{\text{BAF}_{\text{hom}}} \qquad (2)$$

This solution has the disadvantage of being affected by the noise amplitude of the BAF, leading to some divergence between observed and real allelic imbalance near values of zero and one. However it has the advantage of being very sensitive to differences relative to other segments of the same sample, without being biased by segment length or the number of heterozygous SNPs per segment. A combined view of segmented log-ratio and allelic imbalance can indicate the most likely absolute copy number (and major and minor allele copy number) associated with each chromosomal segment (Figure 2). This works similarly regardless of technology platform (SNP array type), as prior knowledge of the exact relationship between DNA abundance and log ratio is not required.

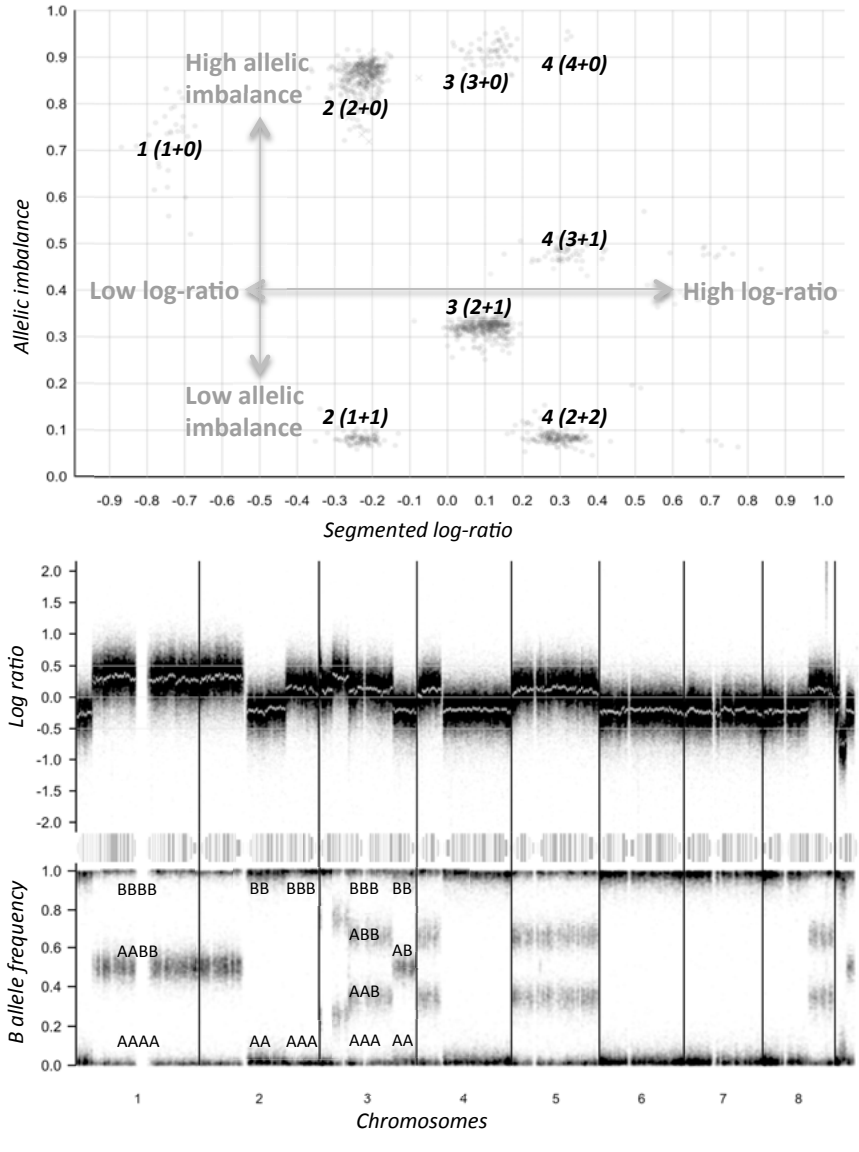

Figure 2. Absolute allele-specific copy numbers of a cancer cell line revealed by plotting log-ratio against allelic imbalance for genomic segments. Dots in the upper panel are genomic segments, which form clusters with certain allele-specific copy numbers. The observed allelic imbalance saturates at values of 0.1 for normal diploid and balanced amplifications, and at 0.7-0.9 for homozygous segments. In between, clusters appear at the correct theoretical allelic imbalance, defined as the difference in allele (homolog)-specific copy number divided by total copy number: 1/3 for a simple duplication and 2/4 for a triplication to 4 copies.

Heterogeneity within the tumour

Not all cancer cells in a tumour are alike. Growing tumours can mutate continuously, leading to genetic diversity in the cell population[4,25]. Most acquired mutations have little or no effect, and in a growth of millions or billions of cells, new mutations are likely to go undetected unless they confer a growth advantage and form a faster-growing subclone. This is part of how cancers develop and progress, and tumour samples often contain observable subclones with additional somatic mutations. Copy number analysis is well suited to detect subclones with a modest difference in their copy number profile, as chromosomal segments containing many markers allow for reliable estimates of DNA abundance relative to the rest of the genome (Figure 3). Unfortunately the possibility of presence of subclonal alterations also complicates estimates of absolute copy numbers, as any cluster of genomic segments, in theory, could be observed as an effect of many different combinations of copy numbers in different cell fractions. When estimating absolute copy numbers it is usually assumed that the simplest possible combination of cell fractions and alterations that would lead to the observation is correct.

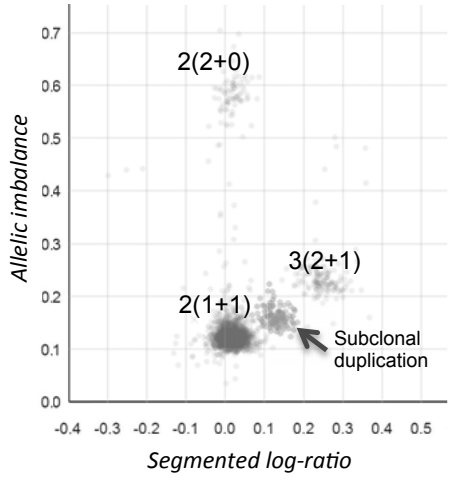

Figure 3. Genetic heterogeneity of the cancer indicated by a subclonal duplication. The fraction of cells in the subclone, assuming it has a simple duplication in addition to the karyotype of the main clone, can be estimated using the log-ratio of homogeneous two and three copies as reference points.

Normal cells in tumour tissue

Not all cells in tumour tissue are part of the clonal expansion of cells that cause the disease. Stromal cells, blood vessels and connective and scar tissue may grow with the tumour and therefore some fraction of cells in a tumour sample contains genomes with none of the mutations that caused the malignancy. Because these genetically normal cells do not have the copy number alterations of the cancer, their DNA serves to dilute the observed effect on log-ratio and BAF of copy number alterations in the tumour cells (Figure 4). Higher normal cell content makes it more difficult to detect smaller copy number-altered segments containing relatively few probes, and a normal cell content above some 90% makes it very hard to estimate absolute copy numbers. An unknown fraction of normal cells also adds to the ambiguity of copy number analysis as shown in Figure 5.

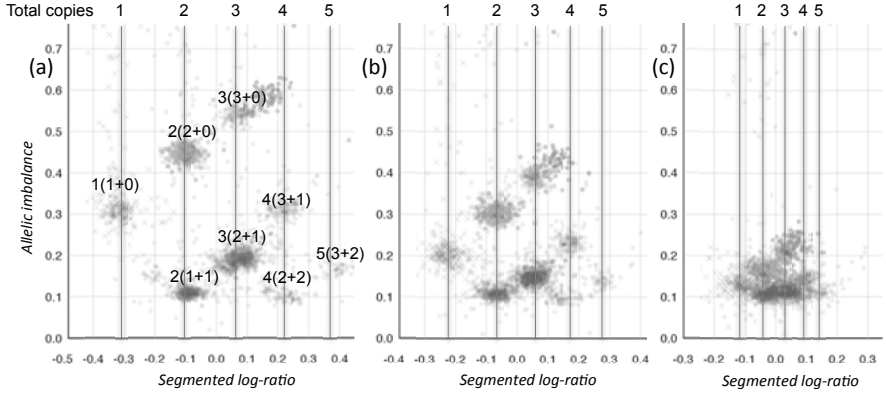

Figure 4. The effect of normal cell content in tumour tissue. (**a**) 55% tumour cells. (**b**) 35% tumour cells. (**c**) 15% tumour cells.

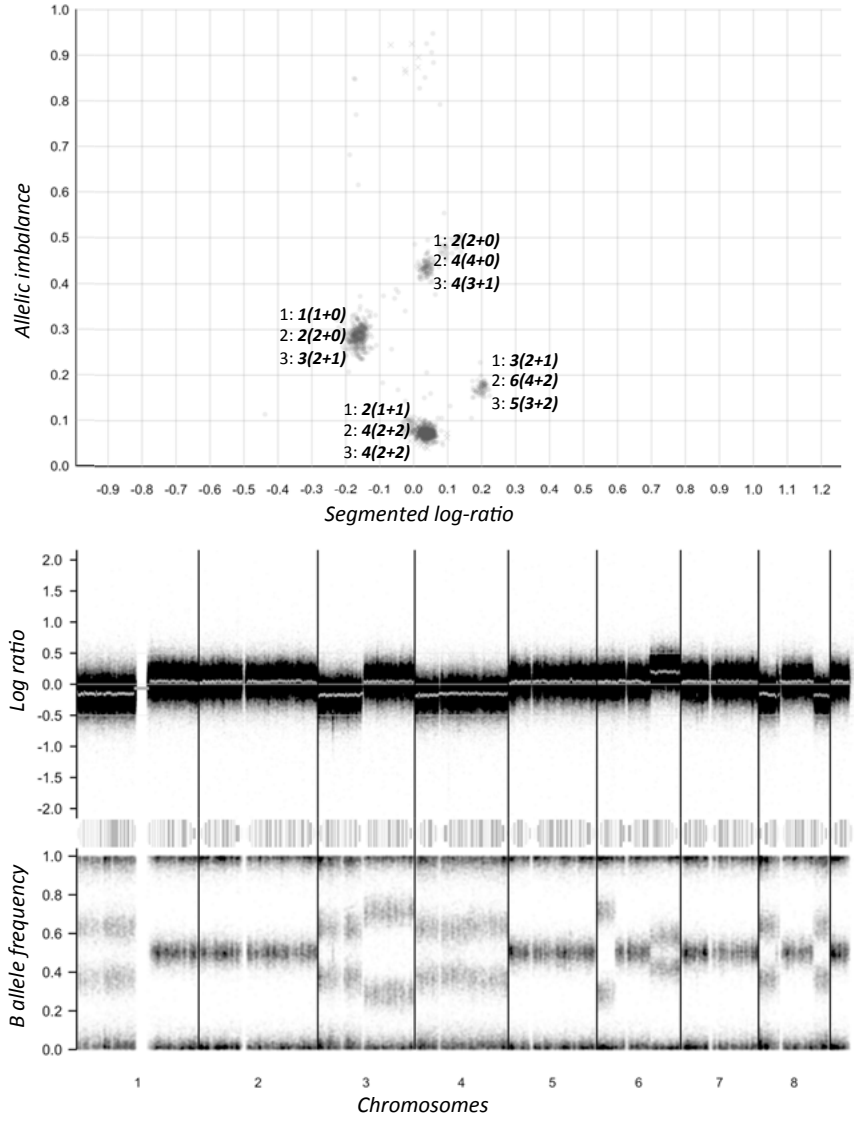

Figure 5. Ambiguity when estimating absolute copy numbers. Three alternative interpretations are presented for each cluster. **Alternative 1:** Purity slightly below 50% would imply presence of LOH and single-copy deletions and duplications in the tumour cells. Allelic imbalance would be near 1/3 for deletion, near 1/2 for LOH and near 1/5 for duplication, corresponding to one cell with 2 non-identical copies for every tumour cell. **Alternative 2:** A perfect doubling of the aforementioned cancer genome would result in identical DNA composition, given purity near 1/3. **Alternative 3:** Instead of the roughly one normal cell with 2 copies per cancer cell in the first alternative, two additional copies could be present in each tumour cell with only a small amount of normal cells in the sample, leading to another identical DNA composition.

Ambiguity in extracted DNA

It is important to remember that some events cannot be detected using this approach. A complete doubling of any genome with no subsequent copy number alterations would be invisible in a fixed amount of extracted DNA, or, in case of presence of normal cell DNA, only alter the apparent purity. More than one set of true copy numbers can lead to the same observation, as exemplified in Figure 5. If tumour cell heterogeneity is considered, a large number of true copy number profiles could lead to the same observation. Knowledge of commonly observed karyotypes in the studied disease helps reduce the risk of error. Prior knowledge of purity also helps, especially when studying cancer cell lines, which can be assumed to not contain patient-matched normal cells.

Downstream analysis

Estimates of allele-specific copy numbers for all genomic segments may be sufficient results for a single-sample, but they are hardly the end result of a research project. More likely, the researcher would like to transform such primary results into a more refined form, which answers a specific question or systematically presents a meaningful summary of a set of samples. One common way to present copy number alterations in a cohort of cancer samples is as a set of recurrently deleted or amplified genomic regions[48]. Such a set of alterations can indicate driver events such as key affected oncogenes and tumour suppressor genes. Such genes are relatively easily spotted for focal high-level amplification (e.g. *NMYC*) and homozygous deletions (e.g. *CDKN2A*) which tend to affect only one or a few genes, as well as for the most common tumour suppressor genes such as *TP53* (indicated by recurrent LOH on 17p). Systematic analysis of the most commonly included genes of recurrently affected genomic regions, over many samples, can come a long way in narrowing down the list of candidate driver genes. GISTIC[19,49] is an example of an algorithm with this aim, freely available and suitable for use with TAPS or Patchwork output.

In the relatively small, selected case-control studies we have been involved in, the aim has been to investigate association between copy number alterations and some phenotype or subgroup exhibited by the cases rather than absolute frequency of alteration or novel driver genes. While this may be performed using common multivariate analysis on presence/absence of each recurrent alteration per sample, it would naturally limit the result to a selected set of recurrent alterations. Copy number alterations are usually of different size in different samples. Therefore a more flexible approach is to calculate the difference in alteration frequency at maximum resolution throughout the genome, separately for each type of alteration (e.g. deletions

and amplifications) and look for peaks (and statistical significance) in that signal. We made a variant of such an analysis is available in TAPS, inspired by the solution used in the commercial software Nexus Copynumber (Biodiscovery Inc.). Unlike Nexus though, which uses a fixed cut-off on log-ratio to determine gains and losses, we used absolute copy number estimates separated into absolute gain, absolute loss, gain relative to individual genome average ploidy, loss relative to individual genome average ploidy, focal gain, focal loss and LOH to produce several sets of frequency difference signals, each representing a certain type of alteration. This has the advantage of distinguishing focal high-level amplification of e.g. *EGFR* from low-level gain of most or all of chromosome 7 (where *EGFR* is located), both of which are recurrent in glioblastoma[19]. Distinguishing between relatively low gene dosage (deletion relative to individual genome average) and LOH (regardless of absolute copy number) can also indicate whether loss of gene dosage or loss of a functional allele is responsible for a peak in deletions and LOH, as we argue in Paper IV.

Looking forward as knowledge about recurrent alterations in cancer increases, eventually the result of a genomic analysis of cancer should be presented as the complete set of variants with any relevance for diagnosis and treatment. This could be reduced to the set of actionable variants in diagnostics and should include rare and previously undocumented alterations in the particular cancer type but with some known relevance in other cancers, as even rare combinations of mutations may indicate excellent response to certain treatments[50]. Recent work in defining a draft set of recurrent copy number alterations in cancer include an analysis by Zack et al[9] based on data from the Cancer Genome Atlas. They presented a large set of alterations recurrent in any of 12 cancer types, with estimated frequencies in each of the 12 types. It is tempting to begin transforming sample-specific sets of genomic segments into presence or absence of each documented recurrent alteration. As long as the list of relevant alterations is not too incomplete it would simplify all downstream analysis, one example being systematic analysis of drug resistance in cell lines by eliminating thousands of genes per sample whose deletion or amplification is mostly a passenger effect.

Paper summaries and discussion

Visualization-based copy number analysis of cancer (Paper I)

In the first paper we describe TAPS, a computational method for copy number analysis of cancer. With this method we are able to estimate the absolute number of copies of each homologous chromosomal copy per cancer cell, at high resolution throughout the genome and in the presence of both aneuploidy and admixture of normal DNA. TAPS works on SNP microarray data and has since been used in both research projects and diagnostics.

Several methods with similar aims including ABSOLUTE[24], ASCAT[51], GAP[52] and others[53–62] have been published. Most of these methods use mathematical models of the expected relationship between array data and copy number, scoring putative combinations of average ploidy and purity and picking the best fit as the most likely truth for each sample. This generally leads to some 70-90% concurrence with validations such as FISH on successfully processed samples, especially if known cancer type-specific expectations are included in the model[24]. TAPS works somewhat differently and is primarily intended to be a visualization tool, with the optional possibility to calculate genome-wide segmented copy numbers based on automatic or manual interpretation, as described in the paper. This has some advantages for optimizing the information gained from each individual sample, which is particularly important in diagnostics. The primary result will not be locked to a definite interpretation of purity and ploidy, reducing the risk of purity overestimation obscuring relevant subclonal alterations. Clinically important implications of alternative interpretations are less likely to go unnoticed. Indeed, even circumstantial evidence of alterations that would impact diagnosis or risk stratification, which differ depending on the disease, should always be made visible to the analyst in a way that maximizes the chance of detection. When the most likely solution is difficult to determine, the analyst may choose to validate ploidy by other means, or at least take the uncertainty into consideration. We believe that it is reasonable to not always achieve confident estimates of absolute copy numbers, as long as certain alterations of interest are detected with the highest possible sensitivity.

The most significant drawback of an analysis dependent on manual input is the risk of misinterpretation of the data leading to systematic errors in the

generated copy number profiles. Interpreting the visualizations is associated with a steep learning curve and just processing a lot of samples does not guarantee improving performance, unless difficult samples are continuously validated. In addition, a DNA extraction cannot be completely relied upon for determining absolute copy numbers per cell. As discussed above several copy number profiles may lead to the same observation. The analysis can therefore at best be strongly indicative of a certain copy number profile, and managing some uncertainty is unavoidable.

An application where the combination of log-ratio and allelic imbalance works very well is the analysis of genomic heterogeneity using multiple patient-matched tumour samples[63]. TAPS visualization is a powerful way to spot clonal and subclonal copy number alterations, especially if multiple patient-matched samples are available. The large number of probes and SNPs contained in most copy number alterations allow for estimation of the cell fraction with a subclonal alteration using either log-ratio or BAF. Copy number analysis may thus be used to gain a general estimate of the genetic heterogeneity of a cancer, which may carry prognostic information, as well as model the phylogeny of multiple patient-matched tumour clones. This may be one of the most important current applications of allele-specific copy number analysis.

Allele-specific copy number analysis of FFPE samples (Paper II)

In collaboration with the Department of Women's and Children's Health, Uppsala University Hospital, we explored early-stage serous ovarian cancer samples for association between copy number alteration and progression-free survival as well as risk of recurrence after surgery. The project used the new Oncoscan microarray set[64] for formalin-fixed paraffin-embedded (FFPE) tissue samples, which uses molecular inversion probe technology to reduce the problems caused by degradation of genomic DNA[65]. TAPS was used for the analysis, and evidence of association between 19q LOH, *TP53* mutation and recurrence was found.

We learned some important lessons from this study, particularly regarding the use of FFPE material for genomic analysis. As described in the paper, the analysis was first attempted using a protocol for analysing FFPE material on the Affymetrix 250k array, with only 20% success rate. Most of the samples that worked well were only a few years old. The Oncoscan array set, developed specifically for use with FFPE material, raised the success rate to about 70%. Analysis failed partially due to problems with DNA quality, especially concerning older samples, and partially due to low tumour cell content. Although an improvement, a total of only 37 successful samples led to limita-

tions in statistical power illustrated by the large number of genomic regions with a big difference in frequency of alteration between the prognostic groups (which were expected to have largely similar alteration frequencies).

Comparing frequency of alteration along the genome usually involves performing a statistical test for a large number of genomic segments, using the union of all break points in the data. Assuming a p-value of 0.05 is required for significance, 5% of independent observations can be expected to appear as false positive discoveries[66]. However the enormous number of tests and the dependency of adjacent segments where observations are near identical make adjusting for multiple testing difficult. P-values can be used to rank the significance discoveries[67] and we opted for such an approach in our study. As in all studies of this kind, our findings should not be considered validated until shown independently in other studies.

We expect that under optimal settings, with an arbitrarily large number of samples in each group, alteration frequency should become significantly different throughout the entire genome, although with mostly small difference in alteration frequency (unless the groups are very different subgroups of the disease). Alteration frequency difference and effect size could then be used to assess the suitability of each alteration as a marker.

Genome sequencing as an alternative to SNP arrays (Paper III)

The third paper presents Patchwork, a computational method similar to the one described in Paper I, though designed for use with whole genome sequencing instead of microarrays. Some methods were already available for copy number analysis based on whole genome sequencing, but most worked on read coverage only. Patchwork was among the first to combine SNP allele frequencies based on read counts with read coverage for copy number analysis. Our intention was to explore and demonstrate allele-specific copy number analysis of cancer using whole-genome sequencing at both high and low coverage. We found that 5-10X coverage is sufficient for allele-specific copy number analysis and that higher-coverage sequencing provides allele specific information of similar quality to SNP microarrays.

Whole genome sequencing provides more information on genomic alterations than microarrays do, and is not subject to the probe specificity and saturation effects discussed in the Method section. Sequence reads can potentially be used to resolve genomic mutations of all kinds, opening the possibility to use copy number information to adjust read count thresholds for detection of single-nucleotide variants, and structural rearrangement junctions at base pair-resolution to detect copy number alterations with much greater sensitivity.

Since second-generation DNA sequencing became commercially available in 2005, the cost of sequencing a complete human genome has dropped steeply from millions of dollars to a few thousands[68]. As computational tools mature and the costs of sequencing and computation continue to drop, sequencing is expected to eventually obsolete the use of microarrays. Main disadvantages with sequencing include the higher cost (currently about ten times more expensive for some 50x coverage of the reference genome) and the larger size of generated data (~1000GB compared to ~100MB).

Sequencing is potentially faster than microarrays depending on library preparation, sequence volume and availability of machines. Microarrays require many hours of hybridization, which is significant in some diagnostic applications. DNA amounts required for whole genome sequencing are still generally higher than for microarrays but are slowly coming down.

Targeted (e.g. exome) sequencing has the advantage of generating much greater read depth relative to cost and required DNA, which is attractive when there is limited DNA available or exceptionally deep sequencing of selected genomic regions is required, but it does not provide high-resolution data for copy number analysis. As sequencing technology and computational methods continue to mature, it is arguably a matter of time before whole genome sequencing becomes standard practice in diagnostics.

A prognostic marker for stages II-III colon cancer (Paper IV)

In the last paper we collaborated with the Department of Surgical Sciences, Uppsala University Hospital, and studied copy number alterations in colon cancer and their association with risk of distant recurrence after surgery. Using high-density (Affymetrix SNP6) arrays and TAPS, we identified association between 1p36 deletion and metastatic dissemination. We could also show that this association applies separately to stage II and stage III patient groups, making the deletion a prognostic marker for distant recurrence.

This was not the first study reporting association between deletion on 1p36 and metastatic disease. One new observation in this study is association between the deletion and metastatic disease as recurrence after surgical resection, rather than only with metastatic disease at the time of diagnosis. We could also show, using our estimates of absolute copy numbers, that deletion relative to the average copy number of the genome should be considered deletion (i.e. presence of the marker for risk of recurrence). This is typically exemplified by deletion having occurred prior to doubling of the genome, after which doubling would have brought the number of copies to two identical, and additional losses of some chromosomes would then reduce the average copy number of many such genomes to about three.

Many prognostic markers with somewhat similar association with clinical outcomes have been presented throughout different types of cancer, and there are some important limitations to their usefulness in clinical practice[69]. The practical value of a prognostic marker depends on how it affects the risk of a particular outcome, the cost of the analysis, whether any action would be taken based on the marker and the associated chance for better patient outcomes. The most intuitive measure of effect size is relative risk (RR), which requires estimating the risk (***p***) of negative outcome given presence or absence of the marker based on the number (***n***) of samples with each observation (Formula 3).

$$RR = \frac{p_{\text{neg.outcome} \mid \text{marker}}}{p_{\text{neg.outcome} \mid \text{no marker}}} = \frac{n_{\text{marker} \cap \text{neg.outcome}} / n_{\text{marker}}}{n_{\text{no marker} \cap \text{neg.outcome}} / n_{\text{no marker}}} \qquad (3)$$

Estimating RR requires the patients included in the study to be unselected for outcome (a cohort study). Case-control studies such as ours are based on a selection of samples with each outcome, disregarding the overall risk of negative outcome in the population. Therefore the frequency of the marker can be estimated in each outcome group, but the overall and marker-stratified risks of negative outcome cannot be estimated. This limits the estimate of effect size to odds ratio (OR) rather than risk difference or relative risk (Formula 4).

$$OR = \frac{p_{\text{neg.outcome} \mid \text{marker}} / p_{\text{pos.outcome} \mid \text{marker}}}{p_{\text{neg.outcome} \mid \text{no marker}} / p_{\text{pos.outcome} \mid \text{no marker}}} = \frac{n_{\text{marker} \cap \text{neg.outcome}} / n_{\text{marker} \cap \text{pos.outcome}}}{n_{\text{no marker} \cap \text{neg.outcome}} / n_{\text{no marker} \cap \text{pos.outcome}}} \qquad (4)$$

In patient groups with a very low risk of negative outcome, OR is a decent approximation of RR, as RR approaches OR ($\boldsymbol{n}_{\text{marker,pos.outcome}} \approx \boldsymbol{n}_{\text{marker}}$ and $\boldsymbol{n}_{\text{no\_marker,pos.outcome}} \approx \boldsymbol{n}_{\text{no\_marker}}$) when the total risk of negative outcome approaches zero[70]. Ideally, the absolute risk of a negative outcome should be known for patients with and without the marker. An unselected study provides such estimates, which are then applicable to patients as long as their total risk of negative outcome is similar to that of the study cohort. Unfortunately the risk of negative outcome can differ between subgroups based on e.g. disease stage, which may not have been stratified for in the study. OR, while less intuitive in nature, is independent of total risk of negative outcome. Let's assume that for each outcome, the frequency of the marker does

not change with the total risk of negative outcome in the patient group. The relative risk of recurrent disease in patients with the marker may then be estimated as a function of the total risk of each outcome in a patient group and the frequency of the marker in each outcome group (Formula 5).

$$RR = \frac{p_{\text{neg.outcome} \mid \text{marker}}}{p_{\text{neg.outcome} \mid \text{no marker}}} = \frac{p_{\text{neg.outcome} \cap \text{marker}} / p_{\text{marker}}}{p_{\text{neg.outcome} \cap \text{no marker}} / p_{\text{no marker}}} = \tag{5}$$

$$\frac{(p_{\text{neg.outcome}} \cdot p_{\text{marker} \mid \text{neg.outcome}}) / (p_{\text{neg.outcome}} \cdot p_{\text{marker} \mid \text{neg.outcome}} + p_{\text{pos.outcome}} \cdot p_{\text{marker} \mid \text{pos.outcome}})}{(p_{\text{neg.outcome}} \cdot p_{\text{no marker} \mid \text{neg.outcome}}) / (p_{\text{neg.outcome}} \cdot p_{\text{no marker} \mid \text{neg.outcome}} + p_{\text{pos.outcome}} \cdot p_{\text{no marker} \mid \text{pos.outcome}})}$$

Similar ways to estimate RR have been in use for some time[71], and carries some risk of confounding results, especially regarding confidence interval estimation[72]. Here it serves to illustrate the relationship between OR which is constant, and RR, which varies with the risk of negative outcome in the patient group. OR is always greater than RR, which approaches OR asymptotically when the risk of negative outcome approaches zero. Figure 10 shows estimated RR as a function of risk of recurrence given a conservative approximation of the effect size observed for 1p36 deletion in our study. For stages II and III, here assumed to have 10% and 30% risk of recurrence, the estimated values of RR become 3.14 and 2.3, respectively.

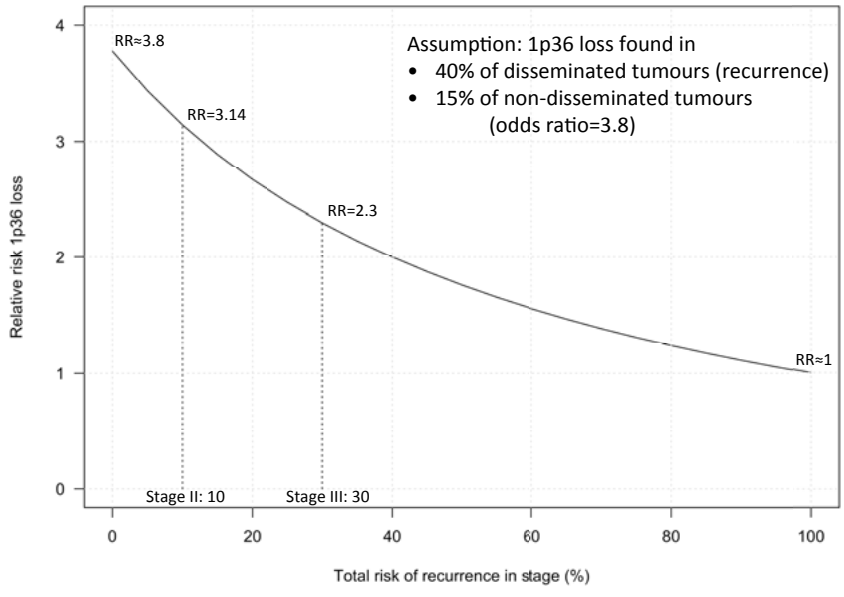

Figure 10. Estimated relative risk of tumour dissemination and recurrence of the disease given presence of 1p36 deletion. Relative risk is presented as a function of total risk of recurrence in the patient group, using Formula 5 and assuming 40% of disseminated and 15% of non-disseminated patients have the deletion as observed in our study.

Similarly, the assumption of constant marker frequency given outcome can be used to achieve a more intuitive presentation of what can be expected from the marker as a function of total risk of recurrent disease. The marker presented in our study has its main potential at a low total risk of recurrence, where the risks and discomfort associated with chemotherapy outweigh the chance of benefit for many patients, as shown in Figure 11:

Treating no patients in the stage II group with 10% risk of recurrence would lead to an expected 10% undertreatment (red dotted line). This could be lowered to 6% (red line) by treating patients with the marker, at the cost of treating 17.5% of tested patients (black line). 13.5% would be overtreated (blue line) and 4% would gain the chance to benefit from the treatment (green line). At a higher risk of recurrence of 30%, it may be more reasonable to compare use of the marker with treating all patients, as 30% undertreatment may be unacceptably high (red dotted line). In that case, the expected overtreatment of 70% (blue dotted line) could be reduced to 10.5% (blue line) at the cost of 18% undertreatment (red line). Treated patients could be reduced from 100% to 22.5% and 12% would potentially benefit from treatment, as they both received treatment and would have had recurrence of the disease.

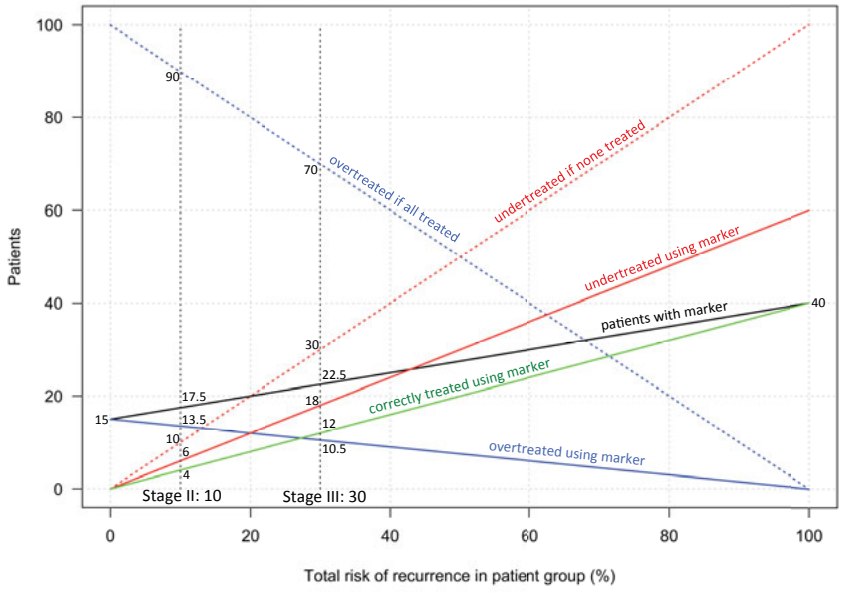

Figure 11. Estimated effects per 100 patients treated with adjuvant chemotherapy based on 1p36 loss as a marker, compared with treating all patients or none. Estimates are presented as a function of total risk of recurrence in the patient group, assuming 40% of disseminated (recurrence after surgery) and 15% of non-disseminated tumours have the deletion as observed in our study. At modest to high risk of recurrence, use of the marker should be compared to treating all patients, which results in overtreatment indicated by the blue dotted line, but no undertreatment. At very low risk of recurrence, use of the marker should be compared to treating none of the patients, which results in undertreatment indicated by the red dotted line, but no overtreatment. Thus an unreasonably high under- or overtreatment could be traded for the effects indicated by the solid lines.

We believe that 1p36 deletion status has an interesting potential as a marker, but that it is somewhat situational. Performing copy number analysis on tissue samples after surgery requires substantial infrastructure, which may and may not be worthwhile relative to other options. We expect that for most patients, other factors than 1p36 deletion status will have more influence on adjuvant chemotherapy decisions and risk assessment, and we suspect that 1p36 deletion will join a growing pool of prognostic markers with circumstantial usefulness in clinical practice. As more studies are performed on colon cancer genomics, we may eventually learn why 1p36 deletions are associated with dissemination and use that information in a more powerful way.

Given the complexity of the interplay between cancer progression and human physiology, we find it reasonable that single alterations confer a weak prognostic effect at most, unless the alteration is associated with a very different subgroup of the disease. We expect single alterations to have more

potential predicting clinical factors directly connected to the cellular process influenced by the particular gene, such as perhaps drug response. It is possible that combinations of multiple mutations, each with a limited predictive power, may be combined to achieve a better association with prognostic outcome, but with hundreds of recurrent alterations it is not feasible to collect enough samples to identify such combinations with common multivariate analysis. Ultimately individualization of diagnosis and treatment based on genomics must come from a progressively deeper understanding of the impact of each mutation on the molecular biology of the tumour.

Closing remarks

It has been a privilege to spend these years working with and learning about the analysis of genomic alterations in cancer. As the cause of the disease and its progression is unravelled in greater detail, we can look forward to individualization of treatment based on genetic profiling as well as deeper insights into cell biology and evolution.

Research, and biomedical research in particular, has transitioned from being the work of a few geniuses into large-scale multidisciplinary industry-like teamwork, where many aspects of a project require contribution from experts. With the transition from the early age of genome sequencing into the age of big data, the field of cancer genomics is set to grow ever more complex and diverse. Global collaboration will be essential to leveraging the full potential of translational cancer genomics, specifically in the form of free, open source method development and freely available primary data for meta-analysis and data mining. We can look forward to staggering amounts of genomic and biomedical data in the decades to come, and somewhere in the chaos of new information lurk the keys to transforming cancer into a curable or chronic disease.

Conclusions

In this thesis project we have developed tools for copy number analysis with a slightly different focus than comparable methods developed at the same time. More than just processing tools, they can be used to visualize the cancer genome in a way that maximizes flexibility with interpretation of the results. That is intended to lead to swift and easy assessment of clinically meaningful alterations and facilitate generation of new hypotheses, while minimizing the risk of important alterations going undetected. Browsing through whole-genome visualizations provides a convenient way to assess aneuploidy, purity and heterogeneity in addition to specific alterations of interest, and is also useful for comparing patient-matched samples. For systematic studies of many samples, tables of copy number estimates can be produced.

We performed two studies where we used our method to investigate cancer samples in new ways, separating LOH from deletions and separating relative gain and loss from absolute. Additional collaborative studies are published or underway, and we have begun using our method for diagnostics. Performing this work has given us valuable tools and a deeper insight into the chromosomal alterations that contribute to cancer. Our current and future projects benefit from these experiences as we keep working for better bioinformatics and better copy number analysis of cancer.

Acknowledgements

During these years I have had the honour to work with and learn from many inspiring people in Uppsala's research community and beyond. In particular my main supervisor Anders Isaksson who deserves a lot of credit for not only listening to all my rants about what we could and should do, but setting aside resources for me explore and develop ideas into something useful. Anders, you have done a great job with support and guidance throughout this PhD project, sharing your many connections in Sweden and abroad, and not losing faith in our work during setbacks.

I would like to thank my current and former colleagues at Uppsala Array Platform whom I have worked closely with for so long. Maria Rydåker, Malin Olsson and Anna Haukkala for entrusting me with minor duties in the laboratories (though, wisely, not many times). Hanna Göransson Kultima for patiently teaching me about microarray analysis in the early days. Sebastian DiLorenzo, Björn Viklund and Martin Dahlö for being hackers. Belinda Fridman, Annsofie Andersson and Anna Dovärn for keeping things running smoothly during parental leaves.

Collaboration with experts in Uppsala's cancer research community has opened my eyes to different aspects of biomedical research. Karolina Edlund, Bengt Glimelius, Helgi Birgisson, Patrick Micke, Johan Botling and Magnus Sundström. Ingiridur Skirnisdottir and Helena Åkerud. Rebeqa Gunnarsson, Larry Mansouri and Richard Rosenquist Brandell. Tanmoy Mondal, Gaurav Pandey and Chandrasekhar Kanduri. Tobias Sjöblom and Lucy Mathot. Joakim Crona. Sven Nelander, Vikki Marinescu and Sathish Baskaran. Inspiring collaborators beyond Uppsala include Anita Grigoriadis and Johnathan Watkins at King's College London and David Gisselsson Nord at Lund's University Hospital.

Colleagues and fellow PhD students (mostly graduates, by now) Christofer Bäcklin, Carl Mårten Lindkvist, Jessica Nordlund, Anna-Karin Hamberg, Henning Karlsson, Daniel Laryea, Jessica Schubert and many more at Clinical Pharmacology and SciLifeLab, thanks to you coffee and a nice chat is never far away.

My great family which has kept supporting my decision to study something so complex and frustrating receive special credit, most of all Kajsa and Toste for reminding me every day of what is best in life.

References

1. Boveri, T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J. Cell Sci.* **121 Suppl 1,** 1–84 (2008).
2. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
3. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100,** 57–70 (2000).
4. Diaz Jr, L. A. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486,** 537–540 (2012).
5. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339,** 1546–1558 (2013).
6. The Genetics of Cancer. *Cancer.Net* at <http://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer>
7. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45,** 1127–1133 (2013).
8. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10,** 551–564 (2009).
9. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45,** 1134–1140 (2013).
10. Nowell, P. & Hungerford, D. Minute Chromosome in Human Chronic Granulocytic Leukemia. *Science* **132,** 1497–1497 (1960).
11. Caspersson, T., Zech, L. & Johansson, C. Differential binding of alkylating fluorochromes in human chromosomes. *Exp. Cell Res.* **60,** 315–319 (1970).
12. Mitelman, F. in *eLS* (John Wiley & Sons, Ltd, 2001). at <http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0005552.pub2/abstract>
13. Janoueix-Lerosey, I. *et al.* Overall Genomic Pattern Is a Predictor of Outcome in Neuroblastoma. *J. Clin. Oncol.* **27,** 1026–1033 (2009).
14. Ambros, I. M., Brunner, C., Abbasi, R., Frech, C. & Ambros, P. F. Ultra-High Density SNParray in Neuroblastoma Molecular Diagnostics. *Front. Oncol.* **4,** (2014).
15. Grundy, P. E. *et al.* Loss of Heterozygosity for Chromosomes 1p and 16q Is an Adverse Prognostic Factor in Favorable-Histology Wilms Tumor: A Report From the National Wilms Tumor Study Group. *J. Clin. Oncol.* **23,** 7312–7321 (2005).
16. Parker, T. L. & Strout, M. P. Chronic Lymphocytic Leukemia: Prognostic Factors and Impact on Treatment. *Discov. Med.* **11,** 115–123 (2011).
17. Pfister, S. *et al.* Outcome Prediction in Pediatric Medulloblastoma Based on DNA Copy-Number Aberrations of Chromosomes 6q and 17q and the MYC and MYCN Loci. *J. Clin. Oncol.* **27,** 1627–1636 (2009).

18. Mitri, Z., Constantine, T. & O'Regan, R. The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. *Chemother. Res. Pract.* **2012,** (2012).
19. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl. Acad. Sci.* **104,** 20007–20012 (2007).
20. Motionbuzz.com. Cytogenetic Testing Methods | University of Florida Health Pathology Laboratories. (2010). at <http://pathlabs.ufl.edu/services/ cytogenetics/cytogenetic-testing-methods>
21. Li, W. & Olivier, M. Current analysis platforms and methods for detecting copy number variation. *Physiol. Genomics* **45,** 1–16 (2013).
22. Bejjani, B. A. & Shaffer, L. G. Application of array-based comparative genomic hybridization to clinical diagnostics. *J. Mol. Diagn. JMD* **8,** 528–533 (2006).
23. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14,** 1–16 (2013).
24. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* (2012). doi:10.1038/nbt.2203
25. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* **1805,** 105–117 (2010).
26. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
27. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291,** 1304–1351 (2001).
28. Garraway, L. A. & Lander, E. S. Lessons from the Cancer Genome. *Cell* **153,** 17–37 (2013).
29. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417,** 949–954 (2002).
30. Samuels, Y. *et al.* High Frequency of Mutations of the PIK3CA Gene in Human Cancers. *Science* **304,** 554–554 (2004).
31. (Chairperson), T. J. H. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
32. Greenman, C. D. *et al.* Estimation of rearrangement phylogeny for cancer genomes. *Genome Res.* **22,** 346–361 (2012).
33. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
34. Stephens, P. J. *et al.* Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144,** 27–40 (2011).
35. Baca, S. C. *et al.* Punctuated Evolution of Prostate Cancer Genomes. *Cell* **153,** 666–677 (2013).
36. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149,** 979–993 (2012).
37. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501,** 328–337 (2013).
38. Kumar, A. *et al.* Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes. *Genome Biol.* **15,** 530 (2014).
39. Dancey, J. E., Bedard, P. L., Onetto, N. & Hudson, T. J. The Genetic Basis for Cancer Treatment Decisions. *Cell* **148,** 409–420 (2012).

40. Bignell, G. R. *et al.* Abstract 2218: Genomic characterisation of 1015 cancer cell-lines. *Cancer Res.* **74,** 2218–2218 (2014).
41. Halper-Stromberg, E. *et al.* Performance assessment of copy number microarray platforms using a spike-in experiment. *Bioinformatics* **27,** 1052–1060 (2011).
42. Ivakhno, S. *et al.* CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* **26,** 3051–3058 (2010).
43. Boeva, V. *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27,** 268 –269 (2011).
44. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40,** e72 (2012).
45. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17,** 1665–1674 (2007).
46. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostat. Oxf. Engl.* **5,** 557–572 (2004).
47. Basics of CNV Calling Algorithms HMM, CBS, Rank Segmentation. *BioDiscovery - Copy Number Variation* at <http://www.biodiscovery.com/2013/05/21/basics-of-cnv-calling-algorithms-hmm-cbs-rank-segmentation/>
48. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
49. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
50. Exceptional Responders Initiative: Questions and Answers. *National Cancer Institute* at <http://www.cancer.gov/newscenter/newsfromnci/2014/ ExceptionalRespondersQandA>
51. Loo, P. V. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107,** 16910–16915 (2010).
52. Popova, T. *et al.* Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol.* **10,** R128 (2009).
53. Attiyeh, E. F. *et al.* Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res.* **19,** 276–283 (2009).
54. Gardina, P. J., Lo, K. C., Lee, W., Cowell, J. K. & Turpaz, Y. Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500K SNP Mapping Arrays. *BMC Genomics* **9,** 489 (2008).
55. Chen, H., Xing, H. & Zhang, N. R. Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. *PLoS Comput. Biol.* **7,** e1001060 (2011).
56. Göransson, H. *et al.* Quantification of normal cell fraction and copy number neutral LOH in clinical lung cancer samples using SNP array data. *PloS One* **4,** e6057 (2009).
57. Bengtsson, H., Neuvial, P. & Speed, T. P. TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics* **11,** 245 (2010).

58. Ortiz-Estevez, M., Bengtsson, H. & Rubio, A. ACNE: a summarization method to estimate allele-specific copy numbers for Affymetrix SNP arrays. *Bioinforma. Oxf. Engl.* **26,** 1827–1833 (2010).
59. Staaf, J. *et al.* Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol.* **9,** R136 (2008).
60. Yau, C. *et al.* A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol.* **11,** R92 (2010).
61. Li, A. *et al.* GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. *Nucleic Acids Res.* **39,** 4928–4941 (2011).
62. Liu, Z., Li, A., Schulz, V., Chen, M. & Tuck, D. MixHMM: Inferring Copy Number Variation and Allelic Imbalance Using SNP Arrays and Tumor Samples Mixed with Stromal Cells. *PLoS ONE* **5,** (2010).
63. Mengelbier, L. H. *et al.* Intratumoral genome diversity parallels progression and predicts outcome in pediatric cancer. *Nat. Commun.* **6,** (2015).
64. Affymetrix Rolls Out OncoScan Service for Profiling FFPE Cancer Samples. *GenomeWeb* at <https://www.genomeweb.com/arrays/affymetrix-rolls-out-oncoscan-service-profiling-ffpe-cancer-samples>
65. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21,** 673–678 (2003).
66. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125,** 279–284 (2001).
67. Network, T. C. G. A. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
68. DNA Sequencing Costs. at <http://www.genome.gov/sequencingcosts/>
69. Pepe, M. S., Janes, H., Longton, G., Leisenring, W. & Newcomb, P. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. *Am. J. Epidemiol.* **159,** 882–890 (2004).
70. Cummings, P. The relative merits of risk ratios and odds ratios. *Arch. Pediatr. Adolesc. Med.* **163,** 438–445 (2009).
71. Zhang, J. & Yu, K. F. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* **280,** 1690–1691 (1998).
72. McNutt, L.-A., Wu, C., Xue, X. & Hafner, J. P. Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. *Am. J. Epidemiol.* **157,** 940–943 (2003).

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1072

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-244361

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2015