MSR1 repeats modulate gene expression and affect risk of breast and prostate cancer

A.M. Rose,[1] A. Krishan,[2] C.F. Chakarova,[1] L. Moya,[3,4] S. Chambers,[5] M. Hollands,[6*] J.C. Illingworth,[6*] S.M.G. Williams,[6*] H.E. James,[7] A.Z. Shah,[1] C.N.A. Palmer,[8] A. Chakravarti,[9] J.N. Berg,[7] J. Batra,[4] S.S. Bhattacharya[1]

[1] UCL Institute of Ophthalmology, University College London, London, UK

[2] Cell Therapy and Regenerative Medicine, CABIMER, Seville, Spain

[3] Australian Prostate Cancer Research Centre – Queensland, Translational Research Institute, Brisbane, 4102, Australia

[4] Cancer Program, School of Biomedical Sciences, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, 4102, Australia

[5] Allied Health Research, Menzies Health Institute Queensland, Griffith University, Brisbane, 4111, Australia

[6] UCL Medical School, University College London, London, UK

[7] Clinical Genetics, Ninewells Hospital & Medical School, University of Dundee, Dundee, UK

[8] Centre for Pharmacogenetics and Pharmacogenomics, Ninewells Hospital and School of Medicine, University of Dundee, Dundee, UK

[9] Johns Hopkins University School of Medicine, Institute of Genetic Medicine, Baltimore, USA

*The indicated authors contributed equally to this work

Correspondence: Dr Anna M. Rose,

UCL Institute of Ophthalmology, Bath Street,

London EC1V 9EL, U.K.

E-mail: anna.rose@ucl.ac.uk

Tel: (+44) 20 7608 6800

ABSTRACT

Background: MSR1 repeats are a 36-38bp minisatellite element that have recently been implicated in the regulation of gene expression, through copy number variation (CNV).

Patients and methods: Bioinformatic and experimental methods were used to assess the distribution of MSR1 across the genome, evaluate the regulatory potential of such elements and explore the role of MSR1 elements in cancer, particularly non-familial breast cancer and prostate cancer.

Results: MSR1s are predominately located at chromosome 19 and are functionally enriched in regulatory regions of the genome, particularly regions implicated in short-range regulatory activities (H3K27ac, H3K4me1, and H3K4me3). MSR1-regulated genes were found to have specific molecular roles, such as serine-protease activity (P=4.80x10^{-7}) and ion channel activity (P=2.7x10^{-4}). The kallikrein locus was found to contain a large number of MSR1 clusters, and at least six of these showed CNV. An MSR1 cluster was identified within *KLK14*, with 9-copies and 11-copies being normal variants. A significant association with the 9-copy allele and non-familial breast cancer was found in two independent populations (P=0.004; P=0.03). In the white British population, the minor allele conferred an increased risk of 1.21 to 3.51-times for all non-familial disease, or 1.7 to 5.3-times in early-onset disease. The 9-copy allele was also found to be associated with increased risk of prostate cancer in an independent population (odds ratio = 1.27-1.56; P =0.009).

Conclusions: MSR1 repeats act as molecular switches that modulate gene expression. It is likely that CNV of MSR1 will affect risk of development of various forms of cancer, including that of breast and prostate. The MSR1 cluster at *KLK14* represents the strongest risk factor identified to date in non-familial breast cancer and a significant risk factor for prostate cancer. Analysis of MSR1 genotype will allow development of precise stratification of disease risk and provide a novel target for therapeutic agents.

KEY WORDS

MSR1, minisatellite, repeat element, gene expression, gene regulation, kallikrein genes, breast cancer, prostate cancer.

KEY MESSAGE

MSR1 repeats are widespread in the genome and copy number variation (CNV) of the element acts as a molecular switch to control gene expression. MSR1 CNV is likely to control expression of >200 genes and drive dysregulation at the kallikrein locus and other cancer-related genes. CNV of MSR1 at the kallikrein locus is an important risk factor for non-familial breast cancer and prostate cancer.

INTRODUCTION

The concept of “junk DNA” arose from an observation: that DNA content is poorly correlated with the complexity of an organism; for example, the onion genome is approximately five times greater than human. To resolve this paradox, geneticists suggested that most DNA has no biological purpose and could be considered “junk”. This dogma has been challenged recently by studies such as the ENCODE project, which estimated that up to 80% of the genome has function [1,2]; although this has been considered by some to be an over-estimate [3,4]. ENCODE has, nevertheless, focussed attention on possible functions for “junk DNA” – including work on repeat elements, such as minisatellites.

Chromosome 19 is unusually rich in repetitive DNA and one repeat element – the 36-38bp minisatellite sequence, MSR1 – has been reported to be specific to this chromosome [5,6]. It was shown that copy number variation (CNV) of MSR1 regulates gene expression, with CNV of MSR1 repeats located upstream to the *PRPF31* promoter affecting transcription of the downstream gene [7]. A cluster of MSR1 elements was identified within the murine *troponin I* (*TnIc*) promoter, and deletion alters *TnIc* expression [8,9]. Furthermore, MSR1 sequence might regulate expression of an aberrant *KLK4* sense-antisense chimera transcript in prostate cancer cells [10].

The emerging influence of CNV of MSR1 on gene expression implicates a role in oncogenesis and, in this work, the regulatory potential of MSR1 repeats was explored, with focus on the oncogenic kallikrein locus.

METHODS

Bioinformatic methods

The MSR1 consensus sequence was identified from Repbase [11]. MSR1 family membership and sequence boundaries were determined by RepeatMasker, using the consensus sequence of the family. A Hidden Markov Model (HMM) profile was constructed using multiple sequence alignment and an extensive whole genome search for profile HMM was performed using nHMMER with strict profile threshold ($E=1x10^{-8}$) [12]. The analysis followed the standard pipeline for transposon-related repeat families [13].

RepeatMasker annotations track for MSR1 family was probed for intersection with several regulatory tracks of ENCODE using Table Browser tool of UCSC genome browser (http://genome.ucsc.edu/) [14]. The regulatory data from 7 cell lines assayed in ENCODE was analysed (GM12878, K562, H1 human embryonic stem cells, HSMM, HUVEC, NHEK and NHLF). These cell lines were used as there is full, robust data of H3K marks and other regulatory signals.

MSR1 repeats were probed *in silico* for their putative effect on genes using three models of gene regulation: (i) Basal-plus-extension; (ii) Two-nearest-gene; (iii) Single-nearest-gene (bioinformatic pipelines available on request).

The list of genes that were possibly regulated by MSR1 repeat elements was analysed to detect functional clustering of protein domain (InterPro domains, https://www.ebi.ac.uk/interpro/), Molecular Signature Database (MSigDB, http://software.broadinstitute.org/gsea/msigdb/index.jsp), DAVID gene ontology (https://david.ncifcrf.gov/) and GREAT software (http://bejerano.stanford.edu/help/display/GREAT/) [15].

Analysis of gene dysregulation in malignancy and the kallikrein locus

Gene dysregulation data from 20 malignancies was obtained from COSMIC database (http://cancer.sanger.ac.uk/cosmic). Data for CNV of MSR1 at the kallikrein locus was obtained through UCSC genome browser (https://genome.ucsc.edu). Specifically, data was analysed for intersection of MSR1 sequence on the RepeatMasker track, with CNV data on the DGV structural variation track.

Vector preparation and dual-luciferase reporter assay

Non-labelled PCR product was amplified using template DNA from a control individual. A modified pGL3-basic vector (Promega) was used, containing a basic thymidine kinase promoter (pTK), and the fragments were cloned upstream to the TK sequence. Dual-luciferase reporter assay was performed in MCF7, RPE-1 cells and HeLa cells by transient transfection, as previously described [**Error! Bookmark not defined.**].

Patient selection

Breast cancer - Peripheral blood DNA was obtained from Tayside Biorepository, these being samples from White-British patients with a diagnosis of primary BrCa without family history suggestive of familial inheritance. The control samples were women from the same ethnic population with no history of BrCa, obtained from Generation Scotland 3D resource or ECACC control panels [16]. Written, informed consent was obtained from all individuals and local ethics approval was gained from the NHS Research Ethics Committee for Scotland A (REC reference number: 06/

MRE00/105).

Prostate cancer - A total of 905 prostate cancer patients were genotyped. From these, 600 patients were recruited via collaborations with The Cancer Council Queensland (ProsCan study) and 305 samples were provided by patients who donated blood to the QLD node of the Australian Prostate Cancer BioResource (APCB) as detailed previously [17,18]. From the 842 healthy controls analysed, 334 were recruited through the Australian Red Cross Blood Services and 508 were enrolled through the Electoral Roll where age- and postal code-matched patients to the ProsCan study were selected [17]. All methods were carried out in accordance with relevant guidelines and regulations, and all experimental protocols were approved by QUT's Human Ethics Committee (Ethics' Approval number: 1000001171), the Australian Red Cross Services (Ethics' Approval number: 2004#17) and Cancer Council Queensland (Ethics' Approval number: 3629H). Only patients who provided informed written consent were included in the study.

Analysis of patient genotype

Breast cancer: In total, 633 patients with non-familial BrCa and 650 controls underwent analysis. PCR was performed using a FAM-labelled forward primer (5'FAM-GAAGCTGGATTGAGGAAACG) and an unlabelled reverse primer (GTGCCTCCGGTCTTGAGTAG). PCR products underwent a standard genotyping reaction and were analysed on ABI 3130x. Data was analysed using Data Collection v30 and Gene Mapper v4.1.

Prostate cancer: A touchdown PCR was optimised using the GoTaq Green Master Mix (Promega, Sydney, Australia). The first 10 cycles of the PCR used an annealing temperature of 63^{o}C while the remaining 30 cycles were performed at 62^{o}C. The rest

of the conditions for both series where identical, with an initial denaturation temperature of 95°C for 30 seconds and a final elongation temperature of 72°C for 45 seconds. A subsequent 10 minutes at 72°C finalised the protocol. The primers used were the same as for breast cancer samples. MSR1 copy number was assessed on agarose gels using a sizing ladder.

Statistical analysis

Breast cancer: Due to small sample size, ▪ ^{2} testing was applied to compare the risk of disease in those carrying only the major allele (homozygous 11 copies), as compared to those carrying minor alleles (8-, 9-, or 10-copies). The odds ratio was estimated using conditional maximum likelihood estimate (CMLE). Dose-response analysis was performed by an Extended Mantel-Haenszel χ-square for linear trend. Sub-type analysis was performed for age of onset and histological sub-type. Statistical analysis was performed using OpenEpi v3.03 (www.openepi.com).

Prostate cancer: Association of MSR1 was analysed for prostate cancer risk using univariate binary logistic regression (IBM SPSS Statistics; 23.0) where the dependant variable is the case-control status. A $P < 0.05$ was considered significant, and OR and 95% confidence interval (CI) are shown. For genotype association analysis the most common allele, homozygous 11/11, was used as a reference. Random sampling with replacement tests has been carried out using the bootstrapping analysis (IBM SPSS Statistics 23.0) using a seed value of 1,000 samples 1,000,000 times. To confirm the results were not age-related, both the univariate binary logistic regression and the bootstrapping were age-corrected. The allele/genotype was used as the categorical value and the age was the categorical co-variate. The prostate cancer/control status was the unique dependant variable.

RESULTS

MSR1 repeat elements are distributed across the genome, with highest density on chromosome 19q

The non-redundant (canonical) MSR1 sequence was found to have 978 hits across the human genome (**Figure 1A**, **Supplementary Table S1**, **Supplementary Figure S1A-C**). Additionally, there were >2000 further instances of degenerate sequence genome-wide (**Supplementary Figure S1D**). The canonical HMM sequence had the highest level of conservation, and was observed most frequently (>550 hits) on chromosome 19 (**Figure 1A**). There were also a reasonable number of occurrences on chromosome 7 (110 occurrences) and chromosome 1 (44 occurrences). All other chromosomes had a small number of occurrences (<30), but MSR1 was not evident on the mitochondrial genome (**Supplementary Table S1**).

MSR1 repeats coincide with genomic markers of gene regulation

To investigate if MSR1 repeats are a global regulator of transcription, it was considered whether the elements coincided with genomic markers of regulation (**Figure 1B**). It was found that 70% of MSR1 elements are located within open chromatin and, furthermore, a large proportion of MSR1 repeats were associated with H3K27ac (70%), H3K4me1 (83%) and/or H3K4me3 (85%). MSR1s were less commonly associated with DNaseI hypersensitivity sites, transcription factor binding sites (TFBS), DNA methylation marks or FAIRE signal.

Serine-protease genes and ion channel genes are putatively regulated by MSR1 elements

Having identified that the majority of MSR1 sequence coincided with markers of gene regulation, it was considered important to study which genes might be regulated by the repeat, and to see whether these genes shared any common features.

Only MSR1 repeats with highest fidelity were studied at this time, as it was felt that these were most likely to be functional. This highlighted 227 genes, mainly on chromosome 19, which are prime candidates for regulation by MSR1 repeats (**Supplementary Figure S2**). The genetic distance between the MSR1 repeat element and transcription start sites (TSS) of putatively regulated genes was then studied, showing that the vast majority of MSR1 elements were located within 50kB of TSS, with a significant proportion located very close to the gene TSS (<5kB) (**Figure 1C-H**). By the basal-plus-extension method, 41.0% were located between 0-5kB of TSS and 52.9% were located at a distance of 5-50kB; by the two-nearest-gene method, the respective figures were 31.8% and 60.6% respectively and for the single-nearest-gene method, 58.9% and 40.5%. To analyse whether MSR1 repeat elements were enriched in regions close to gene TSS, 5 random 50kB regions containing MSR1 elements were chosen at random.

Gene ontology analysis of the functional clustering of genes putatively regulated by MSR1 elements showed enrichment for a number of molecular functions (**Supplementary Table S2**). In particular, MSR1-containing genes were strongly enriched for serine peptidase activity ($P<4.8\times10^{-7}$) and ion channel activity ($P=2.7\times10^{-4}$) (**Figure 2A**). This was confirmed on analysis of InterPro signatures,

which showed strong enrichment for peptidase domains (P=1.5x10^{-9}) and voltage-gated ion channels (P=3.2x10^{-3}) (**Figure 2B**).

Genes containing MSR1 clusters are frequently dysregulated in cancer

The 227 genes that were identified as candidates for regulation by MSR1 repeats were analysed for dysregulation in 20 common malignancies. Gene expression data in malignancy was available for all but 13 of the genes. This showed that all the analysed genes putatively controlled by MSR1 were over-expressed in at least 5% of some cancer types, and 50 of the genes were dysregulated in a large proportion of at least one tumour type (≥15%) (**Figure 3, Supplementary Table S3**). Under-expression of genes regulated by MSR was far less common, with only 78 of the genes being reported as under-expressed, and only 3 genes being reported as frequently under-expressed (**Supplementary Table S4**).

The kallikreins are a family of serine-proteases encoded by a series of tandem-array genes located at chromosome 19q13.4 – totalling 15 genes and one pseudo-gene (**Figure 4A**). Gene ontology analysis had indicated a strong functional clustering for serine-peptidases, due to functional clustering of 15 genes (*CAPN1*, *HPN*, *KLK1*, *KLK10*, *KLK13*, *KLK14*, *KLK15*, *KLK2*, *KLK3*, *KLK4*, *KLK6*, *KLK7*, *KLK8*, *KLK9*, *PSENEN*). Further analysis of the kallikrein locus mapped a large number of MSR1 clusters (**Table 1**). Analysis of deep-sequencing data showed that at least six of the clusters harboured CNV in a control population, these clusters lying closest to *KLK4*, *KLK7*, *KLK14* and *KLK15*.

Copy number variation of the *KLK14* MSR1 cluster alters gene transcription *in vitro*

A cluster of MSR1 repeats located within the 3'UTR of *KLK14* was selected for further study (**Figure 4B**). This cluster was chosen because it demonstrated CNV in previously published population data and *KLK14* has associations with breast and prostate cancer.

The cluster of interest was amplified from control individuals of Northern European descent, demonstrating four alleles (8-, 9-, 10- or 11-copies). The 11-copy allele was the major allele, but 9-copy was reasonably frequent (MAF = 13.3%). The 8- and 10-copy alleles existed at <1% frequency and were considered rare variants.

The two common alleles (11- or 9-MSR1 copies) were cloned upstream to a basic promoter in both forward and reverse strand orientation, and dual luciferase reporter activity was tested in MCF7, HeLa and RPE-1 cell lines (**Figure 4C-E**, **Supplementary Table S5 and S6**). It was demonstrated that the presence of MSR1 elements enhanced basal transcription, but that the 9-copy allele had a significantly greater effect on reporter activity than the 11-copy allele.

In the forward strand orientation, the 9-copy allele (KLK14-9) had 4.9 to 9.1-fold induction over pTK; similarly, in the reverse strand orientation, KLK14-9 had 5.0 to 6.9-fold induction over pTK. In contrast, the 11-copy allele (KLK14-11) had a lesser enhancing effect on transcription, with 2.5 to 5.5-fold induction (forward) or 3.0 to 3.5-fold induction (reverse). The difference between 9-copy and 11-copy alleles was statistically significant for both forward and reverse strand orientations in all three cell lines (P<1x10-5). It can, therefore, be considered that the observed polymorphism

causes a functional change and that the 9-copy allele drives relatively higher expression of *KLK14* than the 11-copy allele.

Copy number variation of MSR1 is a major risk for non-familial breast cancer

As a functional role of the MSR1 CNV was demonstrated, and *KLK14* dysregulation is associated with BrCa, a case-control analysis was performed for non-familial disease (**Figure 4F-K**, **Supplementary Table S7**). The risk of developing BrCa was studied by calculating the dose response to the 9-copy *KLK14* allele; that is, the risk of carrying heterozygous exposure or homozygous exposure, as compared to baseline exposure of no rare alleles (11,11 homozygotes).

There is a strong stratified association between the number of 9-copy alleles and the relative risk of BrCa (**Table 2**). Heterozygotes (9,11) have 1.21-times higher risk than 11-copy homozygotes, whilst 9-copy homozygotes have a 3.51-times higher risk (MH $\chi 2 = 8.25$; $P=0.004$).

The observed effect was strongest for individuals with early-onset BrCa (less than 50 years), with heterozygous carriers having 1.65-times higher risk and, remarkably, homozygous carriers having 5.34-times increased risk (MH $\chi 2 = 10.71$; $P=0.001$). The trend did not reach statistical significance for patients with onset of disease at 50 years and older (MH $\chi 2 = 2.98$; $P=0.08$). The association was significant for the estrogen receptor (ER) positive subtype, with 1.22 to 3.42-times increased risk in heterozygous and homozygous carriers, respectively (MH $\chi 2 = 5.62$; $P=0.02$). The association between the 9-copy allele and ER-negative histological subtype did not reach significance.

There were insufficient individuals carrying the rare alleles (8- or 10-copy) and so it was not possible to perform a dose-response analysis. Instead, ▪ ^{2} analysis was performed; this did not show significant association between the rare alleles and risk of BrCa (*P* >0.05).

The association was replicated by re-analysis of previously published data [19]. Although only a small number of cases and controls were available (24 of each group), analysis of dose-response replicated the association of the 9-copy allele with non-familial BrCa in a second, independent population (*P* =0.03, **Table 2**).

Copy number variation of MSR1 is a major risk for prostate cancer

The association between MSR1 CNV at the KLK14 locus and PrCa was then assayed in an Australian case-control cohort, as a further replication and validation of the association between MSR1 CNV and malignancy. The two most common alleles were the 9 and 11 repeats, the 11-copy being the major allele (11-copy allele frequency = 79.6%; 9-copy allele frequency = 17.1%). Interestingly, a greater variety of rare alleles were observed in the Australian population: 6-copies (0.2%), 8-copies (1.3%), 10-copies (1.2%), 12-copies (0.4%) and 13-copies (0.1%) (**Supplementary Table S8**).

The 9-copy allele was shown to be significantly associated with prostate cancer risk at allele level (OR = 1.3, CI = 1.1 – 1.6, *p* =0.001) and genotype level (OR = 1.3, CI = 1.04 – 1.6, *p* =0.001). The results were age independent and confirmed by bootstrapping (**Table 2**, **Table 3**). The major 11-copy allele was associated with a protective effect (OR = 0.65, CI = 0.5 – 0.9, p =0.025) and the results were also age-independent and confirmed by bootstrapping (**Table 3**).

DISCUSSION

The ENCODE project predicted that many "junk DNA" sequences, including repetitive elements, would have a molecular function. MSR1 is a minisatellite sequence that was previously considered chromosome 19 specific [**Error! Bookmark not defined.**]. Recently, MSR1 repeat elements were implicated in regulation of gene expression – with CNV of MSR1 at the *PRPF31* locus modulating expression of the upstream gene [**Error! Bookmark not defined.**].

In this work, bioinformatic methods were used to establish the distribution of MSR1s within the human genome, to explore whether MSR1 acted as a global regulator of gene expression, and to investigate how regulation of gene expression by MSR1 influences malignancy.

First, this work has shown that the MSR1 sequence is not exclusively located on chromosome 19 – but is found on all chromosomes, except the mitochondrial genome. Most MSR1 repeat elements were found to be located close to gene TSS – this being suggestive of a role in short-range regulation. Furthermore, more than 70% of MSR1 elements intersect with H3K marks. H3K marks tend to be associated with short-range enhancers; as opposed to FAIRE signal, which is associated with long-range elements and with which there was little intersection. This is supportive of previously published work that looked at the putative mechanism of action of MSR1 [7]. This work showed that spatial relation of the MSR1 cluster to the promoter sequence was critical for the effect of CNV and that the element did not bind transcription factors, nor was it a target for epigenetic regulation [7]. It was proposed, therefore, that 3D DNA conformational change (e.g. DNA looping) was the mechanism by which MSR1 affected gene transcription.

Next, gene ontology and functional clustering analyses showed that MSR1 sequence chiefly regulated genes with ion-channel and serine-peptidase function, including the kallikrein locus. Dysregulation of protease activity is one of the key events implicated in tumour growth and progression, both at primary and metastatic disease sites [20]. Our analysis found that serine-protease genes were strongly enriched for MSR1 elements ($P<4.80\text{x}10^{-7}$) due to clustering in the kallikrein locus. Dysregulation of kallikrein genes has been implicated in several malignancies; particularly, endocrine cancers (prostate, breast, ovarian) but also adenocarcinoma of the colon, lung and pancreas, neuroendocrine tumours, and acute lymphoblastic leukaemia [21,22]. It was found that at least four kallikrein genes harboured MSR1 CNV in a normal population: *KLK4*, *KLK7*, *KLK14* and *KLK15.* However, the structural variation data available for analysis was from a small sample of South Asian Malay individuals and is, as such, likely to be an under-estimate of true CNV at the kallikrein locus [23].

It was surmised that MSR1 CNV element might cause over-expression of *KLK14* in BrCa. Reporter assay showed that presence of the 3'UTR MSR1 cluster acted as an enhancer element. Critically, *in vitro*, the 9-copy allele induces a greater enhancement of transcription than the 11-copy allele. This was consistently demonstrated in three different cell lines – including MCF7, which is a BrCa cell line. If the same phenomenon occurred *in vivo*, it strongly suggests that the 9-copy allele induces higher levels of *KLK14* gene expression as compared to the 11-copy allele. It was postulated, therefore, that the 9-copy allele might be associated with higher risk of BrCa.

In our case-control cohort, it was shown that the 9-copy allele was associated with a significantly increased risk of disease. Increasing presence of the 9-copy allele was associated with a greater risk of BrCa: thus, as compared to 11-copy homozygotes,

the heterozygous carriers of the 9-copy allele had an odds ratio of 1.2 and, of greater influence, those homozygous for the 9-copy allele had an odds ratio of 3.5. This effect was even greater in those with early-onset disease, with a remarkably high effect (1.7 to 5.3-fold increased risk). Furthermore, it seems the risk (9-copy) allele is strongly associated with the ER positive histological subtype, but not the ER negative subtype.

In an association study, it is critical to show reproducibility in a second population. Here, we were able to re-analyse previously published data from a small cohort from Toronto, Canada [19]. Our analysis confirmed a statistically significant association between the 9-copy allele and risk of non-familial BrCa. Given the small *n*, however, caution would be advised in interpretation of the odds ratios generated from this analysis.

As high-penetrance alleles are mainly associated with familial forms of cancer, it is to be expected that a study of non-familial BrCa would display an allele with low- or moderate-penetrance cancer susceptibility. It is crucial to note, however, that with an odds ratio of 1.2-3.5 (1.7-5.3 in early-onset disease), the *KLK14*-MSR1 cluster is by far the most influential susceptibility allele identified to date in non-familial BrCa. Crucially, even the identification of this single risk allele with such a high influence is a major step towards stratification of risk of non-familial BrCa in women in the general population, as even this one locus could be clinically useful for a predictive test. However, it is predicted that study of other kallikreins and MSR1-containing loci would allow formulation of an even more powerful method for risk stratification in non-familial BrCa – this facilitating enrolment into effective screening and early prevention programmes for those at highest risk.

To further confirm the association of MSR1 CNV with malignancy, a case-control analysis was performed between the *KLK14*-MSR1 cluster and prostate cancer. This confirmed that the higher-expressing, 9-copy allele was a robust and significant risk factor for prostate cancer in the population studied. The 9-copy allele was associated with prostate cancer risk at allele level (OR = 1.3; p =0.001) and genotype level (OR = 1.3; p =0.001), these results being age-independent and confirmed by bootstrapping. Furthermore, the major 11-copy allele was associated with a significant protective effect (OR = 0.65; p =0.025).

The association between the *KLK14* higher-expressing allele and risk of prostate cancer makes logical sense, as *KLK14* expression is significantly higher in cancerous compared to non-cancerous prostatic tissue (**Figure 3**, [24]). Furthermore, up-regulation of *KLK14* is observed in advanced and more aggressive tumours, and has been associated with disease progression [24,25].

It would be valuable to assess genotype of all MSR1 clusters across the kallikrein locus in a case-control cohort of patients with breast, prostate and other endocrine-related malignancies. Subsequently, MSR1 elements outside the kallikrein locus could be studied, in malignancy and other diseases associated with these genes. Furthermore, we predict that hypermutation of the MSR1 clusters might occur within tumour cells, propagating further gene dysregulation. Finally, protease inhibitors are an emerging class of chemotherapeutic agents and, accordingly, pharmaceutical targeting of MSR1 might represent a new anti-cancer chemotherapeutic approach.

In summary, this study provides persuasive evidence that MSR1 repeats are global regulators of gene expression and that this prototypical “junk DNA” element can influence many important human diseases. We anticipate that study of MSR1 will

allow development of tools for screening, diagnosis and prognostication for many diseases, and hope that long-term this will allow early intervention to predict and prevent devastating diseases, such as breast, prostate and other cancers.

ACKNOWLEDGEMENTS

The authors are grateful to the individuals, both cases and controls, who consented to take part in this research study. Thanks is also given to Fiona Carr and Beverley Scott for technical assistance, and to Professor Geoffrey E. Rose and Dr Channa Jayasena for critical reading of the manuscript. The authors are grateful to the Australian Prostate Cancer BioResource members who contributed to the collection and pathological review of the biospecimens and associated annotated data: J.Clements, P.Saunders, A.Eckert, P.Heathcote, G.Wood, G.Malone, H.Samaratunga, A.Collins, M.Turner and K.Kerr.

FUNDING

This project received no specific funding.

DISCLOSURE

AMR holds patent number PCT/GB2014/050107.

REFERENCES

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57-74.

2. Ecker JR, Bickmore WA, Barroso I, et al. Genomics: ENCODE explained. Nature. 2012; 489:52-5.

3. Doolittle WF. Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci USA. 2013; 110:5294-300.

4. Graur D, Zheng Y, Price N, et al. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013; 5:578-90.

5. Grimwood J, Gordon LA, Olsen A, et al. The DNA sequence and biology of human chromosome 19. Nature. 2004; 428:529-35.

6. Jurka J, Walichiewicz J, Milosavljevic A. Prototypic sequences for human repetitive DNA. J Mol Evol. 1992; 5:286-91.

7. Rose AM, Shah AZ, Venturini G, et al. Transcriptional regulation of PRPF31 gene expression by MSR1 repeat elements causes incomplete penetrance in retinitis pigmentosa. Sci Rep. 2016; 6:19450.

8. Bhavsar PK, Brand NJ, Yacoub MH, Barton PJ. Isolation and characterization of the human cardiac troponin I gene (TNNI3). Genomics. 1996; 35:11-23.

9. Cullen ME, Dellow KA, Barton PJ. Structure and regulation of human troponin genes. Mol Cell BioChem. 2004; 263:81-90.

10. Lai J, Lehman ML, Dinger ME, et al. A variant of the KLK4 gene is expressed as a cis sense-antisense chimeric transcript in prostate cancer cells. RNA. 2010; 16:1156-66.

11. Jurka J, Kapitonov VV, Pavlicek A, et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005; 110:462-7.

12. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013; 29:2487-9.

13. Wheeler TJ, Clements J, Eddy SR, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nuc acids res. 2013; 41:D70-D82.

14. Karolchik D, Hinrichs AS, et al. The UCSC Table Browser data retrieval tool. Nuc acids res. 2004; 32:D493-D496.

15. McLean CY, Bristor D, Hiller M, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat biotechnol. 2010; 28:495-501.

16. Kerr SM, Liewald DC, Campbell A, et al. Generation Scotland: Donor DNA Databank; A control DNA resource. BMC Med Genet. 2010; 11:166.

17. Batra J, Lose F, O'Mara T, et al. Association between Prostinogen (KLK15) genetic variants and prostate cancer risk and aggressiveness in Australia and a meta-analysis of GWAS data. PloS one. 2011; 6:e26527.

18. Lose F, Srinivasan S, O'Mara T, et al: Genetic Association of the KLK4 Locus with Risk of Prostate Cancer. PloS one. 2012; 7:e44520.

19. Yousef GM, Bharaj BS, Yu H, Poulopoulos J, Diamandis EP. Sequence analysis of the human kallikrein gene locus identifies a unique polymorphic minisatellite element. Biochem Biophys Res Commun. 2001; 285(5):1321-9.

20. Eatemadi A, Aiyelabegan HT, Negahdari B, et al. Role of protease and protease inhibitors in cancer pathogenesis and treatment. Biomed Pharmacother. 2017; 86:221-231.

21. Borgoño CA, Michael IP, Diamandis EP. Human tissue kallikreins: physiologic roles and applications in cancer. Mol Cancer Res. 2004; 2:257-80.

22. Emami N, Diamandis EP. New insights into the functional mechanisms and clinical applications of the kallikrein-related peptidase family. Mol Oncol. 2007;1:269-87.

23. Wong LP, Ong RT, Poh WT, et al. Deep whole-genome sequencing of 100 southeast Asian Malays. Am J Hum Genet. 2013; 92:52-66.

24. Yousef GM, Stephan C, Scorilas A, et al. Differential expression of the human kallikrein gene 14 (KLK14) in normal and cancerous prostatic tissues. Prostate. 2003;56:287-92.

25. Rabien A, Fritzsche F, Jung M, et al. High expression of KLK14 in prostatic adenocarcinoma is associated with elevated risk of prostate-specific antigen relapse. Tumour Biol. 2008;29:1-8.

TABLE LEGENDS

TABLE 1

Location and orientation of MSR1 clusters within the kallikrein locus, with the nearest gene and the copy number variants identified in a previously published database of whole genome deep-sequencing in a control population. The clusters with reported CNV in a test population are highlighted in grey.

Table 2 Univariate binary logistic regression for the three most common genotypes of the MSR1 cluster located in the 3'UTR of *KLK14* in either heterozygous (9,11 genotype) or homozygous form (9,9 genotype), as compared to the major allele homozygotes (11,11 genotype

The relative risk of BrCa is calculated as the Extended Mantel-Haenszel χ^2 for linear trend with odds ratio. [a]OR Calculated using a binary logistic regression, [b]bootstrap (two-tailed). The 11/11 repeats was used as reference for genotype analysis (IBM SPSS Statistic Processor; [b]P value calculated by Extended Mantel-Haenszel χ2 for linear trend; *% with respect all the genotypes observed (see Supplementary Table S8 for full values); CI: confidence of interval.

Table 3

Results of association between the *KLK14* MSR1 CNVs and prostate cancer in a large case-control cohort. Association was calculated using (a) binary logistic regression, (b) bootstrap (two-tailed), (c) binary logistic regression (age corrected), (d) bootstrap (two-tailed) age corrected. The 11/11 repeats was used as reference for genotype analysis (IBM SPSS Statistic Processor; 23). OR: odds ratio; ns: not significant; CI: confidence interval (see Supplementary Table S8 for full values).

FIGURE LEGENDS

FIGURE 1

(**A**) The distribution of MSR1 repeat elements across the genome, shown as occurrences (hit count) per 1 Mb of chromosome.

(**B**) The coincidence of MSR1 repeat elements, by percentage, with recognised genomic marks of regulatory potential. The assays analysed were: (i) chromatin accessibility assays and chromatin interaction assays, (ii) layered H3K27Ac mark on 7 cell lines from ENCODE, (iii) layered H3K4Me1 mark on 7 cell lines from ENCODE, (iv) layered H3K4Me3 mark on 7 cell lines from ENCODE, (v) digital DNaseI Hypersensitivity Clusters from ENCODE, (vi) transcription levels assayed by RNA-seq on 7 Cell Lines from ENCODE and ChIP Transcription Factor ChIP-seq from ENCODE, (vii) formaldehyde-assisted isolation of regulatory elements (FAIRE)-seq (viii) DNA methylation assays. Data from regulation supertracks (composite regulatory signals from 7 cell lines (GM12878, K562, H1 HESC, HSMM, HUVEC, NHEK, NHLF) was integrated into all 8 analytic sub-groups, thus highlighting generality of these regulatory signals.

Relative position (**C**) and absolute distance (**D**) to gene TSS for genes detected by "basal-plus-extension" methods; relative position (**E**) and absolute distance (**F**) to gene TSS for genes detected by "two-nearest-genes" method; and relative position (**G**) and absolute distance (**H**) to gene TSS for genes detected by "single-nearest-gene" method. In (C), (E) and (G), the x-axis distances are 1: <-500kB, 2: -500 to -50kB, 3: -50 to -5 kB, 4: <-5kB, 5: 0-5kB, 6: 5-50kB, 7: 50-500kB, 8: >500kB.

FIGURE 2

Functional clustering analysis of the genes putatively regulated by MSR1 repeats showed important molecular functions (**A**), protein domains (**B**) and pathways (**C**) that are enriched.

Statistical significance of enrichment denoted by asterisk – ****P*=<0.001, ***P*=<0.01, **P*=<0.05.

FIGURE 3

Analysis of 20 common malignancies for over-expression of 214 genes putatively controlled by MSR1 repeat elements. Colour-coded for percentage of tumour samples reported to over-express the gene: beige 1-4.9%, yellow 5-9.9%, orange 10-14.9%, red 15-19.9%, maroon 20-29.9%, purple 30-39.9%, navy ≥40%.

FIGURE 4

(**A**) Schematic representation of the kallikrein cluster at chromosome 19q13, showing relative gene position and size (intergenic distances not to scale). Arrows represent gene orientation. (**B**) UCSC genome browser data, showing structure of *KLK14*, with genomic co-ordinates (hg19), exons and MSR1 repeat elements clusters (turquoise, the cluster assayed is arrowed).

Results of dual-luciferase reporter assay in MCF7 cells (**C**), HeLa cells (**D**) and RPE-1 cells (**E**). Data presented as the average ratio of pGL3-insert to pTK, error bars represent one standard deviation; the positive control (pTK) and negative control (pGL3) are also shown. Orientation of MSR1 sequence indicated by + (forward strand) or – (reverse strand).

Cluster	Start	End	+/-	Size	Nearest gene	Gene position	Deletions?	Insertions?	Inversions?
1	50838533	50838945	-	414	KLK15	Promoter	Y	N	N
	50838964	50839068	-	108					
	50839086	50839216	-	138					
2	50839388	50839641	-	255	KLK15	Promoter	Y	N	N
3	50858321	50858405	-	111	KLK3	Intron	N	N	N
4	50906457	50906862	-	411	KLK4	3' UTR	Y	N	N
5	50968950	50969313	-	373	KLK5	Promoter/intron 1	N	N	N
6	50974768	50975026	-	237	KLK7	3' UTR	N	N	N
	50975045	50975743	-	652					
7	50979979	50980231	-	256	KLK7	Intron	N	N	N
8	50982481	50983661	+	1204	KLK7	5' UTR	Y	N	Y
9	50999193	50999292	-	106					
10	51000301	51000392	+	108	KLK8	Intron	N	N	N
	51001293	51001473	-	175					
11	51001710	51002157	-	440	KLK8/KLK9	Promoter	N	N	N
12	51003325	51003415	+	92	KLK9	Intron	N	N	N
	51003447	51003619	+	172					
13	51015634	51015849	+	210	KLK10	Intron	N	N	N
14	51077561	51077973	-	412	KLK14	3' UTR	Y	N	N
15	51078986	51079439	+	462	KLK14	Intron	Y	N	N

Table 2

	MSR1 genotype	Exposure level	Cases (%)*	Controls (%)*	Total	Odds Ratio[a] (95% CI)	MH $\chi 2$	*P*-value[b]
All cases (BrCa)								
	11,11	Level 0 vs 0	418 (66)	470 (72)	888	1		
	9,11	Level 1 vs 0	177 (28)	164 (25)	341	1.21 (0.94 – 1.56)		
	9,9	Level 2 vs 0	25 (4)	8 (1)	33	3.51 (1.6 – 7.9)	8.25	0.004
Age of onset								
<50 years	11,11	Level 0 vs 0	66	470	536	1		
	9,11	Level 1 vs 0	38	164	202	1.65		
	9,9	Level 2 vs 0	6	8	14	5.34	10.71	0.001
≥50 years	11,11	Level 0 vs 0	248	470	718	1		
	9,11	Level 1 vs 0	95	164	259	1.1		
	9,9	Level 2 vs 0	13	8	21	3.08	2.98	0.08
Histological subtype								

ER -	11,11	Level 0 vs 0	58	470	528	1		
	9,11	Level 1 vs 0	25	164	189	1.24		
	9,9	Level 2 vs 0	4	8	12	4.05	2.73	0.10
ER +	11,11	Level 0 vs 0	206	470	676	1		
	9,11	Level 1 vs 0	88	164	252	1.22		
	9,9	Level 2 vs 0	12	8	20	3.42	5.62	0.02
Replication set (BrCa)								
	11,11	Level 0 vs 0	6	17	23	1		
	9,11	Level 1 vs 0	17	5	22	9.63		
	9,9	Level 2 vs 0	1	2	3	1.42	4.49	0.034
All Cases (PrCa)								
	11,11	Level 0 vs 0	538 (59)	546 (65)	1084	1		
	9,11	Level 1 vs 0	292 (32)	233 (28)	525	1.27 (1 – 1.6)		
	9,9	Level 2 vs 0	37 (4)	24 (3)	61	1.56 (0.9 – 2.7)	6.77	0.009

Table 3

Genotype	Controls (n) (%)	Cases (%)	OR (95% CI)[a]	p-value[a]	p-value[b]	p-value[c]	p-value[d]
	1 (0.1)	0	-	-	-	-	-
	1 (0.1)	0	-	-	-	-	-
	1 (0.1)	0	-	-	-	-	-
	6 (0.7)	2 (0.2)	-	-	-	-	-
	14 (1.7)	9 (1)	-	-	-	-	-
	24 (3)	37 (4)	-	ns	-	-	-
	1 (0.1)	0					
	233 (28)	292 (32)	1.3 (1.04 – 1.6)	0.021	0.018	0.029	0.023
	8 (1)	17 (1.9)	-	-	-	-	-
	3 (0.4)	7 (0.8)	-	-	-	-	-
	546 (65)	538 (59)	Reference	-	-	-	-
	1 (0.1)	1 (0.1)	-	-	-	-	-
	3 (0.4)	0	-	-	-	-	-
	0	2 (0.2)	-	-	-	-	-
	3 (0.2)	0	-	-	-	-	-
	22 (1.3)	11 (0.6)	-	-	-	-	-
	288 (17)	368 (20)	1.3 (1.1 – 1.6)	0.001	0.003	0.003	0.004
	20 (1.2)	41 (2)	-	-	-	-	-
	1344 (80)	1385 (77)	0.65 (0.5 – 0.9)	0.025	0.025	0.042	0.041
	6 (0.4)	0	-	-	-	-	-
	1 (0.1)	5 (0.3)	-	-	-	-	-

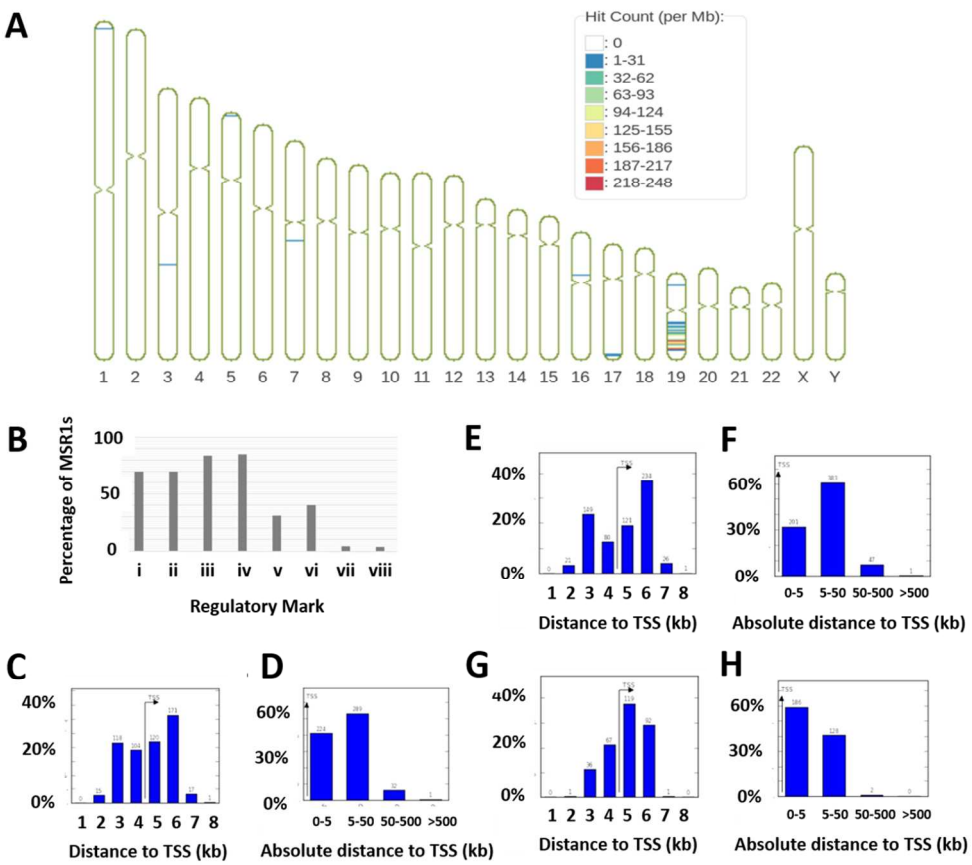

Figure 1

334x297mm (96 x 96 DPI)

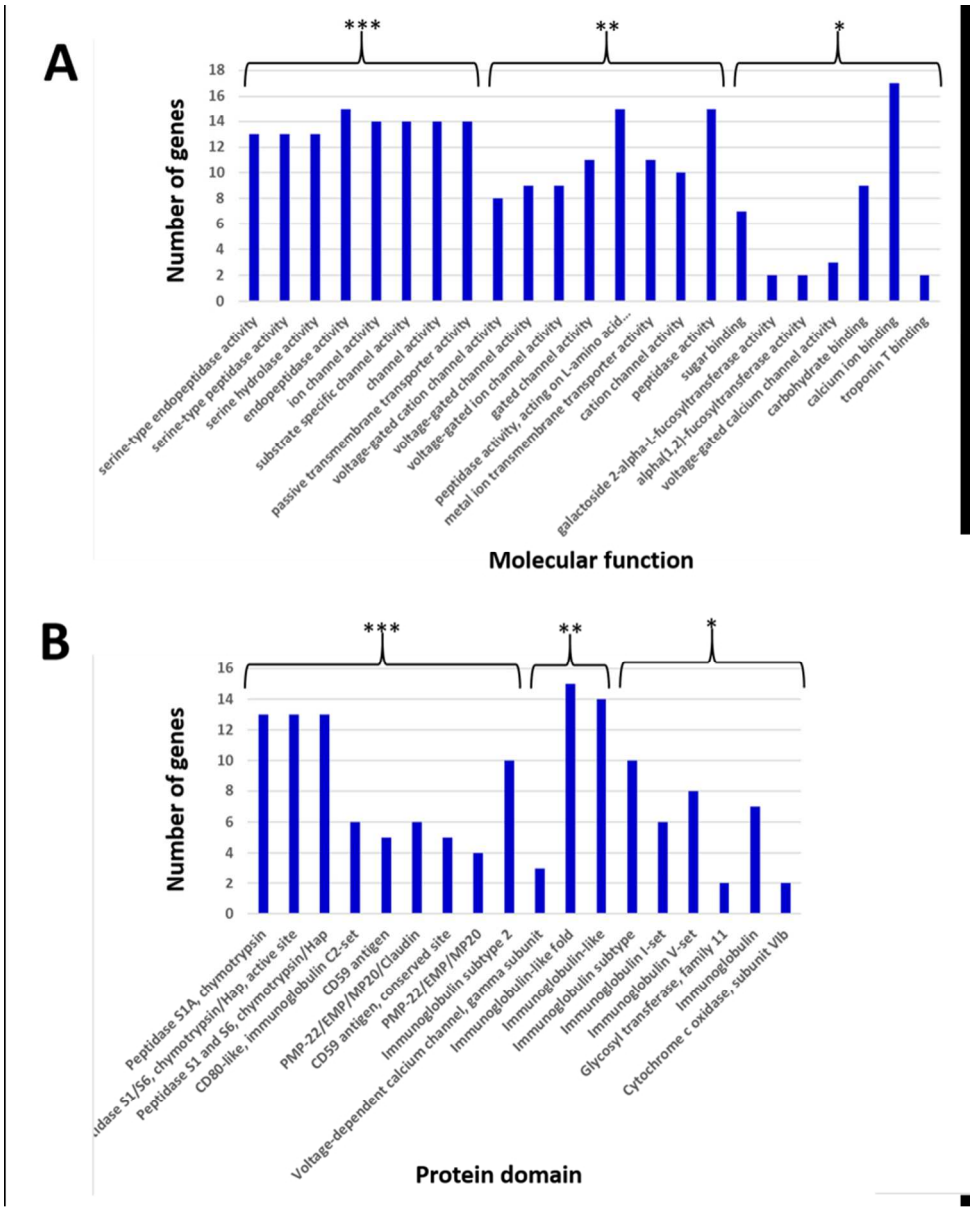

Figure 2

239x297mm (96 x 96 DPI)

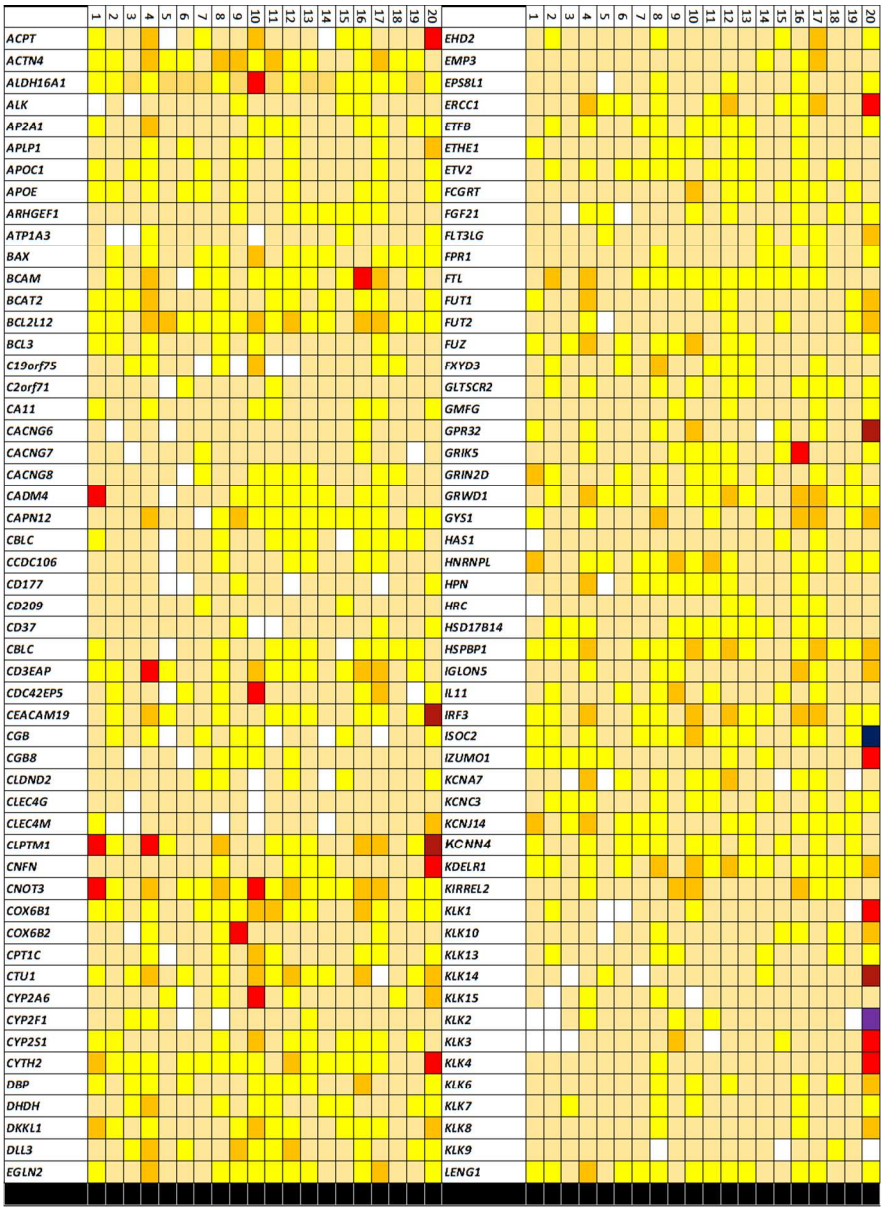

Figure 3

297x407mm (96 x 96 DPI)

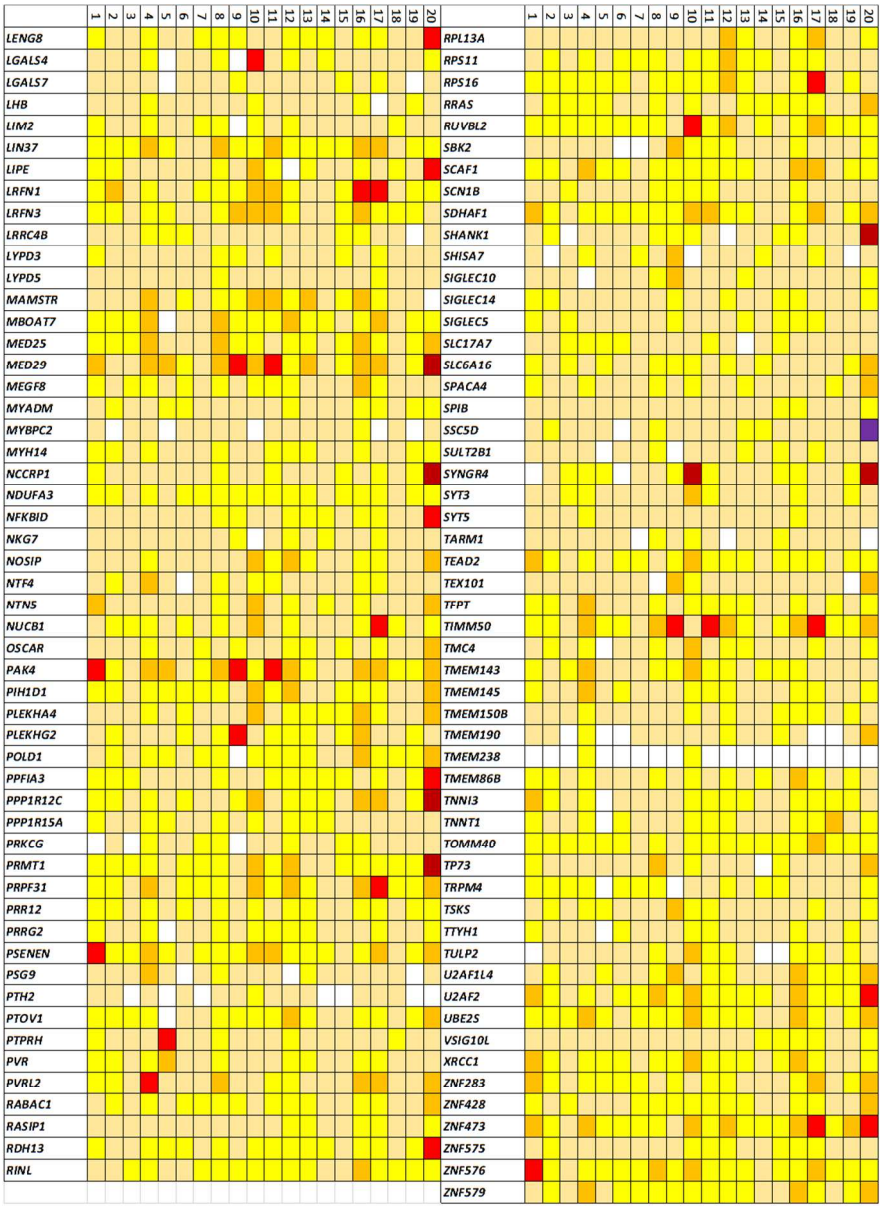

Figure 3-2

298x408mm (96 x 96 DPI)

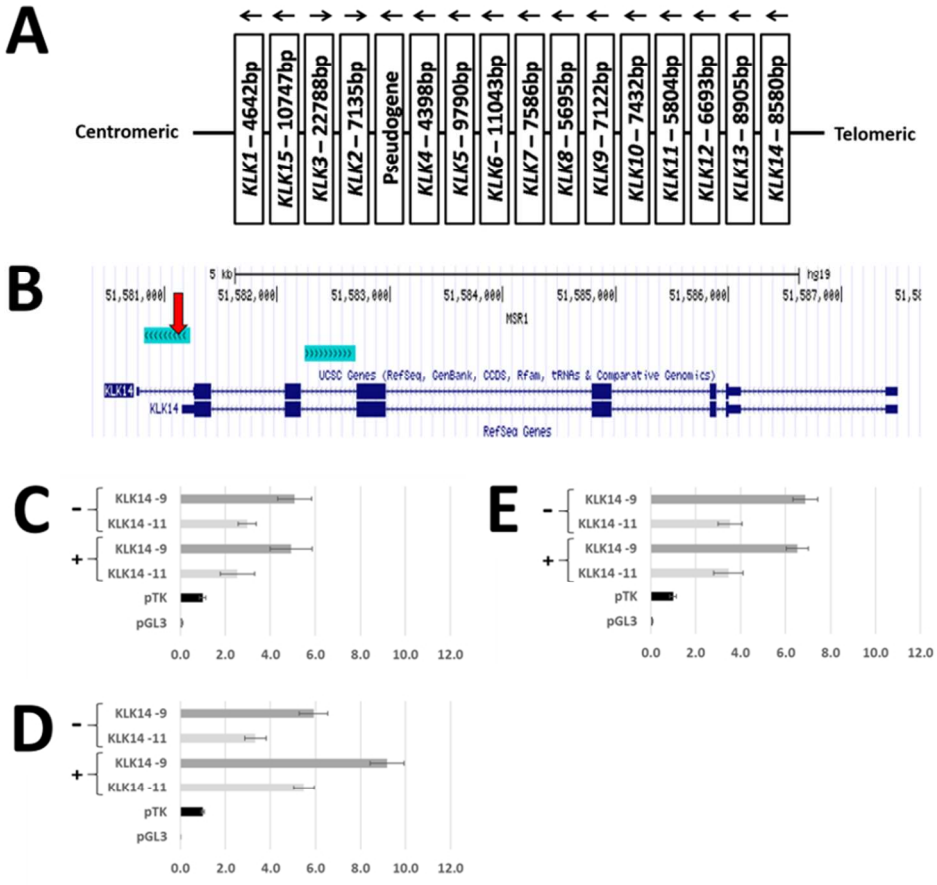

Figure 4

248x230mm (96 x 96 DPI)