UPPSALA
UNIVERSITET

UPTEC X 14 007
Examensarbete 30 hp
Maj 2017

Timing of chromosomal alterations during tumour development

Björn Viklund

UPPSALA
UNIVERSITET

**Teknisk- naturvetenskaplig fakultet
UTH-enheten**

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Timing of chromosomal alterations during tumour development

Björn Viklund

During cancer development, tumour cells will accumulate a lot of both somatic point mutations and copy number alterations. It is not unusual that affected genes have a copy number that differs from the usual two. Due to the loss of DNA repair mechanisms the cells can mutate independent from each other which gives rise to different subclones within the tumour. A tumour cell and its future daughter cells that gets an advantage in cell division speed compared to its competing neighbours, will eventually make up a large portion of the tumour. All the mutations that the subclone's most recent common ancestor acquired until the expansion will be shared across the subclone.

In this project, we have developed a method using the mutation frequencies from publicly available whole genome sequencing data, to quantify the amount of competing subclones in a sample and determining the time to its copy number duplications. This method could be further developed to be an extension to regular copy number analysis.

A heterogeneous tumour can grow faster and be more resistant to treatment. Therefore, it is important to learn more about cancer development and get a greater understanding of the order in which copy number alterations occur.

Handledare: Anders Isaksson
Ämnesgranskare: Tobias Sjöblom
Examinator: Jan Andersson
ISSN: 1401-2138, UPTEC X 14 007

Populärvetenskaplig sammanfattning

En cell som så småningom utvecklas till en cancercell samlar på sig många somatiska mutationer, det vill säga mutationer som enbart kommer existera i den cellen och dess dotterceller. Mutationerna varierar från enbaspolymorfier till att hela kromosomer dupliceras eller går förlorade. Det innebär att antalet kopior av de påverkade generna ofta avviker från det normala två. Ett mål med projektet var att utnyttja helgenomsekvensering för att med hjälp av punktmutationer bestämma ungefär när kopietalsförändringar har inträffat.

I en tumör är det vanligt att vissa mutationer bara finns i en subklon av tumörcellerna. Om någon av dessa mutationer ger en ökad tillväxthastighet kommer subklonen utgöra en allt större andel av tumörcellerna.

Vi har utvecklat en metod som kan användas till att hitta eventuella subkloner, avgöra ungefär när de började växa samt hur stor andel av cellerna de utgör. Detta kan vidareutvecklas och bli en del av förbättrad kopietalsanalys.

När metoden användes för att analysera cellinjer och ett tumörprov, observerade jag att tidpunkterna för kromosomförändringar varierade mellan proverna. Fortsatta studier krävs för att utröna vilken relevans detta har för tumörutvecklingen och kliniskt beteende hos tumörerna.

Contents

1 Introduction

Every cell in the body is exposed to mutations. Only mutations that occur in germ cells are inherited to the offspring. All other mutations are somatic mutations, they only affect the individual cell in which the mutation occurs and its daughter cells. During cancer development, the cell that will eventually transform into a fast growing cancer cell, will have accumulated a lot of somatic mutations. These somatic mutations range from single nucleotide substitutions to whole chromosome duplications or deletions. Deletions and duplications will cause the number of copies of the genome to locally deviate from the normal two[1,2].

To initiate tumour growth in colorectal cancer, it has been shown that a series of point mutation events have to occur. An initial point mutation that regularly occur in the Adenomatous polyposis coli (APC) gene will cause the cell to ease its restriction of cell division speed, giving it a slight edge compared to its neighboring cells. If another point mutation occurs in a gene such as the Kirsten rat sarcoma viral oncogene homolog (KRAS) gene, will give the cell an even greater advantage in division speed[3]. The next step is, the disabling of tumour suppressor genes such as the TP53 gene. This is often accomplished when one of the alleles is affected by a point mutation that removes its function, and the other is deleted. This is referred as the "Two hit hypothesis"[4]. Each time a cell gains an advantageous mutation over its siblings a new subclone will form, this can occur numerous times during cancer development (Figure 1).

To study these somatic point mutations, a sample from the tumour is extracted. A tumour can contain billions of cells, including cancer cells and regular cells such as blood vessel cells. During preparation of genomic DNA to be whole genome sequenced, cells in the sample are lysed and mixed together, making the resulting DNA a mix of tumour and normal DNA. The DNA is then fragmented and sequenced as short reads, that are assembled against a reference genome. The more reads that map to a specific region the higher is the sequence coverage. Nucleotides or regions that differ from the reference genome are considered genomic variants[6].

Data from whole genome sequencing can be used to identify the absolute copy numbers throughout the cancer genome, based on the two non-identical homologous chromosomes that are normally found[5]. Which means that the number of copies for each specific allele across the genome is known.

Access to whole genome sequence and copy number data opens the possibility to estimate when a specific copy number alteration has occurred. There is an existing method that estimates the order in which rearrangements of genomic segments have occurred, by using graph theory methods and somatic mutations[7]. This method is encouraged to use with high sequence coverage, upwards of 180 was recommended to achieve good performance[2]. Unfortunately the sequence coverage that was available during this work was on around 52. Therefore in this project we have focused on determining the time to duplication (TTD) of copy-number altered regions, and to identify the time until subclonal growth in the tumour.

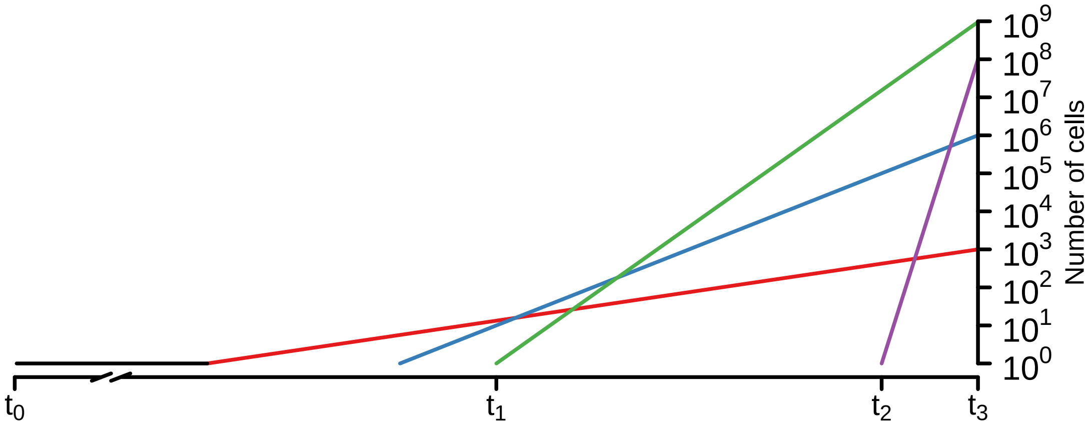

Figure 1: Schematic illustration of tumour progression into cancer. Here we assume exponential growth for each subclone and do no take any growth limiting factors into account (nearby blood vessels and nutrition, et cetera). The interval t_0 until t_3 spans from the time of initial somatic mutation t_0 until the date of tumour sample collection t_3. During t_0 until the red line, a single cell accumulates somatic mutations during a time period (up to several decades). The red line starts the moment a rate limiting mutation in a gene, such as the APC gene, affects the cell's growth rate. This initiates the cells uncontrolled proliferation into a tumour. The blue line starts when a cell gained another mutation in for example the KRAS gene and caused the cell to proliferate faster than the red clone. At t_1, a cell from either the red or blue clone gained yet another mutation that further increased the growth rate and formed the major clone at the time the samples was taken. At time t_2 a new rate limiting mutation occurred in one of the cells. At time t_3 the sample was taken, since the green clone had the largest number of cells it is considered as the primary clone. The purple clone is only one-tenth of the green clone, and therefore defined as a subclone that occupies 10% of the sample. Due to the small proportions the remaining subclones will not be detected.

2 Method development

Publicly available whole genome sequencing data was retrieved for the breast cancer cell lines HCC1187 and HCC2218 and patient-matched normal cell lines. Genomic DNA from four samples of human colon cancer, A01_167, C01_203, E01, G01_278 and patient-matched blood or normal tissue were sequenced by Complete Genomics, see Table 1. Complete Genomics provides sequencing and reference genome assembly with their own pipeline.

The flowchart in Figure 2 shows an overview of the major steps needed for this project.

Table 1: Description of samples used in the study. MSI/MSS describes if the samples are microsatellite instable or microsatellite stable.

Name	Source	Median coverage	Mean ploidy	#point mutations	MSI/MSS
HCC1187	Breast cancer cell line	51	2.8N	18644	MSS
HCC2218	Breast cancer cell line	46	2.1N	23894	MSS
A01_167	Primary colon cancer	53	3.5N	18711	MSS
C01_203	Primary colon cancer	51	3N	18776	MSS
E01	Primary colon cancer	60	2N	15305	MSS
G01_278	Primary colon cancer	53	2.8N	21893	MSS

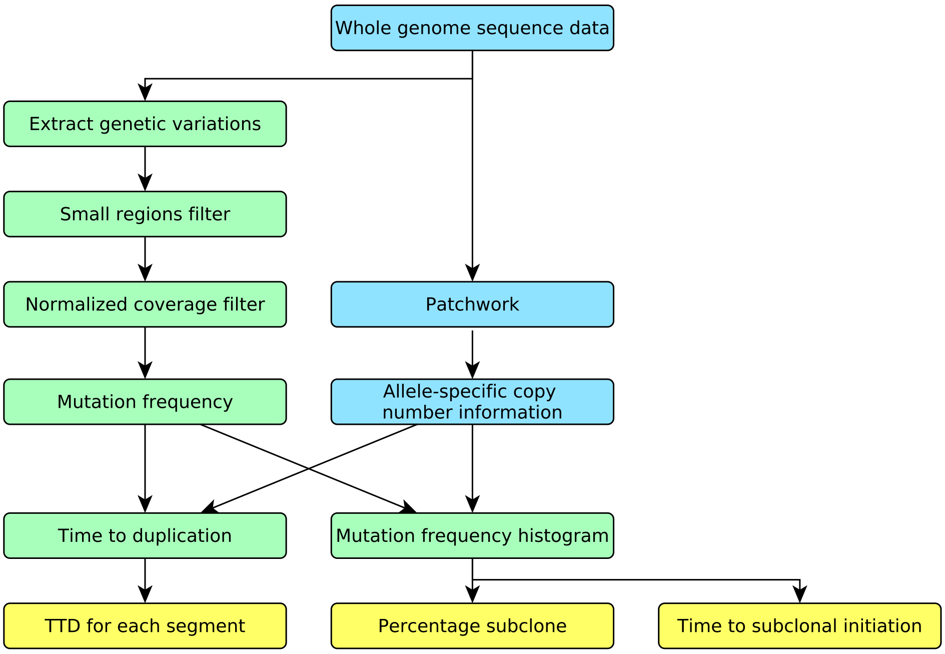
Whole genome sequence data
Extract genetic variations
Small regions filter
Normalized coverage filter
Patchwork
Mutation frequency
Allele-specific copy number information
Time to duplication
Mutation frequency histogram
TTD for each segment
Percentage subclone
Time to subclonal initiation

Figure 2: Flowchart describing the project. The blue boxes indicate already existing tools/data. Methods that had to be developed are marked green and the output is coloured yellow. Patchwork was used to identify the copy numbers. Genetic variations such as somatic point mutations and SNPs, and sequence coverage were extracted from the whole genome sequence data. Due to an unexpected non-random distribution of point mutations in a few of regions per sample, two filters were developed to cope with skewed results (See filtering). The mutation allele frequency is a ratio between zero and one that describes how much of the coverage each point mutation occupy. With the copy number information and the mutation frequencies, the time to duplication for each segment was estimated. The mutation allele frequency and normal diploid genomic segments were used to both determine the time to subclonal initiation, that is the time the cell that eventually grew into a subclone spent acquiring somatic mutations until a new rate limiting mutation occurred, and to estimate the proportion of subclone in the tumour.

2.1 Time to duplication

TTD of a segment is the proportion of time between t_0 and t_1 when the aberration occurred (Figure 1). A TTD-value near zero indicates an early duplication whereas a late duplication will have a TTD-value near one.

To obtain TTD-values, point mutations before and after gene duplication were extracted. If the point mutation occurred before duplication, then it may be amplified and there could be two exact copies of the point mutation. A point mutation that occurred before duplication is called as an early point mutation. A point mutation that occurred after duplication will affect only one chromosome, and is marked as a late point mutation (Figure 3). During a period of time, the segment will have accumulated several point mutations, both before and after a duplication event. If the duplication happened early, there would be a lot more late than early point mutations. If it happened late it would have few late point mutations and considerably more early point mutations.

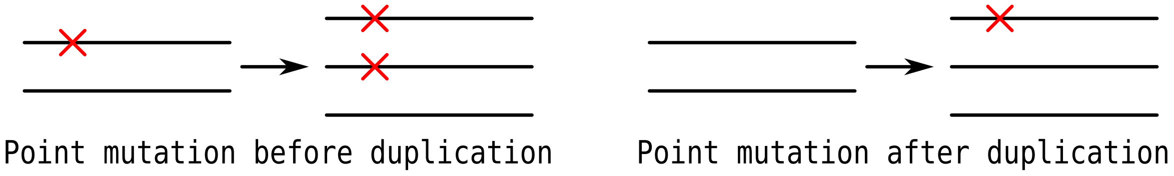
Point mutation before duplication
Point mutation after duplication

Figure 3: Visualization of point mutations before and after duplication of a segment. *Left*: Early point mutations result in amplification of the point mutation upon duplication. *Right*: Late point mutations will not be copied as they are post-duplication.

Assuming even exposure to point mutations over time, the TTD can be estimated by the ratio between the exposure of early mutations and the total exposure of mutations (Equation 1). The more copies of a segment the greater is the exposure to somatic mutations. To compensate for varying exposure, the copy number before and after duplication were taken into account (Equation 2). Figure 4 illustrates a duplication example.

$$time\ to\ duplication = \frac{early\ mutation\ exposure}{total\ mutation\ exposure} \tag{1}$$

$$time\ to\ duplication = \frac{\frac{\#early\ mutations}{early\ copy\ number}}{\frac{\#late\ mutations}{copy\ number} + \frac{\#early\ mutations}{early\ copy\ number}} \tag{2}$$

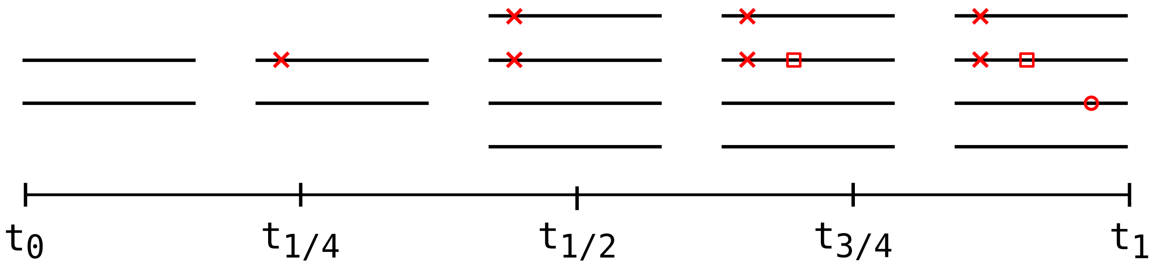

Figure 4: Visualization of time to duplication. The time t_0 indicates when the cell started to accumulate somatic mutations, and t_1 the time the cell was sequenced. The two segments in focus were duplicated approximately halfway along the timeline. Only one point mutation occurred before the duplication. After the duplication the mutation exists on two out of four chromosomes. During the rest of the time, two more point mutations occurred, and could only affect either one of the chromosomes. At t_1 there was a total of three mutations. The exposure of point mutations were assumed to be constant per amount of DNA. Equation 2 reveals the time to duplication to be one-half.

The method described above will only discover the early mutations that have occurred on a segment that was amplified. Any unamplified segments' point mutations will only affect one copy regardless of when they occur, and therefore would not be informative. As point mutations occur randomly, an equal number should occur on both homologous chromosomes before duplication. The observed number of duplicated early mutations was used to estimate the number of early unduplicated mutations.

2.2 Mutation allele frequency

To determine if a point mutation occurred early or late with whole genome sequencing data, the mutation allele frequency was used. The frequency was defined as the ratio between the coverage of the mutation and the total coverage on that position. If a mutation exists on two out of three segments it will have an expected mutation allele frequency of about 2/3, as opposed to 1/3 if the mutation was on the unamplified segment. This is visualized in Figure 5 where there are two distinct bands when point mutation allele frequencies are plotted against their position on the chromosome.

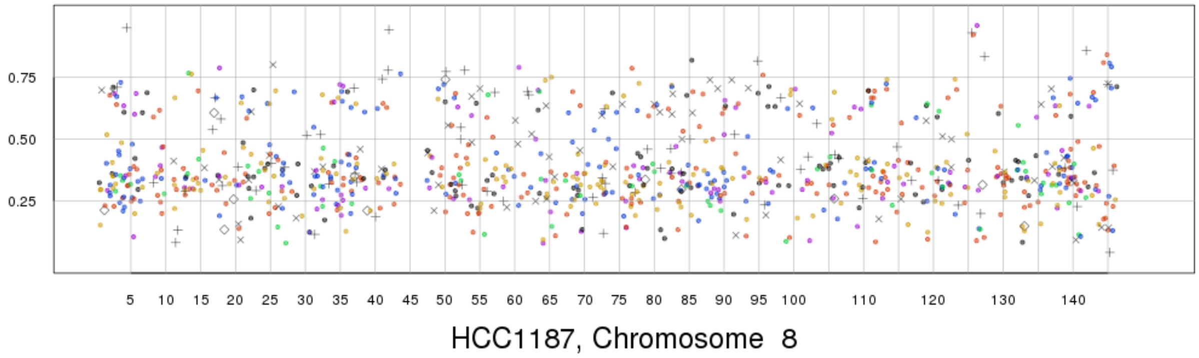

Figure 5: HCC1187, chromosome 8. The x-axis shows the position on the chromosome, the y-axis shows the mutation allele frequency. The whole chromosome has copy number three. Each point mutation with its corresponding mutation allele frequency is plotted against its position on the chromosome. Early mutations have a mutation allele frequency at approximately 0.66 or 0.33, and 0.33 for late mutations.

2.3 Extract somatic mutations

To identify somatic point mutations in the tumour cells, tumor DNA was compared to patient-matched normal DNA. All variations from the reference genome that were not also found in the normal DNA were classified as potential somatic mutations. Complete Genomics gives every somatic mutation a so-called somatic score, that ranks each somatic mutation by quality. A score higher than -10 was considered as a real somatic mutation, all these were extracted from the data.

2.4 Filtering

It is the random occurrence of point mutations that make it feasible to use this approach in determining the TTD. When plotting each mutation allele frequency to its corresponding position on the chromosome, most of the point mutations were evenly distributed along the chromosome. Some regions had large clusters of point mutations very near each other, indicating a deviation in mutation exposure. These regions stood out from the rest and introduced a large source of error. To determine the TTD as accurate as possible, the small region filter and the normalized coverage filter were used.

2.4.1 Small region filter

Longer regions produce more accurate TTD due to the greater amount of point mutations, while shorter tend to do the opposite. The data contained small regions that both split larger regions into smaller ones and contained too few point mutations to be reliable. Therefore regions shorter than two million base pairs and their point mutations were removed (Figure 6), and the adjacent larger regions with the same copy number were merged (Table 2 and Table 3).

Table 2: Unfiltered copy number region table. Showing a 100 kb long region that breaks a large segment.

HCC1187, chromosome 2			
Cn	**Start**	**Stop**	**Length**
3	10000	5700000	5.69e6
4	5700000	5800000	1.00e5
3	5800000	10200000	4.40e6

Table 3: Filtered copy number region table. Showing the extended segment.

HCC1187, chromosome 2			
Cn	**Start**	**Stop**	**Length**
3	10000	10200000	1.02e7

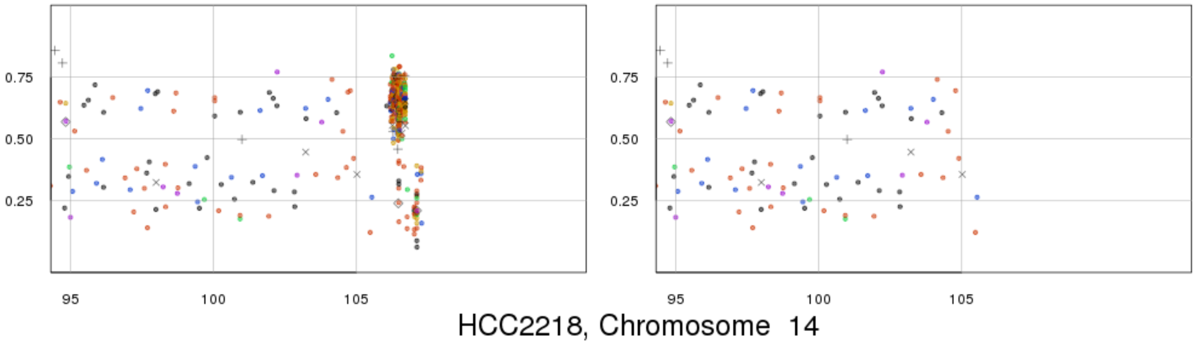

Figure 6: HCC2218, chromosome 14. The x-axis shows the position on the chromosome, the y-axis shows the mutation allele frequency. Some regions had a dense cluster of point mutations, these were a source of error. With the removal of short regions, the dense point mutation clusters were also removed. *Left*: A large accumulation of point mutations at the end of chromosome 14. *Right*: Point mutations after the removal of short regions.

2.4.2 Normalized coverage filter

The normal genome contains very short regions where the copy number deviates from two. These will remain in the cancer genome. As Patchwork uses the tumour-normal coverage ratio, they appear to have the same copy number as the surrounding region. This contributed to a source of error when a small part of a segment contained the wrong copy number and would have been exposed differently to somatic mutations. A higher copy number correspond to a higher coverage, resulting that regions with different copy numbers than the majority in the segment stood out when comparing their coverages. Point mutations within those regions were removed (Figure 7).

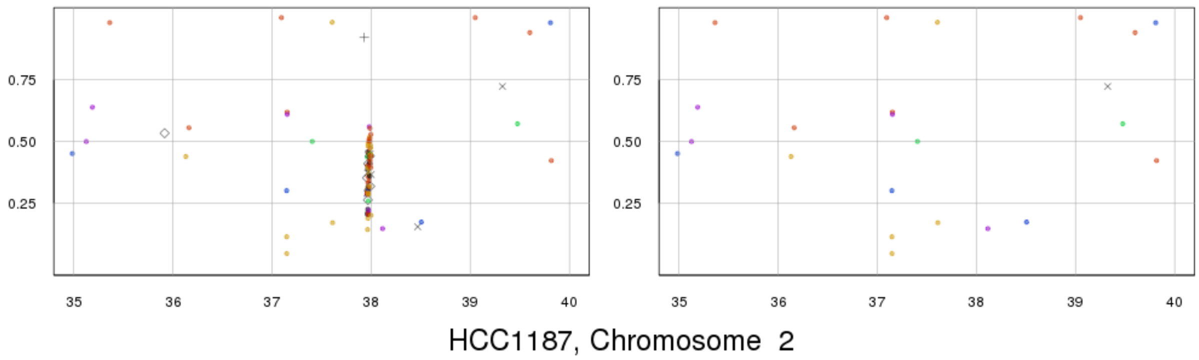

Figure 7: HCC1187, chromosome 2. The x-axis shows the position on the chromosome, the y-axis shows the mutation allele frequency. *Left*: A cluster of point mutations within a region with a different copy number than the surrounding ones. *Right*: Every mutation within the region is removed by the normalized coverage filter.

2.5 Histogram

The time to subclonal initiation is the time between t_1 and t_2 in Figure 1, that is the time the cell that eventually grew into a subclone took to gain another rate limiting mutation. The time is relative to the interval t_0 until t_1. The ratio between number of point mutations accumulated during the subclone and the primary clone development (Equation 3) was used to estimate the time to subclonal initiation. The point mutations that were acquired during the primary clone development were inherited by all cells in the tumour.

In genomic segments where no copy number variations have occurred, point mutations would only affect one of the two chromosomes, resulting in a mutation frequency of around one half. Point mutations that occurred in the cell that grew into a subclone were present at a lower mutation allele frequency than the point mutations present in all cells. Because we did not measure the mutation allele frequency exactly, these point mutations would give rise to a distribution with a lower mutation allele frequency than the point mutations present in all tumour cells. From the bimodal distribution in the histogram of the mutation allele frequencies in Figure 8, the number of mutations in the two distributions was estimated by mirroring the side of the distribution that was facing outwards, reducing the overlap between the two as much as possible.

$$time\ until\ subclone\ initiation\ = \frac{\#subclonePointMutations}{\#pointMutationsInAllCells} \tag{3}$$

The proportion of subclone cells at the time of analysis were obtained by using the mutation allele frequency at the peak of the subclone distribution (Equation 4). A greater proportion of subclone increases the expected mutation allele frequency for subclonal point mutations, until the maximum value of one half is reached.

$$\%subclone = \frac{peakSubclone}{0.5} \tag{4}$$

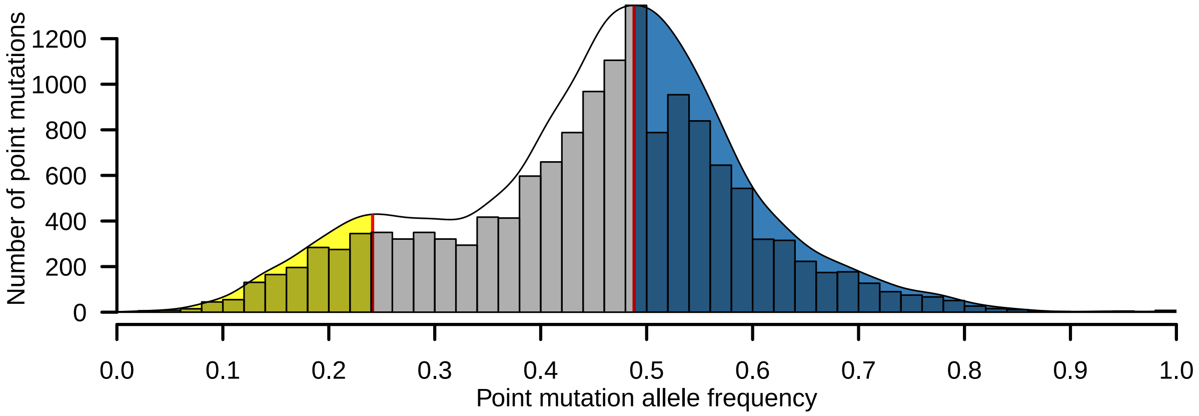

Figure 8: Histogram of the mutation allele frequencies in regions without copy number alterations in HCC2218. The red lines indicate mutation allele frequency peaks, the left peak is the subclone's peak and the right is the primary clone's peak. The subclone's peak is located at 0.242 and the primary clone's peak is 0.486. Using Equation 4 results in a subclone content of around 50%. The yellow and blue areas were used when calculating the number of point mutation in each clone. The number of point mutations under each color were doubled to estimate the opposite side of the distribution to minimize the interference between the two distributions.

2.6 Estimation of the time t_0 until t_1 in years

The time t_0 until t_1 for each sample was translated into years using the number of point mutations per megabase pairs (Mbp) divided by the point mutation rate per Mbp per year.

The point mutation rate in the exome in colon cancer is estimated to $7.6 * 10^{-10}$ mutations per base pair per cell cycle, and a colon cancer cell's self-renewal rate is estimated to one per week[9]. The point mutation rate is higher in the genome compared to the exome due to purifying selection. In our primary colon cancer samples there was a 36.3% higher mutation rate in the genome compared to the exome. This was compensated by increasing the exome point mutation rate with 36.3%. Using Equation 5 the point mutation rate per Mbp per year was 0.0539.

To identify the number of point mutations per Mbp in each sample, segments from copy number one and two were used. The number of point mutations for each copy numbers were extracted using the same technique used in Section 2.5, subclonal mutations were excluded. Point mutations with extreme coverage were also excluded using the normalized coverage filter discussed in section 2.4.2. The number of point mutations is proportional to the copy number. This effect was compensated for. Using Equation 6 and 7 the time for each sample was calculated.

$$\text{number of point mutations per Mbp per year } = \text{point mutation rate} * 1.363 * 52 * 10^{6} \tag{5}$$

$$\text{number of point mutations per Mbp} = \frac{\text{number of point mutations}}{\text{segment lengths} * 10^{6}} \tag{6}$$

$$\text{year} = \frac{\text{number of point mutations per Mbp}}{\text{number of point mutations per Mbp per year}} \tag{7}$$

2.7 Application to real sequence data

The complexity of the genome in the samples A01_167, C01_203 and G01_278 made the copy number analysis unfeasible and therefore they were not used. Table 4 shows the samples' time between t_0 until t_1, the time to duplication, and the subclone content. Figure 9 visualizes the TTD for each sample. It shows three different scenarios, in HCC1187 the majority of copy number alterations appear to have happened halfway through, in HCC2218 TTDs are more evenly spread out in the interval and in E01 the greater part of the events happened relative late.

Table 4: Each samples' point mutations per mega base pairs (translated into years between the parentheses), time to subclone initiation (translated into years between the parentheses) and the subclone percentage.

Sample	**MutPerMb (t_0 to t_1)**	**TTSI**	**Subclone(%)**
HCC1187	1.48 (37.6 Yr)	0.09 (5.3 Yr)	37
HCC2218	2.2 (55.6 Yr)	0.17 (14.5 Yr)	50
E01	1.83 (46.4 Yr)	0.06 (4.2 Yr)	37

2.8 Point mutations present in genes with loss of heterozygozity

According to the "two hit hypothesis", one or many point mutations in tumour suppressor genes may have been involved in the disabling of the gene. Therefore, the point mutations in known tumour suppressor genes[8] who had lost either one of their alleles and contained at least one point mutation in an exon were extracted. In regions with greater copy numbers than one it was possible to determine if the point mutation occurred before or after the first duplication by using the same technique discussed in section 2.1. See appendix A for a complete list of the point mutations that were found.

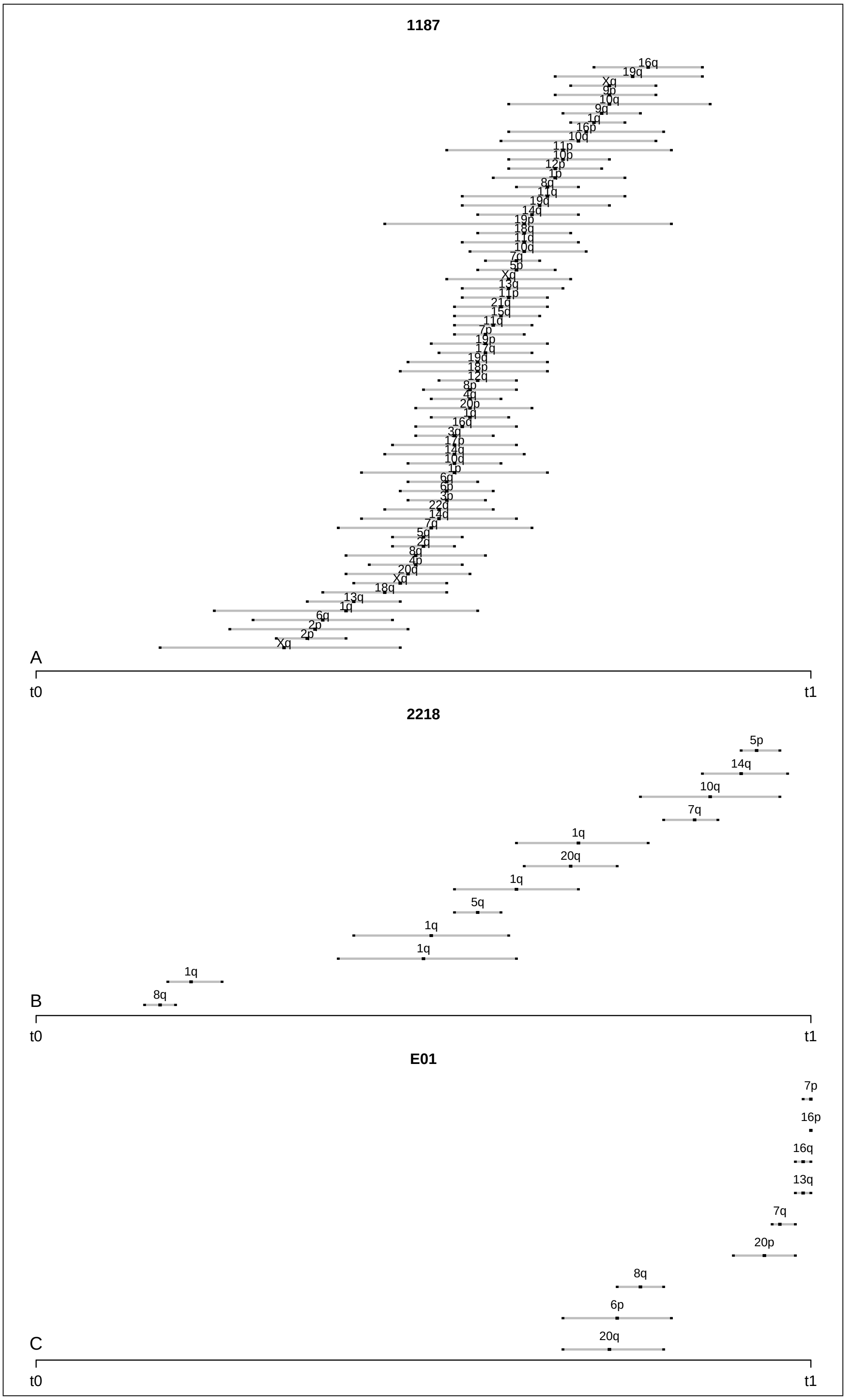

Figure 9: Illustration of TTDs with a 95% confidence interval. The label on top of the TTD is the chromosome position of each region. *A*:HCC1187, *B*: HCC2218 and *C*: E01.

3 Discussion and conclusion

I have developed a bioinformatic method that can be used to increase the knowledge about cancer and cancer development. We only had access to two breast cancer cell lines and one primary colon cancer sample, this made it difficult to draw conclusions about recurrent patterns. This will be interesting when more primary colon cancer samples are analyzed. Different subgroups of colon cancer may have distinct developmental patterns, were different regions gets amplified at different stages. Combining the TTD, the time between t_0 and t_1, time to subclonal initiation and the subclone content gives us new ways to describe tumour development. These parameters can also be used to improve methods for allele specific copy number estimation like Patchwork. This method can also be used with junction information in further development of digital karyotyping, to be able to have access to the karyotype of a sample just from its sequence instead of using spectral karyotyping, where the chromosomes are coloured with florescence and is quantified with a light microscope. The digital karyotype gives a more complete picture of the actual karyotype compared with what can be achieved with a light microscope.

I encountered many obstacles during the course of the project that needed to be resolved. The non-random distribution of point mutations was by far the largest problem, but due to the success of the two filters the analysis was feasible.

Greenman *et al.*'s method is significantly more advance and is used on smaller regions. That method requires a very high sequence coverage. Therefore that method is not applicable for validation of our results[7].

Limitations: The method that we used to estimate the number of point mutations from the mutation allele frequency distributions in Section 2.5 is rough but good enough compared to other methods such as Mixture Models, which are more accurate but is more difficult to automate.

Non-synonymous somatic mutations in the exome affect the resulting translated protein, which in some cases will remove its function and reduce the specific growth rate of the cell. The result is a lower observed mutation rate in the exome. The difference varies between individual tumours but is consistently higher in the whole genome compared to the exome. Vogelstein *et al.* used exome data to identify the point mutation frequency per self-renewal. By combining these two pieces of information we can estimate the time t_0-t_2 in years. To get a more accurate estimate it would be better if the mutation rate per nucleotide per cell division was estimated directly from the whole genome.

Conclusion: This method builds upon recent technological advances in bioinformatics and whole genome sequence data to provide new insights into cancer development. We are looking forward to analyzing new whole genome sequenced colon cancer samples. With more data sets available, our method will provide a more detailed view of copy number alterations in cancer. This will lead to a greater understanding how cancer evolves and might find differences between different types of colon cancer.

4 Acknowledgements

I would like to thank my supervisor Anders Isaksson, who gave me the opportunity to work in his group and help me with the project, Markus Mayrhofer for the genetic expertise, Christofer Bäcklin for R support, Martin Dahlö for UPPMAX support and Sebastian DiLorenzo and Hanna Göransson Kultima for brainstorming with me.

References

[1] Michael R. Stratton, Peter J. Campbell and P. Andrew Futreal, *The cancer genome*, Nature 458, 719-724, 2009

[2] Serena Nik-Zainal, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, Adam Shlien, Susanna L. Cooke, Jonathan Hinton, Andrew Menzies, Lucy A. Stebbings, Catherine Leroy, Mingming Jia, Richard Rance, Laura J. Mudie, Stephen J. Gamble, Philip J. Stephens, Stuart McLaren, Patrick S. Tarpey, Elli Papaemmanuil, Helen R. Davies, Ignacio Varela, David J. McBride, Graham R. Bignell, Kenric Leung, Adam P. Butler, Jon W. Teague, Sancha Martin, Goran Jönsson, Odette Mariani, Sandrine Boyault, Penelope Miron, Aquila Fatima, Anita Langerød, Samuel A.J.R. Aparicio, Andrew Tutt, Anieta M. Sieuwerts, Åke Borg, Gilles Thomas, Anne Vincent Salomon, Andrea L. Richardson, Anne-Lise Børresen-Dale, P. Andrew Futreal, Michael R. Stratton and Peter J. Campbell, *The Life History of 21 Breast Cancers*, Cell 149, 994-1007, May 2012

[3] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr., Kenneth W. Kinzler, *Cancer Genome Landscapes*, Science 339, 1546-1558, 2013

[4] Alfred G. Knudson, *Hereditary cancer: Two hits revisited*, Journal of Cancer Research and Clinical Oncology 122, 135-140, 1996

[5] Markus Mayrhofer, Sebastian DiLorenzo and Anders Isaksson, *Patchwork: allele-specific copy number analysis of whole genome sequenced tumor tissue*, Genome Biology, 14:R24, 2013

[6] Pauline C. Ng, Ewen F. Kirkness, *Whole Genome Sequencing*, Methods in Molecular Biology 628, 215-226, 2010

[7] Chris D. Greenman, Erin D. Pleasance, Scott Newman, Fengtang Yang, Beiyuan Fu, Serena Nik-Zainal, David Jones, King Wai Lau, Nigel Carter, Paul A.W. Edwards, P. Andrew Futreal, Michael R. Stratton and Peter J. Campbell, *Estimation of rearrangement phylogeny for cancer genomes*, Genome Research 22, 346–361, 2012

[8] Min Zhao, Jingchun Sun, Zhongming Zhao, *TSGene: a web resource for tumor suppressor genes*, Nucleic Acids Research 22, doi:10.1093/nar/gks937, 2013 Database Issue

[9] Cristian Tomasetti, Bert Vogelstein, Giovanni Parmigiani, *Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation*, PNAS vol. 110 no. 6, 1999–2004, 2013

A List of point mutations present in genes with loss of heterozygozity

Table 5: Point mutations in tumour suppressor genes where one of the copies was lost and at least one point mutation in its exome, sample HCC1187.

Chr	Start	Stop	Gene	HGNC	Copy number	MutAlFreq	Occurrence
2	225414044	225414046	CUL3	8452	2	0.90	pre
2	228099860	228099876	COL4A3	1285	2	1.00	pre
3	71323909	71323910	FOXP1	27086	2	1.00	pre
3	71461732	71461733	FOXP1	27086	2	1.00	pre
4	126388749	126388750	FAT4	79633	2	0.35	post
6	148809264	148809265	SASH1	23328	2	1.00	pre
6	166825416	166825417	RPS6KA2	6196	2	0.39	post
6	166901505	166901506	RPS6KA2	6196	2	1.00	pre
6	167137504	167137505	RPS6KA2	6196	2	0.98	pre
9	8484613	8484614	PTPRD	5789	3	0.20	post
9	8772135	8772136	PTPRD	5789	3	0.68	post
9	8775420	8775421	PTPRD	5789	3	0.30	post
9	9030414	9030415	PTPRD	5789	3	0.97	pre
9	9238314	9238315	PTPRD	5789	3	0.31	post
9	9447429	9447430	PTPRD	5789	3	1.00	pre
9	9783316	9783317	PTPRD	5789	3	0.72	post
9	9819906	9819907	PTPRD	5789	3	0.96	pre
9	9984857	9984859	PTPRD	5789	3	0.98	pre
9	10203169	10203170	PTPRD	5789	3	0.85	pre
9	10573198	10573199	PTPRD	5789	3	0.72	post
9	10584288	10584289	PTPRD	5789	3	0.36	post
9	95988961	95988962	WNK2	65268	3	0.71	post
9	96049666	96049667	WNK2	65268	3	1.00	pre
10	95824364	95824365	PLCE1	51196	2	0.98	pre
12	94575207	94575208	PLXNC1	10154	2	0.84	pre
12	94602668	94602669	PLXNC1	10154	2	0.17	post
13	21243289	21243290	IFT88	8100	3	0.97	pre
13	33848027	33848027	STARD13	90627	3	0.96	pre
16	72820296	72820296	ZFHX3	463	2	0.09	post
16	72917922	72917922	ZFHX3	463	2	0.54	post
16	72939812	72939812	ZFHX3	463	2	0.10	post
16	77439130	77439131	ADAMTS18	170692	2	0.91	pre
16	78232849	78232850	WWOX	51741	2	0.40	post
16	78688999	78689000	WWOX	51741	2	0.96	pre
17	7579362	7579365	TP53	7157	2	1.00	pre
17	34062356	34062357	RASL10B	91608	2	0.97	pre
17	39925273	39925274	JUP	3728	2	0.13	post
19	1526077	1526078	PLK5P	126520	2	0.52	post

20	59843664	59843665	CDH4	1002	2	0.46	post
20	59906670	59906671	CDH4	1002	2	0.21	post
20	59906674	59906675	CDH4	1002	2	0.67	post
20	59924840	59924841	CDH4	1002	2	0.37	post
20	60023314	60023315	CDH4	1002	2	1.00	pre
20	60349376	60349378	CDH4	1002	2	0.97	pre
20	60398401	60398402	CDH4	1002	2	0.40	post
22	26140619	26140620	MYO18B	84700	2	0.68	post
22	45071140	45071140	PRR5	55615	2	0.33	post

Table 6: Point mutations in tumour suppressor genes where one of the copies was lost and at least one point mutation in its exome, sample HCC2218.

Chr	Start	Stop	Gene	HGNC	Copy number	MutAlFreq	Occurence
3	142163601	142163602	XRN1	54464	1	1.00	-
3	142256827	142256828	ATR	545	1	1.00	-
7	77226449	77226449	PTPN12	5782	1	0.03	-
7	146686240	146686241	CNTNAP2	26047	3	1.00	pre
7	146699918	146699919	CNTNAP2	26047	3	1.00	pre
7	146809968	146809969	CNTNAP2	26047	3	0.99	pre
7	147010173	147010174	CNTNAP2	26047	3	0.94	pre
7	147088762	147088763	CNTNAP2	26047	3	0.97	pre
7	147260356	147260357	CNTNAP2	26047	3	1.00	pre
7	147406905	147406906	CNTNAP2	26047	3	1.00	pre
7	147441123	147441124	CNTNAP2	26047	3	1.00	pre
7	147727986	147727987	CNTNAP2	26047	3	1.00	pre
7	147935621	147935621	CNTNAP2	26047	3	0.39	post
7	147949725	147949726	CNTNAP2	26047	3	0.99	pre
8	17556041	17556042	MTUS1	57509	1	1.00	-
8	17560702	17560703	MTUS1	57509	1	0.89	-
16	72825876	72825877	ZFHX3	463	1	1.00	-
16	72897111	72897112	ZFHX3	463	1	0.14	-
16	72969632	72969633	ZFHX3	463	1	1.00	-
16	73021735	73021736	ZFHX3	463	1	1.00	-
16	73025830	73025831	ZFHX3	463	1	1.00	-
16	73086606	73086607	ZFHX3	463	1	1.00	-
16	77369455	77369456	ADAMTS18	170692	1	1.00	-

Table 7: Point mutations in tumour suppressor genes where one of the copies was lost and at least one point mutation in its exome, sample E01.

Chr	Start	Stop	Gene	HGNC	Copy number	MutAlFreq	Occurence
4	96102511	96102512	UNC5C	8633	1	0.62	-
4	96233953	96233954	UNC5C	8633	1	0.48	-
4	96270509	96270509	UNC5C	8633	1	0.04	-
4	126297136	126297137	FAT4	79633	1	0.41	-
4	126308271	126308272	FAT4	79633	1	0.71	-
8	12992855	12992856	DLC1	10395	1	0.47	-
8	15511933	15511934	TUSC3	7991	1	0.12	-
8	15524967	15524968	TUSC3	7991	1	0.27	-
8	22918947	22918948	TNFRSF10B	8795	1	0.50	-
15	53067578	53067579	ONECUT1	3175	1	0.37	-
18	49882858	49882859	DCC	1630	1	0.52	-
18	50063369	50063370	DCC	1630	1	0.54	-
18	50353450	50353451	DCC	1630	1	0.14	-
18	50568184	50568185	DCC	1630	1	0.16	-
18	50661777	50661777	DCC	1630	1	0.53	-
18	50855065	50855066	DCC	1630	1	0.59	-
18	50894253	50894254	DCC	1630	1	0.53	-
X	65828357	65828359	EDA2R	60401	1	0.56	-