EUROPEAN COMMISSION

Community Research

SEVENTH FRAMEWORK PROGRAMME

Systems Biology of Colorectal Cancer

SYSCOL
258 236

Final report

FP7 Collaborative Project

20/02/2016

SYSCOL final publishable summary report

Executive summary

Colorectal cancer (CRC) is one of the most commonly occurring cancers, affecting over half a million individuals in Western countries each year. In Europe, it is the single most important cancer in terms of severity, number of cases and cost for society. At present, there are no effective screening programs or therapies and despite much research there is no clear readily avoidable cause. At the point that our research project was launched, several genes that increase risk for colorectal cancer had been identified, but as large-scale genomic data was not available, and whole genomes of CRCs had not been sequenced, many mechanisms that contribute to this common disease had not been identified.

The SYSCOL project was specifically set up to combine basic and clinical research, in order to generate and integrate genomic data on colorectal cancer. During the course of the project we have genetically characterised colorectal tumor patient samples and corresponding normal tissue to identify and validate inherited genetic alterations that predispose some individuals to colorectal cancer. Some of these have undergone analyses in which the mechanisms leading to uncontrolled cell growth have been identified and validated. We have also mapped modifications to the DNA itself that can affect gene activity, and have analysed how these contribute to CRC.

The generated data has been integrated into a computational network-model. The model has additionally been fed with available information regarding risk-variants and mutations known to be critical for colorectal cancer formation. The model describes and explains the information flow from inherited or acquired mutations, via the intermediate steps that lead to activation of downstream genes that drive uncontrolled cell growth and tumor formation. Using this model we have developed a first draft of a computational predictor to predict an individual's risk of developing CRC based on his/her genome sequence.

The work within the project has resulted in major scientific achievements with clear medical importance. The mutual contribution of all partners has generated approximately 56 peer-reviewed publications, nine of which are in top-journals such as *Cell, Science, Nature and Nature Genetics*.

Several of the project findings have the potential to improve the future treatment and survival of patients with CRC. Novel risk variants identified within SYSCOL have already been translated to the clinical setting, and are now routinely used to identify individuals with increased risk. In addition to identifying rare variants that dramatically increase individuals' risk to develop CRC, we also developed a quantitative model predicting the combined effect of risk variants that are very common in the population. Although this model needs to be further validated in the clinic, the results indicate that it represents a successful first step towards assessing risk of developing CRC in the population.

The acquired data has also been used to classify CRC into five molecular subgroups and to identify a set of prognostic biomarkers for each subtype. The classification of CRC has the potential to guide treatment decisions and predict prognosis within the near future. We have also developed methods for minimally-invasive detection of tumor DNA in blood samples, identified mechanisms of treatment sensitivity and resistance and evaluated new treatment approaches in colorectal cancer. We expect that the new avenues for therapeutic intervention identified within the project will have clinical benefits once they have been tested in clinical trials. Finally, we anticipate that the significantly increased mechanistic understanding of CRC will help in identifying novel therapeutic targets, which will have a long-term impact on human health.

Project context

Colorectal cancer (CRC) is the third most common cancer world wide, and the second leading cause of cancer death in the developed countries. Over a million patients are diagnosed with colorectal cancer each year and the five-year survival is only at around 55 %. The risk of developing colorectal cancer increases with age and the incidence is likely to increase as populations' age and more of the world adopts a Western life style.

Although a large amount of work concentrated on understanding the underlying cause have resulted in the identification of a set of genes that increase risk for colorectal cancer, many mechanisms that contribute to this common disease have not been identified yet.

Colorectal cancer

The lower part of the digestive system is known as the colon, of which the last 15 cm is called the rectum. Cancer of the colon and rectum – colorectal cancer – begin as polyps that grow on the inner lining of the large intestine.

Most sporadic cases of colorectal cancer are believed to develop from benign adenomas (polyps) to carcinoma by the accumulation of genetic abnormalities (**Figure 1A**). However, only a small percentage of adenomas progress to carcinomas and the time period required for this development is lengthy, with a minimum of 5-10 years.

The majority of all colorectal cancers occur sporadically without any known cause, but certain groups of people have a predisposition to the development of cancer of the large intestine. These people may carry specific genetic mutations or have relatives with the condition. Approximately 15% of all colorectal cancer cases are familial, with the most common inherited conditions being familial adenomatous polyposis (FAP) and hereditary non-polyposis colorectal cancer (HPNCC) (fig. 1B). The familial cancers FAP and HNPCC both have an early onset, whereas the sporadic cases appear much later, around 60 years of age.

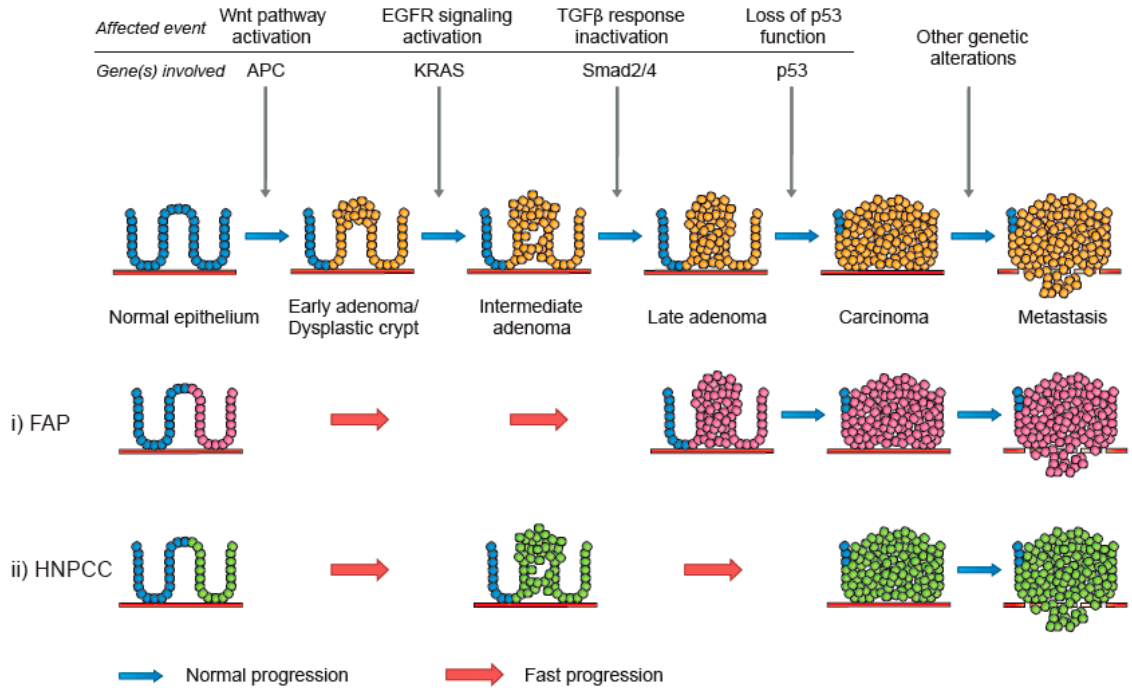

Figure 1. Progression from normal epithelium to colorectal cancer requires an accumulation of mutations in particular genes that affect the balance between proliferation and apoptosis. A. The steps in development of sporadically occurring cancer in a normal colon epithelium. However, not all colorectal tumours exhibit all the mutations shown in the above figure. B. Individuals with an hereditable cancer predisposition. i) FAP: germline inactivation of one APC allele. Adenoma formation is faster, but progression from adenoma to carcinoma has the same rate as sporadic colorectal cancer. ii) HNPCC: germ line inactivation of one allele of either of the mismatch repair genes MSH2 or MLH1 in combination with somatic inactivation of the other allele leads to an increase in the mutation rate, which in turn speeds up the adenoma to carcinoma progression. *Adapted from Davies, R. J., et al. 2005. Colorectal cancer screening: prospects for molecular stool analysis, Nature Review Cancer 5:199-209*

Screening and treatment

Colorectal cancer can take many years to develop and early detection greatly improves the chances of a cure. Screening helps to detect cancer, polyps and smaller lesions in the colon. If the screening reveals a problem, diagnosis and treatment can occur promptly. There are several screening methods available, including a rectal exam, fecal blood test and colonoscopy.

Colonoscopy (endoscopic examination of the lower part of the digestive system) is highly effective but it is expensive and invasive and population screening programs are therefore primarily based on Fecal Occult Blood Screening (detection of small amounts of blood in the stool) which has lower sensitivity.

The treatment of choice depends on how far the cancer has progressed. Surgery is the primary treatment for the removal of polyps and/or cancer tumors. Chemotherapy can be used to limit the spread of the cancer to other parts of the body. Radiation therapy is less common in colon cancer, but is sometimes used to treat rectal cancer, as the rectum does not move as much as the colon. However, despite advances in the management of colorectal cancer over the last 25 years, 5-year survival remains at around 55 %.

The molecular basis of colorectal cancer

Sets of genes that control a specific cellular function are called pathways. There are two types of pathway involved in driving colorectal cancer development. The first type contains so called tumor suppressors and oncogenes. Tumor suppressors are genes whose protein products have a negative effect on cell growth or can promote cell death. Oncogenes are genes whose protein products, when activated, cause the cells to divide uncontrollably or prevent cells from undergoing apoptosis. In the normal colon the balance between cell birth and cell death is tightly controlled to ensure that the total number of cells is constant. When the genes controlling cell growth pathways are altered through mutations, the balance is shifted and can result in uncontrolled growth and tumor formation.

The other type of pathway contains so called stability genes, involved in controlling the integrity of the genome and controlling the rate of mutations. Defects in these pathways result in the accumulation of mutations and damage to the genome, which in turn accelerates the tumor progression.

A number of genes belonging to both types of pathway have been identified to play a role in colorectal cancer formation, including the APC/Wnt pathway, KRAS and EGFR signalling pathways, the BMP and TGF-Beta signalling pathways and p53.

The SYSCOL project

To advance the understanding of colorectal cancer, we proposed to use a systems biology approach to model the complex biology of colorectal tumor formation. The project we outlined was based on patient samples and data gathered using both computational and experimental methods. The overall strategy was to collect and characterise CRC patient samples and use these samples to systematically generate genetic information in order to identify mutations and variations that increase an individuals risk to develop cancer. In parallel, computational tools were to be developed, that could predict how the variants can contribute to cancer.

The genetic variants and computational predictions would be used to identify the mechanisms that are required for colorectal cancer growth and to develop a mechanistic model for CRC formation. The identified mechanisms and predictions would then be functionally tested and the result used to guide the data generation efforts. Some of the findings would also undergo clinical validation. The final integrated CRC model would be used to predict CRC risk based on an individuals genome sequence and allow a classification of CRC into a number of subtypes.

Our multidisciplinary research project SYSCOL brought together a strong team of researchers with complementary expertise in computer science, statistics, medical genetics, genomics and systems biology from across Europe and the US. A list of the 11 project partners and their contact details is attached.

Objectives

The overall objective for the SYSCOL project was to advance the understanding of the formation of colorectal cancer and increase the likelihood of the development of effective prediction tools and therapies, with the ultimate aim of reducing mortality rates.

SYSCOL aimed to systematically map out the changes and variations in the genetic code that increase individuals' risks of developing colorectal cancer, using Systems Biology tools. The variants were to be used to identify the mechanisms that are required for colorectal cancer growth and to develop a model for colorectal cancer formation. The model would describe the cellular pathways that contribute to tumor formation and explain in detail how the genetic disposition of an individual can activate the expression of genes that cause uncontrolled cell growth and lead to cancer. This model could subsequently be used to discover novel therapeutic targets, guide genetic screening in order to identify individuals with a heightened risk for developing colorectal cancer and to classify patients into subgroups in order to personalise medical treatments. This information could be used to identify novel treatment targets that are more susceptible to drugs than the currently known pathways.

Main S&T results/foregrounds

The work within the SYSCOL project was divided into nine different work packages, each headed by one of the partners. The different work packages are tightly connected to and dependent on each other and on the exchange of samples and data generated. An overview of the workflow is shown in **Figure 2**. The work and achievements of the different WPs is described below.

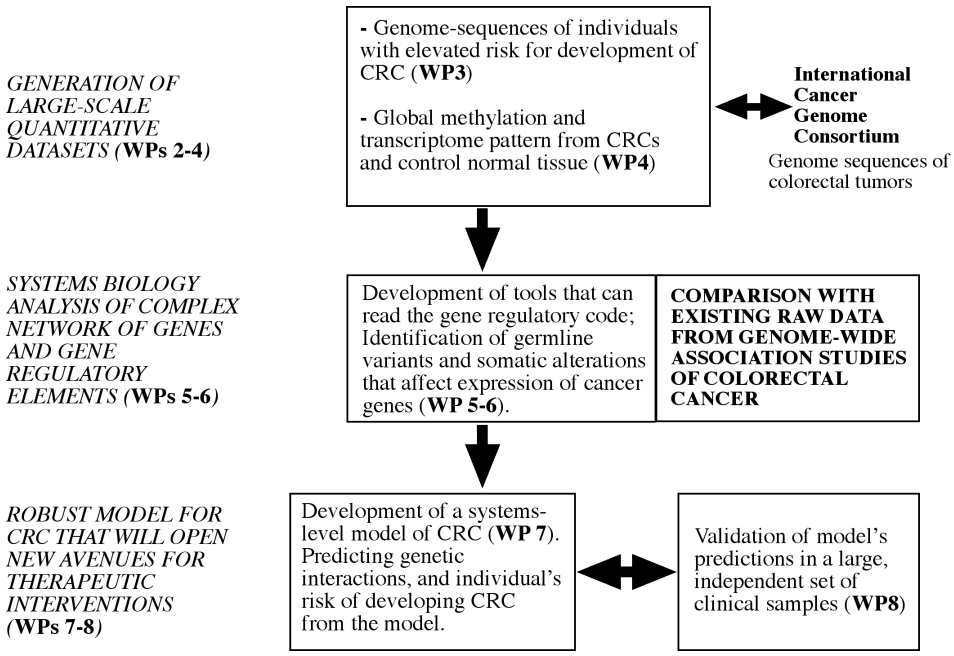

Figure 2. SYSCOL workflow.

WP2 Sample collection

Collecting large and well-characterized materials is a prerequisite for every successful project relying on patient samples. The goal for this work package was to collect colorectal carcinoma (CRC) samples and their respective normal tissue (colon/rectum epithelia) for SYSCOL molecular analyses. The aim was to collect 300 tumor-normal pairs per year for four years, i.e. 1200 pairs and the collection was performed by Partners Ørntoft, Aaltonen and Velculescu.

The samples underwent basic characterisation and stratification. Samples with known high-penetrance CRC predisposition mutations in genes such as *MLH1, MSH2, MSH6, LKB1*, or *APC* were excluded from the study.

Since the start of the project, samples were collected from a total of 1600 patients, fulfilling and exceeding the original goal. The samples have been processed to generate high quality DNA and RNA for molecular analyses in WPs 3 and 4 (for more details see under the respective WP). Additional samples from the collection, together with samples from existing biobanks, have been used for the functional validation in WP8. By completing the intended sample collection and combining the set with the existing resources, we significantly increased the statistical power to test our hypotheses and ensured that adequate numbers of samples from both hereditary and sporadic CRC were included.

Data on the patient samples have been placed on a project database that will be operated according to existing rules and regulations. Clinical data were carefully recorded, to profoundly characterize the natural history (such as typical age at onset, etc.) of the cancer predisposition phenotypes detected during the study. All data in the database was anonymised before entering, and personal information thus protected.

Taken together, WP2 was very successful and resulted in a sample base significantly larger than that envisioned at the start of the effort. The samples were forwarded to WPs3 and 4 for molecular characterisation and to WP8 for clinical validation.

WP3 - Genotyping and sequencing

Workpackage 3 was principally concerned with how inherited factors (genes) influence the risk of an individual developing colorectal cancer (CRC). Variation in certain genes has been known for some years to alter the chance of developing CRC and the major aim of WP3 has been to identify new CRC risk genes. It was initially envisaged that there would be two types of new CRC gene: (i) genes in which rare variants conferred a high risk of cancer; and (ii) genes influenced by common variants that individually contributed a low or modest amount to cancer risk. A third intermediate class of gene, in which uncommon variants caused a moderately increased cancer risk, was also investigated during the project. The information gained contributes not only to our understanding of CRC in general, but also to its classification for purposes such as patient survival after treatment. The genetic information generated in in this work package was integrated with the data generated in WP4 and transferred to WPs 6 and 8 for functional and clinical validation and to WP7 where it was integrated in the CRC model.

New high-risk genes for colorectal cancer

New technology meant that sequencing the whole genome or the part of the genome that coded for proteins (exome) was a viable way of identifying new high-risk CRC genes. The expense of the method meant that focussing on patients most likely to carry these genes (strong family history of CRC, multiple bowel tumours/polyps, or CRC at an early age) was a useful strategy. To date, several hundred such patients have been exome- or genome-sequenced within SYSCOL (Partners Tomlinson, Aaltonen, Velculescu and Houlston) to identify genes that predispose to CRC.

Our first finding was that a small number of patients carried inherited mutations in one of the previously-identified CRC genes, such as APC and a set of 4 genes involved in repairing DNA (mismatch repair (MMR) genes). Although the patients had been screened for these genes previously, older methods were not capable of searching for all cancer-causing changes, and we were therefore expecting to find some of these changes. However, <5% of our patients carried changes in these known genes.

We then focussed on our families who met all three of our criteria for study: the most extensive family history of CRC; several individuals with tens of bowel polyps with the potential for becoming cancerous; and one or more patients affected at <35 years of age. This led to the identification of inherited cancer-causing changes in two related genes, POLE and POLD1 (**Figure 3**) (Palles et al. Nat Genet 2012). These genes make proteins that are centrally involved in copying DNA when cells divide, a process that occurs billions of times every day to keep the body healthy. The specific changes in POLE and POLD1 did not, however, prevent this copying. Instead, they allowed more errors than usual to occur, causing the body to acquire mutations that eventually cause cancer. Patients can develop not only CRC, but also cancer of the uterus and, less often, cancers of other sites. Since the SYSCOL data were reported, our findings have been validated by other groups, and families are now tested routinely for POLE and POLD1 mutations in the clinic, alongside APC, MMR genes, and other CRC genes. We have set up a database of inherited POLE and POLD1 changes (*www.lovd.nl/POLE, www.lovd.nl/POLD1).*

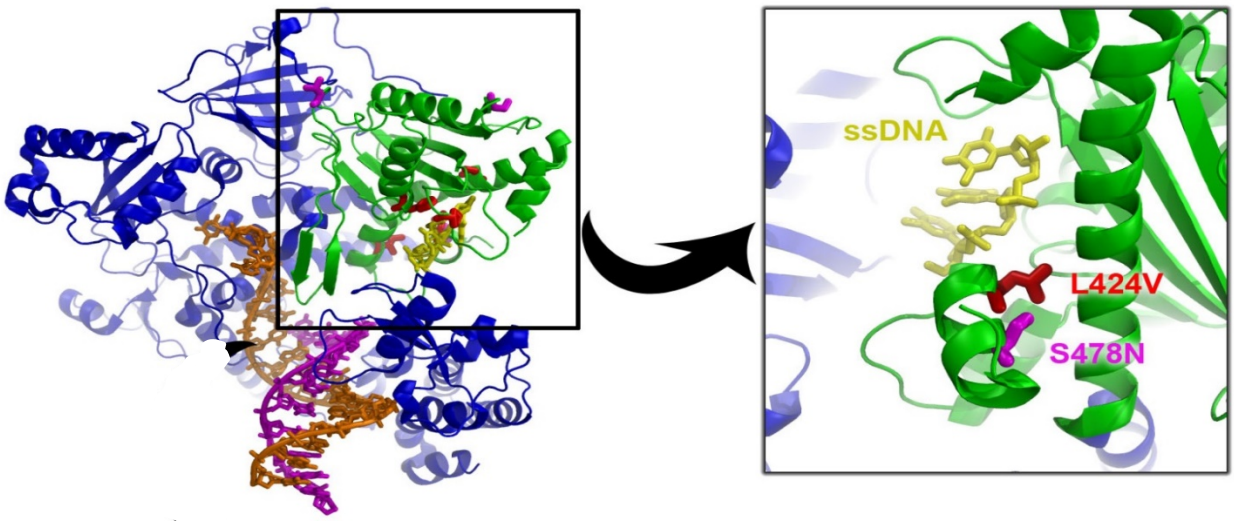

Figure 3. A representation of the Pole and Pold1 proteins showing the two inherited mutations found in SYSCOL, POLE L424V (red) and POLD1 S478N (purple). The yellow represents an error in DNA copying. The two mutant sites lie close to this DNA, consistent with less effective removal of the error and an increased number of mutations.

In parallel with our work, non-inherited *POLE* mutations acquired during colon cancer growth were reported by other groups. We also reported non-inherited *POLE* mutations in endometrial cancers. For unknown reasons, *POLD1* mutations almost always seem to occur in the inherited setting. We

have shown that cancers with *POLE* mutations usually have a very good prognosis, and we are making attempts to use this to spare some patients unnecessary chemotherapy or radiotherapy.

Analysis of the SYSCOL sequencing data continues. Several candidate CRC genes have been reported, and we hope to validate these and identify additional high-risk genes in the coming months.

Common genetic variants and colorectal cancer risk

Prior to the start of SYSCOL, we had identified a set of about 15 sites in the genome at which the DNA sequence varied among members of the population. These variants (called single nucleotide polymorphisms, or SNPs) comprised single letter (ACGT) differences between people, such that these differences were associated with a small (typically 10% per variant) increase or decrease in the chance of developing CRC. In contrast to the high-risk genes, almost everyone has a mixture of high and low risk variants. SYSCOL provided an important opportunity to take this work forward, with the eventual aim of identifying as many of these common CRC SNPs as possible.

Quite early in the Project, Partners Tomlinson, Aaltonen and Houlston, working with other groups outside SYSCOL, identified three more of these SNPs (Dunlop et al. *Nat Genet* 2012). Although the SNPs do not typically occur within the genes themselves, and therefore do not affect protein sequence, they probably affect protein levels. After analysis of tens of thousands of individuals, additional CRC SNPs were found (see Table below). The new SNPs found in SYSCOL account for ~5-10% of the inherited risk of CRC, suggesting that many additional CRC SNPs remain to be found.

SNP	Nearby gene(s)	Comments
rs3824999	POLD3	Involved in copying DNA. Closely related to POLD1 (see above)
rs1321312	CDKN1A	Central to cell response to DNA damage or stress
rs5934683	SHROOM2	Gene function not well known
rs1035209	NKX2-3	Gene function not well known
rs3217810	CCND2	Involved in causing cell to divide
rs10911251	LAMC1	Controls tissue structure
rs72647484	WNT4, CDC42	Involved in growth and repair of intestine
rs16941835	FOXL1	Involved in causing cell to divide
rs992157	PNKD, TBIM1	Involved in cell death
rs3184504	SH2B3	Involved in inflammation

We also provided data for an examination in WP4, 5 and 7 of how SNPs affect gene expression (when a gene is turned on and off).

One of the findings from our work has been that there are often multiple SNPs that affect CRC risk independently or semi-independently around a gene. We have performed particularly detailed examination of this possibility near a gene called Gremlin1, and found three SNPs that affect CRC risk (Lewis et al. *Cell reports* 2014). In fact, at least 4 other CRC risk SNPs exist at genes that interact with

Gremlin1, showing that these probably affect a function of special importance for preventing CRC in the human body.

Search for intermediate risk genetic variants

A search was performed for uncommon genetic variants (typically present in 0.1-1% of the population) and CRC. This search, which involved SYSCOL and external groups, yielded the *SH2B3* CRC SNP shown in the Table above. It also identified (specifically in Dutch families) a candidate gene *NTHL1* shown by others in parallel studies to be a new high-risk CRC gene. *NTHL1* mutations are extremely uncommon in countries such as the UK and Finland, but the first cases have been reported by Partners Aaltonen and Houlston in these countries.

WP4 – Methylome and transcriptional state of tumors

It has been shown that modifications to the DNA itself, so called epigenetic changes, can also contribute to cancer. These types of changes that include the addition of methyl groups to the DNA molecule and modifications of the histone proteins that the DNA is packed around can affect gene expression, i.e. how often the information in the gene is used to make the protein it encodes. The aim of this WP was to determine whether changes in methylation, transcription factor expression or packing of the DNA correlate with altered gene expression and the development of colorectal cancer. The work was performed by Partners Esteller, Dermitzakis, Velculescu and Ørntoft.

The DNA methylome

To identify DNA methylation patterns related to CRC development two complementary strategies were used.

Using a high throughput bisulfite sequencing approach Partner Esteller obtained a detailed and comprehensive insight in the DNA methylation variation between normal and primary tumor tissue, and metastatic tumor from one CRC patient. Our results suggest that there is a progressive loss of DNA methylation (hypomethylation) as the cancer progresses. We detected a more unevenly methylated genome in the metastatic tissue as compared to normal tissue and primary tumor (**Figure 4A**). We also detected large less methylated areas in chromosomes 13 and 21. Overall we identified 173,816 regions that were differentially methylated in the normal tissue as compared to the tumor sample. Despite the global loss of methylation in the tumor, hypermethylation (increase in methylation) was observed in promoter zones, the regions of the DNA that regulate the expression of genes (**Figure 4B**). These results strengthen the idea that methylation changes in regulatory promoter regions (mainly, hypermethylation of tumor suppressor genes) plays a crucial role in colon cancer development, through gene silencing.

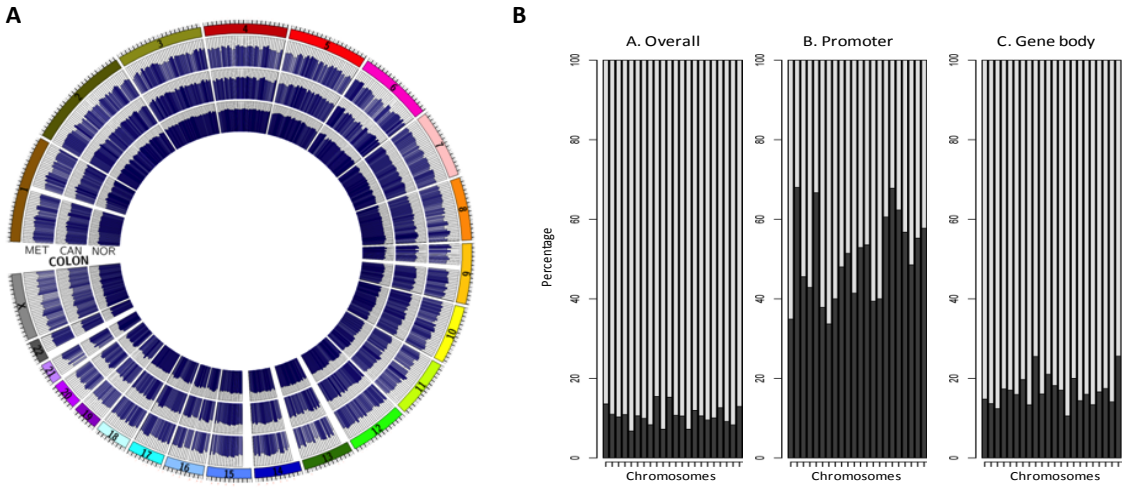

Figure 4. (A) Circos plot of genome-wide DNA methylation levels in the normal, CRC primary tumor and liver metastasis from the same individual at all individual chromosomes. **(B)** Percentage of DMRs that gained/lost in different genomic features: overall (B.A); promoter regions (B.B) and Gene Body (B.C).

To obtain a global picture of the DNA methylome in CRC, we used the Infinium Human Methylation450 Bead Chip platform from Illumina to study the genome wide methylation profile of a large series of CRC patients. In total we profiled the methylation pattern of 292 colorectal tumors and 28 non-tumor paired normal colon tissue. Strikingly, within each sample group (tumor and controls) the overall patterns of methylation were similar, indicating that the DNA methylation profile is altered during tumor development (**Figure 5**). We identified 31,909 differently methylated regions (DMR) including 17,047 hypomethylated and 14,862 hypermethylated DMRs in tumor samples compared to normal. As we had observed using the Whole-genome bisulfite sequencing approach, hypermethylated CpGs were mainly located in promoters, whereas hypomethylated CpGs were situated in intergenic regions. This observation is in concordance with previous observations, where it has been described that the global DNA hypomethylation is associated with genome instability and chromosome fragility and that it is accompanied by hypermethylation at specific promoter regions of tumor suppressor genes. To conclude, our methylome analysis has allowed the identification of hypermethylation DMRs in known and potential tumor suppressors. We also obtained a list of hypomethylated DMRs that may be correlated with potential oncogenes. Our genome wide data provides new clues about the regulation of potential tumor suppressor and oncogenes in colorectal cancer and can provide relevant clues for understanding gene regulation.

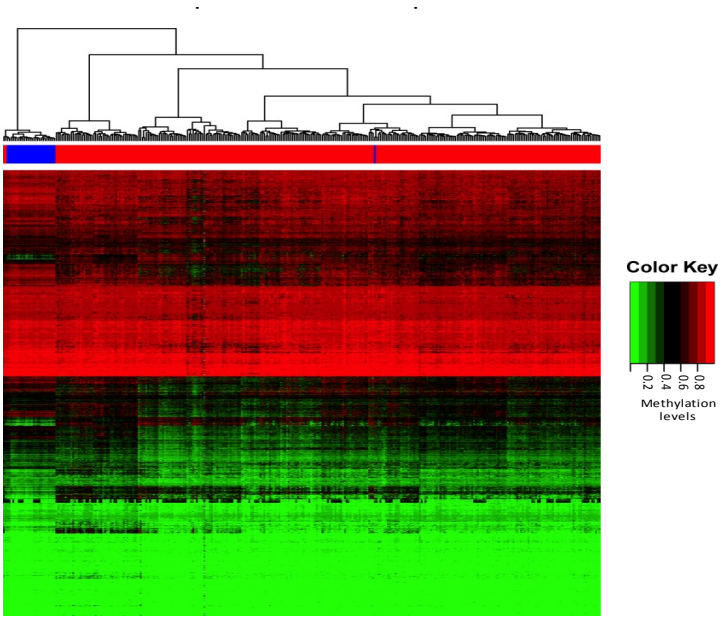

Figure 5. Unsupervised hierarchical clustering. Colorectal (red) and control (blue) samples are clustered separately. On the right, a scale for methylation levels is provided.

The DNA methylation data was forwarded to WP7 and used for the colorectal cancer model and also to WP8 where it contributed to the characterization of CRC sub-classes and identification of prognostic biomarker candidates.

Gene expression analysis

Colorectal cancer results from a combination of genetic alterations (mutations). Mutations can occur in both the protein coding genes, that make up approximately 2 % of the DNA, or in the non-coding DNA. It is relatively easy to predict the target genes for mutations that occur in gene coding regions. However, many of the CRC risk variants identified so far are located in the non-coding regions of the DNA. How these mutations lead to cancer growth is more difficult to predict, in particular when the causative variants are located far from genes.

So called enhancers or regulatory elements lie within the non-coding DNA. These are regulatory DNA sequences that, when bound by specific proteins called transcription factors, control how often an associated gene is turned on or off, and that different proteins are expressed during specific cellular processes or in specific types of cells. Recent studies have shown that mutations in these regulatory regions can lead to aberrant expression of target genes that in turn contribute to cancer.

In WP4 we (Partner Dermitzakis) examined this large non-coding regulatory region of the genome, in search of non-coding regulatory elements that impact CRC development. To our knowledge this is the first study examining the noncoding genome on this scale. In collaboration with Partners Ørntoft, Esteller, Tomlinson and Houlston, we found two classes of mutations with an impact on CRC development that were previously unknown (Ongen et al. Nature 2014). Firstly, we discovered regulatory variants inherited by one's parents, which are not active in healthy tissue, but become activated in tumours and potentially drive the progression of cancer. Secondly, using a novel approach of analysing gene expression from normal and tumour tissues, we detected the effect of somatic (acquired) regulatory mutations in the non-coding regions of the genome that affect CRC tumorigenesis. Moreover we have improved this method of detecting the effects of somatic

regulatory mutations so that it can be applied to cohorts where the normal and the tumour samples are not matched. Both of these classes of genomic changes are novel with respect to cancer.

Our study highlights a novel methodology that can be used to discover regulatory effects that drive cancer progression and are of comparable importance to the protein altering mutations previously described. This approach may also be applied to other cancer types and will ultimately lead to a more comprehensive understanding of cancer development and open the door to a wider range of options for treatment.

The data generated in WP4 has been the foundation for the predictions and models built in WP7 and for the investigation of the interplay between the genome, epigenome, and gene expression and their combined effect on CRC development.

WP5 Regulatory element models

The goal of WP5 was to improve our understanding of the role of regulatory elements (described in WP4) in colorectal cancer. In order to achieve this, we (Partners Segal, Taipale, Dermitzakis, Ukkonen, Stunnenberg and Kel) developed computational approaches for the identification of regulatory variants (mutations) in the human genome, which are predicted to affect gene expression or colorectal cancer (CRC) development. The research focused on constructing a framework for predicting variations in gene expression from genetic variation among humans. Such an ability to predict expression from the genetic make up is key to understanding which variants affect expression, and how. We successfully integrated computational predictions based on various DNA sequence features and other genomic data, and the framework to predict variants that affect gene expression in CRC patients.

Among our key achievements were the establishment of and improvements to databases and tools used to analyse extremely large datasets efficiently, further developing and improving tools for identifying short DNA sequences that are recognized by DNA binding proteins, and the development of an integrative tool incorporating information of transcription factor binding and DNA packaging with gene expression, thereby linking epigenetic mechanisms with transcription factor binding (Manor and Segal *PLoS Genetics* 2013). The tools were used successfully on SYSCOL data from WPs 3 and 4 to identify a set of regulatory SNPs that are involved in colorectal cancer and to analyse regulatory DNA alterations around oncogenes. Most of the tools are publically available via open-access web-based services and can be accessed via links posted on the SYSCOL website.

As part of WP5 we (Partner Ukkonen) further developed and improved our novel tool for footprinting short DNA sequences that bind DNA binding molecules (EEL). EEL works by first predicting regulatory elements by aligning human (or other chosen centre species) with all other species using the pairwise EEL enhancer model (**Figure 6**) (https://www.cs.helsinki.fi/u/kpalin/EEL/).

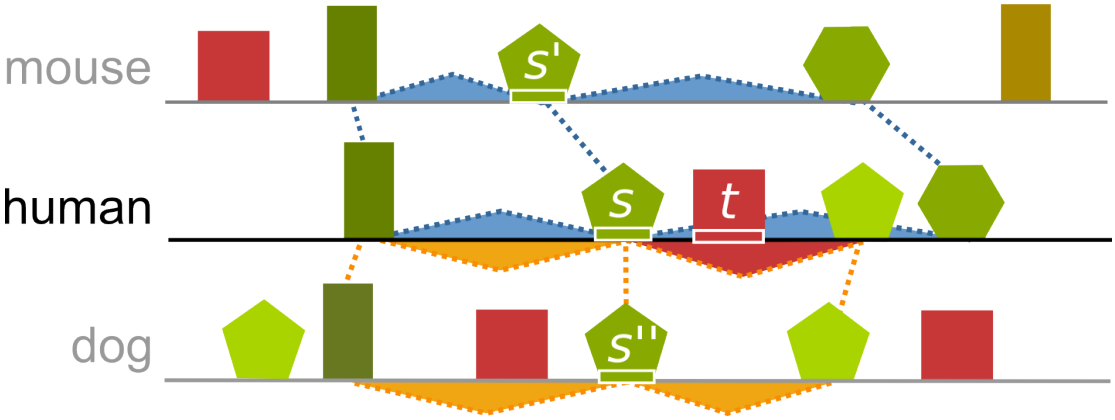

Figure 6. Schematic view of EEL footprinting. The site 's' in the human genome will have a high footprinting score as it is part of mouse (blue) and dog (orange) alignments that approximately preserve the order and the distances of sites. On the other hand, the site 't' would a have low footprint score even though the transcription factor (red box) has a binding site in all orthologous sequences.

In addition, we developed several novel algorithms for finding accurate binding affinity models for transcription factors binding to the DNA in pairs. These models have the potential to give insight to how mutations in the regulatory cluster may change the expression of the related gene.

Experimental methods to test the predicted roles of genetic variations in the regulatory regions in a high-throughput manner were also developed and used (Partners Segal and Stunnenberg).

In summary, this WP was highly successful in achieving the tasks as set out, including developing models for how alterations in the genome affect the properties of a cell, models for integrating both regulatory and epigenetic information, models for identifying functional enhancers within the human genome and models for being able to 'read' the regulatory information encoded in the DNA sequence. The predictions were used in the validation and integration of novel CRC models in WP7.

WP6 Validation of CRC genes and regulatory elements

The primary goal of WP6 was to validate novel findings (oncogenes, regulatory elements and genetic variants) identified in and outside the project, and to identify the proteins that actually bind to the predicted causative regulatory variants identified in WP5. The work was performed as collaboration between partners Taipale, Velculescu, Stunnenberg and Houlston, and the generated data has been and will continue to be used to interpret cancer genome data generated both in and outside the project. The results have also been used to improve the computational tools developed within WP5 and for the CRC-model in WP7.

Validation of novel CRC oncogenes

Work within the SYSCOL project has led to the identification of several novel germline mutations and risk-variants associated with colorectal cancer. In order for these findings to be translated to the clinical setting we need to understand the functional consequences of the identified risk variants. A variety of assays were used to assess the roles that the mutated genes/risk-associated variants play in CRC. A few examples of the validation studies are described below.

POLE and POLD1 encode proteins that are centrally involved in copying DNA and correcting potential errors when cells divide. The mutations we identified in WP3 were mapped to sites in the proofreading domain and were predicted to impair the correction of mispaired bases during DNA copying. Functional validation in yeast revealed that the POLD1 mutation resulted in a 12-fold higher mutation rate. Consistent with these findings, *POLD1*- and *POLE*-mutant tumors had an increased rate of base substitution mutations (Palles et al. *Nature Genet* 2012).

Another risk variant is located in a region associated with increased risk for developing colorectal and prostate cancers, but whose mechanism of action was unclear. Although this variant increases cancer risk by only 20 per cent, it is extremely common and therefore accounts for more inherited cancer than any other currently known genetic variant or mutation. To validate the function of this variant we removed the gene region containing the risk variant from the mouse genome, and found that as a result the mice were healthy but displayed a small decrease in the expression of a nearby cancer gene, called MYC. However, when these mice were tested for the ability to form tumours after activation of an oncogenic signal that causes colorectal cancer in humans, they showed dramatic resistance to tumor formation. The removed gene region thus appears to act as an important gene switch that promotes cancer, and without it tumors develop much more rarely (Sur et al. *Science* 2012).

The functional studies also resulted in the identification of a number of candidate CRC causing genes, in addition to the already known common CRC drivers such as APC, TP53, PIK3CA and SMAD4. These findings were included in the CRC model in WP7 and used in WP8 to identify connections between genomic changes, clinical characteristics and novel therapies (Bertotti et al. *Nature* 2015).

Identification of proteins that bind to genetic variants

A high throughput screening platform was developed to investigate if the CRC risk-variants identified in WP3 are located in regions bound by transcription factors, and to identify the proteins that actually bind to the predicted causative regulatory variants (Hubner et al. J Proteome Res 2015). Using this platform, 116 risk variants have been screened and more than 10 possibly functional sites identified.

Validation of regulatory elements and TF binding sites

Another important aim of this WP was to identify and validate the regulatory elements and transcription factor binding sites (identified in WP5) required for colorectal cancer (CRC) growth. This data will help to interpret cancer genome data generated both outside the project (ICGC) and within WP3. The generated results were also used to improve the computational tools developed within WP5 and the CRC-model in WP7.

The location of elements that regulate gene expression, can be seen by using a set of transcription factor binding markers. We performed genome-wide location analyses to map a set of markers and key CRC transcription factor locations in several colorectal cancer cell lines and patient samples. This information is essential to describe the transcriptional regulation in colon cancer cells. We found that less than 1% of the colorectal cancer cell genome is occupied by transcription factors. The protein cohesin is present in almost all of the TF clusters and thus can be used to mark the regions where most of the CRC-driving mutations are likely to be found (Yan et al. Cell 2013).

Moreover, we identified the DNA sequences that bind to over four hundred proteins that control gene expression (Jolma et al. *Cell* 2013) and long range interactions between regulatory elements and distant target genes (Jäger et al. *Nat Commun* 2015). This knowledge is required to understand how the differences in individuals' genomes affect their risk of developing disease

Overall, through the work in WP6 we have gained insights into the functions of a number of genes and regulatory elements important in CRC development. The validated findings were forwarded to and included in the CRC model described in WP7.

WP7 Modelling

The aim of WP7 was to combine existing information about inherited risk-variants and tumor mutations with the information generated in **WPs 3-6** into a model of CRC which explains how the inherited and tumor genotypes of individual CRC cases can cause uncontrolled cell growth and cancer. To ensure that the model was built on solid knowledge and that high quality information on signalling pathways that are known to be important for CRC formation was incorporated, an expert curated CRC model was built in parallel to the computational model. Partners Taipale, Segal, Ukkonen, Aaltonen, Velculescu and Kel formed the core of the modelling team.

The CRC model

The resulting CRC model consists of several different modules, explaining the different aspects of CRC (**Figure 7**).

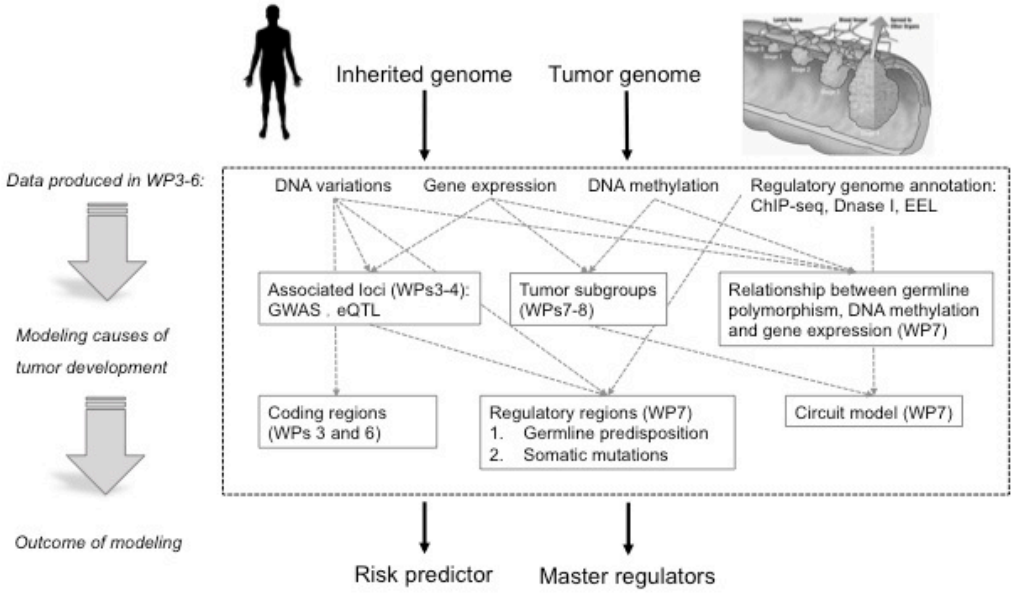

Figure 6. Schematic overview of the data integration for the CRC network model

1. Modelling of regulatory mutations

This part of the model explains in molecular detail how the genetic make up of an individual tumor in concert with the inherited genetic make up leads to activation of target genes, which in turn results in uncontrolled growth and cancer.

By combining the molecular data generated in WPs 3 and 4 with the CRC risk data from WP3 and literature, and the predicted and experimentally validated regulatory elements from WPs 5 and 6, we can predict target genes that are affected by variants in the noncoding DNA. We also modelled the mutational process that occurs in colorectal cancer to unravel the underlying mechanisms of regulatory mutations in cancer.

2. Modelling of the relationships between inherited mutations, DNA methylation and gene expression

Changes in DNA methylation can affect gene expression and thereby play a role in cancer formation. To model the relationship between genetic information, gene expression and methylation, we integrated the detailed DNA methylation pattern information generated in WP4 with gene expression profiles and genetic information obtained from the same patients as in WPs 3-4. We found that methylation changes can cause both gene expression variation or be the consequence of gene expression levels. However, in the majority of cases DNA methylation has a passive role.

3. Circuit model - from input to output

Using the geneXplain platform (http://genexplain.com/genexplain-platform-1) we have built a CRC circuit model, which works as a wiring diagram that describes the information flow from input to output, i.e. how cancer driving mutations lead to aberrant expression of target genes that in turn results in uncontrolled cell growth and cancer.

Within WP8 we have identified five molecular and clinically homogenous subtypes of CRC and shown that the prognosis of the subtypes were significantly different, indicating that they should not be managed and treated in the same way. The subclassification (described in more detail under WP8)

was based on differences in gene expression and DNA methylation patterns as well as on clinical information. With these subclasses as a starting point we constructed molecular networks that describe the altered signalling pathways, the upstream molecules, the CRC target genes, intermediate genes that transfer the signal and the master regulators for each subclass. Each network contains feedback loops and contains information of the types of biological processes the network includes.

4. Risk predictor

The model has also been extended with a risk predictor function, which uses genetic make up as input and gives as output the relative risk of an individual to develop CRC. The predictor is based on project derived and published CRC associated single nucleotide variants. Although the risk predictor needs further testing, the results indicate that this is a successful first step towards CRC risk assessment that can be used in the clinical setting.

5. Expert curated model

We have modelled the different cellular processes that when altered can lead to selective growth advantage and colorectal cancer by collecting the project generated and published information on genetic risk variants, mutations accumulated in CRC primary and metastatic tumors and classifying the commonly affected genes into signalling pathways.

The integration of molecular data generated both in and outside of the SYSCOL project and the development of a CRC model has significantly increased our understanding of the molecular mechanisms of colorectal cancer. In addition, the model has the potential to be further developed and used in the clinical setting to assess individual's risk of developing CRC. We also expect that the increased knowledge of the pathways and regulatory regions whose activity is required for CRC cell growth will help to identify novel druggable treatment targets.

WP8 Clinical validation

The aim of WP 8 "Clinical validation" was to validate and translate the basic research findings of the SYSCOL project into clinically useful tools. Despite advances in the management of CRC over the last 25 years, only around 55% of the people diagnosed with CRC are alive 5-years after diagnosis. We (Partners Ørntoft, Velculescu, Aaltonen and Houlston) have developed a range of tools and approaches that have the potential to improve treatment and survival of CRC patients in the coming years, these are outlined below.

Improving the cost-efficiency of CRC screening by targeting it to those most likely to benefit

The majority of CRCs develop over many years. They begin as polyps that sequentially grow and develop into cancer. As the earlier stages are more treatable and there is a strong correlation between stage and survival, early detection through screening has the potential to significantly improve survival. Accordingly screening programs have been or are being implemented in many countries. In order to make screening programs more cost-effective, today most of them are targeted towards the aging population (typically those above the age of 50). In SYSCOL we have identified genetic variants associated with an increased risk of developing CRC. Individuals carrying these variants in their genomes have up to 7.7 times higher risk of developing CRC than individuals not carrying these variants. In WP 8 we investigated the potential of using this information to target screening towards those individuals most likely to benefit. The results indicate that personalized screening programs combining genetic risk markers and age have the potential to greatly reduce the number of individuals screened, while still detecting nearly as many cases (Frampton et al. Ann Oncol

2015). However, real-world 'road testing' is needed to establish true efficiency and to address the technical and operational issues involved with genetic testing.

Resolving CRC heterogeneity

The clinicians managing CRC patients are facing several challenges; one of these being that colorectal cancer is a surprisingly heterogeneous disease entity. Patients with apparently similar tumors may experience very different disease courses and their tumors may respond very differently to the same treatment. In SYSCOL we generated detailed molecular information on a large number of CRC tumors and established a mathematical framework for interpreting the information. Having this information and framework allowed us to develop a tool for subclassifying CRC tumors into five molecular and clinically homogenous subtypes. We showed that the prognosis of the subtypes were significantly different, indicating that they should not be managed and treated in the same way. Moreover, our findings indicated that the subtypes will likely also respond differently to the same treatment. The established tool represents the best way of resolving the heterogeneity of CRC available today. In the future determining the subtype of a given patients tumor will likely become an essential part of deciding which treatment strategy to apply. We expect the benefit of the subtypes to be confirmed in clinical trials within near future.

Development of tools for blood based minimally-invasive detection and monitoring of CRC tumor burden

It is well known that solid tumors (including CRC) release DNA fragments into the blood circulation. Blood samples represent an attractive source of tumor DNA, because they can be drawn with minimal risk to the patients. Moreover, they can be drawn serially over time making them highly suited for longitudinal surveillance analysis. Tumor DNA found in blood is called circulating tumor DNA (ctDNA). To detect tumor DNA in a blood sample one needs to be able to distinguish it from the normal DNA. In SYSCOL we have performed extensive genomic characterizations on a large number of tumors, which has enabled us to identify tumor specific genomic alterations that can be used as markers of ctDNA. Furthermore, we have developed a number of different approaches for measuring these markers in patient blood samples. We also have developed approaches for screening blood samples for signs of tumor DNA without having any prior knowledge of specific tumor markers. These latter approaches open up the possibility of screening for CRC using blood samples, rather than the faecal samples as used today. Potentially, such a change could lead to an increased participation in CRC screening (today only 60% of invitees choose to participate) as investigations indicate that screening participants prefer blood tests over faecal tests (Lerary et al. Sci Transl Med 2012, Bettegowda et al. Sci Transl Med 2014)

By analysing blood samples for ctDNA we have been able to monitor CRC patients longitudinally during treatment to assess how their tumors responded to treatment. We have also used this approach to monitor patients following tumor surgery. For a subset of patients we identified ctDNA in the blood after surgery and all these patients later experienced a clinical recurrence of disease, indicating that detection of ctDNA is a strong predictor of future disease relapse. On average we were able to detect relapsing disease 10 months before the relapse was clinically acknowledged using standard follow-up procedures. Consequently, ctDNA analysis has the potential to provide the patients and clinicians with a critical window of opportunity for treating the relapse at an early time-point where it is still possible to cure the patient. By analysing blood samples we have also been able to identify tumor specific genomic alterations that can be targeted by specific treatments. In the future this opens up the possibility that the treatment choice could be determined based on a simple blood analysis.

The tools we have developed for minimally-invasive detection of tumor DNA, are not restricted to CRC but can also be applied to other cancer types. The potential clinical uses of the tools are many, including CRC screening, estimating prognosis after surgery, early detection of relapsing disease, tracking resistance to therapy, and detecting mutations that can be therapeutically targeted. The results are very promising and though much remains to be done in order to enable clinical usage of the developed minimally-invasive tools, we are confident that the tools are poised to radically change the way we manage patients with cancer.

Identification of novel genes and mechanisms involved therapeutic response

In SYSCOL we have also established a series of laboratory model systems of human CRC. Using these models we have been able to evaluate the impact of genetic variants (normal occurring as well as tumor specific) on how tumor cells respond to specific treatments. In this way we have identified a number of novel mechanisms of therapy resistance (Bertotti et al. Nature 2015). We have furthermore used the same models to provide proof of principle demonstration that personalized treatment, meaning targeting the specific genetic alterations found in the cancer cells, can have a very strong antitumor effect. We have furthermore been able to show that combination therapies were often more effective than single-agent treatments. The models represent unique tools for assessing the effect of novel treatments before testing them on humans.

In summary

Work package 8 has been very successful in translating the many findings from the SYSCOL project into a range of potentially useful clinical tools. We have developed tools with the potential to be used for: optimizing CRC screening, resolving CRC heterogeneity, improving predictions of disease course and response to therapy –guiding choice of treatment strategy, non-invasive detection of CRC and longitudinal monitoring of tumor burden – monitoring response to therapy and early detection of relapsing disease, detecting novel mechanisms of therapy resistance, and assessing the effect of novel therapies. All in all, the SYSCOL project has provided a significant amount of results that have the potential to improve the future treatment and survivability of patients with CRC.

Potential impact including socio-economic impact and main dissemination activities and exploitation of results

Socio-economic impact

The identification of new high-risk cancer genes often has clear, immediate benefits for families with cancer. In our case, owing to the SYSCOL work, *POLE* and *POLD1* were added to the list of around 12 high-risk genes for CRC predisposition. It therefore became possible to test families with many people who had CRC, as well as patients with early-onset CRC or multiple polyps, for these genes as a routine part of the diagnosis and investigation of such patients by Clinical Geneticists. If a disease-causing mutation is found, other family members are usually tested to see whether they carry the gene and therefore have a high CRC (and endometrial) cancer risk. If they are gene carriers, they can be offered regular screening with the aim of reducing the risk to a level close to that of the general population. Non-carriers can be reassured that their cancer risk is similar to that of the general population.

Although commercial providers offer tests for common, low-risk CRC SNPs, such testing has been slow to catch on in the regulated clinical setting. There is a good reason for that, namely that the proportion of total risk accounted for by such factors is low. It follows that whilst the risk of CRC for any individual who does not have one of the high-risk mutations can be estimated by SNP testing, the true risk lies within such wide limits that the estimate is almost meaningless. If we find many more CRC SNPs or other genetic or non-genetic risk factors, those estimates may eventually improve.

However, a more promising use for SNP testing is in the setting of cancer screening programmes, where most people are screened unnecessarily (in theory, 100% of the population is screened whereas only 5-6% will ever develop CRC and only half of those will die from it). Here, targeting screening to groups of people at higher risk based on their genetics has the potential to improve screening, by directing screening to those most likely to benefit. The number of individuals to be screened will then be greatly reduced, while nearly as many cases will still be detected.

The risk predictor tool we developed has the long term potential to be used to assess the genetic risk profile of a given individual. Although the predictor is still in its early development and needs further evaluation and testing before clinical use, genetic profiles can potentially be used to optimize the efficiency of population-wide screening programs.

We have also developed a range of tools for guiding the choice of oncological therapy, and for assessing the effect of therapy. Such tools have the potential to improve the effectiveness of therapy, to save the patients the unnecessary side-effects of ineffective treatments, and at the same time reduce health-care costs as the number ineffective treatments will be reduced. We have also developed non-invasive blood based tools that have the potential to optimize the post-operative management of CRC patients, both in terms of identifying who is likely to benefit from adjuvant chemotherapy, and by earlier detection of relapsing disease than is possible today. The latter potentially provides a window-of-opportunity for curatively-intended intervention, which is only rarely possible in the treatment of relapsing disease today. We foresee that the knowledge and tools developed in work package 8 will lead to improved treatment and survivability of CRC patients and at the same time make the clinical management of CRC more cost-efficient

The vast amount of patient derived molecular and clinical data that we have generated and integrated into the CRC model has significantly increased our understanding of the mechanisms behind colorectal cancer. This model can be used to find novel, druggable therapeutic targets that will have a long term impact on human health.

The Systems biology tools and computational models developed within the project have lead to a better understanding of the complex network of genes and regulatory systems behind colorectal cancer. As genetic variation in regulatory regions have been shown to be important in predicting the risk of individuals to develop most common diseases, we expect that the methods and tools developed and shared via the SYSCOL website, will be widely applicable to the study of other complex disorders.

The project will thus have broader implications on research, prevention and treatment of all disorders that depend on multiple genomic events.

Main dissemination activities and exploitation of results

Scientific publications and press releases

The work within the project has resulted in major scientific achievements with clear medical importance. The mutual contribution of all partners has generated approximately 56 peer-reviewed publications, nine of which are in top-journals such as *Cell, Science, Nature and Nature Genetics*. Many of the publications are joint efforts of several collaborating project partners.

Press releases for new discoveries emanating from the SYSCOL project have been published by the individual Partners institutions, on the SYSCOL website, on the individual Partners web pages, on Horizonhealth.eu and on CORDIS. Links to the publications and press events are available via the SYSCOL website. SYSCOL results have also been presented by partners at conferences, scientific and clinical meetings and workshops,

Systems biology data, tools and models shared via open access

To make the data resources and computational tools available for other researchers working in similar fields of technological applications, they have been uploaded on open-access web-based services. Links to these Systems biology tools and models are also available via the SYSCOL webpage, under the "links" tab. Much of the generated project data has been deposited in different repositories and is available through links in the respective publications.

Organisation of courses and workshops

To share project generated knowledge and expertise, several courses and workshops have been arranged, both centrally and by individual partners. The project management team has arranged four Systems biology symposia that were open to the general public. At these events SYSCOL partners presented their project work and invited experts from similar scientific fields to share their knowledge and experience. To strengthen the links with other FP7 networks using Systems Biology approaches to cancer medicine, representatives from some of these projects were invited to present their work and discuss common interests at two of the Symposia. We have also co-organised a Nobel conference in Stockholm, a student/postdoc workshop in Greece and a High Throughput Biology meeting in Stockholm.

Clinical and industrial networks

Several of the project partners are part of large, functional clinical networks through which the project results have been disseminated. In addition, the clinical sample collection in **WP2** and the clinical validation in **WP8** have led to the establishment of additional networks of clinicians with whom the project partners collaborate. This extensive network helped to introduce the genetic testing for POLE/POLD1 germline mutations and will help to translate the findings generated within the project to the clinical setting.

Our small/medium enterprise partner geneXplain, who is leading in regulatory genomics and Systems Biology areas in Europe, provide a comprehensive platform of bioinformatics, cheminformatics and Systems Biological Tools for personalized medicine and pharmacogenomics. geneXplain have been using their substantial industrial network to reach out to industry and investigate the possible exploitation of SYSCOL results.

Public awareness and participation

Raising public awareness of our research and showing how within this collaborative project we have generated excellent scientific achievements that have increased our understanding of colorectal cancer, and how our results have the potential to improve the clinical management of colorectal cancer has been central. The management team and partners have used a variety of sources to reach the general public.

The SYSCOL website is a dedicated project website that functions as the main external communication tool for the project: http://syscol-project.eu
The webpage contains general project information in the form of an introduction to colorectal cancer aimed at the general public, a press room with links to press releases and findings, links to SYSCOL publications, computational tools and resources developed within the project and information about SYSCOL related events and meetings. A *Partner* page contains information for each of the SYSCOL partners, their affiliations and contact details and a link to their own web pages, where they describe their contributions to research.

Some of the findings within the SYSCOL project have attracted major global and local News channels, including Daily Newspapers, TV news and local radio. Project results have also been featured via YouTube videos. These sites are extremely valuable for disseminating the research performed within SYSCOL to the general public, as they are sites that people use every day and therefore they do not have to search for information and it is disseminated to a very wide audience.

Public website

http://syscol-project.eu/